Machine Learning Prediction of the Redox Activity of Quinones

Ilia Kichev , Lyuben Borislavov * , Alia Tadjer * , Radostina Stoyanova

Posted Date: 3 October 2023

doi: 10.20944/preprints202310.0103.v1

Keywords: quinones; machine learning; ridge regression; decision tree; ensemble methods; density functional theory; organic electrode materials

Article

Machine Learning Prediction of the Redox Activity of Quinones

Ilia Kichev [1,2], Lyuben Borislavov [1]*, Alia Tadjer [2]* and Radostina Stoyanova [1]

[1] Institute of General and Inorganic Chemistry, Bulgarian Academy of Sciences, Sofia 1113, Bulgaria
[2] Faculty of Chemistry and Pharmacy, University of Sofia, 1164 Sofia, Bulgaria
* Correspondence: lborislavov@svr.igic.bas.bg, tadjer@chem.uni-sofia.bg

Abstract: The redox properties of quinones underlie their unique characteristics as organic battery components that outperform the conventional inorganic ones. Furthermore, the redox properties could be precisely shaped by using different substituent groups. Machine learning and statistics, on the other hand, have proven to be very powerful approaches for efficient in silico design of novel materials. Herein, we demonstrated the machine learning approach for the prediction of the redox activity of quinones that potentially can serve as organic battery components. For the needs of the present study, a database of small quinone-derived molecules was created. A large number of quantum chemical and chemometrics descriptors was generated for each molecule and subsequently different statistical approaches were applied to select the descriptors that most prominently characterize the relationship between structure and redox-potential. Various machine-learning methods for screening of prospective organic battery electrode materials were deployed to select the most trustworthy strategy for machine learning aided design of organic redox materials. It was found that ridge regression models perform better than regression decision tree and decision tree based ensemble algorithms.

Keywords: quinones; machine learning; ridge regression; decision tree; ensemble methods; density functional theory; organic electrode materials

1. Introduction

In recent years the global demand for effective energy storage materials constantly grows [1]. Traditionally, the widely used electrode materials in metal-ion batteries are inorganic compounds capable of reversible redox transformations [2, 3]. Organic electrode materials, on the other hand, have some gainful properties such as structural diversity and flexibility, synthetic tunability, lower price, and harmless recyclability [4-7]. Among the organic compounds considered for battery electrode materials research, quinones have engendered the most ubiquitous expectations and extensive investigation. Quinones are a class of organic compounds derived from aromatic dioles whose redox capacity makes them interesting for designing novel organic electrode materials [8]. Quinones with low molecular weight, such as 1,4-benzoquinone, have a relatively high redox potential [9] and, in case the two-electron redox reaction of benzoquinone takes place, a high capacity could be expected. However, due to sublimation and dissolution of benzoquinone in the organic electrolyte solvents, poor capacity is observed in practice [10]. These problems can be overcome by immobilizing benzoquinone on nanoparticles [11], by using various polymers containing benzoquinone fragments [12-14] or by introducing different substituent groups [15]. The redox potential of the quinones is dependent on the substituent type: electron-withdrawing substituents, such as halogen, carbonyl, nitro, and carboxylate groups, make quinones stronger oxidants, while electron donating groups, such as amine, hydroxyl and alkoxy groups, turn quinones into weaker oxidants [16]. In the present study, a dataset of quinones with electron-withdrawing substituents was constructed since this class of materials exhibits a fairly high redox potential.

Machine learning and statistics approaches have successfully been applied for capturing complex relationships between materials structure and different properties of interest [17]. This kind of approaches have also effectively been employed in the design of novel energy storage materials: Joshi et al [18] demonstrated that deep neural networks (DNN), support vector regression (SVR) and

 doi:10.20944/preprints202310.0103.v1

kernel ridge regression (KRR) can be used to predict redox potential of inorganic electrode materials extracted from the Materials Project Database; Zhang et al. [19] have used a Crystal Graph Convolutional Neural Network (CGCNN) to creatively build a interpretable deep learning model that predicts redox potential based on inorganic material crystal structures [19]. Machine learning algorithms have also been productively applied in the design of organic electrode materials: Allam et al. [20] have developed a prescreening procedure that relies on density functional theory to compute both redox potential of organic electrode materials and molecular descriptors such as HOMO-LUMO gap and electron affinity to be used as input features of artificial neural networks (ANNs), gradient boosting regression (GBR) and KRR. A major disadvantage is that the density functional theory, which is comparatively computationally expensive, is used for descriptors computation. Tutte et al. [21] propose a Hammet-like approach to model quinone solubility in organic electrolytes that are typically used in lithium-ion batteries (the organic electrode materials must have low solubility in the battery electrolyte). Machine learning screening has also been applied for the design of quinone electrolytes for redox flow batteries [22]: Wang et al. have created a dataset by generating various disubstituted quinones by replacing hydrogens in different quinone backbones with a predefined set of substituents and subsequently have utilized extreme gradient boosting algorithm to build a model for screening HOMO-LUMO gap and free energy of solvation. In the current study, different linear and nonlinear regression models were built to predict electrode potential of substituted quinones.

Dataset construction plays a central role in any data-driven study. In this report, two tactics for creation of application-specific datasets were combined. Firstly, a top-down approach was used: molecular structures that satisfy some application-specific conditions (i.e., contain a quinone fragment) were extracted from PubChem [23] (a large publicly available database). Next, a bottom-up approach was applied: the dataset of molecular structures produced in the first step was expanded with systematically generated derivatives of the already selected species. This strategy guarantees that the final dataset created is structurally consistent.

2. Materials and Methods

2.1. Dataset Construction

2.1.1. Molecular Structures Generation

To construct the dataset, 100 benzoquinone derivatives were extracted from the PubChem database [23] as SMILES strings. The SMILES strings were converted into 3D structures using the OpenBabel software package [24] and subsequently the DerGen software [25] was used to generate all possible derivatives of those compounds containing a -CN or a -C≡CMe group. 494 structures were produced in total. This dataset construction procedure guarantees that molecules generated are structurally similar and hence makes it easier to establish the structure-electrode potential relationship for a quinone series with electron withdrawing substituents – a group of compounds that is particularly interesting for organic energy storage materials.

2.1.2. Dataset Splitting

The dataset was shuffled and split into a training set (395 compounds, 80 % of the whole dataset) and a test set (99 compounds, 20 % of the whole dataset). To avoid data leakage [26], the descriptor selection and hyperparameter optimization were performed on the training set. Average R^2 metric over 5-fold cross-validation was used for model performance assessment during the descriptor selection and hyperparameter optimization.

