Materials Today Energy 17 (2020) 100482

Contents lists available at ScienceDirect

Materials Today Energy

journal homepage: www.journals.elsevier.com/materials-today-energy/

# Molecular structure–redox potential relationship for organic electrode materials: density functional theory–Machine learning approach

O. Allam [a, b, g], R. Kuramshin [a, c, g], Z. Stoichev [a, b, g], B.W. Cho [b], S.W. Lee [b], S.S. Jang [a, d, e, f, *]

[a] Computational NanoBio Technology Laboratory, School of Materials Science and Engineering, Georgia Institute of Technology, Atlanta, GA, 30332-0245, USA
[b] The George W. Woodruff School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, GA, 30332-0405, USA
[c] College of Computing, Georgia Institute of Technology, Atlanta, GA, 30332-0765, USA
[d] Institute for Electronics and Nanotechnology, Georgia Institute of Technology, Atlanta, GA, 30332, USA
[e] Parker H. Petit Institute for Bioengineering and Bioscience, Georgia Institute of Technology, Atlanta, GA, 30332, USA
[f] Strategic Energy Institute, Georgia Institute of Technology, Atlanta, GA, 30332, USA

ARTICLE INFO

*Article history:*
Received 24 May 2020
Received in revised form
22 June 2020
Accepted 9 July 2020
Available online 15 July 2020

*Keywords:*
Electrochemistry
Li-ion battery
Quantum mechanics
Molecular modeling
Molecular design

ABSTRACT

In this study we develop a high-throughput screening method by employing a density functional theory (DFT) - machine learning (ML) framework for the design of novel organic electrode materials. For this purpose, DFT modeling is performed to calculate basic electronic properties of various organic compounds, namely redox potential, electron affinity, highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO), which are used in conjunction with basic molecular descriptors to train three machine learning models (ML): artificial neural networks (ANN), gradient-boosting regression (GBR), and kernel ridge regression (KRR) through three different protocols. These three protocols, or ***pipelines***, are developed in order to enhance each model's capability to learn the data and make predictions. The first two ***pipelines*** utilize the original features only, while the third ***pipeline*** utilizes composite features which are screened by a least absolute shrinkage and selection operator (LASSO). Particularly, the second and third ***pipelines*** employ a Pearson correlation analysis in conjunction with recursive feature elimination (RFE). From this study, the most important features to predict redox potential are identified as the electron affinity and the number of bound Li atoms. After optimizing machine learning models in each ***pipeline***, it is found that KRR predicts the redox potential with the highest accuracy.

## 1. Introduction

As global interest and investment in renewable energy resources have grown very rapidly, the demand for effective energy storage technologies has also been enormously increased [1]. Lithium-ion batteries have become the most popular and widespread form of energy storage devices due to their high stability and storage capacity [2–7]. Despite their great success for a wide variety of applications, especially for portable electronic devices, the large-scale installation of Li-ion batteries has been hindered due to the utilization of rare metals such as cobalt. Furthermore, low lithium diffusion in conventional transition metal based electrode materials has been noted as a limiting factor responsible for low power capacity. Therefore, in order to overcome those obstacles in current Li-ion battery technology, it is very desirable to explore a multitude of material candidates to identify promising alternative electrode materials which possess higher sustainability and enhanced performance.

Among a variety of candidate materials, organic materials have several benefits relative to the conventional materials for electrode applications. First, organic materials are composed of abundant and relatively inexpensive elements such as carbon, oxygen, sulfur, hydrogen and nitrogen [8]. Second, redox-active organic molecules such as quinones offer a higher capacity compared to conventional materials [9–11]. Third, organic electrode materials have higher

[*] Corresponding author.
*E-mail address:* SeungSoon.Jang@mse.gatech.edu (S.S. Jang).
[g] These authors contributed equally to this work.

https://doi.org/10.1016/j.mtener.2020.100482
2468-6069/© 2020 Elsevier Ltd. All rights reserved.

structural degrees of freedom than their transition metal based electrode counterparts; allowing for the fine tuning of their electrochemical properties. Furthermore, these organic materials can be combined with carbon nanomaterials such as graphene/graphite derivatives to provide even more opportunities in material design with higher performance [12–22]. However, due to the expensive and time-consuming nature of experimental research in the development of new Li-ion batteries with various organic materials, a more reliable and efficient high-throughput approach should be employed in order to accelerate the exploration of a large number of candidates.

As a promising candidate for such reliable and efficient throughput screening, first-principles computational methods have gained a great deal of attention due to ever increasing computational power and algorithmic breakthroughs. In our previous studies, we have developed a DFT-based protocol to predict the redox potential of organic electrode materials with high accuracy, in which the predicted redox potentials have uncertainties of around 0.2 V [13–21]. However, it should be noted that high efficacy DFT modeling still requires significant computational time and thus is not ideal for the vast screening of candidate materials.

Through recent progress in machine learning (ML), a new path has been paved for capturing and learning complicated. Relations among input data sets [23], demonstrating that ML can be used in materials science in order to predict complex behaviors of materials, and thereby to help explore a vast chemical space for new materials discovery [23–25]. For instance, once a machine learning model is trained, it can provide an immediate estimate of complex electronic and electrochemical properties of materials, which otherwise would have taken a significantly longer time to obtain experimentally or computationally from first-principles.

In this study, we develop machine learning (ML) models to i) accurately predict the redox potential of various organic materials with a high efficiency, and to ii) analyze how various molecular descriptors of interest affect the redox potential. Three different learning models are trained, namely artificial neural networks (ANN) [26], kernel ridge regression (KRR) [27], and gradient-boosting regression (GBR) [28,29], under three different strategies, which we call *pipelines*, to provide an advanced ML scheme for the accurate prediction of redox potential.

As summarized in Fig. 1, the molecules in our data set include various derivatives of functionalized graphene flakes, ketones, quinones, corannulenes, and coronenes. These molecules were the subjects of our past investigations in which we examined the effect of various structural variables such as the presence of functional groups, heteroatoms, and bound lithium atoms on the cathodic activity [13–21]. A full accounting of our data set can be found in our past works [13–21]. Furthermore, after training the learning models, the redox potential of 17 sumanene derivatives were predicted to confirm the predictive capability of the model. Furthermore, we assessed the relative contribution of several primary and composite molecular features to the redox potential prediction. Such insight into key relationships between the primary molecular characteristics and the redox potential can lead to more directed molecular structure design for tuned performance.

$$\begin{tikzcd}
\mathrm{R(gas)+e^-} \arrow[r, "\Delta G^{red}(R,gas)"] \arrow[d, "\Delta G^{solv}(R)"'] & \mathrm{R^-\,(gas)} \arrow[d, "\Delta G^{solv}(R^-)"] \\
\mathrm{R(sol)+e^-} \arrow[r, "\Delta G^{red}(R,sol)"] & \mathrm{R^-\,(sol)}
\end{tikzcd}$$

**Fig. 2.** Thermodynamic cycle to calculate the equilibrium redox potential in the solution phase.

## 2. Computational methods

### 2.1. Redox potential calculation

DFT is used to prepare the data set for training the ML models in this study. After organic molecules are geometrically optimized as shown in Fig. 1, basic electronic properties such as highest occupied molecular orbital (HOMO), lowest unoccupied molecular orbital (LUMO), and redox potential (RP), are calculated. The redox potentials of the organic molecules in the solution phase were calculated using the thermodynamic cycle suggested by Truhlar et al. as described in Fig. 2 [30,31]. Computational details of the DFT computations used to predict the redox potential are found in the Supporting Information [13–21].

