Bioactivity Prediction by a Heterogeneous Siamese Neural Network (SNN)

Pham Hai Nam, Nguyen Son, Doan Huu Thanh, Le Mau Minh Phuc*

* University of Science and Technology of Hanoi

Abstract—**The Bioactivity Determination while taking medicine causes challenges leading to drug failure. To address this risk and enhance compound activities, Bioactivity Classes during lead optimization are necessary based on the foundation of Structure-activity Relationships as the connection between the Chemical Compounds and the Bioactivity. This study has two major objectives: (i) to evaluate the BioAct-Het (A Siamese Neural Network) model by taking 7 predicted and actual side-effects drugs according to evaluation metrics to ensure the trustworthiness and (ii) to find the Bioactivity Classes of 24 marketed drugs, including both COVID-19 and non-COVID-19 substances, through BioAct-Het. Since the confirmation of trustworthiness of BioAct-Het results in greater than 75%, the model is believed to be an accurate model. It is moreover used to examine and analyze the results of 25 drug inputs. Therefore, 5 drugs are both concluded as at least side effects and requiring careful use, leading to some particular side effects including immune system disorders, vascular disorders, nervous system disorders, metabolism and nutrition disorders, and others. Last but not least, with these concluded drugs, Paxlovid (Nirmatrelvir) and Interferon-beta 1a (IFN-β1a) are recommended for use in COVID-19 treatment combined with the prescription of doctors. The data and code for this study are available on https://github.com/Xernnn/Bioact-Pred-Het-SNN. However, the datasets were obtained from publicly available sources.**

Keywords: BioAct-Het, Bio-Prof, bioactivity, drug side-effects, Siamese neural networks, COVID-19 treatment, structure-activity relationships.

I. Introduction

A. Context of Computational Methods in Drug Discovery

The drug-likeness of a compound must be identified and optimized with a demanding and lengthy process [12]. Generally, it takes over a decade and costs billions to bring a new chemical compound to market as a drug. Predicting a compound's bioactivities, which relies on the effects of living organisms including both positive and negative outcomes like side effects and toxicity, is crucial in drug discovery. During lead optimization, the objective is to improve the structure of leads while maintaining the therapeutic properties to enhance bioactivity Hughes et al. [3].

However, experimental evaluation related to bioactivity is challenging due to the number of potential analogs and the high screening costs Reymond et al. [8]. Computational techniques, to explore the structure-activity relationship (SAR) with demonstrated compounds with similar structures often share bioactivities, are a viable alternative to expensive screening methods at laboratories Nantasenamat et al. [7].

Application of computation methods to predict the bioactivity of novel compounds encounters two challenges, including compound representation and data scarcity. Compound representation, firstly, involves selecting an appropriate structure representation. SMILES, fingerprints, and graphs with traditional algorithms have been used to represent a compound in the bioactivity class Weininger [11]. Moreover, the lack of data scarcity to train a reliable model necessitates the creation of varied datasets consisting of chemical substances with a range of bioactivities. Ratios that are out of balance between positive and negative data can also impact the performance of prediction. Torres et al. introduced an additional one-shot classification approach using a Siamese neural network architecture, which employs a Convolutional Neural Network (CNN) to identify novel compound features, particularly for less-represented classes. This model, with the use of the Tox21 database, represents compounds using the one-hot encoding of canonical SMILES structures and make into groups based on chemical structures for training Weininger [11].

The bioactivity class prediction (BCP) problem can be addressed by many research methods that rely on the central hypothesis of SAR studies. These approaches aim to identify patterns in drug chemical structures associated with specific bioactivity classes and can be trained on pairs of compounds sharing the same bioactivity class to learn similarities and pairs with bioactivities to obtain the differences, and used to predict whether new chemical compounds belong to which particular bioactivity class Vella and Ebejer [10]

The evaluation of previous studies demonstrates that the relationship between chemical compounds and bioactivities is extremely complex due to a significant relationship challenge in the BCP problem. To address these difficulties, this study implements a method called BioAct-Het, which attempts to determine the likelihood of association between a compound and a bioactivity class. BioAct-Het uses a heterogeneous SNN to map chemical compounds and bioactivity classes into a unified latent space, which is capable of representing bioactivity classes based on related substances. The performance of BioAct-Het is evaluated using both supervised learning and meta-learning approaches on three databases: SIDER, Tox21, MUV, and the self-constructed dataset Heyrati et al. [2].

B. Related Work

BioAct-Het is a heterogeneous SNN (Siamese Neural Network) that solves the bioactivity class prediction problem, meaning whether the compound causes the bioactivity or not. The input of the model consists of two branches, which are embedding compounds and bioactivity classes. Since this is a binary classification, the output will be 1 if the compound causes the bioactivity class, otherwise, the prediction will be 0.

1) Chemical representation as feature extractors - Graph Convolutional Networks (GCN)

In the experiment protocol, Graph Convolutional Networks (GCN) are used to represent the chemical structure of the compounds, which is one of the two inputs for the BioAct-Het architecture. Initially, the SMILES structure of these compounds is transformed into graph representations, where each node corresponds to an atom, and the edges represent the bonds between these atoms Li et al. [5]. The graphs are then passed into the GCN to represent the chemical structures of the compounds as vectors.

The GCN architecture contains two operations, the first one involves calculating the weighted sum of the representations of nodes in a graph. The sum is then processed through a linear layer following a sigmoid function. The second operation is to use an element-wise maximum to the representations of nodes in a graph, which is then passed through a Multilayer perceptron for the final vector representations of a chemical compound. For this process, two pre-trained models are carried out: GCN-Canonical GCN-AttentiveFP Heyrati et al. [2].

GCN-Canonical gets the information from the nodes and the following neighbors by considering how many connections or degrees each node has. The method ensures that every node is calculated equally based on the number of connections to the other nodes Kipf and Welling [4]. On the other hand, GCN-AttentiveFP is used to identify the important connections between any nodes, while still keeping the layout of the molecule. The model is designed to capture critical relationships in the molecular structures Xiong et al. [14].

2) *Bio-Prof as a bioactivity class model*

The Bio-Prof model is used to represent the bioactivity classes based on the substructures of the chemical compounds, which is the other input of BioAct-Het. To represent the bioactivity classes, the Morgan fingerprint, a binary string representation encoding function, is used for this purpose by examining which molecule structures in the compounds that are known to be active or inactive to the bioactivity classes. Each bit in the fingerprint represents the presence or absence of a particular substructure within a compound for each bioactivity class Morgan [6].

3) *BioAct-Het as a bioactivity prediction model*

As previously mentioned, BioAct-Het is a heterogeneous SNN (Siamese Neural Network), which is different from homogeneous SNNs, in which the inputs are dual-branch networks with tied weights. Heterogeneous SNN consists of distinctive networks in two various branches and a different learning component. The branches transform the initial representation of the chemical compounds and bioactivity classes into a unified latent space to share properties and become comparable. To bring the similarity of bioactivity classes and chemical compounds close together in the latent space, BioAct-Het calculates the likelihood of association through the distances Bemis and Murcko [1].

First of all, by taking the results from the GCN models and the Bio-Prof model, BioAct-Het passes the results to the dense layers containing two nonlinear functions, ReLU as the activation functions and kl-divergence loss functions. The aim is to transform chemical compound embedding vectors and bioactivity class embedding vectors into having the same dimension, creating a unified latent space for chemical compounds and bioactivities to produce a distribution.

In the second step, BioAct-Het predicts the association between the compound and bioactivity class based on the distance in the latent space. The distance vector is then passed through a dense layer containing a nonlinear function, ReLU as the activation function and sigmoid function with a threshold of 0.5. Since the output layer contains one node, the result will be 1 if the compound causes the bioactivity class, otherwise, the prediction will be 0, with the loss function being binary cross-entropy function. After that, minimizing the loss function is required to increase the model's effectiveness in predictions Wu et al. [13].