2.2. Molecular Descriptors

Representing molecular structures in an unambiguous machine-readable format is not a trivial task. Many different molecular representations have been developed [27]. Molecular structures can be represented as:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 October 2023 doi:10.20944/preprints202310.0103.v1

- strings – for example the SMILSES representation that contains information about atom types and connectivity [28];
- connection table formats [29]: tabular formats that provide information about atom counts, atom types, connectivity matrix, bonded pairs of atoms, chirality, etc.; an example for such molecular representation format is the MDL molfile;
- vectors of features: a molecule can be represented either as a vector of molecular properties (descriptors) such as molecular weight, molecular volume, numbers of certain atom types, topology, etc. [30] or as a molecular fingerprint: a bitstring (can be regarded as vector of ones and zeros) is derived from the molecular structure according to a predefined set of rules [31] – one of the most employed fingerprints are the Extended-Connectivity Fingerprints (ECFP) based on the Morgan's algorithm [32], since they are specially designed for establishing structure-property relationships [33];
- computer learned representations: in recent years a large number of machine learning based molecular representation were developed – those methods rely on convolutional neural networks (CNN) and/or recurrent neural networks (RNN) to transform a molecule represented as a SMILES string or by 3D cartesian atom coordinates to a low dimensional latent space [34, 35] that can be used both for property prediction and for generation of new molecular structures [36].

In the current study the PaDEL [37] software package was employed to generate a multitude (750 descriptors per molecule) of cheminformatics-based molecular descriptors and the MOPAC [38] software package was used to produce semi-empirical descriptors such as HOMO and LUMO energies and dipole moments of the reducible compounds.

2.2.1. Descriptor Selection

Feature selection is a key step in any data-driven study [39]. The objective of feature selection is to sort out features that have strong correlation with the target variable. In the current work the following steps were taken:

- Low-variance descriptors were removed: descriptors whose value equals descriptor mode for 60% or more of the molecules in the dataset were discarded
- Descriptors that have weak correlation with the target value were discarded. Correlation with covariance between normalized descriptors and normalized target values of less than 0.25 is considered a weak correlation. The normalization was performed as follows:

$$V_{norm} = \frac{V - V_{mean}}{\sigma_V},$$

- where Vnorm is the normalized value, V is the unnormalized value, Vmean is the mean of V in the dataset and σV is the standard deviation of V in the dataset.
- Strongly correlated descriptors (covariance between normalized descriptors greater than 0.7) were discarded. After this operation, 52 descriptors were left.
- Backward stepwise regression [40] was used for further descriptor reduction (Figure 1). Finally, 32 descriptors remained (Table 1)

Table 1. Molecular descriptor names and descriptions.

Description	Descriptor name
Lowest partial charge weighted BCUTS [41]	BCUTc-1l
Highest partial charge weighted BCUTS [41]	BCUTc-1h
Total number of double bonds (excluding aromatic bonds)	nBondsD2
Triply bound carbon bound to another carbon	C1SP1
Doubly bound carbon bound to three other carbons	C3SP2
A topological descriptor combining distance and adjacency information [42]	ECCEN
Count of atom-type H E-State: H on aaCH, dCH2 or dsCH* [43]	nHother
Count of atom-type E-State: =C< [43]	ndssC

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 October 2023 doi:10.20944/preprints202310.0103.v1

Count of atom-type E-State: aaC- [43]	naasC
Count of atom-type E-State: N≡ [43]	ntN
Sum of E-States for weak hydrogen bond acceptors [43]	SwHBa
Sum of atom-type H E-State: =CH- [43]	SHdsCH
Sum of atom-type H E-State: H on aaCH, dCH2 or dsCH [43]	SHother
Sum of atom-type E-State: =C< [43]	SdssC
Sum of atom-type E-State: aaC- [43]	SaasC
Sum of atom-type E-State: N≡ [43]	StN
Minimum atom-type H E-State: H on aaCH, dCH2 or dsCH [43]	minHother
Minimum atom-type E-State: aaC- [43]	minaasC
Minimum atom-type E-State: =O [43]	mindO
Maximum atom-type H E-State: H on aaCH, dCH2 or dsCH [43]	maxHother
Maximum atom-type E-State: aaC- [43]	maxaasC
Mean intrinsic state values I [43]	meanI
Maximum negative intrinsic state difference in the molecule (related to the nucleophilicity of the molecule) [44]	MAXDN2
Maximum positive intrinsic state difference in the molecule (related to the electrophilicity of the molecule) [44]	MAXDP2
Complexity of the system [45]	fragC
Number of rings	nRing
Topological diameter (maximum atom eccentricity)	topoDiameter
Mean topological charge index of order 2 [46]	JGI2
Topological polar surface area	TopoPSA
van der Waals volume calculated using the method proposed in Zhao et al. *JACS* **2003**, *68*, 7368-7373] [47]	VABC
Molecular weight	MW
Energy of the lowest unoccupied molecular orbital estimated by PM6 [eV]	LUMO

*a=aromatic; s=single; d=double.

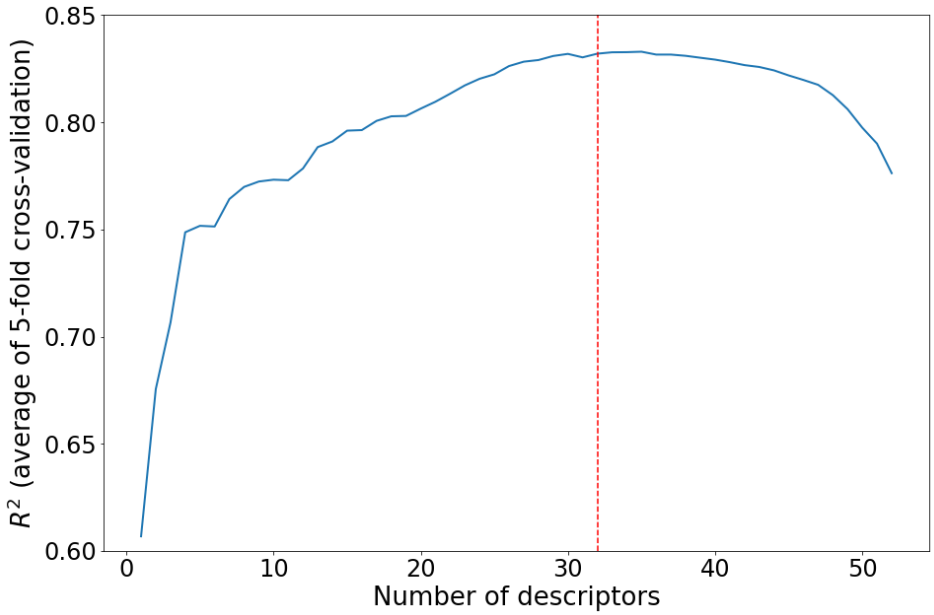
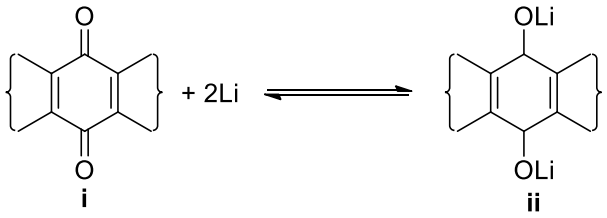


