

Investigation on low-performance tuned-regressor of inhibitory concentration targeting the SARS-CoV-2 polyprotein 1ab

Daniel Febrían Sengkey^{1,5,7}, Angelina Stevany Regina Masengi^{2,5}, Alwin Melkie Sambul^{1,5},
Trina Ekawati Tallei^{3,4,5}, Sherwin Reinaldo Unsratdianto Sompie^{1,6}

¹Department of Electrical Engineering, Faculty of Engineering, Universitas Sam Ratulangi, Manado, Indonesia

²Department of Pharmacology and Therapy, Faculty of Medicine, Universitas Sam Ratulangi, Manado, Indonesia

³Department of Biology, Faculty of Mathematics and Natural Science, Universitas Sam Ratulangi, Manado, Indonesia

⁴Department of Biology, Faculty of Medicine, Universitas Sam Ratulangi, Manado, Indonesia

⁵Biomolecular Laboratory, Universitas Sam Ratulangi, Manado, Indonesia

⁶Information and Communication Technology Academic Support Unit, Universitas Sam Ratulangi, Manado, Indonesia

⁷Directorate of Research, Development, and Innovation, Indonesia Artificial Intelligence Society, Jakarta, Indonesia

Article Info

Article history:

Received Oct 19, 2024

Revised Jun 12, 2025

Accepted Jul 10, 2025

Keywords:

Hyperparameter tuning

Inhibitory concentration 50

Murcko fragments

Quantitative structure-activity
relationship

SARS-CoV-2 polyprotein 1ab

ABSTRACT

Hyperparameter tuning is a key optimization strategy in machine learning (ML), often used with GridSearchCV to find optimal hyperparameter combinations. This study aimed to predict the half-maximal inhibitory concentration (IC₅₀) of small molecules targeting the SARS-CoV-2 replicase polyprotein 1ab (pp1ab) by optimizing three ML algorithms: histogram gradient boosting regressor (HGBR), light gradient boosting regressor (LGBR), and random forest regressor (RFR). Bioactivity data, including duplicates, were processed using three approaches: untreated, aggregation of quantitative bioactivity, and duplicate removal. Molecular features were encoded using twelve types of molecular fingerprints. To optimize the models, hyperparameter tuning with GridSearchCV was applied across a broad parameter space. The results showed that the performance of the models was inconsistent, despite comprehensive hyperparameter tuning. Further analysis showed that the distribution of Murcko fragments was uneven between the training and testing datasets. Key fragments were underrepresented in the testing phase, leading to a mismatch in model predictions. The study demonstrates that hyperparameter tuning alone may not be sufficient to achieve high predictive performance when the distribution of molecular fragments is unbalanced between training and testing datasets. Ensuring fragment diversity across datasets is crucial for improving model reliability in drug discovery applications.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Daniel Febrían Sengkey

Department of Electrical Engineering, Faculty of Engineering, Universitas Sam Ratulangi

Kampus Unsrat Street, Bahu, Manado 95115, Indonesia

Email: danielsengkey@unsrat.ac.id

1. INTRODUCTION

The COVID-19 pandemic was one of the forces that drove the surge in computer-aided drug discovery (CADD adoption). Related studies during this period target either the host target, such as the transmembrane protease serine 2 (TMPRSS2) [1], or the part of the virus, such as the 3-chymotrypsin-like protease (3CLpro) or the main protease (M^{pro}) [2]–[9]. Huang *et al.* [1] utilized molecular docking to examine the drugs with positively charged guanidinobenzoyl and/or aminidinobenzoyl groups to inhibit TMPRSS2 at the host. Molecular docking was also used to assess the potential to repurpose approved drugs,

such as quinoline [2] as well as isavuconazonium, α -LI, and pentagastrin [6] to inhibit the virus's main protease. While also targeting the main protease, natural products were assessed with molecular docking as alternative pharmacotherapy options [4], [8], [9].

The adoption of machine learning (ML) is a variation in CADD, known as machine learning-aided drug discovery (MLDD). Classification is a common task in MLDD, where the target of the classification uses either a known interaction, coded as a binary value [10]–[12] or categories based on the discretized half-maximal inhibitory concentration (IC_{50}) value [7], [13], [14]. Despite the discretization of the IC_{50} being a common approach as demonstrated in the mentioned studies, however, this approach is discouraged in general epidemiology studies due to the loss of information within the numeric variable [15]. Based on two meta-analyses, it is found that continuous, rather than discrete, measurements could improve validity and reliability [16]. For instance, Gao *et al.* [17] build regression models using random forest (RF), and support vector machine (SVM) with some optimization to predict the IC_{50} of the [1,2,3] triazolo [4,5-d] pyrimidine derivatives (1,2,3-TPD) to inhibit the replication of the MGC-803, the gastric cancer cell in humans. In contrast to the work in [13], [18] utilized SVM, artificial neural network (ANN), k nearest neighbor (KNN), and RF to build regression models for predicting the IC_{50} towards multiple hepatitis C virus (HCV) non-structural proteins. Similarly, the work of Fiat *et al.* [19] used random forest regression (RFR) and gradient boosting regression (GBR) to develop ML models to predict the IC_{50} , targeting the HCV genotype 1a (Isolate 1). Support vector regression (SVR) is used in predicting the inhibition of small molecules to beta-secretase 1 (BACE1), which is an enzyme related to Alzheimer's disease (AD) [20]. In another study, multiple linear regressions (MLR) was found as the best algorithm compared to SVR, classification and regression (CART), and ANN in predicting the compound binding free energy (BFE) towards the SARS-CoV-2 main protease [21].

In our previous study [22], we experimented with 42 ML regression algorithms to predict the IC_{50} of bioactive compounds, targeting the polyprotein 1ab (pp1ab) of the SARS-CoV-2, which comprises the virus's non-structural protein (NSP) 12 to NSP16 [23], [24]. The default hyperparameters were used without any tuning process involved. The features were derived from the compounds by using PubChem fingerprints. Out of the 42 experimented algorithms, three algorithms: RFR, light gradient boosting machine regression (LGBR), and histogram gradient boosting machine regression (HGBR) were found as the most stable for this combination based on the R^2 values. Hyperparameter tuning is a technique in ML that is used to optimize the model performance by tweaking the hyperparameters of the algorithm [25]–[28]. It is commonly used with GridSearchCV, which combines a large hyperparameter search space and cross-validation to obtain the optimal generalizable model for the algorithm. Therefore, in this study, we extended the experiment with these algorithms, which also fall into the ensemble tree-based category, and investigated the impacts of data distribution, especially the Murcko fragments of the compounds, on the model performance.

