

Uporaba AI alata AlphaFolda 2 za modeliranje strukture proteina

Dilber, Ivana

Master's thesis / Diplomski rad

2023

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Rijeka / Sveučilište u Rijeci**

Permanent link / Trajna poveznica: <https://urn.nsk.hr/urn:nbn:hr:193:077207>

Rights / Prava: [In copyright](#) / [Zaštićeno autorskim pravom.](#)

Download date / Datum preuzimanja: **2024-05-12**

Repository / Repozitorij:



[Repository of the University of Rijeka, Faculty of Biotechnology and Drug Development - BIOTECHRI Repository](#)



University of Rijeka
Department of Biotechnology
University Graduate Programme
'Biotechnology in medicine'

Ivana Dilber

The use of the AI tool AlphaFold 2 for protein structure modeling

Master's thesis

Rijeka, 2023.

University of Rijeka
Department of Biotechnology
University Graduate Programme
'Biotechnology in medicine'

Ivana Dilber

The use of the AI tool AlphaFold 2 for protein structure modeling

Master's thesis

Rijeka, 2023

Mentor: dr. sc. Višnja Stepanić

Co-mentor: assoc. prof. dr. sc. Nela Malatesti

Sveučilište u Rijeci
Odjel za biotehnologiju
Diplomski sveučilišni studij
„Biotehnologija u medicini“

Ivana Dilber

Uporaba AI alata AlphaFolda 2 za modeliranje strukture proteina

Diplomski rad

Rijeka, 2023.

Mentor: dr. sc. Višnja Stepanić

Komentor: izv. prof. dr. sc. Nela Malatesti

Mentor: dr. sc. Višnja Stepanić

Co-mentor: assoc. prof. dr. sc. Nela Malatesti

The thesis was defended on October 26, 2023 before the committee:

1. doc. dr. sc. Željka Maglica – president of the committee
2. doc. dr. sc. Ivan Gudelj – member of the committee
3. dr. sc. Višnja Stepanić – mentor

This thesis has 42 pages, 22 figures, and 33 references.

Zahvala

Iznimno sam zahvalna svojoj mentorici, dr. sc. Višnji Stepanić, na svim dobronamjernim savjetima, razumijevanju, strpljenju i prenesenom znanju. Hvala Vam na danoj prilici i nesebičnosti koju ste mi pružili.

Također se želim zahvaliti i svojoj komentorici, izv. prof. dr. sc. Neli Malatesti, kao i sestri Moniki te ostatku obitelji, dečku Dinu i prijateljicama, posebno Mariji, Marti i Doroteji koje su bile uz mene u najboljim, ali i onim manje dobrim trenucima studiranja i pružili mi motivaciju i podršku za daljnji napredak.

Hvala i mojim čupavim prijateljicama Niki i Zoey, koje su mi pravile društvo i bile najbolje cimerice koje sam mogla poželjeti.

Abstract

Proteins are biological macromolecules composed of amino acids linked by peptide bonds. Their three-dimensional (3D) structures are still challenging to determine and the number of proteins with resolved tertiary structures is rather small compared to the number of known protein sequences. The 3D structures of proteins are essential for understanding their function, and thus biological processes orchestrating health and diseases. The 3D protein structure allows us to identify "binding pockets" and functionally relevant regions of the protein. Nowadays innovative approaches have been developed for fast determination of protein conformations. These include computer algorithms that predict the 3D structure of the protein from its polypeptide primary sequence. In this thesis, we use AlphaFold 2, an open-source software that uses available protein datasets and artificial intelligence (AI), to predict the 3D structure of proteins. In this study, AlphaFold 2 structure models were analyzed for randomly generated amino acid sequences and for well-known industrial biocatalysts halohydrin dehalogenases HheC and HheA. The random sequences were generated by the tool RandSeq, while the FASTA inputs for HheA and HheC were formed from the crystal structures 1ZMO and 1ZMT, respectively, downloaded from the Protein Data Bank. The AlphaFold 2 conformations were analyzed using PyMOL and ChimeraX visualization software. While AlphaFold 2 could not reliably predict the structures of random sequences, as expected, the structures of the enzymes HheA and HheC in their monomeric and tetrameric states were predicted with high reliability. However, structural peculiarity like the entry of the C-terminal tail into the diagonal subunit of the HheC tetramer was not predicted. This study shows that AlphaFold 2 structures can be good starting conformations for molecular dynamics simulations while their use for molecular docking calculations should be taken with caution.

Keywords: AlphaFold 2, 3D protein structure, haloalcohol/halohydrin dehalogenase, HheA, HheC

Sažetak

Proteini su biološke makromolekule sastavljene od aminokiselina povezanih peptidnim vezama. Njihove trodimenzionalne (3D) strukture još uvijek je teško odrediti, a broj proteina s rješanim tercijarnim strukturama prilično je malen u usporedbi s brojem poznatih proteinskih sekvenci. 3D strukture proteina bitne su za razumijevanje njihove funkcije, a time i bioloških procesa koji upravljaju zdravljem i bolestima. 3D proteinska struktura omogućuje nam identificiranje "veznih džepova" i funkcionalno relevantnih regija proteina. Danas se razvijaju inovativni pristupi za brzo određivanje konformacija proteina, a to uključuje i računalne algoritme koji predviđaju 3D strukturu proteina iz primarne sekvence polipeptida. U ovom diplomskom radu koristimo AlphaFold 2, softver otvorenog koda koji koristi dostupne skupove podataka o proteinima i umjetnu inteligenciju (AI) za predviđanje 3D strukture proteina. U ovoj studiji pomoću AlphaFolda 2 predviđene se strukture za nasumično generirane sekvence aminokiselina i za dobro poznate industrijske biokatalizatore halohidrin dehalogenaza HheC i HheA. Nasumične sekvence generirao je alat RandSeq, dok su FASTA ulazi za HheA i HheC formirani iz kristalnih struktura 1ZMO, odnosno 1ZMT, preuzetih iz baze Protein Data Bank (PDB). Predviđene konformacije analizirane su pomoću softvera za vizualizaciju PyMOL i ChimeraX. Iako AlphaFold 2 nije mogao pouzdano predvidjeti strukture nasumičnih sekvenci, kao što se i očekivalo, strukture enzima HheA i HheC u njihovim monomernim i tetramernim stanjima predviđene su s visokom pouzdanošću. Međutim AlphaFold 2 nije predvidio strukturnu osobitost ulaska C-terminalnog repa u dijagonalnu podjedinicu tetramera HheC. Ova studija pokazuje da strukture predviđene AlphaFoldom 2 mogu biti dobre početne konformacije za simulacije molekulske dinamike, dok njihovu upotrebu za izračune molekuskog uklapanja treba uzeti s oprezom.

Ključne riječi: AlphaFold 2, 3D struktura proteina, haloalkohol/halohidrin dehalogenaze, HheA, HheC

Table of Contents

1. Introduction.....	1
1. 1. The importance of three-dimensional structures of proteins.....	1
1. 2. 3D protein structure prediction by AlphaFold 2.....	2
1. 2. 1. CASP organization and introduction to AlphaFold2 – How it all began...	2
1. 2. 2. Architectural modules of AlphaFold 2.....	3
1. 2. 3. Limitations of AlphaFold 2	5
1. 3. Protein structure similarity metrics.....	6
1. 4. Enzymes and their importance	7
2. Aims of Thesis.....	10
3. Materials and Methods.....	11
4. Results	16
4. 1. AlphaFold 2 predicts 3D protein structures poorly from randomly generated amino acid sequences	16
4. 2. Amino acid substitutions do not significantly affect the conformations of monomeric subunits of the halohydrin dehalogenase enzymes HheA and HheC .	18
4. 2. 1. AlphaFold 2 predictions for the HheA monomer.....	18
4. 2. 2. AlphaFold 2 predictions for the HheC monomer.....	24
4. 3. AlphaFold 2 successfully predicts symmetric homotetrameric conformations	28
5. Discussion	31
6. Conclusion.....	37
7. References	39
8. Curriculum Vitae.....	43

1. Introduction

1. 1. The importance of three-dimensional structures of proteins

In 1838 Jacob Berzelius, a Swedish chemist, proposed the name protein for the material produced by plants as a food for animals. (1) Nowadays it is known that there are numerous proteins with a variety of functions such as cellular structural support, immune protection, reaction catalysis, transduction of cell signals, participating in DNA transcription, etc. The biological function of proteins depends on their tertiary and quaternary structures which result from the folding of polypeptide sequences and their mutual assembling, respectively. The three-dimensional structures (3D) of proteins are essential for understanding their function, and thus biological processes orchestrating health and diseases. The representation of the 3D protein structure allows us to identify "drug binding pockets" and functionally relevant regions of the protein. (2) 3D tertiary structures of proteins are experimentally determined by methods such as X-ray crystallography, nuclear magnetic resonance (NMR), and cryo-electron microscopy. However, these methods are complex, expensive, and time-consuming, and the resulting protein structure(s) represents one or a small number of protein conformations. The number of proteins with experimentally determined tertiary structures is rather small compared to the number of known protein sequences (210,180 PDB structures compared to 251,600,768 UniProt entries). Nowadays innovative approaches have been developed for fast determination of protein conformations. Such are computational models that predict the 3D structure of a protein from its polypeptide primary sequence. (3)

1. 2. 3D protein structure prediction by AlphaFold 2

1. 2. 1. CASP organization and introduction to AlphaFold2 – How it all began

In 2018, Google's start-up, DeepMind presented an open source software AlphaFold at the Critical Assessment of Structure Prediction (CASP) competition of prediction structures of proteins from their primary sequences, that is solving the problem of protein folding. The CASP is an organization whose goal is to find a solution to the problem of 3D protein structure based on amino acid sequence. This competition has three categories: TBM (template-based modeling), FM (free modeling), and FM/TBM. (4,5) Contestants must submit their 3D models for proteins whose experimental structures have not yet been published, and they are analysed by independent reviewers. All models and structural analyses are publicly available. (4,6)

AlphaFold is a deep learning (DL) code based on a convolutional neural network (NN) which, from the target protein's amino acid sequence and multiple sequence alignment (MSA) features and statistics, outputs the protein structure in the form of a distogram (histogram showing inter-residue distances). AlphaFold also included following multiple Gradient Descent optimizations to find the structure that corresponds to a minimum at a potential energy surface. The basis of the AlphaFold approach is the observation that residues that are in spatial contact tend to show patterns of correlated mutations. (7) Although proteins mutate and evolve, the structures of related proteins remain relatively similar.

In 2020, the next version of AlphaFold, AlphaFold 2, was presented at the next CASP14 competition. This artificial intelligence (AI) system won the CASP14 competition by predicting 3D protein structures from amino acid sequences with accuracy at the atomic (that is experimental) level (8) and with much greater precision than any of the other competing methods.

AlphaFold 2 is significantly different from its previous version; it uses amino acids as an input sequence to construct an MSA based on several protein sequence databases, to determine which parts of the sequence are prone to mutation and find the correlation between those parts. It also identifies proteins of similar structure that are used to build an initial representation of the target sequence (pair representation). (9)

1. 2. 2. Architectural modules of AlphaFold 2

The AlphaFold 2 code is divided into three main modules: Search and Embedding, Evolutionary transformer (Evoformer), and Structural module (Figure 1). (10)

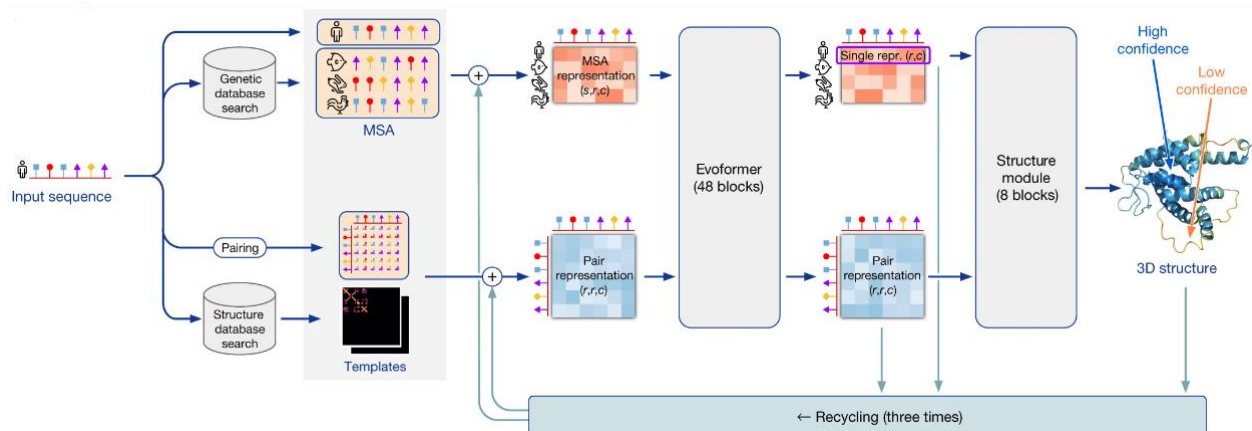


Figure 1. A simplified architecture of the AlphaFold 2 model. The flow of information between 3 modules is described by arrows; s - number of sequences, r - number of residues, c - number of channels. (Figure taken from (10))

The Search and Embedding is an input module. Based on the amino acid sequence of a target protein in the sequence databases, it finds (evolutionary) similar sequences through MSA and available 3D structures/templates using various databases (Uniref90, Uniclust30,

Mgnify, Big Fantastic Database, PDB70...). This is followed by the generation of the MSA representation and the pair representation and by embedding evolutionary, physical and geometric constraints of protein structures into the architecture.

Evoformer is the second AlphaFold 2 module, the so-called coder that contains 48 NN blocks, each of which has two inputs –a processed MSA and an $N_{res} \times N_{res}$ array that represents residue pairs. Evoformer blocks have an attention-based architecture. (10) In comparison to the convolutional neural network (CNN), an advantage of Evoformer blocks lies in the fact that information can be exchanged, i.e. transferred among the MSA display and the 3D template display, meaning that the improvement of the MSA estimation in return physically correctly modifies the protein structures represented by the templates by taking directly the spatial and evolutionary relationships.

