DOI: 10.14529/jsfi250408

Characterization of the Role of Amino Acid Residues in the 2-Hydroxybiphenyl 3-Monooxygenase Catalysis Based on Bioinformatic Analysis of the Flavin-dependent Monooxygenases and Supercomputer Modeling of the Structure of Mobile Fragments Applying Variational Autoencoders

Kirill E. Kopylov[1,2], ***Maxim A. Shchepetov***[2,3], ***Vytas K. Švedas***[1,2,3]

By modeling of predominant conformations of mobile loops in previously unresolved regions of 2-hydroxybiphenyl 3-monooxygenase structure (PDB ID: 5BRT) using GPU-accelerated metadynamics simulations integrated with artificial intelligence and high-performance computing the full-length protein model was built. Combined with bioinformatic analysis of the flavin-dependent monooxygenases it allowed to propose the functional role of amino acid residues in the 2-hydroxybiphenyl 3-monooxygenase catalysis. Three subfamily-specific residues Glu359, Lys339, Arg360 and the Asp332 residue, conservative throughout the entire family of flavin-dependent monooxygenases, form salt bridges Glu359-Lys339 and Arg360-Asp332, which stabilize alpha helices preserving the integrity of the Rossmann fold of the FAD-binding domain; subfamily-specific residues Trp338 and Glu359 provide the correct positioning of alpha-helices by interacting with two conservative residues Asp557 and Arg555 from the hydroxylase domain.

NAD binding pocket is formed by a number of subfamily-specific residues Trp38, Ser40, Ser42, Arg46, Ser47, Ala180, Asn205, Ser291, Trp293 located in an elongated pocket adjacent to the FAD binding site. The Asp313 residue, conservative in the entire family of flavin-dependent monooxygenases, directly interacts with FAD through hydrogen bonding with 2'-OH-ribitol, contributing to the binding and orientation of the cofactor. The Arg46, Ser47, Gly202, Ser203, Asn205, Arg242, Val253, Trp293, Met321, and Pro320, conservative for the entire family, play a crucial role forming the substrate binding site. The binding of cofactors and substrate in a quaternary complex and their orientation due to interactions with subfamily-specific positions Arg46, Ala180, His181 and Trp293 allows to perform the hydride transfer to the substrate stereospecifically. The triple stacking interaction between the FAD isoalloxazine ring, NADH nicotinamide ring and the subfamily-specific residue Trp293 leads to the formation of a highly stable charge-transfer complex and preferential Pro-S position in 2-hydroxybiphenyl 3-monooxygenase catalysis.

Keywords: flavin-dependent monooxygenases, 2-hydroxybiphenyl 3-monooxygenase from Pseudomonas azelaica, mobile loop structure prediction, full-length protein modeling, bioinformatics analysis, functional amino acid residues.

Introduction

Establishing the relationship between the structure and function of proteins, and enzymes in particular, is one of the most important tasks of modern biology. Despite the development of experimental techniques in structural biology (e.g., X-ray crystallography, NMR and cryo-EM), currently the PDB database [`https://rcsb.org`, `https://wwpdb.org` [3]] contains information on experimentally determined structures of 246,005 proteins (1,068,557 structures are presented in a separate section of Calculated Structure Models), although Uniprot [`https://uniprot.org` [1]] contains 199,006,239 amino acid sequences (mainly with unknown function, whereas

[1]Lomonosov Moscow State University, Belozersky Institute of Physicochemical Biology, Moscow, Russia
[2]Lomonosov Moscow State University, Research Computing Center, Moscow, Russia
[3]Lomonosov Moscow State University, Faculty of Bioengineering and Bioinformatics, Moscow, Russia

annotated are 573,661), and effective approaches need to be developed to bridge the gap between these information arrays.

Expectations of a breakthrough in this area are associated with the use of computer modeling and artificial intelligence. Indeed, the development of methods for predicting protein structure based on a known amino acid sequence has allowed to propose multiple spatial models (more than 200 million structures in the AlphaFold Protein Structure Database [`https://alphafold.ebi.ac.uk/`, [46]]), but the success of prediction is largely due to the presence of homologues of the protein under study and similar structures in training sets. The complexity of the problem also lies in the fact that many experimentally determined protein structures contain unresolved regions. Therefore, one of the not yet solved difficult problems is the prediction of the state of mobile elements, including loops, in the protein structure. The presence of such "blank spots" in the available information greatly complicates the search for the relationship between a full-length structure and a function. The mobile elements of the structure play an important role in the functioning of the protein/enzyme, ensuring its conformational plasticity and controlling access of the regulatory ligand or substrate to the complementary binding site, on the one hand, as well as their retention in a bound state, isolation of the formed complex from the environment and ensuring the specific microenvironment in the complex.

Since among the experimentally determined structures accumulated in the Protein Data Bank (PDB) over half contain unresolved segments (that often represent loops [38, 48]), accurate modeling of their structure is of high importance for preparing full-length protein models. Even in the post-AlphaFold era loop positioning remains one of the most challenging yet indispensable tasks in full-length protein modeling. Loops are often located on the protein surface, exhibit dynamic movement and adopt multiple distinct conformations that are not represented in PDB structures or adequately predicted [2, 30]. Accounting for loop dynamics has been recognized to be a key issue studying protein-ligand as well as protein-protein or antigen-antibody interaction and allosteric regulation, protein design, etc. [14, 16, 24, 29, 32]. Loop dynamics in enzyme structure has been recognized as an extremely important factor for catalytic activity, specificity as well as stability, evolution and design of improved enzyme variants [5, 7, 8, 25–28, 31, 49]. Flexible loop modeling using computational approaches in combination with artificial intelligence can provide an essential progress studying loop structural organization and their flexibility (appear to be open direction for exploration). Separate publications on the use of machine learning methods on this topic began to appear starting in the mid-80s, but a noticeable increase in research activity has been observed over the past 20 years (Fig. 1). These approaches allow to solve a number of problems and indicate the promise of supercomputer molecular modeling methods combined with the use of artificial intelligence approaches [21, 45].

