Toward crowdsourced drug discovery: start-up of the volunteer computing project SiDock@home

Natalia Nikitina[1][0000-0002-0538-2939]✉, Maxim Manzyuk[2][000-0002-6628-0119],
Marko Jukić[3,4][0000-0001-6083-5024], Črtomir Podlipnik[5][0000-0002-8429-0273],
Ilya Kurochkin[6][0000-0002-0399-6208], and Alexander Albertian[6][0000-0002-6586-8930]

[1] Institute of Applied Mathematical Research, Karelian Research Center
of the Russian Academy of Sciences, Petrozavodsk, Russia,
`nikitina@krc.karelia.ru`

[2] Internet portal BOINC.ru, Moscow, Russia,
`hoarfrost@rambler.ru`

[3] Chemistry and Chemical Engineering, University of Maribor, Maribor, Slovenia

[4] Faculty of Mathematics, Natural Sciences and Information Technologies,
University of Primorska, Koper, Slovenia
`jukic.marko@gmail.com`

[5] Faculty of Chemistry and Chemical Technology,
University of Ljubljana, Ljubljana, Slovenia,
`crtomir.podlipnik@gmail.com`

[6] Federal Research Center "Computer Science and Control"
of the Russian Academy of Sciences, Moscow, Russia,
`qurochkin@gmail.com, assa@4ip.ru, assa@isa.ru`

Abstract. In this paper, we describe the experience of setting up a computational infrastructure based on BOINC middleware and running a volunteer computing project on its basis. We characterize the first series of computational experiments and review the project's development in its first six months. The gathered experience shows that BOINC-based Desktop Grids allow to to efficiently aid drug discovery at its early stages.

Keywords: Desktop Grid · Distributed computing · Volunteer computing · BOINC · Virtual drug screening · Molecular docking · SARS-CoV-2

1 Introduction

Among the variety of high-performance computing (HPC) systems, Desktop Grids hold a special place due to their enormous potential and, at the same time, high availability. Desktop Grids combine non-dedicated geographically distributed computing resources (typically, desktop computers) connected to the central server by the Internet or a local access network. The nodes perform computations for the Desktop Grid in their idle time. The resources are usually provided either by the volunteer community or by individuals and organizations related to the performed research. Such a computational infrastructure allows to efficiently solve computationally intensive scientific problems in the areas of mathematics [1], physics [2], astronomy [3] and others.

Submitted to the International Conference Russian Supercomputing Days 2021, RSD-2021, https://russianscdays.org/, on May 15, 2021

Upon the onset of COVID-19 pandemic in 2020, bio-medicine computational problems received a particular attention of scientists all over the world. At the same time, the public interest to such problems considerably raised, allowing the scientists to gather unprecedented amounts of computational resources.

In particular, a volunteer computing project Folding@home gathered the resources of 2.4 exaflops in early 2020, becoming the first world's exascale system, more powerful than the Top500 supercomputers altogether [4]. The overall potential of Desktop Grids is estimated as hundreds of exaflops [5], much more than the total power of all existing supercomputers.

Folding@home is a prominent example of a world-wide volunteer computing project, one of the many run by the world's leading research institutions. Computational capacities provided by the volunteer community allow such projects to perform large-scale computational experiments and process large amounts of data for solving urgent problems of social importance and wide public interest. This is the case of the fight against novel coronavirus disease since the beginning of 2020 [6–10].

In contrast, an organization may employ its own idle computational resources within an enterprise-level Desktop Grid. Such approach is particularly useful at the very early stages of research when computational experiments are irregular or subject to significant changes. Enterprise Desktop Grids support the research of a localized importance such as studying rare or neglected diseases [11].

To organise and manage Desktop Grid-based distributed computations, a number of software platforms are used. The most popular platform among them is BOINC (Berkeley Open Infrastructure for Desktop Computing) [5]. Among the 157 active largest projects on volunteer computing, 89 are based on BOINC [12]; that is, BOINC can be considered a *de-facto* standard for the operation of volunteer computing projects. The BOINC platform is an actively developing Open Source software and provides rich functionality for running projects both at the global and at the enterprise level.

This paper addresses the start-up of a BOINC-based volunteer computing project SiDock@home aimed at drug discovery. The main computationally intensive problem to solve in the process of drug discovery is the virtual screening, an *in silico* alternative to high-throughput screening. We describe the problem of the virtual screening and a series of computational experiments held within SiDock@home at its first mission: the fight against SARS-CoV-2.

