Journal Pre-proof



Machine learning in antiviral drug design

Anja Kolarič, Marko Jukič, Urban Bren

PII: S0968-0896(25)00367-0

DOI: https://doi.org/10.1016/j.bmc.2025.118426

Reference: BMC 118426

To appear in:

Received date: 28 July 2025

Revised date: 10 September 2025

Accepted date: 29 September 2025

Please cite this article as: A. Kolarič, M. Jukič and U. Bren, Machine learning in antiviral drug design, (2024), https://doi.org/10.1016/j.bmc.2025.118426

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2025 Published by Elsevier Ltd.

Journal Pre-proof

Machine Learning in Antiviral Drug Design

Anja Kolarič[a], Marko Jukič[a,b,*], Urban Bren[a,b,c,*]

[a] Laboratory of Physical Chemistry and Chemical Thermodynamics, Faculty of Chemistry and Chemical Engineering, University of Maribor, Smetanova ulica 17, SI-2000 Maribor, Slovenia

[b] Faculty of Mathematics, Natural Sciences and Information Technologies, University of Primorska, Glagoljaška ulica 8, SI-6000 Koper, Slovenia

[c] Institute of Environmental Protection and Sensors, Beloruska ulica 7, SI-2000 Maribor, Slovenia

---

Abstract

Viral infections pose a significant health threat worldwide. Due to the high mutation rates of many viruses and their reliance on host cellular machinery, the development of effective antiviral therapies is particularly difficult. As a result, only a limited number of antiviral agents is currently available. In parallel to modern vaccines, traditional antiviral drug development is both time-consuming and costly, underscoring the need for faster, more efficient approaches. In recent years, particularly since the beginning of the COVID-19 pandemic, machine learning (ML) together with broader artificial intelligence (AI), have emerged as powerful methodologies for drug discovery and offer the potential to accelerate the identification and development of antiviral agents. This review examines the application of ML in the early stages of antiviral drug discovery, with a particular focus on recent studies where ML methods have successfully identified hit compounds with experimentally demonstrated activity in biological assays. By highlighting these successful case studies, the review illustrates the growing impact of ML in advancing the discovery of urgently needed novel antivirals.

Keywords: machine learning, artificial intelligence, antiviral compounds, biological activity

Journal Pre-proof

1. Introduction

Viral infections represent a major global public health challenge. Although hundreds of pathogenic viruses pose threats to human health, effective antiviral drugs are available for only a limited number of them.[1] Developing antivirals is particularly difficult because viruses depend on the host cellular machinery for replication, making it challenging to identify drug targets that can selectively inhibit the virus without damaging host cells. This lack of selective toxicity remains a significant obstacle in the antiviral drug development.[2] A further complication in the development of antiviral drugs represents the high mutation rate of viruses, which can lead to treatments becoming less effective or even completely ineffective. Traditional antiviral drug discovery and development is both time-consuming and expensive, factors that become especially critical when facing rapidly mutating viruses or emerging pathogens for which immediate therapeutic options are needed.[3] This issue has been underscored by the outbreaks of highly pathogenic viruses such as Zika, Ebola, and severe acute respiratory syndrome (SARS, SARS-CoV-2) virus, which have caused substantial morbidity and mortality, yet are still lacking in the number of approved, effective antiviral treatments.[1] These limitations highlight the urgent need for faster, more efficient approaches to antiviral drug development. The latter has been clearly demonstrated by the COVID-19 pandemic that started in late 2019.

In recent years, artificial intelligence (AI) has emerged as a transformative technology in drug discovery, becoming an integral component of modern pharmaceutical research. Machine learning (ML) in particular has attracted great interest due to its potential to accelerate drug discovery and reduce costs. ML can be used for various tasks, such as identifying drug candidates, repurposing existing drugs, and developing new compounds with desired properties. By analyzing large and complex datasets, ML models are capable of predicting biological activity, off-target effects, and interactions with specific protein targets. Despite the increasing number of studies incorporating ML into drug discovery pipelines, many of these efforts remain at the theoretical stage. In most cases, predicted hit compounds have not been validated in biological assays, limiting their further practical utility.[4-6] The outbreak of the COVID-19 pandemic in 2019 has significantly accelerated the application of ML in drug discovery, particularly in the search for antiviral therapies against SARS-CoV-2.[7] Numerous reviews have collected SARS-CoV-2 antiviral agents identified with ML.[7-14] At the same time, the application of ML has expanded to the discovery of other viral and non-viral drugs.

Given the urgent need for effective antiviral agents and the growing integration of ML into drug discovery, this review investigates antiviral compounds that have been identified or developed using ML techniques. To demonstrate the practical utility of ML in identifying biologically active antiviral compounds, we compiled and summarized studies in which ML-based methods led to the discovery of antiviral agents that were subsequently validated through biological experiments. We discuss the ML approaches employed in these studies, the types of viral targets addressed, and the outcomes of experimental evaluations. While much of the recent ML-assisted antiviral drug discovery has focused on agents targeting SARS-CoV-2, we also included studies involving antiviral candidates for other viral pathogens. A related review published in 2022 examined ML-identified hit compounds that were biologically evaluated across a range of diseases, including viral infections.[15] Therefore, our review focuses specifically on antiviral hit compounds identified using ML from 2022 onward, including the ones reported in 2022. To support readers in the

understanding of the methodologies, we also provide a concise overview of the most commonly used ML techniques in drug discovery.

2. Machine Learning (ML)

ML represents a subfield of AI that involves the development of algorithms capable of learning from large datasets by identifying patterns and relationships. These learned patterns are then used to make predictions or informed decisions on new, unseen data.[8] ML approaches can be broadly categorized based on the degree of human involvement required, particularly in feature engineering (Figure 1). Conventional ML models typically rely on a significant human input to manually extract and select relevant features from the data prior to training. In contrast, deep learning automates the feature extraction process through the application of multi-layered artificial neural networks. This allows deep learning models to learn hierarchical and complex representations directly from the raw data, reducing the need for manual preprocessing. Compared to traditional ML methods, deep learning models are often better suited for capturing complex, non-linear relationships in large and high-dimensional datasets, which can lead to the improved predictive performance.[14]

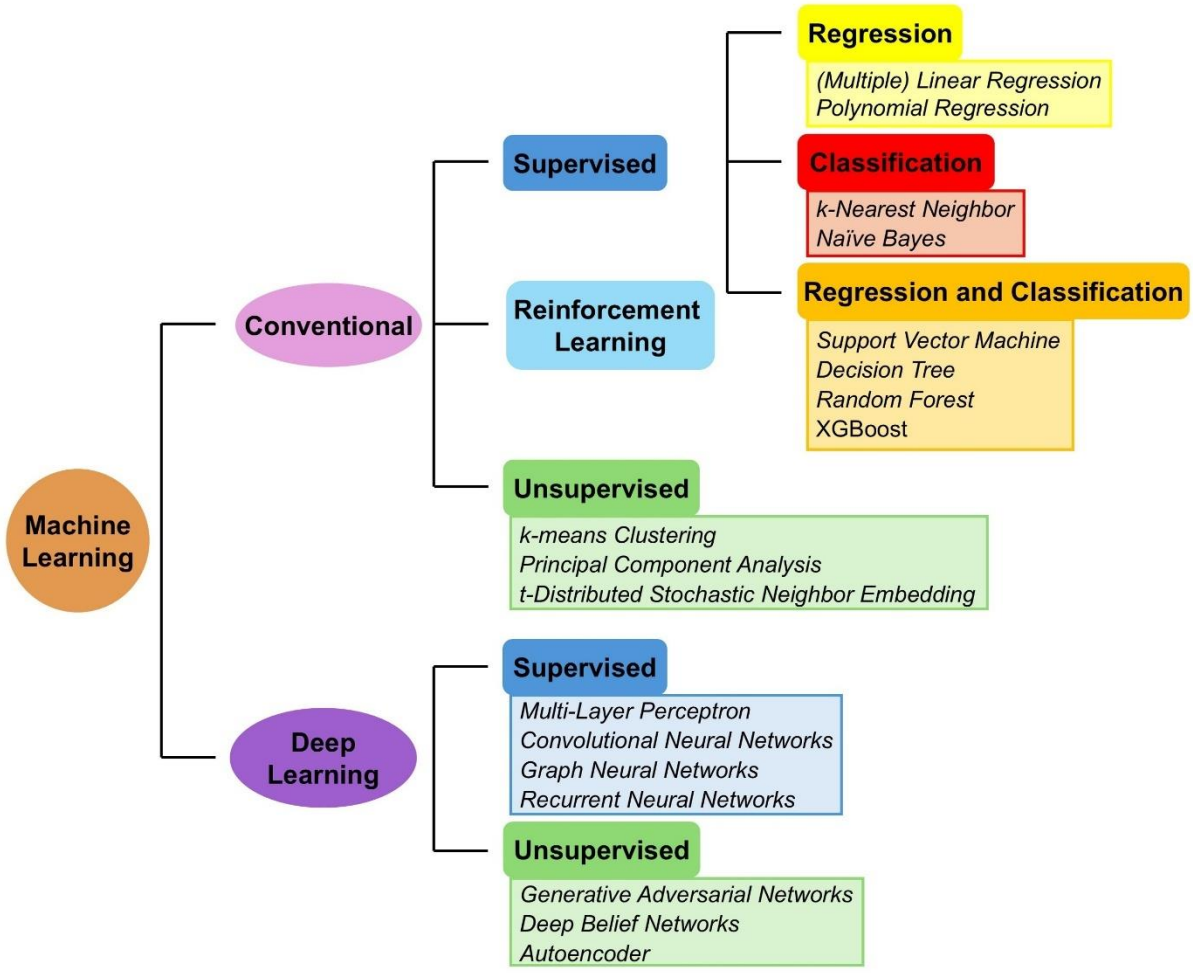


Figure 1. Graphical overview of the most common ML methods in drug discovery.

2.1. Conventional Machine Learning Methods

Traditional ML models operate by analyzing input data to detect patterns and generate predictive models based on them. The performance of such models is highly dependent on the quality of data preprocessing and the relevance of the features extracted, a process that typically requires a substantial human expertise. Effective feature engineering is crucial for facilitating the model to capture meaningful relationships within the data. In general, the performance of ML models improves with larger training datasets and appropriate training durations. More data provides a broader basis for identifying generalizable patterns, while sufficient training time allows the model to optimize its internal parameters for more accurate predictions. This capacity for iterative learning and improvement represents a core strength of ML, allowing models to adapt over time and to perform effectively in dynamic or evolving environments.[4,5,7,8] ML algorithms are typically categorized into three main types of learning paradigms: supervised, unsupervised, and reinforcement learning.[4,5,7,16] In supervised learning, the model is trained on labeled data. In unsupervised learning, the model works with unlabeled data and seeks to uncover hidden structures or groupings within the dataset.[5,7] Reinforcement learning, in contrast, involves an agent learning to make decisions through trial and error, guided by a system of rewards and penalties based on its actions within a defined environment.[4,5]

2.1.1. Supervised Learning

In supervised learning, a ML model is developed and trained using input data that include corresponding output labels. The model analyzes patterns within these input-output pairs and once trained on a sufficient volume of data, can make accurate predictions on unseen inputs. [5,7,14,16] Supervised learning can be broadly categorized into two main tasks: regression and classification.[4,7,16] Regression models are applied to predict continuous numerical values for new data points based on patterns learned from the training data. Classification models, on the other hand, are used to assign new data points to predefined categories or classes. These models learn discriminative features from the input data during training and apply this knowledge to classify new observations accurately.[7,16]

2.1.1.1. Overview of the Most Common Regression Algorithms in Antiviral Design Literature

The most common applications of regression algorithms include the prediction of biological and chemical features, such as biological activity, physicochemical properties, toxicity predictions, ADMET (absorption, distribution, metabolism, excretion, and toxicity) predictions, etc.[7,17] Typical applications of ML algorithms in antiviral drug discovery, along with their key strengths and limitations, are shown in Table 1.

(Multiple) Linear Regression

Linear regression models are among the simplest ML algorithms and are used to model the linear relationship between dependent and independent variables. When the model uses only one independent variable to predict the dependent variable, it is referred to as a simple linear regression. In contrast, when two or more independent variables are used, the model is referred to as a multiple linear regression.[7,16] Linear regression is often used in the antiviral design to predict biological activity and molecular properties.[18,19]

Polynomial Regression

Polynomial regression extends linear regression by allowing for a more complex, non-linear relationship between the independent and dependent variables. While it is still considered a form of linear model in terms of the parameters being estimated, it differs by incorporating

polynomial terms of the independent variables. In polynomial regression, the original linear equation is transformed into a polynomial of degree *n*, which allows the model to fit curved patterns in the data. For example, a second-degree polynomial includes squared terms, enabling the model to capture quadratic relationships. Higher-degree polynomials allow for increasingly flexible curves, which can model more intricate trends within the data. This approach is particularly useful when the data exhibit a non-linear trend that cannot be captured adequately by a simple or multiple linear regression. However, caution is needed, as using very high-degree polynomials can lead to overfitting and to a low performance on new data.[7,16] Polynomial regression is often used in the development of antiviral drugs to model complex relationships between molecular structures to predict physicochemical properties. [18,20]

2.1.1.2. Overview of the Most Common Classification Algorithms in Antiviral Design Literature

k-Nearest Neighbor

The k-nearest neighbor (k-NN) algorithm represents an effective classification method used in ML. In k-NN, the class of a new data point is predicted based on the majority class among its *k* nearest neighbors in the training dataset. The fundamental assumption of the method is that data points with similar features tend to belong to the same class. Unlike many other ML algorithms, k-NN does not involve the training of a predictive model or estimating parameters. Instead, it is a non-parametric, instance-based learning method in which all training data points are stored. When a new data point needs to be classified, the algorithm computes the distance between this point and all instances in the training set, commonly using Euclidean distance or other distance metrics. The *k* closest points are identified, and the new data point is assigned to the class most frequently represented among these neighbors. As k-NN performs all computations at prediction time rather than during the training phase, it can be computationally expensive, particularly for large datasets. One of the most critical hyperparameters in k-NN is the choice of *k*. A small value of *k* can make the model sensitive to noise and outliers, whereas a large *k* may smooth out predictions and overlook important local patterns.[16,21] In antiviral design, k-NN has been successfully applied to a range of tasks, including drug target interaction prediction, drug classification, ADMET prediction, and drug repurposing.[22] In the study by Dong *et al*.[23], the k-NN algorithm served as an initial screening tool to eliminate compounds from the classification dataset that are unlikely to have an inhibitory activity against the SARS-CoV-2 main protease (M^{pro}).

Naïve Bayes

The Naïve Bayes algorithm is based on Bayes' theorem. It describes how the probability of a hypothesis can be updated when new evidence becomes available. The Naïve Bayes algorithm classifies data points under the simplifying assumption that the features in the data are completely independent of each other. This means that the algorithm treats each feature as if it has no relationship with the remaining ones, which simplifies the problem. This makes the calculations much faster and easier, and the algorithm can process high-dimensional data efficiently. However, the assumption of complete independence between the features is often not true in real datasets, as the features are often correlated with each other. Moreover, the algorithm is not always accurate when it comes to estimating probabilities, as these are not always reliable. Despite these limitations, Naïve Bayes is often used in drug discovery due to its robustness, speed, and versatility.[5,16] The application of the Naïve Bayes

algorithm in antiviral design includes the prediction of the biological activity of compounds, the selection of drug candidates, the estimation of experimental results, as well as the prediction of protein-protein, drug-drug and drug-protein interactions.[5]

2.1.1.3. Overview of the Most Common Regression and Classification Algorithms in Antiviral Design Literature

Support Vector Machine

Support vector machines, or SVM, represent ML algorithms designed to find an optimal hyperplane that separates data points in a high-dimensional space. For classification tasks, the goal is to determine a boundary that effectively distinguishes between different classes. In regression tasks, SVM aims to identify a hyperplane that best fits the data. In both cases, the algorithm maximizes the margin, which is the distance between the hyperplane and the closest data points, known as support vectors.[5,7,16] One of the key strengths of SVM lies in its ability to handle high-dimensional data and to detect complex patterns, making it particularly useful for analyzing large and noisy datasets.[5] In antiviral design, SVM has been applied in several areas, including the prediction of biological activity and molecular properties, drug–target interaction prediction, and target identification.[5,17] Xu *et al*.[24] used RF to classify predictions by a QSAR (quantitative structure–activity relationship) model aimed at identifying inhibitors for the SARS-CoV main protease (M^{pro}). In contrast, Dong *et al*.[23] applied RF in a regression task to estimate the binding affinities of candidate compounds against the SARS-CoV-2 M^{pro}.

Decision Tree

The decision tree represents a ML algorithm that recursively partitions a dataset into smaller, more homogeneous subsets. The process begins at the root node, which forms the entire dataset. At each decision point, the algorithm selects an input feature that best splits the data into branches, aiming to group together the data points with similar characteristics. In classification tasks, the decision tree attempts to assign data points to discrete classes based on feature values. In regression tasks, the algorithm predicts continuous numerical values by forming branches that minimize the variation within each group. The final output is a tree-like model composed of internal nodes (decision points), branches (outcomes of decisions), and leaf nodes (final predictions). A central challenge in using decision trees is finding the appropriate complexity. While deeper trees can fit training data more accurately, they risk overfitting and making the model less generalizable to new data.[7,16] In antiviral discovery, decision trees have been applied to predict biological activity, drug–target interactions, toxicity, and ADMET properties.[22]

Random Forest and XGBoost

Random Forest (RF) represents an ensemble ML method that builds and combines multiple decision trees to improve the predictive performance and to reduce the risk of overfitting. Each tree in the forest is trained on a random subset of the data and features, a process known as bootstrap aggregation or bagging. Once trained, each tree makes an independent prediction.[5,7,16] For classification tasks, the final output is determined by a majority voting across all trees. For regression tasks, the prediction is calculated as the average of all tree outputs.[4] By aggregating the results of many uncorrelated decision trees, RF reduces variance and typically achieves a higher accuracy and better generalization than a single

decision tree. This ensemble approach is particularly effective in handling complex datasets and noisy features.[5]

XGBoost, short for Extreme Gradient Boosting, is yet another popular tree-based ensemble method widely used in ML. Unlike RF, which builds trees in parallel, XGBoost builds trees sequentially, with each new tree learning from the residual errors of the previous ones. This technique, known as gradient boosting, focuses on minimizing the loss function more efficiently by giving a greater attention to data points that were poorly predicted by earlier models. XGBoost includes built-in regularization, which helps preventing overfitting, and is known for its scalability, speed, and performance, especially in large and structured datasets.[25]

In antiviral drug design, both RF and XGBoost have been applied to a wide range of tasks. These include the prediction of chemical and drug properties, virtual screening, identification of drug-target interactions, as well as protein-related predictions such as protein pKa values, protein-protein binding affinities, and functional classification of proteins.[5,17] Xu *et al*.[24] used RF for QSAR modeling to identify inhibitors of SARS-CoV-2 virus entry. Saar *et al*.[26] used RF to predict and rank the activity of antiviral compounds targeting the SARS-CoV-2 M^{pro}. Pal *et al*.[27] used XGBoost to develop a classification model for the identification of potential inhibitors of the SARS-CoV-2 papain-like protease (PL^{pro}).

2.1.2. Unsupervised Learning

In unsupervised learning, an ML model identifies patterns or structures in data without the guidance of known output labels. Since there are no predefined target values, the algorithm explores the inherent structure of the dataset to uncover meaningful relationships among data points.[4,5,7] One of the most commonly used and known techniques in unsupervised learning is clustering, which involves grouping data points based on their similarity. Data points within the same cluster share more common features with each other than with those in other clusters. Clustering helps to reveal natural groupings within the data and is often used for exploratory data analysis.[4,5,7,14,16] Unlike supervised learning, unsupervised methods are not suitable for tasks such as regression or classification that require labeled input–output pairs.[7] However, they are highly valuable for discovering hidden patterns, reducing dimensionality, detecting anomalies, and identifying latent features that may inform further analysis or downstream predictive modeling.[5]

2.1.2.1. Overview of the Most Common Clustering and Dimensionality Reduction Techniques in Antiviral Design

k-means Clustering

k-means clustering represents a ML method that classifies data points into a certain number of clusters. The process starts with the random selection of k centroids, which represent the center of each cluster in the dataset. Each point in the dataset is then assigned to the nearest centroid based on distance. The centroids are subsequently recalculated by moving them closer to the average position of the points in their respective clusters. This process is repeated, adjusting the centroids and reassigning the data points to the nearest centroid, until the centroids no longer change, or a maximum number of iterations is reached. A major advantage of the k-means applies in its ability to manage high-dimensional data and to recognize complex patterns, especially in large and noisy datasets. However, the performance of the k-means can be affected by the initial selection of centroids. Despite these challenges, the k-means algorithm remains an indispensable tool in drug discovery

due to its simplicity, scalability and flexibility.[5,7] In the study by Xu *et al*.[24] the k-means algorithm was used for clustering the compounds with the highest predicted probabilities of inhibitory activity against SARS-CoV-2 entry and M^{pro} to ensure structural diversity.

Principal Component Analysis

Principal component analysis, or PCA, represents a widely used dimensionality reduction technique that transforms a large set of correlated variables into a smaller set of uncorrelated variables called principal components. These components capture the maximal possible variance in the data using fewer dimensions, allowing the dataset to be simplified while retaining most of its important information. The primary objective of PCA is to reduce data complexity, making it easier to visualize, interpret, and process, particularly in high-dimensional spaces. Although dimensionality reduction may result in a certain loss of information, PCA often achieves a favorable balance between simplification and data fidelity. This reduction not only improves computational efficiency but also helps mitigate issues such as overfitting in ML models. In antiviral design, PCA plays a significant role in simplifying complex datasets. It is applied to reduce the dimensionality of molecular descriptors, to group antiviral compounds based on structural or physicochemical similarities, and to support predictions of antiviral efficacy. PCA is also applied in the analysis of protein–ligand interactions, optimization of antiviral formulations, and assessment of compound toxicity.[28]

t-Distributed Stochastic Neighbor Embedding (t-SNE)

t-distributed stochastic neighbor embedding, or t-SNE, represents a dimensionality reduction technique designed to transform high-dimensional data into a lower-dimensional representation, typically two or three dimensions, for visualization purposes. It works by measuring pairwise similarities between data points in the high-dimensional space and then mapping them into a low-dimensional space in such a way that similar points remain close to each other, preserving the local structure of the data. t-SNE is particularly effective at uncovering complex, non-linear relationships in data that may be missed by linear methods. It is widely used for visualizing patterns, grouping structures, and relationships within large and complex datasets. This makes it especially valuable in exploratory analysis and in applications where intuitive understanding of the data's structure is important. However, t-SNE can be computationally intensive, especially with large datasets, and it may struggle to accurately represent global structures or clusters that occur at different scales. Additionally, the performance and output of t-SNE are sensitive to hyperparameters such as perplexity and learning rate, which require careful tuning to achieve meaningful results. In antiviral design, t-SNE has been applied to tasks such as compound clustering, drug-target relationship analysis, molecular representation, and the selection of candidate molecules based on the structural or functional similarity.[5] In the study by Ali *et al*.[29], t-SNE was used to cluster and identify potential nucleoside analogs among the compounds predicted as potential antiviral agents against SARS-CoV-2.

2.1.3. Reinforcement Learning

Reinforcement learning represents a type of ML in which a system learns optimal decision-making through interactions with a dynamic environment. The learning process is based on trial and error: the system receives positive rewards for beneficial actions and penalties for unfavorable ones. Unlike supervised learning, reinforcement learning does not rely on

labeled input–output pairs. Instead, it continuously gathers feedback from the environment and adjusts its strategy to maximize cumulative rewards over time. This approach is particularly well-suited for problems involving sequential decision-making and adaptive behavior. In the context of antiviral design, reinforcement learning has shown success in several key applications. These include molecular docking, where it can guide ligand positioning to optimize binding affinity, *de novo* compound design, where it helps generate novel antiviral candidates with desired properties; and the optimization of pharmacokinetic and pharmacodynamic characteristics of antiviral agents.[5,16] In the study by Alkan *et al*.[30], reinforcement learning based on the Markov decision process was used to predict the bioactivity of small compounds against the Rift Valley fever virus.

2.2. Deep Learning Methods

Deep learning represents a subset of ML that relies on artificial neural networks (ANNs) to recognize complex patterns in data. ANNs are inspired by the structure and function of the human brain and consist of layers of interconnected units called neurons. These neurons are organized into an input layer, one or more hidden layers, and an output layer. The input layer receives raw data, which is transformed through the hidden layers via a series of weighted mathematical operations. Each neuron applies an activation function to its input to introduce non-linearity, enabling the network to learn intricate relationships. The output layer produces the final prediction or classification. Deep learning models operate in two main stages: feedforward and backpropagation. During the feedforward phase, data flows through the network from the input to the output layer. In the backpropagation phase, the model evaluates prediction errors and adjusts the weights in the network to minimize them, thereby optimizing performance. This process is repeated iteratively across many training cycles. The depth and complexity of an ANN, determined by the number of layers and neurons, allow deep learningmodels to capture highly non-linear patterns, that traditional ML methods may miss. Deep neural networks are particularly effective in modeling high-dimensional, unstructured data. Deep learning algorithms can be trained in a supervised or unsupervised settings, depending on whether labeled data are available. A third approach, known as hybrid learning, combines aspects of both supervised and unsupervised learning to enhance the flexibility and performance.[5,14]

Deep learning has significantly advanced the field of generative modeling, a class of machine learning algorithms that are now predominantly based on deep learning techniques. These models have overtaken conventional approaches and have become the cornerstone of modern computational methods. Generative models have proven to be a powerful tool for the de novo design of novel molecules enabling the development of new compounds within a defined chemical space. They are trained using large databases of known compounds to learn the underlying patterns and structural rules of molecular architecture. After training, a generative model can be used to design new molecules or fragments that have not yet been explored. These molecules can be tailored to have specific physicochemical properties or biological activities.[5,7] Molecular representations in generative models can take various forms, such as chemical fingerprints, Simplified Molecular Input Line Entry System (SMILES) strings representation, molecular diagrams or three-dimensional structures.[31] Recent advances include design of 3D molecular structures that are structurally compatible with the target to fit the binding pockets of target proteins and produce drug-like molecules with high binding affinity.[32] In addition, a new way of linking chemical names to molecular structures has been adopted, enabling design and optimization by editing IUPAC names and adding and modifying functional groups so that chemists can easily customize the functional groups. Subsequently, these modifications are translated into SMILES strings for computational analysis, leading to compelling results in the production of unique, diverse and synthesizable

compounds that are particularly useful for antiviral drug development.[33] These models are rapidly changing the landscape of drug design as they allow access to previously inaccessible areas of chemical space with greater speed and precision. Sun *et al*.[34] used generative modeling techniques to design a novel inhibitor targeting the SARS-CoV-2 main protease Mpro, which has entered phase I clinical trials

In antiviral design, deep learning has been used in a variety of applications, including the prediction of antiviral activity, analysis of protein–ligand interactions, *de novo* generation of antiviral candidates, and modeling of viral genome features.

