

Article

# Machine Learning Application for Medicinal Chemistry: Colchicine Case, New Structures, and Anticancer Activity Prediction

Damian Nowak <sup>1,\*</sup> , Adam Huczyński <sup>2</sup> , Rafał Adam Bachorz <sup>3,4</sup>  and Marcin Hoffmann <sup>1,\*</sup> 

<sup>1</sup> Department of Quantum Chemistry, Faculty of Chemistry, Adam Mickiewicz University in Poznan, Uniwersytetu Poznańskiego 8, 61-614 Poznan, Poland

<sup>2</sup> Department of Medical Chemistry, Faculty of Chemistry, Adam Mickiewicz University in Poznan, Uniwersytetu Poznańskiego 8, 61-614 Poznan, Poland; adam.huczynski@amu.edu.pl

<sup>3</sup> Institute of Medical Biology of Polish Academy of Sciences, Lodowa 106, 93-232 Lodz, Poland; rbachorz@cbm.pan.pl

<sup>4</sup> Institute of Computing Science, Faculty of Computing, Poznań University of Technology, Piotrowo 2, 60-965 Poznań, Poland

\* Correspondence: damian.nowak@amu.edu.pl (D.N.); mmh@amu.edu.pl (M.H.)

**Abstract:** In the contemporary era, the exploration of machine learning (ML) has gained widespread attention and is being leveraged to augment traditional methodologies in quantitative structure–activity relationship (QSAR) investigations. The principal objective of this research was to assess the anticancer potential of colchicine-based compounds across five distinct cell lines. This research endeavor ultimately sought to construct ML models proficient in forecasting anticancer activity as quantified by the  $IC_{50}$  value, while concurrently generating innovative colchicine-derived compounds. The resistance index (RI) is computed to evaluate the drug resistance exhibited by LoVo/DX cells relative to LoVo cancer cell lines. Meanwhile, the selectivity index (SI) is computed to determine the potential of a compound to demonstrate superior efficacy against tumor cells compared to its toxicity against normal cells, such as BALB/3T3. We introduce a novel ML system adept at recommending novel chemical structures predicated on known anticancer activity. Our investigation entailed the assessment of inhibitory capabilities across five cell lines, employing predictive models utilizing various algorithms, including random forest, decision tree, support vector machines, k-nearest neighbors, and multiple linear regression. The most proficient model, as determined by quality metrics, was employed to predict the anticancer activity of novel colchicine-based compounds. This methodological approach yielded the establishment of a library encompassing new colchicine-based compounds, each assigned an  $IC_{50}$  value. Additionally, this study resulted in the development of a validated predictive model, capable of reasonably estimating  $IC_{50}$  values based on molecular structure input.

**Keywords:** machine learning; colchicine; drug discovery; anticancer activity; QSAR; molecular docking; in silico screening



**Citation:** Nowak, D.; Huczyński, A.; Bachorz, R.A.; Hoffmann, M. Machine Learning Application for Medicinal Chemistry: Colchicine Case, New Structures, and Anticancer Activity Prediction. *Pharmaceuticals* **2024**, *17*, 173. <https://doi.org/10.3390/ph17020173>

Academic Editor: Krzysztof Marciniak

Received: 20 November 2023

Revised: 2 January 2024

Accepted: 12 January 2024

Published: 29 January 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Machine learning (ML) methods are being investigated to speed up the discovery of new bioactive chemical structures. Current methods are aiming to propose novel chemical structures with desired properties based on already known chemical structures [1–7]. ML models learn from the experimentally determined biological activities and molecular descriptors or other mathematical descriptors of a chemical molecule [8]. The molecular features can be easily computed with the application of the RDKit [9] and Mordred [10] libraries, which take advantage of the linear representation of a structure, called SMILES [11]. With the gathering of all the knowledge given above, we can propose new chemical structures and assign the potential biological activity to them, which can be useful information for selection for experimental verification [12].

The objects of the studies, from the chemical point of view, are colchicine-based derivatives. These can be potentially used as anticancer treatments [13]. The biological activity feature, in this study, is defined as  $IC_{50}$  (nM), which is defined as a quantitative measure that indicates how much of a particular inhibitory substance (e.g., drug) is needed to inhibit, in vitro, a given biological process or biological component by a half [14]. This study aims to predict  $IC_{50}$  values for various cell lines, namely:

1. A549—adenocarcinomic human alveolar basal epithelial cells, lung cancer related [15];
2. BALB/3T3—detection of the carcinogenic potential of chemicals [16];
3. LoVo/DX—human colon adenocarcinoma doxorubicin-resistant cell line [17];
4. LoVo—human colon adenocarcinoma cell line, colorectal cancer related [18];
5. MCF-7—breast-cancer-related cell line [19].

The resistance index (RI) was calculated to assess the resistance of the cancer cell lines given above. The RI index indicates how many times a resistant subline is chemoresistant relative to its non-resistant cell line. It is calculated with the application of the following formula:  $RI = IC_{50}$  for LoVo/DX cell line divided by  $IC_{50}$  for LoVo cell line. When the RI value is in the range from 0 to 2, the cells are sensitive to the compound, an RI in the range 2–10 means moderate sensitivity, and an RI above 10 indicates strong resistance [20].

The selectivity index (SI) serves as a metric for evaluating the selectivity of a novel colchicine-based compound. It is determined for each specific cell line through the utilization of the following formula:  $SI = IC_{50}$  for the normal cell line divided by  $IC_{50}$  for the corresponding cancerous cell line. An SI value greater than unity indicates that the drug exhibits enhanced efficacy against tumor cells in comparison to its toxicity towards normal cells. For instance, in the case of the MCF-7 cancer cell line, the SI is calculated as follows:  $SI = IC_{50}$  for the BALB/3T3 normal cell line divided by  $IC_{50}$  for the MCF-7 cancer cell line [21].

The first level of selection can be conducted in silico, as it requires a low-resource approach. Then, from the selected compounds, some can be tested in vitro. The presented cell lines can be treated with new colchicine-based derivatives and tested for the required biological activity and safety of the new compounds in vitro. The aim is to have as low an  $IC_{50}$  value as possible for the cancer cell lines and higher activities for the BALB/3T3 cell line. If the colchicine-based compound has met the requirements for both good biological activity and safety, the evaluation of the compound can go to the next stage, the in vivo stage.

The data underpinning this study were gathered from diverse sources, and the compounds under scrutiny underwent rigorous in vitro assessments. Specifically, the data stem from the experimental efforts of medicinal chemists (Czerwonka, Krzywik et al.) as documented in multiple sources [22–26]. These sources serve as pivotal components of this study, likely containing comprehensive details about colchicine-based compounds. They encompass information regarding synthesis methodologies and the compounds' anticancer attributes, denoted by their half-maximal inhibitory concentration ( $IC_{50}$ ) values in nanomolars (nM) against different cell lines. Additionally, this study refers to the availability of other relevant data at the PubChem database [27], augmenting the dataset with additional insights into the compounds or their properties, thereby enhancing the research's comprehensiveness.

The realm of machine learning (ML) spans a spectrum of methodologies, ranging from elementary approaches like linear regression [28] to more intricate models such as multiple linear regression (MLR) [29], decision trees (DTs) [30], random forests (RFs) [31], k-nearest neighbors (KNNs) [32], support vector machines (SVMs) [33], XGBoost [34], and neural networks [35,36]. The complexity of an ML model often correlates with the volume of required data. While simpler models can suffice for smaller datasets, their scope might be limited. Models exhibiting behaviors of overfitting and underfitting require careful consideration; overfitting occurs when a model learns patterns too precisely from training data but performs poorly on new data, while underfitting reflects inadequate learning to generalize on the dataset [37].

Additionally, molecular docking emerges as a crucial selection procedure for constructing new libraries of colchicine-based derivatives. The AutoDock Vina (AD Vina) algorithm [38], used in this study, boasts accuracy akin to its predecessor, AutoDock 4 (AD4) [39]. Plewczynski et al.'s studies [40] affirm that AD4 accurately redocked 93% of ligand–protein complexes. Despite the relatively simple scoring functions, software of this nature proves invaluable as a supporting tool in drug design processes.

Our study can be divided into two pathways, which can be run separately. The first route is the construction of new colchicine-based compounds with the application of a previously prepared algorithm [4,12]. This pathway lets us create a library of new colchicine-based structures after a few steps of selection. The second pathway is related to the biological activity, using machine learning (ML) models' predictions. To obtain this target, we need to build and test several ML models and select the "best" of them for each of the biological activity targets. Then, with the "best" models, we can assign target features,  $IC_{50}$  values, for the newly created colchicine-based structures. At the end, we have a library of in silico-tested compounds that have a predicted  $IC_{50}$  value, binding affinity to the 1SA0 protein domain [41], and SYBA score [42], which indicates the potential difficulty of synthesis. The selection of the 1SA0 protein domain is predicated on the recognition that colchicine exhibits binding affinity to  $\beta$ -tubulin, consequently inducing microtubule destabilization by rendering the colchicine-bound dimers incapable of assembly. Notably, in eukaryotic cells, which are characterized by the expression of various isoforms of  $\beta$ -tubulin,  $\beta I$  (one of the isoforms of  $\beta$ -tubulin) consistently stands out as the predominant and most prominent target for drug binding interactions [20].

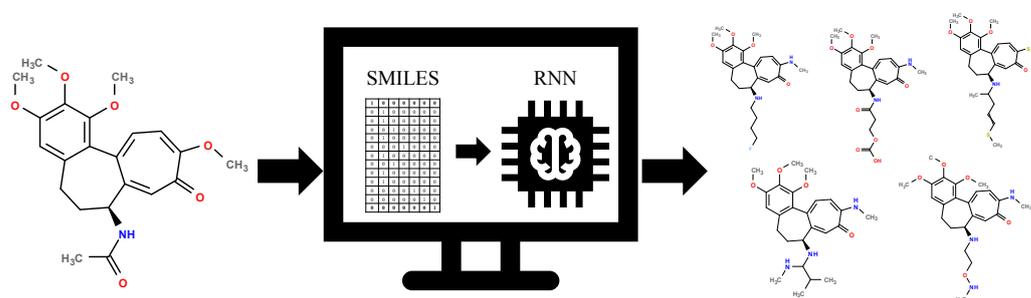
## 2. Results and Discussion

### 2.1. Training Data

The training dataset methodology proposed can be efficiently used for small datasets. This lets us use machine learning techniques for issues that are new, and thus, for which there is a lack of data, meaning chemical structures with assigned biological activity. This simple workaround gives us the opportunity to create new chemical structures from a small starting database [12]. Supplementary Files S1–S3 are related to this section.

### 2.2. Generative Neural Network

The generative RNN method [12] is illustrated conceptually in Figure 1. This technique enables the creation of a diverse library of novel colchicine-based structures by leveraging existing ones. By employing this approach, we have been able to develop a repository of fresh colchicine-based structures, which can undergo additional processing. For instance, selection procedures can be applied to refine and enhance these structures further.



**Figure 1.** The general scheme of new colchicine-based structures library creation.

With the 120 starting structures, we trained the neural network, namely, the recurrent neural network (RNN), with the goal of chemical space exploration. This let us teach the RNN how to reconstruct chemical structures, and with its application, new structures were proposed (Supplementary Files S4–S7). Figure 2 shows the decreasing loss value, indicating that the model's performance becomes better and better as the learning time increases (Supplementary File S8).

The proposed approach enabled us to create 1786 chemical structures (Supplementary Files S9–S11). Their creation was quite random, and due to that, data selection (Sections 2.4 and 3.4) was performed. These structures are distinct from each other. This let us keep more similar structures to the starting ones, as machine learning model predictions are more accurate in the closer chemical space.

Selected structures are shown below (Table 1). The table indicates the capabilities of a generative machine learning model, the recurrent neural network (RNN) [4]. Table 1 contains two starting structures with experimentally measured  $IC_{50}$  values for various cell lines and their SMILES codes. The second part of Table 1 shows some of the newly created chemical structures, their Tanimoto similarity [43] to the starting structure above, and, following the application of trained machine learning models, their predicted  $IC_{50}$  values. It can be observed that a small change in structure may lead to profound changes in biological activity (Supplementary File S41).

**Table 1.** The structures of selected AI-generated colchicine-based compounds. The table contains two starting structures with the assigned anticancer activities, Tanimoto similarity, and SMILES codes. For the newly generated structures  $IC_{50}$  values are predicted and not tested yet.

