JPPIPA 8(6) (2022)

Jurnal Penelitian Pendidikan IPA

Journal of Research in Science Education

http://jppipa.unram.ac.id/index.php/jppipa/index

Dimensional Reduction of QSAR Features Using a Machine Learning Approach on the SARS-Cov-2 Inhibitor Database

Azizah Munaya[1], Arry Yanuar[1*], Firdayani[2]

[1]Department of Biomedical Computation and Drug Design Laboratory, Faculty of Pharmacy, Universitas Indonesia, Kampus UI, Depok 16424, West Java, Indonesia.
[2]Research Center for Vaccine and Drugs, Research Organization for Health, National Research and Innovation Agency, West Java, Indonesia.

Received: October 30, 2022
Revised: December 25, 2022
Accepted: December 30, 2022
Published: December 31, 2022

Corresponding Author:
Arry Yanuar
arry.yanuar@ui.ac.id

© 2022 The Authors. This open access article is distributed under a (CC-BY License)

DOI: 10.29303/jppipa.v8i6.2432

Abstract: Quantitative Structure-Activity Relationship (QSAR) is a method that relates the chemical composition of a molecule to its biochemical, pharmaceutical and biological activities. The characteristics of a molecule's chemical constituents, such as chemical descriptors and fingerprints, are necessary to create a good QSAR model. Dimensionality reduction can alleviate the issue of several unnecessary and redundant chemical descriptors and chemical fingerprints in a high-dimensional feature-number data set by shrinking the high-dimensional original space to a low-dimensional intrinsic space. There are two categories of dimensional reduction techniques: feature extraction and feature selection. The dimension reduction approach can be utilized as a starting step in running a QSAR Virtual Screening Model on a dataset of SARS-CoV-2 inhibitor medications to create novel treatments for Covid-19 cases based on machine learning (ML) and the idea of medicinal repurposing. Fe extraction and feature selection are crucial to determining which feature sets should be applied to a specific classification process in QSAR modeling to produce reliable virtual screening results. The SARS-Cov-2 inhibitor drug database's chemical descriptor and chemical fingerprint were extracted using a simple, quick, and accurate method in this work. The total number of selected features is 12122 features. PCA, Missing values, and Random Forest are the techniques employed. The Xgboost Tree Ensemble, Naive Bayes, Support Vector Machine, Random Forest, and Deep Learning (Artificial Neural Network/Multilayer Perceptron) were used to classify the QSAR modeling on the training and test data. The Random Forest approach, when applied to all chemical descriptors and chemical fingerprint features, along with the XGBoost algorithm, yields the best feature selection results (accuracy value of 0.845 and AUC of 0.904). There are 233 characteristics for the regression QSAR approach and 273 features for the feature selection-based QSAR method of classification. Next, virtual screening of QSAR modeling of prospective drugs for Covid-19 therapy can be done utilizing the outcomes of the characteristics that have been chosen using the Random Forest approach.

Keywords: QSAR; PCA; Missing values; Random forest; SARS-Cov-2

Introduction

Quantitative structure-activity relationship (QSAR) is a method that compares the chemical composition of a molecule to its biochemical, pharmacological, and biological activities (Bastikar et al., 2022) and can be used to identify compounds with enhanced biological activity (Ishola et al., 2021). According to a report by the Eastern Research Group (ERG), developing a novel molecular entity can take between 10 and 15 years, with a success rate of only 2.01% (Xue et al., 2018) and a cost that can approach $3 billion USD (DiMasi et al., 2016). The QSAR model can assist the development of new compounds by saving costs and time in terms of synthesis and manufacture of molecules as well as in vitro and in vivo molecular testing using the concept of drug repurposing (drug reuse) (Bender et al., 2021) use computer Aided Drug Design (CADD) (Paul et al., 2021) through Artificial Intelligence-based virtual screening (Cavasotto et al., 2021).