The rest of this article is organized as follows: in section 2, we present the dataset as well as the methods we used for data curation, treatments in pre-processing, model training, validation, and performance evaluation. Then, in section 3, we compare the performance between the treatments, as well as investigate the distribution of compound characteristics in training and testing datasets. Last, in section 4, this paper is concluded, and directions for future work are presented.

2. METHOD

The research methodology mainly follows the core activities of data science methodology, as shown in Figure 1 mainly consists of three parts. The preprocessing part is related to fashioning the compounds' bioactivity data for ML training. The pipeline part is where we use custom pipelines that feed into the hyperparameter tuning process. The pipelined approach will ensure no data leakage, hence guaranteeing that the model has never seen the data used for its performance evaluation. Last, in the result analysis and documentation part, the experiment results are analyzed and compared. Mainly, we used Python version 3.10 and scikit-learn [29] version 1.5.1 in the modeling and analysis phases.

2.1. Preprocessing

The data preparation phase begins with data acquisition, specifically inhibitory bioactivity data. By using the ChEMBL web service [30], we acquired, in total, 1,455 compounds with known IC_{50} to the SARS-CoV-2 pp1ab (ChEMBL4523582), heavily increased from our previous study [22]. In this dataset, compounds are represented in simplified molecular input line entry system (SMILES) format. The data cleaning also includes standardizing the SMILES notation of each compound and converting the IC_{50} to the respective negative logarithmic scale, pIC_{50} , hence narrowing the scale. Following the cleaning steps, we continue with treating the duplicates. In drug discovery experiments, different approaches and different laboratory settings might yield different IC_{50} values, despite the use of the same compound. In our

experiments, we tried several approaches to handle the duplicated data. First, we left them as is; second, we aggregated them by taking the average of the pIC_{50} value; and last, we dropped all duplicated compounds.

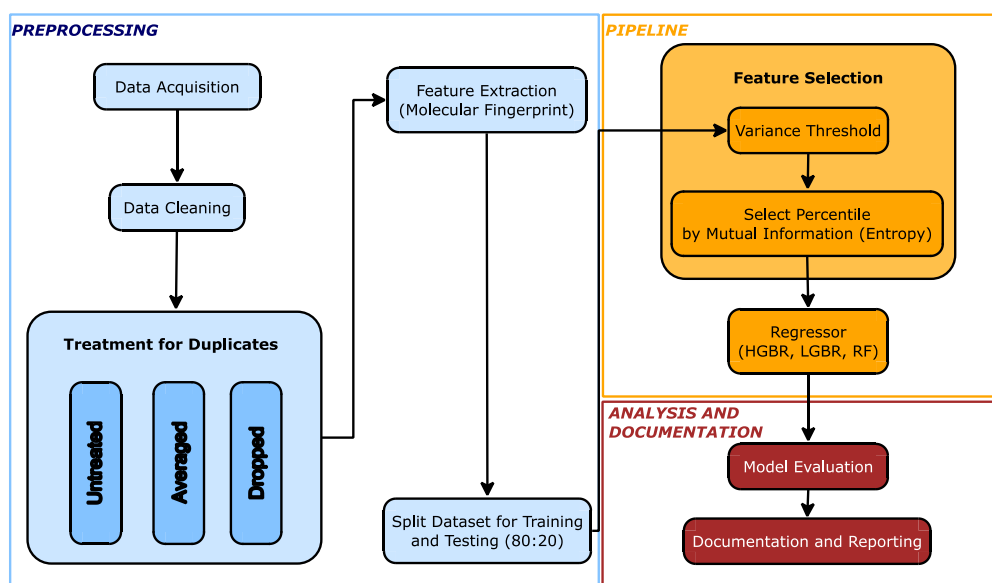


Figure 1. Course of research

After the duplicates were treated, we continued by transforming the chemical compounds (in SMILES) into molecular fingerprints (descriptors), resulting in a table for each fingerprint we used. The molecular fingerprints represent the characteristics of a chemical compound. For each compound, a fingerprint is a series of bits, where each bit is a Boolean, representing a specific chemical characteristic, and, as a whole, describes the compound. For instance, the PubChem fingerprint, the first bit shows whether the respective compound possesses four or more hydrogen atoms. The transformations to the fingerprints are done using PaDEL software [31]. In total, there are 12 variants of feature sets. The description of each fingerprint and the number of molecular features it has are provided in Table 1. Since 12 types of molecular fingerprints are in use and three treatments for duplicates, 36 datasets are used for the experiments. Then, using an 80:20 ratio of training and testing data, respectively, each dataset is split using the function available in scikit-learn.

2.2. Pipeline and hyperparameter tuning

To ensure the reliability and the continuity of model training and, later, utilize them for inferencing, the feature selection processes are coupled with the regressors as pipelines. The first feature selection method is the variance threshold. This feature selection method drops features with variance under the specified level. The rest of the features are then fed into the second feature selection method, the mutual information (entropy). We set the features selector to use only the top certain percentile, according to the features' entropy score. The post-feature selection dataset will then be used to train the regressor. As described earlier, three ML regression algorithms were explored alternately: HGBR, RFR, and LGBR.

As a development from our previous approach in [22], the current method employs hyperparameter tuning using GridSearchCV, to exhaustively test each combination of the hyperparameters in the search space. To ensure the generalizability of the hyperparameters with the best performance during training, 5-fold cross-validation is used. Table 1 lists the steps and modules in the pipelines, and the search space used for hyperparameter tuning.

2.3. Analysis and documentation

In this part of the research, we evaluate the performance of the models by comparing the performance of the trained models and applying it to infer the labels in the testing dataset. Performance metrics used are R^2 and the root mean squared error (RMSE). Statistical analyses and figures are done using the R statistical software version 4.4.1 [32].

