Correlation between the Stability of Substituted Cobaltocenium and Molecular Descriptors

Shehani T. Wetthasinghe,[†] Chunyan Li,[‡] Huina Lin,[†] Tianyu Zhu,[†] Chuanbing Tang,[†] Vitaly Rassolov,[†] Qi Wang,[*,‡] and Sophya Garashchuk[*,†]

[†]*Department of Chemistry and Biochemistry, University of South Carolina, Columbia, SC*

[‡]*Department of Mathematics, University of South Carolina, Columbia, SC*

E-mail: qwang@mailbox.sc.edu; garashchuk@sc.edu

Abstract

Metallocenium cations, used as a component in an anion exchange membrane of a fuel cell, demonstrate excellent thermal and alkaline stability, which can be improved by the chemical modification of the cyclopentadienyl rings with substituent groups. In this work the relation between the bond dissociation energy (BDE) of the cobaltocenium ($CoCp_2^+$) derivatives, used as a measure of the cation stability, with chemistry-informed descriptors obtained from the electronic structural calculations, is established. The analysis of 12 molecular descriptors for 118 derivatives reveals a non-linear dependence of the BDE on the electron donating-withdrawing character of substituent groups coupled to the energy of frontier molecular orbitals. A chemistry-informed feed-forward neural network trained using k-fold cross-validation over the modest dataset is able to predict the BDE from the molecular descriptors with the mean absolute error of about 1 kcal/mol. The theoretical analysis suggests some promising modifications of cobaltocenium for experimental research. The results demonstrate that even for modest data sets incorporation of the chemistry knowledge into the neural network architecture, e.g.

through mindful selection and screening of the descriptors and their interactions, paves the way to gain new insight into molecular properties.

1 Introduction

Polyelectrolyte-based anion-exchange membranes (AEMs) have a broad range of applications in energy conversion and storage devices, such as alkaline fuel cells, redox flow batteries, electrodialysis and electrolyzers,[1–3] due to their ability to bind small ions (like hydroxide, halide, carbonate ions) and to their redox stability under operational potentials. Yet, the development of advanced membranes with sufficient chemical stability and high anionic conductivity under extremely harsh, highly basic conditions in alkaline fuel cells, remains an outstanding challenge. Derivatives of cobaltocenium, i.e. of [bis(cyclopentadienyl)cobalt(III)] $CoCp_2^+$, where Cp:=cyclopentadienyl, are regarded as promising AEM components (Fig. 1), because of their excellent thermal and alkaline stability under operating conditions of fuel cells.

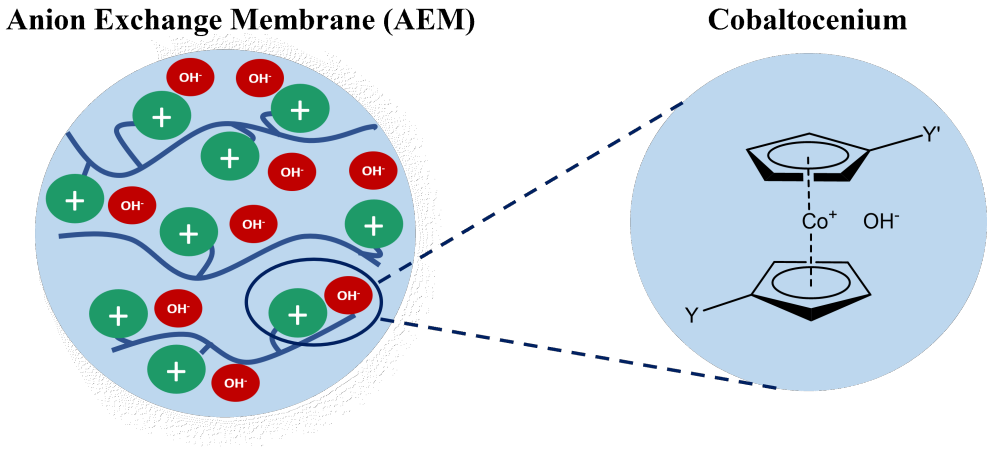

Figure 1: The anion exchange membrane (AEM) of a fuel cell with the cobaltocenium cation in a polymer side chain. Modifications of the cobaltocenium cation by one or two substituent groups, Y and Y′, are considered, including Y′ =H (mono-substituted set), Y′ =Y (di-substituted set) and Y′ = CH_3 (hetero-substituted set).

For example, permethyl-$CoCp_2^+$ has been found ultra-stable compared to the unsubsti-

tuted $CoCp_2^+$, i.e. 8.5% degradation after six weeks (at 140° in 1 M $NaOD/D_2O$ solution) for the former, while the latter compound fully degraded after one week.[4] Degradation of the substituted cobaltocenium cations has been also measured in the Tang group[5] (Cl^- as a counterion, 80° in 5 M KOD/D_2O solution) for five substituted species with one, two and four methyl groups, and with either one or two tert-butyl groups per Cp ring. Increase in the number of the substituent groups has led to increased stability of the cations for both methyl and tert-butyl derivatives sequences, which is consistent with the electronic effect; in addition, superior stability of the cobaltocenium with four tert-butyl groups compared to the octa-methyl cobaltocenium (8.2% and 18.5% cation loss after 1025 hours, respectively) has been attributed to the steric effect of the tert-butyl group. Given that the synthesis and characterization of the full range of possible derivatives are impractical, to gain further insight into the substituent effect on stability useful in IEM design, we have performed a computational modeling study, described below.