The characteristics of a molecule's chemical components is one of the requirements for developing a

How to Cite:
Azizah, M., Yanuar, A., & Firdayani, F. (2022). Dimensional Reduction of QSAR Features Using a Machine Learning Approach on the SARS-Cov-2 Inhibitor Database. *Jurnal Penelitian Pendidikan IPA*, *8*(6), 3095–3101. https://doi.org/10.29303/jppipa.v8i6.2432

reliable QSAR model. This attribute helps establish the distinctions between compounds that give each its own properties. The QSAR model uses chemical descriptors and chemical fingerprints as features. Chemical descriptors are phrases that characterize certain information about the examined molecule. Utilization of physicochemical properties Because the chemical descriptor provides a bridge between the molecular structure and the biological activity of the molecule, its nature has a substantial impact on the interpretation of the QSAR model (Roy et al., 2015). Chemical fingerprints are characteristics that represent a compound's substructure. Typically, a collection of substructures is organized in hashtable format. The fingerprint is represented in binary bit strings, which represent the distinctive properties of each molecule. Each bit in the fingerprint reflects a feature that is unavailable (0) or available (1) (Jasial et al., 2016).

In general, several irrelevant and redundant characteristics in chemical descriptors and chemical fingerprints raise the difficulty of data processing, knowledge mining, and pattern categorization. As a crucial solution to this issue, dimensionality reduction can eliminate noise and information overload by transforming the original high-dimensional space into the intrinsic low-dimensional space (M. Li et al., 2020). In general, dimensional reduction techniques can be categorized into two distinct categories: feature extraction and feature selection. Principal Component Analysis (PCA), Multi-Dimensional Scaling (MDS), Isometric Mapping (ISOMAP), and Local Linear Embedding (LLE) are dimensionality reduction techniques based on feature extraction. The feature selection approach ranks the original features based on predetermined criteria and chooses the highest-ranked features to generate a subset. There are three primary feature selection models: filter, wrapper, and embedding (M. Li et al., 2020). Random Forest is a generalized machine learning method focusing on feature selection utilizing ensemble methods like bagging (Jain et al., 2021).

The feature extraction and selection methodology can be utilized as a first step in executing a QSAR Virtual Screening Model on a dataset of SARS-CoV-2 inhibitor compounds to produce novel medications in machine learning (ML)-based Covid-19 scenarios using the notion of drug repurposing. As of October 27, 2022, the World Health Organization (WHO) stated that 24 countries were witnessing a surge in cases, with 626,090,028 confirmed positive cases of Covid-19 and 6,564,556 deaths worldwide (WHO, 2022). Thus, the initial stage in feature extraction and selection is crucial for determining which feature set should be employed in a specific classification process in QSAR modeling to produce correct virtual screening results (Mendes Junior et al., 2020).

Several research on the reduced dimensions of SARS-Cov-2 inhibitor drugs has been undertaken. Extraction and feature selection performed by García et al. (2021) and Erlina et al. (2020) necessitates advanced programming skills, specifically using the paDel descriptor, Python, and protr R, Research (Rajput et al., 2021). Only the extraction and selection of chemical descriptors were performed. In this study, chemical descriptors and chemical fingerprints from the SARS-Cov-2 inhibitor drug database were extracted using a straightforward, rapid, and accurate method. Utilized techniques include PCA, Missing values, and Random Forest. The training and test data were then classified using the Xgboost Tree Ensemble, Naive Bayes, Support Vector Machine, Random Forest, and Deep Learning (Artificial Neural Network/Multilayer Perceptron) so that predictions of the activity of various molecules in large databases can be made more quickly than with the virtual method. Other presentations Using the KNIME Analysis Platform, the Artificial Intelligence application utilized in this work is simple to implement in academia and the drug raw material sector. An open-source platform that allows for the flexible creation of workflows without the need for significant programming abilities, hence simplifying and reducing analysis time (Jain et al., 2021).

Method

Virtual feature extraction and selection screening QSAR modeling with a Machine Learning approach is applied as the method (Figure 1).

