INDUSTRIAL AND SYSTEMS ENGINEERING

ISE

Machine Learning for the Redox Potential Prediction of Molecules in Organic Redox Flow Battery

PEIYUAN GAO[1], DIDEM KOCHAN[1,2], YU-HANG TANG[3], XIU YANG[2], AND EMILY G. SALDANHA[4]

[1]Physical and Computational Directorate, Pacific Northwest National Laboratory, Richland, WA, 99352, USA

[2]Department of Industrial and Systems Engineering, Lehigh University, Bethlehem, PA, 18015 USA

[3]NVIDIA Corporation, 2788 San Tomas Expy, Santa Clara, CA, 95051, USA

[4]National Security Directorate, Pacific Northwest National Laboratory, Richland, WA, 99352, USA

ISE Technical Report 25T-012

LEHIGH
UNIVERSITY

Journal of Power Sources 629 (2025) 236035

Contents lists available at ScienceDirect

Journal of Power Sources

journal homepage: www.elsevier.com/locate/jpowsour

Machine learning for the redox potential prediction of molecules in organic redox flow battery

Peiyuan Gao [a,*], Didem Kochan [a,b], Yu-Hang Tang [c], Xiu Yang [b,**], Emily G. Saldanha [d,***]

[a] Physical and Computational Directorate, Pacific Northwest National Laboratory, Richland, WA, 99352, USA
[b] Department of Industrial and Systems Engineering, Lehigh University, Bethlehem, PA, 18015, USA
[c] NVIDIA Corporation, 2788 San Tomas Expy, Santa Clara, CA, 95051, USA
[d] National Security Directorate, Pacific Northwest National Laboratory, Richland, WA, 99352, USA

HIGHLIGHTS

- A comprehensive experimental redox potential database with pH value for ORFB.
- Lightweight Graphics machine learning models with GPU acceleration for ORFB.
- Accurate prediction on both experimental and computational data of ORFB.

ABSTRACT

Organic redox flow batteries (ORFB) are recognized as an innovative technology for the large-scale storage of renewable energy. The redox potential of organic redox-active molecules plays a vital role in their performance. Advanced screening techniques like high-throughput experiment and machine learning (ML) have significantly enhanced organic material performance and transformed the field of ORFB. However, the scarcity of experimental data poses a considerable challenge for ML model development in this domain. In our study, we developed lightweight graph-based Gaussian process regression (GPR) models with GPU-accelerated marginalized graph kernel and hybrid kernel to predict the redox potentials of organic redox-active molecules for ORFBs, specifically focusing on small datasets. To evaluate model accuracy, we created a new experimental database of organic redox-active molecules by the data from hundreds of published papers and assembled previous computational datasets. We also considered some key parameters, such as pH conditions and solvent type, to assess their impact on redox potential prediction. Our GPR model predicted redox potentials with high accuracy across all datasets using minimal training data. The study provides powerful tools for molecule screening and design and delivers valuable guidance on designing training datasets for costly experiments.

1. Introduction

Redox flow battery (RFB) is a type of rechargeable battery where energy is stored in liquid electrolytes containing electroactive species. These electrolytes are stored in external tanks and circulated through an electrochemical cell stack during charge and discharge cycles. Redox flow batteries are particularly suitable for large-scale energy storage applications due to their scalability, flexibility, and long cycle life. There are various RFB technologies including vanadium redox flow battery (VRFB), zinc-bromine flow battery, organic redox flow battery (ORFB) and iron-chromium flow battery [1–3]. Among these redox flow batteries, ORFBs, including both aqueous and nonaqueous ORFB, are a type of redox flow battery that uses organic molecules dissolved in water or organic solvent as the electroactive species [4–6]. This battery technology is being developed to provide a more sustainable, cost-effective, and environmentally friendly option for large-scale energy storage compared to traditional batteries like lithium-ion batteries. Instead of using metal-based ions (like in vanadium redox flow batteries), ORFBs use organic molecules. Organic materials can be cheaper and more abundant than the metals used in other flow batteries [6,7]. Organic materials can be sourced from renewable resources, making the technology more environmentally friendly. Also, these organic materials can be designed and synthesized to have desirable electrochemical properties, such as higher solubility and cell voltage, and stable redox states

* Corresponding author.
** Corresponding author.
*** Corresponding author.
E-mail address: peiyuan.gao@pnnl.gov (P. Gao).

https://doi.org/10.1016/j.jpowsour.2024.236035
Received 14 October 2024; Received in revised form 26 November 2024; Accepted 5 December 2024
Available online 16 December 2024

[8]. ORFBs generally have lower energy densities compared to some other battery technologies. To enhance the energy density of ORFBs, the key challenge is to identify organic molecules with appropriate redox potential values. Whether in theoretical or practical flow battery systems, a higher redox potential for the catholyte and a lower redox potential for the anolyte are always desired [9–11]. Predicting the redox potentials of molecules is also essential in other fields such as electrocatalysis, medical chemistry and environmental science [12–14]. For these relevant applications, choosing the best candidates from an essentially infinite chemical space for experimental testing of the redox potential property requires efficient screening approaches. Many organic compounds such as quinones, viologens, flavins, thiazines, imides, and their derivatives have been investigated for redox-active species in both aqueous and non-aqueous RFBs. Among them, quinone and phenazine which are structural analogues to anthraquinones, their reduction mechanism depends on the media, occurring either via a 2-electron reduction (ET) in two steps in aprotic solvents, or via a $2e^{-}/2H^{+}$ proton-coupled electron transfer (PCET) mechanism in a single step in protic media. These chemical reactions will lead to different outcomes in redox potential. Additionally, their redox potentials strongly depend on the nature of functional groups in the molecular structure [15,16].

With the development of computational power, quantum chemistry calculations are crucial in material science because they offer a fundamental understanding of the electronic structure and properties of materials at the atomic and molecular levels [17]. Methods such as Density Functional Theory (DFT) have been widely used to characterize the electronic and structural parameters of various compounds in RFB. By predicting the behavior of these materials, quantum chemistry calculation enables researchers to design and optimize new materials with desired properties. This predictive capability is essential for developing advanced materials for applications ranging from sustainable energy solutions and catalysis to electronic devices and nanotechnology. Moreover, the use of quantum chemistry calculations can significantly reduce the need for time-consuming and costly experimental procedures, thereby accelerating the pace of innovation and discovery in material science. For example, Asenjo-Pascual et al. performed Natural Bond Orbitals (NBO) and Atom Dipole Correction Hirshfeld (ADCH) charge distribution analyses in DFT calculations to evaluate stability of the compounds in ORFB [18]. Achazi et al. designed a multi-step procedure with DFT calculations for molecules screening in organic radical polymer anodes [19]. The redox potential property of a molecule, including the oxidation and reduction potentials, can also be determined through DFT calculation. The oxidation/reduction potential of molecule in solution is a sum of adiabatic ionization energy/adiabatic electron affinity and the corresponding solvation free energy. In calculation, redox potentials can be estimated either directly using continuum solvent models or via Born–Haber thermodynamic cycles of either a half- or full reaction. Accurate reproduction of experimental redox potentials is therefore a benchmark test for the methods of computational chemistry. Also, some DFT calculation datasets have been built [20], such as the ROP313 [21], RedDB [22], OREDOX159 [23] and CompBatPET dataset [24]. However, due to the high computational cost of high accuracy quantum chemistry calculation, most calculations have been performed using approximated methods, such as continuum solvation models. Unfortunately, to date, accurate predictions of this crucial property based on first principles calculation remain challenging, with typical prediction errors around 0.5 V. This is because for common functionals in DFT, the average free energy difference in gas phase calculation could be 1–10 kcal/mol, and the error of free energy calculations in solvent might be even larger [25,26]. Based on the thermodynamic cycle and the Nernst equation, the error of redox potential would be 0.04 V if the free energy calculation deviation is 1 kcal/mol. To reduce the error, it will require expensive quantum chemistry composite methods like G4 [27]. The accurate calculation of solvation free energy is another problem. To be useful for applications, for example, in catalysis, errors should not exceed 0.2 V (3 pK units at ambient temperature) [28]. Although a few machine learning (ML) models have been developed to correct the error on the free energy, the error is still large, and the models that provide correction need additional training for a new dataset [29–31]. Therefore, ML models utilizing experimental data would be more directly applicable in both lab and industrial contexts.

