Bridging the experiment-calculation divide: machine learning corrections to redox potential calculations in implicit and explicit solvent models

Eugen Hruska, Ariel Gale, and Fang Liu*

Emory University, Department of Chemistry, Atlanta, Georgia 30322, United States

E-mail: fang.liu@emory.edu

Abstract

Prediction of redox potentials is essential for catalysis and energy storage. Although density functional theory (DFT) calculations have enabled rapid redox potential predictions for numerous compounds, prominent errors persist compared to experimental measurements. In this work, we develop machine learning (ML) models to reduce the errors of redox potential calculations in both implicit and explicit solvent models. Training and testing of the ML correction models are based on the diverse ROP313 dataset with experimental redox potentials measured for organic and organometallic compounds in a variety of solvents. For the implicit solvent approach, our ML models can reduce both the systematic bias and the number of outliers. ML corrected redox potentials also demonstrate less sensitivity to DFT functional choice. For the explicit solvent approach, we significantly reduce the computational costs by embedding the microsolvated cluster in implicit bulk solvent, obtaining converged redox potential results with a smaller solvation shell. This combined implicit-explicit solvent model, together with GPU-accelerated quantum chemistry methods, enabled rapid generation of a large dataset

of explicit-solvent-calculated redox potentials for 165 organic compounds, allowing detailed investigation of the error sources in explicit solvent redox potential calculations.

1 Introduction

Redox potential is a fundamental thermodynamic property that describes the tendency of a chemical species to lose or acquire electrons. It is essential in mechanism studies of catalysis[1,2] and energy storage.[3] Design and discovery in these fields needs rapid predictions of redox potentials of thousands or even millions of candidate compounds, which can hardly be achieved with cyclic voltammetry measurements. Recent developments in accelerated quantum mechanical (QM) methods and solvent models enabled rapid curation of computational redox potentials datasets with thousands of molecules. Machine learning (ML) models trained on these computational data sets can accurately reproduce the QM calculated redox potentials, enabling the exploration of multi-million compound spaces for redox flow battery design[4] and biochemical discovery.[5]

However, the accuracy of QM predicted redox potentials relative to experiment strongly depends on the QM methods and solvent-model-related parameters. Highly accurate results ($\leq$ 65 mV) are only found in small batch studies with specific combinations of methods and parameters applied to a specific class of compounds.[6] This uncertainty in QM calculations can impact the accuracy of curated computational redox potential datasets, and hence, the predictivity of ML models trained on these computational datasets used for design and discovery. Therefore, there is an urgent need to improve the accuracy of the QM prediction of redox potentials.

Due to the complexity of the solvent environment, errors in QM calculations of redox potentials can be attributed to many different sources. One prominent source is the solvation-free energy of species involved in the redox process. Most redox potential calculations were conducted in implicit solvent models, such as the conductor-like polarizable continuum models (COSMO,[7] C-PCM,[8–10] GCOSMO,[11] IEF-PCM[12–14]) and its variant COSMO-RS,[15] due to their computational efficiency. However, prominent errors have been observed in implicit solvent redox potential calculations,

especially for highly charged compounds.[16,17] Systematic bias of the calculated redox potentials relative to the experiment is often corrected with a simple linear regression,[18] which, however, cannot deal with the different error sizes for differently charged species, nor can it remove large-error outliers. A few strategies have been proposed to correct the errors associated with highly charged species and verified to be effective on small data sets (ca. 20 compounds), such as the Pseudo-counterion Solvation Scheme,[16] and the variable-temperature H-atom addition/abstraction.[17]

Explicit solvent approaches are shown to be more accurate for some challenging systems, such as transition metal complexes.[18–20] However, they can hardly be used in high-throughput computational design and discovery due to their significantly higher computational costs. The high computational costs are caused by the fact that converged redox potential results can only be obtained by including a large enough explicit solvent shell into QM calculation.[21] The optimal solvent shell size varies among different studies on different solute molecules.[22–24]

In this work, we develop ML models to reduce the errors of QM redox potential calculations in both implicit and explicit solvent models. Training and testing of the ML correction models are based on the large, diverse ROP313 dataset[25] with experimental redox potentials of organic and organometallic compounds measured in different solvents. For the implicit solvent approach, we trained ML models to improve the accuracy relative to experiments across a diverse set of molecules regardless of the DFT functionals chosen. For the explicit solvent approach, we significantly reduce the computational costs by embedding the microsolvated cluster in C-PCM bulk solvent, obtaining converged redox potential results with fewer solvation shells. This combined implicit-explicit solvent model, together with GPU-accelerated QM methods, enabled rapid generation of a large dataset of explicit-solvent-calculated redox potentials for 165 redox couples, allowing us to investigate the source of errors in explicit solvent redox potential calculations.

2 Theory

The general formula for redox potential calculation is given by the Nernst equation

$$\mathrm{E}^{\circ} = -\frac{\Delta G^{\mathrm{EA}}_{(\mathrm{sol})}}{n_{\mathrm{e}}\mathrm{F}} - \mathrm{E}^{\circ}(\mathrm{REF}), \tag{1}$$

where $\Delta G^{\mathrm{EA}}_{(\mathrm{sol})}$ is the free energy change associated with reduction at standard conditions in the solution phase, n_{e} is the number of electrons, and F is the Faraday constant. In this work, the one-electron ferrocenium/ferrocene (Fc^{+}/Fc) couple is used as the reference, because Fc^{+}/Fc is the internal reference used in the experimental measurement of the redox potential of ROP313 dataset[25] and is known to be useful to reduce both experimental[26,27] and computational errors,[28,29] especially when varying the solvent in which the redox potential is measured.

2.1 Implicit solvent calculation of redox potential

In the implicit solvent models, $\Delta G^{\mathrm{EA}}_{(\mathrm{sol})}$ is calculated based on the Born-Haber cycle,[21] given by

$$\Delta G^{\mathrm{EA}}_{(\mathrm{sol})} = G_{\mathrm{PCM}}(\mathrm{red}) - G_{\mathrm{PCM}}(\mathrm{ox}) + \Delta H^{\mathrm{T}} - T\Delta S(g). \tag{2}$$

Here $G_{\mathrm{PCM}}(\mathrm{red})$ and $G_{\mathrm{PCM}}(\mathrm{ox})$ are the free energy of the reduced and oxidized species obtained from PCM calculation including electron energy and solvent-solute interaction energy, whereas ΔH^{T} and $-TS(g)$ are the gas-phase enthalpy and entropy contributions to the Gibbs free energy. Recent studies on large datasets show that the gas-phase enthalpy and entropy contributions have very limited influence on the accuracy of redox potential prediction, and can be omitted to avoid computational costs related to vibrational frequency calculation.[4] Hence, the equation for redox potential calculation in implicit solvent in this work is simplified as

$$\Delta G^{\mathrm{EA}}_{(\mathrm{sol})} = G_{\mathrm{PCM}}(\mathrm{red}) - G_{\mathrm{PCM}}(\mathrm{ox}). \tag{3}$$

2.2 Explicit solvent calculation of redox potential

The redox potential of the explicitly solvated molecules is calculated with the thermodynamic integration (TI) method.[30] Specifically, we use the linear response (LR) approximation of TI[31–33] to avoid simulations of the nonphysical superposition state of the reduced and oxidized forms of the system.[19] For each redox couple, two QM/MM molecular dynamics simulations are performed separately for the reduced and oxidized states. The free energy difference of the reduced and oxidized states is evaluated by thermally averaging the vertical energy gap of the reduced and oxidized states (Eq.4).

$$\Delta G_{\mathrm{TI}} = \frac{1}{2}\left(\langle G_{\mathrm{red}} - G_{\mathrm{ox}}\rangle_{\mathrm{ox}} + \langle G_{\mathrm{red}} - G_{\mathrm{ox}}\rangle_{\mathrm{red}}\right) \tag{4}$$

The validity of the TI method and the LR approximation can be estimated by comparing the estimated reorganization energies, λ^{St} and λ^{var}, according to the Marcus theory of electron transfer.[34] Here, λ^{St} and λ^{var} are called Stokes reorganization energy and variance reorganization energy, respectively, defined as

$$\lambda^{\mathrm{St}} = \frac{\langle \Delta G\rangle_{\mathrm{ox}} - \langle \Delta G\rangle_{\mathrm{red}}}{2n_{\mathrm{e}}F}, \tag{5}$$

$$\lambda^{\mathrm{var}} = \frac{\sigma^2_{\mathrm{ox}} + \sigma^2_{\mathrm{red}}}{4k_{\mathrm{b}}T}, \tag{6}$$

where σ is the variance of the vertical energy gap, and k_{b} is the Boltzmann constant. The reorganization energies λ^{St} and λ^{var} should reach identical values when the LR approximation holds.[33]

