

# RedPred, a machine learning model for the prediction of redox reaction energies of the aqueous organic electrolytes

Murat Cihan Sorkun<sup>1,2</sup>, Elham Nour Ghassemi<sup>1</sup>, Cihan Yatbaz<sup>1</sup>, J. M. Vianney A. Koelman<sup>1,2</sup>, and Süleyman Er<sup>1\*</sup>

Aqueous Organic Redox Flow Batteries (AORFBs) are considered as one of the most appealing technologies for large-scale energy storage due to their electroactive organic materials, which are abundant, easy to produce, and recyclable. A prevailing challenge for the redox chemistries applied in AORFBs is to achieve high power and energy density. The chemical design and molecular engineering of the electroactive compounds is an effective approach for the optimization of their physicochemical properties. Among them, the reaction energy of redox couples is often used as a proxy for the measured potentials. In this study, we present RedPred, a machine learning (ML) model that predicts the one-step two-electron two-proton redox reaction energy of redox-active molecule pairs. RedPred comprises an ensemble of Artificial Neural Networks, Random Forests, and Graph Convolutional Networks, trained using the RedDB database, which contains over 15,000 reactant-product pairs for AORFBs. We evaluated RedPred's performance using six different molecular encoders and five prominent ML algorithms applied in chemical science. The predictive capability of RedPred was tested on both its training chemical space and the chemical space outside its training domain using two separate test datasets. We released a user-friendly web tool with open-source code to promote software sustainability and broad use.

## INTRODUCTION

Efficiently integrating intermittent renewable energy sources into the electricity grid requires advanced energy storage solutions to balance supply and demand<sup>1</sup>. The promise of design flexibility, high scalability, low maintenance costs, and long cycle-life renders redox flow batteries (RFBs) as a viable short- and medium-term stationary storage technology for the excess electricity generated by renewables<sup>2</sup>. A typical RFB consists of two main components: tanks for storing the negative and positive electrolytes and a fuel cell where redox reactions occur during charging and discharging cycles. Electrolytes are pumped into the cell and collected back into the tanks following the electrochemical conversion. With this unique design that decouples energy and power, RFBs offer high flexibility in chemical design and high scalability for grid-scale energy storage.

The performance of an RFB is strongly dependent on the redox activity of the electrolyte materials. The first generation of RFBs contained inorganic electrolyte materials. For instance, the first successfully commercialized<sup>2</sup> RFBs were vanadium RFBs (VRFBs)<sup>3</sup>, which applied the reversible electrochemistry between different oxidation states of vanadium. However, the

---

<sup>1</sup>DIFFER -- Dutch Institute for Fundamental Energy Research, De Zaale 20, Eindhoven 5612 AJ, the Netherlands.

<sup>2</sup>Center for Computational Energy Research, Department of Applied Physics, Eindhoven University of Technology, Eindhoven 5600 MB, the Netherlands.

\*email: s.er@diffen.nl

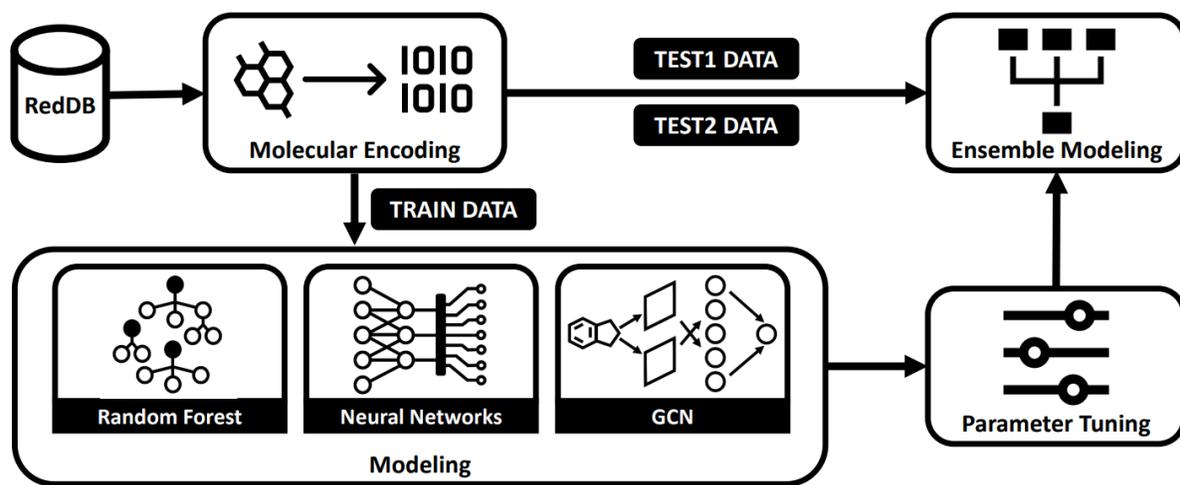
cost of vanadium hampered the large-scale applicability of VRFBs. In addition to economic reasons related to the cost of raw materials, the toxicity and harsh chemical nature of the inorganics drove researchers towards finding alternative low-cost, abundant, and safe redox-active materials.