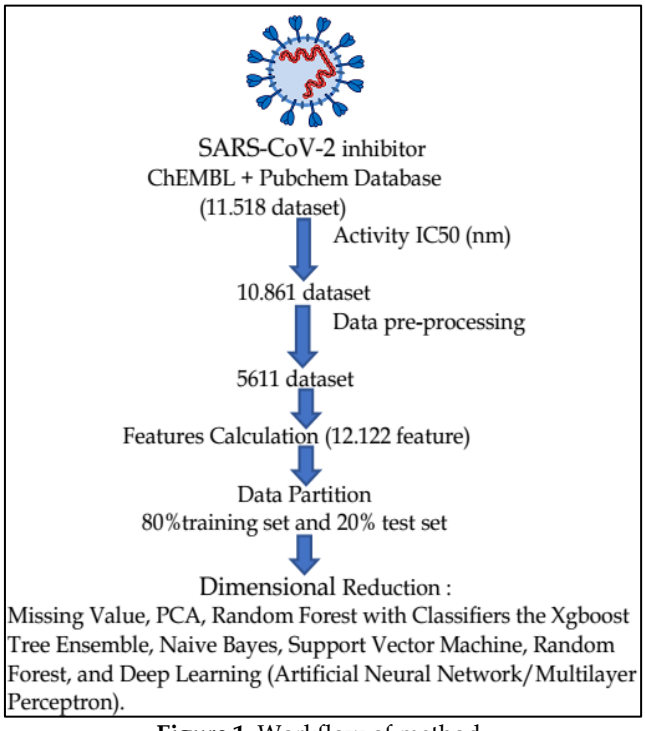

Figure 1. Workflow of method

Dataset

The data used in this study is a SARS-CoV-2 inhibitor compound molecule which was downloaded from the ChEMBL chemical molecular bioactivity database site on March 11, 2022, in CSV format, which can be accessed via (https://www.ebi.ac.uk/ chembl) and the PubChem website on October 7, 2022, in CSV format accessible via (https://pubchem.ncbi.nlm.nih.gov/#query=covid-19&tab=bioassay).

Data Pre-processing

The dataset was screened for IC50 activity; then, the data set was curated with the activity unit being nM and an empty activity value (missing value). Then standardization was carried out in canonical SMILES, the duplicate compounds and salt were removed (using the RDKit salt stripper node), and the presence of free hydrogen (using the hydrogen remover node).

Calculation of QSAR Features

After data preparation, the next step is the calculation of the chemical descriptors and chemical fingerprint features (Figure 2).

Figure 2. Workflow of features extraction

Chemical descriptors: The nodes used are RDKit Descriptor Calculation and molecular properties descriptor.

Chemical fingerprint: The nodes used are RDKit and CDK fingerprint. RDKit fingerprint consists of fingerprint MACCS, Morgan, Avalon, Feat Morgan, Atom pair, Rdkit, Torsion, and Layered; As for the CDK, fingerprint consists of standard, Extended, Estate, Pubchem, MACCS, and Circular fingerprints.

Dataset Transform

The inhibition activity of the compound dataset is presented as data on its biological activity IC50 (nM), which is then converted into the PIC50 value using equation (1) below:

$$PIC_{50} = (9 - \text{Log } IC_{50}) \quad (1)$$

Feature Extraction and Selection

Feature selection uses 3 (three) methods, namely PCA, dividing into 9 (nine) principal components, the Missing Values method, and the Random Forest method. The features of the selected compounds are classified into 8 (eight) categories, as shown in Table 1.

Table 1. The Group of Feature Selection

The Group of Feature Selection
RDKit Descriptor
Descriptor molecular properties
RDKit fingerprint
CDK fingerprint
Combination of RDKit Descriptor and Descriptor molecular properties
Combination of RDKit Descriptor and RDKit fingerprint
Combination of RDKit Descriptor and CDK fingerprint
Combination of RDKit Descriptor, Descriptor molecular properties RDKit fingerprint, and CDK fingerprint

Table 2. Model QSAR

	Naive Bayes	XGBoost	Random Forest	ANN
RDKit Descriptor	Des-NB	Des-XGB	Des-RF	Des-ANN
Descriptor molecular properties	MP-NB	MP-XGB	MP-RF	MP-ANN
RDKit fingerprint	FP-NB	FP-XGB	FP-RF	FP-ANN
CDK fingerprint	CDK-NB	CDK-XGB	CDK-RF	CDK-ANN
Combination of RDKit Descriptor and Descriptor molecular properties	DesMP-NB	DesMP-XGB	DesMP-RF	DesMP-ANN
Combination of RDKit Descriptor and RDKit fingerprint	DesFP-NB	DesFP-XGB	DesFP-RF	DesFP-ANN
Combination of RDKit Descriptor and CDK fingerprint	RDKCDK-NB	RDKCDK- XGB	RDKCDK-RF	RDKCDK-ANN
Combination of RDKit Descriptor, Descriptor molecular properties RDKit fingerprint, and CDK fingerprint	ALL-NB	ALL-XGB	ALL-RF	ALL-ANN

