

# The design of compounds with desirable properties - The anti-HIV case study

---

**Novak, Jurica; Pathak, Prateek; Grishina, Maria A.; Potemkin, Vladimir A.**

*Source / Izvornik:* **Journal of Computational Chemistry, 2022, 44, 1016 - 1030**

**Journal article, Accepted version**

**Rad u časopisu, Završna verzija rukopisa prihvaćena za objavljivanje (postprint)**

<https://doi.org/10.1002/jcc.27061>

*Permanent link / Trajna poveznica:* <https://um.nsk.hr/um:nbn:hr:193:158194>

*Rights / Prava:* [In copyright](#) / [Zaštićeno autorskim pravom.](#)

*Download date / Datum preuzimanja:* **2025-02-19**

*Repository / Repozitorij:*



[Repository of the University of Rijeka, Faculty of  
Biotechnology and Drug Development - BIOTECHRI  
Repository](#)



# The Design of Compounds with Desirable Properties – the Anti-HIV Case Study

Jurica Novak,<sup>1,2,3</sup> Prateek Pathak,<sup>4</sup> Maria A. Grishina,<sup>5</sup> and Vladimir A. Potemkin<sup>6</sup>

Correspondence to: Jurica Novak (E-mail: [jurica.novak@biotech.uniri.hr](mailto:jurica.novak@biotech.uniri.hr)) and Maria A. Grishina (E-mail: [grishinama@susu.ru](mailto:grishinama@susu.ru))

<sup>1</sup> Department of Biotechnology, University of Rijeka, Rijeka, Croatia, 51000

<sup>2</sup> Center for Artificial Intelligence and Cybersecurity, University of Rijeka, Rijeka, Croatia, 51000

<sup>3</sup> Scientific and Educational Center "Biomedical Technologies", Higher Medical & Biological School, South Ural State University, Chelyabinsk, Russia, 454080

<sup>4</sup> Laboratory of Computational Modelling of Drugs, Higher Medical & Biological School, South Ural State University, Chelyabinsk, Russia, 454080

<sup>5</sup> Laboratory of Computational Modelling of Drugs, Higher Medical & Biological School, South Ural State University, Chelyabinsk, Russia, 454080

<sup>6</sup> Laboratory of Computational Modelling of Drugs, Higher Medical & Biological School, South Ural State University, Chelyabinsk, Russia, 454080

## ABSTRACT

Efficacy and safety are among the most desirable characteristics of an ideal drug. The tremendous increase in computing power and the entry of artificial intelligence into the field of computational drug design are accelerating the process of identifying, developing, and optimizing potential drugs. Here, we present novel approach to design new molecules with desired properties. We combined various neural networks and linear regression algorithms to build models for cytotoxicity and anti-HIV activity based on Continual Molecular Interior analysis (CoMIn) and Cinderella's Shoe (CiS) derived molecular descriptors. After validating the reliability of the models, a genetic algorithm was coupled with the Des-Pot Grid algorithm to generate new molecules from a predefined pool of molecular fragments and predict their bioactivity and cytotoxicity. This combination led to the proposal of 16 hit molecules with high anti-HIV activity and low cytotoxicity. The anti-SARS-CoV-2 activity of the hits was predicted.

## Introduction

The highly active antiretroviral therapy (HAART), the combination of reverse integrase inhibitors, HIV protease inhibitors and an integrase inhibitors which suppresses HIV replication, is recognized as the most effective therapy for AIDS.<sup>1</sup> According to the World Health Organization, 25.4 million people (67 % of people living with HIV) were receiving antiretroviral treatment by end 2019.<sup>2</sup> Unfortunately, there are dark sides of lifelong HAART leading to the systemic complications involving heart, bone, kidney and other organs.<sup>3–5</sup> From the perspective of the human immunodeficiency virus, the entity with the highest

reported mutation rate,<sup>6</sup> it successfully fights continuous drug selective pressure, leading to the emergence of drug resistant forms.<sup>7</sup> Due to the absence of a successful anti-HIV vaccine and drawbacks of presently approved anti-HIV drugs, a design of a new drug candidates, with increased potency and less side effects (less toxic, increased pharmacokinetic properties) remains hot topic in scientific and pharmaceutical community,<sup>8–24</sup> even 30 years after the first registered drug, azidothymidine.<sup>25</sup>

Over the past 20 years, myriad of a quantitative structure-activity relationship (QSAR) studies were performed with the aim to design novel anti-HIV compounds. For example, Zakariazadeh *et al.*<sup>12</sup>

designed four different QSAR models combining quantum and molecular mechanical descriptors for naphthyridine derivatives against HIV-1 integrase (HIV-IN) activity. Tong *et al.*<sup>26</sup> build 3D-QSAR models using random sampling analysis on molecular surface and translocation comparative molecular field vector analysis. They designed 18 new compounds, whose activity against HIV-1 protease (HIV-PR) were predicted to be higher than the activity of the template molecule, and explored the mechanism of action by molecular docking. Descriptors based on the radial distribution function weighted by the number of valence shell electrons were used by Potemkin's group to establish a model relating descriptors with the inhibition constant for a series of HIV-PR inhibitors.<sup>20,23</sup> Additionally, they exploited those models to study influence of point mutations of HIV-PR to the inhibition constant.<sup>21</sup> 3D-QSAR models for thiazolidinones of Ravichandran *et al.*<sup>27</sup> identified the steric and electrostatic properties, as well as the hydrogen bond acceptor, hydrogen bond donor, and hydrophobic properties correlate the most with inhibition potency toward HIV-1 reverse transcriptase (HIV-RT). One of the limitations of the most of QSAR studies is that they took into consideration the homogeneous series of compounds against only one protein. The approach by Speck-Planche *et al.*<sup>9</sup> try to circumvent this restriction by developing multi target QSAR models constructed from a heterogeneous database of compounds targeting seven essential proteins identified as crucial for the HIV infection. Combining the ligand-based approach and their QSAR models, they proposed six molecules as potential anti-HIV agents. The same authors proposed multitasking model for *in silico* design of compounds with high anti-HIV activity and desirable ADMET properties.<sup>28</sup>

The enormous increase in computer power in the last decades coupled with the artificial intelligence algorithms revolutionize the computationally aided drug design.<sup>29,30</sup> Wei *et al.*<sup>31</sup> calculated 2120 geometrical, topological and electronic properties descriptors for the set of 381 HIV-PT inhibitors and 9866 decoys. Then, using genetic algorithm (GA) for feature selection, they develop a support vector machine (SVM) classification model for HIV-PT

inhibitors, with 90 % prediction accuracy for independent validation set. After screening National Cancer Institute database using the proposed classifier and testing in an *in vitro* HIV-1 protease inhibitory assay 6 hit molecules, two molecules are proposed as potential inhibitors. Darnag *et al.* compared the performance of multiple linear regression (MLR), artificial neural network (ANN) and SVM in QSAR study of 38 cyclic-urea derivatives, which have been confirmed as HIV-PR inhibitors.<sup>32</sup> Models obtained by SVM showed better quality and higher generalization capabilities compared to linear regression methods. Xuan *et al.* used 551 HIV-IN inhibitors with different scaffolds to build models for predicting anti-HIV-IN activity.<sup>33</sup> After selection of 20 molecular descriptors, they used a Kohonen's self-organizing map and SVM to obtain the model with correlation coefficient of 0.93 for test set. Zorn *et al.* conducted a machine learning study based on data from wild type HIV-1 cell based and HIV-RT DNA polymerase inhibition assay.<sup>34</sup> Comparing predictive abilities of seven models trained using different machine learning algorithms, they demonstrated comparable performance of support vector machine and deep neural network approaches. Recently, Stolbov *et al.* published web resource for prediction of anti-HIV activity based on the structural formula using GUSAR and PASS models.<sup>22</sup>

More than 6.4 million deaths are connected to new infection COVID-19, as of 22 August according to data available from WHO. COVID-19 is a disease caused by SARS-CoV-2 virus. Due to its high virulence and problems in global vaccine campaign, efficient anti-COVID-19 drugs are needed. One of the attractive targets is SARS-CoV-2 3-chymotrypsin like protease (3CLpro), identified as crucial enzyme mediating viral replication and transcription.<sup>35,36</sup> Various studies tried to repurpose existing drugs, with special attention given to antiviral drugs and HIV protease inhibitors.<sup>37-44</sup> Motivated by those studies and encouraged by successful designed, validation and application of QSAR model predicting anti-3CLpro activity,<sup>42</sup> for all proposed HIV-1 protease inhibitors their possibility to stop the activity of the 3CLpro will be predicted.