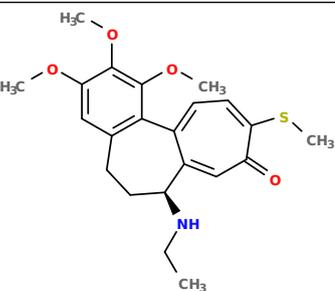
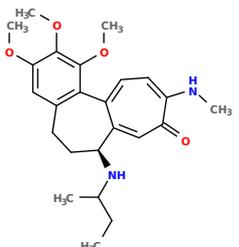
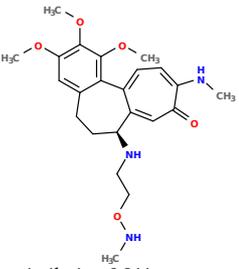
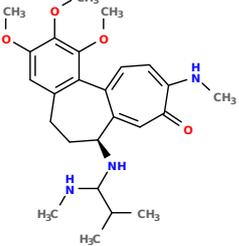
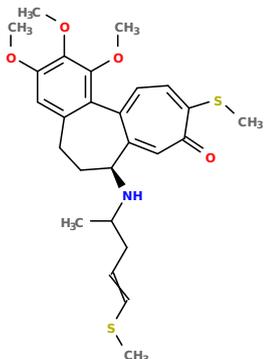
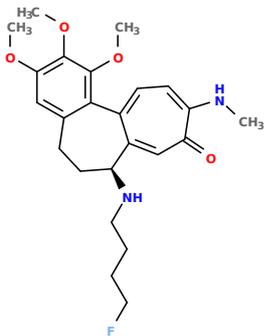
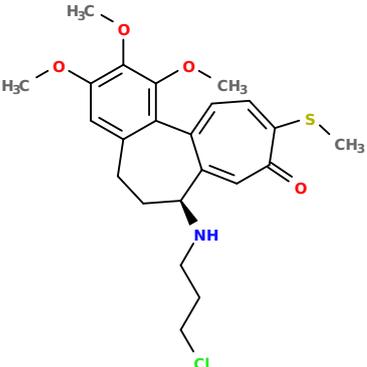
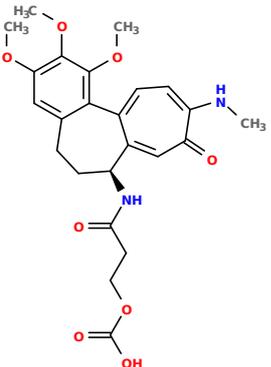
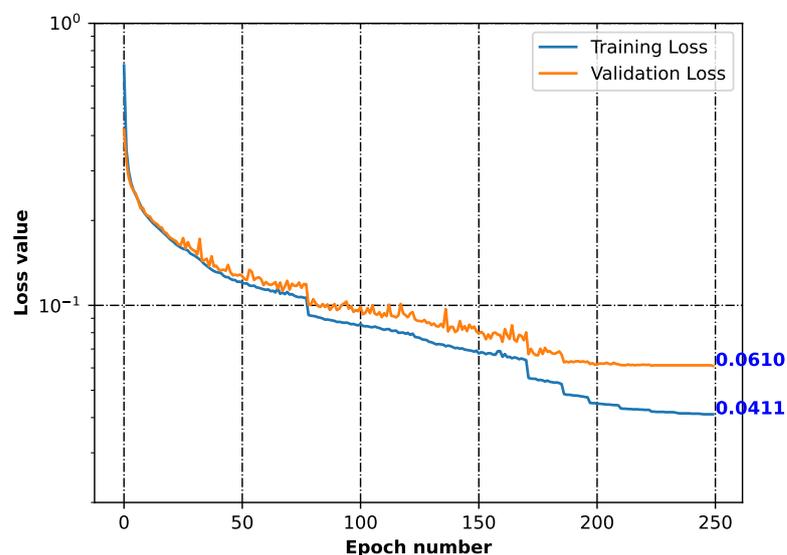
Starting structures <sup>1</sup>	
 <p><math>IC_{50}</math> (nM): A549: <math>10.8 \pm 1.4</math>, MCF-7: <math>10.3 \pm 0.4</math>, LoVo: <math>6.5 \pm 1.9</math>, LoVo/DX: <math>54.9 \pm 22.0</math>, BALB/3T3: <math>10.2 \pm 1.9</math> SMILES: <chem>COc2c3C1=CC=C(S(C)C(=O)C=C1[C@H](CCc3cc(OC)c2OC)NCC</chem></p>	 <p><math>IC_{50}</math> (nM): A549: <math>9.6 \pm 1.3</math>, MCF-7: <math>9.7 \pm 1.5</math>, LoVo: <math>7.8 \pm 1.0</math>, LoVo/DX: <math>8.5 \pm 1.1</math>, BALB/3T3: <math>7.5 \pm 1.5</math> SMILES: <chem>CC(C)N[C@H]2CCc3cc(OC)c(OC)c(OC)c3C1=CC=C(NC)C(=O)C=C12</chem></p>
Newly proposed structures <sup>1</sup>	
 <p>Tanimoto similarity: 0.844 <math>IC_{50}</math> (nM): A549: 8.3, MCF-7: 8.4, LoVo: 6.2, LoVo/DX: 94.3, BALB/3T3: 10.9 SMILES: <chem>COC1=C2C3=CC=C(NC)C(=O)C=C3[C@H]1(CCC2=CC(OC)=C1OC)NCCONC</chem></p>	 <p>Tanimoto similarity: 0.931 <math>IC_{50}</math> (nM): A549: 11.3, MCF-7: 7.9, LoVo: 9.6, LoVo/DX: 129.6, BALB/3T3: 21.3 SMILES: <chem>CNC(C(C)C)N[C@H]1C=2C(C3=C(C=O)C(OC)=C3OC)C1=CC=C(C(=O)C=2)NC</chem></p>
 <p>Tanimoto similarity: 0.953 <math>IC_{50}</math> (nM): A549: 11.2, MCF-7: 12.7, LoVo: 14.5, LoVo/DX: 71.1, BALB/3T3: 10.2 SMILES: <chem>C1=2[C@H]1(CCC3=CC(OC)=C(OC)C(OC)=C3C1=CC=C(S(C)C(=O)C=2)NC(C)CC=CSC</chem></p>	 <p>Tanimoto similarity: 0.978 <math>IC_{50}</math> (nM): A549: 15.9, MCF-7: 8.6, LoVo: 9.6, LoVo/DX: 21.6, BALB/3T3: 23.5 SMILES: <chem>N([C@H]1)CCC2=CC(OC)=C(OC)C(OC)=C2C3=CC=C(NC)C(C=C3)=O)CCCCF</chem></p>

Table 1. Cont.

	
<p>Tanimoto similarity: 0.983</p> <p><math>IC_{50}</math> (nM): A549: 10.6, MCF-7: 12.2, LoVo: 14.5, LoVo/DX: 18.1, BALB/3T3: 11.5</p> <p>SMILES: COC1=C2C3=CC=C(S(C)C(=O)C=C3[C@H]1)CC(C2=CC(OC)=C1OC)NCCCCl</p>	<p>Tanimoto similarity: 0.931</p> <p><math>IC_{50}</math> (nM): A549: 900, MCF-7: 4259, LoVo: 345, LoVo/DX: 9490, BALB/3T3: 835</p> <p>SMILES: O=C(N[C@H]1)C(C)C2=CC(OC)=C(OC)C(OC)=C2C3=CC=C(NC)C(=O)C=C3)CCOC(=O)O</p>

<sup>1</sup> Graphical representations of chemical structures were prepared with Open Babel software version 3.1.1 [44,45].



**Figure 2.** Training and validation losses of the neural network minimization. Both parameters dropping indicates that the model was learning how to generalize from the input.

### 2.3. Machine Learning Models for Anticancer Activity

The biological activity, given by  $IC_{50}$  (nM), was transformed with Equation (1) (Section 3.3). This is shown in File S12, and the results were saved in Supplementary File S13. The correlation of the molecular descriptors [46] with the  $IC_{50}$  values is shown in Supplementary File S14.

Five different machine learning methodologies were applied (Supplementary Files S15–S29). Namely, multiple linear regression (MLR) [29], decision tree (DT) [30], random forest (RF) [31], k-nearest neighbors regression (KNN) [32], and support vector machines (SVMs) [33]. Each of them was investigated through various molecular descriptors as training features. The lower the number of features, the higher the correlation of the features to the biological activity parameter, given by  $IC_{50}$  (nM).

Figures 3–7 (created within Supplementary File S30) exhibit the quality measurements for each methodology investigated across individual cell lines. The term ‘random state’ [47] denotes the initial seed employed by a pseudo-random number generator (PRNG) within machine learning algorithms. Its function is to ensure reproducibility by establishing a consistent starting point each time the code is executed, thereby ensuring uniform results during experiments or model training.

The random state value signifies variations in both the training and testing data while maintaining their specified percentages, as detailed in Section 3.3. The distribution of these datasets can be observed in Supplementary Files S55–S70. The utilization of different random states during data splitting leads to alterations in the composition of the training and testing datasets. Consequently, distinct data points are employed for the training and testing procedures at each random state.

Two scoring parameters were explored to select the best final models for predicting the activity of each of the cell lines. The first parameter is the correlation coefficient (R) [48], and the second is the root mean square error (RMSE) [49]. According to a study conducted by D. Chicco et al., the coefficient of determination  $R^2$  (where the correlation coefficient is the square root of  $R^2$ ) provides more informative insights than the RMSE.  $R^2$  does not possess the interpretative limitations of the MSE, RMSE, MAE, and MAPE. Therefore, we suggest using R-squared as the standard metric for evaluating regression analyses in any scientific domain [50]. Conversely, B. Lucic has demonstrated that R can present an overly optimistic measure of agreement between predicted values and corresponding experimental values, leading to an excessively optimistic conclusion about the model’s predictive ability [51]. The RMSE provides insights into absolute prediction errors, which could be more informative in practical scenarios compared to R-squared, which focuses on explaining variance. Since our aim is to select the best machine learning model for each cell line, both parameters were taken into consideration.

As Figures 3–7 depict, we can draw the following conclusion: The “best” methodology for determining each of the cell lines’ biological activity is the random forest (RF) [31] methodology. In every instance, it exhibits the highest correlation coefficient across both the training and testing datasets, along with the lowest RMSE. This explains why the RF methodology was chosen as the final model for each of the biological activities, as the RF methodology learns best how to generalize on the given dataset. The “best” models are shown in Supplementary File S31 and saved in Supplementary Files S32–S36, respectively, for each of the cell lines.

Bias in ML algorithms skews results in favor of or against an idea, in either direction. This is a systematic error brought on by false assumptions made throughout the ML learning process. This can have an impact on how an ML model is built [52].

The biological activity, given in  $IC_{50}$  (nM), for the A549 cell line machine learning investigation results are stored in Supplementary Files S15–S17 and S32. Figure 3 depicts the “best” performance of simple machine learning models at the lowest number of features (five features). The two best methodologies observed are DT and RF. It should be stated that the RF model is composed of many DT models. The worst-performing approaches are the other ML methods employed (MLR, KNN, SVM).

The R values (higher values indicate better performance) of the DT and RF ML models on the training datasets are similar, but RF demonstrates superior performance on the test set of data. While other ML models perform adequately on the training data, they significantly underperform on the testing data, as evident from their R values, indicating underfitting [37] to the test data. The RMSE values, depicted in Figure 3, suggest that DT and RF are the most effective methods for predicting the A549  $IC_{50}$  (nM) parameter. These ML models exhibit the lowest RMSE values compared to MLR, KNN, and SVM. Both the R values and RMSE consistently support the use of a random state of 15 in the RF model as the optimal choice. Details of the data splitting for this model can be found in Supplementary File S56.

The molecular descriptors utilized for predicting A549  $IC_{50}$  (nM) values include the following: AMID\_O (average molecular ID on O atoms), EState\_VSA5 (EState VSA Descriptor 5 with a range of 1.17 to less than 1.54), MDEO-12 (molecular distance edge between primary O and secondary O), SaasC (sum of aasC), and VSA\_EState5 (VSA EState Descriptor 5 ( $5.74 \leq x < 6.00$ )) [53]. The rationale behind their selection is detailed in Supplementary Files S15–S17, S31 and S55.