3 Computational details

3.1 Data set

We employ the ROP313[25] data set to compare the different computational approaches for predicting redox potential. This benchmark set with experimental redox potentials for 313 individual

molecules is composed of two subsets of 193 organics (OROP) and 120 organometallics (OMROP), respectively. It allows a representative comparison of the accuracy of different computational methods due to the diverse and considerable number of systems. The systems are medium-sized with the number of atoms ranging from 5 to 82 and total electrons from 22 to 427 (Figure 1). The data set contains four different solvents spanning a wide range of dielectric constants, including acetonitrile (MeCN, ε=35.69), water (ε=78.36), dichloromethane (ε=8.93), and dimethylformamide (DMF, ε=37.22). The experimental redox potentials were converted to an internal reference of Fc^{+}/Fc in an identical solvent.[35,36] The oxidized forms of benchmark molecules have charges ranging from -4 to 2. The organometallics possess a variety of 3rd row (Ti, V, Cr, Mn, Fe, Ni, Co) and 4th row (Ru, Rh, Os, Ir) transition metal centers. The accuracy of redox potential calculations for OROP and OMROP are analyzed separately due to the typically larger errors in the calculation of organometallic systems.[25]

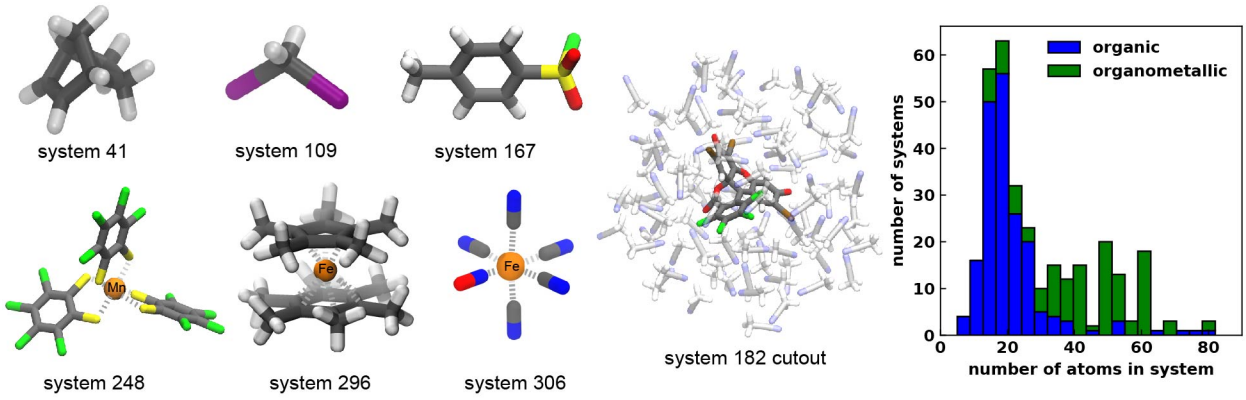

Figure 1: Representative systems from ROP313 dataset used in this work. (top left) Three representative organics including the cyclic aliphatic Norbornane, the halocarbon diiodomethane, and the aromatic 4-methylbenzylsulfonyl chloride. (bottom left) Three representative organometallics with bidentate (3,4,5,6-tetrachlorobenzene-1,2-dithiolate), sandwich (Pentamethylcyclopentadiene), and monodentate (CN^{-}, CO) ligands. (middle) Example structure of a microsolvated cluster used in our explicit solvent calculation. (right) The distribution of the number of atoms in each system. All atoms are colored by elements [C in grey, H in white, O in red, N in blue, Cl in green, S in yellow, I in purple, metal (Mn, Fe) in orange].

3.2 DFT calculation

All geometry optimizations and single point energy calculations were performed with the GPU-accelerated quantum chemistry package TeraChem.[37] Unless otherwise specified, the density functional theory (DFT) calculations use the B3LYP hybrid functional with DFT-D3 empirical dispersion correction,[38] combined with either 6-31G* basis set or LANL2DZ[39] effective core potentials for the transition metals and I and Br. To test the robustness of our machine learning models to different functionals, a set of hybrid (B3LYP-D3, PBE0-D3) and range-corrected hybrid (ωb97-D3, ωb97X-D3, ωPBEh-D3, CAM-B3LYP-D3) functionals commonly used for redox potential calculations are also employed to calculated single point energies to obtain redox potentials. The geometries optimized with B3LYP-D3 are used in all cases.

For implicit solvent model calculations, geometry optimizations were carried out with the TRIC[40] optimizer using default tolerances of 4.5×10^{-4} hartree/bohr for the maximum gradient and 1×10^{-6} hartree for the change in self-consistent field (SCF) energy between steps. Initial structures, charge, and spin states were obtained from the ROP313 data set originally optimized with ωB97x-c3. Level-shifting[41] values of 0.3 H_a for virtual orbitals were applied if the calculation did not converge without level-shifting. Solvation energies were obtained from single point energies with a conductor-like polarizable continuum model (C-PCM)[42,43] as implemented in TeraChem. The solute cavity was built using defaults available for nonmetals in TeraChem (i.e., 1.2$\times$ Bondi's van der Waals radii,[44]) and we provided standard van der Waals radii[45] for metals, which were also scaled by 1.2.

3.3 Explicit solvent calculation

Due to the significantly higher computational costs of the explicit solvation calculations, only 165 organic redox couples solvated in MeCN are calculated (details in Supporting Information/SI). Initial explicit solvation configurations are generated with Packmol[46] in a cubic box of side length 56 Å with Na^+ and Cl^- counterions. A multi-step approach is chosen to equilibrate the solvated box. First, molecular dynamics parameters are generated for the organic solute molecules with Amber-

tools[47] using GAFF force field.[48] The solvent MeCN molecules are described with a customized 6-site model from literature.[49] The explicitly solvated reference Fc^+/Fc redox couple is set up in a similar approach, but the parameters for the Fc^+/Fc solute are determined from MCPB.py[50] due to the existence of metal-ligand bonds. These non-polarizable MM force fields allowed long-time dynamics in Amber 20[47] without bond-breaking at minimal computational costs. Second, the solvated system is minimized, and then slowly heated to 300 K over 20 ps with a Langevin thermostat with a collision frequency of 2 ps^{-1} and a nonbonded cutoff of 8 Å, and pressure equilibrated to 1 bar over 600 ps with a Berendsen barostat with a pressure relaxation time of 1 ps. This long pressure equilibration is chosen to improve the density on the interface between the solute and solvent, which is critical for the explicit redox potential calculation. This multi-step solvation of each system is automated by AutoSolvate, an in-house python script developed in our group, which is available upon request. The resulting pressure-equilibrated system is the initial structure for the QM/MM[51,52] simulation with TeraChem and Amber 20. The QM region is then solute treated with B3LYP-D3/6-31G*, and the MM region is the explicit solvent. The QM electrostatic cutoff is 8 Å. The QM/MM simulation involves an initial energy minimization, followed by 0.5 ps of temperature equilibration with Langevin thermostat at 298.15K with a collision frequency of 5 ps^{-1}. The following 5 ps of QM/MM NVT dynamics trajectory are used for TI. For each of the two charge states, 200 snapshots are extracted from the 5 ps QM/MM trajectories. The large number of snapshots is required to average out the variance of the vertical energy gap.[53] Each snapshot includes the solute and dozens of solvent molecules. To conduct DFT single point calculations needed by TI on these snapshots with reasonable computational costs, a cutoff needs to be applied to generate a microsolvated cluster that is small enough to calculate with DFT but large enough to generate a converged redox potential. We tested various cutoff values from 2 Å to 10 Å. We also investigated how C-PCM implicit solvent applied around these microsolvated clusters impacts the calculated redox potential.

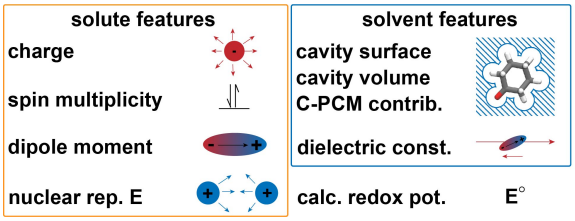

Figure 2: ML features used for training machine learning models to correct errors in redox potential calculations in this study.