Figure 1. Results of the backward stepwise regression for descriptor selection.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 October 2023 doi:10.20944/preprints202310.0103.v1

2.3. Redox Potential Calculation

The redox potential was calculated with the Densify Functional theory (DFT):



Geometry optimization was performed on all quinone derivatives in the dataset (**i)** and their respective reduced forms (**ii)** using the B3LYP functional in combination with the 6-311++G(2df,2p) basis set as implemented in the Gaussian 16 software package [48]. The electrode potential was calculated by the Nernst equation:

$$\Delta E = \frac{-\Delta G}{nF}, \quad (1)$$

where n is the number of exchanged electrons, F is the Faraday's constant and the reaction free energy ΔG is calculated as follows:

$$\Delta G = G_{ii} - G_i - 2G_{Li}. \quad (2)$$

G_{ii} and G_i were obtained from the B3LYP/6-311++G(2df,2p) calculation as follows [49]:

$$H_X = E_0 + ZPE + H_{trans} + H_{rot} + H_{vib} + RT \quad (3a)$$

$$S_X = S_{trans} + S_{rot} + S_{vib} + S_{el} \quad (4b)$$

$$G_X = H_X - TS_X \ , \quad (5c)$$

where E_0 is the total electronic energy, ZPE is the unscaled zero-point energy, H_{trans}, H_{rot}, and H_{vib} are, correspondingly, the translational, rotational, and vibrational shares in the enthalpy, S_{trans}, S_{rot}, S_{vib} and S_{el} are, respectively, the rotational, translational, vibrational and electronic motion contributions to the entropy. RT represents the work term converting the internal energy into enthalpy (T = 298 K). G_{Li} is the free energy of lithium in the gas phase.

2.4. Machine Learning Methods Used

Different machine learning methods were deployed to investigate the relationship between molecular structure and electrode potential.

2.4.1. Ridge Regression

Ridge regression is a method for estimation the coefficients of l2-regularized multiple linear regression models:

$$\boldsymbol{X\beta} = \boldsymbol{y}, \quad (6)$$

where for a dataset consisting of n molecules, each represented as a m-dimensional vector, $\boldsymbol{X}$ is a $n \times (m+1)$ matrix of n-dimensional column vectors $\boldsymbol{x}_j$ ($\boldsymbol{x}_1$ is $[1\ 1\ \dots\ 1]^T$, while $x_2, x_3, \dots, x_{(m+1)}$ are the values of the corresponding descriptors), known as explanatory variables, $\boldsymbol{\beta}$ is a $(m+1)$ dimensional vector of parameters, where β_1 is the intercept term and $\boldsymbol{y}$ is the vector of observed values (redox potentials in the current study). The ridge estimator of $\boldsymbol{\beta}$ is given by [50]:

$$\boldsymbol{\beta} = \left(\boldsymbol{X^T X} + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{X^T y}, \quad (7)$$

where λ is a regularization coefficient and $\boldsymbol{I}$ is the identity matrix. Ridge regression is known to perform better than linear regression in case of mutually correlated explanatory variables (molecular descriptors in our case) [50].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 October 2023 doi:10.20944/preprints202310.0103.v1

2.4.2. Decision Tree

First introduced in 1987 [51] decision trees are hierarchical supervised machine learning models that logically combine a sequence of decisions based on simple tests and their possible outcomes. This is achieved by optimizing the simple test condition threshold during the training process [52]. In the course of training, all possible data splits are considered:

$$Q_m^l = \{(x, y) | x_j < t_m\} \tag{6a}$$

$$Q_m^r = Q_m \setminus Q_m^l, \tag{6b}$$

where Q_m is the data at node m, Q^l_m and Q^r_m are the candidate splits, $\mathbf{x}$ are the training data vectors, $\boldsymbol{y}$ is the target variable vector. The threshold condition is optimized by comparing the quality of the splits by using an appropriate cost function. For regression decision trees the mean square error (MSE – eqn. (8a)) or the Poisson deviance (eqn. (8b)) can be used as cost functions [51]:

$$\overline{y}_m = \frac{1}{n_m} \sum_{y \in Q_m} y \tag{7}$$

$$H(Q_m) = \frac{1}{n_m} \sum_{y \in Q_m} \left(y - \overline{y}_m\right)^2 \tag{8a}$$

$$H(Q_m) = \frac{1}{n_m} \sum_{y \in Q_m} \left(ylog\left(\frac{y}{\overline{y}_m}\right) - y - \overline{y}_m\right)^2 \tag{9b}$$

This splitting operation is performed for all the features and the feature split that leads to the largest decrease in the cost function is kept at node m. This allows the estimation of the feature importance – the more a feature split decreases the cost function, the more important the feature.

It should be noted, that due to their structure of sequential simple tests, decision trees are able to capture nonlinear dependencies between the explanatory variables and the measured property. Decision trees have been successfully utilized to solve both classification and regression problems [53-55]. There exist numerous algorithms for decision tree construction: ID3, C4.5, CART, MARS, and CHAID [56] In the present study, the CART algorithm with mean square error cost function was used.

2.4.3. Random Forest

Random forests are ensemble machine learning algorithms that can be used for classification and regression. Multiple decision trees are constructed using randomly selected explanatory variables (molecular descriptors in our case) and each tree is trained on different bootstrapped sample (sampling, allowing multiple selection of same items) of the training set. When a prediction is made, the average result of all trees is returned [57].

2.4.4. Extra-Trees

The Extra-trees algorithm [58] is similar to the Random forest algorithm – a multitude of decision trees are used; however, the individual decision trees are trained on subsamples of the training set taken without replacement (in contrast to bootstrapping). Another important difference is that in the Extra-trees algorithm the cut point is selected randomly, while in the Random forest algorithm the optimal split is chosen. These differences generally lead to reduction of bias and variance. The random choice of a cut point also makes the algorithm faster (in the Random forest algorithm the optimal split is found by computing some impurity metric for all possible splits).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 October 2023 doi:10.20944/preprints202310.0103.v1

2.4.5. Gradient Boosting

The gradient boosting relies on fitting of a sequence of weak prediction models (decision trees in this case) on repeatedly altered versions of training data. [59] The predictions of all individual weak predictors are combined as a sum:

$$\hat{y}_i = F_M(\boldsymbol{x}_i) = \sum_{m=1}^{M} h_m(\boldsymbol{x}_i), \quad (9)$$

where $\hat{y}_i$ is the model prediction, $\boldsymbol{x}_i$ is a vector of all features that describes the i-th object in the dataset (in our case, all descriptors used to represent a molecule), M is the number of weak estimators and $h_m(\boldsymbol{x}_i)$ is the prediction of the m-th weak estimator. From eqn. (9) it follows that:

$$F_m(\boldsymbol{x}_i) = F_{m-1}(\boldsymbol{x}_i) + h_m(\boldsymbol{x}_i)\ . \quad (10)$$

The weak predictor $h_m(\boldsymbol{x}_i)$ in eq. (10) is fitted to minimize a sum of cost functions Cm:

$$h_m = argmin_h(C_m) = argmin_h\left(\sum_{i=1}^{n} c(y_i, F_{m-1}(\boldsymbol{x}_i) + h_m(\boldsymbol{x}_i))\right), \quad (11)$$

where n is the number of training entries and $c(y_i, F(\boldsymbol{x}_i))$ is a cost function, such as mean square error (MSE, eqn. (8a)).

