Supercomputing Frontiers and Innovations

2026, Vol. 13, No. 1

Scope

- Future generation supercomputer architectures
- Exascale computing
- Parallel programming models, interfaces, languages, libraries, and tools
- Supercomputer applications and algorithms
- Novel approaches to computing targeted to solve intractable problems
- Convergence of high performance computing, machine learning and big data technologies
- Distributed operating systems and virtualization for highly scalable computing
- Management, administration, and monitoring of supercomputer systems
- Mass storage systems, protocols, and allocation
- Power consumption minimization for supercomputing systems
- Resilience, reliability, and fault tolerance for future generation highly parallel computing systems
- Scientific visualization in supercomputing environments
- Education in high performance computing and computational science

Editorial Board

Guest Editors' Introduction for Special Issue on

Supercomputing Challenges in Molecular Modeling in Life and Material Sciences and Astrochemistry

Rapid developments in supercomputing hardware and software have made it possible to utilize molecular modeling as a powerful tool that complements experimental studies. It accompanies investigations in materials science, life sciences, and astrochemistry. The exploitation of GPU-accelerated architectures has revolutionized molecular dynamics simulations. A few years ago, supercomputing facilities were used to obtain tens-of-nanoseconds trajectories for model systems. Microsecond trajectories can now be obtained on a single GPU in the same amount of time. CPU development has mostly contributed to quantum chemical calculations, making it possible to utilize more precise methods for larger molecular systems.

Machine learning and artificial intelligence methods with respect to molecular modeling are also developing. In 2024, the Nobel Prize in Chemistry was awarded for computational protein design and, jointly, for protein structure prediction, which revolutionized biochemical and biophysical studies. AI augmented approaches accelerate the screening and design of novel therapeutics, catalysts, and materials, as well as the understanding of astrochemical events.

This special issue comprises studies at the edge of supercomputing and molecular modeling. The included papers address algorithmic innovations, scalable implementations, and applications. Collectively, they illustrate how frontier supercomputing assists in and clarifies experimental studies.

Guest Editors
Prof. Maria Khrenova,
Prof. Andrey Stolyarov,
Chemistry Department,
Lomonosov Moscow State University,
Moscow, Russia

Contents

DOI: 10.14529/jsfi260101

Effective Algorithms of the RI Approximation for the CIS Method: an Example of Application of the High-Memory Strategy in the *Ab Initio* Calculations

Maksim D. Zhukov[1], ***Ilya O. Glebov***[1]

Two variants of the CIS methods with the RI approximation have been implemented. Both methods employ the high-memory strategy: the first, RI-CIS(1), is based on the full storage of the decomposed electronic repulsion integrals (ERI) tensor and CIS Hamiltonian, while the second, RI-CIS(2), stores only the decomposed ERI tensor. Both variants of the RI-CIS were tested for parallelism, performance, and precision. The results are compared with the default CIS method and RIJCOSX approximation. The considered methods demonstrate higher performance compared to their analogs and higher precision compared to the RIJCOSX approximation. Even the worse scaling of the RI methods did not lead to the lower performance in the conducted test calculations. The reported algorithms show that the performance of the quantum chemistry calculations is limited not only by the CPU power but also by the availability of RAM. The large volume of available memory can significantly increase the speed of the calculation by employing more effective but memory-consuming algorithms.

Keywords: Configuration interaction, resolution of identity, ab initio, electronic repulsion integrals, Davidson diagonalization.

Introduction

Configuration interaction (CI) methods [1] are key approaches for the electronic structure calculations, especially in the cases of strong electronic correlation. These methods are used for calculation of the excited states in the modeling of photochemical reactions and calculations of the electronic spectra. However, the computational costs of these methods rapidly grow with the increase of the excitation number and active space size. This fact limits the applicability of the CI methods for the large molecular systems without any additional approximations.

Configuration interaction with single excitations (CIS) is the simplest method within the framework of CI theory. It can be considered as the extension of the Hartree—Fock methods for the description of the excited states by taking the single electronic transition into consideration.

The CIS method does not take into account double and higher excitations and, thus, gives lower precision for systems with a strong electronic correlation and a high contribution of biradical configurations. However, it is a helpful tool to achieve the preliminary information for the calculations with more complicated methods, especially in case of large molecules. The TD-DFT method [2] uses similar basics and has the same limitations but can achieve higher precision if the proper exchange-correlation functional is chosen. Both of these methods are useful for the simulations of large systems due to their numerical efficiency (calculation time grows only $O(N^4)$). However, the performance of these methods can be further increased if approximations such as Cholesky Decomposition (CD) [3] and Resolution of Identity (RI) [4] are used. Currently, the attention of quantum chemical software developers is focused on the more complicated advanced methods, while the simpler ones are undeservedly forgotten because their optimization seems unnecessary. For instance, the implementation of the restricted Hartree—Fock (RHF) method with the RI approximation was not optimal in the frequently used quantum chemical packages until recently [5], despite the theoretical background being well-known [6]. The realization of CIS in

[1]Chemistry Department, Lomonosov Moscow State University, 1-3 Leninskie Gory, Moscow, 119991, Russia

the ORCA package [7, 8] contains the RI only for the Coulomb contribution (RIJ), while the exchange term can be simplified only with the chain of spheres (COSX) [9] approximation. The previous successful realization of the RI-RHF method encouraged us to perform a similar job for other basic quantum chemical methods, such as CIS.

The most time-consuming part of the CIS calculations, like that of most other quantum chemical methods, is the calculation of electronic repulsion integrals (ERI). The obvious way is to calculate them, store in memory and use in the iterative calculation cycles. However, the total amount of ERIs is N_{AO}^4 (where N_{AO} is the number of atomic orbitals), which is too high to be stored for any practically valuable calculation. The use of the symmetry of the ERI tensor and different techniques that avoid storing negligibly small ERIs do not solve this problem. Thus, the direct method, which is based on the recalculation of ERI at each iteration, was proposed [10]. However, this approach leads to a significant increase in calculation time due to the multiple recalculations of ERIs.

The approximations based on the ERI tensor decomposition, such as CD [3] and RI [4], can reduce both computational cost and memory requirement of the calculation. They are successfully applied to boost methods such as RHF [5], MP2 [11], CASPT2 [12], CCSD [13], and many others. The possibility to store the decomposed ERI tensor helps to avoid multiple recalculations of the most time-consuming parts of the calculation. It should be noted that memory required for storing the CIS Hamiltonian is comparable to that required for the decomposed ERI tensor. Thus, the use of modern computers makes the application of high-memory algorithms possible, and no direct method with multiple recalculations of data is really needed in the present time.

Two variants of the CIS method with the RI approximation are presented in the article:

- the first one with storing both RI-ERI tensor and CIS Hamiltonian in RAM;
- the second one with storing RI-ERI tensor and without storing CIS Hamiltonian in RAM.

The paper is organized as follows: first, the basic CIS theory and the modifications needed for the RI approximation are discussed; then, the testing objects are presented, and the details of the calculations are provided; finally, the results of the benchmark calculations and the comparison of different algorithms are shown.

1. Theory

1.1. CIS Method

The CIS wavefunction is constructed as a linear combination of the configuration state functions (CSF) formed as all the possible single excitations of the ground state RHF wavefunction:

$$|\Psi_{\mathrm{CIS}}\rangle = \sum_{i}^{\mathrm{occ}} \sum_{a}^{\mathrm{virt}} C_i^a |\Phi_i^a\rangle = \sum_{i}^{\mathrm{occ}} \sum_{a}^{\mathrm{virt}} C_i^a \cdot \frac{1}{\sqrt{2}} (a_{a\alpha}^{\dagger} a_{i\alpha} + a_{a\beta}^{\dagger} a_{i\beta}) |\Phi_0\rangle, \tag{1}$$

where $a_{a\alpha}^{\dagger}$ and $a_{i\alpha}$ are the creation and annihilation operators corresponding to the addition of the electron with α spin to the a-th virtual orbital and the removal of the electron with α spin from the i-th core orbital. Hereinafter, indices i, j, k, and l are used for the core subspace; a and b for the virtual; and p, q, etc. for any of them.

The coefficients C_i^a of the decomposition of the CI wavefunction $|\Psi_{\mathrm{CIS}}\rangle$ by the CSFs $|\Phi_i^a\rangle$ can be found by solving the linear variational problem – CI Hamiltonian diagonalization. The

Hamiltonian matrix elements can be found as:

$$H_{ijab} = \langle \Phi_i^a | \hat{H} | \Phi_j^b \rangle = \begin{cases} 2(ai|bj) - (ab|ij), & a \neq b, i \neq j \\ F_{ab} + 2(ai|bi) - (ab|ii), & a \neq b, i = j \\ -F_{ij} + 2(ai|aj) - (aa|ij), & a = b, i \neq j \\ F_{aa} - F_{ii} + 2(ai|ai) - (aa|ii) + E_{core}, & a = b, i = j \end{cases}, \quad (2)$$

where $(ai|bj)$ is the electronic repulsion integral (3), F_{pq} is the element of the Fockian matrix (4), E_{core} is the core energy containing the internuclear repulsion V_{nn}, the diagonal elements of the one electron Hamiltonian ($\hat{h}$) in the core subspace, Coulomb and exchange integrals (5):

$$(ai|bj) = \iint \frac{\varphi_i(r_1)\varphi_a(r_1) \cdot \varphi_j(r_2)\varphi_b(r_2)}{|r_1 - r_2|} dr_1 dr_2 \; ; \quad (3)$$

$$F_{pq} = h_{pq} + \sum_{k \in \text{core}} \Big[2(pq|kk) - (pk|qk) \Big] \; ; \quad (4)$$

$$E_{core} = V_{nn} \; + 2 \sum_{k \in \text{core}} h_{kk} \; + \sum_{k,l \in \text{core}} \Big[2(kk|ll) - (kl|kl) \Big]. \quad (5)$$

The full diagonalization of the Hamiltonian matrix is a very time-consuming task. However, only a small number of the lowest eigenvalues is needed because most of the higher eigenvalues do not have any physical meaning. Lowest eigenvalues can be found using the Davidson algorithm [14, 15], which does not require the full storage of the Hamiltonian matrix H. Only the procedure of multiplication of the trial vector by the matrix is needed:

$$H|\Psi_n\rangle = \sum_J d_{nJ} |\Phi_J\rangle = \sum_J \left[\sum_I C_{nI} H_{IJ} \right] |\Phi_J\rangle = \sum_{I,J} C_{nI} H_{IJ} |\Phi_J\rangle, \quad (6)$$

where matrix elements H_{IJ} can be calculated "on the fly". At first, the ERIs in the atomic orbital basis $(\alpha\beta|\gamma\delta)$ are calculated. Then they are used to calculate ERIs for the molecular orbitals $(pq|rs)$. Finally, the matrix elements H_{IJ} are calculated (2) and used for the multiplication (6). The Hamiltonian matrix can be stored in case of small molecules or large amount of available RAM. It requires storage of $n_\text{c}^2 \cdot n_\text{v}^2$ elements (n_c and n_v are the numbers of core and virtual orbitals). However, in any case, the calculation of $(\alpha\beta|\gamma\delta)$ is needed. The ERI tensor containing N_AO^4 elements can not be stored in RAM for any valuable calculation. The recalculation of $(\alpha\beta|\gamma\delta)$ on each iteration significantly increases the computational time, which is an inevitable cost of avoiding storage of the integrals.

1.2. Resolution of Identity (RI) Approximation

The resolution of identity approximation [6], which is also called Density Fitting (DF), is widely used to speed up the ERI calculation. This approach gives two significant advantages. First, it reduces the scaling of the calculations from $O(N^4)$ to $O(N^3)$ due to the change of the four-center integrals to the three-center ones. Second, it reduces the memory requirements and makes the recalculation of the ERIs at each iteration unnecessary [5].

The key idea of RI is to use the approximation of the product of two basis functions (pair density) as a linear combination of functions from the auxiliary basis set $\{\rho_\mu(r)\}$:

$$\rho_{ij} = \chi_i(r)\chi_j(r) \approx \widetilde{\rho}_{ij} = \sum_\mu C_\mu^{ij}\rho_\mu(r), \tag{7}$$

where $\{\rho_\mu(r)\}$ is a special set of auxiliary basis functions that has the dimension higher than the dimension of atomic orbital basis but significantly lower than the number of pair products $\chi_i(r)\chi_j(r)$ ($N_{\text{aux}} \approx 3N_{\text{AO}} \div 5N_{\text{AO}}$). C_μ^{ij} coefficients are determined by the minimization of the approximation error. The most suitable for ERI is RI-V or Coulomb fitting, which is based on the minimization of the Coulomb self-repulsion of the density difference:

$$\Delta_{ij} = \iint \frac{\left[\rho_{ij}(r_1) - \widetilde{\rho}_{ij}(r_1)\right]\left[\rho_{ij}(r_2) - \widetilde{\rho}_{ij}(r_2)\right]}{|r_1 - r_2|} dr_1 dr_2 = \\ = (\rho_{ij} - \widetilde{\rho}_{ij}|\rho_{ij} - \widetilde{\rho}_{ij}) = \|\rho_{ij} - \widetilde{\rho}_{ij}\|^2 \ \to \min; \tag{8}$$

$$\left\| \chi_i \cdot \chi_j \ - \ \sum_\mu C_\mu^{ij}\rho_\mu(r) \right\| \ \to \min . \tag{9}$$

This is a standard problem of linear algebra that leads to a system of equations:

$$\sum_\mu C_\mu^{ij}(\rho_k|\rho_\mu) = (\rho_k|\chi_i\chi_j) \quad \Leftrightarrow \quad \sum_\mu V_{k\mu}C_\mu^{ij} = (ij|k), \tag{10}$$

where $(ij|k)$ is the three-center integral (11) containing one auxiliary and two orbital basis functions; $V_{k\mu}$ is the two-center coulomb integral (12) of two auxiliary functions:

$$(ij|k) = \iint \frac{\chi_i(r_1)\chi_j(r_1)\cdot\rho_k(r_2)}{|r_1 - r_2|} dr_1 dr_2 \ ; \tag{11}$$

$$V_{k\mu} = \iint \frac{\rho_k(r_1)\cdot\rho_\mu(r_2)}{|r_1 - r_2|} dr_1 dr_2. \tag{12}$$

The solution of eq. (10) is the following:

$$C_\mu^{ij} = \sum_k (V^{-1})_{\mu k}(ij|k). \tag{13}$$

The substitution of (7) into the $(\alpha\beta|\gamma\delta)$ gives:

$$(\alpha\beta|\gamma\delta) = (\chi_\alpha\chi_\beta|\chi_\gamma\chi_\delta) \approx \left(\sum_\mu C_\mu^{\alpha\beta}\rho_\mu \middle| \sum_\nu C_\nu^{\gamma\delta}\rho_\nu\right) = \sum_{\mu,\nu} C_\mu^{\alpha\beta}C_\nu^{\gamma\delta}(\rho_\mu|\rho_\nu) = \sum_{\mu,\nu} C_\mu^{\alpha\beta}C_\nu^{\gamma\delta}V_{\mu\nu} \ . \tag{14}$$

The further substitution of coefficients C_μ^{ij} (13) in (14) leads to:

$$(\alpha\beta|\gamma\delta) \approx \sum_{\mu,\nu}\left[\sum_k (V^{-1})_{\mu k}(\alpha\beta|k)\right] V_{\mu\nu}\left[\sum_l (V^{-1})_{\nu l}(\gamma\delta|l)\right] = \\ = \sum_k\sum_l\left[(\alpha\beta|k)(\gamma\delta|l)\sum_{\mu,\nu}(V^{-1})_{\mu k}V_{\mu\nu}(V^{-1})_{\nu l}\right]. \tag{15}$$

V is a symmetric, positive definite matrix, so the same is true for V^{-1}, and:

$$\sum_{\mu,\nu}(V^{-1})_{\mu k}V_{\mu\nu}(V^{-1})_{\nu l} = (V^{-1})_{kl} \ . \tag{16}$$

Thus, the four-center ERIs can be expressed through two- and three-center integrals as follows:

$$(\alpha\beta|\gamma\delta) \approx \sum_{k,l}(\alpha\beta|k)(V^{-1})_{kl}(l|\gamma\delta). \tag{17}$$

The precision of the resulting RI approximation depends on the quality of the auxiliary basis set $\{\rho_\mu\}$. These sets are specially optimized for the exact representation of the pair densities $\rho_{ij}(r)$ and their Coulomb interactions. There are standard auxiliary basis sets specially designed for the calculations with the particular orbital basis sets (e.g., cc-pVXZ-RI [6], def2/JKFIT, def2/RIFIT [16]). The calculation error is negligible if the appropriate auxiliary basis set is used, but it can also be reduced by taking a larger auxiliary basis set.

1.3. RI-CIS Variants

A series of testing calculations using the CIS method with the RI approximation was conducted. The Davidson diagonalization [14, 15] was used to solve the variational CI problem. Two different implementations of the RI-CIS method based on the mentioned above formulae were realized. These two methods employ different strategies for the calculation and storage of the H_{CIS} matrix:

1) The first one rejects the recalculation of the Hamiltonian on the fly and employs storage of the full Hamiltonian matrix. Let us call it RI-CIS(1). Despite the fact that this method can be inapplicable for the large systems where the Hamiltonian storage is impossible, it must have higher performance for the smaller systems.
2) The second variant is a direct RI-CIS where the Hamiltonian matrix is recalculated on the fly during each Davidson iteration. However, it is not a commonly used direct CI because the three-center ERI tensor is stored but not recalculated. Let us call it RI-CIS(2). This approach must be less effective than RI-CIS(1) for the small molecules, but it would be the only choice if the H_{CIS} can not be stored in memory.

Each of the considered algorithms leads to the reduction of the number of operations and calculation time, that can be different for different size of molecule and used basis set. Thus, the more efficient algorithm can be determined only by the testing calculations.

2. Computational Details

Both reported algorithms RI-CIS(1) and RI-CIS(2) were implemented in the **NOPT** – the author's software written in C++ for the calculation of the **N**on-**O**rhogonal CI [17] and its extension by the **P**erturbation **T**heory [18, 19] (available upon request from the authors). It was compiled with the Libint [20] library version 2.9.0 for the atomic integrals and BLAS for the matrix multiplications. Both algorithms were then tested for accuracy, performance, parallelism, and scaling.

The benchmark calculations were done using the ORCA 5.0.3 [7, 8] package. The approximation error was estimated by the comparison with the default variant of CIS, and the performance was compared with the CIS in the RIJCOSX approximation. The comparison was performed for the calculation of the same molecule using the same basis set.

The testing calculations were performed for the polyene molecules C_2H_4, C_4H_6, C_6H_8, C_8H_{10}, $C_{10}H_{12}$ with the correlation consistent basis sets cc-pVnZ (n = D, T, Q) and their analogs aug-

mented by diffuse functions aug-cc-pVnZ. The corresponding auxiliary basis sets (cc-pVnZ-RI and aug-cc-pVnZ-RI) were taken for the RI approximation. All the basis sets were downloaded from the Basis Set Exchange database (BSE) [21–23]. The structures of all the molecules were taken to be planar, all double C=C bonds were in the trans configuration, and all single C–C bonds were in the s-trans-conformation. All the angles were 120°, and the bond lengths of C–H and two types of C–C were taken from the butadiene microwave geometry [24]. The structure of $C_{10}H_{12}$ is shown in Fig. 1.

All the testing calculations were carried out on the computer with two central processors Intel Xeon E5-2697 v3 with 2.60 GHz frequency and 14 cores per CPU. The computer had 196 GB of RAM, which was sufficient for all the calculations.

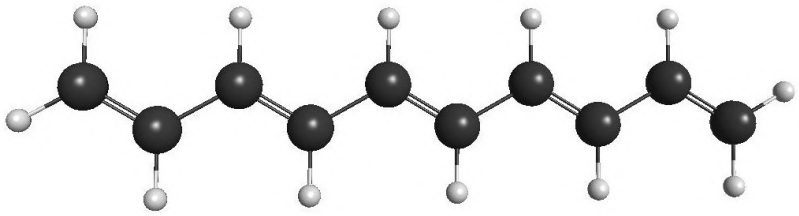


Figure 1. Structure of $C_{10}H_{12}$. Image was made using wxmacmolplt software [25]. Carbon atoms are black, hydrogen atoms are gray. Two different types of C–C bonds are drawn as single and double

3. Results and Discussion

3.1. Accuracy Tests

The tests of the accuracy of the excitation energies calculated by RI-CIS and RIJCOSX-CIS were conducted by comparing of the 10 lowest values calculated for the polyene molecules with the values calculated by default CIS. The excitation energies calculated by both variants of RI-CIS are equal; the deviation of $\sim 10^{-8}$ Hartree is caused by the numerical error and can be considered negligible. The maximum deviations for the excitation energies calculated for C_2H_4 and $C_{10}H_{12}$ with different basis sets by different approximate CIS methods are given in Tab. 1. The RIJCOSX-CIS was used with two variants of the grid: DefGrid2 and DefGrid3.

These deviations in the excitation energies show that the error of the RI approximation decreases with the increase of the basis set, especially with the augmentation. At the same time, the RIJCOSX error is higher than the RI error for small molecules. The use of the DefGrid3 does not significantly increase precision. The increase of the basis set in the cc-pVnZ series does not affect the RIJCOSX error, while the augmentation leads to larger errors. Both approximations show a decrease in error with the increase in molecule size. In the case of large molecule and compact basis set, the errors are similar for RI and RIJCOSX.

The decrease of the RIJCOSX error with the increase in the molecule size is probably caused by the method of grid construction. The grid points are generated around the atoms. The electronic density for the large polyenes is delocalized along the chain of the conjugated carbon bonds, while the density for the small polyenes is more delocalized in the orthogonal direction. Thus, the

Table 1. Errors of the excitation energies of different molecules calculated by different methods with different basis sets

Molecule	Basis set	Δ, cm^{-1}		
		RI-CIS	RIJCOSX-CIS	
			DefGrid2	DefGrid3
C_2H_4	cc-pVDZ	104.0	173.7	155.8
	aug-cc-pVDZ	18.8	278.3	255.6
	cc-pVTZ	30.1	155.5	186.9
	aug-cc-pVTZ	24.6	242.4	252.6
	cc-pVQZ	15.1	162.4	189.6
	aug-cc-pVQZ	21.6	236.9	244.6
$C_{10}H_{12}$	cc-pVDZ	60.9	75.6	70.2
	aug-cc-pVDZ	13.1	113.5	93.2

localization of the grid points better fits electronic density for the larger polyenes, and the energy errors become lower.

The deviations of the excitation energies of C_2H_4 and $C_{10}H_{12}$ calculated by RI-CIS and RIJCOSX-CIS with DefGrid2 and DefGrid3, as a function of the state number, are presented in Fig. 2. The figure shows that the RIJCOSX error depends significantly on the electronic state, and the picture is the same with any basis set. Contrarily, the RI error is significantly smaller if a larger and more diffuse basis is used.

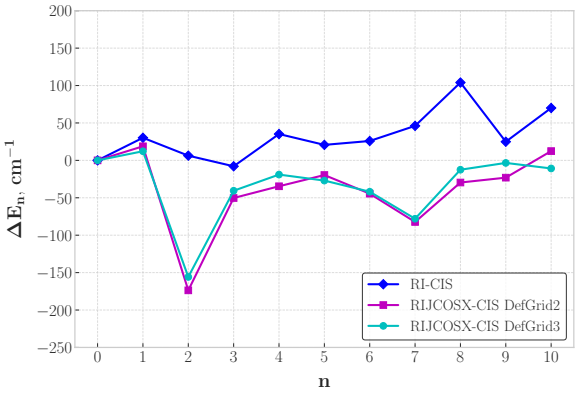
3.2. Performance Tests

A series of calculations with the standard and approximate variants of the CIS method was carried out. The analysis of the total computational time and the time of one Davidson iteration as a function of molecule size (the number of carbon atoms) and basis set dimension was performed. The use of different convergers and initial CIS guesses in ORCA and NOPT resulted in a different number of iterations. However, the number of iterations also depends on the convergence criteria, which can vary in different situations. So, the most representative parameter is the time per iteration, and all the following comparisons are based on this value.

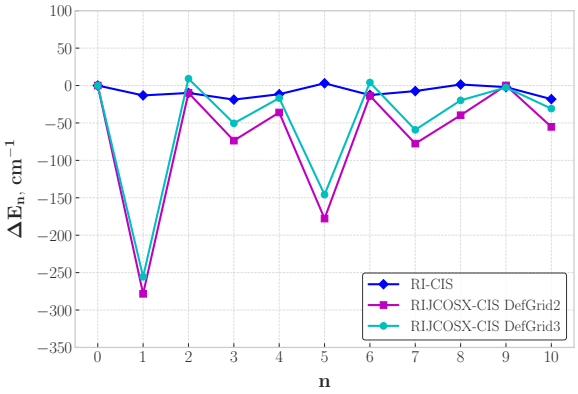
3.2.1. Parallelism test

The parallelism tests were performed for all the considered CIS methods by the calculations of the $C_{10}H_{12}$ molecule with the cc-pVDZ and aug-cc-pVDZ basis sets using different numbers of threads. The results are presented in Fig. 3.

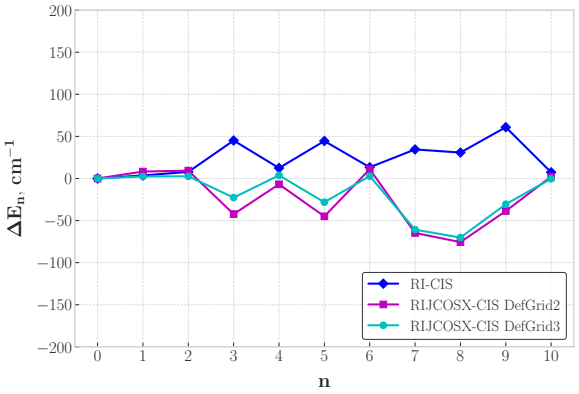
Similar slopes of the curves indicate that all the considered methods have a similar parallelism efficiency. This is true for both compact and augmented basis sets. The efficiency decreases with the increase in the number of threads for all methods. However, this decrease is lower for the CIS(2), and it becomes faster than CIS-RIJCOSX with DefGrid3.



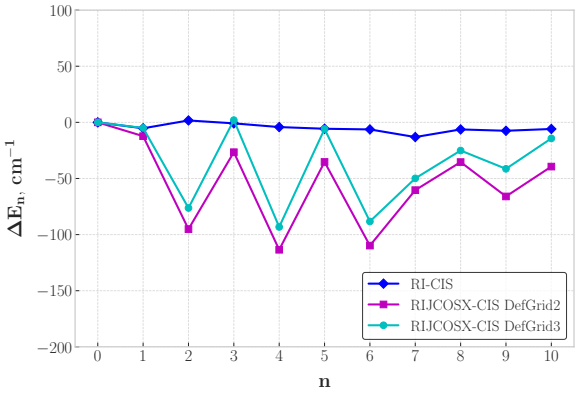
(a) Error for C_2H_4 with cc-pVDZ basis



(b) Error for C_2H_4 with aug-cc-pVDZ basis

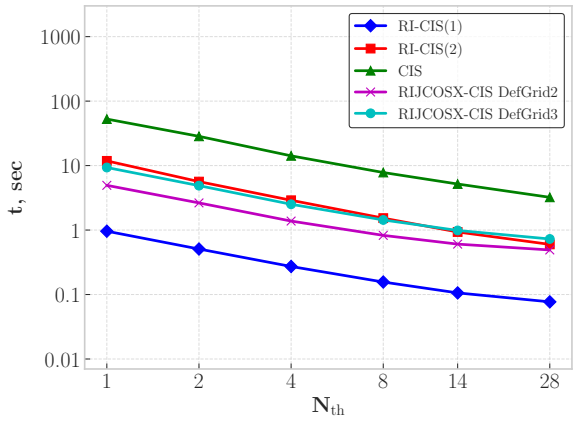


(c) Error for $C_{10}H_{12}$ with cc-pVDZ basis

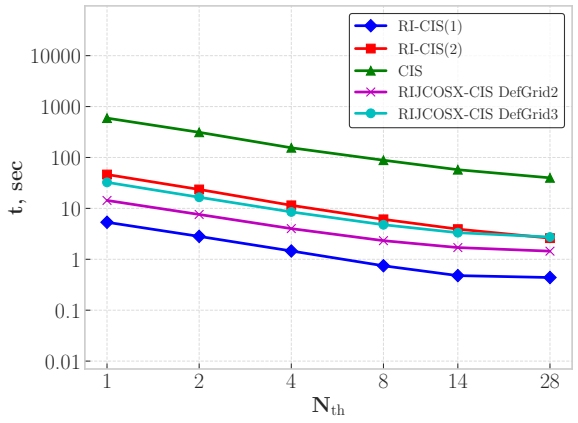


(d) Error for $C_{10}H_{12}$ with aug-cc-pVDZ basis

Figure 2. Error of the excitation energy calculated by the approximate CIS methods as a function of the number of state ($\Delta E_n = E_n^{\text{appr.CIS}} - E_n^{\text{CIS}}$)



(a) cc-pVDZ basis set



(b) aug-cc-pVDZ basis set

Figure 3. Dependence of the time per iteration t on the number of threads N_{th} with different CIS variants. Axes are in the logarithmic scale

The quantitative characteristic of parallelism efficiency can be obtained by fitting the shown curves by the equation:

$$t \sim N_{\text{th}}^{-\alpha} \ . \tag{18}$$

α and R^2 of the fitted curves are given in Tab. 2. The fitting is done for two intervals $N_{\text{th}} \in [1; 14]$, where only one CPU can be engaged in the calculation, and $N_{\text{th}} \in [1; 28]$, where both CPUs are engaged at the last point.

Table 2. The fitted α and R^2 of the equation (18) for the calculations of the $C_{10}H_{12}$ with cc-pVDZ and aug-cc-pVDZ basis sets

Basis set	cc-pVDZ				aug-cc-pVDZ			
Method	$N_{\text{th}} \in [1; 14]$		$N_{\text{th}} \in [1; 28]$		$N_{\text{th}} \in [1; 14]$		$N_{\text{th}} \in [1; 28]$	
	α	R^2	α	R^2	α	R^2	α	R^2
RI-CIS(1)	0.840	0.997	0.772	0.988	0.925	0.999	0.799	0.967
RI-CIS(2)	0.959	0.999	0.904	0.994	0.946	0.998	0.881	0.991
CIS	0.892	0.998	0.850	0.995	0.891	0.997	0.827	0.990
RIJCOSX-CIS DefGrid2	0.807	0.992	0.712	0.971	0.820	0.993	0.714	0.966
RIJCOSX-CIS DefGrid3	0.861	0.996	0.783	0.984	0.873	0.995	0.768	0.972

3.2.2. Scaling test (basis set)

A series of test calculations was carried out for the C_2H_4 molecule with cc-pVnZ and aug-cc-pVnZ (n = D, T, Q) basis sets using different variants of the CIS method. The dependence of the time per iteration on the number of basis functions was analyzed. The results are shown in Fig. 4.

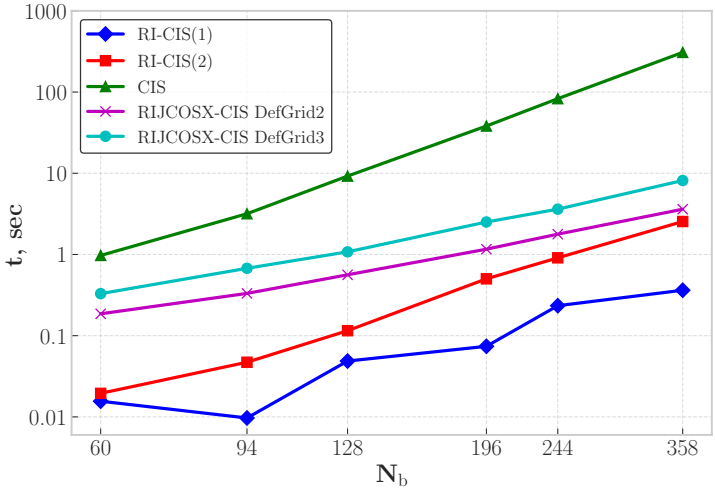


Figure 4. Dependence of the time per iteration t on the basis set dimension N_{b} for C_2H_4 calculated with $N_{\text{th}} = 1$. Axes are in the logarithmic scale

The figure shows that both RI-CIS methods demonstrate higher performance for any of the considered basis sets, with RI-CIS(1) having the highest one. The RIJCOSX methods are slower

than RI for the considered molecule, and the use of the denser grid (DefGrid3) leads to a decrease in performance compared to that with DefGrid2. The default CIS method is slower than any approximate variant.

The dependence of the calculation time per iteration on the dimension of the basis set was fitted with the function:

$$t \sim N_{\mathrm{b}}^{\beta} \ . \tag{19}$$

The fitted β and R^2 parameters of equation (19), presented in Tab. 3, demonstrate that, despite the higher performance, RI-CIS methods have worse scaling compared to RIJCOSX. The better scaling of the RIJCOSX methods can be explained by the slower increase in the number of integrals per point and the constant number of grid points. However, the scaling of RI-CIS is better than that of the default CIS.

Table 3. The fitted β and R^2 of the equation (19) for calculations of the C_2H_4 in different basis sets

Method	Parameter	
	β	R^2
RI-CIS(1)	2.050	0.876
RI-CIS(2)	2.833	0.993
CIS	3.264	0.998
RIJCOSX-CIS DefGrid2	1.678	0.995
RIJCOSX-CIS DefGrid3	1.791	0.997

Table 4. The fitted γ and R^2 of the equation (20) for the calculations of $C_{2n}H_{2n+2}$, $n = 1, 2, 3, 4, 5$ polyenes with cc-pVDZ and aug-cc-pVDZ basis sets

Basis set	cc-pVDZ		aug-cc-pVDZ	
Method	γ	R^2	γ	R^2
RI-CIS(1)	2.639	0.957	4.136	0.978
RI-CIS(2)	4.030	0.989	4.260	1.000
CIS	2.591	0.996	3.294	0.996
RIJCOSX-CIS DefGrid2	2.057	0.999	2.370	0.999
RIJCOSX-CIS DefGrid3	2.119	0.999	2.416	1.000

3.2.3. Scaling test (molecule size)

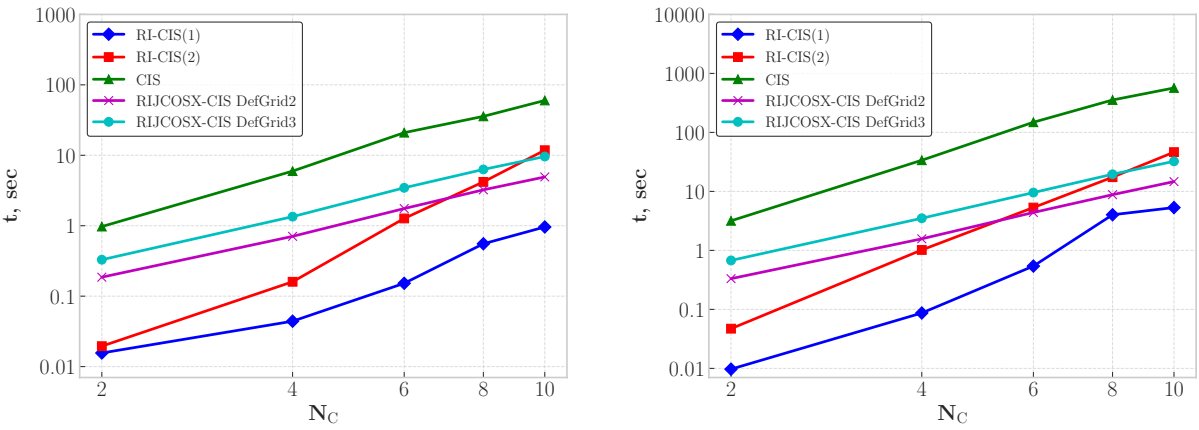
A series of test calculations was carried out for the $C_{2n}H_{2n+2}$ with n = 1, 2, 3, 4, 5 molecules with cc-pVDZ and aug-cc-pVDZ basis sets using different variants of CIS method. The results are shown in Fig. 5.

The figure shows that RI-CIS(1) has the highest performance for any considered molecule size. Both RI-CIS methods have smaller computation times compared to the default CIS methods. RI-CIS(2) is faster than RIJCOSX with any grid for small molecules but becomes slower for the larger ones.

The dependence of the calculation time per iteration on the number of carbon atoms N_{C} was fitted with the function:

$$t \sim N_{\mathrm{C}}^{\gamma} \ . \tag{20}$$

The fitted γ and R^2 parameters of equation (20), presented in Tab. 4, demonstrate that both RI-CIS methods have worse scaling compared to RIJCOSX. Default CIS has intermediate γ. Two variants of RI-CIS have different behaviors: the scaling of RI-CIS(2) is worse but more stable with the basis augmentation. RI-CIS(1) has γ values close to those of the RIJCOSX variants



(a) cc-pVDZ basis set

(b) aug-cc-pVDZ basis set

Figure 5. Dependence of the time per iteration t on the number of carbon atoms N_C for the calculations by the different CIS variants. Axes are in the logarithmic scale

with the compact basis but much higher with the augmented one. However, despite having worse scaling, RI-CIS(1) demonstrates higher performance for all the considered molecules. Moreover, the better precision of the RI-CIS methods makes them preferable, even in those cases where they may become slower due to the worse scaling.

Conclusion

The testing of the accuracy and performance of the considered CIS methods can be summarized as follows:

- RI approximation can introduce a relatively small error to the CIS results. This error is caused by the approximate calculation of the ERIs and is less than 150 cm^{-1}. The precision can be improved by employing a larger auxiliary basis set.
- RI approximation demonstrates a higher precision compared to RIJCOSX, especially for smaller molecules and augmented basis sets.
- Implementation of the considered algorithm has parallelism similar to that of the variants implemented in ORCA.
- RI-CIS methods have higher performance compared to RIJCOSX method for small and moderate-sized molecules. The worse scaling may make RI methods slower for the case of large molecules and basis sets. However, the better precision makes RI approximation preferable.
- RI-CIS(1) is the best choice as long as the available memory is sufficient.

The considered realization of the RI-CIS demonstrates the efficiency of the high-memory strategy in quantum chemical calculations. The employment of algorithms for the decomposition of the ERI tensor decreases the memory requirements, making it possible to store them in the RAM of modern computers. This situation is typical for the quantum chemical calculations, where a huge amount of intermediate data is calculated but cannot be stored and must be recalculated at each step. The use of the high memory strategy helps to avoid multiple recalculations and significantly increases performance.

The availability of a high volume of RAM makes this strategy applicable. However, the efficient technologies of large tensor decomposition, such as resolution of identity, Cholesky decomposition, etc., are even more important than the physical volume of memory.

Acknowledgements

The study was conducted under the state assignment of Lomonosov Moscow State University, project No. 121031300173-2.

