Machine Learning the Redox Potentials of Phenazine Derivatives: A Comparative Study on Molecular Features

Siddharth Ghule*[1], Sayan Bagchi[1], Kumar Vanka*[1]

[1]*CSIR-National Chemical Laboratory, Pune, India*

ABSTRACT

Electricity generation is a major contributing factor for greenhouse gas emissions. Energy storage systems available today have a combined capacity to store less than 1% of the electricity being consumed worldwide. Redox Flow Batteries (RFBs) are promising candidates for green and efficient energy storage systems. RFBs are being used in renewable energy systems, but their widespread adoption is limited due to high production costs and toxicity associated with the transition-metal-based redox-active species. Therefore, cheaper and greener alternative organic redox-active species are being investigated. Recent reports have shown organic molecules based on phenazine are promising candidates for redox-active species in RFBs. However, the large number of available organic compounds makes the conventional experimental and DFT methods impractical to screen thousands of molecules in a reasonable amount of time. In contrast, machine-learning models have low development time, short prediction time, and high accuracy; thus, are being heavily investigated for virtual screening applications. In this work, we developed machine-learning models to predict the redox potential of phenazine derivatives in DME solvent using a small dataset of 185 molecules. 2D, 3D, and Molecular Fingerprint features were computed using readily available and easy-to-use python libraries, making our approach easily adaptable to similar work. Twenty linear and non-linear machine-learning models were investigated in this work. These models achieved excellent performance on the unseen data (i.e., $R^2 > 0.98$, MSE < 0.008 V^2 and MAE < 0.07 V). Model performance was assessed in a consistent manner using the training and evaluation *pipeline* developed in this work. We showed that 2D molecular features are most informative and achieve the best prediction accuracy among four feature sets. We also showed that often less preferred but relatively faster linear models could perform better than non-linear models when the feature set contains different types of features (i.e., 2D, 3D, and Molecular Fingerprints). Further investigations revealed that it is possible to reduce the training and inference time without sacrificing prediction accuracy by using a small subset of features. Moreover, models were able to predict the previously reported promising redox-active compounds with high accuracy. Also, significantly low prediction errors were observed for the functional groups. Although some functional groups had only one compound in the training set, best-performing models could achieve errors (MAPE) less than 10%. The major source of error was a lack of data near-zero and in the positive region. Therefore, this work shows that it is possible to develop accurate machine-learning models that could potentially screen millions of compounds in a short amount of time with a small training set and limited number of easy to compute features. Thus, results obtained in this report would help in the adoption of green energy by accelerating the field of materials discovery for energy storage applications.

*Corresponding authors:
Siddharth Ghule (ss.ghule@ncl.res.in), Kumar Vanka (k.vanka@ncl.res.in)

INTRODUCTION

Today, ~85% of the World's Energy demand is being fulfilled by Fossil fuels [1], [2]. The limited supply of fossil fuels and the ever-increasing population has raised concerns that we might run out of fossil fuels sooner than expected [1], [3]. Furthermore, electricity production from fossil fuels is one of the major factors responsible for greenhouse gas emissions [4]. In this age, humanity faces two significant challenges of balancing increased energy demand while reducing the environmental impact associated with energy production. In past decades, investments and research efforts in green technology have been increased to overcome these challenges [5]. Significant progress has already been made to access renewable energy sources [6], [7]. Renewable energy sources being intermittent require efficient energy storage [4]. Improvements in the energy storage technology would not only help in the adoption of renewable energy but also help in making efficient use of non-renewable energy sources. Historically, it has been more expensive to store energy than to expand energy generation for handling increased demand [8]. Thus, grid systems employed today are likely to fail when additional energy could not be generated during peak demand. Massive Texas Blackout in February, 2021 is an example of such failure [9]. It suggests that efficient energy storage technology is urgently required. Unfortunately, only 1% of the energy consumed worldwide could be stored with the energy storage technology accessible today [10]. Furthermore, the contribution of electrochemical batteries to energy storage capacity is less than 2%, even though most of the devices we use every day include batteries [8], [10]. Li-ion batteries are widely used today due to their high energy density, high specific energy, long cycle life, fast charge-discharge cycle [4], [8], [11]. Unfortunately, Li-ion batteries suffer from high production costs, safety, and environmental impact [2], [12]. Redox Flow Batteries (RBFs) have the potential to overcome drawbacks of Li-ion batteries owing to their high storage capacity, independent control over storage capacity and power, fast responsiveness, ease of scaling, room temperature operation, cost-effectiveness, high round trip efficiency, safety, and negligible environmental impact [13]–[15]. RBFs are increasingly being used as the energy storage devices in renewable energy systems, thereby helping in the adoption of green energy [15], [16]. A schematic diagram of the typical redox flow battery is shown in Figure 1. Redox Flow Battery (RFB) consists of two storage tanks containing cathode and anode redox-active species dissolved in an electrolyte solution. Electrolyte solution in the positive and negative compartment is termed catholyte and anolyte, respectively. These storage tanks are connected to an electrochemical cell (or current collector) via pumps. The electrochemical cell consists of porous electrodes separated by an ion-selective membrane. During operation, electrolytes containing redox-active species are pumped to the electrochemical cell, where redox-active species undergo either oxidation or reduction depending on the charge/discharge cycle. Then, electrolytes are circulated back to their storage tanks [13], [17]. So far, transition metal-based redox flow batteries (e.g., vanadium, iron, and chromium) have found some commercial success. However, their widespread adoption has been limited mainly due to high production cost, toxicity, and cell component corrosion associated with the use of transition metal salts [18], [19]. Therefore, redox flow batteries containing organic redox-active species are being heavily investigated due to their low production cost, access to a massive space of electroactive compounds, and low environmental impact [19], [20]. Many organic compounds such as quinones, viologens, flavins, thiazines, imides, and their derivatives have been investigated for redox-active species in both aqueous and non-aqueous RFBs [18], [21], [22]. However, non-aqueous RFBs offer large operating voltage [21]. Recently, Phenazine derivatives have shown

to be promising candidates for high-voltage, high-density redox-active species in non-aqueous RFBs. Therefore, phenazine derivatives are currently being investigated for novel redox-active species [18], [23].

These investigations remain primarily experimental. Unfortunately, the vast chemical space offered by organic compounds cannot be explored using experimental procedures. Quantum mechanical DFT computations have been used heavily in chemistry research due to high accuracy but are very slow and cannot screen millions of molecules in a reasonable amount of time. Therefore, a fast and reliable method to screen millions of compounds without compromising accuracy is required. In this regard, Machine-learning algorithms have shown excellent predictive accuracies along with short prediction time [24]–[28]. Therefore, Machine learning models have been used extensively to screen millions of molecules in materials science and drug discovery [29]–[33]. Machine learning models generally require a large amount of data for accurate prediction. When the quantity of data is limited, feature engineering is employed to generate the most informative features. These features are expected to capture the appropriate molecular information required to predict the target variable. Feature engineering requires domain knowledge, relying on having access to the experts [34]–[36]. In small datasets, DFT-based or experimentally determined features have been used due to their high accuracy. However, some reports also explore simple features based on molecular structure [37]–[42].

The goal of this study was to develop machine-learning models that predict redox potential in a short amount of time while maintaining high accuracy. Therefore, we did not compute any features from DFT calculations or experimental studies. Features used in this study were computed from molecular structures using readily available, easy-to-use python libraries such as RDKit [43] and DeepChem [44]. These libraries have been used in other studies as well [45]–[48]. Previous studies to predict redox potential using machine-learning investigate only a small number of non-linear models [49]–[53]. Furthermore, none of the previous studies use easily computable features from RDKit and DeepChem libraries. This study investigates twenty different linear and non-linear machine-learning models to predict the redox potential of phenazine derivatives in DME (Dimethoxyethane) solvent. Linear models are generally faster to train but may not capture complex relations between features and target variables, whereas non-linear models can capture these complex relationships but may need a considerable amount of training time. Total 3510 features containing 2D, 3D and Molecular Fingerprints were generated using RDKit, and DeepChem. Models were trained on four feature sets described in Table 1 to obtain high prediction accuracy. Moreover, to understand which feature set had the best prediction accuracy, a detailed analysis of model performance was carried out using the *pipeline* developed in this work (see methods section). The *pipeline* was developed to make training and evaluation easy, consistent, and automatic for all models. *Pipeline* combines different model training and evaluation steps into a single, convenient sub-routine. Then, the feature importance analysis was performed to identify the most important features in each feature set. Next, model performance was analyzed on small subsets of the most important features to reduce training and inference time for large datasets. Then, the promising redox-active candidates were identified using the predictions from best-performing models. Also, the prediction accuracy across different functional groups was analyzed. Finally, the sources of errors were identified. We believe that methods used in this work are easily

adaptable, and the results obtained in this study would help accelerate the discovery of novel redox-active species for energy storage applications.
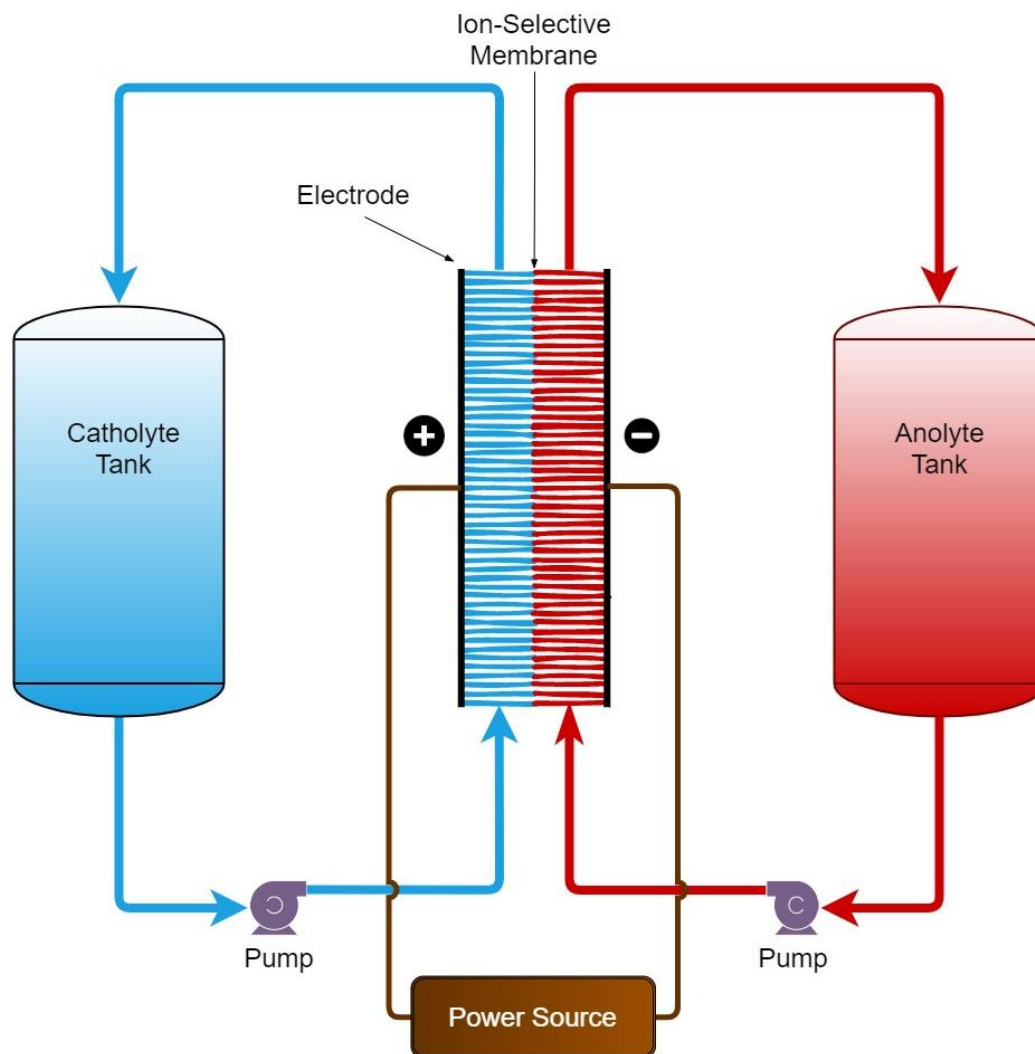

Figure 1. Schematic diagram of a typical Redox Flow Battery

DATA CURATION

Data used in this study was obtained from ref. [18]. The redox potentials of 189 phenazine derivatives in DME were provided. These potentials were computed using DFT. Phenazine derivatives contain twenty unique electron-withdrawing and donating functional groups (–N(CH3)2, –NH2, –OH, –OCH3, –P(CH3)2, –SCH3, –SH, –CH3, –C6H5, –CH=CH2, –F, –Cl, –CHO, –COCH3, –CONH2, –COOCH3, –COOH, –CF3, –CN and –NO2). Optimized 3D structures of derivatives in neural and in anionic states were also provided. Only neutral structures were used for the feature computation. However, not all compounds were provided with their neutral structure. Therefore, compounds missing neutral structures were removed. Thus, we ended up with 185 compounds in the final dataset. Next, 3510 different types of features were generated using RDKit and DeepChem libraries as described in the methods section. Finally, the whole dataset was shuffled and split randomly into a train-set and a test-set in a 7:3 ratio. This resulted in 129 samples in the train-set and 56 samples in the test-set.

METHODS

Feature Generation

For each compound, three types of features were generated: (i) 2D, (ii) 3D, and (iii) Molecular Fingerprints. A list of all features used in this study is given in Table S1. Ten 2D features were generated from the raw data (features with the word *'basic'* in suffix). The rest of the 2D features were computed using RDKit [43]. All 3D features were computed using RDKit. Molecular Fingerprints were computed using RDKit and DeepChem [44] libraries. These features were grouped into four sets, as shown in Table 1. Molecular Fingerprints and some of the 3D features are 1D vectors. In this study, we consider each component of the vectorial feature as a separate feature. Therefore, a small number of 3D and Molecular Fingerprint features results in a large number of features.

