Benchmarking Feature Selection Algorithms in the Molecular Property Prediction Space

Alejandro Corrochano
dept. Computer Science
University of Gothenburg
Gothenburg, Sweden
guscorral@student.gu.se

Azadeh Karimisefat
dept. Computer Science
University of Gothenburg
Gothenburg, Sweden
guskarimaz@student.gu.se

Dipti Aswal
dept. Computer Science
University of Gothenburg
Gothenburg, Sweden
gusaswdi@student.gu.se

Yossra Gharbi
dept. Computer Science
University of Gothenburg
Gothenburg, Sweden
gusghayo@student.gu.se

***Abstract*—When there is no drug available for a specific disease or clinical condition, medical companies initiate a drug discovery program. Usually, for a drug to be permitted in the market, its pipeline is assessed by a large variety of time- and cost-intensive in vitro and in vivo tests in order to protect humans from presumably harmful effects. Consequently, companies are trying to utilize computational methods to develop more powerful and less time-consuming approaches. Fortunately, with the emergence of Machine Learning (ML), researchers are leveraging their computational infrastructure to build, enhance, and deploy prediction models across the drug discovery process. In this paper, we present the initial stages of drug discovery, discussing some of the possible actions that can be taken to increase the success rate. Subsequently, we cover one of the most common ML approaches in the lead identification stage by applying several models on molecular properties, including lipophilicity and Ames mutagenicity, to predict how potential compounds behave within the organism. This is known as Absorption, Distribution, Metabolism, Excretion, and Toxicity prediction (ADMET). Our objective is to retrieve the features that are more relevant in the molecular property prediction of the compounds, to improve model's interpretability. Among all models utilized, Support Vector Regressor (SVR) and Random Forest Classifier (RFC) exhibited best performance in the lipophilicity and Ames mutagenicity properties prediction, respectively. Besides, the identification of the most relevant features boosted the model efficacy and reduced the complexity of the model.**

***Index Terms*—Drug discovery, Ames mutagenicity, Lipophilicity, Machine Learning, Feature Extraction, in vitro, in vivo.**

I. Introduction

A drug discovery program initiates mostly because there is a disease or some clinical condition that has not been yet covered by any medical solution [1]. To successfully fulfil this underlying motivation, the drug must go through many tests and clinical trials before it is launched. As all stages are tedious, costly and highly susceptible to failure, it is vital for companies to reduce the probability of failure among them. With the emergence of Machine Learning (ML), researchers are able to bring down the failure rates across all areas, specifically, in drug discovery field. The molecular properties' prediction of potential drug candidates, for instance, can help avoid the repetition of clinical trials due to inaccurate assumptions about these properties [2]. The first part of this paper briefly introduces the steps followed in the drug discovery process along with the molecular properties specifically, lipophilicity and Ames mutagenicity that will be later predicted to identify potential drug candidates. Modeling selected properties and evaluating each model based on recent ML algorithms is the next step, which is then followed by several feature extraction methods. Finally, we outline how feature extraction can affect model performance in this field.

II. Background theory

A. *Drug Discovery*

Bioinformatics' approaches can help researchers to identify, select and prioritize potential disease targets based on the information extracted from publications, patent information, gene expression data, proteomics data, transgenic phenotyping and compound profiling data [1]. However, the initial stages in the drug discovery pipeline are usually time-consuming and cannot be considered as routine activity. Therefore, the quality of the starting point and the expertise of the team members are the key determinants of a successful outcome of this phase. On an average, for each project 200,000 to more than 10^6 compounds are screened and during the next step, around 100 compounds are selected for pre-clinical testing and finally, only one or two candidate molecules enter clinical trials. It is estimated that only 10% of small molecule projects within industry progress from pre-clinic stages to being a potential candidate and the toxicity of compounds is one of the key reasons of failures at this step [1].

*1) **Target Identification and Validation**:* Targets are biological units, such as proteins or genes, that hold a strong relationship with a certain disease. Selecting targets based on available evidence is commonly named as target identification which is the initiation of the drug discovery process. In this phase, drugs should not only fulfil safety requirements but also be efficacious against the target. In the current stage, genetic, cellular and in vivo mechanisms (experiments on living organism setups) are used to identify and validate targets. Identification might include mRNA examination to assess whether they are present in a disease or somehow responsible for the advancement of the disease. Another common approach is to study the genetic associations and

identify the role of gene polymorphism in the progression of a disease [1].

Validation techniques involve using in-vitro approaches, where experiments are conducted with microorganisms, cells, or molecules outside of their normal environments. Transgenic, i.e., genetically modified animals, are also used for this purpose. Recent studies have found that chemical genomics, which is the study of genomic responses to chemical compounds, have found to be efficient in the high-speed discovery of novel drugs and drug targets [1].

*2) **The Hit Discovery**:* Compound screening assays are developed during the hit identification and lead discovery phase. Any compounds which has the desired activity in a compound screen and whose activity is confirmed upon retesting is detected at this stage [1]. There are several screening paradigms amongst researchers to identify hit molecules, such as High Throughput Screening (HTS), Fragment screening and Focused, or knowledge-based screening. Specifically, HTS screens the entire compound library against the target, hence involving the use of complex laboratory automation [1].

The Hit discovery begins with the development of biological assays and a posterior screening of compound libraries. These screens are executed to identify molecules that interact with the drug target. The pharmaceutical companies aim to identify targets, assemble compound collections and the associated infrastructure for compound screening and hit molecules and ultimately optimize those screening Hits into clinical candidates [1].