References

1. Shavitt, I.: The history and evolution of configuration interaction. Molecular Physics 94(1), 3–17 (May 1998). https://doi.org/10.1080/002689798168303

2. Adamo, C., Jacquemin, D.: The calculations of excited-state properties with Time-Dependent Density Functional Theory. Chem. Soc. Rev. 42(3), 845–856 (2013). https://doi.org/10.1039/c2cs35394f

3. Beebe, N.H.F., Linderberg, J.: Simplifications in the generation and transformation of two-electron integrals in molecular calculations. International Journal of Quantum Chemistry 12(4), 683–705 (Oct 1977). https://doi.org/10.1002/qua.560120408

4. Vahtras, O., Almlöf, J., Feyereisen, M.: Integral approximations for LCAO-SCF calculations. Chemical Physics Letters 213(5–6), 514–518 (Oct 1993). https://doi.org/10.1016/0009-2614(93)89151-7

5. Glebov, I.O., Poddubnyi, V.V.: An effective algorithm of the Hartree–Fock approach with the storing of two-electron integrals in the resolution of identity approximation. Russian Journal of Physical Chemistry A 98(4), 617–625 (Apr 2024). https://doi.org/10.1134/s0036024424040101

6. Weigend, F.: A fully direct RI-HF algorithm: Implementation, optimised auxiliary basis sets, demonstration of accuracy and efficiency. Physical Chemistry Chemical Physics 4(18), 4285–4291 (Aug 2002). https://doi.org/10.1039/b204199p

7. Neese, F.: The ORCA program system. WIREs Computational Molecular Science 2(1), 73–78 (Jun 2011). https://doi.org/10.1002/wcms.81

8. Neese, F.: Software update: the ORCA program system, version 4.0. WIREs Computational Molecular Science 8(1) (Jul 2017). https://doi.org/10.1002/wcms.1327

9. Neese, F., Wennmohs, F., Hansen, A., Becker, U.: Efficient, approximate and parallel Hartree–Fock and hybrid DFT calculations. a 'chain-of-spheres' algorithm for the Hartree–Fock exchange. Chemical Physics 356(1–3), 98–109 (Feb 2009). https://doi.org/10.1016/j.chemphys.2008.10.036

10. Almlöf, J., Faegri, K., Korsell, K.: Principles for a direct SCF approach to LICAO–MO ab-initio calculations. Journal of Computational Chemistry 3(3), 385–399 (Sep 1982). https://doi.org/10.1002/jcc.540030314

11. Weigend, F., Häser, M., Patzelt, H., Ahlrichs, R.: RI-MP2: optimized auxiliary basis sets and demonstration of efficiency. Chemical Physics Letters 294(1–3), 143–152 (Sep 1998). https://doi.org/10.1016/s0009-2614(98)00862-8

12. Aquilante, F., Malmqvist, P.Å., Pedersen, T.B., *et al.*: Cholesky decomposition-based multiconfiguration second-order perturbation theory (CD-CASPT2): Application to the spin-state energetics of Co^{III}(diiminato)(NPh). Journal of Chemical Theory and Computation 4(5), 694–702 (Apr 2008). https://doi.org/10.1021/ct700263h

13. Folkestad, S.D., Kjønstad, E.F., Koch, H.: An efficient algorithm for Cholesky decomposition of electron repulsion integrals. The Journal of Chemical Physics 150(19) (May 2019). https://doi.org/10.1063/1.5083802

14. Davidson, E.R.: The iterative calculation of a few of the lowest eigenvalues and corresponding eigenvectors of large real-symmetric matrices. Journal of Computational Physics 17(1), 87–94 (Jan 1975). https://doi.org/10.1016/0021-9991(75)90065-0

15. Crouzeix, M., Philippe, B., Sadkane, M.: The Davidson Method. SIAM Journal on Scientific Computing 15(1), 62–76 (Jan 1994). https://doi.org/10.1137/0915004

16. Weigend, F.: Accurate Coulomb-fitting basis sets for H to Rn. Physical Chemistry Chemical Physics 8(9), 1057 (2006). https://doi.org/10.1039/b515623h

17. Glebov, I.O., Kozlov, M.I., Poddubnyy, V.V.: Comparison of the Coulomb and non-orthogonal approaches to the construction of the exciton Hamiltonian. Computational and Theoretical Chemistry 1153, 12–18 (Apr 2019). https://doi.org/10.1016/j.comptc.2019.02.010

18. Glebov, I.O., Poddubnyy, V.V., Khokhlov, D.V.: Perturbative expansion of non-orthogonal product approach for charge transfer states. The Journal of Physical Chemistry A 126(34), 5800–5813 (Apr 2022). https://doi.org/10.26434/chemrxiv-2022-9jb6k

19. Glebov, I.O., Poddubnyy, V.V., Khokhlov, D.: Perturbation theory in the complete degenerate active space (CDAS-PT2). The Journal of Chemical Physics 161(2) (Jul 2024). https://doi.org/10.1063/5.0211210

20. Valeev, E.F.: Libint: A library for the evaluation of molecular integrals of many-body operators over Gaussian functions. http://libint.valeyev.net/ (2025), version 2.11.2

21. Pritchard, B.P., Altarawy, D., Didier, B., *et al.*: New basis set exchange: An open, up-to-date resource for the molecular sciences community. Journal of Chemical Information and Modeling 59(11), 4814–4820 (Oct 2019). https://doi.org/10.1021/acs.jcim.9b00725

22. Feller, D.: The role of databases in support of computational chemistry calculations. Journal of Computational Chemistry 17(13), 1571–1586 (Oct 1996). https://doi.org/10.1002/(sici)1096-987x(199610)17:13<1571::aid-jcc9>3.0.co;2-p

23. Schuchardt, K.L., Didier, B.T., Elsethagen, T., *et al.*: Basis set exchange: A community database for computational sciences. Journal of Chemical Information and Modeling 47(3), 1045–1052 (Apr 2007). https://doi.org/10.1021/ci600510j

24. Craig, N.C., Groner, P., McKean, D.C.: Equilibrium structures for butadiene and ethylene: Compelling evidence for π-electron delocalization in butadiene. The Journal of Physical Chemistry A 110(23), 7461–7469 (May 2006). https://doi.org/10.1021/jp060695b

25. Bode, B.M., Gordon, M.S.: Macmolplt: a graphical user interface for GAMESS. Journal of Molecular Graphics and Modelling 16(3), 133–138 (Jun 1998). https://doi.org/10.1016/s1093-3263(99)00002-9

DOI: 10.14529/jsfi260102

High-Performance Computing in the Molecular Dynamics of Tubulin Cytoskeleton Polymers

Ilya B. Kovalenko[1], *Vladimir A. Fedorov*[1], *Ekaterina P. Vasyuchenko*[1], *Ekaterina G. Kholina*[1], *Svetlana Yu. Kovalenko*[1], *Andrey B. Rubin*[1]

High-performance computing is one of the most essential tools fueling the advancement of computational biology. The article discusses the application of the full-atom molecular dynamics (MD) method to study the dynamic behavior of filaments formed by the protein tubulin, and presents the results of testing the calculation performance depending on the latest models of central processors and video accelerators. Our comparative performance analysis of GPU-based computing architectures for all-atom MD simulations of biomolecular systems not only provides guidance on choosing the best computing solution in terms of price-performance ratio, but also shows the maximum potential computational performance that modern CPUs and GPUs can provide. For example, MD of the biomolecular system containing a tubulin protofilament in an explicitly specified solvent consisting of more than 300 thousand atoms can be studied with performance of 232 ns/day at time step 2 fs when using single-node computer with the latest CPU and GPU generation architecture. Constantly evolving computing resources coupled with modern software enable us to solve increasingly complex problems in life sciences.

Keywords: molecular dynamics, tubulin, microtubule, CPU, GPU, computing performance.

Introduction

Tubulins and tubulin-like proteins form cellular structures that play a crucial role in numerous cellular processes, including cell division. For example, tubulins form microtubules, which are involved in chromosome search and separation. This is possible due to the unique property of microtubules to spontaneously polymerize and depolymerize, a property known as dynamic instability [6]. Another protein, FtsZ (Filamenting temperature-sensitive mutant Z), is a tubulin-like protein that forms filaments with a repeating arrangement of subunits. These filaments form a ring (so-called Z-ring) around the longitudinal midpoint, or septum, of a bacterial cell [9]. FtsZ is essential for cell division in almost all bacteria and in many but not all archaea [8].

Although eukaryotic microtubules are a well-studied subject, the specific molecular mechanisms underlying and governing their dynamic instability remain unclear. One of the questions concerning microtubules is whether individual microtubule protofilaments assume a straight or curved shape in solution. We know even less about the specific molecular mechanisms underlying the dynamic behavior of FtsZ filaments that form the Z-ring. In any case, it is known that FtsZ protofilaments have polarity and move in one direction by treadmilling [9].

To study the dynamic behavior of both individual protofilaments and entire microtubules formed by tubulin, as well as filaments of the FtsZ protein, it is convenient to use methods of molecular computer modeling, in particular, the classical molecular dynamics (MD) method. However, studying such large molecular systems as microtubules and even individual tubulin protofilaments or FtsZ filaments requires extremely high computational resources. The situation is further complicated by the fact that both tubulins and FtsZ possess long, unstructured C-terminal regions or tails (in the case of FtsZ, this unstructured tail can reach a hundred amino acid residues in length), the role of which in dynamic behavior remains unclear. Because these

[1]Lomonosov Moscow State University, Moscow, Russia

regions lack a defined structure and their dynamic behavior is highly variable, the building of a molecular model requires increasing the reaction volume and obtaining long MD trajectories.

High-performance computing (HPC) is one of the most essential tools that fuels the advancement of computational biology. Since our latest overview of HPC capabilities, benchmarked on simulations of tubulin dynamics in 2023 [4], new version of CPU and GPU became commercially available. This article discusses the issue of choosing the optimal computational architecture for studying large biological systems using the MD method and the gains in computing speed that are provided by the latest models of central and graphic processors. Since 2018, when NVIDIA released the first RTX series GPUs, four generations of these accelerators have been released, and each new generation increased the speed of calculations. New, more powerful central processing units (CPUs) have also been designed by Intel, and although their contribution to the increase in MD calculation performance is less than that of GPUs, we also discuss them in this article, including Intel Ultra series CPU.

The article is organized as follows. Section 1 is devoted to the computational methods used in the paper. In Section 2 we provide the results of molecular dynamics performance tests using different computer architectures. The conclusion summarizes the study.

1. Methods

All tests were performed using the all-atom explicit solvent MD. The calculations were executed using the GROMACS 2022.5 or GROMACS 2025.2 software package [2] which facilitates parallel computing on hybrid architectures and incorporates the CHARMM27 force field [10]. Each benchmark simulation was run for a duration of 30 minutes, employing the TIP3P water model. The tubulin tetramer structure was obtained from the Protein Data Bank (PDB id 5SYF [7]). The dimensions of the simulation volume were selected to ensure that the distance from the protein's surface to the nearest boundary of the simulation box was no less than two nanometers. Long-range electrostatic interactions were taken into account using the particle mesh Ewald method [3]. Both Coulomb and Lennard-Jones cutoffs were configured to 1.25 nm. Molecular dynamics (md) integrator was used. For the water box systems, the time step was 1 fs and no restraints were used. For the tubulin protofilament, the time step was 2 fs and constraints were imposed on the bonds of atoms with hydrogens. The "GPU-resident" option was tested, which updates coordinates on the GPU when all force and coordinate data remain resident on the GPU for a number of steps [1]. Specifications of MD systems used for benchmarking are summarized in Tab. 1.

Table 1. Specification of molecular dynamics systems used in the benchmarks

MD systems	MD system name	Number of atoms	System size, nm
Water box (WB)	WB-10	10206	$4.7 \times 4.7 \times 4.7$
	WB-80	80232	$9.3 \times 9.3 \times 9.3$
	WB-200	203415	$12.7 \times 12.7 \times 12.7$
	WB-500	500076	$17.2 \times 17.2 \times 17.2$
	WB-1000	1000005	$21.7 \times 21.7 \times 21.7$
Tubulin tetramer	Tub-4	307453	$11.3 \times 12.5 \times 22.2$

2. Results and Discussion

2.1. Performance Tests

We have tested the performance of all-atom MD calculations for water box systems of different sizes as well as tubulin tetramer in water for both considered Intel CPUs (Core i9-14900KF and Core Ultra 9 285K) and a variety of NVIDIA RTX graphic accelerators. The table with all performance tests is available via this link. The results of the performance tests are also presented selectively in Fig. 1 and Tables 2–4.

2.2. Dependence of Performance on System Size

To understand how the performance of MD calculations depends on the size of the molecular system, we tested five water box systems of different sizes (10, 80, 200, 500, and 1000 thousands of atoms) for both considered Intel CPUs with and without NVIDIA RTX5090 graphic accelerator. We normalized the performance value to the system size and the performance of the Core i9-14900KF CPU only, Fig. 1. The graph shows that performance normalized to the number of atoms is virtually independent of system size when calculations are performed solely on the CPU, without a GPU. However, when using the top-of-the-line NVIDIA RTX5090 GPU for calculations, the situation changes greatly. First, performance with a GPU is significantly higher than without it, reaching its maximum (more than 20 times faster than with a CPU only) on large molecular systems (half a million and a million atoms). Second, on smaller molecular systems, normalized performance drops significantly. For example, for a system size of 80000 atoms, it drops by a factor of 2 on an Intel Ultra 9 285K processor, and by approximately 20% on a Core i9-14900KF processor. These results will be discussed in more detail in the following sections.

The cause of the relatively poor normalized performance for a small system (less than a hundred thousand atoms) may be the ineffective use of GPU computing resources. However, small systems are computationally fast, the computing performance of a system of 10000 atoms reaches 1 μs/day. In addition, biological systems of interest often contain a much larger number of atoms.

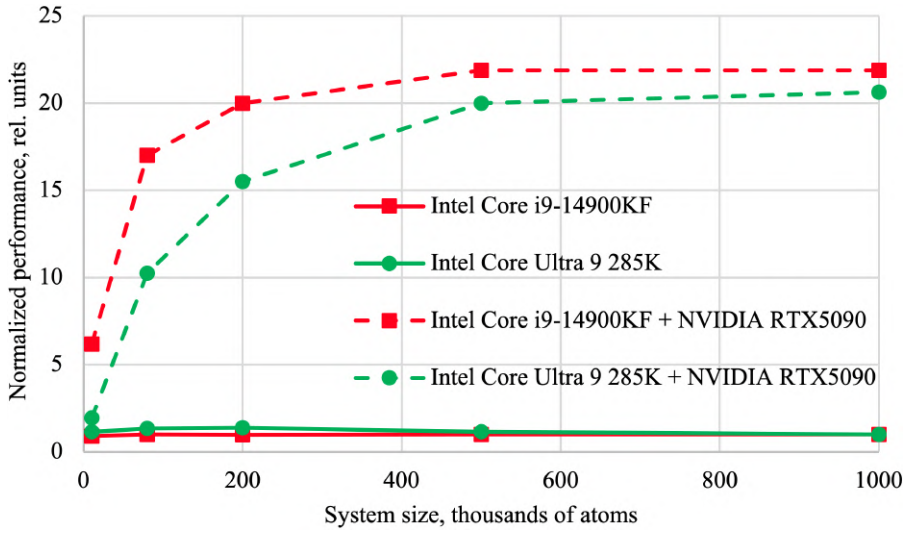


Figure 1. Normalized performance of MD calculation for systems of different sizes both for CPU only and CPU+GPU computer systems

2.3. Comparison of the Latest Intel Ultra 9 285K and i9-14900KF Processors in CPU-only Mode and in Combination with the Latest NVIDIA RTX5090 Graphics Accelerator

If the GPU is not used for molecular dynamics calculations, but only the CPU, the latest Intel Ultra 9 285K processor demonstrates almost 1.5 times higher performance than the Intel i9-14900KF on small systems of approximately 10000 atoms (Tab. 2, the third and fourth columns). However, the larger the molecular system, the smaller the performance gain. For a system consisting of a million atoms, no gain is observed, and these two CPUs demonstrate identical performance. It should be noted that the two versions of Gromacs compared, using only the CPU, produce virtually identical results, especially for the Intel i9-14900KF processor.

For large molecular systems with 200000 or more atoms, the latest NVIDIA RTX5090 graphics accelerator provides a 20-fold increase in performance when running MD simulations on the latest version of Gromacs 2025.2, or when using version 2022.5 with the GPU-resident option (Tab. 2). This observation roughly corresponds to data obtained three years ago (2022) on a previous-generation computer with an Intel i9-13900K processor and an NVIDIA RTX4090 graphics accelerator: back then, the graphics accelerator increased the computing speed by 14–22 times, depending on the system size [4]. Of course, a 2025 computer system with the graphics accelerator is about 1.3 times faster than the previous one anyway. It is worth adding that the new Intel i9-14900KF CPU delivers slightly faster, more stable and predictable GPU-based MD performance than the latest Intel Ultra 9 285K, especially for large molecular systems (Tab. 2).

Table 2. Single-node performance (ns/day) for water box systems of different size depending on various combinations of CPUs and RTX5090 GPU and version of Gromacs MD software

Molecular system	Intel Core CPU model	2022.5 No GPU	2025.2 No GPU	2022.5 GPU	2022.5 GPU-resident	2025.2 GPU	2025.2 GPU-resident
WB-10	i9-14900KF	142	143	768	1053	990	1016
WB-80	i9-14900KF	20	20	197	365	340	350
WB-200	i9-14900KF	7.8	7.8	85	160	160	160
WB-500	i9-14900KF	3.2	3.2	35	69	70	70
WB-1000	i9-14900KF	1.6	1.6	14	35	35	35
WB-10	Ultra 9 285K	205	185	336	348	310	319
WB-80	Ultra 9 285K	30	27	28	218	205	205
WB-200	Ultra 9 285K	12	11	28	127	124	123
WB-500	Ultra 9 285K	3.9	3.7	24	65	64	64
WB-1000	Ultra 9 285K	1.7	1.6	15	35	34	33

2.4. Notes on Using the GPU-resident Option in the Latest Version of Gromacs

In our previous paper [4], we argued that the GPU-resident option significantly (up to 2.3 times) improved the performance of MD calculations. Using the latest version of Gromacs, this option no longer provides any performance gain, at least on large systems over 200000 atoms (Tab. 2, the last two columns).

Table 3. Single-node Gromacs 2025.2 molecular dynamics performance (ns/day) for tubulin tetramer on various GPUs for Intel i9-14900KF CPU (results with GPU-resident option on and off are equal)

Test number	GPU	Series name	Release date	Performance, ns/day
1	No GPU	—	—	11
2	RTX2080ti	Turing	Sep 2018	45
3	RTX3080	Ampere	Sep 2020	65
4	RTX3090	Ampere	Sep 2020	75
5	RTX4080	Ada Lovelace	Nov 2022	121
6	RTX4090	Ada Lovelace	Oct 2022	167
7	RTX5080	Blackwell	Jan 2025	146
8	RTX5090	Blackwell	Jan 2025	232

2.5. Contribution of the Graphics Accelerator to the Performance of MD Calculations

We tested the performance of MD calculations on the latest version of Gromacs and the new Intel i9-14900KF processor, depending on the model of the NVIDIA RTX (Ray Tracing shader eXtreme) graphics accelerator, starting with the most advanced RTX20 models (Turing series), RTX2080ti, and ending with the latest model RTX5090 (Blackwell series), see Tab. 3. The performance test was conducted on a full-atom model of a real biological system, which is a tetramer of the protein tubulin in explicit water.

Immediately striking is the significant increase in MD performance starting with the 40 series (Ada Lovelace). For example, RTX4080 with its 121 ns/day is more than 1.5 times faster than the previous top-end model, RTX3090 (78 ns/day). Moreover, the top-end version of the 40 series (167 ns/day) is actually more than twice as fast as the top-end version of the 30 series.

As for the latest 50 series of NVIDIA graphics accelerators, they do not show such impressive results compared to the 40 series, although the top version of the 50 series RTX5090 with its fantastic 232 ns/day is nevertheless 1.4 times faster than RTX4090. The younger model of the 50 series, RTX5080, shows significantly worse performance even than the top card of the previous 40 series. Overall, in the past 7 years, starting in 2018, the MD performance of top-end RTX series graphics cards has increased more than 5 times. If we compare the performance of a computer with a top-end graphics accelerator versus one without a graphics accelerator at all, the performance gain is more than 20 times.

2.6. Performance of the Latest Intel Ultra 9 Processor in MD Calculations When Using the New Top-End NVIDIA RTX Graphics Accelerators

We tested the performance of MD calculations on the latest version of Gromacs and the newest Ultra series of Intel desktop CPUs, the Ultra 9 285K processor, depending on the NVIDIA RTX graphics accelerator model (Tab. 4). Performance tests were produced on the same biological system as for Intel i9-14900KF processor – a tubulin tetramer in explicit water. Since the test results were quite unexpected, we also performed the test on a larger molecular system, namely WB-1000.

Using NVIDIA RTX30-series graphics accelerators, the computational performance of both types of the latest Intel processors (Ultra 9 285K and i9-14900KF) is virtually identical. The same situation is repeated with the lower-end graphics cards of the next generations – RTX4080 and RTX5080 – although in these cases the Ultra processor demonstrates slightly better results. However, the test results become completely unpredictable and different when it comes to the top-end GPUs, namely RTX4090 and RTX5090. With an Intel Ultra processor, these graphics cards demonstrate significantly lower performance even compared to their lower-end counterparts, not to mention the fantastic speed they demonstrated with the Intel i9-14900KF. Indeed, for RTX5090, the MD computation performance with the Ultra 9 285K processor dropped more than 2-fold compared to i9-14900KF! This seems incredible and may be explained by the significantly different internal architecture of Intel's new Ultra series CPU, in which the processor consists of multiple dies. Instead of a single monolithic die, Intel Core Ultra uses a disaggregated chiplet design and features a neural processing unit (NPU) for AI acceleration. This architecture separates components like the CPU, GPU, and NPU into "tiles" that are manufactured on optimal processes before being combined, leading to power efficiency.

However, for the WB-1000 molecular system, three times larger and consisting of a million atoms, the Ultra 9 285K processor again outperforms the i9-14900KF when using the 5090 graphics card (38 ns/day versus 33 ns/day). The result regarding the computational slowdown when switching from the 80th to the 90th GPU also does not hold for the WB-1000 system (Tab. 4, last column). Indeed, the calculation speed of the 5090 graphics card is twice that of the 5080.

The Intel Core Ultra 9 processor performs slower with memory than the i9-14900K, as the latter has higher clock speeds and better multi-core performance, which is critical for memory performance in applications. Indeed, the Ultra series processor has 24 threads, while the i9-14900K has 32. At the same time, the smaller the size of the molecular system, the more frequent data exchange with the video card occurs. This may explain why the Ultra 9 285K performance drops more sharply compared to the i9-14900K as the biological system size decreases (Tab. 2). Note that performance drops only when using the top-end 40- and 50-series graphics accelerators. This is likely due to their higher memory bus widths (384 and 512 bits, respectively) and larger memory capacities (24 GB and 32 GB, respectively) than their lower-end counterparts. It can be concluded that Ultra 9 series processors, despite their powerful integrated graphics, energy efficiency, and AI capabilities, are focused on different tasks than high-performance desktop processors like the i9-14900K, which is aimed at maximum performance in traditional computing tasks.

Conclusion

Over the past seven years, since the introduction of the NVIDIA RTX series of graphics accelerators in 2018 and the release of the latest versions of Gromacs, the speed of all-atom MD calculations has increased more than fivefold. Moreover, since 2014, MD performance has increased by almost 35 times when comparing a top-of-the-line gaming CPU Intel Core i7-4790K and a top video accelerator NVIDIA GTX 980, both released in 2014 [5]. All this has become possible thanks to the rapid and continuous development of the computing capabilities of central and, in particular, graphic processors.

Our comparative performance analysis of GPU-based computing architectures for all-atom MD simulations of biomolecular systems not only provides guidance on choosing the best computing solution in terms of the price-performance ratio, but also shows the maximum potential

Table 4. Single-node Gromacs 2025.2 molecular dynamics performance (ns/day) for tubulin tetramer and WB-1000 molecular systems on various GPUs for Intel Ultra 9 285K CPU (results with GPU-resident option on and off are equal)

Test number	GPU	Series name	Release date	Performance for tubulin tetramer, ns/day	Performance for WB-1000, ns/day
1	No GPU	—	—	15	1.6
2	RTX3080	Ampere	Sep 2020	65	11
3	RTX3090	Ampere	Sep 2020	77	12
4	RTX4080	Ada Lovelace	Nov 2022	126	18
5	RTX4090	Ada Lovelace	Oct 2022	99	25
6	RTX5080	Blackwell	Jan 2025	136	17
7	RTX5090	Blackwell	Jan 2025	104	33

computational performance that modern CPUs and GPUs can provide. In addition, we show how the latest software versions and computational options can improve the performance of the calculations. MD of the biomolecular system containing a tubulin protofilament in an explicitly specified solvent consisting of more than 300 thousand atoms can be studied with a performance of 232 ns/day at time step 2 fs when using a single-node computer with the latest CPU and GPU generation architecture (Intel Core i9-14900KF and Nvidia RTX5090, respectively).

Acknowledgments

This study was supported by the Russian Science Foundation, project No. 24-74-00002, https://rscf.ru/en/project/24-74-00002/.

References

1. Abraham, M., Alekseenko, A., Bergh, C., *et al.*: GROMACS 2023.2 Manual (Jul 2023). https://doi.org/10.5281/zenodo.8134388

2. Abraham, M.J., Murtola, T., Schulz, R., *et al.*: GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX 1, 19–25 (2015). https://doi.org/10.1016/j.softx.2015.06.001

3. Ewald, P.P.: Die Berechnung optischer und elektrostatischer Gitterpotentiale. Annalen der Physik 369(3), 253–287 (1921). https://doi.org/10.1002/andp.19213690304

4. Fedorov, V.A., Kholina, E.G., Gudimchuk, N.B., Kovalenko, I.B.: High-performance computing of microtubule protofilament dynamics by means of all-atom molecular modeling. Supercomputing Frontiers and Innovations 10(4), 62–68 (2023). https://doi.org/10.14529/jsfi230406

5. Fedorov, V.A., Kholina, E.G., Kovalenko, I.B., Gudimchuk, N.B.: Performance analysis of different computational architectures: molecular dynamics in application to protein assemblies, illustrated by microtubule and electron transfer proteins. Supercomputing Frontiers and Innovations 5(4), 111–114 (2018). https://doi.org/10.14529/jsfi180414

6. Gudimchuk, N.B., McIntosh, J.R.: Regulation of microtubule dynamics, mechanics and function through the growing tip. Nature Reviews Molecular Cell Biology 22(12), 777–795 (2021). https://doi.org/10.1038/s41580-021-00399-x

7. Kellogg, E.H., Hejab, N.M., Howes, S., *et al.*: Insights into the distinct mechanisms of action of taxane and non-taxane microtubule stabilizers from cryo-EM structures. Journal of Molecular Biology 429(5), 633–646 (2017). https://doi.org/10.1016/j.jmb.2017.01.001

8. Liao, Y., Ithurbide, S., Evenhuis, C., *et al.*: Cell division in the archaeon Haloferax volcanii relies on two FtsZ proteins with distinct functions in division ring assembly and constriction. Nature Microbiology 6, 594–605 (2021). https://doi.org/10.1038/s41564-021-00894-z

9. Loose, M., Mitchison, T.J.: The bacterial cell division proteins FtsA and FtsZ self-organize into dynamic cytoskeletal patterns. Nature Cell Biology 16(1), 38–46 (2014). https://doi.org/10.1038/ncb2885

10. MacKerell Jr, A.D., Feig, M., Brooks, C.L.: Improved treatment of the protein backbone in empirical force fields. Journal of the American Chemical Society 126(3), 698–699 (2004). https://doi.org/10.1021/ja036959e

DOI: 10.14529/jsfi260103

Ionic and Water-Saturated Clusters in Self-Healing Polydimethylsiloxanes Modelled by Molecular Dynamics

Tatiana M. Makarova[1], *Ekaterina V. Bartashevich*[1]

Elaboration of new self-healing polymer materials with improved properties such as room-temperature healing and sufficient mechanical performance is a complicated task. In our study, we present the groundwork for construction of multicomponent systems for polydimethylsiloxane-based polymers via condensation in molecular dynamics resembling the simulated annealing protocol. Upon the self-organization according to the force field that model the intermolecular interactions, all the compounds of the "siloxane equilibration" system reproducibly assembles in a polydisperse structure with ionic and water-saturated aggregates. Around these aggregates, the negatively charged polymer terminal groups are oriented, along with ions of initiators and residual water. At the temperature of self-healing, the outer layers of the aggregates intensively exchange with each other. Thus, the molecular dynamic simulations shed light on crucial structural and dynamical properties on a molecular level that can influence the self-healing process, which would be useful in further targeted development of these materials having a prospect in cable products.

Keywords: self-healing materials, PDMS, "siloxane equilibration", MD simulation.

Introduction

Nowadays the requirement for high-tech polymer-contained devices in various areas permanently increases, entailing the problems of labored reparation of them in case of mechanical or electrical damage. This problem particularly manifests in electronic devices, when a breakdown can distort the mechanical integrity of a dielectric material disabling thus a rather expensive device. To overcome this problem, polymers with self-healing properties, i.e., able to heal the occurring mechanical defects in it, have attracted significant interest as promising participants of a sustainable development concept [3, 27]. Self-healing materials are comprised by various types of polymers possessing exchangeable bonds or interactions with relatively low activation energy threshold, which are able to reversibly cleavage and reform, providing mobility of the chains and regaining mechanical strength and integrity after the damage. These exchangeable bonds may be noncovalent (hydrogen bonds [9, 52, 57], ionic interactions [12, 29, 37] etc.) or covalent [46, 55] (for example, Diels—Alder reversible condensation [6, 17], donor-acceptor metal coordination [15, 28, 34, 51], imine bond [47, 60], disulfide bond [35, 42], dynamic urea bond [56], boron ester bond [7, 10, 58], oxime ester [20] and many others), or the polymer can contain a combination of both types.

Polysiloxane-based self-healing materials are comprised by chains of the Si–O– sequence with the feature of recombination of Si–O bonds between two different chains. High flexibility and low rotation energy barrier of the polysiloxane chains [32], provide excellent mobility of the chains and thus high efficacy of self-healing, however, along with low mechanical performance. Nevertheless, in case of electronic devices including those in biomedicine, the effectiveness, rate and autonomy of self-healing is a priority. Also, such peculiarities of self-healing polysiloxanes as optical transparency, flexibility, high environmental and thermal stability provide their wide potential application in electronics and wearable devices [19, 26, 53], soft robotics [58], microfluidics [39] and special indispensability for biomedical devices due to their complete biocompatibility [33, 50, 54]. Extremely low (about -120°C) glass transition temperature, T_g, of polysiloxanes

[1]South Ural State University, Chelyabinsk, Russian Federation

makes them a very attractive object for modifications in order to approach the room temperature effective self-healing via, for example, adding of a silicone exchange catalyst [8, 24] or additional type of reversible bonds with lower activation energy [14, 25, 40]. However, these methods may significantly increase the material cost and, what is more important, decrease one of the most valuable unique polysiloxane properties such as biocompatibility. Therefore, development of the room temperature silicone self-healing materials with minimal chemical interventions is a challenging task. To find the pathway of such modification, a detailed picture of the self-healing mechanism at the molecular level would be indispensable.

In all the available literature, the self-healing of polyorganosiloxanes is attributed to the "siloxane equilibration" reaction, in which an electronegative terminus of one chain attacks a middle of another chain by S_N2 mechanism, upon which the second chain is shortened, and the first one is lengthened. The S_N2 mechanism of this reaction is confirmed both by quantum chemical calculations using DFT methods [13] and experimental methods [36, 44]. Indeed, it is the presence of charged $–Si(CH_3)_2O^-$ terminal groups that imparts self-healing properties to siloxane materials, while neutralization of the charged chain ends deprives the materials of this ability [59]. However, detailed atomic and molecular structure of the multicomponent polymeric system could provide crucial information for understanding the self-healing process and its relation with desired mechanical performance.

In our study, for computational investigations, the "siloxane equilibration" compounds, previously obtained as follows [38, 41], were selected as the most simple compounds of their class. These polymers were synthesized by anionic polymerization of the D_4 reagent (octamethyltetracyclosiloxane) with a small amount of its dimer, *bis*-D_4 reagent, every molecule of which contains C–C bond and, being included into the polymer, provides ethylene crosslinks between two linear PDMS chains. The polymerization is initiated by an anion, typically OH^-, and the most common initiators are $N(CH_3)_4OH$ (TMAOH) [59] or KOH. It was found that the quality and type of the initiator can influence the self-healing temperature and other material properties. This and other features can hardly be explained without understanding of the PDMS-based "siloxane equilibration" material organization at the molecular level. The reliable structural and dynamical modeling could reveal some crutial heterogeneity in its structure and its role in the self-healing processes.

Therefore, molecular dynamics (MD) simulations were performed for investigation of this issue. Though this method does not imply rearrangements of covalent bonds, it is indispensable to model the polymer organisation at the molecular level and assess the possibility of a certain reaction according to the distance between the reaction centers. Task-oriented material design with purpose properties, attempted to reach certain values by variation of chemical composition and structure, is more efficient when it is based on a structure-property relationship model. The MD simulation method has approved itself to be appropriate and conclusive for investigation of irregular structure of polymer materials, including those with self-healing abilities [23]. In this research, we exploited it to obtain the three-dimensional dynamic picture of the "siloxane equilibration" materials and to establish the processes guiding its self-healing.

Virtual simulation provides insights into systems with different parameters easily fine-tuning in the model compared to the synthesis, such as the initiator effectiveness or a small water impurity in the investigated polymer. To address this issue, we designed several theoretical systems along with experimentally based varying parameters such as the counterion type, the initiator effectiveness or an average chain length.

To establish the detailed structure of the self-healing PDMS materials, a simulation protocol that would chemically-friendly guide the system self-organization is required. If any initial guess for the irregular multicomponent polymer structure is absent, it is possible to start from the cell with randomly distributed components in it, which is anyway far from the real structure and thus requires the surmounting of a lot of substantial energy barriers on the way to the potential minimum with the most reliable structure of the material organization. The simulated annealing protocol [43] provides such kind of a global optimization strategy which at the beginning point supplies the system with an energy sufficient to overcome barriers on its energetic landscape and then gradually reduces the additional energy which steers the system towards the global energy minimum. In our study, we adopted a similar strategy to the MD simulation of PDMS-based "siloxane equilibration" systems containing ions of different initiators, $N(CH_3)_4OH$ or KOH, in various concentration, and water molecules. The results of the proposed modelling protocol and the influence of the concentration of different ions were thoroughly analyzed.

The present article is organized into three main sections. The "Methods" section is devoted to computational protocol details, the "Results and discussion" section contains general description of MD simulation data and their interpretation. Conclusion points out the significance of the obtained results for understanding of the self-healing mechanism and potential directions of material development for its improvement.

1. Methods

1.1. System Design and Construction of the Polymer MM Models

We assumed that upon polymerization, each initiator anion creates one unbranched PDMS chain, while each *bis*-D_4 crosslink consumes two initiator anions for every opening ring, creating one cross-linked polymer molecule. Therefore, the average chain length in the system is defined by the ratios between the initial components (D_4, *bis*-D_4 and the initiator, KOH or $(CH_3)_4NOH$) and one more value – the initiator efficiency, i.e., the percentage of the initiator ions that participated in the nucleophilic substitution reaction and created an "opened" $-(CH_3)_2Si-OH$ group. Modeling of the systems with M_w according to the NMR and GPC data (58 kDa and 66 kDa) implied certain KOH efficacy and thus the presence of a certain number of unreacted OH^- in the system along with the initiator counterions. The corresponding amount of these ions were added to the systems (Tab. 1). Also, systems with different initiator efficiencies and calculated from it average chain lengths were similarly constructed.

The minimal number of chains was selected in order to provide the unfolded conformation of them in the simulation. For systems with large M_w amount above 40 chains was sufficient for this, while in the systems with a higher initiator efficiency we had opportunity to simulate a larger number chains. The lengths of individual chains were generated according to the Poisson distribution with a given mean, and these chains were arranged within a cubic cell at equal distances from each other in each of the three dimensions. Crosslinks were introduced into some of the chains according to the mass fraction of *bis*-D_4. The chain and the location within it for each crosslink were also chosen randomly, with the only assumption being that the crosslinks or chain ends must be separated from each other by at least six DMS residues.

Though PDMS is a highly hydrophobic material, it was supposed to absorb some amount of water. Thus, certain excess of water in the amount of 0.5% mass. was included into the models as a preliminary upper bound of water content.

Systems with $(CH_3)_4NOH$ were constructed in a similar manner, but K^+ ions were replaced with $(CH_3)_4N^+$ ions using an authoring Python script with the *cmd* library, and are denoted hereafter as PDMS-TMA systems (in names like PDMS-TMAOH-13 the last number means initiator effectiveness, here 13%). Otherwise, systems with KOH as an initiator are denoted as PDMS-KOH.

The full list of the constructed systems is the following:

- PDMS-K-100 with the initiator effectiveness of 100%;
- PDMS-K-30 with the initiator effectiveness of 30%;
- PDMS-K-13 (13.2% effectiveness, corresponding to the experimental M_w of 58 kDa);
- PDMS-K-10 (10.4% effectiveness, corresponding to the experimental M_w of 66 kDa);
- PDMS-TMA-13.

For them, the quantitative compound and other details are given in the Tab. 1. In it, M_n is given in a number of DMS residues, while nOH^- denotes the number of free unreacted anions in the system. After a number of chains and "=" its quantitative distribution along every dimension is given.

Table 1. The quantitative (in a number of molecules) composition of the "siloxane equilibration" material simulation cells

System name	Polymer chains	Cell, nm	M_n	Crosslinks	nOH^-	Cations	Water
PDMS-KOH-100	$216 = 2 \times 6 \times 18$	39	53	12	0	228	222
PDMS-KOH-30	$64 = 1 \times 2 \times 32$	58	197	12	177	253	248
PDMS-KOH-13	$64 = 1 \times 4 \times 16$	46	782	47	732	843	986
PDMS-KOH-10	$48 = 1 \times 4 \times 12$	52	890	41	756	845	843
PDMS-TMAOH-13	$64 = 1 \times 4 \times 16$	46	782	47	732	843	986

The construction of the systems was performed using a Python script with the *cmd* module. For systems with low initiator effectiveness (10 and 13%) long chains were folded into "rolls" from an unfolded D_4 structure with Si–O–Si–O and O–Si–O–Si torsion angles $\tilde{1}76°$ – the minimal value that allowed the chains to be folded without steric interference (Fig. 1a). After this, the cell was filled with water, and the required amount of randomly distributed ions and water molecules were added using the *gmx solvate* and *gmx genion* instruments of GROMACS software, respectively.

The molecular mechanical models of the PDMS residues, both for the linear and branched fragments, were prepared using the General Amber force field [49]. Partial charges were evaluated according to the RESP model [1]. The lacking parameters of covalent bonds and angles for DMS and crosslink residues were calculated by DFT method with M06–2X functional and 6–$311^{++}G^{**}$ basis set [30] along with parameters from [16].

1.2. Molecular Dynamics Simulation Protocol

We began the simulation at high temperature (about 800–1000 K) which ensured an easy overcoming of all the conformational barriers, and then linearly decreased it down to the room temperature, providing thus the system movement towards the most energetically optimal structure guided by forces of intermolecular interactions from the forcefield.

The MD simulations were carried out using GROMACS 2019.4 software [22, 48]. The time step of integration for equations of motion was 2 fs. The system coordinates were written to the *xtc* trajectory file every 20 ps. The lengths of all bonds were controlled using the LINCS

algorithm [21]. The temperature was maintained by the velocity rescaling thermostat with additional stochastic correction [4] with 0.1 ps coupling time. The pressure in NPT simulations was controlled by Berendsen barostat [2] with the coupling time of 5 ps. Ewald particle mesh with the sixth order of interpolation and a 0.1 nm grid step was used to threat long-range electrostatic interactions [11].

Every MD simulation was preceded by a geometry optimisation by the Broyden—Fletcher—Goldfarb—Shanno algorithm [5]. The sequence of all calculation procedures upon the system modelling is given in Tab. 1.

1.2.1. Results and discussion

At the initial stage of the system modelling, initial unit cells with PDMS chains were constructed. We applied the software previously developed by us for modeling of lesser polymer cells [31] that was based on various probabilistic algorithms which essentially randomly distributed components in the cell, creating a system that reproduces the required ratios of initial reagents and the average chain length (with the individual chain lengths distributed according to the Poissons law). Thus constructed systems with PDMS chains, ions, residual water and D_4 reactant underwent a procedure of *in silico* condensation to achieve a deep equilibration during the system self-organization.

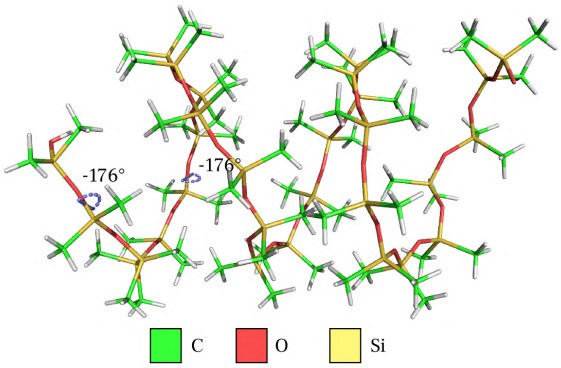
In this procedure, after potential energy optimization, MD simulation is to be carried out at constant volume and high temperature that is significantly higher than the physical polymer could sustain (up to 1073K, see Tab. 1). During this simulation, the polymer chains and other components intensively mix with each other, remaining distributed throughout the cell without significant condensation in a confined region.

At this step, for systems with long PDMS chain a crutial problem arose: the polymer chains began to "curl" folding on themselves (which would result into unnatural conformations upon further condensation) rather then interact with each other due to large interchain distances. Prevention of the curling by mere temperature increase entailed a high probability of triggering the LINCS algorithm, aborting the simulation. A solution was found in reducing the free space in the initial cell by folding the polymer chains into artificial helices as tightly as the PDMS structure allowed without a steric clash of the van der Waals radii of the atoms (Fig. 1a). The dense PDMS chains packed into such kind of "rolls" did not inflict tense system conformation in further simulations: the chains unrolled during the first tens of the NVT trajectory (Fig. 1c, left). Therefore, this chain twisting was used to construct the systems with long polymer chains.

