Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Computational Biology and Chemistry 98 (2022) 107656

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/cbac

Machine learning prediction of 3CL^{pro} SARS-CoV-2 docking scores

Lukas Bucinsky [a,*], Dušan Bortňák [b], Marián Gall [c,d], Ján Matúška [a], Viktor Milata [b], Michal Pitoňák [d,e], Marek Štekláč [a], Daniel Végh [b], Dávid Zajaček [a]

[a] Institute of Physical Chemistry and Chemical Physics, Faculty of Chemical and Food Technology, Slovak University of Technology in Bratislava, Radlinského 9, SK-81237 Bratislava, Slovak Republic
[b] Institute of Organic Chemistry, Catalysis and Petrochemistry, Faculty of Chemical and Food Technology, Slovak University of Technology in Bratislava, Radlinského 9, SK-81237 Bratislava, Slovak Republic
[c] Institute of Information Engineering, Automation and Mathematics, Faculty of Chemical and Food Technology, Slovak University of Technology in Bratislava, Radlinského 9, SK-81237 Bratislava, Slovak Republic
[d] Computing Center, Centre of Operations of the Slovak Academy of Sciences, Dúbravská cesta č. 9, SK-84535 Bratislava, Slovak Republic
[e] Department of Physical and Theoretical Chemistry, Faculty of Natural Sciences, Comenius University in Bratislava, Mlynská dolina, Ilkovičova 6, SK-84215 Bratislava, Slovak Republic

ARTICLE INFO

Keywords:
AutoDock molecular docking
3CLpro Mpro 6WQF
Machine learning
TensorFlow XGBoost SchNetPack
COVID19
SARS-CoV-2

ABSTRACT

Molecular docking results of two training sets containing 866 and 8,696 compounds were used to train three different machine learning (ML) approaches. Neural network approaches according to Keras and TensorFlow libraries and the gradient boosted decision trees approach of XGBoost were used with DScribe's Smooth Overlap of Atomic Positions molecular descriptors. In addition, neural networks using the SchNetPack library and descriptors were used. The ML performance was tested on three different sets, including compounds for future organic synthesis. The final evaluation of the ML predicted docking scores was based on the ZINC *in vivo* set, from which 1,200 compounds were randomly selected with respect to their size. The results obtained showed a consistent ML prediction capability of docking scores, and even though compounds with more than 60 atoms were found slightly overestimated they remain valid for a subsequent evaluation of their drug repurposing suitability.

1. Introduction

The novel zoonotic coronavirus, classified by the International Committee on Taxonomy of Viruses (ICTV) on February 11, 2020 as Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), has first surfaced in late December 2019 in Wuhan province China Wu et al. (2020); Li et al. (2020a,b); Zhou et al. (2020); Wang et al. (2020); Shereen et al. (2020); Bogoch et al. (2020). SARS-CoV-2 is the causative agent of the disease COVID19, which manifests with various respiratory symptoms and can lead to death in immunocompromised patients Wu et al. (2020); Islam et al. (2020); Mallah (2021). Furthermore, post-COVID symptoms, which include respiratory, gastrointestinal, heart, and neural difficulties, are reported Nalbandian et al. (2021); Pesce et al. (2021); Yan et al. (2021). Although the means of first transmission to humans are open for discussion, the virus's impact is beyond argument. The World Health Organization was forced to declare an outbreak of COVID19 worldwide pandemic in March 2020, with about 120,000 cases detected in more than 110 countries around the world, warning of the sustained risk of further global spread Zhang et al. (2020). Eighteen months later [October 07, 2021], there have been more than 236,580,675 confirmed cases with 4,829,816 deceased worldwide Dong et al. (2020); JHU (2020).

In an immediate action, scientists around the world boosted contra action in the effort to develop possible drugs and/or identify potential targets for drug refurbishment and repurposing Tejera et al. (2020); Fischer et al. (2020); Hall and Ji (2020); Wu et al. (2020); Hosseini and Amanlou (2020); Jin (2020); Smith and Smith (2020); Batra et al. (2020); Nagar et al. (2021); Acharya et al. (2020). Currently, vaccine development by pharmaceutical companies offers great action against virus spread and harm Liu et al. (2020); Bernal et al. (2021); Nanduri et al. (2021). To this day [October 07, 2021], more than 6.4 billion vaccines Dong et al. (2020); JHU (2020) have been administered worldwide with more than 22.6 % of the world population receiving at least one dose Mathieu et al. (2021); Our World in Data, (2020).

* Corresponding author.
E-mail address: lukas.bucinsky@stuba.sk (L. Bucinsky).

https://doi.org/10.1016/j.compbiolchem.2022.107656
Received 22 October 2021; Received in revised form 23 February 2022; Accepted 24 February 2022
Available online 26 February 2022

However, new mutations have become identified Mahase (2021) and can reverse the success of the vaccination campaigns around the world. Hence, targeted drug development/refurbishment to treat, slow down, and/or deactivate the virus replication and cell entry protocol are still valid approaches. In this respect, *in silico*-based tactics appear as a very reasonable choice to shrink a large database of chosen compounds to tens or hundreds of compounds that can be potentially active and cleave a certain protein active site. In addition, we have to take into account the possibility that this is not the last pandemic. Therefore, the tools of trade still need to be developed and established Casbarra and Procacci (2021) to provide a much faster response to treat new potential pandemic risks Cho et al. (2021); Llanos et al. (2021); Muratov et al. (2021); Zev et al. (2021). The popularity of molecular docking and connected software were slowly diminishing in the 2010s due to the simplicity of its approach, while methods based on a more robust interpretation of drug-target interactions gained more popularity. However, this changed drastically with the ongoing COVID19 pandemic and with the need to screen tens of thousands of compounds. In this respect, many more research teams turned to molecular docking. Nowadays, molecular docking is mainly used for *in silico* drug design against various SARS-CoV-2 inhibition targets Elfiky (2020); Joshi et al. (2020a); Hall and Ji (2020); Kong et al. (2020); Meyer-Almes (2020); Jimenez-Alberto et al. (2021); Das et al. (2021); Guedes et al. (2021); Elmezayen et al. (2021). Still, information exchange is not always taken care of, meaning that the complete data is often not accounted for in high-throughput works published with only a few exceptions Smith and Smith (2020); Tejera et al. (2020); Steklac et al. (2021); Guedes et al. (2021). In addition, a docking protocol or a molecular dynamics verification takes considerable time and computational demands. Therefore, methods for an in-few-seconds assessment of possible evaluation (feasibility with respect to, e.g., the docking score evaluation) of large compound sets need to be developed. This is indeed a well-suited task for machine learning (ML) protocols Batra et al. (2020); Gentile et al. (2020); El-Behery et al. (2021). One of the promising machine learning techniques is Deep Tensor Neural Networks (DTNNs) implemented in the SchNetPack package Schutt et al. (2018), It has been used successfully to predict not only the energies of molecules Schutt et al. (2019a) but even the shape of the Schrödinger wavefunction Schutt et al. (2019b). In addition, other existing ML approaches and libraries, such as TensorFlow Abadi et al. (2015) and XGBoost Chen and Guestrin (2016), are to be mentioned. The latter approach has already been used for docking score prediction and/or in protein-ligand affinity studies Li et al. (2019); Lu et al. (2019); Zhang et al. (2019); Yang et al. (2019). Deep learning protocols have emerged useful in screening large sets of compounds (1.3 billion) for evaluation of targeted drug likeness Gentile et al. (2020), including COVID19 research Ton (2020); Joshi et al. (2020b); Santana and Silva-Jr (2021a,b); Acharya et al. (2020). It is worth mentioning that the employed ML protocols involve SMILES (Simplified Molecular-Input Line-Entry System) strings of molecules and a subsequent fingerprint generation upon these. Furthermore, ML protocols not only aim at predicting the docking scores (pharmacological efficacy), but can also be applied in the inverse molecular design Sanchez-Lengeling and Aspuru-Guzik (2018).

SARS-CoV-2, like most other coronaviruses, has a genomic code for 16 non-structural proteins (nsps) that are responsible for the transcription and subsequent replication of the virus, and four essential structural proteins: spike glycoprotein, envelope protein, nucleocapsid protein, and membrane protein. These proteins are responsible for the viru entry to human cells, the virion shape, the pathogenesis, and the release of viral particles, respectively Li and Kang (2020). Of these four essential proteins, the spike protein is of great interest to target to slow down the virus cell entry mechanism Wu et al. (2020); Islam et al. (2020). In addition, the spike protein is a targeted site for antibodies, which offers a possible means of COVID19 vaccination Islam et al. (2020). After the entry of the virus into the cell, deposition of the nucleocapsid into the cytoplasm occurs, initiating the translation of the viral genome into the replicase polyprotein by means of messenger RNA (mRNA) Petushkova and Zamyatnin (2020). Non-structural proteins encoded in the SARS-CoV-2 genome include various enzymes, such as 3-chymotrypsin--like protease ($3CL^{pro}$), which is sometimes denoted as M^{pro} (main protease), papain-like protease (PL^{pro}), helicase, RNA-dependent polymerase (RdRp) and primase Zumla et al. (2016). PL^{pro} and $3CL^{pro}$ are responsible for the cleavage of polyproteins translated from viral RNA into enzymes that are vital for the RNA replication: RdRp, helicase (nsp13), and nucleoside triphosphatase Petushkova and Zamyatnin (2020). Furthermore, PL^{pro} is a multifunctional protease with deubiquitinating and deISGylating activities Li and Kang (2020). Therefore, these essential proteins are considered viable targets for the treatment of COVID19.

This work focuses on $3CL^{pro}$ as a potential target for the development of antivirals, as inhibition of its activity can/will cease the pace of viral replication Li and Kang (2020). $3CL^{pro}$ is a three-domain protease, consisting of ca. 306 amino acid residues, and is highly conserved among coronaviruses. Domains I (residues 8-101) and II (residues 102-184), which form a beta-barrel secondary structure, are connected by a long loop (residues 185-200) with domain III (residues 201-303) formed by alpha-helices. The $3CL^{pro}$ substrate binding region is located around its catalytic dyad, which is capable of hydrolyzing peptide bonds in enzymes and consists of nucleophilic Cys145 located in domain II and its proton acceptor counterpart His41 located in domain I Tejera et al. (2020). This catalytic dyad along with the Thr25 amino acid residue forms a subside denoted S2. Additional S1 subside, formed by His41, Phe140, Glu 143, His163, Glu166, and His172 amino acid residues, can be found in its proximity. The S1 and S2 amino acids are mainly involved in hydrophobic and electrostatic interactions. Furthermore, there are three additional shallow subsides S3-S5 located nearby. These subsides, formed by His41, Met49, Met165, Glu166, and Gln189 amino acid residues, can tolerate different functionalities offering various options for the inhibitor-protease complex stabilization Khan et al. (2021); Jin (2020); Lu et al. (2006).

Herein, two sets of compounds were chosen as the basis of semi-flexible docking into the $3CL^{pro}$ structure with the PDB code 6WQF. The smaller set **S** contained 866 compounds Steklac et al. (2021) (COVID19 related compounds from the external PubChem database Kim et al. (2018)). The larger set **L** contained 8,696 compounds from the QSAR, docking and molecular dynamics study targeting the inhibition of $3CL^{pro}$ of Tejera *et al.* Tejera et al. (2020). The ML validation set (**V**) was based on 100 additional compounds from Tejera *et al.* Tejera et al. (2020). Several previous ML studies had employed the SMILES strings (Simplified Molecular-Input Line-Entry System) Batra et al. (2020); Gentile et al. (2020); Tejera et al. (2020); Ton (2020); Joshi et al. (2020b); Santana and Silva-Jr (2021a,b) and utilized, e.g., Morgan fingerprints Riniker and Landrum (2013) of the ligands (compounds) as a basis for the molecular descriptors employed. Herein, the direct Cartesian coordinates space approach according to the xyz (mol2/sdf/pdb) file format was chosen. Subsequently, three different machine learning (ML) approaches were applied and their potential for docking score prediction was presented. The trained ML protocols, using **S** and **L** training sets, were further validated against compounds with the best docking scores of other studies (set **B**) Wu et al. (2020); Hosseini and Amanlou (2020); Adem et al. (2020); Shah et al. (2020); Fischer et al. (2020), a set of new compounds proposed for organic synthesis (set **O**), and against a set of *in vivo* compounds from the ZINC15 database Irwin et al. (2012) (denoted the production set, **P**). To assess the quality of the ML prediction capacity of **P** data set docking scores. This set was split into 12 subsets of compounds with a defined number of atoms (1-10, 11-20, etc.) and the docking scores were calculated for a random selection of 100 compounds from each subset, designated as **P'**. The flowchart of the work presented in this paper is shown in Fig. 1.

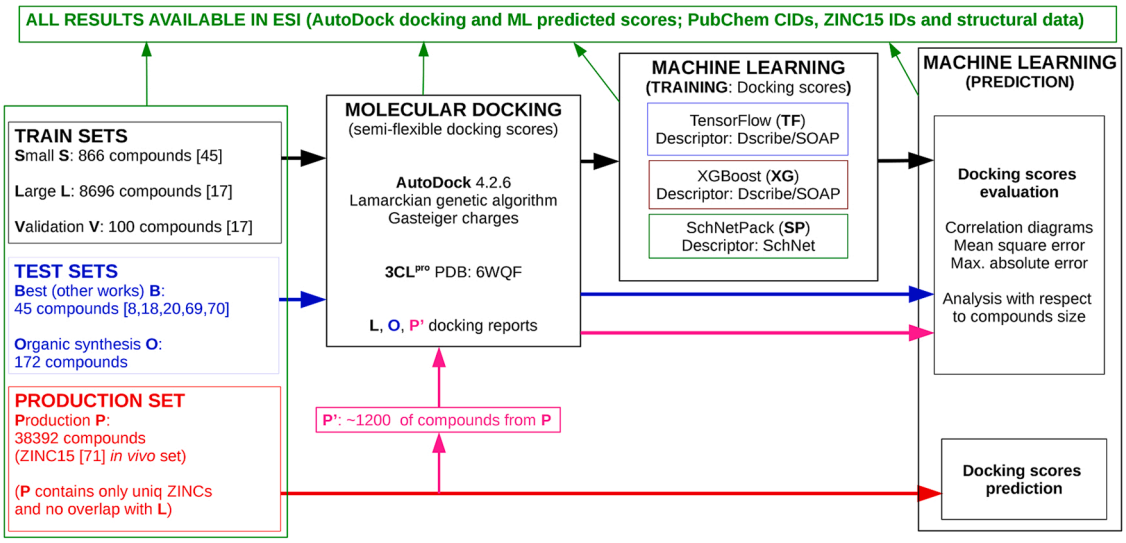

Fig. 1. Flowchart of the presented work, including the employed data sets, docking protocol details, machine learning (ML) methods, and the evaluation of ML docking scores prediction.

2. Methods

2.1. Data sets of compounds

The 3-dimensional structures of COVID19 related compounds (as previously used in the study of Štekláč *et al.* Steklac et al. (2021)) were downloaded from the external PubChem database Kim et al. (2018) [October 23, 2020]. After metal complexes (salts) and silicon containing compounds were removed, the first training set contains 866 compounds and is denoted as **S** (small).

For the second training set of compounds, a total of 10,246 structures of Tejera *et al.* Tejera et al. (2020) were downloaded from the PubChem database [January 11, 2021]. Compounds containing atoms/ions and uncrecognized molecular patterns by the AutoDock 4.2.6 force field, Morris et al. (2009); Wojciechowski (2017) as well as those that were already contained within the first 866 compounds of set **S** or were not present as 3D structures in PubChem (due to their size) were removed, leading to a total number of 8,796 compounds in the second set. Subsequently, this set was split into 8,696 compounds, denoted **L** (large set) and 100 randomly selected compounds, denoted **V** (validation set). The validation set was used to control the convergence of ML training processes.