Aqueous Organic RFBs (AORFBs), based on new organic redox species and innovative designs, have increasingly been researched in recent years. The use of organic electrolytes offers several advantages, including abundance, structural diversity, sustainability, modest manufacturing costs, and recycling opportunities<sup>4</sup>. Molecular engineering has proven to be an effective approach for devising new organic electroactive materials with desirable properties for AORFBs, as it allows for high tunability in the structure and composition of organics. For example, the addition of chemically effective functional groups to existing redox-active molecule structures has been found to be effective in adjusting the battery-relevant properties to desired levels. To identify the candidate materials for AORFBs, high-throughput virtual screening (HTVS) approaches have become increasingly popular<sup>5-11</sup>. These approaches allow for agile explorations of designated chemical spaces, in contrast to the time-intensive trial-and-error methods. Nevertheless, due to the high computational demands of first-principles simulations, the current HTVS efforts are constrained to examining a few thousand candidates.

Recent advancements in artificial intelligence (AI) and machine learning (ML) technologies have led to significant improvements in the discovery of energy-efficient materials by their time efficiency and prediction accuracy<sup>10-17</sup>. This is also promising for the acceleration of RFB materials development by rapid property predictions, extending the search space, or, more ambitiously, the *de novo* design of energy compounds with desired properties.

The long-term operational performance of an RFB is dependent on its energy density and cycling stability<sup>18</sup>. The achievable energy density is often constrained by the water solubility of active species and the electrochemical potentials they can attain<sup>19</sup>. Due to being a key molecular property for various fields of chemistry and the availability of openly accessible datasets<sup>20-22</sup>, ML-based solubility predictions have been increasingly employed<sup>23-27</sup>. In contrast, ML-based predictions of redox potentials for electrolyte materials have not been sufficiently well investigated. Although some recent studies demonstrated promising results for the applicability of ML-based screening of redox-active materials<sup>28,29</sup>, accurate and comprehensive ML models are still needed to accelerate the screening of candidate AORFB electrolyte materials from diverse chemical spaces.

In the present work, we developed an ML model, namely RedPred, that predicts the redox reaction energy which is a highly correlated descriptor with the experimental redox potential of AORFB electrolyte materials<sup>30</sup>. To train our models, we used RedDB<sup>9</sup>, a recently published materials dataset that contains candidate AORFB reactant-product pairs calculated by accurate quantum chemical methods. We evaluated the performance of the proposed model on two reserved test sets where the first set covers the same chemical space of the training set and the second set falls outside the chemical space to assess its extrapolation ability. We compared the performances of the models trained on different ML algorithms and molecular encoders. Based on the new learnings, we developed RedPred which comprises an ensemble of Artificial Neural Networks (ANN), Random Forests (RF), and Graph Convolutional Networks (GCNs). RedPred showed promising results not only within the training chemical space but also outside of it. Importantly, we provided insights about the difficulties of extrapolating outside of the training



**Figure 1.** The workflow diagram of ensemble ML model for predicting the reaction energy of AORFB electrolyte materials.

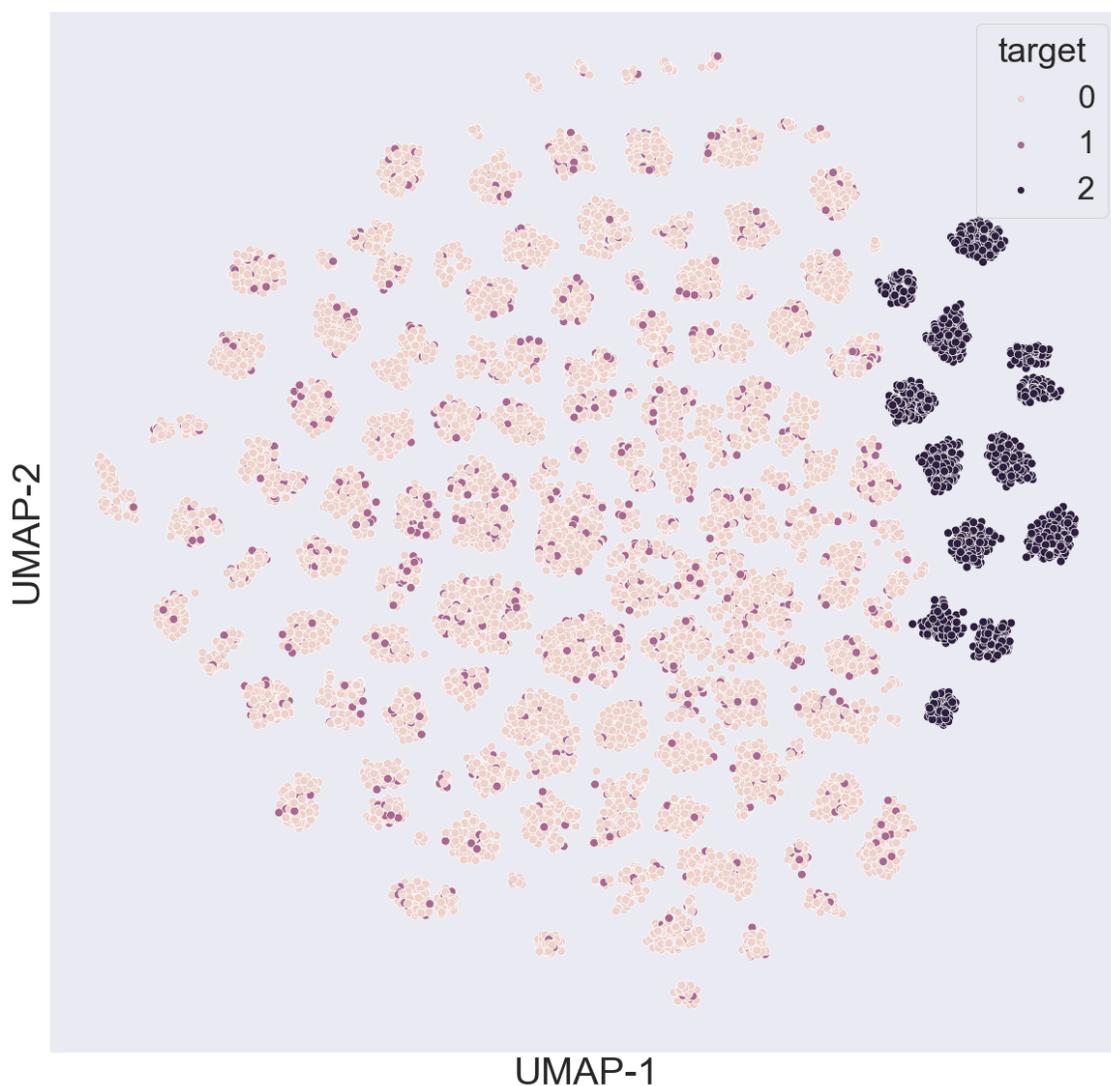
domains. We released RedPred as an online web tool, along with its source code, for public use with the aim of expediting the search of prospective AORFB electrolyte materials as a complementary tool.