These features were tested on various machine learning methods with supervised machine learning classification algorithms, such as Xgboost Tree Ensemble, Naive Bayes, Support Vector Machine, Random Forest, and Deep Learning (Artificial Neural Network/Multilayer Perceptron), as shown in table 2.

Data Partition

The SARS-CoV-2 inhibitor compound dataset with Chemical descriptors and Chemical fingerprint features was partitioned into two parts with a ratio of 80% as a training set and 20% as a test set.

Evaluation of Feature Selection Results

All models were analyzed with statistical parameters with the concept of confusion matrix, mainly accuracy and Plot Receiver Operating Characteristic (ROC).

Result and Discussion

The dataset of SARS-CoV-2 inhibitor compounds generated from the ChEMBL database consisted of 10,465 compounds; after screening for biological activity, the number decreased to 9.808. According to the PubChem database, 1053 biologically active chemicals were obtained. Therefore, the total number of biologically active chemicals in the dataset was 10.861. The dataset is then curated and standardised by removing duplicate, salt, and free hydrogen components, obtaining 5611 compounds which are eligible for extraction and feature selection (Figure 3).

The feature calculations yields 118 chemical descriptors, 42 molecular properties descriptors, 7735 RDKit fingerprint features, and 4217 CDK fingerprint features, bringing the overall number of features to 12122. The obtained range of PIC50 values is between 0 and 9.046 (Figure 3.). The pIC50 value of biological activity offers better precision than the IC50 value (Attiq et al., 2022). In addition, the data were classified into active and inactive compounds based on the pIC50 limit value, with pIC50 ≤ 4.698 suggesting inactive compounds and > 5.25 suggesting active compounds. This number was chosen based on the range of observed pIC50 values for the full data set in order to maximize the chemical space representation for each structure class (active and inactive). The structures with pIC50 values between 5.250 and 4.698 (range 0.55 units) were omitted to prevent edge effects and improve the prediction power of the model by limiting potential changes in activity caused by experimental errors and procedures (Janeiro, 2018).

Three extraction and feature selection techniques were selected because they represent a variety of types, especially extraction type (PCA), filter type (missing value), and bagging type (Random Forest). PCA was chosen because it is a straightforward nonparametric method for extracting a large number of important features from a set of redundant or noisy data. PCA compute, which calculates the components of the reduced variable, and PCA Apply, which displays the PCA group, which is the projected variable from the original column reduction, are the nodes utilized by the PCA technique. The missing values method was selected because it is simple, quick, and effective in filtering out missing data. While the Random Forest Method was chosen because it uses various possible Decision Trees that segment features randomly, initially with the widest segmentation range and reducing as it approaches, so that the results seem to be more specific.