2 The project SiDock@home

2.1 Virtual drug screening

HPC tools assist drug discovery at its first stages [13,14] which is natural considering the complex nature of biochemical processes, enormous sizes of the libraries of existing and synthesizable chemical compounds and fragments.

In this work, we consider structure-based virtual screening, a computational technique based on molecular docking of a library of small compounds against a specified therapeutic target. Molecular docking itself is a complex and computationally demanding procedure [15] performed using a variety of software tools (see, e.g., [16] for a detailed review). As this is a computer-aided simulation of a biochemical process, there are yet a number of aspects to improve, and new approaches are being developed and implemented.

In the presented project, we employ a developing molecular docking software CmDock [17] which started in 2020 as a fork of an open-source software RxDock [18], aimed at optimisation, implementation of new features and utilisation of modern hardware, namely GPU computational resources.

Following the course of research of COVID.SI, we considered a set of 59 targets to screen first of all (see Table 1). We use 3D structural models of targets generated by the D-I-TASSER/C-I-TASSER pipeline [19] as well as PDB databse and a uniquely designed chemical library of a billion of compounds.

Table 1. Targets for the first set of computational experiments in SiDock@home.

Target ID	The protein	Organism	Source of structure	PDB Code
1-21	3CL Pro	SARS-2	Snapshots from MD trajectory	
26-34	Spike Protein	SARS/ MERS/ SARS-2	Crystalographic structures	2AJF,2DD8, 3SCL, 5X58,6ACK,6LZG, 6M0J,6M17,6VW1
35-37	DHODH	Human	Crystalographic structures	4IGH,4JTU,4OQV
41-48	PL Pro	SARS/ MERS/ SARS-2	Crystalographic structures	2FE8,3MP2,4OW0, 6W9C,6WRH,6WUU, 6WX4,6WZU
49-50	FURIN	Human	Crystalographic structures	5JXH, 5MIM
51-54	Methyl Transferase	SARS-2	Crystalographic structures	6W4H,6W61, 7C2I,7C2J
55-56	E Protein	SARS/ SARS-2	NMR/ Homology model	5X29 (SARS) 5X29 Homology (SARS-2)
58-59	PL Pro	SARS-2	Homology models	based on 3E9S, 5E6J, 6W9C

In the next subsections, we provide more details on the work being performed in the project since its beginning.

2.2 Start of the project

SiDock@home [20] stems from a citizen science project "*Citizen science and the fight against the coronavirus*" (COVID.SI) [21] which originally performed virtual screening of a library of ten million of small molecules against multiple potential therapeutic targets, and based on an original middleware platform. For the purposes of the project popularization and scaling, SiDock@home was started as a BOINC-based extension of COVID.SI.

In [22], we overview the drug discovery problem being solved and the project's place among other BOINC-based projects fighting against SARS-CoV-2. In [23], we provide the performance dynamics of the project's Desktop Grid during the first six months of its work.

Fig. 1 presents the capacity dynamics and illustrates the growth of available computational resources. The maximal theoretical capacity of all registered computers is depicted by the dashed line filled with light gray. Due to the principles of the Desktop Grid operation, a maximum of a computer's resources is unlikely to be available to a BOINC project. The solid line filled with dark gray represents the capacity actually available to BOINC tasks, considering the limitations imposed by the clients.

To summarize, the project's Desktop Grid has reached the scale of a modern supercomputer in the first six months and keeps growing as the community's interest to the project rises.

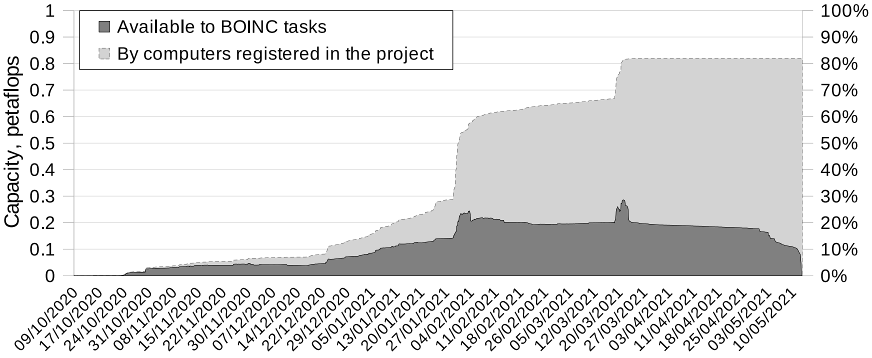

Fig. 1. Capacity dynamics of the project's Desktop Grid in 6 months.