Table 1. Feature sets

Sr. No	Feature Set	Description	Number of Features
1	2d+3d+fp	Contains 2D and 3D features generated using raw data and RDKit. Also contains Molecular Fingerprints generated using RDKit and deepchem	3510
2	2d	Contains only 2D features generated using raw data and RDKit	151
3	3d	Contains only 3D features generated using RDKit	869
4	fp	Contains Molecular Fingerprints generated using deepchem and RDKit	2490

Pre-processing

Features having NaN (Not a Number) value for any compound were removed. Also, features having identical values for all compounds were removed as they do not contain any useful information. All 2D and 3D features computed from the RDKit library were scaled using the *'StandardScaler'* class of scikit-learn library [54], which removes mean and scales each feature to unit variance.

Machine-learning Models

Twenty linear and non-linear machine-learning models were investigated in this study. Machine-learning models were implemented with the scikit-learn python library [54]. A list of all models is given in Table 2.

Table 2. Model list

Sr. No.	Model Name	Alias	Model Type
1	Linear Regression	linear_reg	Linear
2	Ridge Regression	ridge	Linear
3	Lasso	lasso	Linear
4	Elastic-Net	elastic_net	Linear
5	LARS Lasso	lasso_lars	Linear
6	Orthogonal Matching Pursuit	omp	Linear
7	Bayesian Ridge Regression	bayesian_ridge	Linear
8	Automatic Relevance Determination Regression	ARDR	Linear
9	Passive Aggressive	PA	Linear
10	Huber Regression	huber	Linear
11	Kernel ridge Regression	kernel_ridge	Non-Linear
12	Support Vector Machines	SVR	Non-Linear
13	Gaussian Processes Regression	gaussian_process	Non-Linear
14	Decision Trees	decision_tree	Non-Linear
15	Bagging meta-estimator	bagging	Non-Linear
16	Random Forest	random_forest	Non-Linear
17	AdaBoost	ada_boost	Non-Linear
18	Gradient Boosting regression	gradient_boosting_reg	Non-Linear
19	Artificial Neural Network	neural_network	Non-Linear
20	Nearest Neighbors Regression	knn_reg	Non-Linear

Hyper-Parameter Tuning

Hyper-Parameter tuning was performed for all models using the *'GridSearchCV'* class of the scikit-learn library. *'GridSearchCV'* performs a systematic search over the grid of parameters to identify the best set of parameters using cross-validation. 10-fold cross-validation with mean squared error (MSE) was used in this study.

Evaluation Metrics

Following metrics were used to evaluate the model performance. In the formulas below, N denotes the number of samples, $\hat{y}_i$ denotes the predicted value of i-th sample and the y_i denotes the corresponding true value.

1. Coefficient of Determination (R^2):

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

$$where, \quad \bar{y} = \frac{\sum_{i=1}^{N} y_i}{N}$$

2. Mean Squared Error (MSE):

$$MSE = \frac{\sum_{i=1}^{N}(y_i - \hat{y})^2}{N}$$

3. Mean Absolute Error (MAE):

$$MAE = \frac{\sum_{i=1}^{N}|y_i - \hat{y}_i|}{N}$$

4. Mean Absolute Percentage Error (MAPE):

$$MAPE = 100 * \frac{\sum_{i=1}^{N} \frac{|y_i - \hat{y}_i|}{[y_i]}}{N}$$

MSE was used as an internal evaluation metric in all cross-validation procedures. Other metrics were used to get more interpretable results. The use of terms 'Accuracy' and 'Performance' in this report is contextual and refers to one or more metrics defined above.

MSE and MAE Threshold

To understand whether the model was learning or not, we determined the approximate upper bound on MSE and MAE for the train and test set. It is expected that MSE and MAE would stay below this threshold if learning were successful. It was observed that when training fails model predicts constant value (i.e., mean of the training data). The threshold value for MSE and MAE was determined using the mean value of training data. Threshold values are shown in Table 3.

Table 3. Threshold values

Sr. No.	Metric	Training-Set Threshold	Test-Set Threshold
1	MSE	0.47	0.44
2	MAE	0.6	0.56

k-Fold Cross-Validation

In a typical k-fold cross-validation procedure, the train-set is split into k sets of approximately equal size. Then, the model is trained on k-1 sets, leaving one set as a test-set. Then, the performance of the trained model is evaluated on the left-out test-set. This procedure is repeated for every fold, and average performance is reported. As every data-point in the training set is evaluated as if it belongs to the test-set, the performance obtained from cross-validation is considered a reasonable estimate of out-of-sample performance. k-fold cross-validation gives

robust out-of-sample performance for the model. It is a crucial evaluation technique, especially when the dataset size is very small and it becomes impractical to partition data into three sets (i.e., train, validation, test). 10-fold cross-validation was used in this study.

Feature Importance Score

Feature importance scores were computed using the optimized Random Forest, AdaBoost, and Gradient Boosting Regression trained on all features from the corresponding feature set.

Pipeline

To assess the model performance, we developed a pipeline that combines all training and evaluation components into a single procedure. Given the train and test sets as inputs, the *pipeline* first performs hyperparameter-tuning for all models, then evaluates the performance of optimized models on CV, train, and test sets, and finally combines necessary results from each step in a single dataframe. *Pipeline* makes the training and evaluation easy, consistent and automatic for all models across different scenarios. A pictorial representation of the *pipeline* is shown in Figure 2. Different steps in the *pipeline* are described below:

1. **Input:** First, training and test data are provided as inputs
2. **Hyper-Parameter Tuning:** In this step, optimized parameters of all twenty models are determined using the train-set as described in this section before.
3. **10-Fold Cross-Validation:** In this step, the cross-validation performance of optimized models is evaluated using 10-fold cross-validation on the train-set. Three metrics (i.e., R^2, MSE, and MAE) are recorded during the cross-validation for all models.
4. **Train and Test set Performance:** In this step, the performance of all optimized models is evaluated on the train and test set. Three metrics (i.e., R^2, MSE, and MAE) are recorded during the evaluation for all models.
5. **Output:** In this step, outputs from steps 2-4 are combined into one dataframe containing the best set of parameters, 10-fold cross-validation performance, train and test set performance of all models.
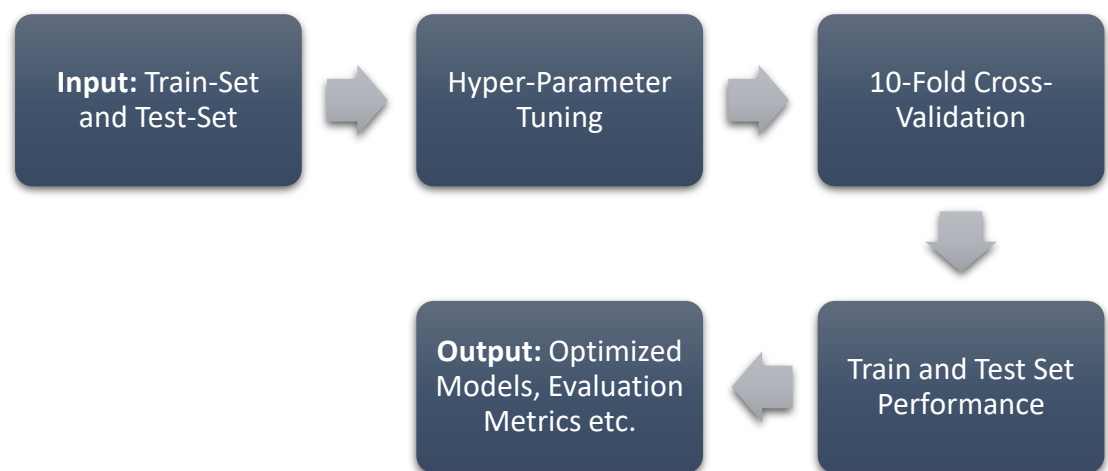
Figure 2. Pictorial representation of model training and evaluation pipeline
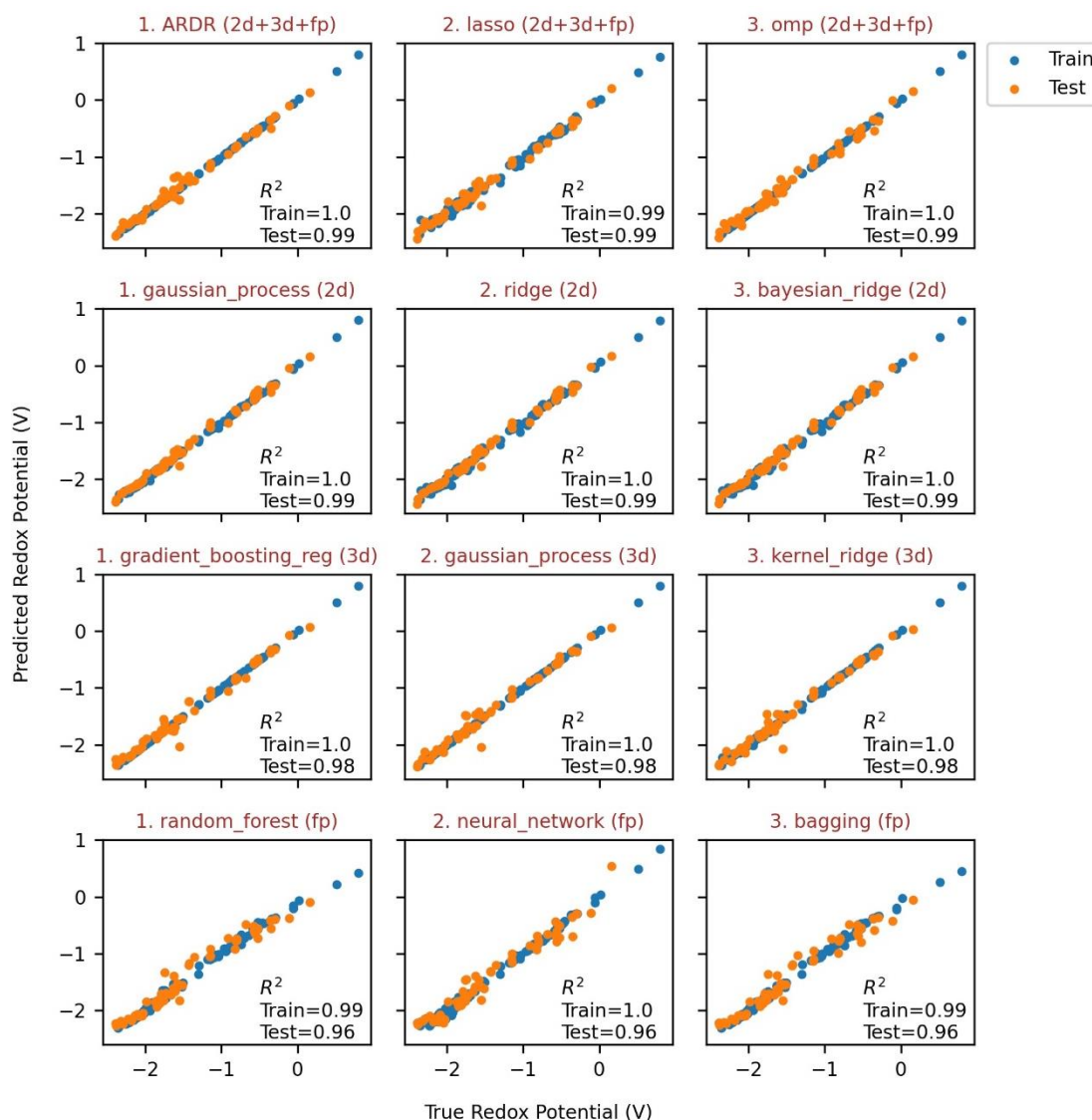
RESULTS AND DISCUSSION

Figure 3. Machine-learning prediction of redox potential vs. True redox potential of the three best-performing models in four feature sets. The title above the plot indicates the model name, its rank, and corresponding feature set in brackets.

Analysis of Best-Performing Machine-learning Models

For accurate prediction of the redox potential of phenazine derivatives, we employed twenty different linear and non-linear machine-learning models, which are listed in Table 2. The whole dataset was shuffled and split randomly into train and test sets in a 7:3 ratio. The size of the train and test set was 129 and 56, respectively. Even though models were trained on a relatively small dataset, they achieved excellent performance on the unseen data (i.e., test-set). Figure 3 shows the redox potential predicted by models on the y-axis and the corresponding true value of redox potential on the x-axis. The majority of models achieved R^2 value of 0.99 on the test-set (R^2 values in the plots were rounded to two decimal places for clarity). Table 4 shows the twenty best-performing models obtained in this study along with their R^2, MSE, and MAE values on cross-validation (CV), train-set, and test-set. All top twenty models not only had an outstanding performance on train-set but also on test-set (i.e., $R^2 > 0.98$, MSE < 0.008 V^2 and MAE < 0.07 V).

Table 4. Twenty best-performing models. Models were trained on all features from the corresponding feature set.