Table 1. Hyperparameter tuning pipeline steps, module, and hyperparameter search space

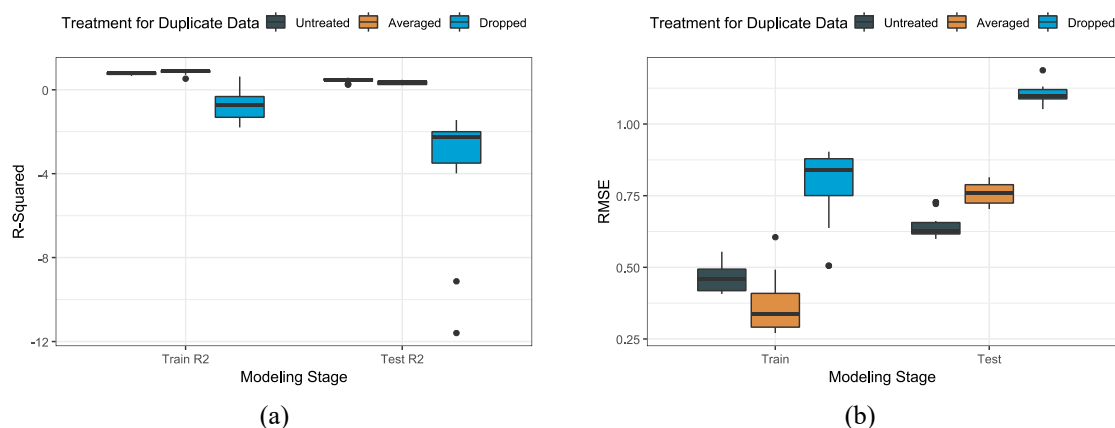
Pipeline step	Module	Hyperparameter search space
Feature selection	Variance threshold	threshold: 0.8(1-0.8)=0.16; 0.9(1-0.9)=0.09
	Select percentile (scoring by mutual information)	percentile: 10, 20, 50, 100
Regressor	HGBR	max_iter: [100, 1000, 10000, 99999999], max_depth: [None, 10, 20, 30, 40, 50], min_samples_leaf: [1, 2, 4, 8, 16, 32, 64], l2_regularization: [0, 0.1, 0.01], learning_rate: [0.01], warm_start: [True, False], early_stopping: [True], n_iter_no_change: [10, 100], random_state: [22],
	LGBR	boosting_type: ['gbdt','rf'], n_estimators: [99999999], max_depth: [-1, 15, 31, 63], learning_rate: [0.01], random_state: [22], num_leaves: [7, 31, 127, 1027, 2047, 4095]
	RFR	early_stopping_rounds: [20] n_estimators: [10, 100, 1000], min_samples_leaf: [1, 2, 4, 8, 16], max_depth: [None, 10, 20, 30, 40, 50], oob_score: [True, False], random_state: [22], warm_start: [True, False], min_samples_split: [2, 3, 4, 8, 16], max_features: ["sqrt", "log2", None]

3. RESULTS AND DISCUSSION

Using the best hyperparameters found for each combination of molecular fingerprints and algorithm for every treatment of duplicated data, we trained models and applied them to the testing dataset. Tables 2 to 4 show the best hyperparameters for HGBR, LGBR, and RFR algorithms, respectively, for each molecular fingerprint.

3.1. Performance metrics

Figure 2 shows the boxplots of the performance metrics, R^2 , and RMSE of the models with the hyperparameters that gave the best performance during the tuning with the 5-fold cross-validation step. It is obvious that by entirely dropping the duplicated bioactivity data, the models performed extremely differently from the other two treatments. When the duplicated data was dropped, the R^2 values dropped and commonly fell under zero with a higher variation, either during training or testing, as can be seen in Figure 2(a). Meanwhile, when these duplicates were left untouched, the R^2 during training was slightly lower than the averaged pIC₅₀ treatment, but the condition was reversed in testing. The boxplot of the loss function, RMSE, in Figure 2(b) indicates the same thing. Performance metrics for each combination of molecular fingerprint and algorithm with the best hyperparameters are shown in Figures 1 and 2.

Figure 2. Boxplots of performance metrics across treatments and modeling stages of (a) R^2 and (b) RMSE

Before statistically comparing the performance metrics, the Shapiro-Wilk test was applied to check the distribution normality of each performance metric. For this test, the data are grouped according to treatments, algorithms, and modeling stages. Therefore, a single distribution tested has 12 performance data. Table 2 shows the p-values of the Shapiro-Wilk test. With $\alpha=0.05$, it is clear that some of the data are not normally distributed, hence non-parametric test should be used for further analysis.

Table 2. P-values of the Shapiro-Wilk test for normality distribution of the performance metrics, grouped by the treatment for duplicates and algorithms. The italicized numbers are those under the $\alpha=0.05$

Treatment	Algorithm	Train R ²	Test R ²	Train RMSE	Test RMSE
Untreated	HGBR	0.038	0.167	0.083	0.124
Untreated	RFR	0.017	0.197	0.047	0.035
Untreated	LGBR	0.018	0.202	0.051	0.036
Averaged	HGBR	0.017	0.574	0.089	0.879
Averaged	RFR	0.012	0.475	0.047	0.842
Averaged	LGBR	0.012	0.360	0.048	0.681
Dropped	HGBR	0.001	<0.001	0.841	0.205
Dropped	RFR	0.255	0.537	0.210	0.649
Dropped	LGBR	0.282	0.629	0.132	0.812

We used the Friedman test for one-way repeated measures analysis of variance to compare each performance metric between the treatments with the same algorithm, by using the molecular fingerprint as the identifier. The results, as shown in Table 3, show that in all comparisons, at least one group of duplicate data treatment has a significantly different distribution of a particular performance metric. Following the one-way repeated measures Friedman test, we carried out the Pairwise Wilcoxon test to compare performance metrics between different treatments of the same algorithms. The Benjamini-Hochberg (BH) method is used for p-value adjustment. The results in Table 4 shows that in most cases, with $\alpha=0.05$, it can be seen that treatments for duplicate data significantly affect the performance. The R² during training with HGBR of the untreated and averaged treatments is the only comparison that is not significantly different. However, its counterpart in testing is significantly different.

Table 3. Results of the repeated measures Friedman test of the performance metrics between treatments

Algorithm	Metrics	n	F	Degree of freedom	p-value
HGBR	Train R ²	12	18.667	2	<0.001
RFR	Train R ²	12	22.167	2	<0.001
LGBR	Train R ²	12	22.167	2	<0.001
HGBR	Test R ²	12	20.667	2	<0.001
RFR	Test R ²	12	22.167	2	<0.001
LGBR	Test R ²	12	22.167	2	<0.001
HGBR	Train RMSE	12	19.500	2	<0.001
RFR	Train RMSE	12	24.000	2	<0.001
LGBR	Train RMSE	12	24.000	2	<0.001
HGBR	Test RMSE	12	24.000	2	<0.001
RFR	Test RMSE	12	24.000	2	<0.001
LGBR	Test RMSE	12	24.000	2	<0.001

3.2. Murcko fragments

In drug discovery, since different fragments lead to different bioactivity between the small molecules and the target, decomposing the compounds into fragments is a common task [33]. The Murcko fragments, proposed by Bemis and Murcko in 1996 [34], is a widely adopted technique, including in MLDD [35], [36]. The method works by ring systems, linkers, and the side chains of the molecules. The Murcko fragments consist of a combination of rings and linkers between them, with all terminal substituents removed. In this part, we compare the characteristics of the Murcko fragments between treatments and modeling stages to identify the cause of the low-performance metrics even after adopting hyperparameter tuning. The Murcko fragments are extracted from the compounds using the R chemistry development kit (RCDK) package version 3.8.1 [37]. The minimum fragment size used in the extraction is three. In total, 551 fragments can be identified from the bioactivity dataset. The fragments are numbered from F001 to F551 according to their frequencies in the dataset. Out of the 551 fragments, 12 with the highest frequencies were selected for further analysis.