C. *Our Contribution*

The study of marketed drugs and COVID-19 side-effect drugs is relatively time-consuming and difficult due to a large range of biological knowledge. Moreover, even though it is available on GitHub under https://github.com/Xernnn/Bioact-Pred-Het-SNN, the source code needs to be improved and re-implemented to evaluate available side-effect drugs in SIDER and predicted side-effect drugs according to confusion metrics. This study is mostly to understand the application of machine learning in Medicine and case studies of COVID-19, instead of only conventional paths, such as training and testing a type of datasets on the internet, and conclude. We, therefore, spent time studying medical cases and new neural network cases that have an objective to real-life applications with a clear further study collaborated with virtual screening.

In detail, the main contributions of this study:

- To construct an experimental design with the relation of SAR.
- To construct the dataset with the study of which marked drugs treating COVID-19.
- To evaluate the model, which considers a compound-bioactivity class pair as a positive pair and a negative pair, with the complex relationship between compounds and bioactivity classes, to learn a unified latent space via a heterogeneous SNN.
- To visualize the outcome to further study.

II. Methods - Experimental Design

The design of each experiment is to predict whether marketed drugs cause some bioactivity classes. This computational model is used initially to forecast the potential bioactivities of marketed drugs before starting with virtual screening or docking studies.

In Figure 1, potential drugs are chosen with the self-constructed dataset treating COVID-19. Moreover, it goes to pre-trained chemical compounds - GCN-AttentiveFP model (feature extractors) to get chemical representations. These representations are then fed to the BioAct-Het model, linked to each side effect class representation that was gained. The model results in the probability of each drug showing side effect classes in the form of embedding vectors in the range of 0 to 1. If the probability is greater than 0.5, the bioactivity class is represented in chemical compounds and vice versa Shih et al. [9].

A. *Data materials*

There are four different datasets for distinct purposes related to side effects, toxicity, and the risk of overfitting to similar compounds. Almost all datasets are used as pre-trained models for less time-consuming re-training processes, while a manual collection dataset is used for testing the model to further analyze results. Table I describes the ratios for training and testing equivalent to each dataset.

1) *A manual collection dataset*

A manual collection dataset is constructed to evaluate the pre-trained model and analyze the further results. It contains 24 marketed drugs with the SMILES structure used to treat COVID-19. These drugs are used to evaluate the model by selecting 7 selected drugs with the 27 side effects in SIDER compared with the prediction of these 7 drugs of an SNN model. Moreover, potential side effects extracted from results can be used to propose which side effects are typical for dosage use

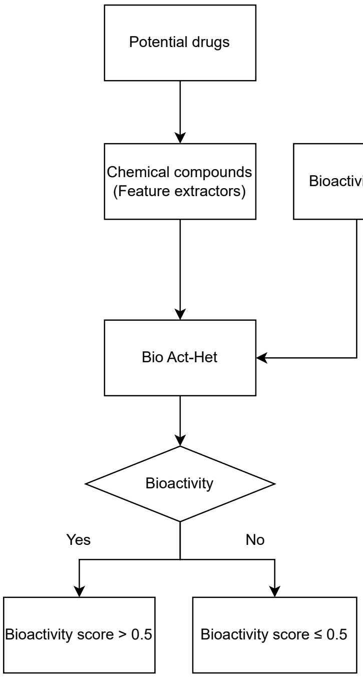


Figure 1. Flowchart

and chemical compound recommendations. Details in Appendix 1.

2) SIDER

The SIDER is a dataset of marketed-drugs side effects. These include 27 classes using Medical Dictionary for Regulatory Activities (MedDRA) classification according to system organ classes. SIDER is a variant in, for example, Hepatobiliary disorders, Metabolism and nutrition disorders, and Product issues. Details in Table I.

3) Tox21

Tox21 is designed to evaluate the toxicity of chemical substances. It consists of 12 toxicological assays following receptors, which forecast the potential risks linked to exposure to these compounds. Details in Table I.

4) MUV

The MUV is indicated as a benchmark for evaluating the performance of virtual screening with 17 challenging tasks and priority of the minimization of the risk of overfitting to similar compounds. Details in Table I.

Table I
SIDER, TOX21, MUV DATASETS INFORMATION

Dataset	Compounds	Bioactivity Classes	(+) Samples	(-) Samples
SIDER	1427	27	21,868	16,661
Tox21	7831	12	5,862	72,084
MUV	93,087	17	489	1,332,593

B. Data representation and pre-trained BioAct-Het Model

As described in related work, GCN-Canonical and GCN-AttentitiveFP are used for feature extractors. Chemical compound representation steps are then gone into Bio-Prof which is based on the profile of chemical structure of related compounds. Moreover, the model called BioAct-Het (a heterogeneous SNN) plays a role in bringing the similar concept of bioactivity classes and chemical compounds in the latent space and inferring the likelihood of association through distances as embedding vectors. The strategies and evaluation approach in pre-trained models are according to Table II.

- The association-based strategy is to determine the relationship between a chemical substance in the data set and a bioactivity class (preferred to use in BioAct-Het).
- The bioactivity class-based strategy aims to uncover chemical compounds that demonstrate new and uncommon bioactivity (preferred to use in Bioactivity class).
- The compound-based strategy is to predict the bioactivity classes of a newly found chemical compound released into the market (preferred to use in Chemical Compounds - feature extractors).

Table II
THE STRATEGIES AND EVALUATION APPROACH IN PRE-TRAINED MODELS

Category	Training	Testing	Purpose
Association-based (Pre-trained)	90%	10%	Predict the unknown association
Bioactivity-based (Pre-trained)	80%	20%	Identify chemical compounds that exhibit a novel type of bioactivity
Compound-Based Strategy (Pre-trained)	90%	10%	Predict the bioactivity classes of a newly found chemical compound

C. Evaluation Metrics

Confusion matrices and other evaluation metrics are used to assess how accurate the model is. These tools are essential for determining the trustworthiness of the model and for further analysis.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Table III
EVALUATION MATRIX EXPLANATION

Prediction	Definition
True Positive (TP)	The number of activated bioactivity classes for a drug predicted correctly by the BioAct-Het.
True Negative (TN)	The number of nonactivated bioactivity classes for a drug predicted correctly by the BioAct-Het.
False Positive (FP)	The number of activated bioactivity classes for a drug predicted wrongly by the BioAct-Het.
False Negative (FN)	The number of nonactivated bioactivity classes for a drug predicted wrongly by the BioAct-Het.

III. RESULTS AND EVALUATION

A. Model Evaluation

To evaluate the performance of the BioAct-Het model, a testing process is carried out. This involves comparing the side effects of 7 actual drugs from the SIDER dataset with those predicted for 7 drugs, among a total of 25 potentially marketed drugs that had ground truth side-effect labels. These results are presented in Table IV. These drugs were:

- **Chloroquine**
- **Famotidine**
- **Guanfacine**

- **Hydroxychloroquine**
- **Oseltamivir**
- **Prednisolone**
- **Ritonavir**

For each drug-side effect pair in the test set, the model made a prediction between 0 and 1 for the likelihood of that side effect occurring. The threshold for classifying predictions is 0.5, which is a positive prediction for the side effects if the result is above the threshold, and below the threshold results in a negative prediction.

We then generated bubble charts and confusion matrices to visualize how well the model's predictions matched each drug's ground truth labels. The bubble charts showed the counts of true positive, false positive, true negative, and false negative predictions for each drug. Overall, the confusion matrices indicated the model accurately predicted both the presence and absence of side effects for most drug-side effect pairs. (Appendix 2)

To quantify the model's performance, we can see that in table IV, the evaluation metrics are being used, such as accuracy, precision, recall, and F1 score calculation based on the true and false positives and negatives. The accuracy was 85.7%, precision was 86.2%, recall was 85.3%, and F1 score was 85.7%. All of these evaluation metrics were well above the 75% threshold, demonstrating the model could reliably predict side effects in a one-shot learning scenario.