3.4 Machine learning models

Machine learning (ML) models are utilized to reduce the errors in computed redox potentials compared to experimental values. Due to the relative small size of the experimental redox potential dataset, we trained ML methods on a small set of 9 physics-inspired features related to solvent effects and redox potential calculation, including 4 features for the solute molecule (charge, dipole moment, spin multiplicity, and nuclear repulsion energy of the oxidized state), 4 features for the solvent model (implicit solvation cavity surface area and volume, C-PCM solvation energy, and dielectric constant of the solvent) plus the predicted redox potential (Figure 2). We compared the performance of different ML models utilizing scikit-learn, including simple linear regression (lin-1), multiple linear regression (lin-m), random forest regression (RF),[54] gradient boost regression (GB),[55] kernel ridge regression (KR),[56] and artificial neural network (ANN).[57] Hyperparameter optimization for all models was carried out with grid-search cross-validation to prevent over-fitting, using a random 80% train/20% test split, with 20% of the training set (16% overall) set aside as the validation subset for hyperparameter selection (SI Text S1). Input features were normalized over the training set to have zero mean and unit variance. Final models with optimal hyperparameters were retrained with the whole training set (80% overall). We repeated the training five times on different 80:20 splittings of the data set, such that the five 20% test sets in total cover all data points. Performance of the ML correction is reported for the combination of the 5 test sets.

3.5 Free Energy landscape and TICA

The convergence of the conformation dynamics of the QM/MM trajectories for our explicit solvent calculations can be confirmed by analyzing the trajectories. One challenge in analyzing the trajectories is the high-dimensional nature with up to hundreds of individual atoms moving. To analyze the solute conformation changes we use the dimension reduction method Time-lagged Independent Component Analysis (TICA),[58] which extracts a few dimensions from the high-dimensional raw trajectories. TICA represents the kinetic behavior of the trajectory better than principal component analysis (PCA). The following equations describe how TICA selects the dimensions with the highest kinetic information. The raw cartesian time-dependent coordinates are converted into rotation- and translation-invariant mean-free features $\mathrm{y}_i(\mathrm{t})$, in this case, distances between atoms. For a lag time τ, the covariance matrix $\mathrm{C}(0)$ and the time-lagged covariance matrix $\mathrm{C}(\tau)$ for the mean-free coordinates are calculated:

$$\mathrm{C}_{ij}(\tau) = \left\langle \mathrm{y}_i(\mathrm{t})\mathrm{y}_j(\mathrm{t}+\tau) \right\rangle_t \tag{7}$$

$$\mathbf{C}(\tau)v_i = \mathbf{C}(0)\mathrm{v}_i\lambda_i \tag{8}$$

The generalized eigenvalue problem allows us to select a few dimensions or eigenvectors v_i with the largest eigenvalues λ_i, which contain most of the kinetic information of this system and have the longest associated timescales. The projection of the raw trajectory on the selected eigenvectors gives a low-dimensional representation of the solute behavior. In this work, only the two slowest dimensions are selected, and the lag time utilized is 5 fs due to the fast conformational dynamics of the systems. In the projection defined by the two selected TICA dimensions, we calculate the free energy, which allows us to visualize both the ensemble of the reached conformations and the transitions between the conformations. The free energy landscape is comparable to the potential energy surface (PES), but with additional degrees of freedom considered[59]. In order to plot the redox potential projection in the TICA dimensions, additional redox potential calcula-

tions were performed for conformations uniformly sampled along the two TICA dimensions from the QM/MM trajectories.

4 Results

In the following subsections, we will investigate the errors in the calculated redox potential compared to experimental results, and build ML models to correct these errors and reduce computational costs. We will first investigate the accuracy of the implicit solvation redox potential calculations and build ML models to correct the errors related to the imbalanced treatment of different charge states in implicit solvent models. We will also demonstrate that the ML corrections improve the robustness of the calculated redox potentials with respect to DFT functionals. For the explicit solvation approach, we investigate the dependency of the calculated redox potential with respect to the solvent shell size and will demonstrate that by applying C-PCM implicit solvent around the explicit solvent molecules, a much smaller explicit solvent shell is required to obtain converged redox potentials, leading to significantly reduced computational costs. We will then discuss the impacts of applying ML correction to the explicit solvent calculated results. The source of errors in the explicit solvation redox potential calculations will be analyzed at last.

4.1 Errors in implicit solvent redox potential calculations

It is known that redox potential calculations in implicit solvent typically have relatively large uncertainties in the solvation free energy since electrochemical half-reactions involve the consumption or generation of charge species.[6] Although low unsigned errors of less than 100 mV were observed in small batch studies, more realistic errors can be much larger when larger and more diverse test sets are considered. In our calculation of the ROP313 dataset with B3LYP-D3 and C-PCM (Figure 3, Figures S1-S4), two types of errors are present: systematic bias and large-error outliers.

Comparing to experimental results, the calculated redox potentials of the OMROP data set show a systematic overestimation, especially in the higher value range (Figure 3). This type of

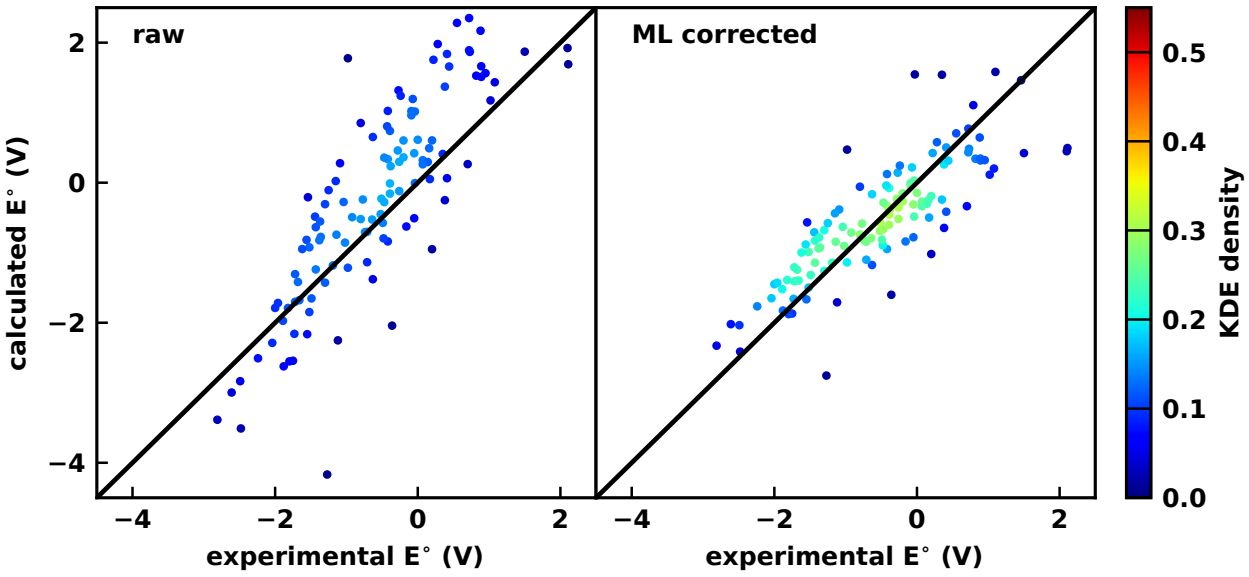

Figure 3: Parity plots of implicit solvent approach calculated vs experimental redox potentials for compounds in the OMROP dataset. Data points are colored by kernel density estimation (KDE) density values, as indicated by the color bar on the right. (left) Raw data from implicit solvation calculations. (right) ML corrected data obtained with the best performing lin-1 model.

systematic bias is commonly seen in implicit solvent redox potential calculations and is often corrected with simple linear fitting between the calculated and experimental redox potentials[21,60]. After applying the linear regression correction for OMROP, the mean absolute error (MAE) is reduced from 0.76 V to 0.44 V (Table 2).

However, the traditional linear fitting correction typically cannot help fixing errors that do not follow a systematic linear trend, including big-error outliers. For example, transition metal complexes with excess positive/negative charges are known to be typical outliers because implicit solvent models usually have an imbalanced treatment of solvation free energy for molecules with different net charges.[16,17] Inspired by this idea, we analyzed the errors of redox couples of different charge states. For the OROP data set, most systems have either +1/0 or 0/-1 charge states. Statistics on the distribution of signed errors of these two groups have different trends (Figure 4). The mean signed errors (MSEs) of both groups are negative, but the MSE of the +1/0 group (-0.25 V) is further from zero than that of the 0/-1 group (-0.08 V). The difference in MAE is even more significant: the 0/-1 group has a much higher MAE of 0.56 V than the 0.35 V of the +1/0 group

due to larger numbers of outliers. Similar results are observed in the OMROP dataset (SI Figure S5).

Because of different distributions of errors in systems with different charges, simple linear fitting cannot effectively correct the errors. A few methods have been proposed in previous works to fix these charge-dependent errors, including the pseudo-counterion solvation (PCIS) scheme that applies a charge-dependent correction formula,[16] and the variable-temperature H-atom addition/abstraction approach that modifies the thermodynamic cycle.[17] Here, we correct the systematic bias and charge-dependent outliers simultaneously by training machine learning models that are aware of the complex mapping between solute/solvent features and the errors in implicit solvent calculations, which will be discussed further in Section 4.2.