The Structural module is a decoder converting the processed MSA representation and pairwise representation into 3D coordinates of all atoms in a protein. It takes each amino acid residue as a separate object and predicts rotations and translations to represent the 3D structure of the protein. This decoder has two input elements; the first contains a linear projection of the first order of the MSA representation, while the second is the output of the pair from Evoformer where each residue is represented as a triangle whose vertex next to the obtuse angle indicates the Ca atom, while the other two peaks indicate the N-atom of the amino group and the C-atom of the carbonic acid. In the beginning, all frames are placed at the same point in the same orientation, and the output of the structural module is the 3D coordinates of all the protein atoms obtained by allowing simultaneous local refinement of all parts of the structure, which is done by the novel equivariant transformer in an iterative way using the whole network. This is termed 'recycling' and is related to approaches in computer

vision. Recycling steps are repeated three times to make the result as accurate as possible. (8)

In 2021, the publicly accessible AlphaFold Protein Structure Database (<https://alphafold.ebi.ac.uk>) was created with more than 360,000 predicted structures from 21 organism proteomes. Today, this database has more than 200 million entries from human and 47 other organisms' proteomes, the predicted structure of which is available to everyone for free, with atomic coordinates in PDB format and Predicted Aligned Error (PAE, about which we will talk later) in JSON format. AlphaFold 2 reduced the number of human proteins without structural coverage to 29 from 5027. The publicly available database of predicted protein structures opens the possibility of easier selection of preclinical models based on the similarity of proteins of different species to human proteins. (11)

1. 2. 3. Limitations of AlphaFold 2

Despite its revolutionary efficiency in predicting the 3D structure of proteins, AlphaFold 2 has limitations, and many questions remain to be resolved. AlphaFold 2, for example, has a harder time predicting intrinsically disordered regions and loops of proteins, which are of great importance for drug design because they are located on the protein surface and are easily accessible to solvents and other proteins. Stevens and He (12) showed that AlphaFold 2 can only predict shorter loops (<20 amino acids) with high accuracy, while Azzaz et al. (13) showed that predicting the structure of membrane proteins with AlphaFold 2 is not reliable due to inconsistencies in the position of transmembrane domains. AlphaFold 2 also cannot predict structures with ligands, complexes with DNA or RNA, or post-translational modifications such as methylation, phosphorylation, or glycosylation. (14) Since its algorithm is based on MSAs and requires

databases of known 3D structures (such as PDB), AlphaFold 2 cannot predict new structures and the use of evolutionary information from larger MSAs requires powerful computer processors, and their structure prediction takes a lot of time as the protein length increases. (5)

1. 3. Protein structure similarity metrics

Various metrics were defined and used to evaluate the predicted protein conformations. The most commonly used metrics for protein structure similarity are the Root-Mean-Square Deviation of atomic positions (RMSD), the Global Distance Test (GDT), and the Local Distance Difference Test (LDDT).

The simplest evaluation parameter is RMSD, usually expressed in angstroms (\AA), which calculates the average deviation between aligned protein structures based on the positions of Ca atoms of residues. However, some of its characteristics, such as insensitivity to missing parts of the model and the dominance of outliers in poorly predicted regions, considerably limit its usefulness for evaluating the quality of structural predictions. (15)

To mitigate the sensitivity of RMSD to regions that deviate significantly between two aligned structures, the dimensionless Global Distance Test (GDT) was introduced. GDT is also calculated over the backbone Ca atoms, but it is calculated with the percentage of Ca atoms that are found within certain cutoff distances of each other. It is expressed as a percentage from 0 % (a meaningless prediction) to 100 % (a perfect prediction), that is structures are more similar with the higher GDT percentage. (16) One of the advantages of GDT is that strongly deviating atoms do not significantly affect the result. The CASP competition uses the Global Distance Test Total Score (GDT_TS), with cut-off distances of 1 \AA , 2 \AA , 4 \AA , and 8 \AA , and whose value can be calculated

via the free online Local Global Alignment tool (LGA), <http://linum.proteinmodel.org/AS2TS/LGA/lga.html>. (17)

To overcome the limitations of RMSD, in addition to GDT, LDDT metric was developed. LDDT provides a superposition-free result and estimates the local distance differences of all atoms in the model including side chain atoms. The result is the average of four fractions calculated using thresholds of 0.5 Å, 1 Å, 2 Å, and 4 Å, expressed as a percentage (from 0 to 100 percent). (15)

1. 4. Enzymes and their importance

Enzymes are biocatalysts that increase the reaction rate and accelerate the conversion of substrates into products, by reducing the activation energy of the reaction. They are divided into oxidoreductases, transferases, hydrolases, lyases, isomerases, and ligases (synthetases). (18) Lyases are enzymes that catalyse the breaking of chemical bonds (known as elimination reaction) between carbon atoms and between carbon and oxygen, sulphur or nitrogen, and creating a new double bond or ring structure. (19) They are found in cellular processes such as the citric acid cycle, and the greatest application of lyases is a production of L-DOPA, a drug used to treat Parkinson's disease. (20)

Haloalcohol dehalogenases, halohydrin hydrogen-halide lyases, or halohydrin dehalogenases are enzymes isolated from bacteria that can grow on vicinal haloalcohols. They use the Ser-Tyr-Arg catalytic triad to deprotonate the haloalcohol oxygen which then attacks the halogen-bearing carbon atom, producing an epoxide and halide ion (Figure 2). (21) It is believed that arginine activates tyrosine which, as a catalytic base, takes a proton from the halohydrin substrate. The residue serine most likely binds the substrate while the adjacent carbon approaches the alcohol oxygen of the substrate, releasing a halogen ion. (22) Currently, six different haloalcohol dehalogenases have

been isolated, which, based on the similarity of the amino acid sequence, are grouped into three subtypes – A, B, or C. From the sequence and structural similarity, it was assumed that the A and C type enzymes have common ancestors, while type B enzymes have different precursors. (23) Their sequences show that they are all evolutionarily related to NAD(P)(H)-dependent short-chain dehydrogenase/reductase (SDR).

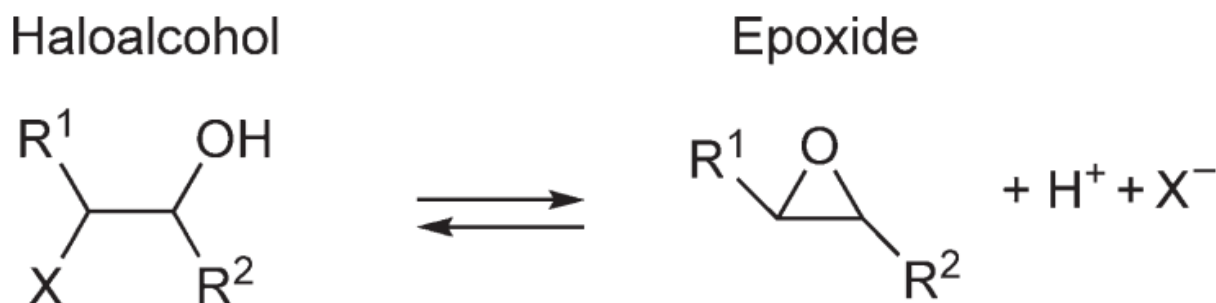


Figure 2. Haloalcohol/halohydrin dehalogenases catalyse alcohol-ketone conversions. Deprotonation of the haloalcohol oxygen-induced attack to the halogen-bearing carbon atom, producing an epoxide and halide ion. (*Figure taken from (22)*)

Haloalcohol dehalogenases can be found in both Gram-negative and Gram-positive bacteria. (23) In the Protein Data Bank (PDB) under PDB ID case 1ZMO is the crystal structure of haloalcohol dehalogenase of A type, HheA, isolated from *Arthrobacter* sp. AD2. It is a tetramer consisting of four identical subunits. (Figure 3A) The structure that has 35% sequence identity with the 1ZMO structure is in the PDB under the number 1ZMT. (Figure 3B) 1ZMT is the crystal structure of C-type haloalcohol dehalogenase, HheC, isolated from *Agrobacterium radiobacter* AD1. (21,24) This enzyme catalyses the dehalogenation of aliphatic and aromatic vicinal haloalcohols such as 1,3-dichloro-2-propanol and 2-phenyl-1-chloro-2-ethanol, thereby producing HCl and corresponding epoxides.

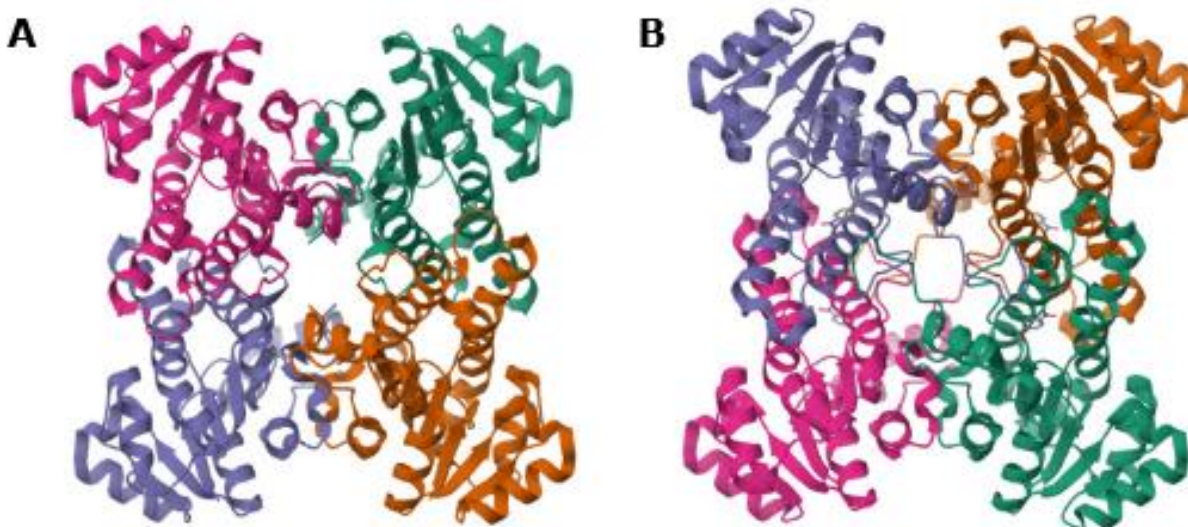


Figure 3. Haloalcohol dehalogenase structures in PDB. (A) 1ZMO - crystal structure of haloalcohol dehalogenase A type, HheA and (B) 1ZMT - crystal structure of C-type haloalcohol dehalogenase, HheC. (*Figures taken from (21,24)*)

These biocatalysts intrigued us to examine the accuracy and precision of structures predicted by AlphaFold 2, with using the 1ZMO and 1ZMT PDB structures as references.

2. Aims of Thesis

AlphaFold 2 is an AI program for the prediction of 3D coordinates of proteins starting from their primary amino acid sequences. Knowing the 3D structure of a protein is essential for finding its active sites and functionally relevant regions/hot spots.

The first goal of the research study performed is to test the ability of AlphaFold 2 to predict protein structure based on a randomized amino acid sequence. The hypothesis is that AlphaFold 2 will fail to predict such structures, given that there are no related sequences to them.

The second goal is to verify the precision and accuracy of AlphaFold 2 in predicting the 3D structures of the (hypothetical) monomers of haloalcohol dehalogenases HheA and HheC, and their mutants. The hypothesis is that AlphaFold 2 will successfully predict the 3D structures of these monomers and their mutants since there are available crystal structures for HheA and HheC as well as other halohydrin dehalogenases in the PDB database, which then facilitates the use of these molecules for research purposes.

The third goal of the performed study is to verify the precision and accuracy of AlphaFold 2 in predicting the 3D structures of tetramers of haloalcohol dehalogenases HheA and HheC. The hypothesis is that AlphaFold 2 will have difficulty predicting the 3D structures of these tetramers, which means that some other method will have to be used or invented to reveal their 3D structure.

3. Materials and Methods

In this study AlphaFold 2 version 2.1.1 was applied. AlphaFold 2 was installed on the Isabella computer cluster. Isabella is a computer cluster of the University Computing Centre (Srce) of the University of Zagreb, which was created in 2002 and has been enabled for non-profit computations. (25) Isabella belongs to the High-Performance Computing (HPC) group of clusters, i.e., clusters with high efficiency, and is intended for large parallel computations that require large amounts of processor cores, graphics processors, and working memory. (26) The user accesses Isabella using the Secure Shell (SSH) protocol, through the `teran.srce.hr` access server. (27) In order to access SSH, it is necessary to install PuTTY. Putty is an open-source software, developed by Simon Tatham. (28)

Input for AlphaFold 2 is a primary sequence in the FASTA format. The FASTA sequences of amino acids for the structures 1ZMO and 1ZMT were downloaded from the RCSB PDB database. These are X-ray structures of wild-type (WT) haloalcohol dehalogenases HheA and HheC, respectively. The resolution of 1ZMO crystal structure is 2.00 Å, while the resolution of 1ZMT structure is slightly better, at 1.70 Å. 1ZMT has a longer amino acid sequence (254) than 1ZMO (244). (21,24) The FASTA representations for the 1ZMO and 1ZMT mutants were also created, as shown in Figures 4 and 5, respectively. The 1ZMO mutant has an amino acid change at position 178, where the amino acid asparagine is replaced by the amino acid alanine (N178A). The 1ZMT mutant is a quadruple mutant, which means that the substitutions occurred at four positions; at position 84, proline was replaced by valine (P84V), at position 86, phenylalanine was replaced by proline (F86P), and threonine was replaced by alanine at position 134 (T134A) and asparagine by alanine at position 176 (N176A).

A >1ZMO_1|Chains A, B, C, D, E, F, G, H|halohydrin dehalogenase|Arthrobacter sp. AD2 (168809)
 MVIALVTHARHFAGPAAVEALTQDGYTVVCHDASFADAAERQRFESENPGTIALAEQKPERLVDATLQHGEAIDTIVSNDYIPRPMNRLPLE
 GTSEADIRQMFEALSIFPILLQLQSAIAPLRAAGGASVIFITSSVGKKPLAYNPLYGPARAATVALVESAAKTLSRDGILLYAIGNPFFNNPT
 YFPTSDWENNPELRRERVDVPLGRLGRPDGMGALITFLASRRAPIVGQFFAFTGGYLP

B >1ZMO_1|Chains A, B, C, D, E, F, G, H|halohydrin dehalogenase|Arthrobacter sp. AD2 (168809)
 MVIALVTHARHFAGPAAVEALTQDGYTVVCHDASFADAAERQRFESENPGTIALAEQKPERLVDATLQHGEAIDTIVSNDYIPRPMNRLPLE
 GTSEADIRQMFEALSIFPILLQLQSAIAPLRAAGGASVIFITSSVGKKPLAYNPLYGPARAATVALVESAAKTLSRDGILLYAIGPAFFNNPT
 YFPTSDWENNPELRRERVDVPLGRLGRPDGMGALITFLASRRAPIVGQFFAFTGGYLP




Figure 4. FASTA formats of the PDB structure 1ZMO and the N178A 1ZMO mutant.

The mutated position is marked with a red arrow.

