

Predicting the Redox Potentials of Phenazine Derivatives using DFT Assisted Machine Learning

*Siddharth Ghule^{*1,2}, Soumya Ranjan Dash^{1,2}, Sayan Bagchi^{1,2}, Kavita Joshi^{*1,2}, Kumar
Vanka^{*1,2}*

¹Physical and Materials Chemistry Division, CSIR-National Chemical Laboratory (CSIR-NCL),
Dr. Homi Bhabha Road, Pashan, Pune 411008, India

²Academy of Scientific and Innovative Research (AcSIR), Ghaziabad 201002, India

ABSTRACT

Here, four machine-learning models were employed to predict the redox potentials of phenazine derivatives in DME using DFT. A small dataset of 189 phenazine derivatives having only one type of functional group per molecule (20 unique groups) was used for the training. Models were validated on the external test-set containing new functional groups and diverse molecular structures and achieved reasonable accuracies ($R^2 > 0.57$). Despite being trained on the molecules with a single type of functional group, models were able to predict the redox potentials of derivatives containing multiple and different types of functional groups with reasonable accuracy ($R^2 > 0.6$). This type of performance for predicting redox potential from such a small and simple dataset of phenazine derivatives has never been reported before. Redox Flow Batteries (RFBs) are emerging as promising candidates for energy storage systems. However, new green and efficient materials are required for their widespread usage. We believe that the hybrid DFT-ML approach demonstrated in this report would help in accelerating the virtual screening of phenazine derivatives saving computational and experimental resources. This approach could potentially identify novel molecules for green energy storage systems such as RFB.

Keywords: Machine learning, Redox potential, Energy storage

1. 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 major challenges: of balancing increased energy demand while reducing the environmental impact associated with energy production. In the 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 cannot be generated during peak demand. The massive Texas Blackout in February, 2021 is an example of such a failure ⁹. It suggests that efficient energy storage technology is urgently required. Unfortunately, only 1.0% of the energy consumed worldwide can be stored with the energy storage technology accessible today ¹⁰. Furthermore, the contribution of electrochemical batteries to energy storage capacity is less than 2.0%, 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, and fast charge-discharge cycle ^{4,8,11}. Unfortunately, Li-ion batteries suffer from high production costs, safety issues, and high environmental impact ^{2,12}. Redox Flow Batteries (RFBs)

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 impact¹³⁻¹⁵. RFBs are increasingly being used as 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. RFB consists of two storage tanks containing cathode and anode redox-active species dissolved in an electrolyte solution. The electrolyte solution in the positive and negative compartments 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 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 (such as 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 voltage²¹. Recently, phenazine derivatives

have been shown to be promising redox-active candidates in non-aqueous RFBs with high voltage and density. Therefore, phenazine derivatives are currently being investigated as candidates 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 development and prediction times 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 predictions. 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 necessary to predict the target variable. Feature engineering requires domain knowledge, relying on having access to 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.

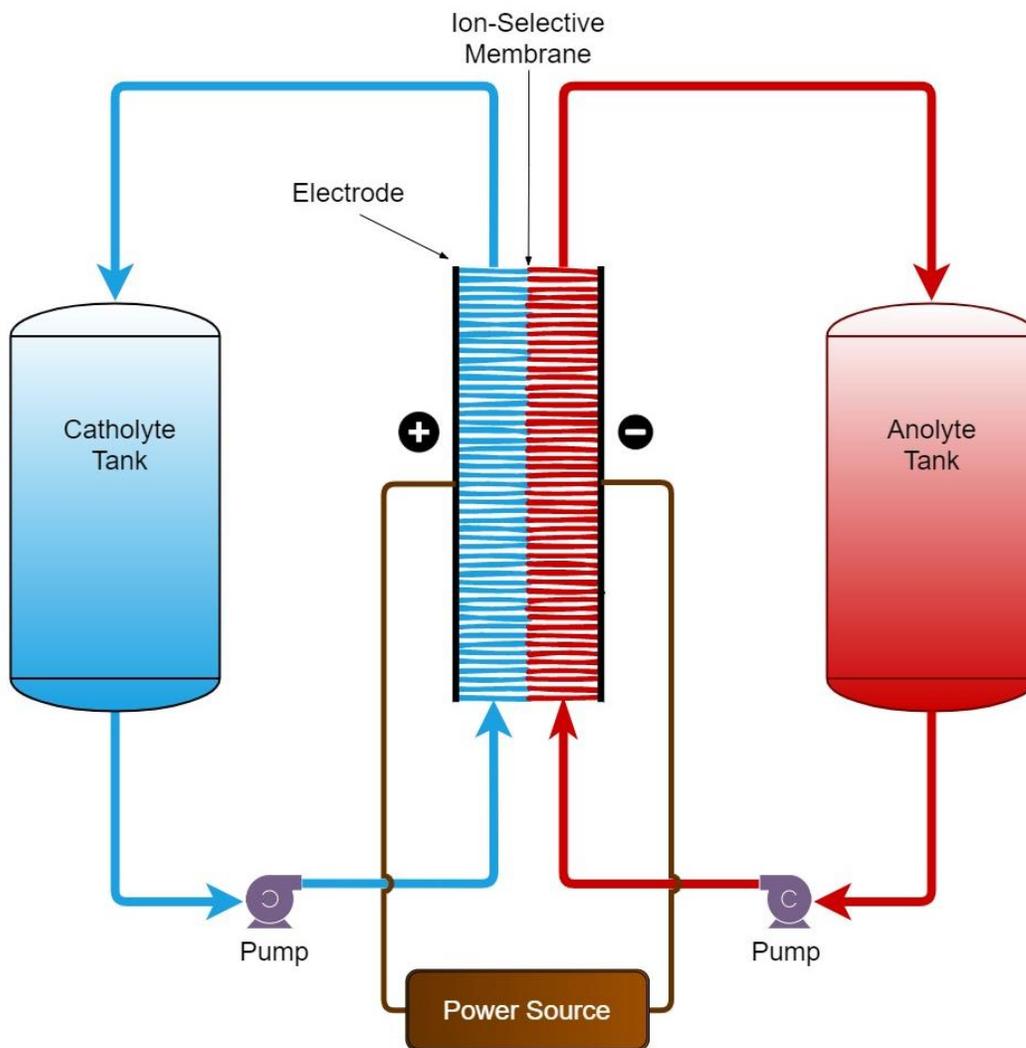


Figure 1. Schematic diagram of a typical redox flow battery

In this work, we investigated four machine learning models to predict the redox potentials of phenazine derivatives in DME solvent (dimethoxyethane). The training dataset was obtained from the previously reported DFT study consisting of 189 phenazine derivatives with only one type of functional group per molecule (20 unique functional groups). Molecular features were computed from the optimized neutral structures using RDKit python library. Then, model

performance was assessed on an external test-set compiled from the literature consisting of new functional groups, multiple functional groups, and diverse structures. Their redox potential was computed using the DFT. Next, the trained models were employed to predict the redox potentials of randomly generated phenazine derivatives with multiple functional groups. Then, we carried out feature importance analysis using Permutation Importance. Finally, we identified promising candidates for anode from the test-sets.

2. MATERIALS AND METHODS

2.1 Computational details

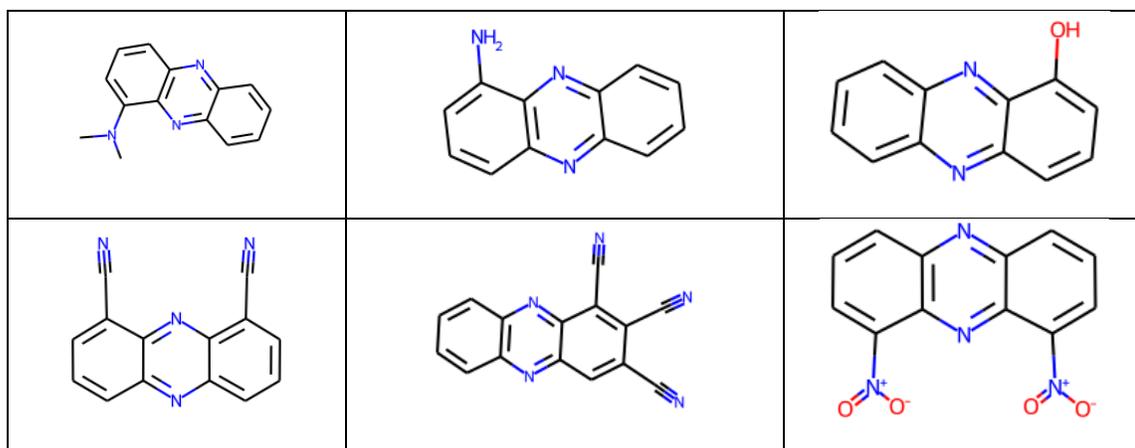
The Redox potential of phenazine derivatives was computed using the DFT workflow described in the paper by Mavrandonakis et al.¹⁸. All DFT calculations (gas phase and DME solvent) were performed with the Gaussian 09 software⁴³. The term ‘Redox Potential’ in this report corresponds to the ‘Reduction Potential’ with respect to the unsubstituted phenazine molecule.