The group contribution method has been one of the most widely used approaches to estimate standard Gibbs energies and redox potentials when there is no experiment data, especially in biochemical process [32]. Advanced mathematical methods, such as multiple linear regression analysis based on chemistry descriptor, have gradually entered this field [33]. Since then, numerous structural and energetic descriptors have been developed for predicting redox potentials [34,35], while the focus has shifted from biologically active to industrially relevant compounds. Recently, ML methods, including traditional and novel ML and deep learning (DL) techniques such as deep neural networks and graph neural networks have also been used in redox potential prediction [36–41]. Despite many advanced approaches that have been developed for redox potential prediction, one key obstacle in developing valuable models is the insufficient amount of high-quality experimental data. As the redox potential depends on many factors like pH value and solvent type, the parameter space of the data is very large, and the data collection would be expensive even with high through-put electrochemical characterization [42–44].

Gaussian process model is a type of lightweight ML model for training data. In this work, we developed molecular graph based Gaussian process regression (GPR) model with graphics processing unit (GPU) acceleration for redox potential prediction of molecules, targeting the organic redox-active species in ORFB. To address the issue of data lack, we built a new comprehensive experimental redox potential database by compiling data from hundreds of ORFB published papers. The database includes more than 500 redox potential measurements of organic redox-active molecules in water and/or organic solvents. To the best of our knowledge, this is currently the largest experimental redox potential database for ORFBs. The experiment parameters of pH value (for aqueous ORFB), type of solvent, chemical identifiers and corresponding reference are contained as well. We perform tests on the subsets including the redox-active molecules that involve one-electron or two-electron transfer mechanism in water and organic solvent and obtain good results. Additionally, we tested our model on a couple of published DFT calculation datasets of typical organic redox-active molecules as quinone and phenazine derivatives. The performance of the trained model is excellent on the computational redox potential datasets. Based on the results, we discuss the efficient training dataset construction and molecule design for the development of high-performance organic material in ORFB.

2. Results and discussion

2.1. Data curation and machine learning model development

In this work, we build a comprehensive database of experimental redox potential data from hundreds of published papers and perform ML model training and test with these data. All the redox potential values were converted with Standard Hydrogen Electrode (SHE) as a reference. For each entry, the pH value in experiment was also included (for aqueous ORFBs) since the redox potential of organic redox-active molecules is dependent on the pH value. The chemical identifiers of these molecules like Simplified Molecular Input Line Entry System (SMILES) string and International Chemical Identifier Key (InChiKey) and reference are provided. Using SMILES string, many other molecular descriptors can be calculated with cheminformatics package as RDkit. For some molecules, multiple different redox potentials values were found. We selected the most reliable one for the user's benefit. Additionally, two computational datasets are included in our tests. The datasets details are listed in Table 1. Dataset 1 is a curated experimental dataset

Table 1
Summary of the datasets in this work.

Index	Dataset type	Number of data points	Number of core structures	Solvent type	pH
1–1	Experimental data	99	12	Water	<7
1–2	Experimental data	96	9	Water	>7
1–3	Experimental data	336	19	Organic solvent	None
2	Computational data	178	1	Organic solvent	None
3	Computational data	408	17	No solvent	None

from hundreds of organic redox flow batteries papers. To conduct a systematic investigation, the experiment database is split into three subsets. Dataset 1-1 contains molecules in acidic solution for aqueous ORFB. While Dataset 1-2 contains molecules in alkaline solution for aqueous ORFB. And the dataset 1-3 contains data for nonaqueous ORFB. Molecules in neutral solutions of ORFBs are not included in the current database due to insufficient data. Datasets 2 and 3 are computational datasets by DFT calculations from other people's work. Dataset 2 includes phenazine derivatives and dataset 3 consists of quinone derivatives, which are both typical redox-active molecules in ORFBs. Detailed discussions of each system will follow in the subsequent sections.

These datasets are used to train and test ML model. As the fidelity of DFT calculation data is different from experiment data, we train the ML models separately for each datasets following the workflow shown in Fig. 1. After the data curation, the process starts with converting the SMILES string of molecules in the dataset into graphs. In this setting, atoms serve as nodes and bonds serve as edges. Then, we apply graph kernel to determine the average similarity by performing simultaneous random walks on pairs of graphs and generating paths. The details of the graph kernel construction are listed in Section 4.1. To validate the model and optimize the hyperparameters in the model, we apply 5 -fold cross-validation. The details are shown in Sections 4.2 and 4.3.

We calculated the mean absolute error (MAE) and root mean square error (RMSE) values to evaluate the performance of the predictions. The MAE and RMSE are defined as

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \widehat{y}_i| \tag{1}$$

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

where n is the number of molecules, y_i is the ground truth for the redox potential of the molecule, and $\widehat{y}_i$ is the prediction of redox potential obtained by the ML model.

Finally, predictive models are built using graph-based GPR by utilizing both the pairwise similarity matrix among the training molecules and the cross-similarity matrix between the new molecule and the training molecules.

2.2. Experimental data test

To evaluate the performance of our model for experimental data, we built a database with ~500 data points of organic redox-active molecules including quinone, phenazine and many other organic redox-active species derivatives in water or organic solvent. The datasets are included in the supplementary data. To conduct systematic tests, the data were split into three subsets according to the type of solvent and pH value in the solution. Using a data separation strategy tailored to the specific applications of redox-active molecules in different ORFBs allows us to exclude the effect of pH value and decrease the dimensionality of input parameters in the GPR model.

2.2.1. Organic redox-active molecules in aqueous ORFB by GPR model with graph kernel

Estimating redox potentials with chemical accuracy, often required to select the best candidate molecule, is a nontrivial task. The redox potential of the species is significantly influenced by the substituents on the core structure. Therefore, the redox potential can be manipulated using proper substitution. We evaluate the performance of the GPR models in acidic and alkaline ORFBs separately. Using our experimental data sets, we identified 99 redox potential data points of organic redox-active species in acidic ORFB from the literature. Most of them are obtained at around pH = 0. According to the Pourbaix diagram of organic redox-active molecules in acidic ORFB, the redox potential value is not sensitive with small change of pH value in this small region around pH = 0. Therefore, the pH value is not used as an input parameter in the ML model training and test process. From the 99 data points gathered, 80

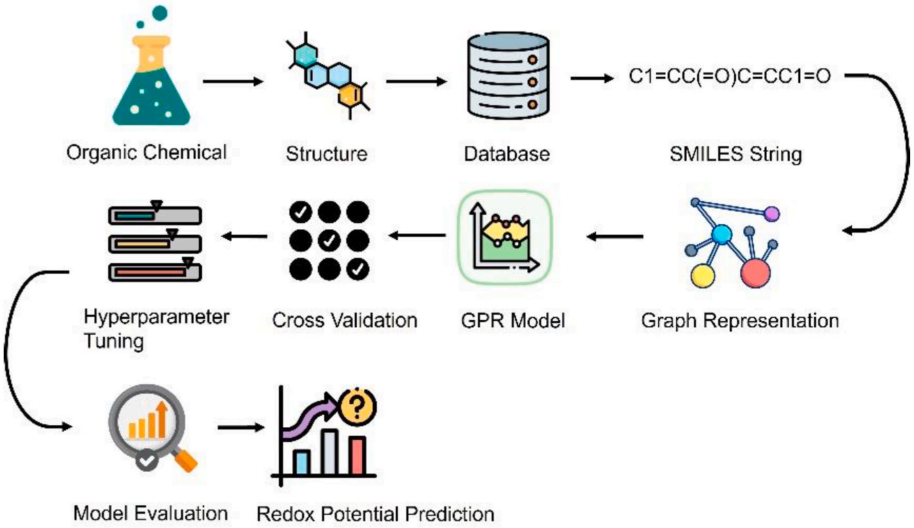

Fig. 1. Representation of the multistep ML workflow for redox potential prediction of organic redox-active molecules in ORFB.

were designated for the training set and 19 for the test set. Fig. 2a and b illustrates the model's exceptional performance on this small dataset of acidic ORFB. The MAE of the training set and test set are 0.038 V and 0.053 V, respectively. The RMSE of the training and the set are 0.063 V and 0.071 V. Additionally, the correlation efficient R^2 are 0.957 and 0.953, indicating a good correlation between the ground truth and prediction. Note that the experimental data is more complex than computational data, as it involves multiple core structures, and a single core structure may encompass various types of functional groups. The measurement noise in experimental data is also significantly greater than that in computational data. Our model is still able to achieve excellent performance despite having such a limited amount of data.