The main goal of the present study is to combine available multitarget anti-HIV model with cytotoxicity

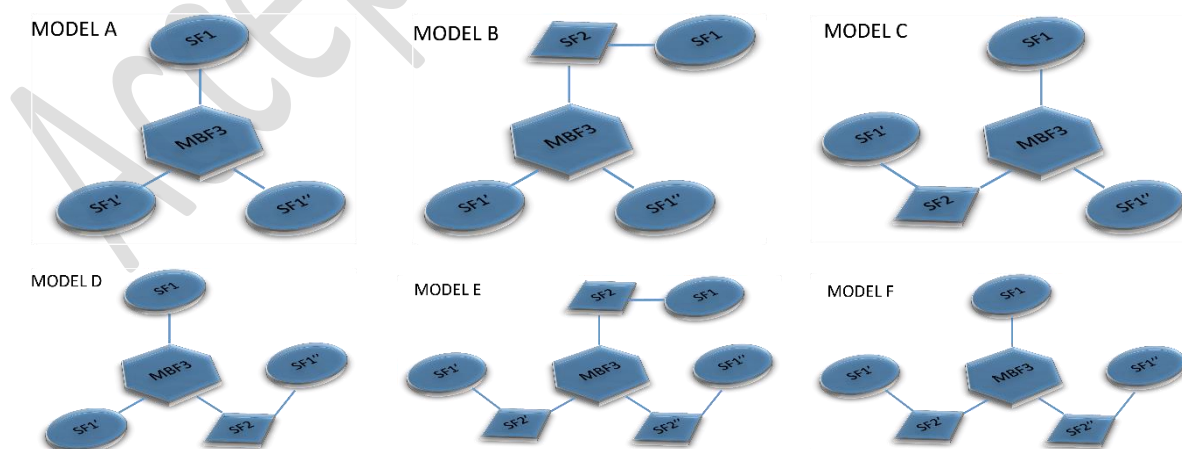
model from computational-aided drug design web platform Chemosophia,<sup>45</sup> and using modified version of recently introduced DesPot-Grid algorithm to design novel potent and non-toxic anti-HIV drugs. Secondary aim includes repositioning of novel anti-HIV compounds as potential SARS-CoV-2 3CLpro inhibitors to fight COVID-19 pandemic. The rest of the paper is organized as follows. In Methods section short overview of anti-HIV activity and cytotoxicity models, together with modification introduced to the DesPot-Grid algorithm are reviewed. In the Results and discussion part, potential hit molecules generated by DesPot-Grid are identified and analyzed. In Conclusion, brief summary of results is presented.

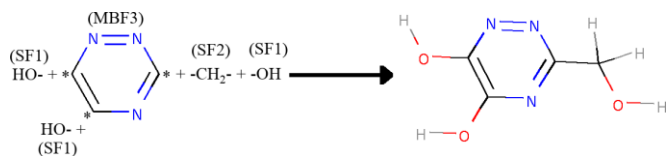
## Methods

### Molecular Fragments and Model Design

For successful generation of novel molecules with high anti-HIV activity and low cytotoxicity using DesPot-Grid approach it is crucial to carefully select molecular fragments. The main block fragments include nine moieties extracted from the molecules confirmed as active in AIDS Antiviral Screen Data<sup>46</sup> database, three triazoles, all five nucleobases (with guanine and uracil with two different substitution possibilities) and 1,3-oxazine-2,6-dione (thymine with substituted NH group by oxygen atom). Those blocks can have substitutions on three positions (MBF3 – main block fragment with 3 open valences; see Supporting Information). The geometries were optimized using approach based on MM3 molecular mechanics force field and AlteQ

quantum-chemical method,<sup>45,47,48</sup> and saved in sdf file format. To fully describe the main block, in separate file atoms' indices where the substitution can occur have to be defined, and sdf file has to be modified to create an open valence. This was done by removing hydrogen atom or methyl group bound to the atom where the substitution is going to be introduced. Fragments-substituents are obtained in similar fashion, combining fragments from confirmed active molecules of AIDS Antiviral Screen Data database and a variety of functional groups or (poly)cyclic moieties. Again, atoms with open valence have to be explicitly defined in separate text file. In this paper, two sets of fragments-substituents were used – with one (FS1) and with two (FS2) open valences. Altogether, there were 20 fragments in MBF3 set, 87 fragments in FS1 and 40 in FS2 set (Figure S11-3). Six ways to combine MBF3 with FS1 or/and FS2, were exploited in this study for design of new anti-HIV compounds (Figure 1). Thus, the number of combinations of structural fragments for the presented models varies from 13,170,060 (model A) to 842,883,840,000 (model E). It is obvious that the activity prognosis for such a large number of structures is inappropriate. The DesPot-Grid algorithm makes it possible to carry out a directed design of active structures from the presented fragments, excluding a lengthy enumeration of all possible combinations. How the design of new molecules within DesPot-Grid algorithm works is schematically presented on Scheme 1 for model B, with 1,2,4-triazine as the main block with three open valences (MBF3), methylene group as the fragment-substituent with two (SF2) and hydroxyl as the fragment-substituent with one (SF1) open valence.



**Figure 1.** Six models used by DesPot-Grid to generate new compounds.**Scheme 1.** An example of new molecule design by DesPot-Grid following model B.

### Dataset Collection

The  $IC_{50}$  values characterizing compounds' cytotoxicity, and  $EC_{50}$  characterizing its anti-HIV effect were collected from National Cancer Institute database.<sup>46</sup> Toxic effect was defined as the concentration necessary to inhibit the growth of uninfected cells ( $IC_{50}$ ). 100 most toxic and 100 least toxic compounds with measured  $IC_{50}$  values as molar concentration entered to training dataset for the cytotoxicity model designing. The  $IC_{50}$  values were converted into  $pIC_{50}$  for minimizing the variation during relationship establishment. The compounds which have  $pIC_{50}$  more than 8 were considered toxic, and if  $pIC_{50}$  was less than 3.3, compound was classified as nontoxic (Table SI1). 3D structures were retrieved from PubChem,<sup>49</sup> and optimized without changing geometries of stereo centers.<sup>45,50,51</sup>

In creating the training dataset for the models of anti-HIV activity, we used the compounds with both confirmed active class (CA) and confirmed inactive class (CI) from the United States National Cancer Institute database. The database contains more than 39000 compounds and corresponding anti-HIV activities, so that the concentrations required to observe a protective effect on the infected cells ( $EC_{50}$  values) are given. In addition, screening results were categorized as CA, CI and CM (confirmed moderately active).

### QSAR Models

One of the main prerequisites of DesPot-Grid are robust and reliable models, capable to identify compounds with high anti-HIV activity and low cytotoxicity. To meet this requirement, 18

models predicting anti-HIV activity and 12 models predicting cytotoxicity are exploited. The final anti-HIV activity score is a geometrical mean of all 18 models (PA), and analogously final cytotoxicity score is geometrical mean of 12 models (PT). This criterion is more stringent rule than arithmetic mean, but our previous studies demonstrate its feasibility.<sup>52</sup> The anti-HIV activity score is a number in range between 0 and 1, where 0 means the compound does not show anti-HIV activity, while 1 indicates the compound has profound anti-HIV activity. In case of the cytotoxicity, 0 indicates that the compound is very toxic, while 1 indicates non-toxicity. Models used to predict anti-HIV activity and cytotoxicity are publicly available on chemosophia.com,<sup>53</sup> an on-line platform for cheminformatics, bioinformatics and drug design. In total, 18 and 12 QSAR models were used to evaluate anti-HIV activity and cytotoxicity, respectively. QSAR models for SARS-CoV-2 3CLpro activity used in this study were recently developed and validated,<sup>42</sup> and are being implemented on chemosophia.com.

Here, only basic concepts behind those models are discussed, while for details reader is referred to original articles. 3D-QSAR Cinderella's Shoe (CiS) algorithm aims to classify compounds as being active or inactive based on molecular exterior.<sup>51,54</sup> The key point is to superimpose molecules of the training dataset to reach the best coincidence of molecular external field and to model pseudo – receptor complementary to the external field of bioactive molecules. The algorithm simulates pseudo-receptor based on Coulomb and van der Waals potentials on the molecular surface representing the molecular field. CiS algorithm calculates the entire spectrum of interaction characteristics, including interaction energies ( $E_j$ ), forces ( $F_j$ ) and force constants ( $k_j$ , elastic component)<sup>52,54</sup>:

$$E_j = \sum_{m=1}^N (E_{jm}^C + E_{jm}^{vdW}) + U_j \quad (1)$$

where  $E_{jm}^C$  and  $E_{jm}^{vdW}$  are Coulomb and Van der Waals energies of interaction of each  $m$ -th atom of the molecule with the  $j$ -th pseudo-atom of the receptor.  $U_j$  is elastic energy of interaction of the molecule with the  $j$ -th pseudo-atom

$$U_j = \frac{k_j \Delta r_j^2}{2} \quad (2)$$

$\Delta r_j$  is deviation of the  $j$ -th pseudo-atom of the receptor from the average position when interacting with the molecule of the dataset. In addition, when equating the force constants to zero,

$$F_{j,x} = \frac{\partial E_j}{\partial x} = 0 \quad (3)$$

the CiS algorithm can simulate an unlimitedly expandable receptor. This property is useful, since it imitates receptor pockets, which can be characterized by a large variation in size, like in HIV-1 protease.

The use of self-consistent field in CiS carries out the optimal arrangement of molecules in the complementary receptor until a constant energy value and values of the forces of intermolecular interactions equal to zero are achieved. In the general case, in the algorithm, the energy includes Coulomb, van der Waals interactions and the elastic energy of intermolecular interactions, which in its turn depends on the force constants that determine the flexibility and the extensibility of the pseudo-receptor.

