

Application of machine learning-based Quantitative Structure-
Activity Relationship in modern chemoinformatic

CHEM0027

Code: GERVASIO2

Supervisor name: Prof Francesco Gervasio

Date of submission: 19/02/2022

Table of contents

Research highlight.....	3
Reflective commentary	5
Literature Review	6
Introduction	6
Quantitative Structure-Activity Relationship.....	6
Ligand-based vs. receptor-based drug design	6
Descriptors	7
Machine learning.....	8
Development of machine learning	8
Foundations of machine learning	8
Supervised machine learning	9
Unsupervised machine learning	10
Receiver Operating Characteristic curve.....	10
Coefficient of determination (R^2 value)	11
Case study: Covid-19 therapeutics.....	12
Case study: Triple-negative Breast Cancer (TNBC)	18
Conclusion.....	22
Reference.....	24

Research highlight

In late autumn 2019, the world was found unprepared for a pandemic that resulted in a global health crisis. Fortunately, the efforts of frontier medical staff and the application of hygiene policy contained the disease, until researchers created their special weapon in a very short time—in the form of vaccines against the threatening storm in its roaring fury. Behind the rapid response, the forwarding progress of drug discovery makes everything possible.

Quantitative Structural-Activity Relationship (QSAR) defines how molecular structure can connect to physical, chemical, and biological activities. As a reinforcement to the traditional process, QSAR is an in-silico method that can reduce the research costs and time of drug discovery to a great extent. It screens through giant databases and outputs the most suitable drugs determined by computer algorithms. QSAR is not a brand-new idea, instead it was proposed by scientists a few decades ago. 'To be willing is to be able,' says the old proverb. The development of artificial intelligence derives many algorithms, also known as Machine Learning, which applies across disciplines. Nowadays, one of the most promising ways to find a therapeutic for Covid-19 is to adopt a machine learning model along with the concepts of QSAR.

QSAR always starts with a giant database. Another reason humans can fight the virus efficiently is that it is not the first time to meet the prefix 'corona-'. SARS-CoV-1 and MERS-CoV. They once showed up in 2003 and 2012 respectively and took away many people's lives at their origin. After the outbreaks, scientists understood the pathogenic pathway and more importantly found that some viral features are highly conserved among coronaviruses. No surprise, '3 chymotrypsin-like proteases (3CLpro) and RNA-dependent RNA polymerase (RdRp) are two ideal protein targets for QSAR modeling,' states Julian Ivanov from American Chemical Society. It means that many drugs would potentially share similar activities against the viruses. In other words, it will act as strong candidates that help to build a database.

There are two databases (FDA-approved drugs and CAS REGISTRY substances) suggested by the authors in addition to the 'SARS-, MERS-, and COVID-19- related documents published since 2003'. By using these databases, algorithms can play a constructive role in the following machine learning stage. Support Vector Classifier and XGBoost are the models due to their high robustness of classifying drugs into 'active' and 'inactive' groups, which is marked by the Receiver Operating Characteristic (ROC) curve. 'This can be attributed to (i) the more separated distributions of actives and inactives and (ii) the high diversity of active and inactive examples in the training data that were prepared by CAS scientists,' says Ivanov. The team also applied specific criteria such as the drug concentration aiming at the protein target (e.g., IC50) to ensure that the labels are well distinguished.

Eventually, 37 candidates out of 1087 from FDA, 970 out of 49,437 from CAS, and 2500 from published documents were labeled as 'active'. QSAR is only one of the first steps of drug discovery. There will be a long way to refine the list, without a doubt. In practice, some candidates are ready to enter the next stage; Some provide assistance such as increasing blood flow against the decreasing oxygen uptake (one of the Covid-19 syndromes); Some are abandoned due to multiple reasons such as being toxic to normal cells. The common functional groups among drug candidates are identified as structural alerts, which will orient scientists in the correct direction by combining the most effective ones.

QSAR analysis will complement human resources to establish a more efficient and strictly controlled drug discovery process. The repeated works are left to machines so that people will be more focused on creation. With the fast development of computational power, humans will finally catch up with the demand of solving health problems.

Reference:

J. Ivanov, D. Polshakov, J. Kato-Weinstein, Q. Zhou, Y. Li, R. Granet, L. Garner, Y. Deng, C. Liu, D. Albaiu, J. Wilson and C. Aultman, *ACS Omega*, 2020, **5**, 27344–27358.

Reflective commentary

The motivation for studying machine learning and QSAR originates from my personal interest in data science. Understanding the frontier application of chemistry will help me realize how people face significant health challenges when the best and most efficient solutions are needed. At the moment, Covid-19 changes its Receptor Binding Domain (RBD) of spike proteins to escape the immune response from the vaccines. Machine learning is a tool that makes the drug production follows the mutation tightly. It is better to understand the robustness of machine-learning-based QSAR for different demands. Therefore, I choose another case study of triple-negative breast cancer which is a dangerous disease without appropriate solutions. The rest of the papers are also important for understanding machine learning mechanisms, the overview of QSAR, and some statistical measures such as ' R^2 '. I want to combine real cases with my knowledge of machine learning to explore the future possibility of solving similar problems independently.

Literature Review

Introduction

Covid-19 pandemic became a global emergency in late autumn 2019. The requirement of fast drug discovery has been warned by this abrupt assault on human health. In recent years, computational techniques were developed at an unprecedented speed that helped boost the process of identifying drug candidates in the treatment of complex diseases. Quantitative Structure-Activity Relationship (QSAR) assumes that the molecular activity is directly related to the molecular structure rather than the chemical properties.¹¹ It is achieved by analyzing historical data and combining existing structural data to predict the possible drug candidates. With the aid of this in-silico process, a large number of structures were found from drug databases such as FDA and CAS. Time and human resources costs are reduced to a considerable degree.