A >1ZMT_1|Chains A, B, C, D|Haloalcohol dehalogenase HheC|Agrobacterium tumefaciens (358)
 MSTAIVTNVKHFGGMGSALRLSEAGHTVACHDESFKQKDELEAFAETYPQLKPMSEQEPAELIEAVTSAYGQVDVLVSNDIFAPEFQPIDKY
 AVEDYRGAVEALQIRPFALVNAVASQMKKRKSGHIIFITSATPFGPWKELSTYTSARAGACTLANALSKELGEYNIPVFAIGPNTLHSEDSP
 YFYPTPEWKTNPEHVAHVKKVTALQRLGTQKELGELVAFLASGSCDYLTGQVFWLAGGFPMIERWPGMPE

B >1ZMT_1|Chains A, B, C, D|Haloalcohol dehalogenase HheC|Agrobacterium tumefaciens (358)
 MSTAIVTNVKHFGGMGSALRLSEAGHTVACHDESFKQKDELEAFAETYPQLKPMSEQEPAELIEAVTSAYGQVDVLVSNDIFAVEPQPIDKY
 AVEDYRGAVEALQIRPFALVNAVASQMKKRKSGHIIFITSAAPFGPWKELSTYTSARAGACTLANALSKELGEYNIPVFAIGPNTLHSEDSP
 YFYPTPEWKTNPEHVAHVKKVTALQRLGTQKELGELVAFLASGSCDYLTGQVFWLAGGFPMIERWPGMPE

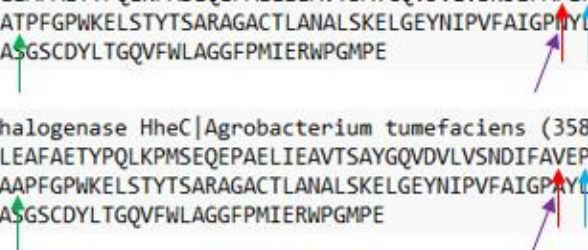


Figure 5. FASTA formats of the structures 1ZMT and corresponding quadrupole mutant. Mutated positions are marked with arrows; red - P84V, blue - F86P, green – T134A and purple - N176A.

After creating the FASTA formats of studied WT dehalogenases and their mutants, a script for running AlphaFold 2 was formed by following the instructions that can be found on the website <https://wiki.srce.hr/display/RKI/Alphafold2>. An example of my script can be seen in Figure 6. It is important that the script name ends with the extension .sge, otherwise, AlphaFold 2 will not give results.

```

#$ -N alphafold2
#$ -pe gpu 1
#$ -l cores=5
#$ -cwd

module load bioinfo/alphafold2/2.1.1

GPUDEVICE=$(cat $TMPDIR/gpu)

run_alphafold.sh -d $ALPHAFOLDDB -o /home/idilber/Tetrameri/1ZMO/Monomer -m monomer -f
/home/idilber/Tetrameri/1ZMO/Monomer/1ZMOMonomer.fasta -t 2022-06-01 -a $GPUDEVICE -n 5

```

Figure 6. Alphafold.sge script used for running *in silico* experiment on the cluster Isabella.

AlphaFold 2 starts running, and the time required to display the results depends on the length of the amino acid sequence, the stages of search and preprocessing of the input, relaxation, and locking of the structure. (10) There is another script, ParallelFold for running Alphafold 2 jobs on Isabella. Parallelfold is a modified script of Alphafold.sge and is used for parsing the CPU (MSA and template searching) and GPU (model prediction) parts. (29) First, it is necessary to run a script for the CPU part, and then for the GPU part (Figures 7A, 7B). This approach enables enhanced visualization of AlphaFold 2 quality/confidence metrics by using a relevant script for visualization, to see graphs for sequence coverage, predicted LDDT (pLDDT) per position, and PAE (Figure 8).

In the case of predictions of structures of tetramers of halohydrin dehalogenases HheA and HheC, in the cpu.sge and gpu.sge scripts (Figure 7), instead of „monomer“ the command „multimer“ is used.

<pre>#!/bin/bash #\$ -N parafold2 #\$ -pe *mpisingle 8 #\$ -cwd module load bioinfo/parallelfold/2.1.1 run_alphafold.sh -d \$ALPHAFOLDDDB -o /home/idilber/Tetrameri/1ZMO/Monomer -f -i 1ZMOMonomer.fasta -p monomer_ptm -t 2022-12-15 --usegpu=False</pre>	A
<pre>#!/bin/bash #\$ -N parafold2 #\$ -pe gpu 1 #\$ -l cores=5 #\$ -cwd module load bioinfo/parallelfold/2.1.1 GPUDEVICE=\$(cat \$TMPDIR/gpu) cuda-wrapper.sh run_alphafold.sh -d \$ALPHAFOLDDDB -o /home/idilber/Tetrameri/1ZMO/Monomer -i 1ZMOMonomer.fasta -p monomer_ptm -t 2022-12-15 -a \$GPUDEVICE -n 5</pre>	B

Figure 7. An example for cpu.sge and gpu.sge scripts. (A) shows an example of cpu.sge script for 1ZMO monomer. The flag marked in red "-usegpu=False" informs us that the GPU is not used, but the CPU is, while -f stops AlphaFold 2 after creating feature.pkl in the output directory. (B) shows an example of gpu.sge script for 1ZMO monomer. We started the same job in the same directory and feature.pkl must be present in the output folder.

```
#!/bin/bash
#$ -N PLLDT_PAE
#$ -pe *mpisingle 2
#$ -cwd

module load bioinfo/parallelfold/2.1.1

run_visualisation.sh -i /home/idilber/Tetrameri/1ZMO/Monomer/1ZMOMonomer/ -o /home/idilber/Tetrameri/1ZMO/Monomer/1ZMOMonomer -n 1ZMOMonomer
```

Figure 8. An example of the visualization.sge script. The image shows the script for the 1ZMO monomer, made according to the rules written on the Srce website. The path to the directory with the results (-i), the path to the directory where the images will be created (-o) and the name of the image (-n) are circled in red.

To investigate the ability of AlphaFold 2 to predict 3D structures from random protein sequences, that are not available in PDB, we used a Random Protein Sequence generator (RandSeq). RandSeq is a free web tool that generates random protein sequences (<https://web.expasy.org/randseq/>).

The results were analysed using free molecular visualization programs, PyMOL and UCSF ChimeraX. In this thesis, structures are represented in PyMOL. Otherwise, the name of the drawing software is given. For the calculation of LDDT, the basic LDDT tool is available at <https://swissmodel.expasy.org/assess>. SWISS-MODEL is a web tool that uses homology modeling to predict the 3D structure of proteins. (30)

4. Results

4. 1. AlphaFold 2 predicts 3D protein structures poorly from randomly generated amino acid sequences

To test the capabilities of AlphaFold 2, we decided to use RandSeq (<https://web.expasy.org/randseq/>) to generate random protein sequences. The hypothesis was that AlphaFold 2 could not predict such structures. Before randomly generating a protein sequence, the length of the sequence (from 200 to 9999 amino acids) and the composition of the sequence, i.e. proportion of amino acids must be determined. 200 amino acids were chosen for the length of the sequence and the specific proportions of individual amino acids were defined in three different ways. The first random amino acid sequence had the average composition of 20 amino acids in the proteins from UniProtKB/Swiss-Prot database (Figure 9A); the second one had the composition that was chosen by slightly and arbitrarily changing average composition (Figure 10A), and the third one was strongly disturbed from the average composition of proteins (Figure 11A). The FASTA format was chosen as the output so that the resulting sequences could directly be run by AlphaFold 2. In all three cases, pLDDT values were low meaning that all predicted conformations are of low reliability. According to pLDDT values the predicted AlphaFold 2 conformations or their parts are categorized as: pLDDT <50 very low confidence, 50-70 low confidence, 70-90 high confidence, >90 very high confidence, i.e. correctly predicted 3D protein structures. (15)

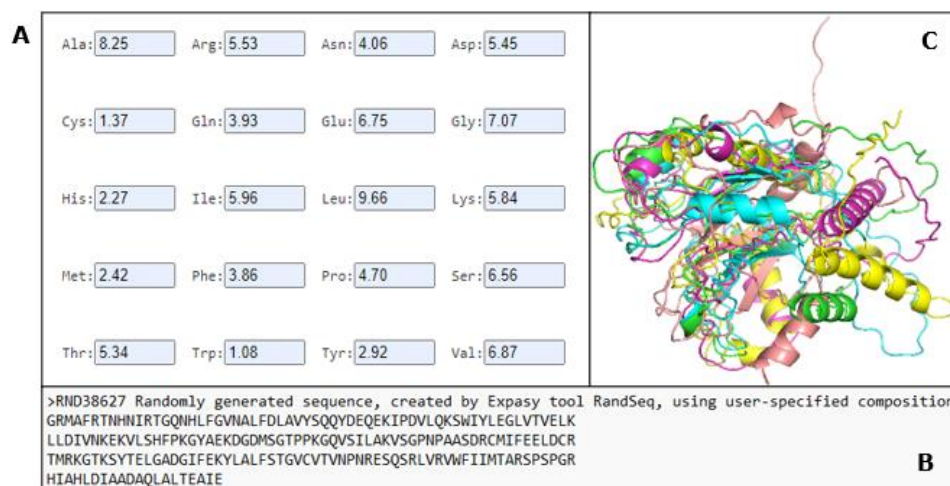


Figure 9. AlphaFold 2 structure outputs for the primary sequence with average amino acid composition are irregular. (A) Average percentages of amino acids suggested by the RandSeq online tool, (B) the FASTA format for the random amino acid sequence with average composition shown in 9A, and (C) the five mutually highly dissimilar 3D structures predicted by AlphaFold 2 for the random sequence shown in 7B, visualized in PyMOL. The average RMSD is 16.50 Å, with pLDDT values ranging from 24.42 % to 28.92 %.

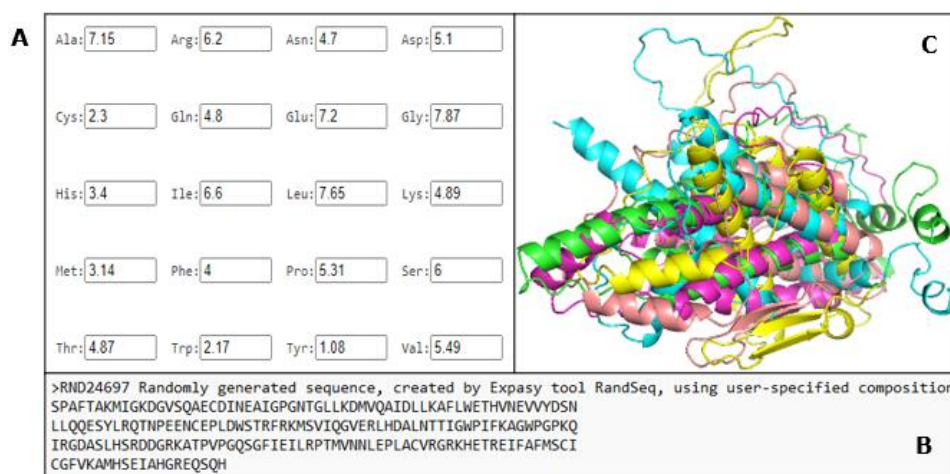


Figure 10. AlphaFold 2 generated 3D protein structures for a primary sequence with arbitrary composition have low pLDDT values and a small percentage of mutual alignment. (A) Small deviation in amino acid proportions from the average percentages shown in Figure 9A, and (B) the corresponding FASTA input. (C) The five 3D structures are shown in different colours with pLDDT values from 34.18 % to 37.57 %, shown by PyMOL. The average RMSD is 16.75 Å.

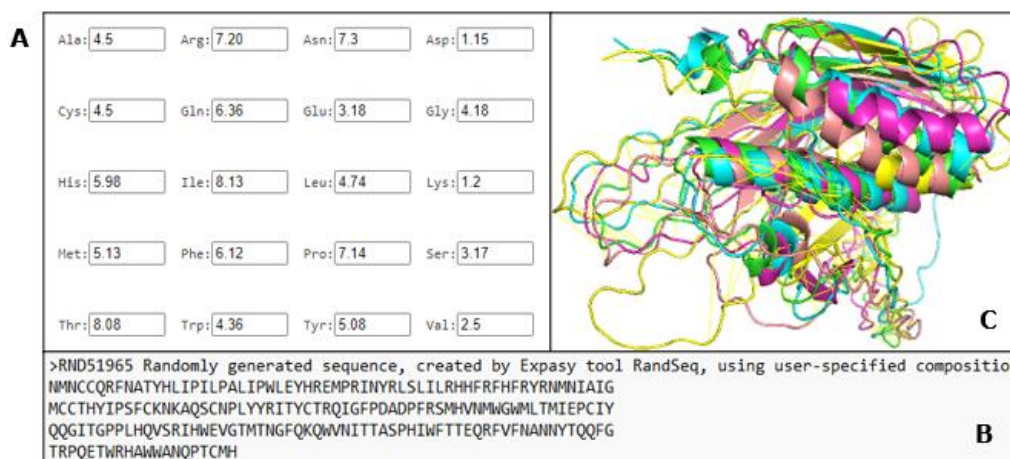


Figure 11. AlphaFold 2 protein structures predicted for the primary sequence with amino acid proportions very different from the average values. (A) Big deviation in amino acid proportions from the average percentages shown in Figure 9A, and (B) the corresponding FASTA input. (C) The 3D structures are shown in PyMOL, with pLDDT scores from 28.70 % to 32.74 %. The average RMSD is 6.12 Å.

4. 2. Amino acid substitutions do not significantly affect the conformations of monomeric subunits of the halohydrin dehalogenase enzymes HheA and HheC

4. 2. 1. AlphaFold 2 predictions for the HheA monomer

In order to determine the reliability of AlphaFold 2 for protein modeling, the sequence of amino acids for the dehalogenase HheA was taken from the PDB structure 1ZMO, and the 3D conformations of its monomeric and tetrameric forms were predicted. The monomer and tetramer conformations were also predicted for the HheA N178A mutant in which the asparagine at position 178 close to the catalytic triad, was replaced by alanine.

The hypothesis was that AlphaFold 2 would more accurately model monomers than multimers and that there would be no significant difference between the predicted conformations of the WT monomer and the N178A mutant.

Figure 12A shows local LDDT scores along the FASTA sequence for the best, top-ranked structure predicted by AlphaFold 2 for the HheA monomer as assessed against the crystal structure 1ZMO as a reference structure. The local LDDT scores in Figure 12A as well as the LDDT total score were calculated using the previously mentioned online tool Structure Assessment Tool (<https://swissmodel.expasy.org/assess>). The tool is designed to compare single chains, which means that the first chain in the structure of the desired model is compared with the first chain of the reference structure. Only amino acid chains connected by peptide bonds are taken into account for the calculation, while everything else is ignored (water, ligands, DNA). (2) Figure 12B depicts the comparison of local pLDDT values for amino acid residues for five predicted monomeric structures for HheA as assessed by AlphaFold 2.