2.3 Computational experiments

The process of structure-based virtual screening includes molecular docking of a library (or a library subset) of chemical compounds against each investigated target. The corresponding computational problem belongs to the class of bag-of-tasks problems efficiently solved using Desktop Grids. In this subsection, we describe the setup of the initial series of computational experiments performed in SiDock@home.

The first mission of SiDock@home is aimed at several targets playing important roles in the life cycle of SARS-CoV-2. Virtual screening for each target defines an independent computational experiment, which is a batch of BOINC tasks to complete. Each BOINC task consists of molecular docking of an independent compound (or a subset of compounds) against the target.

For the molecular docking, we have employed native applications RxDock [18] and CmDock [17] using the BOINC wrapper program [24]. The resulting applications `rxdock-boinc`, `cmdock-boinc` and `cmdock-boinc-zip` have been implemented, so far, for Linux 64-bit, Windows 64-bit and Mac OS 64-bit but ARM binaries of CmDock are available for the future.

In each application, input files for a BOINC task include a package of small compound models (ligands), a target model and a description of the screening protocol. In the third application, `cmdock-boinc-zip`, packages of ligands are transferred in a compressed form to save disc space on the server.

The sizes of the input packages of ligands were selected so as to comply with conventional principles of BOINC projects where a task takes on average 1-2 hours to complete on an average desktop computer. However, the runtime may depend also on other factors such as the size of the supposed binding site.

Fig. 2 and Fig. 3 illustrate the difference between sizes of binding sites of targets 3CL^{pro} (1st computational experiment) and Eprot (5th computational experiment). Task runtimes are directly correlated to cavity volumes and the numbers of grid points.

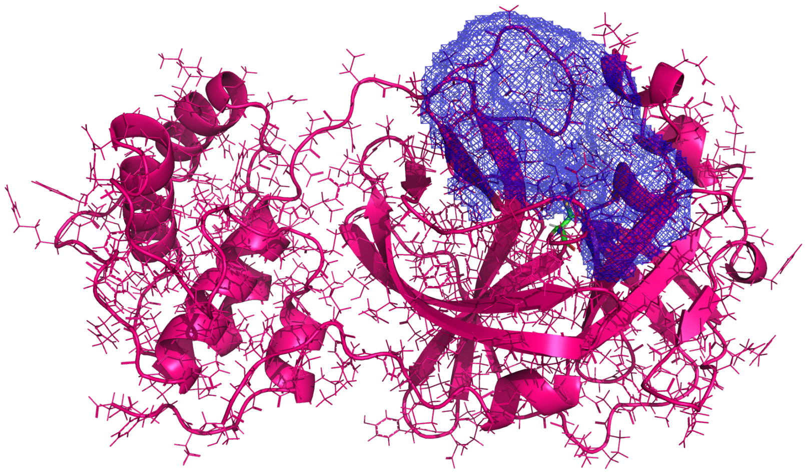

Fig. 2. The protease 3CL^{pro}, the target of the first computational experiment, with a ligand docked at the binding site of volume $3106.25A^3$ (24 850 points).

2.4 Server

The detailed description of BOINC architecture, functions and mechanisms is provided in [5]. In brief, the computational process may be described as follows. BOINC middleware has a server-client architecture. The server generates a large number of tasks that are mutually independent parts of a computationally-intensive problem such as virtual screening. The clients are the computers of any supported architecture voluntarily provided by the community.