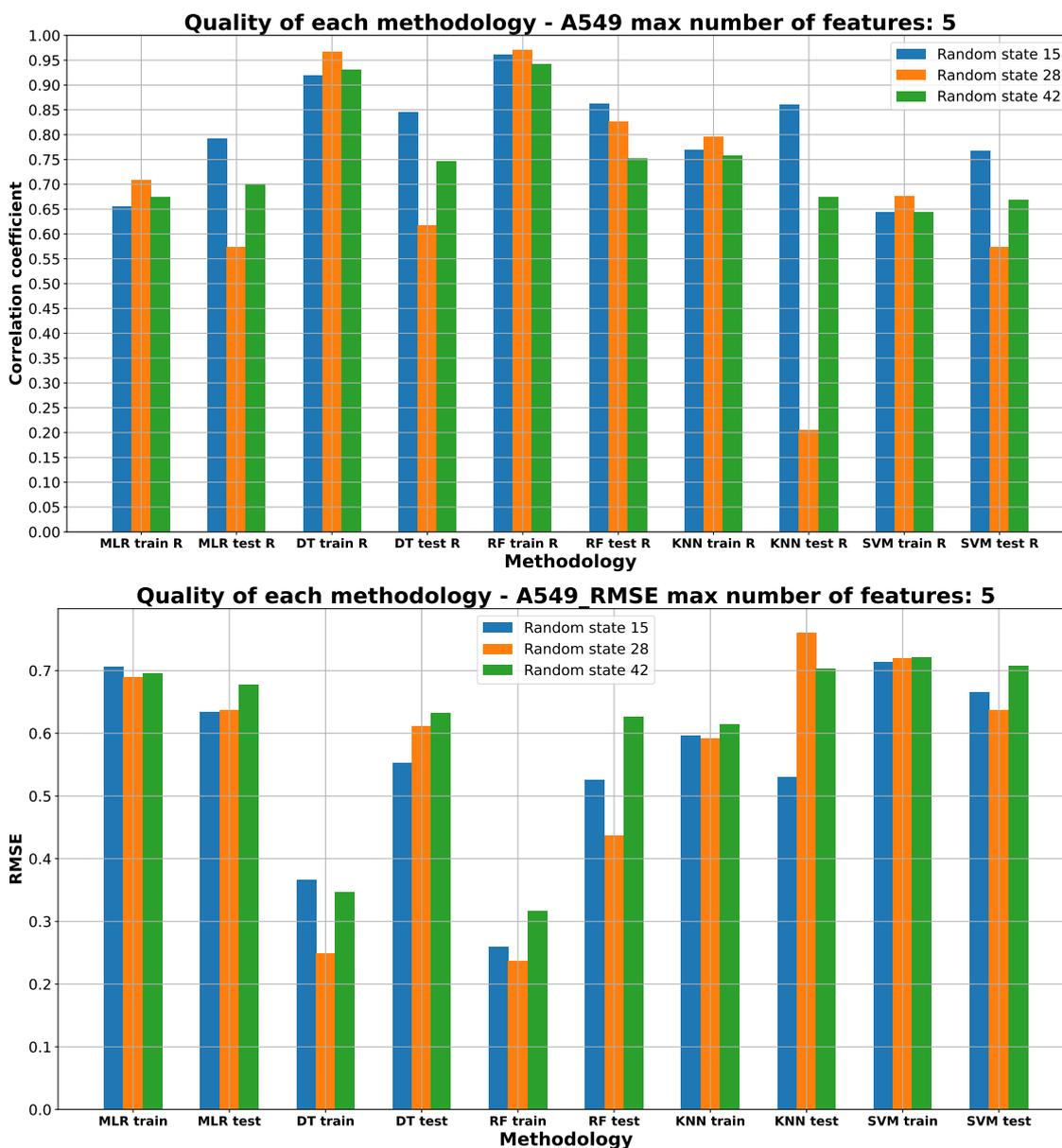


Figure 3. Machine learning models' performance on A549 cell line.

The biological activity, given in  $IC_{50}$  (nM), for the BALB/3T3 cell line machine learning investigation results are stored in Supplementary Files S18–S20 and S33. Figure 4 depicts the “best” performance of simple machine learning models at the lowest number of features (six features). The two best methodologies are observed to be DT and RF. The worst-performing approaches are the MLR, KNN, and SVM machine learning approaches.

The R values for the DT and RF ML models on the training datasets show a comparable performance, with RF slightly outperforming DT on the test set. Other ML models perform reasonably well on the data, but are worse than RF model. The KNN is reflecting the signs of underfitting [37] to the testing data.

The RMSE values presented in Figure 4 highlight that the most effective models for predicting the BALB/3T3  $IC_{50}$  (nM) parameter are DT and RF. These ML models show the lowest RMSE values compared to other methods such as MLR, KNN, and SVM. Both the R values and RMSE consistently endorse the RF model with a random state of 15 as the optimal choice. Detailed information regarding the data splitting for this model is available in Supplementary File S61.

The molecular descriptors used to predict the BALB/3T3  $IC_{50}$  (nM) values are as follows: AMID\_O (averaged molecular ID on O atoms), EState\_VSA5 (EState VSA Descriptor 5 ( $1.17 \leq x < 1.54$ )), GATS2c (Geary coefficient of lag 2 weighted by Gasteiger charge), MDEO-12 (molecular distance edge between primary O and secondary O), NdssC (number of dssC), and VSA\_EState5 (VSA EState Descriptor 5 ( $5.74 \leq x < 6.00$ )) [53]. The rationale behind their selection is detailed in Supplementary Files S18–S20, S31 and S55.

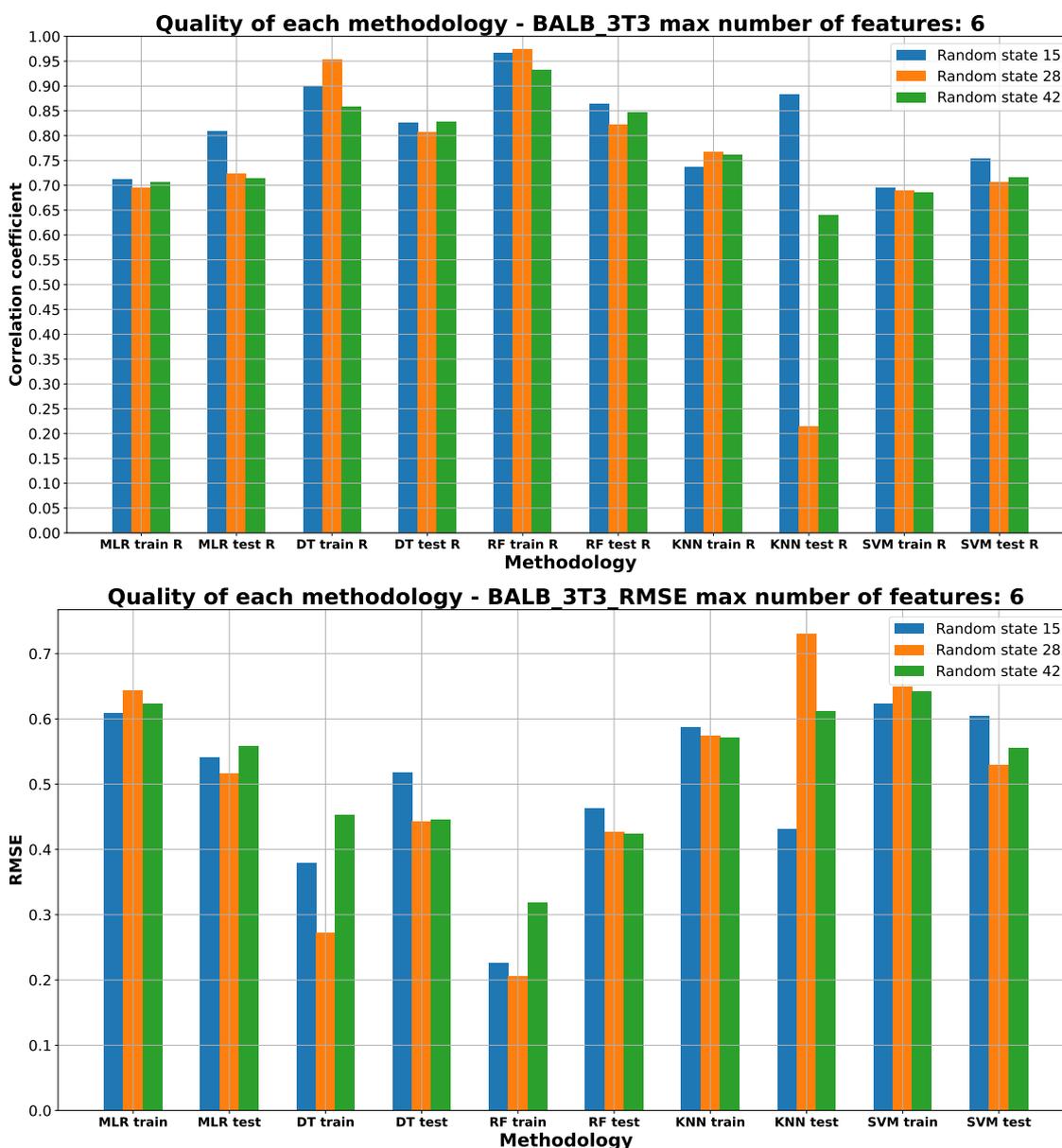


Figure 4. Machine learning models' performance on BALB/3T3 cell line.

The biological activity, given in  $IC_{50}$  (nM), for the LoVo/DX cell line machine learning investigation results are stored in Supplementary Files S21–S23 and S34. Figure 5 depicts the “best” performance of simple machine learning models at the lowest number of features (five features). The two best methodologies are observed to be DT and RF. The DT methodology training correlation coefficient equals 1.00, but this methodology was not chosen as the final one due to the overfitting [37] to the training data. The final method is RF with random state 42, which performs quite well on training and testing data.

The R values of the DT and RF ML models on the training datasets exhibit a similar performance, with RF displaying slightly superior performance on average when evaluated on the test dataset. Notably, the DT models reveal signs of overfitting [37], as indicated by R values of 1.00, leading to comparatively larger errors based on RMSE compared to the RF ML model. Interestingly, the RMSE for the DT training dataset with a random state of 28 is notably low, primarily due to overfitting [37] to the training data.

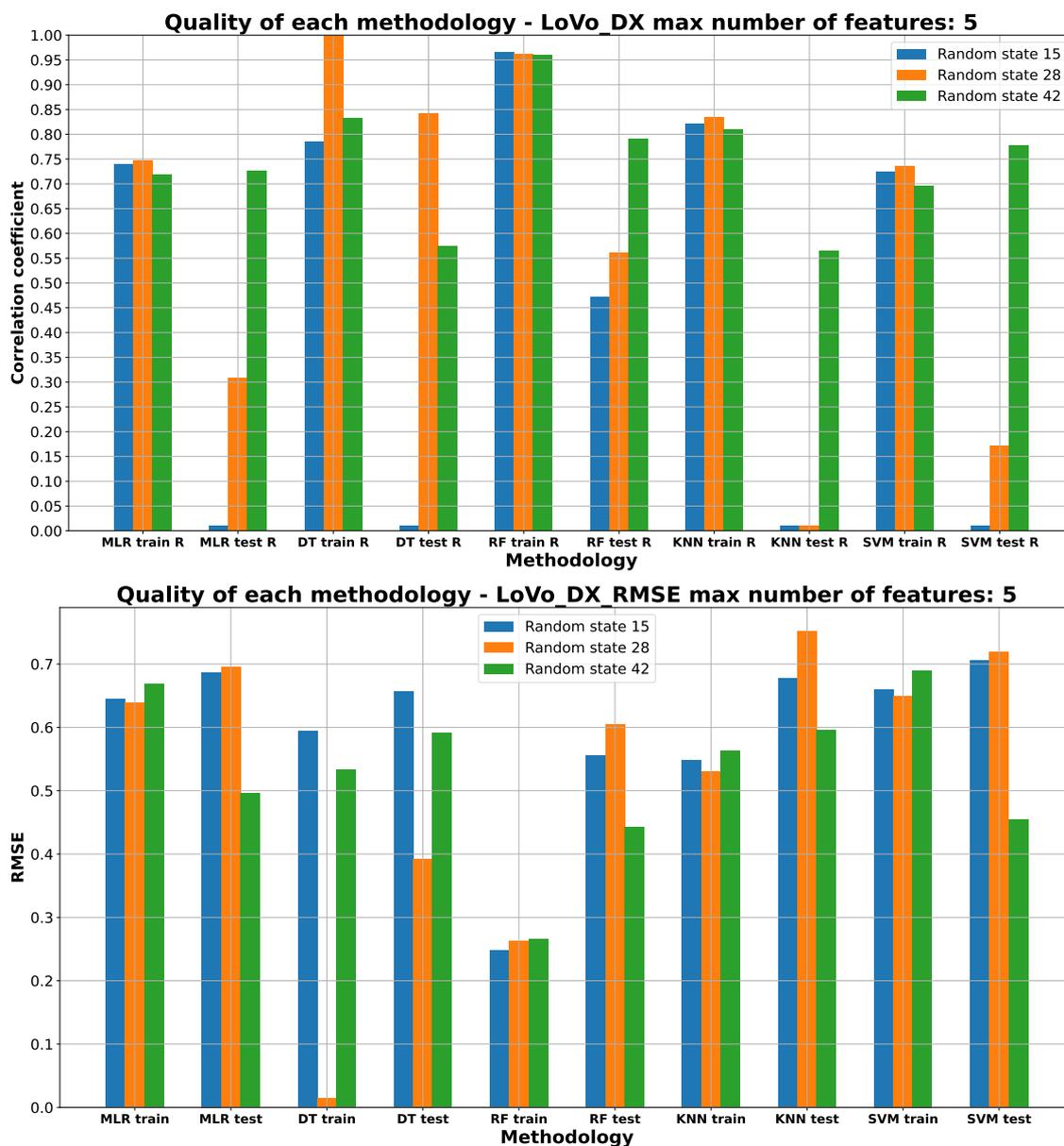


Figure 5. Machine learning models’ performance on LoVo/DX cell line.