Flavin-dependent monooxygenases are involved in a wide range of biological processes often playing a key role in the catabolism of natural and anthropogenic compounds or participating in the biosynthesis of numerous physiologically active compounds like hormones, vitamins and antibiotics [33]. Enzymes of this superfamily can be used as catalysts for large scale practical applications starting from lignin degradation or detoxification of xenobiotic compounds to the biocatalytic synthesis of key intermediates in pharmaceutical industry [9, 15]. The detailed mechanistic and structural studies of the FMO family enzymes are therefore of fundamental and practical interest. In this work the role of key amino acid residues in the catalytic mechanism of 2-hydroxybiphenyl 3-monooxygenase, a flavoprotein from *Pseudomonas azelaica* was analyzed

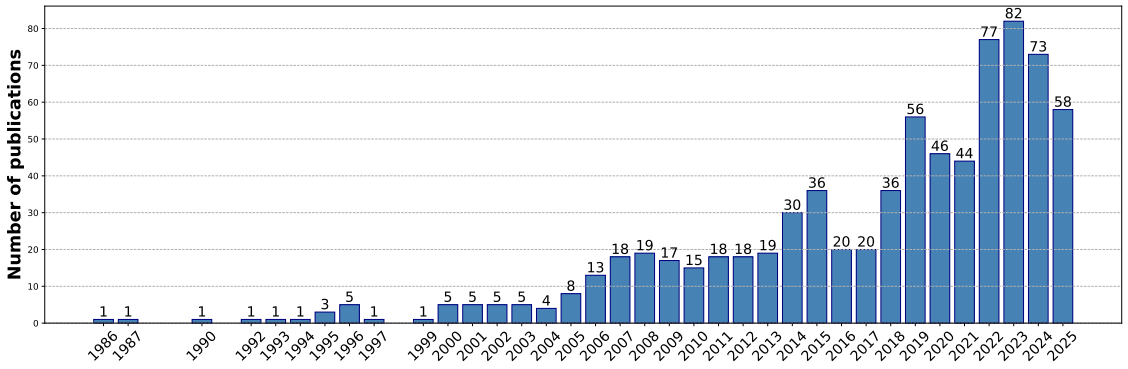

Figure 1. Publications on the topics of machine learning / artificial intelligence / neural networks and protein unresolved/non resolved/unstructured/non structured region modelling (PubMed)

based on the bioinformatics analysis of flavin-dependent monooxygenase family enzymes as well as supercomputer molecular modeling applying methods of artificial intelligence.

1. Materials and Methods

1.1. Bioinformatic Analysis

Bioinformatic analysis of a large representative set of flavin-dependent monooxygenases was used to study their structural and sequence variability by implementing a combination of protein sequence and structure comparison algorithms to account for structural and functional variability within this superfamily. The multiple structure-guided sequence alignment was constructed using the *Mustguseal* web-server (`https://biokinet.belozersky.msu.ru/mustguseal`) [42]. The server sequentially utilizes several bioinformatics algorithms: structural similarity search is used to find evolutionarily distant related proteins that have acquired functional/regulatory diversity due to significant changes in amino acid sequences and structures during natural selection (step 1); structural alignment to superpose the identified evolutionarily distant proteins that represent different families within a large superfamily (step 2); amino acid sequence similarity search to find evolutionarily close proteins, representatives of the selected families, and subsequent amino acid sequence alignment to superpose the sequences of evolutionarily close proteins (representatives of the same family, step 3). In the final step, the alignment of available structures of evolutionarily distant homologs is used as a framework for aligning the amino acid sequences of all collected representatives of the superfamily (step 4). Thus the selected PDB structures and their corresponding complete amino acid sequences were used to perform a structural alignment. The obtained structural core alignment was further used by the Mustguseal web-server to construct a larger alignment by incorporating all available sequences of related proteins. Each representative protein was used as a query for a sequence similarity search in Swiss-Prot and TrEMBL databases to collect its evolutionary close relatives. The obtained sequence sets were further filtered using the default parameters to remove redundant entries (at the 95% pairwise sequence identity threshold) within each family and superimposed using the core structural alignment of the representative proteins as a guide. The final structure-guided sequence alignment contained 4417 sequences and structures of monooxygenases with high structural, but low sequence similarity to 2-hydroxybiphenyl

3-monooxygenase from *Pseudomonas azelaica*. This representative set of homologs was automatically clustered into groups by maximizing sequence conservation within the groups and sequence variability between the groups using the *Zebra2* web-server (`https://biokinet.belozersky.msu.ru/zebra2`) [40]. The predicted clusters corresponded to different subfamilies of flavin-dependent monooxygenases that can be further analyzed according to the available functional annotation of members in each predicted group. Web-servers *pocketZebra* (`http://biokinet.cmm.msu.ru/index.php/pocketzebra`) and *visualCMAT* (`http://biokinet.cmm.msu.ru/visualcmat`) were used to find the NADH binding site and evaluate the function of residues in this site [39, 42].

1.2. Molecular Modeling

Neural network models with the architecture of hyperspherical variational autoencoders (S-VAE) were used as a component in modelling of the position of mobile loops. A critical choice in VAE design is the structure of the latent space, which significantly affects the model's capacity to represent underlying data properties [19]. For systems described by dihedral or torsional angles, such as molecular structures, a hyperspherical latent space offers an elegant solution [10]. The autoencoder was trained on the set of dihedral angles of unresolved loops, obtained by molecular dynamics. Dihedral spaces represent angular relationships between atoms in a molecular system. These angles, ranging from $[0, 2\pi)$, are inherently periodic. The periodicity implies that a dihedral space is better represented as a flat torus. The mathematical relationship between flat tori and hyperspheres further substantiates this choice. Embedding a flat torus in a hypersphere allows the model to naturally preserve properties such as smoothness and periodicity, ensuring that adjacent angular values in the data space (e.g., $2\pi - \epsilon$ and $0 + \epsilon$) remain adjacent in the latent representation [20]. The hypersphere inherently respects the periodic boundary conditions of these angles. This feature contrasts with Euclidean latent spaces, which are not naturally suited to handle periodicity or modular arithmetic, often leading to discontinuities or distortions in the latent representation.

The autoencoder is based on multiple narrowing convolutional layers and uses reparameterization of the three-dimensional von Mises-Fisher distribution. A weighted sum of the cumulative data reconstruction function (cosine distance) and Kullback–Leibler divergence is used as a metric for the error function. KL divergence assumes the hyperspherical nature of the latent space being calculated between the determined three-dimensional von Mises–Fisher distribution and a two-dimensional uniform hyperspherical distribution. The coefficient for accounting for KL divergence in the error function is 1.2, and the learning rate is $1 \cdot 10^{-2}$. The AdamW optimizer was used to train the model during 500 epochs. To find the optimal structure of protein sites, GPU accelerated metadynamics simulations were performed, using the Amber 20 molecular dynamics package [35] in conjunction with the Plumed 2.9 metadynamics package [12, 43]. The collective variable was defined using the `PYTORCH_MODEL` function [4], which consisted of compiled JIT PyTorch model (encoding part of the S-VAE) and the corresponding set of dihedral angles. Metadynamics was performed on three outputs of the model using a "well-tempered" scheme. The initial height of the Gaussian was set to 6.0 kJ/mol, the width to 0.06 for each measurement. A new potential was added every 400 steps. The bias factor was set to 10.0 and the simulation temperature to 300 K. For one computational experiment, 10 tasks ("walkers") were run simultaneously with a common metadynamics potential. All classical parameters for molecular dynamics modeling were as described above.

