

VILNIUS UNIVERSITY
LIFE SCIENCES CENTRE
INSTITUTE OF BIOCHEMISTRY



MATAS TIŠKUS

Molecular Biotechnology

**Molecular modelling and the study of the structure-function relationship of
cytidine deaminases**

Master thesis

Supervisors: dr. Nina Urbelienė

prof. Rolandas Meškys

Student: Matas Tiškus

Vilnius, 2022

Table of Contents

LIST OF ABBREVIATIONS	3
INTRODUCTION.....	4
1 LITERATURE OVERVIEW	6
1.1 CYTIDINE DEAMINASES	6
1.1.1 Overall cytidine deaminase structure.....	8
1.1.2 Cytidine deaminase active site structure.....	12
1.1.3 Cytidine deaminase reaction mechanism.....	14
1.2 CYTIDINE DEAMINASE INTERACTION WITH SUBSTRATES.....	16
1.2.1 CDA interaction with the sugar moiety	17
1.2.2 CDA interactions with the base moiety	18
1.3 PROTEIN ENGINEERING METHODS.....	19
1.3.1 Examples of engineered cytidine deaminase enzymes	20
2 MATERIALS AND METHODS	23
2.1 MATERIALS	23
2.2 METHODS	25
1.2.1 Site-directed mutagenesis primer creation.....	25
1.2.2 Site-directed mutagenesis primer phosphorylation.....	26
1.2.3 Site-directed mutagenesis PCR reaction.....	26
1.2.4 DNA electrophoresis and DNA fragment purification for agarose gels.....	26
1.2.5 Linear DNA fragments ligation	27
1.2.6 Competent cell preparation.....	27
1.2.7 Electroporation.....	27
1.2.8 Chemical transformation.....	28
1.2.9 Plasmid purification.....	28
1.2.10 Enzymatic activity testing using agar selective medium	28
1.2.11 Protein expression and purification	29
1.2.12 Proteins fractionation by SDS-PAGE.....	29
1.2.13 Bradford assay for protein concentration determination	29
1.2.14 Enzyme activity assessment spectrophotometrically.....	30
1.2.15 Thin-layer chromatography	30
1.2.16 Enzyme kinetic parameter determination	30
1.2.17 Protein structure modelling.....	30
1.2.18 Molecular dynamics.....	31
1.2.19 Molecular docking.....	32
1.2.20 Assessing enzyme binding pocket SASA relationship with substrate selectivity	32
3 RESULTS AND DISCUSSION	33
3.1 SEQUENCE ANALYSIS AND RELATIONSHIP WITH ENZYME SPECIFICITY	34
3.2 STRUCTURE MODELLING	35
3.3 CDA_F14 INTERACTIONS WITH BZDC	39
3.4 CDA_F14 MUTATIONAL ANALYSIS	40
3.5 CDA_F14 ENZYME KINETICS AND SUBSTRATE SPECIFICITIES	42
CONCLUSIONS	46
SUMMARY	47
SANTRAUKA	48
LITERATURE	50

LIST OF ABBREVIATIONS

CDA – cytidine deaminases acting on free pyrimidine nucleosides

dCDA – dimeric CDA

tCDA – tetrameric CDA

C – cytidine

U – uridine

dC – 2'-deoxycytidine

dU – 2'-deoxyuridine

BzdC – *N*⁴-benzoyl-2'-deoxycytidine

THU – tetrahydro-2'-deoxyuridine

dUMP – 2'-deoxyuridine monophosphate

dTMP – 2'-deoxythymidine monophosphate

ssDNA – single-stranded DNA

APOBEC – Apolipoprotein B mRNA editing catalytic polypeptide-like, acts on ssDNA or RNA

AID – activation induced cytidine deaminase, belonging to the APOBEC family

SNP – single nucleotide polymorphism

MD – molecular dynamics

RMSD – root mean square deviation

RMSF – root mean square fluctuation

INTRODUCTION

Cytidine deaminases are a large family of enzymes sharing the same structural core, but performing many different tasks in our organisms. The task of tetrameric and dimeric cytidine deaminases is the recycling of pyrimidines by converting free cytidine into uridine which then can also be turned into thymidine. APOBEC family cytidine deaminases also serve an important role in higher animals as a defence mechanism against viruses and enablers of the vast pool of possible antibody variations our bodies can produce. The ability to deaminate cytidine also gives cytidine deaminases the ability to interact with various chemotherapy agents based on pyrimidines. These interactions can make it difficult to create an effective treatment regimen without side effects because cytidine deaminase activity modulates the efficacy of these drugs (Frances & Cordelier, 2020). This modulating effect can also come from exogenous sources of cytidine deaminase activity as is the case with intratumor bacteria conferring resistance to treatment (Geller & Straussman, 2018).

Even though cytidine deaminases are important for effective cancer treatment, most of the studies published on cytidine deaminases are focused on using the ability of the APOBEC family cytidine deaminases to deaminate cytidine in single-stranded DNA or RNA molecules. This enables an efficient editing mechanism that doesn't rely on double-stranded breaks and homology repair mechanisms to create specific modifications. Although only a single kind of alteration can be performed, a C to T transition, this method reduces errors that can be caused by non-homologous end joining and enables specific mitochondrial DNA editing without reliance on mitochondria DNA degradation, because it doesn't require RNA, which is difficult to transport into mitochondria, as a guide system like in the case of CRISPR-Cas9 (Huang et al., 2021).

This study, on the other hand, focuses on cytidine deaminase acting on free cytidine. Several tetrameric cytidine deaminases, discovered while developing metagenomic libraries for amidohydrolase selection, exhibited before unseen catalytic capabilities. These enzymes were able to deamidate N^4 -acyl-/ N^4 -alkyl-, N^4 -carboxy, S^4 -alkyl- and O^4 -alkoxy- substrates with various substitutes. It has been known for a long time that the *Escherichia coli* dimeric cytidine deaminase can deaminate N^4 -methylcytidine to uridine (Cohen & Wolfenden, 1971), but in this study, a wide range of cytidine deaminase substrates was shown. This work aims to understand what determines this broad spectrum of catalytic activities of the studied enzymes. A deeper understanding of the factors contributing to discovered substrate specificities could enable novel prodrugs activated by the discovered enzymes to be used, this, for example, can reduce or eliminate the impact varying degrees of cytidine deaminase activity has on the efficacy of cancer treatment. To this end, these enzymes were modelled and using docking and molecular dynamics simulations regions of interest were selected. These selected regions were then mutated to evaluate their effect on the discovered

enzymatic activities. The effects were evaluated by substrate spectrum and kinetic parameters determination.

The goal of the thesis:

To identify regions of tetrameric cytidine deaminases, which affect the catalysis of nucleophilic substitution reaction at the 4th position of the heterocyclic ring in pyrimidine nucleoside analogues.

Objectives:

1. To model the tertiary and quaternary structures of studied cytidine deaminases.
2. To select the regions influencing the ability to deaminate N^4 substituted pyrimidine analogues by using molecular dynamics and docking simulations.
3. To create mutant cytidine deaminase variants of the selected regions and to evaluate the effect of mutations on the enzyme's ability to deaminate N^4 substituted pyrimidines.

1 LITERATURE OVERVIEW

1.1 Cytidine deaminases

Cytidine deaminases (CDA) (EC 3.5.4.5) are enzymes that participate in the recycling of pyrimidines by deaminating cytidine (C) and deoxycytidine (dC) into uridine (U) and deoxyuridine (dU) respectively. Pyrimidine nucleotides are necessary for various biological processes like DNA and RNA synthesis, and maintenance of both nuclear and organellar genomes, in humans deficiencies in CDA activity lead to disorders like Bloom syndrome (Pedroza-García et al., 2019). Both C and dC can be converted into uracil and used for nucleotide synthesis, alternatively, uracil can be catabolized into β -alanine, such conversion provides cells with both carbon and nitrogen (Figure 1.1). CDA's involved in nucleotide salvage belong to the cytidine deaminase superfamily, require zinc-binding for their catalytic activity and are found in all living organisms. The zinc ion is coordinated by either a histidine and two cysteines or three cysteine residues. CDA's acting on free C or dC are either homodimeric (dCDA) (*Escherichia coli* CDA) or homotetrameric (tCDA) (human CDA) in nature (Chung et al., 2005; Frances & Cordelier, 2020).

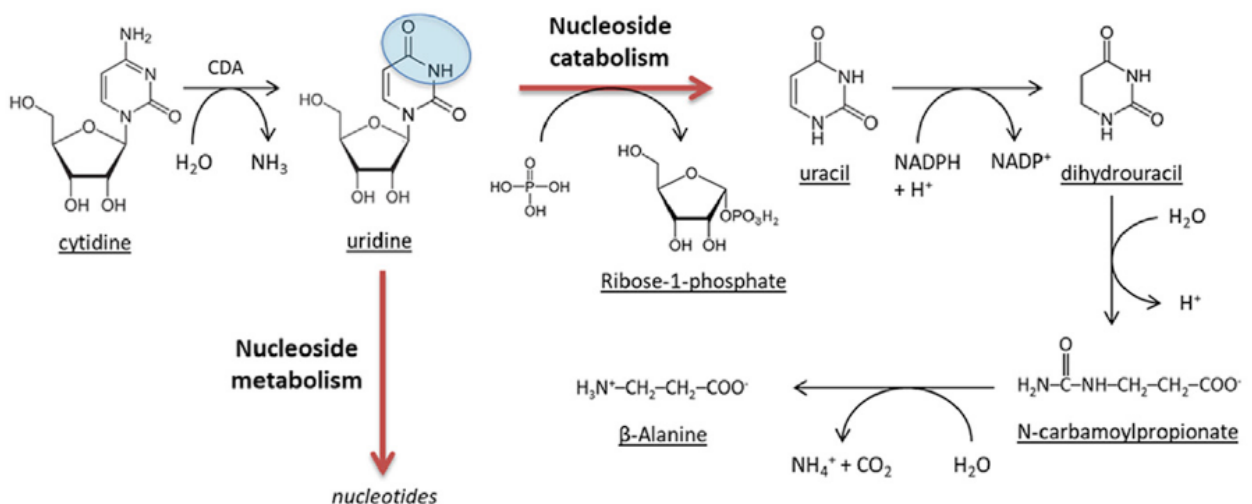
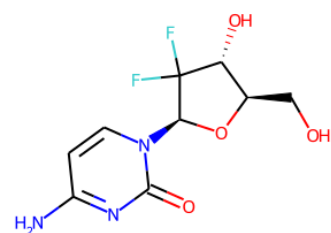


Figure 1.1 Cytidine recycling pathway by CDA (Frances & Cordelier, 2020)

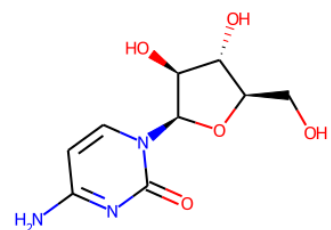
Apolipoprotein B mRNA editing catalytic polypeptide-like (APOBEC) (EC 3.5.4.1) family of cytidine deaminases also belong to the deaminase superfamily, but in contrast to CDA's, they act on single-stranded DNA (ssDNA) or RNA molecules. There is a total of eleven APOBEC family proteins in humans which have a wide spectrum of functions. For example, APOBEC1 (A1) regulates lipid uptake by introducing an early stop codon into apolipoprotein B mRNA and regulating the levels of low-density lipoproteins (LDL), which are heart disease risk factors (Chen, 2021). APOBEC3G

restricts the replication of human immunodeficiency virus (HIV), hepatitis B virus and retroelements by cytidine deamination in ssDNA, thus hypermutating guanosine to adenosine, or by RNA binding (Holden et al., 2008). Another member of the APOBEC family – activation-induced cytidine deaminase (AID), is an essential enzyme for the immune system through its role in somatic hypermutation (SHM), class switch recombination (CSR), variable-diversity-joining (VDJ) and gene conversion. All of these processes enable the production of diverse immunoglobulins (Navaratnam & Sarwar, 2006). In addition, AID plays a role in DNA demethylation through its ability to deaminate 5-methylcytosine (Rios et al., 2020). Deregulation of APOBEC expression or cellular localization is also proven to lead to genome hypermutation and cancer (Henderson & Fenton, 2015).

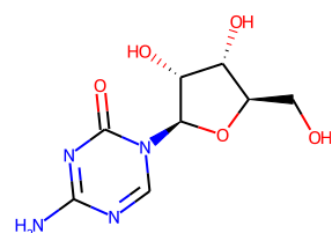
In humans CDA's are mostly expressed in the liver, bone marrow and placenta, they are also found in mature neutrophils (Micozzi et al., 2014). CDA's are interesting because of their role in cancer treatment efficacy. Various pyrimidine analogues are used for cancer treatment. These include gemcitabine (Lvarez et al., 2012), cytarabine (J. Liu et al., 2016), azacytidine (Hattori et al., 2019), decitabine (Dhillon, 2020), and capecitabine (Walko & Lindley, 2005) (Figure 1.2). These drugs work mainly by incorporating into RNA and DNA molecules, interfering with protein and nucleic acid synthesis and inhibiting the thymidylate synthase, which converts 2'-deoxyuridine monophosphate (dUMP) to 2'-deoxythymidine monophosphate (dTMP) and is important in the pyrimidine salvage pathway (Serdjebi et al., 2015; Warner et al., 2014). Gemcitabine, cytarabine, azacytidine and decitabine are inactivated by deamination by cytidine deaminase, whereas one of the steps to metabolize capecitabine into its active form, 5-fluorouracil, requires deamination by cytidine deaminase. Deoxycytidine analogue drugs like torcitabine and (+)-b-2',3'-dideoxy-5-fluoro-3'-thiacytidine used for viral treatment are also inactivated by cytidine deaminases (Jansen et al., 2011). The *tCDA* gene (gene map locus 1p36.2-p35) in humans is rather polymorphic



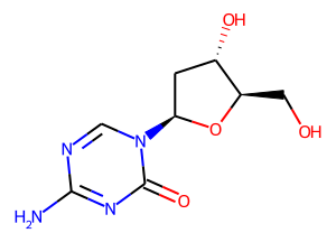
Gemcitabine



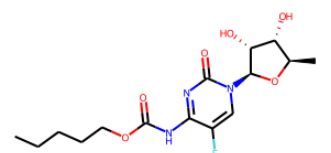
Cytarabine



Azacytidine



Decitabine



Capecitabine

Figure 1.2 Pyrimidine analogues used for chemotherapy

and has multiple single nucleotide polymorphisms (SNP) that can change the enzymes activity levels and thus influence treatment efficacy. For example, Q27K/A70T variant of human tCDA has low activity towards cytosine based chemotherapeutic agents, consequently patients harboring this tCDA variant might be more responsive to treatment using these drugs, on the other hand, the dose also needs to be adjusted, as lower tCDA activity can lead to severe drug toxicity (Micozzi et al., 2014). Intratumor *Escherichia coli* and *Mycoplasma hyorhinitis* bacteria in pancreatic and colorectal tumours can also induce resistance to gemcitabine through its deamination, this further complicates treatments with pharmaceuticals susceptible to CDA inactivation (Geller et al., 2017; Geller & Straussman, 2018).

1.1.1 Overall cytidine deaminase structure

There are quite a few tCDA and dCDA structures published. Examples of tetrameric CDA include *Bacillus subtilis* tCDA (PDB: 1JTK) (Johansson et al., 2002), mouse tCDA (PDB: 2FR6) (Teh et al., 2006), *Mycobacterium tuberculosis* tCDA (PDB: 3IJF) (Zilpa A. Sánchez-Quitian et al., 2011), *Saccharomyces cerevisiae* tCDA (PDB: 1R5T) (Xie et al., 2004) and human tCDA (PDB: 1MQ0) (Chung et al., 2005). Dimeric CDA structures include *Klebsiella pneumoniae* dCDA (PDB: 6K63) (W. Liu et al., 2019), *Arabidopsis thaliana* dCDA (PDB: 6L08) (Wang et al., 2020), *Escherichia coli* dCDA (PDB: 1ALN) (Xiang et al., 1996). Published APOBEC protein structures include APOBEC2 (PDB: 2NYT) (Prochnow et al., 2007) and APOBEC3G (PDB: 3IQS) (Holden et al., 2008).

The structures of tCDA and dCDA are quite similar and it is possible to superimpose them with an RMSD value between the alpha carbon atoms that is lower than 1.7 Å (Johansson et al., 2002). Each domain in the tCDA is around 15 kDa and 140 amino acids long. Overall, the protein maintains a 222 symmetry. The core monomer structure contains 5 alpha spirals and 5 beta sheets, where the 5 beta sheets are in between one alpha spiral from the N terminus and the remaining four alpha spirals from the other side forming an $\alpha/\beta/\alpha$ sandwich structure which is a typical CDA domain structure (Figure 1.3). Each monomer in tCDA has an active site and each active site is made from three surrounding monomers. It is noted that around ten residues at the C end of the monomer do not have a defined structure, because this region is located near the active site entrance of adjacent monomers it might play a role in substrate binding by opening and closing the active site (Johansson et al., 2002; Teh et al., 2006). The interactions between the tCDA subunits are mostly of a hydrophobic nature (Zilpa A. Sánchez-Quitian et al., 2010). The conserved tCDA amino acids corresponding to Ser25, Arg93, Gln94, Glu98, Leu118 in *Mycobacterium tuberculosis*, Ser22, Arg90, Gln91, Glu95, Leu124 in *Bacillus subtilis* and Ser34, Arg103, Gln104, Glu108, Leu132 in Human

tCDA are shown to be important for the interactions between subunits (Carlow et al., 1999). They have also been shown to form hydrogen bonds that help maintain quaternary protein structure, for example, Gly90 to Gln94 and Arg93 to Pro120 in *Mycobacterium tuberculosis* tCDA. Other conserved amino acids like Tyr21/24/51 in *Mycobacterium tuberculosis* tCDA also seem to be important for the interaction between subunits through π - π interactions and hydrogen bonds. Tyr51 interacts with the same Tyr51 from a subunit across through π - π stacking (Figure 1.3 A – subunit A interaction with subunit D). Tyr21 and Tyr24 form hydrogen bonds with conserved Glu108 from an adjacent subunit (Figure 1.3 A – subunit A interaction with subunit B) (Zilpa A. Sánchez-Quitian et al., 2010, 2011). Tyr60 Y60G mutant in human tCDA (Tyr51 equivalent in *Mycobacterium tuberculosis*) had similar V_{max} values to wild-type enzyme but the K_m values were significantly increased for both C and dC. The reduction in affinity

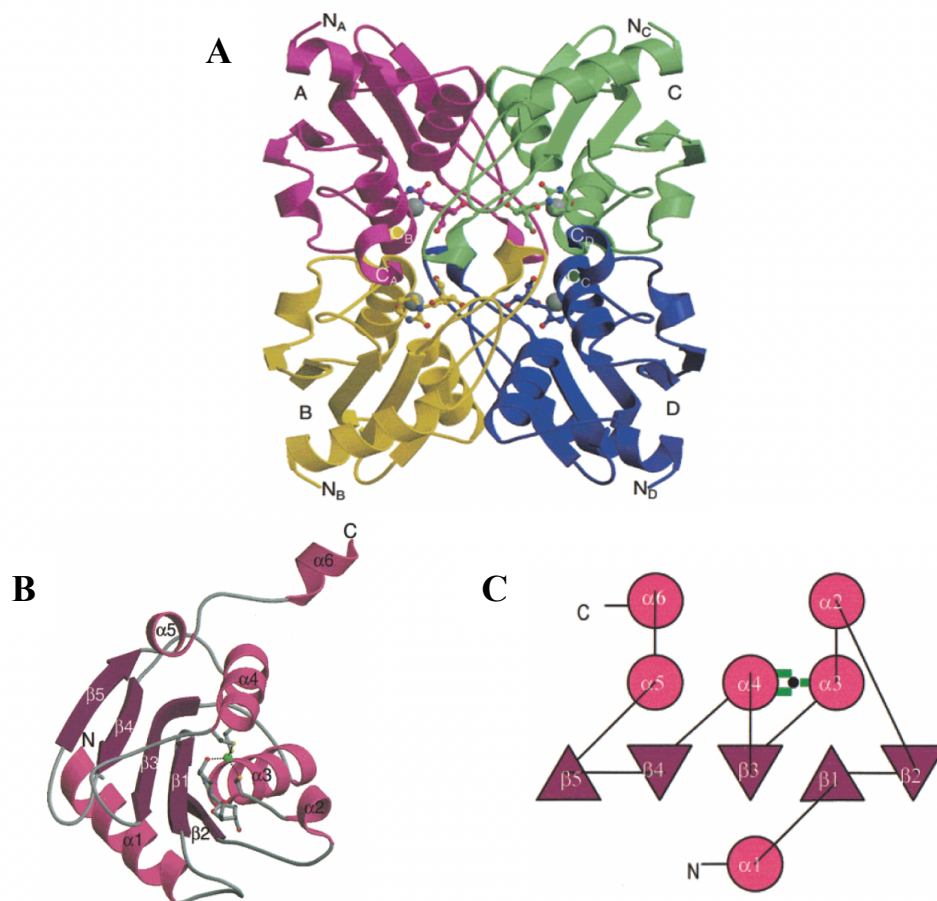


Figure 1.3 *Bacillus subtilis* tCDA structure (PDB: 1JTK). **A** – *Bacillus subtilis* tCDA quaternary structure, with zinc in grey and a bound inhibitor 3,4,5,6-tetrahydro-2'-deoxyuridine represented in stick and ball representation, **B** – *Bacillus subtilis* tCDA monomer tertiary structure, **C** – schematic of *Bacillus subtilis* tCDA secondary structure arrangement, active site cysteines shown as green rectangles (Johansson et al., 2002)

for substrates in the Y60G variant might come from the disturbed hydrogen bonding of the substrate ribose 5'-OH and the Tyr60 backbone -NH, because the mutation into glycine causes a more compact quaternary structure, as a side effect this mutation also infers high resistance to tetrameric structure dissociation by SDS (Costanzi et al., 2011). Tyr33 in human tCDA (Tyr24 equivalent in

Mycobacterium tuberculosis tCDA) was proven to be necessary for tCDA tetrameric structure formation, a Y33G mutation led to protein aggregation and degradation, without added chaperones, showing Tyr33 importance for correct folding and tetrameric structure formation of tCDA (Micozzi et al., 2010). Another study also investigated Y33F and Y33S mutations, both variants were also inactive and formed insoluble inclusion bodies further showcasing the importance of this position for the correct tetrameric structure formation (Costanzi et al., 2011). Conserved Asn48 in *Mycobacterium tuberculosis* tCDA also forms a hydrogen bond with a conserved Gln94 from an adjacent subunit just like Tyr21 and Tyr24 (Zilpa Adriana Sánchez-Quitian et al., 2015).