While the other ML models perform reasonably well on the training data, they demonstrate notably inferior performance on the testing data, evident from their R values, signifying potential underfitting [37] to the test data. The RMSE values depicted in Figure 5 suggest that DT and RF are the most effective methods for predicting the LoVo/DX  $IC_{50}$  (nM) parameter. These ML models exhibit the lowest RMSE values compared to other models like MLR, KNN, and SVM.

Both the R values and RMSE consistently support the RF model with a random state of 42 as the optimal choice. Detailed information regarding the data splitting for this model can be located in Supplementary File S67.

The molecular descriptors used to predict LoVo/DX  $IC_{50}$  (nM) values are the following: GATS2c (Geary coefficient of lag 2 weighted by Gasteiger charge), MATS2c (Moran coefficient of lag 2 weighted by Gasteiger charge), NdssC (number of dssC), RNCG (relative negative charge), and TopoPSA(NO) (topological polar surface area (use only nitrogen and oxygen)) [53]. The rationale behind their selection is detailed in Supplementary Files S21–S23, S31 and S55).

The biological activity, given in  $IC_{50}$  (nM), for the LoVo cell line machine learning investigation results are stored in Supplementary Files S24–S26 and S35. Figure 6 depicts the “best” performance of simple machine learning models at the lowest number of features (two features). The two best methodologies are observed to be DT and RF. As the performance of the DT and RF methodologies is similar, by our choice, the RF methodology was used to build a final predictive model, as it can cover a wider space of predictive possibilities compared to the simple DT model. The worst-performing approaches are the MLR, KNN, and SVM machine learning approaches.

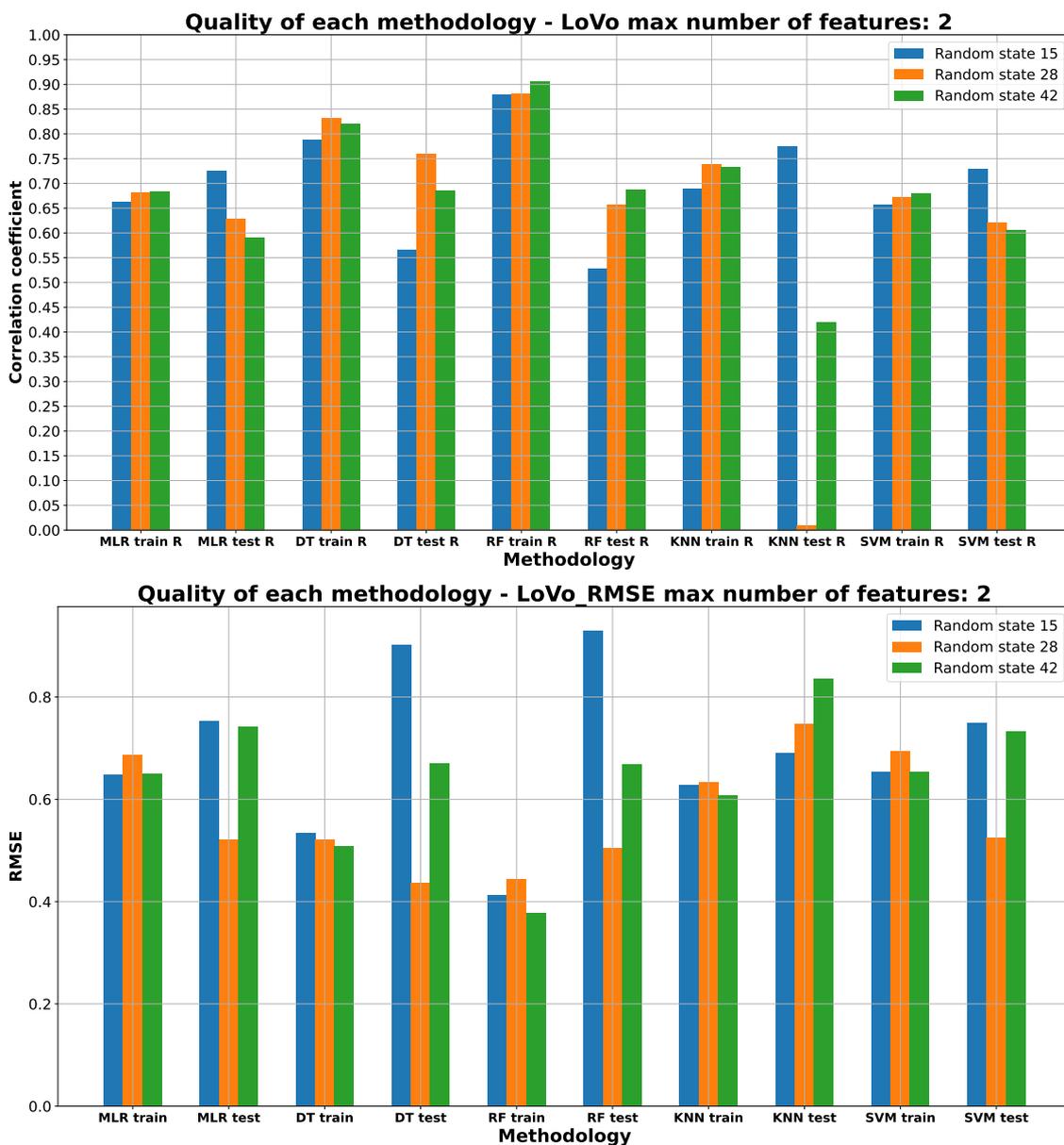
The R values, indicating model performance, show comparable results between the DT and RF ML models on the training datasets, with RF demonstrating a slightly better average performance on the test dataset. Interestingly, while other ML models perform reasonably well on the training data, they notably underperform on the testing data, as indicated by lower R values, hinting at potential underfitting [37] to the test dataset.

In terms of predicting the LoVo  $IC_{50}$  (nM) parameter, the RMSE values displayed in Figure 6 highlight that DT and RF emerge as the most effective methods. These ML models display the lowest RMSE values compared to alternative models such as MLR, KNN, and SVM, signifying their superior predictive accuracy.

The advantage of RMSE is also apparent in Figure 6, where the correlation coefficient alone proves insufficient as a quality parameter. The RMSE indicates that the models are comparable in terms of quality, aligning with its assessment, unlike the sole reliance on the correlation coefficient (R).

Consistently, both the R values and RMSE support the RF model with a random state of 28 as the optimal choice. For more detailed information regarding the data division used for this model, please refer to Supplementary File S63.

The molecular descriptors utilized for predicting the LoVo  $IC_{50}$  (nM) values include the following: EState\_VSA5 (EState VSA Descriptor 5 ( $1.17 \leq x < 1.54$ ), and MDEO-12 (molecular distance edge between primary O and secondary O) [53]. The rationale behind their selection is detailed in Supplementary Files S24–S26, S31 and S55).



**Figure 6.** Machine learning models' performance on LoVo cell line.

The biological activity, given in  $IC_{50}$  (nM), for the MCF-7 cell line machine learning investigation results are stored in Supplementary Files S27–S29 and S36. Figure 7 depicts the “best” performance of simple machine learning models at the lowest number of features (four features).

While the R values for the DT and RF ML models on the training datasets are fairly similar, RF demonstrates superior performance on the average test set. Notably, the performance of other ML models appears robust on the training data but markedly deteriorates on the testing data (KNN), revealing significant underfitting [37] issues, as indicated by lower R values.

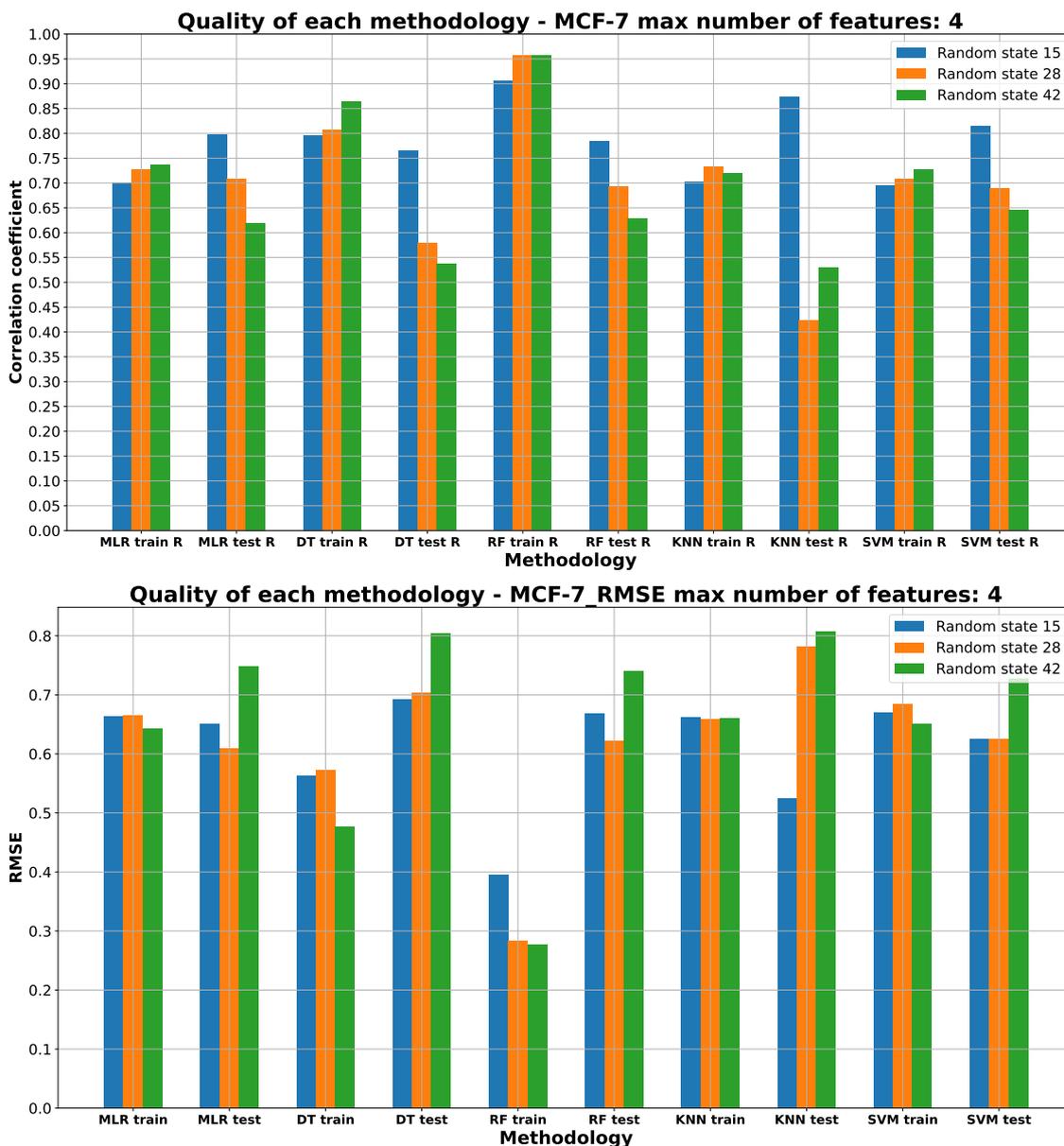
Analyzing the RMSE values depicted in Figure 7, it becomes evident that DT and RF stand out as the optimal methods for predicting the LoVo  $IC_{50}$  (nM) parameter. These ML models exhibit the lowest RMSE values compared to alternative models such as MLR, KNN, and SVM, highlighting their superior predictive accuracy.