1.3. Mobile Loop Structure Prediction

The AlphaFold 2 model [17] was used to predict the structure of missing sites. An enzyme sequence from the UniProt database with the identifier O06647 was provided to the model. The model was run with standard settings in the "monomer" mode, producing 5 predicted structures using 5 different models from AlphaFold.

Rosetta [37] and MODELLER [13] were used as well for modeling of the missing sections in the 5BRT [18] structure. Using the MODELLER software, a Gln195-Ser201 segment from the 6EM0 [6] crystal structure was inserted into the original structure. Using Rosetta Remodel, a fully atomic model with the starting positions of the loops from the monomer of the protonated original structure was created. All protein atoms that were not included in the modeled regions were fixed in their positions, and the method of preserving the initial protein's protonation state was used. In the next step, 100 models of the 2-hydroxybiphenyl 3-monooxygenase monomer with alternative loop positions were generated using the Rosetta LoopModel. The coordinates of FAD and 2-hydroxybiphenyl were added to the model as user-defined residues. Next Generation KIC method was used to model 10000 loop conformations. The top 10 structures were selected based on Rosetta energy score and conformational reasonableness.

2. Results and Discussion

2.1. Bioinformatic Analysis

A search for homologous enzymes was performed using the CATH superfamily database [11]. The 2-hydroxybiphenyl 3-monooxygenase belongs to the CATH 3.50.50.60 superfamily. Since the CATH database is updated with a delay, a structural similarity search in the PDB database was additionally used to search for new homologues of the 2-hydroxybiphenyl 3-monooxygenase enzyme using the superpose algorithm (PDBeFOLD service) [23]. For comparative analysis of specific proteins, pairwise structural alignments were constructed using the Combinatorial Extension method [36]. The PDB structure 5BRT of 2-hydroxybiphenyl 3-monooxygenase, a flavoprotein from *Pseudomonas azelaica*, was used as a query for a structure similarity search in the PDB database to collect a nonredundant set of 16 proteins which shared 20% pairwise sequence similarity with 5BRT and represented different families within the superfamily (8FHJ,2R0G,3IHG,4K2X,2DKH,6UI5,4ICY,7YJ0,1PN0,6AIN,8JQP,3GMB,5EOW,6SW2,5TUI). Bioinformatic analysis revealed six distinct subfamilies of enzymes within the flavin-dependent monooxygenase family. The number of identified subfamilies corresponds well to the classification of external flavoprotein monooxygenases developed two decades ago on the basis of amino acid sequence, tertiary structure and cofactor preference when the 309 annotated bacterial genomes available at that time in the NCBI (`http://www.ncbi.nlm.nih.gov/BLAST/`) were screened for the presence of monooxygenase homologs using the prototype protein sequence and the BLAST tool. Review by Van Berkel W.J., Kamerbeek N.M. and Fraaije M.W. provided an inventory of known flavoprotein monooxygenases belonging to these different subclasses and highlighted the biocatalytic potential of this family of enzymes [44].

Currently, thanks to the accumulation of vast amounts of structural data, including even unannotated protein sequences, it has become possible to investigate distant evolutionary relationships within protein superfamilies and identify key amino acid residues critical to the entire family, as well as variable residues responsible for functional divergence and being specific to subfamilies of homologous enzymes. Thus, using bioinformatics analysis of protein

families, it is possible to identify amino acid residues conserved across the entire family or only within individual subfamilies (subfamily-specific positions in multiple structure-based sequence alignments) [40]. As a result, 21 amino acid residues were identified as conservative ones for the entire family (eight Gly residues in positions 13, 15, 18, 62, 176, 179, 312, 325, three Leu residues in positions 25, 336, 340, three Asp residues in positions 178, 313, 332, two Ala185 and Ala314, one residue of Ser182, Arg307, His316, Pro320 and Tyr356) (Fig. 2). These residues are located in the most structurally conservative FAD/NAD binding domain D1.

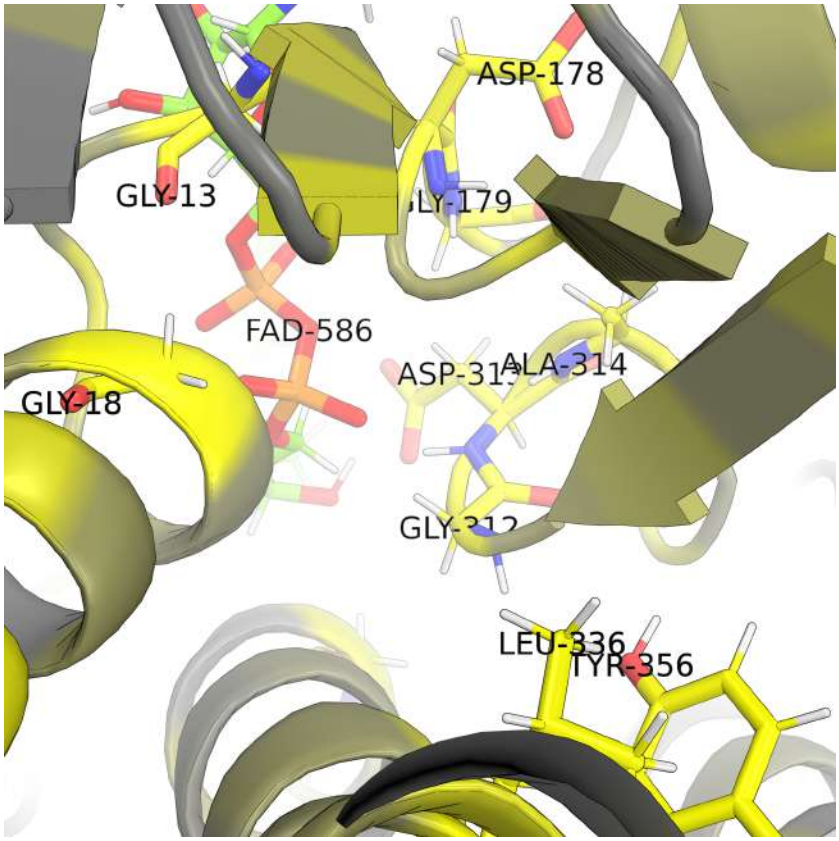

Figure 2. Conservative positions in the FAD-binding domain

2.2. Functional Role of Amino Acid Residues Identified by Bioinformatics Analysis. Binding of Substrates and Cofactors

Since the focus of this article is on the role of amino acid residues in the structure and catalytic mechanism of 2-hydroxybiphenyl 3-monooxygenase from *Pseudomonas azelaica*, we will discuss the information obtained from bioinformatic analysis in relation to this protein (Fig. 3).