The dimeric cytidine deaminase might have arisen from a duplication of the *tCDA* gene because the dCDA monomer mimics the asymmetric unit of the tCDA (Teh et al., 2006). The dCDA monomer is formed from a catalytic N-terminal domain and a non-catalytic C-terminal domain. Both

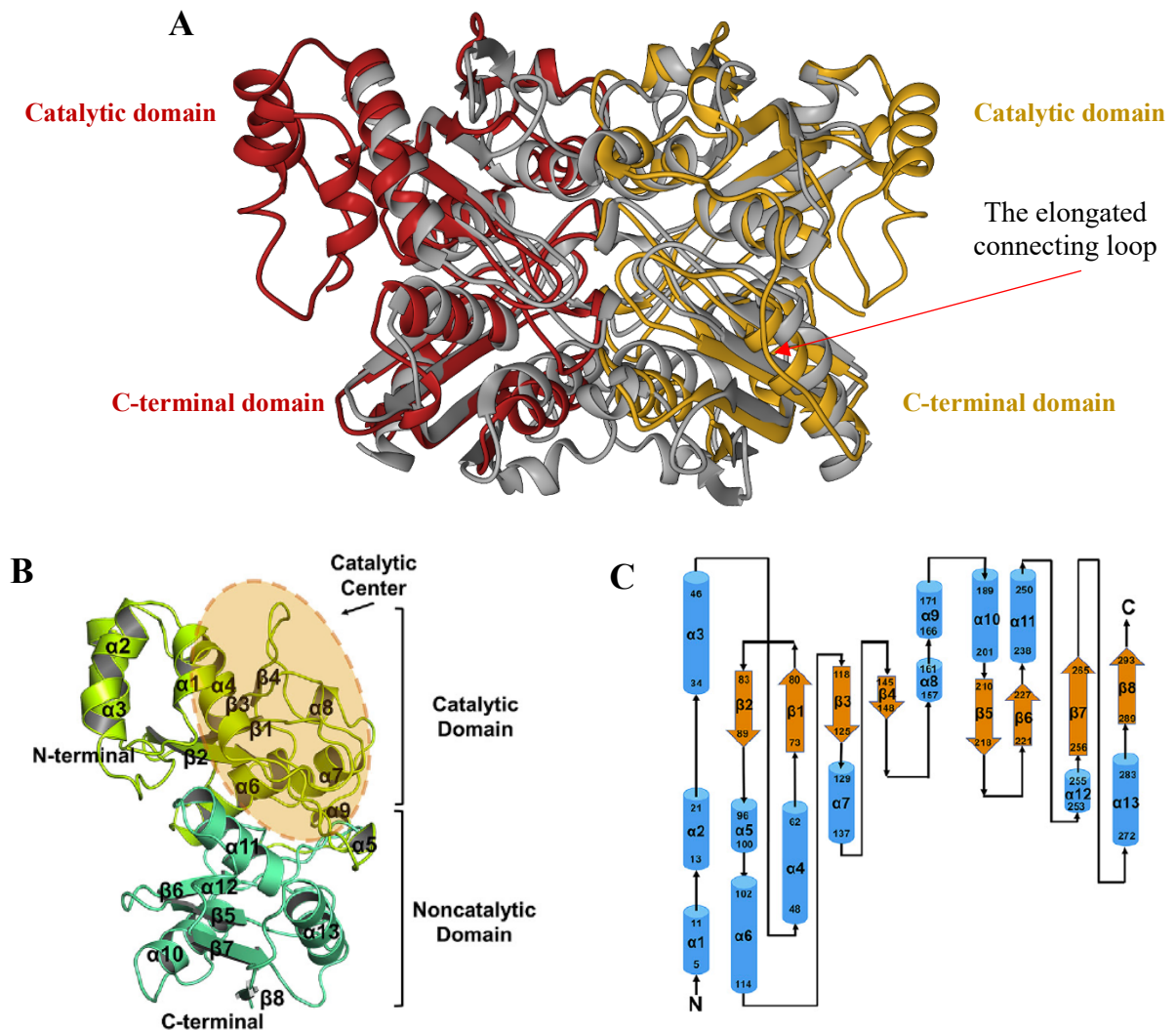


Figure 1.4 *Klebsiella pneumoniae* dCDA structure (PDB: 6K63). **A** – *Klebsiella pneumoniae* dCDA quaternary structure (red, yellow) superimposed onto *Escherichia coli* tCDA structure (grey), **B** – *Klebsiella pneumoniae* dCDA monomer structure, **C** – schematic of *Klebsiella pneumoniae* dCDA secondary structure arrangement (W. Liu et al., 2019)

domains of the monomer adopt the same structure of beta sheets being sandwiched in between alpha spirals, the classical CDA fold structure. The two domains of a dCDA monomer are connected through an elongated C terminal domain loop that is also found in tCDA (Figure 1.4), this loop connects the fourth beta strand with the eighth alpha spiral. The monomers usually contain 13 alpha helices and 8 beta sheets. The alpha helix number is higher and the beta sheet number lower than what you would expect from a fusion of two tCDA subunits, this is because dCDA has an elongated N terminal domain and a shorter C terminal domain than tCDA. Consequently, dCDA loses two helices and a beta strand corresponding to $\alpha 5$, $\alpha 6$ and $\beta 5$ in tCDA (Figure 1.3 C) (Johansson et al., 2002; W. Liu et al., 2019). The main difference between dCDA and tCDA is that even though dCDA mimics the overall structure of tCDA it only has two active sites compared to four in tCDA. The C terminal domain of dCDA lacks the necessary amino acids for coordinating a zinc ion and ensuring the catalytic reaction thus it is inactive. There are also more differences in the amino acids that coordinate the zinc ion, in dCDA the zinc ion is coordinated by two cysteines and a single histidine, whereas in tCDA it is coordinated by three cysteine residues. But the mechanism of deamination and the main interactions with substrates are the same for both dCDA and tCDA enzymes (Wang et al., 2020).

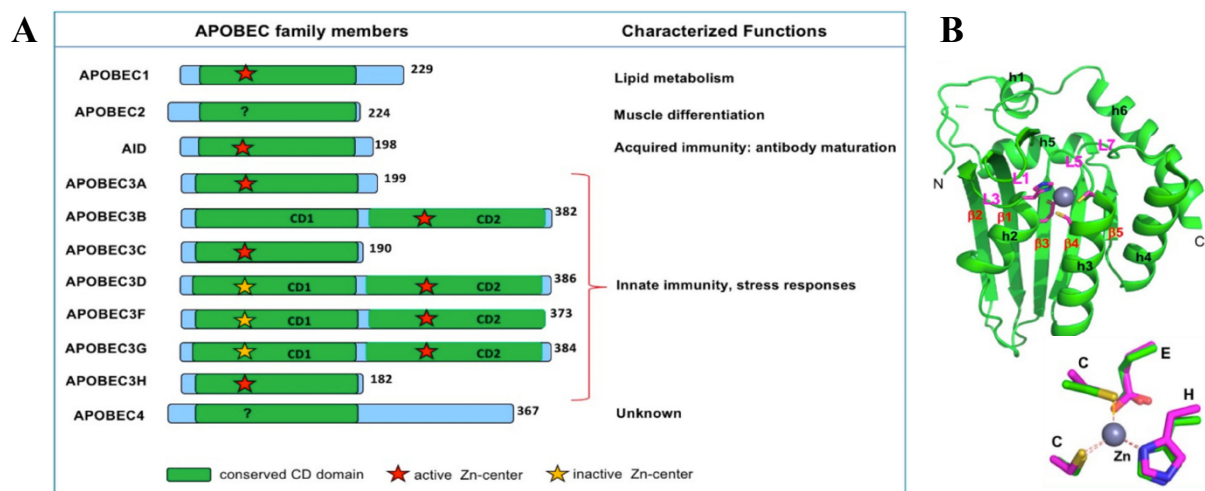


Figure 1.5 APOBEC family proteins. A – APOBEC family proteins with their function and domain organisation, B – APOBEC2 (A2) (PDB: 2NYT) tertiary structure and active site residues (Chen, 2021)

The APOBEC family is thought to have originated in jawless fish lymphocytes 500 million years ago. Although the functions of the APOBEC family CDA substantially differ from CDA's involved in nucleotide salvage pathways they still maintain the overall $\alpha/\beta/\alpha$ sandwich structure. Many APOBEC family members form higher oligomeric structures together and complexed with RNA. AID and APOBEC (AC3) seem to be monomeric from crystal structure analysis, but other APOBEC family proteins with known structures (A1, A2, A3A, A3H, A3G) were crystallized as

dimers with the potential to organize into even higher oligomeric states. Usually, APOBEC proteins have a single active site but some also have additional inactive sites, all the active sites, including the inactive ones, contain a conserved **H**-[P/A/V]-**E**-X[23–28]-**P**-**C**-X[2-4]-**C** (X any amino acid) motif (Figure 1.5 A). The zinc ion just like in dCDA is coordinated by two cysteines and a single histidine residue (Figure 1.5 B). The catalytic activity and substrate selection are determined mainly by loops and the features of the secondary structure near the active site (Chen, 2021; Salter et al., 2016).

1.1.2 Cytidine deaminase active site structure

As mentioned earlier tCDA has four active sites and each active site is made through the interaction of three monomers, for example, the active site of monomer A in figure 3A consists of the monomers A, B and C. Even though the active site is made of multiple monomers, the amino acids essential for catalytic activity are all found in only one of the monomers making up the active site. In tCDA, these essential amino acids are three cysteines and a glutamate residue, which are found in two conserved regions **CAERXA** (X usually S or T) and **PCG[A/I]CRQV[L/M]XE** (X any amino acid) (Zilpa A. Sánchez-Quitian et al., 2011). The cysteines tetrahedrally coordinate a zinc ion, which binds a water molecule. The glutamate is required for shuttling protons during the deamination reaction (Figure 1.6 A). During the reaction, the water molecule bound to the zinc ion is deprotonated, thus creating a nucleophilic hydroxyl ion. The pK_a value of the bound water molecule is affected by both the metal ion and the surrounding amino acids. Although in tCDA the zinc ion is coordinated by three negatively charged residues, tCDA enzymatic activity is like that of dCDA, where the zinc ion

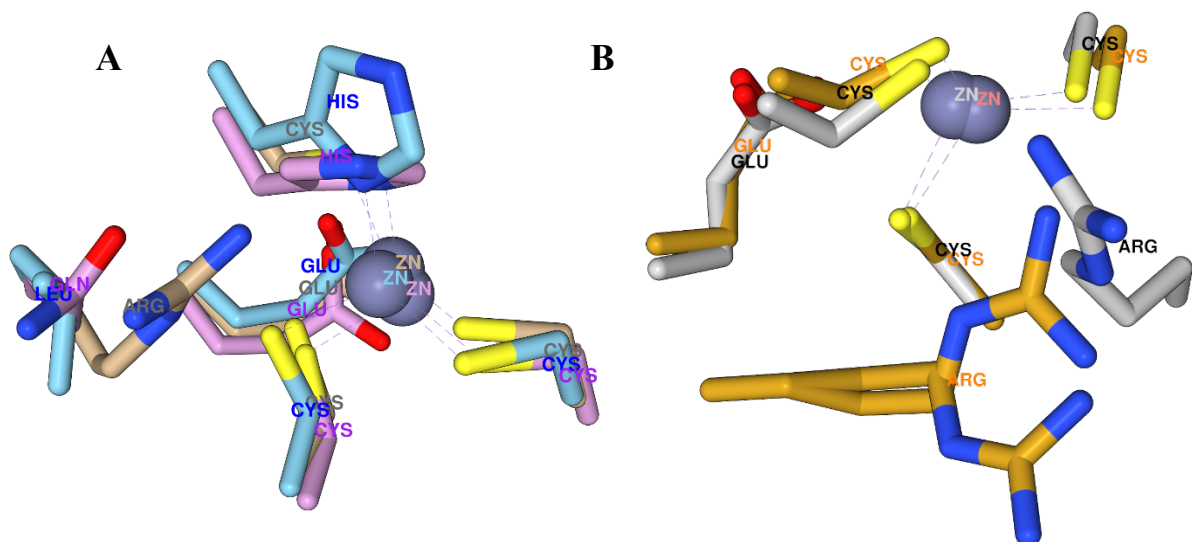


Figure 1.6 Cytidine deaminase active sites. **A** – Superimposed cytidine deaminase active sites: **Tan** – *Bacillus subtilis* tCDA active site (PDB: 1JTK), **Violet** – *Escherichia coli* dCDA active site (PDB: 1ALN), **Blue** – APOBEC3G active site (PDB: 4ROW, note Leu instead of Arg), **B** – Different arginine positions in the active site: **Grey** - *Mycobacterium tuberculosis* tCDA active site, the arginine is found on another loop, **Gold** – mouse tCDA active site, arginine had two observed conformations

is coordinated by a positive histidine and two negative cysteines. A possible explanation for this could be a conserved arginine found in tCDA, it forms hydrogen bonds with two of the three cysteines, this, together with the cysteines being located on the positive dipole ends of alpha spirals, acts to lower the negative charge exerted on the zinc ion and thus lowers the pK_a value of bound water, making it easier to break the bond between the proton and the hydroxyl ion (Johansson et al., 2002).

Zinc ions are very common in metalloenzymes that catalyse hydrolysis or hydration reactions, they are found in carbonic anhydrases (Mickevičiūtė et al., 2018), carboxypeptidases (Greenblatt et al., 1998), phosphatases (Sunden et al., 2017) and more, but other divalent metal ions are also found in enzymes interacting with substrates having a cytosine ring. For example, in cytosine deaminases (EC: 3.5.4.1) that deaminate free cytosine. The *Escherichia coli* cytosine deaminase (PDB: 1K6W, 3O7U) is a hexameric protein with (αβ)₈ barrel structure. Instead of only binding zinc ions, this cytosine deaminase can also bind iron ions, which are coordinated by four histidine and a single aspartate residue (Hall et al., 2011; Ireton et al., 2002). The bound iron ion can also be exchanged for other divalent metal ions like manganese, cobalt or zinc using *o*-phenanthroline, although these exchanges lower the effectiveness of the enzyme (k_{cat} s⁻¹ values decreased from 185 to 88, 50 and 32 respectively) (Porter & Austin, 1993). Similar experiments with exchanging metal ions were also performed with *Bacillus subtilis* tCDA. The zinc ions were unbound with *p*-hydroxymercuriphenyl sulfonate (PMPS) titration followed by the addition of EDTA and DTT. The apoenzyme could be reconstituted with the addition of cadmium or cobalt ions, but only 20% of native activity was maintained, and the cobalt substituted enzyme lost activity after 24 h. Other metals like iron, copper, nickel, or magnesium did not restore enzymatic activity. It was also shown that the zinc ion was not essential to maintaining the tertiary enzyme structure – even though the tetrameric structure disassembled after titration with PMPS the tetramer reassembled with the addition of DTT without added zinc ions (Mejlhede & Neuhard, 2000).

The negative charge reducing Arg56 in *Bacillus subtilis* tCDA was shown to be required for substrate deamination by mutating it into alanine, glutamine, and glutamate. Both R56A and R56Q mutants had decreased V_{max} values without significant reduction in K_m values, showing that the positive dipole of alpha spirals decreases the negative charge of the two cysteine residues enough to maintain catalytic activity, even though it was diminished, wild-type tCDA V_{max} with cytidine was 184 ± 18 μmol/min, R56A – 7.5 ± 0.5 μmol/min, R56Q – 29 ± 4 μmol/min. Changing the positively charged arginine into a negatively charged glutamate residue on the other hand introduced too much negative charge, the R56D mutant had a five-time reduction in bound zinc ions, meaning that the binding site of zinc was destabilized. The R56D mutant also showed no observable enzymatic activity (Johansson et al., 2004). In human tCDA, the conserved Arg68 was also shown to be important for the catalysis of cytidine. Two mutant variants were made R68G and R68Q. The R68Q mutation

increased both K_m and V_{max} values, but the V_{max}/K_m ratio was 1.2 compared to wild-types 1.7, meaning that this mutation didn't have a drastic effect on overall enzyme activity. The R68G mutation on the other hand showed only marginal activity levels with a V_{max}/K_m ratio of 0.06 (Vincenzetti et al., 2008). In the mouse tCDA, it was shown that the Arg68 could also serve a role in removing uridine from the active site after the deamination of cytidine. Unlike in crystal structures of tCDA from *Bacillus subtilis* or *Saccharomyces cerevisiae*, the Arg68 in mouse tCDA was observed in two conformations (Figure 6B). The first gauche⁺ conformation maintained the hydrogen bonds with the Cys65 and Cys102 residues, but the second gauche⁻ conformation interacted with the backbone of Leu62 from an adjacent monomer breaking the hydrogen bonds with the cysteine residues. This second conformation seems to be counterproductive for cytidine deamination by decreasing the net positive charge of the zinc ion, but the more negative zinc at the same time might facilitate easier clearing of uridine products by weakening the bond with O4 of uridine. The most likely reason why the Arg56 is found in two conformations only in the mouse tCDA is because the mouse tCDA has enough space around the Arg56 for it to rotate, whereas in other tCDA structures the arginine residue is restricted by the surrounding amino acids (Teh et al., 2006). Interestingly in *Mycobacterium tuberculosis* tCDA the arginine responsible for lowering the negative charge is moved to another alpha spiral compared with other tCDA enzymes and cysteine is found instead in its usual place (Figure 1.6 B) (Timmers et al., 2012).

The cysteine residues that coordinate the zinc ion were shown to be important for catalysis by mutating one of the cysteines into a histidine to mimic the active site of dCDA. The single C53H mutant of *Bacillus subtilis* tCDA formed inclusion bodies when expressed and did not show any enzymatic activity. Another double C53H/R56Q mutation was also performed, this mutant fully mimics the active site of dCDA where the zinc ion is bound by two cysteines and one histidine, and the arginine is replaced with glutamine. The double mutant did show some residual activity, but it was 500 times less active than the wild-type tCDA enzyme (Johansson et al., 2004).

1.1.3 Cytidine deaminase reaction mechanism

During the deamination reaction, the substrate C or dC and a water molecule are converted into U or dU and ammonia. Cytidine deaminases do not deaminate cytosine, likely because the interactions with the cytidine ribose moiety are needed for the substrate stabilization (Costanzi et al., 2011). But it is known that the *Escherichia coli* dCDA can deaminate N4-methylcytosine to uridine (Cohen & Wolfenden, 1971) and that tCDA can deaminate pyrimidines with fused five-member heterocycle rings at the C5 and C6 positions (Ludford et al., 2021). The reaction mechanism is most likely conserved across all cytidine deaminase superfamily enzymes (Salter et al., 2016).

The reaction starts with the cleavage of the water O-H bond, the remaining hydroxyl ion stays bound to the zinc ion and the proton binds to the glutamate -COO⁻ group (Figure 1.7 steps E to A). In the next step, the hydroxyl ion makes a nucleophilic attack on the C4 atom of cytidine, and the proton bound to glutamate is transferred to the N3 atom of cytidine breaking the double bond between N3 and C4 (Figure 1.7 steps A to B). This state in the B step is referred to as the tetrahedral intermediate. The proton from the -OH group then migrates to glutamate (Figure 1.7 steps B to C) and finally from the glutamate it is transferred onto the -NH₂ of cytidine. This forms ammonia, which is released and at the same time, the lone electron from oxygen makes a double bond with the C4 atom of cytidine (Figure 1.7 steps C to D). The rate-determining step in the reactions is considered to be the proton

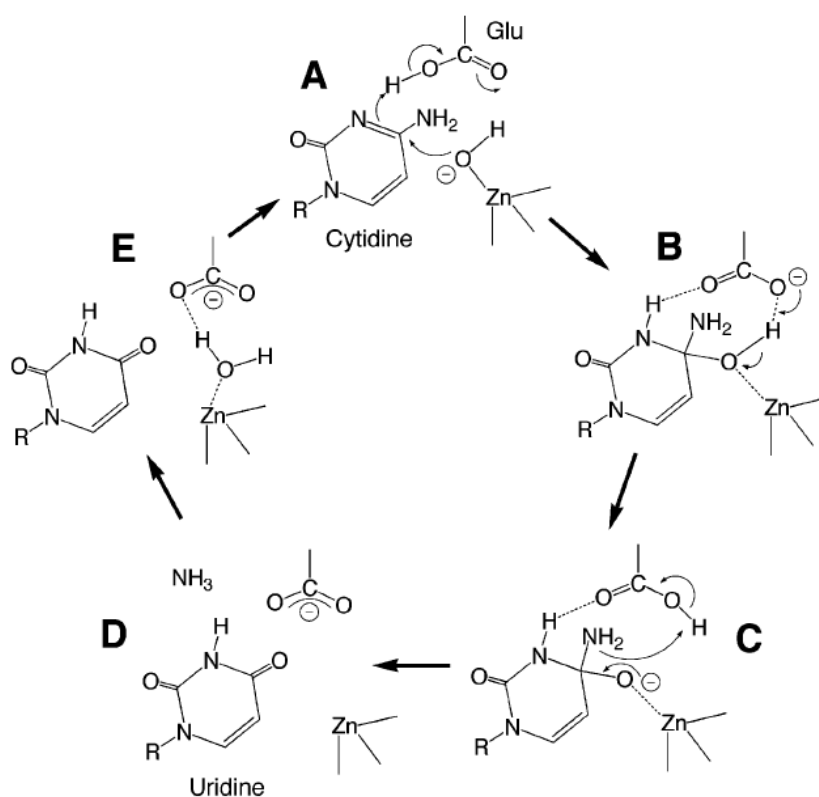


Figure 1.7 Cytidine deaminase reaction catalytic cycle (Matsubara et al., 2006).

transfer between the tetrahedral intermediates -OH group to its -NH₂ group (Matsubara et al., 2006). Theoretically, it was calculated that an extra water molecule in the active site of tCDA could enhance its catalytic efficiency by creating a hydrogen bond network and lowering the energy barrier of the proton transfer (Matsubara et al., 2006), although in some crystal structure results it is noted that the substrates are bound tightly in the binding sites of tCDA and there is no access for extra solvent molecules (Chung et al., 2005; Johansson et al., 2002).

1.2 Cytidine deaminase interaction with substrates

tCDA binds substrates through two moieties– the sugar moiety and the pyrimidine moiety. It was proposed that the interaction with substrates in tCDA happens through three main, relatively conserved, regions. The first two regions are located at the tetramer interface and so also contribute to the quaternary structure formation. In the human tCDA, these three regions are $_{32}\text{PYSHF}_{36}$, $_{54}\text{NIENACYP}_{61}$ and $_{131}\text{ELLPSSF}_{137}$, in mouse tCDA – $_{32}\text{PYSRF}_{36}$, $_{54}\text{NIENACYP}_{61}$ and $_{131}\text{ELLPASF}_{137}$, in *Bacillus subtilis* tCDA – $_{20}\text{PYSKF}_{24}$, $_{42}\text{NIENAAYS}_{49}$ and $_{119}\text{ELLPGAF}_{125}$, in *Mycobacterium tuberculosis* tCDA – $_{23}\text{PYSRF}_{27}$, $_{45}\text{NIENVSYG}_{52}$ and $_{117}\text{DLLPDAF}_{123}$. In the first region only the fourth position changes between arginine, histidine, and lysine, but all these amino acids have similar chemical properties. The second region has more variation, notably, the last residue is changed into serine in *Bacillus subtilis* instead of glycine or proline which both imply a turn in the structure. The third region hosts a conserved phenylalanine and has some variability in the fifth and sixth positions. Molecular dynamics (MD) simulations on the *Mycobacterium tuberculosis* tCDA with and without substrate also highlighted flexibility of these regions, interestingly the results also show high mobility in the 85-90 amino acid positions, these amino acids correspond to a loop close to the active site (Figure 1.8) (Zilpa A. Sánchez-Quitian et al., 2011).

