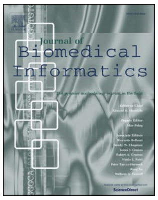
Journal of Biomedical Informatics 119 (2021) 103821

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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



Special Communication

Machine learning enabled identification of potential SARS-CoV-2 3CLpro inhibitors based on fixed molecular fingerprints and Graph-CNN neural representations☆

Jacek Haneczok [a,*], Marcin Delijewski [b,**]

[a] Erste Group IT, Am Belvedere 1, 1100 Vienna, Austria
[b] Department of Pharmacology, Faculty of Medical Sciences in Zabrze, Medical University of Silesia, Katowice, Poland

ARTICLE INFO

Keywords
AI drug repurposing
Machine learning
Graph Convolutional Neural Network
Molecular property prediction
SARS-CoV-2
COVID-19

ABSTRACT

Aim: Rapidly developing AI and machine learning (ML) technologies can expedite therapeutic development and in the time of current pandemic their merits are particularly in focus. The purpose of this study was to explore various ML approaches for molecular property prediction and illustrate their utility for identifying potential SARS-CoV-2 3CLpro inhibitors.

Materials and methods: We perform a series of drug discovery screenings based on supervised ML models operating in different ways on molecular representations, encompassing shallow learning methods based on fixed molecular fingerprints, Graph Convolutional Neural Network (Graph-CNN) with its self-learned molecular representations, as well as ML methods based on combining fixed and Graph-CNN learned representations.

Results: Results of our ML models are compared both with respect to the aggregated predictive performance in terms of ROC-AUC based on the scaffold splits, as well as on the granular level of individual predictions, corresponding to the top ranked repurposing candidates. This comparison reveals both certain characteristic homogeneity regarding chemical and pharmacological classification, with a prevalence of sulfonamides and anticancer drugs, as well as identifies novel groups of potential drug candidates against COVID-19.

Conclusions: A series of ML approaches for molecular property prediction enables drug discovery screenings, illustrating the utility for COVID-19. We show that the obtained results correspond well with the already published research on COVID-19 treatment, as well as provide novel insights on potential antiviral characteristics inferred from *in vitro* data.

1. Introduction

Among various techniques from the fields of artificial intelligence (AI) and machine learning (ML), the applications to the problem of molecular property prediction are of central significance for the drug discovery process, starting from the early screening phase in which potential promising drug candidates can be identified [1]. In the current urgent need to fight the global COVID-19 pandemic the merits of AI and ML are particularly in focus [2–10], taking into account that *in silico* results are still subject to additional *in vitro* and *in vivo* experiments and further clinical trials to ensure their safety and efficacy [11].

The current pandemic crisis is caused by a novel coronavirus, named severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), that emerged in 2019 in Wuhan, China and rapidly turned out to be a life-threatening pathogen. On March 11, 2020, the World Health Organisation declared COVID-19 a pandemic, leading to nearly 100 million COVID-19 cases and over 2 million deaths till the moment of writing this article [12,13]. The novel pathogen belongs to the family Coronaviridae and it is an enveloped virus with a positivesense, single-stranded RNA genome. SARS-CoV-2 belongs to the genus Betacoronavirus, together with SARS-CoV-1 [14] and both are zoonotic pathogens that have caused fatal epidemics or pandemics after entering the human population. For these highly pathogenic viruses pharmacotherapeutic interventions are still needed to be improved [15].

☆ The views expressed in this article are those of the authors and not necessarily those of Erste Group IT.

* Corresponding author.

** Corresponding author.

E-mail addresses: jacek.haneczok@gmail.com (J. Haneczok), mdelijewski@sum.edu.pl (M. Delijewski).

https://doi.org/10.1016/j.jbi.2021.103821
Received 8 February 2021; Received in revised form 18 April 2021; Accepted 16 May 2021
Available online 28 May 2021

Drug repurposing, or repositioning, defined as finding alternative indications for approved or investigational drugs outside their primary registration, could be a possible way to overcome the time limitation of research and development needed to design a new drug. Repurposed drugs have lower risk of failure and require lower investments compared to de novo drug development, [16], which has a very low success rate, reaching about 6.2% [1,17] while taking typically 12 to 15 years [18].

The purpose of our study was to identify the best repurposing candidates among the Food and Drug Administration (FDA) approved drugs, based on their predicted antiviral activity against SARS-CoV-2. To this end we have trained supervised machine learning models based on data from a large crystallographic fragment screen against SARS-CoV-2 3CL protease (3CLpro). The 3CLpro of SARS-CoV-2 known as the main protease is an enzyme which has essential role in processing the polyproteins that are translated from the viral RNA. The 3CLpro operates at no fewer than 11 cleavage sites on the large polyprotein 1ab (replicase 1ab) and inhibition of the activity of this enzyme blocks viral replication. The big advantage of the supposed inhibitors is that such molecules are unlikely to be toxic, as no human proteases with a similar cleavage specificity are known [19].

Similar studies in the context of *in silico* AI and ML applications for identifying drug candidates for the treatment of COVID-19 have been reported in the literature, based on different ML approaches, as well as different training datasets and molecular representation methods. Kowalewski et al. [20] used shallow learners (random forest and SVM) trained on data for 65 target human proteins known to interact with the SARS-CoV-2 proteins, including the ACE2 receptor, and showed volatile candidates as novel inhaled therapeutics. Beck et al. [6] applied a deep learning-based drug-target interaction model (Molecule Transformer-Drug Target Interaction) pre-trained on the Drug Target Common (DTC) database and BindingDB database (with manual curations) and predicted antiviral properties among known antiviral drugs in order to identify the most promising anti-SARS-CoV-2 candidates. Ke et al. [4] utilized a deep neural network model trained on two different databases, one of them including in particular next to SARS-CoV active drugs also drugs active against human immunodeficiency virus and influenza virus and indicated the best drug candidates after confirming their activity against a feline infectious peritonitis (FIP) virus. The results obtained in comparable *in silico* approaches show repurposing candidates from very diverse pharmacological groups, including, among others, antibiotics, antivirals, painkillers, diuretics, antihistamines, drugs acting on respiratory tract, circulatory and cardio-vascular systems as well as anticancer drugs [20,6,4,21,5].

The main contribution of our study is twofold. Firstly, we explore various ML approaches, operating in different ways on molecular representations, encompassing:

- Shallow learning methods based on fixed molecular fingerprints,
- Graph Convolutional Neural Network (Graph-CNN) model with its self-learned molecular representations,
- ML methods based on combining fixed and Graph-CNN learned molecular representations.

Secondly, we describe a series of drug discovery screenings based on these approaches, and illustrate their utility for identifying novel groups of potential drug candidates against COVID-19. We show that the obtained results both correspond well with the already published results on COVID-19 treatment, as well provide novel insights on potential antiviral characteristics inferred from *in vitro* data obtained using crystallography techniques. An illustrative overview of the considered ML enabled screening pipelines is given in Fig. 1.

2. Materials and methods

2.1. Datasets

The dataset used to train our models consists of molecular samples from the fragments screened for binding with SARS-CoV-2 3CL protease (3CLpro) using crystallography techniques. Data is sourced from the Diamond Light Source group [22] and deposited in the Protein Data Bank, with appropriate protocols and experimental details. Data was released on the 18th March, 2020, and contained ~ 880 samples, thereof 78 hits, 58 on the active site and 39 which are covalently bound [23–25]. The screening was based on the crystal structure of 3Clpro at 2.16 Å in complex with a covalent inhibitor [26]. This structure of the SARS-CoV-2 3CLpro at high resolution (PDB ID: 6YB7) was used to conduct a large crystallographic fragment screen against it. The full length protein was cloned as described in [27] for the SARS main protease, which yielded crystals of the unliganded enzyme that diffracted to high resolution (1.25 Å) on beamline I04-1, in a different space group to the inhibitor complex. The structure was then determined and refined, the active site was empty and solvent accessible, building a setup for screening, including fragment-based drug discovery (XChem), where small chemical fragments can be soaked into drug targets, leading to release of set of 80 hit structures which were fully modelled and refined [28]. Due to the fact that SARS-CoV-2 3CLpro is an integral component in the viral replication process the considered dataset composes to our knowledge the largest and most actual currently available sample of this type, on which a ML model can be trained to provide predictions for inferring the antiviral activity against SARS-CoV-2. Compounds that inhibit SARS-CoV-2 3CLpro are tagged as active (antiviral activity is positive). The obtained combined predictions offer structural and reactivity information for on-going structure-based drug design against

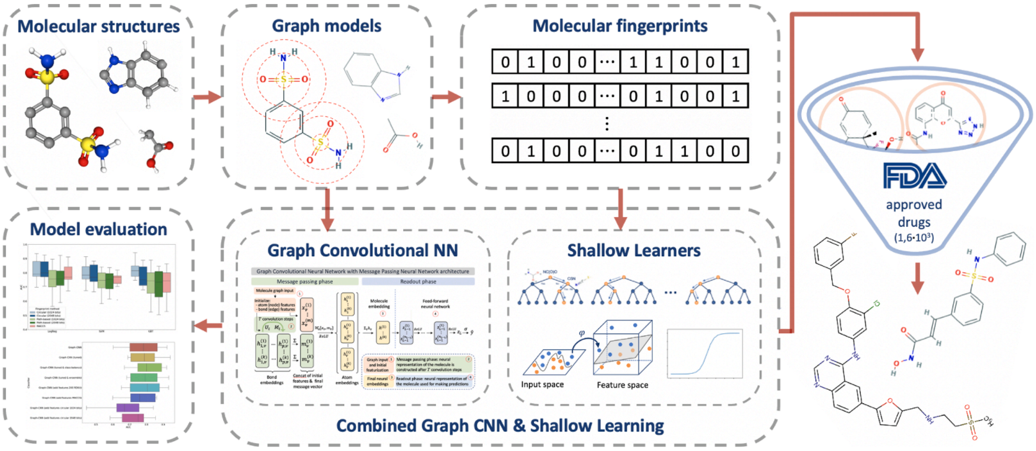


Fig. 1. A schematic illustration of the utilized machine learning enabled screening pipelines based on fixed molecular fingerprints and Graph-CNN neural representations.