2.2.1. Supervised Learning

In a supervised learning, a neural network is trained using input data and the corresponding outputs. It processes the inputs to generate outputs and then compares them with the actual output. Any differences between the predicted and actual outputs generate an error signal that the network uses to adjust its parameters. This process is repeated until the network has minimized the errors and achieved a satisfactory level of accuracy.[4,5,14] Supervised neural networks can be applied for regression and classification tasks.[4] When the model is used for classification, it assigns the input data to a specific class. When applied for regression, it predicts numerical values. The most common supervised learning methods include Multi-Layer Perceptron, Convolutional Neural Networks, and Recurrent Neural Networks.[5,14]

Multi-Layer Perceptron

Multi-Layer Perceptron (MLP) represents an ANN, which consists of an input layer, one or more hidden layers, and an output layer. The purpose of a MLP is to model complex relationships between inputs and outputs, making it a powerful tool for various classification and regression tasks.[4,5] MLP can be used in drug discovery for tasks such as *de novo* drug design, prediction of biological activity, as well as prediction of drug-drug and drug-target interactions.[4] Ang *et al*.[35] used MLP to predict the binding probability of compounds to SARS-CoV-2 M^{pro}. Bess *et al*.[36] used MLP to classify the similarity of drugs with antiviral peptides (AVPs) in repurposing against SARS-CoV-2.

Convolutional Neural Networks

A convolutional neural network (CNN) forms a type of feedforward deep ANN that is particularly effective at recognizing patterns in input data, especially in images. CNNs consist of different layers that transform input data into more useful features and work together to learn important patterns. The first key layer is the convolutional layer, which applies filters that slide over the input data and perform a mathematical operation at each data point to extract the features. Next is the pooling layer, which simplifies the data by retaining only its most significant features. This step helps to reduce computational costs and the amount of storage space required. The last, fully connected layer processes the extracted features and makes the final prediction. The network improves its performance over time by optimizing filters and weights using backpropagation. Input data in the form of 2D representations of chemical compounds can be used for CNNs, which are used in drug discovery to predict biological activity, chemical properties, the behavior of compounds, and their interaction with biological targets. It can also be applied for grouping molecular structures based on their similarities. [5,37] Komatsu *et al*.[38] used CNN to analyze protein-ligand interactions and provide activity prediction probabilities for potential inhibitors of SARS-CoV-2 M^{pro}. Anwaar *et al*.[39] used CNNs to predict ligand-protein binding affinities against several SARS-CoV-2 viral proteins. Gkekas *et al*.[40] used a 3D-CNN to re-evaluate protein-ligand complexes from molecular docking and predict the likelihood of compounds

that block the interaction between the SARS-CoV-2 spike protein and the human ACE2 receptor. Xiang *et al*.[41] used the CNN-based model DeepFrag to analyze protein-ligand complexes and predict fragment modifications that could enhance binding to the SARS-CoV-2 nucleocapsid protein. Kovalishyn *et al*.[42] used a consensus QSAR model combining Trans-CNN and Trans-CNF to predict the antiviral activity of compounds against varicella zoster virus.

Recurrent Neural Networks

Recurrent neural networks (RNNs) represent a special type of ANNs designed for processing sequential data, such as time series or natural language. RNNs differ significantly in how the information flows. Conventional feedforward neural networks process data in the direction from input to output without storing information from previous steps. In contrast, RNNs can remember previous information by feeding the output of a previous step into the network for the next step. In this way, the network can take previous input into account when making predictions. RNNs process data step by step, using the same weights at each time step. After the network makes a prediction, it checks for errors and adjusts the weights through backpropagation through time by calculating the errors across all time steps. In this way, RNNs can capture patterns that depend on long-term relationships in sequential data. RNNs can be used for compound optimization, *de novo* drug design, prediction of drug–target interactions, synthesis and efficacy testing of compounds, as well as for predicting compound properties. [5,37] Saramago *et al*.[43] used an Average Stochastic Gradient Descent Weight-Dropped Long Short-Term Memory (AWD-LSTM) neural network to generate fragment-like compounds with potential activity against SARS-CoV-2 M^{pro}, fine-tuned the model with known M^{pro} inhibitors and applied a classifier to predict bioactivity.

Graph Neural Networks

Graph Neural Networks (GNNs) form a type of ANNs developed for processing graphically structured data. Such data consists of nodes, which represent objects with features, and edges, which represent relationships between them. GNNs learn through a process called message passing, in which the nodes exchange information with their neighbors. After receiving information, the nodes update their features, and this process is repeated across multiple layers, allowing the network to learn the global structure of a graph. Once the network has learned sufficiently meaningful representations, it can be used for predictions. In drug discovery, small compounds, drug-protein interactions, and drug-disease interactions can be represented as graphs in which objects such as atoms, proteins or drugs are represented as nodes and chemical bonds between atoms or interactions between drugs, proteins, or diseases are represented as edges. Some applications of GNNs in drug discovery include molecular property predictions, protein structure predictions, and drug repurposing. GNNs can be applied for the generation of new compounds in drug discovery, for associations between drugs and diseases, for drug repurposing, and for drug response predictions.[6] Varikoti *et al*.[44] used a GNN to predict the binding affinity of compounds to the crystal structure of SARS-CoV-2 M^{pro}. Van der Horst *et al*.[45] used GNN to predict the antiviral activity of compounds against SARS-CoV-2 and ranked the candidates according to their probability of inhibiting angiotensin-converting enzyme 2 (ACE2), M^{pro} or transmembrane protease serine 2 (TMPRSS2). Ali *et al*.[29] used a graph convolutional neural network (GCNN) to predict the antiviral activity of compounds against SARS-CoV-2. Nguyen *et al*.[46] used a GCNN to predict binding free energies of compounds as potential neuraminidase inhibitors for influenza A and B.

2.2.2. Unsupervised Learning

As with conventional ML, the model in unsupervised learning is trained with data that has no known outputs. The main goal is for the model to recognize patterns in the input data, examine them, and organize the data based on similarities using techniques such as clustering (grouping data with similar features) and association (how different features relate to each other).[5,47,48] Unsupervised learning includes Generative Adversarial Networks, Deep Belief Networks, and Autoencoder.[5,14]

Generative Adversarial Networks

Generative Adversarial Networks (GANs) consist of two neural networks: a generator and a discriminator. The generator takes random vectors as input and generates data points, which are then presented to the discriminator. The discriminator receives both real data points from the training set as well as the data points generated by the generator and attempts to distinguish between the two. During training, the generator and discriminator compete with each other, with the generator improving his ability to generate realistic data, while the discriminator refines its ability to accurately classify real and fake data points. This process continues until both networks have found optimal parameters where the generator produces highly realistic data points and the discriminator can accurately distinguish between them.[5,37] In the context of drug discovery, GANs are applied for the predictions and investigation of drug properties and binding affinities, the analysis of image data from assays, the prediction of drug–drug and drug–target interactions, as well as for the *de novo* drug design.[5]

Deep Belief Networks

Deep Belief Networks (DBNs) consist of several layers of Restricted Boltzmann Machines stacked on top of each other. In this network, the first layer learns to recognize the features of the input data. Once the first layer is trained, its output is used as the input for the next layer, which is then trained in the same way. This process, known as pre-training, is repeated for each layer of the network. After unsupervised pre-training, the entire network is fine-tuned using supervised backpropagation (when data with known outputs is available), adjusting the weights and parameters of the network to optimize its performance. This approach helps in an effective training of deep networks by initially learning simple features in the first layers and gradually capturing more complex representations in deeper layers.[5,18] DBNs are useful in drug discovery, especially in defining molecular features and accurately predicting the biological activity of novel compounds.[5]

Autoencoder

Autoencoder (AE) represents a type of ANN that consists of an encoder, that compresses the input data into a compact representation, while capturing the most important features, and a decoder, that recovers the compressed data to make them as similar as possible to the original input. In drug discovery, AEs have proven useful in predicting drug binding affinity, developing new compounds and predicting drug–protein interactions.[5]

Table 1. Typical applications of ML algorithms in antiviral drug discovery, along with their key strengths and limitations.

ML algorithm	Typical use in antiviral design	Strengths	Weaknesses	Reference
Supervised Regression Algorithms				
(Multiple) Linear Regression	Prediction of biological activity and molecular properties	Simple, fast, and easy to interpret	Not suitable for non-linear data	18, 19

Polynomial Regression	Prediction of biological activity and molecular properties	Captures more complex, non-linear relationships	Prone to overfitting	18, 20
Supervised Classification Algorithms				
k-Nearest Neighbor	Prediction of drug–target interaction drug classification, ADMET prediction, and drug repurposing	No training required, the model postpones the calculation until the time of prediction	Computationally expensive Requires careful choice of *k* value	16, 21
Naïve Bayes	Prediction of biological activity, protein-protein, drug-drug and drug-protein interactions, selection of drug candidates, and estimation of experimental results	Fast and easy calculations Can process high-dimensional data	Assumes feature independence, which is rarely true and affects accuracy	5
Supervised Regression and Classification Algorithms				
Support Vector Machine	Prediction of biological activity, molecular properties, drug–target interaction, and target identification	Handles high-dimensional data Recognize complex patterns, especially in large and noisy datasets	Sensitive to kernel function and its parameters May require balancing of unbalanced datasets	5, 7
Decision Tree	Prediction of biological activity, drug–target interactions, toxicity, and ADMET properties	Easy to interpretate	Prone to overfitting	7
Random Forest and XGBoost	Prediction of molecular properties, protein-protein binding affinities protein pKa values, identification of drug-target interactions, virtual screening, and functional classification of proteins	Reduces overfitting Increases accuracy and achieves better generalization	Harder to interpretate	5, 7, 25
Unsupervised Clustering and Dimensionality Reduction Techniques				
k-means Clustering	Prediction of molecular properties, calculation of similarities between compounds, clustering	Handles high-dimensional data	Performance affected by the initial selection of k centroids	5,7

Journal Pre-proof

	of compound properties	Recognize complex patterns, especially in large and noisy data		
Principal Component Analysis	Reduction of molecular descriptor dimensionality, grouping of antiviral compounds based on structural or physicochemical similarities, predictions of antiviral efficacy, analysis of protein–ligand interactions, optimization of antiviral formulations, and assessment of compound toxicity	Simplifies data complexity	Dimensionality reduction may lead to loss of information	28
t-Distributed Stochastic Neighbor Embedding	Compound clustering, drug-target relationship analysis, molecular representation, and selection of candidate molecules	Effective at uncovering complex, non-linear relationships	Computationally intensive May misrepresent global structures Sensitive to hyperparameters	5
Reinforcement Learning	Molecular docking, de novo compound design, and the optimization of pharmacokinetic and pharmacodynamic characteristics	Learn optimal decision-making through interactions with a dynamic environment	Computationally intensive Requires lots of data	5
Supervised Learning				
Multi-Layer Perceptron	De novo drug design, prediction of biological activity, drug-drug and drug-target interactions	Simple and fast	Training is challenging Requires careful parameter selection	4,5
Convolutional Neural Networks	Prediction of biological activity, chemical properties, the behavior of compounds, and their interaction with biological targets, molecular clustering	Effectively learns complex features from visual data	Computationally intense Requires lots of data	5, 26

Journal Pre-proof

Recurrent Neural Networks	Compound optimization, de novo drug design, prediction of drug–target interactions, compound properties, synthesis and efficacy testing of compounds	Captures long-term patterns in sequential data	Struggles to retain initial information in long sequences	5, 37
Graph Neural Networks	Generation of new compounds, associations between drugs and diseases, drug repurposing, and drug response predictions	Learns large-scale patterns and relationships Functions are order-independent	Computationally expensive for large graphs Sensitive to noisy or incomplete data	6
Unsupervised Learning				
Generative Adversarial Networks	Predictions and investigation of drug properties and binding affinities, the analysis of image data from assays, the prediction of drug–drug and drug–target interactions, de novo drug design	New data closely resemble real-world data	Unstable training makes it difficult to balance generator and discriminator	5, 37
Deep Belief Networks	Molecular features defining and prediction of biological activity	Recognize complex patterns in the input data	Complex training requires lots of data and computing resources	5
Autoencoder	Prediction of drug binding affinity, and drug–protein interactions, development of new compounds	Effectively compresses key input features	Sensitive to nosy data	5

3. Application of Machine Learning for the Identification of Biologically Active Antiviral Agents

Figure 2 illustrates a typical machine learning workflow in antiviral drug design, which includes data collection and splitting, model training and deployment, and experimental validation to identify and optimize potential antiviral agents.

Journal Pre-proof

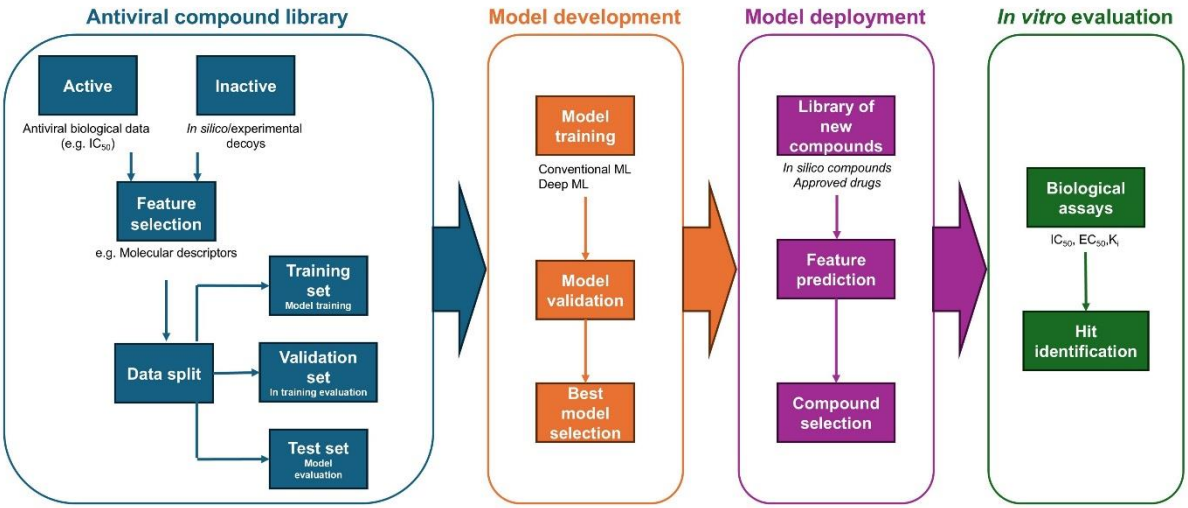


Figure 2. Common ML workflow in antiviral drug design.

An overview of ML studies on antiviral compound identification from 2022 onwards is provided in Table 2. Information on molecular targets, computational tools, and ML algorithms is also included along with corresponding references

Table 2. An overview of studies based on ML for antiviral agent identification with information on molecular targets, computational tools, ML algorithms, biological assay type, and antiviral activity from 2022 onwards.

Journal Pre-proof

Author	Target	Model/toolkit/ platform/software	ML algorithm	Assay type	Antiviral activity[1]	Reference
Iype *et al.*	SARS-CoV-2 M^{pro}		KRR	FRET-based enzyme inhibition assay	**bromocriptine** IC_{50} = 0.13 μM **avoralstat** IC_{50} = 2.16 μM	49
Jo *et al.*	SARS-CoV-2 M^{pro}	$SOM^{AI}PRO$ https://sombiotech.com/	not specified	FRET-based enzyme inhibition assay	**prexasertib** IC_{50} = 1.99 μM **cynarin** IC_{50} = 1.82 μM, **eravacyclin** IC_{50} = 1.65 μM	52
				immunofluorescence-based infection assay of SARS-CoV-2 in VeroE6 cells	**eravacyclin** IC_{50} = 30.61 μM	
Xu *et al.*	SARS-CoV-2 entry and M^{pro}		RF and SVM	luminescence-based entry inhibition assay in HEK293-ACE2 cells	**NCGC00390584** IC_{50} = 3.1 nM	24
				flurescence-based enzyme inhibition assay	**NCGC00390337** IC_{50} = 0.92 μM	
				cytopathic effect inhibition assay of SARS-CoV-2 in HEK293-ACE2 cells	**NCGC00345807** AC_{50} = 0.20 ± 0.07 μM **NCGC00161621** AC_{50} = 1.41 ± 0.39 μM **MLS000703078** AC_{50} = 1.78 ± 0.00 μM **NCGC00390625** AC_{50} = 3.55 ± 1.12 μM **NCGC00599688** AC_{50} = 3.98 ± 0.00 μM	

Journal Pre-proof

					NCGC00017063 AC_{50} = 4.14 ± 1.41 μM	
Ang *et al.*	SARS-CoV-2 M^{pro}	DeepPurpose https://github.com/kexinhuang12345/DeepPurpose	MLP	fluorescence-based enzyme inhibition assay	**CNP0061237** IC_{50} = 6.88 μM **CNP0375828** IC_{50} = 22.4 μM **CNP0381522** IC_{50} = 8.24 μM **CNP0366487** IC_{50} = >452 μM	35
Dong *et al.*	SARS-CoV-2 M^{pro}		k-NN and SVM	FRET-based enzyme inhibition assay	**1** IC_{50} = 0.714 μmol/L **2** IC_{50} = NA[2] **3** IC_{50} = 38.33 μmol/L **4** IC_{50} = 6.217 μmol/L **5** IC_{50} = 9.763 μmol/L **6** IC_{50} = 4.992 μmol/L **7** IC_{50} =NA[2] **8** IC_{50} = 0.172 μmol/L **9** IC_{50} = 0.420 μmol/L **10** IC_{50} = 5.006 μmol/L	23
Joshi *et al.*	SARS-CoV-2 M^{pro}	3D-Scaffold https://github.com/PNNL-CompBio/3D_Scaffold	TNN	FRET-based enzyme inhibition assay	**A** IC_{50} = 0.30 ± 0.03 μM[3] K_i = 11.5 ± 3.2 μM **B** IC_{50} = 0.40 ± 0.13 μM[3] K_i = 6.0 ± 4.0 μM **C** IC_{50} = 0.16 ± 0.03 μM [3] K_i = 5.3 ± 2.1 μM **D** IC_{50} = 0.41 ± 0.06 μM[3] K_i = ≥ 62.50 μM	50
Varikoti *et al.*	SARS-CoV-2 M^{pro}	3D-Scaffold https://github.com/PNNL-CompBio/3D_Scaffold	TNN, SVM, and GNN	FRET-based enzyme inhibition assay	**MCULE-7471308738** IC_{50} = 2.95 ± 0.0017 μM **MCULE-4926166920** IC_{50} = 3.41 ± 0.0015 μM	44

Journal Pre-proof

Saar *et al.*	SARS-CoV-2 M^{pro}		RF	fluorescence-based enzyme inhibition assay	**1** IC_{50} = 0.34 μM	26
				cytopathic effect inhibition assay of SARS-CoV-2 in Vero E6 cells	**1** EC_{50} = 120 nM	
Saramago *et al.*	SARS-CoV-2 M^{pro}		AWD-LSTM neural network	FRET-based enzyme inhibition assay	**818** IC_{50} = 1.51 μM **737** IC_{50} = 51.9 μM **183** IC_{50} = 116.8 μM **048** IC_{50} = 139.7 μM **281** IC_{50} = 241.5 μM	43
				cytopathic effect inhibition assay of SARS-CoV-2 in Vero E6 cells	**818** EC_{50} = 1.1 μM **737** EC_{50} =NI μM **183** EC_{50} =NI μM	
				cytopathic effect inhibition assay of SARS-CoV-2 in Calu-3 cells	**818** EC_{50} = 3.7 μM **737** EC_{50} = 2.4 μM **183** EC_{50} = 1.8 μM	
Zhang *et al.*	SARS-CoV-2 M^{pro}	ChemBERTa https://deepchem.io/tutorials/transfer-learning-with-chemberta-transformers/	transfer learning	FRET-based enzyme inhibition assay	**T2730** IC_{50} = 67.6 μM **T2844** IC_{50} = 235.8 μM	57
Komatsu *et al.*	SARS-CoV-2 M^{pro}	INTENDD http://www.interprotein.com/intendd.html	CNN	SPR	**vorapaxar** K_d = 27 μM **dasabuvir** K_d = 3.1 μM	38
				imunofluorescence-based inhibition assay of SARS-CoV-2 in Vero E6 cells	**vorapaxar** EC_{50} = 11 μM	
Juárez-Mercado *et al.*	SARS-CoV-2 M^{pro}	Assay Central Software	seven ML algorithms in Assay	FRET-based enzyme inhibition assay	**13** IC_{50} = 3.5 μM **13a** IC_{50} = > 100 μM	51

Journal Pre-proof

		https://collaborations.com/assay-central	Central Software (not specified) and RF		**13b** IC_{50} = 29.1 μM **13c** IC_{50} = 1.8 μM	
				cytopathic effect inhibition assay of SARS-CoV-2 in Vero CCL81cells	**13** EC_{50} = > 100 μM **13a** EC_{50} = 35.3 μM **13b** EC_{50} = 59.9 μM **13c** EC_{50} = 57.0 μM	
Sun *et al*.	SARS-CoV-2 M^{pro}	Chemistry42 http://chemistry42.com	over 40 generative models	fluorescence-based enzyme assay	**1A** IC_{50} = 61 ± 1 nM **1B** IC_{50} = 121 ± 7 nM **2** IC_{50} = 16 ± 1 nM **3** IC_{50} = 14 ± 2 nM	34
				imunofluorescence-based inhibition assay of SARS-CoV-2 in Vero E6 cells	**1A** EC_{50} = 2.15 μM **1B** EC_{50} = 1.13 μM **2** EC_{50} = 0.068 μM **3** EC_{50} = 0.071 μM	
Anwaar *et al.*	SARS-CoV-2 PL^{pro}	DeepDTA https://github.com/hkmztrk/DeepDTA	CNN	ubiquitin-rhodamine 110 Gly fluorescent intensity assay	**rifapentine** IC_{50} = 15.18 μM **flavin adenine dinucleotide disodium** IC_{50} = 12.39 μM	39
Garnesey *et al.*	SARS-CoV-2 PL^{pro}		neural network (type not specified)	FRET-based enzyme inhibition assay	**4** K_i = 2 nM	63
				cytopathic effect inhibition assay of SARS-CoV-2 in dNHBE cells	**4** EC_{50} = 13.9 nM	
Pal *et al.*	SARS-CoV-2 PL^{pro}		XGBoost	fluorescence-based enzyme inhibition assay	**NCGC00473156** IC_{50} = 9.88 μM **NCGC00473291** IC_{50} = 6.24 μM	27
Teshima	RBD-ACE2	INTENDD	CNN	immunoassay-based	**demethoxycurcumin**	59

Journal Pre-proof

et al.	binding	http://www.interprotein.com/intendd.html		SARS-CoV-2 Spike-ACE2 interaction inhibition assay	IC_{50} = 5.0 μM **bisdemethoxycurcumin** IC_{50} = 2.5 μM **curcumin sulfate** IC_{50} = 2.4 μM **monodemethylcurcumin** IC_{50} = 8.4 μM **bisdemethylcurcumin** IC_{50} = 8.1 μM	
				imunofluorescence-based inhibition assay of SARS-CoV-2 in Vero E6 cells	**demethoxycurcumin** EC_{50} = 7.2 μM **bisdemethoxycurcumin** EC_{50} = 4.9 μM **curcumin sulfate** EC_{50} = 18.6 μM **monodemethylcurcumin** EC_{50} = 6.2 μM **bisdemethylcurcumin** EC_{50} = 23.0 μM	
Gkekas *et al.*	RBD-ACE2 binding		3D-CNN	LuTHy interaction inhibition assay in HEK293T cells	**CKP-22** IC_{50} = 55 μM	40
				fluorescence-based SARS-CoV-2 pseudovirus inhibition entry assay in SH-SY5Y cells	**CKP-22** IC_{50} = 49 ± 2.5 μM	
				cytopathic effect inhibition assay of SARS-CoV-2 in Vero E6 cells	**CKP-22** IC_{50} = 50 μM **CKP-25** IC_{50} = 3.5 μM	
Yang *et*	SARS-CoV-2	Chemprop	D-MPNN	fluorescence-based	**Mg-132** IC_{50} = 12.28 nM	53

