

# Supercomputing Frontiers and Innovations

2020, Vol. 7, No. 3

## Scope

- Enabling technologies for high performance computing
- Future generation supercomputer architectures
- Extreme-scale concepts beyond conventional practices including exascale
- Parallel programming models, interfaces, languages, libraries, and tools
- Supercomputer applications and algorithms
- Distributed operating systems, kernels, supervisors, and virtualization for highly scalable computing
- Scalable runtime systems software
- Methods and means of supercomputer system management, administration, and monitoring
- Mass storage systems, protocols, and allocation
- Energy and power minimization for very large deployed computers
- Resilience, reliability, and fault tolerance for future generation highly parallel computing systems
- Parallel performance and correctness debugging
- Scientific visualization for massive data and computing both external and in situ
- Education in high performance computing and computational science

## Editorial Board

### Editors-in-Chief

- **Jack Dongarra**, University of Tennessee, Knoxville, USA
- **Vladimir Voevodin**, Moscow State University, Russia

### Editorial Director

- **Leonid Sokolinsky**, South Ural State University, Chelyabinsk, Russia

### Associate Editors

- **Pete Beckman**, Argonne National Laboratory, USA
- **Arndt Bode**, Leibniz Supercomputing Centre, Germany
- **Boris Chetverushkin**, Keldysh Institute of Applied Mathematics, RAS, Russia
- **Alok Choudhary**, Northwestern University, Evanston, USA

- **Alexei Khokhlov**, Moscow State University, Russia
- **Thomas Lippert**, Jülich Supercomputing Center, Germany
- **Satoshi Matsuoka**, Tokyo Institute of Technology, Japan
- **Mark Parsons**, EPCC, United Kingdom
- **Thomas Sterling**, CREST, Indiana University, USA
- **Mateo Valero**, Barcelona Supercomputing Center, Spain

## Subject Area Editors

- **Artur Andrzejak**, Heidelberg University, Germany
- **Rosa M. Badia**, Barcelona Supercomputing Center, Spain
- **Franck Cappello**, Argonne National Laboratory, USA
- **Barbara Chapman**, University of Houston, USA
- **Yuefan Deng**, Stony Brook University, USA
- **Ian Foster**, Argonne National Laboratory and University of Chicago, USA
- **Geoffrey Fox**, Indiana University, USA
- **Victor Gergel**, University of Nizhni Novgorod, Russia
- **William Gropp**, University of Illinois at Urbana-Champaign, USA
- **Erik Hagersten**, Uppsala University, Sweden
- **Michael Heroux**, Sandia National Laboratories, USA
- **Torsten Hoefler**, Swiss Federal Institute of Technology, Switzerland
- **Yutaka Ishikawa**, AICS RIKEN, Japan
- **David Keyes**, King Abdullah University of Science and Technology, Saudi Arabia
- **William Kramer**, University of Illinois at Urbana-Champaign, USA
- **Jesus Labarta**, Barcelona Supercomputing Center, Spain
- **Alexey Lastovetsky**, University College Dublin, Ireland
- **Yutong Lu**, National University of Defense Technology, China
- **Bob Lucas**, University of Southern California, USA
- **Thomas Ludwig**, German Climate Computing Center, Germany
- **Daniel Mallmann**, Jülich Supercomputing Centre, Germany
- **Bernd Mohr**, Jülich Supercomputing Centre, Germany
- **Onur Mutlu**, Carnegie Mellon University, USA
- **Wolfgang Nagel**, TU Dresden ZIH, Germany
- **Alexander Nemukhin**, Moscow State University, Russia
- **Edward Seidel**, National Center for Supercomputing Applications, USA
- **John Shalf**, Lawrence Berkeley National Laboratory, USA
- **Rick Stevens**, Argonne National Laboratory, USA
- **Vladimir Sulimov**, Moscow State University, Russia
- **William Tang**, Princeton University, USA
- **Michela Taufer**, University of Delaware, USA
- **Andrei Tchernykh**, CICESE Research Center, Mexico
- **Alexander Tikhonravov**, Moscow State University, Russia
- **Eugene Tyrtshnikov**, Institute of Numerical Mathematics, RAS, Russia
- **Roman Wyrzykowski**, Czestochowa University of Technology, Poland
- **Mikhail Yakobovskiy**, Keldysh Institute of Applied Mathematics, RAS, Russia

## Technical Editors

- **Yana Kraeva**, South Ural State University, Chelyabinsk, Russia
- **Mikhail Zymbler**, South Ural State University, Chelyabinsk, Russia
- **Dmitry Nikitenko**, Moscow State University, Moscow, Russia

## Contents

<b>Accounting of Receptor Flexibility in Ultra-Large Virtual Screens with VirtualFlow Using a Grey Wolf Optimization Method</b> C. Gorgulla, K. Fackeldey, G. Wagner, H. Arthanari .....	4
<b>Perspectives on Supercomputing and Artificial Intelligence Applications in Drug Discovery</b> J. Xu, J. Ye .....	13
<b>Computational Modeling of the SARS-CoV-2 Main Protease Inhibition by the Covalent Binding of Prospective Drug Molecules</b> A.V. Nemukhin, B.L. Grigorenko, I.V. Polyakov, S.V. Lushchekina .....	25
<b>Computational Characterization of the Substrate Activation in the Active Site of SARS-CoV-2 Main Protease</b> M.G. Khrenova, V.G. Tsirelson, A.V. Nemukhin .....	33
<b>In Search of Non-covalent Inhibitors of SARS-CoV-2 Main Protease: Computer Aided Drug Design Using Docking and Quantum Chemistry</b> A.V. Sulimov, D.C. Kutov, A.S. Taschilova, I.S. Ilin, N.V. Stolpovskaya, K.S. Shikhaliev, V.B. Sulimov .....	41
<b>Computational Approaches to Identify a Hidden Pharmacological Potential in Large Chemical Libraries</b> D.S. Druzhilovskiy, L.A. Stolbov, P.I. Savosina, P.V. Pogodin, D.A. Filimonov, A.V. Veselovsky, K. Stefanisko, N.I. Tarasova, M.C. Nicklaus, V.V. Poroikov .....	57



This issue is distributed under the terms of the Creative Commons Attribution-Non Commercial 3.0 License which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is properly cited.

# Accounting of Receptor Flexibility in Ultra-Large Virtual Screens with VirtualFlow Using a Grey Wolf Optimization Method

*Christoph Gorgulla*<sup>1,2,5</sup> , *Konstantin Fackeldey*<sup>3,4</sup> ,  
*Gerhard Wagner*<sup>2</sup> , *Haribabu Arthanari*<sup>2,5</sup> 

© The Authors 2020. This paper is published with open access at SuperFri.org

Structure-based virtual screening approaches have the ability to dramatically reduce the time and costs associated to the discovery of new drug candidates. Studies have shown that the true hit rate of virtual screenings improves with the scale of the screened ligand libraries. Therefore, we have recently developed an open source drug discovery platform (VirtualFlow), which is able to routinely carry out ultra-large virtual screenings. One of the primary challenges of molecular docking is the circumstance when the protein is highly dynamic or when the structure of the protein cannot be captured by a static pose. To accommodate protein dynamics, we report the extension of VirtualFlow to allow the docking of ligands using a grey wolf optimization algorithm using the docking program GWOVina, which substantially improves the quality and efficiency of flexible receptor docking compared to AutoDock Vina. We demonstrate the linear scaling behavior of VirtualFlow utilizing GWOVina up to 128 000 CPUs. The newly supported docking method will be valuable for drug discovery projects in which protein dynamics and flexibility play a significant role.

*Keywords:* ultra-large virtual screening, molecular docking, drug discovery, COVID-19, structure-based drug design, CADD, computer aided drug design, AutoDock, grey wolf optimization, cloud computing.

## Introduction

In structure based drug design one common goal is to design and optimize a small molecule (compound), such that it fits optimally, with favorable energetics into the binding pocket of a target protein. The conventional approach is to test the compounds individually in an experimental wet lab setting via high throughput screens. Besides these experiments, computer-aided drug design (CADD) has been established, which allows to compute the binding affinity between the small molecule and the target protein. In the realm of structure-based virtual screenings, for a given binding pocket on the surface of the protein, ligands from a large databases of prospective candidate molecules are *screened*, i.e. the binding strength of individual ligands from the large database to a given target are computationally predicted by molecular docking. Docking programs such as AutoDock Vina [22] fit these candidate molecules into the binding pocket and score the resulting binding geometry with regard to the binding affinity. The scoring function associated with the docking program assigns a docking score, that is a marker of the calculated binding affinity, to each tested interaction geometry (Fig. 1a). The docking score is a measure of the predicted binding affinity of the small molecule to the protein (Fig. 1b). The interaction geometry (conformation) with the best docking score is the final score for the compound. Docking methods should estimate the binding affinity as precisely as possible on the one hand, and as quickly as possible on the other hand. Even with fast docking programs, ultra-large virtual

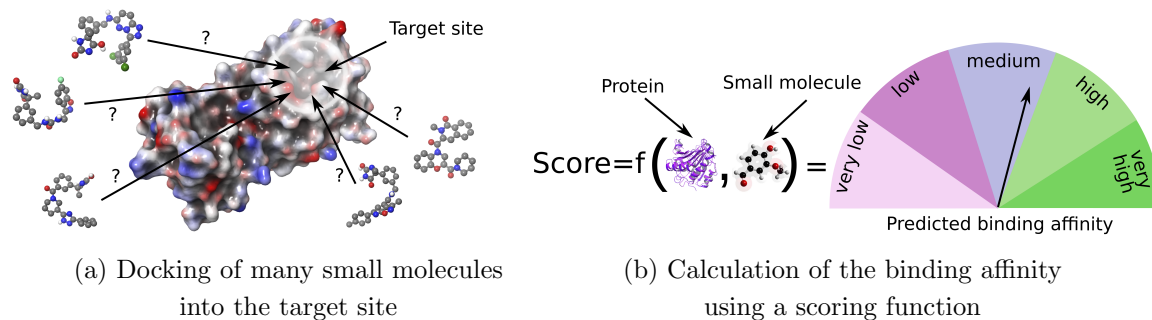
<sup>1</sup>Department of Physics, Harvard University, Cambridge, USA

<sup>2</sup>Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, USA

<sup>3</sup>Institute of Mathematics, Technical University Berlin, Berlin, Germany

<sup>4</sup>Zuse Institute Berlin, Berlin, Germany

<sup>5</sup>Department of Cancer Biology, Dana Farber Cancer Institute, Boston, USA



**Figure 1.** Principle of virtual screening and scoring of small molecules

screenings involving hundreds of millions or billions of molecules require a massive amount of computation time with millions of CPU hours. Thus resources such as supercomputer or cloud-based computational platforms are required. With increased scale of the ligand library screened, the true hit rate (the fraction of hits which bind to the target protein in experimental binding assays) of the screening improves, as was shown by theoretical and experimental studies [8, 9, 13]. Since the chemical space of small molecules suitable for drug discovery is estimated to contain more than  $10^{60}$  molecules [1], even billion of compounds represent only a minuscule fraction of the possible chemical space to be explored.

We describe here an extension of a platform allowing to screen billions of compounds (VirtualFlow), focusing on the new GWOVina docking program which is based on the novel global optimization algorithm.

## 1. Materials and Methods

VirtualFlow is an open source parallel workflow platform that we recently developed for executing virtual screenings [9, 23], which for the first time allowed to routinely screen billions of compounds. VirtualFlow can be employed on any type of Linux-based computer clusters managed by a batchsystem, as well as on cloud computing platforms such as the Google Cloud. VirtualFlow consists of two different modules, one for the ligand preparation (VirtualFlow for Ligand Preparation – VFLP) and one for the virtual screenings (VirtualFlow for Virtual Screening – VFVS). Although both methods are used for different tasks, they share the same core technology.

### 1.1. VirtualFlow for Ligand Preparation

Before the ligands can be docked to the target protein, they must be prepared into a ready-to-dock format. VFLP is dedicated for this task, and is able to prepare ultra-large ligand libraries which can be readily used by VFVS. In addition to the correct file format and the three dimensional spatial structure of the ligand, it can compute the tautomerization and protonation states of the molecules via tools like ChemAxon’s JChem package [2] and Open Babel [17]. Once a library is prepared, it can be used again and again with VFVS. However, since VFLP can prepare the molecules in almost any output format, the prepared ligand libraries can also be used for any other purpose and other docking platforms. We have previously used VFLP to prepare the REAL library from Enamine (2018 version) containing 1.4 billion compounds, as well as the ZINC15 library (2018 version) containing 1.2 billion compounds into a ready-to-dock format

(VirtualFlow versions of these libraries) [9, 21]. The VirtualFlow versions of these libraries are freely available on our website [23].

## 1.2. VirtualFlow for Virtual Screening

A primary goal of virtual screenings is to create a hit list which contains the top scoring compounds ranked by their docking score, essentially sorting the molecules in the library based on their propensity to bind to the target site on the protein. The docking score correlates with the predicted binding affinity to the target protein. The scoring (and thus the ranking) is based on the calculation of the free energy of the small molecules to the receptor. While for a single compound the docking can be done relatively quickly, for a large ligand library containing millions or billions of compounds, this requires a substantial amount of computing capacity and time. The computational costs associated with the docking routines dramatically increase further if docking procedures with high accuracy (e.g. including receptor flexibility, exhaustive search of the docking space) are used. To solve the first challenge caused by the large number of ligands, VirtualFlow uses a massive parallelization approach to speed up the virtual screening (*vide infra*). To deal with the second challenge involving the increased computational costs in the high accuracy implementation, VirtualFlow can be deployed in a multi-staged manner. In a multi-staged virtual screen, the entire collection of ligands that are planned to be screened is at first docked with a fast method at reduced accuracy. In the next stage, the top X% of compounds of the first stage are transferred to the next stage, and screened with higher accuracy. In principle, any number of stages with subsequent increase in computational time, which in turn increases accuracy, can be employed.

One of the key characteristics of VirtualFlow is that it is able to scale very efficiently up to hundreds of thousands of CPUs, which it achieves by employing an embarrassingly/perfectly parallelization strategy. In this context, VirtualFlow uses an advanced task list approach, which completely eliminates the need for any communication between individual workers. The tasks are distributed at the beginning in advance to the individual workers by a workload balancer, which removes the need to access the task list during the runtime of a job. To reduce the number of tasks, the ligands which need to be processed are grouped into collections (of e.g. 1000 ligands), and each collection represents a task in the central task list. The input and output databases of the workflow consist of a multi-level file structure, which involves folders, and (compressed and uncompressed) tar archives.

VFVS supports different docking programs such as AutoDock Vina [22], Smina [11], Vina-Carb [16], and VinaXB [10]. All of them are based on AutoDock Vina and improve different aspects. Vina-Carb for instance improves the accuracy of carbohydrate docking. Here, we have added support for GWOVina [24], which is able to handle protein side chain flexibility more efficiently than AutoDock Vina. It is able to efficiently handle (in terms of computational speed) considerably more number of flexible side chains compared to AutoDock Vina. In addition, in terms of the quality of the results, GWOVina samples the conformational space of the side chains much more effectively due to the utilization of a new swarm-optimization-based optimization algorithm, the grey wolf optimizer (GWO) [24].

### 1.3. Grey Wolf Optimization-Based Docking Algorithm

We have added support of GWOVina [24] to VFVS. GWOVina uses the recently developed grey wolf optimization algorithm, which has turned out to be highly efficient for flexible ligand docking and other types of tasks [12, 14, 24]. Finding the optimal or best orientation of the ligand in the target site of the protein is equivalent to finding the absolute minimum of an energy landscape in the high-dimensional conformational space, similar to the protein folding problem [18]. The conformational space is defined by the degrees of freedom of both the ligand and the flexible side chains of the individual amino acid that constitute the docking site on the receptor, and all the degrees of freedom are treated equally by the algorithm. The degrees of freedom of the receptor side chains which are selected to be flexible consists of the torsion angles around the rotatable bonds. The degrees of freedom of the ligands includes the translation and rotation in three dimensions in addition to the torsion angles around the rotatable bonds.

Inspired by the hunting behavior of the grey wolf pack, the grey wolf algorithm is an optimization algorithm using swarm intelligence, i.e. it takes advantage of collective behavior of a self-organized system (wolf pack). Members of a grey wolf pack are either  $\alpha$ ,  $\beta$ ,  $\delta$ , or  $\omega$  wolves, and each of them has a different function during the hunting of a prey. Within the hierarchy, the  $\alpha$  wolf is the dominant wolf, making the decisions when hunting. The  $\beta$  wolf supports the  $\alpha$  wolf in its decisions and also in the enforcement of the orders of the  $\alpha$  wolf. The  $\delta$  wolves can be considered as a collection of specialist such as scouts (monitoring the territory), hunters (helping  $\alpha/\beta$  wolves in hunting) and elders (former experienced  $\alpha$  and  $\beta$  wolves) are among with them. The  $\delta$  wolves are subordinates of the  $\alpha$  and  $\beta$  wolves. The  $\omega$  wolves are the lowest wolves which have to submit to the  $\alpha$ ,  $\beta$  and  $\delta$  wolves. In the GWO algorithm [14], which models the hunting strategy of  $m$  grey wolves, each wolf represents a search agent, and the prey is the optimum (in our case the energy minimum, i.e. the “best” orientation of the ligand and the flexible site chains in the target site). More precisely the  $\alpha$ ,  $\beta$  and  $\delta$  wolves guide the hunting and the  $\omega$  wolves follow. We outline the core algorithm below.

Let  $x(t) \in \mathbb{R}^n$  be the position of a wolf and  $x_p(t) \in \mathbb{R}^n$  the position of the prey at time step  $t$ . The distance vector of the wolf from the prey  $p$  is then set by

$$d = |c \odot x_p(t) - x(t)|. \quad (1)$$

Here, the product  $c \odot x_p$  as well as the absolute value is understood element-wise (component-by-component), implying the former is the Hadamard product. The vector  $c = 2(\mathbf{rand}[0, 1])^n$ , where  $(\mathbf{rand}[0, 1])^n$  is a vector with  $n$  random numbers between 0 and 1 as its entries. Thus  $d$  is a vector with positive entries. With this, the position  $x(t + 1)$  of the wolf in the next time step is given by

$$x(t + 1) = x_p(t) - a \odot d, \quad (2)$$

where  $a = 2b \odot (\mathbf{rand}[0, 1])^n - b$  and  $b$  is a vector with entries decreasing linearly from 2 to 0 in each iteration step. Note, that the absolute value of  $a$  determines the moving direction of the wolf with respect to the prey. In case of  $|a| > 1$  the wolf veers away from the prey and in case  $|a| < 1$  it moves towards the prey. However in our context, the position of the optimum (prey) is not known. In the first step the energy (fitness) for each wolf position is evaluated and arranged in increasing order of fitness. The algorithm then assigns the first three  $\alpha$ ,  $\beta$  and  $\delta$  to the positions with the lowest energy since it is supposed that they are the closest to the prey

(optimum). In a next step the three best updates are computed via

$$d_\alpha = |c_1 \odot x_\alpha - x|, d_\beta = |c_2 \odot x_\beta - x|, d_\delta = |c_3 \odot x_\delta - x|, \quad (3)$$

where  $c_1, c_2$ , and  $c_3$  are defined analogue to  $c$ . With this the estimated prey positions are obtained by

$$x_1 = x_\alpha - a_1 \odot d_\alpha, x_2 = x_\beta - a_2 \odot d_\beta, x_3 = x_\delta - a_3 \odot d_\delta, \quad (4)$$

where  $a_1, a_2, a_3$  are defined analogue to  $a$ . The next position of the wolf is then given as

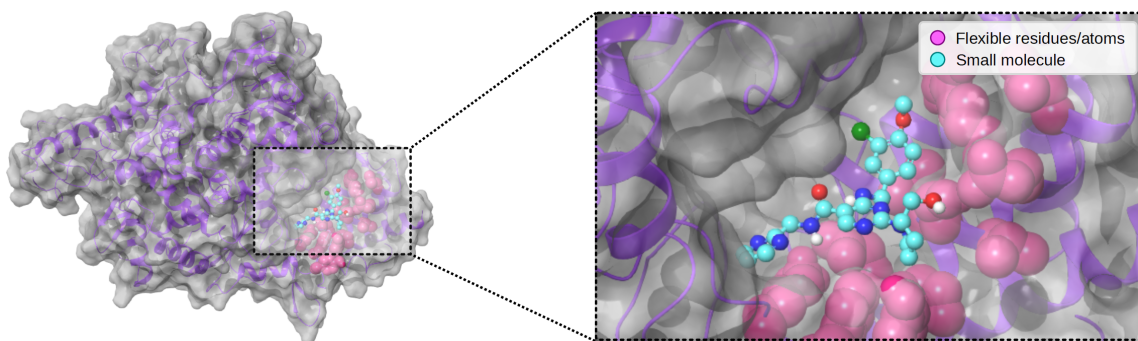
$$x(t+1) = \frac{x_1 + x_2 + x_3}{3}. \quad (5)$$

The original GWO algorithm was extended in GWOVina by an additional random walk mechanism which is sometimes employed for the movement of a wolf to improve the docking program.

For running the algorithm, the number  $m$  of search agents (wolves), the objective function (energy function), the dimension  $n$  as well as the search space (conformational space) has to be provided. The GWO algorithm replaces the global Monte Carlo-based optimizer used by the original AutoDock Vina, while the original scoring function of AutoDock Vina is used as the objective function. The computation time needed by GWOVina is proportional to the number of wolves used according to the authors, and the quality of the docking results increases with the number of wolves due to the more elaborate exploration of the conformation space [24]. To fully harness the capabilities of GWO, at least four wolves need to be employed when using this docking algorithm, as the grey wolf algorithm is based on a four level hierarchy of wolves ( $\alpha, \beta, \delta$ , and  $\omega$  wolves). GWOVina was previously compared with AutoDock Vina via several benchmarks [24], and GWOVina has shown a substantial advantage over AutoDock Vina in terms of finding better docking poses in less time.

## 2. Results

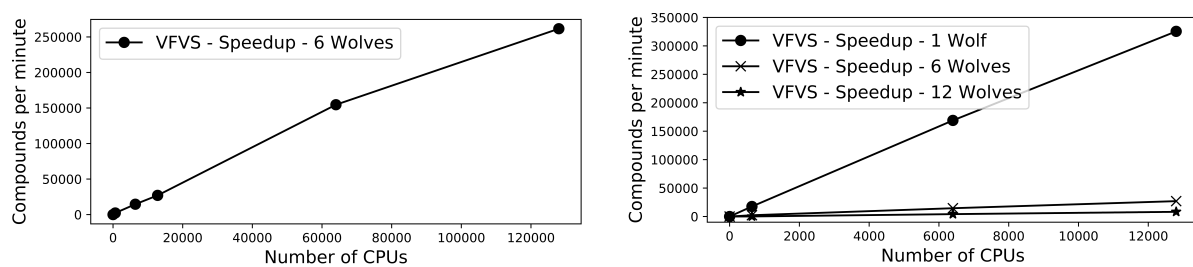
Due to the large surface area that constitutes the RNA binding interface of nsp12 (Fig. 2a), there are a large number of protein side chains to account for at the interaction surface that are critical in engaging the RNA. Most docking programs cannot handle the flexibility of such large numbers of side chains efficiently, but for GWOVina this does not pose a challenge. We have allowed a total of 12 residue side chains at the RNA binding interface to be flexible during



(a) RNA-dependent RNA polymerase (b) Small molecule docked to the RNA binding region

**Figure 2.** Biomolecular test system consisting of nsp12 used for the benchmarks





(a) Scaling behavior of VFVS using GWOVina as the docking program, where the population size of the wolf pack was set to 6

(b) The scaling behavior for different values of the wolf population size

**Figure 3.** Scaling behavior of VFVS using GWOVina as the docking program

the benchmarks (Lys593, Phe594, Tyr595, Leu854, Glu857, Arg858, Val860, Ser861, Leu862, Ile864, Asp865, Tyr915), though GWOVina could handle significantly more [24]. An example compound docked to the target site can be seen in Fig. 2b). The computations were run on a Slurm cluster in the Google Cloud [7], and the compute nodes which were used employed second generation AMD EPYC Rome CPUs. The cluster file system which was used is an Elastifile Cloud File System [3], which is a Network File System (NFS) server. For this benchmark we have created a test library, which consists of compounds from the REAL library from Enamine which we had previously prepared (see above). The entire test library had a size of 1 billion compounds, which is large enough for not being depleted during the benchmarks. The test library consists of 10 metatranches, each containing 1000 tranches, of which each tranche contains 10 000 collections, and each collection contains 10 different compounds. The 10 compounds are the same in each collection, meaning each collection in the test library is identical but nonetheless treated independently by VirtualFlow, which makes the benchmark more reproducible. The 10 distinct test molecules are relatively flexible, each containing between 9 and 10 rotatable bonds, and have a molecular weight between 400 and 425 daltons.

We have tested the scaling behavior and virtual screening speed of VirtualFlow using GWOVina with up to 128 000 CPUs, and the speedup was roughly linear up to the maximum number of CPUs tested (Fig. 3a). The test system which was used is the RNA-dependent RNA polymerase (RdRP) of the SARS-CoV-2 virus, and the targeted site on this protein is the RNA binding interface (Fig. 2). RdRPs have been effectively targeted to develop antiviral therapeutics in several viral infections in the past such as HCV, Zika virus (ZIKV), and HCoV-229E this an attractive target for therapeutic intervention of COVID-19 [4–6]. The receptor structure which was used is the cryo-EM structure with PDB code 7BV1 [25]. The RNA was removed, and the structure prepared with Maestro from Schrödinger by adding hydrogen atoms at physiological pH value [19]. The command line tool `prepare_flexreceptor4.py` of AutoDockTools was used to merge nonpolar hydrogen atoms, to split the receptor into rigid and flexible parts, and to convert the two parts into the PDBQT format [15]. The number of wolves used in the first benchmark is one, because this setting puts the most stress on the computational infrastructure, such as the cluster file system, due to the minimal docking time per ligand and thus maximum amount of file transfers and related activities.

We have also tested the virtual screening speed for different numbers of wolves which are utilized by the grey wolf optimization algorithm of GWOVina. With only one wolf the docking

speed is by far the fastest, while the computational costs of using six wolves is roughly double as fast compared to that of using 12 wolves (Fig. 3b).

## Conclusion

Virtual screenings have an enormous potential in making drug discovery faster and more affordable in the future, and in allowing to find cures for diseases which so far were incurable. With ultra-large virtual screenings now in the accessible computational range, their power has substantially increased, and they could soon become a standard approach for finding new initial hit and lead compounds. Furthermore, due to the vast chemical spaces which can be screened in comparison to traditional approaches, tight binders to even highly challenging targets can be identified. And with that, the generally more challenging class of protein-protein interactions can be targeted via virtual screening approaches. Protein-protein interactions play a role in almost any disease, and are expected to become the most important class of targets in the future. For any docking efforts the choice of the protein structure and the target site on the protein is critical to the success of the screen. Normally the starting points for the docking efforts are structures derived from X-ray crystallography, cryo-EM or NMR. These structures normally represent a single snapshot of the protein and in most cases the lowest energy conformation. However in solution, inside the cell, where the hit molecule from the screen should function, the structure could be dynamic and accommodating this information about the protein dynamics will substantially improve the true hit rate. Capturing protein dynamics is not a trivial task. NMR studies can provide information about protein dynamics over a wide timescale (nanoseconds to hours), but it is not simple to relate it back to precise structural information. Molecular dynamics simulations can provide structural information but are limited in the timescale they can sample, typically up to a few microseconds. The interesting conformational changes and allosteric changes happen in the microsecond to millisecond time regime. There have been a few simulation that have been extended to milliseconds by the group of DE Shaw using the custom designed Anton computer [20]. In the absence of detailed structural information to capture dynamics, which is the case in most efforts, incorporating side chain dynamics is a good alternative to account for dynamics. The combination of GWOVina and VirtualFlow unifies the best of both worlds, to account for dynamics and identify genuine hits, facilitated by the power of supercomputing platforms. VFVS, including the new feature, is available on GitHub (<https://github.com/VirtualFlow/VFVS>).

## Acknowledgments

We thank ChemAxon for a free academic license for the JChem package, and Google for computing time on the Google Cloud. H.A. acknowledges funding from the Claudia Adams Barr Program for Innovative Cancer Research. G.W. acknowledges support from NIH grants CA200913 and AI037581. K.F. would like to thank the Math+. We would like to thank Arthur Jaffe for his support. This research was supported in part by grant TRT 0159 from the Templeton Religion Trust and by ARO Grant W911NF1910302 to Arthur Jaffe.

Conflicts of interest: G.W. and C.G. are cofounders of the company Virtual Discovery, Inc., which provides virtual screening services. G.W. serves as the director of this company.