Furthermore, 45 best scoring compounds from five different publications Wu et al. (2020); Hosseini and Amanlou (2020); Adem et al. (2020); Shah et al. (2020); Fischer et al. (2020) were compiled in the test set denoted as **B** (best). Their 3D structures were downloaded from the PubChem database (31), the ZINC database (10) Sterling and Irwin (2015); Irwin et al. (2012), and the 3D geometries of four compounds without a 3D structure in either database were optimized in the MMFF94 force field Halgren (1996a,b,c); Halgren and Nachbar (1996); Halgren (1996d) using the OpenBabel chemical toolbox 2.3.2. O'Boyle et al. (2011). In addition, the 3D geometries of 172 compounds proposed by the coauthors from the Department of Organic Chemistry of the Slovak University of Technology (SUT) were optimized at the MMFF94 level of theory, using OpenBabel, forming the test set, denoted as **O** (organic synthesis). The selection of structures for the application of computational methods to model their potential affinity to $3CL^{pro}$ was made upon the molecular skeleton/backbone patterns, which had been previously found to exhibit antiviral and/or biological activity. Adamantylamine derivatives belong to the group of the most well-known antivirals De Clercq (2011); Yet (2018); Colalto (2020). Therefore, new adamantylamine derivatives were suggested for further synthesis with either activated enol ethers or via coupling of this fragment with quinoline or 1,3,5-triazines. In addition, the possible biological activity of amine derivatives containing the same secondary diamines as chloroquine and hydroxychloroquine was taken into account Radl (2020). Pyrazole derivatives are known for their biological and antiviral effects Khan et al. (2016). Furthermore, the coauthors from the Department of Organic Chemistry at SUT are experienced with their synthesis Tarabova et al. (2014); Bortnak et al. (2018). Hence, pyrazole derivatives were also included in this docking study. One of the desired properties behind choosing the compounds in the **O** set was their volatility. In this sense, the potential antiviral activity Buhner (2013); da Silva et al. (2020) is also likely to be enhanced by selecting compounds with higher volatility from this set for subsequent synthesis (breathable medicine offers better means of drug delivery, release profile, absorption, distribution, and efficacy to treat pulmonary diseases Buhner (2013); da Silva et al. (2020)). The **O** set compounds are compiled in the Supplementary Material as two separate files. The first (csv) file includes SMILES codes, as well as overlap with the ZINC and PubChem compounds (showing the particular IDs). Docking scores and structural formulas are provided in the second (pdf) file.

Finally, the *in vivo* set of compounds from the ZINC database Irwin et al. (2012) has been chosen as the production set for the trained ML approaches, denoted **P** (production). This ZINC set (downloaded on 23th March 2021), containing 60,407 compounds, joins sets of FDA approved drugs, drugs approved only outside the US, compounds currently in any phase of clinical trials, substances with human exposure that are not approved or in clinical trials, and compounds tested *in vivo* Sterling and Irwin (2015). Compounds with the same ZINC codes and those which overlap with the other data sets were removed, leading to 38,392 unique compounds in the production **P** set.

To critically assess the performance of the employed machine learning protocols, the **P** data set was split into 12 subsets of compounds with a defined number of atoms (1-10, 11-20, etc.) and the docking scores of 100 randomly selected compounds from each subset were evaluated. This set, denoted **P'**, contains 1,181 compounds as there were fewer than 100 compounds in the subsets with the number of atoms 1-10 (89 compounds) and 111-120 (92 compounds). These 12 subsets were subsequently merged to six subsets (1-20, 21-40, 41-60, etc.) to seek brevity of the results presentation.

We present two series of ML results for the two distinguished training sets **S** and **L**. The first series serves as a demonstration of predictive capacity of the ML models used. Here, the **S** data set is used as the

training set and the remaining data sets, **L**, **B**, and **O**, as the test sets. In the second series, we present our "best results" using the large data set **L** as the training set, and the other data sets as the test/production sets. For both series, the same **V** data set was used for validation in the ML training process.

In general, the data sets employed consist of compounds that contain the following organogenic elements and halogens: N, O, F, P, S, Cl, Br, and I. The number of atoms in the studied compounds varies from 4 to 145 and the molecular weight ranges from 30 to 1283 Da. The number of compounds and the range of their atom counts for each data set are given in Table 1. The normalized distributions of all data sets with respect to the number of atoms in the compounds, split into subsets per 10 atoms, are shown in Fig. 2.

2.2. $3CL^{pro}$ structure

The 3D structure of the SARS-CoV-2 $3CL^{pro}$ protease determined at room temperature was downloaded from the RCSB Protein Data Bank Berman et al. (2000) (PDB ID: 6WQF) Kneller et al. (2020). This structure was further edited in the AutoDockTools software Morris et al. (2009); Sanner (1999), to remove water molecules. One single water molecule situated between His41 and Asp187 within the pocket of the $3CL^{pro}$ active site was retained. It had previously been reported that this water molecule participates in charge stabilization interactions of the neighbouring residues Kneller et al. (2020).

2.3. Molecular Docking

Semi-flexible dockings were performed with Autodock 4.2.6 software Morris et al. (2009); Wojciechowski (2017). Gasteiger charges were added to all compounds studied using the AutoDock utility scripts Morris et al. (2009). Potential maps were calculated within a grid box of 90 x 90 x 90, with a resolution of 0.275 Å, centered at x, y, z = (-20 Å, -5 Å, 15 Å). The Lamarckian genetic algorithm as implemented in Autodock 4.2.6 software was employed, with the total number of docking runs set to 50. Each generation contained 300 individuals, the maximum number of energy evaluations was set to 30,000,000, and the number of maximum populations was set to 27,000. Other parameters such as crossover and mutation rates, as well as parameters describing the Solis & Wets local search algorithm, were kept at their default values. All resulting docked poses were clustered with 2.0 Å tolerance and analyzed with the AutoDockTools utility Morris et al. (2009); Sanner (1999). The hydrogen bond pattern was analyzed with the LigPlot+ software Wallace et al. (1995) with the maximum acceptor-donor distance set to 3.35 Å. LigPlot+ was also used to create 2D schematic diagrams of putative docking modes.

2.4. Neural networks with TensorFlow and Keras

In the present study, we chose the Smooth Overlap of Atomic Positions (SOAP) molecular descriptor Bartok et al. (2013a,b,c). This descriptor uses a local expansion of a smeared Gaussian atomic density for the description of atomic regions in molecules and employs orthonormal density functions based on spherical harmonics and radial basis functions. The SOAP descriptor was used as implemented in the DScribe package Himanen et al. (2020), which is an easy-to-use Python code library Van Rossum and Drake (1995). The following SOAP parameters were chosen: rcut=9 (cutoff for the local region), nmax=8 (radial basis functions per atom), lmax=8 (degree of spherical harmonics), sigma=1.0 (standard deviation of the Gaussians used) and average='outer' (averaging over the power spectrum of different sites); leading to a feature vector with a dimension of 29,160 for all structures studied.

Keras Chollet (2015) and TensorFlow Abadi et al. (2015) neural networks (NN) were chosen as the first machine learning model within this study, abbreviated as TF (see Table 2). Keras is a deep learning Application Programming Interface (API) written in Python running on top of the machine learning platform TensorFlow. Several network topology models and activation functions were tested. Finally, standard softplus activation functions were used for all neurons as suggested in the paper of Profitt and Pearson Profitt and Pearson (2019). The exception to this was the last layer in which the linear activation functions were employed. The network topology was set to three dense layers, the first containing 100 neurons, the second 50 neurons, and the last with five neurons Profitt and Pearson (2019). Other neural network topologies were tested as well, taking into account up to four hidden dense layers and 200 neurons in each of the hidden layers. It was found that the huge number of input parameters per compound (29,160) entering the neural networks leads to an unpredictable dependence of the obtained results on the neural network topology (not shown). All weights in the neural networks employed were initialized randomly using the standard RandomUniform Keras function. Other parameters were kept at their default values. The NN model was trained for a fixed number of 100 total epochs using the optimizer of Kingma and Ba (2015), and an initial learning rate of 10^{-3} (other parameters were kept at their default values). The validation set **V** was used during the TF training process to choose the best NN parameters within the training process.

2.5. Gradient boosted decision trees with XGBoost

As a complementary machine learning approach to neural networks, gradient boosted decision trees (GBDT) were selected. The results presented in this work were obtained using the optimized distributed gradient boosting library XGBoost Chen and Guestrin (2016), abbreviated XG (see Table 2). The SOAP descriptor was also used with the XG approach, with only slightly modified parameters compared to TF: 7Å for the local region cutoff (rcut) and 3.0 standard deviations of the Gaussians (sigma) used to expand the atomic density. Based on the results of a coarse grain manual hyper-parameter optimization (not shown), the maximum tree depth of six and 100 boosting rounds were used in GBDT training and inference. For a comprehensive list of all XG GBDT parameters, see Table S1. A finer hyper-parameter tuning was not beneficial due to inherent biases (and possible overfitting issues) of the training sets.

2.6. Neural networks with SchNetPack

In addition, the deep learning architecture SchNet was used, as implemented in the computational package SchNetPack Schutt et al. (2018) (abbreviated SP; see Table 2) to predict the docking scores of the compounds studied. SchNet is a variant of Deep Tensor Neural Networks (DTNNs), where interactions are modeled by continuous-filter convolutions with filter-generating networks. The SchNet representation was used as a descriptor of the atomistic system of compounds in each data set. Atomic representations in SchNet had a 128-feature dimension, with

Table 1
Summary of the employed sets of compounds (number of compounds and the span of the number of atoms).

Set name	Set abbreviation	Number of compounds	Range of atom counts
Small Steklac et al. (2021)	**S**	866	6-120
Large Tejera et al. (2020)	**L**	8,696	4-117
Validation Tejera et al. (2020)	**V**	100	14-105
Best Wu et al. (2020); Hosseini and Amanlou (2020); Adem et al. (2020); Shah et al. (2020); Fischer et al. (2020)	**B**	45	26-145
Organic synthesis	**O**	172	15-120
Production Sterling and Irwin (2015)	**P**	38,392	4-119

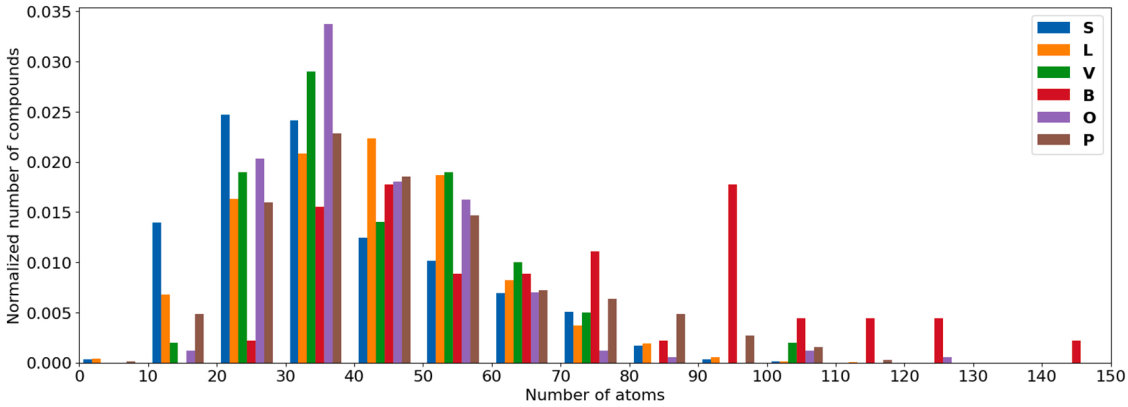

Fig. 2. Normalized distribution of data sets accounted in this study (splits per 10 atoms).

Table 2
Summary of ML methods used.

Abbreviation	Package	Approach	Descriptor
TF	TensorFlow	NN	DScribe/SOAP
XG	XGBoost	GBDT	DScribe/SOAP
SP	SchNetPack	DTNN	SchNet

six interaction blocks to incorporate the local environment of atoms within the atomic representation of molecules. The SchNet prediction block evaluates the atom-wise contributions to the property studied with two available options. In the first one, the contributions are summed up, and the result represents the desired property. This is an ideal approach for the description of extensive physical properties (e.g., atomization energy Jung et al. (2020)) and is later referred to as the 'summed' approach. In the second option, the contributions are averaged and the result is suitable to predict intensive physical properties (e.g., atomization energy per atom Stocker et al. (2020)), this approach is referred to as 'averaged'. The parameters of the DTNN prediction block were unchanged with respect to the defaults provided by SchNetPack, i.e., the underlying neural network consists of two hidden layers with sizes of 128 and 64 neurons, respectively. The output layer was set as a single neuron with the value of the predicted property.

3. Results

3.1. Molecular Docking

The studied compounds and their docking poses were analyzed using two criteria. First and foremost, the free energy of binding (referred to as the docking score) as defined by the AutoDock scoring function was chosen for this purpose. Second, the number of putative hydrogen bonds formed between the compound and $3CL^{pro}$ was considered.

The compounds in our initial **S** data set have calculated free energies of binding (docking score) ranging from -3.35 kcal/mol to -15.06 kcal/mol with a median score of -8.09 kcal/mol Steklac et al. (2021). Compounds from the second **L** data set, together with the validation set **V**, achieved docking scores from -1.69 kcal/mol to -15.96 kcal/mol with a median score of -8.97 kcal/mol. Compounds proposed by the Department of Organic Chemistry at SUT (data set **O**) achieved docking scores ranging from -4.70 kcal/mol to -15.01 kcal/mol with a median score of -9.09 kcal/mol. An improvement in the median docking score and its variance is expected, as the compounds were suggested with the aim of increasing the binding affinity towards SARS-CoV-2 $3CL^{pro}$. The lowest median value of the docking score was observed for the data set **B** with -11.75 kcal/mol. The docking scores for this set ranged from -7.78 kcal/mol to -15.71 kcal/mol. However, this ought to be expected as the **B** data set contains a selection of the best scoring compounds from several publications Wu et al. (2020); Hosseini and Amanlou (2020); Adem et al. (2020); Shah et al. (2020); Fischer et al. (2020).

A two percent of the compounds from our sets (**S**,**L**,**B**,**O**,**V**) reached excellent docking scores below -13 kcal/mol. This holds especially true for the compounds in the **B** data set, where as many as 13 out of 45 compounds achieved this or lower score Steklac et al. (2021). Seven additional compounds (out of 172) proposed in the **O** data set achieved a comparable docking score and can be considered as a basis for further targeted synthesis and subsequent *in vitro* experiments, see Fig. 3.

Hydrogen bond analysis of combined data from all data sets used revealed that the majority of compound-protein complexes are stabilized by six and fewer hydrogen bonds, with only 330 complexes stabilized by more than six hydrogen bonds, see Fig. 4. The most commonly predicted participants in the formation of hydrogen bonds include S1 subside amino acids Glu166 (56.05 % of observed cases), His163 (17.90 %) and Gly143 (15.01 %). Only about 8 % of the examined compounds formed hydrogen bonds with the catalytic dyad His41-Cys145 found in the S2 subside, indicating that the mode of action of most compounds is to block the access to the catalytic dyad rather than its direct inhibition. Other amino acids frequently predicted to participate in the formation of hydrogen bonds include Thr190 (38.38 %), Gln192 (27.07 %), and Asn142 (15.70 %).