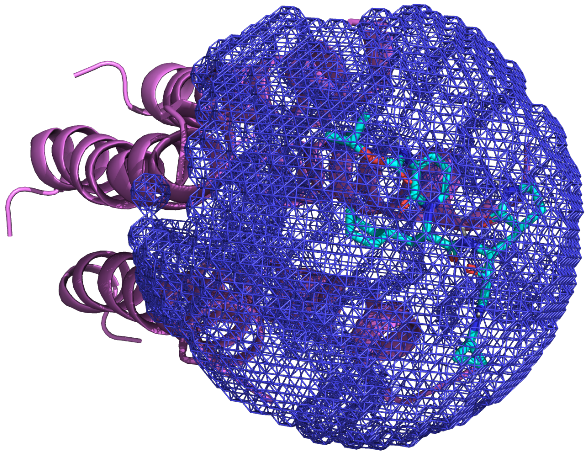

Fig. 3. The envelope protein (E) pentamer, the target of the fifth computational experiment, with a ligand docked at the binding site of volume $5068.25 A^3$ (40 546 points).

When a client computer is idle, it requests work from the server, receives tasks, and processes them independently. In SiDock@home, a single task performs molecular docking of 2 000 ligands against a specified target and pre-filters the obtained results according to the specified HTVS protocol. Another replica of the same task is sent to another computer of another user to facilitate the check of results. Such a replication mechanism provided by BOINC allows one to balance between speed and accuracy of computations. Although task replication is known to be most efficient in the end phase of a computational experiment [25], it also allows to reduce the time of an initial phase when the application is subject to implementation errors.

Upon finishing, the client reports results back to the server. The results are checked for correctness and validity, and stored for further usage such as post-filtering. In SiDock@home, we consider a result *correct* if it reports a positive number of ligands successfully docked. The results of two replicas of the same task are considered *valid* if they report the same set of ligands successfully docked. Such checks allow to automate the processing of the most common types of errors and to speed up the computational process.

Mechanisms implemented in BOINC middleware allow efficient scaling of the project as new clients join a project. At the same time, increase of the number of clients causes increase of the Desktop Grid throughput, and, consequently, necessitates scaling of the server resources to process all the workflow.

In Table 2, we summarize the information on the servers used in SiDock@home. Initially, the project's server part was deployed in an Ubuntu 18.04 LTS-based machine referred to as **Cloud**. In six months of operation, two hardware servers were installed: **Humpback** (Russia) and **Gray** (Slovenia).

Table 2. Servers of the project SiDock@home as of May 2021.

Server	Functions	Characteristics
Cloud	BOINC server (6 months); development; storage and processing of I/O files (rxdock-boinc, cmdock-boinc)	Cloud virtual machine at 4 virtual cores of Xeon 6140, 8 Gb RAM, 32 Gb SSD, 512 Gb HDD
Humpback	BOINC server; storage and processing of I/O files (cmdock-boinc-zip)	HP DL 380 Gen8; 2x Intel(R) Xeon(R) CPU E5-2620 v2 @ 2.10GHz (12 cores, 24 threads), 32 Gb RAM, 2x HDD 4Tb SAS, RAID 1.
Gray	Auxilary server; storage of the library; archivation; computations	2x Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz (14 cores, 28 threads), 64 Gb RAM, 2x 10 Tb HDD, RAID 1.

2.5 Clients

Unlike computational clusters and supercomputers, Desktop Grids are devoid of high-speed interconnection between computational nodes, homogeneity, reliability, and a scheduled availability of the nodes. These disadvantages restrict the class of the computational problems solved on Desktop Grids and impose difficulties when organizing the computational process.

However, the practice of many BOINC projects has proven the high efficiency of the Desktop Grids built with help of the community. Essential features of the Desktop Grid technology are their affordability, adaptability and a high potential in the quick attraction of a large number of voluntarily provided computing resources.

BOINC maintains the record of the performed work in the form of credit [26] which, in the most common case, is calculated as follows. Each host $h \in H$ is assigned a value a_h, a peak performance of its CPU flops, estimated with an internal BOINC benchmark. When a task τ has been executed on host h, BOINC registers the elapsed time $T_{h\tau}$. The amount of credit the host would get for the task is $C_{h\tau} = T_{h\tau} \cdot a_h \cdot CS$. If the result passes a validity check on the server and the quorum has been met, the host is awarded $C_{h\tau}$ (or an appropriately adjusted value if quorum exceeds 1).

Here, $CS = \frac{200}{86\,400} \times 10^9$ *(a Cobblestone)* is a constant unifying the effective work of heterogeneous computers with a reference one that would do one gigaflop/s based on the Whetstone benchmark and receive 200 credits a day.