2.2 Data Generation

- **Training-set and Internal test-test:** These datasets were obtained from work reported by Mavrandonakis and co-workers¹⁸. In the report, the redox potential of 189 phenazine derivatives in DME were computed using DFT. These DFT redox potentials were used as a target variable in this work during training and testing. Twenty unique electron-withdrawing and electron-donating functional groups were present in the dataset (-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). It should be noted that phenazine derivatives in this dataset contain only one type of functional group per molecule. Optimized 3D structures of derivatives in neutral and in anionic states were also provided. However, only neutral structures were used in this study. Unfortunately, not all compounds were supplied with their neutral structure, those compounds were modeled, and their optimized structure was added to the dataset. Next, 208 different types of features were generated using RDKit python library⁴⁴. The

list of all features is given in the Table S1 of supporting information. The features were scaled using the *'StandardScaler'* class of the scikit-learn library⁴⁵, which removes the mean and scales each feature to unit variance. Finally, the whole dataset was shuffled and split randomly into a training-set and test-set in an 8:2 ratio (151 samples in the training-set and 38 samples in the test-set). A few phenazine derivatives from the training-set/internal test-set are shown in Table 1.

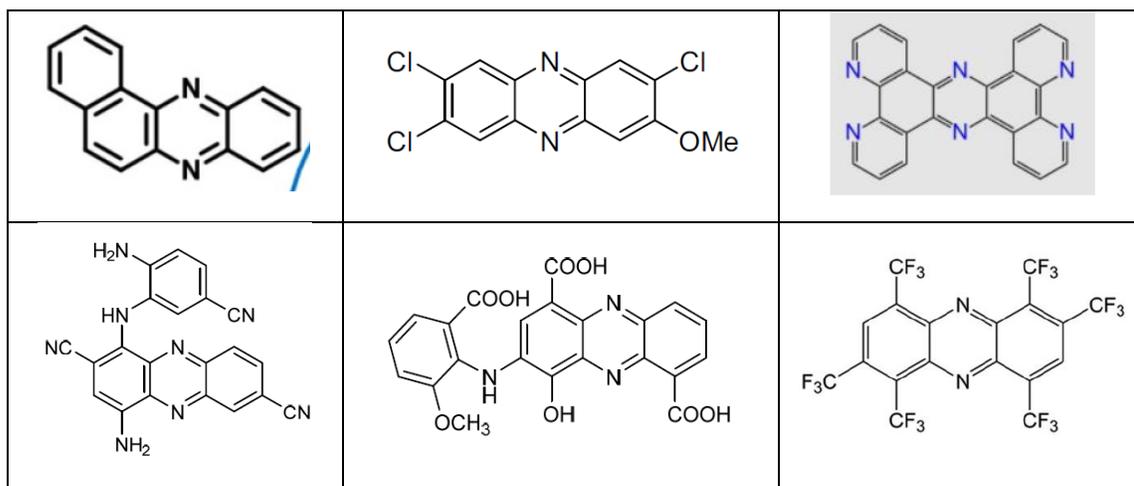
Table 1. Some representative structures from training-set/internal test-set



- **External test-set:** This dataset was compiled from different reports studying various properties of phenazine derivatives⁴⁶⁻⁵⁰. Their redox potentials were computed using DFT and used as a target variable during testing. We gathered a total of 30 phenazine derivatives. Derivatives containing five or more substituted rings were removed. Also, derivatives having drastically different neutral and anion structures were removed. In the end, the external test-set had 22 very diverse phenazine derivatives with multiple types of functional groups. Table 2 shows some of the structures from this dataset. It

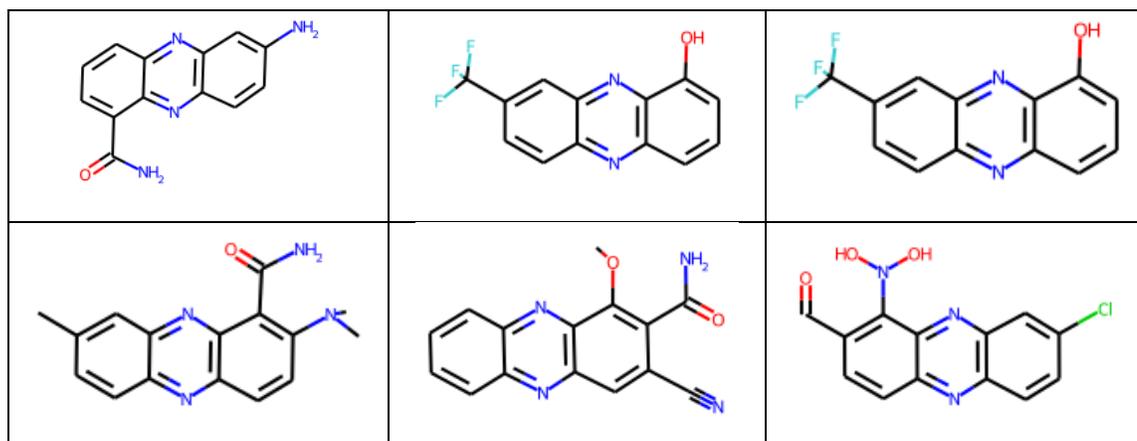
can be seen that this dataset contains unique and different structures from the training-set.

Table 2. Some representative structures from external test-set



- Multiple functional groups dataset:** This dataset was generated by randomly choosing the position and the type of functional group from this list - ($-\text{N}(\text{CH}_3)_2$, $-\text{NH}_2$, $-\text{OH}$, $-\text{OCH}_3$, $-\text{P}(\text{CH}_3)_2$, $-\text{SH}$, $-\text{CH}_3$, $-\text{C}_6\text{H}_5$, $-\text{CH}=\text{CH}_2$, $-\text{F}$, $-\text{Cl}$, $-\text{CHO}$, $-\text{COCH}_3$, $-\text{CONH}_2$, $-\text{COOCH}_3$, $-\text{COOH}$, $-\text{CF}_3$, $-\text{CN}$ and $-\text{NO}_2$). The phenazine derivatives with two and three functional groups and 20 molecules each were generated. Their redox potentials were computed using DFT and used as a target variable during testing. The term ‘Multiple’ refers to the derivatives with different types and more than one functional group in this report. A few representative structures from these datasets are shown in Table 3.

Table 3. Some representative structures from multiple functional groups dataset



2.3 Machine-learning Models

Following four machine-learning models were investigated in this study. These models were chosen due to their ability to generalize from small datasets. Models were implemented with the scikit-learn python library⁴⁵. Hyperparameters of the models were optimized using the ‘*GridSearchCV*’ class of the scikit-learn library. 5-fold cross-validation with mean squared error (MSE) was used for the optimization. The grid of hyperparameters for each model is given the Table S2 of supporting information.

- **Automatic Relevance Determination Regression (ARDR):** This is the probabilistic model related to the sparse Bayesian learning (SBL) framework. It assumes axis-parallel, elliptical Gaussian distribution for each coefficient. The precision of each Gaussian distribution is drawn from the prior distribution (gamma distribution); therefore, it can lead to sparser coefficients. Thus, it is an effective tool to remove irrelevant features^{51,52}.