### 2.2. Overview of machine learning models

Recently Artificial Neural Networks (ANN) have been extensively used to uncover the structure-property relationships in a wide range of materials [12,32,33]. ANNs are a subset of machine

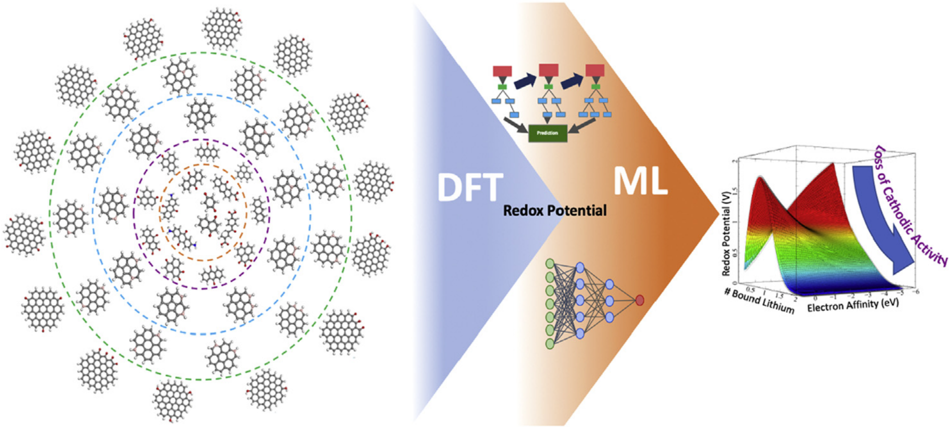


**Fig. 1.** Schematic of some of the organic molecules that have been investigated in this study: derivatives of Functionalized Graphene Flakes [21]; Ketone [20]; Quinone [17]; Corannulene [14]; Coronene [13]. The redox potential values from our prior works [13,14,17,20,21] are employed as a training set for the ML models.

learning algorithms that are inspired by the way the brain processes information via highly interconnected nodes (artificial neurons). Each node applies an activation function to a sum of weighted inputs from incoming connectors through which the result proceeds to the nodes in the following layer. The output of a network with *L* layers and *n* nodes in layer *L* can be expressed in the general form:

$$y_n^L = \left[ f\left( \sum_m w_{nm}^{L-1} \left[ \dots \left[ f\left( \sum_j w_{kj}^2 \left[ f\left( \sum_i w_{ji}^1 x_i + b_j^1 \right) \right] + b_k^2 \right) \right] \dots \right]_m + b_n^{L-1} \right) \right]_n \tag{1}$$

where $f(x)$ denotes the activation function associated with each node, $w_{nm}^{L-1}$ the weight on the connector between node *m* in layer *L-1* and node *n* in layer *L*, and $b_n^{L-1}$ the bias value at node *n*. The ANN is iteratively trained by adjusting the weight on each neuron based on the gradient of the error, which is commonly known as back-propagation.

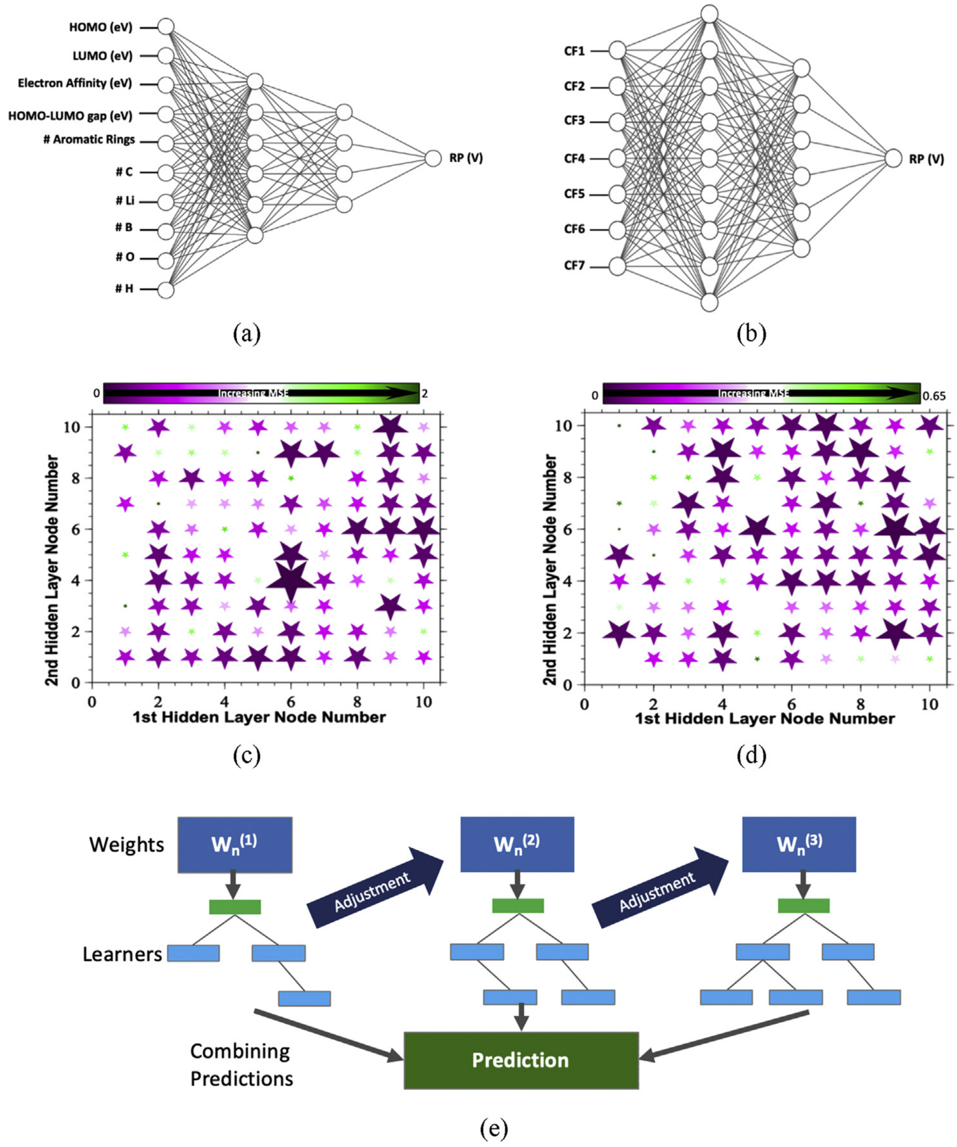


**Fig. 3.** ANN architecture optimized for (a) the original raw input descriptors and for (b) the composite input descriptors. The MSE results of the hidden layer grid search for the original raw inputs and the composite inputs are also shown in (c) and (d), in which the optimal architectures are determined to be "6–4" and "9–6" respectively. The size of the marker is inversely correlated to their error; which is also depicted by the color map where deep purple is for lowest MSE. (e) Sequential learning model in GBR is illustrated.

The optimal number of nodes (displayed in Fig. 3a and b) in each hidden layer can be determined by a grid search which takes into account every possible network architecture for a given number of hidden layers; this approach can be rather inefficient for large numbers of hidden layers, thus requiring different hyperparameter optimization schemes such as random search. The ANN models in this study were trained using the Quasi Newton Method [34] in the MATLAB Neural Network Toolkit [35].

Another highly promising learning model is Gradient Boosting Regression (GBR), which is an ensemble ML technique that minimizes a loss function across many weak learners through a process known as boosting [28,29]; the overall process is represented in Fig. 3e. We used a GBR model which fits up to 500 individual regression trees trained according to an optimized loss function. Each tree is added to the model sequentially, and the input weights are adjusted to minimize the error. Once a regression tree is fitted, the coefficient values are unchanged during the rest of the convergence process. A prediction is produced by an additive model where the predictions of all the individual weak learners are summed to reduce the error of each sequential tree.

The third machine learning model which we consider is Kernel Ridge Regression (KRR), which is an extension of the ridge regression model that learns a space created by applying the kernel method to the input data by minimizing a squared loss term [27].

$$\widehat{y} = y^T(K + \lambda I)^{-1}\kappa \tag{2}$$

where $y^T$, $K$ and $\kappa$ denote the transpose of the dependent variable, the train sample based kernel, and test sample based kernel matrix, respectively. The advantage of KRR comes from its relative simplicity and the addition of the kernel trick which allows KRR to fit even non-linearly correlated data. Initially, two kernels were considered, the linear kernel and the radial basis function (RBF) kernel [27]. The radial basis function (RBF) kernel is employed because of its capability to capture non-linear correlations in the data. Both KRR and GBR were implemented using the Scikit-Learn package.

