Stability Trends in Di-Substituted Cobaltocenium Based on the Analysis of the Machine Learning Models

Shehani T. Wetthasinghe,* Sophya V. Garashchuk,* and Vitaly A. Rassolov*

Department of Chemistry and Biochemistry, University of South Carolina, Columbia, SC 29208, USA

E-mail: shehani@email.sc.edu; garashchuk@sc.edu; rassolov@mailbox.sc.edu

Abstract

Cobaltocenium derivatives have shown great potential as components of anion exchange membranes in fuel cells because they exhibit excellent thermal and alkaline stability under operating conditions, while allowing for high anion mobility. The properties of the cobaltocenium-anion complexes can be chemically tuned through the substituent groups on the cyclopentadienyl (Cp) rings of the cation $CoCp_2^+$. However, synthesis and characterization of the full range of possible derivatives are very challenging and time-consuming, and while the computational tools can greatly expedite this process, full screening of the electronic structure at a high level of theory is still computationally intensive. Therefore, in this work we consider the machine learning (ML) modeling as a tool of predicting stability of di-substituted $[CoCp_2]OH$ complexes measured by their bond-dissociation energy (BDE). The relevant process here is the dissociation of the cobaltocenium-hydroxide complex into fragments $[CoCpY']OH$ and CpY, where Y and Y' each represent one out of 42 substituent groups of experimental

interest. In agreement with the previous ML study of 120 of mono- and selected di-substituted species [Wetthasinghe *et al.* J. Chem. Phys. A (2022) **126**], our analysis of the dataset expanded to all possible di-substituted cobaltoceniums, points to the highest occupied and lowest unoccupied molecular orbitals, along with the Hirshfeld charge on the singly-substituted benzene, to be the key features predicting the BDE of the unseen complexes. Based on examination of the outliers, the acidity of substituents ($(CO)NH_2$ in our case) is found of special significance for the cobaltocenium stability and for the model development. Moreover, we demonstrate that upon the dataset refinement, the conventional ML models are capable of predicting the BDE close to 1 kcal/mol based on the properties of just the fragments, thereby greatly reducing the total number of species and of the computational time of each calculation. Such fragment-based 'combinatorial' approach to the BDE modeling is noteworthy, since the geometry optimization of complexes in solution is conceptually challenging and computationally demanding, even when leveraging high-performance computing resources.

1 Introduction

The Machine Learning (**ML**) methods addressing problems of high computational complexity, are being rapidly adopted and increasingly employed in computational chemistry[1,2], even though the limits of their applicability are still quite uncertain[3,4], and the rational understanding of the behavior of systems subjected to ML algorithms is often elusive. Certain constraints on the accuracy of the ML predictions for real systems come from the approximations inherent to the computational chemistry methods. In particular, driven by high interest in extended molecular systems and materials, the Density Functional Theory (**DFT**) has become a go-to electronic structure method of computing energetic and other properties of stable molecules – especially when sizable datasets are needed – as its scaling properties yield an acceptable balance of accuracy and computational cost. However, the transition

metal (**TM**) compounds, especially for the first-row TM (in the periodic table), still present a considerable challenge for the DFT, and the development of the density functionals that target TM compounds remains an active area of research. A complementary approach is to use ML to improve the accuracy of DFT[5]. Nevertheless, the ML has already made significant contributions to the TM chemistry[2]. For instance, ML has played a crucial role in predicting the HOMO-LUMO gap for TM complexes, uncovering structure-property relationships to aid in the rational design of hetero-bimetallic TM complexes[6], predicting relative-energy and total-energy values for organic and TM-containing molecules[7], and even predicting the line-shape of first-row transition metal K-edge x-ray absorption near-edge structure (**XANES**) spectra[8] using advanced ML methods, such as multiple linear regression, kernel ridge regression, and deep learning techniques.

This work is motivated by the experimental interest in the TM-containing cations as components of anion exchange membranes (**AEM**) in fuel cells[9–11], because these cations exhibit excellent thermal and alkaline stability under operating conditions, while allowing for high anion mobility, and because the properties of the cobaltocenium-anion complexes can be chemically tuned through the substituent groups on the cyclopentadienyl rings (**Cp**) of the cation, $CoCp_2^+$[12]. However, synthesis and characterization of the full range of possible derivatives are very challenging and time-consuming, and while the computational tools can greatly expedite this process, full screening of the electronic structure at a high level of theory, including the DFT methods, is still computationally intensive. Therefore, we are interested in ML modeling as a tool of predicting stability of substituted cobaltocenium-hydroxide complexes, [CoCpY′CpY]OH. As a proxy for the cation stability, we use the bond-dissociation energy (BDE) of [CoCpY′CpY]OH dissociating into fragments [CoCpY′]OH and CpY, where Y and Y′ represent one out of 42 substituent groups of experimental interest.

In Refs[13,14], a team of researchers, including these authors, describes the development of the Neural Networks (**NN**) models predicting the BDE within 1 kcal/mol error, based on a dataset of 118 substituted cobaltocenium molecules. Computations of TM molecular

compounds is challenging due to multiple spin states interfering with the electronic and geometric convergence, which, along with experimental considerations, explains the small by ML standards dataset in the study. The chemically-useful accuracy of the ML predictions was enabled by careful selection of 12-15 input features facilitated by the experimental chemistry expertise, and by the advanced ML techniques, specifically by the Chemistry-Informed and Quadratic NN models introduced in Ref.[14].

In this work, we analyze an expanded dataset to 903 species, which includes all the Y/Y$'$ combinations of the di-substituted cobaltocenium, generated in the automated fashion, and the focus is shifted to a *performance of common ML methods*, i.e. Linear Regression[15], Decision Tree[16,17], Bagged Tree[18,19], K-Nearest Neighbors (**KNN**)[20], Random Forest[21], Support Vector Regression (**SVR**)[22] and Extreme Gradient Boosting (**XGBoost**)[23]. The human, as opposed to the machine, learning is focused on the analysis of the outliers in the ML predictions, and rationalizing them in terms of underlying chemical properties. Specifically, according to our analysis, the acidity of protons on the substituent groups and their steric accessibility during the deprotonation reaction are associated with the unusually high stability relative to the trends within the dataset captured by the ML modeling.

The paper is organized as follows: the molecular model, electronic structure methods and the data generation procedure are described in Section 2. The ML models, their analysis and a discussion of the chemical reasons behind the outliers are presented in Section 3. Section 4 provides a summary and an outlook.

2 Methods and Data generation

2.1 Molecular models and methods

Our main goal is prediction of the bond dissociation energy (**BDE**) of substituted cobaltocenium [Bis(cyclopentadienyl)cobalt(III)] used as a proxy for the stability of the cobaltocenium under basic conditions, i.e. the stability of the cobaltocenium/hydroxide complex with re-

spect to 'splitting' into Cp-containing fragments in aqueous medium as shown in Fig. 1,

$$[\mathrm{CoCpY'CpY}]\mathrm{OH} \rightarrow [\mathrm{CoCpY'}]\mathrm{OH} + \mathrm{CpY}.$$

The relevant BDE is defined as

$$\mathrm{BDE} = E([\mathrm{CoCpY'}]\mathrm{OH}) + E(\mathrm{CpY}) - E([\mathrm{CoCpY'CpY}]\mathrm{OH}). \quad (1)$$

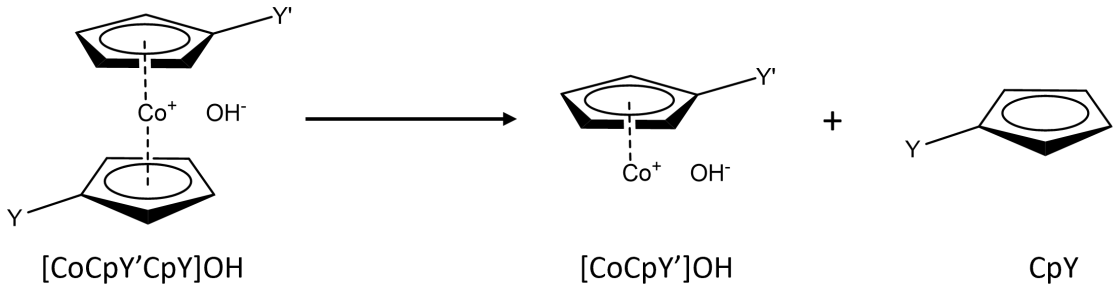


Figure 1: Dissociation of $[\mathrm{CoCpY'CpY}]\mathrm{OH}$ into Cp-containing fragments.

We consider all the cobaltocenium species with a single substituent group, labeled Y or Y′, on each of the two cyclopentadienyl (Cp) rings in all combinations of Y and Y′. The list of substituents is comprised of 42 groups shown in Fig. 2. Following Ref.[13], the $[\mathrm{CoCpY'CpY}]^+$ ion is explicitly solvated with the hydroxyl ion OH^-, and implicitly solvated within water using Polarized Continuum Model (**PCM**), as implemented in Q-Chem[24,25]. The geometry optimization and property calculations are performed at the B3LYP-D3/m6-31G** theory level, which involves the modified 6-31G* basis set for cobalt[26], and 6-31G** basis for all other atoms[27]. The lowest energy spin states of the $[\mathrm{CoCpY'CpY}]\mathrm{OH}$, $[\mathrm{CoCpY'}]\mathrm{OH}$ and CpY are the singlet, quartet, and doublet, respectively. The choice of the electronic structure method has been previously validated through comparison to several other theoretical methods and experiments[13]. In all cases the energies and other molecular descriptors used in the ML models are computed at the optimized geometries. Because the geometry optimization can be problematic for the system with low binding energies, the species with computed

BDE $<$ 3 kcal/mol are excluded from the dataset. For future reference, let us note here, that in addition to the set of 42 substituents in Fig. 2, four additional groups have been considered, namely OH, COOH, NH_2 and $NHCH_3$. Instead of forming complexes with the hydroxide, the corresponding cobaltocenium derivatives got deprotonated during geometry optimization. These four groups were, therefore, excluded from the ML modeling.

To check the influence of the chosen basis set on the computed energies, we have performed additional calculations for three substituents using 6-311++G(d,p) and G3MP2Large[28] sets. The observed energy deviations of 5-15 kcal/mol, summarized in Table S4[au], are consistent with the DFT calculations for the transition metal compounds[28].

The ML model development in this work is based on the molecular descriptors or features associated with the properties of just the dissociation fragments for both conceptual and practical potential considerations. First of all, the lowest energy structure is less ambiguous and the property calculation is more accurate for the fragments ($\sim$ 20 atoms) compared to the cobaltocenium/hydroxide complex ($\sim$ 40 atoms), especially, given the multiple binding sites of the hydroxide. Furthermore, the scaling of the DFT cost with the system size means that a single point energy calculation on a complex is about 10 times faster than on a fragment. The most significant potential computational saving though, would be the sheer number of the complexes ($42 \times 43/2$) compared to that of the fragments (42×2).