Table IV
MODEL EVALUATION WITH 7 TESTING DRUGS

Drug Name	Accuracy	Precision	Recall	F1 Score
Chloroquine	85.71%	88.57%	85.71%	85.08%
Famotidine	89.29%	92.50%	89.29%	89.82%
Guanfacine	96.43%	93.05%	96.43%	94.69%
Hydroxychloroquine	78.57%	76.07%	78.57%	77.26%
Oseltamivir	85.71%	82.77%	85.71%	83.99%
Prednisolone	82.14%	79.01%	82.14%	79.79%
Ritonavir	96.43%	92.99%	96.43%	94.68%
Overall	**87.76%**	**85.42%**	**87.76%**	**86.52%**

Ritonavir's confusion matrix, for example, is represented in figure 2 and 3, it is clear to notice the distribution of true positive and false negative is relatively significant enough to make this model accurate.

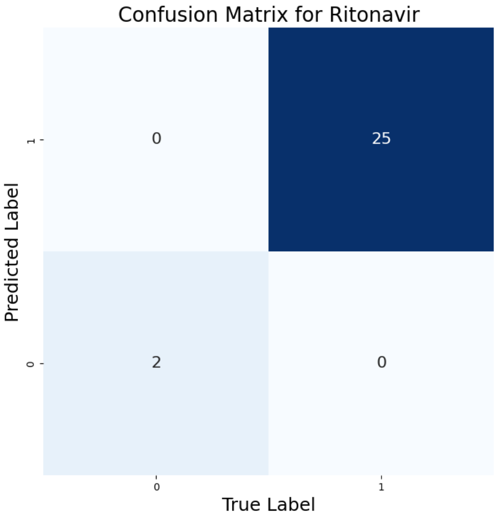


Figure 2. Ritonavir Confusion Matrix

Notably, through table IV, four of the drugs tested - Hydroxychloroquine, Oseltamivir , Chloroquine and Ritonavir

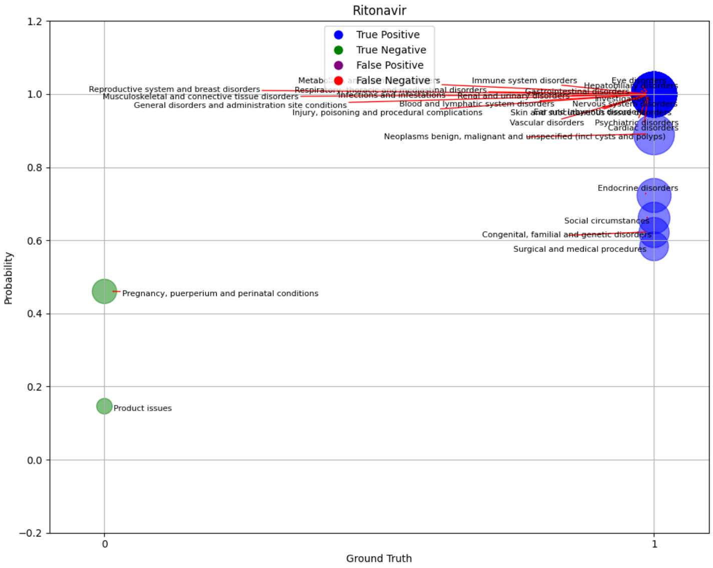


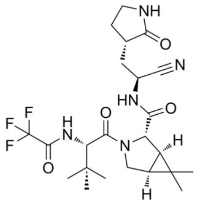
Figure 3. Ritonavir Confusion Matrix

- are currently used as COVID-19 treatments, and the model achieved especially high scores on these drugs. *This suggests the model may be particularly suited for evaluating potential COVID-19 drug candidates.*

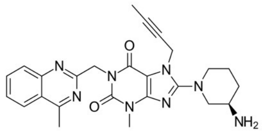
B. Drug Discovery Results Analysis

Next, to investigate the safest drug options, an identification process is carried out, which results in the 5 predicted drugs having the fewest side effects by the model:

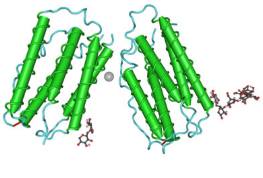
- **Paxlovid (Nirmatrelvir)** with **12** predicted side effects.
- **Linagliptin** with **11** predicted side effects.
- **Interferon-beta-1a** with **11** predicted side effects.
- **Famotidine** with **10** predicted side effects.
- **Hydroxychloroquine** with **10** predicted side effects.



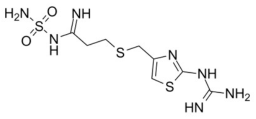
Paxlovid (Nirmatrelvir)
(nP < 0.5: 12)



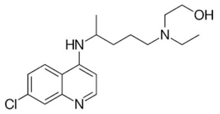
Linagliptin
(nP < 0.5: 11)



Interferon-beta 1a
(nP < 0.5: 11)



Famotidine
(nP < 0.5: 10)

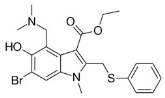


Hydroxychloroquine
(nP < 0.5: 10)

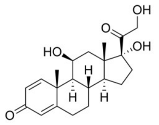
Paxlovid, Interferon-beta-1a, and Hydroxychloroquine are relevant to treating COVID-19. Therefore, these three drugs will be the focus of the next stage of analysis. However, Hydroxychloroquine is reported as an untrusted drug due to a wide range number of side effects. Therefore, only Paxlovid and Interferon-beta-1a are issued to select as the potential drugs to treat COVID-19 disease.

On the other hand, the 5 drugs predicted to have the most side effects, over 20 each, were:

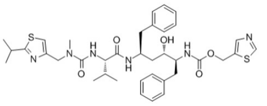
- **Umifenovir** with **21** predicted side effects.
- **Prednisolone** with **22** predicted side effects.
- **Ritonavir** with **25** predicted side effects.
- **Cyclosporine** with **26** predicted side effects.
- **Ivermectin-B1a** with **26** predicted side effects.



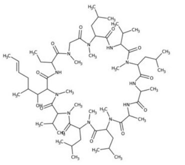
Umifenovir (nP > 0.5): 21



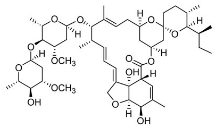
Prednisolone (nP > 0.5): 22



Ritonavir (nP > 0.5): 25



Cyclosporine (nP > 0.5): 26



Ivermectin-B1a (nP > 0.5): 26

Notably, Ritonavir, Umifenovir, and Ivermectin-B1a are currently used to treat COVID-19, so patients taking these medications should be counseled on potential side effects and closely monitored.

IV. Discussion

In this section, the effectiveness of the proposed model is evaluated according to evaluation metrics. Resulting in 7 predicted side-effects drugs compared with the actual side-effect drugs in SIDER, the model is trustworthy or effective due to the average percentage of accuracy, prediction, F1-score, and recall is entirely greater than 85% even though the quality paper was also proposed as a significantly accurate model due to almost all trained models that were greater than 75%, except in MUV dataset. The usefulness in the real world is significantly applicable, particularly in discovering the future bioactivities of developed chemical compounds. COVID-19 is a mostly common disease, which can be used initially in this discovery to find potential drugs with at least side effects for future drug-taken dosage. The detail of toxicological and MUV datasets with a wide range of classes seem not to be enough to predict due to a large number of toxicology and similarity of chemical compounds, which hardly count with substances' complexity.

COVID-19 is a common disease that caused numerous deaths, 7,040,264, in 2024 with a fifth rank compared with other pandemics. This makes scientists and researchers spend a long time investigating and conducting experiments, offering effective treatments and vaccines. With the different medications or newly introduced drugs, the prediction of bioactivities, such as side effects, is crucial before administering these marketed drugs to patients to ensure patient safety and combine future treatments to gain better outcomes in the decision-making process for use in clinical practice. Hydroxychloroquine is commonly known as a COVID-19 drug. However, severe side effects cause patients including cardiac toxicity. As a result, it stopped being released in most countries.