## RESULTS

In this section, we explain the processes and results of ML development for predicting the reaction energy of AORFB electrolyte materials. These processes include data preprocessing, selection of molecular encoder, ML model development, and ensemble modeling. Figure 1 illustrates the processes that this study implements as a workflow diagram.

### Data Preprocessing

The data used in this study was collected from RedDB<sup>9</sup> version 1.0, which is a recently published computational database focused on the candidate electroactive compounds for AORFBs. In addition to calculated atomic and molecular data, RedDB contains 15,932 reaction pairs and their calculated reaction energies (see Equation (1)) that are used as the target property in the current study. From the published database, we removed the instances that contain missing reaction energy values and ended up with 15,794 reaction pairs. The chemical space of the data visualized by ChemPlot<sup>31</sup> where each dot corresponds to a single molecule in 2-dimensions is shown in Figure 2. The molecules were placed based on their structural similarities extracted from their SMILES notations and the dimensions were reduced by the UMAP<sup>32</sup> algorithm. Based on the chemical space occupation, we divided the data into three disjunct sets: Training, Test-1, and Test-2. In order to evaluate the performance distinctively on the training- and extrapolated-domain, we selected Test-1 instances from within the chemical space of the training set, while the Test-2 instances reside outside the chemistry space covered by the training set. The training set was used to train and tune the models while the test sets were reserved for the evaluation of ML model performance. As a result, 12,707 (80.5%) instances were reserved for Training, 1,607 (10.2%) for Test-1, and 1,480 (9.3%) for Test-2.



**Figure 2.** The chemical space of the data is visualized by ChemPlot using structural similarity and UMAP options. The colors corresponding to the following numbers refer to the different parts of the dataset; (0):Training, (1):Test-1, and (2):Test-2.

### Selection of Molecular Encoders

In RedDB, the molecules are represented in SMILES notation. In order to train ML models, first, it is required to convert the SMILES representations to latent space variables. To find the best-suited representation for our aim, we compared the available molecular encoders. For this purpose, we employed various molecular encoders which are structural fingerprints (ECFC<sup>33</sup>, ECFP<sup>33</sup>, SECFP<sup>34</sup>, and MACCS), NLP embeddings (Mol2vec<sup>35</sup>), and physico-chemical descriptors (Mordred<sup>36</sup>). To compare the performances of the molecular encoders, we trained separate Light Gradient Boosting Machine (LGBM) models under the same conditions by encoding the training set using the different molecular encoders. The results are shown in Table 1 for each molecular encoder. Among the models, the one encoded by ECFC showed the best performance on both Test-1 and Test-2 datasets. The model encoded by ECFP showed slightly poor performance compared to ECFC on both Test-1 and Test-2. The model encoded by SECFP showed a close performance to the ECFC and ECFP on Test-1 but the performance dropped

on Test-2. Models encoded by Mol2vec and Mordred showed good performance on Test-1 but performed poorly on Test-2. Models encoded by MACCS showed relatively poor performance on both datasets. As a result, we decided to use the ECFC encoder in our experiments. Further details about the molecular encoders are provided in the Methods section.