Quantitative Structure-Activity Relationship

Ligand-based vs. receptor-based drug design

A hit discovery is a process of confirming the binding affinity of drug candidates to their biological targets. QSAR is often used in the next drug discovery step following this confirmation, which is also known as hit-to-lead analysis.⁸ The quality of hit-to-lead optimization determines drug discovery orientation because it considerably condenses the number of drug candidates through screening. The concept of linking molecular activity to the molecular structure was proposed by Corwin Hansch and Toshio Fujita in 1961. As shown in figure 1, the Hansch equation $Biological\ activity = Function(parameters)$ assumes multivariate linearity in the model and allows scientists to extract patterns from physiochemical descriptors to better understand the underlying principles of biological activity.⁴

NADH Oxidation, Inhibition by Barbiturates (3)

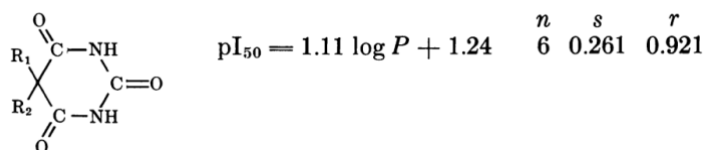


Figure 1: Inhibition of NADH Oxidation by Barbiturates, Hansch equation. pI_{50} is the negative logarithm of molar concentration inhibiting enzyme activity by 50%. (Adapted from Valkenburg *et al.*, 1974)

There is another approach that defines the way of ligand-receptor interaction. The receptor-based drug design, using methods such as Molecular dynamics (MD) simulation, develop in parallel with the ligand-based drug design (QSAR) due to the fast improvement of computational power. The structural properties of ligand strongly depend on the interactions with the receptor, such as the stretch and bend of chemical bonds. In contrast to MD, QSAR takes advantage of interpreting drug efficacy without understanding the corresponding receptor.⁶ It makes QSAR more flexible in finding a new ligand or repurposing it from previous knowledge.

Descriptors

QSAR is often used to predict biological and chemical activities. For example, biological activities such as binding affinity and toxicity, chemical activities such as melting point and solubility. According to the Hansch equation, the inputs are physiochemical parameters, also known as molecular descriptors. In the past decades, chemical descriptors have evolved into structural features belonging to different data types and dimensions. 1D descriptors such as molecular weight and chain length are usually degenerate, indicating that compounds can share the same values of these features. To a higher dimension, the structural details further differentiate thus it is less likely the compounds are mapped into the same value, such as 2D molecular fingerprint and 3D quantum chemical descriptors.⁶

Machine learning

Development of machine learning

The development of machine learning is closely correlated to the timeline of artificial intelligence. In the 1980s, the rise of the expert system is a tremendous breakthrough for artificial intelligence. It could perform multiple tasks such as language translation and image interpretation with logic as a real expert. However, it only performed tasks in a specific field. Without the support of fast calculation and the capacity of databases, it was limited by mathematics theories. Nowadays, machine learning is used to perform regression and classification tasks by different algorithms. These algorithms are categorized into supervised and unsupervised learning, depending on whether there are labels on the training data. In QSAR, data are usually trained with the labels 'active' and 'inactive' so that machine can perform classification tasks that would hopefully separate the labels as better as possible. In the therapeutics discovery of Covid-19, both Support Vector Machine and XGBoost can perform supervised classification tasks.

Foundations of machine learning

'Machine learning is a field of study that gives computers the ability to learn without explicitly explained,' it is the first definition of machine learning claimed by Arthur Samuel in the 1950s. A more formal and modern definition was provided by Tom Mitchell: A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P , if its performance at tasks in T , as measured by P , improves with experience E . For the checkers playing examples, the experience E would be the experience of having the program play tens of thousands of games itself. Task T would be playing checkers, and the performance measure P will be the probability that wins the next game of checkers against some new opponent.

Supervised machine learning

In supervised learning, a data set is given, and people already know what correct output should look like, having the idea that there is a relationship between the input and the output. In an example of breast cancer, people would try to predict breast cancer tumors as malignant or benign based on medical records. A data set with correct answers is the training set that is used to generate a general pattern for breast cancer. The machine learning task contains two variables shown as tumor size and age in figure 2.

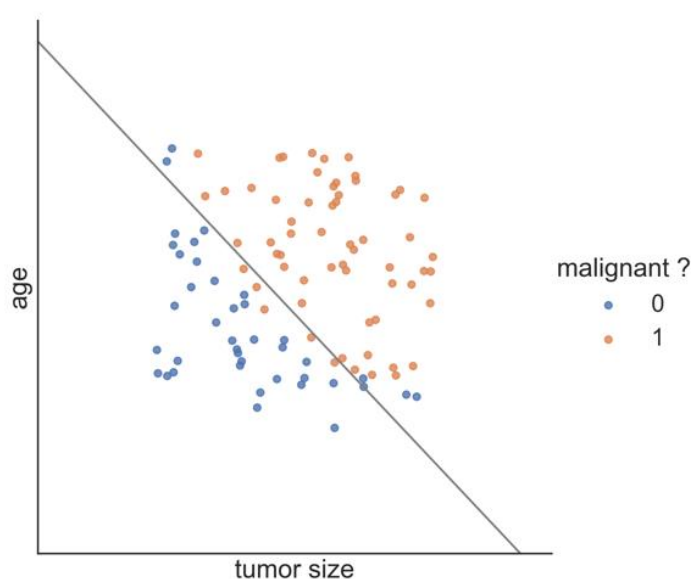


Figure 2: Supervised learning: classification of breast cancer tumors. (Produced by author)

The color of data labels indicates the types of tumors. A decision boundary separates the malignant and benign tumors in the training set. After the training with many medical records, judgment can be made to a new patient according to her age and tumor size. If her data lies above the decision boundary, she may tragically have breast cancer. Figure 2 is a straightforward implication of supervised learning using logistic regression. In the case study of Covid-19 therapeutics, the variables become chemical descriptors. Support vector machine is an algorithm that can deal with hundreds of descriptors. It is suitable for large databases and usually generates robust results.

Unsupervised machine learning

Unsupervised learning allows people to approach problems with little idea of what results should look like. Structures are derived from data where the effects of the variables are not necessarily understood. It is usually done by clustering the data based on relationships among the variables in the data. In the case study of breast cancer, Deep Neural Network (DNN) is proved to perform excellent structure-based drug design even lacking knowledge of the targeting core structure also the best descriptors are challenging to find.