2.2 Data generation

The computational procedure was automated using bash and Python scripting. Generation of the initial structures for the geometry optimization of the complexes is illustrated in Fig. 3. As a 'template' structure we use the optimized geometry of the unsubstituted $[CoCp_2]OH$ complex, resulting in a symmetric orientation of the Cp rings sketched. As follows from Ref.[13], the lowest energy arrangement of two substituents (one per Cp) is the trans configuration, i.e. positions C1 and C2 as shown Fig. 3(a,c). First, the basic structure of the complex is created by placing Co at the origin of the Cartesian system of

coordinates, and rotating the $CoCp_2$ unit to have the average position of the Cp rings in the xy-plane. Then, the hydroxide is placed along the y-axis with the oxygen atom separated from the cobalt atom by 4.0 Å as shown in Fig 3(a). Next, one H-atom of a Cp ring is replaced with a substituent group Y, and the geometry of Y is optimized while keeping the remaining atomic positions frozen. The resulting coordinates of the substituent atoms are used to generate the initial geometries for all di-substituted species, as illustrated in Fig. 3(c). Finally, the prepared script running through the list of 42 substituents produces `*.xyz` files of initial geometries of 903 complexes. The input files for the full geometry optimization are prepared by combining the structure files with the Q-Chem job specification block using another Python code. Then, the geometry optimization calculations are submitted to a high performance computer in a serial manner using a bash script. Calculations for two fragments with 42 substituents each are setup and executed in a similar manner. The output files are parsed for error messages, and those jobs are examined and resubmitted manually. We encountered errors in only 6 derivatives. A sample Q-Chem input file with all user-specified parameters is provided (Table S1-S3).

Once the optimization calculations are completed, the data used in the ML analysis are extracted from the output files to an excel sheet using the developed python script. The following features, which have been previously identified as the most relevant to the BDE prediction[14] are collected:

(i) energy of [CoCpY′CpY]OH;

(ii) HOMO energy (E_{HOMO}) of [CoCpY′CpY]OH;

(iii) LUMO energy (E_{LUMO}) of [CoCpY′CpY]OH;

(iv) energy of [CoCpY′]OH fragment;

(v) energy of CpY fragment;

(vi) E_{HOMO} of [CoCpY′]OH fragment;

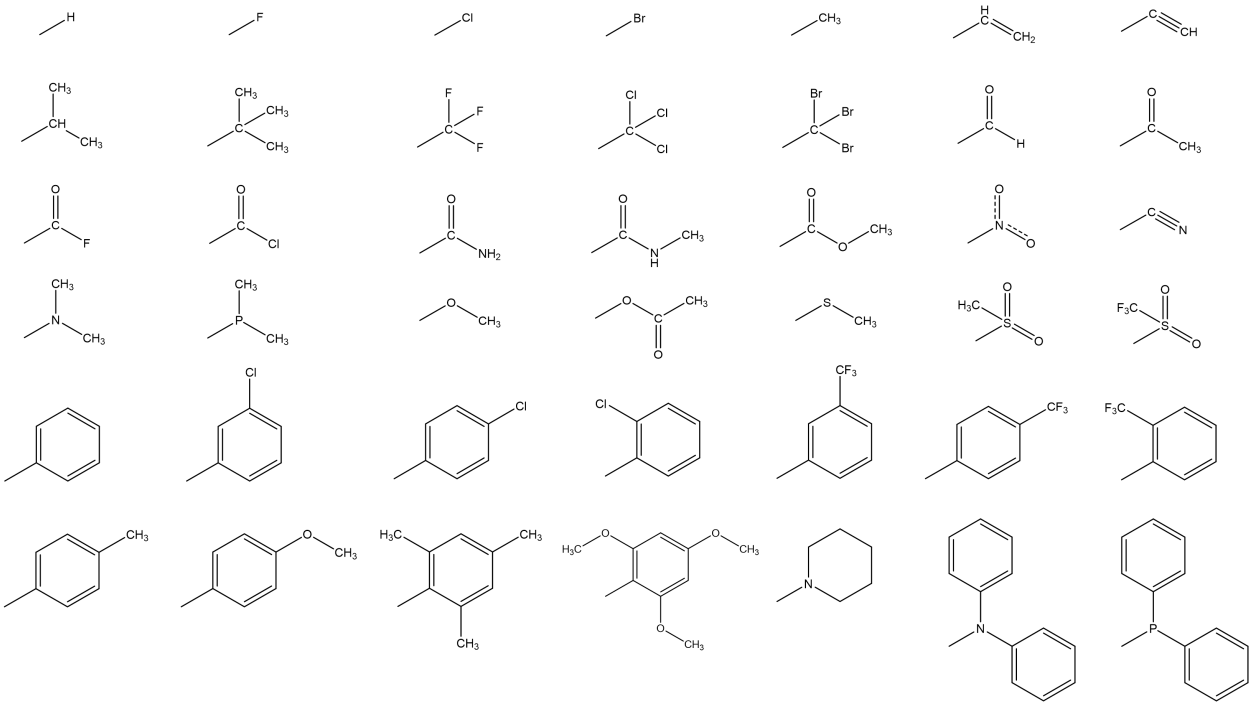


Figure 2: The substituent groups (denoted as Y or Y$'$ throughout) used in chemical modification of the cobaltocenium [CoCpY$'$CpY]$^+$.

(vii) E_{LUMO} of [CoCpY$'$]OH fragment;

(viii) E_{HOMO} of CpY fragment;

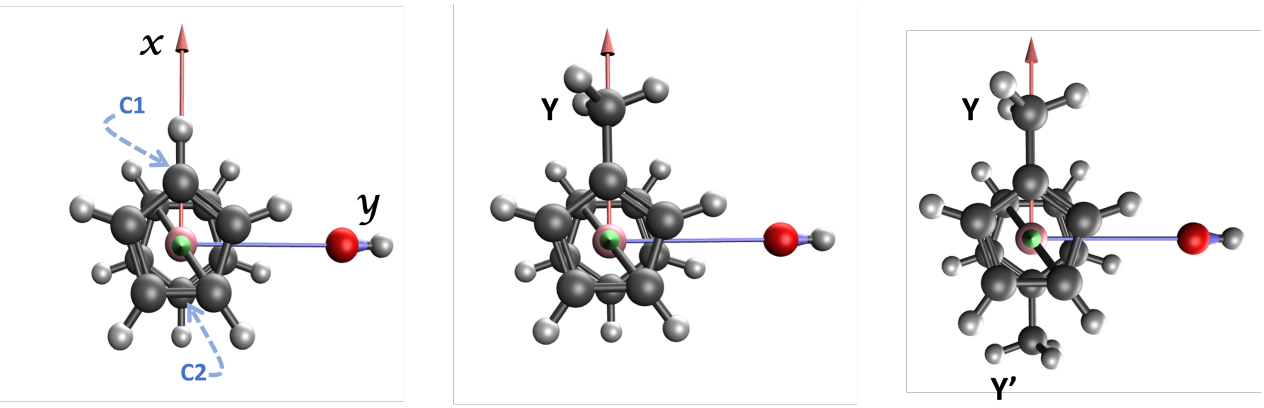
(ix) E_{LUMO} of CpY fragment.

These are augmented by the features derived from the calculations on the substituent groups described in previous work[13]:

(x) sum of the Hirshfeld charges on C_6H_5 from the electronic structure of C_6H_5Y';

(xi) sum of the Hirshfeld charges on C_6H_5 from the electronic structure of C_6H_5Y.

The last two features are introduced as a measure of the electron donating/withdrawing character of the substituents: a hydrogen atom in the aromatic system is replaced by a substituent Y, which leads to redistribution of electrons on the ring quantified by the electron population analysis. We use specifically the Hirshfeld charge, rather than Mulliken or Löwdin charges, because the former is less basis set dependent[29]. The BDE of a complex is computed from Eq. (1) using features (i), (iv) and (v). The ML modeling of the BDE is based on the



(a) (b) (c)

Figure 3: The procedure of generating initial geometries of [CoCpY′CpY]OH. (a) Co of the unsubstituted complex is placed at the origin of the Cartesian coordinates and OH is aligned with y-axis. (b) H-atom on C1 is replaced with the substituent Y. (c) H-atom on C2 is replaced with Y′ whose positions were reflected with respect to the origin.

features (v)-(xi) that are the HOMO and LUMO energies of the fragments. The computed dataset is available through the Open Science Framework[30]. Before proceeding with the ML analysis, the species characterized by the BDE below 3 kcal/mol are removed from the dataset, as being below the DFT accuracy, which reduces the dataset size from 903 to 873 molecules. Figure S2 illustrates the distribution of each feature after removing the initial outliers.

3 Results and discussion

3.1 The machine learning models

Now let us turn to the ML model development and analysis. Seven ML regression models have been constructed: 80% of randomly selected data was used for training the models while remaining 20% was used for testing. Since the energy and charge features have different units, the data were normalized prior to modeling according to the following relation,

$$x_i^{normalized} = \frac{x_i^{original} - \mu}{\sigma}, \tag{2}$$

where $x_i^{original}$, $x_i^{normalized}$, μ and σ are, respectively, the original value, the normalized value, mean and standard deviation of a chemical property, calculated for ith derivative. The following models have been considered: Linear Regression[15], Decision Tree[16,17], Bagged Tree[18,19], KNN[20], Random Forest[21], SVR[22] and XGBoost[23], as implemented in `scikit-learn` library in Python[31]. Once the models have been trained, their accuracy is assessed from the root mean squared error (RMSE) and from the Pearson's Correlation coefficient (R^2 Score). As

Table 1: Performance of machine learning models with their default parameters

Model	Train R^2 Score	Test R^2 Score	Train RMSE (kcal/mol)	Test RMSE (kcal/mol)
Linear Regression	0.4598	0.4003	4.8570	5.1703
Decision Tree	1.0000	0.5010	0.0000	4.7163
Bagged Tree	0.9686	0.7585	1.1718	3.2806
KNN	0.7000	0.4287	3.6200	5.0462
Random Forest	0.9821	0.7951	0.8845	3.0223
SVR	0.4793	0.3785	4.7688	5.2633
XGBoost	0.9050	0.8440	2.0366	2.6366

seen from Table 1 listing both measures, in all cases the test R^2 scores are significantly lower than the train R^2 scores. Thus, all models suffer from overfitting, as confirmed by large test RMSE scores relative to the train RMSE. We also note that based on the test scores XGBoost, Random Forest and Bagged Tree perform the best, and we select these three out of seven models for further optimization via tuning hyperparameters and conducting five-fold cross validation. The default and optimized model parameters are given in Table S5.