Bioinformatic analysis takes into account the evolutionary relationships in protein families, and amino acid residues conservative in subfamilies may indicate functionally important, but variable positions in the protein structure.

The most typical for the entire family is a FAD binding subsite. On the one side, this area is bounded by two alpha-helices. Three subfamily-specific positions (Glu359, Lys339, Arg360) and the Asp332 residue, conservative throughout the entire family, form two structurally important salt bridges Glu359-Lys339 and Arg360-Asp332, which stabilize alpha helices in the dynamic

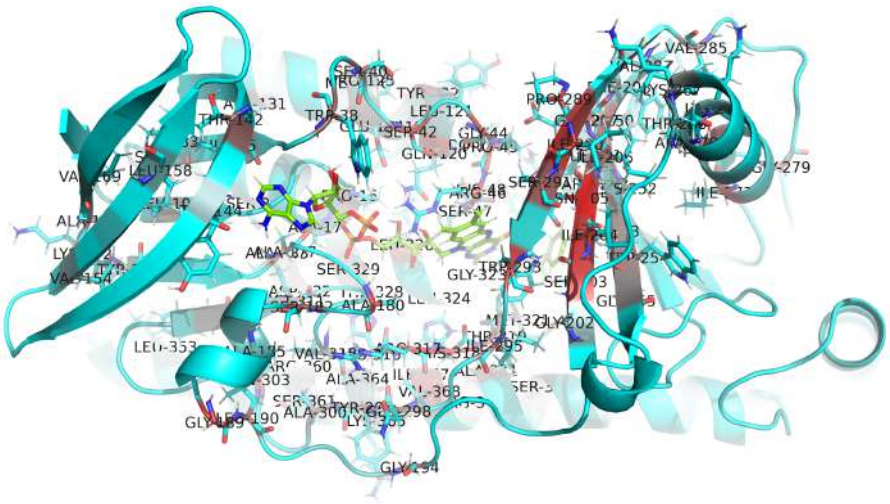

Figure 3. The subfamily-specific positions shown in 2-hydroxybiphenyl 3-monooxygenase from *Pseudomonas azelaica*

protein structure, preserving the integrity of the Rossmann fold of the FAD-binding domain D1 (Fig. 4). Such salt bridges are present in most subfamilies, however, in some cases Arg360 is replaced by Cys or Met, Lys339 by Cys, Ala or Glu, and Glu359 by Asn or Thr. One more subfamily-specific position, Trp338, participates in the correct positioning of these alpha-helices by interacting with the Asp557 along with the additional interaction of Glu359 with the Arg555 residue, i.e., two conservative residues from the hydroxylase domain.

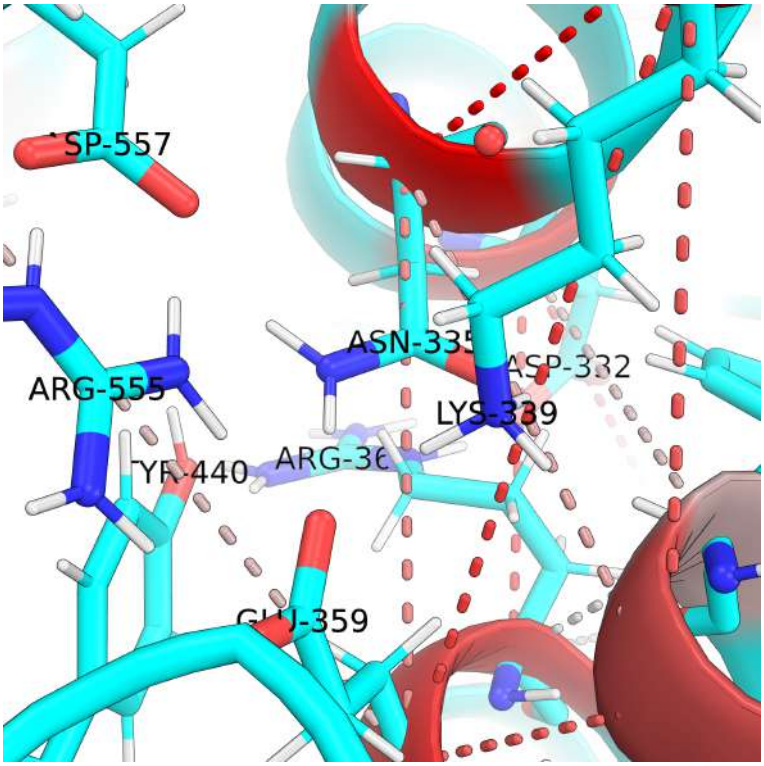

Figure 4. Salt bridges formed by subfamily-specific and correlating amino acid residues in 2-hydroxybiphenyl 3-monooxygenase from *Pseudomonas azelaica*

On the other side, in the conformationally flexible region limiting the FAD binding site, there are two glycine residues: Gly325, a conservative position for the entire family, and Gly323, a specific position for the 2-hydroxybiphenyl-3-monooxygenase subfamily. Molecular dynamics simulations indicate that their role is to provide conformational flexibility for the transfer of the isoalloxazine ring of FAD occurring in the catalytic mechanism. Only the Asp313 residue, which is conservative in the entire family, is directly interacting with the FAD molecule through hydrogen bonding with 2'-OH-ribitol, thereby contributing to the binding and orientation of the cofactor both in the static structure and in molecular dynamics simulations.