B. ADMET

The *Absorption, Distribution, Metabolism, Elimination, and Toxicity* properties of drug candidates are important for their efficacy and safety as therapeutics. Up to half of clinical trial failures have been attributed to discrepancies in ADMET properties [3]. Therefore, an increased focus on the optimization of ADMET properties, along with potency and selectivity can lead to the reduction in failure rates [4].

Predicting ADMET properties have become essential in computational chemistry in recent times. Parameters such as lipophilicity, solubility, and metabolic stability can be measured in a high throughput manner in vitro and are therefore used as early ADMET screens [5]. As for a more precise definition, lipophilicity measures the ability of a drug to dissolve in a lipid (e.g. fats, oils) environment. High values often leads to a higher rate of metabolism, poor solubility, high turn-over, and low absorption. On the other hand, Ames mutagenicity is the ability of a drug to induce genetic alterations. Such drugs can cause damage to the DNA and can even result in cell death or other severe adverse effects.

C. Drug Discovery Challenges

The process from Hit generation to pre-clinical candidate selection is time-consuming and cannot be considered as a routine activity. Moreover, intellectual inputs are required from scientists at each step. The quality of starting point and the expertise of the available team are the key factors for a successful outcome at this phase. Per studies, only 10% of small molecule projects within industry become potential candidates, and failures can arise due to the problems such as (i) configuring a reliable assay; (ii) no developable outputs from the HTS; (iii) different behaviour of compounds in secondary or native tissue assays; (iv) toxicity of compounds in in vitro or in vivo [1]. In addition, as Hughes et al. [1] discussed, once a drug candidate reaches the clinical stage and becomes public, the termination of the project can affect company's reputation. Since the risk of failure at each step of the process is high, prior studies before clinical development, such as molecule identification, may help companies to handle these challenges.

D. Feature Selection

Identifying the most relevant features at the beginning of ML process can facilitate solving the problem more efficiently by enabling the algorithm to train faster. It also reduces the complexity of a model hence improves the interpretability [6]. As outlined in [7], to reduce the dimensionality of the input space, it is required to first, identify all the relevant features, determine the correlations between them and then choose a minimum subset of features or alternative subsets to maximize the efficiency.

III. Method

A. Dataset and Data Pre-processing

Therapeutics Data Commons (TDC) is an open-science platform with ML-ready dataset and learning tasks for therapeutics, spanning the discovery and development of safe and effective medicines. Data in TDC are in a Simplified Molecular-Input Line-Entry System (SMILES) encoding to allow users to represent a chemical structure in a string format thus, making it easy for a computer to deal with. The dataset employed in the project consists of 4200 and 7278 compounds to predict lipophilicity and Ames mutagenicity molecular properties respectively. Each of them is split into train, validation, and test sets according to a 70:10:20 ratio given by default. Additionally, these steps were carried out to pre-process the data:

- Standardize compounds
- Remove compounds containing null feature values
- Remove compounds with less than 5 heavy atoms
- Remove duplicate compounds

As for the lipophilicity dataset, there was only one compound containing null values and another with less than 5 heavy atoms. On the other hand, in the Ames mutagenicity dataset 20 duplicated atoms were identified which shared the same label.

B. Input Data

The first step in molecular ML is encoding the structure of the molecules in a form that is amenable to ML. By doing

this, we can perform statistical analysis and computational techniques on the set of molecules to gain new insights.

*1) **Descriptors**:* Molecular descriptors are the mathematical representations of molecules' properties that are generated by algorithms [8]. These numerical features are used to quantitatively describe the physical and chemical information of the molecules. For instance, the LogP, which is a quantitative representation of the lipophilicity of the molecules, is obtained by measuring the partitioning of the molecule between an aqueous phase and a lipophilic phase which consists usually of water/n-octanol. The molecular descriptors help in predicting the ADMET properties of molecules and can be classified as one-dimensional (1D), two-dimensional (2D), or three-dimensional (3D) descriptors.

*2) **Fingerprints**:* Molecular Fingerprints are representations of chemical structures which were essentially designed to assist in chemical database substructure searching. These structures are later used for analysis tasks, such as similarity searching, clustering, and classification. Extended-Connectivity Fingerprints (ECFPs) are a recently developed fingerprint methodology explicitly designed to capture molecular features relevant to molecular activity [9].

C. Baseline Model

In this project, we evaluate regression and classification models based on the ***lipophilicity*** and ***Ames mutagenicity*** of compounds before and after feature selection. For the classification task, which is about classifying compounds as being toxic (having Ames mutagenicity) and nontoxic (without Ames mutagenicity), we have investigated Support Vector Classifier (SVC) [10], Multi-layer Perceptron classifier (MLPClassifier) [11], Random Forest Classifier (RFC) [12], XGBoost Classifier (XGBC) [13], and Decision Tree Classifier (DTC) [14] models individually at each step. On the other hand, for predicting lipophilicity, models including XGBoost Regression (XGBR), Multi-layer Perceptron Regressor (MLPRegressor), Random Forest Regressor (RFR), Decision Tree Regressor (DTR), and Support Vector Regressor (SVR) have been used.

*1) **Multi-layer Perceptron**:* Is a kind of feed-forward Artificial Neural Network (ANN) and generally solves the XOR problem by using network design. This model can be applied for both classification and regression problems in a supervised learning approach and it is sometimes colloquially referred to as "vanilla" neural networks, especially when it has a single hidden layer.