The performance of optimized models is summarized in Table 2 and illustrated in Fig. S3. Clearly, XGBoost, characterised by the highest R^2 and lowest RMSE test scores, is the best performing model. Yet, even after the model parameter optimization, the test scores are lower than the train scores. To identify the reason and reduce overfitting, let us examine the outliers. From the test results in Fig. S1 showing predicted vs actual BDE for each model, it is easy to notice that the points deviating from the trend line tend to have high actual BDEs, BDE $>$ 30 kcal/mol. These outlier species, responsible for the lowering of test scores relative to the train scores, call for further investigation.

Table 2: Summary of optimized model performance

Model	Train R^2 Score	Test R^2 Score	Train RMSE (kcal/mol)	Test RMSE (kcal/mol)
XGBoost	0.9663	0.8809	1.2237	2.2252
Random Forest	0.9819	0.8230	0.8971	2.7135
Bagged Tree	0.9821	0.8251	0.8935	2.6970

3.2 The outlier analysis

First, the outliers common to all three optimized models were identified by calculating the error defined as $\epsilon =$ | BDE(actual) - BDE(predicted) | where $\epsilon > 3$ kcal/mol. Out of the top 8 outliers, shown in Table 3 and in Fig. S4, five contain amide group $(CO)NH_2$.

Table 3: The substituent groups in [CoCpY$'$]OH and CpY fragments for the top eight outliers common to the three optimized models. The error in kcal/mol is averaged over these models.

Substituent Groups		Error
Y$'$	Y	(kcal/mol)
$(CO)NH_2$	$P(CH_3)_2$	14.6233
$(CO)NH_2$	Ph	14.0758
$(CO)NH_2$	p-$PhOCH_3$	13.6500
$(CO)NH_2$	$(CO)CH_3$	7.5766
op-$Ph(CH_3)_3$	op-$Ph(CH_3)_3$	6.3572
$(CO)NH_2$	$S(CH_3)$	5.9637
$N(CH_3)_3$	NC_5H_{10}	4.5597
m-$PhCF_3$	p-PhCl	3.9388

Therefore, we separate out all 42 amide derivatives from the dataset and reanalyze the data distribution with respect to each input feature. The results are shown in Fig. S5 with the corresponding values for the amide derivatives (magenta) overlaid over the distribution for each feature. The $(CO)NH_2$ derivatives do not appear as clear outliers within the feature distributions, which means that at a different molecular property is associated with the outliers, and this feature should be included into the analysis. Thus, we have considered six additional properties, not included in the ML models disucssed so far: the energy of the complexes, [CoCpYCpY$'$]OH, and of the fragments, [CoCpY$'$]OH and CpY, the HOMO/LUMO energies of the complex, and the BDE itself. The distributions of the outliers relative to these

properties are shown in Fig. S6. It is evident that most of the $(CO)NH_2$ derivatives appear as significant outliers in the distribution of the HOMO energy of the complex. Closer examination of the BDE dependence on the HOMO energy of the substituted complexes, shown in Fig. 4, reveals that the amide derivatives data can be grouped into two clusters – a larger cluster containing $(CO)NH_2$ derivatives with BDE > 20 kcal/mol and $E(HOMO) < -0.21$ hartree, and a smaller one with lower BDE values (< 20 kcal/mol) and higher HOMO energy values $E_{HOMO} > -0.20$ hartree. In summary, this analysis has shown that most (33 out of 42) of the $(CO)NH_2$ derivatives appear as outliers, and that this trend is observed only in the HOMO energy of the complex, which was not used as an input feature to the ML models.

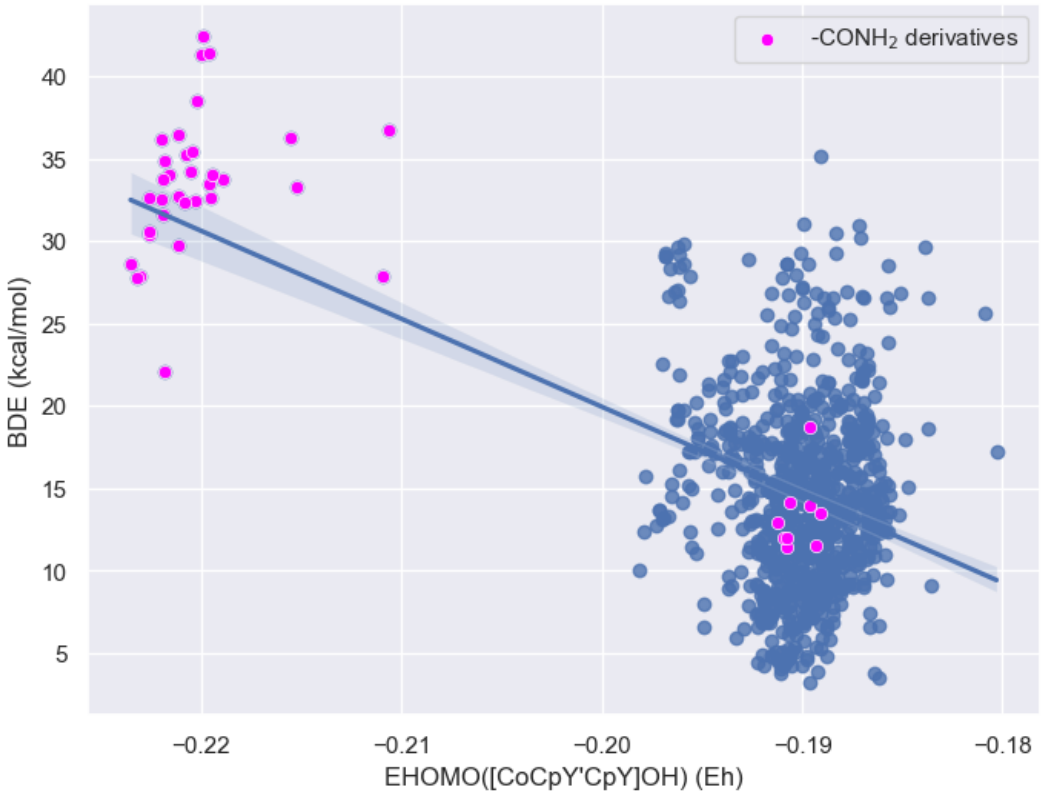


Figure 4: The outlier analysis: BDE vs HOMO energy of the complex. The blue shaded area indicates 95% confidence interval.

In order to confirm that the relatively poor performance of the ML models is attributable to the $(CO)NH_2$ derivatives, they are removed from the dataset, which is then reanalyzed

following the same procedure as described above. First, the models were evaluated with the default parameters and then with the optimized parameters. Table S6 shows the default and the optimized parameters for the models when the refined data set is used.

The performance of the three selected ML models with the refined data is presented in Fig. S7 and the summary is provided in Table 4. Both train and test R^2 scores are significantly improvement in all three ML models after the removal of $(CO)NH_2$ outliers from the data set. The best preforming model, i.e. optimized XGBoost, gives prediction accuracy for unseen data of 96%, with the RMSE of 1 kcal/mol. This indicates that aside from $(CO)NH_2$ outliers, the implemented ML models themselves do not exhibit any problems. Next, we propose a chemical explanation for the appearance of $(CO)NH_2$ derivatives as outliers.

Table 4: Summary of optimized model performance for refined data set

Model	Train R^2 Score	Test R^2 Score	Train RMSE (kcal/mol)	Test RMSE (kcal/mol)
XGBoost	0.9928	0.9623	0.4860	1.0104
Random Forest	0.9908	0.9411	0.5495	1.2639
Bagged Tree	0.9908	0.9406	0.5503	1.2685

3.3 Chemistry of amide outliers

At the start of the project, in addition to the 42 substituent groups (Fig. 2), we have also considered OH, COOH, NH_2, and $NHCH_3$ groups. However, the geometry optimizations in PCM did not yield a coordinated cobaltocenium-hydroxide complex, because the hydroxide deprotonated those substituents. Thus, a reasonable assumption is that the outlier behavior of compounds with $(CO)NH_2$ derivatives is also related to the acidity of amide protons. Acidity, or ease of detachment of a proton from the molecule, is related to the degree of polarization of its chemical bond, and thus can be correlated with the partial charge on the hydrogen in a protonated molecule. To validate this hypothesis, we calculated the Mulliken charge on H atoms of substituent groups, Y, attached to the mono-substituted complex, [CoCpCpY]OH after the partial optimization of the Y positions while the remaining atoms

fixed in space. Selected substituents from the dataset and several known acidic substituents were analyzed. Additionally, we have done the Hirshfeld charge analysis on H atoms of the same substituent groups attached to a benzene ring. The results are summarized in Fig. S8.

In the case of known acidic substituents, the Mulliken charge ranges between 0.31 and 0.39, while the Hirshfeld charge falls within the range of 0.12 to 0.22. Among all the tested groups, the H atoms attached to N atoms of $(CO)NH_2$ and $(CO)NHCH_3$ exhibit charges within this range. Out of these two substituent types, $(CO)NH_2$ demonstrates the highest charges among its H atoms. To expand the analysis to all substituent groups with H atoms, we calculated the deprotonation ability using benzene (C_6H_6) as the probe. The deprotonation ability is defined as follows:

$$Deprotonation\ ability = E(C_6H_5Y) - E(C_6H_5[Y-H]^-) \tag{3}$$

where $E(C_6H_5Y)$ is the energy of benzene substituted with Y and $E(C_6H_5[Y-H]^-)$ represents the energy of deprotonated version of C_6H_5Y. To reduce the computational cost, only C_6H_6 probe has been considered. It is important to note that acidity is inversely proportional to the deprotonation ability. The calculated deprotonation abilities are displayed in Fig. 5, where they are arranged from the least to the greatest acidity. The formamide group $(CO)NH_2$ shows the highest acidity among the substituent groups considered, while the N-methylacetamide group $(CO)NHCH_3$ shows the second highest acidity. Reexamining the optimized geometries of these compounds demonstrates the proximity of OH^- to a formamide hydrogen, as opposed to solvation of the $CoCp_2^+$ core. To determine if this is an artefact of the PCM model, we have also performed geometry optimizations for a few $(CO)NH_2$ derivatives in the gas phase, and the same behavior was still observed. Therefore, it has been confirmed that the trapping of the OH^- anion around the $(CO)NH_2$ group is attributed to the presence of acidic hydrogen atom. Interestingly, despite the high acidity of $(CO)NHCH_3$, steric hindrance by the methyl CH_3 group prevents the association of OH^-

with amide hydrogens. This explains why N-methylacetamide does not pose a problem to the ML model. Therefore, we recommend to evaluate the acidity of new substituent groups using deprotonation ability calculation or Mulliken/Hirshfeld charge analysis, to ensure applicability of the ML models investigated in the current work.