Friedman [59] proposed a regularization strategy, based on scaling the contribution of each new weak predictor by a learning rate γ:

$$F_m(\boldsymbol{x}_i) = F_{m-1}(\boldsymbol{x}_i) + \gamma h_m(\boldsymbol{x}_i). \quad (12)$$

It has been demonstrated [60] that in many cases Gradient boosting outperforms other ensemble methods such as Random forests and Extra trees.

In the present study, all machine learning algorithms were exploited as implemented in the scikit-learn library [61].

3. Results and Discussion

The redox potential distribution (Figure 2) over the entire dataset (494 compounds) shows that the redox potential of the majority of compounds lies between 0.75 V and 1.60 V which makes them suitable for cathode materials.

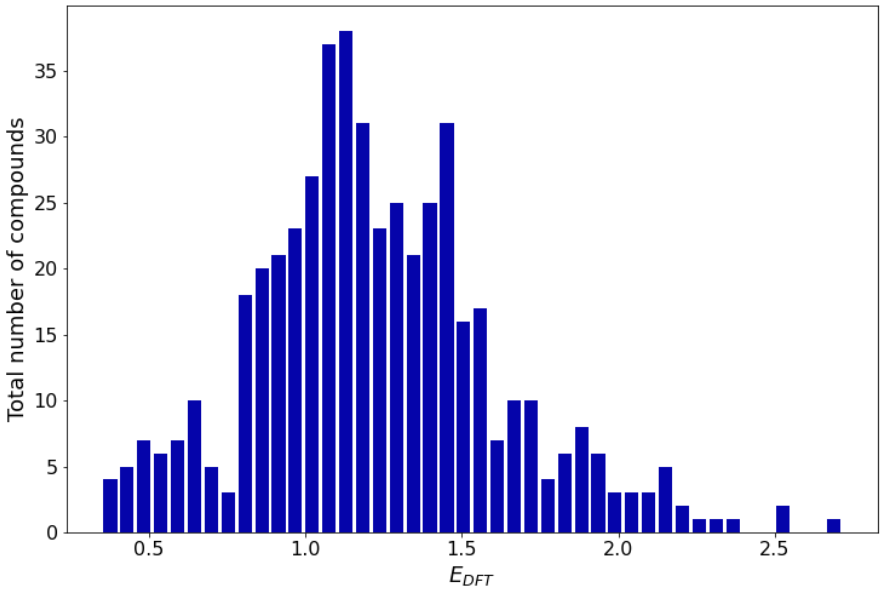


Figure 2. Redox potential distribution histogram.

In order to find an optimal approach for machine learning modelling of structure-redox potential relationship the following algorithms were tested: Ridge regression, Decision tree, Random

 doi:10.20944/preprints202310.0103.v1

forest, Extra-trees and Gradient boosting. Artificial neural networks were not considered since they are prone to overfitting, especially when trained on an insufficient amount of data [62].

In order to attain maximal predictive ability, the hyperparameters (parameters that control the learning process) of each of the machine learning models were optimized by a grid search. The model performance was evaluated by the averaged coefficient of determination R^2 [63] of the 5-fold cross validation over the training set. The training R^2 was also taken into account, since the difference between validation and training R^2 can be used to judge whether the model is overfitted.

The l2-regularization value (λ in eqn. (5)) in Ridge regression does not have a significant impact on the model performance (Figure 3a): increasing of the l2-regularization value leads to a decrease (by the almost equal amount) of the training and validation R^2. It should be noted that the difference between training and validation R^2 reached a minimum at $\lambda = 0.1$, and hence, this value of lambda results in an optimal (neither underfitted, nor overfitted) Ridge regression model.

Decision tree maximal allowed depth plays a central role in determining whether the decision tree underfits or overfits the training data: larger maximal allowed depth results in a deeper tree that fits the training data better; however, if a tree is too deep, the noise in the training data is also learned, i.e., the decision tree overfits. In the present work the maximal tree depth was varied from 2 to 20 (Figure 3b). Optimal algorithm performance was attained when maximal tree depth was 3. A serious advantage of the Decision trees is their ability to visualize the learning process (Figure 4). Furthermore, the Decision tree algorithm enables examination of the descriptor significance. It was found that the most significant descriptors: MAXDN2, LUMO, SaasC, SHdsCH, and BCUTc-1H, are all related to electronic structure of the molecules – quinones that contain more CN and C≡C-Me groups (lower LUMO, large MAXDN2 due to CN-groups) exhibit larger redox potential.

To examine the predictive ability of maximal Random forest regression and Extra-trees regression, the depth of decision tree estimator was set to 3 (since we established that this value of maximal depth ensures maximal learning performance) and the number of estimators was optimized to achieve maximal coefficient of determination over the validation set (Figures 3c and 3d): 10 and 15 estimators were chosen for Random forest and Extra-trees, respectively.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 October 2023 doi:10.20944/preprints202310.0103.v1

