UNIVERSITÀ
DEGLI STUDI
DI PADOVA

TESI DI DOTTORATO

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Chimiche

CORSO DI DOTTORATO DI RICERCA IN: SCIENZE MOLECOLARI
CURRICOLO: SCIENZE CHIMICHE
CICLO XXXVI

Structural characterization of different proteins as potential drug targets

Tesi redatta con il contributo finanziario della Fondazione Cariparo

Coordinatore: Ch.mo Prof. Stefano Corni
Supervisore: Ch.mo Prof. Roberto Battistutta

Dottorando : Emanuele Fornasier

I would like to dedicate this thesis to my loving wife and my beloved children.

Abstract

Structure determination of proteins is key to understanding their function and mechanism of action. In this thesis the structural characterization of three proteins by X-ray crystallography is described.

The human Aryl hydrocarbon Receptor (AhR) is an important transcription factor which has been known for a long time to be involved in several regulatory processes and immune-system adaptation. Its first structural characterizations only emerged in very recent years as its stability in solution is precarious. Despite our numerous attempts to produce a soluble construct, it was not possible to conduct any structural characterization of AhR given its very low stability in solution.

The Main Protease (Mpro) of SARS-CoV-2 is an essential enzyme for the virus maturation and replication. We extensively characterized this protein both in its wild-type variant and in several mutants developed in our laboratory with the aim to better understand its enzymatic mechanism and binding dynamics. A novel conformation of the wild-type variant of Mpro and several structures of mutants in presence of inhibitors and endogenous substrates are hereby described.

The first experimental structure of murine Threonine Aldolase (Tha-1) enzyme was determined in presence and in absence of its PLP cofactor. These structures were important for the validation of a reverse-docking algorithm and provide the first structural evidence of a mammalian Threonine Aldolase enzyme.

Table of Contents

List of Figures

List of Tables

Part 1

The Aryl Hydrocarbon Receptor (AhR)

Introduction

AhR and the bHLH-PAS transcription factors

The Aryl hydrocarbon Receptor (AhR) is a transcription factor of the basic Helix-Loop-Helix Per-ARNT-Sim (bHLH-PAS) family. bHLH-PAS transcription factors are proteins that share a common architecture including – starting from the *N*-terminus – a bHLH DNA binding domain, two tandemly positioned PAS domains (PAS-A and PAS-B) and a highly variable, protein-specific transactivation or transrepression domain (TAD)[1,2]. These transcription factors are widely spread among vertebrates and invertebrates and are involved in several regulatory processes. Hypoxia-inducible factor 1α (HIF-1α), for example, is involved in angiogenesis and other physiological responses to low oxygen levels. CLOCK (Circadian locomotor output cycles kaput) instead plays a central role in circadian rhythm regulation. bHLH-PAS proteins can be divided in two classes (I and II, Table 1): all proteins belonging to Class I must form heterodimers with a member of Class II to be transcriptionally active. Furthermore, protein complexes that include ARNT adopt quaternary motifs that are structurally different from the ones that include BMAL[1].

Table 1. Classes of the bHLH-PAS transcription factor family.

Class I	Class II
HIF-1α	ARNT (HIF-1β) ARNT2
HIF-2α	
HIF-3α	

AhR	
AhRR	
NPAS1	
NPAS3	
NPAS4	
SIM1	
SIM2	
CLOCK	BMAL1 (ARNTL) BMAL2 (ARNTL2)
NPAS2	

bHLH transcription factors recognize an hexanucleotide sequence CANNTG (known as Ephrussi-Box, or E-box) where N can be any nucleotide[3]. In the transcription factor dimer, each monomer recognizes the first or the second half of the E-box. Substrate affinity is determined by direct DNA binding of the first two, *N*-terminal alpha helices of the bHLH domain, while the other two, following helices are involved in protein dimerization.

The Aryl hydrocarbon receptor (AhR) is expressed in a variety of tissues across the human body. As its name suggests, AhR mainly binds to polycyclic aromatic hydrocarbons (PAHs), like dioxin or tryptophan metabolites. Historically known as an environmental sensor (it was known in the past as *dioxin receptor*), it is involved in the modulation of the immune response and in several inflammatory diseases[4]. The PAS-A domain of all bHLH-PAS proteins in involved in the recognition and the dimerization of the two factors that form a given complex. In the case of AhR, the PAS-B domain is the environmental sensor itself, as it contains the ligand binding pocket.

AhR is mainly located in the cell cytosol in an inactive form in association with other proteins, with which it forms the receptor complex[5]. Such complex is constituted by the co-chaperone p23, XAP2 (also known as AhR-interacting protein, AIP), two molecules of the chaperone Hsp90 (Heat-Shock Protein 90 kDa), and the protein kinase SRC (Figure 1). Upon ligand entry in the cytosol and successive binding with the PAS-B domain of AhR, the protein complex enters the cell nucleus. AhR then dissociates from its associated proteins and dimerizes with the Aryl hydrocarbon Receptor Nuclear Translocator (ARNT) protein. Successively, the AhR-ARNT transcriptionally active complex promotes transcription for a variety of genes, most of which are involved in PAHs oxidation like CYP1A1 of the cytochrome P450 superfamily.

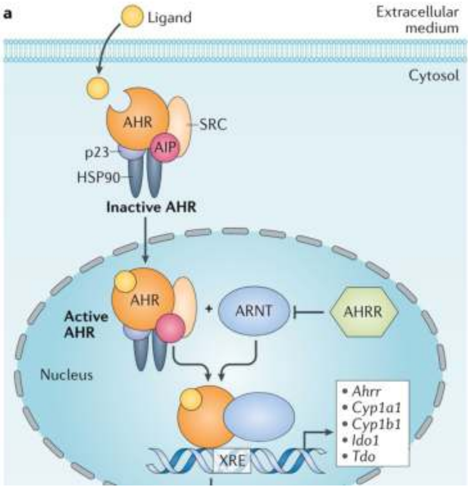


Figure 1. Activation mechanism of AhR. Upon ligand entry in the cytosol and successive binding with the PAS-B domain of AhR, the protein complex enters the cell nucleus. AhR then dissociates from its associated proteins and dimerizes with the Aryl hydrocarbon Receptor Nuclear Translocator (ARNT) protein. Successively, the AhR-ARNT transcriptionally active complex promotes transcription for a variety of genes. Image from Rothhammer *et al*[5].

Some bHLH-PAS transcription factors have been characterized by X-ray crystallography, alone and in complex with their DNA responsive element (DRE)[1,2,6]. In all available structures, one or more domains among bHLH and PAS are resolved, while the intrinsically disordered TAD domain has not been structurally characterized for any transcription factor so far. In the case of AhR two, independent crystallographic studies were published in 2017 by Schulte *et al.* and by Seok *et al.*[7,8]. In both cases the resolved structure included the bHLH and PAS-A domains of AhR in complex with the corresponding domains of ARNT and the DNA responsive element (Figure 2). AhR crystal structures published in 2017 both lack the PAS-B domain. Schulte *et al.* explicitly mentioned in their work that *"Since all attempts to produce a soluble AhR construct including the PAS-B domain failed, we co-expressed truncated human AhR and mouse ARNT constructs without the PAS-B and TAD".*[7] Since the AhR PAS-B domain contains the ligand binding pocket that activates the receptor complex in the cytoplasm and promotes the AhR-related transcription program, the knowledge of its 3D structure is crucial for the comprehension of the activation mechanism and for the rational development of future drugs targeting AhR. Since different molecules can act either as an agonist or an antagonist toward AhR, a considerable amount of research has been dedicated in elucidating the key features that determine ligand-induced effects on AhR transcription. Two computationally-derived homology models of the PAS-B domain have been published in 2007 and 2011 by the Bonati group[9.10]. The two models are based on *apo* and *holo* HIF-2α experimental structures, respectively.

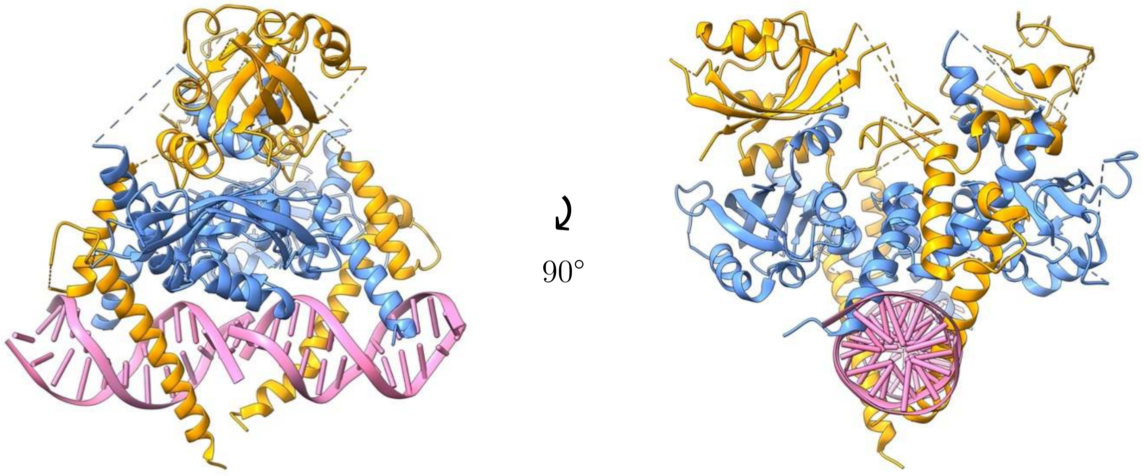


Figure 2. Structure of the AhR-ARNT complex interacting with the DRE. Side view (on the left) and front view (right) of AhR (in blue) in complex with ARNT (orange) and DRE (pink). On the front view, the two PAS-A domains of AhR and ARNT are located on the upper left side, while the two bHLH domains are located on the right side, interacting with the DNA. Structure from Schulte et al.[7] PDB ID: 5NJ8.

HIF-2α is another member of the bHLH-PAS superfamily and it is the protein with the highest sequence identity and similarity compared to AhR (31% and 62%, respectively). More recently, Denison *et al.* exploited site-directed mutagenesis to pinpoint key residues in the PAS-B domain which are important in substrate recognition and agonist/antagonist discernment[11]. The most important breakthrough came in 2022, when it was reported the first structure of the (partially) incomplete cytosolic AhR complex by cryogenic electron microscopy (Cryo-EM)[12]. Such structure was the first experimental structure containing the PAS-B domain of human AhR, but lacked its PAS-A, bHLH and TAD domains, as well as the p23 co-chaperone (Figure 3). More recently, a new Cryo-EM structure was published with a more complete cytosolic AhR complex that included p23 but still only contained only the PAS-B domain of AhR[13].

In the same period of time, some new structures were reported of the PAS-B domain of AhR from *D. melanogaster* (Figure 4) by Dai and co-workers[14]. In their article, they screened the recombinant expression in *E. coli* of over ten AhR PAS-B homologs from different species (human and murine among them) but found that *"only the PAS-B domain of Drosophila […] is soluble"*. This is, to date, the only crystallographic structure of an AhR PAS-B domain, although it comes from an organism that is quite distant from the mammals from an evolutionary point of view. The sequence similarity and identity for such domain, when compared to the human homolog, are 67% and 44%, respectively.

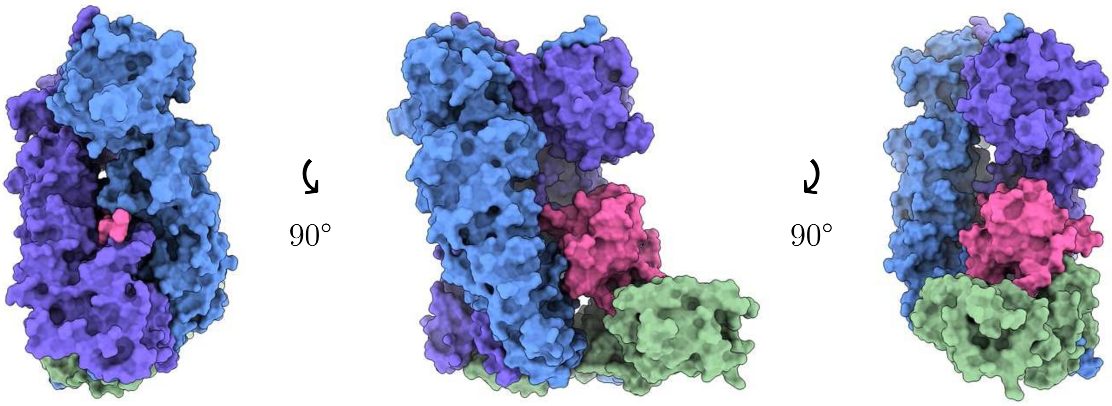


Figure 3. Cryo-EM structure of the AhR cytosolic complex. The PAS-B domain of AhR (in pink) is interwoven with the HSP90 dimer (in blue and purple) and with XAP2 (in green). Structure from Gruszczyk *et al.*[12] PDB ID: 7ZUB.

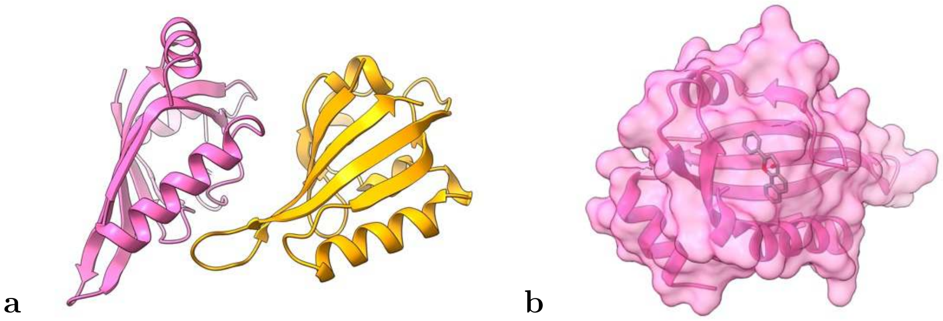


Figure 4. Structure of the PAS-B domain of *D. melanogaster*. (a) The PAS-B domain of *Drosophila* AhR forms a dimer with the PAS-B domain of murine ARNT (PDB ID: 7VNI). (b) α-Naphthoflavone is completely buried in the ligand binding pocket of *Drosophila* AhR (PDB ID: 7VNH).

AhR as a drug target

AhR is involved in homeostasis as well as in a wide variety of inflammatory and neoplastic disorders, and as such has recently emerged as a promising target for cancer therapy and other diseases[5,15]. It is the only known member of the bHLH-PAS family to be activated by ligand (*via* the binding to the PAS-B domain). It was also found to be particularly important in adaptive immunity, because it creates a direct link between external insults (both endogenous and exogenous, including pollutants) and immune system response[4]. A wide variety of AhR agonists and antagonists are known (Table 2), together with selective AhR modulators (SAhRMs) which are AhR agonists that do not show oestrogen-receptor affinity (which is instead common for the other AhR agonists).

Table 2. Compounds that affect AhR activity (from Murray *et al.*[15]).

Agonists	Xenobiotic	*Halogenated aromatic hydrocarbons:* • 2,3,7,8-tetrachlorodibenzo-*p*-dioxin • Dibenzofurans • Biphenyls • Polyaromatic hydrocarbons • 3-methylcholanthrene • Benzo[a]pyrene • Benzanthracenes • Benzoflavones *Pharmaceuticals:* • Tranilast • Leflutamide • Omeprazole
	Dietary	*Flavonoids:* • Quercetin • Galangin *Indoles:* • Indole-3-carbinol • 3,3′-diindoylmethane • Indolo[3,2b]carbazole
	Endogenous	• Kynurenic acid • Kynurenine • 6-formylindolo[3,2b]carbazole • Indoxyl sulfate
	Microflora	• Indirubin • 7-ketocholesterol • 3-methylindole • Trypthantrin • 1,4-dihydroxy-2-naphtoic acid • Malassezin

Antagonists	Xenobiotic	• 6,2′,4′,-trimethoxyflavone • GNF351 • CH-223191 • StemRegenin 1
	Dietary	• Resveratrol
Selective AhR modulators	Xenobiotic	• SGA360 • 3′,4′-dimethoxy-α-naphthoflavone • 6-methyl-1,3,8-trichlorodibenzofuran

AhR is involved in all stages of cancer development (initiation, promotion, progression and metastasis) but its role is emerging in increasing complexity. AhR overexpression is common in many tumor types, but the tumor development can be either favored or hampered by AhR antagonists (as it can be with agonists). Furthermore, *in vivo* studies are complicated by the fact that some ligands show different activities in mice compared to human cell lines[15]. The AhR-mediated response is dependent not only by ligand type, but it is cell-type specific and context-specific. On one side, the fact that AhR is involved in a plethora of metabolic pathways and malignant processes makes it a very attractive target for small-molecule drug design. On the other hand, its vast network of interactions and its ambivalent behavior towards different types of stimuli makes it hard to develop effective drugs with limited side-effects.

Final remark. At the beginning of my PhD thesis, the main scope of my thesis was the resolution of the PAS-B domain of AhR. The first Cryo-EM structure of the human PAS-B domain was published in late 2022[12], at the beginning of the third (and last) year of my PhD program. Our main focus for this project, however, did not change; we analyzed such structure and tried to extract some useful information to successfully create a new human PAS-B construct with potentially useful mutations and well-defined *N*- and *C*-termini that would show a good solubility. Such construct would be - ideally - easier to produce and to handle compared to the whole AhR cytosolic complex; a small and soluble AhR PAS-B construct may well behave in biophysical assays and could crystallize easily, enabling faster and high-throughput structural studies to be conducted.

Results and discussion

PAS-B-only constructs

Initial construct design was based on PAS-B structures available in 2019, in particular those regarding AhR most similar proteins: ARNT and HIF-2α. Supposing that AhR may suffer from low solubility and low expression levels, we chose the pET-SUMO expression vector. This expression system has many advantages compared to more classical expression plasmids of the pET series. It is a low-copy number, *lac*-promoted and kanamycin-resistant plasmid that is designed to produce chimeric proteins with an *N*-terminal 6xHistidine tag (HisTag), followed by the SUMO (Small Ubiquitine-like MOdifier protein) sequence and the desired protein sequence at the *C*-terminus. The main advantages of this expression system are the increased solubility of the chimeric protein and the possibility to cleave the 6His-SUMO moiety from the desired protein sequence at the exact end of the SUMO sequence (no amino acid linker is needed between SUMO and the desired protein). This precise cleavage is carried out by a highly specific enzyme (ULP-1 peptidase from *S. cerevisiae*, EC 3.4.22.68) which recognizes the 3-dimensional structure of SUMO and releases the desired protein without extra amino acids overhangs. For most bHLH-PAS proteins it was reported the co-crystallization of heterodimers composed of one PAS-B domain of Class I and one PAS-B domain of Class II. Since the PAS-B domain of murine ARNT was already well characterized and was known to crystallize well, we cloned its sequence from M354 to E470 (ARNTa1, Table 3). For AhR we chose to test the PAS-B domain of both murine and human origin. Constructs AhRa1 and AhRb1 comprise the canonical PAS-B domain of human, while AhRa2 and AhRb2 have the same *N*-terminus but include about 30 more amino acids at the *C*-terminus. This longer constructs were designed on the basis of secondary- and tertiary structure computational predictions with the Phyre2 engine[16], the Rosetta suite[17] and, later, by the Alphafold algorithm[18]. All algorithms converged in predicting that part or all the ~30 extra amino acids of the longer constructs would arrange in an α-helix ordered structure, thereby possibly leading to a more stable construct.

Table 3. PAS-B-only ARNT and AhR constructs

Code	Source	Sequence span	Vector
ARNTa1	murine	M354-E470	pET-SUMO
AhRa1	human	I280-E389	
AhRa2	human	I280-N418	
AhRb1	murine	I274-E383	
AhRb2	murine	I274-S412	

PAS-B-only AhR and ARNT constructs were cloned in linearized pET-SUMO plasmid with the TA cloning method (procedure details at p.77). This cloning method is fast and avoids the use of restriction enzymes. The desired DNA inserts are produced using appropriate primers and the Taq polymerase in a PCR reaction. The linearized pET-SUMO vector presents thymine overhangs at the 3' ends, while the insert presents 5' adenine overhangs because it is produced with the Taq DNA polymerase (Figure 5). The linearized plasmid and the T4 DNA ligase are then added to the PCR product yielding the complete, circular plasmid. The main drawback of this process is the need for a linearized plasmid with 5' thymine overhangs and the possibility on inserts being inserted in reverse direction.

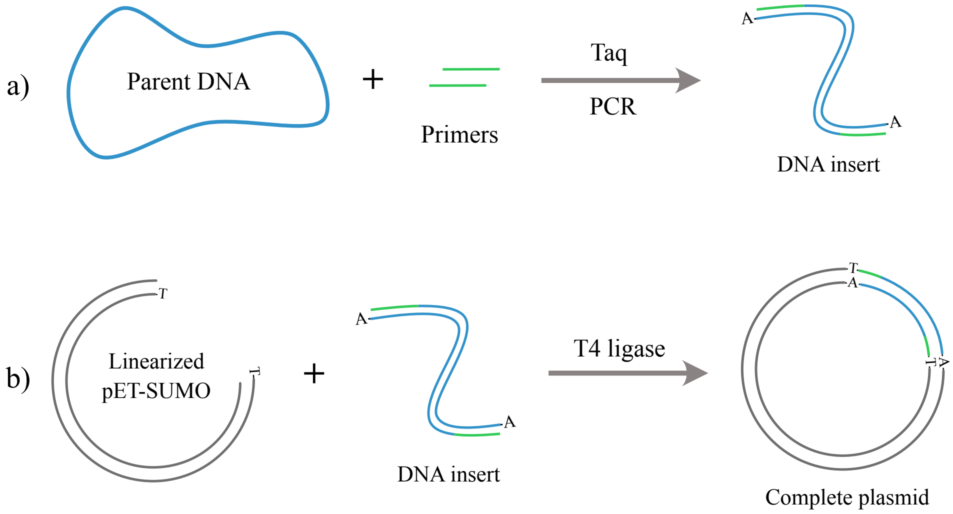


Figure 5. TA cloning procedure scheme. The DNA insert with adenine overhangs on 5' is obtained in a PCR reaction employing Taq polymerase. b) The DNA insert is added to the linearized pET-SUMO plasmid and ligated by T4 DNA ligase.

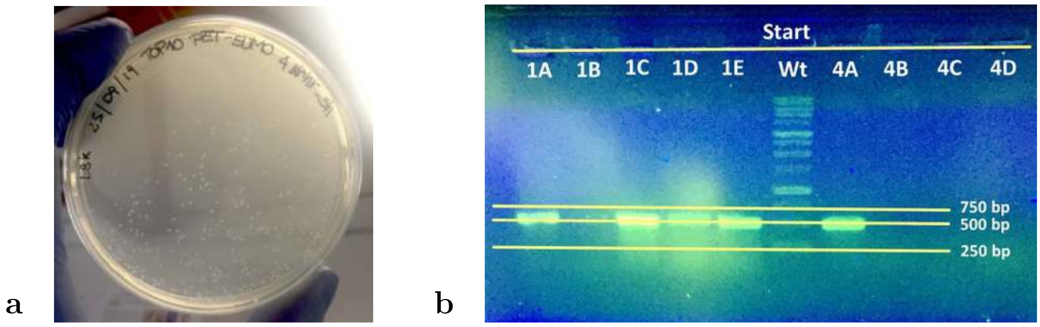


Figure 6. Selection of cloning results of the PAS-B-only constructs. a) Colonies visible on a Petri selection plate for AhRb1. b) Agarose gel of the colony PCR experiment with positive and negative samples.

At the end of the cloning procedure, the ligated product is transformed in cloning-grade, highly competent *E. coli* bacterial cells (TOP10 strain) with the heat shock method at 42 °C. Finally bacterial cells are plated on LB-agar Petri dishes with kanamycin resistance and incubated overnight at 37 °C (Figure 6, a). Bacterial colonies appear by the next day and are screened by colony PCR (procedure details at p.79, Figure 6, b). The colonies that test positive at the colony PCR procedure are then sequenced to confirm the correct insertion of the DNA insert in the destination vector. In the case of TA cloning, it is to be expected a 50% incidence of false positives in the colony PCR procedure as the DNA insert has the same probability of entering the destination vector in the correct or in the reverse direction.

The cloning procedure for all constructs displayed in Table 3 proceeded without major delays except for AhRa2, for which no colonies appeared at the end of the procedure. For such construct, it was determined that the annealing temperature of the PCR cycle had to be lowered from 55 ° to 50 °C (even though all the primers used in this experiment had the same theoretical melting temperature).

When all the constructs were confirmed positive, the plasmids were transformed in expression-grade *E. coli* cells (BL21(DE3) strain) and glycerol stocks were made to speed up subsequent experiments. Expression experiments were set up for all constructs (See details at page 81). Three main parameters have been screened:

- Optical density at time of induction (0.6 – 1)
- IPTG concentration (0.2 – 1 mM)
- Expression temperature (20 °C, 30 °C, 37 °C)

While ARNT expressed in very high yields under all conditions, the expression levels of all PAS-B-only AhR constructs were always low at 30 °C and 37 °C, while they were higher at 20 °C (Figure 7). Ideal IPTG concentration was found to be 0.5 mM, while the optical density at time of induction was not found to influence the expression levels significantly within the tested range.

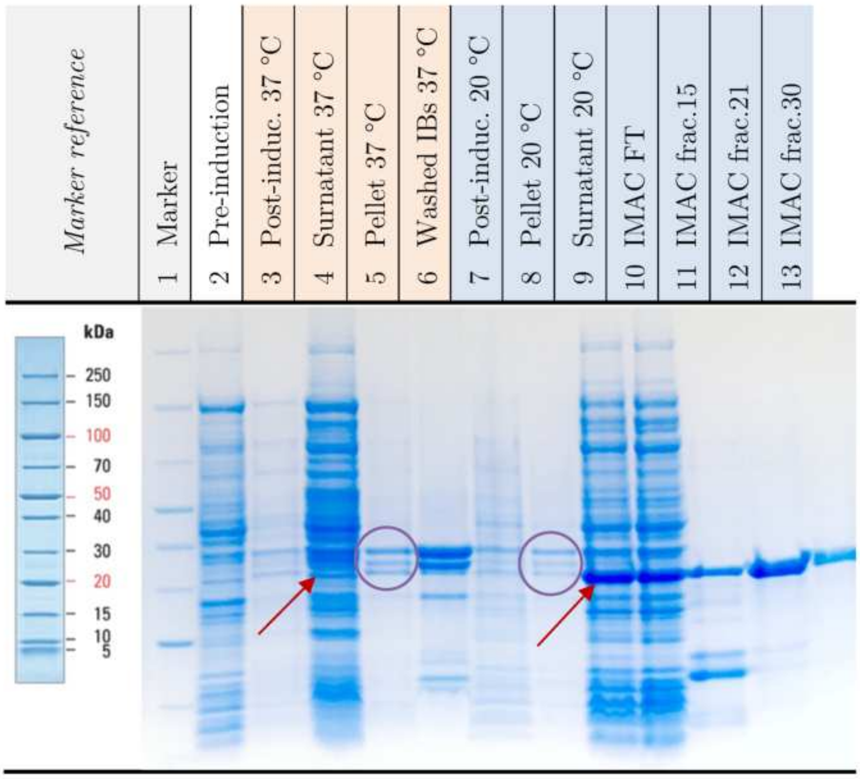


Figure 7. Effect of expression temperature on recombinant AhRa2 yield. The bacterial culture was the same before induction. At the moment of induction it was split in two: the first one was then grown at 37 °C (lanes 3-6), the second one at 20 °C (lanes 7-13). Notice how the AhR protein isolated after the lysis from the soluble fraction (red arrows) is much higher for the 20 °C synthesis, while it is very similar in the insoluble fraction (purple circles).

All AhR constructs showed high expression levels (an example is shown in Figure 8, on the left) but unfortunately their solubility was very low. In particular, we found the protein at high concentrations both in the soluble fraction (surnatant) and in the insoluble fraction (pellet) of the bacterial lysate (Figure 8, on the left, lanes 3-5). This indicates that part of AhR is already aggregated in the bacterial cells prior to its lysis, and most likely located inside the so-called inclusion bodies (IBs). The amount of AhR in the soluble fraction was still very high, however, and it was predominant with respect to all the other endogenous proteins present in the cell lysate. What surprised us, is that very little – if any – protein was isolated by the immobilized-cation affinity chromatography step (IMAC), which should capture all the HisTag-containing protein. We noticed that the cell lysate surnatant, which is clarified by centrifugation and then filtered at 0.22 µm, slowly began to scatter light before and during injection in the IMAC column, indicating early signs of aggregation. When the sample exited the column (i.e. the flow-through of the IMAC step) it became increasingly turbid in a matter of minutes and flocculated in a few hours. To investigate the nature of the phenomenon, we centrifuged the flow-through to separate the precipitate from the surnatant and we prepared two separate samples for the SDS-PAGE run (the precipitate readily re-dissolved in the denaturing SDS sample buffer). It emerged clearly that the precipitant is almost pure His-SUMO-AhR recombinant protein, while all the other bacterial, endogenous proteins remain in solution (Figure 8, on the left, lanes 6-8). The same phenomenon emerged for all PAS-B-only AhR constructs, leading to very small amounts of recombinant protein being isolated at the end of the IMAC purification step.

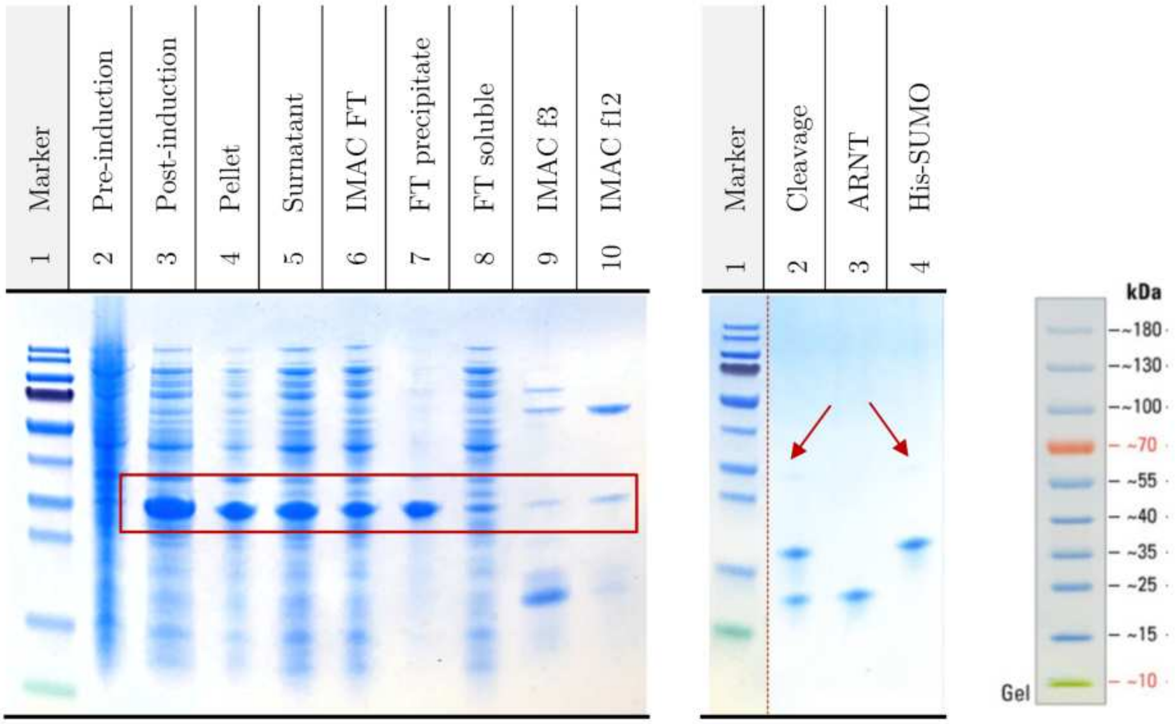


Figure 8. SDS-PAGE of selected PAS-B-only constructs purifications. On the left, gel representing the purification of AhRa1 (band highlighted in the red box). The expression level is very high (lane 3) but a large part of the recombinant protein is lost in the pellet (lane 4). The IMAC chromatography can isolate only a very small amount of the total protein (lanes 9, 10).

In the center, cleavage of the ARNTa1 protein with ULP-1. ARNT and His-SUMO were separated by IMAC chromatography. Red arrows indicate small traces of uncleaved His-SUMO-ARNT.

On the right are the reference weights of the molecular marker.

The synthesis of ARNTa1 proceeded flawlessly instead and was characterized by high expression levels, good solubility, and very limited losses along the purification process. His-SUMO-ARNTa1 was also successfully cleaved by ULP-1 to yield the desired ARNTa1 protein (Figure 8, on the right). Overall yields of the purified protein were in the order of 5-10 mg per liter of bacterial culture.

Since the precipitation of AhR constructs happened in a dynamic fashion after the lysis of the bacteria, we tested the possibility of stabilizing such constructs with the aid of ARNT. The rationale behind this approach is that while the PAS-B domain of AhR alone may not be stable on its own, it may remain in solution if it forms a dimer with its natural partner ARNT. This process is normally done by co-expression approaches, that is, by transforming DNA sequences that codify both proteins of interest inside the same bacterial cell and then inducing the synthesis of both proteins at the same time (Figure 9, **ii** and **iii**). With this approach, known as co-expression, both proteins are produced in each cell simultaneously.

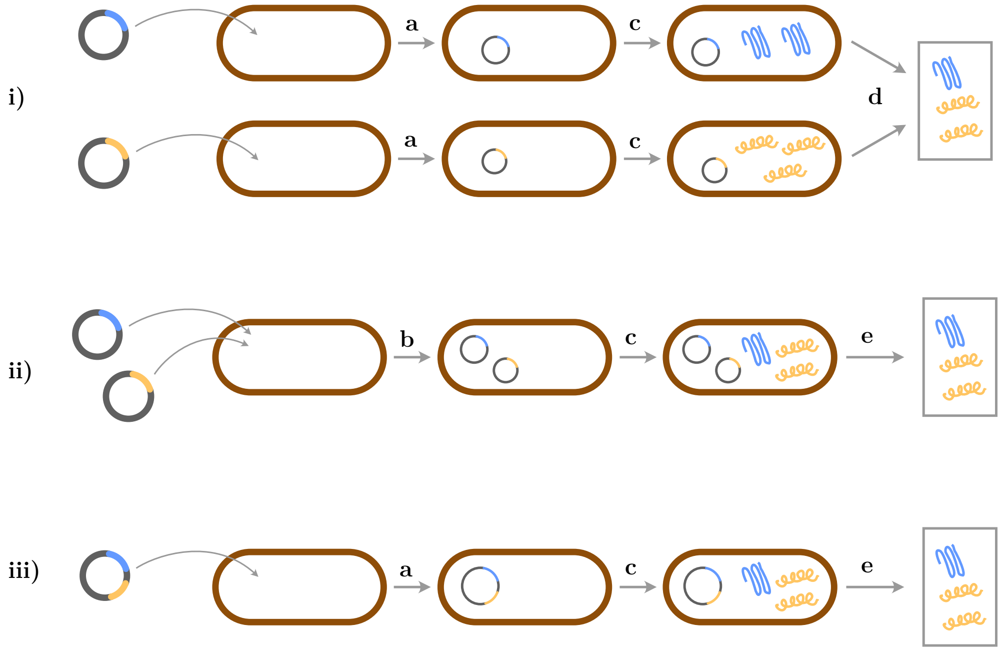


Figure 9. Co-expression and co-lysis approaches. In the co-lysis approach (i), the two plasmids are transformed in two different bacterial cultures (a), which are then allowed to grow and express the two proteins separately (c). At the end of the expression, the two bacterial cultures are mixed and co-lysed together (d). Co-expression can be achieved in two different ways: ii) the two proteins are coded in two different plasmids that are transformed in the same bacterial cell (a, co-transformation). The two proteins are then expressed at the same time (c) and finally lysed (d). iii) the two proteins are coded in the same plasmid (on two different cloning sites) and are transformed (a) in the cells, which will express both proteins at the same time (c).

Another approach is to grow two different bacterial cultures, each producing one of the desired proteins, and then mixing the bacterial cells prior to their lysis (Figure 9, **i**). With this approach, the two proteins come in contact with each other at the moment of the lysis procedure. We call this second method *co-lysis*. While this second approach may seem more cumbersome, it is simpler to operate as the overall procedure is a simple parallelization of two, simple experiments. The co-expression, on the other hand, requires cloning the two different proteins in two different plasmids with different antibiotic selections. The co-expression thus requires more effort during the cloning procedure but is simpler at the expression stage. One additional benefit of this approach is that both proteins are produced inside every bacterial cell at the same time. This is particularly useful if one of the two proteins precipitates before the lysis procedure, which is of common occurrence. In our case, we wanted to test the potential stabilization effect of ARNT after the lysis procedure, so we opted to try the co-lysis approach with our existing plasmids. Unfortunately, our expectations were not met as we were not able to isolate significant amounts of AhR. In particular, at the end of the co-lysis experiments we only obtained pure ARNT, while AhR continued to precipitate after the lysis and during the purification procedures (data not shown). Since only AhR precipitated during our experiments in presence of ARNT, we hypothesized that they could not dimerize effectively, otherwise they would either co-precipitate or form stable dimers in solution.

pDUET co-expression approach

Since our attempts to obtain a soluble AhR-ARNT complex with the co-lysis approach failed, we decided to test the co-expression method. We opted to use the pDUET cloning platform, as it allows to clone two different DNA sequences in a single plasmid that contains two multiple cloning sites (MCS). This approach is more complex on the cloning side compared to either the co-lysis method or the co-expression with two different plasmids. The TA cloning method is not viable – for example – as it is not possible to have a linearized plasmid with two different cloning sites. A common option is the use of different restriction enzymes for the two cloning sites (which indeed present several different restriction sites each). We instead opted for the so-called restriction-free cloning (RF cloning) which – as its name suggests – avoids the usage of restriction enzymes and is instead based on two PCR cycles and the exploitation of high-performance, high-fidelity DNA polymerase enzymes (experimental details at page 78). By using the pDUET expression plasmid, we lost the benefits of the SUMO fusion proteins (enhanced solubility and no-leftover HisTag removal), aiming at the putative solubility enhancement provided by the dimerization with ARNT.

Since the co-lysis method suggested that the binding affinity of the PAS-B domains of AhR and ARNT may not be strong enough to yield a stable dimer, we opted to clone larger constructs in the pDUET plasmid. The rationale behind this choice is that by adding the PAS-A domain (and for some constructs even the bHLH domain), the affinity of the two proteins will increase, as the intermolecular forces attracting the two are generally additive (as are the possible interface areas). As for the length of the PAS-B domain, we chose three different lengths: the first two being the same as AhRa1 and AhRa2 (sequences ending with E389 and N418, respectively), the third one being a longer version ending with S474 (Table 4). We avoided to test the murine homolog of AhR in pDUET, as it demonstrated to be less soluble than its human counterpart in our previous experiments. For ARNT we only designed two constructs: one with the PAS-A and PAS-B domains (ARNTc4) the other also encompassing the bHLH domain (ARNTc1). We designed AhR constructs both with (AhRc1…6) and without (AhRc7…12) the HisTag, with the idea of trying to isolate the latter ones as a complex with the ARNT counterpart, which instead comprised the HisTag.

Table 4. pDUET-based AhR and ARNT constructs. 6His indicates the presence of the 6xHisTag at the *N*-terminus. bHLH, PAS-A and PAS-B indicate the presence of the corresponding domain in the construct sequence. Notice that the PAS-B domain have been cloned in three different lengths (ending with E389, N418 or S474 respectively). Constructs ARNTb1-b2 were prepared as a backup.

Code	Source	Sequence span	6His	bHLH	PAS-A	PAS-B	Vector
ARNTb1	murine	Q81-E470	✓	✓	✓	✓	pET-SUMO
ARNTb2	murine	K155-E470	✓	-	✓	✓	
AhRc1	human	G30-E389	✓	✓	✓	✓	pDUET (MCS1)
AhRc2	human	G30-N418	✓	✓	✓	✓	
AhRc3	human	G30-S474	✓	✓	✓	✓	
AhRc4	human	G109-E389	✓	-	✓	✓	
AhRc5	human	G109-N418	✓	-	✓	✓	
AhRc6	human	G109-S474	✓	-	✓	✓	
AhRc7	human	G30-E389	-	✓	✓	✓	
AhRc8	human	G30-N418	-	✓	✓	✓	
AhRc9	human	G30-S474	-	✓	✓	✓	
AhRc10	human	G109-E389	-	-	✓	✓	
AhRc11	human	G109-N418	-	-	✓	✓	
AhRc12	human	G109-S474	-	-	✓	✓	
ARNTc1	murine	Q81-E470	✓	✓	✓	✓	pDUET (MCS2)
ARNTc4	murine	K155-E470	✓	-	✓	✓	

We started the cloning procedure of constructs AhRc4-c6 with the successful synthesis of the corresponding megaprimers (Figure 10). We were then able to insert the megaprimers in the cloning site number 1 (MCS1) and confirm the current sequence by DNA sequencing. We also synthesized the ARNTc4 megaprimer successfully but failed multiple times to insert it into the pDUET plasmid containing any of the AhRc4-c6 sequences. Optimization of the megaprimer amount (100, 200 or 400 ng) and of the annealing temperature of the PCR cycle of the RF cloning did not yield any positive result. A close inspection of the flanking regions of the megaprimers confirmed that they were correctly designed, while the DNA sequencing of the destination pDUET vector confirmed the expected target sequence. We have not been able to track down the exact cause of the failed RF cloning of ARNT, and finally gave up un such attempts when we realized that the AhR sequences cloned in pDUET had very low expression levels (see next paragraph).

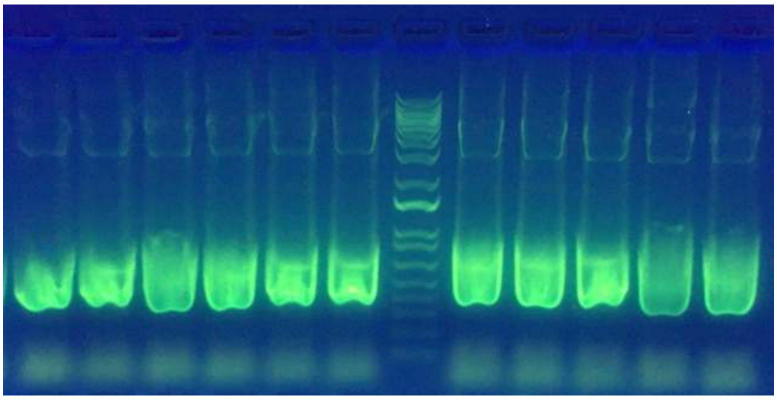


Figure 10. Agarose gel of the AhRc4-c5-c6 megaprimers. Each megaprimer occupies three lanes as the volume of the PCR reaction (50 µL) exceeds the well size (~20 µL). Lane #7 is the DNA size marker (DNA ladder). At the end of the electrophoretic run, the megaprimer bands are excised from the agarose gel and the megaprimers are extracted with a commercially available kit.

Since it was not possible to insert the correct sequence of ARNT in the MCS2 of pDUET, we started testing the expression levels of AhR alone in the pDUET plasmid. For these expression tests, another protein named NDUFAF (NADH:ubiquinone oxidoreductase complex assembly factor), was co-expressed with AhR as its sequence was located in the MCS2 of pDUET plasmid that we used as a template. NDUFAF is completely insoluble, and its presence has to be expected in the pellet fraction at a molecular weight of 38 kDa. Unfortunately, AhR expression levels were extremely low and very variable from batch to batch for all AhRc4-c5-c6 constructs. We then cloned the ARNTb2 construct (Table 4) which contained the PAS-A and PAS-B domains of ARNT (thus the one corresponding to the AhRc4-c5-c6 constructs) but in the pET-SUMO plasmid instead of the MCS2 of pDUET. Since the pET-SUMO and the pDUET plasmids have different antibiotic resistances (kanamycin for the former, chloramphenicol for the latter) we were able to co-transform both plasmid in the same bacterial cells. It was thus possible to execute the co-expression of both proteins using the double plasmid method (see Figure 9, **ii** on page 13). In this experiment we obtained a good expression level for ARNT and NDUFAF only (Figure 11, left panel), but unfortunately no AhR overexpression was detected. ARNT was isolated in good yield at the IMAC step (Figure 11, top-right panel). Being the elution peak quite broad, we took three samples from fractions 14, 16 and 18 of the eluate and run those samples in a Superdex 200 10/300 GL analytical size-exclusion chromatography (SEC) column (Figure 11, bottom-right panel). All the fractions produced a very similar peak with the same retention time of ~14 mL, compatible with the 50 kDa mass of the His-SUMO-ARNTb2 construct. All the peaks isolated from the SEC runs produce two distinct – but slightly overlapping – bands on the SDS-PAGE (Figure 11, left panel, lanes 8-10). We interpret this phenomenon as an artifact of the electrophoresis run, as the SEC cromatographies – which were conducted in native conditions – exclude both the possibility of an ARNT homodimerization or a heterodimerization with AhR.

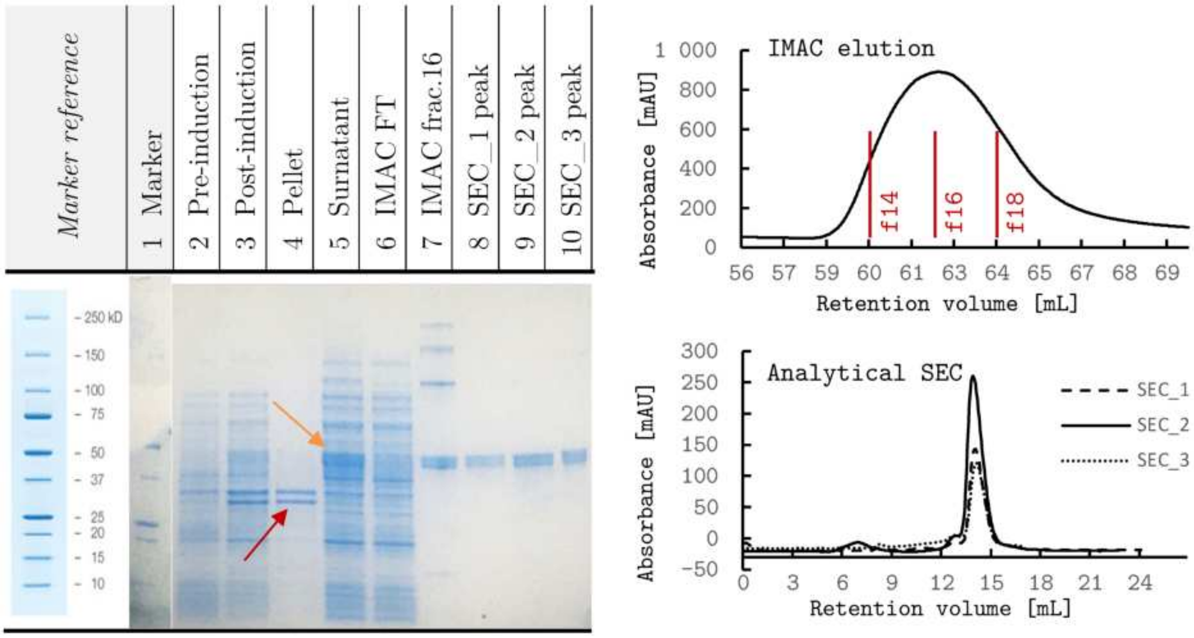


Figure 11. Co-expression experiment of AhRc6 and ARNTb2. On the left, SDS-PAGE of the various expression and purification steps. AhRc6 is not detected, while NDUFAF is located in the insoluble fraction (red arrow). ARNTb2 remains in the surnatant and is later isolated by the IMAC and SEC chromatographic steps. The fractions 14, 16 and 18 of the IMAC have been characterized with the SEC 1, 2 and 3 runs.

We also tested the expression levels of the same two constructs in two separate cultures (one for each construct). The results were quite clear; AhRc6 was not expressed at detectable levels, NDUFAF was found in the insoluble fraction of the bacterial lysate (likely located in the inclusion bodies) (Figure 12). ARNTb2 expressed at lower levels instead (the overexpression is not even visible in the whole bacterial lysate) but it is completely located in the soluble fraction of the lysate and it is easily isolated by the IMAC step. The expression levels of the AhRc6 construct is much lower than that of previous, PAS-

B-only constructs, which could benefit from the increased solubility given by the SUMO moiety (Cf. Figure 7, Figure 8). For those constructs, even though part of the protein precipitated after the lysis procedure, part of it was still isolated in the soluble fraction, which instead did not happen for AhRc6. Given the unpromising results obtained with the pDUET-based constructs (scarce expression levels for AhR, hampered cloning for ARNT) we decided to move on to new constructs based on a new vector: pET-21d(+).

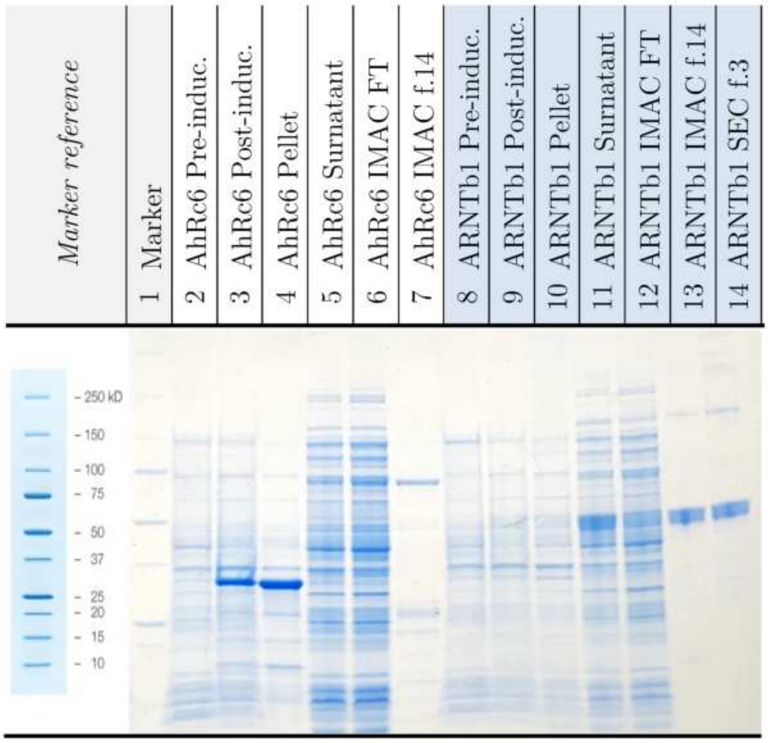


Figure 12. Expression of separate AhRc6 and ARNTb2 cultures. There is a clear overexpression in the AhRc6 culture, but it is from the NDUFAF protein located in the MCS2 of pDUET. The expression level of AhR, in absence of the SUMO moiety, is not detectable (lanes 3-5), while the expression levels of ARNT are lower but it is very soluble and easily isolated (lanes 9-13).

pET-21d(+) and pET-SUMO multi domain constructs

We decided to clone the same AhR construct of the AhRc1-c6 series into the pET-21d(+) plasmid (AhRd1-d6 series, Table 5). This plasmid has a single cloning site (like pET-SUMO) but it does not produce fusion proteins. It has an ampicillin resistance, so it can be co-transformed with both the pET-SUMO and the pDUET plasmids. It is designed to produce a non-cleavable HisTag at the protein *C*-terminus, but we have prevented it during the cloning process by inserting a stop codon between the end of our construct and the HisTag sequence. The idea behind this choice was to produce the

AhR-ARNT complex by the two-plasmid co-expression method and then isolate such complex with the cleavable HisTag of the ARNT construct.

The cloning process was executed with the restriction-free cloning method (experimental details at page 78) and started experimenting with constructs AhRd4-d5-d6, which contained the PAS-A and PAS-B domains of human AhR.

Table 5. pET-21d(+)-based AhR constructs. The HisTag at the *C*-terminus is prevented by a stop codon insertion. bHLH, PAS-A and PAS-B indicate the presence of the corresponding domain in the construct sequence.

Code	Source	Sequence span	6His	bHLH	PAS-A	PAS-B	Vector
AhRd1	human	G30-E389	-	✓	✓	✓	pET-21d(+)
AhRd2	human	G30-N418	-	✓	✓	✓	
AhRd3	human	G30-S474	-	✓	✓	✓	
AhRd4	human	G109-E389	-	-	✓	✓	
AhRd5	human	G109-N418	-	-	✓	✓	
AhRd6	human	G109-S474	-	-	✓	✓	

After the successful cloning of the AhRd4-d5-d6 construct, we co-transformed them with ARNTb2 in *E. coli* BL21(DE3) competent cells. We have thus produced three cultures, each containing ARNTb2 and one of the aforementioned AhR constructs. We induced the co-expression of both proteins under AhR-optimal conditions (e.g. temperature of 20 °C, 0.5 mM of IPTG, etc.), lysed the cells and analyzed the soluble and insoluble fractions of all cultures. We have then run the soluble fractions through an IMAC chromatography in order to capture the AhR-ARNT complex *via* the ARNT HisTag. While we obtained a good over-expression of ARNT (in line with previous experiments), we were not able to detect any sign of AhR recombinant expression for any of the AhRd4-d5-d6 constructs (Figure 13). Recombinant expression of the AhRd4-d5-d6 constructs without ARNT co-expression also did not produce any detectable over-expression (data non shown).

We were puzzled by the absence of detectable expression levels of AhR constructs based on both pDUET and pET-21d(+) plasmids. The presence of the AhR plasmids in the host bacteria is confirmed by the fact that the bacterial cultures grew rapidly under double antibiotic resistance (bacterial cultures grew in presence of chloramphenicol and ampicillin when transformed with pDUET and pET-21d(+) plasmids, respectively). Furthermore, pET-SUMO, pDUET and pET-21d(+) vectors all have the same *lac* inducer (activated by IPTG) and the same T7 promoter, thus the concomitant expression of ARNT (based on pET-SUMO) and missing expression of AhR is not related to protein synthesis induction issues. Finally, any cloning artifact have been excluded by careful and complete sequencing of all cloned DNA sequences). We argue that the best explanation for the missing expression levels of such AhR construct must be the missing SUMO moiety. SUMO may not only stabilize the AhR constructs in solution, it but may also reduce its toxicity towards bacterial cells, thus increasing the expression levels.

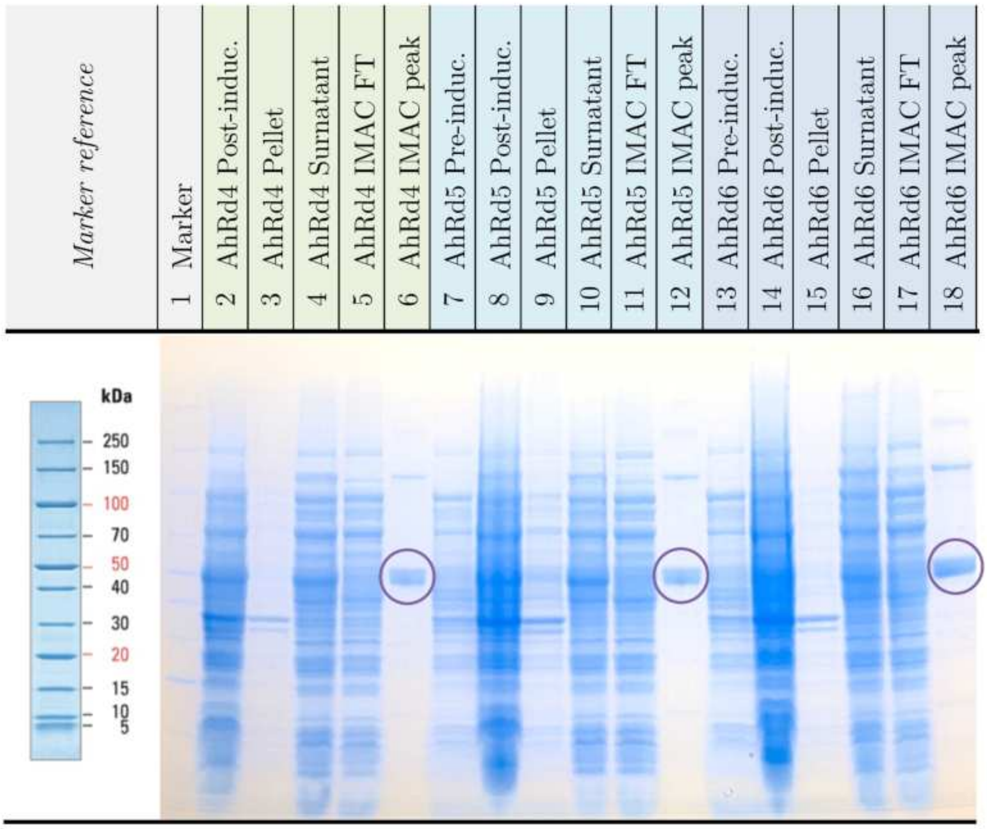


Figure 13. Co-expression of AhRd4-d5-d6 constructs with ARNTb2. The purple circles indicate the protein isolated from the IMAC, which is composed only by ARNT (~50 kDa in size, vs the 32~41 kDa size of the AhR constructs)

We decided to clone the same AhR sequences of the pDUET and pET-21d(+) plasmids into the pET-SUMO vector (Table 6). At the same time, we cloned the corresponding ARNT sequences with and without HisTag into the pET-21d(+) plasmid (Table 7) in order to conduct co-expression experiments with the new AhR constructs in pET-SUMO. Due to technical difficulties and time restraints during the cloning process, we were only able to prepare constructs from AhRe1 to AhRe4 for AhR, and only ARNTd1 among the 4 designed ARNT constructs.

Table 6. pET-SUMO-based multidomain AhR constructs. bHLH, PAS-A and PAS-B indicate the presence of the corresponding domain in the construct sequence.

Code	Source	Sequence span	6His	bHLH	PAS-A	PAS-B	Vector
AhRe1	human	G30-E389	✓	✓	✓	✓	pET-SUMO
AhRe2	human	G30-N418	✓	✓	✓	✓	
AhRe3	human	G30-S474	✓	✓	✓	✓	
AhRe4	human	G109-E389	✓	-	✓	✓	
AhRe5	human	G109-N418	✓	-	✓	✓	
AhRe6	human	G109-S474	✓	-	✓	✓	

Table 7. pET-21d(+)-based multidomain ARNT constructs. bHLH, PAS-A and PAS-B indicate the presence of the corresponding domain in the construct sequence.

Code	Source	Sequence span	6His	bHLH	PAS-A	PAS-B	Vector
ARNTc1	murine	Q81-E470	-	✓	✓	✓	pET-21d(+)
ARNTc2	murine	K155-E470	-	-	✓	✓	
ARNTd1	murine	Q81-E470	✓	✓	✓	✓	
ARNTd2	murine	K155-E470	✓	-	✓	✓	

In the first experiment we tested the co-expression of all available AhR+ARNT constructs and the expression of the AhR constructs alone. When expressed alone, all AhR construct showed very high expression levels, in line with the predicted positive effect of the SUMO fusion partner (Figure 14, lanes 2-9). At the same time, the co-expressions of the same AhR constructs with ARNT (under identical and concomitant experimental conditions) was not detectable (Figure 14, lanes 10-17). This result is in part positive because it demonstrates that the pET-SUMO vector is able to express difficult AhR constructs at high levels, but it puzzled us the negative effect of the co-expression with ARNT. We expected a slight decrease in the expression levels of AhR because of the increased metabolic burden given to the bacterial cells producing not one but two recombinant proteins at the same time. To give credit to this hypothesis, we should see at least some overexpression of ARNT in the range of 44 kDa in the SDS-PAGE, of which unfortunately there is no trace. We then tested the recombinant expression of ARNTd1 alone, finding that even in absence of AhR synthesis, there is no trace of ARNT overexpression (data not shown). Since ARNTd1 is the only ARNT construct that is not present in a pET-SUMO plasmid, and the only one not overexpressing, our best guess is that, similarly to AhR, also ARNT constructs are quantitatively overexpressed only in the presence of the SUMO fusion moiety. ARNTb1 and ARNTb2 which contain the same sequences of ARNTd1 and ARNTd2 in the pET-SUMO vector (Table 4), were found to both co-express at high levels and be very soluble.

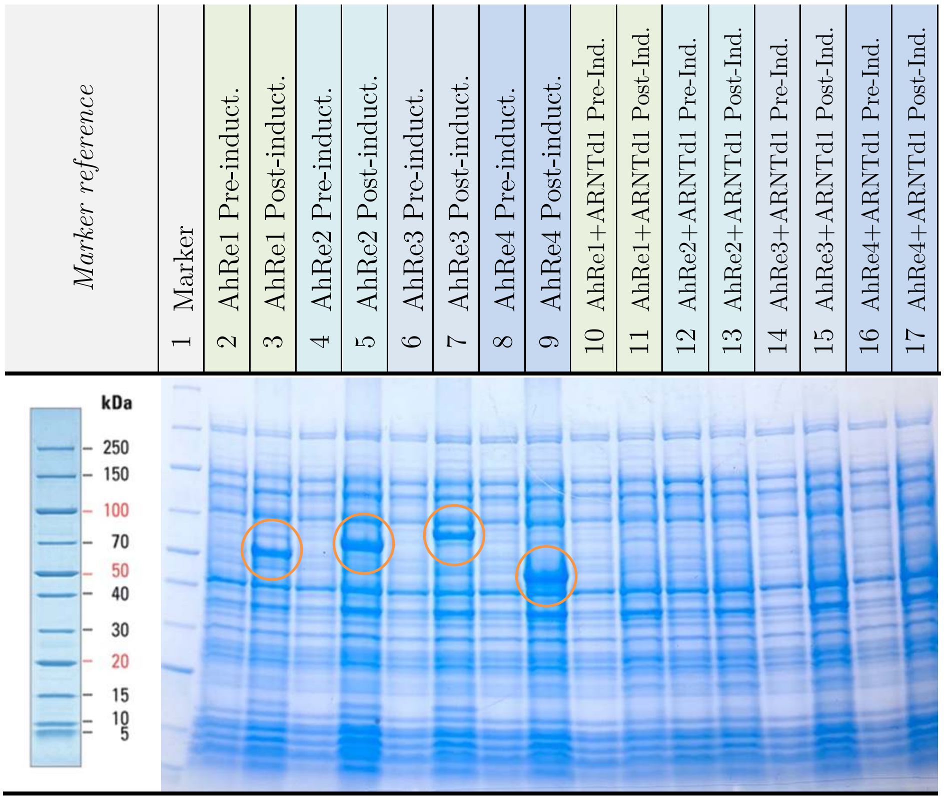


Figure 14. Co-expression of AhRe1-e2-e3-e4 constructs with ARNTd1. Over-expression of AhR constructs is only visible when they were expressed alone (lanes 2-9, AhR bands circled in orange), while in the co-expressions with ARNT (lanes 10-17) neither AhR of ARNT are detectable.

Structure-inspired AhR constructs

As it was previously described in the introduction section, the first experimental structures of a PAS-B domain of AhR were published in 2022. The first structure was the crystallographic structure PAS-B domain from *D. melanogaster* alone and in complex with ARNT[14], the other was the semi-complete cytosolic complex of human AhR solved by Cryo-EM[12]. We took inspiration from both these structures in order to design new constructs with enhanced solubility. Starting from the *Drosophila* structure, we noticed a strong, hydrogen-bond-mediated interaction between the *C*-terminal Arginine 381 and residues Asp311, Asp312 and Ala333 (Figure 15). Such interactions likely stabilize the *C*-terminal α-helix of the AhR PAS-B domain and possibly stabilize the PAS-B domain as a whole. We thus designed by sequence alignment two PAS-B-only AhR constructs with the human (AhRf1) and murine (AhRf2) sequences that have the same *N*- and

C-termini as the reported Drosophila structure (Table 8). We additionally designed mutants of AhRf1 with mutation L331E (AhRg1), I338Q (AhRg2) and with both mutations (AhRg3). These mutations are intended to reduce the hydrophobicity of the F-α-helix of AhR, which interacts with XAP2 in the cytosolic complex of AhR but is exposed to the solvent in the free PAS-B domain. The Cryo-EM structure of the AhR cytosolic complex also confirms that the C-terminal α-helix of the human PAS-B ends with Arg398, while the following protein sequence is not structured.

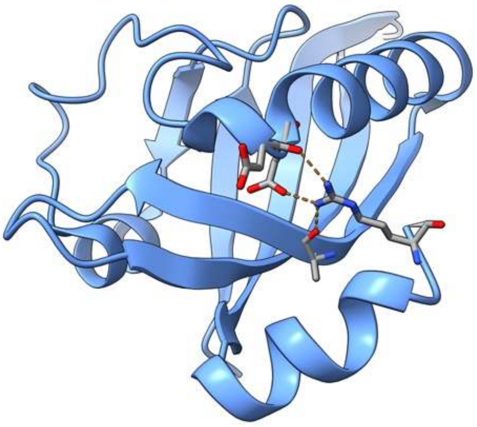


Figure 15. Arg381 interactions in AhR PAS-B from *D. melanogaster*. The *C*-terminal Arg381 residue creates hydrogen bonds with residues Asp311, Asp312 and Ala333, likely contributing to the stabilization of the *C*-terminal α-helix.

Table 8. Structure-inspired AhR PAS-B constructs. AhRf1-f2 mimic the *N*- and *C*-termini of the reported *Drosophila* PAS-B domain. AhRg1-g2-g3 are point-mutated versions of AhRf1.

Code	Source	Sequence span	PAS-B	Mutations	Vector
AhRf1	human	N284-R398	✓	WT	pET-SUMO
AhRf2	murine	N278-R392	✓	WT	
AhRg1	human	N284-R398	✓	L331E	
AhRg2	human	N284-R398	✓	I338Q	
AhRg3	human	N284-R398	✓	L331E + I338Q	

Constructs AhRf1-f2 were obtained with the restriction-free cloning method, as previously described, while the mutants AhRg1-g2-g3 were obtained by site-directed mutagenesis starting from the AhRf1 construct. The double mutant AhRg3, having two very close mutation sites, was obtained in a single mutagenesis reaction using a long primer encompassing both mutation sites (see "List of primers" on page 92).

The non-mutated AhRf1 and AhRf2 constructs both overexpressed well. The human homolog was better overexpressed and more soluble compared to the murine construct, in accordance to previous results obtained from PAS-B-only AhR constructs. The amount of protein that is possible to isolate from the soluble fraction of the bacterial lysate is still very low and accounts to a minimal percentage of the overall recombinant protein (Figure 16). Nonetheless, construct AhRf1, together with construct AhRa2, showed the highest overexpression and yield in the soluble fraction of the bacterial lysate. For this reason, their extraction and purification were attempted with alternative approaches, as it will be described in the following section.

Of the AhRg1-2-3 series, we started by cloning the double mutant AhRg3 with the site-directed mutagenesis method. Contrary to our expectations, subsequent and repeated experiments demonstrated that AhRg3 overexpression was low and the construct still showed a very low stability as it quickly precipitated out of the soluble fraction of the bacterial lysate (Figure 17). Additionally, most of AhRg3 was located in the insoluble fraction of the bacterial lysate, indicating that it probably aggregates inside the cells prior to their lysis. A similar behavior was found for construct AhRg1 (L331E single mutant), while AhRg2 was not cloned due to time restraints. We concluded that the designed mutations diminished the recombinant protein expression while not increasing the protein stability and were thus discarded for further experimentation.

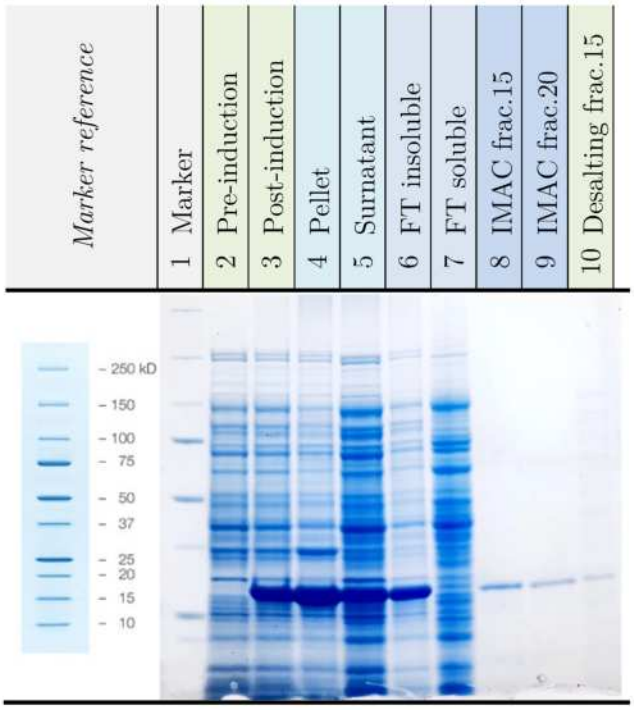


Figure 16. Expression and purification of AhRf1. The overexpression is very high (lane 3), but the protein is highly localized in the insoluble fraction (4). Only minimal

amount of protein is recovered by IMAC chromatography (8,9), while the majority of it precipitates out of the IMAC flow-through fraction (6).

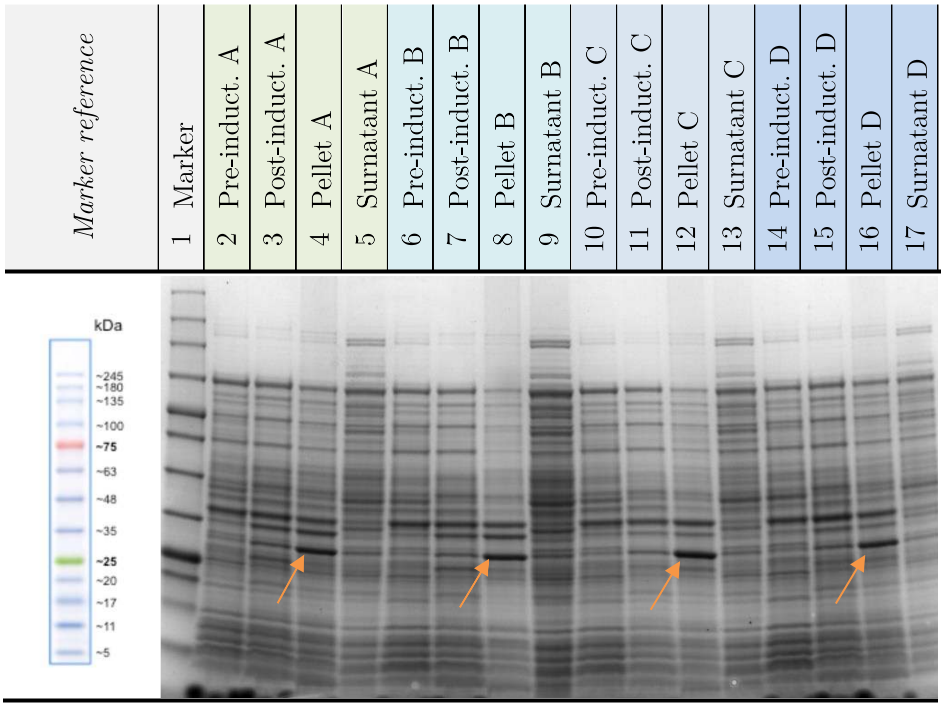


Figure 17. Recombinant expression of AhRg3. Condition A: 20 °C, 1 mM IPTG, overnight expression (ON). Condition B: 20 °C, 0.5 mM IPTG, ON. Condition C: 37 °C, 0.5 mM IPTG, ON. Condition D: 37 °C, 1 mM IPTG, 4 hours. Arrows indicate that most recombinant protein is located in the insoluble fraction (pellet) of the bacterial lysate.

Alternative approaches

All of the AhR constructs described so far showed very serious expression and solubility issues. For the best overexpressing proteins (namely AhRa2 and AhRf1), despite all of the aforementioned attempts, we recognized that the protein was not stable after the bacterial lysis, since it quickly precipitates out of the extraction buffer. For these constructs – however – a small amount of chimeric protein with SUMO was isolated at the end of the IMAC 1 and Desalting chromatographies and samples suitable for circular dichroism (CD) were prepared. The CD spectra were recorded, together with the spectrum of His-SUMO alone (Figure 18). The spectra indicate that AhR is indeed present in an ordered state, and that its secondary structure is well defined and in line with its PAS domain nature.

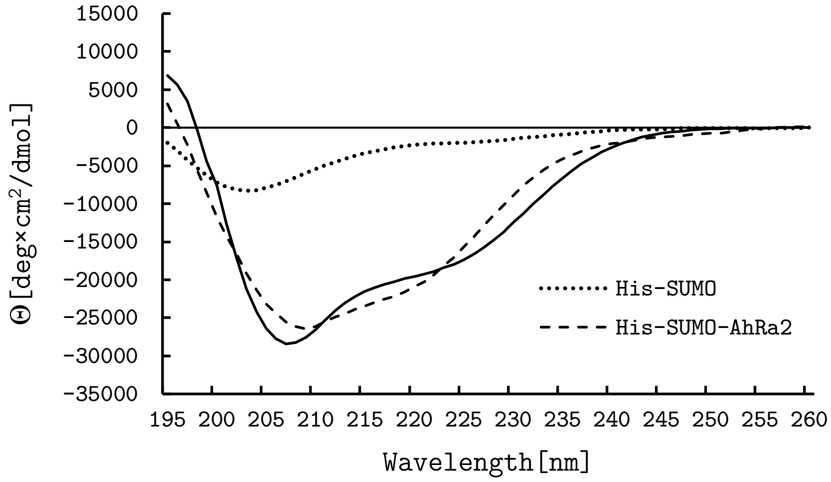


Figure 18. CD spectra of selected AhR chimeric constructs. The secondary structure of His-SUMO-AhRa2 and of His-SUMO-AhRf1 is similar, with a superposition of α-helix and β-sheet signals. The His-SUMO moiety was measured separately and showed mostly a random coil conformation, confirming that the signal seen for the SUMO-AhR chimeras is attributable mostly to the AhR moiety.

The structured state of these AhR construct, together with their good expression levels, suggest that there could be the possibility of isolating the protein in good yields if their precipitation after the bacterial lysis is avoided. For this reason, we envisaged different expression and purification approaches that could allow us to isolate the protein in reasonable amounts.

Zinc supplementation. After being able to isolate a small amount of chimeric 6His-SUMO-AhRa2 (cleavage of the SUMO moiety consistently resulted in complete AhR precipitation) we tested such construct in a thermal shift assay using the differential scanning fluorimetry method (DSF). We initially tested - in 96-well microplates – several chemical environments including various buffer systems, pH values, salt concentrations and various ions and organic additives. We also tested the DSF signal of 6His-SUMO protein alone to confirm that the melting curve is indicative of the AhR unfolding and not of SUMO. Initial results reported that salt (NaCl) concentration did not affect AhR stability significantly, while the optimal pH was 8 (although 7 and 7.5 were very similar). Of all tested additives, only zinc sulfate ($ZnSO_4$, 2 mM concentration) showed a remarkable stabilization effect, increasing the melting temperature from 37 °C to 55 °C (Figure 19). The same effect was found in presence of zinc chloride ($ZnCl_2$), excluding the SO_4^{2-} anion as the determining stabilization factor. Other common divalent cations like magnesium (Mg^{2+}) and calcium (Ca^{2+}) did not affect the melting temperature. Since the zinc cation is very like to be complexed by the HisTag, we also tested copper and nickel sulfates ($CuSO_4$, $NiSO_4$) but found that they instead lowered the melting temperature of

6His-SUMO-AhRa2 to ~33 °C. To further test the validity of our DSF data, we tested the 6His-SUMO protein alone and found that it does not show any DSF signal (Figure 19) likely because of its intrinsically disordered structure. It must be highlighted, remarkably, that the melting temperature of 6His-SUMO-AhRa2 is indeed very low, in accordance to all previously presented data, confirming its elevated instability.

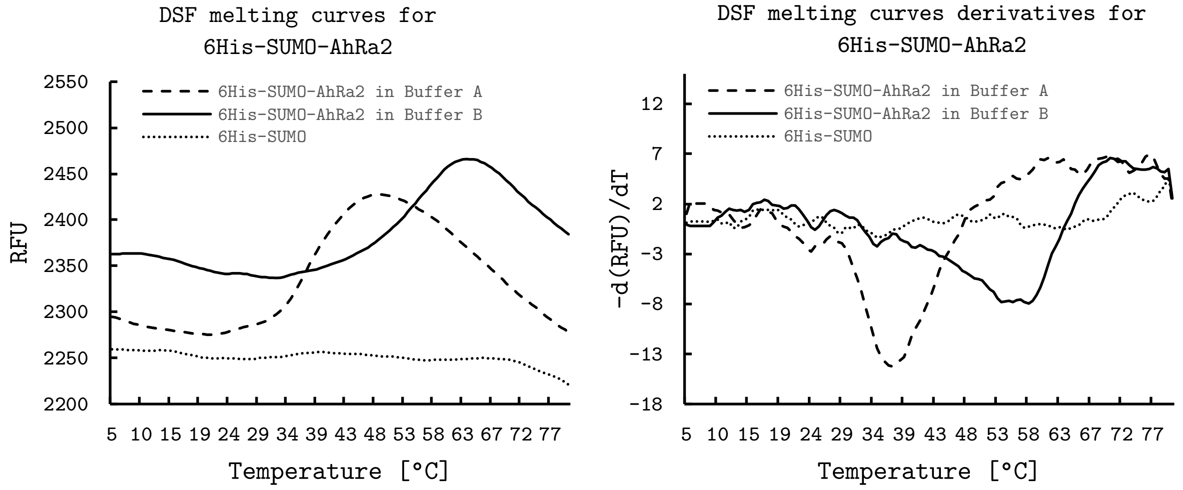


Figure 19. DSF melting curves of 6His-SUMO-AhRa2. Buffer A: 20 mM TRIS pH 8, 150 mM NaCl. Buffer B: 20 mM TRIS pH 8, 150 mM NaCl, 2 mM $ZnSO_4$. The presence of the zinc salt increases the melting temperature from 37 °C to 55 °C for 6His-SUMO-AhRa2. 6His-SUMO does not show any melting process.

The exploitation of the zinc stabilization in the protein synthesis, extraction and purification process is not straightforward. The idea was to supplement the zinc during some stage of the AhR preparation, in order to increase its solubility and thus its yield. Adding zinc to any of such stages, unfortunately, presents its challenges:

- Adding zinc in the bacterial broth during protein expression: zinc uptake by *E. coli* is tightly regulated, thus zinc concentration in the bacterial cytosol is likely not affected.
- Adding zinc to the lysis buffer: zinc causes the DNA to precipitate, co-precipitating AhR (which is very basic and strongly binds to DNA).
- Adding zinc after lysate centrifugation, prior to IMAC injection: the DNA is mostly removed, but the HisTag is completely saturated by zinc ions, making the IMAC isolation not possible.
- Adding zinc after IMAC chromatography: not useful, as most of the protein is already lost at this stage.

All of the above strategies were separately tested but were found to be ineffective as the protein yield was lower compared to similar experiments lacking the zinc supplementation (data not shown). We thus experimented adding the zinc after the lysate centrifugation, when the DNA is mostly removed, and then isolating the protein by cationic exchange chromatography. Even under these conditions, we were only able to isolate a minimal amount of AhR protein of low purity, while most of it precipitated out of the lysate even in presence of the zinc ions (data not shown). The same result was obtained with the AhRf1 construct. We concluded that the stabilization effect seen on the DSF technique was either a specific experimental artifact or that it was not exploitable or effective on the whole bacterial lysate, compared to the isolated AhR protein.

ITE supplementation. It has been reported in literature that the PAS-domain-containing, bacterial protein *LasR* (which is unrelated to AhR, even though it contains a PAS domain) could be recombinantly expressed in *E. coli* only in presence of a suitable, high-affinity ligand in the expression broth[19]. Such ligand was then retained during all the purification procedure and eventually found at high occupancy in the LasR crystallographic structures.