For most of the data obtained from alkaline solution, i.e., pH > 7, the range of pH value is 12–14. Like the above dataset, in this range, the redox potential value is almost independent with the pH value. Therefore, we did not include pH value in our model input. Out of the 96 data points, 80 were allocated to the training set and 16 to the test set. As the dataset includes various core structures, the complexity of this dataset is obviously higher than the computational dataset by DFT calculation. Fig. 2c and d displayed the prediction results. The MAE values for the training and test sets are 0.024 V and 0.085 V, respectively, with corresponding RMSE values of 0.041 V and 0.098 V. This confirms the outstanding performance of our model on a realistic dataset. The $R^2 = 0.961$ and 0.897 for the training and test data set, respectively. The MAE and RMSE are a bit higher than that of acidic ORFB. To further improve it, we tried another method, which is discussed in the next section.

2.2.2. Organic redox-active molecules dataset in alkaline ORFB by GPR model with hybrid kernel

To further improve the prediction of the GPR model in alkaline ORFB, we try a hybrid kernel to introduce additional information of physicochemical property. The graph kernel only includes structural information of the molecule. While the redox potential is a type of physicochemical property. Previous papers showed that some quantum chemistry descriptors, such as the frontier molecular orbital energies, electron affinities and ionization energies have strong correlation with redox potential [45–48]. In theory, according to Koopman's theorem, the oxidation/reduction potential is proportional to the highest occupied molecular orbital (HOMO)/lowest unoccupied molecular orbital (LUMO) energy if the solvation free energy and thermodynamic correction of free energy are ignored, and the vertical process approximation is applied. These are simple descriptors that can be obtained from DFT calculation as they only require computing the frontier orbital energies in the neutral compounds in either the gas phase or in solution. We try integrating the molecular orbital energy into the training data with the hybrid kernel including the graph kernel and the radial basis function (RBF) kernel. While more sophisticated kernels could be explored, here we just start with a basic one to determine whether the additional information can improve the model performance. The LUMO energies descriptor is calculated by DFT and semiempirical methods. We try the semiempirical method because previous paper reported that a good correlation has been found between computed frontier molecular orbital energies and electron affinities or ionization energies [20]. Fig. 3a presents the relationship between the DFT/xTB LUMO energy calculation data and experimental redox potential data. Generally, the LUMO energy data exhibit some correlation with the redox potential data, but some data points exhibit large fluctuations. Although the absolute LUMO energy values are different with the two functionals ωB97X-D3 and B3LYP-D3, the R^2 values are very close (0.404 vs 0.442). That indicates is the correlation is not sensitive to the selection of the functional in DFT calculation. The R^2 between xTB calculation and experimental data is 0.419. Additionally, the xTB calculation results show greater fluctuations. In the following machine learning model training, we only used the B3LYP-D3 calculation result.

As shown in Fig. 3b and c, we found that incorporating LUMO energy through another kernel has almost no effect on the MAE values (0.085 V and 0.086 V), while it improves the RMSE slightly. The RMSE value decreases by 0.01 V with the RBF kernel and LUMO energy descriptor calculated by DFT. Additionally, we found that the prediction result is a bit sensitive to the data quality with the RBF kernel. The MAE with the

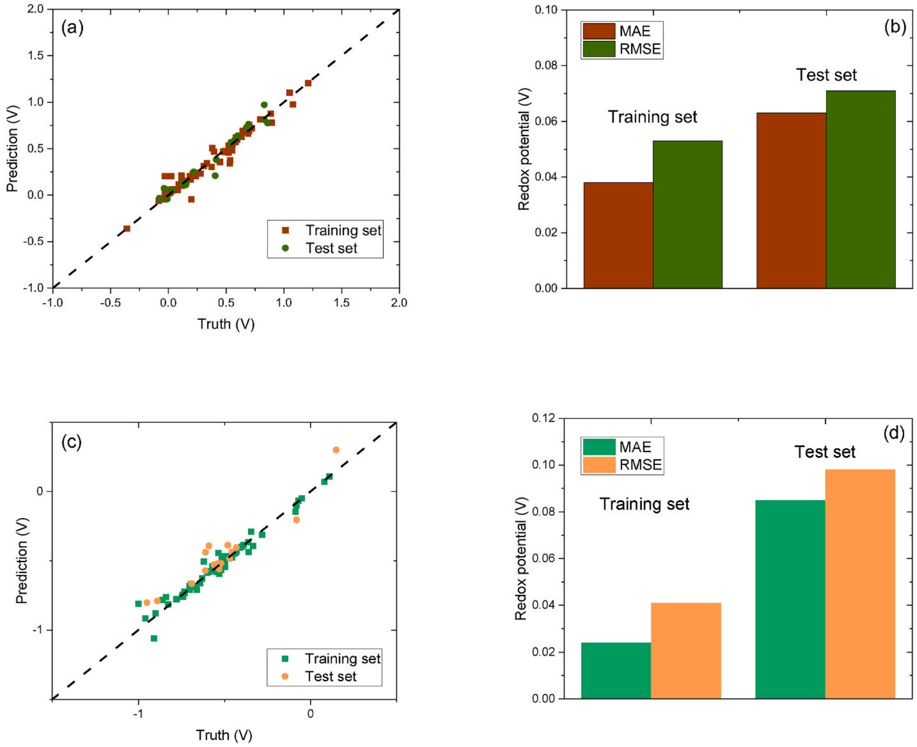

Fig. 2. (a) Parity plots of training data and test data and (b) MAE and RMSE of training set and test set in the experimental dataset of organic redox-active molecules in acidic ORFB. (c) Parity plots of training data and test data and (d) MAE and RMSE of training data and test data in alkaline ORFB by GPR model with graph kernel.

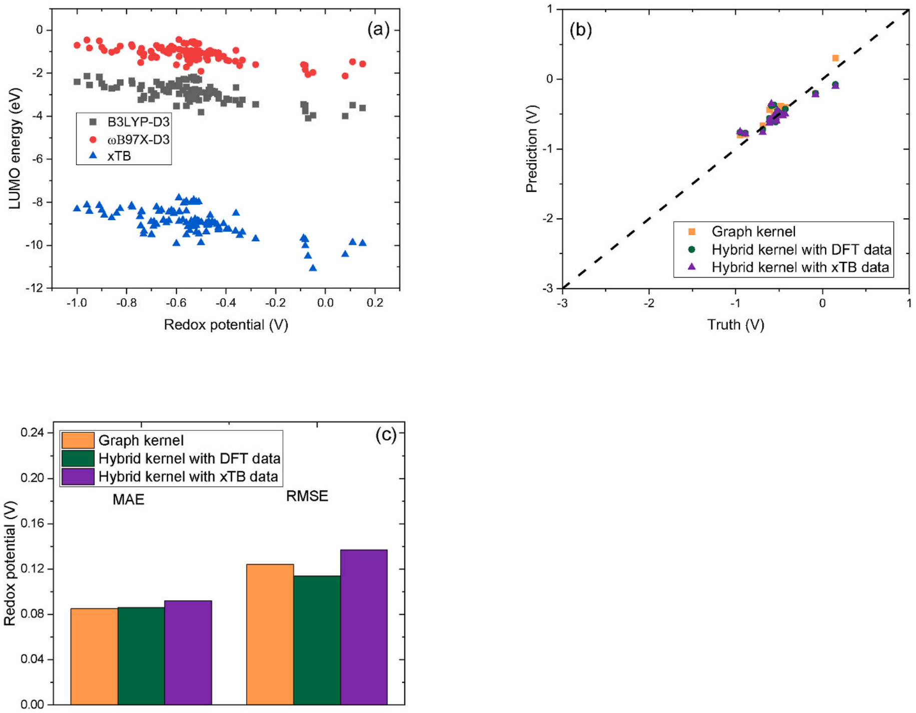

Fig. 3. (a) The correlation between experimental redox potential and LUMO energy in calculations of the organic redox-active molecules. (b) Parity plots of the test data and (c) MAE and RMSE of the test data in the dataset of organic redox-active molecules in water (alkaline solution) by GPR model with hybrid kernel.