Table 3 shows the top five ranked compounds from each separate data set along with their free energies of binding and the amino acids that participate in the formation of hydrogen bonds. The extended version of this table that includes up to 20 compounds from each data set can be found in Table S2. (Note that all docking scores and their predictions are compiled in the csv files of the Supplementary Material zip container.) The putative binding modes of the two top scoring compounds to SARS-CoV-2 $3CL^{pro}$ can be found in Figure S1.

Data sets **S** and **B**: Compounds in the **S** and **B** data sets are part of the previously reported study of Štekláč *et al.* Steklac et al. (2021). However, differences arise between the results presented herein and Štekláč *et al.* Steklac et al. (2021) due to different LGA settings (previously, the total number of docking runs was set to 50 and the number of energy evaluations was 50,000,000). Due to the stochastic nature of AutoDock protocols, certain compounds have achieved different docking scores under different parameters. These differences vary up to 2.5 kcal/mol, which is comparable with the AutoDock scoring function standard error Morris et al. (2009). Nevertheless, the results achieved in these two instances show consistency, as 17 out of the top 20 scoring compounds for each data set remain the same with only slight variations in their order. The interested reader is pointed to the original article Steklac et al. (2021), where the binding affinities as well as the docking poses are thoroughly discussed.

Data sets **L** and **V**: The overall best scoring compounds with the lowest free energy of binding can be found in the most robust data set, in which as many as 40 compounds exhibit $3CL^{pro}$ binding affinities similar

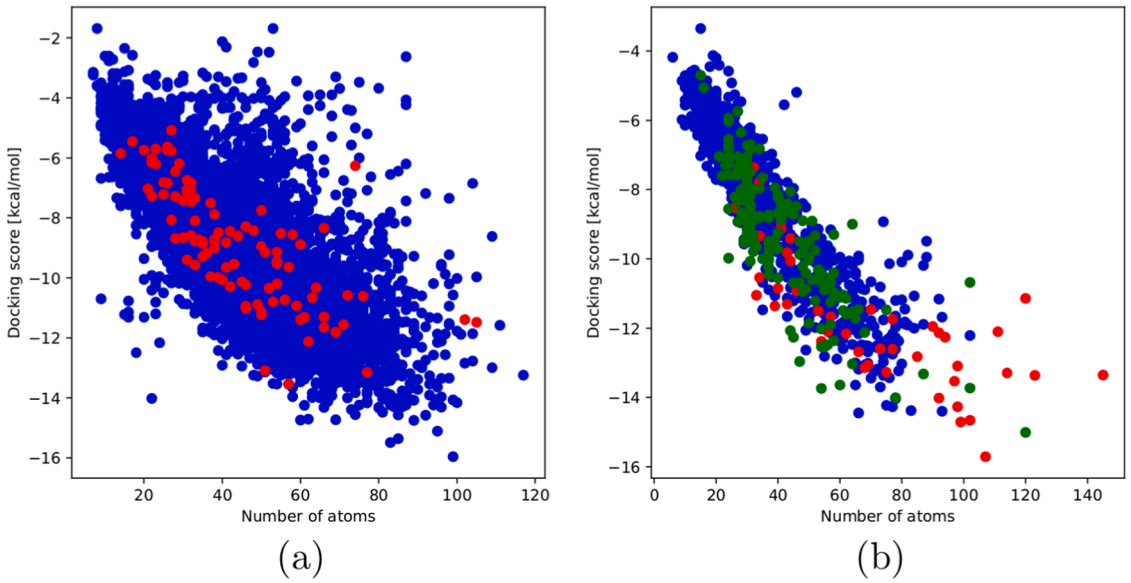

Fig. 3. Docking scores of data sets with respect to the size of the compounds (number of atoms). (a) blue/red - compounds from **L**/**V** set, respectively. (b) blue/red/green - compounds from **S**/**B**/**O** set, respectively.

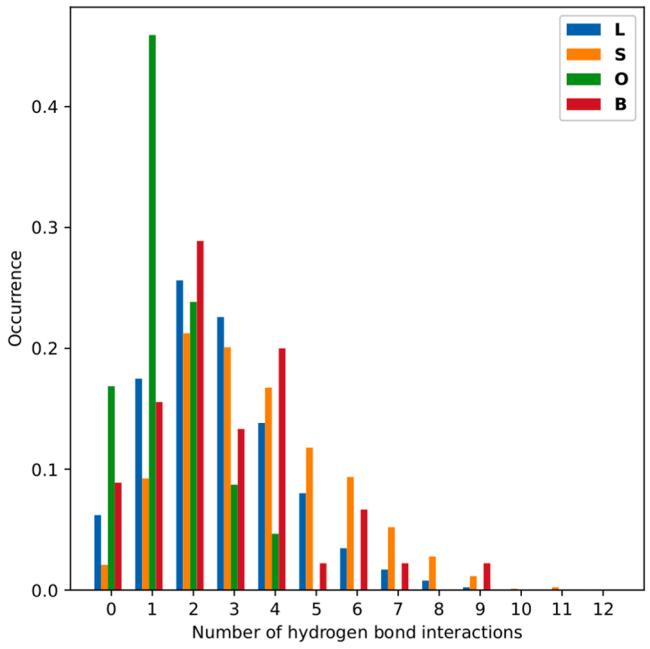

Fig. 4. Frequency of predicted hydrogen bonds formed between the docked compounds and 3CL^{pro} across (**S**, **L**/**V**, **B**, **O**) data sets.

to the top five compounds from other sets (below -14 kcal/mol). Plazomicin (-14.75 kcal/mol) is a broad-spectrum aminoglycoside antibiotic used for the treatment of severe urinary tract infections. Its effect on COVID19 treatment has not yet been explored/reported. A high binding affinity (-15.11 kcal/mol) was also reported in the case of an yet uninvestigated compound (CID: 5289508). Eldecalcitol (-15.36 kcal/mol), a D-vitamin analogue, is used for the treatment of osteoporosis Sanford and McCormack (2011). Its primary effect is the reduction of calcium reabsorption into the body from bone, leading to increased bone mineral density while decreasing the risk of further bone fractures Hatakeyama et al. (2010). An excellent docking score of Eldecalcitol is in agreement with our previously reported study Steklac et al. (2021), in which the pair of D-vitamin analogues calcitriol and calcifediol had shown high potential for the inhibition of SARS-CoV-2 3CL^{pro}. Here one can find a further reason (besides the immunomodulatory role) Chiodini et al. (2021); Bouillon et al. (2018) why D-vitamin is suggested to be administered to patients during the COVID19 treatment. The second highest scoring compound, neladenoson bialanate (-15.49 kcal/mol), was part of separate clinical trials as a potential drug for the treatment of systolic and diastolic heart failures Voors et al. (2019). No repurposing study of any of the four aforementioned compounds (plazomicin, eldecalcitol, compound with CID: 5289508, neladenoson bialanate) was previously published in connection with possible inhibition of SARS-CoV-2 replication through 3CL^{pro} deactivation. Bicotrizole (-15.96 kcal/mol) is a phenolic benzotriazole with broad-spectrum ultraviolet radiation absorbing potential. It has previously been reported to have the highest inhibitory potential against 3CL^{pro} from compounds selected in Jiménez-Alberto *et al.* Jimenez-Alberto et al. (2020), but it has been disregarded from further inspection based on its properties which make it unsuitable for therapeutic use (poor water solubility) Mavon et al. (2007).

Data set **O**: Compounds from the data set proposed by the Department of Organic Chemistry at SUT showed lower binding affinity (less negative docking score) to the SARS-CoV-2 3CL^{pro} structure than compounds from other sets. The notable exception is the compound denoted as O_099, with a docking score of -15.01 kcal/mol. Inspection of the hydrogen bond patterns reveals that the docking score is largely driven by the shape complementarity of this compound with the targeted cavity. From the 20 best scoring compounds of the data set **O**, 13 form H-bonds with the protein structure. The recurring pattern in compounds from the **O** data set is the presence of multiple large substituents on the main molecular skeleton, such as phenyl groups substituted to various degrees and/or adamantyl groups. For instance, the aforementioned compound O_099 contains three benzimidazole groups, each with one adamantyl substituent. Other **O** compounds shown in Table 3 contain at least two aromatic groups (e.g., imidazole, benzene, furane, and benzimidazole rings) that offer various modes of substitution. These observations will be further utilized in the synthesis of compounds with enhanced binding affinity towards the SARS-CoV-2 3CL^{pro} unit.

3.2. Neural networks with TensorFlow and Keras

The TF neural network training is shown in Figures S2a,b. The Mean Square Error (MSE) evaluated during the training process of both sets (**S**

Table 3
Top five scoring compounds from each data set for the $3CL^{pro}$ 6WQF structure with their docking scores (DS) and amino acids participating in predicted H-bond formation.

Data set	CID	Name	DS [kcal/mol]	H-bond forming amino acids
S	118628567	Subasumstat	-14.45	Glu166, Thr190
	53472683	Vazegepant	-14.40	Thr190
	6918155	Ciclesonide	-14.38	Thr26, Glu166, Gln189
	5281040	Montelukast	-14.27	Asn142, Glu166
	5459840	20-Hydroxyecdysone	-14.04	Thr26, Gly143, Glu166, Arg188, Thr190, Gln192
B	25151504	Cobicistat	-15.71	Asn142
	24873435	Simeprevir	-14.71	Ser46, Glu166, Gln189
	3010818	Telaprevir	-14.66	Thr26, **His41**, Asn142, Glu166
	45110509	Paritaprevir	-14.27	Phe140, Gly143, **Cys145**, Gln189
	5362440	Indinavir	-14.02	Glu166
L	3571576	Bisoctrizole	-15.96	Thr26, Asn142, Gly143
	56848985	Neladenoson bialanate	-15.49	Thr26, Asn142, Glu166
	6918141	Eldecalcitol	-15.36	Thr25, Thr26, Cys44, Asn142, Arg188, Thr190, Gln192
	5289508[a]	-	-15.11	Thr26, His163, Gln189
	42613186	Plazomicin	-14.75	His41, Phe140, Glu166, Arg188, Gln189, Thr190, Gln192
O	O_099[a]	-	-15.01	-
	O_104[a]	-	-14.01	Gly143, Gln189
	O_120[a]	-	-13.74	Gln189, Thr190, Gln192
	O_101[a]	-	-13.73	Glu166
	O_095[a]	-	-13.64	Glu166
P'	ZINC000049593065[a]	-	-15.81	Leu141, Asn142, Ser144, His163, Glu166
	ZINC000028472118[a]	-	-15.56	Ser46, Asn142, Gly143, Thr190, Gln192
	ZINC000072190175[a]	-	-15.15	Asn142
	ZINC000049942448[a]	-	-15.11	Leu141, Gly143, Ser144, His163
	ZINC000095535846[a]	-	-14.96	Asn142, His164, Glu166, Gln189

[a] Compound does not have a trivial name

and **L**) shows that the validation set **V** becomes well assessed within 20 NN epochs (orange lines). The training set itself can also be considered well balanced with respect to the docking score prediction within the 20 epochs (blue lines), see Figures S2a,b. The final (predicted vs. expected) correlation diagrams of the **S** and **L** TF training sets are depicted in Figures S3a,b.

The test results for both TF trained NNs are depicted in Fig. 5. Here (see Fig. 5), the **S** trained TF NN is tested against **L**, **B**, and **O** data sets and the **L** trained one is tested against **S**, **B**, and **O** data sets. Furthermore, the linear regression fits (slope, intercept, standard errors, and R^2), mean square error (MSE), and individual maximum absolute error (MAE) outliers are compiled in Table 4. These test results show that both TF trained NNs are capable to predict the docking scores with reasonable accuracy, i.e., the MSE is below 3 kcal/mol and mostly below 1 kcal/mol, see Fig. 5. Thus, taking into account the docking score evaluation error of 2-3 kcal/mol Morris et al. (2009), these results are to be considered reliable. It is actually of no surprise that the best MSE was obtained for the combination of the training set **S** and the test set **S** (MAE outlier of 2.06 kcal/mol was found for the compound with CID: 5291). Compared to the combination of train/test set **L** (MAE outlier of 6.91 kcal/mol was found for the compound with CID: 123966), the MSE for the train/test set **S** is lower by more than a half. The explanation of this result appears rather trivial, the **S** data set is ten times smaller than **L**, and hence the TF NN should be capable to adapt to a smaller set more accurately than to a larger one. On the other hand, the **L** trained TF model is more successful in docking score prediction for the external test sets compared to the **S** model. This is expected because a larger manifold of data allows for a better tuned ML protocol with a superior predictive capacity. This is confirmed with respect to the MSE, MAE outlier, R^2, or slope values, see Table 4. The actual MAE outliers for the **S** trained TF NN are 8.48 (CID: 10945) for test set **L**, 4.99 (CID: 123794) for test set **B**, and 2.75 (compound O_15) for test set **O** (all values are in kcal/mol). In the case of **L** trained TF NN, the MAE outliers are 2.87 (CID: 5459840) for the **S** test set, 3.39 (CID: 123794) for the **B** test set, and 2.59 (compound O_99) for the **O** test set (all values are in kcal/mol).

All linear regression slopes are positive, and their values deviate from one by a small margin (none of the slopes is below 0.75). The intercepts are well defined with respect to the zero value in the ideal case and/or the formal docking score error. Only test set **B** shows slopes above one, and the largest MSE and MAE outlier values (and/or the worst R^2) among the predicted TF NN scores. These prediction results can be attributed to the presence of a non-negligible amount of compounds with a larger number of atoms than those present in training sets **L** and **S**, see Fig. 2. Thus, the ML prediction of docking scores for a given compound (or a set of compounds) correlates strongly with the number of atoms, see below. This can also be seen in the subsequent XG and SP predictions, see below.

3.3. Gradient boosted decision trees with XGBoost

The same **V** data set as used for the validation (training progress) during training of both NN (training sets **S** and **L**) was also employed to control the termination of the XG GBDT parameter optimization process. Optimization of the XG GBDT parameter stopped after 10 iterations during which the RMSE (or the MSE) of the validation set did not improve. The XG MSE of the validation set during parameter optimization is shown in Figures S2c,d (orange line), including the MSE of the actual docking score prediction for the given training set (blue line). The correlation of XG predicted and expected docking scores of both training sets is shown in Figures S3c,d and Table 4.