The ligand-binding, PAS-B domain of AhR is a relatively small structure of ~110 amino acids. It does not have a solvent-exposed binding pocket, but instead - it was initially supposed - a binding site completely buried inside the PAS fold. The first experimental structures of AhR in 2022 later confirmed such hypothesis (Cf. Figure 4 on page 5). We hypothesized that the PAS-B domain of AhR may be stabilized, when not bound to other proteins comprising the cytosolic complex, by the presence of a small molecule located in its binding pocket. To validate this hypothesis, we tested the AhR expression and extraction in presence the commercially available AhR agonist 2-(1′H-indole-3′-carbonyl)-thiazole-4-carboxylic acid methyl ester (ITE, for short, Figure 20). ITE was chosen because it is one of the best known and strongest AhR binders, with a reported in-vitro PAS-B inhibition constant of 3 nM[20]. At the same time, contrary to many AhR ligands, it also has a “reasonable” water solubility.

We initially tested the recombinant expression of AhRa2 with ITE added to the bacterial broth at a 1 mg/L concentration but found no significant increase in protein yield (data not shown). In later experiments we expressed AhRa2 in absence of ITE, but we added ITE to the lysis buffer in order for the ligand and the protein to combine before protein precipitation. In both cases the recombinant expression was high, but the protein being isolated at the end of the IMAC chromatography was very low and, most importantly, did not increase compared to experiments made in absence of ITE. We did not have the means to discern if AhR precipitated because ITE did not bind to the AhR (either in the bacterial cells for the first experiment, or in the bacterial lysate in the second experiment) or if ITE did in fact bind to AhR but it could not prevent its precipitation.

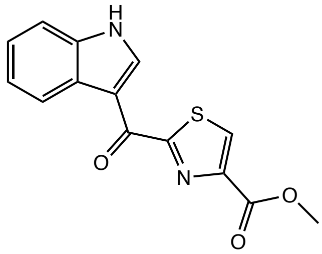


Figure 20. Structure of ITE (2-(1′H-indole-3′-carbonyl)-thiazole-4-carboxylic acid methyl ester)

Refolding. The refolding process is a well-established protocol for the expression and purification of recombinant proteins of low solubility[21,22]. On the expression step of protein production, the precipitation of the recombinant protein inside the bacterial inclusion bodies (IBs) is desired and favored rather than prevented. This is usually achieved by inducing the protein synthesis with high concentrations of chemical inducer (e.g. of IPTG in case of *lac* promoters) and conducting the protein synthesis at high temperature (e.g. 37 °C for *E. coli*). The IBs are located in the insoluble fraction of the bacterial lysate, together with genomic DNA, insoluble proteins, bacterial membranes and more. IBs are isolated from the rest of the insoluble fraction by multiple washes with aqueous buffers containing small amounts of detergent (Figure 21). The washed IBs are then solubilized in a denaturing buffer containing guanidinium chloride or urea to bring the insoluble proteins in solution in a denatured state. The concentration of the denaturing agent is then slowly lowered in a matter of minutes or hours using dialysis tubes or chromatographic techniques in order for the protein to refold on its own. We attempted the refolding process with constructs AhRa2, AhRf1 and AhRg3, all expressed at 37 °C. We attempted the synthesis of AhRa2 in IBs by inducing the synthesis at 37 °C and 20 °C. In the first case the protein was located primarily in the insoluble fraction of the bacterial lysate, as expected, but its overall yield was very low, while at 20 °C the protein yield was higher in the soluble fraction and comparable in the insoluble fraction (Figure 7 on page 11). This result was not expected as a higher expression temperature is typically associated with higher expression levels. For all constructs, the inclusion bodies were resuspended in 6 M guanidinium chloride and loaded in an IMAC column. Subsequently, the denaturant concentration was gradually lowered with a linear gradient of 100 column volumes lasting 14 hours. The refolded protein was then eluted with a buffer containing imidazole. Contrary to our expectations, the amount of protein isolated in AhRa2 and AhRf1 experiments was very low, much smaller than the amount obtained from a standard IMAC done in native conditions for the same constructs. This result is coherent with the fact that such construct tended to remain in the soluble fraction of the bacterial lysate. Construct AhRg3, on the other hand, had an intrinsically lower solubility than AhRa2 and AhRf1, and was found primarily in the pellet fraction at all temperatures tested (cf. Figure 17 on page 25). For this reason, the refolding process was more successful with this construct. We successfully isolated AhRg3-rich inclusion bodies from a culture grown at 37 °C for 6 hours (Figure 22) and subsequently isolated and refolded in an IMAC column the desired fusion protein (His-SUMO-AhRg3). After removal of the imidazole used to release the fusion protein, we cleaved the His-SUMO moiety by incubation with the Ulp1 protease. Unfortunately, all the released AhRg3 precipitated out of solution while only His-SUMO was visible in the second IMAC chromatography.

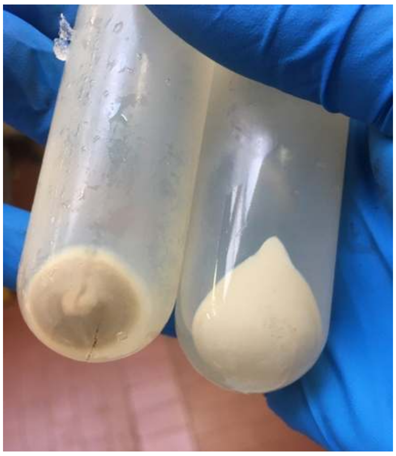


Figure 21. Washing process of the inclusion bodies. The crude insoluble fraction of the bacterial lysate (left tube) is washed multiple times in presence of detergent to isolate the inclusion bodies (tube on the right).

In conclusion, while for the AhRa2 and AhRf1 the refolding method proved to be less efficient than the extraction from the soluble fraction, in the case of AhRg3 the isolation and refolding from the IBs was successful. Unfortunately, even in the case of AhRg3, the refolded and cleaved protein is still very unstable and readily precipitates in absence of the SUMO fusion partner.

Chaperone co-expression. Foldases are a group of molecular chaperones that assist the folding process of nascent proteins. We experimented the co-expression of AhR construct AhRf1 with two foldases systems: GroEl-GroES and dnaK-dnaJ-grpE. For each system, the two-plasmid co-expression systems was used (Figure 9, ii, on page 13) employing commercially available chaperone-containing plasmids. Such plasmids contained a different antibiotic selection (essential for the co-transformation process) and a different inducer for recombinant protein expression (*ara* promoter compared to the *lac* promoter of the AhR constructs). These features enable the chaperones to be expressed separately from AhR constructs, more specifically they can be produced *prior* to AhR in order to obtain an optimal folding-assistance and protein-stabilization effects.

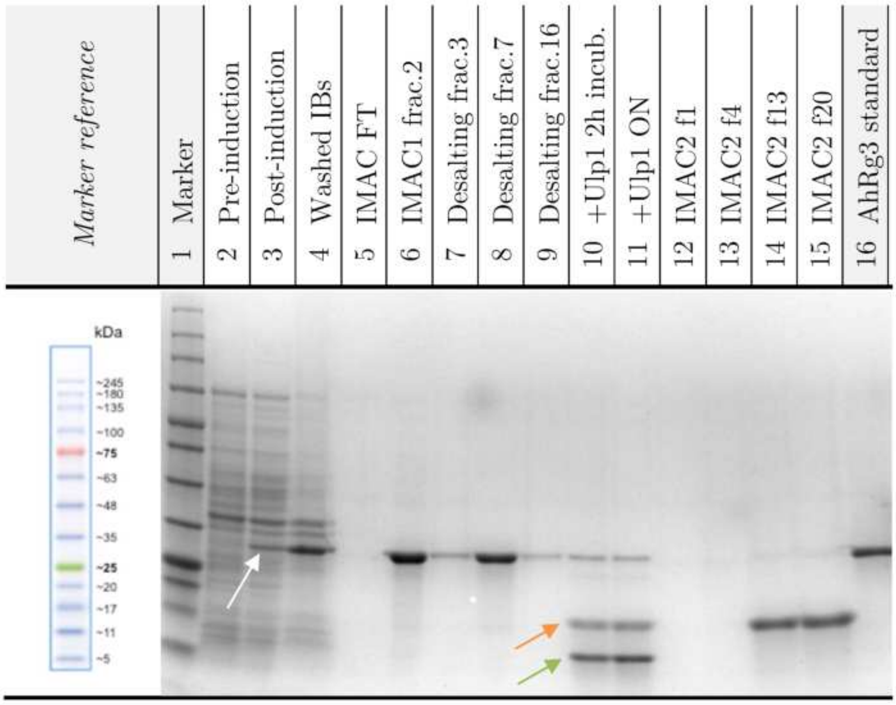


Figure 22. Isolation and refolding of AhRg3 from inclusion bodies. Overexpression of AhRg3 (white arrow) is visible after an induction at 37 °C for 6 hours. The insoluble fraction of the bacterial lysate was washed multiple times to isolate the inclusion bodies (lane 4). AhRg3 was successfully isolated and refolded in the IMAC1 chromatography (lane 6). Imidazole was subsequently removed with the desalting column (lanes 7-9). Ulp1 protease cleaved the His-SUMO-AhRg3 construct in 2 hours (lane 10, orange arrow: His-SUMO, green arrow: AhRg3), while further incubation overnight did not process the remaining uncleaved protein (lane 11). AhRg3 was lost after cleavage by simple centrifugation as it was not present in the flow-through of the IMAC2 chromatography (lanes 12-13), while His-SUMO remained in solution and was captured by the IMAC column (lanes 14-15).

The expression levels of AhRf1 in presence of any of the two foldases systems were very low or undetectable, while the overexpression of the foldases were high. This result is in part explained because the foldases synthesis was induced first with addition of arabinose during the early log phase of bacteria growth (OD=0.1) while AhR synthesis was induced with IPTG in the late log phase (OD=0.6). Only trace amounts of AhR were isolated by subsequent IMAC chromatographic steps, however. This result is similar to the attempted co-expression of AhRe1-2-3-4 with ARNTd1 (Figure 14 at page 22). In that case AhR expressed well in absence of ARNT while no detectable overexpression was seen in co-expression experiments. In the case of the experiments employing foldases, it was AhR that did not co-express in presence of the foldases, while the same constructs expressed well in absence their absence. Also in this case it is difficult to determine the exact cause of this behavior. Once again it is possible that the metabolic burden induced by multiple, concomitant protein expressions determines a lower yield of some proteins (some of which may be more susceptible than others).

Conclusions

Several approaches were tested in order to obtain a stable and soluble construct of AhR which also encompassed its PAS-B domain. Initial attempts were based on the synthesis of various constructs which only contained the PAS-B domain fused to the SUMO protein. Most of these chimeric constructs demonstrated a high overexpression and were localized in the soluble fraction of the bacterial lysate. Their fate – however – was to quickly precipitate out of the solution during subsequent chromatographic steps. For the most soluble construct (AhRa2), which could be isolated in small amounts and keeping its concentration low, the cleavage of the SUMO moiety resulted in the complete precipitation of the remaining protein. Nonetheless, this construct was the only one that we could initially isolate (even though in small amounts and as SUMO fusion proteins) to conduct some partial characterization such as the differential scanning fluorimetry that hinted to a possible stabilization effect of the zinc ions (Zn^{2+}) towards AhR. Unfortunately, such effect was either an artifact of the technique or it could not be fully exploited in our experimental conditions.

Our approach then focused on the co-expression of AhR with its natural partner ARNT in an attempt to isolate our desired protein in a dimeric complex. From these experiments we demonstrated that SUMO is essential not only for the solubility of AhR, but also for its overexpression. The co-expression experiments of AhR and ARNT proved challenging both in the cloning and the expression processes. All the AhR constructs in these experiments overexpressed less than their PAS-B-only counterparts and, more importantly, we noticed no sign of dimerization with ARNT. We could not conclude if this effect was caused because AhR precipitated before being able to bind to ARNT, or if the binding occurred but their association constant was not sufficient to counteract the AhR precipitation. We noticed, however, that ARNT never co-precipitated with ARNT.

When the first structures of AhR including the PAS-B domain were reported, we took inspiration from them to design new constructs and take advantage of the newly available structural information. All the constructs optimized on the basis of known structures did not yield the desired results, albeit one of them (AhRf1) demonstrated comparable performances to AhRa2. The double mutant AhRg3 was the construct that most overexpressed in the insoluble fraction of the bacterial lysate, and showed interesting results with the refolding method, but its solubility was still too low for any serious biophysical – not to mention structural – characterization.

Additional approaches like the addition of the strong agonist ITE and the co-expression with chaperones did not yield a soluble AhR PAS-B construct.

Even though the complete AhR cytosolic complex has been solved structurally by Cryo-EM, we believe that the availability of a small PAS-B-containing construct of AhR would still be very beneficial in the research activity around this important receptor. Moreover, there is not, to date, any experimental structure of the AhR-ARNT transcriptionally active complex which includes the PAS-B domains. Thus, an intrinsically stable AhR construct, which may be stable without the Hsp90 chaperone of the cytosolic complex would be desirable.

Part 2

The SARS-CoV-2 Main Protease (Mpro)

Introduction

Between the end of 2019 and the first months of 2020, a new virus quickly spread all over the world starting from mainland China. The term coronavirus, once common only among virologists, quickly became ordinary. The family of *Coronaviridae* is a family of positive-sense single-stranded RNA (+ssRNA) viruses known since the 1960s and includes several viruses that infect primarily mammalian hosts, some of which (MERS-CoV, SARS-CoV-1, SARS-CoV-2 and others) are infectious to humans[23]. Some coronaviruses such as MERS-CoV are highly lethal, while others (such as HCoV 229E) circulate annually and are associated with the common cold symptoms, others – like SARS-CoV-2 – lie in between. The quick outbreak of SARS-CoV-2 in 2020 found our society largely unprepared and inflicted large human and economic losses. Together with common hygiene and isolation practices, the most effective containment weapon developed against the new virus have been vaccines (especially those based on liposome-encapsulated messenger RNAs)[24,25]. These vaccines have the great advantage of being very quick to develop and have a low cost per dose. They have the drawback of being more sensible to virus mutations, however, as the antibodies produced after their administration can be escaped by new viral variants (as it has happened with the *delta* and *omicron* variants). Antiviral drugs commonly target enzymes that are vital for the virus replication or maturation; they take longer to develop and have a higher cost per treatment, but in turn they are more resistant to virus mutations. Our research group, following many others around the globe, began to study one of the most important proteins of SARS-CoV-2: its main protease (Mpro), one of the key enzymes involved in the viral maturation process and thus an ideal target for antiviral drugs. Our research work on the SARS-CoV-2 main protease hereby described is a minuscule, yet relevant share among the gargantuan discoveries that the pandemic stimulated in the scientific community worldwide. We believe that the results we obtained are meaningful and can support a deeper understanding of Mpro and can help the future discovery on new antiviral drugs.

An essential enzyme

Coronaviruses produce most of their proteins as two large polyproteins (pp1a and pp1ab) from two, large open reading frames (ORF1a and ORF1b, respectively)[26]. Such polyproteins are then post-translationally cleaved by two proteases (which are part of the polyproteins themselves), namely the main protease (Mpro, or 3C-like protease 3CLpro) and the papain-like protease (PLpro). These proteases cleave specific recognition sequences and release the individual proteins in a process known as polyprotein maturation (Figure 23). The proteins released from pp1a and pp1ab are non-structural proteins (NSPs), to differentiate them from the 4 structural proteins: spike (S), envelope (E), membrane (M) and nucleocapsid (N), which are synthesized as single proteins.

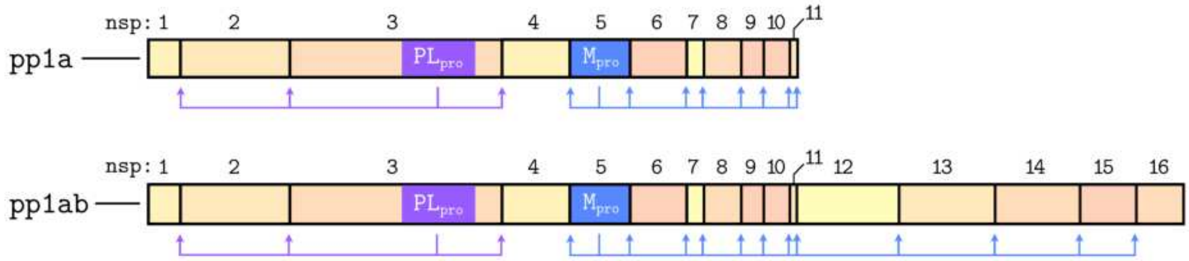


Figure 23. Autocleavage process of pp1a and pp1ab polyproteins. Mpro (nsp5) and PLpro (a domain of nsp3) operate the cleavage of the pp1a and pp1ab, releasing the individual non-structural proteins (NSPs) in a process known as polyprotein maturation.

Most NSPs are necessary for the virus to create the replication and transcription complex, but the function of some NSPs is still not known or fully understood, because most of them do not have homologs whose function is known. In the case of Mpro and PLpro their function if well established, and the absence of an homolog in mammals is a key advantage from a drug design point of view, as the possibility of off-target activity of an Mpro inhibitor is inherently low[27].

The recognition sequences of Mpro are highly variable but the amino acids closest to the cleavage site are mostly conserved (Table 9). Only for the central portion of the recognition sequence – namely Leu-Gln↓(Ser/Ala) – a clear pattern can be traced, with the glutamine in position P_1 being conserved among all *Coronaviridae*. To accommodate for such promiscuous substrates, Mpro has to adapt its shape to meet the necessary steric requirements, demanding a notable plasticity in the area around the active site[28,29].

Table 9. Recognition sequences of Mpro. Nsp4/5 indicates the cleavage sequence extending from the nsp4 *C*-terminus and the nsp5 *N*-terminus and so on. Notice how only the glutamine in position P_1 is completely conserved, while residues in position P_2 and P_1' are mildly conserved. Beyond these positions, no trend is found. The cleavage (indicated by ↓) occurs between P_1 and P_1'.

	P_6	**P_5**	**P_4**	**P_3**	**P_2**	**P_1**	↓	**P_1'**	**P_2'**	**P_3'**	**P_4'**	**P_5'**
nsp4/5	T	S	A	V	L	**Q**	↓	S	G	F	R	K
nsp5/6	S	G	V	T	F	**Q**	↓	S	A	V	K	R
nsp6/7	K	V	A	T	V	**Q**	↓	S	K	M	S	D
nsp7/8	N	R	A	T	L	**Q**	↓	A	I	A	S	E
nsp8/9	S	A	V	K	L	**Q**	↓	N	N	E	L	S
nsp9/10	A	T	V	R	L	**Q**	↓	A	G	N	A	T
nsp10/11-12	R	E	P	M	L	**Q**	↓	S	A	D	A	Q
nsp12/13	P	H	T	V	L	**Q**	↓	A	V	G	A	C
nsp13/14	N	V	A	T	L	**Q**	↓	A	E	N	V	T
nsp14/15	T	F	T	R	L	**Q**	↓	S	L	E	N	V
nsp15/16	F	Y	P	K	L	**Q**	↓	S	S	Q	A	W

Architecture of Mpro

The SARS-CoV-2 main protease adopts a structure remarkably similar to its SARS-CoV-1 homolog, with which it shares a 96% of sequence identity[30]. Mpro forms a homodimer in solution (K_d≈2.5 µM) with two identical monomers oriented roughly perpendicular to each other with a ~90° tilt (Figure 24). Each monomer is composed by 306 amino acids, weights 33.8 kDa and can be divided in 3 domains. Domains I (residues 8-101) and II (residues 102-184) are dominated by antiparallel β-sheet secondary structures and adopt a chymotrypsin fold, similar to other viral 3C proteases. Domain III (residues 201-303) is instead found exclusively in coronaviral proteases, it is structured with 5 α-helices and is mainly involved in the dimerization of the two monomers. Domains II and III are connected with a long, relatively unstructured linker of 16 amino acids (residues 185-200, Figure 24). A notable feature is the protrusion of the protein *N*-terminus (also known as "*N*-finger") of each monomer into the domain II of the other monomer, pointing to the active site. This deep insertion is a key structural feature of this enzyme, as it helps to shape the active site (in particular the subsite S_1, interacting with P_1). The Arg4 of the *N*-finger of each monomer also creates a salt bridge with Asp290 of the opposite monomer, increasing the association constant of the dimer. Key mutations on the dimerization interface (R298A in SARS-CoV-1)[31] or the deletion of Domain III in SARS-CoV-2 (our work) prevent the dimerization of Mpro leading to its complete loss of enzymatic activity.

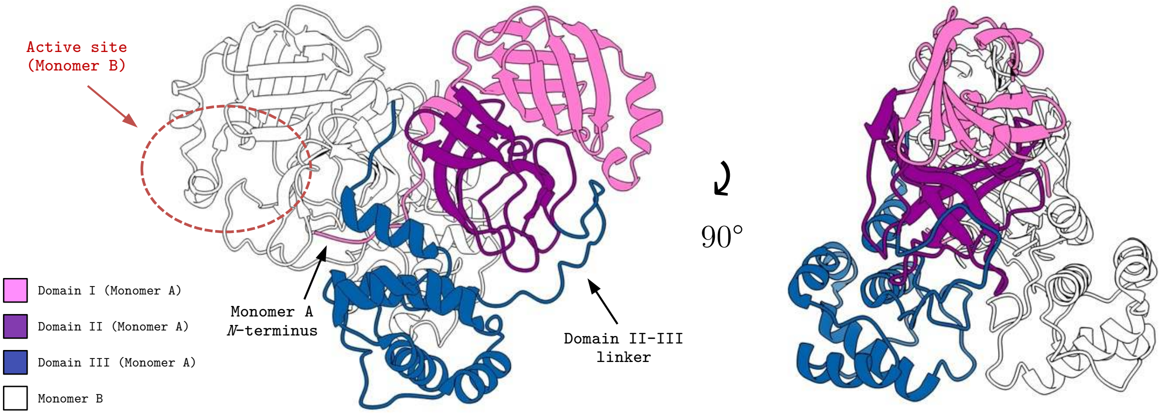


Figure 24. Domain architecture of Mpro. The main protease is present in solution as an homodimer. Monomer B is drawn in white for clarity. β-sheet-rich domains I and II (pink and purple for monomer A, respectively) adopt a classical chymotrypsin fold, while the α-helical domain III (in blue for monomer A) is mainly involved in the dimerization interface. The active site of monomer B is circled in red and is located between domains I and II.

Mpro is a cysteine protease which operates the catalytic cycle mainly through the catalytic dyad composed of Cys145 and His41 (Figure 25). Following the substrate approach, Cys145 is at first deprotonated by Nε-His41, then it operates the nucleophilic attack on the carbonyl of the peptide bond connecting position P_1 and P_1' of the substrate. Asp187 and His164 adjuvate this step increasing the basicity of His41 via a "catalytic" water molecule, facilitating the Cys145 deprotonation. The amidic nitrogen atom of P_1' receives a proton from Nε-His41 and another from the catalytic water molecule, leading the release of the P_1'-side hydrolysis product and the formation of the acyl-enzyme. The latter then de-acylates in an intramolecular rearrangement leading to the release of the P_1-side hydrolysis product.

The catalytic efficiency (k_{cat}/K_M) of SARS-CoV-2 Mpro for the NSP4/5 substrate is not very high (28500 $M^{-1}s^{-1}$), but is comparable to that of SARS-CoV-1 (26500 $M^{-1}s^{-1}$)[30]. Catalytic efficiencies for the other NSPs are lower and vary in a substrate-specific manner, suggesting that the recognition sequences have evolved together with the protease to regulate the polyprotein cleavage[32].

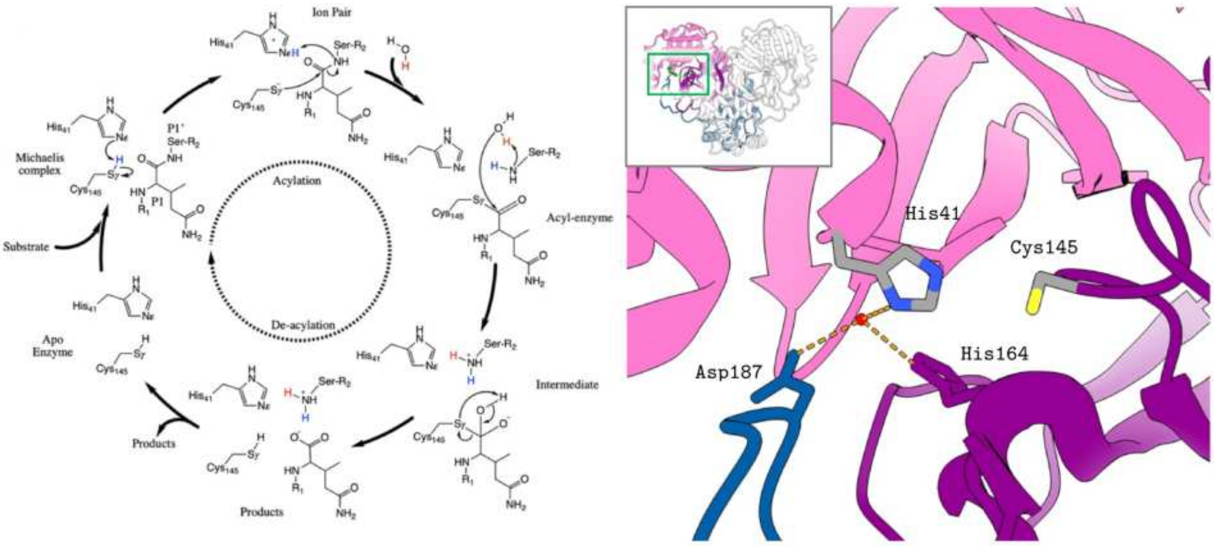


Figure 25. Catalytic cycle and active site of Mpro. On the left, the catalytic cycle of Mpro is depicted (scheme from Guzmán *et al.*[33]); His41 deprotonates Cys145, which operates the nucleophilic attack (acylation) on the peptide bond connecting residues P_1 and P_1' of the substrate. On the right, crystallographic structure of the active site of apo Mpro (our work). Asp187 and His164 increase the basicity of His41 *via* a "catalytic" water molecule (in red). Catalytic water enters the catalytic cycle as a proton and hydroxyl ion donor between the acylation and de-acylation steps.

Results and discussion

A new conformation for the WT Mpro

The first structure of the main protease of the SARS-CoV-2, was first reported by Zhang *et al.* in March 2020, about 3 months after the discovery of the new coronavirus[30]. We started the production of Mpro in our laboratories in summer 2020, thanks to the generous gift of prof. Hilgenfeld who sent us the wild-type Mpro sequence (Mpro^{WT}) cloned inside the pGEX-6P-1 expression vector (Table 14 on page 90). We started the recombinant production of such protein starting from their reported protocol and later adapting it to our necessities, leading to serious improvements in protein yield (from 2-3 mg/L of culture to 5-10 mg/L). The main modifications were the change in the expression medium (from YT broth to LB) and the switch for the expression vector from pGEX-6P-1 to pET-SUMO. The former construct produces a recombinant product with – starting from the *N*-terminus – a GST fusion protein, an Mpro auto-cleavage sequence, Mpro^{WT} sequence, a 3C PreScission cleavage sequence and an HisTag (Figure 26, on the left). The Mpro auto-cleavage sequence resembles the nsp4/5 recognition sequence and allows the recombinant protein to auto-cleave itself *in vivo* leaving an authentic *N*-terminus and releasing the GST protein which is discarded. Subsequent purification is aided by the *C*-

terminal HisTag, which can later be cleaved with a suitable 3C-protease, leading to an Mpro construct with authentic *N*- and *C*-termini. After a few months of successful expression of Mpro^{WT} with this construct, we cloned the same Mpro^{WT} sequence in the pET-SUMO vector (Figure 26, on the right). In this case the vector presents the HisTag at the *N*-terminus, followed by SUMO and finally by Mpro at the *C*-terminus. The cleavage site of the ULP-1 protease (SUMO protease) in indicated for clarity in Figure 26; it does not correspond to a sequence-based cleavage site, however, as ULP-1 recognizes the three dimensional structure of SUMO rather than a specific consensus sequence. Therefore, also for the pET-SUMO-based construct the *N*- and *C*-termini of the recombinant protein are authentic at the end of the purification process. The key advantages of this expression vector are the increased yield (thanks to the SUMO fusion protein, which acts as an expression and solubility enhancer) and the possibility to produce catalytically inactive Mpro constructs, which is not possible with the pGEX-based vector that relies on the auto-cleavage process. For this reason, in addition to the wild-type Mpro construct, all the subsequent constructs we prepared, which are mostly catalytically incompetent, have been cloned in the pET-SUMO vector (Table 14 on page 90).

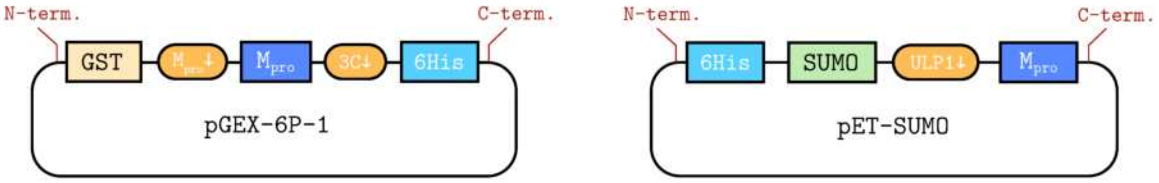


Figure 26. Comparison of Mpro expression vectors. On the left, the construct based on pGEX-6P-1 (from Zhang *et al.*)[30] produces a recombinant product with – starting from the *N*-terminus – a GST fusion protein, an M_{pro} auto-cleavage sequence, Mpro^{WT} sequence, a 3C PreScission cleavage sequence and an HisTag. The pET-SUMO-based construct (our work), on the right, presents – in the same order – an HisTag, the SUMO fusion protein, the Mpro^{WT} sequence. The ULP-1 (SUMO-protease) cleavage site is indicated between SUMO and Mpro for clarity, even though it does not correspond to a sequence-based cleavage site. Both methods produce the Mpro sequence with authentic *N*- and *C*-termini, but only the pET-SUMO-based platform is applicable to catalytically inactive Mpro sequences.

The first batches of Mpro^{WT} were produced transforming *E. coli* BL21(DE3) competent cells with the pGEX-6P-1 plasmid, then growing the transformed bacteria in YT medium and inducing the recombinant protein production with IPTG (Experimental details at page 81). At the end of the purification protocol, the protein was concentrated to 12 mg/mL and crystallization screening was started with PACT Premiere and Morpheus commercial kits (crystallization protocols at page 84). Initial hits obtained with the aid of a crystallization robot were later optimized by manual reproduction. The micro-seeding technique was determined to be essential for the obtainment of diffraction quality crystals (see Figure 58 on page 85). The best crystallization conditions for Mpro^{WT} (and for all the other Mpro constructs, except Mpro^{Cdel}) were: 0.1 M MMT (DL-Malic acid, MES monohydrate, Tris at 1:2:2 molar ratio), pH 7, PEG 1500 20%. We initially crystallized the Mpro^{WT} in *apo* condition (that is, in absence of any ligand) and in presence of known, commercially available Mpro inhibitors reported in the literature from drug repurposing (Figure 27, entries 1-7).

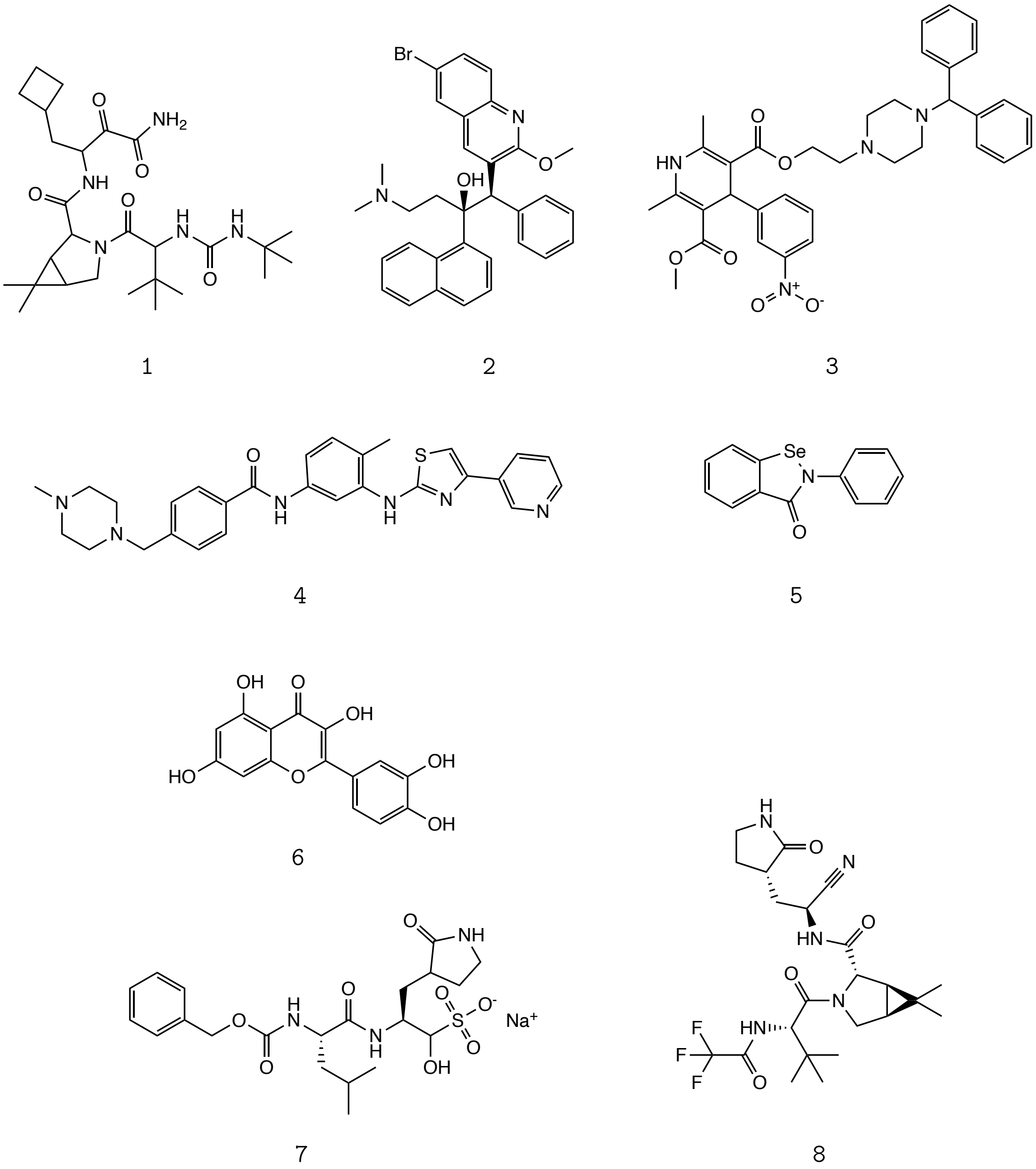


Figure 27. Known Mpro inhibitors from the literature, tested in our experiments. (1) Boceprevir, (2) Bedaquiline, (3) Manidipine, (4) Masitinib, (5) Ebselen, (6) Quercetine, (7) GC-376, (8) Nirmatrelvir.

The ligand-protein complexes were crystalized with the co-crystallization technique; a ligand stock solution in DMSO at 100 mM concentration was added to the protein solution at a 5% concentration (5 mM final concentration of ligand, vs 0.38 mM of Mpro, 13 molar excesses for the ligand). The ligand-protein complexes were incubated at 4 °C for 16 h prior to the crystallization experiments, which were carried out in parallel for all *apo* and *holo* conditions. After incubation with masitinib, manidipine and bedaquiline, a white precipitate appeared and the solutions were clarified by centrifugation at 17000 g; as the protein concentration was essentially unchanged after centrifugation, we concluded that the precipitate is composed of the inhibitors, which are poorly soluble in water. The fact that the protein was later crystallized under the same conditions as described for the free form further confirmed that its concentration was not altered by the centrifugation process. Mpro crystals appeared overnight for all the tested *apo* and *holo* conditions and completed growth in 2 days. Crystals were then flash-frozen and X-ray diffraction data were collected at beamlines ID23-1 and ID23-2 at the European Synchrotron Radiation Facility (ESRF, Grenoble, France).

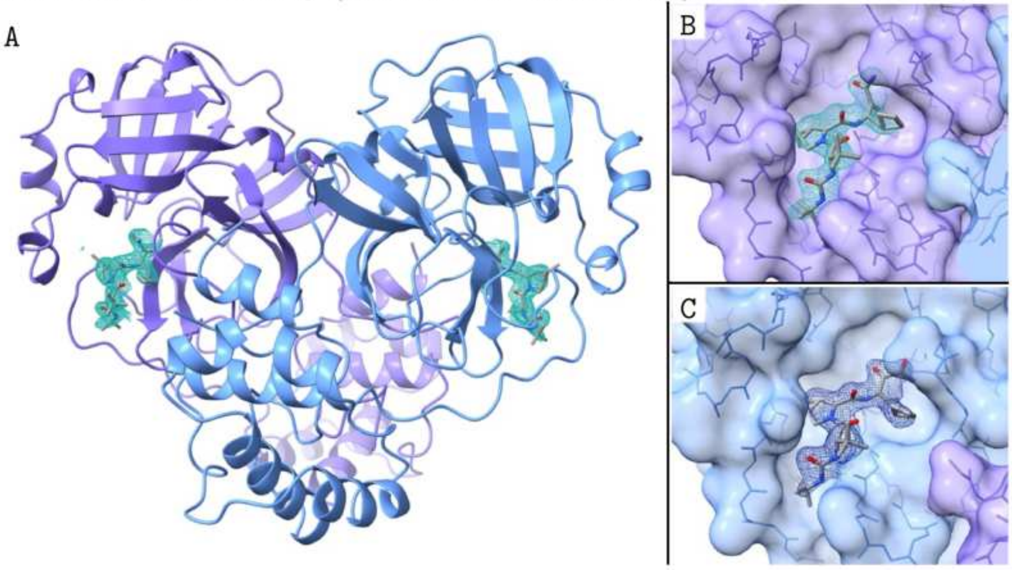


Figure 28. Mpro$^{\mathrm{WT}}$ in complex with boceprevir. Structure obtained by co-crystallization of boceprevir with Mpro$^{\mathrm{WT}}$ at 13 molar excesses. Panel A: overview of the structure; 2F$_{\mathrm{O}}$-F$_{\mathrm{C}}$ map is highlighted for boceprevir only for clarity (1.0 σ level). The inhibitor is present in the active site of both monomers at full occupancy (panels B and C).

The crystals were monoclinic (space group C2, with unit-cell parameters a=113.1 Å, b=54.7 Å, c=44.8 Å, α=90.0°, β=101.3°, γ=90.0°) with one monomer per asymmetric unit; the dimer is formed by the crystallographic two-fold axis. The phase problem for all structures was solved by molecular replacement (MR) with the previously reported SARS-CoV-2 Mpro structure by Zhang *et al.*[30] (PDB: 6Y2E). After successful molecular replacement and a first round of refinement, most structures obtained (including the complex with boceprevir) were virtually identical to those reported in the literature

(6Y2E for the *apo* form; 7W40, Andi *et al.*[34], for that in complex with boceprevir, Figure 28) and the electron density was clearly visible for the entire protein sequence. For ten structures, however, the electron density was of much lower quality or even absent in selected portions of the protein. Residues 1–3 of the *N*-finger, residues 139–144 (known as the *oxyanion loop*), and the side chain of His163 - all of which comprise the active site of the enzyme – did not match the MR model.

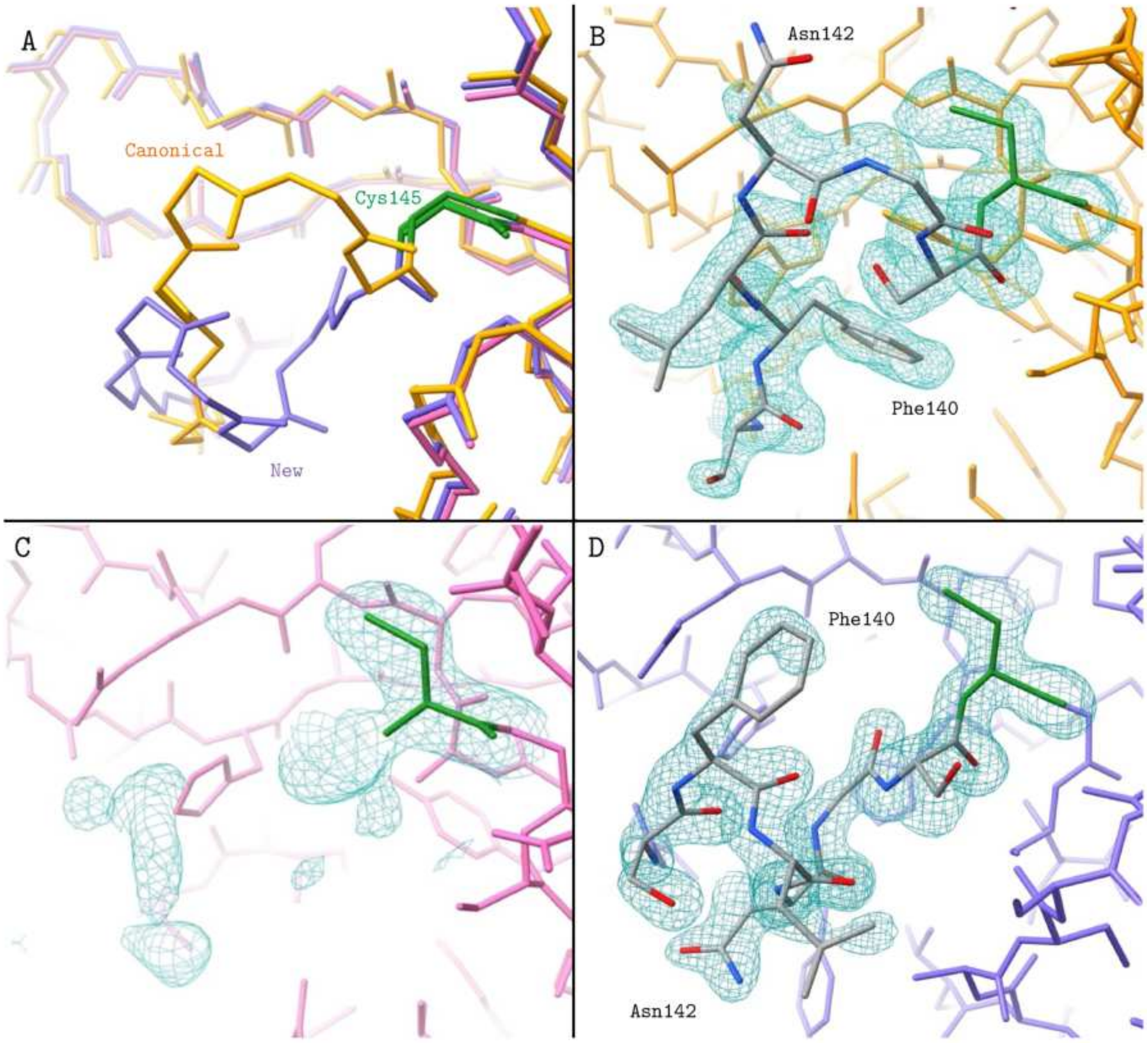


Figure 29. Experimental conformations of the oxyanion loop of Mpro^{WT}. Panel A represents the superposition of the conformations (side chains are omitted for clarity) in three scenarios (new conformation: purple, destabilized conformation: pink, canonical conformation: orange). Catalytic Cys145 is colored in green in all models. Panel B: electron density map (2F$_O$-F$_C$) and model for the canonical conformation. Panel C: electron density and model for the destabilized conformation (residues 139-144 are not present as the electron density is discontinuous). Panel D: electron density map and model for the *new* conformation obtained by co-crystallization with masitinib. Electron density is at 1.0 σ level for all maps and is restricted to the oxyanion loop for clarity.

Since the molecular replacement method used to solve the phase problem can introduce bias in the resulting structure, we repeated the MR with the same model but deprived of residues of ambiguous regions (Mpro$^{\Delta 1\text{-}3,139\text{-}144,\text{H164A}}$), as a search model. For most of the structures, a clear electron density was visible with all residues unambiguously in the active conformation (Figure 29, panel B). In some cases, the electron density was not sufficiently well-defined to consistently trace the mobile zones (Figure 29, panel C). For four structures, strikingly, it was possible to clearly model residues 1-3, 139-144, and the side chain of His163 in *new* conformations (Figure 29, panel D). We call this conformation *'new'* because there are no equivalents in Mpro structures deposited in the PDB. This novel conformation differs from the active conformations (both ours and previously reported ones), but it also differs from the literature-reported, inactive conformations from SARS-CoV-1 Mpro (including PDB entry 2QCY, where the oxyanion loop adopts a 3_{10}-helix conformation).

Mpro was found in the *new* conformation only in crystals grown from the enzyme pre-incubated with inhibitors masitinib, manidipine or bedaquiline. At the same time, the incubation with such inhibitors was not a determinant for the *new* conformation to develop, as Mpro adopted the active or the destabilized conformations in several crystals grown under the same conditions. Interestingly, no clear trace of such ligands was found in any experimental structure (neither in the active site, or elsewhere in the crystal structure), although their absence can be justified by their relatively high IC_{50} values in the range of 2.5-19 µM and low solubility[35,36]. We hypothesize that the *new* conformation and the canonical conformation are both thermodynamically stable, and that the presence of selected small ligands is able to shift the equilibrium towards the new conformation.

By close inspection the *new* conformation (Data collection and refinement details are reported in Table 20 at page 96) of the oxyanion loop approximately appears as a twisted version of the canonical, more linear conformation. The structure is not α-helical, but rather there are two successive β-turns with hydrogen bonds between Leu141 CO and Ser144 NH and between Ser144 CO and Ser147 NH and an α-turn with an hydrogen bond between Ser139 CO and Gly143 NH (Figure 30). This network of H-bonds clearly stabilizes the new conformation from a thermodynamic point of view. The twisting movement that is necessary for the transition from the canonical conformation to the *new* conformation also comes with a destabilizing effect; the hydrophobic side-chain of Phe140 is moved from a buried position to a solvent-exposed one (C_α is moved by 7.5 Å), while the opposite is true for the polar Asn142, whose C_α moves by 9.8 Å from solvent-exposed to a buried position (Figure 29, panels A, B and D).

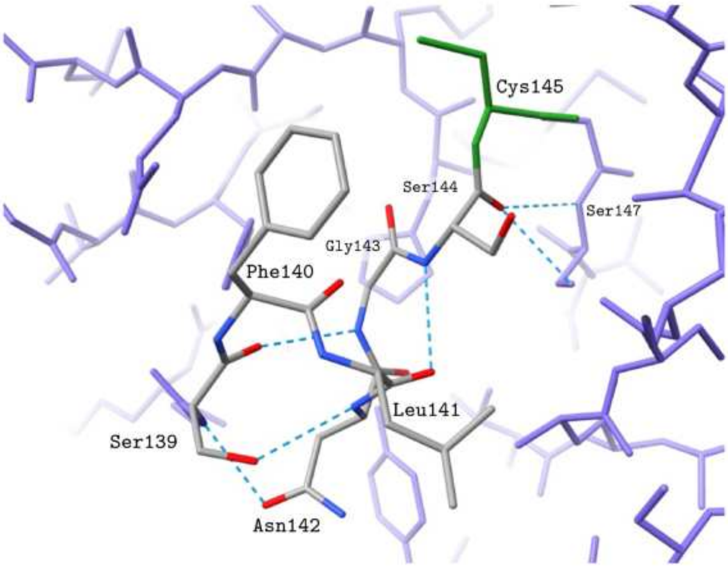


Figure 30. Hydrogen-bond interactions stabilizing the *new* conformation of Mpro^{WT}.

It is worth noticing that the positions of the residues of the catalytic dyad (His41 and Cys145) are virtually unchanged, similarly to the position of the catalytic water. It should be noted, however, that the new conformation of Mpro^{WT} is enzymatically inactive as the positions of several key residues - mainly those of the oxyanion loop - do not support the enzyme proteolytic activity. More specifically, the oxyanion loop is involved in shaping the subsite of the active site which is involved in the recognition of the P_1 position of the substrate (which is always a glutamine, see Table 9 at page 36). More importantly, the oxyanion loop is essential for the stabilization of the tetrahedral oxyanion intermediates of the acyl enzyme (hence its name) and for the reduction of the energy of the transition states associated with its formation (Figure 31).

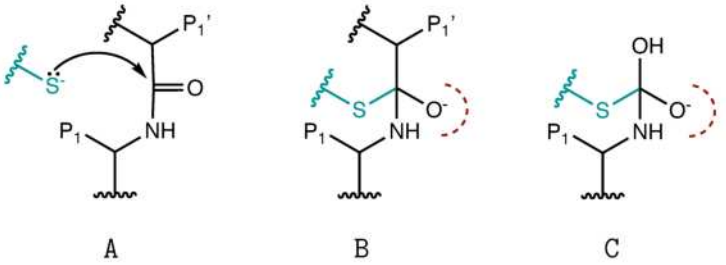


Figure 31. Formation and structures of the Mpro acyl enzyme intermediates. Active thiolate group from catalytic Cys145 (colored green) attacks the carbonyl group of the amidic bond between residues P_1 and P_1' of the substrate (A). The oxyanion loop (schematically represented by a red half circle) stabilize the negatively charged, tetrahedral oxyanion of the acyl enzyme intermediates of the catalytic cycle (B and C).

The perturbation of the oxyanion loop on a given monomer also modifies both the *N*-finger and the *C*-terminus of the other monomer, as they both lie in its close proximity. In active Mpro Ser1 of the *N*-finger is directly connected to Phe140 of the oxyanion loop and Glu166, but in the *new* conformation both interactions are lost, with the *N*-finger laterally moved away from the oxyanion loop. The perturbation of the *C*-terminus is even more pronounced as the six *C*-terminal residues (301-306) are not rigidly structured (as in the canonical conformation) and are not visible in the electron density in the *new* conformation, indicating a strong mobility.

The new conformation also differs from the canonical one in other aspects; the interface area between the two chains of the dimer is reduced from 1661 to 1273 Å^2, as evaluated by PISA analysis, which also assesses a sharp reduction in the number of hydrogen bonds and salt bridges. Nonetheless, key interactions that lead to Mpro dimerization, such as Glu290/Arg4' (salt bridge), Tyr126/Met6' (hydrophobic interaction) and Arg298/*N*-finger are present in the new structure as well, which is indeed dimeric.

Even though the *new* conformation of Mpro shows some radical changes in the oxyanion loop and in the vicinity of the active site, there is a clear indication that it can still bind natural substrates without major rearrangements. Superposition of the conformation of the *new* conformation Mpro^{WT} to the double-mutant (H41A,C145A), catalytically inactive Mpro^{DM} in complex with NSP4/5 (our work) do not show major protein-substrate clashes (Figure 32, Panel A). Small scale molecular dynamics simulations protracted by our colleagues at the Department of Pharmacy confirmed that only minor side chain movements are needed for the *new* conformation to accommodate the substrate.

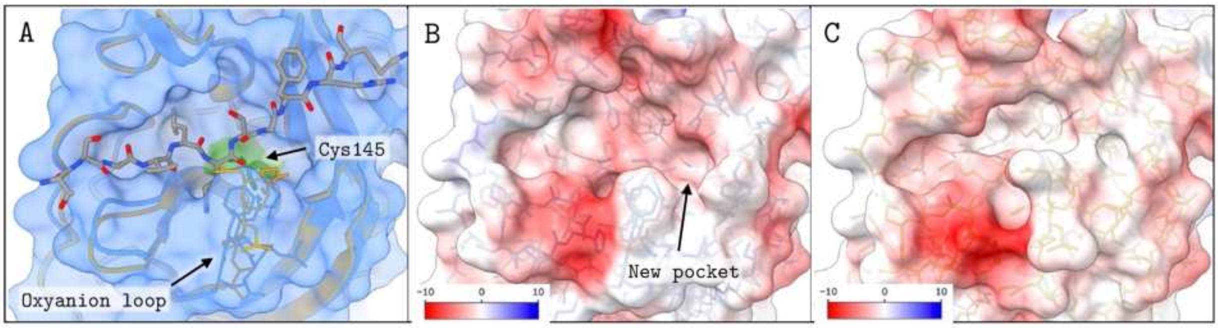


Figure 32. Active site surface of the *new* conformation of Mpro. Panel A: Superposition of the new conformation of Mpro^{WT} to Mpro^{DM} in complex with NSP4/5. Despite major conformational modifications in the oxyanion loop, the *new* conformation can still bind the substrate without major rearrangements. In panels B and C are depicted the electrostatic potential surfaces of the *new* and the canonical conformation, respectively. Notice in particular the new pocket that is formed in close proximity to the active site in the *new* conformation.

The rearrangement of the oxyanion loop in the *new* conformation also opens a new cavity in very close proximity to the catalytic Cys145 (Figure 32, panels B and C). The newly formed pocket is particularly interesting from a drug-design point of view for two reasons: firstly, it adds a new cavity for inhibitors to bind, which is particularly useful in the case of the relatively shallow Mpro active site. Secondly, an inhibitor that could take advantage of such cavity would lock the protein in an enzymatically inactive conformation, as the oxyanion loop would not be able to return to the canonical (enzymatically active) conformation.

Our colleagues at the Department of Pharmacy took this opportunity to design novel Mpro inhibitors based on the *new* conformation. After a docking-based initial screening and a molecular dynamics secondary selection, they proposed a selection of ~30 compounds (in two different series) which were then tested in a FRET-based enzymatic assay (a selection is reported in Figure 33). The compounds were then co-crystallized with Mpro^{WT}, and in case it failed, the soaking technique was also used.

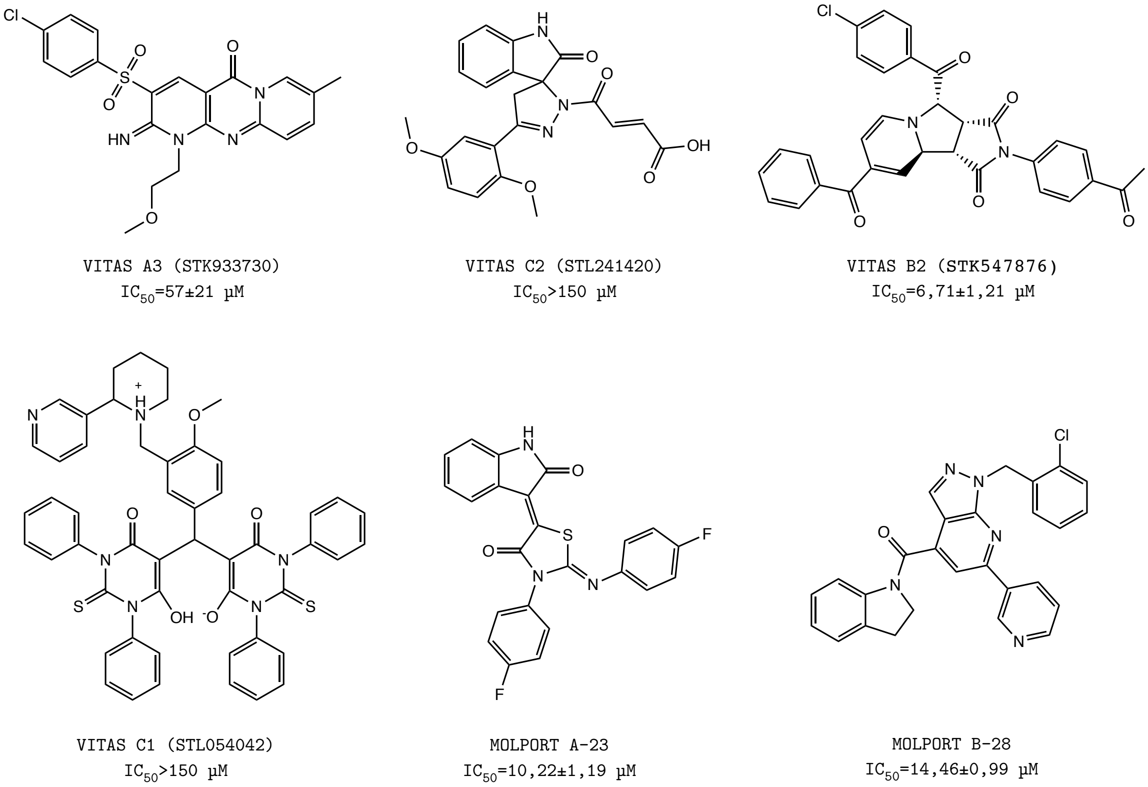


Figure 33. Selection of compounds designed to bind the *new* conformation of Mpro. Compounds Vitas C1 and C2 were the only ones that co-crystallized with Mpro, but they also showed the least inhibitory activity in FRET-based enzymatic assays. Compounds Vitas B2 and Molport A-23, on the other hand, were found to be strong inhibitors but we did not find them bound to Mpro either in co-crystallization of soaking crystallization experiments.

Of all tested compounds, unfortunately, only two of them (Vitas C1 and C2) were found in the electron density of the crystallographic structures obtained from crystallization experiments (for both inhibitors only part of the overall molecule was visible). For the remaining compounds, only the enzyme in the *apo* form was present. The biggest discrepancy, however, was that the compounds that showed the highest inhibitory activity in the enzymatic assay (e.g. Molport A-23 or Vitas B2) did not form complexes in co-crystallization experiments and did not bind to Mpro even in subsequent soaking experiments. At the same time, the two compounds that we found covalently bound to Mpro in co-crystallization experiments did not show any appreciable inhibitory effect in the enzymatic assays. Even though the IC_{50} value and the co-crystallization outcome should be correlated, the two experiments were indeed conducted under very different experimental conditions: 0.03 µM enzyme concentration and 0.25-150 µM compound concentration in the FRET-based enzymatic assay versus a ~300 µM enzyme concentration and 5 mM compound concentration in the co-crystallization experiment.

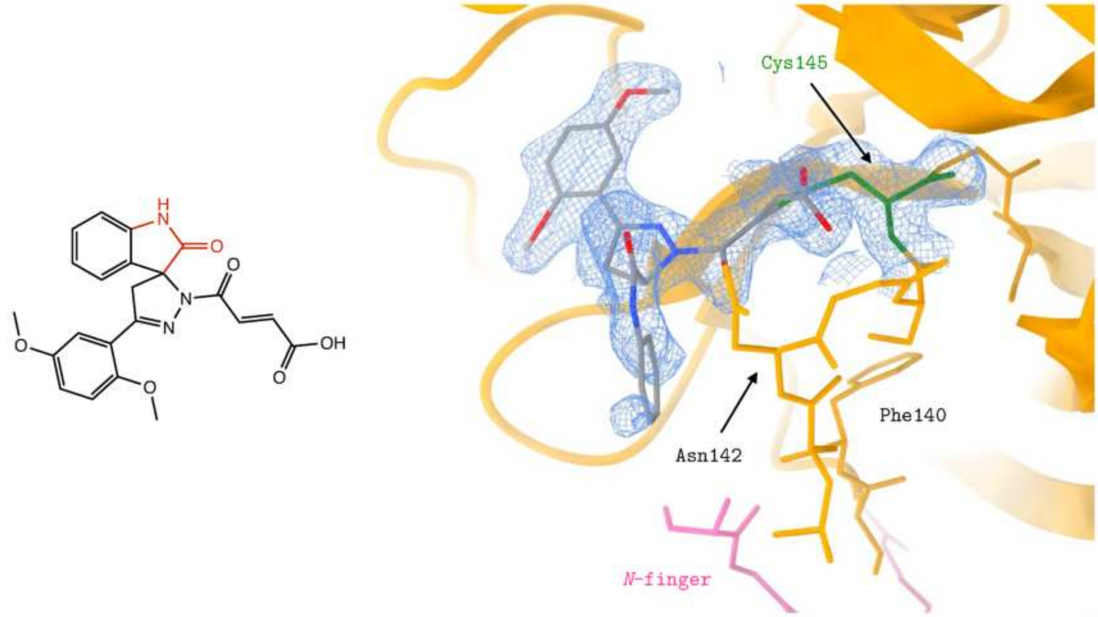


Figure 34. Mpro^{WT} in complex with compound Vitas C2. The electron density is contoured at the 1.0 σ level. The inhibitor is covalently bound to catalytic Cys145. The oxyanion loop and the *N*-finger adopt the canonical conformation. Of the whole Vitas C2 compound, a small portion (colored red in the schematic structure on the left) is not present in the electron density.

The most striking result was that in the two structures of Mpro^{WT} in complex with Vitas C1 and Vitas C2, the enzyme adopted the canonical conformation both in the oxyanion loop and the *N*-finger. A possible explanation for this result is that the computational models that predicted a strong binding to the *new* conformation of Mpro did not consider the possibility that also the canonical conformation could be a suitable target. Another plausible explanation is that some of the molecules used in our experiments were degraded (prior to our use or in experimental conditions), as only parts of them were visible in the electron density. In any case, given that a large portion – rather than a small fragment – of the compounds was clearly present in the active site of the enzyme in very close proximity of the oxyanion loop, we deemed unlikely that even the unfragmented molecules could induce Mpro in the *new* conformation.

The *new* conformation of Mpro^{WT} here presented clearly represents a *local* minimum in potential energy as it could be crystallized, and the crystals obtained produced an unambiguous electron density in the region of the oxyanion loop. At the same time, the *global* minimum in potential energy is found in the canonical conformation, as it is the only form found in *apo* condition and it is also the predominant form when substrates are present. As our experiments demonstrated, however, the *new* conformation is intrinsically inactive and can be reliably induced by the presence of suitable small molecules. These inhibitors would be part of a new class on their own and would open new possibility in the development of new drugs against SARS-CoV-2.

The results obtained regarding the new conformation of MproWT have been published in an article[37], a copy of which is attached in the final pages of this thesis.

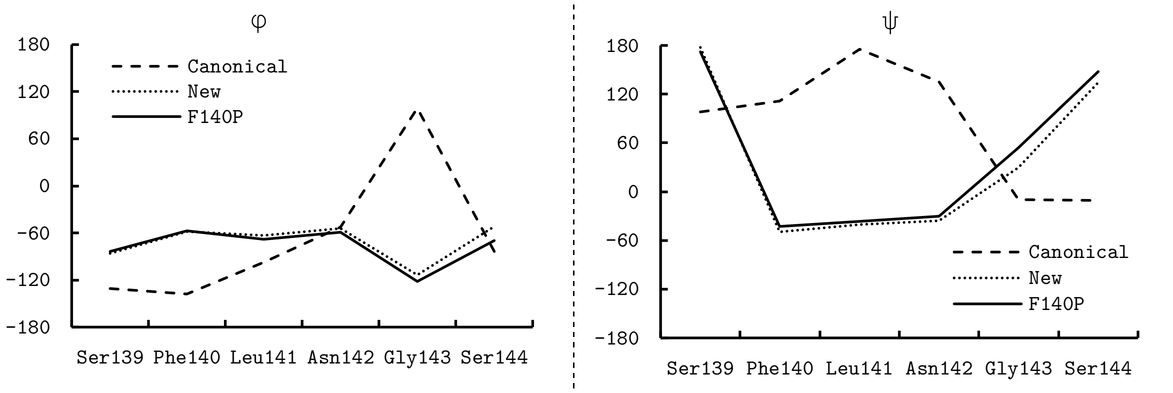
A helical oxyanion loop: Mpro^{F140P}

Since it was not possible to reproduce Mpro^{WT} in the *new* conformation without co-crystallizing it with selected small-molecule inhibitors (masitinib, bedaquiline or manidipine) we designed a new mutant that would arrange in the *new* conformation in *apo* condition. The φ dihedral angle of Phe140 in the canonical conformation is -137.8, while in the *new* conformation is -58.0°, which is very close to the -60° angle of the conformationally-constrained proline residue (Table 10). The ψ dihedral angles of the canonical and the *new* conformation are also very different (+111.2° and -49.5°, respectively) but are both in the Ramachandran-allowed region of the proline amino acid. We hypothesized that the Mpro^{F140P} mutant would adopt the *new* conformation of the oxyanion loop in an irreversible way, due to the conformational constraints of the proline amino acid. To assess possible perturbations of the oxyanion loop, we decided to clone Mpro^{F140P} and Mpro^{F140A} mutants, with the latter acting as a control.

We successfully obtained both mutants in the pET-SUMO expression vector using the site-directed mutagenesis method starting from the Mpro^{WT} construct (see Table 14 on page 90 and Table 17 on page 93). The constructs were then transformed and expressed in *E. coli* bacteria of the BL21(DE3) strain. Both constructs showed good overexpression levels, with yields comparable to the wild-type variant (~5 mg of purified protein per liter of bacterial culture). The last purification step of size-exclusion chromatography confirmed that both proteins are present in solution as dimers around the 1 mg/mL concentration range. The enzymatic activities of the two constructs were tested with the same protocol used for the wild-type variant (a FRET-based enzymatic assay); while Mpro^{F140A} showed a reduced activity compared to Mpro^{WT}, Mpro^{F140P} did not show any enzymatic activity. The two proteins were then employed in crystallization experiments both alone and in presence of the same inhibitors tested for the Mpro^{WT} (Figure 27 at page 41 and Figure 33 at page 47). For both constructs various crystallization conditions were tested using commercially-available crystallization kits (PACT premiere and Morpheus, Molecular Dimensions), but it was found that the ideal condition was identical to Mpro$^{\text{WT}}$ [0.1 M MMT (DL-Malic acid, MES monohydrate, Tris at 1:2:2 molar ratio), pH 7, PEG 1500 20% or 25%] with the exception of the protein concentration, which was 7 mg/mL for $\text{M}_{\text{pro}}{}^{\text{F140A}}$ and 9 mg/mL for Mpro$^{\text{F140P}}$ versus 12 mg/mL for Mpro$^{\text{WT}}$. Both proteins crystallize only using the micro-seeding technique, using a homologous seed stock solution or one obtained from Mpro$^{\text{WT}}$ crystals. The alanine mutant was found to be harder to crystallize, with crystals appearing as thin plates of small size. The proline mutant, however, crystallized more easily, almost on par with the wild-type mutant.

Table 10. Backbone dihedral angles of the oxyanion loop. From left to right, the dihedral angles (φ and ψ) of the oxyanion region for the canonical conformation of Mpro$^{\text{WT}}$, the *new* conformation of Mpro$^{\text{WT}}$, and for Mpro$^{\text{F140P}}$. The dihedral angles of residue 140 (Phe or Pro) are highlighted in red. Notice the high similarity of the *new* conformation and Mpro$^{\text{F140P}}$, as opposed to the canonical conformation. Graphical plots of the dihedral angles are displayed below.

Canonical conformation			*New* conformation			$\text{M}_{\text{pro}}{}^{\text{F140P}}$		
Residue	**φ**	**ψ**	**Residue**	**φ**	**ψ**	**Residue**	**φ**	**ψ**
Ser139	-130.5	+98.5	Ser139	-86.2	+177.5	Ser139	-83.2	+172.1
Phe140	-137.8	+111.2	Phe140	-58.0	-49.5	Pro140	-57.3	-42.8
Leu141	-97.7	+175.3	Leu141	-63.2	-40.4	Leu141	-68.0	-36.8
Asn142	-52.2	+135.1	Asn142	-54.5	-35.6	Asn142	-59.2	-30.5
Gly143	+99.0	-9.9	Gly143	-113.5	+29.8	Gly143	-121.8	+54.4
Ser144	-83.3	-10.8	Ser144	-51.8	+134.6	Ser144	-69.4	+147.5



While both constructs were crystallized in presence of inhibitors and peptides mimicking the natural substrates of Mpro, all crystals obtained were in *apo* form. In the crystals of both mutants the protein tended to arrange in a crystallographic dimer in the same space group (C2) and the same unit cell as Mpro^{WT}, with only one monomer per asymmetric unit and the other monomer being recreated by the twofold rotation symmetry axis. Even though no ligand was found in the electron density of the mutant crystals, the analysis of the obtained structures was quite revealing. The dihedral angles of the oxyanion loop (residues 139-144) of Mpro^{F140P} are superimposable to the *new* conformation of Mpro^{WT} (Table 10) with an overall RMSD of 0.266 Å for the protein backbone.

The similarities of Mpro^{F140P} and the *new* conformation of Mpro^{WT} are not limited to the oxyanion loop, as both the backbone and the sidechains of the whole sequence are in very similar positions (RMSD 0.418 Å for the backbone and 0.664 Å for the side chains). Other peculiar features of the *new* conformation of Mpro^{WT} are found in the proline mutant as well, such as the flexibility of the *C*-terminus (residues 301-306 are not resolved) and the position of the *N*-finger.

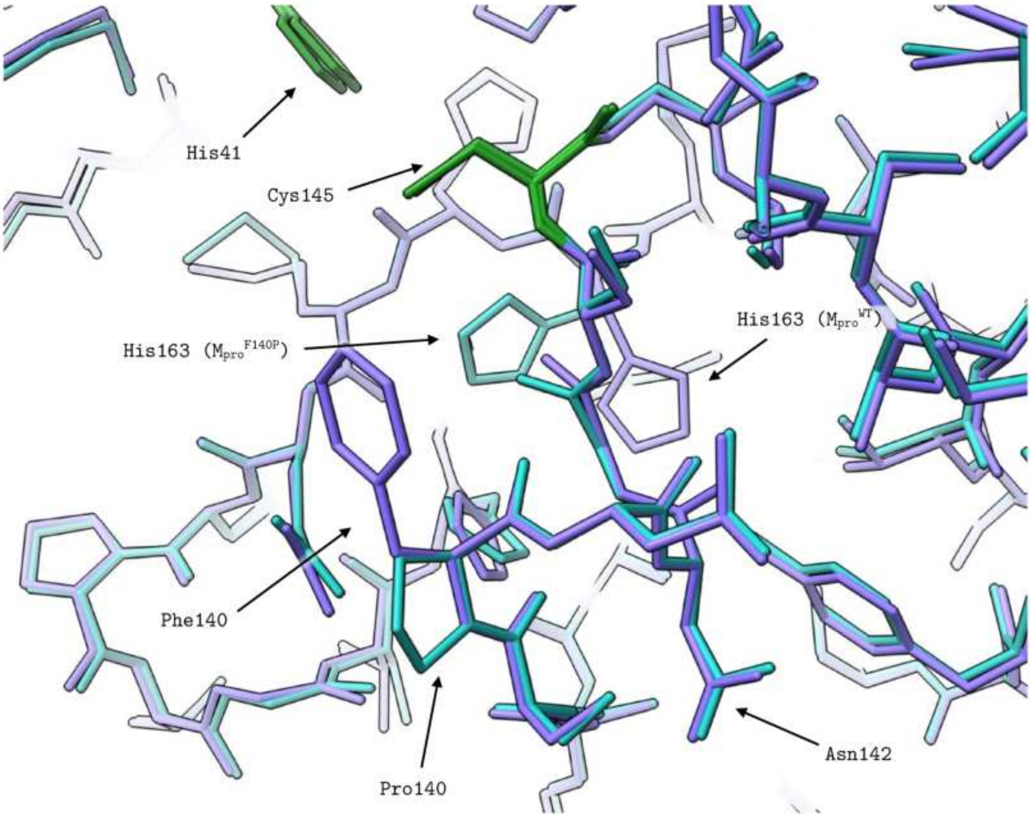


Figure 35. Structural details of Mpro^{F140P}. The proline mutant (in cyan) is compared with the *new* conformation of Mpro^{WT} (in purple). The backbone trace of the oxyanion loop is superimposable (RMSD 0.266 Å), as the only difference is visible in the side chain of the mutated residue 140. The catalytic amino acids His41 and Cys145 (colored in green) are in identical positions. The side chain of His163 is the most perturbed between the two structures (RMSD 3.501 Å) and the only significant difference if non-solvent-exposed residues are considered.

The active sites of the two structures are also very similar, as both the catalytic dyad and the oxyanion loop both occupy virtually identical positions (Figure 35). The side chain of His163 is the most perturbed between the two structures (RMSD 3.501 Å for the side chain, 0.342 Å for the backbone) and the only significant difference if non-solvent-exposed residues are considered.

Protein crystals were obtained for the Mpro^{F140A} mutant as well, though its propensity to crystallize was markedly lower compared to Mpro^{F140P} and Mpro^{WT}, generating smaller and less diffracting crystals. All crystallographic structures obtained of Mpro^{F140A} were in the *apo* form, even though several experiments were conducted in presence of small-molecule inhibitors and peptidic enzymatic substrates both by co-crystallization and by soaking. The Mpro^{F140A} mutant consistently crystallized in the C2 space group with one molecule per asymmetric unit. The fold of the alanine mutant is superimposable to the canonical conformation of Mpro^{WT} (RMSD = 0.286 Å). Key residues are missing in the electron density, however, namely residues 140-146 which comprise most of the oxyanion loop and the catalytic Cys145.

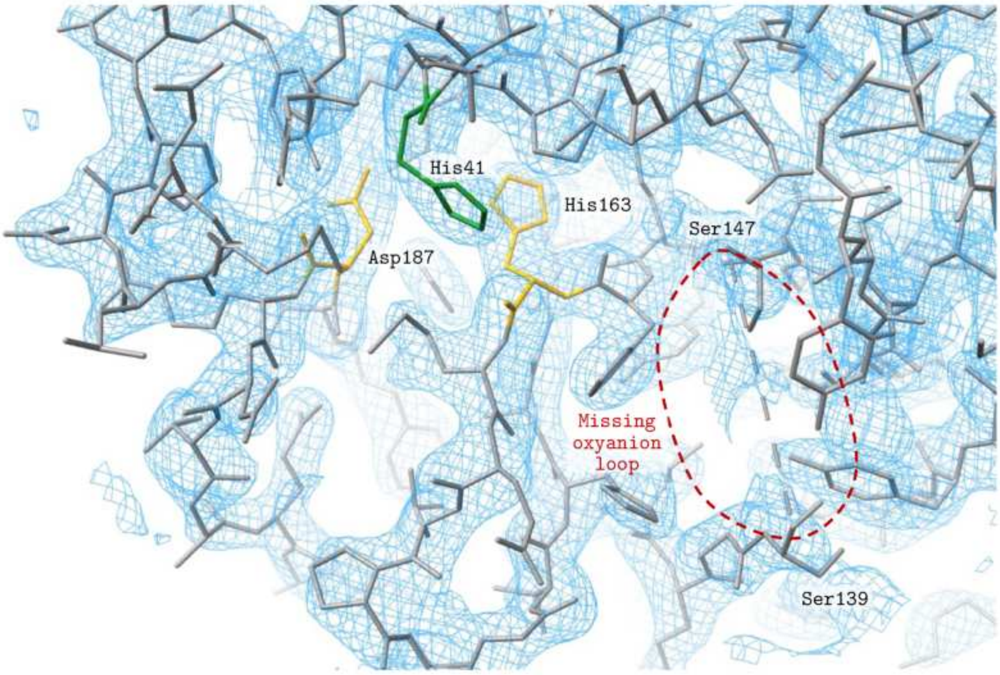


Figure 36. Crystal structure of Mpro^{F140A} in the vicinity of its active site. The electron density of the $2F_O$-F_C map is drawn at the 1σ level. Only residue His41 (colored in green) of the catalytic dyad is present in the electron density. Residues 140-146, which comprise most of the oxyanion loop and the other catalytic residue (Cys145), are not visible instead, indicating a large destabilization effect induced by the F140A mutation. Tyr118 is also missing in the electron density. Other important residues for the active site, such as Asp187 and His163 (in orange) – which coordinate the *catalytic* water with His41 – are present in their canonical position.

This marked destabilization effect is both more pronounced (as it regards also residues 145-146) and more ubiquitous (as it does occur also in absence of ligands) than what we observed for Mpro^{WT} (see Figure 29, Panel C, at page 43). Since this destabilization effect is not dependent on the crystallization condition and is locally confined to the oxyanion loop, we hypothesize that it is likely caused by the F140A mutation.

An Mpro with no partner: Mpro^{Cdel}

Since the domain III of the SARS-CoV-2 main protease is necessary for the homodimerization of the enzyme and is thought to be essential to retain its catalytic activity, we cloned a new construct deprived of such domain. The new construct, named Mpro^{Cdel}, lacks the α-helical domain III but retains the long linker that connects it to domain II (see Figure 24 on page 38). The linker was maintained as one of its residues, namely Asp187, coordinates the catalytic water in the active site of the enzyme (see Figure 25 at page 39).

Mpro^{Cdel} was cloned in the pET-SUMO expression vector with the restriction-free cloning technique (experimental details at page 78). The purification protocol was analogous to the other Mpro constructs and did not present significant differences. It was found, however, that the protein was present in solution uniquely as a monomer, as it eluted from the size-exclusion chromatographic step at higher retention volumes compared to the other Mpro constructs (Figure 37).

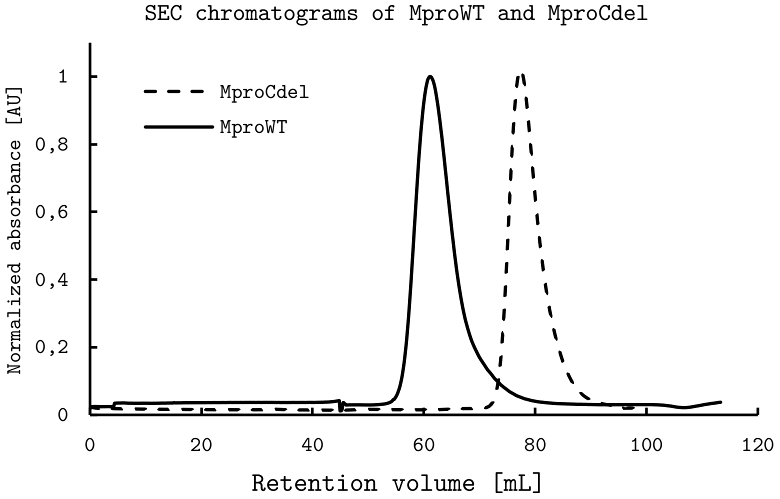


Figure 37. SEC chromatograms of Mpro^{Cdel} and Mpro^{WT}. The molecular weights of the Mpro^{WT} dimer and the Mpro^{Cdel} monomer are 67.6 kDa and 21.9 kDa, respectively. The retention times for the two peaks correspond to ~65 kDa for the first one and ~20 kDa for the second one, thus confirming that Mpro^{WT} is present as a dimer and Mpro^{Cdel} as a monomer. The chromatograms were obtained using an HiLoad 16/600 Superdex 75 size-exclusion chromatography column.

Mpro^{Cdel} was subsequently employed in crystallization experiments. The protein was found to be the only Mpro construct not crystallizing with the same precipitant mixture as the other constructs (0.1 M MMT pH 7, PEG 1500 20-25%). A large number of different crystallization conditions was tested thanks to crystallization screenings (Index HT, Crystal Screen HT and PEG-RX HT from Hampton Research; JCSG-plus, PACT premier and Morpheus from Molecular Dimensions; JBScreen Basic HTS from Jena Bioscience). Crystalline material appeared only in a handful of conditions, and a single crystal was obtained in condition G3 of the Morpheus screening after about 3 months of incubation. Further attempts to recreate the crystallization condition either with the same or with other precipitants did not lead to new crystals. Attempts to use the microseeding technique from available crystalline material also failed. It was possible, however, to resolve the *apo* structure of Mpro^{Cdel} from the only available, diffraction-quality crystal. Diffraction data of the crystal structure revealed that the protein arranged in the $P2_12_12_1$ space group with two molecules per asymmetric unit (Figure 38). Data collection and model statistics are reported on Table 19 at page 95. The two subunits are arranged in a crystallographic dimer – rather than a constituent one – as their arrangement and reciprocal orientation is different from those of Mpro^{WT}.

