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.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Machine learning prediction of 3CL\textsuperscript{pro} SARS-CoV-2 docking scores

Lukas Bucinsky\textsuperscript{a,}\textsuperscript{*}, Dušan Bortnák\textsuperscript{b}, Marián Gall\textsuperscript{c,d}, Ján Matuška\textsuperscript{a}, Viktor Milata\textsuperscript{b}, Michal Pito\textsuperscript{e,d}, Marek Šteklác\textsuperscript{e}, Daniel Végh\textsuperscript{b}, Dávid Zajacek\textsuperscript{a}

\textsuperscript{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
\textsuperscript{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
\textsuperscript{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
\textsuperscript{d} Institute of Physical and Theoretical Chemistry, Faculty of Natural Sciences, Comenius University in Bratislava, Mlynská dolina, Ilkovičova 6, SK-84215 Bratislava, Slovak Republic

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

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 6.8 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).
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 proteins: spike glycoprotein, envelope protein, nucleocapsid protein, Flow Abadi et al. (2015) and XGBoost Chen and Guestrin (2016), are to techniques is Deep Tensor Neural Networks (DTNNs) implemented in 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 possibility that this is not the last pandemic. Therefore, the tools of trade proteins: spike glycoprotein, envelope protein, nucleocapsid protein, Flow Abadi et al. (2015) and XGBoost Chen and Guestrin (2016), are to techniques is Deep Tensor Neural Networks (DTNNs) implemented in the cytoplasm occurs, initiating the translation of the viral genome into the

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 proteins: spike glycoprotein, envelope protein, nucleocapsid protein, Flow Abadi et al. (2015) and XGBoost Chen and Guestrin (2016), are to techniques is Deep Tensor Neural Networks (DTNNs) implemented in 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 possibility that this is not the last pandemic. Therefore, the tools of trade proteins: spike glycoprotein, envelope protein, nucleocapsid protein, Flow Abadi et al. (2015) and XGBoost Chen and Guestrin (2016), are to techniques is Deep Tensor Neural Networks (DTNNs) implemented in the cytoplasm occurs, initiating the translation of the viral genome into the...
3. The selection of structures for the application of computational methods to model their potential affinity to 3CL level of theory, using OpenBabel, forming the test set, denoted as Slovak University of Technology (SUT) were optimized at the MMFF94 by the coauthors from the Department of Organic Chemistry of the et al. (2011). In addition, the 3D geometries of 172 compounds proposed Halgren (1996d) using the OpenBabel chemical toolbox 2.3.2. O MMFF94 force field Halgren (1996a, b, c); Halgren and Nachbar (1996); without a 3D structure in either database were optimized in the PubChem database (31), the ZINC database (10) Sterling and Irwin (2015). Compounds with the same ZINC codes and those which are 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 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 unrecognized 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.

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 23rd 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 11-120 (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, 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. 3CLpro structure

The 3D structure of the SARS-CoV-2 3CLpro protease determined at room temperature was downloaded from the RCSB Protein Data Bank Berman et al. (2000) (PDB ID: 6WQF). The 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 3CLpro 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 Å, 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 orthogonal 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), max=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
Underlying neural network consists of two hidden layers with sizes unchanged with respect to the defaults provided by SchNetPack, i.e., the 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 considered for this purpose. Second, the number of putative hydrogen bonds formed between the compound and SARS-CoV-2 3CL\textsuperscript{pro} was calculated.

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\textsuperscript{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 %), Glu192 (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 top two scoring compounds to SARS-CoV-2 3CL\textsuperscript{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 Steklac et al. Steklac et al. (2021). However, differences arise between the results presented herein and Steklac 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\textsuperscript{pro} binding affinities similar.
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 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 3CLpro. 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 3CLpro 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 3CLpro from compounds selected in 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 3CLpro 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 3CLpro 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
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: 123996), 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 O15) 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 O99) 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 S2c, 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