Prediction results for the **S**, **L**, **B**, and **O** test sets of both training sets are depicted in Figs. 6 and S3c,d. As found for TF, the XG approach is successful in the prediction of the expected docking scores. The MSE and MAE outlier values (including R^2) are reasonable when taking into account the error of the AutoDock scoring function (2-3 kcal/mol) Morris et al. (2009), and comparable with the TF NN performance results. The **L** test set evaluated for the XG **S** training set shall be discussed in more detail as an example, see Fig. 6a. The correlation (represented by a linear fit - red line) is skewed because of the 0.745 slope, hence the predicted docking score values are overestimated compared to the reference

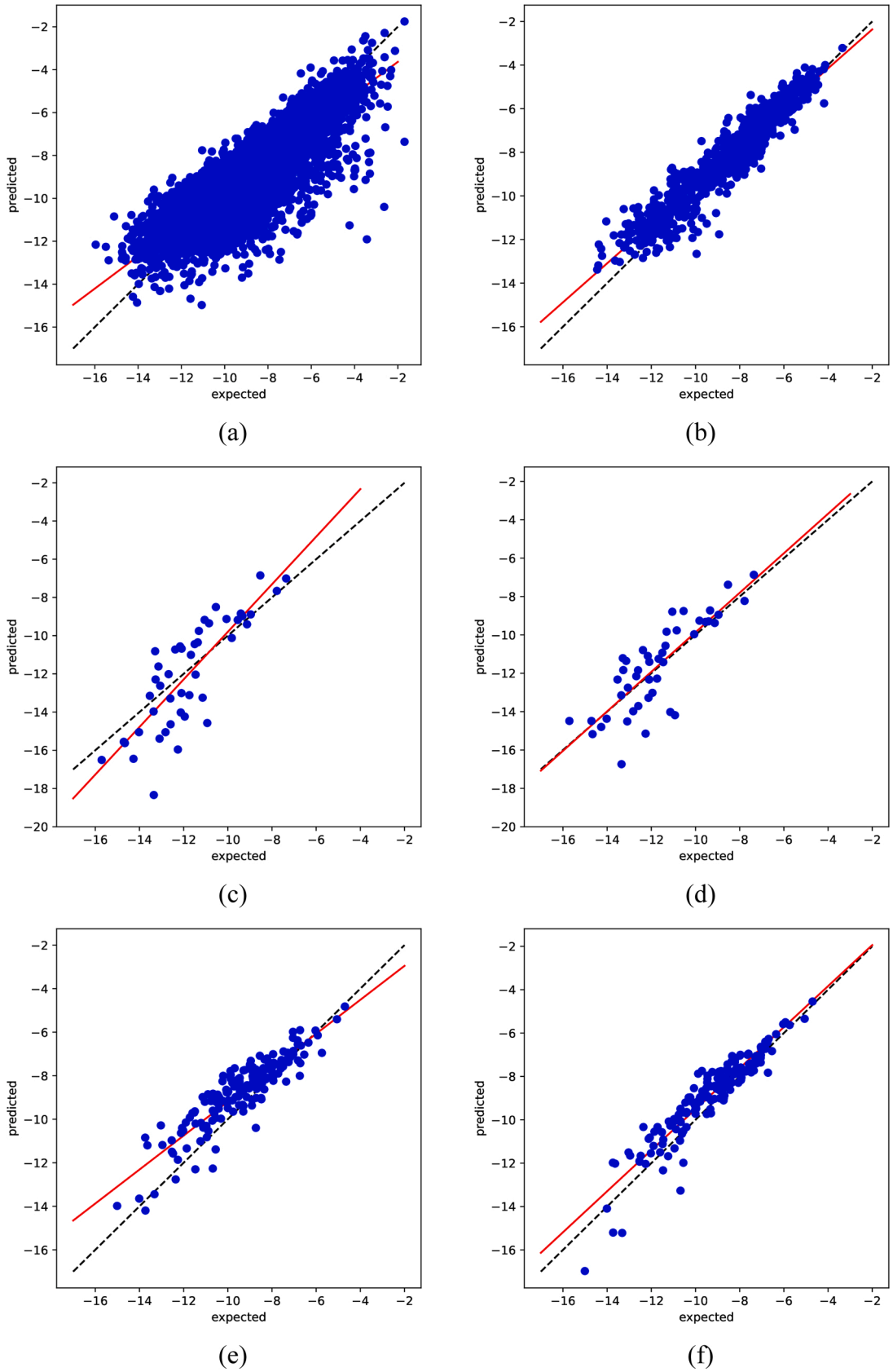

Fig. 5. Correlation between TF predicted and AutoDock calculated (expected) docking scores for: (a) **L** / **S**, (b) **S** / **L**, (c) **B** / **S**, (d) **B** / **L**, (e) **O** / **S**, (f) **O** / **L** (test/training set). The black dashed line represents the ideal correlation between the predicted and expected results and the red solid lines represent linear regression fits between the expected (docking) results and the predicted ones.

Table 4
Machine learning results: slope, intercept, slope stderr (σ_s), intercept stderr (σ_i), R^2, MSE, and MAE outlier.

TF training set	test set	slope	intercept	σ_s	σ_i	R^2	MSE	MAE
S	**S**	0.887	-0.619	0.006	0.055	0.958	0.342	2.058
S	**L**	0.755	-2.126	0.005	0.042	0.757	1.124	8.479
S	**B**	1.245	2.648	0.132	1.576	0.673	2.855	4.989
S	**O**	0.780	-1.394	0.030	0.282	0.803	1.098	2.894
L	**L**	0.837	-1.452	0.004	0.037	0.835	0.761	6.914
L	**S**	0.894	-0.577	0.010	0.087	0.904	0.565	2.870
L	**B**	1.030	0.435	0.109	1.300	0.674	1.774	3.388
L	**O**	0.946	-0.051	0.028	0.268	0.869	0.671	2.587
XG training set	test set	slope	intercept	σ_s	σ_i	R^2	MSE	MAE
S	**S**	0.893	-0.452	0.005	0.045	0.972	0.371	2.326
S	**L**	0.745	-2.198	0.005	0.043	0.749	1.163	7.953
S	**B**	0.605	-3.379	0.064	0.763	0.674	2.740	4.577
S	**O**	0.749	-1.658	0.035	0.329	0.733	1.366	4.823
L	**L**	0.850	-1.331	0.003	0.029	0.896	0.490	3.725
L	**S**	0.866	-0.728	0.010	0.092	0.887	0.713	3.716
L	**B**	0.640	-3.132	0.072	0.853	0.649	2.409	3.591
L	**O**	0.758	-2.037	0.024	0.230	0.852	0.580	2.399
SP training set	test set	slope	intercept	σ_s	σ_i	R^2	MSE	MAE
S	**S**	0.916	-0.378	0.008	0.073	0.933	0.436	2.666
S	**L**	0.816	-1.603	0.005	0.048	0.739	1.240	8.889
S	**B**	1.542	5.262	0.222	2.642	0.528	9.510	10.395
S	**O**	0.828	-1.058	0.037	0.352	0.746	1.183	3.758
L	**L**	0.849	-1.329	0.004	0.038	0.832	0.775	5.781
L	**S**	0.916	-0.346	0.011	0.092	0.897	0.632	3.436
L	**B**	1.143	1.565	0.174	2.067	0.501	4.545	8.044
L	**O**	0.928	-0.486	0.034	0.322	0.816	0.708	4.675
SP training set	test set	slope	intercept	σ_s	σ_i	R^2	MSE	MAE
S 'averaged'	**S**	0.752	-1.682	0.012	0.108	0.813	1.107	3.607
S 'averaged'	**L**	0.650	-2.998	0.005	0.045	0.669	1.549	8.140
S 'averaged'	**B**	0.504	-4.205	0.065	0.776	0.581	4.124	4.718
S 'averaged'	**O**	0.662	-2.622	0.030	0.289	0.736	1.202	4.406
L 'averaged'	**L**	0.863	-1.123	0.004	0.032	0.875	0.587	5.068
L 'averaged'	**S**	0.879	-0.507	0.010	0.092	0.891	0.794	5.114
L 'averaged'	**B**	0.920	-0.054	0.096	1.142	0.681	2.164	4.285
L 'averaged'	**O**	0.787	-1.475	0.024	0.230	0.862	0.746	5.322

docking scores below ca. -9 kcal/mol and underestimated for compounds with the highest inhibition potential found by AutoDock. However, the resulting MSE of 1.163 kcal/mol indicates a reasonable performance when considering the score span of about 14 kcal/mol in this test set, yet certain outliers are clearly visible. The MAE outlier is fairly large, 7.95 kcal/mol (CID: 123966).

The inference of the **S** trained XG predicted scores from the **B** and **O** test sets is shown in Figs. 6c and 6e. The correlation for the **B** test set is clearly the worst, the slope of the linear fit is 0.605, the MSE is 2.740 kcal/mol, and the MAE outlier is 4.58 kcal/mol (compound CP9 from Fischer *et al.* Fischer et al. (2020)), see below and in Table 4. It is worth noting that the **B** test set shows the worst prediction reliability for TF NNs as well. Interestingly, the TF slopes for the **B** test set are larger than one. The XG (**S** training set) results of the **O** test set practically copy those of the **L** test set, the slope value is 0.742, the MSE is 1.301 kcal/mol, and the MAE outlier is of 4.82 kcal/mol (compound O_102) which is comparable to the **B** test set MAE outlier. As already found for TF ML, the better performance of the **O** test set compared to **B** can be attributed to a well-defined composition of the **O** test set with respect to the number of atoms in (the molecular weight of) the compounds. Furthermore, the **O** test set does not show a large variation in the compounds' structure.

The results of the trained XG GBDT with the **L** training set are more optimistic than in the case of the **S** training set (being in close resonance with the TF results). Although the **B** test set improves only slightly in terms of the slope, intercept, MSE, and MAE outlier, the R^2 values become worse, see Table 4 and below. Compared to the results shown in Fig. 6a, Fig. 6b depicts the reversed train-test (**L**-**S**) setup order, which leads to a decrease of the MSE by almost a factor of two (0.713) in the **L** training set XG predictions, including the improvement of the MAE outlier and R^2, see Table 4. There is also a notable improvement of the slope (intercept) of the linear fit, 0.866 (-0.728 kcal/mol), indicating a tendency to slightly underestimate the top scores. The MAE outlier is decreased to 3.72 kcal/mol (CID: 53477854) and the outliers are now biased by a much narrower error bar and distributed more regularly with higher prediction accuracy achieved for the less potent compounds. It has been mentioned (see Fig. 2) that the count of compounds with a lower number of atoms (up to 70) is well accounted for in the training sets and the docking score prediction for such compounds is more accurate than for larger compounds (like those in test set **B**). Additionally, as already stated (subsection 3.1, Docking scores), only 40 compounds of the **L** set exhibit binding affinities (docking scores) below -14 kcal/mol, hence it can be assumed that the desired property will suffer from a certain bias (e.g., XG GBDT tends to underestimate the best scoring compounds). Figs. 6d and 6f show the inference accuracy of **L** training set with respect to the **B** and **O** test sets. No significant improvement of the predictive power of the **L** training set is observed for the **B** test set XG docking scores, when compared to the **S** training set. The MSE decreases from 2.740 to 2.409 kcal/mol, the slope increases from 0.605 to 0.640, and MAE outlier decreases to 3.59 kcal/mol (CID: 25151504), which is similar to the **S** training set (3.72 kcal/mol, CID: 53477854) and still larger than in the case of the **O** test set (2.40 kcal/mol, O_95) of the **L** trained XG GBDT. In contrast, the description of the **O** test set (Fig. 6f) by the **L** trained XG GBDT is superior to the **S** one (Fig. 6e). The MSE decreases to less than half (0.596 kcal/mol), although the actual slope is almost unchanged, 0.745, and the MAE outlier is reduced by a factor of two, respectively, see Table 4. A better prediction of the docking scores of compounds from the **O** test set by the **L** trained XG GBDT (including the improved statistics) is further reflected in the shift of the crossing points between the ideal correlation of expected and predicted correlation lines towards the better scoring compounds.

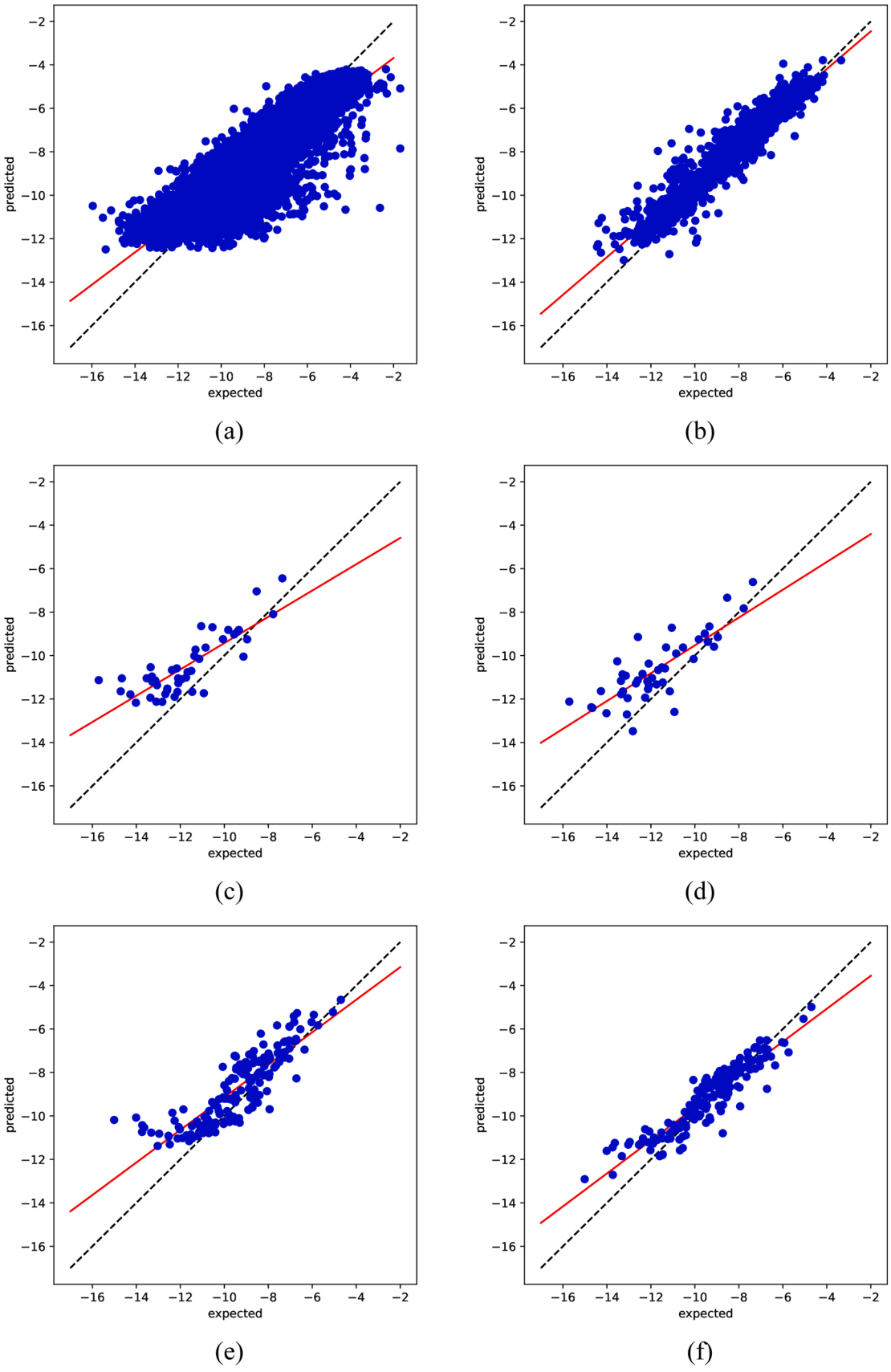

Fig. 6. Correlation between XG predicted and AutoDock calculated (expected) docking scores for: (a) **L** / **S**, (b) **S** / **L**, (c) **B** / **S**, (d) **B** / **L**, (e) **O** / **S**, (f) **O** / **L** (test/training set). The black dashed line represents the ideal correlation between the predicted and expected results and the red solid lines represent linear regression fits between the expected (docking) results and the predicted ones.

3.4. Neural networks with SchNetPack

Two different ways to aggregate the results from the prediction block of SchNetPack (SP) were used, denoted as 'summed' and 'averaged' and employed for both the **S** and **L** training sets. This results in four distinct neural networks, which ought to be considered. Herein, we will focus in more detail on the pictorial representation of the SP 'summed' results given in Fig. 7 and the complementary details about SP 'averaged' results are given in Figure S4. The linear regression fits (as well as the MSE and MAE outlier values) of the training and test sets [**S**, **L**, **B**, **O**] are given in Table 4. The MSE of the SP 'summed' training process is shown in Figures S2e-f. It can be seen that the training lasted for almost 600 and 200 epochs for **S** and **L** sets with respect to a stable MSE value of the validation **V** set (orange line), respectively. The MSE of each training set stabilizes after *ca.* 200 SP 'summed' epochs, see blue lines in Figures S2e-f. The SP 'averaged' validation and training MSEs are similar to the 'summed' approach, and the magnitude of MSE fluctuations in the validation sets and the MSE convergence with the number of epochs are both consistent, see Figures S2g-h.