Earlier identified NAD binding pocket located next to the FAD binding subsite [22] is also formed by a number of subfamily-specific residues (Trp38, Ser40, Ser42, Arg46, Ser47, Ala180, Asn205, Ser291, Trp293). These residues are located in an elongated pocket adjacent to the FAD binding site, some of them are also involved in the formation of the substrate delivery tunnel (Fig. 5). Subfamily-specific positions Arg46, Ser47, Gly202, Ser203, Asn205, Arg242, Val253, Trp293, Met321, and Pro320, conservative for the entire family, play a crucial role forming the substrate binding site in 2-hydroxybiphenyl 3-monooxygenase structure. Functional role of some of them (Arg46, Arg242, Trp293) was earlier described [18, 22], the others (Ser201, Met223, Trp225, Val236, Val238, Ala240, Trp 254, Leu 375, Leu428) have not been observed in the course of bioinformatic analysis.

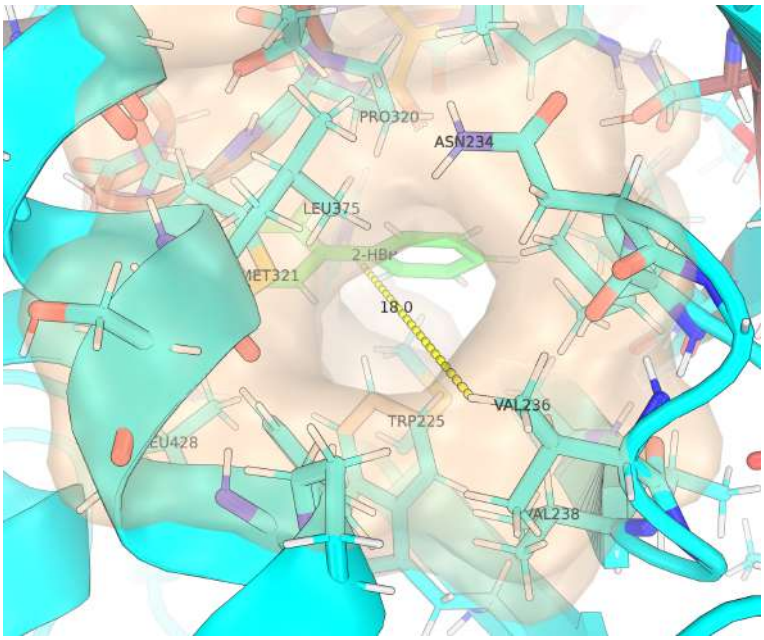

Figure 5. A tunnel for the entrance of the substrate to the active site of 2-hydroxybiphenyl 3-monooxygenase from *Pseudomonas azelaica*. The sticks highlight functionally important residues. The green color shows the substrate 2-hydroxybiphenyl

2.3. Subfamily-specific Amino Acid Residues in a Catalytic Mechanism

2-Hydroxybiphenyl 3-monooxygenase is capable to carry out stereospecific transformations because two hydrogen atoms at the C4 carbon of the dihydropyridine ring of NADH or NADPH as well as the sides of FAD isoalloxazine ring are not equivalent; the binding of cofactors and substrate in a quaternary complex and their orientation due to interactions with subfamily-specific positions Arg46, Ala180, His181 and Trp293 allows to perform the hydride

transfer to the substrate stereospecifically. In particular, the triple stacking interaction between the FAD isoalloxazine ring, NADH nicotinamide ring and the subfamily-specific residue Trp293 leads to the formation of a highly stable charge-transfer complex and preferential Pro-S position in 2-hydroxybiphenyl 3-monooxygenase catalysis. As in the course of catalytic conversion the Trp293 residue is pushed aside, forming a cavity for the NADH nicotinamide ring, it is important to control dynamic changes in the active site. The efficiency of the stacking interaction is regulated by the Arg46 and Trp293 residues, and their relationship with the catalytic His48 residue located in the substrate binding site controls the interaction between the centers of the oxidative and reduction stages of the reaction mechanism [22].

2.4. Flexible Elements of the Structure

Along with the structure of the binding sites of cofactors and substrates, mobile structural elements play an important role in the catalytic mechanism, ensuring enzymes conformational plasticity thus controlling entrance to the channels and delivery of substrates and cofactors to the active site, as well as their retention in a bound state, isolation from the environment and ensuring the unique microheterogenicity of the medium for the effective catalytic transformations. In most cases, when constructing a full-scale structural model of an enzyme, it is necessary to determine the state of the mobile loops so that molecular modeling of the catalytic stages of the enzymatic reaction can be carried out adequately. There are four mobile regions in the structure of the substrate binding domain of 2-hydroxybiphenyl 3-monooxygenase formed by the polypeptide chain fragments Gln195-Ser201, Arg219, Tyr256-Ile265, and Arg228-Val236. Three loops, Gln195-Ser201, Tyr256-Ile265, and Arg228-Val236, are located in close proximity to the active site (Fig. 6).

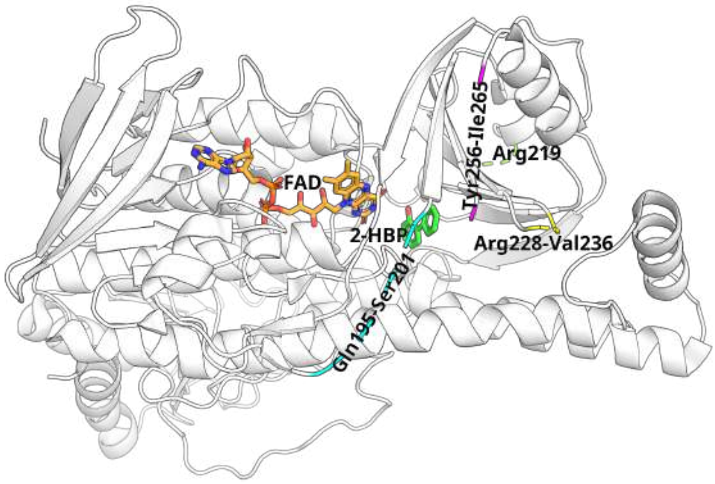

Figure 6. Unresolved structural fragments in the structure (PDB ID 5BRT) of 2-hydroxybiphenyl 3-monooxygenase from *Pseudomonas azelaica*: the blue color shows the *Gln195-Ser201* region, green – *Arg219*, pink – *Tyr256-Ile265*, yellow – *Arg228-Val236*. The FAD cofactor is shown in orange, the 2-hydroxybiphenyl substrate is shown in green

Full-length 2-hydroxybiphenyl 3-monooxygenase models with different positions of the loops predicted using different methods: LoopModel module of the Rosetta package, Modeller software, Alphafold v2.3.0 [13, 17, 34] are presented below (Fig. 7). Obviously, there is a

remarkable discrepancy in the predicted positions of flexible loops. Subfamily-specific residues are not present in these regions, however Gly194 and Gly202 as well as Gly255 and Thr266 are located next to flexible loops and can make a remarkable impact on conformational dynamics of 2-hydroxybiphenyl 3-monooxygenase structure. We suggested to apply an integrated approach combining metadynamics simulations in latent space using variational autoencoders on supercomputers to explore initial approximations of these flexible region structures derived from modeling tools such as AlphaFold, RosettaFold, Modeller, SwissModel, etc. [21]. The predominant conformations of previously unresolved mobile regions in the active site of flavin-dependent 2-hydroxybiphenyl 3-monooxygenase identified using this approach are shown in Fig. 7.