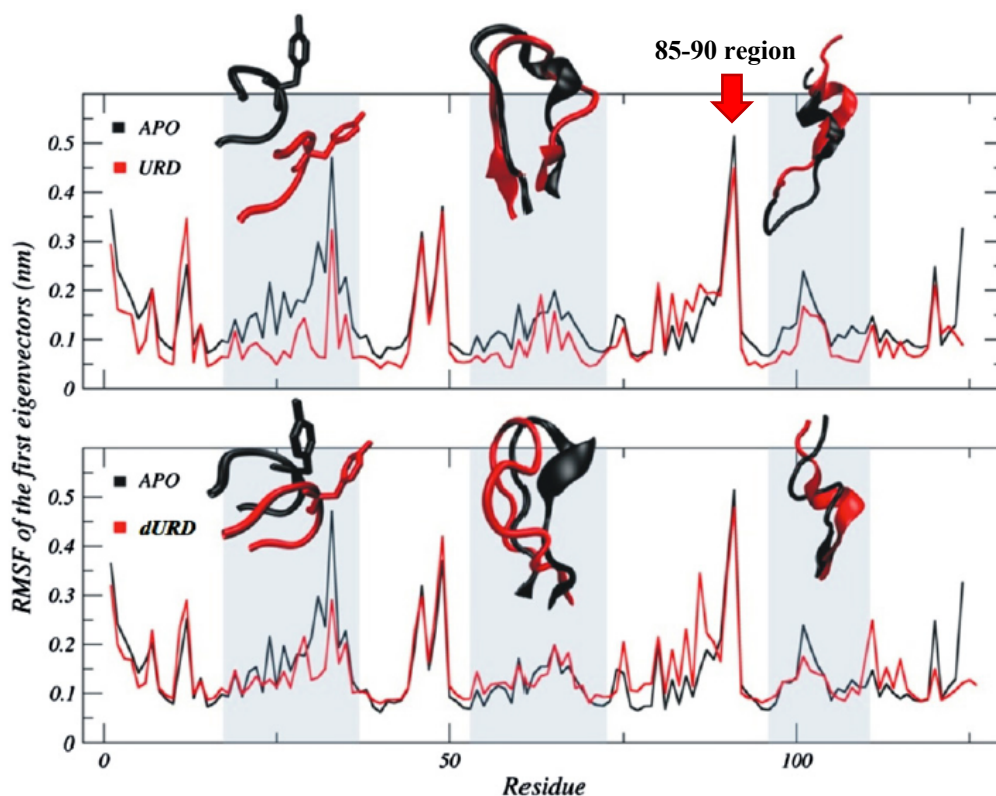


Figure 1.8 *Mycobacterium tuberculosis* tCDA structure fluctuation for the components of the first eigenvector. The displayed regions are Val22-Phe27, Thr42-Cys56, Val110-Phe123, the 85-90 region is highlighted with a red arrow, URD – enzyme with uridine, dURD – enzyme with 2'-deoxyuridine, APO – enzyme without substrate. (Zilpa A. Sánchez-Quitian et al., 2011)

1.2.1 CDA interaction with the sugar moiety

The interaction of CDA's with the sugar moiety of cytidine is performed through a hydrogen bond between the 3'-OH of ribose and conserved glutamate and asparagine residues (Asn42 and Glu44 in *Bacillus subtilis* tCDA) (Figure 1.9 A). The 3'-OH needs to be in β -conformation for the bond to form, ribose analogues with 3'-OH in the α -position do not bind tCDA (Costanzi et al., 2011). The 2'-OH group ribose does not seem to have interaction partners in tCDA, this is most likely the reason why both C and dC are equally good substrates for tCDA, but the size of possible substitutions in this position is restricted as it is found right next to one of the cysteine residues required for catalysis (Costanzi et al., 2011; Johansson et al., 2002). The importance of glutamate binding to 3'-OH was tested experimentally by creating five mutant variants, E47A, E47D, E47L, E47H and E47Q of *Mycobacterium tuberculosis* tCDA. The E47A and E47H variants were not soluble. The rest were soluble, and the mutations did not seem to affect zinc ion binding or tertiary and quaternary structure formation. The E47L variant did not have any detectable catalytic activity. The E47D and E47Q variants did show catalytic activity, their K_m values were not drastically affected but the k_{cat} values decreased by 37 and 19 times respectively. The decrease only in k_{cat} values seems to suggest a catalytic role for Glu47, likely by indirectly affecting the orientation of the pyrimidine moiety and impacting proton transfer needed for deamination (Zilpa Adriana Sánchez-Quitian et al., 2015). The importance of 3'-OH stabilization by hydrogen bond is also shown in pH profiles of *Mycobacterium tuberculosis* tCDA, a single ionizable group with a pK_a of 4.3 (± 1) could reduce the catalytic activity

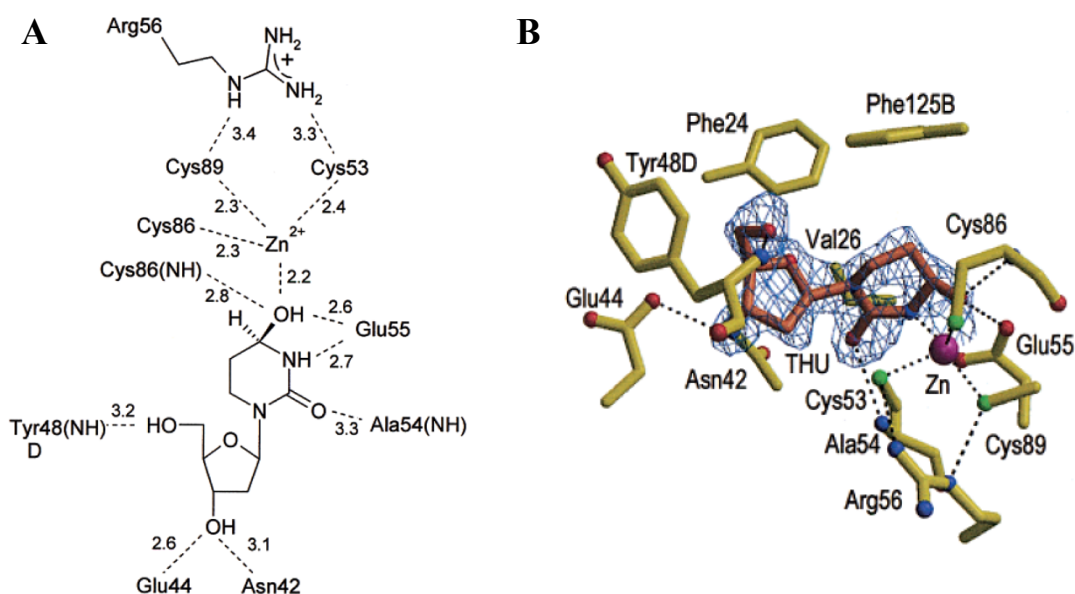


Figure 1.9 *Bacillus subtilis* tCDA interaction with the CDA inhibitor THU. A – diagram of hydrogen bonds with lengths between THU and tCDA B – stick representation of THU in the active site of tCDA (Johansson et al., 2002)

of the enzyme, it is speculated that this might be caused by glutamate protonation and the loss of the hydrogen bond to 3'-OH (Zilpa A. Sánchez-Quitian et al., 2010).

The 5'-OH group of ribose forms a bond with the conserved Tyr48 (*Bacillus subtilis* tCDA) (Figure 1.9 A) backbone -NH group from an adjacent subunit (Figure 1.3 A – substrate in subunits A active site with subunits D tyrosine) (Johansson et al., 2002). It is also noted that in CDA the ribose needs to be in C2'-endo orientation for efficient deamination, deamination rate for substrates with locked C3'-endo orientation were 100000 times lower. Interestingly adenosine deaminases prefer the ribose in C3'-endo orientation and have only a 100 lower deamination rate for C2'-endo locked substrates. This is a confirmation for the lack of evolutionary homology between cytidine and adenosine deaminases (Marquez et al., 2009).

1.2.2 CDA interactions with the base moiety

For interactions with the base part, the ability for the catalytic glutamate to form hydrogen bonds with atoms in the N3 position is important, because the N3 position is used for proton transfer during the deamination reaction and the N3 position protonation is essential for tetrahedral intermediate formation (Figure 1.7). When binding transition state analogues as inhibitors, for example, tetrahydro-2'-deoxyuridine (THU) the 4-OH group of THU can also form a hydrogen bond with the catalytic glutamate, this interaction also seems important, especially for inhibitors like THU binding, but not as critical as the hydrogen bond with the N3 position for substrate binding. One of the active sites cysteines backbone -NH group (Cys88 in *Bacillus subtilis*) can also form a hydrogen bond with the 4-OH of THU (Zilpa A. Sánchez-Quitian et al., 2011). The 2-O atom of cytidine was hydrogen-bonded to the backbone -NH group of alanine corresponding to the conserved Ala54 in *Bacillus subtilis* (Chung et al., 2005; Johansson et al., 2004; Teh et al., 2006).

Other non-bonded interactions with the base come from the conserved phenylalanine (Phe125 in *Bacillus subtilis* tCDA) residue at the C-end of tCDA (Figure 9B). This residue comes from an adjacent subunit (Figure 1.3 A – subunits A C-end interacts with subunits B active site) and stabilizes the substrate through T shaped π - π stacking. Substitutions in the C5 or C6 pyrimidine positions to nitrogen, as in 5/6-azacytidine, disrupt this π - π interaction by changing the partially positive charges of carbon to partially negative charges of nitrogen (Costanzi et al., 2011). Phe137 (Phe125 in *Bacillus subtilis* tCDA equivalent) mutation into alanine in human tCDA produces an inactive enzyme that also was not able to form tetramers and even the monomer structure seemed to be affected (Vincenzetti et al., 2008).

1.3 Protein engineering methods

Directed evolution and rational protein engineering are the two main methods for protein modification. They can be employed when there is a need to change the substrate specificities of an enzyme, adapt the enzyme for large scale industrial use, or other applications where the wild-type enzyme cannot be used. This adaptation can be achieved by changing its thermostability, adapting it to new pH ranges, making the catalysis more efficient in general, and changing the enzyme's enantio- and regioselectivity (Bornscheuer & Höhne, 2018).

Directed evolution aims to generate improved protein variants by accelerating natural evolution. Mutations as a result of natural evolution are rare and usually don't improve the enzyme in ways that would benefit us (Packer & Liu, 2015). The easiest way to speed up the mutation rate is to use mutagens such as ethyl methane sulfonate or ultraviolet light, alternatively turning off the natural genome repair mechanism also increases the mutation rate (Lai et al., 2004). An alternative is using systems like the MP6 plasmid that expresses proteins disturbing natural DNA error repair mechanisms (Badran & Liu, 2015). The drawbacks of this approach are that the mutations affect the whole organism and the chances of positive results in the gene of our interest are slim. Another simple but more efficient way of random mutagenesis is mutagenic PCR. Using high magnesium or manganese concentrations and mutagenic nucleotides the normal 10^{-9} DNA replication error rate can be increased to 10^{-3} (Packer & Liu, 2015). Random mutagenesis vastly accelerates new protein variant production and can be applied without any prior knowledge about the protein structure or mechanism of action. But an even more efficient approach to random mutagenesis is targeting specific regions which could be the most important for achieving the results we want. These regions can be selected by analysing sequence alignments with similar proteins, knowing the tertiary structure and the catalytic mechanism of the enzyme. Then the regions can be mutated using mutagenic PCR primers (Reetz et al., 2005). After creating these mutated proteins, the next step of directed evolution is selecting the improved mutant variants. This can be done by using a chromogenic substrate that changes colour after an enzymatic reaction or by using fluorescent reporter genes and monitoring the fluorescence (Packer & Liu, 2015). Other higher throughput methods include using auxotrophic host strains to screen $10^6 - 10^9$ mutant variants in a single petri dish, employing fluorescently activated cell sorting (FACS), microfluidic techniques and phage display (Agresti et al., 2010; Bornscheuer et al., 2019; Urbelienė et al., 2020; Vallejo et al., 2020).

Rational protein engineered, differently than random mutagenesis, uses the information about the enzymes structure, catalysis mechanism and interaction with substrates to make specific changes in the protein. The first example of rational protein engineering is a truncated bovine ribonuclease that maintained its catalytic activity (Gutte, 1975). Usually for single or few amino acid substitutions, deletions or insertions, site-directed mutagenesis is used. This basic method uses PCR with altered

primers that introduce the desired alterations into the protein sequence. For mutations in live systems technologies like CRISPR-Cas9 or CDA enzymes fused to transcription activator-like effectors (TALE) can be used (Biot-Pelletier & Martin, 2016; Mok et al., 2020). Alternatively, if many adjustments are needed the process can be accelerated by using *de novo* DNA synthesis which is becoming more robust and affordable (Kosuri & Church, 2014). But the first step of rational protein engineering is deciding what changes to make. The simplest method for deciding this would be manually analysing the protein sequence and tertiary structure if it is available. When the protein of interest structure is not experimentally solved molecular modelling tools like Modeller or AlphaFold2 can be employed (Jumper et al., 2021; Webb & Sali, 2016). To figure out possible interactions with substrates molecular docking and MD simulations are useful. MD simulations can help find mobile regions of the protein and visualize transition between structure configurations, together with molecular docking this helps to see which amino acids interact with the protein and if the proposed complex or protein is even stable (Childers & Daggett, 2017).

In some cases, an underlying enzyme to modify doesn't have to exist. The most important factor in biochemical reactions is the geometry of the substrate and catalytic molecules, this geometry can be determined using quantum mechanical calculations and then superimposed onto existing protein scaffolds using software like SABER or RosettaMatch (Leman et al., 2020; Nosrati & Houk, 2012). If matching scaffolds are found the next step is optimizing the active site around the amino acids and substrates for the desired enzyme, this can be done with software like RosettaDesign (Leman et al., 2020). The problem of determining the amino acid sequence that could form an arbitrary tertiary structure was proven to be easier to solve than the classical protein folding problem of determining the tertiary protein sequence from an amino acid sequence as shown by 93 residues α/β protein Top7 (Kuhlman et al., 2003). But limitations exist for these *de novo* methods of protein engineering for example *in vitro/vivo* water molecules in the active site could hinder the reaction. Because of these drawbacks *de novo* designs are best followed with additional rational design or directed evolution rounds to produce good results (Vaissier Welborn & Head-Gordon, 2019). Machine learning methods, specifically deep learning methods, are also being used in both protein structure prediction and design. Generative adversarial networks are being used to learn protein representations from actual enzymes and then generate millions of variants than can be further screened for desired activities (Strokach & Kim, 2022).

1.3.1 Examples of engineered cytidine deaminase enzymes

Most examples of engineered CDA enzymes come from the APOBEC family. Because they naturally act on single-stranded DNA (ssDNA) and RNA substrates APOBEC family proteins are

being used for C to T conversion as a gene-editing tool. CDA's have an advantage over CRISPR-Cas9 methods because they don't need to induce double-stranded breaks (DSBs) to make corrections. DSBs are usually repaired by non-homologous end-joining which leads to insertions, deletions, duplications, and other sequence alterations. Adding donor DNA can help stimulate homology-directed repair (HDR), but even then, the process results in a mixture of intended modifications and various indels from end-joining processes, also microorganisms like *Streptomyces* have low HDR capability and mechanisms that rely on HDR are inefficient in them (Zhao et al., 2019). Furthermore, methods for effective single-nucleotide editing are desired because most genetic disorders are caused by single nucleotide polymorphisms (Huang et al., 2021). First efforts to use CDA's for base editing were done by fusing a rat APOBEC1 to a Cas9 nickase, for targeting, and an Uracil DNA glycosylase inhibitor (UGI) to inhibit base excision repair (BER). This construct was able to induce C to T conversions in 37% of treated human cells with an indel rate of 1.1% (Komor et al., 2016). Later improvement to this construct by exchanging *Streptococcus pyogenes* Cas9 nickase with *Staphylococcus aureus* Cas9 nickase and using a double mutant W90Y, R126E of rat APOBEC1 achieved ~50-75% conversion rate and narrowed the editing window from ~5 nucleotides to 1-2 nucleotides (Kim et al., 2017). Another group also tried fusing human AID to zinc fingers (ZF) and TALE but only achieved an editing efficiency of 13% in *Escherichia coli* and 2.5% in human cells. The higher efficiency of Cas9 fused variants was probably caused by the ssDNA loop which forms from Cas9 binding, ssDNA being natural substrates for AID and APOBEC, and by nicking the non-edited strand to bias the eukaryotic mismatch repair to use the edited DNA strand as a template for repair (Yang et al., 2016). Another AID fused to a deactivated Cas9 system was created for *Escherichia coli* genome editing and achieved up to 95.1% editing rate for some targets (Banno et al., 2018).

Other examples of engineered CDA's are an *Escherichia coli* dCDA optimized for synthesis of COVID-19 drug Molnupiravir intermediate – *N*-Hydroxy-cytidine (Figure 1.10). This was achieved through multiple rounds of saturation mutagenesis targeted at regions surrounding the active site of the dCDA. The engineered dCDA prefers to bind NH₂OH in the active site over H₂O and achieved an 85% yield in 3 h only using 0.001 mol% of the purified enzyme (Burke et al., 2022).

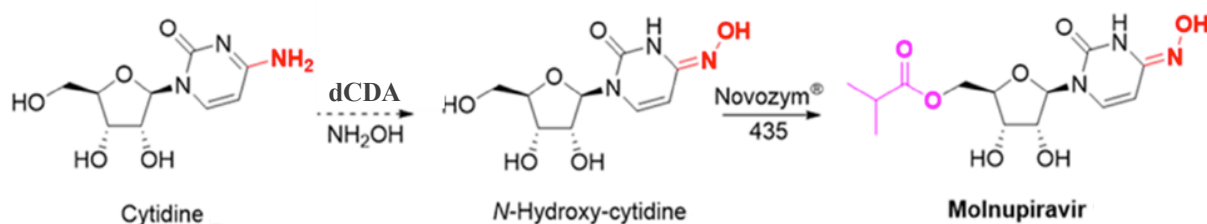


Figure 1.10 Synthesis of Molnupiravir with engineered dCDA and Novozym 435. (Burke et al., 2022)

Further examples are a C to T bacterial mutagenesis system acting on defined DNA regions engineered by fusing a CDA to a T7 RNA polymerase (Moore et al., 2018). A CDA inhibitor zebularine resistant human tCDA engineered through random mutagenesis to help prevent myelosuppression in patients when treated with combination therapy of zebularine and cytosine arabinoside after hematopoietic stem cell transplantation (Ruan et al., 2016), a novel cytidine deaminase from *Burkholderia cenocepacia* that can deaminate double-stranded DNA fused to TALE for mitochondrial DNA (mtDNA) editing. Fusion to TALE instead of Cas9 is beneficial for mtDNA editing because it doesn't require guide RNA transportation into the mitochondria, also mitochondria don't have efficient DSB repair mechanisms so mtDNA cut by Cas9 systems is rapidly degraded, meaning previously editing could only be achieved through heteroplasmy shifting (Mok et al., 2020).

2 MATERIALS AND METHODS

2.1 MATERIALS

Table 2.1 Used substrates

Capecitabine Cytosine β -D-arabinofuranoside	Sigma-Aldrich (Germany)
<i>N</i> ⁴ -acetyl-2'-deoxy-5'-O-DMT-cytidine <i>N</i> ⁴ -benzoylcytidine <i>N</i> ⁴ -benzoyl-5-methylcytidine 2'-deoxy-5-hydroxycytidine 2'-deoxy-5-hydroxymethylcytidine 5-hydroxymethylcytidine 2'-deoxy-5-propynylcytidine Pseudoisocytidine Isocytidine 5-fluorocytidine 2-thiocytidine 2'-deoxy-5-methylcytidine 2',5'-dideoxycytidine 2',3'-dideoxycytidine 2'- <i>O</i> -methylcytidine 3'-azido- <i>N</i> ⁴ -benzoyl-2',3'-dideoxycytidine	Carbosynth (UK)
3'-levulinyl- <i>N</i> ⁴ -benzoyl-2'-deoxycytidine 5'-levulinyl- <i>N</i> ⁴ -benzoyl-2'-deoxycytidine 3'-acetyl- <i>N</i> ⁴ -benzoyl-2'-deoxycytidine 2'-deoxycytidine	Jena Bioscience (Germany)
<i>N</i> ⁴ -acetylcytidine <i>N</i> ⁴ -acetyl-2'-deoxycytidine <i>N</i> ⁴ -benzoyl-2'-deoxycytidine <i>N</i> ⁴ -isobutyl-2'-deoxycytidine	Combi-Blocks (USA)
<i>N</i> ⁴ -hexanoyl-2'-deoxycytidine <i>N</i> ⁴ -nicotinoyl-2'-deoxycytidine <i>N</i> ⁴ -(2-acetyl-benzoyl)-2'-deoxycytidine <i>N</i> ⁴ -(3-acetyl-benzoyl)-2'-deoxycytidine <i>N</i> ⁴ -(4-acetyl-benzoyl)-2'-deoxycytidine <i>N</i> ⁴ -(2-benzoyl-benzoyl)-2'-deoxycytidine <i>N</i> ⁴ -(3-benzoyl-benzoyl)-2'-deoxycytidine <i>N</i> ⁴ -(4-benzoyl-benzoyl)-2'-deoxycytidine 4-thio-methyl-U 4-thio-ethyl-U 4-thio-benzyl-U 5-F-4-thio-methyl-U 5-F-4-thio-ethyl-U 5-F-4-thio-benzyl-U 5-F-4-thio-phenyl-U 5-F-4-methoxy-U 5-F-4-butoxy-U 5-F-4-benzyloxy-U <i>N</i> ⁴ -methylcytidine <i>N</i> ⁴ - <i>N</i> ⁴ -dimethylcytidine <i>N</i> ⁴ -2-hydroxyethyl-2'-dC	Synthesized in the Molecular Microbiology and Biotechnology Department (MMBD)

Table 2.1 Used substrates

<i>N</i> ⁴ -aminoethyl-2'-dC 4-(4-morpholinyl)-2'-dU 5-F-4-(4-morpholinyl)-2;-dU <i>N</i> ⁴ -hexyl-2'-dC <i>N</i> ⁴ -(indolin-1-yl)-2'-dC <i>N</i> ⁴ -(2,3,4,5,6-pentahydroxyhexyl)-2'-dC <i>N</i> ⁴ -(1 <i>H</i> -indol-6-yl)methyl)-2'-dC 4-(4-morpholinyl)-5-Fluoro-2'-deoxyuridine 4-(4-morpholinyl)-2'-deoxyuridine	
---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	--

Table 2.2 Used mutagenic primers

Clone	Forward primer (5'->3')	Reverse primer (5'->3')
F14_T51G	TTACGGAGCGGGTAATTGCGG	GAAGCATTTCGATATTTGCG CCTAAAAAC
F14_G81I	TGATAGGGTCATCGCACCTTGCG	CCATCGGTTACGATCGCCAAC G
F14_G85L	TGATAGGGTCATCGCACCTTGCG	CCATCGGTTACGATCGCCAAC G
F14_R56L	ATTGCGGTGAATTAAGTGCCATTTTC GC	TGGTCGCTCCGTAAGAAGCAT TTTCG
F14_F126A	GATGCGAGCAACGAAAGAGGATCTT TTAG	GGCAACAATTCATCGATCGTT TTTTCTAACG
F14_F126W	GATGCGATGGACGAAAGAGGATCTT TTAG	GGCAACAATTCATCGATCGTT TTTTCTAACG
F14_C88H_ C91H	GCACCTCACGGAATCCACCGTCAAG	ACCGACCCTATCACCATCGGT TAC
F14_C53H_ R56Q	GCGACCAATCATGGTGAACAAAGTG CC	TCCGTAAGAAGCATTTCGAT ATTTGCGCC
F14_del83- 85	GCACCTTGCGGAATCTGCC	ATCACCATCGGTTACGATCGC C
F14_mutdel 83-85	NNNNNNNNNGCACCTTGCGGAATCT GCC	ATCACCATCGGTTACGATCGC C
F14_del127- 130	CTTTTAGGCCATCACCATCACCAC	GAATCGCATCGGCAACAATTC ATC
F14_mutdel 127-130	NNNNNNNNNNNNNCTTTTAGGCCATC ACCATCACCAC	GAATCGCATCGGCAACAATTC ATC
Pco_G70T	TCCCGTCTACCCTGTGCGCG	ACGCCGCGTTTTCTGGTTCG
Pco_I108A	TGAAATCTCCGGCGTCTCCGTGC	GGAAACCGTTACCGTTACGCG CC
Ppo_C50T	CGGTCTGACCAACTGCGCGG	TAAGACGCGTTTTCAACGTTG CAACCG
Ppo_V82L	GAAGGTCCGCTGTCTCCGTGC	GGTGTCCGCCGCAACCG
Tar_I85A	AAAACCGGCGTTCCCGTGCG	CCGGCGCGATAGAAGAAGAG ATCGC
Lsp_A82I	TAACGCGGACATCGCGCCGTG	CCGTTGCAAACAACCGCCAGC G

Used *Escherichia coli* strains:

DH5α: F⁻, ϕ 80*dlacZ* Δ *M15*, *recA1*, *endA1*, *gyrAB*, *thi-1*, *hsdR17*(rK⁻, mK⁺), *supE44*, *relA1*, *deoR*, Δ (*lacZYA-argF*) *U169*, *phoA*.