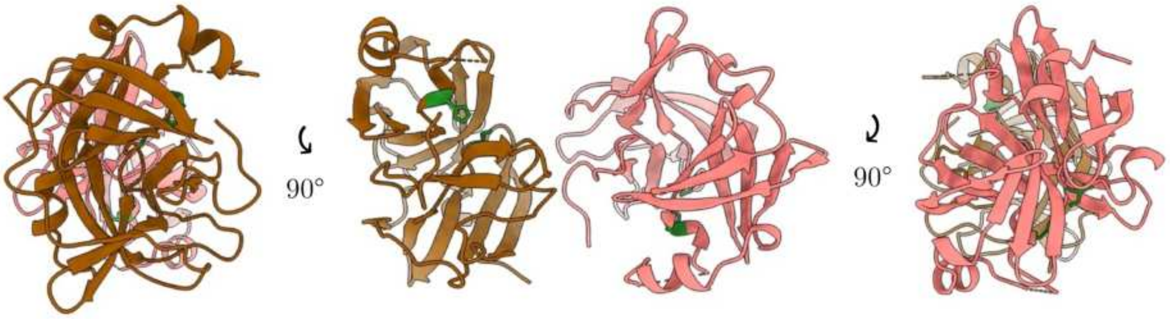


Figure 38. Arrangement of the two molecules of Mpro^{Cdel} in the crystal ASU. The arrangement and the reciprocal orientation (~180°) of the two chains is different from those of Mpro^{WT}. The catalytic dyad is colored in green.

For each molecule, the electron density of three sections of the protein chain was missing due to their mobility (Figure 39, panel A), namely residues 1-4 (*N*-terminus), 47-50 (loop section) and 188-199 (Domain II-III linker) for subunit A and residues 1-5 (*N*-terminus), 45-51 (loop section) and 188-199 (Domain II-III linker) for chain B. The mobility of the *N*-terminus and the Domain II-III linker can be explained by the absence of the dimerization partner, as both sections are stabilized by the dimerization partner of the opposite chain in wild-type Mpro. The loop section spanning residues 45-51 is not involved in crystal contacts and is exposed to the solvent, thus the absence of the electron density in this area is related to the loop's mobility.

A notable feature of the crystallographic structure of Mpro$^{\text{Cdel}}$ is the presence of residue Asp187 in subunit A and its absence in subunit B. Such residue is essential for the coordination of the so-called catalytic water, together with His41 and His164 (see Figure 25 at page 39). In chain A of Mpro$^{\text{Cdel}}$, Asp187 forms three hydrogen bonds with Arg40 and Tyr54, and coordinates the catalytic water as well (Figure 39, panel B). Interestingly, the electron density of Asp187 is not visible in chain B. The last visible residue at the *C*-terminus of chain B (Val186) is pointing towards the solvent, indicating that in this case Asp187 is highly mobile and is not kept in place by the interactions with Arg40 and Tyr54. The catalytic water for chain B is still present, though, indicating that the coordination by His41 and His164 is sufficient to keep it in place.

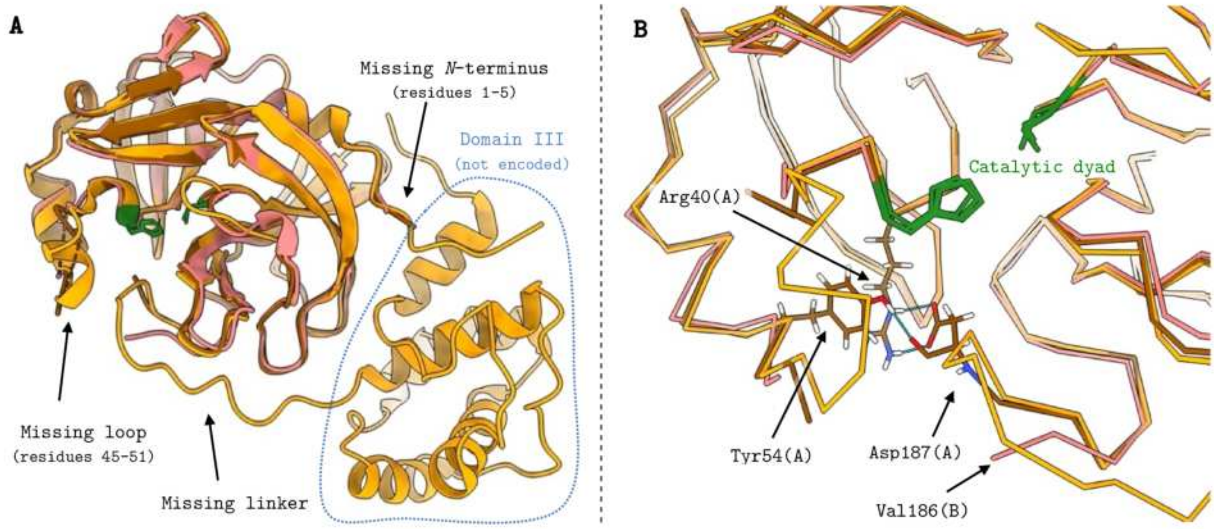


Figure 39. Architecture of Mpro$^{\text{Cdel}}$ in comparison with Mpro$^{\text{WT}}$. Panel A: Chains A (in brown) and B (in pink) of the X-Ray structure of Mpro$^{\text{Cdel}}$ are aligned with the canonical structure of Mpro$^{\text{WT}}$ (in orange). Several portions of Mpro$^{\text{Cdel}}$ are missing from the electron density due to their mobility, namely the *N*-terminus, the loop comprising residues 45-51, and the linker connecting Domain II and III. Domain III is naturally absent from the Mpro$^{\text{Cdel}}$ structure as it was not encoded in the expression vector. Panel B: Small differences are present in the missing residues of Chain A and B of Mpro$^{\text{Cdel}}$; in chain A the last visible residue at the *C*-terminus is Asp187 which creates three hydrogen bonds with Arg40 and Tyr54, while in chain B Asp187 is not visible as the last visible residue at the *C*-terminus is Val186, which points towards the solvent.

Another interesting feature of the crystallographic structure of Mpro$^{\text{Cdel}}$ is the conformation of the oxyanion loop. For this region the electron density of all residues (139-144) is well defined for both the backbone and the sidechains, with no significant differences between subunits A and B (Figure 40, panel A). The configuration adopted by the oxyanion loop by Mpro$^{\mathrm{Cdel}}$ is different from that of the canonical conformation of Mpro$^{\mathrm{WT}}$ but is different from the *new* conformation as well (Figure 40, panels B and C, respectively). The RMSD between the oxyanion loop residues is 8.36 Å and 6.83 Å, respectively, to be compared with an RMSD<1 Å for the backbone of the whole chain when compared to both Mpro$^{\mathrm{WT}}$ conformations. This result, together with the notable features of the oxyanion loop found in the *new* conformation of Mpro$^{\mathrm{WT}}$ and the destabilization induced by the F140A mutation, indicate that this region plays a crucial role in the enzyme functioning and it is deeply influenced by the presence of small molecules (in Mpro$^{\mathrm{WT}}$), the presence of the dimerization partner (Mpro$^{\mathrm{Cdel}}$) or by non-drastic point mutations (Mpro$^{\mathrm{F140A}}$).

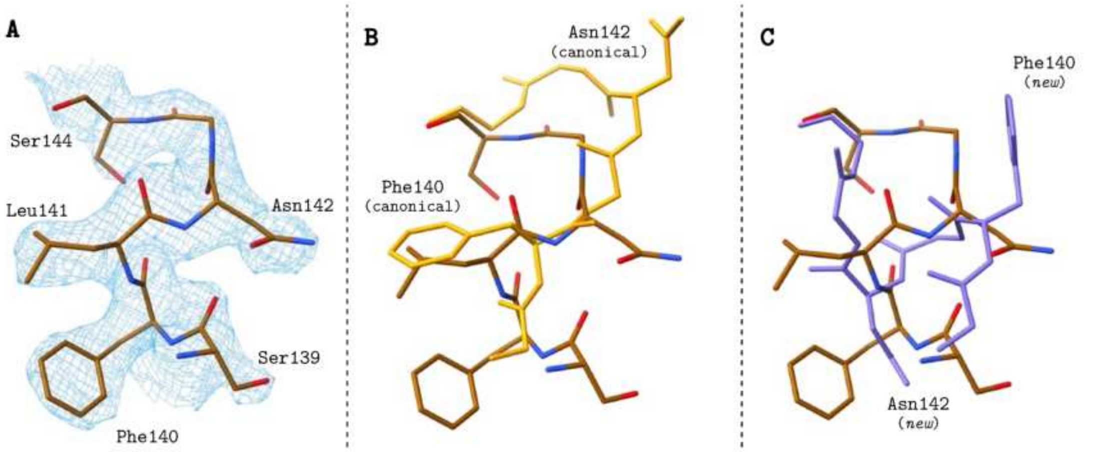


Figure 40. Oxyanion loop conformation of Mpro$^{\mathrm{Cdel}}$. The configuration of the oxyanion loop (residues 139-144) of chain A of Mpro$^{\mathrm{Cdel}}$ with the electron density map at the 1σ level is depicted in panel A. The oxyanion loops of chains A and B of Mpro$^{\mathrm{Cdel}}$ adopt the same conformation. For comparison, the oxyanion loops of Mpro$^{\mathrm{WT}}$ in the canonical and in the *new* conformation are presented in panels B and C, respectively. The arrangement of the oxyanion loop residues of Mpro$^{\mathrm{Cdel}}$ is different from both conformations of Mpro$^{\mathrm{WT}}$, with a RMSD between the residues of 8.36 Å and 6.83 Å, respectively.

An incompetent Mpro: Mpro^{DM}

To study the interaction of Mpro with its substrates, we cloned a double-mutant Mpro (Mpro^{DM}) deprived of the catalytic dyad (H41A, C145A) by the site-directed mutagenesis method starting from the wild-type sequence inserted in the pET-SUMO vector. The resulting plasmid was sequenced to verify the successful mutation process. The protein was then produced and purified using the same protocol as Mpro^{WT} with slightly higher yields, possibly because the wild-type, catalytically active construct may be relatively toxic to the *E. coli* cells. Mpro^{DM} eluted from the size-exclusion chromatography step at the same retention volume of the wild-type counterpart, indicating that the double-mutant is present uniquely as a dimer in solution.

Four different peptidic substrates were synthesized by solid-phase peptide synthesis by our colleagues of the Biondi research group at the Department of Chemistry. The 11-mer substrates mimicked the NSP4/5, NSP5/6, NSP7/8 and NSP14/15 recognition sequences of Mpro (see Table 9 at page 36) and were capped at both extremities; acetylation at the *N*-terminus and amidation at the *C*-terminus, to mimic the peptide bonds and make the peptides appear more like native protein. Acetylation and amidation also help to stabilize peptides, minimizing aminopeptidase and carboxypeptidase degradation, respectively. Such peptides were employed both in biophysical assays such as isothermal titration calorimetry (ITC) to determine their affinity to the enzyme, and in crystallization experiments to obtain the structure of the protease in complex with its substrates. Some experiments were hampered by the low solubility of some peptides, which varies widely depending on the sequence; NSP4/5 was the most soluble, followed by NSP5/6 and NSP14/15 which were reasonably soluble and finally NSP7/8 which was so insoluble that it could not be employed in any experiment.

The crystallographic structure of Mpro^{DM} in the free form was obtained under the same conditions of the wild-type variant. Subsequently, co-crystallization and soaking experiments were carried out with all available NSPs (including NSP7/8), and four different crystallographic structures of Mpro^{DM} in complex with its substrates were obtained: the Mpro^{DM}-NSP4/5 complex obtained by co-crystallization (1) and by soaking (2), and the co-crystallization complexes with NSP5/6 (3) and NSP14/15 (4).

Structure of apo Mpro^{DM}. Mpro^{DM} in its apo form crystallized in the same crystal packing as the wild-type variant (C2 space group, very similar unit cell), with one molecule per asymmetric unit (the dimer is crystallographic, as the other monomer is generated by the two-fold rotation symmetry axis). Data collection and refinement statistics are reported on Table 21 at page 97.

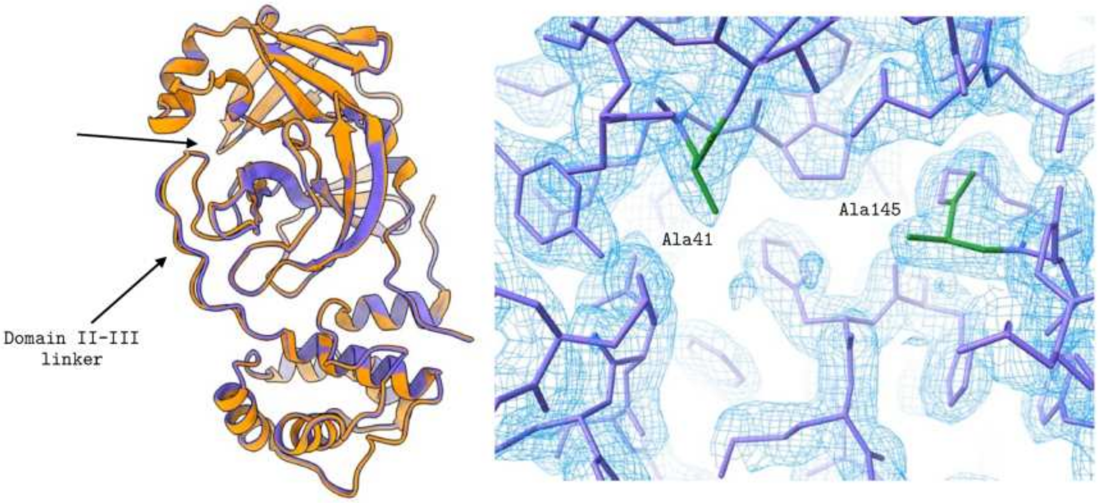


Figure 41. Crystallographic structure of Mpro^{DM}. On the left, superposition of the crystallographic structures of Mpro^{DM} (in purple) and Mpro^{WT} (canonical conformation, in orange). The two structures are very similar with the exception of the domain II-III linker which show a slight conformational change. On the right, 2F$_O$-F$_C$ electron density map (1σ level) of the active site of Mpro^{DM} with the mutated residues colored in green.

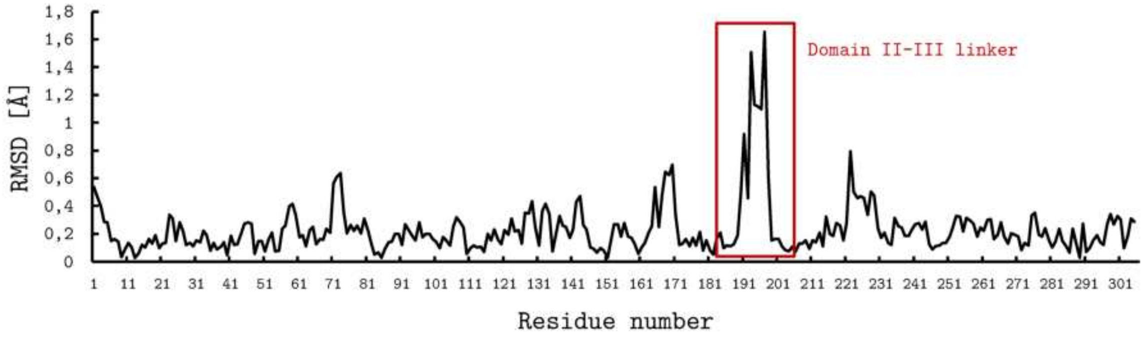


Figure 42. Pairwise alignment RMSDs of Mpro^{WT} vs Mpro^{DM}. The deviation between the double-mutant Mpro and its wild-type homolog is very small, with the exception of the domain II-III linker (residues 190-198) which is a relatively mobile region.

The resulting Mpro^{DM} crystallographic structure is superimposable to that of canonical Mpro^{WT} (overall RMSD between C$_\alpha$ 0.310 Å, Figure 41), with the partial exception of the domain II-III linker (residues 190-198, RMSD between C$_\alpha$ 0.4-1.7 Å, Figure 42). Most importantly, the active site region is not altered by the double mutation, and the catalytic water is still present in its position even in the absence of His41, as it is kept in place by Asp187 and His164.

Structure of Mpro^{DM} in complex with NSP4/5. The complex obtained by co-crystallization of Mpro^{DM} with the peptide NSP4/5 crystallizes in a different space group, the triclinic P1, with two molecules (that is, one dimer) in the asymmetric unit (data collection and model refinement statistics are reported at page 97). Both active sites are occupied by the peptide, which binds in similar ways but with small but significant differences, especially at the extremities (Figure 43). The oxyanion loop adopts the classical conformation, as in all Mpro^{DM}-peptide complexes later described.

Gln-P1 of the peptide is the strongest interactor, with 6 different interactions (Figure 44): the amide group of the side chain interacts with the N$_\varepsilon$2 of His163, the CO of Phe140, the carboxylate of Glu166 and with Asn142 side chain via a water-mediated interaction; the backbone CO of Gln-P1 interacts with the NH of Gly143, Ser144 and Ala145 of the oxyanion loop, while the backbone NH interacts with the CO of His164. The hydrophobic side chain of Leu-P2 fills the hydrophobic S2 subsite (formed by Met49, Met165, Pro52 and Ala41), while its backbone NH is anchored to the side chain of Gln189. The side chain of Val-P3 points towards the solvent but the backbone is anchored through the NH and the CO to the CO and the NH of Glu166, respectively. Ala-P4 is inserted in the partially hydrophobic S4 subsite (formed by the side chains of Gln192, Met165 and Leu167 and the CO of Arg188) with the NH bound to the CO of Thr190. Ser-P5 and Thr-P6 are exposed and weakly interact with the protein, with the two carbonyls making water-mediated interactions with backbone 166-168 and 192, respectively.

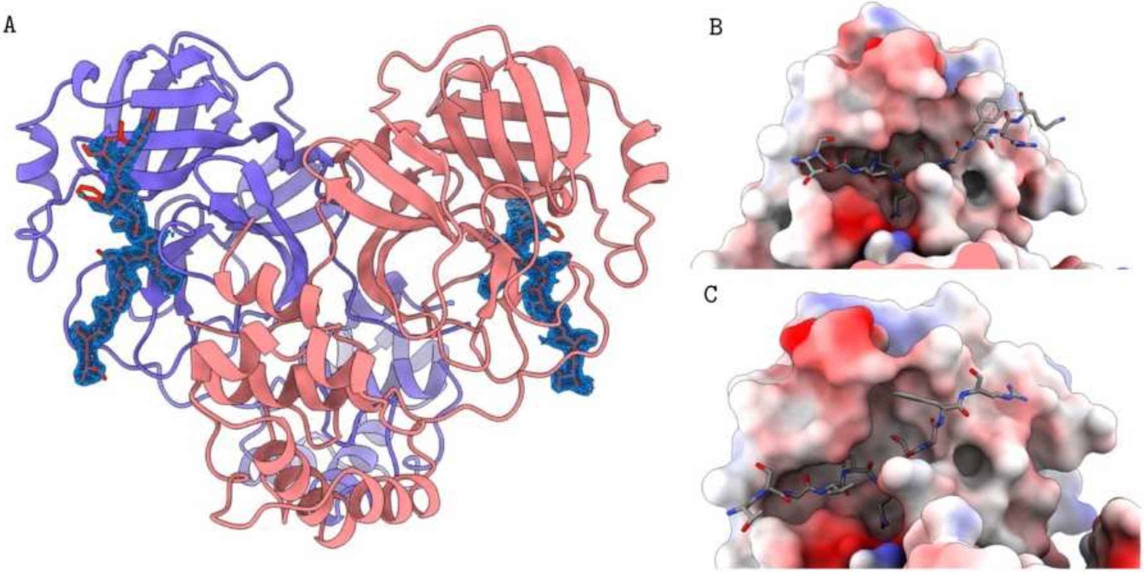


Figure 43. Structure of Mpro^{DM} in complex with NSP4/5 (co-crystallization). Panel A: overview of the dimer, with the electron density map of the peptides highlighted at the 1σ level. Panels B and C: poses of the peptide on the two subunits (A and B, respectively) of Mpro^{DM} (calculated electrostatic surface).

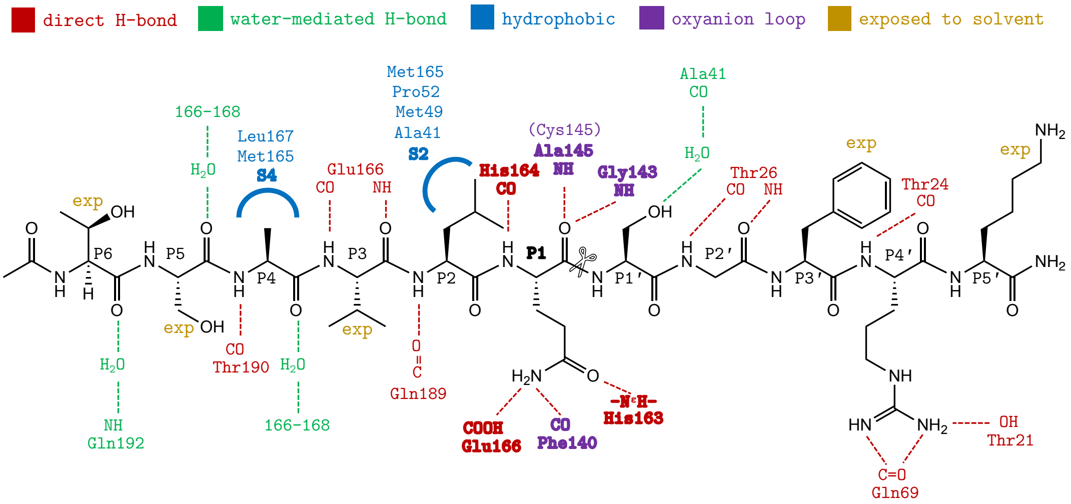


Figure 44. Interactions of peptide NSP4/5 with Mpro^{DM}. The ✂ symbol indicates the cleavable bond. A dense network of interactions contributes to the strong binding of the substrate to the enzyme, with polar interaction being the main driver.

At the *C*-terminal end of the interacting peptide, the Ser-P1' side chain interacts with the CO of Ala41 *via* a water molecule. The NH and CO groups of Gly-P2' make two hydrogen bonds with the backbone of Thr26. Phe-P3’ and Lys-P5’ are solvent-exposed and interact poorly with the protein. Arg-P4' interacts strongly, forming a hydrogen bond between the backbone NH and the carbonyl group of Thr24 and, more importantly, *via* three hydrogen bonds that connect its guanidinium group to the side chains of Gln69 and Thr21.

The interactions of the NSP4/5 peptide in subunit B are essentially the same. Notable differences are present for Thr-P6, which is slightly shifted, and for Lys-P5’, which is not visible in the electron density.

The P-value computed by the PISA analysis on the interaction between NSP4/5 and Mpro^{DM} is 0.46, meaning that the binding is mainly led by polar interactions, in accordance to the aforementioned description.

Even though the binding mode of the two subunits with the peptidic substrate is very similar, the Mpro dimer is not symmetric. The main differences are found in the different perturbation pattern of Domain I (which has higher B-factors in chain A – 59.0 – than in chain B – 54.2) and in the *C*-terminal tail (starting from residue number 300). Since these differences occur in regions of the enzyme that do not face the binding site of Mpro, it can be conclude that the asymmetry of the dimer is likely induced by the ligand binding.

The pairwise alignment RMSDs measured between the Mpro^{DM} complex and the *apo* wild-type homolog are a useful measure to assess the ligand-induced conformational variations of the enzyme. For the complex of Mpro^{DM} with NSP4/5, several regions show a distinct perturbation effect: beta-turn 22-26 (due to interactions with positions P2’ and

P4' of the peptide), loop 166-172 (interactions with P-positions of the peptide) and loop 44-50 (due to the insertion of Leu-P2 in sub-pocket S2). The domain II-III linker (residues 187-199) is modified as well, but a similar movement was also present for the *apo* Mpro^{DM} enzyme and it is not possible to clearly distinguish the effect induced by the peptide in this region. Interestingly, the oxyanion loop (residues 139-144) does not show relevant perturbations despite the numerous interactions with the peptide.

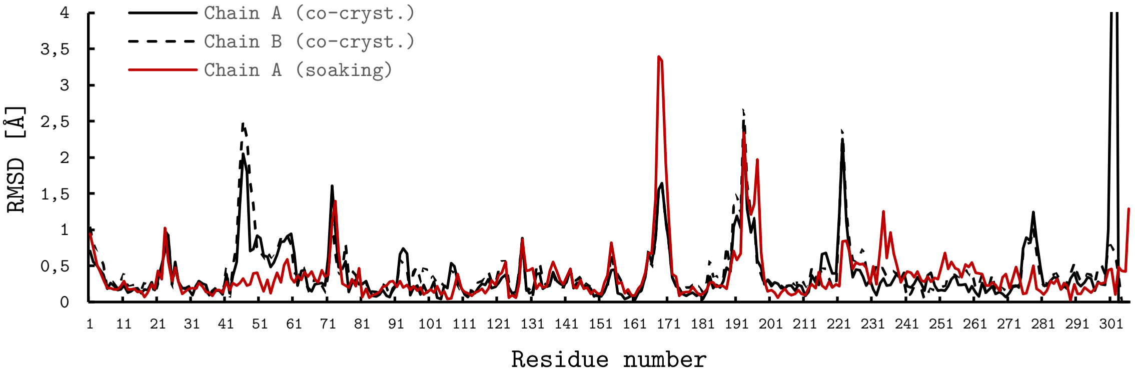


Figure 45. Alignment RMSDs of Mpro^{DM} (NSP4/5 complexes) vs *apo* Mpro^{WT}. In black, the pairwise RMSDs of the Mpro^{DM}-NSP4/5 complex obtained by co-crystallization (two chains are present as the ASU contains the M$_{pro}$ dimer). In red, the pairwise RMSDs of the M$_{pro}$^{DM}-NSP4/5 complex obtained by soaking (only one chain is present as the dimer is crystallographic). Notice how the perturbation pattern is different between the two complexes, likely because in the soaking experiment the enzyme lacks several conformational degrees of freedom.

The complex of Mpro^{DM} with NSP4/5 was also obtained by soaking, starting from the *apo* enzyme crystallized in the C2 space group. In this structure the dimer is symmetric and crystallographic, as the ASU contains a single unit of the enzyme (Figure 46). The conformation of both the enzyme and the peptide are different from the complex obtained by co-crystallization. Regarding the protein, the perturbation pattern induced by NSP4/5 shows significant variations, especially in loop 44-50 and at the *C*-terminal tail (now completely resolved) which are now very similar to the wild-type apo enzyme (Figure 45, red trace). Additionally, while perturbations in regions 22-25, 166-172 and 187-199 are similar in magnitude, they differ in geometry compared to the co-crystallization complex. The peptide network of hydrogen bonds with the enzyme is mostly conserved, but only the central amino acids Gln-P1 and Ser-P1' are superimposable to the co-crystallization complex (Figure 46, panel C). Pronounced differentiation is visible at the extremities of the peptide, especially in the P' positions, with the Arg-P4' in a completely different position, with the side chain anchored to the CO of Gly23 instead of the side chain of Thr24.

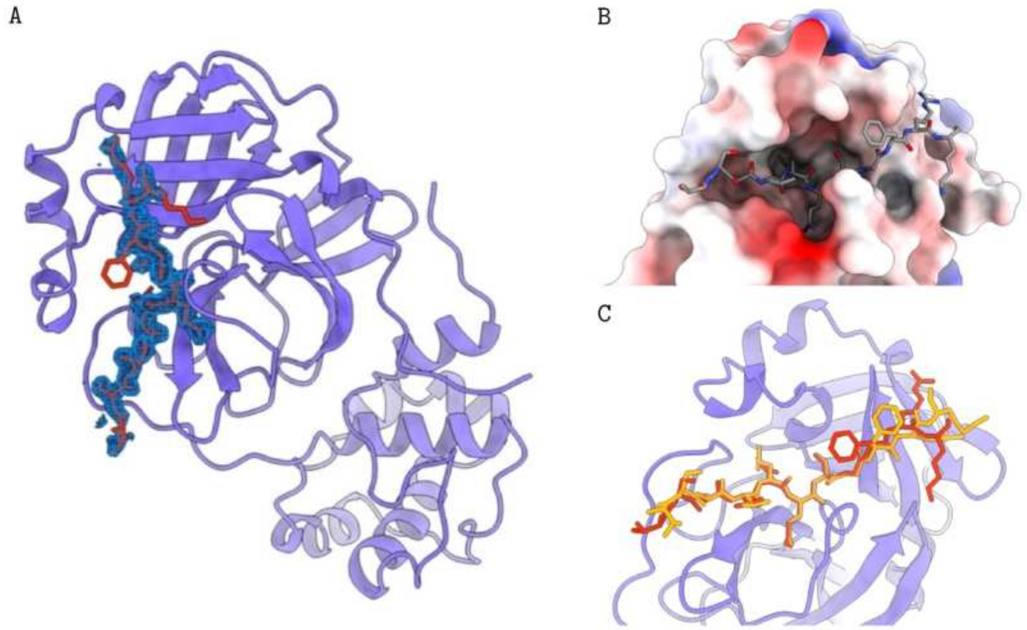


Figure 46. Structure of Mpro^{DM} in complex with NSP4/5 (soaking). Panel A: overview of the structure (only one subunit is present as the dimer is crystallographic), with the electron density map of the peptide highlighted at the 1σ level. Panel B: pose of the peptide on Mpro^{DM} (calculated electrostatic surface). Panel C: comparison of the NSP4/5 arrangements obtained in the soaking complex (in red) and in the co-crystallization complex (in orange); the two models of the peptide differ increasingly moving from the central part towards the extremities.

The widespread differences between the Mpro^{DM}-NSP4/5 complexes obtained by co-crystallization and by soaking are likely caused by the conformational constraints (dictated by the crystal packing) to which the latter must submit. Although the binding of the peptide does not trigger very pronounced rearrangements, it does require the concerted movement of several regions of the protein, some of which are not in close proximity to the binding site. Consequently, Mpro complexes obtained by soaking are likely not reliable for structural analysis and binding predictions.

Structure of Mpro^{DM} in complex with NSP5/6. The complex, obtained by co-crystallization, was found in the orthorhombic $P2_12_12_1$ space group (data collection and modelling statistics at page 97). Two molecules (that is, one dimer) were found in the asymmetric unit, as in the co-crystallization of Mpro with the other NSPs. Differently from the complex with NSP4/5, this time the active site of only one subunit is occupied, while in the other one just a small portion of the substrate is visible (Figure 47).
NSP4/5 and NSP5/6 differ significantly in the interaction network established with the enzyme, with increasingly higher deviations moving from the central section (positions P1 and P1' are essentially identical) towards the extremities (where there is a very low homology both in the peptide sequence and in its interactions). As an example, in position

P2 the hydrophobic leucine of NSP4/5 is substituted in a similar way by the phenylalanine in NSP5/6 filling the S2 subsite (Figure 48).

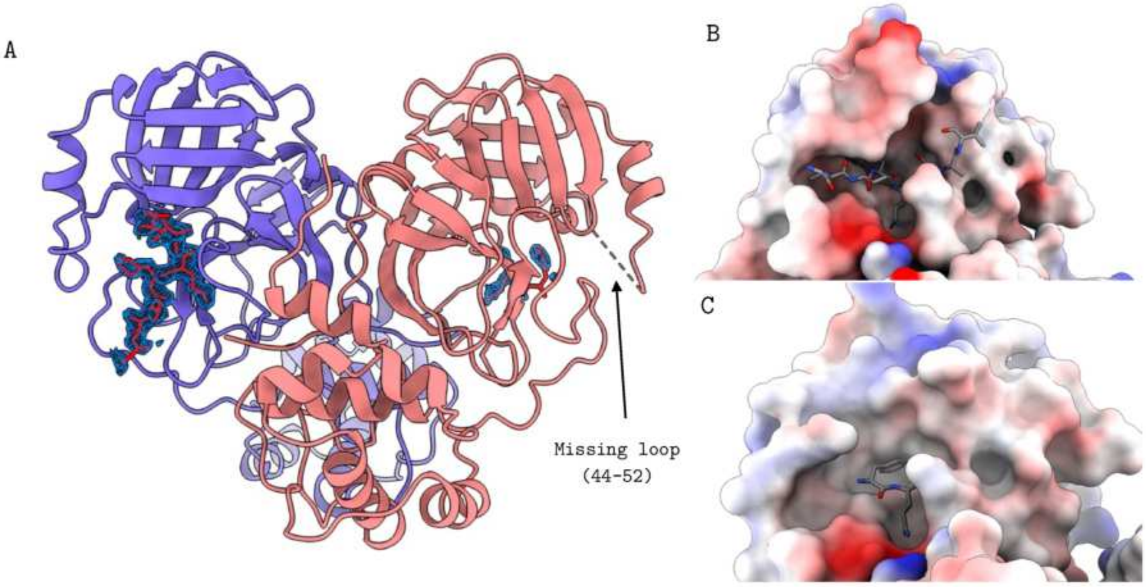


Figure 47. Structure of Mpro^{DM} in complex with NSP5/6. Panel A: overview of the dimer, with the electron density map of the peptides highlighted at the 1σ level. In subunit A (in purple) NSP5/6 is visible from residue P5 to P3', while in subunit B (in pink) only residues P2 and P1 are slightly visible in the electron density. Panels B and C: poses of the peptide on the two subunits (A and B, respectively) of Mpro^{DM} (calculated electrostatic surface).

In position P4, instead, a bulkier valine (in NSP5/6) replaces an alanine (of NSP4/5) and causes modification of the backbone with the loss of and H-bond, which translates in a high mobility for Ser-P6, which is not visible in the electron density. A similar effect is seen at the P' extremity (residues P3' to P5'), with only P3' having a sufficiently clear electron density to be included in the model. Residue Arg-P4' of NSP4/5 (which was locked in place by three H-bonds on the side chain) is now replaced by a flexible lysine. The PISA P-value for the peptide-enzyme interface is 0.45, indicating that the reduction in the strength of polar forces was not balanced by an increase in hydrophobic interactions. The loss of interactions at the terminal regions of the peptide is likely the main driving force that leads to their mobility.

In subunit B of the complex the active site is almost empty, with some faint electron density signal that allowed the tentative modelling of residues Phe-P2 and Asn-P1.

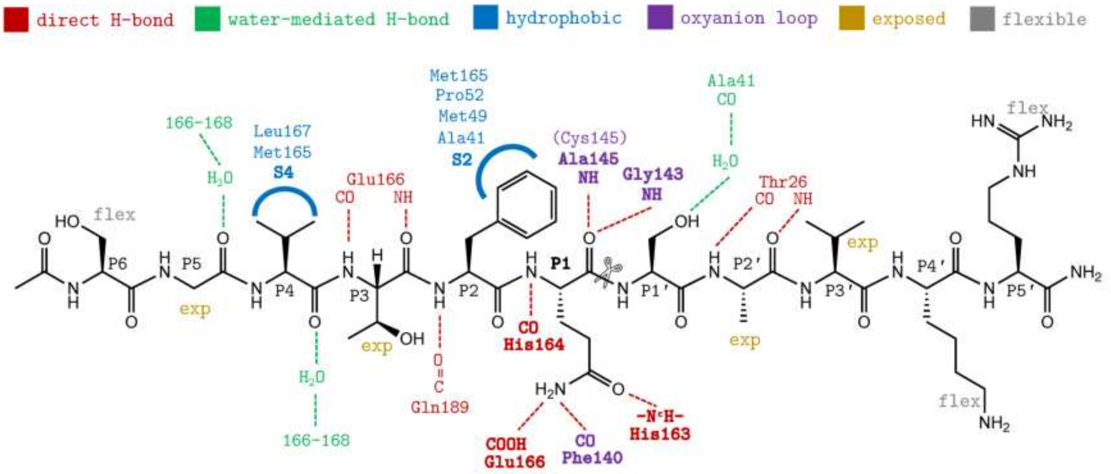


Figure 48. Interactions of peptide NSP5/6 with Mpro^{DM}. The ✂ symbol indicates the cleavable bond. Compared to NSP4/5, two (out of 10) direct hydrogen bonds are lost in the backbone at the extremities, as well as the three hydrogen bonds which locked Arg-P4' in place, now replaced by a mobile lysine residue. The loss of interactions at the terminal regions of the peptide is likely the main driving force that leads to their mobility.

Similarly to the MproDM-NSP4/5 complex, also in this structure the substrate-induced protein perturbations are spread throughout the dimer structure and not limited to the binding regions. Moreover, different scenarios emerge for the two dimer subunits: loop 44-52, for example, is well defined but very shifted in subunit A compared to *apo* Mpro^{WT} (RMSD ~2Å), while in subunit B the loop is not visible in the electron density (Figure 47, Figure 49). A completely different arrangement is also visible for the C-terminal tail of the two subunits. Additionally, the average perturbation over the complete length of the protein chain is different, with an RMSD of 0.713 and 1.152 Å for subunits A and B, respectively. In other words, while the binding event is almost completely limited to the subunit A of the complex, subunit B is the most perturbed one, indicating that the former is able to induce a conformational change in the latter.

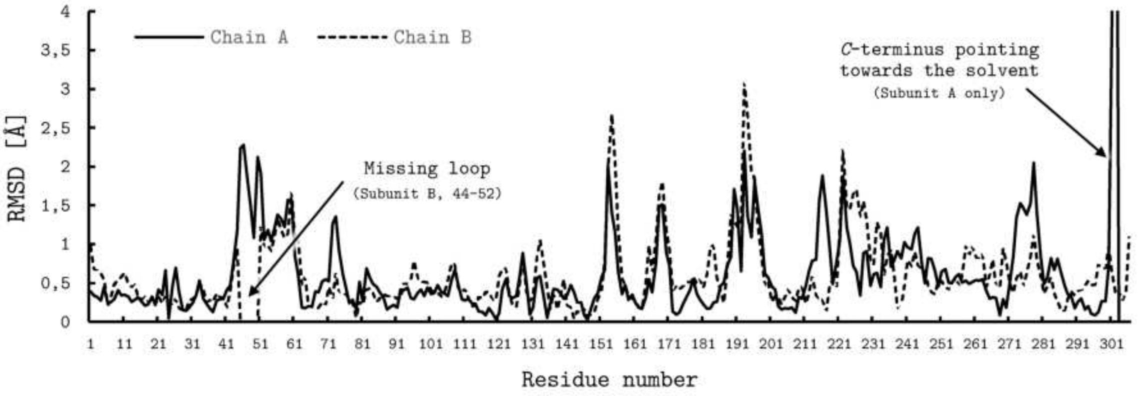


Figure 49. Alignment RMSDs of Mpro^{DM} (NSP5/6 complex) vs *apo* Mpro^{WT}.

Structure of Mpro^{DM} in complex with NSP14/15. The complex of MproDM with NSP14/15 was obtained by co-crystallization and resulted in a $P2_12_12_1$ space group crystal, similar to the complex of NSP5/6 (data collection and modelling statistics at page 97). In this structure only subunit A is occupied by the peptide, while no trace of a binding event is visible in the active site (or elsewhere) of subunit B (Figure 50).

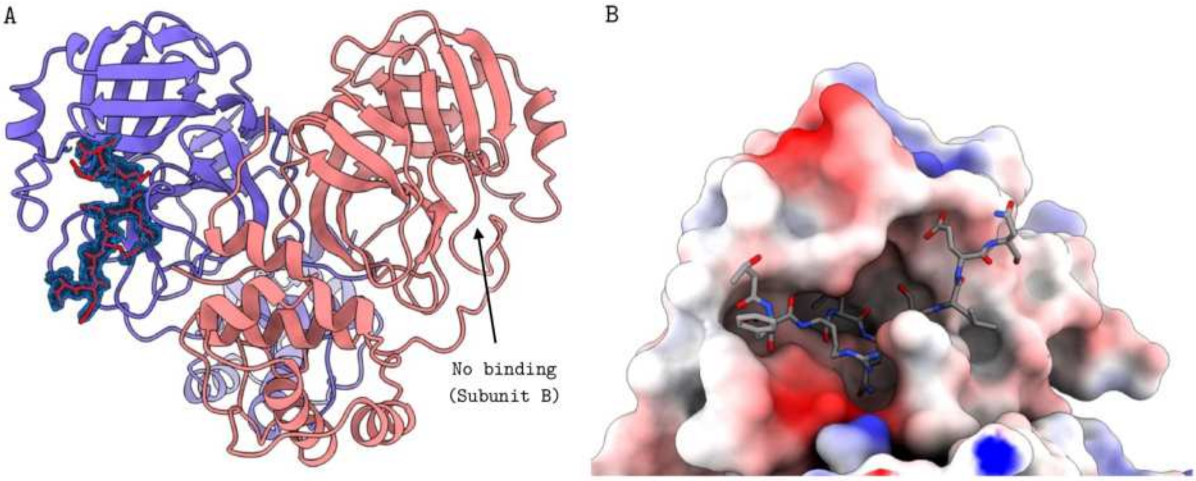


Figure 50. Structure of Mpro^{DM} in complex with NSP14/15. Panel A: overview of the dimer, with the electron density map of the peptide highlighted at the 1σ level. The substrate (in red) is present only in subunit A (colored in purple). Panel B: pose of the peptide in subunit A of Mpro^{DM} (calculated electrostatic surface).

The binding mode is analogous to that of NSP5/6, with the backbone in a similar pose from position P3 to P3'. Compared to NSP4/5, the best superposed residues are the centermost ones (Gln-P1 and Ser-P1'), which integrally conserve the interactions with the protein (Figure 51). Leu-P2 fits the S2 hydrophobic subsite as in NSP4/5. Arginine in position P3 (instead of a valine and threonine in NSP4/5 and NSP5/6, respectively) is solvent-exposed and mobile, but the backbone hydrogen bonds are conserved. Position P4 is now occupied by a threonine (instead of an alanine and valine in NSP4/5 and NSP5/6, respectively) which pushes away the peptide backbone from the S4 subsite, causing the loss of the H-bond involving the backbone NH. Interestingly, a new hydrogen bond is now formed between the Thr-P4 side chain hydroxyl and the NH of Gln192, which gets closer to the substrate peptide. Position P5 (phenylalanine) and P6 (threonine) are clearly visible in the electron density, because of a stabilizing interaction between the side chain hydroxyl of the Thr-P6 and the CO of Gln 189. Among the three NSP substrates studied, this is the only one with a clear interaction between position P6 and the enzyme. Compared to NPS4/5, Thr-P4 pushes apart the peptide from the protein, as in NSP5/6, but, unlike in the last, Thr-P6 (instead of a serine) is able to fold back the chain due to the side chain interaction. This turn is stabilized by four hydrogen bonds internal to the peptide, involving positions P4, P5 and P6.

Leu-P2' (instead of a glycine and an alanine in NSP4/5 and NSP5/6, respectively) is exposed to solvent. The side chain of the following glutamate in P3' is anchored to the carbonyl of Thr24 and the side chain hydroxyl of Thr25. The following Asn-P4’ is well anchored to Thr24, with hydrogen bonds involving both the backbone and the side chain of the two residues. Therefore, it seems that position P4’ is crucial for the successful anchoring of the substrate to the enzyme; the anchoring is possible for NSP4/5 (*via* an arginine) and for NSP14/15 (*via* a glutamine) but not for NSP5/6, which present a flexible lysine in such position.

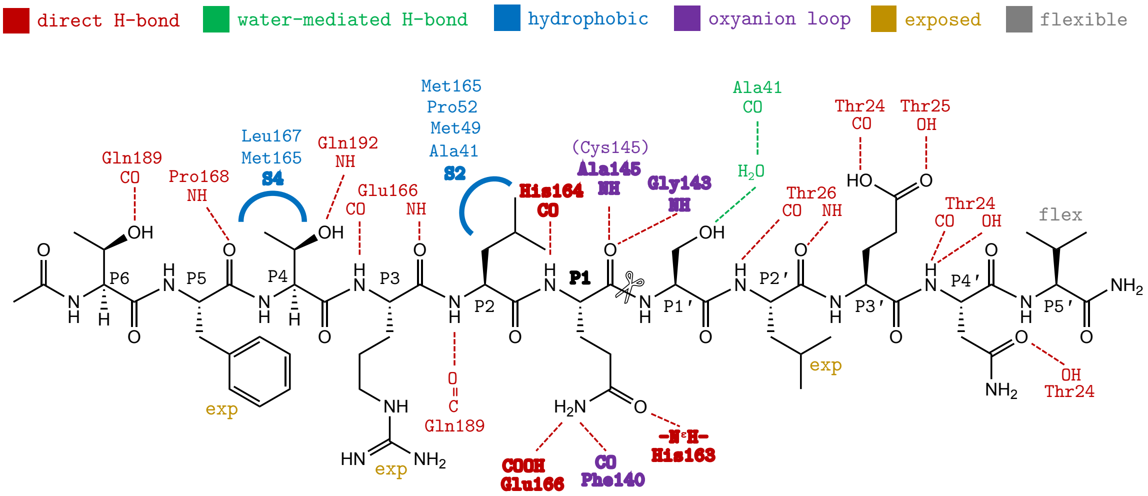


Figure 51. Interactions of peptide NSP14/15 with Mpro^{DM}. The ✂ symbol indicates the cleavable bond.

The mode of interaction of this region in NSP4/5 and NSP14/15 is different, due to the different properties of the residues involved (arginine and glutamine) but also due to the nature of the residue at position P3', a mobile phenylalanine in NSP4/5 and a fixed glutamate in NSP14/15. The PISA P-value of 0.42 is similar to those of the other substrates.

The perturbation signature for the two subunits follows a similar pattern to that of NSP5/6 (Figure 52). Residues 44-64 of both subunits, with loop 44-52 in particular, confirm to have the highest deviation compared to *apo* Mpro^{WT}. A different behavior for the *C*-terminal tail is once again visible for this complex. Also in this case, the average overall perturbation of the enzyme is higher in subunit B (which does not show any evidence of substrate binding) than in subunit A (in which the binding process is very clear).

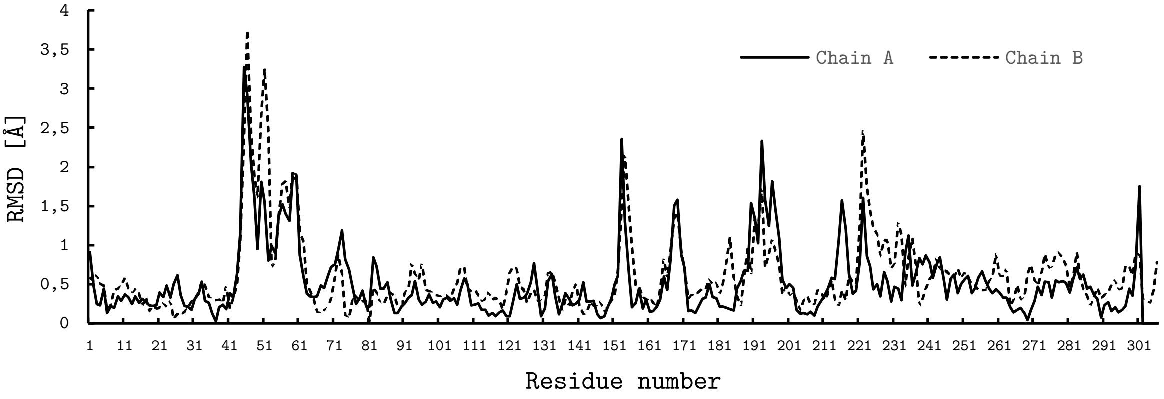


Figure 52. Alignment RMSDs of Mpro^{DM} (NSP14/15 complex) vs *apo* Mpro^{WT}.

The results discussed so far on the Mpro^{DM}-NSP complexes are in line with the literature-reported structures[29,38,39]. A brief overview of published structures is presented on Table 11. Considering both our structures and the ones reported in Table 11, we can conclude that the full perturbation effect induced by the peptide binding is visible with peptides of at least 11 amino acids in length (positions P6 to P5') and, most importantly, in co-crystallization experiments only.

Table 11. Mpro-NSP structures reported in the literature. Only structures with NSP4/5, NSP5/6 and NSP14/15 were included.

PDB-ID	Substrate	Method	Notes
7T70	NSP4/5	Co-crystallization	A longer, 12-mer peptide was used (additional residue at the *C*-terminus), the pose of the substrate is identical. Asymmetrical dimer.
7N89	NSP4/5	Co-crystallization	A shorter, 8-mer peptide was used (Ser-P5 to Phe-P3'). Arg-P4' extensive interactions are not present as such residue was omitted.
7MGS	NSP4/5	Co-crystallization	A shorter, 9-mer peptide was used (Ala-P4 to Lys-P5'). Substrate is visible only up to Phe-3'. Mpro dimer is crystallographic (one subunit per ASU) meaning that the two subunits are identical.
7DVP	NSP4/5	Soaking	A longer, 20-mer peptide was used with the soaking method but is visible only up to Arg-P4'. Mpro dimer is crystallographic (one subunit per ASU), in line with our complex obtained with the soaking method.
7T8M	NSP5/6	Co-crystallization	A longer, 12-mer peptide was used (additional residue at the *C*-terminus). The pose of the substrate is identical, but the peptide is present in both subunits. This effect can be

			explained with the increased solubility of the substrate used which - unlike ours - was not capped with protection groups at the extremities. The dimer is still asymmetric, and the same perturbations are visible when compared to our structure.
7DVW	NSP5/6	Soaking	A longer, 20-mer peptide was used with the soaking method. Visible from Ser-P6 to Val P3', as in our model. Mpro dimer is crystallographic (one subunit per ASU), with two identical subunits.
7DW0	NSP14/15	Soaking	A longer, 20-mer peptide was used with the soaking method. Mpro dimer is crystallographic (one subunit per ASU), with two identical subunits.

The complete characterization of the structures of Mpro^{DM} in complex with NSP4/5, NSP5/6 and NSP14/15 peptides is described in an article in preparation.

Conclusions

A variety of Mpro constructs were studied with different techniques, with a strong emphasis on their structural characterization. Besides the lessons learned on the single constructs discussed so far, a few important, more general considerations can be taken regarding all of them. The first conclusion is that the oxyanion loop of the active site is of crucial importance for the enzyme's correct functioning: it has an intrinsic instability (as confirmed by several structures in which its mobility prevents its electronic density to be seen), but at the same time it interacts with the most conserved residues of the substrates cleaved by Mpro (especially Gln-P1). The *new* conformation, in which we found Mpro^{WT} in presence of selected small molecules, proved to be evanescent and not easily druggable. Nonetheless, we were successful in faithfully reproducing its structure in an irreversible way thanks to the Mpro^{F140P} mutant; such mutant constitutes a convenient and powerful platform for future studies regarding the oxyanion loop.

The second important lesson was that relatively localized perturbations tend to spread out and influence distant parts of the protein, sometimes in unpredictable ways. While this conclusion is certainly valid for many proteins, its extent has likely not been fully understood for Mpro, for which - for example - no proof of allosteric modulation is known to date. The protein-wide domino-effect of Mpro perturbations is clear with the analysis of Mpro^{Cdel} mutant, where the absence of the dimerization-promoting Domain III forced the protein to: (1) remain in solution as a monomer, (2) to lose its catalytic activity, and also (3) to adopt a novel conformation of the oxyanion loop (neither *new* nor canonical), even though it is not in proximity of Domain III or crystal contacts. Another example is the perturbation of the Mpro$^{\mathrm{DM}}$ structure induced by the binding of peptide substrates: for such structures it is evident how the binding event of the peptide in the active site of a given subunit of the Mpro dimer can spread out to distant regions of the same chain but also (sometimes predominantly) to the other subunit.

Part 3

Threonine Aldolase 1 (Tha1)

Introduction

Scientific development in recent years provided a large amount of information regarding gene identification (genomics) and protein function (proteomics), yet the structural information associated with several proteins did not keep up[40,41]. A familiar example is the Aryl hydrocarbon receptor (AhR, see first chapter), whose function, expression regulation and interaction network were known much earlier than the first experimental structural evidence. One of the consequences of the mismatch between functional and structural information is that some metabolic pathways present “holes” for which the protein or the gene responsible for its fulfillment is not known. One of such “metabolic holes” is the second step of the carnitine synthesis, namely the conversion of 3-hydroxy-N^{ε}-trimethyl-L-lysine to 4-*N*-trymethylaminobutyraldeide by an enzyme with 3-hydroxy-N^{ε}-trimethyl-L-lysine aldolase (HTMLA) activity[42] (Figure 53).

In some mammalians (such as rodents) HTMLA activity is thought to be carried out by the enzyme Threonine Aldolase 1 (Tha-1), while in other species (including humans) the gene associated with such enzyme is silent[43]. The activity of HTMLA in humans is likely replaced by serine hydroxymethyltransferase (SHMT1 and SHMT2, cytosolic and mitochondrial version, respectively). In both cases the HTMLA activity is protracted by enzymes known for other catalytic activities.

To assess the identity of mouse and human HTMLA, our colleagues from the University of Parma (M. Malatesta, *et al.*) developed a reverse docking algorithm which fitted the transition state of reaction II of the carnitine synthesis onto a set of enzyme structures. Since the reaction is known to be dependent on the presence of pyridoxal 5-phosphate (PLP), the enzyme candidate list is limited to proteins which can bind PLP. In parallel, several ranking methods were tested with validated enzyme-substrate pairs to elect the best method. The transition state of reaction II of the carnitine synthesis was docked to a list of computationally generated enzyme structures from Alphafold[18], and the poses obtained were ranked accordingly. Human SHMT1 and SHMT2 obtained a total of 116 and 102 favorable poses, respectively, while murine Tha-1 obtained only 51 favorable poses (

Table **12**).

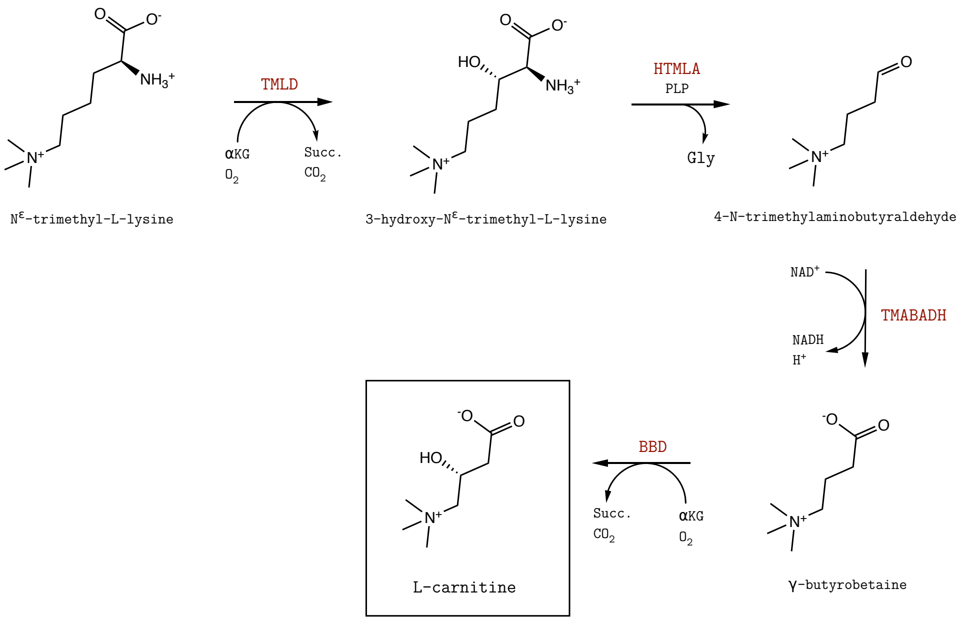


Figure 53. Overview of the carnitine synthetic pathway. N^{ε}-trimethyl-L-lysine (TML) is converted by N^{ε}-trimethyl-L-lysine dioxygenase (TMLD) to 3-hydroxy-N^{ε}-trimethyl-L-lysine (HTML). A PLP-dependent enzyme, named 3-hydroxy-N^{ε}-trimethyl-L-lysine aldolase (HTMLA) then converts HTML to 4-*N*-trymethylaminobutyraldeide (TMABA), which is then converted to γ-butyrobetaine by 4-*N*-trymethylamino-butyraldeide dehydrogenase (TMABADH). Finally, reaction of γ-butyrobetaine with γ-butyrobetaine dioxygenase (BBD) yields L-carnitine. The true identity of HTMLA in humans and other species is not known to date.

Table 12. Enzymatic activities and docking scores of HTMLA candidates. The docking score of the reverse docking score of Tha-1 increases dramatically with the new experimental structure obtained. The new score justifies the enzymatic activity measured experimentally.

Source	Enzyme	Docking score	k_{cat}/K_M ($M^{-1}s^{-1}$)
Human	SHMT1	116 (Alphafold structure)	32.4±4.3
Human	SHMT2	102 (Alphafold structure)	6.3±0.7
Murine	Tha1	51 (Alphafold structure)	13700±890
Murine	Tha1	91 (Experimental structure)	

The reliability of the Alphafold model of murine Tha-1 was questioned, as only one experimental structure of Threonine aldolase was present on the PDB, from a distant bacteria species (*Thermotoga maritima*). The availability of Tha-1 structures from a single organism severely limits the reliability of the model inferred on other organisms (mouse, in this case), especially if those organisms are evolutionarily distant.

Results and discussion

Murine Threonine aldolase was recombinantly produced in *E. coli* by the laboratory of Malatesta *et al.*, which provided us the concentrated and purified protein for crystallization experiments. Initial experiments of concentration optimization for crystallization (see experimental details at page 85) determined that a concentration of 2.5-3.5 mg/mL was ideal. Using both the Morpheus and PACT premier crystal screens (Molecular Dimensions), diffraction-quality crystals were obtained using the sitting-drop vapor diffusion method. Diamond-shaped crystals appeared in conditions G9 (0.2 M Potassium sodium tartrate tetrahydrate, 0.1 M Bis-Tris propane pH 7.5, 20% w/v PEG 3350), H5 (0.2 M sodium nitrate, 0.1 M Bis-Tris propane pH 8.5, 20% w/v PEG 3350) and H9 (0.2 M Potassium sodium tartrate tetrahydrate, 0.1 M Bis-Tris propane pH 8.5, 20% w/v PEG 3350) of the PACT Premiere HT-96 screen within 24 hours. The crystallization experiment was then reproduced with larger drops (1 µL protein + 1 µL of precipitant, versus 0.2+0.2 µL used in the screening experiment) but no significant increase in crystal size was observed (about 20-30 µm in diameter). The employment of the crystal microseeding technique did not improve the crystal size either, suggesting that the crystal growth may be hampered by some mobile regions of the protein construct, such as the histidine tag and the linker connecting it to Tha-1 (about 20 amino acids in total).

Tha1 crystallized in two space groups: orthorhombic F222 and monoclinic C2, with one molecule and two molecules in the ASU, respectively (data collection and refinement statistics on Table 22 at page 98). The PLP cofactor is visible only in the monoclinic structure; however, the active site is very similar in the two cases, with only minor differences. The expected tetrameric quaternary structure is formed by crystallographic symmetries, with four identical units in F222 (related by a 222 symmetry) and two identical dimers in C2 (related by a two-fold axis). The RMSD values between the single units (around 0.26-0.28 Å) indicate that the monoclinic and orthorhombic structures are very similar (the tetrameric assembly is also conserved in the two space groups). As indicated by data from PISA[44] analysis, the interface between units A and B (analogous to that between units C and D) is contributing stronger to the stability of the quaternary structure in comparison with the interface between units A and C (analogous to that between units B and D) (Figure 54, panels a, b). Hence, the tetramer can be considered a dimer (AB+CD) of dimers (A+B and C+D), with the first dissociation being ABCD to AB+CD. A comparison with the structure of the *Thermotoga maritima* threonine aldolase (PDB code 1M6S) returned RMSD values between 1.03 and 1.17 Å for the single units, indicating a significant structure difference even though the secondary structures and the whole quaternary assembly are conserved. The major structural difference is related to an insertion of 13 residues in Tha1 between positions 346-260, residues absent in the *T. maritima* threonine aldolase. In the two enzymes the position of the PLP cofactor is mostly conserved: PLP is covalently bound to Lys242 in murine Tha1 and to Lys199 in bacterial Tha1 in a very similar pose (Figure 54, panel c). The Alphafold model of murine Tha1 is also very similar to our experimental structure, with minor movements of the side chains (Figure 54, panel d). The computational model lacks the PLP cofactor.

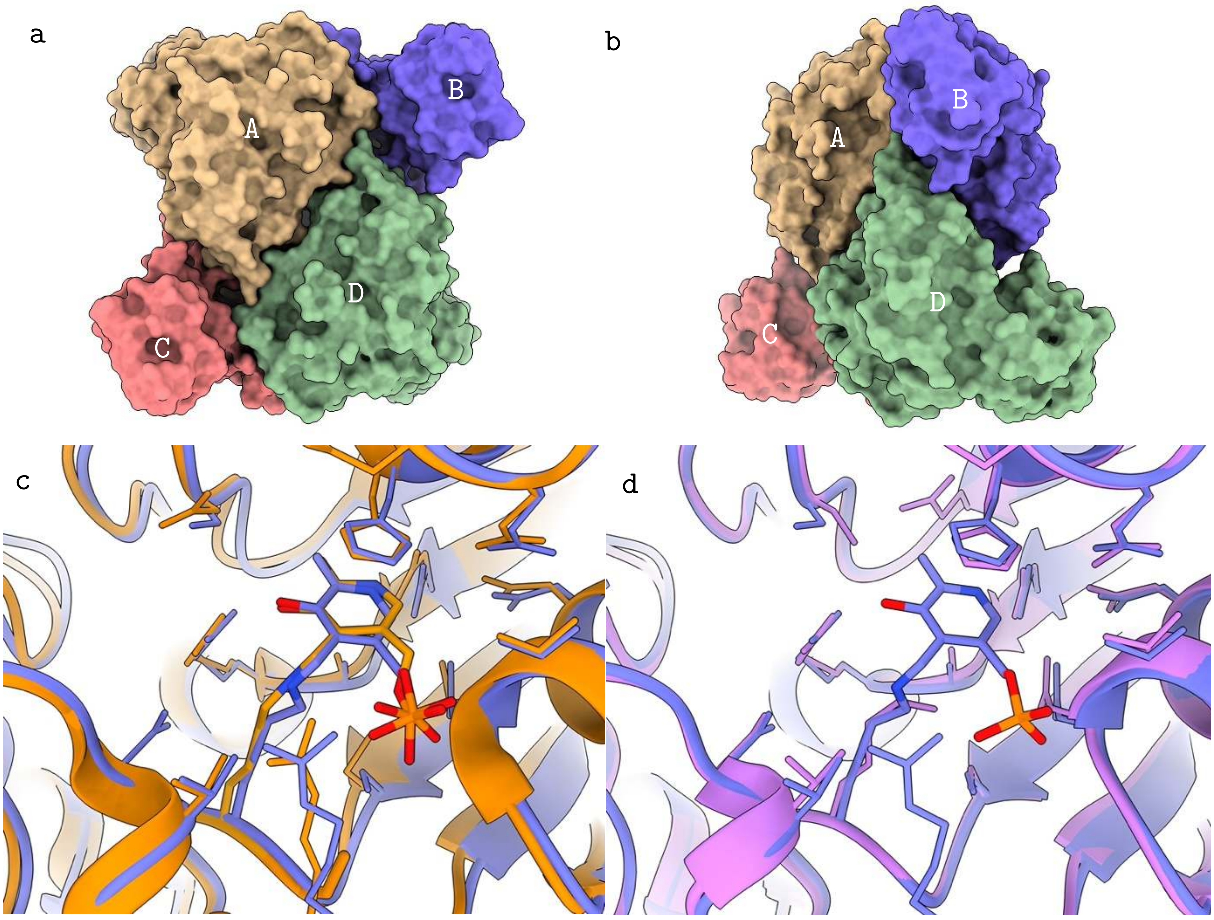


Figure 54. Experimental structure of murine threonine aldolase. In panels A and B the front view and side view of the Tha1 tetramer are depicted, respectively. Panel C: superposition of the active site of Tha1 from *T. maritima* (in orange) and mouse (in purple); PLP is covalently bound to Lys242 in murine Tha1 and to Lys199 in bacterial Tha1 in a very similar pose . Panel D: superposition of the active site of Tha1 of the Alphafold model (in pink) and mouse (in purple).

The docking screening was repeated by including in the data set the crystallographic structure of Tha1. The docking score of the catalytic cluster based on the new experimental structure jumped to 91, compared with 51 in the previous analysis (

Table **12**). By comparing the two structures, some differences are observed in the side chains of the substrate binding residues. There are minor differences in the chain containing the catalytic lysine (e.g. Arg372A), while differences in the position of the residues contributed by the other chains (Tyr168C, Tyr69B) are more pronounced, suggesting that they result mainly from subunit assembly.

Conclusions

We presented the first experimental structure of murine Threonine Aldolase at atomic resolution. We determined such structure both in presence and in absence of its co-factor PLP. The structure is overall similar to both the *T. maritima* homolog and the Alphafold model, but some small yet relevant differences are present. Such variations justify the measured HTML aldolase enzymatic activity and concurrently provide a validation for the Malatesta *et al.* reverse docking algorithm. The results hereby presented on murine Tha1 and further details on the characterization of enzymes exhibiting threonine HTMLA activity are fully described in an article under revision on Nature Communications.

Experimental methods

In following chapter, the most common experimental methods employed in the different projects are described. The overall procedure is the same, while some specific details may vary in a subject-specific manner.

Cloning methods

TA-cloning

The template of murine and human AhR and murine ARNT was generously provided by the Perugia laboratory (F. Fallarino) and were diluted 100-fold to a final concentration of 20 ng/µL for mARNT, 38 ng/µL for hAhR, 22 ng/µL for mAhR. Primers with code 1-8 (see Primers table on page 92) were provided by Invitrogen and a stock solution was produced at 10 µM concentration.

PCR sample composition: 5 µL of template DNA, 0.75 + 0.75 µL of the forward and reverse primers (10 µM), 1.5 µL of $MgCl_2$ 25 mM, 1 µL of 10 mM dNTPs mix, 2.5 µL of 10x PCR buffer, 0.25 µL of GRS TAQ polymerase 5 U/µL, 13.25 µL of water.

PCR cycling conditions: initial denaturation at 95 °C for 5 minutes followed by 30 cycles each comprising: 95 °C for 30 seconds (denaturation), 55 °C for 30 seconds (annealing) and 72 °C for 30 seconds (elongation). Finally, incomplete fragment completion was performed for 10 minutes at 72 °C.

Ligation. Sample composition: 3 µL of PCR-amplified sample, 2 µL of linearized pET-SUMO plasmid (0.025 µg/µL), 1 µL of ligation buffer 10x, 1 µL of T4 DNA ligase (5 U/µL), 3 µL of water. The samples were placed in the thermocycler at 15 °C overnight to allow the ligation reaction to proceed.

Transformation. For each sample, 7 µL of the ligated sample was added to a frozen aliquot of *E. coli* TOP10 competent cells. The samples were incubated for 30 minutes in ice, then heat shock was applied for 60 seconds at 42 °C, then again in ice for 30 seconds. Finally, 500 µL of SOC liquid culture were added and the sample was stirred for 1 h at 37 °C. The liquid cultures were plated on 1.5% LB agar Petri dishes containing 30 µg/mL of kanamycin and incubated overnight at 37 °C. Colonies appeared by the next day and were screened by colony PCR.

Restriction-free (RF) cloning

DNA primers for restriction-free cloning are composed of two parts: the first one is complementary to the DNA sequence that needs to be amplified (and is identical to the one used for the TA cloning). The second part is complementary to the destination vector and allows the DNA insert to anneal to the correct position of the destination plasmid.

PCR 1 (megaprimer synthesis): 2 µL + 2 µL of the forward and reverse primers (10 µM), 0.1-0.2 µL of template DNA (50-100 ng total), 1 µL of 10 mM dNTPs mix, 5 µL of 10x Thermopol PCR buffer, 0.5 µL of Vent DNA polymerase 2 U/µL, water to 50 µL.

PCR 1 cycling conditions: initial denaturation at 95 °C for 2 minutes. Then 30 cycles each comprising: 95 °C for 30 seconds (denaturation), 50-60 °C (depending on primers) for 30 seconds (annealing) and 72 °C for 90 seconds (elongation). Finally, incomplete fragment completion was performed for 5 minutes at 72 °C.

Agarose gel electrophoresis and megaprimer extraction. An agarose gel with a concentration of 0.8-1.5% is prepared based on the expected megaprimer size. Just prior to gel casting, the EuroSafe dye was added. At the end of the PCR reaction the loading buffer (comprising glycerol and bromothymol blue) was added to each PCR tube. Each sample and a suitable DNA marker (ladder) are then loaded to the agarose gel, which is then run at constant voltage and visualized under UV light. The bands containing the megaprimers are then excised and the DNA is extracted with a suitable agarose gel extraction kit. Typical yields are in the range of 500-1500 ng of purified megaprimer per 50 µL of PCR 1 reaction volume.

PCR 2 (megaprimer insertion): 100-300 ng of megaprimer (this parameter needs to be optimized for every construct), 0.1-0.2 µL of template DNA (50 ng total), 0.5 µL of 10 mM dNTPs mix, 5 µL of 5x Phusion HF PCR buffer, 0.5 µL of Phusion DNA polymerase 2 U/µL, water to 25 µL.

PCR 2 cycling conditions: initial denaturation at 95 °C for 1 minute. Then 30 cycles each comprising: 95 °C for 20 seconds (denaturation), 50-60 °C (depending on primers) for 45 seconds (annealing) and 72 °C for 5 minutes (elongation). Finally, incomplete fragment completion was performed for 10 minutes at 72 °C.

DpnI digestion and transformation: 1 µL of DpnI enzyme (20 U/µL) are added to 10 µL of the reaction solution of PCR 2. The cleavage of methylated DNA is allowed to proceed for 37 °C for 1 hour. Of the obtained solution, 5 µL are then used to transform *E. coli* competent cells according to the transformation protocol (experimental details at p. 80).

Colony PCR

Colonies that appear on the selection plate at the end of the cloning procedure (either TA or RF cloning) are screened by colony PCR prior to sequencing. First, a PCR master mix solution is prepared for each construct to be tested. Such solution contains adequate primers that bind only to the DNA insert being cloned in the destination vector and not to the vector itself.

PCR sample composition: 0.5 + 0.5 µL of the forward and reverse primers (10 µM), 1 µL of 10 mM dNTPs mix, 2 µL of 10x Thermopol PCR buffer, 0.3 µL of Vent DNA polymerase 2 U/µL, water to 20 µL. Each sample is placed in a single PCR tube. Each of the colonies to be tested is inoculated with a sterile loop in the PCR tube containing the master mix. The initial denaturation step of the PCR cycle breaks the bacterial cells releasing the plasmids contained therein and allowing the PCR screening to take effect.

PCR cycling conditions: initial denaturation at 95 °C for 5 minutes. Then 20 cycles each comprising: 95 °C for 30 seconds (denaturation), 50-55 °C (depending on primers) for 30 seconds (annealing) and 72 °C for 30 seconds (elongation). Finally, incomplete fragment completion was performed for 10 minutes at 72 °C.

Agarose gel electrophoresis. An agarose gel with a concentration of 0.8-1.5% is prepared based on the expected PCR amplicon size. Just prior to gel casting, the EuroSafe DNA dye was added. At the end of the PCR reaction the loading buffer (comprising glycerol and bromothymol blue) was added to each PCR tube. Each sample and a suitable DNA marker (ladder) are then loaded to the agarose gel, which is then run at constant voltage and finally visualized under UV light (Figure 55). Since the amplicon size is known from the DNA cloning design, positive colonies must appear at the expected DNA size (a suitable positive control or a DNA ladder can be used for this purpose).

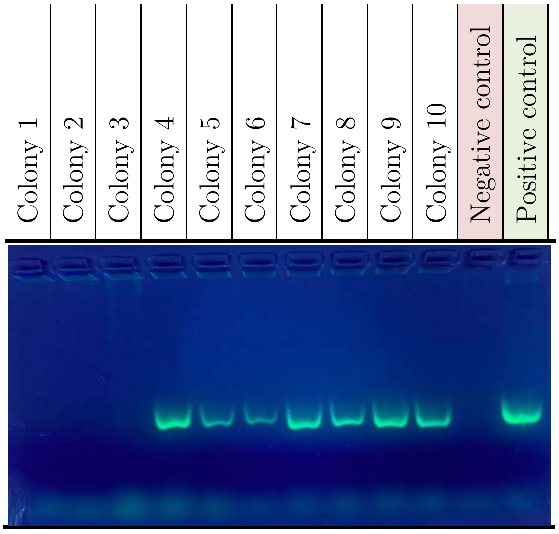


Figure 55. Example of agarose gel of a colony PCR. Colonies 4-10 appear positive at the colony PCR test. Colonies 4 and 7 were expanded and their plasmid DNA

sequenced. Both RF cloning procedures were successful, but colony 4 presented a non-silent point mutation caused by an error of the DNA polymerase, which can only be detected by DNA sequencing. Furthermore, in our experience, bands with a fainter fluorescence (lanes 5 and 6) are typically confirmed to be false positives by DNA sequencing.

Bacterial transformation

In our laboratory we normally use both commercially available and "home-made" *E. coli* competent cells. We produce the home-made cells with the calcium chloride medium: this method is fast and inexpensive, but the obtained competency is relatively low. This makes these cells ideal to be transformed with already-purified plasmids (typically with 50 ng of plasmid per cell aliquot). We use instead the TOP10 (Invitrogen) or E. cloni competent cells (Biosearch laboratories) for the transformation at the end of the TA- or RF-cloning. In such cases the competency of the cells must be very high as the amount of plasmid at the end of the cloning procedure is very low. The transformation process is the same for both types of cells.

Transformation process. The centrifuge tube containing the competent cells is taken from the -80 °C freezer and put in ice for 20 minutes. The purified plasmid (about 50 ng) or the cloning product from the DpnI digestion of T4 ligation (5 µL of reaction mix) is added to the competent cells and allowed to incubate in ice for 20 more minutes. The tube is then transferred for 60 seconds in a heat block at 42 °C (heat shock), then again in ice for 3 minutes. 800 µL of recovery medium are added to the transformed cells and the tube is put under agitation for 60 minutes at 37 °C. The tube is then centrifuged at 5000 *g* for 1 minute to precipitate bacteria. About ¾ of the supernatant is removed, while the precipitated bacteria are resuspended in the remaining medium (~200 µL total) to be plated on a suitable Petri selection plate. The Petri dish is then incubated at 37 °C overnight. If the transformation process is successful, bacterial colonies appear overnight (Figure 56) and are ready to be tested (e.g. colony PCR) or amplified (in a liquid broth).

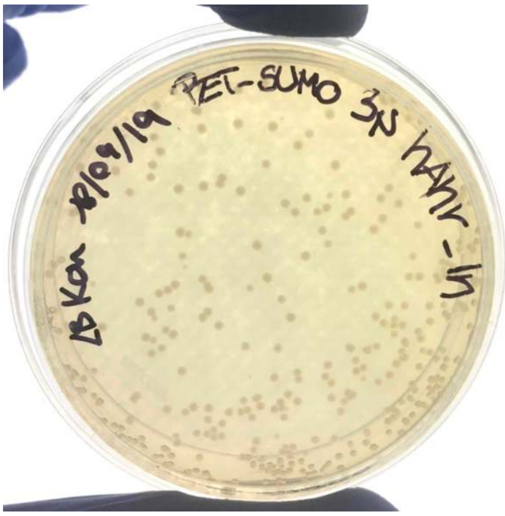


Figure 56. Bacterial colonies on a Petri dish. Colonies typically appear after an incubation period overnight at 37 °C.

Protein expression in *E. coli*

A bacterial glycerol stock is expanded in a preculture of 20 mL of LB with added antibiotic selection (kanamycin for pET-SUMO, ampicillin for pGEX, pGro7 and pET-21d, chloramphenicol for pDUET) and allowed to grow overnight at 37 °C under agitation. The following day, part of the preculture is diluted in fresh LB broth (typically 500 mL for expression tests and 1-2 L for larger scale production) at a dilution of 1:50 to 1:200 depending on the experiment. The expression culture is then incubated at 37 °C under agitation until the desired optical density (OD, which is directly proportional to bacterial concentration) is reached. When the desired OD is reached (typically between 0.6-1), Isopropyl β-D-1-thiogalactopyranoside (IPTG) is added to the bacterial culture and expression is allowed to proceed at the desired temperature (typically 20 °C – 37 °C) for a given period of time (typically 4 h for expression at 37 °C and overnight for expressions at 20 °C).

Protein extraction and purification

Bacterial lysis

Note: throughout all the steps described in this section, the sample is kept in ice. At the end of the protein expression, bacteria are pelleted from the culture by centrifugation at 5000 g for 20 minutes at 6 °C. The supernatant (spent broth) is discarded, and the bacteria are resuspended in the lysis buffer. The composition of the lysis buffer largely depends on the recombinant protein being extracted, while its amount is proportional to the quantity of bacteria being processed. The amount of lysis buffer used is normally 10 to 20 mL per liter of culture, depending on the final optical density of the culture. Since the first purification step for all proteins described in this thesis is the immobilized cation affinity chromatography (IMAC), the lysis buffer coincides with the binding buffer of the IMAC itself plus different optional additives. We normally add DNAseI (Sigma-Aldrich) (plus a small amount of $MgCl_2$ as a co-factor) to digest most of the DNA present in the bacterial lysate and simplify subsequent filtration steps. DNAse was essential in AhR purifications, because in its absence DNA associated with AhR was isolated in the IMAC binding step (AhR has an isoelectric point between 7.5 and 9.0 depending on the construct). Cellular lysis is performed using the French press (Thermo-Fisher scientific) technique with 2 cycles at ~10000 psi (~690 bar). The cellular lysate is then centrifuged at 18000 g for 30 minutes at 6 °C to separate the soluble (supernatant) and insoluble (pellet) fractions. Depending on the experimental design, the recombinant protein is either mainly located in the insoluble fraction (in the case of the refolding method) or in the supernatant. In the latter case, the supernatant is filtered at 0.45 µm and injected in the IMAC column.

Variation in case of refolding. In the case of the refolding method, the intracellular precipitation of the recombinant protein is favored inside the so-called inclusion bodies (IBs). The IBs precipitate together with the cellular debris in the pellet fraction after the lysis. The pellet is then subject to a series of washing steps (inclusion bodies purifications, see Figure 21 on page 30). The purified inclusion bodies are the resuspended in a buffer containing 6 M guanidinium chloride for a few hours or overnight. The resuspended protein can then be refolded by gradually decreasing the guanidinium concentration using dialysis tubes or by immobilizing the denatured protein on an IMAC column (in the case of a protein with the HisTag) and applying a gradient with decreasing concentration of guanidinium ions.

Purification process

IMAC-1. All chromatographic purifications were carried out with ÄKTA purifier chromatographers (GE healthcare). For most of the constructs described in this thesis, the first chromatographic step consisted in an immobilized metal affinity chromatography (IMAC), more specifically the HisTrap HP of HisTrap FF columns (Cytiva). The bacterial lysate is injected through the column. After sample injection, the column is washed for several minutes with a 98/2 ratio of binding and elution buffers (Buffer A and B, respectively). The mixture corresponds to an imidazole concentration in the eluent mixture of 10 mM, which is necessary to reduce the amount of non-specific binding to the column. The other eluent components can vary based on the construct being purified, but they all contain NaCl (in variable concentrations), a buffering compound (typically phosphate, TRIS, HEPES, etc.) and a reducing agent, if necessary, to reduce protein oxidation for delicate samples (e.g. DTT). After the column is washed, the recombinant protein is eluted by increasing the percentage of buffer B, either with a linear gradient or by concentration steps. The eluted protein is then collected and fractionated.

Desalting. Imidazole contained in the eluted protein of the IMAC-1 step must be removed to engage the protein in the following chromatographic steps. The process is basically a buffer exchange operated with an HiPrep Desalting column (GE healthcare). This column is a special size-exclusion chromatography (SEC) column optimized for buffer exchange, which separates all molecular entities above ~5 kDa (basically most proteins) from their buffers (whose components are smaller), replacing it with the buffer in which the column was equilibrated (i.e. a buffer not containing imidazole).

Fusion protein/affinity tag cleavage. The fusion protein (i.e. SUMO, for pET-SUMO constructs) or the affinity tag (the bare HisTag, for pET-21d, pDUET, pGEX) are removed from the recombinant protein by means of a suitable protease (Ulp-1 for SUMO, 3C-protease for pGEX, etc.). The protein and the protease are incubated overnight at 4 °C or 12 °C, depending on the construct.