Based on the results presented in Figs. 7 and S4, the SP approach is successful in predicting the docking scores as found for the previous ML models. As was already the case in the previous two ML techniques, the **S** train set for **S** test set prediction yields lower MSE and the slope (R^2) values are closer to the ideal value compared to the **L** train **L** set set combination, see Table 4 and Figures S3e,f. The MAE outliers of the 'summed' SP approach for the **S** and **L** training sets are 2.67 (CID: 118628567) and 5.78 kcal/mol (CID: 5280899), respectively, and the SP 'averaged' MAE outliers are 3.61 (CID: 5281040) and 5.07 kcal/mol (CID: 9811704), respectively. Nevertheless, the SP predictive power for the other test sets is better for the **L** training set. The MSE values are consistently lower for both test sets (**O** and **B**) in the case of the **L** vs. **S** training set comparison (0.820 and 4.545 kcal/mol vs. 1.002 and 9.510 kcal/mol, respectively). It is fair to note that the 'summed' SP MAE outlier of **O** test set is smaller for the **S** training set (3.76 kcal/mol, O_98) compared to **L** training set (4.68 kcal/mol, O_99). The respective MAE SP 'averaged' outlier values are 4.41 (O_120) for the **S** and 5.32 kcal/mol (O_95) for the **L** training sets. Nonetheless, considering the 2-3 kcal/mol error of the AutoDock scoring function, both SP predicted scores for both training sets provide the verification of a reasonable choice of the training set and neural network model. When comparing the 'summed' and 'averaged' SP results, the 'summed' SP procedure shows a better prediction behavior for the **S**, **L**, and **O** test sets when considering the slope, R^2, MSE and MAE outlier statistics, see Table 4, Figs. 7 and S4. However, this trend becomes reversed for the **B** test set. The **B** test set MSE SP 'averaged' value for the **S** training set is only 4.124 kcal/mol compared to the value of the SP 'summed' approach of 9.510 kcal/mol and only 2.164 kcal/mol vs. 4.545 kcal/mol for the **L** training set. The SP 'summed' MAE outliers (train **S** 10.40 kcal/mol, CID: 5361 and train **L** 8.04 kcal/mol, CID: 5361) are larger than in the 'averaged' scenario (train **S** 4.72 kcal/mol, CID: 25151504 and train **L** 4.29 kcal/mol, CID: 9854073). Interestingly, the slope of the **B** test set is larger than one for the 'summed' SP NN (as is the case of the TF NNs), while the 'averaged' SP NN slope is below one (as is the case in the XG GBDT approach).

4. Discussion

After testing three different **S** and **L** trained machine learning models, it was found that the larger training set **L** leads to a better docking score prediction of external test sets as expected (lower MSE and MAE outlier values and the linear regression parameters are closer to their ideal values). In addition, the **B** test set indicates an issue with compounds whose size (i.e., the number of atoms) notably exceeds the compounds' size present in the training sets (the distribution of the number of atoms in the compounds of test set **B** is shifted to larger values, see Fig. 2). Thus, the results for the **B** test set are the worst (MSE, MAE outlier, and linear regression results). Note that all calculated and predicted docking scores are available in a zip container of the Supplementary Material.

To further explore the prediction behavior of the trained ML models, the results for the production set **P** will be discussed below. The **L** trained TF, XG, and SP predictions of the docking scores for the **P** data set are shown in Figs. 8a,c,e (blue dots) to present the absolute performance of the ML models used. The comparison of the predicted docking scores of the **P** data set (Fig. 8) with the AutoDock scores for the **S**, **L**, **B**, and **O** data sets (Fig. 3) looks very alike. The docking scores tend to decrease monotonically up to compounds with 80 atoms, after which TF and XG predictions plateau at -15.3 (ZINC000096300428) and -13.2 kcal/mol (ZINC000058485982), respectively. In the case of the SP 'summed' model, lower docking score predictions are obtained compared to TF and XG, namely -20.9 kcal/mol (ZINC000072266997), see Figs. 8a,c,e (ZINC000072266997 is actually omitted in Fig. 8e).

To assess the quality of the ML prediction capacity ca. 1200 compounds from the **P** data set were randomly chosen and merged into the **P'** data set, see the Data sets of compounds subsection of the Methods section. The ML prediction of docking scores for the **P'** data set is shown in Figs. 8a,c,e (red dots), including the 20 top ranking compounds of **P'** (black dots) as found by AutoDock, see Table S2. Below, the best Autodock scoring compounds of **P'** will be highlighted first, subsequently, the performance of the ML prediction will be considered together with further options for ML protocol improvements discussed as last.

P' docking scores: Compounds from the data set **P'** have achieved docking scores ranging from -2.68 kcal/mol to -15.81 kcal/mol. The docking score variance should not be surprising, since this data set contains compounds with a regular distribution of the number of atoms (compounds with fewer atoms tend to achieve worse docking scores). It is important to note that there were 27 compounds with a docking score above -1 kcal/mol. The common trait among them is their size (number of atoms), 24 of these compounds have more than 101 atoms, while the remaining three compounds are in the 71-100 atoms number range. Thus, it can be surmised that the unusually small docking score is a result of noncomplementarity of their shape with the protein cavity. Excluding these 27 compounds, which are left out of any further discussion, the median docking score of **P'** was -8.71 kcal/mol. The 20 best scoring compounds from the **P'** data set include 11 compounds tested *in vivo* only, six compounds with human exposure, and three compounds in clinical trials, see Tables 3 and S2. A common pattern among these compounds (7 out of 20) is their observed/predicted activity at Mu-type and Delta-type opioid receptors based on ChEMBL 20 Gaulton et al. (2016). Three out of five top scoring compounds (ZINC000049593065, ZINC000072190175, ZINC000049942448) exhibit this activity. There are six compounds in the selection of the 20 best scoring compounds with neither observed nor predicted activities based on ChEMBL 20. The common identifiers of these structures are the high number of aromatic rings and amine and ketone functional groups, offering various means of hydrogen bond formation. The hydrogen bond pattern for this production set loosely follows the observations made for the data sets discussed in the molecular docking results section, with the identified hydrogen bonds to Asn142, Glu166, Gln189 and Thr190.

P' docking score prediction: The correlation between the expected and predicted docking scores of the **P'** data set shows consistency with up to 60 atoms in a compound, see Fig. 8 and Table S3. The MSE for the docking score predictions for compounds with fewer than 61 atoms are lower than 2 kcal/mol, see Table S3. Still, the presence of outliers with an absolute error above 2 kcal/mol between the predicted and expected docking scores for subsets of compounds with 20-41 and 40-61 atoms leads to worse linear regression slopes. In total, there are 36, 21, and 34 such 2 kcal/mol outliers for compounds with up to 61 atoms in the TF, XG, and SP 'summed' results of data set **P'**, respectively. In addition, only one outlier is found in the SP 'summed' prediction with an absolute error above 4 kcal/mol. In the case of compounds with the number of atoms larger than 60, the MSE values are worse than 4 kcal/mol. Nevertheless, 105, 126, and 103 compounds have the absolute

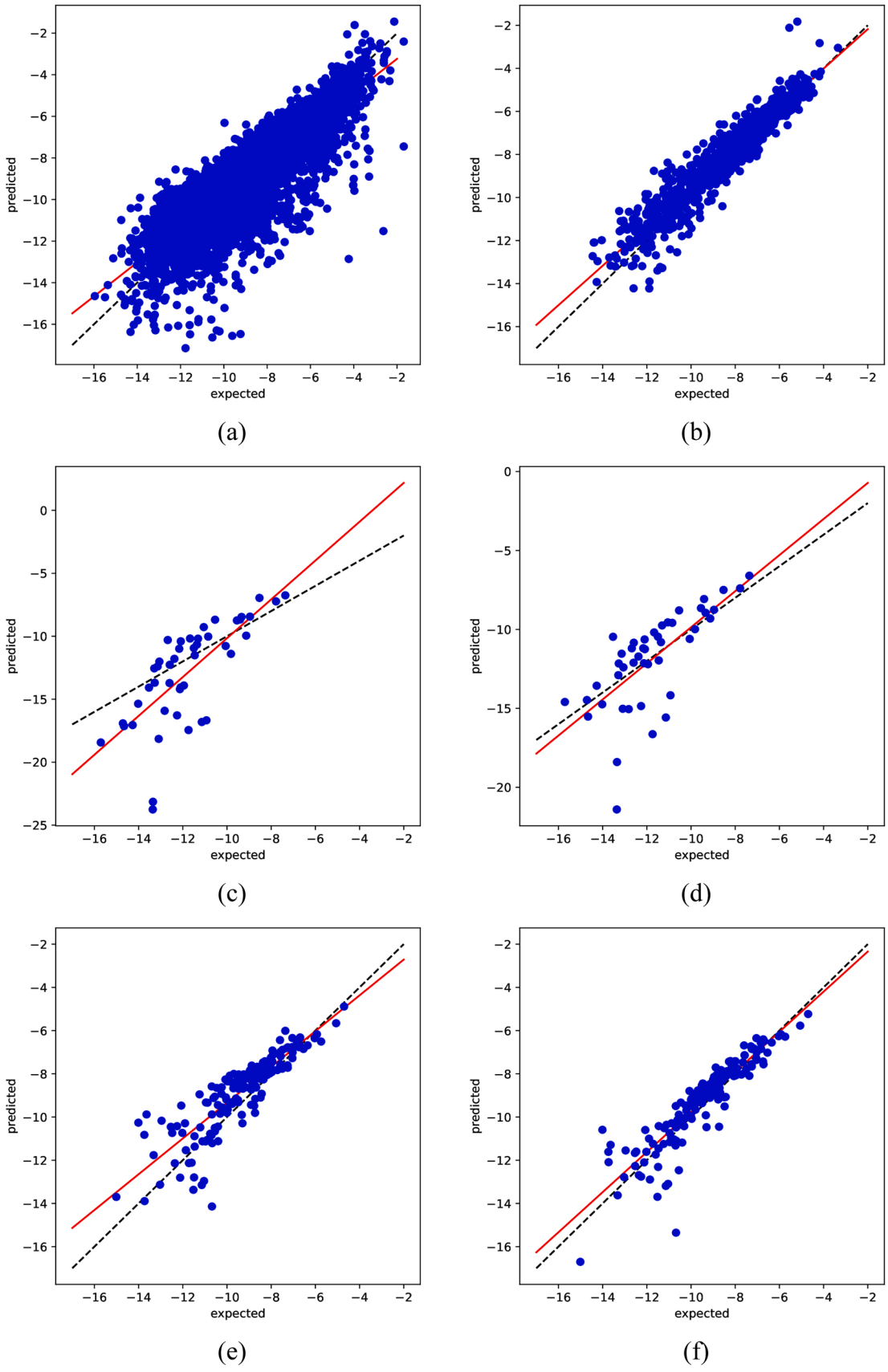

Fig. 7. Correlation between SP 'summed' predicted and AutoDock calculated (expected) docking scores for: (a) **L / S**, (b) **S / L**, (c) **B / S**, (d) **B / L**, (e) **O / S**, (f) **O / L** (test/training set). The black dashed line represents the ideal correlation between the predicted and expected results and the red solid lines represent linear regression fits between the expected (docking) results and the predicted ones.

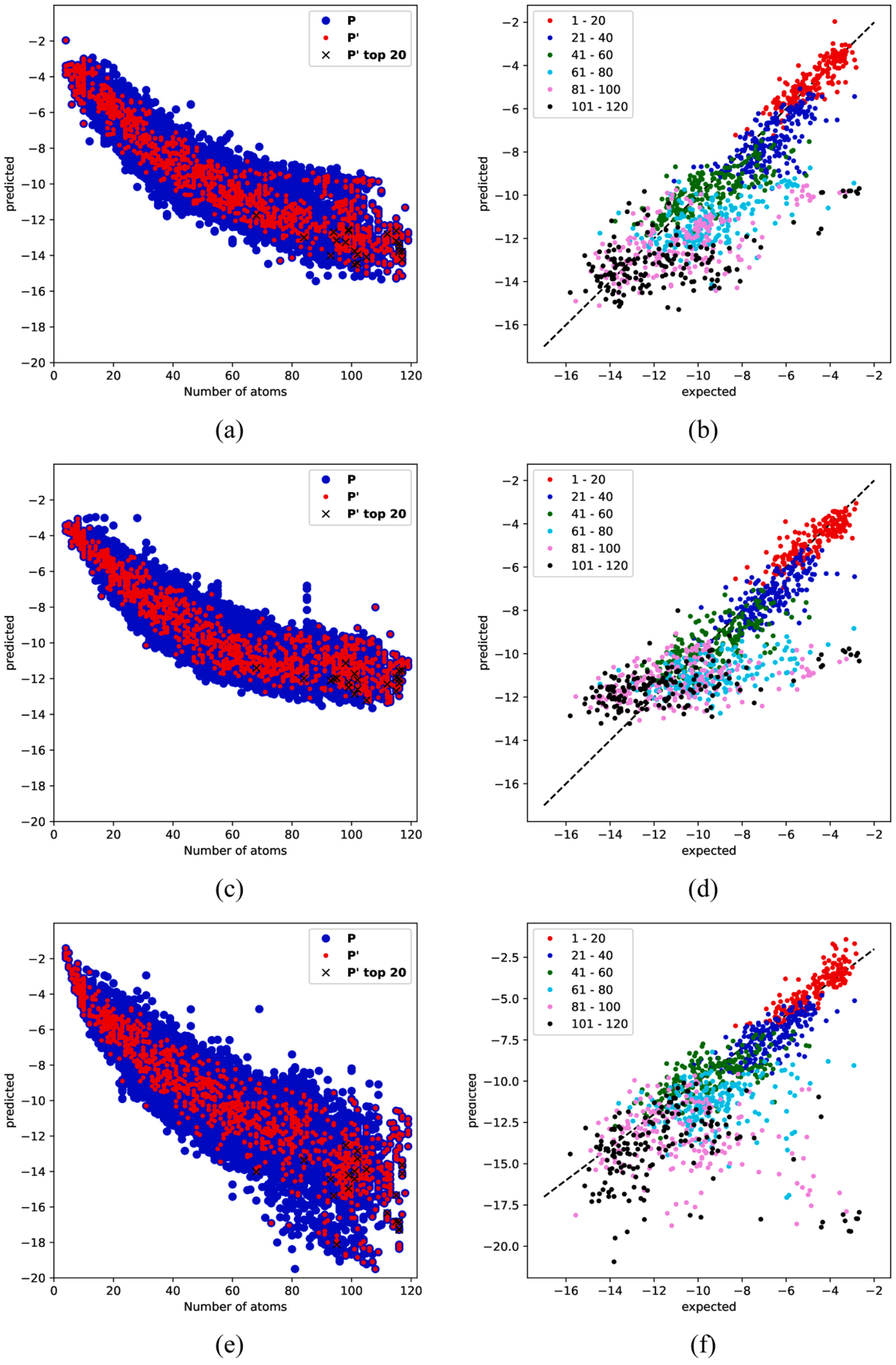

Fig. 8. The performance of ML docking score predictions for set **P** (a) TF; (c) XG; and (e) SP. Correlation of the predicted and expected docking scores of **P'** per 20 atoms batches with the size of the compounds (b) TF; (d) XG; and (f) SP.

prediction error below 2 kcal/mol in the case of TF; 130, 140, and 103 compounds in the case of XG; and 111, 93, and 87 compounds in the case of the SP 'summed' predicted docking scores of 61-80, 81-100, and 101-120 atoms in the subsets of **P'** compounds, respectively. The prediction outliers above 2 kcal/mol tend to be overestimated, see Fig. 8. This means that the ML approaches would suggest these compounds as suitable and their rejection would be based on the subsequent evaluation of the docking score. This also manifests itself in the top 20 AutoDock scoring compounds of **P'** which, albeit shift to compounds with the number of atoms above 80, are being well assessed by the trained ML models, see Figs. 8a,c,e (black dots). Hence, the trained ML models incline towards false positive results for docking score prediction. Still, the essential motivation for ML docking score prediction, i.e., the reduction of the required CPU time and resources by rejection of less suited compounds, is accomplished. When considering the time requirements for the direct docking score evaluation of a single compound, which is 3-5 CPU hours, the ML protocol is orders of magnitude faster, i.e, the descriptor preparation and docking score prediction for hundreds to thousands of compounds happens in a matter of seconds.