*2) **Random Forest**:* Is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. This model is very helpful for tabular data however, data characteristics can affect their performance. In classification, the RFC returns the class selected by the majority of individual trees, and in regression, the RFR returns the mean prediction of the trees.

*3) **Decision Tree**:* Is a tree-like model and decision support tool which is commonly used in operations research, specifically in decision analysis, but is also a popular tool in ML. It is useful for tabular data and can be applied either in regression as DTR or classification as DTC problems.

*4) **XGBoost**:* Is a decision-tree-based ensemble ML algorithm that uses a gradient boosting framework. In prediction problems involving unstructured data (images, text, etc.) ANNs tend to outperform all other algorithms or frameworks. However, when it comes to small-to-medium structured/tabular data, decision tree based algorithms are considered best-in-class.

*5) **Support Vector Machine (SVM)**:* Is a supervised learning model based on statistical learning frameworks and one of the most robust prediction methods. When a classification task wants to be performed, SVC is employed. It tries to maximize the gap between different regions, which differs in the label. Once this is done, it maps unseen data into regions and returns the predicted category/label for each data point. However, for regression tasks, SVR is performed. In this case, a line or hyperplane in higher dimensions, will be found to fit the training data based on some loss error threshold.

D. Feature Selection Methods

In this report, three feature selection methods were evaluated. These methods return the most important features for the both classification and regression models.

*1) **Minimum Redundancy Maximum Relevance (MRMR)**:* This method seeks to find a set of features giving the best possible predictive performance based on the reasonable trade-off between relevance and redundancy. It selects a subset of features having most relevance with the target class and less redundancy with other features.

*2) **Relief**:* In comparison with other algorithms, Relief method assumes the dependencies between features and tries to estimate the quality of attributes based on how well the attribute can distinguish between instances that are near to each other. It is a filter based algorithm that does not search through all feature combinations but uses the concept of the nearest neighbours to derive feature statistics hence faster than the other methods.

*3) **Mutual Info Regression/Classification (MIR/MIC)**:* This algorithm calculates the mutual information between two variables and measures the reduction in uncertainty for one variable provided the value of the other variable is known. In simple terms, it is a measure of mutual dependence between different features.

E. Model Assessment

*1) **Regression**:*

- **R Squared ($\mathbf{R}^2$)**
  R^2 also known as coefficient of determination, is a measure of variance for a dependent variable explained by the independent variables. It is similar to the correlation coefficient, R but correlation explains the strength of the relationship between an independent and dependent

variable, whereas R^2 describes the variance between two variables.

$$1 - \sum \frac{(y_i - \hat{y}_i)^2}{(y_i - \bar{y}_i)^2} \tag{1}$$

- ***Mean Squared Error (MSE)***
  It is an estimator which measures the average of the squares of the errors in regression models. Here, error signifies the difference between the estimated and the actual values. This measurement is usually used for evaluating the performance of the regression models. Generally, it is always a positive value with the error decreasing as the error approaches zero.

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2}$$

- ***Mean Absolute Error (MAE)***
  It is a measurement for comparing predicted versus observed values in regression tasks. All absolute errors between paired observations expressing the same phenomenon are summed up and divided by the number of existing phenomenon in order to assess the prediction model.

$$\frac{\sum_{i=1}^{n}|y_i - x_i|}{n} \tag{3}$$

*2) **Classification**:*

- ***Building Blocks of Confusion Matrix:***
  Considering two classes as '*Positive*' or '*Negative*' and predictions as '*False*' or '*True*' in the confusion matrix:
  TP (True Positive): The number of "Positive" class which is classified "correctly"
  TN (True Negative): The number of "Negative" class which is classified "correctly"
  FP (False Positive): The number of "Negative" class which is classified "incorrectly"
  FN (False Negative): The number of "Positive" class which is classified "incorrectly"

- ***Precision:***
  Is the proportion of relevant results in the list of all returned predicted results. This is measured by fraction of positive classes which are classified correctly over all positive classes predicted by the model.

$$\frac{TP}{TP + FP} \tag{4}$$

- ***Recall:***
  Refers to the percentage of total relevant results correctly classified by the model. This is done by the fraction of positive classes which are classified correctly over all existing positive samples.

$$\frac{TP}{TP + FN} \tag{5}$$

- ***Accuracy:***
  Measures how the model could classify both classes correctly. It is a commonly used criteria for model performance evaluation.

$$\frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

- ***Matthews correlation coefficient (MCC):***
  Despite the accuracy, this criterion can be used even if the classes are of very different sizes.

$$\frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{7}$$

- ***Area Under the Curve (AUC):***
  Measures the ability of a classifier to distinguish between classes. The higher the AUC, the better the performance of the model would be, which means that the model could distinguish the positive and negative classes very well.

IV. Results

A. Lipophilicity

*1) **Descriptors**:*

- ***Baseline Model:***
  First, all the descriptors of each compound are fed into the different models. The goal is to investigate the performance of the baseline models after basic data cleaning techniques. To do that, we plot a grouped bar chart. Each model is evaluated by three different metrics (Fig 1).