Journal Pre-proof

al.	CTSL	https://github.com/chemprop/chemprop		enzyme inhibition assay	**Z-FA-FMK** IC_{50} = 54.87 nM **leupeptin hemisulfate** IC_{50} = 5.77 nM **Mg-101** IC_{50} = 5.77 nM **calpeptin** IC_{50} = 43.98 nM **daptomycin** IC_{50} = 7.87 μM	
				luciferase-based inhibition assay of SARS-CoV-2 pseudovirus in Huh7 cells	**Mg-132** EC_{50} = 212.50 nM **Z-FA-FMK** EC_{50} = 701.20 nM **leupeptin hemisulfate** EC_{50} = 39.29 μM **Mg-101** EC_{50} = 3.12 μM **calpeptin** EC_{50} = 29.82 μM **daptomycin** EC_{50} = 207.27 μM	
				luciferase-based inhibition assay of SARS-CoV-2 pseudovirus in A549 cells	**Mg-132** EC_{50} = 159.20 nM **Z-FA-FMK** EC_{50} = 1.13 μM **leupeptin hemisulfate** EC_{50} = 4.26 μM **Mg-101** EC_{50} = 1.15 μM **calpeptin** EC_{50} = 34.81 μM **daptomycin** EC_{50} = 98.81 μM	
				luciferase-based inhibition assay of SARS-CoV-2	**Mg-132** EC_{50} = 147.20 nM	

Journal Pre-proof

				B.1.351 pseudovirus in Huh7 cells	**Z-FA-FMK** EC_{50} = 884.70 nM **leupeptin hemisulfate** EC_{50} = 21.54 μM **Mg-101** EC_{50} = 0.96 μM **calpeptin** EC_{50} = 13.53 μM **daptomycin** EC_{50} = 218.84 μM	
				immunofluorescence-based inhibition assay of SARS-CoV-2 virus in Huh7 cells	**Mg-132** EC_{50} = 0.21 μM **Z-FA-FMK** EC_{50} = 9.61 μM **leupeptin hemisulfate** EC_{50} = 10.55 μM	
Xiang *et al.*	SARS-CoV-2 N protein	EMPIRE DeepFrag https://durrantlab.pitt.edu/deepfragmodel/	AE CNN	SPR	**38** K_d = 0.35 μM	41
				microscale thermophoresis	**38** K_d = 0.73 μM	
				MTT assay in Vero E6 cells	**38** EC_{50} = 11.3 μM	
Ali *et al.*	SARS-CoV-2	DeepChem https://deepchem.io/	GCNN and t-SNE	ELISpot assay in Vero E6 cells	**C3** EC_{50} = 1 ± 2.1 μM **C5** EC_{50} = 1 ± 1.9 μM **C7** EC_{50} = 3.9 ± 1.8 μM **C8** EC_{50} = 1.5 ± 1.7 μM **C9** EC_{50} = 6.6 ± 1.8 μM **C10** EC_{50} = 5.3 ± 1.7 μM **C11** EC_{50} = 6.7 ± 2.1 μM **C19** EC_{50} = 1.7 ± 2 μM azathioprine	29
				immunofluorescence staining, plaque assay and Western blot in SARS-CoV-2	NA[2]	

Journal Pre-proof

				infected Calu-3 cells		
Bess *et al.*	SARS-CoV-2 entry, fusion and replication	eVir https://github.com/wanwenzeng/deepdrug	SNN and MLP	immunostaining assay in SARS-CoV-2 infected Calu-3 cells	**mebendazole** IC_{50} = 48.55 nM **gefitinib** IC_{50} = 62 nM **topotecan** IC_{50} = 9.362 nM **niclosamide** IC_{50} = 982.5 nM **carfilzomib** IC_{50} = 5.741 nM	36
van der Horst *et al.*	SARS-CoV-2		GNN	qPCR analysis	Vero E6 **PKI-179** IC_{50} = 0.3 μM **MTI-31** IC_{50} = < 5 μM Calu-3 **PKI-179** IC_{50} = 0.4 μM **MTI-31** IC_{50} = 0.15 μM	45
Kovalishyn *et al.*	varicella zoster virus	OCHEM https://ochem.eu/home/show.do	Trans-CNN and Trans-CNF	cytopathic effect inhibition assay of HCMV in human foreskin fibroblast	**3** EC_{50} = 4.01 μM **9** EC_{50} = 0.10 μM	42
Kovalishyn *et al.*	human cytomegalovirus	OCHEM https://ochem.eu/home/show.do	WEKA-RF, DLCA, ASNNs, and XGBOOST	cytopathic effect inhibition assay of varicella zoster virus in human foreskin fibroblast	**1** EC_{50} = 0.18 μM **5** EC_{50} = 5.58 μM	67
				virus yield reduction assay of HCMV in human foreskin fibroblast	**1** EC_{50} = > 20 μM	
Alkan *et al.*	Rift Valley fever virus Gc	AtomNet	CNN	plaque inhibition assay of RVFM in	**RGc-B05** IC_{50} = 10.98 μM	30

Journal Pre-proof

	protein	https://www.atomwise.com/how-we-do-it/		Vero cells	**RGc-F06** IC_{50} = 20.61 μM **RGc-F12** IC_{50} = 11.90 μM	
Izmailyan *et al.*	neuraminidase of influenza A and B viruses		MDP	fluorescence-based enzyme inhibition assay	Influenza A: **DS-22-inf-009** IC_{50} = 24.8 μM **DS-22-inf-021** IC_{50} = 11.2 μM Influenza B: **DS-22-inf-009** IC_{50} = 21.2 μM **DS-22-inf-021** IC_{50} = 9.1 μM	69
				cytopathic effect inhibition assay of IAV and IBV in MDCK cells	EC_{50} < 2.5 μM	
Nguyen *et al.*	neuraminidase of influenza A and B viruses		GCNN	fluorescence-based enzyme inhibition assay	NA[2]	46

[1]AC_{50}, half-maximal activity concentration; IC_{50}, half-maximal inhibitory concentration EC_{50}, half-maximal effective concentration, K_i, inhibition constant; K_d, dissociation constant; [2]NA, not available; [3]IC_{50} measured after 2 h.

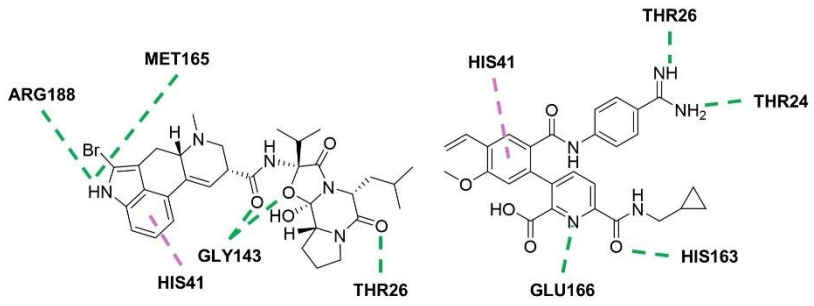
Journal Pre-proof

3.1. Antiviral Agents against SARS-CoV-2

3.1.1. Antiviral Agents targeting SARS-CoV-2 Main Protease

Most studies have focused on the identification of antiviral agents against the M^{pro} of SARS-CoV-2. SARS-CoV-2 M^{pro} is a homodimeric enzyme that is crucial for the cleavage of the viral polyprotein into functional, non-structural proteins that are critical for viral replication. Catalytic activity is mediated by the catalytic dyad of residues His41 and Cys145. Ligands that interact with Cys145 and/or His41 therefore reduce the activity of SARS-CoV-2 M^{pro}.[48-51]

In the study by Iype *et al.*,[49] the authors applied the DrugBank database to identify potential inhibitors of the SARS-CoV-2 M^{pro}. They randomly selected 2,307 compounds from the DrugBank dataset for molecular docking into M^{pro}. Using the 3D molecular structure of each molecule and the average binding energy from molecular docking, they developed a kernel ridge regression (KRR) model that is applied for modeling nonlinear relationships between variables. This model was then used to predict the binding energies of compounds from the DrugBank database and led to the selection of two promising candidates, **bromocriptine** and **avoralstat** (Figure 3), which were then tested *in vitro* with fluorescence resonance energy transfer (FRET)-based enzymatic assay on SARS-CoV-2 $M^{pro.}$ **Bromocriptine** exhibited an IC_{50} of 0.13 μM, while **avoralstat** had an IC_{50} of 2.16 μM. Molecular docking of **bromocriptine** into M^{pro} crystal structure (PDB ID: 6LU7) showed that it forms five hydrogen bonds with Thr26, Gly143, Met165 and Arg188 residues of M^{pro}. In contrast, **avoralstat** showed four hydrogen bonds with Thr26, Thr24, His163 and Glu166 residues of M^{pro}, suggesting that both drugs bind to the active site. Furthermore, molecular dynamics simulations of 100 ns showed that both drugs remained stable within the M^{pro} binding pocket and indicated that the ligands maintained their binding conformation in the binding pocket with no apparent rotation. MM-PBSA analysis during the final 20 ns trajectory revealed that **bromocriptine** (-29.39 kcal/mol) had a lower binding free energy than **avoralstat** (-6.91 kcal/mol). The large difference in binding energy is due to van der Waals interactions, which are more pronounced in **bromocriptine**. In addition, hydrogen bonding analysis in the last 20 ns of the simulation showed that **bromocriptine** frequently interacts with the catalytic residues His41 and Cys145. Several residues of the substrate binding site, including Gln192, Ser46, Arg188, Met49, Asn142, and Glu166, were also involved in hydrogen bonding with bromocriptine. **Avoralstat** mainly forms hydrogen bonds with Thr26, Gly23 and His64 and does not interact with the two catalytic residues, which may explain its lower inhibitory activity.



bromocriptine
IC_{50} ($SARS\text{-}CoV\text{-}2\ M^{pro}$) = 0.13 μM

avoralstat
IC_{50} ($SARS\text{-}CoV\text{-}2\ M^{pro}$) = 2.16 μM

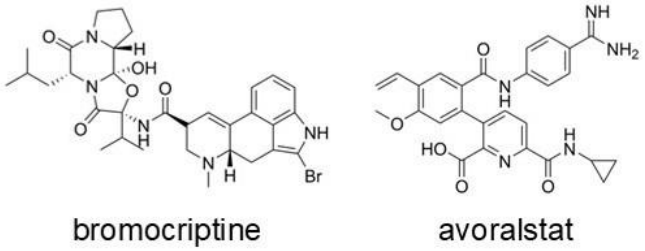


Figure 3. Identification of active SARS-CoV-2 M^{pro} inhibitors with a KRR model. Green dotted lines represent hydrogen bonds, while violet dotted lines indicate pi-stacking interactions.

Jo *et al.*[52] used SOM^{AI}PRO, an AI-based ligand-based virtual screening platform developed by SOM Biotech, to identify potential inhibitors of SARS-CoV-2 $M^{pro.}$ SOM^{AI}PRO is designed to identify compounds that are functionally similar to a reference compound but structurally different. The authors applied the experimentally confirmed SARS-CoV-2 M^{pro} inhibitor as the reference compound and used SOM^{AI}PRO to screen a library of approximately 9,000 FDA (Food and Drug Administration)-approved and investigational drugs to find candidates that are similar to the reference compound based on the molecular fields. The details of SOM^{AI}PRO's AI algorithms are not disclosed. Once SOM^{AI}PRO identified candidate compounds, its selection was refined by covalent and non-covalent molecular docking. 30 compounds with the highest rank were selected and applied in an *in vitro* FRET-based enzyme inhibition assay against SARS-CoV-2 M^{pro}. **Cynarin**, **eravacyclin** and **prexasertib** (Figure 4) showed inhibitory activity with IC_{50} values of 1.82 μM, 1.65 μM and 1.99 μM, respectively. Binding mode analysis by molecular docking (PDB ID: 6LU7) revealed that these compounds most likely interact with residues in catalytic dyad (residues His41 and Cys145) in the active site of M^{pro}. Cys145 was predicted to be involved in binding with all three chemicals, along with Glu166 in the active site, which was also predicted to be a key residue interacting through hydrogen bonds and a salt bridge interaction. **Cynarin** and **eravacyclin** are thought to form a covalent bond with Cys145, while **prexasertib** forms a hydrogen bond with this residue. This was confirmed in kinetic studies, which showed that **cynarin** and **eravacyclin** form irreversible covalent bonds with Cys145, while **prexasertib** acts as a competitive non-covalent inhibitor. Further immunofluorescence-based cellular assays revealed that **eravacyclin** effectively inhibits SARS-CoV-2 infection in Vero E6 cells (kidney epithelial cells of the African green monkey) with an IC_{50} of 30.61 μM and shows no effect on cell viability (CC_{50} > 50 μM), highlighting its potential as a antiviral against COVID-19.

Journal Pre-proof

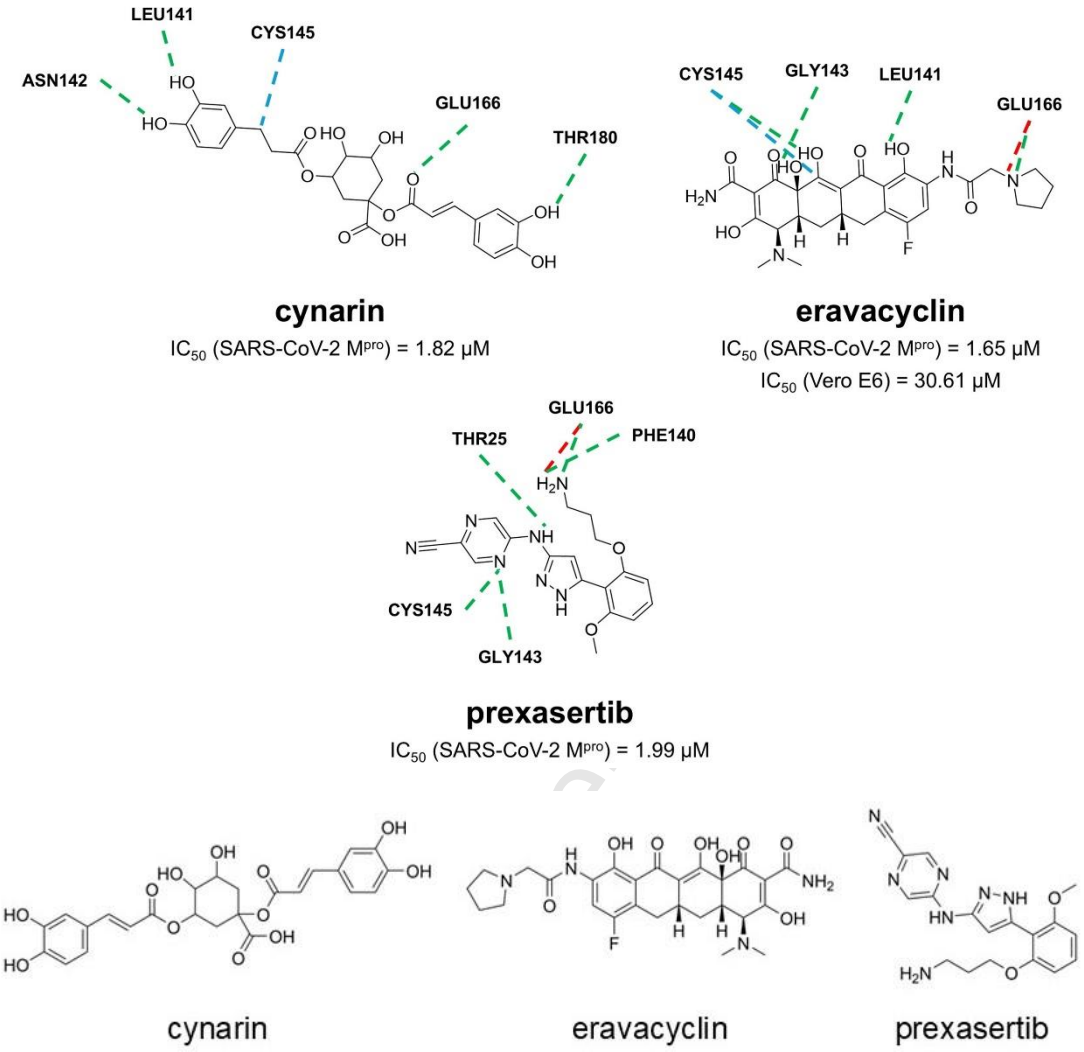


Figure 4. Identification of active SARS-CoV-2 M^{pro} inhibitors using SOM^{AI}PRO. G Green dotted lines represent hydrogen bonds, blue dotted lines indicate covalent bonds, and red dotted lines denote salt bridges.

A high-throughput screening of 8,149 compounds from the *in-house* collection of the National Center for Advancing Translational Sciences (NCATS) was conducted by Xu *et al.*[24] using two SARS-CoV-2-related assays: the SARS-CoV-2 pseudoparticle entry assay and the M^{pro} assay. The compounds identified by the screening were classified as either active or inactive based on the AC_{50}, maximal response (efficacy) values and shape of concentration−response curve. This classification was applied to develop predictive QSAR and biological activity-based models (BABM) for inhibitors of viral entry and M^{pro}. The QSAR model for viral entry inhibitors was built using the RF algorithm, while the QSAR model for M^{pro} inhibitors was built using the SVM algorithm, with molecular fingerprints as input features. The BABM models used activity profiles as input features. The BABM models were developed using the weighted feature significance method, a feature selection, and ranking approach. Although the BABM models performed better than the QSAR models, both were applied to screen approximately 360,000 compounds from the NCATS in-house library. For the QSAR approach, the compounds with the highest prediction probabilities (compounds with probabilities >0.5 were considered active) were clustered using the k-means algorithm to ensure structural diversity, and the best candidates from each cluster were selected for subsequent experimental validation. For the BABM models, compounds that exceeded certain thresholds for weighted feature scores were selected for experimental testing. By

combining the selections from both models, 1,972 predicted viral entry inhibitors and 1,493 predicted M^{pro} inhibitors were biologically evaluated. Viral entry inhibitors were detected in luminescence-based HEK293-ACE2 cell assay, while M^{pro} inhibitors were detected in a fluorescence-based enzyme assay. The most potent viral entry inhibitor identified was **NCGC00390584** (exatecan) with an IC_{50} of 3.1 nM, while the most active M^{pro} inhibitor was **NCGC00390337** (Z-DQMD-FMK) with an IC_{50} of 0.92 μM (Figure 5). 578 viral entry inhibitors and 150 M^{pro} inhibitors were further tested in a cell-based SARS-CoV-2 cytopathic effect reduction assay, with 22 compounds showing potency below 5 μM, including compounds **NCGC00345807** (CAA-0225; AC_{50} = 0.20 ± 0.07 μM), **NCGC00161621** (cepharanthine; AC_{50} = 1.41 ± 0.39 μM), **MLS000703078** (AC_{50} = 1.78 ± 0.00 μM), **NCGC00390625** (maropitant ;AC_{50} = 3.55 ± 1.12 μM), **NCGC00599688** (AQ-13; AC_{50} = 3.98 ± 0.00 μM), and **NCGC00017063** (amodiaquine; AC_{50} = 4.14 ± 1.41 μM).

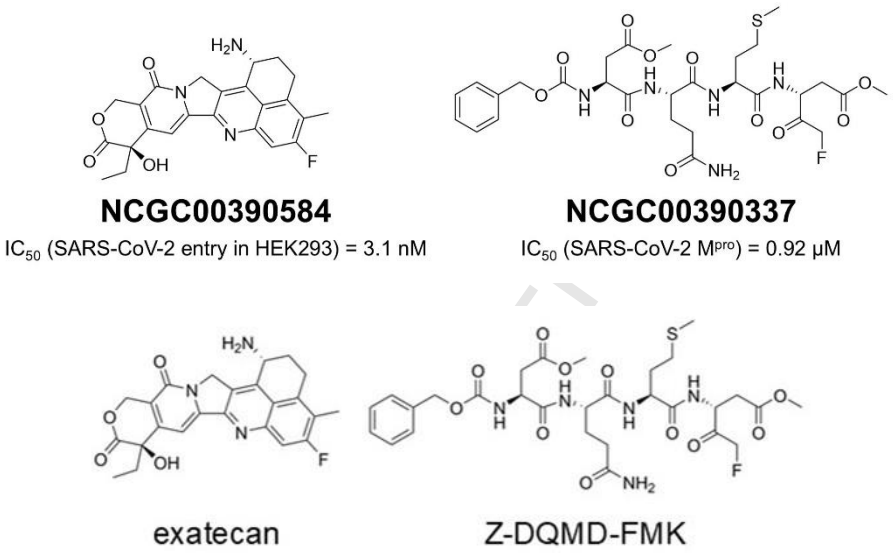


Figure 5. Identification of SARS-CoV-2 entry and M^{pro} inhibitors with RF and SVM-based QSAR models as well as BAMB models.

Ang *et al.*[35] focused on identifying natural compounds that can inhibit SARS-CoV-2 M^{pro}. They used DeepPurpose, a deep learning toolkit specifically designed for predicting interactions between compounds and their target proteins.[54] In this study, DeepPurpose was applied to predict protein-ligand binding by a MLP model. Public database BindingDB, which contains information on the binding affinity of proteins and small compounds, was used to train the model. The input features were encoded representations of both the drug molecules (Simplified Molecular Input Line Entry Specification (SMILES) and the target protein (amino acid sequence). Subsequently, the COlleCtion of Open Natural prodUcTs (COCONUT) library, which contains 400,000 natural compounds, and the amino acid sequence of SARS-CoV-2 M^{pro} served for binding predictions. 431 compounds with a predicted binding probability of at least 50% and desirable pharmacokinetic properties were selected for subsequent molecular docking on SARS-CoV-2 M^{pro} (PDB ID: 6LU7). Of the top 20 candidates from the molecular docking, seven compounds that showed favorable binding affinity were tested *in vitro* with fluorescence-based assay against SARS-CoV-2 M^{pro} enzyme. Four compounds (**CNP0061237**, **CNP0375828**, **CNP0381522**, and **CNP0366487** (Figure 6)) showed inhibition with IC_{50} values of 6.88, 22.4, 8.24 and >452 μM, respectively. Compounds **CNP0381522**, **CNP0375828** and **CNP0366487** are indole-based, while **CNP0061237** represents a benzene derivative, indicating potential for further development

of SARS-CoV-2 treatment. All four compounds **CNP0381522**, **CNP0375828**, **CNP0061237** and **CNP0366487** were predicted to form hydrogen bonds with the catalytic residue Cys145. With the exception of compound **CNP0375828**, all compounds were also observed to interact with residue Glu166, which is important for protein dimerization, so inhibition of protein dimerization was proposed as a possible mechanism.

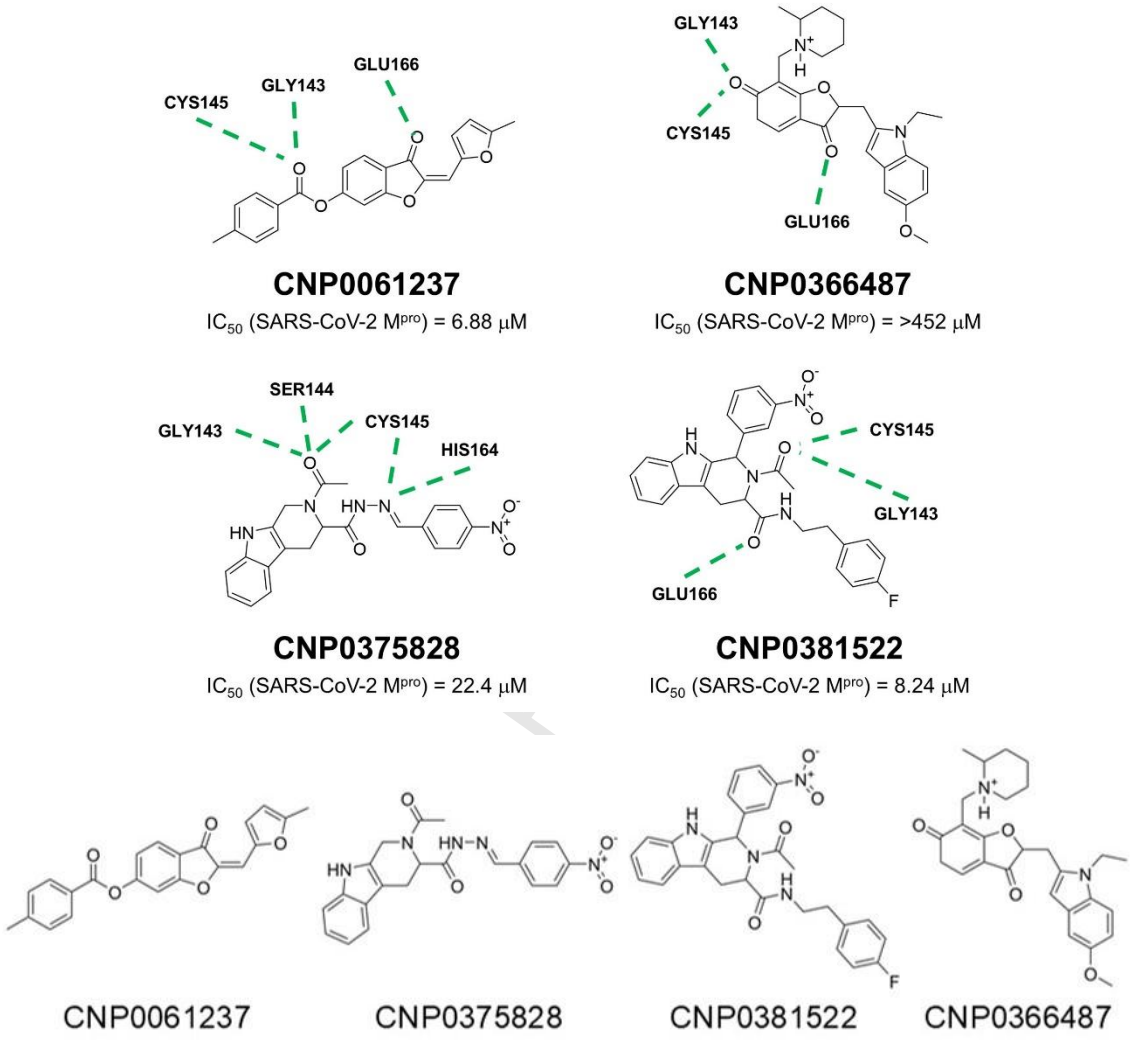


Figure 6. Identification of SARS-CoV-2 M^{pro} inhibitors using the MLP model from the DeepPurpose toolkit. Green dotted lines represent hydrogen bonds.