In regards to pIC₅₀ as the regression target and the nature of the Murcko fragments as a fragment that appears in related compounds, which in turn affects the compounds' characteristics, then their molecular fingerprints which are used as features for the regression algorithms, imply that compounds with the same Murcko fragment should have similar pIC₅₀. Figure 3 shows the distributions of the pIC₅₀ of the selected

Murcko fragments for training and testing in all three treatments. From the 12 sampled Murcko fragments, it can be seen from Figure 3 some fragments have different pIC_{50} distributions, so the trend is more pronounced when the duplicate bioactivity data are dropped. For instance, the Murcko fragments F001, F002, F003, and F005 have different pIC_{50} distributions. Still in the dropped row, since it has fewer data, there are cases where certain Murcko fragments only exist in either dataset, such as happened with F010 and F011. Despite the Murcko fragment F010 also only appearing in one of two datasets in the averaged treatment, it can be seen that the boxplots in the respective row have similar pIC_{50} distributions.

Table 4. Results of the two-sided Pairwise Wilcoxon test with the BH adjustment on the performance metrics between treatments

Algorithm	Metrics	Treatment group 1	Treatment group 2	n1	n2	W	p-value	Adjusted p-value
HGBR	Train R ²	Untreated	Averaged	12	12	32	0.622	0.622
HGBR	Train R ²	Untreated	Dropped	12	12	78	<0.001	<0.001
HGBR	Train R ²	Averaged	Dropped	12	12	78	<0.001	<0.001
RFR	Train R ²	Untreated	Averaged	12	12	1	0.001	0.001
RFR	Train R ²	Untreated	Dropped	12	12	78	<0.001	<0.001
RFR	Train R ²	Averaged	Dropped	12	12	78	<0.001	<0.001
LGBR	Train R ²	Untreated	Averaged	12	12	1	0.001	0.001
LGBR	Train R ²	Untreated	Dropped	12	12	78	<0.001	<0.001
LGBR	Train R ²	Averaged	Dropped	12	12	78	<0.001	<0.001
HGBR	Test R ²	Untreated	Averaged	12	12	71	0.009	0.009
HGBR	Test R ²	Untreated	Dropped	12	12	78	<0.001	<0.001
HGBR	Test R ²	Averaged	Dropped	12	12	78	<0.001	<0.001
RFR	Test R ²	Untreated	Averaged	12	12	76	0.001	0.001
RFR	Test R ²	Untreated	Dropped	12	12	78	<0.001	<0.001
RFR	Test R ²	Averaged	Dropped	12	12	78	<0.001	<0.001
LGBR	Test R ²	Untreated	Averaged	12	12	76	0.001	0.001
LGBR	Test R ²	Untreated	Dropped	12	12	78	<0.001	<0.001
LGBR	Test R ²	Averaged	Dropped	12	12	78	<0.001	<0.001
HGBR	Train RMSE	Untreated	Averaged	12	12	67	0.027	0.027
HGBR	Train RMSE	Untreated	Dropped	12	12	0	<0.001	0.001
HGBR	Train RMSE	Averaged	Dropped	12	12	0	<0.001	0.001
RFR	Train RMSE	Untreated	Averaged	12	12	78	<0.001	<0.001
RFR	Train RMSE	Untreated	Dropped	12	12	0	<0.001	<0.001
RFR	Train RMSE	Averaged	Dropped	12	12	0	<0.001	<0.001
LGBR	Train RMSE	Untreated	Averaged	12	12	78	<0.001	<0.001
LGBR	Train RMSE	Untreated	Dropped	12	12	0	<0.001	<0.001
LGBR	Train RMSE	Averaged	Dropped	12	12	0	<0.001	<0.001
HGBR	Test RMSE	Untreated	Averaged	12	12	0	<0.001	<0.001
HGBR	Test RMSE	Untreated	Dropped	12	12	0	<0.001	<0.001
HGBR	Test RMSE	Averaged	Dropped	12	12	0	<0.001	<0.001
RFR	Test RMSE	Untreated	Averaged	12	12	0	<0.001	<0.001
RFR	Test RMSE	Untreated	Dropped	12	12	0	<0.001	<0.001
RFR	Test RMSE	Averaged	Dropped	12	12	0	<0.001	<0.001
LGBR	Test RMSE	Untreated	Averaged	12	12	0	<0.001	<0.001
LGBR	Test RMSE	Untreated	Dropped	12	12	0	<0.001	<0.001
LGBR	Test RMSE	Averaged	Dropped	12	12	0	<0.001	<0.001

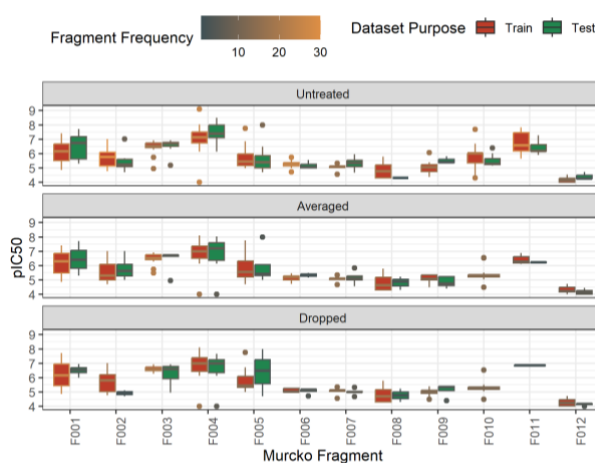


Figure 3. The boxplots of the pIC_{50} distributions for the selected Murcko fragments

Figure 4 shows the 12 selected Murcko fragments plotted as line structure, followed by the name, and statistics for each treatment. As the splitting strategy used an 80:20 proportion for training and testing, respectively, it can be seen that not all of these selected fragments are evenly distributed regarding the proportion. For instance, in each treatment, there 30 compounds share Murcko fragment F001. In the untreated duplicates dataset, the split is exactly 80:20 (24:6), but in the averaged and dropped, the splits are slightly shifted to 86.67:13.33 (26:4). F002 is another frequent Murcko fragment, that split with a ratio 78.94:21.06 (30:8), 80:20 (20:5), and 90:10 (18:2) at the untreated, averaged, and dropped duplicate treatments, respectively. The ratio for the Murcko fragment F002 at the dropped treatment has a major deviation from the expected split ratio. The deviations of the split ratio are even more noticeable for the selected Murcko fragments with less frequency, such as F010 and F011. Murcko fragment F010 was distributed with a ratio of 75:5 (21:7) for the untreated duplicate and 100:0 for the other two treatments.