BOINC credit system and leader boards serve for unifying the contributions made by the geographically distributed, highly heterogeneous computing nodes.

As of the middle of May 2021, the number of active participants is about 2 000 with about 8 000 computers. These numbers are subject to dynamic changes due to a BOINC community competition held in the beginning of May. After its ending, the number of active participants and computers is expected to decrease.

At the longer time range, however, these numbers are expected to grow as the project develops.

Since the beginning of the project, the total gained credit has approached 753 463 286 Cobblestones calculated by users, 752 484 194 by computers. The difference is due to the fact that some participants occasionally delete their computers from the project; however, the data accumulated by computers allows to evaluate the credit dynamics with a high accuracy.

Let us consider the computers participating in SiDock@home in more detail. Table 3 provides the statistics on CPUs with non-zero credit. As of the middle of May 2021, CPUs of 1 293 models participate in the project. One observes a parity between Intel® and AMD vendors; the total credit is of the same order. The number of AMD-based computers is almost two times less than Intel®-based ones, but the number of cores and the recent average credit are higher.

Overall, the presented data testify that, despite of the fact that AMD takes about 20% of the world market, BOINC volunteer community (which highly appreciates the CPU performance) rates AMD high and tends to purchase the latest CPU models.

Table 3. CPUs participating in SiDock@home.

Vendor	**Number of computers**	**Number of cores**	**Total credit**	**Recent average credit**
<not detected>	69	552	1 969 693	24 214
ARM	49	268	335 590	17 117
AMD	3 627	118 303	362 273 292	6 508 317
CentaurHauls	1	2	664	0
Intel®	9069	111 263	388 131 880	6 267 857

Apart from the total computational performance, a project's scale may be characterized by the number of active participants and the relative distribution of their contributions. In Fig. 4 and 5, we graphically depict the leader board of participants by the total credit of the projects SiDock@home and RakeSearch [**?**] of a different size. The diagrams illustrate the heterogeneity of the contributions.

But what is more important, one observes the difference in the roles that individual contributions play. The top users are definitely making significant contributions, but their aggregated fraction becomes less with the community's growth, even despite the fact that the large "group users" (accounts that combine multiple computers, for example, belonging to the same organization) join the project and actively participate. The project of a larger scale (SiDock@home) bases mostly on a large number of relatively small contributions, while in the project of a smaller scale (RakeSearch), large contributions dominate.

Finally, in Table 4, we provide the distributions of individual contributions in comparison with a project's top participant. The data show that individual contributions have more variance in a larger-scale project.

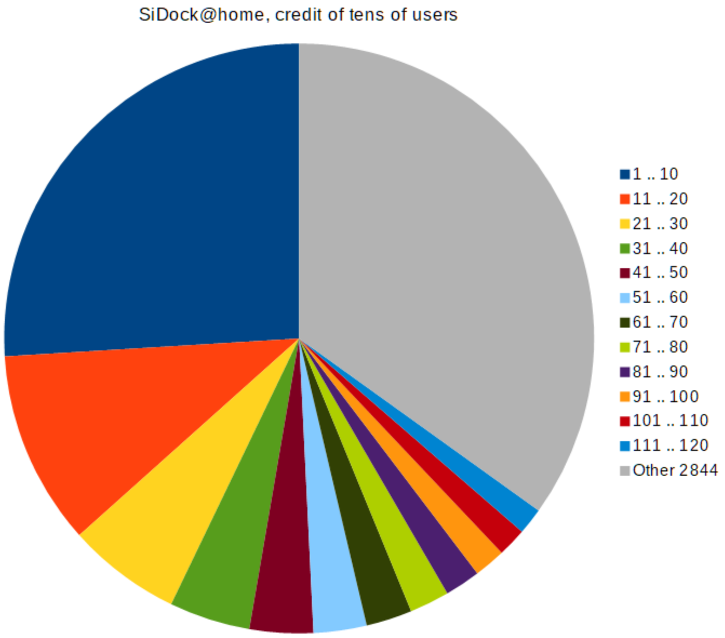

Fig. 4. User credit statistics in the project SiDock@home (2 964 participants).