In the research, 24 potential COVID-19 drugs, such as Vancomycin or Paxlovid (Nirmatrelvir), were collected after a long time investigation. These include antimalarial, antibiotic, or immunosuppressive drugs. It is an FDA-approved medication that has demonstrated efficacy in reducing hospitalization and serious consequences in COVID-19 patients. This method, which is in line with the compound-based strategy, will enable us to determine if the model can reliably forecast bioactivities and aid in the identification of novel drug bioactivities, particularly side effects, in the range of 0 to 1.

To obtain the representations of the chemical compounds, the pre-trained GCN-AttentiveFP model was used on SIDER. The BioAct-Het model is then given these representations, along with each side effect class representation that was derived. The Medical Dictionary for Regulatory Activities (MedDRA) provides a standardized medical language for classifying and coding medical history, clinical trial data, and adverse events in drug development and postmarket. This is where the side effects classes come from. Over 20,000 preferred words (PT), or low-level basic classes, are included in MedDRA and are arranged into 27 system organ groupings.

The results give 5 drugs with the least side effects including Paxlovid (Nirmatrelvir), Interferon-beta 1a, Linagliptin, Famotidine, and Hydroxychloroquine. However, Hydroxychloroquine, Linagliptin, and Famotidine are removed due to the report of a large number of side effects and unavailable COVID-19 drugs. Paxlovid (Nirmatrelvir) includes Injury, poisoning and procedural complications, Hepatobiliary Disorders, Metabolism and Nutrition Disorders, Investigations, Musculoskeletal and Connective Tissue Disorders, Gastrointestinal Disorders, Immune System Disorders, Reproductive System and Breast Disorders, General Disorders and Administration Site Conditions, Vascular Disorders, Blood and Lymphatic System Disorders, Skin and Subcutaneous Tissue Disorders, Infections and Infestations, Renal and urinary disorders, Nervous system disorders, Injury, poisoning, and procedural complications, Injury, poisoning and procedural complications; Interferon-beta 1a contains Metabolism and Nutrition Disorders, Investigations, Musculoskeletal and Connective Tissue Disorders, Gastrointestinal Disorders, Immune System Disorders, General Disorders and Administration Site Conditions, Vascular Disorders, Blood and Lymphatic System Disorders, Skin and Subcutaneous Tissue Disorders, Infections and Infestations, Respiratory, thoracic and mediastinal disorders, Psychiatric disorders, Renal and urinary disorders, Cardiac disorders, Nervous system disorders, Injury, poisoning and procedural complications. Nevertheless, more investigation is required to completely comprehend potential side effects and assess the medication's safety and effectiveness in treating COVID-19 and other illnesses, as there have been few studies conducted on it.

The results give 5 drugs requiring careful use Ritonavir, Umifenovir, Cyclosporine, Prednisolone and Ivermectin-B1a. Three of these, such as Ritonavir, Umifenovir and Ivermectin-B1a, are known as COVID-19 drugs; however, the dosage of these drugs should be noticed to patients through prescriptions while obligatorily taking medicine.

Based on the results of 25 drugs, some particular side effects consist of Immune System Disorders, Vascular Disorders, Nervous system disorders, Metabolism and Nutrition Disorders, Gastrointestinal Disorders and Musculoskeletal and Connective Tissue Disorders. Compared with 5 drugs least side effects, some particular side effects are included in table V.

Table V
COMMON SIDE EFFECTS LISTING FOR DRUG COMPARISON

No.	Common Side Effects of the 25 Drugs	Included in Top 5 with Least Side Effects	Paxlovid (Nirmatrelvir)	Interferon-beta 1a
1	Immune System Disorders	Yes	Yes	Yes
2	Vascular Disorders	Yes	Yes	Yes
3	Nervous System Disorders	Yes	Yes	Yes
4	Metabolism and Nutrition Disorders	Yes	Yes	Yes
5	Gastrointestinal Disorders	Yes	Yes	Yes
6	Musculoskeletal and Connective Tissue Disorders	Yes	Yes	Yes
7	General Disorders and Administration Site Conditions	Yes	Yes	Yes
8	Injury, Poisoning and Procedural Complications	Yes	Yes	Yes
9	Renal and Urinary Disorders	Yes	Yes	Yes
10	Skin and Subcutaneous Tissue Disorders	Yes	Yes	Yes
11	Infections and Infestations	Yes	Yes	Yes
12	Investigations	Yes	Yes	Yes

V. CONCLUSION

This study applies the BioAct-Het model to address the BCP problem. The main contributions of this study can be attributed to two key factors. Firstly, a novel approach of using Bio-Prof (Bioactivity class) to represent bioactivity classes as input of 25 investigated potentially marketed drugs. Moreover, the use of a heterogeneous SNN named BioAct-Het is motivated by the complexity and diversity of the relationships between chemical substances and bioactivity classes, resulting in embedding vectors. The model contains two branches for embedding compounds and bioactivity classes in similar concepts captured in a unified latent space. A limitation of this study is the focus on small molecules, which seems to be difficult in larger and complex molecules. Therefore, this factor should be considered in future research endeavors. The use of BioAct-Het as a pre-trained model in three strategies: Association-based, Bioactivity-based, and Compound-based approaches. If research time is extended, this model can be compared with other techniques. After considering under 25 potential drug inputs, Paxlovid (Nirmatrelvir) and Interferon-beta 1a (IFN-β1a) are recommended for use in COVID-19 treatment; however, this study should further investigate to test in vitro, in vivo, combined with the results of in silico.

Associated Content

Data Availability Statement

The data and code underlying this article are available publicly at https://github.com/Xernnn/Bioact-Pred-Het-SNN. However, data sets were derived from sources in the public domain.

Author Contributions

Table VI
AUTHOR CONTRIBUTIONS

Name	Task	Responsibility
Doan Huu Thanh	Studied and conducted COVID-19 diseases and marketed drugs with SMILES structure	Supervised by Pham Hai Nam
Nguyen Son	Developed and implemented the method	Interpreted results, wrote manuscript on LaTeX platform, administered project
Le Mau Minh Phuc	Experimented and conceptualized using the Siamese Neural Network (SNN)	Edited manuscript
Pham Hai Nam	Supervised Doan Huu Thanh, confirmed medical information related to COVID-19	Interpreted results, wrote manuscript, administered project

BIBLIOGRAPHY

[1] G. W. Bemis and M. A. Murcko. The properties of known drugs. *American Chemical Society*, 39(15):2887–2893, 1965. doi: https://doi.org/10.1021/jm9602928.

[2] M. P. Heyrati, Z. Ghorbanali, M. Akbari, G. Pishgahi, and F. Zare-Mirakabad. Bioact-het: A heterogeneous siamese neural network for bioactivity prediction using novel bioactivity representation. *American Chemical Society*, 8(47): 44375–45163, 2023. doi: https://doi.org/10.1021/acsomega.3c05778.

[3] J. Hughes, S. Rees, S. Kalindjian, and K. Philpott. Principles of early drug discovery. *British Journal of Pharmacology*, 162(6):1239–1249, 2011. doi: https://doi.org/10.1111/j.1476-5381.2010.01127.x.

[4] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. *CoRR*, abs/1609.02907, 2016. doi: https://doi.org/10.48550/arXiv.1609.02907.

[5] M. Li, J. Zhou, J. Hu, W. Fan, Y. Zhang, Y. Gu, and G. Karypis. Dgl-lifesci: An open-source toolkit for deep learning on graphs in life science. *American Chemical Society*, 6(41):26822–27599, 2021. doi: https://doi.org/10.1021/acsomega.1c04017.

[6] H. Morgan. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *American Chemical Society*, 5(2):107–113, 1965. doi: https://doi.org/10.1021/c160017a018.

[7] C. Nantasenamat, C. Isarankura-Na-Ayudhya, and V. Prachayasittikul. Advances in computational methods to predict the biological activity of compounds. *Expert Opinion on Drug Discovery*, 5(7):633–654, 2010. doi: https://doi.org/10.1517/17460441.2010.492827.

[8] J.-L. Reymond, L. Ruddigkeit, L. Blum, and R. van Deursen. The enumeration of chemical space. *WIREs Computational Molecular Science*, 2(5):717–733, 2012. doi: https://doi.org/10.1002/wcms.1104.