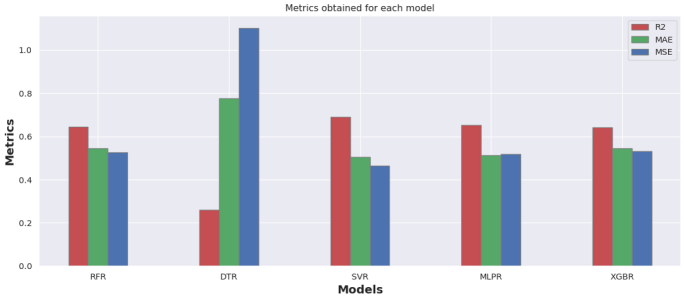

Fig. 1. The comparison of five baseline models for descriptors in lipophilicity

From the results we obtained after fitting the models on the training set and evaluated on the validation set, we can think of eliminating the DTR since it performs poorly compared to the other models and conserving SVR as our final model as it outperforms other models. However, advanced data cleaning can dramatically change the results. Thus, we refine the data and compare the models.

- ***Manual Data Cleaning:***
  In this second step, for the purpose of decreasing the computational cost and improving the models performance, we remove all zero-variance features and the

highly correlated features with each other and with our target. This is an important step since numerous ML algorithms are affected by correlated features. Removing those features will reduce unnecessary bias. Fig 2 highlights the effect of manual data cleaning on each model. Although the number of features has already

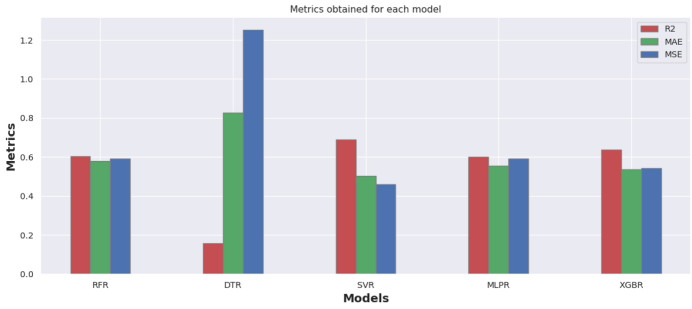

Fig. 2. The comparison of five models for descriptors in lipophilicity after manual data cleaning

dropped remarkably leading to lower computational cost of modelling, we see no major improvement in the results after applying basic techniques to reduce the number of features. Therefore, more advanced methods are needed to explore the sets of features that contribute most to our prediction variables. This is the third step.

- ***Advanced Feature Selection Algorithms:***
  In this third step, we deployed multiple feature selection algorithms and each time we calculated MSE and R^2 metrics of each baseline model based on the features selected by each algorithm. Then, we plotted the MSE and R^2 scores of the best model against the number of features by each feature selection algorithm. This process of identifying the subset of features that are most relevant in predicting our target variable is an essential step in improving the quality of our models.

  MRMR: Each baseline model has been the subject of the MRMR algorithm investigation. Table 1 shows the results of the performance of each baseline model after applying the feature identification by MRMR.

TABLE I
THE PERFORMANCE OF BASELINE MODELS AFTER APPLYING MRMR

	R^2	MAE	MSE
DFR	0.64	0.55	0.54
DTR	0.20	0.80	1.19
SVR	0.70	0.50	0.46
MLPR	0.62	0.55	0.57
XGBR	0.66	0.53	0.50

SVR is the model that has the best performance metrics with R2 value equal to approximately 0.69. We also wanted to investigate further the behavior of MSE and R^2 of SVR in terms of the number of features selected by MRMR. The result is shown in Fig 3.

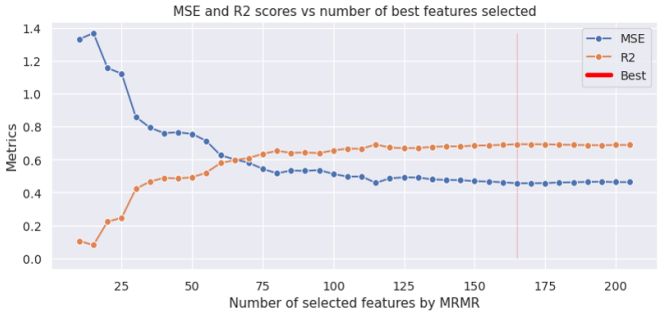

Fig. 3. The behavior of MSE and R^2 of the model SVR in terms of the number of selected features by MRMR

We can see that the linear curve of R2 is increasing rapidly in the interval 0 to approximately 70, then it becomes almost constant even though the number of selected features is increasing. MSE is inversely proportional to R2. 165 is the number of features that induces SVR optimal capability in prediction.

Relief: Each baseline model has been the subject of the Relief algorithm investigation. Table 2 is the results of the performance of the baseline models after applying the feature identification by Relief.

TABLE II
THE PERFORMANCE OF BASELINE MODELS AFTER APPLYING RELIEF

	R^2	MAE	MSE
DFR	0.65	0.55	0.53
DTR	0.20	0.80	1.19
SVR	0.70	0.49	0.45
MLPR	0.68	0.52	0.48
XGBR	0.67	0.52	0.49

SVR is also the model that has the best performance metrics with R2 value equal to approximately 0.7. According to Fig 4, there is a slight improvement in the performance of the SVR model with the Relief feature selection. 165 is the number of features that induces SVR optimal prediction performance.