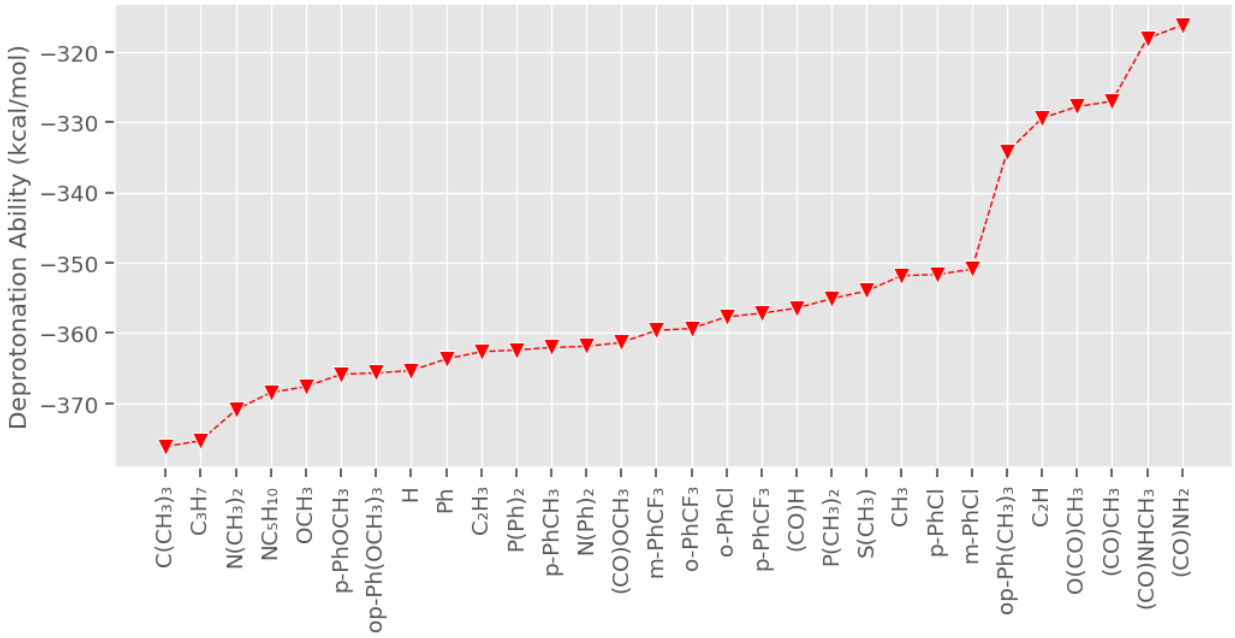


Figure 5: Deprotonation ability on substituent (Y) of C_6H_5Y computed in PCM at B3LYP-D3/6-31G** theory level.

4 Conclusion

In this project we have developed a machine learning model to predict the stability of di-substituted $[CoCp_2]OH$ derivatives based on the BDE of their dissociation into $[CoCpY']OH$ and CpY fragments. To accomplish this, we have computed a large dataset of these derivatives by automating the process of generating and submitting input files to a high-performance cluster, and of extracting the relevant data for property calculations from the output files using python scripts. According to the previous work, we observed that the computational time for the simulations of complexes was significantly higher compared to its fragments. Therefore, the main objective of this project was to create a machine learning model that

utilizes fragment properties to provide predictions with high accuracy and reduced computational time.

After testing several standard machine learning models, three models were selected for further optimizations based on the evaluation scores. However, optimizing the parameters of models did not resolve overfitting noticed with the default model parameters. Therefore, the data set was further investigated for unusual behavioral patterns. After carefully examining the outliers shared by all three optimized machine learning models, we have discovered that the substitution of amide $(CO)NH_2$ group in one of the Cp ring caused a significant deviation of the predicted BDE value from the corresponding actual BDE value. Surprisingly, the same species did not manifest themselves as outliers in the input features. To gain deeper understanding, as the next step, the energy values defining the actual BDEs and HOMO/LUMO energies of the complexes were examined. This analysis revealed that the amide substituted complexes appeared as outliers in HOMO energy of the complexes. Based on data in Fig. 4, the majority of $(CO)NH_2$ derivatives have higher BDE values with lower HOMO energies, which implies that $(CO)NH_2$ stabilizes the cobaltocenium/hydroxide complex by acting as the electron withdrawing group. However, this observation appears to contradict our previous work and experimental results, which showed that the electron-donating groups stabilize $[CoCp_2]OH$ derivatives. Consequently, we conducted further analysis to understand why $(CO)NH_2$ derivatives exhibited such different behavior.

Based on the challenges we encountered at the beginning of the project, we hypothesized that the acidity might influence this distinct behavior. To investigate this, we performed Mulliken/Hirshfeld charge analysis and calculated the deprotonation ability of the H atom in the substituent groups. The results indicated that amide is the most acidic group among the 42 substituent groups. This acidity leads to the trapping of OH^- anions around the H atom of the substituent, preventing their interaction with the $CoCp_2^+$ cation. This behavior explains why $(CO)NH_2$ derivatives exhibited unique characteristics. Also, this suggests a substituent deprotonation is a possible degradation pathway for the AEM. Consequently,

we re-optimized the machine learning models by excluding $(CO)NH_2$ derivatives from the dataset, achieving the highest accuracy with the XGBoost model, which demonstrated 96 % performance for unseen data with the accuracy of 1 kcal/mol.

The modified and optimized XGBoost model we developed can be employed to predict the stability of $[CoCp_2]OH$ derivatives containing less acidic substituent groups compared to $(CO)NH_2$. Therefore, we recommend performing deprotonation ability calculation or Hirshfeld charge analysis of the substituted benzene prior to utilizing the machine learning model. If the calculated deprotonation ability or Hirshfeld charge exceeds the level of $(CO)NH_2$, the model may not accurately predict the stability for that specific substituent group. Our optimized machine learning model is recommended for predicting the stability of less acidic derivatives of $[CoCp_2]OH$ compared to $(CO)NH_2$.

5 Supporting Information

Table S1. Sample input file for geometry optimization of $[CoCp_2]OH$.

Table S2. Sample input file for geometry optimization of [CoCp]OH.

Table S3. Sample input file for geometry optimization of Cp fragment.

Table S4. Basis sets investigation.

Table S5. Default and new parameters after optimization for the complete dataset.

Table S6. Default and new parameters after optimization for refined data set.

Figure S1. The train and test results for the optimized ML models.

Figure S2. The distribution of each input features. The bins show the distribution as a histogram plot while the line indicates the kernel density estimate (KDE) plot.

Figure S3. (a) Evaluation of optimized models with R^2 Scores, (b) Evaluation of optimized models with RMSE.

Figure S4. Actual and predicted BDE for top 8 outliers.

Figure S5. Distribution of $(CO)NH_2$ derivatives over input features. The blue bins show the

distribution of all derivatives as a histogram plot and blue line indicates its KDE plot. The magenta colored bins show the distribution of $(CO)NH_2$ derivatives as a histogram plot.

Figure S6. Distribution of $(CO)NH_2$ derivatives over additional features. The blue bins show the distribution of all derivatives as a histogram plot and blue line indicates its KDE plot. The magenta colored bins show the distribution of $(CO)NH_2$ derivatives as a histogram plot.

Figure S7. The train and test results of the optimized models for the refined dataset.

Figure S8. The Mulliken charge on H of the Y-group of [CoCpHCpY]OH, and the Hirshfeld charge on H on the Y-group of C_6H_5Y. The examined H-atom is highlighted with red in chemical structures. The values highlighted with red has the highest charge on H atoms which imply high acidity levels.

6 Acknowledgment

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. Additional support comes from the National Science Foundation of U.S.A. CHE-1955768. We acknowledge computational resources of the ACCESS (Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support) program available through allocation TG-DMR110037.

References

(1) Badillo, S.; Banfai, B.; Birzele, F.; Davydov, I. I.; Hutchinson, L.; Kam-Thong, T.; Siebourg-Polster, J.; Steiert, B.; Zhang, J. D. An Introduction to Machine Learning. *Clin. Pharmacol. Ther.* **2020**, *107*, 871–885.

(2) Nandy, A.; Duan, C.; Taylor, M. G.; Liu, F.; Steeves, A. H.; Kulik, H. J. Computational

discovery of transition-metal complexes: from high-throughput screening to machine learning. *Chem. Rev.* **2021**, *121*, 9927–10000, PMID: 34260198.

(3) Korolev, V.; Mitrofanov, A.; Korotcov, A.; Tkachenko, V. Graph Convolutional Neural Networks as "General-Purpose" Property Predictors: The Universality and Limits of Applicability. *J. Chem. Inf. Model.* **2020**, *60*, 22–28, PMID: 31860296.

(4) Duan, C.; Janet, J. P.; Liu, F.; Nandy, A.; Kulik, H. J. Learning from Failure: Predicting Electronic Structure Calculation Outcomes with Machine Learning Models. *J. Chem. Theory Comput.* **2019**, *15*, 2331–2345, PMID: 30860839.

(5) Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. Predicting reaction performance in C-N cross-coupling using machine learning. *Science* **2018**, *360*, 186–190.

(6) Taylor, M. G.; Nandy, A.; Lu, C. C.; Kulik, H. J. Deciphering Cryptic Behavior in Bimetallic Transition-Metal Complexes with Machine Learning. *J. Phys. Chem. Lett.* **2021**, *12*, 9812–9820, PMID: 34597514.

(7) Husch, T.; Sun, J.; Cheng, L.; Lee, S. J. R.; Miller, I., Thomas F. Improved accuracy and transferability of molecular-orbital-based machine learning: Organics, transition-metal complexes, non-covalent interactions, and transition states. *J. Chem. Phys.* **2021**, *154*, 064108.

(8) Rankine, C. D.; Penfold, T. J. Accurate, affordable, and generalizable machine learning simulations of transition metal x-ray absorption spectra using the XANESNET deep neural network. *J. Chem. Phys.* **2022**, *156*, 164102.

(9) Rowland, T. G.; Sztáray, B.; Armentrout, P. B. Metal–Cyclopentadienyl Bond Energies in Metallocene Cations Measured Using Threshold Collision-Induced Dissociation Mass Spectrometry. *J. Phys. Chem. A* **2013**, *117*, 1299–1309, PMID: 23215634.

(10) Pawar, M. S.; Zha, Y.; Disabb-Miller, M. L.; Johnson, Z. D.; Hickner, M. A.; Tew, G. N. *Polymer Composites for Energy Harvesting, Conversion, and Storage*; 2014; Chapter 6, pp 127–146.