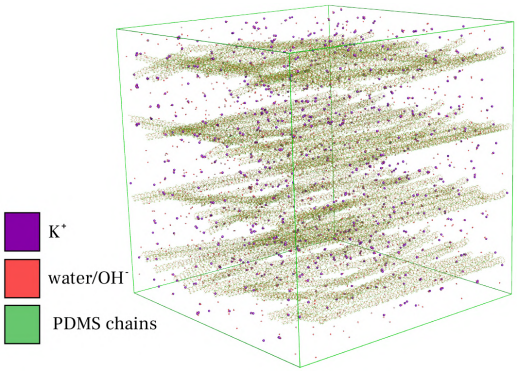
At the next step of the simulation protocol, the system was slowly cooled to ambient or 100°C temperature. Upon this process, molecules and ions aggregated according to their interaction forces, thus adopting the most natural conformation and mutual arrangement (see Fig. 1c, right).

At the final stage, the systems at 25 or 100°C were extracted from the condensation trajectory after stabilization as starting systems for further simulations. After the geometry optimization, final trajectories of 500 ns length were obtained for each type of the system (all kinds of simulated systems are described in Methods). The final density that the modeled systems reached was about 0.95 (Fig. 1), which falls within the range of 0.91–1.00 for PDMS materials [45].

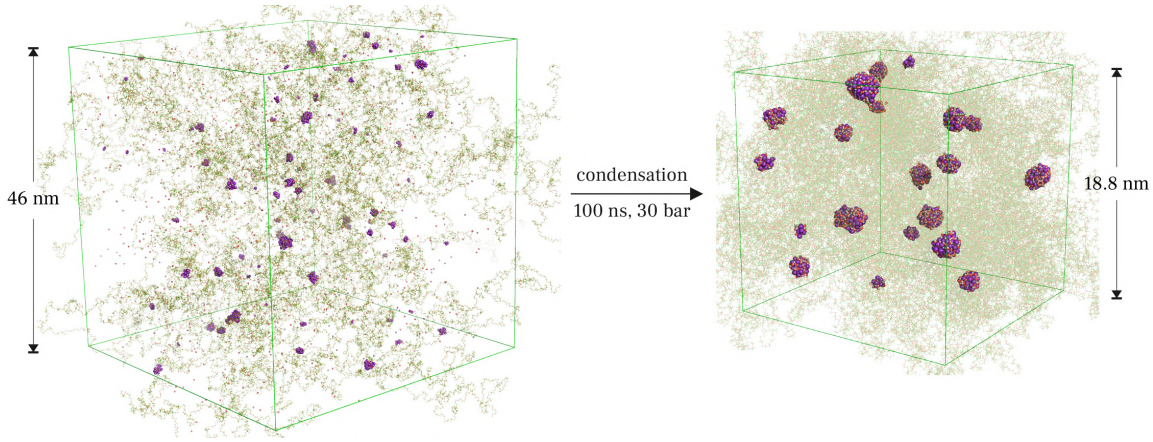
In all the simulated systems we observed an aggregation of positively charged counter-ions and the self-organization of charged and polar groups and molecules around these aggregates (Fig. 2). These aggregates assembled reproducibly in various cells with different compositions and in all the MD runs with different ratios of components (Tab. 1).



(a) Initial compacted conformation of PDMS chains



(b) General view of the starting system with "rolled" PDMS chains and ions

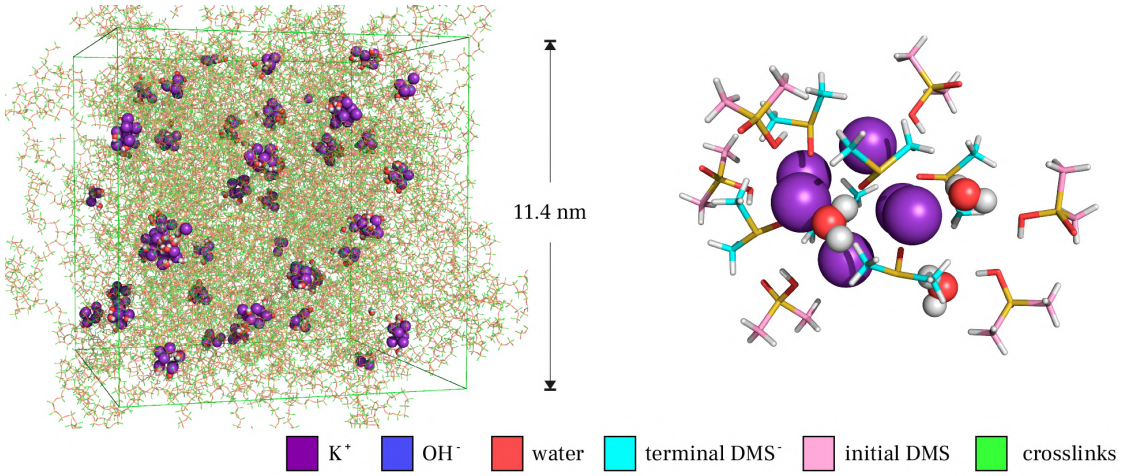


(c) Condensation of the polymer material cell after short NVT simulation

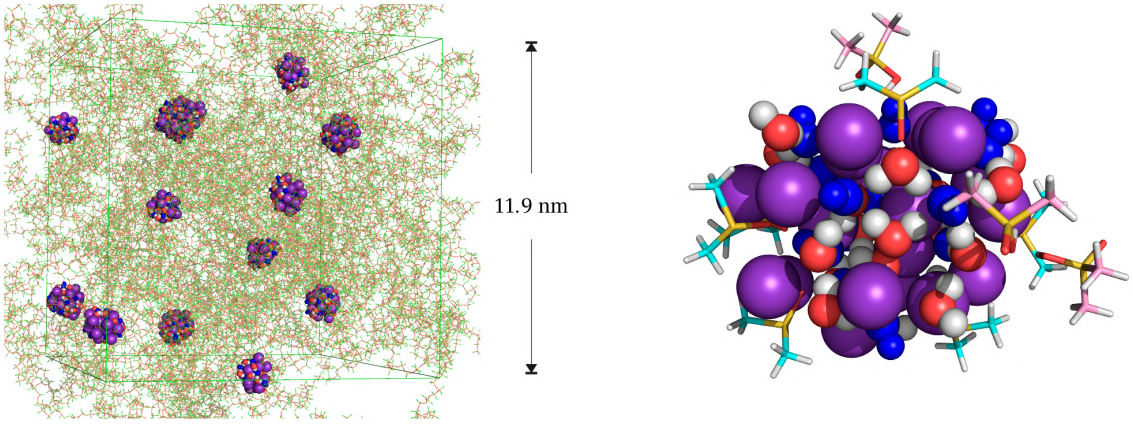
Figure 1. Construction and condensation of polymer material simulation cells

We observed that the size and distribution of the aggregates heavily depended on the initiator effectiveness. For a system with 100% effectiveness (without free OH^- in the polymer), the average ionic aggregate size was about 5–6 cations (Fig. 2a). As initiator efficiency decreased, the aggregate size increased due to residual OH^- anions incorporating between cations, thereby enhancing their retention (Fig. 2b). The greater was the fraction of unreacted initiator remnants, the larger was the average size of the aggregates, increasing up to more than 40 K^+ cations at 10–13% effectiveness (Tab. 1). With extremely low initiator effectiveness (10–13%), the number of K^+ and OH^- ions in the system was nearly equal, which provided formation of stable ionic aggregates of many tens of ions (Fig. 2c).

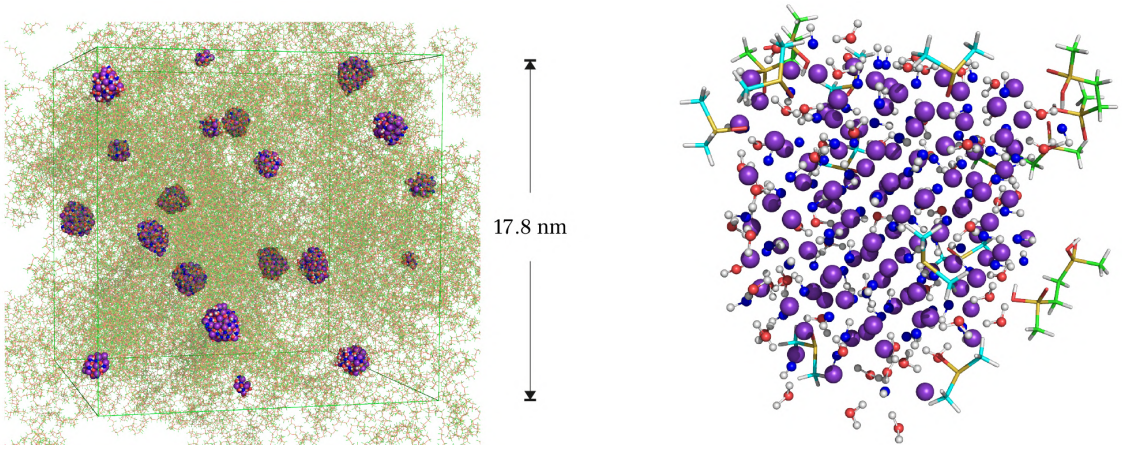
The aggregates in our systems had a more complex structure than any other ionic aggregates previously described [18]. All negatively charged termini of the polysiloxane chains were "anchored" by cationic aggregates. All water molecules, as well as polar hydrophilic groups at the beginning of the chains at the site of nucleophilic OH^- attachement, were also coordinated around the ionic aggregates, so that the space between the aggregates was filled only with low-polarity lipophilic PDMS chains. At larger aggregate sizes, they attracted even less polar residues, such as forming ethylene crosslinks in the chains (Fig. 2c, right). Thus, all the negatively charged groups that initiate the siloxane exchange reaction appeared to be trapped on the surface of the aggregates in our model. This should be taken into account for future investigation of the siloxane equilibration kinetics.



(a) Initiator effectiveness 100%, no OH^- ions in the system



(b) Initiator effectiveness 30%



(c) Initiator effectiveness 13%

Figure 2. Condensed simulation cells with ionic aggragates (on the left) and organization of individual aggregates (on the right, the common color scheme is given on the top) at different effectivenes of KOH polymerization initiator

In the obtained MD simulations trajectories integrity of K^+ and OH^- aggregates and their coupling with water molecules or terminal PDMS residues were investigated. It was expected that the dynamics of inner layers (consisting of K^+ and OH^- ions) and outer layers (water, PDMS residues) of the aggregates could provide an insight into how the structure determines the materials properties, including the self-healing capacity.

As it was mentioned above, an increase in the amount of unreacted OH^- at reduced initiator effectiveness, enabled a formation of increasingly larger ionic clusters, where K^+ and OH^- alternate with each other. Thus, the considered decrease in initiator efficiency led to the formation of increasingly larger and rarified ionic aggregates in the bulk of polymeric cell. Additionally, significant differences in the dynamics of the aggregates were detected. During the simulation, the aggregates and their environment underwent two types of dynamic changes. The first one involved the aggregate "core" itself – the cluster of K^+ and OH^- ions – which could split into two parts with their subsequent separation (however, they could further merge again). These fission events were the more common, the higher the initiator efficiency was, i.e., the smaller were the clusters themselves. For the systems without any free OH^-, aggregates easily disintegrated, exchanged their core ions, or merged, as they were extremely small (with an average of 5 K^+ ions) and located close to each other (Tab. 1). As the ionic clusters grew with increasing of OH^- content, the aggregates became more stabilized. In particular, in systems with 10–13% efficiency, ion exchange between aggregates was practically non-existent; however, another interesting process prevailed.

The outer layer of such clusters, formed by water and chain end groups, proved to be even more mobile. Individual groups in this layer could reversibly separate from the cluster, and if the distance became large enough, return not to their own cluster, but to a neighbouring one. Thus, during the simulations, an exchange of outer layer components occurred between clusters. Expectedly, this exchange became more intense with increasing temperature. For PDMS-KOH-13 at room temperature, no such exchange events were observed within 0.5 μs of simulation, but as the temperature increased up to 100°C, the "hopping" of the PDMS termini between aggregates occurred many times within the same trajectory span (which is illustrated on Fig. 2). Since self-healing is always related with the redistribution of reversible interactions in the material, which provides the relative mobility of its chains, the redistribution of terminal PDMS groups between the aggregates in the MD trajectory might influence the self-healing mechanism of the PDMS-KOH material.

Water expectedly appeared to be the most mobile component of the outer layer of the aggregates, so the water exchange between aggregates was the most frequent event (see Tab. 1). It is to be established in further modeling, whether water exchange alleviates the inter-aggregate "hopping" of the PDMS residues and thus is an important active participant of the self-healing process, or this is an independent process.

With substitution of the KOH initiator with $N(CH_3)_4OH$ in the PDMS-TMAOH-13 systems, the behaviour of the aggregates was completely different. The aggregates themselves became unstable, when one part of them could shift and separate from another; however, such separations of the aggregates were reversible. What is more important, no exchange of PDMS residues between the outer aggregate layers was detected at either 25°C or 100°C. However, at all these temperatures, the real PDMS-TMAOH system provides self-healing. Therefore, the self-healing ability of PDMS-TMAOH systems should be mainly attributed to another physicochemical processes: general mobility of the PDMS chains coupled with ionic aggregates and "siloxane equilibration" reaction.

To the contrast, PDMS-KOH systems demonstrate a unique behaviour: rapid redistribution of internal noncovalent interactions between stable "anchor" aggregates. This process presumably can provide self-healing ability along with low viscosity, and thus could explain self-healing of PDMS-KOH with relatively good mechanical performance.

Conclusion

Molecular dynamic modelling of self-healing PDMS-based "siloxane equilibration" materials enabled to obtain the details of polydisperse structure depending on their composition. The main distinguishing feature of the systems with KOH or $N(CH_3)_4OH$ initiators was that initiator counterions are concentrated into aggregates, surrounded by negatively charged groups from the ends of the polymer chains. As initiator efficiency decreased, these aggregates became increasingly larger due to the intercalation of residual OH^- between potassium ions, additionally stabilizing these aggregates. They also coordinated the polar groups of the system around themselves: water molecules, as well as the initial polymer residues that carried the attached OH^- ion and the cross-linking residues. During the MD simulation of materials with KOH as initiator with 10–13% effectiveness at 100°C, these polar groups extensively migrated between stable ionic aggregates. Water was especially mobile in this regard, but the initial residues also underwent several "switches" during the 500 ns trajectory. At 25°C, these groups had insufficient kinetic energy to switch from one aggregate and attach to another one. These simulation data are in line with the experimental fact that the material with KOH as initiator undergoes self-healing only at the temperature above 100°C.

Another noteworthy point following from the MD results deserves special attention for the future study of self-healing properties. It is the balance between two interrelated processes: the "siloxane equilibration" reaction and the dynamic of ionic particles in self-healing process. If the reaction requires free $–(CH_3)_2SiO^-$ termini, which, according to the simulation data, are preferably bound with ionic aggregates, then the siloxane equilibration should be considered in the light of the aggregate structure. The aggregate itself, its size, compound and stability is expected to impact the kinetics of the "siloxane equilibration" reaction, and it is a comprehensive task for further investigations.

Thus, the structural models obtained via MD simulation condensation protocol discovered the existence of stable ionic aggregates in the PDMS polymer bulk. It enatails that the self-healing process can be determined by the structure and dynamics of these ionic aggregates, which would open new perspectives of modifications by controlling the size, distribution or kinetic parameters of group exchange of water-ionic aggregates in polydisperse multicomponent system.

Acknowledgements

This research was funded by the Ministry of Science and Higher Education of the Russian Federation (grant FENU 2024-0003).

The authors would like to express their acknowledgement to the Artificial Intelligence and Quantum Technologies (SEC "AIQ") Scientific and Educational Center of South Ural State University.

References

1. Bayly, C.I., Cieplak, P., Cornell, W.D., *et al.*: A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: The RESP model. Journal of Physical Chemistry 97(40), 10269–10280 (1993). https://doi.org/10.1021/j100142a004

2. Berendsen, H.J., Postma, J.P., Van Gunsteren, W.F., *et al.*: Molecular dynamics with coupling to an external bath. The Journal of Chemical Physics 81(8), 3684–3690 (1984). https://doi.org/10.1063/1.448118

3. Bergman, S.D., Wudl, F.: Re-Mendable Polymers. In: Self Healing Materials. Springer Series in Materials Science, pp. 45–68. Springer (2007). https://doi.org/10.1007/978-1-4020-6250-6_3

4. Bussi, G., Donadio, D., Parrinello, M.: Canonical sampling through velocity rescaling. Journal of Chemical Physics 126(1) (2007). https://doi.org/10.1063/1.2408420

5. Byrd, R.H., Lu, P., Nocedal, J., *et al.*: A Limited Memory Algorithm for Bound Constrained Optimization. SIAM Journal on Scientific Computing 16(5), 1190–1208 (1995). https://doi.org/10.1137/0916069

6. Chen, X., Dam, M.A., Ono, K., *et al.*: A thermally re-mendable cross-linked polymeric material. Science 295(5560), 1698–1702 (2002). https://doi.org/10.1126/science.1065879

7. Chen, Y., Tang, Z., Liu, Y., *et al.*: Mechanically Robust, Self-Healable, and Reprocessable Elastomers Enabled by Dynamic Dual Cross-Links. Macromolecules 52(10), 3805–3812 (2019). https://doi.org/10.1021/acs.macromol.9b00419

8. Cho, S.H., White, S.R., Braun, P.V.: Room-temperature polydimethylsiloxane-based self-healing polymers. Chemistry of Materials 24(21), 4209–4214 (2012). https://doi.org/10.1021/cm302501b

9. Cordier, P., Tournilhac, F., Soulié-Ziakovic, C., *et al.*: Self-healing and thermoreversible rubber from supramolecular assembly. Nature 451(7181), 977–980 (2008). https://doi.org/10.1038/nature06669

10. Cromwell, O.R., Chung, J., Guan, Z.: Malleable and Self-Healing Covalent Polymer Networks through Tunable Dynamic Boronic Ester Bonds. Journal of the American Chemical Society 137(20), 6492–6495 (2015). https://doi.org/10.1021/jacs.5b03551

11. Darden, T., York, D., Pedersen, L.: Particle mesh Ewald: An $N \cdot \log(N)$ method for Ewald sums in large systems. The Journal of Chemical Physics 98(12), 10089–10092 (1993). https://doi.org/10.1063/1.464397

12. Das, A., Sallat, A., Böhme, F., *et al.*: Ionic Modification Turns Commercial Rubber into a Self-Healing Material. ACS Applied Materials and Interfaces 7(37), 20623–20630 (2015). https://doi.org/10.1021/acsami.5b05041

13. Debsharma, T., Nguyen, L.T., Maliszewski, B.P., *et al.*: Eliminating creep in vitrimers using temperature-resilient siloxane exchange chemistry and N-heterocyclic carbenes. Chemical Science 16(21), 9337–9347 (2025). https://doi.org/10.1039/d4sc06278g

14. Deriabin, K.V., Filippova, S.S., Islamova, R.M.: Self-Healing Silicone Materials: Looking Back and Moving Forward. Biomimetics 8(3), 286 (2023). https://doi.org/10.3390/biomimetics8030286

15. Deriabin, K.V., Ignatova, N.A., Kirichenko, S.O., *et al.*: Structural features of polymer ligand environments dramatically affect the mechanical and room-temperature self-healing properties of cobalt(ii)-incorporating polysiloxanes. Organometallics 40(15), 2750–2760 (Jul 2021). https://doi.org/10.1021/acs.organomet.1c00392

16. Dong, X., Yuan, X., Song, Z., *et al.*: The development of an Amber-compatible organosilane force field for drug-like small molecules. Physical Chemistry Chemical Physics 23(22), 12582–12591 (2021). https://doi.org/10.1039/d1cp01169c

17. Du, G., Mao, A., Yu, J., *et al.*: Nacre-mimetic composite with intrinsic self-healing and shape-programming capability. Nature Communications 10(1) (2019). https://doi.org/10.1038/s41467-019-08643-x

18. Eisenberg, A., Hird, B., Moore, R.B.: A New Multiplet-Cluster Model for the Morphology of Random Ionomers. Macromolecules 23(18), 4098–4107 (1990). https://doi.org/10.1021/ma00220a012

19. Ghosh, K., Morgan, A., Garcia-Casas, X., *et al.*: Tailoring of Self-Healable Polydimethylsiloxane Films for Mechanical Energy Harvesting. ACS Applied Energy Materials 7(19), 8185–8195 (2024). https://doi.org/10.1021/acsaem.4c01275

20. He, C., Shi, S., Wang, D., *et al.*: Poly(oxime-ester) Vitrimers with Catalyst-Free Bond Exchange. Journal of the American Chemical Society 141(35), 13753–13757 (2019). https://doi.org/10.1021/jacs.9b06668

21. Hess, B., Bekker, H., Berendsen, H.J., *et al.*: LINCS: A Linear Constraint Solver for molecular simulations. Journal of Computational Chemistry 18(12), 1463–1472 (1997). https://doi.org/10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H

22. Hess, B., Kutzner, C., Van Der Spoel, D., *et al.*: GRGMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. Journal of Chemical Theory and Computation 4(3), 435–447 (2008). https://doi.org/10.1021/ct700301q

23. Karatrantos, A.V., Couture, O., Hesse, C., *et al.*: Molecular Simulation of Covalent Adaptable Networks and Vitrimers: A Review. Polymers 16(10), 1373 (2024). https://doi.org/10.3390/polym16101373

24. Krug, D.J., Asuncion, M.Z., Laine, R.M.: Facile Approach to Recycling Highly Cross-Linked Thermoset Silicone Resins under Ambient Conditions. ACS Omega 4(2), 3782–3789 (2019). https://doi.org/10.1021/acsomega.8b02927

25. Lai, J.C., Jia, X.Y., Wang, D.P., *et al.*: Thermodynamically stable whilst kinetically labile coordination bonds lead to strong and tough self-healing polymers. Nature Communications 10, 1164 (2019). https://doi.org/10.1038/s41467-019-09130-z

26. Latif, S., Amin, S., Haroon, S.S., *et al.*: Self-healing materials for electronic applications: An overview. Materials Research Express 6(6) (2019). https://doi.org/10.1088/2053-1591/ab0f4c

27. Li, B., Cao, P.F., Saito, T., *et al.*: Intrinsically Self-Healing Polymers: From Mechanistic Insight to Current Challenges. Chemical Reviews 123(2) (2023). https://doi.org/10.1021/acs.chemrev.2c00575

28. Li, C.H., Wang, C., Keplinger, C., *et al.*: A highly stretchable autonomous self-healing elastomer. Nature Chemistry 8(6), 618–624 (2016). https://doi.org/10.1038/nchem.2492

29. Luo, F., Sun, T.L., Nakajima, T., *et al.*: Oppositely charged polyelectrolytes form tough, self-healing, and rebuildable hydrogels. Advanced Materials 27(17), 2722–2727 (2015). https://doi.org/10.1002/adma.201500140

30. Makarov, G.I., Makarova, T.M.: General AMBER force field parameters for modeling polyalkylsiloxane chains. Mendeleev Communications 35(2), 221–223 (2025). https://doi.org/10.71267/mencom.7580

31. Makarova, T.M., Bartashevich, E.V.: Construction of Self-Healing PDMS Materials Models by Supercomputer MD Simulations. In: 19th Proceedings of the International Scientific Conference, PCT'2025, Moscow, Russia, April 8-10, 2025, pp. 36–43. Chelyabinsk: Publishing of the South Ural State University (2025). https://doi.org/10.14529/pct2025

32. Mark, J.E.: Some interesting things about polysiloxanes. Accounts of Chemical Research 37(12), 946–953 (2004). https://doi.org/10.1021/ar030279z

33. Miranda, I., Souza, A., Sousa, P., *et al.*: Properties and applications of PDMS for biomedical engineering: A review. J. Funct. Biomater. 13(1), 2 (2022). https://doi.org/10.3390/jfb13010002

34. Oh, J.Y., Son, D., Katsumata, T., *et al.*: Stretchable self-healable semiconducting polymer film for active-matrix strain-sensing array. Science Advances 5(11) (2019). https://doi.org/10.1126/sciadv.aav3097

35. Oku, T., Furusho, Y., Takata, T.: A concept for recyclable cross-linked polymers: Topologically networked polyrotaxane capable of undergoing reversible assembly and disassembly. Angewandte Chemie - International Edition 43(8), 966–969 (2004). https://doi.org/10.1002/anie.200353046

36. Osthoff, R.C., Bueche, A.M., Grubb, W.T.: Chemical Stress-Relaxation of Polydimethylsiloxane Elastomers. Journal of the American Chemical Society 76(18), 4659–4663 (1954). https://doi.org/10.1021/ja01647a052

37. Peng, Y., Zhao, L., Yang, C., *et al.*: Super tough and strong self-healing elastomers based on polyampholytes. Journal of Materials Chemistry A 6(39), 19066–19074 (2018). https://doi.org/10.1039/c8ta06561f

38. Prokudin, A.V., Dziuba, M.A., Safonov, V.I., *et al.*: Self-healing "siloxane equilibrium" materials after low-power electric breakdown in small-volume cells. Mendeleev Commun. 36(3), 308–310 (2026). https://doi.org/10.71267/mencom.7926

39. Raj M, K., Chakraborty, S.: PDMS microfluidics: A mini review. Journal of Applied Polymer Science 137(27), 48958 (2020). https://doi.org/10.1002/app.48958

40. Rao, Y.L., Chortos, A., Pfattner, R., *et al.*: Stretchable self-healing polymeric dielectrics cross-linked through metal-ligand coordination. Journal of the American Chemical Society 138(18), 6020–6027 (2016). https://doi.org/10.1021/jacs.6b02428

41. Rashevskii, A.A., Deriabin, K.V., Parshina, E.K., *et al.*: Self-healing redox-active coatings based on ferrocenyl-containing polysiloxanes. Coatings 13(7), 1282 (Jul 2023). https://doi.org/10.3390/coatings13071282

42. Rekondo, A., Martin, R., Ruiz De Luzuriaga, A., *et al.*: Catalyst-free room-temperature self-healing elastomers based on aromatic disulfide metathesis. Materials Horizons 1(2), 237–240 (2014). https://doi.org/10.1039/c3mh00061c

43. Rutenbar, R.A.: Simulated annealing algorithms: An overview. IEEE Circuits and Devices Magazine 5(1), 19–26 (1989). https://doi.org/10.1109/101.17235

44. Saed, M.O., Terentjev, E.M.: Siloxane crosslinks with dynamic bond exchange enable shape programming in liquid-crystalline elastomers. Scientific Reports 10(1), 6609 (2020). https://doi.org/10.1038/s41598-020-63508-4

45. Seethapathy, S., Górecki, T.: Applications of polydimethylsiloxane in analytical chemistry: A review. Analytica Chimica Acta 750, 48–62 (2012). https://doi.org/10.1016/j.aca.2012.05.004

46. Sharma, H., Rana, S., Singh, P., *et al.*: Self-healable fiber-reinforced vitrimer composites: overview and future prospects. RSC Adv. 12(50), 32569–32582 (2022). https://doi.org/10.1039/d2ra05103f

47. Taynton, P., Yu, K., Shoemaker, R.K., *et al.*: Heat- or water-driven malleability in a highly recyclable covalent network polymer. Advanced Materials 26(23), 3938–3942 (2014). https://doi.org/10.1002/adma.201400317

48. Van Der Spoel, D., Lindahl, E., Hess, B., *et al.*: GROMACS: Fast, flexible, and free. J. Comput. Chem. 26, 1701–1718 (2005). https://doi.org/10.1002/jcc.20291

49. Wang, J., Wolf, R.M., Caldwell, J.W., *et al.*: Development and testing of a general Amber force field. Journal of Computational Chemistry 25(9), 1157–1174 (2004). https://doi.org/10.1002/jcc.20035

50. Wang, P., Wang, Z., Liu, L., *et al.*: Self-Healable and Reprocessable Silicon Elastomers Based on Imine–Boroxine Bonds for Flexible Strain Sensor. Molecules 28(16), 6049 (2023). https://doi.org/10.3390/molecules28166049

51. Weng, G., Thanneeru, S., He, J.: Dynamic Coordination of Eu–Iminodiacetate to Control Fluorochromic Response of Polymer Hydrogels to Multistimuli. Advanced Materials 30(11), 1706526 (2018). https://doi.org/10.1002/adma.201706526

52. Wu, J., Cai, L.H., Weitz, D.A.: Tough Self-Healing Elastomers by Molecular Enforced Integration of Covalent and Reversible Networks. Advanced Materials 29(38), 1702616 (2017). https://doi.org/10.1002/adma.201702616

53. Yang, X., Huang, W., Dong, H., *et al.*: Smart Polydimethylsiloxane Materials: Versatility for Electrical and Electronic Devices Applications. Advanced Materials 37(17), 2500472 (2025). https://doi.org/10.1002/adma.202500472

54. Yao, Y., Tai, H., Wang, D., *et al.*: One-pot preparation and applications of self-healing, self-adhesive PAA-PDMS elastomers. Journal of Semiconductors 40(11) (2019). https://doi.org/10.1088/1674-4926/40/11/112602

55. Ye, G., Wang, C., Guo, Y., *et al.*: Vitrimer as a Sustainable Alternative to Traditional Thermoset: Recent Progress and Future Prospective. ACS Polymers Au 5(5), 445–457 (2025). https://doi.org/10.1021/acspolymersau.5c00081

56. Ying, H., Zhang, Y., Cheng, J.: Dynamic urea bond for the design of reversible and self-healing polymers. Nature Communications 5 (2014). https://doi.org/10.1038/ncomms4218

57. Yoshida, S., Ejima, H., Yoshie, N.: Tough Elastomers with Superior Self-Recoverability Induced by Bioinspired Multiphase Design. Advanced Functional Materials 27(30) (2017). https://doi.org/10.1002/adfm.201701670

58. Zhang, K., Sun, J., Song, J., *et al.*: Self-Healing Ti3C2MXene/PDMS Supramolecular Elastomers Based on Small Biomolecules Modification for Wearable Sensors. ACS Applied Materials and Interfaces 12(40), 45306–45314 (2020). https://doi.org/10.1021/acsami.0c13653

59. Zheng, P., McCarthy, T.J.: A surprise from 1954: Siloxane equilibration is a simple, robust, and obvious polymer self-healing mechanism. Journal of the American Chemical Society 134(4), 2024–2027 (2012). https://doi.org/10.1021/ja2113257

60. Zou, Z., Zhu, C., Li, Y., *et al.*: Rehealable, fully recyclable, and malleable electronic skin enabled by dynamic covalent thermoset nanocomposite. Science Advances 4(2) (2018). https://doi.org/10.1126/sciadv.aaq0508

DOI: 10.14529/jsfi260104

Comparative Molecular Dynamics Study of E- and Z-Biliverdin-IXα Binding to Human Serum Albumin

Igor V. Polyakov[1], *Maria G. Khrenova*[1]

Human serum albumin (HSA) is the main transporter of a wide range of endogenous ligands, including linear tetrapyrrolic bile pigments. Despite many experimental and theoretical studies, the detailed binding modes of linear tetrapyrroles in HSA remain not fully understood. Here, we investigate the interaction of 4Z,15E and 4Z,15Z biliverdin-IXα with HSA by classical molecular dynamics and machine learning. Starting from the crystallographic complex of HSA with 4Z,15E bilirubin-IXα, we construct models for both biliverdin isomers and explore their conformational space in the initial binding site. We analyse protein-ligand contacts, conformational flexibility, and the populations of distinct binding poses using clustering of interaction fingerprints. The results reveal both shared and isomer-specific interaction patterns between biliverdin and HSA. Several conserved contacts are maintained in both complexes, while distinct differences in contact occupancies and binding pocket conformations are observed between the E- and Z-isomers. Overall, this study provides a consistent molecular level picture of how biliverdin isomers interact with HSA and demonstrates a practical workflow for analysing flexible protein-ligand complexes by combining molecular dynamics with interaction fingerprint clustering.

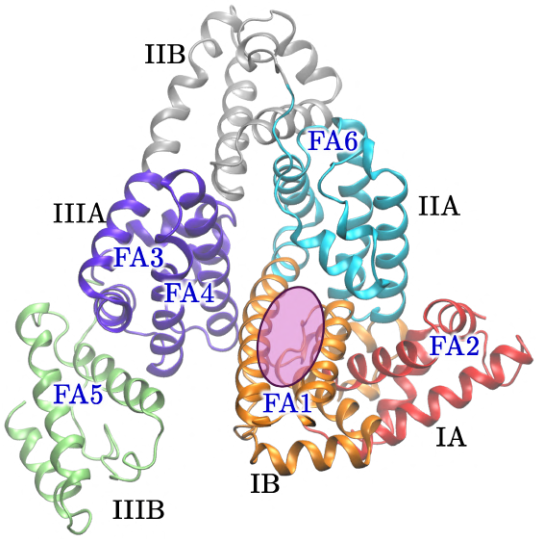
Keywords: albumin, biliverdin, bilirubin, molecular dynamics, clustering.

Introduction

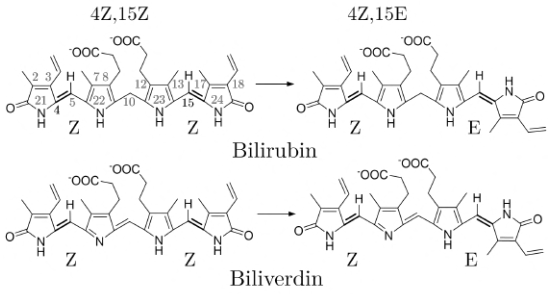
Human serum albumin (HSA) is the main transport protein in blood plasma which carries out an important task of binding and carrying a wide range of compounds, including fatty acids, bile pigments, hormones, and many drugs [2]. Because of its flexible structure, albumin can form stable complexes with ligands of different chemical nature, significantly influencing their bioavailability, distribution, and metabolism in the organism [5, 9]. HSA is a monomeric protein composed of three domains (I–III), and each domain is divided into two subdomains (A and B). Several functionally important binding sites are located inside these domains (Fig. 1), and the most studied ones are Sudlow site I and Sudlow site II in subdomains IIA and IIIA, which are mainly associated with aromatic and hydrophobic ligands [10]. Albumin also contains additional hydrophobic pockets and cavities (Fig. 1) that can bind large and flexible molecules, and the protein conformation can adapt to the chemical nature of the ligand. Bilirubin and biliverdin are endogenous bilins, tetrapyrrolic chromophores formed during heme catabolism. Biliverdin is produced in the oxidative cleavage of heme by heme oxygenase and is then reduced to bilirubin by biliverdin reductase. In blood plasma, bilirubin is almost completely bound to albumin [21], which prevents its precipitation and toxic effects on tissues. The bilirubin molecule is highly conformationally flexible, forms intramolecular hydrogen bonds [15], and can exist in different isomeric forms, mainly as 15Z and 15E isomers (Fig. 1). Isomeric states of bilin chromophores can significantly influence their physicochemical properties and the mode of interaction with proteins, which is important both in biophysical and clinical contexts, for example for phototherapy of hyperbilirubinemia [14].

Experimental studies of bilirubin binding to human serum albumin usually employ spectroscopic methods (such as fluorescence spectroscopy, circular dichroism, and absorption) and kinetic measurements, which provide estimates of binding constants and confirm the presence

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



(a) Different ligand binding sites in albumin, the heme binding site is highlighted



(b) Structures of the main isomers of bilirubin and biliverdin

Figure 1. Protein and ligands

of a high affinity bilirubin-albumin interaction [3, 14]. These approaches give important quantitative parameters, but they have limited ability to resolve the three dimensional architecture of the complex, to describe in detail the dynamic structural changes of the ligand and the protein during binding, and to clarify the specific role of individual amino acid residues in complex stabilization. In this work, we inspected the current state of structural data in the Protein Data Bank (PDB), which contains many experimentally determined three dimensional structures of human serum albumin complexes with different endogenous and exogenous ligands, including fatty acids, steroid hormones, and various pharmacologically active molecules. At the same time, structural data on albumin complexes with bilin chromophores are very limited. At present, the only experimentally resolved atomic structure of a human serum albumin complex with bilirubin is the structure with PDB ID 2VUE, where albumin is crystallised with 4Z,15E bilirubin IXα [28] (the identified binding site is highlighted in Fig. 1).

It was shown that in the albumin complex bilirubin is located in a deep, mainly hydrophobic cavity formed by amino acid residues of subdomain IB. Structural analysis indicates that the bilirubin binding disturbs the Sudlow site I region, but the geometry of the cavity and the interaction pattern differ significantly from typical albumin-drug complexes. In particular, bilirubin adopts an extended and curved conformation inside the cavity, which requires local rearrangement of amino acid side chains and demonstrates pronounced conformational plasticity of the protein. Compared to complexes with fatty acids or small aromatic ligands, bilirubin interacts with albumin not only through nonpolar contacts but also via hydrogen bonds with nearby polar and charged residues, which stabilise its conformation in the bound state. The authors of the crystallographic study clearly show [28] that under crystallisation conditions bilirubin binds in one dominant site. However, they also note that several earlier studies suggest a more complex binding behaviour that may include additional, lower affinity or transient binding modes. The absence of direct structural observation of these states can be related to their low population and high conformational mobility of the ligand; therefore, alternative bilirubin-albumin binding

modes that may exist in solution remain outside the scope of experimental structural data but can be explored by atomistic molecular modelling.

Spectroscopic studies have demonstrated that, like bilirubin, biliverdin can also bind to HSA, thus affecting the intrinsic tryptophan fluorescence and influencing the binding of other ligands by competition for the site [16, 26]. However, no experimentally resolved crystal structures of HSA-biliverdin complexes are currently available in the literature, so the exact localisation and binding pattern of biliverdin remain unknown.

Molecular modelling and molecular dynamics (MD) [27] simulations are now widely and routinely used to study protein-ligand and protein-protein complexes. Nevertheless, existing computational studies of albumin complexes with bilin chromophores are very limited [7, 19, 23]. These works usually rely on relatively short molecular dynamics trajectories and do not include a systematic analysis of the ligand conformational ensemble and its interactions with the protein. Modern microsecond scale atomistic models of HSA-bilirubin/biliverdin complexes, with statistically reliable averaging over multiple trajectories, are not yet available. For this reason, the application of current molecular modelling techniques to investigate the structure and dynamics of the HSA-biliverdin complex is a relevant task. Such an approach can reveal specific features of ligand conformational behaviour, differences between its isomeric forms, and the key amino acid residues and interactions that determine the stability of the protein-ligand complex. An additional important aspect is the development of a workflow that allows for automated analysis of such protein-ligand systems. The article is organized as follows. First, we describe the molecular model setup for the HSA-biliverdin system followed by a description of the molecular dynamics protocol. The analysis and discussion of the obtained results is carried out with both statistical approaches and machine learning algorithms.

1. Protein-Ligand Complex Model Setup

As the starting structure for our model, we used the heavy atom coordinates from the Protein Data Bank entry 2VUE [28], since bilirubin is the ligand most similar in structure to biliverdin and no HSA-biliverdin structures are available in the PDB. Using HSA structures with small ligands as starting points was considered problematic, because docking of extended and flexible tetrapyrrolic ligands is a nontrivial task, especially when taking into account the intrinsic mobility of the albumin structure. In the original crystal structure, a region with missing electron density corresponding to residues 79–88 is present. Although this fragment is not part of the bilirubin binding site and does not directly affect complex formation, it was rebuilt using the tools of the Rosetta molecular modelling package [1]. Since PDB ID 2VUE contains only a small number of crystallographic water molecules, we performed additional solvation with the Dowser++ program [20], which is designed to predict energetically favourable positions of water molecules in protein cavities. A total of 365 water molecules were placed in the structure according to predicted positions.

We used the CHARMM36m [12] force field to describe HSA, since it is widely applied for protein simulations and provides good accuracy for structural and dynamical properties. Parameters for bilirubin were generated with the CHARMM general force field CGenFF [25], which is designed for small organic molecules within the CHARMM framework. Because of the high conformational flexibility of biliverdin and the lack of previously published parameters, we performed an additional validation of the CGenFF parameters. For this purpose, we prepared a separate test system containing a single biliverdin molecule embedded in a box of 200 TIP3 water

molecules. The biliverdin structure was first optimised at the molecular mechanics level and then refined by a hybrid QM/MM approach using the NAMD/TeraChem software and modified interface [18, 24]. The QM part comprised the biliverdin molecule, while water molecules were treated by molecular mechanics. QM part was computed with the wB97X D3/6 31G** [17] level of theory. The good agreement between the optimised geometries indicates that the selected parameters provide an adequate description of biliverdin.