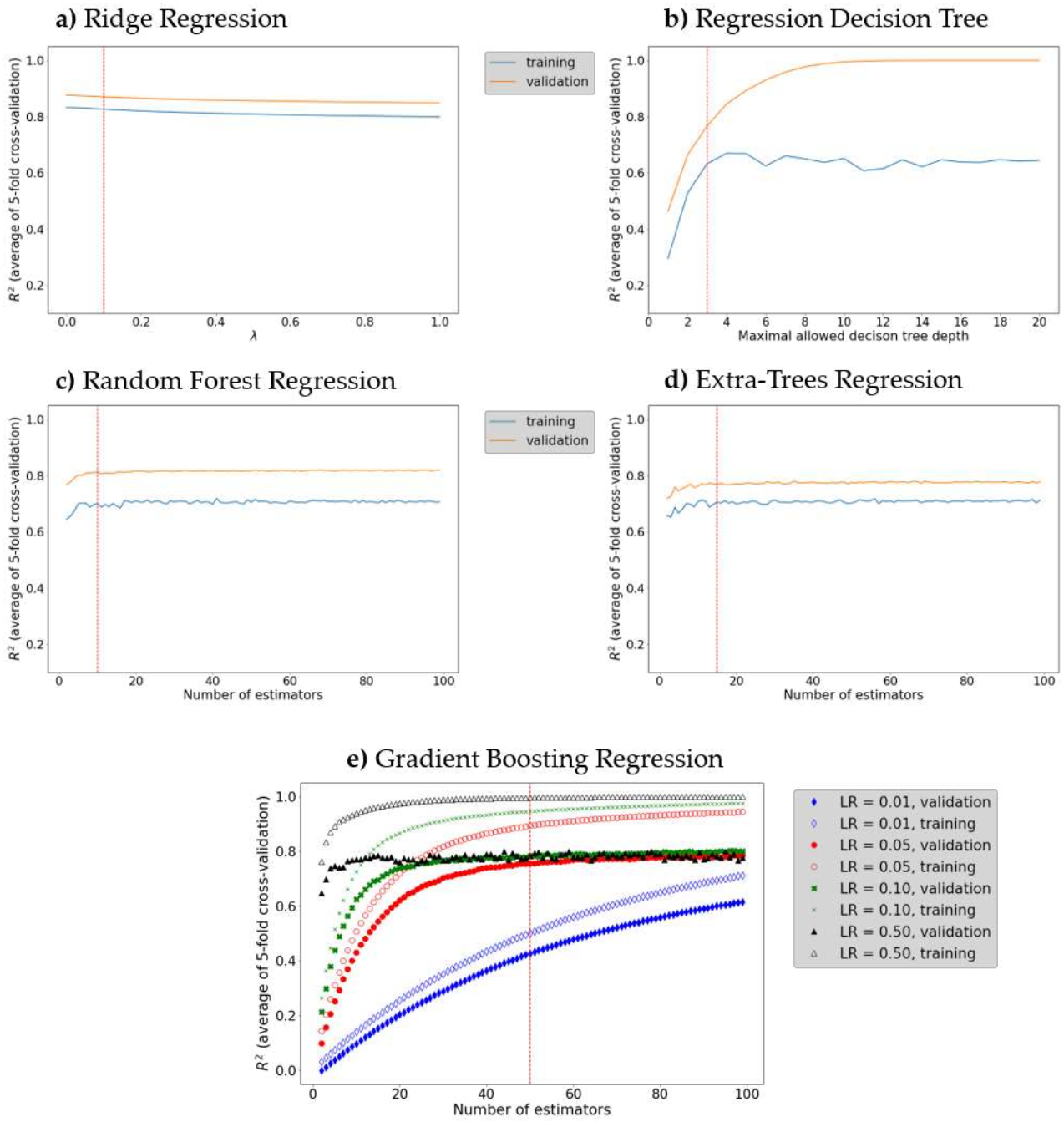


Figure 3. Model tuning by grid search for: (**a**) optimal learning rate in Ridge regression; (**b**) optimal maximal Decision tree depth; (**c** and **d** respectively) optimal number of decision tree estimators in Random forest regression and Extra-trees regression; (**e**) optimal number of decision tree estimators and learning rate in Gradient boosting.

It was found that the Extra-trees algorithm is less prone to overfitting: R^2 value over the validation set is closer to the R^2 value over the test set. Random forest and Extra-trees algorithms can also be used to estimate the descriptors importance - the most significant descriptors for the Decision tree (described above) are found among the ten most significant descriptors of both algorithms, which confirms that the descriptors related to the electronic structure, such as the LUMO energy, and descriptors derived from electronegativity, such as SaasC, SHdsCH, MAXDN2, and meanI, are important for machine-learning prediction of the redox potential of organic energy storage materials. As expected, we found that the Gradient boosting regression exhibits a better predictive ability and is less prone to overfitting than the other ensemble methods used (Random forest regression and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 October 2023 doi:10.20944/preprints202310.0103.v1

Extra trees regression). Learning rate (γ in eqn. (12)) and number of weak predictors values of 0.05 and 50, respectively, were found by grid searching (Figure 3e).

The prediction models performance, as evaluated by the average coefficient of determination over a 5-fold cross-validation (R^2_{CV}), increases in the following order: regression Decision tree (R^2_{CV} = 0.632), Random forest regression (R^2_{CV} = 0.705), Extra trees regression (R^2_{CV} = 0.715), Gradient boosting regression (R^2_{CV} = 0.756) and Ridge regression (R^2_{CV} = 0.832).

All machine learning algorithms were evaluated on the test set. To visualize model performance, scatter plots of redox potential, calculated by density functional theory, versus redox potential, estimated by the corresponding machine learning algorithms, are drawn (Figure 5). Linear regression was implemented to construct a trendline in the (E_{model}, E_{DFT}) – space (Figure 5, red lines) and the slope, intercept and coefficient of determination (R^2) of the trendline were calculated. When the model ideally fits the data, the trendline slope is supposed to have one and zero for slope and trendline, respectively, and the R^2-value should be close to one. It was found that the models performance on test set does not differ significantly from the models performance observed upon the 5-fold cross-validation, which means that the models fit the data fairly well (i.e., models are not significantly overfitted or underfitted). All of the models tend to give worse prediction for large voltages: a possible explanation is that in the training set there are less molecules exhibiting a high redox potential.