*This paper is distributed under the terms of the Creative Commons Attribution-Non Commercial 3.0 License which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is properly cited.*

## References

1. Bohacek, R.S., McMartin, C., Guida, W.C.: The art and practice of structure-based drug design: A molecular modeling perspective. *Medicinal Research Reviews* 16(1), 3–50 (1996), DOI: 10.1002/(SICI)1098-1128(199601)16:1;3::AID-MED1;3.0.CO;2-6
2. ChemAxon: JChem Suite 18.20.0. <https://chemaxon.com/products/jchem-engines>, accessed: 2020-09-06
3. Elastifile Cloud File System. <https://console.cloud.google.com/marketplace/details/elastifile/elastifile-cloud-file-system>, accessed: 2020-09-06
4. Elfiky, A.A.: Zika viral polymerase inhibition using anti-HCV drugs both in market and under clinical trials. *Journal of Medical Virology* 88(12), 2044–2051 (2016), DOI: 10.1002/jmv.24678
5. Elfiky, A.A., Elshemey, W.M.: IDX-184 is a superior HCV direct-acting antiviral drug: a QSAR study. *Medicinal Chemistry Research* 25(5), 1005–1008 (2016), DOI: 10.1007/s00044-016-1533-y
6. Ganesan, A., Barakat, K.: Applications of computer-aided approaches in the development of hepatitis C antiviral agents. *Expert Opinion on Drug Discovery* 12(4), 407–425 (2017), DOI: 10.1080/17460441.2017.1291628
7. Google Cloud. <https://cloud.google.com>, accessed: 2020-09-06
8. Gorgulla, C.: Free Energy Methods Involving Quantum Physics, Path Integrals, and Virtual Screenings. Ph.D. thesis, Freie Universität Berlin (2018), DOI: 10.17169/refubium-11597
9. Gorgulla, C., Boeszoermenyi, A., Wang, Z.F., et al.: An open-source drug discovery platform enables ultra-large virtual screens. *Nature* 580(7805), 663–668 (2020), DOI: 10.1038/s41586-020-2117-z
10. Koebel, M.R., Schmadeke, G., Posner, R.G., et al.: AutoDock VinaXB: implementation of XBSF, new empirical halogen bond scoring function, into AutoDock Vina. *Journal of Cheminformatics* 8(1), 27 (2016), DOI: 10.1186/s13321-016-0139-1
11. Koes, D.R., Baumgartner, M.P., Camacho, C.J.: Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. *Journal of Chemical Information and Modeling* 53(8), 1893–1904 (2013), DOI: 10.1021/ci300604z
12. Lal, D.K., Barisal, A., Tripathy, M.: Grey Wolf Optimizer Algorithm Based Fuzzy PID Controller for AGC of Multi-area Power System with TCPS. *Procedia Computer Science* 92, 99–105 (2016), DOI: 10.1016/j.procs.2016.07.329
13. Lyu, J., Wang, S., Balias, T.E., et al.: Ultra-large library docking for discovering new chemotypes. *Nature* 566, 224–229 (2019), DOI: 10.1038/s41586-019-0917-9

14. Mirjalili, S., Mirjalili, S., Lewis, A.: Grey wolf optimizer. *Advances in Engineering Software* 69, 46–61 (2014), DOI: 10.1016/j.advengsoft.2013.12.007
15. Morris, G.M., Huey, R., Lindstrom, W., et al.: AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of Computational Chemistry* 30(16), 2785–2791 (2009), DOI: 10.1002/jcc.21256
16. Nivedha, A.K., Thieker, D.F., Makeneni, S., et al.: Vina-Carb: Improving Glycosidic Angles during Carbohydrate Docking. *Journal of Chemical Theory and Computation* 12(2), 892–901 (2016), DOI: 10.1021/acs.jctc.5b00834
17. O’Boyle, N.M., Banck, M., James, C.A., et al.: Open Babel: An open chemical toolbox. *Journal of Cheminformatics* 3(1), 1–14 (2011), DOI: 10.1186/1758-2946-3-33
18. Papoian, G.A., Wolynes, P.G.: The physics and bioinformatics of binding and folding – an energy landscape perspective. *Biopolymers* 68(3), 333–349 (2003), DOI: 10.1002/bip.10286
19. Schrödinger LLC, New York: Maestro Release 2020-3. <https://www.schrodinger.com/maestro>, accessed: 2020-09-06
20. Shaw, D.E., Deneroff, M.M., Dror, R.O., et al.: Anton, a special-purpose machine for molecular dynamics simulation. *Commun. ACM* 51(7), 91–97 (2008), DOI: 10.1145/1364782.1364802
21. Sterling, T., Irwin, J.J.: ZINC 15 – ligand discovery for everyone. *Journal of Chemical Information and Modeling* 55(11), 2324–2337 (2015), DOI: 10.1021/acs.jcim.5b00559
22. Trott, O., Olson, A.J.: AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry* 31(2), 455–461 (2010), DOI: 10.1002/jcc.21334
23. VirtualFlow. <https://virtual-flow.org/> (2020), accessed: 2020-09-06
24. Wong, K.M., Tai, H.K., Siu, S.W.I.: GWOVina: A grey wolf optimization approach to rigid and flexible receptor docking. *Chemical Biology & Drug Design* (2020), DOI: 10.1111/cbdd.13764
25. Yin, W., Mao, C., Luan, X., et al.: Structural basis for inhibition of the RNA-dependent RNA polymerase from SARS-CoV-2 by remdesivir. *Science* 368(6498), 1499–1504 (2020), DOI: 10.1126/science.abc1560

# Perspectives on Supercomputing and Artificial Intelligence Applications in Drug Discovery

*Jun Xu*<sup>1,2</sup>, *Jiming Ye*<sup>1</sup>

© The Authors 2020. This paper is published with open access at SuperFri.org

This review starts with outlining how science and technology evaluated from last century into high throughput science and technology in modern era due to the Nobel-Prize-level inventions of combinatorial chemistry, polymerase chain reaction, and high-throughput screening. The evolution results in big data accumulated in life sciences and the fields of drug discovery. The big data demands for supercomputing in biology and medicine, although the computing complexity is still a grand challenge for sophisticated biosystems in drug design in this supercomputing era. In order to resolve the real-world issues, artificial intelligence algorithms (specifically machine learning approaches) were introduced, and have demonstrated the power in discovering structure-activity relations hidden in big biochemical data. Particularly, this review summarizes on how people modernize the conventional machine learning algorithms by combing non-numeric pattern recognition and deep learning algorithms, and successfully resolved drug design and high throughput screening issues. The review ends with the perspectives on computational opportunities and challenges in drug discovery by introducing new drug design principles and modeling the process of packing DNA with histones in micrometer scale space, an example of how a macrocosm object gets into microcosm world.

*Keywords: drug discovery, big data, artificial intelligence, HPC.*

## 1. Big Data and Supercomputing Challenges in Drug Discovery

In the last century, three cutting-edge inventions, which were combinatorial chemistry (CC), polymerase chain reaction (PCR), and high-throughput screening (HTS), significantly changed biomedical science and technology. CC was invented by Robert Bruce Merrifield who won 1984 Nobel Prize for Solid Synthesis [26], and made high throughput syntheses (a method for scientific experimentation using robotics, data processing/control software, liquid handling devices, and sensitive detectors allows a researcher to quickly make millions of chemicals for biological tests) become possible [19]. PCR was invented by Kary Banks Mullis who won 1993 Nobel Prize [31], and expedited human gene project. HTS was invented by Donald J. Cram, Jean-Marie Lehn and Charles J. Pedersen, who jointly won 1987 Nobel Prize in chemistry for their development and use of molecules with structure-specific interactions of high selectivity. HTS significantly accelerated screening huge number of compounds against biological targets. These inventions triggered high throughput science and technology and revolutionized pharmaceutical discovery and development. Because people now can make chemical compounds, biopolymers and validate their biological properties in high throughput manner. Consequently, human being is facing big data and supercomputing challenges in modern time.

Drug discovery and development involve in the following major processes: molecular design, biological or chemical syntheses, molecular structural elucidations and pharmaceutical analyses, pharmaceutical target identification and validation, drug screening, preclinic experiments and clinic trials, pharmacokinetics (PK) and pharmacodynamics (PD) analyses, disease diagnoses, and clinic drug applications. Each process involves instrumental measurements that result in big data. These data are not only “big” (volume from GB to PB), but stored in many different formats (variety) and required prompt analyses (velocity).

<sup>1</sup>School of Biotechnology and Health Sciences/Center for Biomedical Data Research, Wuyi University, China

<sup>2</sup>Research Center for Drug Discovery, School of Pharmaceutical Sciences, Sun Yat-Sen University, China

There are mainly four sources contributing to the big data in drug discovery:

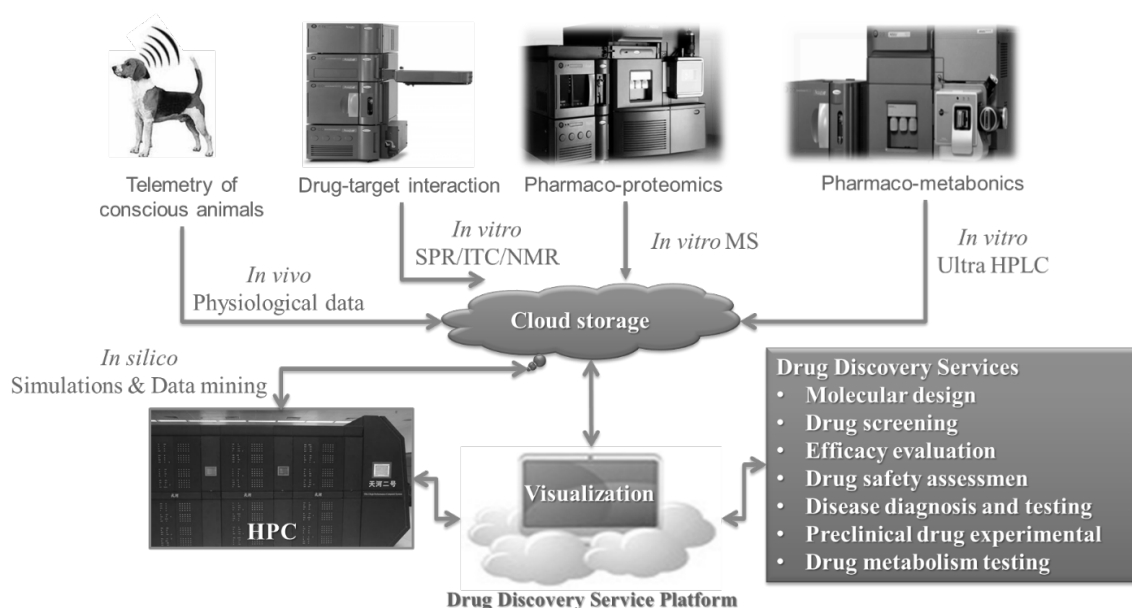
1. **High throughput experiments.** High throughput syntheses can generate many data describing molecular structures and properties and high throughput screening campaigns can generate many data regarding the relations of the compounds and their biological targets.
2. **Health information / office automation.** These resources contain patients information regarding demographic, administrative, health status / risks, medical history, current management of health conditions, and outcomes data.
3. **Scientific publications, patents, and databases.** Publications in life sciences grow rapidly. PubMed collects more than 30 million biomedical articles from more than 7,000 journals; by August 2020, American Chemical Abstracts (ACS) collects more than 100 millionth compound, 64 million gene sequences. The number of US patent applications reaches 15394762 since 1963. Databases ChEMBL, PubChem, ChemSpider, ZINC, and SureChEMBL collect 2, 90, 63, 980, and 17 million compounds [34]. These data cannot be utilized or digested without computational or artificial intelligence approaches.
4. **Simulations.** Along with the increasing computing power, we can simulate greater and more complicated biological systems. Taking molecular dynamics simulation as examples, each nanosecond simulated conformational trajectories resulted in 2 GB data averagely; it would generate millions of conformations ( $\sim 2$  TB data) for micro-second time scale.

These big data bring in following challenges:

1. **Data storage.** Petabyte ( $10^{15}$  bytes) of digital information relies on cloud storage; chemical and biological data annotation/curation and quality assurance are challenging.
2. **Visualization.** Small molecules or biopolymers are described in graphs *per se*. These objects are usually converted into numbers (descriptors). Thus, a molecule is defined as a point in multi-dimensional space that requires dimension reduction approaches (such as principal component analysis (PCA), and nonlinear dimensionality reduction techniques), metadata generation techniques.
3. **Data mining.** Based on the high dimensional data, scientists are facing classification problems. Molecules are classified into two or more clusters corresponding to their phenotypes. Moreover, people need to understand the relations between the key factors/features/chemotypes and a specific phenotype(s). The real challenges are (a) the relations between the features and phenotypes are not of classical analytic function relations; (b) the features for an entire molecule are not related to its phenotypic property in the most of situation; (c) the local feature(s)/substructure(s) for an molecule can be the key to a phenotypic property, but there are uncountable ways to partition a molecular structure into substructures. That is why so many data mining tools have been developed (such as clustering algorithms, decision trees, supporting vector machines (SVM), artificial neural networks (ANN)).
4. **Computational complexity.** The most precise theory to study a molecular system is quantum chemistry. However, the computational complexity of different quantum chemistry algorithms is so difficult that even a quantum computer will be unable to solve [38]. When we deal with a huge number of molecules interacting a protein, the situations are worse. To identify drug targets for a drug lead, multisequence alignment techniques are required. The computational complexity of sequence alignment algorithms ranges from  $O(m * n)$  to  $O(n^2)$  [5]. To identify privileged substructures responsible for a biological activity, sub-

structure match algorithms are applied. The computing complexity of these algorithms are usually polynomial [39].

Traditional drug discovery did not generate big data, however, modern instrumentation and automation changed the situations. With micro-chip technology, people can collect *in vivo* data from model animals 24 hours a day to monitor a drug action *in situ*. New drug discovery technologies such as Surface Plasmon Resonance (SPR, measuring protein-ligand affinity with optics) [18], Isothermal titration calorimetry (ITC, measuring protein-ligand affinity with entropy and enthalpy) [29], and Saturation Transfer Difference NMR spectroscopy (STD-NMR, measuring protein-ligand affinity with nuclear magnetic resonance) [27] allow us to acquire unprecedented protein-ligand interaction data. Omices (such as Pharmaco-proteomics, Pharmaco-metabonomics), High performance computing, and cloud technologies are indispensable components for modern drug discovery service platform (Fig. 1).



**Figure 1.** Drug discovery service platform with HPC and cloud storage

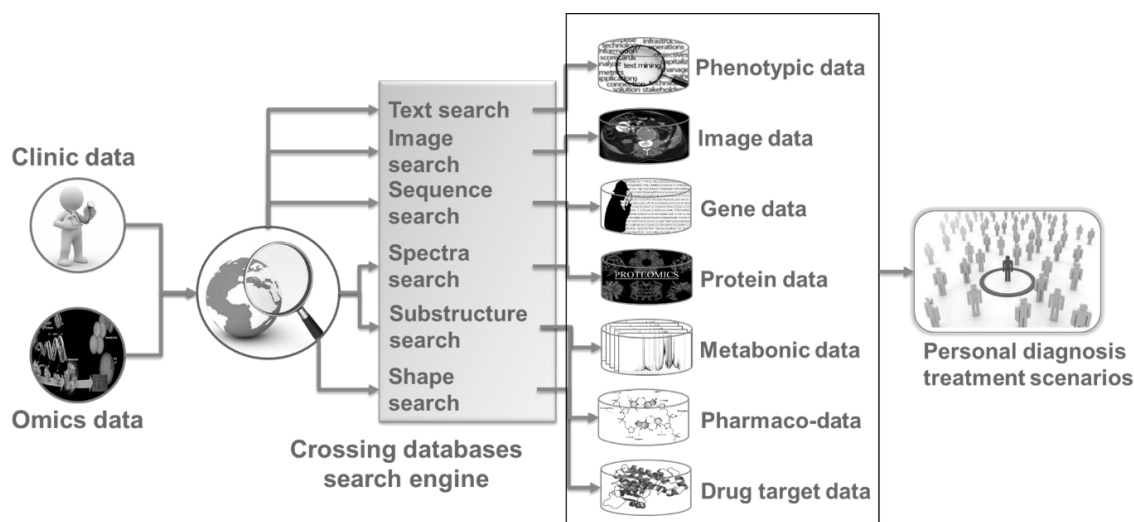
There are millions of available compounds from vendors and billions of virtual compounds for virtual screening. For a structure-based virtual screening approach (such as molecular docking), billions of conformations have to be enumerated for millions molecular structures requiring petabytes (PB) storage. If the docking results were validated by 50 ns MD simulations, each ligand-protein complex would need  $\sim 25$  GB space.

To constantly monitor the pharmaceutical efficacy of a compound *in vivo*, 2–3 months continuous administration will result in 3.5 PB physiological and pharmacological data, from which the efficacy, dosage, and toxicity can be determined.

Constantly monitoring cell changes (such as the effect of drugs on cell activity, the drug distribution, the alter cell behavior, proliferation or apoptosis) while cells incubated with a compound will result in 1 PB data for tracking 10 traits in 1000 cells for 24-hours.

Drug discovery processes involve in the data derived from patients to various devices in many different formats; these data require different search engines and approaches to retrieve and elucidate; and eventually result in personal diagnosis and treatment scenarios (Fig. 2).

Intrinsically, modern drug discovery is to discover macrocosmic solutions by simulation microcosmic phenomena with many experimental data. Therefore, this is a multi-scales simulation



**Figure 2.** Data, search engine and mining tools involved in drug discovery process

process, which covers time scale (from femto-seconds to hours/days), space scale (from nano-meters to meters) at changing resolutions (from electron orbitals to molecular machines) and various theories/methods (from density function theory to biopolymer physics) [11].

Therefore, the computing complexity in drug discovery is due to the complexity of the molecular systems. In many cases, the computing complexity issue can be reduced by parallel computing technology (*aka* high-performance computing) if the problem is parallelizable. For example, employing molecular dynamics-based virtual screening (MDBV), a state of the art HPC can be 600 times faster than an eight-core PC server is in screening a typical drug target (which contains about 40 K atoms). Also, careful design of the GPU/CPU architecture can reduce the HPC costs [15].

A successful virtual drug screening campaign relies on a properly selected compound library. Brutal random virtual screening can lead failure even one has the highest performance computing facility. Therefore, we desperately develop artificial intelligence (AI) applications in pharmaceutical studies.

## 2. Artificial Intelligence and Drug Discovery

The essence of drug discovery is to identify a molecule that interacts its designated biological target from a compound library that have millions of molecules. To do this, we have to understand the relation of molecular structure and activity (SAR). Here, the structure in SAR is actually substructure. A drug molecule can be considered as a molecular machine consists of various functional parts (also termed as substructures, fragments, or chemotypes). How to define the functional parts has been puzzling for many years. Many methods, such as empiric-based method [13], and computational rule based methods [7, 12, 28, 40] were proposed. There is no perfect way to partition substructures from a compound library. Therefore, People also explored other methods, such as molecular descriptors [30], atomic pairs [8], and fingerprints [6].

Conventionally, in order to predict whether a species (for example, a natural substance) has a biological activity, scientists have to extract moieties from the substance, determine chemical structures (represented in topologies, 3D shapes or static surfaces) of the active ingredients; then to covert the chemical structures into a numeric array (called as molecular descriptors or fingerprints). Then, various mathematic models are applied on the data to generate predictive



models. Finally, the models result in the prognosis whether the substance is a candidate to become a drug (Fig. 3).

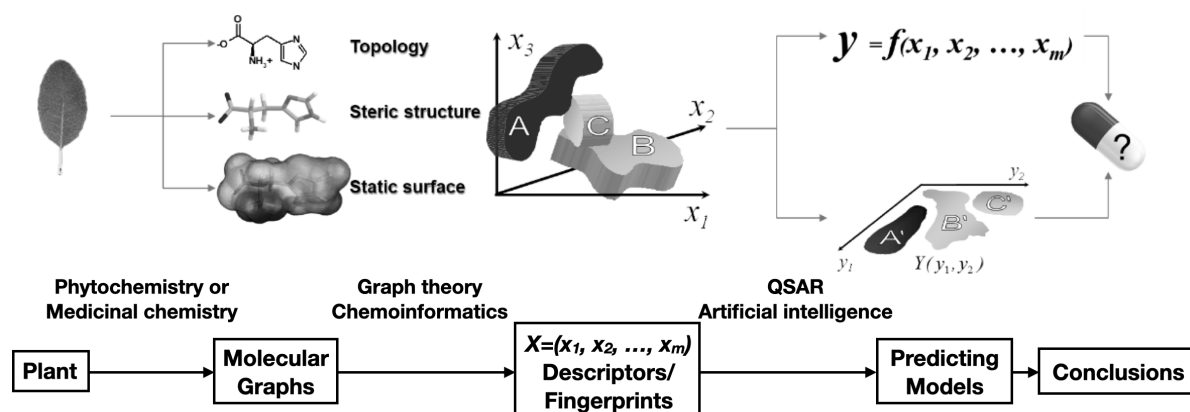


Figure 3. Flow-chart for conventional structure-activity predictions

Molecular structure representations can be converted into various molecular descriptors such as sub-structural fragments, scaffolds, atom pairs (paths), topologic indexes, physical/biological/chemical properties, and fingerprints (bit-maps). The combinations of the descriptors will be figured out based on two principles: (1) a descriptor in the combination has to be significantly associated with the property to be predicted; (2) descriptors within the combination should be orthogonal to each other. Based upon the descriptor combination data, one can build predictive models with learning methods as shown in Fig. 4.

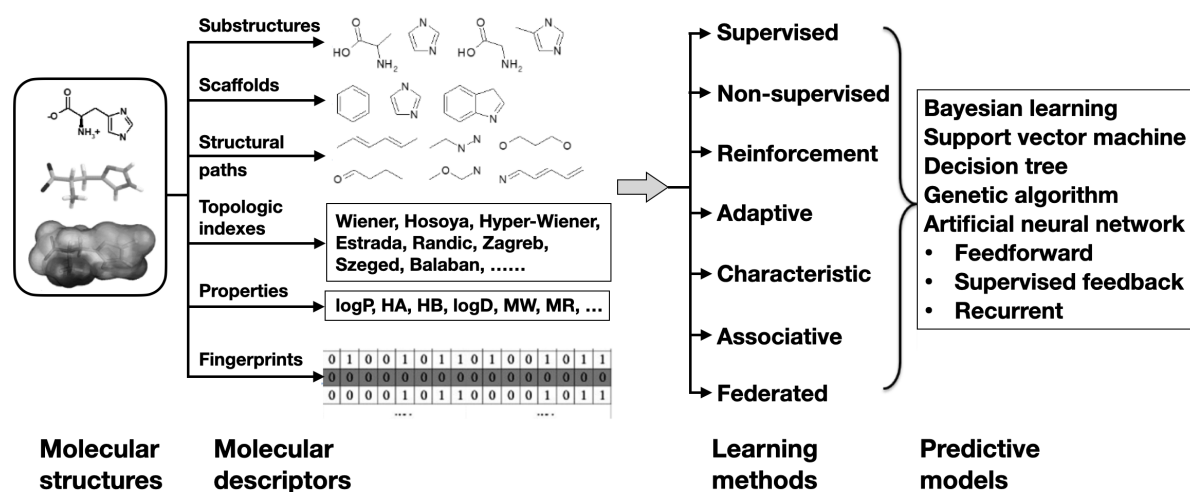


Figure 4. Conventional flow-chart for applying AI methods in predicting pharmaceutical properties for molecules in drug discovery process

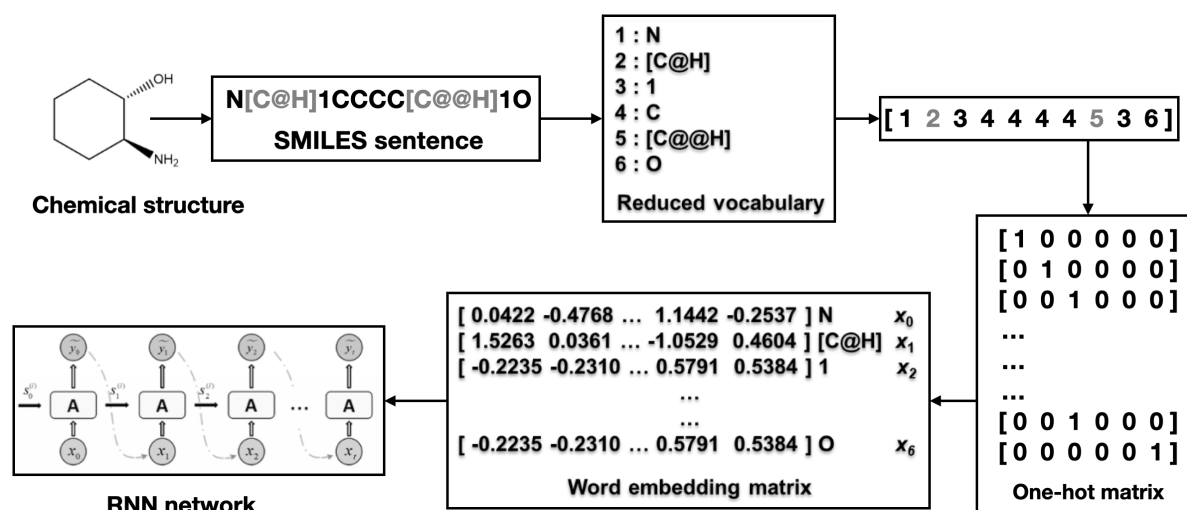
The cores of AI are pattern recognitions that are divided into numeric and non-numeric pattern recognitions. Markush structure or substructure recognitions are non-numeric; self-organizing map (SOM) (*aka* Kohonen network), support vector machine, hierarchical cluster tree, or random forests (*aka* random decision forests) are numerical. The common defect of the conventional machine learning algorithms is that the model performance highly relies on how a modeler selects and combines the molecular descriptors. Unfortunately, there is not rational rules to choose and combine molecular descriptors. In order to make up for this defect, people tried

many approaches, such as rule-embedded naive Bayesian learning [24], multiple machine learning models [23], and combining recursive partitioning with Nave Bayesian learning approaches [35].

Now, people realize that deriving substructures that related to activities from a molecule or molecular library depends on related drug target. In the earlier time of chemoinformatics, a number of molecular structure linear notions were developed due to the lack of computer graphic terminals in that period. Weininger developed the linear notations system called as SMILES (simplified molecular-input line-entry system) that are well accepted internationally [37]. SMILES is an accurate language for molecules, a SMILES notation/sentence precisely describes the atomic connectivity in a molecule. Thus, a compound library can be “written” as an article composed in SMILES sentences. A focused compound library for a specific biological target can be viewed as an article written in SMILES sentences under the same title.

This concept is important because we can derive substructures and activities relations (SAR) without predefining substructures. With deep learning approaches, we can figure out the SAR or predict drug targets with syntax pattern recognition techniques [25].

As shown in Fig. 5, a chemical structure is converted into a SMILES sentence, which is then transformed to a reduced vocabulary, eventually a word embedding matrix is calculated and finally sent to recurrent neural network (RNN) to train a learning model.



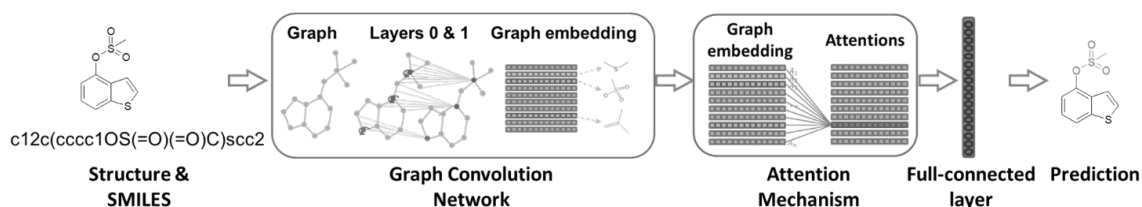
**Figure 5.** Training deep learning models using SMILES without predefining substructures

With self-attention mechanism, structure-activity/property relations (SAR/SPR) can be discovered through chemical linear notation (for example, SMILES) syntax analyses using an interpretable deep learning architecture. The syntax pattern recognition approach has been applied in predicting chemical properties, toxicology, and bioactivity from experimental data sets [2, 3, 9, 10, 17, 36, 44].

With the syntax pattern recognition protocol, drug-like, lead-like, or quasi-biogenic molecules can be proposed by a deep learning program. A quasi-biogenic molecule generator (QBMG) to compose virtual quasi-biogenic compound libraries by means of gated recurrent unit recurrent neural networks has been reported. The library includes stereo-chemical properties, which are crucial features of natural products. QMBG can reproduce the property distribution of the underlying training set, while being able to generate realistic, novel molecules outside of the training set. The proposed compounds were associated with known bioactivities. Therefore,

with a given focused compound library for a biological target, a computer can generate novel compounds that are promising to be active against the target [43].

A property of a molecule can associate with one or more substructures in its structure. For chemical structure stability prediction, if one substructure is found responsible for the instability, it will be enough to conclude the molecule is unstable. A model (DeepChemStable) [22] employing an attention-based graph convolution network based on the COMDECOM data (experimental chemical compound instability data set [45]) was implemented to predict a compound instability. The main advantage of this method is that is an end-to-end model, which does not predefine structural fingerprint features, but instead, dynamically learns structural features and associates the features through the learning process of an attention-based graph convolution network. The previous ChemStable program (with conventional machine learning approach) [24] relied on a rule-based method to reduce the false negatives. DeepChemStable, on the other hand, reduces the risk of false negatives without using a rule-based method minimizing the rate of false negatives, which is a greater concern for instability prediction (Fig. 6).



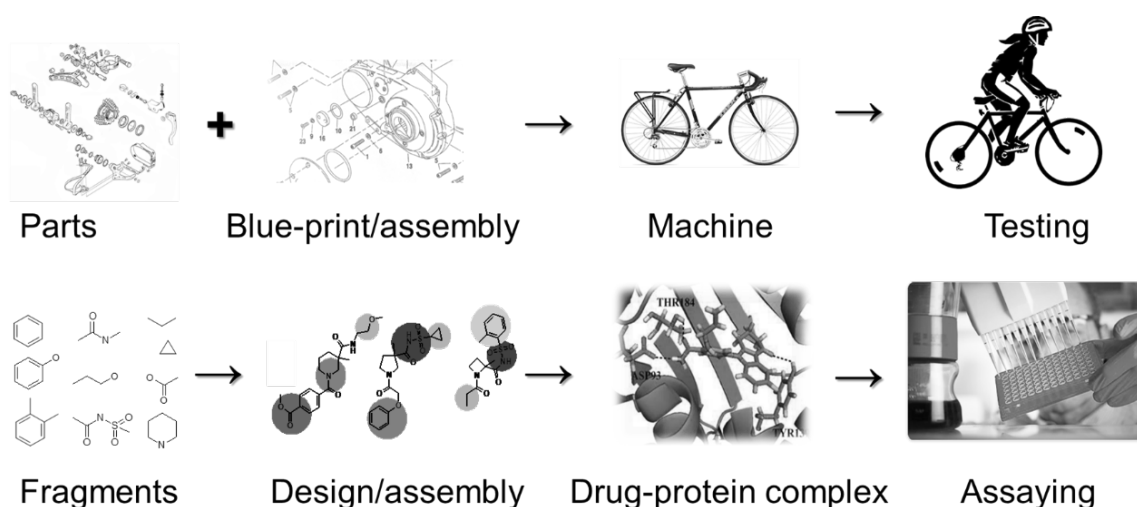
**Figure 6.** Flow-chart of an attention-based graph convolution network that predict a compound chemical instability

Fragment-based drug design (FBDD) [16] gains great achievements these years. Linking fragments to generate a focused compound library for a specific drug target is still puzzling. A program named SyntaLinker that is based on a syntactic pattern recognition with deep conditional transformer neural networks was reported recently. The state-of-the-art transformer links molecular fragments automatically by learning from known structures in medicinal chemistry databases (such as ChEMBL). Linking the fragments was viewed as connecting substructures that were predefined by empirical rules in the past. In SyntaLinker, however, the rules of linking fragments can be learned implicitly from the known chemical structures by recognizing syntactic patterns embedded in SMILES notations. With deep conditional transformer neural networks, SyntaLinker can generate molecular structures based on a given pair of fragments and additional restrictions [41].

Syntactic pattern has also been applied in predicting chemical reaction feasibility. Copper(I)-catalyzed alkyneazide cycloaddition (CuAAC) reaction is a main click chemistry reaction [20] and widely employed in drug discovery. However, the success rate of the CuAAC reaction is not satisfactory as expected. A recurrent neural network (RNN) model was reported to predict its feasibility. Authors designed and synthesized a structurally diverse library of 700 compounds with the CuAAC reaction to obtain experimental data. Then, a bidirectional longshort-term memory with a self-attention mechanism (BiLSTM-SA) model was built. The model achieved total accuracy of 80%. Density functional theory investigations were conducted to provide evidence for the correlation between bromo- $\alpha$ -C hybrid types and the success rate of the reaction [32].

### 3. Perspectives on Computational Opportunities and Challenges in Drug Discovery

The Nobel Prize in Chemistry 2016 was awarded jointly to Jean-Pierre Sauvage, Sir J. Fraser Stoddart and Bernard L. Feringa for the design and synthesis of molecular machines [33]. This can be viewed as an overture for artificial molecular machine era. So far, chemists focus on the mechanical aspects artificial molecular machines [1, 42]. Actually, a drug molecule can also be viewed as an artificial molecular machine that consists of a number of parts (fragments) for regulating biological targets. Thus, the essential questions for drug design methodology becomes (1) what are the fragments for a drug molecule for its target? (2) how to assembly the fragments to make (synthesize) a drug molecule? (3) how to biologically validate the assembled molecules. FBDD, click chemistry (combinatorial chemistry), and HTS are the current answers to these questions respectively. Drug discovery process is similar a machine invention process (Fig. 7).



**Figure 7.** Comparison of processes of a machine discovery and a drug discovery

Drug discovery is much more sophisticated than design and make a machine in macrocosm due to a drug molecule has to regulate even more complicated biological machines in microcosm [14, 21]. The main challenges to a drug designer are: (1) the designed plan for assembling fragments is not necessarily chemically feasible; and (2) the designed molecules against a target is not necessarily functioned as expected. Because most of the mechanisms of actions in life are not well understood to us. Therefore, new drug design approaches and *in silico* experiments are demanding to deal with the big data and computing complexity problems.

Artificial intelligence (AI) techniques will continue to demonstrate their power in drug discovery. Especially, deep learning (DL) techniques have shown the usefulness in deriving SAR from big biochemical data. However, DL assumes the positives and negatives are evenly distributed in a training set, and the number of the samples is big enough. However, typical medicinal chemistry data mainly contain positives with no or minor negatives.

Drug discovery involves multi-scale computation issues. For example, the length of a typical human DNA molecule is about 1.8 meters (visible in macrocosm) has to be tightly packed up to fit in the micro-meter-scale space of cell nucleus (in microcosm). We dont have a convincing theory to explain how a DNA enters microcosm world from macrocosm world with the help of histone proteins. It is a grand computational challenge to generate a model and simulate this

process. Interestingly, recent report claimed that the histones are not just used for packing DNA, they are enzymes that may have helped power eukaryote evolution [4].