SARS-CoV-2 3CLpro [28].

For a screening set of candidate molecules, among which the best repurposing candidates are identified based on their predicted inhibiting potential against SARS-CoV-2, we employed the FDA set of all approved drugs [29]. The set of FDA approved drugs is an important resource for medical practice and consists of compounds that are safe and efficacious drug products, approved by the FDA for use in the USA [30].

2.2. *Machine Learning Task*

In this study we consider supervised ML approaches applied to the task of molecular property prediction [31] based on a training set given as a set of pairs $\{(x_i,y_i) : i = 1,\ldots,n\}$, where x_i is the representation of the molecular structure of the ith compound, taken as the starting point for the learning algorithm, and y_i is its property or activity score, in our context binary label corresponding to the activity or inactivity of the compound.

2.3. *Molecular representations*

In general, two types of representations of the molecular structure x_i can be considered:

- Fixed molecular descriptors or fingerprints where $x_i \in \mathbb{R}^p$ are precomputed numerical vectors of features or $x_i \in \mathbb{N}^p$ are nonnegative count vectors or, more typically, $x_i \in \{0,1\}^p$ are binary vectors representing the presence or absence of particular substructures (or other characteristics of the compound),
- Molecular graph encodings (such as SMILES [32,33]) x_i, which, inspired by representational learning [34], are further jointly converted into fixed-length embeddings $\widetilde{x}_i \in \mathbb{R}^d$ using sequence or graph models, most typically based on neural networks, yielding self-learned neural fingerprints or embeddings, jointly learned through back-propagation.

Accordingly, ML models operating on the fixed pre-computed molecular fingerprints are typically shallow learning methods, whereas deep neural networks are used for constructing the self-learned molecular representations. In general it remains an open research area to find out which of the two paradigms is superior for which task and for which characteristics of the underlying dataset [1,35].

2.4. *Shallow learning on fixed representations*

Based on fixed pre-computed molecular fingerprints providing representations of the molecular structures, shallow learning approaches apply further transformations. Fixed molecular fingerprints are typically either dictionary-based or hash-based. Whereas dictionary-based fingerprints rely on a predefined dictionary of substructures or features, hash-based approaches use hashing algorithms to combine large number of substructures into unique fingerprints. Depending on how the substructures are enumerated, the hash-based fingerprints can be further divided into topological or path-based and circular fingerprints. We use in our experiments the following methods: i) dictionary-based MACCS keys [36], ii) hashed path-based RDKit's implementation inspired by the Daylight fingerprint [37] and iii) hashed circular fingerprints generated using a variant of the Morgan algorithm [38].

We include in our study the current state-of-the-art shallow learners such as decision tree ensembles and SVMs and compare them with regularized logistic regression.

2.4.1. *Regularized Logistic Regression*

Logistic regression (LogReg) model [39] involves modeling the conditional distribution of the response given the predictors $p(y|x)$ using the logistic function. For the two-class classification problem coded via $y \in \{-1,1\}$ the logistic regression model is of the form $p(y = 1|x) = e^{\beta_0+\beta^T x_i}/(1 + e^{\beta_0+\beta^T x_i})$, where β's are unknown parameters. A regularized logistic regression with L2 penalty is fitted by minimizing the following cost function

$$\min_{\beta_0,\beta}(\frac{1}{2}\beta^T\beta + C\sum_{i=1}^{n}(\log(\exp(-y_i(\beta_0+\beta^T x_i))+1)),$$

where $C > 0$ is a hyper-parameter corresponding to the inverse of the regularization strength.

2.4.2. *Support Vector Machine (SVM)*

Support vector machine (SVM) [40,39] is a kernel-based classification method attempting to identify an optimal decision boundary between the observations belonging to two categories. Decision boundaries are found by mapping the training data x to a high-dimensional version $\varphi(x)$, where φ is a mapping (called the feature map) from the input space into some Hilbert space $\mathscr{F}$, called the feature space.[1] The SVM algorithm finds in the feature space $\mathscr{F}$ a decision boundary as a linear hyperplane with the maximal margin, where the margin maximization problem can be conveniently reformulated as the following convex optimization problem

$$\min_{\alpha_0,\boldsymbol{\alpha}}(\sum_{i=1}^{n}\max(0,1-y_i f(x_i)) + \frac{\lambda}{2}\boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha}),$$

with $f(x) = \alpha_0 + \sum_{i=1}^{n}\alpha_i k(x,x_i)$, where $y_i \in \{-1,1\}$, $\mathbf{K}$ is the matrix of kernel evaluations for all pairs of training points and $\lambda = \frac{1}{C}$ with a tuning parameter $C > 0$ controlling the cost of violation to the separation margin. Given the solutions $\widehat{\alpha_0}$ and $\widehat{\boldsymbol{\alpha}}$ the decision function is given by $\mathrm{sign}(\widehat{f}(x)) = \mathrm{sign}(\widehat{\alpha_0} + \sum_{i=1}^{n}\widehat{\alpha}_i k(x,x_i))$.

2.4.3. *Gradient Boosted Tree Ensemble (GBT)*

The approach based on Gradient Boosted Trees (GBT) adopted in this study is based on [41] and relies on ensembling $m = 1,\ldots,M$ decision trees $T(x;\Theta^{[m]})$, whose predictions are given by

$$T(x;\Theta^{[m]}) = \sum_{j=1}^{J}\gamma_j^{[b]}I(x \in R_j^{[m]}),$$

where $R_j^{[m]}$ are the terminal nodes (representing the disjoint regions of the feature space) and $\gamma_j^{[m]}$ are estimates of class probabilities assigned to each terminal node. In the mth step, given the predictions $\widehat{y}^{[m-1]}$ based on the current ensemble consisting o $(m-1)$ trees, the following optimization problem is solved

$$\min_{\Theta^{[m]}}\sum_{i=1}^{n}l(y_i,\widehat{y}^{[m-1]} + T(x_i;\Theta^{[m]})) + \Omega(T(x;\Theta^{[m]}))$$

for the terminal nodes and their values $\Theta^{[m]} = \{R_j^{[m]},\gamma_j^{[m]}\}_{j=1}^{J^{[m]}}$ of the mth tree, where l is the log-loss function and $\Omega(T(x;\Theta^{[m]}))$ is the penalty term[2]. Hence, the ensemble model is greedily updated by adding the new tree $T(x;\Theta^{[m]})$ that most improves the overall model. The prediction of the final GBT ensemble is given by

$$\widehat{y} = \sum_{m=1}^{M}T(x;\Theta^{[m]}).$$

[1] Every feature map φ defines a positive definite kernel via $k(x,x') = \langle\varphi(x),\varphi(x')\rangle_{\mathscr{F}}$, which is typically interpreted as a measure of dissimilarity between the inputs x and x'.

[2] The penalty term is given by $\Omega(T(x;\Theta^{[m]})) = \gamma J^{[m]} + \frac{1}{2}\lambda\sum_{j=1}^{J}\gamma_j^{[m]^2}$, with γ and λ being the regularization hyper-parameters.

2.5. Graph Convolutional Neural Network (Graph-CNN)

A Graph Convolutional Neural Network (Graph-CNN) operates on graph data, where nodes represent atoms, edges represent bonds, and the process of jointly encoding the molecular substructures and aggregating or pooling the information into fixed-length embeddings is similar to the one used in Convolutional Neural Networks (CNNs). Similarly as in case of CNNs, layers that come earlier in the Graph-CNN model extract low-level generic features (representing molecular substructures) and layers that are higher up extract higher-level, more abstract features (representing more elaborate substructures) towards the Graph-CNN predictive objective. For non-Euclidean data[3], such as graph data, defining the operations of convolution or pooling is not straightforward or even possible and Graph Neural Networks (GNNs) are a deep learning approach for addressing these difficulties [42,43].

A unifying framework for Graph-CNNs, generalizing several GNNs and CNNs approaches, proposed by Google research scientists in [44], are so called Message Passing Neural Networks (MPNNs). An MPNN operates on an undirected graph[4] G with features x_v representing the vth node (atom) and e_{vw} representing the edge (bond) between nodes v and w and the forward pass consist of two phases: a message passing phase consisting of T steps (convolutions) creating the molecular representations (self-learned fingerprints) and a readout phase using the final representations for making predictions. The message passing phase is initiated by mapping atom features x_v to another set of vectors h_v^0 termed hidden states. In the tth step a message m_v^{t+1} is created according to

$$m_v^{t+1} = \sum_{w \in N(v)} M_t(h_v^t, h_w^t, e_{vw}),$$

where $N(v)$ is the set of neighbors of v in graph G and M_t is a message function, and used further for updating the hidden states by

$$h_v^{t+1} = U_t(h_v^t, m_v^{t+1}),$$

where U_t is a vertex update function. The readout phase uses a readout function R to map the final hidden states representing the molecule to the final output of the neural network

$$\hat{y} = R(\{h_v^T : v \in G\}).$$

We adopt the directed version of MPNN [45] with the implementation used in [46], where edge (bond) based hidden states h_{vw}^t and messages m_{vw}^t are used rather than the node (atom) based states h_v^t and messages m_v^t counterparts[5], as illustrated in Fig. 2. Hidden states are initialized with

$$h_{vw}^0 = ReLU(W_0[x_v, e_{vw}]),$$

where $ReLU$ is the rectifier activation function, $W_0 \in \mathbb{R}^{h \times h_0}$ is a weight matrix and $[x_v, e_{vw}] \in \mathbb{R}^{h_0}$ is the concatenation of the atom features x_v and bond features e_{vw}. The message passing update equation is taken as

$$m_{vw}^{t+1} = \sum_{k \in N(v) \setminus w} M_t(x_v, x_w, h_{kv}^t) = \sum_{k \in N(v) \setminus w} h_{kv}^t$$

and the hidden state updates are calculated using the same function at each step t according to

$$h_{vw}^{t+1} = U_t(h_{vw}^t, m_{vw}^{t+1}) = U(h_{vw}^t, m_{vw}^{t+1}) = ReLU(h_{vw}^0 + W_m m_{vw}^{t+1}),$$

where $W_m \in \mathbb{R}^{h \times h}$ is a weight matrix. After the final convolution step T the final representation of the vth atom of the molecule is calculated as

$$h_v = ReLU(W_a[x_v, m_v]),$$

where $m_v = \sum_{k \in N(v)} h_{kv}^T$ and $W_a \in \mathbb{R}^{h \times h_a}$ is a weight matrix with h_a such that $[x_v, m_v] \in \mathbb{R}^{h_a}$. In the readout phase, the final hidden states h_v representing the molecule are summed to produce a single embedding vector for the molecule $h = \sum_{v \in G} h_v$ and the predictions are generated via $\hat{y} = f(h)$, where f is a feed-forward neural network. After calculating the predictions, the loss function is computed over a batch of molecules based on model predictions and the ground truth values. The gradients of the loss with respect to the network weights is calculated by means of back-propagation and used by the optimizer to iteratively update the weights.