LUMO energy calculated by DFT approach is 0.086 V. While the MAE with the LUMO energy calculated by xTB is a bit larger, i.e., 0.092 V. Also, the RMSE values increases by 0.023 V. This is consistent with the data quality of the LUMO energies. Although the xTB calculation is much faster than the DFT calculation, the low-quality data does not work well in the hybrid kernel. On the other hand, the high-quality data with higher correlation of the target property might further improve the prediction. By incorporating additional corrections such as renormalization energy and solvation effects, the descriptor's accuracy could be enhanced [49] and the correlation may be stronger. That will be explored in our future work.

2.2.3. Organic redox-active molecules dataset with one-electron transfer in organic solvent by GPR model with graph kernel

In this section, we test our model on organic redox-active molecule in nonaqueous ORFB. Prince et al. reported the redox potential of the two half quinone derivative couples, $Q/Q^{\cdot -}$ and $Q^{\cdot -}/Q^{=}$, of 11 parent quinones and 118 substituted 1,4-benzoquinones, 91 1,4-naphthoquinones, and 107 9,10-anthraquinones in dimethylformamide (DMF) solvent [50]. Quinones participate in two-electron, two-proton reactions in ORFB, but the proton transfer and electron transfer may be not synchronized. Proton transfer often lags behind the electron transfers. The reaction mechanism is very complicated. Understanding the redox potential change for various quinones in one-electron reaction is beneficial for experimentalists seeking to comprehend the reaction mechanisms of various quinones, which is crucial for optimizing reaction conditions and guiding molecular design. As the number of $Q^{\cdot -}/Q^{=}$ data is limited, we only test our model on the redox potential of quinones with $Q/Q^{\cdot -}$ one-electron transfer. Some molecules with large functional groups such as decyl group were not included in the dataset because the graph generation of these molecules is slow. Finally, we selected 336 data points. These data were divided into a training set of 300 samples and a test set of 36 samples. Fig. 4 shows the prediction results. This dataset includes a greater variety of core structures and functional group types, and the number of molecules containing multiple functional groups is noticeably higher than the other datasets. Despite these challenges, the prediction of the GPR model still achieves good performance. The MAE for the training set and test set are 0.073 V and 0.118 V, respectively, which are both less than 0.2 V. The RMSE for the training set is 0.111 V and for test set is 0.189 V. The R^2 values for the training and test datasets

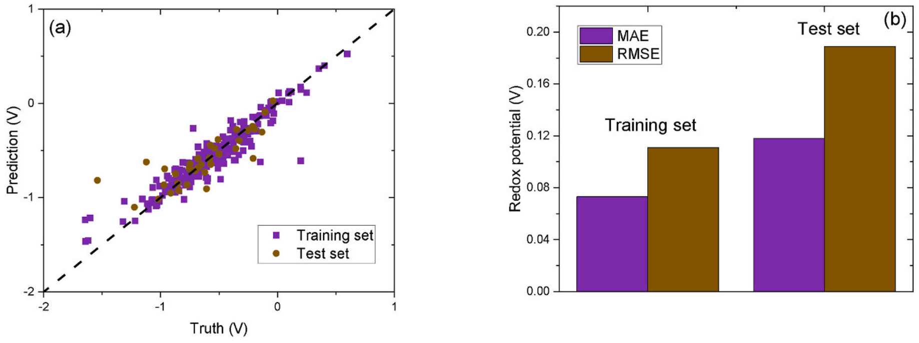

Fig. 4. (a) Parity plots of training data and test data and (b) MAE and RMSE of training data and test data in the dataset of quinones with one-electron transfer in DMF solvent by GPR model with graph kernel.

are 0.886 and 0.683, respectively. It is also observed that a few points had significant deviations, indicating that more data are necessary for the high diversity dataset.

2.3. Computational dataset test

2.3.1. Phenazine derivatives

We also evaluate our GPR approach on two computational datasets. The first is the redox potential data of phenazine derivatives in aprotic solvent (Dataset 2). It has been verified that a phenazine-based organic redox-active material compound showed an exceptionally high reversible capacity that exceeds 90 % of its theoretical value in ORFB [51]. Phenazines comprise a large group of redox-active nitrogen-containing heterocyclic anthracene skeletons, sharing similar chemical, electrochemical and physical properties to those of the quinone family, as shown in Fig. 5a. de la Cruz et al. investigated the redox potentials of ~200 phenazine derivatives in dimethoxyethane (DME) solvent with DFT calculation [52]. There are 20 types of electron-donating or -withdrawing groups at different positions. Note that most of the molecules in this dataset are with one substituent. Only a minor fraction of the molecules has all hydrogens replaced. They identified promising candidates for both the negative and positive sides of organic-based flow batteries. By adding an appropriate number of functional groups at the specific targeted positions, the redox potentials can be modified up to −0.65 V (for the electron-donating amino groups) and to +2.25 V (for the electron-withdrawing cyano groups) compared to the parent phenazine. Also, introducing electron-donating groups at appropriate positions through partial functionalization can lead to a redox potential as negative as or more negative than that achieved with full functionalization. Totally, 189 DFT calculation data were generated in their work. In this work, we retained 178 data after removing duplicate and the molecules with the functional group that only appear once. The whole data set was shuffled and split randomly into a training-set and test-set in an 8:2 ratio (150 samples in the training set and 28 samples in the test set).

Fig. 5c and d shows the result of the ML model. The R^2 between the ground truth and the prediction for the training set is 0.997. The R^2 of the test set is 0.989. Despite the small dataset, the prediction result is excellent. As presented in Fig. 5c and d, the MAE for the training data set and test dataset are 0.025 V and 0.047 V, respectively. The RMSE for the training data set and test dataset are 0.033 V and 0.073 V. This implies that for DFT calculation data, the task of predicting redox potential for the same core structure with one substituted functional group is not difficult for our GPR model. On the other hand, the potential energy surface in the parameter space by DFT calculation might be smoother than experiment. This is similar to other DFT datasets of electrolyte additives in previous work on Li ion batteries [53,54]. Ghule et al. also showed that the performance of several ML models trained on phenazine derivatives with a single type of functional groups is very good at predictions on the training set ($R^2 \geq 0.98$). However, the model's accuracy in predicting the redox potentials of derivatives containing multiple and varied functional groups is lower ($R^2 = 0.74$–0.89) [55]. Our model has shown a high degree of accuracy in predicting data derived from DFT calculations. By comparing the original dataset with Ghule's new dataset, it is found that some molecules with new functional groups were added into the test set in Ghule's work. That would be one of the reasons that affect the prediction of their models. Given that the diversity of functional groups is much greater than that of the core structure, we recommend including a few samples with the same functional groups when building the training dataset to enhance the prediction accuracy of a new test set. This may significantly improve the results. We also generated two-dimensional embeddings with Uniform Manifold Approximation and Projection package (UMAP) [56]using the distance in the graph kernel feature space is used as input. As shown in Fig. 5b, there are some clusters. It is reasonable because all the molecules feature an identical core structure, and the functional groups can be categorized. It also indicates that the distances between some molecules are very close. Therefore, we can expect good prediction with a small dataset.

