DOI: 10.14529/jsfi200306

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

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

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]

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

[2] Laboratory of Cancer Immunometabolism, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Frederick, Maryland, USA

[3] 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, TMPRSS2, 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 workflow 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 reflect 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.

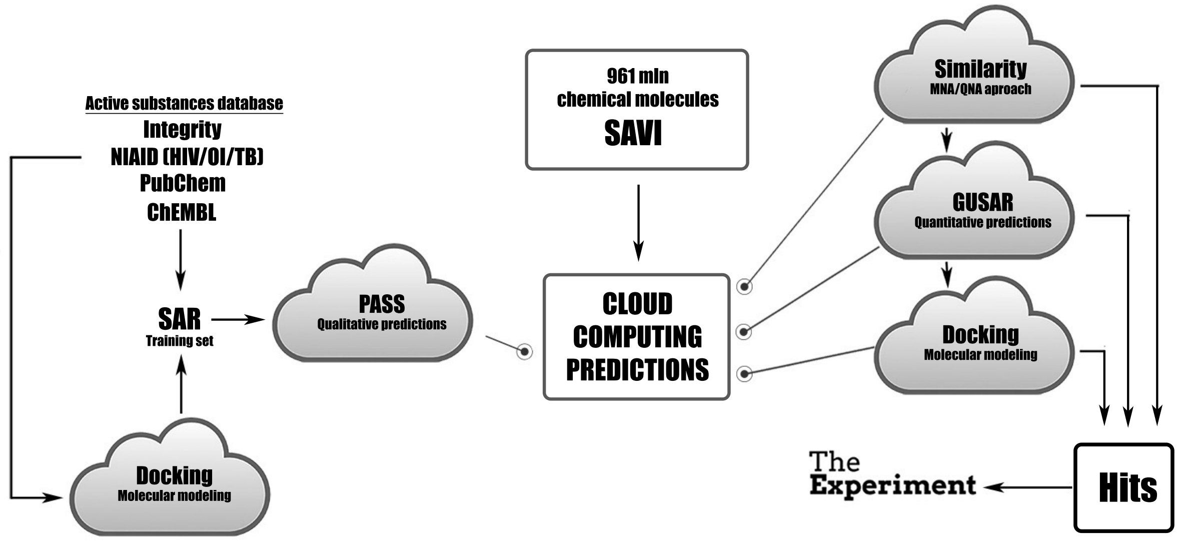


Figure 1. General workflow of the large database analysis for identification of potential pharmacological substances

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:

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

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:

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

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:

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

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

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

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:

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

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 bioactivites 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: Pa, the probability of belonging to the subset of "actives"; and Pi, the probability of belonging to the subset of "inactives". By default, all compounds, for which PASS predicts $Pa > Pi$, 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 $< P, Q >$ [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 sp3-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.

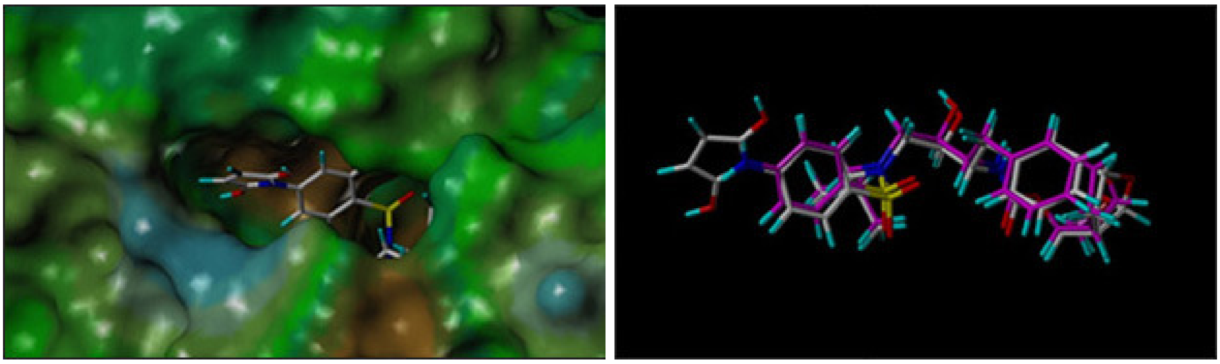
6. Identification of New Potential Anti-HIV Agents by Analysis of SAVI

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).