The aim of Dong *et al.*[23] was to identify inhibitors of the SARS-CoV-2 M^{pro}. In their approach, they used a virtual screening pipeline with classification and regression models to identify potential inhibitors of the M^{pro}. For the regression model, a dataset was created that included compounds from the literature with experimentally determined IC_{50} values against M^{pro}, that were converted to K_i, as well as compounds from the BindingDB database with experimentally determined K_i values against M^{pro}. For the classification model, compounds from different literature sources were collected and labeled as hits or non-binders. Compounds with $pK_i > 5.0$ were classified as hits, while compounds with $pK_i \leq 5.0$ were considered non-binders. The input features were 20-dimensional MACAW embeddings derived from structural fingerprints and molecular similarity metrics, used to represent each compound. In the first phase, a k-NN model was trained on the classification dataset to filter out compounds that were unlikely to be effective. In the second phase, a SVM model was trained on the regression dataset to predict binding affinities. The trained models were then

applied to a virtual library of commercially available compounds consisting of over 400,000 lead-like compounds. 105 compounds with a hit probability of ≥ 70% by the classification model and a predicted binding affinity pK_i ≥ 6.5 by the regression model were identified and submitted for the prediction of ADMET properties. Compounds were categorized into clusters based on their structural similarity, and 10 compounds from different clusters with the ADMET score ≥ 25 were selected for biological evaluation with FRET-based protease assay. Eight compounds showed a clear inhibitory effect on M^{pro} with an IC_{50} of 0.18-18.82 μM. Compounds **2** and **7** showed the strongest M^{pro} inhibition at low concentrations. However, an unexpected increase in fluorescence signal was observed at higher concentrations. This is due to autofluorescence in compound **2**, while an unidentified phenomenon occurs in compound **7**. As a result, the compounds show an atypical inhibition profile, that prevents an accurate determination of their IC_{50} value. However, both compounds were retained for further investigation to evaluate the cytotoxicity and antiviral activity of the selected compounds on human lung fibroblast cells (MRC-5). Antiviral properties were evaluated using MRC-5 cells infected with human coronavirus hCoV-229, but no values were reported. Cytotoxicity results indicated cytotoxicity of compound **9**, while compound **8** showed limited cytotoxicity and compounds **1** and **7** showed minimal cytotoxicity. Exact values were not reported. However, compound **7** (Figure 7) was recognized as the most promising antiviral hit against SARS-CoV-2, which exhibited dose-dependent inhibition of viral cell infectivity and minimal toxicity. Binding of compound **7** to M^{pro} was further examined by molecular docking which revealed a pi-H interaction with Glu166. Since no obvious interaction with the catalytic dyad was observed, allosteric binding was also considered. A binding site search was performed, and the compound was docked. A pi-H interaction with His246 was observed, suggesting a possible effect at this allosteric site.

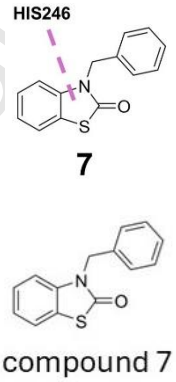


Figure 7. Identification of SARS-CoV-2 M^{pro} inhibitor using k-NN classification and SVM regression models. Violet dotted lines indicate pi-stacking interactions.

Joshi *et al.*[50] developed an AI pipeline to design electrophilic warhead-based covalent candidates with the aim of discovering covalent inhibitors of SARS-CoV-2 M^{pro}. They used their previously developed 3D-Scaffold generative model based on a tensor neural network (TNN) that generates molecular structures around a desired scaffold by iteratively placing one atom after another in 3D space.[55] The model was trained with 5,000 drug-like compounds from FDA and Enamine databases containing different scaffold structures, using their 3D molecular coordinates and the corresponding SMILES strings as input features to guide the generation of new compounds in three-dimensional space. Subsequently, 1,000 new valid and synthesizable compounds were created with piperazine and pyrrolidine scaffolds to which different electrophilic warheads were attached. The generated compound library was then applied for covalent docking, where specific electrophilic warheads were

covalently docked to the M^{pro} binding site targeting Cys145. The top 25 compounds ranked based on highest predicted binding affinity, binding pose, and most favorable interactions with M^{pro} were selected for biological evaluation on SARS-CoV-2 M^{pro} using FRET-based assay. Four chloroacetamide warhead compounds (**A**-**D**; Figure 8) with piperazine (**A**, **B**, **D**) and pyrrolidine (**C**) scaffolds showed inhibitory activity with IC_{50} values below 20 μM. Further investigation of their binding mechanism indicated covalent inhibition. The most potent inhibitor, compound **C**, exhibited an IC_{50} of 160 nM after a two-hour pre-incubation compared to 2.39 ± 0.09 μM before incubation. Kinetic studies showed that compounds **A**-**C** exhibited the expected saturation kinetics with low micromolar K_is for M^{pro}, with compound **C** displaying the lowest K_i of 5.3 ± 2.1 μM. Mass spectrometry studies confirmed the formation of inhibitor-M^{pro} complexes, with compound **C** showing the highest number of protein-inhibitor complexes. X-ray crystal structures of M^{pro} in complex with these inhibitors indeed displayed covalent bonds between them and Cys145. Molecular dynamics simulations suggested that protein dynamics is yet another important factor for further studies. Compound **C** was identified as the most potent M^{pro} inhibitor; however, all compounds exhibited cytotoxicity at concentrations up to 10 μM in Vero E6 cells. X-ray crystallographic studies of compounds **A** (PDB IDs: 8DL9), **C** (PDB IDs: 8DLB) and **D** (PDB IDs: 8DMD) bound to M^{pro} revealed that all compounds formed covalent bonds between the catalytic Cys145 and the chloroacetamide warheads as well as hydrogen bonds with nearby residues such as Gly143 and Cys145. The pi-H interaction was also observed with the key residue His41, being strongest with compound **D** and weakest with compound **C**, while it was not observed with compound **D**.

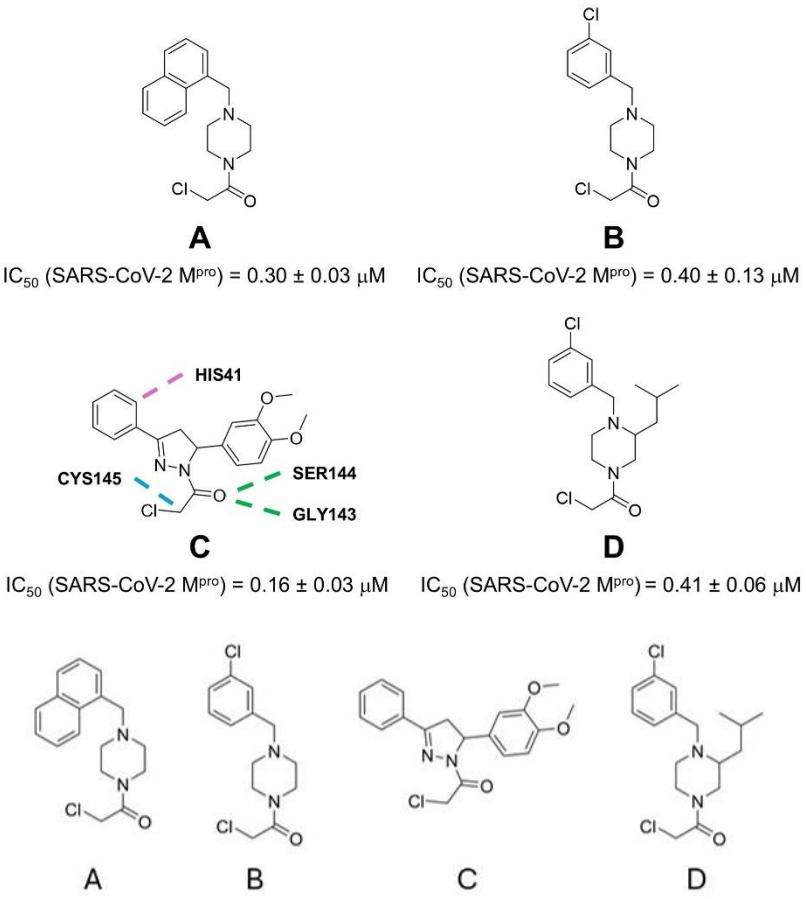
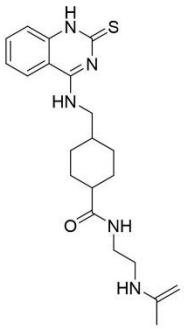


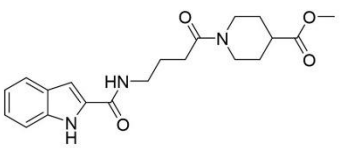
Figure 8. Identification of covalent SARS-CoV-2 M^{pro} inhibitors using the 3D-Scaffold model based on TNN. Green dotted lines represent hydrogen bonds, blue dotted lines represents covalent bonds, while violet dotted lines indicate pi-stacking interactions.

A similar approach was again applied by the same research group,[44] in which the authors used the 3D-Scaffold model to create a library of covalent and non-covalent inhibitors based on the scaffolds of compounds showing promising antiviral activity against the M^{pro} of SARS-CoV-2 with and without warheads. The model was trained with a dataset of 5,000 drug-like compounds from the FDA and Enamine databases (their 3D molecular coordinates and the corresponding SMILES strings used as input features), resulting in the generation of 1,000 new compounds with different scaffold structures. These new compounds not only had similar fingerprints to the training set but also exhibited the same properties as the input scaffold. To ensure drug-like properties, these newly generated compounds were subjected to a filtering process that narrowed down the selection to less than 500 candidates. Two molecular docking protocols were then applied, one targeting covalent and the other non-covalent interactions. Compounds that exhibited favorable binding scores and orientations were selected for the subsequent optimization stage. This included toxicity predictions, QSAR analyzes and multi-parameter optimization. The QSAR model was developed based on a SVM model to predict the activity of the compounds against M^{pro}. After creating a model and training it with a library of experimentally validated M^{pro} ligands from PostEra and the scientific literature, the model was analyzed to determine which molecular descriptors had the greatest impact on the predictions. From the set of 288 compounds predicted to be active, the selection was further refined by selecting 158 compounds that had similar values of the key descriptors. Multi-parameter optimization showed that the generated compounds were diverse yet shared important properties with the known inhibitors. These selected compounds were then analyzed using a GNN model. The model was trained on the data from the PDBbind and used to predict the binding affinity of the 158 compounds studied to the crystal structure of $M^{pro.}$ This analysis resulted in the 20 top candidates, 8 of which were commercially available for experimental validation. In FRET-based M^{pro} enzyme inhibition assays, compounds **MCULE-7471308738** and **MCULE-4926166920** (Figure 9) displayed the strongest inhibitory effects with IC_{50} values of 2.95 ± 0.0017 and 3.41 ± 0.0015 μM, respectively. Further investigations with time dependent inhibition experiments confirmed the non-covalent binding of these two compounds. Native mass spectrometry estimated their binding affinities and showed weak interactions in the range of 0.2-0.3 mM. Molecular dynamics simulations (PDB ID: 6WQF) revealed that **MCULE-7471308738** induces structural changes throughout the protein, suggesting that it may have allosteric effects, whereas **MCULE-4926166920** appears to stabilize the protein more uniformly. For **MCULE-7471308738**, hydrogen bonding with several M^{pro} active site residues was observed during the simulation, although it is not specified which interactions are involved.



MCULE-7471308738

IC_{50} (SARS-CoV-2 M^{pro}) = 2.95 ± 0.0017 μM



MCULE-4926166920

IC_{50} (SARS-CoV-2 M^{pro}) = 3.41 ± 0.0015 μM

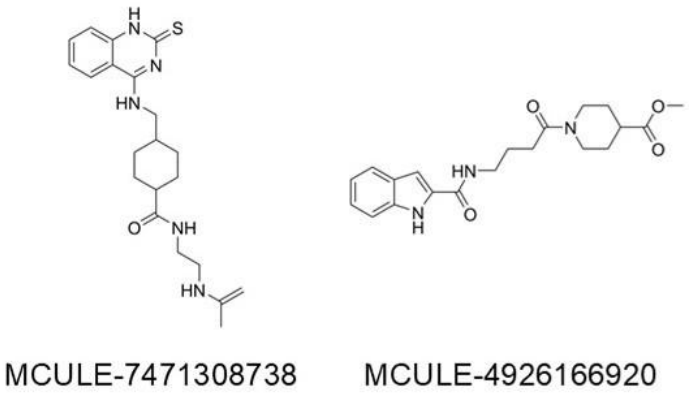


Figure 9. Identification of SARS-CoV-2 M^{pro} inhibitors using successive ML methods: a TNN-based 3D-Scaffold model, a SVM-based QSAR model and a GNN model.

In the study by Saar *et al.*,[26] the authors developed a RF model that was trained on the PDBBind dataset of protein-ligand complexes with known activity. Using this dataset, a model was trained to learn the relationship between the molecular descriptors extracted from the corresponding protein-ligand complex structure and the IC_{50} activity of the ligand so that it could rank compounds based on their predicted activity. The model was fine-tuned using SARS-CoV-2 data. To evaluate the model's performance in ranking novel ligands, the authors used a scaffold-split approach, in which an entire scaffold is removed from the training and included in the test set. The authors confirmed that the model is able to accurately rank unknown ligands without the need to include the structures of their specific scaffold in the training set. Ligands from the COVID Moonshot project, which contains 236 compounds with a measured activity against M^{pro}, were applied for molecular docking to generate predicted binding poses for these compounds within the M^{pro} binding site. The predicted complexes were used in the model along with the antiviral activity, and two compounds with an amine handle were selected as starting points for library design. A virtual library was designed that can be synthesized in a single reaction step by amide formation and reductive amination. This resulted in 15,720 compounds in the virtual amid library and 2,664 compounds in the reductive amination library. The authors docked these compounds and predicted binding complexes, that were passed to the ranking model to evaluate the ligands. The top 15 compounds from the final amide library ranking and the top 25 from the reductive amination library were synthesized and tested both in laboratory experiments and on living cells. Compound **1** (Figure 10) showed the strongest binding to M^{pro} in fluorescence-based assay with IC_{50} of 0.34 μM and was also tested in Vero E6 cells infected with SARS-CoV-2 using viral cytopathic effect inhibition assay, where it exhibited an EC_{50} value of 120 nM. Importantly, the compound displayed no cytotoxicity in Vero E6 cells at 10 μM. Structural analysis confirmed that compound **1** binds to a key pocket in M^{pro}, supporting its antiviral potential. Although it remained stable in metabolic experiments in both humans and rats, it exhibited poor permeability, suggesting that modifications may be required to improve absorption in the body. Molecular docking analyzes showed that compound **1** predictably forms hydrogen bonds with residues Glu166 and His163. Pi-stacking interactions were also observed with residues His163 and His41.

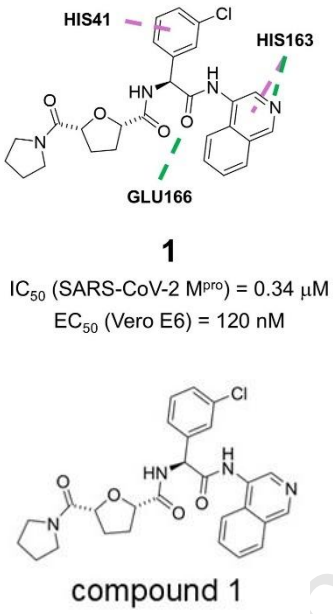
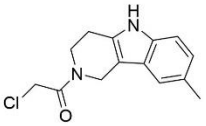


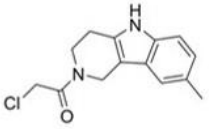
Figure 10. Identification of SARS-CoV-2 M^{pro} inhibitor with an RF model. Green dotted lines represent hydrogen bonds, while violet dotted lines indicate pi-stacking interactions.

A deep learning approach[56] was used in the study by Saramago *et al.*[43] to generate fragment-like compounds with potential inhibitory activity against SARS-CoV-2 M^{pro}. First, they developed a generative model based on the AWD-LSTM neural network, a type of RNN, and trained it with over 1.6 million unique compounds from the ChEMBL database. The input features for the model were SMILES strings of compounds. This model was able to generate new chemical structures and learn general chemical features. The model was then fine-tuned using a specific dataset of compounds from PubChem with a known activity against SARS-CoV-1 M^{pro} to incorporate information unique to M^{pro} inhibitors. The model generated 50,000 compounds. In a final step, a classifier was added to predict the bioactivity of new compounds against M^{pro}. 37 fragments that were classified as active by the model and met the rule of three for fragments were commercially available and structurally diverse, with none of the structures matching a compound that was already known to be active against M^{pro}. Of the 34 fragments tested in a FRET-based M^{pro} inhibition assay, five structurally diverse fragments (**048**, **183**, **281**, **737**, and **818**) showed statistically significant inhibition of M^{pro} activity and were identified as hits. Compounds **818**, **737**, and **183** were the most potent M^{pro} inhibitors with IC_{50} values of 1.5, 51.9, and 116.8 μM, respectively, and were further evaluated for their antiviral effects on SARS-CoV-2 replication and for associated cytotoxicity in Vero-E6 and Calu-3 cells. The most active compound, **818** (Figure 11), inhibited the replication of SARS-CoV-2 B.1 strain in Vero E6 cells and Calu-3 cells with an EC_{50} of 1.1 and 3.7 μM, respectively, and displayed low cytotoxicity. Therefore, the compounds can serve as a starting point for further development of SARS-CoV-2 M^{pro} inhibitors. The possible binding mode for compound **818** was investigated with molecular docking restricted to the binding site of the catalytic site (PDB ID: 6W79) and with molecular docking that examined the entire substrate binding site (PDB ID: 7K40). Molecular docking predicted different bindings, which also indicates different mechanisms.



818

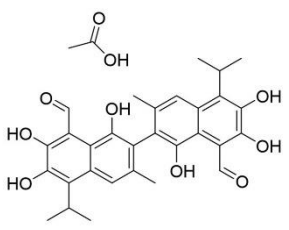
IC_{50} (SARS-CoV-2 M^{pro}) = 1.5μM
EC_{50} (Vero E6) = 1.1 mM
EC_{50} (Calu-3) = 3.7 mM



818

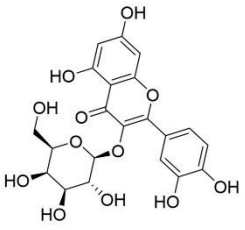
Figure 11. Identification of SARS-CoV-2 M^{pro} inhibitor using AWD-LSTM neural network-based model.

Zhang *et al.*[57] used a pre-trained ChemBERTa model, a transfer learning model for molecular property prediction,[58] to identify novel inhibitors for SARS-CoV-2 M^{pro}. They improved the performance of the model by supplementing the dataset with previously identified inhibitors for SARS-CoV-1 M^{pro} of 629 active and 288,940 inactive compounds. The input features used to train the model were Morgan fingerprints derived from SMILES strings of compounds. The SARS-CoV-1 inhibitors were used due to high sequence conservation and similar substrate binding sites between SARS-CoV-1 and SARS-CoV-2 $M^{pro.}$ The fine-tuned model was then applied to predict the bioactivity of the Targetmol library of 2,364 natural compounds and a library of 66,392 *de novo* generated compounds created by Santana and Silva.[38] 385 natural compounds and 66 *de novo* generated compounds were predicted to be active. The top 20 compounds from each library were subjected to molecular docking, with 10 natural compounds and two *de novo* generated compounds showing favorable binding interactions. Finally, 18 natural compounds from the top 20 active compounds predicted by deep learning and 9 natural compounds from molecular docking were tested for inhibition of SARS-CoV-2 M^{pro} activity in a FRET-based assay. The experiments revealed that compounds **T2730** (gossypol acetic acid) and **T2844** (hyperoside) had IC_{50} values of 67.6 and 235.8 μM, respectively, indicating their potential as SARS-CoV-2 M^{pro} inhibitors (Figure 12).



T2730

IC_{50} (SARS-CoV-2 M^{pro}) = 67.6 μM



T2844

IC_{50} (SARS-CoV-2 M^{pro}) = 235.8 μM

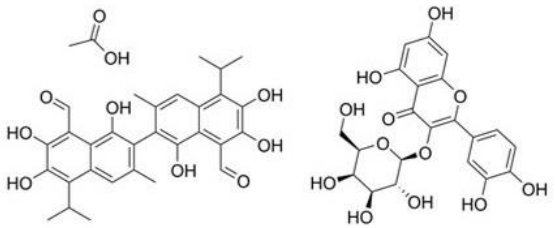
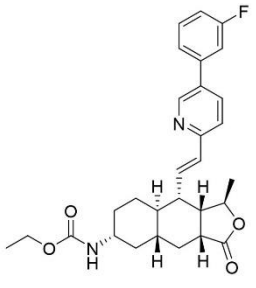


Figure 12. Identification of SARS-CoV-2 M^{pro} inhibitors using a transfer learning ChemBERTa model.

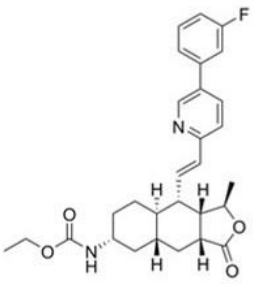
An AI-assisted drug discovery platform Artificial Intelligence-guided INTerprotein's Engine for New Drug Design (AI-guided INTENDD) was employed in the study by Komatsu *et al*.[38] to identify potential inhibitors of SARS-CoV-2 M^{pro} from a library of FDA-approved drugs. The INTENDD system uses CNN to analyze protein-ligand interactions based on the provided protein binding pocket coordinates and structural information on the compounds, including atomic types, 3D positions, interaction patterns, and functional groups, and provides predicted binding poses along with scores that indicate probabilities of the activity prediction. The authors performed molecular docking of 1,741 suitable FDA-approved drugs to SARS-CoV-2 M^{pro} with 20 or fewer poses per compound. The generated binding poses and structural information on M^{pro} were subjected to AI-guided INTENDD, which analyzed their poses and suggested twenty compounds as binding candidates for further experimental validation. The affinity of these compounds was measured by surface plasmon resonance (SPR), with 13 compounds exhibited M^{pro} binding signals in the SPR analysis. Remarkably, six of these compounds had not previously been tested against SARS-CoV-2 $M^{pro.}$ **Vorapaxar** and **dasabuvir** (Figure 13) showed a relatively high affinity for M^{pro}, with dissociation constants (K_d) of 27 and 3.1 μM, respectively. Due to the high cytotoxicity of **dasabuvir** (cytotoxicity at concentrations of 3 μM and higher in the preliminary experiments), only **vorapaxar** was tested for its cell-based antiviral activity in cells infected with SARS-CoV-2 using immunofluorescence detection. **Vorapaxar** inhibited the replication of SARS-CoV-2 with an EC_{50} of 11 μM and showed cytotoxicity with CC_{50} >50 μM in Vero E6 cells. The binding pose of **vorapaxar** proposed by AI-guided INTENDD suggests that it can bind to the active site of M^{pro} (PDB ID: 6Y2G) but does not fully occupy the binding region of the enzyme. The predicted binding interactions were not shown or discussed. Therefore, the authors concluded that modifications of **vorapaxar** to better interact with the M^{pro} binding pocket could enhance its inhibitory effect.



vorapaxar

Kd (SARS-CoV-2 M^{pro}) = 27 μM

EC_{50} (Vero E6) = 11 μM



vorapaxar

Figure 13. Identification of SARS-CoV-2 M^{pro} inhibitor using a CNN-based model of the AI-guided INTENDD platform.

Juárez-Mercado *et al.*[51] performed a high-throughput virtual screening of a large compound library (about 17 million compounds) using molecular fingerprinting to identify the compounds that bare structural similarities to known anti-coronavirus agents with different mechanisms of action. Next, they focused on identifying inhibitors of SARS-CoV-2 M^{pro}. From the screened hits, only commercially available compounds (4,423 compounds) were selected and subjected to seven ML algorithms within the Assay Central software[60] and to QSAR based on a RF algorithm developed for the identification of SARS-CoV-2 M^{pro} inhibitors.[61] These models used molecular descriptors as input features to predict compound activity. The authors did not specify which seven Assay Central Software algorithms were used. The compounds predicted as active by the consensus scoring of all seven ML algorithms and the compounds labeled as positive hits by the QSAR model were considered hits. This process led to the identification of eighteen potential M^{pro} inhibitors. Of these, twelve compounds were purchased for the evaluation of their inhibitory effect on M^{pro} using FRET-based assay. The most potent compound, **13**, inhibited M^{pro} activity with an IC_{50} of 3.5 μM and served as the lead compound in the subsequent analog search. Three analogs (**13a**, **13b** and **13c**; Figure 14) were identified, with analogs **13b** and **13c** inhibiting the enzymatic activity of M^{pro} with IC_{50} values of 29.1 and 1.8 μM, respectively. Further studies of cytopathic effect on SARS-CoV-2-infected Vero CCL81 cell cultures revealed that compounds **13a**, **13b** and **13c** exhibited antiviral *in vitro* activity in the mid-micromolar range, while compound **13** showed no antiviral activity in cells. The compounds also showed no detectable cytotoxicity in the concentration ranges tested (0.78–100 μM for compounds **13**, **13a** and **13b** and 0.39-50 μM for compound **13c**) in Vero CCL81 cells. Molecular docking of the compounds to the catalytic site of SARS-CoV-2 M^{pro} (PDB ID: 7JPZ) revealed pi-stacking interactions between the pyridone motif of the compounds and the side chain of His41. Other predicted interactions include interactions with the polar residues Gly143 and His163, the hydrophobic residue Leu141 and the negatively charged residue Glu166 which could be involved in enzymatic inhibition as they would interfere with substrate binding. However, molecular dynamics simulation of compound **13c** in complex with SARS-CoV-2 M^{pro} revealed changes in ligand conformations and possible displacement from the cavity.

