New QM/MM implementation of the MOPAC2012 in the GROMACS

A.O. Zalevsky[1], R.V. Reshetnikov[2,3], and A.V. Golovin[1,3,4]

[1] Faculty of Bioengineering and Bioinformatics
[2] Institute of Gene Biology, Russian Academy of Sciences
[3] Apto-Pharm LLC
[4] Sechenov First Moscow State Medical University

Abstract. Hybrid QM/MM simulations augmented with enhanced sampling techniques proved to be advantageous in different usage scenarios, from studies of biological systems to drug and enzyme design. However, there are several factors that limit the applicability of the approach. First, typical biologically relevant systems are too large and hence computationally expensive for many QM methods. Second, a majority of fast non *ab initio* QM methods contain parameters for a very limited set of elements, which restrains their usage for applications involving radionuclides and other unusual compounds. Therefore, there is an incessant need for new tools which will expand both type and size of simulated objects. Here we present a novel combination of widely accepted molecular modelling packages GROMACS and MOPAC2012 and demonstrate its applicability for design of a catalytic antibody capable of organophosphorus compound hydrolysis.

Keywords: qmmm, hpc, molecular modelling, rational design

1 Introduction

Hybrid QM/MM simulations were introduced in 1976 [1] and only somewhat recently began to be actively used in the molecular dynamics (MD) simulations, becoming a popular tool for studying biomolecular systems. Now the method allows to gather a ms-scale statistics on protein dynamics, where thermal motion could significantly contribute to chemical reactivity and conformational space of the system [2]. An additional momentum to the rise of hybrid QM/MM applications in biosystems' studies was given by Parinello and colleagues who developed the "metadynamics" method [3–5]. This method allows to scan a conformational space of biomolecular systems searching for rare events, such as reactions catalyzed by protein enzymes. Generally, a quantum subsystem in metadynamic modeling of enzymatic reactions consists of up to 1000 atoms, and to simulate a thermal movement of the system, hundreds of thousands steps of energy gradients and geometry optimization calculations can be made. It requires considerable computing resources even at a low level of QM calculations. Optimization of computational tools for the task is the key factor affecting the progress of enzyme design and an engineering of other biopolymers with desired properties.

A number of software packages capable of hybrid simulations have been developed. The popular MD program package GROMACS [6] has a QM/MM interface to several quantum chemistry software tools. Significant drawback of the original interface implementation is that the data exchange between the programs occurs via the file system, which imposes a significant performance limitation when used on cluster computing systems with distributed storage environment. Here, we propose the modification of the widely accepted MOPAC-2012 [7] semi-empirical QM tool allowing to use it as a GROMACS library. The resulting software features a high speed computing with OpenMP and CUDA acceleration options for both QM and MM hybrid subsystems. An important advantage of the proposed implementation is a wide range of supported chemical elements and a variety of available semi-empirical QM parameters for biological systems.

2 Methods

2.1 ONIOM

The ONIOM approach [8] allows to divide a system into several layers with an independent description of intra-layer interactions. In QM/MM case, force gradients are first evaluated for the isolated QM subsystem using a selected semi-empirical or *ab initio* model. Next, the gradients and the total potential energy of the system are calculated using the corresponding MM force field and added to the ones obtained for the isolated QM subsystem. Finally, in order to avoid duplicated contribution from the QM subsystem a molecular mechanics calculation is performed for the isolated QM region and the result is subtracted from total sum:

$$E_{tot} = E_I^{QM} + E_{I+II}^{MM} - E_I^{MM} \tag{1}$$

where the subscripts I and II refer to the QM and MM subsystems, respectively. The superscripts indicate at what level of theory the energies are computed. Physical separation between systems is achieved by introduction of linking atoms (LA) which are rendered as hydrogens in quantum part and do not introduce additional interactions into mechanical part. [9].

2.2 Implementation of the QM/MM interface

Using the implementation of the GROMACS/ORCA interface as a reference, we modified corresponding parts of GROMACS (MD engine, input-output module and CMake configuration files) and MOPAC2012 (input-output and parameters verification modules). In the reference implementation the exchange between GROMACS and ORCA packages is performed through text files, which limits precision: data is printed and read with the “%10.7f” pattern, which provides only 7 decimal digits, corresponding to single precision, while internally ORCA uses double precision. Our implementation uses MOPAC2012, which was compiled and assembled into static library and directly linked into GROMACS static binary during compilation allowing direct data transfer between MOPAC2012 and GROMACS with double precision. To optimize the control of QM calculations, the `GMX_QM_MOPAC2012_KEYWORDS` environment variable was added, which holds MOPAC2012 keywords and is read at the initial simulation step (Figure 1).

