DOI: 10.14529/jsfi220201

Computational Characterization of N-acetylaspartylglutamate Synthetase: From the Protein Primary Sequence to Plausible Catalytic Mechanism

Igor V. Polyakov[1,2], *Artem E. Kniga*[1,2], *Alexander V. Nemukhin*[1,2]

The methods of supercomputer molecular modeling are applied to characterize structure and dynamics of one of the key human brain enzymes, N-acetylaspartylglutamate synthetase. The three-dimensional all-atom models of the enzyme with the reactants in the active site are constructed in several steps, starting from pilot protein structure in the *apo*-form obtained with the AlphaFold2 from the protein primary sequence. Deposition of reactant molecules into the protein cavity, construction of the reaction intermediate and relaxation of the complex are carried out with the help of large-scale classical molecular dynamics calculations. On the top of the construct, molecular dynamics simulations with the quantum mechanics/molecular mechanics interaction potentials are performed for the most promising conformations of the model system. Analysis of the latter allows us to propose plausible catalytic mechanisms of chemical reactions in the enzyme active site. The applied computational strategy opens the way towards *ab initio* enzymology using modern supercomputer simulations.

Keywords: molecular dynamics, quantum mechanics/molecular mechanics, QM/MM MD, GPU-accelerated algorithms, N-acetylaspartylglutamate synthetase, enzyme-substrate complexes, reaction intermediates.

Introduction

High-performance computing plays an increasingly important role in life sciences, including simulations of chemical reactions in enzymes using advanced modeling methods based on the quantum mechanics/molecular mechanics (QM/MM) theory [1, 13, 14]. A practical goal of these simulations is to exploit the obtained information on structures and dynamics in protein systems for prediction of novel prospective drugs to fight human diseases [2]. Usually, such calculations are based upon the available experimental results, such as three-dimensional structures of macromolecules deposited in the Protein Data Bank (PDB) [6]. However, even in this case, raw crystallographic or nuclear magnetic resonance spectroscopy data undergo significant computational refinement [23] before deposition in the structure databanks.

The amount of available protein sequences and structural data enabled a number of tools to predict 3D protein structures [15] including the neural network-based models such as AlphaFold2 [12], which is claimed to be "demonstrating accuracy competitive with experimental structures in a majority of cases and greatly outperforming other methods" in the challenging 14th Critical Assessment of protein Structure Prediction (CASP14). However, very limited information if any is available, whether such structures are useful to enable accurate simulations using the QM/MM-based approaches to obtain enzymatic reaction mechanisms, locate reaction intermediates and evaluate energy profiles. Classical molecular dynamics (MD) simulations are routinely used to prepare structures for the QM/MM simulations and to refine the computationally predicted structures [10]. The downside of computational prediction of accurate protein structure and dynamics by using the all-atom classical MD is the amount of computational ef-

[1] Emanuel Institute of Biochemical Physics, Russian Academy of Sciences, Moscow, Russian Federation

[2] Department of Chemistry, Lomonosov Moscow State University, Moscow, Russian Federation

fort required to obtain reliable results, because a swarm of trajectories that cover microsecond timescales are needed [10].

On the upside, the recent developments in the MD simulation software with the GPU-resident version of NAMD 3.0 [19] promise great advances for the classical MD with the GPU-heavy supercomputers. Specifically, each trajectory can be efficiently calculated on a single GPU without significant load on the CPU and interconnect. Thus, many MD trajectories can be executed at the same time with high performance and efficient hardware utilization. The follower of the classical MD, i.e. molecular dynamics simulations with QM/MM potentials (QM/MM MD), benefit greatly from such GPU systems as well, if the NAMD/Terachem software stack is used [16, 22]. Original implementation of the NAMD QM/MM script interface to Terachem used to have some pitfalls, which have been recently fixed [13]. In this work, we use the high-performance classical MD and QM/MM MD calculations in order to characterize structure and dynamics of the acylphosphate reaction intermediate in the catalytic cycle of the key human brain enzyme, N-acetylaspartylglutamate synthetase (NAAGS), responsible for the formation of the most abundant brain dipeptide N-acetylaspartylglutamate (NAAG) [4]. This dipeptide acts as a retrograde neurotransmitter selectively localized in the glutamatergic synapses; it plays an essential role in cognition and memory consolidation underlying the novel object recognition task [5].

We have previously described computer simulations aimed to predict the reaction mechanism of the related enzyme, N-acetylglutamate synthase [20] and to characterize formation of a very abundant human neuropeptide N-acetylaspartate (NAA), one of the reactants in NAAG synthesis [21]. An essential feature of the present project is that no relevant structure of NAAG synthetase enzyme is available in the PDB. Therefore, we ought to apply modern computational methods to construct a full-atom three-dimensional structure of the enzyme on the base of its primary amino acid sequence and to deposit reaction species into the enzyme active site. This is a heavy computationally demanding task, which requires the use of supercomputer facilities.