## Acknowledgements

The author would like to thank the National Key R&D Program of China (2017YFB0203403), the National Natural Science Foundation of China (81870608), the science and technology program of Guangzhou (201604020109), the Guangdong Provincial Key Lab. of New Drug Design and Evaluation (Grant 2011A060901014) for funding.

*This paper is distributed under the terms of the Creative Commons Attribution-Non Commercial 3.0 License which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is properly cited.*

## References

1. Aprahamian, I.: The future of molecular machines. *ACS Central Science* 6(3), 347–358 (2020), DOI: 10.1021/acscentsci.0c00064
2. Arús-Pous, J., Johansson, S.V., Prykhodko, O., et al.: Randomized smiles strings improve the quality of molecular generative models. *Journal of Cheminformatics* 11(1), 71 (2019), DOI: 10.1186/s13321-019-0393-0
3. Arús-Pous, J., Patronov, A., Bjerrum, E.J., et al.: SMILES-based deep generative scaffold decorator for de-novo drug design. *Journal of Cheminformatics* 12(1), 38 (2020), DOI: 10.1186/s13321-020-00441-8
4. Attar, N., Campos, O.A., Vogelauer, M., et al.: The histone H3-H4 tetramer is a copper reductase enzyme. *Science* 369(6499), 59–64 (2020), DOI: 10.1126/science.aba8740
5. Baichoo, S., Ouzounis, C.A.: Computational complexity of algorithms for sequence comparison, short-read assembly and genome alignment. *Bio Systems* 156-157, 72–85 (2017), DOI: 10.1016/j.biosystems.2017.03.003
6. Banegas-Luna, A.J., Cern-Carrasco, J.P., Pérez-Sánchez, H.: A review of ligand-based virtual screening web tools and screening algorithms in large molecular databases in the age of big data. *Future Medicinal Chemistry* 10(22), 2641–2658 (2018), DOI: 10.4155/fmc-2018-0076
7. Bemis, G.W., Murcko, M.A.: The properties of known drugs. 1. Molecular frameworks. *Journal of Medicinal Chemistry* 39(15), 2887–2893 (1996), DOI: 10.1021/jm9602928
8. Carhart, R.E., Smith, D.H., Venkataraghavan, R.: Atom pairs as molecular features in structure-activity studies: definition and applications. *Journal of Chemical Information and Computer Sciences* 25(2), 64–73 (1985), DOI: 10.1021/ci00046a002
9. Chen, H., Engkvist, O., Wang, Y., et al.: The rise of deep learning in drug discovery. *Drug Discovery Today* 23(6), 1241–1250 (2018), DOI: 10.1016/j.drudis.2018.01.039
10. Chen, J., Cheong, H.H., Siu, S.W.I.: Bestox: A convolutional neural network regression model based on binary-encoded SMILES for acute oral toxicity prediction of chemical

- compounds. In: Martín-Vide, C., Vega-Rodríguez, M.A., Wheeler, T. (eds.) *Algorithms for Computational Biology*. pp. 155–166. Springer International Publishing, Cham (2020), DOI: 10.1007/978-3-030-42266-0\_12
11. Dans, P.D., Walther, J., Gómez, H., Orozco, M.: Multiscale simulation of DNA. *Current Opinion in Structural Biology* 37, 29–45 (2016), DOI: 10.1016/j.sbi.2015.11.011
  12. Dehaspe, L., Toivonen, H., King, R.D.: Finding frequent substructures in chemical compounds. In: Agrawal, R., Stolorz, P.E., Piatetsky-Shapiro, G. (eds.) *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, KDD-98*, 27–31 August 1998, New York City, New York, USA. pp. 30–36. AAAI Press (1998)
  13. Durant, J.L., Leland, B.A., Henry, D.R., Nourse, J.G.: Reoptimization of MDL keys for use in drug discovery. *Journal of Chemical Information and Computer Sciences* 42(6), 1273–1280 (2002), DOI: 10.1021/ci010132r
  14. García-López, V., Chen, F., Nilewski, L.G., et al.: Molecular machines open cell membranes. *Nature* 548(7669), 567–572 (2017), DOI: 10.1038/nature23657
  15. Ge, H., Wang, Y., Li, C., et al.: Molecular dynamics-based virtual screening: Accelerating the drug discovery process by high-performance computing. *Journal of Chemical Information and Modeling* 53(10), 2757–2764 (2013), DOI: 10.1021/ci400391s
  16. Hajduk, P.J., Greer, J.: A decade of fragment-based drug design: strategic advances and lessons learned. *Nature Reviews Drug Discovery* 6(3), 211–219 (2007), DOI: 10.1038/nrd2220
  17. Hirohara, M., Saito, Y., Koda, Y., et al.: Convolutional neural network based on SMILES representation of compounds for detecting chemical motif. *BMC Bioinformatics* 19(19), 526 (2018), DOI: 10.1186/s12859-018-2523-5
  18. Homola, J.: Surface plasmon resonance sensors for detection of chemical and biological species. *Chemical Reviews* 108(2), 462–493 (2008), DOI: 10.1021/cr068107d
  19. Itoh, H., Tokumoto, K., Kaji, T., et al.: Development of a high-throughput strategy for discovery of potent analogues of antibiotic lysocin E. *Nature Communications* 10(1), 2992 (2019), DOI: 10.1038/s41467-019-10754-4
  20. Kolb, H.C., Finn, M.G., Sharpless, K.B.: Click chemistry: Diverse chemical function from a few good reactions. *Angewandte Chemie International Edition* 40(11), 2004–2021 (2001), DOI: 10.1002/1521-3773(20010601)40:11;2004::AID-ANIE2004j3.0.CO;2-5
  21. Lancia, F., Ryabchun, A., Katsonis, N.: Life-like motion driven by artificial molecular machines. *Nature Reviews Chemistry* 3(9), 536–551 (2019), DOI: 10.1038/s41570-019-0122-2
  22. Li, X., Yan, X., Gu, Q., et al.: DeepChemStable: Chemical stability prediction with an attention-based graph convolution network. *Journal of Chemical Information and Modeling* 59(3), 1044–1049 (2019), DOI: 10.1021/acs.jcim.8b00672
  23. Li, Y., Wang, L., Liu, Z., et al.: Predicting selective liver X receptor beta agonists using multiple machine learning methods. *Mol Biosyst* 11(5), 1241–1250 (2015), DOI: 10.1039/c4mb00718b

24. Liu, Z., Zheng, M., Yan, X., et al.: ChemStable: a web server for rule-embedded naïve Bayesian learning approach to predict compound stability. *Journal of Computer-Aided Molecular Design* 28(9), 941–950 (2014), DOI: 10.1007/s10822-014-9778-3
25. Mayr, A., Klambauer, G., Unterthiner, T., et al.: Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chem. Sci.* 9(24), 5441–5451 (2018), DOI: 10.1039/C8SC00148K
26. Merrifield, R.B.: Solid phase synthesis (Nobel lecture). *Angewandte Chemie International Edition in English* 24(10), 799–810 (1985), DOI: 10.1002/anie.198507993
27. Meyer, B., Peters, T.: NMR spectroscopy techniques for screening and identifying ligand binding to protein receptors. *Angewandte Chemie International Edition* 42(8), 864–890 (2003), DOI: 10.1002/anie.200390233
28. Peng, H., Liu, Z., Yan, X., et al.: A de novo substructure generation algorithm for identifying the privileged chemical fragments of liver X receptorbeta agonists. *Scientific Reports* 7(1), 11121 (2017), DOI: 10.1038/s41598-017-08848-4
29. Rajarathnam, K., Rösger, J.: Isothermal titration calorimetry of membrane proteins – progress and challenges. *Biochimica et Biophysica Acta (BBA) – Biomembranes* 1838(1, Part A), 69–77 (2014), DOI: 10.1016/j.bbamem.2013.05.023
30. Sahoo, S., Adhikari, C., Kuanar, M., Mishra, B.K.: A short review of the generation of molecular descriptors and their applications in quantitative structure property/activity relationships. *Current Computer-Aided Drug Design* 2(3), 181–205 (2016), DOI: 10.2174/1573409912666160525112114
31. Saiki, R.K., Bugawan, T.L., Horn, G.T., et al.: Analysis of enzymatically amplified beta-globin and HLA-DQ alpha DNA with allele-specific oligonucleotide probes. *Nature* 324(6093), 163–166 (1986), DOI: 10.1038/324163a0
32. Su, S., Yang, Y., Gan, H., et al.: Predicting the feasibility of copper(i)-catalyzed alkyne–azide cycloaddition reactions using a recurrent neural network with a self-attention mechanism. *Journal of Chemical Information and Modeling* 60(3), 1165–1174 (2020), DOI: 10.1021/acs.jcim.9b00929
33. Van Noorden, R., Castelvechi, D.: World’s tiniest machines win chemistry Nobel. *Nature* 538(7624), 152–153 (2016), DOI: 10.1038/nature.2016.20734
34. Walters, W.P.: Virtual chemical libraries. *Journal of Medicinal Chemistry* 62(3), 1116–1124 (2019), DOI: 10.1021/acs.jmedchem.8b01048
35. Wang, L., Chen, L., Liu, Z., et al.: Predicting mTOR inhibitors with a classifier using recursive partitioning and naïve Bayesian approaches. *PLOS ONE* 9(5), 1–15 (2014), DOI: 10.1371/journal.pone.0095221
36. Wang, S., Guo, Y., Wang, Y., et al.: SMILES-BERT: Large scale unsupervised pre-training for molecular property prediction. In: *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. pp. 429–436. BCB ’19, Association for Computing Machinery, New York, NY, USA (2019), DOI: 10.1145/3307339.3342186

37. Weininger, D.: SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences* 28(1), 31–36 (1988), DOI: 10.1021/ci00057a005
38. Whitfield, J.D., Love, P.J., Aspuru-Guzik, A.: Computational complexity in electronic structure. *Phys. Chem. Chem. Phys.* 15(2), 397–411 (2013), DOI: 10.1039/C2CP42695A
39. Xu, J.: GMA: A generic match algorithm for structural homomorphism, isomorphism, and maximal common substructure match and its applications. *Journal of Chemical Information and Computer Sciences* 36(1), 25–34 (1996), DOI: 10.1021/ci950061u
40. Xu, J.: A new approach to finding natural chemical structure classes. *Journal of Medicinal Chemistry* 45(24), 5311–5320 (2002), DOI: 10.1021/jm010520k
41. Yang, Y., Zheng, S., Su, S., et al.: Syntalinker: automatic fragment linking with deep conditional transformer neural networks. *Chem. Sci.* 11(31), 8312–8322 (2020), DOI: 10.1039/D0SC03126G
42. Zhang, L., Marcos, V., Leigh, D.A.: Molecular machines with bio-inspired mechanisms. *Proceedings of the National Academy of Sciences* 115(38), 9397–9404 (2018), DOI: 10.1073/pnas.1712788115
43. Zheng, S., Yan, X., Gu, Q., et al.: QBMG: quasi-biogenic molecule generator with deep recurrent neural network. *Journal of Cheminformatics* 11(1), 5 (2019), DOI: 10.1186/s13321-019-0328-9
44. Zheng, S., Yan, X., Yang, Y., Xu, J.: Identifying structure–property relationships through SMILES syntax analysis with self-attention mechanism. *Journal of Chemical Information and Modeling* 59(2), 914–923 (2019), DOI: 10.1021/acs.jcim.8b00803
45. Zitha-Bovens, E., Maas, P., Wife, D., et al.: Comdecom: predicting the lifetime of screening compounds in DMSO solution. *J Biomol Screen* 14(5), 557–565 (2009), DOI: 10.1177/1087057109336953



# Computational Modeling of the SARS-CoV-2 Main Protease Inhibition by the Covalent Binding of Prospective Drug Molecules

*Alexander V. Nemukhin*<sup>1,2</sup>, *Bella L. Grigorenko*<sup>1,2</sup>, *Igor V. Polyakov*<sup>1,2</sup>,  
*Sofya V. Lushchekina*<sup>2</sup>

© The Authors 2020. This paper is published with open access at SuperFri.org

We illustrate modern modeling tools applied in the computational design of drugs acting as covalent inhibitors of enzymes. We take the Main protease (M<sup>Pro</sup>) from the SARS-CoV-2 virus as an important present-day representative. In this work, we construct a compound capable to block M<sup>Pro</sup>, which is composed of fragments of antimalarial drugs and covalent inhibitors of cysteine proteases. To characterize the mechanism of its interaction with the enzyme, the algorithms based on force fields, including molecular mechanics (MM), molecular dynamics (MD) and molecular docking, as well as quantum-based approaches, including quantum chemistry and quantum mechanics/molecular mechanics (QM/MM) methods, should be applied. The use of supercomputers is indispensably important at least in the latter approach. Its application to enzymes assumes that energies and forces in the active sites are computed using methods of quantum chemistry, whereas the rest of protein matrix is described using conventional force fields. For the proposed compound, containing the benzoisothiazolone fragment and the substitute at the uracil ring, we show that it can form a stable covalently bound adduct with the target enzyme, and thus can be recommended for experimental trials.

*Keywords:* SARS-CoV2 main protease, QM/MM, molecular docking, molecular dynamics, covalent inhibitor.

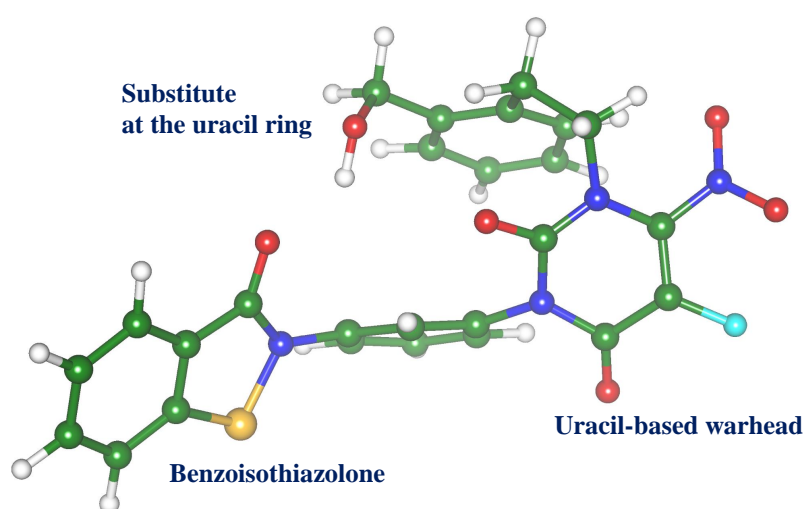
## Introduction

Under the current public health emergency caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), it is imperative to propose prospective drugs for medical treatment of COVID-19. The cysteine protease, called the Main protease (M<sup>Pro</sup>), from the SARS-CoV-2 is considered as one of the targets in current research [11]. The goal is to inhibit the enzyme and to prevent its function required for viral replication and transcription. In this paper, we describe practical steps in computational modeling of the mechanism of M<sup>Pro</sup> inhibition by a chemical compound capable to form a covalent bond with the catalytic amino acid residue cysteine in the active site of the enzyme, which causes the inhibition. This approach differs from a common strategy to propose inhibiting molecules, which form non-covalently bound complexes with an enzyme, thus blocking the entrance to the active site for natural substrates. From the computational side, the latter strategy requires an application of algorithms of molecular docking and classical molecular dynamics, which are successfully used in high performance computer simulations [12]. If our aim is to predict a covalent inhibitor, a broader range of computational tools is necessary, including quantum chemistry based methods [1], which require supercomputer resources. These approaches are discussed in the Methodology section. A specific drug candidate capable to inhibit SARS-CoV-2 M<sup>Pro</sup> by the covalent binding is proposed in this work by the following reasons. The search of prospective candidates is often based on screening the data bases of drugs already approved for treatment of other diseases in attempts to find a new application for a corresponding chemical compound. In particular, the known antimalarial drugs

<sup>1</sup>Department of Chemistry, Lomonosov Moscow State University, Moscow, Russia

<sup>2</sup>Emanuel Institute of Biochemical Physics, Russian Academy of Sciences, Moscow, Russia

are among those actively investigated to fight COVID-19. At preliminary steps, we attempted to attach the molecule of the famous hydroxychloroquine compound to  $M^{Pro}$ . This substance is widely known, for example, from mass media communications. Our studies, including docking and molecular dynamics simulations, did not show positive trends in respect of its interaction with  $M^{Pro}$ . More promising species were found among other proposed antimalarial drugs with the benzothiazolone moiety [10]. Some of these compounds could be docked to  $M^{Pro}$  with reasonable binding energies. Second, a recent study of covalent inhibitors of another cysteine protease [5] suggests a set of molecules with the building blocks, which can be efficient covalent inhibitors of  $M^{Pro}$ . Correspondingly, we designed computationally a compound by joining a benzothiazolone fragment, which is capable to dock at  $M^{Pro}$ , with a moiety capable to interact with the active site of  $M^{Pro}$  forming a stable adduct. A molecular model of this compound is shown in Fig. 1.



**Figure 1.** Molecular model of the compound considered as a covalent inhibitor of SARS-CoV-2  $M^{Pro}$ . The uracil-based warhead is responsible for the chemical reaction with the enzyme, the benzothiazolone fragment and the substitute at the uracil ring provide the enhanced binding to the protein. Green – carbon atoms, red – oxygen, blue – nitrogen, yellow – sulfur, cyan – fluorine

It should be pointed out that actually a set of molecules can be derived on the base of the used template, namely, a construct from the benzothiazolone moiety and cysteine protease inhibitors, which differ by substitutes at the uracil warhead. Below, we describe the computational approaches, which are utilized in modeling mechanisms of the covalent inhibition of  $M^{Pro}$  taking this particular compound, but these approaches are similar for the entire series of prospective drugs.

## 1. Methodology

The following steps should be performed to predict computationally a covalent inhibitor of an enzyme.

**Step 1.** One should construct molecular models of the compound and the protein. To design a molecule of the type shown in Fig. 1, computer programs of molecular mechanics (MM) and quantum chemistry (QC) are the major tools. Typically, organic molecules of prospective drugs contain up to a hundred of atoms, and MM-based constructors do not require expensive

calculations. There are a plenty of algorithms and computer programs employing conventional force field parameters, which allow one to create a preliminary model. If necessary, equilibrium geometry parameters of the molecule and the atomic charges can be cleaned in QC calculations. Although a high accuracy of these parameters is not expected in this application, and expensive QC algorithms can be avoided, nevertheless, supercomputer resources can be requested if a problem of multiple conformations of polyatomic flexible molecule should be solved. Initial coordinates of an enzyme macromolecule usually are taken from a suitable crystal structure deposited to the Protein Data Bank [2]. To convert these raw materials to a relevant molecular model one should add hydrogen atoms, which are not recognized in crystallography experiments, and often restore missing amino acid residues or correct artificial molecular groups needed to perform measurements. Also, the protein molecule should be surrounded by a proper amount of solvent water molecules. These actions are carried out using either MM-based computer programs, or classical molecular dynamics (MD) simulations. The latter can require supercomputer resources because of an enormous amount of atoms in a model molecular system.

**Step 2.** Molecular docking is one of the most popular molecular modeling techniques, particularly wide spread in drug design. While in certain cases docking can be performed using inexpensive equipment, the use of the corresponding programs for virtual screening of large databases containing many thousands of ligands requires supercomputer resources. Moreover, for flexible ligands with many torsion degrees of freedom (like, for instance for the compound shown in Fig. 1), the multi-processor performance of the docking program at several dozen or hundred computing cores of a supercomputer is important [8, 12]. Therefore, molecular docking of the selected compound to the enzyme surface is a major step in characterization of non-covalent inhibitors and it constitutes a preliminary step in the design of covalent inhibitors. These algorithms are very helpful in finding location of the compound near the enzyme active site to obtain a first approach to construct an enzyme-substrate (ES) complex. Usually, the ES complex, which is required in subsequent calculations of reaction profiles, differs considerably from a hint in docking experiments; however, analysis of such obtained protein-ligand configurations is practically important.

**Step 3.** Making and breaking chemical bonds, which accompany formation of covalently bound adducts in modeling of covalent inhibition of enzymes, cannot be described at the level of force fields and require application of quantum-based algorithms. The most suitable approach, the hybrid quantum mechanics / molecular mechanics (QM/MM) method, was proposed almost half-century ago. Its application to enzymes assumes that energies and forces in the active sites are computed using methods of quantum chemistry, whereas the rest of protein matrix is described using conventional force fields. Since first applications, this methodology developed rapidly and presently several computer packages can be used for QM/MM calculations. For instance, we make a reference to the work [3], in which the entire reaction energy profile in enzyme catalysis from the ES complex to the enzyme-product (EP) complex was evaluated using QM/MM. In our case, the ES complex corresponds to the structure optimized in QM/MM calculations initiated by the atomic coordinates preliminary obtained in molecular docking and classical molecular dynamics simulations. The optimization means that all geometry parameters of the complex corresponding to the minimum energy point on the potential energy surface are evaluated using the QM/MM approach, which requires multiple calculations of the energy and energy gradients with respect to coordinates of nuclei. If reliable QM/MM approaches are applied, including a reasonable size of the QM-subsystem, trustworthy functionals in the density

functional theory (DFT), and well-defined basis sets to represent molecular orbitals, then the use of supercomputer resources is inevitable. Next, a reaction coordinate should be specified to describe the pathway from ES to EP. Multiple series of constrained QM/MM minimizations, i.e. optimizations of all structural parameters for each value of the reaction coordinate should be carried out to find the points of reaction intermediates and transition states. Needless to say, that those are extremely time- and resource-consuming calculations. In our case, an important quantity is the energy difference between the ES and EP points, which indicates, how stable is the covalently bound adduct, i.e. the product of the interaction of a possible drug molecule with an enzyme.

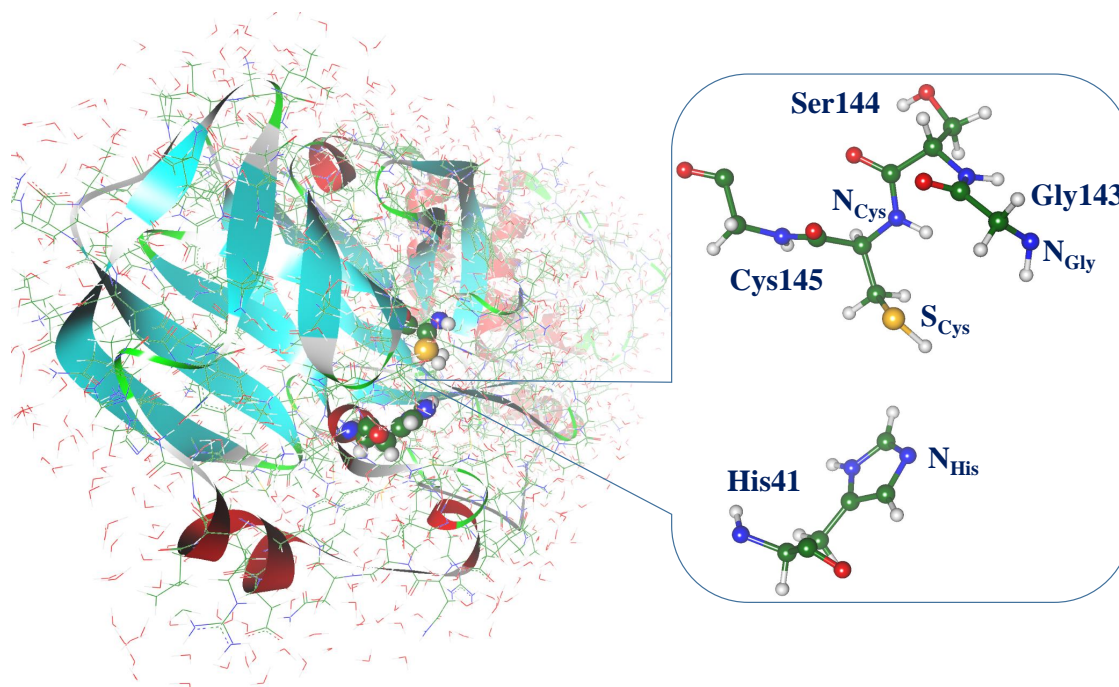
## 2. Implementation for the M<sup>pro</sup> Covalent Inhibition

This paper describes all steps listed in the Methodology section for the case of the covalent inhibition of the SARS-CoV-2 M<sup>pro</sup>.

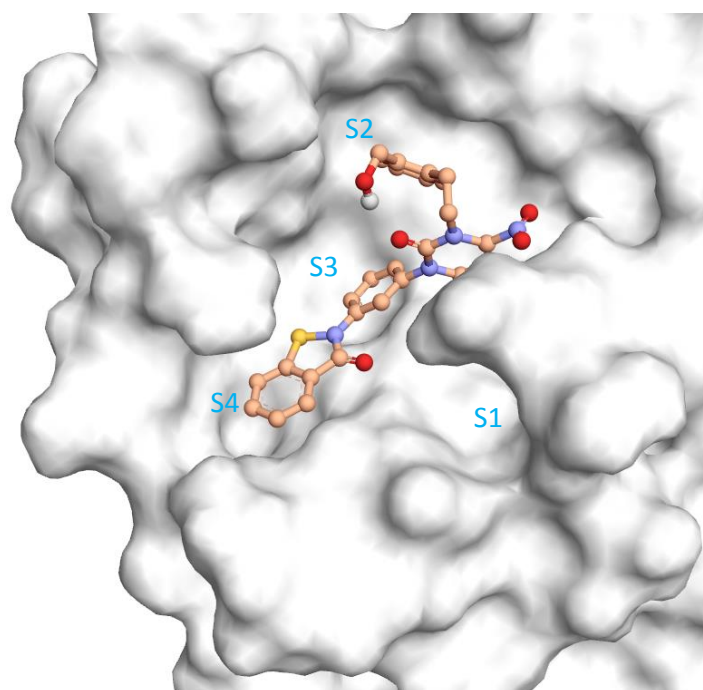
**Step 1.** The compound shown in Fig. 1 was constructed using the Discovery Studio Visualizer MM-based program. QC calculations at the DFT PBE0/6-31G\* level using the NWChem program [13] allowed us to correct few geometry parameters of the molecule. Initial set of atomic coordinates was taken from the structure of the ligand-free SARS-Cov-2 M<sup>pro</sup> deposited to the Protein Data Bank [2] as the 6yb7 entry. Restoration of the three-dimensional all-atom molecular model was preformed using the MM and MD tools. Solvent water molecules were added using the Visual Molecular Dynamics (VMD) package [4] to prepare a solvent box border at least 1000 pm from the protein. Relaxation of the structure was performed in MD simulations using the NAMD program [9] with the CHARMM36 force field [14] for the protein and TIP3P for water molecules with the NPT ensemble at 298 K temperature and 1 atm pressure. Positions of all atoms found in the X-ray structure, including oxygen atoms of crystallographic water molecules, were fixed. First, added solvent shells were equilibrated during 1 ns MD simulation, until cell volume stabilized. Next, the system was minimized during 1000 steps. In QM/MM calculations the solvent shell was reduced to the distance of 600 pm from the protein. Figure 2 illustrates a ligand-free model protein system showing the enzyme active site in the inset. The mechanism of this enzyme includes the nucleophilic attack of S<sub>Cys</sub> on the carbon atom of a substrate coupled with the proton transfer from the S-H group of Cys145 to the nitrogen atom N<sub>His</sub> of His41. The inset in the right part in Fig. 2 shows the catalytic dyad, Cys145 and His41, and the groups from the 143-145 chain forming the oxyanion hole, N<sub>Cys-H</sub> and N<sub>Gly-H</sub>.

**Step 2.** The active site of M<sup>pro</sup> spans far beyond the catalytic site illustrated in Fig. 2. Four subsites (pockets) S1–S4 are distinguished on the enzyme surface [16]. Hence the molecular docking grid box included the whole area with the all these pockets. This requires 80 grid points in each direction with grid spacing 375 pm (30000 pm × 30000 pm × 30000 pm). Affinity maps for this grid box were prepared with Autogrid 4.2.6 tool of AutoDock suite [8]. Considering an amount of torsional degrees of freedom of the considered compound and a size of the grid box, the following parameters of the main Lamarckian Genetic Algorithm [7] were used: 256 runs, 25×106 evaluations, 27×104 generations, and a population size of 3000. Docking was performed with AutoDock 4.2.6 program [8]. The protein was fully rigid during molecular docking. Analyzing the obtained docked poses, a particular attention was payed to the distance between the carbon atom of the inhibitor and the sulfur atom of Cys145 of the enzyme, i.e. the nucleophilic attack distance. Position with the shortest distance of 450 pm was in TOP-5 poses, ranged by estimated binding free energies, difference with them was within 0.2 kcal/mol and the binding

energy was equal to  $-9.4$  kcal/mol. Thus, we conclude that there are several binding poses competing with each other, and one of them is productive. With the warhead (see Fig. 2) in the active site, an inhibitor occupies S2, S3 and S4 pockets of the active site (Fig. 3).



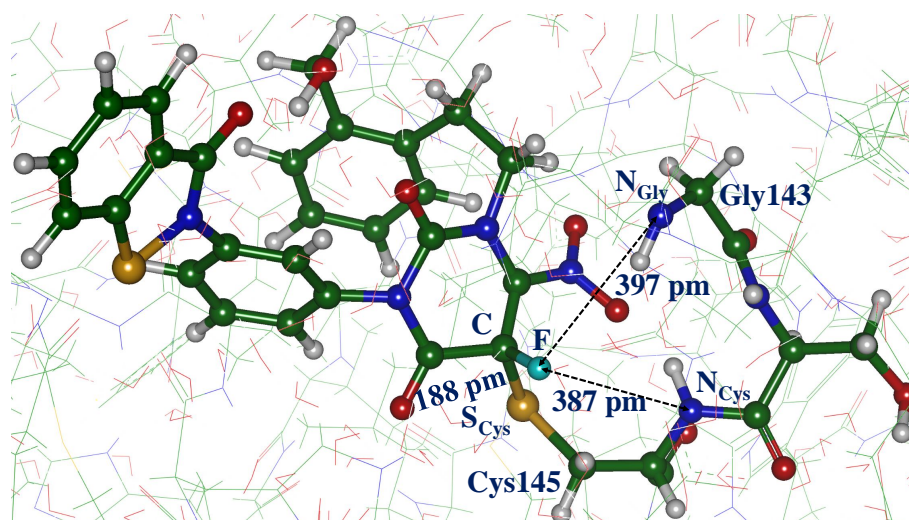
**Figure 2.** A model system for the SARS-CoV-2 M<sup>pro</sup>. Left: a view on the ligand-free protein. Right: the substrate-free active site. Green – carbon atoms, red – oxygen, blue – nitrogen, yellow – sulfur



**Figure 3.** Position of the inhibitor on the M<sup>pro</sup> according to the docking studies

**Step 3.** QM/MM calculations of the reaction energy profile were carried out on the “Lomonosov-2” supercomputer [15]. The NWChem software package [13] was compiled man-

ually with the Intel<sup>®</sup> Parallel Studio XE 2019 Update 5 and linked with Intel<sup>®</sup> Math Kernel Library and OpenMPI 4.0.1 message-passing library. Our previous experience benchmarking the NWChem [6] proved that a storage-based approach was much faster than the “direct” self-consistent field (SCF) procedure. The only disadvantage is that one needs an adequate amount of nodes to hold the intermediate data (mostly, the two-electron integrals) in the fast memory (RAM in this case, the nodes are diskless). Hence, we used 8 nodes per task to be able to hold all the data in RAM and achieve a proper performance. For example, a typical calculation with 1250 basis set functions, 50 optimization cycles (each consists of 5 steps in the quantum subsystem and 100 classical steps) takes around 12 hours of wall time or 1344 CPU×hours. In total, it took over 150000 CPU×hours to compute the entire reaction energy profile. The selected reaction coordinate is a distance between  $S_{Cys}$  of the catalytic Cys145 of the enzyme and the carbon atom covalently bound to the fluorine atom in the inhibitor molecule (Fig. 1). Figure 4 illustrates the QM/MM optimized structure of the stable reaction intermediate on the reaction route with the covalent bond between the inhibitor molecule and catalytic cysteine Cys145 of the enzyme. The adduct is tightly bound in the protein matrix. The C- $S_{Cys}$  distance is 188 pm corresponding to the newly formed chemical bond. The F-N distances (387 pm and 397 pm) evidence that the adduct enters the oxyanion hole.