- **Gaussian Process Regression (GP):** It is the nonparametric Bayesian model. The nonparametric Bayesian model provides the probability distribution of parameters over all possible functions that fit the data. Thus, prior in a Gaussian process is specified on function space. Gaussian process prior is a multivariate normal distribution whose mean is obtained from the data, and covariance is specified using the kernel function. The hyperparameters of the kernel are optimized during the training ^{53,54}. We used a combination of *WhiteKernel* and *RBF* kernel. *WhiteKernel* is used for specifying noise level and *RBF* kernel is a very popular kernel used in many algorithms.
- **Kernel ridge regression (KRR):** It is the extension of ridge regression with kernel trick. In ridge regression, a linear model is learned with the l_2 -norm regularization. Using the kernel trick, KRR learns a linear function in the high dimensional non-linear space without actually transforming the data ⁵⁵.
- **Support Vector Regression (SVR):** This model is the regression form of support vector machine (SVM), a popular algorithm for classification tasks. Analogous to SVM, SVR depends on the subset of training data and ignores the points whose prediction is close to their true value. Therefore, SVM also utilizes kernel trick and learns a hyperplane in the high dimensional space ⁵⁶.

2.4 Evaluation Metrics

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

- Coefficient of Determination (R^2):

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

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

- Mean Squared Error (MSE):

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

- Mean Absolute Error (MAE):

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

The use of terms ‘Accuracy’ and ‘Performance’ in this report is contextual and refers to one or more metrics defined above.

2.5 Feature Importance Analysis

The feature importance analysis was performed using the technique known as Permutation Importance. In this technique, the values of the feature to be assessed are randomly shuffled (permuted). Then, prediction accuracy is computed on the shuffled dataset. Shuffling of feature values is equivalent to replacing the feature with noise, thereby removing its information from the dataset. Therefore, the model is expected to perform poorly on the shuffled dataset if the feature is important. The degree of importance depends on the amount of variation in the accuracy. This technique does not retrain the model; therefore, a trained model is required. The permutation importance was computed using ‘*permutation_importance*’ class of the Scikit-learn library and the training-set⁵⁷. This procedure was repeated 100 times to obtain reliable estimates. The feature importance scores were rescaled between 0 to 1. The mean and standard deviation of

the feature scores were reported. The mean feature score was used for the ranking of individual features. The terms 'Feature' and 'Descriptor' are used interchangeably in this report.

3. RESULTS AND DISCUSSION

3.1 Test-set Performance

We assessed the generalizability of the trained models (i.e., performance on the unseen data) using internal and external test-sets. Please refer to section 2 for the preparation of internal and external test-sets. As the internal test-set comes from the same source, it is very similar to the training-set and contains derivatives with only one type of functional group per molecule. Whereas external test-set is compiled from multiple sources, it has very diverse phenazine derivatives with different types of functional groups. It also contains functional groups and structures not present in the training-set (e.g., -NHPh, -Br, extended conjugation). Figure 2 shows the performance on the internal test-set, and Figure 3 shows the performance on the external test-set. It can be seen that all models have excellent accuracy on internal test-set ($R^2 > 0.97$) and reasonable accuracy on external test-set. Gaussian Process Regression (GP) achieved the highest R^2 of 0.77 on the external test set. After deep analysis in section 3.3, it was revealed that GP is not a stable model while relatively low performing models KRR ($R^2 = 0.58$) and SVR ($R^2 = 0.66$) are more stable. Therefore, one should be careful while using the high performing model, and stability of the model should also be considered. The values of performance metrics on internal and external tests are shown in Table 4. Such a performance on the external test-set is surprising as models were trained on the phenazine derivatives having only one type of functional group. These results show that machine learning models are capable of generalizing from a very small and simple dataset.

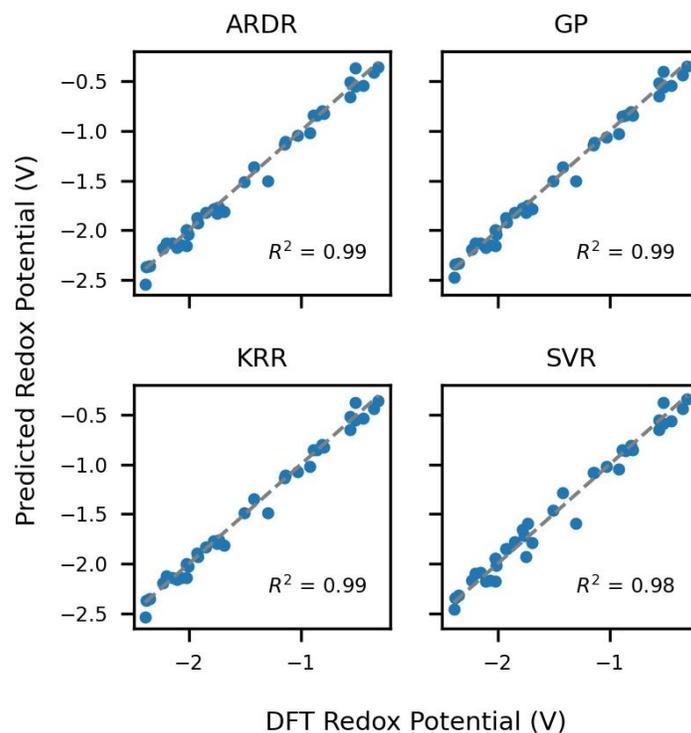


Figure 2. Plots showing machine learning predictions on internal test-set (y-axis) vs. DFT redox potentials (x-axis). Gray dash line corresponds to the perfect predictions.

Table 4. Values of performance metrics on internal and external test-sets. Number were rounded upto two decimals

Model name	Internal test-set			External test-set		
	R^2	MSE	MAE	R^2	MSE	MAE
ARDR	0.99	0.01	0.05	0.57	0.1	0.24
GP	0.99	0	0.05	0.77	0.05	0.16
KRR	0.99	0	0.05	0.58	0.1	0.23
SVR	0.98	0.01	0.08	0.66	0.08	0.22

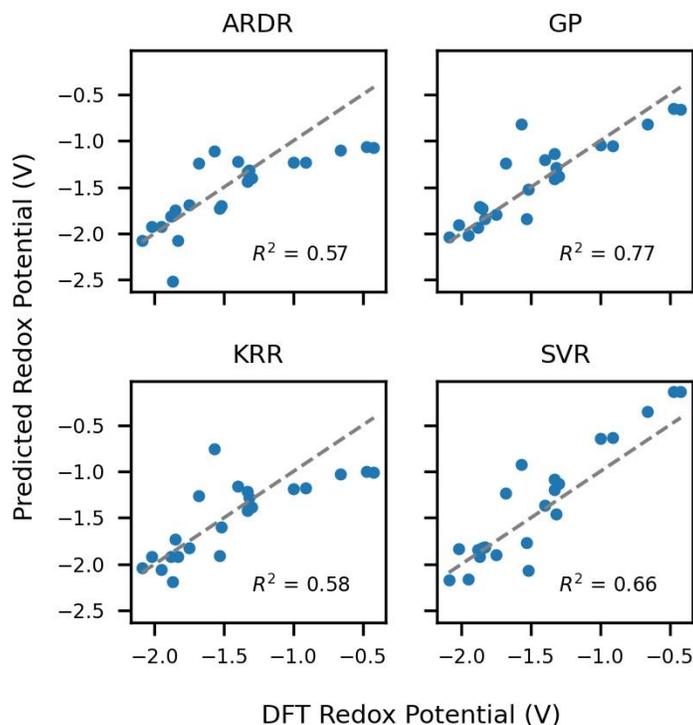


Figure 3. Plots showing machine learning predictions on external test-set (y-axis) vs. DFT redox potentials (x-axis). Gray dash line corresponds to the perfect predictions.

3.2 Prediction on Multiple Functional Groups

Next, we assessed the model performance on the phenazine derivatives substituted with different types of functional groups per molecule. The dataset was generated randomly; please refer to section 2 for the generation of this dataset. Figure 4 and Figure 5 show the performance on the derivatives containing two and three different functional groups, respectively. It can be seen that the models performed reasonably well ($R^2 > 0.6$) even though molecules used for the training had only one type of functional group. In particular, ARDR model achieved the highest performance of $R^2 = 0.7$ and $R^2 = 0.6$ on two and three functional group datasets, respectively. A deeper analysis of ARDR in section 3.3 suggests that ARDR is not a very reliable model.

Although KRR and SVR have relatively low performance, they are more reliable. Therefore, one should be careful while using a high-performing model, and the model's reliability should also be considered. Nevertheless, these results again show the surprising generalization power of machine learning models.