Our study is designed around experimentally relevant substitutions considered by the Tang group, yielding a set of about 118 species. The resulting dataset is small by the machine-learning standards, but a few promising approaches of combining the machine learning techniques with small data sets have been reported recently: some exploit the power of transfer learning[6,7] (a small dataset augments a related but different large dataset), and some incorporate the domain knowledge into the neural network design, such as integrating deep learning algorithms with the d-band theory of chemisorption at transition-metal surfaces.[8] The latter approach, though less general, leads to interpetable machine learning model, an obviously desirable feature.[9] In some other examples of NN applications based on small datasets[10,11] (down to 211 and 176 experimental samples, respectively), the descriptor screening and careful Deep NN and Artificial NN construction are performed to predict OH^- conductivity and stability of the anion exchange membranes depending on the polymer composition. In this work, focused specifically on cobaltocenium derivatives, we take a simpler approach of imparting chemical knowledge at an early stage through the selection and

screening of the molecular descriptors, subsequently used in neural network models. This strategy mitigates overfitting despite the small dataset size, and enables chemical insight into the stability of cobaltocenium-hydroxide complexes.

The remainder of the paper is organized as follows. The system, its molecular representation and theoretical model are described in Section 2. The molecular descriptors and the conventional and neural network analysis are discussed in Section 3. Summary and outlook are presented in Section 4.

2 The computational model and methods

To generate a coherent dataset we start by establishing a well-defined and practical computational model for the cobaltocenium derivatives relevant to experiments. The process central to this model is the dissociation of the $CoCp_2^+ \cdots OH^-$ complexes in aqueous environment; one or both cyclopentadienyl (Cp) rings are modified by one out of 42 substituent groups. The corresponding bond dissociation energy (BDE) is taken as a measure of the cobaltocenium stability, which is supported by the correlation of the computed BDE with the measured degradation of the cobaltocenium derivatives.[5] The BDEs along with twelve molecular properties (or features) are obtained from the electronic structure calculations, and their correlations are analyzed firstly using the conventional linear regression analysis. The results are rationalized in chemical terms and further used to develop the non-linear Least Squares Fit (LSF) and the chemistry-informed feed-forward neural network (CIFNN) models whose predictive performance is assessed through validation. The electronic structure calculations are performed with Q-Chem 5.3[12] and Spartan'18;[13] the deep neural network models are implemented with PyTorch.[14]

In general, performance of an AEM depends on many factors including its morphology, hydration level, ion concentration and temperature. Its fundamental component is the polyelectrolyte consisting of the cations tethered or embedded into a polymeric chain and

of the mobile hydroxide ions, all in the aqueous environment. The positive charge on the cobaltocenium is balanced by the hydroxide anion, forming $CoCp_2^+ \cdots OH^-$ complexes which should dissociate as $CoCp_2OH \rightarrow CoCp_2^+ + OH^-$, enabling the desired OH^- transport. An alternative process which degrades an AEM is the bond dissociation of the cobaltocenium cation itself, such as

$$CoCp_2OH \rightarrow Cp + CoCpOH. \quad (1)$$

Thus, the stability of the complexes with respect to the cobaltocenium dissociation is essential to a functioning AEM, and understanding the effect of possible synthetic modification of the cobaltocenium as a tool of increasing its stability is highly desirable. The relevant BDE is,

$$BDE = E(CoCpOH) + E(Cp) - E(CoCp_2OH). \quad (2)$$

Here and throughout $CoCp_2^+$ and $CoCp^+$ are referred to the cobaltocenium with or without substitutions.

We have considered a set of mono- and di-substituted cobaltocenium cations covering a range of electron-donating and electron-withdrawing substituent groups, Y, shown in Fig. S1 and listed on the horizontal axes in Fig. 3. Y and Y′ label the substituent groups on the two Cp rings (one per Cp) as illustrated on Fig. 1. Altogether, the dataset consists of 40 mono-substituted species (Y′=H) including the pristine cobaltocenium, 38 di-substituted species of the type (Y′=Y) and 40 hetero-substituted species (Y′=CH_3).

To examine the electronic structure of a few hundred complexes and dissociation fragments one needs an atomistic model which is realistic, unambiguous and computationally practical. A basic DFT gas-phase description of the cobaltocenium-hydroxide complex results in a deprotonation of one of the Cp rings, which is not expected in the aqueous IEM environment. The addition of up to six explicit water molecules stabilizes the $CoCp_2^+OH^-$ complex, but the BDE values show significant variation on their number, most likely due to increased chances of trapping in the local minima (Fig. 2). Thus, for consistency within

the computed data, we settle on representing the aqueous environment within the Polarized Continuum Model (PCM),[15] which also prevents deprotonation of Cp, and reduces ambiguity in the geometry minimization. The van der Waals radius of Co was set to 2.0 Å,[16] the Q-Chem default radii have been used otherwise. Additional information on the $CoCp_2^+OH^-$ dissociation in PCM is given in Table S2.