### Table 3

| Data set | CID         | Amino acids | DS [kcal/mol] | H-bond forming amino acids |
|----------|-------------|-------------|---------------|----------------------------|
| S        | 118628567   | Gln166, Thr190 | -14.45        |                             |
|          | 53472683    | Thr190      | -14.40        |                             |
|          | 6918155     | Thr26, Gln189 | -14.38        |                             |
|          | 5281040     | Glu186      | -14.27        |                             |
|          | 5459840     | Gln186, Arg188, Gln192 | -14.04 | Glu189, Arg188, Gln192 |
| B        | 25151504    | Gln189, Arg188, Gln192 | -15.71 | Gln189, Arg188, Gln192 |
|          | 24873435    | Ser46, Glu166 | -14.71        |                             |
|          | 3010818     | Thr26, His41, Gln189 | -14.66  |                      |
|          | 45110509    | Gln189, Arg188, Gln192 | -14.27 | Glu186, Arg188, Gln192 |
| L        | 5362440     | Gln186      | -14.02        |                             |
|          | 3571576     | Gly143      | -15.96        |                             |
|          | 5684985     | Thr26, Arg188, Gln189 | -15.49  |                      |
|          | 6918141     | Cys145, Gly143, Gln189 | -15.36 | Gln189, Arg188, Gln192 |
|          | 5289508     | Thr26, His163, Gln189 | -15.11  |                      |
|          | 42613186    | Phe140, Gly143, Gln189 | -14.75 | Gln189, Arg188, Gln192 |
| O        | 0.090*      | Gln186      | -15.01        |                             |
|          | 0.104*      | Gln186      | -14.01        |                             |
|          | 0.120       | Gln189      | -13.74        |                             |
|          | 0.101*      | Gln186      | -13.73        |                             |
|          | 0.096*      | Gln186      | -13.64        |                             |
|          | P*          | Gln186      | -15.81        |                             |

* Compound does not have a trivial name
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.
fairly large, 7.95 kcal/mol (CID: 123966). This test set, yet certain outliers are clearly visible. The MAE outlier is performance when considering the score span of about 14 kcal/mol in NNs as well. Interestingly, the TF slopes for the noting that the Fischer et al. (2020)), see below and in Table 4. It is worth structures. Figs. 6d and 6f show the inference accuracy of the training set XG GBDT is superior to the training set with respect to the 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 training set. The MSE decreases from 2.740 to 2.409 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 the MAE outlier decreases to 3.59 kcal/mol (CID: 25151504), which is comparable to the S training set (2.40 kcal/mol, O_95) of the L training set. In contrast, the description of the S 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, 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, 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, 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).
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) 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 (R²) 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: 52808899), 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, R², MAE 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 8.20.40 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 (MAE, 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 (red 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 (black dots) as found by AutoDock, see Table S2.

Below, the best Autodocking 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 of out five top scoring compounds (ZINC00049593065, 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, Gin189 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 4 kcal/mol between expected and predicted scores is evident. For example, around 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\textsuperscript{+} per 20 atoms batches with the size of the compounds (b) TF; (d) XG; and (f) SP.
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 Borňák: Investigation, review draft; Marian Gall: Investigation, Methodology TensorFlow, Writing original draft, review and editing, Validation; Ján Matiška: Investigation, Methodology SchNetPack, Writing original draft, review and editing, Validation; Viktor Milata: Investigation, Review draft; Michal Pitonák: Investigation, Methodology XGBoost, Writing original draft, review and editing, Validation; Marek Stekláč: Investigation, Methodology Docking, Writing original draft, review and editing, Validation; Daniel Veľa: Investigation, review draft; Dávid Zajacek: 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.
Tejera, E., Munteanu, C.R., L. Tarabova, D., Soralova, S., Breza, M., Fronc, M., Holzer, W., Milata, V., 2014. Use of Stekl 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/10.1039/c8ra01294j.

Sanchez-Lengeling, B., Aspuru-Guzik, A., 2018. Inverse molecular design using machine learning: Generative models for mass engineering. arXiv Science 368 (6489), 409.

Schütt, K., Kössel, P., Gastegger, M., Nicoli, K.A., Tkatchenko, A., Müller, K.-R., 2019b. Simplified AutoDock force field for hydrated binding sites. Nat. Commun. 10 (1), 5024. https://doi.org/10.1038/s41467-019-12875-2.

Schütt, K.T., Kössel, P., Gastegger, M., Nicoli, K.A., Tkatchenko, A., Müller, K.-R., 2019h. SchNetPack: A Deep Learning Toolbox For Atomsimics. J. Chem. Theory Comput. 15 (1), 448–455. https://doi.org/10.1021/acs.jctc.9b00598.

Shah, B., Modi, P., Sagar, S.R., 2020. In silico studies on therapeutic agents for COVID-19. Drug Repurposing Using QSAR, Docking and Molecular Dynamics for New Drugs to Be Found? J. Mol. Struct. 1245, 130968 https://doi.org/10.1016/j.molstruc.2020.130968.

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.

Shreiner, 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.

Shi, Z., 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/s41586-020-2085-z.

Shi, Z., 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/s41586-020-2085-z.

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.