A tentative scheme of NAAG synthesis is outlined in Fig. 1; however, no attempts are known to specify the reaction mechanism. Thus, our computationally derived data present the first approach to characterize this important process.

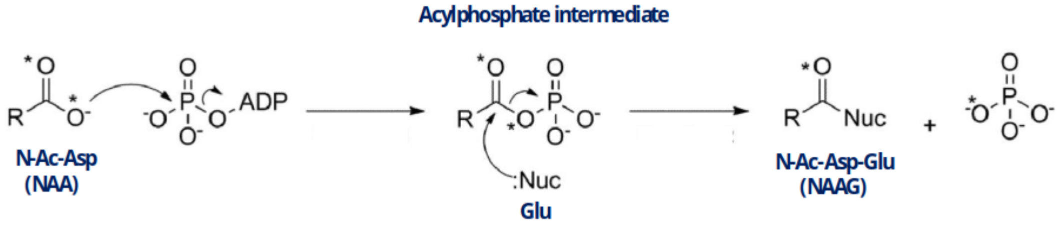

Figure 1. Expected reaction mechanism of the synthesis of N-acetylaspartylglutamate (NAAG).N-acetylaspartate (NAA), glutamate (Glu) as the nucleophile (Nuc) and adenosine triphosphate are the reactants

The left panel in Fig. 2 shows the model system to be used for simulation of the reaction. Chemical formulae of three reactants are presented in the right panels.

We describe, in this work, the initial step of the construction of the acylphosphate reaction intermediate (see the central panel in Fig. 1). Rationalization of structural and dynamical features of this intermediate is a necessary stage of the entire project, because attempts to build

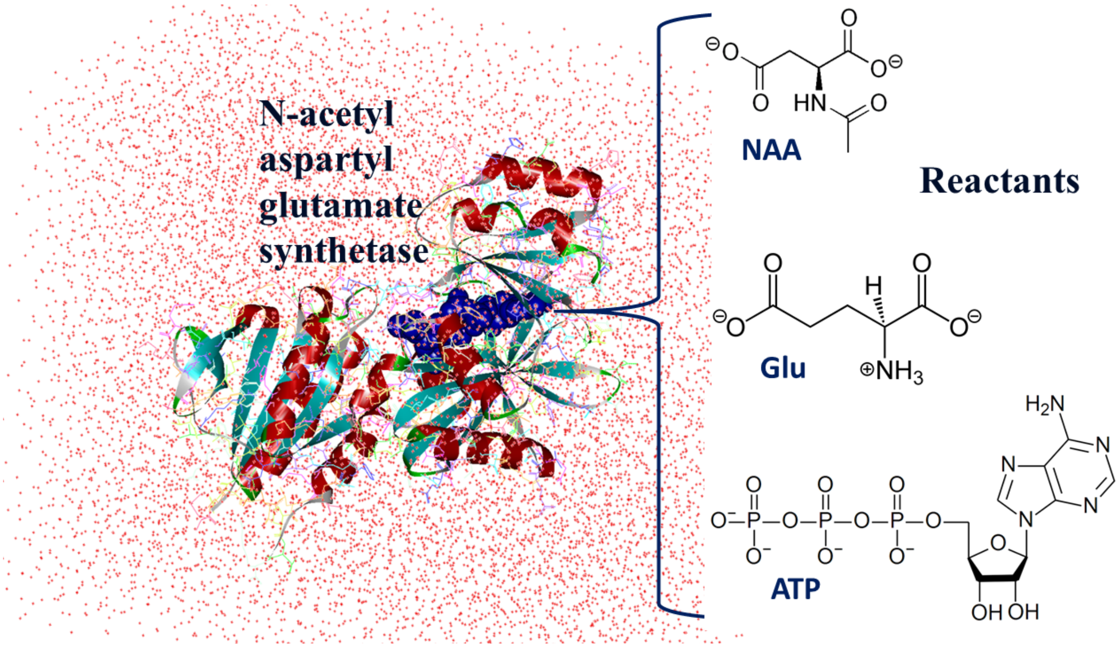

Figure 2. Computationally derived model system. Conventionally, α-helices of the protein are shown in red, β-sheets in cyan. The reactants are schematically shown as the dark blue space-filled spheres. The red dots surrounding the protein refer to solvation water shells

in silico enzyme-substrate complexes without knowledge of relevant structures from PDB would most probably lead to highly uncertain starting points in modelling the reaction mechanism.

1. Computational Methods

The primary amino acid sequence of the NAAGS sequence coded by the RIMKLA gene (Ribosomal Modification Protein RimK Like Family Member A) was taken from the UniProtKB [3] record Q8IXN7. Multiple sequence alignment was obtained with MMSEQ2 search [24] using a local installation of ColabFold pipeline featuring AlphaFold2 model for protein structure prediction [17]. The pipeline was run with a default preset of parameters for monomeric proteins.