## 2.3. Hyperparameter selection

Hyperparameters represent a set of variables in a machine learning algorithm that govern how the model behaves. They must be set before the model is trained, and unlike regular parameters, may not be optimized or altered during the training process. Their purpose is to allow a single algorithm to be effective on a wide range of input data sets. This creates a challenge of selecting the optimal set of hyperparameters that leads to the best performance on the input data as defined by the model creator. A common approach for hyperparameter selection involves exploring a predefined hyperparameter space and evaluating the results on a validation set. For instance, in order to determine the optimal architecture within the two hidden layers in ANN (represented in Fig. 3a and b), a comprehensive grid search is performed by creating 100 unique neural networks where the number of nodes varies from 1 to 10 for each hidden layer. Then, the optimal node configuration that produces the minimum mean squared error (MSE) value for the validation set is chosen as shown in Fig. 3c and d.

To train our ML models (ANN, GBR, and KRR), the initial data set of 108 organic molecules was first randomized and normalized, and then 80% of the data set was used for training ML models, while 20% was used for validation. Subsequently the trained ML models were used to predict the redox potentials of six organic molecules in a test set which have not been used in training. In this study, cross-validation was used during hyperparameter selection in order to evaluate the relative model performance given by each set of hyperparameters and during recursive feature elimination to select the optimal subset of input features. More specifically, the training data is split into five groups, or folds, where four folds are used to train the model and one is used to evaluate its performance. As such, to evaluate the performance of an algorithm on a set of possible parameters we train it five separate times on combinations of four training folds and one validation fold. The goal of this process is to reduce model bias by effectively evaluating a model against a more representative subset of our training data without exposing it to any of the hold out test data, which is completely left out of the model. Further details on hyperparameter selection for all three learning models can be found in the proceeding sections.

### 2.3.1. Artificial neural network

A grid search over two hidden layers yielded a “6–4” node configuration and a “9–6” node configuration for the original 10 primary input features and the post-LASSO composite features, respectively, as seen in Fig. 3c and d. In order to ensure that two hidden layers are optimal, the same test is performed for the case of having three hidden layers. When the number of hidden layers is changed to three, 1,000 unique neural networks are produced. It is found that the optimal neural network configurations for 3 hidden layers are “10-9-5” and “8-6-3” for the primary input features and the post-LASSO composite features, respectively. However, these configurations lead to a very slight improvement in the performance. Since the addition of hidden layers and neurons can potentially induce a greater risk of memorization, it is decided to use only two hidden layers for our ANN models in this study. Additionally, as a defense against overfitting, an early stopping criteria is implemented by terminating the training if the validation performance degrades for 10 consecutive epochs (Fig. S1a). Furthermore, the hyperbolic tangent sigmoid function is employed as the activation function with a constant learning rate of 0.01 and a data spread of 80% training and 20% validation.

### 2.3.2. Gradient boosting regression

It is generally known that GBR is one of the most difficult models to optimize due to the large number of hyperparameters involved. This complexity can be attributed to the boosting algorithm in GBR, meaning that the ensemble method and the individual learners have their own sets of parameters. The optimal hyperparameters are set to use a learning rate of 0.125, LAD loss, mass tree depth of 5, minimum samples per leaf of 2, minimum samples split of 0.5, and 500 estimators. Once an optimal model is selected, the feed-forward time is comparable to the other two algorithms.

The learning rate controls the rate at which the model converges. Thus a learning rate that is too high would lead to the model potentially converging on a non-optimal solution (local minimum), whereas a rate that is too low would not converge at all. In addition to hyperparameters defining the gradient boosting model, hyperparameters of the regression trees are also tuned. The learning rate was optimized through a grid search between 0.075 and 0.125, in which the learning rate was tuned by incrementally building and evaluating a model for parameter values within the specified range.

### 2.3.3. Kernel ridge regression

The two hyperparameters for the KRR model are the kernel and alpha value. Among the two considered kernels, namely linear and RBF, the RBF kernel proves to be more robust because of its reduced MSE from the validation sets. The alpha value is determined to be 0.0028 by performing a grid search of 100 samples on a range from 0 to 1 on a logarithmic scale.

### 2.4. Primary input features

DFT is used to calculate basic electronic properties including adiabatic electron affinity (EA), highest occupied molecular orbital (HOMO), lowest unoccupied molecular orbital (LUMO), and the HOMO-LUMO gap, which have been used as input features (or descriptors) to train the machine learning models [12]. Additional basic structural features are also included to account for the structural variation of the molecular compounds: number of carbon, boron, oxygen, lithium and hydrogen atoms, and the number of aromatic rings, present in the molecule [12].

### 2.5. Composite input features

Prior to creating any composite features, the primary input features must be normalized. During this normalization process, raw input data is adjusted to have a common scale, facilitating subsequent optimization processes to run efficiently even with inputs of dissimilar ranges. We perform this normalization in two steps: first, we standardize the data to have the standard normal distribution, and then we rescale the inputs to have a range from $-1$ to 1.

Furthermore, the composite feature development involves three major steps. The first step is to transform our input features using a set of common functions, with the aim of non-linearizing the significance of each input to allow for the selection of a subset of features that are highly correlated with the output property [36]. This is implemented by applying the following four functions $x^2$, $x^{1/2}$, log $(2+x)$, and $e^x$, where $x$ represents each of the primary input features. Given 10 primary input features, this procedure creates 40 transformed features. The second step is to create new features by combining primary and transformed features by systematically multiplying 2 and 3 input features. Given 50 starting features (10 primary and 40 transformed), the combination of 2 and 3 features generates 1,225 and 19,600 composite features, respectively. The second step is designed to mitigate the fact that each single primary feature may not have a clear direct correlation with the output; and thus the composite feature generation aims to help the learning models find the cooperative relationships among the features to capture complicated correlations with the target. Similar to how an RBF kernel in KRR can add an extra dimension to the original features, the composite feature approach aims to increase the dimensionality of the input features (with respect to the primary features) in order to perform non-linear regression.

In the third step, a smaller subset of the most valuable descriptors are selected out of the 20,875 features (10 original, 40 transformed, 20,825 composite) generated so far, which are used to train our machine learning models. This third step was accomplished by excluding features with zero regression coefficients as produced from training a least absolute shrinkage and selection operator (LASSO) model [37]. LASSO is based on the least-squares regression method, which is much less intensive in computation than the other machine learning methods used in this study [37]. By constraining the sum of model parameters, LASSO penalizes the coefficients of the regression variables, reducing some of these coefficients to zero. Using LASSO, the total number of features is reduced from 20,875 to 7. This result was produced by performing hyperparameter selection using LASSO and taking the model that has the lowest MSE on the validation set [37]. By assuming LASSO can converge on the optimal feature subset with given hyperparameters, the most important subset of input can be determined by choosing the most accurate LASSO model. These resulting features are then used to train the final model for each ML algorithm. Now that we have identified the tools that can be utilized for optimizing the feature space, we will describe how these tools are implemented in three different optimization strategies.

## 3. Results and discussion

The three ***pipelines*** used for each of the three models, GBR, KRR, and ANN, are summarized in Fig. 4. In ***Pipeline 1***, each ML model is directly trained using the primary features, whereas ***Pipeline 2*** employs a relative contribution analysis (RCA) and recursive feature elimination (RFE) to improve the performance of the models. In ***Pipeline 3***, the feature space is expanded from the primary 10 to 20,875 features, which is followed by a LASSO screening that results in 7 composite features are used in the training of the learning models.