It is of no doubt that the inclusion of compounds containing a higher number of atoms within the training sets is desirable to improve the docking score prediction. Essentially, two strategies (and/or hypotheses of the way in which larger molecules are included in the training set) can be considered. First is a training set of molecules with a normal, i.e., Gaussian, distribution tailored towards the larger compounds (up to 150 atoms). The second approach accounts for an exponential distribution with respect to the size of compounds, due to the increase in their structural variability with the number of atoms. The advantage of the first approach is the promise of a well-defined accuracy for the entire size span of the compounds, with the danger that the larger compounds are still not sufficiently represented (overestimation of docking scores). The second approach should be much better suited for the prediction of docking scores of larger molecules, while the lighter compounds (e.g., below 40 or 60 atoms) might be predicted less accurately. Although the protein cavity was not explicitly accounted for within the ML process, the inclusion of compounds that are not able to bind to the active site in the training set is also desirable. Thus, the ML model can comprehend this information, which is very relevant for compounds with more than 100 atoms. This is the case for the **P'** set where 12 compounds with docking scores above -6 kcal/mol that are considerably overestimated in the ML predictions, see Figs. 8b,d,f. In addition, the suitability of the application of an ML model for the prediction of a docking score of a particular compound could be automated on the basis of the evaluation of structural similarity with compounds from the training set. A further step towards a more robust prediction of docking scores employs an iterative (add on what you have) extension of the training set Gentile et al. (2020); Ton (2020) with respect to compounds whose docking scores were available or were not assessed accordingly (i.e., taking into account outliers and inclusion of compounds with up to 120 atoms). The aforementioned points are worth the effort in the forthcoming studies. These should improve the robustness of ML docking score prediction, allowing to extend these ML models to additional protein targets of the virus in similar drug repurposing strategies, and help in the development of ML assisted drug design. Herein, the focus was to explore the performance of different ML models to predict the docking scores of compounds using a direct Cartesian coordinates space (xyz/mol2/sdf/pdb) file format for the generation of ML descriptors.

5. Conclusions

All three ML approaches are successfully trained to predict docking scores. Naturally, the larger training set (**L**) yields better predictions of the docking scores compared to the smaller one (**S**).

It is worth pointing out that the worst ML prediction was obtained for test set **B** that contains compounds with a higher number of atoms than those present in the training sets.

The potential of the trained ML models to predict docking scores has been evaluated for the *in vivo* ZINC data set (**P**). Our ML approaches are suited for an accurate docking score prediction of compounds containing up to 60 atoms (ca. 600 Da). In addition, the ML predicted docking scores of larger compounds are overestimated (i.e., false positive), which means that these compounds are classified as potential candidates for docking score validation when employing these ML processes in a drug repurposing strategy.

Hence, the direct Cartesian coordinates space-based descriptors are proved useful in the ML based docking score prediction. Thus, they are a valid alternative with respect to the single string SMILES representation, descriptors of which are based on molecular fingerprints. In addition, the direct Cartesian coordinates space-based representation of compounds appears to be a relevant choice for the inclusion of the protein cavity effects (direct interaction) within the ML protocol in the future.

CRediT authorship contribution statement

Lukas Bucinsky Investigation, Methodology, Writing original draft, review and editing, Validation, Supervision; **Dušan** Bortňák Investigation, review draft; **Marián** Gall Investigation, Methodology TensorFlow, Writing original draft, review and editing, Validation; **Ján** Matúška Investigation, Methodology SchNetPack, Writing original draft, review and editing, Validation; **Viktor Milata** Investigation, Review draft; **Michal Pitoňák** Investigation, Methodology XGBoost, Writing original draft, review and editing, Validation; **Marek Štekláč** Investigation, Methodology Docking, Writing original draft, review and editing, Validation; **Daniel Végh** Investigation, review draft; **Dávid** Zajaček Investigation, Methodology Docking.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This article was written thanks to the generous support under the Operational Program Integrated Infrastructure for the project: "Strategic research in the field of SMART monitoring, treatment and preventive protection against coronavirus (SARS-CoV-2)", Project no. 313011ASS8, co-financed by the European Regional Development Fund (ERDF). This work was supported by the Science and Technology Assistance Agency under the contract nos. APVV-20-0213, APVV-20-0127, APVV-19-0087, and APVV-17-0513. Further financial support was obtained from the Slovak Grant Agency VEGA under contract nos. 1/0718/19, 1/0777/19, and 1/0139/20. In addition, the Ministry of Education, Science, Research and Sport of the Slovak Republic is acknowledged for the funding within the scheme "Excellent research teams". The authors thank the HPC center at the Slovak University of Technology in Bratislava and the Computing Centre of the Slovak Academy of Sciences, which are part of the Slovak Infrastructure of High Performance Computing (SIVVP) project funded by ERDF (ITMS codes 26210120002 and 26230120002).

Supplementary Material

The following files are available free of charge.

- mmc2.pdf: Tables: XGBoost parameters; Top scoring compounds from each data set; **P'** machine learning results per 20 atom subsets in the size of compounds. Figures: Putative binding modes of the two best scoring compounds from each data set; Validation for **S** and **L** training process; Results for **S** and **L** training process; The **S** and **L** training set of the 'averaged' SchNet architecture performance.

- csv files with docking scores and/or their predictions for all sets; csv file with SMILES codes and overlap with PubChem and ZINC databases for the **O** set compounds; pdf file (mmc3.pdf) with docking scores and structural formulas for the **O** set compounds.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.compbiolchem.2022.107656.

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X., 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, software available from tensorflow.org. ⟨https://www.tensorflow.org/⟩.

Acharya, A., Agarwal, R., Baker, M.B., Baudry, J., Bhowmik, D., Boehm, S., Byler, K.G., Chen, S.Y., Coates, L., Cooper, C.J., Demerdash, O., Daidone, I., Eblen, J.D., Ellingson, S., Forli, S., Glaser, J., Gumbart, J.C., Gunnels, J., Hernandez, O., Irle, S., Kneller, D.W., Kovalevsky, A., Larkin, J., Lawrence, T.J., LeGrand, S., Liu, S.-H., Mitchell, J.C., Park, G., Parks, J.M., Pavlova, A., Petridis, L., Poole, D., Pouchard, L., Ramanathan, A., Rogers, D.M., Santos-Martins, D., Scheinberg, A., Sedova, A., Shen, Y., Smith, J.C., Smith, M.D., Soto, C., Tsaris, A., Thavappiragasam, M., Tillack, A.F., Vermaas, J.V., Vuong, V.Q., Yin, J., Yoo, S., Zahran, M., Zanetti-Polzi, L., 2020. Supercomputer-Based Ensemble Docking Drug Discovery Pipeline with Application to Covid-19. J. Chem. Inf. Model. 60 (12), 5832–5852. https://doi.org/10.1021/acs.jcim.0c01010.

Adem, S., Eyupoglu, V., Sarfraz, I., Rasul, A., Ali, M., 2020. Identification of Potent COVID-19 Main Protease (Mpro) Inhibitors from Natural Polyphenols: An in Silico Strategy Unveils a Hope against CORONA., Preprints, 2020030333. https://doi.org/10.20944/:http://arXiv.org/abs/s202003.0333.v1.

Bartók, A.P., Kondor, R., Csányi, G., 2013a. On representing chemical environments. Phys. Rev. B 87, 184115. https://doi.org/10.1103/PhysRevB.87.184115. ⟨https://link.aps.org/doi/10.1103/PhysRevB.87.184115⟩.

Bartók, A.P., Kondor, R., Csányi, G., 2013b. Publisher's Note: On representing chemical environments [Phys. Rev. B 87, 184115 (2013b)]. Phys. Rev. B 87, 219902. https://doi.org/10.1103/PhysRevB.87.219902. ⟨https://link.aps.org/doi/10.1103/PhysRevB.87.219902⟩.

Bartók, A.P., Kondor, R., Csányi, G., 2017. Erratum: On representing chemical environments [Phys. Rev. B 87, 184115 (2013c)]. Phys. Rev. B 96, 019902. https://doi.org/10.1103/PhysRevB.96.019902. ⟨https://link.aps.org/doi/10.1103/PhysRevB.96.019902⟩.

Batra, R., Chan, H., Kamath, G., Ramprasad, R., Cherukara, M.J., Sankaranarayanan, S.K. R.S., 2020. Screening of Therapeutic Agents for COVID-19 Using Machine Learning and Ensemble Docking Studies. J. Phys. Chem. Lett. 11 (17), 7058–7065. https://doi.org/10.1021/acs.jpclett.0c02278.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The Protein Data Bank. arXiv Nucleic Acids Res 28 (1), 235–242. https://doi.org/10.1093/nar/28.1.235. arXiv. ⟨https://academic.oup.com/nar/article-pdf/28/1/235/9895144/280235.pdf⟩.

Bernal, J.L., Andrews, N., Gower, C., Gallagher, E., Simmons, R., Thelwall, S., Stowe, J., Tessier, E., Groves, N., Dabrera, G., Myers, R., Campbell, C.N.J., Amirthalingam, G., Edmunds, M., Zambon, M., Brown, K.E., Hopkins, S., Chand, M., Ramsay, M., 2021. Effectiveness of Covid-19 Vaccines against the B.1.617.2 (Delta) Variant. N. Engl. J. Med. 385 (7), 585–594. https://doi.org/10.1056/NEJMoa2108891.

Bogoch, I.I., Watts, A., Thomas-Bachli, A., Huber, C., Kraemer, M.U.G., Khan, K., 2020. Pneumonia of unknown aetiology in Wuhan, China: potential for international spread via commercial air travel, 01 J. Travel Med. 27 (2), taaa008. https://doi.org/10.1093/jtm/taaa008, 01. ⟨https://academic.oup.com/jtm/article-pdf/27/2/taaa008/32902439/taaa008.pdf⟩.

Bortnak, D., Milata, V., Sofranko, J., Vegh, D., Fronc, M., Herich, P., Kozisek, J., Hrivnakova, V., Soral, M., 2018. On the formation of uncommon pyrazoloazepines from 5-aminopyrazoles as by-products in the Clauson-Kaas reaction. JOURNAL OF MOLECULAR STRUCTURE 1166, 243–251. https://doi.org/10.1016/j.molstruc.2018.04.034.

Bouillon, R., Marcocci, C., Carmeliet, G., Bikle, D., White, J.H., Dawson-Hughes, B., Lips, P., Munns, C.F., Lazaretti-Castro, M., Giustina, A., Bilezikian, J., 2018. Skeletal and Extraskeletal Actions of Vitamin D: Current Evidence and Outstanding Questions. arXiv Endocrine Reviews 40 (4), 1109–1151. https://doi.org/10.1210/er.2018-00126. arXiv. ⟨https://academic.oup.com/edrv/article-pdf/40/4/1109/28933998/er.2018-00126.pdf⟩.

Buhner, S.H., 2013. Herbal Antivirals: Natural Remedies for Emerging & Resistant Viral Infections. Storey Publishing LLC, North Adams, MA.

Casbarra, L., Procacci, P., 2021. Binding free energy predictions in host-guest systems using Autodock4. A retrospective analysis on SAMPL6, SAMPL7 and SAMPL8 challenges. J. Comput.-Aided Mol. Des 35 (6), 721–729. https://doi.org/10.1007/s10822-021-00388-4.

Chen, T., Guestrin, C., 2016. XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, ACM, New York, NY, USA 785–794. https://doi.org/10.1145/2939672.2939785.

Chiodini, I., Gatti, D., Soranna, D., Merlotti, D., Mingiano, C., Fassio, A., Adami, G., Falchetti, A., Eller-Vainicher, C., Rossini, M., Persani, L., Zambon, A., Gennari, L., 2021. Vitamin d status and sars-cov-2 infection and covid-19 clinical outcomes. Frontiers in Public Health 9. https://doi.org/10.3389/fpubh.2021.736665. ⟨https://www.frontiersin.org/article/10.3389/fpubh.2021.736665⟩.

Cho, E., Rosa, M., Anjum, R., Mehmood, S., Soban, M., Mujtaba, M., Bux, K., Moin, S.T., Tanweer, M., Dantu, S., Pandini, A., Yin, J., Ma, H., Ramanathan, A., Islam, B., Mey, A.S.J.S., Bhowmik, D., Haider, S., 2021. Dynamic Profiling of beta-Coronavirus 3CL M-pro Protease Ligand-Binding Sites. J. Chem. Inf. Model. 61 (6), 3058–3073. https://doi.org/10.1021/acs.jcim.1c00449.

Chollet, F., 2015. Keras. https://keras.io.

Colalto, C., 2020. Volatile molecules for COVID-19: A possible pharmacological strategy? Drug Dev. Res 81, 950–968. https://doi.org/10.1002/ddr.21716.

Das, S., Sarmah, S., Lyndem, S., Roy, A.S., 2021. An investigation into the identification of potential inhibitors of SARS-CoV-2 main protease using molecular docking study. pMID: 32362245. arXiv J. Biomol. Struct. Dyn. 39 (9), 3347–3357. https://doi.org/10.1080/07391102.2020.1763201. pMID: 32362245. arXiv. ⟨https://doi.org/10.1080/07391102.2020.1763201⟩.

De Clercq, E. (Ed.), 2011. Antiviral drug Strategies, 50. Wiley WH-Verlag and CO. KGaA, Weinheim Germany.

Dong, E., Du, H., Gardner, L., 2020. An interactive web-based dashboard to track COVID-19 in real time, Lancet Inf. Dis 20 (5), 533–534. https://doi.org/10.1016/S1473-3099(20)30120-1.

El-Behery, H., Attia, A.-F., El-Fishawy, N., Torkey, H., 2021. Efficient machine learning model for predicting drug-target interactions with case study for Covid-19. Comput. Biol. Chem. 93, 107536 https://doi.org/10.1016/j.compbiolchem.2021.107536. ⟨https://www.sciencedirect.com/science/article/pii/S1476927121001031⟩.

Elfiky, A.A., 2020. Ribavirin, Remdesivir, Sofosbuvir, Galidesivir, and Tenofovir against SARS-CoV-2 RNA dependent RNA polymerase (RdRp): A molecular docking study. Life Sci. 253, 117592 https://doi.org/10.1016/j.lfs.2020.117592. ⟨https://www.sciencedirect.com/science/article/pii/S0024320520303404⟩.

Elmezayen, A.D., Al-Obaidi, A., Şahin, A.T., Yelekçi, K., 2021. Drug repurposing for coronavirus (COVID-19): in silico screening of known drugs against coronavirus 3CL hydrolase and protease enzymes. pMID: 32306862. arXiv J. Biomol. Struct. Dyn. 39 (8), 2980–2992. https://doi.org/10.1080/07391102.2020.1758791. pMID: 32306862. arXiv. ⟨https://doi.org/10.1080/07391102.2020.1758791⟩.