First step of CoMIn (Continual Molecular Interior analysis) algorithm, a molecular interior based approach, includes overlaying molecules from the training dataset with the condition to maximize coincidence of the potentials or the quantum functions at the junctions of the generalized lattice (the mold of the superimposed dataset)<sup>52,54</sup>. The potentials can be Coulomb and van der Waals potentials, potentials of hydrogen bonds, distribution of

MERA atomic "matter" (eq. 4), its derivative (eq. 5) and their products with different weight factors ( $w_i$ )<sup>52,54</sup>:

$$\varphi_j = w_{ij} \alpha_j e^{-\beta_j r_{jm}^2} \quad (4)$$

$$\varphi'_j = -2w_{ij} \beta_j r_{jm} \alpha_j e^{-\beta_j r_{jm}^2} \quad (5)$$

where  $w_{ij}$  is  $i$ -th weight factor of atom  $j$  (atomic charge, volume, number of occupied atomic orbits, number of unoccupied atomic orbits, HOMO and LUMO energies as well as the products of these weight factors),  $r_{jm}$  is a distance of the atom  $j$  from the lattice junction  $m$ , and  $\alpha_j$  and  $\beta_j$  are explained in<sup>51</sup>. Those potentials are descriptors used to design QSAR models. In CiS and CoMIn algorithms, the approaches used to create relationships between bioactivity and descriptors include linear reaction of neural network (LNN), or neural network with sigmoid neurons (NNSN), or linear regression model (LRM). Then computed bioactivity (BA) is transformed to the probability of bioactivity expressed as a desirability function:

$$p = \exp[-\exp(a - b \times BA)] \quad (6)$$

In total, 18 QSAR models with cross-validation coefficient of determination (cross-R<sup>2</sup>) in range 0.88 - 0.95 for the prognosis of probability of anti-HIV bioactivity ( $p_{\text{anti-HIV}} = 0.5$  when  $pEC_{50} = 6.1$ ) were created. Table SI2 and Table SI3 summarize details of 18 anti-HIV activity and 12 cytotoxicity models, respectively, including the methods and potentials used for the feature calculations, the algorithms used to generate the QSAR models, and the results of the cross-R<sup>2</sup> validation. 12 QSAR models with cross-R<sup>2</sup> in range 0.91 - 0.99 for the prognosis of being non-toxic ( $p_{\text{tox}} = 0$  when  $pIC_{50} > 8$ , and  $p_{\text{tox}} = 1$  when  $pIC_{50} < 3.3$ ,  $p_{\text{tox}} = 0.5$  when  $pIC_{50} = 5.4$ ) were created.

### DesPot-Grid

The DesPot-Grid algorithm was used for the targeted design of promising anti-HIV drugs. In DesPot-Grid algorithm, the structure of the

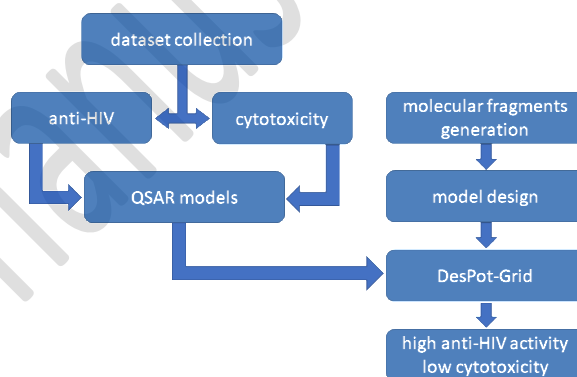
molecule is represented by a set of connected blocks. Each block corresponds to a set of structural fragments (radicals), each of which has its own unique serial number. Each fragment (block) has free valence(s), which are used to connect to neighboring fragments (blocks), *via* single, double, or triple bond. Since the algorithm uses genetic algorithm for optimizing the target function (high anti-HIV activity, low cytotoxicity), the number of genes in chromosome is equal to the number of structural fragments constituting the molecule. As a result, the genetic code (chromosome) of the molecule is represented by a set of serial numbers of structural fragments (radicals) in each block (Table 1). A genetic algorithm also generates novel, promising structures. 2D structures of fragments used in this investigation, together with the positions of open valences, are presented in Supporting Information (Figures SI1-3).

Table 1. An example of crossover and mutation in the generation of an offspring with genetic algorithm implemented in the DesPot-Grid.

	MBF3	SF2	SF1	SF1'	SF1''
Parent1	6	25	10	19	73
Parent2	2	39	46	24	6
Child1	2	25	10	19	73
Child2	6	39	46	24	6
Mutant1	2	25	10	19	79
Mutant2	6	39	46	60	6

At the first step, an initial population formation of 100 parent structures is generated by random selection of the main block and fragments according to the selected model (Figure 1). Each structure is represented by its own genetic code and the probability of their prospects. All geometries are optimized using approach based on MM3 molecular mechanics force field and AlteQ method,<sup>45,50,51</sup> followed by predictions of anti-HIV activity and cytotoxicity. Then, at the stage of the production of the offspring of the next generation, the processes of crossover and mutation are carried out. An example of crossover and mutation for two parents from

model B is presented in Table 1. From chromosomes of two parent molecules two offspring (child) chromosomes are generated. In addition to them, two more children with mutations are formed, where the value of one gene change randomly. All four child chromosomes are converted into the 3D structure, structures were optimized and the goodness score is calculated. In case if newly generated structure has higher goodness score than parent molecules used for its generation, the chromosome of weaker parent is replaced by the child chromosome. The workflow from the design of the QSAR model and molecular fragments to the final hit compounds with the desired properties is shown in Scheme 2.



**Scheme 2.** The workflow for identifying compounds with low cytotoxicity high and anti-HIV activity.

### Novelty testing of hits

The topological molecular fingerprint (FP) were calculated for the best set of compounds generated by DesPot-Grid algorithm and for FDA approved HIV medicines (Table SI4). We adopted extended-connectivity fingerprints (ECFPs) method,<sup>55</sup> as implemented in RDKit package (2020.03.1),<sup>56</sup> with radius of the fingerprint set to four. The molecular similarity was estimated using Tanimoto coefficient,  $T_c$ ,<sup>57</sup>

$$T_c(A, B) = \frac{c}{a + b - c} \quad (7)$$

where  $a$  and  $b$  are numbers of features present in compounds  $A$  and  $B$ , respectively.  $c$  is number of features shared by compounds  $A$  and  $B$ . The Tanimoto coefficient was calculated for each DesPot-Grid compound - FDA approved HIV medicine pair and for DesPot-Grid compound - 3CLpro inhibitors. The 3CLpro inhibitors' set included ten the most active compounds from the training set used to build QSAR model from the reference<sup>42</sup> and direct antiviral drugs from DrugBank database.<sup>58,59</sup>

### Molecular Docking

Open Babel, an open chemical toolbox,<sup>60</sup> was used to convert sdf files with 3D structures of compounds (compounds generated by DesPot-Grid algorithm and twenty FDA approved HIV medicines listed in Table S14) to pdb file format. Python script `prepare_ligand4.py`, a part of AutoDock Tools,<sup>61</sup> converted pdb files to pdbqt file format, a suitable format for running molecular docking experiments.

From Protein Data Bank<sup>62</sup> 3D structures of nevirapine - HIV-1 reverse transcriptase (1VRT, resolution 2.20 Å) and darunavir - HIV-1 protease (4DQB, resolution 1.50 Å) were downloaded. Since there are missing residues in 1VRT structure, they were modelled through the Chimera's<sup>63</sup> interface to Modeller program<sup>64</sup>. All missing residues are far from catalytic site, where the nevirapine is bound. PDB2PQR web server<sup>65</sup> was used to predict protonation states of residues' side chains. We utilize AutoDock Tools to add Gasteiger charges to each atom, to merge nonpolar hydrogens, to determine atom types, and to save the structures of the prepared receptors as pdbqt files. All water molecules were removed. The center of the grid box was at the center of the mass of the ligand, with Cartesian coordinates 1.5, -36.7, 22.3 Å for 1VRT and 19.9, 29.7, 14.0 Å for 4DQB, and the size of the grid box was set to 25 × 25 × 25 Å and 20 × 20 × 20 Å for 1VRT and 4DQB, respectively. The number of modes and the exhaustiveness were set to 100. All structures within 4 kcal mol<sup>-1</sup> relative to the conformation with the best binding score were saved. The plausibility of the

conformation was checked by visual inspection. Docking was performed using the AutoDock Vina suite.<sup>66</sup>

## Results and Discussion

### DesPot-Grid Algorithm

The main goal of this paper is to propose novel molecules as potent anti-HIV drugs, with high anti-HIV activity and low possibility of unwanted side effects. To minimize probability of side effects, the goodness score ( $g$ ) combining anti-HIV activity and cytotoxicity is introduced:

$$g = \left( \prod_{i=1}^{n=18} p_{anti-HIV,i} \prod_{j=1}^{m=12} p_{tox,j} \right)^{\frac{1}{n+m}} \quad (8)$$