The obtained AlphaFold model structure did not contain any ligands or water molecules; hence, an extensive active site reconstruction was required. To assist this step, the crystal structures PDB ID: 1GSA [9] and 2DLN [8] were taken as reference structures to construct the acylphosphate intermediate, namely, to introduce two magnesium ions, ADP, acylphosphate (NAA-PO_3) and Glu. We aligned and positioned ADP with the hydrogen bonds to Val192, Lys190, Gly161 main chain atoms; Gln189, Lys154, Asp199, Lys111 side chain atoms; magnesium ions and their coordination spheres with the contacts to ADP, NAA-PO_3, Glu273, Asp260, Asn275 and water molecules; NAA-PO_3 was coordinated by Arg160, Arg201, magnesium atoms and water molecules. Unlike for other ligands, no clear reference could be found for the glutamate position; hence, several starting positions were chosen to assist the C-N bond formation during the reaction progress (Fig. 1). Water molecules were initially added to the AlphaFold model by the Dowser++ [18] software. Several water molecules were removed to avoid clashes with the ligands in the active site. The model was solvated and made charge-neutral through adding the counter-ions with VMD [11]. The model contains 39346 atoms in total.

Classical molecular dynamics simulation with NAMD [19] was chosen as an approach to refine the designed structures before running the QM/MM MD calculations. Simulations were carried out assuming the isothermalisobaric (NPT) ensemble at P = 1 atm and T = 300 K using the Nos-Hoover Langevin piston pressure control and Langevin dynamics, integration step was set to 1 fs. Periodic boundary conditions along with the particle mesh Ewald method to account for the long-range electrostatic interactions were employed.

At first, harmonic restraint potential was applied to all protein backbone atoms to equilibrate the initial models. This is a standard practice [10] in the MD refinement protocols, the simulations were run for at least 100 ns each. All the constraints were released at the second step of the model refinement. At this step, we faced certain difficulties, because long trajectories ended up with the protein reorientation towards a shorter periodic cell dimension, which resulted in an artefact of MD simulations with the imposed boundary conditions, the formation of thread-like megastructures accounting for the periodicity. To overcome this difficulty, we restarted calculations from the constrained trajectories with most of the constraints released except for the CA atoms of the β-sheets (although the CA atoms of the β-sheets near the active site were constrained). In this approach, the protein conformational flexibility is greatly enhanced, as compared to initial restrained trajectories, but the periodicity artefacts can be avoided.

NAMD 3.0 was employed in order to fully harvest computational capabilities of the DGX2 supercomputer. A DGX2 node enabled us to run parallel swarms of trajectories of up to 16 simultaneously with all the GPUs and only 16 CPU cores utilized with the GPU-resident version of NAMD with CUDASOAintegrate option set to “on”. A total amount of 1.3 Tb of classical MD trajectories was produced, which accounts for 29 000 ns, whereas 1600 ns can be produced per day using all the Tesla V100 GPUs of a DGX2 node.

Four frames from the significantly different and stable classical MD trajectories were selected to start QM/MM MD simulations. The QM part contained 135 atoms described by the density function theory with the PBE0 hybrid functional, D3 dispersion corrections and 6-31G** basis set with 1395 basis functions in total. The QM system included NAA-PO_3, ADP (cut on the C4’-C5’ bond), two magnesium atoms; Arg160, Arg201, Arg215, Glu273, Asp260, Asn275 side chains and water molecules.

QM/MM simulations were run with the NAMD/Terachem [16, 22] software and the modified interface [13]. Combined length of computed QM/MM MD trajectories was over 50 ps, with a performance of ≈1.5 ps/day per GPU.

2. Analysis of the Designed Protein Structure

As described above, we designed computationally the model system in the conformation of the reaction intermediate containing phosphorylated NAA (see the central panel in Fig. 1). This structure is the best starting point for future simulations of the full energy profile, because attempts to begin construction from the enzyme-substrate complex, that is from the enzyme with the reactants (the left panel in Fig. 1), should be prohibitively expensive due to the expected huge conformation flexibility of the protein with reactants.

The model system used in simulations is composed of the protein with the embedded ligands, phosphorylated NAA, Glu, and ADP (see Fig. 1), surrounded by the shell of water molecules. The ligands are sandwiched between two β-strands formed by the amino acid residues #163-168, 185-191 on the one side and by five β-strands that include residues from the range of 197 to 277. The two loops consisting of residues #158–162 and 218-230 form a gate-like contact between these β-sheets.

Figure 3 illustrates a typical MD frame of the system showing the atoms of the reaction intermediate. The specified distances between Glu and NAA-PO_3 (the C-N distance, 3.4 Å), between NAA-PO_3 and ADP (the distance 3.1 Å between the phosphorus atom of the γ-phosphate group covalently bound to NAA and the oxygen atom of the β-phosphate), are given only to illustrate geometry parameters, which are fluctuating along MD trajectories.