Molecular Encoder	Training			Test-1			Test-2		
	MAE	RMSE	R <sup>2</sup>	MAE	RMSE	R <sup>2</sup>	MAE	RMSE	R <sup>2</sup>
ECFP	0.004	0.008	0.958	0.005	0.008	0.955	0.009	0.012	0.835
ECFC	0.004	0.007	0.968	0.004	0.008	0.964	0.009	0.011	0.845
SECFP	0.004	0.008	0.962	0.005	0.009	0.953	0.013	0.016	0.694
MACCS	0.010	0.015	0.864	0.010	0.015	0.863	0.013	0.021	0.472
Mol2vec	0.005	0.008	0.964	0.007	0.010	0.933	0.021	0.023	0.326
Mordred	0.006	0.010	0.937	0.007	0.011	0.925	0.015	0.024	0.295

**Table 1.** The performance comparison of the models trained using different molecular encoders in predicting reaction energy (in Hartree).

### Development of Machine Learning Models

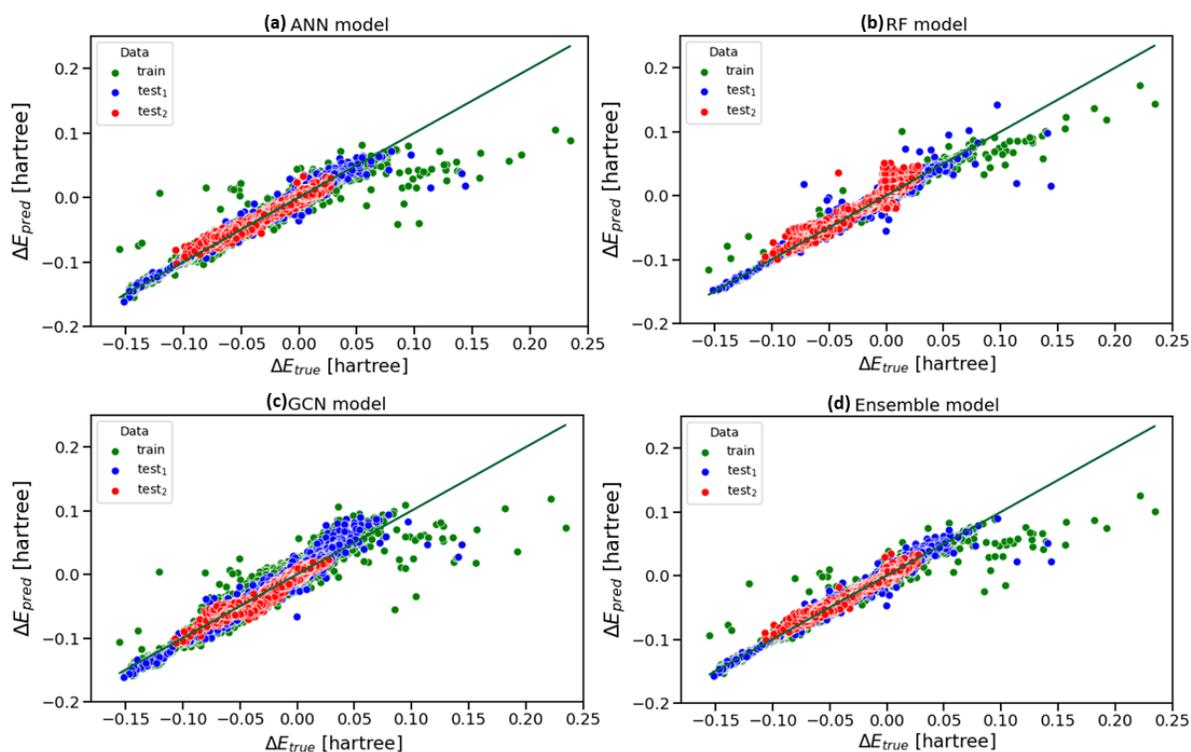
In this step, we developed ML models by implementing state-of-the-art algorithms to compare their performances on predicting reaction energies. We applied five methods including, ANN, RF, GCN, LGBM, and eXtreme Gradient Boosting (XGB). All the models are trained using ECFC encoder except GCN which contains its own integrated encoder. We trained and tuned the parameters of the models using the training set. The best-performing configurations for each method are tested on the reserved test sets. While all models showed excellent performance on Test-1, ANN and GCN models showed significantly better performance on Test-2. Among all the ML models, ANN yielded the top performance on both datasets. Finally, by combining the top-performing three models (ANN, RF, and GCN), we developed a weighted ensemble model (see Equation (2)). The ensemble model improved the performance for both test sets. The test results of singular models and the ensemble model are shown in Table 2 given as MAE, RMSE, and R<sup>2</sup>. Figure 3 shows the true versus the predicted reaction energy values of Training, Test-1, and Test-2 data points for the top-performing models and the ensemble model. Further details of training, tuning, and ensembling the models are provided in the Methods section.

Method	Training			Test-1			Test-2		
	MAE	RMSE	R <sup>2</sup>	MAE	RMSE	R <sup>2</sup>	MAE	RMSE	R <sup>2</sup>
ANN	0.004	0.007	0.968	0.004	0.008	0.964	0.005	0.006	0.951
RF	0.001	0.003	0.993	0.004	0.008	0.964	0.008	0.011	0.848
GCN	0.007	0.010	0.940	0.007	0.010	0.934	0.006	0.008	0.926
LGBM	0.004	0.008	0.961	0.005	0.008	0.959	0.009	0.011	0.850
XGB	0.004	0.006	0.975	0.005	0.008	0.957	0.010	0.012	0.826
Ensemble	-	-	-	0.004	0.007	0.969	0.004	0.006	0.959

**Table 2.** The performance comparison of the models trained using different ML methods and the ensemble model in predicting reaction energy (in Hartree).