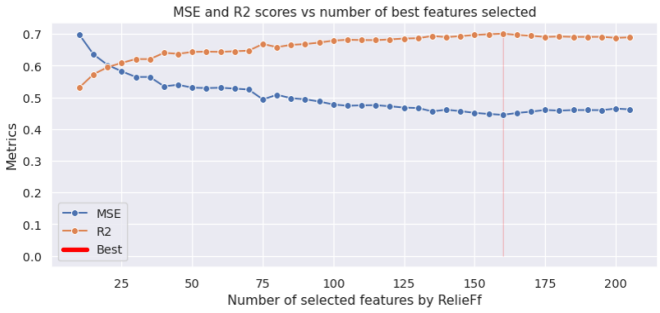

Fig. 4. The behavior of MSE and R^2 of the model SVR in terms of the number of selected features by Relief

MIR: Each baseline model has been the subject of the Mutual Information Regression algorithm investigation. Table 3 describes the results of the performance of each baseline model after applying the features identification by the Mutual Information Regression.

TABLE III
THE PERFORMANCE OF BASELINE MODELS AFTER APPLYING MIR

	R^2	MAE	MSE
DFR	0.64	0.55	0.53
DTR	0.24	0.76	1.13
SVR	0.69	0.51	0.46
MLPR	0.64	0.55	0.54
XGBR	0.65	0.54	0.52

SVR is again the model that has the best performance metrics with R^2 value equal to 0.69. There is no improvement in the quality of the model by MIR (Fig 5). SVR performs best with a number of features equal to 125.

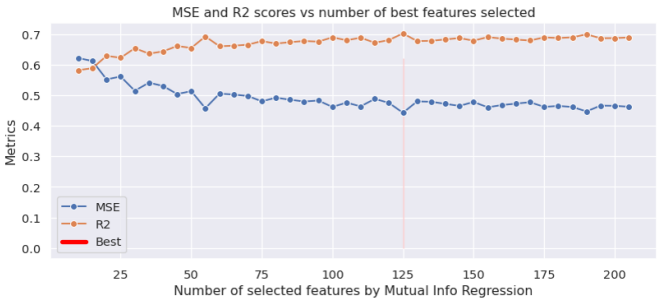

Fig. 5. The behavior of MSE and R^2 of the model SVR in terms of the number of selected features by MIR

Finally, we plotted the performance of these three feature selection methods to select the best of them for the final evaluation (Fig 6 and Fig 7).

Fig. 6. Comparative performance of the utilized feature selection algorithms in terms of R^2 and number of features

- ***Hyperparameter Tuning:***
  Based on the previous results, we can conclude that SVR is the best fit for the descriptors dataset. The evaluation metric R^2 varies from 0.69 to 0.7 depending on the feature selection algorithm. The next step is hyperparameters tuning. The pipeline consists of fitting the model to the training data with different combinations of hyperparameter values and checking how it performs on the validation data by looking at the R^2 and MSE metrics until the performance is satisfactory. This pipeline is repeated on the SVR model with different features that are selected by each feature selection algorithm. Table 4 marks that the best configuration for SVR is post-hyperparameter tuning and the features are selected by Relief.

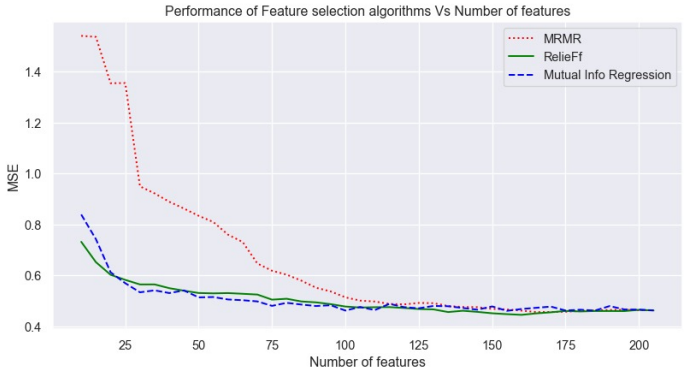

Fig. 7. Comparative performance of the utilized feature selection algorithms in terms of MSE and number of features

TABLE IV
COMPARATIVE RESULTS OF THE UTILIZED FEATURE SELECTION ALGORITHMS BEFORE VERSUS AFTER HYPERPARAMETER TUNING (AHT) ON THE SVR MODEL.

	NO. of features	R2	MSE
Baseline	207	0.69	0.46
MRMR	165	0.69	0.46
AHT MRMR	165	0.75	0.37
RELIEF	160	0.70	0.45
AHT RELIEF	160	0.75	0.37
MIR	125	0.68	0.47
AHT MIR	125	0.74	0.39

- ***Evaluation of the Final Model:***
  Now that the final model is ready, it is evaluated on the test dataset. A test set is utilized for an unbiased evaluation of the model. It mimics unseen real world data and gives an idea on how well our model will perform. The majority of the points are close to the regressed diagonal line although some prediction values have distantly deviated from actual values (Fig 8). Overall, R^2 is equal to 0.72, which is a moderate result.

*2) **Fingerprints**:* Unlike the descriptors dataset, there is no improvement that we can add to the fingerprints in order to enhance the performance of the model since each molecule is represented by a sequence of 0s and 1s. We simply fed the dataset to the models. Each model adjusts its parameters based on the training set to best predict the target variable

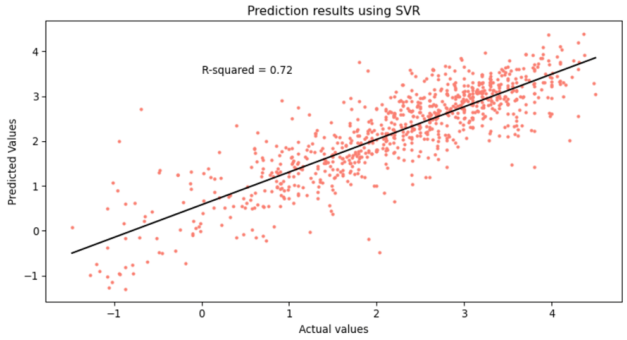

Fig. 8. Scatter plot of actual values versus predicted values

critically on unseen data. As the Fig 9 shows, the results obtained from the prediction on the validation set underlines that SVR is the best model performed on the fingerprint dataset with an R^2 equal to 0.56.