2.3.2. Quinone derivatives

We additionally tested another DFT calculation dataset of quinones derivatives with a constrained range of target property, referred to as dataset 3. As shown in Fig. 6, this dataset includes 17 quinone core

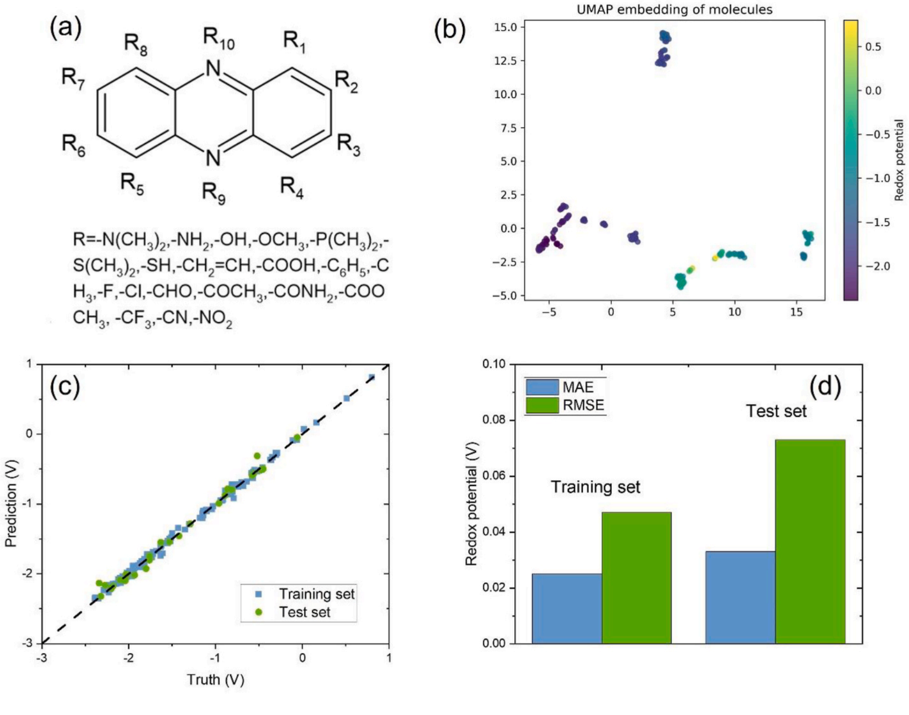

Fig. 5. (a) Chemical structures of phenazine core structure with functional groups in DFT dataset (b) UMAP projection of the 181 data in the DFT calculation dataset. (c) Parity plots of training data and test data and (d) MAE and RMSE of training data and test data in the dataset of phenazine derivatives in DME solvent.

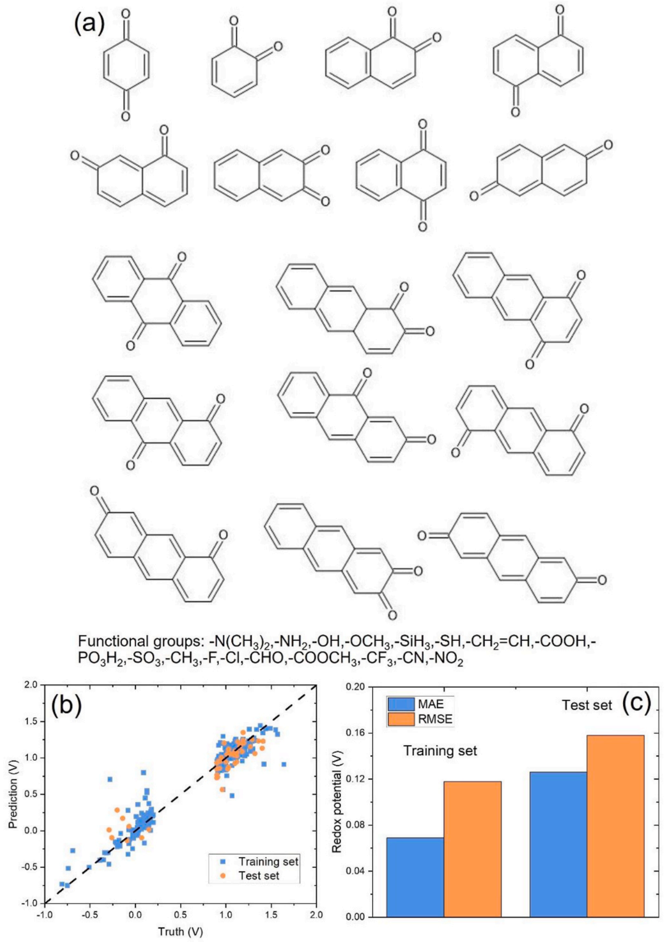

Fig. 6. (a) Chemical structures of quinone core structures with functional groups in dataset 3. (b) Parity plots of training data and test data and (c) MAE and RMSE of training data and test data in the dataset of quinone and quinone derivatives in dataset 3.

structures with 18 functional groups [57]. This dataset only included the molecules where one substituent or all hydrogens are replaced. Note that the authors claimed that totally 1710 molecules were calculated, but only the molecules with redox potential range <0.2 V or >0.9 V in the DFT calculation data were listed in the paper resulting in 409 data points. In Fig. 6c, we can see that the MAE of the training set and test set are 0.069 V and 0.118 V, respectively. The RMSEs are 0.126 V and 0.158 V, respectively. The R^2 values are 0.940 for training set and 0.892 for test set. Even with some limitations in the dataset, the performance of our model remains strong. These results demonstrate that the prediction is still accurate, though not as accurate as the phenazine derivatives case. One reason is due to having multiple core structures in this dataset. Although the amount of data is greater than the previous dataset, there are 17 core structures in this dataset and only one core structure in previous dataset, which increases the complexity of the prediction task. Another reason could be the absence of samples in the chemical parameter space, resulting from the constraints imposed by the target property range. As a filter was applied to the range of the target property in the original work, it greatly decreases the overall number of samples (1710–409). On the other hand, the result also indicates that reducing the range of the target property on a small dataset does not contribute to better model prediction. Ghule et al. mentioned when the phenazine derivatives contain more than one functional group that differ in their type, the redox potential shift is determined by the group showing the highest absolute shift in the corresponding single functional group derivative [55]. The ML results suggest that additional training data might be necessary for the ML model to handle chemicals with two or more functional groups.

3. Conclusion

Incorporating computational strategies into flow battery research represents a significant step forward in commercializing large-scale energy storage technologies. The computational strategies, including high-throughput screening and ML, has dramatically improved material performance and revolutionized the field of ORFBs. The shortage of experimental data is a significant challenge for ML model development in this field. In this work, we develop graph-based GPR models with GPU acceleration to predict redox potentials of organic redox-active molecules for ORFB, which targets the ability to predict using small datasets for training. Also, we built a comprehensive experimental redox potential database with more than 500 data points from published organic redox flow batteries papers. It consists of multiple typical core structures and their derivatives such as quinones and phenazines. Some key experimental parameters such as the pH condition and the type of solvent are also included. To the best of our knowledge, this is currently the largest experimental redox potential database for ORFBs. Also, a couple of computational datasets for typical organic redox-active molecules in ORFB were tested by our model.

We found that our GPR model with the marginalized graph kernel can predict redox potential at high accuracy (MAE 0.03–0.09 V) for experimental aqueous ORFB data. By employing a data separation strategy based on the intended use of redox-active molecules in various ORFBs, the dimensionality of input parameters can be reduced for the GPR model. The GPR model with hybrid kernel, i.e., chemical structure information + physical properties, has the potential further improve the redox potential prediction. Further investigation is still needed into the type of physical descriptor and the form of kernel function. For a more complicated dataset of the quinones in organic solvent with one-electron transfer case, the MAE is 0.07–0.12 V. This indicates that our GPR model can handle the uncertainty that was generated by solvent and/or conformation change. The performance of our GPR model is excellent on DFT calculations datasets with single phenazine core structure (MAE 0.02–0.04 V). It is also effective on a DFT calculation dataset for quinones (MAE 0.06–0.12V), even when the input data is not continuous in the parameter space. These results demonstrate the current models perform well for some core structures and can be extended to new core structure with a small amount of data. The GPR models are simple and fast to be deployed and produce predictions with low errors. Therefore, these models can be considered as efficient alternatives to experiment for fast screening and inverse design of organic material in ORFB. They can also be applied to a wide variety of molecules that can be easily described with structural features. This study provides useful tools for molecule screening and design in ORFB and some insights on designing training datasets from costly experiments.