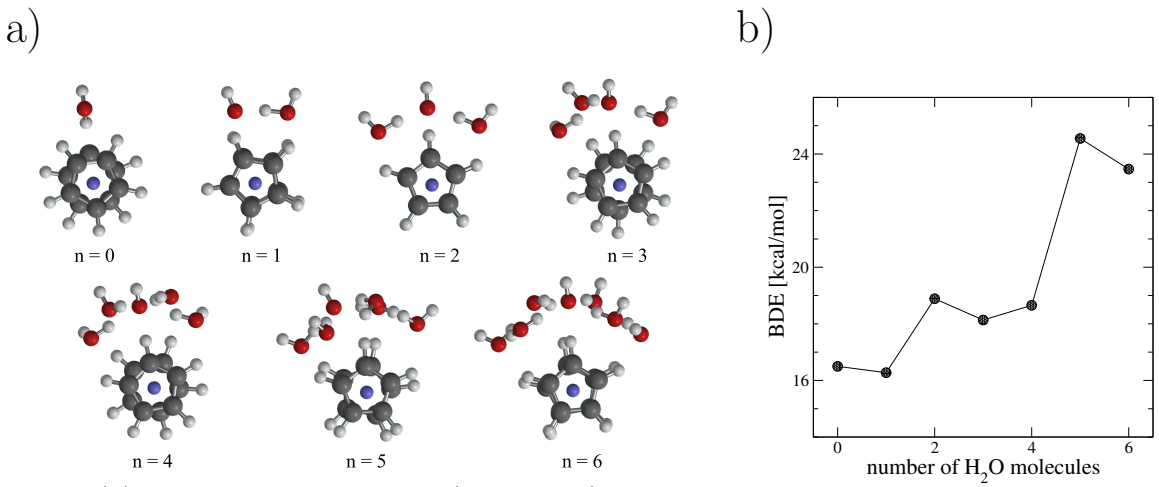

Figure 2: (a) Optimized structures (top views) for systems consisting of the unsubstituted $CoCp_2^+$, OH^- and n water molecules $n = [0, 6]$ and (b) the corresponding BDEs as a function of n, computed at B3LYP-D3/6-31G* level in gas phase.

The density functional theory (DFT) is employed for practical reasons. After testing several popular functionals and basis sets, the B3LYP-D3 functional[17–19] combined with m6-31G*[20] for Co and 6-31G** for C, O, H and other atoms are chosen: this reproduces the previously reported BDE trend[21] for the multi-substituted derivatives of $CoCp_2^+$ (Fig. S2) and yields the BDE of 110.4 kcal/mol for the unsubstituted $CoCp_2^+$ in gas phase, which is in better agreement with the experimental value of 105.4 ± 3.5 kcal/mol,[22] compared to the method of Ref.[21] Within this model, the mono-, di- and hetero-substituted cobaltocenium-hydroxide complexes dissociate into C_5H_4Y and $C_5H_4Y'CoOH$, where $Y' = \{H, Y, CH_3\}$ (exceptional case: when $Y=C(CH_3)_3$ and $Y'= CH_3$, it dissociates to C_5H_4Y' and C_5H_4YCoOH). In agreement with Ref.[22] the lowest energy spin-states for $CoCp_2^+$, $CoCp^+$ and Cp are the singlet, quartet and doublet, respectively. Additional calculations performed at the MP2 level (Table S1) yield the same spin-states.

3 Results and discussion

3.1 The molecular properties and their correlation to the BDE

The electronic structure calculation provides a wealth of information about a molecule, including those relevant to its stability. Since we work with a modest set of 118 molecules, we have carefully selected twelve chemically-meaningful properties, computed at the geometries optimized in solvent unless stated otherwise. Using abbreviations HOMO and LUMO for the highest occupied molecular orbital and the lowest unoccupied molecular orbitals, respectively, these properties, or *molecular descriptors*, include those of the substituted complexes, dissociation fragments and of the substituent groups themselves: (**1**) the solvation energy of $CoCp_2OH$, E_{solv}; (**2**) its dipole moment, μ; (**3**) the distance between Co and O of the hydroxide, d_{CoO}; (**4**) the distance between the center-of-mass for the carbons of two Cp rings, d_{Cp}; (**5**) the HOMO and (**6**) LUMO energies of the complexes, (**7**) the HOMO and (**8**) LUMO energies of the CoCpOH, (**9**) the HOMO and (**10**) LUMO of the Cp's. E_{solv} quantifies stabilization due to solvent; μ, d_{CoO} and d_{Cp} are the 'gross' features related to the reactivity and stability; the half-sum and the difference of the HOMO/LUMO energies are related to the Fermi-level and the chemical hardness both associated with the stability of a molecule. To characterize the electron donating (ED) and electron-withdrawing (EW) properties of the substituent groups and capture their effect on the complex stability, we also include (**11**) the polarizability, α, of a substituent and, finally, (**12**) the Hirshfeld charge, H, induced by a substituent on a 'probe' molecule. The latter, less-commonly-used property was considered for either benzene or Cp used as such probes. The Hirshfeld charge is defined as the sum of atomic charges on C_6H_5/C_5H_4 of a substituted benzene/Cp, and has been shown to be less basis-dependent compared to the Mulliken or Lowdin charges.[23] As seen in Fig. 3, both versions of the H-charge display a similar correlation with the generally accepted ranking of the ED/EW character of the substituents. The full cobaltocenium data set is listed in Tables S3-S5. Having a realistic chemically meaningful target accuracy for the

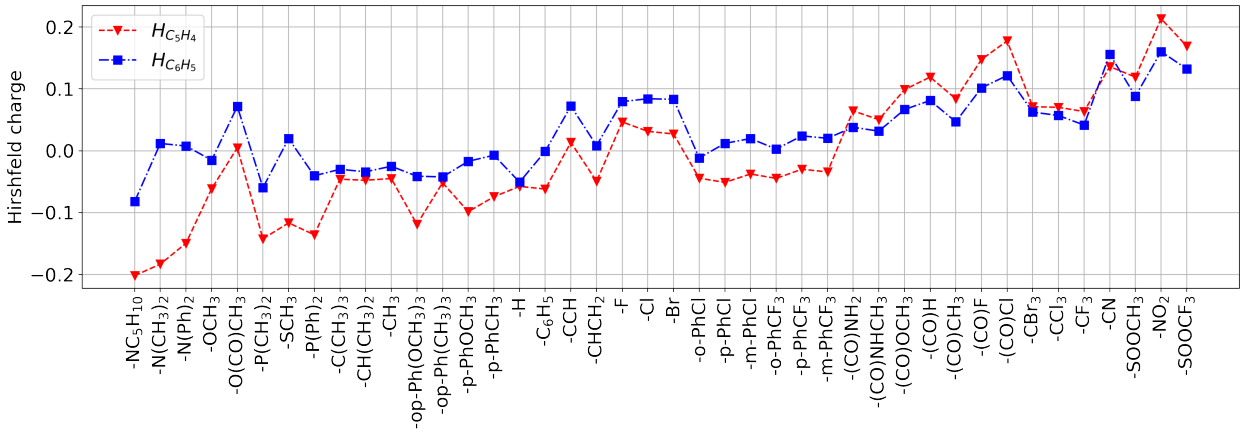

Figure 3: The sum of the Hirshfeld charges on C_6H_5 and C_5H_4 of C_6H_5Y and C_5H_4Y as a function of substituents, Y, arranged from the strongest electron donating to the strongest electron withdrawing effect according to the general organic chemistry.