The benefit of the RMSE is observable in Figure 7, highlighting the inadequacy of solely using the correlation coefficient as a quality measure. The RMSE indicates comparability

among the models in terms of quality, supporting its evaluation, unlike relying solely on the correlation coefficient (R).

Consistently, both the R values and RMSE endorse the RF model with a random state of 15 as the most effective choice. For a detailed breakdown of the data splitting used for this model, please refer to Supplementary File S68.

The molecular descriptors used to predict MCF-7  $IC_{50}$  (nM) values are the following: AMID\_O (averaged molecular ID on O atoms), EState\_VSA5 (EState VSA Descriptor 5 ( $1.17 \leq x < 1.54$ ), MDEO-12 (molecular distance edge between primary O and secondary O), and EState\_VSA6 (EState VSA Descriptor 6 ( $1.54 \leq x < 1.81$ )) [53]. The rationale behind their selection is detailed in Supplementary Files S27–S29, S31 and S55).



**Figure 7.** Machine learning models' performance on MCF-7 cell line.

The final model quality measurements are presented in Table 2. The table presents the machine learning methodology selected for the final predictive model construction, the number of features used to build the model, the correlation of molecular descriptors to the target data, the overall R score [48] (calculated by asking the model to predict the target value on the whole dataset), the mean squared error (MSE) [54], the mean absolute

error (MAE) [55] and, lastly, the root mean square error (RMSE) [49], all calculated on the whole dataset.

**Table 2.** Predictive model approaches investigated for the prediction of given cell lines' biological activities (Supplementary File S31).

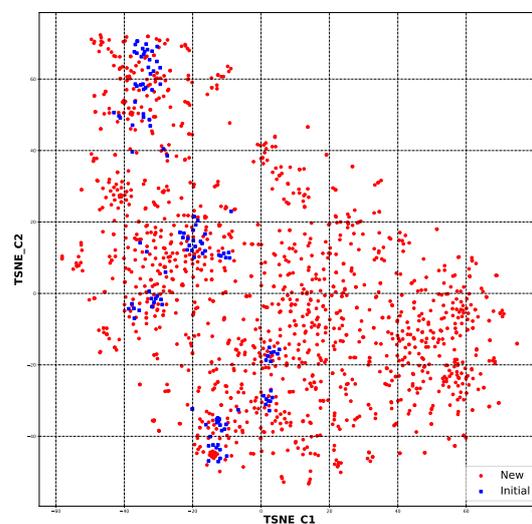
Target	Methodology	Random State	Number of Features	Correlation Threshold	Overall R Score	MSE	MAE	RMSE
A549	RF	15	5	0.51	0.944	0.099	0.205	0.314
BALB/3T3	RF	15	6	0.51	0.950	0.075	0.183	0.274
LoVo/DX	RF	42	5	0.63	0.950	0.089	0.217	0.299
LoVo	RF	28	2	0.54	0.865	0.206	0.320	0.453
MCF-7	RF	15	4	0.51	0.883	0.200	0.258	0.447

#### 2.4. Data Selection

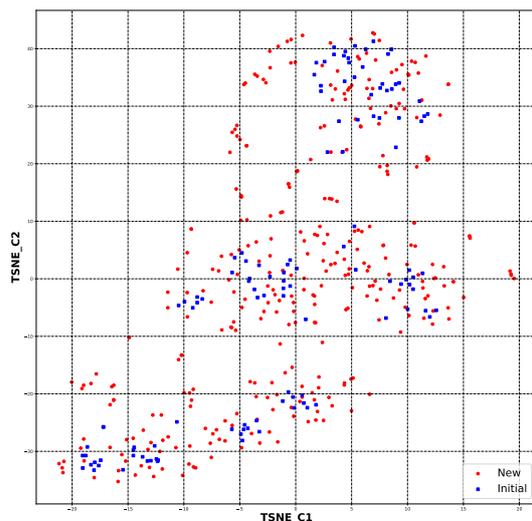
The first step of selection was conducted using Tanimoto similarity [43,56,57] (Section 3.4). This step let us cut out 430 structures that were too different in the meaning of chemical similarity calculated based on molecular fingerprints [56]. A total of 1356 new structures remained after the first step (Supplementary Files S37–S39).

The second step of selection was the employment of the SYBA classifier [42] (Section 3.4). This provided an estimation of the potential difficulty of synthesis. Figure 8 shows the chemical space of the considered species with blue-marked starting structures and red-marked new structures. As can be seen in Figure 9, after the application of the SYBA classifier, only structures more similar to the starting ones remain. This step sieved out 1021 structures, leaving 335 structures (Supplementary Files S40 and S41).

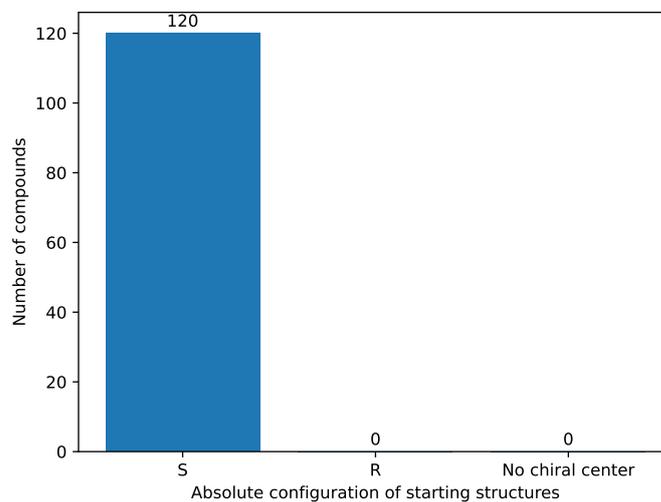
In the final step of the selection process (Section 3.4), we employed a rigorous selection criterion for the absolute configuration of the compounds, as the starting structures have an 'S' absolute configuration on the seventh carbon atom [22–26,58]. Figure 10 provides a visual representation of the distribution of absolute configurations in the initial set of structures. Notably, all of these structures exhibited the 'S' absolute configuration. Subsequently, Figure 11 illustrates the distribution of absolute configurations in the newly generated colchicine-based structures. The application of this selective criterion resulted in the elimination of 50 structures from the initial dataset, reducing the total number of structures under examination to 285, as shown in Supplementary File S42. This stringent selection process ensured that only structures conforming to the desired 'S' absolute configuration on the seventh carbon atom were included in our study.



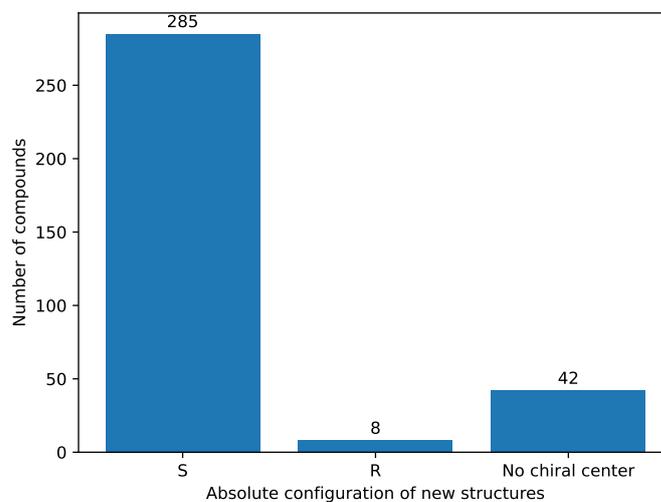
**Figure 8.** The chemical space for all the newly generated structures (1356) and initial ones (120) based on molecular fingerprints (Supplementary File S43).



**Figure 9.** The chemical space for the SYBA-selected newly generated structures (335) and initial ones (120) based on molecular fingerprints (Supplementary File S43).



**Figure 10.** The absolute configuration of starting chemical structures' distribution.

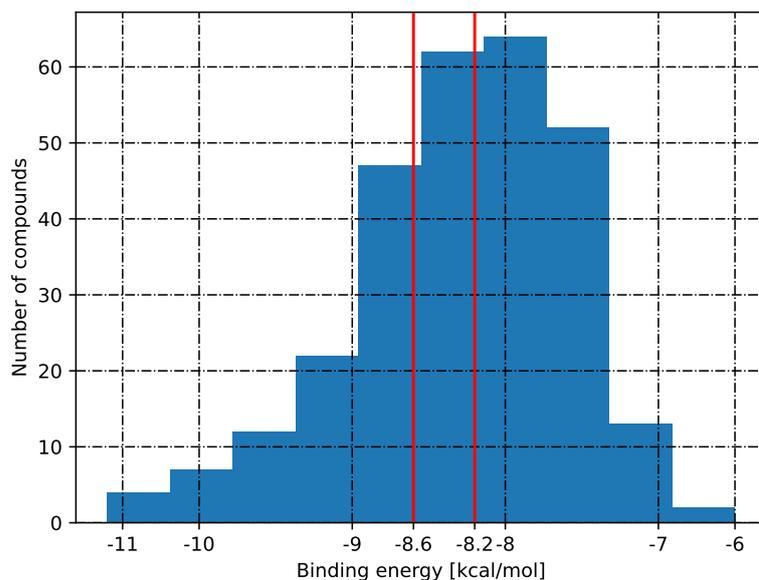


**Figure 11.** The absolute configuration of newly generated chemical structures' distribution.

The PubChem search (Supplementary Files S44–S47) provided information that four structures out of the AI-created colchicine-based structures were found in the PubChem database of known compounds. Their (PubChem Compound Identifications) CIDs were 6351355, 162648725, 164628185, and 162672356, respectively. This indicates that the vast majority of the AI-proposed structures are new, and from them, we can pick the best candidates based on machine learning models and molecular docking. It also shows the generative capabilities of the machine learning models proposed.

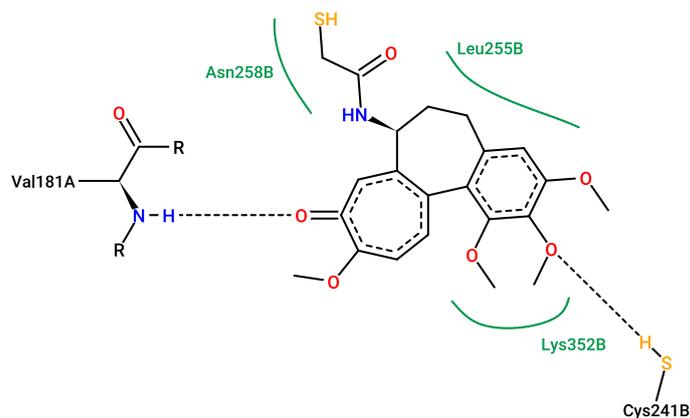
### 2.5. Molecular Docking

Molecular docking studies conducted indicate that the AI-proposed structures have similar affinity to the 1SA0 (PDB ID) protein domain. Figure 12 depicts the distribution of the affinity of selected (285) structures to the protein domain. The first red line shows the raw colchicine affinity ( $-8.2$  kcal/mol) and the second natively present structure in the raw PDB file ( $-8.6$  kcal/mol). The combination of biological activity prediction based on ML models combined with molecular docking can enhance the selection of structures for experimental verification and potentially speed up new bioactive system discoveries. Of the structures, 131 have an affinity greater than  $-8.2$  kcal/mol, and 78 have an affinity lower than  $-8.6$  kcal/mol. The number of structures with an affinity equal to  $-8.6$  kcal/mol is 14, and 15 structures have an affinity equal to  $-8.2$  kcal/mol.

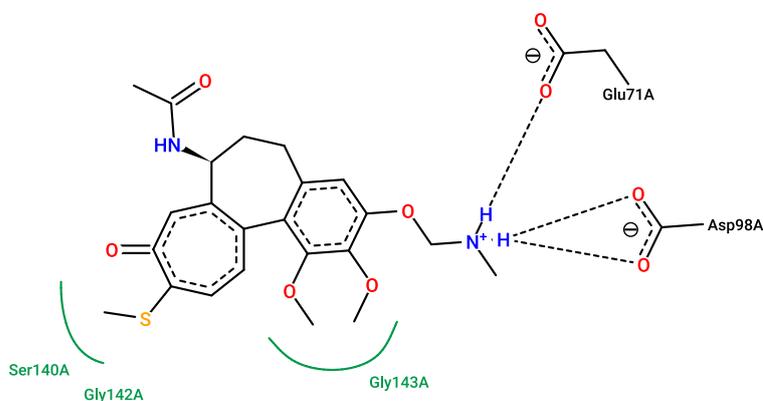


**Figure 12.** The binding energies of 285 structures that were selected from the previous step (Sections 2.4 and 3.4). Red lines indicates calculated binding energy of ligand with the protein target.