HMS174 *ΔpyrFΔcdd*: F⁻ *recA1*, *hsdR* (rK12⁻ mK12⁺) (DE3) (Rif R), *ΔpyrFΔcdd*:Km, (constructed in MMBD).

Growth medium used:

LB medium: 0.5% peptone, 0.5% NaCl, 0.3% yeast extract.

M9 medium: 33.9 g/L Na₂HPO₄, 15 g/L KH₂PO₄, 5 g/L NH₄Cl, 2.5 g/L NaCl, 0.2% (w/v) glucose, 0.2% casamino acids, 1 mM IPTG, 15 g/L agar, 0.02 mg/ml uridine or its analogues.

SOB medium: 0.5% yeast extract, 2% tryptone, 10 mM NaCl, 2.5 mM KCl, 20 mM MgSO₄.

SOC medium: SOB medium + 0.2% glucose.

The components were mixed in distilled water for all mediums, and the pH value was adjusted to 7.0. Mediums were autoclaved for 30 min at 121 °C, and 1 atm pressure.

Plasmids used:

pLATE31 – Thermo Fisher Scientific

2.2 METHODS

1.2.1 Site-directed mutagenesis primer creation

Various mutations were performed to investigate the importance and function of single amino acids and regions of the CDA proteins. For creating specific amino acid substitutions site-directed mutagenesis was performed using simple PCR principles. One of the primers used for mutagenic PCR had a few non-complementary nucleotides in the middle to change the amino acid encoded. The non-complementary nucleotides in the primer were surrounded by 10-15 fully complementary nucleotides making the total primer length around 30 bases. Attention was paid to make sure that the 3' end of the primers ended in a C or G nucleotide to create a 3' clamp. The annealing temperature of the primers was between 65 and 72 °C. Annealing temperatures were calculated using the T_m calculator (Thermo Scientific web tools).

The active sites C53H, C88H, C91H and R56Q mutants were created in three rounds. First, the C53H/R56Q and C88H/C91H double mutants were made, each using a single primer with two mutation sites, then the C53H/R56Q/C88H/C91H mutant was made by again mutating the C53H/R56Q mutant plasmid with the C88H/C91H mutant primers. All other considerations mentioned for the single mutation primers are also applicable to double mutation primers.

For the cases when deletion of a region was needed the primers were made to surround the region to be deleted, after amplification of the whole plasmid the region is left out of the product. The filling of the deleted region with random amino acids was done by adding a random nucleotide overhang to the 5' end of the forward primer used for deletion of the region, it was acknowledged that this method could produce premature stop codons, but because only four amino acids were

changed at the maximum the ~20% probability of a stop codon was deemed acceptable. All other considerations mentioned for the single mutation primers are also applicable to these primers.

1.2.2 Site-directed mutagenesis primer phosphorylation

Primers used for site-directed mutagenesis were phosphorylated before mutagenesis PCR reactions. The reaction mixture is given in the Table 2.3 The reaction was incubated at 37 °C for 30 min. After the phosphorylation reaction, the T4 polynucleotide kinase was inactivated by incubation at 75 °C for 10 min.

Table 2.3 DNA primer phosphorylation reaction mixture.

T4 polynucleotide kinase	1 µL
T4 DNA ligase buffer 10x	2.5 µL
ATP (10 mM)	2.5 µL
Primer	1.25 µL (100 µM)
H ₂ O	17.75 µL

1.2.3 Site-directed mutagenesis PCR reaction

PCR reactions were done using either “Phusion Green Hot Start II High-Fidelity PCR Master Mix” or “Phusion™ Plus PCR Master Mix” (Thermo Fisher Scientific). The reaction was performed in a 20 µl volume, its contents are given in the table 2.4, and the reaction conditions are given in the Table 2.5

Table 2.4 PCR reaction mixture.

“Phusion Green/Plus PCR Master Mix”	10 µL
Primers	0.5 µM each
DNA	0.5 µL (100 ng)
DMSO	0.3 µL
H ₂ O	up to 20 µL

Table 2.5 PCR reaction conditions.

Step	Temperature	Time	Number of cycles
Initial denaturation	98 °C	30 s	x1
Denaturation	98 °C	10 s	x25
Primer annealing	65 - 72 °C	20 s	
Primer extension	72 °C	25 s/kb	
Final extension	72 °C	5 min	x1

1.2.4 DNA electrophoresis and DNA fragment purification for agarose gels

For DNA electrophoresis 1% agarose TAE buffer gels were used. Gels were run at 120 V for 10 min, then suspended in 0.05% ethidium bromide solution for 5 min and ran again at 120 V for another 2-5 min. For fragment size determination “GeneRuler DNA Ladder Mix” (Thermo Fisher Scientific) was used.

DNA fragments were purified from agarose gels using “GeneJET Gel Extraction Kit”

(Thermo Fisher Scientific) according to the manufacturer's recommendations, with minor modifications – the cut fragments were always suspended in 600 μL of Binding Buffer.

1.2.5 Linear DNA fragments ligation

Linear fragments got from site-directed mutagenesis reactions were ligated over 18 h at 4 °C. The reaction mixture is given in the Table 2.6 After ligation, the reaction mixture was desalted using the “GeneJET PCR Purification Kit” (Thermo Fisher Scientific) as this step was found to increase the transformation efficiency using electroporation. Prepared DNA was either immediately used for transformation or stored at -20 °C for later use.

Table 2.6 DNA ligation reaction mixture

T4 DNA ligase buffer 10x	1 μL
T4 DNA ligase (5 U/ μL)	0.5 μL
PEG 4000	1 μL
DNA	5 μL
H ₂ O	2.5 μL

1.2.6 Competent cell preparation

Electro-competent cells were prepared by growing the selected cell strain in 25 mL LB medium at 37 °C with aeration until the cell density at 600 nm reached 0.6 – 0.7. Then the cells were centrifuged at 4000g 4 °C for 10 min. The supernatant was discarded, and the cells were resuspended in 10 mL of ice-cold 10% glycerol solution and centrifuged again at 4000g 4 °C for 10 min, after centrifugation the supernatant was again discarded. The above step was repeated three more times. After the last centrifugation, the cell biomass was resuspended in 2 mL of ice-cold 10% glycerol solution and 95 μL were dispensed into 1.5 mL reaction tubes. Cells prepared this way were either used immediately or stored at -80 °C for later use.

Competent cells for chemical transformation were prepared again by growing the selected cell strain in 25 mL LB medium at 37 °C with aeration until the cell density at OD₆₀₀ reached 0.6 – 0.7. The cells were then pelleted by centrifugation but instead of 10% glycerol, they were washed in an ice-cold 0.2 M CaCl₂ solution four times using the same conditions as for electro-competent cells. Prepared cells were either used immediately or, if they were stored for later use at -80 °C, glycerol was added to a total concentration of 10%.

1.2.7 Electroporation

95 μL of electro-competent cells were mixed with 1 – 2.5 μL of plasmid DNA and incubated on ice for 5 min. After incubation, the mixture was pipetted into a frozen electroporation cuvette (Electroporation cuvette Eppendorf). Electroporation was performed at 1800 mV/cm with 3.9 – 5.4 ms impulse length (Electroporator 2510 Eppendorf). Immediately after electroporation the cells were

suspended in 900 μ L of SOC medium and incubated for 30 min at 37 °C. After incubation the cells were pelleted by centrifugation at 10000g for 10 s, most of the medium was discarded and the pellet was resuspended in the remaining 100 – 200 μ L volume. The cell suspension was spread onto prepared LB agar plates with all necessary additives and grown at 37 °C.

1.2.8 Chemical transformation

95 μ L of competent cells were mixed with 1 – 5 μ L of plasmid DNA and incubated on ice for 10 min. After incubation, the cells were heat-shocked at 42 °C in an air thermostat for 2 min. After heat shock, the cells were suspended in 900 μ L of SOC medium and incubated at 37 °C for 30 min. After incubation, the cells were pelleted and plated on LB agar plates with necessary additives same as electroporated cells.

1.2.9 Plasmid purification

Plasmids were purified and sequenced to make sure the mutagenesis was successful. Up to three colonies were selected after plasmid transformation and plating. These colonies were grown in 5 mL of LB medium at 37 °C with aeration until the cell density at OD₆₀₀ reached ~0.9. Plasmids were purified using the “ZymoPURE II Plasmid Midiprep Kit” (Zymo Research) according to the manufacturer’s specifications. If required sequencing was performed in Macrogen Europe (Netherlands) using the T7 promoter (5’-TAATACGACTCACTATAGGG-3’) sequence as a primer, the results were analysed using Benchling (Biology Software, <https://benchling.com>) to select good mutagenesis results.

1.2.10 Enzymatic activity testing using agar selective medium

For some mutations enzyme activity was first evaluated using a selective M9 medium. Plasmids harbouring mutations were transformed into HMS174 Δ *pyrF* Δ *cdd*:Km cells. This strain is a uridine auxotroph that cannot grow without added uracil or cytidine. The selective M9 medium doesn’t provide cells with either but modified cytidine analogues like BzdC are added to it. If the mutated protein can remove the modification the cell auxotrophy is complemented and cell colonies form on the medium if the modifications are not removed the auxotrophy is not complemented and cell colonies do not form.

Cell colonies with the mutated plasmid genes are first grown on LB agar medium, then several colonies are streaked onto the M9 agar medium plates, and the plates are incubated at 37 °C for 24 h. Three M9 plates are used for one sample: a plate without any additives (negative control), HMS174 Δ *pyrF* Δ *cdd* cells should not grow on this plate, a plate with added dC (positive control), HMS174 Δ *pyrF* Δ *cdd* cells should grow on this plate irrespective of the plasmid they have, and a plate with

BzdC, HMS174 *ΔpyrFΔcdd* cells grow on this plate only if the plasmid has a protein that can remove the benzoyl group from BzdC. Apart from cytidine analogues, all M9 plates are supplemented with the appropriate antibiotic and 0.1 mM IPTG to induce enzyme expression.

1.2.11 Protein expression and purification

Selected strain cells were grown until the optical density reached $OD_{600} \sim 1.0$, protein expression was induced by adding 0.1 mM IPTG and the cells were grown for 18 h at 30 °C with aeration. After incubation, the cells were pelleted by centrifugation at 4000 g 4 °C for 10 min and resuspended in buffer A (20 mM Tris-HCl buffer, 100 mM NaCl, pH 7.5). The cells were lysed with ultrasound for 5 min in total with 5 s breaks every 3s of sonification using 30% power (Branson Digital Sonifier SFX 250 (Emerson)). The lysate was centrifuged at 10000 g 4 °C for 5 min and the supernatant was used for protein purification.

All proteins were tagged with a C-end 6xHis-tag. Proteins were purified either using “HisPur™ Ni-NTA Purification Kit” (Thermo Fisher Scientific) spin columns according to the manufacturer’s instruction or by Äkta Purifier 100 system (GE Healthcare) using a 1 mL Ni²⁺ HiTrap chelating HP column (GE Healthcare). Purification with Äkta consisted of two phases: protein sample application and elution. Before protein application, the column was washed with 5 CV of buffer A, then the sample was applied over 15 CV and washed with another 5 CV of buffer A. The bound proteins were eluted with a linear gradient of buffer B (20 mM Tris-HCl buffer, 100 mM NaCl, 0.3 M imidazole, pH 7.5) over 10 CV. After purification, the protein was dialysed against buffer A at 4 °C for 24 h with a 200:1 ratio of buffer A to protein volume. If the protein was used for crystallization, it was additionally purified using size exclusion chromatography through Superdex™ 200 (Cytiva) using buffer A.

1.2.12 Proteins fractionation by SDS-PAGE

Protein electrophoresis was performed in a vertical stand. The protein sample was mixed with Laemmli Sample Buffer containing 5% 2-mercaptoethanol and 2% SDS, and heated in a boiling water bath for 10 min to denature. The SDS-polyacrylamide gel was composed of a 5% stacking and 14% separating layers. Electrophoresis had two phases: 20 min at 64V, 20 mA and 40 min at 200V, 22 mA. After electrophoresis, the gel was dyed with Coomassie Brilliant Blue G-250 and bleached by boiling in water with 10% acetic acid.

1.2.13 Bradford assay for protein concentration determination

4 μL of the protein sample were mixed with 200 μL of Pierce™ Coomassie Plus (Bradford) Assay Reagent (Thermo Fisher Scientific) and after mixing incubated at room temperature for 5 min.

Optical density was measured at 595 nm. The protein concentration was determined using a calibration curve calculated using predetermined concentration BSA samples optical density (Bradford, 1976).

1.2.14 Enzyme activity assessment spectrophotometrically

Enzyme activity was determined in 96 well UV plates. 5 μ L of the enzyme (5-30 μ g) were mixed in 100 μ L of 50 mM potassium phosphate buffer with 100 mM NaCl pH 7.5 and the substrate was added to a final concentration of 0.4 mM. The control reaction contained 5 μ L of enzyme storage buffer instead of the enzyme. After incubation at 30 °C for 1 h and 18h, optical absorptions were determined in the 240 – 320 nm range every 2 nm. Hydrolysis was determined by comparing the control and enzyme reaction optical absorption curves.

1.2.15 Thin-layer chromatography

2 μ L (2-12 μ g) of the enzyme were mixed within 20 μ L of 50 mM potassium phosphate buffer with 100 mM NaCl pH 7.5 and the substrate was added to a final concentration of 2 mM. The reaction was incubated at 30 °C for 1 h. After incubation 1 μ L was transferred onto the “TLC Silica gel 60” (Merck Millipore) plate. The mobile phase consisted of CHCl_3 and CH_3OH mixed in a 5:1 ratio. Results were analysed under 254 nm UV light.

1.2.16 Enzyme kinetic parameter determination

Selected enzyme kinetic parameters were determined for dC and BzdC by monitoring absorption values at 290 nm ($\Delta\epsilon$ BzdC = 11000 $\text{M}^{-1} \text{cm}^{-1}$) and 310 nm ($\Delta\epsilon$ dC = 1600 $\text{M}^{-1} \text{cm}^{-1}$) respectively. For dC concentrations of 0.05 mM, 0.10 mM, 0.20 mM, 0.25 mM, 0.40 mM, 0.50 mM, 0.80 mM and 1.00 mM were tested. For BzdC concentrations of 12.50 μ M, 18.75 μ M, 25 μ M, 37.50 μ M, 50 μ M, 75 μ M, 100 μ M and 150 μ M were tested. Substrates were mixed in 50 mM potassium phosphate buffer with 100 mM NaCl, pH 7.5. 6 μ L (6-35 μ g) of protein solution were mixed in a UV cuvette with 594 μ L of the substrate solution and the change in absorbance was monitored for 1 min. This was repeated three times for each substrate concentration. K_m and V_{max} and k_{cat} values were determined by fitting Michaelis–Menten enzyme kinetics model onto the generated data using lmfit software version 1.0.3 (<https://lmfit.github.io/lmfit-py/>) (Newville et al., 2014). For the CDA_F14 T51G Lineweaver–Burk transformation was used to reduce parameter errors.

1.2.17 Protein structure modelling

CDA_F14 homology modelling was carried out using either the Bioinformatics Toolkit available at the Max Planck Institute for Developmental Biology (Tübingen, Germany;

<https://toolkit.tuebingen.mpg.de>) (Gabler et al., 2020; Zimmermann et al., 2018), Robetta available at (<https://robetta.bakerlab.org/>) (Baek et al., 2021; Hiranuma et al., 2021) or AlphaFold2 API notebook available at (<https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb>) (Jumper et al., 2021). When modelling with the Bioinformatics Toolkit homologous templates were found using HHpred. Good structures (probability > 95% and identity > 40%, resolution < 2.5 Å) were selected for homology modelling using MODELLER. When modelling with Robetta or AlphaFold2 the default parameters were used. The best model overall was selected by comparing the quality of models produced by different methods using VoronMQA available at (<https://bioinformatics.lt/wtsam/voromqa>) (Dapkūnas et al., 2018; Olechnovič & Venclovas, 2017) and checking for model agreement with known structures.

Homology models done using CDA_F14 as a template were performed using the standalone MODELLER software version (10.2) (Webb & Sali, 2016).

1.2.18 Molecular dynamics

Molecular dynamics were performed using AMBER16 software (<https://ambermd.org/>). Protein structures were prepared using TLEAP, substrates were prepared using ANTECHAMBER. The protein structures were parameterized using the ff14sb forcefield and the substrates if used were parametrized using the GAFF forcefield.

The generalized Born solvation model was used for implicit water simulations to increase the simulation and conformation space sampling speeds. The system was neutralized by adding the required number of Na⁺ or Cl⁻ ions. The implicit water simulation had four steps. First, the system was minimized using sander, then heated to 300 K over 500 ps and equilibrated for another 500 ps using pmemd. The production simulations were performed for at least 50 ns also using pmemd. For heating, equilibration and production simulations the non-bonded cut-off was infinite, the temperature was maintained using Langevin dynamics with collision frequency 0,5 ps⁻¹, and the trajectory was integrated every 2 fs using the SHAKE algorithm for bond length control.

For explicit water simulations, TIP3P molecular water was used. The enzyme-substrate complex was solvated in a water box of 35 Å and the system charge was neutralized by adding the required number of Na⁺ or Cl⁻ ions. The simulation had five steps. First, the system was minimized with sander, then heated to 300 K over 1 ns, then the system pressure was equilibrated to 1 bar over 2 ns and the system was equilibrated for a further 2 ns. The production simulation was run for 100 ns. Simulations were performed in constant volume periodic boundary conditions with isotropic pressure scaling. For heating, equilibration, and production simulations the non-bonded cut-off was set to 12 Å, the temperature was maintained using Langevin dynamics with a collision frequency of

2 ps⁻¹, the pressure was maintained using the Berendsen barostat. The trajectory was also integrated every 2 fs with the SHAKE algorithm for bond length control. Analysis of all trajectories was performed using CPPTRAJ.

1.2.19 Molecular docking

Molecular docking was performed using AutodockVina (Trott & Olson, 2009). The tetrahedral intermediate structures were docked into enzyme poses obtained from MD simulations every 0,5 ns. The tetrahedral intermediate was chosen because it was the most straightforward way of accounting for the H₂O molecule that is involved in the deamination reaction. Protein structures were prepared for docking using USCF Chimera DockPrep software. Substrate structures were prepared using Avogadro software, minimized using GAFF forcefield and protonated to 7,5 pH. Molecular docking was performed into each cytidine deaminase active site separately. Binding boxes were centred on Zn²⁺ ions found in the active site and their dimensions were determined by the size of the substrate. Parameters used for docking: exhaustiveness = 100, num_modes = 15, energy_range = 20. Docked structures were sorted according to the distance between relevant residues and binding energy. Selected poses were used for molecular dynamics simulations.

Standalone docking of a substrate into an active site was performed using the same parameters as for the MD simulation pose screening docking.

1.2.20 Assessing enzyme binding pocket SASA relationship with substrate selectivity

Monomers of CDA_EH, CDA_Lsp, CDA_Ppo, CDA_Pin, CDA_Pco, CDA_Smo, CDA_Tar, CDA_Dfa, CDA_Hfi, CDA_Mtu and crystal structures of CDA_F14 and CDA_Bsu (PDB: 1JTK) were superimposed onto each other. The mentioned CDAs were modelled by using AlphFold2 and MODELLER programs and CDA_F14 crystal structure as a reference (only for MODELLER). Mouse tCDA monomer with bound cytidine (PDB: 2FR6) was also superimposed onto the structures, the cytidine was used to select atoms which belong to the binding pocket. The atoms, which were within 5 Å of the cytidine were considered to belong to the binding pocket. Per atom solvent accessible surface area (SASA) was calculated using the Shrake-Rupley algorithm (Shrake & Rupley, 1973) implemented in the Biopython package version 1.79 (<https://biopython.org/>) (Cock et al., 2009). Binding pocket SASA was determined by summing SASA of atoms that were considered to belong to the binding pocket. Substrate volume was calculated using RDKit version 2022.03.1 (<https://www.rdkit.org/>). The Pearson correlation between the binding pocket SASA and substrate volume was determined using NumPy version 1.22.3 (<https://numpy.org/doc/stable/index.html>).