Rank	Feature Set	Model Name	R² (CV)	MSE (CV) / V²	MAE (CV) / V	R2 (Train-set)	MSE (Train-set) / V²	MAE (Train-set) / V	R² (Test-set)	MSE (Test-set) / V²	MAE (Test-set) / V
1	2d	gaussian_process	0.9738	0.0078	0.0559	0.9991	0.0004	0.0136	0.9921	0.0035	0.0428
2	2d	ridge	0.9767	0.0069	0.0541	0.9960	0.0019	0.0298	0.9916	0.0037	0.0454
3	2d	bayesian_ridge	0.9751	0.0072	0.0537	0.9964	0.0017	0.0282	0.9915	0.0037	0.0452
4	2d	neural_network	0.9644	0.0108	0.0649	0.9988	0.0005	0.0133	0.9909	0.0040	0.0447
5	2d	kernel_ridge	0.9738	0.0077	0.0586	0.9947	0.0025	0.0372	0.9896	0.0046	0.0501
6	2d	omp	0.9160	0.0546	0.0798	0.9940	0.0028	0.0389	0.9876	0.0055	0.0567
7	2d	ARDR	0.9768	0.0074	0.0582	0.9956	0.0020	0.0333	0.9874	0.0055	0.0540
8	2d+3d+fp	ARDR	0.9519	0.0132	0.0600	0.9999	0.0000	0.0052	0.9873	0.0056	0.0473
9	2d+3d+fp	lasso	0.9826	0.0064	0.0593	0.9937	0.0029	0.0429	0.9868	0.0058	0.0519
10	2d+3d+fp	omp	-0.5387	0.3913	0.2317	1.0000	0.0000	0.0000	0.9861	0.0061	0.0628
11	2d	lasso	0.9768	0.0078	0.0680	0.9901	0.0046	0.0550	0.9857	0.0063	0.0599
12	2d+3d+fp	gaussian_process	0.9849	0.0058	0.0509	1.0000	0.0000	0.0021	0.9855	0.0064	0.0478
13	2d+3d+fp	bayesian_ridge	0.9851	0.0055	0.0509	0.9993	0.0003	0.0114	0.9853	0.0064	0.0483
14	2d+3d+fp	ridge	0.9858	0.0053	0.0499	0.9989	0.0005	0.0144	0.9849	0.0066	0.0494
15	2d+3d+fp	gradient_boosting_reg	0.9713	0.0141	0.0673	1.0000	0.0000	0.0009	0.9849	0.0066	0.0546
16	2d+3d+fp	linear_reg	0.9812	0.0069	0.0576	1.0000	0.0000	0.0000	0.9836	0.0072	0.0546
17	2d+3d+fp	kernel_ridge	0.9857	0.0052	0.0536	0.9960	0.0019	0.0314	0.9834	0.0073	0.0559
18	2d	gradient_boosting_reg	0.9753	0.0112	0.0672	0.9989	0.0005	0.0170	0.9834	0.0073	0.0587
19	2d+3d+fp	random_forest	0.9622	0.0209	0.0772	0.9963	0.0017	0.0257	0.9833	0.0073	0.0629
20	2d+3d+fp	bagging	0.9661	0.0186	0.0755	0.9970	0.0014	0.0233	0.9827	0.0076	0.0655

Assessing Model Performance on Four Feature Sets

The performance of machine-learning models depends on the type and the quality of features used for training. Therefore, it is important to identify the best set of features that achieve high prediction accuracy. Hence, we assessed the model performance on four sets of features shown in Table 1. The goal here was to understand how different types of molecular features affect model performance. The list of all features is given in Table S1. Model performance on each feature set was assessed using the *pipeline* described in the methods section. Twenty best-performing models in Table 4 were obtained after assessing model performance on all four feature sets. The '*Feature Set*' column in Table 4 shows the corresponding feature set used for the training. Gaussian Processes Regression trained on 2D features achieved the highest prediction accuracy in this study. The negative value of R^2 (CV) for the *Orthogonal Matching Pursuit (omp)* model in Table 4 shows that it may not generalize well to the unseen data. A similar trend was observed for a few other linear models (Figure 4). Nine models, including the top seven models in Table 4, were trained on the '*2d*' feature set. The rest of the models in Table 4 were trained on the '*2d+3d+fp*' feature set. We did not observe any model trained on either '*3d*' or '*fp*' in Table 4. The best-performing model in each feature set, along with their test-set performance (i.e., R^2, MSE, and MAE), is shown in Table 5. Additionally, ten best-performing models in each feature set, along with their CV, train-set, and test-set performance (i.e., R^2, MSE, and MAE), are shown in Tables S2-S5. Similarly, the best-performing feature set for each model, along with their CV, train-set, and test-set performance, is shown in Table S6. Observations from Table 4, Table 5 and Tables S2-S6 show that the prediction accuracy with respect to feature set follows this order: *2d > 2d+3d+fp > 3d > fp*. Therefore, we conclude that 2D features are more informative than 3D and Molecular Fingerprint features to predict the redox potential of phenazine derivates in DME. Tables S2-S5 also show linear models (e.g., *ARDR, lasso, omp, ridge, bayesian_ridge*) perform better than non-linear models on '*2d+3d+fp*' and '*2d*' feature sets, whereas non-linear models (e.g., *gradient_boosting_reg, gaussian_process, random_forest, neural_network*) perform better than linear models on '*2d*', '*3d*' and '*fp*' feature sets. This observation suggests that linear models should be preferred when the feature set consists of different features (i.e., 2D, 3D, and Molecular Fingerprints). Non-linear models should be preferred when the feature set contains either 3D or Molecular Fingerprint features. Any model could be used for 2D features. Linear models are generally faster than non-linear models due to their simple structure but are not prefered due to low accuracy. The results obtained here show that linear models could give accurate predictions than non-liearn models on certain combination of feature set. Utilizing linear models in thses scenarios could significantly reduce the training and inference time.

Table 5. Best-performing models in four feature sets. Models were trained on all features from the corresponding feature set.

Sr. No.	Featue Set	Model Name	R^2 (Test-set)	MSE (Test-set) / V^2	MAE (Test-set) / V
1	2d+3d+fp	ARDR	0.9873	0.0056	0.0473
2	2d	gaussian_process	0.9921	0.0035	0.0428
3	3d	gradient_boosting_reg	0.9788	0.0093	0.0573
4	fp	random_forest	0.9583	0.0183	0.1012

Cross-Validation and Out-of-Sample Performance

10-fold cross-validation (CV) performance (i.e., R^2, MSE, and MAE) obtained from the *pipeline* is shown in Figure 4 for all twenty models. Cross-validation gives a reasonable estimate of out-of-sample performance (i.e., performance on unseen data). *'2d+3d+fp', '2d'* and *'3d'* feature sets had the acceptable CV performance (i.e., MSE and MAE below their threshold value) on most models except on four linear models. The computation of threshold values for MSE and MAE is described in the methods section. These four linear models (i.e., *linear_reg, omp, PA, huber*) had negative R^2 value, high MSE, and high MAE (i.e., close to threshold) for at least one feature set. *'fp'* feature set had the worst CV performance on all models. Three linear models (i.e., *omp, PA, huber*) had poor CV performance on the *'3d'* feature set. Performance on test-set (i.e., R^2, MSE, and MAE) obtained from the *pipeline* is shown in Figure 5 for all twenty models. As models never see a test-set, test-set gives an even better estimate of out-of-sample performance than cross-validation. All feature sets had an acceptable test-set performance on all models except Linear Regression. Linear Regression had a poor test-set performance on *'2d'* and *'fp'* feature sets. Furthermore, the averaged train and test performance (i.e., R2, MSE, and MAE values were averaged over all models) for each feature set is shown in Table 6. Values of Linear Regression were not considered in the average due to very high errors. In Table 6, *'2d+3d+fp'* and *'fp'* feature sets had better train-set performance and poor test-set performance than the *'2d'* feature set. This indicates that 2D features are better at generalizing to unseen data than 3D and Molecular Fingerprint features. Trend analysis of Figure 4, Figure 5, and Table 6 revealed previously observed order of feature set performance, *2d > 2d+3d+fp > 3d > fp*. Plots similar to Figure 4 for the train-set are shown in Figure S1.

Table 6. Averaged train and test set performance of four models. Performance was aveaged over all models except Linear Regression. Models were trained on all features from the corresponding feature set.

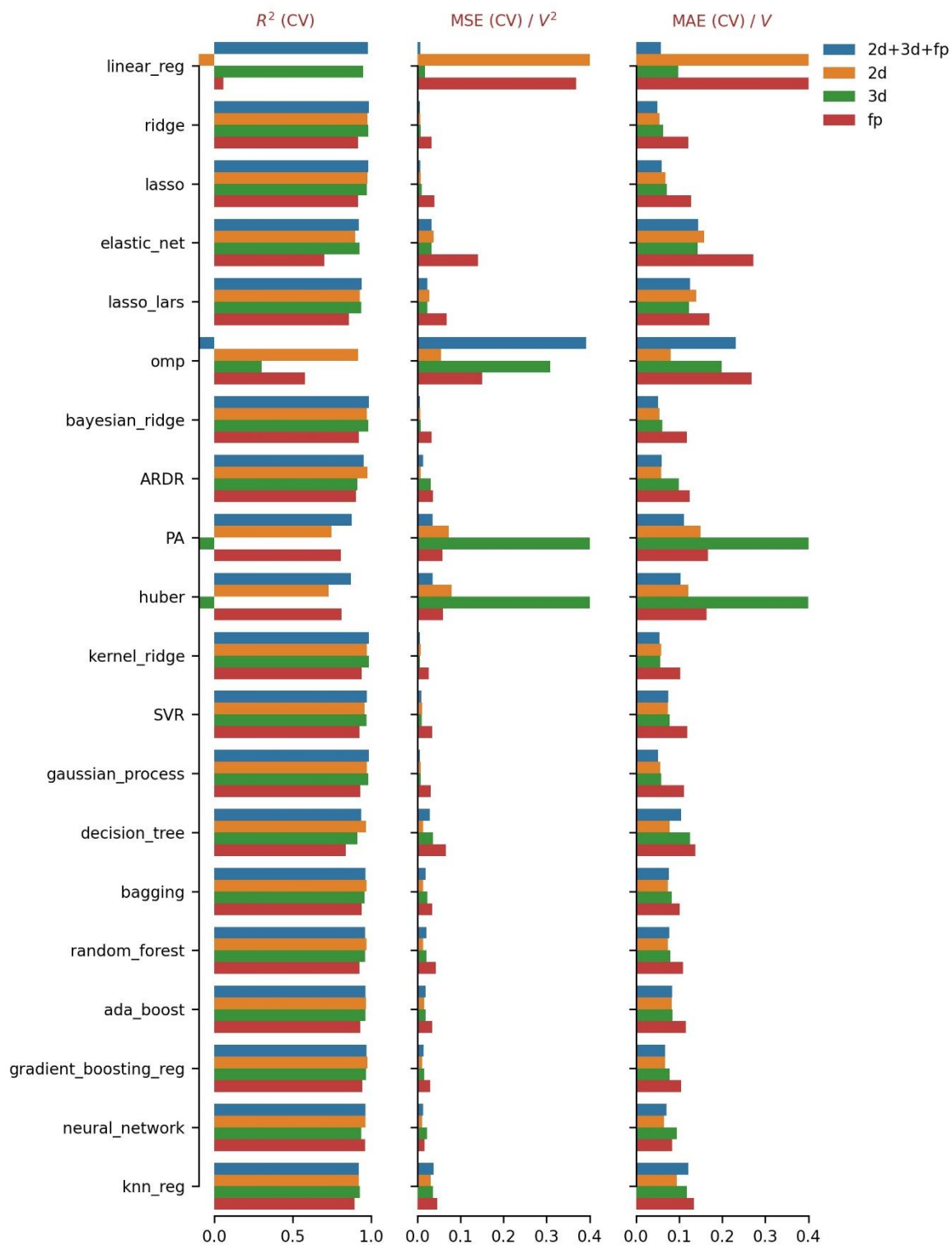| Sr. No.	Feature Set	R^2 (Train-set)	MSE (Train-set) / V^2	MAE (Train-set) / V	R^2 (Test-set)	MSE (Test-set) / V^2	MAE (Test-set) / V
1	2d+3d+fp	0.9926	0.0035	0.0301	0.9718	0.0124	0.0733
2	2d	0.9876	0.0058	0.0426	0.9729	0.0119	0.0734
3	3d	0.9917	0.0039	0.0335	0.9535	0.0204	0.0899
4	fp	0.9718	0.0132	0.0506	0.9028	0.0427	0.1511

Figure 4. 10-Fold Cross-Validation performance of twenty models. Models were trained on all features from the corresponding feature set.
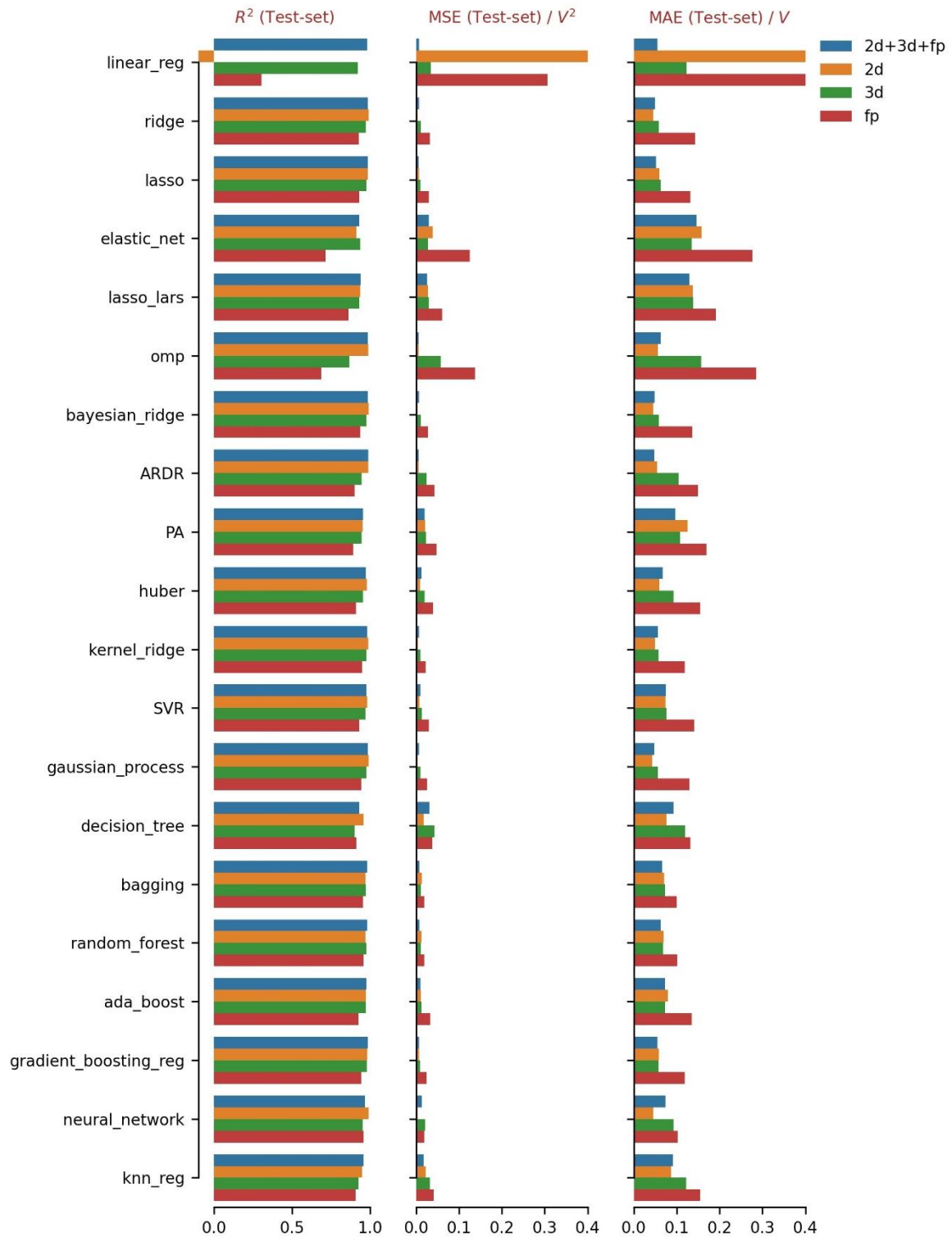
Figure 5. Test-set performance of twenty models. Models were trained on all features from the corresponding feature set.