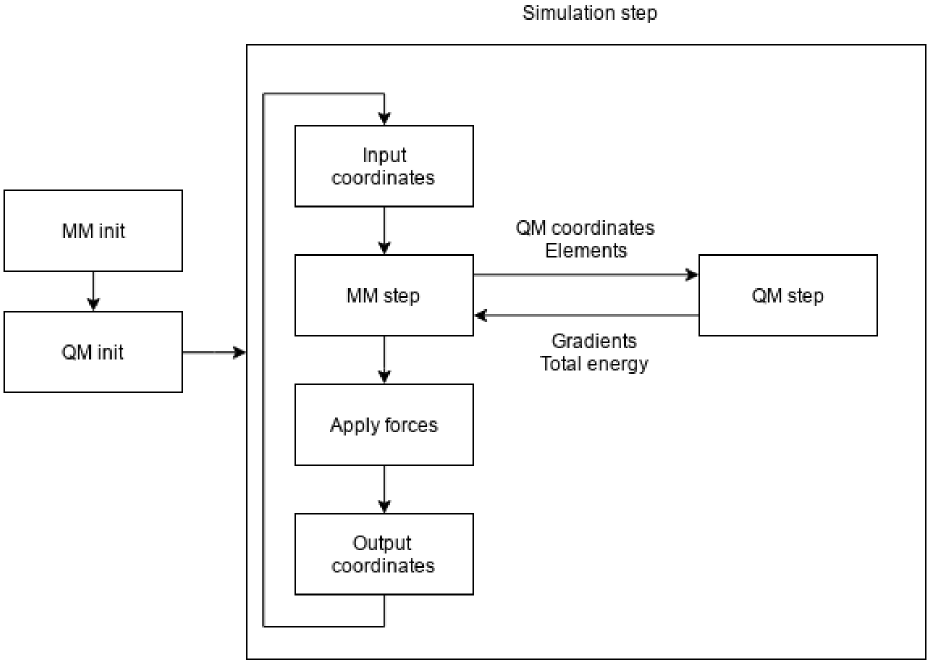


Fig. 1. Scheme of a generic QM/MM interface

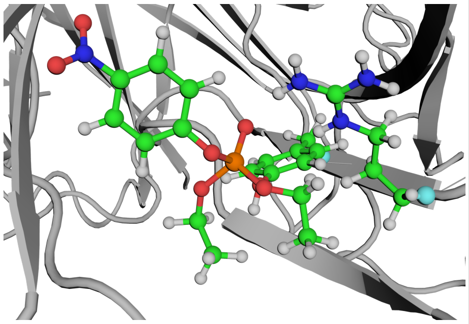
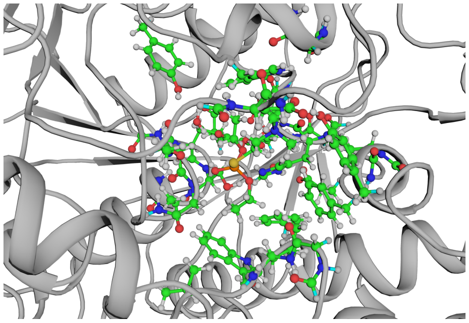
2.3 Building environment

Compilation of MOPAC2012 and GROMACS 5.0.7 into statically linked binaries was performed on local workstation (Intel(R) Core(TM) i7-6700K CPU @ 4.00GHz, NVIDIA GTX1070, 32Gb RAM, A-Data XPG SX8000NP 128Gb SSD, Ubuntu 16.04.4 amd64). Due to the MOPAC dependencies from the Intel®MKL libraries, Intel®Composer suite 2015 update 1 packaged with MKL version 11.2 was used. We also compiled GROMACS versions with GPU support (NVIDIA® CUDA toolkit version 7.5 on the local system and version 8.0 on the "Lomonosov-2" supercomputer). Several GROMACS/ORCA and GROMACS/MOPAC versions were prepared: single precision, single precision with GPU support and double precision (double precision with GPU is not supported). MPI support was disabled and OpenMP support was enabled in all versions. We used ORCA version 4.2.1 distributed as precompiled binaries.