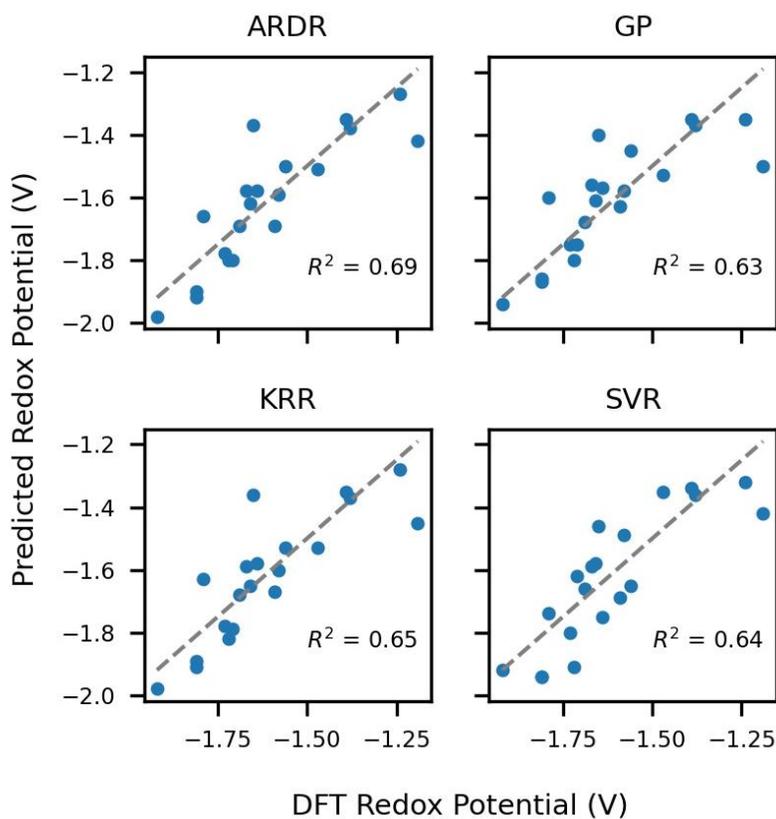


Figure 4. Plots showing machine learning predictions on two different functional groups dataset (y-axis) vs. DFT redox potentials (x-axis). Gray dash line corresponds to the perfect predictions.

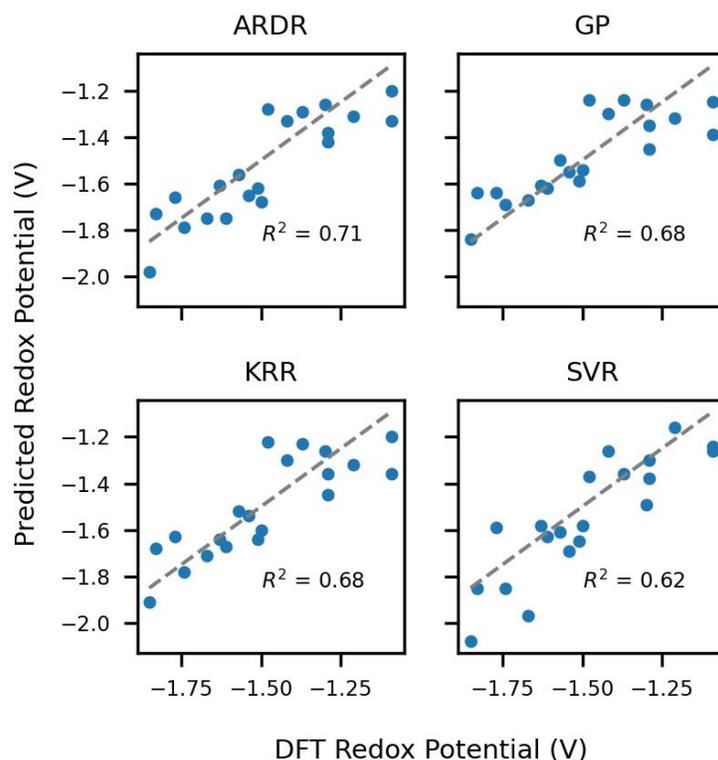


Figure 5. Plots showing machine learning predictions on three different functional groups dataset (y-axis) vs. DFT redox potentials (x-axis). Gray dash line corresponds to the perfect predictions.

Furthermore, we added these randomly generated 20 derivatives containing two different types of functional groups to the training-set and retrained the models on this new dataset of 171 derivatives. The predictive performance of this combined dataset was assessed on the same dataset of 20 derivatives containing three different types of functional groups. The results of this analysis are shown in Figure 6. It can be seen that the model performance has improved significantly with the addition of more data in the training-set.

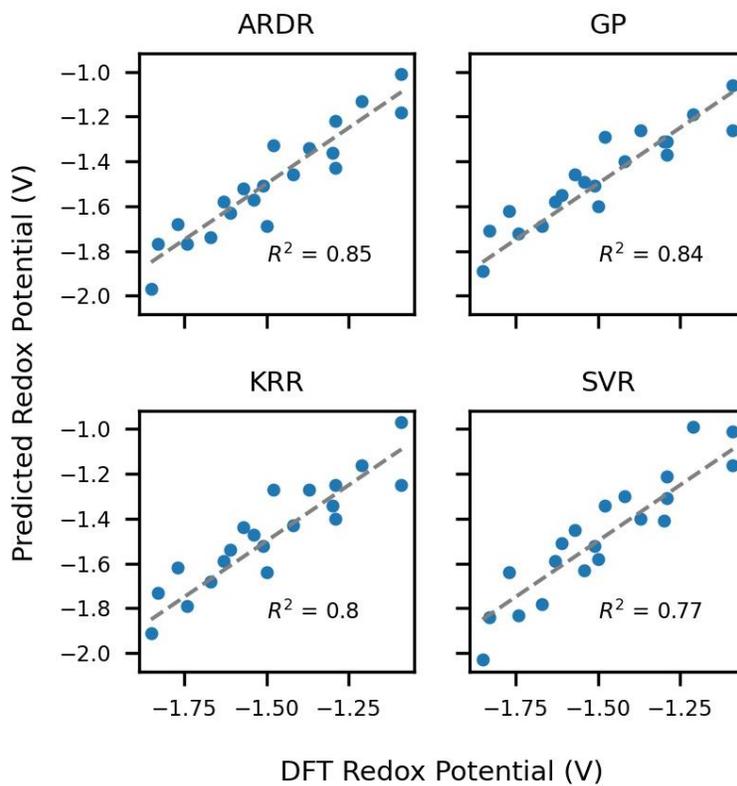


Figure 6. Plots showing machine learning predictions on three different functional groups dataset (y-axis) vs. DFT redox potentials (x-axis). The combined dataset (training-set + two different functional group dataset) was used for the training. Gray dash line corresponds to the perfect predictions.

3.3 Feature Importance Analysis

We carried out feature importance analysis using Permutation Importance. Please refer to section 2 for the details on the technique. In order to understand how model performance changes with the number descriptors, we retrained the models on the subset of features and assessed their performance on the internal test-set. Top 50 features based on their permutation importance score were used. R^2 was used as a performance metric. The result of this analysis is shown in Figure 7. It can be seen that most of the models show a jump in the R^2 and have $R^2 > 0.8$ around the top 10 features. The unusual behavior of GP model is attributed to the instability of the model for a small number of features. The plots in Figure 8 show the histograms of the top ten important features from each model. It is interesting to note that most of the features in ARDR have a very small weight as ARDR tries to prune the large number of irrelevant features leading to a sparse model^{52,58}. Six out of ten features - '*MaxAbsPartialCharge*', '*MinPartialCharge*', '*NHOHCount*', '*PEOE_VSA1*', '*SMR_VSA6*', '*fr_aniline*' are common to all models. Other variations in the feature importance scores could be attributed to the difference in the internal structures of the models. Here, we discuss some of the common features from Figure 8.

- **BCUT2D_MWLOW:** This is the lowest eigenvalue of the connectivity matrix having atomic mass as diagonal elements. Off-diagonal elements depend on the bond order⁵⁹. BCUT2D descriptors have proven very effective for QSAR/QSPR studies due to their good discriminative power^{60,61}. Being dependent on the connectivity matrix and atomic mass, it contains information on molecular size and topology.