Feature Importance Analysis

Here, we performed feature importance analysis for each feature set to identify the most important features. We used Random Forest, AdaBoost, and Gradient Boosting Regression to calculate feature importance score. Figure 6, Figure 7, Figure 8, and Figure 9 show feature importance histograms for *'2d', '3d', 'fp' and '2d+3d+fp'* feature sets, respectively. Only twenty features with the highest scores are shown in the histograms. The most important features in the *'2d'* feature set are *SlogP_VSA4, fr_NH0, VSA_Estate3,* and *VSA_Estate4*. *SlogP_VSA4* includes the LogP [55] and Van der Waals surface area contributions from all atoms in the molecule. *fr_NH0* is the number of Tertiary amines. *VSA_Estate3* and *VSA_Estate4* are calculated using EState indices [56] and Van der Waals surface area contributions of all atoms in the molecule. Many graph-based features like *Kappa2, BertzCT, Chi1, Chi2n, HallKierAlpha* [57] and some chemically intuitive features like *fr_ArN* (i.e., number of N functional groups attached to aromatic ring), *MinPartialChargge, MaxAbsPartialCharge* are also observed in the top twenty features. In the case of *'3d'* feature set, *RDF_120, RDF_90, RDF_125 WHIM_90, WHIM_86* are among the top 3D features (*RDF, WHIM* are 1D vectors) [58]. The number at the end of the feature name denotes its position in the corresponding feature vector. Some components of *MORSE* and *GETAWAY* feature vectors [58] are also observed in the top twenty features. Only two components of *Autocorr3D* were observed in one of the histograms (i.e., AdaBoost histogram) suggesting *Autocorr3D* [58] is a relatively less important 3D feature. None of the scalar 3D features were observed in feature importance histograms, suggesting scalar 3D features are less important than vector 3D features. We had only three types of fingerprints in the *'fp'* feature set (i.e., RDKit, ECFP4, MACCS keys). RDKit Fingerprints are daylight-like fingerprints computed from hashing molecular subgraphs [59]. ECFP4 [60] or Extended Connectivity Circular Fingerprints are computed from bag-of-word representation of local molecular neighbourhood. Four in ECFP4 denotes the radius of local neighbourhood. MACCS keys are computed using SMARTS-based implementation of the 166 public MACCS keys [61]. Components only from RDKit and ECFP4 were among the top features whereas only one component from MACCS keys was observed in one of the histograms (i.e., Gradient Boosting Regression histogram). This indicates MACCS key fingerprint does not contain enough molecular information to predict redox potential. Figure 9 shows the feature importance histograms for the *'2d+3d+fp'* feature set. This feature set contains all the features from *'2d', '3d',* and *'fp'* feature sets. The most important features in this feature set were also the most important features in their respective set. Top features are mainly from *'3d'* and *'2d'* feature sets, and only one component from the ECFP4 feature vector was observed in the lower end of the Gradient Boosting Regression histogram. This again shows Molecular Fingerprints are the least informative among all features. Feature importance histogram of *'2d+3d+fp'* feature set contains a few *'2d'* features and predominantly *'3d'* features. This can be attributed to how fast importance scores diminish from most important features to the least important features. Feature importance scores in the *'2d'* feature set (Figure 6) diminish faster than the *'3d'* feature set (Figure 7). This also indicates that very few 2D features are required to predict redox potential compared to 3D features.
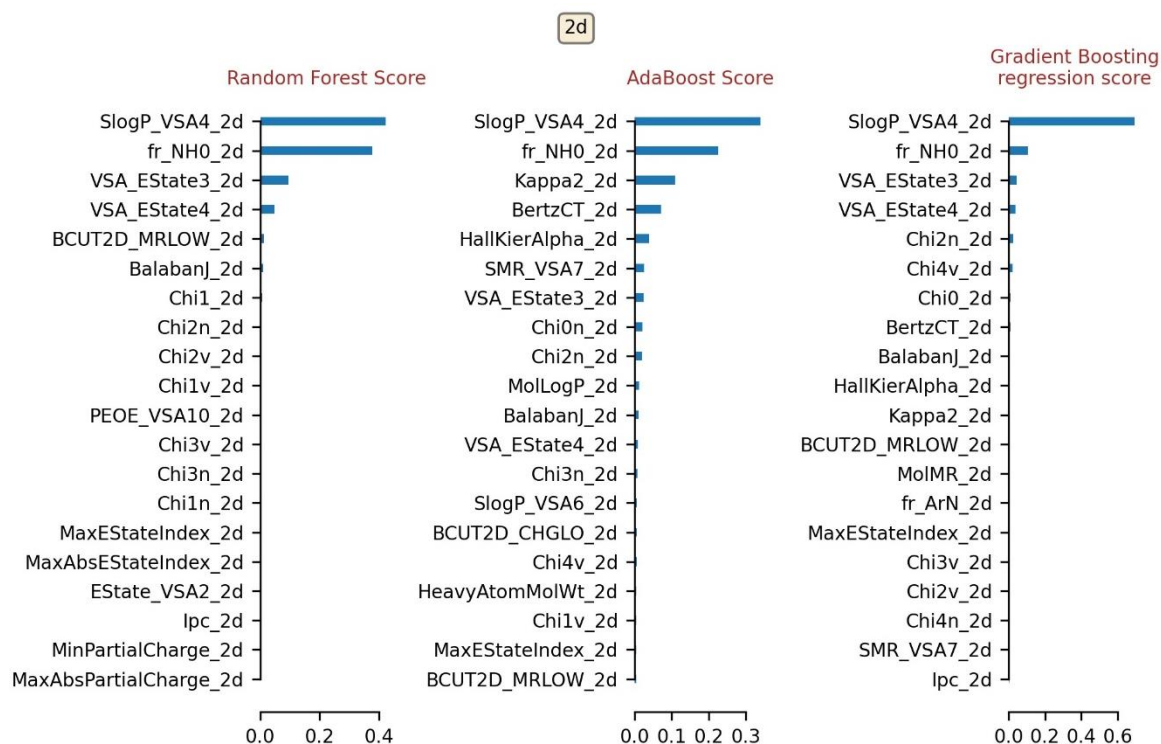
Figure 6. Feature importance histograms of *'2d'* feature set. Models were trained on all features from the corresponding feature set.
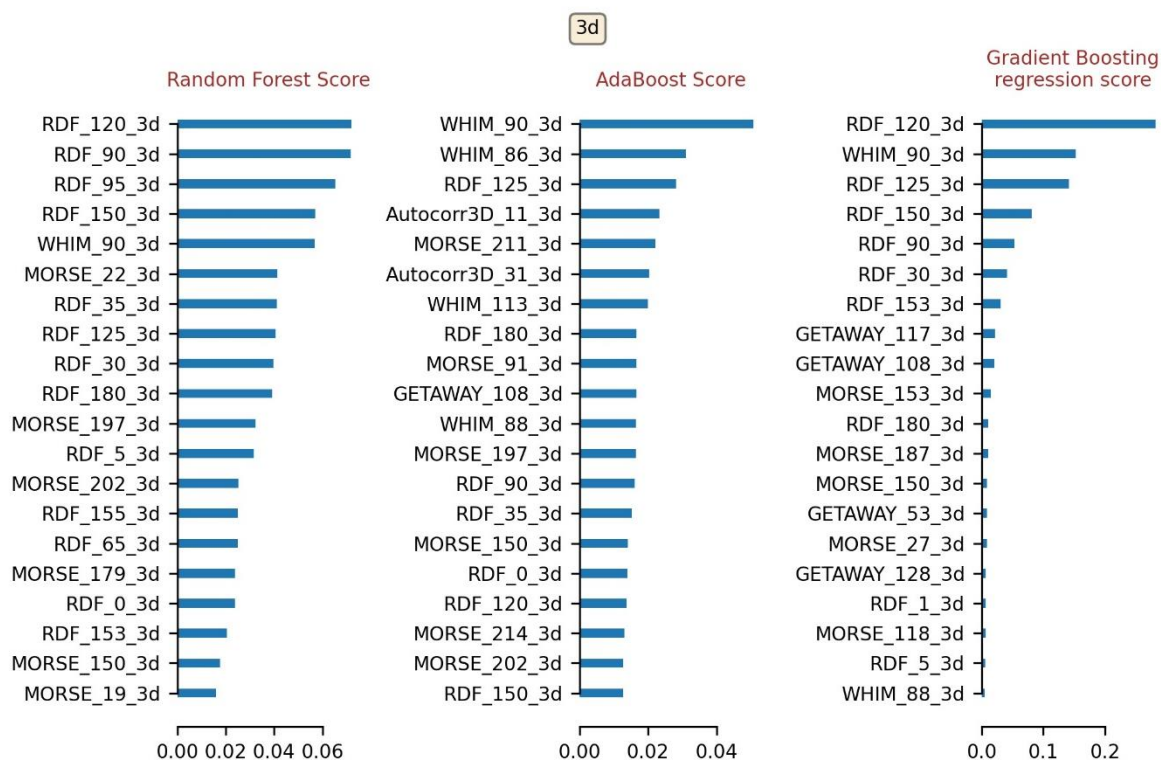