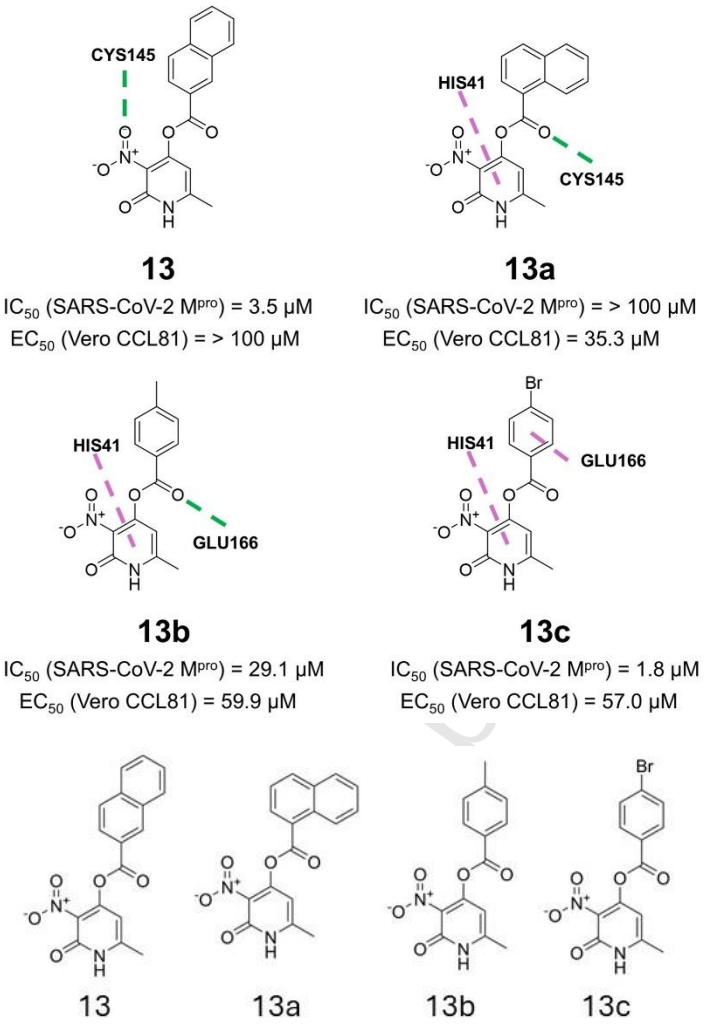


Figure 14. Identification of SARS-CoV-2 M^{pro} inhibitors using Assay Central software and QSAR based on a RF algorithm developed for the identification of SARS-CoV-2 M^{pro} inhibitors. Green dotted lines represent hydrogen bonds, while violet dotted lines indicate pi-stacking interactions.

In the study by Sun *et al.*[34] the authors used the generative chemistry platform Chemistry42 for the *de novo* design of novel small molecule inhibitors targeting the SARS-CoV-2 M^{pro}. Chemistry42 utilizes over 40 generative models, including generative autoencoders, generative adversarial networks, flow-based approaches, evolutionary algorithms, language models, and others using various molecular representations such as strings, graphs, and 3D shapes. The performance of the model is driven by scoring systems that assess how well each molecule could bind to the target, how easy it would be to synthesize, and other drug-like properties. Their goal was to develop covalent inhibitors, and the development of novel M^{pro} inhibitors was conducted in three rounds. In the first round, a structure-based design strategy was applied to identify promising molecules based on the key features of a known co-crystallized ligand (PDB ID: 6W63). The Chemistry42 platform was used to search for a good starting hit from a pre-selected set of commercially available compounds, and the experiment was guided by the use of 4 pharmacophores of co-crystalized ligand, pocket features (mandatory H-bond interaction with His67), a self-organized map (SOM) trained on cysteine protease inhibitors from ChEMBL, and synthetic feasibility as the driving force of the reward function. A small group of compounds was prioritized, and α-chloroacetamides showed strong submicromolar activity against SARS-CoV-2 M^{pro}. In the second round of the drug discovery process, Chemistry42 platform was used to expand and optimize the α-

chloroacetamide hit. Unlike the first round, this round did not rely on previous training data of known inhibitors and did not utilize the SOM module, allowing for broader investigation. The docked pose of the best hit was used as a template, with its pharmacophore and shape features determining the reward function. Due to the high reactivity of the chloroacetamide warhead, it was replaced with a less reactive α-chlorofluoroacetamide alternative. The third round was similar to the second round, but with the α-chlorofluoroacetamide warhead. Among the newly generated molecules, compound **1A** (Figure 15) proved to be the most potent inhibitor in fluorescence-based SARS-CoV-2 M^{pro} assay with an IC_{50} of 61 nM. However, **1A** showed significant off-target activity by inhibiting several CYP450 enzymes as well as low cellular efficacy and poor bioavailability in mice. Therefore, compound **1B** (IC_{50} = 121 nM; Figure 15), which was also developed in this round, exhibited a better ADME profile and was selected for further optimization. In the further optimization phase, the researchers aimed to improve the drug-like properties of compound **1B** by improving its permeability, minimizing the inhibition of CYP enzymes and increasing its oral bioavailability. By introducing a methyl group, they developed compounds **2** and **3**, which showed stronger enzymatic inhibition of SARS-CoV-2 M^{pro} in fluorescence-based assay with IC_{50} values of 16 nM and 14 nM, respectively. Compound **3** was particularly promising as it showed minimal CYP inhibition (IC_{50} = >50 μM) for all isoforms tested, indicating the lowest risk of metabolic disruption. Compounds **2** and **3** (Figure 15) showed good cellular activity in imunofluorescence-based inhibition assay of SARS-CoV-2 infected VeroE6 cells with EC_{50} = 68 nM and 71 nM, respectively. Compound **3** (**ISM3312**) possessed better *in vivo* pharmacokinetic and was therefore selected for further evaluation. Moreover, **ISM3312** exhibited low cytotoxicity of 86.8 μM in Vero E6 cells. To investigate the binding mechanism of **ISM3312**, the compound was crystallized in complex with SARS-CoV-2 M^{pro} (PDB ID: 8WTS). The fluorine group interacts only weakly with Cys145. Hydrogen bonds were observed with the residues Gly143, His163 and Glu166 and a pi stacking interaction with His41. These interactions stabilize the molecule in the active site. **ISM3312** also binds covalently to the catalytic residue Cys145, resulting in irreversible inhibition, as confirmed by enzymatic dilution experiments showing sustained activity over four hours. **ISM3312** showed broad spectrum activity in fluorescence-based M^{pro} enzymes assay of various human coronaviruses including Omicron P132H (IC_{50} = 22 nM) SARS-CoV-2 (IC_{50} = 14 nM) SARS-CoV (IC_{50} = 14 nM) MERS-CoV (IC_{50} = 52 nM) HCoV-229E (IC_{50} = 11 nM) HCoV-OC43 (IC_{50} = 10 nM) HCoV-NL63 (IC_{50} = 14 nM) HCoV-HKU1 (IC_{50} = 4 nM). Importantly, **ISM3312** has higher activity than Nirmatrelvir, a SARS-CoV-2 inhibitor in clinical use, and Ensitrelvir, a known non-covalent SARS-CoV-2 inhibitor. **ISM3312** also demonstrated superior or comparable antiviral cell activity against a spectrum of SARS-CoV-2 variants in cell-based SARS-CoV-2 cytopathic effect assay in Vero E6 cells, including strains Delta (EC_{50} = 0.1 μM), BA.2.3 (EC_{50} = 0.06 μM), BA.5 (EC_{50} = 0.03 μM), XBB.1 (EC_{50} = 0.04 μM) and EG.5 (EC_{50} = 0.08 μM). The antiviral activity of **ISM3312** against various human coronaviruses was further investigated using immunofluorescence-based assays in different cells. It showed lower EC_{50} values against two MERS-CoV variants (EMC-2012 (EC_{50} = 40 nM) and MA30 (EC_{50} = 50 nM) in Huh7 cells), 229E (EC_{50} = 19 nM in Huh7 cells), NL63 (EC_{50} = 167 nM in Huh7-hACE2 cells) and OC43 (EC_{50} = 56 nM Huh7 cells) strains compared to nirmatrelvir and ensitrelvir, indicating stronger antiviral activity. In addition, **ISM3312** showed potent broad spectrum antiviral activity at nanomolar concentrations in a physiologically relevant human airway organoid model and broad spectrum antiviral activity against SARS-CoV-2 variants and other human coronaviruses independent of P-glycoprotein modulation, as well as against various mouse coronavirus models. The SARS-CoV-2 FRET-based SARS-CoV-2 M^{pro} enzyme assay showed that **ISM3312** exhibited higher inhibitory activity against mutants resistant to nirmatrelvir. When the virus was repeatedly exposed to

ISM3312, the resulting mutations did not occur at the binding site of the compound on the viral enzyme. This finding suggests that **ISM3312** may be less likely to induce resistance mutations at the binding sites. In addition, although **ISM3312** inhibited nirmatrelvir resistant viruses, nirmatrelvir retained some inhibition against **ISM3312** resistant viruses. Therefore, sequential treatment with **ISM3312** and nirmatrelvir could combat drug resistant strains. These favorable pharmacological and safety profiles led to the selection of **ISM3312** as a preclinical candidate for COVID-19 treatment, which has successfully completed Phase I studies in China.

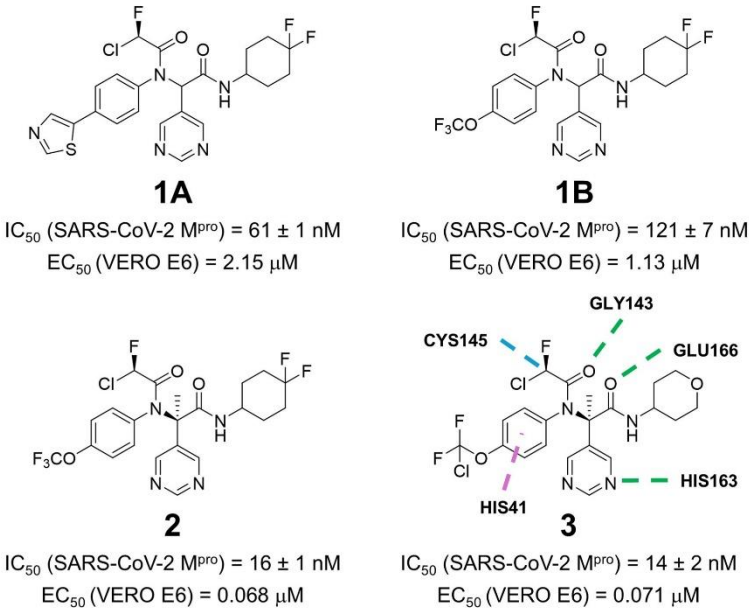


Figure 15. Identification of SARS-CoV-2 M^{pro} inhibitor using a generative chemistry platform Chemistry42. Green dotted lines represent hydrogen bonds, blue dotted lines represents covalent bonds, while violet dotted lines indicate pi-stacking interactions.

Comparison of the most potent compounds identified as SARS-CoV-2 M^{pro} inhibitors revealed strong enzyme inhibition, with most compounds exhibiting activity in the low micromolar range, as confirmed by cell-based assays. Despite their different chemical structures, the compounds showed similarities in interaction patterns. Two types of M^{pro} inhibition were investigated: reversible inhibition by non-covalent compounds and ireversible covalent inhibition. Covalent inhibitors formed a covalent bond with the catalytic residue Cys145, whereas non-covalent inhibitors typically interacted with this residue via hydrogen bonds. Other most common interactions included pi-stacking interactions with the catalytic residue His41 and hydrogen bonding with Glu166. These predicted interactions suggest that the compounds inhibit enzyme activity by interacting, either reversibly or irreversibly, with Cys145 and His41. They may also be involved in the dimerization of the protein, as the interaction of the compounds with Glu166 was suggested to be important for the dimerization of the enzyme.[62]

3.1.2. Antiviral Agents targeting SARS-CoV-2 Papain-Like Protease

Like SARS-CoV-2 M^{pro}, papain-like protease (PL^{pro}) is an enzyme responsible for polyprotein processing by cleavage, which is required for viral replication and evasion of the host immune response.[27,61] The active site contains a catalytic triad consisting of Cys112, His273, and Asp287.[61] Inhibiting the catalytic activity of PL^{pro} could interfere with viral processes, making PL^{pro} a promising target for antiviral drug discovery.

Anwaar *et al.*[39] employed a modified deep learning model called Deep Drug-Target Binding Affinity Prediction (DeepDTA), which utilizes CNNs to predict ligand-protein binding affinities. The model was trained using the KIBA dataset, which consists of experimentally evaluated inhibitors from the kinase protein family. The successfully trained model was then applied to 2,440 FDA-approved and 8,168 investigational drugs from DrugBank, and the binding affinities of these compounds were predicted against 24 SARS-CoV-2 viral proteins as KIBA scores. The model used SMILES strings of ligands and FASTA sequences of proteins as input features to learn and predict binding affinities. 168 drugs with high predicted KIBA scores were chosen for molecular docking at all protein targets. Among the compounds from different drug classes with the best combination of KIBA scores from the model and binding affinity values from molecular docking, 16 compounds showed the highest predicted inhibitory potential against key SARS-CoV-2 viral proteins. Moreover, the inhibitory effect of 5 compounds on SARS-CoV-2 PL^{pro} was experimentally evaluated using ubiquitin-rhodamine 110 Gly fluorescent intensity assay, of which the **rifapentine** and **flavin adenine dinucleotide disodium** (Figure 16) showed the highest inhibition of PL^{pro} activity with an IC_{50} of 15.18 and 12.39 μM, respectively. The binding modes predicted by molecular docking showed that the binding pocket of PL^{pro} (PDB ID: 6W9C) consists mainly of hydrophilic residues, which is why hydrogen bonding of **rifapentine** with residues Arg911 and Tyr1009 were predicted. Similarly, **flavin adenine dinucleotide disodium** was predicted to hydrogen bond with residues Asp909 and Gln1014.

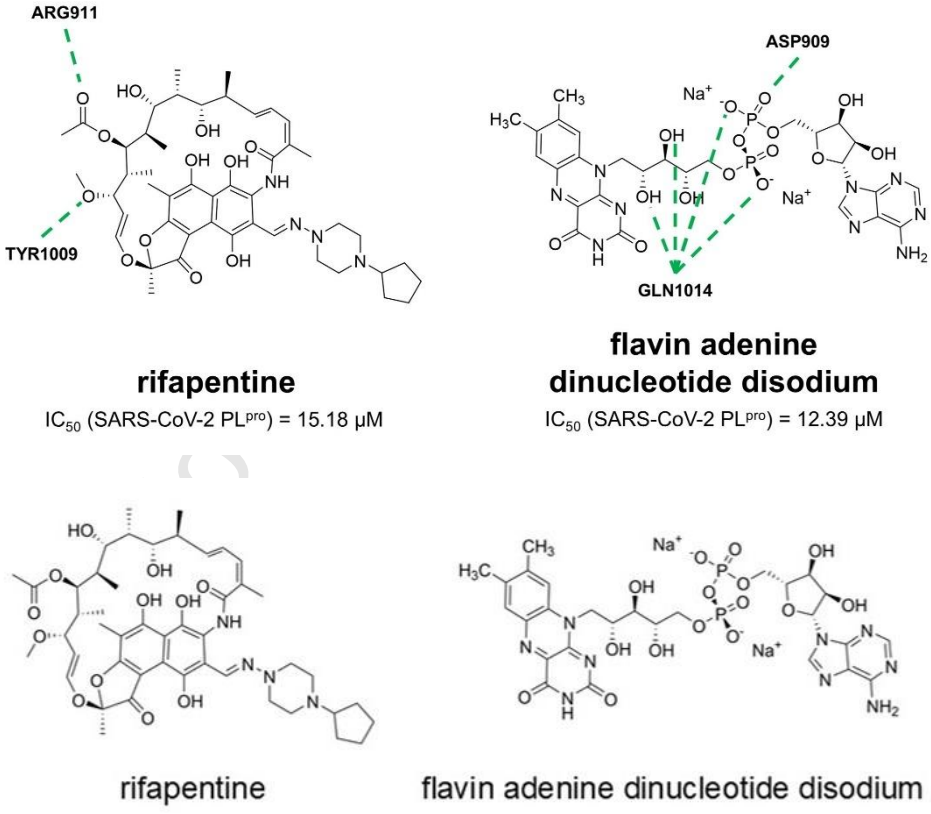


Figure 16. Identification of active SARS-CoV-2 PL^{pro} inhibitors using DeepDTA, a CNN-based model. Green dotted lines represent hydrogen bonds.

Garnsey *et al.*[63] conducted a study to identify inhibitors of SARS-CoV-2 PL^{pro}. They started the study with the ChEMBL database, which contains compounds with experimentally determined biological activities, and fragmented them into core structures and fragments. The cores and fragments were then categorized based on their effects on the compound's

activity. A neural network classification model was subsequently trained to predict which fragments might be active when bound to a particular core structure. The core structure of **GRL0617**[64] was applied as the basis for developing new compounds to which alternative fragments were bound based on the neural network predictions. The type of neural network used was not specified. The corresponding compounds were then synthesized and tested biologically. The result of this protocol was the lead compound **PF-07957472** (Figure 17), a potent and selective PL^{pro} inhibitor (K_i of 2 nM determined with FRET-based PL^{pro} inhibition assay) with favorable pharmacokinetic properties and potent cellular antiviral activity determined in SARS-CoV-2 infected differentiated normal human bronchial epithelial (dNHBE) cells by cytopathic effect assay (EC_{50} = 13.9 nM) as well as *in vivo* efficacy. The compound was also effective against the SARS-CoV virus. Crystallographic studies (PDB ID: 9CSY) showed that **PF-07957472** could bind to the same binding pocket as **GRL0617** but displayed optimized interactions leading to stronger and more stable binding. It forms important hydrogen bonds with Asp164, Glu167, Tyr264 and Gln269 and pi-stacking interactions with Pro247, Pro248 and Tyr268 (it is not clear which part of the compounds is involved in the interactions).

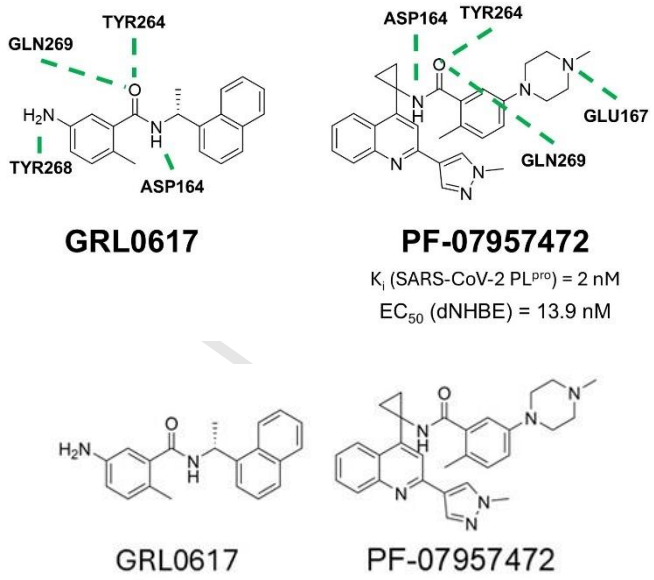


Figure 17. Identification of SARS-CoV-2 PL^{pro} inhibitor using neural network classification model. Green dotted lines represent hydrogen bonds.

The authors of the study by Pal *et al.*[27] aimed to identify inhibitors of the SARS-CoV-2 PL^{pro}. First, they performed a quantitative high-throughput screening to experimentally validate a subset of 15,093 compounds from the in-house library against the PL^{pro} enzyme. This subset was then used to develop and train several ML models, with the XGBoost classification model showing the best performance. The models used molecular fingerprints and molecular descriptors as input features. The model was applied to screen the in-house Genesis library of 101,220 compounds to predict whether the compounds are active or inactive. 125 virtual screening compounds predicted to be active were tested with fluorescence-based PL^{pro} inhibition assay, and 25 compounds were identified as SARS-CoV-2-PL^{pro} inhibitors, with IC_{50} values ranging from 1.75 to 36 μM. Subsequently, 100 analogous compounds were selected for further *in vitro* evaluation, and 20 of these analogs showed inhibitory activity against SARS-CoV-2 PL^{pro}, with IC_{50} values ranging from 9 to less than 23 μM. Finally, all 45 hits from the initial screening and similarity search were distributed among 14 different chemotypes, indicating chemical diversity. Based on these results, special attention was paid to compounds (**NCGC00473156** and **NCGC00473291**; Figure 18) with the

benzo[1,4]oxazepin-3(2H)-one-based chemotype (chemotype 7), which has an activity of less than 10 μM. This chemotype was prioritized for hit-to-lead optimization. Various analogs of chemotype 7 were synthesized without significantly improving the IC_{50} values.

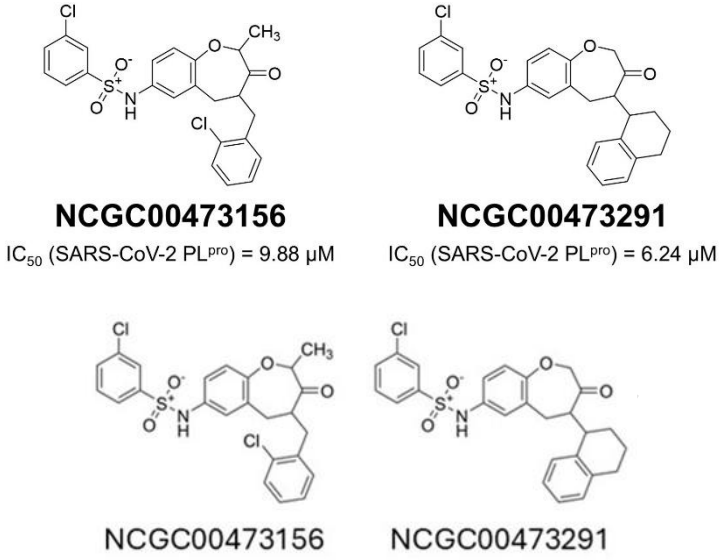


Figure 18. Identification of SARS-CoV-2 PL^{pro} inhibitors using an XGBoost classification model.

Conventional and deep learning ML methods again identified promising PL^{pro} inhibitors with antiviral activity in the low micromolar range. Analysis of their reported binding modes suggests that these compounds do not occupy the same binding pocket. It also appears that none of the compounds interact directly with the catalytic residues. The crystal structure of **PF-07957472** (Figure 16) shows that the binding overlaps with the substrate binding site but does not extend into the catalytic triad. For compounds reported in other studies, no detailed discussions of their mechanisms of action were provided. This may suggest that these inhibitors may function via different mechanisms of action, as the compounds are structurally different.

3.1.3. Antiviral Agents targeting SARS-CoV-2 Spike RBD–ACE2 Interaction

Infection by SARS-CoV-2 begins with the binding of the viral spike protein via its receptor binding domain (RBD) to the human protein ACE2.[39] Prevention of viral entry is therefore the first interruption of viral transmission.

The research group by Teshima *et al*.[59] applied the AI-guided INTENDD platform to search for curcumin analogs capable of inhibiting the binding of the SARS-CoV-2 RBD to the human ACE2 receptor. They selected 334 curcumin analogs from PubChem and subjected them to molecular docking at the RBD pocket of SARS-CoV-2, that plays a crucial role in RBD-ACE2 binding. The input features were obtained by preprocessing coordinate information on protein-ligand binding mechanisms (atomic species, 3D atomic positions, interactions and functional groups) using data from the Protein Data Bank and information on ligand activity from PDBbind. The complexes from the molecular docking were then entered into AI-guided INTENDD, and 13 commercially available compounds were selected based on the predicted probability. They were tested for RBD-ACE2 binding inhibition using immunoassay and antiviral activity in SARS-CoV-2–infected Vero E6 cells using an immunofluorescence-based cell assay. Five analogs, **demethoxycurcumin**, **bisdemethoxycurcumin**, **curcumin sulfate**, **monodemethylcurcumin** and **bisdemethylcurcumin** (Figure 19), together with **curcumin**, showed RBD-ACE2 binding inhibition with IC_{20} below 30 μM and antiviral activity

with EC_{50} below 30 μM. No cytotoxicity was observed in Vero E6 cells up to 30 μM for most compounds (**monodemethylcurcumin** had CC_{50} of 25.2 μM). Structure-activity relationship analyses showed that the double bonds in the carbon chain connecting the two phenolic groups are prerequisites for activity. Although the authors suspected a binding site (PDB ID: 6M0J) of the compounds predicted hydrogen bonding with Tyr453 and Asn501 and pi-stacking with Tyr505 as key interactions with RBD, further crystallographic studies are required confirm.