$g$  is a geometrical mean of a scores of our previously described 18 anti-HIV activity prediction models ( $p_{anti-HIV}$ ) and 12 cytotoxicity models ( $p_{tox}$ ) from the Chemosophia web service. In addition, the probability of being active (PA) and the probability of not being toxic (PT) were calculated for each compound:

$$PA = \left( \prod_{i=1}^{n=18} p_{anti-HIV,i} \right)^{\frac{1}{n}} \quad (9)$$

$$PT = \left( \prod_{j=1}^{m=12} p_{tox,j} \right)^{\frac{1}{m}} \quad (10)$$

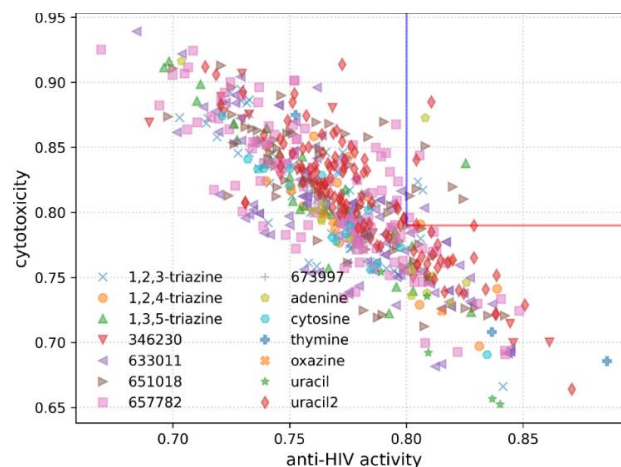
As described in Methods section, PA is a number between 0 and 1, and if PA is above 0.74, the compound is classified as active. PT, describing the cytotoxicity, is also a number between 0 and 1, where low PT values are characteristic for toxic compounds, while non-toxic compounds have PT value close to 1.

DesPot-Grid calculations were performed on Intel® Core™ i7-6700K CPU @ 4.00 GHz, 32 GB RAM, and 64-bit Windows 10 Pro operating system. At the moment, the code is not parallelized, so on average, one DesPot-Grid



calculation was running for 14 days on one core. It includes evolution of 10000 generations of new molecules within genetic algorithm, together with global minima search, anti-HIV activity and cytotoxicity predictions for each molecule that has been generated (child and mutants). The computation time is dependent on the size and the floppiness of the generated molecules, and the bottleneck is a geometry optimization step.

For each of six models for generation of new molecules (Figure 1), 100 most potent compounds after 10000 genetic algorithm generations are saved and closely analyzed (Figure 2). 16 novel compounds, together with 2D structures, goodness (g), predicted anti-HIV activity (PA) and predicted cytotoxicity (PT) scores are reported in Table 2. The criteria for selecting those molecules were PA score above 0.80 and PT score above 0.79, to maximize the probability of identifying active compounds with low cytotoxicity. On Figure S14 is presented comparison of PA and PT scores for initially generated parent molecules and for final population, obtained after 10000 generation for model B. As expected, the initial population is dominated by cytotoxic molecules with low anti-HIV activity. Average PA and PT scores of the initial population are 0.54 and 0.04, respectively. Genetic algorithm simulating natural processes of mutations and cross overs, successfully optimized population, what is reflected by increase of average PA to 0.77 and PT to 0.81. At the end, all molecules are classified as non-cytotoxic and 79 out of 100 compounds in the population have PA score above 0.74, what classify them as anti-HIV active molecules. Another consequence of 'natural selection' is survival of only the fittest main blocks, with the reduction of initial 18 to final eight main blocks in the population.



**Figure 2.** Predicted anti-HIV activity and cytotoxicity for 600 compounds identified by DesPot-Grid algorithm as potent and non-toxic drugs against HIV. 16 hit molecules are within square defined by anti-HIV activity above 0.80 (blue line) and cytotoxicity above 0.79 (red line). Legend is referring to the structure of the main block.

**Table 2.** Hit molecules with 2D structures and the goodness (g), anti-HIV activity (PA) and cytotoxicity (PT) scores.

Molecule	2D structure	g	PA	PT
1		0.840	0.811	0.885
2		0.833	0.808	0.873
3		0.830	0.825	0.838
4		0.825	0.808	0.851

5		0.82 1	0.80 6	0.84 4
6		0.82 0	0.82 0	0.82 0
7		0.81 9	0.82 4	0.81 0
8		0.81 9	0.80 4	0.84 0
9		0.81 2	0.80 5	0.82 4
10		0.81 1	0.80 7	0.81 7
11		0.81 0	0.80 4	0.81 8
12		0.80 6	0.80 8	0.80 3
13		0.80 6	0.80 1	0.81 3
14		0.80 0	0.80 6	0.79 1
15		0.79 9	0.80 1	0.79 6
16		0.79 9	0.80 4	0.79 0

### SARS-CoV-2 Repurposing Study

With the emergence of COVID-19 pandemic, drug repurposing studies included several

antiviral drugs, including a well-known HIV-1 protease inhibitor lopinavir.<sup>42,67–70</sup> Among the FDA approved antiviral drugs, it has the highest predicted activity against 3CLpro.<sup>42</sup> Unfortunately, it was demonstrated by Zhang *et al.*<sup>71</sup> that although it is good inhibitor of SARS-CoV-2 3CLpro *in vitro*, it is not effective *in vivo* due to very low concentration of free lopinavir, unbound to plasma proteins. The same QSAR model as in reference<sup>42</sup> was used to estimate inhibition power of 600 compounds designed by DesPot-Grid algorithm against SARS-CoV-2 3CLpro. Top 12 molecules ranked by its potential to inhibit 3CLpro are presented in Table 3. Compound **17** has the highest predicted inhibition power, with PA equal to 0.971, with low cytotoxicity (PT = 0.886). To put those numbers into broader perspective, let us mention that lopinavir have PA score of 0.991 and atazanavir 0.865. Additional investigation is needed to reveal a mechanism of inhibition, ADMET properties and to experimentally confirm our predictions.

**Table 3.** Hit molecules with 2D structures, activity against 3CLpro (PA) and cytotoxicity (PT).

Molecule	2D structure	PA	PT
<b>17</b>		0.971	0.886
<b>18</b>		0.949	0.789
<b>19</b>		0.944	0.884
<b>20</b>		0.942	0.761
<b>21</b>		0.937	0.856

22		0.934	0.806
23		0.925	0.872
24		0.921	0.864
10		0.919	0.817
25		0.919	0.865
26		0.907	0.821
27		0.902	0.763

### Novelty testing of hits

Although ECFPs were initially developed for capturing molecular features relevant to molecular activity, they proved useful in similarity searching, virtual screening and clustering.<sup>55</sup> Here, we exploited ECFPs and Tanimoto index to estimate molecular similarity between newly generated anti-HIV compounds and approved HIV medicines. This information might give initial insight into the mechanism of action of potential drug.

Three the most similar compounds within hit set are compounds **4**, **6** and **7** (Figure S15). The Tanimoto's similarity index between **4** and **6** is 0.73, the only difference being the additional hydroxyl group on ethyl moiety of **6**. Compound **7** differs from **4** having piperidine instead piperazine bound to nitrogen atom of pyrrole ring. Second subset of similar compounds constitutes compounds **1**, **8** and **16**. *t*-butyl

moiety directly bound to azo group in **1** is replaced by *n*-propyl group in **8**. This substitution results with slightly lower anti-HIV bioactivity and slightly higher cytotoxicity. The substitution of hydrazyl in **1** with methoxy group (**16**) is reflected mainly by the increase of cytotoxicity, while predicted bioactivity remains practically the same.

The FDA approved HIV medicines can be classified as nucleoside reverse transcriptase inhibitors (NRTIs) (abacavir, emtricitabine, lamivudine, tenofovir disoproxil fumarate, and zidovudine), non-nucleoside reverse transcriptase inhibitors (NNRTIs) (doravirine, efavirenz, etravirine, nevirapine, and rilpivirine), protease inhibitors (PIs) (atazanavir, darunavir, fosamprenavir, ritonavir, saquinavir, and tipranavir), CCR5 antagonists (maraviroc), attachment inhibitors (AI) (fostemsavir) and integrase inhibitors (II) (dolutegravir, raltegravir). In general, our compounds have low molecular similarity compared to FDA approved HIV medicines (Figure S16). For example, zidovudine, a nucleoside reverse transcriptase inhibitor, is the medicine with the highest molecular similarity index – 0.24, 0.22 and 0.21 with compounds **16**, **1** and **8**, respectively. Compound **15** is the most similar with nevirapine (NNRTI), and has the highest Tanimoto's index with five out of six protease inhibitors (fosamprenavir being the only exception). It also has the highest similarity with nevirapine and rilpivirine (both NNRTIs).