Figure 7. Feature importance histograms of *'3d'* feature set. Models were trained on all features from the corresponding feature set.
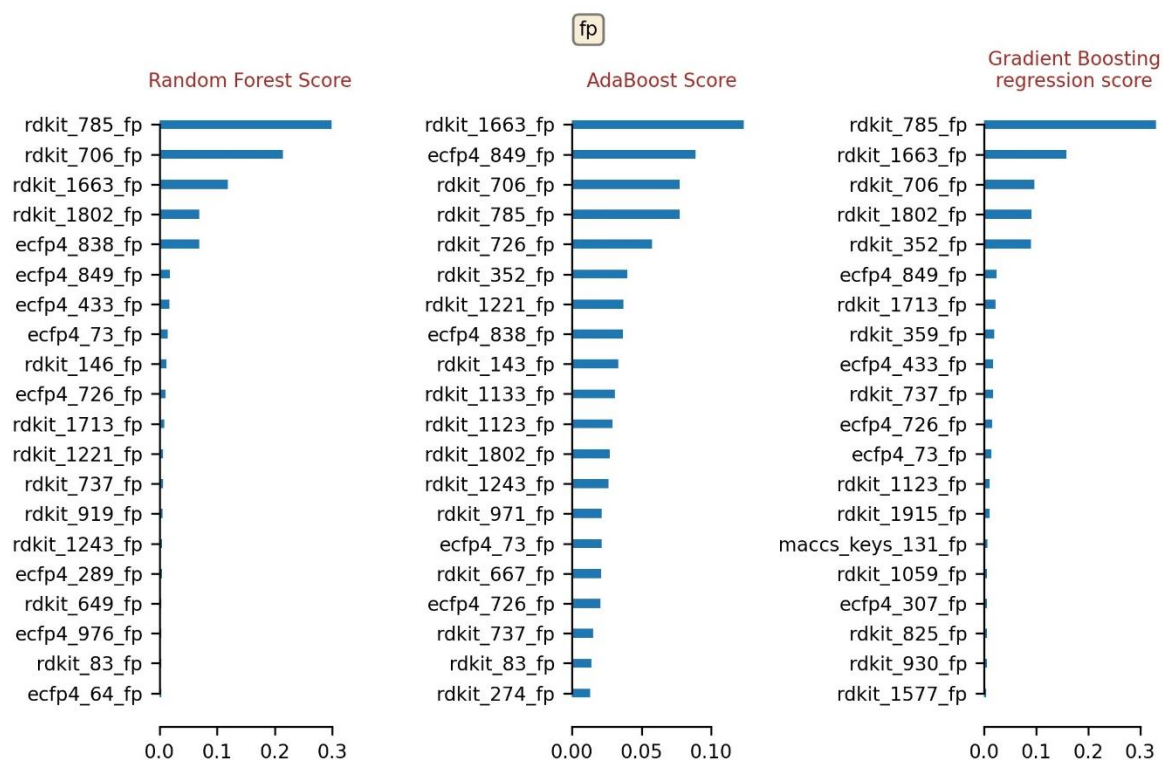
Figure 8. Feature importance histograms of *'fp'* feature set. Models were trained on all features from the corresponding feature set.
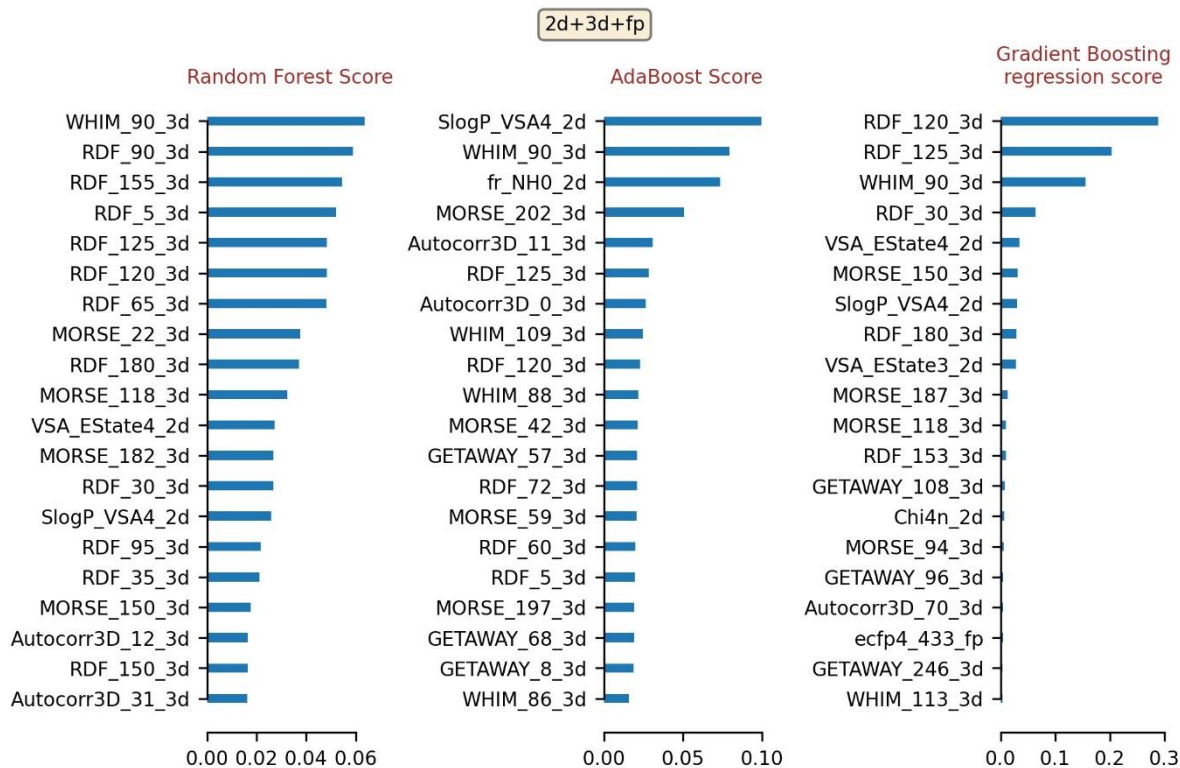

Figure 9. Feature importance histograms of *'2d+3d+fp'* feature set. Models were trained on all features from the corresponding feature set.

Effect of Feature Size on Model Performance

Previous feature importance study suggests that not all features may be necessary to achieve high predictive performance. To confirm this hypothesis, two best-performing models in each feature set were selected and re-trained on the subset of features. Features were sorted in
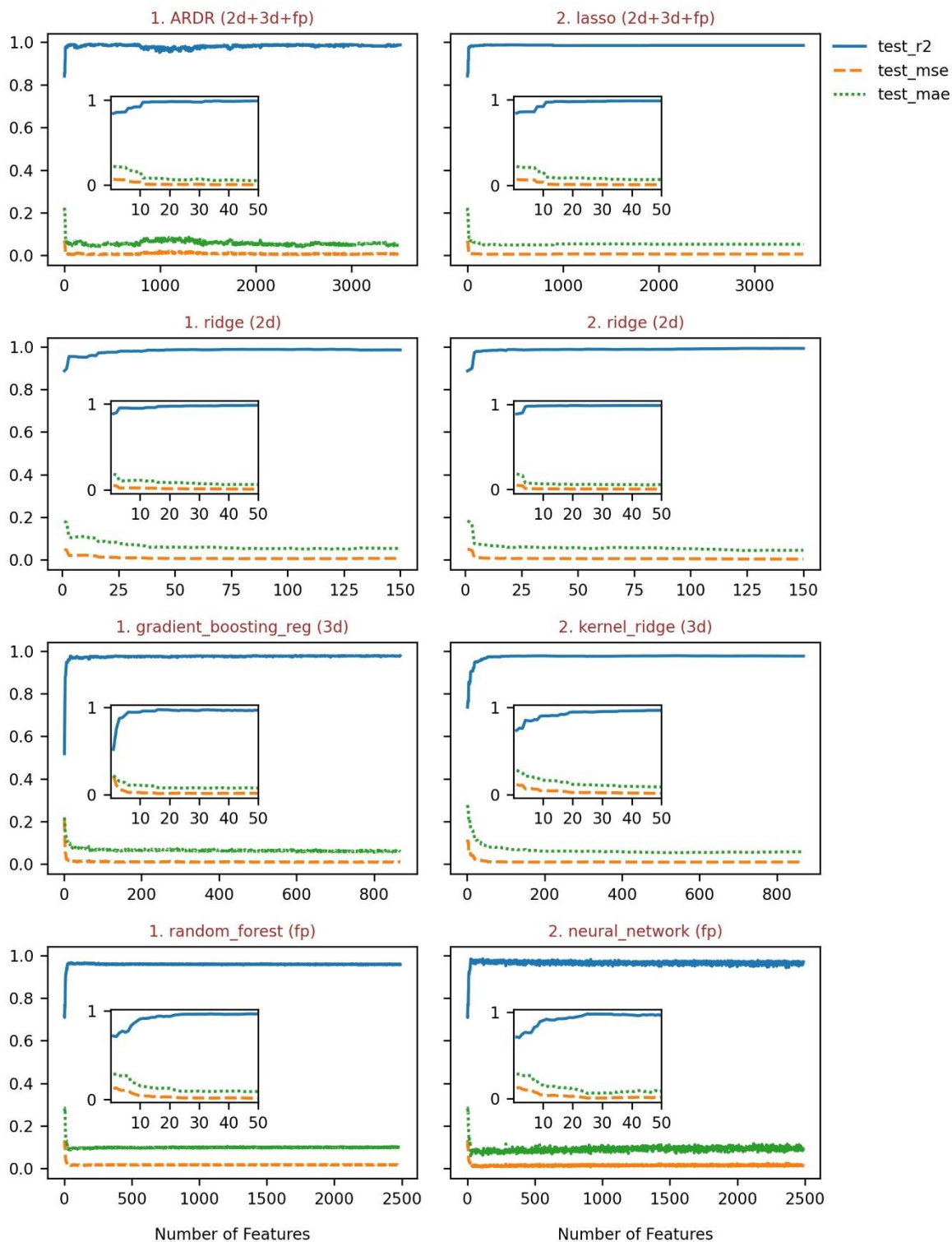
Figure 10. Model performance vs. Number of features

descending order based on the Random Forest scores. Models were re-trained staring with a single feature to the full set of features, and three metrics (i.e., R^2, MSE, MAE) on test-set were recorded. The results for each feature set are shown in Figure 10. Inset plots show the same data for the first 50 features. Plots corresponding to the *'2d'* feature set quickly saturate (after ~ 5 features), suggesting only a small number of 2D features are required to predict redox potential accurately. Plots of the *'3d'* feature set seem to saturate after twenty features, whereas plots corresponding to the *'fp'* feature set saturate slowly and require more than twenty features to achieve similar performance. In the case of the *'2d+3d+fp'* feature set, plots seem to saturate around 15-20 features and look approximatly similar to a linear combination of the plots from the *'2d', '3d'*, and *'fp'* feature sets. These plots clearly show that not all features are required to attain a high level of prediction accuracy.

Assessing Model Performance on Limited Number of Features

To gain insight into the quality of predictions when models are trained on the limited number of features, all models were re-trained on the subset of features using the *pipeline* (see methods section). Features were chosen from the array of features sorted based on the Random Forest score. The number of features was varied from five to twenty in a step of five. Figure 11 shows the test-set performance of models when trained on a small number of features from the *'2d+3d+fp'* feature set. Performance of the full set of features is also shown for reference. We observed that model performance generally increases with the number of features. A few exceptions were also observed. Some models (e.g., *PA, huber, neural network,* and *knn_reg*) had better performance on the top twenty features than a full set of features. Moreover, *decision_tree* performed better on the top fifteen features than a full set of features. Similar trends were observed for other feature sets, too (Figures S2-S4). However, in the *'fp'* feature set, *SVR* and *knn_reg* showed better performance on the top twenty features than a full set of features. We also analyzed which feature set was able to achieve the highest accuracy when models were trained on a small number of features. Figure 12 shows the test-set performance of all models when trained on only the top five features from each feature set. *'2d'* feature set achieved R^2 value as high as 0.9869 with only five features with the *'bagging'* model. Performance of *'2d+3d+fp'* and *'3d'* feature set was similar but sub-par to *'2d'* feature set. The similarity in performance of *'2d+3d+fp'* and *'3d'* feature sets could be attributed to the similarity in their top five features (see Figure 7 and Figure 9). *'fp'* feature set performed poorly with only five features. These results were consistent across all models. Similar results were also obtained for the models trained on the top ten, fifteen, and twenty features which are shown in Figures S5-S7. Furthermore, a significant descreasing in the training and inference time was observed while using limited number of features for the training. These results suggest that training and inference time could be reduced by using a small number of features while maintaining a good accuracy level.
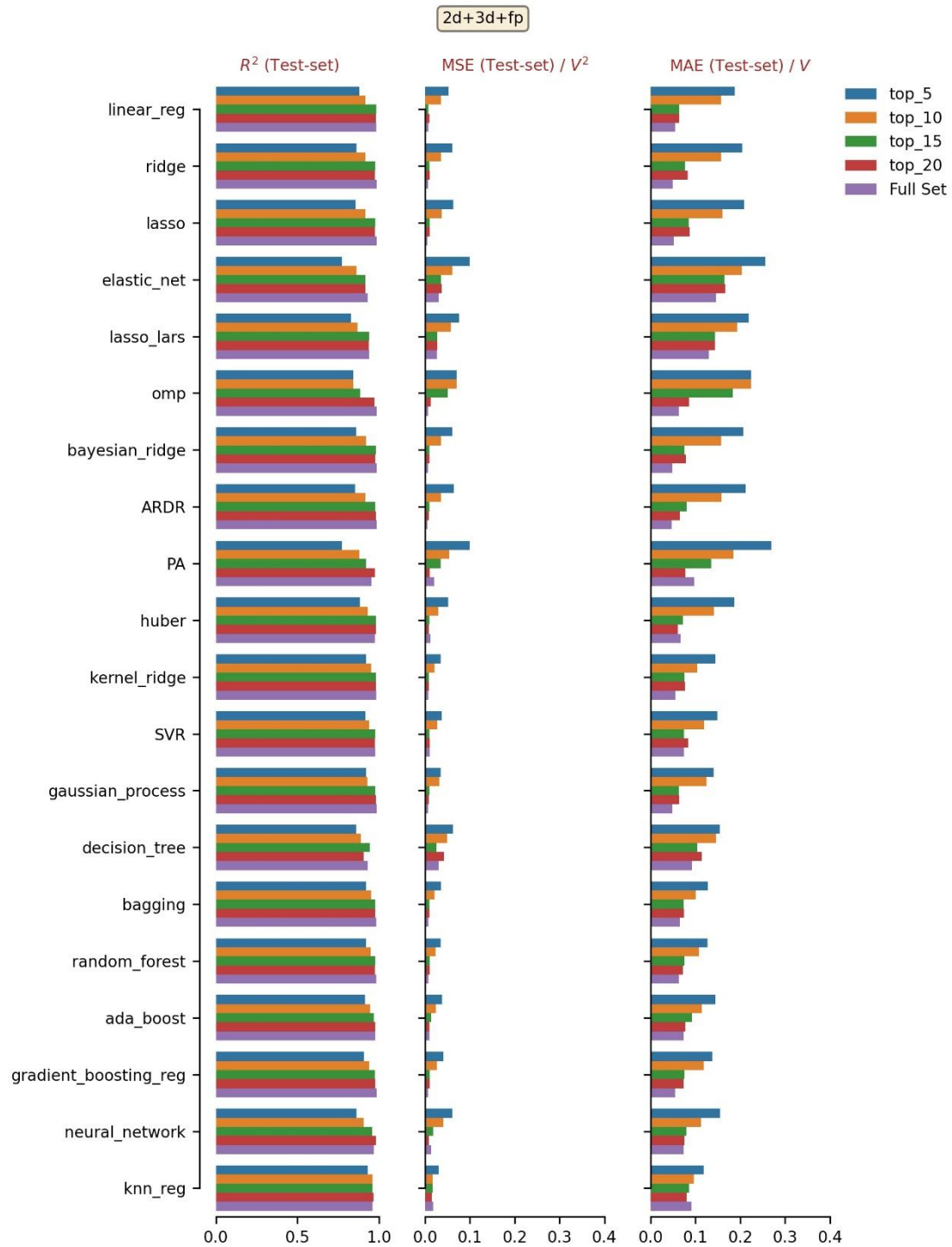
Figure 11. Test-set performance of twenty models trained on top-5, 10, 15, and 20 features from *'2d+3d+fp'* feature set. Top most important features were selected based on the Random Forest score. Test-set performance of the models trained on all features from the *'2d+3d+fp'* features (full set) is also shown for reference.
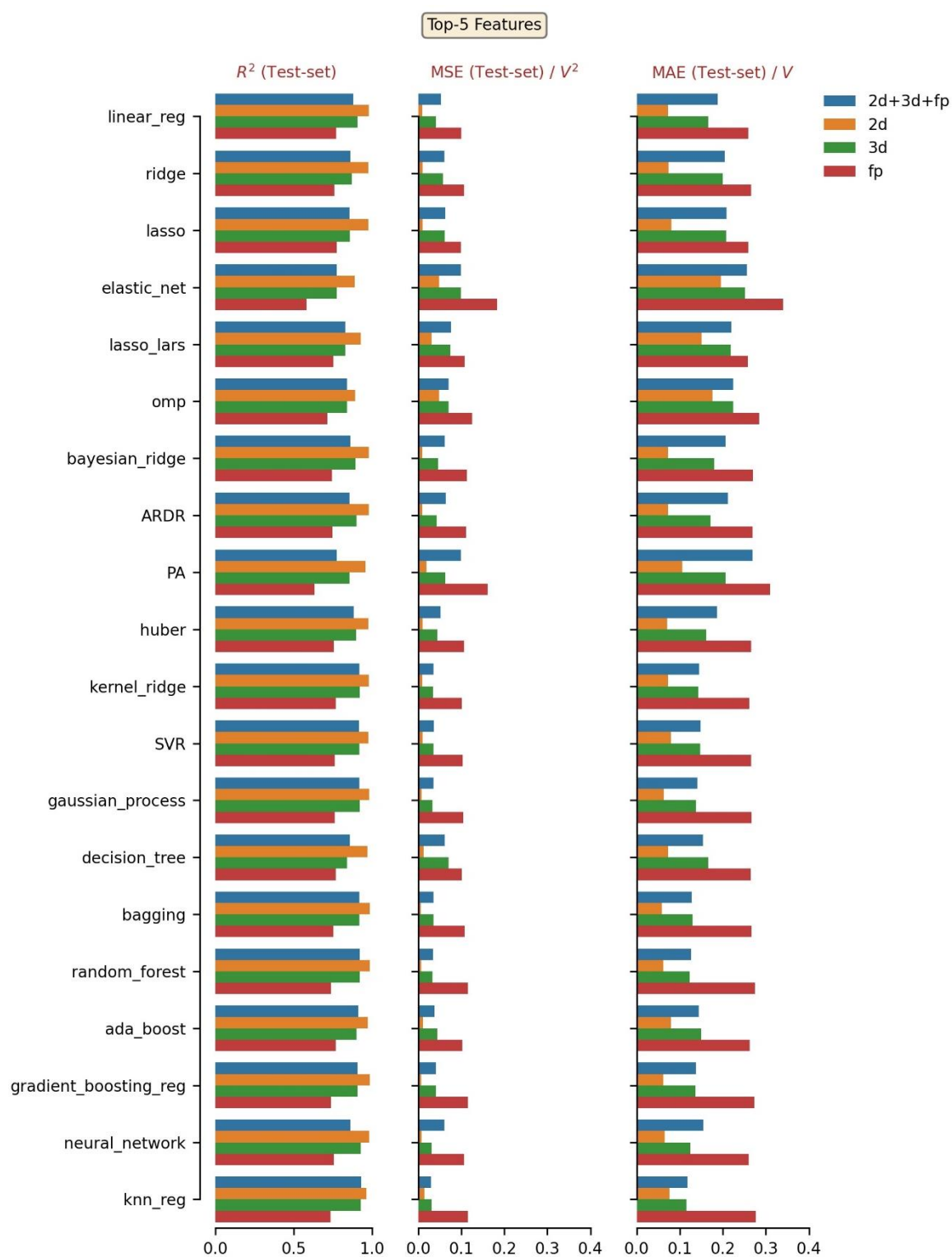
Figure 12. Test-set performance of twenty models trained on five most important features from the corresponding feature set. Five most important features were selected based on the Random Forest score.