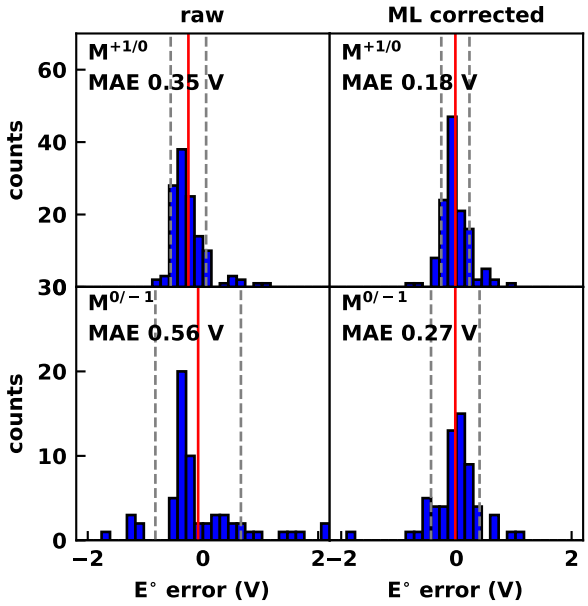

Figure 4: Implicit solvation redox potential errors in V for different system charges are shown in the histograms. The red vertical lines indicate the mean signed error for each charge. The gray dashed vertical lines show the size of the standard deviation. Results shown are for organic systems, organometallic systems are shown in SI Figure S5. The top two histograms show systems with a charge state of +1/0 and the bottom two histograms show a charge state of 0/-1. The left two histograms show raw implicit solvation results, and the mean is non-zero and different for each of the charge states. The two right histograms show results after ML correction (KR), where the mean bias is removed for both charge states.

4.2 ML corrections for implicit solvent calculations

We train ML models to correct the errors in implicit solvent redox potential calculations, focusing on reducing the uncertainty in solvation free energy. Since the C-PCM solvation free energy is determined by the charge distribution of the solute, the solvent dielectric constant, and the cavity shape, we use features that represent these physical properties and can be easily extracted from calculation output files (Figure 2). Specifically, net charge, dipole moment, and spin multiplicity of the solute represent the solute charge distribution; solvent dielectric constant describes the electrostatic screening strength; cavity surface area and volume describe the cavity shape; nuclear repulsion energy of the solute contains implicit information about the solute structure; and the calculated redox potential is also included in the features to allow for various corrections in different ranges of the redox potential.

For the B3LYP-D3 results discussed in Section 4.1, we observed reduced MAE after corrections with all types of ML models (Table 1). For the OROP dataset, all methods except for the linear models have significantly reduced the MAE by 7%-52%, with KR as the best performing model reducing the MAE from 0.30 V to 0.21 V. Even more prominent improvement is observed in the OMROP dataset, where all ML models present a MAE reduction of over 33%, with RF as the best performing model reducing the MAE to 0.43 V.

Further analysis shows that the error reduction is achieved by fixing both the systematic bias and the outliers discussed in Section 4.1. Compared to the raw calculated redox potential for OMROP, the ML-corrected results distribute more evenly on both sides of the diagonal of the parity plot (Figure 3), significantly removing the overestimation trend. More importantly, the more challenging charge-state dependent errors are reduced simultaneously (Figure 4). The MSEs of both the 0/-1 group and +1/0 group are both closer to 0 after applying ML correction (Figure 4). The MAEs of both groups are reduced by about 50%, demonstrating balanced correction to differently charged groups. Mitigation of large-error outliers is even more obviously shown by the reduced number of systems with an error above 1V. After applying the ML model, such systems are reduced from 14 to 2 for OROP dataset, and from 35 to 11 for OMROP dataset.

4.3 Robustness of the ML correction to DFT functional choice

In this section, we demonstrate the robustness of the ML corrections developed in Section 4.2 with respect to the choice of DFT exchange correlation (XC) functional. It is widely known that the accuracy of redox potential calculation in the implicit solvent is highly sensitive to the choice of XC functional, and the optimal functional is usually system dependent, leading to the lack of general strategy to obtain highly accurate computational redox potentials for large, diverse datasets.[25] This high sensitivity is also observed in our calculations. For the OROP dataset, the MAEs of the best-performing PBE0-D3 (0.30 V) and the worst-performing ωb97-D3 (0.64 V) differ by 0.34 V (Table 1). An even larger variation (0.76 V - 1.35 V) in the MAEs of different functionals is observed for the OMROP set (Table 2).

Table 1: DFT functional-dependence of the test set MAE (in V) of redox potentials predicted by implicit solvent model in combination with various ML correction models for the OROP dataset. For each functional, the best ML correction model with the smallest MAE (in V) is shown in the last column. For each column (ML model), functional sensitivity denotes the MAE difference between the best and worst performing functionals, and best functional denotes the MAE of the best performing functional.

MAE (V)	No ML	lin-1	lin-m	KR	GB	RF	ANN	best model
B3LYP-D3	0.43	0.33	0.32	0.22	0.27	0.27	0.25	KR/0.22
B3LYP	0.41	0.34	0.32	0.23	0.27	0.26	0.27	KR/0.23
ωb97-D3	0.64	0.55	0.36	0.33	0.32	0.30	0.33	RF/0.30
ωb97X-D3	0.47	0.48	0.35	0.29	0.33	0.27	0.30	RF/0.27
ωPBEh-D3	0.44	0.44	0.34	0.26	0.33	0.27	0.30	KR/0.26
PBE0-D3	0.30	0.31	0.31	0.21	0.25	0.25	0.27	KR/0.21
CAM-B3LYP-D3	0.40	0.42	0.34	0.26	0.31	0.28	0.25	ANN/0.25
functional sensitivity	0.34	0.23	0.05	0.11	0.08	0.05	0.08	RF/0.05
best functional	0.30	0.31	0.31	0.21	0.25	0.25	0.25	KR/0.21

In contrast, ML corrected results are always less sensitive to the choice of DFT functional. For the OROP dataset, any of the non-linear ML models can reduce the MAEs of each functional and reduce the performance difference among functionals to no more than 0.11V (Table 1). KR is generally the most efficient ML model for OROP, reducing the organic MAE down to 0.21 V. With KR correction, the MAE of each functional is reduced by at least 0.08 V, and a greater improvement is observed for the originally worse-performing functionals. As a result, the performance difference

Table 2: DFT functional-dependence of the test set MAE (in V) of redox potentials predicted by implicit solvent model in combination with various ML correction models for the OMROP dataset. For each functional, the best ML correction model with the smallest MAE (in V) is shown in the last column. For each column (ML model), functional sensitivity denotes the MAE difference between the best and worst performing functionals, and best functional denotes the MAE of the best performing functional.

MAE (V)	No ML	lin-1	lin-m	KR	GB	RF	ANN	best model
B3LYP-D3	0.76	0.44	0.46	0.46	0.50	0.47	0.46	lin-1/0.44
B3LYP	0.82	0.46	0.47	0.47	0.50	0.45	0.47	RF/0.45
ωb97-D3	1.35	0.67	0.70	0.75	0.79	0.74	0.73	lin-1/0.67
ωb97X-D3	1.27	0.60	0.63	0.60	0.71	0.67	0.77	KR/0.60
ωPBEh-D3	1.10	0.58	0.60	0.61	0.62	0.60	0.64	lin-1/0.58
PBE0-D3	0.83	0.45	0.47	0.45	0.47	0.43	0.48	RF/0.43
CAM-B3LYP-D3	1.32	0.59	0.61	0.63	0.72	0.67	0.65	lin-1/0.59
functional sensitivity	0.59	0.23	0.25	0.30	0.33	0.32	0.31	lin-1/0.23
best functional	0.76	0.44	0.46	0.45	0.47	0.43	0.46	RF/0.43

between any two functionals is reduced to less than 0.11 V from 0.34 V. (Figure 5). Similar reduction of functional sensitivity after ML-correction is observed for the OMROP dataset (Table 2). However, due to smaller OMROP dataset size, the linear ML correction performs similarly to non-linear ML corrections. Overall, RF has the best performance, reducing the OMROP MAE down to 0.43 V, whereas the lowest sensitivity to function choice [MAE(worst)-MAE(best)=0.23 V] is observed with lin-1. Although the functional sensitivity is much smaller after ML correction, B3LYP-D3 and PBE0-D3 are overall the best for both OROP and OMROP datasets. Therefore, B3LYP-D3 will be used in the explicit solvent model calculations in Sections 4.4 and 4.5.