- **MaxAbsPartialCharge:** This is the maximum value of the absolute Gasteiger partial charges present in the molecule. In 1980, Gasteiger and Marsili gave the procedure to calculate the partial charges in a molecule. That procedure is known as Partial Equalization of Orbital Electronegativities (PEOE). In this method, the charge is transferred between bonded atoms until equilibrium. The Gasteiger partial charges depend on connectivity and orbital electronegativity, thus capturing the electron-donating and withdrawing power of the atoms ⁶². Electronegativity is essential information as electron-donating groups decrease the redox potential and electron-withdrawing groups increase the redox potential ¹⁸.
- **MinPartialCharge:** This is the minimum value of the Gasteiger partial charges present in the molecule. Please refer to the discussion of *MaxAbsPartialCharge* for the properties of Gasteiger partial charges.
- **PEOE_VSA1:** This is the sum of the approximate accessible van der Waals surface area (i.e., VSA in Å²) of the atoms having partial charge in a specified range ^{63,64}. The partial charges are computed using the PEOE method developed by Gasteiger and Marsili in 1980. Please refer to the discussion of *MaxAbsPartialCharge* for the PEOE method. Thus, this descriptor captures the information related to molecular size and the number of functional groups having partial charge in a specified range.
- **SMR_VSA6:** This is the sum of the approximate accessible van der Waals surface area (i.e., VSA in Å²) of the atoms having molar refractivity in a specified range ^{63,64}. The molar refractivity is computed using the model developed by Crippen in 1999 ⁶⁵. This descriptor contains information on molecular size and polarizability.

- **NHOHCount:** It is the number of N-H and O-H bonds present in the molecule.
- **fr_NH0:** It is the number of tertiary amines present in the molecule.
- **fr_aniline:** It is the number of anilines moieties present in the molecule.

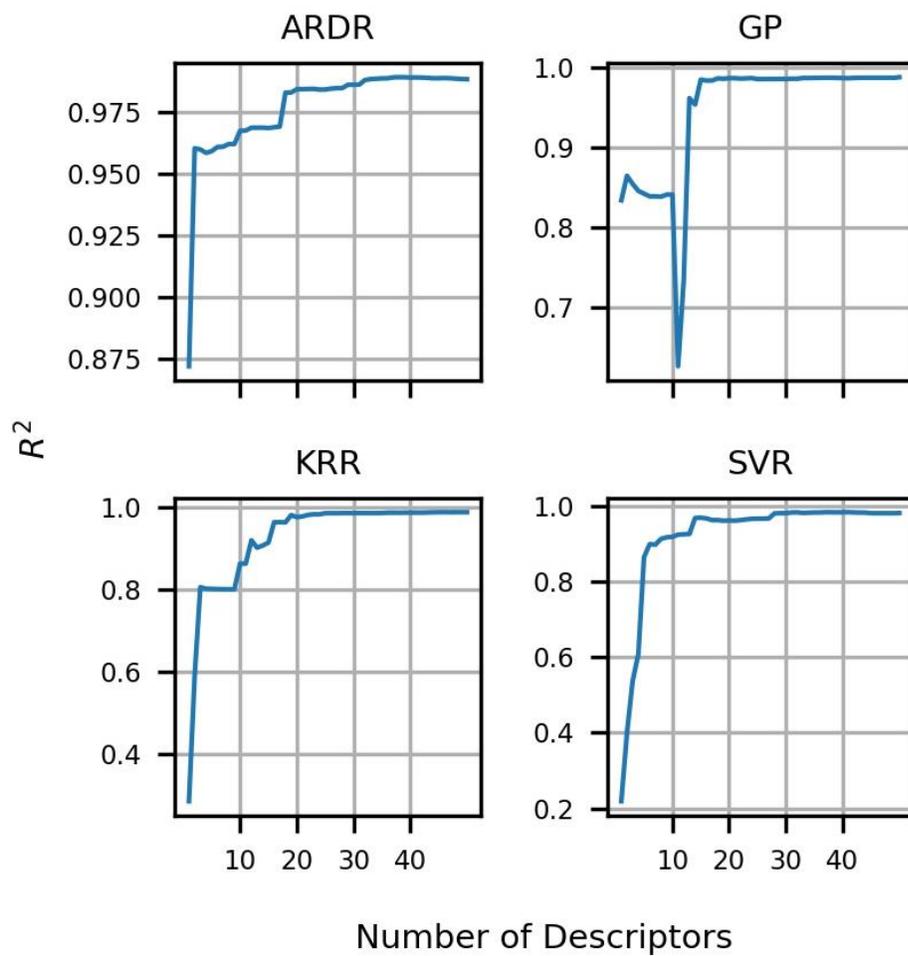


Figure 7. R^2 vs. number of descriptors. R^2 was computed using the internal test-set.

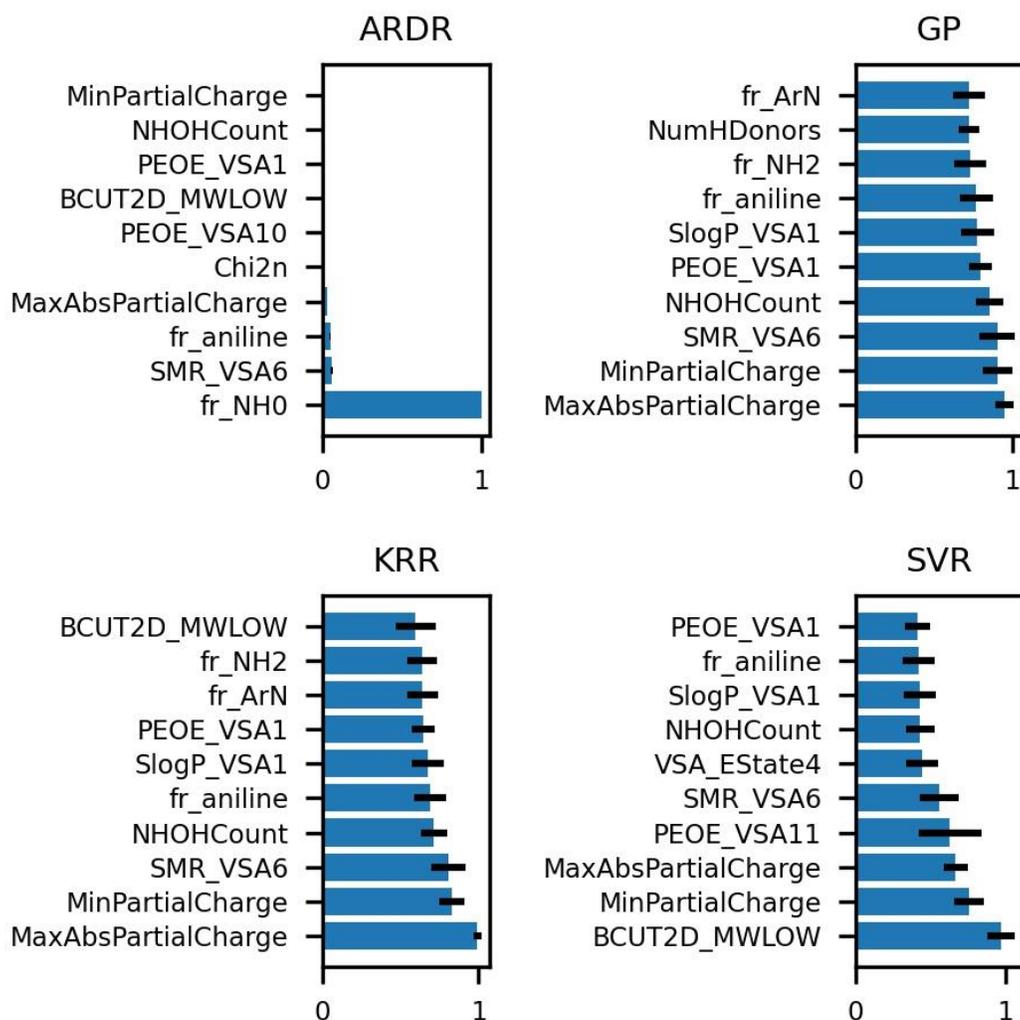


Figure 8. Top ten features (y-axis) vs. mean feature importance score (x-axis). Feature importance scores were rescaled between 0 to 1. Error bars represents standard deviation from 100 repetitions.