3 RESULTS AND DISCUSSION

During development of metagenomic libraries for selection of amidohydrolases able to deaminate N^4 -benzoyl-2'-deoxycytidine (BzdC) to 2'-deoxycytidine (dC) several clones were found to have *tCDA* genes instead of true amidohydrolase genes. This was a surprising finding as previously no CDA enzymes were known to be able to deaminate these kinds of bulky substrates, only other known similar activity in CDA was *Escherichia coli* dCDA able to deaminate N^4 -methylcytidine (Cohen & Wolfenden, 1971). Further analysis revealed that some of the selected enzymes were also able to convert N^4 -acyl-/ N^4 -alkyl-, N^4 -carboxy-, S^4 -alkyl- and O^4 -alkoxy- cytidine derivatives into uridine (Figure 3.1). Of the studied CDA enzymes CDA_F14 and CDA_EH were selected from metagenomic libraries using uridine auxotrophic cells *Escherichia coli* DH10B Δ *pyrFEC::Km* and N^4 -benzoyl-2'-deoxycytidine on minimal M9 medium by dr. Nina Urbelienė (Urbelienė et al., 2019). Other CDA's included in the study were selected from the human microbiota (*Holdemania* – CDA_Hfi, *Lachnoclostridium* sp. – CDA_Lsp, *Solobacterium* - Smo, *Dielma* – CDA_Dfa, *Prevotella copri* – CDA_Pco), pathogenic organisms (*Prevotella intermedia* – CDA_Pin, *Mycobacterium tuberculosis* – CDA_Mtu), soil microorganism (*Paenibacillus polymyxa* – CDA_Ppo), and the archaea kingdom (*Thaumarchaeota* – CDA_Tar). *Escherichia coli* dCDA, *Bacillus subtilis* tCDA (CDA_Bsu) and commercially available human tCDA (Human_CDA) were also used for comparisons. The analysis in this study focuses mostly on CDA_F14 and its ability to deamidate N^4 substituted substrates like BzdC.

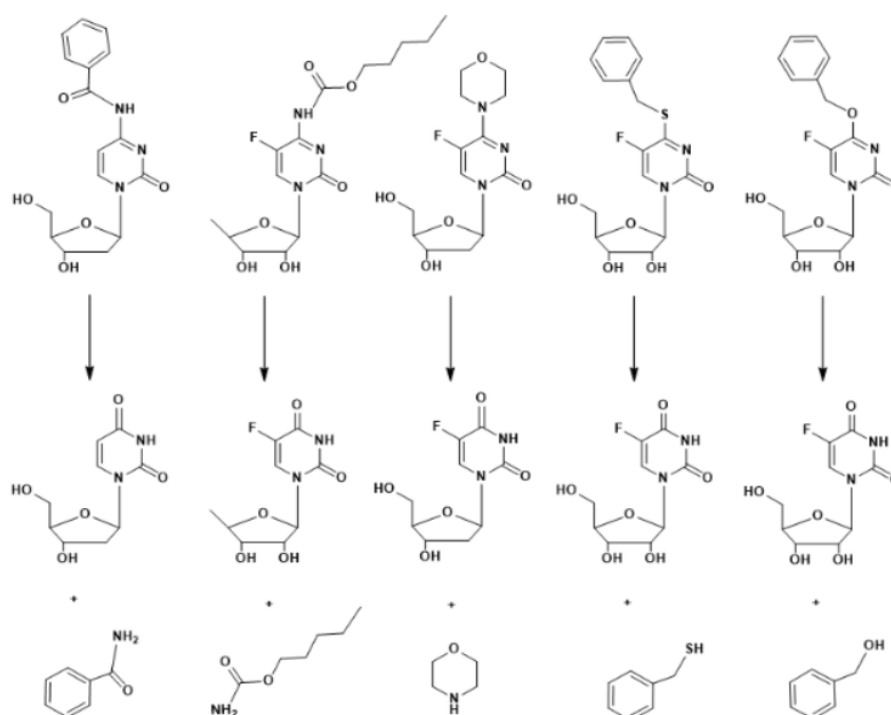


Figure 3.1 Substrates of the discovered tCDA enzymes. From left to right: BzdC, Capecitabine, 4-(4-morpholinyl)-5-Fluoro-2'-deoxyuridine, 4-Thio-benzyl-5-fluoro-uridine, 4-benzyloxy-5-fluoro-uridine.

3.1 Sequence analysis and relationship with enzyme specificity

Initial BLAST analysis showed that CDA_F14 has 85% identity with a cytidine deaminase from *Firmicutes bacterium*, Pfam analysis also showed that the enzyme belongs to cytidine deaminases, HHpred analysis further confirmed that CDA_F14 is highly similar to other tCDA enzymes. Next a sequence alignment with other selected tCDA enzymes was made using the MAFFT algorithm with standard parameters (Figure 3.2). All tCDA's had the two conserved regions, $_{72}C(A/G)ERXA_{77}$ (X – polar uncharged (Ser, Thr, Asn or hydrophobic Ala) and $_{111}PC(G/M)(A/D)CRQV(V/L/I/M)XE_{121}$ (X – any amino acid), where the amino acids required for deamination are found, the only exception was CDA_Mtu in which the active site arginine is located in the 114th position instead of the 75th position (Timmers et al., 2012). Additionally the Tyr37, Tyr40, Glu121 and Asn64 which were all shown to be important to quaternary structure formation are also conserved in all of the tCDA enzymes (Zilpa A. Sánchez-Quitian et al., 2011). Gln61 and Glu63 which interact with the sugar moiety of the substrate as well as Phe155 are also fully conserved. Main differences can be seen in CDA_Pco and CDA_Pin which have a longer sequence in a couple of loops (positions 100-105 and 125-129) and together with Human_CDA and Mouse_CDA have a longer loop in the N-terminal.

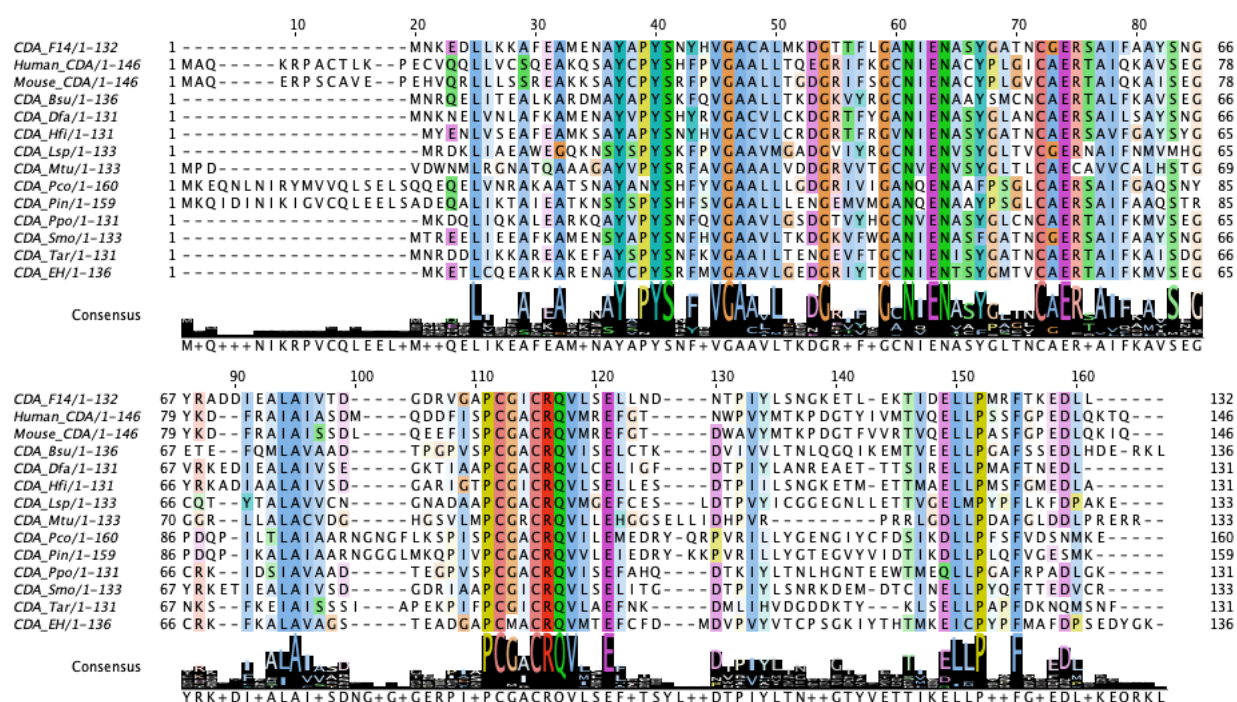


Figure 3.2 Sequence alignment of selected CDA enzymes

But even though the sequences of these tCDA's are very similar they have quite different substrate specificities towards N^4 -acyl-/ N^4 -alkyl-/ S^4 -alkyl-/ O^4 -alkoxy-nucleosides harbouring short, bulky aliphatic, aromatic and aryl groups (Supplementary figure 2.). CDA_F14, CDA_EH, CDA_Bsu and CDA_Lsp particularly have a wide range of substrate specificity. CDA_F14 and

CDA_Lsp can convert even substrates with two added benzoyl groups like N^4 -(3-benzoyl-benzoyl)-2'-deoxycytidine to dU (Supplementary table 2.). CDA_F14, CDA_EH, CDA_Bsu and CDA_Lsp were also equally able to convert oxy- and thio- substituted uridines. The most surprising activity was the ability of CDA_F14 to deaminate N^4 -acetyl-2'-deoxy-5'-O-DMT-cytidine, it is not understood how this reaction can happen as the three benzoyl groups on 5'-OH have no place in the active site. Docking simulations with this substrate did not give possible results, in all poses in the active site the sugar moiety exchanged places with the pyrimidine moiety, and catalysis cannot happen in this orientation. The structures of molecules are given in supplementary figure 1.

3.2 Structure modelling

CDA_F14 was both homology modelled and crystalized. The crystallization was performed by dr. Nina Urbelienė, and the structure was solved by dr. Giedrė Tamulaitienė. By the structure results, CDA_F14 has a canonical tCDA structure (Figure 3.3 A, B). The homology modelled structure compared with the *Bacillus subtilis* tCDA had $C\alpha$ root mean squared deviation (RMSD) value of 3.868 Å and the crystalized structure had a RMSD of 3.846 Å. The RMSD difference between all atoms of homology modelled and crystalized CDA_F14 was 0.871 Å, showing that the structures are overall very similar. Looking at previously determined cytidine deaminase models (PDB: 2FR6, 1R5T, 1JTK, 1MQ0, 3IJF, 1UX1), with and without substrates, there doesn't seem to be enough space in the active site to accommodate cytidine derivatives with bulky N^4 substitutes. Moreover, it is noted that some of these cytidine deaminases (PDB: 1JTK, 1MQ0) also fully engulf the substrate leaving it inaccessible to solvent (Chung et al., 2005; Johansson et al., 2002). Homology modelled CDA_F14 also did not seem to have enough space in the binding pocket to fit substrates like BzdC, although it was larger than when compared with the mouse or *Bacillus subtilis* tCDA enzymes (Figure 3.3 A). Docking of BzdC into the homology modelled structure was also only successful after MD simulations, but correct substrate

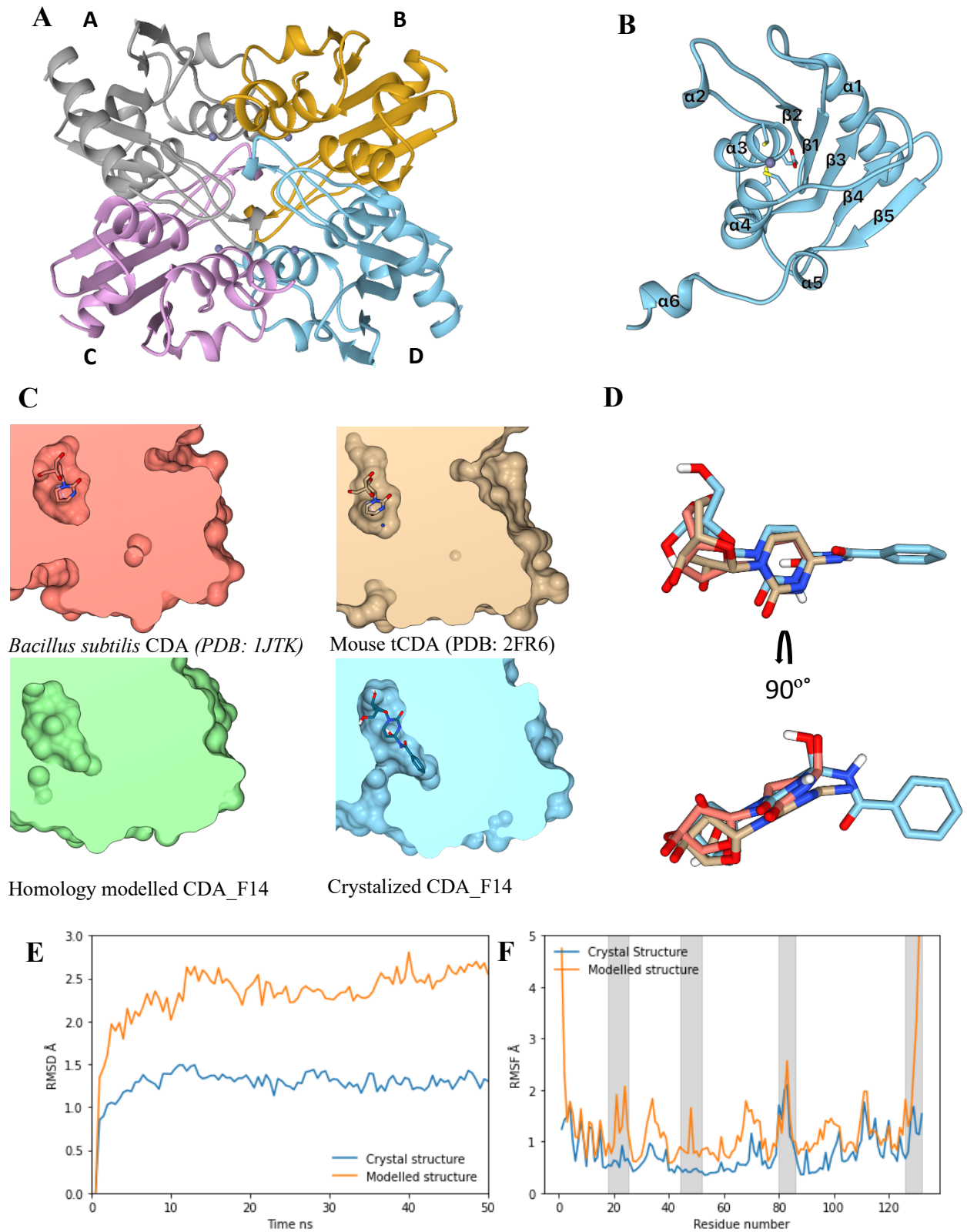


Figure 3.3 **A** – CDA_F14 crystal quaternary structure, **B** – CDA_F14 crystal monomer structure with marked alpha spirals and beta sheets, **C** – visualization of tCDA binding pocket space, moude tCDA has C, *Bacillus subtilis* has THU, CDA_F14 has BzdC in the active site, **D** – comparison of docked BzdC pose (blue) with C from mouse tCDA (tan) and THU from *Bacillus subtilis* CDA (orange red) **E** – CDA_F14 MD simulations RMSD values, **F** – CDA_F14 MD simulation RMSF values, the highlighted regions are: Tyr18-His25, Glu44-Asn52, Asp80-Ala86, Phe126-Leu132

poses according to known tCDA substrate complexes were not achieved. Performing MD simulations on both the modelled and crystalized CDA_F14 it was observed that the homology modelled tCDA had almost double C α atom RMSD values throughout the simulation and took longer to equilibrate (Figure 3.3 E). This shows that the homology modelled structure was further from a relaxed conformation than the crystal structure. The crystal CDA_F14 on the other hand had a noticeably larger binding pocket and it was possible to dock BzdC with a pose close to the expected orientation of the substrate in the active site of tCDA (Figure 3.3 C, D).

Molecular dynamics simulations revealed two main mobile regions in CDA_F14 that could also interact with the substrate. One was the C-end with the conserved phenylalanine – Phe126-Leu132, the other was a loop near the active site – Asp80-Ala86. The mobility of the loop near the active site was observed both in the MD simulation root mean square fluctuation (RMSF) data (Figure 3.3 F), and in the x-ray diffraction data where the residues Asp80-Arg83 exhibited mean B factors of 25.38 Å² compared to the average B factors of 17.57 Å² for the whole structure. Two other regions around the active site – Tyr18-His25 and Glu44-Asn52, that are known to interact with substrates and have higher RMSF values in *Mycobacterium tuberculosis* tCDA also showed relatively high RMSF values in CDA_F14 (Figure 3.3 F) (Zilpa A. Sánchez-Quitian et al., 2011).

Logically binding of larger substrates like BzdC requires more space in the active site. One possible criterion to separate which tCDA's are able of converting these bulky substrates could be checking the area of the active site i.e., solvent accessible surface area (SASA). Because most of the tCDA's selected in the study didn't have determined tertiary structures, their structures were modelled using deep learning structure prediction tools. Looking at the results, SASA doesn't strongly correlate with the volume of the enzymes substrates (Figure 3.4). This might be because the published structures of tCDA enzymes all seem to have smaller active sites or are crystalized with regular substrates like dC or equivalent pyrimidine like inhibitors. To check the influence of the template chosen on SASA, the structures were also modelled using CDA_F14 crystal structure as the reference. These results show a clearer relationship between the substrate volume and the active site SASA, but without knowing the true structures of the enzymes this approach doesn't provide a clear answer to whether the tCDA enzyme will work on large substrates like BzdC, because differently modelled structures have large differences in active site SASA's. Another interesting point is that judging by the *Bacillus subtilis* tCDA (PDB: 1JTK) active site size (Figure 3.3 A) it also wouldn't be able to fit BzdC, but it has been shown experimentally that CDA_Bsu can deamidate BzdC, although BzdC was a poor substrate. 1JTK has been crystalized together with the inhibitor THU and it might be that when

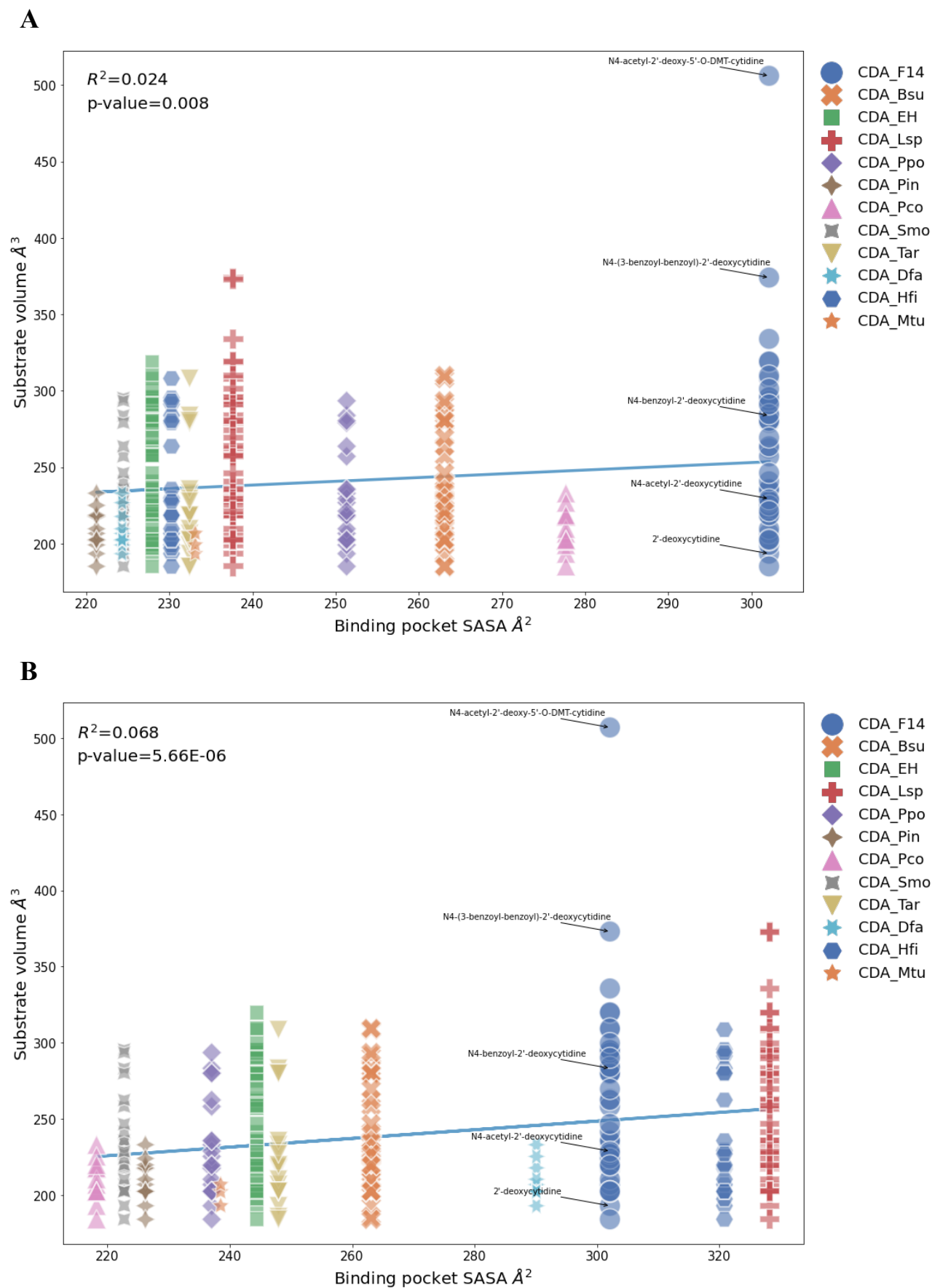


Figure 3.4 Substrate volume dependence on SASA. A – modelled tCDA structure SASA, **B** – modelled tCDA structure, using CDA_F14 crystal structure as a template, SASA

it binds a substrate the surrounding structure closes in on the substrate, the seemingly mobile Asp80-Ala86 loop might play a role in this structure tightening.