In comparison with the crystal structure of the 1ZMO subunit, the greatest deviation in the ranked_0 model (Figure 12A) is observed for the amino acid residue Tyr143. The difference in positions of Tyr143 in the crystal 1ZMO structure and in the modeled structure is shown in Figure 13. Figure 13C provides a visual representation of the excellent alignment of the secondary structure elements between crystal and ranked_0 structures of monomer. In Figure 13D the regions with the lower local LDDT values corresponding to amino acids 190-205 and the C-terminal tail (Figure 12) are circled.

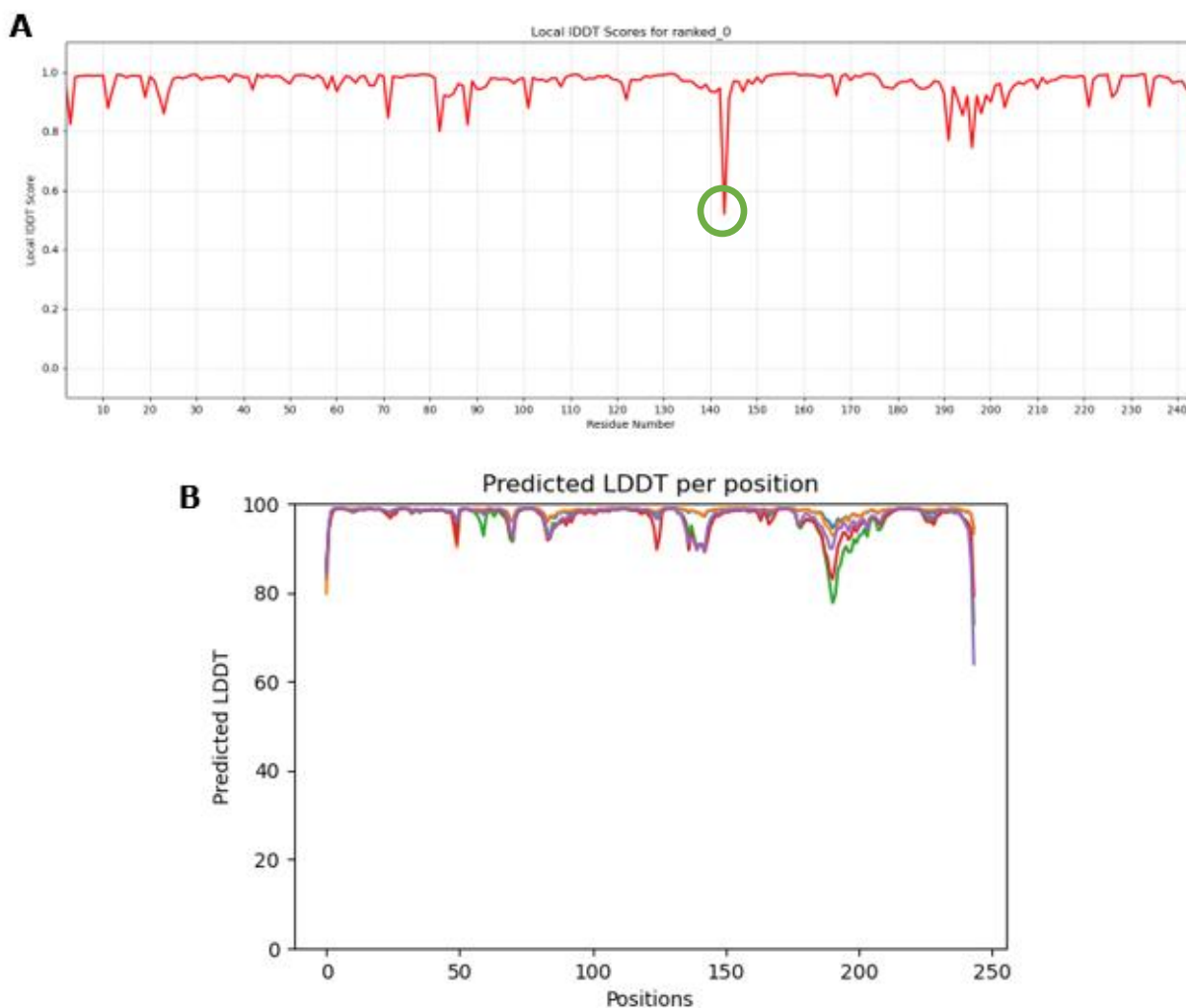


Figure 12. AlphaFold 2 conformations for hypothetical HheA monomer have high predicted pLDDT scores. (A) Local LDDT values of monomer structure obtained by AlphaFold 2 were assessed against a monomer subunit from the PDB structure 1ZMO of the WT HheA enzyme as a reference. Deviation in (A) circled green corresponds to the 143rd amino acid residue, tyrosine, with a local LDDT value of 51.90 %. Global LDDT is 96.13 %. (B) Variation of AlphaFold 2 pLDDT values for five structures predicted for HheA monomer.

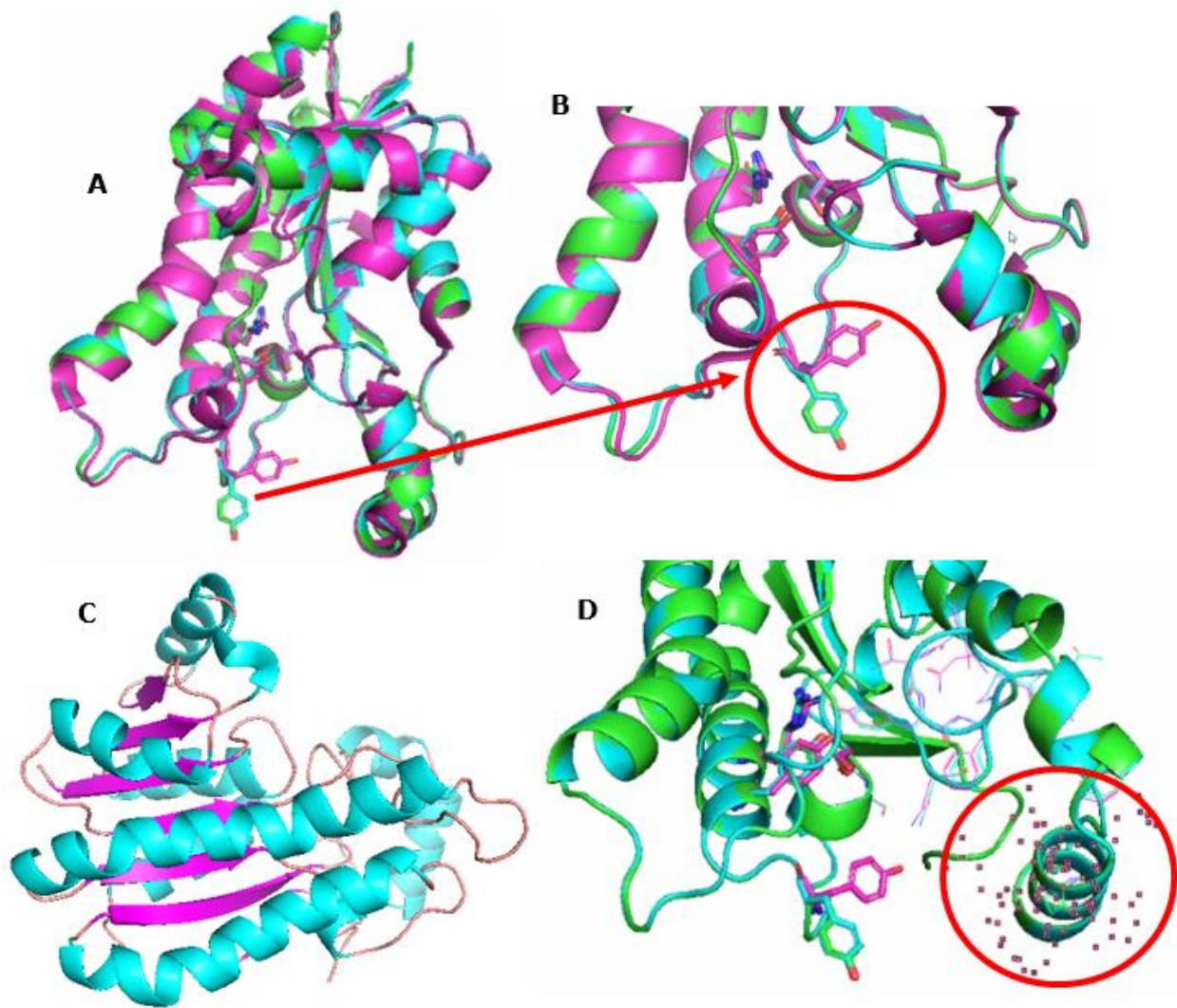


Figure 13. AlphaFold 2 correctly predicts the conformation of the 1ZMO monomeric subunit. (A) Alignment of the experimental 1ZMO subunit A (pink) with AlphaFold 2 predicted WT (cyan) and N178A mutant (green) monomer. RMSD values are ~ 0.23 Å. (B) The amino acid Tyr143 is circled red. (C) Alpha helices (cyan), beta sheets (purple), and loops (pink) are well reproduced by AlphaFold 2. (D) The alpha coil and the C-terminal tail represent the dynamic parts of HheA monomer. In addition to Tyr143, the catalytic residues Ser134, Tyr147 and Arg151 are shown in stick representation.

In the case of the N178A mutant, the smaller local LDDT value with 1ZMO structure of WT HheA as a reference has been observed for the mutated

amino acid residue which is illustrated in Figures 14A and 14B obtained using the online Structure Assessment Tool, <https://swissmodel.expasy.org/assess>.

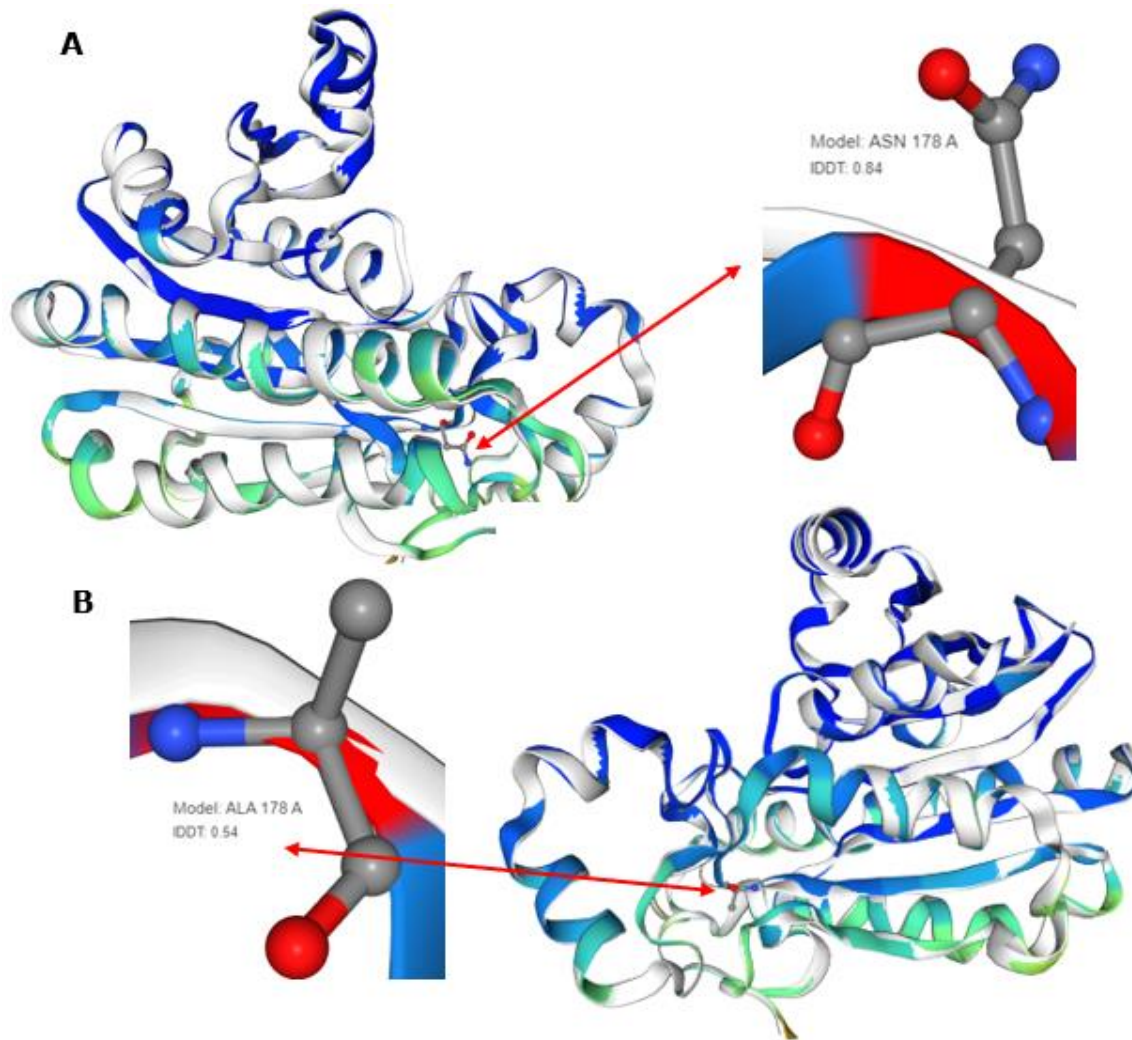


Figure 14. The amino acid N178A mutation leads to a significantly lower IDDT score for the mutated residue while the rest of the conformation is the same in WT and N178A mutant. (A) The arrow indicates the amino acid asparagine (ASN), located at the 178th position on chain A, and its local IDDT value of 84 % as compared to 1ZMO monomer conformation as a reference. (B) The arrow indicates the amino acid alanine (ALA), which replaced the amino acid asparagine of the WT HheA enzyme. It is associated with a significantly lower local IDDT value of 54 %. Areas marked in dark blue have a high IDDT score (70-100), while areas marked in green indicate low local IDDT scores (50-70).