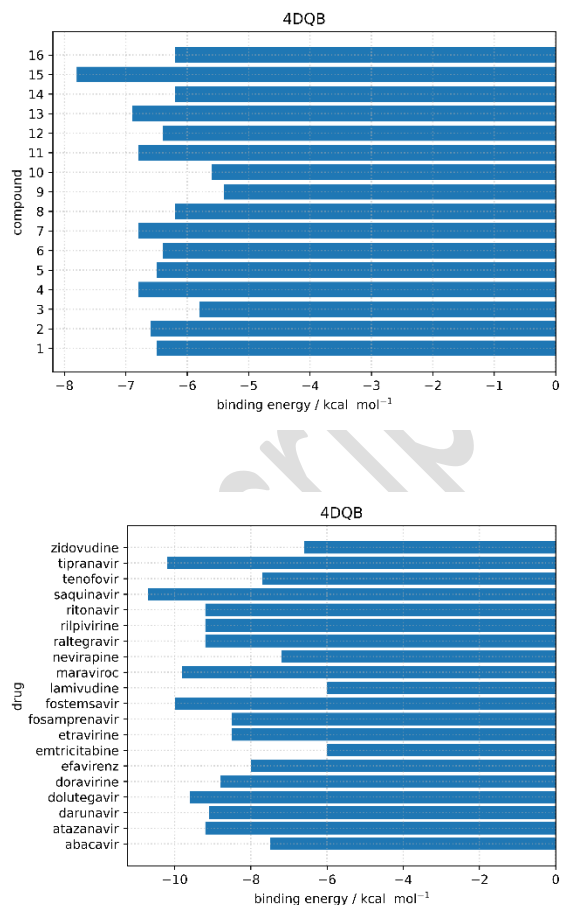
Top 12 molecules identified as potential 3CLpro inhibitors have low similarity index with ten the most active compounds from the training set from reference<sup>42</sup> (Figure S17), with the highest similarity being just above 0.12. However, they have slightly higher similarity with direct antiviral drugs (Figure S18). For example, compound **17** have Tanimoto's indices of 0.19 and 0.18 with Vidarabine and Famciclovir, respectively. Both Vidarabine and Famciclovir are analogues of purine nucleotide bases, while **17** has 1,3,5-triazine ring as the main building block.

Favipiravir, a drug inhibiting replication of influenza A and B, and compound **24** are the most similar, with Tanimoto's index equal to 0.23. They have in common six membered heterocyclic ring with two nitrogen atoms and one of the substituents being halogen atom. On May 11, there are 42 reported clinical studies on [clinicaltrials.gov](http://clinicaltrials.gov) investigating the potential of favipiravir against COVID-19, supporting potential usage of compound **24** as anti-COVID-19 medicine.

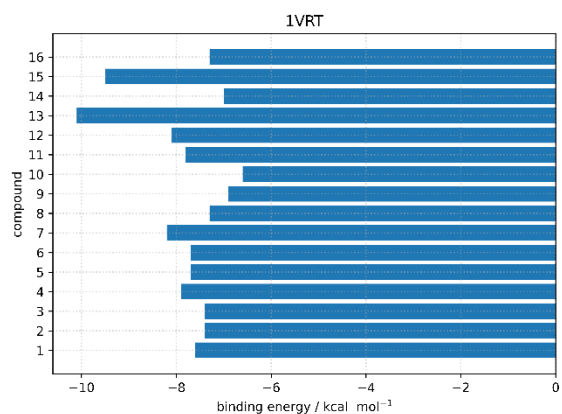
### Molecular Docking

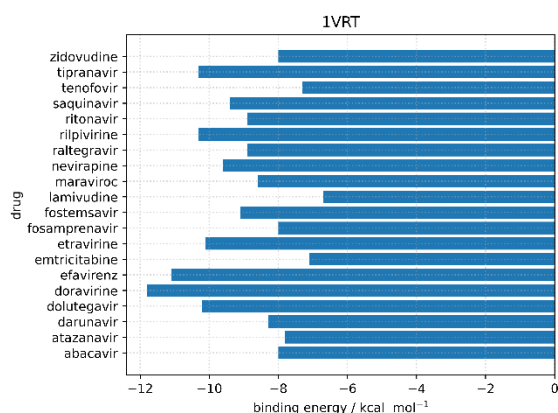
Based on molecular similarity, as target molecules for molecular docking studies we selected enzymes from the complexes of drugs nevirapine and darunavir with wild type HIV-1 reverse transcriptase and wild type HIV-1 protease, respectively.

FDA approved protease inhibitors have better affinity towards HIV-1 protease, compared to compounds designed by DesPot-Grid algorithm (Figure 3). For example, drugs saquinavir and darunavir have binding energy of  $-10.7 \text{ kcal mol}^{-1}$  and  $-9.1 \text{ kcal mol}^{-1}$ . The calculated RMSD is  $0.873 \text{ \AA}$  for poses of darunavir within the catalytic pocket for structures obtained experimentally by X-ray diffraction (4DQB) and predicted docking experiment (Figure SI9). The hydroxyl group of darunavir is in both cases favorably oriented in the proximity of neutral aspartic acid (ASH25) side chain to form hydrogen bond. Saquinavir follows the same interaction pattern (data not shown). Compound **15** has the lowest binding energy of all proposed new anti-HIV medicines ( $-7.8 \text{ kcal mol}^{-1}$ ). It is oriented in different fashion within the pocket. Oxygen atom from ether group is hydrogen bonded to hydrogen atom from peptide bond of chain B isoleucine residue (ILE50). The shortest distance to ASP25/ASH25 is  $3.67 \text{ \AA}$ , so it is evident that possible inhibition of HIV-1 protease by compound **15** includes interaction with residues outside the conserved catalytic triad (ASP25, THR26, GLY27) (Figure SI10, Table S15).



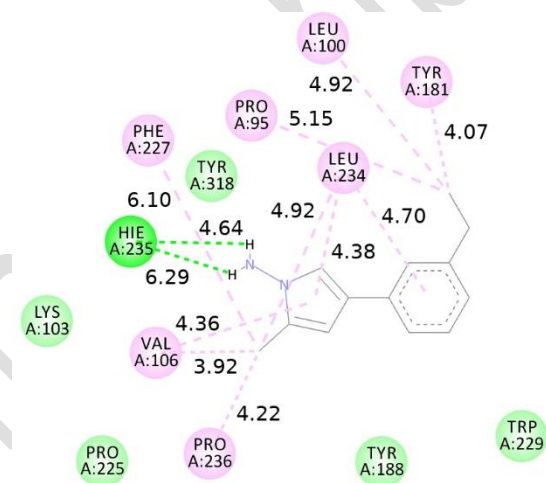
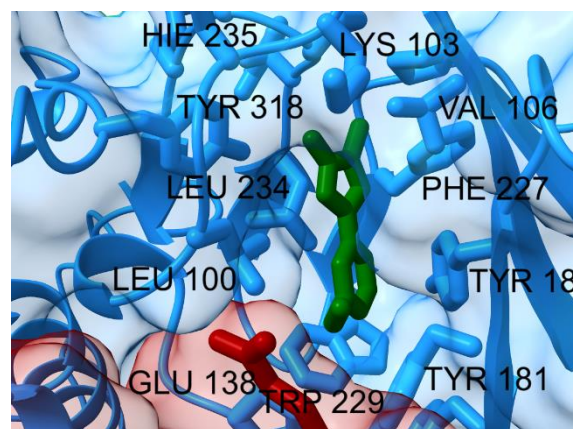
**Figure 3.** Binding energies obtained by docking experiments of novel compounds (top) and FDA approved anti-HIV medicines (bottom) to the wild type HIV-1 protease (PDB ID 4DQB).





**Figure 4.** Binding energies obtained by docking experiments of novel compounds (top) and FDA approved anti-HIV medicines (bottom) to the wild type HIV-1 reverse transcriptase (PDB ID 1VRT).

Two compounds with the best binding scores to wild type HIV-1 reverse transcriptase are compounds **13** (-10.1 kcal mol<sup>-1</sup>) and **15** (-9.5 kcal mol<sup>-1</sup>). Their score is comparable with the scores of several approved RT inhibitors (rilpivirine, etravirine, nevirapine), and slightly worse than doravirine and efavirenz (Figure 4). According to the analysis of the interactions between the drug and HIV-1 RT based on X-ray resolved 3D structure of HIV-1 RT – nevirapine complex,<sup>72</sup> and our docking study, nevirapine interacts with the enzyme dominantly *via* hydrophobic interactions with LEU100, VAL106, VAL179, TYR181, TRP229, LEU234 and TYR318 (Figure S11). Hydrophobic interactions also play crucial role in positioning of compound **13** into the binding pocket (Figure 5). Additional stabilizing factor is weak hydrogen bond between amino group of **13** and carboxyl oxygen from peptide bond of histidine 235 of A chain. With favorable orientation of hydroxyl group of tyrosine, nitrogen atom of **13** can serve as proton acceptor and form one extra hydrogen bond.



**Figure 5.** Insight into the catalytic site of wild type HIV-1 reverse transcriptase (1VRT) with docked compound **13** (green, top) and interaction pattern (bottom).

## Conclusions

The dream of anyone involved in drug development is to develop a drug that has high activity against the selected target with high selectivity and low cytotoxicity. In this article, we combine CoMIn and CiS algorithms with machine learning to propose new models for cytotoxicity and anti-HIV activity prediction. The cytotoxicity model was developed based on 100 of the most toxic and 100 of the least toxic compounds from the US National Cancer Institute. Anti-HIV training set was extracted from the *AIDS Antiviral Screen Data*, where the EC<sub>50</sub> concentrations required to observe a protective

effect on the infected cells were provided. Neural networks and linear regression algorithms were used to create twelve cytotoxicity models that yielded a cytotoxicity score based on a desirability function. The model was validated by 10-fold cross-validation, and the  $R^2$  values are within an interval of 0.91 to 0.99.