IMAC-2. The aim of the IMAC-2 chromatography is to separate the cleaved recombinant protein from the undesired leftovers, namely: the uncleaved protein, the protease, the fusion protein/the affinity tag. An IMAC chromatography is ideal as all the leftovers can be captured by the IMAC column (all proteases used for this purpose contain a suitable affinity tag themselves), while the cleaved, recombinant protein does not bind to the column and is collected in the column flow-through.

SEC. As a last chromatographic step a size-exclusion chromatography (SEC) is carried out with HiLoad 16/600 Superdex 75 or Superdex 200 columns (Cytiva). This procedure fulfills several necessities at the same time: it separates the desired proteins from all its aggregates and from contaminants of different molecular weight, it gives an estimate of the molecular weight of the recombinant protein assembly (thus informing on its monomeric/multimeric nature) and it can be used for a final buffer exchange, if needed.

Final concentration. The protein is subject to a concentration step to meet the requirements of subsequent characterization techniques. For crystallization experiments, a concentration of ~8.0-12.0 mg/mL was reached for the various Mpro constructs and 3.5 mg/mL for Tha1. The concentration step is done with centrifugal concentrators with a suitable molecular weight cutoff (MWCO), which was at least twice as small as the protein molecular weight (i.e. a 10000 MWCO concentrator was used for proteins of at least 20000 Da in weight). The concentrated protein is then dispensed in 20 µL aliquots, flash-frozen in liquid nitrogen and stored at -80 °C.

Crystallization techniques

A frozen aliquot of the concentrated protein is thawed in ice and centrifuged for two minutes at 17000 *g* to remove insoluble precipitates. The protein concentration is then adjusted if needed, by using the same buffer in which the protein is dissolved.

Crystallization screening. The reservoirs of a MRC2 96-well two-drop crystallization plate (Swiss-CI) are filled with 40 µL of the desired crystallization screening (e.g. Morpheus HT-96, Index HT, PACT premiere, etc.) with an 8-channel micropipette. The MRC2 plates (Figure 57, on the left) are standard 8x12-wells microplates, ideal for high-throughput, sitting-drop vapor diffusion crystallization experiments. One or two protein solutions can be tested for each plate as two wells are available for each reservoir. If two crystallization drops are used, they can either be two different proteins, or the same protein at different concentrations. The crystallization drops are dispensed with and Oryx Nano robot (Douglas Instruments, Figure 57, on the right) with 0.2 µL of protein solution + 0.2 µL of precipitant solution, which is retrieved from the reservoir. At the end of the dispensing procedure, the plate is manually covered in a protective film and incubated at controlled temperature (+9 °C, +18 °C) to allow crystals to grow.

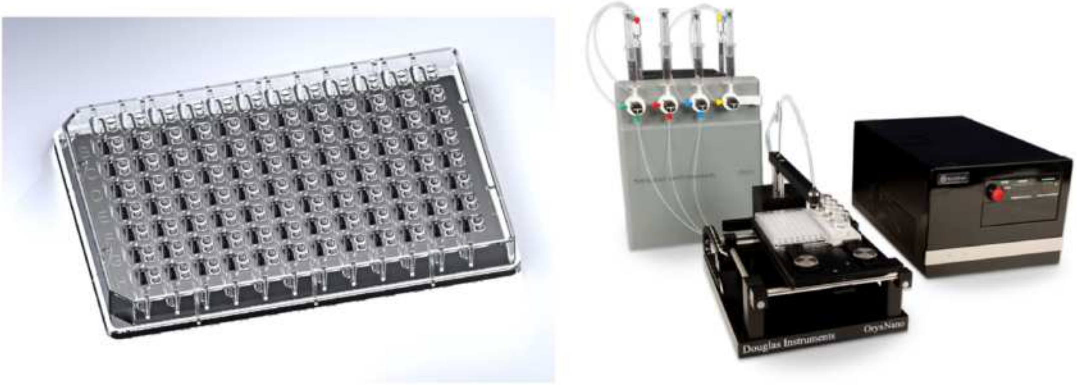


Figure 57. Crystallization screening setup. On the left, a Swiss-CI MRC2 crystallization plate, on the right an Oryx Nano crystallization robot (Douglas Instruments).

Manual crystallization. When initial crystallization conditions are determined by large scale screening, manual reproduction and optimization of the crystallization leads follow. All the manual manipulation is made with the aid of an optical stereoscopic microscope. Manually laid, vapor-diffusion crystallization drops are dispensed on 24-Well, “Linbro” style plates using the hanging-drop vapor diffusion method. The drops are laid on glass cover slips, which are sealed onto the plate wells by means of silicone vacuum grease. Crystallization drops are composed of 1 µL of protein solution + 1 µL of precipitant solution + (optional) 0.2 µL of crystal seed solution. At the end of the dispensing procedure, the plate is incubated at controlled temperature (+4 °C, +9 °C or +18 °C) to allow crystals to grow.

Crystal micro-seeding. This technique allows to obtain high quality single crystals from otherwise difficult-to-crystallize proteins. This procedure consists in adding a micro-seed stock solution to the crystallization drop during its dispensation. Such solution is created by crushing previously obtained, low quality crystals (i.e. needles, spherulites, etc., which are unsuitable for single crystal X-ray diffraction) to sub-micrometric size to reduce the large, irregular crystals to very small, monolithic crystals. The seed stock solution is then diluted and tested in fresh crystallization drops at several dilutions, typically 1:1000 to 1:10000. The seed stock is then flash-frozen in liquid nitrogen or in an ethanol/dry-ice bath and stored at -80 °C for future use. The new crystals grown with the seed stock added typically grow in a more regular morphology, compared to the crystals used for the creation of the seed stock solution itself. If the new crystals present a better morphology but are not yet suitable for X-ray diffraction (e.g. they are too small, too thin, twinned, etc.) a new seed stock can be created from such crystals in a recursive way, in order to gradually improve the crystal quality. In the case of Mpro constructs, the use of a suitable seed stock solution was paramount for the obtainment of diffraction-quality crystals (Figure 58), while for Tha-1 the seed stock solution did not improve the quality of the crystals.

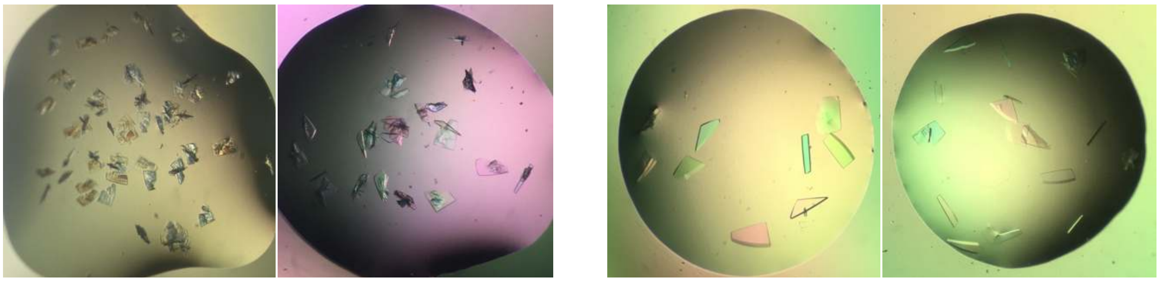


Figure 58. Effect of micro-seeding in Mpro^{WT} crystallization drops. Comparison of manual crystallization drops of apo Mpro^{WT} obtained in absence (on the left) and in presence (on the right) of the seed-stock solution.

Concentration optimization. To test the optimal protein concentration for subsequent crystallization screenings or manual experiments, a specific experiment is dispensed using the Oryx Nano robot. Using a Swiss-CI MRC2 96-well plate, a grid of different crystallization conditions is laid down (Figure 59). On the x axis the protein concentration is varied by using different protein/precipitant volume ratios. On the y axis the precipitant concentration is varied dispensing different ratios of precipitant and diluent (the latter having the same composition of the former - i.e. the same buffer and additives - but deprived of the precipitant molecules). In this way a high number of unique crystallization conditions are tested in a single experiment, and the optimal protein and precipitant concentration are determined.

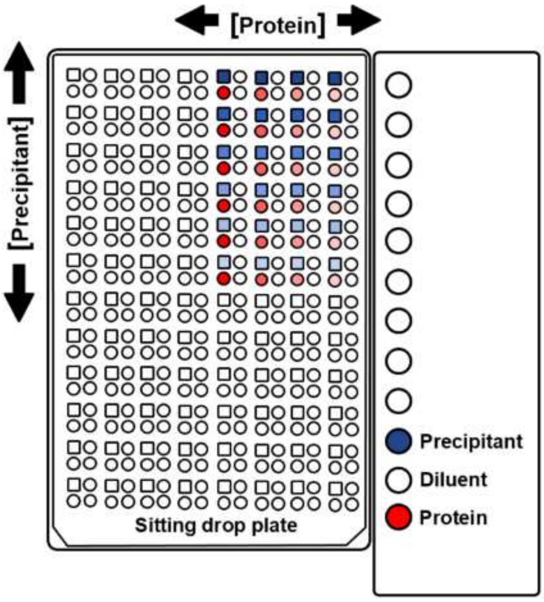


Figure 59. Concentration optimization experiment. On the x axis the protein concentration is varied by using different protein/precipitant volume ratios. On the y axis the precipitant concentration is varied dispensing different ratios of precipitant and diluent (the latter having the same composition of the former - i.e. the same buffer and additives - but deprived of the precipitant molecules).

Co-crystallization and soaking. The crystallization of ligand-protein complexes can be achieved by two methods: co-crystallization and soaking. In the first technique the ligand-protein complex is created in solution by mixing the protein solution with a ligand stock solution and then protracting the crystallization experiment as usual. In the soaking experiment crystals of the protein in *apo* form (that is, in absence of ligands) are created first, then the crystallization wells are open, a small amount of ligand stock solution is added and the well is sealed again. The ligand then slowly diffuses into the crystal lattice through the so-called solvent channels forming the ligand-protein complex.

Both methods are widely employed as they both possess peculiar advantages. The soaking method is ideal in case of high throughput studies, while the co-crystallization method is less prone to artifacts as the protein can rearrange freely in solution upon ligand binding, while in the case of the soaking experiment it would be conformationally restrained by the already-formed crystal contacts. In the case of the complexes of Mpro^{DM} with peptidic substrates, the two techniques produced different structures, with the one obtained with the co-crystallization experiments being more representative of the real complex that is formed in solution.

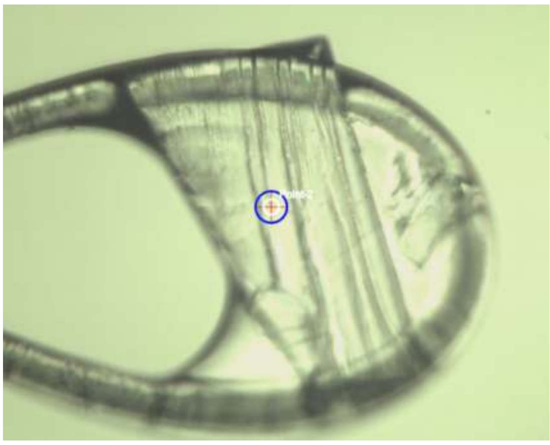


Figure 60. Crystal damage induced by the soaking experiment. The soaking of the NSP4/5 peptide solution to the apo MproDM crystal caused the formation of longitudinal cracks along the crystal structure, indicating a certain degree of internal stress. Despite the macroscopic damage, the diffraction quality of the crystal was sufficient for the resolution of the structure at high resolution.

Crystal fishing and freezing. After the complete growth of the crystals in their mother liquor and a few days before data collection at the synchrotron facility, crystals are fished from their crystallization wells and flash-cooled in liquid nitrogen. A brief dip of the fished crystal in a cryoprotectant solution is sometimes done just prior to the flash-freezing, but it was found to be negatively impact the quality of the crystal diffraction for most our samples. The large majority of the crystals that we tested were not cryo-protected prior to their freezing.

Diffraction data analysis. Data reduction of the diffraction data is protracted with XDS[45] automatically by the ESRF automatic data analysis pipeline. Only for a limited amount of samples data reduction was done locally using the same software starting from the diffraction images with manual settings. For all datasets the molecular replacement method was used to solved the phase method with Phaser[46] of the Phenix suite[47]. In the case of Mpro structures, the model for the molecular replacement was derived from deposited structures of the WT, *apo* enzyme (e.g. PDB ID 6Y2E). In the case of Tha1, the model used to solve the murine Tha1 structure was the deposited structure of Tha1 from *T. maritima* (e.g. PDB ID 1M6S). Refinement was finally done alternating automatic refinement with phenix.refine[48] and manual refinement in Coot[49].

Image rendering. All images presented in this thesis work were prepared by the author, unless otherwise stated. Protein structure illustrations were all prepared with the ChimeraX[50] software.

List of plasmid constructs

Table 13. AhR project plasmid constructs. *The G189R mutation of the AhRe2 construct was casually obtained by an error of the DNA polymerase during the cloning process.

#	Name	source	Span	6His	bHLH	PAS-A	PAS-B	Mutations
1	ARNTa1	murine	M354-E470	✓	-	-	✓	WT
2	AhRa1	human	I280-E389	✓	-	-	✓	WT
3	AhRa2	human	I280-N418	✓	-	-	✓	WT
4	AhRb1	murine	I274-E383	✓	-	-	✓	WT
5	AhRb2	murine	I274-S412	✓	-	-	✓	WT
6	ARNTb1	murine	Q81-E470	✓	✓	✓	✓	WT
7	ARNTb2	murine	K155-E470	✓	-	✓	✓	WT
8	AhRc1	human	G30-E389	✓	✓	✓	✓	WT
9	AhRc2	human	G30-N418	✓	✓	✓	✓	WT
10	AhRc3	human	G30-S474	✓	✓	✓	✓	WT
11	AhRc4	human	G109-E389	✓	-	✓	✓	WT
12	AhRc5	human	G109-N418	✓	-	✓	✓	WT
12	AhRc6	human	G109-S474	✓	-	✓	✓	WT
14	AhRc7	human	G30-E389	-	✓	✓	✓	WT
15	AhRc8	human	G30-N418	-	✓	✓	✓	WT
16	AhRc9	human	G30-S474	-	✓	✓	✓	WT
17	AhRc10	human	G109-E389	-	-	✓	✓	WT
18	AhRc11	human	G109-N418	-	-	✓	✓	WT
19	AhRc12	human	G109-S474	-	-	✓	✓	WT
20	ARNTc1	murine	Q81-E470	-	✓	✓	✓	WT
21	ARNTc4	murine	K155-E470	-	-	✓	✓	WT
22	AhRd1	human	G30-E389	-	✓	✓	✓	WT
23	AhRd2	human	G30-N418	-	✓	✓	✓	WT
24	AhRd3	human	G30-S474	-	✓	✓	✓	WT
25	AhRd4	human	G109-E389	-	-	✓	✓	WT
26	AhRd5	human	G109-N418	-	-	✓	✓	WT
27	AhRd6	human	G109-S474	-	-	✓	✓	WT
28	AhRe1	human	G30-E389	✓	✓	✓	✓	WT
29	AhRe2	human	G30-N418	✓	✓	✓	✓	WT, G189R*
30	AhRe3	human	G30-S474	✓	✓	✓	✓	WT
31	AhRe4	human	G109-E389	✓	-	✓	✓	WT
32	AhRe5	human	G109-N418	✓	-	✓	✓	WT
33	AhRe6	human	G109-S474	✓	-	✓	✓	WT
34	ARNTc1	murine	Q81-E470	-	✓	✓	✓	WT
35	ARNTc2	murine	K155-E470	-	-	✓	✓	WT
36	ARNTd1	murine	Q81-E470	✓	✓	✓	✓	WT
37	ARNTd2	murine	K155-E470	✓	-	✓	✓	WT
38	AhRf1	human	N284-R398	✓	-	-	✓	WT
39	AhRf2	murine	N278-R392	✓	-	-	✓	WT

#	Name	source	Span	6His	bHLH	PAS-A	PAS-B	Mutations
40	AhRg1	human	N284-R398	✓	-	-	✓	L331E
41	AhRg2	human	N284-R398	✓	-	-	✓	I338Q
42	AhRg3	human	N284-R398	✓	-	-	✓	L331E + I338Q

Table 14. M$_{\text{pro}}$ project plasmid constructs.

#	Name	Source	Span	Vector	Mutations
1	Mpro$^{\text{WT}}$	SARS-CoV-2	S3264-Q3569	pGEX-6P-1	WT
2	Mpro$^{\text{WT}}$	SARS-CoV-2	S3264-Q3569	pET-SUMO	WT
3	Mpro$^{\text{DM}}$	SARS-CoV-2	S3264-Q3569	pET-SUMO	H41A+C145A
4	Mpro$^{\text{Cdel}}$	SARS-CoV-2	S3264-T3462	pET-SUMO	WT
5	Mpro$^{\text{F140A}}$	SARS-CoV-2	S3264-Q3569	pET-SUMO	F140A
6	Mpro$^{\text{F140P}}$	SARS-CoV-2	S3264-Q3569	pET-SUMO	F140P

Table 15. AhR project cloning and mutagenesis primers

#	name	Vector	Forward primer	Reverse primer
1	ARNTa1	pET-SUMO	mARNT.fw	mARNT.rv
2	AhRa1	pET-SUMO	hAhR.fw	hAhR_sh.rv
3	AhRa2	pET-SUMO	hAhR.fw	hAhR_ln.rv
4	AhRb1	pET-SUMO	mAhR.fw	mAhR_sh.rv
5	AhRb2	pET-SUMO	mAhR.fw	mAhR_ln.rv
6	ARNTb1	pET-SUMO	pET-SUMO_Arnt_1.fw	pET-SUMO_Arnt_1.rv
7	ARNTb2	pET-SUMO	pET-SUMO_Arnt_2.fw	pET-SUMO_Arnt_1.rv
8	AhRc1	pDUET (MCS1)	His_AhR_1.fw	His_AhR_1.rv
9	AhRc2	pDUET (MCS1)	His_AhR_1.fw	His_AhR_2.rv
10	AhRc3	pDUET (MCS1)	His_AhR_1.fw	His_AhR_3.rv
11	AhRc4	pDUET (MCS1)	His_AhR_4.fw	His_AhR_1.rv
12	AhRc5	pDUET (MCS1)	His_AhR_4.fw	His_AhR_2.rv
12	AhRc6	pDUET (MCS1)	His_AhR_4.fw	His_AhR_3.rv
14	AhRc7	pDUET (MCS1)	AhR_7.fw	His_AhR_1.rv
15	AhRc8	pDUET (MCS1)	AhR_7.fw	His_AhR_2.rv
16	AhRc9	pDUET (MCS1)	AhR_7.fw	His_AhR_3.rv
17	AhRc10	pDUET (MCS1)	AhR_10.fw	His_AhR_1.rv
18	AhRc11	pDUET (MCS1)	AhR_10.fw	His_AhR_2.rv
19	AhRc12	pDUET (MCS1)	AhR_10.fw	His_AhR_3.rv
20	ARNTc1	pDUET (MCS2)	Arnt_1.fw	Arnt_1.rv
21	ARNTc4	pDUET (MCS2)	Arnt_4.fw	Arnt_1.rv
22	AhRd1	pET-21d(+)	pET-21d_bHLH_AHR.fw	pET-21d_AhR_1.rv

#	name	Vector	Forward primer	Reverse primer
23	AhRd2	pET-21d(+)	pET-21d_bHLH_AHR.fw	pET-21d_AhR_2.rv
24	AhRd3	pET-21d(+)	pET-21d_bHLH_AHR.fw	pET-21d_AhR_3.rv
25	AhRd4	pET-21d(+)	pET-21d_AhR_1.fw	pET-21d_AhR_1.rv
26	AhRd5	pET-21d(+)	pET-21d_AhR_1.fw	pET-21d_AhR_2.rv
27	AhRd6	pET-21d(+)	pET-21d_AhR_1.fw	pET-21d_AhR_3.rv
28	AhRe1	pET-SUMO	pET-SUMO_AhR_1.fw	pET-SUMO_AhR_1.rv
29	AhRe2	pET-SUMO	pET-SUMO_AhR_1.fw	pET-SUMO_AhR_2.rv
30	AhRe3	pET-SUMO	pET-SUMO_AhR_1.fw	pET-SUMO_AhR_3.rv
31	AhRe4	pET-SUMO	pET-SUMO_AhR_2.fw	pET-SUMO_AhR_1.rv
32	AhRe5	pET-SUMO	pET-SUMO_AhR_2.fw	pET-SUMO_AhR_2.rv
33	AhRe6	pET-SUMO	pET-SUMO_AhR_2.fw	pET-SUMO_AhR_3.rv
34	AhRf1	pET-SUMO	pET-SUMO_AhRf1.fw	pET-SUMO_AhRf1.rv
35	AhRf2	pET-SUMO	pET-SUMO_AhRf2.fw	pET-SUMO_AhRf2.rv
36	ARNTc1	pET-21d(+)	pET-21d(+)_ARNT_1_fw	pET-21d(+)_ARNT_1_rv
37	ARNTc2	pET-21d(+)	pET-21d(+)_ARNT_2_fw	pET-21d(+)_ARNT_1_rv
38	ARNTd1	pET-21d(+)	pET-21d(+)_ARNT_1_fw	pET-21d(+)_ARNT_2_rv
39	ARNTd2	pET-21d(+)	pET-21d(+)_ARNT_2_fw	pET-21d(+)_ARNT_2_rv
40	AhRg1	pET-SUMO	AhRg1.fw	AhRg1.rv
41	AhRg2	pET-SUMO	AhRg2.fw	AhRg2.rv
42	AhRg3	pET-SUMO	AhRg3.fw	AhRg3.rv

List of primers

Table 16. Primer sequences for the AhR project. Stop codons are highlighted in red, start codons in green and mutations in purple. The vertical separator \| highlights the formal separation between the flanking regions corresponding to the vector (plasmid) and the insert (for primers used in RF cloning method).

#	Name	Sequence
1	mArnt.fw	5’-ATGAGTAACATTTGTCAGCCAAC-3’
2	mArnt.rv	3’-**TTA**TTCCTGGCTAGAGTTCTTCAC-5’
3	hAhr.fw	5’-ATCCGGACCAAAAATTTTATC-3’
4	hAhr_sh.rv	3’-**TTA**CTCATCTGTTAGTGGTCTCTGAG-5’
5	hAhr_ln.rv	3’-**TTA**GTTGGTTGCCTCATACAACAC-5’
6	mAhr.fw	5’-ATTCGAACCAAAAACTTCATC-3’
7	mAhr_sh.rv	3’-**TTA**TTCATCCGTCAGTGGTCTC-5’
8	mAhr_ln.rv	3’-**TTA**GCTGGAGATCTCGTACAACAC-5’
9	pET-SUMO_Arnt_1.fw	5’-GCTCACAGAGAACAGATTGGTGGT|CAGAGCTCTGCGGATAAAGAGAGACT-3’
10	pET-SUMO_Arnt_2.fw	5’-GCTCACAGAGAACAGATTGGTGGT|AAGCCATCTTTCCTCACTGATCAGG-3’
11	pET-SUMO_Arnt_1.rv	3’-GCCGAATAAATACCTAAGCTTGTCT|**TTA**TTCCTGGCTAGAGTTCTTCACATTGG-5’
12	His_AhR_1.fw	5’-GGCAGCAGCCATCACCATCATCACCAC|GGAATCAAGTCAAATCCTTCCAAGC-3’
14	His_AhR_4.fw	5’-GGCAGCAGCCATCACCATCATCACCAC|GGCCTGAACTTACAAGAAGGAGAATTC-3’
15	AhR_7.fw	5’-AACTTTAATAAGGAGATATACC|**ATG**GGAATCAAGTCAAATCCTTCCAAGC-3’
16	AhR_10.fw	5’-AACTTTAATAAGGAGATATACC|**ATG**GGCCTGAACTTACAAGAAGGAGAATTC-3’
17	Arnt_1.fw	5’-GTTAAGTATAAGAAGGAGATATACAT|**ATG**CAGAGCTCTGCGGATAAAGAGAGACT-3’
18	Arnt_4.fw	5’-GTTAAGTATAAGAAGGAGATATACAT|**ATG**AAGCCATCTTTCCTCACTGATCAGG-3’
19	His_AhR_1.rv	3’-ATGCGGCCGCAAGCTTGTCGACCTGCAGG|**TTA**CTCATCTGTTAGTGGTCTCTGAGTTAC-5’
20	His_AhR_2.rv	3’-ATGCGGCCGCAAGCTTGTCGACCTGCAGG|**TTA**GTTGGTTGCCTCATACAACACAGC-5’
21	His_AhR_3.rv	3’-ATGCGGCCGCAAGCTTGTCGACCTGCAGG|**TTA**ACTTGAAGCAGGATAGAGATAAATAGAC-5’
22	Arnt_1.rv	3’-GGTTTCTTTACCAGACTCGAGGGTACC|**TTA**TTCCTGGCTAGAGTTCTTCACATTGG-5’
23	pET-21d_bHLH_AHR.fw	5’-TTTAAGAAGGAGATATACCATGGGA|GGAATCAAGTCAAATCCTTCCAAGC-3’
24	pET-21d_AhR_1.fw	5’-TTTAAGAAGGAGATATACCATGGGA|GGCCTGAACTTACAAGAAGGAGAATTC-3’
25	pET-21d_AhR_1.rv	3’-TCAGTGGTGGTGGTGGTGGTGCTCGAG|**TTA**CTCATCTGTTAGTGGTCTCTGAGTTAC-5’
26	pET-21d_AhR_2.rv	3’-TCAGTGGTGGTGGTGGTGGTGCTCGAG|**TTA**GTTGGTTGCCTCATACAACACAGC-5’
27	pET-21d_AhR_3.rv	3’-TCAGTGGTGGTGGTGGTGGTGCTCGAG|**TTA**ACTTGAAGCAGGATAGAGATAAATAGAC-5’
28	pET-SUMO_AhRf1.fw	5’-GCTCACAGAGAACAGATTGGTGGT|AATTTTATCTTTAGAACCAAACACA-3’
29	pET-SUMO_AhRf1.rv	3’-GCCGAATAAATACCTAAGCTTGTCT|**TTA**TCGTTTTCGTAAATGCTC-5’
30	pET-SUMO_AhRf2.fw	5’-GCTCACAGAGAACAGATTGGTGGT|AACTTCATCTTCAGGACC-3’
31	pET-SUMO_AhRf2.rv	3’-GCCGAATAAATACCTAAGCTTGTCT|**TTA**TCGCTTCTGTAAATGCTC-5’
32	pET-21d(+)_ARNT_1_fw	5’-TTTAAGAAGGAGATATACCATGGGA|CAGAGCTCTGCGGATAAAGAGAGACT-3’
33	pET-21d(+)_ARNT_1_rv	3’-TCAGTGGTGGTGGTGGTGGTGCTCGAG**TTA**TTCCTGGCTAGAGTTCTTCACATTGG-5’
34	pET-21d(+)_ARNT_2_fw	5’-TTTAAGAAGGAGATATACCATGGGA|AAGCCATCTTTCCTCACTGATCAGG-3’
35	pET-21d(+)_ARNT_2_rv	3’-TCAGTGGTGGTGGTGGTGGTGCTCGAGTTCCTGGCTAGAGTTCTTCACATTGG-5’
36	AhRg1.fw	5’-GCAGCTGATATG**GAA**TATTGTGCCGAG-3’

#	Name	Sequence
37	AhRg1.rv	3’-CTCGGCACAATA**TTC**CATATCAGCTGC-5’
38	AhRg2.fw	5’-GCCGAGTCCCAT**CAA**CGAATGATTAAGACTGGA-3’
39	AhRg2.rv	3’-TCCAGTCTTAATCATTCG**TTG**ATGGGACTCGGC-5’
40	AhRg3.fw	5’-GCAGCTGATATG**GAA**TATTGTGCCGAGTCCCAT**CAA**CGAATGATTAAGACTGGA-3’
41	AhRg3.rv	3’-TCCAGTCTTAATCATTCG**TTG**ATGGGACTCGGCACAATA**TTC**CATATCAGCTGC-5’

Table 17. Primer sequences for the $\mathrm{M_{pro}}$ project.

#	Name	Sequence
1	Mpro_H41A.fw	5’-GTCTATTGCCCTCGT**GCT**GTCATCTGCACCTCT-3’
2	Mpro_H41A.rv	3’-AGAGGTGCAGATGAC**AGC**ACGAGGGCAATAGAC-5’
3	Mpro_C145A.fw	5’-TTCCTTAATGGCAGC**GCT**GGTTCGGTGGGCTTT-3’
4	Mpro_C145A.rv	3’-AAAGCCCACCGAACC**AGC**GCTGCCATTAAGGAA-5’
5	Mpro_F140P.fw	5’-ACGATCAAAGGCAGC**CCC**CTTAATGGCAGCTGT-3’
6	Mpro_F140P.rv	3’-ACAGCTGCCATTAAG**GGG**GCTGCCTTTGATCGT-5’
7	Mpro_F140A.fw	5’-ACGATCAAAGGCAGC**GCC**CTTAATGGCAGCTGT-3’
8	Mpro_F140A.rv	3’-ACAGCTGCCATTAAG**GGC**GCTGCCTTTGATCGT-5’
9	Mpro_DM.fw	5’-GCTCACAGAGAACAGATTGGTGGT|TCGGGGTTTCGCAAAATGGCGTTTCCG-3’
10	Mpro_DM.rv	3’-CCGAATAAATACCTAAGCTTGTCT**TTA**|CTGAAACGTGACACCGCTACACTG-5’

List of crystallographic data collections

Table 18. Summary of X-Ray data collections.

#	Date	Beamline	Nr. of crystals	Constructs
1	13-14/11/2020	ID23-2	32	MproWT + inhibitors
2	09-10/12/2020	ID23-1	32	MproWT + inhibitors
3	11-12/12/2020	ID30A3	-	Same samples as above (recovery session)
4	01-02/07/2021	ID23-2	32	MproWT + inhibitors (Masitinib, Vitas)
5	02-03/03/2022	ID23-1	38	MproWT + inhibitors (Molport)
6	04-05/06/2022	ID23-1	30	MproWT, MproDM + Inhibitors, peptides
7	19-20/11/2022	ID23-1	26	MproF140P, MproF140A + Inhibitors, peptides
8	28-29/01/2023	ID23-1	42	MproF140A, MproDM + Inhibitors, peptides
9	11-12/03/2023	ID23-2	9	Tha1
10	09-10/06/2023	ID23-2	48	MproF140P + Inhibitors

All data collections were conducted remotely at the European Synchrotron Radiation Facility (ESRF), Grenoble, France. Crystals were manually fished from the crystallization drops under a microscope and flash-frozen in liquid nitrogen. The frozen crystals mounted on their nylon loops were arranged either in SPINE pucks or in UNIPUCKs, put in a shipping dewar (dry shipper) and sent to ESRF. The crystals were maintained at 77 K from the flash-freezing process until the data collection at the beamline.

Data collection and structure statistics

Table 19. Data collection and structure statistics for $M_{pro}{}^{Cdel}$

Resolution range	28.47 - 2.17 (2.248 - 2.17)
Space group	P 21 21 21
Unit cell	52.963 62.183 120.582 90 90 90
Total reflections	39417 (3640)
Unique reflections	21157 (1966)
Multiplicity	1.9 (1.9)
Completeness (%)	97.16 (91.85)
Mean I/sigma(I)	9.19 (1.23)
Wilson B-factor	45.19
R-merge	0.05122 (0.5588)
R-meas	0.07244 (0.7903)
R-pim	0.05122 (0.5588)
CC1/2	0.997 (0.626)
CC*	0.999 (0.877)
R-work	0.2158 (0.3446)
R-free	0.2891 (0.4413)
Number of non-hydrogen atoms	2775
protein	2712
ligands	5
solvent	58
Protein residues	352
RMS(bonds)	0.014
RMS(angles)	1.34
Ramachandran favored (%)	90.70
Ramachandran allowed (%)	7.85
Ramachandran outliers (%)	1.45
Rotamer outliers (%)	2.61
Average B-factor	54.57
protein	54.56
solvent	49.85

Table 20. Data collection and structure statistics for $\mathrm{M_{pro}}^{\mathrm{WT}}$ (*new* conformation)

X-Ray source	ID23-2, ESRF
Wavelength (Å)	0.873130
Resolution range	55.44 - 1.58 (1.61 - 1.58)
Space group	C2
Unit cell	113.07 54.71 44.84 90.0 101.30 90.0
Total reflections	145297 (7276)
Unique reflections	36653 (1847)
Multiplicity	4.0 (3.9)
Completeness (%)	99.4 (99.5)
Mean I/sigma(I)	9.2 (1.0)
Wilson B-factor	23.7
R-merge	0.070 (1.305)
R-meas	0.081 (1.505)
R-pim	0.040 (0.739)
CC1/2	0.998 (0.357)
R-work	0.1771
R-free	0.2031
Number of non-hydrogen atoms	2568
protein	2350
solvent	218
RMS(bonds)	0.008
RMS(angles)	0.868
Ramachandran favored (%)	97.99
Ramachandran allowed (%)	2.01
Ramachandran outliers (%)	0.00
Average B-factor	
protein	32.6
solvent	43.1
PDB code	7NIJ

Table 21. Data collection and structure statistics for $M_{pro}{}^{DM}$

X-ray source	ESRF ID23-1				
Wavelength (Å)	0.8856				
Sample	Free form (apo)	NSP4/5 co-cryst.	NSP4/5 soak.	NSP5/6 co-cryst.	NSP14/15co-cryst.
Space group	C2	P1	C2	$P2_12_12_1$	$P2_12_12_1$
N° molecule per ASU	1	2	1	2	2
Cell dimensions					
a, b, c (Å)	114.35, 53.24, 44.70	53.56, 61.36, 67.90	113.69, 52.19, 45.15	67.75, 99.04, 101.99	67.90, 99.03, 100.84
α, β, γ (°)	90.00, 102.90, 90.00	92.21, 109.10, 108.38	90.00, 103.31, 90.00	90.00, 90.00, 90.00	90.00, 90.00, 90.00
Resolution range (Å)	48.04-1.80 (1.84-1.80)	46.88-1.78 (1.81-1.78)	47.20-1.70 (1.73-1.70)	49.52-1.80 (1.84-1.80)	55.95-1.87 (1.91-1.87)
R_{merge}	0.076 (1.502)	0.038 (0.003)	0.068 (1.103)	0.077 (1.4607)	0.057 (0.907)
R_{meas}	0.083 (1.625)	0.045 (1.195)	0.075 (1.278)	0.081 (1.726)	0.067 (1.064)
R_{pim}	0.031 (0.609)	0.024 (0.652)	0.031 (0.633)	0.025 (0.603)	0.034 (0.547)
Total number of observations	170902 (8756)	240323 (7094)	146742 (3491)	668786 (25605)	213122 (12099)
Total number unique	24239 (1303)	68784 (2308)	26488 (911)	64002 (3564)	55875 (3408)
Mean(I)/σ(I)	11.3 (1.1)	13.8 (0.9)	11.9 (1.1)	16.1 (1.0)	10.5 (1.2)
$CC_{1/2}$	0.999 (0.495)	0.999 (0.665)	0.998 (0.487)	0.998 (0.485)	0.998 (0.536)
Completeness (%)	99.4 (91.4)	93.1 (55.0)	93.3 (60.1)	99.5 (95.2)	98.4 (95.2)
Multiplicity	7.1 (6.7)	3.5 (3.1)	5.5 (3.8)	10.4 (7.2)	3.8 (3.6)
Wilson B estim. ($Å^2$)	35.97	38.47	26.27	32.30	37.98
R_{work}/R_{free} (%)	18.68/20.40	17.40/19.93	15.87/18.55	17.93/21.78	18.86/22.02
Number of atoms					
Protein	2365	4651	2376	4675	4710
Water	163	331	302	335	186
Average B, all atoms	39.89	50.76	30.34	40.67	50.99
RMSD					
Bond lengths (Å)	0.007	0.018	0.004	0.014	0.006
Bond angles (°)	0.695	1.481	0.721	1.233	0.822
Ramachandran statistics					
Favored (%)	97.69	98.06	97.44	97.49	96.31
Allowed (%)	1.98	1.94	2.56	2.35	3.52
Outliers (%)	0.33	0	0	0.17	0.17

Table 22. Data collection and structure statistics for murine Tha1

Data collection		
X-ray source	ESRF ID23-2	
Wavelength (Å)	0.8731	
Space group	F222	C2
N° of molecules per ASU	1	2
Cell dimensions		
a, b, c (Å)	83.69, 100.52, 171.26	83.97, 101.64, 95.96
α, β, γ (°)	90.00, 90.00, 90.00	90.00, 116.14, 90.00
Resolution range (Å)	25.71 - 2.26 (2.33 - 2.26)	43.07 - 2.60 (2.72 - 2.60)
R_{merge}	0.084 (0.842)	0.104 (0.731)
R_{meas}	0.088 (0.887)	0.115 (0.804)
R_{pim}	0.027 (0.272)	0.047 (0.325)
Total number of observations	168470 (15409)	109798 (14100)
Total number unique	17051 (1561)	21160 (2662)
Mean(I)/σ(I)	16.6 (1.5)	8.9 (1.4)
$CC_{1/2}$	0.998 (0.881)	0.996 (0.786)
Completeness (%)	99.7 (99.9)	94.8 (98.0)
Multiplicity	9.9 (9.9)	5.2 (5.3)
Wilson B estimate (Å^2)	46.4	65.3

Refinement		
Resolution range (Å)	25.71 - 2.26	43.07 - 2.60
R_{work}/R_{free} (%)	21.0/24.0	23.4/25.9
Number of atoms		
Protein	2806	5536
Water	57	29
Average B, all atoms (Å^2)	67.0	83.0
RMSD		
Bond lengths (Å)	0.003	0.002
Bond angles (°)	0.596	0.560
Ramachandran statistics		
Favored (%)	95.4	95.8
Allowed (%)	4.1	3.9
Outliers (%)	0.5	0.3
PDB entry	8PUS	8PUM

References

(1) Wu, D.; Rastinejad, F. Structural Characterization of Mammalian bHLH-PAS Transcription Factors. *Curr Opin Struct Biol* **2017**, *43*, 1–9. https://doi.org/10.1016/j.sbi.2016.09.011.

(2) Fribourgh, J. L.; Partch, C. L. Assembly and Function of bHLH–PAS Complexes. *Proceedings of the National Academy of Sciences* **2017**, *114* (21), 5330–5332. https://doi.org/10.1073/pnas.1705408114.

(3) de Martin, X.; Sodaei, R.; Santpere, G. Mechanisms of Binding Specificity among bHLH Transcription Factors. *International Journal of Molecular Sciences* **2021**, *22* (17), 9150. https://doi.org/10.3390/ijms22179150.

(4) Neavin, D. R.; Liu, D.; Ray, B.; Weinshilboum, R. M. The Role of the Aryl Hydrocarbon Receptor (AHR) in Immune and Inflammatory Diseases. *International Journal of Molecular Sciences* **2018**, *19* (12), 3851. https://doi.org/10.3390/ijms19123851.

(5) Rothhammer, V.; Quintana, F. J. The Aryl Hydrocarbon Receptor: An Environmental Sensor Integrating Immune Responses in Health and Disease. *Nat Rev Immunol* **2019**, *19* (3), 184–197. https://doi.org/10.1038/s41577-019-0125-8.

(6) Kolonko, M.; Greb-Markiewicz, B. bHLH–PAS Proteins: Their Structure and Intrinsic Disorder. *International Journal of Molecular Sciences* **2019**, *20* (15). https://doi.org/10.3390/ijms20153653.

(7) Schulte, K. W.; Green, E.; Wilz, A.; Platten, M.; Daumke, O. Structural Basis for Aryl Hydrocarbon Receptor-Mediated Gene Activation. *Structure* **2017**, *25* (7), 1025-1033.e3. https://doi.org/10.1016/j.str.2017.05.008.

(8) Seok, S.-H.; Lee, W.; Jiang, L.; Molugu, K.; Zheng, A.; Li, Y.; Park, S.; Bradfield, C. A.; Xing, Y. Structural Hierarchy Controlling Dimerization and Target DNA Recognition in the AHR Transcriptional Complex. *Proceedings of the National Academy of Sciences* **2017**, *114* (21), 5431–5436. https://doi.org/10.1073/pnas.1617035114.

(9) Pandini, A.; Denison, M. S.; Song, Y.; Soshilov, A. A.; Bonati, L. Structural and Functional Characterization of the Aryl Hydrocarbon Receptor Ligand Binding Domain by Homology Modeling and Mutational Analysis. *Biochemistry* **2007**, *46* (3), 696–708. https://doi.org/10.1021/bi061460t.

(10) Motto, I.; Bordogna, A.; Soshilov, A. A.; Denison, M. S.; Bonati, L. New Aryl Hydrocarbon Receptor Homology Model Targeted To Improve Docking Reliability. *J. Chem. Inf. Model.* **2011**, *51* (11), 2868–2881. https://doi.org/10.1021/ci2001617.

(11) Soshilov, A. A.; Denison, M. S. Ligand Promiscuity of Aryl Hydrocarbon Receptor Agonists and Antagonists Revealed by Site-Directed Mutagenesis. *Mol Cell Biol* **2014**, *34* (9), 1707–1719. https://doi.org/10.1128/MCB.01183-13.

(12) Gruszczyk, J.; Grandvuillemin, L.; Lai-Kee-Him, J.; Paloni, M.; Savva, C. G.; Germain, P.; Grimaldi, M.; Boulahtouf, A.; Kwong, H.-S.; Bous, J.; Ancelin, A.; Bechara, C.; Barducci, A.; Balaguer, P.; Bourguet, W. Cryo-EM Structure of the Agonist-Bound Hsp90-XAP2-AHR Cytosolic Complex. *Nat Commun* **2022**, *13* (1), 7010. https://doi.org/10.1038/s41467-022-34773-w.

(13) Wen, Z.; Zhang, Y.; Zhang, B.; Hang, Y.; Xu, L.; Chen, Y.; Xie, Q.; Zhao, Q.; Zhang, L.; Li, G.; Zhao, B.; Sun, F.; Zhai, Y.; Zhu, Y. Cryo-EM Structure of the Cytosolic AhR Complex. *Structure* **2023**, *31* (3), 295-308.e4. https://doi.org/10.1016/j.str.2022.12.013.

(14) Dai, S.; Qu, L.; Li, J.; Zhang, Y.; Jiang, L.; Wei, H.; Guo, M.; Chen, X.; Chen, Y. Structural Insight into the Ligand Binding Mechanism of Aryl Hydrocarbon Receptor. *Nat Commun* **2022**, *13* (1), 6234. https://doi.org/10.1038/s41467-022-33858-w.

(15) Murray, I. A.; Patterson, A. D.; Perdew, G. H. Aryl Hydrocarbon Receptor Ligands in Cancer: Friend and Foe. *Nat Rev Cancer* **2014**, *14* (12), 801–814. https://doi.org/10.1038/nrc3846.

(16) Kelley, L. A.; Mezulis, S.; Yates, C. M.; Wass, M. N.; Sternberg, M. J. E. The Phyre2 Web Portal for Protein Modeling, Prediction and Analysis. *Nat Protoc* **2015**, *10* (6), 845–858. https://doi.org/10.1038/nprot.2015.053.

(17) Conchúir, S. Ó.; Barlow, K. A.; Pache, R. A.; Ollikainen, N.; Kundert, K.; O'Meara, M. J.; Smith, C. A.; Kortemme, T. A Web Resource for Standardized Benchmark Datasets, Metrics, and Rosetta Protocols for Macromolecular Modeling and Design. *PLOS ONE* **2015**, *10* (9), e0130433. https://doi.org/10.1371/journal.pone.0130433.

(18) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2.

(19) Bottomley, M. J.; Muraglia, E.; Bazzo, R.; Carfì, A. Molecular Insights into Quorum Sensing in the Human Pathogen Pseudomonas Aeruginosa from the Structure of the Virulence Regulator LasR Bound to Its Autoinducer *. *Journal of Biological Chemistry* **2007**, *282* (18), 13592–13600. https://doi.org/10.1074/jbc.M700556200.

(20) Song, J.; Clagett-Dame, M.; Peterson, R. E.; Hahn, M. E.; Westler, W. M.; Sicinski, R. R.; DeLuca, H. F. A Ligand for the Aryl Hydrocarbon Receptor Isolated from Lung. *Proceedings of the National Academy of Sciences* **2002**, *99* (23), 14694–14699. https://doi.org/10.1073/pnas.232562899.

(21) Li, M.; Su, Z.-G.; Janson, J.-C. In Vitro Protein Refolding by Chromatographic Procedures. *Protein Expr Purif* **2004**, *33* (1), 1–10. https://doi.org/10.1016/j.pep.2003.08.023.

(22) Yamaguchi, H.; Miyazaki, M. Refolding Techniques for Recovering Biologically Active Recombinant Proteins from Inclusion Bodies. *Biomolecules* **2014**, *4* (1), 235–251. https://doi.org/10.3390/biom4010235.

(23) Holmes, K. V. CORONAVIRUSES (CORONAVIRIDAE). *Encyclopedia of Virology* **1999**, 291–298. https://doi.org/10.1006/rwvi.1999.0055.

(24) Hillary, V. E.; Ceasar, S. A. An Update on COVID-19: SARS-CoV-2 Variants, Antiviral Drugs, and Vaccines. *Heliyon* **2023**, *9* (3). https://doi.org/10.1016/j.heliyon.2023.e13952.

(25) Kumari, M.; Lu, R.-M.; Li, M.-C.; Huang, J.-L.; Hsu, F.-F.; Ko, S.-H.; Ke, F.-Y.; Su, S.-C.; Liang, K.-H.; Yuan, J. P.-Y.; Chiang, H.-L.; Sun, C.-P.; Lee, I.-J.; Li, W.-S.; Hsieh, H.-P.; Tao, M.-H.; Wu, H.-C. A Critical Overview of Current Progress for COVID-19: Development of Vaccines, Antiviral Drugs, and Therapeutic Antibodies. *Journal of Biomedical Science* **2022**, *29* (1), 68. https://doi.org/10.1186/s12929-022-00852-9.

(26) V'kovski, P.; Kratzel, A.; Steiner, S.; Stalder, H.; Thiel, V. Coronavirus Biology and Replication: Implications for SARS-CoV-2. *Nat Rev Microbiol* **2021**, *19* (3), 155–170. https://doi.org/10.1038/s41579-020-00468-6.

(27) Narayanan, A.; Narwal, M.; Majowicz, S. A.; Varricchio, C.; Toner, S. A.; Ballatore, C.; Brancale, A.; Murakami, K. S.; Jose, J. Identification of SARS-CoV-2 Inhibitors Targeting Mpro and PLpro Using in-Cell-Protease Assay. *Commun Biol* **2022**, *5* (1), 1–17. https://doi.org/10.1038/s42003-022-03090-9.

(28) Kneller, D. W.; Phillips, G.; O'Neill, H. M.; Jedrzejczak, R.; Stols, L.; Langan, P.; Joachimiak, A.; Coates, L.; Kovalevsky, A. Structural Plasticity of SARS-CoV-2 3CL Mpro Active Site Cavity Revealed by Room Temperature X-Ray Crystallography. *Nat Commun* **2020**, *11* (1), 3202. https://doi.org/10.1038/s41467-020-16954-7.

(29) Shaqra, A. M.; Zvornicanin, S. N.; Huang, Q. Y. J.; Lockbaum, G. J.; Knapp, M.; Tandeske, L.; Bakan, D. T.; Flynn, J.; Bolon, D. N. A.; Moquin, S.; Dovala, D.; Kurt Yilmaz, N.; Schiffer, C. A. Defining the Substrate Envelope of SARS-CoV-2 Main Protease to Predict and Avoid Drug Resistance. *Nat Commun* **2022**, *13* (1), 3556. https://doi.org/10.1038/s41467-022-31210-w.

(30) Zhang, L.; Lin, D.; Sun, X.; Curth, U.; Drosten, C.; Sauerhering, L.; Becker, S.; Rox, K.; Hilgenfeld, R. Crystal Structure of SARS-CoV-2 Main Protease Provides a Basis for Design of Improved α-Ketoamide Inhibitors. *Science* **2020**, *368* (6489), 409–412. https://doi.org/10.1126/science.abb3405.

(31) Shi, J.; Sivaraman, J.; Song, J. Mechanism for Controlling the Dimer-Monomer Switch and Coupling Dimerization to Catalysis of the Severe Acute Respiratory Syndrome Coronavirus 3C-Like Protease. *Journal of Virology* **2008**, *82* (9), 4620–4629. https://doi.org/10.1128/jvi.02680-07.

(32) MacDonald, E. A.; Frey, G.; Namchuk, M. N.; Harrison, S. C.; Hinshaw, S. M.; Windsor, I. W. Recognition of Divergent Viral Substrates by the SARS-CoV-2 Main Protease. *ACS Infect. Dis.* **2021**, *7* (9), 2591–2595. https://doi.org/10.1021/acsinfecdis.1c00237.

(33) Ramos-Guzmán, C. A.; Ruiz-Pernía, J. J.; Tuñón, I. Unraveling the SARS-CoV-2 Main Protease Mechanism Using Multiscale Methods. *ACS Catal.* **2020**, *10* (21), 12544–12554. https://doi.org/10.1021/acscatal.0c03420.

(34) Andi, B.; Kumaran, D.; Kreitler, D. F.; Soares, A. S.; Keereetaweep, J.; Jakoncic, J.; Lazo, E. O.; Shi, W.; Fuchs, M. R.; Sweet, R. M.; Shanklin, J.; Adams, P. D.; Schmidt, J. G.; Head, M. S.; McSweeney, S. Hepatitis C Virus NS3/4A Inhibitors and Other Drug-like Compounds as Covalent Binders of SARS-CoV-2 Main Protease. *Sci Rep* **2022**, *12* (1), 12197. https://doi.org/10.1038/s41598-022-15930-z.

(35) Ghahremanpour, M. M.; Tirado-Rives, J.; Deshmukh, M.; Ippolito, J. A.; Zhang, C.-H.; Cabeza de Vaca, I.; Liosi, M.-E.; Anderson, K. S.; Jorgensen, W. L. Identification of 14 Known Drugs as Inhibitors of the Main Protease of SARS-CoV-2. *ACS Med. Chem. Lett.* **2020**, *11* (12), 2526–2533. https://doi.org/10.1021/acsmedchemlett.0c00521.

(36) Drayman, N.; DeMarco, J. K.; Jones, K. A.; Azizi, S.-A.; Froggatt, H. M.; Tan, K.; Maltseva, N. I.; Chen, S.; Nicolaescu, V.; Dvorkin, S.; Furlong, K.; Kathayat, R. S.; Firpo, M. R.; Mastrodomenico, V.; Bruce, E. A.; Schmidt, M. M.; Jedrzejczak, R.; Muñoz-Alía, M. Á.; Schuster, B.; Nair, V.; Han, K.; O'Brien, A.; Tomatsidou, A.; Meyer, B.; Vignuzzi, M.; Missiakas, D.; Botten, J. W.; Brooke, C. B.; Lee, H.; Baker, S. C.; Mounce, B. C.; Heaton, N. S.; Severson, W. E.; Palmer, K. E.; Dickinson, B. C.; Joachimiak, A.; Randall, G.; Tay, S. Masitinib Is a Broad Coronavirus 3CL Inhibitor That Blocks Replication of SARS-CoV-2. *Science* **2021**, *373* (6557), 931–936. https://doi.org/10.1126/science.abg5827.

(37) Fornasier, E.; Macchia, M. L.; Giachin, G.; Sosic, A.; Pavan, M.; Sturlese, M.; Salata, C.; Moro, S.; Gatto, B.; Bellanda, M.; Battistutta, R. A New Inactive Conformation of SARS-CoV-2 Main Protease. *Acta Cryst D* **2022**, *78* (3), 363–378. https://doi.org/10.1107/S2059798322000948.

(38) Zhao, Y.; Zhu, Y.; Liu, X.; Jin, Z.; Duan, Y.; Zhang, Q.; Wu, C.; Feng, L.; Du, X.; Zhao, J.; Shao, M.; Zhang, B.; Yang, X.; Wu, L.; Ji, X.; Guddat, L. W.; Yang, K.; Rao, Z.; Yang, H. Structural Basis for Replicase Polyprotein Cleavage and Substrate Specificity of Main Protease from SARS-CoV-2. *Proceedings of the National Academy of Sciences* **2022**, *119* (16), e2117142119. https://doi.org/10.1073/pnas.2117142119.

(39) Kneller, D. W.; Zhang, Q.; Coates, L.; Louis, J. M.; Kovalevsky, A. Michaelis-like Complex of SARS-CoV-2 Main Protease Visualized by Room-Temperature X-Ray Crystallography. *IUCrJ* **2021**, *8* (6), 973–979. https://doi.org/10.1107/S2052252521010113.

(40) Redfern, O. C.; Dessailly, B.; Orengo, C. A. Exploring the Structure and Function Paradigm. *Current Opinion in Structural Biology* **2008**, *18* (3), 394–402. https://doi.org/10.1016/j.sbi.2008.05.007.

(41) Kustatscher, G.; Collins, T.; Gingras, A.-C.; Guo, T.; Hermjakob, H.; Ideker, T.; Lilley, K. S.; Lundberg, E.; Marcotte, E. M.; Ralser, M.; Rappsilber, J. Understudied Proteins: Opportunities and Challenges for Functional Proteomics. *Nat Methods* **2022**, *19* (7), 774–779. https://doi.org/10.1038/s41592-022-01454-x.

(42) VAZ, F. M.; WANDERS, R. J. A. Carnitine Biosynthesis in Mammals. *Biochemical Journal* **2002**, *361* (3), 417–429. https://doi.org/10.1042/bj3610417.

(43) Edgar, A. J. Mice Have a Transcribed L-Threonine Aldolase/GLY1 Gene, but the Human GLY1 Gene Is a Non-Processed Pseudogene. *BMC Genomics* **2005**, *6* (1), 32. https://doi.org/10.1186/1471-2164-6-32.

(44) Krissinel, E.; Henrick, K. Inference of Macromolecular Assemblies from Crystalline State. *Journal of Molecular Biology* **2007**, *372* (3), 774–797. https://doi.org/10.1016/j.jmb.2007.05.022.

(45) Kabsch, W. XDS. *Acta Cryst D* **2010**, *66* (2), 125–132. https://doi.org/10.1107/S0907444909047337.

(46) McCoy, A. J.; Grosse-Kunstleve, R. W.; Adams, P. D.; Winn, M. D.; Storoni, L. C.; Read, R. J. Phaser Crystallographic Software. *J Appl Cryst* **2007**, *40* (4), 658–674. https://doi.org/10.1107/S0021889807021206.

(47) Liebschner, D.; Afonine, P. V.; Baker, M. L.; Bunkóczi, G.; Chen, V. B.; Croll, T. I.; Hintze, B.; Hung, L.-W.; Jain, S.; McCoy, A. J.; Moriarty, N. W.; Oeffner, R. D.; Poon, B. K.; Prisant, M. G.; Read, R. J.; Richardson, J. S.; Richardson, D. C.; Sammito, M. D.; Sobolev, O. V.; Stockwell, D. H.; Terwilliger, T. C.; Urzhumtsev, A. G.; Videau, L. L.; Williams, C. J.; Adams, P. D. Macromolecular Structure Determination Using X-Rays, Neutrons and Electrons: Recent Developments in Phenix. *Acta Cryst D* **2019**, *75* (10), 861–877. https://doi.org/10.1107/S2059798319011471.

(48) Afonine, P. V.; Grosse-Kunstleve, R. W.; Echols, N.; Headd, J. J.; Moriarty, N. W.; Mustyakimov, M.; Terwilliger, T. C.; Urzhumtsev, A.; Zwart, P. H.; Adams, P. D. Towards Automated Crystallographic Structure Refinement with Phenix.Refine. *Acta Cryst D* **2012**, *68* (4), 352–367. https://doi.org/10.1107/S0907444912001308.

(49) Emsley, P.; Cowtan, K. Coot: Model-Building Tools for Molecular Graphics. *Acta Cryst D* **2004**, *60* (12), 2126–2132. https://doi.org/10.1107/S0907444904019158.

(50) Meng, E. C.; Goddard, T. D.; Pettersen, E. F.; Couch, G. S.; Pearson, Z. J.; Morris, J. H.; Ferrin, T. E. UCSF ChimeraX: Tools for Structure Building and Analysis. *Protein Science* **2023**, *32* (11), e4792. https://doi.org/10.1002/pro.4792.

STRUCTURAL BIOLOGY

ISSN 2059-7983

A new inactive conformation of SARS-CoV-2 main protease

Emanuele Fornasier,[a] Maria Ludovica Macchia,[b] Gabriele Giachin,[a] Alice Sosic,[b] Matteo Pavan,[c] Mattia Sturlese,[c] Cristiano Salata,[d] Stefano Moro,[c] Barbara Gatto,[b] Massimo Bellanda[a,e] and Roberto Battistutta[a,e]*

[a]Department of Chemical Sciences, University of Padua, Via F. Marzolo 1, 35131 Padova, Italy, [b]Department of Pharmaceutical and Pharmacological Sciences, University of Padua, Via F. Marzolo 5, 35131 Padova, Italy, [c]Molecular Modeling Section, Department of Pharmaceutical and Pharmacological Sciences, University of Padua, Via F. Marzolo 5, 35131 Padova, Italy, [d]Department of Molecular Medicine, University of Padua, Via Gabelli 63, 35121 Padova, Italy, and [e]Institute of Biomolecular Chemistry of CNR, Padua Unit, Via F. Marzolo 1, 35131 Padova, Italy. *Correspondence e-mail: roberto.battistutta@unipd.it

Received 17 November 2021

Accepted 27 January 2022

Edited by G. Cingolani, Thomas Jefferson University, USA

Keywords: SARS-CoV-2; main protease; COVID-19; M^{pro}; crystal structure; inactive conformation.

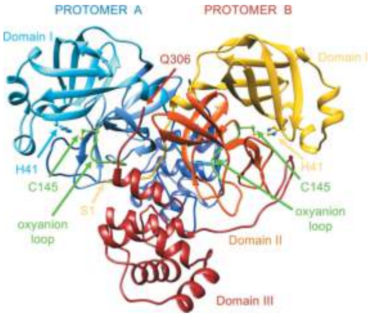
PDB reference: SARS-CoV-2 main protease in a novel conformational state, 7nij

Supporting information: this article has supporting information at journals.iucr.org/d

The SARS-CoV-2 main protease (M^{pro}) has a pivotal role in mediating viral genome replication and transcription of the coronavirus, making it a promising target for drugs against the COVID-19 pandemic. Here, a crystal structure is presented in which M^{pro} adopts an inactive state that has never been observed before, called new-inactive. It is shown that the oxyanion loop, which is involved in substrate recognition and enzymatic activity, adopts a new catalytically incompetent conformation and that many of the key interactions of the active conformation of the enzyme around the active site are lost. Solvation/desolvation energetic contributions play an important role in the transition from the inactive to the active state, with Phe140 moving from an exposed to a buried environment and Asn142 moving from a buried environment to an exposed environment. In new-inactive M^{pro} a new cavity is present near the S2′ subsite, and the N-terminal and C-terminal tails, as well as the dimeric interface, are perturbed, with partial destabilization of the dimeric assembly. This novel conformation is relevant both for comprehension of the mechanism of action of M^{pro} within the catalytic cycle and for the successful structure-based drug design of antiviral drugs.

1. Introduction

To face the global COVID-19 pandemic, besides prevention via the use of vaccines, it is also essential to develop targeted therapeutic options for patients infected by the SARS-CoV-2 betacoronavirus. In general, one of the most promising classes of antiviral drug candidates are protease inhibitors, small molecules that are able to inhibit enzymes involved in virus replication within the cell. Very low sequence identity with human proteases and distinct cleavage-site specificities suggest that viral enzymes can be inhibited with very low associated toxic effects (‘off-target’ effects), if any. Indeed, protease inhibitors have already been efficient in the treatment of viral pathogens such as hepatitis C virus (Pol & Corouge, 2014) and human immunodeficiency virus (HIV; Skwarecki *et al.*, 2021). In coronaviruses, the main protease, M^{pro}, is a cysteine peptidase that is essential for the replication cycle of positive-sense, single-stranded RNA coronaviruses (Xia & Kang, 2011), including SARS-CoV-2. It is also known as 3C-like protease or 3CL^{pro} from the similarity of its active site and its substrate specificity to those of the picornavirus 3C protease (Anand *et al.*, 2002). M^{pro} is involved in the proteolytic processing of the two overlapping polyproteins pp1a and pp1ab, with the formation of individual mature nonstructural



proteins (Snijder *et al.*, 2016), and as such it is a validated antiviral drug target (Dai *et al.*, 2020; Günther *et al.*, 2021; Ullrich & Nitsche, 2020). Currently, there are at least two SARS-CoV-2 M^{pro} inhibitors in phase I clinical trials as candidates with potent antiviral activity: the orally administered PF-07321332 (Pavan *et al.*, 2021) and the intravenously administered PF-00835231 (Ahmad *et al.*, 2021).

SARS-CoV-2 M^{pro} (nsp5), a 306-amino-acid polypeptide of molecular weight 33.8 kDa (Wu *et al.*, 2020), shares 96% sequence identity and a very similar 3D structure with SARS-CoV M^{pro} [0.53 Å r.m.s.d. between PDB entries 6y2e (Zhang *et al.*, 2020) and 2bx4 (Tan *et al.*, 2005)]. Very similar 3D structures have also been found for other coronaviral M^{pro}s such as those from Porcine transmissible gastroenteritis virus (TGEV), which was the first structure of a coronaviral M^{pro} (Anand *et al.*, 2002), Human coronavirus (HCoV) strain 229E (Anand *et al.*, 2003), Infectious bronchitis virus (IBV; Xue *et al.*, 2008) and MERS-CoV (Ho *et al.*, 2015). This structural similarity, which is particularly relevant around the active site, leads to the possibility of the development of pan-coronaviral drugs.

M^{pro} exists in an equilibrium between a monomer and a homodimer (with the two protomers roughly perpendicularly oriented; Fig. 1*a*), with an apparent K_d of between 0.8 and 14 µ*M* for the SARS-CoV enzyme, depending on the experimental conditions (Chen *et al.*, 2006). For SARS-CoV-2 M^{pro}, the K_d has been estimated to be 2.5 µ*M* by analytical ultracentrifugation (Zhang *et al.*, 2020) and 0.14 µ*M* by native mass spectrometry (El-Baba *et al.*, 2020). Unlike 3C protease, only the SARS-CoV M^{pro} dimer shows enzymatic activity (Anand *et al.*, 2002) and the correct shape of the substrate-binding site, particularly of the S1 subsite; the correct conformation for productive catalytic events is linked to the dimerization process. It has been proposed that the dimerization process has a direct regulatory role of the activity of M^{pro} during the coronaviral replication process (Hsu *et al.*, 2005; Li *et al.*, 2016). Given the high structural similarity, particularly at the dimeric interface, it was reasoned that dimerization of the enzyme is also necessary for the catalytic activity of SARS-CoV-2 M^{pro} (Zhang *et al.*, 2020),

Each M^{pro} protomer is composed of three structural domains (Fig. 1*a*; Anand *et al.*, 2002). The chymotrypsin-like and 3C protease-like β-barrel domains I (residues 1–99) and II (residues 100–182) directly control the catalytic event. The substrate-binding site is between these two domains and comprises several subsites for substrate binding (from S1 to S6 and from S1′ to S3′), corresponding to the P1–P6 and P1′–P3′ amino-acid positions of the substrates (according to the convention P6–P5–P4–P3–P2–P1↓P1′–P2′–P3′, where ↓ indicates the hydrolyzed peptide bond; Anand *et al.*, 2003). Enzymatic proteolysis by SARS-CoV-2 M^{pro} at the 11 cleavage sites on the viral polyprotein occurs on the C-terminal side of a conserved glutamine in position P1, with the most common consensus sequence being Leu-Gln↓(Ser/Ala), indicating that specificity is determined mostly by the P2, P1 and P1′ positions (Ullrich & Nitsche, 2020). Glutamine in position P1 is fully conserved not only for SARS-CoV-2 but also in substrates of SARS-CoV and MERS-CoV. Prime recognition sites at the C-terminus of P1′ are not conserved. M^{pro} subsites S4, S2, S1 and S1′ have been identified as the most relevant

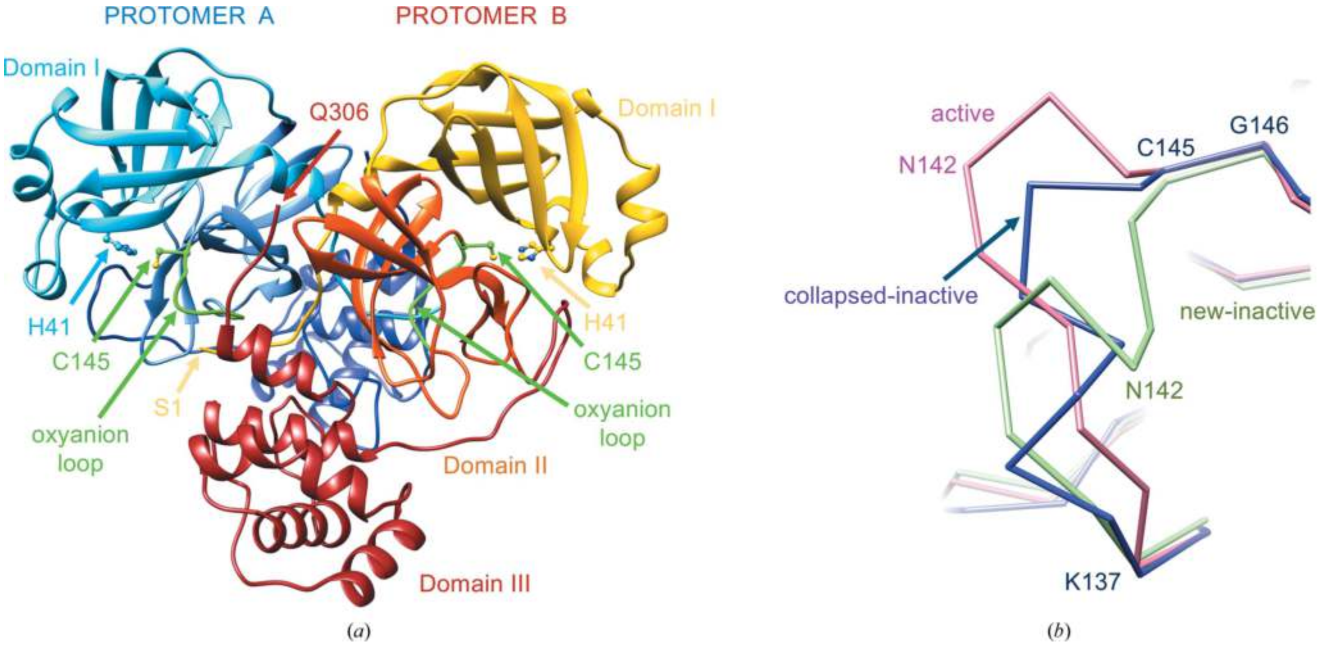


Figure 1
SARS-CoV-2 M^{pro} architecture, free form (PDB entry 6y2e). (*a*) Dimeric assembly of the protease with the main structural features discussed in the text highlighted. Protomer *A* is in blue-based colors and protomer *B* is in yellow/red-based colors. The two oxyanion loops and the two catalytic cysteines 145 are shown in green. (*b*) Comparison between different oxyanion-loop conformations of M^{pro}: active in SARS-CoV-2 M^{pro} (PDB entry 6y2e) in pink, collapsed-inactive in SARS-CoV M^{pro} (PDB entry 1uj1 chain *B*) in magenta and new-inactive in SARS-CoV-2 M^{pro} (this work) in green.

subsites for substrate binding, with regions in the S5, S4 and S2 sites showing considerable conformational flexibility upon binding different chemical groups (Kneller, Galanie *et al.*, 2020). The chymotrypsin-like fold, including domains I and II, is connected by a 16-residue flexible loop to the extra α-helical domain III (residues 198–306; Fig. 1*a*). Domain III is absent in other RNA virus 3C-like proteases and plays a key role in enzyme dimerization and activity regulation of M^{pro} (Anand *et al.*, 2002; Shi & Song, 2006).

At variance with the classical catalytic triad of chymotrypsin-like proteases, coronaviral M^{pro} has a catalytic dyad, consisting of His41 and Cys145 in SARS-CoV-2 (Fig. 1*a*); a conserved water molecule occupies a position analogous to that of the side chain of the third member of the catalytic triad (for instance, aspartate in chymotrypsin and asparagine in papain) and forms hydrogen bonds to the side chains of His41, His164 and Asp187. It has been proposed that this conserved water is involved in the catalytic event (Anand *et al.*, 2002).

A key role in the proper function of the enzyme is also played by the N-finger (residues 1–7) as the N-terminal tail of one protomer interacts and stabilizes the binding site (S1 subsite) of the other protomer (Verschueren *et al.*, 2008). Indeed, deletion of the N-finger hampers dimerization in solution and abolishes the proteolytic activity. Both the N-finger and the C-terminus are results of the autoproteolytic processing of M^{pro}. Accordingly, in the mature dimeric enzyme both termini of one protomer face the active site of the other.

The important conserved residues Phe140, Leu141, Asn142 and Ser144 (SARS-CoV-2 numbering) are part of a structural element that is essential for a productive catalytic event, the so-called oxyanion loop comprising residues 138–145, which globally lines the binding site for glutamine P1. The central role of the oxyanion loop in the catalytic reaction mechanism of serine proteases and cysteine proteases has been extensively characterized (Frey & Hegeman, 2007). The correct positioning of the oxyanion hole, which is part of the oxyanion loop (formed by the backbone of Gly143, Ser144 and Cys145 in SARS-CoV-2 M^{pro}), is essential for stabilization of the transient tetrahedral acyl (oxyanion) transition state via the hydrogen-bond donor properties of the amides (Anand *et al.*, 2002; Lee *et al.*, 2020; Verschueren *et al.*, 2008). In the known crystal structures of SARS-CoV and SARS-CoV-2 M^{pro}, the oxyanion loop adopts essentially the same ‘active’ conformation; here, we take PDB entry 6y2e as a reference for this conformation (Douangamath *et al.*, 2020; Jin, Du *et al.*, 2020; Jin, Zhao *et al.*, 2020; Zhang *et al.*, 2020). A specific conformation is defined to be active when the amino acids known to participate in the chemical reaction catalyzed by the enzyme are properly positioned and oriented for the reaction to proceed. We also term this conformation catalytically competent.

Variations from the active conformation of the oxyanion loop are found in a few forms of the enzyme, which were consequently considered to be inactive or catalytically incompetent, as in protomer *B* of SARS-CoV M^{pro} (PDB entries 1uj1 and 1uk2; Yang *et al.*, 2003), in the monomeric R298A mutant of SARS-CoV M^{pro} (PDB entry 2qcy; Shi *et al.*, 2008) and in the C172A mutant of $3C^{pro}$ from the picornavirus hepatitis A virus (Allaire *et al.*, 1994), as well as in IBV $3CL^{pro}$ (PDB entries 2q6f and 2q6d; Xue *et al.*, 2008). In the inactive monomeric R298A mutant (PDB entry 2qcy), the region of the oxyanion loop, Ser139-Phe140-Leu141, is converted into a short 3_{10}-helix. In PDB entry 1uj1 (SARS-CoV M^{pro} crystallized at pH 6) the oxyanion loop of one of the two protomers exists in a ‘collapsed’ conformation (similar to that found in PDB entry 2qcy), which is considered to be catalytically incompetent, in which the hydrogen bond between Glu166 and His172 that is important for activity is broken (Yang *et al.*, 2003). In the following, we will refer to these two inactive conformations with similar oxyanion-loop conformations as collapsed-inactive (Fig. 1*b*).

In the vast majority of SARS-CoV and SARS-CoV-2 M^{pro} crystal structures, the dimer is crystallographic (Jaskolski *et al.*, 2021); that is, there is only one molecule in the asymmetric unit and therefore the two protomers are perfectly identical. In the very few inactive structures, apart from the artificially induced monomeric forms, the dimer is formed by two different molecules present in the asymmetric unit, one of which is in the inactive state and the other of which is in the active state. Based on molecular-dynamics simulations coupled to activity data in solution, it was suggested that only one protomer at a time is active in the dimer (Chen *et al.*, 2006).

Here, we describe a new inactive structure (called new-inactive) of the main protease of SARS-CoV-2 that is clearly distinct from both the active and the known collapsed-inactive structures, with an oxyanion-loop conformation that is very different from those previously described (Fig. 1*b*). In Section 4, we argue that this conformation has an important functional role as part of the catalytic cycle of coronaviral M^{pro}.

2. Materials and methods

2.1. Recombinant protein production and purification

The plasmid PGEX-6p-1 encoding SARS-CoV-2 M^{pro} (Zhang *et al.*, 2020) was a generous gift from Professor Rolf Hilgenfeld, University of Lübeck, Lübeck, Germany. Recombinant protein production and purification were adapted from Zhang *et al.* (2020) (where the structure of M^{pro} in the active form was presented; PDB entry 6y2e). The expression plasmid was transformed into *Escherichia coli* strain BL21 (DE3) and then precultured in YT medium at 37°C (100 µg ml^{-1} ampicillin) overnight. The preculture was used to inoculate fresh YT medium supplemented with antibiotic and the cells were grown at 37°C to an OD_{600} of 0.6–0.8 before induction with 0.5 m*M* isopropyl β-D-1-thiogalactopyranoside (IPTG). After 5 h at 37°C, the cells were harvested by centrifugation (5000*g*, 4°C, 15 min) and frozen. The pellets were resuspended in buffer *A* (20 m*M* Tris, 150 m*M* NaCl pH 7.8) supplemented with lysozyme, DNase I and PMSF for lysis. The lysate was clarified by centrifugation at 12 000*g* at 4°C for 1 h and loaded onto a HisTrap HP column (GE Healthcare) equilibrated with 98% buffer *A*/2% buffer *B* (20 m*M* Tris, 150 mM NaCl, 500 mM imidazole pH 7.8). The column was washed with 95% buffer A/5% buffer B, and His-tagged M$^{\rm pro}$ was then eluted with a linear gradient of imidazole from 25 to 500 mM. Pooled fractions containing the target protein were subjected to buffer exchange with buffer A using a HiPrep 26/10 desalting column (GE Healthcare). Next, PreScission protease was added to remove the C-terminal His tag (20 µg of PreScission protease per milligram of target protein) at 12°C overnight. The protein solution was loaded onto a HisTrap HP column connected to a GSTrap FF column (GE Healthcare) equilibrated in buffer A to remove the GST-tagged PreScission protease, the His tag and the uncleaved protein. M$^{\rm pro}$ was finally purified using a Superdex 75 prep-grade 16/60 SEC column (GE Healthcare) equilibrated with buffer C (20 mM Tris, 150 mM NaCl, 1 mM EDTA, 1 mM DTT pH 7.8). Fractions containing the target protein with high purity were pooled, concentrated to 25 mg ml^{-1} and flash-frozen in liquid nitrogen for storage in small aliquots at −80°C.

2.2. Protein characterization and enzymatic kinetics

The correctness of the M$^{\rm pro}$ DNA sequence was verified by sequencing the expression plasmid. The molecular mass was determined as follows: recombinant SARS-CoV-2 M$^{\rm pro}$, diluted in 50% acetonitrile with 0.1% formic acid, was analyzed by direct infusion electrospray ionization (ESI) on a Xevo G2-XS QTOF mass spectrometer (Waters). The detected species displayed a mass of 33 796.64 Da, which very closely matches the value of 33 796.81 Da calculated from the theoretical full-length protein sequence (residues 1–306). A representative ESI-MS spectrum is shown in Supplementary Fig. S1. To characterize the enzymatic activity of our recombinant M$^{\rm pro}$, we adopted a FRET-based assay using the substrate 5-FAM-AVLQ↓SGFRK(DABCYL)K (Proteogenix). The assay was performed by mixing 0.05 µM M$^{\rm pro}$ with various concentrations of substrate (1–128 µM) in a buffer composed of 20 mM Tris, 100 mM NaCl, 1 mM EDTA, 1 mM DTT pH 7.3. Fluorescence intensity (excitation at 485 nm and emission at 535 nm) was monitored at 37°C with a VictorIII microplate reader (Perkin Elmer). A calibration curve was created by measuring multiple concentrations (from 0.001 to 5 µM) of free fluorescein in a final volume of 100 µl reaction buffer. Initial velocities were determined from the linear section of the curve, and the corresponding relative fluorescence units per time unit (ΔRFU s^{-1}) were converted to the amount of cleaved substrate per time unit (µM s^{-1}) by fitting to the calibration curve of free fluorescein. The catalytic efficiency $k_{\rm cat}/K_{\rm m}$ was 4819 ± 399 s^{-1} M^{-1}, which is in line with literature data (Ma *et al.*, 2020; Zhang *et al.*, 2020).

2.3. Crystallization and data collection

A frozen aliquot of M$^{\rm pro}$ was thawed in ice, diluted in a 1:2 ratio with buffer C (20 mM Tris, 150 mM NaCl, 1 mM EDTA, 1 mM DTT pH 7.8) to a final concentration of 12.5 mg ml^{-1} and clarified by centrifugation at 16 000*g*. The inhibitors masitinib, manidipine, bedaquiline and boceprevir were dissolved in 100% DMSO to a concentration of 100 mM. The protein was crystallized both in the free form and in the presence of inhibitors by co-crystallization. In all cases, final crystal growth was obtained by microseeding starting from small crystals of the free enzyme. The protein in the free form was crystallized using the sitting-drop vapor-diffusion method at 18°C, mixing 1.0 µl M$^{\rm pro}$ solution with 1.0 µl precipitant solution [0.1 M MMT (DL-malic acid, MES and Tris base in a 1:2:2 molar ratio) pH 7.0, 25% PEG 1500] and 0.2 µl seed stock (diluted 1:500, 1:1000 or 1:2000 with precipitant solution) and equilibrating against a 300 µl reservoir of precipitant solution. Crystals appeared overnight and grew for 48 h after the crystallization drops had been prepared. In the case of co-crystallization, M$^{\rm pro}$ was incubated for 16 h at 8°C with a 13-fold molar excess of inhibitor (final DMSO concentration 5%). After incubation with masitinib, manidipine or bedaquiline, a white precipitate appeared and the solutions were clarified by centrifugation at 16 000*g*; as the protein concentration was essentially unchanged after centrifugation, we concluded that the precipitate is composed of the inhibitors, which are poorly soluble in water. The fact that the protein was later crystallized under the same conditions as described for the free form further confirmed that its concentration was not altered by the centrifugation process. For data collections, crystals were fished from the drops, cryoprotected by a quick dip into 30% PEG 400 (with 5 mM inhibitor in the case of co-crystals) and flash-cooled in liquid nitrogen. The crystals were monoclinic (space group C2), isomorphous to the crystals of the free enzyme (PDB entry 6y2e), with one monomer in the asymmetric unit; the dimer is formed by the crystallographic twofold axis.

2.4. Structure determination, refinement and analysis

Data were collected on beamlines ID23-2 and ID23-1 at the ESRF. Diffraction data integration and scaling were performed with *XDS* (Kabsch, 2010) and data reduction and analysis were performed with *AIMLESS* (Evans & Murshudov, 2013). Initially, structures were solved by molecular replacement (MR) with *Phaser* (McCoy *et al.*, 2007) in *Phenix* (Liebschner *et al.*, 2019) using PDB entries 6y2e and 5rel (M$^{\rm pro}$ in complex with PCM-0102340; Douangamath *et al.*, 2020) as search models. To limit MR model bias in critical zones (namely residues 139–144, 1–3 and the side chain of His163) we then performed new MR runs using PDB entry 6y2e without residues 139–144 and 1–3, and with an alanine instead of a histidine at position 163, as the search model. Only for co-crystallization experiments with boceprevir was electron density for the ligand clearly visible from the beginning of the refinement (Supplementary Figs. S2 and S3), and the three final structures, modeled from residues 1 to 306 (compared with the 'new' structure modeled to residue 301), are virtually identical to those deposited in the PDB (Fu *et al.*, 2020). In all of the other cases, no electron density indicating the presence of the inhibitors masitinib, manidipine or bedaquiline in the active site (or elsewhere) was detectable. For four structures, it was possible to efficiently model residues 139–144, 1–3 and the side chain of His163 in 'new' conformations. The final struc-

Table 1
X-ray diffraction data-processing and model-refinement statistics.