### Analysis of Poorly Predicted Molecule Pairs

As shown in Figure 3, some data points far from the diagonal line were predicted extremely poorly by all different ML models. Moreover, most of the erroneous predictions belong to the Training set which was unexpected. Therefore, we investigated



**Figure 3.** The DFT-calculated versus the ML-predicted reaction energy data as obtained by using different ML models: (a) ANN, (b) RF, (c) GCN, and (d) the ensemble model. Green, blue, and red data points represent the Training, Test-1, and Test-2 datasets, respectively. The proximity of the data points to the diagonal line indicates their accuracy.

the reason behind these outlier molecules by analyzing the related molecule pairs. First, we selected 39 molecule pairs that have an error over 0.05 Hartree predicted by the ensemble model. Then, for every selected molecule pair, we searched the most similar pairs from RedDB as described in the Methods section. Using the similar molecule pairs found by the similarity search, we compared their DFT computed reaction energies. We calculated the deviation of the reaction energies of the selected pair and the similar molecule pairs from their mean reaction energy. As a result of the comparison, we found that the reaction energies of selected pairs were much different than their similar pairs while the reaction energy values of similar pairs used for comparison were close to each other. This result pointed out a possible error that is coming from the DFT computations for the selected molecule pairs. We further investigated these molecule pairs by inspecting their 3D geometries. We confirmed that 25 molecules from the inspected pairs had significant structural distortions in their DFT-optimized geometries, which is a sign of chemical instability. Therefore, these pairs of molecules have been removed from RedDB in its next version.

## DISCUSSION

The applied five ML methods, three of which are tree-based and two are NN-based, performed well on the Test-1 instances as the Test-1 scores were close to the Training scores. However, tree-based models had a significant performance drop when extrapolating on the Test-2 domain compared to NN-based models. This result shows that NN-based models are much more robust on extrapolation than tree-based models. We also compared the effect of different encoding methods (see Table 2).

Among them, fingerprint-based methods (ECFC, ECFP, and SECFP) showed the top performances. Although these three fingerprints are calculated in a similar manner, ECFC contains not only the substructures but also their number of occurrences. This additional information positively contributed to the performance of the models.

One of the key outcomes of this study is that it reveals the influence of the investigated chemical space of the molecules on the evaluation of the performance of the models. Although neither test set participated in the training process, the models performed relatively poorly in Test-2 (outside the training domain) whereas they did very well in Test-1 (within the training domain). It is important to note that, although the Test-2 data contains instances that have no overlap with the chemical space domain of the training set, it is still in relatively close proximity to this domain considering the vast chemical space of all organic compounds. This result shows that data-driven models not only tend to overfit on training data but also tend to overfit on their training domain. It points out that the ML models tested using only a test set sharing the same chemical space with the training set are very likely to fail on extrapolation. Therefore, it is crucial to have an additional test set that is not covered by the chemical space of the training set in order to determine the extrapolation capability of the proposed models.

A recent study<sup>29</sup> also demonstrated the applicability of ML-driven approaches for the discovery of AORFB compounds. They trained a GCN model using around 45,000 reaction pairs extracted from QM9/G4(MP2)<sup>37</sup> dataset. Their model performed very well on five-fold cross-validation, however, showed poor performance on the external test set from the NIST database, especially on structures that were underrepresented in the training data. It is not possible to compare their results with ours directly since the models are trained using different datasets covering different chemical spaces. QM9/G4(MP2) dataset contains compounds up to nine atoms and comprise five elements (C, H, O, N, and F) while RedDB contains AORFB-focused compounds up to 46 atoms and comprise six different elements (C, H, O, N, F, and S) at max. Nevertheless, both studies show the possibility of data-driven discovery of candidate AORFB compounds by their accurate predictions on a defined chemical space. They also underline the danger of the inapplicability of data-driven models on the outside of this chemical space.

Another essential outcome of the study is detecting the structural inconsistencies in the DFT-optimized data based on the predictions of ML models. As described in the Results section, we identified 39 reaction pairs with large differences between the DFT-calculated and the ML-predicted reaction energies. It turned out that 25 molecules from the identified pairs had noticeable structural distortions. This result showed that ML predictions can also be used for the validation of data produced by computational and experimental methods.

## METHODS

### Data

The data used in this study was collected from RedDB-v1<sup>9</sup>, which contains 31,677 DFT calculated electroactive candidate compounds for AORFBs. The compounds in the RedDB library had been generated from the chemical functionalization of 52 different core structures that belong to two foremost studied classes of organic electroactive compounds: quinones and aza-aromatics. In RedDB, the redox couples are matched by assuming a reversible two-electron two-proton mechanism in which the product molecule, MH<sub>2</sub>, is generated from the reactant

molecule,  $M$ . Accordingly, the redox reaction energy ( $\Delta E_{\text{rxn}}$ ) is calculated as

$$\Delta E_{\text{rxn}} = E(\text{MH}_2) - [E(M) + E(\text{H}_2)] \quad (1)$$

where  $E(M)$ ,  $E(\text{MH}_2)$ , and  $E(\text{H}_2)$  are the total energies of reactant and product molecules, and hydrogen molecule, respectively<sup>9</sup>.