The effectiveness of the ML-corrections in reducing MAE is also verified for DFT calculations with different basis set choices (SI Figure S10). After ML correction, calculations obtained with larger basis sets with diffuse functions show higher accuracy for both OROP and OMROP (Figures S2, S4, and S10), but the improvement to 6-31G* is only up to 0.04 V. Without ML correction, we observe an increased MAE for OMROP when larger basis sets are used, consistent with previous reports for transition metal complexes.[61]

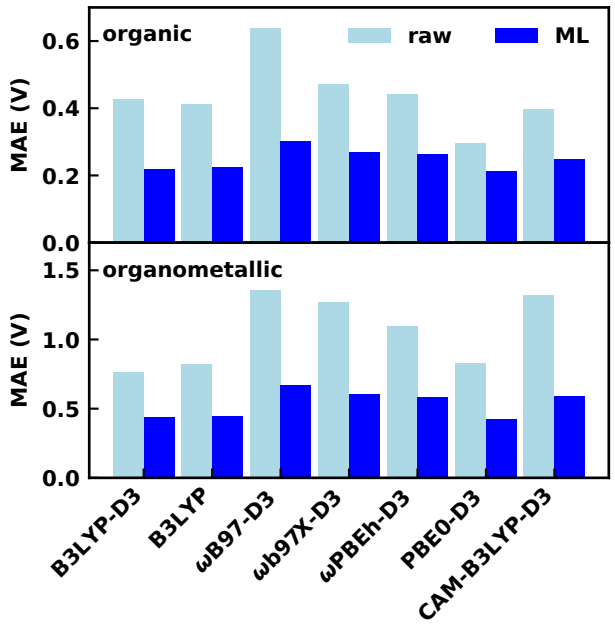

Figure 5: ML improvement dependence on the chosen functional. Values and description in Table 1. Only the best ML approach for each functional is shown. Values for OROP (top) and OMROP (bottom) data sets.

4.4 Convergence of explicit solvent redox potential calculations

Explicit solvent models are expected to provide a more accurate prediction of redox potential, but their application is limited by the higher computational costs compared to implicit solvent calculations. It was shown in previous studies that multiple layers of solvent shells are needed to obtain converged redox potential, requiring QM calculations of large microsolvated systems with hundreds of atoms.[22,62] Here we demonstrate that embedding the microsolvated cluster in C-PCM can effectively reduce the number of explicit solvent molecules needed without affecting the accuracy.

We first investigated the dependence of calculated redox potential values on the solvation shell size for system 141, Cyclohexanone, a typical system picked from ROP313 dataset (17 atoms, 54 electrons, Figure 6). For the traditional explicit solvent model without C-PCM, the redox potential varies strongly with solvent shell size and only converges above 10 Å (Figure 6). The existence of multiple plateaus at around 5, 8, and 10 Å is likely to be caused by the boundaries of solvent lay-

ers. In contrast, the redox potential obtained with the microsolvated cluster embedded in C-PCM converges rapidly as the solvent shell size increases and remains unchanged after 4 Å (Figure 6). The faster convergence of explicit solvent model embedded in C-PCM can be explained by a more efficient representation of long-range and polarization interactions with the bulk solvent compared to including only a few explicit solvation shells. The explicit solvation redox potentials calculated with and without C-PCM converge to different values, which is likely to be caused by the inherent deficiencies of C-PCM model in describing solvation free energy. We will show that these potential artifacts caused by C-PCM can be effectively removed with ML corrections, as demonstrated with the implicit solvent approach in Section 4.2-4.3.

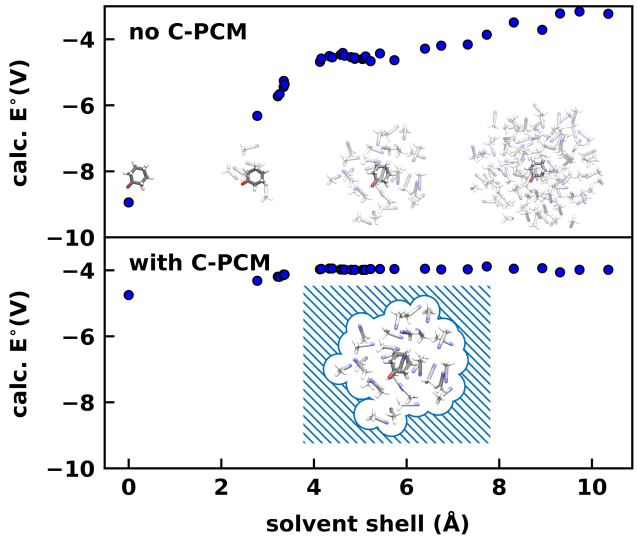

Figure 6: Convergence of redox potential with respect to the radius of the explicit solvation shell for system 141. (top) Explicit solvation shell only. Different explicit solvent shell sizes without pcm are shown at the top. (bottom) Explicit solvation shell together with pcm. The range of redox potentials with pcm is significantly smaller. Schema of combination explicit+implicit solvation at the bottom.

To demonstrate that the radius convergence holds for different systems, we repeat the investigation to determine the optimal solvent shell size for general redox potential calculations in the explicit solvent model with and without the C-PCM embedding for two additional OROP systems (SI Text S2). For each system, we calculate the absolute deviation of the redox potential at a given solvent shell size compared to the reference value obtained at 10 Å. The averaged absolute

deviations over the 3 systems at solvent shell sizes 2, 4, 6, and 8 Å are 5.7, 2.5, 2.3, and 1.8 V, respectively, for the explicit solvent model without C-PCM. In contrast, these values are 0.65, 0.03, 0.05, and 0.03 V for the explicit solvent model embedded in C-PCM. In summary, the explicit solvent model without C-PCM does not converge even at 8 Å, whereas a 4 Å solvent shell embedded in C-PCM reaches an accuracy and convergence sufficient for most use cases. The embedding of microsolvated clusters in C-PCM reduces the required solvation shell size and computational cost while maintaining accuracy.

4.5 Accuracy of explicit solvent calculations and ML corrections

Thanks to the reduced computational costs enabled by C-PCM embedding, we are able to rapidly curate a dataset of redox potentials of 165 organic systems calculated with explicit solvent model. Compared to previous studies with explicit solvent models typically involving only a few systems, this larger dataset allows us to summarize some statistically meaningful trends in this type of calculation. Similar to implicit solvent calculations, both systematic biases and large-error outliers are present in our explicit solvent calculations (Figure 7). The predicted values are systematically below the experimental values, and in the low redox potential region ($E^\circ \leq 0$ V), there are more outliers with overestimated redox potential, similar to the trend in implicit solvent calculations (Figure 3).

Table 3: Performance of explicit solvent approach before and after various ML corrections. The mean absolute error (MAE) in V and the number of systems with an error above 1V are shown for the analyzed 165 systems from OROP. The "No ML" column shows the explicit solvation results without any ML correction. The Implicit and Implicit+ML columns show the results of implicit solvent approach without ML correction and with the best performing model (KR). All calculations are performed with B3LYP/6-31G*.

	No ML	lin-m	KR	GB	RF	ANN	Best ML	Implicit	Implicit+ML
MAE (V)	0.64	0.22	0.20	0.22	0.19	0.24	RF/0.19	0.40	0.20
Errors >1V	3	3	2	3	4	4	GB/2	11	1

We then train machine learning models to correct the systematic bias and outliers in the explicit solvent redox potential calculations, using the same procedure as the implicit solvent ML

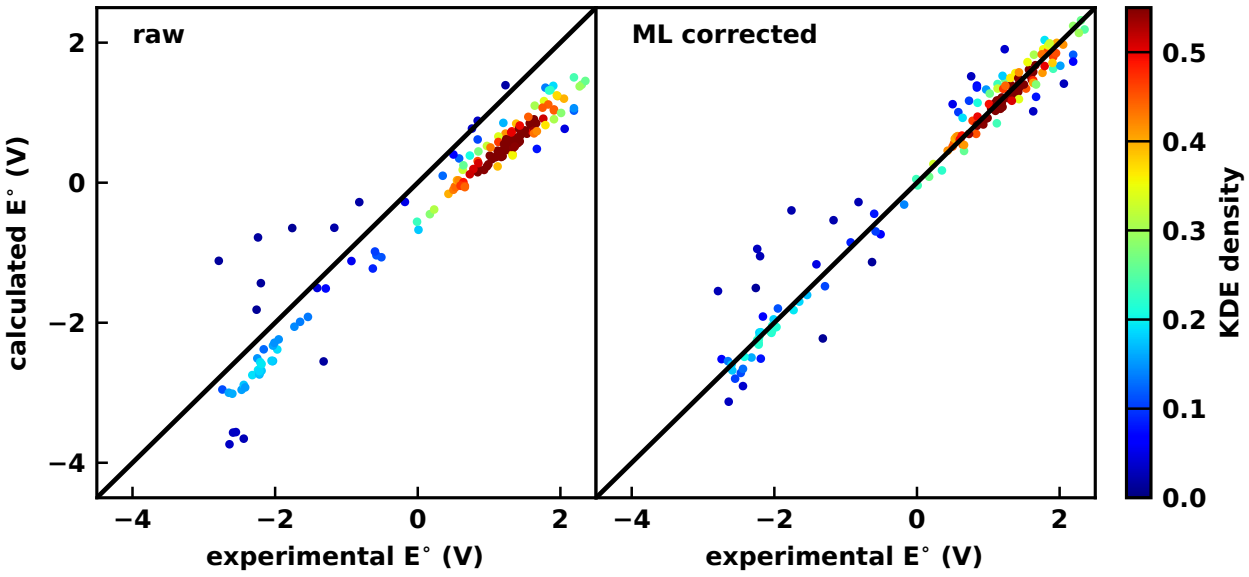

Figure 7: Accuracy of the predicted redox potentials with the explicit solvent approach with the miscrosolvated cluster embedded in C-PCM. (left) Raw results without ML correction show a systematic bias. (right) Results after ML correction with RF.