[9] R. D. Shih, H. M. Johnson, D. G. Maki, and C. H. Hennekens. Hydroxychloroquine for coronavirus: The urgent need for a moratorium on prescriptions. *The American Journal of Medicine*, 133(9):1007–1008, 2020. doi: https://doi.org/10.1016/j.amjmed.2020.05.005.

[10] D. Vella and J.-P. Ebejer. Few-shot learning for low-data drug discovery. *American Chemical Society*, 63(1):27–42, 2023. doi: https://doi.org/10.1021/acs.jcim.2c00779.

[11] D. Weininger. Smiles, a chemical language and information system. *American Chemical Society*, 28(1):31–36, 1988. doi: https://doi.org/10.1021/ci00057a005.

[12] O. J. Wouters, M. Mckee, and J. Luyten. Estimated research and development investment needed to bring a new medicine to market, 2009-2018. *British Journal of Pharmacology*, 323(9):844–853, 2020. doi: https://doi.org/10.1001/jama.2020.1166.

[13] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, and V. Pande. Moleculenet: a benchmark for molecular machine learning. *Chem. Sci.*, 9:513–530, 2018. doi: http://dx.doi.org/10.1039/C7SC02664A.

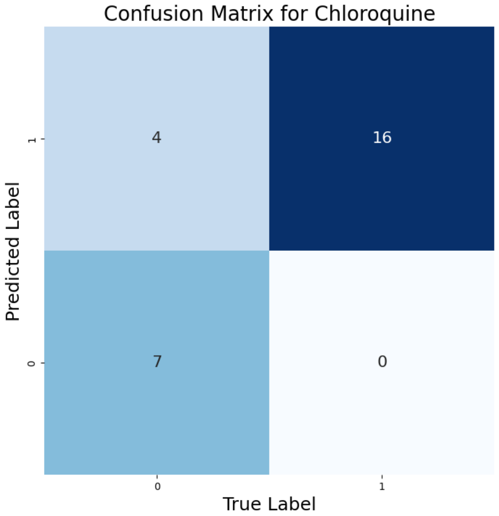
[14] Z. Xiong, D. Wang, X. Liu, F. Zhong, X. Wan, X. Li, Z. Li, X. Luo, K. Chen, H. Jiang, and M. Zheng. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *American Chemical Society*, 63(16):8749–8760, 2020. doi: https://doi.org/10.1021/acs.jmedchem.9b00959.