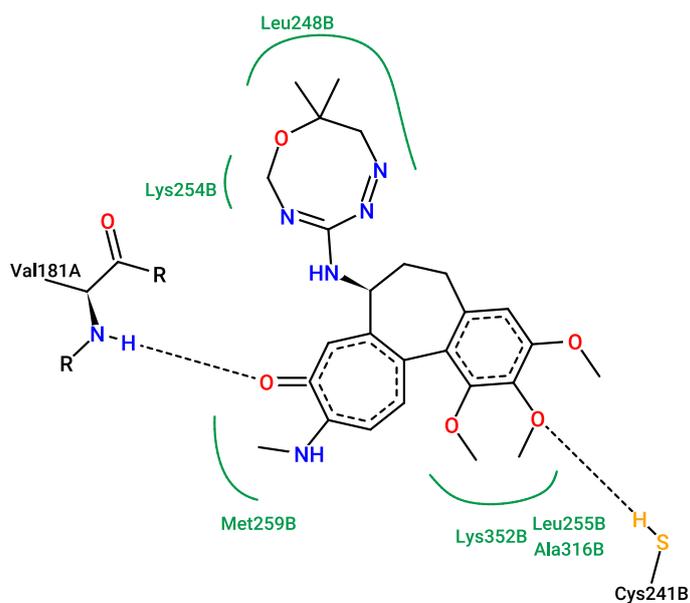
Figures 13–16 show the interactions that can be formed between colchicine-based structures and the binding site of 1SA0. The green solid lines indicate hydrophobic contact, and the black dashed lines indicate possible hydrogen bond formation. All of them are stabilizing ligands in the pocket of the binding site of the protein. Indeed, molecular docking is a good tool to enhance the possibilities of biological activity mechanism explanation [59].



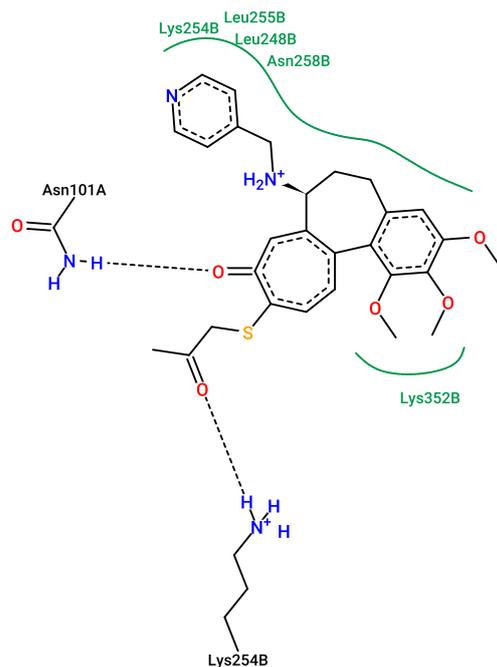
**Figure 13.** Redocked native structure with binding energy of  $-8.6$  kcal/mol. The picture depicts possible interactions between the natively present structure in 1SA0 protein's active site and the active site itself.



**Figure 14.** New structure number 113 with binding energy  $-7.4$  kcal/mol. The picture depicts possible interactions between AI-created structure number 113 and the active site of the 1SA0 protein domain.



**Figure 15.** New structure number 73 with binding energy  $-10.2$  kcal/mol. The picture depicts possible interactions between AI-created structure number 73 and the active site of the 1SA0 protein domain.



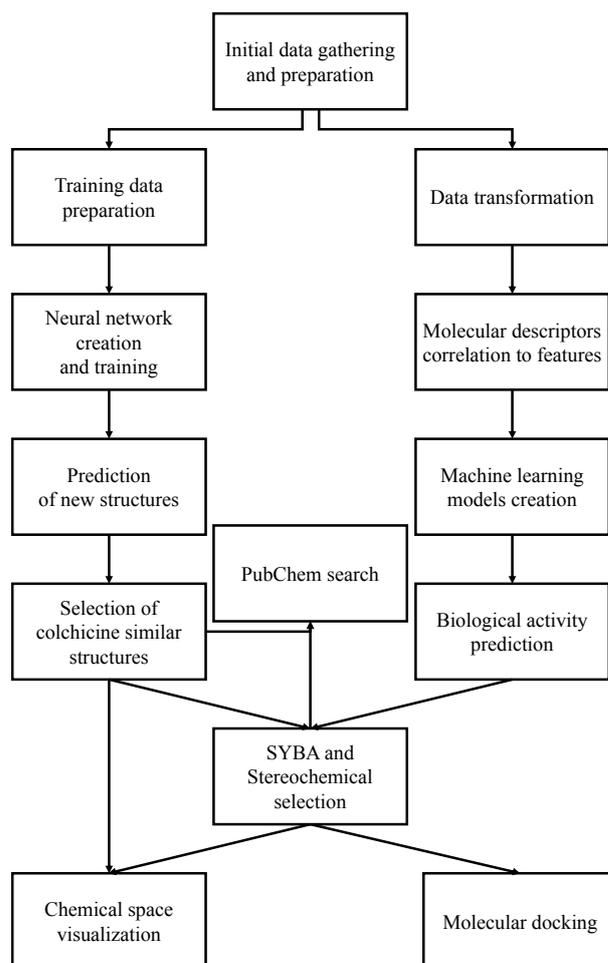
**Figure 16.** New structure number 162 with binding energy  $-8.6$  kcal/mol. The picture depicts possible interactions between AI-created structure number 162 and the active site of the 1SA0 protein domain.

Additional 3D interaction visualizations between ligands from Figures 13–16 can be found in the Supplementary Material, called additional supplementary figures. Figure S1 illustrates the native 1SA0 ligand situated within the protein’s binding site. Figure S2 displays AI-generated ligand number 73 occupying the binding site of the 1SA0 protein. Likewise, Figure S3 exhibits AI-generated ligand 113 also within the identical binding site of the same protein. In a similar vein, Figure S4 portrays AI-created ligand 162 occupying this binding site. Figures S5 and S6 depict all the previously mentioned ligands simultaneously occupying the binding site within the 1SA0 protein domain. The visualizations were performed using the Chimera tool, version 1.16 [60].

Molecular dynamic simulations (MDSs) play a crucial role in advancing molecular biology and facilitating the discovery of new drugs. However, this study did not encompass this aspect, leaving ample space for undeniable enhancements to the presented method.

### 3. Materials and Methods

The files which are mentioned in the manuscript are attached in the Supplementary Materials, and their names are stored in the “Files” attachment. The overall workflow is presented in Figure 17. The files pertaining to the conducted study have been consolidated on GitHub’s platform and are accessible through this link: [https://github.com/XDamianX-coder/Colchicine\\_ML](https://github.com/XDamianX-coder/Colchicine_ML) (accessed on 18 January 2023). All files can also be viewed within the compiled Supplementary Materials, namely, Supplementary code.pdf containing all the code and Jupyter notebooks, and Supplementary Data.xlsx, where the Excel (Version 2312) results are stored. The file numbering has been maintained to ensure the project’s readability.



**Figure 17.** The overall workflow of presented studies.

### 3.1. Training Data

The training data for both the generative neural network creation and creating the machine learning models for anticancer activity prediction were collected from previously published material [22–26]. From this material, we collected 120 structures, SMILES codes [11], with experimentally assigned  $IC_{50}$  activities towards various cell lines, such as A549, BALB/3T3, LoVo/DX, LoVo, and MCF-7. All the data can be viewed in Supplementary File S1.

Based on this collection, we proposed new chemical structures and their estimated activity for each of the evaluated cell lines. It also should be stated that the  $IC_{50}$  parameter unit in this case is given in units of nM for both the starting and newly created structures.

We have created various SMILES [11] representations of each of the colchicine derivatives. This was performed in Supplementary Files S2 and S3 was created as the training dataset, with 118,070 SMILES representations.

### 3.2. Generative Neural Network

The vectorization procedure is necessary for feeding the neural network with the chemical structures' data, as chemical structures are not an easily accessible representation for computers, but vectors are. Vectorization is the process of converting a computer-unreadable representation of data through mathematical processing into computer-readable objects known as mathematical vectors [61]. The format of the chemical representation that was vectorized was the so-called molecular sequence [4], which was derived from the SELFIES [62] depiction of a molecule. This was conducted within Supplementary File S4.

The generative neural network was built with the application of the previously proposed architecture [4,12] with little modification. As we did not have thousands of struc-

tures, we encoded and vectorized the same molecule many times using different SMILES code. This was achieved with the support of the RDKit library [9]. In that way, we were able to create a sufficient number of data points (118,070) for neural network training (106,263) and validation (11,807). This was performed inside Supplementary File S4. The model is stored in Supplementary Files S5–S7. The neural network performance is recorded in Supplementary File S8.

With the application of the neural network, we proposed a number of new structures (Supplementary File S9), which are close in the meaning of the chemical space, as our recurrent neural network (RNN) [1–3,12] learns how to reconstruct vectorized chemical structures. The quality of the model and the loss function were measured using categorical cross-entropy [63]. The structures proposed by the machine learning model are stored in Supplementary Files S10 and S11.

### 3.3. Machine Learning Models for Anticancer Activity

Based on the collected data (120 data points), we proposed very simple models for revealing the biological activity of unknown compounds. The following methodologies were investigated: multiple linear regression (MLR) [29], decision tree (DT) [30], random forest (RF) [31], k-nearest neighbors regression (KNN) [32], and support vector machines (SVMs) [33].

Firstly, the target features, biological activities, were transformed with the following equation:

$$pIC_{50} = -1 \times \log\left(\frac{IC_{50}}{10^9}\right) \quad (1)$$

This was performed to reduce the skewness of the data points (see Supplementary File S12). The transformed data are stored in Supplementary File S13.

Then, the correlation of molecular descriptors [46], calculated with RDKit [9] and Mordred [10], to each of the biological activities was provided in the nM unit of  $IC_{50}$ . This was conducted in Supplementary File S14. It gives us information about the number of features that can be used at certain thresholds of correlation between molecular descriptors and biological activity.

The machine learning methodologies were investigated with three different random seed values, namely, 15, 28, and 42, which were randomly chosen. Each of the models evaluated was tested with different correlation thresholds of biological activity,  $IC_{50}$  values towards certain cell lines:

1. A549 (Supplementary Files S15–S17);
2. BALB/3T3 (Supplementary Files S18–S20);
3. LoVo/DX (Supplementary Files S21–S23);
4. LoVo (Supplementary Files S24–S26);
5. MCF-7 (Supplementary Files S27–S29).

and molecular descriptors. The higher the correlation value, the fewer molecular descriptors were available for model construction. We aimed to construct the simplest models possible for each of the target cell lines.

At the end, the “best” models were chosen based on quality measurements, such as correlation threshold, mean squared error (MSE) [54], mean absolute error (MAE) [55], and root mean square error (RMSE) [49] (Supplementary Files S30 and S30a). Then, the final models were built (Supplementary File S31) and saved for each cell line separately (Supplementary Files S32–S36). The model was used to predict the biological activities of the newly generated structures, and the outputs were recorded (Supplementary File S41).

### 3.4. Data Selection

As the neural network can create a variety of structures, subsequent data selection must be performed. This was performed in the following way:

1. Preservation of structures that are highly similar to the colchicine core. The Tanimoto similarity [43,56,57] threshold value was set to the lowest similarity found among the starting structures to the colchicine core, namely, 0.257 (Supplementary Files S37–S39).
2. The SYBA selection process, as documented in the study by Vorsilak et al. [42], this stage serves the purpose of eliminating structures that could pose challenges during the synthesis process. The SYBA algorithm yields a numerical SYBA score, wherein higher values indicate greater feasibility for molecular synthesis. The algorithm computes SYBA scores for the initial set of structures, and the lowest recorded score, set at 19.48 (documented in Supplementary File S40), is subsequently employed as the threshold for evaluating newly generated structures. The results of this analysis are stored in Supplementary File S41.
3. Stereochemistry [9] selection was performed as the third step; it assumes compounds have the 'S' absolute configuration on the seventh carbon [58]. The 'R' absolute configuration structures do not tend to be biologically active [22–26], thus they are removed from further consideration. The results of this step are stored in Supplementary File S42.
4. The process of RI and SI selection is conducted subsequent to the prediction of  $IC_{50}$  values, as described in Section 3.3. This pivotal step enables the identification and retention of AI-generated colchicine-based structures that satisfy the prerequisites concerning drug resistance (RI) and specificity towards cancer cell lines (SI). The resultant indices are stored within Supplementary File S54.