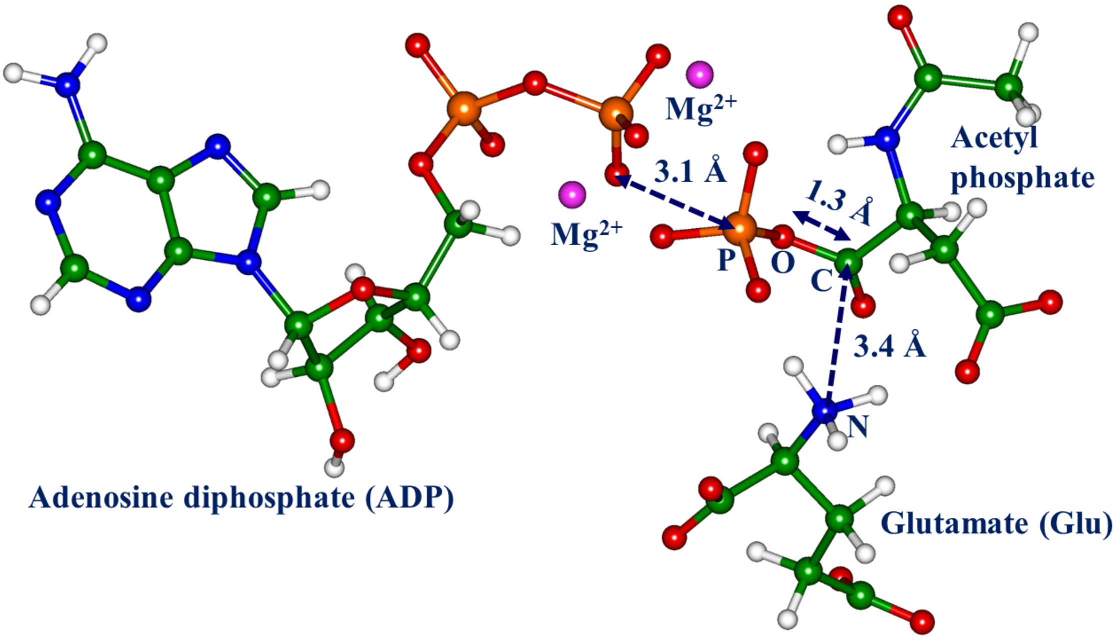

Figure 3. A structure of the reaction intermediate containing the acetylphosphate species with the phosphorylated NAA. The dashed dark blue arrows indicate the direction of nucleophilic attack along the N-C distance and the break of the C-O bond to form NAAGS, as well as the coordinate of the bond P-O breakage in the initial ATP molecule

Both visual MD trajectory inspection and trajectory RMSD (root mean square deviation) analysis (Fig. 4) show that the overall protein structure remains stable during all classical MD trajectories. The RMSD from the predicted model calculated over all the CA atoms is in the range of 3-3.5 Å.

Maintaining the original protein fold is not enough for the model to be useful in further QM/MM MD calculations, which strongly depend on the conformation of the active site. While all four presented trajectories show somewhat different but stable conformations within their active sites, only trajectories 1 and 2 maintain the active site conformation during 400 ns. Trajectories 3 and 4 show degradation of the active site. This is not evident from the Fig. 4, but the RMSF (root mean square fluctuation) analysis (Fig. 5) provides additional data. Many movements observed via the RMSF analysis were found in all trajectories, and do not contribute to destabilize configurations of the active site. For example, those are large deviations near the C- and N-terminus; the regions #132-150, 170-180, which are covered by the α-helices connected by loops to the β-strands: they are expected to be more flexible than the β-sheets. The residues #158-162 and 218-230 from a gate-like contact, connecting the β-sheet structures (Fig. 6), show the most variance (see also Fig. 5) for the trajectory 3, and this movement correlates with the active site destabilization. For the trajectory 4, no such correlation in the RMSF analysis is observed. The visual inspection of this trajectory reveals that the movement of the side chain of Arg160, which coordinates NAA-PO_3, is associated with the active site conformational change. In many other trajectories, which are not shown in the manuscript, the discussed issues, namely,