BDE of $\sim$ 2 kJ/mol, 8 species with the BDEs below 3 kJ/mol have been excluded from the modeling (Table S6). An additional data set (Table S6) was generated for the di-substituted rhodocenium complexes ($RhCp_2OH$) employing fit-SBKJC effective core potential and the matching basis method with the B3LYP-D3 functional. The van der Waals radius of 2.0 Å has been used for Rh in the PCM.[16] The lowest energy spin-states for $RhCp_2^+$ and $RhCp^+$ are the singlet and doublet, respectively. The BDEs for all substituted $RhCp_2^+OH^-$ complexes are higher by about 20 kJ/mol than the cobaltocenium counterparts.

3.2 Modeling stability of complexes

To unravel the BDE dependence on molecular descriptors we start by computing the Pearson linear correlation coefficients for the 38 di-substituted species. The coefficients are listed in Table 1 using H-charge on benzene. For the cobaltocenium the highest correlation of the BDE is with the LUMO (0.531) energy and with the Hirshfeld charge (0.625), but surprisingly not with the HOMO. These correlations are even higher (0.836 and 0.867 respectively) for the rhodocenium complex. This can be understood as follows: the considered substituent groups are charge-neutral and, therefore, relative to the electron-donating H atom, they have the EW effect.[23] Upon the substitution of the H-atom of the pristine

$CoCp_2^+$, an electron is pulled from the HOMO level of $CoCp_2^+$ which is 'transformed' into the LUMO level of the substituted $CoCp_2^+$. For the ED group, we would expect the opposite, i.e. an electron is pulled from an ED group to the LUMO of the pristine $CoCp_2^+$, correlating with the HOMO of the substituted complex. Therefore, the BDE is sensitive to the HOMO or LUMO energy *modulated* by H, and it is useful to incorporate a switching function of H into the functions modeling dependence of the BDE on the HOMO/LUMO energies. A logistic function $f(H)$ – effectively a one-layer network – with adjustable parameters k and H_0 –

$$f(H) = \frac{1}{1 + e^{-k(H-H_0)}}, \tag{3}$$

is used in the following Least Square Fit (LSF) models.[24]

Table 1: Correlation coefficients of the BDE with the descriptors for $CoCp_2^+OH^-$ and $RhCp_2^+OH^-$.

complex	E_{solv}	μ	d_{CoO}	d_{Cp}	E_{HOMO}	E_{LUMO}	α	H
$CoCp_2^+OH^-$	-0.213	-0.094	0.266	-0.009	-0.109	0.531	0.146	-0.625
$RhCp_2^+OH^-$	-0.179	0.002	0.240	0.334	0.392	0.836	0.490	-0.867

Since the full cobaltocenium data set includes mono-, di- and hetero-substituted species with distinct derivative groups Y and $Y' = \{H, Y, CH_3\}$ on the two Cp rings, some stability models include the HOMO/LUMO energy of the fragments. The following fitting functions

have been considered:

$$\begin{aligned}
\hat{y}_1 : &= c_0 + (c_1 + c_2 f(H^{(2)}))E^{(1)}_{\rm HOMO} + (c_3 + c_4 f(H^{(2)}))E^{(1)}_{\rm LUMO}, && (4)\\
\hat{y}'_1 : &= c_0 + c_1 E^{(1)}_{\rm HOMO} + c_2 E^{(1)}_{\rm LUMO} + c_3 H^{(2)}, && (5)\\
\hat{y}_2 : &= c_0 + \left(c_1 + c_2 f(H^{(2)})\right)E^{(2)}_{\rm HOMO} + \left(c_3 + c_4 f(H^{(2)})\right) E^{(2)}_{\rm LUMO}\\
&+ \left(c_5 + c_6 f(H^{(3)})\right) E^{(3)}_{\rm HOMO} + \left(c_7 + c_8 f(H^{(3)})\right) E^{(3)}_{\rm LUMO}, && (6)\\
\hat{y}'_2 : &= c_0 + c_1 E^{(2)}_{\rm HOMO} + c_2 E^{(2)}_{\rm LUMO} + c_3 E^{(3)}_{\rm HOMO} + c_4 E^{(3)}_{\rm LUMO} + c_5 H^{(2)} + c_6 H^{(3)}, && (7)\\
\hat{y}_3 : &= c_0 + c_1 E^{(1)}_{\rm HOMO} + c_2 E^{(1)}_{\rm LUMO}\\
&+ \left(c_3 + c_4 f(H^{(2)})\right) E^{(2)}_{\rm HOMO} + \left(c_5 + c_6 f(H^{(2)}\right) E^{(2)}_{\rm LUMO}\\
&+ \left(c_7 + c_8 f(H^{(3)})\right) E^{(3)}_{\rm HOMO} + \left(c_9 + c_{10} f(H^{(3)})\right) E^{(3)}_{\rm LUMO}. && (8)
\end{aligned}$$

In Eqs (4-8), the superscripts ($i = 1, 2, 3$) refer to $\mathrm{Co(C_5H_4Y')(C_5H_4Y)}$, and to $\mathrm{CoC_5H_4Y'}$ and $\mathrm{C_5H_4Y}$ fragments, respectively. The LSFs $\{\hat{y}'_i\}$ are linear, while the LSFs $\{\hat{y}_i\}$ incorporate the non-linear switching function of Eq. (3). Both versions of H (defined either on the benzene probe or on the Cp probe) are considered.