After the successful design of the cytotoxicity model, it was coupled with the model for predicting anti-HIV activity and integrated into the DesPot-Grid method. 18 QSAR models obtained by linear regression and the neural network algorithm and CiS and CoMIn descriptors were combined to calculate the probability of anti-HIV bioactivity. The 10-fold cross-validation and  $R^2$  values above 0.89 confirmed the high predictive power of the newly developed anti-HIV activity models. DesPot-Grid is a program for the development of new potential drugs based on a genetic algorithm. It creates molecules from a predefined base of fragments linked by predefined open valences. Each molecule is optimized and its activity and cytotoxicity values are predicted. We proposed to calculate the goodness score as the geometric mean of predicted (anti-HIV) activity and predicted cytotoxicity. The potential of the introduced methods is demonstrated by the development of new compounds with anti-HIV activity. In total, six different schemes with 20 main blocks with three possible substitution sites and 127 substituents with one and two open valences yielded 600 compounds with the desired properties. The genetic algorithm was shown to enrich the initial population of toxic and non-active molecules with non-toxic molecules with high anti-HIV activity potential. While for the initial pool of 100 randomly generated compounds the average values for cytotoxicity and anti-HIV activity scores were 0.04 (very toxic) and 0.54 (not active), the scores of the optimized pool after 10000 generations were 0.81 (non-toxic) and 0.77 (with anti-HIV activity), clearly demonstrating the potential of the proposed protocol. Our approach can be easily

extended by adding additional QSAR models (e.g., for metabolic prediction) and is transferable to other areas of chemistry for which suitable QSAR models are available (e.g., for pesticide development).

The low molecular similarity of the top 16 hit molecules to FDA-approved anti-HIV drugs suggests that the developed compounds belong to a new class of molecules with anti-HIV activity. Molecular docking experiments were used to test two potential mechanisms of action, namely the possibility of acting as HIV-1 protease inhibitors and as HIV-1 reverse transcriptase inhibitors. Compound **15** has a slightly lower docking score than the approved HIV-1 protease inhibitors, while compounds **13** and **15** have comparable docking scores to the approved HIV-1 reverse transcriptase inhibitors. These results should be taken with caution,<sup>73</sup> as further computational (molecular dynamics simulations) or experimental (*in vitro* and *in vivo*) studies are needed.

Inspired by drug repurposing studies, we predicted the potential of newly developed anti-HIV compounds as SARS-CoV-2 3CLpro inhibitors. To this end, the QSAR model based on the reconstruction of a model receptor molecular field (CiS algorithm) was used to evaluate the suitability of the compounds as 3CLpro inhibitors. Twelve molecules were identified to have high inhibitory potential against the main SARS-CoV-2 protease, with compound **17** ((E)-4-(3,3-dimethyltriaz-1-en-1-yl)-1,3,5-triazin-2-amine) having the highest activity score.

### Data Availability Statement

The data that support the findings of this study are available from M.A.G. (E-mail: grishinama@susu.ru) upon reasonable request.

### Acknowledgments

This work was supported by the RFBR, DST, CNPq and SAMRCA under Grant 20-53-80002.

**Author Contributions:** Conceptualization: V.A.P., M.A.G., J.N., Methodology: V.A.P., M.A.G., J.N., Software: V.A.P., M.A.G., J.N., Validation: J.N., Formal analysis: J.N., Investigation: J.N., Data curation: P.P., J.N., Writing – J.N., P.P., V.A.P., Visualization: J.N., Funding acquisition: V.A.P.

**Keywords:** HIV-1 protease; cytotoxicity; QSAR; drug repurposing; 3CLpro

Additional Supporting Information may be found in the online version of this article.

## References and Notes

- (1) Menéndez-Arias, L. Molecular Basis of Human Immunodeficiency Virus Type 1 Drug Resistance: Overview and Recent Developments. *Antiviral Res* **2013**, *98* (1), 93–120. <https://doi.org/10.1016/j.antiviral.2013.01.007>.
- (2) World Health Organization. <https://www.who.int/hiv/data/en/> (accessed 2019-09-16).
- (3) Kramer-Hämmerle, S.; Rothenaigner, I.; Wolff, H.; Bell, J. E.; Brack-Werner, R. Cells of the Central Nervous System as Targets and Reservoirs of the Human Immunodeficiency Virus. *Virus Res* **2005**, *111* (2 SPEC. ISS.), 194–213. <https://doi.org/10.1016/j.virusres.2005.04.009>.
- (4) Palmisano, L.; Vella, S. A Brief History of Antiretroviral Therapy of HIV Infection: Success and Challenges. *Ann Ist Super Sanita* **2011**, *47* (1), 44–48. [https://doi.org/10.4415/ANN\\_11\\_01\\_10](https://doi.org/10.4415/ANN_11_01_10).
- (5) Maartens, G.; Celum, C.; Lewin, S. R. HIV Infection: Epidemiology, Pathogenesis, Treatment, and Prevention. *The Lancet* **2014**, *384* (9939), 258–271. [https://doi.org/10.1016/S0140-6736\(14\)60164-1](https://doi.org/10.1016/S0140-6736(14)60164-1).
- (6) Cuevas, J. M.; Geller, R.; Garijo, R.; López-Aldeguer, J.; Sanjuán, R. Extremely High Mutation Rate of HIV-1 In Vivo. *PLoS Biol* **2015**, *13* (9), 1–19. <https://doi.org/10.1371/journal.pbio.1002251>.
- (7) Pillay, D.; Taylor, S.; Richman, D. D. Incidence and Impact of Resistance against Approved Antiretroviral Drugs. *Rev Med Virol* **2000**, *10* (4), 231–253. [https://doi.org/10.1002/1099-1654\(200007/08\)10:4<231::AID-RMV290>3.0.CO;2-P](https://doi.org/10.1002/1099-1654(200007/08)10:4<231::AID-RMV290>3.0.CO;2-P).
- (8) Damm, K. L.; Ung, P. M. U.; Quintero, J. J.; Gestwicki, J. E.; Carlson, H. A. A Poke in the Eye: Inhibiting HIV-1 Protease through Its Flap-Recognition Pocket. *Biopolymers* **2008**, *89* (8), 643–652. <https://doi.org/10.1002/bip.20993>.
- (9) Speck-Planche, A.; Kleandrova, V. v.; Luan, F.; Cordeiro, M. N. D. S. A Ligand-Based Approach for the in Silico Discovery of Multi-Target Inhibitors for Proteins Associated with HIV Infection. *Mol Biosyst* **2012**, *8* (8), 2188–2196. <https://doi.org/10.1039/c2mb25093d>.
- (10) Frańczek, T.; Siwek, A.; Paneth, P. Assessing Molecular Docking Tools for Relative Biological Activity Prediction: A Case Study of Triazole HIV-1 NNRTIs. *J Chem Inf Model* **2013**, *53* (12), 3326–3342. <https://doi.org/10.1021/ci400427a>.
- (11) Gu, W.-G.; Zhang, X.; Yuan, J.-F. Anti-HIV Drug Development Through Computational Methods. *AAPS J* **2014**, *16* (4), 674–680. <https://doi.org/10.1208/s12248-014-9604-9>.
- (12) Zakariazadeh, M.; Barzegar, A.; Soltani, S.; Aryapour, H. Developing 2D-QSAR Models for Naphthyridine Derivatives against HIV-1 Integrase Activity. *Medicinal Chemistry Research* **2015**, *24* (6), 2485–2504. <https://doi.org/10.1007/s00044-014-1305-5>.