Figure 15 demonstrates the usage of the metric Predicted Alignment Error (PAE) in AlphaFold 2. PAE indicates the expected positional error at residue x if the predicted and actual structures are aligned at residue y (using C α , N, and C atoms), that is a distance error for every pair of residues. (31)

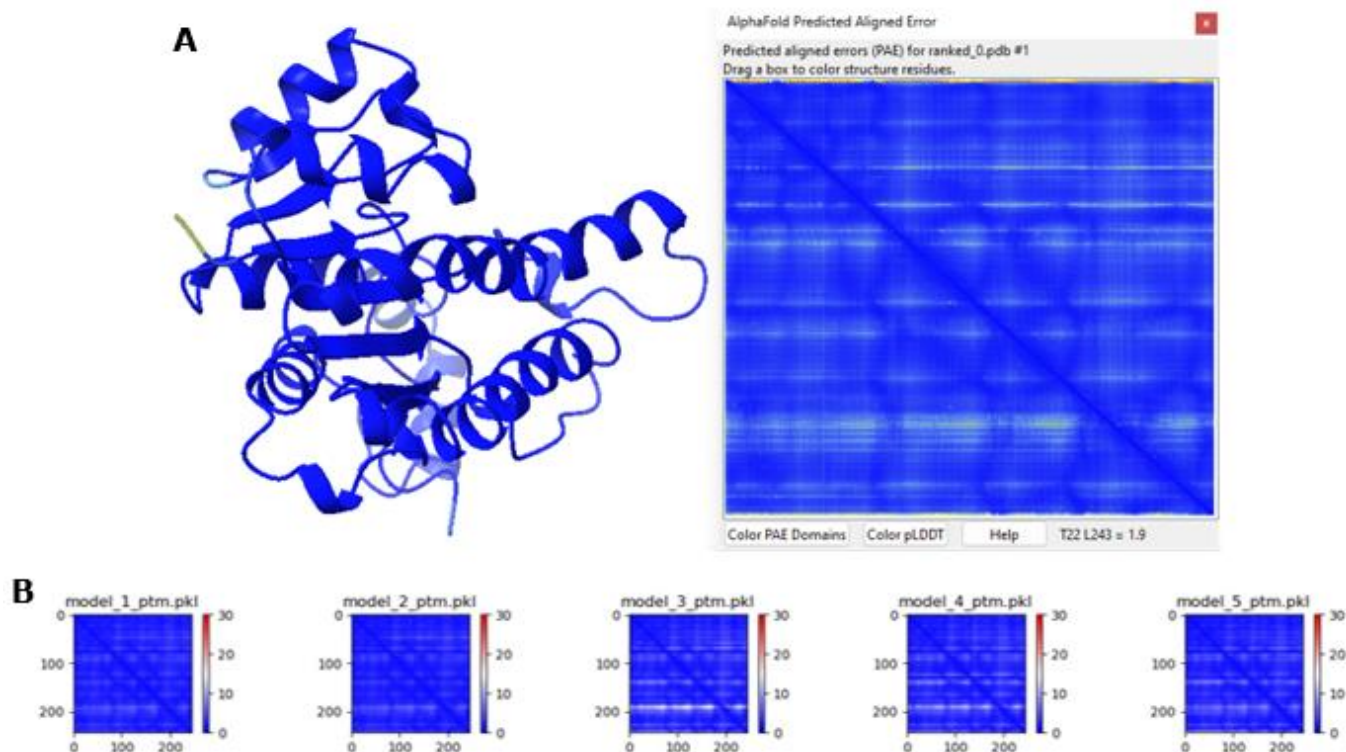


Figure 15. Predicted Alignment Error (PAE) plots for WT HheA monomeric models.
 (A) PAE output of model 1 shown in ChimeraX – structure colored according to pLDDT values.
 (B) PAE plots for the first 5 models were obtained with AlphaFold 2. The coloring is according to PAE values 0 - 31.75 Å.

4. 2. 2. AlphaFold 2 predictions for the HheC monomer

In the PDB, another halohydrin dehalogenase biocatalyst HheC is found under the PDB ID 1ZMT, which shares 35% sequence identity with 1ZMO, that is HheA isoform. (21) The analogous calculations and analysis of the predicted conformations were conducted for the enzyme HheC as for the dehalogenase HheA (Figures 16, 17, 19 and 20). In addition, considering the essential role of the catalytic triad Ser-Tyr-Arg, we also checked the reliability of AlphaFold 2 in predicting its conformation in the case of HheC (Figure 18).

In Figures 16A and B the local IDDT and pLDDT values were shown for top ranked_0 structure and all five models, respectively, obtained by AlphaFold 2 for the monomeric conformation of WT HheC. The local IDDT values were calculated with the structure of the subunit/chain A from the crystal structure 1ZMT as a reference. In Figure 17, the explanation of the low IDDT value that is poor agreement with the crystal position of the amino acid residue Pro84 (Figure 16A), is provided through visualization using PyMOL. Figure 18 shows the excellent reproduction of positions of the catalytic amino acid residues in the ranked_0 model, with 1ZMT structure as a reference. Figure 19 represents an evaluation of the AlphaFold2 results for the quadrupole mutant P84V/F86A/T134A/N176A and Figure 10 shows PAE diagrams for five output models of HheC monomeric structure.

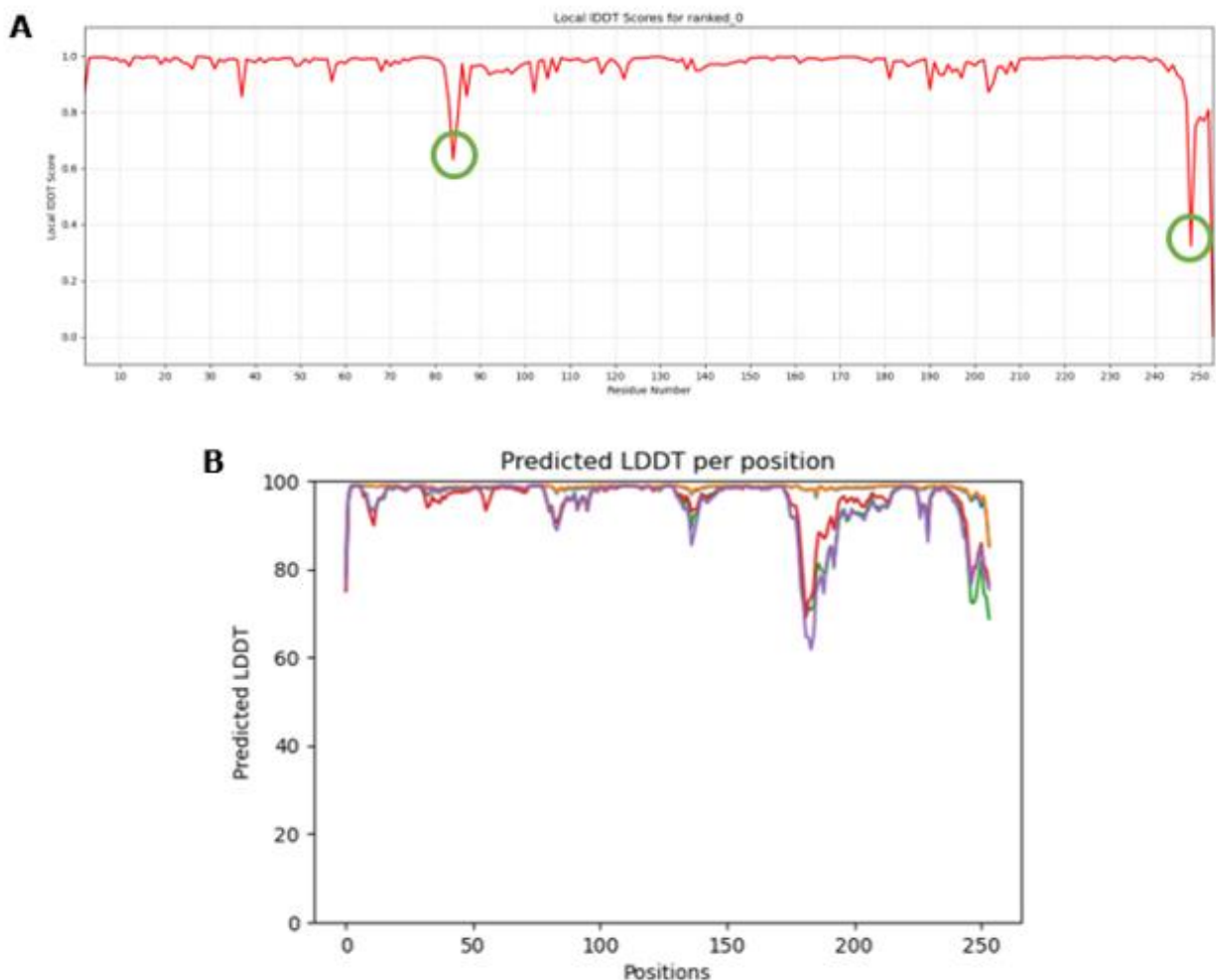


Figure 16. AlphaFold 2 predicted the monomer structure of HheC with a high global LDDT score. (A) Local IDDT values of monomer structure obtained by AlphaFold 2 assessed by a monomer subunit from the PDB structure 1ZMT of the WT HheC enzyme as a reference. Deviations in (A) circled green correspond to the 84th and 248th amino acid residue, proline and arginine, with local LDDT values of 63.10 % and 32.34 %. Global LDDT is 97.25 %. (B) AlphaFold 2 pLDDT values for five modeled structures for the HheC monomer.

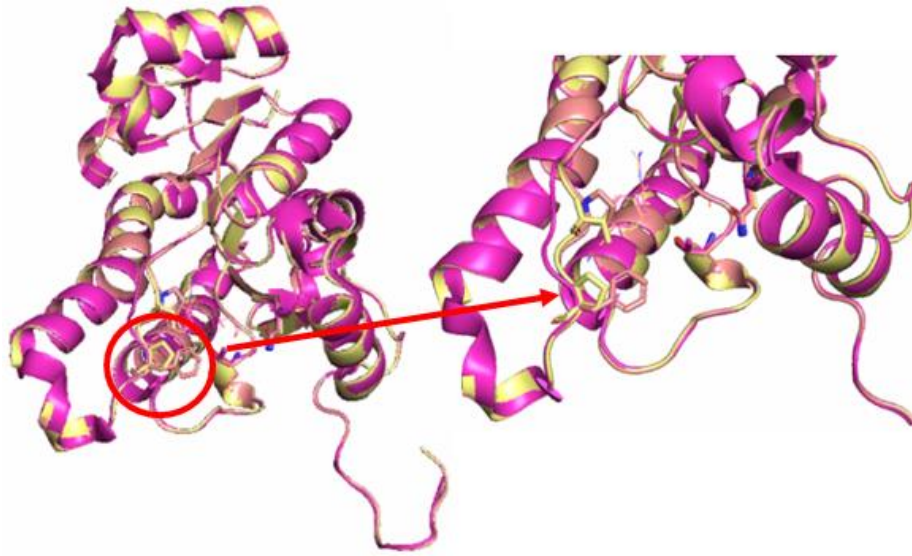


Figure 17. Incorrectly predicted proline position in the predicted WT monomer conformation with the 1ZMT structure as a reference. Alignment of the experimental 1ZMT subunit A (purple), AlphaFold 2 predicted WT monomer (pink), and quadrupole monomer mutant (yellow) and. The wrongly predicted Pro84 is shown on the right, which is also confirmed by the local LDDT plot (Figure 16A). Arg248 is not shown in the figure, but it is located on the C-terminal tail.

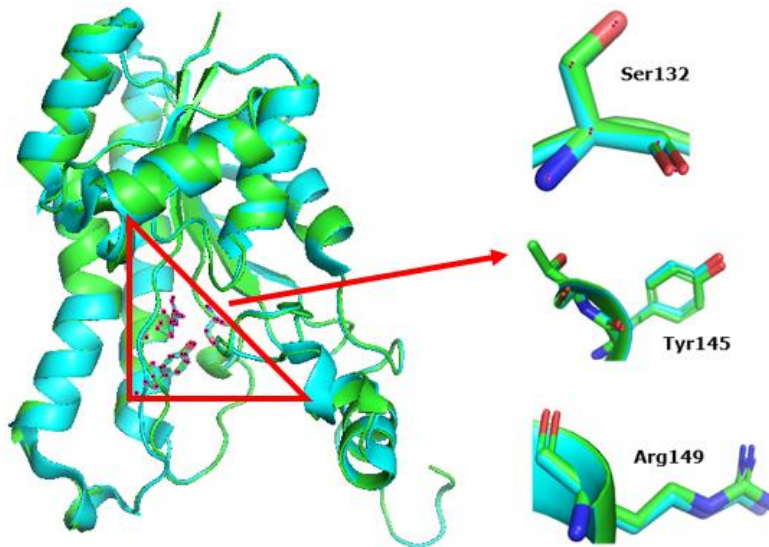


Figure 18. AlphaFold 2 correctly predicts the positions of the catalytic amino acid residues in WT and mutant monomeric structures with 1ZMT as a reference. Alignment of WT 1ZMT monomer (green) and AlphaFold 2 predicted 1ZMT monomer mutant (blue) indicates that the positions of the catalytic sites Ser132, Tyr145, and Arg149 do not change with P84V/F86A/T134A/N176A mutations. RMSD is 0.18 Å.

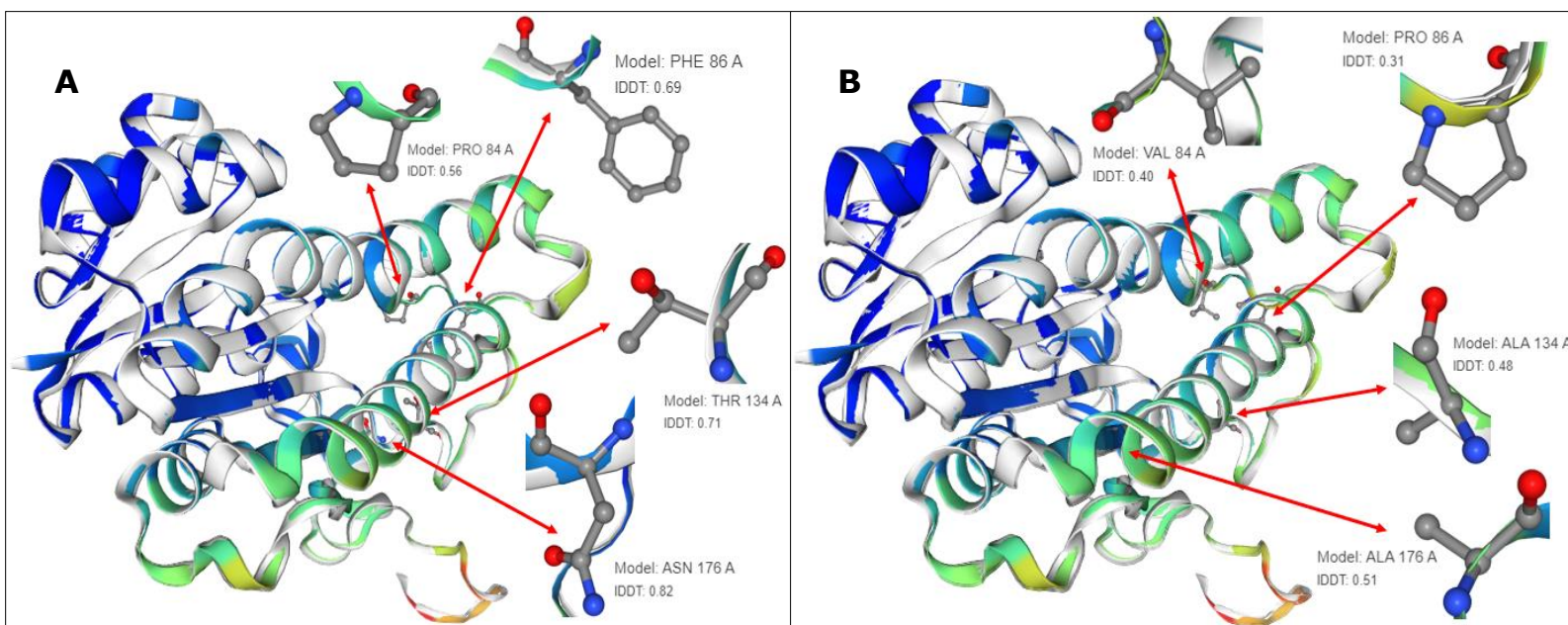


Figure 19. The change in amino acids leads to their significantly lower local IDDT scores as calculated and visualized by the Structure Assessment Tool. (A) Arrows indicate the amino acids proline (PRO) at position 84, phenylalanine (PHE) at position 86, threonine (THR) at position 134 and asparagine at position 176. (B) Arrows show valine (VAL) instead of proline at position 84, proline instead of phenylalanine at position 86, alanine instead of threonine at position 134, and asparagine at position 176. Each of them has an IDDT value associated with it. The chain is colored according to the values of the local IDDT score - areas marked in dark blue have a high IDDT score (70-100), while areas marked in green indicate low IDDT scores (50-70).