Please note that Pearson correlation analysis is added to ***Pipelines 2 and 3*** to enhance the learning models by removing redundant and mutually-correlated features. Here, “mutually-correlated features” corresponds to input features which have strong correlations with other input features and thereby make little contribution to the training. Removing these mutually-correlated features can lessen the redundancy of the input feature space and enhance the efficiency of ML model. In the case of ***Pipelines 2 and 3***, the correlation filter does not eliminate any features, but it ensures that the features selected by LASSO, in the case of ***Pipelines 3,*** were not mutually-correlated.

The accuracy of the ML models within each ***pipeline*** in predicting redox potential is evaluated by MSE. Here, it is important to note that MSE is used in this context as a metric to evaluate the relative performances of models, not necessarily as an absolute measure of the performance. From Table 1, it is demonstrated that ***Pipeline 3*** outperforms both ***Pipelines 1 and 2***. As shown in Fig. 4, ***Pipeline 2*** represents the addition of feature selection to ***Pipeline 1***, while ***Pipeline 3*** represents the addition of composite feature generation and LASSO. Thus the main differences of performance among pipelines in Table 1 should be attributed to the feature selection or the composite feature protocol.

The recursive feature elimination (RFE) is performed by ranking the relative importance of each feature in predicting redox potential using a relative contribution analysis (RCA) and then cumulatively removing each feature in order of ascending importance. Lastly, only the collection of features minimizing the MSE are kept. This RFE procedure ensures that we utilize the minimum number of features necessary for the learning models.

Another noteworthy point in Table 1 is that all ML models are improved significantly with the addition of LASSO in ***Pipeline 3***. Please note that KRR demonstrates the best performance with an MSE of 0.025 in ***Pipeline 3*** (vs. 0.032 and 0.045 for ANN and GRB), whereas it shows the lowest performance out of the three ML models in ***Pipeline 1*** and ***2***. This highlights the differences in the capability of each ML model in predicting non-linear correlations between the inputs and outputs. Given the increased sophistication of ANN and GBR compared to KRR, they are able to extract the complicated correlation between the original input features and the output more effectively. On the contrary, KRR is converged to an optimal model with the aid of composite feature approach and LASSO selection under ***Pipeline 3***. It should be noted that ANN could outperform KRR if given a larger data set and a larger hyperparameter space including more hidden layers. However, ANN is also potentially more vulnerable to memorization if the number of hidden layer size is increased. Thus two hidden layers are deemed sufficient for this study.

In Fig. 5, it is revealed that EA and #Li are important features for all three ML models. KRR is the only model that does not recognize EA as the most significant feature. This is likely because the KRR model is not fully optimized in ***Pipeline 2*** compared to the other

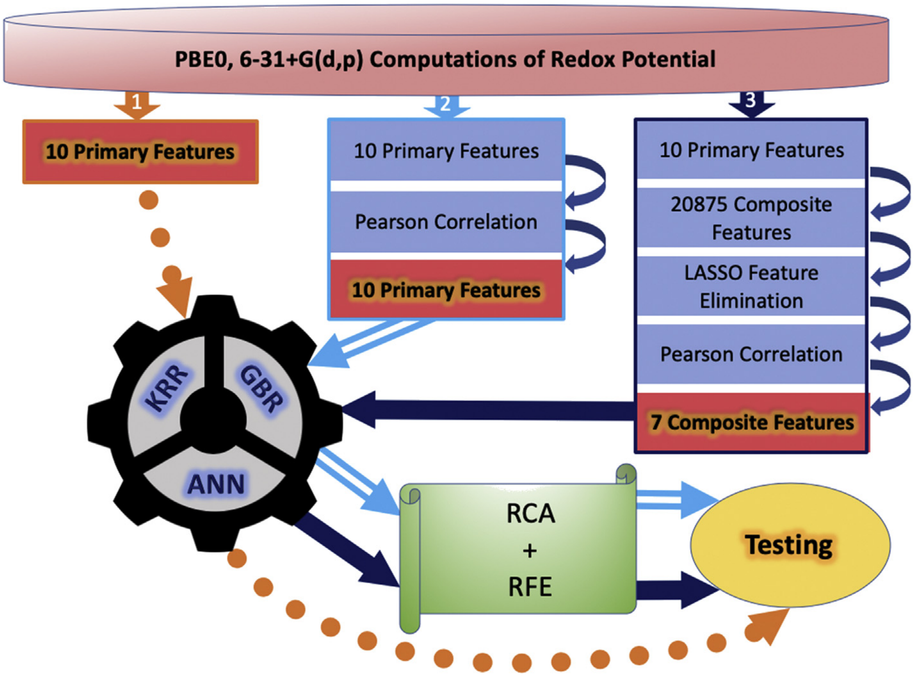


**Fig. 4.** Overall breakdown of the three ***pipelines*** for all three learning models. ***Pipeline 1*** represents the base protocol, in which the models were trained directly using the 10 primary features. ***Pipeline 2*** depicts the placement of a Pearson Correlation filter, in addition to a relative contribution analysis (RCA) and recursive feature elimination (RFE). Lastly, ***Pipeline 3*** depicts the addition of composite features and feature elimination using LASSO.

models, as seen in Table 1. Furthermore, it is important to note that a more reliable assessment of the contribution of lithium binding can be obtained by increasing the amount of cases in the testing set that consider more than one bound lithium. Nevertheless, the even higher importance associated with EA is expected due to the fact that EA is a measure of a materials capability to attain electrons, which is a direct indicator of a material's electrochemical tendency for reduction. Thus all models agree that EA has a relatively higher contribution than other features with respect to predicting the redox potential.

RFE for ANN, GBR, and KRR in ***Pipeline 2*** are shown in Fig. 6. The purpose of performing RFE is to find the minimal collection of input features needed to optimize the model performance. In each plot in Fig. 6, more features are eliminated moving along the horizontal axis; the most important feature is not included in the plot as it is always retained. The figures illustrate how MSE and the coefficient of determination are changed as a function of such feature elimination. The coefficient of determination ($R^2$) is a measure of the variance extent in a dependent variable that is predictable from the independent variable. For both KRR and ANN, no features are eliminated from ***Pipeline 2*** because the elimination of “None” achieves the lowest MSE as shown in Fig. 6. On the contrary, in the case of GBR, all the features are eliminated from RFE except for #Li and EA, so that MSE of GBR is reduced via ***Pipeline 2*** in comparison to ***Pipeline 1***.

**Table 1**
The overall performances (MSE) of all three machine learning models under the three Pipeline models. KRR under the third Pipeline results in the lowest MSE.

| Machine Learning Models | Pipeline 1 | Pipeline 2 | Pipeline 3 |
|---|---|---|---|
| ANN | 0.099 | 0.099 | 0.032 |
| GBR | 0.103 | 0.097 | 0.045 |
| KRR | 0.130 | 0.130 | 0.025 |

Moving on to ***Pipeline 3***, it is observed that all composite features down selected via LASSO contain EA as shown in Table 2. Furthermore, Fig. 7a shows that all ML models have the ***Composite Feature 1*** (***CF1***) as the feature with the highest rank, which validates the optimization of the ML models via ***Pipeline 3***. Please note that this ***CF1*** feature contains EA and #Li, confirming the results of RCA for the original features in ***Pipeline 2*** (Fig. 5). The inclusion of EA in the LASSO-selected composite features is expected since EA is an intrinsic characteristic of a material's tendency to attain an

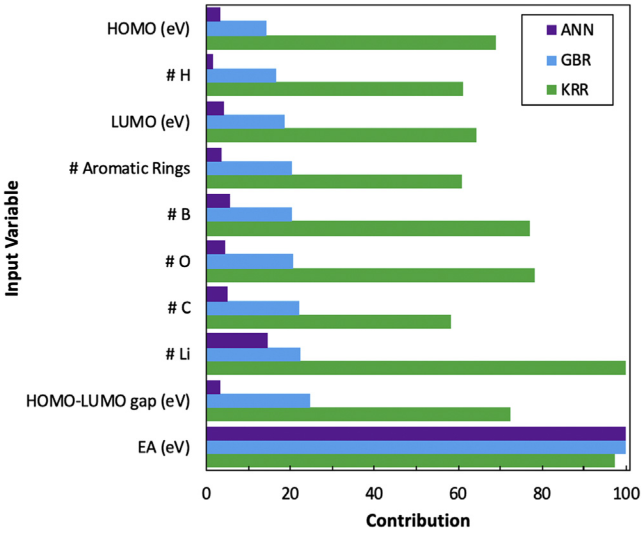


**Fig. 5.** The relative contribution of the original features is computed using ***Pipeline 2*** for ANN, GBR, and KRR. All the features within each model are given a score as a proportion relative to the highest ranking feature, which is given a contribution of 100. Noticeably, KRR is the only algorithm where the contribution of each feature is more equally distributed.