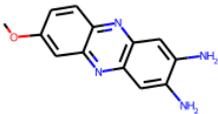
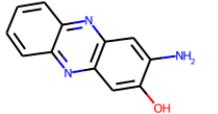
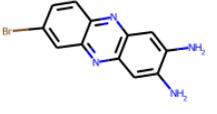
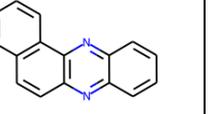
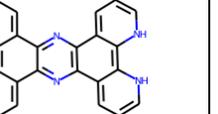
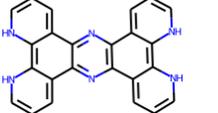
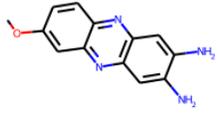
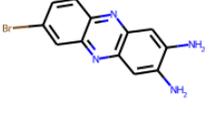
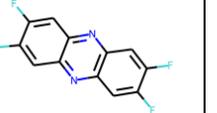
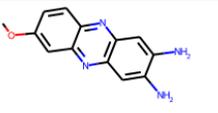
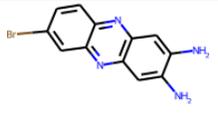
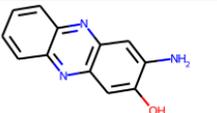
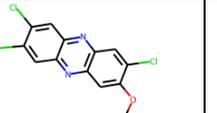
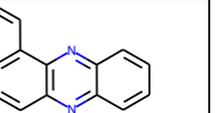
From the analysis in this section, we realized that there are some issues with the ARDR and GP which are outlined below. One should be very careful while using ARDR and GP models.

- **Issues with ARDR model:** As ARDR is related to the sparse Bayesian learning (SBL) framework, it reduces the number of irrelevant features. Unfortunately, in this case, ARDR has put a lot of weight on only one feature, i.e., fr_NH0 (Figure 8). Surprisingly, ARDR also archives an accuracy of more than 0.95 R^2 only with the two features (Figure 7). Although it has shown good performance on the dataset used in this work, it may not work for the broad chemical space. This type of behaviour reduces the reliability of the model.
- **Issues with GP model:** From Figure 7, it can be seen that the model's accuracy decreases with more features, and at around ten features, there is a significant drop in the performance. We also encountered divided by zero errors in the kernel function during this analysis. This shows that GP is not a very stable model.

3.4 Identification of Promising Phenazine Derivatives for Anode

In this section, we identify the top five promising candidates for anode using the trained machine learning models. Electron-donating molecules with negative redox potential are preferred candidates for the anode. As KRR and SVR are stable models, the predictions here are based on them. The values of redox potentials are averaged over 100 independent iterations of data splitting and model training. Table 5 lists the top five phenazine derivatives from the external test-set with the most negative redox potentials obtained from DFT and two machine learning models. 4 out of 5 predictions from KRR and SVR match with DFT predictions. The predictions for other test-sets are shown supporting information in Table S3-S5.

Table 5. Top five anode candidates predicted using DFT, KRR and SVR from the external test-test. SVR and KRR were trained on the phenazine derivatives containing single type of functional group per derivative. Mol ID, and redox potential predicted from DFT and ML models are shown below the respective candidates. Mol IDs were assigned to identify derivatives from the corresponding test-set. Derivatives are arranged in increasing order of redox potential. Redox potential is in the units of Volts.

DFT	 Mol ID: 13 DFT: -2.09	 Mol ID: 29 DFT: -2.02	 Mol ID: 12 DFT: -1.95	 Mol ID: 1 DFT: -1.88	 Mol ID: 5 DFT: -1.87
KRR	 Mol ID: 5 ML: -2.15 DFT: -1.87	 Mol ID: 13 ML: -2.03 DFT: -2.09	 Mol ID: 12 ML: -2.03 DFT: -1.95	 Mol ID: 2 ML: -1.96 DFT: -1.53	 Mol ID: 29 ML: -1.95 DFT: -2.02
SVR	 Mol ID: 13 ML: -2.09 DFT: -2.09	 Mol ID: 12 ML: -2.07 DFT: -1.95	 Mol ID: 29 ML: -1.87 DFT: -2.02	 Mol ID: 3 ML: -1.86 DFT: -1.52	 Mol ID: 1 ML: -1.84 DFT: -1.88

4. CONCLUSIONS

In this study, we trained four machine learning models to predict the redox potentials of phenazine derivatives in DME solvent. Models were trained on a small dataset of 151 phenazine derivatives having only one type of functional group (20 unique functional groups). Trained models showed reasonable accuracy on internal and external test-sets containing a diverse set of phenazine derivatives. We also showed that despite being trained on derivatives with a single type of functional group, models were able to predict the redox potentials of the derivatives having multiple and different types of functional groups to a reasonable accuracy. A small addition of 20 derivatives in the training set significantly improves the accuracy. Finally, we carried out a feature importance study and discussed their essential properties. Deeper analysis also showed that one shouldn't rely only on the performance but also investigate the stability and reliability of the models before prediction. This study shows that it is possible to develop reasonably accurate machine learning models for a relatively complex quantity such as redox potential using a small and simple dataset.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge and contains the following:

Table with the list of all features, Table with the grid of parameters used during hyperparameter optimization. Tables containing the top five candidates for anode from the multiple functional group test-sets.

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

Kavita Joshi - 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/0000-0001-6079-4568> ;Email: k.joshi@ncl.res.in ; Phone: +91-20-25902476

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

Soumya Ranjan Dash - 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

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

Author Contributions

SG and SD have carried out the research work and written the manuscript with the guidance and supervision of KV, KJ and SB.

Notes

The authors declare no competing financial interest.

DATA AND CODE AVAILABILITY

The MOL files of all phenazine derivatives and compiled DFT data are available at https://github.com/siddharth-ncl-work/ml_redox_potential-DATA.git.

ACKNOWLEDGMENTS

K.V. is grateful to the Department of Science and Technology (DST) (EMR/2014/000013) for providing financial assistance. S.B. acknowledges SERB India (EMR/2016/000576). S.G. acknowledges Council of Scientific and Industrial Research (CSIR) for providing Research Fellowship. SD acknowledges CSIR-NCL (MLP101026) for providing a Fellowship. The support and the resources provided by ‘PARAM Brahma Facility’ under the National Supercomputing Mission, Government of India at the Indian Institute of Science Education and Research (IISER) Pune are gratefully acknowledged.

REFERENCES

- (1) Shafiee, S.; Topal, E. When Will Fossil Fuel Reserves Be Diminished? *Energy Policy* **2009**, *37* (1), 181–189. <https://doi.org/10.1016/j.enpol.2008.08.016>.
- (2) Dehghani-Sani, A. R.; Tharumalingam, E.; Dusseault, M. B.; Fraser, R. Study of Energy Storage Systems and Environmental Challenges of Batteries. *Renew. Sustain. Energy Rev.* **2019**, *104* (January), 192–208. <https://doi.org/10.1016/j.rser.2019.01.023>.
- (3) Höök, M.; Tang, X. Depletion of Fossil Fuels and Anthropogenic Climate Change-A Review. *Energy Policy* **2013**, *52*, 797–809. <https://doi.org/10.1016/j.enpol.2012.10.046>.
- (4) Gür, T. M. Review of Electrical Energy Storage Technologies, Materials and Systems: Challenges and Prospects for Large-Scale Grid Storage. *Energy Environ. Sci.* **2018**, *11* (10), 2696–2767. <https://doi.org/10.1039/c8ee01419a>.
- (5) Chu, W. S.; Chun, D. M.; Ahn, S. H. Research Advancement of Green Technologies. *Int. J. Precis. Eng. Manuf.* **2014**, *15* (6), 973–977. <https://doi.org/10.1007/s12541-014-0424-8>.
- (6) Balat, H. Green Power for a Sustainable Future. *Energy Explor. Exploit.* **2007**, *25* (1), 1–25. <https://doi.org/10.1260/014459807781036403>.
- (7) Demirbas, A. Electrical Power Production Facilities from Green Energy Sources. *Energy Sources, Part B Econ. Plan. Policy* **2006**, *1* (3), 291–301. <https://doi.org/10.1080/15567240500400648>.
- (8) Dunn, B.; Kamath, H.; Tarascon, J. M. Electrical Energy Storage for the Grid: A Battery