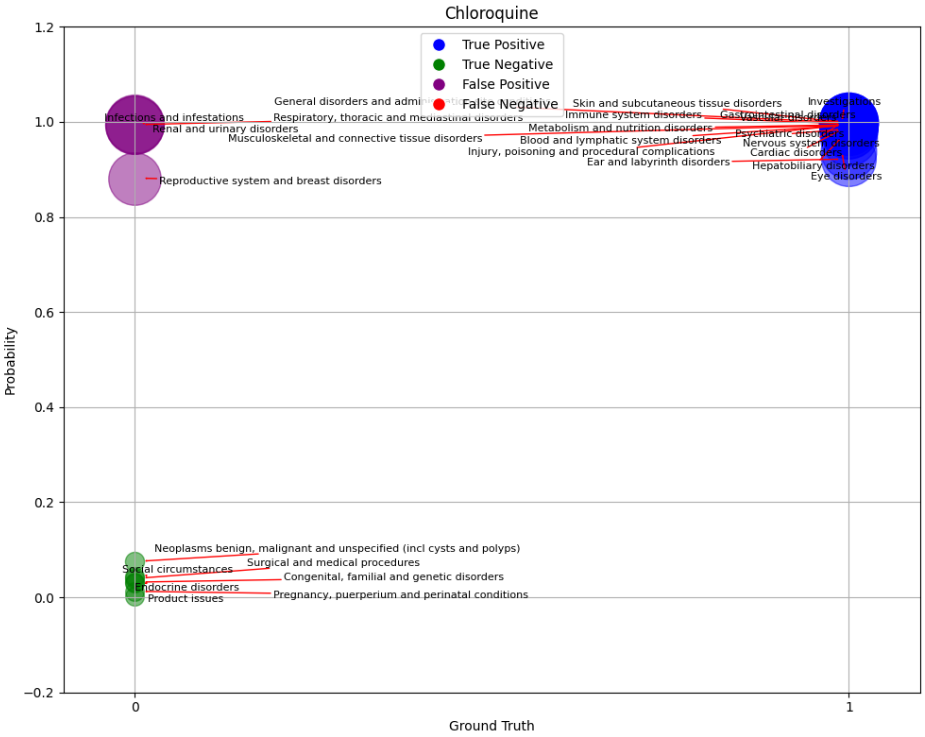
APPENDICES

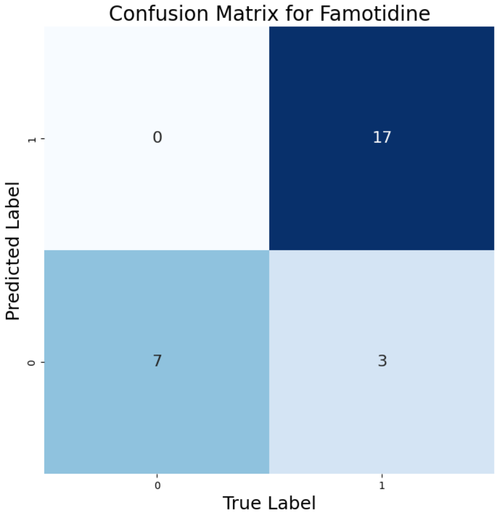
APPENDIX 1: Our Dataset

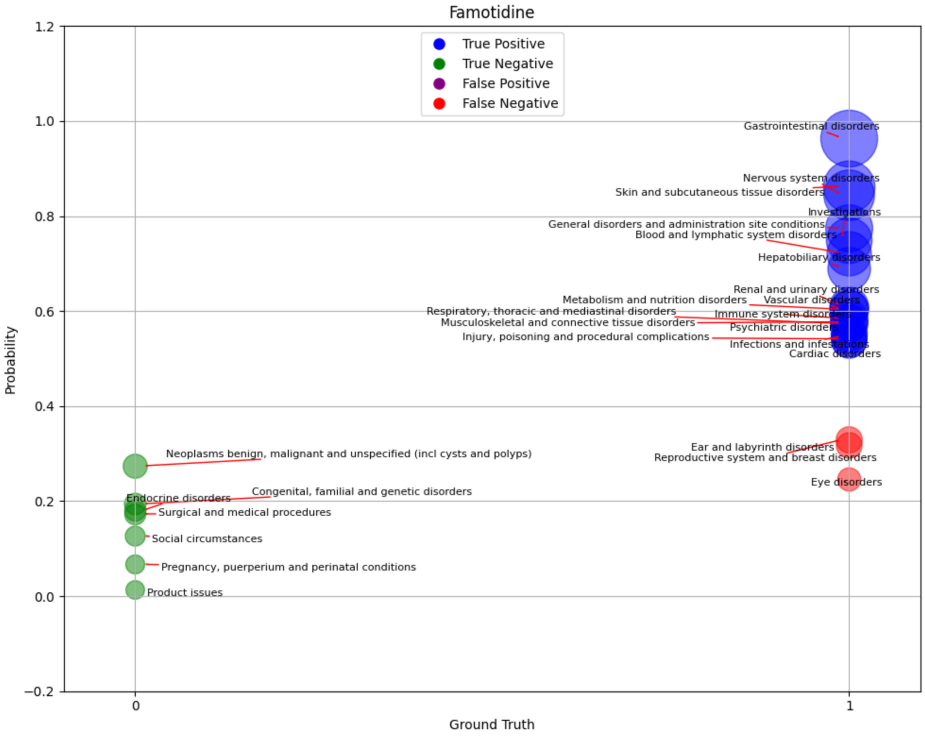
Drug	Chemical Compound	SMILES Structure
Vancomycin	$C_{66}H_{75}Cl_{2}NO_{21}$	Cl.CN[C@H](CC(C)C)C(=O)N[C@@H]1[C@H](O)C2=CC=C(OC3=C(O[C@@H]4O[C@H](CO)[C@@H](O)[C@H](O)[C@H]4O[C@H]4C[C@](C)(N)[C@H](O)[C@H](C)O4)C4=CC(=C3)[C@@H](NC(=O)[C@H](CC(N)=O)NC1=O)C(=O)N[C@@H]1C3=CC(=C(O)C=C3)C3=C(O)C=C(O)C=C3[C@H](NC(=O)[C@@H](NC1=O)[C@H](O)C1=CC(Cl)=C(O4)C=C1)C(O)=O)C(Cl)=C2
Cyclosporine	$C_{56}H_{77}N_{11}O_{11}$	CC[C@@H]1NC(=O)[C@H]([C@H](O)[C@H](C)C\C=C\C)N(C)C(=O)[C@H](C(C)C)N(C)C(=O)[C@H](CC(C)C)N(C)C(=O)[C@H](CC(C)C)N(C)C(=O)[C@@H](C)NC(=O)[C@H](C)NC(=O)[C@H](CC(C)C)N(C)C(=O)[C@@H](NC(=O)[C@H](CC(C)C)N(C)C(=O)CN(C)C1=O)C(C)C
Fingolimod	$C_{19}H_{33}NO_{2}$	Cl.CCCCCCCCC1=CC=C(CCC(N)(CO)CO)C=C1
Remdesivir	$C_{27}H_{35}N_{5}O_{8}P$	CCC(CC)COC(=O)C(C)NP(=O)(OCC1C(C(C(O1)(C#N)C2=CC=C3N2N=CN=C3N)O)O)OC4=CC=CC=C4
Oseltamivir	$C_{16}H_{31}N_{3}O_{8}$	CCC(CC)OC1C=C(CC(C1NC(=O)C)N)C(=O)OCC
Ritonavir	$C_{37}H_{47}N_{3}O_{7}S$	CC(C)C1=NC(=CS1)CN(C)C(=O)NC(C(C)C)C(=O)NC(CC2=CC=CC=C2)CC(C(CC3=CC=CC=C3)NC(=O)OCC4=CN=CS4)O
Molnupiravir	$C_{16}H_{16}F_{3}N_{3}O_{5}$	CC(C)C(=O)OCC1C(C(C(O1)N2C=CC(=NC2=O)NO)O)O
Paxlovid (Nirmatrelvir)	$C_{36}H_{45}N_{5}O_{7}S$	CC1(C2C1C(N(C2)C(=O)C(C(C)(C)C)NC(=O)C(F)(F)F)C(=O)NC(CC3CCNC3=O)C#N)C
Guanfacine	$C_{9}H_{9}Cl_{2}N_{3}O$	C1=CC(=C(C(=C1)Cl)CC(=O)N=C(N)N)Cl
Liraglutide	$C_{166}H_{250}N_{40}O_{45}S$	CCCCCCCCCCCCCCCC(=O)NC(CCC(=O)NCCCCC(C(=O)NC(CCC(=O)O)C(=O)NC(CC1=CC=CC=C1)C(=O) NC(C(C)CC)C(=O) NC(C)C(=O)NC(CC2=CNC3=CC=CC=C32)C(=O)NC(CC(C)C)C(=O)NC(C(C)C)C(=O)NC(CCCNC(=N)N)C(=O)NCC(=O)NC(CCCNC(=N)N)C(=O)NCC(=O)O)NC(=O)C(C)NC(=O)C(C)NC(=O)C(CCC(=O)N)NC(=O)CNC(=O)C(CCC(=O)O)NC(=O)C(CC(C)C)NC(=O)C(CC4=CC=C(C=C4)O)NC(=O)C(CO)NC(=O)C(CO)NC(=O)C(C(C)C)NC(=O)C(CC(=O)O)NC(=O)C(CO)NC(=O)C(C(C)O)NC(=O)C(CC5=CC=CC=C5)NC(=O)C(C(C)O)NC(=O)CNC(=O)C(CCC(=O)O)NC(=O)C(C)NC(=O)C(CC6=CN=CN6)N)C(=O)O
Linagliptin	$C_{25}H_{28}N_{7}O_{2}$	CC#CCN1C2=C(N=C1N3CCCC(C3)N)N(C(=O)N(C2=O)CC4=NC5=CC=CC=C5C(=N4)C)C
Baricitinib-Phosphate	$C_{16}H_{17}N_{6}O_{3}S$	CCS(=O)(=O)N1CC(C1)(CC#N)N2C=C(C=N2)C3=C4C=CNC4=NC=N3.OP(=O)(O)O
Dexamethasone-Sodium-Phosphate	$C_{22}H_{99}FO_{5}P$	CC1CC2C3CCC4=CC(=O)C=CC4(C3(C(CC2(C1(C(=O)COP(=O)([O-])O-])O)C)O)FC.[Na+].[Na+]
Tocilizumab	$C_{66}H_{76}N_{10}O_{16}S$	CCNC(=O)[C@@H]1CC[C@@H]2C@@HCC[C@@H]2C1
Remdesivir (Veklury)	$C_{27}H_{35}N_{5}O_{8}P$	CCC(CC)COC(=O)C(C)NP(=O)(OCC1C(C(C(O1)(C#N)C2=CC=C3N2N=CN=C3N)O)O)OC4=CC=CC=C4
Anakinra	$C_{76}H_{105}N_{19}O_{19}S_{2}$	CC(C)C@HC(=O)NCC(=O)N
Chloroquine	$C_{18}H_{26}ClN_{3}$	CCN(CC)CCCC(C)NC1=C2C=CC(=CC2=NC=C1)Cl
Hydroxychloroquine	$C_{18}H_{26}ClN_{3}O_{3}$	CCN(CCCC(C)NC1=C2C=CC(=CC2=NC=C1)Cl)CCO
Famotidine	$C_{8}H_{15}N_{3}O_{2}S$	C1=C(N=C(S1)N=C(N)N)CSCCC(=NS(=O)(=O)N)N
Umifenovir	$C_{22}H_{28}ClN_{3}O_{5}S$	CCOC(=O)C1=C(N(C2=CC(=C(C(=C21)CN(C)C)O)Br)C)CSC3=CC=CC=C3
Ivermectin-B1a	$C_{50}H_{76}O_{14}$	CCC(C)C1C(CCC2(O1)CC3CC(O2)CC=C(C(C(C=CC=C4COC5C4(C(C=C(C5O)C)C(=O)O3)O)C)OC6CC(C(C(O6)C)OC7CC(C(C(O7)C)O)OC)OC)C)C
Prednisolone	$C_{21}H_{26}O_{5}$	CC12CC(C3C(C1CCC2(C(=O)CO)O)CCC4=CC(=O)C=CC34C)O
Fluvoxamine	$C_{15}H_{21}F_{3}N_{2}O_{2}$	COCCCCC(=NOCCN)C1=CC=C(C=C1)C(F)(F)F
Baxdrostat	$C_{22}H_{25}N_{3}O_{2}$	CCC(=O)NC1CCCC2=C1C=NC=C2C3=CC4=C(C=C3)N(C(=O)CC4)C

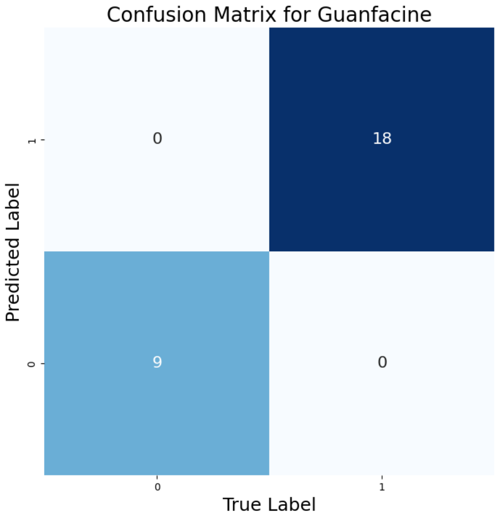
APPENDIX 2: The Bubble Charts and the Confusion Matrices of the 7 Selected Drugs