Based on t-distributed stochastic neighbor embedding (t-SNE) analysis [64], a dimensionality reduction algorithm, the chemical space of the created structures was compared to the initial structures. This approach let us separate data that could not be divided by a straight line, hence the name “nonlinear dimension reduction”. It gave us the opportunity to comprehend high-dimensional information and transfer it into a low-dimensional space. It reduced the size of each molecule's molecular fingerprint and presented further similarities between the new structures and the starting structures (Supplementary File S43).

The generated structures were searched for in the PubChem database with the application of the PubChemPy pythonic library [65]. This gave information about whether the structure generated could be found in the PubChem database (Supplementary Files S44–S47).

### 3.5. Molecular Docking

The molecular docking studies were conducted via the AutoDock Vina [38] solution. The target protein domain was 1SA0 [41] (PDB [66] ID). The natively present structure and the colchicine itself were docked to the same protein domain, so the result of the calculation conducted could be used as a reference. The investigated active site of the protein domain was the same for each of the structures. The search parameters are given here: center  $[x, y, z] = [119.743, 92.779, 10.765]$ , size  $[x, y, z] = [44, 44, 60]$ .

The OpenBabel tool [44,45] was used to create \*.pdbqt files, which are necessary for AutoDock Vina, from 3D structures that were previously created with the application of RDKit [9] functionalities. In this manner, it can be performed automatically for many ligands, rather than manually. The protein domain was prepared with the AutoDock Tools 1.5.7 [67].

The following procedure was conducted for the molecular docking:

1. Raw 1SA0.pdb structure was downloaded from the PDB;
2. Native ligand present inside the pocket was saved separately;
3. 3D structure of raw colchicine was prepared (Supplementary File S48);
4. All the selected new colchicine-based structures were transformed into 3D objects and prepared for molecular docking procedures (Supplementary Files S49 and S50);

5. Molecular docking was conducted (Supplementary File S51), and the results are saved in Supplementary File S52. The visualization of the results is in Supplementary File S53. The final results are stored in Supplementary File S54.

The 2D graphs of the interactions between selected colchicine-based structures have been depicted with the ProteinsPlus web application [68–76], which lets users create 2D maps of protein–ligand interactions efficiently.

#### 4. Conclusions

The proposed approach opens the opportunity to create a library of new colchicine-based compounds with assigned biological activity for each of the investigated cell lines: A549, BALB/3T3, LoVo/DX, LoVo, and MCF-7. The library created can be in vitro investigated for testing our predictive model capabilities. The methodology presented shows that we can create a large library of structures and conduct multi-step selection. We can use different discriminators, such as similarity of compounds, difficulty of synthesis classification, or chirality of compounds created.

These findings suggest a significant potential for the deliberate selection of chemical structures that align with specific criteria. The RI and SI indices, through their computational calculations, can serve as supplementary criteria for the meticulous curation of AI-generated colchicine-based compounds, facilitating their subsequent synthesis and experimental validation.

Our methodology shown here can be used in other quantitative structure–activity relationship (QSAR) studies. In this study, we evaluated various ML approaches: the RF, DT, MLR, KNN, and SVM ML models. Therefore, we could select the best solution for predicting the half-maximal inhibitory concentration value ( $IC_{50}$ ), for five cell lines of the compounds proposed. The created RF models performed quite well with training and testing data, although the distribution was not pure Gaussian. Surprisingly, it was found that some capabilities for the recognition of  $IC_{50}$  patterns were gathered by the RF models for each of the cell lines analyzed.

This model was designed so that it works with colchicine-based compounds, so it has not been used with other structures. The machine learning (ML) models have much higher certainty in the results for more similar structures. If one wanted to predict  $IC_{50}$  values for totally different compounds, these would be less certain. This means that we can be more sure of the models' predictions if the structure we are considering is closer, in chemical space, to the training data. This is the limitation of the model.

**Supplementary Materials:** The following supporting information can be downloaded at: [www.mdpi.com/xxx/s1](http://www.mdpi.com/xxx/s1).

**Author Contributions:** Conceptualization, D.N., A.H. and M.H.; methodology, D.N., R.A.B. and M.H.; software, D.N.; validation, D.N. and R.A.B.; formal analysis, D.N., A.H., R.A.B. and M.H.; investigation, D.N.; resources, D.N. and A.H.; data curation, D.N. and A.H.; writing—original draft preparation, D.N. and M.H.; writing—review and editing, D.N., A.H., R.A.B. and M.H.; visualization, D.N.; supervision, M.H.; project administration, D.N. and M.H.; funding acquisition, R.A.B. and M.H. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** All publication-related information can be accessed at this address: [https://github.com/XDamianX-coder/Colchicine\\_ML](https://github.com/XDamianX-coder/Colchicine_ML) (accessed on 18 January 2023). The Supplementary Materials can also be downloaded using the link provided above. Molecular docking results will be shared on request.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial intelligence
ML	Machine learning
RNN	Recurrent neural network
SMILES	Simplified molecular-input line-entry system
SELFIES	Self-referencing embedded strings
MLR	Multiple linear regression
DT	Decision tree
RF	Random forest
KNN	K-nearest neighbors
SVM	Support vector machines
R	Correlation coefficient
MSE	Mean squared error
MAE	Mean absolute error
RMSE	Root mean square error
PDB	Protein Data Bank
RI	Resistance index
SI	Selectivity index

## References

1. Bjerrum, E.J.; Threlfall, R. Molecular Generation with Recurrent Neural Networks (RNNs). *arXiv* **2017**, arXiv:1705.04612. [[CrossRef](#)]
2. Kotsias, P.C.; Arús-Pous, J.; Chen, H.; Engkvist, O.; Tyrchan, C.; Bjerrum, E.J. Direct steering of de novo molecular generation with descriptor conditional recurrent neural networks. *Nat. Mach. Intell.* **2020**, *2*, 254–265. [[CrossRef](#)]
3. Segler, M.H.S.; Kogej, T.; Tyrchan, C.; Waller, M.P. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent. Sci.* **2018**, *4*, 120–131. [[CrossRef](#)]
4. Nowak, D.; Bachorz, R.A.; Hoffmann, M. Neural Networks in the Design of Molecules with Affinity to Selected Protein Domains. *Int. J. Mol. Sci.* **2023**, *24*, 1762. [[CrossRef](#)]
5. Ranjan, A.; Shukla, S.; Datta, D.; Misra, R. Generating novel molecule for target protein (SARS-CoV-2) using drug–target interaction based on graph neural network. *Netw. Model Anal. Health Inform. Bioinform.* **2022**, *11*, 6. [[CrossRef](#)]
6. Chen, Y.; Wang, Z.; Wang, L.; Wang, J.; Li, P.; Cao, D.; Zeng, X.; Ye, X.; Sakurai, T. Deep generative model for drug design from protein target sequence. *J. Cheminform.* **2023**, *15*, 38. [[CrossRef](#)]
7. Li, J.N.; Yang, G.; Zhao, P.C.; Wei, X.X.; Shi, J.Y. CProMG: Controllable protein-oriented molecule generation with desired binding affinity and drug-like properties. *Bioinformatics* **2023**, *39*, i326–i336. [[CrossRef](#)]
8. Bachorz, R.A.; Pastwińska, J.; Nowak, D.; Karaś, K.; Karwaciak, I.; Ratajewski, M. The application of machine learning methods to the prediction of novel ligands for ROR $\gamma$ /ROR $\gamma$ T receptors. *Comput. Struct. Biotechnol. J.* **2023**, *21*, P5491–P5505. [[CrossRef](#)]
9. Landrum, G. RDKit: Open-Source Cheminformatics Software. 2016. Available online: <https://zenodo.org/records/7415128> (accessed on 16 August 2023).
10. Moriwaki, H.; Tian, Y.S.; Kawashita, N.; Takagi, T. Mordred: A molecular descriptor calculator. *J. Cheminform.* **2018**, *10*, 4. [[CrossRef](#)]
11. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Model.* **1988**, *28*, 31–36. [[CrossRef](#)]
12. Nowak, D.; Babijczuk, K.; Jaya, L.O.I.; Bachorz, R.A.; Mrówczyńska, L.; Jasiewicz, B.; Hoffmann, M. Artificial Intelligence in Decrypting Cytoprotective Activity under Oxidative Stress from Molecular Structure. *Int. J. Mol. Sci.* **2023**, *24*, 11349. [[CrossRef](#)]
13. Kumar, A.; Sharma, P.R.; Mondhe, D.M. Potential anticancer role of colchicine-based derivatives: An overview. *Anti-Cancer Drugs* **2017**, *28*, 250–262. [[CrossRef](#)]
14. Aykul, S.; Martinez-Hackert, E. Determination of half-maximal inhibitory concentration using biosensor-based protein interaction analysis. *Anal. Biochem.* **2016**, *508*, 97–103. [[CrossRef](#)]
15. Foster, K.A.; Oster, C.G.; Mayer, M.M.; Avery, M.L.; Audus, K.L. Characterization of the A549 Cell Line as a Type II Pulmonary Epithelial Cell Model for Drug Metabolism. *Exp. Cell Res.* **1998**, *243*, 359–366. [[CrossRef](#)]
16. Aaronson, S.A.; Todaro, G.J. Development of 3T3-like lines from Balb-c mouse embryo cultures: Transformation susceptibility to SV40. *J. Cell Physiol.* **1968**, *72*, 141–148. [[CrossRef](#)]
17. Grandi, M.; Geroni, C.; Giuliani, F.C. Isolation and characterization of a human colon adenocarcinoma cell line resistant to doxorubicin. *Br. J. Cancer* **1986**, *54*, 515–518. [[CrossRef](#)]
18. Hsu, H.H.; Chen, M.C.; Day, C.H.; Lin, Y.M.; Li, S.Y.; Tu, C.C.; Padma, V.V.; Shih, H.N.; Kuo, W.W.; Huang, C.Y. Thymoquinone suppresses migration of LoVo human colon cancer cells by reducing prostaglandin E2 induced COX-2 activation. *World J. Gastroenterol.* **2017**, *23*, 1171–1179. [[CrossRef](#)]