- of Choices. *Science* (80-). **2011**, 334 (6058), 928–935.
<https://doi.org/10.1126/science.1212741>.
- (9) Chung, E. What caused the deadly power outages in Texas and how Canada’s grid compares <https://www.cbc.ca/news/technology/power-outages-texas-canada-1.5920833> (accessed Mar 30, 2021).
- (10) Larcher, D.; Tarascon, J. M. Towards Greener and More Sustainable Batteries for Electrical Energy Storage. *Nat. Chem.* **2015**, 7 (1), 19–29.
<https://doi.org/10.1038/nchem.2085>.
- (11) Koochi-Fayegh, S.; Rosen, M. A. A Review of Energy Storage Types, Applications and Recent Developments. *Journal of Energy Storage*. Elsevier Ltd February 1, 2020, p 101047. <https://doi.org/10.1016/j.est.2019.101047>.
- (12) Deng, D. Li-ion Batteries: Basics, Progress, and Challenges. *Energy Sci. Eng.* **2015**, 3 (5), 385–418. <https://doi.org/10.1002/ese3.95>.
- (13) Skyllas-Kazacos, M.; Chakrabarti, M. H.; Hajimolana, S. A.; Mjalli, F. S.; Saleem, M. Progress in Flow Battery Research and Development. *J. Electrochem. Soc.* **2011**, 158 (8), R55. <https://doi.org/10.1149/1.3599565>.
- (14) Leung, P.; Li, X.; Ponce De León, C.; Berlouis, L.; Low, C. T. J.; Walsh, F. C. Progress in Redox Flow Batteries, Remaining Challenges and Their Applications in Energy Storage. *RSC Adv.* **2012**, 2 (27), 10125–10156. <https://doi.org/10.1039/c2ra21342g>.
- (15) Sánchez-Díez, E.; Ventosa, E.; Guarnieri, M.; Trovò, A.; Flox, C.; Marcilla, R.; Soavi, F.;

- Mazur, P.; Aranzabe, E.; Ferret, R. Redox Flow Batteries: Status and Perspective towards Sustainable Stationary Energy Storage. *J. Power Sources* **2021**, *481*, 228804. <https://doi.org/10.1016/j.jpowsour.2020.228804>.
- (16) Alotto, P.; Guarnieri, M.; Moro, F. Redox Flow Batteries for the Storage of Renewable Energy: A Review. *Renew. \& Sustain. ENERGY Rev.* **2014**, *29*, 325–335. <https://doi.org/10.1016/j.rser.2013.08.001>.
- (17) Qi, Z.; Koenig, G. M. Review Article: Flow Battery Systems with Solid Electroactive Materials. *J. Vac. Sci. Technol. B, Nanotechnol. Microelectron. Mater. Process. Meas. Phenom.* **2017**, *35* (4), 040801. <https://doi.org/10.1116/1.4983210>.
- (18) De La Cruz, C.; Molina, A.; Patil, N.; Ventosa, E.; Marcilla, R.; Mavrandonakis, A. New Insights into Phenazine-Based Organic Redox Flow Batteries by Using High-Throughput DFT Modelling. *Sustain. Energy Fuels* **2020**, *4* (11), 5513–5521. <https://doi.org/10.1039/d0se00687d>.
- (19) Gentil, S.; Reynard, D.; Girault, H. H. Aqueous Organic and Redox-Mediated Redox Flow Batteries: A Review. *Current Opinion in Electrochemistry*. Elsevier B.V. June 1, 2020, pp 7–13. <https://doi.org/10.1016/j.coelec.2019.12.006>.
- (20) Leung, P.; Shah, A. A.; Sanz, L.; Flox, C.; Morante, J. R.; Xu, Q.; Mohamed, M. R.; Ponce de León, C.; Walsh, F. C. Recent Developments in Organic Redox Flow Batteries: A Critical Review. *Journal of Power Sources*. Elsevier B.V. August 31, 2017, pp 243–283. <https://doi.org/10.1016/j.jpowsour.2017.05.057>.

- (21) Cao, J.; Tian, J.; Xu, J.; Wang, Y. Organic Flow Batteries: Recent Progress and Perspectives. *Energy and Fuels* **2020**, *34* (11), 13384–13411. <https://doi.org/10.1021/acs.energyfuels.0c02855>.
- (22) Li, M.; Rhodes, Z.; Cabrera-Pardo, J. R.; Minter, S. D. Recent Advancements in Rational Design of Non-Aqueous Organic Redox Flow Batteries. *Sustain. Energy Fuels* **2020**, *4* (9), 4370–4389. <https://doi.org/10.1039/d0se00800a>.
- (23) Elena I. Romadina, Denis S. Komarov, K. J. S.; A.Troshin, P. New Phenazine Based Anolyte Material for High Voltage Organic Redox Flow Batteries. *Chem. Commun.* **2021**, 57 (24). <https://doi.org/10.1039/D0CC07951K>.
- (24) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine Learning for Molecular and Materials Science. *Nature* **2018**, *559* (7715), 547–555. <https://doi.org/10.1038/s41586-018-0337-2>.
- (25) Schmidt, J.; Marques, M. R. G.; Botti, S.; Marques, M. A. L. Recent Advances and Applications of Machine Learning in Solid-State Materials Science. *npj Comput. Mater.* **2019**, *5* (1). <https://doi.org/10.1038/s41524-019-0221-0>.
- (26) Wei, J.; Chu, X.; Sun, X.; Xu, K.; Deng, H.; Chen, J.; Wei, Z.; Lei, M. Machine Learning in Materials Science. *InfoMat* **2019**, *1* (3), 338–358. <https://doi.org/10.1002/inf2.12028>.
- (27) Pilia, G.; Wang, C.; Jiang, X.; Rajasekaran, S.; Ramprasad, R. Accelerating Materials Property Predictions Using Machine Learning. *Sci. Rep.* **2013**, *3*, 1–6. <https://doi.org/10.1038/srep02810>.

- (28) Batra, R. Accurate Machine Learning in Materials Science Facilitated by Using Diverse Data Sources. *Nature* **2021**, *589* (7843), 524–525. <https://doi.org/10.1038/d41586-020-03259-4>.
- (29) Gómez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Duvenaud, D.; Maclaurin, D.; Blood-Forsythe, M. A.; Chae, H. S.; Einzinger, M.; Ha, D. G.; Wu, T.; Markopoulos, G.; Jeon, S.; Kang, H.; Miyazaki, H.; Numata, M.; Kim, S.; Huang, W.; Hong, S. I.; Baldo, M.; Adams, R. P.; Aspuru-Guzik, A. Design of Efficient Molecular Organic Light-Emitting Diodes by a High-Throughput Virtual Screening and Experimental Approach. *Nat. Mater.* **2016**, *15* (10), 1120–1127. <https://doi.org/10.1038/nmat4717>.
- (30) Hautier, G.; Fischer, C. C.; Jain, A.; Mueller, T.; Ceder, G. Finding Natures Missing Ternary Oxide Compounds Using Machine Learning and Density Functional Theory. *Chem. Mater.* **2010**, *22* (12), 3762–3767. <https://doi.org/10.1021/cm100795d>.
- (31) Faber, F. A.; Lindmaa, A.; Von Lilienfeld, O. A.; Armiento, R. Machine Learning Energies of 2 Million Elpasolite (ABC2D6) Crystals. *Phys. Rev. Lett.* **2016**, *117* (13), 2–7. <https://doi.org/10.1103/PhysRevLett.117.135502>.
- (32) Carrasquilla, J.; Melko, R. G. Machine Learning Phases of Matter. *Nat. Phys.* **2017**, *13* (5), 431–434. <https://doi.org/10.1038/nphys4035>.
- (33) Cavasotto, C. N.; Di Filippo, J. I. Artificial Intelligence in the Early Stages of Drug Discovery. *Arch. Biochem. Biophys.* **2021**, *698*, 108730. <https://doi.org/https://doi.org/10.1016/j.abb.2020.108730>.