2.4 Test systems

Performance test of the QM/MM tools was done on two model systems: "small" - artificial enzyme with paraoxone substrate [10] and "big" — complex of butyril choline esterase with echotiophate. The QM subsystem size was 64 and 408 atoms for the "small" and "big" systems, correspondingly.

Table 1. Composition of QM systems. Gray color represents MM part and colored QM part. Linking atoms are colored in cyan.

Small	**Big**
	
67 atoms 2 linking atoms	408 atoms 31 linking atoms

"Small" system The starting conformation of the protein-ligand enzymatic complex was taken from the studies described previously [11]. Simulation system was filled with TIP3P water molecules [12], the total charge was neutralized with Na^+ or Cl^- ions. Water and ions were equilibrated around the protein-paraoxon complex with a 100-ps MD simulation with a restrained positions of protein and paraoxon atoms. For the MM subsystem we used the parameters from parm99 force field with corrections[13]. The QM subsystem was described with semi-empirical Hamiltonian PM3 [14] and consisted of paraoxone, Arg35 and Tyr37 side-chain atoms and the two closest water molecules.

"Big" system Coordinates of esterase were taken from PDB ID 1LXW[15]. Missing residues V377, D378, D379, Q380 and C66 were added using PDB entry 2XMD [16] as a template. Protonation state was reconstructed according to table values with the `pdb2gmx` tool from the GROMACS package.

Molecular docking experiments were performed using Autodock Vina[17] to place the ligand in the active site of the enzyme. Preparation of the input PDBQT files and output processing were done with AutoDock Tools[18]. Initial echotiophate structure with partial Gasteiger charges was created using Avogadro software[19]. Docking cell contained whole protein with a margin of 5 from the edge atoms. The “exhaustiveness” parameter, which affects sampling, was set to 64 because of a moderate number of torsion angles. We performed 20 independent docking runs with freezed protein atoms remained and flexible ligand.

For further calculations one random configuration was selected with the distance between the echotiophate phosphorus atom and catalytic core less than 3.5.

Further equilibration was performed as for the previous system. Final QM system contained backbone parts of Tyr111-Phe115, full residues (with NH or CO parts for the terminal aminoacids) from Gly193-Gly197 loop, Met434-Ile439 loop, side chains of Gln220, Ser221, Asn319, Glu322, Tyr416, whole ligand Ech527 and 3 water molecules in the catalytic area.

Production run The prepared systems were subjected to QM/MM simulation with the modified GROMACS/ORCA or GROMACS/MOPAC2012 package [6, 11]. The time step used was 0.2 fs. Temperature coupling with Nos-Hoover scheme allowed observation of the behavior of systems at human body temperature, 310 K. The total length of simulations was set to 100 steps and 10 independent replicas were calculated for each system.

QM parameters Following parameters were used for QM calculations: for GROMACS (common part for both MOPAC2012 and ORCA)

```
QMMM         = yes
QMMM-grps    = QM
QMMMscheme   = ONIOM
QMmethod     = RHF           ; required but ignored
QMbasis      = STO-3G        ; required but ignored
QMcharge     = XX
QMmult       = 1
```

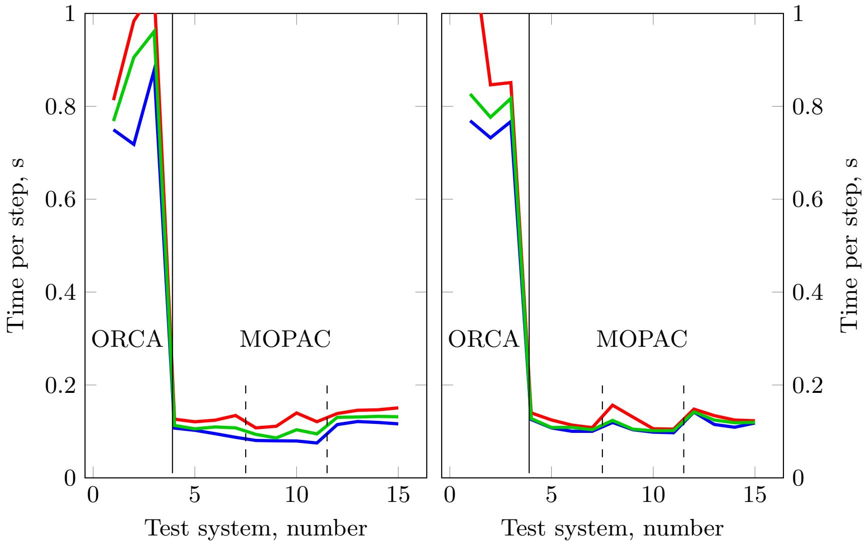
for MOPAC2012:

```
PM3 1SCF GRADIENTS CHARGE=XX singlet THREADS=YY
```