3.3 CDA_F14 interactions with BzdC

CDA_F14 interaction with BzdC was investigated using molecular docking and MD simulations. After examining tCDA structures crystalized with substrates, the most likely structure feature able to influence the bonding of large N^4 substituted substrates was the Asp80-Ala86 loop. The simulation studies revealed that the sugar and the pyrimidine moieties of BzdC bind the same way as described in previous studies of crystalized tCDA enzymes from *Bacillus subtilis*, *Mycobacterium tuberculosis* and Human tCDA (Figure 3.5) (Chung et al., 2005; Johansson et al., 2002; Zilpa Adriana Sánchez-Quitian et al., 2015). The 5'-OH is hydrogen bonded to Tyr450 (D)

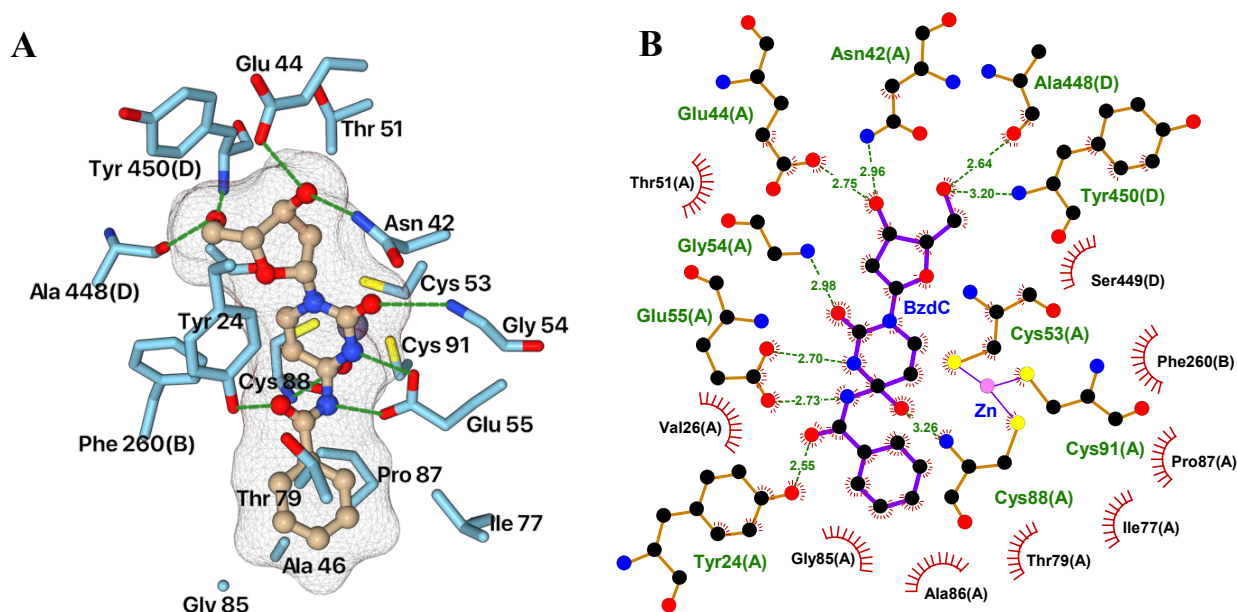


Figure 3.5 CDA_F14 interactions with BzdC in the active site. A – 3D view of BzdC in the active site, B – schematic of BzdC interactions with amino acids in the active site.

backbone -NH group and Ala448 backbone -O group, the 3'-OH is hydrogen bonded to Asn42 and Glu44. The base moiety interacts with the catalytic Glu55 through hydrogen bonds with the pyrimidine N3 and the amide bond N. The O2 is bound to Gly54 backbone -NH and the hydroxy group of the intermediate BzdC structure makes a hydrogen bond with the catalytic Cys88 backbone -NH group. One extra hydrogen bond comes from the amide bond oxygen binding with Tyr24. Tyrosine in this position is also found in CDA_Dfa and CDA_Hfi, other selected tCDA had a phenylalanine residue in this position (Figure 3.2). This tyrosine might help with amide binding, but it isn't necessary as enzymes with phenylalanine in this position are also able to deamidate amide substrates. Non-bonded interactions through π - π stacking with the pyrimidine moiety came from the

conserved Phe260, other residues close to the sugar moiety were Thr51 and Ser449. The benzoyl moiety of BzdC was in a hydrophobic pocket that was mainly made by the Asp80-Ala86 loop or residues close to it, interactions included Gly85, Ala86, Pro87, Thr79, Ile77 and Val26. The most important hydrogen bond interactions, made by Glu44, Asn42, Glu55 and Ala448, were maintained throughout most of the protein-substrate complex MD simulation, while other interactions were more transient (Table 3.1).

Table 3.1 Hydrogen bond interactions throughout the MD simulations of CDA_F14 and docked BzdC. **Acceptor/Donor** – acceptor and donor heavy atoms for the hydrogen bond, **Frames** – total number of frames, out of 200, the bond was formed, **Frac** – fractions of total frames the bond was formed, **AvgDist** – average distance of the formed hydrogen bond, **AvgAng** – average angle of the hydrogen bond

Acceptor	Donor	Frames	Frac	AvgDist	AvgAng
GLU_55@OE2	BzdC@O2	197	0.985	2.6242	166.5948
GLU_44@OE2	BzdC@O3'	195	0.975	2.6849	168.0457
ALA_448@O	BzdC@O5'	191	0.955	2.7092	157.8677
GLU_55@OE1	BzdC@N1	186	0.93	2.7894	150.8163
GLU_55@OE2	BzdC@N2	156	0.78	2.7866	144.8356
BzdC@O3'	ASN_42@ND2	129	0.645	2.8714	158.0506
BzdC@O	GLY_54@N	68	0.34	2.9115	162.4482
BzdC@O5'	TYR_450@N	33	0.165	2.9208	163.2453
BzdC@O1	TYR_24@OH	31	0.155	2.7251	163.8727
GLU_55@OE1	BzdC@O2	14	0.07	2.8251	143.4366
BzdC@O2	CYM_88@N	5	0.025	2.9486	165.3881

3.4 CDA_F14 mutational analysis

Based on the CDA_F14 and other tCDA structure models, docking of BzdC into these models and MD simulations of the enzyme and the enzyme substrate complexes several regions were chosen as sites for mutagenesis. Firstly, the Asp80-Ala86 loop was targeted, to check the presumptions of its importance to binding large N4 substituted substrates. 85th residue in CDA_F14 is in a prime position to influence BzdC binding. Human and mouse tCDA's have an isoleucine in this position, which looking at the crystal structure directly blocks the benzoyl group of BzdC. Moreover, the sequence alignment of tCDA's (Figure 3.2) showed that most enzymes active towards BzdC have either a glycine or alanine residue in this location, CDA_Bsu and CDA_Ppo are exceptions with a valine residue in this position. CDA_Tar is the only active tCDA with a leucine residue in this position, but it is noted that CDA_Tar is a thermophilic enzyme, it was most active in 60 °C. The higher temperature might change the enzyme structure and allow BzdC to enter the active site despite the leucine. Another chosen residue was the 81st, this residue is in the distal end of the binding site, but larger hydrophobic substitutions in this position should also stuff the active site and reduce space for the benzoyl moiety of BzdC. For comparison Human tCDA has a methionine and the mouse tCDA

has a leucine in this position. To test the importance of the Asp80-Ala86 loop region mutation of amino acids in the 81st and 85th positions were made. Mutations in the 81st position were CDA_F14 G81L and CDA_F14 G81L/G85I, in the 85th position CDA_F14 G85I, CDA_Lsp A82I, CDA_Pco I108A, CDA_Ppo V82L, CDA_Tar I85A. The logic for the changes were that large hydrophobic substitutions in this position would inhibit the ability to bind BzdC and vice versa – small amino acid residues would support substrate binding. The Arg83-Gly85 region was also deleted and randomly mutated. The conserved Phe126 was also chosen as a target to test the π - π stacking importance. Therefore, two variants – CDA_F14 F126A and F126W were constructed. The influence of C-end residues after the conserved Phe126 were tested with random mutagenesis and deletion of the Thr127-Asp130 region. Possible influence of Thr51 for interaction with the ribose moiety was tested by creating CDA_F14 T51G, CDA_Pco G70T, CDA_Ppo C50T mutants. Finally, the active site cysteine and arginine residues were tested by creating CDA_F14 R56L, CDA_F14 C53H/R56Q, CDA_F14 C88H/C91H and CDA_F14 C53H/R56Q/C88H/C91H mutants. The C53H/R56Q variant is meant to resemble the active of dCDA like in *Escherichia coli* dCDA and the C53H/R56Q/C88H/C91H variant is meant to mimic the active site of carbonic anhydrases where the zinc ion is coordinated by three histidine residues.

All mutant activity against BzdC was first checked on M9 selective medium using a uridine auxotrophic *Escherichia coli* strain (Figure 3.5). For the Asp80-Ala86 loop point mutants no complete activity reversals were observed, but the deletion mutant CDA_F14 del 83-85 lost its catalytic activity towards BzdC. The CDA_Tar I85A variant became active at 37 °C whereas wild-type CDA_Tar clone didn't complement uridine auxotrophic phenotype with BzdC as a uridine source. The active site CDA_F14 C88H/C91H and C53H/R56Q/C88H/C91H mutants lost all enzymatic activity, only the CDA_F14 C53H/R56Q variant showed some deaminase activity against dC. The random mutagenesis mutants CDA_F14 ₈₃SML₈₅, ₈₃HSL₈₅, ₈₃QQS₈₅, ₁₂₇HSSG₁₃₀, ₁₂₇CLYR₁₃₀ also retained their activity against BzdC, showing that leucine as well as isoleucine and serine in the 85th position don't completely block BzdC binding and that the residues after the conserved Phe126 also most likely do not have much influence on overall enzymatic activity. All other mutants complemented uridine auxotrophic phenotype on the M9 medium with Bzdc.

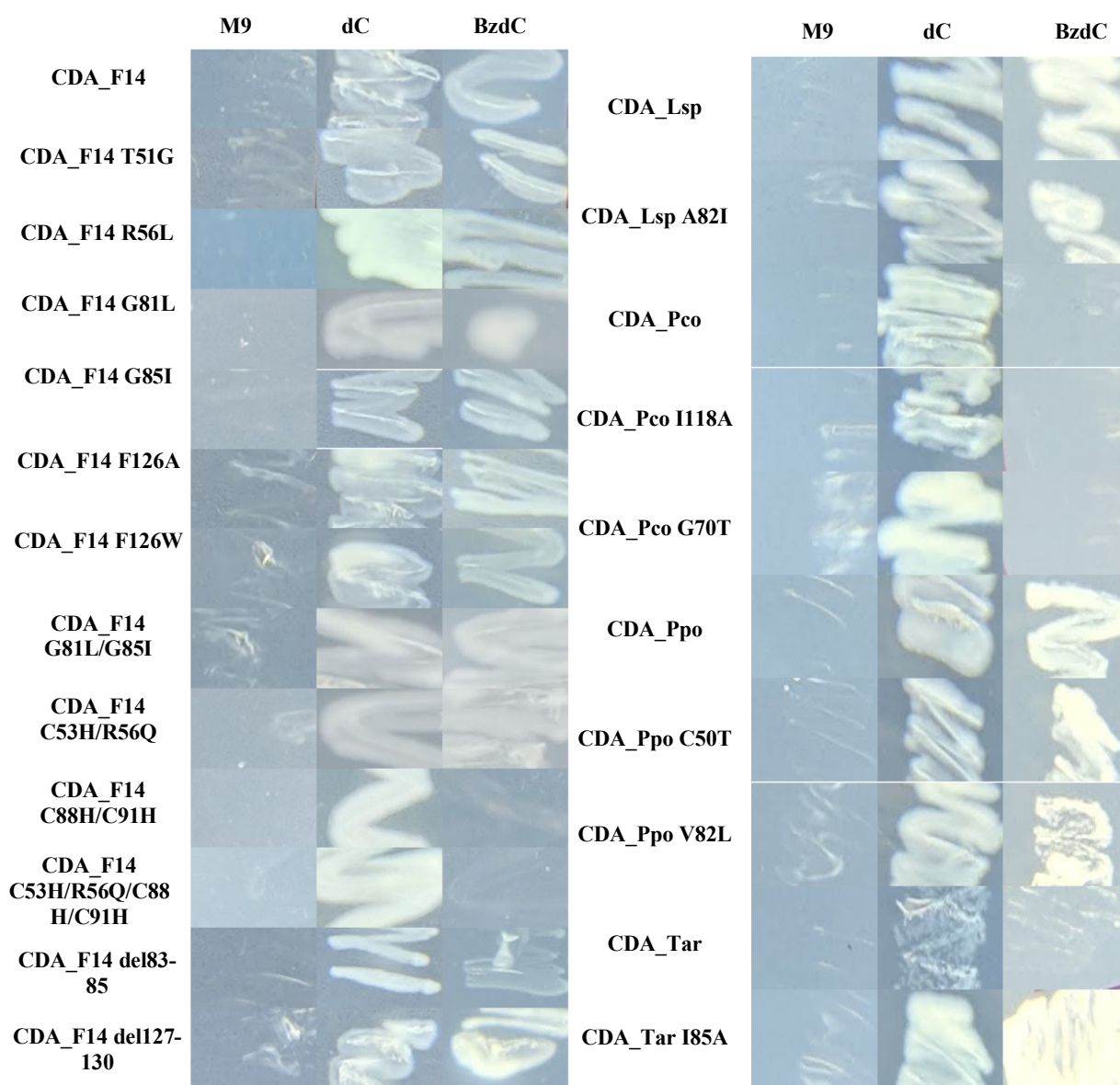


Figure 3.5 tCDA mutant activity testing on M9 medium. M9 – no added pyrimidine source (negative control), dC – added dC as a pyrimidine source (positive control), BzdC – added BzdC as pyrimidine source (growth only if the enzyme deamidate BzdC)

3.5 CDA_F14 enzyme kinetics and substrate specificities

To evaluate differences in activity, CDA_F14 mutants G81L, G85I, G81L/G85I, F126A, F126W, T51G, R56L, del83-85, del127-130, C53H/R56Q, C88H/C91H and C53H/R56Q/C88H/C91H mutants were purified, and their kinetic parameters (Table 3.2) and substrate specificities were determined (Supplementary table 2.). The mutants CDA_Lsp A82I, CDA_Ppo C50T and V82L, CDA_Pco G70T and I108A, CDA_Tar I85A and random mutagenesis mutants CDA_F14 ⁸³SML₈₅, ⁸³HSL₈₅, ⁸³QQS₈₅, ¹²⁷HSSG₁₃₀ and ¹²⁷CLYR₁₃₀ were also purified and their substrate specificities were determined (Supplementary table 2.).

Table 3.2 Kinetic parameters of the wild-type and mutant CDA_F14 towards BzdC and dC

	Substrate	K_m (M)	k_{cat} (s^{-1})	k_{cat}/K_m (M^{-1}/s^{-1})
CDA_F14	BzdC	$(1.15 \pm 0.16) \times 10^{-4}$	$(5.04 \pm 0.4) \times 10^{-1}$	$(4.36 \pm 3.61) \times 10^3$
	dC	$(1.95 \pm 0.36) \times 10^{-4}$	$(24.4 \pm 1.71) \times 10^{-1}$	$(12.5 \pm 1.2) \times 10^3$
CDA_F14 T51G	BzdC	$(4.25 \pm 2.65) \times 10^{-4}$	$(8.90 \pm 5.90) \times 10^{-3}$	$(2.09 \pm 5.14) \times 10^{-1}$
	dC	$(0.81 \pm 3.67) \times 10^{-2}$	$(0.57 \pm 2.57) \times 10^1$	$(3.10 \pm 2.60) \times 10^2$
CDA_F14 G81L	BzdC	$(8.66 \pm 3.8) \times 10^{-4}$	$(7.89 \pm 3.08) \times 10^{-1}$	$(9.11 \pm 0.3) \times 10^2$
	dC	$(2.27 \pm 0.15) \times 10^{-4}$	$(1.97 \pm 0.05) \times 10^{-1}$	$(8.69 \pm 0.33) \times 10^2$
CDA_F14 G85I	BzdC	$(1.56 \pm 0.1) \times 10^{-4}$	$(2.36 \pm 0.01) \times 10^{-1}$	$(1.52 \pm 0.08) \times 10^3$
	dC	$(2.96 \pm 0.88) \times 10^{-4}$	$(6.87 \pm 0.80) \times 10^{-1}$	$(2.32 \pm 0.32) \times 10^3$
CDA_F14 R56L	BzdC	$(1.33 \pm 0.17) \times 10^{-4}$	$(1.22 \pm 0.09) \times 10^{-2}$	$(9.20 \pm 0.42) \times 10^1$
	dC	$(1.67 \pm 0.19) \times 10^{-4}$	$(2.58 \pm 0.09) \times 10^{-2}$	$(15.5 \pm 1.1) \times 10^1$
CDA_F14 G81L/G85I	BzdC	$(4.93 \pm 1.3) \times 10^{-4}$	$(4.14 \pm 0.88) \times 10^{-2}$	$(8.4 \pm 0.3) \times 10^1$
	dC	$(2.93 \pm 0.5) \times 10^{-4}$	$(6.2 \pm 0.44) \times 10^{-1}$	$(2.12 \pm 0.2) \times 10^3$
CDA_F14 del127-130	BzdC	$(1.46 \pm 0.01) \times 10^{-4}$	$(5.41 \pm 0.23) \times 10^{-2}$	$(3.70 \pm 0.09) \times 10^2$
CDA_F14 F126A	BzdC	$(2.59 \pm 0.73) \times 10^{-4}$	$(3.88 \pm 0.80) \times 10^{-2}$	$(1.50 \pm 0.091) \times 10^2$
CDA_F14 F126W	BzdC	$(2.11 \pm 0.38) \times 10^{-4}$	$(10.4 \pm 1.3) \times 10^{-2}$	$(4.92 \pm 0.24) \times 10^2$

Main changes observed in substrate specificity changes were CDA_Lsp A82I losing ability to deamidate N^4 -(4-benzoyl-benzoyl)-2'-dC and N^4 -nicotinoyl-2'-dC, this goes in line with the hypothesis that large aliphatic amino acids in the 85th position hinder substrate with bulky N^4 group binding. All CDA_F14 mutants together with CDA_Lsp A82I also lost the ability to deamidate Capecitabine and N^4 -(2,3,4,5,6-pentahydroxyhexyl)-2'-dC. CDA_Lsp A82I also lost activity towards some thio- and alkoxy- compounds. CDA_Ppo V82L lost activity towards N^4 -acetyl-2'-dC, which was surprising because it was still active towards BzdC. CDA_Pco G70T surprisingly lost activity towards N^4 -methylcytidine. Deletion of the 127-130 residues in CDA_F14 did not have a large effect on substrate specificity except for N^4 -(2,3,4,5,6-pentahydroxyhexyl)-2'-dC, whereas the deletion of the residues 83-85 strongly affected activity towards most substrates. Other tCDA variants didn't have major substrate specificity changes.

The C53H/R56Q, C88H/C91H, C53H/R56Q/C88H/C91H, del83-85 variants kinetic activity was not evaluated because their rate of hydrolysis was too low to measure spectrophotometrically. Overall, all mutated variants had lower enzymatic efficiency with BzdC and dC compared to wild-type CDA_F14. The G81L mutations reduced the catalytic efficiency (k_{cat}/K_m) by ~5 times, the G85I only reduced the catalytic efficiency by ~3 times, but the double G81L/G85I mutant reduced it by ~53 times. This shows that by themselves these mutations are not too impactful, but together they effectively inhibit both binding and deamidation of BzdC. The R56L mutant didn't reduce affinity for either BzdC or dC but it reduced the turnover rate (k_{cat}) by ~41 and ~94 times respectively, these results agree with previous findings that exchanging the positive arginine to an uncharged residue

doesn't impact K_m values much but severely decreases the reaction speed (Johansson et al., 2004). The deletion of residues 127 through 130 also reduced the catalytic efficiency toward BzdC by ~14 times compared with the wild-type enzyme, but this effect might have been caused by bringing the histidine tag used for purification closer to the active site. The F126A and F126W mutants negatively impacted both substrate affinity (K_m) and the turnover rate, the F126A variant reduced catalytic efficiency by ~30 times and the F126W variant reduced it by ~10, this difference probably comes from the ability of tryptophan to still provide π - π stacking, even though its bulk could have made these interactions harder to achieve because of steric hindrances. Unlike in the human tCDA the F126A mutant of CDA_F14 was soluble and still exhibited measurable catalytic activity (Vincenzetti et al., 2008). The T51G mutant results were hard to interpret because of the high error values, judging from the reaction curve, which was linear instead of hyperbolic (Figure 3.6), it seemed this variant had a very high K_m value and even the highest substrate concentrations were still below it. The high K_m might have been caused by an induced cooperativity between subunit as the Thr51 also maintained a hydrogen bond with the Tyr48 backbone oxygen atom for a total of 66 ns out of a 100 ns of the MD simulation of the substrate and enzyme. Cooperativity in tCDA hasn't been shown previously even though there are tight interactions between the monomers of tCDA (Vincenzetti et al., 2008). The overall decrease in the enzymatic activities of the mutated CDA_F14 variants is well illustrated by plotting the catalytic turnover values for different substrate concentrations (Figure 3.6). Another interesting observation is that the decrease in activity was proportionally higher for dC than it was for BzdC even for a mutant like G85I which was meant to mainly impact interactions with BzdC. This shows that the changes made most likely had a nonlocal effect that disturbed the protein structure and affected the interactions and geometries of substituted and unsubstituted substrates in the active site. The only exception to this observation was the double G81L/G85I variant, its activity against dC was comparable to the G85I variant, and higher than the G81L variant. It might be that the G85I mutation complemented the G81L mutation and restored some of the enzyme's activity towards dC.