corrections, as shown in Figure 7. All tested ML models significantly reduce the test set MAE, with the best-performing RF reducing the MAE from 0.64 V to 0.19 V (Table 3). The number of outliers stays almost unchanged after applying all ML corrections. The comparison of the error histograms for explicit and implicit solvation is shown in SI Figure S11. Before ML correction, the explicit solvent approach has a higher systematic bias than the implicit solvent approach, which is significantly reduced by the ML correction. The ML features with the highest importance are the raw predicted redox potential, the nuclear repulsion energy, and the C-PCM solvation energy (SI Figure S9).

We then compare and contrast the performance of the explicit and implicit solvent approaches on the same subset of 165 organic systems. Although previous small batch studies indicate that the computationally more expensive explicit solvent approach tends to be more accurate than the implicit solvent approach, our data show exceptions. Before ML correction, the explicit solvent approach has a larger MAE (0.64 V) than the implicit solvent approach (0.40 V) due to a larger systematic bias. However, it produces fewer outliers (3 vs. 11), which can be attributed to the

more accurate representation of solvent-solute interactions in explicit solvation. After ML correction, the explicit solvent approach has a significantly improved MAE (0.19 V) similar to the implicit solvent approach (0.20 V), but its advantage in generating fewer outliers disappears due to the significant reduction of outliers for the implicit solvent approach. These counter-intuitive observations motivate us to further investigate the errors in our explicit solvent redox potential calculations in Section 4.6.

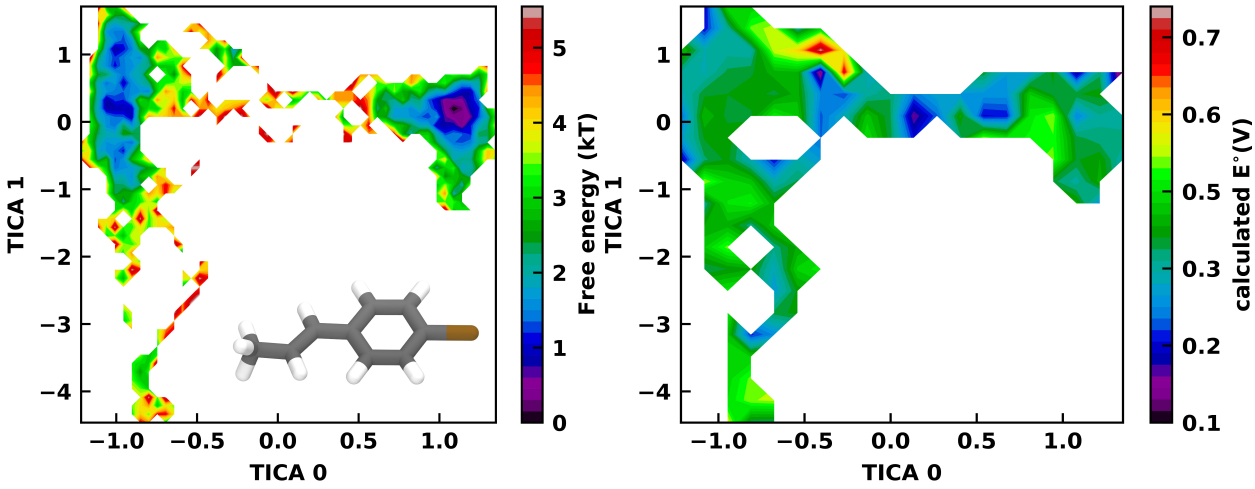

Figure 8: Dimension reduced landscape for system 29. (left) Free energy landscape along the slowest two TICA dimensions shows three minima. (right) The averaged redox potential along the slowest two TICA dimensions. The redox potential varies significantly across the conformational space.

4.6 Analysis of error sources in explicit solvent calculations

To understand why the explicit solvent approach did not outperform the implicit solvent model as expected, we further analyze potential sources of errors, and which systems are more likely to suffer from these errors. A typical error source for explicit solvent model is the sampling of redox couple configurations, so we picked a representative system (system 29) and investigated the variation of its calculated redox potential projected onto a TICA dimension-reduced landscape of the solute conformations (Figure 8). The projected redox potential is estimated only for the oxidized charge state since the free energy landscapes for the two charge states are generally different and

separate. Additional systems are shown in SI Figures S6-7. The dimension-reduced landscape is generated from the QM/MM trajectory used for TI calculation of redox potential, as described in Section 3.5. The free energy landscape of system 29 has three local free energy minima, with a fast transition time in the femtosecond range (Figure 8, left). The sparse sampling of the transition regions indicates that longer QM/MM trajectories are necessary to converge the sampling. Although conformations around the local minima have similar redox potential values of about 0.4 V, the transition region between the minima has some conformations with redox potential values as high as 0.7 V or as low as 0.1 V (Figure 8, right). For systems with multiple local minimas, insufficient sampling or analyzing only one conformation as in the implicit solvation approach can be one of the contributions to the observed redox potential errors and outliers.

Another potential error source is the LR approximation utilized in TI for the explicit solvent redox potential calculation. We compare the estimated reorganization energies λ^{St} and λ^{var} to validate the LR approximation. For the 165 explicit solvated systems calculated, the median λ^{St} is 0.37 V and the median λ^{var} is 0.30 V. The small difference between the two estimates indicates that the LR approximation is generally valid for the investigated systems. Histograms of the reorganization energies show that most systems have λ^{St} and λ^{var} below 0.5 V (SI Figure S8), similar to reported values of previous small batch studies.[33] However, some outlier systems have λ^{St} or λ^{var} up to around 2.5 V. For most systems the λ^{St} and λ^{var} show similar values, but we observe a systematic bias of λ^{var} smaller than λ^{St} by around 0.04 V. We also observe outlier systems where the λ^{St} and λ^{var} don't match, possibly caused by the limited number of snapshots utilized. Differences between λ^{var} and λ^{St} have been previously been reported for metalloproteins,[31,63] with non-parabolic shapes of the free-energy surfaces,[63] polarizability,[64] nonergodic effects[65] and inner-sphere/solute or outer-sphere/solvent effects[66] discussed as causes.

Furthermore, we observe a correlation between the size of reorganization energy and the size of redox potential errors: systems with higher reorganization energies tend to have higher redox potential errors. The ML corrected MAE for systems with $\lambda^{\mathrm{St}} > 0.5$ V is 0.38 V, significantly higher than the 0.15 V MAE for systems where $\lambda^{\mathrm{St}} < 0.5$ V. Alternatively utilizing λ^{var} as the

criterion gives consistent results (SI Table S1).

Finally, we examine the counterions as a possible source of error. The counterions are originally added for neutralizing the charge of the system during MM dynamics in the periodic boundary condition. For most systems, the counterions stayed far from the solute and were not included in the microsolvated clusters used for TI. However, for some solutes with higher net charges, the counterions may migrate to be close to the solute and be included in the TI snapshots. Although we suspect the electrostatic interaction between the counterions and the solute may cause artifacts in the calculated redox potential, we see only a small increase of MAE when counterions are present in the microsolvated clusters (SI Table S1).

5 Conclusion

This work exploits machine learning to reduce the errors relative to experimental measurements in redox potential calculations in both implicit and explicit solvents.

For the implicit solvent approach, we apply ML corrections to mitigate the systematic biases and reduce the number of outliers. This approach enabled us to reach an MAE of 0.21 V for the OROP dataset and 0.43 V for the OMROP dataset without any systems excluded, demonstrating improved accuracy compared to previously reported calculations on the same datasets without ML correction.[25] More importantly, the ML correction decreases the sensitivity of predicted redox potential values to functional choice, a long-standing challenge affecting the redox potential prediction accuracy for large, diverse datasets. With more experimental redox potential data points reported and larger experimental datasets available in the future, our ML corrections are expected to achieve even better performance through model retraining.