The complete simulation system was assembled with the PSFGEN module of VMD [13]. Protonation of amino acid side chains was carried out according to a neutral pH: histidine residues were treated as neutral, aspartate and glutamate residues as deprotonated, lysine and arginine residues were protonated. Disulfide bridges were manually introduced between the following cysteine residue pairs: 53–62, 75–91, 90–101, 124–169, 168–177, 200–246, 245–253, 265–279, 278–289, 316–361, 360–369, 392–438, 437–448, 461–477, 476–487, 514–559, and 558–567. The protonated HSA-biliverdin complex, including the predicted structural water molecules, was then placed in a cubic simulation water box such that the minimum distance from any protein atom to the box boundary was at least 10 Å. To mimic physiological conditions, sodium and chloride ions were added to reach a NaCl concentration of 0.15 mol/L, and their total number was adjusted to ensure overall charge neutrality of the system.

2. Molecular Dynamics Protocol

Molecular dynamics simulations were performed with the NAMD software version 3.0.2 [22]. Model systems were simulated in the isothermal-isobaric ensemble (NPT, P = 1 atm, T = 300 K) using Langevin dynamics. The SHAKE and SETTLE algorithms were utilized to constraint bonds involving hydrogen atoms, which allowed for the use of 2 fs integration time step. Periodic boundary conditions were enforced: long range electrostatic interactions were treated by the particle mesh Ewald algorithm, and a cutoff of 12 Åwas applied for exact electrostatics.

The equilibration protocol consisted of several consecutive stages. In the first stage, only the solvent was relaxed, while the coordinates of the protein and ligand were kept fully fixed for 5 ns, in order to remove unfavourable contacts of water with the protein and ligand surface. Next, the fixed positional constraints were replaced by harmonic restraints with a force constant of 1 kcal/(mol·Å^2) applied to all heavy atoms of the protein and biliverdin for the next 10 ns of simulation to allow for a "soft" start. In the production stage, all constraints were removed except for harmonic restraints with the same force constant applied to the Cα atoms of five N terminal and five C terminal residues of the protein chain, which were used to prevent global drift of the protein and excessive motion of termini.

To study the dynamics of the HSA-biliverdin complex, several independent molecular dynamics trajectories for different isomeric forms of the ligand were computed. For the 15E isomer, which corresponds to the configuration present in the PDB ID 2VUE crystal structure [28], four independent trajectories were calculated, and four trajectories were also obtained for the native 15Z isomer. Each trajectory was 18 hundred nanoseconds long, thus the total simulation time exceeded 14 μs. The use of multiple independent trajectories provided statistically meaningful averaging and a more complete description of the conformational ensemble of the protein-ligand complex.

3. Results and Discussion

To enable the analysis of molecular dynamics trajectories, we first aligned all frames to the protein backbone, which removes the effect of global protein motions and allows focusing on local dynamics of the complex. However, the total amount of data included hundreds of thousands of frames, so manual inspection of such volume is not possible even for a human expert. Therefore, we applied computational methods to extract useful information from the molecular statistics, including analysis of ligand conformations, mobility of protein residues, and protein-ligand contacts. The root-mean-square deviation (RMSD) values for the protein backbone in the E and Z models were 3.25 ± 0.67 Å and 3.68 ± 0.77 Å, respectively, while for heavy atoms of biliverdin they were 3.49 ± 1.05 Å and 3.44 ± 0.68 Å. The calculated RMSF values per protein residue are shown in Fig. 2, and per atom groups of biliverdin – in Fig. 3. The

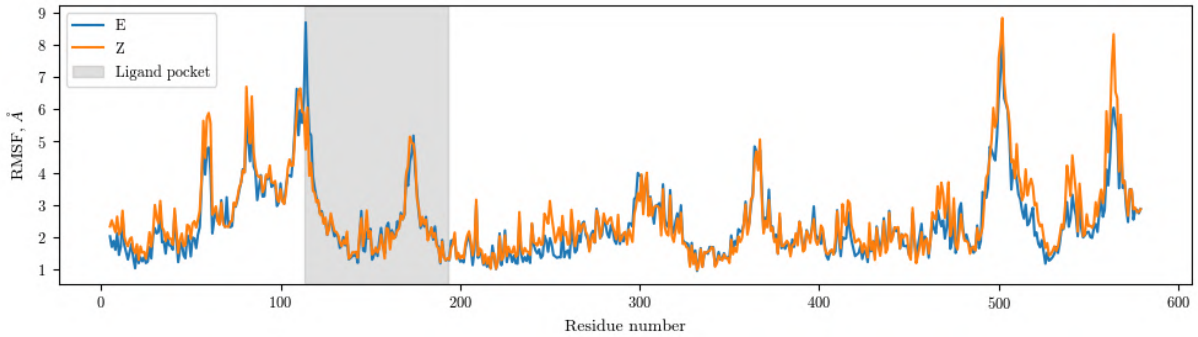


Figure 2. Protein residues RMSF values for the E- and Z-models. The original ligand pocket is highlighted

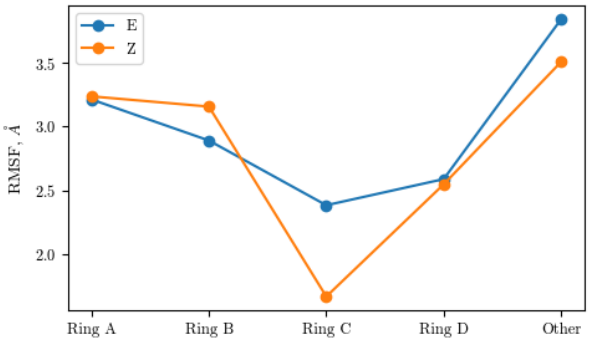
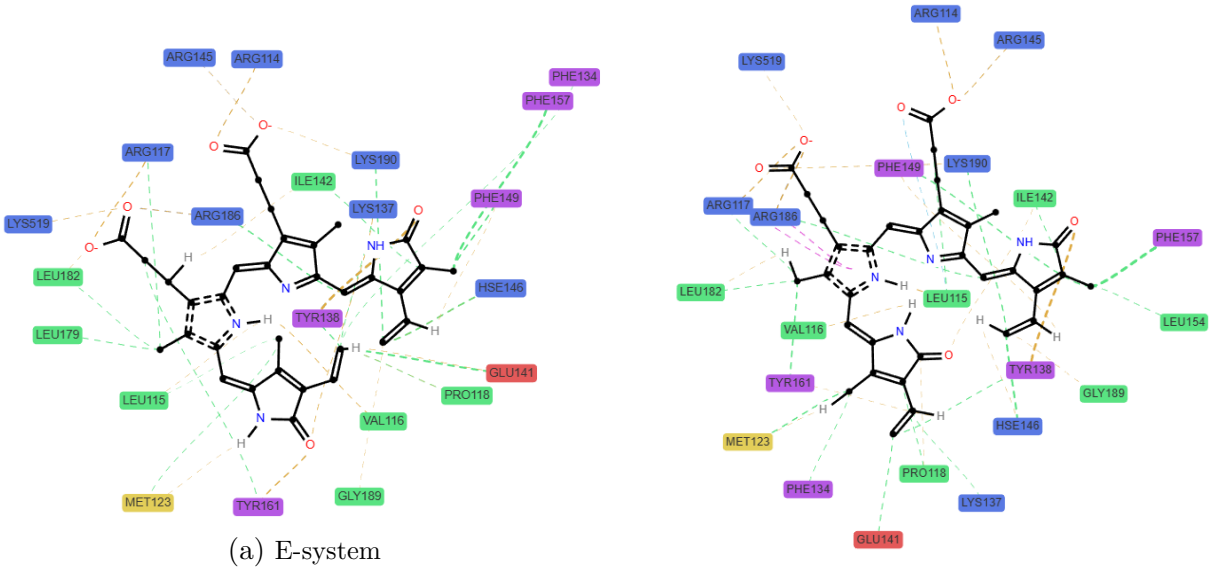


Figure 3. RMSF values for different atom groups of biliverdin

most noticeable difference between the models is observed in the RMSF of residue Arg114 in the chromophore pocket, located on the loop connecting parts of domain I (IA and IB), as well as ring C of biliverdin – however, this information is clearly not enough for comparative evaluation of the models. The most direct and simple method to assess conformational diversity of the studied complexes is QT-clustering [6] (Quality Threshold Clustering) implemented in VMD [13], which can be performed based on RMSD of heavy atoms of the ligand and surrounding amino acid residues. Unfortunately, since biliverdin is a rather flexible ligand and its conformation and position in the protein pocket change significantly along the trajectories, this simple method did not yield satisfactory results – a more precise tool is required to identify patterns of protein-ligand contacts.

Therefore, to detect key interactions of biliverdin with HSA amino acid residues, we used the MDAnalysis [11] package together with the ProLIF module [4]. The algorithm allows for efficient extraction and analysis of various contact descriptors for long MD trajectories. Specifically: Anionic – an interaction where the ligand acts as an anion (negatively charged) and the protein residue as a cation (for example, interaction of the ligand carboxyl group with lysine or arginine); HBDonor – the ligand donates a proton in a hydrogen bond, while the protein accepts it; HBAcceptor – the ligand has an electronegative atom with a lone pair that attracts hydrogen from the protein; Hydrophobic – hydrophobic contacts between nonpolar atoms (carbon, sulfur, halogens), recorded by default at interatomic distances up to 4.5Å; PiCation – interaction between a positively charged group and an aromatic ring; VdWContact – the least specific descriptor that records simple approach of any atoms to the sum of their van der Waals radii with a small tolerance. The detected contacts are shown schematically in Fig. 4, where the colour coding corresponds to the different interaction types described above (VdWContact – golden, PiCation – purple, HBDonor/Acceptor – blue, Anionic – dark blue), the line thickness reflects contact frequency, and contacts observed in less than 20% of trajectory frames were excluded from the scheme. The schemes in Fig. 4 show that several charged residues (Glu141,



(a) E-system

(b) Z-system

Figure 4. Schematic representation of protein-ligand contacts detected in the molecular dynamics trajectories

Arg114/117/145/186, Lys190) and aliphatic residues (Val116, Leu115/182, Pro118, Ile142), as well as aromatic residues of the protein (Tyr138/161, Phe134/149/157), make contacts with both 15Z and 15E biliverdin. These data are quantified in Tab. 1 and Tab. 2, which list the most similar and most different contacts between the systems, considering the average occupancy based on the score. The score is calculated in percent as the average occupancy multiplied either by the distance from the modulus of occupancy difference to unity (Tab. 1) or by the modulus of occupancy difference (Tab. 2). Although this type of analysis reveals both average similar and average different interactions with residues of the protein chromophore pocket, it still does not provide a clear description of conformational diversity in the systems and does not allow for selection of representative structures for further comparative analysis. Therefore, we performed clustering based on the extracted contact data.

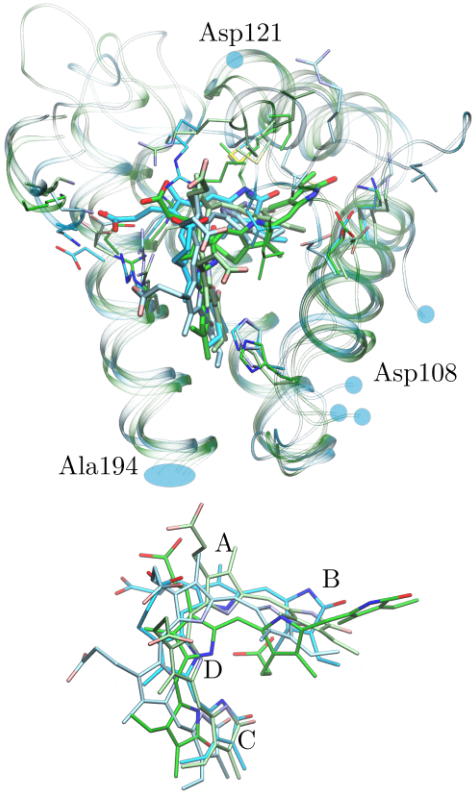
Table 1. Top score similar contacts along the trajectories for albumin model systems with 15E- and 15Z-biliverdin isomers

residue	His146	Ile142	Phe149	Lys190	Tyr138	Arg186	Pro118	Arg117	Phe157	Leu115
E_occ	95	99	91	84	84	77	72	74	66	58
Z_occ	97	94	97	91	81	93	83	95	83	58
score	93	92	88	81	80	72	69	67	61	58

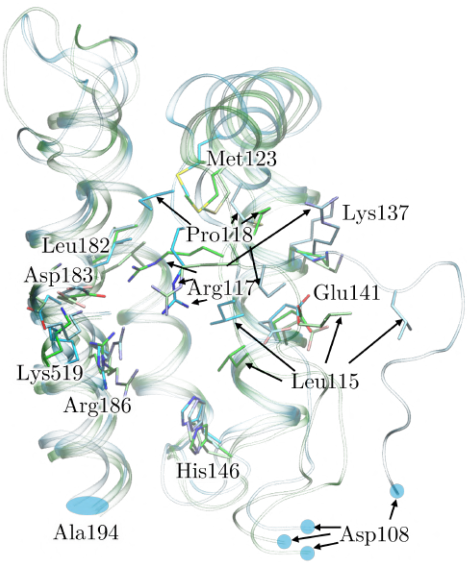
Table 2. Top score different contacts along the trajectories for albumin model systems with 15E- and 15Z-biliverdin isomers

residue	Glu141	Tyr161	Met123	Arg117	Arg186	Tyr161	Phe157	Tyr138
E_occ	91	43	54	71	55	66	66	92
Z_occ	36	78	84	95	77	38	83	79
score	35	21	21	20	15	14	13	11

For each system, we selected a subset of frames by uniform subsampling to limit the total number of analysed frames to 3000. For each selected frame, a binary interaction fingerprint describing ligand-protein contacts was computed using the ProLIF framework, which produces a bit vector encoding the presence or absence of specific interaction types between the ligand and protein residues. Pairwise similarity between fingerprints was quantified with the Tanimoto coefficient, and a distance matrix was constructed as $d_{ij} = 1 - S_{ij}$. Hierarchical agglomerative clustering was carried out using the complete linkage method. The optimal number of clusters was found by maximising the silhouette score over cluster numbers from 2 to 11. For each cluster, the population was calculated as the fraction of frames assigned to that cluster. A representative frame was taken as the structure with maximum total similarity to all other frames in the same cluster; these structures were utilized for visualisation. The process resulted in 4 clusters for the E-system and 5 for the Z-system, with populations of 0.25, 0.37, 0.15, 0.23 and 0.47, 0.12, 0.04, 0.22, 0.15 for the E- and Z-system, respectively. These populations indicate significant conformational diversity for the Z-isomer, where the dominant cluster accounts for less than half of all frames, while for the E-isomer the main cluster is close to one third of frames. Alignment of representative frames from molecular dynamic simulations for both systems is shown in Fig. 5 and Fig. 6, black arrows indicate residues that considerably change spatial location. The performed analysis reveals significant conformational diversity both between the E- and Z-systems and within each of these systems. For example, in the most populated clusters of the E-system (cluster 2 [0.37] and cluster 1 [0.25]), the main differences (modulus of contact occupancy difference over 0.5) are observed for contacts with Arg114 (0.7), Met123 (0.6), Phe134 (0.7), Lys137 (0.6), Tyr138 (1.0), Phe157 (0.9), Tyr161 (0.7), Leu179 (0.6), and Leu182 (0.9). Similar values for the Z-system clusters (1 [0.47] and 4 [0.22]) include: Asn111 (0.7), Pro113 (0.6), Arg114 (1.0), Leu115 (1.0), Val116 (0.5), Arg117 (0.9), Lys137 (0.5), Tyr138 (0.7), Arg145 (1.0), His146 (0.6), Leu182 (0.6), Arg186 (0.8), Lys190 (1.0), Lys519 (0.7), Glu520 (0.8), Ile523 (0.6). When comparing the most populated clusters between E (2 [0.37]) and Z (1 [0.47]) systems, the largest absolute differences in contact frequencies were found for Glu141 (0.7), Leu115 (0.6), Leu179 (0.6), Met123 (0.6), and Tyr161 (0.6). Since we believe in the importance of transparency and reproducibility for any computational publication, including this work, the complete workflow used here, as well as the trajectory data and representative frames, are available in the public Zenodo repository [8].

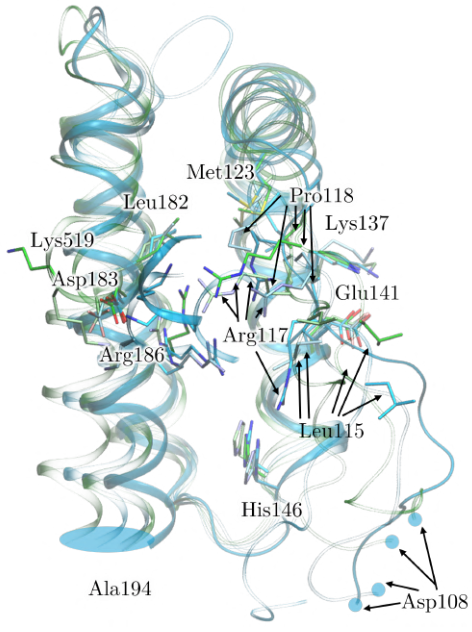


(a) Alignment of entire structures by protein backbone

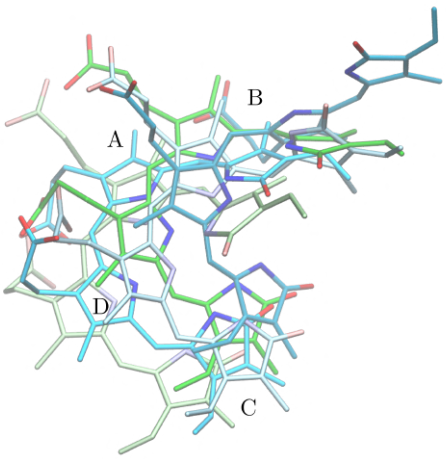


(b) Amino acid residues that interact with biliverdin

Figure 5. E-system representative frames from clustering



(a) Amino acid residues that interact with biliverdin



(b) Biliverdin molecules in the aligned structures

Figure 6. Z-system representative frames from clustering

Conclusion

In this work we calculated a set of long molecular dynamics trajectories of the human serum albumin complex with the biliverdin ligand. A thorough analysis of the trajectories was conducted for both 15E- and 15Z-biliverdin isomers showing that biliverdin remains located in the original albumin bilirubin cavity as defined in the PDB ID 2VUE experimental structure throughout the simulation. The biliverdin ligand demonstrates significant positional fluctuations within the cavity and adopts different conformations for both isomeric forms of biliverdin, 15E and 15Z. Examination of the protein-ligand contact network along the trajectories revealed top similar and top different contacts for both model systems. Hierarchical agglomerative clustering based on the pairwise similarity between fingerprints allowed us to quantify not only differences between the binding modes of the isomers but also substantial variation of contacts within individual trajectories and extract representative frames for structural comparison. Therefore, the molecular modeling methods and trajectory analysis techniques applied here can be used in future studies to model properties such as absorption spectra of protein-ligand systems, since accurate prediction of these properties requires consideration of the full conformational ensemble rather than a single structure. The computational workflow we developed, based on the MDAnalysis and ProLIF libraries, simplifies the analysis of protein-ligand system molecular dynamics trajectories. The workflow is available free of charge in the public Zenodo repository [8].

Acknowledgements

The work was conducted under the state assignment of Lomonosov Moscow State University 121031300176-3. The research was carried out using the equipment of the shared research facilities of HPC computing resources at Lomonosov Moscow State University including Istok computing system (Agreement 075-15-2025-541).

References

1. Alford, R.F., Leaver-Fay, A., Jeliazkov, J.R., *et al.*: The Rosetta all-atom energy function for macromolecular modeling and design. Journal of Chemical Theory and Computation 13(6), 3031–3048 (2017). https://doi.org/10.1021/acs.jctc.7b00125

2. Ashraf, S., Qaiser, H., Tariq, S., *et al.*: Unraveling the versatility of human serum albumin – a comprehensive review of its biological significance and therapeutic potential. Current Research in Structural Biology 6, 100114 (2023). https://doi.org/10.1016/j.crstbi.2023.100114

3. Berde, C., Hudson, B., Simoni, R., Sklar, L.: Human serum albumin. Spectroscopic studies of binding and proximity relationships for fatty acids and bilirubin. Journal of Biological Chemistry 254(2), 391–400 (1979). https://doi.org/10.1016/S0021-9258(17)37930-9

4. Bouysset, C., Fiorucci, S.: ProLIF: a library to encode molecular interactions as fingerprints. Journal of Cheminformatics 13(1) (Sep 2021). https://doi.org/10.1186/

s13321-021-00548-6

5. Catalano, C., Lucier, K.W., To, D., *et al.*: The CryoEM structure of human serum albumin in complex with ligands. Journal of Structural Biology 216(3), 108105 (2024). https://doi.org/10.1016/j.jsb.2024.108105

6. Daura, X., Conchillo-Solé, O.: On quality thresholds for the clustering of molecular structures. Journal of Chemical Information and Modeling 62(22), 5738–5745 (2022). https://doi.org/10.1021/acs.jcim.2c01079

7. Díaz, N., Suárez, D., Sordo, T.L., Merz, K.M.: Molecular Dynamics Study of the IIA Binding Site in Human Serum Albumin: Influence of the Protonation State of Lys195 and Lys199. Journal of Medicinal Chemistry 44(2), 250–260 (2001). https://doi.org/10.1021/jm000340v

8. European Organization for Nuclear Research, OpenAIRE: Zenodo (2013). https://doi.org/10.5281/zenodo.18630027

9. Fasano, M., Curry, S., Terreno, E., *et al.*: The extraordinary ligand binding properties of human serum albumin. IUBMB Life 57(12), 787–796 (2005). https://doi.org/10.1080/15216540500404093

10. Ghuman, J., Zunszain, P.A., Petitpas, I., *et al.*: Structural basis of the drug-binding specificity of human serum albumin. Journal of Molecular Biology 353(1), 38–52 (2005). https://doi.org/10.1016/j.jmb.2005.07.075

11. Gowers, R., Linke, M., Barnoud, J., *et al.*: MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. In: Proceedings of the 15th Python in Science Conference. p. 98–105. SciPy (2016). https://doi.org/10.25080/majora-629e541a-00e

12. Huang, J., Rauscher, S., Nawrocki, G., *et al.*: CHARMM36m: an improved force field for folded and intrinsically disordered proteins. Nature Methods 14(1), 71–73 (Jan 2017). https://doi.org/10.1038/nmeth.4067

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

14. Jasprova, J., Dal Ben, M., Vianello, E., *et al.*: The biological effects of bilirubin photoisomers. PLOS ONE 11(2), 1–16 (02 2016). https://doi.org/10.1371/journal.pone.0148126

15. Jóźwiak, K., Ogrin, P., Urbic, T., Filarowski, A.: Molecular dynamics and density functional theory studies of conformational stability of bilirubin and biliverdin. Journal of Molecular Liquids 391, 123287 (2023). https://doi.org/10.1016/j.molliq.2023.123287

16. Lemli, B., Lomozová, Z., Huber, T., *et al.*: Effects of Heme Site (FA1) Ligands Bilirubin, Biliverdin, Hemin, and Methyl Orange on the Albumin Binding of Site I Marker Warfarin: Complex Allosteric Interactions. International Journal of Molecular Sciences 23(22), 14007 (Nov 2022). https://doi.org/10.3390/ijms232214007

17. Lin, Y.S., Li, G.D., Mao, S.P., Chai, J.D.: Long-range corrected hybrid density functionals with improved dispersion corrections. Journal of Chemical Theory and Computation 9(1), 263–272 (2013). https://doi.org/10.1021/ct300715s

18. 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

19. Moosavi-Movahedi, Z., Bahrami, H., Zahedi, M., *et al.*: A theoretical elucidation of bilirubin interaction with HSA's lysines: First electrostatic binding site in IIA subdomain. Biophysical Chemistry 125(2), 375–387 (2007). https://doi.org/10.1016/j.bpc.2006.09.013

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

21. Petersen, C.E., Ha, C.E., Harohalli, K., *et al.*: A dynamic model for bilirubin binding to human serum albumin *. Journal of Biological Chemistry 275(28), 20985–20995 (Jul 2000). https://doi.org/10.1074/jbc.M001038200

22. Phillips, J.C., Hardy, D.J., Maia, J.D.C., *et al.*: Scalable molecular dynamics on CPU and GPU architectures with NAMD. The Journal of Chemical Physics 153(4), 044130 (07 2020). https://doi.org/10.1063/5.0014475

23. Radibratovic, M., Minic, S., Stanic-Vucinic, D., *et al.*: Stabilization of human serum albumin by the binding of phycocyanobilin, a bioactive chromophore of blue-green alga spirulina: Molecular dynamics and experimental study. PLOS ONE 11(12), 1–18 (12 2016). https://doi.org/10.1371/journal.pone.0167973

24. Seritan, S., Bannwarth, C., Fales, B.S., *et al.*: TeraChem: Accelerating electronic structure and ab initio molecular dynamics with graphical processing units. The Journal of Chemical Physics 152(22), 224110 (06 2020). https://doi.org/10.1063/5.0007615

25. Vanommeslaeghe, K., MacKerell, A.D.J.: Automation of the CHARMM General Force Field (CGenFF) I: Bond Perception and Atom Typing. Journal of Chemical Information and Modeling 52(12), 3144–3154 (2012). https://doi.org/10.1021/ci300363c

26. Wei, Y.l., Li, J.q., Dong, C., *et al.*: Investigation of the association behaviors between biliverdin and bovine serum albumin by fluorescence spectroscopy. Talanta 70(2), 377–382 (Sep 2006). https://doi.org/10.1016/j.talanta.2006.02.052

27. Wu, N., Zhang, R., Peng, X., *et al.*: Elucidation of protein–ligand interactions by multiple trajectory analysis methods. Phys. Chem. Chem. Phys. 26, 6903–6915 (2024). https://doi.org/10.1039/D3CP03492E

28. Zunszain, P.A., Ghuman, J., McDonagh, A.F., Curry, S.: Crystallographic Analysis of Human Serum Albumin Complexed with 4Z,15E-Bilirubin-IX. Journal of Molecular Biology 381(2), 394–406 (2008). https://doi.org/10.1016/j.jmb.2008.06.016

DOI: 10.14529/jsfi260105

MPI+OpenMP Implementation of Resolution-of-the-Identity Hartree—Fock Method Exploiting Permutational Symmetry of Three-Center Electron Repulsion Integrals

Iurii V. Kashpurovich[1,2], *Alexander V. Oleynichenko*[3,2], *Vladimir V. Stegailov*[1,2,4]

We report a high-performance implementation of the resolution-of-the-identity Hartree—Fock method that fully exploits the permutational symmetry of three-center electron repulsion integrals (ERIs). The present implementation adopts a hybrid MPI+OpenMP parallelization strategy. Two different algorithmic approaches (with and without the preliminary transformation of ERIs) are analyzed and compared. A custom data layout introduced previously is employed. Designed to efficiently leverage the permutational symmetry of ERIs, it minimizes not only inter-node communication but also local memory traffic. Other extensive low-level and algorithmic optimizations are proposed and discussed. Reasonable parallel scaling is demonstrated by performance benchmarks on a chlorophyll dimer $(C_{55}H_{72}O_5N_4Mg)_2$ in an aqueous environment of 48 molecules (322 atoms overall, 3700 and 11896 functions in main and auxiliary basis sets, respectively). Peak speedups of 84× and 71× on 128 threads are achieved for the ERI calculation and the exchange matrix construction, respectively, within the algorithm involving the preliminary transformation.

Keywords: restricted Hartree—Fock method, resolution-of-the-identity, density fitting, three-center electron repulsion integrals, MPI, OpenMP.

Introduction

The Hartree—Fock (HF) method, also known as the self-consistent field (SCF) method [8], is a basic approach to solving the many-body electronic structure problem in modern quantum chemistry. It serves both as a starting point for more advanced electron correlation methods and as a conceptual and software base for density functional theory (DFT). The computational complexity of a straightforward SCF algorithm scales as $O(N^4)$ with a system of N atoms, and numerous computational techniques and their program implementations with reduced scaling were proposed in the last decades to overcome this SCF deficiency. One of the most prominent modern approaches is the resolution-of-the-identity (RI) approximation, also known as the density fitting technique (DF) [4, 5, 9, 10, 12, 15, 16, 20, 28, 30, 31, 33, 34, 36, 38, 39, 42, 44, 47, 49]. Working expressions of the RI-HF theory employ three- and two-center electron repulsion integrals (ERIs) defined as

$$(\mu\nu|B) = \int \frac{\chi_\mu(\mathbf{r}_1)\chi_\nu(\mathbf{r}_1)\phi_B(\mathbf{r}_2)}{|\mathbf{r}_1 - \mathbf{r}_2|} d\mathbf{r}_1 d\mathbf{r}_2, \tag{1}$$

$$V_{BC} = \int \frac{\phi_B(\mathbf{r}_1)\phi_C(\mathbf{r}_2)}{|\mathbf{r}_1 - \mathbf{r}_2|} d\mathbf{r}_1 d\mathbf{r}_2, \tag{2}$$

where the basis functions χ_μ, $\mu = 1, \ldots, N_{\text{AO}}$ and the auxiliary basis functions ϕ_B, $B = 1, \ldots, N_{\text{aux}}$ are typically represented by atom-centered Gaussian-type orbitals (atomic or-

[1]Joint Institute for High Temperatures of Russian Academy of Sciences, Moscow, Russian Federation
[2]Moscow Institute of Physics and Technology, Moscow, Russian Federation
[3]Petersburg Nuclear Physics Institute named by B.P. Konstantinov of NRC "Kurchatov Institute", Gatchina, Russian Federation
[4]HSE University, Moscow, Russian Federation

bitals, AOs). Array (1) is also referred to as RI tensor. In addition to reducing the computational scaling of the SCF method to $O(N^\alpha)$ ($\alpha \in [3.0, 4.0)$), the RI approximation allows one to store all ERIs (including $(\mu\nu|B)$) in local or distributed random access memory (RAM) for most practical problems of interest [15, 39, 45] as well as aggregate high-performance tensor contraction engines (such as BLAS [3] libraries) and hardware accelerators (such as GPUs [4, 5, 42, 53]).

The first distributed RI-SCF implementation [12] stored all ERIs (1) to disk using a compression. Then the integrals were fully unpacked at each SCF iteration and computations with dense tensors were performed. Up to now, permutational symmetry of $(\mu\nu|B)$ with respect to indices μ and ν was accounted for only to efficienly store the RI tensor in memory. In the recently reported implementation presented by our group [25] it was also exploited for the Fock matrix assembly stage, although at the cost of additional technical complications. There are other challenges as well. For example, for systems consisting of hundreds of atoms conventional HF calculations (not employing the RI approximation) can be even faster than those using RI-HF [22]. It is because of difficulties to deal with the sparsity of intermediate arrays within RI-SCF procedure, while the scaling of the straightforward "conventional" HF approach can be reduced to $O(N^2)$, if one employs a proper integral screening [7, 14, 18]. Other issues are the lack of GPU-accelerated RI-HF implementations [4, 5, 42, 53] and of massively-parallel implementations of the two-component quasirelativistic RI-SCF method.

Our goal is to gradually fill the specified gaps within the new BUFO program package for quantum chemistry simulations. This paper reflects the current status of our progress and summarizes several previous developments [25–27] aimed at the efficient parallel implementation of the RI-HF method. Careful benchmarking of parallel algorithms for electronic structure calculations is instrumental in achieving the best performance of supercomputer systems [41]. Here we present two different MPI+OpenMP algorithms of the RI-HF method and conduct performance tests using two HPC systems: a 2-socket server with shared memory and a 32-node supercomputer with distributed memory. The key feature of the presented implementation is storing three-center ERIs (1) in the distributed RAM using the permutationally-adapted memory layout in order to minimize not only communications between processes, but also local memory movement [25].

The paper is organized as follows. Section 1 outlines the theoretical background of the spin-restricted RI-HF method. Different approaches for its implementation for distributed memory architectures and actual algorithms are described in Section 2. In Section 3 we highlight the parallel computers employed to test the scalability of the developed implementations and also describe molecular system for benchmark calculations in details. Results are presented in Section 4. Conclusion provides some remarks and specifies future directions of work.

1. Theoretical Background

Throughout this paper, we adopt the following index labeling conventions:

1. i: occupied molecular orbitals, with range N_{occ};
2. μ, ν, ρ, σ: atomic basis functions (AOs), with range N_{AO};
3. A, B, P: auxiliary RI basis functions, with range N_{aux}.

For the sake of simplicity, restricted Hartree—Fock (RHF) theory is considered further, though our approaches can be extended to unrestricted and two-component quasirelativistic version of the method as well. The primary goal of the SCF algorithm is to determine expansion

coefficients $C_{i\mu}$ defining spatial molecular orbitals (MOs) $\varphi_i(\mathbf{r})$,

$$\varphi_i(\mathbf{r}) = \sum_{\mu=1}^{N_{\text{AO}}} C_{i\mu}\chi_\mu(\mathbf{r}). \tag{3}$$

The coefficients $C_{i\mu}$ are needed to construct density matrix $D_{\mu\nu}$

$$D_{\mu\nu} = 2\sum_{i=1}^{N_{\text{occ}}} C_{i\mu}C_{i\nu}. \tag{4}$$

These coefficients are obtained by diagonalizing the Fock matrix $F_{\mu\nu}$

$$F_{\mu\nu} = h_{\mu\nu} + J_{\mu\nu} - K_{\mu\nu}, \tag{5}$$

where $h_{\mu\nu}$ stands for the core Hamiltonian. The Coulomb and exchange matrices $J_{\mu\nu}$ and $K_{\mu\nu}$ within the RI approximation are constructed according to the expressions

$$J_{\mu\nu} = \sum_{\rho,\sigma}^{N_{\text{AO}}}\sum_{B,C}^{N_{\text{aux}}} D_{\rho\sigma}(\mu\nu|B)V^{-1}_{BC}(C|\rho\sigma), \tag{6}$$

$$K_{\mu\nu} = \frac{1}{2}\sum_{i}^{N_{\text{occ}}}\sum_{\rho,\sigma}^{N_{\text{AO}}}\sum_{B,C}^{N_{\text{aux}}} C_{i\sigma}C_{i\rho}(\mu\sigma|B)V^{-1}_{BC}(C|\rho\nu). \tag{7}$$

The V^{-1}_{BC} matrix is a symmetric positive definite matrix, so it can be factorized using the Cholesky decomposition with factor L_{BP}. The latter can be used to transform three-center ERIs

$$\widetilde{(\mu\nu|P)} = \sum_{B}^{N_{\text{aux}}} (\mu\nu|B)L_{BP} \tag{8}$$

in order to get simplified expressions for Coulomb and exchange matrices

$$J_{\mu\nu} = \sum_{\rho,\sigma}^{N_{\text{AO}}}\sum_{P}^{N_{\text{aux}}} D_{\rho\sigma}\widetilde{(\mu\nu|P)}\widetilde{(P|\rho\sigma)}, \tag{9}$$

$$K_{\mu\nu} = \frac{1}{2}\sum_{i}^{N_{\text{occ}}}\sum_{\rho,\sigma}^{N_{\text{AO}}}\sum_{P}^{N_{\text{aux}}} C_{i\sigma}C_{i\rho}\widetilde{(\mu\sigma|P)}\widetilde{(P|\rho\nu)}. \tag{10}$$

2. Algorithm Design

2.1. High-Level Description

In this work, two different implementations of the RI-RHF method are presented. The first (RI-JK) refers to Eqs. (6) and (7) further, while the second one (RI-TJK) to Eqs. (9) and (10). Corresponding high-level view to both algorithms is sketched in Algorithm 1, so they are composed of stages:

- (line 1) scheduling of integration;
- (line 2) calculating ERIs: note that we adopt the row-major ordering convention in this work, so the last index is the fastest-varying; two algorithms initially store three-center ERIs using different layouts; the upper index p identifies the dimension being partitioned between processes;

- (line 3) performing the transformation (8) and the global transposition of $(\widetilde{\mu\nu|B})$ only in case of the RI-TJK (i.e., $(\widetilde{B|\mu\nu}) \equiv (B|\mu\nu)$ for the RI-JK) and calculating the inverse V_{BC}^{-1} and its Cholesky decomposition L_{BC} for both algorithms;
- (line 4) SCF procedure.

Each step is discussed in detail in the following subsections. However, the focus is on the three-center ERIs, as they constitute the core of the RI approximation.

2.2. Scheduling ERIs Calculation

First of all, let us note that modern highly efficient libraries for the calculation of integrals (like those used in this work `libcint` [43] and `libgrpp` [37]) operate on shells of basis functions, where the shell groups functions sharing the same exponent and angular momentum. We will further assume that N_{AO} basis functions are grouped into $N_{\text{AO}}^{\text{sh}}$ shells, while N_{aux} functions into $N_{\text{aux}}^{\text{sh}}$.

Scheduling is needed to overcome challenges of the integration:

- computational workload balancing;
- fixed pattern of storing integrals.

The better the first one is resolved, the better scalability would be achieved. Then it seems that all $(N_{\text{AO}}^{\text{sh}})^2 N_{\text{aux}}^{\text{sh}}$ triplets of shells should be distributed dynamically between processes. Although this approach can easily lead to $O(N_{\text{AO}}^2 N_{\text{aux}})$ communication volume. However, dealing with the second challenge allows to reduce this communication, while imposing restrictions on the granularity of the distribution of shell triplets. Now, these triplets should be grouped in order to fill the whole block of the RI tensor. These blocks are composed of subblocks depicted in Fig. 1 as regions bounded by red lines (note that integrals are stored row-wise). So, depending on the algorithm, the blocks can be:

- Groups of rows of size N_{aux} in case of the RI-JK (see Fig. 1a). Note that red lines in Fig. 1a also show removed rows (they are removed in order to store permutationally-unique integrals only).
- Groups of triangles of the matrix $N_{\text{AO}} \times N_{\text{AO}}$ in case of the RI-TJK (see Fig. 1b). The calculation in this case may start and proceed from either upper or lower triangle. Each group is formed appropriately.

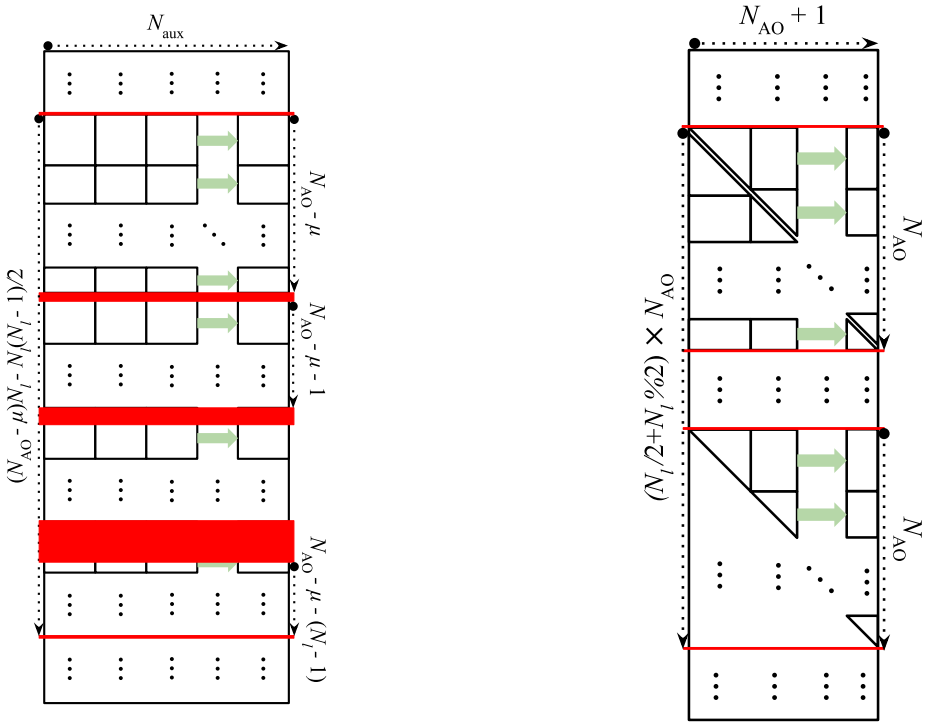
Thus, either $N_{\text{AO}}^{\text{sh}}$ (or even $(N_{\text{AO}}^{\text{sh}})^2$) or $N_{\text{aux}}^{\text{sh}}$ basis shells can be distributed initially in a static manner like in this work: a shell of $N_{\text{AO}}^{\text{sh}}$ or $N_{\text{aux}}^{\text{sh}}$ composed of the largest number of functions is identified and pushed to the process task queue with the smallest current number of ERIs expected to be calculated. This distribution provides predictable data placement of integrals in memory and a compromise between task granularity and subsequent communication volume. As a result of scheduling stage, shells are specifically ordered in the global workspace, so the

Algorithm 1 RI-RHF method

Input: Number of main basis shells: $N_{\text{AO}}^{\text{sh}}$; number of auxiliary basis shells: $N_{\text{aux}}^{\text{sh}}$

Output: Coefficients of MOs (Eq. (3)): $C_{i\mu}$

1: `scheduler`, $\texttt{ord}_{\text{AO}}$, $\texttt{ord}_{\text{aux}}$ ← ScheduleGlobalIntegrationShellOrder($N_{\text{AO}}^{\text{sh}}$, $N_{\text{aux}}^{\text{sh}}$)
2: $h_{\mu\nu}$, $(\mu^{(p)}\nu|B)$ or $(B^{(p)}|\mu\nu)$, V_{BC} ← CalculateERIs(`scheduler`, $\texttt{ord}_{\text{AO}}$, $\texttt{ord}_{\text{aux}}$)
3: $(\widetilde{B^{(p)}|\mu\nu})$, V_{BC}^{-1}, L_{BC} ← TransformERIs($(\mu^{(p)}\nu|B)$ or $(B^{(p)}|\mu\nu)$, V_{BC})
4: $C_{i\mu}$ ← SCFProcedure($h_{\mu\nu}$, $(\mu\nu|\widetilde{B^{(p)}})$, V_{BC}^{-1}, L_{BC})



(a) Row-majored $(\mu\nu|B)$. N_l is the size of the shell μ_l; μ is the number of the first function from μ_l in the basis

(b) Permutationally-adapted $(B|\mu\nu)$. N_l is the size of the shell B_l

Figure 1. Memory layout of the RI tensor (1) and scheme of ERI calculation in batches

corresponding $\texttt{ord}_{\text{AO}}$ and $\texttt{ord}_{\text{aux}}$ arrays of them are returned to perform integration (Alg. 1, line 1) within `CalculateERIs` function (Alg. 1, line 2).