In the current study, we used a total of 15,794 redox couples and their DFT-calculated reaction energy data from RedDB. The following set of operations is executed to separate the datasets into three parts: Training, Test-1, and based on their chemical space occupation. First, the chemical space of the dataset is generated by ChemPlot<sup>31</sup> version 0.1.2 using structural similarity and UMAP dimensionality reduction methods with the default parameters except for the *random state* set to zero for the reproduction. Next, Test-2 is selected from the dataset based on four criteria; (1): closer clusters that are far from the center of the chemical space are selected, (2): all the instances within the selected cluster are included, (3): chemical space of Test-2 does not intersect the chemical space of the remaining data, (4): approximately 10% of the instances are selected. Finally, around 10% is selected by stratified random sampling as Test-1 from the remaining data. Only the reactant molecule SMILES data is used both for the separation of the dataset and for the training of ML models. Figure 2 shows the coverage of the separated data sets in the visualized chemical space.

#### **ECFP encoder**

ECFP encodes the molecules as the bit vector where each dimension represents the existence of a particular substructure in the molecules. ECFP determines the substructures of the molecules by starting from each non-hydrogen atom and attaching the neighbor atoms until a specified radius is reached. In this study, we used RDKit<sup>38</sup> implementation of ECFP with a vector length of 2,048 and radius of 2 adjacent atoms.

#### **ECFC encoder**

ECFC uses the same algorithm as ECFP but creates a count vector instead of a bit vector. Each dimension of the count vector represents the number of a particular substructure that exists in the molecules. In this study, we used RDKit<sup>38</sup> implementation of ECFC with a vector length of 2,048 and radius of 2 adjacent atoms.

#### **SECFP encoder**

SECFP encodes the molecules as bit-vector using the same extended connectivity principle of ECFC but using the MHFP<sup>34</sup> implementation. In this study, we implemented SECFP encoder with a vector length of 2,048 and a radius of 3 adjacent atoms from MHFP<sup>34</sup> library.

#### **MACCS encoder**

MACCS encodes the molecules as bit-vector based on the existence of predefined 166 substructures. In this study, we used RDKit<sup>38</sup> implementation of MACCS encoder.

#### **Mol2vec encoder**

Mol2vec<sup>35</sup> is a Python library that allows learning vector representations of molecules using machine learning. Similar to the Word2vec models used in natural language processing, Mol2vec converts molecules into ECFP representation and then treats substructures as words in sentences. Using an unsupervised approach, Mol2vec creates the feature vectors. We used a pretrained model on 20 million compounds provided by the library. The pretrained model converts the given molecules into a 300-dimensional feature vector.

### Mordred descriptors

Mordred library<sup>36</sup> contains more than 1,800 2D and 3D molecular descriptors in its catalog. In this study, we employed 201 physico-chemical 2D descriptors which contain topological descriptors and the number of atoms, bonds, and rings.

### ANN model configuration

We used the Keras framework to train the ANN models. We employed a five-layered architecture (2048, 128, 32, 8, 1). Using grid search, we optimized the parameters of the ANN models with the configurations given below. To optimize the given parameters, we used 5-fold cross-validation and also early stopping with a patience of 5 by reserving 20 percent of the training set for validation. The selected parameters for the final ANN model that shows the best performance are given below in bold. The remaining parameters were used with default values.

- Activation function: (sigmoid, **relu**, tanh)
- Batch size: (**16**, 32, 64, 128)
- Dropout rate: (**0**, 0.1, 0.2)
- Learning rate: (**0.001**, 0.005, 0.01)
- Optimizer: (RMSprop, **Adam**, SGD)

### RF model configuration

We used scikit-learn<sup>39</sup> library to train the RF models. We optimized the parameters of the RF models using grid search. The selected parameters for the final RF model that shows the best performance are given below in bold. The remaining parameters were used with default values.

- Number of estimators: (**100**, 1000)
- Max depth: (**no limit**, 5, 10, 15, 20)
- Bootstrap: (**True**, False)

### LGBM model configuration

We used lightGBM<sup>40</sup> library to train the LGBM models. Using grid search, we optimized the parameters of the LGBM models with the configurations given below. To optimize the given parameters, we used 5-fold cross-validation and also early stopping for a maximum of 500 iterations with a patience of 3 by reserving 20 percent of the training set for validation. The selected parameters for the final LGBM model that shows the best performance are given below in bold. The remaining parameters were used with default values.

- Max depth: (1, 3, 5, 7)
- Learning rate: (0.01, 0.03, 0.05, **0.1**)
- Regularization alpha(L1): (**0**, 0.1, 0.2, 0.3, 0.4 )
- Regularization lambda(L2): (0, 0.1, **0.2**, 0.3, 0.4 )

### XGB model configuration

We used XGBoost<sup>41</sup> library to train the XGB models. Using grid search, we optimized the parameters of the XGB models with the configurations given below. To optimize the given parameters, we used 5-fold cross-validation and also early stopping for a maximum of 500 iterations with a patience of 3 by reserving 20 percent of the training set for validation. The selected parameters for the final XGB model that shows the best performance are given below in bold. The remaining parameters were used with default values.