4. Method

4.1. GPR method with graph kernel and hybrid kernel

A GPR model is a nonparametric Bayesian approach utilized for regression tasks. This model provides a robust framework for determining the probability distribution of parameters over an extensive range of functions that accurately fit the observed data. Within a GPR model, the prior is defined directly on the function space, representing the relationship between input and output variables. The Gaussian process prior itself is a multivariate normal distribution, where the mean function is derived from the data, and the covariance structure is dictated by a specified kernel function. This kernel function encodes assumptions about the function's smoothness and other properties, enabling the GPR model to capture complex patterns and deliver probabilistic predictions with associated uncertainty. In this study, we employ a graph representation for each molecule in the dataset and construct a GPR model using pairwise similarity matrix between the molecules to match for the corresponding target values. The graph kernel is the same as we used in our previous work [58]. In order to evaluate the similarity, we convert the molecules into graphs in which atoms are assigned to the nodes and atomic distances are represented by edges. After that, we apply the marginalized graph kernel to compute the average similarity across all paths generated by random walks on every pair of graphs. In this context, we represent a molecule of n atoms as an undirected and unweighted graph $G = \{V = \{v_i\}, E = \{e_{ij}\}, i,j \in \{1, 2, \cdots, n\}\}$. Each atom i is denoted by a node v_i, characterized by a feature vector $\phi(v_i)$ that contains information about chemical elements, such as charge, aromaticity, and hydrogen count. The marginalized graph kernel $K(G,G')$ which quantifies the expectation of path similarity in the simultaneous random walk can be formulated as

$$\begin{aligned} K(G,G') = \sum_{l=1}^{\infty}\sum_{h}\sum_{h}&\left(p_s(h_1)\prod_{i=2}^{l}p_t(h_1|h_{i-1})p_q(h_l)\right)\\ &\times\left(p'_s(h'_1)\prod_{j=2}^{l}p'_t\left(h'_j\middle|h'_{j-1}\right)p'_q(h'_l)\right)\\ &\times K_v\left(v_{h_1}v'_{h'_1}\right)\prod_{k=2}^{l}K_v\left(v_{h_1}v'_{h'_1}\right)K_e\left(e_{h_{k-1}h_k}, e'_{h'_{k-1}h'_k}\right) \end{aligned} \tag{3}$$

In this equation, l denotes the length of the path, h and h' are paths on the graphs represented by length-l vectors of node labels, p_s is the starting probability of the random walk on each node, p_q is the stopping probability of the random walk on each node at any given step, p_t is the transition probability between a pair of nodes, K_v is an elementary kernel that calculates the similarity between two nodes, while K_e is another elementary kernel that calculates the similarity between a pair of edges, i.e., bonds. Leveraging the dynamic programming, Equation (3) can be reformulated as a linear system.

$$K(G,G^{'}) = \sum_{h_1\in V, h_1^{'}\in V^{'}} p_s(h_1)p_s^{'}(h_1^{'})K_v(h_1,h_1^{'})R_\infty(h_1,h_1^{'}), \tag{4}$$

where R_∞ is the solution of a linear system with a Kronecker product structure, which is

$$r_\infty = \mathrm{q}\otimes\mathrm{q}^{'} + \left[(\mathrm{P}\otimes\mathrm{P}^{'})\odot\left(\mathrm{E}\otimes^{K_e}\mathrm{E}^{'}\right)\right]\bullet diag\left(\mathrm{v}\otimes^{K_v}\mathrm{v}^{'}\right)\bullet r_\infty. \tag{5}$$

The notation is as follows:
v: node label vector of G with $\mathrm{v}_i = v_i$,
p: the starting probability vector of G with $\mathrm{p}_i = p_s(v_i)$,
q: the stopping probability vector of G with $\mathrm{q}_i = p_q(v_i)$,
P: the transition probability matrix of G,
E: the edge label matrix of G with $\mathrm{E}_{ij} = e_{ij}$,
$\otimes^{K_v}$: kernelized Kronecker product between v and v′ with respect to K_v,
$\otimes^{K_e}$: kernelized Kronecker product between E and E′ with respect to K_v, and
v′, p′, q′ and E′: the vectors and matrices that correspond to G′.

Now, we can adapt this formulation in our model to perform the GPR prediction. Given a training set D containing m molecules, i.e., m graphs $\{(G_1, G_2, \cdots, G_m)\}$, a quantity of interest $\{(E_1, E_2, \cdots, E_m)\}$ and a marginalized graph kernel, the GPR prediction for a test set containing n molecules can be formulated as

$$\mathrm{E}^*: = \left[E_1^*, E_2^*, \cdots, E_n^*\right]^T = \mathrm{K}_{\mathrm{D}^*}^{\mathrm{T}}\mathrm{K}_{\mathrm{DD}}^{-1}y_D, \tag{6}$$

where $\mathrm{y}_D = (E_1, E_2, \cdots, E_m)$ stands for the column vector containing the property of each molecule in the training set. Here, the predictive uncertainty of GPR can be represented by the posterior matrix.

$$\Sigma^*: = \mathrm{K}_{**} - \mathrm{K}_{\mathrm{D}^*}^{\mathrm{T}}\mathrm{K}_{\mathrm{DD}}^{-1}K_{D^*}, \tag{7}$$

where K_{DD} is an $m \times m$ covariance matrix of the training set, K_{**} is an $n \times n$ covariance matrix of the test set, and K_D is $m \times n$ training-test cross covariance matrix. The graph kernel was implemented by the graph dot package [59,60].

To construct the hybrid kernel, we multiplied the graph kernel with the radial basis function (RBF) kernel. The RBF kernel is defined as

$$K(x,x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right) \tag{8}$$

And it was implemented by scikit-learn package.

4.2. Hyperparameter optimization and cross-validation

For the GPR model with graph kernel, we employ a standard cross-validation approach to identify the hyperparameters in the ML model, ensuring consistent model selection. This approach is applied uniformly across all our data sets for GPR model with single kernel. Specifically, each data set is split into training-validation and testing parts as described in the following section. To provide a reliable representation of the test data given the limited number of measurements, we use 5-fold cross-validation (CV). The molecules in the training-validation set of each data set are divided into five subsets based on the sequence of their International Chemical Identifier Key (InChIKey). Each subset is iteratively used as a validation set, while the training set consists of the remaining four subsets. Consequently, a 5-fold CV task involves five independent training and validation runs, where the training and validation sets have relative sizes of 80 % and 20 %, respectively. We use the Scikit-Learn library to implement the CV task and conduct an extensive grid search for hyperparameter tuning. The optimal hyperparameter set is determined by the configuration that results in the minimum averaged MAE across the 5-fold CV. There are seven hyperparameters in the graph kernel, which are listed in Table 2. In the grid search, five values (0.1, 0.3, 0.5, 0.7, 0.9) were selected for the hyperparameters aromatic, charge, element, hcount, and edge order. If a particular optimized value is 0.1 or 0.9, the ranges (0, 0.1) or (0.9, 1) will be further explored with a step size of 0.01. The learning rates were set to 0.01, 0.1 and 0.5 and lmin = 0 or 1. For more information about the meaning of the hyperparameters, we recommend reviewing the manual of graphdot on GitHub (https://github.com/yhtang/GraphDot). After determining the best hyperparameters, the models are refitted using the full training dataset and then used to calculate the performance on the test set. The optimized hyperparameter sets for each dataset are provided in Table 3. We can see that the hyperparameters vary across different datasets.

For the GPR model with hybrid kernel, we used Genetic Algorithms (GA) to search the best hyperparameter set. The initial range of hyperparameters in the graph kernel was set to ±0.2, which are based the optimized results by the single kernel GPR model. The range of the other two parameters in the RBF kernel were set to 0.001 to 10. The GA function was implemented by scikit-opt package.

4.3. Descriptor calculation by DFT and semiempirical methods

The LUMO energy descriptor was obtained by DFT and semiempirical calculations. DFT calculations were performed using the

Table 2
Hyperparameters in graph kernel.

Name	Range
aromatic	0–1
charge	0–1
element	0–1
hcount	0–1
edge order	0–1
learning rate	0.01–0.5
lmin	0 or 1

Table 3
Optimized hyperparameter sets for each dataset.

Index	Optimized hyperparameter set
2	(0.1, 0.5, 0.9, 0.9, 0.3, 0.01, 0)
3	(0.9, 0.9, 0.9, 0.5, 0.3, 0.1, 0)
1-2	(0.5, 0.7, 0.9, 0.9, 0.3, 0.1, 0)
1-1	(0.1, 0.1, 0.99, 0.99, 0.5, 0.5, 0)
1-3	(0.9, 0.5, 0.9, 0.9, 0.3, 0.5, 0)

B3LYP(-D3) and ωB97X-D3 functionals with ORCA package [61,62]. The geometries of the molecules were optimized with def2-SVP basis set (ma-def2-SVP for anion). Vibrational frequencies were calculated for validation of stable configuration. An effect of implicit solvent model with water was included via Conductor-like Polarizable Continuum Model (CPCM) model [63]. Semiempirical calculations were perform with xTB with GFN2 method [64]. Vibrational frequencies were also computed to ensure the stable configuration. An analytical linearized Poisson–Boltzmann (ALPB) implicit solvation model was used to mimic the solvation of water [65].