for ORCA:

```
! RHF PM3 NOFROZENCORE CONV HUECKEL
%rel SOCType  1 end
%elprop Dipole false end
%scf MaxIter 5000 end
%output PrintLevel Nothing end
```

where charge and number of MOPAC2012 threads were altered depending on the test system.



.

Fig. 2. Performance of QM/MM packages for the “small” system on local workstation (left) and supercomputer ”Lomonosov-2” wit Lustre filesystem (right). Red line: maximum time per step value across ten replicas. Green: median time per step value across ten replicas. Blue: minimal time per step value across ten replicas. Description of systems is given in the Table 2

Table 2. Test system descriptions. Threads counts concerns to both MM and QM susbsystems. Precision description describe Gromacs precision. MM GPU concerns to Gromacs acceleration of MM subsystem.

System Number	Description
1	Orca single precision
2	Orca single precision and MM GPU
3	Orca double precision
4	Mopac single precision; 1 thread
5	Mopac single precision; 2 threads
6	Mopac single precision; 4 threads
7	Mopac single precision; 8 threads
8	Mopac single precision; 1 thread and MM GPU
9	Mopac single precision; 2 threads and MM GPU
10	Mopac single precision; 4 threads and MM GPU
11	Mopac single precision; 8 threads and MM GPU
12	Mopac double precision; 1 thread
13	Mopac double precision; 2 threads
14	Mopac double precision; 4 threads
15	Mopac double precision; 8 threads

3 Results and Discussions

Computational times Hybrid QM/MM calculation of the “small” system with MOPAC library showed an eight-fold acceleration in comparison with the ORCA binary (Figure 3, the test systems decoding is given in Table 2). The performance difference was mainly associated with the necessity of the data exchange between GROMACS and ORCA via file system; the QM subsystem energy gradients were calculated with a similar speed by MOPAC and ORCA tools. The use of double precision version of GROMACS slowed the calculation speed on the local workstation approximately 1.5 times, but the difference was completely leveled when the calculations were performed on the supercomputer. Surprisingly, the maximum performance was achieved on the local computer with a fast file system (NVMe 1.2). The use of threads for QM and MM subsystems gave a performance gain of 10%, the most significant effect was observed for the slow SCF (Self-Consistent-Field) steps on the ”Lomonosov-2” supercomputer. This observation indicates the effectiveness of Intel libraries parallelization while getting the wavefunction to converge.

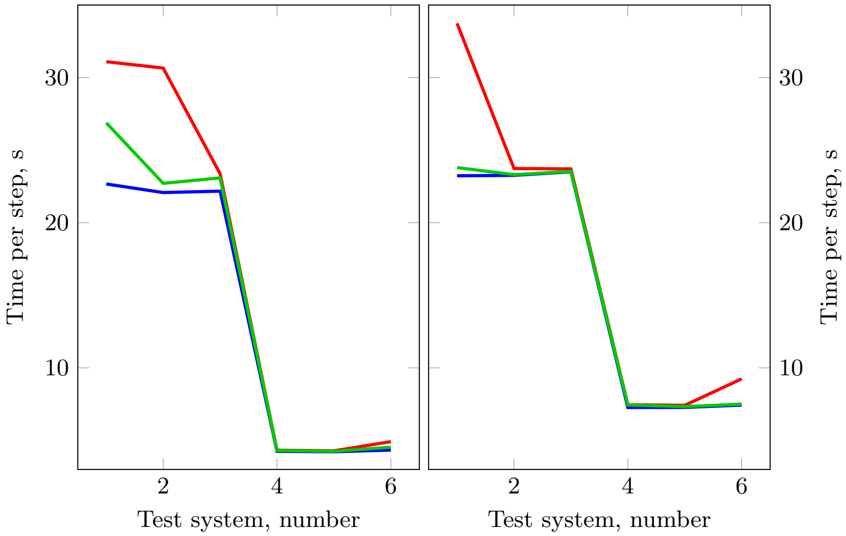


Fig. 3. Performance of test systems for the “big” system on local workstation (left) and supercomputer ”Lomonosov-2” wit Lustre filesystem. Red line maximum time value across ten replicas. Green is median time values across ten replicas. Blue is minimal time values across ten replicas. Description of systems are given in the Table 3.