Fischer, A., Sellner, M., Neranjan, S., Smieško, M., Lill, M.A., 2020. Potential Inhibitors for Novel Coronavirus Protease Identified by Virtual Screening of 606 Million Compounds. Int. J. Mol. Sci. 21 (10) https://doi.org/10.3390/ijms21103626. ⟨https://www.mdpi.com/1422-0067/21/10/3626⟩.

Gaulton, A., Hersey, A., Nowotka, M., Bento, A.P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L.J., Cibrián-Uhalte, E., Davies, M., Dedman, N., Karlsson, A., Magariños, M.P., Overington, J.P., Papadatos, G., Smit, I., Leach, A.R., 2016. The ChEMBL database in 2017. arXiv Nucleic Acids Res. 45 (D1), D945–D954. https://doi.org/10.1093/nar/gkw1074. arXiv. ⟨https://academic.oup.com/nar/article-pdf/45/D1/D945/8846762/gkw1074.pdf⟩.

Gentile, F., Agrawal, V., Hsing, M., Ton, A.-T., Ban, F., Norinder, U., Gleave, M.E., Cherkasov, A., 2020. Deep Docking: A Deep Learning Platform for Augmentation of Structure Based Drug Discovery. arXiv ACS Cent. Sci 6 (6), 939–949. https://doi.org/10.1021/acscentsci.0c00229. arXiv. ⟨https://doi.org/10.1021/acscentsci.0c00229⟩.

Guedes, I.A., Costa, L.S.C., dosSantos, K.B., Karl, A.L.M., Rocha, G.K., Teixeira, I.M., Galheigo, M.M., Medeiros, V., Krempser, E., Custodio, F.L., Barbosa, H.J.C., Nicolas, M.F., Dardenne, L.E., 2021. Drug design and repurposing with DockThor-VS web server focusing on SARS-CoV-2 therapeutic targets and their non-synonym variants. MAR 10 Sci. Rep. 11 (1). https://doi.org/10.1038/s41598-021-84700-0.

Halgren, T., 1996a. Merck molecular force field. 1. Basis, form, scope, parameterization, and performance of MMFF94. J. Comput. Chem. 17 (5-6), 490–519 doi:10.1002/(SICI)1096-987X(199604)17:6<490::AID-JCC1>3.3.CO;2-V.

Halgren, T., 1996b. Merck molecular force field. 2. MMFF94 van der Waals and electrostatic parameters for intermolecular interactions. J. Comput. Chem. 17 (5-6), 520–552 doi:10.1002/(SICI)1096-987X(199604)17:6<520::AID-JCC2>3.3.CO;2-W.

Halgren, T., 1996c. Merck molecular force field. 3. Molecular geometries and vibrational frequencies for MMFF94. J. Comput. Chem. 17 (5-6), 553–586.

Halgren, T., 1996d. Merck molecular force field. 5. Extension of MMFF94 using experimental data, additional computational data, and empirical rules. J. Comput. Chem. 17 (5-6), 616–641 doi:10.1002/(SICI)1096-987X(199604)17:6<616::AID-JCC5>3.3.CO;2-3.

Halgren, T., Nachbar, R., 1996. Merck molecular force field. 4. Conformational energies and geometries for MMFF94. J. Comput. Chem. 17 (5-6), 587–615 doi:10.1002/(SICI)1096-987X(199604)17:6<587::AID-JCC4>3.0.CO;2-P.

Hall, D.C., Ji, H.-F., 2020. A search for medications to treat COVID-19 via in silico molecular docking models of the SARS-CoV-2 spike glycoprotein and 3CL protease, Travel. Med. Infect. Dis 35, 101646. https://doi.org/10.1016/j.tmaid.2020.101646.

Hatakeyama, S., Yoshino, M., Eto, K., Takahashi, K., Ishihara, J., Ono, Y., Saito, H., Kubodera, N., 2010. Synthesis and preliminary biological evaluation of 20-epi-eldecalcitol [20-epi-1α,25-dihydroxy-2β -(3-hydroxypropoxy)vitamin D3: 20-epi-ED-71]. J. Steroid Biochem. Mol. Biol. 121 (1), 25–28. https://doi.org/10.1016/j.jsbmb.2010.03.041. ⟨https://www.sciencedirect.com/science/article/pii/S0960076010001329⟩.

Himanen, L., Jäger, M.O., Morooka, E.V., Federici Canova, F., Ranawat, Y.S., Gao, D.Z., Rinke, P., Foster, A.S., 2020. DScribe: Library of descriptors for machine learning in

materials science. Comp. Phys. Commun. 247, 106949 https://doi.org/10.1016/j.cpc.2019.106949. ⟨https://www.sciencedirect.com/science/article/pii/S0010465519303042⟩.
Hosseini, F.S., Amanlou, M., 2020. Anti-HCV and anti-malaria agent, potential candidates to repurpose for coronavirus infection: Virtual screening, molecular docking, and molecular dynamics simulation study. Life Sci. 258, 118205 https://doi.org/10.1016/j.lfs.2020.118205. ⟨https://www.sciencedirect.com/science/article/pii/S0024320520309577⟩.
Irwin, J., Sterling, T., Mysinger, M., Bolstad, E., Coleman, R., ZINC, 2012. A Free Tool to Discover Chemistry for Biology. J. Chem. Inf. Model. 52, 1757–1768. https://doi.org/10.1021/ci3001277.
Islam, H., Rahman, A., Masud, J., Shweta, D., Araf, Y., Ullah, M., AlSium, S., Sarkar, B., 2020. A generalized overview of SARS-COV-2: Where does the current knowledge stand? Dec. Electron. 17 (6) https://doi.org/10.29333/ejgm/8258.
JHU CSSE COVID-19. 2020}, howpublished=(https://github.com/CSSEGISandData/COVID-19) (accessed October 07 2021).
Jiménez-Alberto, A., Ribas-Aparicio, R.M., Aparicio-Ozores, G., Castelán-Vega, J.A., 2020. Virtual screening of approved drugs as potential SARS-CoV-2 main protease inhibitors. Comput. Biol. Chem. 88, 107325 https://doi.org/10.1016/j.compbiolchem.2020.107325. ⟨https://www.sciencedirect.com/science/article/pii/S1476927120304813⟩.
Jiménez-Alberto, A., Ribas-Aparicio, R.M., Aparicio-Ozores, G., Castelán-Vega, J.A., 2020. Virtual screening of approved drugs as potential SARS-CoV-2 main protease inhibitors. Comput. Biol. Chem. 88, 107325 https://doi.org/10.1016/j.compbiolchem.2020.107325 https://pubmed.ncbi.nlm.nih.gov/32623357. ⟨https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7316061/⟩.
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., 2020. Structure of M-pro from SARS-CoV-2 and discovery of its inhibitors. Nature 582 (7811), 289–293. https://doi.org/10.1038/s41586-020-2223-y.
Joshi, T., Joshi, T., Pundir, H., Sharma, P., Mathpal, S., Chandra, S., 2020b. Predictive modeling by deep learning, virtual screening and molecular dynamics study of natural compounds against SARS-CoV-2 main protease. pMID: 32752947. arXiv J. Biomol. Struct. Dyn. 0 (0), 1–9. https://doi.org/10.1080/07391102.2020.1802341. pMID: 32752947. arXiv. ⟨https://doi.org/10.1080/07391102.2020.1802341⟩.
Joshi, T., Joshi, T., Sharma, P., Mathpal, S., Pundir, H., Bhatt, V., Chandra, S., 2020a. In silico screening of natural compounds against COVID-19 by targeting Mpro and ACE2 using molecular docking. Eur. Rev. Med. Pharmacol. Sci. 24 (8), 4529–4536. https://doi.org/10.26355/eurrev_202004_21036.
Jung, H., Stocker, S., Kunkel, C., Oberhofer, H., Han, B., Reuter, K., Margraf, J.T., 2020. Size-extensive molecular machine learning with global representations. arXiv Chem. Systems Chem. 2 (4), e1900052. https://doi.org/10.1002/syst.201900052. arXiv. ⟨https://chemistry-europe.onlinelibrary.wiley.com/doi/pdf/10.1002/syst.201900052⟩.
Karki, N., Verma, N., Trozzi, F., Tao, P., Kraka, E., Zoltowski, B., 2021a. Predicting Potential SARS-COV-2 Drugs–In Depth Drug Database Screening Using Deep Neural Network Framework SSnet, Classical Virtual Screening and Docking. Int. J. Mol. Sci. 22 (4) https://doi.org/10.3390/ijms22041573. ⟨https://www.mdpi.com/1422-0067/22/4/1573⟩.
Khan, M.F., Alam, M.M., Verma, G., Akhtar, W., Akhter, M., Shaquiquzzaman, M., 2016. The therapeutic voyage of pyrazole and its analogs: A review. Eur. J. Med. Chem. 120, 170–201. https://doi.org/10.1016/j.ejmech.2016.04.077. ⟨https://www.sciencedirect.com/science/article/pii/S0223523416303956⟩.
Khan, S.A., Zia, K., Ashraf, S., Uddin, R., Ul-Haq, Z., 2021. Identification of chymotrypsin-like protease inhibitors of SARS-CoV-2 via integrated computational approach. pMID: 32238094. arXiv J. Biomol. Struct. Dyn. 39 (7), 2607–2616. https://doi.org/10.1080/07391102.2020.1751298. pMID: 32238094. arXiv. ⟨https://doi.org/10.1080/07391102.2020.1751298⟩.
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., 2018. PubChem 2019 update: improved access to chemical data. arXiv Nucleic Acids Res. 47 (D1), D1102–D1109. https://doi.org/10.1093/nar/gky1033. arXiv. ⟨https://academic.oup.com/nar/article-pdf/47/D1/D1102/27437306/gky1033.pdf⟩.
Kingma, D.P., Ba, J., 2015. Adam: A Method for Stochastic Optimization. CoRR abs/1412.6980.
Kneller, D., Phillips, G., O'Neill, H., Jedrzejczak, R., Stols, L., Langan, P., Joachimiak, A., Coates, L., Kovalevsky, A., 2020. Structural plasticity of SARS-CoV-2 3CL Mpro active site cavity revealed by room temperature X-ray crystallograph. Nat. Commun. 11, 3202. https://doi.org/10.1038/s41467-020-16954-7.
Kong, R., Yang, G., Xue, R., Liu, M., Wang, F., Hu, J., Guo, X., Chang, S., 2020. COVID-19 Docking Server: a meta server for docking small molecules, peptides and antibodies against potential targets of COVID-19. arXiv Bioinformatics 36 (20), 5109–5111. https://doi.org/10.1093/bioinformatics/btaa645. arXiv. ⟨https://academic.oup.com/bioinformatics/article-pdf/36/20/5109/35065253/btaa645.pdf⟩.
Li, H., Liu, S.-M., Yu, X.-H., Tang, S.-L., Tang, C.-K., 2020b. Coronavirus disease 2019 (COVID-19): current status and future perspectives. Int. J. Antimicrob. Agents 55 (5), 105951. https://doi.org/10.1016/j.ijantimicag.2020.105951.
Li, H., Peng, J., Sidorov, P., Leung, Y., Leung, K.-S., Wong, M.-H., Lu, G., Ballester, P.J., 2019. Classical scoring functions for docking are unable to exploit large volumes of structural and interaction data. Bioinformatics 35 (20), 3989–3995. https://doi.org/10.1093/bioinformatics/btz183.
Li, Q., Kang, C., 2020. Progress in Developing Inhibitors of SARS-CoV-2 3C-Like Protease. aug Microorganisms 8 (8). https://doi.org/10.3390/microorganisms8081250.
Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., Ren, R., Leung, K.S., Lau, E.H., Wong, J.Y., Xing, X., Xiang, N., Wu, Y., Li, C., Chen, Q., Li, D., Liu, T., Zhao, J., Liu, M., Tu, W., Chen, C., Jin, L., Yang, R., Wang, Q., Zhou, S., Wang, R., Liu, H., Luo, Y., Liu, Y., Shao, G., Li, H., Tao, Z., Yang, Y., Deng, Z., Liu, B., Ma, Z., Zhang, Y., Shi, G., Lam, T.T., Wu, J.T., Gao, G.F., Cowling, B.J., Yang, B., Leung, G.M., Feng, Z., 2020a. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. pMID: 31995857. arXiv N. Engl. J. Med. 382 (13), 1199–1207. https://doi.org/10.1056/NEJMoa2001316. pMID: 31995857. arXiv. ⟨https://doi.org/10.1056/NEJMoa2001316⟩.
Liu, C., Zhou, Q., Li, Y., Garner, L., Watkins, V, S.P., Carter, L.J., Smoot, J., Gregg, A.C., Daniels, A.D., Jervey, S., Albaiu, D., 2020. Research and Development on Therapeutic Agents and Vaccines for COVID-19 and Related Human Coronavirus Diseases. ACS Cent. Sci. 6 (3), 315–331. https://doi.org/10.1021/acscentsci.0c00272.
Llanos, M.A., Gantner, M.E., Rodriguez, S., Alberca, L.N., Bellera, C.L., Talevi, A., Gavernet, L., 2021. Strengths and Weaknesses of Docking Simulations in the SARS-CoV-2 Era: the Main Protease (Mpro) Case Study. J. Chem. Inf. Model. 61 (8), 3758–3770. https://doi.org/10.1021/acs.jcim.1c00404.
Lu, I.-L., Mahindroo, N., Liang, P.-H., Peng, Y.-H., Kuo, C.-J., Tsai, K.-C., Hsieh, H.-P., Chao, Y.-S., Wu, S.-Y., 2006. Structure-Based Drug Design and Structural Biology Study of Novel Nonpeptide Inhibitors of Severe Acute Respiratory Syndrome Coronavirus Main Protease. pMID: 16913704. arXiv J. Med. Chem. 49 (17), 5154–5161. https://doi.org/10.1021/jm060207o. pMID: 16913704. arXiv. ⟨https://doi.org/10.1021/jm060207o⟩.
Lu, J., Hou, X., Wang, C., Zhang, Y., 2019. Incorporating Explicit Water Molecules and Ligand Conformation Stability in Machine-Learning Scoring Functions. J. Chem. Inf. Model. 59 (11), 4540–4549. https://doi.org/10.1021/acs.jcim.9b00645.
Mahase, E., 2021. Covid-19: How many variants are there, and what do we know about them? bmj BM J-British Medical Journal 374. https://doi.org/10.1136/bmj.n1971.
Mallah, S., Ghorab, I, O.K., Al-Salmi, S., Abdellatif, O.S., Tharmaratnam, T., Iskandar, M. A., Sefen, J.A.N., Sidhu, P., Atallah, B., El-Lababidi, R., Al-Qahtani, M., 2021. COVID-19: breaking down a global health crisis. Ann. clin. microbiol. 20 (1), 35. https://doi.org/10.1186/s12941-021-00438-7.
Mathieu, E., Ritchie, H., Ortiz-Ospina, E., Roser, M., Hasell, J., Appel, C., Giattino, C., Rodes-Guirao, L., 2021. A global database of COVID-19 vaccinations. Nat. Hum. Behav 5 (7), 947–953. https://doi.org/10.1038/s41562-021-01122-8.
Mavon, A., Miquel, C., Lejeune, O., Payre, B., Moretto, P., 2007. In vitro Percutaneous Absorption and in vivo Stratum Corneum Distribution of an Organic and a Mineral Sunscreen. Skin Pharmacol. Physiol. 20 (1), 10–20. https://doi.org/10.1159/000096167. ⟨https://www.karger.com/DOI/10.1159/000096167⟩.
Meyer-Almes, F.-J., 2020. Repurposing approved drugs as potential inhibitors of 3CL-protease of SARS-CoV-2: Virtual screening and structure based drug design. Comput. Biol. Chem. 88, 107351 https://doi.org/10.1016/j.compbiolchem.2020.107351. ⟨https://www.sciencedirect.com/science/article/pii/S1476927120307040⟩.
Morris, G.M., Huey, R., Lindstrom, W., Sanner, M.F., Belew, R.K., Goodsell, D.S., Olson, A.J., 2009. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. arXiv J. Comput. Chem. 30 (16), 2785–2791. https://doi.org/10.1002/jcc.21256. arXiv. ⟨https://onlinelibrary.wiley.com/doi/pdf/10.1002/jcc.21256⟩.
Muratov, E.N., Amaro, R., Andrade, C.H., Brown, N., Ekins, S., Fourches, D., Isayev, O., Kozakov, D., Medina-Franco, J.L., Merz, K.M., Oprea, T., Poroikov, I, V., Schneider, G., Todd, M.H., Varnek, A., Winkler, D.A., Zakharov, A., Cherkasov, V, A., Tropsha, A., 2021. A critical overview of computational approaches employed for COVID-19 drug discovery. Chem. Soc. Rev. 50 (16), 9121–9151. https://doi.org/10.1039/d0cs01065k.
Nagar, P.R., Gajjar, N.D., Dhameliya, T.M., 2021. In search of SARS CoV-2 replication inhibitors: Virtual screening, molecular dynamics simulations and ADMET analysis. J. Mol. Struct. 1246, 131190 https://doi.org/10.1016/j.molstruc.2021.131190. ⟨https://www.sciencedirect.com/science/article/pii/S002228602101320X⟩.
Nalbandian, A., Sehgal, K., Gupta, A., Madhavan, M.V., McGroder, C., Stevens, J.S., Cook, J.R., Nordvig, A.S., Shalev, D., Sehrawat, T.S., Ahluwalia, N., Bikdeli, B., Dietz, D., Der-Nigoghossian, C., Liyanage-Don, N., Rosner, G.F., Bernstein, E.J., Mohan, S., Beckley, A.A., Seres, D.S., Choueiri, T.K., Uriel, N., Ausiello, J.C., Accili, D., Freedberg, D.E., Baldwin, M., Schwartz, A., Brodie, D., Garcia, C.K., Elkind, M.S.V., Connors, J.M., Bilezikian, J.P., Landry, D.W., Wan, E.Y., 2021. Post-acute COVID-19 syndrome. Nat. Med 27 (4), 601–615. https://doi.org/10.1038/s41591-021-01283-z.
Nanduri, S., Pilishvili, T., Derado, G., Soe, M.M., Dollard, P., Wu, H., Li, Q., Bagchi, S., Dubendris, H., Link-Gelles, R., Jernigan, J.A., Budnitz, D., Bell, J., Benin, A., Shang, N., Edwards, J.R., Verani, J.R., Schrag, S.J., 2021. Effectiveness of Pfizer-BioNTech and Moderna Vaccines in Preventing SARS- CoV-2 Infection Among Nursing Home Residents Before and During Widespread Circulation of the SARS-CoV-2 B.1.617.2 (Delta) Variant National Healthcare Safety Network, March 1-August 1, 2021. MMWR-Morb. Mortal. Wkly. Rep. 70 (34), 1163–1166.
O'Boyle, N., Banck, M., James, C., Morley, C., Vandermeersch, T., Hutchison, G., 2011. Open Babel: An open chemical toolbox. J. Cheminform. 3, 33. https://doi.org/10.1186/1758-2946-3-33.
Our World in Data, 2020. Coronavirus (COVID-19) Vaccinations. (https://ourworldindata.org/covid-vaccinations) (accessed October 07 2021).
Pesce, M., Agostoni, P., Botker, H.-E., Brundel, B., Davidson, S.M., De Caterina, R., Ferdinandy, P., Girao, H., Gyongyosi, M., Hulot, J.-S., Lecour, S., Perrino, C., Schulz, R., Sluijter, J.P., Steffens, S., Tancevski, I., Gollmann-Tepekoylu, C., Tschoepe, C., van Linthout, S., Madonna, R., 2021. COVID-19-related cardiac complications from clinical evidences to basic mechanisms: opinion paper of the ESC Working Group on Cellular Biology of the Heart. Cardiovasc. Res. 117 (10), 2148–2160. https://doi.org/10.1093/cvr/cvab201.