CRediT authorship contribution statement

Peiyuan Gao: Writing – original draft, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Didem Kochan:** Writing – original draft, Software, Investigation, Formal analysis. **Yu-Hang Tang:** Writing – review & editing, Software, Methodology. **Xiu Yang:** Writing – review & editing, Software, Methodology. **Emily G. Saldanha:** Writing – review & editing, Resources, Project administration, Funding acquisition.

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 work was supported by the Energy Storage Materials Initiative (ESMI), which is a Laboratory Directed Research and Development Project at Pacific Northwest National Laboratory (PNNL). PNNL is a multiprogram national laboratory operated for the U.S. Department of Energy (DOE) by Battelle Memorial Institute under Contract no. DE-AC05-76RL01830.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jpowsour.2024.236035.

Data availability

I have shared the data in supplementary materials.

References

[1] Y.K. Zeng, T.S. Zhao, L. An, X.L. Zhou, L. Wei, A comparative study of all-vanadium and iron-chromium redox flow batteries for large-scale energy storage, J. Power Sources 300 (2015) 438–443, https://doi.org/10.1016/j.jpowsour.2015.09.100.

[2] C. Zhang, Z. Yuan, X. Li, Designing better flow batteries: an overview on fifty years' research, ACS Energy Lett. 9 (2024) 3456–3473, https://doi.org/10.1021/acsenergylett.4c00773.

[3] A. Dinesh, et al., Iron-based flow batteries to store renewable energies, Environ. Chem. Lett. 16 (2018) 683–694, https://doi.org/10.1007/s10311-018-0709-8.

[4] F.L. Zhu, W. Guo, Y.Z. Fu, Functional materials for aqueous redox flow batteries: merits and applications, Chem. Soc. Rev. 52 (2023) 8410–8446, https://doi.org/10.1039/d3cs00703k.

[5] M. Shoaib, et al., Advances in redox flow batteries - a comprehensive review on inorganic and organic electrolytes and engineering perspectives, Adv. Energy Mater. 14 (2024), https://doi.org/10.1002/aenm.202400721.

[6] J.A. Luo, B. Hu, M.W. Hu, Y. Zhao, T.L. Liu, Status and prospects of organic redox flow batteries toward sustainable energy storage, ACS Energy Lett. 4 (2019) 2220–2240, https://doi.org/10.1021/acsenergylett.9b01332.

[7] R. Feng, et al., Reversible ketone hydrogenation and dehydrogenation for aqueous organic redox flow batteries, Science 372 (2021) 836–840, https://doi.org/10.1126/science.abd9795.

[8] D. Emmel, et al., Benchmarking organic active materials for aqueous redox flow batteries in terms of lifetime and cost, Nat. Commun. 14 (2023) 6672, https://doi.org/10.1038/s41467-023-42450-9.

[9] A.A. Howard, T. Yu, W. Wang, A.M. Tartakovsky, Physics-informed CoKriging model of a redox flow battery, J. Power Sources 542 (2022) 231668, https://doi.org/10.1016/j.jpowsour.2022.231668.

[10] Y. Fu, et al., Physics-guided continual learning for predicting emerging aqueous organic redox flow battery material performance, ACS Energy Lett. 9 (2024) 2767–2774, https://doi.org/10.1021/acsenergylett.4c00493.

[11] W. Chen, Y. Fu, P. Stinis, Physics-informed machine learning of redox flow battery based on a two-dimensional unit cell model, J. Power Sources 584 (2023) 233548, https://doi.org/10.1016/j.jpowsour.2023.233548.

[12] A. Das, Y.V. Grinkova, S.G. Sligar, Redox potential control by drug binding to cytochrome P450 3A4, J. Am. Chem. Soc. 129 (2007) 13778–13779, https://doi.org/10.1021/ja074864x.

[13] E.S. Rountree, B.D. McCarthy, T.T. Eisenhart, J.L. Dempsey, Evaluation of homogeneous electrocatalysts by cyclic voltammetry, Inorg. Chem. 53 (2014) 9983–10002, https://doi.org/10.1021/ic500658x.

[14] T. Borch, et al., Biogeochemical redox processes and their impact on contaminant dynamics, Environ. Sci. Technol. 44 (2010) 15–23, https://doi.org/10.1021/es9026248.

[15] K.P. Castro, et al., Incremental tuning up of fluorous phenazine acceptors, Chem. Eur J. 22 (2016) 3930–3936, https://doi.org/10.1002/chem.201504122.

[16] C. Wang, et al., Molecular design of fused-ring phenazine derivatives for long-cycling alkaline redox flow batteries, ACS Energy Lett. 5 (2020) 411–417, https://doi.org/10.1021/acsenergylett.9b02676.

[17] A. Benayad, et al., High-throughput experimentation and computational freeway lanes for accelerated battery electrolyte and interface development research, Adv. Energy Mater. 12 (2022) 2102678, https://doi.org/10.1002/aenm.202102678.

[18] J. Asenjo-Pascual, et al., DFT calculation, a practical tool to predict the electrochemical behaviour of organic electrolytes in aqueous redox flow batteries, J. Power Sources 564 (2023) 232817, https://doi.org/10.1016/j.jpowsour.2023.232817.

[19] A.J. Achazi, et al., Development of a multi-step screening procedure for redox active molecules in organic radical polymer anodes and as redox flow anolytes, J. Comput. Chem. 45 (2024) 1112–1129, https://doi.org/10.1002/jcc.27299.

[20] M. Korth, Large-scale virtual high-throughput screening for the identification of new battery electrolyte solvents: evaluation of electronic structure theory methods, Phys. Chem. Chem. Phys. 16 (2014) 7919–7926, https://doi.org/10.1039/C4CP00547C.

[21] H. Neugebauer, F. Bohle, M. Bursch, A. Hansen, S. Grimme, Benchmark study of electrochemical redox potentials calculated with semiempirical and DFT methods, J. Phys. Chem. 124 (2020) 7166–7176, https://doi.org/10.1021/acs.jpca.0c05052.

[22] E. Sorkun, Q. Zhang, A. Khetan, M.C. Sorkun, S. Er, RedDB, a computational database of electroactive molecules for aqueous redox flow batteries, Sci. Data 9 (2022) 718, https://doi.org/10.1038/s41597-022-01832-2.

[23] L. Jia, et al., Predicting redox potentials by graph-based machine learning methods, J. Comput. Chem. 45 (2024) 2383–2396, https://doi.org/10.1002/jcc.27380.

[24] A. Hashemi, et al., Density functional theory and machine learning for electrochemical square-scheme prediction: an application to quinone-type molecules relevant to redox flow batteries, Digital Dis. 2 (2023) 1565–1576, https://doi.org/10.1039/D3DD00091E.

[25] A.V. Marenich, C.J. Cramer, D.G. Truhlar, Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions, J. Phys. Chem. B 113 (2009) 6378–6396, https://doi.org/10.1021/jp810292n.

[26] A.A. Voityuk, S.F. Vyboishchikov, Fast and accurate calculation of hydration energies of molecules and ions, Phys. Chem. Chem. Phys. 22 (2020) 14591–14598, https://doi.org/10.1039/D0CP02667K.

[27] S. Maier, B. Thapa, K. Raghavachari, G4 accuracy at DFT cost: unlocking accurate redox potentials for organic molecules using systematic error cancellation, Phys. Chem. Chem. Phys. 22 (2020) 4439–4452, https://doi.org/10.1039/C9CP06622E.

[28] X. Liu, J. Cheng, M. Sprik, Aqueous transition-metal cations as impurities in a wide gap oxide: the Cu2+/Cu+ and Ag2+/Ag+ redox couples revisited, J. Phys. Chem. B 119 (2015) 1152–1163, https://doi.org/10.1021/jp506691h.

[29] E. Hruska, A. Gale, F. Liu, Bridging the experiment-calculation divide: machine learning corrections to redox potential calculations in implicit and explicit solvent models, J. Chem. Theor. Comput. 18 (2022) 1096–1108, https://doi.org/10.1021/acs.jctc.1c01040.

[30] P.P. Fehér, Á. Madarász, A. Stirling, Prediction of redox power for photocatalysts: synergistic combination of DFT and machine learning, J. Chem. Theor. Comput. 19 (2023) 4125–4135, https://doi.org/10.1021/acs.jctc.3c00286.