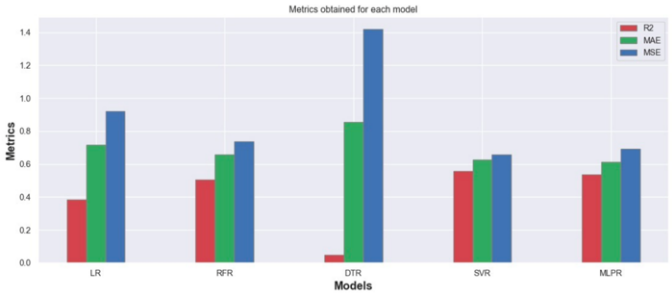

Fig. 9. The comparison of five baseline models for fingerprints in lipophilicity

B. *Ames Mutagenicity*

The same pipeline applied to lipophilicity is used for Ames mutagenicity. The only difference is that lipophilicity is a regression task, whereas Ames mutagenicity is a classification task.

*1) **Descriptors**:*

- ***Baseline Model:***
  All descriptors are fed into the different models after basic data cleaning. The models are performing equally well on the training dataset. There is an insignificant difference in their evaluation metrics. Random Forest Classifier is roughly the best model among the baseline models (Fig 10).
- ***Manual Data Cleaning:***
  There is no major improvement in the results after removing the highly correlated features (Fig 11).
- ***Advanced Feature Selection Algorithms:***
  We deployed the same feature selection algorithm used in lipophilicity on the Ames dataset and each time we calculated MCC, AUC, accuracy, precision, and recall metrics of each baseline model based on the features selected by each algorithm. Then, we plotted the MCC and AUC scores of the best model against the number of features by each feature selection algorithm.

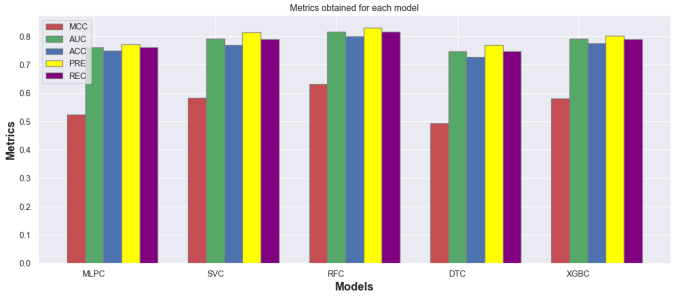

Fig. 10. The comparison of five baseline models for descriptors in Ames mutagenicity

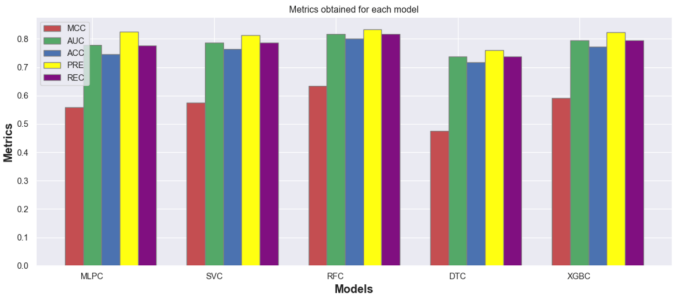

Fig. 11. The comparison of five models for descriptors in Ames mutagenicity after manual data cleaning

MRMR: Each baseline model has been the subject of the MRMR algorithm investigation. Table 5 outlines the results of the performance of each baseline model after applying the feature identification by MRMR. RFC is the model that slightly has the best performance metrics compared to other baseline models.

TABLE V
THE PERFORMANCE OF BASELINE MODELS AFTER APPLYING MRMR

	MCC	AUC	Precision	Recall	Accuracy
MLPC	0.54	0.77	0.77	0.75	0.77
SVC	0.52	0.76	0.73	0.81	0.76
RFC	0.65	0.82	0.81	0.83	0.82
DTC	0.53	0.76	0.75	0.77	0.76
XGBC	0.62	0.81	0.79	0.83	0.81

Relief: Each baseline model has been the subject of the Relief algorithm investigation. Table 6 presents the results of the performance of the baseline models after applying the feature identification by Relief. RFC is also the model that has the best performance metrics. There is a slight improvement in the performance of the SVR model with the Relief feature selection.
We notice that RFC performs practically the same with a number of features superior to 90 (Fig 13).