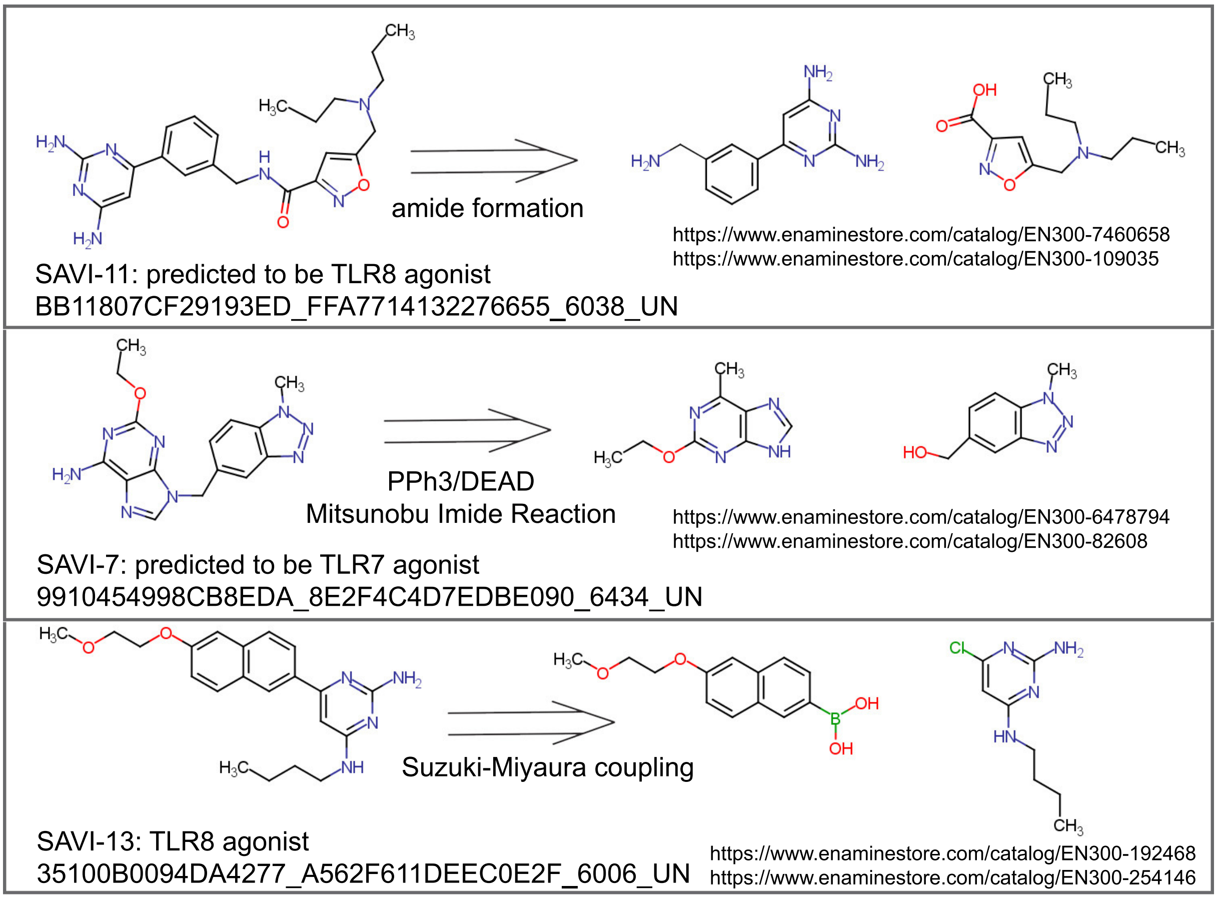
(a) Location of a SAVI molecule in the HIV-1 protease active center (colored according to hydrophobic potential)

(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).



SAVI-11: predicted to be TLR8 agonist
BB11807CF29193ED_FFA7714132276655_6038_UN

https://www.enaminestore.com/catalog/EN300-7460658
https://www.enaminestore.com/catalog/EN300-109035

SAVI-7: predicted to be TLR7 agonist
9910454998CB8EDA_8E2F4C4D7EDBE090_6434_UN

https://www.enaminestore.com/catalog/EN300-6478794
https://www.enaminestore.com/catalog/EN300-82608

SAVI-13: TLR8 agonist
35100B0094DA4277_A562F611DEEC0E2F_6006_UN

https://www.enaminestore.com/catalog/EN300-192468
https://www.enaminestore.com/catalog/EN300-254146

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.