Receiver Operating Characteristic curve

Receiver Operating Characteristic curve (ROC) acts as an essential indicator for the quality of machine learning algorithms.

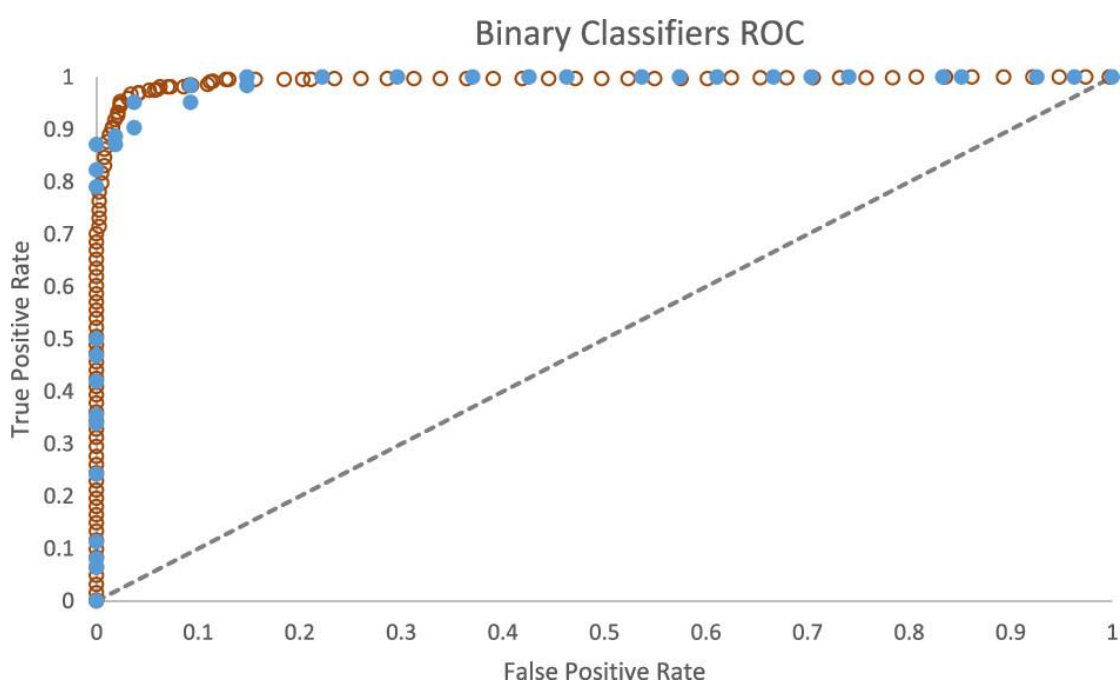


Figure 3: Receiver operating Characteristic curve. (Adapted from Ivanov *et al.*, 2020)

As shown in figure 3, a ROC has a false positive rate (FPR, also known as Sensitivity) on the horizontal axis and a true positive rate (TPR, also known as Recall) on the vertical axis. As

shown in figure 4, True positive is when a positive result is correctly predicted as positive. False positive is when a negative result is wrongly predicted as positive. False positive rate is the partition of false positive in all the negative 'true label'. True positive rate is the partition of true positive in all the positive 'true label'.

pred_label \ true_label	Positive	Negative
Positive	TP	FP
Negative	TN	FN

$$FPR = \frac{FP}{FP+FN} \quad TPR = \frac{TP}{TP+TN}$$

Figure 4: Four types of results in a binary classification problem. (Produced by author)

A threshold value can be set in the example of breast cancer tumors. If the probability of a malignant tumor is higher than the threshold, the prediction on the patient is 'positive'. In figure 3, as the threshold decreases, the TPR and FPR will increase from (0,0) to (1,1) because more and more patients are predicted to have a malignant tumor. The dashed linear line is the random chance (without algorithms) of having a malignant tumor. For an ideal classification model, the optimal probability threshold has TPR=1 and FPR=0, and the ROC equivalently has an area of 1.

Coefficient of determination (R^2 value)

The coefficient of determination (R^2 value) is one way to examine the fitting of data.¹⁰ It measures how well the actual outcomes (e.g., malignant/benign tumor, active/inactive drugs) can be explained by the changes of molecular descriptors. The upper part of the fraction is given by the sum of difference between the experimental result and the predicted result for the i-th

drug in the test sets. The lower part is the variance of the test sets. The better the fitting, the closer the $R^2=1$.

$$R^2 = 1 - \frac{\sum_{i=1}^{n_{test}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{test}} (y_i - \bar{y}_i)^2}$$

Q_{F3}^2 is modified from the original R^2 value.³ It emphasizes the connection between training sets and test sets. The lower part of the fraction changes to the variance of the training sets.

$$Q_{F3}^2 = 1 - \frac{\sum_{i=1}^{n_{test}} (y_i - \hat{y}_i)^2 / n_{test}}{\sum_{j=1}^{n_{tr}} (y_j - \bar{y}_{tr})^2 / n_{tr}}$$

Q_{F3}^2 is applied when different models with the same training sets are compared.⁵

Case study: Covid-19 therapeutics

The medical treatments for SARS-CoV-2 (Covid-19) were mainly palliative two years ago, such as oxygen concentrators and anti-inflammatory drugs. At present, direct treatment has been put into production, such as the antiviral medicines (Sotrovimab and Molnupiravir) offered by the National Health Service (NHS) in the UK. Traditionally, it requires a decade from a drug's blueprint to end the whole circle with clinical trials. The significant therapeutic improvement would not be made without the assistance of machine-learning-based QSAR.