Identifying Promising Redox-Active Candidates

Compounds having a significant difference in their redox potential are considered to be potential redox-active species. In this study, trained machine learning models were able to predict previously reported redox-active candidates. In particular, promising cathode redox-active compounds such as tetra-amino-phenazine (TAPZ), hexa-amino-phenazine (HAPZ), and octa-amino-phenazine (OAPZ) were predicted with high accuracy. The absolute difference between predicted (using the best-performing model) and actual redox potential of TAPZ, HAPZ, and OAPZ derivatives was less than 0.07 V ($< 3\%$). Furthermore, our best-performing model was able to predict the redox potential of octa-nitro-phenazine (ONPZ) with an error less than 0.004 V ($<2.4\%$) even though the dataset used in this study contains very few compounds with positive redox potential. ONPZ is considered a promising anode redox-active candidate for RBFs containing phenazine derivatives [18]. Additionally, most models could predict the correct relative order of compounds with respect to redox potential, clustering potential cathode redox-active candidates in the negative region, and potential anode redox-active candidates in the region close to zero. From these results, we see that promising cathode redox-active species contain electron-donating groups (i.e., NH2), whereas promising anode redox-active species contain electron-withdrawing groups (i.e., CN, NO2). These results are in good agreement with the previous report [18].

Analysis of Predictive Performance on Individual Function Groups

Here, we evaluate the prediction accuracy of the best-performing models in each feature set with respect to different functional groups attached to the phenazine ring. Twenty different functional groups were present in the phenazine derivatives used in this study. Train-set and Test-set Predictions were obtained from the best performing model in each feature set (see Table 5). Figure 13 shows the Mean Absolute Percentage Error (MAPE) for each functional group (FG) present in the test-set. MAPE in Figure 13a is averaged over FG and four feature sets. Whereas, MAPE in Figure 13b is averaged over only FG. MAPE of the functional groups present in the train-set is shown in Figure S8. MAPE for the majority of the functional groups was well below 10%. Even though some functional groups appeared only once in the training set (see Figure 14), most of the models were able to predict their redox potential with minimal error. Functional group COCH3 was present only in the test-set that means models never saw
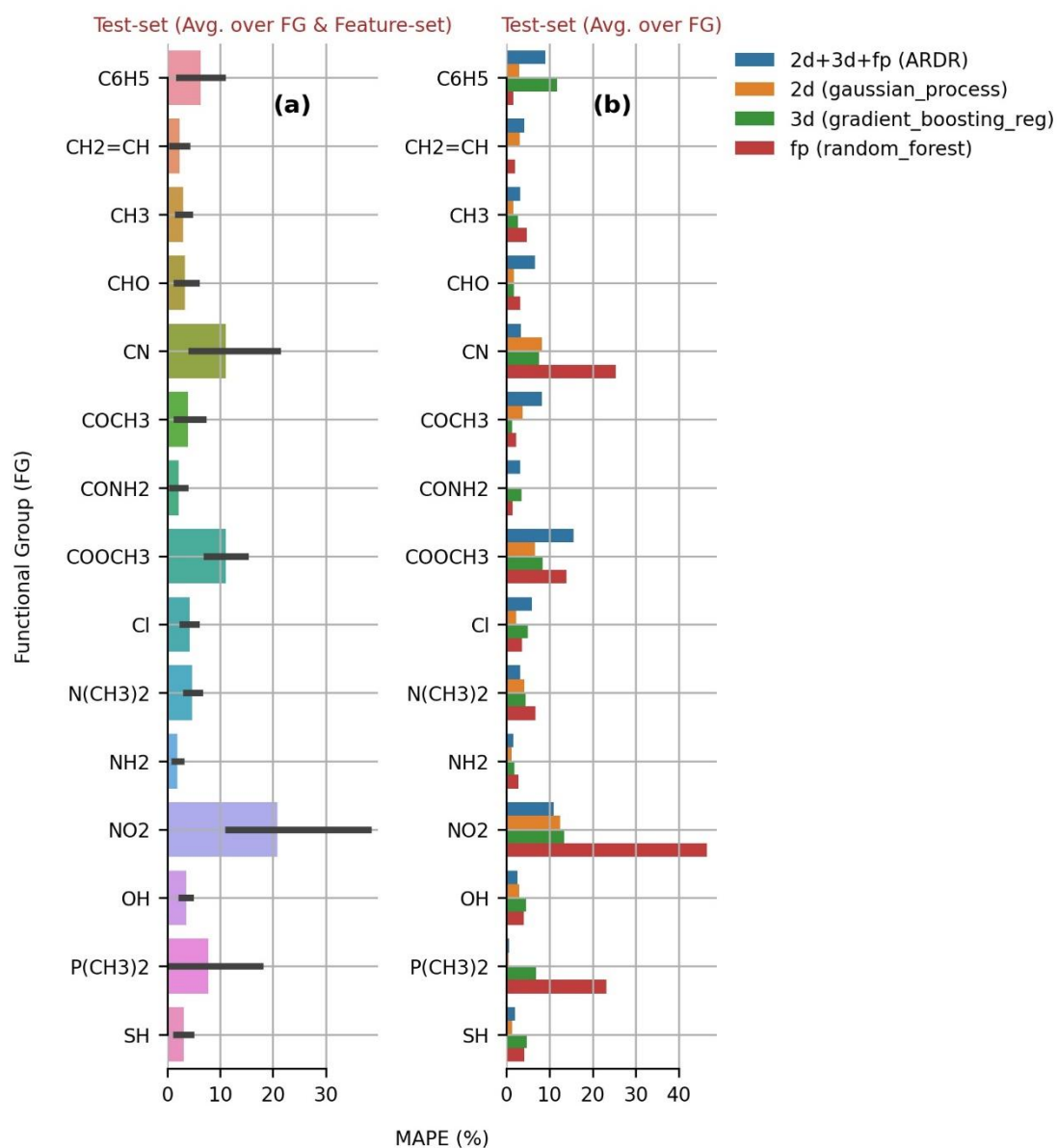
Figure 13. Functional group (FG) vs. Mean Absolute Percentage Error (MAPE) of test-set. (a) MAPE is averaged over FG and feature sets. (b) MAPE is averaged over FG. Test-set predictions were obtained from the best-performing model in the corresponding feature set.

this functional group during the training; still, the error in its prediction was less than 5%. This shows that models successfully learned the hidden patterns between features and redox potential from the training set, resulting in low generalization errors.
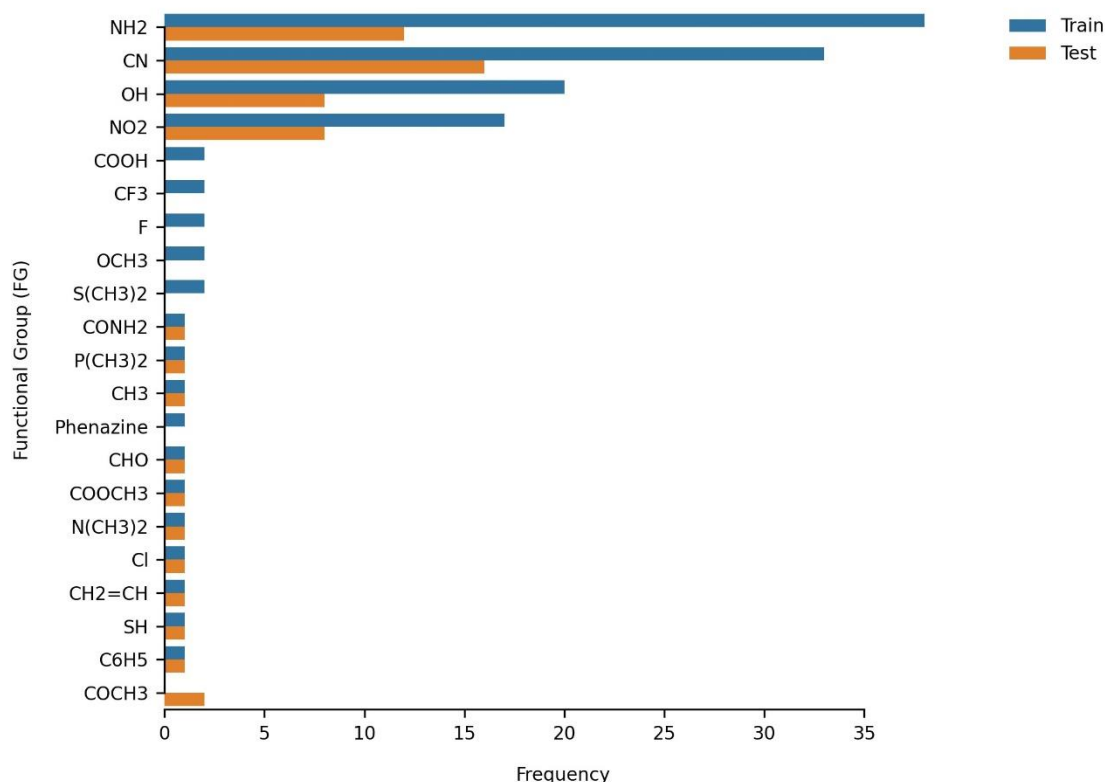
Figure 14. Distribution of functional groups in train and test sets.