- (34) Peyton, B. G.; Briggs, C.; D’Cunha, R.; Margraf, J. T.; Crawford, T. D. Machine-Learning Coupled Cluster Properties through a Density Tensor Representation. *J. Phys. Chem. A* **2020**, *124* (23), 4861–4871. <https://doi.org/10.1021/acs.jpca.0c02804>.
- (35) Seko, A.; Hayashi, H.; Nakayama, K.; Takahashi, A.; Tanaka, I. Representation of Compounds for Machine-Learning Prediction of Physical Properties. *Phys. Rev. B* **2017**, *95* (14), 1–11. <https://doi.org/10.1103/PhysRevB.95.144110>.
- (36) Sahoo, S.; Adhikari, C.; Kuanar, M.; Mishra, B. A Short Review of the Generation of Molecular Descriptors and Their Applications in Quantitative Structure Property/Activity Relationships. *Curr. Comput. Aided-Drug Des.* **2016**, *12* (3), 181–205. <https://doi.org/10.2174/1573409912666160525112114>.
- (37) Zeiri, Y.; Fisher, D.; Lukow, S. R.; Berezutskiy, G.; Gil, I.; Levy, T. Machine Learning Improves Trace Explosive Selectivity: Application to Nitrate-Based Explosives. *J. Phys. Chem. A* **2020**, *124* (46), 9656–9664. <https://doi.org/10.1021/acs.jpca.0c05909>.
- (38) Nayak, S.; Bhattacharjee, S.; Choi, J.-H.; Cheol Lee, S. Machine Learning and Scaling Laws for Prediction of Accurate Adsorption Energy. *J. Phys. Chem. A* **2019**, *124* (1), 247–254. <https://doi.org/10.1021/acs.jpca.9b07569>.
- (39) Wei, Y.; Chin, K.; M. Barge, L.; Perl, S.; Hermis, N.; Wei, T. Machine Learning Analysis of the Thermodynamic Responses of In Situ Dielectric Spectroscopy Data in Amino Acids and Inorganic Electrolytes. *J. Phys. Chem. B* **2020**, *124* (50), 11491–11500. <https://doi.org/10.1021/acs.jpcc.0c09266>.

- (40) L. Nisbet, M.; M. Pendleton, I.; M. Nolis, G.; J. Griffith, K.; Schrier, J.; Cabana, J.; J. Norquist, A.; R. Poeppelmeier, K. Machine-Learning-Assisted Synthesis of Polar Racemates. *J. Am. Chem. Soc.* **2020**, *142* (16), 7555–7566. <https://doi.org/10.1021/jacs.0c01239>.
- (41) Wexler, R. B.; Mark P. Martirez, J.; M. Rappe, A. Chemical Pressure-Driven Enhancement of the Hydrogen Evolving Activity of Ni₂P from Nonmetal Surface Doping Interpreted via Machine Learning. *J. Am. Chem. Soc.* **2018**, *140* (13), 4678–4683. <https://doi.org/10.1021/jacs.8b00947>.
- (42) Lee, M. H. Identification of Host-Guest Systems in Green TADF-Based OLEDs with Energy Level Matching Based on a Machine-Learning Study. *Phys. Chem. Chem. Phys.* **2020**, *22* (28), 16378–16386. <https://doi.org/10.1039/d0cp02871a>.
- (43) M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ort, and D. J. F. Gaussian 09. Gaussian, Inc.: Wallingford CT 2016.
- (44) Landrum, G. RDKit: Open-source cheminformatics <https://www.rdkit.org/>.
- (45) Pedregosa FABIANPEDREGOSA, F.; Michel, V.; Grisel OLIVIERGRISEL, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Vanderplas, J.; Cournapeau, D.; Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Thirion, B.; Grisel, O.; Dubourg, V.; Passos, A.; Brucher, M.; Perrot and Édouardand, M.; Duchesnay, A.; Duchesnay EDOUARDDUCHESNAY,

- Fré. Scikit-Learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot. *J. Mach. Learn. Res.* **2011**, *12* (85), 2825–2830.
- (46) Nakagawa, R.; Nishina, Y. Simulating the Redox Potentials of Unexplored Phenazine Derivatives as Electron Mediators for Biofuel Cells. *J. Phys. Energy* **2021**, *3* (3), 034008. <https://doi.org/10.1088/2515-7655/ABEBC8>.
- (47) Miao, L.; Liu, L.; Zhang, K.; Chen, J. Molecular Design Strategy for High-Redox-Potential and Poorly Soluble n-Type Phenazine Derivatives as Cathode Materials for Lithium Batteries. *ChemSusChem* **2020**, *13* (9), 2337–2344. <https://doi.org/10.1002/CSSC.202000004>.
- (48) Sousa, A. C.; Martins, L. O.; Robalo, M. P. Laccases: Versatile Biocatalysts for the Synthesis of Heterocyclic Cores. *Mol. 2021, Vol. 26, Page 3719* **2021**, *26* (12), 3719. <https://doi.org/10.3390/MOLECULES26123719>.
- (49) Castro, K. P.; Clikeman, T. T.; DeWeerd, N. J.; Bukovsky, E. V.; Rippy, K. C.; Kuvychko, I. V.; Hou, G.-L.; Chen, Y.-S.; Wang, X.-B.; Strauss, S. H.; Boltalina, O. V. Incremental Tuning Up of Fluorous Phenazine Acceptors. *Chem. – A Eur. J.* **2016**, *22* (12), 3930–3936. <https://doi.org/10.1002/CHEM.201504122>.
- (50) Wang, C.; Li, X.; Yu, B.; Wang, Y.; Yang, Z.; Wang, H.; Lin, H.; Ma, J.; Li, G.; Jin, Z. Molecular Design of Fused-Ring Phenazine Derivatives for Long-Cycling Alkaline Redox Flow Batteries. *ACS Energy Lett.* **2020**, *17*, 411–417.

<https://doi.org/10.1021/ACSENERGYLETT.9B02676>.

- (51) 1.1. Linear Models — scikit-learn 1.0 documentation https://scikit-learn.org/stable/modules/linear_model.html#bayesian-regression (accessed Oct 23, 2021).
- (52) Wipf, D.; Nagarajan, S. A New View of Automatic Relevance Determination. *Adv. Neural Inf. Process. Syst.* **2007**, *20*.
- (53) 1.7. Gaussian Processes — scikit-learn 1.0 documentation https://scikit-learn.org/stable/modules/gaussian_process.html (accessed Oct 23, 2021).
- (54) Quick Start to Gaussian Process Regression | by Hilarie Sit | Towards Data Science <https://towardsdatascience.com/quick-start-to-gaussian-process-regression-36d838810319> (accessed Oct 23, 2021).
- (55) 1.3. Kernel ridge regression — scikit-learn 1.0 documentation https://scikit-learn.org/stable/modules/kernel_ridge.html (accessed Oct 23, 2021).
- (56) 1.4. Support Vector Machines — scikit-learn 1.0 documentation <https://scikit-learn.org/stable/modules/svm.html#svm-regression> (accessed Oct 23, 2021).
- (57) 4.2. Permutation feature importance — scikit-learn 1.0 documentation https://scikit-learn.org/stable/modules/permutation_importance.html#permutation-importance (accessed Oct 23, 2021).
- (58) 1.1. Linear Models — scikit-learn 1.0 documentation https://scikit-learn.org/stable/modules/linear_model.html (accessed Oct 21, 2021).

- (59) Burden, F. R. Molecular Identification Number for Substructure Searches. *J. Chem. Inf. Comput. Sci.* **1989**, *29* (3), 225–227. <https://doi.org/10.1021/CI00063A011>.
- (60) Stanton†, D. T. Evaluation and Use of BCUT Descriptors in QSAR and QSPR Studies. *J. Chem. Inf. Comput. Sci.* **1998**, *39* (1), 11–20. <https://doi.org/10.1021/CI980102X>.
- (61) B, P.; SD, P. Classification of Kinase Inhibitors Using BCUT Descriptors. *J. Chem. Inf. Comput. Sci.* **2000**, *40* (6), 1431–1440. <https://doi.org/10.1021/CI000386X>.
- (62) Gasteiger, J.; Marsili, M. Iterative Partial Equalization of Orbital Electronegativity—a Rapid Access to Atomic Charges. *Tetrahedron* **1980**, *36* (22), 3219–3228. [https://doi.org/10.1016/0040-4020\(80\)80168-2](https://doi.org/10.1016/0040-4020(80)80168-2).
- (63) Landrum, G. Getting Started with the RDKit in Python — The RDKit 2020.03.1 documentation <https://www.rdkit.org/docs/GettingStartedInPython.html#list-of-available-descriptors> (accessed Mar 31, 2021).
- (64) QuaSAR-Descriptor <http://www.cadaster.eu/sites/cadaster.eu/files/challenge/descr.htm> (accessed Oct 22, 2021).
- (65) Wildman, S. A.; Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (5), 868–873. <https://doi.org/10.1021/ci9903071>.

For Table of Contents Use Only

Predicting the Redox Potentials of Phenazine Derivatives using DFT Assisted Machine Learning

Siddharth Ghule*, Soumya Ranjan Dash, Sayan Bagchi, Kavita Joshi*, Kumar Vanka*