Similar to most antiviral drugs in the market, the orientation of this research is to inhibit some functional proteins inside the coronavirus. 3 chymotrypsin-like protease (3CLpro) is the enzyme of coronavirus that cuts off the polypeptide chains, especially important in processing RNA-dependent RNA polymerase (RdRp). Coronavirus belongs to the family of positive-stranded RNA (+ssRNA) viruses. Viral genomic RNA functions as mRNA that can be translated to form viral proteins immediately upon infection of the host cells. RdRp is the enzyme that catalyzes the replication of viral genomic RNA in host cells. Either breaking the function of

3CLpro or RdRp can reduce coronavirus activity. More importantly, the primary structure (amino acid sequence) of 3CLpro and RdRp are approximately 90% identical to that of the previous human coronavirus.¹¹ It indicates that diverse drug candidates can be repurposed from the treatment of previous diseases, such as SARS-CoV-1 and MERS-CoV.

468 and 1212 active chemicals in the training stage are mainly collected from the recorded study of SARS-CoV-1 and hepatitis C virus (HCV), respectively, to establish the training sets for 3CLpro and RdRp models. The training sets consist of compounds with different core structures (also known as structural alerts, Figure 5).

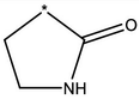
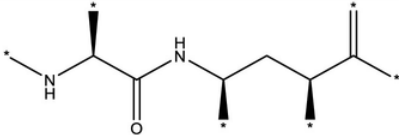
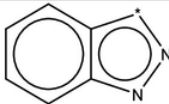
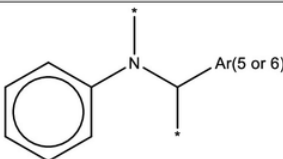
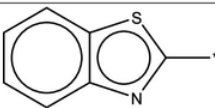

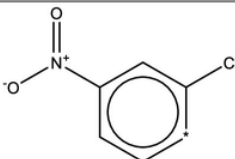
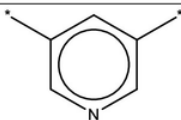
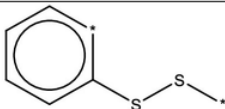
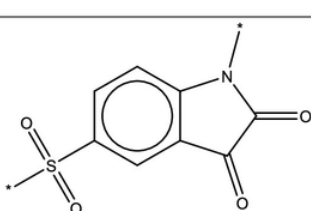
Structural alert	Active substances in training set (N=468)
	138
	85
	60
	52
	40
	35
	26
	24
	18
	10

Figure 5: Structural alerts for the training set. (Adapted from Ivanov *et al.*, 2020)

The structural alerts are the functional fragments that show inhibitory effects for 3CLpro and RdRp. Each training set is partitioned into five parts. Every time the models are trained out of 4 parts, the last part is used to obtain the area under the ROC curve (AUC). After careful

selection of hundreds of descriptors, training sets are loaded. The concentration of inhibitors to achieve a 50% inhibitory effect is the IC₅₀ values. Active and inactive drugs are defined by 'IC₅₀<10 μ M' and 'IC₅₀>100 μ M', respectively. Support Vector Machine (SVM) and XGBoost show the highest robustness (AUC=0.99) among all the machine learning algorithms. As shown in Figure 6, SVM expands the classification task using logistic regression (Figure 2). XGBoost is an enhanced version decision tree that is specially designed for binary classification tasks.

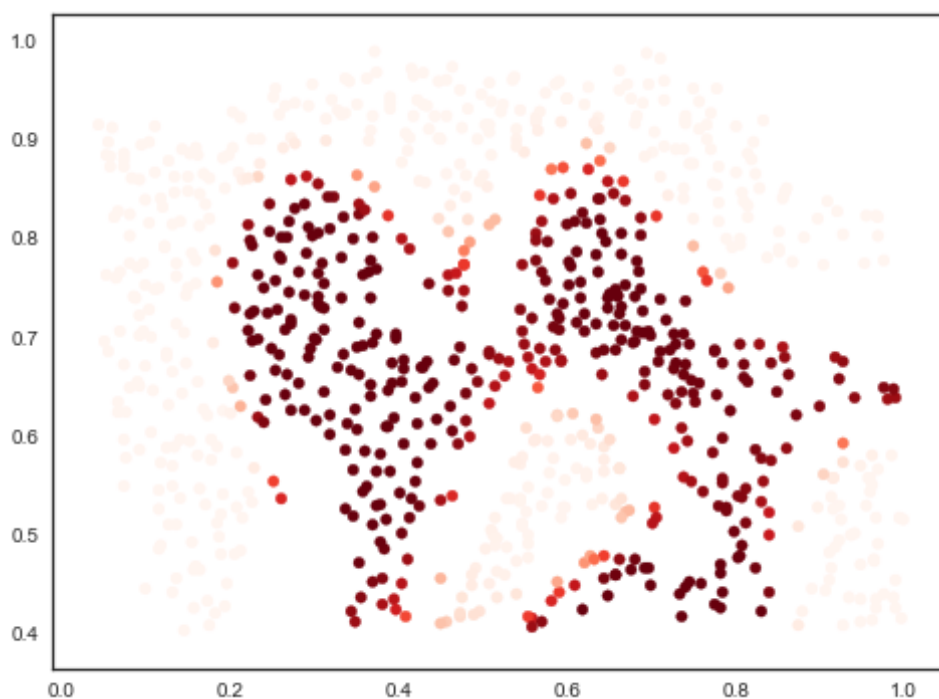


Figure 6: Support Vector Machine (Radial Kernel). (Produced by author)

For the results, databases such as FDA-approved drugs and CAS REGISTRY substances provided 1087 and 50000 input as the drug candidates for therapeutics. SVM and XGBoost predicted over 3000 active candidates against 3CLpro and 21000 against RdRp.¹¹