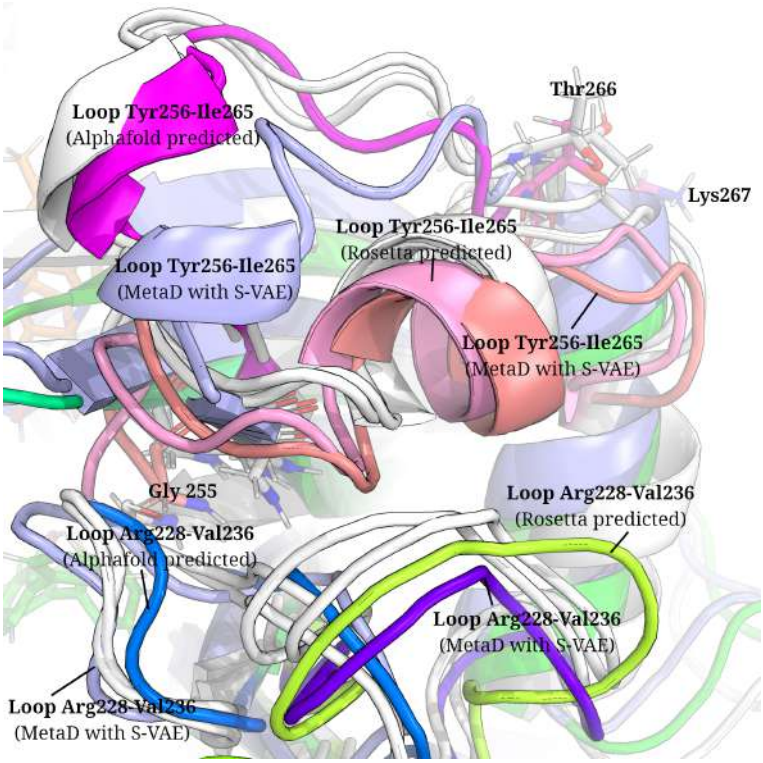

Figure 7. The positions of the loops in the active site of 2-hydroxybiphenyl 3-monooxygenase from *Pseudomonas azelaica*, predicted by Alphafold and Rosetta and optimized using molecular dynamics with hyperspherical variational autoencoder (S-VAE)

The coevolutionary relationship between the subfamily-specific residues Arg46 and Trp293, which regulate the efficiency of the stacking interaction between FAD and NADH was established using the *visualCMAT* server [41], as well as their relationship with the catalytic His48 residue located in the substrate binding site, which controls the interaction between the centers of the oxidation and reduction stages of the reaction mechanism. The important role of amino acid residues of the active site in the stabilization of the transition state should be noted: Trp293 forms a stacking interaction with the NADH nicotinamide ring, and His318 in the protonated form participates in stabilization due to the hydrogen bond formed between the backbone carbonyl oxygen and the NADH amide group. However, the catalytic mechanism of 2-hydroxybiphenyl 3-monooxygenase has not been fully understood yet. From two parts of the catalytic cycle, the reduction and oxidation half-reaction, only the first one was characterized on a molecular level: the reduction of FAD with NADH leading to the formation of $FADH^-$ anion,

i.e., the step when hydrogen atoms are transferred between the coenzymes and the substrate. Stabilization of the transition state in 2-hydroxybiphenyl 3-monooxygenase catalysis occurs due to the induced fit formation of a specific pocket for binding the nicotinamide ring of NADH, which is connected to the active site by interacting subfamily-specific residues Arg46 and Trp293 [22]. Thus, bioinformatic analysis has helped to elucidate or suggest previously unknown roles for some amino acid residues, as well as to identify a list of residues that appear to play an important role in the evolution of this enzyme family, although the influence of these positions on the structural and functional properties is still unknown and should be investigated.

Conclusion

The bioinformatic analysis of the flavin-dependent monooxygenases used together with the modeling of predominant conformations of mobile loops in previously unresolved regions of 2-hydroxybiphenyl 3-monooxygenase structure using GPU-accelerated metadynamics simulations integrated with artificial intelligence and high-performance computing allowed to elucidate previously unknown functional roles for some amino acid residues in the reduction half-reaction. Three subfamily-specific residues Glu359, Lys339, Arg360 and the Asp332 residue, conservative throughout the entire enzyme family, were shown to form structurally important salt bridges Glu359-Lys339 and Arg360-Asp332, which stabilize alpha helices preserving the integrity of the Rossmann fold of the FAD-binding domain; subfamily-specific residues Trp338 and Glu359 were shown to provide the correct positioning of alpha-helices by interacting with two conservative residues Asp557 and Arg555 from the hydroxylase domain. Subfamily-specific residues Trp38, Ser40, Ser42, Arg46, Ser47,Ala180, Asn205, Ser291, Trp293 form the elongated NAD binding pocket adjacent to the FAD binding site. The conservative Asp313 residue makes important contribution to the binding and orientation of the cofactor through hydrogen bonding with 2'-OH-ribitol of FAD. The Arg46, Ser47, Gly202, Ser203, Asn205, Arg242, Val253, Trp293, Met321, and conservative Pro320 residue play a crucial role forming substrate binding site. The binding of cofactors and substrate in a quaternary complex and their orientation due to interactions with subfamily-specific residues Arg46, Ala180, His181 and Trp293 allows to perform the hydride transfer to the substrate stereospecifically. The triple stacking interaction between the FAD isoalloxazine ring, NADH nicotinamide ring and the subfamily-specific residue Trp293 leads to the formation of a highly stable charge-transfer complex and preferential Pro-S position in 2-hydroxybiphenyl 3-monooxygenase catalysis. A primary challenge for the future remains the role of amino acid residues in the mechanism of the oxidation half-reaction of the complete catalytic cycle of 2-hydroxybiphenyl 3-monooxygenase.

Acknowledgements

The study was conducted under the state assignments of Lomonosov Moscow State University for Belozersky Institute of Physicochemical Biology, topic 119042590056-2, and for Research Computing Center, topic 121011690003-6. The research was carried out using the equipment of the shared research facilities of HPC computing resources at Lomonosov Moscow State University [47].