**Figure 4.** Stable reaction intermediate formed upon the interaction of the inhibitor with the  $M^{Pro}$ . Green – carbon atoms, red – oxygen, blue – nitrogen, yellow – sulfur, cyan – fluorine

The reaction pathway includes the points of the enzyme-substrate (ES) complex, reaction intermediate (INT) and the enzyme-product (EP) complex. The ES complex is a non-covalent complex of a substrate in the active site of an enzyme; the reaction intermediate is a transient covalent complex of these two species and the EP is the final stable covalent complex. By using the DFT approach PBE0/6-31G\* in the QM-subsystem and the AMBER force field parameters in the MM-subsystem we obtain that the energy of the reaction intermediate shown in Fig. 4 is lower than the level of ES by about 15 kcal/mol, whereas the enzyme-product (EP) complex, separated from ES by energy barriers of about 10 kcal/mol, is lower than the level of ES by about 19 kcal/mol. These energy gaps are large enough to conclude that the covalently bound adduct is a fairly stable complex, and the enzyme cannot function after capturing the inhibitor. This indicates that the goal of simulations is achieved, and the computationally designed compound is able to block  $M^{Pro}$ .

## Conclusion

We show that the computationally predicted compound (Fig. 1) created by motifs of the known antimalarial drug molecules and cysteine protease inhibitors can be considered as a prospective drug capable to inhibit the critical enzyme M<sup>Pro</sup> from the SARS-CoV-2 virus. Molecular dynamics and molecular docking approaches evidence that this molecule can be attached to the surface of the enzyme and obstruct delivery of natural substrates. A more intriguing option is a covalent binding of the compound to the catalytic cysteine residue from the M<sup>Pro</sup> active site. Results of high performance QM/MM calculations show that the energy of the enzyme-product complex is about 19 kcal/mol lower than the level of the enzyme-substrate complex, which leads to complete blocking the enzyme.

## Acknowledgements

We thank Dr. Maria G. Khrenova for valuable help. This work was supported by the Russian Foundation for Basic Research (project № 19-03-00043). This research is carried out using the equipment of the shared research facilities of HPC computing resources at Lomonosov Moscow State University.

*This paper is distributed under the terms of the Creative Commons Attribution-Non Commercial 3.0 License which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is properly cited.*

## References

1. Arodola, O.A., Soliman, M.E.: Quantum mechanics implementation in drug-design workflows: does it really help? *Drug Design, Development and Therapy* 11, 2551–2564 (2017), DOI: 10.2147/DDDT.S126344
2. Berman, H.M.: The Protein Data Bank. *Nucleic Acids Research* 28(1), 235–242 (2000), DOI: 10.1093/nar/28.1.235
3. Grigorenko, B.L., Khrenova, M.G., Nilov, D.K., Nemukhin, A.V., Švedas, V.K.: Catalytic cycle of penicillin acylase from *Escherichia coli*: QM/MM modeling of chemical transformations in the enzyme active site upon penicillin G hydrolysis. *ACS Catalysis* 4(8), 2521–2529 (2014), DOI: 10.1021/cs5002898
4. Humphrey, W., Dalke, A., Schulten, K.: VMD: Visual molecular dynamics. *Journal of Molecular Graphics* 14(1), 33–38 (1996), DOI: 10.1016/0263-7855(96)00018-5
5. Klein, P., Johe, P., Wagner, A., et al.: New cysteine protease inhibitors: Electrophilic (het)arenes and unexpected prodrug identification for the *Trypanosoma* protease rhodesain. *Molecules* 25(6), 1451 (2020), DOI: 10.3390/molecules25061451
6. Mironov, V.A., Grigorenko, B.L., Polyakov, I.V., Nemukhin, A.V.: Benchmarking quantum chemistry methods in calculations of electronic excitations. *Supercomputing Frontiers and Innovations* 5(4), 62–66 (2018), DOI: 10.14529/jsfi180405
7. Morris, G.M., Goodsell, D.S., Halliday, R.S., et al.: Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *Jour-*

- nal of Computational Chemistry 19(14), 1639–1662 (1998), DOI: 10.1002/(SICI)1096-987X(19981115)19:14<1639::AID-JCC10>3.0.CO;2-B
8. Morris, G.M., Huey, R., Lindstrom, W., et al.: AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of Computational Chemistry* 30(16), 2785–2791 (2009), DOI: 10.1002/jcc.21256
  9. Phillips, J.C., Braun, R., Wang, W., et al.: Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry* 26(16), 1781–1802 (2005), DOI: 10.1002/jcc.20289
  10. Price, K.E., Armstrong, C.M., Imlay, L.S., et al.: Molecular mechanism of action of antimalarial benzoisothiazolones: species-selective inhibitors of the Plasmodium spp. MEP pathway enzyme, IspD. *Scientific Reports* 6(1), 36777 (2016), DOI: 10.1038/srep36777
  11. Sivasankarapillai, V.S., Pillai, A.M., Rahdar, A., et al.: On facing the SARS-CoV-2 (COVID-19) with combination of nanomaterials and medicine: Possible strategies and first challenges. *Nanomaterials* 10(5), 852 (2020), DOI: 10.3390/nano10050852
  12. Sulimov, A., Kutov, D., Sulimov, V.: Supercomputer docking. *Supercomputing Frontiers and Innovations* 6(3), 26–50 (2019), DOI: 10.14529/jsfi190302
  13. Valiev, M., Bylaska, E., Govind, N., et al.: NWChem: A comprehensive and scalable open-source solution for large scale molecular simulations. *Computer Physics Communications* 181(9), 1477–1489 (2010), DOI: 10.1016/j.cpc.2010.04.018
  14. Vanommeslaeghe, K., MacKerell, A.D.: Charmm additive and polarizable force fields for biophysics and computer-aided drug design. *Biochimica et Biophysica Acta* 1850(5), 861–871 (2015), DOI: 10.1016/j.bbagen.2014.08.004
  15. 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), DOI: 10.14529/jsfi190201
  16. Zhang, L., Lin, D., Sun, X., et al.: Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved alpha-ketoamide inhibitors. *Science* 368(6489), 409–412 (2020), DOI: 10.1126/science.abb3405



# Computational Characterization of the Substrate Activation in the Active Site of SARS-CoV-2 Main Protease

*Maria G. Khrenova*<sup>1,2</sup>, *Vladimir G. Tsirelson*<sup>3,4</sup>, *Alexander V. Nemukhin*<sup>1,5</sup>

© The Authors 2020. This paper is published with open access at SuperFri.org

Molecular dynamics simulations with the QM(DFT)/MM potentials are utilized to discriminate between reactive and nonreactive complexes of the SARS-CoV-2 main protease and its substrates. Classification of frames along the molecular dynamic trajectories is utilized by analysis of the 2D maps of the Laplacian of electron density. Those are calculated in the plane formed by the carbonyl group of the substrate and a nucleophilic sulfur atom of the cysteine residue that initiates enzymatic reaction. Utilization of the GPU-based DFT code allows fast and accurate simulations with the hybrid functional PBE0 and double-zeta basis set. Exclusion of the polarization functions accelerates the calculations 2-fold, however this does not describe the substrate activation. Larger basis set with d-functions on heavy atoms and p-functions on hydrogen atoms enables to disclose equilibrium between the reactive and nonreactive species along the MD trajectory. The suggested approach can be utilized to choose covalent inhibitors that will readily interact with the catalytic residue of the selected enzyme.

*Keywords:* SARS-CoV-2 main protease, QM/MM MD, GPU-accelerated algorithms, substrate activation.

## Introduction

The combined quantum mechanics/molecular mechanics (QM/MM) approach is a proper tool to study chemical reactions in the active sites of enzymes. It allows considering chemical reaction at the QM level taking into account the electrostatic field and steric constraints coming from the rest of the protein and polar water solvent. Reliable results can be obtained only if a relatively high level of theory for the QM subsystem is applied. It was demonstrated that utilization of the density functional theory (DFT) with the hybrid functionals [23, 24] is suitable for such purpose, whereas simplified semiempirical methods like, for example, DFTB fails [25]. DFT-based QM/MM calculations require utilization of supercomputer facilities and many CPUs. Recently, considerable efforts have been performed to develop a GPU-based DFT code [8, 9, 11, 12, 18, 19]. It was implemented in the TeraChem program package [17]. Later, the interface for the QM/MM calculations was proposed [10]. It is based on the powerful NAMD program [16] and the corresponding supporting features can be utilized in QM/MM simulations. This implementation demonstrates dramatic speedup: the same calculations of energy gradient for the system with about 1000 basis functions can be performed on a GPU in less than 2 minutes or on 100 CPUs in about 5 minutes. This allows not only geometry optimization requiring hundreds of gradient calculations for each stationary point on the potential energy surface, but thousands of calculations to obtain a molecular dynamics trajectory.

Computational studies of molecular mechanisms of enzymatic reactions are important for biomedical applications. Analysis of the substrate specificity of proteases can be utilized to construct covalent inhibitors. Those form covalent bonds with the catalytic residues and com-

<sup>1</sup>Department of Chemistry, Lomonosov Moscow State University, Moscow, Russia

<sup>2</sup>Bach Institute of Biochemistry, Federal Research Center "Fundamentals of Biotechnology" of the Russian Academy of Sciences, Moscow, Russia

<sup>3</sup>Mendeleev University of Chemical Technology, Moscow, Russia

<sup>4</sup>South Ural State University (National Research University), Chelyabinsk, Russia

<sup>5</sup>Emanuel Institute of Biochemical Physics, Russian Academy of Sciences, Moscow, Russia

pletely abolish enzymatic activity. This is of great importance for the main protease from the SARS-CoV-2 as it is responsible for partitioning of a synthesized viral polypeptides to the specific fragments that forms a set of proteins required for the virus replication. Elimination of enzymatic activity of the main protease leads to the virus death.

It was recently demonstrated for the main protease from the SARS-CoV-2 that the analysis of QM/MM MD trajectories of the enzyme-substrate complexes can explain the substrate specificity of this enzyme [6]. Also, it was shown that the proper theory level is crucial to obtain valuable results. Utilization of the hybrid DFT functional PBE0 [1] allows one to discriminate reactive and nonreactive enzyme-substrate complexes along MD trajectory, whereas GGA-type functional PBE [14, 15] that lacks exact HF exchange fails.

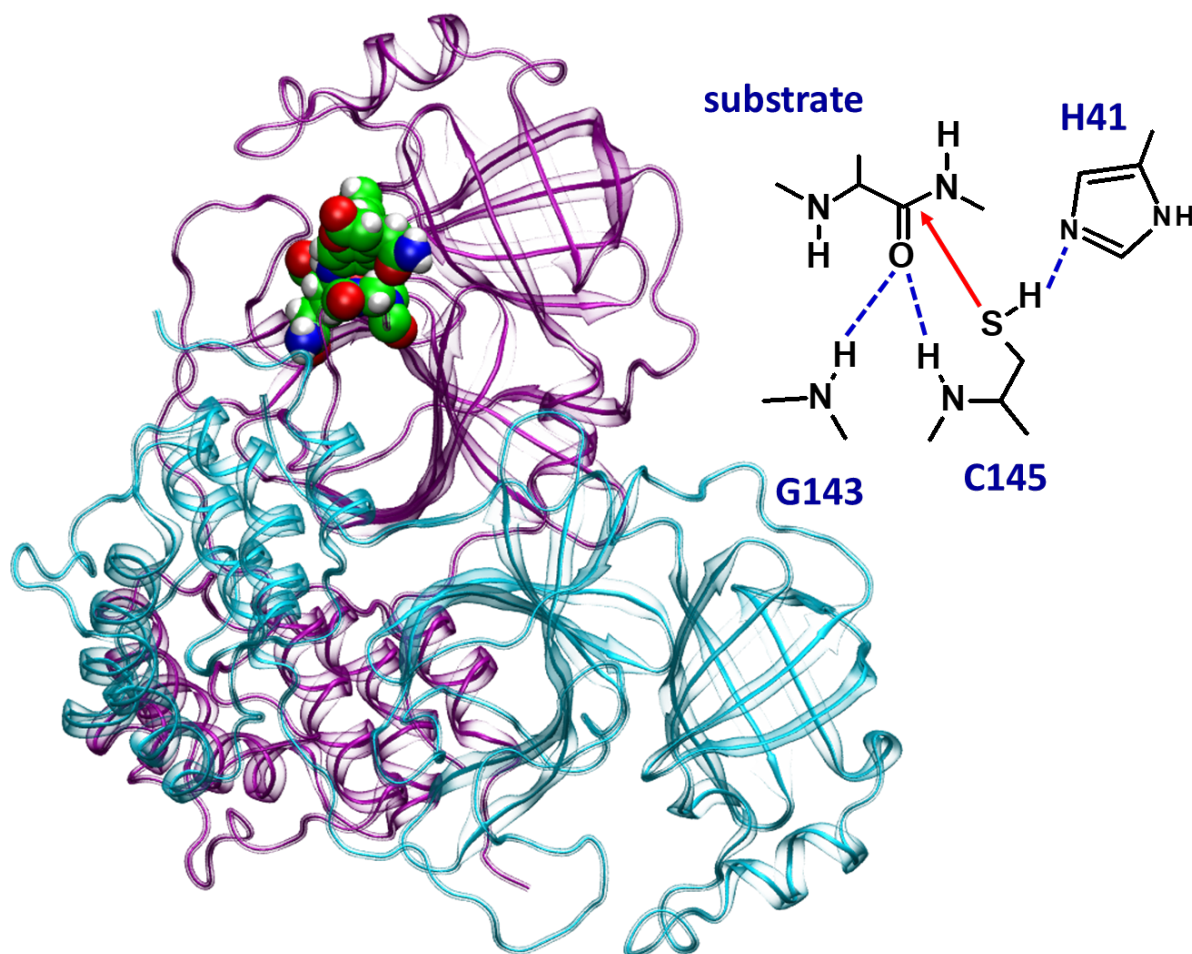
In this communication, we address an important question of the theory level required for the proper description of the dynamic behaviour of the enzyme-substrate complex taking a topical example of the SARS-CoV-2 main protease and its substrate. We test the influence of the basis set, namely presence of polarization functions (d-functions on heavy atoms and p-functions on hydrogen atoms) in QM calculations, as their exclusion speeds up calculations about 2-fold.

## 1. Methodology

We borrowed the model system of the SARS-CoV-2 main protease with its oligopeptide substrate from the recent study [6]. The main protease exists in the dimeric form and the active site of one of monomers is considered (Fig. 1). The substrate is involved in three key interatomic interactions with the active site: two hydrogen bonds with the oxyanion hole of the enzyme formed by the NH groups of the backbones of glycine G143 and cysteine C145; specific interaction formed prior the nucleophilic attack between the carbonyl carbon atom of the substrate and a sulfur atom of C145 (Fig. 1). Another important quantity considered in this study is the interatomic distance between the hydrogen atom of the SH group of C145 and nitrogen atom of histidine H41 (Fig. 1). This distance characterizes whether the proton forms a covalent bond with the nitrogen or sulfur atom along a MD trajectory. Details of the molecular dynamic protocol can be found in ref [6]. Shortly, the QM/MM MD simulations were performed in NPT ensemble at  $T = 300$  K and  $p = 1$  atm. The integration time step was set to 1 ps and the trajectory length was 15 ps for each system. The CHARMM36 [2, 3] and CGenFF [20–22] force fields parameters were utilized for the enzyme and substrate molecules and TIP3P [5] parameters for water molecules. The QM subsystem was treated with the hybrid PBE0 [1] DFT functional with empirical dispersion correction D3 [4] with the 6-31G(d,p) or 6-31G basis sets. The QM/MM MD simulations were performed with TeraChem [17] and NAMD [16] programs and their interface proposed in ref [10]. The electron density analysis, namely the Laplacian of electron density,  $\nabla^2\rho(r)$ , calculations were performed in the Multiwfn program [7].

## 2. Results and Discussion

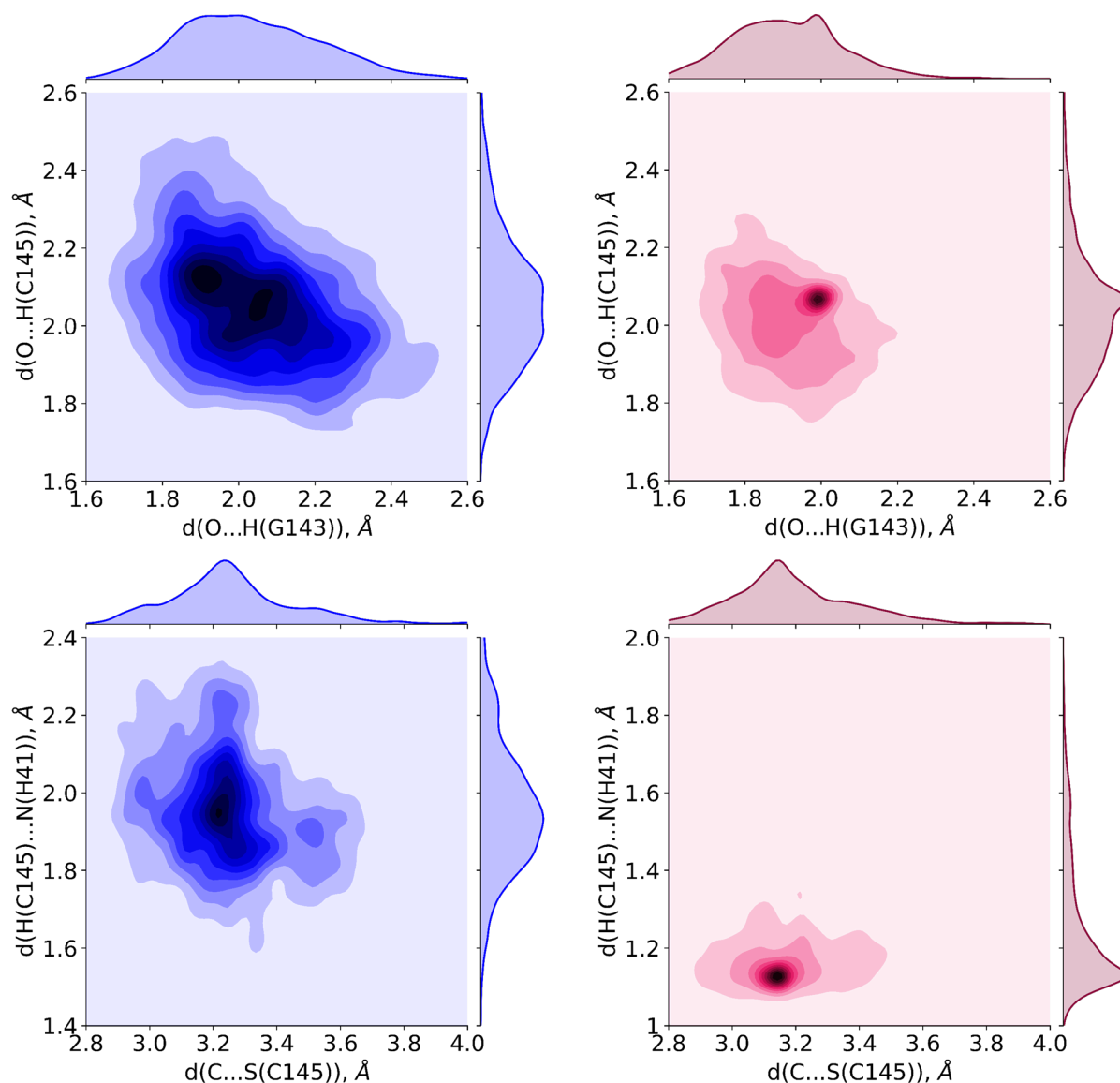
The first step of chemical reaction in the active site of cysteine protease is the nucleophilic attack accompanied with the proton transfer from the SH group of the catalytic cysteine residue (C145) to a histidine residue (H41) [13] (Fig. 1). The oxyanion hole formed by the NH fragments of the backbones of C145 and G143 in case of SARS-CoV-2 is supposed to be responsible for the substrate activation [13] (Fig. 1). Therefore, we start with the analysis of these four interatomic distances distributions to compare MD trajectories simulated at the QM(PBE0-



**Figure 1.** Dimeric main protease from SARS-CoV-2. Monomers are colored cyan and magenta. The active site is shown in colored van der Waals spheres. The 2D structure of the part of substrate and catalytically important parts of the active site are shown. Red arrow depicts the direction of the nucleophilic attack and hydrogen bonds are shown by blue dashed lines

D3/6-31G(d,p))/MM and QM(PBE0-D3/6-31G)/MM levels (Fig. 2, Tab. 1). The hydrogen bonds between the oxygen atom of the substrate and oxyanion hole are distributed similarly in both MD trajectories. The distribution is slightly more narrow and the mean values are slightly shifted to the smaller values in case of QM(PBE0-D3/6-31G)/MM. Similar features are observed for the distribution of the distances of the nucleophilic attack,  $d(\text{C}\dots\text{S}(\text{C145}))$ . Contrary, the  $d(\text{H}(\text{C145})\dots\text{N}(\text{H41}))$  distributions are notably different. In the QM(PBE0-D3/6-31G(d,p))/MM MD trajectory the covalent bond is always formed between a proton and a sulfur atom of C145. In the QM(PBE0-D3/6-31G)/MM MD trajectory, the proton is usually located closely to the nitrogen atom of H41 and the covalent bond is predominantly formed between them.

We analyzed the origin of considerable differences in the dynamic behaviour of model systems treated at different levels of theory. First, we selected the representative MD frame from the QM(PBE0-D3/6-31G(d,p))/MM corresponding to the reactive complex (Fig. 3). The Laplacian of electron density,  $\nabla^2\rho(r)$ , depicts areas of electron density concentration ( $\nabla^2\rho(r) < 0$ ) and depletion ( $\nabla^2\rho(r) > 0$ ), i.e. electrophilic and nucleophilic sites, respectively. If the electron



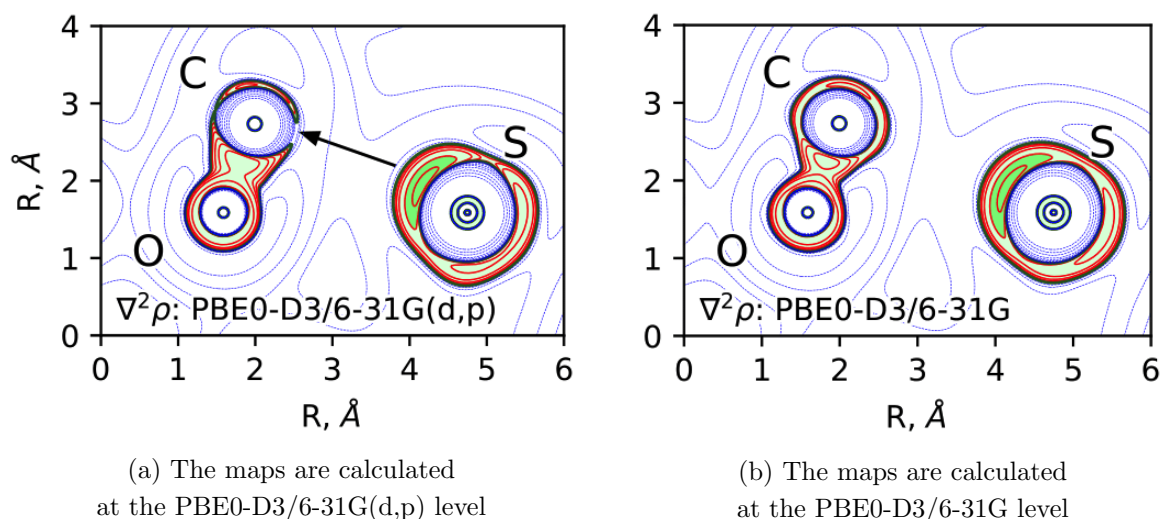
**Figure 2.** Distributions of key interatomic distances obtained along MD trajectories with the QM(PBE0-D3/6-31G(d,p))/MM (blue) and QM(PBE0-D3/6-31G)/MM (red) potentials

**Table 1.** The mean values and standard deviations for the key interatomic distances obtained at different levels of theory

Interatomic distance	QM(PBE0-D3/6-31G(d,p))/MM	QM(PBE0-D3/6-31G)/MM
$d(C...S(C145)), \text{ \AA}$	$3.26 \pm 0.20$	$3.19 \pm 0.19$
$d(O...H(G143)), \text{ \AA}$	$2.04 \pm 0.18$	$1.91 \pm 0.14$
$d(O...H(C145)), \text{ \AA}$	$2.06 \pm 0.16$	$2.03 \pm 0.16$
$d(H(C145)...N(H41)), \text{ \AA}$	$1.96 \pm 0.16$	$1.25 \pm 0.17$

density is calculated at the PBE0-D3/6-31G(d,p) level with polarization functions on all atoms, the electrophilic site on the carbonyl carbon atom of the substrate is observed (Fig. 3a). It is the region of positive values of the  $\nabla^2\rho(r)$  in the direction of nucleophilic attack between the C and S atoms. If the Laplacian of electron density is recalculated at the same geometry

configuration without polarization functions, no substrate activation is observed (Fig. 3b). The nucleophilic attack is facilitated by the electron lone pair of the sulfur atom that is observed on both  $\nabla^2\rho(r)$  maps. It is due to the fact that the substrate activation is a more refined electron density feature than the lone pair electron concentration, and polarization functions are required for its description. To further support this idea, we calculated electron density difference,  $\Delta\rho$ , maps in the same plane (Fig. 4). The difference is mostly pronounced around the carbonyl carbon atom. Electron density is considerably higher in the direction of the nucleophilic attack if calculated without polarization functions. This explains the absence of the electron density depletion region around C atom on the  $\nabla^2\rho(r)$  maps calculated at the PBE0-D3/6-31G level.



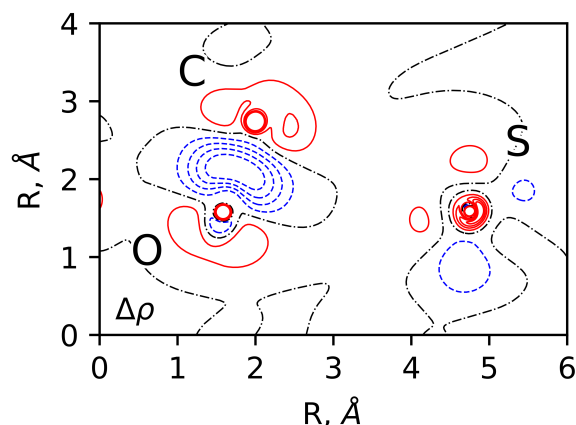
**Figure 3.** The Laplacian of electron density,  $\nabla^2\rho(r)$ , maps in the plane S(Cys145)...CO (the carbonyl group of substrate) of the representative reactive enzyme-substrate complex obtained at the QM(PBE0-D3/6-31G(d,p))/MM level. Contour lines are  $\pm(2; 4; 8) \times 10^n$  au,  $-2 \leq n \leq 1$ , blue dashed contour lines indicate the electron density depletion areas ( $\nabla^2\rho(r) > 0$ ) and red solid lines identify the electron density concentration ( $\nabla^2\rho(r) < 0$ ), green solid line is  $\nabla^2\rho(r) = 0$ . The area with  $\nabla^2\rho(r) < 0$  is colored in light green and the lone pair on the sulfur atom ( $\nabla^2\rho(r) < -0.2$ ) is highlighted green. Black arrow indicates the direction of the nucleophilic attack

We also analysed  $\nabla^2\rho(r)$  maps at several MD frames from the trajectory calculated at the QM(PBE0-D3/6-31G)/MM level. We selected a frame with the short distances of the nucleophilic attack and hydrogen bonds with the oxyanion hole and a frame where the proton from the SH group is transferred to the histidine residue (in this case the nucleophile is a negatively charged sulfur). In both cases Laplacian of electron density maps demonstrates no substrate activation.

Combining these computational experiments together, we can conclude that the applied computational strategy is useful in predicting substrate specificity of enzymes, if a proper level of theory is applied.

## Conclusion

Novel DFT code implemented in the TeraChem [17] and optimized for the GPU discloses new opportunities for the QM(DFT)/MM molecular dynamics simulations. It allows fast and accurate simulations of interactions occurring between the substrate and enzyme in its active site.



**Figure 4.** The electron density difference,  $\Delta\rho$ , maps between  $\rho$  calculated at the PBE0-D3/6-31G and PBE0-D3/6-31G(d,p) levels calculated in the same plane and at the same MD frame as on Fig. 3. Blue dashed isolines are  $-0.028$  au,  $-0.021$  au,  $-0.014$  au and  $-0.007$  au, red solid isolines are  $0.007$  au,  $0.014$  au,  $0.021$  au and  $0.028$  au, black dash-dotted isoline corresponds to  $\Delta\rho = 0$

We take the main protease from the SARS-CoV-2 and its substrate to analyze the influence of the basis set on the quality of interaction description. Despite the 2-fold speedup of the calculations, elimination of polarization functions results in the wrong description of the substrate activation and cannot be recommended for such analysis. Utilization of the 6-31G(d,p) basis set enables to disclose equilibrium between the reactive and nonreactive species along the MD trajectory. We proposed the methodology based on the QM/MM MD simulations followed by the electron density analysis that allow one to determine substrate activation in the active site of a protease and we demonstrated the importance of utilization of a proper QM theory level for reliable evaluation of this quantity. This approach can be utilized to choose covalent inhibitors that will readily interact with the catalytic residue of the selected enzyme.

## Acknowledgements

This work was supported by the Russian Foundation for Basic Research (project № 18-29-13006). We acknowledge the use of supercomputer resources of the Joint Supercomputer Center of the Russian Academy of Sciences and the equipment of the shared research facilities of HPC computing resources at Lomonosov Moscow State University [26].