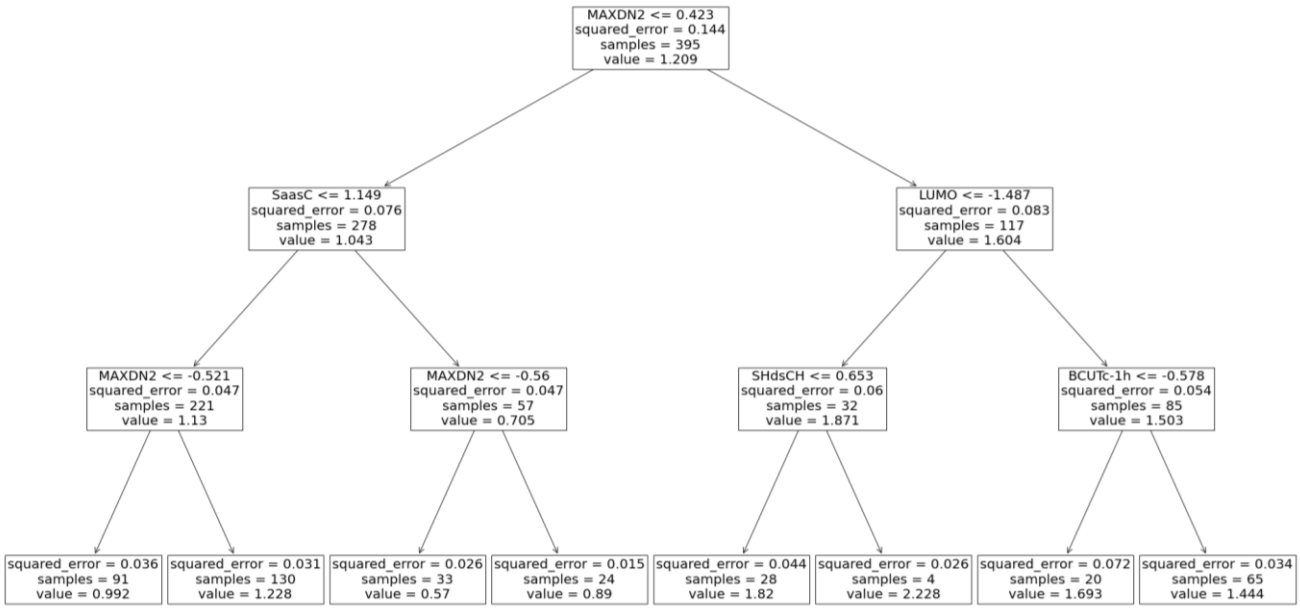


Figure 4. Regression decision tree with maximal depth of 3 chart.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 October 2023 doi:10.20944/preprints202310.0103.v1

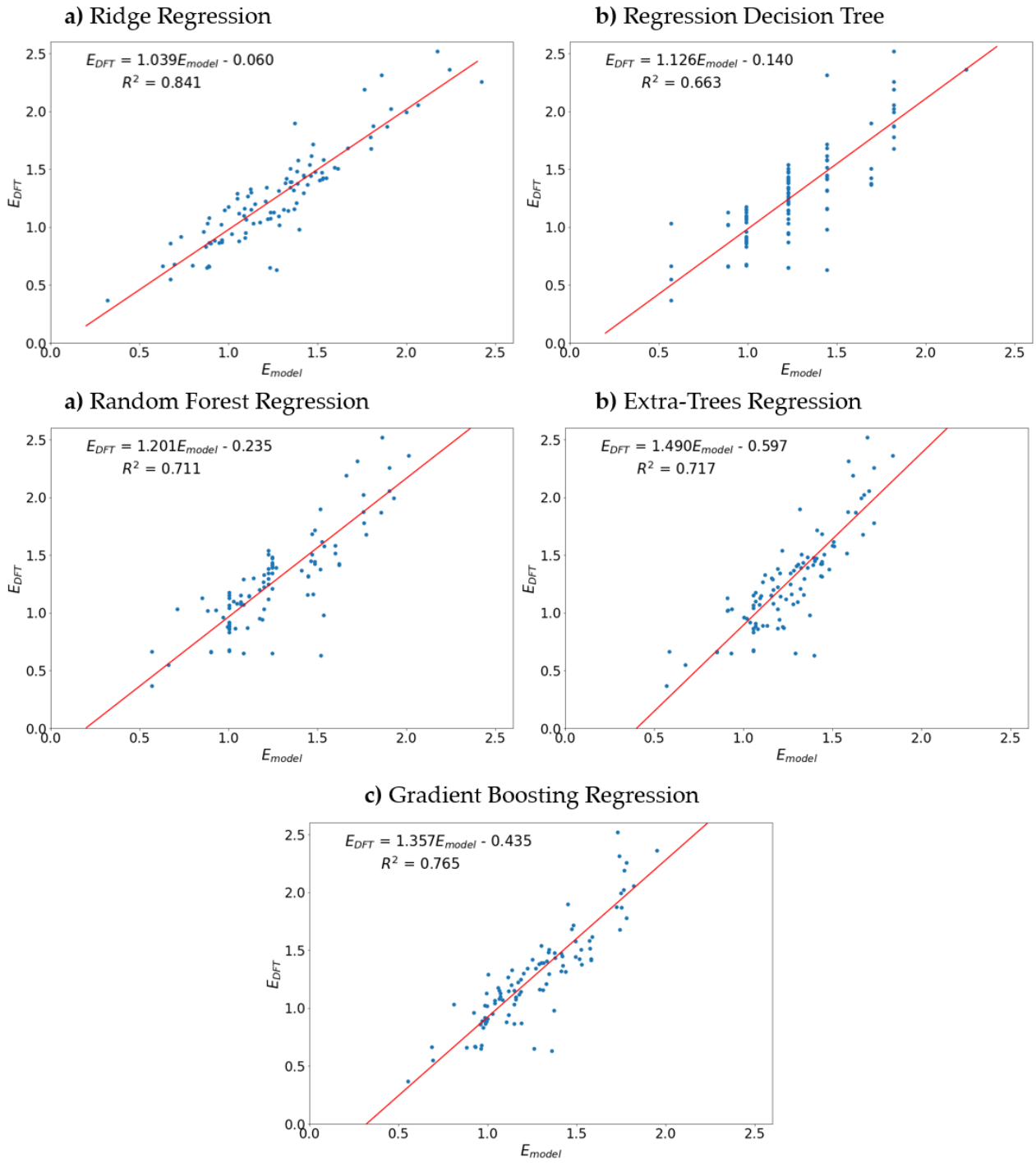


Figure 5. Models performance on the test set.

4. Conclusions

We have constructed a dataset of 494 potential organic electrode materials by automated generation of derivatives of 100 quinones extracted from a general purpose public database (PubChem). A descriptor selection procedure that combines low-variance descriptor removal with covariance matrix analysis and stepwise linear regression for finding uncorrelated descriptors, on which the redox potential of the molecules in the dataset depends, was devised. Due to the comparatively small dataset size, deep learning approaches were not deployed as inappropriate, since they are prone to overfitting when trained on small amounts of data. Five different supervised machine learning models for regression that tend to give better results for smaller datasets were built. The hyperparameters of all those models were tuned to attain maximal electrode potential predictive

ability. The models performance was evaluated on a test set containing molecules that are completely unknown to the model. It was established that model performance increases in the following order: Regression decision tree < Random forest regression < Extra trees regression, Gradient boosting regression < Ridge regression. It turned out that the linear model, i.e., the Ridge regression outperforms the decision tree based algorithms, known to be able to capture nonlinear dependencies between descriptors and target variable. This is an implication that the relationship between the electrode potential and some chemical properties is most probably linear. In particular, it was found that descriptors related to the electronic structure (LUMO and E-state descriptors) have a large significance. In addition, the Ridge regression is an excellent method for screening of databases as it is a very fast and computationally inexpensive approach, compared to other machine learning algorithms.