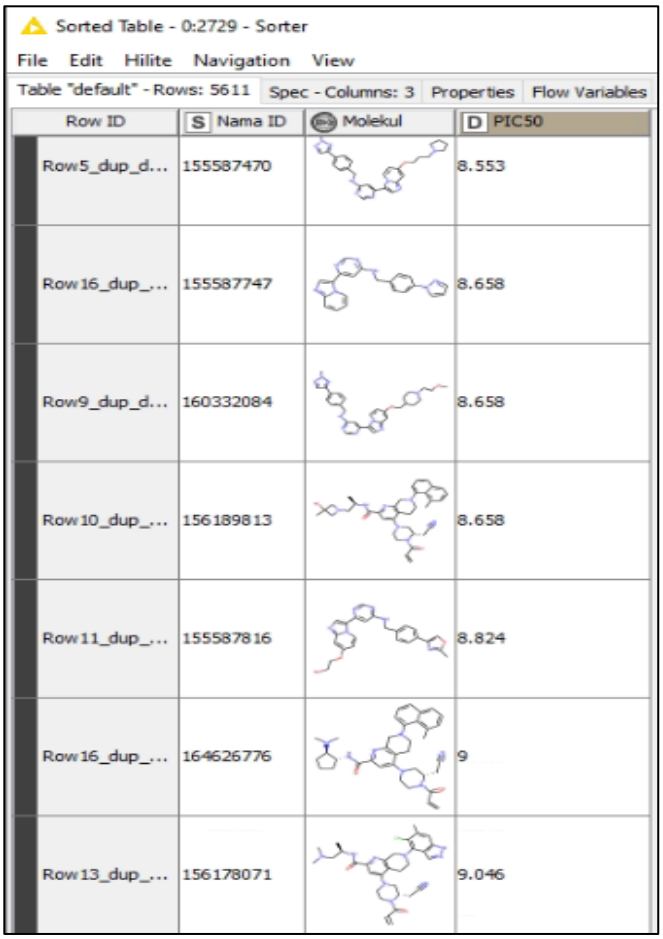

Figure 3. Pre-processing result data

Each of these methods was evaluated within five classification-based supervised learning algorithms. The methods with the highest accuracy were the Missing Values method for the chemical descriptor feature with the ANN algorithm (0.961), the Missing Values method for the chemical descriptor feature with the XGBoost algorithm (0.948), and the Random Forest method with the CDK fingerprint feature with the XGBoost algorithm (0.858). Combining all descriptor and fingerprint characteristics, the XGBoost algorithm (0.845).

Table 3. Accuracy Value

	Naive Bayes	XGBoost	Support Vector Machine	ANN
RDKit Descriptor				
Missing Value	0.781	0.948	0.553	0.961
PCA	0.480	0.630	0.566	0.742
Random Forest	0.520	0.605	0.605	0.703
Molecular properties Descriptor				
Missing Value	0.553	0.643	0.540	0.631
PCA	0.536	0.596	0.738	0.631
Random Forest	0.476	0.451	0.549	0.558
RDKit fingerprint				
Missing Value	0.712	0.811	0.566	0.785
PCA	0.562	0.579	0.527	0.704
Random Forest	0.527	0.785	0.785	0.753
CDK fingerprint				
Missing Value	0.708	0.841	0.566	0.753
PCA	0.605	0.618	0.549	0.742
Random Forest	0.755	0.858	0.553	0.776
RDKit Descriptor and Molecular Properties Descriptor				
Missing Value	0.690	0.832	0.562	0.729
PCA	0.605	0.832	0.562	0.729
Random Forest	0.643	0.818	0.553	0.781
RDKit Descriptor and RDKit fingerprint				
Missing Value	0.708	0.811	0.566	0.759
PCA	0.562	0.579	0.527	0.708
Random Forest	0.665	0.785	0.532	0.742
RDKit Descriptor and CDK fingerprint				
Missing Value	0.649	0.819	0.566	0.789
PCA	0.618	0.618	0.549	0.682
Random Forest	0.712	0.739	0.533	0.799
Combination of RDKit Descriptor, Descriptor molecular properties RDKit fingerprint, and CDK fingerprint				
Missing Value	0.626	0.794	0.566	0.785
PCA	0.635	0.485	0.605	0.738
Random Forest	0.515	0.845	0.536	0.716

The greatest accuracy findings suggest that the missing value and random forest feature selection method is a feasible method that can generate accurate features for the virtual screening of QSAR models. However, the missing value method's accuracy is only high for one feature, the RxDKit descriptor and CDK fingerprint. As for Random forest, the combination of chemical descriptors and chemical fingerprint features has a good accuracy value. Thus, the Random Forest technique is the best for all features.

The PCA method implies a linear relationship between variables that will reduce the prediction accuracy of the QSAR model (P. Li et al., 2022) and PCA only retains the main components from an informatics perspective, but the main chemical features related to QSAR is ignored causing overfitting and low accuracy (J. Li et al., 2021).

The graph of the accuracy results in the 3 (three) methods is shown in Figure 4.