2.6. Combining fixed and Graph-CNN learned representations

This section provides a description of considered approaches for combining fixed representations (fingerprints) and self-learned representations based on the Graph-CNN model.

2.6.1. Graph-CNN enhanced with additional molecular features

Graph-CNN approach described in 2.5 can be further extended by adding additional, auxiliary molecular features in order to enhance the self-learned representations. To this end the vector of self-learned representations h is concatenated with the additional features vector h_{aux} in the readout phase and the predictions are generated via

$$\hat{y} = f([h, h_{aux}]),$$

where f is a feed-forward neural network.

2.6.2. Shallow learners enhanced with Graph-CNN self-learned neural embeddings

As another alternative for combining fixed pre-computed molecular fingerprints and self-learned molecular representations we consider regularized logistic regression and GBT models described in 2.4.1 and 2.4.3 trained on the concatenation of feature vectors $[x, h]$, where x are pre-computed fingerprints and h are the neural embeddings extracted from Graph-CNN readout phase as sum over the the atom embeddings of the molecule $h = \sum_{v \in G} h_v$, as described in Section 2.5.

2.6.3. Stacking ensemble of Graph-CNN and shallow learners

Stacking ensemble (sometimes called stacked generalization) [49] is a technique for combining multiple different learning algorithms, referred to as base models, by training a new learning algorithm on the predictions generated by the base models. The base-models are sometimes also called level-0 models and the new model combining their predictions is referred to as the meta-model or level-1 model. We employ the following basic stacking algorithm using 2-fold cross-validation:

1. Split the train set in two parts,
2. Train each of the base-models on the first part of the train set and generate predictions for the second part,
3. Train each of the base-models on the second part of the train set and generate predictions for the first part,

[3] One of the keys to the success of CNNs and other deep NNs is their ability to capitalize on the statistical properties such as stationarity and compositionality through local statistics, which are exhibited by e.g. natural image or video data. In the context of CNNs and computer vision tasks stationarity stems from shift-invariance and composionality from the multi-resolution data structure and these properties are exploited by e.g. alternating convolutional and pooling layers. More generally, for data with underlying Euclidean or grid-like structure the invariances of underlying data structures are the key properties built into network architecture.

[4] The formalism of MPNN can be easily extended to directed multigraphs.

[5] The motivation for this model design is to avoid messages that loop back to their preceding node, which can introduce noise [47,46]. For similarities between this edge-based message passing design and belief propagation in probabilistic graphical models see [45,48]

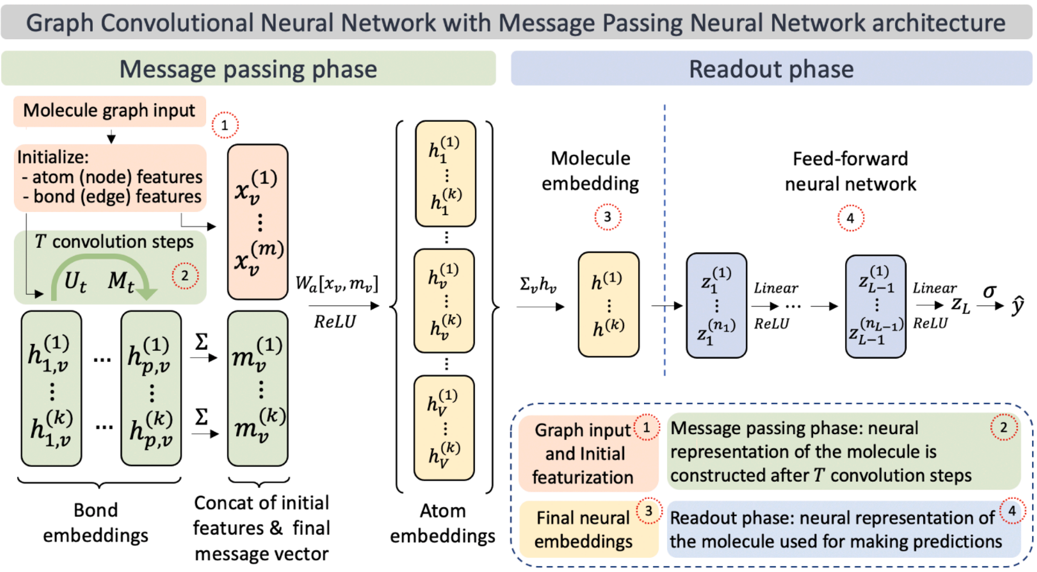

Graph Convolutional Neural Network with Message Passing Neural Network architecture
Message passing phase
Readout phase
Molecule graph input
1
Initialize:
- atom (node) features
- bond (edge) features
T convolution steps
U_t M_t
2
$x_v^{(1)}$ ⋮ $x_v^{(m)}$
$W_a[x_v, m_v]$
$ReLU$
$h_{1,v}^{(1)}$ ⋮ $h_{1,v}^{(k)}$
…
$h_{p,v}^{(1)}$ ⋮ $h_{p,v}^{(k)}$
Σ
Σ
$m_v^{(1)}$ ⋮ $m_v^{(k)}$
Bond embeddings
Concat of initial features & final message vector
$h_1^{(1)}$ ⋮ $h_1^{(k)}$
$h_v^{(1)}$ ⋮ $h_v^{(k)}$
$h_V^{(1)}$ ⋮ $h_V^{(k)}$
Atom embeddings
$\Sigma_v h_v$
Molecule embedding
3
$h^{(1)}$ ⋮ $h^{(k)}$
Feed-forward neural network
4
$z_1^{(1)}$ ⋮ $z_1^{(n_1)}$
Linear
ReLU
$z_{L-1}^{(1)}$ ⋮ $z_{L-1}^{(n_{L-1})}$
Linear
ReLU
z_L
σ
$\hat{y}$
Graph input and Initial featurization 1
Message passing phase: neural representation of the molecule is constructed after T convolution steps 2
Final neural embeddings 3
Readout phase: neural representation of the molecule used for making predictions 4

Fig. 2. A schematic illustration of the described Graph Convolutional Neural Network (Graph-CNN) based on Message Passing Neural Network (MPNN) architecture. Message passing phase (left): k-dimensional atom embeddings h_v, for all $v = 1, \ldots, V$ atoms (vertices of graph G representing the molecule), calculated as the $ReLU$-transformed products of the weight matrix W_a and the concatenation of the initial atom features x_v and the sums m_v of the final values of the bond (edge) states h_{kv} for $k \in N(v)$, obtained after the final convolution step T. Readout phase (right): atom embeddings h_v summed up to build the k-dimensional neural representation h of the molecule are further passed as an input to the feed-forward neural network with L-layers, followed by a sigmoid activation function σ outputting a binary response.

4. Use out-of-sample predictions from steps 2. and 3. to train the meta-model
5. Re-train each of the base models on the full train set for generating predictions for the held-out test set.

As we aim at building a model combining fixed molecular representations with neural self-learned representations we consider two pairs of base learners: i) LogReg and Graph-CNN, and ii) GBT and Graph-CNN. The first pair represents a stacking of Graph-CNN with arguably the simplest shallow learner and the latter a stacking of state-of-the-art shallow classifier with Graph-CNN. As a meta-model we employ a logistic regression model.

2.7. Model training setups

All utilized fingerprints are calculated using the Python's API to the `RDKit` library. Our shallow learning pipelines are based on `scikit−learn` Python library [50]. Regularized logistic regression is implemented using scikit-learn's `LogisticRegressionCV` using the quasi-Newton `lbfgs` optimizer. SVM is implemented using scikit-learn's `SVC` classifier with Gaussian radial basis function (RBF) kernel. The GBT implementation is based on the scikit-learn wrapper interface for `XGBoost` [41]. The hyper-parameter tuning for the shallow learners was performed with the grid-search method, using 5-fold and 10-fold stratified cross-validation, resulting in a nested cross-validation setup. The Graph-CNN model is based on `Chemprop`'s implementation using `PyTorch`, described in [46], including the same initialization of the feature vectors (for atoms: atomic number, number of bonds, formal charge, chirality, number of bonded hydrogen atoms, hybridization, aromaticity, atomic mass; for bonds: bond type, conjugation, ring membership and stereochemistry features). For performing stochastic gradient-based optimization in the training stage the Adam algorithm [51] was used and adjustments of the learning rate were performed using the Noam scheduler with piecewise linear increase and exponential decay, inspired by [52]. During training, the model was evaluated with respect to AUC on a holdout validation set containing 10% of the training molecules and the early stopping technique was employed. Hyper-parameters of Graph-CNN were tuned via Bayesian optimization method utilizing `Hyperopt` Python library, using the implementation described in [46].