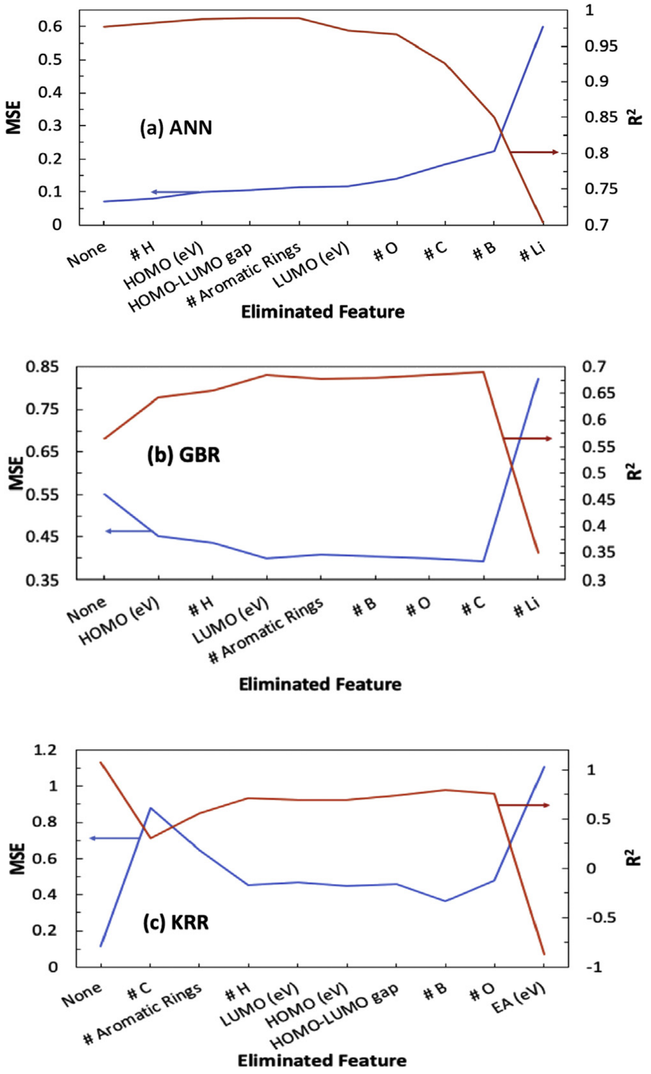


**Fig. 6.** The recursive feature elimination under ***Pipeline 2***, in which features are systematically removed in ascending order of their relative contribution to find the feature set with the minimal number of features and the minimal error: (a) ANN; (b) GBR; (c) KRR.

electron for electrochemical reduction. Furthermore, Fig. 7b demonstrates a small number of mutually-correlated composite features, indicating that highly correlated features were successfully eliminated by LASSO. Additionally, ANN (Fig. 7c) and KRR (Fig. 7e) show a reduction in error after the removal of CF7, whereas GBR (Fig. 7d) was optimized by the removal of CF5.

The varying accuracy of the learning models, mainly between KRR and the rest, can be primarily attributed to the differences in how each ML model handles data. For instance, while ANN and GBR are fundamentally non-linear ML models, the capability of KRR to perform non-linear regression is dependent on the choice of kernel [38]. For instance, KRR would be unable to model non-linear relationships entirely using a linear kernel. Although the RBF kernel implemented in this study is a non-linear kernel, modeling data which do not have a linear relationship, is not easily achieved using KRR. This is clearly presented in ***Pipeline* 1**, where GBR and ANN achieve lower error than KRR (0.103 eV and 0.099 eV vs. 0.130 eV) as they are able to more readily extract non-linear correlations between the primary input features and redox potential. Therefore, to improve the performance of KRR for capturing complicated correlations, we develop composite features as described previously. When trained on composite features screened with LASSO, KRR's error decreases lower than the other two learning models.

**Table 2**
Composite features generated from the original descriptors by implementing LASSO in ***Pipeline 3***.

| Feature Index | Composite Feature |
|---|---|
| CF1 | $\sqrt{EA}\cdot exp(EA)\cdot exp(\#Li)$ |
| CF2 | $\sqrt{EA}\cdot exp(EA)\cdot exp(LUMO)$ |
| CF3 | $\sqrt{EA}\cdot exp(EA)\cdot exp(\#B)$ |
| CF4 | $\sqrt{EA}\cdot exp(EA)\cdot exp(HOMO - LUMO\ gap)$ |
| CF5 | $EA\cdot exp(\#B)\cdot exp(\#Li)$ |
| CF6 | $\sqrt{EA}\cdot exp(EA)\cdot exp(\#H)$ |
| CF7 | $exp(EA)\cdot exp(\#H)\cdot exp(No.\ of\ Aromatic\ Rings)$ |

The correlation obtained from the training sets and test sets using the three ML models under ***Pipeline 3*** are presented in Fig. 8 (the correlations for Pipeline 1 and 2 can be found in Figs. S2 and S3, respectively). A common trend that is observed is the low accuracy in predicting the negative redox potential for the naphthoquinone with two lithium organic molecules (depicted by the red circle in Fig. 8b, d and 8f), whereas the positive redox potentials are predicted more accurately by all the models. This is most likely due to the presence of a small number of molecules with negative redox potential in the training set (only 5 data sets out of the 108). Furthermore, in the case of ANN and KRR in Fig. 8, the $R^2$ values for the test set are higher than that for the training set, which can indicate that these models likely do not overfit since the performance in the test set exceeds that in the training set. In contrast, the trend is the opposite for GBR. As illustrated in Fig. S1b, the MSE for the validation set is not significantly reduced after 20 epochs. Although this does not necessarily indicate overfitting, it suggests that the size of the training data is not sufficient for GBR to learn the problem.

After the model is trained using KRR through ***Pipeline 3***, the capability of KRR to predict the redox potential of organic molecules was tested against 17 sumanene derivatives [39]. Although both the test set, which was used to evaluate the three models and three pipelines, and the sumanene data are unknown to the ML models, the test set is a sample of the dataset which our group developed. On the other hand, the sumanene data is extracted from an external study. Additionally, please note that the sumanene set, unlike the test set, has no representative derivatives in the training set [39]. The sumanene derivatives are a class of $\pi$-conjugated fullerene fragments that consist of a central benzene ring surrounded by alternating cyclopentadiene and benzene rings [39]. It turns out that the model predicts the redox potential, shown in Fig. 9a, with an average error (discrepancy) and a Pearson correlation of 3.94% and ~97%, respectively, between the DFT and KRR predicted redox potentials. This result indicates that the KRR model trained via ***Pipeline 3*** can predict the redox potential of a wide range of organic materials.