- (13) Li, N.; Ainsworth, R. I.; Ding, B.; Hou, T.; Wang, W. Using Hierarchical Virtual Screening To Combat Drug Resistance of the HIV-1 Protease. *J Chem Inf Model* **2015**, *55* (7), 1400–1412. <https://doi.org/10.1021/acs.jcim.5b00056>.
- (14) Hosseini, A.; Alibés, A.; Noguera-Julian, M.; Gil, V.; Paredes, R.; Soliva, R.; Orozco, M.; Guallar, V. Computational Prediction of HIV-1 Resistance to Protease Inhibitors. *J Chem Inf Model* **2016**, *56* (5), 915–923. <https://doi.org/10.1021/acs.jcim.5b00667>.
- (15) Ghosh, A. K.; Osswald, H. L.; Prato, G. Recent Progress in the Development of HIV-1 Protease Inhibitors for the Treatment of HIV/AIDS. *J Med Chem* **2016**, *59* (11), 5172–5208. <https://doi.org/10.1021/acs.jmedchem.5b01697>.
- (16) Zhan, P.; Pannecouque, C.; de Clercq, E.; Liu, X. Anti-HIV Drug Discovery and Development: Current Innovations and Future Trends. *J Med Chem* **2016**, *59* (7), 2849–2878. <https://doi.org/10.1021/acs.jmedchem.5b00497>.
- (17) Inthajak, K.; Khamsemanan, N.; Nattee, C.; Toochinda, P.; Lawtrakul, L. A Prediction Approach for Anti-HIV Activity of HEPT Compounds Using Random Forest Technique. *Monatshefte für Chemie - Chemical Monthly* **2017**, *148* (10), 1697–1709. <https://doi.org/10.1007/s00706-017-1945-5>.
- (18) Kaiser, T. M.; Burger, P. B.; Butch, C. J.; Pelly, S. C.; Liotta, D. C. A Machine Learning Approach for Predicting HIV Reverse Transcriptase Mutation Susceptibility of Biologically Active Compounds. *J Chem Inf Model* **2018**, *58* (8), 1544–1552. <https://doi.org/10.1021/acs.jcim.7b00475>.
- (19) Chen, J.; Peng, C.; Wang, J.; Zhu, W. Exploring Molecular Mechanism of Allosteric Inhibitor to Relieve Drug Resistance of Multiple Mutations in HIV-1 Protease by Enhanced Conformational Sampling. *Proteins: Structure, Function and Bioinformatics* **2018**, *86* (12), 1294–1305. <https://doi.org/10.1002/prot.25610>.
- (20) Novak, J.; Grishina, M. A.; Potemkin, V. A.; Gasteiger, J. Performance of Radial Distribution Function-Based Descriptors in the Chemoinformatic Studies of HIV-1 Protease. *Future Med Chem* **2020**, *12* (4), 299–309. <https://doi.org/10.4155/fmc-2019-0241>.
- (21) Novak, J.; Grishina, M. A.; Potemkin, V. A. Novel Radial Distribution Function Approach in the Study of Point Mutations: The HIV-1 Protease Case Study. *Future Med Chem* **2020**, *12* (11), 1025–1036. <https://doi.org/10.4155/fmc-2020-0042>.
- (22) Stolbov, L.; Druzhilovskiy, D.; Rudik, A.; Filimonov, D.; Poroikov, V.; Nicklaus, M. AntiHIV-Pred: Web-Resource for in Silico Prediction of Anti-HIV/AIDS Activity. *Bioinformatics* **2020**, *36* (3), 978–979. <https://doi.org/10.1093/bioinformatics/btz638>.
- (23) Novak, J.; Grishina, M. A.; Potemkin, V. A. The Influence of Hydrogen Atoms on the Performance of Radial Distribution Function-Based Descriptors in the Chemoinformatic Studies of HIV-1 Protease Complexes with Inhibitors. *Curr Drug Discov Technol* **2021**, *18* (3), 414–422. <https://doi.org/10.2174/1570163817666200102130415>.
- (24) Wang, Y.; Lv, Z.; Chu, Y. HIV Protease Inhibitors: A Review of Molecular Selectivity and Toxicity. *HIV/AIDS - Research and Palliative Care* **2015**, *7*, 95. <https://doi.org/10.2147/HIV.S79956>.
- (25) Brook, I. Approval of Zidovudine (AZT) for Acquired Immunodeficiency Syndrome. *JAMA* **1987**, *258* (11), 1517.



- <https://doi.org/10.1001/jama.1987.03400110099035>.
- (26) Tong, J.; Zhan, P.; Bai, M.; Yao, T. Molecular Modeling Studies of Human Immunodeficiency Virus Type 1 Protease Inhibitors Using Three-Dimensional Quantitative Structure-Activity Relationship, Virtual Screening, and Docking Simulations. *J Chemom* **2016**, *30* (9), 523–536. <https://doi.org/10.1002/cem.2809>.
- (27) Ravichandran, V.; Prashantha Kumar, B. R.; Sankar, S.; Agrawal, R. K. Predicting Anti-HIV Activity of 1,3,4-Thiazolidinone Derivatives: 3D-QSAR Approach. *Eur J Med Chem* **2009**, *44* (3), 1180–1187. <https://doi.org/10.1016/j.ejmech.2008.05.036>.
- (28) Kleandrova, V. V.; Speck-Planche, A. Multitasking Model for Computer-Aided Design and Virtual Screening of Compounds With High Anti-HIV Activity and Desirable ADMET Properties. In *Multi-Scale Approaches in Drug Discovery*; Elsevier, 2017; pp 55–81. <https://doi.org/10.1016/B978-0-08-101129-4.00003-5>.
- (29) Lo, Y. C.; Rensi, S. E.; Torng, W.; Altman, R. B. Machine Learning in Chemoinformatics and Drug Discovery. *Drug Discov Today* **2018**, *23* (8), 1538–1546. <https://doi.org/10.1016/j.drudis.2018.05.010>.
- (30) Mak, K. K.; Pichika, M. R. Artificial Intelligence in Drug Development: Present Status and Future Prospects. *Drug Discov Today* **2019**, *24* (3), 773–780. <https://doi.org/10.1016/j.drudis.2018.11.014>.
- (31) Wei, Y.; Li, J.; Chen, Z.; Wang, F.; Huang, W.; Hong, Z.; Lin, J. Multistage Virtual Screening and Identification of Novel HIV-1 Protease Inhibitors by Integrating SVM, Shape, Pharmacophore and Docking Methods. *Eur J Med Chem* **2015**, *101*, 409–418. <https://doi.org/10.1016/j.ejmech.2015.06.054>.
- (32) Darnag, R.; Minaoui, B.; Fakir, M. QSAR Models for Prediction Study of HIV Protease Inhibitors Using Support Vector Machines, Neural Networks and Multiple Linear Regression. *Arabian Journal of Chemistry* **2017**, *10*, S600–S608. <https://doi.org/10.1016/j.arabjc.2012.10.021>.
- (33) Xuan, S.; Wu, Y.; Chen, X.; Liu, J.; Yan, A. Prediction of Bioactivity of HIV-1 Integrase ST Inhibitors by Multilinear Regression Analysis and Support Vector Machine. *Bioorg Med Chem Lett* **2013**, *23* (6), 1648–1655. <https://doi.org/10.1016/j.bmcl.2013.01.081>.
- (34) Zorn, K. M.; Lane, T. R.; Russo, D. P.; Clark, A. M.; Makarov, V.; Ekins, S. Multiple Machine Learning Comparisons of HIV Cell-Based and Reverse Transcriptase Data Sets. *Mol Pharm* **2019**, *16* (4), 1620–1632. <https://doi.org/10.1021/acs.molpharmaceut.8b01297>.
- (35) Anand, K.; Ziebuhr, J.; Wadhvani, P.; Mesters, J. R.; Hilgenfeld, R. Coronavirus Main Proteinase (3CLpro) Structure: Basis for Design of Anti-SARS Drugs. *Science (1979)* **2003**, *300* (5626), 1763–1767. <https://doi.org/10.1126/science.1085658>.
- (36) Jin, Z.; Du, X.; Xu, Y.; Deng, Y.; Liu, M.; Zhao, Y.; Zhang, B.; Li, X.; Zhang, L.; Peng, C.; Duan, Y.; Yu, J.; Wang, L.; Yang, K.; Liu, F.; Jiang, R.; Yang, X.; You, T.; Liu, X.; Yang, X.; Bai, F.; Liu, H.; Liu, X.; Guddat, L. W.; Xu, W.; Xiao, G.; Qin, C.; Shi, Z.; Jiang, H.; Rao, Z.; Yang, H. Structure of Mpro from SARS-CoV-2 and Discovery of Its Inhibitors. *Nature* **2020**, *582* (7811), 289–293. <https://doi.org/10.1038/s41586-020-2223-y>.
- (37) Mahdi, M.; Mótyán, J. A.; Szojka, Z. I.; Golda, M.; Miczi, M.; Tózsér, J. Analysis of the Efficacy of HIV Protease Inhibitors