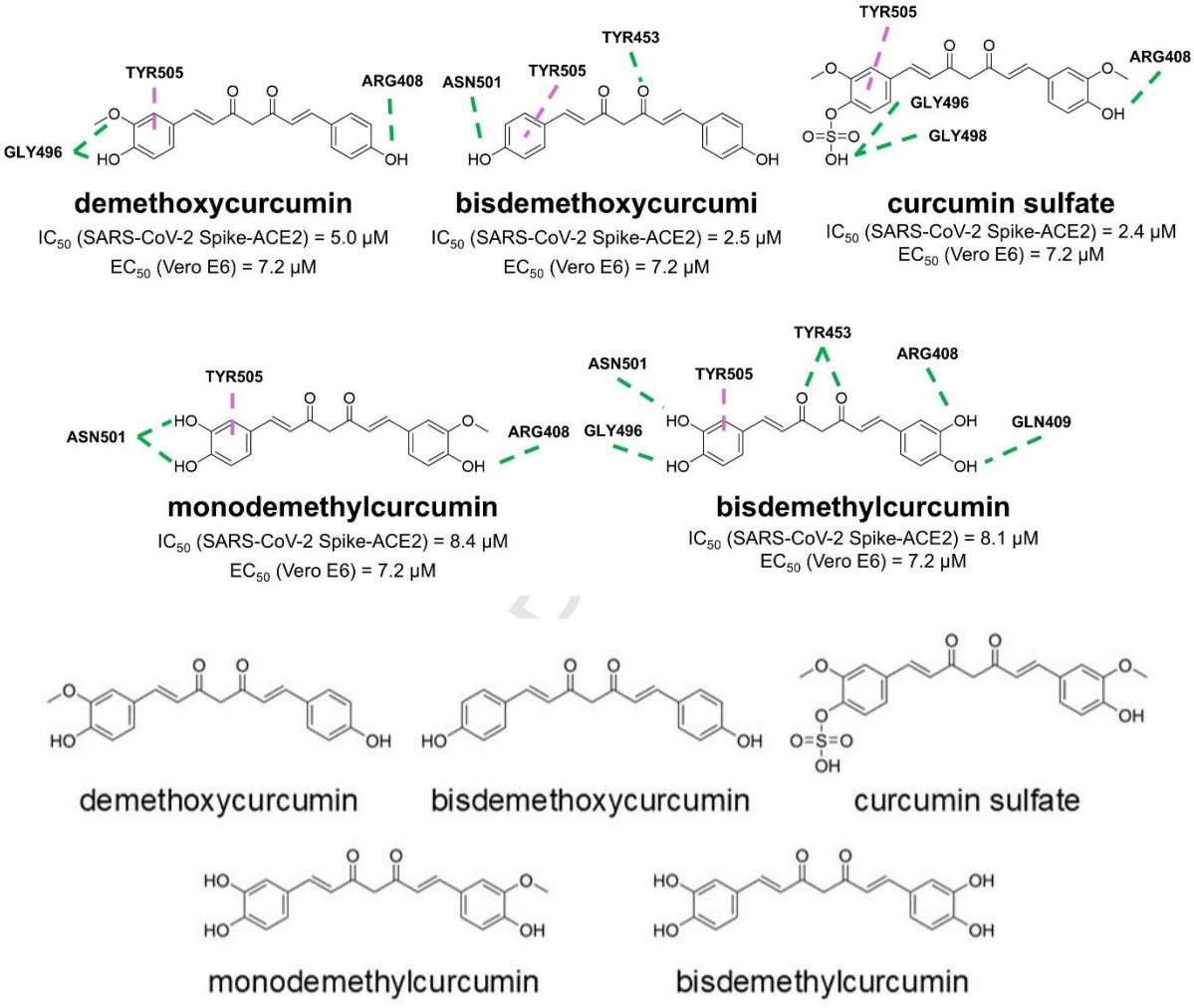


Figure 19. Identification of SARS-CoV-2 RBD-ACE2 binding inhibitors using a CNN-based model of the AI-guided INTENDD platform. Green dotted lines represent hydrogen bonds, while violet dotted lines indicate pi-stacking interactions.

Compounds that could block the interaction between the SARS-CoV-2 spike RBD and the human ACE2 receptor were identified by Gkekas *et al.*[40] They developed a pipeline that combined molecular docking with Smina and an AI-based rescoring function based on a 3D-CNN trained on a Database of Useful Decoys, Enhanced (DUD-E). This pipeline was applied to screen commercial libraries from ChemBridge Corporation. First, molecular docking was performed with Smina targeting the spike protein at the interaction site where spikes interact with ACE2 (PDB ID: 6M0J) to predict binding affinities, generating 50 different poses for each compound. The resulting protein-ligand complexes were then re-scored using an AI-based rescoring function to capture the binding mechanism by analyzing 3D structural features such as atomic species, spatial atomic positions, interatomic interactions, and functional groups, and to distinguish the most promising complexes. To further improve the

prediction accuracy, a fusion approach was applied by combining the affinities predicted by Smina with the AI-based scores. Finally, potential ligand hits were categorized into groups based on their structural similarity to ensure diversity. 31 hit compounds were experimentally validated for their inhibitory effect on the spike RBD/ACE2 interaction using the cell-based LuTHy assay (double-readout bioluminescence-based two-hybrid technology) in HEK293T cells. 6 compounds suppressed spike RBD/ACE2 interaction in a dose-dependent manner. All compounds were evaluated in a fluorescence-based cell entry inhibition assay with the SARS-CoV-2 pseudovirus in SH-SY5Y cells. Compounds 5979349 (**CKP-22**) and **6659989** reduced infections in a dose-dependent manner with IC_{50} values of 49 ± 2.5 and 3.2 ± 0.2 μM, respectively. Finally, both compounds were tested in an assay to reduce the cytopathic effect in SARS-CoV-2 infected Vero-E6 cells. The benzimidazole-based compound **CKP-22** (Figure 20) being the most potent, inhibiting the original SARS-CoV-2 virus *in vitro* with an IC_{50} of 50 μM. The only interaction predicted by molecular docking (PDB ID: 6M0J) of **CKP-22** was pi-stacking with Tyr505. The subsequent structural optimization of **CKP-22** led to the synthesis of 29 derivatives. Among them, 4-methoxy-1-(naphthalen-2-ylmethyl)-benzimidazole (**CKP-25**; Figure 19) proved to be the most potent, which effectively inhibited the spike-RBD/ACE2 interaction and reduced the cytopathic effect of SARS-CoV-2 in Vero E6 cells with an IC_{50} of 3.5 μM. Citotoxicity studies showed no toxicity in Vero E6 cells at active concentration. In addition, the viral load was reduced. It also proved to be a weak inhibitor of the CYP450 system, but is not biodegradable, is not metabolized in the liver and has no safety issues. It can, therefore, be considered a lead compound against COVID-19 infection. According to the molecular docking, **CKP-25** binds in the same way as **CKP-22** and forms additional hydrogen bonds (without specifying with which residue). Another molecular dynamic simulation at 100 nm showed stable binding due to pi-stacking with Tyr505 and hydrogen bonding mainly with Tyr496 or Arg403.

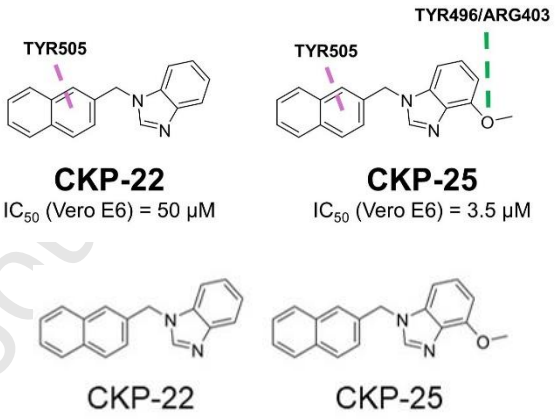


Figure 20. Identification of SARS-CoV-2 inhibitors of RBD-ACE2 binding based on 3D-CNN model. Green dotted lines represent hydrogen bonds, while violet dotted lines indicate pi-stacking interactions.

Inhibitors that prevent the binding of SARS-CoV-2 RBD to ACE2 have been identified with CNN. Curcumin derivatives and **CKP-25** (Figure 19) show strong antiviral cell activity in the low micromolar range. However, due to the use of different assay formats, a direct comparison of their efficacy may not be entirely reliable. From the comparison of the predicted binding modes, it could be concluded that the compounds from both studies occupy the same binding pocket, as interactions with the same residues were observed. Pi-stacking of compounds with Tyr505 and hydrogen bonding with residue Tyr496 are common among the compounds, suggesting that these compounds may be critical for activity. However, the compounds reported by Teshima *et al.*[59] exhibit a larger chemical structure,

which allows them to occupy a broader region within the binding pocket.**3.1.4. Antiviral Agents targeting Other Known SARS-CoV-2 Targets**

Several host proteases, such as M^{pro}, PL^{pro} and TMPRSS2, have been extensively studied as potential targets for SARS-CoV-2 inhibition. Another host protease that has been targeted is cathepsin L (CTSL), which is involved in the cleavage of the spike protein of SARS-CoV-2. In addition to protease inhibitors, other therapeutic approaches targeting viral structural proteins such as the SARS-CoV-2 nucleocapsid protein, which plays a crucial role in RNA binding and packaging of the viral genome, have also been explored.

Using Chemprop, a directed message-passing neural network (D-MPNN), Yang *et al.*[53] aimed to identify compounds that could potentially inhibit CTSL, a promising target for SARS-CoV-2. The model was trained with a library of compounds from PubMed that were evaluated against CTSL. Each molecule was translated from its SMILES strings into a molecular graph structure, with the nodes representing the atoms and the edges representing the chemical bonds. A graph network model was then trained to predict whether a molecule inhibits the activity of CTSL. The best performing model was subsequently used to identify potential CTSL inhibitors from the library of bioactive selleck compounds and the ZINC15 *in vitro* database. Compounds were ranked according to their prediction scores to inhibit CTSL activity, and the top 50 predicted compounds were selected for further biological experiments in a fluorescence-based CTSL activity assay. The top five compounds, namely **Mg-132**, **Z-FA-FMK**, **leupeptin hemisulfate, Mg-101**, and indeed **calpeptin** inhibited CTSL activity with IC_{50} values of 12.28, 54.87, 5.77, 5.77 and 43.98 nM, respectively. All 5 compounds also suppressed SARS-CoV-2 pseudovirus infection in human hepatoma cells (Huh7) cells with half-maximal effective concentrations (EC_{50}) of 212.50 nM, 701.20 nM, 39.29 μM, 3.12 μM, and 29.82 μM, respectively. The results were confirmed using human lung epithelial carcinoma cell line (A549) with adenovirus-mediated overexpression of human ACE2 to increase pseudovirus infection rates with all 5 compounds suppressing SARS-CoV-2 pseudovirus infection with EC_{50} values of 159.20 nM, 1.13 μM, 4.26 μM, 1.15 μM, and 34.81 μM, respectively. Moreover, the efficacy of these compounds against the SARS-CoV-2 B.1.351 (beta) variant of the pseudovirus was investigated in Huh7 and similar or enhanced inhibitory effects were observed. All 5 molecules also suppressed infection with the SARSCoV-2 B.1.351 pseudovirus in Huh7 cells and in A549 cells overexpressing TMPRSS2 with similar EC_{50} values, suggesting that these drugs remain effective even after TMPRSS2 overexpression, which is preferentially used to CTSL. All pseudovirus assays were performed using luciferase assay. The cytotoxicity results in Huh7 cells showed that **Mg-132**, **Z-FA-FMK**, **leupeptin hemisulfate**, **Mg-101** and **calpeptin** had CC_{50} values of 100.22 μM, 328.10 μM, 6.00 mM, 64.32 μM, and 115.70 μM, respectively. The three most potent compounds, **Mg-132**, **Z-FA-FMK**, and **leupeptin hemisulfate** (Figure 21), were further tested for validation of antiviral activity with live SARS-CoV-2 viruses in Huh7 using an immunofluorescence assay. All three compounds inhibited SARS-CoV-2 infection in the micromolar range with EC_{50} values of 0.21, 9.61, and 10.55 μM, respectively. The ML model was further applied to a library of 3,177 FDA-approved compounds, with **daptomycin** (Figure 21) identified as the most promising candidate for the inhibition of CTSL. In a fluorescence-based enzyme assay, **daptomycin** exhibited an IC_{50} of 7.87 μM. Subsequent cell-based assays with SARS-CoV-2 pseudovirus infection in Huh7 cells showed antiviral activity with EC_{50} values of 207.27 μM and comparable efficacy against the SARS-CoV-2 B.1.351 pseudovirus in Huh7 cells (EC_{50} = 218.84 μM). **Daptomycin** also showed stronger inhibition in A549 cells (EC_{50} = 98.81 μM) and retained its antiviral activity after overexpression of TMPRSS2 in both Huh7 (EC_{50} = 226.16 μM) and A549 (EC_{50} = 89.69 μM) cell lines. Toxicity studies in Huh7 cells showed low cytotoxicity for **daptomycin** with a CC_{50}

value of 3.57 mM. Molecular docking was performed to predict the binding pattern of these inhibitors to CTSL (PDB ID: 5MQY). **Z-FA-FMK**, **Mg-132**, and **leupeptin hemisulfate** are known covalent inhibitors, therefore covalent docking was performed for these compounds, while non-covalent docking was performed for daptomycin. Hydrogen bonds were predicted for all compounds, with **Mg-132** forming two hydrogen bonds with residues Gln19 and Gly68, **Z-FA-FMK** forming two hydrogen bonds with residues Gly68 and Asp162, **leupeptin hemisulfate** forming three hydrogen bonds with residues Gln19, Gly68 and Asp162, while **daptomycin** forming three hydrogen bonds with residues Asp160, Met161 and Asp162. Covalent bonds were predicted for the covalent inhibitors as shown in Figure 20. Gln19, Gly68 and Asp162 were concluded as key residues within the CTSL binding pocket.

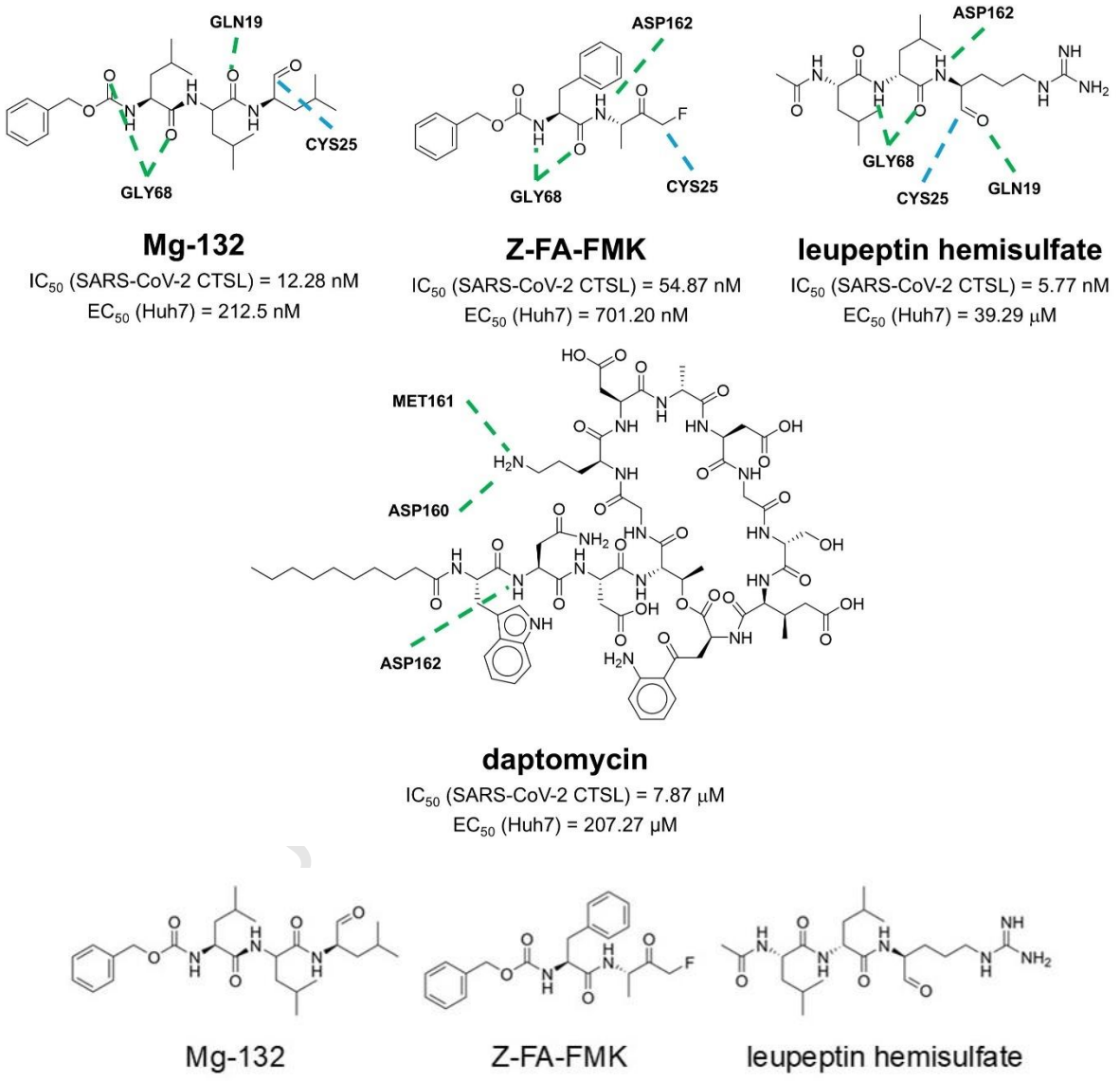


Figure 21. Identification of active SARS-CoV-2 CTSL inhibitors using D-MPNN. Green dotted lines represent hydrogen bonds, while blue dotted lines indicate covalent bonds.

Instead of focusing on functional proteins, which are usually studied, Xiang *et al.*[41] aimed to identify the inhibitors of the structural proteins of SARS-CoV-2. They targeted the SARS-CoV-2 nucleocapsid protein, which plays a critical role in binding to viral RNA, a crucial step in packaging the viral genome into new virions. The authors applied the structure generator called “Exhaustive Molecular Library Production In a Scaffold-REtained Manner” (EMPIRE)[65] and CNN-based model DeepFrag[66] to develop new compounds. EMPIRE combines a

variational AE and a building block-based method to generate new fragments that are attached to the scaffold structure, while DeepFrag relies on a deep CNN to optimize fragment placement. The EMPIRE structure generator was used to create analogs of a known nucleocapsid protein inhibitor with a phenanthridine scaffold by attaching a fragment library of benzene substructures. Input features for EMPIRE included the SMILES strings of phenanthridine derivatives, selected connection points on the scaffold, and fragment-based substructures derived from a curated library. This resulted in approximately 1,000 new compounds. Separately, the DeepFrag model analyzed the structure of the phenanthridine scaffold-ligand complex with the nucleocapsid protein and predicted which fragments could improve interactions with the binding site, suggesting the most suitable fragments to attach to the phenanthridine scaffold. DeepFrag generated over 10,000 optimized analogs. DeepFrag used the 3D structures of the protein and the molecule, binding pocket locations, and surrounding amino acid residue properties as input features to propose modifications to the fragment that could improve the binding of the drug to the SARS-CoV-2 nucleocapsid protein. Molecular docking was subsequently applied to screen compounds from both approaches, identifying 44 compounds with suitable binding potentials. These compounds were synthesized and tested for their binding activity to the SARS-CoV-2 nucleocapsid protein using SRP. The results showed that compound **38** (Figure 22) exhibited significant binding potential with a K_d of 0.35 μM, which was further validated by microscale thermophoresis, which yielded a K_d of 0.73 μM. These results confirm the strong binding of compound **38** to the SARS-CoV-2 nucleocapsid protein. *In vitro* antiviral activity assays revealed that compound **38** can inhibit the replication of SARS-CoV-2 virus in Vero E6 cells, with an EC_{50} of 11.3 μM determined in MTT (3-[4,5-dimethylthiazol-2-yl]-2,5 diphenyl tetrazolium bromide) assay, with no cytotoxicity observed (CC_{50} > 200 μM) in Vero E6 cells. The microscale thermophoresis assay confirmed that compound **38** had a higher binding affinity to the N protein than to the viral RNA and prevented the RNA from efficiently binding to the nucleocapsid protein. At a concentration of 20 μM, compound **38** reduced the interaction between viral RNA and nucleocapsid protein, and at 100 μM, compound **38** almost completely blocked this interaction. These data confirm that compound 38 effectively inhibits the binding of nucleocapsid protein to RNA. Molecular docking (PDB ID: 6VYO) mainly predicted van de Waals and pi-stacked interactions of the nucleocapsid protein with compound **38** involving Arg68, Ile131 and Trp132. The authors identified mutations of these residues as critical for the binding of compound **38**.

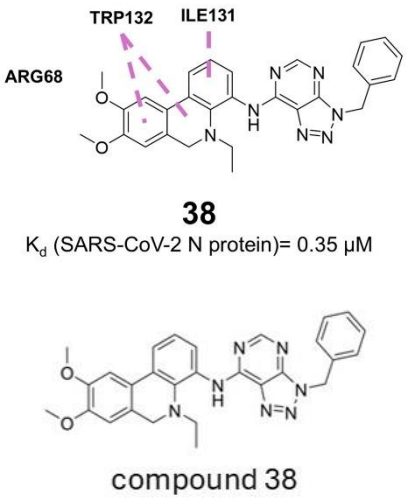
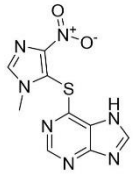


Figure 22. Identification of SARS-CoV-2 N-protein inhibitor using EMPIRE structure generator and CNN-based model DeepFrag. Violet dotted lines indicate pi-stacking interactions.

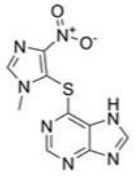
3.1.5. Antiviral Agents targeting Unidentified SARS-CoV-2 Targets

Several studies have identified potential SARS-CoV-2 antiviral agents that have shown activity in cell-based assays; however, their specific targets and mechanisms of action remain unclear.

The open-source drug discovery platform DeepChem was used by Ali *et al.*[29] to provide a prediction model based on a GCNN. The model was trained with compounds experimentally tested against SARS-CoV-2 from the NCATS repository that were classified as active or inactive. Compounds were presented as SMILES strings and each molecule was transformed into a molecular graph in which the atoms were represented as nodes and the bonds as edges, and the input features included atomic-level chemical properties and bond connectivity. The model was trained to map the structural features of the compounds to their functional activity and was then applied to approximately 700,000 compounds from the Korea Chemical Bank to predict their potential against SARS-CoV-2, with compounds having a higher score showing a higher probability of antiviral activity. The top 100 compounds were purchased, and their antiviral activity was tested *in vitro* on Caco-2 cells (human lung cancer cell line) using enzyme-linked immunosorbent spot (ELISpot) assay. 19 compounds showed inhibition of cell infection by SARS-CoV-2 by up to 80%. These 19 compounds were further tested in Vero E6 cells using enzyme-linked immunosorbent spot (ELISpot) assay, with 8 compounds showing dose-dependent antiviral activity with EC_{50} values between 1 and 7 μM. In addition, the compounds exhibited cytotoxicity higher than 50 μM in Vero E6 cells. To elucidate their molecular mechanism, the researchers applied a computational method called the similarity ensemble approach to predict the biological target based on structural similarity to other compounds with known activities for specific targets. They concluded that their compounds may exert their antiviral effects indirectly by targeting specific host cell proteins and act as indirect antiviral agents by modulating host pathways. To identify direct-acting antiviral compounds, they improved their model by including nucleoside analogue inhibitors such as remdesivir and molnupiravir with known direct antiviral mechanisms in the training group. To identify potential nucleoside analogues among the first 1,000 compounds predicted by the model, they used the t-SNE clustering technique. Nine compounds were tested to quantify intracellular virus concentrations in Calu-3 cells infected with SARS-CoV-2 using immunofluorescence staining. One compound, **azathioprine** (Figure 23), showed antiviral activity similar to established nucleoside analogue inhibitors and its antiviral effect was confirmed in Calu-3 cells infected with SARS-CoV-2 by plaque assay and Western blot. Low cytotoxicity was also observed in Calu-3 cells, although no exact values for antiviral activity or cytotoxicity were reported for **azathioprine**. As **azathioprine** is a prodrug, its metabolite, **thioinosinic acid**, was also investigated. Both compounds displayed *in vitro* activity against SARS-CoV-2, with **thioinosinic acid** also significantly reducing the viral load in the lungs of mice.



azathioprine



azathioprine

Figure 23. Identification of SARS-CoV-2 nucleoside analogue inhibitor using DeepChem platform with models based on GCNN and t-SNE.

"eVir", an advanced AI platform was utilized by Bess *et al.*[36] to identify FDA-approved drugs for repurposing against SARS-CoV-2. The main goal of their study was to find drugs that mimic the action of AVPs. AVPs are fragments of human proteins that are known to interfere with important viral processes such as entry, fusion, and replication. The idea is that drugs, which function similarly to AVPs, could also inhibit the virus. To predict how drugs might interact with the virus, eVir first created representations of both drugs and AVPs and assessed their effects on these viral mechanisms. Input features included drug-protein interaction networks, protein-protein interaction networks, mechanistic AVP profiles, and node embeddings generated with Node2Vec that captured topological and functional relationships within the biological network. The predicted effects of the drugs were then compared with those of the AVPs. Siamese neural networks (SNNs) were used to assess the similarity between the drugs and the AVPs based on their predicted effects. Three separate SNNs were created for each viral mechanism (entry, fusion, and replication). This allowed the system to evaluate how closely a drug resembles the action of an effective AVP for each of these key viral mechanisms. To further refine the model, MLP models were created for each antiviral mechanism to predict whether a drug could be classified as similar to AVPs. The predictions of SNNs and MLPs were then combined into a composite score reflecting the predicted impact of each drug on viral entry, fusion, and replication. This final score was applied to identify promising antiviral candidates. Drugs that performed the best in this analysis included **zinc**, **mebendazole**, **levomenol**, **gefitinib**, **niclosamide**, and **imatinib**. These compounds were tested against SARS-CoV-2 infected Calu-3 cells using an immunostaining assay, with **mebendazole** (IC_{50} of 48.55 nM), **gefitinib** (IC_{50} of 62 nM), **topotecan** (IC_{50} of 9.362 nM), **niclosamide** (IC_{50} of 982.5 nM), and **carfilzomib** (IC_{50} of 5.741 nM) showing strong antiviral activity (Figure 24). Their IC_{50} values were lower than those of remdesivir, and they displayed minimal to no cytotoxic effects in Calu-3 cells as indicated by CC_{50} values: **mebendazole** (not determined), **gefitinib** (4.4 μM), **topotecan** (169 nM), **niclosamide** (20 μM), and **carfilzomib** (184 nM).. Moreover, **mebendazole** (Figure 16) was tested in mice infected with SARS-CoV-2 and exhibited a 44.2% reduction in viral load in the lungs compared to an untreated placebo control.