Author Contributions: Conceptualization, A.T. and R.S.; methodology, I.K. and L.B.; software, I.K. and L.B.; validation, L.B.; formal analysis, L.B.; resources, L.B.; data curation, I.K.; writing—original draft preparation, I.K. and L.B.; writing—review and editing, A.T.; visualization, L.B.; supervision, R.S. and A.T.; project administration, R.S.; funding acquisition, R.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by CARIM-VIHREN, grant number КП-06-*Д*В-6/2019 and the APC was funded by European Twinning on Materials Chemistry Enabling Clean Technologies (TwinTeam), grant number D01-272/10.2020.

Data Availability Statement: The computed data is available from the authors on request.

Acknowledgments: European Regional Development Fund within the Operational Programme "Science and Education for Smart Growth 2014–2020" under the Project CoE "National center of mechatronics and clean technologies" (BG05M2OP001-1.001-0008) – for computational facilities; I.K. is grateful to Prof. Pascal Friederich for internship hosting; thanks are due to Dr. Nina Markova and Yanislav Danchovsky for the initial dataset (100 compounds) construction.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Poizot, P.; Dolhem, F. Clean energy new deal for a sustainable world: from non-CO2 generating energy sources to greener electrochemical storage devices *Energy Environ. Sci.* **2011**, *4*, 2003-2019
2. Larcher, D.; Tarascon, J.M. Towards greener and more sustainable batteries for electrical energy storage. *Nat. Chem.* **2015**, *7*, 19–29.
3. Poizot, P.; Gaubicher, J.; Renault, S.; Dubois, L.; Liang, Y.; Yao, Y. Opportunities and Challenges for Organic Electrodes in Electrochemical Energy Storage. *Chem. Rev*. **2020**, *120*, 6490–6557
4. Schon, T.B.; McAllister, B.T.; Li, P.-F.; Seferos, D.S. The rise of organic electrode materials for energy storage. *Chem. Soc. Rev*. **2016**, *45*, 6345–6404
5. Lu, Y.; Zhang, Q.; Li, L.; Niu, Z.; Chen, J. Design Strategies toward Enhancing the Performance of Organic Electrode Materials in Metal-Ion Batteries. *Chemistry* **2018**, *4*, 2786–2813
6. Lu, Y.; Chen, J. Prospects of organic electrode materials for practical lithium batteries. *Nat. Rev. Chem.* **2020**, *4*, 127–142.
7. Esser, B.; Dolhem, F.; Becuwe, M.; Poizot, P.; Vlad, A.; Brandell, D. A perspective on organic electrode materials and technologies for next generation batteries. *J. Power Sources* **2021**, 482, 228814
8. Yan, L.; Zhao, C.; Sha, Y.; Li, Z.; Liu, T.; Ling, M.; Zhou, S.; Liang, C. Electrochemical redox behavior of organic quinone compounds in aqueous metal ion electrolytes. *Nano Energy* **2020** *73* 10476
9. Tobishima, S.; Yamaki, J.; Yamaji A. . Cathode Characteristics of Organic Electron. Acceptors for Lithium Batteries *J. Electrochem. Soc*. **1984**, *131*, 57-63
10. Senoh, H.; Yao, M.; Sakaebe, H.; Yasuda, K.; Siroma, Z. A two-compartment cell for using soluble benzoquinone derivatives as active materials in lithium secondary batteries. *Electrochimica acta* **2011,** *56*, 10145–10150
11. Genorio, B.; Pirnat, K.; Cerc-Korosec, R.; Dominko, R.; Gaberscek, M. Electroactive Organic Molecules Immobilized onto Solid Nanoparticles as a Cathode Material for Lithium-Ion Batteries *Angew. Chem. Int. Ed*. **2010,** *49*, 7222-7224
12. Foos, J. S.; Erker, S.M.; Rembetsy, L.M. Synthesis and Characterization of Semiconductive Poly-l,4-Dirnethoxybenzene and Its Derived Polyquinone *J. Electrochem. Soc*. **1986,** *133*, 836-840

13. Häringer, D.; Novák, P.; Haas, O.; Piro, B.; Pham, M.-C. Poly(5-amino-1,4-naphthoquinone), a Novel Lithium-Inserting Electroactive Polymer with High Specific Charge *J. Electrochem. Soc.* **1999,** *146*, 2393-2396
14. Gall, T.L.; Reiman, K.H.; Grossel, M.C.; Owen, J.R. Poly(2,5-dihydroxy-1,4-benzoquinone-3,6-methylene): a new organic polymer as positive electrode material for rechargeable lithium batteries, Journal of Power Sources *J. Power Sources* **2003,** *119–121* 316-320
15. Son, E. J.; Kim, J. H.; Kim, K.; Park, C. B. Quinone and its derivatives for energy harvesting and storage materials. *J. of Mat. Chem. A*, **2016**, *4(29)*, 11179–11202
16. J. Q. Chambers, *Quinonoid Compounds*, 1-st ed. John Wiley & Sons Ltd: Hoboken, NJ, USA, 2010
17. Mueller, T.; Kusne, A.G.; Ramprasad R. Machine learning in materials science: recent progress and emerging applications In *Rev. Comput. Chem.*, Vol. 29; Parrill, A. L.; Lipkowitz, B. K., Eds, John Wiley & Sons, Inc.: Indianapolis, Indiana, USA 2016 p.p. 186-273
18. Joshi, R. P.; Eickholt, L.; Li, L.; Fornari, M.; Barone, V.; Peralta, J. E. Machine Learning the Voltage of Electrode Materials in Metal-ion Batteries. *ACS Appl. Mater. Interfaces* **2019**, *11(20)*, 18494–18503
19. Zhang, X.; Zhou, J.; Lu, J.; Shen, L. Interpretable learning of voltage for electrode design of multivalent metal-ion batteries. *npj Comput Mater* **2022,** *8*, 175
20. Allam, O.; Kuramshin, R.;Stoichev, Z.; Cho, B. W.; Lee, S. W.; Jang, S. S. Molecular structure–redox potential relationship for organic electrode materials: density functional theory–Machine learning approach *Materials Today Energy* **2020** *17*,100482
21. Tuttle, M. R.; Brackman, E. M.; Sorourifar, F.; Paulson, J.; Zhang, S. Predicting the Solubility of Organic Energy Storage Materials Based on Functional Group Identity and Substitution Pattern *J. Phys. Chem. Lett*. **2023**, *14*, 1318−1325
22. Wang, F.; Li, J.; Liu, Z.; Qiu, T.; Wu, J. Lu, D. Computational design of quinone electrolytes for redox flow batteries using high-throughput machine learning and theoretical calculations *Front. Chem. Eng*. **2022**, *4*
23. Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2023 update. *Nucleic Acids Res.* **2023**, *51*, D1373–D1380
24. Open Babel development team. (2016). Open Babel. Retrieved from http://openbabel.org/wiki/Main_Page
25. Kichev, I.; Borislavov, L.; Tadjer, A. Automated generation of molecular derivatives – DerGen software package *Materials Today: Proceedings* **2022,** *61*, 1287–1291
26. Nayak, S.K.; Ojha, A.C. (). Data Leakage Detection and Prevention: Review and Research Directions. *In Machine Learning and Information Processing. Advances in Intelligent Systems and Computing*, vol 1101.; Swain, D.; Pattnaik, P.; Gupta, P. Eds., Springer, Singapore. 2020, p.p. 203–212
27. Wigh, D. S.; Goodman, J. M.; Lapki, A. A. A review of molecular representation in the age of machine learning, *WIREs Comput Mol Sci*. **2022**, *12*, e1603.
28. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules *J. Chem. Inf. Comput. Sci*. **1988**, *28(1)*, 31–36
29. Dalby, A.; Nourse, J. G.; Hounshell, W. G. ;Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited, *J. Chem. Inf. Comput. Sci.* **1992**, *32(3)*, 244–255
30. Todeschini, R.; Consonni Handbook of Molecular Descriptors. Methods and Principles in Medicinal Chemistry. WILEY-VCH Verlag GmbH: Weinheim (Federal Republic of Germany) 2000
31. Cereto-Massagué, A.; Ojeda, M.J; Valls, C.; Mulero, M.; Garcia-Vallvé, S.; Pujadas, G. Molecular fingerprint similarity search in virtual screening, *Methods* **2015,** *71* 58-63
32. Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965,** *5 (5),* 107–113
33. Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints *J. Chem. Inf. Model.* **2010,** *50 (5)* 742–754
34. Elton, D. C.; Boukouvalas, Z.; Fuge, M. D.; Chung, P. W. Deep learning for molecular design—a review of the state of the art. *Mol. Syst. Des. Eng*. **2019,** *4*, 828-849
35. Kuzminykh, D.; Polykovskiy, D.; Kadurin, A.; Zhebrak, A.; Baskov, I.; Nikolenko, S.; Shayakhmetov, R.; Zhavoronkov, A. 3D Molecular Representations Based on the Wave Transform for Convolutional Neural Networks *Mol. Pharmaceutics* **2018,** *15(10)*, 4378–4385
36. Skalic, M.; Jiménez Luna, J.; Sabbadin, D.; De Fabritiis, G Shape-Based Generative Modeling for de-novo Drug Design *J. Chem. Inf. Model.* **2019**, *59(3)* 1205–1214
37. Yap, C. W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem.* **2011,** *32(7)* 1466-1474
38. MOPAC2016, James J. P. Stewart, Stewart Computational Chemistry, Colorado Springs, CO, USA, HTTP://OpenMOPAC.net (2016)