2.8. Evaluation Method

The quality of the classifier outputs is assessed using Receiver Operating Characteristic curve-Area Under the Curve (ROC-AUC) as the primary metric. Performance evaluation is based on repeated random sub-sampling cross-validation using 10 randomly seeded 80:20 scaffold splits of training and testing data. Comparing to commonly used random splits, the utilized scaffold splits offer a superior and more challenging evaluation setup, testing model's ability to generalize to new chemical spaces, critical for drug repurposing applications [35]. The utilized scaffold splitting approach follows [53,35] and relies on partitioning of the molecular structures based on their Murcko scaffolds calculated using RDKit. In order to test the models' capacity to generalize to new molecular structures based on the desired test set size and ensuring that the test set is not too homogeneous, partitions with molecule count that would exceed the half of the test set size are allocated to the train set. The remaining partitions are randomly allocated to the train and test sets, until the desired split ratio is achieved.

2.9. Generating Final Predictions

The final predictions of the antiviral activity of the FDA approved drugs are generated after re-training the models on the whole dataset and ranking the repurposing candidates based on their predicted probabilities of activity. To reduce the variability inherent in predictions due to stochastic nature of the algorithms, we build meta-ensembles consisting of 10 models, each trained with a different random seed. The final rank of each drug used for ordering the repurposing candidates is obtained by taking the median over the individual rankings from the meta-ensemble.

3. Results

As the ML methods employed in this study are divided into the classes of: i) shallow learning methods based on fixed molecular fingerprints, ii) Graph-CNN utilizing self-learned representations and iii) methods based on combining i) and ii), we begin with a summary of the results based on all ML approaches, followed by dedicated subsections providing more details.

3.1. Results summary

For shallow learning models we have compared the ROC-AUC performance values under the evaluation scheme described in Section 2.8, based on all considered types of fingerprints listed in Section 2.4. We observe that 1024-bit circular fingerprints show the highest performance, as illustrated in Fig. 3, and hence we proceed in further experiments with this fixed fingerprint setup.

Different setups for Graph-CNN model compared in terms of ROC-AUC are shown in Fig. 4. First, we compared the default Graph-CNN with a version with tuned hyperparameters and additionally enforced equal number of active and inactive molecules in each batch (class balance). Then, we run experiments with ensembling 5 instances of the Graph-CNN model as well as versions with the following additional features, combined as described in Section 2.6.1: version with additional 200 RDKit features (following [46]), version with MACCS fingerprints as additional features and versions with 1024-bit and 2048-bit circular fingerprints as additional features. For generating final predictions using Graph-CNN we employ the version with tuned hyperparameters and class balance.

Comparing the performance of shallow learners to Graph-CNN and our methods based on combining fixed and Graph-CNN-learned representations we generally observe that after the above mentioned selection of fingerprint method, all models operate on roughly the same level of performance in terms of ROC-AUC, as shown in Fig. 5. However, although the aggregated performance in terms of ROC-AUC does not differ too much, we observe significant and interesting differences in characteristics on granular level of individual predictions, on which we comment further in the dedicated subsections below. The obtained ROC-AUC values for our base (non-combined) models are: LogReg 0.82 (±0.08), SVM 0.79 (±0.07), GBT 0.82 (±0.06), Graph-CNN 0.81 (±0.09). The corresponding ROC-AUC values for our final versions of combined models are: LogReg and GBT enhanced with Graph-CNN self-learned neural embeddings 0.82 (±0.08) and 0.79 (±0.06), respectively, and the stacking ensembles based on Graph-CNN and LogReg as well as Graph-CNN and GBT, both 0.82 (±0.07). The approach based on enhancing Graph-CNN with additional molecular features described in 2.6.1 was generally observed to result in somewhat lower (as well as more dispersed) average performance values with ROC-AUC ranging from 0.67 (±0.12) to 0.79 (±0.11), hence we drop this approach from the further analysis and proceed with approaches for combining Graph-CNN with fixed molecular embeddings described in Sections 2.6.2 and 2.6.3.

Generally, the predictions obtained by all considered models show certain characteristic homogeneity regarding both chemical and pharmacological classification, with prevalence of sulfonamides and anticancer drugs. Interestingly, based on all considered models, among the top 3 rank-ordered repurposing candidates for COVID-19, either sulfonamides or anticancer nitrogen mustard derivatives, or both, are identified.

Our top rank-ordered repurposing candidates obtained by LogReg, GBT and Graph-CNN models are provided in Tables 1–3, respectively. In particular, among repurposing candidates that were identified based on at least three ML approaches used in this study, are: famotidine and bosentan, which both belong to sulfonamide derivatives, as well as imatinib (benzanilides) and melphalan (alkylating nitrogen mustard), which both belong to anticancer drugs. Notably, the relationship of famotidine, bosentan and imatinib to the course of COVID-19 has been already discussed in the literature [54–56].

In the remainder of this section we report in more detail on the results obtained by the individual ML approaches.

3.2. Shallow learning on fixed representations

The results obtained with the LogReg model, listed in Table 1, show strong homogeneity in reference to both chemical and pharmacological classification, with predominant representation of sulfonamides[6] and anticancer drugs[7]. Moreover, LogReg identifies the three mentioned above drugs, already discussed in literature in the context of COVID-19: bosentan, famotidine, imatinib. In order to investigate further the characteristic chemical and pharmacological pattern that emerged in the predictions obtained by the LogReg model, we follow with the SVM model. The top rank-ordered predictions from the SVM model are also observed to be characterized by a certain dominance of sulfonamides[8] and anticancer[9] drugs, however to a lesser extent compared to the LogReg predictions. Moreover, apart from the already listed drugs identified by our models and already discussed in the literature in the context of COVID-19, additional drugs appearing both in our SVM results and already published studies are: rivaroxaban [57,58] and sildenafil [59]. In order to further verify the dominance of specific drug categories indicated by our shallow models, we proceed with the GBT results, shown in Table 2. Here, as deeper interaction effects between the molecular features are taken into account, we observe a greater chemical and pharmacological diversity among the obtained repurposing candidates for COVID-19. Beside the so far identified characteristic drug classes[10], the novel pharmacological groups discovered by the GBT model consist of macrolide antibiotics representation including azithromycin, clarithromycin and erythromycin as well as by vaborbactam, a beta-lactamase inhibitor. Among drugs which are already discussed in the literature in the context of COVID-19 and identified additionally by the GBT model are melphalan (clinical trial), ibrutinib [60], azithromycin, clarithromycin [61] as well as linagliptin, a dipeptidyl peptidase 4 (DPP-4) inhibitor for the treatment of type II diabetes [62].

3.3. Graph-CNN

Drug indications obtained by the Graph-CNN are generally consistent with those obtained with our shallow models based on fixed fingerprints, both in terms of chemical and pharmacological classification, with strong representation of sulfonamides and anticancer drugs, as shown in Table 3. Among results obtained with the Graph-CNN model, which have been already discussed in the literature in reference to COVID-19 are melphalan (clinical trial) and famotidine [54]. Interestingly, the novel repurposing candidates additionally indicated by the

[6] Sulfonamide derivatives are represented here by sulfadiazine, sulfasalazine, sulfanilamide and mafenide with antimicrobial activity as well as bumetanide and furosemide, belonging to diuretics, belinostat, an anticancer drug, famotidine, with antihistaminergic mechanism of action, bosentan, a dual endothelin receptor antagonist, as well as thiothixene, an antipsychotic agent.

[7] Anticancer drugs are represented here by ifosfamide and estramustine, which belong to the class of nitrogen mustard compounds, carmustine, belonging to nitrosoureas, belinostat, a sulfonamide derivative, imatinib, a benzanilides representative, lapatinib, a quinazolinamines derivative, as well as mitoxantrone, an antraquinone derivative.

[8] Beside the common indications which emerge both in SVM and LogReg results (11 drugs), additional sulfonamide derivatives are represented by lifitegrast, a lymphocyte function–associated antigen-1 antagonist and sildenafil, a selective inhibitor of cGMP specific phosphodiesterase type 5 (PDE5).

[9] Additional anticancer drug to those already indicated by LogReg model is panobinostat, a tryptamine derivative.

[10] Beside the common indications emerging both in GBT, SVM and LogReg results, additional sulfonamide derivative is tamsulosin, a selective alpha-1A and alpha-1B adrenoreceptor antagonist. Also selexipag, belonging to aminosulfonyl compounds, which is used for the treatment of pulmonary arterial hypertension (PAH) has been among the best repurposing candidates. Additional anticancer drug to those already indicated by LogReg and SVM models are melphalan and chlorambucil, belonging to nitrogen mustards as well as ibrutinib, a diphenylethers analogue.

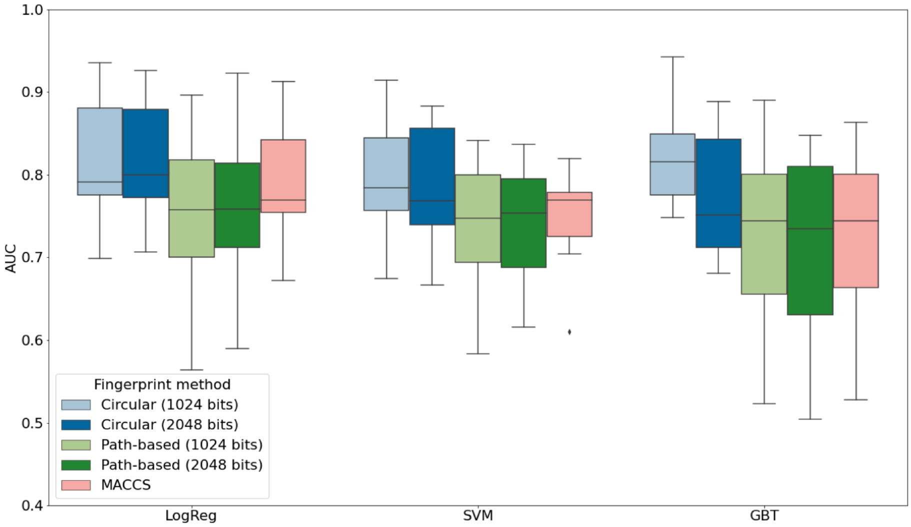


Fig. 3. Evaluation of shallow learning performance based on different versions of molecular fingerprint methods: boxplots of ROC-AUC based on 10 randomly seeded 80:20 scaffold splits of training and testing data.