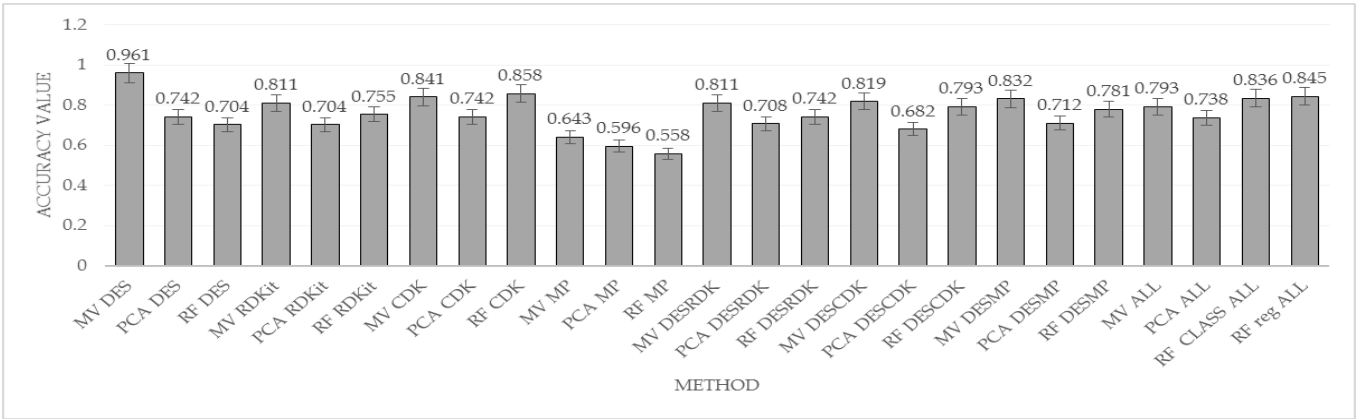

Figure 4. The graph of accuracy results

Furthermore, the largest Area Under Curve (AUC) of the ROC analysis was obtained for the missing values method for the chemical descriptor feature (0.987), the Random Forest CDK fingerprint method (0.915), and the Random Forest method combined with all chemical descriptor features and chemical fingerprints (0.915) and (0.904) (Figure 5). The curve demonstrates that AUC results are linear with accuracy outcomes, with Random Forest being the optimal method for all characteristics.

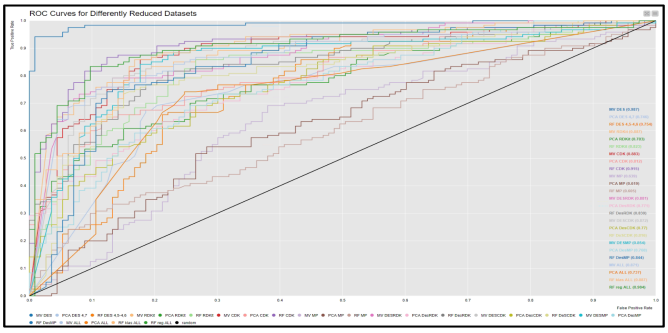

Figure 5. ROC Curve of feature selection

The feature selection of the random forest method is adapted from Hermansyah et al. (2021), in which training and validation data and test results are used to identify essential features with the parameters 2000 and 10 for tree depth. The QSAR method of classification generated from feature selection obtains 273 features consists of RDKit fingerprints (Morgan, Feat Morgan, Torsion, Atom pair, Avalon, Layered and MACCS) anad CDK fingerprints (Standard, Extended, Pubchem, MACCS, and Circular). The QSAR method of regression obtains 233 features consists of Molecular Quantum Numbers (MQN), RDKit fingerpints (Morgan, Feat Morgan, Torsion, Atom pair, Avalon, Layered and MACCS), CDK fingerprints (Standard, Extended, Pubchem, MACCS, and Circular). Most of the features in the classification method are hashed fingerprints where certain substructures are hashed into bit strings, these fingerprints are most useful in classification methods when used with molecules that are likely to be covered by a given structural key (Cereto-Massagué et al., 2015). As for the selected features in regression methods, in addition to hashed fingeprints, there are also MQN features that can describe the number of five- or six-membered rings, topological surface area, cyclic trivalent and tetravalent vertices, and vertices and edges shared by more than two rings (Batra et al., 2020).