(11) Phung, Q. M.; Vancoillie, S.; Pierloot, K. Theoretical Study of the Dissociation Energy of First-Row Metallocenium Ions. *J. Chem. Theor. Comp.* **2014**, *10*, 3681–3688, PMID: 26588513.

(12) 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, PMID: 31846313.

(13) Wetthasinghe, S. T.; Li, C.; Lin, H.; Zhu, T.; Tang, C.; Rassolov, V.; Wang, Q.; Garashchuk, S. Correlation between the Stability of Substituted Cobaltocenium and Molecular Descriptors. *J. Phys. Chem. A* **2022**, *126*, 80–87.

(14) Li, C.; Wetthasinghe, S. T.; Lin, H.; Zhu, T.; Tang, C.; Rassolov, V.; Wang, Q.; Garashchuk, S. Stability Analysis of Substituted Cobaltocenium[Bis(cyclopentadienyl)cobalt(III)] Employing Chemistry-Informed Neural Networks. *J. Chem. Theor. Comp.* **2022**, *18*, 3099–3110.

(15) Maulud, D.; Abdulazeez, A. M. A Review on Linear Regression Comprehensive in Machine Learning. *J. Appl. Sci. Technol. Trends* **2020**, *1*, 140–147.

(16) Somvanshi, M.; Chavan, P.; Tambade, S.; Shinde, S. V. A review of machine learning techniques using decision tree and support vector machine. *2016 International Conference on Computing Communication Control and Automation (ICCUBEA)* . 12-13 Aug 2016, Pune, India, pp 1–7.

(17) Pathak, S.; Mishra, I.; Swetapadma, A. An Assessment of Decision Tree based Clas-

sification and Regression Algorithms. *2018 3rd International Conference on Inventive Computation Technologies (ICICT)*. 15-16 Nov. 2018, Coimbatore, India, pp 92–95.

(18) Iverson, L.; Prasad, A.; Liaw, A. New machine learning tools for predictive vegetation mapping after climate change: Bagging and Random Forest perform better than regression tree analysis. *Proceedings, UK-International Association for Landscape Ecology, Cirencester, UK* . 2004; pp 317–320.

(19) Breiman, L. Bagging predictors. *Machine Learning* **1996**, *24*, 123–140.

(20) Kramer, O. *K-nearest neighbors*; Springer, 2013; pp 13–23.

(21) Segal, M. R. Machine learning benchmarks and random forest regression. https://escholarship.org/uc/item/35x3v9t4, (accessed Nov 3, 2023).

(22) Awad, M.; Khanna, R. *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*; Apress: Berkeley, CA, 2015; pp 67–80.

(23) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA, 2016; p 785–794.

(24) Lange, A. W.; Herbert, J. M. A smooth, nonsingular, and faithful discretization scheme for polarizable continuum models: The switching/Gaussian approach. *J. Chem. Phys.* **2010**, *133*, 244111.

(25) Epifanovsky, E.; Gilbert, A. T. B.; Feng, X.; Lee, J.; Mao, Y.; Mardirossian, N.; Pokhilko, P.; White, A. F.; Coons, M. P.; Dempwolff, A. L. et al. Software for the frontiers of quantum chemistry: An overview of developments in the Q-Chem 5 package. *J. Chem. Phys.* **2021**, *155*, 084801.

(26) 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.

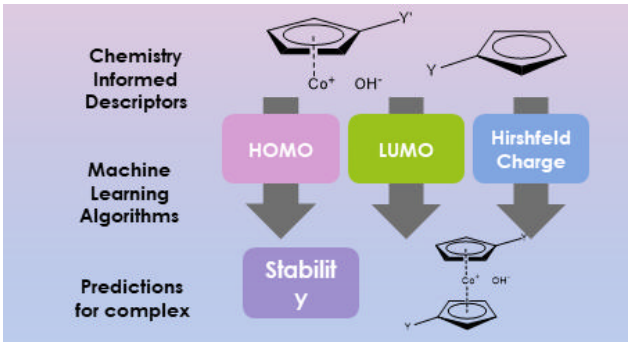
(27) Hehre, W. J.; Ditchfield, R.; Pople, J. A. Self-Consistent Molecular Orbital Methods. XII. Further Extensions of Gaussian-Type Basis Sets for Use in Molecular Orbital Studies of Organic Molecules. *J. Chem. Phys.* **1972**, *56*, 2257.

(28) Mayhall, N. J.; Raghavachari, K.; Redfern, P. C.; Curtiss, L. A.; Rassolov, V. Toward accurate thermochemical models for transition metals: G3Large basis sets for atoms Sc–Zn. *J. Chem. Phys.* **2008**, *128*, 144122.

(29) 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.

(30) Garashchuk, S. V.; Wetthasinghe, S. T. Stability and Properties of Substituted Cobaltocenium. https://doi.org/10.17605/OSF.IO/6ZA8C, 2022; OSF. (accessed September 10, 2023).

(31) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.


Chemistry Informed Descriptors
Machine Learning Algorithms
Predictions for complex
HOMO
LUMO
Hirshfeld Charge
Stabilit y
Co+
OH-
Y'
Y

Supplemental Information: Stability Trends in Di-Substituted Cobaltocenium Based on the Analysis of the Machine Learning Models

by Shehani T. Wetthasinghe, Sophya V. Garashchuk, Vitaly A. Rassolov

Table S1: Sample input file for geometry optimization of $[CoCp_2]OH$.

$molecule			
0 1			
C	0.95704	-1.66305	-0.7215
C	0.97269	-1.66015	0.70764
C	-0.38164	-1.64429	1.16373
C	-1.23354	-1.63982	0.01686
C	-0.40674	-1.64938	-1.14809
H	1.8241	-1.6266	-1.36722
H	1.85367	-1.62061	1.33409
H	-0.70486	-1.59301	2.19478
H	-0.75234	-1.60227	-2.17207
Co	0.00000	0.00000	0.00000
C	-0.95705	1.66305	0.72148
C	-0.97269	1.66015	-0.70764
C	0.38164	1.64429	-1.16374
C	1.23353	1.63982	-0.01687
C	0.40673	1.64939	1.14808
H	-1.8241	1.6266	1.36721
H	-1.85367	1.6206	-1.3341
H	0.70486	1.59301	-2.19479

H	0.75234	1.60227	2.17206
O	0.00000	0.00000	4.00000
H	0.00000	0.00000	4.97596

#insert substituent1

#insert substituent2

$end

$rem

MEM_TOTAL	20000
MEM_STATIC	5000
METHOD	B3LYP
DFT_D	D3_ZERO
BASIS	GEN !use m631g* for Co
PURECART	11
SYMMETRY_IGNORE	1
unrestricted	FALSE
SCF_GUESS	SAD
JOBTYPE	OPT
MAX_SCF_CYCLES	2000
SCF_CONVERGENCE	7
GEOM_OPT_MAX_CYCLES	500
SOLVENT_METHOD	PCM !include H2O solvent

$end

$pcm

Theory	CPCM
Method	SWIG

Solver	Inversion
HeavyPoints	194
HPoints	194
Radii	read
vdwScale	1.2
$end	

$solvent

Dielectric	78.39

$end

$van_der_waals

1

1	1.2
6	1.92
7	1.55
8	1.52
9	1.5138
15	1.8657
16	1.8153
17	1.782
27	2
35	1.8855

$end

$basis

#modified 6-31G* basis set

Co 0

S 6 1.00

66148.99	0.00175979	
9933.077	0.01348162	
2262.816	0.06649342	
637.9154	0.2307939	
204.4122	0.4792919	
69.82538	0.3514097	

SP 6 1.00

1378.841	0.00237628	0.00397149
328.2694	0.0316745	0.03108174
106.0946	0.1262888	0.1357439
39.83275	-0.02584552	0.3476827
16.18622	-0.6183491	0.462634
6.667788	-0.4567008	0.2051632

SP 6 1.00

54.52355	-0.003993	-0.00729077
18.29783	0.07409663	-0.02926027
7.867348	0.2542	0.0656415
3.340534	-0.2921657	0.4000652
1.393756	-0.7318703	0.4950236
0.551326	-0.2040784	0.175824

SP 3 1.00

2.151947	0.05379843	-0.2165496
0.811063	0.2759971	0.1240488
0.121017	-1.129692	0.9724064

SP 1 1.00

0.043037 1 1

D 3 1.00

21.33427761 0.10682753

5.57681943 0.426687871

1.598512718 0.669867704

D 1 1.00

0.399532909 1

F 1 1.00

0.8 1

****

H 0

6-31g**

****

C 0

6-31g**

****

O 0

6-31g**

****

F 0

6-31g**

****

N 0

6-31g**

****

S 0

6-31g**

****

P 0

6-31g**

****

Cl 0

6-31g**

****

Br 0

6-31g**

****

$end

Table S2: Sample input file for geometry optimization of [CoCp]OH.

$molecule

0 4

#insert xyz coordinates of [CoCp]OH.

$end

$rem

MEM_TOTAL	20000	
MEM_STATIC	5000	
METHOD	B3LYP	
DFT_D	D3_ZERO	
BASIS	GEN	!use m631g* for Co
PURECART	11	
SYMMETRY_IGNORE	1	
UNRESTRICTED	TRUE	

SCF_GUESS	sad	
JOBTYPE	OPT	
MAX_SCF_CYCLES	2000	
SCF_CONVERGENCE	7	
GEOM_OPT_MAX_CYCLES	500	
SOLVENT_METHOD	PCM	!include H2O solvent
$end		

$pcm	
Theory	CPCM
Method	SWIG
Solver	Inversion
HeavyPoints	194
HPoints	194
Radii	read
vdwScale	1.2
$end	

$solvent	
Dielectric	78.39
$end	

$van_der_waals	
1	
1	1.2
6	1.92
7	1.55

8 1.52

9 1.5138

15 1.8657

16 1.8153

17 1.782

27 2

35 1.8855

$end

$basis

Co 0

insert modified 6-31G* as shown in S1

****

H 0

6-31g**

****

C 0

6-31g**

****

O 0

6-31g**

****

F 0

6-31g**

****

N 0

6-31g**

****

S 0

6-31g**

****

P 0

6-31g**

****

Cl 0

6-31g**

****

Br 0

6-31g**

****

$end

Table S3: Sample input file for geometry optimization of Cp fragment

$molecule

0 2

#insert xy coordinates for Cp fragment

$end

$rem

MEM_TOTAL	20000
MEM_STATIC	5000
METHOD	B3LYP
DFT_D	D3_ZERO
BASIS	6-31g**
SYMMETRY_IGNORE	1
UNRESTRICTED	TRUE
SCF_GUESS	sad
JOBTYPE	opt
MAX_SCF_CYCLES	1000
SCF_CONVERGENCE	7
GEOM_OPT_MAX_CYCLES	1000
SOLVENT_METHOD	PCM !include H2O solvent