For the explicit solvent approach, we embedded the microsolvated clusters in C-PCM to reduce the computational costs. The redox potential calculated with the combined implicit/explicit model converges with a 4 Å solvent shell, significantly smaller than the around 10 Å shell required by the explicit-solvent-only counterpart. The resulting acceleration allowed rapid explicit solvent

redox potential calculation for 165 OROP compounds, with a smaller number of outliers than implicit solvent calculations on the same set of systems. We then adapted the aforementioned ML correction to these combined implicit/explicit calculations to reduce potential artifacts brought in by C-PCM, and obtained very similar performance compared to the implicit solvent approach after ML correction.

Finally, we analyzed the potential error sources of the explicit solvent approach. Dimension reduction analysis of explicitly solvated trajectories indicates a primary source of errors to be the insufficient thermodynamic sampling of solute conformations. Although the explicit solvent approach overcomes the drawback of the implicit solvent approach that considers only one optimized geometry, solutes with multiple minima may not have been sufficiently sampled due to limited trajectory lengths. The second error source is the limited validity of the LR approximation made in our TI protocol, as revealed by the discrepancy between the reorganization energies λ^{St} and λ^{var} in our dataset. We observed that counterions are not a large error source since only a small difference in accuracy was observed between the groups with and without counterions.

We expect that the ML-based correction strategies proposed in this work will enable rapid curation of computational redox potential datasets with an improved agreement with experimental measurements. The increased robustness to QM method choice will allow automated workflows for redox potential calculation without post-adjusted, system-specific calculation parameters. These will enhance the accuracy and efficiency of high-throughput computational design and discovery in catalysis and energy storage.

Acknowledgement

This work was supported by start-up funds provided by Emory University. This work used the Extreme Science and Engineering Discovery Environment[67] (XSEDE) Bridges-2 at Pittsburgh Supercomputing Center through allocation CHE200099, which is supported by National Science Foundation grant number ACI-1548562. The authors would like to thank Prof. Carlos Jaime for

useful discussions about acetonitrile force field parameters.

Supporting Information Available

See the supplementary material for additional implicit redox potential plots; additional redox potential projections; reorganization energies, ML results; hyperparameter ranges; description of ZIP file (PDF) and the geometry optimized xyz coordinates for all structures; the raw and ML corrected redox potential predictions for both implicit and explicit solvation; and utilized system subsets (ZIP).

References

(1) Prier, C. K.; Rankic, D. A.; Macmillan, D. W. C. Visible Light Photoredox Catalysis with Transition Metal Complexes: Applications in Organic Synthesis. **2013**,

(2) Connelly, N. G.; Geiger, W. E. *Chemical Redox Agents for Organometallic Chemistry*; 1996.

(3) Wang, W.; Sprenkle, V. *Energy storage: Redox flow batteries go organic*; 2016.

(4) Janet, J. P.; Ramesh, S.; Duan, C.; Kulik, H. J. Accurate Multiobjective Design in a Space of Millions of Transition Metal Complexes with Neural-Network-Driven Efficient Global Optimization. *ACS Central Science* **2020**, *13*, 2.

(5) Jinich, A.; Sanchez-Lengeling, B.; Ren, H.; Harman, R.; Aspuru-Guzik, A. 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 Central Science* **2019**, *5*, 1199–1210.

(6) Marenich, A. V.; Ho, J.; Coote, M. L.; Cramer, C. J.; Truhlar, D. G. Computational electrochemistry: Prediction of liquid-phase reduction potentials. *Physical Chemistry Chemical Physics* **2014**, *16*, 15068–15106.

(7) Klamt, A.; Schuurmann, G. Cosmo: a New Approach to Dielectric Screening in Solvents with Explicit Expressions for the Screening Energy and Its Gradient. *J. Chem. Soc., Perkin Trans. 2* **1993**, 799.

(8) Barone, V.; Cossi, M. Quantum Calculation of Molecular Energies and Energy Gradients in Solution by a Conductor Solvent Model. *J. Phys. Chem. A* **1998**, *102*, 1995.

(9) Amovilli, C.; Barone, V.; Cammi, R.; Cancès, E.; Cossi, M.; Mennucci, B.; Pomelli, C. S.; Tomasi, J. Recent Advances in the Description of Solvent Effects with the Polarizable Continuum Model. *Advances in Quantum Chemistry* **1998**, *32*, 227–261.

(10) Lange, A. W.; Herbert, J. M. A smooth, nonsingular, and faithful discretization scheme for polarizable continuum models: The switching/Gaussian approach. *Journal of Chemical Physics* **2010**, *133*.

(11) Truong, T. N.; Stefanovich, E. V. A New Method for Incorporating Solvent Effect into the Classical, Ab Initio Molecular Orbital and Density Functional Theory Frameworks for Arbitrary Shape Cavity. *Chem. Phys. Lett.* **1995**, *240*, 253.

(12) Mennucci, B.; Cancès, E.; Tomasi, J. Evaluation of Solvent Effects in Isotropic and Anisotropic Dielectrics and in Ionic Solutions with a Unified Integral Equation Method: Theoretical Bases, Computational Implementation, and Numerical Applications. *J. Phys. Chem. B* **1997**, *101*, 10506.

(13) Cancès, E.; Mennucci, B.; Tomasi, J. A New Integral Equation Formalism for the Polarizable Continuum Model: Theoretical Background and Applications to Isotropic and Anisotropic Dielectrics. *J. Chem. Phys.* **1997**, *107*, 3032.

(14) Tomasi, J.; Mennucci, B.; Cancès, E. The IEF Version of the PCM Solvation Method: an Overview of a New Method Addressed to Study Molecular Solutes at the QM Ab Initio Level. *J. Mol. Struct.: THEOCHEM* **1999**, *464*, 211.

(15) Klamt, A.; Jonas, V.; Bürger, T.; Lohrenz, J. C. Refinement and parametrization of COSMO-RS. *Journal of Physical Chemistry A* **1998**, *102*, 5074–5085.

(16) Matsui, T.; Kitagawa, Y.; Shigeta, Y.; Okumura, M. A density functional theory based protocol to compute the redox potential of transition metal complex with the correction of pseudo-counterion: General theory and applications. *Journal of Chemical Theory and Computation* **2013**, *9*, 2974–2980.

(17) Bím, D.; Rulíšek, L.; Srnec, M. Accurate Prediction of One-Electron Reduction Potentials in Aqueous Solution by Variableerature H-Atom Addition/Abstraction Methodology. *Journal of Physical Chemistry Letters* **2016**, *7*, 7–13.

(18) Noodleman, L.; Case, D. A.; Mouesca, J. M.; Lamotte, B. Valence electron delocalization in polynuclear iron-sulfur clusters. *Journal of Biological Inorganic Chemistry* **1996**, *1*, 177–182.

(19) Wang, L. P.; Van Voorhis, T. A polarizable QM/MM explicit solvent model for computational electrochemistry in water. *Journal of Chemical Theory and Computation* **2012**, *8*, 610–617.

(20) Zeng, X.; Hu, H.; Hu, X.; Cohen, A. J.; Yang, W. Ab initio quantum mechanical/molecular mechanical simulation of electron transfer process: Fractional electron approach. *The Journal of Chemical Physics* **2008**, *128*, 124510.

(21) Li, J.; Fisher, C. L.; Chen, J. L.; Bashford, D.; Noodleman, L. Calculation of Redox Potentials and pKa Values of Hydrated Transition Metal Cations by a Combined Density Functional and Continuum Dielectric Theory. *Inorganic Chemistry* **1996**, *35*, 4694–4702.

(22) Basdogan, Y.; Groenenboom, M. C.; Henderson, E.; De, S.; Rempe, S. B.; Keith, J. A. Machine Learning-Guided Approach for Studying Solvation Environments. *Journal of Chemical Theory and Computation* **2020**, *16*, 633–642.

(23) Zhang, J.; Zhang, H.; Wu, T.; Wang, Q.; Van Der Spoel, D. Comparison of Implicit and Explicit Solvent Models for the Calculation of Solvation Free Energy in Organic Solvents. *Journal of Chemical Theory and Computation* **2017**, *13*, 1034–1043.

(24) Boereboom, J. M.; Fleurat-Lessard, P.; Bulo, R. E. Explicit Solvation Matters: Performance of QM/MM Solvation Models in Nucleophilic Addition. *Journal of Chemical Theory and Computation* **2018**, *14*, 1841–1852.

(25) Neugebauer, H.; Bohle, F.; Bursch, M.; Hansen, A.; Grimme, S. Benchmark Study of Electrochemical Redox Potentials Calculated with Semiempirical and DFT Methods. *J. Phys. Chem. A* **2020**, *124*, 20.