As previously noted, the most important composite feature, **CF1** in ***Pipeline 3***, contains the raw features of EA and #Li. The dependence of redox potential on these two features is displayed as a 2-D contour map in Fig. 9b using KRR (Fig. S4 displays the color map using GBR). It is shown that the most positive redox potential (i.e. increased cathodic activity) is attained in the region with more negative EA and no bound lithium atoms. On the other hand, redox potential is predicted to become negative (shown by the blue regime) when more than one lithium atom is present. This is in agreement with our prior finding that quinone derivatives lose their cathodic property with respect to lithium

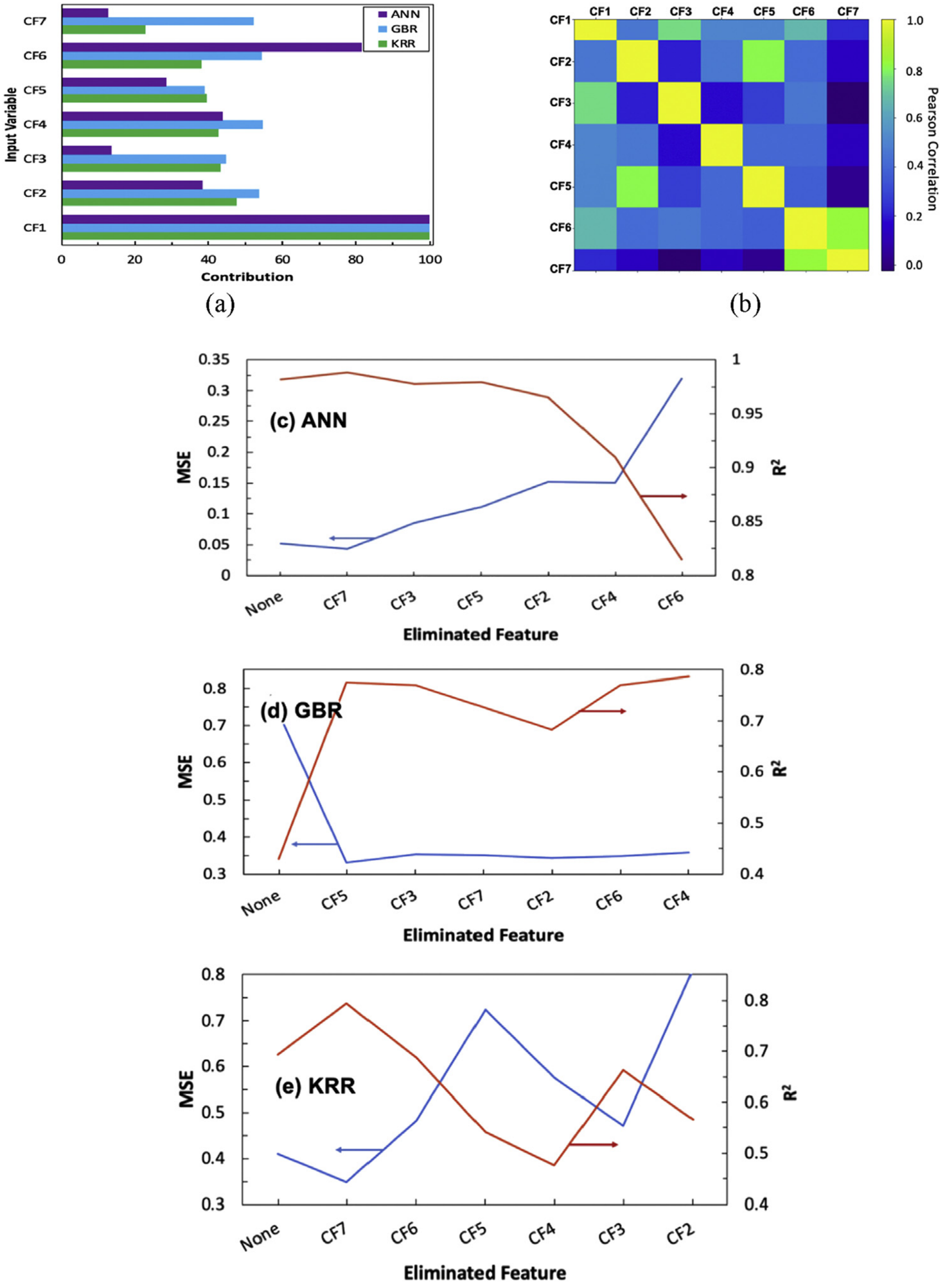


**Fig. 7.** (a) The relative contribution of each composite feature in ***Pipeline 3*** for ANN, GBR, and KRR. The contribution score for the features is relative to the most important feature, which has a score of 100. CF1 is the feature with the highest relative contribution for all three learning models; (b) the coefficient of determination is computed between every pair of composite features. The recursive feature elimination in ***Pipeline 3***: (c) ANN; (d) GBR; (e) KRR.

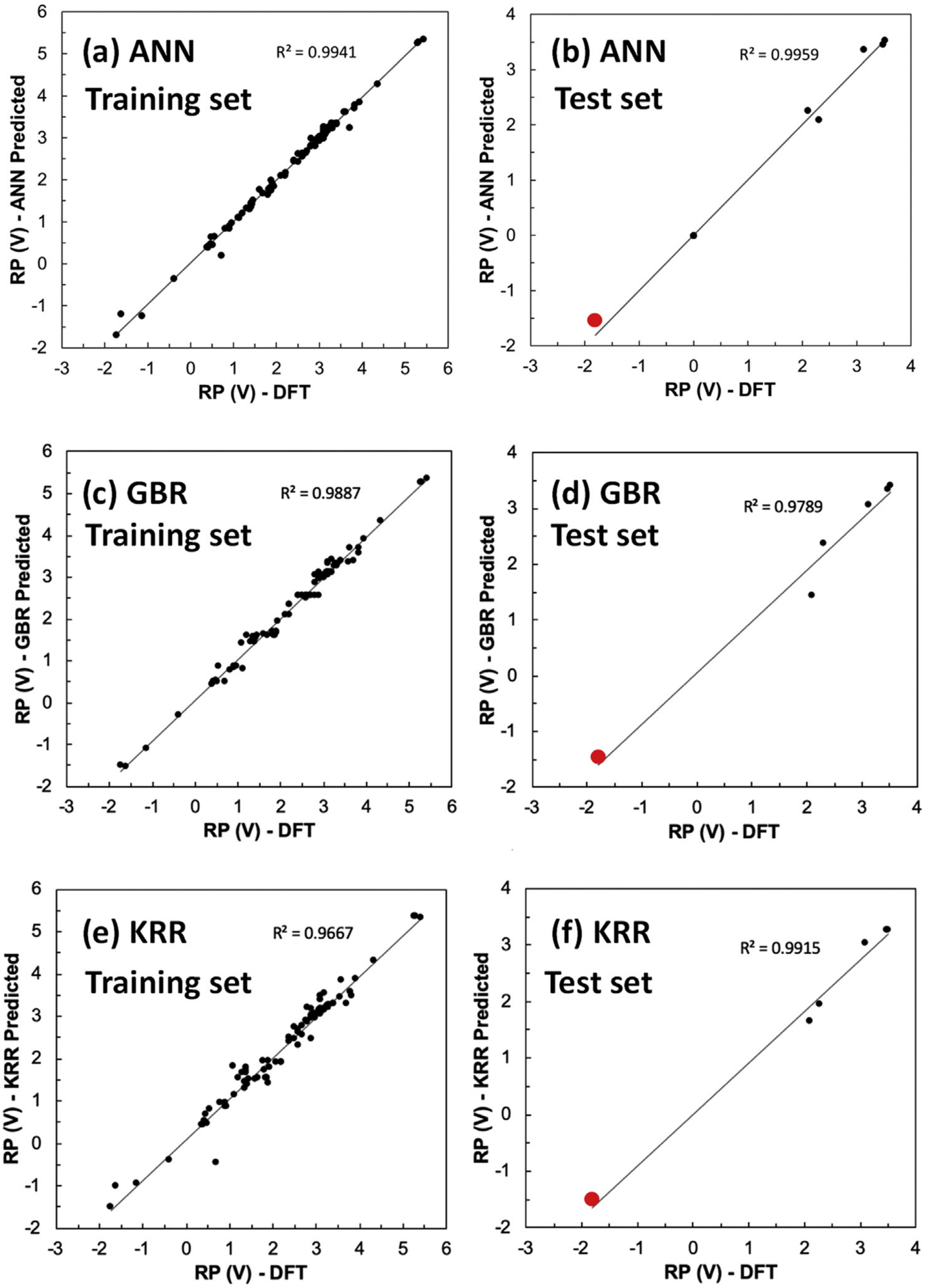
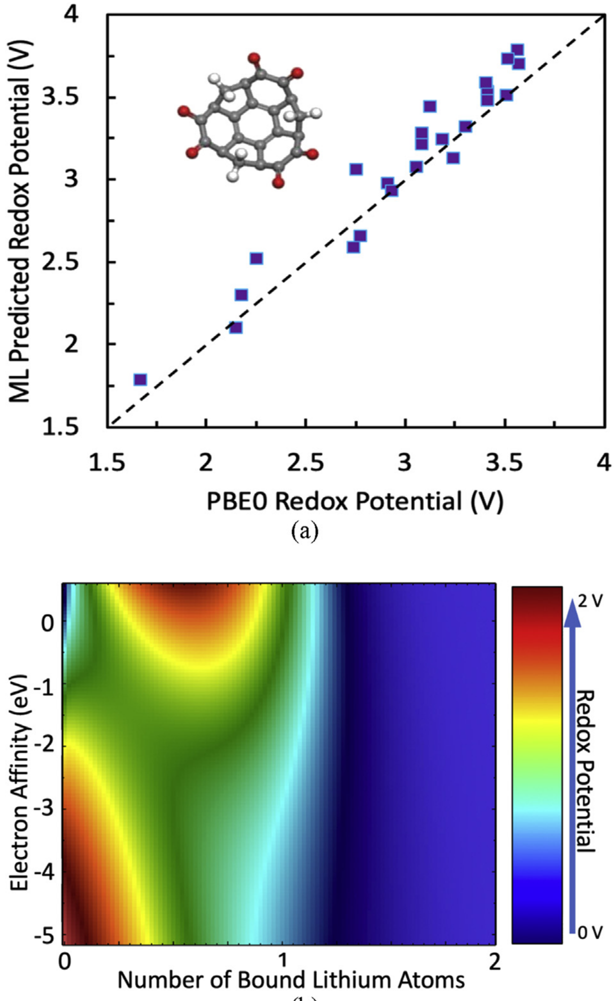