However Zhang et al. (2022) reported that no fingerprints can outperform the others considering all targets and that different fingerprint types are effective on different targets and different fingerprints take different active compounds, and the combination of multiple fingerprints gives the best performance (Xie et al., 2020).

Conclusion

Dimension reduction methods, including feature extraction and feature selection, are essential steps to evaluate which feature sets should be used in the virtual screening process of QSAR modeling so that accurate virtual screening results will be obtained (Kabir et al., 2022). The best feature selection method obtained is the Random Forest method against a combination of all chemical descriptors and chemical fingerprints with the XGBoost algorithm (accuracy value of 0.845 and AUC value of 0.904). Then the results of the features that have been selected using the Random Forest method can be used for the next step, namely virtual screening of QSAR modeling of potential compounds for Covid-19 therapy.

Acknowledgements

The authors are grateful thanks to the Biomedical Computation and Drug Design Laboratory, Pharmacy Faculty, University of Indonesia. This study was funded by Ministry of Education, Culture, Research And Technology-Pascasarjana Grant NKB-893/UN2.RST/HKP.05.00/2022

References

Attiq, N., Arshad, U., Brogi, S., Shafiq, N., Imtiaz, F., Parveen, S., Rashid, M., & Noor, N. (2022). International Journal of Biological Macromolecules Exploring the anti-SARS-CoV-2 main protease potential of FDA approved marine drugs using integrated machine learning templates as predictive tools. *International Journal of Biological Macromolecules*, *220*(September), 1415–1428. https://doi.org/10.1016/j.ijbiomac.2022.09.086

Bastikar, V., Bastikar, A., & Gupta, P. (2022). Chapter 10 - Quantitative structure–activity relationship-based computational approaches. In A. Parihar, R. Khan, A. Kumar, A. K. Kaushik, & H. Gohel (Eds.), *Computational Approaches for Novel Therapeutic and Diagnostic Designing to Mitigate SARS-CoV-2 Infection* (pp. 191–205). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-323-91172-6.00001-7

Bender, A., & Cortés-Ciriano, I. (2021). Artificial intelligence in drug discovery: what is realistic, what are illusions? Part 1: Ways to make an impact, and why we are not there yet. *Drug Discovery Today*, *26*(2), 511–524. https://doi.org/10.1016/j.drudis.2020.12.009

Cavasotto, C. N., & Di Filippo, J. I. (2021). Artificial intelligence in the early stages of drug discovery. *Archives of Biochemistry and Biophysics*, *698*(November 2020), 108730. https://doi.org/10.1016/j.abb.2020.108730

DiMasi, J. A., Grabowski, H. G., & Hansen, R. W. (2016). Innovation in the pharmaceutical industry: New estimates of R&D costs. *Journal of Health Economics*, *47*, 20–33. https://doi.org/10.1016/j.jhealeco.2016.01.012

Erlina, L., Paramita, R. I., Kusuma, W. A., Fadilah, F., Tedjo, A., Pratomo, I. P., Ramadhanti, N. S., Nasution, A. K., Surado, F. K., Fitriawan, A., Istiadi, K. A., & Yanuar, A. (2020). *Virtual Screening on Indonesian Herbal Compounds as COVID-19 SupportiveTherapy: Machine Learning and Pharmacophore Modeling Approaches*. *May*. https://doi.org/10.21203/rs.3.rs-29119/v1

García, R., Hussain, A., Koduru, P., Atis, M., Wilson, K., Park, J. Y., Toby, I., Diwa, K., Vu, L., Ho, S., Adnan, F., Nguyen, A., Cox, A., Kirtek, T., García, P., Li, Y., Jones, H., Shi, G., Green, A., & Rosenbaum, D. (2021). Identification of potential antiviral compounds against SARS-CoV-2 structural and non structural protein targets : A pharmacoinformatics study of the CAS COVID-19 dataset. *Computers in Biology and Medicine*, *133*, 104364. https://doi.org/10.1016/j.compbiomed.2021.104364

Ishola, A. A., Adedirin, O., Joshi, T., & Chandra, S. (2021). QSAR modeling and pharmacoinformatics of SARS coronavirus 3C-like protease inhibitors. *Computers in Biology and Medicine*, *134*, 104483. https://doi.org/https://doi.org/10.1016/j.compbiomed.2021.104483

Jain, R., & Xu, W. (2021). RHDSI: A novel dimensionality reduction based algorithm on high dimensional feature selection with interactions. *Information Sciences*, *574*, 590–605. https://doi.org/https://doi.org/10.1016/j.ins.2021.06.096

Janeiro, R. De. (2018). *Classification Models Based on Machine Learning for The Prediction of mPGES-1 Inhibitor*. *309*, 7–8.