Values in parentheses are for the highest resolution shell.

Data collection	
X-ray source	ID23-2, ESRF
Wavelength (Å)	0.873130
Space group	*C*2
a, *b*, *c* (Å)	113.07, 54.71, 44.84
α, β, γ (°)	90.00, 101.30, 90.00
Resolution range (Å)	55.44–1.58 (1.61–1.58)
$R_{\rm merge}$	0.070 (1.305)
$R_{\rm meas}$	0.081 (1.505)
$R_{\rm p.i.m.}$	0.040 (0.739)
Total No. of observations	145297 (7276)
No. of unique observations	36653 (1847)
Mean $I/\sigma(I)$	9.2 (1.0)
$CC_{1/2}$ (%)	99.8 (35.7)
Completeness (%)	99.4 (99.5)
Multiplicity	4.0 (3.9)
Wilson *B* estimate (Å^2)	23.7
Refinement	
Resolution range (Å)	55.44–1.58
$R_{\rm work}/R_{\rm free}$ (%)	17.71/20.31
No. of atoms	
Protein	2350
Water	218
B factors (Å^2)	
Protein	32.6
Water	43.1
R.m.s.d.	
Bond lengths (Å)	0.008
Bond angles (°)	0.868
Coordinate error (maximum-likelihood-based by *Phenix*) (Å)	0.21
Ramachandran statistics	
Favored (%)	97.99
Allowed (%)	2.01
Outliers (%)	0.00
PDB code	7nij
Ensemble refinement	
No. of models	60
$R_{\rm work}/R_{\rm free}$ (%)	15.47/20.80

tures were obtained by alternating cycles of manual refinement with *Coot* (Emsley *et al.*, 2010) and automatic refinement with *phenix.refine* (Afonine *et al.*, 2012). At the end, the model was submitted to *phenix.ensemble_refinement* (Burnley *et al.*, 2012) with default parameters. Data-collection and refinement statistics for the structure obtained by a co-crystallization experiment with masitinib (which was not visible in the final electron density) are reported in Table 1. Secondary-structure analysis was performed with *DSSP* (Kabsch & Sander, 1983; Touw *et al.*, 2015). Local energetic frustration analysis was performed with the *Frustratometer* server (http://frustratometer.qb.fcen.uba.ar; Parra *et al.*, 2016). Interface analysis was performed using *PISA* (Krissinel & Henrick, 2007).

2.5. Molecular modeling

The majority of the computational work was performed on a Linux desktop workstation (Intel Xeon CPU E5-1620 3.60 GHz) running Ubuntu 16.04 LTS. Molecular-dynamics trajectories were collected on a heterogeneous Nvidia GPU cluster composed of 20 GPUs with models spanning from GTX1080 to RTX2080Ti. For structure preparation, coordinates of the active conformation of SARS-CoV-2 M$^{\rm pro}$ were retrieved from the Protein Data Bank (PDB entry 6y2e). Coordinates for both the active and the new-inactive conformation were processed with the aid of the *Molecular Operating Environment* (*MOE*) 2019.01 (Chemical Computing Group) structure-preparation tool. Initially, the functional unit of the protease (the dimeric form) was restored by applying a symmetric crystallographic transformation to each asymmetric unit. Residues with alternate conformations were assigned to the highest occupancy alternative. Moreover, missing residues that are present in the primary sequence were added using the *MOE* Loop Modeler tool. The *MOE* Protonate3D tool was used to assign the most probable protonation state to each residue (pH 7.4, T = 310 K, i.f. = 0.154). Partial charges were then assigned using the AMBER10 force field and H atoms were energy-minimized until the gradient was below 0.1 kcal mol^{-1} Å^{-2}. Finally, ions and all co-crystallized molecules except for water were removed before saving the structures. The system setup for the MD simulations was carried out using the *antechamber*, *parmchk* and *tleap* software implemented in the *AmberTools*14 suite (Case *et al.*, 2005). AMBER ff14SB (Maier *et al.*, 2015) was adopted for system parametrization and attribution of partial charges. Protein structures were explicitly solvated in a rectangular prismatic TIP3P (Jorgensen *et al.*, 1983) periodic water box with borders placed at a distance of 15 Å from any protein atom. Na$^+$ and Cl$^-$ ions were added to neutralize the system until a salt concentration of 0.154 *M* was reached. MD simulations were then performed using *ACEMD*3 (Harvey *et al.*, 2009), which is based upon an OpenMM 7.4.2 engine (Eastman *et al.*, 2017). Initially, 1000 steps of energy minimization were executed using the conjugate-gradient algorithm. A two-step equilibration procedure was then carried out: the first step consisted of a 1 ns canonical ensemble (NVT) simulation with 5 kcal mol^{-1} Å^{-2} harmonic positional constraints applied to each protein atom, while the second step consisted of a 1 ns isothermal–isobaric (NPT) simulation with 5 kcal mol^{-1} Å^{-2} harmonic positional constraints applied only to protein C$^\alpha$ atoms. The production phase consisted of three independent MD replicas for each protein conformation. Each simulation had a duration of 1 µs and was performed using the NVT ensemble at a constant temperature of 310 K with a timestep of 2 fs. For both the equilibration and the production stage, the temperature was maintained constant using a Langevin thermostat. During the second step of the equilibration stage, the pressure was maintained at a fixed value of 1 atm with a Monte Carlo barostat. MD trajectories were aligned using protein C$^\alpha$ atoms from the first trajectory frame as a reference, wrapped into an image of the system under periodic boundary conditions (PBC), and subsequently saved using a 200 ps interval between each frame and removing any ions and water molecules using *Visual Molecular Dynamics* 1.9.2 (*VMD*; Humphrey *et al.*, 1996). The protein radius of gyration ($R_{\rm g}$), the root-mean-square deviation (r.m.s.d.) and the root-mean-square fluctuation (r.m.s.f.) of atomic positions along the trajectory were calculated for protein C$^\alpha$ atoms exploiting the

MDAnalysis (Gowers *et al.*, 2016; Michaud-Agrawal *et al.*, 2011) Python module. Secondary-structure analysis was carried out with the *STRIDE* package (Frishman & Argos, 1995) as implemented in *VMD* 1.9.2. The collected data were then plotted using the Matplotlib Python library (Hunter, 2007).

Furthermore, two classic MD simulations were performed on the complexes obtained by superposing the coordinates of peptide ligands from PDB entries 2q6g and 7khp on the new-inactive conformation of SARS-CoV-2 M^{Pro} using *MOE* 2019.01. For each peptide–ligand complex, a two-stage equilibration protocol followed by a single productive simulation was carried out. The first equilibration step consisted of a 0.1 ns canonical ensemble (NVT) simulation with 5 kcal mol^{-1} Å^{-2} harmonic positional constraints applied to each protein atom, while the second equilibration step consisted of a 0.5 ns isothermal–isobaric (NPT) simulation with 5 kcal mol^{-1} Å^{-2} harmonic positional constraints applied only to protein C$^{\alpha}$ atoms. For both equilibration simulations, the temperature was maintained constant (T = 310 K) using a Langevin thermostat, while during the second equilibration stage the pressure was kept at a constant value of 1 atm using a Monte Carlo barostat. The productive simulation was carried out for 10 ns in the NVT ensemble (T = 310 K).

3. Results

3.1. Identification of a new-inactive conformation of M^{pro}

In a campaign to obtain structural insights into SARS-CoV-2 M^{pro}, we analyzed 27 different data sets to determine crystal structures of M^{pro} in complex with different inhibitors, among which were masitinib, manidipine and bedaquiline (Ghahremanpour *et al.*, 2020). As 'positive' controls (*i.e.* structures that were already known), we considered ligand-free M^{pro} and M^{pro} in complex with the known α-ketoamide covalent reversible inhibitor boceprevir, an approved HCV drug that is also able to bind to SARS-CoV-2 M^{pro} (Fu *et al.*, 2020). M^{pro} samples were produced and crystallized in parallel, with very similar experimental procedures, analogous to those of the active enzyme (PDB entry 6y2e; Zhang *et al.*, 2020; see Section 2). Almost all tested crystals were monoclinic (space group $C2$, with unit-cell parameters $a \simeq 113.1$, $b \simeq 54.7$, $c \simeq 44.8$ Å, $\alpha = 90.0$, $\beta \simeq 101.3$, $\gamma = 90.0°$), isomorphous to the crystals of the free active enzyme (PDB entry 6y2e; Zhang *et al.*, 2020) and to most of the deposited M^{pro} structures, signifying the same crystal contacts. After successful molecular replacement and a first round of refinement, in most cases (including the complex with boceprevir) electron density was clearly visible for the entire sequence, indicating a protein matrix with a very similar structure to the search models (PDB entries 6y2e and 5rel; Douangamath *et al.*, 2020). However, there were a significant number of cases, around ten, in which the electron density was of much lower quality or was even absent in particular portions of the protein, namely residues 139–144 of the oxyanion loop, residues 1–3 of the N-finger and the side chain of His163 in the S1 specificity subsite, all of which are residues that are part of the active site. To cope with the known molecular-replacement bias problem and to correctly rebuild the ambiguous parts, we performed new MR runs using PDB entry 6y2e deprived of residues 139–144 and 1–3, and with an alanine instead of a histidine at position 163 (to remove the His side chain), as a search model. This allowed us to confirm perturbations in the conformation of the selected areas for ten structures, while clear electron density was visible for the remaining cases with the oxyanion loop unambiguously in the active conformation (Supplementary Figs. S2 and S3). In some cases, the electron density was so poor that the tracing of the chain was very problematic, and it was not possible to reliably rebuild the mobile zones entirely (Supplementary Fig. S2*b*). For four structures, it was possible to efficiently model residues 139–144, residues 1–3 and the side chain of His163 in 'new' conformations ('new' because there are no equivalents in M^{pro} structures deposited in the PDB) that differ from the active conformations and also from the collapsed-inactive conformations, including PDB entry 2qcy, where the oxyanion loop adopts a 3_{10}-helix conformation (Supplementary Fig. S2*c*). In this regard, comprehensive analyses of the available SARS-CoV and SARS-CoV-2 M^{pro} crystal structures have recently appeared in the literature (Behnam, 2021; Brzezinski *et al.*, 2021; Jaskolski *et al.*, 2021; Wlodawer *et al.*, 2020). In no case was a conformation analogous to that presented here described, confirming our assessment of a new-inactive state. The most relevant structures discussed here are reported in Supplementary Table S1.

In summary, we found three different conformational states for the oxyanion loop: active (Supplementary Fig. S2*a*), flexible (*i.e.* with poor electron density; Supplementary Fig. S2*b*) and, strikingly, a new-inactive state (Supplementary Fig. S2*c*). A comparison of the known active and collapsed-inactive conformations with the new-inactive conformation presented here is shown in Fig. 1(*b*).

The new-inactive structures were derived solely from crystals obtained using M^{pro} pre-incubated with the inhibitors masitinib, manidipine or bedaquiline, but in no case was electron density indicating the presence of the inhibitors detected. This is explainable by the medium/high IC$_{50}$ (in the range 2.5–19 µM; Drayman *et al.*, 2021; Ghahremanpour *et al.*, 2020) and the very low aqueous solubility of the molecules (when inhibitors in 100% DMSO were added to the protein solution, visible white precipitates appeared). It is tempting to speculate that the presence of these inhibitors in solution plays a role in favoring the selection of the new-inactive conformation by the crystallization process. Some structures of crystals from co-crystallization experiments with masitinib or manidipine, again without any evidence for the presence of the ligand in the binding site, show the oxyanion active conformation. This indicates that these molecules, although favoring the new state, are not strict determinants for its formation. In the free form of the enzyme (from crystallization experiments with no ligands), we obtained structures with very clear electron density for the oxyanion loop, as shown in Supplementary Fig. S2(*a*), with low local B factors in the refined model, but also structures with a very 'destabilized', mobile oxyanion loop, as in Supplementary Fig. S2(*b*), with much higher *B* factors in the final model. This suggests that the high flexibility of the oxyanion loop is an intrinsic property of the free enzyme and is not artificially induced by the presence of ligands in the crystallization experiments.

Here, we describe only one of the structures of M^{pro} determined in the new-inactive conformation, which was obtained by co-crystallization experiments with masitinib (no relevant differences exist among the four new-inactive M^{pro} structures). Data-collection and final model statistics are reported in Table 1; final electron densities for the most relevant regions discussed in the text are shown in Fig. 2. Unlike in other inactive structures of the enzyme, in which only one protomer adopts the inactive conformation, the dimeric arrangement of the new structure is due to a crystallographic symmetric axis, and the two subunits are therefore identical and both inactive.

3.2. The oxyanion loop adopts a novel inactive conformation

The most striking property of the new structure is the significantly different conformational state of the oxyanion loop (Figs. 1 and 3), which is essential for stabilization of the tetrahedral acyl (oxyanion) transition state during the catalytic cycle. The loop backbone is stabilized by many hydrogen bonds in the new state (Fig. 3*a*). According to the *DSSP* standardized secondary-structure assignment (Kabsch & Sander, 1983; Touw *et al.*, 2015), in the new oxyanion loop

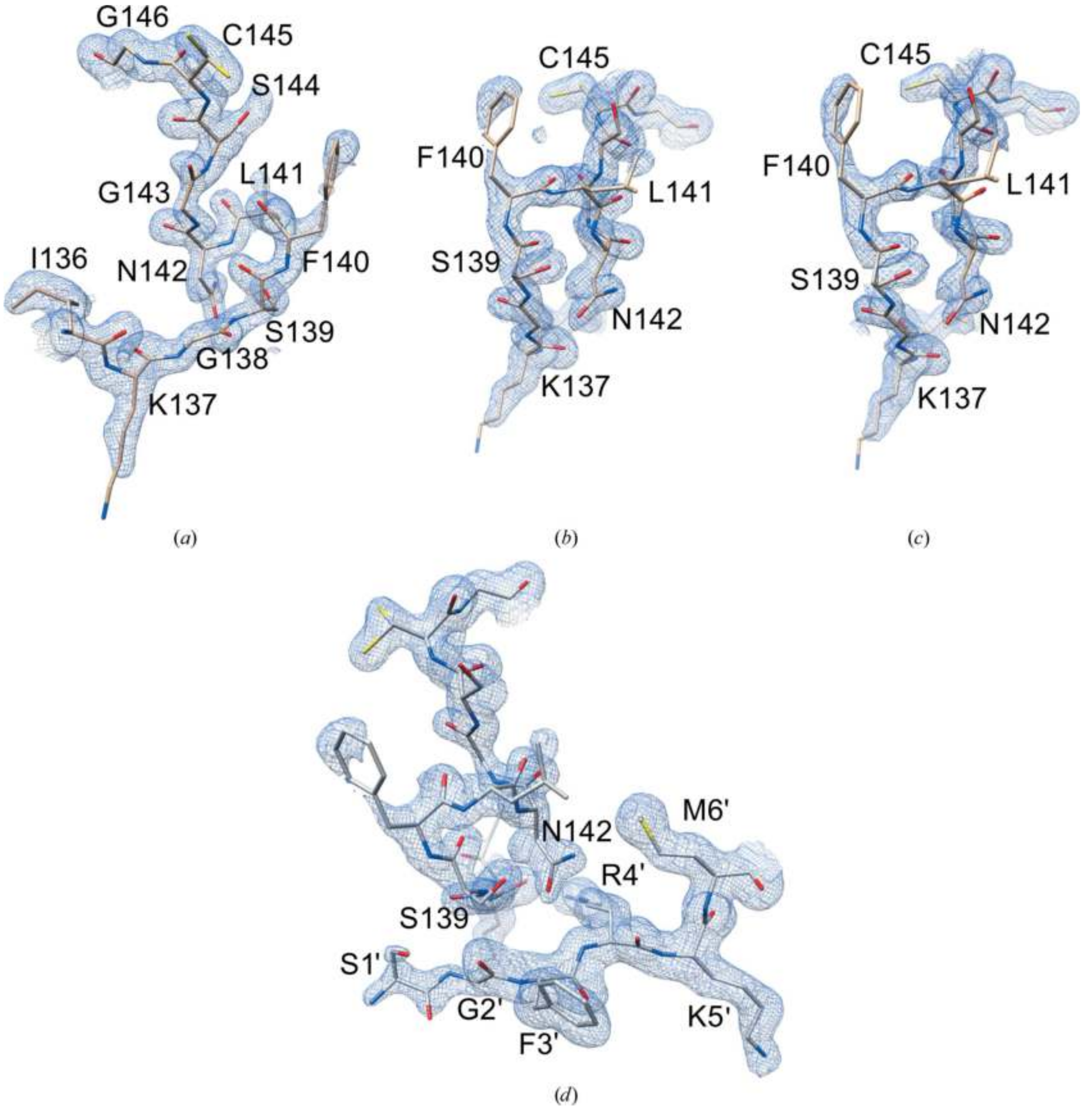


Figure 2
Final electron densities for the most relevant regions of new-inactive M^{pro}. $2F_o - F_c$ maps contoured at the 1.0σ level are shown. (*a*) and (*b*) show two views of the final electron density for the oxyanion loop in the new conformation. Leu141 and the solvent-exposed Phe140 and Lys137 side chains have incomplete densities indicating various degrees of flexibility. (*c*) Simulated-annealing omit map (oxyanion-loop residues 138–146 were omitted) viewed as in (*b*). (*d*) Electron density in the inter-protomer (intra-dimer) interaction area between the oxyanion loop of one protomer and the N-finger of the other protomer (residues Ser1′–Met6′).

there are two consecutive ‘3-turns’ (β-turns) with hydrogen bonds between Leu141 CO and Ser144 NH and between Ser144 CO and Ser147 NH. This region is further stabilized by a ‘4-turn’ (α-turn) with a hydrogen bond between Ser139 CO and Gly143 NH. *DSSP* does not recognize any 3_{10}-helical segments in the oxyanion loop (as present in the inactive PDB entry 2qcy).

There are other hydrogen bonds involving the backbone that stiffen the oxyanion loop: between Cys145 CO and Asn28 NH, between His163 CO and Gly146 NH and between Ser147 CO and His163 NH (Fig. 3*a*). As a result, the new conformation appears to be quite stable and rigid, as confirmed by the good quality of the local electron density (Fig. 2 and Supplementary Fig. S2*c*).

To analyze the energetics of the local contacts, we performed an energetic frustration analysis (Parra *et al.*, 2016) on the active and new-inactive conformations. The concept of local frustration in protein structure refers to possible residual energetic conflicts in local interactions in folded proteins, using a ‘frustration index’ that measures how favorable a particular contact is relative to the set of all possible contacts in that location (Chen *et al.*, 2020). The ‘principle of minimal frustration’ assumes that proteins find their native state by minimizing the internal energetic conflicts within their polypeptide chain (Bryngelson & Wolynes, 1987). The degree of frustration is therefore dependent on the type of amino acids involved in the interaction. Local violations of this principle have been recognized to be important to exert the proper biological functions, specifically around the active sites of protein enzymes (Freiberger *et al.*, 2019). Analysis of the local configurational frustration of the most interesting contacts around the active site of active and new-inactive M$^{\rm pro}$ is shown in Supplementary Table S2. In both conformations, the catalytic Cys145 is a minimally frustrated ‘hub’ (here we call a position with $\geq$10 minimally frustrated interactions a minimally frustrated hub), with a small prevalence of interactions in the active conformation. On the other hand, the difference for Phe140 is striking: eight minimally frustrated interactions are present in active M$^{\rm pro}$ (where it is buried in a hydrophobic pocket) as opposed to no interactions in new-inactive M$^{\rm pro}$ (where it is solvent-exposed). Differences between the two structures are also evident for other amino acids of the oxyanion loop, namely Leu141, Gly143 and Ser144, indicating their diverse involvement in the local energetic contributions. The oxyanion loop of inactive M$^{\rm pro}$ has a larger number of minimally frustrated interactions with Cys117. This residue is a minimally frustrated hub in both conformations; however, given the higher number of minimally frustrated interactions in new-inactive M$^{\rm pro}$ (18 versus ten), Cys117 seems to play an important role in the stabilization of the new-inactive conformation. Internal to the oxyanion loop there is also a highly frustrated (unfavorable) interaction involving Leu141, with Ser139 in new-inactive M$^{\rm pro}$ and with Ser144 in active M$^{\rm pro}$. This suggests that Leu141 may be important in switching between the two conformations.

3.3. Many key interactions of the active enzyme are lost in new-inactive M$^{\rm pro}$

The correct location of Phe140, Leu141, Asn142, Ser144, Tyr161, His163, Met165, Glu166 and His172 (as seen in the active PDB entry 6y2e, for instance) is an absolute requirement for the reaction catalyzed by M$^{\rm pro}$ to properly proceed, with special reference to stabilization of the tetrahedral acyl-intermediate (Anand *et al.*, 2002; Lee *et al.*, 2020; Verschueren *et al.*, 2008). Notably, all of these residues are conserved among known coronaviral M$^{\rm pro}$s, underlining their importance. In the new structure of M$^{\rm pro}$ most of these residues move away from the ‘active location’: Phe140, Leu141, Asn142 and Ser144 because of displacement of the oxyanion loop (Fig. 3*b*) and His163 and His172 because of rotation of their side chains (Fig. 4).

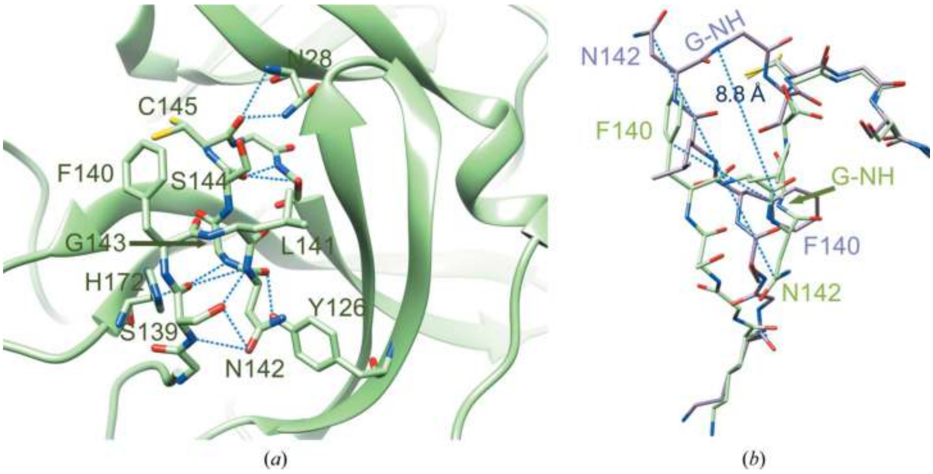


Figure 3
Details of the hydrogen-bond interactions in the oxyanion region of new-inactive M$^{\rm pro}$. (*a*) The new conformation of the oxyanion loop is stabilized by several backbone hydrogen bonds (blue dashed lines) as described in the main text. The side chain of catalytic Cys145 has a double conformation. (*b*) Comparison between the new-inactive (green) and active (light magenta; PDB entry 6y2e) oxyanion loops. There are large movements (blue dashed lines) of the side chains of Asn142 and Phe140. In the new-inactive conformation, Asn142 moves from an exposed position with an ASA of 153.74 Å^2 to a buried position with an ASA of 49.00 Å^2 and Phe140 moves from a buried position with an ASA of 14.79 Å^2 to an exposed position with an ASA of 143.29 Å^2. Gly143 NH (G-NH) of the oxyanion hole, which is involved in the stabilization of the tetrahedral intermediate, moves 8.8 Å away.

Specifically, Asn142 C$^{\alpha}$ and the side chain of Phe140 are remarkably shifted from the active position by 9.8 and 7.5 Å, respectively (Fig. 3*b*). Phe140, which is buried in a hydrophobic cleft in active M$^{\rm pro}$ with as accessible surface area (ASA) of 14.79 Å^2), is now exposed to the solvent (ASA 143.29 Å^2), while

Asn142, which is exposed in active M^{pro} (ASA 153.74 Å^2), is now buried (ASA 49.00 Å^2). The side chain of Asn142 is locked in the new position by hydrogen bonds to the side-chain O$^{\gamma}$ and backbone NH of Ser139. Markedly, the oxyanion hole Gly143 NH, the correct positioning of which is essential for the stabilization of the tetrahedral oxyanion intermediate during catalysis, is moved 8.8 Å away.

As a consequence, many interactions that are recognized to be important for stabilization of the active conformation are lost, namely hydrogen bonds between Glu166 and His172 and between Tyr161 and His163, as well as the aromatic stacking between His163 and Phe140 (Verma *et al.*, 2020). The rotation of the side chain of His163 (located at the very bottom of the S1 subsite), the hydrogen-bond properties of which seem to be very important in determining both substrate specificity and proper inhibitor binding (Deshmukh *et al.*, 2021), is a noteworthy characteristic of this new conformation of M^{pro}. His163 is no longer available for substrate binding as it rotates away to avoid steric clashes with Gly143 CO (Fig. 4). Its position is now 'functionally' occupied by His172, which moves towards the S1 subsite (Fig. 4). The other three important residues, Tyr161, Met165 and Glu166, essentially maintain the same position as adopted in active M^{pro}. Despite the large displacement of the oxyanion loop, the position of the catalytic dyad His41 and Cys145 is not significantly altered, especially in the backbone, even though the Cys145 side chain now shows a double conformation (Fig. 5). The conserved water molecule near His41 is still present in the same position, making hydrogen bonds to the side chains of His41, His164 and Asp187 as in active SARS-CoV-2 M^{pro}.

3.4. The N-finger, the C-terminal tail and the dimeric interface are perturbed in new-inactive M^{pro}

In new-inactive M^{pro}, the dimeric interface is altered compared with that of the active conformation. *PISA* analysis of the interface shows that in new-inactive M^{pro} the interface area is reduced (from 1661 to 1273 Å^2), as are the number of hydrogen bonds (from 33 to six) and the number of salt bridges (from 12 to six). However, structural features that are important for stabilization of the dimeric form are essentially conserved, namely (i) the salt bridge between Glu290 of one protomer and Arg4′ of the other (Anand *et al.*, 2002), (ii) the hydrophobic aromatic interaction between Tyr126 and Met6′ (Wei *et al.*, 2006) and (iii) the interaction of Arg298 with the N-finger and the C-terminus (Shi *et al.*, 2008). This suggests that although new-inactive M^{pro} is still able to form dimers, the dimeric state is less stable compared with that of active M^{pro}.

At the dimeric interface, relevant changes in both the N- and C-termini are present. In active M^{pro}, the N-finger of one protomer interacts and stabilizes the S1 subsite of the other protomer (Verschueren *et al.*, 2008). For instance, in active SARS-CoV-2 M^{pro} (PSB entry 6y2e) Ser1 of one protomer is hydrogen-bonded both to the carboxylate group of Glu166 and to the main chain of Phe140 of the other protomer. In the new-inactive structure, these interactions are lost as a consequence of the different oxyanion conformation of one protomer that 'pushes away' residues 1–3 of the N-finger of the other protomer (Fig. 6), with Gly2′ CO now at 3.2 Å from Ser139 NH. The rearrangement of the oxyanion loop of one protomer also influences the C-terminal tail of the other protomer, the electron density of which is no longer visible from residue 301 onwards, indicating high flexibility (Figs. 6*b* and 7). Among the residues of the oxyanion loop, Leu141 shows major changes at the level of the dimeric interface (Fig. 7*b*), also causing rotation of the side chain of Tyr118 to avoid steric clashes, further supporting its possible central role in switching between the new-inactive and active conformations.

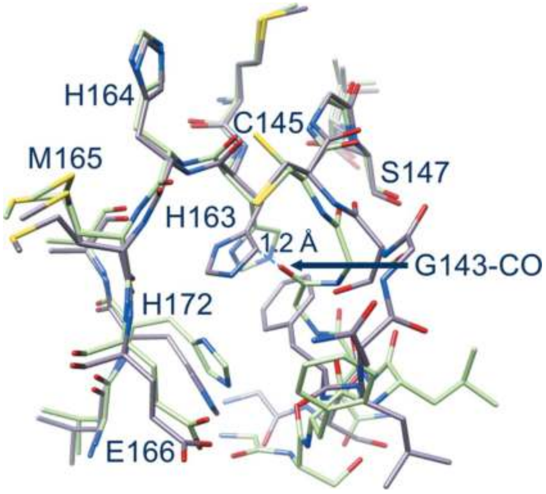


Figure 4
Comparison between new-inactive (green) and active (light magenta) M^{pro}. In the new structure the side chain of His163 rotates away to avoid steric clashes with the oxyanion loop: in the active conformation (PDB entry 6y2e) the His163 side chain would be 1.2 Å from the new position of Gly143 CO. Note also the movement of His172.

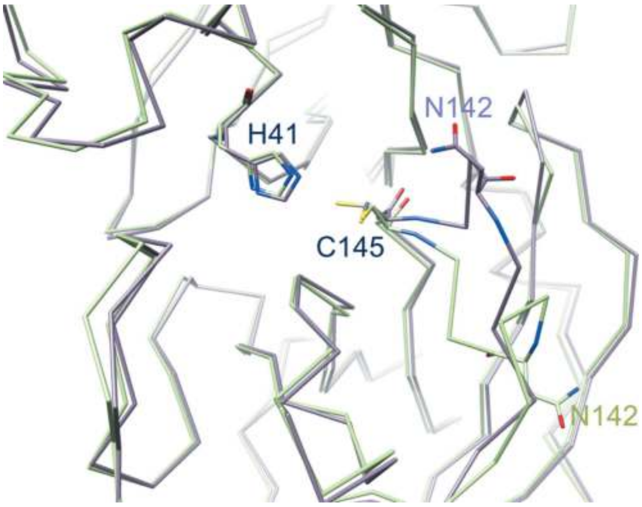


Figure 5
Catalytic dyad. In new-inactive M^{pro} (green) the position of the catalytic dyad His41 and Cys145 is similar to that in the active enzyme (PDB entry 6y2e, light magenta) despite the large shift of residues 138–144. In new-inactive M^{pro} Cys145 adopts a double conformation.

3.5. New-inactive M^{pro} can still bind substrates

Having established that the new structure is catalytically incompetent, we tried to understand whether it is still able to bind natural substrates. Superposition of the new-inactive conformation with either the active conformation in complex with the C-terminal acyl-intermediate (PDB entry 7khp; Lee *et al.*, 2020) or the SARS-CoV M^{pro} active conformation in complex with its 11-mer substrate complex (PDB entry 2q6g; Xue *et al.*, 2008) does not show evident steric clashes for the substrate. This is also valid for superposition of the new-inactive conformation with two recent complexes between SARS-CoV-2 M^{pro} and two peptide substrates corresponding to the nsp4/5 (Kneller *et al.*, 2021) and nsp8/9 (MacDonald *et al.*, 2021) cleavage sites. Additionally, a short molecular-dynamics refinement of the complexes of the new-inactive conformation of SARS-CoV-2 M^{pro} with either the C-terminal acyl-intermediate or the 11-mer peptide substrate reveal compatible binding modes, with only minor side-chain rearrangements (Fig. 8). The reshaped S1 site of the new-inactive M^{pro} could still host a P1 glutamine, although the rearrangement causes the loss of its interactions with Glu166 O$^{\varepsilon}$ and Phe140 CO in favor of a single hydrogen bond

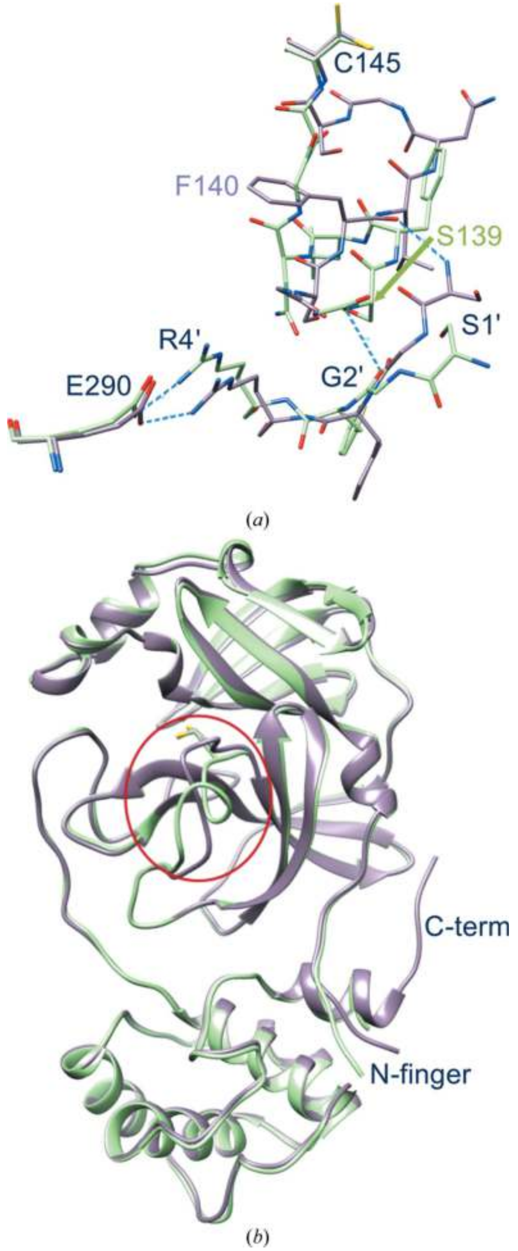


Figure 6
Displacements at the intra-protomer interface. New-inactive M^{pro} is in green and active M^{pro} is in magenta. (*a*) The new oxyanion loop of one protomer pushes away residues 1′–3′ of the other protomer; however, the key salt bridge between Arg4′ and Glu290, which is important for dimer stabilization, is conserved. (*b*) Overall superposition of active and new-inactive M^{pro} shows that besides those in the oxyanion loop (red ellipsoid), major differences are located in the N-finger and in the C-terminal tail, which is not visible in new-inactive M^{pro}.

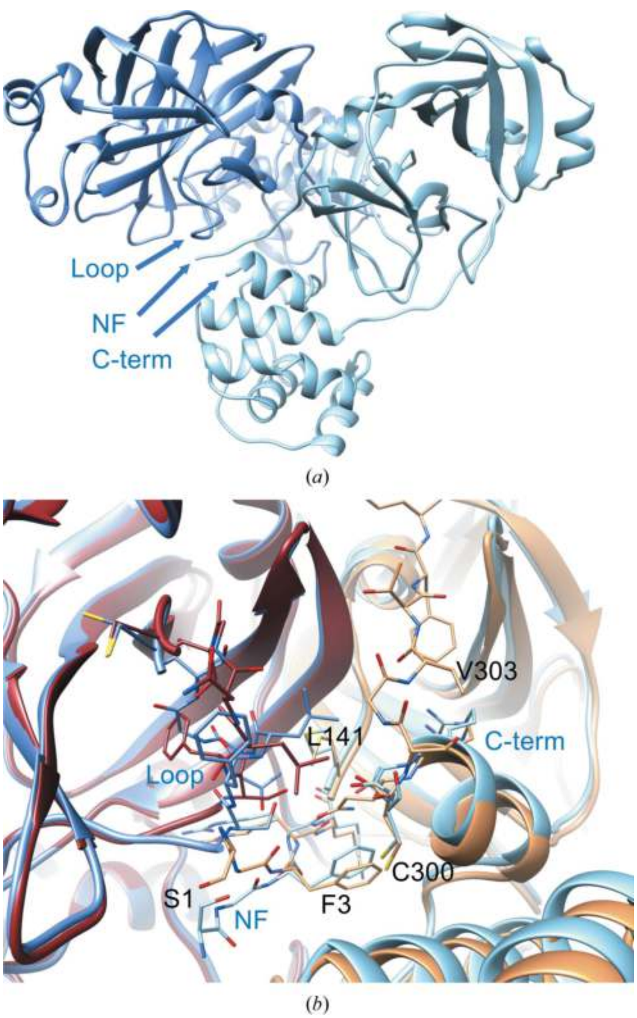


Figure 7
Dimeric architecture of new-inactive M^{pro}. (*a*) The new conformation of the oxyanion loop (labeled 'loop') causes changes in the interface between protomer *A* (blue) and protomer *B* (light blue) at the level of the N-finger (labelled 'NF') and the C-terminal tail (labeled 'C-term'). (*b*) Local differences between the new structure [blue-based colors as in (*a*)] and the canonical structure (PDB entry 6y2e; brown-based colors, with intact C-terminus): the shift of the Leu141 side chain seems to have major effects in destabilizing the C-terminal tail of the new structure.

to Gly143 CO (Fig. 9). Aside from the alterations of the S1 subsite, which alter the recognition profile of the P1 glutamine, the other interaction features are retained, namely the hydrogen bonds to Glu166 and Gln189 and the hydrophobic interactions of the P2 phenylalanine within the S2 subpocket. This is a quite remarkable observation because it suggests that the new conformation could be inactive not necessarily because it is incapable of recognizing the substrate, but because the catalytic machinery is not properly organized for an efficient catalytic event, particularly in the oxyanion-hole region, and is unable to stabilize the tetrahedral acyl intermediate. The new conformation of the oxyanion loop generates a new cavity near position S2′, as evident from comparison of the new structure with the SARS-CoV-2 acyl-enzyme (PDB entry 7khp; Lee *et al.*, 2020) and the SARS-CoV 11-mer substrate complex (PDB entry 2q6g; Xue *et al.*, 2008) (Fig. 8).

3.6. The new-inactive conformation is stable and is in equilibrium with the active conformation in solution

For SARS-CoV M^{pro}, it has been shown that the active-site loops are very dynamic and sensitive to variations in the environmental conditions (Lee *et al.*, 2005; Tan *et al.*, 2005; Xue *et al.*, 2007, 2008; Yang *et al.*, 2003; Zheng *et al.*, 2007). Similarly, the oxyanion loop of SARS-CoV-2 M^{pro} showed conformational flexibility as deduced from room-temperature X-ray crystallography (Kneller, Phillips, Weiss *et al.*, 2020; Kneller, Phillips, O'Neill *et al.*, 2020). To test the stability and to model the dynamics of new-inactive M^{pro}, specifically of the

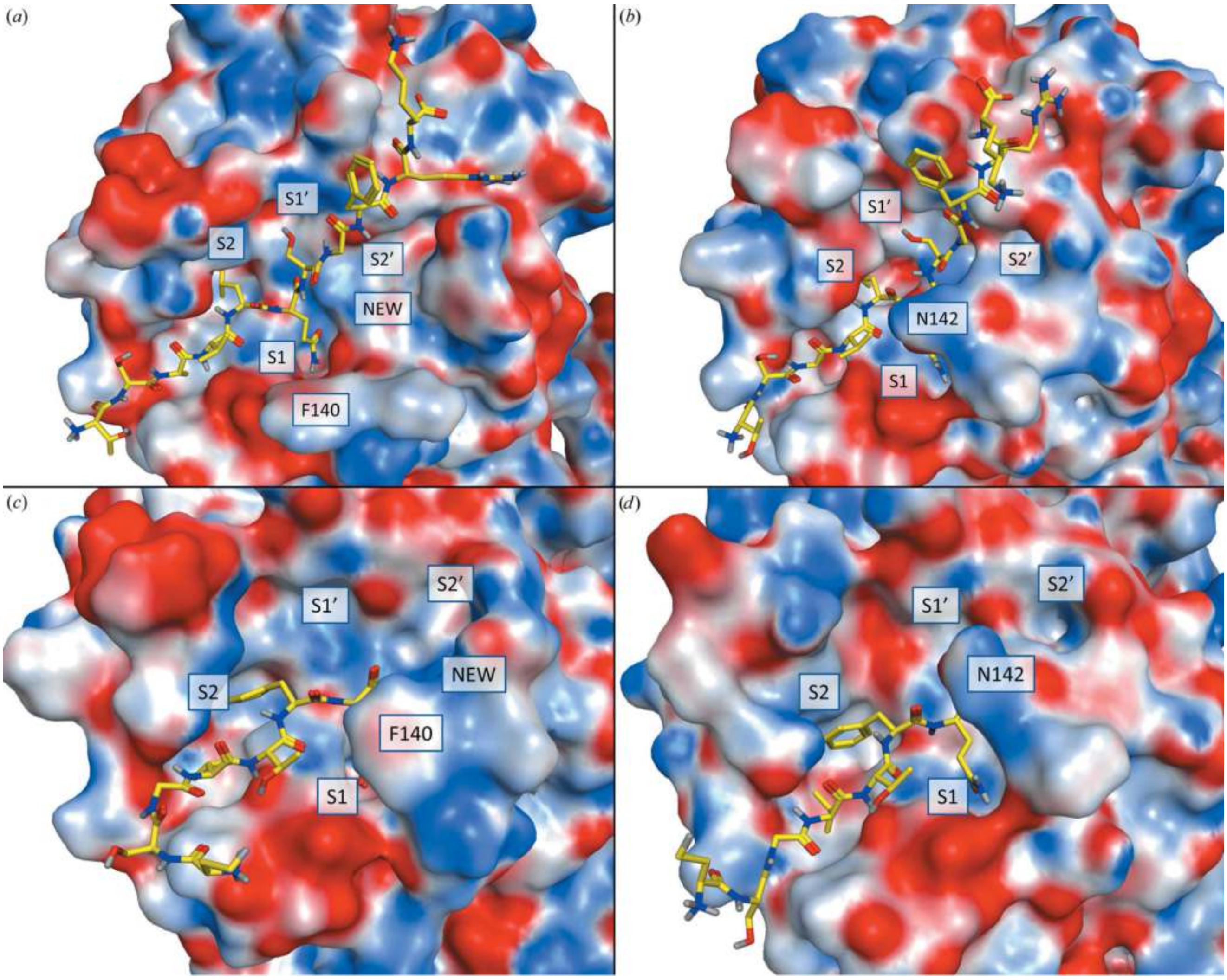


Figure 8
Reshaping of the S1 and S2′ subsites. Molecular-dynamics modeling of the hypothetical interaction of new-inactive M^{pro} with substrates is shown. Top, putative interaction with the 11-mer pseudo-substrate peptide from PDB entry 2q6g: (*a*) new-inactive M^{pro}, (*b*) SARS-CoV M^{pro} from PDB entry 2q6g. Bottom, putative interaction with the acyl-intermediate of the M^{pro} C-terminal autoprocessing site: (*c*) new-inactive M^{pro}, (*d*) M^{pro} in PDB entry 7khp. As a result of the rearrangement of the oxyanion loop, a new cavity near the S2′ site, labeled 'NEW', is formed.

oxyanion loop and regions involved in substrate binding, we performed crystallographic ensemble refinement (Burnley *et al.*, 2012) and MD simulations.

The 60 structures generated by ensemble refinement of new-inactive M^{pro} compatible with the crystallographic restraints confirm the new conformation of the oxyanion loop and reveal that its flexibility is comparable to that of other portions of the substrate-binding region (residues 43–51 in domain I and residues 188–198 in the flexible linker connecting domains II and III; Fig. 10), as also found in the literature. In four out of 60 structures the oxyanion-loop conformation is similar to that in the active form, which is in line with the experimental observation of a residual electron density compatible with the presence of a small fraction of the oxyanion loop and of the side chain of His163 in the active conformation in the crystal state. In this respect, all structures determined here, including new-inactive M^{pro}, were obtained from batches of correctly autoprocessed protein (*i.e.* catalytically active towards itself at the N-terminus) which displayed normal catalytic activity in solution towards substrate peptides.

This strongly suggests the presence of a dynamic equilibrium in solution with the coexistence of different conformations, including inactive conformations. In other words, exhibition of the correct catalytic activity on the macroscopic level (with the full ensemble of conformational states available in solution for M^{pro}) does not contrast with the possibility of selection by the crystallization process (in this case probably favored by the presence of certain small molecules) of a subpopulation of a catalytically incompetent form of the enzyme as shown here and for the previous structure with PDB code 1uj1. The conclusion that the dynamic equilibrium in solution includes both the active and the new-inactive conformation is supported by comparing the results of ensemble refinement of the structure in the free state with very poor electron density for the oxyanion loop (Supplementary Fig. S2*b*). The refined ensemble conformations show a highly dynamic oxyanion loop, with 20% of conformations similar to the active conformation, 23% of conformations similar to the inactive conformation and 57% of conformations in intermediate states.

To assess the structural stability of the new-inactive conformation of SARS-CoV-2 M^{pro} and to compare it with the active conformation, three independent 1 µs classical molecular-dynamics simulations were performed for both conformations. For the active state, PDB entry 6y2e was taken as a reference. As depicted in Fig. 11, which summarizes the principal geometric analysis performed along the MD trajectories, the two structures show a similar degree of stability. The backbone r.m.s.d. profile for PDB entry 7nij (Fig. 11*b*), representing the new-inactive conformation of M^{pro}, displays moderately higher fluctuations with respect to the active state (Fig. 11*a*). As can be seen in the per-residue r.m.s.f. plots (Figs. 11*c* and 11*d*), this difference can mainly be attributed to major structural fluctuations in the same regions that were marked as flexible by the crystallographic data, namely the three flexible loops 43–51, 188–198 and 272–279 and the C-terminus (299–306), while the rest of the structure is quite stiff, as in the active state. Specifically, the C-terminus in the new-inactive conformation of M^{pro} shows the highest amplitude of movement, as denoted by the high r.m.s.f. values associated with these residues. This result agrees with the absence of electron density for residues 301–306, which indicates high flexibility of this region. Instead, the N-terminus (residues 1–4) shows more limited fluctuations for both M^{pro} conformations, which is in agreement with the presence of well defined electron density in both structures. The overall structural stability of the new-

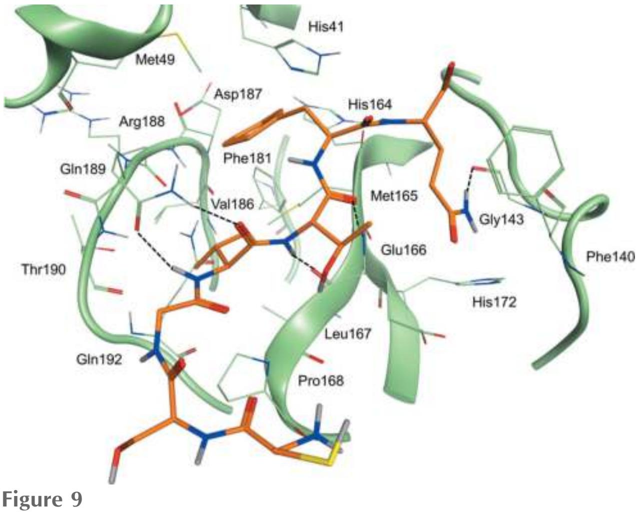


Figure 9
Details of the putative interaction between new-inactive M^{pro} (green) and the C-terminal acyl-intermediate peptide substrate from PDB entry 7khp (orange). Hydrogen bonds between the substrate and the binding site are depicted as dashed black lines. Aside from the P1 glutamine and its interactions with the P1 pocket, other common interaction features such as hydrogen bonds to Glu166 and Gln189 and hydrophobic interactions of the P2 phenylalanine side chain within the S2 subpocket are retained.

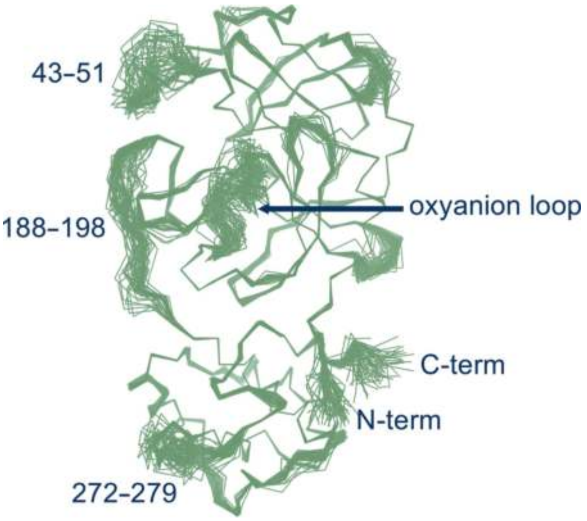


Figure 10
Ensemble refinement. The 60 structures generated by ensemble refinement highlight the mobile regions of new-inactive M^{pro}. The oxyanion loop, which is confirmed in the new conformation, has a flexibility similar to those of residues 43–51 and 188–198 involved in substrate recognition as the S3 and S4 sites.

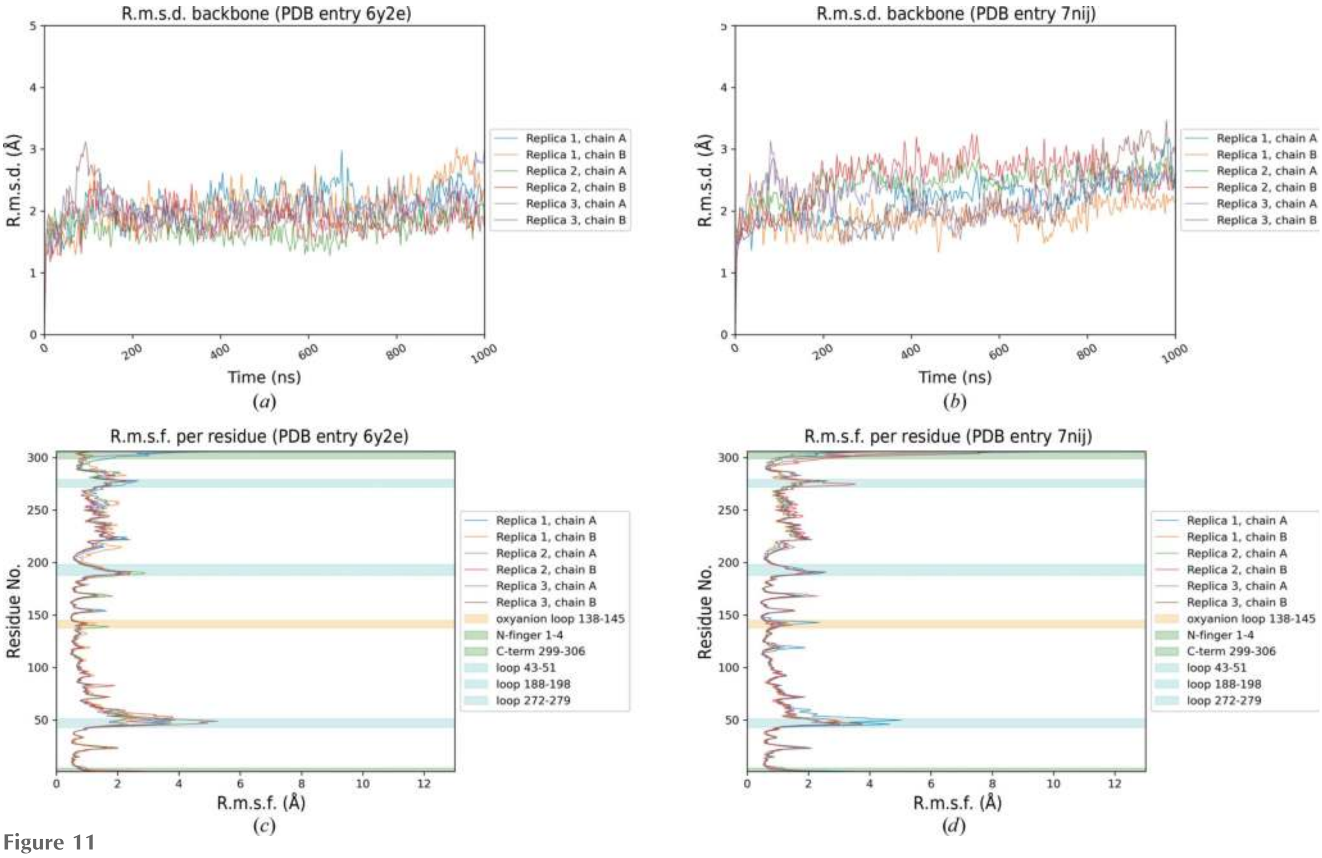


Figure 11

Results of MD simulations. Summary of the key geometric analysis performed along the MD trajectories for both the active (PDB entry 6y2e) and new-inactive (PDB entry 7nij) conformations of SARS-CoV-2 M^{pro}. (*a*) and (*b*) highlight the time-dependent variation of the protein root-mean-square deviation (r.m.s.d.) of C^{α} atomic positions for PDB entries 6y2e and 7nij, respectively. (*c*) and (*d*) summarize the per-residue mean root-mean-square fluctuation (r.m.s.f.) of atomic positions of protein C^{α} atoms for PDB entries 6y2e and 7nij, respectively. The most relevant regions of the protein are highlighted in the plot for visualization clarity as described in the legend. For both r.m.s.d. and r.m.s.f. analyses, each chain composing the crystallographic dimer is considered separately.

inactive conformation of M^{pro} is also confirmed by the time-dependent evolution of both secondary-structure elements and the protein radius of gyration (R_g), with only minor oscillations, similar to those seen in the active conformation (Supplementary Figs. S4, S5 and S6). Despite the slightly higher fluctuations observed in the inactive conformation, no sufficient motions were observed to shed light on a possible transition mechanism between the two conformations. It is not surprising that such rearrangement was not sampled even on a 1 µs scale, since such collective motions in proteins usually involve longer timescales (*i.e.* millisecond to microsecond; Orellana, 2019).

4. Discussion

We had the opportunity to capture a new and stable (as seen in MD simulations) inactive state of M^{pro}, called new-inactive, expanding the knowledge of the conformational space accessible to the enzyme. Altogether, the movements in the substrate-binding region and near the catalytic site result in a significant reshaping of the reaction center (Figs. 3, 4 and 8) that has never previously been observed and is much more pronounced than in the previously described collapsed-inactive M^{pro} conformation. The conformation adopted by residues 139–144 of the oxyanion loop is potentially catalytically incompetent. The backbones of key residues in the oxyanion hole are 8–10 Å away from the catalytically competent position. Fundamental interactions for the proper function of the enzyme are broken or absent, as illustrated in the previous section. Among the residues of the oxyanion loop, Phe140, Leu141 and Asn142 play a major role in the shift between the new-inactive and active conformations. The new state of the oxyanion loop of one protomer pushes the N-finger of the second protomer away from the position adopted in the active enzyme. The last six residues of the C-terminal tail are not visible in the electron-density map and were confirmed to be fully flexible by MD simulations. The novel conformations of the oxyanion loop and of the N- and C-termini result in a weakening of the dimeric architecture, as shown by decreases in the interaction surface area and in the number of inter-protomer interactions. Major variations in the dimeric interface are connected to Leu141 of the oxyanion loop.

This new structure is relevant for the analysis of the M^{pro} catalytic cycle, which was recently investigated using biodynamics theory under non-equilibrium conditions (Selvaggio &

Pearlstein, 2018), using the available crystal structures, which show M$^{\rm pro}$ in different conformational states (Wan *et al.*, 2020). This novel approach tries to mimic *in vivo* conditions, which depend on non-equilibrium structure–kinetics relationships. From this analysis a substrate-induced M$^{\rm pro}$ activation mechanism was developed, suggesting the existence of a complex substrate-binding activation mechanism in both SARS-CoV and SARS-CoV-2. The proposed catalytic cycle involves transition from the collapsed-inactive conformation of the oxyanion loop, represented by the free form of monomeric M$^{\rm pro}$ (PDB entry 2qcy), to the putative substrate-bound form of monomeric M$^{\rm pro}$, represented by one monomer of PDB entry 2q6g (with an active oxyanion loop), and finally to the dimeric fully active state, represented by dimeric M$^{\rm pro}$ (PDB entry 6m03; very similar to PDB entry 6y2e). The new-inactive structure presented here shows a new conformational state with an accessible oxyanion loop, adding novel important pieces of information to the structural dynamics of the substrate-induced activation of M$^{\rm pro}$ in the context of its catalytic cycle. In the non-equilibrium model, it was hypothesized that transition of the oxyanion loop from the inactive to the active conformation is triggered mainly by solvation/desolvation effects. This also applies to transitions involving our new-inactive structure, where, for activation, Phe140 moves from an exposed position (with no minimally frustrated interactions) to a buried position (with eight minimally frustrated interactions), while Asn142 moves from a buried position to an exposed position. In the context of the conformational dynamics of M$^{\rm pro}$, the intriguing possibility esists that the remodeling of the S2′ subsite can be correlated with the large amino-acid variation in position P2′ of SARS coronaviral nonstructural protein (nsp) cleavage sites, M$^{\rm pro}$ autoprocessing included. Despite being catalytically incompetent, this new state (with a novel cavity in position S2′) seems to be able to bind natural substrates of M$^{\rm pro}$ (see Figs. 8 and 9). Among the 11 substrates of SARS-CoV-2 M$^{\rm pro}$, position P2′ is highly variable, hosting nine different amino acids with very different chemical and structural properties: small, such as Gly and Ala, bulky hydrophobic, such as Ile, Val and Leu, positively charged, such as Lys, negatively charged, such as Glu, and polar and hydrogen-bond donor/acceptor, such as Ser and Asn. It is conceivable that the flexibility of the oxyanion-loop conformation is correlated to this variability of the substrates, specifically in position P2′, and to the necessity to accommodate the different substrates during the maturation process of the pp1a and pp1ab polypeptides, in the correct succession of proteolytic events. We suggest that this new conformational state is that preferred by the enzyme to efficiently host substrates with bulky hydrophobic residues in position P2′, for instance for the processing of nsp7/8 (Ile), nsp12/13 (Val) and nsp14/15 (Leu) cleavage sites. According to the M$^{\rm pro}$ reaction scheme proposed by Wan *et al.* (2020), the substrate-binding event triggers the conformational switch of the oxyanion loop, which adopts the necessary conformation for a productive catalytic event. Overall, the following scheme can be proposed: (i) for the initial binding, specific substrates (with bulky residues in position P2′) select the new-inactive conformation among a complex ensemble of different conformations of M$^{\rm pro}$ in mutual equilibrium, (ii) the binding event causes conformational changes of the oxyanion loop and, mainly, of the side chains of Glu166, His172 and His163, (iii) the dimeric architecture is stabilized because of rearrangements of the N-finger and the C-terminus and (iv) the resulting activated enzyme is ready to properly hydrolyze the substrate.

The new-inactive structure is also important for the structure-based drug-discovery process that is currently being applied to M$^{\rm pro}$ (Deshmukh *et al.*, 2021). The approach of 'repurposing' already known drugs via classical docking methodologies on the 3D structure of the protein target is interesting because, methodologically, it is potentially fast and the safety profiles of the tested compounds are already known. This justifies the large amount of research devoted to repurposing known antiviral drugs against M$^{\rm pro}$ (Cannalire *et al.*, 2016). Obviously, the success rate of these campaigns would greatly benefit from the possibility of targeting significantly different, stable, conformations. In this respect, the discovery of the new stable inactive conformation of M$^{\rm pro}$ presented here, with the remodeling of the S1 subsite and the formation of the nearby new cavity near subsite S2′ (poorly explored until now as known inhibitors usually span the enzyme S1–S4 subsites), offers solid attractive possibilities for the design of completely new classes of antiviral drugs targeting M$^{\rm pro}$. Indeed, a putative binder of the new-inactive form could reduce the population of the active conformation by stabilizing the inactive conformation. Also, a ligand able to bind the novel, readapted site around the catalytic cysteine could sterically hamper the recognition of the substrate. In addition, the possibility of targeting a novel subpocket could increase the affinity by establishing novel contacts and interactions. Most of the more promising M$^{\rm pro}$ inhibitors were developed by optimizing starting hits that were further decorated to explore the subpockets located around the catalytic center, following the classic route of fragment maturation in fragment-based lead discovery (Yang & Yang, 2021). One notable example is represented by the optimization of portions of parampanel on S1 and S1′ and its engagement of S3–S4, which lead to a fourfold boost in IC$_{50}$ activity (Zhang *et al.*, 2021).

In conclusion, the new-inactive structure of M$^{\rm pro}$ is relevant for better understanding of the function and mechanism of action of this fundamental enzyme for SARS-CoV-2 replication in the cell, with a particular accent on the dynamics within the catalytic cycle of the enzyme, which explores different conformational states including that presented here for the first time. Further, the discovery of this unprecedented inactive conformation of M$^{\rm pro}$ provides a unique opportunity for the more successful design of antiviral drugs with improved pharmacological properties using both classical docking-based and innovative non-equilibrium-based approaches.

Acknowledgements

The authors are grateful to Professor Rolf Hilgenfeld, University of Lübeck, Lübeck, Germany for the free availability of the plasmid encoding SARS-CoV-2 M^{pro} (Zhang *et al.*, 2020). We thank ESRF beamline ID23-2 local contact Daniele De Sanctis and beamline ID23-1 local contact Romain Talon for help and assistance with data collection. The MMS laboratory is very grateful to Chemical Computing Group, OpenEye and Acellera for the scientific and technical partnership. The MMS laboratory gratefully acknowledges the support of Nvidia Corporation in the donation of the Titan V GPU used for this research. We thank Professor Stefano Mammi for the careful reading of the manuscript and his constructive remarks. Author contributions were as follows. Recombinant protein production and purification were planned and performed by MLM and MB, mass spectrometry by AS and BG, activity measurements by AS, MLM and BG, crystallization experiments and data collection by EF, GG and RB, data processing, structure solution and refinement by RB, MD simulation by MP, MS and SM, overall design of the research by MB, BG, SM, CS and RB, and manuscript writing by RB. The authors declare that they have no conflicts of interest. Open Access Funding provided by Universita degli Studi di Padova within the CRUI-CARE Agreement.

Funding information

This work was supported by funding from the CARIPARO Foundation (‘Progetti di ricerca sul Covid-19’ No. 55812 to BG and PhD scholarship to EF), from the Department of Chemical Sciences (project P-DiSC #01BIRD2018-UNIPD to MB) and from the University of Padua (Starting Grant STARS@UNIPD – call 2019 to GG).

References

Afonine, P. V., Grosse-Kunstleve, R. W., Echols, N., Headd, J. J., Moriarty, N. W., Mustyakimov, M., Terwilliger, T. C., Urzhumtsev, A., Zwart, P. H. & Adams, P. D. (2012). *Acta Cryst.* D**68**, 352–367.

Ahmad, B., Batool, M., Ain, Q., Kim, M. S. & Choi, S. (2021). *Int. J. Mol. Sci.* **22**, 9124.

Allaire, M., Chernaia, M. M., Malcolm, B. A. & James, M. N. G. (1994). *Nature*, **369**, 72–76.

Anand, K., Palm, G. J., Mesters, J. R., Siddell, S. G., Ziebuhr, J. & Hilgenfeld, R. (2002). *EMBO J.* **21**, 3213–3224.

Anand, K., Ziebuhr, J., Wadhwani, P., Mesters, J. R. & Hilgenfeld, R. (2003). *Science*, **300**, 1763–1767.

Behnam, M. A. M. (2021). *Biochimie*, **182**, 177–184.

Bryngelson, J. D. & Wolynes, P. G. (1987). *Proc. Natl Acad. Sci. USA*, **84**, 7524–7528.

Brzezinski, D., Kowiel, M., Cooper, D. R., Cymborowski, M., Grabowski, M., Wlodawer, A., Dauter, Z., Shabalin, I. G., Gilski, M., Rupp, B., Jaskolski, M. & Minor, W. (2021). *Protein Sci.* **30**, 115–124.

Burnley, B. T., Afonine, P. V., Adams, P. D. & Gros, P. (2012). *eLife*, **1**, e00311.

Cannalire, R., Barreca, M. L., Manfroni, G. & Cecchetti, V. (2016). *J. Med. Chem.* **59**, 16–41.

Case, D. A., Cheatham, T. E., Darden, T., Gohlke, H., Luo, R., Merz, K. M., Onufriev, A., Simmerling, C., Wang, B. & Woods, R. J. (2005). *J. Comput. Chem.* **26**, 1668–1688.

Chen, H., Wei, P., Huang, C., Tan, L., Liu, Y. & Lai, L. (2006). *J. Biol. Chem.* **281**, 13894–13898.

Chen, M., Chen, X., Schafer, N. P., Clementi, C., Komives, E. A., Ferreiro, D. U. & Wolynes, P. G. (2020). *Nat. Commun.* **11**, 5944.

Dai, W., Zhang, B., Jiang, X., Su, H., Li, J., Zhao, Y., Xie, X., Jin, Z., Peng, J., Liu, F., Li, C., Li, Y., Bai, F., Wang, H., Cheng, X., Cen, X., Hu, S., Yang, X., Wang, J., Liu, X., Xiao, G., Jiang, H., Rao, Z., Zhang, L., Xu, Y., Yang, H. & Liu, H. (2020). *Science*, **368**, 1331–1335.

Deshmukh, M. G., Ippolito, J. A., Zhang, C.-H., Stone, E. A., Reilly, R. A., Miller, S. J., Jorgensen, W. L. & Anderson, K. S. (2021). *Structure*, **29**, 823–833.

Douangamath, A., Fearon, D., Gehrtz, P., Krojer, T., Lukacik, P., Owen, C. D., Resnick, E., Strain-Damerell, C., Aimon, A., Ábrányi-Balogh, P., Brandão-Neto, J., Carbery, A., Davison, G., Dias, A., Downes, T. D., Dunnett, L., Fairhead, M., Firth, J. D., Jones, S. P., Keeley, A., Keserü, G. M., Klein, H. F., Martin, M. P., Noble, M. E. M., O’Brien, P., Powell, A., Reddi, R. N., Skyner, R., Snee, M., Waring, M. J., Wild, C., London, N., von Delft, F. & Walsh, M. A. (2020). *Nat. Commun.* **11**, 5047.

Drayman, N., DeMarco, J. K., Jones, K. A., Azizi, S., Froggatt, H. M., Tan, K., Maltseva, N. I., Chen, S., Nicolaescu, V., Dvorkin, S., Furlong, K., Kathayat, R. S., Firpo, M. R., Mastrodomenico, V., Bruce, E. A., Schmidt, M. M., Jedrzejczak, R., Muñoz-Alía, M., Schuster, B., Nair, V., Han, K., O’Brien, A., Tomatsidou, A., Meyer, B., Vignuzzi, M., Missiakas, D., Botten, J. W., Brooke, C. B., Lee, H., Baker, S. C., Mounce, B. C., Heaton, N. S., Severson, W. E., Palmer, K. E., Dickinson, B. C., Joachimiak, A., Randall, G. & Tay, S. (2021). *Science*, **373**, 931–936.

Eastman, P., Swails, J., Chodera, J. D., McGibbon, R. T., Zhao, Y., Beauchamp, K. A., Wang, L.-P., Simmonett, A. C., Harrigan, M. P., Stern, C. D., Wiewiora, R. P., Brooks, B. R. & Pande, V. S. (2017). *PLoS Comput. Biol.* **13**, e1005659.

El-Baba, T. J., Lutomski, C. A., Kantsadi, A. L., Malla, T. R., John, T., Mikhailov, V., Bolla, J. R., Schofield, C. J., Zitzmann, N., Vakonakis, I. & Robinson, C. V. (2020). *Angew. Chem. Int. Ed.* **59**, 23544–23548.

Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Acta Cryst.* D**66**, 486–501.

Evans, P. R. & Murshudov, G. N. (2013). *Acta Cryst.* D**69**, 1204–1214.

Freiberger, M. I., Guzovsky, A. B., Wolynes, P. G., Parra, R. G. & Ferreiro, D. U. (2019). *Proc. Natl Acad. Sci. USA*, **116**, 4037–4043.

Frey, P. A. & Hegeman, A. D. (2007). *Enzymatic Reaction Mechanisms.* Oxford University Press.

Frishman, D. & Argos, P. (1995). *Proteins*, **23**, 566–579.

Fu, L., Ye, F., Feng, Y., Yu, F., Wang, Q., Wu, Y., Zhao, C., Sun, H., Huang, B., Niu, P., Song, H., Shi, Y., Li, X., Tan, W., Qi, J. & Gao, G. F. (2020). *Nat. Commun.* **11**, 4417.

Ghahremanpour, M. M., Tirado-Rives, J., Deshmukh, M., Ippolito, J. A., Zhang, C.-H., Cabeza de Vaca, I., Liosi, M.-E., Anderson, K. S. & Jorgensen, W. L. (2020). *ACS Med. Chem. Lett.* **11**, 2526–2533.

Gowers, R., Linke, M., Barnoud, J., Reddy, T., Melo, M., Seyler, S., Domański, J., Dotson, D., Buchoux, S., Kenney, I. & Beckstein, O. (2016). *Proceedings of the 15th Python in Science Conference (SCIPY 2016)*, edited by S. Benthall & S. Rostrup, pp. 98–105. https://conference.scipy.org/proceedings/scipy2016/oliver_beckstein.html.

Günther, S., Reinke, P. Y. A., Fernández-García, Y., Lieske, J., Lane, T. J., Ginn, H. M., Koua, F. H. M., Ehrt, C., Ewert, W., Oberthuer, D., Yefanov, O., Meier, S., Lorenzen, K., Krichel, B., Kopicki, J.-D., Gelisio, L., Brehm, W., Dunkel, I., Seychell, B., Gieseler, H., Norton-Baker, B., Escudero-Pérez, B., Domaracky, M., Saouane, S., Tolstikova, A., White, T. A., Hänle, A., Groessler, M., Fleckenstein, H., Trost, F., Galchenkova, M., Gevorkov, Y., Li, C., Awel, S., Peck, A., Barthelmess, M., Schlünzen, F., Lourdu Xavier, P., Werner, N., Andaleeb, H., Ullah, N., Falke, S., Srinivasan, V., França, B. A., Schwinzer, M., Brognaro, H., Rogers, C., Melo, D., Zaitseva-Doyle, J. J., Knoska, J., Peña-Murillo, G. E., Mashhour, A. R., Hennicke, V., Fischer, P., Hakanpää, J., Meyer, J., Gribbon, P., Ellinger, B., Kuzikov, M., Wolf, M., Beccari, A. R., Bourenkov, G., von Stetten, D., Pompidor, G., Bento, I., Panneerselvam, S., Karpics, I., Schneider, T. R., Garcia-Alai, M. M., Niebling, S., Günther, C., Schmidt, C., Schubert, R., Han, H., Boger, J., Monteiro, D. C. F., Zhang, L., Sun, X., Pletzer-Zelgert, J., Wollenhaupt, J., Feiler, C. G., Weiss, M. S., Schulz, E.-C., Mehrabi, P., Karničar, K., Usenik, A., Loboda, J., Tidow, H., Chari, A., Hilgenfeld, R., Uetrecht, C., Cox, R., Zaliani, A., Beck, T., Rarey, M., Günther, S., Turk, D., Hinrichs, W., Chapman, H. N., Pearson, A. R., Betzel, C. & Meents, A. (2021). *Science*, **372**, 642–646.

Harvey, M. J., Giupponi, G. & Fabritiis, G. D. (2009). *J. Chem. Theory Comput.* **5**, 1632–1639.

Ho, B.-L., Cheng, S.-C., Shi, L., Wang, T.-Y., Ho, K.-I. & Chou, C.-Y. (2015). *PLoS One*, **10**, e0144865.

Hsu, M.-F., Kuo, C.-J., Chang, K.-T., Chang, H.-C., Chou, C.-C., Ko, T.-P., Shr, H.-L., Chang, G.-G., Wang, A. H.-J. & Liang, P.-H. (2005). *J. Biol. Chem.* **280**, 31257–31266.

Humphrey, W., Dalke, A. & Schulten, K. (1996). *J. Mol. Graph.* **14**, 33–38.

Hunter, J. D. (2007). *Comput. Sci. Eng.* **9**, 90–95.

Jaskolski, M., Dauter, Z., Shabalin, I. G., Gilski, M., Brzezinski, D., Kowiel, M., Rupp, B. & Wlodawer, A. (2021). *IUCrJ*, **8**, 238–256.

Jin, Z., Du, X., Xu, Y., Deng, Y., Liu, M., Zhao, Y., Zhang, B., Li, X., Zhang, L., Peng, C., Duan, Y., Yu, J., Wang, L., Yang, K., Liu, F., Jiang, R., Yang, X., You, T., Liu, X., Yang, X., Bai, F., Liu, H., Liu, X., Guddat, L. W., Xu, W., Xiao, G., Qin, C., Shi, Z., Jiang, H., Rao, Z. & Yang, H. (2020). *Nature*, **582**, 289–293.

Jin, Z., Zhao, Y., Sun, Y., Zhang, B., Wang, H., Wu, Y., Zhu, Y., Zhu, C., Hu, T., Du, X., Duan, Y., Yu, J., Yang, X., Yang, X., Yang, K., Liu, X., Guddat, L. W., Xiao, G., Zhang, L., Yang, H. & Rao, Z. (2020). *Nat. Struct. Mol. Biol.* **27**, 529–532.

Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. (1983). *J. Chem. Phys.* **79**, 926–935.

Kabsch, W. (2010). *Acta Cryst.* D**66**, 125–132.

Kabsch, W. & Sander, C. (1983). *Biopolymers*, **22**, 2577–2637.

Kneller, D. W., Galanie, S., Phillips, G., O'Neill, H. M., Coates, L. & Kovalevsky, A. (2020). *Structure*, **28**, 1313–1320.

Kneller, D. W., Phillips, G., O'Neill, H. M., Jedrzejczak, R., Stols, L., Langan, P., Joachimiak, A., Coates, L. & Kovalevsky, A. (2020). *Nat. Commun.* **11**, 3202.

Kneller, D. W., Phillips, G., Weiss, K. L., Pant, S., Zhang, Q., O'Neill, H. M., Coates, L. & Kovalevsky, A. (2020). *J. Biol. Chem.* **295**, 17365–17373.

Kneller, D. W., Zhang, Q., Coates, L., Louis, J. M. & Kovalevsky, A. (2021). *IUCrJ*, **8**, 973–979.

Krissinel, E. & Henrick, K. (2007). *J. Mol. Biol.* **372**, 774–797.

Lee, J., Worrall, L. J., Vuckovic, M., Rosell, F. I., Gentile, F., Ton, A.-T., Caveney, N. A., Ban, F., Cherkasov, A., Paetzel, M. & Strynadka, N. C. J. (2020). *Nat. Commun.* **11**, 5877.

Lee, T.-W., Cherney, M. M., Huitema, C., Liu, J., James, K. E., Powers, J. C., Eltis, L. D. & James, M. N. G. (2005). *J. Mol. Biol.* **353**, 1137–1151.

Li, C., Teng, X., Qi, Y., Tang, B., Shi, H., Ma, X. & Lai, L. (2016). *Sci. Rep.* **6**, 20918.

Liebschner, D., Afonine, P. V., Baker, M. L., Bunkóczi, G., Chen, V. B., Croll, T. I., Hintze, B., Hung, L.-W., Jain, S., McCoy, A. J., Moriarty, N. W., Oeffner, R. D., Poon, B. K., Prisant, M. G., Read, R. J., Richardson, J. S., Richardson, D. C., Sammito, M. D., Sobolev, O. V., Stockwell, D. H., Terwilliger, T. C., Urzhumtsev, A. G., Videau, L. L., Williams, C. J. & Adams, P. D. (2019). *Acta Cryst.* D**75**, 861–877.

Ma, C., Sacco, M. D., Hurst, B., Townsend, J. A., Hu, Y., Szeto, T., Zhang, X., Tarbet, B., Marty, M. T., Chen, Y. & Wang, J. (2020). *Cell Res.* **30**, 678–692.

MacDonald, E. A., Frey, G., Namchuk, M. N., Harrison, S. C., Hinshaw, S. M. & Windsor, I. W. (2021). *ACS Infect. Dis.* **7**, 2591–2595.

Maier, J. A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K. E. & Simmerling, C. (2015). *J. Chem. Theory Comput.* **11**, 3696–3713.

McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. (2007). *J. Appl. Cryst.* **40**, 658–674.

Michaud-Agrawal, N., Denning, E. J., Woolf, T. B. & Beckstein, O. (2011). *J. Comput. Chem.* **32**, 2319–2327.

Orellana, L. (2019). *Front. Mol. Biosci.* **6**, 117.

Parra, R. G., Schafer, N. P., Radusky, L. G., Tsai, M.-Y., Guzovsky, A. B., Wolynes, P. G. & Ferreiro, D. U. (2016). *Nucleic Acids Res.* **44**, W356–W360.

Pavan, M., Bolcato, G., Bassani, D., Sturlese, M. & Moro, S. (2021). *J. Enzyme Inhib. Med. Chem.* **36**, 1646–1650.

Pol, S. & Corouge, M. (2014). *Med. Mal. Infect.* **44**, 449–454.

Selvaggio, G. & Pearlstein, R. A. (2018). *PLoS One*, **13**, e0202376.

Shi, J., Sivaraman, J. & Song, J. (2008). *J. Virol.* **82**, 4620–4629.

Shi, J. & Song, J. (2006). *FEBS J.* **273**, 1035–1045.

Skwarecki, A. S., Nowak, M. G. & Milewska, M. J. (2021). *ChemMedChem*, **16**, 3106–3135.

Snijder, E. J., Decroly, E. & Ziebuhr, J. (2016). *Adv. Virus Res.* **96**, 59–126.

Tan, J., Verschueren, K. H. G., Anand, K., Shen, J., Yang, M., Xu, Y., Rao, Z., Bigalke, J., Heisen, B., Mesters, J. R., Chen, K., Shen, X., Jiang, H. & Hilgenfeld, R. (2005). *J. Mol. Biol.* **354**, 25–40.

Touw, W. G., Baakman, C., Black, J., te Beek, T. A. H., Krieger, E., Joosten, R. P. & Vriend, G. (2015). *Nucleic Acids Res.* **43**, D364–D368.

Ullrich, S. & Nitsche, C. (2020). *Bioorg. Med. Chem. Lett.* **30**, 127377.

Verma, N., Henderson, J. A. & Shen, J. (2020). *J. Am. Chem. Soc.* **142**, 21883–21890.

Verschueren, K. H. G., Pumpor, K., Anemüller, S., Chen, S., Mesters, J. R. & Hilgenfeld, R. (2008). *Chem. Biol.* **15**, 597–606.

Wan, H., Aravamuthan, V. & Pearlstein, R. A. (2020). *ACS Pharmacol. Transl. Sci.* **3**, 1111–1143.

Wei, P., Fan, K., Chen, H., Ma, L., Huang, C., Tan, L., Xi, D., Li, C., Liu, Y., Cao, A. & Lai, L. (2006). *Biochem. Biophys. Res. Commun.* **339**, 865–872.

Wlodawer, A., Dauter, Z., Shabalin, I. G., Gilski, M., Brzezinski, D., Kowiel, M., Minor, W., Rupp, B. & Jaskolski, M. (2020). *FEBS J.* **287**, 3703–3718.

Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., Hu, Y., Tao, Z.-W., Tian, J.-H., Pei, Y.-Y., Yuan, M.-L., Zhang, Y.-L., Dai, F.-H., Liu, Y., Wang, Q.-M., Zheng, J.-J., Xu, L., Holmes, E. C. & Zhang, Y.-Z. (2020). *Nature*, **579**, 265–269.

Xia, B. & Kang, X. (2011). *Protein Cell*, **2**, 282–290.

Xue, X., Yang, H., Shen, W., Zhao, Q., Li, J., Yang, K., Chen, C., Jin, Y., Bartlam, M. & Rao, Z. (2007). *J. Mol. Biol.* **366**, 965–975.

Xue, X., Yu, H., Yang, H., Xue, F., Wu, Z., Shen, W., Li, J., Zhou, Z., Ding, Y., Zhao, Q., Zhang, X. C., Liao, M., Bartlam, M. & Rao, Z. (2008). *J. Virol.* **82**, 2515–2527.

Yang, H. & Yang, J. (2021). *RSC Med. Chem.* **12**, 1026–1036.

Yang, H., Yang, M., Ding, Y., Liu, Y., Lou, Z., Zhou, Z., Sun, L., Mo, L., Ye, S., Pang, H., Gao, G. F., Anand, K., Bartlam, M., Hilgenfeld, R. & Rao, Z. (2003). *Proc. Natl Acad. Sci. USA*, **100**, 13190–13195.