*This paper is distributed under the terms of the Creative Commons Attribution-Non Commercial 3.0 License which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is properly cited.*

## References

1. Adamo, C., Barone, V.: Toward reliable density functional methods without adjustable parameters: The PBE0 model. *The Journal of Chemical Physics* 110(13), 6158 (1999), DOI: 10.1063/1.478522
2. Anisimov, V.M., Lamoureux, G., Vorobyov, I.V., et al.: Determination of electrostatic

- parameters for a polarizable force field based on the classical drude oscillator. *Journal of Chemical Theory and Computation* 1(1), 153–168 (2005), DOI: 10.1021/ct049930p
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  $\phi$ ,  $\psi$  and side-chain  $\chi_1$  and  $\chi_2$  dihedral angles. *Journal of Chemical Theory and Computation* 8(9), 3257–3273 (2012), DOI: 10.1021/ct300400x
  4. Grimme, S., Antony, J., Ehrlich, S., Krieg, H.: A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *The Journal of Chemical Physics* 132(15), 154104 (2010), DOI: 10.1063/1.3382344
  5. 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), DOI: 10.1063/1.445869
  6. Khrenova, M.G., Tsirelson, V.G., Nemukhin, A.V.: Dynamical properties of enzymesubstrate complexes disclose substrate specificity of the SARS-CoV-2 main protease as characterized by the electron density descriptors. *Physical Chemistry Chemical Physics* 22(34), 19069–19079 (2020), DOI: 10.1039/D0CP03560B
  7. Lu, T., Chen, F.: Multiwfn: A multifunctional wavefunction analyzer. *Journal of Computational Chemistry* 33(5), 580–592 (2012), DOI: 10.1002/jcc.22885
  8. Luehr, N., Ufimtsev, I.S., Martínez, T.J.: Dynamic precision for electron repulsion integral evaluation on graphical processing units (GPUs). *Journal of Chemical Theory and Computation* 7(4), 949–954 (2011), DOI: 10.1021/ct100701w
  9. Manathunga, M., Miao, Y., Mu, D., et al.: Parallel implementation of density functional theory methods in the quantum interaction computational kernel program. *Journal of Chemical Theory and Computation* 16(6), 4315–4326 (2020), DOI: 10.1021/acs.jctc.0c00290
  10. Melo, M.C.R., Bernardi, R.C., Rudack, T., et al.: NAMD goes quantum: An integrative suite for QM/MM simulations. *Nature methods* 15(5), 351 (2018), DOI: 10.1038/nmeth.4638
  11. Miao, Y., Merz, K.M.: Acceleration of electron repulsion integral evaluation on graphics processing units via use of recurrence relations. *Journal of Chemical Theory and Computation* 9(2), 965–976 (2013), DOI: 10.1021/ct300754n
  12. Miao, Y., Merz, K.M.: Acceleration of high angular momentum electron repulsion integrals and integral derivatives on graphics processing units. *Journal of Chemical Theory and Computation* 11(4), 1449–1462 (2015), DOI: 10.1021/ct500984t
  13. Otto, H.H.: Cysteine proteases and their inhibitors. *Chemical Reviews* 97(1), 133–172 (1997), DOI: 10.1021/cr950025u
  14. Perdew, J.P., Burke, K., Ernzerhof, M.: Generalized gradient approximation made simple. *Physical Review Letters* 77(18), 3865–3868 (1996), DOI: 10.1103/PhysRevLett.77.3865
  15. Perdew, J.P., Burke, K., Ernzerhof, M.: Generalized gradient approximation made simple [Phys. Rev. Lett. 77, 3865 (1996)]. *Physical Review Letters* 78(7), 1396–1396 (1997), DOI: 10.1103/PhysRevLett.78.1396

16. Phillips, J.C., Braun, R., Wang, W., et al.: Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry* 26(16), 1781–1802 (2005), DOI: 10.1002/jcc.20289
17. TeraChem v 1.9, PetaChem, LLC, [www.petachem.com](http://www.petachem.com)
18. Ufimtsev, I.S., Martínez, T.J.: Quantum chemistry on graphical processing units. 1. Strategies for two-electron integral evaluation. *Journal of Chemical Theory and Computation* 4(2), 222–231 (2008), DOI: 10.1021/ct700268q
19. Ufimtsev, I.S., Martinez, T.J.: Quantum chemistry on graphical processing units. 2. Direct self-consistent-field implementation. *Journal of Chemical Theory and Computation* 5(4), 1004–1015 (2009), DOI: 10.1021/ct800526s
20. Vanommeslaeghe, K., Hatcher, E., Acharya, C., et al.: CHARMM general force field (CGenFF): A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *Journal of Computational Chemistry* 31(4), 671–690 (2010), DOI: 10.1002/jcc.21367
21. Vanommeslaeghe, K., MacKerell, A.D.: 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), DOI: 10.1021/ci300363c
22. Vanommeslaeghe, K., Raman, E.P., MacKerell, A.D.: Automation of the CHARMM general force field (CGenFF) II: Assignment of bonded parameters and partial atomic charges. *Journal of Chemical Information and Modeling* 52(12), 3155–3168 (2012), DOI: 10.1021/ci3003649
23. Vasilevskaya, T., Khrenova, M.G., Nemukhin, A.V., Thiel, W.: Mechanism of proteolysis in matrix metalloproteinase-2 revealed by QM/MM modeling. *Journal of Computational Chemistry* 36(21), 1621–1630 (2015), DOI: 10.1002/jcc.23977
24. Vasilevskaya, T., Khrenova, M.G., Nemukhin, A.V., Thiel, W.: Methodological aspects of QM/MM calculations: A case study on matrix metalloproteinase-2. *Journal of Computational Chemistry* 37(19), 1801–1809 (2016), DOI: 10.1002/jcc.24395
25. Vasilevskaya, T., Khrenova, M.G., Nemukhin, A.V., Thiel, W.: Reaction mechanism of matrix metalloproteinases with a catalytically active zinc ion studied by the QM(DFTB)/MM simulations. *Mendeleev Communications* 26(3), 209–211 (2016), DOI: 10.1016/j.mencom.2016.05.010
26. Voevodin, 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), DOI: 10.14529/jsfi190201



# In Search of Non-covalent Inhibitors of SARS–CoV–2 Main Protease: Computer Aided Drug Design Using Docking and Quantum Chemistry

Alexey V. Sulimov<sup>1,2</sup>, Danil C. Kutov<sup>1,2</sup>, Anna S. Taschilova<sup>1,2</sup>,  
Ivan S. Ilin<sup>1,2</sup>, Nadezhda V. Stolpovskaya<sup>3</sup>, Khidmet S. Shikhaliev<sup>3</sup>,  
Vladimir B. Sulimov<sup>1,2,4</sup>

© The Authors 2020. This paper is published with open access at SuperFri.org

Two stages virtual screening of a database containing several thousand low molecular weight organic compounds is performed with the goal to find inhibitors of SARS–CoV–2 main protease. Overall near 41000 different 3D molecular structures have been generated from the initial molecules taking into account several conformers of most molecules. At the first stage the classical SOL docking program is used to determine most promising candidates to become inhibitors. SOL employs the MMFF94 force field, the genetic algorithm (GA) of the global energy optimization, takes into account the desolvation effect arising upon protein-ligand binding and the internal stress energy of the ligand. Parameters of GA are selected to perform the meticulous global optimization, and for docking of one ligand several hours on one computing core are needed on the average. The main protease model is constructed on the base of the protein structure from the Protein Data Bank complex 6W63. More than 1000 ligands structures have been selected for further postprocessing. The SOL score values of these ligands are more negative than the threshold of  $-6.3$  kcal/mol obtained for the native X77 ligand docking. Subsequent calculation of the protein-ligand binding enthalpy by the PM7 quantum-chemical semiempirical method with COSMO solvent model have narrowed down the number of best candidates. Finally, the diverse set of 20 most perspective candidates for the *in vitro* validation are selected.

*Keywords:* docking, global optimization, quantum docking, inhibitors, CADD, SARS–CoV–2, COVID–19,  $M^{pro}$ .

## Introduction

The recent outbreak of the COVID–19 pandemic caused by the SARS–CoV–2 coronavirus (CoV) makes the development of appropriate drugs and vaccines extremely urgent. Currently, there are no direct-acting drugs for the SARS–CoV–2 virus that are specifically designed to inhibit the proteins of this particular coronavirus. At the same time, there are already three-dimensional structures of the main protease ( $M^{pro}$  or  $3CL^{pro}$  – 3-chymotrypsin-like protease) of SARS–CoV–2 with high resolution. The molecular mechanisms of the SARS–CoV–2 replication life cycle are largely understood, and  $M^{pro}$  is considered an important therapeutic target for new drugs. Different test systems for *in vitro* experiments to determine the activity of compounds to inhibit  $M^{pro}$  of this coronavirus have been developed. Thus, there are all the necessary prerequisites for the development of new anti-SARS–CoV–2 drugs using computer modeling based on the known structure of the main protease of SARS–CoV–2.

Developing direct anti-CoV drugs was conducted long before the COVID–19 outbreak. Different CoVs cause a variety of respiratory diseases from the common cold [36] to the Severe Acute Respiratory Syndrome (SARS)–CoV [10, 29]. The coronavirus similar to SARS–CoV, the Mid-

<sup>1</sup>Research Computer Center of Lomonosov Moscow State University, Moscow, Russia

<sup>2</sup>Moscow Center of Fundamental and Applied Mathematics, Moscow, Russia

<sup>3</sup>Department of Organic Chemistry, Faculty of Chemistry, Voronezh State University, Voronezh, Russia

<sup>4</sup>Dimonta Ltd., Moscow, Russia

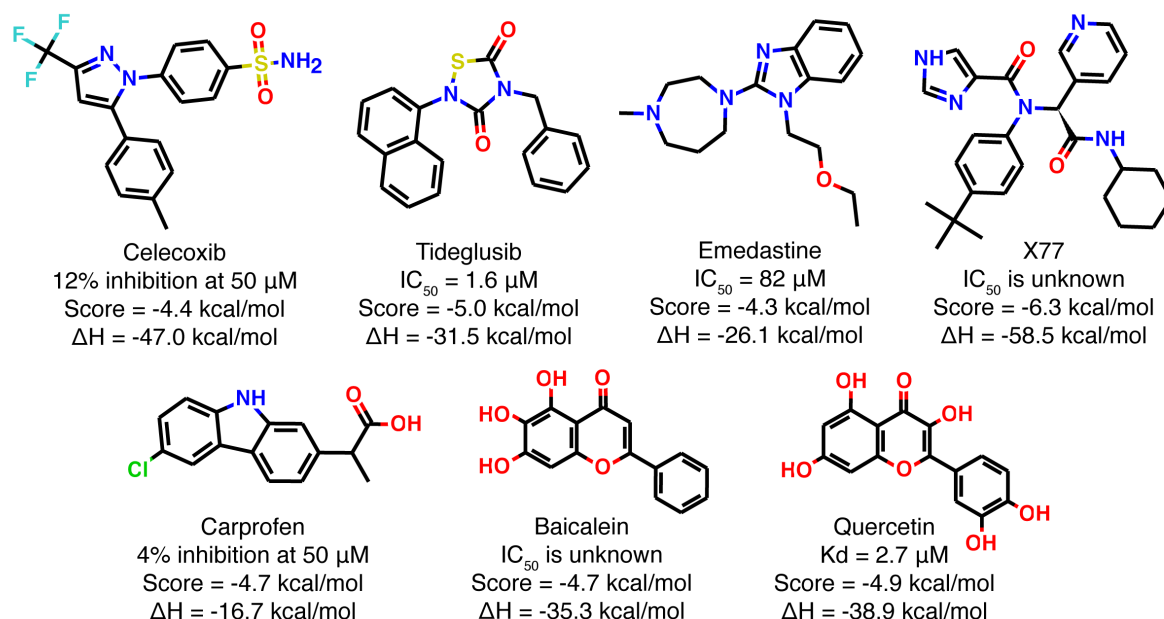
dle East respiratory syndrome coronavirus (MERS-CoV), was identified in 2012 [67]. Soon after discovery SARS-CoV the spatial atomistic models of its main protease have been determined by homology [3] and from experimental crystal structure [65] and the key role of M<sup>pro</sup> [2] in the SARS-CoV and MERS-CoV replication has been revealed. During 2003–2015 many reversible and irreversible inhibitors of the main protease of SARS-CoV have been found [41], but most of them were weak inhibitors, and none of them reached clinical trials. However, these attempts prepared the basis for anti-SARS-CoV-2 computer aided structural based drug design. The main protease and several other proteins of SARS-CoV-2 have been rapidly identified as the therapeutic target for anti-COVID-19 drugs. Many high quality 3D structures of SARS-CoV-2 M<sup>pro</sup> in the apo form and with different inhibitors have been already deposited in Protein Data Bank [5]. SARS-CoV-2 M<sup>pro</sup> has a cysteine-histidine catalytic dyad: *Cys145* and *His41*.

Docking is the most widely used molecular modeling method in the search of SARS-CoV and SARS-CoV-2 main protease inhibitors. Results of a broad search for SARS-CoV-2 non-covalent M<sup>pro</sup> inhibitors are presented in [59] where the Deep Docking (DD) platform [16] is used. The effective combination of quantitative structure-activity relationship (QSAR) methods with docking by the Glide program [13, 57] allows to explore the ZINC15 library containing 1.36 billion compounds. The M<sup>pro</sup> model is prepared on the base of the PDB 6LU7 structure containing the M<sup>pro</sup> co-crystallized with a covalent inhibitor. DD uses a deep neural network (DNN) which is trained with docking scores calculated for a set of randomly selected ligands. The trained DNN predicts scores for other ligands and these estimates allow to avoid wasting time for docking of ligands with bad scores. The procedure is organized in four iterative steps with improving the training set at each iteration and docking of three million of best ligands at the final iteration. As a result 1000 ligand have been revealed with the best scores for experimental validation of their inhibitory activity but the experimental confirmation has not been published yet.

In the frame of the drug repurposing strategy the search of M<sup>pro</sup> inhibitors is performed among existing and approved drugs. Such approach is possessed of two advantages. First, such compounds have acceptable solubility in water and this facilitates experimental measurements of their inhibitory activity *in vitro*. Second, it is easier and faster to perform all necessary toxicity studies, preclinical and clinical trials for approving the new drug on the base of such inhibitors of a new target. In the study [32], clinically approved drugs from the DrugBank database were virtually screened against M<sup>pro</sup> by using the Libdock docking program (Discovery Studio 3.5, Accelrys Software Inc). The M<sup>pro</sup> model is constructed by homology using the SARS-CoV M<sup>pro</sup> structure (PDB ID: 1UJ1) and ten best candidates for the experimental verification are identified. Among them are *colistin* (a polymyxin antibiotic), *valrubicin* (an anthracycline antibiotic), *icatibant* (an antagonist of bradykinin receptors) and *bepotastine* (belongs to the family of antihistamines). Those are still waiting an experimental validation. The MOE docking software is used in for the search of SARS-CoV-2 M<sup>pro</sup> inhibitors among 16 antiviral drugs targeting viral proteases and in in-house database, overall there are about 8000 molecules [26]. As a result two *flavone* and *coumarin* derivatives from the in-house database and three approved protease inhibitors (*Remdesivir*, *Daraunvir*, and *Saquinavir*) have been revealed for subsequent experimental validation. Motonory Tsuji performs virtual screening [61] of 1485144 known bioactive compounds by two docking programs, rDock [44] and AutoDock Vina [60]. The screened database is extracted from the ChEMBL26 [9] database and it includes 13308 approved drugs. The M<sup>pro</sup> model is based on the structure with PDB ID 6Y2G and optimization of the protein is

performed with the AMBER99 force field and subsequent molecular dynamics low-temperature annealing. First, rDock reveals 64 ligands candidates to become M<sup>PRO</sup> inhibitors. Second, these best ligands are docked with AutoDock Vina and 29 hits are determined. Among these hits is only one approved drug (*eszopiclone*). The author considers that AutoDock Vina is more accurate than rDock but we notice the lack of Coulomb interactions in AutoDock Vina. No results on experimental confirmation of these findings are published. Authors of [17] go further in the screening strategy and use three docking programs: Glide, FRED [33] and AutoDock Vina. Only the equivalent high affinity binding modes predicted simultaneously by the three docking programs were considered to correspond to bioactive poses. A complex of M<sup>PRO</sup> with the ligand Glide pose is minimized by applying the MM-GBSA minimization in the Prime module (Schrödinger, LLC). Two libraries of approved drugs were screened: eDrug3D [40] and Reaxys-marketed [12]. On the base of docking results and the visual inspection of best ligand poses in the active site of M<sup>PRO</sup> seven candidates to become inhibitors of M<sup>PRO</sup> are selected: *perampanel*, *carprofen*, *celecoxib*, *alprazolam*, *trovafloxacin*, *sarafloxacin* and *ethyl biscoumacetate*. Two of them, *carprofen* and *celecoxib*, demonstrate 4% and 12% inhibition of M<sup>PRO</sup> at 50  $\mu$ M respectively – certainly, it is too weak activity. AutoDock Vina is used also in [6] where four SARS-CoV-2 proteins, including M<sup>PRO</sup>, are targeted by 16 antiviral compounds. The atomistic models of targets are constructed by homology modelling with the I-Tasser server [66]. Docking into M<sup>PRO</sup> is performed with nine moveable residues in the active site. *Simeprevir*, a HCVNS3/4A protease inhibitor, is found the best for M<sup>PRO</sup> inhibition. AutoDock Vina is used for docking of 62 alkaloids and 100 terpenoids from African plants into the M<sup>PRO</sup> model constructed using the complex with PDB ID 6LU7 [19]. The two best alkaloids are *10-hydroxyusambarensine* and *cryptoquinoline*, and two best terpenoids are *6-oxoisoiguesterin* and *22-hydroxyhopan-3-one*. Authors of [15] in search of M<sup>PRO</sup> inhibitors screen marine natural products (14064 compounds) by a combination of the pharmacophore filter, molecular docking and molecular dynamics (MD) simulations. The model of M<sup>PRO</sup> is constructed by the Pharmit server using the complex with PDB ID 6LU7. After the pharmacophore filter 180 compounds are docked by AutoDock Vina. 17 best docked compounds are subjected to MD simulation and the most promising candidates to become M<sup>PRO</sup> inhibitors are phlorotannins which are oligomers of *phloroglucinol* (1,3,5-trihydroxybenzene). Among these phlorotannins is *Dieckol* which has been already proven to be the M<sup>PRO</sup> inhibitor with IC<sub>50</sub> = 2.7  $\mu$ M. SARS-CoV-2 M<sup>PRO</sup> inhibitors are found in [11] but without an experimental validation. Thirty four approved and on-trial inhibitors of different proteases are docked by AutoDock 4.2 into the SARS-CoV-2 M<sup>PRO</sup> models prepared from the PDB complex 6LU7 and from the PDB complex 6M2N. Several drugs are identified as candidates to become SARS-CoV-2 M<sup>PRO</sup> inhibitors, within the classes of the HCV protease, DPP-4,  $\alpha$ -thrombin and coagulation Factor Xa inhibitors. Several conclusions can be made from this short review. First, the bottleneck of new inhibitors discovery is the experimental *in vitro* validation of predicted inhibitors of SARS-CoV-2. Actually such validation will possible be made soon and new SARS-CoV-2 M<sup>PRO</sup> inhibitors will be published. Second, after selection of best compounds using results of fast docking more accurate and slow docking programs are used for the identification of best ligand-candidates for experimental validation. Third, for the acceleration of screening of large databases some a priori considerations are widely used. They are different pharmacophore filtering, narrowing sets of molecules for screening by other methods including “scientific intuition” as well as discarding non-active compounds using predictions on the base of neural networks as in Deep Docking platform (see above).

All existing inhibitors targeting SARS-CoV-2 M<sup>Pro</sup> can be classified into two unequal groups: larger set of covalent inhibitors and smaller group of non-covalent inhibitors. The chemical description of compounds in the first group is mainly based on the nature of the reactive group (a warhead) forming a covalent bond with *Cys145*. Development of covalent inhibitors has its own specific features but we focus here only on non-covalent inhibitors. At the time of this writing only several weak non-covalent inhibitors of SARS-CoV-2 M<sup>Pro</sup> are published (Fig. 1): *tideglusib* [23], *emedastine* [14], *X77* [35], *carprofen* and *celecoxib* [17], *quercetin* [1], and *baicalein* [42].



**Figure 1.** Non-covalent inhibitors of SARS-CoV-2 M<sup>Pro</sup>

*Tideglusib* is naphthalene-containing compound with thiadiazolidine-3,5-dione as a scaffold. *Emedastine* is an antihistamine drug consisting of a diazepane ring and benzimidazole. Compound *X77* – is an only non-covalent drug-like inhibitor crystallized with M<sup>Pro</sup>. It possesses X-like shape with imidazole, *i*Pr-benzene, pyridine and cyclohexene as structural moieties. *Carprofen* is a non-steroidal anti-inflammatory drug now used in veterinary which is simply di-substituted carbazole. *Celecoxib* is also a non-steroidal anti-inflammatory agent based on pyrazole scaffold substituted with methylbenzene and benzenesulfonamide. *Quercetin* and *baicalein* both belong to a group of flavonoids and contain a flavone fragment substituted with a few hydroxyl groups.

We present here the results of virtual screening of a database containing more than 40 000 structures of low molecular weight ligands using our own SOL [53, 56] docking program. For the ligands with best docking scores the binding enthalpy is calculated using the PM7 quantum-chemical semiempirical method and the implicit COSMO solvent model. The choice of thresholds separating active compounds from inactive ones is made with respect to the docking score and the binding enthalpy obtained by the same methods for the native ligand crystallized with M<sup>Pro</sup> in the respective Protein Data Bank. 20 compounds have been selected for the further experimental *in vitro* validation on the base of best docking scores as well as best binding enthalpy values.

## 1. Materials and Methods

Virtual screening requires construction of the atomistic 3D model of the target protein, preparation of 3D structures of all ligands from the database to be screened. The protein model as well as ligand models define accuracy of docking calculations and they should be prepared carefully. The target protein models are constructed on the base of protein-ligand complexes which 3D structures are stored in Protein Data Bank (PDB). These structures contain Cartesian coordinates of all heavy, i.e. non-hydrogen, atoms. Several thousand hydrogen atoms should be added to the protein taking into account the pH condition of solvent where the protein is working. Protonation states of amino acid residues are well known for any pH (usually it is neutral condition  $\text{pH} = 7.4$ ) if the influence of neighboring atoms in the protein globule is not taken into account. Even for equal protonation states different programs add hydrogen atoms in different ways resulting in unequal hydrogen atoms spatial positions. The latter leads to varied results of docking performance, different best ligand poses and diverse scores predicting the protein-ligand binding affinity [30]. Ligands in most of databases are stored as 2D structures and some programs should be used to construct low energy conformers, tautomers and protonation states for each given ligand. Below we explain how all these preliminary works are done in the present case of virtual screening as well as we describe shortly the docking program and quantum-chemical method used.

### 1.1. Protein Spatial Model

Many structures of SARS-CoV-2 main protease have been already deposited in the Protein Data Bank. The structures with best quality have following PDB ID: 5R7Z, 5R83 and 6W63. These are structures of M<sup>Pro</sup> crystallized with non-covalent inhibitors, they do not contain missing residues or atoms and have a good resolution  $< 2.2 \text{ \AA}$ . Protein models are prepared from these PDB structures by removing all atoms, ions and molecules which do not belong to the protein, and then hydrogen atoms are added by our APLITE program [30]. Hydrogen atoms are added to the native ligands extracted from the PDB structures by Avogadro [21]. After local optimization of the energy of these complexes in the frame of the MMFF94 force field [20] with the variation of Cartesian coordinates of all native ligand atoms the position of the ligand does not change significantly: RMSD values calculated over all ligand atoms between native crystallized ligand pose and the optimized ligand pose are less than  $1.6 \text{ \AA}$  in all these three models. This is the simplest check of applicability of the MMFF94 force field used in SOL to modeling of protein-ligand interactions in these complexes. Next, we check the ability of SOL to reproduce the native ligand pose crystallized with the protein. For the M<sup>pro</sup> model prepared from the 5R7Z complex the native docking fails: RMSD between best docked ligand pose and the crystallized one is more than  $7 \text{ \AA}$ . For models constructed on the base of 5R83 and 6W63 complexes the native docking is successful:  $\text{RMSD} = 1.19 \text{ \AA}$  and  $\text{RMSD} = 1.31 \text{ \AA}$ . The visual inspection of the active sites of these protein models reveals that one residue (MET49) of the active site is mobile and adapts to the bound inhibitor. Finally, the model based on the 6W63 PDB structure (the X77 native ligand has 7 torsions) has been selected as the target protein model for virtual screening because the protein active site is more open than one of the model used 5R83 PDB complex (the native ligand has 4 torsions). The SOL score of the native ligand docked into the 6W63 model is equal to  $-6.3 \text{ kcal/mol}$ . This value gives a threshold

for the selection of best ligands after the docking step of virtual screening. The Lomonosov supercomputer [62] of Lomonosov Moscow State University is employed.

## 1.2. Database and Ligand 3D Structure Preparation

The database of compounds of the Department of Organic Chemistry of Voronezh State University [58] is used for the present virtual screening. A wide range of nitrogen-, oxygen- and sulfur-containing heterocyclic compounds are presented in the database. The compounds are small drug-like molecules. Among them there are hydroquinoline derivatives with antibacterial, antifungal, anticoagulant activity [22, 25, 34, 37, 38], aminopyrimidines and pyrrolo[3,2,1-ij]quinolin-2-ones, which are factor Xa and protein kinases inhibitors [50, 54], various plant growth stimulants of the getarylcarboxylic acid class [63, 64].

The database contains approximately 19 000 molecules, and after 2D→3D transformation we obtain 41 000 molecular 3D-structures. The main source of using different 3D-structures for one molecule is different low energy conformations of non-aromatic rings including macro-cycles. The SOL docking program treats the ligand flexibility, as many other docking programs do, only by variations of all torsions, i.e. degrees of freedom describing internal rotations of ligand molecular groups around ordinary covalent bonds, keeping fixed covalent bond lengths and valence angles. Different conformations of non-aromatic rings and macro-cycles of one and the same molecule should be docked as different molecules. The ligands from the database are protonated with ChemAxon Protonation module [7] at pH = 7.4. OpenBabel was used for generating 3D coordinates.

## 1.3. SOL Docking Program

SOL [53, 56] is a classic docking program with many features used in popular docking programs [57]. However, developing this program we tried to make as few model simplifications as possible and to take into account the most important effects determining the accuracy of docking. Performance of SOL is based on the docking paradigm connecting docking with the global optimization problem: the best ligand pose should be near the global energy minimum of the protein-ligand complex. The energy is calculated in the frame of the MMFF94 force field [20] almost without any simplifications and the genetic algorithm (GA) is used for the global optimization. The preliminary calculated grid of potentials describing interactions of a probe ligand atom with the rigid protein is used. Coulomb and van der Waals interactions as well as desolvation energy potentials are calculated in nodes of the grid covering the docking cube (its edge is equal to 22 Å) with the distance between neighboring nodes of 0.22 Å. The desolvation potentials are calculated in the frame of a simplified Generalized Born implicit solvent model. In the process of the global optimization the ligand internal strain energy is calculated also using MMFF94. In the frame of the genetic algorithm niching is used. Niching provides diversity of probe ligand poses preventing premature concentration of best ligand poses near a local energy minimum. The default values of main parameters of GA are sufficiently large: the population size 30 000 and the number of generation 1000. 50 independent runs (by default) of GA are performed, the clustering of 50 solutions is made and the relatively large occupation number of the cluster containing ligand poses with lowest energies is the indication of the successful finding of the ligand pose with the lowest energy. SOL was used for CSAR 2011–2012 docking competition with other docking programs Gold, AutoDock, AutoDock Vina, ICMVLS, Glide

and others and for two of three target-proteins SOL results are among the best ones [8, 53]. SOL is successfully used for development of new inhibitors of several targets with the experimental *in vitro* confirmation: thrombin [45], urokinase (uPA) [4, 55] and the blood coagulation factors Xa and XIa [22, 38, 54]. SOL is adapted to screen large databases of ligands on Lomonosov supercomputer [62] of M.V. Lomonosov Moscow State. Although SOL is a parallel program it is more effective to perform virtual screening of large ligand databases distributing jobs of ligand docking over hundreds and thousands of computing cores, one ligand per one core. Some auxiliary scripts and programs are utilized to submit and queue up docking jobs and to analyze results. Docking of one ligand per one core needs from 1 to several hours depending of the size and flexibility of the ligand. Usually, the docking SOL score values for inhibitors of different target proteins are in the range from  $-5$  kcal/mol to  $-7$  kcal/mol.

#### 1.4. Protein-ligand Binding Enthalpy

Quantum-chemical semiempirical methods are developed since 70–90th of 20th century but their main feedback was until recently the description of non-polar intermolecular interactions including H-bonds formations. But recently the new PM6 [43, 46] and PM7 [47] methods have been developed in the frame of the NDDO (Neglect of Diatomic Differential Overlap) approximation utilizing ideas which have been used in DFT methods [18, 24] for the proper description of dispersion interactions and the hydrogen and halogen bonds formation. Moreover, the PM7 method includes these novelties at the parameterization stage which is based on an extremely large set of molecular data. PM7 as well as PM6 with respective corrections are realized in the MOPAC package [48] and the validation on a broad set of test molecular systems demonstrates that these methods work no worse than DFT ones. In addition, MOPAC includes the MOZYME module [49], where the localized molecular orbitals method is used instead of the LCAO (Linear Combination of Atomic Orbitals) approximation. This allows calculations of whole protein-ligand complexes and to do this quickly. The solute-solvent interaction is responsible for the desolvation energy which is a large term in the protein-ligand binding energy. Due to a large electrostatic permittivity of water (78.5) desolvation term screens strongly Coulomb interactions between protein and ligand atoms. The use of solvent model is extremely important for docking and the binding-energy estimations [51, 52]. So, the binding enthalpy calculations we use here PM7 with the COSMO [27, 28] implicit solvent model implemented also in MOPAC.

After virtual screening by docking and determination of ligands with best docking scores for these best ligands the binding enthalpy  $\Delta H_{bind}$  is calculated as follows:

$$\Delta H_{bind} = H(PL) - H(P) - H(L), \quad (1)$$

where  $H(X)$  is the enthalpy of formation of the molecular system X (where X=PL, P or L) calculated by MOPAC: PL, P and L are the protein-ligand complex, the unbound protein and the unbound ligand, respectively. The protein-ligand binding free energy consists of enthalpy and entropy terms. The latter is positive in most cases due to the loss of degrees of freedom when the ligand is bound to the protein. So, we consider that the more negative is the binding enthalpy the more negative is the binding free energy and respective protein-ligand affinity.

The  $H(PL)$  is calculated as follows. The best docked ligand pose in the protein-ligand complex is used as the initial ligand position at the local optimization of the energy of the protein-ligand complex. The optimization is made by the gradient L-BFGS (limited-memory Broyden-Fletcher-Goldfarb-Shanno) method implemented in MOPAC, positions of all ligand

atoms are varied while all protein atoms are kept fixed. During this optimization the energy of the complex is calculated by PM7 without solvent. The final enthalpy of formation of the complex in the local minimum is recalculated by PM7 with the COSMO solvent (PM7+COSMO) without optimization using 1SCF keyword of MOPAC. The initial conformation of the unbound ligand is generated by the Open Babel program [39] and the local energy optimization is performed by PM7 with variations of positions of all ligand atoms, and the enthalpy of formation  $\Delta H(L)$  is recalculated by PM7+COSMO for the optimized configuration. The enthalpy of formation of the unbound protein is calculated by PM7+COSMO without energy optimization.