In comparison with “small” system we observed non-linear increase in computational time. While system size expanded in 6.5 times, computational time increased in 30 times for both ORCA and MOPAC2012 versions. Neverthe-

Table 3. Test system descriptions. Threads counts concerns to both MM and QM susbsystems. Precision description describe Gromacs precision. MM GPU concerns to Gromacs acceleration of MM subsystem.

System Number	Description
1	Orca single precision
2	Orca single precision and MM GPU
3	Orca double precision
4	Mopac single precision; 1 thread
5	Mopac single precision; 1 thread and MM GPU
6	Mopac double precision; 1 thread

less internal performance ration between ORCA and MOPAC2012 remained the same.

Reproducibility Because implementations of computational protocols in different packages can vary it's very important to verify reproducibility of results across different software. To verify that our interface is producing correct results we compared energy values between corresponding runs with MOPAC2012 and ORCA. Typical error is less than 0.1% at timescale of 100 steps (Fig. 3) and can be explained by difference in precision of data transfer (single in ORCA vs double for MOPAC2012).

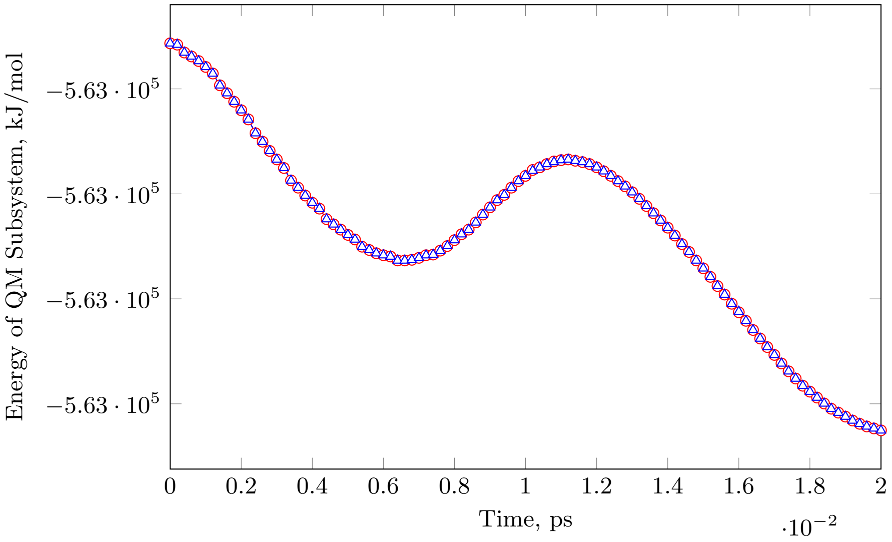


Fig. 4. Comparison of energy values for QM subsystem for MOPAC/Gromacs: red circles and ORCA/Gromacs: blue triangles.

4 Conclusions

We developed a new implementation of QM/MM interface between the MD program package GROMACS and the QM semi-empirical MOPAC tool that performs eight times faster than the original GROMACS interface to ORCA. Linking MOPAC as a GROMACS library allows to use the tool in supercomputer environment with Lustre distributed storage file system without input-output delays. The ability of MOPAC to inexpensively simulate molecular systems of large size allows efficient application of this tool to modern problems in life sciences. Our implementation of QM/MM interface provides a base for even better performance of hybrid simulations considering extensive ongoing development of GROMACS and MOPAC tools.

It pledged the way for combining Gromacs with next MOPAC2016 release with a better support of multithreading as well as GPU acceleration which will allow even to include into QM part even bigger regions which is very important for life science problems.