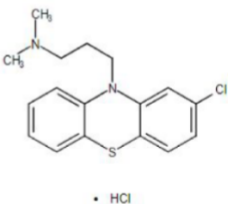
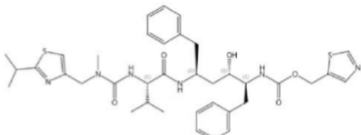
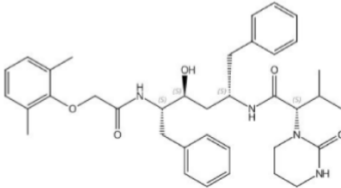
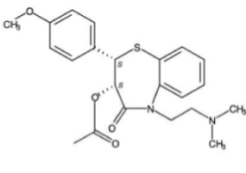
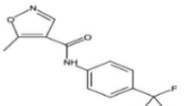
Substance ^{34, 63}	CAS Registry Number	Structure	Mode of action	Current use	In COVID-19 clinical trials ⁶⁵⁻⁶⁷	Data Set ^{34,35}
Chlorpromazine hydrochloride (Thorazine hydrochloride)	69-09-0		dopamine receptor blocker ^{34,64}	treat schizophrenia ^{34,64}	Yes	FDA
Ritonavir (Norvir)	155213-67-5		HIV-1 protease ^{34,64}	treat HIV-1 infection/AIDS ^{34,64}	Yes	FDA
Lopinavir	192725-17-0		HIV-1 protease ^{34,64}	treat HIV-1 infection/AIDS ^{34,64}	Yes	FDA
Diltiazem hydrochloride (Cardizem)	33286-22-5		calcium channel blocker ^{34,64}	treat high blood pressure, angina, and certain heart arrhythmias ^{34,64}	Yes	FDA
Leflunomide (Arava)	75706-12-6		dihydroorotate dehydrogenase inhibitor ^{34,64}	treat rheumatoid and psoriatic arthritis ^{34,64}	Yes	FDA

Figure 7: Examples of active substances predicted by the 3CLpro model. (Adapted from Ivanov et al., 2020)

One of the advantages of machine learning is that the diversity of databases would predict substances with various modes of action. It is noted that the majority of successful prediction comes from FDA-approved drugs. Also, many active substances tested in bioassays (Figure 7) are shown to be HIV-1 protease inhibitors. It leads to the next stage of covid-19 therapeutics discovery that looks for drugs repurposed from HIV-1 protease inhibitors. Furthermore, Diltiazem hydrochloride and Leflunomide provide an ancillary effect for treatment because they act on host cells. The treatment of high blood pressure by Diltiazem hydrochloride reduces

the syndrome from Covid-19 comorbidities such as hypertension. Leflunomide is an inhibitor for dihydroorotate dehydrogenases, the potential protein targets under viral attacks.¹¹

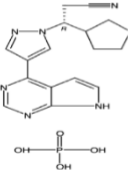
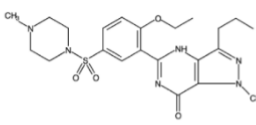
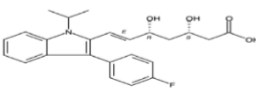
Substance	CAS Registry Number	Structure	Mode of action	Current Use	In COVID-19 clinical trials ⁶⁵⁻⁶⁷	Data Set ^{34,35}
Ruxolitinib phosphate (Jakafi)	1092939-17-7		protein tyrosine kinases JAK 1 and 2 inhibitor ⁷⁷	treat intermediate or high-risk myelofibrosis and polycythemia vera; binding of ruxolitinib to kinases may lead to reduction of inflammation and inhibition of cellular proliferation ⁷⁷	Yes	FDA
Sildenafil (Revatio)	139755-83-2		phosphodiesterase type 5 inhibitor ⁸⁰	treat male erectile dysfunction; improve ability to exercise in adults with pulmonary arterial hypertension by relaxing blood vessels in lungs ⁸⁰	Yes	FDA
Fluvastatin sodium (Lescol XL)	93957-55-2		HMG-CoA reductase inhibitor ⁸¹	reduce risk of heart attack and stroke by reducing LDL cholesterol and triglycerides and increasing HDL cholesterol in blood ⁸¹	No	FDA

Figure 8: Examples of active substances predicted by the RdRp model. (Adapted from Ivanov et al., 2020)

The leading group of active substances predicted by the RdRp model shows structural similarity to Remdesivir (RDV), a nucleotide analogue that has been confirmed for the treatment of covid-19. Typically, RdRp combines with a viral RNA single-strand and simultaneously binds nucleotide triphosphate (NTP) to form the complementary strand by translocation. The activated triphosphate form of RDV (Figure 9) competes with NTP (especially adenosine triphosphate (ATP)) and it is incorporated by the RdRp into the growing RNA chain.^{9,12} It terminates the elongation after three more additions of nucleotides due to the steric hindrance of its cyano group that stalls the RdRp. As predicted by machine learning, Ruxolitinib also contains a cyano group, which is a potent therapeutic for dangerous respiratory complications. Although Fluvastatin sodium fails in covid-19 clinical trials, it still helps lower the risk of cardiovascular disease by controlling HMG-CoA and cholesterol, which is known to be the risk factor for covid-19 infection.

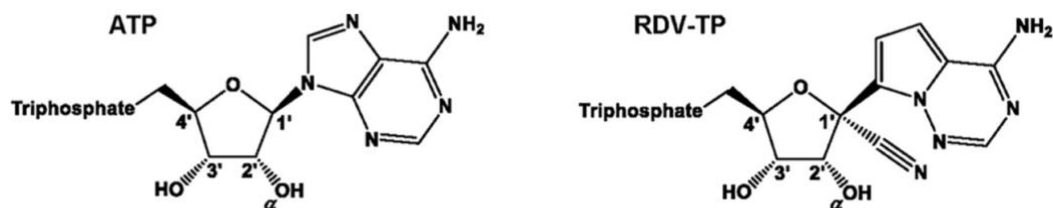


Figure 9: Comparison of adenosine triphosphate and Remdesivir triphosphate. (Adapted from Gordon *et al.*, 2020)

Cinanserin is one of the drugs predicted to be active for covid-19 (Figure 10), which had been evaluated in previous clinical trials for other coronaviruses. The successful prediction of these drugs further validates the robustness of SVM and XGBoost algorithms.