3 Conclusion

Drug discovery is a time-demanding and resource-demanding process that typically lasts for 10-12 years. At its first stage, virtual screening may assist in selection of prospective compounds *in silico* and reduce the consumption of time and money. In this paper, we describe the start-up of the drug discovery project SiDock@home and a series of computational experiments on performing virtual screening in the fight against SARS-CoV-2. Furthermore the project is evolving towards other areas of medicinal chemistry where active compounds for the study of Ebola, Malaria and targets relevant to oncology will be examined.

Citizen science initiatives and, in particular, BOINC projects on bio-medicine have always attracted many volunteers due to their socially important subjects. With the onset of a pandemic of COVID-19, the community's interest raised up to an unprecedented level.

We believe that BOINC community is able to provide the scientists with even more computational resources so as to support the research at all scales.

Funding

This work was partly supported by the Scholarship of the President of the Russian Federation for young scientists and graduate students (project "Game-

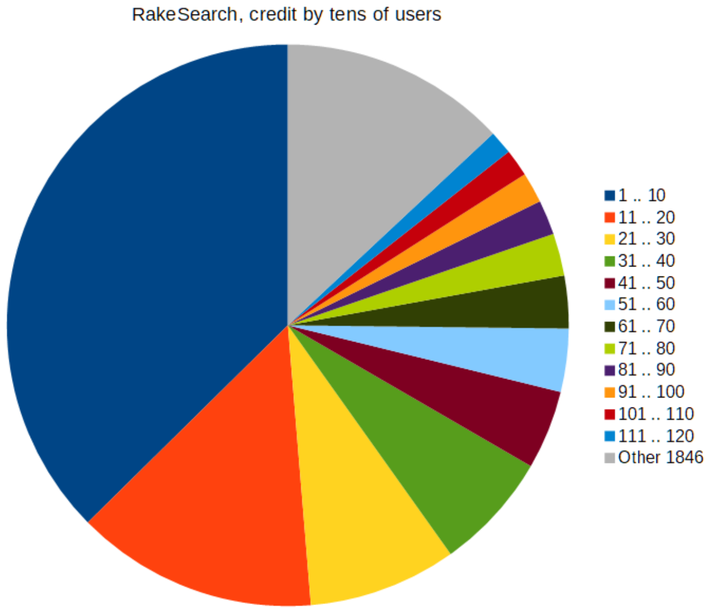

Fig. 5. User credit statistics in the project RakeSearch (1 966 participants).

theoretical mathematical models and algorithms for scheduling in high-performance heterogeneous computational systems”), the Slovenian Ministry of Science and Education infrastructure, project grant HPC-RIVR, and by the Slovenian Research Agency (ARRS), programme P2-0046.

Acknowledgements

The first initial library (one billion of compounds) was prepared with the generous help of Microsoft that donated computational resources in the Azure cloud platform [27]. We all from COVID.SI are grateful and looking forward to future collaborations.

We would like to thank all volunteers who provide their computers to the project. Discussions and advice on the project forum are greatly appreciated.

References

1. Gerasim@home main page. https://gerasim.boinc.ru/, 2021. [Online; accessed 12-May-2021].
2. home — LHC@home. https://lhcathome.web.cern.ch, 2020. [Online; accessed 12-May-2021].
3. Einstein@Home. https://einsteinathome.org, 2021. [Online; accessed 12-May-2021].

Table 4. Distribution of the total credit among participants of two BOINC projects.

RakeSearch, small-scale project		
Credit, % of the leader	**# of participants with larger credit**	**% of participants with larger credit**
0,00%	2964	98,37%
10,00%	23	0,76%
20,00%	11	0,37%
30,00%	5	0,17%
40,00%	2	0,07%
50,00%	2	0,07%
60,00%	2	0,07%
70,00%	1	0,03%
80,00%	1	0,03%
90,00%	1	0,03%
100,00%	1	0,03%
SiDock@home, medium-scale project		
Credit, % of the leader	**# of participants with larger credit**	**% of participants with larger credit**
0,00%	3905	97,82%
10,00%	38	0,95%
20,00%	18	0,45%
30,00%	9	0,23%
40,00%	7	0,18%
50,00%	5	0,13%
60,00%	3	0,08%
70,00%	3	0,08%
80,00%	2	0,05%
90,00%	1	0,03%
100,00%	1	0,03%