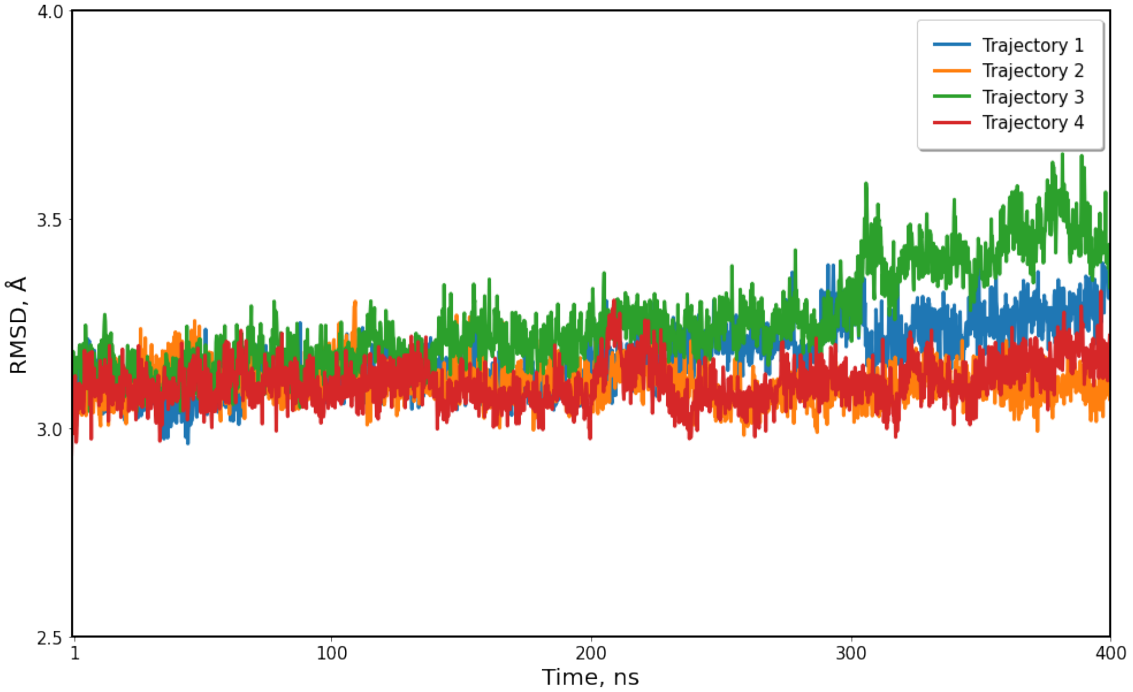

Figure 4. RMSD of the CA atoms compared to the original AlphaFold2 model during selected classical MD runs

lack of stable loop contact forming a gate or destabilization of Arg160 side chain contact to the NAA-PO_3, are associated with the active cite structure degradation or destruction. It is important to follow, whether Glu drifts away or the NAA-PO_3 position changes to the extent that prevents NAAG formation (Fig. 1).