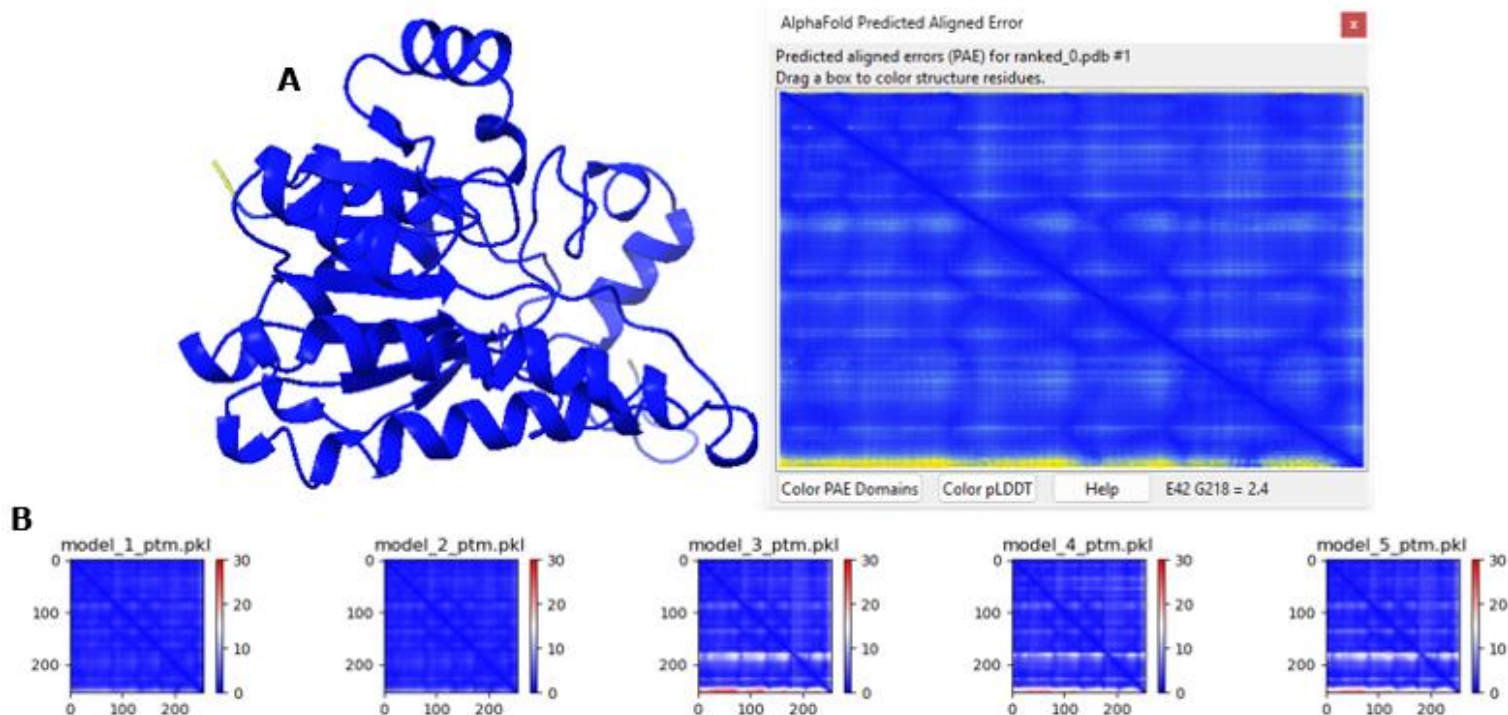


Figure 20. Predicted Alignment Error (PAE) plots for predicted WT HheC monomeric structures. (A) PAE output of model 1 shown in ChimeraX – conformation colored according to pLDDT values. (B) PAE values for the first 5 models obtained with AlphaFold 2.

4. 3. AlphaFold 2 successfully predicts symmetric homotetrameric conformations

After testing AlphaFold 2 in predicting monomer/subunit structures, we decided to find out how it would perform in predicting structures of tetramers of halohydrin dehalogenases HheA and HheC. The hypothesis is that AlphaFold 2 may have difficulty predicting the 3D structures of these multimers. Figure 19 shows local IDDT values calculated for predicted tetramers of HheA and HheC using the X-ray structures 1ZMO and 1ZMT as references. In Figure 20 it is visible that for HheA the tetrameric structure is predicted better than in the case of HheC.

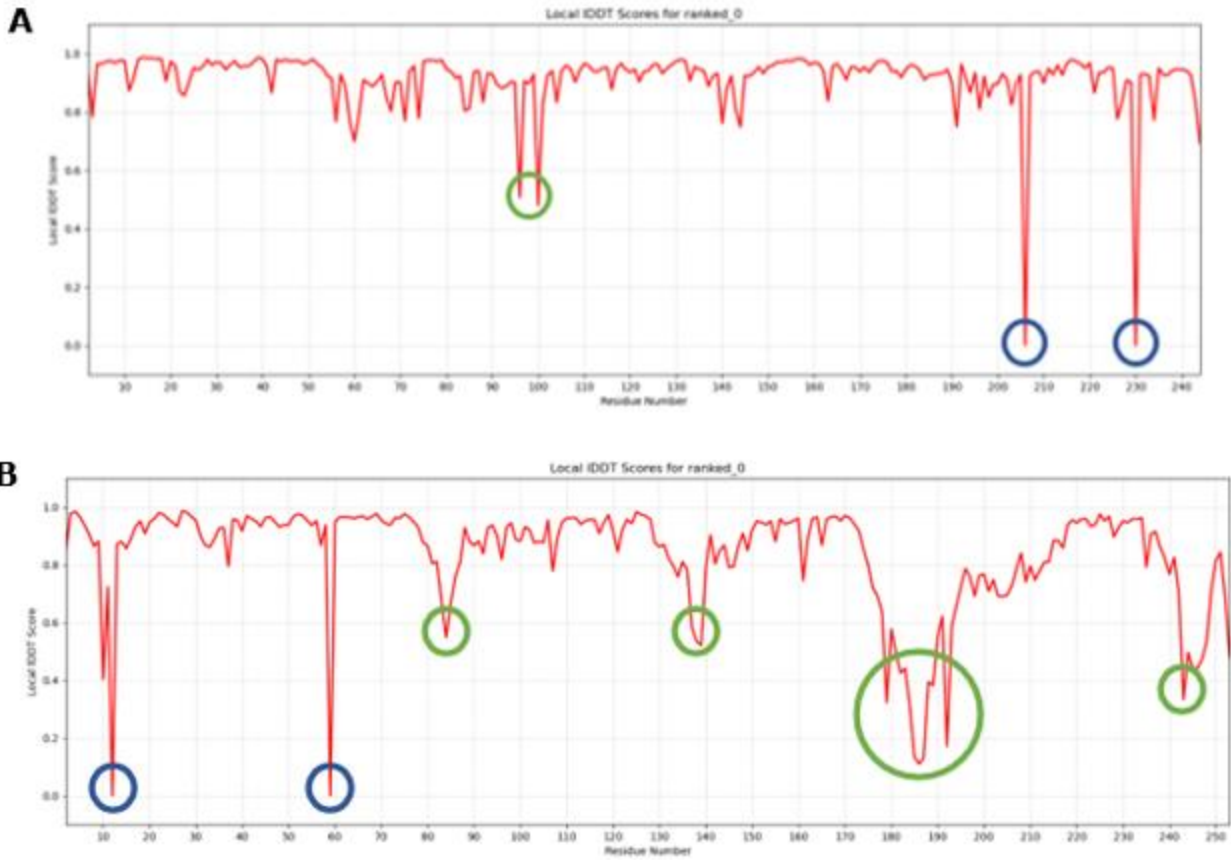


Figure 21. Local IDDT scores for the homotetrameric structures of HheA and HheC.

(A) Local IDDT of WT HheA tetramer and (B) WT HheC tetramer modeled by AlphaFold 2 and compared with their PDB structures 1ZMO and 1ZMT, respectively. Deviations circled green represent wrong predictions of amino acid positions that deviate considerably from their crystal orientations, while those in blue are wrong backbone predictions. Global IDDT is (A) 92.12 % and (B) 82.04 %.

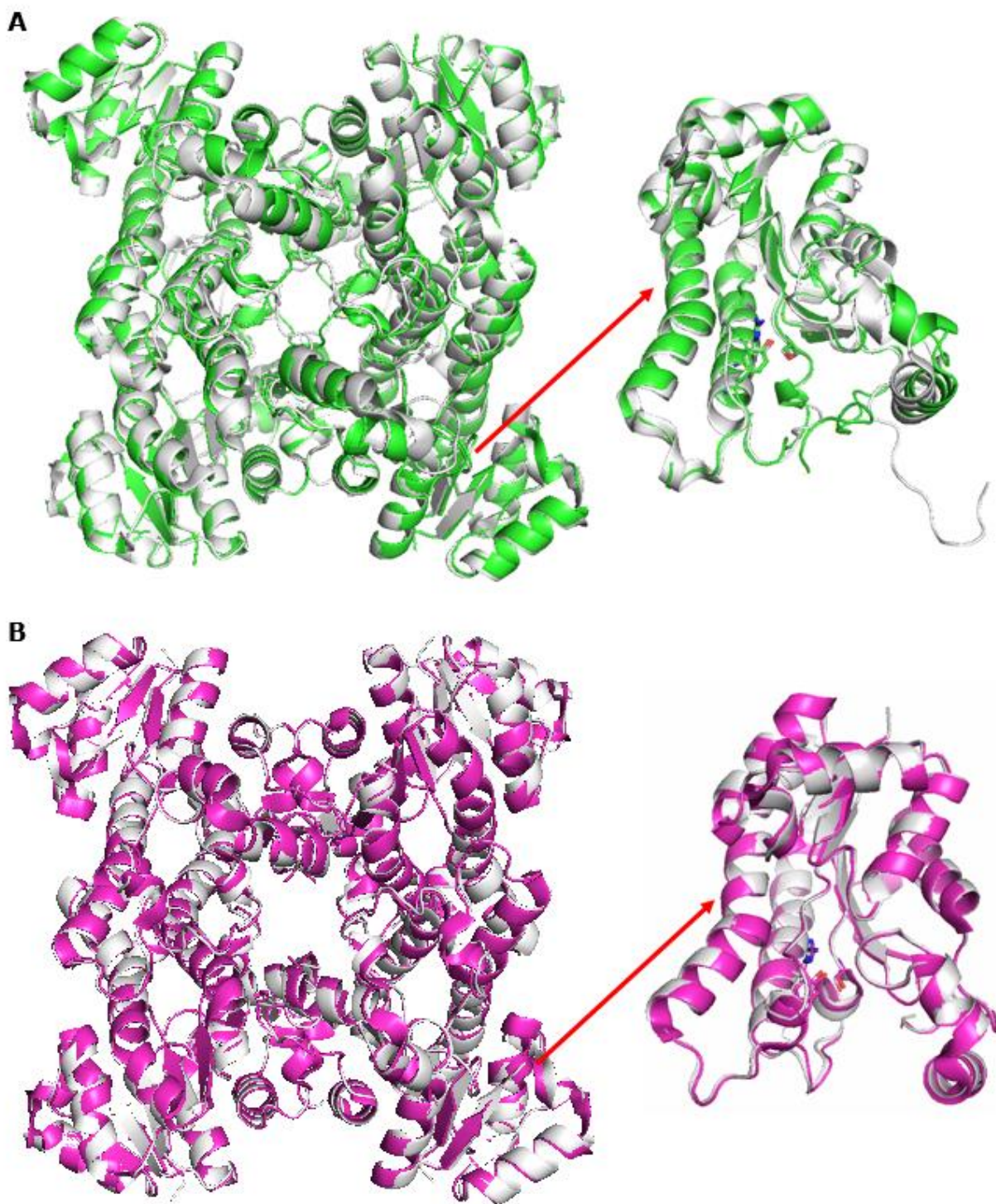


Figure 22. Comparison of WT AlphaFold 2 tetrameric top-ranked models with the crystal structures. (A) For HheC - AlphaFold 2 predicted structure (green) and 1ZMT structure (white) and (B) HheA - AlphaFold 2 predicted multimer (white) and 1ZMO structure (pink). The RMSD values are 0.373 Å for the 1ZMO multimer and 0.620 Å for the 1ZMT multimer.

5. Discussion

The determination of 3D structures of proteins, as well as the discovery of drugs interacting and binding to them, is a comprehensive and expensive research process that requires a great deal of time and money. To save time in the lab and money on the necessary experiments, we focused on predicting protein conformations using artificial intelligence and machine learning. AlphaFold2 is an AI software we decided to test for the purpose of predicting enzyme structures and functionally relevant regions of the protein.