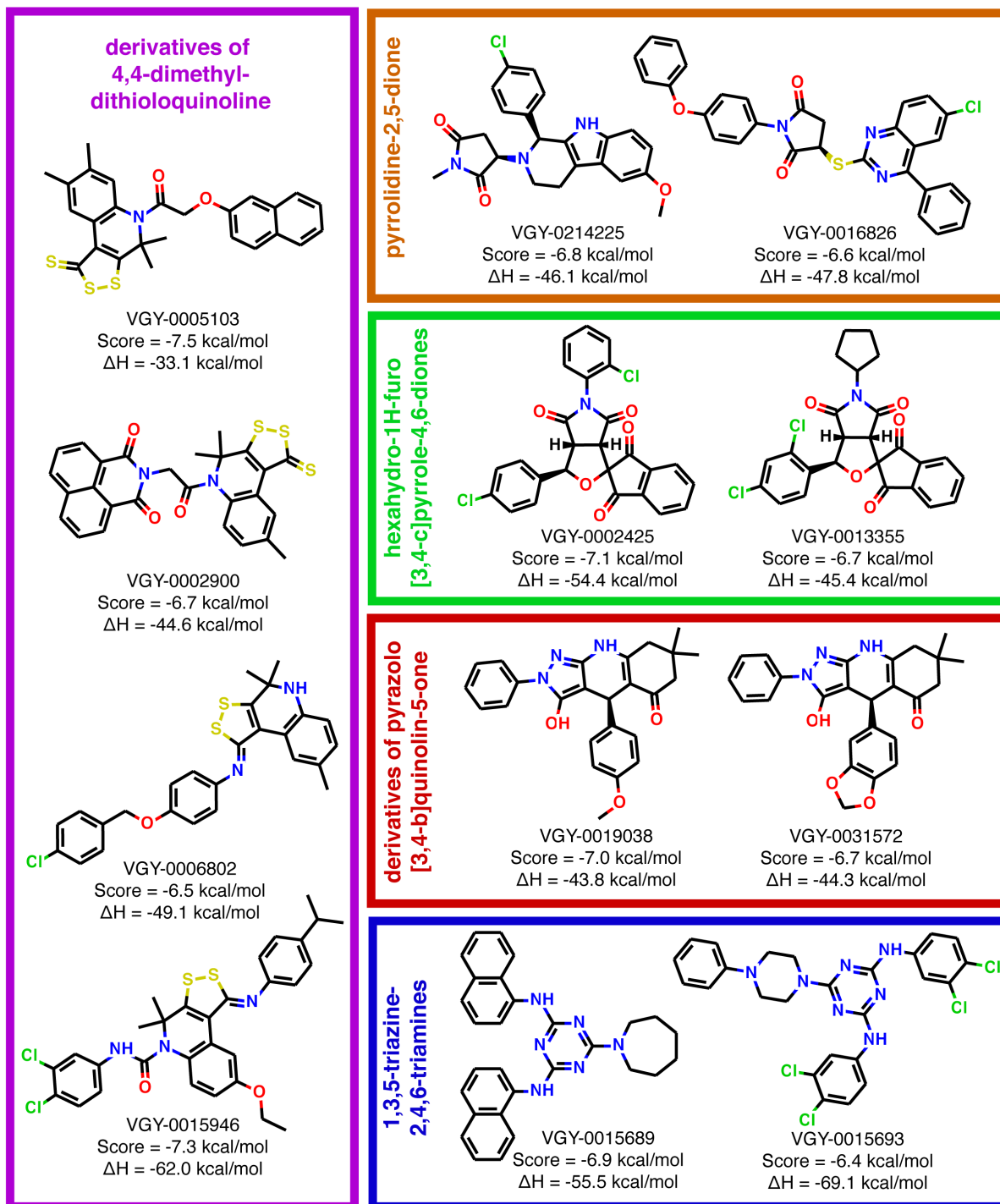
Overall about 178000 CPU\*hours have been spent for this virtual screening.

## 2. Results

Screening of the prepared database results in 1045 primary docking hits with SOL score values more negative than  $-6.3$  kcal/mol which is the value of the SOL score for the native ligand of the complex 6W63 (see above). All these virtual hits are subjected to quantum-chemical postprocessing to perform local optimization of best docked poses and to estimate their binding enthalpies with a more rigorous computational method. According to calculated enthalpies, the list of potential hits is narrowed down – only those compounds are kept which have binding enthalpies better (more negative) or slightly worse than binding enthalpy calculated for the X77 native ligand ( $-58.5$  kcal/mol). Selection of molecules with worse enthalpies is justified by values obtained for other known non-covalent M<sup>Pro</sup> inhibitors which turn to be dramatically less negative than enthalpy of X77 (see Fig. 1). For the final set of 87 candidates to become M<sup>Pro</sup> inhibitors, predicted bound conformations are visually checked for the presence of specific contacts with SARS-CoV-2 M<sup>Pro</sup> as well as the ability to block its catalytic dyad. Moreover, similarity between these compounds is also visually assessed to form a list of best chemically diverse candidates. The inspection of similarity reveals the high number of hexahydro-1H-furo[3,4-c]pyrrole-4,6-diones: 37 compounds of 87 virtual hits possess this fragment. The second most common chemical class among top compounds is derivatives of pyrazolo[3,4-b]quinolin-5-one: seven compounds contain this scaffold. Alike compounds with similar binding modes are removed keeping the best one in terms of the presence of specific contacts with M<sup>Pro</sup> and calculated values of the SOL score and the binding enthalpy.

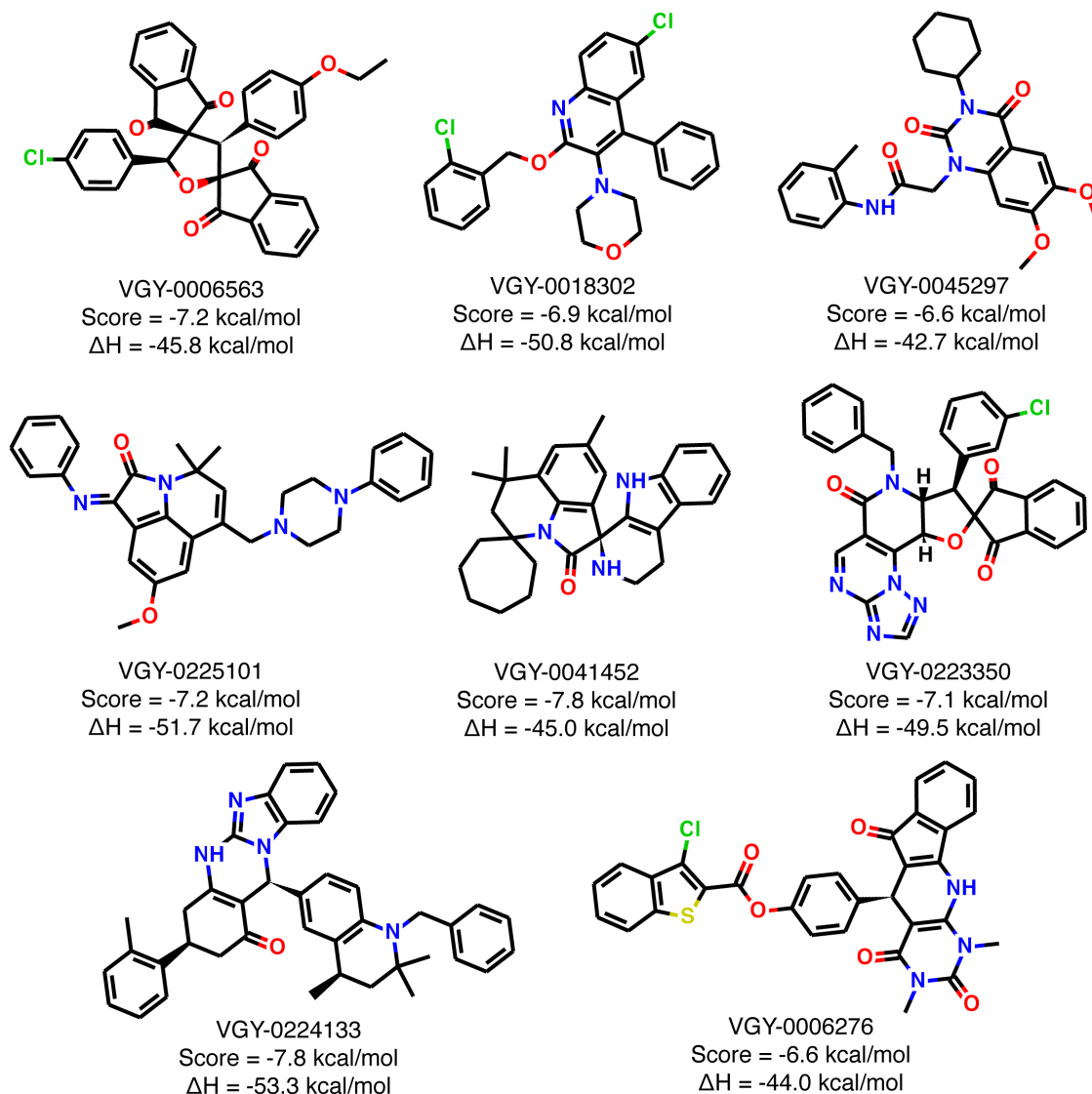
Relying upon results of modeling, observations of docking poses and visual estimation of similarity, we select 20 best candidate molecules as potential inhibitors of M<sup>Pro</sup>. All compounds in its predicted conformations with SARS-CoV-2 M<sup>Pro</sup> block the catalytic dyad of the enzyme. Their structures are listed in Fig. 2 and Fig. 3. Some molecules can be grouped into clusters according to the nature of their scaffold. Four compounds (VGY-0002900, VGY-0006802, VGY-0005103, and VGY-0015946) constitute class of 4,4-dimethyl-dithioquinoline derivatives. Two compounds (VGY-0002425, VGY-0013355) are related to derivatives of hexahydro-1H-furo[3,4-c]pyrrole-4,6-dione. The similar scaffold, pyrrolidine-2,5-dione, can be found in other two candidates: VGY-0016826, VGY-0214225. It is noteworthy that one of known non-covalent M<sup>Pro</sup> inhibitors, *tideglusib* [23] (see Fig. 1), contains the similar aliphatic ring with a nitrogen atom placed between two carbonyl groups. Two candidates (VGY-0015689, VGY-0015693) contain 1,3,5-triazine-2,4,6-triamine as a central fragment. Two other compounds (VGY-019038 and VGY-0031572) belong to derivatives of pyrazolo[3,4-b]quinolin-5-one. Other selected molecules are unique and constitute singletons.





**Figure 2.** Structures of best potential  $M^{PrO}$  inhibitors grouped according to the structure of the scaffold

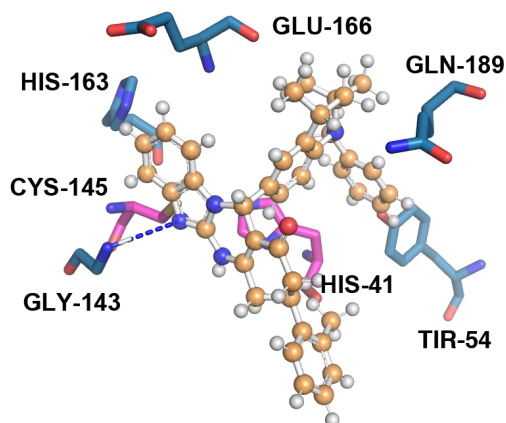
VGY-0224133 represents the best scored compound and belong to derivatives of imidazo[2,1-b]quinazolin-6-one. In its bound conformation predicted by docking and subsequent local optimization by PM7, it forms one H-bond with *Gly143* and pi-stacking interactions with three residues: *His41*, *Tyr54*, and *His163*. Its docking pose is shown in Fig. 4. The compound with the best predicted binding enthalpy, VGY-0015693, contains triazine-2,4,6-triamine scaffold linked to two benzene rings disubstituted with chlorine atoms which makes the molecule symmet-



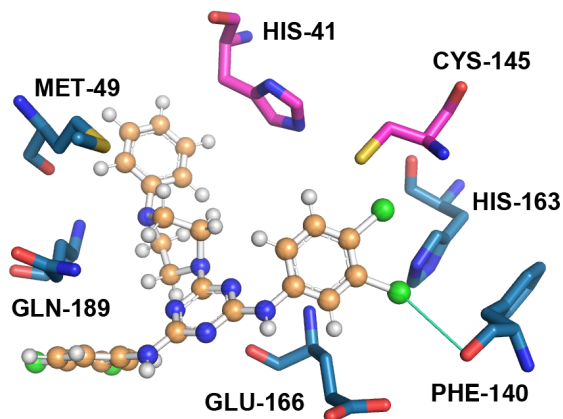
**Figure 3.** Structures of best potential M<sup>pro</sup> inhibitors with no grouping to any cluster

rical. According to geometry complex predicted by docking, VGY-0015693 forms three specific contacts: anion-pi interaction with *Glu166*, a halogen bond with *Phe140-O*, two pi-stacking with *His41* and *His163* (see Fig. 5).

Another peculiarity related to selected candidates is that above mentioned derivatives of 4,4-dimethyl-dithioloquinoline: VGY-0002900, VGY-0006802, VGY-0005103, VGY-0015946, share a disulfide moiety which similar to one found in disulfiram – a known covalent inhibitor of SARS-CoV-2 M<sup>pro</sup> with high thiol-reactiveness against cysteine residues [23, 31]. Because of this fact these compounds can be possible covalent modulators of M<sup>pro</sup> activity. Two other compounds (VGY-0015693, VGY-0225101) contain the piperazine fragment, the common moiety among potential virtual inhibitors of SARS-CoV-2 M<sup>pro</sup> we found by drug repurposing strategy with using the same computational protocol as we applied here. Activity of selected candidates is supposed to be checked *in vitro* against SARS-CoV-2 M<sup>pro</sup>.



**Figure 4.** Docking pose of VGY-0224133 after local optimization by PM7



**Figure 5.** Docking pose of VGY-0015693 after local optimization by PM7

## Conclusion

In search of non-covalent inhibitors of SARS-CoV-2 main protease virtual screening of the database of drug-like molecules is performed. There are two criteria of the selection of most promising molecules candidates to become inhibitors. First criterion is the sufficiently negative value of the SOL docking score and the second one is the sufficiently negative value of the binding enthalpy calculated by the PM7 quantum-chemical semiempirical method with the COSMO implicit solvent model. Thresholds of the criteria are defined by the respective values of the native inhibitor crystallized with the SARS-CoV-2 M<sup>Pro</sup> in the PDB complex 6W63. 20 compounds are selected for further experimental validation which molecules are satisfied the two criteria both.

## Acknowledgments

The research is carried out using the equipment of the shared research facilities of HPC computing resources at Lomonosov Moscow State University, including the Lomonosov super-computer [62].

*This paper is distributed under the terms of the Creative Commons Attribution-Non Commercial 3.0 License which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is properly cited.*

## References

1. Abian, O., Ortega-Alarcon, D., Jimenez-Alesanco, A., et al.: Structural stability of SARS-CoV-2 3CL<sup>Pro</sup> and identification of quercetin as an inhibitor by experimental screening. *International journal of biological macromolecules* 164, 1693–1703 (2020), DOI: 10.1016/j.ijbiomac.2020.07.235
2. Anand, K., Yang, H., Bartlam, M., et al.: Coronavirus main proteinase: target for antiviral drug therapy BT – Coronaviruses with Special Emphasis on First Insights Concerning SARS, pp. 173–199. Birkhäuser Basel (2005), DOI: 10.1007/3-7643-7339-3\_9

- Anand, K., Ziebuhr, J., Wadhwani, P., et al.: Coronavirus main proteinase (3CL<sup>pro</sup>) structure: basis for design of anti-SARS drugs. *Science (New York, N.Y.)* 300(5626), 1763–1767 (2003), DOI: 10.1126/science.1085658
- Beloglazova, I.B., Plekhanova, O.S., Katkova, E.V., et al.: Molecular modeling as a new approach to the development of urokinase inhibitors. *Bulletin of Experimental Biology and Medicine* 158(5), 700–704 (2015), DOI: 10.1007/s10517-015-2839-3
- Berman, H.M., Westbrook, J., Feng, Z., et al.: The Protein Data Bank. *Nucleic Acids Research* 28(1), 235–242 (2000), DOI: 10.1093/nar/28.1.235
- Calligari, P., Bobone, S., Ricci, G., Bocedi, A.: Molecular Investigation of SARS-CoV-2 Proteins and Their Interactions with Antiviral Drugs. *Viruses* 12(4), 445 (2020), DOI: 10.3390/v12040445
- ChemAxon software. <http://www.chemaxon.com>, accessed: 2020-10-01
- Damm-Ganamet, K.L., Smith, R.D., Dunbar Jr., J.B., et al.: CSAR Benchmark Exercise 2011–2012: Evaluation of Results from Docking and Relative Ranking of Blinded Congeneric Series. *J. Chem. Inf. and Model.* 53, 1853–1870 (2013), DOI: 10.1021/ci400025f
- Davies, M., Nowotka, M., Papadatos, G., et al.: ChEMBL web services: streamlining access to drug discovery data and utilities. *Nucleic acids research* 43(W1), W612–W620 (2015), DOI: 10.1093/nar/gkv352
- Drosten, C., Günther, S., Preiser, W., et al.: Identification of a Novel Coronavirus in Patients with Severe Acute Respiratory Syndrome. *New England Journal of Medicine* 348(20), 1967–1976 (2003), DOI: 10.1056/NEJMoa030747
- Eleftheriou, P., Amanatidou, D., Petrou, A., Geronikaki, A.: In Silico Evaluation of the Effectivity of Approved Protease Inhibitors against the Main Protease of the Novel SARS-CoV-2 Virus. *Molecules* 25(11) (2020), DOI: 10.3390/molecules25112529
- Elsevier. Reaxys Database. 2020. <https://www.reaxys.com>, accessed: 2020-10-01
- Friesner, R.A., Banks, J.L., Murphy, R.B., et al.: Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *Journal of Medicinal Chemistry* 47(7), 1739–1749 (2004), DOI: 10.1021/jm0306430
- Gao, J., Zhang, L., Liu, X., et al.: Repurposing Low-Molecular-Weight Drugs against the Main Protease of Severe Acute Respiratory Syndrome Coronavirus 2. *The Journal of Physical Chemistry Letters* 11(17), 7267–7272 (2020), DOI: 10.1021/acs.jpcllett.0c01894
- Gentile, D., Patamia, V., Scala, A., et al.: Putative Inhibitors of SARS-CoV-2 Main Protease from A Library of Marine Natural Products: A Virtual Screening and Molecular Modeling Study. *Marine drugs* 18(4) (2020), DOI: 10.3390/md18040225
- Gentile, F., Agrawal, V., Hsing, M., et al.: Deep Docking: A Deep Learning Platform for Augmentation of Structure Based Drug Discovery. *ACS Central Science* 6(6), 939–949 (2020), DOI: 10.1021/acscentsci.0c00229

17. Gimeno, A., Mestres-Truyol, J., Ojeda-Montes, M.J., et al.: Prediction of Novel Inhibitors of the Main Protease (M-pro) of SARS-CoV-2 through Consensus Docking and Drug Reposition. *International Journal of Molecular Sciences* 21(11), 3793–3821 (2020), DOI: 10.3390/ijms21113793
18. Grimme, S.: Accurate description of van der Waals complexes by density functional theory including empirical corrections. *Journal of Computational Chemistry* 25(12), 1463–1473 (2004), DOI: 10.1002/jcc.20078
19. Gyebi, G.A., Ogunro, O.B., Adegunloye, A.P., et al.: Potential inhibitors of coronavirus 3-chymotrypsin-like protease (3CL<sup>Pro</sup>): an *in silico* screening of alkaloids and terpenoids from African medicinal plants. *Journal of Biomolecular Structure and Dynamics* pp. 1–13 (2020), DOI: 10.1080/07391102.2020.1764868
20. Halgren, T.A.: Merck molecular force field. *Journal of Computational Chemistry* 17(5-6), 490–641 (1996), DOI: 10.1002/(SICI)1096-987X(199604)17:5/6<490::AID-JCC1>3.0.CO;2-P
21. Hanwell, M.D., Curtis, D.E., Lonie, D.C., et al.: Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. *Journal of Cheminformatics* 4(1), 17 (2012), DOI: 10.1186/1758-2946-4-17
22. Ilin, I.S., Lipets, E.N., Sulimov, A.V., et al.: New factor Xa inhibitors based on 1,2,3,4-tetrahydroquinoline developed by molecular modelling. *Journal of Molecular Graphics and Modelling* 89, 215–224 (2019), DOI: 10.1016/j.jmgm.2019.03.017
23. Jin, Z., Du, X., Xu, Y., et al.: Structure of M<sup>Pro</sup> from SARS-CoV-2 and discovery of its inhibitors. *Nature* 582(7811), 289–293 (2020), DOI: 10.1038/s41586-020-2223-y
24. Jurecka, P., Cerny, J., Hobza, P., Salahub, D.R.: Density functional theory augmented with an empirical dispersion term. Interaction energies and geometries of 80 noncovalent complexes compared with *ab initio* quantum mechanics calculations. *Journal of Computational Chemistry* 28(2), 555–569 (2007), DOI: 10.1002/jcc.20570
25. Kartsev, V., Shikhaliev, K.S., Geronikaki, A., et al.: Appendix A. dithioloquinolinethiones as new potential multitargeted antibacterial and antifungal agents: Synthesis, biological evaluation and molecular docking studies. *European Journal of Medicinal Chemistry* 175, 201–214 (2019), DOI: 10.1016/j.ejmech.2019.04.046
26. Khan, S.A., Zia, K., Ashraf, S., et al.: Identification of chymotrypsin-like protease inhibitors of SARS-CoV-2 via integrated computational approach. *Journal of Biomolecular Structure and Dynamics* pp. 1–10 (2020), DOI: 10.1080/07391102.2020.1751298
27. Klamt, A., Schuurmann, G.: COSMO: a new approach to dielectric screening in solvents with explicit expressions for the screening energy and its gradient. *Journal of the Chemical Society, Perkin Transactions 2* (5), 799–805 (1993), DOI: 10.1039/P29930000799
28. Klamt, A.: Conductor-like Screening Model for Real Solvents: A New Approach to the Quantitative Calculation of Solvation Phenomena. *The Journal of Physical Chemistry* 99(7), 2224–2235 (1995), DOI: 10.1021/j100007a062











29. Ksiazek, T.G., Erdman, D., Goldsmith, C.S., et al.: A novel coronavirus associated with severe acute respiratory syndrome. *The New England journal of medicine* 348(20), 1953–1966 (2003), DOI: 10.1056/NEJMoa030781
30. Kutov, D.C., Katkova, E.V., Kondakova, O.A., et al.: Influence of the method of hydrogen atoms incorporation into the target protein on the protein-ligand binding energy. *Bulletin of the South Ural State University, Ser. Mathematical Modelling, Programming & Computer Software* 10(3), 94–107 (2017), DOI: 10.14529/mmp170308
31. Lin, M.H., Moses, D.C., Hsieh, C.H., et al.: Disulfiram can inhibit MERS and SARS coronavirus papain-like proteases via different modes 150, 155–163 (2018), DOI: 10.1016/j.antiviral.2017.12.015
32. Liu, X., Wang, X.J.: Potential inhibitors against 2019-nCoV coronavirus M protease from clinically approved medicines. *Journal of Genetics and Genomics* 47(2), 119–121 (2020), DOI: 10.1016/j.jgg.2020.02.001
33. McGann, M.: FRED and HYBRID docking performance on standardized datasets. *J Comput Aided Mol Des* 26(8), 897–906 (2012), DOI: 10.1007/s10822-012-9584-8
34. Medvedeva, S.M., Potapov, A.Y., Gribkova, I.V., et al.: Synthesis, docking, and anticoagulant activity of new factor-Xa inhibitors in a series of pyrrolo[3,2,1-ij]quinoline-1,2-diones. *Pharmaceutical Chemistry Journal* 51(11), 975–979 (2018), DOI: 10.1007/s11094-018-1726-4
35. Mesecar, A.D.: Structure of COVID-19 main protease bound to potent broad-spectrum non-covalent inhibitor X77, DOI: 10.2210/pdb6w63/pdb, accessed: 2020-10-01
36. Myint, S.H.: *Human Coronavirus Infections BT – The Coronaviridae*, pp. 389–401. Springer US (1995), DOI: 10.1007/978-1-4899-1531-3\_18
37. Novichikhina, N.P., Shestakov, A.S., Potapov, A.Y., et al.: Synthesis of 4H-pyrrolo[3,2,1-ij]quinoline-1,2-diones containing a piperazine fragment and study of their inhibitory properties against protein kinases. *Russ Chem Bull* 4, 787–792 (2020), DOI: 10.1007/s11172-020-2834-3
38. Novichikhina, N., Ilin, I., Tashchilova, A., et al.: Synthesis, Docking, and In Vitro Anticoagulant Activity Assay of Hybrid Derivatives of Pyrrolo[3,2,1-ij]Quinolin-2(1H)-one as New Inhibitors of Factor Xa and Factor XIa. *Molecules* 25(8), 1889 (2020), DOI: 10.3390/molecules25081889
39. O’Boyle, N.M., Banck, M., James, C.A., et al.: Open Babel: An open chemical toolbox. *Journal of Cheminformatics* 3(1), 33 (2011), DOI: 10.1186/1758-2946-3-33
40. Pihan, E., Colliandre, L., Guichou, J.F., Douguet, D.: e-Drug3D: 3D structure collections dedicated to drug repurposing and fragment-based drug design. *Bioinformatics* 28(11), 1540–1541 (2012), DOI: 10.1093/bioinformatics/bts186
41. Pillaiyar, T., Manickam, M., Namasivayam, V., et al.: An Overview of Severe Acute Respiratory Syndrome-Coronavirus (SARS-CoV) 3CL Protease Inhibitors: Peptidomimetics and Small Molecule Chemotherapy. *Journal of Medicinal Chemistry* 59(14), 6595–6628 (2016), DOI: 10.1021/acs.jmedchem.5b01461

42. Rathnayake, A.D., Zheng, J., Kim, Y., et al.: 3C-like protease inhibitors block coronavirus replication in vitro and improve survival in MERS-CoV-infected mice. *Science Translational Medicine* 12(557), eabc5332 (2020), DOI: 10.1126/scitranslmed.abc5332
43. Rezac, J., Fanfrlik, J., Salahub, D., Hobza, P.: Semiempirical Quantum Chemical PM6 Method Augmented by Dispersion and H-Bonding Correction Terms Reliably Describes Various Types of Noncovalent Complexes. *J Chem Theory Comput* 5(7), 1749–1760 (2009), DOI: 10.1021/ct9000922
44. Ruiz-Carmona, S., Alvarez-Garcia, D., Foloppe, N., et al.: rDock: A Fast, Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids. *PLOS Computational Biology* 10(4), 1–7 (2014), DOI: 10.1371/journal.pcbi.1003571
45. Sinauridze, E.I., Romanov, A.N., Gribkova, I.V., et al.: New Synthetic Thrombin Inhibitors: Molecular Design and Experimental Verification. *PLOS ONE* 6(5), 1–12 (2011), DOI: 10.1371/journal.pone.0019969
46. Stewart, J.J.: Optimization of parameters for semiempirical methods V: modification of NDDO approximations and application to 70 elements. *J Mol Model* 13(12), 1173–1213 (2007), DOI: 10.1007/s00894-007-0233-4
47. Stewart, J.J.: Optimization of parameters for semiempirical methods VI: more modifications to the NDDO approximations and re-optimization of parameters. *Journal of Molecular Modeling* 19(1), 1–32 (2013), DOI: 10.1007/s00894-012-1667-x
48. Stewart, J.J.P.: Stewart Computational Chemistry. MOPAC2016. <http://openmopac.net/MOPAC2016.html> (2016), accessed: 2020-10-01
49. Stewart, J.J.P.: Application of localized molecular orbitals to the solution of semiempirical self-consistent field equations. *International Journal of Quantum Chemistry* 58(2), 133–146 (1996), DOI: 10.1002/(SICI)1097-461X(1996)58:2<133::AID-QUA2>3.0.CO;2-Z
50. Stolpovskaya, N.V., Kruzhillin, A.A., Zorina, A.V., et al.: Synthesis of Substituted Aminopyrimidines as Novel Promising Tyrosine Kinase Inhibitors. *Russian journal of organic chemistry* 55(9), 1322–1328 (2019), DOI: 10.1134/S1070428019090094
51. Sulimov, A.V., Kutov, D.C., Katkova, E.V., Sulimov, V.B.: Combined docking with classical force field and quantum chemical semiempirical method PM7. *Advances in Bioinformatics* 2017 (2017), DOI: 10.1155/2017/7167691
52. Sulimov, A.V., Kutov, D.C., Katkova, E.V., et al.: New generation of docking programs: Supercomputer validation of force fields and quantum-chemical methods for docking. *Journal of Molecular Graphics and Modelling* 78, 139–147 (2017), DOI: 10.1016/j.jmgm.2017.10.007
53. Sulimov, A.V., Kutov, D.C., Oferkin, I.V., et al.: Application of the docking program SOL for CSAR benchmark. *Journal of Chemical Information and Modeling* 53(8), 1946–1956 (2013), DOI: 10.1021/ci400094h
54. Sulimov, V.B., Gribkova, I.V., Kochugaeva, M.P., et al.: Application of molecular modeling to development of new factor Xa inhibitors. *BioMed Research International* 2015 (2015), DOI: 10.1155/2015/120802

55. Sulimov, V.B., Katkova, E.V., Oferkin, I.V., et al.: Application of molecular modeling to urokinase inhibitors development. *BioMed Research International* 2014, 625176 (2014), DOI: 10.1155/2014/625176
56. Sulimov, V.B., Ilin, I.S., Kutov, D.C., Sulimov, A.V.: Development of docking programs for Lomonosov supercomputer. *Journal of the Turkish Chemical Society Section A: Chemistry* 7(1), 259–276 (2020), DOI: 10.18596/jotcsa.634130
57. Sulimov, V.B., Kutov, D.C., Sulimov, A.V.: Advances in docking. *Current Medicinal Chemistry* 26(42), 7555–7580 (2019), DOI: 10.2174/0929867325666180904115000
58. The Department of Organic Chemistry of Voronezh State University. <http://www.chem.vsu.ru/?req=/department/5/application/page.html>, accessed: 2020-10-01
59. Ton, A.T., Gentile, F., Hsing, M., et al.: Rapid Identification of Potential Inhibitors of SARS-CoV-2 Main Protease by Deep Docking of 1.3 Billion Compounds. *Molecular Informatics* 39, 2000028 (2020), DOI: 10.1002/minf.202000028
60. Trott, O., Olson, A.J.: AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry* 31(2), 455–461 (2010), DOI: 10.1002/jcc.21334
61. Tsuji, M.: Potential anti-SARS-CoV-2 drug candidates identified through virtual screening of the ChEMBL database for compounds that target the main coronavirus protease. *FEBS Open Bio* 10(6), 995–1004 (2020), DOI: 10.1002/2211-5463.12875
62. 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), DOI: 10.14529/jsfi190201
63. Vostrikova, T.V., Kalaev, V.N., Medvedeva, S.M., et al.: Synthesized organic compounds as growth stimulators for woody plants. *Periodico tche quimica* 17(35), 327–337 (2020)
64. Vostrikova, T.V., Kalaev, V.N., Potapov, A.Y., et al.: Use of new compounds of the quinoline series as effective stimulants of growth processes. *Periodico tche quimica* 17(35), 781–790 (2020)
65. Yang, H., Yang, M., Ding, Y., et al.: The crystal structures of severe acute respiratory syndrome virus main protease and its complex with an inhibitor. *Proceedings of the National Academy of Sciences of the United States of America* 100(23), 13190–13195 (2003), DOI: 10.1073/pnas.1835675100
66. Yang, J., Zhang, Y.: I-TASSER server: new development for protein structure and function predictions. *Nucleic acids research* 43(W1), W174–W181 (2015), DOI: 10.1093/nar/gkv342
67. Zaki, A., Boheemen, S., Bestebroer, T., et al.: Isolation of a Novel Coronavirus from a Man with Pneumonia in Saudi Arabia. *The New England journal of medicine* 367 (2012), DOI: 10.1056/NEJMoa1211721



# Computational Approaches to Identify a Hidden Pharmacological Potential in Large Chemical Libraries

*Dmitry S. Druzhilovskiy*<sup>1</sup> , *Leonid A. Stolbov*<sup>1</sup> , *Polina I. Savosina*<sup>1</sup> ,  
*Pavel V. Pogodin*<sup>1</sup> , *Dmitry A. Filimonov*<sup>1</sup> ,  
*Alexander V. Veselovsky*<sup>1</sup> , *Karen Stefanisko*<sup>2</sup> , *Nadya I. Tarasova*<sup>2</sup> ,  
*Marc C. Nicklaus*<sup>3</sup> , *Vladimir V. Poroikov*<sup>1</sup> 

© The Authors 2020. This paper is published with open access at SuperFri.org

To improve the discovery of more effective and less toxic pharmaceutical agents, large virtual repositories of synthesizable molecules have been generated to increase the explored chemical-pharmacological space diversity. Such libraries include billions of structural formulae of drug-like molecules associated with data on synthetic schemes, required building blocks, estimated physical-chemical parameters, etc. Clearly, such repositories are “Big Data”. Thus, to identify the most promising compounds with the required pharmacological properties (hits) among billions of available opportunities, special computational methods are necessary. We have proposed using a combined computational approach, which combines structural similarity assessment, machine learning, and molecular modeling. Our approach has been validated in a project aimed at finding new pharmaceutical agents against HIV/AIDS and associated comorbidities from the Synthetically Accessible Virtual Inventory (SAVI), a 1.75 billion compound database. Potential inhibitors of HIV-1 protease and reverse transcriptase and agonists of toll-like receptors and STING, affecting innate immunity, were computationally identified. The activity of the three synthesized compounds has been confirmed in a cell-based assay. These compounds belong to the chemical classes, in which the agonistic effect on TLR 7/8 had not been previously shown. Synthesis and biological testing of several dozens of compounds with predicted antiretroviral activity are currently taking place at the NCI/NIH. We also carried out virtual screening among one billion substances to find compounds potentially possessing anti-SARS-CoV-2 activity. The selected hits’ information has been accepted by the European Initiative “JEDI Grand Challenge against COVID-19” for synthesis and further biological evaluation. The possibilities and limitations of the approach are discussed.