**Fig. 8.** Comparison between ML predictions and DFT calculations for redox potentials. Naphthoquinone with 2 Li ions is marked by the red circle.

electrode when two lithium atoms are bound [17]. It is important to note that this observation is with respect to monomeric units of organic molecules, such as quinones, which were examined in this study. In fact, the ranges of values of EA and #Li in Fig. 9b are limited to the ranges which correspond to the cases in the training set. An extrapolation of this finding beyond these limits is possible by expanding the model further to include polymerized units (i.e. with more than two carbonyl groups) which would be capable of accommodating more than 2 lithium atoms. Thus this learning model can be used to perform an expeditious high throughput screening of the redox potentials of a wide range of candidate organic monomers.

## 4. Conclusion

We have examined multiple pathways for establishing a high-fidelity DFT-machine learning framework with the dual goals of 1) analyzing how certain molecular characteristics can be modified to tune the redox potential of organic electrode materials and 2) identifying a protocol that can be readily used for datasets with a limited size. Using basic electronic and molecular features, three learning models, namely artificial neural network, gradient boosting regression, and kernel ridge regression, were trained through three different ***pipelines*** implementing three different levels of sophistication. It was found that KRR, along with a series of



(a)

(b)

**Fig. 9.** (a) Prediction performance of the KRR optimized via ***Pipeline 3*** for 17 newly introduced sumanene derivatives obtained from an external source [39]. (b) 2-D color map depicting the variation of the redox potential as a function of the two most significant features, as predicted by KRR via ***Pipeline 3***. The blue region indicates cathodic deactivation. It should be noted that non-integer quantities of bound lithium signify an uneven number of lithium atoms shared among organic moieties.

processes including composite feature generation, LASSO feature selection, relative contribution analysis, and recursive feature elimination, yielded the highest accuracy prediction of redox potential out of all the models and protocols considered. Furthermore, the optimized model delivered a remarkable performance in predicting the redox potential of a class of organic molecules that was not included in the input data set, suggesting the model's possible utility in the prediction of redox potential for candidate organic molecules beyond the subset which was used in its training. As such, it is concluded that the approach presented in this study is generic in its application for electrochemical properties of organic materials, which makes high-throughput preliminary screening available and useful for a wide range of materials.

## Credit author statement

**Omar Allam**: Invesitgation, Methodology, Formal analysis, Validation, Writing - original draft. **Robert Kuramshin**: Invesitgation, Methodology, Formal analysis, Validation, Writing - original draft. **Zlatomir Stoichev**: Invesitgation, Methodology, Formal analysis, Validation, Writing - original draft. **Byung Woo Cho**: Invesitgation, Formal analysis, Validation. **Seung Woo Lee**: Invesitgation, Formal analysis, Validation. **Seung Soon Jang**: Conceptualization, Methodology, Formal analysis, Validation, Resources, Writing - review & editing, Supervision, Project administration

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgement

This material is based upon the work supported by the National Science Foundation under Grant no. 1805052.

## Appendix ASupplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.mtener.2020.100482.

## References

[1] P. Poizot, F. Dolhem, Clean energy new deal for a sustainable world: from non-CO2 generating energy sources to greener electrochemical storage devices, Energy Environ. Sci. 4 (2011) 2003–2019, https://doi.org/10.1039/C0EE00731E.

[2] L.J. Aaldering, C.H. Song, Tracing the technological development trajectory in post-lithium-ion battery technologies: a patent-based approach, J. Clean. Prod. 241 (2019) 118343, https://doi.org/10.1016/j.jclepro.2019.118343.

[3] M.D. Bhatt, C. O'Dwyer, Recent progress in theoretical and computational investigations of Li-ion battery materials and electrolytes, Phys. Chem. Chem. Phys. 17 (2015) 4799–4844, https://doi.org/10.1039/c4cp05552g.

[4] P.G. Bruce, B. Scrosati, J.M. Tarascon, Nanomaterials for rechargeable lithium batteries, Angew. Chem. 47 (2008) 2930–2946, https://doi.org/10.1002/anie.200702505.

[5] H. Chen, M. Armand, G. Demailly, F. Dolhem, P. Poizot, J.M. Tarascon, From biomass to a renewable Li(x)C(6)O(6) organic electrode for sustainable Li-ion batteries, ChemSusChem 1 (2008) 348–355, https://doi.org/10.1002/cssc.200700161.

[6] H.S. Chen, T.N. Cong, W. Yang, C.Q. Tan, Y.L. Li, Y.L. Ding, Progress in electrical energy storage system: a critical review, Prog. Nat. Sci. 19 (2009) 291–312, https://doi.org/10.1016/j.pnsc.2008.07.014.

[7] N. Nitta, F.X. Wu, J.T. Lee, G. Yushin, Li-ion battery materials: present and future, Mater. Today 18 (2015) 252–264, https://doi.org/10.1016/j.mattod.2014.10.040.

[8] Z.P. Song, H.S. Zhou, Towards sustainable and versatile energy storage devices: an overview of organic electrode materials, Energy Environ. Sci. 6 (2013) 2280–2301, https://doi.org/10.1039/C3EE40709H.

[9] Y.L. Liang, Z.L. Tao, J. Chen, Organic electrode materials for rechargeable lithium batteries, Adv. Energy Mater. 2 (2012) 742–769, https://doi.org/10.1002/aenm.201100795.