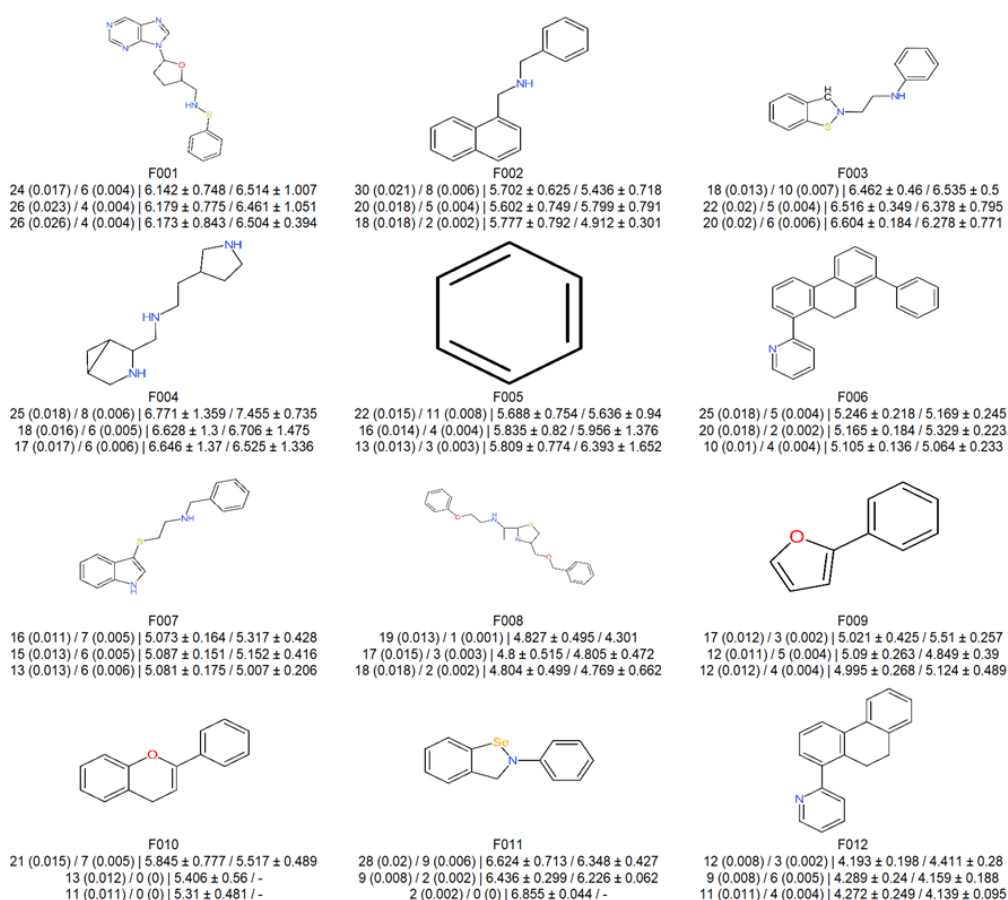


Figure 4. Selected Murcko fragments

The first line in each cell shows the fragment number (F###). Then the second, third, and fourth lines show the proportion and pIC_{50} statistics in untreated duplicates, averaged pIC_{50} , and dropped duplicates, respectively. In each line, the numbers show frequencies and proportions of the respective fragment in the training/testing dataset, followed by the respective average and standard deviation of pIC_{50} in the training/testing dataset.

3.3. Discussions

Typically, hyperparameter tuning is applied to gain higher ML model performance such as demonstrated in previous studies [38]. However, even with a large hyperparameters search space, in this particular study, we found the regressors' performances were not as expected. Therefore, by conducting further analyses, we applied statistical tests to the performance data, grouped by the treatments for duplicated bioactivity data. The results of the repeated measures Friedman test show that the differences in the data preparation significantly impact model performance, regardless of the algorithms. This finding is consistent

with previous studies on hyperparameter optimization. A study by Schratz *et al.* [39] on hyperparameter tuning in the field of ecological modeling, it was found that the results of hyperparameter tuning might be negligible for RF. Similarly, Sipper [40] evaluated many algorithms and datasets and found that considerable gains could not always be expected from hyperparameter tuning. The study also found that RFR, which was also used in our study, is one algorithm expected to gain less from hyperparameter tuning.

Splitting the dataset for training and testing is a standard practice in ML. In classification tasks, ensuring the balance between the labels or classes is an important consideration in data preparation since the diversity of the samples in each class brings considerable influence to the model performance [41]. In another study of heart disease classification with ensemble algorithms, the preserved distribution in train-test splitting brought considerable impacts to the overall performance [42]. Prediction tasks such as regressions do not share this dataset imbalance problem due to the different nature of the target. However, the representativeness of the data characteristics distribution in both training and testing datasets has to be considered. This implies that the fairness of data characteristics in the train-test split has to be considered, as proposed in the study by Salazar *et al.* [43]. In this study, regardless of the hyperparameter tuning with an exhaustive search space on various combinations of treatments of duplicates, feature extraction using various molecular fingerprints as descriptors, and several algorithms, the best models still have low performance. As the Murcko fragment represents the core structural framework of a molecule, including its rings and linkers, with the side chains or terminal substituents excluded, it is central to the molecular structure and often considered as the scaffold on which various functional groups are attached. Our investigation of the Murcko fragments distributions in the train and test datasets found that some of them were not equally distributed in both datasets, resulting in a fragments imbalance between the datasets, therefore, the features learned by the models are different from those in the test dataset. This issue should be considered further with an expanded list of algorithms and bioactivity targets.