Firstly, we generated randomly 200 amino acid long sequences to test whether AlphaFold 2 could predict meaningful structures based on arbitrary primary sequences. Figure 9C shows the structure obtained from the sequence (Figure 9B) which was created by the composition of the average amino acids in the protein suggested by the RandSeq site (Figure 9A). It is visible that the 3D models predicted by AlphaFold 2 are badly mutually aligned and that there is no agreement among them. This is indicated by the high average RMSD value of 16.50 Å (well-predicted structures have RMSD value <1) and pLDDT values ranging from 24.42 % to 28.92 %. According to pLDDT values, the predicted AlphaFold 2 conformations or their parts are categorized as pLDDT <50 very low confidence, 50-70 low confidence (excluding 70), 70-90 high confidence (including 70), >90 very high confidence, i.e., correctly predicted 3D protein structures. This means that the predicted structures are of very low confidence. We see almost the same performance in Figure 10, where the structure under C was produced from a sequence (Figure 10B) created by a small change in the average proportions of amino acids (Figure 10A). The average RMSD value of 16.75 Å is similar to the previous one, while the pLDDT values are slightly higher and range from 34.18 % to 37.57 %, which is again a sign of low confidence and failed prediction by AlphaFold 2. We had a little better luck with the structure visible in Figure 11C, which we obtained by completely perturbing the average amino acid composition of the protein.

Despite a slightly better average RMSD of 6.12 Å, the pLDDT values range from 28.70 % to 32.74 %. The models we received were disappointing but also expected. Namely, AlphaFold 2 needs templates like those from the PDB database for the proteins with high MSA scores, which it uses as guidelines for predicting structures. As an AI tool, AlphaFold 2 learns from data, and since the protein sequences are not random, there are no templates available for randomly generated primary sequences of amino acids. The underlying reasons for the lack of accuracy in the case of random amino acid sequences are the lower matches from the MSA and the input sequences, and the lack of quality and representative structural templates. In short, in all three cases, the problem is that not enough related sequences and structures are known, so AlphaFold 2 did not have enough "prior knowledge" for quality prediction.

We also used AlphaFold 2 for the prediction of structures of the well-known biocatalysts halohydrin dehalogenase HheA and HheC in their tetrameric and hypothetical monomeric states. (23) For the former extensive molecular dynamics simulations have already been done. (32,33) The FASTA inputs of haloalcohol dehalogenase HheA and HheC were generated from their crystal structure 1ZMO and 1ZMT, respectively, available in the PDB database. Both dehalogenases are symmetric homotetramers. We used AlphaFold 2 to predict the structures of monomers and tetramers of these two dehalogenases.

Figure 12 shows two graphs for the dehalogenase HheA; Figure 12A shows the local IDDT values of top ranked_0 monomeric structure calculated using the online tool Structure Assessment Tool (<https://swissmodel.expasy.org/assess>) with regard to the crystal 1ZMO subunit A as a reference. Figure 12B shows the predicted pLDDT values calculated by AlphaFold 2 on the MSA basis. Along with the plot, the Structure Assessment Tool offers results in the table form, from which the local LDDT score of each amino acid can be read, as well as its position. The low local

IDDT value of 51.90 % is observed for tyrosine which is in 143rd place in the primary sequence. For comparison, almost all pLDDT values are greater than 70 (except for the C terminal end). The difference in positions of Tyr143 in the ranked_0 and crystal 1ZMO structure is illustrated in Figure 13. It is located at the very entrance to the binding site on the monomer surface. Other structural elements of the HheA subunit are well predicted as visualized in Figure 13C by alignment of various secondary structure elements (alpha coils – cyan, beta-sheet – purple, loops – pink). In Figure 13D the regions with the low local LDDT values are marked by the red circle. They correspond to alpha helix formed by amino acids 190-205 and the C-terminal tail. Corresponding structural elements have been shown in dehalogenase HheC to be the most dynamic parts. (32)

Figure 14 was also obtained using the Structure Assessment Tool for the purpose of visualizing the change in the local LDDT value due to single-point mutation N178A. In Figure 14A we see the modelled monomeric structure for WT HheA with asparagine at position 178 and its LDDT score is equal to 84 %. In Figure 14B the AlphaFold 2 model for the N178A HheA mutant is shown, the IDDT score for Ala178 is significantly lower and is 54 %. In a global view, the 3D structure of the HheA monomer subunit is predicted with high confidence. The LDDT total score is 96.13 %, while the RMSD is 0.23 Å, which is a sign of high confidence (LDDT > 90 %, RMSD < 1 Å).

Figure 15 shows the PAE plots of 5 models of WT 1ZMO monomer, focusing on the first result_model_1. Predicted Alignment Error (PAE) is an output of the AlphaFold 2 system (limited to 31.75 Å) that researchers can use to estimate the reliability of the relative position and orientation of different parts of the protein model (e.g., of two domains). Groupings on the PAE plot point to the existence of different structurally distinguished parts and domains within protein. It is usually presented as a heatmap, with residue numbers shown along the vertical (y = aligned residue) and horizontal (x =

scored residue) axes and the color of each pixel indicating the PAE value for the corresponding residue pair. As mentioned before, PAE indicates the expected positional error at residue x if the predicted and actual structures are aligned at residue y (using Ca, N, and C atoms). The predicted conformation visible in Figure 15A was colored according to LDDT values and generated in ChimeraX. Expected position error is expressed in Ångströms, and the structure is considered well-predicted and realistic if the PAE values are low (0-10, blue color), while it should not be given biological and structural interpretation if the PAE values are high (20-30, red color). (31) It is clearly visible from the picture that the structure is predicted correctly (high confidence in position and orientation), and somewhat higher PAE scores can be seen for residues around position 195, which is in accordance with Figure 12B, where pLDDT is around 75 % for these amino acids. The PAE output of the 1ZMO N178A mutant is not shown, as it is almost identical.

For the haloalcohol dehalogenase HheC which is the biocatalyst widely used in industry, as a reference for predicted structures we used the X-ray structure with the PDB ID 1ZMT. In the case of the HheC enzyme, in addition to estimating the quality of predicted structures, we also focus on details regarding the position and orientation of its catalytic residues Ser132, Tyr145 and Arg149. The 3D structure of HheC is predicted with high confidence. The LDDT total score is 97.25 %. The largest deviations from the structures of subunits in the crystal structure 1ZMT were observed for the amino acid residues Pro84 and Arg248 (Figure 16A). The LDDT values for Pro84 and Arg248 are 63.10 % and 32.34 %, respectively, which is a sign of low confidence. In difference of unconfidently predicted positions of these amino acid residues (Figure 17), positions and orientations of the catalytic residues Ser132, Tyr145 and Arg149 are predicted with high confidence (Figure 16) in monomeric models of both WT and quadrupole P84V/F86A/T134A/N176A mutant of dehalogenase HheC, despite the fact that the quadruple mutations

are close to the catalytic sites. (33) The positions of mutated amino acid residues are predicted mostly with low to very low confidence (IDDT < 70) in WT and mutant HheC proteins (Figure 19). Figure 20 shows PAE plots of 5 predicted models of the WT HheC monomer, focusing on the first resulted model. It is clearly visible from the picture that the structure is correctly predicted (high confidence in position and orientation), and high PAE scores can also be seen around 245th residue, which is confirmed in Figure 16A, where the LDDT for Arg248 is 32.34 % for. The PAE output for the monomeric structures predicted for the quadruple mutant of HheC is not shown, as it is almost identical.

We also used AlphaFold 2 to predict multimeric structures of the studied halohydrin dehalogenases HheA and HheC which are both natural tetramers. Figure 21 shows the local IDDT values of HheA and HheC tetramer subunits calculated with their PDB structures 1ZMO and 1ZMT, respectively, as references. Deviations in (A) circled green correspond to the 96th and 100th surface amino acid residues, glutamic acid and arginine, with local LDDT values of 50.69 % and 48.01 %, respectively. The incorrect backbone prediction is indicated by the blue circles at the 206th and 230th positions in the sequence. The deviations under (B) for HheC are various, and most of them indicate incorrectly predicted amino acids of proline (Pro59, Pro84, Pro138 and Pro244) and phenylalanine (Phe12, Phe186 and Phe243). AlphaFold 2 predicts the tetramer of HheA with a high global IDDT score of 92.19% and the RMSD values with the referent crystal structure 1ZMO of 0.373 Å (Figure 22B). The global IDDT for HheC is lower. It is 82.04 %. Figure 22A shows the alignment of the top-ranked AlphaFold 2 tetramer model (green) with the crystal structure 1ZMT of WT HheC (white), and the corresponding RMSD is 0.620 Å. AlphaFold 2 failed to predict correctly position of the C-terminal tail which is in the predicted structure placed at the entrance of the binding site of its chain, while in the crystal structure, it is positioned

at the entrance of the binding site of the diagonal subunit. Consequently, conformations of close structural elements are also predicted with lower confidence (Figure 22A). WT HheA and HheC tetramer mutants showed analogous results and are, therefore, not shown.

The use of the AI algorithm AlphaFold 2 requires a high-performance computer cluster for running and specialized algorithms for analyzing its outputs, as well as extensive knowledge and experience in bioinformatics and structural biology. For analysis of AlphaFold 2 results the open accessed visualization software PyMOL and ChimeraX were used as well as the freely available online server Structure Assessment Tool.

In the future, more attention should be paid to improving algorithms and expanding the availability of high-quality experimental data for model training. Such extensions would contribute greatly to fields such as biotechnology and personalized medicine, where the focus is on customized treatment based on an individual's unique protein structure and genetic variation.

6. Conclusion

This thesis is based on investigating the ability of AlphaFold 2 to accurately and precisely predict the 3D structures of proteins. The research began by predicting peptide structures having random amino acid sequences generated in three ways. The resulting AlphaFold 2 conformations were unconfident and unreliable. Such results are to be expected as AlphaFold 2 is an AI software that learns from the available experimental data from various databases such as UniProt or PDB, emphasizing the importance of having high-quality data. In addition, this also reflects that proteins' sequences are not arbitrary.

We then predicted the structures of two catalytic enzymes, haloalcohol dehalogenases HheA and HheC, which share 35% sequence identity. AlphaFold 2 successfully predicted their 3D structures, particularly of their monomeric subunits, as evidenced by high global LDDT scores and low RMSD and PAE values. The predicted structures for dehalogenases HheA and HheC in their monomeric and natural tetrameric forms, including catalytic residue positions, are consistent with their PDB conformations. However, a structural peculiarity of intruding of the C-terminal tail to the diagonal subunit of the HheC tetramer is not predicted.

With this work, we aimed to demonstrate the ability of AlphaFold 2 in predicting 3D protein structures, but also to point out shortcomings that should be resolved in subsequent versions of this valuable software. AlphaFold 2 helps us to quickly get to the 3D structures of proteins, after which we visualize the structures, using available tools like PyMOL or ChimeraX. This can be important because they help us understand what to focus further on in performing either *in silico* (e.g., molecular dynamics simulations) and/or wet experiments.

The future of AlphaFold 2 is likely to be a combination of improving the program itself and exploring its various applications in different scientific and industrial fields. As research in structural biology and bioinformatics advances, AlphaFold 2 is expected to play a significant role in unlocking the secrets of protein structure and function, leading to numerous breakthroughs in science and technology.

7. References

1. Wisniak J. Jöns Jacob Berzelius A Guide to the Perplexed Chemist. The Chemical Educator. 2000 Dec;5(6):343–50.
2. Borkakoti N, Thornton JM. AlphaFold2 protein structure prediction: Implications for drug discovery. Curr Opin Struct Biol. 2023 Feb;78:102526.
3. Varadi M, Velankar S. The impact of AlphaFold Protein Structure Database on the fields of life sciences. Proteomics. 2023 Sep 27;23(17).
4. Moult J, Fidelis K, Kryshtafovych A, Tramontano A. Critical assessment of methods of protein structure prediction (CASP)—round IX. Proteins: Structure, Function, and Bioinformatics. 2011 Jan 14;79(S10):1–5.
5. Bertoline LMF, Lima AN, Krieger JE, Teixeira SK. Before and after AlphaFold2: An overview of protein structure prediction. Frontiers in Bioinformatics. 2023 Feb 28;3.
6. Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J. Critical assessment of methods of protein structure prediction (CASP)—Round XIII. Proteins: Structure, Function, and Bioinformatics. 2019 Dec 23;87(12):1011–20.
7. Kuhlman B, Bradley P. Advances in protein structure prediction and design. Nat Rev Mol Cell Biol. 2019 Nov 15;20(11):681–97.
8. Yang Z, Zeng X, Zhao Y, Chen R. AlphaFold2 and its applications in the fields of biology and medicine. Signal Transduct Target Ther. 2023 Mar 14;8(1):115.
9. Skolnick J, Gao M, Zhou H, Singh S. AlphaFold 2: Why It Works and Its Implications for Understanding the Relationships of Protein Sequence, Structure, and Function. J Chem Inf Model. 2021 Oct 25;61(10):4827–31.
10. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021 Aug 26;596(7873):583–9.

11. Porta-Pardo E, Ruiz-Serra V, Valentini S, Valencia A. The structural coverage of the human proteome before and after AlphaFold. *PLoS Comput Biol*. 2022 Jan 24;18(1):e1009818.
12. Stevens AO, He Y. Benchmarking the Accuracy of AlphaFold 2 in Loop Structure Prediction. *Biomolecules*. 2022 Jul 14;12(7):985.
13. Azzaz F, Yahia N, Chahinian H, Fantini J. The Epigenetic Dimension of Protein Structure Is an Intrinsic Weakness of the AlphaFold Program. *Biomolecules*. 2022 Oct 20;12(10):1527.
14. Perrakis A, Sixma TK. AI revolutions in biology. *EMBO Rep*. 2021 Nov 4;22(11).
15. Mariani V, Biasini M, Barbato A, Schwede T. IDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*. 2013 Nov 1;29(21):2722–8.
16. Zemla A. LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res*. 2003 Jul 1;31(13):3370–4.
17. Li W, Schaeffer RD, Otwinowski Z, Grishin N V. Estimation of Uncertainties in the Global Distance Test (GDT_TS) for CASP Models. *PLoS One*. 2016 May 5;11(5):e0154786.
18. Manubolu M, Goodla L, Pathakoti K, Malmlöf K. Enzymes as direct decontaminating agents—mycotoxins. In: *Enzymes in Human and Animal Nutrition*. Elsevier; 2018. p. 313–30.
19. Paul PEV, Sangeetha V, Deepika RG. Emerging Trends in the Industrial Production of Chemical Products by Microorganisms. In: *Recent Developments in Applied Microbiology and Biochemistry*. Elsevier; 2019. p. 107–25.