[10] Z.P. Song, H. Zhan, Y.H. Zhou, Anthraquinone based polymer as high performance cathode material for rechargeable lithium batteries, Chem. Commun. (2009) 448–450, https://doi.org/10.1039/B814515F.

[11] H. Mitome, T. Ishizuka, Y. Shiota, K. Yoshizawa, T. Kojima, Controlling the redox properties of a pyrroloquinolinequinone (PQQ) derivative in a ruthenium(II) coordination sphere, Dalton Trans. 44 (2015) 3151–3158, https://doi.org/10.1039/C4DT03358B.

[12] O. Allam, B.W. Cho, K.C. Kim, S.S. Jang, Application of DFT-based machine learning for developing molecular electrode materials in Li-ion batteries, RSC Adv. 8 (2018) 39414–39420, https://doi.org/10.1039/c8ra07112h.

[13] Y.T. Zhu, K.C. Kim, S.S. Jang, Boron-doped coronenes with high redox potential for organic positive electrodes in lithium-ion batteries: a first-principles density functional theory modeling study, J. Mater. Chem. A. 6 (2018) 10111–10120, https://doi.org/10.1039/c8ta01671b.

[14] J. Kang, K.C. Kim, S.S. Jang, Density functional theory modeling-assisted investigation of thermodynamics and redox properties of boron-doped corannulenes for cathodes in lithium-ion batteries, J. Phys. Chem. C 122 (2018) 10675–10681, https://doi.org/10.1021/acs.jpcc.8b00827.

[15] P. Sood, K.C. Kim, S.S. Jang, Electrochemical and electronic properties of nitrogen doped fullerene and its derivatives for lithium-ion battery applications, J. Energy Chem. 27 (2018) 528–534, https://doi.org/10.1016/j.jechem.2017.11.009.

[16] P. Sood, K.C. Kim, S.S. Jang, Electrochemical properties of boron-doped fullerene derivatives for lithium-ion battery applications, ChemPhysChem 19 (2018) 753–758, https://doi.org/10.1002/cphc.201701171.

[17] K.C. Kim, T.Y. Liu, S.W. Lee, S.S. Jang, First-principles density functional theory modeling of Li binding: thermodynamics and redox properties of quinone derivatives for lithium-ion batteries, J. Am. Chem. Soc. 138 (2016) 2374–2382, https://doi.org/10.1021/jacs.51313279.

[18] T.Y. Liu, K.C. Kim, R. Kavian, S.S. Jang, S.W. Lee, High-density lithium-ion energy storage utilizing the surface redox reactions in folded graphene films, Chem. Mater. 27 (2015) 3291–3298, https://doi.org/10.1021/acs.chemmater.5b00314.

[19] T. Liu, K.C. Kim, B. Lee, Z.M. Chen, S. Noda, S.S. Jang, S.W. Lee, Self-polymerized dopamine as an organic cathode for Li- and Na-ion batteries, Energy Environ. Sci. 10 (2017) 205–215, https://doi.org/10.1039/c6ee02641a.

[20] J.H. Park, T.Y. Liu, K.C. Kim, S.W. Lee, S.S. Jang, Systematic molecular design of ketone derivatives of aromatic molecules for lithium-ion batteries: first-principles DFT modeling, ChemSusChem 10 (2017) 1584–1591, https://doi.org/10.1002/cssc.201601730.

[21] S. Kim, K.C. Kim, S.W. Lee, S.S. Jang, Thermodynamic and redox properties of graphene oxides for lithium-ion battery applications: a first principles density functional theory modeling approach, Phys. Chem. Chem. Phys. 18 (2016) 20600–20606, https://doi.org/10.1039/c6cp02692c.

[22] S. Er, C. Suh, M.P. Marshak, A. Aspuru-Guzik, Computational design of molecules for an all-quinone redox flow battery, Chem. Sci. 6 (2015) 885–893, https://doi.org/10.1039/c4sc03030c.

[23] T. Mueller, A.G. Kusne, R. Ramprasad, Machine learning in materials science: recent progress and emerging applications, Rev. Comput. Chem. 29 (29) (2016) 186–273, https://doi.org/10.1002/9781119148739.ch4.

[24] C. Kim, G. Pilania, R. Ramprasad, Machine learning assisted predictions of intrinsic dielectric breakdown strength of ABX(3) perovskites, J. Phys. Chem. C 120 (2016) 14575–14580, https://doi.org/10.1021/acs.jpcc.6b05068.

[25] O. Allam, C. Holmes, Z. Greenberg, K.C. Kim, S.S. Jang, Density functional theory - machine learning approach to analyze the bandgap of elemental halide perovskites and ruddlesden-popper phases, ChemPhysChem 19 (2018) 2559–2565, https://doi.org/10.1002/cphc.201800382.

[26] W.S. McCulloch, W. Pitts, A logical calculus OF the ideas immanent IN nervous activity (reprinted from bulletin OF mathematical BIOPHYSICS, vol 5, pg 115-133, 1943), Bull. Math. Biol. 52 (1990) 99–115, https://doi.org/10.1016/s0092-8240(05)80006-0.

[27] K.R. Muller, S. Mika, G. Ratsch, K. Tsuda, B. Scholkopf, An introduction to kernel-based learning algorithms, IEEE Trans. Neural Network. 12 (2001) 181–201, https://doi.org/10.1109/72.914517.

[28] J.H. Friedman, Stochastic gradient boosting, Comput. Stat. Data Anal. 38 (2002) 367–378, https://doi.org/10.1016/s0167-9473(01)00065-2.

[29] K.A. Gould, The elements of statistical learning (2nd edition): data mining, inference, and prediction, Dimens. Crit. Care Nurs. 35 (2016), https://doi.org/10.1007/978-0-387-84858-7, 52-52.

[30] P. Winget, C.J. Cramer, D.G. Truhlar, Computation of equilibrium oxidation and reduction potentials for reversible and dissociative electron-transfer reactions in solution, Theor. Chem. Acc. 112 (2004) 217–227, https://doi.org/10.1007/s00214-004-0577-0.

[31] A. Lewis, J.A. Bumpus, D.G. Truhlar, C.J. Cramer, Molecular Modeling of environmentally important processes: reduction potentials, J. Chem. Educ. 81 (2004) 596–604, https://doi.org/10.1021/ed081p596.

[32] J. Behler, M. Parrinello, Generalized neural-network representation of high-dimensional potential-energy surfaces, Phys. Rev. Lett. 98 (2007) 146401, https://doi.org/10.1103/PhysRevLett.98.146401.

[33] H.A. Chen, C.W. Pao, Fast and accurate artificial neural network potential model for MAPbI(3) perovskite materials, ACS Omega 4 (2019) 10950–10959, https://doi.org/10.1021/acsomega.9b00378.

[34] P. Hennig, M. Kiefel, Quasi-Newton methods: a new direction, J. Mach. Learn. Res. 14 (2013) 843–865.

[35] H.B. Demuth, M, User's Guide for Neural Network Toolbox for Use with MATLAB4, 2004.

[36] L.M. Ghiringhelli, J. Vybiral, S.V. Levchenko, C. Draxl, M. Scheffler, Big data of materials science: critical role of the descriptor, Phys. Rev. Lett. 114 (2015) 105503, https://doi.org/10.1103/PhysRevLett.114.105503.

[37] M. Dash, H. Liu, Feature selection for classification, Intell. Data Anal. 1 (1997) 131–156, https://doi.org/10.3233/IDA-1997-1302.

[38] P. Exterkate, Model selection in kernel ridge regression, Comput. Stat. Data Anal. 68 (2013) 1–16, https://doi.org/10.1016/j.csda.2013.06.006.

[39] K.H. Jung, K.C. Kim, Insights on redox properties of sumanene derivatives for high-performance organic cathodes, ACS Appl. Mater. Interfaces 12 (2020) 8333–8341, https://doi.org/10.1021/acsami.9b21991.