References

1. Apweiler, R., Bairoch, A., Wu, C.H., *et al.*: UniProt: the Universal Protein knowledgebase. Nucleic Acids Res 32(Database issue), D115–D119 (Jan 2004). https://doi.org/10.1093/nar/gkh131

2. Barozet, A., Bianciotto, M., Vaisset, M., *et al.*: Protein loops with multiple meta-stable conformations: A challenge for sampling and scoring methods. Proteins 89(2), 218–231 (Feb 2021). https://doi.org/10.1002/prot.26008

3. Berman, H.M., Westbrook, J., Feng, Z., *et al.*: The Protein Data Bank. Nucleic Acids Research 28(1), 235–242 (Jan 2000). https://doi.org/10.1093/nar/28.1.235

4. Bonati, L., Rizzi, V., Parrinello, M.: Data-driven collective variables for enhanced sampling. The Journal of Physical Chemistry Letters 11(8), 2998–3004 (2020). https://doi.org/10.1021/acs.jpclett.0c00535

5. Borges, P.T., Brissos, V., Hernandez, G., *et al.*: Methionine-Rich Loop of Multicopper Oxidase McoA Follows Open-to-Close Transitions with a Role in Enzyme Catalysis. ACS Catal. 10(13), 7162–7176 (Jul 2020). https://doi.org/10.1021/acscatal.0c01623

6. Bregman-Cohen, A., Deri, B., Maimon, S., *et al.*: Altering 2-Hydroxybiphenyl 3-Monooxygenase Regioselectivity by Protein Engineering for the Production of a New Antioxidant. ChemBioChem 19(6), 583–590 (2018). https://doi.org/10.1002/cbic.201700648

7. Corbella, M., Pinto, G.P., Kamerlin, S.C.L.: Loop dynamics and the evolution of enzyme activity. Nat Rev Chem 7(8), 536–547 (Aug 2023). https://doi.org/10.1038/s41570-023-00495-w

8. Crean, R.M., Biler, M., van der Kamp, M.W., *et al.*: Loop Dynamics and Enzyme Catalysis in Protein Tyrosine Phosphatases. J. Am. Chem. Soc. 143(10), 3830–3845 (Mar 2021). https://doi.org/10.1021/jacs.0c11806

9. Crozier-Reabe, K., Moran, G.R.: Form follows function: structural and catalytic variation in the class A flavoprotein monooxygenases. International Journal of Molecular Sciences 13(12), 15601–15639 (2012). https://doi.org/10.3390/ijms131215601

10. Davidson, T.R., Falorsi, L., De Cao, N., *et al.*: Hyperspherical variational auto-encoders. arXiv preprint arXiv:1804.00891 (2018). https://doi.org/10.48550/arXiv.1804.00891

11. Dawson, N.L., Lewis, T.E., Das, S., *et al.*: CATH: an expanded resource to predict protein function through structure and sequence. Nucleic Acids Research 45(D1), D289–D295 (2017). https://doi.org/10.1093/nar/gkw1098

12. Drobot, V.V., Kirilin, E.M., Kopylov, K.E., Švedas, V.K.: PLUMED plugin integration into high performance pmemd program for enhanced molecular dynamics simulations.

Supercomputing Frontiers and Innovations 8(4), 94–99 (2021). https://doi.org/10.14529/jsfi210408

13. Eswar, N., Webb, B., Marti-Renom, M.A., *et al.*: Comparative protein structure modeling using MODELLER. Current Protocols in Protein Science 50(1), 2–9 (2007). https://doi.org/10.1002/0471250953.bi0506s15

14. Fiser, A., Do, R.K.G., Šali, A.: Modeling of loops in protein structures. Protein Science 9(9), 1753–1773 (Jan 2000). https://doi.org/10.1110/ps.9.9.1753

15. Huijbers, M.M.E., Montersino, S., Westphal, A.H., *et al.*: Flavin dependent monooxygenases. Arch Biochem Biophys 544, 2–17 (Feb 2014). https://doi.org/10.1016/j.abb.2013.12.005

16. Jiang, H., Jude, K.M., Wu, K., *et al.*: De novo design of buttressed loops for sculpting protein functions. Nat Chem Biol 20(8), 974–980 (Aug 2024). https://doi.org/10.1038/s41589-024-01632-2

17. Jumper, J., Evans, R., Pritzel, A., *et al.*: Highly accurate protein structure prediction with AlphaFold. Nature 596(7873), 583–589 (2021). https://doi.org/10.1038/s41586-021-03819-2

18. Kanteev, M., Bregman-Cohen, A., Deri, B., *et al.*: A crystal structure of 2-hydroxybiphenyl 3-monooxygenase with bound substrate provides insights into the enzymatic mechanism. Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics 1854(12), 1906–1913 (2015). https://doi.org/10.1016/j.bbapap.2015.08.002

19. Kingma, D.P., Welling, M.: An Introduction to Variational Autoencoders. Foundations and Trends in Machine Learning 12(4), 307–392 (Nov 2019). https://doi.org/10.1561/2200000056

20. Kirilin, E.M., Švedas, V.K.: Study of the Conformational Variety of the Oligosaccharide Substrates of Neuraminidases from Pathogens using Molecular Modeling. Moscow Univ. Chem. Bull. 73(1), 39–45 (Jan 2018). https://doi.org/10.3103/S0027131418020050

21. Kopylov, K., Kirilin, E., Voevodin, V., Švedas, V.: Characterization of conformational flexibility in protein structures by applying artificial intelligence to molecular modeling. Journal of Structural Biology 217(2), 108204 (Jun 2025). https://doi.org/10.1016/j.jsb.2025.108204

22. Kopylov, K., Kirilin, E., Švedas, V.: Conformational transitions induced by NADH binding promote reduction half-reaction in 2-hydroxybiphenyl-3-monooxygenase catalytic cycle. Biochemical and Biophysical Research Communications 639, 77–83 (2023). https://doi.org/10.1016/j.bbrc.2022.11.066

23. Krissinel, E., Henrick, K.: Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallographica Section D: Biological Crystallography 60(12), 2256–2268 (2004). https://doi.org/10.1107/S0907444904026460

24. Kundert, K., Kortemme, T.: Computational design of structured loops for new protein functions. Biol Chem 400(3), 275–288 (Feb 2019). https://doi.org/10.1515/hsz-2018-0348