For each model, the parameters $\{c_1 \dots\}$ and $\{k^{(2)}, H_0^{(2)}, k^{(3)}, H_0^{(3)}\}$ for the LSF minimize the mean square error (MSE),

$$MSE = \frac{1}{K}\sum_k (y_k - \hat{y}_k)^2. \qquad (9)$$

The parameter values are listed in Tables S8 and S9. The LSFs of Eqs (4) and (5) are performed for the di-substituted derivatives (38 species), while LSFs of Eqs (6-8) are performed for the mono- and di-substituted derivatives (78 species), referred to as the *training* set of each case. The predictive properties of the fragment based LSFs are evaluated by computing the BDE of the hetero-substituted set of the derivatives (40 species), referred to as the *testing* set. The resulting errors, including the root mean square (RMS) error, RMS = $(\mathrm{MSE})^{1/2}$, are listed in Table 2.

Table 2: The mean square error (MSE) and the root mean square (RMS) error of the LSF models.

model	H	train MSE [(kcal/mol)^2]	train RMS [kcal/mol]	test MSE [(kcal/mol)^2]	test RMS [kcal/mol]
LSF1 $\hat{y}_1$	C_6H_5	31.850	5.644	-	-
	C_5H_4	35.280	5.940	-	-
LSF1 $\hat{y}'_1$	C_6H_5	35.205	5.933	-	-
	C_5H_4	41.199	6.419	-	-
LSF2 $\hat{y}_2$	C_6H_5	6.590	2.567	15.930	3.991
	C_5H_4	10.082	3.175	18.473	4.298
LSF2 $\hat{y}'_2$	C_6H_5	8.570	2.927	17.191	4.146
	C_5H_4	13.743	3.707	22.168	4.708
LSF3 $\hat{y}_3$	C_6H_5	5.294	2.301	11.347	3.369
	C_5H_4	7.956	2.821	11.299	3.361

Let's evaluate the performance of the three types of the LSF models, involving the frontier MO energies of the complex (LSF1), fragments (LSF2), and complex + fragments (LSF2). Our first observation is that H defined on benzene as a probe gives better accuracy than the one defined on Cp, which is rationalized as follows: while Cp is obviously more relevant to the cobaltocenium than benzene, being a radical Cp is unevenly disturbed by the substituent groups, and is therefore not as a good EW/ED property probe as the benzene ring. The second observation is that both, the training and the testing errors, are reduced as we go from LSF1 to LSF3. The third observation is that the RMS in models with the switching function are lower than that in the linear models by about 1 kcal/mol. The reduction of the training error is expected as we introduce additional fitting parameters. The reduction in testing MSE and RMS is encouraging, as it means that there is not much over-fitting. However, the training error is 2-3 times smaller than the testing error in all cases, which means that these models are not sufficiently accurate. Furthermore, in the LSF models the training and testing data sets are comprised of the distinct families of species, i.e. mono- and di-substituted for training, and hetero-substituted derivatives for testing. The accuracy of the fit and predictions is expected to improve upon bias-free partitioning of the data into training and testing sets. Both shortcomings – the limitations of the analytic form of the LSF and the training/testing partition bias – are overcome by the Chemistry-Informed

Feed-forward Neural Network (CIFNN).

The models LSF2 and LSF3 are generalized to the CIFNN models by putting the same molecular descriptors into a neural network structure, noticing that the non-linear LSF models, in fact, incorporate a single 'neuron' through the sigmoid function of H. The input variable X,

$$X = [\mathrm{E}^{(1)}_{\mathrm{HOMO}}, \mathrm{E}^{(1)}_{\mathrm{LUMO}}, \mathrm{E}^{(2)}_{\mathrm{HOMO}}, \mathrm{E}^{(2)}_{\mathrm{LUMO}}, \mathrm{E}^{(3)}_{\mathrm{HOMO}}, \mathrm{E}^{(3)}_{\mathrm{LUMO}}, H^{(2)}, H^{(3)}], \tag{10}$$

consists of 8 descriptors used in the LSF3 model. Then, the n-level CIFNN3 is given explicitly by

$$h_1 = W_1^T X + b_1, \tag{11}$$

$$h_j = \sigma\left(W_j h_{j-1} + b_j\right), j = 2, ..., n, \tag{12}$$

$$h_o = W_o^T h_n + b_o \in \mathbf{R}^N, \tag{13}$$

where $h_i, i = 1, ..., n$ is the i-th hidden layer, h_o is the output, N is the total number of derivatives in training set, W_o, W_i and $b_o, b_i, i = 1, ..., n$ are weights and bias of the model. The activation function in Eq. (12) is chosen among tanh and variants of the ReLU functions. CIFNN2 is obtained by assigning zero to variable $\mathrm{E}^{(1)}_{\mathrm{HOMO}}$ and $\mathrm{E}^{(1)}_{\mathrm{LUMO}}$. Applying CIFNN2 and CIFNN3 to the two version of H (computed for C_6H_5 and C_5H_4 probes) we develop a total of four neural network models. The BDEs and descriptors in the data set of 118 cobaltocenium samples are standardized using Z-score method.[25] The samples are randomly divided into the training set ($K = 106$) and the testing set (12). The k-fold cross validation[26] is used for a bias-free partitioning of the data via multiple random selections. By comparing the models obtained from different partitions (9 subsets as training set + 1 subset as testing set), the 10-fold cross validation process selects the best performing one as the final solution which presumably reduces the generalization error on new data. In the training process, the cost function is defined as the MSE of the BDE. In the cross-validation, the best model is

chosen according to the relative MSE (RMSE) on the test set,

$$\text{RMSE} := \frac{\sum_{k=1}^{K}(y_k - \hat{y}_k)^2}{\sum_{k=1}^{K}(y_k - \bar{y})^2},\ \bar{y} = \frac{1}{K}\sum_{k=1}^{K} y_k. \tag{14}$$