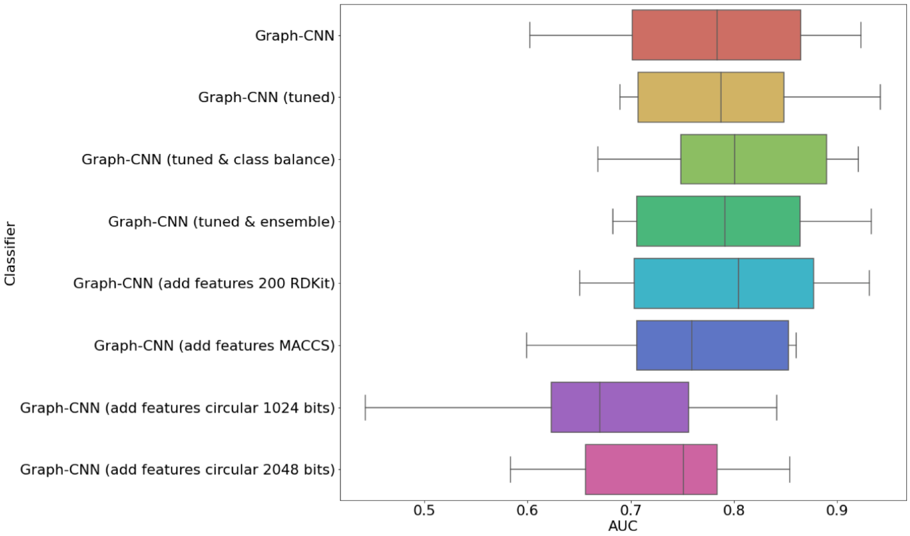


Fig. 4. Evaluation of different setups for a Graph-CNN model. The boxplots show ROC-AUC based on 10 randomly seeded 80:20 scaffold splits of training and testing data for the following configurations (from top to bottom): default Graph-CNN, version with tuned hyperparameters, version with tuned hyperparameters and an equal number of active and inactive molecules in each batch (class balance), an ensemble of 5 Graph-CNN models, version with additional 200 RDKit features [46], version with MACCS fingerprints as additional features, versions with 1024-bit and 2048-bit circular fingerprints as addi.tional features.

Graph-CNN model are 2-mercaptoethanesulfonic acid (coenzyme M) and thiosulfuric acid, which are both used together with anticancer drugs in order to reduce their toxicity.

3.4. Combining fixed and Graph-CNN self-learned representations

First, we report on the results based on LogReg and GBT models enhanced with Graph-CNN self-learned embeddings, as described in Section 2.6.2. These results remain generally consistent in terms of the prevalence of sulfonamides[11] and anticancer[12] drugs. Stacking ensembles combining Graph-CNN with the GBT and the LogReg models, respectively, as described in 2.6.3, give indications which are generally consistent with so far identified drug classes, encompassing sulfonamides and anticancer nitrogen mustard derivatives. Beside the common indications identified by both of considered stacking ensembles, including sulfadiazine, sulfasalazine, sulfanilamide, famotidine, belinostat and mafenide, both models add specific indications, which are consistent with previously found predictions referring to sulfonamides and anticancer drugs. These indications include: diclofenamide, chlorothiazide, hydrochlorothiazide furosemide, carmustine, ifosfamide, chlorambucil, melphalan, cyclophosphamide and imatinib obtained by the stacking ensemble of Graph-CNN and LogReg models, as well as bosentan, bumetanide, lapatinib, estramustine, mitoxantrone and ibrutinib, which are obtained with the stacking ensemble of Graph-CNN and GBT models.

An interesting observation based on the results from the combination of the Graph-CNN and GBT models in a stacking ensemble, is the identification of the following additional drugs, being either closely related to other similar derivatives with documented relation to COVID-19 outcome, or being explicitly discussed in this context: macitentan, an endothelin receptor antagonist, cytarabine, an antineoplastic anti-

[11] In comparison to the LogReg model, additional sulfonamide derivatives obtained after the enhancement with neural embeddings are: diclofenamide, chlorothiazide, hydrochlorothiazide, and belinostate, also an anticancer representative. In comparison to the GBT model, additional sulfonamide derivatives obtained after the enhancement by neural embeddings are: diclodfenamide, chlorothiazide and hydrochlorothiazide.

[12] Additional anticancer drugs obtained after enhancing the LogReg model with neural embeddings are: chlorambucil, cyclophosphamide and mitotane. In comparison to the GBT model, additional anticancer drugs obtained after the enhancement by neural embeddings are: carmustine, mafenide and mitotane.

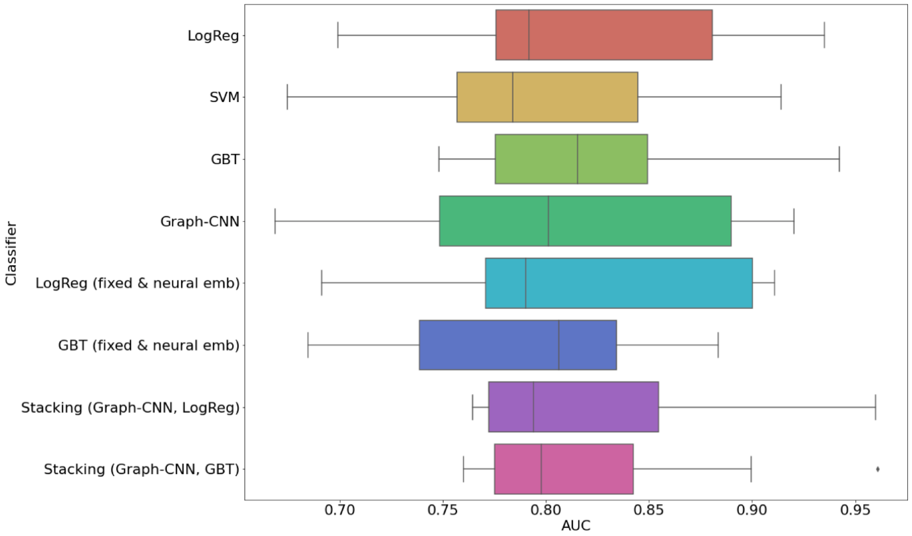


Fig. 5. Evaluation of the performance of shallow (LogReg, SVM, GBT), Graph-CNN and combined approaches based on concatenated fixed fingerprints with neural embeddings as well as stacking of Graph-CNN and shallow models: boxplots of ROC-AUC based on 10 randomly seeded 80:20 scaffold splits of training and testing data.

Table 1
Top repurposing candidates based on the screening generated using the regularized logistic regression model and fixed molecular fingerprint features. Next to IDs and drug names the corresponding compound classes and pharmacological classes are shown.

ChEMBL ID	Drug Name	Compound class	Pharmacological class
CHEMBL439	Sulfadiazine	aminobenzenesulfonamides	antimicrobial
CHEMBL421	Sulfasalazine	benzenesulfonamides	antiinflammatory/antimicrobial
CHEMBL58	Mitoxantrone	anthraquinones	anticancer
CHEMBL1987518	Ensulizole	organosulfonic acids	selective UV-B filter
CHEMBL554	Lapatinib	quinazolinamines	anticancer
CHEMBL1072	Bumetanide	benzenesulfonyl compounds	diuretics
CHEMBL21	Sulfanilamide	aminobenzenesulfonamides	antimicrobial
CHEMBL1024	Ifosfamide	nitrogen mustards	anticancer
CHEMBL419	Mafenide	benzenesulfonamides	antimicrobial
CHEMBL941	Imatinib	benzanilides	anticancer
CHEMBL957	Bosentan	benzenesulfonamides	anti-pulmonary hypertension
CHEMBL408513	Belinostat	benzenesulfonamides	anticancer
CHEMBL902	Famotidine	sulfonamide	antihistaminergic
CHEMBL3317857	Vaborbactam	oxaborine derivatives	beta-lactamase inhibitor
CHEMBL1575	Estramustine	nitrogen mustards	anticancer
CHEMBL35	Furosemide	aminobenzenesulfonamides	diuretics
CHEMBL1201	Thiothixene	organosulfonamides	antipsychotic agent
CHEMBL1198857	Vilanterol	benzylethers	selective β2-adrenergic agonist
CHEMBL513	Carmustine	nitrosoureas	anticancer

metabolite and doxycycline, a tetracycline antibiotic [63,64].

4. Discussion

The SARS-CoV-2 pandemic requires a fast-track of drug development as the course of the disease in many cases is still unpredictable and constitutes a great challenge. One may suggest a deep analysis of currently available potential pharmacological agents through artificial intelligence-enabled screening of chemical space with particular emphasis on repurposing candidates. The FDA approved drugs constitute a good starting point in the context of repurposing officially approved and safe drugs against COVID-19 [10].

The preliminary comparative analysis of the results obtained in our study reveals existence of two major classes of drugs, sharing similarities both in terms of chemical and pharmacological classification, namely the class of anticancer drugs and the class of sulfonamides. These two families of drugs are identified already by the logistic regression model, being the simplest of the considered models. Although the considered more sophisticated models lead to roughly the same level of aggregated performance in terms of ROC-AUC, we observe interesting differences on the level of individual predictions. In particular, due to automatically taken into account deeper interactions between the molecular features, the GBT model is able to discover structures characterized by higher variety, for example a new representation of antibiotics, from the group of macrolides consisting of azithromycin, clarithromycin and erythromycin. The macrolide antibiotics are supposed to improve the course of viral infections, at least through indirect mechanisms including anti-inflammatory or immunomodulatory (or both) effects, however there is no clear evidence of clinical efficacy of macrolides in coronaviruses infections till now [61].