Jasial, S., Hu, Y., Vogt, M., & Bajorath, J. (2016). Activity-relevant similarity values for fingerprints and implications for similarity searching. *F1000Research*, *5*. https://doi.org/10.12688/f1000research.8357.2

Kabir, M. F., Chen, T., & Ludwig, S. A. (2022). A performance analysis of dimensionality reduction algorithms in machine learning models for cancer prediction. *Healthcare Analytics*, *3*, 100125. https://doi.org/https://doi.org/10.1016/j.health.2022.100125

Li, J., Luo, D., Wen, T., Liu, Q., & Mo, Z. (2021). Representative feature selection of molecular descriptors in QSAR modeling. *Journal of Molecular Structure*, *1244*, 131249. https://doi.org/https://doi.org/10.1016/j.molstruc.2021.131249

Li, M., Wang, H., Yang, L., Liang, Y., Shang, Z., & Wan, H. (2020). Fast hybrid dimensionality reduction method for classification based on feature selection and grouped feature extraction. *Expert Systems with Applications*, *150*, 113277. https://doi.org/https://doi.org/10.1016/j.eswa.2020.113277

Li, P., Zhang, W., Lu, C., Zhang, R., & Li, X. (2022). Robust kernel principal component analysis with optimal mean. *Neural Networks*, *152*, 347–352. https://doi.org/https://doi.org/10.1016/j.neunet.2022.05.005

Mendes Junior, J. J. A., Freitas, M. L. B., Siqueira, H. V, Lazzaretti, A. E., Pichorim, S. F., & Stevan, S. L. (2020). Feature selection and dimensionality reduction: An extensive comparison in hand gesture classification by sEMG in eight channels armband approach. *Biomedical Signal Processing and Control*, *59*, 101920. https://doi.org/https://doi.org/10.1016/j.bspc.2020.101920

Paul, D., Sanap, G., Shenoy, S., Kalyane, D., Kalia, K., & Tekade, R. K. (2021). Artificial intelligence in drug discovery and development. *Drug Discovery Today*, *26*(1), 80–93. https://doi.org/10.1016/j.drudis.2020.10.010

Rajput, A., Thakur, A., Mukhopadhyay, A., Kamboj, S., Rastogi, A., Gautam, S., Jassal, H., & Kumar, M. (2021). Prediction of repurposed drugs for Coronaviruses using artificial intelligence and machine learning. *Computational and Structural Biotechnology Journal*, *19*, 3133–3148. https://doi.org/10.1016/j.csbj.2021.05.037

Roy, K., Kar, S., & Das, R. N. (2015). QSAR/QSPR Modeling: Introduction. In *A Primer on QSAR/QSPR Modeling: Fundamental Concepts* (pp. 1–36). Springer International Publishing. https://doi.org/10.1007/978-3-319-17281-1_1

WHO. (2022). *https://covid19.who.int/WHO Coronavirus (COVID-19) Dashboard | WHO Coronavirus (COVID-19) Dashboard With Vaccination Data, 2022*. *Diakses pada tanggal 27 Oktober 2022*.

Xue, H., Li, J., Xie, H., & Wang, Y. (2018). Review of drug repositioning approaches and resources. *International Journal of Biological Sciences*, *14*(10), 1232–1244. https://doi.org/10.7150/ijbs.24612

Zhang, H., Zhang, T., Saravanan, K. M., Liao, L., Wu, H., Zhang, H., Zhang, H., Pan, Y., Wu, X., & Wei, Y. (2022). DeepBindBC: A practical deep learning method for identifying native-like protein-ligand complexes in virtual screening. *Methods*, *205*, 247–262. https://doi.org/https://doi.org/10.1016/j.ymeth.2022.07.009