*Keywords:* drug discovery, chemical-pharmacological space, big data analysis, similarity assessment, machine learning, molecular modeling, virtual screening, HIV/AIDS, SAVI, COVID-19.

## Introduction

Discovery of new pharmaceutical agents is an unabated task of biomedical science because (a) there are no effective and safe drugs against many human diseases; (b) many existing drugs have a narrow therapeutic window due to severe side effects and toxicity; (c) application of drugs can lead to acquired resistance; (d) idiosyncratic and adverse effects restrict the use of specific therapies in particular patients [70].

The number of launched pharmaceutical substances is estimated at 15,000 worldwide, with several dozen new medicines approved every year [51]. About a million biologically active substances are under active study, but many belong to the same chemical series [14]. To increase the chemical-biological diversity of the investigated substances, in addition to the millions of already synthesized drug-like compounds [2, 11], a number of attempts to generate virtual libraries of the so-called “synthesizable molecules” have been carried out in recent years ([57, 59]

<sup>1</sup>Institute of Biomedical Chemistry (IBMC), Moscow, Russian Federation

<sup>2</sup>Laboratory of Cancer Immunometabolism, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Frederick, Maryland, USA

<sup>3</sup>Chemical Biology Laboratory, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Frederick, Maryland, USA

and some others). Such repositories of enumerated molecules include over billion structural formulae of products jointly with data on the possible synthetic routes, building blocks, estimated physical-chemical properties, cost of preparation, etc. The massive number of different chemical data offered by those libraries allows one to categorize them as “Big Data”. Since the number of known pharmacological targets is several thousand, the possible chemical-biological space’s dimensionality achieves about ten to the thirteenth power. Exploring such volumes of data requires developing particular computational methods, allowing to operate (store, retrieve and analyze) over all this structural information for identification of potential pharmacological agents with the required biological activity profiles.

We have developed an approach for analyzing large chemical databases and selecting promising substances based on the combined application of structural similarity assessment, analysis of the structure-activity relationships using machine learning, and molecular docking. This technology has been validated in our project dedicated to finding new biologically active compounds against HIV/AIDS and associated comorbidities in the Synthetically Accessible Virtual Inventory (SAVI) [59]. We showed that its application allows detecting the already known antiretroviral agents, which were found by overlap analysis of SAVI with PubChem [55]. This technology allowed us to select from SAVI some potential HIV-1 proteins inhibitors and TLR-7, TLR-8, and STING agonists, which affect the innate immunity. Activity of three predicted Toll-like receptor agonists that were synthesized has been experimentally confirmed; as of this writing, the NCI/NIH carries out synthesis and biological evaluation of the several dozens of other compounds.

The developed technology could be widely used to search for new pharmacological substances. In particular, in the context of the SARS-CoV-2/COVID-19 pandemic, we have conducted virtual screening of more than one billion accessible substances as part of the Joint European Disruptive Initiative (JEDI) Grand Challenge against COVID-19 to find compounds potentially possessing anticoronavirus activity [33]. Based on the prediction results, we selected potential inhibitors of SARS-CoV-2 proteins, including the main protease 3CLpro, papain-like protease PLpro, RNA dependent RNA polymerase RdRp, and human serine protease, TM-PRSS2, which is involved in virus-host interaction. Information about the selected compounds was passed on to the organizers of the JEDI Grand Challenge. We were included in the top 20 out of 130 participating groups; consequently, compounds proposed by our team were selected for the synthesis and biological activity evaluation.

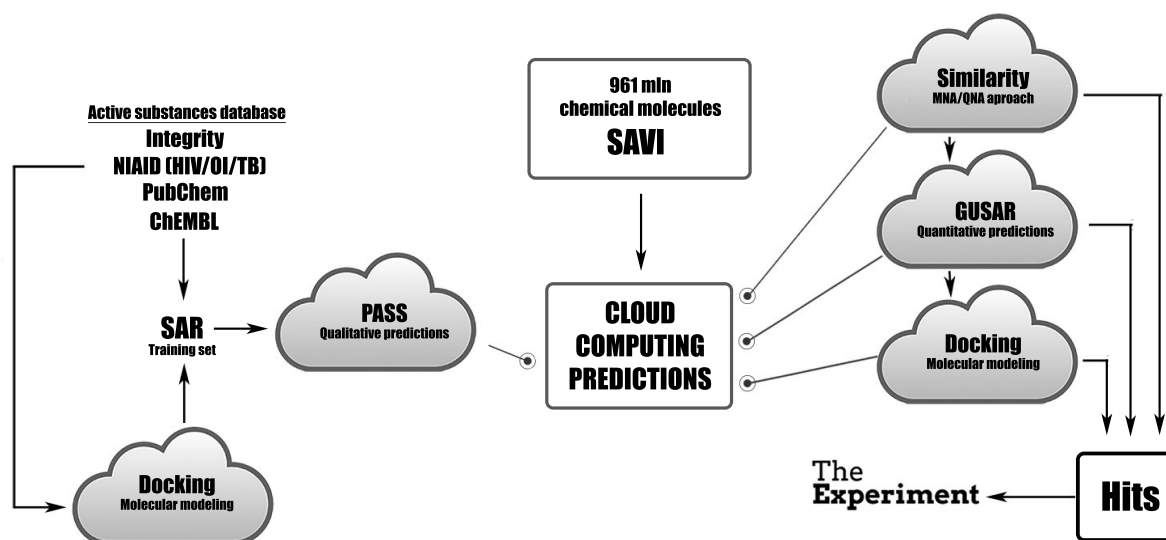
Our approach for *in silico* analysis of big chemical-pharmacological space and its practical validation is described below.

In section 1, we present the general workow that includes: (1) a storage system for a large library of chemical compounds; (2) procedure for creating the training sets for PASS based on publicly and commercially available databases on biologically active compounds and grouping the ligands according to different binding modes identified by supercomputer docking; and (3) selection of the most promising compounds with desirable biological activity (hits) by combining similarity assessment, machine learning, and docking. Section 2 describes our similarity assessment approach based on the original descriptors, which reect the essential structural and physical features of the ligand-target interactions providing the truthful structure-activity relationships analysis for heterogeneous datasets. Section 3 provides a machine learning method to elucidate the structure-activity relationships by analyzing the training sets, including the information about known biologically active compounds. Molecular docking as a method for

verification of selected hits, to predict the binding poses and estimate the affinity we described in Section 4. Section 5 presents the technical realization of data analysis, including the storage and processing systems of big data, software, and virtual environments involved in the study. In section 6, we report the *in silico* selection of new potential anti-HIV agents and innate immunity inducers with the experimental validation of our findings. Our attempt to identify novel antiviral agents, which could be investigated as potential medicines for SARS-CoV-2/COVID-19 therapy, is described in section 7. In the Conclusions section, we summarize the results of the study and point directions for further investigations.

## 1. General Workflow of the Approach

The general workflow of the proposed approach is presented in Fig. 1. The critical part of the process is computer program PASS (Prediction of Activity Spectra for Substances) [24], PASS currently predicts several thousand biological activities based on the analysis of a training set that includes over one million known biologically active compounds. To keep up with the state of biomedical and pharmaceutical science, we regularly update the training set, extracting new information about pharmaceutical agents from different databases, some public (ChEMBL [10], PubChem [55], etc.), others commercial (Clarivate Analytics CDDI [14], etc.). The training procedure includes leave-one-out cross-validation, which provides accuracy estimates for the obtained structure-activity relationships (SAR). To estimate the predictivity of those SAR models, 20-fold cross-validation is performed. In the standard version of PASS, both average accuracy and predictivity exceed 95%. The prediction's reliability can be improved by docking of ligands into a particular binding pocket and selecting best scoring compounds for the training set. It is particularly effective for protein targets with extended or multiple pockets as it allows selecting compounds binding to the same site of the protein.



**Figure 1.** General workflow of the large database analysis for identification of potential pharmacological substances

To select the most promising molecules in large virtual databases of synthesizable compounds (e.g., SAVI [59]) for synthesis and biological testing, three sequential *in silico* methods were applied. The work with the large volumes of data (see a more detailed description of SAVI in section 5) requires using cloud computing infrastructure for data storage and processing.

It should be emphasized that our task was not only to select molecules that have the desired types of biological activity, but also to enrich the initial SAVI library with new knowledge on structure-activity relationships, which increased the actual disk space requirements. In order to improve the quality of the storage environment for chemical information, an HP 3PAR hardware storage system was used. Using a fully connected full-mesh all-active cluster architecture, HP 3PAR provided stable performance in cases of load increase to the disk array and even load of array controllers. This solution allows simultaneous processing of data and metadata, and the use of SAS 15K drives made it possible to significantly speed up access to data stored with good performance.

Applying algorithms for biological activity prediction often requires filtering out of compounds from the general data set by appropriate threshold for molecular weight, amount of hydrogen donors/acceptors, stereochemistry, etc. in order to reduce the computational time for molecular modeling. Studying 3D structures of protein-ligand complexes of known ligands with the biological targets in question can suggest preliminary hypotheses about structural characteristics of the desired molecules. Such filtering approaches may significantly reduce the number of substances under study at the initial stage, but require the use of data indexing in order to increase the speed of selecting the lead compounds from billions of molecules with tens of billions of data items. Therefore, application of high-performance server solutions with parallel computing systems and SQL infrastructure deployed on them is necessary (see the description of technical realization in section 5).

We applied three methods to identify hits with the required biological activity: structural similarity assessment, prediction of biological activity with machine learning methods, and docking. The docking procedure requires significant computational resources; thus, at the first and the second stages of analysis, we used similarity assessment and machine learning to reduce the number of compounds that had to be analyzed by molecular modeling. The advantages and limitations of the methods are described in more detail below.

## 2. Similarity Assessment

“Similar molecules exert similar biological activities” [36]. Despite the occasionally observed violation of this rule in the case of so-called activity cliffs [17], it is widely used in medicinal chemistry to study the analogs of already known pharmaceutical agents having their pharmacological effect/biological target in mind [76]. Moreover, it is the “method-of-the-choice” in the case of novel pharmacological targets having a tiny number of known ligands to generate the (Q)SAR model.

There is no universal method for assessing the similarity between molecules belonging to different chemical classes and having various biological activities [6, 63]. In this study, we develop the method for similarity estimation based on our descriptors named Multilevel Neighborhoods of Atoms (MNA) [22] and Quantitative Neighborhoods of Atoms (QNA) [25]. These descriptors reflect the essential structural and physical features of the ligand-target interactions, as confirmed in many successful cases of structure-activity relationships analysis in heterogeneous datasets [51]. MNA and QNA descriptors differ from most other descriptors [71] because they are presented as unordered sets; in the case of MNA as character strings, i.e. linear notations of atoms with their neighborhoods; in the case of QNA as pairs of real numbers,  $P$  and  $Q$ , for each atom of the molecule.  $P$  and  $Q$  are calculated based on the connectivity matrix and

the standard values of the ionization potential and electron affinity of atoms in a molecule as described earlier [25].

For MNA descriptors, the well-known measure of similarity of two discrete sets:

$$T(A, B) = \frac{n(A \cap B)}{n(A \cup B)} \equiv \frac{n(A \cap B)}{n(A) + n(B) - n(A \cap B)}, \quad (1)$$

may be used where  $n(A \cap B)$  is the number of MNA descriptors at the intersection of the sets of descriptors of molecules  $A$  and  $B$ ;  $n(A \cup B)$  equivalently in their union.  $T(A, B)$  is a Jaccard measure proposed in 1901 [32] and also known as Tanimoto's similarity measure [69].

The peculiarities of QNA descriptors (each structure is described as the set of tuples having mutually dependent members) do not allow the straightforward use of the conventional similarity measures. Therefore, to assess similarity based on QNA descriptors, it is necessary to use other approaches to evaluate the similarity between complex chemical systems, e.g., those proposed by Todeschini [43]. To calculate the similarity of sets by Todeschini, the maximum similarity of each element of the set with elements of another set is used. The maximum contributions of all set elements are summed and then averaged over the total number of elements in both sets.

We propose an estimate  $F(A, B)$  of the similarity of molecular structures by QNA descriptors, using the Todeschini approach and Tanimoto's similarity measure, defined as:

$$F(A, B) = \frac{n(A \cap B)}{n(A) + n(B) - n(A \cap B)}, \quad (2)$$

where  $n(A) = N_A$  and  $n(B) = N_B$  is the number of pairs  $P$  and  $Q$  of the QNA descriptors of molecules  $A$  and  $B$ , respectively,  $n(A \cap B)$  is calculated as:

$$n(A \cap B) = \frac{1}{2} \left( \sum_A \max_{b \in B} [s_{ab}] + \sum_B \max_{a \in A} [s_{ba}] \right), \quad (3)$$

$$s_{ab} = \text{Exp}(-12N_B((P_a - P_b)^2 + (Q_a - Q_b)^2)), \quad (4)$$

$$s_{ba} = \text{Exp}(-12N_A((P_a - P_b)^2 + (Q_a - Q_b)^2)), \quad (5)$$

where  $s_{ab}$  and  $s_{ba}$  are the pairwise similarity of the QNA descriptor of atom  $a$  of molecule  $A$  and the QNA descriptor of atom  $b$  of molecule  $B$ ,  $P_a$  and  $Q_a$  are the QNA descriptor of atom  $a$  in molecule  $A$ ,  $P_b$  and  $Q_b$  are the QNA descriptor of atom  $b$  in molecule  $B$ . The multipliers  $12N_B$  and  $12N_A$  in the exponent have been chosen empirically. The proposed estimates of the similarity of the structures of drug-like compounds  $A$  and  $B$  based on our QNA descriptors are entirely new and do not have analogs.

To obtain quantitative estimates of biological activity for compounds based on these similarity estimates, we used the  $K$  nearest neighbor method, kNN, with weighting by the values of the similarity coefficients  $T(A, B)$  (1) and  $F(A, B)$  (2) according to the equations:

$$\hat{E}_T(A) = \frac{\sum_B T(A, B)E(B)}{\sum_B T(A, B)}, \quad \hat{E}_F(A) = \frac{\sum_B F(A, B)E(B)}{\sum_B F(A, B)}, \quad (6)$$

where  $\hat{E}_T(A)$  and  $\hat{E}_F(A)$  are the estimates of the biological activity of molecule  $A$  according to the amounts of known biological activity  $E(B)$  of molecule  $B$ , the summation is on the  $K$  nearest neighbors (maximum values of similarity), i.e. the set of molecules  $B$ , of the molecule  $A$ .

We had investigated the applicability of the proposed approach to the assessment of activity by similarity for 16,770 inhibitors of HIV-1 protease, reverse transcriptase, and integrase [66].

For all three targets, using both MNA and QNA descriptors, the best values of the mean square deviation (RSMD) and the coefficient of determination of the prediction ( $Q^2$ ) were obtained for the five nearest neighbors, 5NN.

**Table 1.** Values of  $Q^2$  based on similarity estimates by MNA and QNA descriptors at 5NN

Target	Number of structures	$Q^2$ , MNA	$Q^2$ , QNA
HIV-1 integrase	4072	0.7895	0.7946
HIV-1 protease	6390	0.8007	0.8052
HIV-1 protease	6308	0.6933	0.6980

The results obtained with the QNA descriptors outperformed those for the MNA descriptors, which may be explained by the better correspondence of the QNA descriptors to molecular recognition physics. The data presented in Tab. 1 are close to the performance of QSAR models, which were also analyzed [66]. However, our results demonstrated for the first time the applicability of a similarity search using QNA and MNA descriptors as an effective method for processing large databases.

### 3. Machine Learning Methods

In contrast to the biological activity prediction based on pairwise structural similarity, machine learning methods elucidate the structure-activity relationships by analysis of the training sets, including the information about known biologically active compounds [12]. To develop the (Q)SAR ((Quantitative) Structure-Activity Relationships) models, structures of the compounds from the training set should be presented as molecular descriptors [71]. If biological activity is described by quantitative values (IC50, EC50, LD50, etc.), regression QSAR models may be created. If only qualitative data on activity is available (the compound is categorized as either “active” or “inactive” categories), classification SAR models may be created. Best practices in creating (Q)SAR models have been described in several publications (see, e.g. [16, 30, 72]); extensive analysis of different QSAR issues have been presented in a recent review [45]. Initially, QSAR studies were performed with training sets of compounds active in one biological assay; in most cases, all compounds belonged to the same chemical classes [45]. Nowadays, multi-target (Q)SAR activity profiling of compounds is performed increasingly often. One of the first attempts to predict many kinds of biological activity *in silico* based on structural formulae is the computer program PASS (Prediction of Activity Spectra for Substances). A brief description of PASS follows.

#### 3.1. PASS Software

The development of PASS started in the late 1980s [9]. Its primary purpose was to develop a computational method for selecting the most promising substances among the drug-like compounds synthesized by different USSR institutions and to identify the most relevant pharmacological assays for the selected compounds. Since the compounds submitted for the State Registry [9] belonged to diverse chemical series and may have very different kinds of biological activity, it was necessary to develop a method for prediction of broad biological activity profiles based only on structural formulae. That is why our software has been described as: “One of the

earliest and most widely used examples of data-mining target elucidation is the continuously curated and expanded Prediction of Activity Spectra for Substances (PASS) software, which was assimilated from the bioactivities of more than 270,000 compound-ligand pairs” [44]. PASS’s current version predicts over five thousand biological activities based on the analysis of structure-activity relationships for 1,025,468 biologically active compounds. It uses MNA descriptors [22] and employs a modified naive Bayes classifier [26]. This method not only allows one to carry out high-accuracy SAR analysis for compounds from the training set but also is robust enough to provide reasonable estimates of the biological activity spectra of new compounds despite the incompleteness of information in the training set [51].

For a submitted compound, PASS estimates two probabilities:  $P_a$ , the probability of belonging to the subset of “actives”; and  $P_i$ , the probability of belonging to the subset of “inactives”. By default, all compounds, for which PASS predicts  $P_a > P_i$ , are considered to be “actives”.

Both an Invariant Accuracy of Prediction (IAP) determined in leave-one-out cross-validation and as well as the predictivity in 20-fold cross-validation (IAP20) exceed 0.96 averaged across all predicted activities. The PASS performance supersedes those of other known methods for predicting biological activity profiles, which has been shown in several benchmarking analyses [4, 28, 46].

The PASS Professional version allows creating new SAR bases, re-training the program to obtain new knowledge, and validating the accuracy and predictivity using leave-one-out and 20-fold cross-validation. Using this version of the program, we created specialized SAR bases for detecting potential anti-HIV agents in SAVI, which comprises inhibitors having similar binding modes against the main HIV-1 targets. Those inhibitors were selected using docking conducted with the ICM software [47] on the NIH Blue Gene supercomputer.

For this purpose, the protease and reverse transcriptase inhibitors, as well as STING, TLR7, and TLR 8 receptor agonists from the ChEMBL [10], NIAID HIV/OI/TB [48] and Cortellis Drug Discovery Intelligence [14] databases were selected. Classification models based on this data were built using PASS as well as regression models with the GUSAR program (see below). We found that the best predictions were achieved using classification models. This result may be explained by the uneven distribution of the available data regarding quantitative characteristics of activity (bias towards highly active compounds). In order to correct for this displacement, we evaluated the spatial similarity of ligands based on docking for certain crystallized protein-ligand complexes from the Protein Data Bank (PDB) [53]. We selected the following 3D complexes for docking: for TLR7: 5GMH and 5ZSJ; for STING: 4LOH and 5BQX; for HIV-1 Reverse Transcriptase: 2ZD1, for HIV-1 Protease: 2R5P and 2O4P.

**Table 2.** Target-specific training sets based on docking

Target	PDB code	Number of compounds before docking	Number of compounds after docking
TLR7	5GMH and 5ZSJ	429	75
STING	4LOH and 5BQX	326	273
Reverse Transcriptase HIV-1	2ZD1	5877	4120
Protease HIV-1	2R5P and 2O4P	2054	1300

Docking by ICM led to a significant decrease of the number of active molecules in the training sets (Tab. 2), which in turn improved the IAP values estimated in 20-fold cross-validation to 0.99 for all training sets.

To analyze the biological potential of large chemical repositories in the billion-compound size range, a special command-line version of PASS (PASS CL) was developed. PASS CL can be applied in parallel to multiple sub-sets of the whole library to estimate biological activity profiles, and then the obtained results are combined.

### 3.2. GUSAR

QSAR (Quantitative Structure-Activity Relationships) methods are appropriate to perform further selection of compounds with the requested biological activity after utilization of similarity assessment and PASS-based prediction. We used our software GUSAR (General Unrestricted Structure-Activity Relationships) [78] for QSAR analyses based on the structural formulae of the compounds and data about their biological activity/property, to predict activity/property for new compounds. It can predict properties of organic compounds belonging to both homogeneous and heterogeneous chemical classes. The GUSAR program uses the QNA descriptors that describe the molecule as a set of tuples composed of real values  $\langle P, Q \rangle$  [25]. The  $P$  and  $Q$  values are calculated for each atom in a molecule under examination using the connectivity matrix and the standard values of the ionization potential and electron affinity of atoms in the molecule. The current version of GUSAR also uses specific physicochemical descriptors and the results of  $Pa-Pi$  prediction using the PASS algorithm for 3,663 types of activity and based on a training set of over 300,000 biologically active organic compounds. The GUSAR algorithm is based on the self-consistent regression (SCR) method [23]. In the current version of GUSAR, this algorithm is used in combination with the nearest neighbors evaluation and a radial basis function artificial neural network (RBF ANN) based on the SCR results to achieve a multiple-model consensus [37, 79]. A comparative study of the first version of the GUSAR program and CoMFA, CoMSIA, Golpe/GRID, HQSAR, and other widely used methods to construct QSAR models demonstrated the advantages of our approach [25]. Recently, in the Collaborative Modeling Project for Androgen Receptor Activity (CoMPARA), GUSAR estimations were shown to be very good [41].

## 4. Molecular Modeling

Molecular docking is widely used in today's virtual screening of new pharmaceutical agents [39, 42, 67]. In contrast to similarity assessment and machine learning, docking requires significantly more computational resources. Thus, we applied this method for the final verification of the limited number (several hundred to several thousand) of selected hits, to predict the binding poses and estimate the affinity (using the scoring function values). Docking was performed using the programs Dock 6.5 [73] and AutoDock Vina [5]. The cutoff of the scoring function for further selection of compounds was chosen as  $-65$  kcal/mol and  $-8.0$  kcal/mol for Dock 6.5 and AutoDock Vina, respectively. The selected binding poses were manually inspected for their ability to occupy accommodate the subpockets in the protein active sites and analyzed the binding features (H-bonds, steric and electrostatic complementarity).

Virtual screening by docking was performed using ICM-Pro software (Molsoft Corp.) [1]. All screens have been run as swamp job on NIH supercomputer Biowulf. Binding pockets have been



defined using ICM pocket finder [21]. Screening of databases larger than 200000 compounds has been performed in a fast mode with thoroughness = 1. Binding score cutoff for a particular pocket was determined by docking a known ligand for this pocket and adding 5 units to the determined score. 300–500 best scoring compounds were redocked in a thorough mode that tested significantly higher number of poses and compound conformations. 30–40 best scoring hits from the thorough screen were subjected to manual docking with pose evaluation. Compounds with lowest scores have been synthesized and tested.

## 5. Technical Realization of the Big Chemical Data Analysis

The work with the large volumes of SAVI data required the use of cloud computing infrastructure for data storage and processing. Each unique compound is characterized by 62 descriptors describing the initial reagents used (identifiers in the Enamine catalog, etc.), the possible reaction (conditions, protection, expected yield, an estimate of the synthesis cost, etc.), and chemical properties' estimations seen as important for drug development (including "rule of three", "Lipinski's rule of five", n-octanol/water partition coefficient, the share of sp<sup>3</sup>-hybridized carbon atoms, topological polar surface area, prediction of genotoxicity, etc.). Thus, the amount of different records in the SAVI chemical library is more than ten billion, and the total amount of SD files requires more than 12 terabytes of the disk array.

It should be emphasized that our task was not only to select molecules that have the desired types of biological activity, but also to enrich the SAVI database with new knowledge on structure-activity relationships, which to large extent increased the actual disk space requirements. In order to improve the performance of the environment for chemical information storage, HP 3PAR hardware storage system was used. Using a fully connected full-mesh all-active cluster architecture, HP 3PAR system provides stable performance in cases of load increase to the disk array and even load of array controllers. This solution allows simultaneous processing of data and metadata, and the use of the SAS 15K drives made it possible to significantly speed up access to the stored data with good performance.

The application of algorithms for biological activity prediction often requires preliminary selection of compounds from the general data set by some meaningful threshold value, such as molecular weight, amount of hydrogen donors/acceptors, stereochemistry, and much more. Studying the data about the interaction of already known chemical compounds with biological targets are based on crystallography methods, preliminary hypotheses can be suggested about the structural characteristics of the desired molecules. This approach can significantly reduce the number of substances under study at the initial stage but requires the use of the data indexing procedure to increase the speed of selecting the lead compounds. Therefore, the application of high-performance server solutions with parallel computing systems and SQL infrastructure deployed on them is necessary.

Cloud solutions from VMWare for server virtualization were used as a computing cluster in IBMC. Hosts based on the 9th generation Hewlett-Packard Enterprise server line with Intel Xeon E5-2600 v4 family processors with 216 cores were used as the physical component of the cloud solution. Direct Fiber Channel switching with a total bandwidth of up to 64 Gbps was deployed between the compute hosts involved in building the cloud SQL infrastructure and the HP 3PAR storage system.

The SQL infrastructure based on the MySQL relational database management system was deployed due to the need to use fields for the BLOB (Binary Large Object) data type as a

container for the MOL format representation of structural formulae. The infrastructure binds ten cloud-based SQL servers containing information structured according to the compounds molecular weight and the types of transformations developed by Lhasa [38], which were used to generate chemical structures. Such data arrangement reduces the load on the cloud solution processing power by distributing the final SQL queries following the type of data. On the other hand, it allows the users, taking into account the particular type of data, to work within one server without affecting server utilization by the others.

At present, the approach we have applied allowed us in 24 hours to upload, standardize and finalize the preliminary data for performing computer prediction within the framework of one type of biological activity using more than one billion virtual molecules.

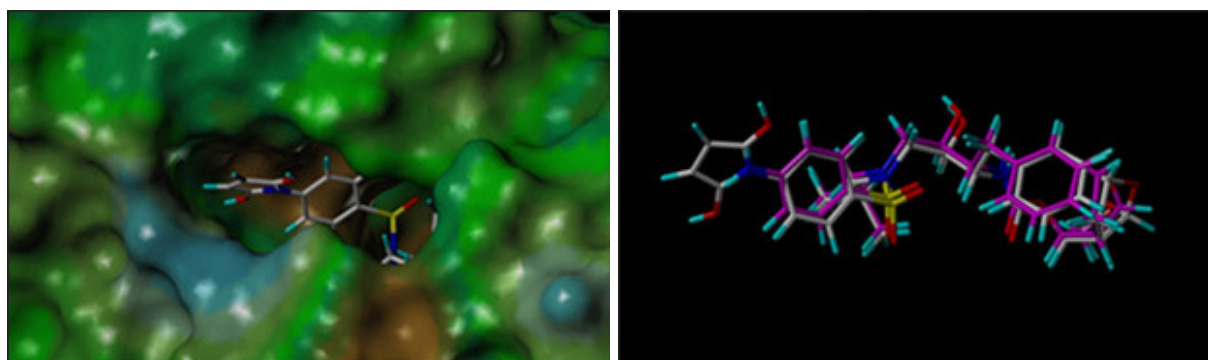
## **6. Identification of New Potential Anti-HIV Agents by Analysis of SAVI**

In 2014, the NCI/NIH Computer-Aided Drug Design (CADD) Group brought together a team of researchers from both academy and industry to launch a project to create the SAVI (Synthetically Accessible Virtual Inventory) library. SAVI is a vast virtual library of new molecules predicted to be easily synthesizable and contains among other data, information on the synthetic routes and the available reagents [49, 59, 60].

Its 2016 first complete file series contained over 283 million structures of new, easily synthesizable organic molecules, meant for *in silico* screening for new pharmacological substances. The current SAVI-2020 release has 1.75 billion generated products with reactions [60]. For this release, 53 transforms were applied to approximately 150,000 building blocks in single-step reactions. A more detailed description of the SAVI project is presented in [49, 59, 60].

Our study aimed to identify substances in SAVI that could be potentially useful in treating HIV/AIDS and HIV-associated disorders based on the prediction of their interaction with molecular targets. We had developed an algorithm for comparing large chemical databases based on the representation of structural formulas in SMILES codes, and evaluated the possibility of detecting new antiretroviral compounds in the SAVI database [61]. By analyzing the intersection of the 283 million 2016 SAVI structures with 97 million structures of the PubChem database [55] we found that only a small part of SAVI (0.015%) is represented in PubChem, which indicates a significant novelty of this virtual library. On the other hand, among those structures, 632 compounds that had been tested for anti-HIV activity were detected, and 41 had the desired activity. A comparison of the structures of these active antiretroviral compounds with the database of commercially available samples in the ZINC database [80] showed that most of these compounds can be obtained from various suppliers. Thus, our studies validated SAVI as a promising source for the search for new anti-HIV compounds [61].

We then analyzed more than 961 million unique structural formulae of drug-like compounds in (an early version of) the SAVI-2020 library using the algorithm presented in Fig. 1. This allowed us to select a number of potential HIV-1 protease inhibitors (53 compounds) and HIV-1 reverse transcriptase inhibitors (48 compounds), as well as TLR 7 receptor (53 compounds), TLR 8 (1378 compounds), and STING (627 compounds) agonists from the SAVI library (TLR and STING agonists affect the innate immunity).