At the forefront of anticancer drugs, melphalan, ifosfamide, cyclophosphamide, carmustine, estramustine, mechlorethamine (mustine) and chlorambucil, belonging to organic nitrogen compound or nitrogen mustard compounds were among the most commonly stated drugs in the majority of used by us models. Beside this group, also other anticancer drugs have been among the best indicated by our models, like diphenylmethanes (mitotane), as well as quinazolinamines (lapatinib), diphenylethers (ibrutinib) and anthraquinones (mitoxantrone). A

Table 2
Top repurposing candidates based on the screening generated using the GBT model and fixed molecular fingerprint features. Next to IDs and drug names the corresponding compound classes and pharmacological classes are shown.

ChEMBL ID	Drug Name	Compound class	Pharmacological class
CHEMBL852	Melphalan	nitrogen mustards	anticancer
CHEMBL33986	Butorphanol	phenanthrenes	analgesics
CHEMBL515	Chlorambucil	nitrogen mustards	anticancer
CHEMBL1575	Estramustine	nitrogen mustard	anticancer
CHEMBL554	Lapatinib	quinazolinamines	anticancer
CHEMBL3317857	Vaborbactam	oxaborine derivatives	beta-lactamase inhibitor
CHEMBL1873475	Ibrutinib	diphenylethers	anticancer
CHEMBL957	Bosentan	benzenesulfonamides	anti-pulmonary hypertension
CHEMBL2105395	Ospemifene	stilbenes	estrogen receptor modulator
CHEMBL21	Sulfanilamide	aminobenzenesulfonamides	antimicrobial
CHEMBL421	Sulfasalazine	benzenesulfonamides	antiinflammatory/antimicrobial
CHEMBL419	Mafenide	benzenesulfonamides	antimicrobial
CHEMBL494753	Estrone Sulfate	sulfated steroids	form of estrogen
CHEMBL439	Sulfadiazine	aminobenzenesulfonamides	antimicrobial
CHEMBL1118	O-desmethylvenlafaxine	cyclohexanols	antidepressant
CHEMBL659	Galantamine	alkaloids	cholinesterase inhibitors
CHEMBL585	Triamterene	pteridines	diuretic
CHEMBL237500	Linagliptin	xanthines	antidiabetics
CHEMBL529	Azithromycin	macrolides	macrolide antibiotic
CHEMBL238804	Selexipag	aminosulfonyl compounds	anti-pulmonary hypertension
CHEMBL1072	Bumetanide	benzenesulfonyl compounds	diuretics
CHEMBL560	Pentazocine	benzomorphans	analgesics
CHEMBL1741	Clarithromycin	macrolides	macrolide antibiotic
CHEMBL1071	Oxaprozin	oxazoles	NSAID
CHEMBL592	Levorphanol	morphinans	analgesic
CHEMBL880	Famciclovir	purines and purine derivatives	antivirals
CHEMBL532	Erythromycin	macrolides	macrolide antibiotic
CHEMBL836	Tamsulosin	benzenesulfonamides	adrenergic antagonist

representative of the nitrogen mustard compounds, melphalan, is under the 2nd phase clinical trial (NCT04380376) that evaluates the efficacy and safety of low-doses of melphalan in patients with pneumonia with confirmed or suspected COVID-19 infection. Similarly, another proposed by us top repurposing candidate, ibrutinib, is already being assessed in a clinical trial to treat COVID-19 patients (NCT04375397) [65]. Ibrutinib, a representative of novel anticancer drugs belonging to the group of the Bruton's tyrosine kinase (BTK) inhibitors is used to treat indolent B-cell malignancies, mantle cell lymphoma, chronic lymphocytic leukemia and chronic graft-versus-host disease (cGVHD) [66]. The inhibition of BTK pathway was assessed as a promising target to reduce the excessively severe immune response in case of COVID-19 [67] and ibrutinib was suggested to be protective against pulmonary injury in SARS-CoV-2 infected patients due to reducing production of proinflammatory and chemoattractant cytokines[13] [60]. It may be worth to mention, that ibrutinib has been among the best results indicated by the GBT and also by the version of GBT enhanced with neural embeddings.

This finding is even more intriguing, taking into account also another similar anticancer drug that was among the best predictions obtained with the LogReg, SVM as well as the stacking ensemble of Graph-CNN and LogReg, namely imatinib, with reported case of SARS-CoV-2 infection successfully treated with this drug [56]. Moreover, the safe and effective administration of these two kinase inhibitors in a patient with concomitant chronic myelogenous leukemia and chronic lymphocytic leukemia has been already reported [70]. Hence, these reports, put together with our result, compose an interesting compilation, drawing attention to the properties of some anticancer drugs that could be of value in case of oncological patients suffering from COVID-19.

Next to anticancer drugs, the second class of drugs with the strongest representation among the best repurposing candidates indentified by our models are sulfonamide derivatives. Sulfonamides are a significant and valuable pharmacological class, with diuretic, hypoglycemic, antithyroid, anticancer as well as antibacterial and antiviral activity. Among sulfonamides indicated by our models, the following subgroups are represented: antibacterial agents, diuretics, dual endothelin receptor antagonist used in the treatment of pulmonary arterial hypertension (PAH), a competitive histamine-2 (H2) receptor antagonist, as well as novel anticancer drugs. This observation is interesting, since the structural patterns of sulfonamides have been already used as a strategy to develop sulfonamide antivirals [71]. Notably, an antiviral activity of several sulfonamide derivatives, both *in vitro* and *in vivo*, has been already reported [71,72].

Specific sulfonamide derivatives, most commonly identified by our models are: sulfadiazine, sulfasalazine, sulfanilamide, dichlorphenamide, chlorothiazide, hydrochlorothiazide and mafenide. Another important sulfonamide identified as top repurposing candidate is bosentan, a dual endothelin receptor antagonist, which is approved for the treatment of pulmonary arterial hypertension (PAH) in New York Heart Association functional classification (NYHA) II-IV and in scleroderma patients. The drug blocks the action of endothelin molecules that promote narrowing of the blood vessels and lead to high blood pressure. Bosentan was suggested to be a treatment candidate for COVID-19 in association with other approved drugs [55], thanks to its potential of improving hemodynamics and prevention of lung fibrosis by reducing profibrotic and proinflammatory cytokines, like IL-2, IL-6, IL-8 and IFN-γ levels.

[13] Moreover, it may have an additional potential of modulating T cells through targeting IL-2-inducible T-cell kinase (ITK), as being a (BTK)/ITK dual inhibitor. As it is known, that SARS-CoV-2 infection triggers lymphocyte apoptosis, which may be associated with the severity of the disease, targeting signaling pathways in T lymphocytes, especially those that preferentially regulate T cell apoptosis and exhaustion over activation, may be a potential strategy for treating patients with severe COVID-19 [65]. It is known, that patients with chronic lymphocytic leukemia treated with ibrutinib displayed an increase in total number of T cells, what may be due to the decrease in activation-induced cell death, which has been shown to be a result of the ITK signaling-mediated up-regulation of Fas ligand (FasL), which promotes activation-induced T cell death [68]. As ITK is highly expressed in T cells and regulates the activation and function of both CD4 + and CD8 + T cells, inhibition of this mechanism may be of significance in case of COVID-19 patients [65] This observation may be even more interesting in the context of recent reports about the high severity of infection in patients with haematological malignancies suffering from COVID-19 [69].

Table 3
Top repurposing candidates based on the screening generated using the Graph-CNN model. Next to IDs and drug names the corresponding compound classes and pharmacological classes are shown.

ChEMBL ID	Drug Name	Compound class	Pharmacological class
CHEMBL17	Diclofenamide	benzenesulfonamides	carbonic anhydrase inhibitor
CHEMBL513	Carmustine	nitrosoureas	anticancer
CHEMBL1670	Mitotane	diphenylmethanes	anticancer
CHEMBL842	Chlorothiazide	organosulfonamides	diuretics
CHEMBL435	Hydrochlorothiazide	organosulfonamides	diuretics
CHEMBL427	Mechlorethamine	nitrogen mustards	anticancer
CHEMBL1201798	Sevelamer	epoxides	phosphate binding drug
CHEMBL130	Chloramphenicol	nitrobenzenes	antimicrobial
CHEMBL88	Cyclophosphamide	nitrogen mustards	anticancer
CHEMBL515	Chlorambucil	nitrogen mustards	anticancer
CHEMBL1577	Methyclothiazide	organosulfonamides	diuretics
CHEMBL21	Sulfanilamide	aminobenzenesulfonamides	antimicrobial
CHEMBL1670	Mitotane	diphenylmethanes	antiinflammatory/anticancer
CHEMBL1098319	Coenzyme M	sulfhydryl (thiol) compound	uroprotective agent
CHEMBL1024	Ifosfamide	nitrogen mustards	anticancer
CHEMBL419	Mafenide	benzenesulfonamides	antimicrobial
CHEMBL852	Melphalan	nitrogen mustards	anticancer
CHEMBL902	Famotidine	sulfonamide	antihistaminergic
CHEMBL1208642	Thiosulfuric Acid	sulfated steroids	adjunct agent for chemotherapy
CHEMBL1373	Modafinil	diphenylmethanes	stimulants
CHEMBL1043	Dapsone	benzenesulfonyl compounds	antimicrobial
CHEMBL439	Sulfadiazine	aminobenzenesulfonamides	antimicrobial
CHEMBL239243	Taurine	sulfonyls	cholinesterase inhibitors

We would also like to underline that another promising candidate identified among top repurposing candidates is famotidine, a competitive histamine-2 (H2) receptor antagonist. This drug in combination with cetirizine, has been already stated to be safe and effective method to reduce the progression in symptom severity in COVID-19 patients, presumably by minimizing the histamine-mediated cytokine storm [54]. Moreover, results of case series suggest that the use of high-dose oral famotidine is associated with improved patient-reported outcomes in non-hospitalised patients with COVID-19 [73].

Furthermore, the combination of famotidine with doxycycline, which is also among the candidates proposed by our stacking ensemble model, has been already reported to state a valuable combination that may provide robust chemoprophylaxis effective against COVID-19 [74].