- against SARS-CoV-2's Main Protease. *Virology* **2020**, *17* (1), 1–8. <https://doi.org/10.1186/s12985-020-01457-0>.
- (38) Mody, V.; Ho, J.; Wills, S.; Mawri, A.; Lawson, L.; Ebert, M. C. C. J. C.; Fortin, G. M.; Rayalam, S.; Taval, S. Identification of 3-Chymotrypsin like Protease (3CLPro) Inhibitors as Potential Anti-SARS-CoV-2 Agents. *Commun Biol* **2021**, *4* (1). <https://doi.org/10.1038/s42003-020-01577-x>.
- (39) Kandagalla, S.; Rimac, H.; Gurushankar, K.; Novak, J.; Grishina, M.; Potemkin, V. Withasomniferol C, a New Potential SARS-CoV-2 Main Protease Inhibitor from the Withania Somnifera Plant Proposed by in Silico Approaches. *PeerJ* **2022**, *10*, e13374. <https://doi.org/10.7717/peerj.13374>.
- (40) Novak, J.; Rimac, H.; Kandagalla, S.; Grishina, M. A.; Potemkin, V. A. Can Natural Products Stop the SARS-CoV-2 Virus? A Docking and Molecular Dynamics Study of a Natural Product Database. *Future Med Chem* **2021**, *13* (4), 363–378. <https://doi.org/10.4155/fmc-2020-0248>.
- (41) Novak, J.; Rimac, H.; Kandagalla, S.; Pathak, P.; Naumovich, V.; Grishina, M.; Potemkin, V. Proposition of a New Allosteric Binding Site for Potential SARS-CoV-2 3CL Protease Inhibitors by Utilizing Molecular Dynamics Simulations and Ensemble Docking. *J Biomol Struct Dyn* **2021**, 1–14. <https://doi.org/10.1080/07391102.2021.1927845>.
- (42) Novak, J.; Potemkin, V. A. A New Glimpse on the Active Site of SARS-CoV-2 3CLpro, Coupled with Drug Repurposing Study. *Mol Divers* **2022**, *26* (5), 2631–2645. <https://doi.org/10.1007/s11030-021-10355-8>.
- (43) Sang, P.; Tian, S. H.; Meng, Z. H.; Yang, L. Q. Anti-HIV Drug Repurposing against SARS-CoV-2. *RSC Adv* **2020**, *10* (27), 15775–15783. <https://doi.org/10.1039/d0ra01899f>.
- (44) Ancy, I.; Sivanandam, M.; Kumaradhas, P. Possibility of HIV-1 Protease Inhibitors-Clinical Trial Drugs as Repurposed Drugs for SARS-CoV-2 Main Protease: A Molecular Docking, Molecular Dynamics and Binding Free Energy Simulation Study. *J Biomol Struct Dyn* **2020**, *0* (0), 1–8. <https://doi.org/10.1080/07391102.2020.1786459>.
- (45) Potemkin, V.; Potemkin, A.; Grishina, M. Internet Resources for Drug Discovery and Design. *Curr Top Med Chem* **2019**, *18* (22), 1955–1975. <https://doi.org/10.2174/1568026619666181129142127>.
- (46) *AIDS Antiviral Screen Data*. <https://wiki.nci.nih.gov/display/NCIDTPdata/AIDS+Antiviral+Screen+Data> (accessed 2020-04-10).
- (47) Potemkin, V.; Grishina, M. Electron-Based Descriptors in the Study of Physicochemical Properties of Compounds. *Comput Theor Chem* **2018**, *1123*, 1–10. <https://doi.org/10.1016/j.comptc.2017.11.010>.
- (48) Potemkin, V.; Palko, N.; Grishina, M. Quantum Theory of Atoms in Molecules for Photovoltaics. *Solar Energy* **2019**, *190* (August), 475–487. <https://doi.org/10.1016/j.solener.2019.08.048>.
- (49) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem in 2021: New Data Content and Improved Web Interfaces. *Nucleic Acids Res* **2021**, *49* (D1), D1388–D1395. <https://doi.org/10.1093/nar/gkaa971>.
- (50) Potemkin, V. A.; Bartashevich, E. V.; Belik, A. V. A New Approach to Predicting the Thermodynamic Parameters of Substances from Molecular

- Characteristics. *Russian Journal of Physical Chemistry* **1996**, *70* (3), 411–416.
- (51) Potemkin, V. A.; Pogrebnoy, A. A.; Grishina, M. A. Technique for Energy Decomposition in the Study of “Receptor-Ligand” Complexes. *J Chem Inf Model* **2009**, *49* (6), 1389–1406. <https://doi.org/10.1021/ci800405n>.
- (52) Potemkin, A. v.; Grishina, M. A.; Potemkin, V. A. Grid-Based Continual Analysis of Molecular Interior for Drug Discovery, QSAR and QSPR. *Curr Drug Discov Technol* **2017**, *14* (3), 1–25. <https://doi.org/10.2174/1570163814666170207144018>.
- (53) Potemkin, V.; Grishina, M. Grid-Based Technologies for In Silico Screening and Drug Design. *Curr Med Chem* **2018**, *25* (29), 3526–3537. <https://doi.org/10.2174/0929867325666180309112454>.
- (54) Potemkin, V.; Grishina, M. Principles for 3D/4D QSAR Classification of Drugs. *Drug Discov Today* **2008**, *13* (21–22), 952–959. <https://doi.org/10.1016/j.drudis.2008.07.006>.
- (55) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J Chem Inf Model* **2010**, *50* (5), 742–754. <https://doi.org/10.1021/ci100050t>.
- (56) RDKit: Open-Source Cheminformatics.
- (57) Bajusz, D.; Rácz, A.; Héberger, K. Why Is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations? *J Cheminform* **2015**, *7* (1), 1–13. <https://doi.org/10.1186/s13321-015-0069-3>.
- (58) Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; Assempour, N.; Iynkkaran, I.; Liu, Y.; Maclejewski, A.; Gale, N.; Wilson, A.; Chin, L.; Cummings, R.; Le, D.; Pon, A.; Knox, C.; Wilson, M. DrugBank 5.0: A Major Update to the DrugBank Database for 2018. *Nucleic Acids Res* **2018**, *46* (D1), D1074–D1082. <https://doi.org/10.1093/nar/gkx1037>.
- (59) DrugBank. <https://www.drugbank.ca/> (accessed 2021-02-02).
- (60) O’Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J Cheminform* **2011**, *3* (1), 33. <https://doi.org/10.1186/1758-2946-3-33>.
- (61) Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J Comput Chem* **2009**, *30* (16), 2785–2791. <https://doi.org/10.1002/jcc.21256>.
- (62) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res* **2000**, *28* (1), 235–242. <https://doi.org/10.1093/nar/28.1.235>.
- (63) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera - A Visualization System for Exploratory Research and Analysis. *J Comput Chem* **2004**, *25* (13), 1605–1612. <https://doi.org/10.1002/jcc.20084>.
- (64) Šali, A.; Blundell, T. L. Comparative Protein Modelling by Satisfaction of Spatial Restraints. *Journal of Molecular Biology*. 1993, pp 779–815. <https://doi.org/10.1006/jmbi.1993.1626>.
- (65) Dolinsky, T. J.; Nielsen, J. E.; McCammon, J. A.; Baker, N. A. PDB2PQR: An Automated Pipeline for the Setup of Poisson-Boltzmann Electrostatics Calculations. *Nucleic Acids Res* **2004**, *32* (WEB SERVER ISS.), 665–667. <https://doi.org/10.1093/nar/gkh381>.
- (66) Trott, O.; Olson, A. J. AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading. *J Comput Chem* **2010**, *31*

- (2), 455–461.  
<https://doi.org/10.1002/jcc.21334>.
- (67) Fehr, A. R.; Channappanavar, R.; Perlman, S. Middle East Respiratory Syndrome: Emergence of a Pathogenic Human Coronavirus. *Annu Rev Med* **2017**, *68*, 387–399.  
<https://doi.org/10.1146/annurev-med-051215-031152>.
- (68) Rafi, M. O.; Bhattacharje, G.; Al-Khafaji, K.; Taskin-Tok, T.; Alfasane, M. A.; Das, A. K.; Parvez, M. A. K.; Rahman, M. S. Combination of QSAR, Molecular Docking, Molecular Dynamic Simulation and MM-PBSA: Analogues of Lopinavir and Favipiravir as Potential Drug Candidates against COVID-19. *J Biomol Struct Dyn* **2020**, *0* (0), 1–20.  
<https://doi.org/10.1080/07391102.2020.1850355>.
- (69) Bolcato, G.; Bissaro, M.; Pavan, M.; Sturlese, M.; Moro, S. Targeting the Coronavirus SARS-CoV-2: Computational Insights into the Mechanism of Action of the Protease Inhibitors Lopinavir, Ritonavir and Nelfinavir. *Sci Rep* **2020**, *10* (1), 20927.  
<https://doi.org/10.1038/s41598-020-77700-z>.
- (70) Liu, J.; Zhai, Y.; Liang, L.; Zhu, D.; Zhao, Q.; Qiu, Y. Molecular Modeling Evaluation of the Binding Effect of Five Protease Inhibitors to COVID-19 Main Protease. *Chem Phys* **2021**, *542* (January), 111080.  
<https://doi.org/10.1016/j.chemphys.2020.111080>.
- (71) Zhang, L.; Liu, J.; Cao, R.; Xu, M.; Wu, Y.; Shang, W.; Wang, X.; Zhang, H.; Jiang, X.; Sun, Y.; Hu, H.; Li, Y.; Zou, G.; Zhang, M.; Zhao, L.; Li, W.; Guo, X.; Zhuang, X.; Yang, X. Lou; Shi, Z. L.; Deng, F.; Hu, Z.; Xiao, G.; Wang, M.; Zhong, W. Comparative Antiviral Efficacy of Viral Protease Inhibitors against the Novel SARS-CoV-2 In Vitro. *Viral Sin* **2020**, *35* (6), 776–784.  
<https://doi.org/10.1007/s12250-020-00288-1>.
- (72) Ren, J.; Esnouf, R.; Garman, E.; Somers, D.; Ross, C.; Kirby, I.; Keeling, J.; Darby, G.; Jones, Y.; Stuart, D.; Stammers, D. High Resolution Structures of HIV-1 RT from Four RT–Inhibitor Complexes. *Nat Struct Mol Biol* **1995**, *2* (4), 293–302.  
<https://doi.org/10.1038/nsb0495-293>.
- (73) Chen, Y. C. Beware of Docking! *Trends Pharmacol Sci* **2015**, *36* (2), 78–95.  
<https://doi.org/10.1016/j.tips.2014.12.001>.

## GRAPHICAL ABSTRACT

Jurica Novak, Prateek Pathak, Maria A. Grishina, and Vladimir A. Potemkin

The Design of Compounds with Desirable Properties – the Anti-HIV Case Study

By applying machine learning methods, new molecules with desired properties can be designed within a reasonable time frame and with low computational costs. The approach is demonstrated with the development of anti-HIV compounds with low cytotoxicity.