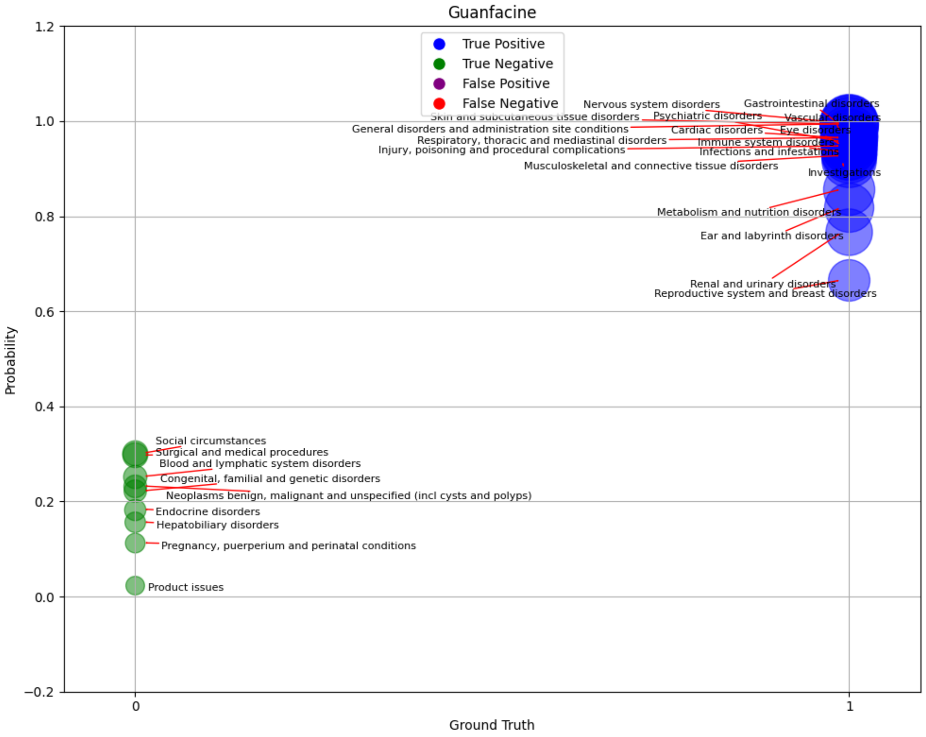


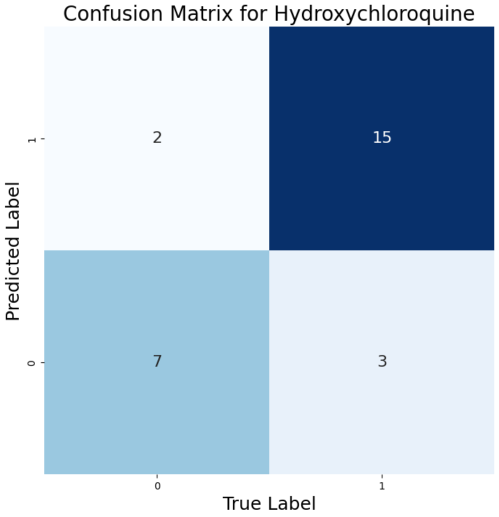


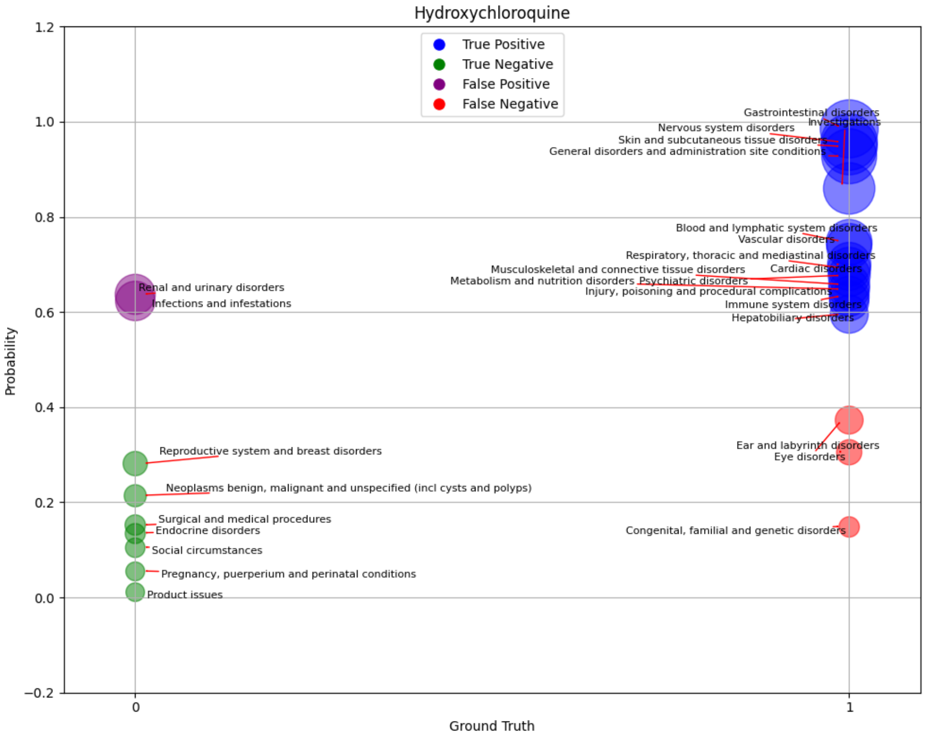


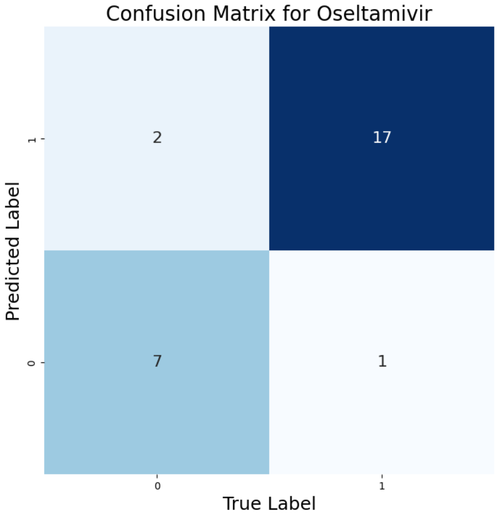


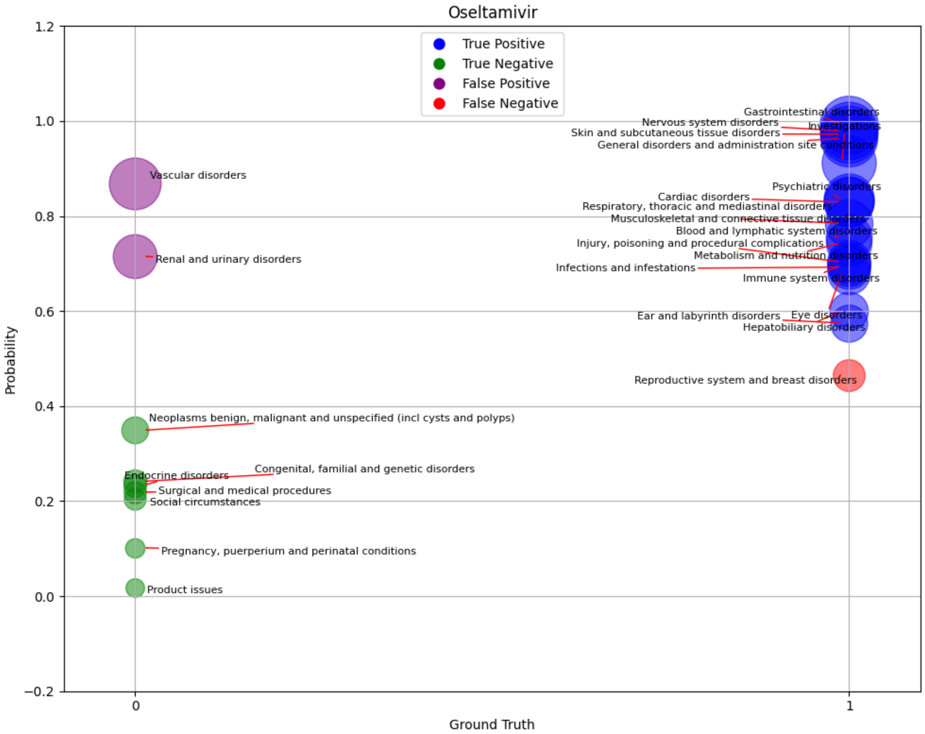

Confusion Matrix for Guanfacine
Predicted Label
True Label
1
0
0
1
0
18
9
0

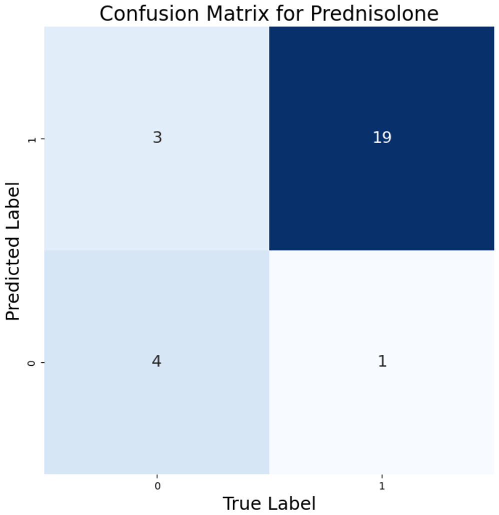

Guanfacine
True Positive
True Negative
False Positive
False Negative
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
Probability
0
1
Ground Truth
Nervous system disorders
Gastrointestinal disorders
Skin and subcutaneous tissue disorders
Psychiatric disorders
Vascular disorders
General disorders and administration site conditions
Cardiac disorders
Eye disorders
Respiratory, thoracic and mediastinal disorders
Immune system disorders
Injury, poisoning and procedural complications
Infections and infestations
Musculoskeletal and connective tissue disorders
Investigations
Metabolism and nutrition disorders
Ear and labyrinth disorders
Renal and urinary disorders
Reproductive system and breast disorders
Social circumstances
Surgical and medical procedures
Blood and lymphatic system disorders
Congenital, familial and genetic disorders
Neoplasms benign, malignant and unspecified (incl cysts and polyps)
Endocrine disorders
Hepatobiliary disorders
Pregnancy, puerperium and perinatal conditions
Product issues

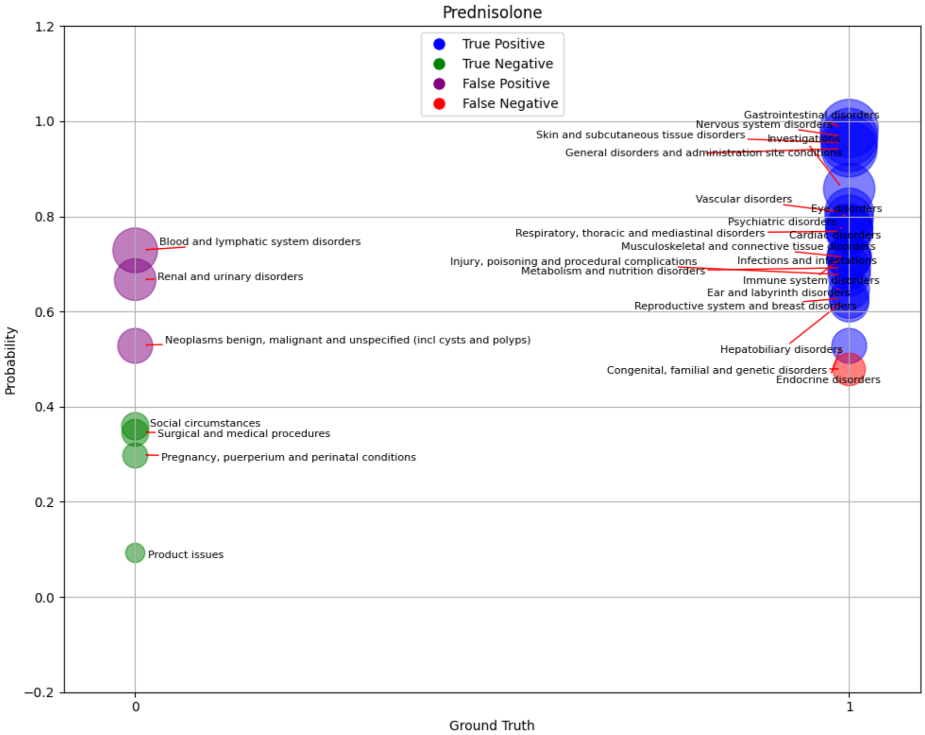

Confusion Matrix for Hydroxychloroquine
Predicted Label
True Label
1
0
0
1
2
15
7
3


Hydroxychloroquine
True Positive
True Negative
False Positive
False Negative
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
Probability
0
1
Ground Truth
Gastrointestinal disorders
Investigations
Nervous system disorders
Skin and subcutaneous tissue disorders
General disorders and administration site conditions
Blood and lymphatic system disorders
Vascular disorders
Respiratory, thoracic and mediastinal disorders
Musculoskeletal and connective tissue disorders
Cardiac disorders
Metabolism and nutrition disorders
Psychiatric disorders
Injury, poisoning and procedural complications
Immune system disorders
Hepatobiliary disorders
Renal and urinary disorders
Infections and infestations
Reproductive system and breast disorders
Ear and labyrinth disorders
Eye disorders
Neoplasms benign, malignant and unspecified (incl cysts and polyps)
Surgical and medical procedures