Ideally, the dynamic `scheduler` should also be returned and then passed to `CalculateERIs` function to balance computational workload during calculation (especially if prescreening techniques are applied [21, 48]). It is not so easy to implement such a `scheduler` efficiently. Different approaches are tried within conventional HF [1, 11, 17, 23, 35, 54] and RI-HF calculations [6, 12, 42]. The best algorithms should combine the features of static and dynamic load-balancing. Nevertheless, our research does not focus on dynamic scheduling only proposing and discussing suitable granularity and initial distribution.

2.3. Calculation of Three-Center ERIs

After distributing shells $N_{\text{AO}}^{\text{sh}}$ or $N_{\text{aux}}^{\text{sh}}$ between processes, each process p has a number of shells either $\mu_l \in \{\mu_{l_0}^{(p)} \dots \mu_{l_n}^{(p)}\}$ in case of RI-JK or $B_l \in \{B_{l_0}^{(p)} \dots B_{l_n}^{(p)}\}$ in case of RI-TJK (in both cases $n = n(p)$, so it depends on initial static distribution of shells). These sets were extracted from the $\texttt{ord}_{\text{AO}}$ and $\texttt{ord}_{\text{aux}}$ arrays. Each shell represents the specific task of performing integration and filling the corresponding subblock of the RI tensor (see Fig. 1: the arrows indicate the progress of filling the RI tensor by batches, shown as rectangles and triangles). Both algorithms have a similar structure that is described in Algorithm 2 for the first of them. For each shell μ_l (line 1) or B_l of size N_l either $(N_{\text{AO}} - \mu)N_l - (N_l - 1)N_l/2$ rows[5] (line 2) or $N_l/2 + N_l\%2$ triangles[6] of the RI tensor should be filled with integrals, respectively. Then a block of `size` rows or triangles

[5] μ is the number of the first function from the shell μ_l in basis

[6] "%" represents the remainder of the division

Algorithm 2 Calculation of three-center ERIs (1)

Input: initial main basis shells distribuiton $\{\mu_{l_0}^{(p)} \dots \mu_{l_n}^{(p)}\}$ on given process p; number of main basis shells: $N_{\text{AO}}^{\text{sh}}$; number of auxiliary basis shells: $N_{\text{aux}}^{\text{sh}}$
Output: $(\mu^{(p)}\nu|B)$, $\mu \leq \nu$

1: **for** $\mu_l \in \{\mu_{l_0}^{(p)} \dots \mu_{l_n}^{(p)}\}$ **do**
2: `size` $\leftarrow$ GetBlockSize(μ_l)
3: **for** $\nu_l \in \{\mu_l \dots N_{\text{AO}}^{\text{sh}}\}, B_l \in \{0 \dots N_{\text{aux}}^{\text{sh}}\}$ **do** in parallel (OMP)
4: `batch` $\leftarrow$ CalculateBatchERIs(μ_l, ν_l, B_l)
5: $(\mu\nu|B^{(p)}) \leftarrow$ PlaceBatchERIsToRITensor($(\mu\nu|B^{(p)})$, `batch`, `offset`, μ_l, ν_l, B_l)
6: **end for**
7: `offset` $\leftarrow$ UpdateOffset(`size`)
8: **end for**

corresponding to shell μ_l or B_l is filled in the `for` loop (lines 3–6) parallelized using dynamic OpenMP policy in current implementations. Finally, after the block is ended, the `offset` is updated (line 7) to continue filling from the correct new position. For the first algorithm it is just increased by `size`, while for the second one lower or upper triangle should be identified as well in order to proceed.

Precise compromise between task granularity and subsequent communication volume can be achieved through grouping shells from external `for` loop (line 1) as well as tiling of the nested `for` loops (line 3), e.g., in a way that task would be characterized not only by shell number μ_l or B_l, but also by groups $\{\mu_{l_{g_1}}^{(p)} \dots \mu_{l_{g_n}}^{(p)}\}$ or $\{B_{l_{g_1}}^{(p)} \dots B_{l_{g_n}}^{(p)}\}$ as well as by tiles $\{\nu_{l_{t_1}}^{(p)} \dots \nu_{l_{t_n}}^{(p)}\}$ or $\{\mu_{l_{t_1}}^{(p)} \dots \mu_{l_{t_n}}^{(p)}\}$, respectively. Work on each task or even on several of them can be parallelized using OpenMP more efficiently then.

As a result of the algorithms part of the three-center ERIs is stored in process p memory. This part is $(\mu^{(p)}\nu|B)$ for RI-TJK and $(B^{(p)}|\mu\nu)$ for RI-JK. However, three-center ERIs are not distributed evenly between computational nodes by default. This is because the integration was scheduled by shells. Ideally, additional balancing of the three-center ERIs distribution during or after integration is needed for maximum efficiency of program execution, although it will be implemented in the future.

Finally, note that the proposed algorithms are easily adaptable to dynamic workflow: shell μ_l or B_l can be pushed to or extracted from the task queue (in fact, noted as $\{\mu_{l_0}^{(p)} \dots \mu_{l_n}^{(p)}\}$ or $\{B_{l_0}^{(p)} \dots B_{l_n}^{(p)}\}$ above) of the process p. This is also true for tasks created in a more complex manner.

2.4. Transformation and Transposition of ERIs

Before starting the SCF procedure, three-center integrals are stored either using permutationally-unique row-major or permutationally-adapted format (see Fig. 1). The purpose of the transformation (8) is to simplify subsequent SCF computations. In turn, this transformation requires $N_{\text{AO}}^2 N_{\text{aux}}^2$ arithmetic operations and subsequent distributed transposition of the $N_{\text{AO}}(N_{\text{AO}}+1)/2 \times N_{\text{aux}}$ matrix. In contrast, the alternative algorithm performs distributed transposition of the $N_{\text{occ}}N_{\text{AO}} \times N_{\text{aux}}$ matrix and $N_{\text{occ}}N_{\text{AO}}N_{\text{aux}}^2$ arithmetic operations subsequently in each iteration of the SCF during the construction of the exchange matrix. Thus, the

Algorithm 3 Transformation of three-center ERIs

Input: non-transformed row-majored three-center integrals $(\mu\nu|B^{(p)})$ (1); Cholesky factor L_{BP} of V_{BC}^{-1} (see Eq. (2)); initial distribution of rows $\overline{(\mu\otimes\nu)^{(p)}\dots(\mu\otimes\nu)^{(p+1)}}$ on process p; size of the main basis N_{AO}; size of the auxiliary basis N_{aux}

Output: $\widetilde{(P^{(p)}|\mu\nu)}$, $\mu \le \nu$

1: **for** $(\mu\otimes\nu) = \overline{(\mu\otimes\nu)^{(p)}\dots(\mu\otimes\nu)^{(p+1)}}$ **do** in parallel (OMP)
2: $\quad\widetilde{(\mu\nu|P^{(p)})} \leftarrow$ `TRMM`$[(\mu\nu|B^{(p)}), L_{BP}]$
3: **end for**
4: $\widetilde{(P^{(p)}|\mu\nu)} \leftarrow \widetilde{(\mu\nu|P^{(p)})}$
5: **for** $P = \overline{P^{(p)}/2\dots P^{(p+1)}/2}$ **do** in parallel (OMP)
6: $\quad$reshape $2\otimes N_{\mathrm{AO}}(N_{\mathrm{AO}}+1)/2$ of $\widetilde{(2P|\mu\nu)}$ into $N_{\mathrm{AO}}\otimes(N_{\mathrm{AO}}+1)$
7: **end for**

transformation (8) is justified if the transposition of $N_{\mathrm{AO}}(N_{\mathrm{AO}}+1)/2 \times N_{\mathrm{aux}}$ matrix is not the bottleneck and the number of SCF iterations is large enough.

The transformation and the transposition of three-center ERIs are described in Algorithm 3. The integrals were calculated and stored in rows of size N_{aux} in the previous stage (Fig. 1a), so now each row can be transformed according to Eq. (8). The corresponding iterations are shown in lines 1–3. They are parallelized with OpenMP. The matrix product is evaluated using the BLAS `TRMM` routine (line 2). Alternatively to parallelizing `for` loop, each `TRMM` operation can be parallelized as well. Then the whole RI tensor is transposed (line 4), so the column-majored RI tensor is obtained. To be stored using permutationally-adapted layout, finally, in lines 5–7 transformed integrals are reshaped into $N_{\mathrm{AO}}\times(N_{\mathrm{AO}}+1)$ matrices composed of a pair of triangles (Fig. 1b). If $N_{\mathrm{aux}}^{(p)}$ is odd, one of the last block triangles is simply filled with zeros.

Note that the algorithm can be greatly optimized by communication computation overlap: `for` loop in lines 1–3 can be divided into rounds, and at the end of each round transformed integrals can be put to the memory of other processes as well as got by given process. Thus, data transfer within the transposition (line 4) would also be divided into rounds being included in `for` loop. The optimal partition of this loop is such that one can perform a computation throughout the time of the message passing. Communication and computation would also be significantly reduced if three-center ERIs sparsity was taken into account. Finally, note that the overall message passing during the whole program execution might not be minimal if the transposition after the transformation (8) is performed, because it affects the amount of memory to be communicated within the SCF procedure. So, some checks need to be done in the future.

2.5. SCF Procedure

The SCF procedure is structured as always: calculation of Coulomb and exchange matrices (Eqs. (6) or (9) and (7) or (10)), construction of the Fock matrix (5) and its diagonalization to obtain molecular orbital expansion coefficients (3). Replicated data approaches as well as distributed ones are possible. For now, our implementations follow the replicated baseline for all objects except the RI tensor (1), which is evenly (or almost evenly) distributed between the computational nodes. The diagonalization of the Fock matrix is performed locally using OpenMP (although it can be implemented with MPI using libraries such as ELPA [32]). Note that iterative diagonalization methods such as Davidson or Lanczos [52] (and refs. therein) can be employed

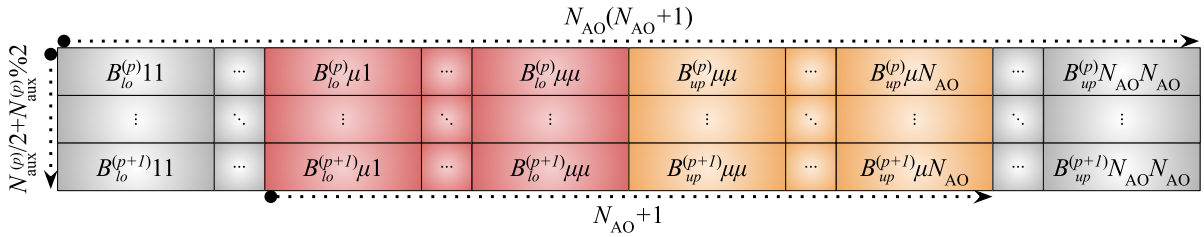


Figure 2. RI tensor representation for contractions within RI-J or RI-TJ (Alg. 4)

Algorithm 4 Coulomb matrix construction according to Eq. (6)

Input: permutationally-adapted three-center integrals $(B^{(p)}|\mu\nu)$ (1) (or $(\widetilde{B^{(p)}}|\mu\nu)$ (8)); density matrix $D_{\mu\nu}$ (4); inverse V_{BC}^{-1} of two-center integrals V_{BC} (2); size of the main basis N_{AO}; size of the auxiliary basis N_{aux}

Output: $J_{\mu\nu}$, $\nu \leq \mu$ (lower triangle)

1: **for** $\mu = \overline{1, N_{\text{AO}}}$ **do** in parallel (OMP)
2: $\quad V_{B_{\text{lo}}}^{(p)} \leftarrow \texttt{GEMV}[(B_{\text{lo}}^{(p)}|\mu\nu),\ D_{\mu\nu}],\ 0 \leq \nu \leq \mu$
3: $\quad V_{B_{\text{up}}}^{(p)} \leftarrow \texttt{GEMV}[(B_{\text{up}}^{(p)}|\mu\nu),\ D_{\mu\nu}],\ \mu \leq \nu \leq N_{\text{AO}}$
4: **end for**
5: $\widetilde{V}_C \leftarrow \texttt{GEMV}[\left(V_{BC}^{-1}\right)^{(p)},\ V_B^{(p)}]$
6: $\widetilde{V}_C^{(p)} \leftarrow \texttt{reduce}\ (\widetilde{V}_C)$ (MPI)
7: **for** $\mu = \overline{1, N_{\text{AO}}}$ **do** in parallel (OMP)
8: $\quad J_{\mu\nu}^{(p)} \leftarrow \texttt{GEMV}[(C_{\text{lo}}^{(p)}|\mu\nu),\ \widetilde{V}_{\text{lo}}^{(p)}],\ 0 \leq \nu \leq \mu$
9: $\quad J_{\nu\mu}^{(p)} \leftarrow \texttt{GEMV}[(C_{\text{up}}^{(p)}|\mu\nu),\ \widetilde{V}_{\text{up}}^{(p)}],\ \mu \leq \nu \leq N_{\text{AO}}$
10: **end for**
11: $J_{\mu\nu} \leftarrow \texttt{Allreduce}[J_{\mu\nu}^{(p)}]$ (MPI)

instead of full diagonalization. But let us focus on the calculation of the Coulomb and exchange matrices as the RI approximation is applied exactly here.

Algorithm 4 outlines the construction of the Coulomb matrix according to Eq. (6) (RI-J). The alternative algorithm (RI-TJ) can be obtained by removing lines 5 and 6, so it requires less communication. The layout used introduces some additional difficulties compared to the straightforward implementation of the algorithm. Figure 2 presents two subblocks of the RI tensor to be contracted within the current iteration of the `for` loop in lines 1–4. These computations are parallelized using OpenMP. Lines 2 and 3 specify the axis for contraction. Elements of the first subblock are placed in lower triangles, while those of the second one are placed in upper triangles (see Figures 1b, 2 and 3). This fact is reflected by the lower indices "lo" and "up" of the auxiliary dimension index $B^{(p)}$. The step results into the local part of the intermediate vector $V_B^{(p)}$ on each process p. For vectorization purposes, it is better to store elements of this subvector with "lo" and "up" indices separately (combining them further if needed). Next, $V_B^{(p)}$ is contracted with the local submatrix $(V_{BC}^{-1})^{(p)}$ obtaining the partial sum of another entire intermediate vector $\widetilde{V}_C$ (line 5). The local parts $\widetilde{V}_C^{(p)}$ of $\widetilde{V}_C$ are then accumulated using the MPI `reduce` routine according to the initial partition along the auxiliary basis dimension $B^{(p)}$ (line 6). Lines 7–10 again refer to the subblocks shown in Fig. 2. Now, columns of these subblocks are contracted with $\widetilde{V}_C^{(p)}$, forming a partial sum $J_{\mu\nu}^{(p)}$. The "lo" and "up" indices were again involved. Note that for vectorization purposes in line 9 the upper term of $J_{\mu\nu}^{(p)}$ can be used (instead of the current approach). `for` loop

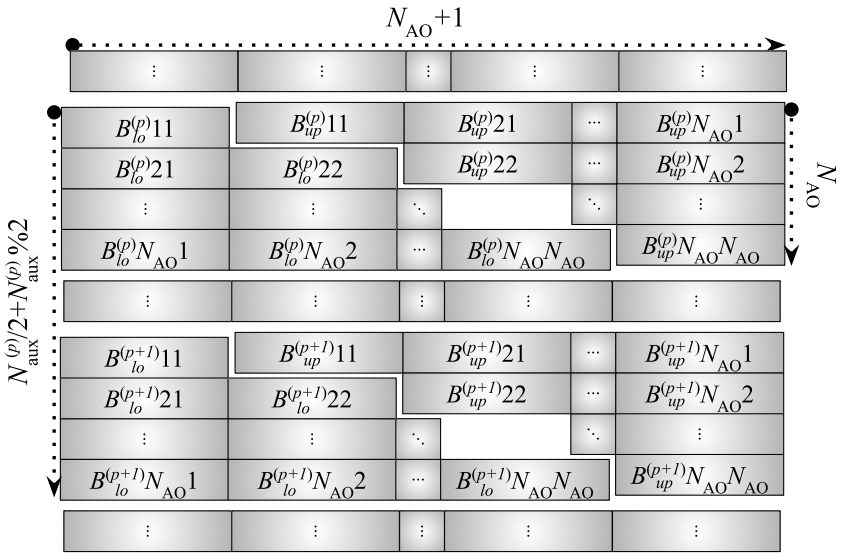


Figure 3. RI tensor representation for contractions within RI-K or RI-TK (Alg. 5)

Algorithm 5 Exchange matrix construction according to Eq. (7)

Input: permutationally-adapted three-center integrals $(B^{(p)}|\mu\nu)$ (1) (or $(\widetilde{P^{(p)}}|\mu\nu)$ (8)); MOs coefficients $C_{i\sigma}$ (3); Cholesky factor L_{BP} of V_{BC}^{-1} (see Eq. (2)); size of the main basis N_{AO}; size of the auxiliary basis N_{aux}

Output: $K_{\mu\nu}$, $\nu \leq \mu$ (lower triangle)

1: **for** $B = \overline{B^{(p)}/2 \ldots B^{(p+1)}/2}$ **do** in parallel (OMP)
2: $(2B|i\mu) \leftarrow$ `SYMM`$[C_{i\sigma}, (2B|\sigma\mu)]$
3: $(2B+1|i\mu) \leftarrow$ `SYMM`$[C_{i\sigma}, (2B+1|\sigma\mu)]$
4: **end for**
5: $(\mu\widetilde{i^{(p)}}|B) \leftarrow (B^{(p)}|i\mu)$
6: $(\mu\widetilde{i^{(p)}}|P) \leftarrow$ `TRMM`$[(\mu\widetilde{i^{(p)}}|B), L_{BP}]$
7: $K_{\mu\nu}^{(p)} \leftarrow$ `SYRK`$[(\mu\widetilde{i^{(p)}}|P)]$
8: $K_{\mu\nu} \leftarrow$ `Allreduce`$[K_{\mu\nu}^{(p)}]$ (MPI)

is parallelized with OpenMP again. Finally, all processes obtain the Coulomb matrix through the MPI `Allreduce` routine.

Algorithm 5 outlines the construction of the exchange matrix according to Eq. (7) (RI-K). The alternative algorithm (RI-TK) can be obtained just by removing lines 5 and 6, although the difference is also that the BLAS `SYRK` routine would be applied to the intermediate tensor part $(\widetilde{P^{(p)}}|i\mu)$ instead of $(\mu\widetilde{i^{(p)}}|P)$. Figure 3 shows a pair of triangles of the RI tensor to be contracted within the current iteration of the `for` loop in lines 1–4, parallelized using OpenMP. If the last triangle is zero-valued, it is more convenient to process the corresponding pair out of the `for` loop. After computations on lines 1–4, global transposition is done (line 5) in order to optimally perform contraction of the intermediate tensor with Cholesky factor L_{BP} of V_{BC}^{-1} (line 6). The partial sum of $K_{\mu\nu}$ is then obtained in line 7. Finally, all processes obtain an exchange matrix using the MPI `Allreduce` routine.

Algorithm 4 can be optimized by overlapping computation in line 5 with communication in line 6, although it is actual for a large number of processes. Algorithm 5 can also be modified in

this way by overlapping computation in lines 1–4 with communication in line 5. Both algorithms can be optimized using only lower triangles of symmetric matrices $J_{\mu\nu}$ and $K_{\mu\nu}$ for the final `Allreduce` operation. The latter can also overlap with the previous computations. All of these modifications are to be implemented in the future, as well as the optimization of Algorithm 3.

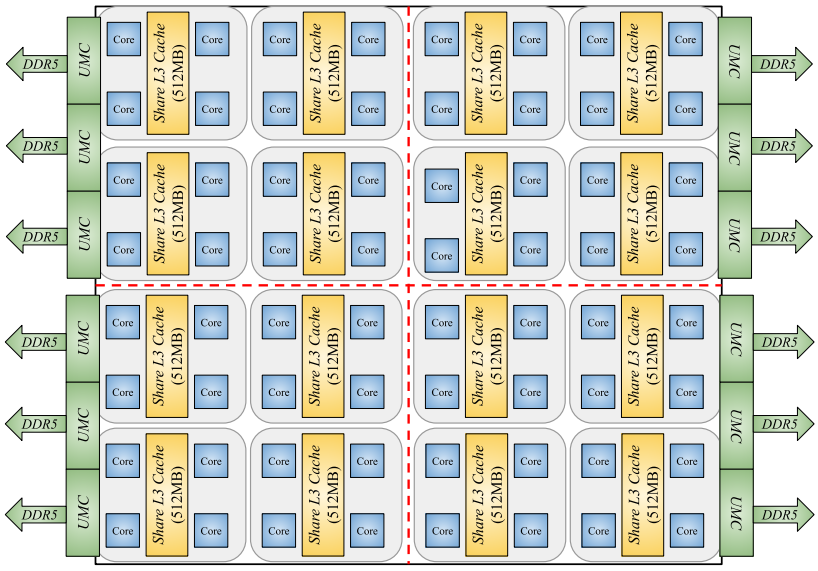
3. Computational Details

The MPI+OpenMP algorithms described in the previous section were implemented in the BUFO quantum chemistry simulation program. The newly developed code was benchmarked on 32× 6-core Intel Xeon E5-1650v3 @ 3.5GHz processors (Desmos at JIHT RAS [24, 40]) and 2× 64-core AMD EPYC 9535 processors (Irbis at MIPT). The first HPC server is a typical distributed memory system (see Fig. 4b and [46]), while the memory of the second workstation is shared between its processors (Fig. 4a and [2, 55]). So, two different strategies were adopted to run the hybrid MPI+OpenMP program using the `map-by` option of the OpenMPI library [13]. OpenMPI 5.0.0 and 4.1.6 were used for the Desmos and Irbis systems, respectively. MPI exchanges in case of the Desmos supercomputer were performed through Infiniband FDR, and in case of the Irbis server through shared memory. For Desmos each MPI process was bound to a processor with 6 cores sharing L3 cache. The number of OpenMP threads was varied only after all proccesors were occupied by MPI ranks. For the Irbis machine, each process was bound to the L3 cache shared by 4 cores of AMD EPYC 9535 processor. Different processor sets are possible for Irbis: NPS1, NPS2 and NPS4[7] (corresponding modes are reflected by red dotted lines in Fig. 4a; see also [55]). The simplest NPS1 was applied, so each processor is represented as a single NUMA node. At first, processes were bound to the caches of the first node. The remaining node was used only for tests with 32 MPI processes (and 1, 2 or 4 OpenMP threads). Alternatively, MPI processes can be gradually (with increase of their number) bound to L3 caches of different nodes again, enabling multithreading after 32 MPI processes would be distributed. Another policy is to reflect 2 processes to the NUMA nodes, increasing the number of threads up to 64 on each node. Although our code might not be optimal for such calculations due to the granularity used to calculate three-center ERIs (see Section 2.2) oriented to distributed computations rather than multithreaded on NUMA.

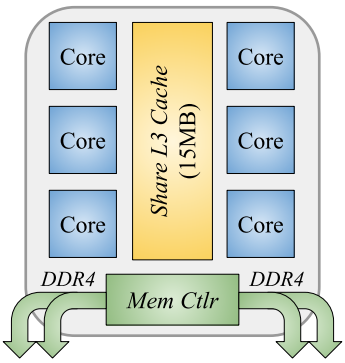
Other differences between the Desmos and Irbis benchmarks lie in the compilers and linear algebra libraries involved. In Desmos, the code was compiled with GCC 12.2.0 and linked with MKL 2025.0, while in Irbis, GCC 13.3.0 and OpenBLAS 3.30 were used. Both compilers support OpenMP 4.5, which is sufficient for all parallel constructs used in this work.

Currently, quantum chemistry calculations are widely used to simulate biomolecules and processes in living matter [29]. In our previous work [26], we already considered a chlorophyll molecule $C_{55}H_{72}O_5N_4Mg$ and its dimer in vacuum; the def2-SVP basis [50] supplemented by the def2-SVP-RIFIT auxiliary basis [19, 51] was used for orbital expansions (3). In this work, the dimer was additionally surrounded by 48 water molecules. Preliminary simulations of the equilibrium conformations of the chlorophyll dimer in a water solution (up to approximately 200 molecules) were performed using molecular dynamics (MD) with the potential [56]. The geometries were provided by Egor Igolnikov; then the number of H_2O molecules was reduced to 48 in order to fit in RAM of our computational facilities. Two main conformations of the chlorophyll dimer were observed in those simulations. The difference in their relative energy is determined

[7] NPS = NUMA (non-uniform memory access) node per socket.



(a) Principal scheme of the 64-core AMD EPYC 9535 processor (see also [2, 55])



(b) Principal scheme of the 6-core Intel® Xeon® E5-1650v3 @ 3.5GHz processor (see also [46])

Figure 4. Principal schemes workstation nodes used in benchmark calculations

mainly by the electrostatic attraction between the porphyrin rings and the interaction of the molecule tails with water. The chlorophyll molecule has a hydrophobic tail and a hydrophilic ring, so it is amphiphilic in general. In water, chlorophylls aggregate in pairs. On the one hand, π-stacking interactions of conjugated systems lead to their rendezvous; on the other hand, it is favorable to place the hydrophobic tails between the porphyrin rings to minimize their contact with water. Although the HF method does not account for electronic correlation, it might be the basis for more advanced calculations. So our tests were performed using one of the equilibrium conformations. The sizes used were $N_{\text{atoms}} = 322$, $N_{\text{occ}} = 722$, $N_{\text{AO}} = 3700$ (grouped into $N_{\text{AO}}^{\text{sh}} = 1792$) and $N_{\text{aux}} = 11896$ (grouped into $N_{\text{aux}}^{\text{sh}} = 4288$).

4. Results and Discussion

Performance tests required about 800 GB of RAM over all processes during the program execution. Our benchmarks on Desmos involved 32 MPI processes with 1 to 6 OpenMP threads. Less processes were not used because of memory limits. In contrast, there were no problems with the available memory workspace on the Irbis machine, so subsequently bindings of MPI processes from 1 to 32 were performed with relevant turn on of OpenMP multithreading.

Records of the type "$\#N_1$" or "$\#N_1$–N_2" in the table column name refer to line or lines of the corresponding algorithm described above in Section 2. "wall" means wall execution time. "Sp. (x)" stands for the speedup of calculation with the increase of the number N of parallel threads involved. The speedup is always evaluated from the wall time.

4.1. Scalability: Calculation of ERIs

Figure 5 and Table 1 show the strong scaling behavior of the integration algorithms within the RI-JK and RI-TJK approaches. Speedups in Tab. 1 are defined by the slowest process. Multithreading on Desmos performs well enough, giving up to 5.0× and 5.6× speedup for RI-JK

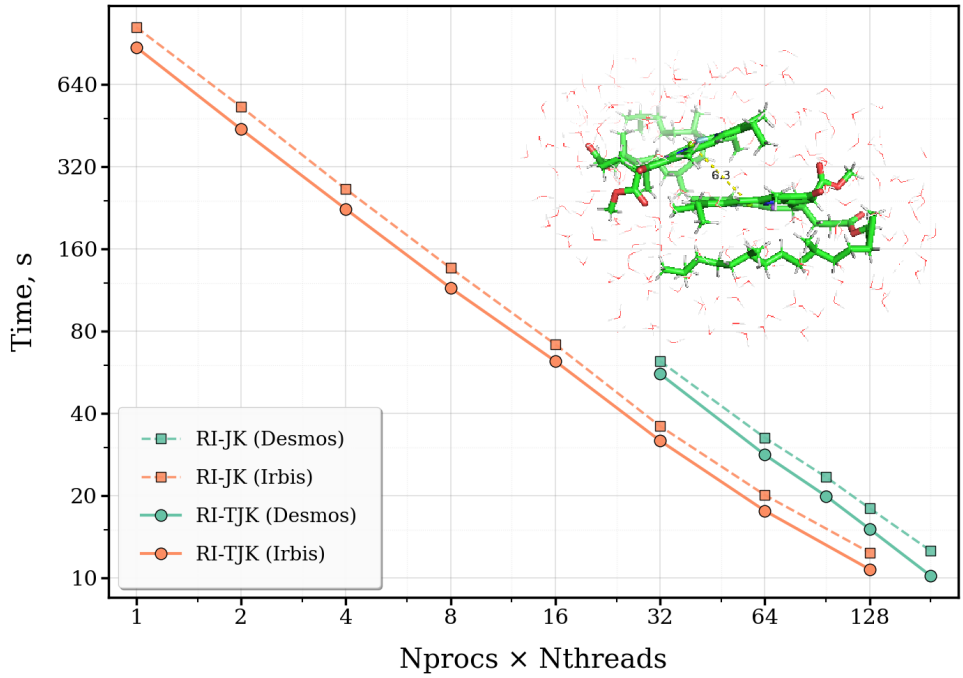


Figure 5. Scalability of the three-center ERIs calculation

Table 1. Breakdown of scalability of the three-center ERIs calculation

N	RI-JK				RI-TJK			
	Desmos		Irbis		Desmos		Irbis	
	Time, s	Sp. (×)	Time, s	Sp. (×)	Time, s	Sp. (×)	Time, s	Sp. (×)
1	–	–	1037.9	1.0×	–	–	877.1	1.0×
2	–	–	530.9	2.0×	–	–	439.7	2.0×
4	–	–	265.3	3.9×	–	–	224.2	3.9×
8	–	–	136.4	7.6×	–	–	115.4	7.6×
16	–	–	71.7	14.5×	–	–	62.2	14.1×
32	62.2	1.0×	36.0	28.8×	55.9	1.0×	31.8	28.0×
64	32.7	1.9×	20.1	51.5×	28.3	2.0×	17.6	49.7×
96	23.4	2.7×	–	–	20.0	2.8×	–	–
128	18.0	3.5×	12.4	83.6×	15.1	3.7×	10.8	81.4×
192	12.6	5.0×	–	–	10.2	5.6×	–	–

and RI-TJK, respectively, when going from 32 to 192 threads. It is not ideal (6.0×) because of the uneven initial static distribution at the stage of scheduling the calculation of three-centers of ERIs (see Alg. 1, line 1). The observed speedup of the RI-TJK implementation is slightly better. The probable reason for this is that $N^{\text{sh}}_{\text{aux}} \sim 2.5 N^{\text{sh}}_{\text{AO}}$, i.e., more integration tasks were distributed between processes. Though this consideration does not primarily apply to the analogous comparison of runnings on Irbis: multithreaded integration (i.e., when more than 32 threads are involved) scales similarly for both algorithms. Interestingly, the speedup of calculations with multithreading on Irbis is worse than the speedup up to 32 processes, when all L3 caches are bound to MPI ranks. Actually, the scalability is good up to 32 processes involved, resulting in 28.8× or 28.0× speedup. Better scalability can be achieved by means of dynamic scheduling, because

there is about 2.5 seconds between the fastest and the slowest processes. Nevertheless, scalability up to 32 processes is greater than that after turning on multithreading. This observation is true for other benchmarks as well (see the following subsections). The exact reason for that is unclear. Although in case of integration the probable reason is a bad decision to parallelize the nested `for` loop (see Alg. 2) with OpenMP instead of the external one.

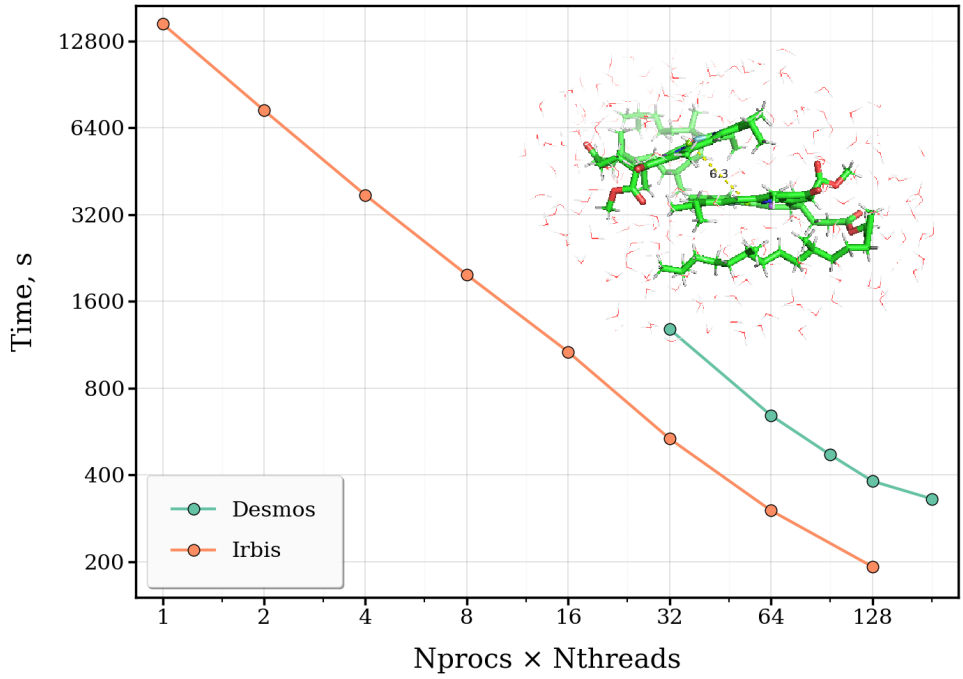


Figure 6. Scalability of the three-center ERIs transformation

Table 2. Breakdown of scalability of the three-center ERIs transformation

N	Desmos Time, s #1–3	Desmos Time, s #4–7	Desmos Time, s wall	Desmos Sp. (×)	Irbis Time, s #1–3	Irbis Time, s #4–7	Irbis Time, s wall	Irbis Sp. (×)
1	–	–	–	–	14435.7	267.4	14703.2	1.0×
2	–	–	–	–	7216.5	148.6	7365.2	2.0×
4	–	–	–	–	3654.4	86.2	3740.6	3.9×
8	–	–	–	–	1926.8	51.9	1978.7	7.4×
16	–	–	–	–	1029.4	39.2	1068.6	13.8×
32	1255.6	25.9	1281.5	1.0×	508.9	25.1	534.1	27.5×
64	624.7	18.4	643.1	2.0×	280.5	20.2	300.7	48.9×
96	451.9	17.5	469.4	2.7×	–	–	–	–
128	366.8	13.6	380.4	3.4×	171.8	20.2	192.1	76.6×
192	318.0	12.3	330.3	3.9×	–	–	–	–

When comparing the Desmos and Irbis performance, it can be seen that the computation time on 32 processes differs by approximately a factor of two. This can be attributed to the twofold advantage of the Irbis CPU cores in terms of FLOP/s due to the AVX-512 instruction set compared to AVX2 available for the Desmos' CPUs.

4.2. Scalability: Transformation of ERIs

Figure 6 and Table 2 show the strong scaling behavior of the three-center ERI transformation (8) and their subsequent transposition for the RI-TJK approach. The `TRMM` operations speedup is great, nearly ideal up to 32 processes, while the scaling of the transposition step is

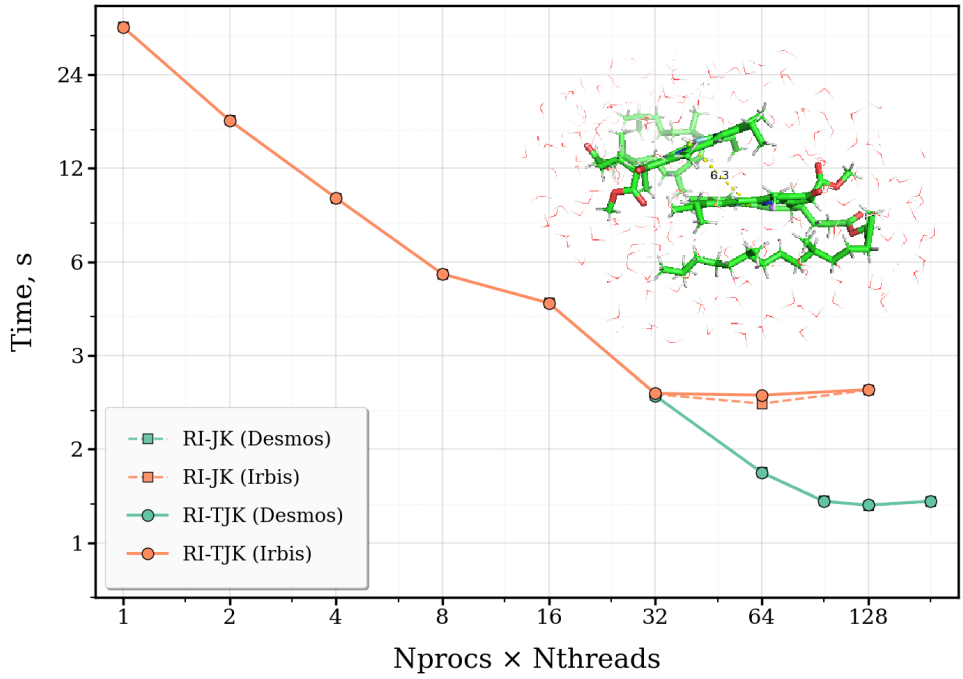


Figure 7. Scalability of the RI-J/TJ algorithm

Table 3. Breakdown of scalability of the RI-J/TJ algorithm

N	RI-J/TJ (Desmos) Time, s #1–4	#7–10	wall	Sp. (×)	RI-J/TJ (Irbis) Time, s #1–4	#7–10	wall	Sp. (×)
1	–	–	–	–	17.21	16.96	34.18	1.0×
2	–	–	–	–	8.61	8.50	17.12	2.0×
4	–	–	–	–	4.86	4.77	9.64	3.6×
8	–	–	–	–	2.71	2.78	5.49	6.2×
16	–	–	–	–	2.18	2.25	4.44	7.7×
32	1.05	1.18	2.23	1.0×	1.10	1.16	2.25	15.2×
64	0.59	0.68	1.27	1.8×	1.00	1.10	2.10	16.3×
96	0.47	0.55	1.02	2.1×	–	–	–	–
128	0.45	0.53	0.99	2.2×	1.08	1.25	2.32	14.7×
192	0.45	0.57	1.02	2.1×	–	–	–	–

not so good. The former operations are not ideally scaled because the rebalancing of the RI tensor after the parallel evaluation of ERIs was not implemented for now. So after the RI tensor is computed, it is still unevenly distributed between MPI processes. The transposition step requires much less time, but can become a bottleneck, when more and more processes are involved. For the computation on 128 threads of Irbis it already takes about 10% of the wall time. This obsta-

cle significantly limits scalability, though the speedup of multithreaded calculations is generally worse again even in spite of the transposition step. The same conclusions about multithreading are also true for Desmos. The reasons for this are currently unclear.

It can be noted that the time needed for the transformation step on 32 processes of Desmos and Irbis again differs by approximately a factor of two. Actually, the transposition step in both tests is done in the same time. So, Irbis offers twice the computational performance of Desmos, but its memory bandwidth is not proportionally higher. As a result, it reaches the memory wall sooner.