Zhang, C.-H., Stone, E. A., Deshmukh, M., Ippolito, J. A., Ghahremanpour, M. M., Tirado-Rives, J., Spasov, K. A., Zhang, S., Takeo, Y., Kudalkar, S. N., Liang, Z., Isaacs, F., Lindenbach, B., Miller, S. J., Anderson, K. S. & Jorgensen, W. L. (2021). *ACS Cent. Sci.* **7**, 467–475.

Zhang, L., Lin, D., Sun, X., Curth, U., Drosten, C., Sauerhering, L., Becker, S., Rox, K. & Hilgenfeld, R. (2020). *Science*, **368**, 409–412.

Zheng, K., Ma, G., Zhou, J., Zen, M., Zhao, W., Jiang, Y., Yu, Q. & Feng, J. (2007). *Proteins*, **66**, 467–479.

STRUCTURAL
BIOLOGY

Volume 78 (2022)

Supporting information for article:

A new inactive conformation of SARS-CoV-2 main protease

Emanuele Fornasier, Maria Ludovica Macchia, Gabriele Giachin, Alice Sosic, Matteo Pavan, Mattia Sturlese, Cristiano Salata, Stefano Moro, Barbara Gatto, Massimo Bellanda and Roberto Battistutta

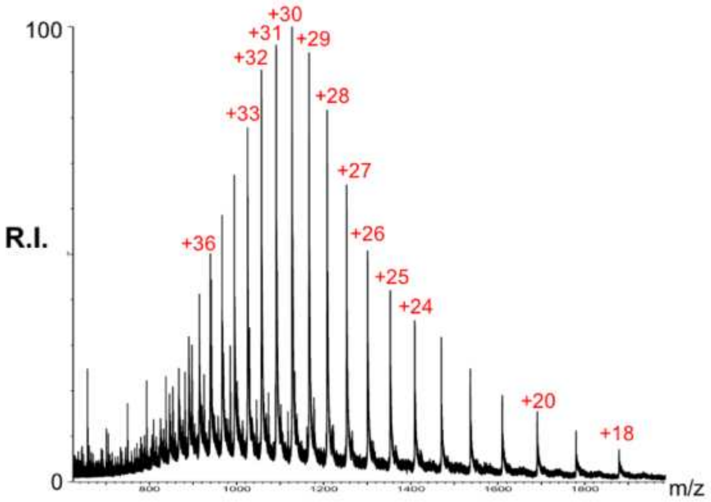


Figure S1

ESI-MS spectrum of full-length recombinant SARS-CoV-2 M^{pro} (residues 1-306). The sample containing 2 μM of the recombinant full-length SARS-CoV-2 M^{pro} was analyzed in 50% acetonitrile added of 0.1% formic acid in positive ion mode by direct infusion ESI-MS. For the sake of clarity, only a few charge states of the detected species are reported in the Figure. The species displayed a deconvoluted mass of 33796.64 Da vs a calculated one of 33796.81 Da.

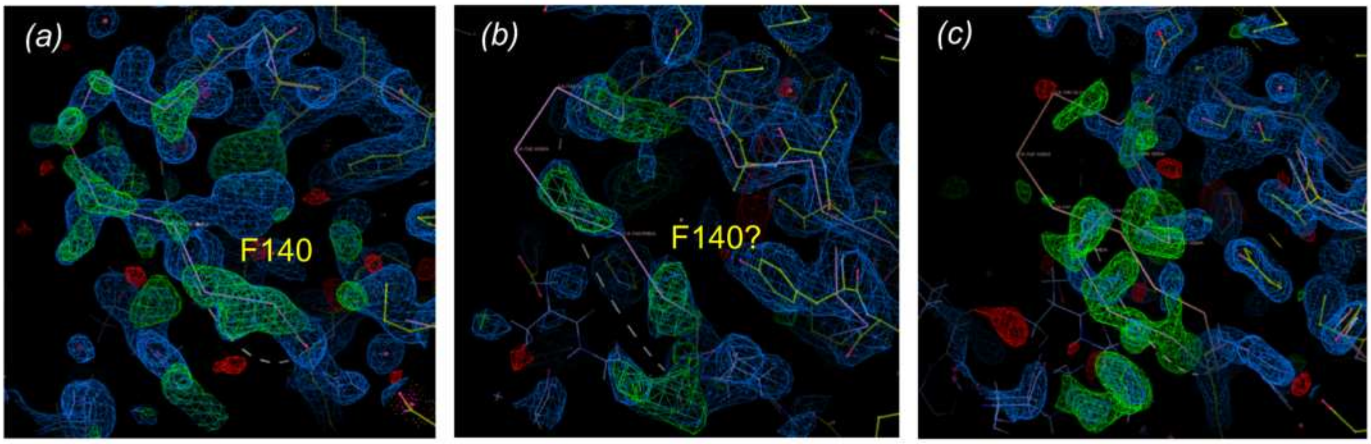


Figure S2

Initial electron densities. Representative situations after MR and a first step of refinement using a model lacking residues 139-144, 1-3 and with H163A substitution (see methods for details). Panel *(a)*, M^{pro} with good density of the oxyanion loop in the active conformation (Cα trace in magenta); side chains are visible, in particular for Phe140. Panel *(b)*, M^{pro} with very poor density for the oxyanion loop; no densities for side chains are visible. Panel *(c)*, M^{pro} electron densities clearly indicate a different conformation for the oxyanion loop (in orange the trace for the active conformation, in magenta the trace for the new conformation, new-inactive M^{pro}). 2F$_O$-F$_C$ maps in blue (contour level 1.0 σ), F$_O$-F$_C$ maps in green/red (contour level 3.0 σ). The white dashed line indicates the missing 139-144 residues.

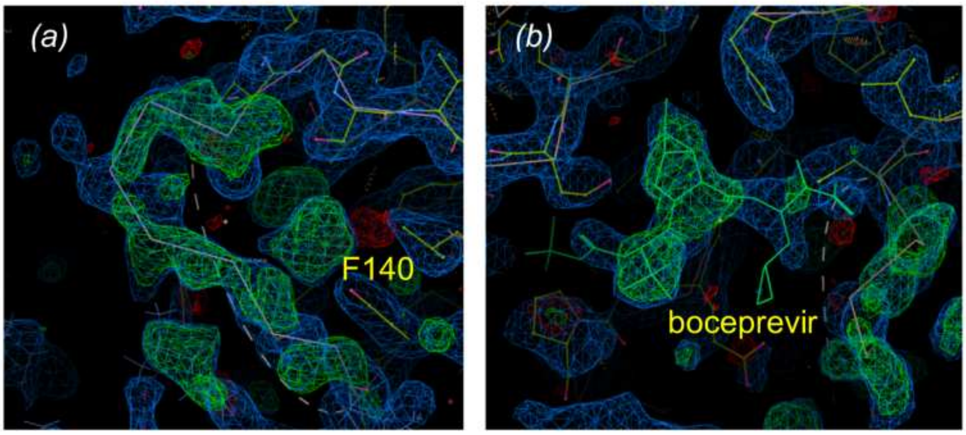


Figure S3

Initial electron densities for the M^{pro}/boceprevir complex. Representative situations after MR and a first step of refinement using a model lacking residues 139-144 and 1-3, with H163A substitution and without boceprevir (see methods for details). Panel *(a)*, M^{pro} with good density of the oxyanion loop in the active conformation (Cα trace in magenta); side chains are visible, in particular for Phe140. Panel *(b)*, electron density indicating the presence of the inhibitor (the boceprevir molecule, still not present at this refinement stage, is shown for reference). 2F$_O$-F$_C$ maps in blue (contour level 1.0 σ), F$_O$-F$_C$ maps in green/red (contour level 3.0 σ). The white dashed line indicates the missing 139-144 residues.

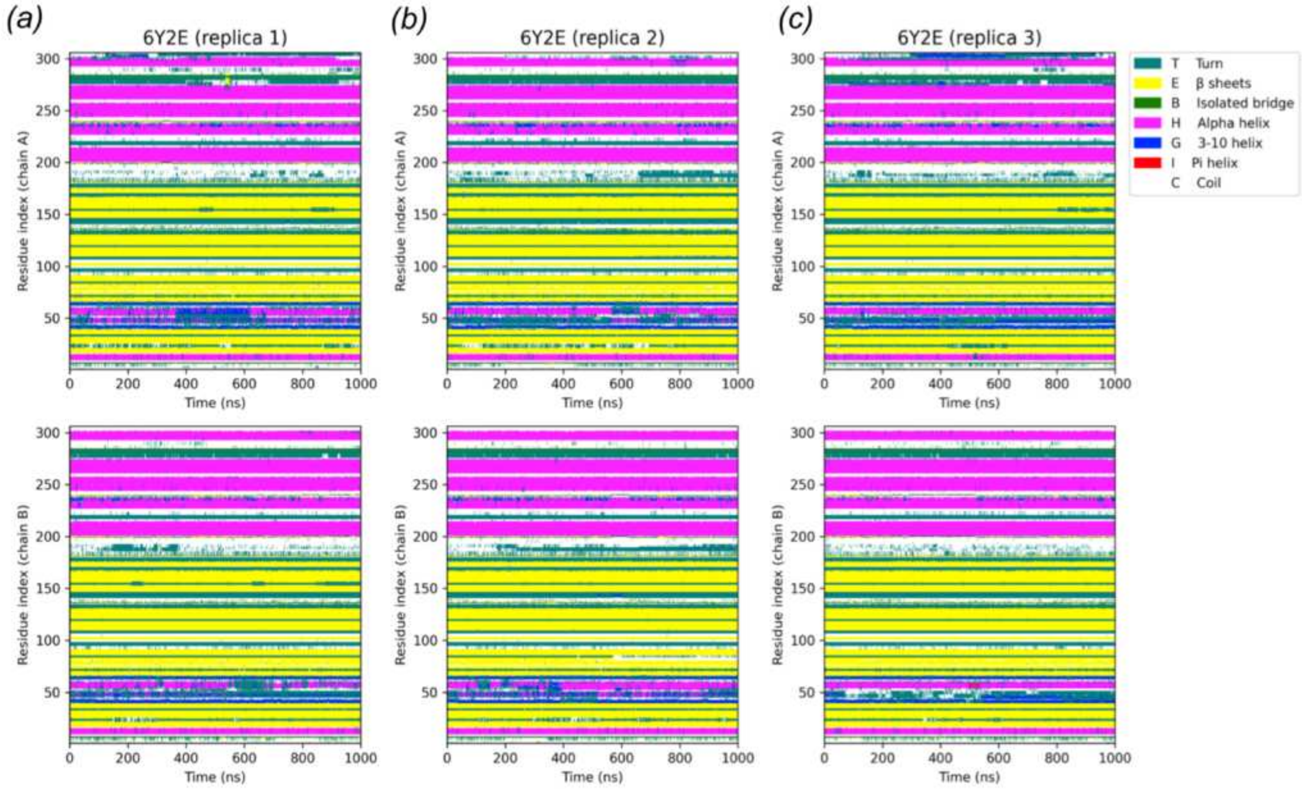


Figure S4

Time-dependent evolution of the secondary structure alongside the three MD simulations (*(a)*, *(b)* and *(c)*) for protein 6Y2E (active conformation). Colors are arranged according to Visual Molecular Dynamics (VMD) convention, as depicted in the legend. Within each subpanel, data referring to chain A of the dimer is reported in the upper part, while data concerning chain B is reported in the lower one.

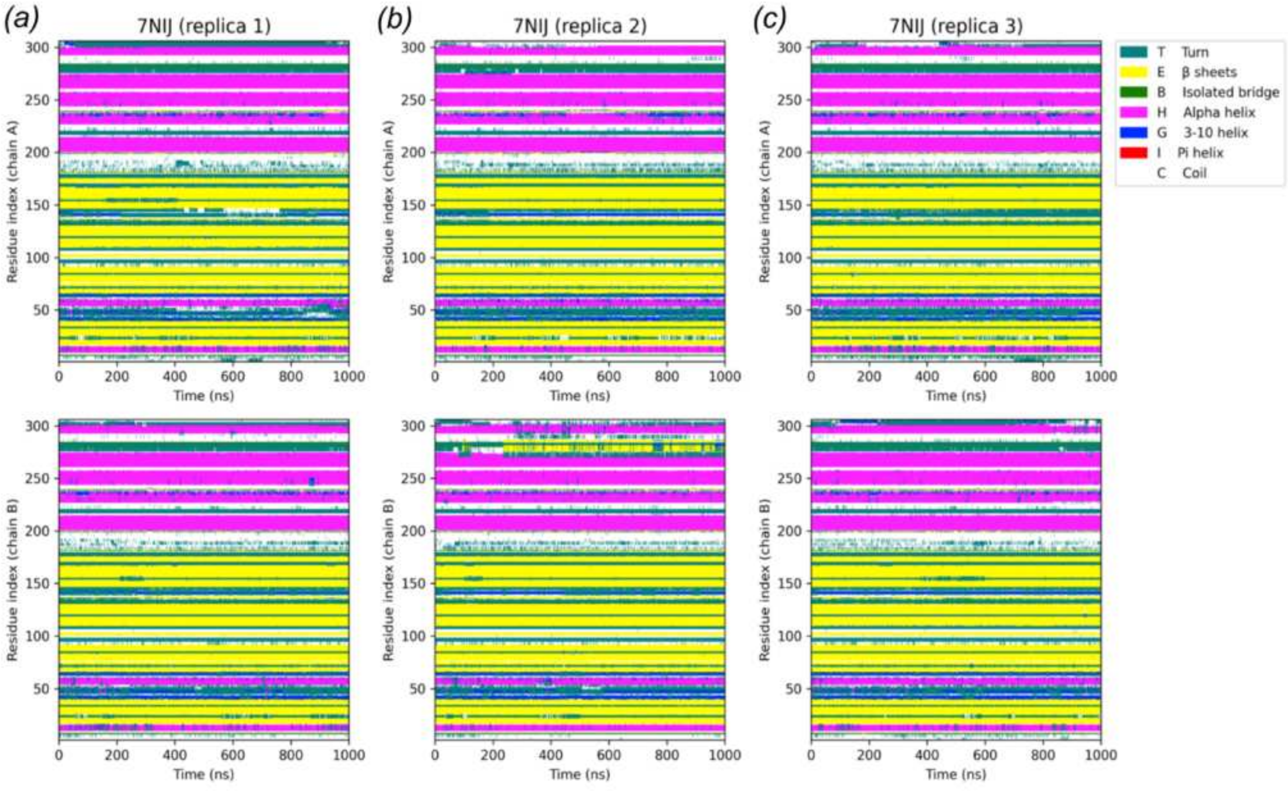


Figure S5

Time-dependent evolution of the secondary structure alongside the three MD simulations (*(a)*, *(b)* and *(c)*) for protein 7NIJ (new-inactive). Colors are arranged according to Visual Molecular Dynamics (VMD) convention, as depicted in the legend. Within each subpanel, data referring to chain A of the dimer is reported in the upper part, while data concerning chain B is reported in the lower one.

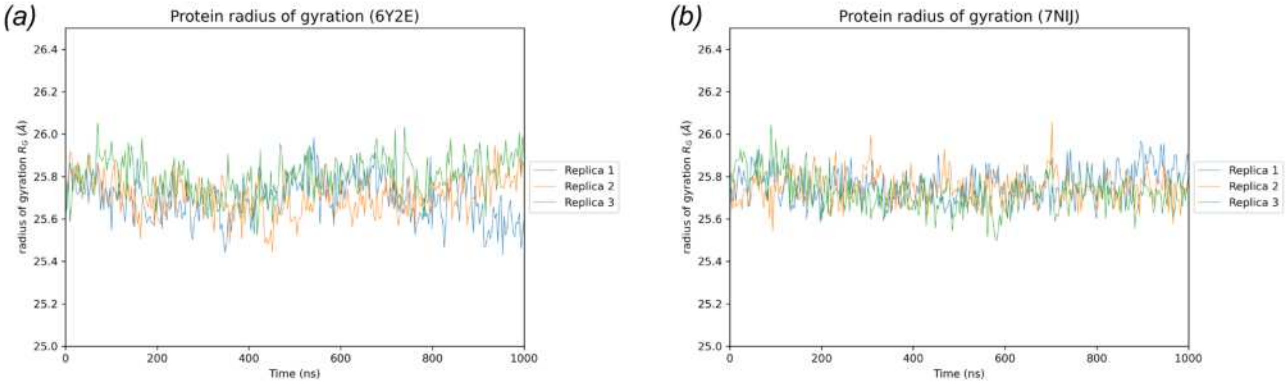


Figure S6

Time-dependent evolution of protein radius of gyration (R_g), for both the active (PDB ID: 6Y2E; panel *(a)*) and new-inactive (PDB ID: 7NIJ; panel *(b)*) conformations of SARS-CoV-2 M^{pro}.

Table S1

PDB codes and some significant pieces of information regarding the relevant structures of SARS-CoV-2 and SARS-CoV M^{pro} discussed in the paper. We used 6Y2E as reference structure for ligand-free, active SARS-CoV-2 M^{pro}. For recent comprehensive analyses of available SARS-CoV and SARS-CoV-2 M^{pro} crystal structures see Behnam, 2021; Brzezinski et al., 2021; Jaskolski et al., 2021; Wlodawer et al., 2020.

*Structure 7NIJ is presented in this paper.

PDB code	Coronavirus	Oxyanion loop conformation	Ligands	Mutations
7NIJ*	SARS-CoV-2	new-inactive	no	no
6Y2E	SARS-CoV-2	active	no	no
6M03	SARS-CoV-2	active	no	no
5REL	SARS-CoV-2	active	PCM-0102340	no
7K40	SARS-CoV-2	active	boceprevir	no
2BX4	SARS-CoV	active	no	no
1UJ1-B	SARS-CoV	collapsed-inactive	no	no
1UK2-B	SARS-CoV	collapsed-inactive	no	no
2QCY	SARS-CoV	collapsed-inactive	no	R298A
2Q6G	SARS-CoV	active	11_{mer} peptide substrate	no
7KHP	SARS-CoV	active	C-terminal acyl-intermediate	no

Table S2
Minimally frustrated energetic interactions involving selected amino acids of the active site in structure 6Y2E ("active") and structure presented in this paper ("new-inactive").

		6Y2E	**new-inactive**
Cys145	Tyr25	O	O
	Tyr26	O	O
	Asn28	O	-
	Gly29	O	O
	Cys38	O	O
	Val42	O	O
	Cys117	O	O
	Phe140	O	-
	Gly143	O	O
	Ser147	O	O
	Met162	O	O
	His163	O	O
	His164	O	O
	Met165	O	O
	Glu166	O	-
	Gly174	O	O
	Total	**16**	**13**
Phe140	Val114	O	-
	Ala116	O	-
	Ile138	O	-
	Cys145	O	-
	His163	O	-
	Met165	O	-
	Glu166	O	-
	His172	O	-
	Total	**8**	**0**
Cys117	Val13	O	-
	Glu14	-	O
	Met17	-	O
	Val18	O	O
	Gln19	O	O
	Leu27	-	O
	Gly29	-	O
	Val114	O	O
	Leu115	O	O
	Asn119	-	O
	Gly120	O	O
	Pro122	-	O
	Ser123	-	O
	Leu141	-	O
	Gly143	-	O
	Ser144	O	O
	Cys145	O	O
	Ser147	O	O
	Val148	O	O
	Total	**10**	**18**
Leu141	Ala116	-	O
	Cys117	-	O
Gly143	Cys117	-	O
	Tyr118	O	-
	Gly138	-	O
	Cys145	O	O
Ser144	Cys117	O	O
	Total	**3**	**6**

One substrate-many enzymes virtual screening uncovers missing genes of carnitine biosynthesis in human and mouse.

Marco Malatesta[1], Emanuele Fornasier[2], Martino Luigi Di Salvo[3], Angela Tramonti[4], Erika Zangelmi[1], Alessio Peracchi[1], Andrea Secchi[1], Eugenia Polverini[5], Gabriele Giachin[2], Roberto Battistutta[2], Roberto Contestabile[3], Riccardo Percudani[1]

[1]Department of Chemistry, Life Sciences and Environmental Sustainability, University of Parma, Parma, Italy

[2]Department of Chemical Sciences, University of Padua, Padova, Italy

[3]Istituto Pasteur Italia-Fondazione Cenci Bolognetti and Department of Biochemical Sciences "A. Rossi Fanelli", Sapienza University of Rome, Rome, Italy

[4]Institute of Molecular Biology and Pathology, Italian National Research Council, Rome, Italy

[5]Department of Mathematical, Physical and Computer Sciences, University of Parma, Parma, Italy

Correspondence to:

Riccardo Percudani: riccardo.percudani@unipr.it
Roberto Contestabile: roberto.contestabile@uniroma1.it

Abstract

The increasing availability of experimental and computational protein structures entices their use for function prediction. We developed an automated procedure to identify enzymes involved in metabolic reactions by evaluating substrate conformations docked to a library of protein structures. By screening AlphaFold-modeled vitamin B6-dependent enzymes, we found that a metric based on catalytically favorable conformations at the enzyme active site performed best (AUROC score=0.84) in identifying genes related to known reactions. Applying this procedure, we identified the mammalian gene encoding hydroxytrimethyllysine aldolase (HTMLA), the second enzyme of carnitine biosynthesis. Experimental validation showed that the top-ranked candidates, serine hydroxymethyl transferase (SHMT) 1 and 2, catalyze the HTMLA reaction. However, a mouse protein (threonine aldolase; Tha1) catalyzes the reaction more efficiently. Tha1 did not rank highest based on the AlphaFold model, but its rank improved to second place using the experimental crystal structure we determined at 2.26 Å resolution. We propose that mouse Tha1 be renamed as Htmla. Our findings suggest that humans have lost a gene involved in carnitine biosynthesis, with HTMLA activity of SHMT partially compensating for its function.

Introduction

In recent years, the enormous progress in the experimental determination [1,2] and computational prediction [3,4] of protein three-dimensional structures is closing the gap between the 1D and 3D protein information. However, there is still a large gap between structural information and knowledge of protein functions [5,6].

Although the function of proteins is determined by their 3D structure, this information is far less used than the sequence to predict protein function. Homology is the main evidence for protein functional annotation, and the 3D structural information is especially used to extend homology and identify residues important for function [7–11].

Yet, there is a well-established use of protein 3D structures in *molecular docking* screening, in which a database of small molecules (ligands) is screened against a protein (receptor) by assessing binding energy and binding mode. This technique is successfully used for large-scale identification of potential drugs [12,13]. In a complementary approach, a library of receptors is screened against a particular ligand. This *reverse docking* technique is mostly used for finding targets of a known drug [14]. Computational models can be used in these screenings in the absence of experimental structures [15,16].

A possible though more challenging use of docking is the matching of enzymes and substrates by predicting the binding of molecules to an enzyme active site [17–19]. Enzymes generally bind with high affinity their substrate molecules [20,21]. However, binding to an enzyme active site is not sufficient to predict that a molecule would undergo reaction. Since the enzymes have greater affinity for the reaction transition state, docking of molecules mimicking the transition state have been proposed in substrate virtual screening [17]. An alternative strategy is to evaluate whether the binding mode of the docked molecule is suitable for catalysis [19].

Enzymes bind specific substrate conformations that are favorable to the catalyzed reaction. According to the principle of stereoelectronic control, a substrate molecule assumes a conformation at the enzyme active site that minimizes the electronic energy of transition state [21]. A textbook example are the enzymes depending on vitamin B6 (pyridoxal 5'-phosphate; PLP), which catalyze different reactions on amino acids by cleaving different Cα bonds. Cleavage of a particular Cα bond by a specific PLP-dependent enzyme depends on the bond orientation relative to the PLP ring [22–25]. This allows to predict which substrate conformations at the active site favor reactions such as, e.g., racemization, decarboxylation, side-chain cleavage.

Enzymatic reactions for which no genes or proteins are known are present in various metabolic pathways [26,27]. The molecular identification of these 'pathway holes' through a reverse docking approach has now become feasible thanks to the availability of high-quality structures at the proteome level [28].

An example of a metabolic pathway involving a reaction that has not yet been assigned an amino acid sequence is carnitine biosynthesis in mammals [29]. Various eukaryotes synthesize the mitochondrial fatty-acid carrier carnitine through a dedicated four-step pathway. At variance with the fungus *Candida albicans*, in humans and other metazoans the molecular identity of 3-hydroxy-N$^\varepsilon$-trimethyllysine aldolase (HTMLA) catalyzing the second step of the pathway is not established, although it is known that the reaction is PLP-dependent [29–31]. This information allows one to restrict the search to a subset of proteins whose full set (PLPome) can be identified by homology [32]. An additional advantage is that the active site of PLP-dependent enzymes is readily identified from the position of the catalytic lysine [33,34].

Here we devised an *in silico* screening procedure (OSMES: one substrate-many enzymes screening) to identify at the structure level enzymes able to bind a given substrate and catalyze a particular PLP-dependent reaction. First of all, using experimentally known enzyme-substrate combinations, we assessed the performance of metrics based on different criteria (binding energy, statistical frequency, catalytically favorable conformation) for the ranking of docked enzyme-substrate complexes. We then applied OSMES with the best performing metric to the identification of HTMLA candidates in the human and mouse PLPomes. The results of our screening and subsequent experimental validation allowed us to identify mammalian genes responsible for HTMLA activity in the carnitine biosynthesis pathway.

Results

One substrate-many enzymes screening (OSMES) for PLP-dependent enzymes

Here we develop an automated procedure to perform a reverse docking screening of a substrate containing a primary amino group bound to PLP cofactor as a Schiff base (external aldimine; substrate), against a set of 3D enzyme structures of a selected PLPome (enzyme set) (**Fig. 1**).

As an enzyme set we used PLPomes of *Homo sapiens* and *Mus musculus* retrieved from the B6 database (B6DB; http://bioinformatics.unipr.it/B6db) composed of 56 and 57 genes respectively. For each RefSeq accession number we obtained the corresponding UniProt ID to download the AlphaFold monomer [28] and mark the position of the catalytic lysine useful for subsequent steps. In this first step, we discarded genes without a conserved catalytic lysine, namely AZIN1, AZIN2, SPTLC1 and PDXDC1 in both sets and Ldc1 in the mouse set, obtaining 105 enzyme targets for our analysis (**Supplementary Table 1**). The vast majority of our targets have AlphaFold models of very high confidence (pLDDT>90 over 90% of residues) for the overall (>80%) and active site (>95%) residues (**Supplementary Fig. 1**).

Since most PLP-dependent enzymes belong to fold-type I, which is characterized by obligate dimeric association forming two identical active sites at the interface, we exploited models available in the SWISS-MODEL Repository (SMR; https://swissmodel.expasy.org/repository) as template to assemble AlphaFold monomers into oligomeric structures (**Fig. 1,** step 2). In our set of enzymes, 96 structures were modeled as oligomers, mostly homomers (79 dimers, 13 tetramers) with the exception of SPTLC2 and SPTLC3, which were modeled as hetero-dimers, both associated with SPTLC1. Once the enzyme set is prepared, the procedure automatically builds the covalent adduct between PLP and a substrate molecule with a given PubChem ID, and creates a 3D coordinate file of the external aldimine for docking screening (**Fig. 1,** step 3). For each enzyme structure, the grid center for docking calculation is positioned at the NZ atom of the catalytic lysine, and the grid size is defined according to the size of the substrate (**Fig. 1,** step 4) (see **Methods**).

As a final step (**Fig. 1,** step 5) the pipeline runs the docking analysis of the substrate against each enzyme structure with AutoDock for Flexible Receptors (ADFR)[35], choosing as flexible residue the same catalytic lysine used to place the grid. The results of the screening are then parsed to rank targets according to different methods (see below).

Evaluation of catalytically favorable conformations is the best performing metric in OSMES

Before proceeding with OSMES to our case study, we assessed the ability of different ranking methods to identify enzymes involved in particular PLP-dependent reactions. We considered 13 different substrates (**Supplemetary Fig. 2**) against the two PLPomes (human and murine) for a total of 26 screenings evaluated with 7 ranking methods (**Fig. 2**). In each screening, one or more positive controls represented by enzymes known to catalyze the examined reaction (validation set) were considered. The validation set consisted of a total of 42 positive controls divided into 14 decarboxylases, 6 aldolases, 14 aminotransferases and 8 other reactions encompassing 4 ammonia-lyases, 2 γ-lyases, and 2 hydrolases (**Supplementary Table 2**).

Among the pose clusters obtained from ADFR analysis, we considered both the lowest-energy cluster (representing the energetically favored cluster; best cluster, BC) and the most populated cluster (representing the statistically favored cluster; largest cluster, LC) (**Fig. 2a**). For both BC and LC, we ranked the results using three different ranking methods: i) the number of *conformations* in the cluster (BCC and LCC); ii) the lowest binding *energy* of the cluster conformations (BCE and LCE); and, to discount the contribution on the constant moiety of the external aldimine, iii) the lowest binding energy of the cluster conformations without the PLP atoms, considering only the *amino acid* (BCaaE and LCaaE).

In addition to these more canonical criteria, we introduced a ranking method that evaluates the number of catalytically favorable conformations (CFC) based on Dunathan's stereoelectronic hypothesis [22]. According to this widely accepted feature of PLP catalysis, when a compound containing a primary amine group binds covalently to the PLP cofactor to form the external aldimine, the reaction proceeds by breaking the bond more parallel to the π orbitals of the cofactor pyrimidine ring, or in other words, more orthogonal to the plane formed by the latter. In the case of an α-amino acid, three different cases are possible (**Fig. 2b, c**), represented by the breaking of the Cα-COOH (as in decarboxylases), Cα-Cβ (as in aldolases), and Cα-Hα bond (as in racemases, aminotransferases, and other lyases). On this basis, for every substrate in our screenings we considered CFC conformations in which the angle (χ) with the PLP ring is maximum for the bond cleaved during the reaction (χ_1 for Cα-COOH; χ_2 for Cα-Cβ; χ_3 for Cα-Hα; **Fig. 2b, c**). As an additional condition for a CFC, we set an upper threshold of 5 Å for the distance between the NZ atom of catalytic lysine and the imine carbon of external aldimine (**Fig. 2c**). The cluster with the maximum number of CFC is considered the “catalytic cluster” (CC) and scored by the number of CFC it contains (CC-CFC) (**Fig. 2d**).

The distribution of the validation test ranked with the 7 different ranking methods shows that with the CC-CFC method the positive controls are generally ranked higher than with other methods (**Fig. 2e**). Within the CC-CFC distribution, a difference in the performance emerged by categorizing positive controls according to the reaction type, with aminotransferases (A) achieving worse results with respect to other reactions (O) that break the Cα-Hα bond or aldolases (B) and decarboxylases (D) (**Supplementary Fig. 3a, b**). The good performance obtained by CC-CFC is supported by the area under the receiver operating characteristic (AUROC) that confirms the CC-CFC as the most performing ranking method, with an AUROC=0.84 compared with 0.7 of LCE, the second best method (**Fig. 2f**).

Application of OSMES to the identification of a missing gene in carnitine biosynthesis

Carnitine biosynthesis begins with release of N^6-trimethyllysine (TML) from the breakdown of post-translationally modified proteins such as histones, calmodulin, cytochrome c, myosin, etc. [36,37], and involves four enzymatic steps (**Fig. 3a**). Reactions 1 and 4 are catalyzed by two Fe^{2+}-dependent dioxygenases: TML dioxygenase (TMLD) and γ-butyrobetaine dioxygenase (BBD), which are related by homology; reaction 3 is catalyzed by trimethylamino butyraldehyde dehydrogenase (TMABADH); reaction 2, the aldol cleavage of HTML to generate glycine and TMABA, is catalyzed by HTMLA. Although there is evidence that this activity requires PLP [38,39], the molecular identity of HTMLA in mammals and other metazoans is unknown.

The pathway described above is not universally present in eukaryotes. For instance, it lacks in species such as the yeast *Saccharomyces cerevisiae* and the darkling beetle *Tenebrio molitor,* which require an external supply of carnitine for fat metabolism [40,41]. The distribution of the genes encoding TMLD and BBD in eukaryotes (**Supplementary Fig. 4**), shows that the known pathway for carnitine biosynthesis is especially present in opisthokonts (fungi and metazoa). However, absence of TMLD and/or BBD in several species, particularly in protostomes, suggests multiple pathway losses, a suitable condition for the identification of missing genes by coevolutionary analysis. This analysis, conducted with a sensitive method of gene coevolution [42] in 1,952 eukaryotic genomes [43] did not reveal an obvious HTMLA candidate, although the best signal among PLP-dependent enzymes was found for an orthogroup annotated as threonine aldolase (**Supplementary Table 3**). Interestingly, a gene belonging to this group has been previously implicated in *Candida albicans* as HTMLA [30]. A gene homologous to threonine aldolase (*Tha1*) is found in several mammals including mice, but not in humans [44] nor in other species capable of synthesizing carnitine (**Supplementary Fig. 4**).

Since homology and coevolutionary analysis provided inconclusive evidence on the identification of mammalian HTMLA, we decided to use OSMES on the full set of PLP-dependent enzymes of human and mouse to identify candidates on a structural basis. To this end, we modeled the external aldimine PLP-HTML complex assuming free rotations around rotatable bonds (**Fig. 3b**) and defined the condition for catalytically favorable conformations of the docked substrate (**Fig. 3c**): a distance of ≤ 5 Å of the PLP aldehyde carbon of the substrate from the NZ atom of catalytic lysine and a relative maximum for the χ_2 angle, as expected for the cleavage between Cα-Cβ that occurs in the HTMLA reaction (**Fig. 3a**).

HTMLA candidates revealed by OSMES in the human and mouse PLPome

The best performing method (CC-CFC) was used to rank the results of HTML-OSMES against human and murine PLPomes (**Fig. 4a**). In the two rankings orthologous enzymes are in similar positions, as confirmed by the correlation between the two sets (Spearman r = 0.83; **Supplementary Fig. 5**).

In both rankings, the first hit is the cytosolic serine hydroxymethyltransferase (SHMT1; Shmt1); its mitochondrial version (SHMT2; Shmt2) ranks just after in second (human) and third (mouse) position, as expected from the strong conservation of active sites residues (**Supplementary Fig. 6**). Interestingly, it has been shown that *E.coli* SHMT can act as an aldolase on β-hydroxylated amino acids, especially with *erythro* configuration [45] that is the configuration adopted by HTML, and it has been proposed that SHMT could be responsible for HTMLA activity in mammals [31]. Descending with the ranking, other potential candidates with tested or predicted aldolase activity and belonging to the same KEGG Reaction Classes as HTMLA (RC00312 and RC00721) are found. These are sphingosine phosphate lyase (SGPL1, Sgpl1; EC: 4.1.2.27), an enzyme anchored to endoplasmic reticulum that catalyzes aldol cleavage forming phosphoethanolamine, and the putative mouse L-threonine aldolase (Tha1), not characterized experimentally but traceable by homology to the yeast low specificity L-threonine aldolase (GLY1, EC: 4.1.2.48). A GLY1 paralog has been genetically characterized as HTMLA in *C. albicans* [30]. Another example of promising candidates is the pair of paralogous enzymes called kynurenine aminotransferases (KYAT1, Kyat1, KYAT3, Kyat3). These enzymes catalyze the transamination of kynurenine into the corresponding α-keto acid. However, they are also able to catalyze β-lyase reactions toward cysteine-S-conjugate substrates (EC: 4.4.1.13), although the reaction mechanism involves deamination unlike HTMLA [46].

In the catalytic clusters of all the mentioned candidates, ADFR is able to position the PLP cofactor in a binding mode similar to that observed in the available experimental structures of homologous enzymes in complex with PLP (**Supplementary Fig. 7**). In all four SHMTs and in Tha1, the lowest-energy conformations of HTML-PLP in the catalytic cluster have the Cα-Cβ bond more perpendicular than in the other enzymes (**Fig. 4b, Supplementary Fig. 8**). By contrast, in the case of both SGPL1 and Sgpl1 (**Supplementary Fig. 8**), and all KYATs (**Fig. 4b, Supplementary Fig. 8**), the Cα-COOH ($\chi_3 < \chi_1 > \chi_2$) and Cα-Hα ($\chi_1 < \chi_3 > \chi_2$) bond, respectively, are the most perpendicular and therefore in an unfavorable conformation for aldol cleavage.

In all four KYATs and Tha1, visual inspection of the docked complexes revealed the presence of an aromatic cage (**Fig. 4b, Supplementary Fig. 8**), characteristic of proteins that bind N-trimethylated substrates, establishing hydrophobic and cation-π interactions with the trimethyl ammonium group [47]. The constant presence of a quaternary amine group in the intermediates of carnitine biosynthesis (**Fig. 3a**), suggests that an aromatic cage could be a structural feature of all enzymes of the pathway, as evidenced by the BBD structure in complex with γ-butyrobetaine [48], the conservation of the corresponding residues in its homologue TMLD, and the binding mode predicted by docking of the substrate in the TMABADH active site (**Supplementary Fig. 9**).

Biochemical validation of HTML-OSMES candidates

For the above reasons, screening candidates KYAT1, SGPL1, SHMT1 and SHMT2 from *Homo sapiens*, and Kyat3 and Tha1 from *Mus musculus* were chosen for the experimental validation. Each protein was produced using optimized conditions in recombinant form to be assayed for HTMLA activity. Recombinant SHMT1, SHMT2, KYAT1, Kyat3 were obtained in pure and soluble form after overexpression in *E. coli* (**Supplementary Fig. 10a-d;** insets). In order to obtain recombinant SGPL1 and Tha1 in the soluble form (**Supplementary Fig. 10e,f;** insets), they were co-expressed with chaperones (GroEL/GroES) as truncated forms without the N-terminal membrane anchor and mitochondrial signal (**Supplementary Fig. 11**; see **Methods**). All the enzymes showed the typical spectrum of protein-bound pyridoxal phosphate with a peak around 400-430 nm (**Supplementary Fig. 10**).

Stereospecific (*2S,3S*) HTML for the activity assays was obtained enzymatically from chemically-synthesized TML (see **Methods**) by exploiting the first reaction of the pathway (**Supplementary Fig. 12**). The activity assays show that SHMT1, SHMT2 and Tha1 catalyze the aldol cleavage of HTML; on the contrary, KYAT1, Kyat3, and SGPL1 are catalytically inactive towards HTML (**Fig. 5**).

Human SHMTs catalyze the aldol cleavage of HTML

In the ^{1}H NMR spectrum of HTML after addition of SHMT1, the increase of a singlet at 3.55 ppm corresponding to glycine α-protons is visible (**Fig. 5a**), clearly appearing after 60 minutes of reaction. TMABA formation is confirmed by 2 distinctive signals at 9.63 ppm and 5.05 ppm of the carbonyl proton and its hydrated form (geminal diol), respectively (**Supplementary Fig. 13a**).

Kinetic characterization of HTML cleavage catalyzed by SHMT1, carried out by a continuous spectrophotometric coupled assay that exploits NAD$^+$ reduction signal at 340 nm in the presence of the third enzyme of the pathway (TMABADH), shows a dependence of the initial velocities on substrate concentrations following Michaelis-Menten kinetics (**Fig. 5b**). The fitting of data to the Michaelis-Menten equation reveals a kinetic efficiency (k_{cat}/K_m) of 32.17 ± 5.34 s^{-1} M^{-1} (**Supplementary Table 4**). We also characterized the enzymatic activity of SHMT2 by spectrophotometric assay (**Supplementary Fig. 14g**), and measured a lower kinetic efficiency (6.23 ± 1.26 s^{-1} M^{-1}) compared to SHMT1 (**Fig. 5c**). In fact, despite a lower K_m (0.80 ± 0.16 mM vs 3.79 ± 0.44 mM) SHMT2 is penalized by a worse k_{cat} (0.005 ± 0.000 s^{-1} vs 0.122 ± 0.006 s^{-1}).

Mouse threonine aldolase (Tha1) shows higher HTMLA activity than human SHMTs

The ^{1}H NMR spectrum of HTML after the addition of Tha1, shows peaks with the same chemical shift observed in the reaction with SHMT1, but in higher quantities (**Supplementary Fig. 13b**), suggesting the same enzymatic activity, but a different efficiency for the two enzymes. A small upfield shift is visible in the main peak of the trimethylated ammonium protons at 3.11 ppm (**Supplementary Fig. 13b**).

Kinetic characterization of Tha1 by the same spectrophotometric assay as SHMT1, and fitting to the Michaelis-Menten equation (**Supplementary Fig. 14a**) resulted in a k_{cat} of 2.311 ± 0.029 s^{-1} and K_m of 0.169 ± 0.009 mM. Comparison with SHMT1 shows better values for both Tha1 constants and a k_{cat}/K_m (1.36 x 10^4 s^{-1} M^{-1}) about a thousand times greater (**Fig. 5c**). To test the substrate specificity of Tha1, we evaluated the activity of the enzyme with other β-hydroxylated amino acids: L-threonine and L-*allo*-threonine (**Fig. 5d**). The enzyme showed activity on both L-threonine and L-*allo*-threonine, but not with the D-enantiomers. However, the preferred substrate of Tha1 is HTML with a catalytic efficiency in the order of 10^4 s^{-1} M^{-1}, followed by L-*allo*-threonine (10^2 s^{-1} M^{-1}) and L-threonine (10^1 s^{-1} M^{-1}) (**Fig. 5e**). These results suggest that Tha1 has a catalytic preference for β-hydroxylated L-amino acids with the *erythro* configuration. With respect to L-*allo*-threonine, the reaction with HTML has a similar k_{cat} but a 50-fold lower K_m (**Supplementary Fig. 14a, b**), suggesting a higher affinity for the intermediate of the carnitine pathway. The two human SHMTs have a similar a preference for substrates with the *erythro* (*S*,*S*) configuration, but are much more efficient with L-*allo*-threonine (~10^4 for SHMT1, ~10^2 for SHMT2) than with HTML (**Fig. 5e;**

Supplementary Fig. 14d, e; Supplementary Table 4), which possesses a bulkier side chain (**Fig. 5d**).

To verify if the preference of Tha1 for the HTML substrate is a feature of threonine aldolase proteins of organisms with the carnitine biosynthesis pathway, we tested the activity of the low-specificity threonine aldolase *e*TA [49] from *E. coli*, which, like other bacteria, does not have carnitine biosynthesis. Recombinant *e*TA was produced in intact form in the homologous host. Characterization of its catalytic efficiency for L-*allo*-threonine and HTML, showed high activity with both substrates with a slight preference for L-*allo*-threonine (**Supplementary Fig. 14f, h**).

HTML is a competitive inhibitor of KYAT1

Although KYAT1 is unable to catalyze the aldol cleavage on HTML, the good binding energies obtained with the screening suggest potential binding at the active site. We thus wanted to test if HTML can inhibit KYAT1 activity on L-kynurenine.

In the presence of an α-keto acid, L-kynurenine is converted by KYAT1 to the corresponding keto acid (4-(2-aminophenyl)-2,4-dioxobutanoate), which rapidly cyclizes to kynurenic acid (**Supplementary Fig. 15a**). By measuring the spectrophotometric signal at 310 nm of the final product, we were able to observe the progress of the reaction in the absence and in the presence of HTML (**Supplementary Fig. 15b, c**). After the addition of 0.5 mM of HTML to the reaction mixture, a slowdown of the reaction is observed (**Supplementary Fig. 15c**), suggesting an inhibitory action. We characterized the initial velocity of kynurenine transamination with increasing concentrations of HTML. The Lineweaver-Burk double reciprocal primary plot shows a family of straight lines intersecting on the y axis, typical of competitive inhibition with a constant V_{max} and an increasing apparent K_m (**Fig. 5f**). A K_i value of 4 mM was determined by the secondary plot (**Supplementary Fig. 15d**).

Crystal structure of mouse Tha1 improves HTML-OSMES results

Although the AlphaFold models in our screening are of high quality overall, there is a disparity in the dataset as evidenced by the different RMSD (root-mean-square deviations) with respect to the templates used for oligomer reconstruction (**Supplementary Fig. 16**). These differences depend on the availability of experimental structures from the same or closely related species. For instance, in the case of KYAT, SGPL, and SHMT, PDB structures are available from various mammals, including humans [50–53] and mouse [54,55], whereas in the case of Tha1, only PDB structures from distant bacterial homologs are available [49,56]. To verify if the results of our screening for Tha1 are confirmed or improved with the availability of an experimental structure, we decided to determine the crystal structure of mouse Tha1.

Mouse Tha1 crystallizes in two space groups, in orthorhombic F222 and in monoclinic C2, with one molecule and two molecules in the ASU, respectively. The PLP cofactor is visible only in the monoclinic structure; however, the active site is very similar in the two cases, with only minor differences. The expected tetrameric quaternary structure is formed by crystallographic symmetries, with four identical units in F222 (related by a 222 symmetry) and two identical dimers in C2 (related by a two-fold axis). SEC-SAXS experiments confirmed the presence of a single component with a MW compatible to that of the sum of 4 units, indicating that the mouse enzyme, despite the lower stability (see **Supplementary Discussion 1** for details), is tetrameric in solution (**Supplementary Fig. 17b**). The RMSD values between the single units (around 0.26-0.28 Å, **Supplementary Table 5**) indicate that the monoclinic and orthorhombic structures are similar. Also the tetrameric assembly is conserved in the two space groups, with two main interfaces (**Fig. 6a**). As indicate by data from PISA analysis (**Supplementary Table 6**), the interface between units A and B (analogous to that between units C and D, termed "main interface") is contributing stronger to the stability of the quaternary structure in comparison with the interface between units A and C (analogous to that between units B and D, termed "secondary interface"). Hence, the tetramer can be considered a dimer (AB+CD) of dimers (A+B and C+D), with the first dissociation being ABCD to AB+ CD (as determined by PISA). A comparison with the structure of the *Thermotoga maritima* threonine aldolase (PDB code 1M6S) returned RMSD values between 1.03 and 1.17 Å for the single units (**Supplementary Table 5**), indicating a significant structure difference even though the secondary structures and the whole quaternary assembly are conserved. The major structural difference is related to an insertion of 13 residues in Tha1 between positions 346-260, residues absent in the *T. maritima* threonine aldolase. In the two enzymes the position of the PLP cofactor is essentially conserved (**Fig. 6b**). Further details on the crystal structures are reported in the **Supplementary Discussion 1**.

We repeated the docking screening by including in the data set the crystallographic structure of Tha1. The HTML-OSMES results show an increase in CFCs compared with what was obtained with the AlphaFold model. In the catalytic cluster, there are 91 CC-CFC within it, compared with 51 in the previous analysis (**Fig. 6d, f; Supplementary Table 7**). By comparing the two structures, some differences are observed in the side chains of the substrate binding residues (**Fig. 6c**). There are minor differences in the chain containing the catalytic lysine (e.g. Arg372A), while differences in the position of the residues contributed by the other chains (Tyr168C, Tyr69B) are more pronounced, suggesting that they result mainly from subunit assembly. In general, it is observed that many more conformations of the entire docking analysis with the crystal structure have the relevant bond nearly perpendicular to the plane of the PLP (0°) (**Fig. 6e, g**), most of which have $|\sin(\chi_2)| \geq 0.95$ (gray area). The number of CC-CFC obtained by HTML-OSMES with the experimental structure would have allowed Tha1 to place second in the mouse ranking. Although to a lesser extent, an increase of CC-CFC value was also obtained by the AlphaFold model[57] built with the addition of the Tha1 experimental structure as template (**Supplementary Fig. 18; Supplementary Table 7**).

Discussion

The experimental characterization of genes and proteins is a severe bottleneck in biology. As the availability of high-quality structural models is now at the proteome scale, there is a need for computational methods able to exploit this information to advance the knowledge of biological functions. Here we show that a structure-based screening can sensitively identify proteins catalyzing a particular metabolic reaction and provide evidence for functional assignments independently from sequence-based methods.

Our OSMES procedure can be directly applied with modifications of the input parameters to the functional identification of proteins catalyzing a restricted subset of enzymatic (PLP-dependent) reactions. Nevertheless, PLP-dependent enzymes constitute a variegated subset of biocatalysts present in a variety of metabolic pathways, responsible for more than 300 distinct activities, about 10% of which without an assigned gene (http://www.kegg.jp/kegg/pathway.html; http://bioinformatics.unipr.it/B6db). By screening known enzyme-substrate combinations, we observed that a ranking based on catalytically favorable conformations performs best in identifying enzymes responsible for given PLP-dependent reactions. Interestingly, however, an acceptable performance (AUROC=0.7) was obtained even by scoring methods based on binding energy or statistical frequency, which are generally applicable to docking screening. On the other hand, although our CFC criterion is specific for enzymes using PLP as a cofactor, there are other classes of enzymes in which catalytically productive substrate conformations at the active site can be devised as for example in tryptophan tryptophylquinone (TTQ) and NAD-dependent dehydrogenases [58–60].

The application of OSMES to the identification of HTMLA candidates in the mammalian carnitine biosynthesis pathway provides a proof-of-concept of the ability of the screening to predict unknown enzyme-substrate associations on a structural basis. The two top-ranked candidates, SHMT1 and SHMT2, were found to be able to catalyze the HTMLA reaction with a measured catalytic efficiency of ~10^1 s^{-1} M^{-1} . However, the Tha1 candidate, which was found in the top-10 mouse ranking and is absent in humans, had a HTMLA catalytic efficiency (~1.4×10^4 s^{-1} M^{-1}) about 3 orders of magnitude higher. At variance with SHMT and other proteins of our set, experimental structures for Tha1 were only available for distant bacterial homologs. Interestingly, the Tha1 ranking in our screening greatly improved by using the crystal structure of the mouse protein or AlphaFold models built taking this information into account.

A surprising result of our experimental validation is that different genes could be responsible for the second step of carnitine biosynthesis in humans and mice. This conclusion, however, is in line with previous observations of greater HTMLA promiscuity than in other reactions of the pathway [31]. In fact, deletion of other genes of the pathway results in the inability of *C. albicans* to grow on fatty acids, whereas deletion of the *gly1* paralog *htmla* only reduces growth on this carbon source, and even the *htmla/gly1* double null strain shows residual growth [30]. Our conclusion that rodents possess a more efficient HTMLA enzyme than humans and other primates is also supported by the observations that administration of TML to humans results in minimal synthesis of carnitine [61], whereas when TML is given to rats, it is nearly entirely converted into carnitine [62]. Also in line with our results is the observation that the HTMLA activity in human tissues is the lowest amongst the enzymes of the pathway, and is mainly observed in the liver [63], where both SHMT1 and 2 are abundantly expressed (https://www.proteinatlas.org/search/shmt).

The *Tha1* phylogeny suggests that this gene has a monophyletic origin in eukaryotes and it has been duplicated only in recent branches of the eukaryotic tree (**Supplementary Fig. 19**). One of these duplications in Saccharomycetales gave rise to the paralogous gene characterized as *htmla* in *C. albicans*. However, most other fungi possess a single copy of the gene (*gly1*), which is probably responsible for HTMLA activity in fungi. On the other hand, the putative animal ortholog *Tha1* could be responsible for HTMLA activity in those animals in which the gene is present together with the other genes of the pathway. As we found that the mouse enzyme has a strong preference (1,000 folds) for HTML towards threonine, the name hydroxytrimethyllysine aldolase (*Htmla*) would be a better descriptor of the mouse gene. It should be noted, however, that orthologous genes are maintained in the budding yeast (YEL046C; *gly1*) and the beetle *Tenebrio molitor* (KAJ3617386) that are known not to produce carnitine [40,41], and in several insect species, such as ants, bees, and wasps, that should lack the biosynthetic pathway as deduced from the absence of TMLD and BBD (see **Supplementary Fig. 4**). This evidence suggests that Tha1 fulfills additional functional roles, as frequently observed in PLP-dependent enzymes [64,65]. As suggested by the enzyme *in vitro* activity (see **Fig. 5**), these additional functions could involve the aldol cleavage of β-hydroxylated L-amino acids with *erythro* configuration.

While *Tha1/Htmla* is present in the majority of eukaryotes, it has been independently lost in various groups of mammals. It is absent in marsupials and some orders of placentals such as Primates and Chiroptera (bats) (see **Supplementary Fig. 4**). The loss of a functional gene in marsupials is presumably ancient as no trace of the gene can be retrieved in their genome by a tblastn search, whereas is more recent in placentals where pseudogenes are readily identified in several species including humans [44] (**Supplementary Fig. 20**). The relatively frequent loss of the gene during mammalian evolution can be explained by sufficient supply of carnitine via the biosynthetic pathway ensured by the HTMLA activity of SHMT and/or sufficient exogenous supply of carnitine via the diet. Interestingly, pseudogenization of TMLD is also observed in bats (**Supplementary Fig. 20**), suggesting loss of carnitine biosynthesis in these species.

Carnitine supply is crucial for energy metabolism as it enables the transport of fatty acids into the mitochondria, where they are oxidized to generate ATP. Although not essential to the body's supply of carnitine, nutritional sources are very important in humans, with about 75% of total body carnitine originating from food sources [66]. The results of our study suggest that humans and some other mammals, having lost a gene coding for an enzyme with efficient HTMLA activity, may have a lower output from the biosynthetic pathway and a higher dietary requirement for carnitine.

Methods

Establishment of the human and mouse PLP enzyme set (step 1 and 2)

The PLPome of the considered organisms (*Homo sapiens* and *Mus musculus*) was obtained with the B6DB (http://bioinformatics.unipr.it/B6db/tmp/) *whole genome analysis* tool, and each RefSeq accession number was converted to the UniProt one with *OSMES.convert_ac*. The enzyme set was built with AlphaFold models downloaded from https://alphafold.ebi.ac.uk/ in monomeric form and then used for the construction of homo-oligomeric structure with the function *OSMES.build_homo_oligo* that use the *super* function of PyMOL (https://pymol.org/) to structurally superimpose the AF monomers to the best template retrieved by the SWISS-MODEL repository (https://swissmodel.expasy.org/repository).
The criteria to choose the best structure used as alignment template were as follows, in order of priority: database source (first PDB, then SWISS-MODEL), oligomeric state (first the template with the higher number of chain, excluding heteromers), structure resolution (Å or QMEAN). Oligomers for SPTLC2 (AC: P97363, O15270) and SPTLC3 (AC: Q8BG54, Q9NUV7) were built with the function *OSMES.build_oligo_manual* due to their heteromeric association both with the same subunit SPTLC1 (AC: O35704, O15269), obtaining the two heterodimers SPTLC2-SPTLC1 and SPTLC3-SPTLC1. Murine Nfs1 (AC: Q9Z1J3) was built manually with PyMOL with the human template due to the absence of a homodimer in the SWISS-MODEL repository. All the models obtained by the procedure were visually inspected with PyMOL.

Substrate preparation (step 3)

The substrates used for the validation of the procedure were selected to represent amino acids with different properties (negative, positive, hydrophobic, aromatic, etc). Amino acids were ligated with a covalent bond (imine) between their Nα groups and the aldehydic group of PLP in the external aldimine state. All the substrates used in the reverse docking screening were constructed with an automated process that retrieves the 3D coordinate file of a given PubChem ID (PID) from PubChem (https://pubchem.ncbi.nlm.nih.gov/), binds a user-selected N atom to the PLP provided in SMILE format (CC1=NC=C(C(=C1O)C)COP(=O)([O-])[O-]), adds hydrogens with the *dimoprhite_df* program (https://github.com/UnixJunkie/dimorphite_dl) assuming a pH of 7.4, creates the PDB file and performs energy minimization using the function *MMFFOptimizeMolecule* from *RDKit* (https://www.rdkit.org/) with MMFF94s as forcefield. Once the PDB file is obtained, charges are assigned according to Gasteiger and converted in the pdbqt format (e.g. *HTL_PLP.pdbqt*) with the *prepare_ligand* script of ADFRsuite. An additional txt (e.g. *HTL_PLP.txt*) file is generated that contains the atom ID for the plane, the bonds for the CFC calculation, and the grid box sizes for AGFR obtained during the substrate preparation (see below). The code for steps 1-3 is available in *OSMES.ipynb*.

Active site positioning and sizing (step 4)

The coordinates for grid positioning are obtained during the enzyme set preparation through i) retrieval of the position of the post translational modification (PTM) lysine in the UniProt database; ii) determination of corresponding the residue number in the PDB file through pairwise alignment (*OSMES.match_fasta_position*) of the UniProt and the PDB sequence converted in FASTA format (*OSMES.pdb2fasta*), and iii) retrieval of the coordinates of the catalytic lysine NZ atom, to be used as the center of the grid. The script output is a tab-separated file (e.g. *Homo_sapiens_coord.tsv*) containing for each pdb file, the coordinates to be used in the AGFR command and required by the OSMES procedure (*OSMES_submit.py*). The box is defined as a cube with a size (*S*) defined during the substrate preparation calculated based on the maximum distance (*maxDist*) among the substrate atoms in the 3D coordinate file, imposing a lower limit of 14 Å .

Reverse-docking procedure (step 5)

The reverse docking procedure uses the files prepared in the previous steps, i.e. a substrate in pdbqt format (*HTL_PLP.pdbqt*) with the corresponding txt file (*HTL_PLP.txt*) and a dataset of enzyme structures in PDB format with the corresponding coordinates file for the grid box center(*Homo_sapiens_coord.tsv*). These input files are defined in a configuration file (*OSMES.config*) along with other docking parameters. This procedure uses the ADFR suite from AutoDock (https://ccsb.scripps.edu/adfr/), to prepare the pdbqt file of the enzyme structure and to run AGFR and ADFR command for every active site in the set. For each enzyme, all the catalytic lysines of the different chains in the structure were used for docking calculations, e.g. for a tetrameric structure we ran 4 different docking analyses for every chain. For our procedure, we use 200 *nbRuns*, 50,000,000 *maxEvals*, 3 Å for *clusteringRMSDCutoff*, 300 *popSize* and 0.2 Å for *spacing*. The procedure was performed in the SkyLake node (4 INTEL XEON E5-6140 2.3GHz, 72 cores, and 384 Gb of RAM) of the HPC facility of the University of Parma. A OSMES analysis took about 17 hours for each organism considered (~100 active sites).

Ranking methods

For the classification of the results, 7 different ranking methods were considered: LCC, LCE, LCaaE, BCC, BCE, BCaaE and CC-CFC. LCC and BCC were obtained directly from the summary output file of ADFR command, and correspond to the number of conformations of the largest and best cluster respectively. LCE and BCE were obtained directly from the summary output file of ADFR command, and correspond to energy of the lowest-energy conformation of the largest and best cluster respectively. LCaaE and BCaaE are based on the non binding energies (VdW and electrostatic interactions) of the amino acid-related atoms, excluding those of PLP. These were obtained with *OSMES.calc_ade* that exploits the utility *ade.py* from ADFRsuite and considers only the atoms named with the three-letter code chosen for every substrate in *OSMES.config*. To define the CFC in the docking results, we took into account two specific conditions through *OSMES.calc_run*: i) the mean distance of the catalytic cluster between the NZ atom of the catalytic lysine and the imine carbon of external aldimine must be less than 5 Å and; ii) |sin(χ)| of the inspected bond should be highest compared to those of the other bonds, and it was calculated by *OSMES.angle_plane_line* with the following formula:

$$|sin(\chi)| = \left| \frac{Aa+Bb+Cc}{\sqrt{a^2+b^2+c^2}\sqrt{A^2+B^2+C^2}} \right|$$

where $Ax + By + Cz = D$ is the plane equation of PLP ring obtained with *OSMES.planeEq* and $ax + by = c$ is the line equation of the inspected bond obtained with *OSMES.lineEq*. $|\sin(\chi)|$ represents the angle between the bond (line) and the PLP ring (plane) and for this reason is [0;1], where 1 is the perfect orthogonality and 0 is the perfect parallelism. For each conformation are calculated 3 different $|\sin(\chi)|$ (i.e. χ_1, χ_2, χ_3) for Cα-COOH, Cα-Cβ and Cα-Hα bonds respectively. All the plots for the analysis of docking results have been obtained with python code available in the *OSMES_result.ipynb* notebook that use Pandas (https://pandas.pydata.org/), Matplotlib (https://matplotlib.org/) and Seaborn (https://seaborn.pydata.org/) libraries.

HTML synthesis

HTML was synthesized starting from TML through enzymatic conversion to the hydroxylated form. TML was obtained by chemical synthesis from (2S)-6-amino-2-{[(tert-butoxy)carbonyl]amino}hexanoic acid (purchased from FCH) using a previously described protocol [67].
For the enzymatic HTML synthesis, we recombinantly produced the TMLD enzyme according to a published protocol [68]. We prepared the reaction using triethanolamine instead of the phosphate buffer, which in the presence of Fe^{2+} ion, immediately precipitates. We prepared 100 mL of reaction mixture (α-keto glutarate 15 mM, ascorbate 5 mM, TML 5 mM, $FeSO_4$ 200 µM, TEA 20 mM, DTT 1 mM, TMLD 10 µM) in a flask agitated for 30 min at 37°C. We finally purified the reaction mixture, that contained the enzyme and the other molecules, with cation exchange chromatography, by exploiting the positively charged N-trimethyl group to isolate HTML from the other negatively charged molecules such as ascorbate, 2-oxoglutarate, and succinate. After reaching pH 5.0 with the addition of HCl, the solution was firstly deprived by the enzyme TMLD through a Vivaspin™ centrifugation, then the flow-through was loaded onto a 50 mL Superloop of ÄKTA pure system FPLC and purified using HiTrap 5 mL SP column. We used 0.2 M HCl to elute the molecule with a gradient of 7 CV. We followed the elution on 210 nm and the fractions of the corresponding peak and flow-through of the column were analyzed by NMR spectra using the setting described below.

Plasmid construction

For the construction of SGLP1 expression plasmid, the SGPL1 (NCBI GeneID: 8879) CDS sequence (XM_006718053.1) inserted into pET-28b vector was purchased from GenScript (USA Inc.). For the construction of Tha1 expression plasmid, the Tha1 (NCBI GeneID: 71776) CDS sequence (NM_027919.4) without the first 40 amino acids corresponding a predicted mitochondrial signal, inserted into pET-28b expression vector was purchased from GenScript (USA Inc.) The constructs were transformed by electroporation into *E. coli* BL21 with pGRO7 plasmid from Takara™ containing GroEL and GroES chaperons. The authenticity of all constructs was verified by sequence analysis.

Protein expression and purification

Human SHMT1 and SHMT2 were recombinantly expressed and purified as previously described [69]. The *E. coli* clone expressing recombinant rat TMABADH was obtained from Ronald J.A. Wanders (University of Amsterdam). TMABADH was recombinantly expressed and purified as previously described [61]. Tha1 and SGPL1 expression was performed by inoculating a single colony of every clone in a Liter of autoinducing LB broth obtained by adding 0.5 g/L glucose and 2 g/L lactose to standard LB medium. Cells were grown at 20°C for 16h after a pre-induction phase at 37°C for 8h. Cell pellets were resuspended in 50 mL of Lysis Buffer (50 mM NaH_2PO_4 pH 7.6, 150 mM NaCl, 20 µM PLP), sonicated (1s on/off alternatively at 40 W for 30 min) and centrifuged (14,000 rpm for 40 minutes at 4°C). Supernatant was loaded onto a 50 mL Superloop of ÄKTA pure system FPLC and purified by Affinity Chromatography (AC) using HisTrap 5 mL FF column. Proteins were washed with Washing Buffer (50 mM NaH_2PO_4 pH 7.4, 100 mM KCl, 10% glycerol, 500 mM sucrose, 20 mM $MgCl_2$, 5 mM ATP, 1 mM DTT) to rid of GroEL which would otherwise be found in the elutions (see lane W in **Supplementary Fig. 14e**); eluted with AC Elution Buffer (20 mM NaH_2PO_4 pH 7.6, 150 mM NaCl, 500 mM imidazole). Protein fractions were collected and concentrated by Vivaspin™ centrifugation for dialysis in a Storage Buffer (50 mM NaH_2PO_4 pH 7.6, 150 mM NaCl, 1 mM DTT, 5 µM PLP). Tha1 was further purified with a size exclusion chromatography (**Supplementary Fig. 17a**) using Superdex 200 10/300 Gl column in 50 mM TEA pH 7.6, 150 mM NaCl, 1 mM DTT for crystallization experiments. Human KYAT1 and mouse Kyat3 were recombinantly expressed and purified as previously described [70]. UV-Vis spectra were collected with JASCO V-750 spectrophotometer and plotted with python code available in **Source Data**. Molar extinction coefficients (ε) for protein quantification were calculated with ProtParam (https://web.expasy.org/protparam/) using the corresponding sequences.

Aldolase activity assays

The HTMLA activity was measured by coupling the aldol cleavage of HTML with the oxidation of TMABA by NAD^+ catalyzed by TMABADH. The rate of the reaction was calculated from the rate of appearance in absorbance at 340 nm, due to the formation of NADH, using a value of ε_{340} = 6,220 $cm^{-1}\cdot M^{-1}$. The reaction was carried out using 1 µM SHMT1 or 4 µM SHMT2 in 50 mM potassium phosphate, pH 7.5, containing 1 mM EDTA, 5 mM 2-mercaptoethanol, 0.5 mM NAD^+ and 8 µM TMABADH, at 37 °C. For Tha1 (1 µM) and *e*TA (0.1 µM) we kept the same conditions but a different temperature (30 °C) due to their low stability at 37°C. The rate of L-*allo*-threonine and threonine cleavage was measured by coupling the reaction with reduction of the product acetaldehyde by NADH and alcohol dehydrogenase [45]. With L-*allo*-threonine as substrate were reduced the amount of SHMT enzymes, 0.15 µM for SHMT1 and 1 µM for SHMT2.The rate of the reaction was calculated from the rate of disappearance in absorbance at 340 nm, due to NADH depletion. Initial velocities were collected in a quartz cuvette (*l* = 1 cm) with JASCO V-750 spectrophotometer and plotted with python code available in **Source Data**.

Inhibition characterization of HTML towards human KYAT1

The rate of aminotransferase activity of KYAT1 was measured with a previously described protocol [71]. Briefly, through a continuous spectrophotometric assay at 310 nm (Δε = 3,625 $M^{-1}\cdot cm^{-1}$) the increase of the signal was monitored due to the higher ε of kynurenic acid compared to that of kynurenine (4,674 $M^{-1}\cdot cm^{-1}$ and 1,049 $M^{-1}\cdot cm^{-1}$, respectively). The reaction mixture contains a saturating concentration of α-ketoglutarate as an acceptor of the amino group. Different concentrations of HTML (0, 2, 4, 8 mM) were used for the Lineweaver-Burk double reciprocal primary plot. Initial velocities were collected in a quartz cuvette (*l* = 0.1 cm) with JASCO V-750 spectrophotometer and plotted with SigmaPlot 14.0 available in **Source Data**.

NMR spectroscopy

^{1}H NMR spectra were acquired with a JEOL ECZ600R spectrometer in no spinning mode at 25°C. Samples were loaded in Wilmad ECONOMY NMR tubes, solved in 600 µL of H_2O:D_2O (9:1) with simple DANTE presat sequence for H_2O suppression. The reactions were monitored using a DANTE presat array sequence with periods of 300 s for 24h. NMR experiments were acquired in 50 mM NaH_2PO_4 pH 7.0 to avoid signals of organic buffers in ^{1}H NMR spectra. NMR spectra were processed and analyzed with MestReNova version 14.2.0 (Mestrelab Research).

Crystallization and data collection

A frozen aliquot of Tha1 was thawed in ice and concentrated to a final concentration of 3.1 mg/mL and cleared by centrifugation at 17000 g. To find the best crystallization conditions crystal screens Pact Premiere HT-96 and Morpheus HT-96 (Molecular Dimensions) were tested mixing 0.2 µL of protein with 0.2 µL of screen against a 40 µL reservoir in MRC 2-lens plates (SWISS-CI) with an Oryx Nano crystallization robot (Douglas instruments). Initial hits appeared in conditions G9 (0.2 M Potassium sodium tartrate tetrahydrate, 0.1 M Bis-Tris propane pH 7.5, 20% w/v PEG 3350), H5 (0.2 M sodium nitrate, 0.1 M Bis-Tris propane pH 8.5, 20% w/v PEG 3350) and H9 (0.2 M Potassium sodium tartrate tetrahydrate, 0.1 M Bis-Tris propane pH 8.5, 20% w/v PEG 3350) of the PACT Premiere HT-96 screen. The protein was finally crystallized using the sitting-drop vapor-diffusion method at 18 °C, mixing 1 µL of Tha1 solution with 1 µL of precipitant solution (0.2 M sodium nitrate, 0.1 M Bis-Tris propane pH 8.5, 20% w/v PEG 3350) and equilibrated against a 80 µL reservoir of precipitant solution in MRC Maxi 48-drops crystallization plates (SWISS-CI). Crystals appeared overnight and finished growing in less than 72 h after the crystallization drops were prepared. For data collections, crystals were fished from the drops and flash-cooled in liquid nitrogen.

SEC-SAXS measurements and analysis

The Tha1 sample was measured by SEC-SAXS at the ESRF bioSAXS beamline BM29, Grenoble, France. A volume of 250 µL of protein sample at 5.5 mg/mL was loaded on a Superdex 200 10/300 GL column (Cytiva) via a high-performance liquid chromatography device (HPLC, Shimadzu) attached directly to the sample-inlet valve of the BM29 sample changer. The sample was measured in buffer C at 20 °C. The column was equilibrated with at least 3 column volumes to obtain a stable background signal before measurement. All parameters for SAXS analysis are described in **Supplementary Table 8**. In the SEC-SAXS chromatogram, frames in the region of stable Rg were selected with CHROMIXS and averaged using PRIMUS to yield a single averaged frame per protein sample. Analysis of the overall parameters was carried out by PRIMUS from ATSAS 3.2.1 package [72]. The pair distance distribution functions, P(r), were used to calculate *ab initio* models in Px symmetry with DAMMIF/N. CRYSOL was used for evaluating and fitting the experimental scattering curve of Tha1 with the corresponding atomic structure solved in this study. Plot and protein model were generated using OriginPro 9.0 and UCSF Chimera software, respectively. SAXS data were deposited into the Small Angle Scattering Biological Data Bank (SASBDB) under accession numbers SASDSU8.

Structure determination, refinement and analysis

Data collections were performed at ESRF beamline ID23-2. Diffraction data integration and scaling were performed with XDS [73] and the DIALS data-processing package [74], data reduction and analysis with Aimless [75]. Structures were solved by Molecular Replacement with Phaser [76] from Phenix [77], using as search model the structure of L-*allo*-threonine aldolase from *Thermotoga maritima*, PDB code 1M6S. Two crystal forms were identified, in space group F222, with one molecule in the asymmetric unit (ASU), and in space group C2, with two molecules in the ASU. The final refined structures were obtained by alternating cycles of manual refinement with Coot [78]and automatic refinement with phenix.refine [79]. Statistics on data collection and refinement are reported in **Supplementary Table 9**. Interface analysis was performed using PISA [80].

Construction of Tha1 models using crystal structure

Different AlphaFold models are obtained by using the best Tha1 crystal structure (F222) as template using the Colab notebook from ColabFold [57]. By setting *pdb_100* to the parameter *template_mode* with the UniProt sequence of Tha1, the PDB templates used for the prediction were retrieved. The corresponding templates were manually downloaded from PDB and Tha1 crystal structure was added. Then changing to *custom* the *template_mode*, two different models were generated. The first by using as template all the structure retrieved by pdb100 mode plus Tha1 crystal structure (AF_F222) and the second by using as template only the Tha1 crystal structure (AF_only_F222).

Data availability

The datasets and computer code used in this study are available in GitHub at the address: https://github.com/lab83bio/OSMES. Crystal structures were deposited in the PDB with accession codes 8PUS and 8PUM for the orthorhombic (F222) and the monoclinic (C2) form, respectively. SAXS data were deposited into the Small Angle Scattering Biological Data Bank (SASBDB) under accession numbers SASDSU8. **Source data** are provided with this paper.

References

1. Li, X. *et al.* Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM. *Nat. Methods* **10**, 584–590 (2013).
2. Nakane, T. *et al.* Single-particle cryo-EM at atomic resolution. *Nature* **587**, 152–156 (2020).
3. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
4. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
5. Kustatscher, G. *et al.* Understudied proteins: opportunities and challenges for functional proteomics. *Nat. Methods* **19**, 774–779 (2022).
6. Rhee, K. Y., Jansen, R. S. & Grundner, C. Activity-based annotation: the emergence of systems biochemistry. *Trends Biochem. Sci.* **47**, 785–794 (2022).
7. Lee, D., Redfern, O. & Orengo, C. Predicting protein function from sequence and structure. *Nat. Rev. Mol. Cell Biol.* **8**, 995–1005 (2007).
8. Redfern, O. C., Dessailly, B. & Orengo, C. A. Exploring the structure and function paradigm. *Curr. Opin. Struct. Biol.* **18**, 394–402 (2008).
9. Gligorijević, V. *et al.* Structure-based protein function prediction using graph convolutional networks. *Nat. Commun.* **12**, 3168 (2021).
10. van Kempen, M. *et al.* Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* (2023) doi:10.1038/s41587-023-01773-0.
11. Rauer, C., Sen, N., Waman, V. P., Abbasian, M. & Orengo, C. A. Computational approaches to predict protein functional families and functional sites. *Curr. Opin. Struct. Biol.* **70**, 108–122 (2021).
12. Schenone, M., Dančík, V., Wagner, B. K. & Clemons, P. A. Target identification and mechanism of action in chemical biology and drug discovery. *Nat. Chem. Biol.* **9**, 232–240 (2013).

13. Bender, B. J. *et al.* A practical guide to large-scale docking. *Nat. Protoc.* **16**, 4799–4832 (2021).

14. Lee, A. & Kim, D. CRDS: Consensus Reverse Docking System for target fishing. *Bioinformatics* **36**, 959–960 (2020).

15. Wong, F. *et al.* Benchmarking AlphaFold -enabled molecular docking predictions for antibiotic discovery. *Mol. Syst. Biol.* **18**, e11081 (2022).

16. Scardino, V., Di Filippo, J. I. & Cavasotto, C. N. How good are AlphaFold models for docking-based virtual screening? *iScience* **26**, 105920 (2023).

17. Hermann, J. C. *et al.* Predicting Substrates by Docking High-Energy Intermediates to Enzyme Structures. *J. Am. Chem. Soc.* **128**, 15882–15891 (2006).

18. Udatha, D. B. R. K. G., Sugaya, N., Olsson, L. & Panagiotou, G. How well do the substrates KISS the enzyme? Molecular docking program selection for feruloyl esterases. *Sci. Rep.* **2**, 323 (2012).

19. Ramírez-Palacios, C., Wijma, H. J., Thallmair, S., Marrink, S. J. & Janssen, D. B. Computational Prediction of ω-Transaminase Specificity by a Combination of Docking and Molecular Dynamics Simulations. *J. Chem. Inf. Model.* **61**, 5569–5580 (2021).

20. Menger, F. M. Analysis of ground-state and transition-state effects in enzyme catalysis. *Biochemistry* **31**, 5368–5373 (1992).

21. Bruice, T. C. & Benkovic, S. J. Chemical Basis for Enzyme Catalysis. *Biochemistry* **39**, 6267–6274 (2000).

22. Dunathan, H. C. Conformation and reaction specificity in pyridoxal phosphate enzymes. *Proc. Natl. Acad. Sci.* **55**, 712–716 (1966).

23. Schneider, G., Käck, H. & Lindqvist, Y. The manifold of vitamin B6 dependent enzymes. *Structure* **8**, R1–R6 (2000).

24. Voet, D. & Voet, J. G. *Biochemistry*. (John Wiley & Sons, 2011).

25. Kessel, A. & Ben-Tal, N. *Introduction to Proteins: Structure, Function, and Motion*. (Chapman and Hall/CRC, 2018). doi:10.1201/9781315113876.

26. Green, M. L. & Karp, P. D. A Bayesian method for identifying missing enzymes in

predicted metabolic pathway databases. *BMC Bioinformatics* **5**, 76 (2004).

27. Karp, P. D. *et al.* Pathway Tools version 23.0 update: software for pathway/genome informatics and systems biology. *Brief. Bioinform.* **22**, 109–126 (2021).

28. Varadi, M. *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022).

29. Vaz, F. M. & Wanders, R. J. A. Carnitine biosynthesis in mammals. *Biochem. J.* **361**, 417–429 (2002).

30. Strijbis, K. *et al.* Identification and characterization of a complete carnitine biosynthesis pathway in *Candida albicans*. *FASEB J.* **23**, 2349–2359 (2009).

31. Strijbis, K., Vaz, F. M. & Distel, B. Enzymology of the carnitine biosynthesis pathway. *IUBMB Life* NA-NA (2010) doi:10.1002/iub.323.

32. Percudani, R. & Peracchi, A. The B6 database: A tool for the description and classification of vitamin B6-dependent enzymatic activities and of the corresponding protein families. *BMC Bioinformatics* **10**, 273 (2009).

33. Eliot, A. C. & Kirsch, J. F. Pyridoxal Phosphate Enzymes: Mechanistic, Structural, and Evolutionary Considerations. *Annu. Rev. Biochem.* **73**, 383–415 (2004).

34. Tramonti, A. *et al.* Characterization of the Escherichia coli pyridoxal 5′-phosphate homeostasis protein (YggS): Role of lysine residues in PLP binding and protein stability. *Protein Sci.* **31**, (2022).

35. Ravindranath, P. A., Forli, S., Goodsell, D. S., Olson, A. J. & Sanner, M. F. AutoDockFR: Advances in Protein-Ligand Docking with Explicitly Specified Binding Site Flexibility. *PLOS Comput. Biol.* **11**, e1004586 (2015).

36. Tanphaichitr, V., Horne, D. W. & Broquist, H. P. Lysine, a Precursor of Carnitine in the Rat. *J. Biol. Chem.* **246**, 6364–6366 (1971).

37. Huszar, G. Tissue-specific biosynthesis of ϵ-N-monomethyllysine and ϵ-N-trimethyllysine in skeletal and cardiac muscle myosin: A model for the cell-free study of

post-translational amino acid modifications in proteins. *J. Mol. Biol.* **94**, 311–326 (1975).

38. Dunn, W. A., Aronson, N. N. & Englard, S. The effects of 1-amino-D-proline on the production of carnitine from exogenous protein-bound trimethyllysine by the perfused rat liver. *J. Biol. Chem.* **257**, 7948–7951 (1982).

39. Cho, Y.-O. & Leklem, J. E. In Vivo Evidence for a Vitamin B-6 Requirement in Carnitine Synthesis. *J. Nutr.* **120**, 258–265 (1990).

40. Carter, H. E., Bhattacharyya, P. K., Weidman, K. R. & Franekel, G. The identity of vitamin BT with carnitine. *Arch. Biochem. Biophys.* **35**, 241–242 (1952).

41. Swiegers, J. H., Dippenaar, N., Pretorius, I. S. & Bauer, F. F. Carnitine-dependent metabolic activities inSaccharomyces cerevisiae: three carnitine acetyltransferases are essential in a carnitine-dependent strain. *Yeast* **18**, 585–595 (2001).

42. Dembech, E. *et al.* Identification of hidden associations among eukaryotic genes through statistical analysis of coevolutionary transitions. *Proc. Natl. Acad. Sci.* **120**, e2218329120 (2023).

43. Kuznetsov, D. *et al.* OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. *Nucleic Acids Res.* **51**, D445–D451 (2023).

44. Edgar, A. J. Mice have a transcribed L-threonine aldolase/GLY1 gene, but the human GLY1 gene is a non-processed pseudogene. *BMC Genomics* **6**, 32 (2005).

45. Contestabile, R. *et al.* ʟ -Threonine aldolase, serine hydroxymethyltransferase and fungal alanine racemase: A subgroup of strictly related enzymes specialized for different functions. *Eur. J. Biochem.* **268**, 6508–6525 (2001).

46. Molecular cloning and expression of a cDNA for human kidney cysteine conjugate *β* -lyase. *FEBS Lett.* **360**, 277–280 (1995).

47. Nagy, G. N. *et al.* Composite Aromatic Boxes for Enzymatic Transformations of Quaternary Ammonium Substrates. *Angew. Chem. Int. Ed.* **53**, 13471–13476 (2014).

48. Leung, I. K. H. *et al.* Structural and Mechanistic Studies on γ-Butyrobetaine Hydroxylase. *Chem. Biol.* **17**, 1316–1324 (2010).