- Max depth: (5, 6, 7)
- Learning rate: (0.01, 0.05, **0.1**, 0.3, 0.5)
- Regularization alpha(L1): (1, 1.1, **1.2**)
- Gamma: (**0**, 0.1, 0.2)

### GCN model configuration

We used DeepChem<sup>42</sup> library to train the GCN models. Unlike other models, GCN uses its own featurization method instead of ECFC. We employed ConvMolFeaturizer class which implements Duvenaud<sup>43</sup> graph convolutions to encode the molecules. We trained the GCN model with the default parameters except for adding a dropout of 0.25 and a batch size of 100. We trained the model for 100 epochs and the best epoch state is selected based on the performance over 20 percent of the reserved validation set from the training set.

### Ensemble model configuration

We configured the ensemble model by combining the top performing three models which are ANN, RF, and GCN. We conducted a weighted ensembling based on inversely proportional to the cube of the mean of Test-1 and Test-2 MAE scores given in Table 2. The calculated weights of the ensemble model are given in Equation (2).

$$Pred_{ENS} = 0.551 \cdot Pred_{ANN} + 0.264 \cdot Pred_{RF} + 0.185 \cdot Pred_{GCN} \quad (2)$$

where  $Pred_{ENS}$ ,  $Pred_{ANN}$ ,  $Pred_{RF}$ , and  $Pred_{GCN}$  are the predictions of Ensemble, ANN, RF, and GCN models, respectively.

### Similarity search

To search for similar molecules from RedDB we used Tanimoto similarity. The Tanimoto similarity is calculated as the ratio of the intersection of the two fixed-size binary vector representations (ECFP encodings) of molecules over their union. The calculation of Tanimoto similarity is given in Equation (3).

$$Tanimoto(U, V) = \frac{|U \cap V|}{|U \cup V|} \quad (3)$$

where  $U$  and  $V$  are binary vector representations of two molecules. We used reactant molecules from reaction pairs to find their similar pairs. From RedDB, the molecules having Tanimoto similarity of 0.6 or higher to the target molecule were selected as similar molecules.

### Code Availability

The reproducibility of the RedPred can be verified by executing the provided scripts on Code Ocean (<https://codeocean.com/capsule/0454424/tree/v1>). The freely accessible RedPred web tool is reachable at: <https://www.amdlab.nl/redpred/>.

### References

1. Dunn, B., Kamath, H. & Tarascon, J.-M. Electrical energy storage for the grid: a battery of choices. *Science* **334**, 928–935 (2011).
2. Sánchez-Díez, E. *et al.* Redox flow batteries: Status and perspective towards sustainable stationary energy storage. *J. Power Sources* **481**, 228804 (2021).
3. Rychcik, M. & Skyllas-Kazacos, M. Characteristics of a new all-vanadium redox flow battery. *J. Power Sources* **22**, 59–67 (1988).

4. Lee, S., Hong, J. & Kang, K. Redox-active organic compounds for future sustainable energy storage system. *Adv. Energy Mater.* **10**, 2001445 (2020).
5. Er, S., Suh, C., Marshak, M. P. & Aspuru-Guzik, A. Computational design of molecules for an all-quinone redox flow battery. *Chem. Sci.* **6**, 885–893 (2015).
6. Cheng, L. *et al.* Accelerating electrolyte discovery for energy storage with high-throughput screening. *The J. Phys. Chem. Lett.* **6**, 283–291 (2015).
7. Tabor, D. P. *et al.* Mapping the frontiers of quinone stability in aqueous media: implications for organic aqueous redox flow batteries. *J. Mater. Chem. A* **7**, 12833–12841 (2019).
8. Fornari, R. P., Mesta, M., Hjelm, J., Vegge, T. & de Silva, P. Molecular engineering strategies for symmetric aqueous organic redox flow batteries. *ACS Mater. Lett.* **2**, 239–246 (2020).
9. Sorkun, E., Zhang, Q., Khetan, A., Sorkun, M. C. & Er, S. Reddb, a computational database of electroactive molecules for aqueous redox flow batteries. *Sci. Data* **9**, 718 (2022).
10. Zhang, Q., Khetan, A., Sorkun, E. & Er, S. Discovery of aza-aromatic anolytes for aqueous redox flow batteries via high-throughput screening. *J. Mater. Chem. A* **10**, 22214–22227 (2022).
11. Zhang, Q. *et al.* Data-driven discovery of small electroactive molecules for energy storage in aqueous redox flow batteries. *Energy Storage Mater.* **47**, 167–177 (2022).
12. Faber, F. A. *et al.* Prediction errors of molecular machine learning models lower than hybrid dft error. *J. Chem. Theory Comput.* **13**, 5255–5264 (2017).
13. Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A. & Kim, C. Machine learning in materials informatics: recent applications and prospects. *npj Comput. Mater.* **3**, 1–13 (2017).
14. Zhou, T., Song, Z. & Sundmacher, K. Big data creates new opportunities for materials research: A review on methods and applications of machine learning for materials design. *Engineering* **5**, 1017–1026 (2019).
15. Himanen, L., Geurts, A., Foster, A. S. & Rinke, P. Data-driven materials science: status, challenges, and perspectives. *Adv. Sci.* **6**, 1900808 (2019).
16. Sorkun, M. C., Astruc, S., Koelman, J. V. A. & Er, S. An artificial intelligence-aided virtual screening recipe for two-dimensional materials discovery. *npj Comput. Mater.* **6**, 1–10 (2020).
17. Chibani, S. & Coudert, F.-X. Machine learning approaches for the prediction of materials properties. *APL Mater.* **8**, 080701 (2020).
18. Singh, V., Kim, S., Kang, J. & Byon, H. R. Aqueous organic redox flow batteries. *Nano Res* **12**, 1988–2001 (2019).
19. Hu, B., DeBruler, C., Rhodes, Z. & Liu, T. L. Long-cycling aqueous organic redox flow battery (aorfb) toward sustainable and safe energy storage. *J. Am. Chem. Soc.* **139**, 1207–1214 (2017).
20. Huuskonen, J. Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology. *J. Chem. Inf. Comput. Sci.* **40**, 773–777 (2000).
21. Delaney, J. S. Esol: estimating aqueous solubility directly from molecular structure. *J. Chem. Inf. Comput. Sci.* **44**, 1000–1005 (2004).
22. Sorkun, M. C., Khetan, A. & Er, S. Aqsolddb, a curated reference set of aqueous solubility and 2d descriptors for a diverse set of compounds. *Sci. Data* **6**, 1–8 (2019).