4.3. Scalability: the SCF Procedure

Figure 7 and Table 3 show the strong scaling behavior of the RI-J and RI-TJ algorithms (lines #5 and #6 are out of interest here, because corresponding transformations were significantly faster than for other steps). Neither algorithm exhibits satisfactory scalability across the full range of the number of threads. The possible reason for performance degradation is the large intermediate buffer used for the reduction of $J^{(p)}_{\mu\nu}$ using OpenMP. Being allocated on the stack in current implementations, this buffer induces significant cache pressure, resulting in frequent cache misses and limited data reuse. Even the twofold superiority typical for Irbis on 32 processes is not observed. This situation is to be resolved in the future, though RI-J and RI-TJ still take much less time than RI-K and RI-TK.

Figure 8 and Tables 4 and 5 show the strong scaling behavior of the RI-K and RI-TK algorithms. Now, the typical picture of nearly ideal speedup up to 27.1× and 28.0× on 32 processes of Irbis is observed, as well as more moderate scalability after that. RI-K scalability is worse, because there is no three-center ERIs rebalancing in current RI-K implemetation and the transposition is memory-bound. The transformation within RI-K takes about 5% of the execution time. Benchmarks performed on Desmos show that multithreading can be efficiently incorpo-

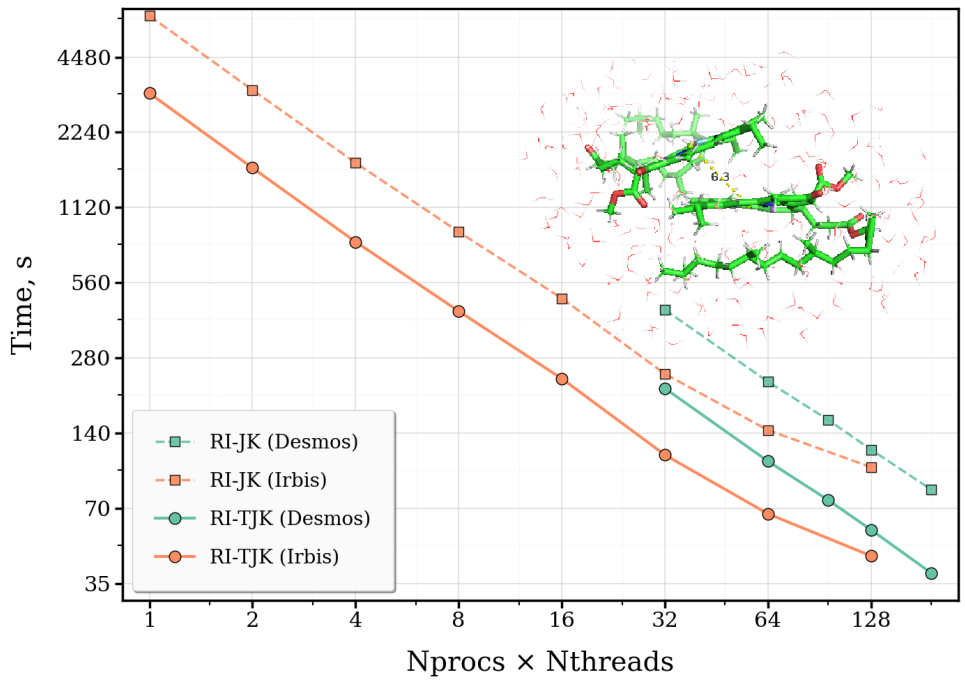


Figure 8. Scalability of the RI-K/TK algorithm

Table 4. Breakdown of scalability of the RI-K algorithm

N	RI-K (Desmos) Time, s					Sp.	RI-K (Irbis) Time, s					Sp.
	#1–4	#5	#6	#7	wall	(×)	#1–4	#5	#6	#7	wall	(×)
1	–	–	–	–	–	–	2138.4	64.3	3296.3	1053.7	6552.7	1.0×
2	–	–	–	–	–	–	1070.7	38.2	1666.7	532.7	3308.2	2.0×
4	–	–	–	–	–	–	543.2	22.6	853.4	272.8	1692.0	3.9×
8	–	–	–	–	–	–	287.1	13.8	452.6	142.7	896.2	7.3×
16	–	–	–	–	–	–	154.2	10.6	242.8	78.1	485.7	13.5×
32	141.8	6.7	220.2	68.4	437.1	1.0×	76.4	6.6	120.5	38.7	242.2	27.1×
64	72.3	5.5	111.4	36.2	225.4	1.9×	44.3	6.2	70.7	23.1	144.2	45.4×
96	50.4	4.9	77.8	25.3	158.4	2.8×	–	–	–	–	–	–
128	38.2	4.6	58.6	19.1	120.6	3.6×	29.8	6.1	50.4	16.4	102.6	63.9×
192	25.5	4.6	40.3	13.0	83.4	5.2×	–	–	–	–	–	–

Table 5. Breakdown of scalability of the RI-TK algorithm

N	RI-TK (Desmos) Time, s			Sp.	RI-TK (Irbis) Time, s			Sp.
	#1–4	#7	wall	(×)	#1–4	#7	wall	(×)
1	–	–	–	–	2138.7	1074.3	3213.0	1.0×
2	–	–	–	–	1070.2	544.2	1614.4	2.0×
4	–	–	–	–	542.9	274.2	817.1	3.9×
8	–	–	–	–	287.0	144.5	431.5	7.5×
16	–	–	–	–	154.1	77.8	231.9	13.9×
32	141.7	70.3	212.1	1.0×	76.3	38.5	114.8	28.0×
64	72.2	36.1	108.3	2.0×	44.1	22.6	66.7	48.2×
96	50.4	25.4	75.7	2.8×	–	–	–	–
128	38.2	19.4	57.6	3.7×	29.8	15.5	45.3	70.9×
192	25.5	13.2	38.7	5.5×	–	–	–	–

rated: up to 5.2× and 5.5× for RI-K and RI-TK, respectively. The reason for the bad results of calculations employing multithreading on Irbis is again unclear. It can be noted that the overall time including all stages (except for the transposition) on 32 processes of Desmos and Irbis differs by approximately a factor of two as before. Also, the transposition step in the RI-K tests is done in the same time on Desmos and Irbis at this number of processes. It is consistent with the coincidence of transposition times observed in the previous section after the transformation of three-center ERIs (8).

Conclusion

We report a new hybrid MPI+OpenMP implementation of the resolution-of-the-identity restricted Hartree—Fock method extensively employing permutational symmetry of the ERI tensor to minimize explicit local memory movement during iterations of SCF procedure. Both versions of the proposed parallel algorithm (RI-JK and RI-TJK) were benchmarked on a system of two

chlorophyll molecules surrounded by 48 water molecules (3700 basis functions supplemented with 11896 auxiliary basis functions) to reveal their performance and scalability features; two different machines with different computer architectures were used throughout to perform these benchmarks.

The RI-JK algorithm allows us to avoid any global transpositions of three-center ERIs at the cost of performing more moderate transposition at each SCF iteration, whereas the RI-TJK algorithm performs one global transposition of these integrals only once before the start of the SCF procedure (see Tab. 2 and Tab. 4 for comparison). Fortunately, in both cases the communication computation overlap can greatly improve the overall efficiency of the algorithms when there would be too many communications. It should be true for the larger number of processes involved, because even for the case of 128 threads communication already takes about 5–10% of wall time. The contractions that can be actually overlapped also differ, because avoiding the pre-transformation (8) of formal complexity $O(N_{\mathrm{AO}}^2 N_{\mathrm{aux}}^2)$ before the SCF procedure leads to including $O(N_{\mathrm{AO}} N_{\mathrm{aux}}^2)$ arithmetic operations at each SCF iteration. Generally, the RI-JK algorithm is recommended when a rather small number of SCF iterations is expected, while RI-TJK is more beneficial for SCF calculations expected to converge slowly. For both scenarios, it was shown that the permutationally-adapted memory layout for storing three-center ERIs can be efficiently used, though some technical problems should be resolved especially in case of RI-J and RI-TJ algorithms. However, the permutationally-adapted memory layout removes the need to perform redundant copy operations in RI-K and RI-TK at each SCF iteration compared to conventional parallel RI-SCF implementations unpacking three-center ERIs each iteration. Another issue to be studied in detail in the future is how to deal with RI tensor sparsity using the permutationally-adapted layout. Performance evaluation shows that the proposed approaches to parallelization of the RI-SCF method are reasonably scalable on modern computational architectures (up to 70–80× speedup on 128 threads), though some problems should be identified and addressed in the future in order to obtain the scalability closer to the ideal one.

Acknowledgements

The authors thank Egor Igolnikov for providing configurations of the chlorophyll dimer in water from MD simulations.

This work was supported by the Ministry of Science and Higher Education of the Russian Federation (State Assignment No. 075-00270-26-00).

References

1. Alexeev, Y., Kendall, R.A., Gordon, M.: The distributed data SCF. Computer Physics Communications 143(1), 69–82 (2002). https://doi.org/10.1016/S0010-4655(01)00439-8

2. AMD: AMD EPYC$^{\mathrm{TM}}$ 9005 Proceessor Architecture Overview. https://docs.amd.com/v/u/en-US/58462_amd-epyc-9005-tg-architecture-overview (2025), accessed: 2026-02-15

3. Blackford, L.S., Demmel, J., Dongarra, J., *et al.*: An updated set of basic linear algebra subprograms (BLAS). ACM Transactions on Mathematical Software 28(2), 135–151 (2002). https://doi.org/10.1145/567806.567807

4. Bussy, A., Schütt, O., Hutter, J.: Sparse tensor based nuclear gradients for periodic Hartree-Fock and low-scaling correlated wave function methods in the CP2K software package:

A massively parallel and GPU accelerated implementation. Journal of Chemical Physics 158(16), 164109 (2023). https://doi.org/10.1063/5.0144493

5. Bussy, A., Hutter, J.: Efficient periodic resolution-of-the-identity Hartree-Fock exchange method with k-point sampling and Gaussian basis sets. Journal of Chemical Physics 160(6), 064116 (2024). https://doi.org/10.1063/5.0189659

6. Calaminici, P., Domínguez-Soria, V.D., Geudtner, G., *et al.*: Parallelization of three-center electron repulsion integrals. Theoretical Chemistry Accounts 115(4), 221–226 (2005). https://doi.org/10.1007/s00214-005-0005-0

7. Dyczmons, V.: No N^4-dependence in the calculation of large molecules. Theoretical Chemistry Accounts 28(3), 307–310 (1973). https://doi.org/10.1007/BF00533492

8. Echenique, P., Alonso, J.L.: A mathematical and computational review of Hartree–Fock SCF methods in quantum chemistry. Molecular Physics 105(23–24), 3057–3098 (2007). https://doi.org/10.1080/00268970701757875

9. Eichkorn, K., Treutler, O., Öhm, H., *et al.*: Auxiliary basis sets to approximate Coulomb potentials. Chemical Physics Letters 242(4), 652–660 (1995). https://doi.org/10.1016/0009-2614(95)00838-u

10. Eichkorn, K., Weigend, F., Treutler, O., Ahlrichs, R.: Auxiliary basis sets for main row atoms and transition metals and their use to approximate Coulomb potentials. Theoretical Chemistry Accounts 97(1–4), 119–124 (1997). https://doi.org/10.1007/s002140050244

11. Foster, I., Tilson, J.L., Wagner, A., *et al.*: Toward high-performance computational chemistry: I. Scalable Fock matrix construction algorithms. Journal of Computational Chemistry 17(1), 109–123 (1996). https://doi.org/10.1002/(SICI)1096-987X(19960115)17:1<109::AID-JCC9>3.0.CO;2-V

12. Früchtl, H.A., Kendall, R.A., Harrison, R.J., Dyall, K.G.: An implementation of RI-SCF on parallel computers. International Journal of Quantum Chemistry 64(1), 63–69 (1997). https://doi.org/10.1002/(SICI)1097-461X(1997)64:1<63::AID-QUA7>3.0.CO;2-%23

13. Gabriel, E., Fagg, G.E., Bosilca, G., *et al.*: Open MPI: Goals, Concept, and Design of a Next Generation MPI Implementation. In: Recent Advances in Parallel Virtual Machine and Message Passing Interface. Lecture Notes in Computer Science, vol. 3241, pp. 97–104. Springer Berlin Heidelberg (2004). https://doi.org/10.1007/978-3-540-30218-6_19

14. Gill, P.M., Johnson, B.G., Pople, J.A.: A simple yet powerful upper bound for Coulomb integrals. Chemical Physics Letters 217(1), 65–68 (1994). https://doi.org/10.1016/0009-2614(93)E1340-M

15. Glebov, I.O., Poddubnyi, V.V.: An effective algorithm of the Hartree–Fock approach with the storing of two-electron integrals in the resolution of identity approximation. Russian Journal of Physical Chemistry A 98(4), 617–625 (2024). https://doi.org/10.1134/S0036024424040101

16. Guidon, M., Hutter, J., VandeVondele, J.: Auxiliary density matrix methods for Hartree-Fock exchange calculations. Journal of Chemical Theory and Computation 6(8), 2348–2364 (2010). https://doi.org/10.1021/ct1002225

17. Harrison, R.J., Guest, M.F., Kendall, R.A., *et al.*: Toward high-performance computational chemistry: II. A scalable self-consistent field program. Journal of Computational Chemistry 17(1), 124–132 (1996). https://doi.org/10.1002/(SICI)1096-987X(19960115)17:1<124::AID-JCC10>3.0.CO;2-N

18. Häser, M., Ahlrichs, R.: Improvements on the direct SCF method. Journal of Computational Chemistry 10(1), 104–111 (1989). https://doi.org/10.1002/jcc.540100111

19. Hellweg, A., Hättig, C., Höfener, S., Klopper, W.: Optimized accurate auxiliary basis sets for RI-MP2 and RI-CC2 calculations for the atoms Rb to Rn. Theoretical Chemistry Accounts 117, 587–597 (2007). https://doi.org/10.1007/s00214-007-0250-5

20. Hollman, D.S., Schaefer, H.F., Valeev, E.F.: Fast construction of the exchange operator in an atom-centred basis with concentric atomic density fitting. Molecular Physics 115(17–18), 2065–2076 (2017). https://doi.org/10.1080/00268976.2017.1346312

21. Hollman, D.S., Schaefer, H.F., Valeev, E.F.: A tight distance-dependent estimator for screening three-center Coulomb integrals over Gaussian basis functions. Journal of Chemical Physics 142(15), 154106 (2015). https://doi.org/10.1063/1.4917519

22. Huang, H., Sherill, C.D., Chow, E.: Techniques for high-performance construction of Fock matrices. Journal of Chemical Physics 152(2), 024122 (2020). https://doi.org/10.1063/1.5129452

23. Ishimura, K., Kuramoto, K., Ikuta, Y., Hyodo, S.A.: MPI/OpenMP Hybrid Parallel Algorithm for Hartree-Fock Calculations. Journal of Chemical Theory and Computation 6(4), 1075–1080 (2010). https://doi.org/10.1021/ct100083w

24. Ismagilov, T., Mukosey, A., Smirnov, F., *et al.*: Towards performance analysis of GPU-aware MPI over Angara interconnect. International Journal of High Performance Computing Applications 40(2), 240–253 (2026). https://doi.org/10.1177/10943420251411961

25. Kashpurovich, I.V., Oleynichenko, A.V., Stegailov, V.V.: Achieving the maximum performance of the resolution of the identity approximation in the Hartree-Fock method. In: Sokolinsky, L., Zymbler, M. (eds.) Parallel Computational Technologies. Communications in Computer and Information Science, vol. 2891, pp. 239–263. Springer (2026)

26. Kashpurovich, I.V., Oleynichenko, A.V., Stegailov, V.V.: Development of strategies for parallel implementation of the Hartree-Fock theory in resolution-of-the-identity approximation. Russian Journal of Physical Chemistry A 100(5), 1013–1036 (2026)

27. Kashpurovich, I.V., Oleynichenko, A.V., Stegailov, V.V.: NUMA-aware OpenMP algorithm for three-center electron repulsion integrals. In: Voevodin, V., Antonov, A., Nikitenko, D. (eds.) Supercomputing. Lecture Notes in Computer Science, vol. 16196, pp. 333–350. Springer, Cham (2026). https://doi.org/10.1007/978-3-032-13127-0

28. Laikov, D.N.: Fast evaluation of density functional exchange-correlation terms using the expansion of the electron density in auxiliary basis sets. Chemical Physics Letters 281(1–3), 151–156 (1997). https://doi.org/10.1016/s0009-2614(97)01206-2

29. Lankin, A.V., Norman, G.E.: Introduction to quantum mechanics of living matter. Russian Journal of Physical Chemistry A 99(6), 1416–1445 (2025). https://doi.org/10.1134/s0036024425700785

30. Le, H.A., Shiozaki, T.: Occupied-orbital fast multipole method for efficient exact exchange evaluation. Journal of Chemical Theory and Computation 14(3), 1228–1234 (2018). https://doi.org/10.1021/acs.jctc.7b00880

31. Manzer, S., Horn, P.R., Mardirossian, N., Head-Gordon, M.: Fast, accurate evaluation of exact exchange: the occ-RI-K algorithm. Journal of Chemical Physics 143(2), 024113 (2015). https://doi.org/10.1063/1.4923369

32. Marek, A., Blum, V., Johanni, R., *et al.*: The ELPA library: scalable parallel eigenvalue solutions for electronic structure theory and computational science. Journal of Physics: Condensed Matter 26(21), 213201 (2014). https://doi.org/10.1088/0953-8984/26/21/213201

33. Mejía-Rodríguez, D., Köster, A.M.: Robust and efficient variational fitting of Fock exchange. Journal of Chemical Physics 141(12), 124114 (2014). https://doi.org/10.1063/1.4896199

34. Merlot, P., Kjærgaard, T., Helgaker, T., *et al.*: Attractive electron–electron interactions within robust local fitting approximations. Journal of Computational Chemistry 34(17), 1486–1496 (2013). https://doi.org/10.1002/jcc.23284

35. Mironov, V.A., Alexeev, Y., Keipert, K., *et al.*: An efficient MPI/OpenMP parallelization of the Hartree-Fock method for the second generation of Intel® Xeon Phi™ processor. In: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC 2017. vol. 31, pp. 1–12. ACM (2017). https://doi.org/10.1145/3126908.3126956

36. Neese, F.: An improvement of the resolution of the identity approximation for the formation of the Coulomb matrix. Journal of Computational Chemistry 24(14), 1740–1747 (2003). https://doi.org/10.1002/jcc.10318

37. Oleynichenko, A.V., Zaitsevskii, A., Mosyagin, N.S., *et al.*: LIBGRPP: A library for the evaluation of molecular integrals of the generalized relativistic pseudopotential operator over Gaussian functions. Symmetry 15(1), 197 (2022). https://doi.org/10.3390/sym15010197

38. Reine, S., Tellgren, E., Krapp, A., *et al.*: Variational and robust density fitting of four-center two-electron integrals in local metrics. Journal of Chemical Physics 129(10), 104101 (2008). https://doi.org/10.1063/1.2956507

39. Shiozaki, T.: BAGEL : Brilliantly Advanced General Electronic-structure Library. Wiley Interdisciplinary Reviews: Computational Molecular Science 8(1), e1331 (2017). https://doi.org/10.1002/wcms.1331

40. Stegailov, V., Dlinnova, E., Ismagilov, T., *et al.*: Angara interconnect makes GPU-based Desmos supercomputer an efficient tool for molecular dynamics calculations. International Journal of High Performance Computing Applications 33(3), 507–521 (2019). https://doi.org/10.1177/1094342019826667

41. Stegailov, V., Smirnov, G., Vecher, V.: VASP hits the memory wall: Processors efficiency comparison. Concurrency and Computation: Practice and Experience 31(19), e5136 (2019). https://doi.org/10.1002/cpe.5136

42. Stocks, R., Palethorpe, E., Barca, J.: Multi-GPU RI-HF energies and analytic gradients – toward high-throughput ab initio molecular dynamics. Journal of Chemical Theory and Computation 20(17), 7503–7515 (2024). https://doi.org/10.1021/acs.jctc.4c00877

43. Sun, Q.: Libcint: An efficient general integral library for Gaussian basis functions. Journal of Computational Chemistry 36(22), 1664–1671 (2015). https://doi.org/10.1002/jcc.23981

44. Sun, Q.: Efficient Hartree-Fock exchange algorithm with Coulomb range separation and long-range density fitting. Journal of Chemical Physics 159(22), 224101 (2023). https://doi.org/10.1063/5.0178266

45. Sun, Q., Berkelbach, T.C., Blunt, N.S., *et al.*: PySCF: the Python-based simulations of chemistry framework. Wiley Interdisciplinary Reviews: Computational Molecular Science 8(1), e1340 (2017). https://doi.org/10.1002/wcms.134

46. Tugov, A.: Dedicated servers based on Intel Xeon E5v3 processors. https://selectel.ru/blog/vydelennye-servery-na-baze-processorov-intel-xeon-e5v3/ (2014), accessed: 2026-02-15

47. Vahtras, O., Almlöf, J., Feyereisen, M.: Integral approximations for LCAO-SCF calculations. Chemical Physics Letters 213(5–6), 514–518 (1993). https://doi.org/10.1016/0009-2614(93)89151-7

48. Valeev, E.F., Shiozaki, T.: Comment on "A tight distance-dependent estimator for screening three-center Coulomb integrals over Gaussian basis functions" [Journal of Chemical Physics 142, 154106 (2015)]. Journal of Chemical Physics 153(9), 097101 (2020). https://doi.org/10.1063/5.0020567

49. Weigend, F.: A fully direct RI-HF algorithm: Implementation, optimised auxiliary basis sets, demonstration of accuracy and efficiency. Physical Chemistry Chemical Physics 4(18), 4285–4291 (2002). https://doi.org/10.1039/b204199p

50. Weigend, F., Ahlrichs, R.: Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. Physical Chemistry Chemical Physics 7(18), 3297–3305 (2005). https://doi.org/10.1039/B508541A

51. Weigend, F., Häser, M., Patzelt, H., Ahlrichs, R.: RI-MP2: optimized auxiliary basis sets and demonstration of efficiency. Chemical Physics Letters 294(1–3), 143–152 (1998). https://doi.org/10.1016/s0009-2614(98)00862-8

52. Windom, Z.W., Bartlett, R.J.: On the iterative diagonalization of matrices in quantum chemistry: reconciling preconditioner design with Brillouin-Wigner perturbation theory. Journal of Chemical Physics 158(13), 134107 (2023). https://doi.org/10.1063/5.0139295

53. Wu, X., Sun, Q., Pu, Z., *et al.*: Enhancing GPU-Acceleration in the Python-Based Simulations of Chemistry Frameworks. Wiley Interdisciplinary Reviews: Computational Molecular Science 15(2), e70008 (2025). https://doi.org/10.1002/wcms.70008

54. Xing, L., Patel, A., Chow, E.: A new scalable parallel algorithm for Fock matrix construction. In: 2014 IEEE 28th International Parallel and Distributed Processing Symposium. pp. 902–914 (2014). https://doi.org/10.1109/IPDPS.2014.97

55. Xu, P., Yang, K.: Balanced Memory Configurations with 5th Generation AMD EPYC Processors. https://lenovopress.lenovo.com/lp2283-balanced-memory-configurations-with-5th-generation-amd-epyc-processors (2025), accessed: 2026-02-15

56. Zhang, L., Silva, D.A., Yan, Y., Huang, X.: Force field development for cofactors in the photosystem II. Journal of Computational Chemistry 33(25), 1969–1980 (2012). https://doi.org/10.1002/jcc.23016

DOI: 10.14529/jsfi260106

pH-Dependent Conformational Analysis of Threonine Using Different Molecular Modeling Methods

Mikhail E. Kuznetsov[1], *Maria G. Khrenova*[1], *Anna M. Kulakova*[1]

Conformational landscape of flexible molecules plays an important role in their reactivity, physicochemical properties and biological functions. The article presents a comparative study of the conformational stability of three protonated forms of threonine ($Thr^{(+)}$, $Thr^{(0)}$, $Thr^{(-)}$) in aqueous solution using classical molecular dynamics (MD), umbrella sampling (US) and metadynamics (MTD) methods. It is shown that classical molecular dynamics fails to achieve ergodic sampling for $Thr^{(0)}$ and $Thr^{(-)}$ due to high rotational energy barriers around the C_α–C_β bond. The US method, despite being slightly more computationally expensive than classical MD, provides the most accurate Gibbs free energy profiles with minimal statistical error. Conventional MTD exhibits an unacceptably high confidence interval (up to 6 kcal/mol), while well-tempered MTD (WT-MTD) yields results that are quantitatively consistent with US (difference less then 0.2 kcal/mol) and an acceptable error margin (∼1 kcal/mol). It was established that at pH < 9.62 ($Thr^{(0)}$ and $Thr^{(+)}$ forms), the trans conformation is the most stable, whereas for the deprotonated $Thr^{(-)}$ form, the $gauche^{(-)}$ conformation is preferred. At the same time, the energy differences between the conformers are small (1–2 kcal/mol), and the transition barriers vary within the range of 3–12 kcal/mol.

Keywords: molecular dynamics, threonine, conformers, NAMD3.

Introduction

The physicochemical properties of molecules can be strongly influenced by their conformational composition, such as acidity/basicity [11] and circular dichroism spectra [10]. Also, in biological processes, certain conformations of active molecules are required for the process to occur. In this work, we investigated the conformations of α-amino acid L-threonine under different pH conditions. This amino acid participates in post-translational modifications and is frequently found in active sites of enzymes. Conformational lability of threonine not only contributes to the tertiary structure stability, but can directly modulate catalytic activity, substrate recognition, and allosteric regulation [16].

Experimental approaches for assessing the relative stability of conformers are applicable to solid [4], gaseous [1], and liquid states [14] of molecules; however, evaluation in aqueous solution or within a protein complex at different pH remains challenging. Numerous computational methods exist for estimating the energies of individual molecular conformations in solution employing descriptions of the system at either quantum or classical molecular mechanics level. Quantum chemistry methods assume solving the Schrödinger equation numerically, which allows us to calculate the relative internal energies of different conformations. Further vibrational analysis provides a way to estimate the Gibbs free energy formally. Molecular mechanics methods, by contrast, treat atoms in molecules as charged spheres connected by springs, with interactions defined strictly by the force field.

In order to evaluate the energy of conformers in solution with explicit solvent molecules, it is necessary to either generate and evaluate a vast number of different configurations or to simulate a trajectory using molecular dynamics with forces derived from either quantum mechanics or molecular mechanics. Quantum calculations are much more computationally intensive, and

[1]Lomonosov Moscow State University, Moscow, Russia

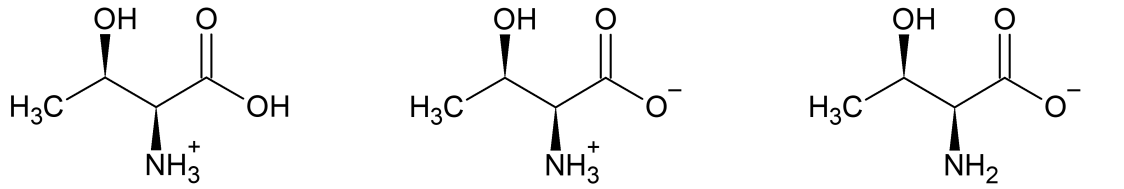
molecular mechanics can be applied to this problem because no chemical bonds are broken or formed during conformational transitions. Therefore, in this study, we perform molecular dynamics simulations using a classical force field to estimate relative Gibbs free energy of stable conformations of threonine at different pH.

The article is organized as follows. Section 1 describes the molecular systems studied and details the computational setup for classical molecular dynamics, umbrella sampling, and metadynamics simulations. Section 2 presents the results obtained by each method, including the analysis of conformational populations, Gibbs free energy profiles, and a comparison of their computational efficiency. Finally, the Conclusion summarizes the key findings, compares the accuracy of the employed techniques, and provides recommendations for choosing the optimal enhanced sampling method for conformational analysis of flexible molecules in solution.

1. Methods

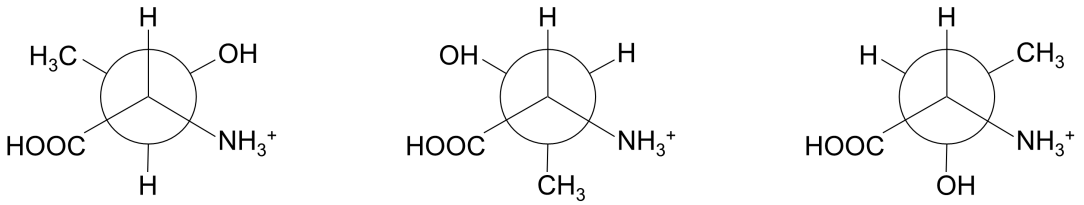
1.1. Molecular Systems

In aqueous solutions, the threonine molecule can exist in three protonation states (Fig. 1), as it contains two ionizable groups: an amino group with pKa = 9.62 and a carboxyl group with pKa = 2.11. Thus, the molecule can adopt predominately a protonated form ($Thr^{(+)}$) at pH < 2.11, a zwitterionic form ($Thr^{(0)}$) at 2.11 < pH < 9.62, and a deprotonated form ($Thr^{(-)}$) at pH > 9.62. Each of these forms has three possible conformations, depending on the value of $CC_\alpha C_\beta C_\gamma$ dihedral, that determines rotation around a single C_α–C_β bond. These conformations are called gauche$^{(-)}$, trans and gauche$^{(+)}$ (Fig. 2).



(a) protonated form $Thr^{(+)}$ (b) zwitterionic form $Thr^{(0)}$ (c) deprotonated form $Thr^{(-)}$

Figure 1. Structural formulas of L-threonine in different protonation states



(a) gauche$^{(-)}$ (b) trans (c) gauche$^{(+)}$

Figure 2. Newman projections showing conformations of the protonated form of L-threonine

Molecular model of threonine zwitterionic form at a neutral pH ($Thr^{(0)}$) was created using the Discovery Studio software and pre-optimized. Two other forms of threonine, protonated ($Thr^{(+)}$, at low pH) and deprotonated ($Thr^{(-)}$, at high pH), were obtained from the zwitterionic form by adding or removing hydrogen atoms. The optimized geometries of $Thr^{(+)}$, $Thr^{(0)}$ and $Thr^{(-)}$ were calculated using density functional theory (DFT) with the B3LYP functional [5]

and the 6-31G** basis set [6], using the conductor-like polarizable continuum model (CPCM) [2] in ORCA 5.0.4 [12].

All three optimized forms of threonine were solvated in a rectangular water box and properly neutralized. The size of the water box was selected so that the distance between any atom of threonine and the cell border exceeded 12 Å. Sodium cations (for $Thr^{(-)}$) or chlorine anions (for $Thr^{(+)}$) were added to model systems for electroneutrality. The preparation of full atomic models, as well as the visualization and analysis of structures, was carried out using the VMD program [8].

Then, the structures were minimized using the steepest descent algorithm for 1000 steps. To relax the solvation shell, 0.5 ns classical molecular dynamics (MD) simulations with fixed threonine atoms were performed for all model systems. MD calculations were carried out in the NPT ensemble with a Langevin thermostat at 298 K and a Berendsen barostat at 1 atmosphere with 1 fs integration step. CHARMM36 force field [3] was used for threonine, and TIP3P for water [9]. Minimization, solvation shell relaxation and further molecular dynamics and metadynamics calculations were carried out using the NAMD3 software [13].

1.2. Classical Molecular Dynamics Simulations

For each of the three forms of threonine 500 ns molecular dynamics trajectories were calculated in the NPT ensemble. The parameters for the molecular dynamics simulation are similar to those used for solvation shell relaxation. Trajectory analysis allowed us to estimate the populations of the $gauche^{(+)}$, $gauche^{(-)}$ and trans conformations for each form of threonine. Assuming that the resulting trajectories are ergodic, we can determine the relative energies of these conformations based on the Gibbs distribution using the following formula: $\Delta\Delta G = \Delta G_1 - \Delta G_2 = -kT \ln\left(\frac{N_1}{N_2}\right)$, where N_1 and N_2 are the number of MD frames of corresponding conformation.

1.3. Umbrella Sampling Simulations

For molecular dynamics simulations using the umbrella sampling method, four starting structures were prepared for each form of threonine. In these structures dihedral angle $H_\alpha C_\alpha C_\beta H_\beta$ was constrained to values 0°, 100°, −100° and 180°. Preparation involved the direct modification of $H_\alpha C_\alpha C_\beta H_\beta$ angle followed by energy minimization as described above.

Umbrella sampling simulations were also performed using the NAMD3 software package with the same MD parameters. The dihedral angle $H_\alpha C_\alpha C_\beta H_\beta$ was selected as a collective variable with a harmonic biasing potential (force constant k = 0.01–0.03 kcal·mol^{-1}·deg^{-2}) applied along. The center of the biasing potential was systematically shifted from −180° to 160° in 20° increments. For each window, a MD trajectory of 10 ns was computed. To enhance the overlap of probability distributions for subsequent free energy analysis, additional simulations were performed with biasing potential centered at ±10° and ±130°.

Weighted histogram analysis method [15] (WHAM) was utilized to reconstruct Gibbs free energy profiles from statistical analysis of $H_\alpha C_\alpha C_\beta H_\beta$ dihedral angle distributions in the umbrella sampling trajectories. In the WHAM program a value of 10^{-4} kcal/mol was used as the energy convergence criterion.

1.4. Metadynamics Simulations

For metadynamics (MTD) calculation the same collective variable ($H_\alpha C_\alpha C_\beta H_\beta$ dihedral angle) was used. All metadynamics (MTD) trajectories were started from value of dihedral angles – $-100°$. Gaussian biasing potentials with initial Gaussian width (at half-height) of 1° and Gaussian height of 1 kcal·mol^{-1}·deg^{-2} were applied. New Gaussians were added every 100 integration steps. A total of ten independent 10 ns metadynamics trajectories were calculated for each of three threonine forms. A bias temperature of 1700 K was selected based on preliminary simulations testing various parameter values.

2. Results and Discussion

2.1. Classical Molecular Dynamics Simulations

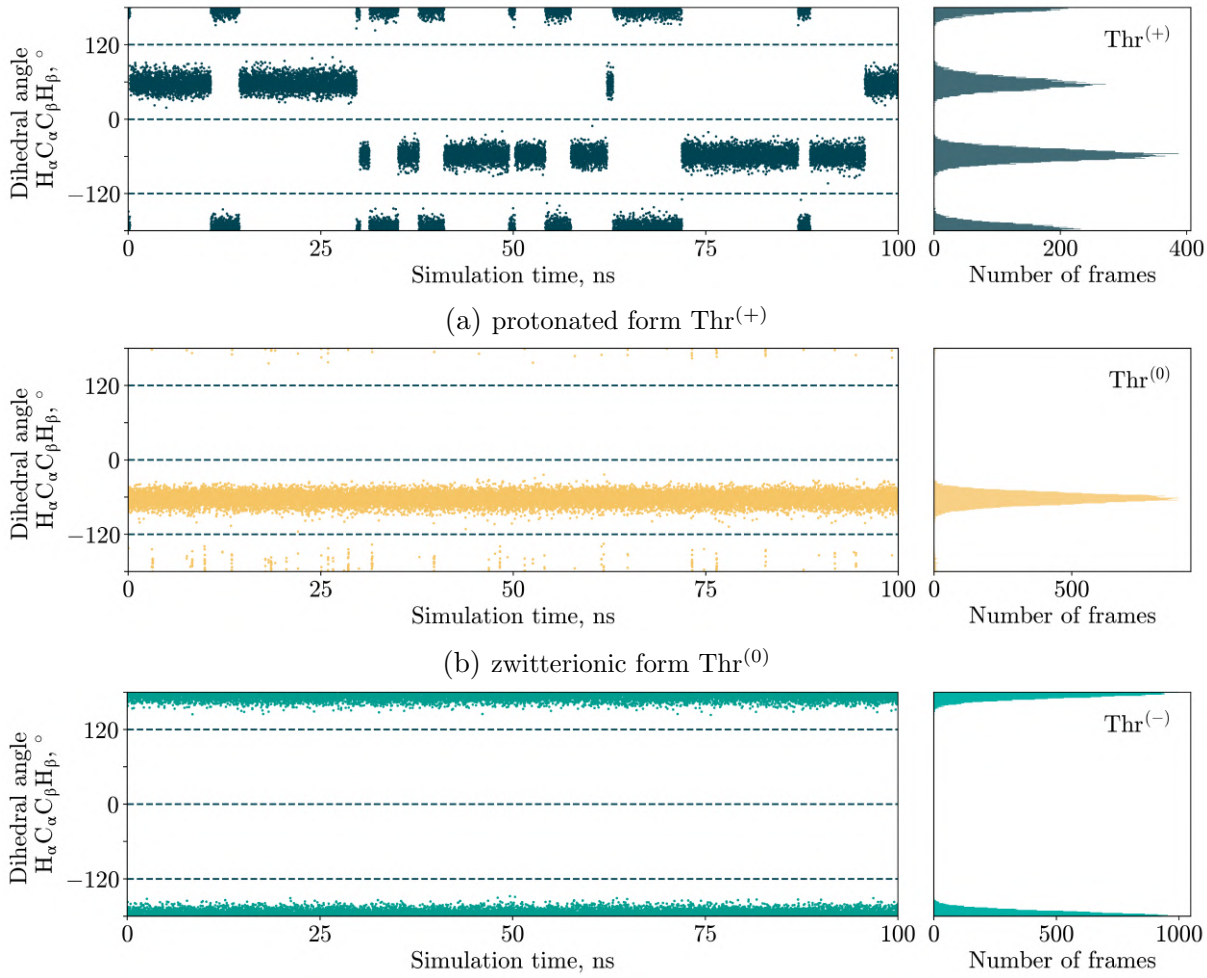
One of the simplest methods to estimate the relative stability of molecule conformation is based on the Gibbs free energy distribution. This method requires to determine gauche$^{(+)}$, gauche$^{(-)}$ and trans conformer population as the number of frames with each conformation. We calculated $H_\alpha C_\alpha C_\beta H_\beta$ dihedral angle at each frame of MD trajectory and determined the corresponding conformation from its value. $H_\alpha C_\alpha C_\beta H_\beta$ value between $-120°$ and 0° corresponds to trans conformation. If dihedral angle ranges from 0° to 120°, we are dealing with a gauche$^{(+)}$ conformation. The ranges from $-180°$ to $-120°$ and from 120° to 180° correspond to the gauche$^{(-)}$ conformation of threonine.

The $H_\alpha C_\alpha C_\beta H_\beta$ distribution was investigated for all three forms of threonine at different pH values (Fig. 3). Despite several different calculations of molecular dynamic modeling for systems with high (Thr$^{(-)}$) and neutral (Thr$^{(0)}$) pH, which started from different initial conformations, it was not possible to obtain an ergodic trajectory. The systems were unable to overcome the high rotational energy barrier. As a result, it is impossible to determine the relative populations for Thr$^{(-)}$ and Thr$^{(0)}$. In contrast, all three conformations were observed for the Thr$^{(+)}$ form, indicating lower rotational barriers. The relative Gibbs free energy was estimated using the energy of the most stable trans conformation as the zero reference point. The results are summarized in Tab. 1.

Table 1. Population and relative Gibbs free energy of Thr$^{(+)}$ conformations from classical molecular dynamics simulations

Conformation	$H_\alpha C_\alpha C_\beta H_\beta$ dihedral angle, deg	Number of frames	Population, %	$\Delta\Delta G$ kcal/mol
gauche$^{(-)}$	$(-180, -120) \cup (120, 180)$	5217	26.1	0.30
trans	$(-120, 0)$	8628	43.1	0
gauche$^{(+)}$	(0, 120)	6155	30.8	0.20

All molecular dynamics simulations were performed using NAMD3 software [13] compiled with GPU acceleration. The primary production runs were carried out using Tesla V100-SXM3-32GB. On this datacenter-grade GPU, the simulation of a small system (∼2000 atoms) achieved a performance rate of 0.000500535 seconds per step, corresponding to an impressive throughput of approximately 172.6 ns per day. For each of three systems (Thr$^{(+)}$, Thr$^{(0)}$ and Thr$^{(-)}$), five independent 100 ns trajectories were generated, resulting in a total simulated time of 1.5 μs



(a) protonated form $\text{Thr}^{(+)}$

(b) zwitterionic form $\text{Thr}^{(0)}$

(c) deprotonated form $\text{Thr}^{(-)}$

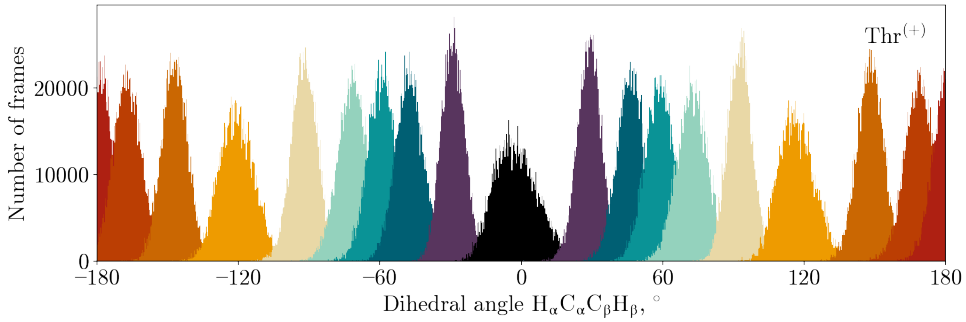
Figure 3. $H_\alpha C_\alpha C_\beta H_\beta$ dihedral angle distribution during classical MD simulations

(500 ns per system). The cumulative computational time required to complete all 15 individual runs amounted to approximately 8.7 days of continuous GPU runtime.