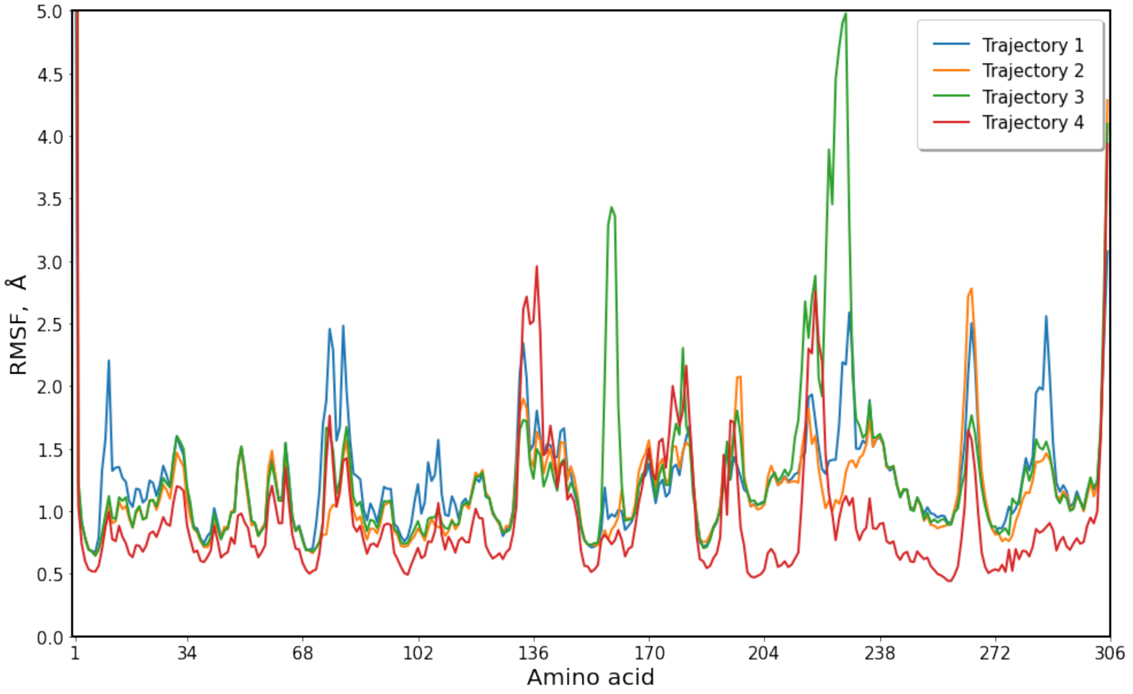

Figure 5. RMSF during selected classical MD runs

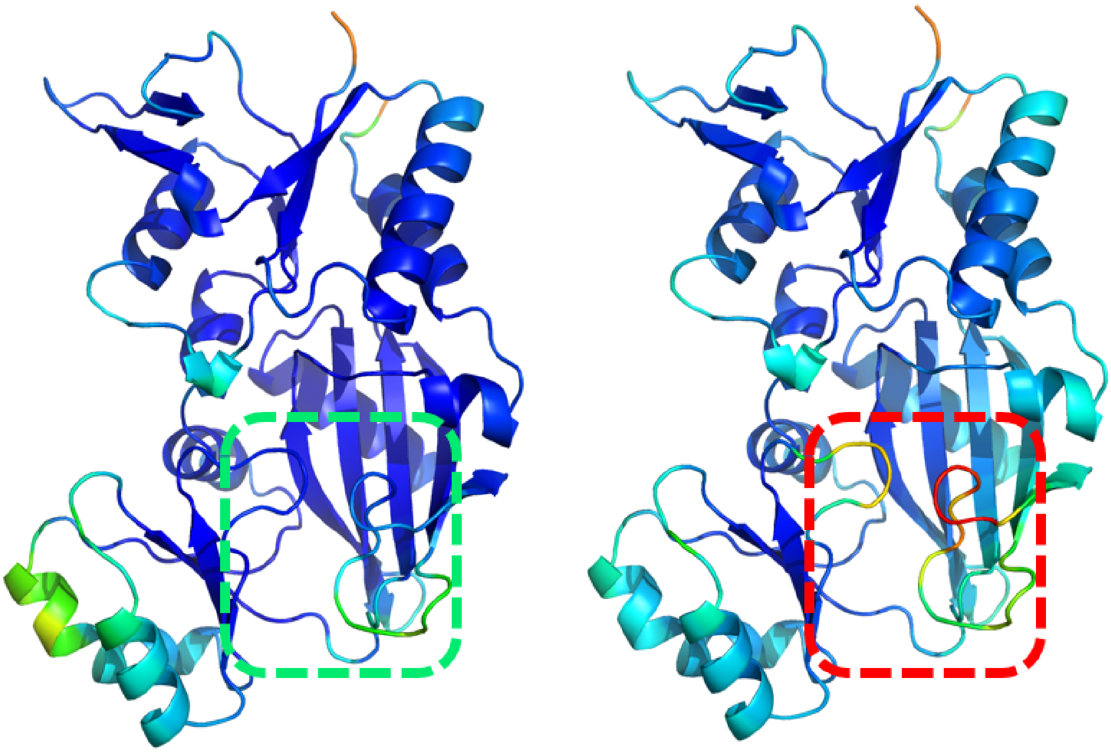

Figure 6. RMSF “coloring” of the 3D protein structure. Less RMSF – blue, cyan, green, yellow, red – more RMSF. Left – trajectory 4, right – trajectory 3. The green and right mark boxes highlight the region of the primary interest

3. Prospective Reaction Mechanism

Following results of the analysis of classical MD trajectories, we can proceed to QM/MM MD calculations. The corresponding frames were selected from those parts of classical trajectories that maintained desired conformations of the active site with the reacting species. The computed graphs in Fig. 7, as well as the data in Fig. 3, present a basis for the consideration of the reaction mechanism.

One of the key aspects of the active site structure of the model protein system is the C-N distance (Fig. 3) between the glutamate and the NAA-PO_3. The short and stable value of this distance during a QM/MM MD trajectory run (Fig. 7) would indicate that the nucleophilic attack (Fig. 1) is favorable given that the H_3N^+ group of glutamate is activated first. This can be achieved by removing a proton to a general base in the active site. Remarkably, there are three candidates for such an acceptor: two carboxyl groups of the attacking glutamate and the carboxyl group of the NAA-PO_3. There are no other carboxyl groups from Glu or Asp residues nearby, which are available to accept a proton. The phosphate groups of NAA-PO_3 and ADP should not be favorable proton acceptors having low pKa values and being coordinated by magnesium atoms in the protein active site. Thus, we speculate, that the NAAG formation reaction is an example of the substrate assisted catalysis [7].

It is important to note that no considerable conformational changes occurred during the short QM/MM MD trajectories except for Arg160 side chain adjustment in all the calculations. The variation of the critical C-N attack distance is rather large between the different trajectories (Fig. 7), ranging from the presumably reactive 3 Å (trajectory 3) to non-reactive 5 Å (trajectory 2). Thus, we note the importance of the computational details of the refinement procedure.