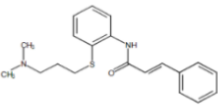
Cinanserin	1166-34-3		SARS-CoV IC ₅₀ = 4.92	CN 1472336; Journal of Virology (2005), 79(11), 7095-7103
------------	-----------	--	-------------------------------------	---

Figure 10: Cinanserin: an inhibitor of 3CLpro of SARS-CoV (2002). (Adapted from Ivanov *et al.*, 2020)

Case study: Triple-negative Breast Cancer (TNBC)

Triple-negative Breast Cancer (TNBC) is a rare disease as it only counts for 15% of all breast cancers. However, it is one of the most aggressive breast cancers that does not exhibit the 3 common receptors (estrogen, progesterone and HER2) found in others. The treatments such as Lumpectomy, Mastectomy, and Radiation therapy inevitably cause physical damage to the human body. This research aims to look for a method of chemotherapy that minimizes the damages to normal epithelial cells and efficiently kills malignant mammalian cells.

In contrast to the drug discovery for covid-19, it requires high sensitivity because it is both limited by the size of databases and the side effects to normal cells. Random Forest (RF) and

Deep Neural Networks (DNN) are used to fit the model after examining the R^2 and ROC among RF, DNN, Partial least square (PLS), and Multiple linear regression (MLR) (Figure 11).

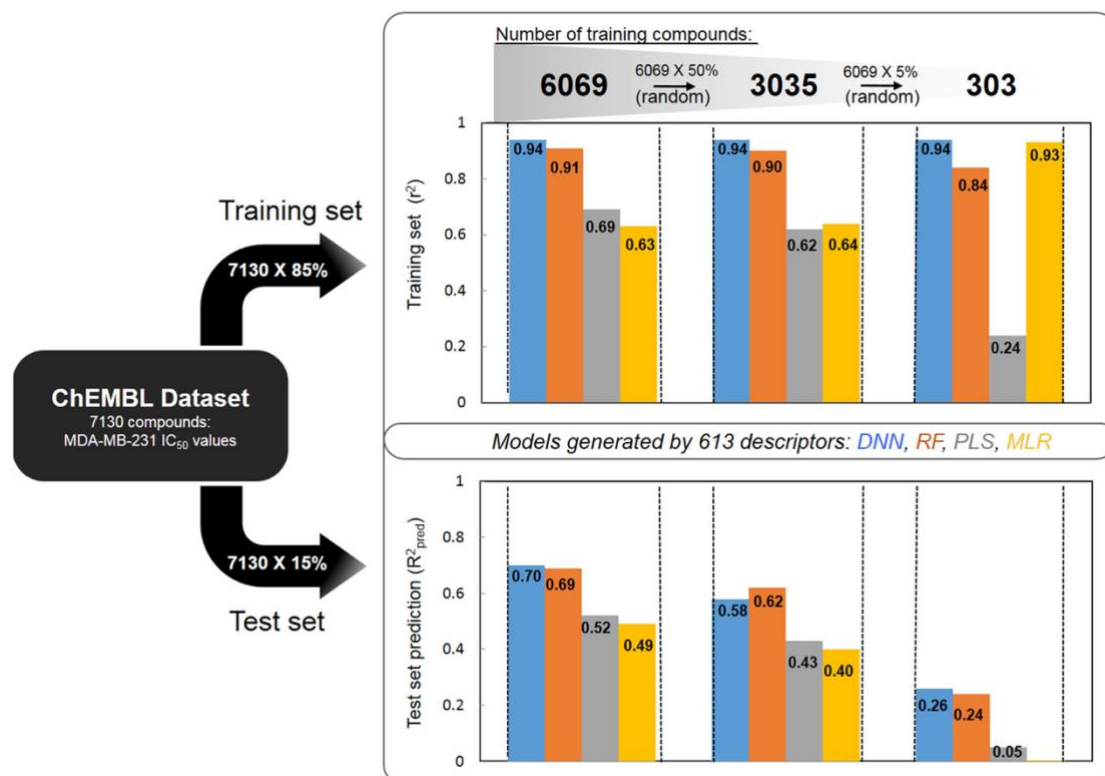


Figure 11: R^2 value for different models in Breast cancer research. (Adapted from Tsou *et al.*, 2020)

Models are generated from 7130 compounds with 613 descriptors that have the inhibitory effect on the unknown TNBC receptors.⁵ RF and DNN hold an R^2 value of 0.91 and 0.94, respectively, with extensive training sets. R^2 is consistent for DNN with a different number of training samples, and it slightly decreased to 0.84 for RF as data sets shrink to 303. PLS and MLR both have an unsatisfactory performance against the size change of the dataset.

DNN overrides other algorithms with high performance at all training sets and a large test set. The reason behind is its extraordinary multi-layer architecture (Figure 12) that simulates the information pathway of the nervous system in humans.

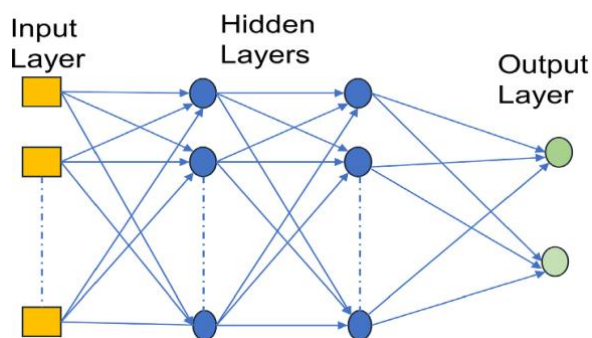


Figure 12: Architecture of multi-layer neural networks. (Adapted from Dara *et al.*, 2021)

The descriptors of a compound are fed into the input layer and always passed to the next layer until an 'active' or 'inactive' label is generated. The neurons are cross-linked by nodes so that DNN can consider the effect of all descriptors simultaneously at every hidden layer. It enables different weights to be added on the descriptors, subject to the fact that some descriptors need to be prioritized to enhance the accuracy. The encoder-decoder model of DNN computes descriptors with different weights as if it extracts the meaning of an English sentence and paraphrases it. This unique function of DNN abandons misleading combinations of features so that it lowers the noise level of the inputs and correspondingly lowers the error in the output.¹

The top 100 predicted compounds with TNBC cell lines (MDA-MB-231) inhibition are selected from 165000 compounds.⁵ Then it undergoes bioassay to test the survival rate of other cell lines (MCF10A, BT-549, and MDA-MB-453) at 10 μ M (Figure 13C). Twelve compounds (1-6 from RF and 7-12 from DNN) are selected with low cytotoxicity to the normal mammalian epithelial cell lines MCF10A. It is noticed that only compounds 3,7,8,10 show direct inhibitory effects to MDA-MB-231 by IC₅₀ values and weak binding to other malignant cell lines (BT-549 and MDA-MB-453). It is proved that DNN is more adaptive to a larger database than RF. Thiazole is a structural alert, as shown in Figure 13B. This finding provides valuable information for the subsequent optimization. Compounds 15-23 are then modified based on a thiazole core (Figure 14), most of which presents potent inhibition towards MDA-MB-231 and mild inhibition towards MCF10A.