Another illustration of the additional value of exploring ML approaches operating on molecular features in various ways, going beyond the logistic regression based on fixed fingerprints, may be seen in the identification of a group of drugs acting on respiratory tract, possibly of particular significance for the treatment of COVID-19. In this group, a potential antiviral activity of the following drugs was identified: bosentan, a dual endothelin receptor antagonist used in the treatment of pulmonary arterial hypertension (PAH), vilanterol, a selective long-acting β2-adrenergic agonist (LABA), selexipag, as well as macitentan, both indicated for the treatment of PAH. A further interesting drug found after segmenting the molecular feature space by the GBT model is linagliptin, a DPP-4 inhibitor used for the treatment of type II diabetes. Linagliptin has been already indicated as a promising inhibitor of main protease of SARS-CoV-2 [75] and its potential to reduce inflammation, thus possibly minimizing the risk for COVID-19 severity, has been also discussed [62].

Interestingly, the presence of sulfonamide structures in the top ranked repurposing candidates predicted by our models, corresponds well with the results obtained by us in a similar study [76], where zafirlukast, containing also a sulfonamide group, was proposed to be the best potential repurposing candidate for COVID-19.

Also, a representation of drugs acting on circulatory and cardiovascular systems has been identified by our study, which is particularly important regarding the specific implications of COVID-19 on patients underlying heart diseases. Here, we report on sildenafil as well as losartan as potential repurposing candidates. Sildenafil, a phosphodiesterase-5 (PDE5) inhibitor, acting by minimising the breakdown of cyclic guanosine monophosphate (cGMP) was suggested to counter the inflammatory cascade and thromboembolic episodes that occur in COVID-19 [59]. Moreover, a pair of drugs identified in our study as potential repurposing candidates, namely sildenafil and ivermectin, were suggested to be beneficial in COVID-19 patients [58].

Regarding the utility of the explored ML techniques, our observations confirm that in particular in the context of limited data sizes both shallow and deep models have their benefits and limitations. While shallow learners rely on pre-computed, fixed molecular representations, suitably designed shallow ML pipelines can provide more stable predictions, compared to the deep learning approaches based on self-learned embeddings. Among the considered types of fixed fingerprints, we observe that for the predictive task at hand the hash-based circular fingerprints showed superior performance and that due to the limited data size the number of bits should not be too high. Although in terms of aggregated performance the considered paradigms lead to comparable level of ROC-AUC, we observe interesting differences on granular level of individual predictions driven by different molecular representations and subsequent transformation steps. Moreover, both deep models utilizing self-learned molecular embeddings as well as shallow models taking into account higher-order feature interactions are observed to lead to an increased models' capacity to identify repurposing candidates for COVID-19 of higher chemical and pharmacological diversity.

While the drug indications presented in this study constitute an interesting discovery, in order to conduct a more in-depth exploration of the predictive performance and reliability of AI-based screenings, the utilization of additional and larger datasets would be necessary. Moreover, exploration of alternative ML approaches and methodologies for representing molecular structures could lead to further interesting insights. Lastly, validating *in vitro* and *in vivo* experiments and clinical trials need to be performed as a next step to ensure the efficacy and other desired properties of the proposed drugs.

5. Conclusions

In this study we explored ML approaches for identification of potential SARS-CoV-2 3CLpro inhibitors based on fixed molecular fingerprints and Graph-CNN neural representations. The resulting drug discovery screenings indicate relevance of two major classes of drugs: sulfonamide derivatives and anticancer drugs. Both these classes emerge from the performed screenings in the form of drug candidates that share similarities both in terms of chemical and pharmacological classification. Moreover, the consideration of multiple different approaches, varying both in terms of employed types of molecular representations as well as learning algorithms transforming them, enables identification of structures characterized by higher variety, resulting in a selection of interesting groups of drugs which are in accordance with already published studies on COVID-19 medications.

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.

Acknowledgements

We are grateful to Simone Vettorel for the insightful discussions and invaluable comments.

References

[1] J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer, et al., Applications of machine learning in drug discovery and development, Nat. Rev. Drug Discovery 18 (2019) 463–477.

[2] A. Zumla, D.S. Hui, E.I. Azhar, Z.A. Memish, M. Maeurer, Reducing mortality from 2019-ncov: host-directed therapies should be an option, The Lancet 395 (2020) e35–e36.

[3] Y. Zhou, F. Wang, J. Tang, R. Nussinov, F. Cheng, Artificial intelligence in COVID-19 drug repurposing, Lancet Digit Health (2020).

[4] Y.Y. Ke, T.T. Peng, T.K. Yeh, W.Z. Huang, S.E. Chang, S.H. Wu, H.C. Hung, T. A. Hsu, S.J. Lee, J.S. Song, W.H. Lin, T.J. Chiang, J.H. Lin, H.K. Sytwu, C.T. Chen, Artificial intelligence approach fighting COVID-19 with repurposing drugs, Biomed J 43 (2020) 355–362.

[5] K. Gao, D.D. Nguyen, J. Chen, R. Wang, G.W. Wei, Repositioning of 8565 Existing Drugs for COVID-19, J Phys Chem Lett 11 (2020) 5373–5382.

[6] B.R. Beck, B. Shin, Y. Choi, S. Park, K. Kang, Predicting commercially available antiviral drugs that may act on the novel coronavirus (SARS-CoV-2) through a drug-target interaction deep learning model, Comput Struct Biotechnol J 18 (2020) 784–790.

[7] J.M. Levin, T.I. Oprea, S. Davidovich, T. Clozel, J.P. Overington, Q. Vanhaelen, C. R. Cantor, E. Bischof, A. Zhavoronkov, Artificial intelligence, drug repurposing and peer review, Nat Biotechnol 38 (2020) 1127–1131.

[8] H. Zhang, K.M. Saravanan, Y. Yang, M.T. Hossain, J. Li, X. Ren, Y. Pan, Y. Wei, Deep Learning Based Drug Screening for Novel Coronavirus 2019-nCov, Interdiscip Sci 12 (2020) 368–376.

[9] S. Mohanty, M. Harun Ai Rashid, M. Mridul, C. Mohanty, S. Swayamsiddha, Application of Artificial Intelligence in COVID-19 drug repurposing, Diabetes Metab Syndr 14 (2020) 1027–1031.

[10] M. Kandeel, M. Al-Nazawi, Virtual screening and repurposing of FDA approved drugs against COVID-19 main protease, Life Sci 251 (2020) 117627.

[11] V. Parvathaneni, V. Gupta, Utilizing drug repurposing against covid-19–efficacy, limitations, and challenges, Life Sci. (2020) 118275.

[12] World Health Organization, Coronavirus disease (covid-19) weekly epidemiological update and weekly operational update., World Health Organization Weekly Epidemiological Update and Weekly Operational Update (2020). https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports.

[13] C. Wang, P.W. Horby, F.G. Hayden, G.F. Gao, A novel coronavirus outbreak of global health concern, The Lancet 395 (2020) 470–473.

[14] D. Kim, J.-Y. Lee, J.-S. Yang, J.W. Kim, V.N. Kim, H. Chang, The architecture of sars-cov-2 transcriptome, Cell (2020).

[15] D. Wrapp, D. De Vlieger, K.S. Corbett, G.M. Torres, N. Wang, W. Van Breedam, K. Roose, L. van Schie, V.-C. COVID, M. Hoffmann, et al., Structural basis for potent neutralization of betacoronaviruses by single-domain camelid antibodies, Cell (2020).

[16] S.S. Cherian, M. Agrawal, A. Basu, P. Abraham, R.R. Gangakhedkar, B. Bhargava, et al., Perspectives for repurposing drugs for the coronavirus disease 2019, Indian J. Med. Res. 151 (2020) 160.

[17] C.H. Wong, K.W. Siah, A.W. Lo, Estimation of clinical trial success rates and related parameters, Biostatistics 20 (2019) 273–286.

[18] J.A. Vernon, J.H. Golec, J.A. DiMasi, Drug development costs when financial risk is measured using the fama–french three-factor model, Health economics 19 (2010) 1002–1005.

[19] L. Zhang, D. Lin, X. Sun, U. Curth, C. Drosten, L. Sauerhering, S. Becker, K. Rox, R. Hilgenfeld, Crystal structure of sars-cov-2 main protease provides a basis for design of improved α-ketoamide inhibitors, Science 368 (2020) 409–412.

[20] J. Kowalewski, A. Ray, Predicting novel drugs for sars-cov-2 using machine learning from a $>$ 10 million chemical space, Heliyon 6 (2020) e04639.

[21] A.K. Verma, R. Aggarwal, Repurposing potential of fda approved and investigational drugs for covid-19 targeting sars-cov-2 spike and main protease and validation by machine learning algorithm, Chemical biology & drug design (2020).

[22] Diamond Light Source group., Main protease structure and xchem fragment screen, Diamond Light Source. Harwell Science and Innovation Campus, Oxfordshire (2020). https://www.diamond.ac.uk/covid-19.html.

[23] A.R. Kinjo, G.-J. Bekker, H. Wako, S. Endo, Y. Tsuchiya, H. Sato, H. Nishi, K. Kinoshita, H. Suzuki, T. Kawabata, et al., New tools and functions in data-out activities at protein data bank japan (pdbj), Protein Sci. 27 (2018) 95–102.

[24] A.R. Kinjo, G.-J. Bekker, H. Suzuki, Y. Tsuchiya, T. Kawabata, Y. Ikegawa, H. Nakamura, Protein data bank japan (pdbj): updated user interfaces, resource description framework, analysis tools for large structures, Nucleic Acids Research (2016) gkw962.

[25] J.Y. Young, J.D. Westbrook, Z. Feng, R. Sala, E. Peisach, T.J. Oldfield, S. Sen, A. Gutmanas, D.R. Armstrong, J.M. Berrisford, et al., Onedep: unified wwpdb system for deposition, biocuration, and validation of macromolecular structures in the pdb archive, Structure 25 (2017) 536–545.

[26] Z. Jin, X. Du, Y. Xu, Y. Deng, M. Liu, Y. Zhao, B. Zhang, X. Li, L. Zhang, C. Peng, et al., Structure of m pro from sars-cov-2 and discovery of its inhibitors, Nature (2020) 1–5.