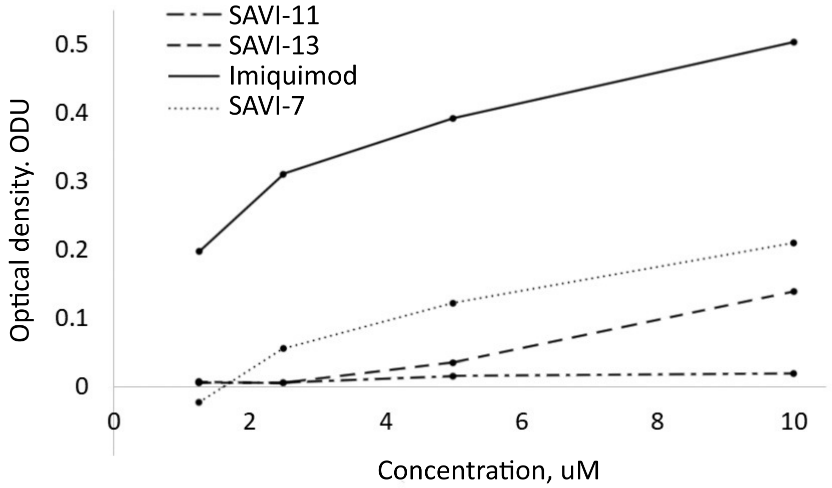


Figure 4. Activation of NF-kB signaling by the tested compounds (structural formulae are given in Fig. 3) in RAW-Blue reporter cells (reference drug – Imiquimod)

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 Challenge 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, Zn2+ 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 TMPRSS2 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 TMPRSS2 inhibitors (Fig. 5).

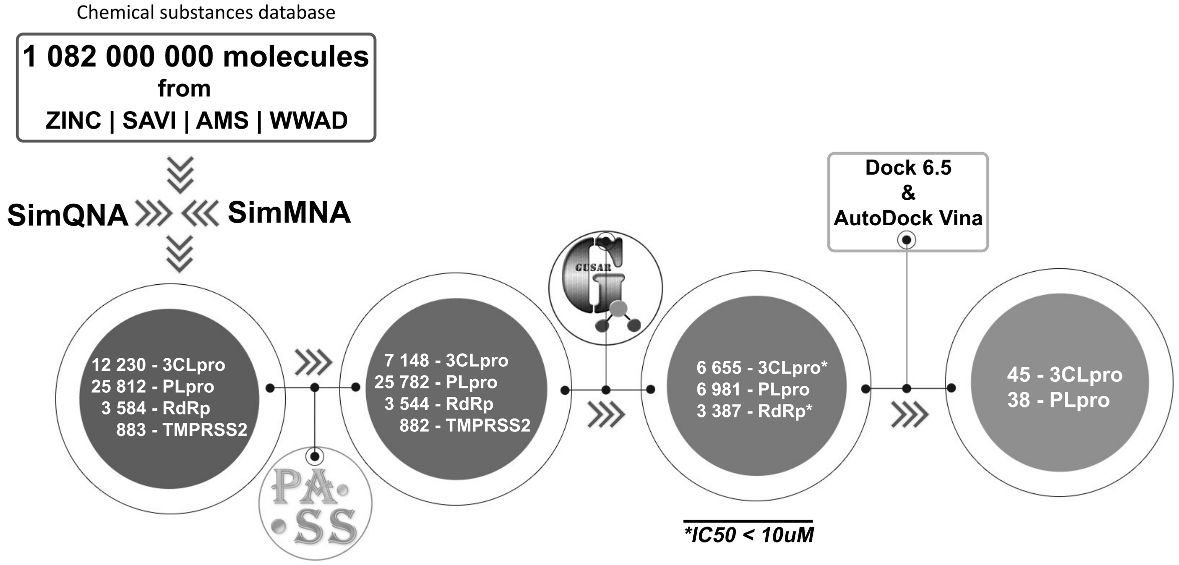


Figure 5. General workflow and results of selection of anti-SARS-CoV-2 hits

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 TMPRSS2 inhibitors.

For TMPRSS2, the spatial structure is not available. Also, for the TMPRSS2 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 TMPRSS2.

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_a 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.

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 SASR-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., Gloriozova, 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 Ki and IC50 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. Murtazalieva, 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`, 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/`, 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, 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