23. Balakin, K. V., Savchuk, N. P. & Tetko, I. V. In silico approaches to prediction of aqueous and dmsolubility of drug-like compounds: trends, problems and solutions. *Curr. Medicinal Chem.* **13**, 223–241 (2006).
24. Schwaighofer, A. *et al.* Accurate solubility prediction with error bars for electrolytes: A machine learning approach. *J. chemical information modeling* **47**, 407–424 (2007).
25. Wang, J. & Hou, T. Recent advances on aqueous solubility prediction. *Comb. Chem. & High Throughput Screen.* **14**, 328–338 (2011).
26. Chevillard, F. *et al.* In silico prediction of aqueous solubility: a multimodel protocol based on chemical similarity. *Mol. Pharm.* **9**, 3127–3135 (2012).
27. Sorkun, M. C., Koelman, J. V. A. & Er, S. Pushing the limits of solubility prediction via quality-oriented data selection. *iScience* **24**, 101961 (2021).
28. Ghule, S., Bagchi, S. & Vanka, K. Machine learning the redox potentials of phenazine derivatives: A comparative study on molecular features. *ChemRxiv preprint* (2021).
29. Barker, J., Berg, L.-S., Hamaekers, J. & Maass, A. Rapid prescreening of organic compounds for redox flow batteries: A graph convolutional network for predicting reaction enthalpies from smiles. *Batter. & Supercaps* **4**, 1482–1490 (2021).
30. Zhang, Q., Khetan, A. & Er, S. Comparison of computational chemistry methods for the discovery of quinone-based electroactive compounds for energy storage. *Sci. Reports* **10**, 1–13 (2020).
31. Sorkun, M. C., Mullaj, D., Koelman, J. V. A. & Er, S. Chemplot, a python library for chemical space visualization. *Chemistry-Methods* e202200005 (2022).
32. McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).
33. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).
34. Probst, D. & Reymond, J.-L. A probabilistic molecular fingerprint for big data settings. *J. Cheminformatics* **10**, 1–12 (2018).
35. Jaeger, S., Fulle, S. & Turk, S. Mol2vec: unsupervised machine learning approach with chemical intuition. *J. Chem. Inf. Model.* **58**, 27–35 (2018).
36. Moriwaki, H., Tian, Y.-S., Kawashita, N. & Takagi, T. Mordred: a molecular descriptor calculator. *J. Cheminformatics* **10**, 1–14 (2018).
37. Kim, H., Park, J. Y. & Choi, S. Energy refinement and analysis of structures in the qm9 database via a highly accurate quantum chemical method. *Sci. Data* **6**, 1–8 (2019).
38. Rdkit: Cheminformatics and machine learning software (2013).
39. Pedregosa, F. *et al.* Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
40. Ke, G. *et al.* Lightgbm: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **30**, 3146–3154 (2017).
41. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794 (2016).
42. Ramsundar, B. *et al.* *Deep Learning for the Life Sciences* (O'Reilly Media, 2019).
43. Duvenaud, D. *et al.* Convolutional networks on graphs for learning molecular fingerprints. *arXiv preprint arXiv:1509.09292* (2015).

## **Acknowledgements**

The authors acknowledge funding from the initiative “Computational Sciences for Energy Research” of Shell and the Netherlands Organisation for Scientific Research (NWO) grant no 15CSTT05. This work was sponsored by NWO Exact and Natural Sciences for the use of supercomputer facilities.

## **Author contributions**

M.C.S., E.N.G., and C.Y. developed codes for preprocessing the data, ML modeling, and analyzing the results. S.E. devised and supervised the project. All authors contributed to the writing of the manuscript.

## **Competing interests**

The authors declare no competing interests.