[31] R. Jinnouchi, F. Karsai, G. Kresse, Machine learning-aided first-principles calculations of redox potentials, npj Comput. Mater. 10 (2024) 107, https://doi.org/10.1038/s41524-024-01295-6.

[32] S.W. Benson, J.H. Buss, Additivity rules for the estimation of molecular properties. Thermodynamic properties, J. Chem. Phys. 29 (1958) 546–572, https://doi.org/10.1063/1.1744539.

[33] M. Kleinová, et al., Antioxidant properties of carotenoids:: QSAR prediction of their redox potentials, Gen. Physiol. Biophys. 26 (2007) 97–103.

[34] A. Jinich, B. Sanchez-Lengeling, H. Ren, R. Harman, A. Aspuru-Guzik, A mixed quantum chemistry/machine learning approach for the fast and accurate prediction of biochemical redox potentials and its large-scale application to 315 000 redox reactions, ACS Cent. Sci. 5 (2019) 1199–1210, https://doi.org/10.1021/acscentsci.9b00297.

[35] K. Nesměrák, A.A. Toropov, in: Alla P. Toropova, Andrey A. Toropov (Eds.), QSPR/QSAR Analysis Using SMILES and Quasi-SMILES, Springer International Publishing, 2023, pp. 139–166.

[36] M.D. Chaka, et al., High-throughput screening of promising redox-active molecules with MolGAT, ACS Omega 8 (2023) 24268–24278, https://doi.org/10.1021/acsomega.3c01295.

[37] P.M. Tagade, et al., Attribute driven inverse materials design using deep learning Bayesian framework, npj Comput. Mater. 5 (2019) 127, https://doi.org/10.1038/s41524-019-0263-3.

[38] T. Nakayama, Y. Igarashi, K. Sodeyama, M. Okada, Material search for Li-ion battery electrolytes through an exhaustive search with a Gaussian process, Chem. Phys. Lett. 731 (2019) 136622, https://doi.org/10.1016/j.cplett.2019.136622.

[39] K. Sodeyama, Y. Igarashi, T. Nakayama, Y. Tateyama, M. Okada, Liquid electrolyte informatics using an exhaustive search with linear regression, Phys. Chem. Chem. Phys. 20 (2018) 22585–22591, https://doi.org/10.1039/C7CP08280K.

[40] S. Manna, S.S. Manna, B. Pathak, Integrated supervised and unsupervised machine learning approach to map the electrochemical windows over 4500 solvents for battery applications, ACS Appl. Mater. Interfaces 16 (2024) 42138–42152, https://doi.org/10.1021/acsami.4c06243.

[41] T. Li, C. Zhang, X. Li, Machine learning for flow batteries: opportunities and challenges, Chem. Sci. 13 (2022) 4740–4752, https://doi.org/10.1039/D2SC00291D.

[42] E.M. Fell, M.J. Aziz, High-throughput electrochemical characterization of aqueous organic redox flow battery active material, J. Electrochem. Soc. 170 (2023), https://doi.org/10.1149/1945-7111/acfcde.

[43] J. Noh, et al., An integrated high-throughput robotic platform and active learning approach for accelerated discovery of optimal electrolyte formulations, Nat. Commun. 15 (2024) 2757, https://doi.org/10.1038/s41467-024-47070-5.

[44] Y. Liang, et al., High-throughput solubility determination for data-driven materials design and discovery in redox flow battery research, Cell Rep. Phys. Sci. 4 (2023), https://doi.org/10.1016/j.xcrp.2023.101633.

[45] R.B. Araujo, et al., Designing strategies to tune reduction potential of organic molecules for sustainable high capacity battery application, J. Mater. Chem. A 5 (2017) 4430–4454, https://doi.org/10.1039/C6TA09760J.

[46] O. Allam, et al., Molecular structure–redox potential relationship for organic electrode materials: density functional theory–Machine learning approach, Mater. Today Energy 17 (2020) 100482, https://doi.org/10.1016/j.mtener.2020.100482.

[47] A. Kuhn, K.G. von Eschwege, J. Conradie, Electrochemical and density functional theory modeled reduction of enolized 1,3-diketones, Electrochim. Acta 56 (2011) 6211–6218, https://doi.org/10.1016/j.electacta.2011.03.083.

[48] C.K. Khor, L.A. Calhoun, J.J. Neville, C.A. Dyker, Experimental and theoretical predictors for redox potentials of bispyridinylidene electron donors, ChemPhysChem 25 (2024) e202400092, https://doi.org/10.1002/cphc.202400092.

[49] V. Pande, V. Viswanathan, Descriptors for electrolyte-renormalized oxidative stability of solvents in lithium-ion batteries, J. Phys. Chem. Lett. 10 (2019) 7031–7036, https://doi.org/10.1021/acs.jpclett.9b02717.

[50] R.C. Prince, P.L. Dutton, M.R. Gunner, The aprotic electrochemistry of quinones, Biochim. Biophys. Acta Bioenerg. 1863 (2022) 148558, https://doi.org/10.1016/j.bbabio.2022.148558.

[51] A. Hollas, et al., A biomimetic high-capacity phenazine-based anolyte for aqueous organic redox flow batteries, Nat. Energy 3 (2018) 508–514, https://doi.org/10.1038/s41560-018-0167-3.

[52] C. de la Cruz, et al., New insights into phenazine-based organic redox flow batteries by using high-throughput DFT modelling, Sustain. Energy Fuels 4 (2020) 5513–5521, https://doi.org/10.1039/D0SE00687D.

[53] Y. Zhang, X. Xu, Machine learning properties of electrolyte additives: a focus on redox potentials, Ind. Eng. Chem. Res. 60 (2021) 343–354, https://doi.org/10.1021/acs.iecr.0c05055.

[54] Y. Okamoto, Y. Kubo, Ab initio calculations of the redox potentials of additives for lithium-ion batteries and their prediction through machine learning, ACS Omega 3 (2018) 7868–7874, https://doi.org/10.1021/acsomega.8b00576.

[55] S. Ghule, S.R. Dash, S. Bagchi, K. Joshi, K. Vanka, Predicting the redox potentials of phenazine derivatives using DFT-assisted machine learning, ACS Omega 7 (2022) 11742–11755, https://doi.org/10.1021/acsomega.1c06856.

[56] L. McInnes, J. Healy, N. Saul, L. Grossberger, UMAP: Uniform Manifold approximation and projection, J. Open Source Softw. 3 (2018) 861.

[57] 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.

[58] P. Gao, et al., Graphical Gaussian process regression model for aqueous solvation free energy prediction of organic molecules in redox flow batteries, Phys. Chem. Chem. Phys. 23 (2021) 24892–24904, https://doi.org/10.1039/D1CP04475C.

[59] Y.-H. Tang, W.A. de Jong, Prediction of atomization energy using graph kernel and active learning, J. Chem. Phys. 150 (2019), https://doi.org/10.1063/1.5078640.

[60] Tang, Y. H., Selvitopi, O., Popovici, D. T. & Buluç, A. in 2020 IEEE International Parallel and Distributed Processing Symposium (IPDPS). 728-738..

[61] F. Neese, F. Wennmohs, U. Becker, C. Riplinger, The ORCA quantum chemistry program package, J. Chem. Phys. 152 (2020), https://doi.org/10.1063/5.0004608.

[62] S. Grimme, J. Antony, S. Ehrlich, H. Krieg, A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu, J. Chem. Phys. 132 (2010) 154104, https://doi.org/10.1063/1.3382344.

[63] V. Barone, M. Cossi, Quantum calculation of molecular energies and energy gradients in solution by a conductor solvent model, J. Phys. Chem. 102 (1998) 1995–2001, https://doi.org/10.1021/jp9716997.

[64] C. Bannwarth, et al., Extended tight-binding quantum chemistry methods, WIREs Computa. Molec. Sci. 11 (2021) e1493, https://doi.org/10.1002/wcms.1493.

[65] S. Ehlert, M. Stahn, S. Spicher, S. Grimme, Robust and efficient implicit solvation model for fast semiempirical methods, J. Chem. Theor. Comput. 17 (2021) 4250–4261, https://doi.org/10.1021/acs.jctc.1c00471.