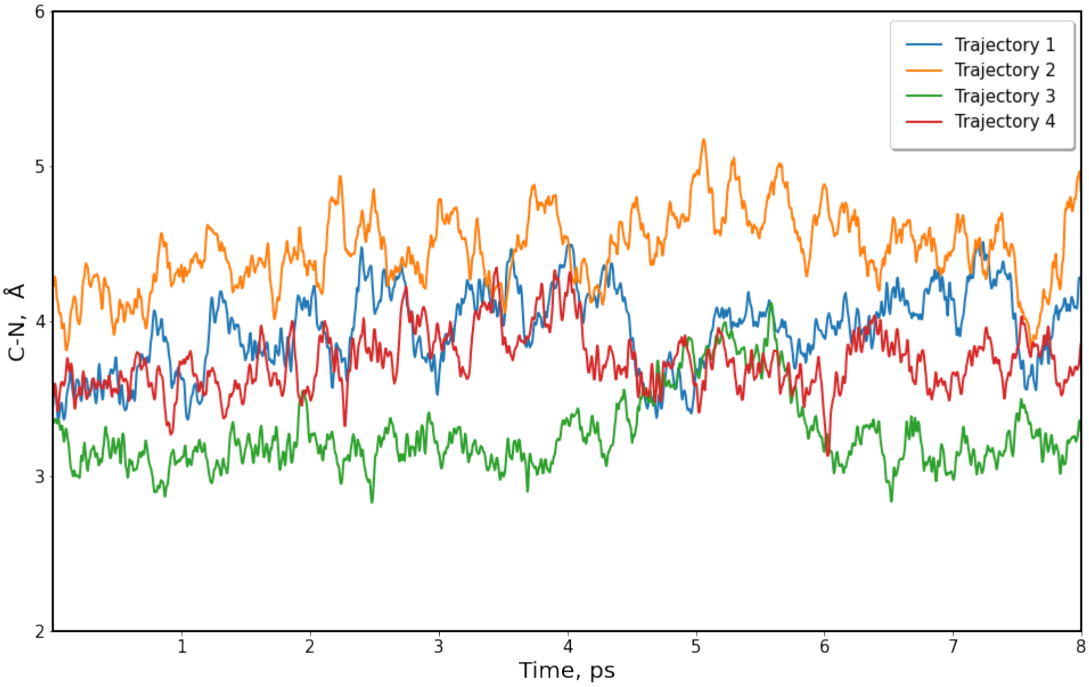

Figure 7. The nucleophilic attack C-N distance in the active site during the QM/MM trajectories. The 8 ps windows was obtained but cutting out the first 2 ps of each trajectory

In order to obtain a single model for QM/MM MD simulation of the reaction, a user should perform an enormous amount of attempts with different active site conformations and the corresponding trajectories. It is of utmost importance that such sampling procedure is carried out in parallel with an efficient software that utilizes modern supercomputer architectures, such as the software stack we described in the Computational approaches section.

Conclusion

In this paper, we describe a strategy to characterize *in silico* the enzyme catalysis, starting from a protein primary sequence without knowing other experimental data usually employed in such computer simulations. In particular, no relevant crystal structures are available in the Protein Data Bank for the adenosine triphosphate dependent binding of N-acetylaspartate and glutamate in the active of N-acetylaspartylglutamate synthetase (NAAGS). The primary sequence can be converted to a pilot structure of the NAAGS protein in the *apo*-form (that is, without reactants in the enzyme active site) with the help of the recently developed algorithms of AlphaFold2. The reacting species are inserted into the enzyme active site using the molecular modeling tools. The analysis of an *in silico* designed structure of NAAGS with the ligands shows that multiple manual corrections are required, which are introduced using the molecular modeling tools. The structure is refined using large-scale classical molecular dynamics simulations as well as molecular dynamics calculations with the QM/MM potentials. The performed analysis of the computationally designed complexes allows us to propose a reaction mechanism in this complicated enzyme-catalyzed chemical reaction, opening the way towards *ab initio* enzymology using modern supercomputer simulations.

Acknowledgements

We thank Russian Science Foundation (project No. 18-13-00030) for the financial support of this work. The study was carried out using the equipment of the shared research facilities of HPC computing resources at Lomonosov Moscow State University.

References

1. Ahmadi, S., Barrios Herrera, L., *et al.*: Multiscale modeling of enzymes: QM-cluster, QM/MM, and QM/MM/MD: A tutorial review. International Journal of Quantum Chemistry 118(9), e25558 (2018). `https://doi.org/10.1002/qua.25558`

2. Aminpour, M., Montemagno, C., Tuszynski, J.A.: An overview of molecular modeling for drug discovery with specific illustrative examples of applications. Molecules 24(9), 1693 (Apr 2019). `https://doi.org/10.3390/molecules24091693`

3. Apweiler, R.: UniProt: the Universal Protein knowledgebase. Nucleic Acids Research 32(90001), D115–D119 (Jan 2004). `https://doi.org/10.1093/nar/gkh131`

4. Becker, I., Lodder, J., Gieselmann, V., Eckhardt, M.: Molecular characterization of N-acetylaspartylglutamate Synthetase. Journal of Biological Chemistry 285(38), 29156–29164 (Sep 2010). `https://doi.org/10.1074/jbc.m110.111765`

5. Becker, I., Wang-Eckhardt, L., Lodder-Gadaczek, J., *et al.*: Mice deficient in the NAAG synthetase II gene Rimkla are impaired in a novel object recognition task. Journal of Neurochemistry 157(6), 2008–2023 (Mar 2021). `https://doi.org/10.1111/jnc.15333`