The result is feedback to the original 7130 compounds for validation. After screening, it is observed that the thiazole-based compounds are clustered within the range of IC_{50} 10 μ M~0.3 μ M. Other properties such as partition coefficient and polar surface area versus total surface area are acceptable.

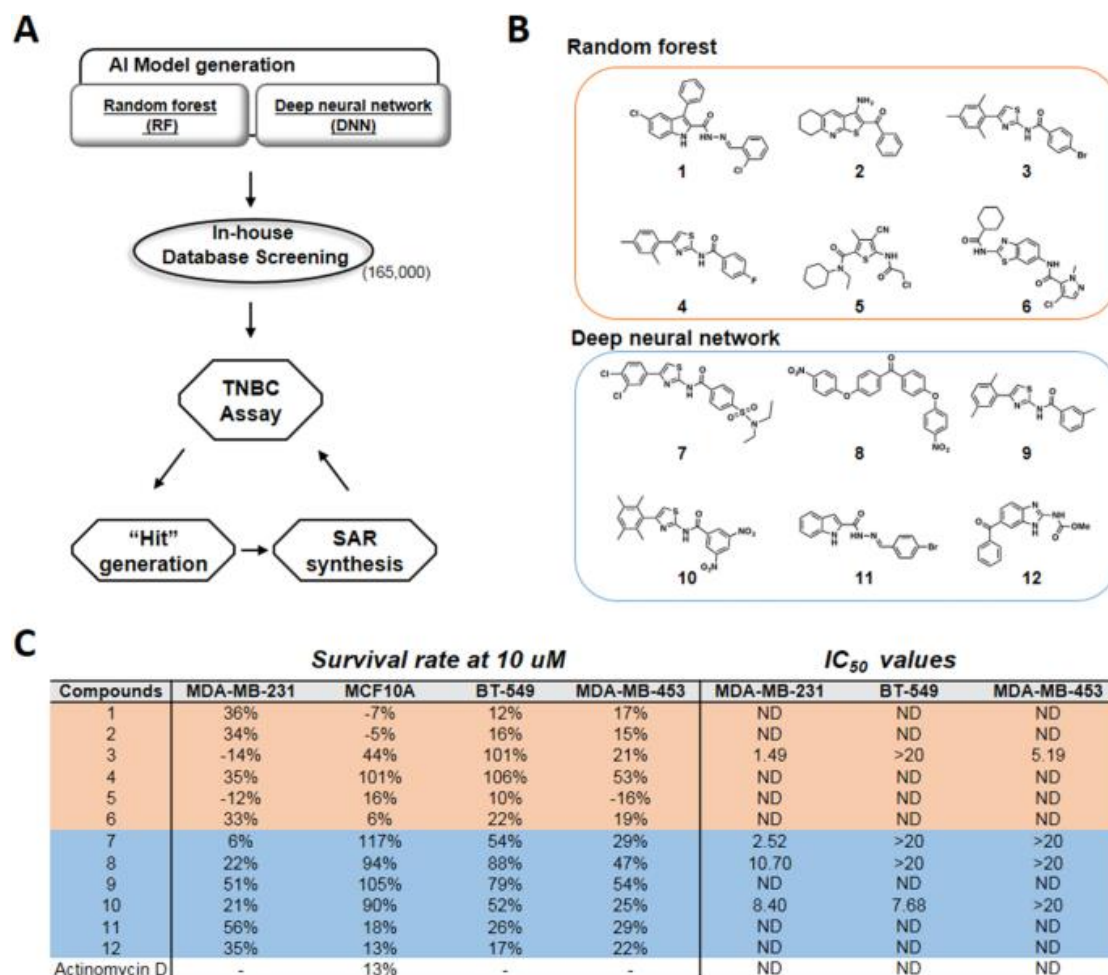
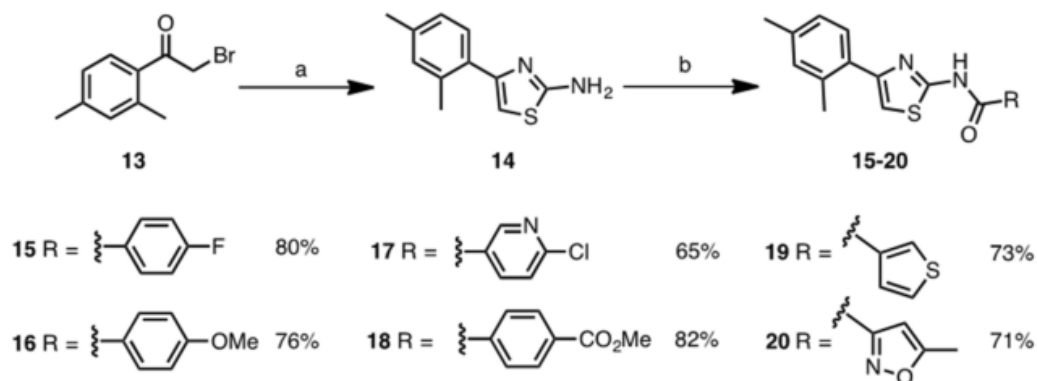
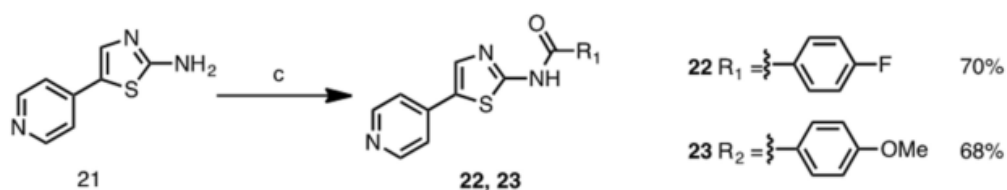