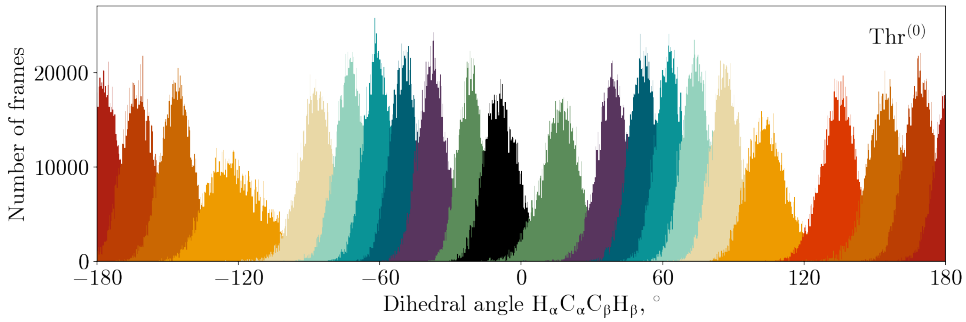
2.2. Umbrella Sampling Simulations

After performing umbrella sampling molecular dynamics with bias potential ($k = 0.01$ kcal·mol^{-1}·deg^{-2}) centered at $H_\alpha C_\alpha C_\beta H_\beta$ values from $-180°$ to $160°$ with $20°$ increments, the overlap of the $H_\alpha C_\alpha C_\beta H_\beta$ distributions was checked. To achieve optimal overlap for the $\text{Thr}^{(-)}$ distributions, two additional MD trajectories with bias potential centered at $10°$ and $-10°$ were calculated. The force constants k were increased to 0.03 kcal·mol^{-1}·deg^{-2} for MD trajectories with bias potential centered at $0°$ and $\pm 10°$. For bias potential centered at $\pm 20°$, $\pm 100°$, $\pm 120°$, and $\pm 140°$ force constants were increased to 0.02 kcal·mol^{-1}·deg^{-2}. To achieve optimal overlap for the neutral form $\text{Thr}^{(0)}$ distributions, three additional MD trajectories with bias potential centered at $130°$, $10°$ and $-10°$ were calculated. The force constants k were increased to 0.02 kcal·mol^{-1}·deg^{-2} for MD trajectories with bias potential centered at $0°$ and $\pm 10°$, $130°$. For $\text{Thr}^{(+)}$ simulations only force constants k were increased to 0.02 kcal·mol^{-1}·deg^{-2} for MD trajectories with bias potential centered at $0°$, $\pm 20°$, $\pm 100°$, $\pm 120°$, and $\pm 140°$. Since the obtained distributions of the dihedral angle overlap well with each

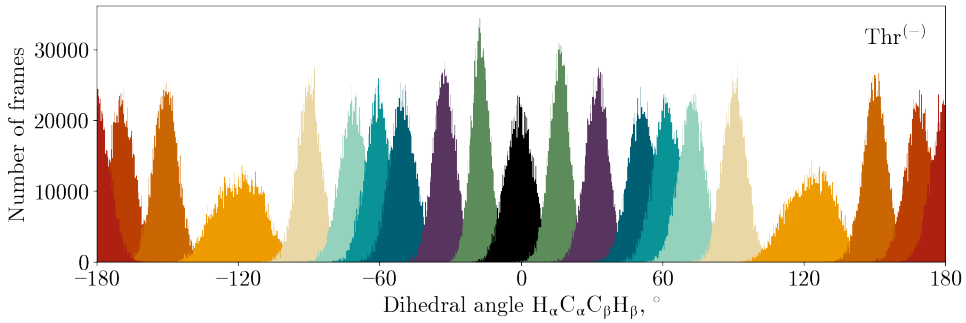
other (Fig. 4), it is possible to calculate Gibbs free energy profiles using the weighted histogram analysis method (Fig. 5).



(a) protonated form $Thr^{(+)}$

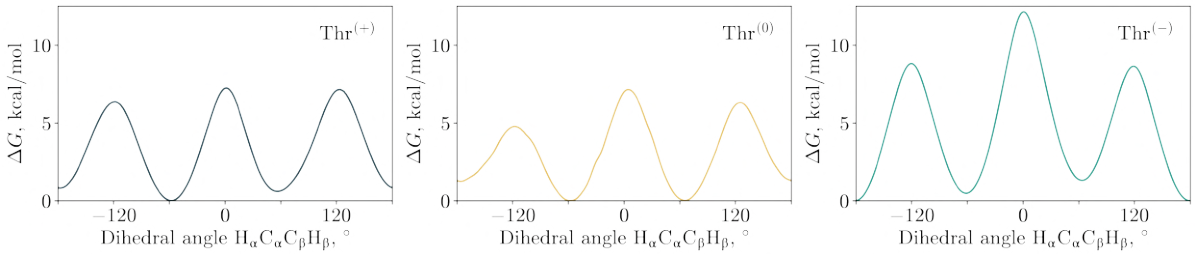


(b) zwitterionic form $Thr^{(0)}$



(c) deprotonated form $Thr^{(-)}$

Figure 4. Distribution of $H_\alpha C_\alpha C_\beta H_\beta$ dihedral angle distributions during umbrella sampling simulations



(a) protonated form $Thr^{(+)}$ (b) zwitterionic form $Thr^{(0)}$ (c) deprotonated form $Thr^{(-)}$

Figure 5. The Gibbs free energy C_α–C_β rotation profiles calculated by umbrella sampling and WHAM methods

The results obtained by the classical molecular dynamics method (Tab. 1) are in good agreement with the results of the umbrella sampling (Tab. 2) for $Thr^{(+)}$ form. The trans-conformation turned out to be the most stable, while the $gauche^{(-)}$ conformation is less stable than the $gauche^{(+)}$ one. In addition, it was observed that the rotation barriers for the deprotonated and zwitterionic forms are on average higher than for the protonated ones. This explains why conformations do not change into each other in classical molecular dynamics. The data obtained also confirm that in the $Thr^{(0)}$ form the rotation barrier from the $gauche^{(-)}$ to the trans conformation is small (2.5 kcal/mol). As a result, classical molecular dynamics quickly leaves the $gauche^{(-)}$ conformation and transforms into a trans conformation. Nevertheless, the molecules sometimes return to the $gauche^{(+)}$ conformation, although the energy barrier for this transition is lower than for the transition from trans to $gauche^{(-)}$ conformation.

Table 2. Relative Gibbs free energy of threonine conformations calculated by umbrella sampling and WHAM methods

Conformation	$H_\alpha C_\alpha C_\beta H_\beta$	$\Delta\Delta G$, kcal/mol		
	dihedral angle, deg	$Thr^{(+)}$	$Thr^{(0)}$	$Thr^{(-)}$
$gauche^{(-)}$	$(-180, -120) \cup (120, 180)$	0.8	2.6	0.0
trans	$(-120, 0)$	0.0	0.0	0.5
$gauche^{(+)}$	$(0, 120)$	0.6	1.9	1.3

For protonated and zwitterionic forms of threonine, trans conformer is the most stable. This is due to compensation of steric difficulties by electrostatic interactions between the -OH group, amino- and carboxyl groups. In the deprotonated form ($Thr^{(-)}$), the electrostatic interaction with the carboxyl group weakens, which makes the $gauche^{(-)}$ conformer more stable.

In contrast to classical molecular dynamics simulations, the umbrella sampling method imposes an external biasing potential to constrain the system along the reaction coordinate, which leads to increased computational cost per step. On the Tesla V100-SXM3-32GB GPU, umbrella sampling simulations achieved a performance of 0.000690609 seconds per step, corresponding to a throughput of 125.107 ns/day – approximately 27.5% slower than classical MD runs on the same hardware. The umbrella sampling calculations were performed separately for each protonation state of threonine. Specifically, 19 independent biased simulations were conducted for $Thr^{(+)}$, 22 for $Thr^{(0)}$, and 21 for $Thr^{(-)}$, each with a trajectory length of 10 ns. The cumulative computational time required to complete all umbrella sampling simulations was approximately 4.96 days of continuous GPU runtime.

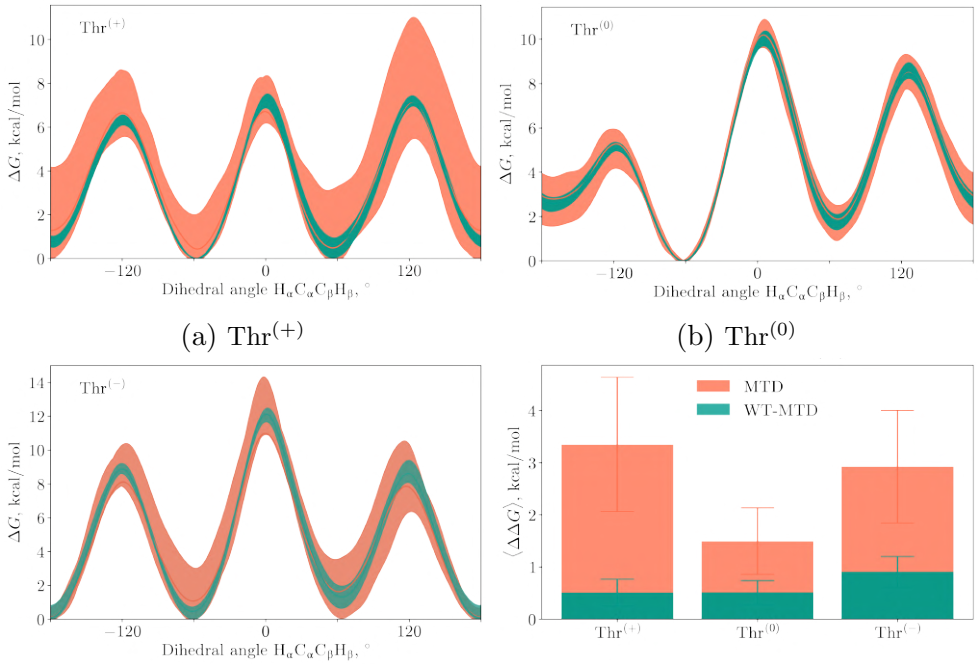
2.3. Metadynamics Simulations

Gibbs free energy profiles were calculated from the metadynamics data using the parameters of the Gaussian potentials at timestep τ (σ – Gaussian width, w – Gaussian height, θ – value of $H_\alpha C_\alpha C_\beta H_\beta$ dihedral angle) by the following formulas (ΔT – bias temperature 1700 K):

$$\Delta G_{\mathrm{MTD}}(\theta) = -w \sum_{\substack{n=1 \\ \tau = n\tau_G}} \exp\left(-\frac{(\theta - \theta(\tau))^2}{2\sigma^2}\right),$$

$$\Delta G_{\text{WT-MTD}}(\theta, \Delta T) = -\frac{T + \Delta T}{\Delta T} \sum_{\substack{n=1 \\ \tau = n\tau_G}} w(\tau) \exp\left(-\frac{(\theta - \theta(\tau))^2}{2\sigma^2}\right).$$

Data were additionally referred to the minima of Gibbs free energy. To estimate the spread of values, we used data obtained from ten different metadynamic (MTD) and ten well-tempered metadynamics (WT-MTD) trajectories. The variation in value was determined as the difference between the maximum and minimum values for each trajectory. The resulting intervals and the average profile are shown in Fig. 6.



(a) Thr$^{(+)}$ (b) Thr$^{(0)}$

(c) Thr$^{(-)}$ (d) error estimation

Figure 6. The Gibbs free energy C_α–C_β rotation profiles and error estimation calculated using metadynamics (MTD) and well-tempered metadynamics (WT-MTD)

Analysis of the obtained data shows that the metadynamics (MTD) gives significant deviations in the profile values, reaching 6 kcal/mol. At the same time, the deviations for the profiles obtained using the WT-MTD method do not exceed 1.5 kcal/mol. The energies of stable conformations for each of the forms were determined similarly to the US results using profiles obtained by the WT-MTD method (Tab. 3).

Table 3. Relative Gibbs free energy of threonine conformations calculated by well-tempered metadynamics

Conformation	$H_\alpha C_\alpha C_\beta H_\beta$ dihedral angle, deg	$\Delta\Delta G$, kcal/mol Thr$^{(+)}$	Thr$^{(0)}$	Thr$^{(-)}$
gauche$^{(-)}$	$(-180, -120) \cup (120, 180)$	0.8 ± 0.2	2.8 ± 0.2	0.0
trans	$(-120, 0)$	0.0	0.0	0.3 ± 0.2
gauche$^{(+)}$	$(0, 120)$	0.6 ± 0.3	2.0 ± 0.2	1.0 ± 0.5

The relative Gibbs free energy values of threonine conformations, obtained by umbrella sampling and well-tempered metadynamics methods, are close to each other. The difference between them is less than 0.2 kcal/mol, which is comparable to the spread of values obtained in the WT-MTD series of calculations.

Metadynamics and well-tempered metadynamics simulations exhibited nearly identical computational performance, as both methods differ from classical molecular dynamics primarily in the periodic deposition of Gaussian bias potentials every 100 simulation steps. On the Tesla V100-SXM3-32GB GPU, the average performance for both metadynamics variants was 0.000572646 seconds per step, corresponding to a throughput of 150.879 ns/day. For each protonation state of threonine, ten independent metadynamics and ten independent well-tempered metadynamics simulations were performed, each with a trajectory length of 10 ns. The cumulative computational time required to complete all metadynamics and well-tempered metadynamics simulations was approximately 3.98 days of continuous GPU runtime.

2.4. Comparison of Methods

The relative Gibbs free energies calculated in this work using classical molecular dynamics, umbrella sampling with WHAM, and metadynamics are in good agreement with each other. All methods employed in this study utilized the CHARMM force field for model description. In the literature, there are articles where the threonine molecule is described using quantum-chemical methods.

In Ref. [14], various conformations of threonine in aqueous solution were investigated using the MP2, B3LYP/6-31G**++, and M062X/6-31G**++ methods, with the IEF-PCM implicit solvent model. A total of 88 conformations were identified, and for the most stable ones, a correlation with experimental data was established. In the experiments, solid samples and solutions of threonine were studied using vibrational circular dichroism, IR, and Raman spectroscopy. The authors demonstrate that: gauche$^{(+)}$ conformation is the most stable for Thr$^{(+)}$ and Thr$^{(0)}$, trans conformation is preferred for Thr$^{(-)}$. The energy differences within the MP2 framework between conformations range from 1.5 (gauche$^{(+)}$ and gauche$^{(-)}$ for Thr$^{(+)}$) to 6.5 kcal/mol (trans and gauche$^{(+)}$ for Thr$^{(-)}$).

In Ref. [7], the conformations of the zwitterionic form of threonine were studied with explicit solvent represented by 7 water molecules using the B3LYP/6-31G*++ method. For the four most stable clusters, the authors observed agreement with experimental IR and Raman spectra. The most stable conformation for Thr$^{(0)}$ is gauche$^{(-)}$, with an energy difference of 0.3 kcal/mol relative to gauche$^{(+)}$, as for the gas phase neutral threonine [1].

Thus, the literature data formally exhibit rather poor agreement both among themselves and with the results obtained in this work, which can be attributed to several factors. Primarily, this discrepancy arises from the use of different methodologies: quantum-chemical methods, which describe electronic structure more accurately, and molecular mechanics methods, which capture dynamics and statistics on scales inaccessible to quantum approaches. Additionally, solvent representation differs, which may also contribute to the differences in energy. It should be noted that the obtained energy differences between conformations generally do not exceed 1–2 kcal/mol, which is comparable to the accuracy levels achievable by both quantum and classical methods.

Among the computational methods used in this study, classical molecular dynamics proved to be the most computationally efficient, achieving a throughput of 172.615 ns/day on one Tesla V100-SXM3-32GB GPU. However, despite its speed, classical MD is inherently limited by high energy barriers and cannot reliably determine transition free energies or complete conformational profiles, as it fails to ensure ergodic sampling of the entire conformational space within feasible simulation times. Umbrella sampling is approximately 27.5% slower than classical MD

and requires careful selection of a sufficient number of windows with adequate histogram overlap to ensure convergence. But US is relatively insensitive to simulation length per window and provides the most accurate Gibbs free energy profiles with negligible statistical error. In contrast, classical metadynamics exhibited unsatisfactory performance, yielding energy value scatter up to 6 kcal/mol – comparable to the barrier heights themselves and far exceeding the energy differences between stable conformations. Well-tempered metadynamics significantly improves upon standard MTD, achieving a throughput of 150.879 ns/day with an acceptable error margin of approximately 1 kcal/mol and quantitative agreement with US results. Despite its slightly higher computational cost, umbrella sampling remains the preferred method due to its superior accuracy and negligible statistical error compared to well-tempered metadynamics.

Conclusion

Classical molecular dynamics simulations principally allow one to determine relative Gibbs free energies of stable conformations under the assumption of ergodicity of the performed simulations, but not energy barriers of transitions between conformations. In our particular system, transition energy barriers are high and, therefore, this condition cannot be achieved even with a 50-fold increase in trajectory simulation time relative to enhanced sampling methods.

Umbrella sampling enables calculation of the entire rotational free energy profile, including both energy barriers and relative energies of stable conformations. Among the methods employed in this work, it exhibits the lowest uncertainty, with negligible statistical error. Since the conformational coordinate in this work ($H_\alpha C_\alpha C_\beta H_\beta$ dihedral angle) is well-defined, this method proves to be the least sensitive to reductions in simulation time.

Classical metadynamics yields unsatisfactory results: the uncertainty in energy values reaches up to 6 kcal/mol, which is comparable with the magnitude of the energy barriers and significantly exceeds the energy differences between stable conformations. Well-tempered metadynamics (WT-MTD) demonstrates an acceptable error margin of approximately 1 kcal/mol and shows quantitative agreement with US results; however, it requires careful parameterization of the applied bias potentials and bias temperature. Furthermore, WT-MTD should have longer trajectories and is characterized by greater uncertainty than umbrella sampling method.

US and WT-MTD methods demonstrate consistent results, showing that at pH values below 9.62, the trans conformation is the more stable form of threonine, whereas for the deprotonated form, the gauche$^{(-)}$ conformation is preferred. However, the relative energies of all conformations are close (∼1–2 kcal/mol), while the energy barriers show greater variation (∼3–12 kcal/mol).

Acknowledgments

The work was conducted under the state assignment of Lomonosov MSU 121031300176-3. The research was carried out using the equipment of the shared research facilities of HPC computing resources at Lomonosov Moscow State University [17] including Istok computing system (Agreement 075-15-2025-541).

References

1. Alonso, J.L., Pérez, C., Sanz, M.E., *et al.*: Seven conformers of L-threonine in the gas phase: a LA-MB-FTMW study. Physical Chemistry Chemical Physics 11(4), 617–627 (2009). https://doi.org/10.1039/b810940k

2. Barone, V., Cossi, M.: Quantum calculation of molecular energies and energy gradients in solution by a conductor solvent model. The Journal of Physical Chemistry A 102(11), 1995–2001 (1998). https://doi.org/10.1021/jp9716997

3. Best, R.B., Zhu, X., Shim, J., *et al.*: Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain $\chi 1$ and $\chi 2$ dihedral angles. Journal of Chemical Theory and Computation 8(9), 3257–3273 (2012). https://doi.org/10.1021/ct300400x

4. Dubey, P., Mukhopadhyay, A., Viswanathan, K.S.: Do amino acids prefer only certain backbone structures? Steering through the conformational maze of l-threonine using matrix isolation infrared spectroscopy and ab initio studies. Journal of Molecular Structure 1175, 117–129 (2019). https://doi.org/10.1016/j.molstruc.2018.07.066

5. Hamprecht, F.A., Cohen, A.J., Tozer, D.J., Handy, N.C.: Development and assessment of new exchange-correlation functionals. The Journal of Chemical Physics 109(15), 6264–6271 (1998). https://doi.org/10.1063/1.477267

6. Hariharan, P.C., Pople, J.A.: The influence of polarization functions on molecular orbital hydrogenation energies. Theoretica Chimica Acta 28(3), 213–222 (1973). https://doi.org/10.1007/BF00533485

7. Hernández, B., Pflüger, F., Adenier, A., *et al.*: Energy maps, side chain conformational flexibility, and vibrational features of polar amino acids L-serine and L-threonine in aqueous environment. The Journal of Chemical Physics 135(5), 055101 (2011). https://doi.org/10.1063/1.3617415

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

9. Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., *et al.*: Comparison of simple potential functions for simulating liquid water. The Journal of Chemical Physics 79(2), 926–935 (1983). https://doi.org/10.1063/1.445869

10. Kundrat, M.D., Autschbach, J.: Computational Modeling of the Optical Rotation of Amino Acids: A New Look at an Old Rule for pH Dependence of Optical Rotation. Journal of the American Chemical Society 130(13), 4404–4414 (2008). https://doi.org/10.1021/ja0782571

11. Lu, H., Chen, X., Zhan, C.G.: First-Principles Calculation of p Ka for Cocaine, Nicotine, Neurotransmitters, and Anilines in Aqueous Solution. The Journal of Physical Chemistry B 111(35), 10599–10605 (2007). https://doi.org/10.1021/jp072917r

12. Neese, F.: Software update: The ORCA program system–version 5.0. Wiley Interdisciplinary Reviews: Computational Molecular Science 12(5), e1606 (2022). https://doi.org/10.1002/wcms.1606

13. Phillips, J.C., Hardy, D.J., Maia, J.D.C., *et al.*: Scalable molecular dynamics on CPU and GPU architectures with NAMD. The Journal of Chemical Physics 153(4), 044130 (2020). https://doi.org/10.1063/5.0014475

14. Quesada-Moreno, M.M., Marquez-Garcia, A.A., Aviles-Moreno, J.R., *et al.*: Conformational landscape of l-threonine in neutral, acid and basic solutions from vibrational circular dichroism spectroscopy and quantum chemical calculations. Tetrahedron: Asymmetry 24(24), 1537–1547 (2013). https://doi.org/10.1016/j.tetasy.2013.09.025

15. Souaille, M., Roux, B.: Extension to the weighted histogram analysis method: combining umbrella sampling with free energy calculations. Computer Physics Communications 135(1), 40–57 (2001). https://doi.org/10.1016/S0010-4655(00)00215-0

16. Szidarovszky, T., Czakó, G., Császár, A.G.: Conformers of gaseous threonine. Molecular Physics 107(7), 761–775 (2009). https://doi.org/10.1080/00268970802616350

17. 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

DOI: 10.14529/jsfi260107

High-Throughput Computational Discovery of Anti-Coronavirus Agents in the COVID-19 Era: Crucial Insights for Combating Emerging Biogenic Threats

Dmitry S. Druzhilovskiy[1], *Dmitry A. Filimonov*[1], *Pavel V. Pogodin*[1], *Anastasia V. Rudik*[1], *Leonid A. Stolbov*[1], *Olga A. Tarasova*[1], *Alexander V. Veselovsky*[1], *Vladimir V. Poroikov*[1]

In May 2020, the Joint European Disruptive Initiative (JEDI) launched the "Billion Molecules against COVID 19" challenge – an extensive open science effort aimed at identifying small molecule inhibitors of SARS-CoV-2 and related human receptors. Our research group joined this initiative among 130 international teams, focusing on the in silico screening for potential anti coronavirus agents that target three viral proteins and one human receptor. The screening campaign covered more than one billion synthetically accessible structures, including approved pharmaceuticals. By July 17, 2020, our team submitted a subset of 10000 prioritized compounds to the organizers for expert evaluation. The results from our selection, together with those from 19 other participating teams, contributed to a pool of approximately 1000 molecules selected for chemical synthesis and bioactivity testing. In total, 878 compounds were successfully synthesized and evaluated for inhibitory activity against various SARS-CoV-2 targets as well as the human serine protease TMPRSS2. Ultimately, 27 compounds – including one proposed by our group – demonstrated measurable anti coronavirus activity. The collective outcomes of these collaborative efforts were reported in the "Molecular Informatics" journal in 2024. In the present study, we summarize our participation in the JEDI challenge and discuss broader methodological and organizational considerations critical for improving the efficiency of rapid scientific responses to future emerging biological threats.

Keywords: COVID-19, JEDI COVID-19 challenge, anti-coronavirus agents, in silico screening, future biogenic threats, effective response.

Introduction

In December 2019, physicians from several hospitals in Wuhan notified the local Center for Disease Control and Prevention of pneumonia cases of unknown etiology. Subsequently, Vision Medicals (Guangzhou) confirmed a novel coronavirus in specimens from Wuhan Central Hospital. In January 2020, the initial SARS-CoV-2 genome sequence was deposited in GISAID [33]. Concurrently, infections were identified beyond China, in Thailand, Japan, and the United States. On January 30, the World Health Organization (WHO) proclaimed a Public Health Emergency of International Concern (PHEIC), reporting 9800 cases and 213 fatalities. The disease received its official designation, "COVID-19", in February 2020; on March 11, WHO classified it as a pandemic, with over 118000 cases in 114 countries and 4291 deaths.

Mitigating the COVID-19 pandemic necessitated a coordinated international effort, comprising border closures and lockdowns, reconfiguration of healthcare systems, rigorous sanitary protocols, advancement of diagnostic tools and vaccines, and identification of new therapeutic modalities [64]. After release of the viral genome sequence, laboratories rapidly designed RT PCR assays targeting SARS-CoV-2, with WHO issuing technical guidance on detection, testing, and case management by January 10, 2020 [48]. National public health laboratories and manufacturers produced and scaled PCR based tests during January–February 2020; by March 2020, many

[1]Institute of Biomedical Chemistry (IBMC), Moscow, Russian Federation

countries had authorized emergency use diagnostic assays. Following viral genome publication, multiple groups initiated vaccine design programs using mRNA, viral vector, inactivated, and protein subunit platforms. In March 2020, at least four vaccine candidates entered first in human (phase I) clinical trials, marking an unprecedented acceleration compared to traditional vaccine timelines [48].

In early 2020, before the authorization of any pharmaceutical agents specifically indicated for COVID-19 treatment, several small-molecule drugs previously employed against other viral infections were proposed for repurposing against SARS-CoV-2. These included lopinavir & ritonavir (in combination), interferon (type I; mainly IFN α/β), ribavirin, chloroquine, hydroxychloroquine, favipiravir, remdesivir, and ivermectin. Except for remdesivir, most of these compounds were subsequently shown to lack clinical efficacy, and some (e.g., chloroquine and hydroxychloroquine) were associated with notable adverse effects. Nevertheless, during the initial stage of the pandemic, they comprised the primary therapeutic approaches available for managing SARS-CoV-2 infection [51].

It became evident that the discovery of new anti coronavirus drugs was urgently needed. In addition to extensive studies conducted by major pharmaceutical companies, numerous academic groups also sought to identify novel promising candidates [37].

In response to the unmet need for effective COVID-19 therapies, the Joint European Disruptive Initiative (JEDI) launched the "Billion Molecules against COVID-19 Grand Challenge" on April 23, 2020 [40]. Following an invitation from our European colleagues, including Prof. Dr. Alexandre Varnek (University of Strasbourg), we chose to participate in this collaborative project.

The terms and conditions of this study were defined as follows:

(1) to conduct virtual screening of more than one billion synthetically accessible compounds, including clinically approved drugs;

(2) to assess their potential interactions with at least three molecular targets implicated in antiviral activity against coronaviruses;

(3) to employ three independent computational methodologies; and

(4) to complete the study within the period of May–June 2020.

The set of targets available for subsequent experimental validation comprised the following proteins: 3C-like protease (3CLpro), papain-like protease (PLpro), transmembrane serine protease 2 (TMPRSS2), spike glycoprotein (S), nucleocapsid protein (N), and RNA-dependent RNA polymerase (RdRp).

Participants were required to submit compound lists containing 10000 molecules for a minimum of three protein targets (amounting to a total of 30000 compounds) using the designated *.csv template. Additionally, a detailed report describing the applied computational methods and performance metrics, selected targets, compound libraries, and obtained results was to be submitted in the *.docx template provided by the organizers.

The consolidated results obtained by participants of the JEDI COVID-19 Challenge are reported in the joint publication [67]. In summary, 31 research teams proposed a total of 639024 candidate compounds with putative activity against the aforementioned anti-coronavirus targets. Among these, 878 compounds were synthesized and subjected to experimental evaluation in the corresponding biological assays. As a result, 27 compounds demonstrating weak inhibitory or binding activity were identified through binding, cleavage, and/or viral suppression assays. It was concluded that the open-science framework adopted in the JEDI COVID-19

Challenge made a measurable contribution to the collective knowledge base, thereby facilitating future drug discovery initiatives aimed at the development of improved therapeutic agents against SARS-CoV-2 [67].

In this manuscript, we describe our methodology, developed within the framework of the JEDI COVID-19 Challenge, for the selection of putative anti coronavirus agents from a chemical space comprising billions of compounds, as well as the results obtained and the insights gained, which may contribute to future strategies for mitigating the emerging biogenic threats.

The article is organized as follows. The Introduction describes the context of the onset of the COVID-19 pandemic and the background of launching the open JEDI "Billion Molecules against COVID-19" initiative, which aimed to urgently search for new antiviral drugs. In Section 1 we detail the proposed computational workflow, which includes three sequential stages: initial hit selection based on structural similarity (MNA and QNA), filtering and ranking using machine learning algorithms (PASS and GUSAR software), and verification of the results via molecular docking. This section also describes the selection process for viral targets and the preparation of a consolidated library of over one billion synthetically accessible compounds. Section 2 provides an analysis of the data obtained at each stage of the virtual screening for potential inhibitors of the 3CLpro and PLpro proteases, RdRp polymerase, and the TMPRSS2 receptor. The authors also share key takeaways from their participation in the JEDI challenge, analyzing the reasons for the low efficiency of mass screening under conditions of initial scarcity and noisiness of the training data. Section 2.2 summarizes the overall results of the study, emphasizing that despite the limitations and inaccuracy of early data, the application of machine learning methods and molecular modeling drastically reduced the material, financial, and time costs of experimental validation, narrowing down the screening of billions of molecules to the synthesis of just 878 compounds.

1. Materials and Methods

Considering the incomplete and occasionally contradictory information available on May 4, 2020, regarding the SARS-CoV-2 virus and its interactions with host cells, we developed a systematic approach for the virtual screening of potential anti-coronavirus hits from extensive chemical libraries.

The proposed workflow comprised three successive stages:

- Initial hit selection. Potential hits were identified among more than one billion compounds by assessing structural similarity to reference molecules with experimentally confirmed anti-coronavirus activity.
- Filtering and ranking. The selected compounds were further filtered and prioritized using machine learning algorithms implemented in the PASS and GUSAR (see Sections 1.2 and 1.3) software packages. The corresponding training sets were continuously expanded and refined throughout the project to improve predictive performance.
- Verification by molecular modeling. Representative compounds from the prioritized set were analyzed using molecular modeling methods to evaluate their potential interactions with SARS-CoV-2 targets. A schematic representation of the overall workflow for selecting compounds with potential anti-coronavirus activity is presented in Fig. 1.

The following sections describe four distinct methods applied across the three stages.

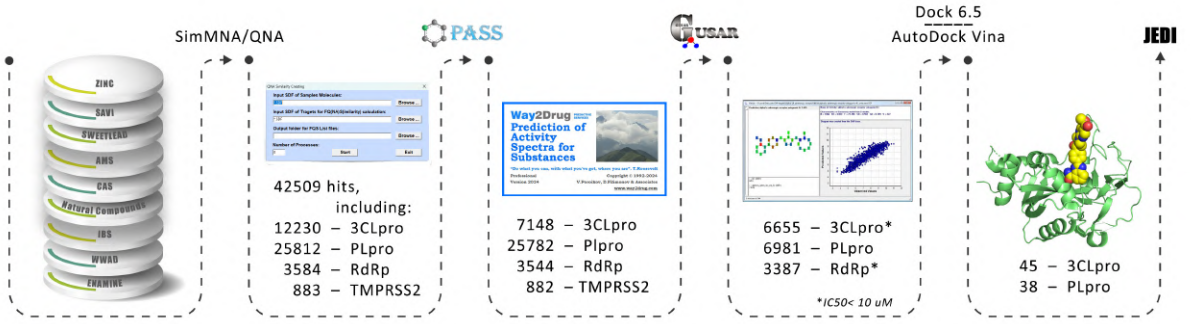


Figure 1. General workflow and results of selection of anti-coronavirus hits

1.1. Similarity Assessment

"Similar molecules tend to exhibit similar biological activities" [45]. Although this principle is occasionally violated in the presence of the so called activity cliffs [18], it remains a cornerstone of medicinal chemistry and is extensively employed in the design of structural analogues presumed to interact with the same target or to elicit comparable pharmacological responses [78]. Furthermore, it constitutes the method of choice in studies involving novel pharmacological targets, particularly when the number of known ligands is insufficient for the construction of a pharmacophore model or the development of a (Quantitative) Structure-Activity Relationship ((Q)SAR) regression or classification models.

At present, no universal method exists for assessing the similarity between molecules belonging to different chemical classes and exhibiting diverse biological activities [6]. Within the framework of the JEDI COVID-19 Challenge, molecular similarity was evaluated using our original molecular descriptors – Multilevel Neighborhoods of Atoms (MNA) [24] and Quantitative Neighborhoods of Atoms (QNA) [25]. These descriptors have demonstrated high efficacy in the analysis of structure-activity relationships for heterogeneous datasets, which is fully consistent with the objectives of the present study, namely the virtual screening of compounds with potential anti-coronavirus activity among more than one billion molecules. Earlier, we examined the applicability of these descriptors for activity prediction based on similarity across a dataset comprising 16770 inhibitors of HIV-1 protease, reverse transcriptase, and integrase [71], and identified both the advantages and limitations of this approach [20].

To conduct the similarity search and identify compounds exhibiting the desired biological activity among a library exceeding one billion molecules, we selected "reference substances", defined as the most potent inhibitors of the four investigated targets reported as of June 2020. These reference substances were subsequently employed as query molecules.

3CLpro. The five most active compounds were collected from various sources and evaluated using different experimental protocols. The activities of GC376, Tideglusib, 11b, and TZDZ-8 were obtained from the respective original publications [16, 41, 49], while MAT-POS-916a2c5a-1 was selected from the PostEra resource [59]. All five compounds were tested against the recombinant SARS-CoV-2 main protease (3CLpro) and exhibited low micromolar inhibitory activities.

PLpro. The PLpro inhibitors 6-thioguanine, GRL0617, 679818, and psoralidin were selected from the corresponding original reports [13, 34, 62] as the most potent inhibitors of the SARS-CoV papain-like protease.

RdRp. The selection of the most active compounds was performed using data from the Stanford Coronavirus Antiviral Research Database [70]. Three compounds were identified as

lead candidates: PubChem CID 44468216 (GS 441524), PubChem CID 121304016 (Remdesivir), and ChEMBL ID CHEMBL2178720 (β-D-N4-Hydroxycytidine). The antiviral activities of GS 441524 and Remdesivir were documented in multiple preprints [9, 17, 61, 63, 66, 77], whereas evidence for β-D-N4-Hydroxycytidine originated from a single study [22]. All three compounds exhibited submicromolar half maximal effective concentrations (EC50) in assays employing SARS-CoV-2 and human cell lines. Notably, both Remdesivir and GS 441524 were also reported to suppress viral RNA expression, corroborating their potent antiviral properties.

TMPRSS2. The selection of the most potent compounds targeting TMPRSS2 was conducted using data retrieved from the ChEMBL database [12]. Three chemical entities exhibiting submicromolar Ki values were identified: CHEMBL1809250, CHEMBL1229259, and CHEMBL1809251. According to the assay description provided in ChEMBL, these compounds were evaluated against the recombinant catalytic domain of TMPRSS2 expressed in Escherichia coli, employing D-cyclohexylalanine-Pro-Arg-AMC as a fluorogenic substrate and fluorescence plate reader analysis for activity quantification. The experimental results were originally reported in reference [68].

In addition to the reference compounds reported in available publications and databases, we also included ligands complexed with SARS-CoV-2 proteins from the Protein Data Bank (PDB) [60] in the similarity search. Ligands from the following six complexes were used: 6LU7 (PRD_002214), 7BRP (HU5), 7BRR (K36), 6Y2G (O6K), 6W63 (X77), and 7BV2 (F86). Ligand IDs are given in parentheses.

1.2. Machine Learning with PASS Computer Program

PASS (Prediction of Activity Spectra for Substances, version 2019) is a computational system developed to predict more than five thousand biological activities with an average Independent Accuracy of Prediction (IAP), quantified as the Receiver Operating Characteristic Area Under the Curve (ROC AUC), of approximately 0.97. These predictions are derived solely from the structural formula of a drug-like compound [57]. The development of PASS began in the late 1980s [10], and over the subsequent three decades, the training datasets have been progressively refined, the range of predictable biological activities expanded, and extensive benchmarking of chemical descriptors and machine learning algorithms conducted [27, 58].

PASS 2019 utilizes structure-activity relationship (SAR) analysis for 1025468 biologically active compounds employing MNA descriptors in combination with a modified nave Bayes classifier [27]. This methodology enables accurate SAR characterization for compounds within the training set and demonstrates sufficient generalizability to provide reliable predictions of biological activity profiles for novel chemical entities, even in the presence of incomplete training data [58].

For each compound under prediction, PASS calculates two probabilities: Pa, representing the likelihood of belonging to the class of "actives", and Pi, the likelihood of belonging to the class of "inactives". By default, compounds for which Pa exceeds Pi are classified as "active".

The performance of PASS exceeds that of other established methods for predicting biological activity profiles, as demonstrated by comparative computational studies [3, 32, 52]. The Professional version of PASS provides functionality for constructing novel training sets, retraining the program to generate an updated SAR knowledge base, and assessing predictive accuracy and reliability through leave-one-out and 20-fold cross-validation procedures, respectively.

Within the framework of the present study, a specialized training set was developed by systematically compiling data from both freely accessible and commercial databases [15], as well as from a wide range of relevant scientific publications. Compounds exhibiting IC50 values below 10 μM were designated as active to serve as the selection threshold. To enhance the representativeness of the chemical space, all newly available information on the structures and activities of anti-coronaviral agents was incorporated into the PASS 2019 training set. Upon completion of the training and validation procedures, an updated SAR knowledge base with the following characteristics was obtained: 1025630 substances, 106828 unique MNA descriptors, 8 selected activities; average IAP equals to 0.9138.

Characteristics of SAR models for each particular activity are given in the Tab. 1. Here, N denotes the number of compounds in the training set that exhibit the given activity; *IAP* (Invariant Accuracy of Prediction) is the predictive performance estimated by leave-one-out cross-validation and is equivalent to the AUC ROC; *20-F IAP* denotes the IAP estimated using 20-fold cross-validation.

Table 1. Characteristics of SAR models for different anti-coronavirus activities

N	**IAP**	**20-F IAP**	**Activity Type**
62	0.9585	0.9587	3C-Like Protease (SARS-CoV) Inhibitors
18	0.9908	0.9909	3C-Like Protease (SARS-CoV-2) Inhibitors
6	0.8296	0.8320	Papain-Like Protease (SARS-CoV-2) Inhibitors
3	0.9970	0.9980	RNA-Directed RNA Polymerase (SARS-CoV-2) Inhibitors
808	0.7535	0.7535	SARS-CoV-2 infection reduction in cell-based assay
5	0.9678	0.9684	SARS-CoV-2 viral Entry Inhibitors
371	0.8129	0.8147	Spike Glycoprotein (S) (SARS-CoV-2)/ACE2 Interaction Inhibitors
3	1.0000	1.0000	Transmembrane Protease Serine 2 (TMPRSS2) Inhibitors

As evident from the data presented above, the accuracy (leave-one-out cross-validation) and predictive performance (20-fold cross-validation) of the developed specialized version of PASS are sufficiently high to support its practical application. This conclusion was also supported by the results of prediction for the reference substances (remdesivir, umifenovir, etc.).