(26) Gagne, R. R.; Koval, C. A.; Lisensky, G. C. Ferrocene as an internal standard for electrochemical measurements. *Inorganic Chemistry* **1980**, *19*, 2854–2855.

(27) Pavlishchuk, V. V.; Addison, A. W. Conversion constants for redox potentials measured versus different reference electrodes in acetonitrile solutions at 25 C. *Inorganica Chimica Acta* **2000**, *298*, 97–102.

(28) Konezny, S. J.; Doherty, M. D.; Luca, O. R.; Crabtree, R. H.; Soloveichik, G. L.; Batista, V. S. Reduction of systematic uncertainty in DFT redox potentials of transition-metal complexes. *Journal of Physical Chemistry C* **2012**, *116*, 6349–6356.

(29) Roy, L. E.; Jakubikova, E.; Guthrie, M. G.; Batista, E. R. Calculation of One-Electron Redox Potentials Revisited. Is It Possible to Calculate Accurate Potentials with Density Functional Methods? *The Journal of Physical Chemistry A* **2009**, *113*, 6745–6750.

(30) Kirkwood, J. G. Statistical mechanics of fluid mixtures. *The Journal of Chemical Physics* **1935**, *3*, 300–313.

(31) Lebard, D. N.; Matyushov, D. V. Ferroelectric hydration shells around proteins: Electrostatics of the protein-water interface. *Journal of Physical Chemistry B* **2010**, *114*, 9246–9258.

(32) Blumberger, J. Recent Advances in the Theory and Molecular Simulation of Biological Electron Transfer Reactions. *Chemical Reviews* **2015**, *115*, 11191–11238.

(33) Falbo, E.; Penfold, T. J. Redox Potentials of Polyoxometalates from an Implicit Solvent Model and QM/MM Molecular Dynamics. *Journal of Physical Chemistry C* **2020**, *124*, 15045–15056.

(34) Marcus, R. A.; Sutin, N. Electron transfers in chemistry and biology. 1985.

(35) Hecht, M.; Fawcett, W. R. Electrochemistry of [V(III)EDTA]- in ethylene glycol-water mixtures. 1. Thermodynamic and transport properties: Solvation of the reactant and product. *Journal of Physical Chemistry* **1996**, *100*, 14240–14247.

(36) Kruger, H. J.; Holm, R. H. Stabilization of Trivalent Nickel in Tetragonal NiS4N2 and NiN6 Environments: Synthesis, Structures, Redox Potentials, and Observations Related to [NiFe]-Hydrogenases. *J. Am. Chem. Soc* **1990**, *112*, 2955–2963.

(37) Götz, A. W.; Wölfle, T.; Walker, R. C. *Annual Reports in Computational Chemistry*; Elsevier BV, 2010; Vol. 6; pp 21–35.

(38) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *Journal of computational chemistry* **2011**, *32*, 1456–1465.

(39) Hay, P. J.; Wadt, W. R. Ab initio effective core potentials for molecular calculations. Potentials for the transition metal atoms Sc to Hg. *The Journal of Chemical Physics* **1985**, *82*, 270–283.

(40) Wang, L.-P.; Song, C. Geometry optimization made simple with translation and rotation coordinates. *The Journal of chemical physics* **2016**, *144*, 214108.

(41) Saunders, V.; Hillier, I. A "Level–Shifting" method for converging closed shell Hartree–Fock wave functions. *International Journal of Quantum Chemistry* **1973**, *7*, 699–705.

(42) York, D. M.; Karplus, M. A smooth solvation potential based on the conductor-like screening model. *Journal of Physical Chemistry A* **1999**, *103*, 11060–11079.

(43) Liu, F.; Luehr, N.; Kulik, H. J.; Martínez, T. J. Quantum Chemistry for Solvated Molecules on Graphical Processing Units Using Polarizable Continuum Models. *Journal of Chemical Theory and Computation* **2015**, *11*, 3131–3144.

(44) Bondi, A. van der Waals Volumes and Radii. *The Journal of Physical Chemistry* **1964**, *68*, 441–451.

(45) Batsanov, S. S. Van der Waals radii of elements. *Inorganic Materials* **2001**, *37*, 871–885.

(46) Martínez, L.; Andrade, R.; Birgin, E. G.; Martínez, J. M. PACKMOL: A package for building initial configurations for molecular dynamics simulations. *Journal of Computational Chemistry* **2009**, *30*, 2157–2164.

(47) D.A. Case K. Belfon, I. Y. B.-S. S. R. B. D. S. C. T. E. C. I. I. I. V. W. D. C. T. A. D. R. E. D. G. G. M. K. G. H. G. A. W. G. R. H. S. I. S. A. I. K. K. A. K. R. K. T. K. T. S. L. S. L. P. L. C. L. J. L. T. L. R.; Kollman, P. A. *Amber 2020*; 2020.

(48) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *Journal of Computational Chemistry* **2004**, *25*, 1157–1174.

(49) Grabuleda, X.; Jaime, C.; Kollman, P. A. *Molecular Dynamics Simulation Studies of Liquid Acetonitrile: New Six-Site Model*; 2000; Vol. 21; pp 901–908.

(50) Li, P.; Merz, K. M. MCPB.py: A Python Based Metal Center Parameter Builder. *Journal of Chemical Information and Modeling* **2016**, *56*, 599–604, PMID: 26913476.

(51) Götz, A. W.; Clark, M. A.; Walker, R. C. An extensible interface for QM/MM molecular dynamics simulations with AMBER. *Journal of Computational Chemistry* **2014**, *35*, 95–108.

(52) Isborn, C. M.; Götz, A. W.; Clark, M. A.; Walker, R. C.; Martínez, T. J. Electronic absorption spectra from MM and ab initio QM/MM molecular dynamics: Environmental effects

on the absorption spectrum of photoactive yellow protein. Journal of Chemical Theory and Computation. 2012; pp 5092–5106.

(53) Weinreich, J.; Browning, N. J.; Von Lilienfeld, O. A. Machine learning of free energies in chemical compound space using ensemble representations: Reaching experimental uncertainty for solvation. *Journal of Chemical Physics* **2021**, *154*, 134113.

(54) Breiman, L. Random forests. *Machine learning* **2001**, *45*, 5–32.

(55) Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics* **2001**, 1189–1232.

(56) Murphy, K. P. *Machine learning: a probabilistic perspective*; MIT press, 2012.

(57) Hinton, G. E. *Machine learning*; Elsevier, 1990; pp 555–610.

(58) Molgedey, L.; Schuster, H. G. Separation of a mixture of independent signals using time delayed correlations. *Physical Review Letters* **1994**, *72*, 3634–3637.

(59) Global optimization of chemical cluster structures: Methods, applications, and challenges. 2021.

(60) Baik, M. H.; Friesner, R. A. Computing redox potentials in solution: Density functional theory as a tool for rational design of redox agents. *Journal of Physical Chemistry A* **2002**, *106*, 7407–7412.

(61) Roy, L. E.; Jakubikova, E.; Graham Guthrie, M.; Batista, E. R. Calculation of one-electron redox potentials revisited. Is it possible to calculate accurate potentials with density functional methods? *Journal of Physical Chemistry A* **2009**, *113*, 6745–6750.

(62) Li, Y.; Hartke, B. Assessing solvation effects on chemical reactions with globally optimized solvent clusters. *ChemPhysChem* **2013**, *14*, 2678–2686.

(63) Seyedi, S. S.; Waskasi, M. M.; Matyushov, D. V. Theory and Electrochemistry of Cytochrome c. *Journal of Physical Chemistry B* **2017**, *121*, 4958–4967.

(64) Dinpajooh, M.; Martin, D. R.; Matyushov, D. V. Polarizability of the active site of cytochrome c reduces the activation barrier for electron transfer. *Scientific Reports 2016 6:1* **2016**, *6*, 1–10.

(65) Jiang, X.; Futera, Z.; Blumberger, J. Ergodicity-Breaking in Thermal Biological Electron Transfer? Cytochrome C Revisited. *The Journal of Physical Chemistry B* **2019**, *123*, 7588–7598.

(66) Weaver, M. N.; Janicki, S. Z.; Petillo, P. A. Ab initio calculation of inner-sphere reorganization energies of arenediazonium ion couples. *Journal of Organic Chemistry* **2001**, *66*, 1138–1145.

(67) Towns, J.; Cockerill, T.; Dahan, M.; Foster, I.; Gaither, K.; Grimshaw, A.; Hazlewood, V.; Lathrop, S.; Lifka, D.; Peterson, G. D. XSEDE: Accelerating Scientific Discovery. *Comput. Sci. Eng.* **2014**, *16*, 62.