39. Pudjihartono, N.; Fadason, T.; Kempa-Liehr, A. W.; O'Sullivan, J. M. A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction *Front. Bioinform.* **2022,** *2*
40. Hocking, R. R. The Analysis and Selection of Variables in Linear Regression *Biometrics*, **1976,** *32*
41. Burden, F.R., Molecular identification number for substructure searches , *J. Chem. Inf. Comput. Sci*., **1989,** *29*, 225-227;
42. Sharma, V.; Goswami, R.; Madan, A.K. Eccentric Connectivity Index: A Novel Highly Discriminating Topological Descriptor for Structure-Property and Structure-Activity Studies *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 273-282
43. Hall, L. H.; Kier, L. B. Electrotopological state indices for atom types: A novel combination of electronic, topological, and valence state information. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1039-1045
44. Gramatica, P.; Corradi, M.; Consonni, V. Modelling and prediction of soil sorption coefficients of non-ionic organic pesticides by molecular descriptors. *Chemosphere* **2000**, *41*, 763-777.
45. Nilakantan, R.; Nunn, D.S.; Greenblatt, L.; Walker, G.; Haraki, K.; Mobilio, D. A family of ring system-based structural fragments for use in structure-activity studies: database mining and recursive partitioning., *J. Chem. Inf. Model* **2006**, *46* 1069-1077
46. Todeschini, R.; Consonni, V. Molecular descriptors for chemoinformatics, WILEY-VCH Verlag GmbH: Weinheim (Federal Republic of Germany) 2009 p.p. 809-812
47. Zhao, Y. H.; Abraham, M. H.; Zissimos, A. M. Fast Calculation of van der Waals Volume as a Sum of Atomic and Bond Contributions and Its Application to Drug Compounds *JACS* **2003**, *68*, 7368-7373
48. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko,B. G.; Gomperts, R.; Mennucci, B. et.al, „Gaussian 16, Revision C.01," Gaussian, Inc., Wallingford CT, 2016.
49. Ochterski, J. W. Thermochemistry in Gaussian, Gaussian, Inc. 2000
50. Hoerl, A. E.; Kennard, R. W. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, **1970**, *12(1)*, 55–67.
51. Breiman, L. Classification and Regression Trees, 1-st ed.; CHAPMAN & HALL/CRC: Boca Raton, FL, USA 1984
52. Kotsiantis, S. B. Decision trees: a recent overview. Artif. Intell. Rev. **2013**, *39*, 261–283.
53. Klekota, J.; Roth, F. P. Chemical substructures that enrich for biological activity. *Bioinformatics* **2008,** 24, 2518–2525
54. Hou, T.; Wang, J.; Li, Y. ADME evaluation in drug discovery. The prediction of human intestinal absorption by a support vector machine. *J. Chem. Inf. Model* **2007**, *47*, 2408–2415
55. Lamanna, C.; Bellini, M.; Padova, A.; Westerberg, G.; Maccari, L. Straightforward recursive partitioning model for discarding insoluble compounds in the drug discovery process. *J. Med. Chem.* **2008**, *51*, 2891–2897
56. Patel, H. H.; Prajapati, P. Study and Analysis of Decision Tree Based Classification Algorithms, *Int. J. Comput. Sci. Eng.* **2018** , *6(10)*, 74-78,
57. Ho, T. K. The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence* **1998**, 20, 832–844
58. Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **2006**, *63*, 3–42
59. Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **2001**, 1189– 1232
60. Piryonesi, S. M.; El-Diraby, T. E. Data Analytics in Asset Management: Cost-Effective Prediction of the Pavement Condition Index *J. Infrastruct. Syst* **2020**, *26(1)*, 04019036
61. Pedregosa, D.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O., Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. et al. Scikit-learn: Machine Learning in Python *J. Mach. Learn. Res.* **2011**, 12, 2825−2830
62. Bejani, M.M.; Ghatee, M. A systematic review on overfitting control in shallow and deep neural networks. *Artif Intell Rev* **2021**, *54*, 6391–6438
63. Wright S. Correlation and causation. Journal of Agricultural Research. **1921,** *XX(7),* 557–585