$end

$pcm

Theory	CPCM
Method	SWIG
Solver	Inversion
HeavyPoints	194

HPoints 194

Radii read

vdwScale 1.2

$end

$solvent

Dielectric 78.39

$end

$van_der_waals

1

1 1.2

6 1.92

7 1.55

8 1.52

9 1.5138

15 1.8657

16 1.8153

17 1.782

35 1.8855

$end

Table S4: Basis sets investigation

Basis set	Medium	Co[CpY][CpY']OH energy (Eh)	Y'	CoCpY' energy (Eh)	Y	CpY energy (Eh)	BDE (kcal/mol)
m6-31g*,6-31g**		-1845.4877		-1651.9874		-193.4741	16.4137
m6-31g*,6-311++G(d,p)	Gas	-1845.6064	H	-1652.0675	H	-193.5193	12.2950
G3MP2Large,G3MP2Large		-1845.8533		-1652.2964		-193.5320	15.6590
m6-31g*,6-31g**		-2305.0822		-1651.9874		-653.0759	11.8696
m6-31g*,6-311++G(d,p)	Gas	-2305.2438	H	-1652.0675	Cl	-653.1497	16.6565
G3MP2Large,G3MP2Large		-2305.4949		-1652.2964		-653.1676	19.3922
m6-31g*,6-31g**		-2018.8076		-1691.3099		-327.4776	12.5660
m6-31g*,6-311++G(d,p)	Gas	-2018.9688	CH_3	-1691.3943	$N(CH_3)_2$	-327.5539	12.8819
G3MP2Large,G3MP2Large		-2019.2254		-1691.6385		-327.5755	7.1571
m6-31g*,6-31g**		-1845.5210		-1652.003971		-193.4746834	26.5672
m6-31g*,6-311++G(d,p)	PCM	-1845.6699	H	-1652.0860	H	-193.5248	37.0337
G3MP2Large,6-311++G(d,p)		-1845.8919		-1652.3134		-193.5248	33.7035
m6-31g*,6-31g**		-2305.1075		-1652.0040		-653.0807	14.3266
m6-31g*,6-311++G(d,p)	PCM	-2305.2881	H	-1652.0860	Cl	-653.1550	29.4970
G3MP2Large,6-311++G(d,p)		-2305.5104		-1652.3134		-653.1550	26.3834
m6-31g*,6-31g**		-2018.8433		-1691.3245		-327.4890	18.7700
m6-31g*,6-311++G(d,p)	PCM	-2019.0325	CH_3	-1691.4158	$N(CH_3)_2$	-327.5669	31.2322
G3MP2Large,6-311++G(d,p)		-2019.2535		-1691.6427		-327.5669	27.5724

Table S5: Default and new parameters after optimization for the complete dataset.

Model	Default Parameter	Optimized Parameter
XGBoost	n_estimators = 100 learning_rate = 0.300000012 max_depth = 6	n_estimators = 125 learning_rate = 0.27 max_depth = 3
Random Forest	n_estimators = 100	n_estimators = 125
Bagged Tree	n_estimators = 10	n_estimators = 125

Table S6: Default and new parameters after optimization for refined data set

Model	Default Parameter	Optimized Parameters
XGBoost	n_estimators = 100 learning_rate = 0.300000012 max_depth = 6	n_estimators = 125 learning_rate = 0.20 max_depth = 4
Random Forest	n_estimators = 100	n_estimators = 99
Bagged Tree	n_estimators = 10	n_estimators = 99

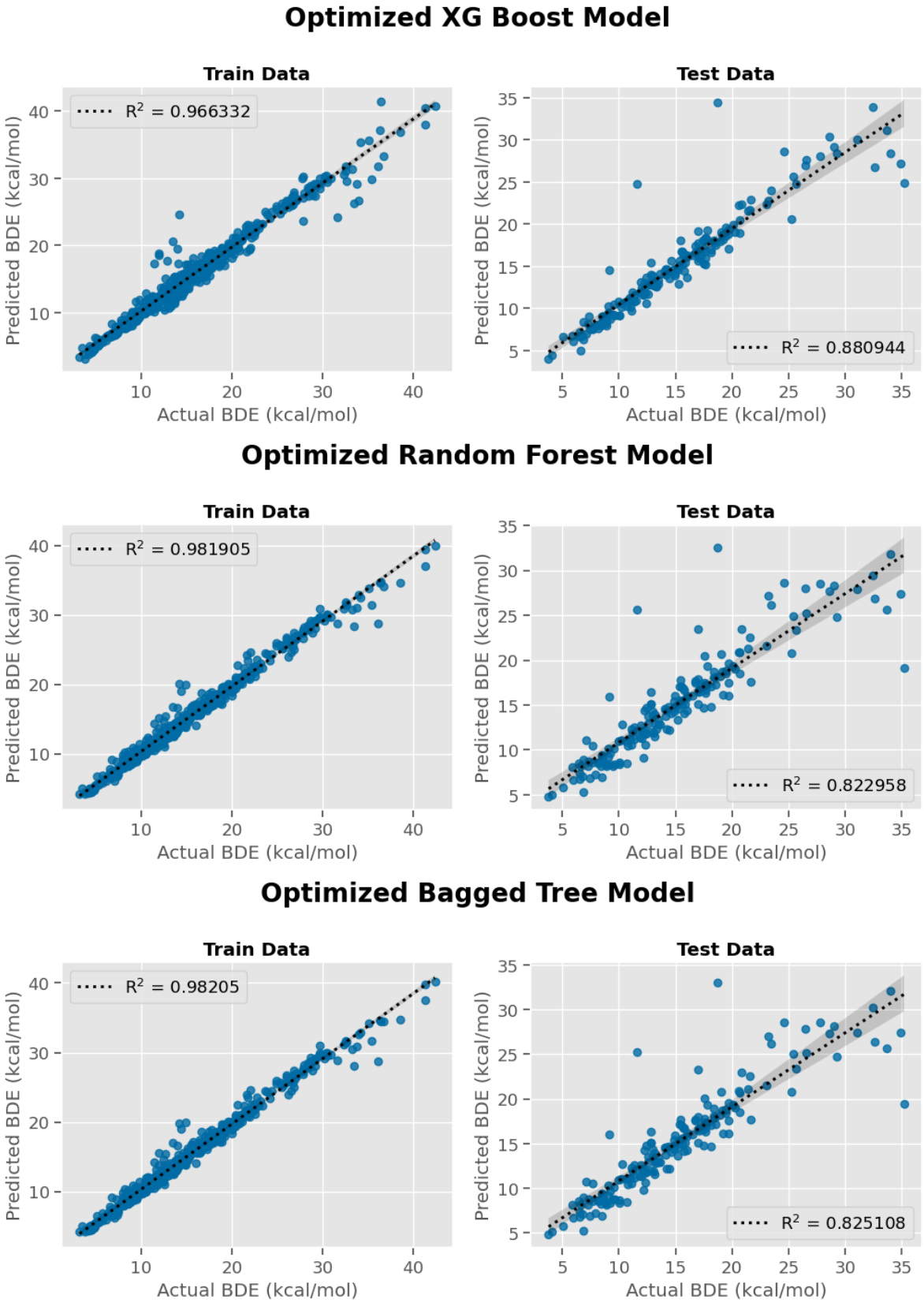


Figure S1: The train and test results for the optimized ML models.