4. CONCLUSION

In this study, we investigated the low performance of the ensemble tree-based regressor algorithms in predicting the IC_{50} of small molecules, targeting the SARS-CoV-2 p1ab. Despite the exhaustive hyperparameter search space, various combinations of treatments of duplicate bioactivity data and molecular fingerprint descriptors as features, none of the resulting models gained a satisfactory number of R^2 and RMSE. Treatment-wise, dropping all the duplicated bioactivity data yielded the worst performance compared to the other two treatments. The R^2 values across modeling stages (train, cross-validation, and test) tend to have similar trends regardless of the molecular fingerprints and algorithms. However, a deeper comparison of the RMSE in each molecular fingerprint shows that the experiments with untreated duplicates tend to yield higher RMSE in test cross-validation than in the real training dataset. At the same time, as a loss function, it should be the other way around. Hence, based on our experiments, treating the duplicates by averaging the pIC_{50} brought more reasonable results. The balanced distribution between labels is an important factor in overall model performance in classification tasks. By having balanced label distribution in both training and testing datasets, the consistency of the data could be preserved, hence, the characteristics faced by the algorithm during model training could also be found when evaluating the model with the testing dataset. Regardless of the nature of the task, the representativeness of the characteristics in the training and testing datasets also influences the model performance. In our study, our investigation of the Murcko fragments distributions in the datasets used for training and testing was not balanced. There are cases where some of the frequent Murcko fragments in the whole dataset were not evenly distributed or did not exist in the testing dataset. This is considered the main cause of the models, despite hyperparameters being tuned with an exhaustive list of search space, which tends to overfit. Future studies should consider the issue of Murcko fragment distribution. When investigating the effect of Murcko fragment distributions in quantitative structure-activity relationship (QSAR) modeling, a wide range of algorithms, targets, tasks, and split ratios must be considered.

ACKNOWLEDGEMENTS

The authors thank the staff of the Information and Communication Technology Academic Support Unit, Universitas Sam Ratulangi, for technical support during the experiments.

FUNDING INFORMATION

This work is funded by the Daftar Isian Pelaksanaan Anggaran (DIPA) Universitas Sam Ratulangi: Riset Dasar Unggulan UNSRAT 2024, contract number: 184/UN12.13/LT/2024.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Daniel Febrian Sengkey	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Angelina Stevany Regina Masengi	✓	✓		✓	✓	✓			✓	✓			✓	✓
Alwin Melkie Sambul	✓	✓	✓	✓			✓			✓			✓	✓
Trina Ekawati Tallei				✓	✓	✓				✓		✓		
Sherwin Reinaldo		✓	✓				✓	✓		✓				
Unsratdianto Sompie														

C : **C**onceptualization

M : **M**ethodology

So : **S**oftware

Va : **V**alidation

Fo : **F**ormal analysis

I : **I**nvestigation

R : **R**esources

D : **D**ata Curation

O : **O**riginal Draft

E : **E**ditings

Vi : **V**isualization

Su : **S**upervision

P : **P**roject administration

Fu : **F**unding acquisition

CONFLICT OF INTEREST STATEMENT

The authors state no conflict of interest.

DATA AVAILABILITY

The bioactivity dataset used in this research was retrieved from the ChEMBL database at <https://ebi.ac.uk/chembl> in April 2024. Preprocessed datasets with extracted fingerprints as features and the different treatments on the duplicate bioactivity are available in https://github.com/danielsengkey/supplementaries/tree/main/pp1lab_ijai2025.

REFERENCES





- [1] X. Huang, R. Pearce, G. S. Omenn, and Y. Zhang, "Identification of 13 guanidinobenzoyl- or amidinobenzoyl-containing drugs to potentially inhibit TMPRSS2 for COVID-19 treatment," *International Journal of Molecular Sciences*, vol. 22, no. 13, Jun. 2021, doi: 10.3390/ijms22137060.
- [2] R. Alexpandi, J. F. De Mesquita, S. K. Pandian, and A. V. Ravi, "Quinolines-based SARS-CoV-2 3CLpro and RdRp inhibitors and spike-RBD-ACE2 inhibitor for drug-repurposing against COVID-19: an in silico analysis," *Frontiers in Microbiology*, vol. 11, Jul. 2020, doi: 10.3389/fmicb.2020.01796.
- [3] A. Khaled and Z. A. El Haliem, "Generative recurrent network for design SARS-CoV-2 main protease inhibitor," in *2022 International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, 2022, pp. 1–6, doi: 10.23919/SoftCOM55329.2022.9911377.
- [4] D. Shaji, S. Yamamoto, R. Saito, R. Suzuki, S. Nakamura, and N. Kurita, "Proposal of novel natural inhibitors of severe acute respiratory syndrome coronavirus 2 main protease: molecular docking and *ab initio* fragment molecular orbital calculations," *Biophysical Chemistry*, vol. 275, Aug. 2021, doi: 10.1016/j.bpc.2021.106608.
- [5] F. Hu, D. Wang, Y. Hu, J. Jiang, and P. Yin, "Generating novel compounds targeting SARS-CoV-2 main protease based on imbalanced dataset," in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, Dec. 2020, pp. 432–436, doi: 10.1109/BIBM49941.2020.9313317.
- [6] I. Achilonu, E. A. Iwuchukwu, O. J. Achilonu, M. A. Fernandes, and Y. Sayed, "Targeting the SARS-CoV-2 main protease using FDA-approved isavuconazonium, a P2–P3 α -ketoamide derivative and pentagastrin: an *in-silico* drug discovery approach," *Journal of Molecular Graphics and Modelling*, vol. 101, Dec. 2020, doi: 10.1016/j.jmgm.2020.107730.
- [7] N. Ferdous *et al.*, "Mpropred: a machine learning (ML) driven web-app for bioactivity prediction of SARS-CoV-2 main protease (M^{pro}) antagonists," *PLOS ONE*, vol. 18, no. 6, Jun. 2023, doi: 10.1371/journal.pone.0287179.
- [8] T. E. Tallei *et al.*, "Potential of plant bioactive compounds as SARS-CoV-2 main protease (M^{pro}) and spike (S) glycoprotein inhibitors: a molecular docking study," *Scientifica*, vol. 2020, pp. 1–18, Dec. 2020, doi: 10.1155/2020/6307457.
- [9] T. E. Tallei *et al.*, "Fruit bromelain-derived peptide potentially restrains the attachment of SARS-CoV-2 variants to hACE2: a pharmacoinformatics approach," *Molecules*, vol. 27, no. 1, Jan. 2022, doi: 10.3390/molecules27010260.
- [10] F. Sulistiawan, W. A. Kusuma, N. S. Ramadhanti, and A. Tedjo, "Drug-target interaction prediction in coronavirus disease 2019 case using deep semi-supervised learning model," in *2020 International Conference on Advanced Computer Science and Information Systems*, IEEE, Oct. 2020, pp. 83–88, doi: 10.1109/ICACSIS51025.2020.9263241.
- [11] L. Erlina *et al.*, "Virtual screening of Indonesian herbal compounds as COVID-19 supportive therapy: machine learning and pharmacophore modeling approaches," *BMC Complementary Medicine and Therapies*, vol. 22, no. 1, Dec. 2022, doi: 10.1186/s12906-022-03686-y.
- [12] N. S. Ramadhanti, W. A. Kusuma, I. Batubara, and R. Heryanto, "Random forest to predict eucalyptus as a potential herb in preventing COVID19," in *2021 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, IEEE, Oct. 2021, pp. 01–05, doi: 10.1109/CIBCB49929.2021.9562940.
- [13] A. A. Malik, C. Phanus-umporn, N. Schaduangrat, W. Shoombuatong, C. Isarankura-Na-Ayudhya, and C. Nantasenamat, "HCVpred: a web server for predicting the bioactivity of hepatitis C virus NS5B inhibitors," *Journal of Computational Chemistry*, vol. 41, no. 20, pp. 1820–1834, Jul. 2020, doi: 10.1002/jcc.26223.