19. Lee, A.V.; Oesterreich, S.; Davidson, N.E. MCF-7 Cells—Changing the Course of Breast Cancer Research and Care for 45 Years. *JNCI J. Natl. Cancer Inst.* **2015**, *107*, djv073. [CrossRef]
20. Klejborowska, G.; Urbaniak, A.; Maj, E.; Wietrzyk, J.; Moshari, M.; Preto, J.; Tuszyński, J.A.; Chambers, T.C.; Huczyński, A. Synthesis, anticancer activity and molecular docking studies of N-deacetylthiocolchicine and 4-iodo-N-deacetylthiocolchicine derivatives. *Bioorg. Med. Chem.* **2021**, *32*, 116014. [CrossRef]
21. Huczyński, A.; Rutkowski, J.; Popiel, K.; Maj, E.; Wietrzyk, J.; Stefańska, J.; Majcher, U.; Bartl, F. Synthesis, antiproliferative and antibacterial evaluation of C-ring modified colchicine analogues. *Eur. J. Med. Chem.* **2015**, *90*, 296–301. [CrossRef]
22. Czerwonka, D.; Sobczak, S.; Maj, E.; Wietrzyk, J.; Katrusiak, A.; Huczyński, A. Synthesis and Antiproliferative Screening of Novel Analogs of Regioselectively Demethylated Colchicine and Thiocolchicine. *Molecules* **2020**, *25*, 1180. [CrossRef] [PubMed]
23. Krzywik, J.; Aminpour, M.; Maj, E.; Mozga, W.; Wietrzyk, J.; Tuszyński, J.A.; Huczyński, A. New Series of Double-Modified Colchicine Derivatives: Synthesis, Cytotoxic Effect and Molecular Docking. *Molecules* **2020**, *25*, 3540. [CrossRef]
24. Czerwonka, D.; Maj, E.; Wietrzyk, J.; Huczyński, A. Synthesis of thiocolchicine amine derivatives and evaluation of their antiproliferative activity. *Bioorg. Med. Chem. Lett.* **2021**, *52*, 128382. [CrossRef]
25. Krzywik, J.; Nasulewicz-Goldeman, A.; Mozga, W.; Wietrzyk, J.; Huczyński, A. Novel Double-Modified Colchicine Derivatives Bearing 1,2,3-Triazole: Design, Synthesis, and Biological Activity Evaluation. *ACS Omega* **2021**, *6*, 26583–26600. [CrossRef]
26. Krzywik, J.; Maj, E.; Nasulewicz-Goldeman, A.; Mozga, W.; Wietrzyk, J.; Huczyński, A. Synthesis and antiproliferative screening of novel doubly modified colchicines containing urea, thiourea and guanidine moieties. *Bioorg. Med. Chem. Lett.* **2021**, *47*, 128197. [CrossRef]
27. Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B.A.; Thiessen, P.A.; Yu, B.; et al. PubChem 2023 update. *Nucleic Acids Res.* **2023**, *51*, D1373–D1380. [CrossRef]
28. Galton, F. Regression Towards Mediocrity in Hereditary Stature. *J. Anthropol. Inst. Great Br. Irel.* **1886**, *15*, 246. [CrossRef]
29. Jobson, J.D. Multiple Linear Regression. In *Applied Multivariate Data Analysis*; Springer Texts in Statistics; Springer: New York, NY, USA, 1991; pp. 219–398. [CrossRef]
30. von Winterfeldt, D.; Edwards, W. *Decision Analysis and Behavioral Research*; Cambridge University Press: Cambridge, UK, 1986.
31. Ho, T.K. Random decision forests. In Proceedings of the 3rd International Conference on Document Analysis and Recognition, San José, CA, USA, 21–26 August 1995; Volume 1, pp. 278–282.
32. Fix, E.; Hodges, J.L. Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties. *Int. Stat. Rev. Rev. Int. Stat.* **1989**, *57*, 238. [CrossRef]
33. Awad, M.; Khanna, R. Support Vector Regression. In *Efficient Learning Machines*; Apress: New York, NY, USA, 2015; pp. 67–80. [CrossRef]
34. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794. [CrossRef]
35. Choi, J.; Park, S.; Ahn, J. RefDNN: A reference drug based neural network for more accurate prediction of anticancer drug resistance. *Sci. Rep.* **2020**, *10*, 1861. [CrossRef] [PubMed]
36. You, Y.; Lai, X.; Pan, Y.; Zheng, H.; Vera, J.; Liu, S.; Deng, S.; Zhang, L. Artificial intelligence in cancer target identification and drug discovery. *Sig. Transduct. Target Ther.* **2022**, *7*, 156. [CrossRef]
37. Everitt, B.; Skrondal, A. *The Cambridge Dictionary of Statistics*, 4th ed.; Cambridge University Press: Cambridge, UK, 2010.
38. Trott, O.; Olson, A.J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **2010**, *31*, 455–461. [CrossRef]
39. Nguyen, N.T.; Nguyen, T.H.; Pham, T.N.H.; Huy, N.T.; Bay, M.V.; Pham, M.Q.; Nam, P.C.; Vu, V.V.; Ngo, S.T. Autodock Vina Adopts More Accurate Binding Poses but Autodock4 Forms Better Binding Affinity. *J. Chem. Inf. Model.* **2020**, *60*, 204–211. [CrossRef]
40. Plewczynski, D.; Łaźniewski, M.; Augustyniak, R.; Ginalska, K. Can we trust docking results? Evaluation of seven commonly used programs on PDBbind database. *J. Comput. Chem.* **2011**, *32*, 742–755. [CrossRef]
41. Ravelli, R.B.; Gigant, B.; Curmi, P.A.; Jourdain, I.; Lachkar, S.; Sobel, A.; Knossow, M. Insight into tubulin regulation from a complex with colchicine and a stathmin-like domain. *Nature* **2004**, *428*, 198–202. [CrossRef]
42. Voršilák, M.; Kolář, M.; Čmelo, I.; Svozil, D. SYBA: Bayesian estimation of synthetic accessibility of organic compounds. *J. Cheminform.* **2020**, *12*, 35. [CrossRef]
43. Maggiora, G.; Vogt, M.; Stumpfe, D.; Bajorath, J. Molecular Similarity in Medicinal Chemistry: Miniperspective. *J. Med. Chem.* **2014**, *57*, 3186–3204. [CrossRef] [PubMed]
44. O’Boyle, N.M.; Banck, M.; James, C.A.; Morley, C.; Vandermeersch, T.; Hutchison, G.R. Open Babel: An open chemical toolbox. *J. Cheminform.* **2011**, *3*, 33. [CrossRef] [PubMed]
45. Open Babel Development Team. Open Babel. Version: 3.1.1. 2020. Available online: [http://openbabel.org/wiki/Main\\_Page](http://openbabel.org/wiki/Main_Page) (accessed on 14 October 2023).
46. Mauri, A.; Consonni, V.; Todeschini, R. Molecular Descriptors. In *Handbook of Computational Chemistry*; Leszczynski, J., Ed.; Springer: Dordrecht, The Netherlands, 2016; pp. 1–29. [CrossRef]
47. Müller, A.C.; Guido, S. *Introduction to Machine Learning with Python: A Guide for Data Scientists*, 1st ed.; O’Reilly Media, Inc.: Sebastopol, CA, USA, 2016; ISBN 978-1-4493-6941-5.
48. Freedman, D.; Pisani, R.; Purves, R. *Statistics (International Student Edition)*, 4th ed.; WW Norton & Company: New York, NY, USA, 2007.

49. Chai, T.; Draxler, R.R. Root mean square error (RMSE) or mean absolute error (MAE)?— Arguments against avoiding RMSE in the literature. *Geosci. Model Dev.* **2014**, *7*, 1247–1250. [[CrossRef](#)]
50. Chicco, D.; Warrens, M.J.; Jurman, G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Comput. Sci.* **2021**, *7*, e623. [[CrossRef](#)] [[PubMed](#)]
51. Lučić, B.; Batista, J.; Bojović, V.; Lovrić, M.; Sović Kržić, A.; Bešlo, D.; Nadramija, D.; Vikić-Topić, D. Estimation of Random Accuracy and its Use in Validation of Predictive Quality of Classification Models within Predictive Challenges. *Croat. Chem. Acta* **2019**, *92*, 379–391. [[CrossRef](#)]
52. Hellström, T.; Dignum, V.; Bensch, S. Bias in Machine Learning—What is it Good for? *arXiv* **2020**, arXiv:2004.00686. [[CrossRef](#)]
53. Mordred Descriptor List. Available online: <https://mordred-descriptor.github.io/documentation/master/descriptors.html> (accessed on 7 December 2023).
54. Schluchter, M.D. Mean Square Error. In *Wiley StatsRef: Statistics Reference Online*, 1st ed.; Balakrishnan, N., Colton, T., Everitt, B., Piegorisch, W., Ruggeri, F., Teugels, J.L., Eds.; Wiley: Hoboken, NJ, USA, 2014. [[CrossRef](#)]
55. Fürnkranz, J.; Chan, P.K.; Craw, S.; Sammut, C.; Uther, W.; Ratnaparkhi, A.; Jin, X.; Han, J.; Yang, Y.; Morik, K.; et al. Mean Absolute Error. In *Encyclopedia of Machine Learning*; Sammut, C., Webb, G.I., Eds.; Springer: New York, NY, USA, 2011; p. 652. [[CrossRef](#)]
56. Cereto-Massagué, A.; Ojeda, M.J.; Valls, C.; Mulero, M.; Garcia-Vallvé, S.; Pujadas, G. Molecular fingerprint similarity search in virtual screening. *Methods* **2015**, *71*, 58–63. [[CrossRef](#)] [[PubMed](#)]
57. Bajorath, J. Selected Concepts and Investigations in Compound Classification, Molecular Descriptor Analysis, and Virtual Screening. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 233–245. [[CrossRef](#)] [[PubMed](#)]
58. Liu, J.; Gao, R.; Gu, X.; Yu, B.; Wu, Y.; Li, Q.; Xiang, P.; Xu, H. A New Insight into Toxicity of Colchicine Analogues by Molecular Docking Analysis Based on Intestinal Tight Junction Protein ZO-1. *Molecules* **2022**, *27*, 1797. [[CrossRef](#)] [[PubMed](#)]
59. Berdzik, N.; Jasiewicz, B.; Ostrowski, K.; Sierakowska, A.; Szlaużys, M.; Nowak, D.; Mrówczyńska, L. Novel gramine-based bioconjugates obtained by click chemistry as cytoprotective compounds and potent antibacterial and antifungal agents. *Nat. Prod. Res.* **2023**, 1–7. [[CrossRef](#)]
60. Pettersen, E.F.; Goddard, T.D.; Huang, C.C.; Couch, G.S.; Greenblatt, D.M.; Meng, E.C.; Ferrin, T.E. UCSF Chimera—A visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004**, *25*, 1605–1612. [[CrossRef](#)] [[PubMed](#)]
61. Jurafsky, D.; Martin, J.H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*; Prentice Hall Series in Artificial Intelligence; Prentice Hall: Upper Saddle River, NJ, USA, 2000.
62. Krenn, M.; Häse, F.; Nigam, A.; Friederich, P.; Aspuru-Guzik, A. Self-Referencing Embedded Strings (SELFIES): A 100% robust molecular string representation. *arXiv* **2019**, arXiv:1905.13741. [[CrossRef](#)]
63. Murphy, K.P. *Probabilistic Machine Learning: An Introduction*; MIT Press: Cambridge, MA, USA, 2022.
64. van der Maaten, L.; Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
65. Swain, M. PubChemPy. 2017. Available online: <https://github.com/mcs07/PubChemPy/> (accessed on 4 August 2023).
66. RCBS Protein Data Bank. 1971. Available online: <https://www.rcsb.org/> (accessed on 9 August 2023).
67. Morris, G.M.; Huey, R.; Lindstrom, W.; Sanner, M.F.; Belew, R.K.; Goodsell, D.S.; Olson, A.J. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.* **2009**, *30*, 2785–2791. [[CrossRef](#)]
68. ProteinsPlus Development Team. ProteinsPlus. Available online: <https://proteins.plus> (accessed on 14 October 2023).
69. Fährrolfes, R.; Bietz, S.; Flachsenberg, F.; Meyder, A.; Nittinger, E.; Otto, T.; Volkamer, A.; Rarey, M. ProteinsPlus: A web portal for structure analysis of macromolecules. *Nucleic Acids Res.* **2017**, *45*, W337–W343. [[CrossRef](#)]
70. Schöning-Stierand, K.; Diedrich, K.; Fährrolfes, R.; Flachsenberg, F.; Meyder, A.; Nittinger, E.; Steinegger, R.; Rarey, M. ProteinsPlus: Interactive analysis of protein–ligand binding interfaces. *Nucleic Acids Res.* **2020**, *48*, W48–W53. [[CrossRef](#)] [[PubMed](#)]
71. Schöning-Stierand, K.; Diedrich, K.; Ehrt, C.; Flachsenberg, F.; Graef, J.; Sieg, J.; Penner, P.; Poppinga, M.; Ungethüm, A.; Rarey, M. Proteins Plus: A comprehensive collection of web-based molecular modeling tools. *Nucleic Acids Res.* **2022**, *50*, W611–W615. [[CrossRef](#)]
72. Stierand, K.; Rarey, M. Drawing the PDB: Protein–Ligand Complexes in Two Dimensions. *ACS Med. Chem. Lett.* **2010**, *1*, 540–545. [[CrossRef](#)]
73. Stierand, K.; Rarey, M. From Modeling to Medicinal Chemistry: Automatic Generation of Two-Dimensional Complex Diagrams. *ChemMedChem* **2007**, *2*, 853–860. [[CrossRef](#)] [[PubMed](#)]
74. Stierand, K.; Maaß, P.C.; Rarey, M. Molecular complexes at a glance: Automated generation of two-dimensional complex diagrams. *Bioinformatics* **2006**, *22*, 1710–1716. [[CrossRef](#)] [[PubMed](#)]
75. Fricker, P.C.; Gastreich, M.; Rarey, M. Automated Drawing of Structural Molecular Formulas under Constraints. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1065–1078. [[CrossRef](#)]
76. Diedrich, K.; Krause, B.; Berg, O.; Rarey, M. PoseEdit: Enhanced ligand binding mode communication by interactive 2D diagrams. *J. Comput. Aided Mol. Des.* **2023**, *37*, 491–503. [[CrossRef](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.