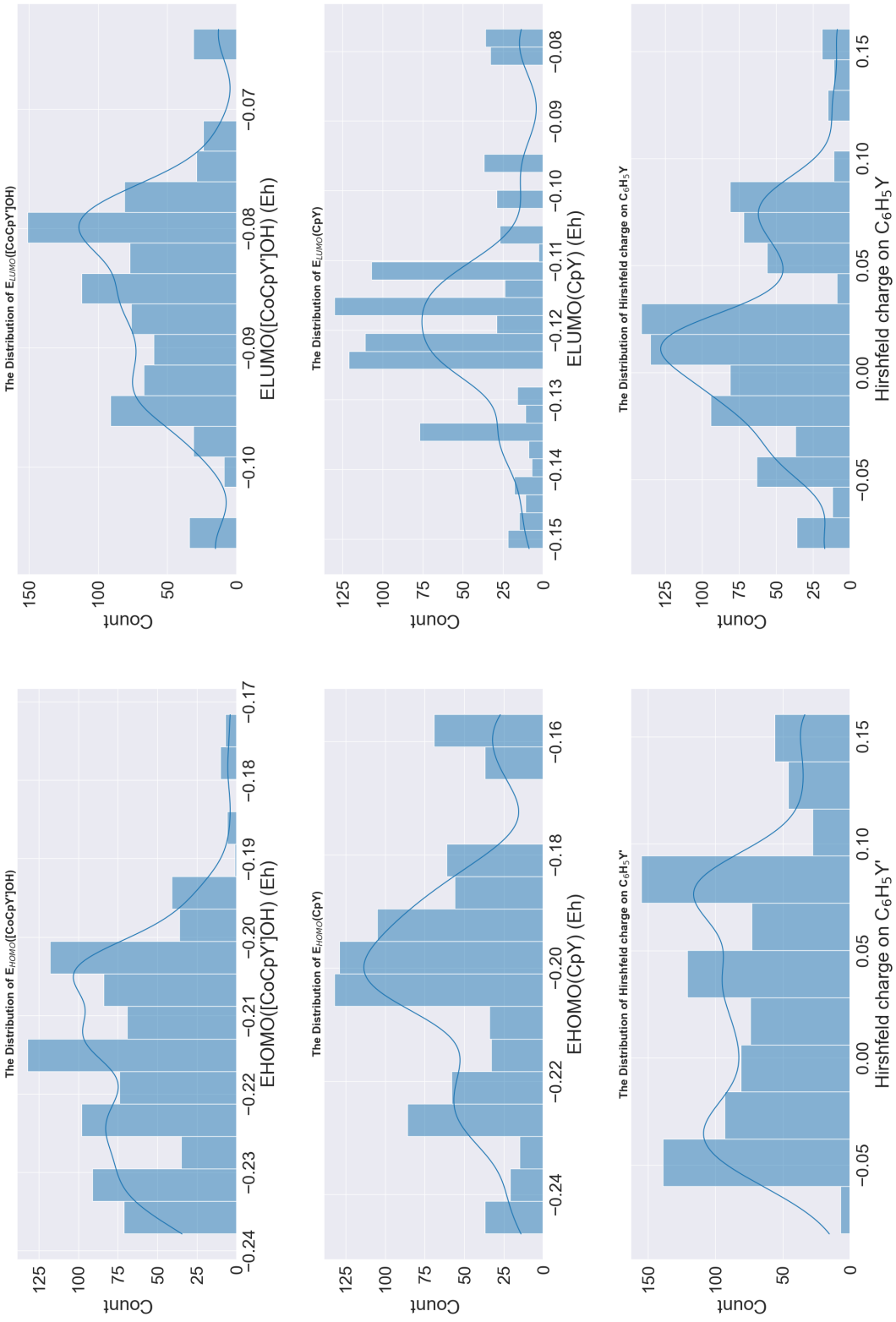
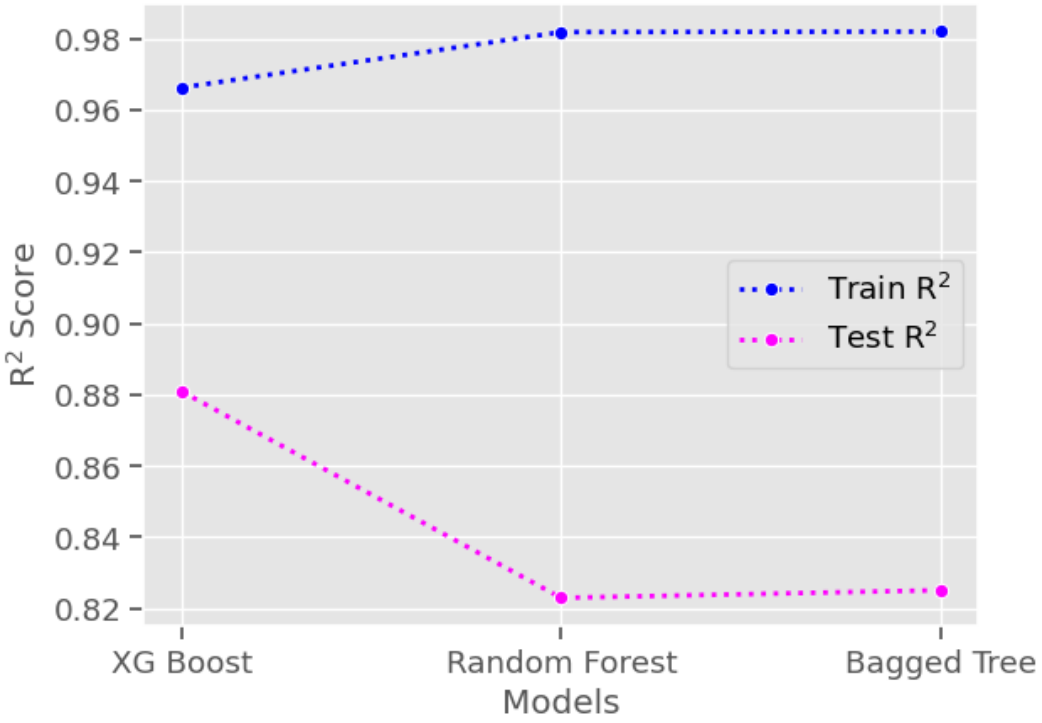
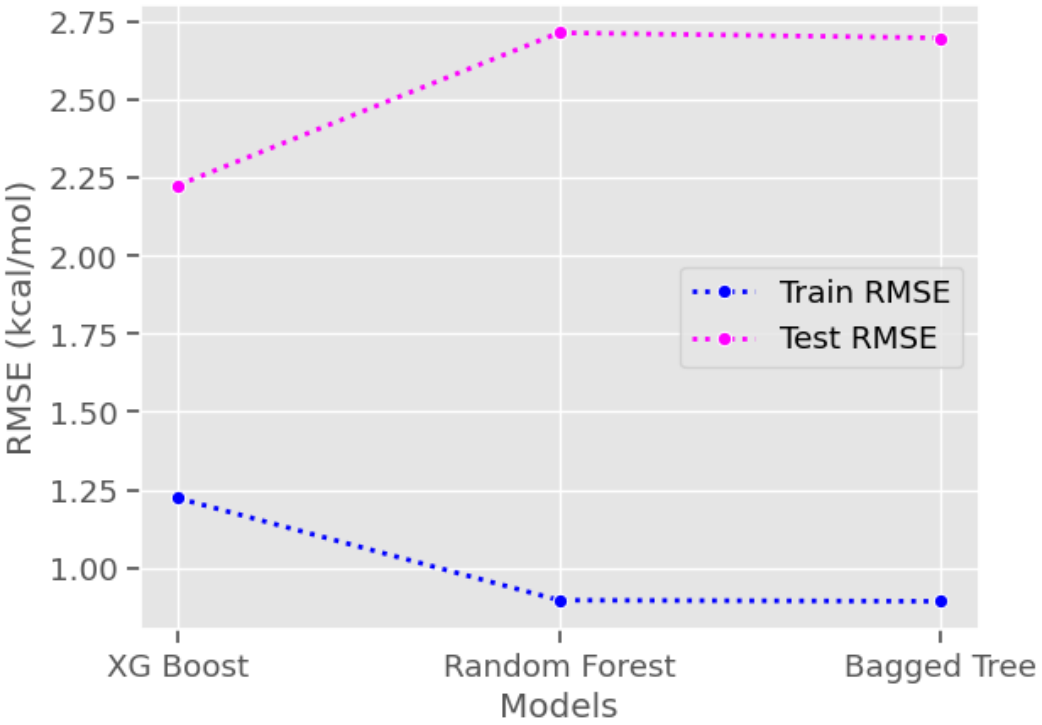


Figure S2: The distribution of each input features. The bins show the distribution as a histogram plot while the line indicates the kernel density estimate (KDE) plot.



(a) fig:R2



(b) RMSE

Figure S3: (a) Evaluation of optimized models with R^2 Scores, (b) Evaluation of optimized models with RMSE