[27] X. Xue, H. Yang, W. Shen, Q. Zhao, J. Li, K. Yang, C. Chen, Y. Jin, M. Bartlam, Z. Rao, Production of authentic sars-cov mpro with enhanced activity: application as a novel tag-cleavage endopeptidase for protein overproduction, Journal of molecular biology 366 (2007) 965–975.

[28] A. Douangamath, D. Fearon, P. Gehrtz, T. Krojer, P. Lukacik, C.D. Owen, E. Resnick, C. Strain-Damerell, P. Ábrányi-Balogh, J. Brandaõ-Neto, et al., Crystallographic and electrophilic fragment screening of the sars-cov-2 main protease, bioRxiv (2020).

[29] T. Sterling, J.J. Irwin, Zinc 15–ligand discovery for everyone, Journal of chemical information and modeling 55 (2015) 2324–2337.

[30] M.S. Kinch, A. Haynesworth, S.L. Kinch, D. Hoyer, An overview of fda-approved new molecular entities: 1827–2013, Drug discovery today 19 (2014) 1033–1039.

[31] J. Shen, C.A. Nicolaou, Molecular property prediction: recent trends in the era of artificial intelligence, Drug Discovery Today: Technologies (2020).

[32] D. Weininger, Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules, Journal of chemical information and computer sciences 28 (1988) 31–36.

[33] R.P. Swanson, The entrance of informatics into combinatorial chemistry, in: The History and Heritage of Scientific and Technological Information Systems: Proceedings of the 2002 Conference, Medford, New Jersey, 2004, pp. 203–211.

[34] I. Goodfellow, Y. Bengio, A. Courville, Y. Bengio, Deep learning, volume 1, MIT press Cambridge, 2016.

[35] K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea, A. Palmer, V. Settels, T. Jaakkola, K. Jensen, R. Barzilay, Analyzing Learned Molecular Representations for Property Prediction, J Chem Inf Model 59 (2019) 3370–3388.

[36] J.L. Durant, B.A. Leland, D.R. Henry, J.G. Nourse, Reoptimization of mdl keys for use in drug discovery, Journal of chemical information and computer sciences 42 (2002) 1273–1280.

[37] Daylight theory: Fingerprints (Accessed Dec 2020). https://www.daylight.com/dayhtml/doc/theory/theory.finger.html.

[38] H.L. Morgan, The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service, Journal of Chemical Documentation 5 (1965) 107–113.

[39] T. Hastie, R. Tibshirani, J. Friedman, The elements of statistical learning: data mining, inference, and prediction, Springer Science & Business Media, 2009.

[40] C. Cortes, V. Vapnik, Support-vector networks, Machine learning 20 (1995) 273–297.

[41] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794.

[42] M. Gori, G. Monfardini, F. Scarselli, A new model for learning in graph domains, in: Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005, volume 2, IEEE, 2005, pp. 729–734.

[43] M.M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, P. Vandergheynst, Geometric deep learning: going beyond euclidean data, IEEE Signal Process. Mag. 34 (2017) 18–42.

[44] J. Gilmer, S.S. Schoenholz, P.F. Riley, O. Vinyals, G.E. Dahl, Neural message passing for quantum chemistry, arXiv preprint arXiv:1704.01212 (2017).

[45] H. Dai, B. Dai, L. Song, Discriminative embeddings of latent variable models for structured data, in: in: International conference on machine learning, 2016, pp. 2702–2711.

[46] K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea, et al., Analyzing learned molecular representations for property prediction, Journal of chemical information and modeling 59 (2019) 3370–3388.

[47] P. Mahé, N. Ueda, T. Akutsu, J.-L. Perret, J.-P. Vert, Extensions of marginalized graph kernels, in: Proceedings of the twenty-first international conference on Machine learning, 2004, p. 70.

[48] D. Koller, N. Friedman, Probabilistic graphical models: principles and techniques, MIT press, 2009.

[49] D.H. Wolpert, Stacked generalization, Neural networks 5 (1992) 241–259.

[50] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[51] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014 arXiv preprint arXiv:1412.6980.

[52] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in neural information processing systems, 2017, pp. 5998–6008.

[53] Z. Wu, B. Ramsundar, E.N. Feinberg, J. Gomes, C. Geniesse, A.S. Pappu, K. Leswing, V. Pande, Moleculenet: a benchmark for molecular machine learning, Chemical science 9 (2018) 513–530.

[54] R.B. Hogan Ii, R.B. Hogan Iii, T. Cannon, M. Rappai, J. Studdard, D. Paul, T. P. Dooley, Dual-histamine receptor blockade with cetirizine-famotidine reduces pulmonary symptoms in covid-19 patients, Pulmonary Pharmacology & Therapeutics 63 (2020) 101942.

[55] S. Javor, A. Salsano, Why not consider an endothelin receptor antagonist against sars-cov-2? Med. Hypotheses 141 (2020) 109792.

[56] A. Morales-Ortega, D. Bernal-Bello, C. Llarena-Barroso, B. Frutos-Pérez, M.Á. Duarte-Millán, V.G. de Viedma-García, A.I. Farfán-Sedano, E. Canalejo-Castrillero, J.M. Ruiz-Giardín, J. Ruiz-Ruiz, et al., Imatinib for covid-19: a case report, Clinical Immunology (Orlando, Fla.) (2020).

[57] A. Fischer, M. Sellner, S. Neranjan, M. Smieško, M.A. Lill, Potential inhibitors for novel coronavirus protease identified by virtual screening of 606 million compounds, Int. J. Mol. Sci. 21 (2020) 3626.

[58] R.I. Horowitz, P.R. Freeman, Three novel prevention, diagnostic and treatment options for covid-19 urgently necessitating controlled randomized trials, Med. Hypotheses (2020) 109851.

[59] L. Mario, M. Roberto, L. Marta, C.M. Teresa, M. Laura, Hypothesis of covid-19 therapy with sildenafil, International journal of preventive medicine 11 (2020).

[60] S.P. Treon, J. Castillo, A.P. Skarbnik, J.D. Soumerai, I.M. Ghobrial, M.L. Guerrera, K.E. Meid, G. Yang, The btk-inhibitor ibrutinib may protect against pulmonary injury in covid-19 infected patients, Blood (2020).

[61] D. Poddighe, M. Aljofan, Clinical evidences on the antiviral properties of macrolide antibiotics in the covid-19 era and beyond, Antiviral Chem. Chemother. 28 (2020), 2040206620961712.

[62] N. Katsiki, E. Ferrannini, Anti-inflammatory properties of antidiabetic drugs: a "promised land" in the covid-19 era? Journal of Diabetes and its Complications 107723 (2020).

[63] P.A. Yates, S.A. Newman, L.J. Oshry, R.H. Glassman, A.M. Leone, E. Reichel, Doxycycline treatment of high-risk covid-19-positive patients with comorbid pulmonary disease, Therapeutic advances in respiratory disease 14 (2020), 1753466620951053.

[64] A. Farouk, S. Salman, Dapsone and doxycycline could be potential treatment modalities for covid-19, Medical hypotheses 140 (2020) 109768.

[65] M.C. McGee, A. August, W. Huang, Btk/itk dual inhibitors: Modulating immunopathology and lymphopenia for covid-19 therapy, J. Leukoc. Biol. (2020).

[66] T. Wen, J. Wang, Y. Shi, H. Qian, P. Liu, Inhibitors targeting bruton's tyrosine kinase in cancers: drug development advances, Leukemia (2020) 1–21.

[67] S. Thibaud, D. Tremblay, S. Bhalla, B. Zimmerman, K. Sigel, J. Gabrilove, Protective role of bruton tyrosine kinase inhibitors in patients with chronic lymphocytic leukaemia and covid-19, Br. J. Haematol. (2020).

[68] M. Long, K. Beckwith, P. Do, B.L. Mundy, A. Gordon, A.M. Lehman, K.J. Maddocks, C. Cheney, J.A. Jones, J.M. Flynn, et al., Ibrutinib treatment improves t cell number and function in cll patients, The Journal of clinical investigation 127 (2017) 3052–3064.

[69] V. Mehta, S. Goel, R. Kabarriti, D. Cole, M. Goldfinger, A. Acuna-Villaorduna, K. Pradhan, R. Thota, S. Reissman, J.A. Sparano, et al., Case fatality rate of cancer patients with covid-19 in a new york hospital system, Cancer discovery (2020).

[70] L.K. Shea, F.M. Mikhail, A. Forero-Torres, R.S. Davis, Concomitant imatinib and ibrutinib in a patient with chronic myelogenous leukemia and chronic lymphocytic leukemia, Clinical Case Reports 5 (2017) 899.

[71] R.N. Dash, A.K. Moharana, B.B. Subudhi, Sulfonamides: Antiviral strategy for neglected tropical disease virus, Curr. Org. Chem. 24 (2020) 1018–1041.

[72] C.T. Supuran, A. Innocenti, A. Mastrolorenzo, A. Scozzafava, Antiviral sulfonamide derivatives, Mini reviews in medicinal chemistry 4 (2004) 189–200.

[73] T. Janowitz, E. Gablenz, D. Pattinson, T.C. Wang, J. Conigliaro, K. Tracey, D. Tuveson, Famotidine use and quantitative symptom tracking for covid-19 in non-hospitalised patients: a case series, Gut (2020).

[74] P.S. Sen Gupta, M.K. Rana, Ivermectin, famotidine, and doxycycline: A suggested combinatorial therapeutic for the treatment of covid-19, ACS Pharmacology & Translational, Science 3 (2020) 1037–1038.

[75] H. Qu, Y. Zheng, Y. Wang, H. Li, X. Liu, X. Xiong, L. Zhang, J. Gu, G. Yang, Z. Zhu, et al., The potential effects of clinical antidiabetic agents on sars-cov-2, Journal of diabetes (1753).

[76] M. Delijewski, J. Haneczok, AI drug discovery screening for covid-19 reveals zafirlukast as a repurposing candidate, Medicine, Drug Discovery (2020) 100077.