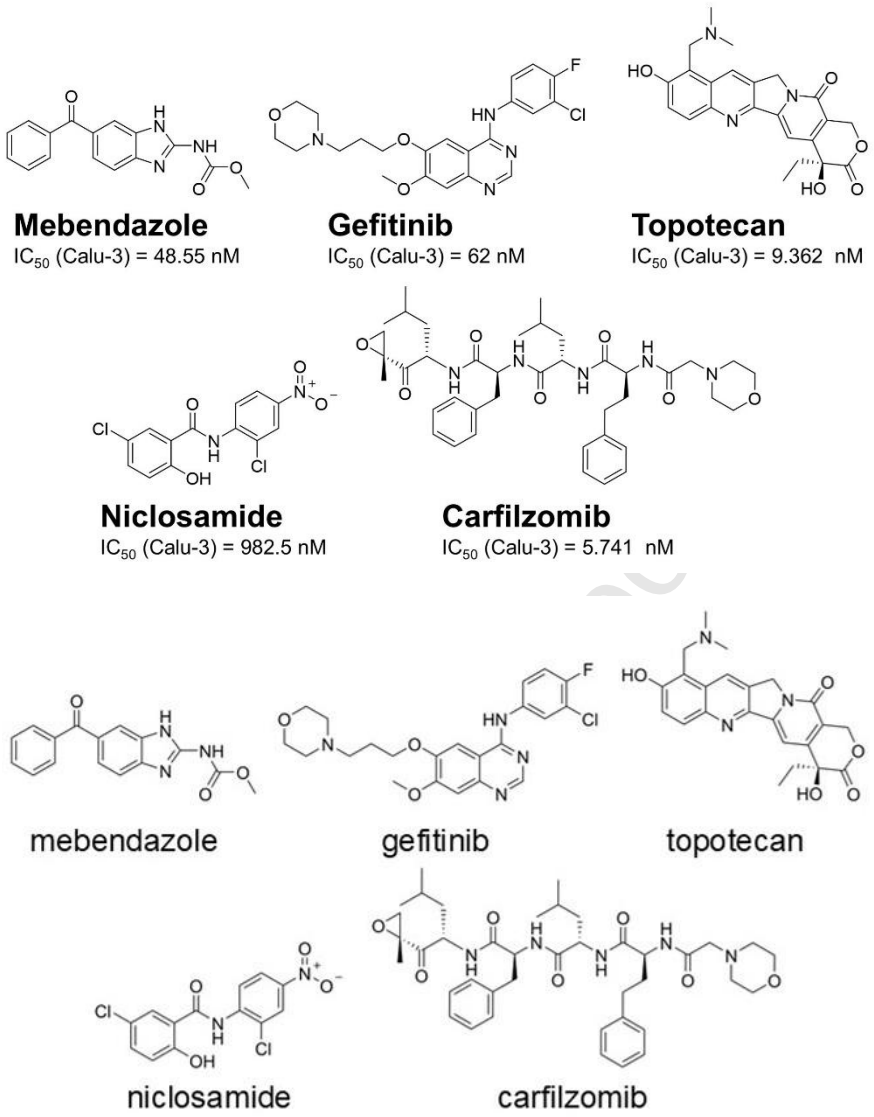


Figure 24. Identification of SARS-CoV-2 inhibitors of viral entry, fusion and replication mechanisms using models based on SNN and MLP from the eVir platform.

Van der Horts *et al.*[45] searched for antiviral agents against SARS CoV-2 among drugs that had undergone at least Phase I trials in humans. In their study, they used various experimentally evaluated compounds from PubChem and ChEMBL, including the ones tested against SARS-CoV-2, to train a GNN model that predicts the probability of activity. Each molecule was represented as a molecular graph structure, with the nodes representing the atoms and the edges representing the chemical bonds, and input features included atom types and bond types. The model was applied to 2,425 compounds that had undergone at least Phase I trials in humans but had not been tested against SARS-CoV-2. These compounds were ranked according to the likelihood of their antiviral activity, filtering out those that could cause adverse effects or cytotoxicity. A scoring function was designed to favor compounds with a high probability of direct inhibition by one of the molecular mechanisms (ACE2, M^{pro}, or TMPRSS2). In this way, six promising compounds were identified for further investigation. Two compounds, **PKI-179** and **MTI-31** (Figure 25), initially identified as inhibitors of the PI3K-mTORC1/2 pathway, have demonstrated significant antiviral activity against SARS-CoV-2 at low micromolar concentrations in Vero E6 cells

expressing human TMPRSS2 (IC_{50} = 0.3 μM and <5 μM, respectively) as well as in Calu-3 cells (IC_{50} = 0.4 μM and 0.15 μM, respectively) as determined by quantitative polymerase chain reaction (qPCR) analysis. Remarkably, both compounds outperformed the known mTOR inhibitor rapamycin. They also showed a broad spectrum of antiviral activity against different SARS-CoV-2 variants (Delta and Omicron) as well as against other coronaviruses (SARS-CoV, MERS-CoV and HCoV-229E). No apparent cytotoxicity was observed for the antiviral active doses in Vero E6 cells expressing human TMPRSS2 and Calu-3 cells. In addition, both compounds exhibited antiviral activity in physiologically relevant primary cultures of human airway epithelia derived from healthy donors, underlining their potential as therapeutic agents against COVID-19.

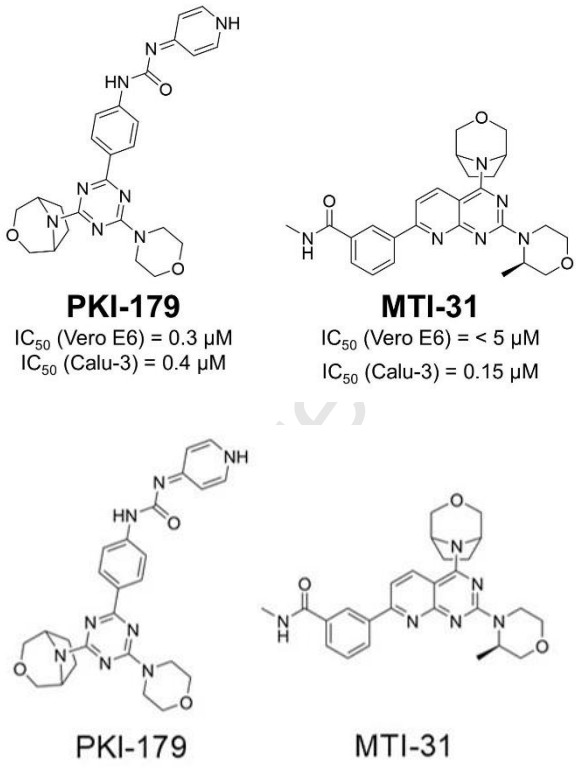


Figure 25. Identification of SARS-CoV-2 agents using a GNN model.

3.2. **Antiviral Inhibitors against Viruses beyond SARS-CoV-2**

While most studies have focused on identifying antiviral agents against SARS-CoV-2, agents against other viruses and their respective targets have also been investigated. These include antiviral agents targeting the neuraminidase of influenza A and B viruses, the cyclin-dependent kinase of varicella zoster virus, the DNA polymerase of human cytomegalovirus and the Gc protein of Rift Valley fever virus.

The aim of the study by Kovalishyn *et al.*[42] was to develop QSAR models to identify new oxazole derivatives with activity against the varicella zoster virus. The Online Chemical Database and Modeling Environment (OCHEM) platform was used for this purpose. A set of 287 compounds from the ChEMBL database evaluated against varicella zoster virus was applied for model development and training. The best QSAR models were built using a transformer convolutional neural network (Trans-CNN) and a transformer convolutional neural fingerprint (Trans-CNF). Therefore, a consensus model representing the average of the two models was applied for the quantitative evaluation of potential antiviral agents. Model used SMILES strings of compounds to retrieve their different information. Fourteen

virtual oxazole derivatives were generated based on available synthetic blocks and reactions, and screened against the consensus model. Seven compounds that were predicted to be the most active were synthesized and tested *in vitro* using cytopathic effect assay. Oxazole derivatives **3** and **9** (Figure 26) showed the highest activity against varicella zoster virus in human foreskin fibroblasts cells, with EC_{50} values of 4.01 and 0.10 μM, respectively. Compound **9** exhibited a strong inhibitory activity but was associated with high cytotoxicity (CC_{50} of 3.11 μM) in the same cell line. In the subsequent molecular docking studies (PDB ID: 4Y72), compound **9** elucidated possible mechanisms of antiviral action by targeting a cyclin-dependent kinase involved in the viral life cycle by forming electrostatic, van der Waals and hydrogen bond interaction (Tyr15) with surrounded residues.

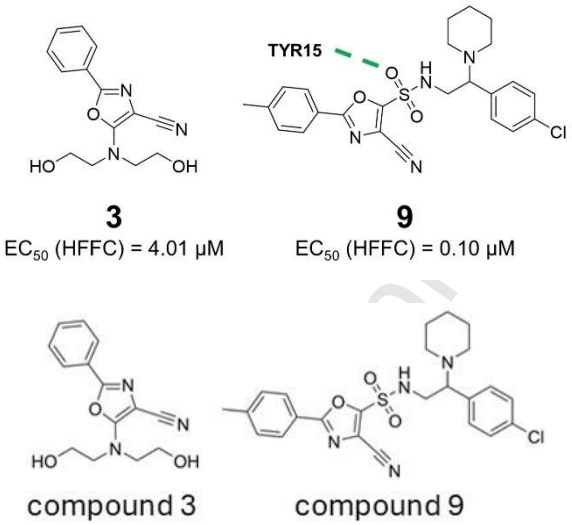


Figure 26. Identification of agents against varicella zoster virus using a consensus QSAR model based on Trans-CNN and Trans-CNF within the OCHEM platform. Green dotted lines represent hydrogen bonds.

The same research group also applied the OCHEM platform to create a QSAR model to predict potential inhibitors against human cytomegalovirus (HCMV).[67] The model was trained and validated using 675 compounds from the ChEMBL database with measured activity against HCMV using molecular descriptors (RDKit) as input features. Four different models were created: knowledge analysis-RF (WEKA-RF), deep learning consensus architecture (DLCA; combines three separate deep learning neural networks), associative neural networks (ASNNs; ensemble of feed-forward backpropagation neural networks with kNN), and XGBOOST. All models showed similar performance, so a consensus model was created as a simple average of the four models. Twenty-four virtual azole derivatives were generated and submitted to the QSAR model for prediction. Five compounds with the highest predicted activity were selected for subsequent synthesis and evaluation of antiviral activity using cell-based cytopathic effect assay. Compounds **1** and **5** (Figure 27) exhibited a high activity against the wild-type strain of HCMV in the human foreskin fibroblast cell line (EC_{50} = 0.18 μM for **1** and EC_{50} = 5.58 μM for **5**) Compounds also exhibited a degree of cytotoxicity in human foreskin fibroblast cells (CC_{50} = 86.23 μM for **1** and CC_{50} = 106.45 μM for **5**). They were less active in further virus yield reduction assay of HCMV in human foreskin fibroblast (EC_{90} = >20 μM for **1** and EC_{90} = 19.76 μM for **5**) and also less active against the resistant HCMV isolate (EC_{90} >20 μM for **1** and **5**). Inverse molecular docking of compounds **1** and **5** identified DNA polymerase as the most likely biotarget. The binding of compound **1** was stabilized by hydrogen bonding with Lys60 and hydrophobic interactions, while four

hydrogen bonds with Lys60 and Val136 and hydrophobic interactions were predicted for compound **5**.

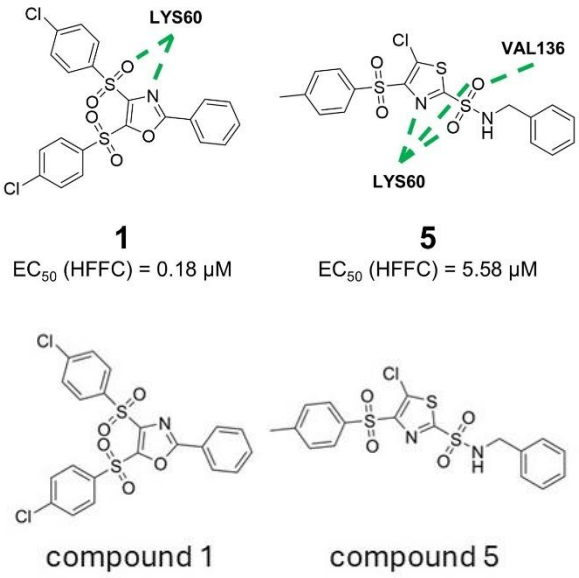


Figure 27. Identification of agents against human cytomegalovirus using consensus QSAR model of WEKA-RF, DLCA, ASNN and XGBOOST within the OCHEM platform. Green dotted lines represent hydrogen bonds.

Alkan *et al.*[30] aimed to identify inhibitors of the Rift Valley fever virus (RVFV), the main cause of the mosquito-borne disease Rift Valley fever, which is listed as a priority disease in the World Health Organization's research and development list. In the study, the authors applied reinforcement learning based on Markov decision process (MDP) using the deep learning-based platform AtomNet[68] to predict the bioactivity of small compounds. AtomNet analyzes the 3D structure of the target protein and the library of compounds and predicts the probability that a given compound will bind effectively to the target protein. The potential drug candidates are then ranked based on their predicted binding affinity. In this study, the authors searched for small-molecule that could block the RVFV fusion process by targeting the Gc protein, which is involved in the fusion of the viral envelope with the host cell membrane. Screening and ranking of approximately 8 million commercially available compounds of Mcule on the post-fusion RVFV Gc protein resulted in 94 potential hits that were selected for biological evaluation. After analyzing virus titers in RVFM infected Vero cells using plaque assay, three compounds, **RGc-B05**, **RGc-F06** and **RGc-F12** (Figure 28), were identified with IC_{50} values of 10.98, 20.61, and 11.90 μM, respectively. These compounds were able to block the viral entry when cells were pretreated prior to infection but showed no effect when applied post-infection. This suggests that they prevent the virus from entering the cell and do not affect the later stages of the viral life cycle. Cytotoxicity was assessed in Vero cells, with **RGc-B05** and **RGc-F06** at 200 μM showing a reduction in cell viability of 8.6% and 6.3% respectively, while **RGc-F06** at the same concentration had no effect on cell viability. Further tests showed that **RGc-F06** also exhibits some antiviral activity against another virus from the Phenuiviridae family, namely the Arumowot virus.

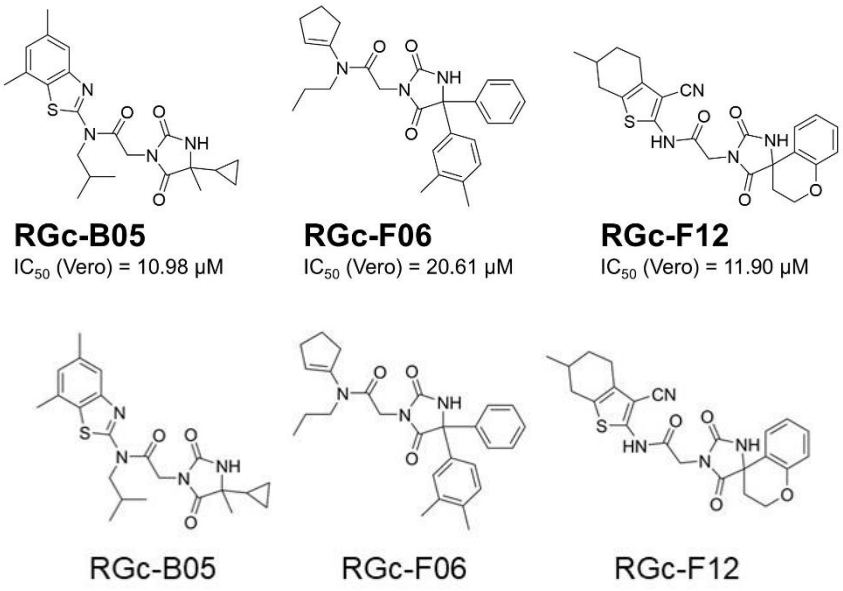


Figure 28. Identification of Rift Valley fever virus inhibitors using MDP model of AtomNet platform.

The authors of the study by Izmailyan *et al.*[69] applied a MDP model to develop a series of compounds targeting the neuraminidase of influenza A and B viruses. Starting from the crystal structures of the neuraminidases of influenza A and B viruses without the co-crystallized ligands, the reinforcement learning algorithm explored all possible modifications of the benzene ring by transforming it into more complex structures, without any prior knowledge of the existing compounds. The result was a total of 60,291 compounds, 86.5% of which achieved a better reward score than the reference compound oseltamivir, indicating the majority of the new compounds outperformed it in terms of desired properties. To narrow down the selection, several filters were applied together with molecular dynamics simulations, resulting in the identification of 31 compounds whose binding free energies were similar or better to the one of oseltamivir. Finally, nine compounds were synthesized and evaluated in *in vitro* experiments using fluorescence-based neuraminidase enzyme inhibition assays. Two compounds, **DS-22-inf-009** and **DS-22-inf-021** (Figure 29), were found to be effective inhibitors of neuraminidases of both influenza A (IC_{50} of 24.8 μM for **DS-22-inf-009** and 11,2 μM for **DS-22-inf-021**) and influenza B viruses (IC_{50} of 21.2 μM for **DS-22-inf-009** and 9.1 μM for **DS-22-inf-021**) and were further investigated both in a cell-based *in vitro* assay and *in vivo*. In a cell-based cytopathic effect assay in infected Madin-Darby canine kidney (MDCK) cells, both compounds inhibited various influenza A and influenza B strains in a dose-dependent manner with EC_{50} < 2.5 μM. The compounds also showed no cytotoxicity to MDCK cells at the concentrations tested (CC_{50} > 50 μM). In addition, animal studies showed that both **DS-22-inf-009** and **DS-22-inf-021** protected mice with survival rates between 65% and 100%, with no signs of toxicity. Molecular dynamic simulations revealed stable binding of the two compounds to the neuraminidase enzymes of influenza A and B during the simulation. In influenza A neuramidase (PDB ID: 6HP0), **DS-22-inf-009** forms hydrogen bonds with Asp151 and Arg368, while in influenza B neuramidase (PDB ID: 4CPY) the hydrogen bonds occur with Arg149, Arg291 and Arg373. **DS-22-inf-021** forms hydrogen bonds with Arg293 and Arg368 in influenza A neuramidase and hydrogen bonds with Arg291 in influenza B neuramidase.

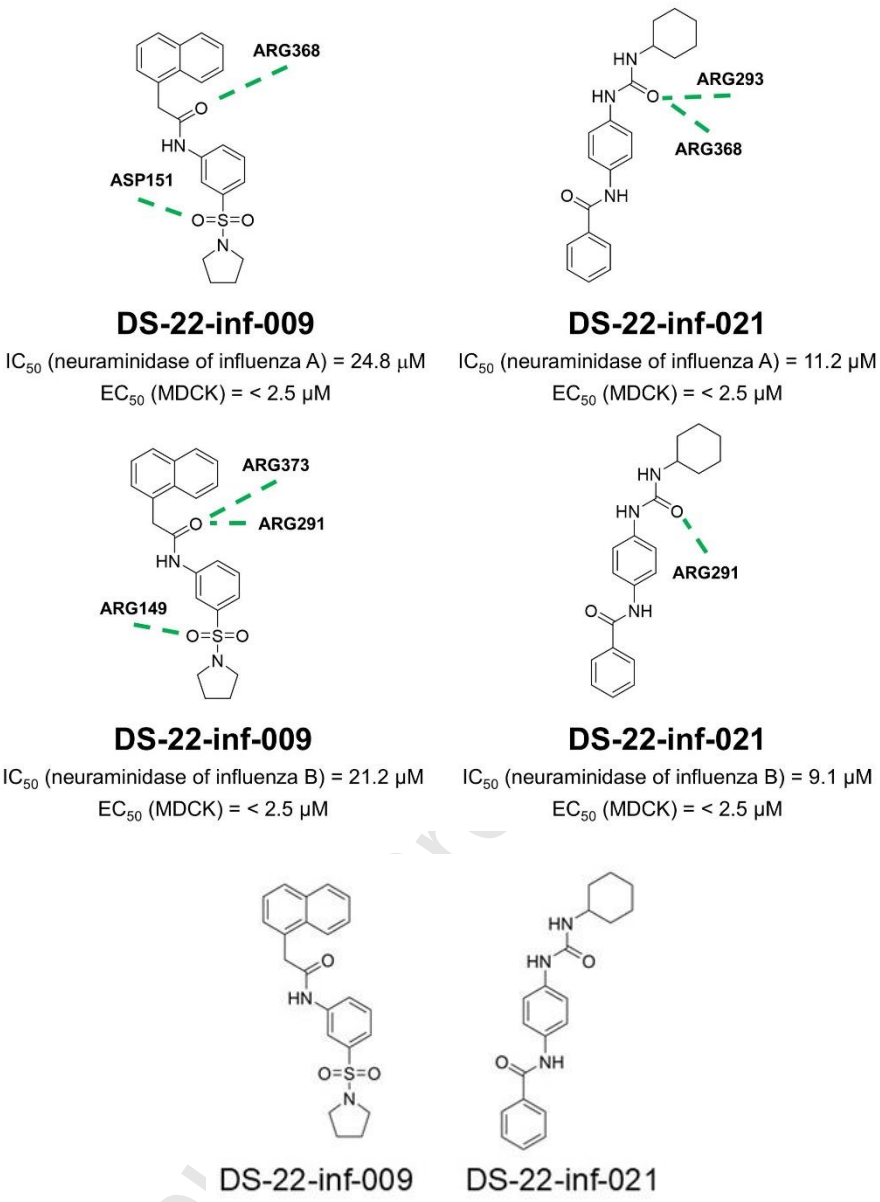


Figure 29. Identification of inhibitors for influenza A and B virus inhibitors using the MDP model. Green dotted lines represent hydrogen bonds.

The aim of the study by Nguyen *et al.*[46] was to identify neuraminidase inhibitors for the inhibition of influenza A and B viruses. They trained different regression models on a dataset of compounds from BindingDB with experimentally determined K_i values for neuraminidase, which were converted into binding free energies. The GCNN model showed the best predictive performance and was, therefore, used to predict the binding free energy of compounds from the CHEMBL dataset. The molecular descriptors calculated with RDKit were used as input features. Subsequently, 11 randomly selected compounds from 400 hit compounds with the lowest predicted binding free energy for neuraminidase were further analyzed for their neuraminidase activity using fluorescence-based inhibition assays. Two compounds showed potential inhibition of neuraminidase, although the exact percentage of inhibition was not reported. However, the exact identity and structure of these two compounds remains unclear. The binding interactions of representative compounds were further investigated by molecular docking and molecular dynamics simulations.

3.3. Biological relevance and implication of the findings

The biological relevance of the ML-driven antiviral discoveries highlighted in this manuscript lies in their ability to target essential viral proteins and host-virus interactions and provide rapid results in the identification of compounds with high antiviral activity that inhibit key viral enzymes such as SARS-CoV-2 M^{pro} and PL^{pro} as well as host factors such as CTSL and ACE2. The use of diverse ML algorithms, from conventional ML methods to deep learning, enables the identification of a broader range of bioactive compounds and mechanisms, improving both prediction accuracy and chemical innovation. These results show that ML has the potential to uncover scaffold diversity beyond traditional medicinal chemistry as well as new mechanisms of action. However, reliance on computational prediction comes with limitations, including dependence on datasets, which are key to successful ML predictions. Future directions should focus on combining ML with experimental testing to augment training datasets with high-quality biological information and develop models that are easier to understand and can facilitate the development of antiviral agents. Ultimately, ML is not a replacement for empirical validation, but a powerful tool that, combined with biological understanding, can accelerate and improve antiviral drug discovery as it was demonstrated in case of compound **ISM3312**, which has entered clinical trials.

4. Limitations and Challenges of ML in Antiviral Drug Design

ML has proven to be a very helpful and successful tool in the search and design of antiviral drugs. It enables the analysis of large datasets and the identification of patterns that might be overlooked by human researchers. However, its application in this field is far from straightforward. One of the most significant challenges is the quality of the data. ML models are highly dependent on the datasets on which they are trained and are therefore highly dependent on the accuracy and completeness of the data. If the data is inconsistent, not cleaned or standardized, not properly validated, etc., the models can produce misleading results. Unlike humans, ML often cannot detect errors in the inputs, which means that errors can influence the final decisions. The issue of data quality is particularly problematic in antiviral drug discovery and suffers from a lack of high-quality data as biological assays can vary in methodology and accuracy. Another challenge is the limited availability of large datasets, which are often not accessible for antiviral agents. This makes it difficult to create accurate models. Furthermore, ML models cannot provide meaningful insights beyond their training data, i.e. they cannot make predictions outside their applicability domain. Another problem is bias, because if the training data favors certain groups or overlooks others, the results can be unrealistic. Additionally, modeling large datasets is computationally intensive and requires significant resources. Therefore, ML should be viewed as an aid in the development of antiviral drugs and not as the ultimate solution or replacement for traditional scientific methods. While it can accelerate discovery and provide valuable insights, its predictions must be thoroughly validated to confirm actual antiviral activity.