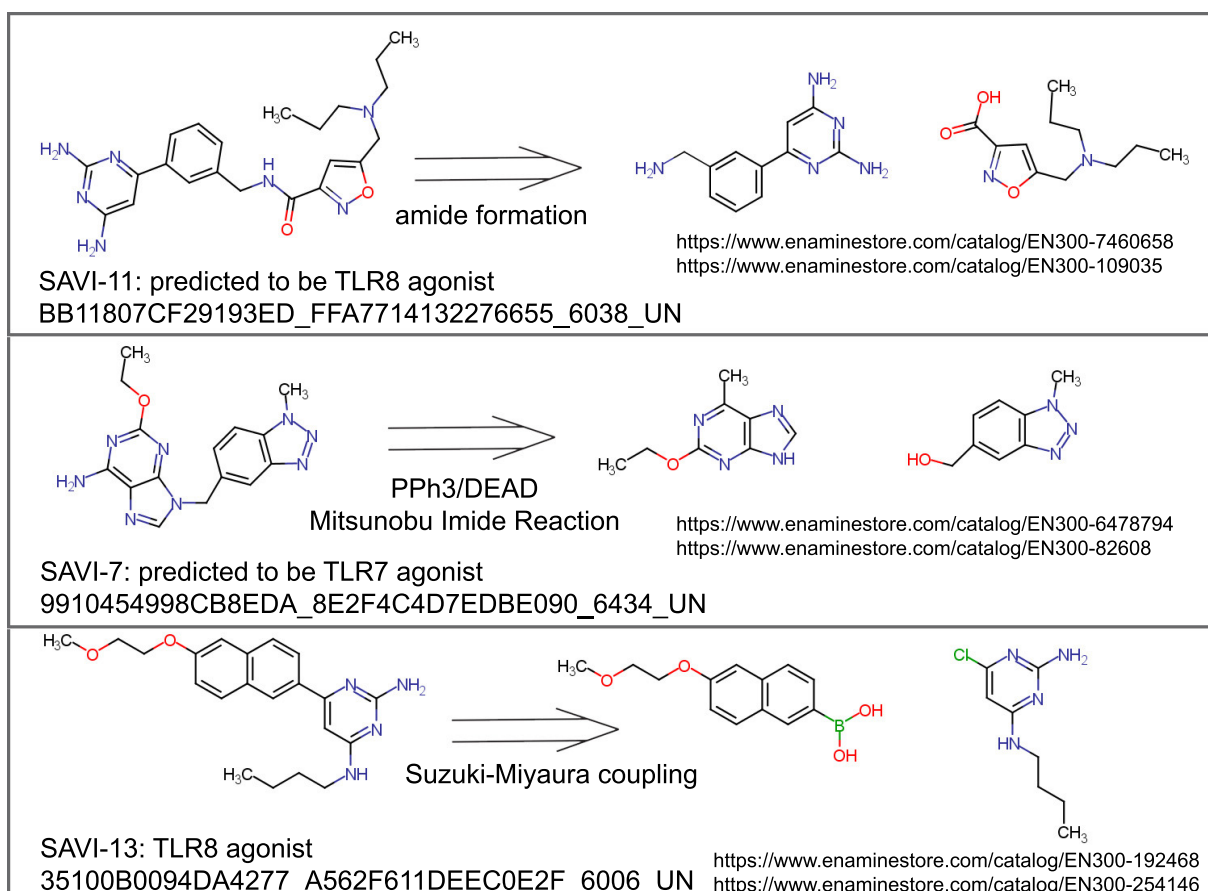
(a) Location of a SAVI molecule in the HIV-1 protease active center (colored according to hydrophobic potential)

(b) Superposition of this molecule with the known HIV-1 protease inhibitor Darunavir (magenta)

**Figure 2.** An example of a binding pose in the active site of HIV-1 protease

An example of a binding pose in the active site of HIV-1 protease of one molecule selected as a hit with is shown in Fig. 2a. The molecule (SAVI ID = 9A6A69BA66D806BA\_98763A2B6A65FDD7\_1031) fits well in the active site.

The superposition of this molecule with well-known HIV-1 protease inhibitor Darunavir is shown in Fig. 2b. Both structures are very similar (Tanimoto coefficient TC = 0.79).

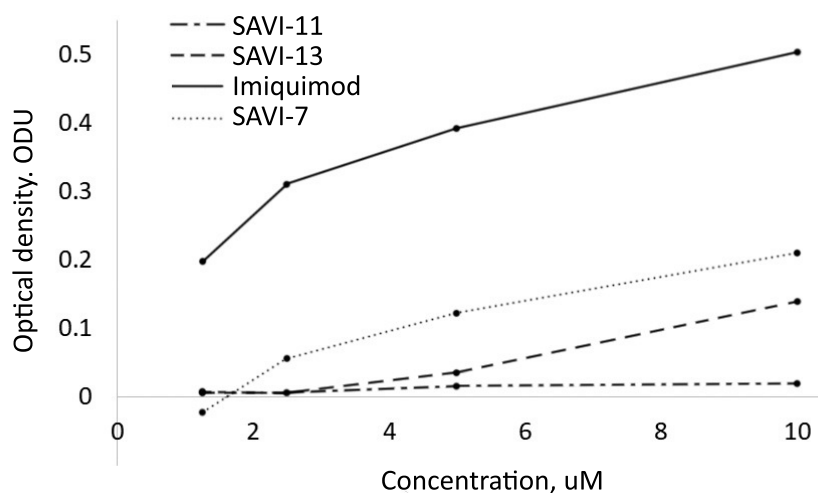


**Figure 3.** Three potential TLR 7/8 agonists selected for experimental testing (on the left – the structures of the products and their identifiers in SAVI; on the right – the starting reagents in the Enamine database; the type of chemical reaction is indicated under the arrow)

Chemical structures selected using our approach had been assessed at NCI/NIH to further synthesis and biological evaluation.

Three potential toll-like receptor 7/8 agonists had been synthesized by Enamines [20] using companys building blocks and synthesis reaction schemes presented in SAVI (Fig. 3).

Cell-based assay revealed that all synthesized compounds induced TLR-mediated activation of NF-kB signaling (Fig. 4). Imiquimod was used as a positive control in the assay since it is a potent TLR7 agonist. However, due to its high toxicity, only local use of this drug is allowed in clinical practice.



**Figure 4.** Activation of NF-kB signaling by the tested compounds (structural formulae are given in Fig. 3) in RAW-Blue reporter cells (reference drug – Imiquimod)

As can be appreciated from Fig. 4, the activity of the three potential agonists of toll-like receptors has been confirmed experimentally. Remarkably, the scaffolds of identified compounds have not been reported to function as TLR7/8 agonists before. Consequently, the finding expands the structural diversity of this important class of immune stimulants thus opening new opportunities for discovery of drugs with better pharmacological profiles. It also demonstrates the power of our ML approach that not just identifies close relatives of already known drugs as is frequently the case but allows for accurate predictions of agonists with diverse structures.

Currently, NCI/NIH continues their studies dedicated to the synthesis and biological testing of several dozen other molecules selected from SAVI using our approach as potential anti-HIV agents.

## 7. Identification of New Potential Anti-SAR-CoV-2 Agents

In 2020, humanity encountered a new global threat, the pandemic of Corona Virus Disease 19 (COVID-19), an infectious disease caused by the SARS-CoV-2 virus. In response to this challenge, many researchers worldwide rapidly initiated the search for medicines that could block the virus interaction with the human organism and its infectivity [7]. We are participating in the Joint European Disruptive Initiative (JEDI) “Grand Challenge against COVID-19” [33]. This call’s principal terms & conditions require performing virtual screening by three independent computational methods among more than one billion available compounds, including launched drugs. Our former experience in computer-aided predictions with SAVI enabled us to find hits with the required biological activities. We applied a similar approach to the JEDI Grand Chal-

lence as described above. We combined the data on structures from several databases, including ZINC [80], SAVI [60], AMS [2], SWEETLEAD [68], Antiviral CAS dataset [3], IBS Natural Compounds Set [31], and World Wide Approved Drugs [77]. After removing the duplicates, structures that do not correspond to the current QSAR applicability criteria [27], and molecules for which there is low chance of obtaining samples for experimental testing, we obtained a combined database of 1,080 billion molecules. This database was used for virtual screening to identify potential inhibitors of any of the targets listed below.

**3-chymotrypsin-like protease (3CLpro/Mpro).** The enzyme 3CLpro, also known as Nsp5, is the main proteolytic enzyme of SAR-CoV-2, playing a major role in its lifecycle. There are many 3D structures of this protease available in PDB [53]. At the beginning of the study structure the structure 6LU7 with inhibitor N3 was only available and it was selected as a target for the docking approach, since it contains the largest inhibitor, which is similar to the natural substrate. In progress of study, the available spatial structures were downloaded from the PDB and analyzed to determine the features participating in the binding of inhibitors. The preparation of the protein structure was done using SYBYL 8.1 suite (Tripos Inc., St. Louis, MO) and included: a) deletion of inhibitor, water, and cocrystallized ions; b) addition of hydrogens; c) atomic charge calculation using the Gasteiger-Hückel method; structure optimization by energy minimization in vacuum using the Tripos force field.

**Papain-like proteinase (PLpro).** PLpro is responsible for the cleavage of the N-terminus of the replicase polyprotein to release Nsp1, Nsp2, and Nsp3. Its function is essential for virus replication. The PDB structure 6WUU was selected as the target for molecular docking. The preparation for docking was the same as for 3CLpro.

**RNA-dependent RNA polymerase (RdRp).** Nsp12, a conserved protein in coronavirus, is an RNA-dependent RNA polymerase (RdRp) and a vital enzyme of coronavirus replication/transcription complex. The PDB structure 7BV2 was used in investigation. The water, Zn<sup>2+</sup> ions, inhibitor and pyrophosphate were deleted. Partial atomic charges were calculated using the Gasteiger-Hückel method; structure was optimized by energy minimization in vacuum using the Tripos force field.

**Human transmembrane peptidase serine 2 (TMPRSS2).** TMPRSS2 cleaves the SARS-CoV-2 spike protein, thus facilitating the infectivity of the virus. Unfortunately, no 3D structure of this protein is currently available. To perform a similarity search and selection of hits with the required biological activity from 1+ billion molecules, we identified the “reference substances” (the most active inhibitors of the four studied targets known in June 2020), used as queries. The following reference substances were used:

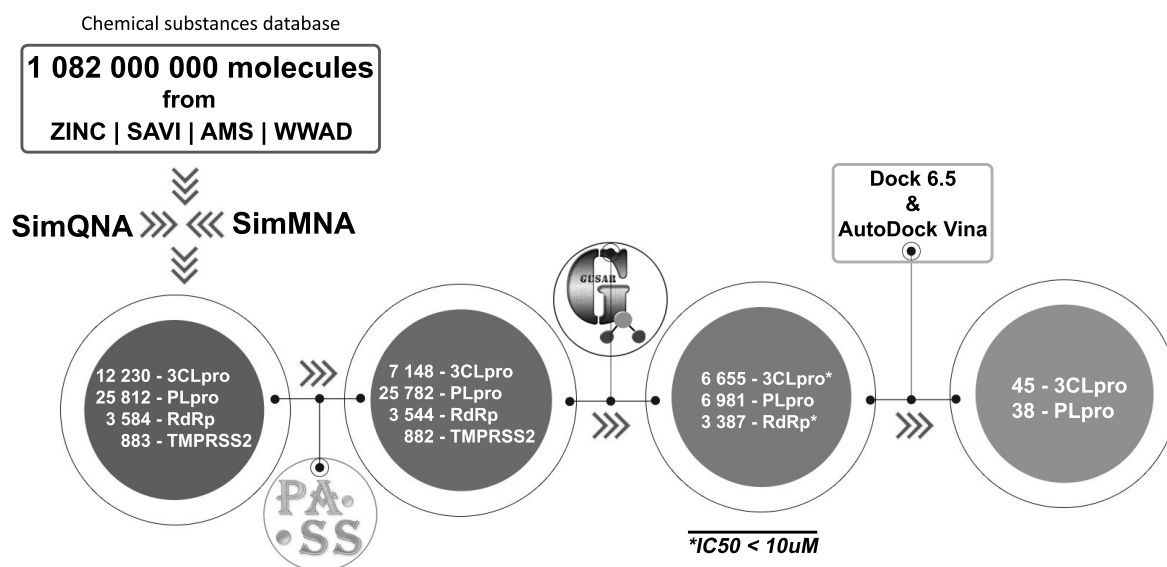
**3CLpro.** The five most active compounds were collected from different sources and tested under different experimental protocols. GC376, Tideglusib, 11b, TZDZ-8 activities were taken from the corresponding original publications [15, 34, 40]. MAT-POS-916a2c5a-1 was selected from the PostEra resource [52]. All of the five compounds were tested using SARS-CoV-2 recombinant main protease and showed low micromolar activities.

**PLpro.** 6-thioguanine, GRL0617, 679818, Psoralidin were taken from the corresponding original publications [13, 29, 35, 56] as most active inhibitors of SARS-CoV Papain-like protease.

**RdRp.** The selection of the most active compounds was carried out in the Stanford Coronavirus Antiviral Research Database [65]. Three chemical compounds were selected, their IDs in widely used databases, and common names are PubChem\_CID: 44468216 (GS-441524), PubChem\_CID: 121304016 (Remdesivir), ChEMBL\_ID: ChEMBL2178720 (Beta-D-

N4-Hydroxycytidine). The activity of GS-441524 and Remdesivir was reported in several preprints [8, 19, 54, 62, 75]. The data on the activity of the Beta-D-N4-Hydroxycytidine originates from a single preprint [62]. All three compounds demonstrated submicromolar activity (EC50) in the tests conducted using SARS-COV-2 and human cell lines to measure antiviral activity. The ability of Remdesivir and GS-441524 to suppress the expression of viral RNA was also studied in addition to the general antiviral effect, and compounds achieved submicromolar EC50 values.

**TMPRSS2.** The selection of the most active compounds was carried out in the ChEMBL database [10]. Three chemical compounds having submicromolar Ki values were found: ChEMBL1809250, ChEMBL1229259, and ChEMBL1809251. According to the assay description from ChEMBL, compounds were tested against the recombinant catalytic domain of TM-PRSS2 expressed in Escherichia coli using D-cyclohexylalanine-Pro-Arg-AMC as substrate by fluorescence plate reader analysis. Results were published in the paper [64]. Based on the assessment of MNA and QNA similarity for the reference molecules described above, we selected 42,509 hits, including 12,230 potential 3CLpro inhibitors; 25,812 potential PLpro inhibitors; 3,584 potential RdRp inhibitors; and 883 potential TM-PRSS2 inhibitors (Fig. 5).



**Figure 5.** General workflow and results of selection of anti-SARS-CoV-2 hits

Further selection was performed based on PASS predictions. As a result, we selected 7,148 potential 3CLpro inhibitors; 25,782 potential PLpro inhibitors; 3,544 potential RdRp inhibitors; and 882 potential TM-PRSS2 inhibitors.

For TM-PRSS2, the spatial structure is not available. Also, for the TM-PRSS2 inhibitors, we could not create both regression and classification models by GUSAR. Thus, this step of the selection was the final step.

Finally, the following potential inhibitors of SARS-CoV-2 proteins were selected: 45 against the main protease 3CLpro, 38 against the papain-like protease PLpro, 3,387 against RNA dependent RNA polymerase RdRp; 882 as potential inhibitors of the human serine protease TM-PRSS2.

Information about the selected compounds was passed on to the organizers of the JEDI Grand Challenge. After expert evaluation, our results were included in the shortlist of 20 out of 130 groups. Thus, compounds selected using our pipeline will be experimentally investigated [18].

## Conclusions

We have proposed an approach for the identification of potential pharmacological substances in very large databases of a billion or more drug-like compounds. The general workflow consists of three stages:

1. Chemical similarity assessment.
2. Prediction of biological activity using machine learning methods.
3. Visual inspection of the binding poses, and estimation of the scoring function using molecular docking.

This approach has been validated in two case studies: (1) identification of compounds potentially inhibiting HIV-1 protease and reverse transcriptase, or being agonists of TLR and STING, which induce the innate immunity, by virtual screening of SAVI; (2) detection of potential anti-SARS-CoV-2 agents by virtual screening of over one billion molecules collected from different available libraries in the context of the JEDI Grand Challenge project against COVID-19.

Synthesis of some selected molecules is currently being performed; these compounds will be evaluated in the appropriate biological assays at NCI/NIH. Three selected TLR 7/8 agonists have already synthesized and tested; experimental results confirmed the computational predictions.

These validations of our approach demonstrates its applicability to the analysis of large databases that significantly extend the available chemical-biological space and opens new opportunities to discover more potent and less toxic pharmaceutical agents.

## Acknowledgments

The work was financially supported by the Russian Foundation of Basic Research, grants No. 17-54-30015-NIH<sub>a</sub> and 20-04-60285. The work of NIT and MCN was supported by the Intramural Research Program of the National Institutes of Health, Center for Cancer Research, National Cancer Institute. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products or organizations imply endorsement by the US Government.

*This paper is distributed under the terms of the Creative Commons Attribution-Non Commercial 3.0 License which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is properly cited.*

## References

1. Abagyan, R., Totrov, M., Kuznetsov, D.: A new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. *J. Comp. Chem.* 15(5), 488–506 (1994), DOI: 10.1002/jcc.540150503
2. Aldrich Market Select (AMS). <https://www.sigmaaldrich.com/chemistry/chemistry-services/aldrich-market-select.html>, accessed: 2020-09-21
3. Antiviral CAS dataset. <https://www.cas.org/covid-19-antiviral-compounds-dataset>, accessed: 2020-09-21
4. Anusevicius, K., Mickevicius, V., Stasevych, M., et al.: Synthesis and chemoinformatics analysis of N-aryl-beta-alanine derivatives. *Research on Chemical Intermediates* 41(10),

7517–7540 (2015), DOI: 10.1007/s11164-014-1841-0

5. AutoDock Vina. <http://vina.scripps.edu/>, accessed: 2020-09-21
6. Bender, A.: How similar are those molecules after all? Use two descriptors and you will have three different answers. *Expert Opinion on Drug Discovery* 5(12), 1141–1151 (2010), DOI: 10.1517/17460441.2010.517832
7. Bobrowski, T., Melo-Filho, C., Korn, D., et al.: Learning from history: do not flatten the curve of antiviral research! *Drug Discovery Today* 25(9), 1604–1613 (2020), DOI: 10.1016/j.drudis.2020.07.008
8. Bojkova, D., McGreig, J., McLaughlin, K., et al.: SARS-CoV-2 and SARS-CoV differ in their cell tropism and drug sensitivity profiles. *bioRxiv* (2020), DOI: 10.1101/2020.04.03.024257
9. Burov, Y., Poroikov, V., Korolchenko, L.: National system for registration and biological testing of chemical compounds: facilities for new drugs search. *Bull. Natl. Center for Biologically Active Compounds* 1, 4–25 (1990)
10. ChEMBL database. <https://www.ebi.ac.uk/chembl/>, accessed: 2020-09-21
11. ChemNavigator. <https://www.chemnavigator.com/>, accessed: 2020-09-21
12. Cherkasov, C., Muratov, E., Fourches, D., et al.: QSAR modeling: where have you been? Where are you going to? *Journal of Medicinal Chemistry* 57(12), 4977–5010 (2014), DOI: 10.1021/jm4004285
13. Chou, C., Chien, C., Han, Y., et al.: Thiopurine analogues inhibit papain-like protease of severe acute respiratory syndrome coronavirus. *Biochemical Pharmacology* 75(8), 1601–1609 (2008), DOI: 10.1016/j.bcp.2008.01.005
14. Cortellis Drug Discovery Intelligence. <https://www.cortellis.com/drugdiscovery>, accessed: 2020-09-21
15. Dai, W., Zhang, B., Jiang, X., et al.: Structure-based design of antiviral drug candidates targeting the SARS-CoV-2 main protease. *Science* 368, 1331–1335 (2020), DOI: 10.1126/science.abb4489
16. Dearden, J., Kaiser, K.: How not to develop a quantitative structure-activity or structure-property relationship (QSAR/QSPR). *SAR and QSAR in environmental research* 20(3-4), 241–266 (2009), DOI: 10.1080/10629360902949567
17. Dimova, D., Bajorath, J.: Advances in Activity Cliff Research. *Molecular informatics* 35(5), 181–191 (2016), DOI: 10.1002/minf.201600023
18. Discord JEDI Chat. <https://discord.com/channels/694851986042126366/694851987208011818>, accessed: 2020-09-21
19. Ellinger, B., Bojkova, D., Zaliani, A., Cinatl, J., et al.: Identification of inhibitors of SARS-CoV-2 in-vitro cellular toxicity in human (Caco-2) cells using a large scale drug repurposing collection. *Research Square Preprint* (2020), DOI: 10.21203/rs.3.rs-23951/v1
20. Enamine Ltd. <https://enamine.net>, accessed: 2020-09-21



21. Fernandez-Recio, J., Totrov, M., Skorodumov, C., Abagyan, R.: Optimal docking area: a new method for predicting protein-protein interaction sites. *Proteins* 58(1), 134–143 (2005), DOI: 10.1002/prot.20285
22. Filimonov, D., Poroikov, V., Borodina, Y., Glorizova, T.: Chemical similarity assessment through multilevel neighborhoods of atoms: definition and comparison with the other descriptors. *Journal of Chemical Information and Computer Sciences* 39(4), 666–670 (1999), DOI: 10.1021/ci980335o
23. Filimonov, D., Akimov, D., Poroikov, V.: Method of self-consistent regression in analysis of quantitative structure-property relationships of chemical compounds. *Pharmaceutical Chemistry Journal* 38(1), 21–24 (2004), DOI: 10.1023/B:PHAC.0000027639.17115.5d
24. Filimonov, D., Poroikov, V., Gloziozova, T., Lagunin, A.: PASS program package, Certificate of Russian State Patent Agency, No. 2006613275 of 15.09.2006
25. Filimonov, D., Zakharov, A., Lagunin, A., Poroikov V.: QNA based “Star Track” QSAR approach. *SAR and QSAR in environmental research* 20(7-8), 679–709 (2009), DOI: 10.1080/10629360903438370
26. Filimonov, D., Druzhilovskiy, D., Lagunin, F., et al.: Computer-aided prediction of biological activity spectra for chemical compounds: opportunities and limitations. *Biomedical Chemistry: Research and Methods* 1(1), e00004 (2018), DOI: 10.18097/bmcrm00004
27. Fourches, D., Muratov, E., Tropsha, A.: Curation of chemogenomics data. *Nature Chemical Biology* 11(8), 535 (2015), DOI: 10.1038/nchembio.1881
28. Geronikaki, A., Druzhilovsky, D., Zakharov, A., Poroikov, V.: Computer-aided predictions for medicinal chemistry via Internet. *SAR and QSAR in environmental research* 19(1-2), 27–38 (2008), DOI: 10.1080/10629360701843649
29. Ghosh, A., 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. *Journal of Medicinal Chemistry* 52(16), 5228–5240 (2009), DOI: 10.1021/jm900611t
30. Gramatica, P.: On the development and validation of QSAR models. *Methods in Molecular Biology* 930, 499–526 (2013), DOI: 10.1007/978-1-62703-059-5\_21
31. InterBioScreen (IBS) Natural Compounds Set. <https://www.ibscreen.com>, accessed: 2020-09-21
32. Jaccard, P.: Distribution de la flore alpine dans le Bassin des Dranses et dans quelques regions voisines. *Bulletin de la Societe Vaudoise des Sciences Naturelles* 37(140), 241–272 (1901), DOI: 10.5169/seals-266440
33. JEDI Grand Challenge Against Covid-19. <https://www.covid19.jedi.group>, accessed: 2020-09-21
34. Jin, Z., Du, X., Xu, Y., et al.: Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. *Nature* 582, 289–293 (2020), DOI: 10.1038/s41586-020-2223-y

35. Kim, D., Seo, K., Curtis-Long, M., et al.: Phenolic phytochemical displaying SARS-CoV papain-like protease inhibition from the seeds of *Psoralea corylifolia*. *Journal of Enzyme Inhibition and Medicinal Chemistry* 29(1), 59–63 (2014), DOI: 10.3109/14756366.2012.753591
36. Kubinyi, H.: Chemical similarity and biological activities. *Journal of the Brazilian Chemical Society* 13(6), 717–726 (2002), DOI: 10.1590/S0103-50532002000600002
37. Lagunin, A., Romanova, M., Zadorozhny, A., et al.: Comparison of Quantitative and Qualitative (Q)SAR Models Created for the Prediction of  $K_i$  and  $IC_{50}$  Values of Antitarget Inhibitors. *Frontiers in Pharmacology* 9, 1138 (2018), DOI: 10.3389/fphar.2018.01136
38. Lhasa Ltd. <https://www.lhasalimited.org>, accessed: 2020-09-21
39. Lushchekina, S., Makhaeva, G., Novichkova, D., et al.: Supercomputer modeling of dual-site acetylcholinesterase (AChE) inhibition. *Supercomputing Frontiers and Innovations* 5(4), 89–97 (2018), DOI: 10.14529/jsfi1804
40. Ma, C., Sacco, M., Hurst, B., et al.: Boceprevir, GC-376, and calpain inhibitors II, XII inhibit SARS-CoV-2 viral replication by targeting the viral main protease. *Cell Research* 30, 678–692 (2020), DOI: 10.1038/s41422-020-0356-z
41. Mansouri, K., Kleinstreuer, N., Abdelaziz, A., et al.: CoMPARA: Collaborative Modeling Project for Androgen Receptor Activity. *Environmental Health Perspectives* 128(2), 27002 (2020), DOI: 10.1289/EHP5580
42. Maslova, V., Reshetnikov, R., Bezugolov, V., et al.: Supercomputer Simulations of Dopamine-Derived Ligands Complexed with Cyclooxygenases. *Supercomputing Frontiers and Innovations* 5(4), 98–102 (2018), DOI: 10.14529/jsfi1804
43. Mauri, A., Ballabio, D., Todeschini, R., Consonni, V.: Mixtures, metabolites, ionic liquids: a new measure to evaluate similarity between complex chemical systems. *Journal of Cheminformatics* 8, 49 (2016), DOI: 10.1186/s13321-016-0159-x
44. Mervin, L., Afzal, A., Drakakis, G., et al.: Target prediction utilising negative bioactivity data covering large chemical space. *Journal of Cheminformatics* 7, 51 (2015), DOI: 10.1186/s13321-015-0098-y
45. Muratov, E., Bajorath, J., Sheridan, R., et al.: QSAR without borders. *Chemical Society reviews* 49(11), 3525–3564 (2020), DOI: 10.1039/d0cs00098a
46. Murtazaliev, K., Druzhilovskiy, D., Goel, R., et al.: How good are publicly available web services that predict bioactivity profiles for drug repurposing? SAR and QSAR in environmental research 28(10), 843–862 (2017), DOI: 10.1080/1062936X.2017.1399448
47. Neves, M., Totrov, M., Abagyan, R.: Docking and scoring with ICM: the benchmarking results and strategies for improvement. *Journal of Computer-Aided Molecular Design* 26(6), 675–686 (2012), DOI: 10.1007/s10822-012-9547-0
48. National Institute of Allergy and Infectious Diseases (NIAID) HIV/OI/TB database. <https://chemdb.niaid.nih.gov>, accessed: 2020-09-21

49. Patel, H., Ihlenfeldt, W., Judson, P., et al.: Synthetically Accessible Virtual Inventory (SAVI). ChemRxiv Preprint (2020), DOI: 10.26434/chemrxiv.12185559.v1
50. Poroikov, V., Filimonov, D., Borodina, Y., et al.: Robustness of biological activity spectra predicting by computer program PASS for non-congeneric sets of chemical compounds. *Journal of Chemical Information and Computer Sciences* 40(6), 1349–1355 (2000), DOI: 10.1021/ci000383k
51. Poroikov, V.: Computer-aided drug design: from discovery of novel pharmaceutical agents to systems pharmacology. *Biochemistry (Moscow), Supplement Series B: Biomedical Chemistry* 14(3), 216–227 (2020), DOI: 10.1134/S1990750820030117
52. PostERA activity data. [https://postera.ai/covid/activity\\_data](https://postera.ai/covid/activity_data), accessed: 2020-09-21
53. Protein Data Bank (PDB). <https://www.rcsb.org>, accessed: 2020-09-21
54. Pruijssers, A., George, A., 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. *bioRxiv* (2020), DOI: 10.1101/2020.04.27.064279
55. PubChem. <https://pubchem.ncbi.nlm.nih.gov>, accessed: 2020-09-21
56. Ratia, K., Pegan, S., Takayama, J., et al.: HA noncovalent class of papain-like protease/deubiquitinase inhibitors blocks SARS virus replication. *Proceedings of the National Academy of Sciences* 105(42), 16119–16124 (2008), DOI: 10.1073/pnas.0805240105
57. REAL database. <https://enamine.net/library-synthesis/real-compounds/real-database>, accessed: 2020-09-21
58. Riva, L., Yuan, S., Yin, X., et al.: A large-scale drug repositioning survey for SARS-CoV-2 antivirals. *bioRxiv* (2020), DOI: 10.1101/2020.04.16.044016
59. SAVI: Synthetically Accessible Virtual Inventory. [https://cactus.nci.nih.gov/download/savi\\_download/](https://cactus.nci.nih.gov/download/savi_download/), accessed: 2020-09-21
60. SAVI-2020 dataset. DOI: 10.35115/37N9-5738
61. Savosina, P., Stolbov, L., Druzhilovskiy, D., et al.: Discovering new antiretroviral compounds in “Big Data” chemical space of the SAVI library. *Biomeditsinskaya Khimiya* 65(2), 73–79 (2019), DOI: 10.18097/PBMC20196502073
62. Sheahan, T., Sims, A., Zhou, S., et al.: An orally bioavailable broad-spectrum antiviral inhibits SARS-CoV-2 and multiple endemic, epidemic and bat coronavirus. *bioRxiv* (2020), DOI: 10.1101/2020.03.19.997890
63. Sheridan, R., Kearsley, S.: Why do we need so many chemical similarity search methods? *Drug Discovery Today* 7(17), 903–911 (2002), DOI: 10.1016/s1359-6446(02)02411-x
64. Sielaff, F., Böttcher-Friebertshäuser, E., Meyer, D., et al.: Development of substrate analogue inhibitors for the human airway trypsin-like protease HAT. *Bioorganic & Medicinal Chemistry Letters* 21(16), 4860–4864 (2011), DOI: 10.1016/j.bmcl.2011.06.033

65. Stanford Coronavirus Antiviral Research Database. <https://covdb.stanford.edu>, accessed: 2020-09-21
66. Stolbov, L., Druzhilovskiy, D., Filimonov, D., et al.: (Q)SAR models of HIV-1 proteins inhibition by drug-like compounds. *Molecules* 25(1), 87 (2020), DOI: 10.3390/molecules25010087
67. Sulimov, A., Kutov, D., Sulimov, V.: Supercomputer docking. *Supercomputing Frontiers and Innovations* 6(3), 25–50 (2019), DOI: 10.14529/jsfi190302
68. SWEETLEAD: A cheminformatics database of medicines, drugs, and herbal isolates. <https://simtk.org/projects/sweetlead>, accessed: 2020-09-21
69. Tanimoto, T.: *An Elementary Mathematical theory of Classification and Prediction*. International Business Machines Corporation (1958)
70. Wermuth, C., Aldous, D., Raboisson, P., et al.: *The Practice of Medicinal Chemistry*. Fourth edition. Academic Press 902 (2015), DOI: 10.1016/B978-0-12-374194-3.X0001-7
71. Todeschini, R., Consonni, V.: *Handbook of Molecular Descriptors*. Wiley-VCH (2008), DOI: 10.1002/9783527613106
72. Tropsha, A.: Best practices for QSAR model development, validation, and exploitation. *Molecular Informatics* 29(6-7), 476–488 (2010), DOI: 10.1002/minf.201000061
73. UCSF Dock. <http://dock.compbio.ucsf.edu>, accessed: 2020-09-21
74. Vuong, W., Khan, M., Fischer, C., et al.: Feline coronavirus drug inhibits the main protease of SARS-CoV-2 and blocks virus replication. *bioRxiv* (2020), DOI: 10.1101/2020.05.03.073080
75. Wang, M., Cao, R., Zhang, et al.: Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) in vitro. *Cell Research* 30(3), 269–271 (2020), DOI: 10.1038/s41422-020-0282-0
76. Wermuth, C.: Similarity in drugs: reflections on analogue design. *Drug Discovery Today* 11(7–8), 348–354 (2006), DOI: 10.1016/j.drudis.2006.02.006
77. World Wide Approved Drugs (WWAD). [http://www.way2drug.com/dr/ww\\_drug\\_approved.php](http://www.way2drug.com/dr/ww_drug_approved.php), accessed: 2020-09-21
78. Zakharov, A., Filimonov, D., Lagunin, A., Poroikov, V.: GUSAR (General Unrestricted Structure-Activity Relationships) program package, Certificate of Russian State Patent Agency, No. 2006613591 of 16.10.2006
79. Zakharov, A., Peach, M., Sitzmann, M., Nicklaus, M.: A new approach to radial basis function approximation and its application to QSAR. *Journal of Chemical Information and Modeling* 54(3), 713–719 (2014), DOI: 10.1021/ci400704f
80. ZINC library. <https://zinc.docking.org>, accessed: 2020-09-21