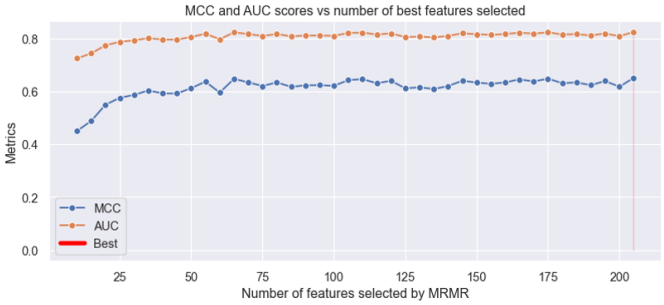

Fig. 12. The behavior of MCC and AUC of the model RFC in terms of the number of selected features by MRMR

TABLE VI
THE PERFORMANCE OF BASELINE MODELS AFTER APPLYING RELIEF

	MCC	AUC	Precision	Recall	Accuracy
MLPC	0.60	0.80	0.78	0.81	0.80
SVC	0.57	0.79	0.76	0.81	0.78
RFC	0.67	0.83	0.81	0.85	0.83
DTC	0.48	0.74	0.71	0.78	0.74
XGBC	0.63	0.82	0.80	0.83	0.82

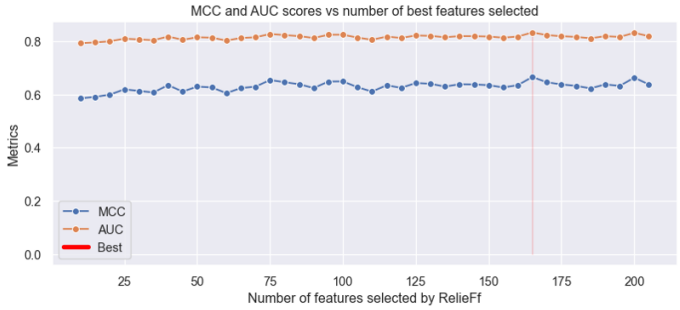

Fig. 13. The behavior of MCC and AUC of the model RFC in terms of the number of selected features by Relief

MIC: Each baseline model has been the subject of the MIC algorithm investigation. The results of the performance of each baseline model after applying the feature identification by the MIC are shown in Table 7. RFC is again the model that has the best performance metrics. There is no improvement in the quality of the model by MIC. RFC performs practically the same with a number of features higher than 95 (Fig 14).

TABLE VII
MODELS PERFORMANCE AFTER THE MIC FEATURE IMPORTANCE

	MCC	AUC	Precision	Recall	Accuracy
MLPC	0.58	0.79	0.75	0.84	0.79
SVC	0.58	0.79	0.77	0.82	0.79
RFC	0.66	0.83	0.81	0.84	0.83
DTC	0.47	0.73	0.72	0.74	0.73
XGBC	0.60	0.80	0.79	0.81	0.80

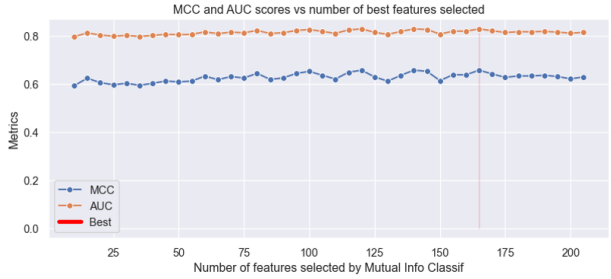

Fig. 14. The behavior of MCC and AUC of the model RFC in terms of the number of selected features by MIC

- ***Hyperparameter Tuning:***
  Based on the previous results, we can conclude that RFC is the best fit for the descriptors dataset although there is no significant difference compared to the other candidate models. The next step is hyperparameters tuning. The tweaking of the hyperparameters is repeated on the RFC model with the different features that are selected by each feature selection algorithm. From the results shown in Table 8, the best configuration for RFC is without hyperparameter tuning and the features are selected by Relief.

TABLE VIII
COMPARATIVE RESULTS OF THE UTILIZED FEATURE SELECTION ALGORITHMS BEFORE VERSUS AFTER HYPERPARAMETER TUNING (AHT) ON THE RFC MODEL

	No. of features	MCC	ACC
Baseline	207	0.63	0.82
MRMR	65	0.65	0.82
AHT(MRMR)	65	0.64	0.82
RELIEF	165	0.67	0.83
AHT(RELIEF)	165	0.64	0.82
Mutual Info Regression	165	0.66	0.83
AHT(MIC)	165	0.64	0.82

- ***Evaluation of the Final Model:***
  Now that the final model is ready, it is evaluated on the test dataset. In the case of Ames mutagenicity, RFC has successfully managed to predict the target variable on the test set with an accuracy equal to approximately 0.9. The confusion matrix in Fig 15 also shows the performance of the final model. The number of uncorrected predictions is low compared to the correct predictions.

2) ***Fingerprints****:* Here, we simply use the fingerprints dataset. No feature selection steps are required. The results obtained from the prediction on the validation set show that SVC is the best performing model on the fingerprint dataset (Fig 16).

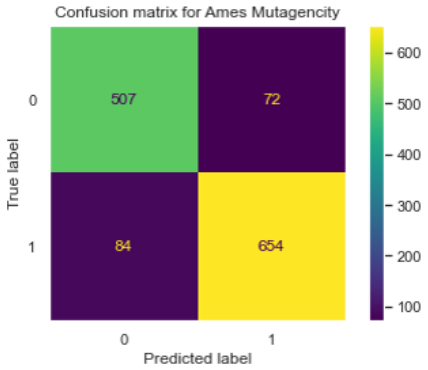

Fig. 15. Confusion matrix from the test set

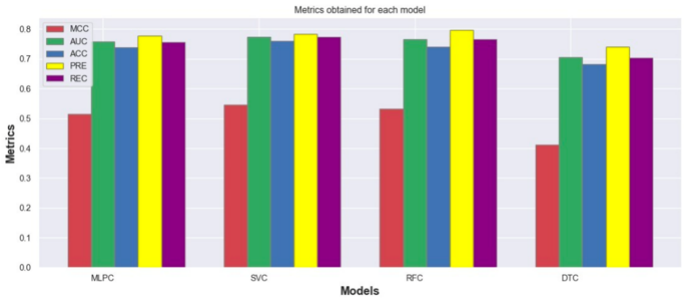

Fig. 16. Metrics results from each model

V. Conclusions

Designing drugs is a multi-parameter optimization task in which several characteristics of compounds including safety, pharmacokinetics and efficacy must be improved to yield drug candidates [15]. In particular, ADMET properties of compounds need to be carefully considered from the beginning as a part of the screening process. Fortunately, recent advances in ML provide numerous opportunities to increase efficiency in the drug discovery pipeline.