Figure 13: **A)** The overview of TNBC inhibitor QSAR. **B)** Structure of predicted compounds 1-12. **C)** Bioassay of compounds 1-12 with actinomycin D as a control. (Adapted from Tsou *et al.*, 2020)

A**i****ii**

Reagents and conditions: (a) thiourea, EtOH, reflux, 0.5 h (b) T₃P, Et₃N, EA, rt, 2h (c) T₃P, Et₃N, EA, rt, 2h

B

	MCF-10A	BT-549	MDA-MB-231	MDA-MB-453
Compounds	Inhibition rate	IC ₅₀ (μ M)	IC ₅₀ (μ M)	IC ₅₀ (μ M)
15	-8%	>10.00	1.89	9.28
16	8%	>10.00	1.4	6.3
17	2%	>10.00	3.41	>10.00
18	4%	>10.00	0.62	>10.00
19	14%	>10.00	2.68	>10.00
20	14%	>10.00	>10.00	>10.00
22	-5%	>10.00	>10.00	>10.00
23	35%	2.6	3.98	7.93

Figure 14: **A)** Synthesis of compounds 15-23 from the available starting material. **B)** Bioassay of compounds 15-23. (Adapted from Tsou *et al.*, 2020)

Conclusion

The case study of Covid-19 and Triple-negative breast cancer has shown a huge potential of machine-learning-based QSAR in identifying and optimizing hits in drug discovery. This is even more important because the alternative receptor-based method for hit optimization such as molecular simulations cannot always satisfy all the needs of drug design due to the

prohibitive computational time scales needed for large molecules such as proteins and peptides. Indeed, the direct calculation of forcefield seems to prioritize the binding affinity, and it is difficult to evaluate the total inhibitory effect. QSAR is not adversely affected by the length of molecules and the characteristics of receptors, as the computational time only depends on the size of databases and the number of descriptors. Although QSAR sometimes provides results to people who lack the knowledge for the detailed mechanism and the mode of drugs (cytotoxic or cytostatic), the traditional QSAR methods with supervised learning (SVM, RF) are enough for the well-known diseases. Core structures are usually generated to optimize hits in a particular direction. Some HIV-1 protease inhibitors are repurposed into 3CLpro inhibitors. Thiazole rings are the functional fragments against TNBC. The mode of action may be deduced from well-known drugs, such as the functional cyano group of Remdesivir. Predicted drugs for covid-19 such as Diltiazem hydrochloride and Leflunomide purpose different modes of action to treat comorbidities. It indicates that QSAR can provide mixed solutions to the associated syndrome in patients. While the traditional QSAR methods are still less able to follow the new disease or those that are not yet understood, such as TNBC, deep neural networks become one way to deal with the new challenges as they can extract the proper features of compounds. The potential for improvements of QSAR with ML is unparalleled because machine learning algorithms are mathematical models that are designed to adapt to different experimental environments. As we learned from the ongoing pandemic, the world needs a flexible tool to quickly develop against fast mutating viruses to ensure human lives are not threatened, and QSAR-ML is one of the viable options to provide this tool.

Reference

- 1 S. Hu, P. Chen, P. Gu and B. Wang, *IEEE J. Biomed. Health Inform.*, 2020, **24**, 3020–3028.
- 2 J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer and S. Zhao, *Nat Rev Drug Discov*, 2019, **18**, 463–477.
- 3 R. Todeschini, D. Ballabio and F. Grisoni, *J. Chem. Inf. Model.*, 2016, **56**, 1905–1913.
- 4 W. Van Valkenburg, Ed., *Biological Correlations—The Hansch Approach*, AMERICAN CHEMICAL SOCIETY, WASHINGTON, D. C., 1974, vol. 114.
- 5 L. K. Tsou, S.-H. Yeh, S.-H. Ueng, C.-P. Chang, J.-S. Song, M.-H. Wu, H.-F. Chang, S.-R. Chen, C. Shih, C.-T. Chen and Y.-Y. Ke, *Sci Rep*, 2020, **10**, 16771.
- 6 J. Mao, J. Akhtar, X. Zhang, L. Sun, S. Guan, X. Li, G. Chen, J. Liu, H.-N. Jeon, M. S. Kim, K. T. No and G. Wang, *iScience*, 2021, **24**, 103052.
- 7 Y.-C. Lo, S. E. Rensi, W. Torng and R. B. Altman, *Drug Discovery Today*, 2018, **23**, 1538–1546.
- 8 S. Dara, S. Dhamercherla, S. S. Jadav, C. M. Babu and M. J. Ahsan, *Artif Intell Rev*, 2021, DOI:10.1007/s10462-021-10058-4.
- 9 G. Kokic, H. S. Hillen, D. Tegunov, C. Dienemann, F. Seitz, J. Schmitzova, L. Farnung, A. Siewert, C. Höbartner and P. Cramer, *Nat Commun*, 2021, **12**, 279.
- 10 V. Consonni, R. Todeschini, D. Ballabio and F. Grisoni, *Mol. Inf.*, 2019, **38**, 1800029.

- 11 J. Ivanov, D. Polshakov, J. Kato-Weinstein, Q. Zhou, Y. Li, R. Granet, L. Garner, Y. Deng, C. Liu, D. Albaiu, J. Wilson and C. Aultman, *ACS Omega*, 2020, **5**, 27344–27358.
- 12 C. J. Gordon, E. P. Tchesnokov, E. Woolner, J. K. Perry, J. Y. Feng, D. P. Porter and M. Götte, *Journal of Biological Chemistry*, 2020, **295**, 6785–6797.