Endocrine disorders
Congenital, familial and genetic disorders
Social circumstances
Pregnancy, puerperium and perinatal conditions
Product issues


Confusion Matrix for Oseltamivir
Predicted Label
1
0
2
17
7
1
0
1
True Label


Oseltamivir
True Positive
True Negative
False Positive
False Negative
Probability
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
0
1
Ground Truth
Gastrointestinal disorders
Nervous system disorders
Investigations
Skin and subcutaneous tissue disorders
General disorders and administration site conditions
Vascular disorders
Psychiatric disorders
Cardiac disorders
Respiratory, thoracic and mediastinal disorders
Musculoskeletal and connective tissue disorders
Blood and lymphatic system disorders
Injury, poisoning and procedural complications
Renal and urinary disorders
Metabolism and nutrition disorders
Infections and infestations
Immune system disorders
Ear and labyrinth disorders
Eye disorders
Hepatobiliary disorders
Reproductive system and breast disorders
Neoplasms benign, malignant and unspecified (incl cysts and polyps)
Congenital, familial and genetic disorders
Endocrine disorders
Surgical and medical procedures
Social circumstances
Pregnancy, puerperium and perinatal conditions
Product issues


Confusion Matrix for Prednisolone
Predicted Label
1
0
3
19
4
1
0
1
True Label


Prednisolone
True Positive
True Negative
False Positive
False Negative
Probability
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
0
1
Ground Truth
Gastrointestinal disorders
Nervous system disorders
Skin and subcutaneous tissue disorders
Investigations
General disorders and administration site conditions
Vascular disorders
Eye disorders
Psychiatric disorders
Blood and lymphatic system disorders
Respiratory, thoracic and mediastinal disorders
Cardiac disorders
Musculoskeletal and connective tissue disorders
Infections and infestations
Injury, poisoning and procedural complications
Metabolism and nutrition disorders
Renal and urinary disorders
Immune system disorders
Ear and labyrinth disorders
Reproductive system and breast disorders
Neoplasms benign, malignant and unspecified (incl cysts and polyps)
Hepatobiliary disorders
Congenital, familial and genetic disorders
Endocrine disorders
Social circumstances
Surgical and medical procedures
Pregnancy, puerperium and perinatal conditions
Product issues