Error Analysis

Even though models had low generalization errors, some functional groups exhibited relatively high MAPE, e.g., CN, NO2, and COOCH3. High errors of CN and NO2 are attributed to a small number of compounds with the redox potential close to zero. Unfortunately, the data used in this study contains a very small number of compounds with redox potential close to zero. The whole dataset contains only 18 compounds (~9.7% of the total data) with redox potential greater than -0.5 V (Figure 15a). Therefore, models had less information to learn from in the region near zero redox potential. This is why test-set predictions near-zero redox potential had relatively high errors (Figure 16). On the other hand, having access to a large enough dataset in the region below -0.5 V, models were able to learn the hidden patterns. That resulted in low prediction errors for the compounds with redox potential below -0.5 V, even for the compounds with less than one sample in the train-set. Functional groups CN and NO2 contain some compounds with redox potential greater than -0.5 V (see Figure 13b) which are responsible for high prediction errors observed in Figure 13. We also observed a slight increase in the errors around -1.5 V in Figure 16, which could be attributed to the relatively low number of data points in the region near -1.5 V. This why COOCH3 (avg. redox potential -1.57 V) also showed slightly high prediction error. The red curve in Figure 16 shows the normalized distribution over redox potential (i.e., density) for the whole dataset.
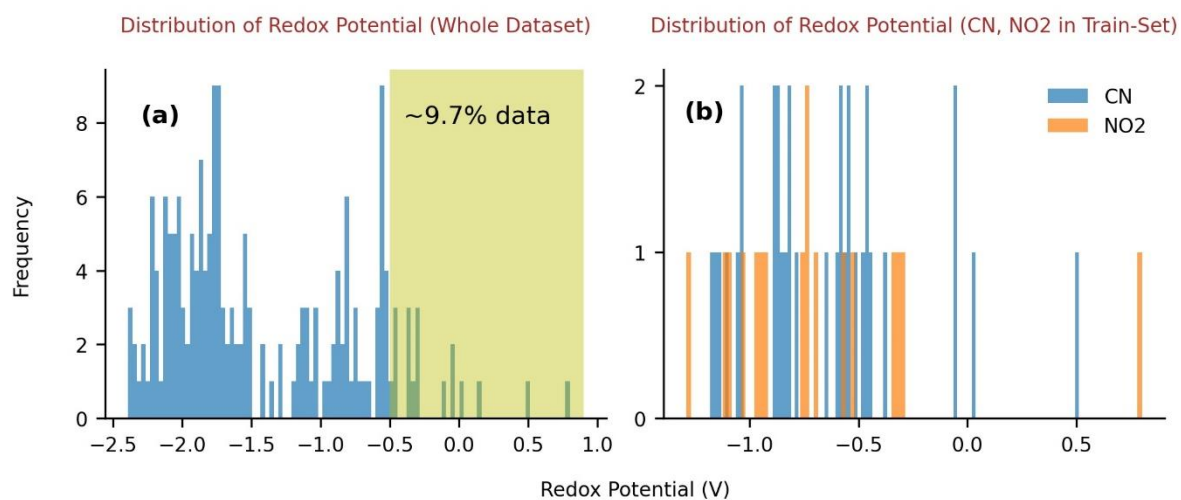
Figure 15. Distribution of redox potential (a) of whole dataset. (b) of CN, NO2 functional groups in train-set
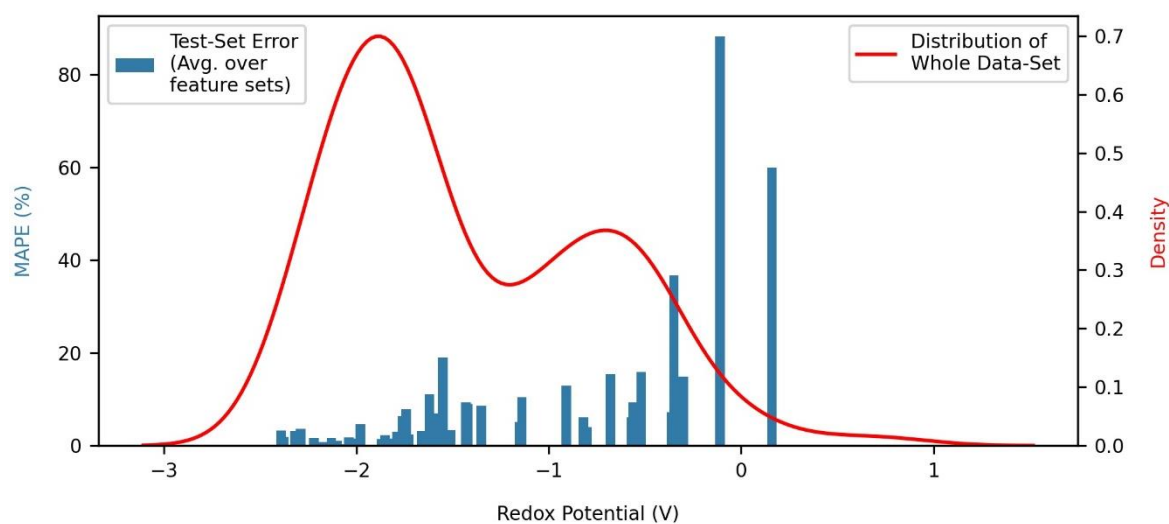

Figure 16. MAPE vs. redox potential. Final MAPE on the y-axis was calculated by averaging the MAPE obtained from the best-performing model in each feature set.

CONCLUSIONS

In this study, we investigated twenty linear and non-linear machine-learning models to predict the redox potential of phenazine derivatives in DME. Both linear and non-linear models trained on a small dataset were able to achieve excellent prediction accuracy on test-set (i.e., $R^2 > 0.98$, MSE < 0.008 V^2, and MAE < 0.07 V). Features used in this study were intentionally chosen to be easily computable from open-source libraries that do not require DFT calculations or experimental measurements, making our approach readily adaptable for similar studies. Model performance was assessed on four feature sets containing different features (i.e., 2D, 3D, and Molecular Fingerprints) using a convenient *pipeline* developed in this work. This *pipeline* combines different training and evaluation components in a single sub-subroutine, making the whole process easy, consistent, and automatic for all models in different scenarios. Gaussian Processes Regression trained on 2D features achieved the highest prediction accuracy. Analysis of model performance on four feature sets revealed the order with respect to prediction accuracy: *2d > 2d+3d+fp > 3d > fp*. Average performance analysis also showed that 2D features are better at generalizing to unseen data than 3D and Molecular Fingerprint features. Therefore, 2D features capture important molecular properties necessary for predicting the redox potential of phenazine derivatives in DME solvent. It was observed that linear models out-perform non-linear models on *'2d+3d+fp'* and *'2d'* feature sets, whereas non-linear models perform better than linear models on *'2d', '3d'* and *'fp'* feature sets. Therefore, linear models should be preferred when the feature set contains different types of features, and non-linear models should be preferred when the feature set contains either 3D or Molecular Fingerprint features. Due to simple structure, linear models have fast training and inference time but suffer from low accuracy. Results obtained here show that lower training and inference times are possible to achieve with the linear models that out-perform non-linear models when data contains different types of features (i.e., 2D, 3D, and Molecular Fingerprints). Feature importance analysis showed features related to Van der Waals surface areas, e.g., *SlogP_VSA4, fr_NH0, VSA_Estate3,* and *VSA_Estate4* were the most important 2D features. *RDF_120, RDF_90, RDF_125, WHIM_90,* and *WHIM_86* were the most important 3D features. *RDKit, ECFP4* were the most important Molecular Fingerprint features. Some features based on molecular structure and charges, e.g., *fr_ArN, MinPartialChargge, and MaxAbsPartialCharge,* were also observed during feature importance analysis. Feature importance analysis also suggested that very few 2D features are required to predict redox potential compared to 3D and Molecular Fingerprint features. This observation was confirmed by re-training models with the subset of features starting from a single feature to a full set of features. Model performance was generally observed to increase with the number of features, but some exceptions were also observed for which the small number of features performed better than a full set of features. Bagging meta-estimator trained on only the top five 2D most important features was able to achieve R^2 value as high as 0.9869. Significant reduction in the training and inference time was observed while maintaining a good level of accuracy. Thus, results obtained in this study would help in reducing the training and inference time for similar future studies on large datasets. MAPE for most functional groups was well below 10%, even for the functional groups with one or zero compounds in the training set. This shows that models were able to successfully learn hidden patterns and generalize quite well to the unseen data. High test errors for three functional groups (CN, NO2, and COOCH3) were observed due to the small number of data points in the region around their average redox potential. With the machine-learning models developed in this study, it is possible to screen millions of new phenazine derivatives in a

reasonable amount of time compared to experimental or DFT methods. Thus, new compounds from large chemical space could be quickly identified as potential redox-active species in RFBs. Furthermore, these models would reduce the number of compounds that need to be analyzed using DFT calculations in hybrid ML-DFT approaches. Thus, we have shown that machine-learning-based approaches can accelerate the discovery of novel materials for energy storage applications.

FUTURE DIRECTIONS

Data used in this study was restricted to only twenty functional groups. Furthermore, redox potentials were obtained in only DME solvent. This restricts the applicability of models to only one solvent. Future studies could expand the dataset to include more functional groups and redox potential determined in various solvents. The primary source of error in this study was the lack of data near zero and positive redox potential. Thus, a large dataset containing compounds with uniform distribution over positive and negative redox potential would significantly reduce generalization errors. This study uses only phenazine derivatives, limiting the applicability of models to phenazine derivatives only. Future studies could combine various compounds to develop a universal model to predict the redox potential of a variety of compounds.

ACKNOWLEDGMENTS

AUTHOR INFORMATION

Corresponding Authors

Siddharth Ghule - Physical and Materials Chemistry Division, CSIR-National Chemical Laboratory (CSIR-NCL), Dr. Homi Bhabha Road, Pashan, Pune 411008, India; Academy of Scientific and Innovative Research (AcSIR), Ghaziabad 201002, India; ORCID iD: https://orcid.org/0000-0003-0864-0777; Email: ss.ghule@ncl.res.in; Phone: +91-20-25903095

Kumar Vanka - Physical and Materials Chemistry Division, CSIR-National Chemical Laboratory (CSIR-NCL), Dr. Homi Bhabha Road, Pashan, Pune 411008, India; Academy of Scientific and Innovative Research (AcSIR), Ghaziabad 201002, India; ORCID iD: http://orcid.org/0000-0001-7301-7573; Email: k.vanka@ncl.res.in; Phone: +91-20-25903095

Authors

Sayan Bagchi - Physical and Materials Chemistry Division, CSIR-National Chemical Laboratory (CSIR-NCL), Pune, 411008, India; Academy of Scientific and Innovative Research (AcSIR), Ghaziabad 201002, India

SUPPORTING INFORMATION

Contains six tables and eight figures

- Tables showing list of all features, ten best-performing models trained on all features from *'2d+3d+fp', '2d', '3d'*, and *'fp'* feature sets, best-performing feature set for each of twenty models
- Figures showing train-set performance of twenty models, test-set performance of twenty models trained on top-5, 10, 15, and 20 features from *'2d', '3d'*, and *'fp'* feature set, functional group (FG) vs. Mean Absolute Percentage Error (MAPE) of train-set

DATA AND CODE AVAILABILITY

The training and test datasets along with the jupyter notebook used for model training is available at https://github.com/siddharth-ncl-work/ML-Redox-Potential.git

REFERENCES

[1]   S. Shafiee and E. Topal, "When will fossil fuel reserves be diminished?," *Energy Policy*, vol. 37, no. 1, pp. 181–189, 2009, doi: 10.1016/j.enpol.2008.08.016.

[2]   A. R. Dehghani-Sanij, E. Tharumalingam, M. B. Dusseault, and R. Fraser, "Study of energy storage systems and environmental challenges of batteries," *Renew. Sustain. Energy Rev.*, vol. 104, no. January, pp. 192–208, 2019, doi: 10.1016/j.rser.2019.01.023.

[3]   M. Höök and X. Tang, "Depletion of fossil fuels and anthropogenic climate change-A review," *Energy Policy*, vol. 52, pp. 797–809, Jan. 2013, doi: 10.1016/j.enpol.2012.10.046.

[4]   T. M. Gür, "Review of electrical energy storage technologies, materials and systems: Challenges and prospects for large-scale grid storage," *Energy Environ. Sci.*, vol. 11, no. 10, pp. 2696–2767, 2018, doi: 10.1039/c8ee01419a.

[5]   W. S. Chu, D. M. Chun, and S. H. Ahn, "Research advancement of green technologies," *Int. J. Precis. Eng. Manuf.*, vol. 15, no. 6, pp. 973–977, 2014, doi: 10.1007/s12541-014-0424-8.

[6]   H. Balat, "Green power for a sustainable future," *Energy Explor. Exploit.*, vol. 25, no. 1, pp. 1–25, 2007, doi: 10.1260/014459807781036403.

[7]   A. Demirbas, "Electrical power production facilities from green energy sources," *Energy Sources, Part B Econ. Plan. Policy*, vol. 1, no. 3, pp. 291–301, 2006, doi: 10.1080/15567240500400648.

[8]   B. Dunn, H. Kamath, and J. M. Tarascon, "Electrical energy storage for the grid: A battery of choices," *Science (80-.).*, vol. 334, no. 6058, pp. 928–935, 2011, doi: 10.1126/science.1212741.

[9]   E. Chung, "What caused the deadly power outages in Texas and how Canada's grid compares," *CBC News*, 2021. https://www.cbc.ca/news/technology/power-outages-texas-canada-1.5920833 (accessed Mar. 30, 2021).

[10]   D. Larcher and J. M. Tarascon, "Towards greener and more sustainable batteries for electrical energy storage," *Nat. Chem.*, vol. 7, no. 1, pp. 19–29, 2015, doi: 10.1038/nchem.2085.

[11]   S. Koohi-Fayegh and M. A. Rosen, "A review of energy storage types, applications and recent developments," *Journal of Energy Storage*, vol. 27. Elsevier Ltd, p. 101047, Feb. 01, 2020, doi: 10.1016/j.est.2019.101047.

[12]   D. Deng, "Li-ion batteries: basics, progress, and challenges," *Energy Sci. Eng.*, vol. 3, no. 5, pp. 385–418, Sep. 2015, doi: 10.1002/ese3.95.