5 Acknowledgements

Authors are grateful to Dr. J. Stewart for sources of MOPAC2012 and initial hints on implementation. Computer resources were provided by the Research Computing Center of Moscow State University. The supercomputer "Lomonosov-2" was used for molecular dynamics simulations. Study was supported by Russian Ministry of Education and Science, #14.576.21.0095

References

1. A. Warshel and M. Levitt. Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J. Mol. Biol.*, 103(2):227–249, May 1976.
2. D. K. Vanatta, D. Shukla, M. Lawrenz, and V. S. Pande. A network of molecular switches controls the activation of the two-component response regulator NtrC. *Nat Commun*, 6:7283, Jun 2015.
3. A. Laio and M. Parrinello. Escaping free-energy minima. *Proc. Natl. Acad. Sci. U.S.A.*, 99(20):12562–12566, Oct 2002.
4. R. Quhe, M. Nava, P. Tiwary, and M. Parrinello. Path Integral Metadynamics. *J Chem Theory Comput*, 11(4):1383–1388, Apr 2015.
5. G. Bussi, A. Laio, and M. Parrinello. Equilibrium free energies from nonequilibrium metadynamics. *Phys. Rev. Lett.*, 96(9):090601, Mar 2006.
6. Schulz R. Larsson P. Bjelkmar P. Apostolov R. Shirts M. R. Smith J. C. Kasson P. M. van der Spoel D. Hess B. Lindahl E. Pronk S., Pall S. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics*, 29(7):845–854, April 2013.
7. J. J. Stewart. Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters. *J Mol Model*, 19(1):1–32, Jan 2013.

8. S. Dapprich, I. Komromi, B. Suzie, K. Morokuma, and M. J. Frisch. A new oniom implementation in gaussian98. part i. the calculation of energies, gradients, vibrational frequencies and electric field derivatives1dedicated to professor keiji morokuma in celebration of his 65th birthday.1. *Journal of Molecular Structure: THEOCHEM*, 461-462:1 – 21, 1999.
9. M. J. Abraham, D. van der Spoel, E. Lindahl, and B. Hess. *GROMACS User Manual version 5.0.7*, 2015.
10. I. Smirnov, E. Carletti, I. Kurkova, F. Nachon, Y. Nicolet, V. A. Mitkevich, H. Debat, B. Avalle, A. A. Belogurov, N. Kuznetsov, A. Reshetnyak, P. Masson, A. G. Tonevitsky, N. Ponomarenko, A. A. Makarov, A. Friboulet, A. Tramontano, and A. Gabibov. Reactibodies generated by kinetic selection couple chemical reactivity with favorable protein dynamics. *Proc. Natl. Acad. Sci. U.S.A.*, 108(38):15954–15959, Sep 2011.
11. I. V. Smirnov, A. V. Golovin, S. D. Chatziefthimiou, A. V. Stepanova, Y. Peng, O. I. Zolotareva, A. A. Belogurov, I. N. Kurkova, N. A. Ponomarenko, M. Wilmanns, G. M. Blackburn, A. G. Gabibov, and R. A. Lerner. Robotic QM/MM-driven maturation of antibody combining sites. *Sci Adv*, 2(10):e1501695, Oct 2016.
12. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of Chemical Physics*, 79(2):926–935, 1983.
13. K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins*, 78(8):1950–1958, Jun 2010.
14. James Stewart. Optimization of parameters for semiempirical methods i. method. *Journal of Computational Chemistry*, 10(2), 1989.
15. F. Nachon, O. A. Asojo, G. E. Borgstahl, P. Masson, and O. Lockridge. Role of water in aging of human butyrylcholinesterase inhibited by echothiophate: the crystal structure suggests two alternative mechanisms of aging. *Biochemistry*, 44(4):1154–1162, Feb 2005.
16. F. Nachon, E. Carletti, M. Wandhammer, Y. Nicolet, L. M. Schopfer, P. Masson, and O. Lockridge. X-ray crystallographic snapshots of reaction intermediates in the G117H mutant of human butyrylcholinesterase, a nerve agent target engineered into a catalytic bioscavenger. *Biochem. J.*, 434(1):73–82, Feb 2011.
17. O. Trott and A. J. Olson. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem*, 31(2):455–461, Jan 2010.
18. G. M. Morris, R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell, and A. J. Olson. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J Comput Chem*, 30(16):2785–2791, Dec 2009.
19. M. D. Hanwell, D. E. Curtis, D. C. Lonie, T. Vandermeersch, E. Zurek, and G. R. Hutchison. Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. *J Cheminform*, 4(1):17, Aug 2012.