4. Folding@home – fighting disease with a world wide distributed super computer. https://foldingathome.org, 2020. [Online; accessed 12-May-2021].
5. David P. Anderson. BOINC: a platform for volunteer computing. *Journal of Grid Computing*, 18:99–122, 2020.
6. Together We Are Powerful – Folding@home. https://foldingathome.org, 2020. [Online; accessed 12-May-2021].
7. Rosetta@home. https://boinc.bakerlab.org, 2020. [Online; accessed 12-May-2021].
8. World Community Grid – home. https://www.worldcommunitygrid.org/, 2020. [Online; accessed 12-May-2021].
9. TN-Grid. http://gene.disi.unitn.it/test/, 2021. [Online; accessed 12-May-2021].
10. IberCIVIS. https://boinc.ibercivis.es/, 2021. [Online; accessed 12-May-2021].
11. Evgeny Ivashko, Natalia Nikitina, and Steffen Möller. High-performance virtual screening in a BOINC-based Enterprise Desktop Grid. *Vestnik Yuzhno-Ural'skogo Gosudarstvennogo Universiteta. Seriya "Vychislitelnaya Matematika i Informatika"*, 4(1):57–63, 2015. [in Russian].

12. Distributed Computing – Computing Platforms. http://distributedcomputing.info/platforms.html, 2020. [Online; accessed 12-May-2021].
13. Savíns Puertas-Martín, Antonio J. Banegas-Luna, María Paredes-Ramos, Juana L. Redondo, Pilar M. Ortigosa, Ol'ha O. Brovarets', and Horacio Pérez-Sánchez. Is high performance computing a requirement for novel drug discovery and how will this impact academic efforts? *Expert Opinion on Drug Discovery*, 15(9):981–985, 2020. PMID: 32345062.
14. Sailu Sarvagalla, Sree Karani Kondapuram, R. Vasundhara Devi, and Mohane Selvaraj Coumar. Chapter 9 - resources for docking-based virtual screening. In Mohane S. Coumar, editor, *Molecular Docking for Computer-Aided Drug Design*, pages 179–203. Academic Press, 2021.
15. Brian K Shoichet, Susan L McGovern, Binqing Wei, and John J Irwin. Lead discovery using molecular docking. *Current opinion in chemical biology*, 6(4):439–446, 2002.
16. Nataraj S Pagadala, Khajamohiddin Syed, and Jack Tuszynski. Software for molecular docking: a review. *Biophysical reviews*, 9(2):91–102, 2017.
17. CmDock. https://gitlab.com/Jukic/cmdock, 2020. [Online; accessed 12-May-2021].
18. Sergio Ruiz-Carmona, Daniel Alvarez-Garcia, Nicolas Foloppe, A Beatriz Garmendia-Doval, Szilveszter Juhos, Peter Schmidtke, Xavier Barril, Roderick E Hubbard, and S David Morley. rDock: A fast, versatile and open source program for docking ligands to proteins and nucleic acids. *PLoS computational biology*, 10(4):e1003571, 2014.
19. Modeling of the SARS-COV-2 Genome using I-TASSER. https://zhanglab.ccmb.med.umich.edu/COVID-19, 2021. [Online; accessed 12-May-2021].
20. SiDock@home. https://sidock.si/sidock, 2020. [Online; accessed 12-May-2021].
21. Home – COVID.SI. https://covid.si/en, 2020. [Online; accessed 12-May-2021].
22. N. Nikitina, M. Manzyuk, Č. Podlipnik, and M. Jukić. Volunteer computing project SiDock@home for virtual drug screening against SARS-CoV-2, 2021. submitted.
23. N. Nikitina, M. Manzyuk, Č. Podlipnik, and M. Jukić. Performance estimation of a BOINC-based Desktop Grid for large-scale molecular docking, 2021. submitted.
24. WrapperApp – BOINC. https://boinc.berkeley.edu/trac/wiki/WrapperApp, 2020. [Online; accessed 12-May-2021].
25. Gaurav D Ghare and Scott T Leutenegger. Improving speedup and response times by replicating parallel programs on a SNOW. In *Workshop on Job Scheduling Strategies for Parallel Processing*, pages 264–287. Springer, 2004.
26. CreditNew – BOINC. https://boinc.berkeley.edu/trac/wiki/CreditNew, 2020. [Online; accessed 12-May-2021].
27. Cloud Computing Services — Microsoft Azure. https://azure.microsoft.com/en-us/, 2020. [Online; accessed 12-May-2021].