25. Li, Z., Meng, S., Nie, K., *et al.*: Flexibility Regulation of Loops Surrounding the Tunnel Entrance in Cytochrome P450 Enhanced Substrate Access Substantially. ACS Catal. 12(20), 12800–12808 (Oct 2022). https://doi.org/10.1021/acscatal.2c02258

26. Li, Z., Xie, D., Song, C., *et al.*: The open-closed transitions within dynamic conformational changes of enzyme loops. Systems Microbiology and Biomanufacturing 6(1), 2 (Nov 2025). https://doi.org/10.1007/s43393-025-00396-7

27. Liao, Q., Kulkarni, Y., Sengupta, U., *et al.*: Loop Motion in Triosephosphate Isomerase Is Not a Simple Open and Shut Case. J. Am. Chem. Soc. 140(46), 15889–15903 (Nov 2018). https://doi.org/10.1021/jacs.8b09378

28. Malabanan, M.M., Amyes, T.L., Richard, J.P.: A Role for Flexible Loops in Enzyme Catalysis. Curr Opin Struct Biol 20(6), 702–710 (Dec 2010). https://doi.org/10.1016/j.sbi.2010.09.005

29. Marks, C., Deane, C.M.: Antibody H3 Structure Prediction. Computational and Structural Biotechnology Journal 15, 222–231 (Jan 2017). https://doi.org/10.1016/j.csbj.2017.01.010

30. Marks, C., Shi, J., Deane, C.M.: Predicting loop conformational ensembles. Bioinformatics 34(6), 949–956 (Mar 2018). https://doi.org/10.1093/bioinformatics/btx718

31. Nestl, B.M., Hauer, B.: Engineering of Flexible Loops in Enzymes. ACS Catal. 4(9), 3201–3211 (Sep 2014). https://doi.org/10.1021/cs500325p

32. Papaleo, E., Saladino, G., Lambrughi, M., *et al.*: The Role of Protein Loops and Linkers in Conformational Dynamics and Allostery. Chem Rev 116(11), 6391–6423 (Jun 2016). https://doi.org/10.1021/acs.chemrev.5b00623

33. Reis, R.A.G., Li, H., Johnson, M., Sobrado, P.: New frontiers in flavin-dependent monooxygenases. Arch Biochem Biophys 699, 108765 (Mar 2021). https://doi.org/10.1016/j.abb.2021.108765

34. Rohl, C.A., Strauss, C.E., Misura, K.M., Baker, D.: Protein structure prediction using Rosetta. Methods in Enzymology 383, 66–93 (2004). https://doi.org/10.1016/S0076-6879(04)83004-0

35. Salomon-Ferrer, R., Götz, A.W., Poole, D., *et al.*: Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh Ewald. Journal of Chemical Theory and Computation 9(9), 3878–3888 (2013). https://doi.org/10.1021/ct400314y

36. Shindyalov, I.N., Bourne, P.E.: Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Engineering 11(9), 739–747 (1998). https://doi.org/10.1093/protein/11.9.739

37. Stein, A., Kortemme, T.: Improvements to Robotics-Inspired Conformational Sampling in Rosetta. PLoS One 8(5), e63090 (2013). https://doi.org/10.1371/journal.pone.0063090

38. Stevens, A.O., He, Y.: Benchmarking the Accuracy of AlphaFold 2 in Loop Structure Prediction. Biomolecules 12(7), 985 (Jul 2022). https://doi.org/10.3390/biom12070985

39. Suplatov, D., Kirilin, E., Arbatsky, M., *et al.*: pocketZebra: a web-server for automated selection and classification of subfamily-specific binding sites by bioinformatic analysis of diverse protein families. Nucleic Acids Research 42(W1), W344–W349 (2014). https://doi.org/10.1093/nar/gku448

40. Suplatov, D., Sharapova, Y., Geraseva, E., Švedas, V.: Zebra2: advanced and easy-to-use web-server for bioinformatic analysis of subfamily-specific and conserved positions in diverse protein superfamilies. Nucleic Acids Res 48(W1), W65–W71 (Jul 2020). https://doi.org/10.1093/nar/gkaa276

41. Suplatov, D., Sharapova, Y., Timonina, D., *et al.*: The visualCMAT: A web-server to select and interpret correlated mutations/co-evolving residues in protein families. Journal of Bioinformatics and Computational Biology 16(02), 1840005 (2018). https://doi.org/10.1142/S021972001840005X

42. Suplatov, D.A., Kopylov, K.E., Popova, N.N., *et al.*: Mustguseal: a server for multiple structure-guided sequence alignment of protein families. Bioinformatics 34(9), 1583–1585 (2018). https://doi.org/10.1093/bioinformatics/btx831

43. Tribello, G.A., Bonomi, M., Branduardi, D., *et al.*: PLUMED 2: New feathers for an old bird. Computer Physics Communications 185(2), 604–613 (2014). https://doi.org/10.1016/j.cpc.2013.09.018

44. Van Berkel, W.J.H., Kamerbeek, N.M., Fraaije, M.W.: Flavoprotein monooxygenases, a diverse class of oxidative biocatalysts. Journal of Biotechnology 124(4), 670–689 (2006). https://doi.org/10.1016/j.jbiotec.2006.03.044

45. Vander Meersche, Y., Cretin, G., Gheeraert, A., *et al.*: ATLAS: protein flexibility description from atomistic molecular dynamics simulations. Nucleic Acids Res 52(D1), D384–D392 (Jan 2024). https://doi.org/10.1093/nar/gkad1084

46. Varadi, M., Anyango, S., Deshpande, M., *et al.*: AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Research 50(D1), D439–D444 (Jan 2022). https://doi.org/10.1093/nar/gkab1061

47. Voevodin, V.V., Antonov, A.S., Nikitenko, D.A., *et al.*: Supercomputer Lomonosov-2: large scale, deep monitoring and fine analytics for the user community. Supercomputing Frontiers and Innovations 6(2), 4–11 (2019). https://doi.org/10.14529/jsfi190201

48. Wang, T., Wang, L., Zhang, X., *et al.*: Comprehensive assessment of protein loop modeling programs on large-scale datasets: prediction accuracy and efficiency. Briefings in Bioinformatics 25(1), bbad486 (Jan 2024). https://doi.org/10.1093/bib/bbad486

49. Zinovjev, K., Guénon, P., Ramos-Guzmán, C.A., *et al.*: Activation and friction in enzymatic loop opening and closing dynamics. Nat Commun 15(1), 2490 (Mar 2024). https://doi.org/10.1038/s41467-024-46723-9