20. Sun H, Zhang H, Ang EL, Zhao H. Biocatalysis for the synthesis of pharmaceuticals and pharmaceutical intermediates. *Bioorg Med Chem*. 2018 Apr;26(7):1275–84.
21. de Jong RM, Kalk KH, Tang L, Janssen DB, Dijkstra BW. The X-Ray Structure of the Haloalcohol Dehalogenase HheA from *Arthrobacter* sp. Strain AD2: Insight into Enantioselectivity and Halide Binding in the Haloalcohol Dehalogenase Family. *J Bacteriol*. 2006 Jun;188(11):4051–6.
22. Hopmann KH, Himo F. Quantum Chemical Modeling of Enzymatic Reactions – Applications to Epoxide-Transforming Enzymes. In: *Comprehensive Natural Products II*. Elsevier; 2010. p. 719–47.
23. Findrik Blažević Z, Milčić N, Sudar M, Majerić Elenkov M. Halohydrin Dehalogenases and Their Potential in Industrial Application – A Viewpoint of Enzyme Reaction Engineering. *Adv Synth Catal*. 2021 Jan 19;363(2):388–410.
24. Schallmey A, Schallmey M. Recent advances on halohydrin dehalogenases—from enzyme identification to novel biocatalytic applications. *Appl Microbiol Biotechnol*. 2016 Sep 8;100(18):7827–39.
25. Belavić M, Imamagić E, Špoljar J. Sveučilišni računski centar Sveučilišta u Zagrebu (Srce). 2022. „Povijest Klastera Isabelle”.
26. <https://www.srce.unizg.hr/napredno-racunanje>. Sveučilišni računski centar Sveučilišta u Zagrebu (Srce). 2023. „Napredno računanje”.
27. Belavić M. Sveučilišni računski centar Sveučilišta u Zagrebu (Srce). 2020. „Tehničke specifikacije Klastera Isabelle”.
28. Tatham S, Lanes A, Harris B, Nevins J. Chiark Greenend. 2023. „PuTTY: a free SSH and Telnet client”.

29. Hrženjak M. Sveučilišni računski centar Sveučilišta u Zagrebu (Srce). 2023. „Parallelfold”.
30. Kiefer F, Arnold K, Kunzli M, Bordoli L, Schwede T. The SWISS-MODEL Repository and associated resources. *Nucleic Acids Res.* 2009 Jan 1;37(Database):D387–92.
31. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* 2022 Jan 7;50(D1):D439–44.
32. Milčić N, Stepanić V, Crnolatac I, Findrik Blažević Z, Brkljača Z, Majerić Elenkov M. Inhibitory Effect of DMSO on Halohydrin Dehalogenase: Experimental and Computational Insights into the Influence of an Organic Co-solvent on the Structural and Catalytic Properties of a Biocatalyst. *Chemistry – A European Journal.* 2022 Oct 7;28(56).
33. Dokli I, Brkljača Z, Švaco P, Tang L, Stepanić V, Majerić Elenkov M. Biocatalytic approach to chiral fluoroaromatic scaffolds. *Org Biomol Chem.* 2022;20(48):9734–41.

8. Curriculum Vitae

Ivana Dilber

Državljanstvo: hrvatsko Datum rođenja: 17/11/1996 Spol: Žensko Telefonski broj: (+385) 916010904

E-adresa: ivanadilber@outlook.com

RADNO ISKUSTVO

STRUČNA PRAKSA

Jadran-galenski laboratorij d.d. (JGL d.d.) [12/06/2023 – 30/06/2023]

Mjesto: Rijeka

Zemlja: Hrvatska

Internetska stranica: <https://www.jgl.hr/>

E-adresa: jgl@jgl.hr

Poduzeće ili sektor: Stručna, Znanstvena I Tehnička Djelatnost

- ispitivanje svojstava lijekova nakon određenog broja mjeseci (3M, 6M, 9M, 12M, 24M, 36M) kako bi se dobio uvid u promjenu njihovih osobina i/ili sastava i donio zaključak o njihovoj sigurnosti
- mjerenje osmolalnosti, dostupnosti volumena, broja pritisa, mase kapi, pH, viskoznosti..

STUDENTSKI LABORATORIJSKI ASISTENT KOLEGIJA ANALITIČKA KEMIJA

Odjel za biotehnologiju, Sveučilište u Rijeci [17/05/2023 – 31/05/2023]

Adresa: Radmile Matejčić 2, 51000 Rijeka (Hrvatska)

Internetska stranica: <https://www.biotech.uniri.hr/hr/>

Poduzeće ili sektor: Stručna, Znanstvena I Tehnička Djelatnost

- priprema potrebne laboratorijske opreme i kemikalija, nadzor i pružanje pomoći studentima prilikom izvođenja laboratorijskih vježbi
- baždarenje laboratorijskog pribora, kvalitativna analiza kationa, aniona i soli, kvantitativna kemijska analiza (neutralimetrija, kompleksometrijske titracije, taložne titracije, oksido-redukcijske titracije), elektroanalitičke metode (potenciometrijska titracija), spektroskopska analiza (UV-VIS), separacijske i kromatografske tehnike (ekstrakcija, TLC)

STRUČNA PRAKSA

Institut Ruđer Bošković, Laboratorij za strojno učenje i reprezentaciju znanja [27/02/2023 – 01/07/2023]

Adresa: Bijelnička Cesta 54, 10000 Zagreb (Hrvatska)

Internetska stranica: <https://www.irb.hr/>

Poduzeće ili sektor: Stručna, Znanstvena I Tehnička Djelatnost

Diplomski rad: The use of the AI tool Alphafold 2 for protein structure modeling (Uporaba AI alata Alphafolda 2 za modeliranje strukture proteina)

- korištenje *in silico* metoda (AlphaFold 2, Linux, PyMOL, ChimeraX, online alati)

STRUČNA PRAKSA

Ljekarna "PrimaPharme" [01/07/2020 – 17/07/2020]

Adresa: Ulica Marka Marulića 9, 52100 Pula (Hrvatska)

E-adresa: pula@primapharme.hr

Poduzeće ili sektor: Zdravstvo I Socijalna Skrb

- izrada pripravaka prema magistralnim receptima te izrada beloderma, rivanola, tekućeg pudera, lanolina, hidrogena, ulja..
- moguć uvid u dnevnik rada na upit

OBRAZOVANJE I OSPOSOBLJAVANJE

MAGISTAR BIOTEHNOLOGIJE U MEDICINI

Odjel za biotehnologiju, Sveučilište u Rijeci [03/10/2021 – Trenutačno]

Adresa: Radmile Matejčić 2, 51000 Rijeka (Hrvatska)

Internetska stranica: <https://www.biotech.uniri.hr/hr/>

Područja obrazovanja: Biotehnologija u medicini

Kolegiji:

- "Metode u DNA tehnologijama", "Genska terapija", "Prirodni spojevi i njihova upotreba u farmakologiji", "Razvoj i registracija lijeka", "Personalizirana medicina", "Koloidi", "Tijevno inženjerstvo", "Stanična terapija", "Sistemska biomedicina", "Nanomedicina", "Metode istraživanja proteina"
- "Molekularna biotehnologija", "Statistika i analiza znanstvenih rezultata", "Uvod u istraživački rad", "Dizajn biološki aktivnih molekula računalnim metodama"

Izborni kolegiji:

- "Genetika ponašanja", "Zakonodavstvo za lijekove", "Kliničko istraživanje u praksi", "Implantacijski materijali u kirurgiji središnjeg živčanog sustava", "Metodologija projektnog upravljanja", "Bioinformatika"

SVEUČILIŠNA PRVOSTUPNICA BIOTEHNOLOGIJE I ISTRAŽIVANJA LIJEKOVA

Odjel za biotehnologiju, Sveučilište u Rijeci [01/09/2017 – 02/09/2021]

Adresa: Radmile Matejčić 2, 51000 Rijeka (Hrvatska)

Internetska stranica: <https://www.biotech.uniri.hr/hr/>

Područja obrazovanja: Biotehnologija i istraživanje lijekova

Konačna ocjena: 4.31 (na skali od 1.00 do 5.00)

Vrsta bodova: ECTS – Broj bodova: 180

Kolegiji:

- "Matematika s osnovama statistike", "Stanična i molekularna biologija", "Znanstvena komunikacija u engleskom jeziku", "Uvod u bioetiku", "Opća kemija", "Fizika", "Analitička kemija", "Informatika"
- "Uvod u bioanorgansku kemiju", "Organska kemija", "Mikrobiologija", "Biokemija", "Opća fiziologija i patofiziologija", "Farmakologija"
- "Imunologija", "Osnove biotehnologije istraživanja lijekova", "Uvod u fizikalnu kemiju", "Kemoinformatika", "Osnove molekularne medicine", "Bioeseji u istraživanju lijekova", "Farmakognozija i prirodni produkti", "Opća toksikologija"

Izborni kolegiji:

- "Drosophila kao model organizam u neuroznanosti", "Bakterijski organizmi u biotehnološkoj proizvodnji", "Predklinička istraživanja u razvoju lijeka", "Patofiziologija aktualnih javnozdravstvenih problema i bolesti", "Biofizika", "Kemijsko računanje u biotehnologiji", "Slobodni radikali u nama i antioksidativni sustavi oko nas"

SREDNJOŠKOLSKO OBRAZOVANJE

Gimnazija Pula [05/09/2011 – 20/05/2015]

Adresa: Trierska ulica 8, 52100 Pula (Hrvatska)

Internetska stranica: <https://www.gimnazijapula.hr/>

Područja obrazovanja: Opća gimnazija

Konačna ocjena: 4.75 (na skali od 1.00 do 5.00)

JEZIČNE VJEŠTINE

Materinski jezik/jezici: **hrvatski**

Drugi jezici:

engleski

SLUŠANJE C2 ČITANJE C1 PISANJE C1

GOVORNA PRODUKCIJA C2

GOVORNA INTERAKCIJA C2

njemački

SLUŠANJE B1 ČITANJE B1 PISANJE B1

GOVORNA PRODUKCIJA B1

GOVORNA INTERAKCIJA A2

Razine: A1 i A2: temeljni korisnik; B1 i B2: samostalni korisnik; C1 i C2: iskusni korisnik

DIGITALNE VJEŠTINE

Microsoft Teams / Google (Google Meet, Google Docs, Google Classroom, Google Forms, Google Drive, Google Slide) / Komunikacijski programi (Whatsapp, Slack, Viber, Skype, Zoom) / Microsoft Office (Microsoft Office Word, Microsoft Office Excel, Microsoft Office Powerpoint)

In silico metode analize

MacMolPlt / Cresset Spark / PyMOL / UCSF Chimera / GROMACS / Marvin / ChemAxon / AutoDock Vina / Linux / Li gPlot+ / VMD – Visual Molecular Dynamics / Marvin Sketch / R/R Studio (temeljni korisnik) / GAMESS / Avogadro

Statistika

Poznavanje rada u programima za statističku obradu podataka (Statistica, MedCalc) / Microsoft Excel

VOZAČKA DOZVOLA

Vozačka dozvola: B

POČASTI I NAGRADE

STIPENDIJA GRADA PULE

Grad Pula [09/01/2023]

Stipendija grada Pule za dodjelu studentskih stipendija za akademsku godinu 2022./2023 studentima koji prvi put upisuju akademsku godinu te su u prethodnoj godini studija ostvarili minimalan prosjek od 4.0.

Poveznica: <https://www.pula.hr/hr/>

STEM stipendija

Ministarstvo znanosti i obrazovanja [28/11/2019]

Državna stipendija za studente u području znanosti, tehnologije, inženjerstva i matematike.

Poveznica: <https://stem.mzo.hr/>

DSD I (Das Deutsches Sprachdiplom)

Njemačke obrazovne vlasti i Ministarstvo vanjskih poslova Savezne Republike Njemačke [27/08/2014]

Službeni certifikat njemačkog jezika njemačkih obrazovnih vlasti i Ministarstva vanjskih poslova Savezne Republike Njemačke kojim se potvrđuju razine znanja njemačkog jezika u srednjim školama. Program je namijenjen poticanju zanimanja za njemački jezik, a ispit se sastoji od čitanja, slušanja, pisanja i usmenog govora.

KOMUNIKACIJSKE I MEĐULJUDSKE VJEŠTINE

Studentski poslovi

Zahvaljujući sezonskim poslovima na kojima sam radila kao prodavačica i blagajnica uvelike sam unaprijedila svoje komunikacijske vještine i sposobnosti te rad s ljudima (kupcima i kolegama). Također već dvije godine radim u Općoj bolnici Pula kao ekonom gdje mi je odgovornost opskrbiti određene odjele sa potrebnom robom i steriliziranim materijalima. Smatram da sam odličan timski, ali i samostalan radnik te vrlo odgovorna, marljiva i sposobna osoba koja brzo uči i voli stjecati nova iskustva i koja je spremna na nove izazove.

VOLONTIRANJE

PROJEKT "Studenti mentori"

[Odjel za biotehnologiju, Sveučilište u Rijeci, 11/09/2018 – Trenutačno]

Zadaća mentora je pomoć pri uhodavanja bruoša (studenti prve godine preddiplomskoga studija) na studijski program, savjetovanje u vezi kolegija i nositelja kolegija, prezentacija postojećih projekata u koji se svaki od bruoša može priključiti, predstavljanje studentske udruge i prostorija udruge. Također, treba im se objasniti kako se bira predstavnik godine i koja je uloga istoga.

PROJEKT "Festival znanosti"

[Odjel za biotehnologiju, Sveučilište u Rijeci, 21/04/2018 – Trenutačno]

Tijekom "Dana Otvorenih vrata" volontirala sam na "Tetragonu" gdje sam skupine srednjoškolaca vodila kroz naš Odjel kako bi sudjelovali u ekipnom natjecanju iz područja matematike, informatike, fizike i biotehnologije. "Otvoreni dan Odjela za biotehnologiju" svake se godine odvija u travnju, u sklopu Festivala znanosti. Cilj "Otvorenih vrata" je popularizacija znanosti te je namijenjena posjetiteljima svih dobi. Neke od radionica koje posjetitelji mogu obići su "Igraonica za najmlađe", "Šarenilo kemije", "Mikroskopiranje", "Svjetleće stanice", "Biotehnologija u prehrani" i slično. Otvorena su i vrata znanstveno-istraživačkih laboratorija kao što su "Laboratorij za hematopoezu", "Laboratorij za molekularnu neurobiologiju", "Laboratorij za genetiku ponašanja" itd.

DODATNO OBRAZOVANJE

Internacionalni krunski kolegij "Patofiziologija aktualnih javnozdravstvenih problema i bolesti"

[08/06/2020 – 24/06/2020]

Upisala sam izborni kolegij (tzv. international capstone course) koji se održavao sa studentima koji pohađaju "St. Cloud State University (Minnesota, SAD)" i sa studentima "Odjela za biotehnologiju, Sveučilište u Rijeci", gdje je cilj bio učiti u timu, naučiti znanstveno pisati, povezati prethodno naučeno gradivo sa novim te usporediti javnozdravstvene, etiopatogenetske, farmakološke i epidemiološke postupke u Americi i Hrvatskoj. Osvojila sam 6 ECTS-a te kolegij položila ocjenom odličan (5).