Two other metrics, mean absolute error (MAE) and mean relative error (MRE) are also used to evaluate the performance:

$$\text{MAE} = \frac{1}{K}\sum_{k}|y_k - \hat{y}_k|,\ \text{MRE} = \frac{1}{K}\sum_{k}\frac{|y_k - \hat{y}_k|}{|y_k|} \times 100\%. \tag{15}$$

After extensive numerical experiments, we choose 5 hidden layers with PReLu as the activation function in each layer in the CIFNN models. Adam optimizer[27,28] is used to minimize the MSE in PyTorch. L^2 regularization and early stopping[26] are employed during the training process to avoid overfitting. The hyperparameters of the neural networks shown in Table S10, are chosen by the grid search method.[26] We finalize the machine-learned, neural network model using the following criteria: (i) we choose the model with RMSE on the training set < 0.1, if no such model satisfies it, we choose the one with the smallest RMSE on the training set; (ii) we choose the model with the smallest RMSE on the test set using the 10-fold validation approach among all the hyperparameters searched. The results are summarized in Table 3. According to the MRE in testing, the C_6H_5 case outperforms the C_5H_4 case by 1.04%, 0.62% in CIFNN2 and CIFNN3 model respectively, and CIFNN3 outperforms CIFNN2 by 0.99%, 1.41%, respectively. All of these are consistent with the LSF results. To further examine the generalization error of the neural network models, we examine the relative error of each training and test instance $e_i = \frac{|y_i - \hat{y}_i|}{|y_i|} \times 100\%, i = 1, \cdots, m$, where m is the size of training or test points, in its frequency map. The frequency map of the relative error for the CIFNN3 model in the C_6H_5 case, given in Fig. 4, shows that rare events in large relative errors can occur, but with a very low frequency in both training and testing. This indicates the deep neural models can be used as a predictive model for the BDE with a low probability that it may fail. Overall, the advantage of the CIFNN models over the

chemistry-informed LSF models is quite significant. All CIFNNs yield comparable errors in training and testing of $\sim$ 1 kcal/mol (outperforming LSF models by a factor of three), demonstrating that feeding the molecular properties into neural networks with a bias-free partition strategy is very effective. In order to examine the robustness of the CIFNN models coupled with k-fold validation in making predictions, we repeated the analysis on several randomly selected subsets of data, consisting of 80% of the original data. Our results do not show any deterioration in the generalization error when the size of the dataset is reduced.

Table 3: Performance of the neural network models. The errors for the nonlinear LSF2 and LSF3 models defined by Eqs (6) and (8), respectively are listed for comparison.

Model	H	RMSE train	RMSE test	train MRE [%]	train MAE [kcal/mol]	test MRE [%]	test MAE [kcal/mol]
CIFNN2	C_6H_5	0.088	0.044	7.609	1.171	**6.663**	**1.009**
	C_5H_4	0.013	0.054	2.266	0.373	7.703	1.082
CIFNN3	C_6H_5	0.082	0.035	8.938	1.323	**5.672**	**0.863**
	C_5H_4	0.036	0.040	5.545	0.857	6.292	0.883
LSF2	C_6H_5					19.278	3.416
	C_5H_4					19.403	3.604
LSF3	C_6H_5					16.148	2.738
	C_5H_4					15.748	2.663

4 Conclusions

In summary, we have analyzed the dependence of the chemical stability of substituted cobaltocenium $CoCp_2^+$, on the computed molecular descriptors using both nonlinear chemistry-informed regression and deep feed forward neural network models. The data set of 118 derivatives based on 42 substituent groups characterized by a range of electron-donating and electron-withdrawing properties is constructed and analyzed. We have considered 12 carefully chosen chemistry-informed descriptors of the complexes and relevant fragments including the key electronic structure (e.g. HOMO and LUMO energy), geometric (e.g. Co to hydroxide distance), and physical (e.g. dipole moment) properties of the molecules. We have observed that the BDE correlates nonlinearly with the HOMO and LUMO energies.