Investigation on low-performance tuned-regressor of inhibitory concentration ... (Daniel Febrian Sengkey)





- [14] T. Lerksuthirat, S. Chitphuk, W. Stitchantrakul, D. Dejsuphong, A. A. Malik, and C. Nantasenamat, "PARP1pred: a web server for screening the bioactivity of inhibitors against DNA repair enzyme PARP-1," *EXCLI Journal*, vol. 22, pp. 84–107, 2023, doi: 10.17179/excli2022-5602.
- [15] C. Bennette and A. Vickers, "Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents," *BMC Medical Research Methodology*, vol. 12, no. 1, Dec. 2012, doi: 10.1186/1471-2288-12-21.
- [16] K. E. Markon, M. Chmielewski, and C. J. Miller, "The reliability and validity of discrete and continuous measures of psychopathology: a quantitative review," *Psychological Bulletin*, vol. 137, no. 5, pp. 856–879, 2011, doi: 10.1037/a0023678.
- [17] Z. Gao, R. Xia, and P. Zhang, "Prediction of anti-proliferation effect of [1,2,3] Triazolo [4,5-d] pyrimidine derivatives by random forest and mix-kernel function SVM with PSO," *Chemical and Pharmaceutical Bulletin*, vol. 70, no. 10, Oct. 2022, doi: 10.1248/cpb.c22-00376.
- [18] S. Kamboj, A. Rajput, A. Rastogi, A. Thakur, and M. Kumar, "Targeting non-structural proteins of hepatitis C virus for predicting repurposed drugs using QSAR and machine learning approaches," *Computational and Structural Biotechnology Journal*, vol. 20, pp. 3422–3438, 2022, doi: 10.1016/j.csbj.2022.06.060.
- [19] D. N. Fiat *et al.*, "Comparative analysis of Hepatitis C virus genotype 1a (Isolate 1) using multiple regression algorithms and fingerprinting techniques," *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 6, no. 4, pp. 478–488, Sep. 2024, doi: 10.35882/jeeemi.v6i4.506.
- [20] T. R. Noviandy, G. M. Idroes, T. E. Tallei, D. Handayani, and R. Idroes, "QSAR modeling for predicting beta-secretase 1 inhibitory activity in alzheimer's disease with support vector regression," *Malacca Pharmaceutics*, vol. 2, no. 2, pp. 79–85, Sep. 2024, doi: 10.60084/mp.v2i2.226.
- [21] G. I. B. Janairo, D. E. C. Yu, and J. I. B. Janairo, "A machine learning regression model for the screening and design of potential SARS-CoV-2 protease inhibitors," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 10, no. 1, Dec. 2021, doi: 10.1007/s13721-021-00326-2.
- [22] D. F. Sengkey and A. Masengi, "Regression algorithms in predicting the SARS-CoV-2 replicase polyprotein lab inhibitor: a comparative study," *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 6, no. 1, pp. 1–10, Dec. 2023, doi: 10.35882/jeeemi.v6i1.338.
- [23] Y. Cárdenas-Conejo, A. Liñan-Rico, D. A. García-Rodríguez, S. Centeno-Leija, and H. Serrano-Posada, "An exclusive 42 amino acid signature in p1lab protein provides insights into the evolutive history of the 2019 novel human-pathogenic coronavirus (SARS-CoV-2)," *Journal of Medical Virology*, vol. 92, no. 6, pp. 688–692, Jun. 2020, doi: 10.1002/jmv.25758.
- [24] R. Yadav *et al.*, "Role of structural and non-structural proteins and therapeutic targets of SARS-CoV-2 for COVID-19," *Cells*, vol. 10, no. 4, Apr. 2021, doi: 10.3390/cells10040821.
- [25] T. Badriyah, D. B. Santoso, I. Syarif, and D. R. Syarif, "Improving stroke diagnosis accuracy using hyperparameter optimized deep learning," *International Journal of Advances in Intelligent Informatics*, vol. 5, no. 3, Nov. 2019, doi: 10.26555/ijain.v5i3.427.
- [26] H. J. P. Weerts, A. C. Mueller, and J. Vanschoren, "Importance of tuning hyperparameters of machine learning algorithms," *arXiv-Computer Science*, pp. 1-17, Jul. 2020.
- [27] B. Bischl *et al.*, "Hyperparameter optimization: foundations, algorithms, best practices, and open challenges," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 13, no. 2, 2023, doi: 10.1002/widm.1484.
- [28] A. Boulesteix, B. Bischl, and P. Probst, "Tunability: importance of hyperparameters of machine learning algorithms," *Journal of Machine Learning Research*, vol. 20, no. 53, 2019.
- [29] F. Pedregosa *et al.*, "Scikit-learn: machine learning in Python," *Journal of Machine Learning Research*, vol. 127, no. 9, pp. 2825–2830, 2019.
- [30] M. Davies *et al.*, "ChEMBL web services: streamlining access to drug discovery data and utilities," *Nucleic Acids Research*, vol. 43, no. W1, Jul. 2015, doi: 10.1093/nar/gkv352.
- [31] C. W. Yap, "PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints," *Journal of Computational Chemistry*, vol. 32, no. 7, pp. 1466–1474, May 2011, doi: 10.1002/jcc.21707.
- [32] R Core Team, *R: a language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2024. [Online]. Available: <https://cran.r-project.org/doc/manuals/r-release/fullrefman.pdf>
- [33] N. N. Ivanov, D. A. Shulga, and V. A. Palyulin, "Decomposition of small molecules for fragment-based drug design," *Biophysica*, vol. 3, no. 2, pp. 362–372, 2023, doi: 10.3390/biophysica3020024.
- [34] G. W. Bemis and M. A. Murecko, "The properties of known drugs. 1. molecular frameworks," *Journal of Medicinal Chemistry*, vol. 39, no. 15, pp. 2887–2893, Jan. 1996, doi: 10.1021/jm9602928.
- [35] A. Kumar, S. Loharch, S. Kumar, R. P. Ringe, and R. Parkesh, "Exploiting cheminformatic and machine learning to navigate the available chemical space of potential small molecule inhibitors of SARS-CoV-2," *Computational and Structural Biotechnology Journal*, vol. 19, pp. 424–438, 2021, doi: 10.1016/j.csbj.2020.12.028.
- [36] T. Yu *et al.*, "Exploring the chemical space of CYP17A1 inhibitors using cheminformatics and machine learning," *Molecules*, vol. 28, no. 4, Feb. 2023, doi: 10.3390/molecules28041679.
- [37] R. Guha, "Chemical informatics functionality in R," *Journal of Statistical Software*, vol. 18, no. 5, 2007, doi: 10.18637/jss.v018.i05.
- [38] K. S. Nugroho, A. Y. Sukmadewa, A. Vidianto, and W. F. Mahmudy, "Effective predictive modelling for coronary artery diseases using support vector machine," *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 11, no. 1, Mar. 2022, doi: 10.11591/ijai.v11.i1.pp345-355.
- [39] P. Schratz, J. Muenchow, E. Iturrutxa, J. Richter, and A. Brenning, "Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data," *Ecological Modelling*, vol. 406, pp. 109–120, Aug. 2019, doi: 10.1016/j.ecolmodel.2019.06.002.
- [40] M. Sipper, "High per parameter: a large-scale study of hyperparameter tuning for machine learning algorithms," *Algorithms*, vol. 15, no. 9, Sep. 2022, doi: 10.3390/a15090315.
- [41] A. Rácz, D. Bajusz, and K. Héberger, "Effect of dataset size and train/test split ratios in QSAR/QSPR multiclass classification," *Molecules*, vol. 26, no. 4, Feb. 2021, doi: 10.3390/molecules26041111.
- [42] D. Mohapatra, S. K. Bhoi, C. Mallick, K. K. Jena, and S. Mishra, "Distribution preserving train-test split directed ensemble classifier for heart disease prediction," *International Journal of Information Technology*, vol. 14, no. 4, pp. 1763–1769, Jun. 2022, doi: 10.1007/s41870-022-00868-2.
- [43] J. J. Salazar, L. Garland, J. Ochoa, and M. J. Pyrcz, "Fair train-test split in machine learning: mitigating spatial autocorrelation for improved prediction accuracy," *Journal of Petroleum Science and Engineering*, vol. 209, Feb. 2022, doi: 10.1016/j.petrol.2021.109885.