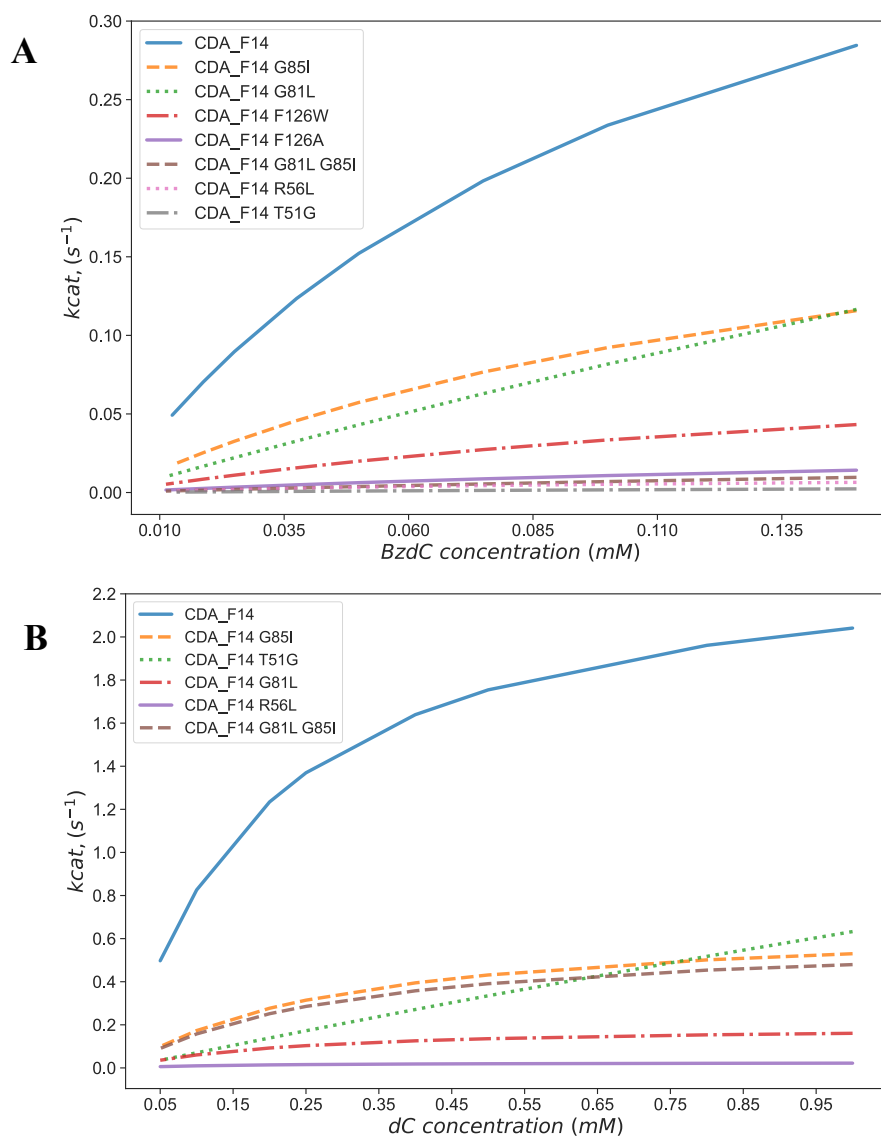


Figure 3.6 Activity graphs of CDA_F14 and its mutants. A – catalytic efficiency of CDA_F14 and its mutants against BzdC, B – catalytic efficiency of CDA_F14 and its mutants against dC

CONCLUSIONS

1. Modelled structures of cytidine deaminases have a similar core of the $\alpha/\beta/\alpha$ deaminase domain. Each subunit consists of a core of five β -strands (β 1- β 5), sandwiched by six α -helices (α 1- α 6).
2. The loop Asp80-Ala86 in CDA_F14 is the main region, which impacts the binding of N^4 substituted pyrimidine analogues.
3. The amino acids in 81st and 85th positions strongly affect the catalysis of N^4 substituted pyrimidine analogues in CDA_F14. Deletion of the 83-85 region drastically reduces the range of substrates and the activity of the enzyme towards them.
4. Mutations of the conserved Phe126 negatively influence catalytic properties of CDA_F14, deletion of the 127-130 region doesn't have a pronounced effect on substrate specificity but impacts enzymes catalytic efficiency towards N^4 -benzoyl-2'-deoxycytidine.
5. Mutations of the active site cysteine residues to histidines results in the loss of enzymatic activity in CDA_F14.
6. The T51G mutation affects activity of CDA_F14 due to the altered interaction between the enzyme's subunits.

SUMMARY

VILNIUS UNIVERSITY

Life Sciences Centre

Institute of Biochemistry

MATAS TIŠKUS

Molecular modelling and the study of the structure-function relationship of cytidine deaminases

Master thesis

During this study, tetrameric cytidine deaminases exhibiting novel nucleophilic substitution activities in the 4th position of the heterocyclic ring i.e., *N*⁴-acyl-/*N*⁴-alkyl, *N*⁴-carboxy, *S*⁴-alkyl and *O*⁴-alkoxy cytidine substrates converting them to uridine and the according amide, amine, carbamate, thiol, or alcohol were investigated. Before this study, these activities of cytidine deaminases were not known. The aim of the study was to, using various molecular modelling techniques, determine structural factors in tetrameric cytidine deaminase CDA_F14, which led to deaminase and deamidase activity against *N*⁴ substituted pyrimidine analogues. Initial results revealed that CDA_F14 has a seemingly mobile loop near the active site which could be mainly responsible for the observed activities. This Asp80-Ala86 loop together with the active site and the C-end of the enzyme were investigated using site-directed mutagenesis. The results revealed that the G81L and G85I mutations, when a small aliphatic residue is exchanged for a larger aliphatic residue, lead to a decrease in enzyme activity against both 2'-deoxycytidine and *N*⁴-benzoyl-2'-deoxycytidine. The active site cysteine mutations into histidines made the enzyme inactive. The T51G mutants exhibited kinetic parameters that may suggest induced cooperativity between the enzyme's subunits. The F126A and F126W variants were both active, but the F126W mutant had higher activity because it could still contribute to π - π stacking. Deletions of amino acids in the 127-130 positions don't lead to major substrate specificity changes, but deletion of the 83-85 residues limits both the enzymes substrate spectrum and overall enzyme efficiency.

SANTRAUKA

VILNIAUS UNIVERSITETAS

Gyvybės mokslų centras

Biochemijos institutas

MATAS TIŠKUS

Citidino deaminazių molekulinis modeliavimas ir jų struktūros-funkcijos ryšio tyrimas

Magistro darbas

Šio darbo metu tirtos tetramerinės citidino deaminazės, kurios katalizuoja nukleofilinę pakeitimo reakciją 4-oje heterociklinio žiedo padėtyje, tai yra N^4 -acil-, N^4 -alkil-, N^4 -karboksi-, S^4 -alkil- ir O^4 -alkoksicitidinai paverčiami uridinu ir atitinkamai amidu, aminu, karbamatu, tioliu arba alkoholiu. Iki šiol nebuvo žinoma, kad citidino deaminazės gali katalizuoti tokio tipo reakcijas. Šio tyrimo tikslas buvo naudojant įvairius molekulinio modeliavimo metodus išsiaiškinti struktūros elementus lemiančius CDA_F14 gebėjimą deamininti ir deamidinti pirimidino nukleozidų darinius su pakaitais 4-oje heterociklinės bazės padėtyje. Pirminiai rezultatai atskleidė, jog CDA_F14 turi galimai judrią kilpą šalia aktyviojo fermento centro, kuri gali nulemti stebėtus fermentinius aktyvumus. Ši Asp80-Ala86 kilpa kartu su aktyviojo centro ir C galo amino rūgštimis buvo tirtos pasitelkiant tikslią mutagenę. Rezultatai atskleidė, jog G81L ir G85I mutacijos, kai maža amino rūgštis pakeičiama didesne alifatine amino rūgšties liekana, lemia sumažėjusį fermentinį aktyvumą reakcijoje su 2'-deoksicitidinu ir N^4 -benzoil-2'-deoksicitidinu. Aktyviojo centro cisteinų mutacijos į histidinus inaktyvuoja fermentinį aktyvumą. T51G mutacija rodo kinetinius parametrus kurie gali būti nulemti šios mutacijos sukulto kooperatyvumo tarp fermento subvienetų. F126A ir F125W mutantai išliko aktyvūs, bet F126W išlaikė aukštesnį aktyvumo lygį, tai galimai lėmė išlaikytos π - π sąveikos F126W mutante. Amino rūgščių 127-130 pozicijose delecija nesukelia esminių pokyčių substratų selektyvumui, o 83-85 pozicijų delecija smarkiai sumažina katalizuojamų substratų spektrą ir bendrą fermento aktyvumą.

Acknowledgements

I am extremely thankful to dr. Nina Urbelienė for her constant support and encouragement throughout my master's studies. I also want to thank prof. Rolandas Meškys for his guidance and the opportunity to work at the molecular microbiology and biotechnology department.

This work was supported by the Research Council of Lithuania. “Selective enzymatic system for prodrug activation” No. 01.2.2-LMT-K-718-03-0082.

List of Publications

R. Meškys, N. Urbelienė, D. Tauraitė, M. Tiškus, V. Preitakaitė “HIDROLAZĖS IR JŲ PANAUDOJIMAS” (“HYDROLASES AND USES THEREOF”). Lithuanian patent application LT2022 514.

LITERATURE

1. Agresti, J. J., Antipov, E., Abate, A. R., Ahn, K., Rowat, A. C., Baret, J. C., Marquez, M., Klibanov, A. M., Griffiths, A. D., & Weitz, D. A. (2010). Ultrahigh-throughput screening in drop-based microfluidics for directed evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(9), 4004–4009. <https://doi.org/10.1073/pnas.0910781107>
2. Aučynaitė, A., Rutkienė, R., Gasparavičiūtė, R., Meškys, R., & Urbonavičius, J. (2018). A gene encoding a DUF523 domain protein is involved in the conversion of 2-thiouracil into uracil. *Environmental Microbiology Reports*, *10*(1), 49–56. <https://doi.org/10.1111/1758-2229.12605>
3. Badran, A. H., & Liu, D. R. (2015). Development of potent in vivo mutagenesis plasmids with broad mutational spectra. *Nature Communications* *2015 6:1*, *6*(1), 1–10. <https://doi.org/10.1038/ncomms9425>
4. Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Dustin Schaeffer, R., Millán, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A., Pereira, J. H., Rodrigues, A. V., Van Dijk, A. A., Ebrecht, A. C., ... Baker, D. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, *373*(6557), 871–876. <https://doi.org/10.1126/>
5. Banno, S., Nishida, K., Arazoe, T., Mitsunobu, H., & Kondo, A. (2018). Deaminase-mediated multiplex genome editing in *Escherichia coli*. *Nature Microbiology*, *3*(4), 423–429. <https://doi.org/10.1038/s41564-017-0102-6>
6. Biot-Pelletier, D., & Martin, V. J. J. (2016). Seamless site-directed mutagenesis of the *Saccharomyces cerevisiae* genome using CRISPR-Cas9. In *Journal of Biological Engineering* (Vol. 10, Issue 1, pp. 1–5). BioMed Central Ltd. <https://doi.org/10.1186/S13036-016-0028-1>
7. Bornscheuer, U. T., Hauer, B., Jaeger, K. E., & Schwaneberg, U. (2019). Directed Evolution Empowered Redesign of Natural Proteins for the Sustainable Production of Chemicals and Pharmaceuticals. *Angewandte Chemie International Edition*, *58*(1), 36–40. <https://doi.org/10.1002/ANIE.201812717>
8. Bornscheuer, U. T., & Höhne, M. (2018). Protein engineering - Methods and protocols. *Methods in Molecular Biology*, *1685*, 1–347. <https://doi.org/10.1007/978-1-4939-7366-8>
9. Bradford, M. M. (1976). A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Analytical Biochemistry*, *72*(1–2), 248–254. [https://doi.org/10.1016/0003-2697\(76\)90527-3](https://doi.org/10.1016/0003-2697(76)90527-3)
10. Burke, A. J., Birmingham, W. R., Zhuo, Y., Thorpe, T. W., Zucoloto da Costa, B., Crawshaw, R., Rowles, I., Finnigan, J. D., Young, C., Holgate, G. M., Muldowney, M. P., Charnock, S. J., Lovelock, S. L., Turner, N. J., & Green, A. P. (2022). An Engineered Cytidine Deaminase for Biocatalytic Production of a Key Intermediate of the Covid-19 Antiviral Molnupiravir. *Journal of the American Chemical Society*, *144*(9), 3761–3765. <https://doi.org/10.1021/jacs.1c11048>
11. Carlow, D. C., Carter, C. W., Mejlhede, N., Neuhard, J., & Wolfenden, R. (1999). Cytidine deaminases

- from *B. subtilis* and *E. coli*: compensating effects of changing zinc coordination and quaternary structure. *Biochemistry*, 38(38), 12258–12265. <https://doi.org/10.1021/BI990819T>
12. Chen, X. S. (2021). Insights into the Structures and Multimeric Status of APOBEC Proteins Involved in Viral Restriction and Other Cellular Functions. *Viruses*, 13(3). <https://doi.org/10.3390/V13030497>
 13. Childers, M. C., & Daggett, V. (2017). Insights from molecular dynamics simulations for computational protein design. *Molecular Systems Design and Engineering*, 2(1), 9–33. <https://doi.org/10.1039/c6me00083e>
 14. Chung, S. J., Fromme, J. C., & Verdine, G. L. (2005). Structure of human cytidine deaminase bound to a potent inhibitor. *Journal of Medicinal Chemistry*, 48(3), 658–660. <https://doi.org/10.1021/jm0496279>
 15. Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & De Hoon, M. J. L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11), 1422–1423. <https://doi.org/10.1093/BIOINFORMATICS/BTP163>
 16. Cohen, R. M., & Wolfenden, R. (1971). Cytidine Deaminase from *Escherichia coli*. *Journal of Biological Chemistry*, 246, 7561–7565. [https://doi.org/10.1016/S0021-9258\(19\)45812-2](https://doi.org/10.1016/S0021-9258(19)45812-2)
 17. Costanzi, S., Vilar, S., Micozzi, D., Carpi, F. M., Ferino, G., Vita, A., & Vincenzetti, S. (2011). Delineation of the molecular mechanisms of nucleoside recognition by cytidine deaminase through virtual screening. *ChemMedChem*, 6(8), 1452. <https://doi.org/10.1002/CMDC.201100139>
 18. Dapkūnas, J., Olechnovič, K., & Venclovas, Č. (2018). Modeling of protein complexes in CAPRI Round 37 using template-based approach combined with model selection. *Proteins: Structure, Function, and Bioinformatics*, 86, 292–301. <https://doi.org/10.1002/PROT.25378>
 19. Dhillon, S. (2020). Decitabine/Cedazuridine: First Approval. *Drugs*, 80(13), 1373–1378. <https://doi.org/10.1007/S40265-020-01389-7>
 20. Frances, A., & Cordelier, P. (2020). The Emerging Role of Cytidine Deaminase in Human Diseases: A New Opportunity for Therapy? *Molecular Therapy*, 28(2), 357–366. <https://doi.org/10.1016/j.ymthe.2019.11.026>
 21. Gabler, F., Nam, S. Z., Till, S., Mirdita, M., Steinegger, M., Söding, J., Lupas, A. N., & Alva, V. (2020). Protein Sequence Analysis Using the MPI Bioinformatics Toolkit. *Current Protocols in Bioinformatics*, 72(1), 1–30. <https://doi.org/10.1002/cpbi.108>
 22. Geller, L. T., Barzily-Rokni, M., Danino, T., Jonas, O. H., Shental, N., Nejman, D., Gavert, N., Zwang, Y., Cooper, Z. A., Shee, K., Thaiss, C. A., Reuben, A., Livny, J., Avraham, R., Frederick, D. T., Ligorio, M., Chatman, K., Johnston, S. E., Mosher, C. M., ... Straussman, R. (2017). Potential role of intratumor bacteria in mediating tumor resistance to the chemotherapeutic drug gemcitabine. *Science (New York, N.Y.)*, 357(6356), 1156–1160. <https://doi.org/10.1126/SCIENCE.AAH5043>
 23. Geller, L. T., & Straussman, R. (2018). Intratumoral bacteria may elicit chemoresistance by metabolizing anticancer agents. *Molecular and Cellular Oncology*, 5(1). <https://doi.org/10.1080/23723556.2017.1405139>
 24. Greenblatt, H. M., Feinberg, H., Tucker, P. A., & Shoham, G. (1998). Carboxypeptidase A: Native, zinc-

- removed and mercury-replaced forms. *Acta Crystallographica Section D: Biological Crystallography*, 54(3), 289–305. <https://doi.org/10.1107/S0907444997010445>
25. Gutte, B. (1975). A synthetic 70 amino acid residue analog of ribonuclease S protein with enzymic activity. *Journal of Biological Chemistry*, 250(3), 889–904.
 26. Hall, R. S., Fedorov, A. A., Xu, C., Fedorov, E. V., Almo, S. C., & Raushel, F. M. (2011). Three-dimensional structure and catalytic mechanism of cytosine deaminase. *Biochemistry*, 50(22), 5077–5085. <https://doi.org/10.1021/BI200483K>
 27. Hattori, N., Sako, M., Kimura, K., Iida, N., Takeshima, H., Nakata, Y., Kono, Y., & Ushijima, T. (2019). Novel prodrugs of decitabine with greater metabolic stability and less toxicity. *Clinical Epigenetics*, 11(1), 1–12. <https://doi.org/10.1186/s13148-019-0709-y>
 28. Henderson, S., & Fenton, T. (2015). APOBEC3 genes: retroviral restriction factors to cancer drivers. *Trends in Molecular Medicine*, 21(5), 274–284. <https://doi.org/10.1016/J.MOLMED.2015.02.007>
 29. Hiranuma, N., Park, H., Baek, M., Anishchenko, I., Dauparas, J., & Baker, D. (2021). Improved protein structure refinement guided by deep learning based accuracy estimation. *Nature Communications 2021 12:1*, 12(1), 1–11. <https://doi.org/10.1038/s41467-021-21511-x>
 30. Holden, L. G., Prochnow, C., Chang, Y. P., Bransteitter, R., Chelico, L., Sen, U., Stevens, R. C., Goodman, M. F., & Chen, X. S. (2008). Crystal structure of the anti-viral APOBEC3G catalytic domain and functional implications. *Nature*, 456(7218), 121–124. <https://doi.org/10.1038/nature07357>
 31. Huang, T. P., Newby, G. A., & Liu, D. R. (2021). Precision genome editing using cytosine and adenine base editors in mammalian cells. *Nature Protocols 2021 16:2*, 16(2), 1089–1128. <https://doi.org/10.1038/s41596-020-00450-9>
 32. Ireton, G. C., McDermott, G., Black, M. E., & Stoddard, B. L. (2002). The structure of Escherichia coli cytosine deaminase. *Journal of Molecular Biology*, 315(4), 687–697. <https://doi.org/10.1006/jmbi.2001.5277>
 33. Jansen, R. S., Rosing, H., Schellens, J. H. M., & Beijnen, J. H. (2011). Deoxyuridine analog nucleotides in deoxycytidine analog treatment: Secondary active metabolites? *Fundamental and Clinical Pharmacology*, 25(2), 172–185. <https://doi.org/10.1111/j.1472-8206.2010.00823.x>
 34. Johansson, E., Mejlhede, N., Neuhard, J., & Larsen, S. (2002). Crystal structure of the tetrameric cytidine deaminase from Bacillus subtilis at 2.0 Å resolution. *Biochemistry*, 41(8), 2563–2570. <https://doi.org/10.1021/bi011849a>
 35. Johansson, E., Neuhard, J., & Willemoe, M. (2004). *Structural , Kinetic , and Mutational Studies of the Zinc Ion Environment in*. 6020–6029.
 36. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Židek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., ... Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature 2021 596:7873*, 596(7873), 583–589. <https://doi.org/10.1038/s41586-021-03819-2>
 37. Kim, Y. B., Komor, A. C., Levy, J. M., Packer, M. S., Zhao, K. T., & Liu, D. R. (2017). Increasing the

- genome-targeting scope and precision of base editing with engineered Cas9-cytidine deaminase fusions. *Nature Biotechnology* 2017 35:4, 35(4), 371–376. <https://doi.org/10.1038/nbt.3803>
38. Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A., & Liu, D. R. (2016). Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* 2015 533:7603, 533(7603), 420–424. <https://doi.org/10.1038/nature17946>
 39. Kosuri, S., & Church, G. M. (2014). Large-scale de novo DNA synthesis: Technologies and applications. In *Nature Methods* (Vol. 11, Issue 5, pp. 499–507). Nature Publishing Group. <https://doi.org/10.1038/nmeth.2918>
 40. Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L., & Baker, D. (2003). Design of a novel globular protein fold with atomic-level accuracy. *Science (New York, N.Y.)*, 302(5649), 1364–1368. <https://doi.org/10.1126/SCIENCE.1089427>
 41. Lai, Y. P., Huang, J., Wang, L. F., Li, J., & Wu, Z. R. (2004). A new approach to random mutagenesis in vitro. *Biotechnology and Bioengineering*, 86(6), 622–627. <https://doi.org/10.1002/bit.20066>
 42. Lemán, J. K., Weitzner, B. D., Lewis, S. M., Adolf-Bryfogle, J., Alam, N., Alford, R. F., Aprahamian, M., Baker, D., Barlow, K. A., Barth, P., Basanta, B., Bender, B. J., Blacklock, K., Bonet, J., Boyken, S. E., Bradley, P., Bystroff, C., Conway, P., Cooper, S., ... Bonneau, R. (2020). Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nature Methods* 2020 17:7, 17(7), 665–680. <https://doi.org/10.1038/s41592-020-0848-2>
 43. Liu, J., Liu, J., Zhao, D., Ma, N., & Luan, Y. (2016). Highly enhanced leukemia therapy and oral bioavailability from a novel amphiphilic prodrug of cytarabine. *RSC Advances*, 6(42), 35991–35999. <https://doi.org/10.1039/C6RA02051H>
 44. Liu, W., Shang, F., Chen, Y., Lan, J., Wang, L., Chen, J., Gao, P., Ha, N. C., Quan, C., Nam, K. H., & Xu, Y. (2019). Biochemical and structural analysis of the *Klebsiella pneumoniae* cytidine deaminase CDA. *Biochemical and Biophysical Research Communications*, 519(2), 280–286. <https://doi.org/10.1016/j.bbrc.2019.08.167>
 45. Ludford, P. T., Li, Y., Yang, S., & Tor, Y. (2021). Cytidine deaminase can deaminate fused pyrimidine ribonucleosides. *Organic and Biomolecular Chemistry*, 19(28), 6237–6243. <https://doi.org/10.1039/D1OB00705J>
 46. Lvarez, P., Marchal, J. A., Boulaiz, H., Carrillo, E., Vélez, C., Rodríguez-Serrano, F., Melguizo, C., Prados, J., Madeddu, R., & Aranega, A. (2012). 5-Fluorouracil derivatives: A patent review. *Expert Opinion on Therapeutic Patents*, 22(2), 107–123. <https://doi.org/10.1517/13543776.2012.661413>
 47. Marquez, V. E., Schroeder, G. K., Ludek, O. R., Siddiqui, M. A., Ezzitouni, A., & Wolfenden, R. (2009). Contrasting Behavior of Conformationally Locked Carbocyclic Nucleosides of Adenosine and Cytidine as Substrates for Deaminases. <https://doi.org/10.1080/15257770903091904>, 28(5–7), 614–632. <https://doi.org/10.1080/15257770903091904>
 48. Matsubara, T., Ishikura, M., & Aida, M. (2006). A quantum chemical study of the catalysis for cytidine deaminase: Contribution of the extra water molecule. *Journal of Chemical Information and Modeling*, 46(3), 1276–1285. <https://doi.org/10.1021/ci050479k>