Conclusions and Future Perspectives

The application of ML algorithms in the discovery and development of novel antiviral agents has witnessed a significant growth in recent years, particularly in response to the COVID-19 pandemic. This global health crisis revealed the urgent need for accelerated strategies in antiviral drug development and exposed the limitations of traditional discovery pipelines. In this context, ML has emerged as a transformative methodology, offering the ability to streamline and enhance various stages of the antiviral drug discovery process. Building on its success across multiple areas of drug development, ML now plays an increasingly important role in antiviral research. ML algorithms can expedite early med-chem stages by rapidly identifying promising antiviral compounds, uncovering novel therapeutic targets, predicting key pharmacokinetic and toxicological properties, and facilitating efficient ultra

high-throughput virtual screening campaigns (uHTVS) of large multi-billion molecular libraries. Recent advances in deep learning architectures, such as convolutional and recurrent neural networks, have further improved the predictive accuracy by capturing complex biological and chemical relationships, including those found in viral genomics and protein-ligand interactions. In addition, ML supports the prediction of resistance-associated mutations and contributes to the identification of potential broad-spectrum antivirals by anaysing conserved motifs on corresponding viral targets across multiple virus variants and families.

Based on our careful review of the scientific literature published from 2022 to the present day, we are confident this paper can inspire the drug design community to examine med-chem approaches that include one or multiple ML-steps. Moreover, we highlight recent studies in which ML was successfully applied to identify biologically active antiviral hits validated through experimental assays. Collectively, the body of the examined work points to the growing potential of ML-assisted antiviral drug discovery and of its ability to reshape how antiviral drug design is performed in the future.

Acknowledgements

Financial resources through the Slovenian Research and Innovation Agency (ARIS) project and program grants P2-0438, I0-E015, P1-0403, Z1-50021, L2-4430, J1-4414, J3-4497, J7-4638, J3-4498, J1-4398, J4-4633, J7-50043, J1-50034, GC-0001, J1-60001, J2-60044, and L7-60161.

CRediT authorship contribution statement

Anja Kolarič: Writing – original draft. **Marko Jukič:** Writing – review and editing. **Urban Bren:** Writing – review and editing.

Data availability

No data was used for the research described in the article.

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.

5. References

1. Lü Z, Dai X, Xu J, Liu Z, Guo Y, Gao Z, Meng F. Medicinal chemistry strategies toward broad-spectrum antiviral agents to prevent next pandemics. *Eur J Med Chem.* 2024;271:116442. https://doi.org/10.1016/j.ejmech.2024.116442.
2. Pardi N, Weissman D. Development of vaccines and antivirals for combating viral pandemics. *Nat Biomed Eng.* 2020;4:1128-1133. https://doi.org/10.1038/s41551-020-00658-w.
3. Kausar S, Said Khan F, Ishaq Mujeeb Ur Rehman M, Akram M, Riaz M, Rasool G, Hamid Khan A, Saleem I, Shamim S, Malik A. A review: Mechanism of action of antiviral drugs. *Int J Immunopathol Pharmacol.* 2021;35:20587384211002621. https://doi.org/10.1177/20587384211002621.

4. Dara S, Dhamercherla S, Jadav SS, Babu CHM, Ahsan MJ. Machine learning in drug discovery: a review. *Artif Intell Rev.* 2022;55:1947-1999. https://doi.org/10.1007/s10462-021-10058-4.
5. Han R, Yoon H, Kim G, Lee H, Lee Y. Revolutionizing medicinal chemistry: the application of artificial intelligence (AI) in early drug discovery. *Pharmaceuticals.* 2023;16:1259. https://doi.org/10.3390/ph16091259.
6. Qureshi R, Irfan M, Gondal TM, Khan S, Wu J, Hadi MU, Heymach J, Le X, Yan H, Alam T. AI in drug discovery and its clinical relevance. *Heliyon.* 2023;9:e17575. https://doi.org/10.1016/j.heliyon.2023.e17575.
7. Williams AH, Zhan CG. Staying ahead of the game: how SARS-CoV-2 has accelerated the application of machine learning in pandemic management. *BioDrugs.* 2023;37:649-674. https://doi.org/10.1007/s40259-023-00611-8.
8. Kandeel M, Al-Mubarak AIA, Alhojaily S, Afzal S, Ismail I, Al-Rasheed M. Revolutionizing antiviral drug discovery: the emerging role of artificial intelligence. *Int J Pharmacol.* 2024;20:536-546. https://doi.org/10.3923/ijp.2024.536.546.
9. Prasad K, Kumar V. Artificial intelligence-driven drug repurposing and structural biology for SARS-CoV-2. *Curr Res Pharmacol Drug Discov.* 2021;2:100042. https://doi.org/10.1016/j.crphar.2021.100042.
10. Lee CY, Chen YPP. New insights into drug repurposing for COVID-19 using deep learning. *IEEE Trans Neural Netw Learn Syst.* 2021;32:4770-4780. https://doi.org/10.1109/TNNLS.2021.3111745.
11. Lv H, Shi L, Berkenpas JW, Dao FY, Zulfiqar H, Ding H, Zhang Y, Yang L, Cao R. Application of artificial intelligence and machine learning for COVID-19 drug discovery and vaccine design. *Brief Bioinform.* 2021;22:bbab320. https://doi.org/10.1093/bib/bbab320.
12. Pires C. A Systematic Review on the Contribution of Artificial Intelligence in the Development of Medicines for COVID-2019. *J Pers Med.* 2021;11:926. https://doi.org/10.3390/jpm11090926.
13. Floresta G, Zagni C, Gentile D, Patamia V, Rescifina A. Artificial intelligence technologies for COVID-19 de novo drug design. *Int J Mol Sci.* 2022;23:3261. https://doi.org/10.3390/ijms23063261.
14. Elkashlan M, Ahmad RM, Hajar M, Al Jasmi F, Corchado JM, Nasarudin NA, Mohamad MS. A review of SARS-CoV-2 drug repurposing: databases and machine learning models. *Front Pharmacol.* 2023;14:1182465. https://doi.org/10.3389/fphar.2023.1182465.
15. Murugan NA, Priya GR, Sastry GN, Markidis S. Artificial intelligence in virtual screening: Models versus experiments. *Drug Discov Today.* 2022;27:1913-1923. https://doi.org/10.1016/j.drudis.2022.05.013.
16. Chadaga K, Prabhu S, Vivekananda BK, Niranjana S, Umakanth S. Battling COVID-19 using machine learning: A review. *Cogent Eng.* 2021;8:1958666. https://doi.org/10.1080/23311916.2021.1958666.
17. Obaido G, Mienye ID, Egbelowo OF, Emmanuel ID, Ogunleye A, Ogbuokiri B, Mienye P, Aruleba K. Supervised machine learning in drug discovery and development: Algorithms, applications, challenges, and prospects. *Mach Learn Appl.* 2024;17:100576. https://doi.org/10.1016/j.mlwa.2024.100576.
18. Suresh M, Tolasa FT, Bonyah E. QSPR/QSAR study of antiviral drugs modeled as multigraphs by using TI's and MLR method to treat COVID-19 disease. *Sci Rep*. 2024;14:1-14. https://doi.org/10.1038/s41598-024-63007-w.
19. Manonmani R, Selvarani P. Multilinear regression analysis of antiviral medications with topological indices. *J Mol Graph Model*. 2025;109104. https://doi.org/10.1016/j.jmgm.2025.109104.

20. Zaman S, Rasheed S, Alamer A. A quadratic regression model to quantify certain latest corona treatment drug molecules based on coindices of M-polynomial. *J Supercomput*. 2024;80:26805-26830. https://doi.org/10.1007/s11227-024-06434-w.
21. Zhou ZH. *Machine learning*. Springer nature; 2021.
22. Lavecchia A. Machine-learning approaches in drug discovery: methods and applications. *Drug Discov Today.* 2015;20:318-331. https://doi.org/10.1016/j.drudis.2014.10.012.
23. Dong J, Varbanov M, Philippot S, Vreken F, Zeng W, Blay V. Ligand-based discovery of coronavirus main protease inhibitors using MACAW molecular embeddings. *J Enzyme Inhib Med Chem.* 2023;38:24-35. https://doi.org/10.1080/14756366.2022.2132486.
24. Xu T, Xu M, Zhu W, Chen CZ, Zhang Q, Zheng W, Huang R. Efficient identification of anti-SARS-CoV-2 compounds using chemical structure-and biological activity-based modeling. *J Med Chem.* 2022;65:4590-4599. https://doi.org/10.1021/acs.jmedchem.1c01372.
25. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD Int Conf on Knowl Discov Data Min. 2016:785–794. https://doi.org/10.1145/2939672.2939785.
26. Saar KL, McCorkindale W, Fearon D, Boby M, Barr H, Ben-Shmuel A, The COVID Moonshot Consortium, London N, von Delft F, Chodera JD, Lee AA. Turning high-throughput structural biology into predictive inhibitor design. *PNAS.* 2023;120:e2214168120. https://doi.org/10.1073/pnas.2214168120.
27. Pal S, Nance KD, Joshi DR, Kales SC, Ye L, Hu X, Shamim K, Zakharov AV. Applications of Machine Learning Approaches for the Discovery of SARS-CoV-2 PLpro Inhibitors. *J Chem Inf Model.* 2025;65:1338–1356. https://doi.org/10.1021/acs.jcim.4c02126.
28. Giuliani A. The application of principal component analysis to drug discovery and biomedical data. *Drug Discov Today*. 2017;22:1069-1076. https://doi.org/10.1016/j.drudis.2017.01.005.
29. Ali M, Park IH, Kim J, Kim G, Oh J, You JS, Kim J, Shin JS, Yoon SS. How deep learning in antiviral molecular profiling identified anti-SARS-CoV-2 inhibitors. *Biomedicines.* 2023;11:3134. https://doi.org/10.3390/biomedicines11123134.
30. Alkan C, O'Brien T, Kenyon V, Ikegami T. Computer-Selected Antiviral Compounds: Assessing In vitro Efficacies against Rift Valley Fever Virus. *Viruses.* 2024;16:88. https://doi.org/10.3390/v16010088.
31. Tong X, Liu X, Tan X, Li X, Jiang J, Xiong Z, Xu T, Jiang H, Qiao N, Zheng M. Generative models for de novo drug design. *J Med Chem.* 2021;64:14011-14027. https://doi.org/10.1021/acs.jmedchem.1c00927.
32. Huang L, Xu T, Yu Y, Zhao P, Chen X, Han J, Xie Z, Li H, Zhong W, Wong K-C, Zhang H. A dual diffusion model enables 3D molecule generation and lead optimization based on target pockets. *Nat Commun.* 2024;*15*:2657. https://doi.org/10.1038/s41467-024-46569-1.
33. Mao J, Wang J, Zeb A, Cho KH, Jin H, Kim J, Lee O, Wang Y, No KT. Transformer-based molecular generative model for antiviral drug design. J Chem Inf Model. 2023;64:2733-2745. https://doi.org/10.1021/acs.jcim.3c00536.
34. Sun J, Sun D, Yang Q, Wang D, Peng J,Guo H, Ding X, Chen Z, Yuan B, Ivanenkov YA,Yuan J, Zagribelnyy BA, He Y, Su J, Wang L,Tang J, Li Z, Li R, Li T, Hu X, Liang X, Zhu A,Wei P, Fan Y, Liu S, Zheng J, Guan X, Aliper A,Yang M, Bezrukov DS, Xie Z, Terentiev VA,Peng G, Polykovskiy DA, Malyshev AS,Malkov MN, Zhu Q, Aspuru-Guzik A, Ding X,Cai X, Zhang M, Zhao J, Zhong N,Ren F, Chen X, Zhavoronkov A, Zhao J. A novel, covalent broad-spectrum inhibitor targeting human coronavirus Mpro. Nat Commun. 2025;16:4546. https://doi.org/10.1038/s41467-025-59870-4

35. Ang D, Kendall R, Atamian HS. Virtual and *in vitro* screening of natural products identifies indole and benzene derivatives as inhibitors of SARS-CoV-2 Main Protease (Mpro). *Biology*. 2023;12:519. https://doi.org/10.3390/biology12040519.
36. Bess A, Berglind F, Mukhopadhyay S, Brylinski M, Alvin C, Fattah F, Wasan KM. Identification of oral therapeutics using an AI platform against the virus responsible for COVID-19, SARS-CoV-2. *Front Pharmacol.* 2023;14:1297924. https://doi.org/10.3389/fphar.2023.1297924.
37. Jing Y, Bian Y, Hu Z, Wang L, Xie XQS. Deep learning for drug design: an artificial intelligence paradigm for drug discovery in the big data era. *AAPS*. 2018;20:1-10. https://doi.org/10.1208/s12248-018-0210-0.
38. Komatsu H, Tanaka T, Ye Z, Ikeda K, Matsuzaki T, Yasugi M, Hosoda M. Identification of SARS-CoV-2 main protease inhibitors from FDA-approved drugs by artificial intelligence-supported activity prediction system. *J Biomol Struct Dyn.* 2023;41:1767-1775. https://doi.org/10.1080/07391102.2021.2024260.
39. Anwaar MU, Adnan F, Abro A, Khan RA, Rehman AU, Osama M, Rainville C, Kumar S, Sterner DE, Javed S, Jamal SB, Baig A, Shabbir MR, Ahsan W, Butt TR, Assir MZ. Combined deep learning and molecular docking simulations approach identifies potentially effective FDA approved drugs for repurposing against SARS-CoV-2. *Comput Biol Med.* 2022;141:105049. https://doi.org/10.1016/j.compbiomed.2021.105049.
40. Gkekas I, Katsamakas S, Mylonas S, Fotopoulou T, Magoulas GE, Tenchiu AC, Dimitriou M, Axenopoulos A, Rossopoulou N, Kostova S, Wanker EE, Katsila T, Papahatjis D, Gorgoulis VG, Koufaki M, Karakasiliotis I, Calogeropoulou T, Daras P, Petrakis S. AI Promoted Virtual Screening, Structure-Based Hit Optimization, and Synthesis of Novel COVID-19 S-RBD Domain Inhibitors. *J Chem Inf Model.* 2024;64:8562-8585. https://doi.org/10.1021/acs.jcim.4c01110.
41. Xiang Z, Fan S, Ren J, Ruan T, Chen Y, Zhang Y, Wang Y, Yu Z, Wang C, Sun X, Hao XJ, Chen DZ. Utilizing artificial intelligence for precision exploration of N protein targeting phenanthridine sars-cov-2 inhibitors: A novel approach. *Eur J Med Chem.* 2024;279:116885. https://doi.org/10.1016/j.ejmech.2024.116885.
42. Kovalishyn V, Severin O, Kachaeva M, Kobzar O, Keith KA, Harden EA, Hartline CB, James SH, Vovk A, Brovarets V. In Silico Design and Experimental Validation of Novel Oxazole Derivatives Against Varicella zoster virus. *Mol Biotechnol.* 2024;66:707-717. https://doi.org/10.1007/s12033-023-00670-w.
43. Saramago LC, Santana MV, Gomes BF, Dantas RF, Senger MR, Oliveira Borges PH, Ferreira VNdS, dos Santos Rosa A, Tucci AR, Dias Miranda M, Lukacik P, Strain-Damerell C, Owen CD, Walsh MA, Ferreira SB, Silva-Junior FP. AI-driven discovery of SARS-CoV-2 main protease fragment-like inhibitors with antiviral activity in vitro. *J Chem Inf Model.* 2023;63:2866-2880. https://doi.org/10.1021/acs.jcim.3c00409.
44. Varikoti RA, Schultz KJ, Kombala CJ, Kruel A, Brandvold KR, Zhou M, Kumar N. Integrated data-driven and experimental approaches to accelerate lead optimization targeting SARS-CoV-2 main protease. *J Comput Aided Mol.* 2023;37:339-355. https://doi.org/10.1007/s10822-023-00509-1.
45. van der Horst D, Carter-Timofte ME, Danneels A, da Costa LS, Kurmasheva N, Thielke AL, Hansen AL, Chorošajev V, Holm CK, Belouzard S, de Weber I, Beny C, Olagnier D. Large-scale deep learning identifies the antiviral potential of PKI-179 and MTI-31 against coronaviruses. *Antiviral Res.* 2024;231:106012. https://doi.org/10.1016/j.antiviral.2024.106012.
46. Nguyen TH, Anh Pham NQ, Thai QM, Vu VV, Ngo ST, Horng JT. Discovering Influenza Virus Neuraminidase Inhibitors via Computational and Experimental Studies. *ACS Omega.* 2024;9:48505-48511. https://doi.org/10.1021/acsomega.4c07194.

47. Fabiyi SD. A review of unsupervised artificial neural networks with applications. *Int J Comput Appl.* 2019;181:22-26. https://doi.org/10.5120/ijca2019918425.
48. Dike HU, Zhou Y, Deveerasetty KK, Wu Q. Unsupervised learning based on artificial neural network: a review. In: 2018 IEEE International Conference on Cyborg and Bionic Systems (CBS). IEEE; 2018:322-327. https://doi.org/10.1109/CBS.2018.8612259.
49. Iype E, Pillai U J, Kumar I, Gaastra-Nedea SV, Subramanian R, Saha RN, Dutta M. In silico and *in vitro* assays reveal potential inhibitors against 3CL pro main protease of SARS-CoV-2. *J Biomol Struct Dyn.* 2022;40:12800-12811. https://doi.org/10.1080/07391102.2021.1977181.
50. Joshi RP, Schultz KJ, Wilson JW, Kruel A, Varikoti RA, Kombala CJ, Kneller DW, Galanie S, Phillips G, Zhang Q, Coates L, Parvathareddy J, Surendranathan S, Kong Y, Clyde A, Ramanathan A, Jonsson CB, Brandvold KR, Zhou M, Head MS, Kovalevsky A, Kumar N. Ai-accelerated design of targeted covalent inhibitors for SARS-CoV-2. *J Chem Inf Model.* 2023;63:1438-1453. https://doi.org/10.1021/acs.jcim.2c01377.
51. Juárez-Mercado KE, Gómez-Hernández MA, Salinas-Trujano J, Córdova-Bahena L, Espitia C, Pérez-Tapia SM, Medina-Franco JL, Velasco-Velázquez MA. Identification of SARS-CoV-2 Main Protease Inhibitors Using Chemical Similarity Analysis Combined with Machine Learning. *Pharmaceuticals*. 2024;17:240. https://doi.org/10.3390/ph17020240.
52. Jo S, Signorile L, Kim S, Kim MS, Huertas O, Insa R, Reig N, Shin DH. A study of drug repurposing to identify SARS-CoV-2 main protease (3CLpro) inhibitors. *Int J Mol Sci.* 2022;23:6468. https://doi.org/10.3390/ijms23126468.
53. Yang WL, Li Q, Sun J, Tan SH, Tang YH, Zhao MM, Li YY, Cao X, Zhao JC, Yang JK. Potential drug discovery for COVID-19 treatment targeting Cathepsin L using a deep learning-based strategy. *Comput Struct Biotechnol.* 2022;20:2442-2454. https://doi.org/10.1016/j.csbj.2022.05.023.
54. Huang K, Fu T, Glass LM, Zitnik M, Xiao C, Sun J. DeepPurpose: a deep learning library for drug–target interaction prediction. *Bioinformatics.* 2020;36:5545-5547. https://doi.org/10.1093/bioinformatics/btaa1005.
55. Joshi RP, Gebauer NWA, Bontha M, Khazaieli M, James RM, Brown JB, Kumar N. 3D-scaffold: a deep learning framework to generate 3D coordinates of drug-like molecules with desired scaffolds. *J Phys Chem B.* 2021;125:12166-12176. https://doi.org/10.1021/acs.jpcb.1c06437.
56. Santana MVS, Silva-Jr FP. De novo design and bioactivity prediction of SARS-CoV-2 main protease inhibitors using recurrent neural network-based transfer learning. *BMC Chem.* 2021;15:8. https://doi.org/10.1186/s13065-021-00737-2.
57. Zhang H, Liang B, Sang X, An J, Huang Z. Discovery of potential inhibitors of SARS-CoV-2 main protease by a transfer learning method. *Viruses.* 2023;15:891. https://doi.org/10.3390/v15040891.
58. Chithrananda S, Grand G, Ramsundar B. ChemBERTa: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv.* 2020;2010.09885. https://doi.org/10.48550/arXiv.2010.09885.
59. Teshima K, Tanaka T, Ye Z, Ikeda K, Matsuzaki T, Shiroma T, Muroya A, Hosoda M, Yasugi M, Komatsu H. Antiviral activity of curcumin and its analogs selected by an artificial intelligence-supported activity prediction system in SARS-CoV-2-infected VeroE6 cells. *Nat Prod Res.* 2024;38:867-872. https://doi.org/10.1080/14786419.2023.2194647.
60. Web-based software for machine learning algorithms. Accessed 10 June 2025. https://www.collaborationspharma.com/assay-central.
61. Alves VM, Bobrowski T, Melo-Filho CC, Korn D, Auerbach S, Schmitt C, Muratov EN, Tropsha A. QSAR Modeling of SARS-CoV M pro Inhibitors Identifies Sufugolix,

Cenicriviroc, Proglumetacin, and Other Drugs as Candidates for Repurposing against SARS-CoV-2. *Mol Inform.* 2021;40:2000113. https://doi.org/10.1002/minf.202000113.
62. Goyal B, Goyal D. Targeting the Dimerization of the Main Protease of Coronaviruses: A Potential Broad-Spectrum Therapeutic Strategy. ACS Comb Sci. 2020;22:297–305. https://doi.org/10.1021/acscombsci.0c00058.
63. Garnsey MR, Robinson MC, Nguyen LT, Cardin R, Tillotson J, Mashalidis E, Yu A, Aschenbrenner L, Balesano A, Behzadi A, Boras B, Chang JS, Eng H, Ephron A, Foley T, Ford KK, Frick JM, Gibson S, Hao L, Hurst B, Kalgutkar AS, Korczynska M, Lengyel-Zhand Z, Gao L, Meredith HR, Patel NC, Polivkova J, Rai D, Rose CR, Rothan H, Sakata SK, Vargo TR, Qi W, Wu H, Liu Y, Yurgelonis I, Zhang J, Zhu Y, Zhang L, Lee A.A. Discovery of SARS-CoV-2 papain-like protease (PLpro) inhibitors with efficacy in a murine infection model. *Sci Adv.* 2024;10:eado4288. https://doi.org/10.1126/sciadv.ado4288.
64. Fu Z, Huang B, Tang J, Liu S, Liu M, Ye Y, Liu Z, Xiong Y, Zhu W, Cao D, Li J, Niu X, Zhou H, Zhao YJ, Zhang G, Huang H. The complex structure of GRL0617 and SARS-CoV-2 PLpro reveals a hot spot for antiviral drug discovery. *Nat Commun.* 2021;12:488. https://doi.org/10.1038/s41467-020-20718-8.
65. Kaitoh K, Yamanishi Y. Scaffold-retained structure generator to exhaustively create molecules in an arbitrary chemical space. *J Chem Inf Model.* 2022;62:2212-2225. https://doi.org/10.1021/acs.jcim.1c01130.
66. Green H, Koes DR, Durrant JD. DeepFrag: a deep convolutional neural network for fragment-based lead optimization. *Chem Sci.* 2021;12:8036–8047. https://doi.org/10.1039/D1SC00163A.
67. Kovalishyn V, Severin O, Kachaeva M, Semenyuta I, Keith KA, Harden EA, Hartline CB, James SH, Metelytsia L, Brovarets V. Design and experimental validation of the oxazole and thiazole derivatives as potential antivirals against of human cytomegalovirus. *SAR QSAR Environ Res.* 2023;34:523-541. https://doi.org/10.1080/1062936X.2023.2232992.
68. Wallach I, Dzamba M, Heifets A. AtomNet: A deep convolutional neural network for bioactivity prediction in structure-based drug discovery. *arXiv preprint arXiv*. 2015;1510.02855. https://doi.org/10.48550/arXiv.1510.02855.
69. Izmailyan R, Matevosyan M, Khachatryan H, Shavina A, Gevorgyan S, Ghazaryan A, Tirosyan I, Gabrielyan Y, Ayvazyan M, Martirosyan B, Harutyunyan V, Zakaryan H. Discovery of new antiviral agents through artificial intelligence: In vitro and in vivo results. *Antiviral Res.* 2024;222:105818. https://doi.org/10.1016/j.antiviral.2024.105818.

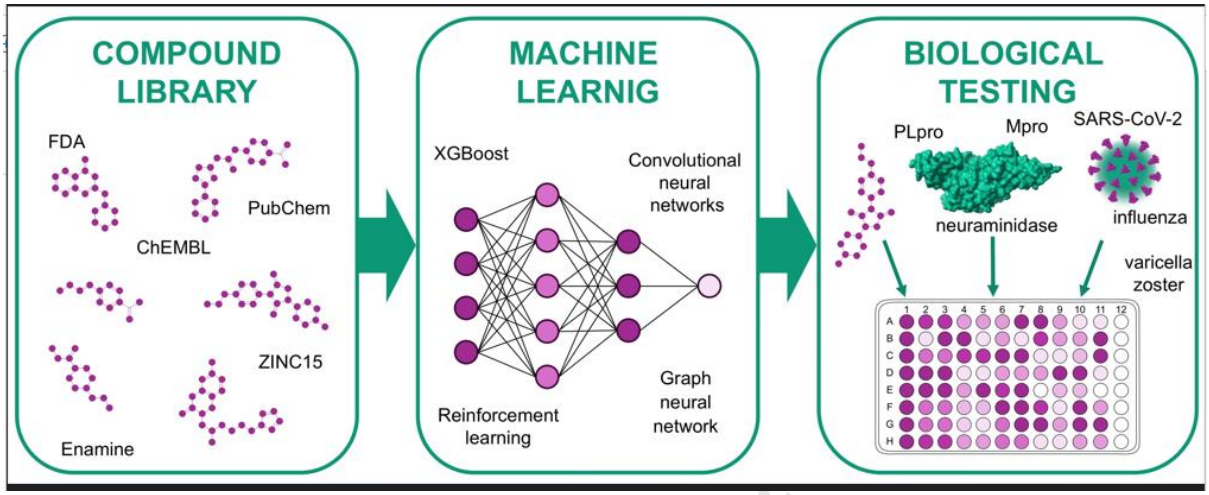
Journal Pre-proof

Declaration of interests

□ 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.

□ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Journal Pre-proof



Graphical abstract

Journal Pre-proof