BIOGRAPHIES OF AUTHORS







Daniel Febrian Sengkey     is a lecturer at the Undergraduate Program in Informatics, Department of Electrical Engineering, Faculty of Engineering, Universitas Sam Ratulangi, Manado-Indonesia. He graduated from the Undergraduate Program in Electrical Engineering of the same department in 2021, then in 2015 achieved his Master's degree in Electrical Engineering from the Department of Electrical Engineering and Information Technology, Faculty of Engineering, Universitas Gadjah Mada, Yogyakarta-Indonesia. His current teaching and research activities are mainly related to the machine learning fundamentals, as well as implementation, especially in biomedical and health informatics. Besides his assignment in the informatics program, he is also a member of the bioinformatics team at the university's Biomolecular Laboratory. His teaching assignments cover fundamental mathematics, machine learning, and bioinformatics. He can be contacted at email: danielsengkey@unsrat.ac.id.







Angelina Stevany Regina Masengi     is currently the Acting Secretary of the Department of Pharmacology and Therapy, Faculty of Medicine, Universitas Sam Ratulangi, and also part of the occupational health and safety team of the Biomolecular Laboratory of the same university. She achieved her Bachelor of Medicine as well as her Medical Doctor from the Faculty of Medicine, Universitas Pelita Harapan, in 2008 and 2010, respectively. Since 2016, she holds a Master's degree in Biomedics from Universitas Indonesia. Since 2018, she has been a tenured lecturer at Universitas Sam Ratulangi, with main teaching assignments in the Undergraduate Program in Medicine. She is also involved in teaching biochemistry courses in the undergraduate nursing, dentistry, and pharmacy programs. She was formerly engaged in the Bioinformatics course at the Undergraduate Program in Informatics. Starting in 2024, she is enrolled in the Doctoral Program in Entomology of the Universitas Sam Ratulangi, focusing on insect-related drug discovery. She can be contacted at email: asrmasengi@unsrat.ac.id.







Dr. Alwin Melkie Sambul     earned his undergraduate degree in Electrical Engineering from Universitas Sam Ratulangi in 2003. His Master's and Ph.D. degrees in Biomedical Engineering, from Kumamoto University, Japan, are completed in 2011 and 2015, respectively. He is currently the Head of the Department of Electrical Engineering at the Faculty of Engineering, and also part of the bioinformatics team at the Biomolecular Laboratory. Both offices he holds are within the Universitas Sam Ratulangi. His research interest is in biomedical engineering, especially in brainwave signaling. He teaches courses such as biomedical informatics, bioinformatics, and algorithms. He can be contacted at email: asambul@unsrat.ac.id.



Prof. Trina Ekawati Tallei     is a Professor at the Department of Biology, Faculty of Mathematics and Natural Sciences, and at the Department of Biology, Faculty of Medicine, Universitas Sam Ratulangi. She is the current head of the latter. She is also the Quality Assurance Manager of the BioMolecular Laboratory, Universitas Sam Ratulangi. She is recognized globally for her contributions to molecular biology and drug discovery. In 2013, she was honored with a millennium development goals award. Notably, she was ranked among the top 2% of the most cited scientists globally in 2023 and 2024. In 2024, she delivered the Sarwono Prawirohardjo Memorial Lecture by the National Research and Innovation Agency. Her research interests encompass molecular sciences, drug discovery, molecular biology, bioinformatics, computer-aided drug design, and metagenomics. She is involved in various teaching activities, including the bioinformatics course in the undergraduate program in informatics. She can be contacted at email: trina_tallei@unsrat.ac.id.



Sherwin Reinaldo Unsratdianto Sompie     is the present Head of the Information and Communication Technology Academic Support Unit (formerly Technical Support Unit) of the Universitas Sam Ratulangi. Before taking office in May 2023, he was the Head of the Department of Electrical Engineering at the Faculty of Engineering, and before that, Head of the Informatics Laboratory of the same faculty. Graduated from the Electrical Engineering Program at Petra Christian University in 2002, he then started his career as a lecturer at the Electrical Engineering Program at Universitas Sam Ratulangi. In 2011, he completed his Master's Degree in Electrical Engineering at Universitas Pelita Harapan. His research interest covers optical systems, renewable energy with photovoltaic systems, and deep learning applications in image processing. He can be contacted at email: aldo@unsrat.ac.id.