6. Berman, H.M.: The protein data bank. Nucleic Acids Research 28(1), 235–242 (Jan 2000). `https://doi.org/10.1093/nar/28.1.235`

7. Dall'Acqua, W., Carter, P.: Substrate-assisted catalysis: Molecular basis and biological significance. Protein Science 9(1), 1–9 (Dec 2008). `https://doi.org/10.1110/ps.9.1.1`

8. Fan, C., Moews, P.C., Walsh, C.T., Knox, J.R.: Vancomycin Resistance: Structure of D-Alanine:D-Alanine Ligase at 2.3 Å Resolution. Science 266(5184), 439–443 (Oct 1994). `https://doi.org/10.1126/science.7939684`

9. Hara, T., Kato, H., Katsube, Y., Oda, J.: A Pseudo-Michaelis Quaternary Complex in the Reverse Reaction of a Ligase: Structure of Escherichia coli B Glutathione Synthetase Complexed with ADP, Glutathione, and Sulfate at 2.0 Å Resolution. Biochemistry 35(37), 11967–11974 (1996). `https://doi.org/10.1021/bi9605245`

10. Heo, L., Feig, M.: Experimental accuracy in protein structure refinement via molecular dynamics simulations. Proceedings of the National Academy of Sciences 115(52), 13276–13281 (2018). `https://doi.org/10.1073/pnas.1811364115`

11. Humphrey, W., Dalke, A., Schulten, K.: VMD: Visual molecular dynamics. Journal of Molecular Graphics 14(1), 33–38 (Feb 1996). `https://doi.org/10.1016/0263-7855(96)00018-5`

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

13. Khrenova, M.G., Polyakov, I.V., Nemukhin, A.V.: Molecular dynamics of enzyme-substrate complexes in guanosine-binding proteins. Khimicheskaya Fizika 41(6), 66–72 (2022). https://doi.org/10.31857/S0207401X22060061

14. Khrenova, M.G., Bulavko, E.S., Mulashkin, F.D., Nemukhin, A.V.: Mechanism of Guanosine Triphosphate Hydrolysis by the Visual Proteins Arl3-RP2: Free Energy Reaction Profiles Computed with Ab Initio Type QM/MM Potentials. Molecules 26(13), 3998 (Jun 2021). https://doi.org/10.3390/molecules26133998

15. Kuhlman, B., Bradley, P.: Advances in protein structure prediction and design. Nature Reviews Molecular Cell Biology 20(11), 681–697 (Nov 2019). https://doi.org/10.1038/s41580-019-0163-x

16. Melo, M.C.R., Bernardi, R.C., Rudack, T., *et al.*: NAMD goes quantum: an integrative suite for hybrid simulations. Nature Methods 15(5), 351–354 (May 2018). https://doi.org/10.1038/nmeth.4638

17. Mirdita, M., Schtze, K., Moriwaki, Y., *et al.*: ColabFold: making protein folding accessible to all. Nature Methods 19(6), 679–682 (may 2022). https://doi.org/10.1038/s41592-022-01488-1

18. Morozenko, A., Stuchebrukhov, A.A.: Dowser++, a new method of hydrating protein structures. Proteins: Structure, Function, and Bioinformatics 84(10), 1347–1357 (Jul 2016). https://doi.org/10.1002/prot.25081

19. Phillips, J.C., Hardy, D.J., Maia, J.D.C., *et al.*: Scalable molecular dynamics on CPU and GPU architectures with NAMD. J. Chem. Phys. 153(4), 044130 (Jul 2020). https://doi.org/10.1063/5.0014475

20. Polyakov, I.V., Kniga, A.E., Grigorenko, B.L., *et al.*: Computer Modeling of N-Acetylglutamate Synthase: From Primary Structure to Elemental Stages of Catalysis. Doklady Biochemistry and Biophysics 495(1), 334–337 (Nov 2020). https://doi.org/10.1134/s1607672920060125

21. Polyakov, I.V., Kniga, A.E., Grigorenko, B.L., Nemukhin, A.V.: Structure of the brain N-acetylaspartate biosynthetic enzyme NAT8l revealed by computer modeling. ACS Chemical Neuroscience 11(15), 2296–2302 (Jul 2020). https://doi.org/10.1021/acschemneuro.0c00250

22. Seritan, S., Bannwarth, C., Fales, B.S., *et al.*: TeraChem: A graphical processing unit-accelerated electronic structure package for large-scale ab initio molecular dynamics. WIREs Computational Molecular Science 11(2) (Jul 2020). https://doi.org/10.1002/wcms.1494

23. Shabalin, I.G., Porebski, P.J., Minor, W.: Refining the macromolecular model – achieving the best agreement with the data from X-ray diffraction experiment. Crystallography Reviews 24(4), 236–262 (Sep 2018). https://doi.org/10.1080/0889311x.2018.1521805

24. Steinegger, M., Söding, J.: MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nature Biotechnology 35(11), 1026–1028 (Oct 2017). https://doi.org/10.1038/nbt.3988