49. Mejlhede, N., & Neuhard, J. (2000). The role of zinc in *Bacillus subtilis* cytidine deaminase. *Biochemistry*, *39*(27), 7984–7989. <https://doi.org/10.1021/BI000542T>
50. Mickevičiūtė, A., Timm, D. D., Gedgaudas, M., Linkuvienė, V., Chen, Z., Waheed, A., Michailovienė, V., Zubrienė, A., Smirnov, A., Čapkauskaitė, E., Baranauskienė, L., Jachno, J., Revuckienė, J., Manakova, E., Gražulis, S., Matulienė, J., Di Cera, E., Sly, W. S., & Matulis, D. (2018). Intrinsic thermodynamics of high affinity inhibitor binding to recombinant human carbonic anhydrase IV. *European Biophysics Journal*, *47*(3), 271–290. <https://doi.org/10.1007/S00249-017-1256-0>
51. Micozzi, D., Carpi, F. M., Pucciarelli, S., Polzonetti, V., Polidori, P., Vilar, S., Williams, B., Costanzi, S., & Vincenzetti, S. (2014a). Human cytidine deaminase: a biochemical characterization of its naturally occurring variants. *International Journal of Biological Macromolecules*, *63*, 64–74. <https://doi.org/10.1016/J.IJBIOMAC.2013.10.029>
52. Micozzi, D., Carpi, F. M., Pucciarelli, S., Polzonetti, V., Polidori, P., Vilar, S., Williams, B., Costanzi, S., & Vincenzetti, S. (2014b). Human cytidine deaminase: A biochemical characterization of its naturally occurring variants. *International Journal of Biological Macromolecules*, *63*, 64–74. <https://doi.org/10.1016/j.ijbiomac.2013.10.029>
53. Micozzi, D., Pucciarelli, S., Carpi, F. M., Costanzi, S., De Sanctis, G., Polzonetti, V., Natalini, P., Santarelli, I. F., Vita, A., & Vincenzetti, S. (2010). Role of tyrosine 33 residue for the stabilization of the tetrameric structure of human cytidine deaminase. *International Journal of Biological Macromolecules*, *47*(4), 471–482. <https://doi.org/10.1016/J.IJBIOMAC.2010.07.001>
54. Mok, B. Y., de Moraes, M. H., Zeng, J., Bosch, D. E., Kotrys, A. V., Raguram, A., Hsu, F. S., Radey, M. C., Peterson, S. B., Mootha, V. K., Mougous, J. D., & Liu, D. R. (2020). A bacterial cytidine deaminase toxin enables CRISPR-free mitochondrial base editing. *Nature*, *583*(7817), 631–637. <https://doi.org/10.1038/s41586-020-2477-4>
55. Moore, C. L., Papa, L. J., & Shoulders, M. D. (2018). A Processive Protein Chimera Introduces Mutations across Defined DNA Regions in Vivo. *Journal of the American Chemical Society*, *140*(37), 11560–11564. <https://doi.org/10.1021/jacs.8b04001>
56. Navaratnam, N., & Sarwar, R. (2006). An overview of cytidine deaminases. *International Journal of Hematology*, *83*(3), 195–200. <https://doi.org/10.1532/IJH97.06032>
57. Newville, M., Stensitzki, T., Allen, D. B., & Ingargiola, A. (2014). *LMFIT: Non-Linear Least-Square Minimization and Curve-Fitting for Python*. <https://doi.org/10.5281/ZENODO.11813>
58. Nosrati, G. R., & Houk, K. N. (2012). SABER: a computational method for identifying active sites for new reactions. *Protein Science : A Publication of the Protein Society*, *21*(5), 697–706. <https://doi.org/10.1002/PRO.2055>
59. Olechnovič, K., & Venclovas, Č. (2017). VoroMQA: Assessment of protein structure quality using interatomic contact areas. *Proteins*, *85*(6), 1131–1145. <https://doi.org/10.1002/PROT.25278>
60. Packer, M. S., & Liu, D. R. (2015). Methods for the directed evolution of proteins. *Nature Reviews Genetics*, *16*(7), 379–394. <https://doi.org/10.1038/nrg3927>
61. Pedroza-García, J. A., Nájera-Martínez, M., Mazubert, C., Aguilera-Alvarado, P., Drouin-Wahbi, J.,

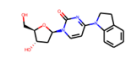
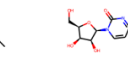
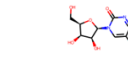
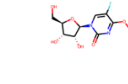
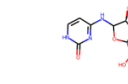
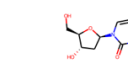
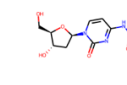
- Sánchez-Nieto, S., Gualberto, J. M., Raynaud, C., & Plasencia, J. (2019). Role of pyrimidine salvage pathway in the maintenance of organellar and nuclear genome integrity. *The Plant Journal*, *97*(3), 430–446. <https://doi.org/10.1111/TPJ.14128>
62. Porter, D. J. T., & Austin, E. A. (1993). Cytosine deaminase. The roles of divalent metal ions in catalysis. *Journal of Biological Chemistry*, *268*(32), 24005–24011. [https://doi.org/10.1016/S0021-9258\(20\)80485-2](https://doi.org/10.1016/S0021-9258(20)80485-2)
63. Prochnow, C., Bransteitter, R., Klein, M. G., Goodman, M. F., & Chen, X. S. (2007). The APOBEC-2 crystal structure and functional implications for the deaminase AID. *Nature*, *445*(7126), 447–451. <https://doi.org/10.1038/nature05492>
64. Reetz, M. T., Bocola, M., Carballeira, J. D., Zha, D., & Vogel, A. (2005). Expanding the Range of Substrate Acceptance of Enzymes: Combinatorial Active-Site Saturation Test. *Angewandte Chemie International Edition*, *44*(27), 4192–4196. <https://doi.org/10.1002/anie.200500767>
65. Rios, L. A. de S., Cloete, B., & Mowla, S. (2020). Activation-induced cytidine deaminase: in sickness and in health. *Journal of Cancer Research and Clinical Oncology*, *0123456789*. <https://doi.org/10.1007/s00432-020-03348-x>
66. Ruan, H., Qiu, S., Beard, B. C., & Black, M. E. (2016). Creation of zebularine-resistant human cytidine deaminase mutants to enhance the chemoprotection of hematopoietic stem cells. *Protein Engineering, Design and Selection*, *29*(12), 573–582. <https://doi.org/10.1093/PROTEIN/GZW012>
67. Salter, J. D., Bennett, R. P., & Smith, H. C. (2016). The APOBEC Protein Family: United by Structure, Divergent in Function. In *Trends in Biochemical Sciences* (Vol. 41, Issue 7, pp. 578–594). Elsevier Ltd. <https://doi.org/10.1016/j.tibs.2016.05.001>
68. Sánchez-Quitian, Zilpa A., Schneider, C. Z., Ducati, R. G., de Azevedo, W. F., Bloch, C., Basso, L. A., & Santos, D. S. (2010). Structural and functional analyses of *Mycobacterium tuberculosis* Rv3315c-encoded metal-dependent homotetrameric cytidine deaminase. *Journal of Structural Biology*, *169*(3), 413–423. <https://doi.org/10.1016/J.JSB.2009.12.019>
69. Sánchez-Quitian, Zilpa A., Timmers, L. F. S. M., Caceres, R. A., Rehm, J. G., Thompson, C. E., Basso, L. A., De Azevedo, W. F., & Santos, D. S. (2011). Crystal structure determination and dynamic studies of *Mycobacterium tuberculosis* Cytidine deaminase in complex with products. *Archives of Biochemistry and Biophysics*, *509*(1), 108–115. <https://doi.org/10.1016/j.abb.2011.01.022>
70. Sánchez-Quitian, Zilpa Adriana, Rodrigues-Junior, V., Rehm, J. G., Eichler, P., Barbosa Trivella, D. B., Bizarro, C. V., Basso, L. A., & Santos, D. S. (2015). Functional and structural evidence for the catalytic role played by glutamate-47 residue in the mode of action of *Mycobacterium tuberculosis* cytidine deaminase. *RSC Advances*, *5*(2), 830–840. <https://doi.org/10.1039/c4ra13748e>
71. Serdjebi, C., Milano, G., & Ciccolini, J. (2015). Role of cytidine deaminase in toxicity and efficacy of nucleosidic analogs. *Expert Opinion on Drug Metabolism and Toxicology*, *11*(5), 665–672. <https://doi.org/10.1517/17425255.2015.985648>
72. Shrake, A., & Rupley, J. A. (1973). Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *Journal of Molecular Biology*, *79*(2), 351–371. <https://doi.org/10.1016/0022->

73. Strokach, A., & Kim, P. M. (2022). Deep generative modeling for protein design. *Current Opinion in Structural Biology*, 72, 226–236. <https://doi.org/10.1016/J.SBI.2021.11.008>
74. Sunden, F., Alsadhan, I., Lyubimov, A., Doukov, T., Swan, J., & Herschlag, D. (2017). Differential catalytic promiscuity of the alkaline phosphatase superfamily bimetallo core reveals mechanistic features underlying enzyme evolution. *Journal of Biological Chemistry*, 292(51), 20960–20974. <https://doi.org/10.1074/JBC.M117.788240>
75. Teh, A. H., Kimura, M., Yamamoto, M., Tanaka, N., Yamaguchi, I., & Kumasaka, T. (2006). The 1.48 Å resolution crystal structure of the homotetrameric cytidine deaminase from mouse. *Biochemistry*, 45(25), 7825–7833. <https://doi.org/10.1021/bi060345f>
76. Timmers, L. F. S. M. E., Ducati, R. G., Sánchez-Quitian, Z. A., Basso, L. A., Santos, D. S., & De Azevedo, W. F. (2012). Combining molecular dynamics and docking simulations of the cytidine deaminase from Mycobacterium tuberculosis H37Rv. *Journal of Molecular Modeling*, 18(2), 467–479. <https://doi.org/10.1007/s00894-011-1045-0>
77. Trott, O., & Olson, A. J. (2009). AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, 31(2), NA-NA. <https://doi.org/10.1002/jcc.21334>
78. Urbelienė, N., Kutanovas, S., Meškienė, R., Gasparavičiūtė, R., Tauraitė, D., Koplūnaitė, M., & Meškys, R. (2019). Application of the uridine auxotrophic host and synthetic nucleosides for a rapid selection of hydrolases from metagenomic libraries. *Microbial Biotechnology*, 12(1), 148–160. <https://doi.org/10.1111/1751-7915.13316>
79. Urbelienė, N., Meškienė, R., Tiškus, M., Stanislauskienė, R., Aučynaitė, A., Laurynėnas, A., & Meškys, R. (2020). A Rapid Method for the Selection of Amidohydrolases from Metagenomic Libraries by Applying Synthetic Nucleosides and a Uridine Auxotrophic Host. *Catalysts*, 10(4), 445. <https://doi.org/10.3390/catal10040445>
80. Vaissier Welborn, V., & Head-Gordon, T. (2019). Computational Design of Synthetic Enzymes. *Chemical Reviews*, 119(11), 6613–6630. https://doi.org/10.1021/ACS.CHEMREV.8B00399/ASSET/IMAGES/MEDIUM/CR-2018-003996_0012.GIF
81. Vallejo, D., Nikoomanzar, A., & Chaput, J. C. (2020). Directed evolution of custom polymerases using droplet microfluidics. *Methods in Enzymology*, 644, 227–253. <https://doi.org/10.1016/BS.MIE.2020.04.056>
82. Vincenzetti, S., Quadrini, B., Mariani, P., De Sanctis, G., Cammertoni, N., Polzonetti, V., Pucciarelli, S., Natalini, P., & Vita, A. (2008). Modulation of human cytidine deaminase by specific aminoacids involved in the intersubunit interactions. *Proteins*, 70(1), 144–156. <https://doi.org/10.1002/PROT.21533>
83. Walko, C. M., & Lindley, C. (2005). Capecitabine: A review. *Clinical Therapeutics*, 27(1), 23–44. <https://doi.org/10.1016/j.clinthera.2005.01.005>
84. Wang, J., Guo, Q., Liu, L., & Wang, X. (2020). Crystal structure of Arabidopsis thaliana cytidine

- deaminase. *Biochemical and Biophysical Research Communications*, 529(3), 659–665.
<https://doi.org/10.1016/j.bbrc.2020.06.084>
85. Warner, D. F., Evans, J. C., & Mizrahi, V. (2014). Nucleotide Metabolism and DNA Replication. *Microbiology Spectrum*, 2(5). <https://doi.org/10.1128/MICROBIOLSPEC.MGM2-0001-2013>
86. Webb, B., & Sali, A. (2016). Comparative Protein Structure Modeling Using MODELLER. *Current Protocols in Bioinformatics*, 54, 5.6.1-5.6.37. <https://doi.org/10.1002/CPBI.3>
87. Xiang, S., Short, S. A., Wolfenden, R., & Carter, C. W. (1996). Cytidine deaminase complexed to 3-deazacytidine: A “valence buffer” in zinc enzyme catalysis. *Biochemistry*, 35(5), 1335–1341.
<https://doi.org/10.1021/bi9525583>
88. Xie, K., Sowden, M. P., Dance, G. S. C., Torelli, A. T., Smith, H. C., & Wedekind, J. E. (2004). The structure of a yeast RNA-editing deaminase provides insight into the fold and function of activation-induced deaminase and APOBEC-1. *Proceedings of the National Academy of Sciences of the United States of America*, 101(21), 8114–8119. <https://doi.org/10.1073/pnas.0400493101>
89. Yang, L., Briggs, A. W., Chew, W. L., Mali, P., Guell, M., Aach, J., Goodman, D. B., Cox, D., Kan, Y., Lesha, E., Soundararajan, V., Zhang, F., & Church, G. (2016). Engineering and optimising deaminase fusions for genome editing. *Nature Communications* 2016 7:1, 7(1), 1–12.
<https://doi.org/10.1038/ncomms13330>
90. Zhao, Y., Tian, J., Zheng, G., Chen, J., Sun, C., Yang, Z., Zimin, A. A., Jiang, W., Deng, Z., Wang, Z., & Lu, Y. (2019). Multiplex genome editing using a dCas9-cytidine deaminase fusion in *Streptomyces*. *Science China Life Sciences* 2019 63:7, 63(7), 1053–1062. <https://doi.org/10.1007/S11427-019-1559-Y>
91. Zimmermann, L., Stephens, A., Nam, S. Z., Rau, D., Kübler, J., Lozajic, M., Gabler, F., Söding, J., Lupas, A. N., & Alva, V. (2018). A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *Journal of Molecular Biology*, 430(15), 2237–2243.
<https://doi.org/10.1016/j.jmb.2017.12.007>

Supplementary information

Supplementary table 1. Substrate molecule 2D representations

					
N4-(2-benzoyl-benzoyl)-2-deoxycytidine	N4-(3-benzoyl-benzoyl)-2-deoxycytidine	N4-(4-benzoyl-benzoyl)-2-deoxycytidine	N4-acetyl-2-deoxy-5-O-DMT-cytidine	5-Levulinyl-N4-benzoyl-2-deoxycytidine	3-Levulinyl-N4-benzoyl-2-deoxycytidine
					
N4-methylcytidine	N4-N4-dimethylcytidine	N4-2-hydroxyethyl-2-deoxycytidine	N4-ethyl-2-deoxycytidine	4-(4-morpholinyl)-2-deoxycytidine	4-(4-morpholinyl)-5-fluoro-2-deoxycytidine
					
N4-hexyl-2-deoxycytidine	4-indol-2-deoxycytidine	N4-2,3,4,5,6-pentahydroxyhexyl-2-deoxycytidine	5-deoxy-5-fluoro-N4-(pentylthio)carbonylcytidine	4-Thio-methyl-uridine	4-Thio-ethyl-uridine
					
4-Thio-n-propyl-uridine	4-Thio-isopropyl-uridine	4-Thio-iso-butyl-uridine	4-Thio-benzyl-uridine	4-(1-thio-methyl-5-fluorouridine	4-Thio-ethyl-5-fluoro-uridine
					
4-Thio-benzyl-5-fluoro-uridine	4-Thio-phenyl-5-fluoro-uridine	4-methoxy-5-fluoro-uridine	4-butoxy-5-fluoro-uridine	4-benzyloxy-5-fluoro-uridine	
					
2-deoxycytidine	Cytidine	2,3-dideoxycytidine	2,5-dideoxycytidine	3-amino-2,3-dideoxycytidine	5-methyl-cytidine
					
2-O-methyl-Cytidine	5-Fluoro-cytidine	isocytidine	pseudocytidine	5-propynyl-2-deoxycytidine	2-deoxy-5-hydroxymethylcytidine
					
5-hydroxymethylcytidine	2-deoxy-5-methylcytidine	2-deoxy-5-hydroxycytidine	2-thio-cytidine	Cytosine	D-arabinofuranoside
					
N4-benzoylcytidine	N4-Benzoyl-5-methylcytidine	N4-acetyl-2-deoxycytidine	N4-isobutyl-2-deoxycytidine	N4-hexanoyl-2-deoxycytidine	N4-benzoyl-2-deoxycytidine
					
N4-ricotinyol-2-deoxycytidine	N4-picolonyl-2-deoxycytidine	N4-isonicotonyl-2-deoxycytidine	N4-(3-acetyl-benzoyl)-2-deoxycytidine	N4-(4-acetyl-benzoyl)-2-deoxycytidine	

Supplementary table 2. Tested CDA catalytic activities. 2 – active, 1- weakly active (activity observable after a 24h period), 0 – inactive, Blank – not tested.

	CDA_F14	F14 del127-130	F14 del83-85	F14 F126A	F14 F126W	F14 T51G	F14 G85I	F14 R56L	F14 G81L	F14 G81L G85I	F14 SML	F14 HSI	F14 QQS	F14 HSSG	F14 CLYR	F14 C53H R56Q	F14 C88H C91H	F14 HQHH	CDA_Pco	CDA_Pco G70T	CDA_Pco I108A	CDA_Ppo	CDA_Ppo CS0T	CDA_Ppo V82L	CDA_Tar	CDA_Tar I85A	CDA_Lsp	CDA_Lsp A82I
2'-dC	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	0	0	2	2	2	2	2	2	2	2	2	2
Cytidine	2										2	2	2	2	2				2			2			2		2	2
2',3'-ddC	1	0	0	0	0	0	0				0	0	0	0	0				0	0	0	1	2	0	2	2	0	0
2',5'-ddC	2	2	0	2	2	2	2				2	2	2	2	2				2	0	2	2	2	2	2	2	2	2
3'-amino 2',3'-ddC	1			0	1	1													1			1			1		1	
5-methyl-C	2																		2			2			2		2	2
2'-O-methyl-C	0			0	0	0					0	0	0	0	0				2			2			2		0	
5-F-cytidine	2																		2			2			2		2	2
isocytidine	0																		0			0			0		0	0
Pseudoisocytidine	1																		2			2			2		2	
5-propynyl-2'-dC	2																		2			2			2		2	
5-hydroxymethyl-2'-dC	2																		1			1			2		2	
5-hydroxymethyl-C	2																		2			2			1		1	
5-methyl-2'-dC	2																		2			2			2		2	2
5-hydroxy-2'-dC	2																		2			2			2		2	2
2-thio-cytidine	0																		0			1			1		0	0
Cytosine β-D-arabinofuranoside	2																		2	2	2	2			2	2	2	0
2'-deoxy-L-cytidine	2																		2			2			2		2	2
N ⁴ -acetyl-C	2																		0			1			1		2	
N ⁴ -benzoyl-C	2																		0			0			0		2	
N ⁴ -Benzoyl-5-methyl-C	2																		0			0			2		2	
N ⁴ -acetyl-2'-dC	2	2	0	2	2	2	2				2	2	2	2	2				0	0	0	2	2	0	2	2	2	2
N ⁴ -isobutyryl-2'-dC	2																		0			1			0		2	
N ⁴ -hexanoyl-2'-dC	2																		0			1			1		2	2
N ⁴ -benzoyl-2'-dC	2	2	0	2	2	1	2	2	2	2	2	2	2	2	2	0	0	0	0	0	0	2	2	2	2	2	2	2
N ⁴ -nicotinoyl-2'-dC	2																		0			2			2		2	0
N ⁴ -(3-acetyl-benzoyl)-2'-dC	2	2	0	0	2	0	0												0	0	0	0	0	0	0		2	0
N ⁴ -(4-acetyl-benzoyl)-2'-dC	2	2	0	2	2	0	0				2	2	2	2	2				0	0	0	0	0	0	0	0	2	2
N ⁴ -(2-benzoyl-benzoyl)-2'-dC	0	0	0	0	0	0	0				0	0	0	0	0				0	0	0	0	0	0	0	0	0	0
N ⁴ -(3-benzoyl-benzoyl)-2'-dC	2																		0			0			0		2	
N ⁴ -(4-benzoyl-benzoyl)-2'-dC	0	0	0	0	0	0	0												0	0	0	0	0	0	0	0	2	0
N ⁴ -acetyl-2'-deoxy-5'-O-DMT-C	2																		0	0	0				0			
5'-Levulinyl-N ⁴ -benzoyl-2'-dC	0																		0			0			0		0	0
3'-Levulinyl-N ⁴ -benzoyl-2'-dC	0																		0			0			0		0	0
3'-acetyl-N ⁴ -benzoyl-2'-dC	0																		0			0			0		0	0
3'-azido-N ⁴ -benzoyl-2',3'-ddC	0										0	0	0	0	0				0			0			0		0	0
4-thio-methyl-U	2																		0	2	2	1	2	2	1	2	2	1
4-thio-ethyl-U	2																		0			0			0		2	2
4-thio-benzyl-U	2	2	1			1	2												0			0			0		2	1
5-F-4-thio-methyl-U	2																		0			0			0		2	0
5-F-4-thio-ethyl-U	2																		0			0			0		2	0
5-F-4-thio-benzyl-U	2																		0			0			0		2	0
5-F-4-thio-phenyl-U	2																		0			0			0			0
5-F-4-methoxy-U	2																		0			2			0		2	0
5-F-4-butoxy-U	2																		0			1			1		2	1

Supplementary table 2. Tested CDA catalytic activities. 2 – active, 1- weakly active (activity observable after a 24h period), 0 – inactive, Blank – not tested.

	CDA_F14	F14 del127-130	F14 del83-85	F14 F126A	F14 F126W	F14 T51G	F14 G85I	F14 R56L	F14 G81L	F14 G81L G85I	F14 SML	F14 HSI	F14 QQS	F14 HSSG	F14 CLYR	F14 C53H R56Q	F14 C88H C91H	F14 HQHH	CDA_Pco	CDA_Pco G70T	CDA_Pco I108A	CDA_Ppo	CDA_Ppo CS0T	CDA_Ppo V82L	CDA_Tar	CDA_Tar I85A	CDA_Lsp	CDA_Lsp A82I
5-F-4-benzoyloxy-U	2																		0			1			1		2	1
N ⁴ -methylcytidine	2	2	2	2	2	2	2												2	0	2	2	2	2	2	2	2	2
N ² -N ⁴ -dimethylcytidine	2																		0			2			0		2	
N ⁴ -2-hydroxyethyl-2'-dC	2																		0			2			2		2	
N ⁴ -aminoethyl-2'-dC	2																		0			0			2		2	
4-(4-morpholinyl)-2'-dU	2	2	0	0	2	0	0												0	0	0	2	0	0	0		2	0
5-F-4-(4-morpholinyl)-2'-dU	2	2	0	2	2	2	2												0	0	0	2	2	2	0		2	2
N ⁴ -hexyl-2'-dC	2																		0			2			0		2	
N ⁴ -(indolin-1-yl)-2'-dC	0	0	0	0	0	0	0												0	0	0	0	0	0	0		0	0
N ⁴ -(2,3,4,5,6-pentahydroxyhexyl)-2'-dC	2	0	0	0	0	0	0												0	0	0	0	0	0	0		2	0
N ⁴ -(1H-indol-6-yl)methyl-2'-dC	2	2	0	2	0	2	2												0	0	0	0	0	0	0		2	2
Capecitabine	2	0	0	0	0	0	0				0	0	0	0	0				0	0	0	0	0	0	0		2	0