1.3. Machine Learning with GUSAR Computer Program

GUSAR (General Unrestricted Structure-Activity Relationships) is a software for the quantitative structure-activity relationship (QSAR) analysis based on compound structural formulas and corresponding activity or property data, and for predicting activities or properties of novel compounds. It enables the development of (Q)SAR models for organic molecules from both homogeneous and heterogeneous chemical classes. GUSAR employs QNA descriptors, which

represent a molecule as a set of tuples of real values P, Q. The P and Q values are calculated for each atom in a molecule using the connectivity matrix together with values of the standard ionization potential and electron affinities of its constituent atoms.

The current version of GUSAR additionally incorporates selected physicochemical descriptors and biological descriptors derived from Pa-Pi predictions produced by the PASS algorithm. The underlying modeling procedure is based on the self-consistent regression (SCR) method [28], which is used in combination with nearest-neighbor evaluation and a radial basis function artificial neural network (RBF ANN) constructed from SCR outputs to obtain a multiple-model consensus [81]. A comparative study of the first GUSAR version with widely used methodologies such as CoMFA, CoMSIA, GOLPE/GRID, and HQSAR demonstrated clear advantages of this approach for QSAR model construction [25]. In the Collaborative Modeling Project for Androgen Receptor Activity (CoMPARA), GUSAR predictions were also found to be robust and accurate, further supporting the reliability of the method [50].

Within the framework of the JEDI COVID-19 Challenge, QSAR models were developed using the GUSAR software for three viral targets: 3CLpro, PLpro, and RdRp. For 3CLpro and RdRp, regression models were obtained with the following characteristics: 3CLpro, N = 45, $R^2 = 0.981$, $Q^2 = 0.836$, F = 7.220, SD = 0.410, V = 11; RdRp, N = 888, $R^2 = 0.875$, $Q^2 = 0.818$, F = 10.570, SD = 0.501, V = 213. Here, N is the number of compounds in the training set; R^2 is the determination coefficient; Q^2 is the leave-one-out cross-validated determination coefficient; SD is the standard deviation; and V is the number of variables.

Due to the small size of the training set, only a classification model was built for PLpro, with the following characteristics: N = 16, Sens = 1.000, Spec = 0.700, BA = 0.850, V = 3. Here, N is the number of compounds in the training set; Sens denotes sensitivity; Spec, specificity; BA, balanced accuracy; and V is the number of variables.

1.4. Molecular Modeling for Verification of Selection

The final validation of a selected subset of hits was conducted using molecular docking to predict ligand binding poses and estimate binding affinities based on docking scores. Docking calculations were performed with DOCK 6.5 [76] and AutoDock Vina [4]. The scoring function cutoffs for compound selection were set to −65 kcal/mol for DOCK 6.5 and −8.0 kcal/mol for AutoDock Vina, respectively. The resulting docking poses were visually inspected to evaluate their compatibility with the subpockets within the protease active sites and to analyze key binding interactions, including hydrogen bonding, steric complementarity, and electrostatic fit.

1.5. Targets Selection

Four out of the six targets proposed by the organizers of the JEDI COVID-19 Challenge against COVID-19 were selected based on the following criteria: (1) the critical role of the target in coronavirus entry into host cells or viral replication; (2) the availability of reference compounds enabling activity assessment by similarity; (3) the availability of data on drug-like compounds suitable for constructing (Q)SAR training sets; and (4) the presence of a resolved three-dimensional structure in the Protein Data Bank. Targets satisfying at least three of these four criteria were deemed suitable for subsequent analysis.

3-chymotrypsin-like protease (3CLpro/Mpro). The 3C-like protease (3CLpro), also referred to as nonstructural protein 5 (Nsp5), is first auto-catalytically cleaved from the viral

polyprotein to yield the mature enzyme. Subsequently, it mediates proteolytic processing of downstream nonstructural proteins at 11 distinct cleavage sites, thereby releasing Nsp4-Nsp16. Numerous three-dimensional structures of this protease are currently available in the Protein Data Bank (PDB). At the initial stage of this study, all available crystallographic structures of 3CLpro were retrieved from the RCSB PDB and analyzed to identify key structural features involved in inhibitor binding. For molecular docking investigations, the crystal structure with PDB ID 6LU7, complexed with the peptide-like inhibitor N3, was selected as the target. This structure was chosen because it contains one of the largest inhibitors, which closely mimics the natural substrate of the protease. Protein structure preparation was carried out using the SYBYL-X 8.1 software suite [73] and involved the following steps: (a) removal of the co-crystallized inhibitor, water molecules, and ions; (b) addition of hydrogen atoms; (c) assignment of atomic charges using the Gasteiger—Hückel method; and (d) energy minimization in vacuum employing the Tripos force field.

Papain-like proteinase (PLpro). PLpro cleaves the N terminal region of the replicase polyprotein to release Nsp1, Nsp2, and Nsp3, an essential step in assembling the viral replicase complex and enabling efficient viral replication. The crystal structure 6WUU was selected as the target for molecular docking, and it was prepared using the same protocol previously applied for 3CLpro.

RNA-dependent RNA polymerase (RdRp). Nsp12, a highly conserved protein among coronaviruses, serves as the RNA-dependent RNA polymerase (RdRp) and constitutes the central catalytic component of the viral replication-transcription complex.

Transmembrane peptidase serine 2 (TMPRSS2). TMPRSS2 mediates proteolytic cleavage of the SARS-CoV-2 spike protein, thereby enhancing viral infectivity. However, the three-dimensional structure of TMPRSS2 has not yet been experimentally determined.

1.6. Libraries

Nine libraries were used for preparation of "billion compounds" set for virtual screening.

Library 1. ZINC [82] included 920839556 structures. Over 750 million compounds were potentially purchasable.

Library 2. SAVI (Synthetically Accessible Virtual Inventory) [65] included about 1.75 billion proposed products structures with reactions generated in the first full enumeration of the SAVI project. Number of the synthesizable compounds is about 976 million (621 million without stereoisomers).

Library 3. SWEETLEAD [72] included 9127 structures (7636 without stereoisomers).

Library 4. AMS (Aldrich Market Select) [1] included 4787319 structures, with samples available in stock of Merck KGaA collected in the framework of the program "Antimicrobial Stewardship".

Library 5. Antiviral CAS dataset [2] included 49408 structures of antiviral compounds and their analogs collected by Chemical Abstracts Services.

Library 6. Natural Compounds Set included 118894 structures of natural compounds collected by our team from several publicly available databases: ChEBI [11], NANP DB [55], NPASS [54], NuBBE DB [56], UNPD [75].

Library 7. IBS Natural Compounds Set [38] included 69034 structures of natural compounds, their analogs and derivatives, which samples could be purchased from InterBioScreen Ltd.

Library 8. WWAD (World Wide Approved Drugs) [79] included 4108 structures of the launched drugs prepared by our team in the framework of our project dedicated to drug repurposing.

Library 9. ENAMINE in-stock compounds [23] included 1.94 million structures that could be obtained from Enamine Ltd.

All data were subjected to pre-processing and standardization procedures using ChemAxon JChem Instant software and the in-house developed program ClearSDF, in full accordance with current methodological recommendations [29–31]. Following the data curation and prioritization of compounds with a higher probability of experimental availability or synthetic feasibility, a final library comprising 1082000000 structures was prepared and subsequently analyzed using the computational approaches described above.

A unified dataset of curated chemical compound structures was constructed by integrating data from nine independent sources, followed by the removal of duplicate entries. Establishing structural uniqueness requires evaluation of molecular graph isomorphism, a NP-complete problem, making pairwise comparisons among approximately two billion structures computationally infeasible. To overcome this, QNA descriptors were calculated for each structure, and a single real-valued parameter – the Q-index (the sum of atomic Q values) – was assigned to each compound. Using the quicksort algorithm, compounds were ordered by increasing Q-index. Only those compounds with Q-index differences below 10^{-9} were subsequently tested for isomorphism. This approach reduced the number of required graph isomorphism checks from N(N-1)/2N(N-1)/2 to a near-linear-scale computation.

2. Results and Discussion

2.1. Selection of Potential Anti-coronavirus Agents

As illustrated in Fig. 1, based on the evaluation of MNA and QNA similarity for the reference compounds described in Section 1, a total of 42509 hits were identified. These included 12230 putative 3CLpro inhibitors, 25812 putative PLpro inhibitors, 3584 putative RdRp inhibitors, and 883 putative TMPRSS2 inhibitors. Subsequent selection was performed using PASS predictions, yielding 7148 potential 3CLpro inhibitors, 25782 potential PLpro inhibitors, 3544 potential RdRp inhibitors, and 882 potential TMPRSS2 inhibitors.

Owing to the absence of a resolved spatial structure for transmembrane peptidase serine 2 (TMPRSS2), and the inability to construct both regression and classification models for its inhibitors using the GUSAR platform, this stage of selection was considered final for that molecular target.

As the probability of TMPRSS2 inhibitory activity predicted by PASS is below 0.4, it may be inferred that the likelihood of detecting this activity experimentally is relatively low. Nevertheless, should the prediction be experimentally confirmed, the identified compound could serve as a lead structure representing a novel chemical class associated with the investigated biological activity (New Chemical Entity) [26].

Subsequent analyses were performed for the remaining three targets employing both regression and classification (Q)SAR models developed using the GUSAR platform. As a result, 6655 potential 3CLpro inhibitors and 3387 potential RdRp inhibitors with estimated IC_{50} values below 10 μM were identified. In the case of PLpro, the classification models yielded 6981 hits predicted to belong to the “active” class.

For RdRp, this stage represented the final step of the selection process. The top five predicted RdRp inhibitors demonstrated markedly high Pa-Pi values, suggesting a high probability of experimental confirmation. Nevertheless, these compounds exhibited strong structural similarity to approved antiviral agents, with four of the five being listed in the CAS antiviral database.

The compounds included in the RdRp training set were nucleotide analogues. Their proposed mechanism of inhibition involves incorporation into the growing RNA chain, thereby terminating RNA elongation by preventing the subsequent addition of nucleotides. The molecular docking programs employed in this study are not appropriate for accurately predicting binding poses or estimating the binding affinities of inhibitors that act through such a mechanism. Consequently, the molecular docking approach was not utilized at the final stage of compound selection for RdRp.

For compounds predicted to possess 3CLpro and PLpro inhibitory activity, additional molecular docking studies were conducted as described above. This analysis resulted in the identification of 45 potential 3CLpro inhibitors and 38 potential PLpro inhibitors. Notably, for these compounds, the computational predictions derived from similarity assessment and Tab. 2 summarizes the three compounds predicted to be the most probable inhibitors for each target.

In consideration of the requirements of the JEDI COVID-19 Challenge (10000 hits per target), the highest-scoring compounds described above were supplemented with additional compounds of lower scores, where such data were available.

2.2. Lessons Learned from Our Participation in the JEDI COVID-19 Challenge

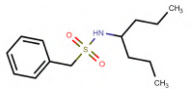
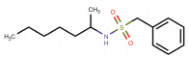
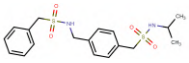
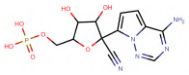
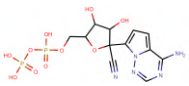
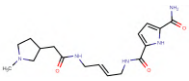
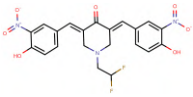
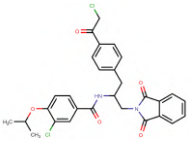
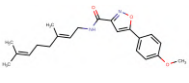
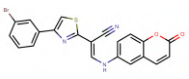
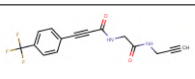
According to the expert evaluation conducted by the JEDI COVID-19 Challenge organizers, thirty-six compounds proposed by our research group were synthesized and assessed in anti-coronavirus bioassays. Among these, one compound exhibited inhibitory activity against PLpro, with an IC_{50} value in the micromolar range [67].

In accordance with the initial guidelines provided by the JEDI COVID-19 Challenge organizers, several approved drugs were also included among the compounds proposed as potential anti-coronavirus agents. However, their inhibitory activity was not examined within the framework of the JEDI COVID-19 Challenge, as multiple studies had already reported the results of screening approved drug libraries against SARS-CoV-2 targets [39, 41, 63, 74].

Several of our predictions concerning potential anti-coronavirus agents among approved drugs were subsequently corroborated by independent investigations. Specifically, inhibition of 3CLpro was demonstrated for narlaprevir, boceprevir, telaprevir [5, 44], carmofur, and disulfiram [46]; inhibition of RdRp for sofosbuvir and gemcitabine [80]; and inhibition of PLpro for dihydroquercetin and hesperetin [42].

The requirement established by the JEDI COVID-19 Challenge organizers to submit 10000 compounds predicted to interact with each of the three SARS-CoV-2 viral targets appears to have been overly stringent. Despite computational screening of approximately one billion synthesizable structures and the application of three independent in silico approaches, our study did not yield that number of hits supported by sufficient evidence. If other teams in the JEDI COVID-19 Challenge faced similar difficulties, it is plausible that the organizers received a large volume of low-confidence data, necessitating substantial additional effort for its further evaluation.

Table 2. The most probable hits for the analyzed targets

Transmembrane peptidase serine 2 (TMPRSS2)		
Name (Database)	Structure	Pa-Pi
ZINC001252905755 (ZINC)		0.315
Z355234742 (Enamine)		0.286
Z198103156 (Enamine)		0.280
RNA-dependent RNA polymerase		
Name (Database)	**Structure**	**Pa-Pi**
BRDWIEOJOWJCLU-LTGWCKQJSA-N (AMS)		0.977
1911578-74-9 (CAS antiviral DB)		0.977
1911578-77-2 (CAS antiviral DB)		0.973
3-chymotrypsin-like protease		
Name (Database)	Structure	Scoring Function
CHUUJOGSXZEWIU-NSCUHMNNSA-N (AMS)		−66.2 (Dock 6.5)
SPSIFTRUXBQBRF-YOENDLTHSA-N (AMS)		−8.4 (AutoDock Vina)
SXCFTBTXHZXEIN-NRFANRHFSA-N (AMS)		−8.6 (AutoDock Vina)
Papain-like proteinase		
Name (Database)	Structure	Scoring Function
NIKRPEWINGWQFH-FOWTUZBSSA-N (AMS)		−8.2 (AutoDock Vina)
DUJJXYLPLPJJQH-RVDMUPIBSA-N (AMS)		−9.7 (AutoDock Vina)
ORPOQLQFKDBKIH-UHFFFAOYSA-N (AMS)		−8.2 (AutoDock Vina)

Another plausible inference from this observation is that the chemical spaces of known antiviral agents and those of currently available synthesizable compounds differ substantially. This finding aligns with earlier work published in 2016 [43], which showed that antiviral compounds from ChEMBL cluster within specific regions of chemical space, with distinct antiviral classes occupying "privileged" zones on GTM (Generative Topographic Mapping) maps that diverge

from the more general drug-like chemical space. Subsequent studies similarly demonstrated that a curated and similarity expanded set of SARS-CoV-2 active compounds occupies a region of chemical space that extends well beyond that of a large commercially available coronavirus focused library (over 20000 molecules) and exhibits distinct scaffold distributions [7].

Currently, the clinically validated direct-acting small-molecule antivirals approved for the treatment of SARS-CoV-2 infection (as opposed to in vitro activity only) include remdesivir, molnupiravir, and the combination nirmatrelvir/ritonavir (Paxlovid). Remdesivir (GS 5734), originally developed by Gilead Sciences, is a nucleotide analogue initially designed for the treatment of severe RNA virus infections with pandemic potential, with an early focus on the Ebola virus [21]. Molnupiravir (EIDD 2801, MK 4482) was originally conceived as an orally bioavailable, broad-spectrum nucleoside analogue for the treatment of alphavirus and influenza virus infections, reflecting its early development as an anti influenza and anti respiratory RNA virus candidate [19]. The combination of nirmatrelvir with ritonavir (Paxlovid) was developed and authorized by Pfizer in record time; notably, nirmatrelvir was not synthesized de novo but rather derived from the earlier optimized SARS-CoV protease inhibitor PF 00835231 [35].

Therefore, it is not surprising that, despite the substantial efforts of 130 research teams comprising approximately 600 experts from leading institutions worldwide (about 45% from Europe, 30% from the Americas, 20% from Asia, and 5% from Africa) participating in the JEDI COVID-19 Challenge and screening billions of synthesizable compounds, only 27 weak inhibitors of SARS-CoV-2 targets were ultimately identified [36]. Clearly, the ambitious goal announced at the outset of the JEDI COVID-19 Challenge "To screen billions of molecules with blocking interactions relevant to SARS-CoV-2, and fast-track the route to a therapeutic treatment" was not achieved and, in all likelihood, could not have been achieved using this approach.

The initial plan was to commence virtual screening on May 4, 2020, and complete it by June 6, 2020. However, this deadline was extended several times, with the final submission date for reports set for July 17, 2020. Aggregation and evaluation of the submitted results were completed by November 18, 2020. In total, 1200 compounds were included in the final list, of which 1000 were selected for synthesis. By April 21, 2021, 878 compounds were synthesized and were tested in anti-SARS-CoV-2 assays by May 24, 2021. At that stage, it became evident that none of the identified inhibitors exhibited activity with an IC50 value of 100 nanomolar or better.

On July 7, 2022, Prof. Thomas Hermans, program manager of the "JEDI Billion Molecules Against COVID 19 Grand Challenge," submitted his Letter of Resignation, outlining the organizational challenges encountered during the project. It was subsequently decided to prepare a joint manuscript representing the collective efforts of the participating community. The manuscript was submitted to the Journal of the American Chemical Society on March 3, 2023, and rejected on April 11, 2023. The revised version was later accepted for publication in Molecular Informatics on October 13, 2023, and published in January 2024 [53].

Just as "One woman can give birth to a child in nine months, but nine women cannot give birth to a child in one month", the discovery of an effective anti-coronavirus drug cannot be hastened merely by increasing the number of researchers, even if they possess substantial expertise in the field.

The testing results for compounds selected by thirty research teams using three independent computational approaches indicated that the accuracy of virtual screening remained low – only a few percent ($27/878 \approx 0.03$). This outcome can be attributed primarily to the pronounced scarcity of reliable data at the initial stage of the project. At that time, all available compounds

with reported ("measured") activity were incorporated into the training sets for the development of (Q)SAR models. Subsequent analyses revealed, however, that some of these compounds lacked genuine biological activity. The inclusion of low-confidence data introduced substantial noise into the training sets, thereby diminishing the predictive performance of the models – a clear case of purgamentum init, exit purgamentum ("garbage in, garbage out").

Nevertheless, even under such conditions, despite the highly limited and noisy training data, the (Q)SAR models developed using machine learning and molecular modeling approaches enabled a substantial reduction in the human, temporal, material, and financial resources required for experimental investigations. Instead of synthesizing and biologically testing billions of molecules, only 878 compounds were synthesized and evaluated in biological assays, leading to the identification of 27 novel anti-coronavirus agents.

Conclusion

On May 5, 2023, the World Health Organization lifted the Public Health Emergency of International Concern (PHEIC) designation, signifying the transition from the acute phase of the COVID-19 pandemic to its long-term endemic management. Nevertheless, the consequences of SARS-CoV-2 infection – particularly long COVID – remain a significant health burden for many patients [69].

The urgent need for a prompt and coordinated response to the COVID-19 pandemic has driven an unprecedented acceleration of scientific and clinical research. The viral genome has been sequenced; diagnostic assays based on PCR and ELISA methodologies have been developed; fundamental mechanisms underlying viral pathogenesis have been elucidated; putative molecular targets have been identified; and experimental models have been established for the in vitro evaluation of potential antiviral agents. Ongoing clinical studies are focused on repurposing existing pharmacotherapies, assessing the safety and efficacy of candidate vaccines, and characterizing the distinctive features of patients responses to infection and therapeutic interventions. A substantial portion of these findings is disseminated almost immediately through online platforms of scientific journals and various specialized research databases.

Certain inconsistencies have been noted between experimental findings and clinical outcomes, highlighting the need for further methodological refinement and validation. Moreover, some studies disseminated through accelerated publication venues lack sufficient methodological rigor and therefore warrant more comprehensive evaluation.

The distinctiveness of the JEDI COVID-19 Challenge lies in its requirement to conduct integrative analyses of experimental and clinical data on SARS-CoV-2/COVID-19 in near real time, in parallel with the rapid emergence of new scientific evidence. Beginning with data on several dozen compounds that demonstrated inhibitory activity against viral infection in cell-based assays, we subsequently assembled training datasets and developed classification and regression (Q)SAR models. These models were then applied for large-scale virtual screening to identify compounds with prospective anti-coronavirus activity among more than one billion drug-like molecules.

Experimental validation of these selected hits is expected to significantly enrich the current knowledge base in this domain, considering that conventional training sets employed in (Q)SAR model development typically comprise only several thousand compounds, whereas the estimated size of the drug-like chemical space approaches approximately 10^{60} molecules [53]. We anticipate that the continued accumulation of experimental data and systematic exploration of the chemical

space will accelerate the discovery of novel therapeutic agents with enhanced safety and efficacy profiles for the treatment of COVID-19.

According to existing estimates, in addition to viruses, there exists a vast array of other potential biogenic threats worldwide, including up to 10^{12} distinct microorganisms, 66 fungal strains, and approximately 391000 plant species. Consequently, new diseases threatening human health may emerge as a result of hypersensitivity reactions to substances secreted by these biological species, as well as from infections or toxic effects they induce [69]. Furthermore, the potential for such biogenic threats has been convincingly demonstrated by Chinese researchers, who conducted systematic studies and identified eight novel pathogenic viruses in rodents inhabiting the tropical island of Hainan [47]. These viruses are considered highly likely to infect humans should they succeed in crossing the species barrier.

As evidenced by the COVID-19 pandemic, the only rapid response to an emerging biogenic threat is the application of the already available drugs – that is, their repurposing for new therapeutic indications. The consequences of previous biogenic threats, such as SARS, MERS, and others, were fortunately less severe; however, this circumstance led to a rapid decline in investments in related research. Consequently, humanity was insufficiently prepared for the COVID-19 pandemic. Considering the lessons learned from this crisis, it is essential to maintain and further develop the research initiatives established during this period, particularly within existing consortia, to continue the discovery of new antiviral agents and to establish robust theoretical, practical, and methodological frameworks for an effective response to future biogenic threats [8].

In view of these considerations, we are in full agreement with the assertion of former NIH Director Francis Collins, who stated: "Perhaps the most valuable lesson that COVID-19 has taught the research community – and hopefully society more broadly – is the importance of collective effort and continuous investment in basic and applied research." [14].

Acknowledgments

The study was conducted in the framework of the Program for Basic Research in the Russian Federation for a long-term period (20212030) (No. 122030100170-5).

References

1. Aldrich Market Select (AMS). `https://www.sigmaaldrich.com/chemistry/chemistry-services/aldrich-market-select.html`, accessed: 2020-07-15

2. Antiviral CAS dataset. `https://www.cas.org/covid-19-antiviral-compounds-dataset`, accessed: 2020-07-15

3. Anusevicius, K., Mickevicius, V., Stasevych, M., *et al.*: Design, synthesis, in vitro antimicrobial activity evaluation and computational studies of new N-(4-iodophenyl)-β-alanine derivatives. Res. Chem. Intermed. 41(10), 7517–7540 (2014). `https://doi.org/10.1007/s11164-014-1841-0`

4. AutoDock Vina. `http://vina.scripps.edu/`, accessed: 2020-07-15

5. Baker, J.D., Uhrich, R.L., Kraemer, G.C., *et al.*: A drug repurposing screen identifies hepatitis C antivirals as inhibitors of the SARS-CoV2 main protease. PLoS One 16(2), e0245962 (2021). https://doi.org/10.1371/journal.pone.0245962

6. Bender, A.: How similar are those molecules after all? Use two descriptors and you will have three different answers. Expert Opin. Drug Discov. 5(12), 1141–1151 (2010). https://doi.org/10.1517/17460441.2010.517832

7. Betow, J.Y., Turon, G., Metuge, C.S., *et al.*: The chemical space spanned by manually curated datasets of natural and synthetic compounds with activities against SARS-CoV-2. Mol. Inform. 44(1), e202400293 (2025). https://doi.org/10.1002/minf.202400293

8. Bobrowski, T., Melo-Filho, C.C., Korn, D., *et al.*: Learning from history: do not flatten the curve of antiviral research! Drug Discov. Today 25(9), 1604–1613 (2020). https://doi.org/10.1016/j.drudis.2020.07.008

9. Bojkova, D., McGreig, J.E., McLaughlin, K.M., *et al.*: SARS-CoV-2 and SARS-CoV differ in their cell tropism and drug sensitivity profiles. bioRxiv preprint (2020). https://doi.org/10.1101/2020.04.03.024257

10. Burov, Yu.V., Poroikov, V.V., Korolchenko, L.V.: National system for registration and biological testing of chemical compounds: facilities for new drugs search. Bulletin of the National Center for Biologically Active Compounds 1, 4–25 (1990).

11. ChEBI. https://www.ebi.ac.uk/chebi/, accessed: 2020-07-15

12. ChEMBL database. https://www.ebi.ac.uk/chembl, accessed: 2026-02-15

13. Chou, C.Y., Chien, C.H., Han, Y.S., *et al.*: Thiopurine analogues inhibit papain-like protease of severe acute respiratory syndrome coronavirus. Biochem. Pharmacol. 75(8), 1601–1609 (2008). https://doi.org/10.1016/j.bcp.2008.01.005

14. Collins, F., Adam, S., Colvis, C., *et al.*: The NIH-led research response to COVID-19. Science 379(6631), 441–444 (2023). https://doi.org/10.1126/science.adf5167

15. Cortellis Drug Discovery Intelligence. https://www.cortellis.com/drugdiscovery/, accessed: 2020-07-15

16. Dai, W., Zhang, B., Jiang, X.M., *et al.*: Structure-based design of antiviral drug candidates targeting the SARS-CoV-2 main protease. Science 368(6497), 1331–1335 (2020). https://doi.org/10.1126/science.abb4489

17. De Meyer, S., Bojkova, D., Cinatl, J., *et al.*: Lack of antiviral activity of darunavir against SARS-CoV-2. Int. J. Infect. Dis. 97, 7–10 (2020). https://doi.org/10.1016/j.ijid.2020.05.085

18. Dimova, D., Bajorath, J.: Advances in activity cliff research. Mol. Inform. 35(5), 181–191 (2016). https://doi.org/10.1002/minf.201600023

19. Do, T.N.D., Abdelnabi, R., Boda, B., *et al.*: Remdesivir: The triple combination of REmdesivir (GS-441524), molnupiravir and ribavirin is highly efficient in inhibiting coronavirus replication in human nasal airway epithelial cell cultures and in a hamster infection model. Antiviral Res. 231, 105994 (2024). https://doi.org/10.1016/j.antiviral.2024

20. Druzhilovskiy, D.S., Stolbov, L.A., Savosina, P.I., *et al.*: Computational approaches to identify a hidden pharmacological potential in large chemical libraries. Supercomputing Frontiers and Innovations 7(3), 57–76 (2020). https://doi.org/10.14529/jsfi200306

21. Eastman, R.T., Roth, J.S., Brimacombe, K.R., *et al.*: Remdesivir: A review of its discovery and development leading to emergency use authorization for treatment of COVID-19. ACS Cent. Sci. 6(5), 672–683 (2020). https://doi.org/10.1021/acscentsci.0c00489

22. Ellinger, B., Bojkova, D., Zaliani, A., *et al.*: A SARS-CoV-2 cytopathicity dataset generated by high-content screening of a large drug repurposing collection. Sci. Data 8(1), 70 (2021). https://doi.org/10.1038/s41597-021-00848-4

23. ENAMINE in-stock compounds. https://enamine.net/, accessed: 2026-02-15

24. Filimonov, D., Poroikov, V., Borodina, Yu., Gloriozova, T.: Chemical similarity assessment through multilevel neighborhoods of atoms: definition and comparison with the other descriptors. J. Chem. Inf. Comput. Sci. 39(4), 666–670 (1999). https://doi.org/10.1021/ci980335o

25. Filimonov, D.A., Zakharov, A.V., Lagunin, A.A., Poroikov, V.V.: QNA-based 'Star Track' QSAR approach. SAR QSAR Environ. Res. 20(7–8), 679–709 (2009). https://doi.org/10.1080/10629360903438370

26. Filimonov, D.A., Lagunin, A.A., Gloriozova, T.A., *et al.*: Prediction of the biological activity spectra of organic compounds using the PASS online web resource. Chem. Heterocycl. Comp. 50(3), 444–457 (2016). https://doi.org/10.1007/s10593-014-1496-1

27. Filimonov, D.A., Druzhilovskiy, D.S., Lagunin, A.A., *et al.*: Computer-aided prediction of biological activity spectra for chemical compounds: opportunities and limitations. Biomedical Chemistry: Research and Methods 1(1), e00004 (2018). https://doi.org/10.18097/bmcrm00004

28. Filimonov, D.A., Akimov, D.V., Poroikov, V.V.: Method of self-consistent regression in analysis of quantitative structure-property relationships of chemical compounds. Pharm. Chem. J. 38(1), 21–24 (2004). https://doi.org/10.1023/B:PHAC.0000027639.17115.5d

29. Fourches, D., Muratov, E., Tropsha, A.: Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. J. Chem. Inf. Model 50(7), 1189–1204 (2010). https://doi.org/10.1021/ci100176x

30. Fourches, D., Muratov, E., Tropsha, A.: Curation of chemogenomics data. Nat. Chem. Biol. 11(8), 535 (2015). https://doi.org/10.1038/nchembio.1881

31. Fourches, D., Muratov, E., Tropsha, A.: Trust, but Verify II: A Practical Guide to Chemogenomics Data Curation. J. Chem. Inf. Model 56(7), 1243–1252 (2016). https://doi.org/10.1021/acs.jcim.6b00129

32. Geronikaki, A., Druzhilovskiy, D., Zakharov, A., Poroikov, V.: Computer-aided predictions for medicinal chemistry via Internet. SAR QSAR Environ. Res. 19(1-2), 27–38 (2008). https://doi.org/10.1080/10629360701843649

33. GIAID. https://gisaid.org/, accessed: 2026-02-15

34. Ghosh, A.K., Takayama, J., Aubin, Y., *et al.*: Structure-based design, synthesis, and biological evaluation of a series of novel and reversible inhibitors for the severe acute respiratory syndrome-coronavirus papain-like protease. J. Med. Chem. 52(16), 5228–5340 (2009). https://doi.org/10.1021/jm900611t

35. Halford, B.: The path to Paxlovid. ACS Cent. Sci. 8(4), 405–407 (2022). https://doi.org/10.1021/acscentsci.2c00369

36. Hodgson, C.L., Broadley, T.: Long COVID – unravelling a complex condition. Lancet Respir. Med. 11(8), 667–668 (2023). https://doi.org/10.1016/S2213-2600(23)00232-1

37. Huang, L., Chen, Y., Xiao, J., *et al.*: Progress in the research and development of anti-COVID-19 drugs. Front. Public Health 8, 365 (2020). https://doi.org/10.3389/fpubh.2020.00365

38. IBS Natural Compounds Set. https://www.ibscreen.com/, accessed: 2020-07-15

39. Jeon, S., Ko, M., Lee, J., *et al.*: Identification of Antiviral Drug Candidates against SARS-CoV-2 from FDA-Approved Drugs. Antimicrob. Agents Chemother. 64(7), e00819–20 (2020). https://doi.org/10.1128/AAC.00819-20

40. JEDI billion molecules against COVID-19 grand challenge. https://www.jedi.foundation/covid19challenge, accessed: 2026-02-15

41. Jin, Z., Du, X., Xu, Y., *et al.*: Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. Nature 582(7811), 289–293 (2020). https://doi.org/10.1038/s41586-020-2223-y

42. Kakhar Umar, A., Zothantluanga, J.H., Luckanagul, J.A., *et al.*: Structure-based computational screening of 470 natural quercetin derivatives for identification of SARS-CoV-2 Mpro inhibitor. PeerJ. 11, e14915 (2023). https://doi.org/10.7717/peerj

43. Klimenko, K., Marcou, G., Horvath, D., Varnek, A.: Chemical space mapping and structure-activity analysis of the ChEMBL antiviral compound set. J. Chem. Inf. Model 56(8), 1438–1454 (2016). https://doi.org/10.1021/acs.jcim.6b00192

44. Kneller, D.W., Galanie, S., Phillips, G., *et al.*: Malleability of the SARS-CoV-2 3CL Mpro active-site cavity facilitates binding of clinical antivirals. Structure 28(12), 1313–1320 (2020). https://doi.org/10.1016/j.str.2020.10.007

45. Kubinyi, H.: Chemical similarity and biological activities. J. Braz. Chem. Soc. 13(6), 717–726 (2002). https://doi.org/10.1590/S0103-50532002000600002

46. Kuzikov, M., Costanzi, E., Reinshagen, J., *et al.*: Identification of inhibitors of SARS-CoV-2 3CL-pro enzymatic activity using a small molecule in vitro repurposing screen. ACS Pharmacol. Transl. Sci. 4(3), 1096–1110 (2021). https://doi.org/10.1021/acsptsci.0c00216

47. Li, Y., Tang, C., Zhang, Y., *et al.*: Diversity and independent evolutionary profiling of rodent-borne viruses in Hainan, a tropical island of China. Virol. Sin. 8(5), 651–662 (2023). https://doi.org/10.1016/j.virs.2023.08.003

48. Listings of WHOs response to COVID-19. https://www.who.int/news/item/29-06-2020-covidtimeline, accessed: 2026-02-15

49. Ma, C., Hurst, B., Hu, Y., *et al.*: Boceprevir, GC-376, and calpain inhibitors II, XII inhibit SARS-CoV-2 viral replication by targeting the viral main protease. Cell Res. 30(8), 678–692 (2020). https://doi.org/10.1038/s41422-020-0356-z

50. Mansouri, K., Kleinstreuer, N., Abdelaziz, A.M., *et al.*: CoMPARA: Collaborative modeling project for androgen receptor activity. Environ Health Perspect. 128(2), 27002 (2020). https://doi.org/10.1289/EHP5580

51. Martinez, M.A.: Lack of effectiveness of repurposed drugs for COVID-19 treatment. Front. Immunol. 12, 635371 (2021). https://doi.org/10.3389/fimmu.2021

52. Murtazalieva, K.A., Druzhilovskiy, D.S., Goel, R.K., *et al.*: How good are publicly available web services that predict bioactivity profiles for drug repurposing? SAR QSAR Environ. Res. 28 (10), 843–862 (2017). https://doi.org/10.1080/1062936X.2017.1399448

53. Muratov, E.N., Bajorath, J., Sheridan, R.P., *et al.*: QSAR Without Borders. Chem. Soc. Rev. 49(11), 3525–3564 (2020). https://doi.org/10.1039/d0cs00098a

54. Natural Product Activity and Species Source (NPASS). https://bidd2.nus.edu.sg/NPASS/, accessed: 2020-07-15

55. Natural Products from Northern African Sources (NANPDB). https://african-compounds.org/nanpdb/, accessed: 2020-07-15

56. Nuclei of Bioassays, Ecophysiology and Biosynthesis of Natural Products Database (NuBBE). https://nubbe.iq.unesp.br/portal/nubbedb.html, accessed: 2020-07-15

57. Poroikov, V.V., Filimonov, D.A., Gloriozova, T.A., *et al.*: Computer-aided prediction of biological activity spectra for organic compounds: the possibilities and limitations. Russ. Chem. Bull. 68(12), 2143–2154 (2019). https://doi.org/10.1007/s11172-019-2683

58. Poroikov, V.V., Filimonov, D.A., Borodina, Yu.V., *et al.*: Robustness of biological activity spectra predicting by computer program PASS for non-congeneric sets of chemical compounds. J. Chem. Inf. Comput. Sci. 40(6), 1349-01355 (2000). https://doi.org/10.1021/ci000383k

59. PostERA activity data. https://postera.ai/covid/activity_data, accessed: 2020-07-15

60. Protein Data Bank. https://www.rcsb.org/, accessed: 2026-02-15

61. Pruijssers, A.J., George, A.S., Schäfer, A., *et al.*: Remdesivir potently inhibits SARS-CoV-2 in human lung cells and chimeric SARS-CoV expressing the SARS-CoV-2 RNA polymerase in mice. Cell Rep. 32(3), 107940 (2020). https://doi.org/10.1016/j.celrep.2020.107940

62. Ratia, K., Pegan, S., Takayama, J., *et al.*: A noncovalent class of papain-like protease/deubiquitinase inhibitors blocks SARS virus replication. PNAS 105(42), 16119–16124 (2008). https://doi.org/10.1073/pnas.0805240105

63. Riva, L., Yuan, S., Yin, X., *et al.*: A large scale drug repositioning survey for SARS-CoV-2 antivirals. Nature 586(7827), 113–119 (2020). https://doi.org/10.1038/s41586-020-2577-1

64. Sachs, J.D., Karim, S.S.A., Aknin, L., *et al.*: The Lancet Commission on lessons for the future from the COVID-19 pandemic. Lancet 400(10359), 1224–1280 (2022). https://doi.org/10.1016/S0140-6736(22)01585-9

65. SAVI. https://cactus.nci.nih.gov/download/savi_download/, accessed: 2020-07-15

66. Sheahan, T.P., Sims, A.C., Zhou, S., *et al.*: An orally bioavailable broad-spectrum antiviral inhibits SARS-CoV-2 and multiple endemic, epidemic and bat coronavirus. bioRxiv preprint (2020). https://doi.org/10.1101/2020.03.19.997890

67. Schimunek, J., Seidl, P., Elez, K., *et al.*: A community effort in SARS-CoV-2 drug discovery. Mol. Inform. 43, e202300262 (2024). https://doi.org/10.1002/minf.202300262

68. Sielaff, F., Böttcher-Friebertshäuser, E., Meyer, D., *et al.*: Development of substrate analogue inhibitors for the human airway trypsin-like protease HAT. Bioorg. Med. Chem. Lett. 21(16), 4860–4864 (2011). https://doi.org/10.1016/j.bmcl.2011.06.03

69. Smith, C.I.E., Bergman, P., Hagey, D.W.: Estimating the number of diseases – the concept of rare, ultra-rare, and hyper-rare. iScience 25(8), 104698 (2022). https://doi.org/10.1016/j.isci.2022.104698

70. Stanford Coronavirus Antiviral Research Database. https://covdb.stanford.edu/, accessed: 2020-07-15

71. Stolbov, L.A., Druzhilovskiy, D.S., Filimonov, D.A., *et al.*: (Q)SAR models of HIV-1 proteins inhibition by drug-like compounds. Molecules 25(1), 87 (2019). https://doi.org/10.3390/molecules25010087

72. SWEETLEAD. https://simtk.org/projects/sweetlead, accessed: 2020-07-15

73. SYBYL-X Suite. https://www.g6g-softwaredirectory.com/bio/proteomics/structure-modeling/20710-Tripos-SYBYL-X-Suite.php, accessed: 2020-07-15

74. Touret, F., Gilles, M., Barral, K., *et al.*: In vitro screening of a FDA approved chemical library reveals potential inhibitors of SARS-CoV-2 replication. Sci. Rep. 10(1), 13093 (2020). https://doi.org/10.1038/s41598-020-70143-6

75. Universal Natural Products Database (UNPD). http://pkuxxj.pku.edu.cn, accessed: 2020-07-15

76. UCSF Dock. http://dock.compbio.ucsf.edu/, accessed: 2020-07-15

77. Wang, M., Cao, R., Zhang, L., *et al.*: Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) in vitro. Cell Res. 30(3), 269–271 (2020). https://doi.org/10.1038/s41422-020-0282-0

78. Wermuth, C.G.: Similarity in drugs: reflections on analogue design. Drug Discov. Today 11(7-8), 348–354 (2006). https://doi.org/10.1016/j.drudis.2006.02.006

79. World Wide Approved Drugs (WWAD) database. https://way2drug.com/wwad/, accessed: 2026-02-15

80. Yuan, C., Goonetilleke, E.C., Unarta, I.C., Huang, X.: Incorporation efficiency and inhibition mechanism of 2'-substituted nucleotide analogs against SARS-CoV-2 RNA-dependent RNA polymerase. Phys. Chem. Chem. Phys. 23(36), 20117–20128 (2021). https://doi.org/10.1039/d1cp03049c

81. Zakharov, A.V., Peach, M.L., Sitzmann, M., Nicklaus, M.C.: A new approach to radial basis function approximation and its application to QSAR. J. Chem. Inf. Model 54(3), 713-719 (2014). https://doi.org/10.1021/ci400704f

82. ZINC. https://zinc.docking.org/, accessed: 2020-07-15