49. Di Salvo, M. L. *et al.* On the catalytic mechanism and stereospecificity of *Escherichia coli*

L -threonine aldolase. *FEBS J.* **281**, 129–145 (2014).

50. Rossi, F., Han, Q., Li, J., Li, J. & Rizzi, M. Crystal Structure of Human Kynurenine Aminotransferase I. *J. Biol. Chem.* **279**, 50214–50220 (2004).
51. Weiler, S. *et al.* Orally Active 7-Substituted (4-Benzylphthalazin-1-yl)-2-methylpiperazin-1-yl]nicotinonitriles as Active-Site Inhibitors of Sphingosine 1-Phosphate Lyase for the Treatment of Multiple Sclerosis. *J. Med. Chem.* **57**, 5074–5084 (2014).
52. Ducker, G. S. *et al.* Human SHMT inhibitors reveal defective glycine import as a targetable metabolic vulnerability of diffuse large B-cell lymphoma. *Proc. Natl. Acad. Sci.* **114**, 11404–11409 (2017).
53. Giardina, G. *et al.* The catalytic activity of serine hydroxymethyltransferase is essential for *de novo* nuclear DTMP synthesis in lung cancer cells. *FEBS J.* **285**, 3238–3253 (2018).
54. Szebenyi, D. M. E., Liu, X., Kriksunov, I. A., Stover, P. J. & Thiel, D. J. Structure of a Murine Cytoplasmic Serine Hydroxymethyltransferase Quinonoid Ternary Complex: Evidence for Asymmetric Obligate Dimers. *Biochemistry* **39**, 13313–13323 (2000).
55. Wlodawer, A. *et al.* Detect, correct, retract: How to manage incorrect structural models. *FEBS J.* **285**, 444–466 (2018).
56. Kielkopf, C. L. & Burley, S. K. X-ray Structures of Threonine Aldolase Complexes: Structural Basis of Substrate Recognition ,. *Biochemistry* **41**, 11711–11720 (2002).
57. Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–682 (2022).
58. Agarwal, P. K., Webb, S. P. & Hammes-Schiffer, S. Computational Studies of the Mechanism for Proton and Hydride Transfer in Liver Alcohol Dehydrogenase. *J. Am. Chem. Soc.* **122**, 4803–4812 (2000).
59. Johannissen, L. O., Scrutton, N. S. & Sutcliffe, M. J. The enzyme aromatic amine dehydrogenase induces a substrate conformation crucial for promoting vibration that

significantly reduces the effective potential energy barrier to proton transfer. *J. R. Soc. Interface* **5**, 225–232 (2008).

60. Enugala, T. R., Morató, M. C., Kamerlin, S. C. L. & Widersten, M. The Role of Substrate-Coenzyme Crosstalk in Determining Turnover Rates in *Rhodococcus ruber* Alcohol Dehydrogenase. *ACS Catal.* **10**, 9115–9128 (2020).
61. Vaz, F. M., Ofman, R., Westinga, K., Back, J. W. & Wanders, R. J. A. Molecular and Biochemical Characterization of Rat ε-N-Trimethyllysine Hydroxylase, the First Enzyme of Carnitine Biosynthesis. *J. Biol. Chem.* **276**, 33512–33517 (2001).
62. Rebouche, C. J., Lehman, L. J. & Olson, L. ϵ-N-Trimethyllysine Availability Regulates the Rate of Carnitine Biosynthesis in the Growing Rat. *J. Nutr.* **116**, 751–759 (1986).
63. Rebouche, C. J. & Engel, A. G. Tissue distribution of carnitine biosynthetic enzymes in man. *Biochim. Biophys. Acta BBA - Gen. Subj.* **630**, 22–29 (1980).
64. Stříšovský, K. *et al.* Mouse brain serine racemase catalyzes specific elimination of L -serine to pyruvate. *FEBS Lett.* **535**, 44–48 (2003).
65. Soo, V. W. C., Yosaatmadja, Y., Squire, C. J. & Patrick, W. M. Mechanistic and Evolutionary Insights from the Reciprocal Promiscuity of Two Pyridoxal Phosphate-dependent Enzymes. *J. Biol. Chem.* **291**, 19873–19887 (2016).
66. Steiber, A. Carnitine: a nutritional, biosynthetic, and functional perspective. *Mol. Aspects Med.* **25**, 455–473 (2004).
67. Baba, R., Hori, Y., Mizukami, S. & Kikuchi, K. Development of a Fluorogenic Probe with a Transesterification Switch for Detection of Histone Deacetylase Activity. *J. Am. Chem. Soc.* **134**, 14310–14313 (2012).
68. Kazaks, A. *et al.* Expression and purification of active, stabilized trimethyllysine hydroxylase. *Protein Expr. Purif.* **104**, 1–6 (2014).
69. Tramonti, A. *et al.* Human Cytosolic and Mitochondrial Serine Hydroxymethyltransferase Isoforms in Comparison: Full Kinetic Characterization and Substrate Inhibition Properties. *Biochemistry* **57**, 6984–6996 (2018).
70. Donini, S. *et al.* Recombinant production of eight human cytosolic aminotransferases

and assessment of their potential involvement in glyoxylate metabolism. *Biochem. J.* **422**, 265–272 (2009).

71. Passera, E. *et al.* Human kynurenine aminotransferase II - reactivity with substrates and inhibitors: Reactivity of kynurenine aminotransferase. *FEBS J.* **278**, 1882–1900 (2011).

72. Manalastas-Cantos, K. *et al. ATSAS 3.0* : expanded functionality and new tools for small-angle scattering data analysis. *J. Appl. Crystallogr.* **54**, 343–355 (2021).

73. Kabsch, W. *XDS*. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 125–132 (2010).

74. Beilsten-Edmands, J. *et al.* Scaling diffraction data in the *DIALS* software package: algorithms and new approaches for multi-crystal scaling. *Acta Crystallogr. Sect. Struct. Biol.* **76**, 385–399 (2020).

75. Evans, P. R. & Murshudov, G. N. How good are my data and what is the resolution? *Acta Crystallogr. D Biol. Crystallogr.* **69**, 1204–1214 (2013).

76. McCoy, A. J. *et al. Phaser* crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).

77. Liebschner, D. *et al.* Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in *Phenix*. *Acta Crystallogr. Sect. Struct. Biol.* **75**, 861–877 (2019).

78. Emsley, P. & Cowtan, K. *Coot* : model-building tools for molecular graphics. *Acta Crystallogr. D Biol. Crystallogr.* **60**, 2126–2132 (2004).

79. Afonine, P. V. *et al.* Towards automated crystallographic structure refinement with *phenix.refine*. *Acta Crystallogr. D Biol. Crystallogr.* **68**, 352–367 (2012).

80. Krissinel, E. & Henrick, K. Inference of Macromolecular Assemblies from Crystalline State. *J. Mol. Biol.* **372**, 774–797 (2007).

Acknowledgements

We thank Alessandro Paiardini and Michel F. Sanner for discussion, Beatrice Cogliati, Domenico Acquotti, and Davide Cavazzini for technical assistance, Andris Kazaks (Latvian Biomedical Research and Study Centre) for the kind gift of the *E. coli* clone expressing recombinant human TMLD, and Ronald J.A. Wanders (University of Amsterdam) for the kind gift of the *E. coli* clone expressing recombinant rat TMABADH. This work benefited from the equipment and framework of the COMP-HUB and COMP-R initiatives, funded by the "Departments of Excellence" program of the Italian Ministry for University and Research (MIUR, 2018-2022 and MUR, 2023-2027), and from the High Performance Computing facility of the University of Parma, Italy. This work was supported by the Italian Ministry for Education, University and Research PRIN grant 2017483NH8 to R.P. and by Sapienza University of Rome grants RP11715C7CB4FED4, RM11916B51484C08, RM120172A76E4B78 and RM12117A610B653E to R.C. This work was also supported by a grant from Istituto Pasteur Italia – Fondazione Cenci Bolognetti (Research Grant "Anna Tramontano" 2020) to R.C.

Author Contributions

R.P. conceived the study. M.M., R.P., M.L.D.S., and R.C. designed the study. M.M. implemented the OSMES procedure. M.M., M.L.D.S., A.T., and E.Z. conducted protein purification and enzyme assays. E.F., G.G., and R.B. conducted protein crystallization and structural analysis. A.S. performed organic synthesis. A.P. and E.P. contributed critical resources and advice. M.M., R.P., and R.B. wrote the initial draft. All authors contributed to the manuscript.

Competing Interests

The authors declare no competing interests.

Figure Legends

Fig.1: One substrate-many enzymes screening (OSMES) workflow.

Scheme of OSMES. The pipeline consists of 5 main steps performed automatically: (i) AlphaFold monomeric models for selected proteins are retrieved; (ii) oligomeric structures are determined with SWISS-MODEL templates; (iii) the substrate is prepared for docking and (iv) used to determine the gridbox size at the active site; (v) finally, the pipeline performs docking analysis and the results are ranked using different methods.

Fig. 2: Evaluation of different ranking methods of OSMES with known substrates of PLP-enzymes.

a, Representation of the 6 ranking methods related to the best cluster (BC; red tones) and the largest cluster (LC; yellow tones). The bar plot represents the 200 conformations of a single docking run clustered with a 3 Å RMSD threshold; LCC and BCC methods consider the number of conformations in the respective cluster. The atoms of the substrate considered in the energy-based ranking methods (BCE, LCE, BCaaE, LCaaE) are highlighted in the insets. **b**, Scheme of the side view of the PLP pyrimidine ring and the three Cα bonds with the respective angles (χ) with respect to the PLP ring plane. **c**, Catalytically favorable conformations (CFC) in the three different PLP-dependent reactions. The conformations from docking analysis are considered CFC if distance (d) between N_ε of catalytic lysine and imine carbon ≤ 5 Å in the catalytic cluster, and the bond cleaved in the expected reaction (superior circumradius) is nearly orthogonal to the PLP ring (plane), that is its angle χ has the maximum relative value (see **Methods**). **d**, Bar plot highlighting in blue the number of CFC in different clusters. Black arrow indicates the Catalytic Cluster (CC) which does not always coincide with BC (red) or LC (yellow). **e**, Letter-value plot showing the distribution of the validation set colored according to the 7 ranking methods. BC related methods are colored in red tones; LC related methods are colored in yellow tones; CC-CFC is colored in blue. Individual dots representing ranking position of positive controls (i.e., enzymes known to act on the substrate) are colored according to substrate identity (legend); black dashed line delimits the top 10 positions. **f**, Receiver operating characteristic curve (ROC) for the different ranking methods colored as in panel **e**; the dotted diagonal represents an area under curve (AUROC) value of 0.5.

Fig. 3: HTMLA, the missing aldolase in animal carnitine biosynthesis.

a, Carnitine biosynthetic pathway in animals. HTMLA, the missing enzyme catalyzing the second step of the pathway is highlighted in yellow. **b**, Atomic model of the energy-minimized conformation of the HTML-PLP external aldimine used for the OSMES procedure. In yellow the carbon atoms of HTML, in white the carbon atoms of PLP. Non-carbon atoms are colored according to CPK convention. **c**, Expected geometry of the catalytic favorable conformation of the docked HTML-PLP substrate. The Cα-Cβ bond is considered labile when $\chi_1 < \chi_2 > \chi_3$ and d ≤ 5 Å.

Fig. 4: HTMLA candidates identified by HTML-OSMES in human and mouse.

a, HTML-OSMES against human and mouse PLPomes ranked with CC-CFC method. Best results (highest for LCC, CC-CFC, $|\sin(\chi_2)|$; lowest for E) in the columns are highlighted with darker colors. |sin(χ_2)|, d and E columns represent the mean values of CC. In orange are highlighted the enzymes with known β-lyase or aldolase activity. **b**, Structural representation of the lowest-energy binding modes among the catalytic clusters obtained by docking of HTML-PLP substrate for SHMT1, Tha1 and KYAT1. Non-carbon atoms are colored according to CPK convention. The conformations are shown with ball-and-sticks and are composed of PLP cofactor (magenta) covalently bound to HTML (yellow), and flexible catalytic lysine (green). The binding site residues (≤ 4.5 Å from HTML-PLP) are shown in lines labeled with one-letter code and number. Polar interactions between substrate and protein are indicated with orange dashes, while cation-π interactions are indicated with olive dashes.

Fig. 5: Experimental validation of HTML-OSMES candidates.

a, Time-resolved ^{1}H NMR spectra of SHMT1 activity in the presence of 5 mM HTML at 0, 35, 65 and 105 minutes. Cα protons singlet of glycine is assigned in the structure. **b**, Nonlinear fitting to the Michaelis Menten equation of the dependency on HTML concentrations of the initial reaction velocity of SHMT1 (1 μM). **c**, Kinetic parameters (k_{cat}, K_m, k_{cat}/K_m) of HTMLA reaction of tested enzymes with mean and standard error values. **d,** Scheme of the broken bond (magenta cross) in the aldol cleavage reactions of HTML, L-*allo*-threonine, L-threonine. In red are the portions common to the three substrates. **e**, Bar plot in log scale of k_{cat}/K_m of different enzymes (Tha1, SHMT1, SHMT2) with different substrates (HTML, L-*allo*-threonine, L-threonine). **f**, Lineweaver-Burk double-reciprocal primary plot of the inhibition by HTML of kynurenine aminotransferase activity of KYAT1. The kynurenine concentration ranged from 0.75 to 3 mM. The concentrations of HTML were 0, 2, 4, and 8 mM.

Fig. 6: Crystal structure of mouse Tha1 improves HTML-OSMES results.

a, Quaternary assembly of Tha1. The four PLP cofactors, one for each unit, are shown in violet ball-and-stick. The main interface, between units A and B (orange/magenta) and the secondary interface between units C and D (teal/pale cyan) are indicated. **b**, Active site of Tha1. The main polar interactions of the PLP cofactor (violet) at the interface between subunits A (carbon atoms in teal) and B (carbon atoms in pale cyan) are indicated. Distances in Å. **c**, Comparison of AlphaFold model (dark colors) and the Tha1 crystal structure (orthorhombic F222, light colors) docked with HTML-PLP substrate. Different chains are colored in different colors. Non-carbon atoms are colored according to CPK convention. HTML-PLP and flexible catalytic lysine are shown in ball-and-sticks. The binding site residues (≤ 4.5 Å from HTML-PLP) are shown in sticks. Polar interactions are indicated with orange dashes, cation-π interactions are indicated with olive dashes. **d-g** , Clustering of the HTML-PLP conformations at the Tha1 active site obtained with HTML-OSMES applied to AlphaFold model (**d, e**) or the crystal structure (**f, g**). Bar plots show the distribution of χ_1 (blue), χ_2 (emerald) and χ_3 (kiwi) angles in each cluster, with the catalytic cluster highlighted in light blue. Circular plots show the cumulative distribution of the three χ angles for all clusters. |sin(χ)| ≥ 0.95 values are defined by gray areas.

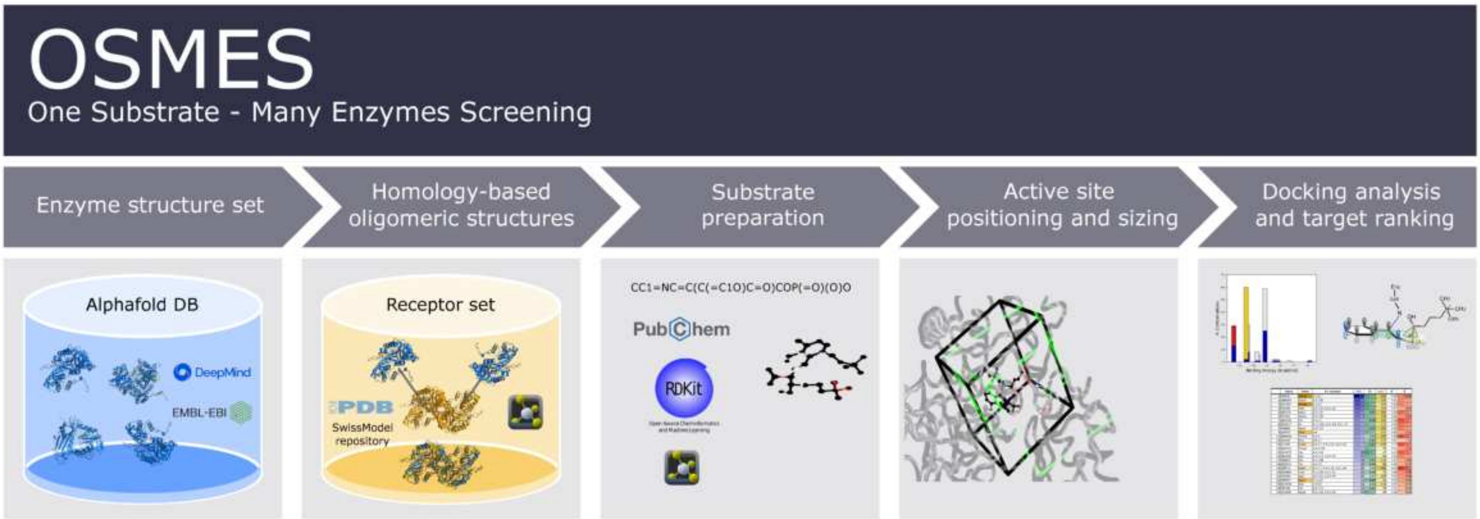

OSMES
One Substrate - Many Enzymes Screening
Enzyme structure set
Homology-based oligomeric structures
Substrate preparation
Active site positioning and sizing
Docking analysis and target ranking
Alphafold DB
DeepMind
EMBL-EBI
Receptor set
PDB
SwissModel repository
CC1=NC=C(C(=C1O)C=O)COP(=O)(O)O
PubChem
RDKit

Fig.1: One substrate-many enzymes screening (OSMES) workflow.

Scheme of OSMES. The pipeline consists of 5 main steps performed automatically: (i) AlphaFold monomeric models for selected proteins are retrieved; (ii) oligomeric structures are determined with SWISS -MODEL templates; (iii) the substrate is prepared for docking and (iv) used to determine the gridbox size at the active site; (v) finally, the pipeline performs docking analysis and the results are ranked using different methods .

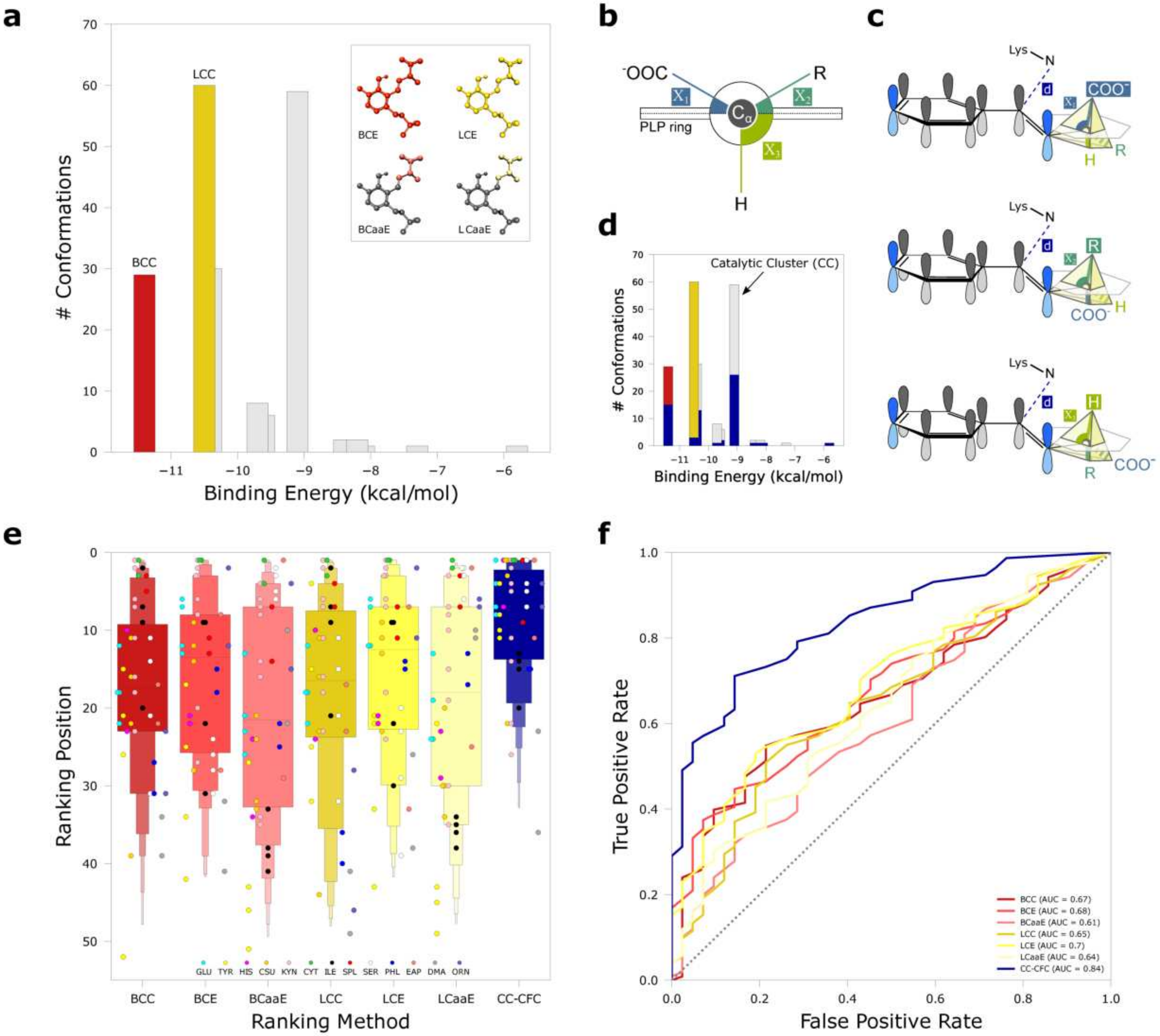


Fig. 2: Evaluation of different ranking methods of OSMES with known substrates of PLP-enzymes.

a, Representation of the 6 ranking methods related to the best cluster (BC; red tones) and the largest cluster (LC; yellow tones). The bar plot represents the 200 conformations of a single docking run clustered with a 3 Å RMSD threshold; LCC and BCC methods consider the number of conformations in the respective cluster. The atoms of the substrate considered in the energy-based ranking methods (BCE, LCE, BCaaE, LCaaE) are highlighted in the insets. **b**, Scheme of the side view of the PLP pyrimidine ring and the three Cα bonds with the respective angles (χ) with respect to the PLP ring plane. **c**, Catalytically favorable conformations (CFC) in the three different PLP-dependent reactions. The conformations from docking analysis are considered CFC if distance (d) between N_ε of catalytic lysine and imine carbon ≤ 5 Å in the catalytic cluster, and the bond cleaved in the expected reaction (superior circumradius) is nearly orthogonal to the PLP ring (plane), that is its angle χ has the maximum relative value (see **Methods**). **d**, Bar plot highlighting in blue the number of CFC in different clusters. Black arrow indicates the Catalytic Cluster (CC) which does not always coincide with BC (red) or LC (yellow). **e**, Letter-value plot showing the distribution of the validation set colored according to the 7 ranking methods. BC related methods are colored in red tones; LC related methods are colored in yellow tones; CC-CFC is colored in blue. Individual dots representing ranking position of positive controls (i.e., enzymes known to act on the substrate) are colored according to substrate identity (legend); black dashed line delimits the top 10 positions. **f**, Receiver operating characteristic curve (ROC) for the different ranking methods colored as in panel **e**; the dotted diagonal represents an area under curve (AUROC) value of 0.5.

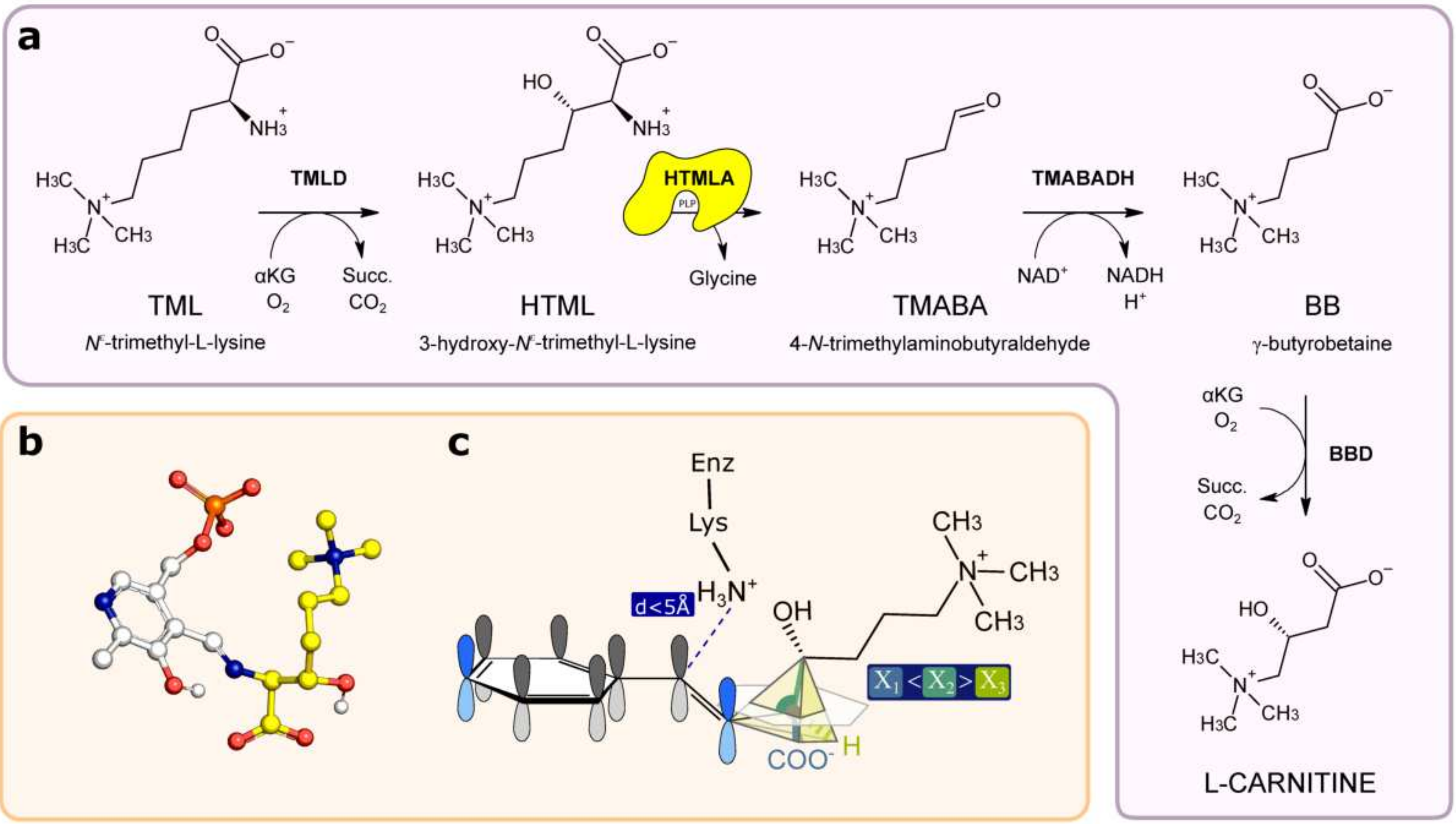


Fig. 3: HTMLA, the missing aldolase in animal carnitine biosynthesis .

a , Carnitine biosynthetic pathway in animals. HTMLA, the missing enzyme catalyzing the second step of the pathway is highlighted in yellow. **b**, Atomic model of the energy -minimized conformation of the HTML-PLP external aldimine used for the OSMES procedure. In yellow the carbon atoms of HTML, in white the carbon atoms of PLP. Non -carbon atoms are colored according to CPK conventio n. **c**, Expected geometry of the catalytic favorable conformation of the docked HTML-PLP substrate. The Cα-Cβ bond is considered labile when $\chi_1 < \chi_2 > \chi_3$ and d ≤ 5 Å.

a

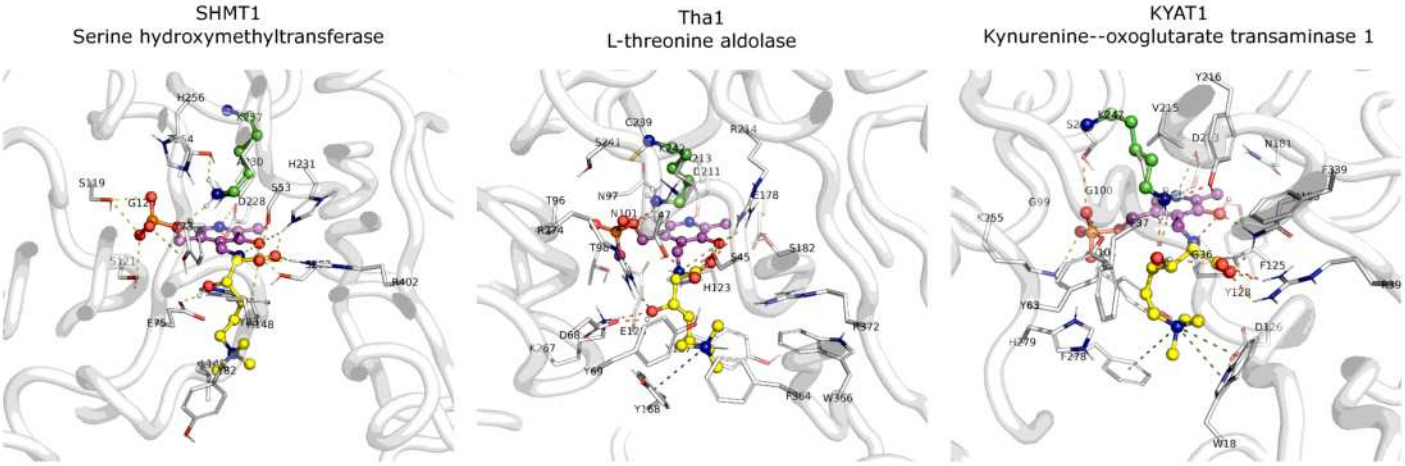
Homo sapiens

	Entry	Gene	EC number	CC-CFC	$\lvert\sin(\chi_2)\rvert$	LCC	d	E (kcal/mol)
1	P34896	SHMT1	2.1.2.1	116	0.81	153	3.74	-10.40
2	P34897	SHMT2	2.1.2.1	102	0.80	142	3.63	-11.10
3	P80404	ABAT	2.6.1.19; 2.6.1.22	75	0.71	142	3.95	-9.40
4	P20711	DDC	4.1.1.28	67	0.69	110	3.81	-8.30
5	P32929	CTH	4.4.1.1	63	0.69	182	3.82	-11.30
5	Q96I15	SCLY	4.4.1.16	63	0.70	144	3.66	-9.10
7	P13196	ALAS1	2.3.1.37	61	0.71	143	3.97	-10.70
8	Q6YP21	KYAT3	2.6.1.7; 4.4.1.13; 2.6.1.63	60	0.66	160	3.68	-12.10
9	Q9NUV7	SPTLC3	2.3.1.50	57	0.74	129	4.03	-10.30
10	O15270	SPTLC2	2.3.1.50	53	0.72	139	4.39	-10.20
11	Q8IUZ5	PHYKPL	4.2.3.134	46	0.72	96	4.79	-9.60
12	P22557	ALAS2	2.3.1.37	45	0.74	100	3.80	-9.00
12	Q9Y697	NFS1	2.8.1.7	45	0.65	89	4.24	-8.90
14	Q8N5Z0	AADAT	2.6.1.39; 2.6.1.7	43	0.57	128	4.18	-11.00
15	Q9GZT4	SRR	5.1.1.18; 4.3.1.18; 4.3.1.17	42	0.69	86	3.93	-10.40
16	P20132	SDS	4.3.1.17; 4.3.1.19	41	0.66	103	3.85	-10.90
17	Q86YJ6	THNSL2	4.2.3.-	39	0.51	175	4.11	-12.10
17	Q16773	KYAT1	2.6.1.7; 4.4.1.13; 2.6.1.64	39	0.59	148	4.00	-11.80
19	P21549	AGXT	2.6.1.51; 2.6.1.44	32	0.52	141	3.97	-11.80
19	P35520	CBS	4.2.1.22	32	0.76	55	4.13	-10.80
21	P04181	OAT	2.6.1.13	31	0.69	69	4.93	-8.60
21	P23378	GLDC	1.4.4.2	31	0.55	88	3.42	-8.50
21	P17174	GOT1	2.6.1.1; 2.6.1.3	31	0.63	82	4.03	-11.90
24	O95470	SGPL1	4.1.2.27	29	0.74	93	3.30	-9.90
25	Q96GA7	SDSL	4.3.1.19; 4.3.1.17	28	0.73	54	3.71	-10.30
26	P54687	BCAT1	2.6.1.42	26	0.57	93	3.74	-10.60
27	Q9Y600	CSAD	4.1.1.29; 4.1.1.11	24	0.63	67	3.82	-10.20
27	Q05329	GAD2	4.1.1.15	24	0.59	114	3.77	-10.30
27	Q99259	GAD1	4.1.1.15	24	0.57	86	3.83	-10.50
30	P11216	PYGB	2.4.1.1	23	0.78	42	4.01	-9.30
31	P19113	HDC	4.1.1.22	22	0.52	100	3.75	-9.50
31	Q6ZQY3	GADL1	4.1.1.11; 4.1.1.29	22	0.56	96	3.72	-10.30
31	Q96EN8	MOCOS	2.8.1.9	22	0.72	75	4.86	-10.60
34	P00505	GOT2	2.6.1.1; 2.6.1.7	19	0.56	91	3.81	-12.00
35	O15382	BCAT2	2.6.1.42	17	0.44	135	3.81	-11.70
35	P17735	TAT	2.6.1.5	17	0.60	62	4.25	-9.80
37	P06737	PYGL	2.4.1.1	15	0.78	131	3.65	-9.70
38	Q8TD30	GPT2	2.6.1.2	13	0.64	53	3.64	-11.00
39	Q8NHS2	GOT1L1	2.6.1.1	11	0.49	92	4.83	-8.20
39	Q9BYV1	AGXT2	2.6.1.44; 2.6.1.40	11	0.73	41	3.44	-8.00
41	Q9Y617	PSAT1	2.6.1.52	10	0.71	54	4.78	-9.40
42	O94903	PLPBP	N/A	9	0.41	69	4.77	-8.10
42	O75600	GCAT	2.3.1.29	9	0.80	116	3.77	-9.60
42	Q8TBG4	ETNPPL	4.2.3.2	9	0.62	54	3.45	-7.30
42	Q4AC99	ACCSL	N/A	9	0.43	57	3.82	-10.10
46	P11217	PYGM	2.4.1.1	6	0.82	137	4.18	-9.70
46	Q96QU6	ACCS	N/A	6	0.34	110	4.03	-10.50
48	P24298	GPT	2.6.1.2	3	0.40	106	3.67	-10.60
48	Q8IYQ7	THNSL1	N/A	3	0.40	117	4.18	-8.90
50	Q16719	KYNU	3.7.1.3	1	0.38	188	3.69	-9.30
51	P11926	ODC1	4.1.1.17	0	0.38	127	6.79	-8.20

Mus musculus

	Entry	Gene	EC number	CC-CFC	$\lvert\sin(\chi_2)\rvert$	LCC	d	E (kcal/mol)
1	P50431	Shmt1	2.1.2.1	113	0.81	152	3.95	-10.00
2	Q8BG54	Sptlc3	2.3.1.50	82	0.76	136	4.61	-10.10
3	Q9CZN7	Shmt2	2.1.2.1	77	0.80	103	3.92	-10.30
4	P61922	Abat	2.6.1.19; 2.6.1.22	71	0.79	104	3.97	-8.90
5	Q8VC19	Alas1	2.3.1.37	64	0.67	155	4.29	-11.40
6	P97363	Sptlc2	2.3.1.50	61	0.70	134	4.62	-9.50
7	Q8VCN5	Cth	4.4.1.1	53	0.68	140	3.44	-12.10
8	Q9QZX7	Srr	5.1.1.18; 4.3.1.18; 4.3.1.17	52	0.71	111	3.90	-11.00
9	Q6XPS7	Tha1	N/A	51	0.72	111	4.83	-9.10
9	P08680	Alas2	2.3.1.37	51	0.67	155	3.65	-10.60
11	Q80W22	Thnsl2	4.2.3.-	46	0.53	155	3.99	-12.30
12	Q9Z1J3	Nfs1	2.8.1.7	45	0.60	117	3.67	-7.60
12	Q71RI9	Kyat3	2.6.1.7; 4.4.1.13; 2.6.1.63	45	0.67	135	3.77	-12.10
14	Q8R1K4	Phykpl	4.2.3.134	42	0.70	76	4.61	-9.10
15	Q91WT9	Cbs	4.2.1.22	38	0.84	47	4.21	-11.70
16	Q8R238	Sdsl	4.3.1.17; 4.3.1.19	37	0.71	92	3.63	-10.30
17	O88533	Ddc	4.1.1.28	34	0.61	110	3.57	-9.00
18	Q8QZR1	Tat	2.6.1.5	28	0.62	88	4.45	-8.70
19	Q8BTY1	Kyat1	2.6.1.7; 4.4.1.13; 2.6.1.64	26	0.46	190	4.01	-13.00
19	Q9DBE0	Csad	4.1.1.29; 4.1.1.11	26	0.60	70	3.85	-9.80
21	O35423	Agxt	2.6.1.51; 2.6.1.44	25	0.47	109	4.39	-9.50
22	Q8R0X7	Sgpl1	4.1.2.27	23	0.68	83	3.34	-10.40
23	P29758	Oat	2.6.1.13	22	0.61	75	3.71	-8.40
23	Q91W43	Gldc	1.4.4.2	22	0.50	113	3.47	-8.70
23	Q3UEG6	Agxt2	2.6.1.44; 2.6.1.40	22	0.75	45	3.52	-8.40
26	Q8VBT2	Sds	4.3.1.17; 4.3.1.19	21	0.68	56	3.55	-10.10
26	Q9JLI6	Scly	4.4.1.16	21	0.65	76	4.10	-7.80
28	P48320	Gad2	4.1.1.15	20	0.60	110	3.76	-10.20
28	P05202	Got2	2.6.1.1; 2.6.1.7	20	0.49	80	4.17	-11.30
30	Q80WP8	Gadl1	4.1.1.11; 4.1.1.29	19	0.56	106	3.68	-11.10
31	O88986	Gcat	2.3.1.29	18	0.43	137	3.83	-10.40
31	P05201	Got1	2.6.1.1; 2.6.1.3	18	0.61	63	4.28	-10.90
33	Q8BH55	Thnsl1	N/A	17	0.65	97	4.58	-7.50
34	O35855	Bcat2	2.6.1.42	15	0.45	96	3.78	-11.90
34	Q3UX83	Accsl	N/A	15	0.77	29	3.55	-9.00
34	Q8BWU8	Etnppl	4.2.3.2	15	0.76	72	4.79	-10.80
34	P23738	Hdc	4.1.1.22	15	0.51	96	3.61	-9.40
38	Q9WVM8	Aadat	2.6.1.39; 2.6.1.7	14	0.33	185	4.18	-11.80
39	P48318	Gad1	4.1.1.15	13	0.52	95	3.66	-10.50
40	Q9ET01	Pygl	2.4.1.1	12	0.74	83	3.55	-10.40
40	Q8BGT5	Gpt2	2.6.1.2	12	0.59	58	4.23	-10.40
42	Q14CH1	Mocos	2.8.1.9	11	0.65	85	4.92	-9.00
42	Q8CI94	Pygb	2.4.1.1	11	0.80	86	3.85	-9.10
44	Q9WUB3	Pygm	2.4.1.1	9	0.75	108	4.02	-9.50
45	A2AIG8	Accs	N/A	7	0.46	69	4.00	-10.20
45	P24288	Bcat1	2.6.1.42	7	0.44	47	3.72	-9.50
47	Q99K85	Psat1	2.6.1.52	5	0.54	81	4.02	-9.70
47	Q7TSV6	Got1l1	2.6.1.1	5	0.55	77	4.78	-7.60
49	Q9CXF0	Kynu	3.7.1.3	2	0.47	145	3.61	-10.40
50	Q8QZR5	Gpt	2.6.1.2	1	0.38	66	3.95	-10.50
50	Q9Z2Y8	Plpbp	N/A	1	0.69	90	4.71	-5.80
52	P00860	Odc1	4.1.1.17	0	0.52	119	6.63	-11.60

b

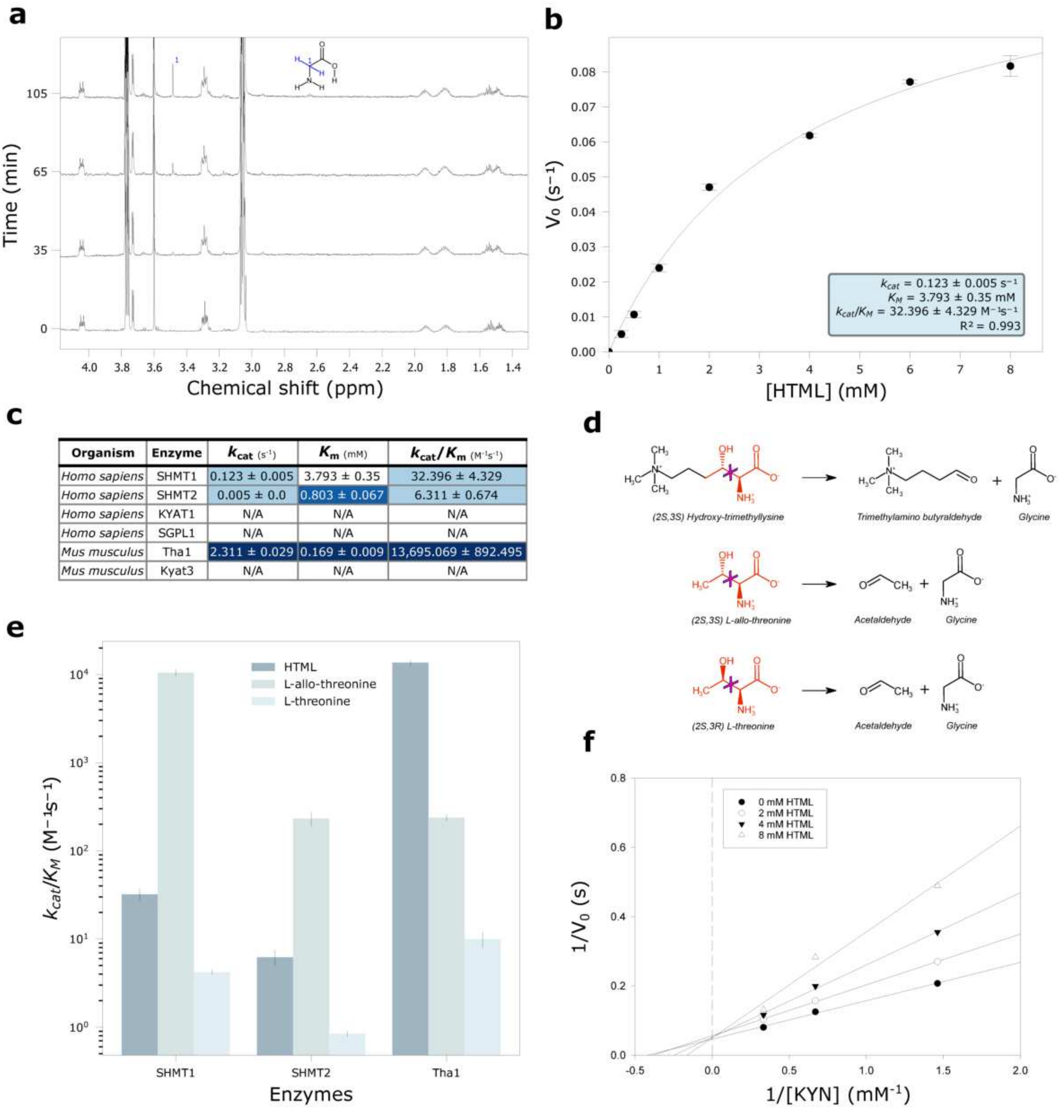

SHMT1
Serine hydroxymethyltransferase

Tha1
L-threonine aldolase

KYAT1
Kynurenine--oxoglutarate transaminase 1

Fig. 4: HTMLA candidates identified by HTML -OSMES in human and mouse.

a, HTML-OSMES against human and mouse PLPomes ranked with CC -CFC method. Best results (highest for LCC, CC-CFC, $|\sin(\chi_2)|$; lowest for E) in the columns are highlighted with darker colors. $|\sin(\chi_2)|$, d and E columns represent the mean values of CC. In orange are highlighted the enzymes with known β-lyase or aldolase activity. **b**, Structural representation of the lowest-energy binding modes among the catalytic clusters obtained by docking of HTML -PLP substrate for SHMT1, Tha1 and KYAT1. Non -carbon atoms are colored according to CPK convention. The conformations are shown with ball-and-sticks and are composed of PLP cofactor (magenta) covalently bound to HTML (yellow), and flexible catalytic lysine (green). The binding site residues (≤ 4.5 Å from HTML-PLP) are shown in lines labeled with one-letter code and number. Polar interactions between substrate and protein are indicated with orange dashes, while cation-π interactions are indicated with olive dashes.



Organism	Enzyme	k_{cat} (s⁻¹)	K_m (mM)	k_{cat}/K_m (M⁻¹s⁻¹)
Homo sapiens	SHMT1	0.123 ± 0.005	3.793 ± 0.35	32.396 ± 4.329
Homo sapiens	SHMT2	0.005 ± 0.0	0.803 ± 0.067	6.311 ± 0.674
Homo sapiens	KYAT1	N/A	N/A	N/A
Homo sapiens	SGPL1	N/A	N/A	N/A
Mus musculus	Tha1	2.311 ± 0.029	0.169 ± 0.009	13,695.069 ± 892.495
Mus musculus	Kyat3	N/A	N/A	N/A

Fig. 5: Experimental validation of HTML-OSMES candidates.

a, Time-resolved ^{1}H NMR spectra of SHMT1 activity in the presence of 5 mM HTML at 0, 35, 65 and 105 minutes. C α protons singlet of glycine is assigned in the structure. **b**, Nonlinear fitting to the Michaelis Menten equation of the dependency on HTML concentrations of the initial reaction velocity of SHMT1 (1 μM). **c**, Kinetic parameters (k_{cat}, K_m, k_{cat}/K_m) of HTMLA reaction of tested enzymes with mean and standard error values. **d,** Scheme of the broken bond (magenta cross) in the aldol cleavage reactions of HTML, L-*allo*-threonine, L-threonine. In red are the portions common to the three substrates. **e**, Bar plot in log scale of k_{cat}/K_m of different enzymes (Tha1, SHMT1, SHMT2) with different substrates (HTML, L-*allo*-threonine, L-threonine). **f**, Lineweaver-Burk double-reciprocal primary plot of the inhibition by HTML of kynurenine aminotransferase activity of KYAT1. The kynurenine concentration ranged from 0.75 to 3 mM. The concentrations of HTML were 0, 2, 4, and 8 mM.

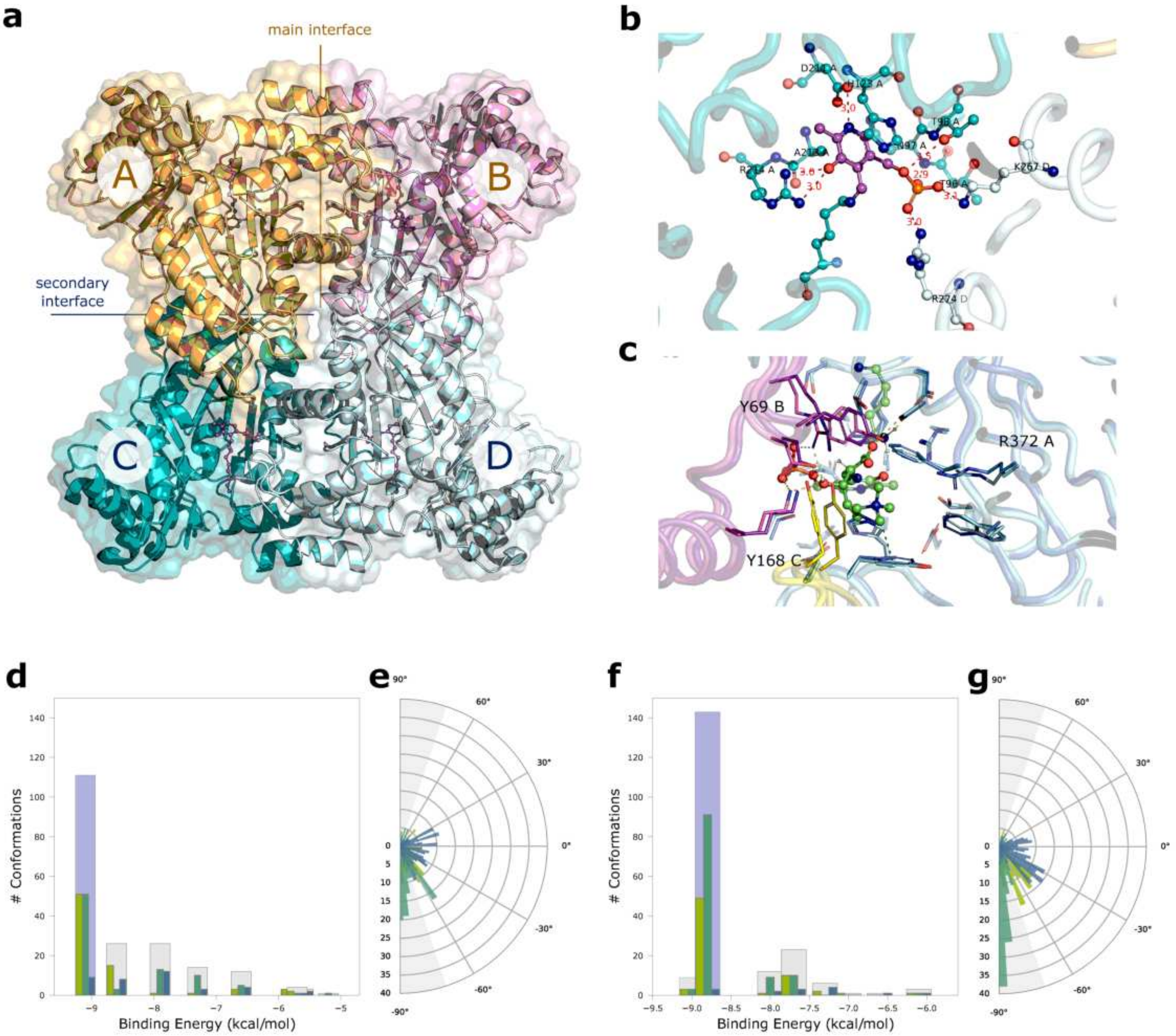


Fig. 6: Crystal structure of mouse Tha1 improves HTML-OSMES results.

a, Quaternary assembly of Tha1. The four PLP cofactors, one for each unit, are shown in violet ball-and-stick. The main interface, between units A and B (orange/magenta) and the secondary interface between units C and D (teal/pale cyan) are indicated. **b**, Active site of Tha1. The main polar interactions of the PLP cofactor (violet) at the interface between subunits A (carbon atoms in teal) and B (carbon atoms in pale cyan) are indicated. Distances in Å. **c**, Comparison of AlphaFold model (dark colors) and the Tha1 crystal structure (orthorhombic F222, light colors) docked with HTML-PLP substrate. Different chains are colored in different colors. Non-carbon atoms are colored according to CPK convention. HTML-PLP and flexible catalytic lysine are shown in ball-and-sticks. The binding site residues (≤ 4.5 Å from HTML-PLP) are shown in sticks. Polar interactions are indicated with orange dashes, cation-π interactions are indicated with olive dashes. **d-g**, Clustering of the HTML-PLP conformations at the Tha1 active site obtained with HTML-OSMES applied to AlphaFold model (**d, e**) or the crystal structure (**f, g**). Bar plots show the distribution of χ_1 (blue), χ_2 (emerald) and χ_3 (kiwi) angles in each cluster, with the catalytic cluster highlighted in light blue. Circular plots show the cumulative distribution of the three χ angles for all clusters. $|\sin(\chi)| \geq 0.95$ values are defined by gray areas.

Supplementary Information for

One substrate-many enzymes virtual screening uncovers missing genes of carnitine biosynthesis in human and mouse.

Marco Malatesta[1], Emanuele Fornasier[2], Martino Luigi Di Salvo[3], Angela Tramonti[4], Erika Zangelmi[1], Alessio Peracchi[1], Andrea Secchi[1], Eugenia Polverini[5], Gabriele Giachin[2], Roberto Battistutta[2], Roberto Contestabile[3], Riccardo Percudani[1]

[1]Department of Chemistry, Life Sciences and Environmental Sustainability, University of Parma, Parma, Italy

[2]Department of Chemical Sciences, University of Padua, Padova, Italy

[3]Istituto Pasteur Italia-Fondazione Cenci Bolognetti and Department of Biochemical Sciences "A. Rossi Fanelli", Sapienza University of Rome, Rome, Italy

[4]Institute of Molecular Biology and Pathology, Italian National Research Council, Rome, Italy

[5]Department of Mathematical, Physical and Computer Sciences, University of Parma, Parma, Italy

Supplementary Discussion 1: Tha1 structure

In Tha1 the PLP cofactor, covalently bound to the side-chain amine of Lys242, is stabilized in the active site by a network of hydrogen bonds and salt bridges with the side chains of Asp211, Arg214 and Thr98 from the same unit, and of Lys267 and Arg274 from the adjacent unit (**Fig. 6b**). His123 is making an aromatic stacking interaction with the PLP pyridine ring (the relative distance between the rings is 3.7 Å). All these residues are conserved in the *T. maritima* enzyme. The PLP phosphate makes a hydrogen bond also with the backbone NH of Asn97; in the *T. maritima* enzyme an analogous interaction is established with the backbone NH of Gly58. While the main interface has similar characteristics for the mouse and the *T. maritima* enzymes (as deduced by PISA analysis, see **Supplementary Table 6**), the secondary interface shows a higher degree of variability. Despite with similar buried area values, around 990-1060 Å^2, it stronger contributes to the stability of the tetrameric assembly in the *T. maritima* enzyme, with much higher values in ΔG^{int} (the solvation free energy gain upon formation of the assembly), ΔG^{int} P-value (a measure of interface specificity, the lower the value the more hydrophobic and interaction-specific the surface) and CSS (the Complexation Significance Score, which indicates how significant for assembly formation the interface is). The secondary interface has a more hydrophobic nature in the *T. maritima* enzyme while it is more polar in the mouse enzyme, with a higher number of hydrogen bonds and salt bridges. As a consequence, the tetrameric assembly of the *T. maritima* enzyme has a higher stability with respect of that of the mouse enzyme, with ΔG^{diss} (the free energy of assembly dissociation) values of 40.9 kcal/mol and 5-6 kcal/mol, respectively (for the ABCD to AB + CD dissociation). The relatively low value of the mouse ΔG^{diss} prompted us to verify the stability of the tetramer in solution. We performed SEC-SAXS experiments that show the presence of a single component with a MW compatible to that of the sum of 4 units, indicating that the mouse enzyme, despite the lower stability, is tetrameric in solution (**Supplementary Fig. 17b**).

Supplementary Table 1: PLP-dependent enzyme set used in OSMES

Entry	**Gene name**	**EC number**	**Organism**
Q3UX83	Accsl	N/A	Mus musculus
Q8VCN5	Cth	4.4.1.1	Mus musculus
P05201	Got1	2.6.1.1; 2.6.1.3	Mus musculus
Q9Z2Y8	Plpbp	N/A	Mus musculus
Q9QZX7	Srr	5.1.1.18; 4.3.1.18; 4.3.1.17	Mus musculus
Q8BH55	Thnsl1	N/A	Mus musculus
Q99K85	Psat1	2.6.1.52	Mus musculus
Q9WUB3	Pygm	2.4.1.1	Mus musculus
Q14CH1	Mocos	2.8.1.9	Mus musculus
Q8CI94	Pygb	2.4.1.1	Mus musculus
Q8QZR1	Tat	2.6.1.5	Mus musculus
A2AIG8	Accs	N/A	Mus musculus
Q8BG54	Sptlc3	2.3.1.50	Mus musculus
Q9ET01	Pygl	2.4.1.1	Mus musculus
Q9JLI6	Scly	4.4.1.16	Mus musculus
Q8R0X7	Sgpl1	4.1.2.27	Mus musculus
O35423	Agxt	2.6.1.51; 2.6.1.44	Mus musculus
Q6P6M7	Sepsecs	2.9.1.2	Mus musculus
P97363	Sptlc2	2.3.1.50	Mus musculus
P29758	Oat	2.6.1.13	Mus musculus
Q8R238	Sdsl	4.3.1.17; 4.3.1.19	Mus musculus
Q8VBT2	Sds	4.3.1.17; 4.3.1.19	Mus musculus
Q9CXF0	Kynu	3.7.1.3	Mus musculus
Q9Z1J3	Nfs1	2.8.1.7	Mus musculus
Q8BGT5	Gpt2	2.6.1.2	Mus musculus
Q71RI9	Kyat3	2.6.1.7; 4.4.1.13; 2.6.1.63	Mus musculus
P08680	Alas2	2.3.1.37	Mus musculus
Q9WVM8	Aadat	2.6.1.39; 2.6.1.7	Mus musculus
Q8VC19	Alas1	2.3.1.37	Mus musculus
Q8QZR5	Gpt	2.6.1.2	Mus musculus
Q91W43	Gldc	1.4.4.2	Mus musculus
Q7TSV6	Got1l1	2.6.1.1	Mus musculus
Q8BWU8	Etnppl	4.2.3.2	Mus musculus
O88986	Gcat	2.3.1.29	Mus musculus
Q8BTY1	Kyat1	2.6.1.7; 4.4.1.13; 2.6.1.64	Mus musculus

Q9CZN7	Shmt2	2.1.2.1	Mus musculus
P61922	Abat	2.6.1.19; 2.6.1.22	Mus musculus
Q8R1K4	Phykpl	4.2.3.134	Mus musculus
P00860	Odc1	4.1.1.17	Mus musculus
Q3UEG6	Agxt2	2.6.1.44; 2.6.1.40	Mus musculus
P50431	Shmt1	2.1.2.1	Mus musculus
Q80WP8	Gadl1	4.1.1.11; 4.1.1.29	Mus musculus
P05202	Got2	2.6.1.1; 2.6.1.7	Mus musculus
P48318	Gad1	4.1.1.15	Mus musculus
O88533	Ddc	4.1.1.28	Mus musculus
Q9DBE0	Csad	4.1.1.29; 4.1.1.11	Mus musculus
P24288	Bcat1	2.6.1.42	Mus musculus
P48320	Gad2	4.1.1.15	Mus musculus
P23738	Hdc	4.1.1.22	Mus musculus
O35855	Bcat2	2.6.1.42	Mus musculus
Q91WT9	Cbs	4.2.1.22	Mus musculus
Q80W22	Thnsl2	4.2.3.-	Mus musculus
Q6XPS7	Tha1	N/A	Mus musculus
P17174	GOT1	2.6.1.1; 2.6.1.3	Homo sapiens
Q4AC99	ACCSL	N/A	Homo sapiens
Q96QU6	ACCS	N/A	Homo sapiens
O95470	SGPL1	4.1.2.27	Homo sapiens
Q96EN8	MOCOS	2.8.1.9	Homo sapiens
O94903	PLPBP	N/A	Homo sapiens
Q9Y697	NFS1	2.8.1.7	Homo sapiens
P06737	PYGL	2.4.1.1	Homo sapiens
Q99259	GAD1	4.1.1.15	Homo sapiens
Q16773	KYAT1	2.6.1.7; 4.4.1.13; 2.6.1.64	Homo sapiens
O75600	GCAT	2.3.1.29	Homo sapiens
P00505	GOT2	2.6.1.1; 2.6.1.7	Homo sapiens
P13196	ALAS1	2.3.1.37	Homo sapiens
Q8N5Z0	AADAT	2.6.1.39; 2.6.1.7	Homo sapiens
Q8TD30	GPT2	2.6.1.2	Homo sapiens
Q8NHS2	GOT1L1	2.6.1.1	Homo sapiens
P11216	PYGB	2.4.1.1	Homo sapiens
P11217	PYGM	2.4.1.1	Homo sapiens
Q9HD40	SEPSECS	2.9.1.2	Homo sapiens

Q9GZT4	SRR	5.1.1.18; 4.3.1.18; 4.3.1.17	Homo sapiens
Q9NUV7	SPTLC3	2.3.1.50	Homo sapiens
Q9Y617	PSAT1	2.6.1.52	Homo sapiens
Q96I15	SCLY	4.4.1.16	Homo sapiens
P20132	SDS	4.3.1.17; 4.3.1.19	Homo sapiens
P04181	OAT	2.6.1.13	Homo sapiens
O15382	BCAT2	2.6.1.42	Homo sapiens
P80404	ABAT	2.6.1.19; 2.6.1.22	Homo sapiens
Q05329	GAD2	4.1.1.15	Homo sapiens
Q6YP21	KYAT3	2.6.1.7; 4.4.1.13; 2.6.1.63	Homo sapiens
Q16719	KYNU	3.7.1.3	Homo sapiens
P34897	SHMT2	2.1.2.1	Homo sapiens
P22557	ALAS2	2.3.1.37	Homo sapiens
P24298	GPT	2.6.1.2	Homo sapiens
Q6ZQY3	GADL1	4.1.1.11; 4.1.1.29	Homo sapiens
P23378	GLDC	1.4.4.2	Homo sapiens
Q9BYV1	AGXT2	2.6.1.44; 2.6.1.40	Homo sapiens
P34896	SHMT1	2.1.2.1	Homo sapiens
P11926	ODC1	4.1.1.17	Homo sapiens
Q8IUZ5	PHYKPL	4.2.3.134	Homo sapiens
P19113	HDC	4.1.1.22	Homo sapiens
Q8TBG4	ETNPPL	4.2.3.2	Homo sapiens
P32929	CTH	4.4.1.1	Homo sapiens
P35520	CBS	4.2.1.22	Homo sapiens
P20711	DDC	4.1.1.28	Homo sapiens
P54687	BCAT1	2.6.1.42	Homo sapiens
Q9Y600	CSAD	4.1.1.29; 4.1.1.11	Homo sapiens
P17735	TAT	2.6.1.5	Homo sapiens
Q96GA7	SDSL	4.3.1.19; 4.3.1.17	Homo sapiens
P21549	AGXT	2.6.1.51; 2.6.1.44	Homo sapiens
O15270	SPTLC2	2.3.1.50	Homo sapiens
Q86YJ6	THNSL2	4.2.3.-	Homo sapiens
Q8IYQ7	THNSL1	N/A	Homo sapiens

Supplementary Table 2: Enzyme-substrate combination used for OSMES validation.

code[a]	Gene	Entry	Organism	Reaction	EC number
CSU	CSAD	Q9Y600	*Homo sapiens*	Decarboxylase	4.1.1.29; 4.1.1.11
	Csad	Q9DBE0	*Mus musculus*	Decarboxylase	4.1.1.29; 4.1.1.11
	GADL1	Q6ZQY3	*Homo sapiens*	Decarboxylase	4.1.1.11; 4.1.1.29
	Gadl1	Q80WP8	*Mus musculus*	Decarboxylase	4.1.1.11; 4.1.1.29
CYT	CTH	P32929	*Homo sapiens*	Other	4.4.1.1
	Cth	Q8VCN5	*Mus musculus*	Other	4.4.1.1
DMA	AGXT2	Q9BYV1	*Homo sapiens*	Aminotransferase	2.6.1.44; 2.6.1.40
	Agxt2	Q3UEG6	*Mus musculus*	Aminotransferase	2.6.1.44; 2.6.1.40
EAP	ETNPPL	Q8TBG4	*Homo sapiens*	Other	4.2.3.2
	Etnppl	Q8BWU8	*Mus musculus*	Other	4.2.3.2
GLU	GAD1	Q99259	*Homo sapiens*	Decarboxylase	4.1.1.15
	GAD2	Q05329	*Homo sapiens*	Decarboxylase	4.1.1.15
	Gad1	P48318	*Mus musculus*	Decarboxylase	4.1.1.15
	Gad2	P48320	*Mus musculus*	Decarboxylase	4.1.1.15
HIS	HDC	P19113	*Homo sapiens*	Decarboxylase	4.1.1.22
	Hdc	P23738	*Mus musculus*	Decarboxylase	4.1.1.22
ILE	BCAT1	P54687	*Homo sapiens*	Aminotransferase	2.6.1.42
	BCAT2	O15382	*Homo sapiens*	Aminotransferase	2.6.1.42
	Bcat1	P24288	*Mus musculus*	Aminotransferase	2.6.1.42
	Bcat2	O35855	*Mus musculus*	Aminotransferase	2.6.1.42
KYN	AADAT	Q8N5Z0	*Homo sapiens*	Aminotransferase	2.6.1.39; 2.6.1.7
	Aadat	Q9WVM8	*Mus musculus*	Aminotransferase	2.6.1.39; 2.6.1.7
	KYAT1	Q16773	*Homo sapiens*	Aminotransferase	2.6.1.7; 4.4.1.13; 2.6.1.64
	KYAT3	Q6YP21	*Homo sapiens*	Aminotransferase	2.6.1.7; 4.4.1.13; 2.6.1.63
	Kyat1	Q8BTY1	*Mus musculus*	Aminotransferase	2.6.1.7; 4.4.1.13; 2.6.1.64
	Kyat3	Q71RI9	*Mus musculus*	Aminotransferase	2.6.1.7; 4.4.1.13; 2.6.1.63
	KYNU	Q16719	*Homo sapiens*	Other	3.7.1.3
	Kynu	Q9CXF0	*Mus musculus*	Other	3.7.1.3
ORN	ODC1	P11926	*Homo sapiens*	Decarboxylase	4.1.1.17
	Odc1	P00860	*Mus musculus*	Decarboxylase	4.1.1.17
PHL	PHYKPL	Q8IUZ5	*Homo sapiens*	Other	4.2.3.134
	Phykpl	Q8R1K4	*Mus musculus*	Other	4.2.3.134
SER	SHMT1	P34896	*Homo sapiens*	Aldolase[b]	2.1.2.1
	SHMT2	P34897	*Homo sapiens*	Aldolase[b]	2.1.2.1
	Shmt1	P50431	*Mus musculus*	Aldolase[b]	2.1.2.1
	Shmt2	Q9CZN7	*Mus musculus*	Aldolase[b]	2.1.2.1
SPL	SGPL1	O95470	*Homo sapiens*	Aldolase	4.1.2.27
	Sgpl1	Q8R0X7	*Mus musculus*	Aldolase	4.1.2.27

TYR	TAT	P17735	*Homo sapiens*	Aminotransferase	2.6.1.5
	Tat	Q8QZR1	*Mus musculus*	Aminotransferase	2.6.1.5
	DDC	P20711	*Homo sapiens*	Decarboxylase	4.1.1.28
	Ddc	O88533	*Mus musculus*	Decarboxylase	4.1.1.28

[a] three-letter code of the substrate used in the pdbqt files.

[b] SHMT catalyzes the transfer of Cβ of serine to tetrahydrofolate (THF) although it can catalyze an aldolase reaction in the absence of THF.

Supplementary Table 3: Significant coevolutionary association of carnitine biosynthesis enzymes TMLD and BBD with PLP-dependent enzymes according to cotr analysis.

5485575at2759 (TMLD or BBD)											
OG1	*OG2*	*species*	*t1*	*t2*	*c*	*d*	*k*	*cotr_score*	*P_value*	*P_value(adj)*	
884390at2759	5485575at2759	1929	43	52	12	0	12	0.145	3.22E-10	7.56E-03	ornithine aminotransferase
345661at2759	5485575at2759	1929	67	52	15	2	13	0.123	7.20E-09	1.66E-01	ethanolamine-phosphate phospho-lyase
177349at2759	5485575at2759	1929	30	52	9	1	8	0.108	5.97E-07	1.00E+00	glycine decarboxylase
178754at2759	5485575at2759	1929	86	52	12	0	12	0.095	1.37E-06	1.00E+00	low-specificity L-threonine aldolase 2
5471916at2759	5485575at2759	1929	57	52	10	0	10	0.101	1.37E-06	1.00E+00	kynureninase
5474881at2759	5485575at2759	1929	53	52	9	0	9	0.094	6.62E-06	1.00E+00	Aminotransferase
TMLD											
OG1	*OG2*	*species*	*t1*	*t2*	*c*	*d*	*k*	*cotr_score*	*P_value*	*P_value(adj)*	
178754at2759	TMLD	1929	86	123	19	0	19	0.100	7.29E-07	1.00E+00	low-specificity L-threonine aldolase 2
884390at2759	TMLD	1929	43	123	13	0	13	0.085	1.09E-06	1.00E+00	ornithine aminotransferase
3024111at2759	TMLD	1929	37	123	12	1	11	0.0738	9.40E-06	1.00E+00	sphingosine-1-phosphate lyase 1
BBD											
OG1	*OG2*	*species*	*t1*	*t2*	*c*	*d*	*k*	*cotr_score*	*P_value*	*P_value(adj)*	
178754at2759	BBD	1929	86	77	16	0	16	109	9.22E-08	1.00E+00	low-specificity L-threonine aldolase 2
884390at2759	BBD	1929	43	77	12	1	11	101	4.10E-07	1.00E+00	ornithine aminotransferase
5487987at2759	BBD	1929	46	77	12	1	11	0.098	8.63E-07	1.00E+00	cystathionine beta-synthase-like
5471916at2759	BBD	1929	57	77	11	0	11	0.089	8.36E-06	1.00E+00	kynureninase
T_B (TMLD and BBD)											
OG1	*OG2*	*species*	*t1*	*t2*	*c*	*d*	*k*	*cotr_score*	*P_value*	*P_value(adj)*	
178754at2759	T_B	1929	86	147	23	0	23	110	2.79E-08	6.38E-01	low-specificity L-threonine aldolase 2

Supplementary Table 4: Kinetic parameters of Tha1, SHMT1 and SHMT2 with different β-hydroxylated amino acids.

Enzyme	**Substrate**	***k_{cat} (s^{-1})***	***K_m (mM)***	***k_{cat}/K_m (s^{-1} M^{-1})***
Tha1	HTML	2.311 ± 0.029	0.169 ± 0.009	13,695.069 ± 892.495
	L-*allo*-threonine	1.96 ± 0.039	8.236 ± 0.496	238.005 ± 19.055
	L-threonine	0.375 ± 0.024	37.531 ± 6.278	10.005 ± 2.304
SHMT1	HTML	0.122 ± 0.006	3.792 ± 0.444	32.173 ± 5.34
	L-*allo*-threonine	2.365 ± 0.034	0.226 ± 0.016	10,464.602 ± 889.49
	L-threonine	0.153 ± 0.008	36.522 ± 4.153	4.199 ± 0.525
SHMT2	HTML	0.005 ± 0.000	0.803 ± 0.162	6.227 ± 1.256
	L-*allo*-threonine	0.312 ± 0.012	1.345 ± 0.190	231.970 ± 41.69
	L-threonine	0.035 ± 0.002	41. 224 ± 3.919	0.845 ± 0.089

Supplementary Table 5: RMSD values (in Å) calculated between Cα-atoms of matched residues.

RMSD (Å)	**C2_chainA**	**C2_chainB**
C2_chainA	-	0.262
F222_chainA	0.255	0.280
1M6S_chainA	1.172	1.148
1M6S_chainB	1.059	1.051
1M6S_chainC	1.119	1.110
1M6S_chainD	1.055	1.031

C2 and F222 refer to the Tha1/HTMLA structures in the two space-groups.
1M6S refers to the structure of *Thermotoga maritima* threonine aldolase.

Supplementary Table 6: Parameters relative to the interaction interfaces as determined by PISA.

Protein	Buried Area (Å^2)	ΔG^{int} (kcal/mol)	ΔG^{int} P-value[a]	HB[a]	SB[a]	CSS[a]	$T\Delta S^{diss}$ (kcal/mol)	ΔG^{diss} (kcal/mol)
ABCD ⇒ AB + CD								
C2	14770	-31.5	-	-	-	-	15.8	5.0
F222	14070	-34.7	-	-	-	-	15.7	6.1
1M6S	13915	-67.6	-	-	-	-	15.5	40.9
AB ⇒ A + B or CD ⇒ C+D (main interface)								
C2	3762.9	-15.0	0.141	29	6	1.000	14.2	14.6
F222	3600.0	-16.8	0.085	28	6	1.000	14.2	15.9
1M6S	3680.0	-14.5	0.253	23	8	1.000	14.0	11.7
AC ⇒ A + C or BD ⇒ B + DC (secondary interface)								
C2	2132.0	-5.3	0.409	6	6	0.225	nd	nd
F222	1986.0	-0.9	0.679	10	4	0.144	nd	nd
1M6S	1910.0	-18.0	0.017	7	0	1.000	13.7	7.4

[a] values determined only for single interfaces.
nd, “not determined” by PISA because of unstable interactions.
C2 and F222 refer to the Tha1/HTMLA structures in the two space-groups. 1M6S refers to the structure of *Thermotoga maritima* threonine aldolase.
HB, number of hydrogen bonds; SB, number of salt bridges.
See main text for the significance of the other parameters.

Supplementary Table 7: CC-CFC with different structure models of Tha1.

rank	**Gene**	**structure**	**CC-CFC**	**$\|\sin(\chi_2)\|$**	**LCC**	**d**	**E (kcal/mol)**
2	Tha1	F222	91	0.754	143	4.55	-8.8
4	Tha1	C2	75	0.681	158	4.26	-9.7
5	Tha1	AF_only_F222[a]	70	0.754	130	4.48	-9.3
6	Tha1	AF_F222[b]	63	0.734	122	4.46	-9.1
7	Tha1	AF_aligned_F222[c]	60	0.659	157	4.27	-9.7
9	Tha1	AF	51	0.724	111	4.83	-9.1

[a] AlphaFold model obtained with F222 as a unique template.

[b] AlphaFold model obtained adding F222 to template list.

[c] AlphaFold model from AFDB aligned to F222 quaternary structure.

Supplementary Table 8: Experimental details on the SAXS experiment and analysis.

(A) Sample details

Sample name	Threonine aldolase 1 (Tha 1)
Organism	*Mus musculus* (Mouse)
UniProt sequence ID	Q6XPS7
Calculated molecular weight (Da)	41497.28 (monomer), 165989.12 (tetramer)
Total frames (frames used)	26
Protein concentration (mg/mL)	5.5 mg/mL
SEC-SAXS column	Superdex 200 10/300 GL Cytiva
Injection volume, flow rate	250 μL, 0.5 mL/min
Protetin buffer	20 mM Tris, 150 mM NaCl, 1 mM EDTA, 1 mM DTT, pH 7.8

(B) SAXS data collection parameters

Instrument	ESRF BM29
Wavelength (Å)	0.99
q-range (Å)	0.004-0.5
Sample-to-detector distance (m)	2.867
Exposure time	0.5 sec/frame
Temperature (° C)	20
Detector	Pilatus3 X 2M (Dectris)
Flux (photons/s)	2 x 10^{12}
Beam size (μm)	100 x 100
Sample configuration	1.8 mm quartz glass capillary
Absolute scaling method	Comparison to water in sample capillary
Normalization	To transmitted intensity by beam-stop counter

(C) Structural parameters

Guinier analysis	
I(0)	21.06
R_g (nm)	3.69 ± 0.03
q-range (nm^{-1}), point range	0.0267-0.1230, 14-50
P(r) analysis	
I(0)	20.88
R_g (nm)	3.62
D_{max} ($R_{max,}$ nm)	10.6
q-range (nm^{-1}), point range	0.0177-3.313, 6-592
Porod volume ($Å^3$)	250961
χ^2 [total estimate from GNOM]	0.9102
Mass estimate based on volume (kDa), ratio to predicted tetramer	167.3, 1

(D) Software employed for SAXS data reduction, analysis and interpretation

SAXS data collection and processing	pyFAI, BsxCuBE and Primus (ATSAS 3.2.1)
Shape/bead modelling	DAMMIN(ATSAS 3.2.1)
Atomic structure modelling	CRYSOL (ATSAS 3.2.1)

3D graphic representation	UCSF Chimera 1.15

(E) Shape model-fitting results

DAMMIN (default parameters)	
q range for fitting (nm^{-1})	0.0177-3.313
Symmetry assumption	P2

(F) Atomistic modelling

CRYSOL (default parameters)	
Starting crystal structure	Tha1 tetramer (this study)
χ^2 of the fit	1.145

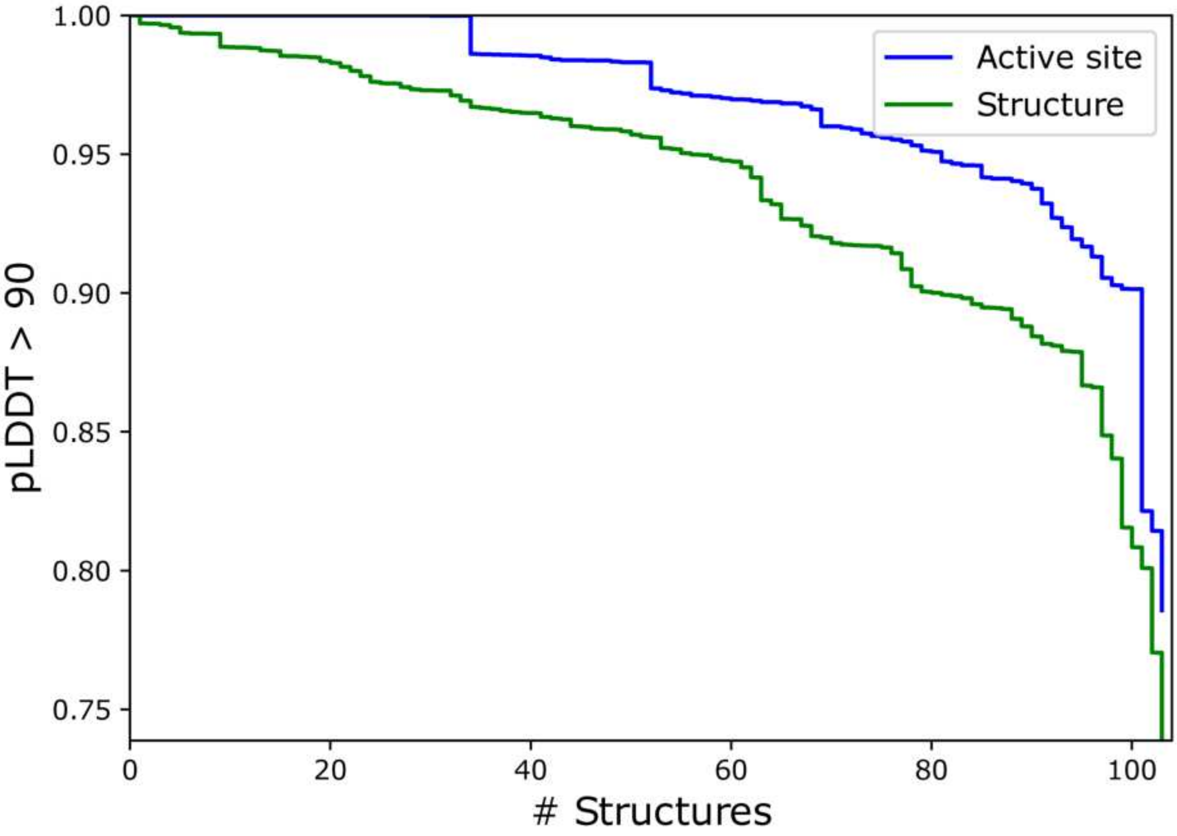
(G) Small Angle Scattering Biological Data Bank (SASBDB)

SASBDB ID*	SASDSU8

Supplementary Table 9: X-ray diffraction data processing and model refinement statistics.