In this project, we were able to exclude molecules with poor physicochemical properties as part of screening process in drug discovery. Taking advantage of the feature selection methods have allowed to build models focusing on the most relevant features. For predicting lipophilicity, we applied LR, MLRegressor, RFR, DTR, and SVR models, whereas for Ames mutagenicity classification, we deployed LR, MLP-Classifier, RFC, DTC, and SVC models. Consequently, we scored the promising models for each task and extracted the appropriate features based on MRMR, Relief and Mutual Information Regression/Classifier. SVR and RFC were the best models for lipophilicity and Ames mutagenicity respectively. Eventually, we tested the efficacy of the best selected model for each task on the test dataset. The results obtained reflects the performance of the models.

In conclusion, accurate prediction of ADMET properties using computational methods has proven to be a promising alternative approach instead of employing intensive time- and cost-consuming experimental methods in drug discovery. This can not only prevent inspection of unwanted chemical regions but also open doors to rather desirable areas thereby making pre-clinical discovery more efficient and constructive. Therefore, drug efficacy should be applied extensively as early as possible. However, while it is true that computational methods have the potential to screen large numbers of compounds in a short span of time and at low costs to yield accurate insights, there are often unintended consequences, primarily related to interpretability and transparency. Feature selection techniques were therefore utilized to address such issues.

References

[1] J. Hughes, "Principles of early drug discovery," *British Journal of Pharmacology*, vol. 162, no. 6, pp. 1239–1249, 2011. [Online]. Available: www.brjpharmacol.org

[2] J. Shen, "Molecular property prediction: recent trends in the era of artificial intelligence," *Drug Discovery Today: Technologies*, vol. 32, pp. 29–36, 2019. [Online]. Available: https://doi.org/10.1016/j.ddtec.2020.05.001

[3] S. Winiwarter, E. Ahlberg, E. Watson, I. Oprisiu, M. Mogemark, T. Noeske, and N. Greene, "In silico adme in drug design - enhancing the impact," *ADMET and DMPK*, vol. 6, p. 15, 03 2018.

[4] E. N. Feinberg, "Improvement in admet prediction with multitask deep featurization," *Journal of Medicinal Chemistry*, vol. 63, p. 88358848, 2020. [Online]. Available: https://dx.doi.org/10.1021/acs.jmedchem.9b02187

[5] L. AP., "Screening for human adme/tox drug properties in drug discovery," *Drug Discov Today*, vol. 7, pp. 357–366, 04 2001.

[6] J. Shen, "Molecular property prediction: recent trends in the era of artificial intelligence, drug discovery today: Technologies, volumes 32–33," *Drug Discovery Today: Technologies*, vol. 32-33, pp. 29–36, 2019.

[7] Y. Liu, "A comparative study on feature selection methods for drug discovery," *Journal of Chemical Information and Computer Sciences*, vol. 44, no. 5, pp. 1823–1828, 2004, pMID: 15446842. [Online]. Available: https://doi.org/10.1021/ci049875d

[8] B. Chandrasekaran, "Computer-aided prediction of pharmacokinetic (admet) properties," vol. 2, pp. 731–755, 2018. [Online]. Available: https://doi.org/10.1016/B978-0-12-814421-3.00021-X

[9] D. Rogers, "Extended-connectivity fingerprints," *Journal of Chemical Information and Modeling*, vol. 50, pp. 742–754, 2010. [Online]. Available: https://doi.org/10.1021/ci100050t

[10] T. Evgeniou and M. Pontil, "Support vector machines: Theory and applications," vol. 2049, 01 2001, pp. 249–257.

[11] F. Murtagh, "Multilayer perceptrons for classification and regression," *Neurocomputing*, vol. 2, no. 5, pp. 183–197, 1991. [Online]. Available: https://doi.org/10.1016/0925-2312(91)90023-5

[12] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston, "Random forest: a classification and regression tool for compound classification and qsar modeling," *Journal of Chemical Information and Computer Sciences*, vol. 43, no. 6, pp. 1947–1958, 2003, pMID: 14632445. [Online]. Available: https://doi.org/10.1021/ci034160g

[13] T. Chen and C. Guestrin, "Xgboost," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug 2016. [Online]. Available: http://dx.doi.org/10.1145/2939672.2939785

[14] V. Podgorelec, P. Kokol, B. Stiglic, and I. Rozman, "Decision trees: An overview and their use in medicine," *Journal of medical systems*, vol. 26, pp. 445–63, 11 2002.

[15] L. D. Pennington and I. Muegge, "Holistic drug design for multiparameter optimization in modern small molecule drug discovery," *Bioorganic Medicinal Chemistry Letters*, vol. 41, p. 128003, 2021. [Online]. Available: https://doi.org/10.1016/j.bmcl.2021.128003