[13]   M. Skyllas-Kazacos, M. H. Chakrabarti, S. A. Hajimolana, F. S. Mjalli, and M. Saleem, "Progress in Flow Battery Research and Development," *J. Electrochem. Soc.*, vol. 158, no. 8, p. R55, 2011, doi: 10.1149/1.3599565.

[14]   P. Leung, X. Li, C. Ponce De León, L. Berlouis, C. T. J. Low, and F. C. Walsh, "Progress in redox flow batteries, remaining challenges and their applications in energy storage," *RSC Adv.*, vol. 2, no. 27, pp. 10125–10156, 2012, doi: 10.1039/c2ra21342g.

[15] E. Sánchez-Díez *et al.*, "Redox flow batteries: Status and perspective towards sustainable stationary energy storage," *J. Power Sources*, vol. 481, p. 228804, Jan. 2021, doi: 10.1016/j.jpowsour.2020.228804.

[16] P. Alotto, M. Guarnieri, and F. Moro, "Redox flow batteries for the storage of renewable energy: A review," *Renew. \& Sustain. ENERGY Rev.*, vol. 29, pp. 325–335, Jan. 2014, doi: 10.1016/j.rser.2013.08.001.

[17] Z. Qi and G. M. Koenig, "Review Article: Flow battery systems with solid electroactive materials," *J. Vac. Sci. Technol. B, Nanotechnol. Microelectron. Mater. Process. Meas. Phenom.*, vol. 35, no. 4, p. 040801, 2017, doi: 10.1116/1.4983210.

[18] C. De La Cruz, A. Molina, N. Patil, E. Ventosa, R. Marcilla, and A. Mavrandonakis, "New insights into phenazine-based organic redox flow batteries by using high-throughput DFT modelling," *Sustain. Energy Fuels*, vol. 4, no. 11, pp. 5513–5521, Nov. 2020, doi: 10.1039/d0se00687d.

[19] S. Gentil, D. Reynard, and H. H. Girault, "Aqueous organic and redox-mediated redox flow batteries: a review," *Current Opinion in Electrochemistry*, vol. 21. Elsevier B.V., pp. 7–13, Jun. 01, 2020, doi: 10.1016/j.coelec.2019.12.006.

[20] P. Leung *et al.*, "Recent developments in organic redox flow batteries: A critical review," *Journal of Power Sources*, vol. 360. Elsevier B.V., pp. 243–283, Aug. 31, 2017, doi: 10.1016/j.jpowsour.2017.05.057.

[21] J. Cao, J. Tian, J. Xu, and Y. Wang, "Organic Flow Batteries: Recent Progress and Perspectives," *Energy and Fuels*, vol. 34, no. 11, pp. 13384–13411, 2020, doi: 10.1021/acs.energyfuels.0c02855.

[22] M. Li, Z. Rhodes, J. R. Cabrera-Pardo, and S. D. Minteer, "Recent advancements in rational design of non-aqueous organic redox flow batteries," *Sustain. Energy Fuels*, vol. 4, no. 9, pp. 4370–4389, 2020, doi: 10.1039/d0se00800a.

[23] K. J. S. Elena I. Romadina, Denis S. Komarov and P. A.Troshin, "New phenazine based anolyte material for high voltage organic redox flow batteries," *Chem. Commun.*, vol. 57, no. 24, 2021, doi: 10.1039/D0CC07951K.

[24] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, "Machine learning for molecular and materials science," *Nature*, vol. 559, no. 7715, pp. 547–555, 2018, doi: 10.1038/s41586-018-0337-2.

[25] J. Schmidt, M. R. G. Marques, S. Botti, and M. A. L. Marques, "Recent advances and applications of machine learning in solid-state materials science," *npj Comput. Mater.*, vol. 5, no. 1, 2019, doi: 10.1038/s41524-019-0221-0.

[26] J. Wei *et al.*, "Machine learning in materials science," *InfoMat*, vol. 1, no. 3, pp. 338–358, 2019, doi: 10.1002/inf2.12028.

[27] G. Pilania, C. Wang, X. Jiang, S. Rajasekaran, and R. Ramprasad, "Accelerating materials property predictions using machine learning," *Sci. Rep.*, vol. 3, pp. 1–6, 2013, doi: 10.1038/srep02810.

[28] R. Batra, "Accurate machine learning in materials science facilitated by using diverse data sources," *Nature*, vol. 589, no. 7843, pp. 524–525, 2021, doi: 10.1038/d41586-020-03259-4.

[29]   R. Gómez-Bombarelli *et al.*, "Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach," *Nat. Mater.*, vol. 15, no. 10, pp. 1120–1127, 2016, doi: 10.1038/nmat4717.

[30]   G. Hautier, C. C. Fischer, A. Jain, T. Mueller, and G. Ceder, "Finding natures missing ternary oxide compounds using machine learning and density functional theory," *Chem. Mater.*, vol. 22, no. 12, pp. 3762–3767, 2010, doi: 10.1021/cm100795d.

[31]   F. A. Faber, A. Lindmaa, O. A. Von Lilienfeld, and R. Armiento, "Machine Learning Energies of 2 Million Elpasolite (ABC2D6) Crystals," *Phys. Rev. Lett.*, vol. 117, no. 13, pp. 2–7, 2016, doi: 10.1103/PhysRevLett.117.135502.

[32]   J. Carrasquilla and R. G. Melko, "Machine learning phases of matter," *Nat. Phys.*, vol. 13, no. 5, pp. 431–434, 2017, doi: 10.1038/nphys4035.

[33]   C. N. Cavasotto and J. I. Di Filippo, "Artificial intelligence in the early stages of drug discovery," *Arch. Biochem. Biophys.*, vol. 698, p. 108730, 2021, doi: https://doi.org/10.1016/j.abb.2020.108730.

[34]   B. G. Peyton, C. Briggs, R. D'Cunha, J. T. Margraf, and T. D. Crawford, "Machine-Learning Coupled Cluster Properties through a Density Tensor Representation," *J. Phys. Chem. A*, vol. 124, no. 23, pp. 4861–4871, 2020, doi: 10.1021/acs.jpca.0c02804.

[35]   A. Seko, H. Hayashi, K. Nakayama, A. Takahashi, and I. Tanaka, "Representation of compounds for machine-learning prediction of physical properties," *Phys. Rev. B*, vol. 95, no. 14, pp. 1–11, 2017, doi: 10.1103/PhysRevB.95.144110.

[36]   S. Sahoo, C. Adhikari, M. Kuanar, and B. Mishra, "A Short Review of the Generation of Molecular Descriptors and Their Applications in Quantitative Structure Property/Activity Relationships," *Curr. Comput. Aided-Drug Des.*, vol. 12, no. 3, pp. 181–205, 2016, doi: 10.2174/1573409912666160525112114.

[37]   D. Fisher, S. R. Lukow, G. Berezutskiy, I. Gil, T. Levy, and Y. Zeiri, "Machine Learning Improves Trace Explosive Selectivity: Application to Nitrate-Based Explosives," *J. Phys. Chem. A*, vol. 124, no. 46, pp. 9656–9664, Nov. 2020, doi: 10.1021/acs.jpca.0c05909.

[38]   S. Nayak, S. Bhattacharjee, J.-H. Choi, and S. Cheol Lee, "Machine Learning and Scaling Laws for Prediction of Accurate Adsorption Energy," *J. Phys. Chem. A*, vol. 124, no. 1, pp. 247–254, Dec. 2019, doi: 10.1021/acs.jpca.9b07569.

[39]   Y. Wei, K. Chin, L. M. Barge, S. Perl, N. Hermis, and T. Wei, "Machine Learning Analysis of the Thermodynamic Responses of In Situ Dielectric Spectroscopy Data in Amino Acids and Inorganic Electrolytes," *J. Phys. Chem. B*, vol. 124, no. 50, pp. 11491–11500, Dec. 2020, doi: 10.1021/acs.jpcb.0c09266.

[40]   M. L. Nisbet *et al.*, "Machine-Learning-Assisted Synthesis of Polar Racemates," *J. Am. Chem. Soc.*, vol. 142, no. 16, pp. 7555–7566, Apr. 2020, doi: 10.1021/jacs.0c01239.

[41]   R. B. Wexler, J. Mark P. Martirez, and A. M. Rappe, "Chemical Pressure-Driven Enhancement of the Hydrogen Evolving Activity of Ni2P from Nonmetal Surface Doping Interpreted via Machine Learning," *J. Am. Chem. Soc.*, vol. 140, no. 13, pp. 4678–4683, Mar. 2018, doi: 10.1021/jacs.8b00947.

[42]   M. H. Lee, "Identification of host-guest systems in green TADF-based OLEDs with

energy level matching based on a machine-learning study," *Phys. Chem. Chem. Phys.*, vol. 22, no. 28, pp. 16378–16386, Jul. 2020, doi: 10.1039/d0cp02871a.

[43] G. Landrum, "RDKit: Open-source cheminformatics." https://www.rdkit.org/.

[44] B. Ramsundar, P. Eastman, P. Walters, V. Pande, K. Leswing, and Z. Wu, *Deep Learning for the Life Sciences*. O'Reilly Media, 2019.

[45] S. Chaube, S. Goverapet Srinivasan, and B. Rai, "Applied machine learning for predicting the lanthanide-ligand binding affinities," *Sci. Rep.*, vol. 10, no. 1, pp. 1–11, 2020, doi: 10.1038/s41598-020-71255-9.

[46] J. A. Pugar, C. M. Childs, C. Huang, K. W. Haider, and N. R. Washburn, "Elucidating the Physicochemical Basis of the Glass Transition Temperature in Linear Polyurethane Elastomers with Machine Learning," *J. Phys. Chem. B*, vol. 124, no. 43, pp. 9722–9733, Sep. 2020, doi: 10.1021/acs.jpcb.0c06439.

[47] A. D. Casey, S. F. Son, I. Bilionis, and B. C. Barnes, "Prediction of energetic material properties from electronic structure using 3D convolutional neural networks," *J. Chem. Inf. Model.*, vol. 60, no. 10, pp. 4457–4473, 2020, doi: 10.1021/acs.jcim.0c00259.

[48] A. J. Minnich *et al.*, "AMPL: A Data-Driven Modeling Pipeline for Drug Discovery," *J. Chem. Inf. Model.*, vol. 60, no. 4, pp. 1955–1968, Apr. 2020, doi: 10.1021/acs.jcim.9b01053.

[49] Y. Okamoto and Y. Kubo, "Ab Initio Calculations of the Redox Potentials of Additives for Lithium-Ion Batteries and Their Prediction through Machine Learning," *ACS Omega*, vol. 3, no. 7, pp. 7868–7874, 2018, doi: 10.1021/acsomega.8b00576.

[50] R. P. Joshi, J. Eickholt, L. Li, M. Fornari, V. Barone, and J. E. Peralta, "Machine Learning the Voltage of Electrode Materials in Metal-Ion Batteries," *ACS Appl. Mater. Interfaces*, vol. 11, no. 20, pp. 18494–18503, 2019, doi: 10.1021/acsami.9b04933.

[51] O. Allam, B. W. Cho, K. C. Kim, and S. S. Jang, "Application of DFT-based machine learning for developing molecular electrode materials in Li-ion batteries," *RSC Adv.*, vol. 8, no. 69, pp. 39414–39420, 2018, doi: 10.1039/c8ra07112h.

[52] Y. Zhang and X. Xu, "Machine Learning Properties of Electrolyte Additives: A Focus on Redox Potentials," *Ind. Eng. Chem. Res.*, vol. 60, no. 1, pp. 343–354, 2021, doi: 10.1021/acs.iecr.0c05055.

[53] O. Allam, R. Kuramshin, Z. Stoichev, B. W. Cho, S. W. Lee, and S. S. Jang, "Molecular structure–redox potential relationship for organic electrode materials: density functional theory–Machine learning approach," *Mater. Today Energy*, vol. 17, p. 100482, 2020, doi: 10.1016/j.mtener.2020.100482.

[54] F. Pedregosa FABIANPEDREGOSA *et al.*, "Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot," 2011. Accessed: Apr. 03, 2021. [Online]. Available: http://scikit-learn.sourceforge.net.

[55] S. A. Wildman and G. M. Crippen, "Prediction of physicochemical parameters by atomic contributions," *J. Chem. Inf. Comput. Sci.*, vol. 39, no. 5, pp. 868–873, 1999, doi: 10.1021/ci990307l.

[56] L. H. Hall, B. Mohney, and L. B. Kier, "The Electrotopological State: Structure

Information at the Atomic Level for Molecular Graphs," 1991. Accessed: Mar. 31, 2021. [Online]. Available: https://pubs.acs.org/sharingguidelines.

[57] L. H. Hall and L. B. Kier, "The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure-Property Modeling," John Wiley & Sons, Ltd, 2007, pp. 367–422.

[58] R. Todeschini and V. Consonni, "Descriptors from Molecular Geometry," in *Handbook of Chemoinformatics*, vol. 3, Weinheim, Germany: Wiley-VCH Verlag GmbH, 2008, pp. 1004–1033.

[59] "Getting Started with the RDKit in Python — The RDKit 2021.03.1 documentation." https://www.rdkit.org/docs/GettingStartedInPython.html#list-of-available-descriptors (accessed Mar. 31, 2021).

[60] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," *J. Chem. Inf. Model.*, vol. 50, no. 5, pp. 742–754, May 2010, doi: 10.1021/ci100050t.

[61] J. L. Durant, B. A. Leland, D. R. Henry, and J. G. Nourse, "Reoptimization of MDL Keys for Use in Drug Discovery," *J. Chem. Inf. Comput. Sci.*, vol. 42, no. 6, pp. 1273–1280, Sep. 2002, doi: 10.1021/ci010132r.