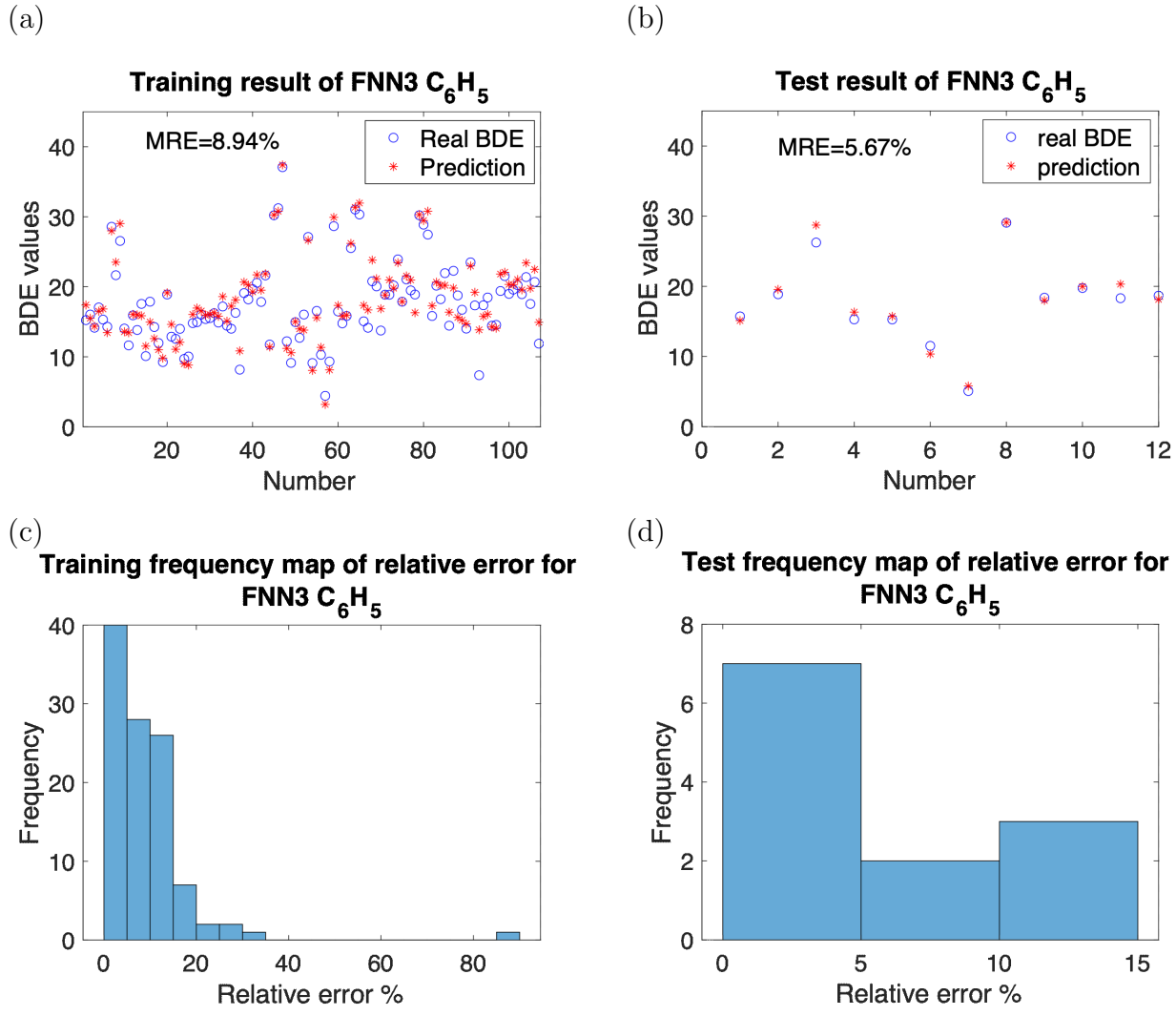

Figure 4: (a) The training and (b) testing errors and the frequency maps of (c) the training and (d) testing relative errors of the CIFNN3 model with the H-charge defined on C_6H_5.

Therefore, within the non-linear regression analysis we have modulated the dependence of the BDE on the HOMO and LUMO by a sigmoid function of the Hirshfeld charge, which yielded test MAE of ~ 3 kcal/mol. The chemical insight gained from the nonlinear regression motivates the use of CIFNN models, which reduce the error to ~ 1 kcal/mol.

The main conclusions from this study are: (i) the Hirshfeld charge is a useful measure of the electron withdrawing-donating character of the substituent groups, with the benzene ring probe leading to better fitting and predictions. (ii) While the models based on the properties of the complexes, fragments and derivatives (LSF3 and CIFNN3) yield slightly lower errors, the models which do not involve the complexes (LSF2 and CIFNN2) are preferred, because the electronic structure calculations are significantly (by a factor of ~ 10) cheaper for the fragments than for the complexes due to both, the scaling of the DFT cost with the system size, and simpler geometry minimization for smaller systems. (iii) Taken together, our modeling suggests that the stability of metallocenium species is enhanced by the electron-donating substituents. Experimental studies verifying the conclusions and modeling approach by integrating chemistry knowledge into the neural network structure will be reported in the near future.

Overall, this work demonstrates that the general concept of incorporating the domain (chemical) insight into the neural network structure extends application of the machine learning techniques to small data sets, ultimately leading to better predictive models compared to the conventional regression analysis that are useful in designing chemical experiments.

5 Acknowledgments

This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences Separation Science program under Award Number DE-SC0020272. Chunyan Li and Qi Wang's research is partially supported by National Science Foundation of US (award DMS-1815921, DMS-1954532 and OIA-1655740) and a

GEAR award from SC EPSCoR/IDeA Program. S.G. acknowledges partial support from the National Science Foundation under Grants CHE-1955768 and OIA-1655740, and GEAR-CRP 20-GC03 award. Calculations were performed in part under an XSEDE[29] allocation TG-DMR110037.

Supporting information. The substituents, the LSF coefficients and the metallocenium data sets.

References

(1) Varcoe, J. R.; Slade, R. C. T. Prospects for Alkaline Anion-Exchange Membranes in Low Temperature Fuel Cells. *Fuel Cells* **2005**, *5*, 187–200.

(2) Couture, G.; Alaaeddine, A.; Boschet, F.; Ameduri, B. Polymeric materials as anion-exchange membranes for alkaline fuel cells. *Progress in Polymer Science* **2011**, *36*, 1521–1557.

(3) Li, X.; Zhang, H.; Mai, Z.; Zhang, H.; Vankelecom, I. Ion exchange membranes for vanadium redox flow battery (VRB) applications. *Energy Environ. Sci.* **2011**, *4*, 1147–1160.

(4) Gu, S.; Wang, J.; Kaspar, R. B.; Fang, Q.; Zhang, B.; Coughlin, E. B.; Yan, Y. Permethyl Cobaltocenium ($Cp^*_2Co^+$) as an Ultra-Stable Cation for Polymer Hydroxide-Exchange Membranes. *Sci. Rep.* **2015**, *5*, 5–11.