Petushkova, A.I., Zamyatnin, A.A.J., 2020. Papain-Like Proteases as Coronaviral Drug Targets: Current Inhibitors, Opportunities, and Limitations. sep. In: Pharmaceuticals, 13. Basel, Switzerland. https://doi.org/10.3390/ph13100277.

Profitt, T.A., Pearson, J.K., 2019. A shared-weight neural network architecture for predicting molecular properties. Phys. Chem. Chem. Phys. 21 (47), 26175–26183. https://doi.org/10.1039/c9cp03103k.

Rádl, S., 2020. A note on chloroquine. Chem. Lett. 114 (7), 426–429.

Riniker, S., Landrum, G., 2013. Open-source platform to benchmark fingerprints for ligand-based virtual screening. J. Cheminform 5, 26. https://doi.org/10.1186/1758-2946-5-26.

Sanchez-Lengeling, B., Aspuru-Guzik, A., 2018. Inverse molecular design using machine learning: Generative models for matter engineering. arXiv Science 361 (6400), 360–365. https://doi.org/10.1126/science.aat2663. arXiv. ⟨https://www.science.org/doi/pdf/10.1126/science.aat2663⟩.

Sanford, M., McCormack, P.L., 2011. Eldecalcitol, Drugs 71 (13), 1755–1770. https://doi.org/10.2165/11206790-000000000-00000.

Sanner, M.F., 1999. Python: a programming language for software integration and development. J. Mol. Graph. Model. 17 (1), 57–61.

Santana, M.V.S., Silva-Jr, F.P., 2021b. De novo design and bioactivity prediction of SARS-CoV-2 main protease inhibitors using recurrent neural network-based transfer learning. BMC Chemistry 15 (1), 8. https://doi.org/10.1186/s13065-021-00737-2.

Schütt, K., Kessel, P., Gastegger, M., Nicoli, K., Tkatchenko, A., Müller, K.-R., 2018. SchNetPack: A Deep Learning Toolbox For Atomistic Systems. J. Chem. Theory Comput. 5 (1), 448–455. https://doi.org/10.1021/acs.jctc.8b00908.

Schütt, K.T., Gastegger, M., Gastegger, A., Müller, K.-R., Maurer, R.J., 2019a. Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions. Nat. Commun. 10 (1), 5024. https://doi.org/10.1038/s41467-019-12875-2.

Schütt, K.T., Kessel, P., Gastegger, M., Nicoli, K.A., Tkatchenko, A., Müller, K.-R., 2019b. SchNetPack: A Deep Learning Toolbox For Atomistic Systems. arXiv J. Chem. Theory Comput. 15 (1), 448–455. https://doi.org/10.1021/acs.jctc.8b00908. arXiv. ⟨https://doi.org/10.1021/acs.jctc.8b00908⟩.

Shah, B., Modi, P., Sagar, S.R., 2020. In silico studies on therapeutic agents for COVID-19: Drug repurposing approach. Life Sci. 252, 117652 https://doi.org/10.1016/j.lfs.2020.117652. ⟨https://www.sciencedirect.com/science/article/pii/S0024320520304008⟩.

Shereen, M.A., Khan, S., Kazmi, A., Bashir, N., Siddique, R., 2020. COVID-19 infection: Emergence, transmission, and characteristics of human coronaviruses. J. Adv. Res. 24, 91–98. https://doi.org/10.1016/j.jare.2020.03.005. ⟨https://www.sciencedirect.com/science/article/pii/S2090123220300540⟩.

da Silva, J.K.R., Figueiredo, P.L.B., Byler, K.G., Setzer, W.N., 2020. Essential Oils as Antiviral Agents, Potential of Essential Oils to Treat SARS-CoV-2 Infection: An In-Silico Investigation. Int. J. Mol. Sci. 21 (10) https://doi.org/10.3390/ijms21103426. ⟨https://www.mdpi.com/1422-0067/21/10/3426⟩.

Smith, M., Smith, J., 2020. Repurposing Therapeutics for COVID-19: Supercomputer-Based Docking to the SARS-CoV-2 Viral Spike Protein and Viral Spike Protein-Human ACE2 Interface. https://doi.org/10.26434/chemrxiv.11871402.v4. ⟨https://europepmc.org/article/PPR/PPR116961⟩.

Štekláč, M., Zajaček, D., Bučinský, L., 2021. 3CLpro and PLpro affinity, a docking study to fight COVID19 based on 900 compounds from PubChem and literature. Are there new drugs to be found? J. Mol. Struct. 1245, 130968 https://doi.org/10.1016/j.molstruc.2021.130968. ⟨https://www.sciencedirect.com/science/article/pii/S0022286021011005⟩.

Sterling, T., Irwin, J., 2015. ZINC 15 - Ligand Discovery for Everyone. J. Chem. Inf. Model. 55, 2324–2337. https://doi.org/10.1021/acs.jcim.5b00559.

Stocker, S., Csanyi, G., Reuter, K., Margraf, J.T., 2020. Machine learning in chemical reaction space. OCT 30 Nat. Commun. 11 (1). https://doi.org/10.1038/s41467-020-19267-x.

Tarabova, D., Soralova, S., Breza, M., Fronc, M., Holzer, W., Milata, V., 2014. Use of activated enol ethers in the synthesis of pyrazoles: reactions with hydrazine and a study of pyrazole tautomerism. Beilstein J. Org. Chem. 10, 752–760. https://doi.org/10.3762/bjoc.10.70.

Tejera, E., Munteanu, C.R., López-Cortés, A., Cabrera-Andrade, A., Pérez-Castillo, Y., 2020. Drugs Repurposing Using QSAR, Docking and Molecular Dynamics for Possible Inhibitors of the SARS-CoV-2 Mpro Protease. Molecules 25 (21). https://doi.org/10.3390/molecules25215172. ⟨https://www.mdpi.com/1420-3049/25/21/5172⟩.

Ton, A.-T., Gentile, F., Hsing, M., Ban, F., Cherkasov, A., 2020. Rapid Identification of Potential Inhibitors of SARS-CoV-2 Main Protease by Deep Docking of 1.3 Billion Compounds. AUG Mol. Inform 39 (8). https://doi.org/10.1002/minf.202000028.

Van Rossum, G., Drake Jr, F. L., 1995. Python tutorial. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands.

Voors, A.A., Bax, J.J., Hernandez, A.F., Wirtz, A.B., Pap, A.F., Ferreira, A.C., Senni, M., van der Laan, M., Butler, J., 2019. Safety and efficacy of the partial adenosine A1 receptor agonist neladenoson bialanate in patients with chronic heart failure with reduced ejection fraction: a phase IIb, randomized, double-blind, placebo-controlled trial. Eur. J. Heart Fail. 21 (11), 1426–1433. https://doi.org/10.1002/ejhf.1591.

Wallace, A.C., Laskowski, R.A., Thornton, J.M., 1995. LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. Protein Eng. 8 (2), 127–134. https://doi.org/10.1093/protein/8.2.127.

Wang, C., Horby, P.W., Hayden, F.G., Gao, G.F., 2020. A novel coronavirus outbreak of global health concern. Lancet 395 (10223), 470–473. https://doi.org/10.1016/S0140-6736(20)30185-9. ⟨https://www.sciencedirect.com/science/article/pii/S0140673620301859⟩.

Wojciechowski, M., 2017. Simplified AutoDock force field for hydrated binding sites. J. Mol. Graph. Model. 78, 74–80. https://doi.org/10.1016/j.jmgm.2017.09.016. ⟨https://www.sciencedirect.com/science/article/pii/S109332631730414X⟩.

Wu, C., Liu, Y., Yang, Y., Zhang, P., Zhong, W., Wang, Y., Wang, Q., Xu, Y., Li, M., Li, X., Zheng, M., Chen, L., Li, H., 2020. Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods. Acta Pharm. Sin. B. 10 (5), 766–788. https://doi.org/10.1016/j.apsb.2020.02.008. ⟨https://www.sciencedirect.com/science/article/pii/S2211383520302999⟩.

Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., Hu, Y., Tao, Z.-W., Tian, J.-H., Pei, Y.-Y., Yuan, M.-L., Zhang, Y.-L., Dai, F.-H., Liu, Y., Wang, Q.-M., Zheng, J.-J., Xu, L., Holmes, E.C., Zhang, Y.-Z., 2020. A new coronavirus associated with human respiratory disease in China. Nature 579 (7798), 265–269. https://doi.org/10.1038/s41586-020-2008-3.

Yan, Z., Yang, M., Lai, C.-L., 2021. Long COVID-19 Syndrome: A Comprehensive Review of Its Effect on Various Organ Systems and Recommendation on Rehabilitation Plans. Biomedicines 9 (8), 966. https://doi.org/10.3390/biomedicines9080966.

Yang, M., Tao, B., Chen, C., Jia, W., Sun, S., Zhang, T., Wang, X., 2019. Machine Learning Models Based on Molecular Fingerprints and an Extreme Gradient Boosting Method Lead to the Discovery of JAK2 Inhibitors. J. Chem. Inf. Model. 59 (12), 5002–5012. https://doi.org/10.1021/acs.jcim.9b00798.

Yet, L., 2018. Privileged Structures in Drug Discovery. Medicinal Chemistry and Syntheses. John Wiley and Sons, Inc, Hoboken, NJ.

Zev, S., Raz, K., Schwartz, R., Tarabeh, R., Gupta, P.K., Major, D.T., 2021. Benchmarking the Ability of Common Docking Programs to Correctly Reproduce and Score Binding Modes in SARS-CoV-2 Protease Mpro. J. Chem. Inf. Model. 61 (6), 2957–2966. https://doi.org/10.1021/acs.jcim.1c00263.

Zhang, L., Lin, D., Sun, X., Curth, U., Drosten, C., Sauerhering, L., Becker, S., Rox, K., Hilgenfeld, R., 2020. Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α -ketoamide inhibitors. arXiv Science 368 (6489), 409–412. https://doi.org/10.1126/science.abb3405. arXiv. ⟨https://science.sciencemag.org/content/368/6489/409.full.pdf⟩.

Zhang, Y., Wang, Y., Zhou, W., Fan, Y., Zhao, J., Zhu, L., Lu, S., Lu, T., Chen, Y., Liu, H., 2019. A combined drug discovery strategy based on machine learning and molecular docking. Chem. Biol. Drug Des. 93 (5), 685–699. https://doi.org/10.1111/cbdd.13494.

Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., Si, H.-R., Zhu, Y., Li, B., Huang, C.-L., Chen, H.-D., Chen, J., Luo, Y., Guo, H., Jiang, R.-D., Liu, M.-Q., Chen, Y., Shen, X.-R., Wang, X., Zheng, X.-S., Zhao, K., Chen, Q.-J., Deng, F., Liu, L.-L., Yan, B., Zhan, F.-X., Wang, Y.-Y., Xiao, G.-F., Shi, Z.-L., 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579 (7798), 270–273. https://doi.org/10.1038/s41586-020-2012-7.

Zumla, A., Chan, J.F.W., Azhar, E.I., Hui, D.S.C., Yuen, K.-Y., 2016. Coronaviruses - drug discovery and therapeutic options. Nat. Rev. Drug Discov. 15 (5), 327–347. https://doi.org/10.1038/nrd.2015.37.