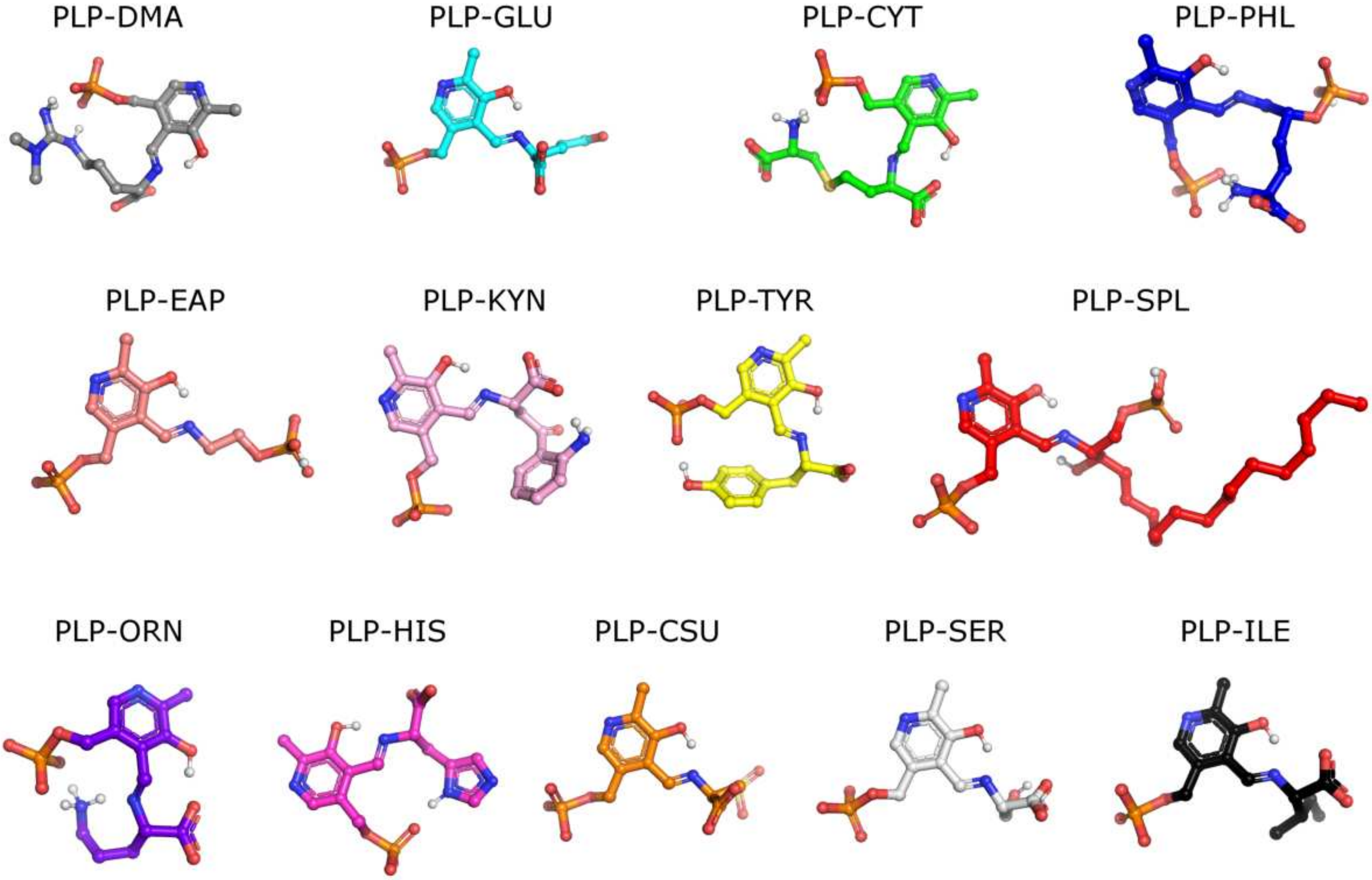
Data collection		
X-ray source	ESRF ID23-2	
Wavelength (Å)	0.8731	
Space group	F222	C2
Cell dimensions		
a, b, c (Å)	83.69, 100.52, 171.26	83.97, 101.64, 95.96
a, b, g (°)	90.00, 90.00, 90.00	90.00, 116.14, 90.00
Resolution range (Å)	25.71 – 2.26 (2.33 – 2.26)	43.07 – 2.60 (2.72 – 2.60)
R_{merge}	0.084 (0.842)	0.104 (0.731)
R_{meas}	0.088 (0.887)	0.115 (0.804)
R_{pim}	0.027 (0.272)	0.047 (0.325)
Total number of observations	168470 (15409)	109798 (14100)
Total number unique	17051 (1561)	21160 (2662)
Mean(I)/s(I)	16.6 (1.5)	8.9 (1.4)
$CC_{1/2}$	0.998 (0.881)	0.996 (0.786)
Completeness (%)	99.7 (99.9)	94.8 (98.0)
Multiplicity	9.9 (9.9)	5.2 (5.3)
Wilson B estimate (Å^2)	46.4	65.3
Refinement		
Resolution range (Å)	25.71 – 2.26	43.07 – 2.60
R_{work}/R_{free} (%)	21.0/24.0	23.4/25.9
Number of atoms		
Protein	2806	5536
Water	57	29
Average B, all atoms (Å^2)	67.0	83.0
r.m.s.d.		
Bond lengths (Å)	0.003	0.002
Bond angles (°)	0.596	0.560
Ramachandran statistics		
Favored (%)	95.4	95.8
Allowed (%)	4.1	3.9
Outliers (%)	0.5	0.3
PDB entry	8PUS	8PUM

Values in parentheses are for the highest-resolution shell.



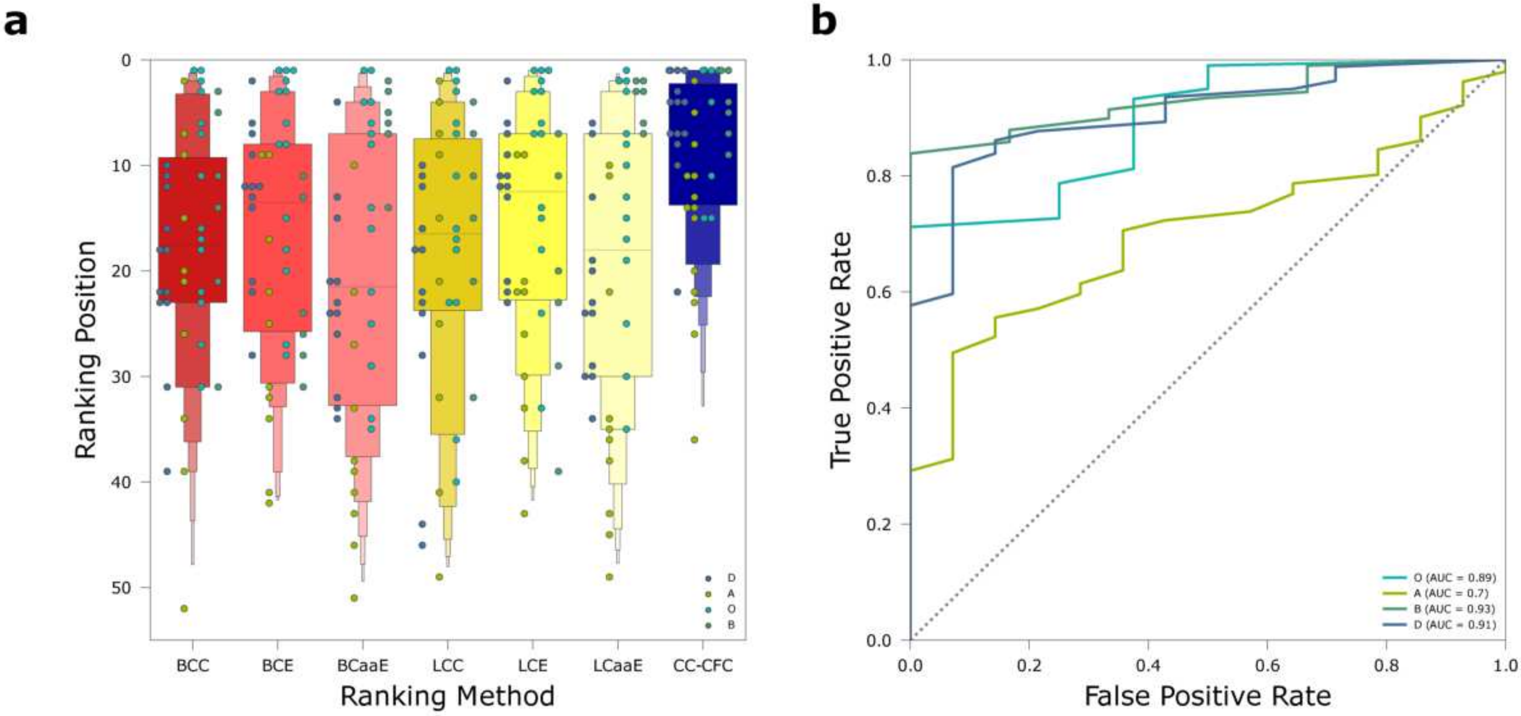
Supplementary Fig. 1: Cumulative frequency curve of pLDDT values in the AlphaFold models for human and mouse PLPomes.

Cumulative frequency curves of the residues with very high confidence (pLDDT > 90) of the whole structure (green) or near the active site (<5Å from catalytic lysine; blue) of the entire enzyme set (103 structures).



Supplementary Fig. 2: Validation library of substrate-PLP complexes.

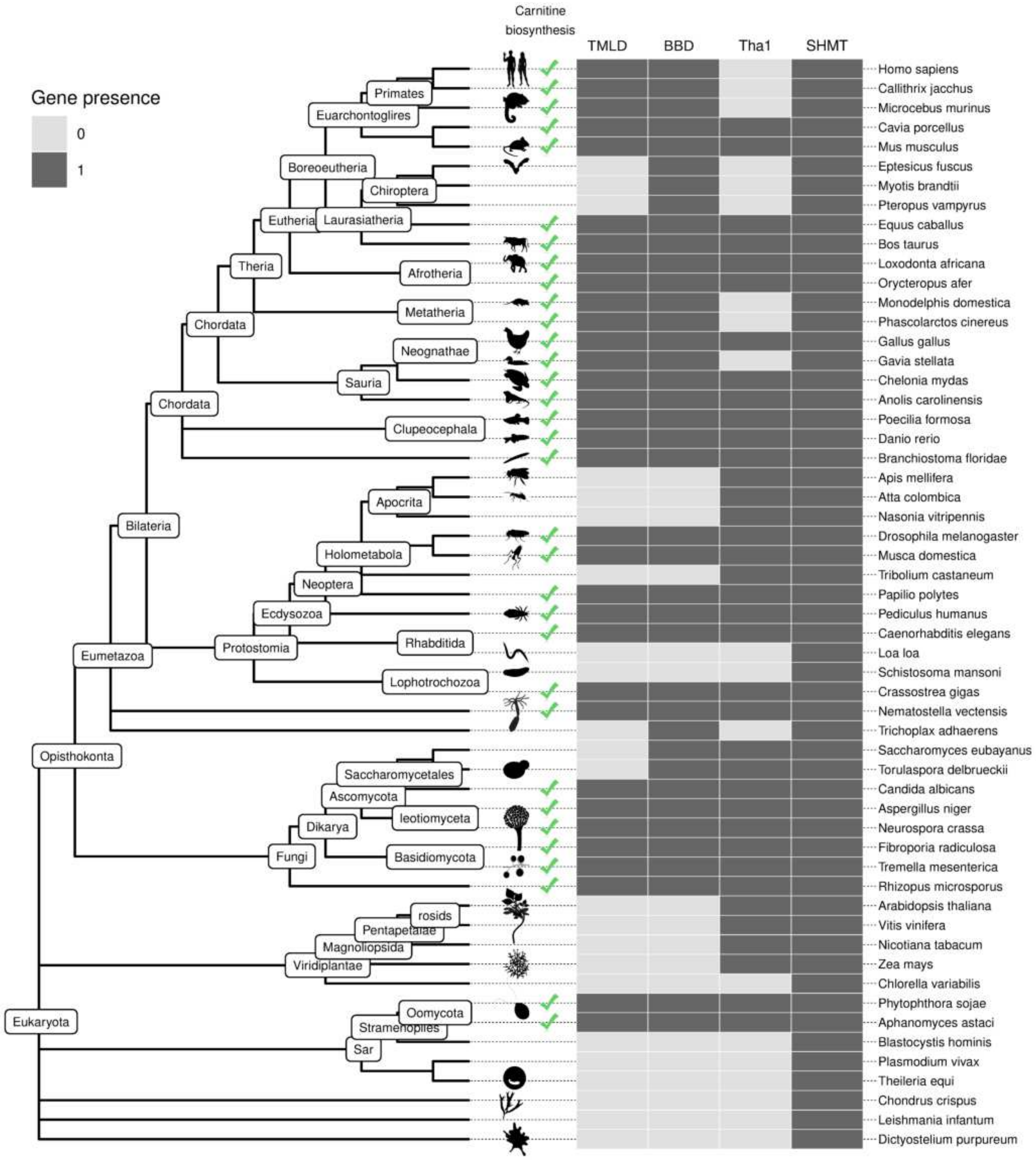
Substrate-PLP complexes (external aldimine) considered for the selection of the best classification method. The energy-minimized conformation of each substrate is shown in ball-and-stick representation. Carbon atoms are colored according to **Fig. 2e**; non-carbon atoms are colored according to CPK. Shown are PLP-bound asymmetric-L-dimethylarginine (PLP-DMA), L-glutamate (PLP-GLU), L-cystathionine (PLP-CYT), 5-phosphohydroxy-L-Lysine (PLP-PHL), phosphoethanolamine (PLP-EAP), L-kynurenine (PLP-KYN), L-tyrosine (PLP-TYR), sphingosine-1-phosphate (PLP-SPL), L-ornithine (PLP-ORN), L-histidine (PLP-HIS), L-cysteinesulfinate (PLP-CSU), L-serine (PLP-SER), L-isoleucine (PLP-ILE).



Supplementary Fig. 3: OSMES performance with positive controls grouped by reaction type.

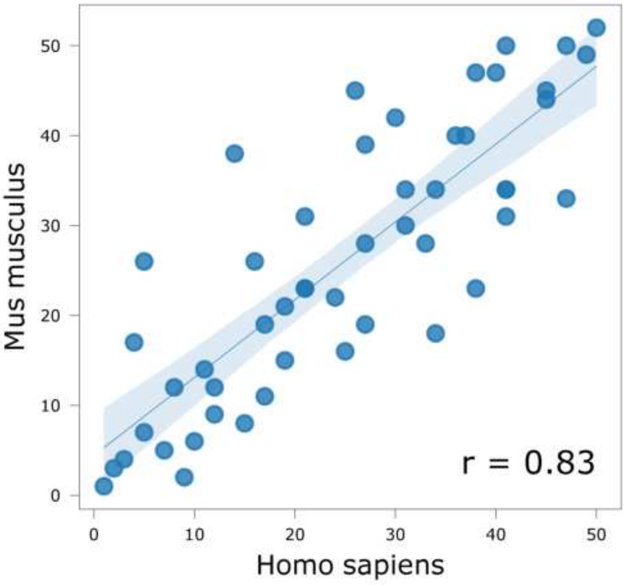
a, Letter-value plot showing the distribution of the validation set colored according to the 7 ranking methods. BC related methods are colored in red tones; LC related methods are colored in yellow tones; CC-CFC is colored in blue.

Individual dots representing ranking positions of positive controls (i.e. enzymes known to act on the substrate) are colored according to reaction type (D: decarboxylase; A: aminotransferase; O: other; B: aldolase, including SHMTs). **b**, Receiver operating characteristic curve (ROC) of the CC-CFC method divided according to reaction type. The dotted diagonal represents an area under curve (AUROC) value of 0.5.

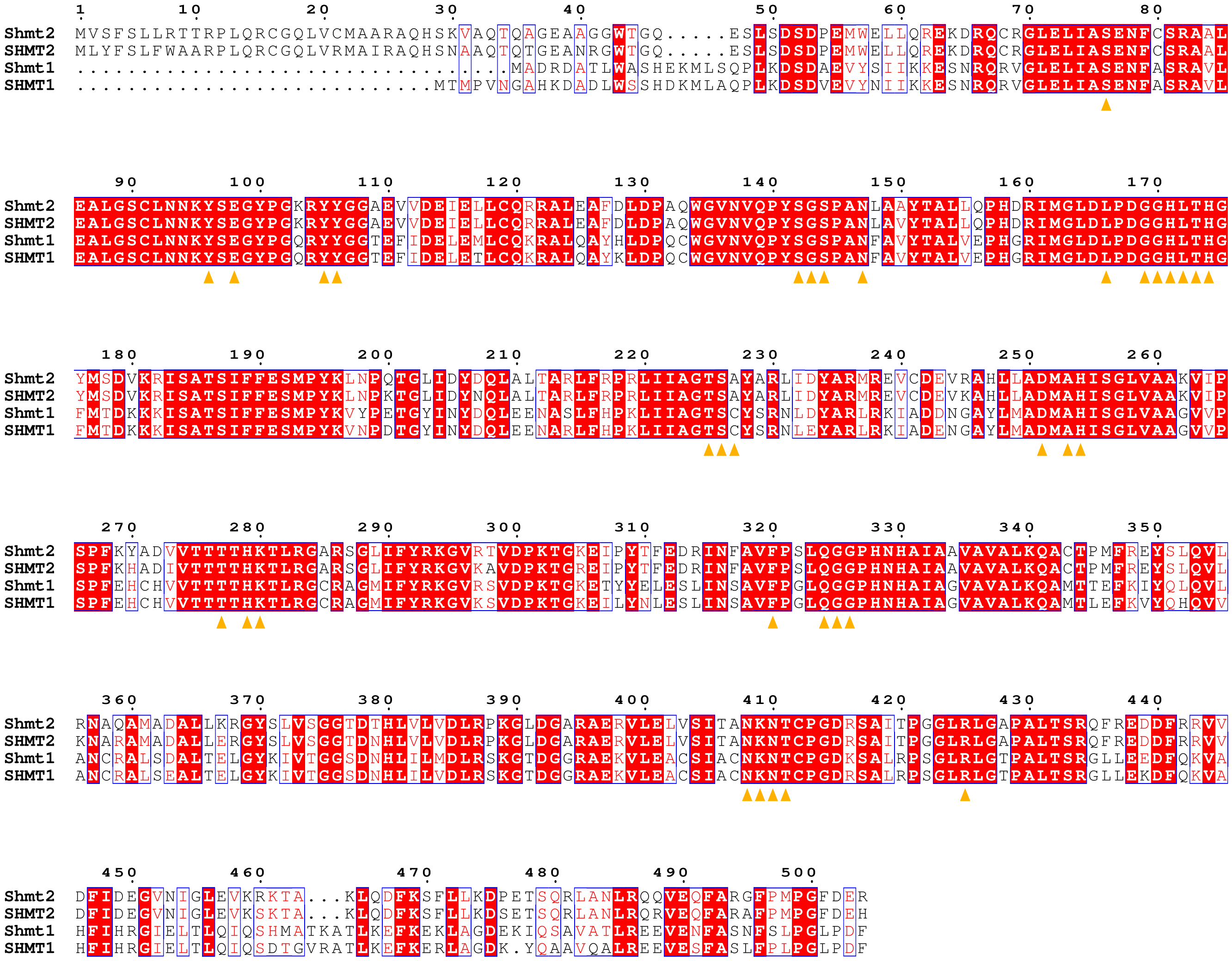


Supplementary Fig. 4: Distribution of genes of the carnitine pathway across eukaryote phylogeny.

Presence of the carnitine biosynthetic pathway in selected eukaryotes as deduced from the co-presence of trimethyllysine dioxygenase (TMLD) and ɣ-butyrobetaine dioxygenase (BBD). The distribution of Threonine aldolase (Tha1) and serine hydroxymethyl transferase characterized in this work for hydroxy trimethyl-lysine aldolase (HTMLA) activity are shown for comparison. The presence of genes was assessed by hmmsearch using Hidden Markov Models built on protein alignments (TMLD, BBD) or extracted from B6DB (Tha1, SHMT). The presence of the carnitine biosynthesis pathway is indicated by a green check mark as deduced by the co-presence of TMLD and BBD. The tree represents a simplified scheme of eukaryote phylogeny according to NCBI Taxonomy.

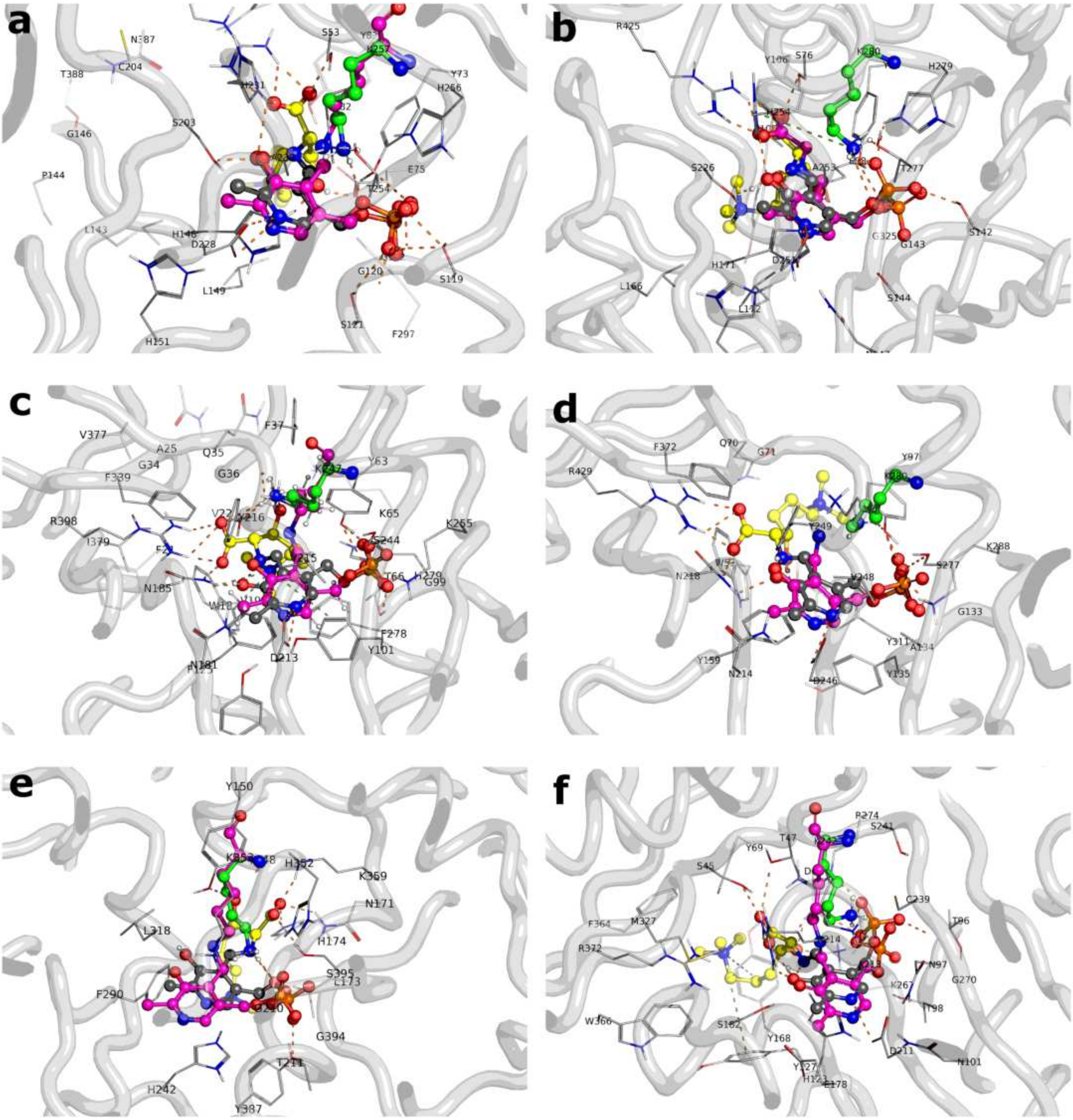


Supplementary Fig. 5: Correlation between the ranking of human and mouse orthologous genes. Scatter plot of the ranking position of each orthologous pair in human (x-axis) and mouse (y axis) CC-CFC ranking. The blue line represents the linear regression of the points; *r* represents the Spearman's rank correlation coefficient.


1 10 20 30 40 50 60 70 80
Shmt2 MVSFSLLRTTRPLQRCGQLVCMAARAQHSKVAQTQAGEAAGGWTGQ.....ESLSDSDPEMWELLQREKDRQCRGLELIASENFCSRAAL
SHMT2 MLYFSLFWAARPLQRCGQLVRMAIRAQHSNAAQTQTGEANRGWTGQ.....ESLSDSDPEMWELLQREKDRQCRGLELIASENFCSRAAL
Shmt1 ..................................MADRDATLWASHEKMLSQPLKDSDAEVYSIIKKESNRQRVGLELIASENFASRAVL
SHMT1 ............................MTMPVNGAHKDADLWSSHDKMLAQPLKDSDVEVYNIIKKESNRQRVGLELIASENFASRAVL
90 100 110 120 130 140 150 160 170
Shmt2 EALGSCLNNKYSEGYPGKRYYGGAEVVDEIELLCQRRALEAFDLDPAQWGVNVQPYSGSPANLAAYTALLQPHDRIMGLDLPDGGHLTHG
SHMT2 EALGSCLNNKYSEGYPGKRYYGGAEVVDEIELLCQRRALEAFDLDPAQWGVNVQPYSGSPANLAVYTALLQPHDRIMGLDLPDGGHLTHG
Shmt1 EALGSCLNNKYSEGYPGQRYYGGTEFIDELEMLCQKRALQAYHLDPQCWGVNVQPYSGSPANFAVYTALVEPHGRIMGLDLPDGGHLTHG
SHMT1 EALGSCLNNKYSEGYPGQRYYGGTEFIDELETLCQKRALQAYKLDPQCWGVNVQPYSGSPANFAVYTALVEPHGRIMGLDLPDGGHLTHG
180 190 200 210 220 230 240 250 260
Shmt2 YMSDVKRISATSIFFESMPYKLNPQTGLIDYDQLALTARLFRPRLIIAGTSAYARLIDYARMREVCDEVRAHLLADMAHISGLVAAKVIP
SHMT2 YMSDVKRISATSIFFESMPYKLNPKTGLIDYNQLALTARLFRPRLIIAGTSAYARLIDYARMREVCDEVKAHLLADMAHISGLVAAKVIP
Shmt1 FMTDKKKISATSIFFESMPYKVYPETGYINYDQLEENASLFHPKLIIAGTSCYSRNLDYARLRKIADDNGAYLMADMAHISGLVAAGVVP
SHMT1 FMTDKKKISATSIFFESMPYKVNPDTGYINYDQLEENARLFHPKLIIAGTSCYSRNLEYARLRKIADENGAYLMADMAHISGLVAAGVVP
270 280 290 300 310 320 330 340 350
Shmt2 SPFKYADVVTTTTHKTLRGARSGLIFYRKGVRTVDPKTGKEIPYTFEDRINFAVFPSLQGGPHNHAIAAVAVALKQACTPMFREYSLQVL
SHMT2 SPFKHADIVTTTTHKTLRGARSGLIFYRKGVKAVDPKTGREIPYTFEDRINFAVFPSLQGGPHNHAIAAVAVALKQACTPMFREYSLQVL
Shmt1 SPFEHCHVVTTTTHKTLRGCRAGMIFYRKGVRSVDPKTGKETYYELESLINSAVFPGLQGGPHNHAIAGVAVALKQAMTTEFKIYQLQVL
SHMT1 SPFEHCHVVTTTTHKTLRGCRAGMIFYRKGVKSVDPKTGKEILYNLESLINSAVFPGLQGGPHNHAIAGVAVALKQAMTLEFKVYQHQVV
360 370 380 390 400 410 420 430 440
Shmt2 RNAQAMADALLKRGYSLVSGGTDTHLVLVDLRPKGLDGARAERVLELVSITANKNTCPGDRSAITPGGLRLGAPALTSRQFREDDFRRVV
SHMT2 KNARAMADALLERGYSLVSGGTDNHLVLVDLRPKGLDGARAERVLELVSITANKNTCPGDRSAITPGGLRLGAPALTSRQFREDDFRRVV
Shmt1 ANCRALSDALTELGYKIVTGGSDNHLILMDLRSKGTDGGRAEKVLEACSIACNKNTCPGDKSALRPSGLRLGTPALTSRGLLEEDFQKVA
SHMT1 ANCRALSEALTELGYKIVTGGSDNHLILVDLRSKGTDGGRAEKVLEACSIACNKNTCPGDRSALRPSGLRLGTPALTSRGLLEKDFQKVA
450 460 470 480 490 500
Shmt2 DFIDEGVNIGLEVKRKTA...KLQDFKSFLLKDPETSQRLANLRQQVEQFARGFPMPGFDER
SHMT2 DFIDEGVNIGLEVKSKTA...KLQDFKSFLLKDSETSQRLANLRQRVEQFARAFPMPGFDEH
Shmt1 HFIHRGIELTLQIQSHMATKATLKEFKEKLAGDEKIQSAVATLREEVENFASNFSLPGLPDF
SHMT1 HFIHRGIELTLQIQSDTGVRATLKEFKERLAGDK.YQAAVQALREEVESFASLFPLPGLPDF

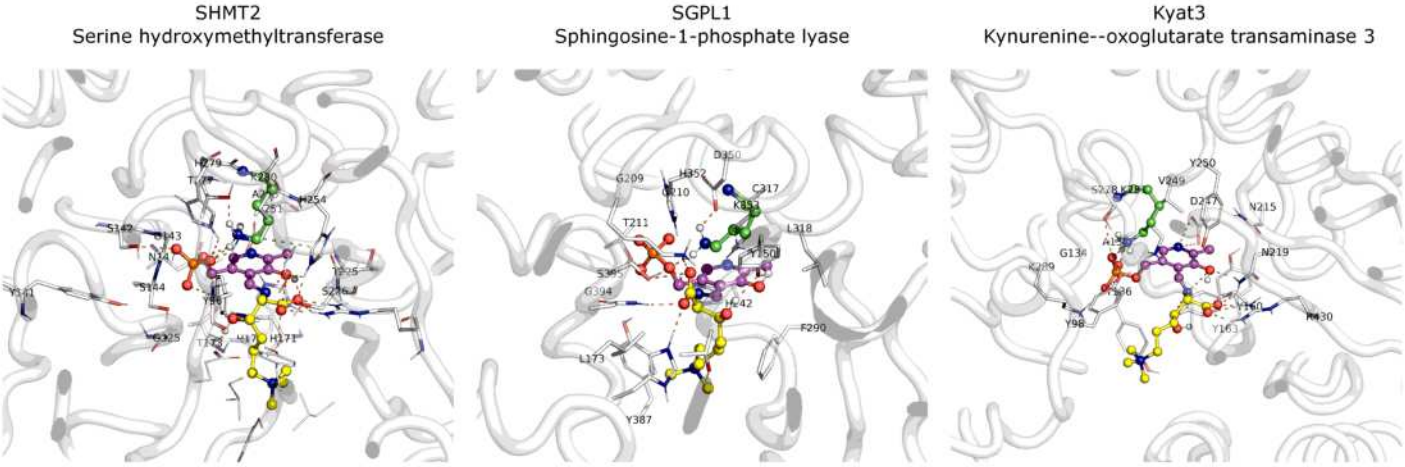
Supplementary Fig. 6: Multiple sequence alignment of human and mouse SHMTs.

Multiple sequence alignment of the main isoforms of human and mouse SHMTs obtained with ClustalΩ and visualized with ESPript. Red shading is according to residue conservation, orange triangles indicate residues at ≤ 3.5 Å from the external aldimine (PDB ID: 6FL5).



Supplementary Fig. 7: Comparison of PLP cofactor positions in docked poses and in experimental structures

Structural representation of the lowest-energy binding modes among the catalytic clusters obtained by docking of PLP-HTML substrate for SHMT1 (**a**), SHMT2 (**b**), KYAT1 (**c**), KYAT3 (**d**), SGPL1 (**e**), Tha1 (**f**) aligned with experimental structures (PDB IDs: 6FL5, 8AQL, 4WLH, 5VEQ, 4Q6R, 1JG8, respectively) containing PLP as external or internal aldimine (purple). Non-carbon atoms are colored according to CPK convention. The conformations are shown with ball-and-sticks composed by PLP cofactor (gray) covalently bound to HTML (yellow), and flexible catalytic lysine (green). The binding site residues (≤ 3.5 Å from PLP-HTML) are shown in lines labeled with one-letter code and number. Polar interactions between substrate and protein are indicated with yellow dashes, while cation-π interactions are indicated with olive dashes.



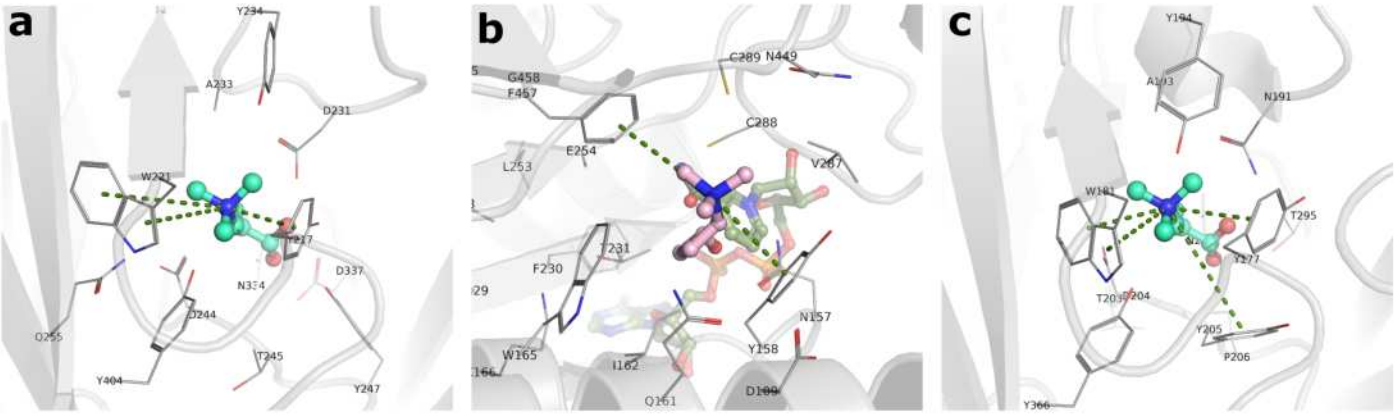
SHMT2
Serine hydroxymethyltransferase

SGPL1
Sphingosine-1-phosphate lyase

Kyat3
Kynurenine--oxoglutarate transaminase 3

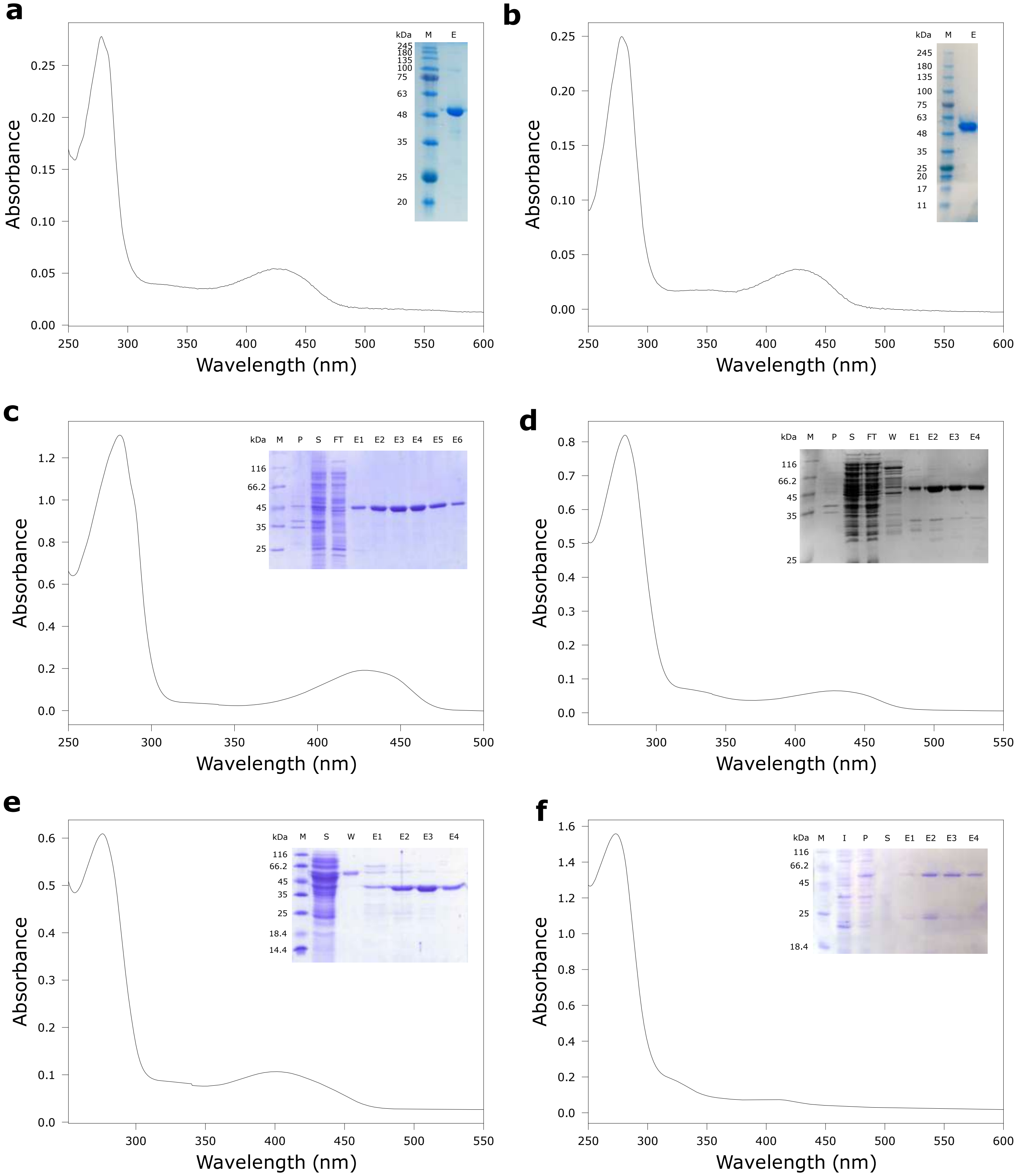
Supplementary Fig. 8: Docking poses of SHMT2, SGPL1 and Kyat3 after HTML-OSMES.

Structural representation of the lowest-energy binding modes among the catalytic clusters obtained by docking of HTML-PLP substrate for SHMT2, SGPL1 and Kyat3. Non-carbon atoms are colored according to CPK. The conformations are shown with ball-and-sticks composed by PLP cofactor (gray) covalently bound to HTML (yellow), and flexible catalytic lysine (green). The binding site residues (≤ 3.5 Å from HTML-PLP) are shown in lines labeled with one-letter code and number. Polar interactions between substrate and protein are indicated with yellow dashes, while cation-π interactions are indicated with olive dashes.



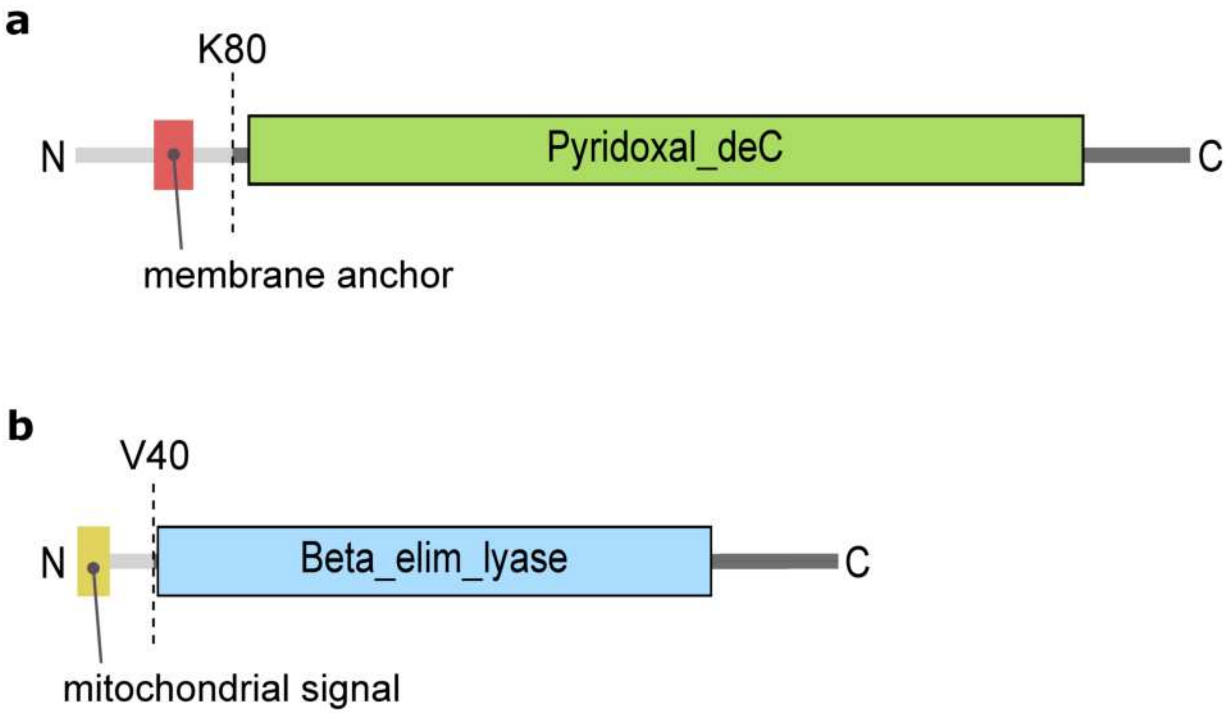
Supplementary Fig. 9: Aromatic cage of enzymes involved in carnitine biosynthesis.

a, Structural representation of aromatic cage in the crystal structure of human BBD crystal structure (PDB ID: 3O2G) in complex with the substrate butyrobetaine. **b**, Structural representation of aromatic cage in the AlphaFold model of human TMLD in complex with butyrobetaine from the structural alignment with BBD structure. **c**, Structural representation of aromatic cage in the crystal structure of human TMABADH crystal structure (PDB ID: 6V6R) in complex with the docked substrate TMABA using ADFR. In all panels, cation-π interactions are indicated with olive dashes.



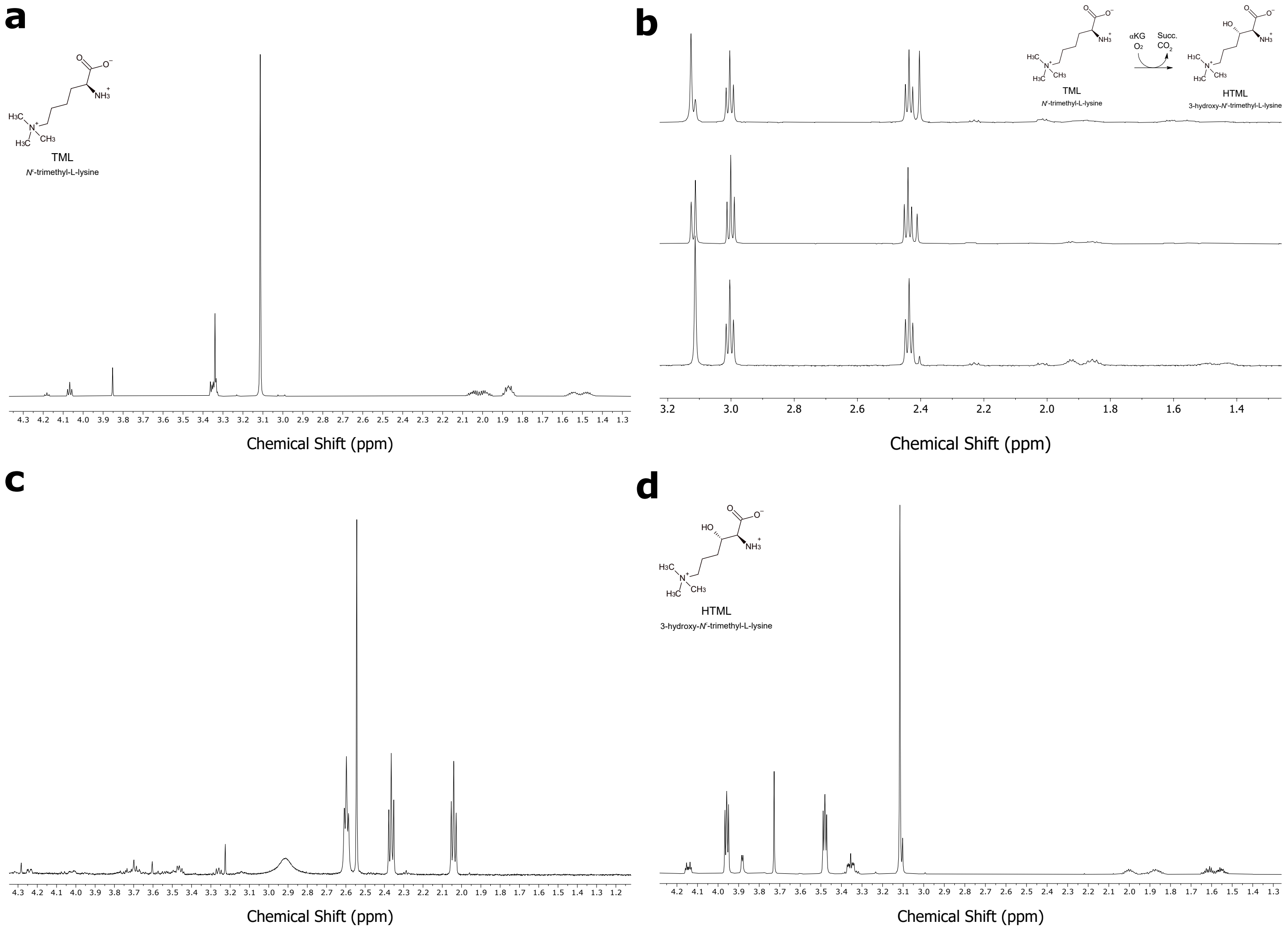
Supplementary Fig. 10: SDS-PAGE and UV-Vis spectra of purified enzymes.

UV -Vis spectra of SHMT1 (**a**), SHMT2 (**b**), KYAT1 (**c**), Kyat3 (**d**), Tha1 (**e**), SGPL1 (**f**) showing a protein peak around 280 nm and a PLP signal around 400-430 nm. Insets show SDS-PAGE of the purification steps (M: marker; I: induced cells; S: soluble fraction; P: pellet; FT: flow-through; W: washing; E: elution).



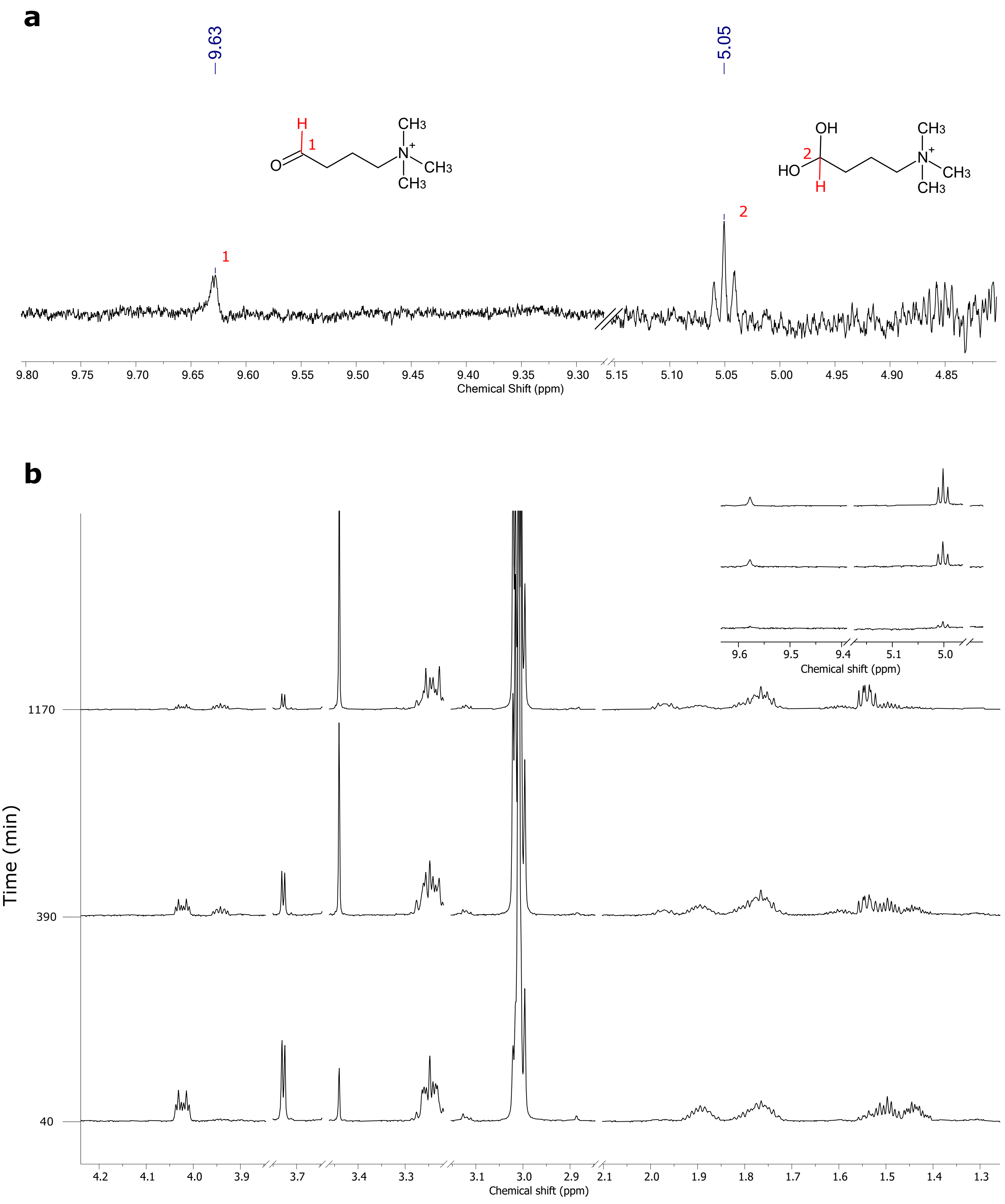
Supplementary Fig. 11: Tha1 and SGPL1 N-terminal truncation

a, SGPL1 domain composition according to PFAM. The dashed line indicates membrane anchor truncation for recombinant protein expression. **b**, Tha1 domain composition according to PFAM. The dashed line indicates mitochondrial signal truncation for recombinant protein expression.



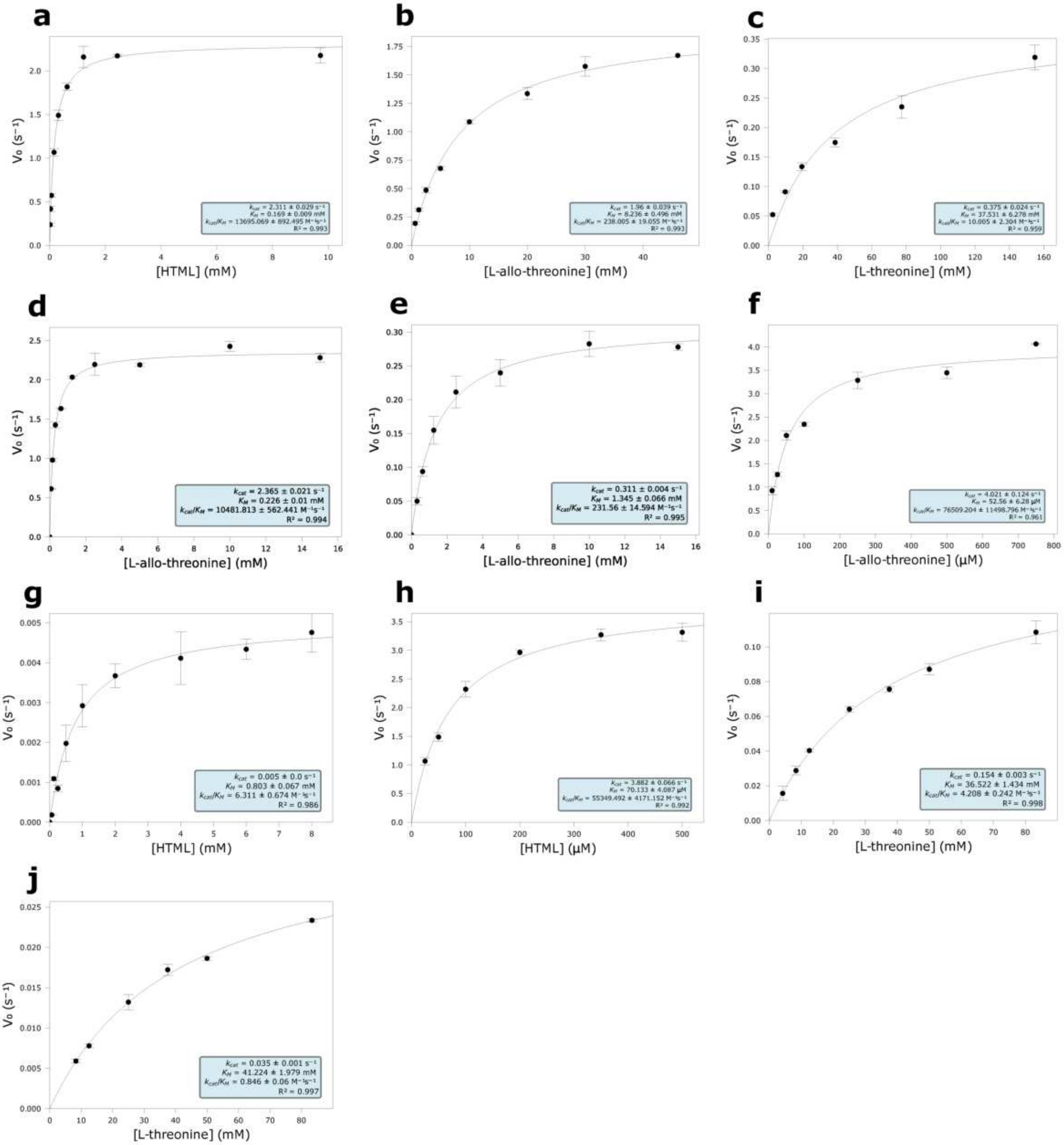
Supplementary Fig. 12: Chemo-enzymatic synthesis of HTML

a, ^{1}H NMR spectrum of chemically synthesized trimethyllysine (TML). **b**, ^{1}H NMR spectra of TML after the addition of TMLD enzyme for the enzymatic synthesis of (S,S) HTML. **c**, ^{1}H NMR spectrum of flow-through after HTML purification from reaction mixture. **d**, ^{1}H NMR spectrum of purified HTML used in the activity assays.

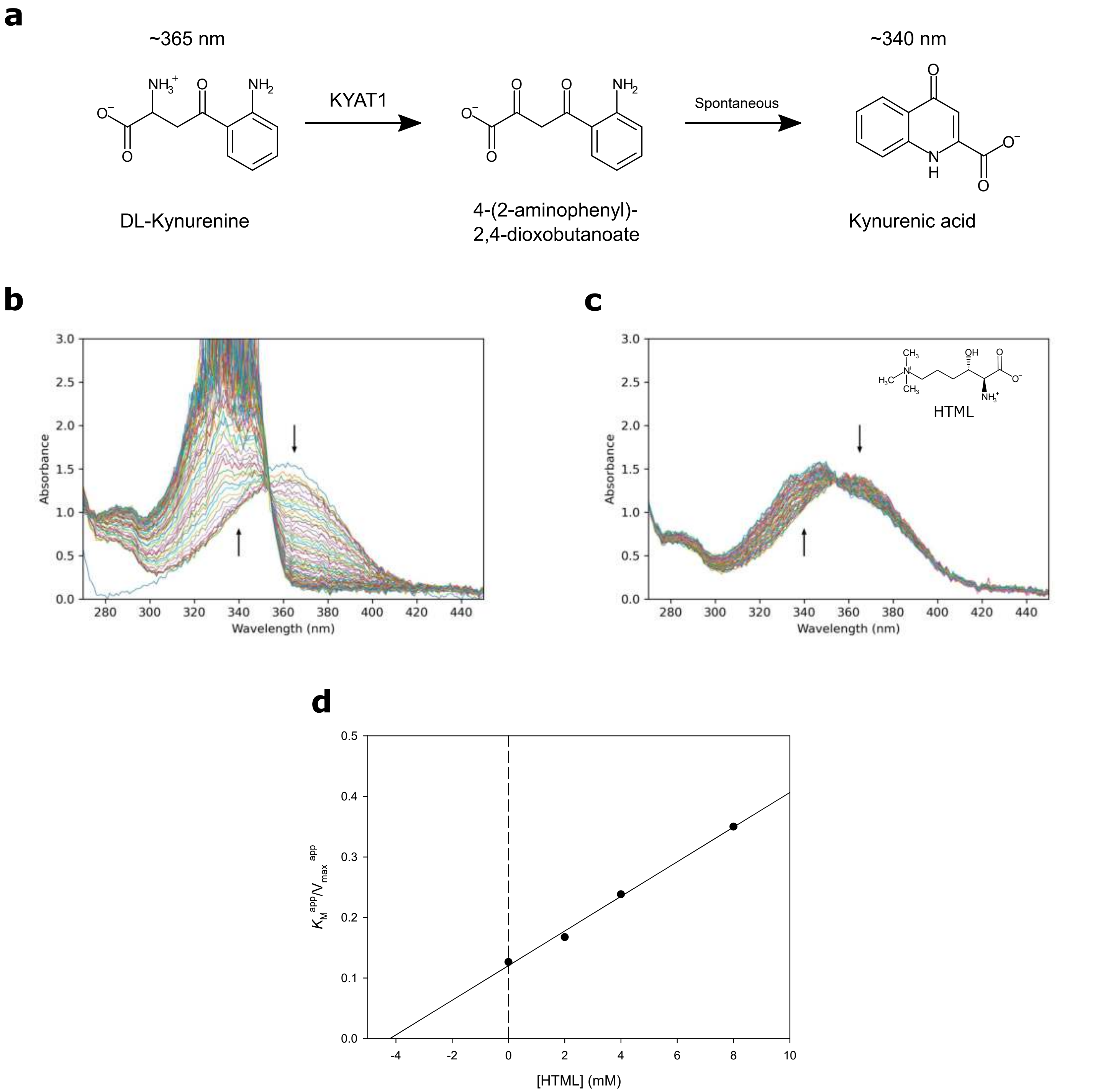


Supplementary Fig. 13: ^{1}H NMR spectra of HTMLA reaction

a, Detail of ^{1}H NMR spectrum of aldehyde proton of TMABA after 1100 minutes of SHMT1 reaction in dehydrated (9.63) and hydrated (5.05 ppm) form. **b**, Time-resolved ^{1}H NMR spectra of Tha1 activity in the presence of 5 mM HTML at 40, 390 and 1170 minutes. Inset shows aldehyde proton of TMABA in dehydrated (9.63) and hydrated (5.05 ppm) form.

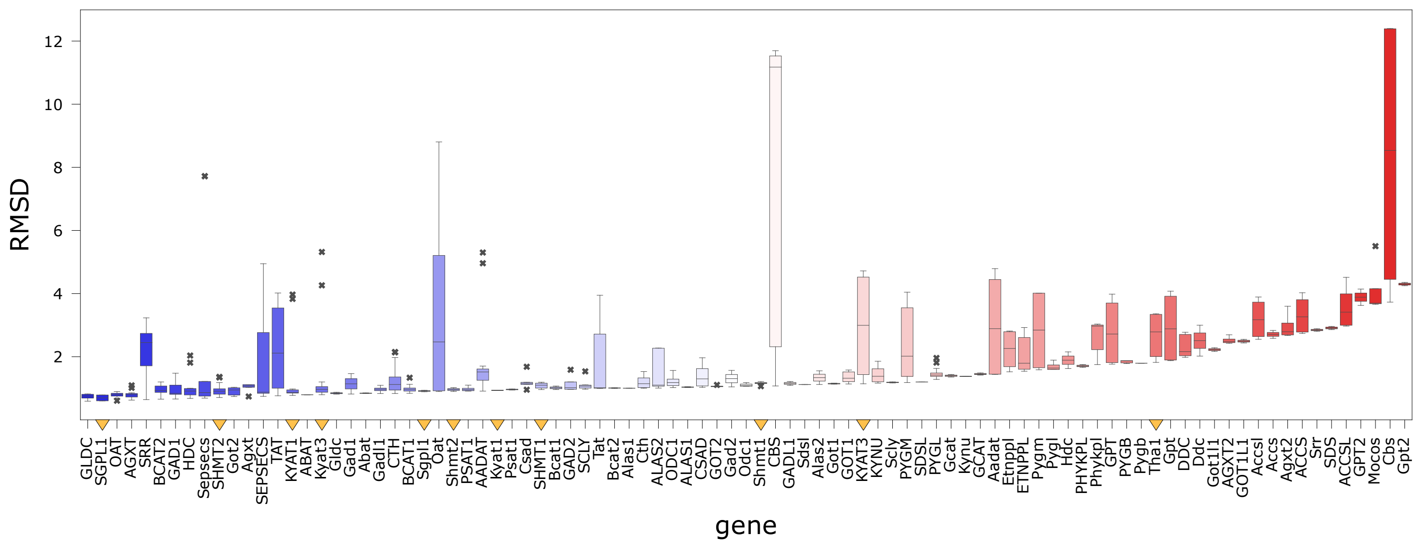


Supplementary Fig. 14: Michaelis-Menten fitting of aldol cleavage of human SHMTs, Tha1 and *e*TA. Michaelis -Menten plots of aldolase activity towards: HTML for Tha1 (**a**), SHMT2 (**g**) and *e*TA (**h**); L-*allo*-threonine for SHMT1 (**d**) Tha1 (**b**), SHMT2 (**e**) and *e*TA (**f**); L-threonine for Tha1 (**c**), SHMT1 (**i**), SHMT2 (**j**).

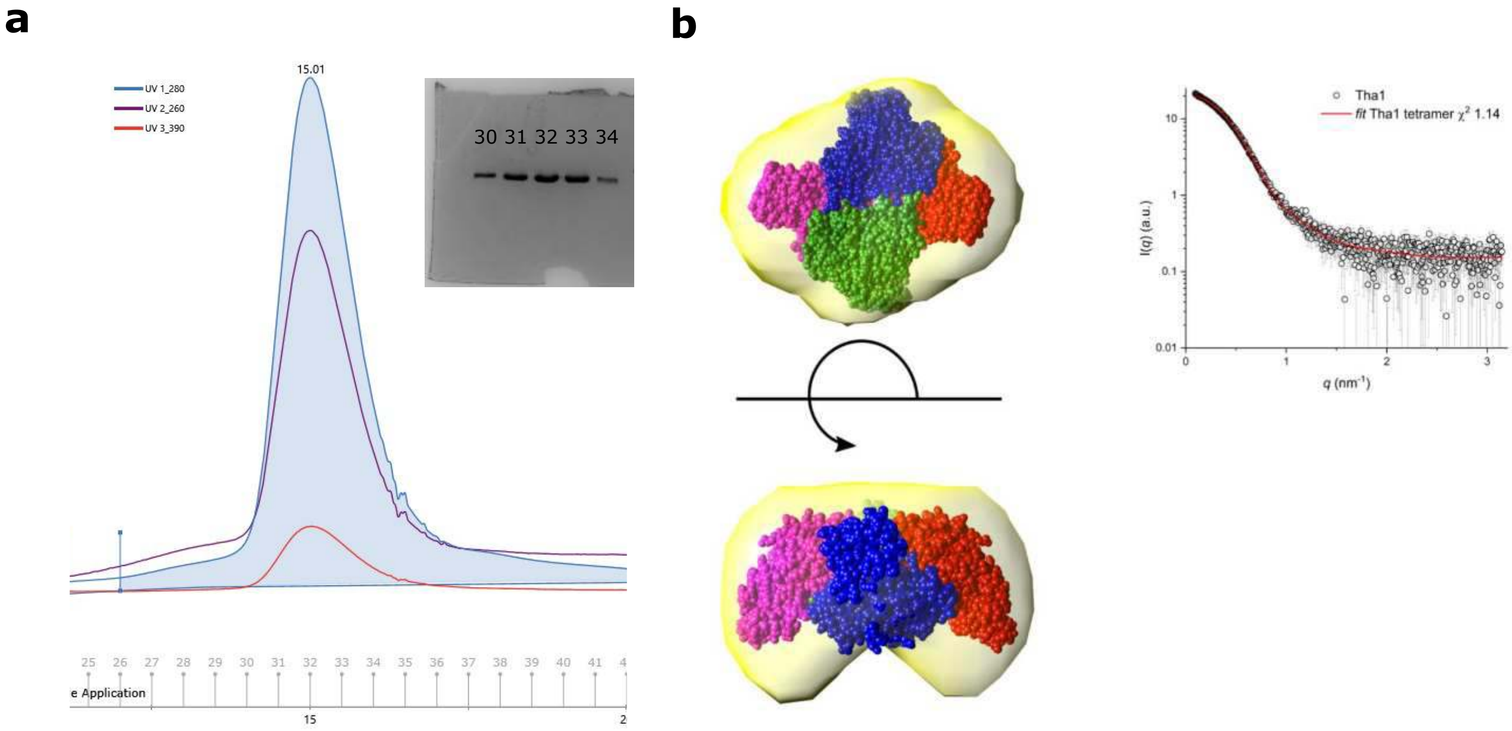


Supplementary Fig. 15: Characterization of HTML as inhibitor of human KYAT 1

a, Transaminase reaction of kynurenine (365 nm) by KYAT1 and subsequent spontaneous reaction forming kynurenic acid (340 nm). **b,c** Time-resolved UV-Vis spectra showing the conversion of DL-kynurenine (0.25 mM) to kynurenic acid. Spectra were collected every 30 seconds for 30 minutes, in the absence (**b**) or in the presence (**c**) of 0.5 mM HTML. **d**, Lineweaver-Burk double-reciprocal secondary plot of the inhibition by HTML of kynurenine aminotransferase activity of KYAT1 .

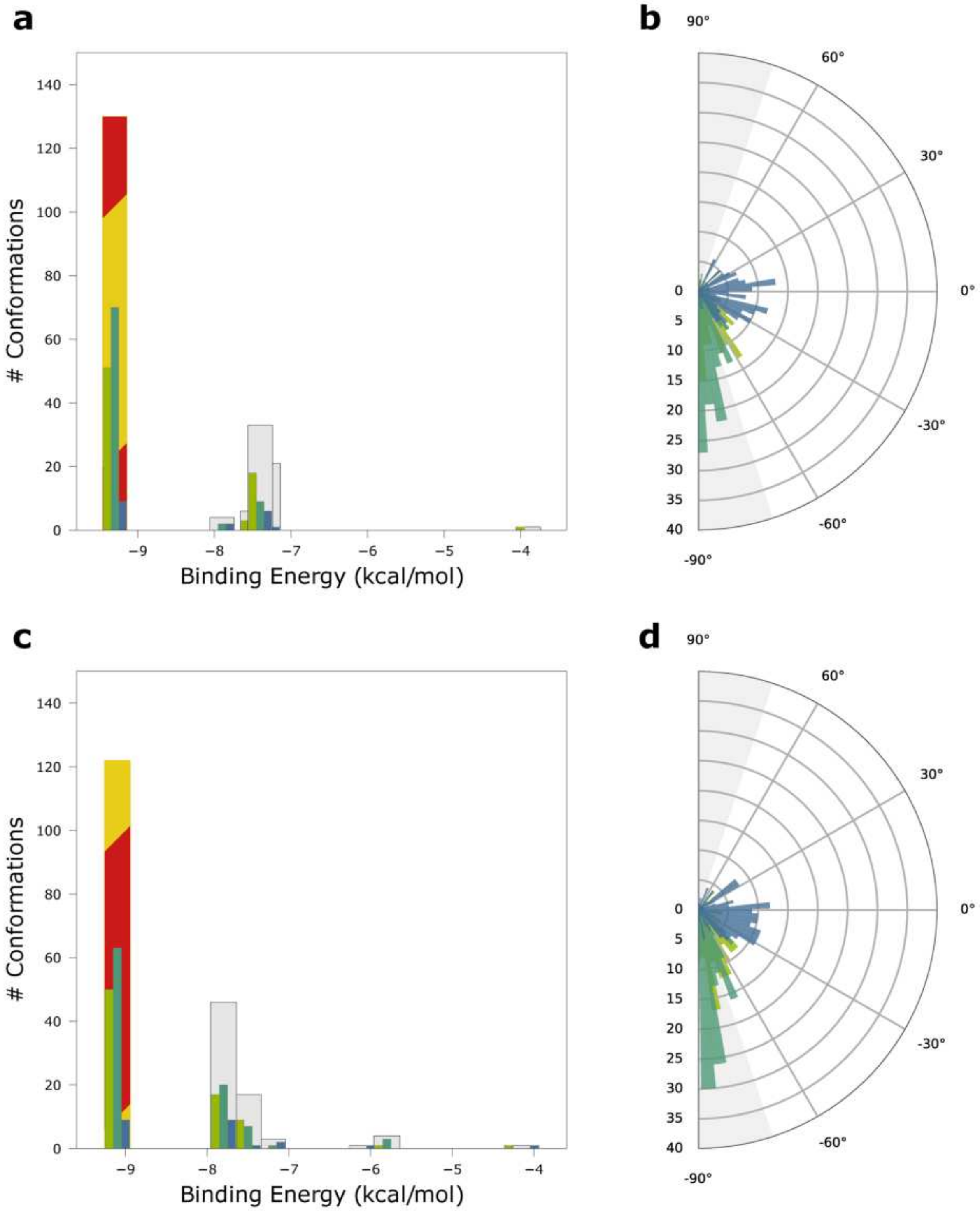


Supplementary Fig. 16: RMSD between AlphaFold and SWISS-MODEL models of enzyme set. Boxplot of Cα RMSDs calculated by the alignment between the AlphaFold models and all the corresponding templates in SMR. X-axis is ordered by minimum values and colored with a blue-white-red gradient. Genes encoding proteins selected for experimental validation are highlighted with yellow arrowheads on the x-axis.



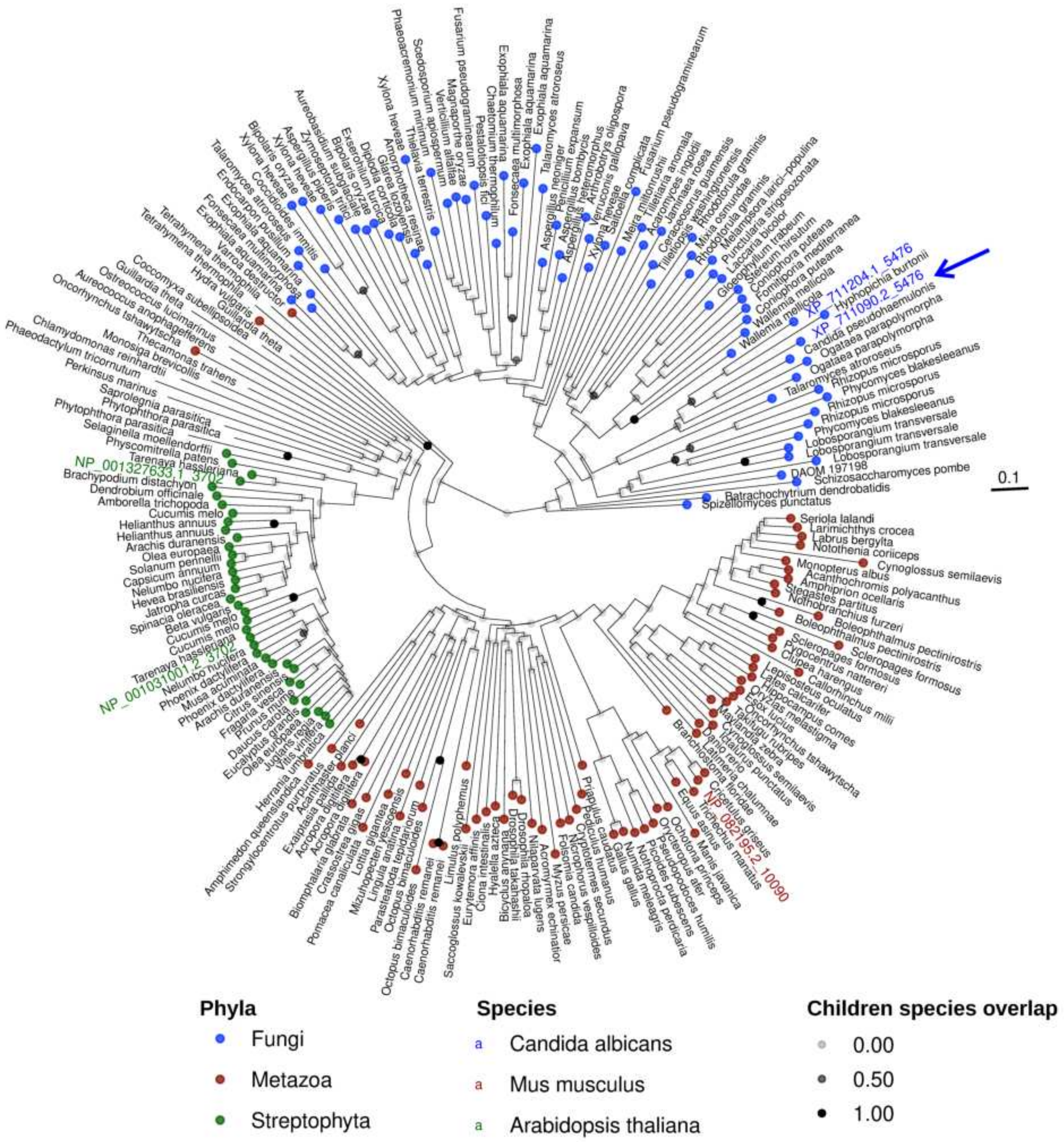
Supplementary Fig. 17: Quaternary structure characterization of mouse Tha1.

a, Size exclusion chromatography of Tha1 for crystallization experiments. Inset shows SDS-PAGE of the fractions. **b**, SAXS analysis of Tha1/HTMLA shows that the enzyme is tetrameric in solution.



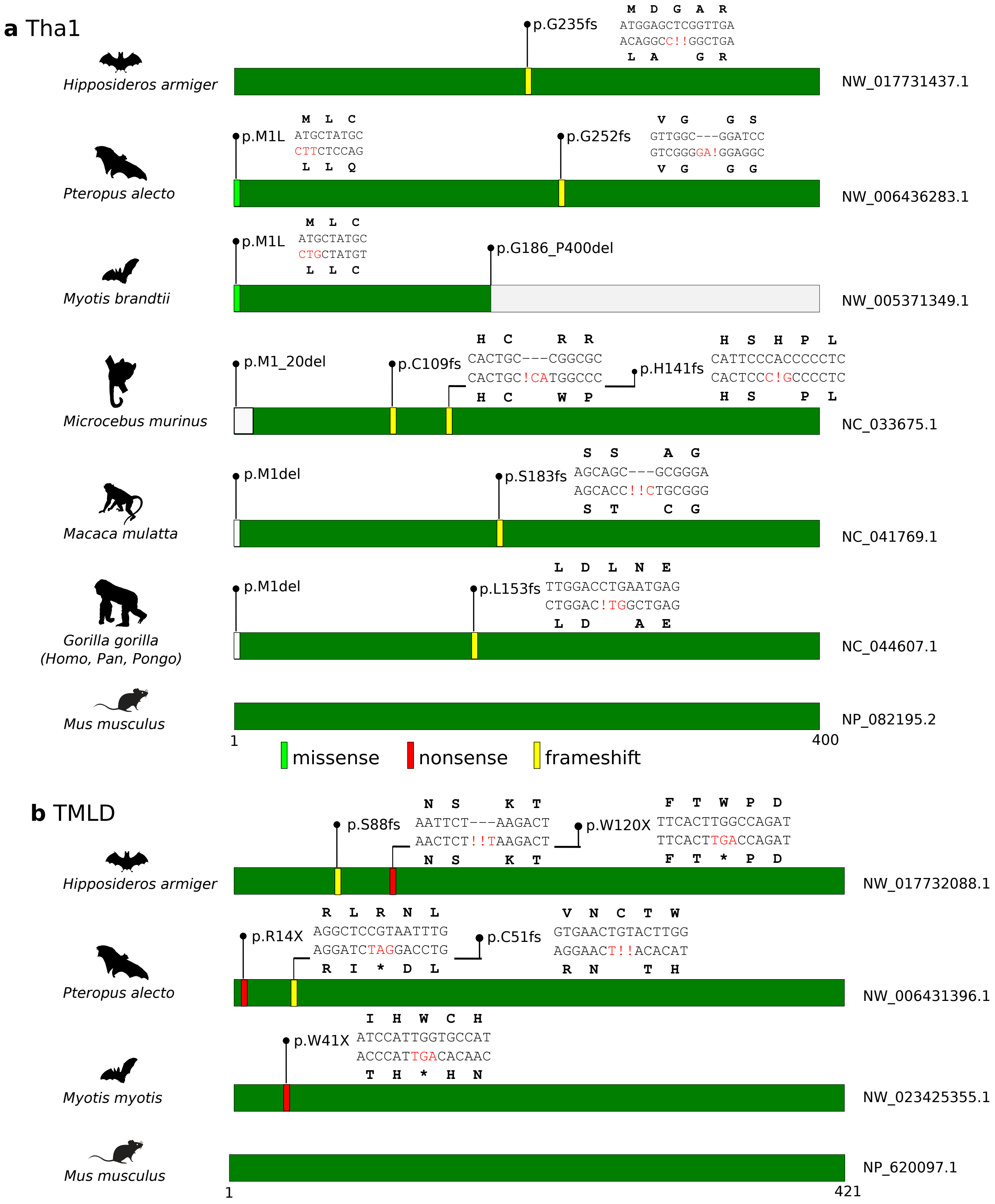
Supplementary Fig. 18: Docking results with different models of Tha1.

Clustering of the HTML-PLP conformations at the Tha1 active site obtained with HTML-OSMES applied to AlphaFold model obtained by adding Tha1 crystal structure as template (**a, b**) or AlphaFold model obtained by using only Tha1 crystal structure as template (**c, d**). Bar plots show the distribution of χ_1 (blue), χ_2 (emerald) and χ_3 (kiwi) angles in each cluster, with the best and the largest cluster highlighted in red and yellow respectively. Circular plots show the cumulative distribution of the three χ angles for all clusters. $|\sin(\chi)| \geq 0.95$ values are defined by gray areas.



Supplementary Fig. 19: Phylogeny of Tha1 in eukaryotes.

Midpoint-rooted neighbor-joining phylogenetic tree of Tha1 in representative eukaryotes. The tree shows segregation of sequences from fungi (blue tips), plants (green tips) and metazoa (red tips). The *Candida albicans* protein XP_711090 previously characterized as HTMLA [1] is indicated with an arrow. Nodes with possible gene duplication, as inferred by the proportion of overlapping species at children nodes [2] are indicated by dark circles.



Supplementary Fig. 20: Pseudogenization of Tha1 and TMLD in mammals

Examples of inactivating mutations in a) Tha1 of Chiroptera and Primates and b) TMLD of Chiroptera. Mutations disrupting the coding sequence are indicated with reference to the mouse Tha1 (NP_082195.2) and TMLD (NP_620097.1) protein sequences according to the Huret nomenclature (http://atlasgeneticsoncology.org/). Codons with disrupting mutations are colored red. Exclamation marks represent frameshift insertions/deletions. Reconstruction of pseudogene sequences was based on tblastn searches followed by GenWise analysis and Macse alignments.

References

1. Strijbis, K. *et al.* Identification and characterization of a complete carnitine biosynthesis pathway in *Candida albicans*. *FASEB J.* **23**, 2349–2359 (2009).
2. Van Der Heijden, R. T., Snel, B., Van Noort, V. & Huynen, M. A. Orthology prediction at scalable resolution by phylogenetic tree analysis. *BMC Bioinformatics* **8**, 83 (2007).