(5) Zhu, T.; Sha, Y.; Firouzjaie, H. A.; Peng, X.; Cha, Y.; Dissanayake, D. M. M. M.; Smith, M. D.; Vannucci, A. K.; Mustain, W. E.; Tang, C. Rational Synthesis of Metallo-Cations Toward Redox- and Alkaline-Stable Metallo-Polyelectrolytes. *Journal of the American Chemical Society* **2020**, *142*, 1083–1089.

(6) Zhang, Y.; Wang, L.; Wang, X.; Zhang, C.; Ge, J.; Tang, J.; Su, A.; Duan, H. Data augmentation and transfer learning strategies for reaction prediction in low chemical data regimes. *Org. Chem. Front.* **2021**, *8*, 1415–1423.

(7) Smith, J. S.; Nebgen, B. T.; Zubatyuk, R.; Lubbers, N.; Devereux, C.; Barros, K.; Tretiak, S.; Isayev, O.; Roitberg, A. E. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nat. Commun.* **2019**, *10*, 1–8.

(8) Wang, S.-H.; Pillai, H. S.; Wang, S.; Achenie, L. E. K.; Xin, H. Infusing theory into deep learning for interpretable reactivity prediction. *Nat. Commun.* **2021**, *12*, 1–9.

(9) Schütt, K. T.; Gastegger, M.; Tkatchenko, A.; Müller, K.-R. Quantum-chemical insights from interpretable atomistic neural networks. 2018.

(10) Zhai, F.-H.; Zhan, Q.-Q.; Yang, Y.-F.; Ye, N.-Y.; Wan, R.-Y.; Wang, J.; Chen, S.; He, R.-H. A deep learning protocol for analyzing and predicting ionic conductivity of anion exchange membranes. *J. Membr. Sci.* **2022**, *642*, 119983.

(11) Zou, X.; Pan, J.; Sun, Z.; Wang, B.; Jin, Z.; Xu, G.; Yan, F. Machine learning analysis and prediction models of alkaline anion exchange membranes for fuel cells. *Energy Environ. Sci.* **2021**, *14*, 3965–3975.

(12) Shao, Y.; Gan, Z.; Epifanovsky, E.; Gilbert, A. T. B.; Wormit, M.; Kussmann, J.; Lange, A. W.; Behn, A.; Deng, J.; Feng, X.; *et al*, Advances in molecular quantum chemistry contained in the Q-Chem 4 program package. *Mol. Phys.* **2015**, *113*, 184–215.

(13) Spartan'18. Wavefunction, Inc. Irvine, CA. `http://www.wavefun.com/products/spartan.html`.

(14) Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; *et al*, In *Advances in Neural Information Processing Systems 32*; Wallach, H., Larochelle, H., Beygelzimer, A., d'Alche Buc, F., Fox, E., Garnett, R., Eds.; Curran Associates, Inc., 2019; pp 8024–8035.

(15) Truong, T. N.; Stefanovich, E. V. A new method for incorporating solveing effect into the classical ab-initio molecular-orbital and density-functional theory frameworks for arbitrary shape cavity. *Chem. Phys. Lett.* **1995**, *240*, 253–260.

(16) Rowland, R. S.; Taylor, R. Intermolecular Nonbonded Contact Distances in Organic Crystal Structures: Comparison with Distances Expected from van der Waals Radii. *J. Phys. Chem.* **1996**, *100*, 7384–7391.

(17) Becke, Axel D., Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **1993**, *98*, 5648–5652.

(18) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **1988**, *37*, 785–789.

(19) Grimme, S.; Hansen, A.; Brandenburg, J. G.; Bannwarth, C. Dispersion-Corrected Mean-Field Electronic Structure Methods. *Chemical Reviews* **2016**, *116*, 5105–5154.

(20) Mitin, A. V.; Baker, J.; Pulay, P. An improved 6-31G* basis set for first-row transition metals. *J. Chem. Phys.* **2003**, *118*, 7775–7782.

(21) Zhu, T.; Xu, S.; Rahman, A.; Dogdibegovic, E.; Yang, P.; Pageni, P.; Kabir, M. P.; Zhou, X.-d.; Tang, C. Cationic Metallo-Polyelectrolytes for Robust Alkaline Anion-Exchange Membranes. *Angew.Chem.* **2018**, *57*, 2388–2392.

(22) Phung, Q. M.; Vancoillie, S.; Pierloot, K. Theoretical Study of the Dissociation Energy of First-Row Metallocenium Ions. *J. Chem. Theory Comput.* **2014**, *10*, 3681–3688.

(23) Liu, S. Where does the electron go? The nature of ortho/para and meta group directing in electrophilic aromatic substitution. *J. Chem. Phys.* **2014**, *141*, 194109.

(24) Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. *Numerical Recipes: The Art of Scientific Computing*, 2nd ed.; Cambridge University Press: Cambridge, 1992.

(25) Kreyszig, E. *Advanced Engineering Mathematics: 10th Ed*; J. Wiley, 2011; p 1037.

(26) Goodfellow, I.; Bengio, Y.; Courville, A. *Deep learning*; MIT press, 2016.

(27) Kingma, D. P.; Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* **2014**,

(28) Bock, S.; Goppold, J.; Weiß, M. An improvement of the convergence proof of the ADAM-Optimizer. *arXiv preprint arXiv:1804.10587* **2018**,

(29) Towns, J.; Cockerill, T.; Dahan, M.; Foster, I.; Gaither, K.; Grimshaw, A.; Hazlewood, V.; Lathrop, S.; Lifka, D.; Peterson, G. D.; *et al*, XSEDE: Accelerating Scientific Discovery. *Computing in Science and Engineering* **2014**, *16*, 62–74.

Graphical TOC Entry