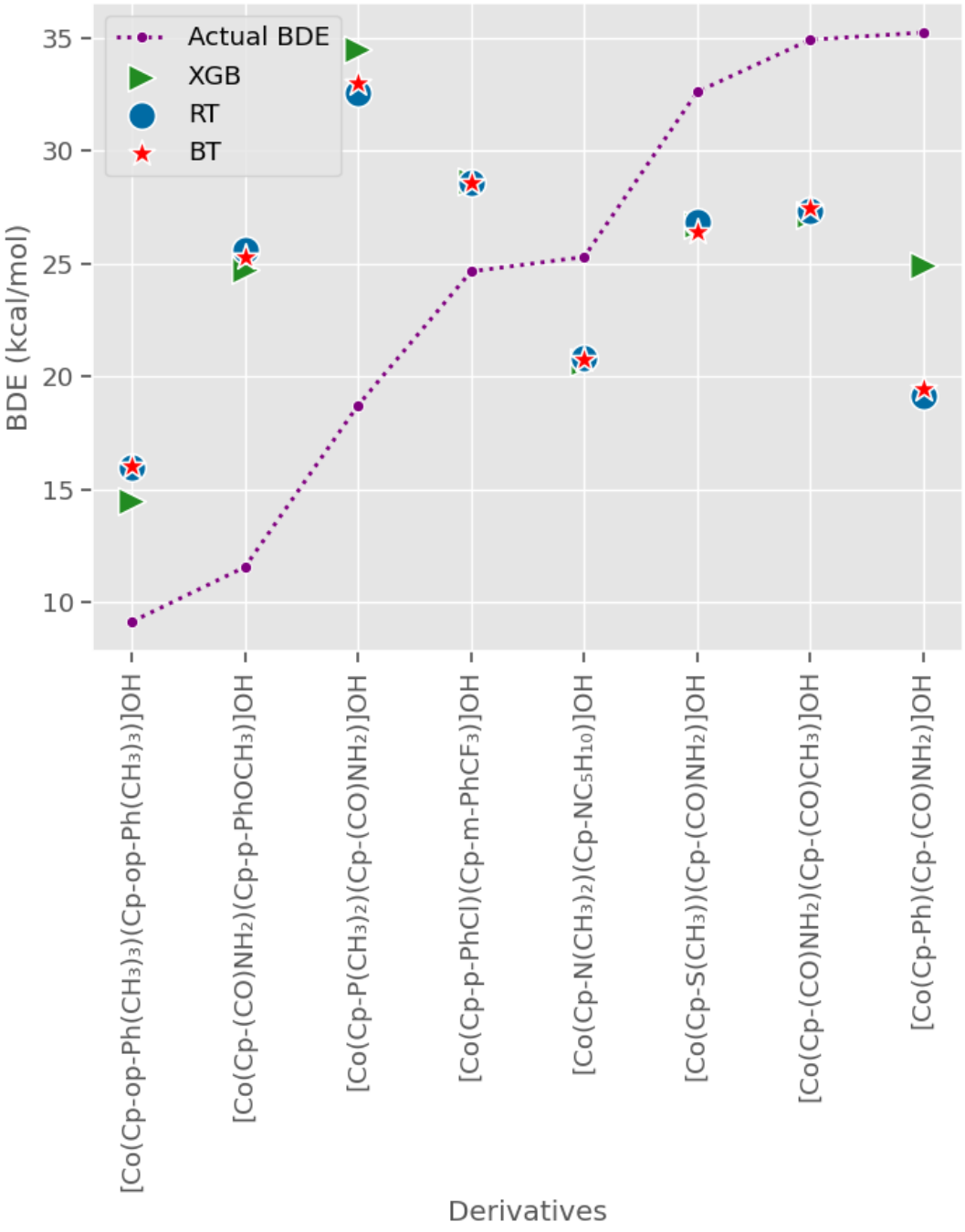


Figure S4: Actual and predicted BDE for top 8 outliers

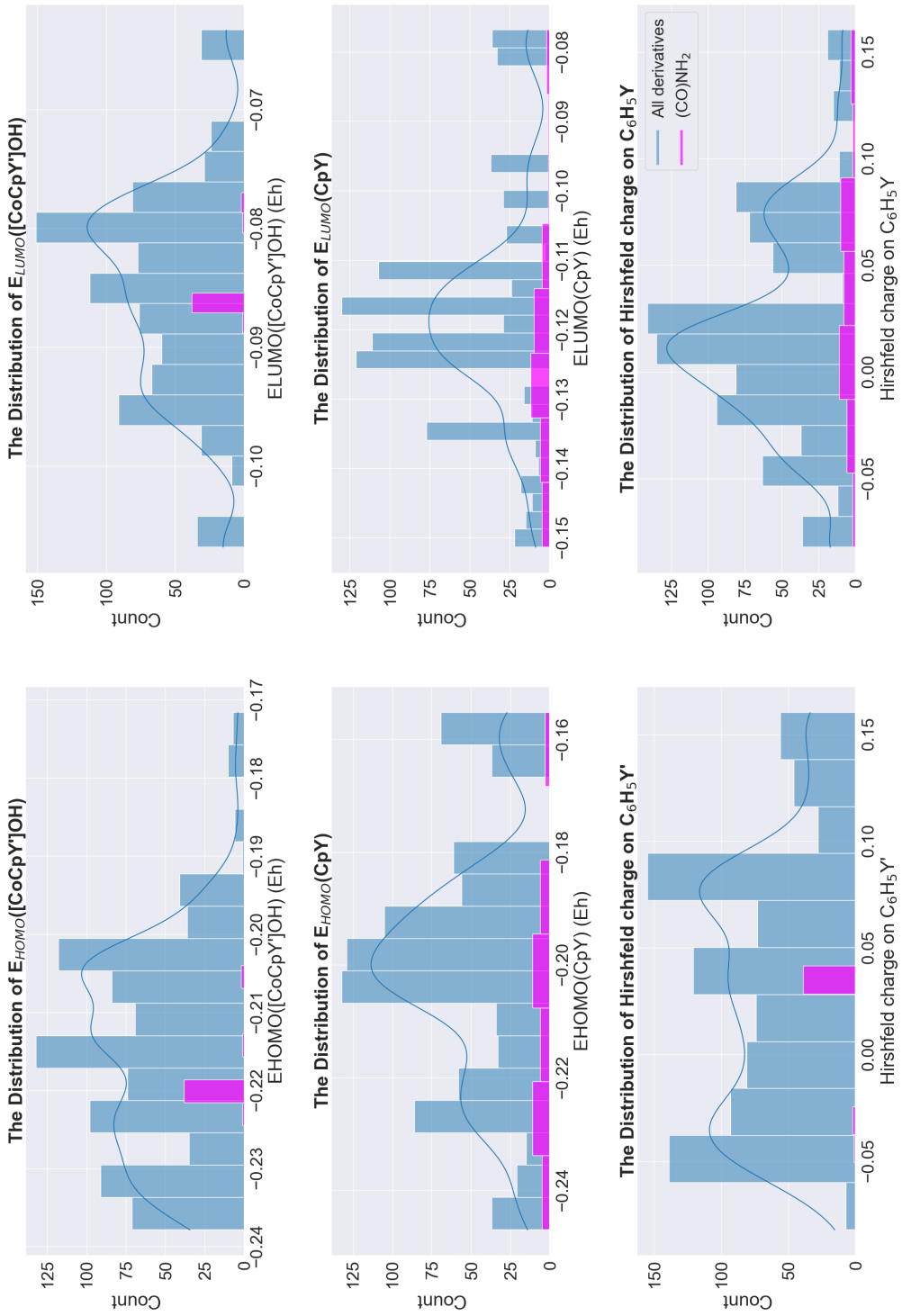


Figure S5: Distribution of $(CO)NH_2$ derivatives over input features. The blue bins show the distribution of all derivatives as a histogram plot and blue line indicates its KDE plot. The magenta colored bins show the distribution of $(CO)NH_2$ derivatives as a histogram plot.

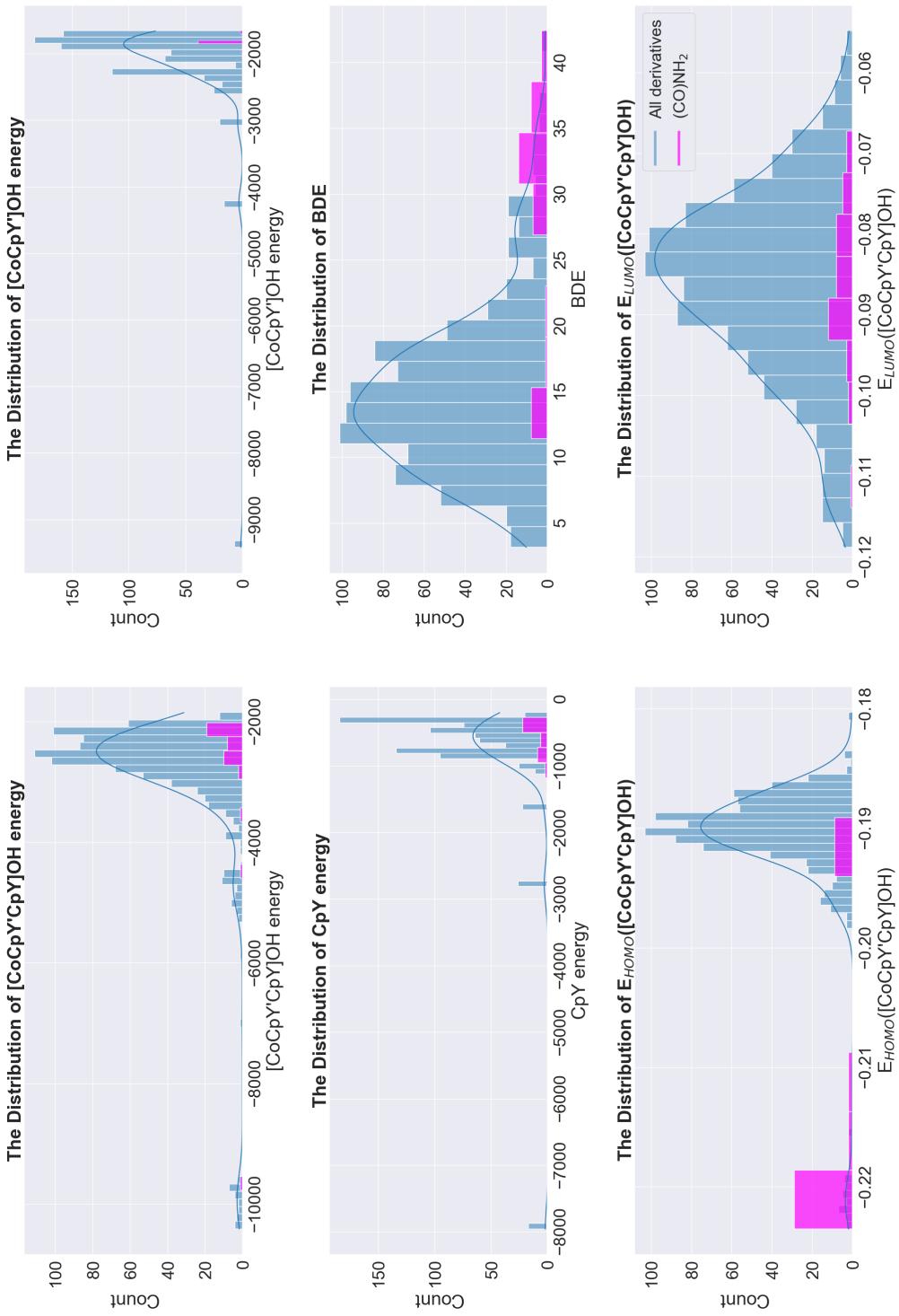


Figure S6: Distribution of $(CO)NH_2$ derivatives over additional features. The blue bins show the distribution of all derivatives as a histogram plot and blue line indicates its KDE plot. The magenta colored bins show the distribution of $(CO)NH_2$ derivatives as a histogram plot.

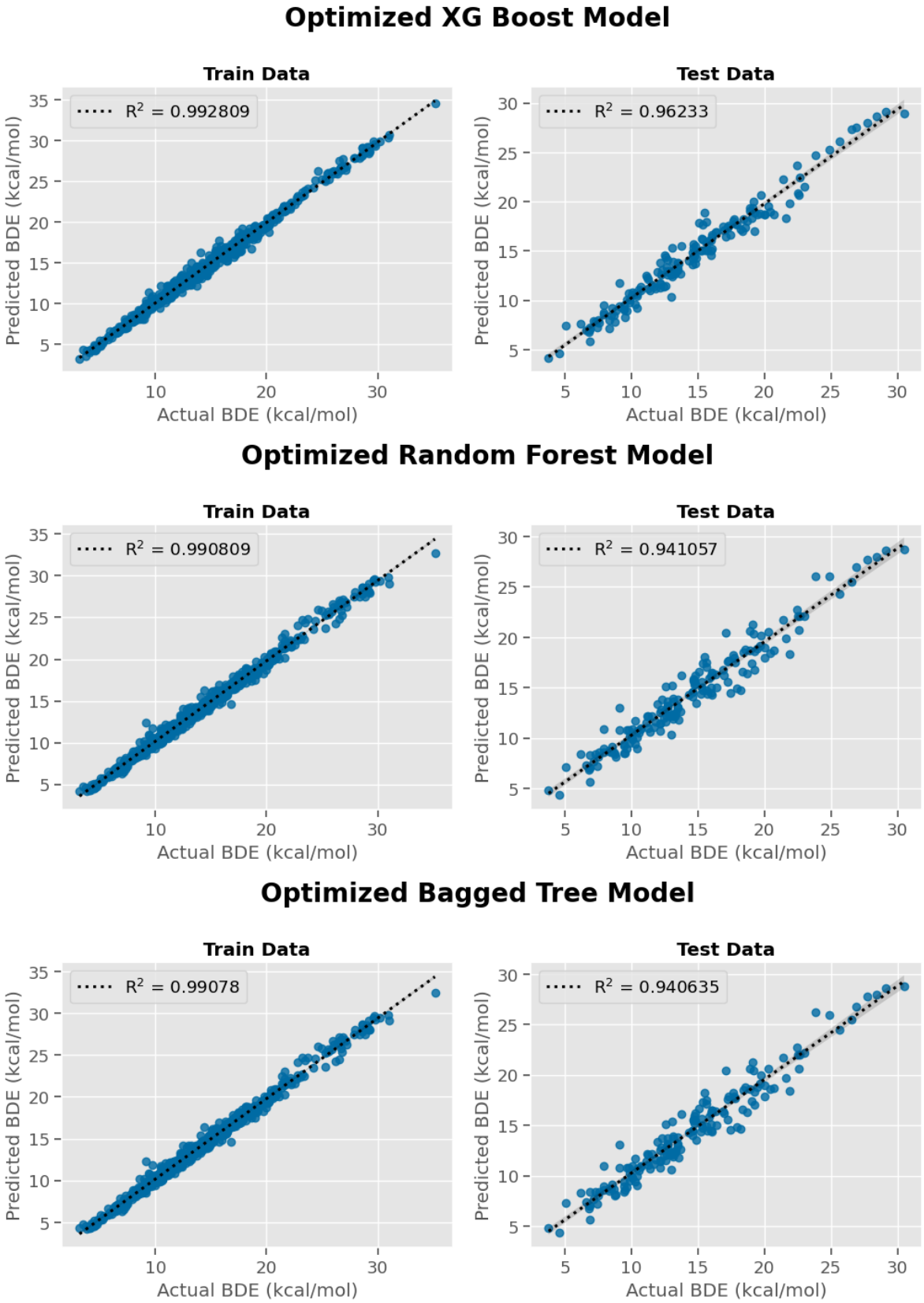


Figure S7: The train and test results of the optimized models for the refined dataset.

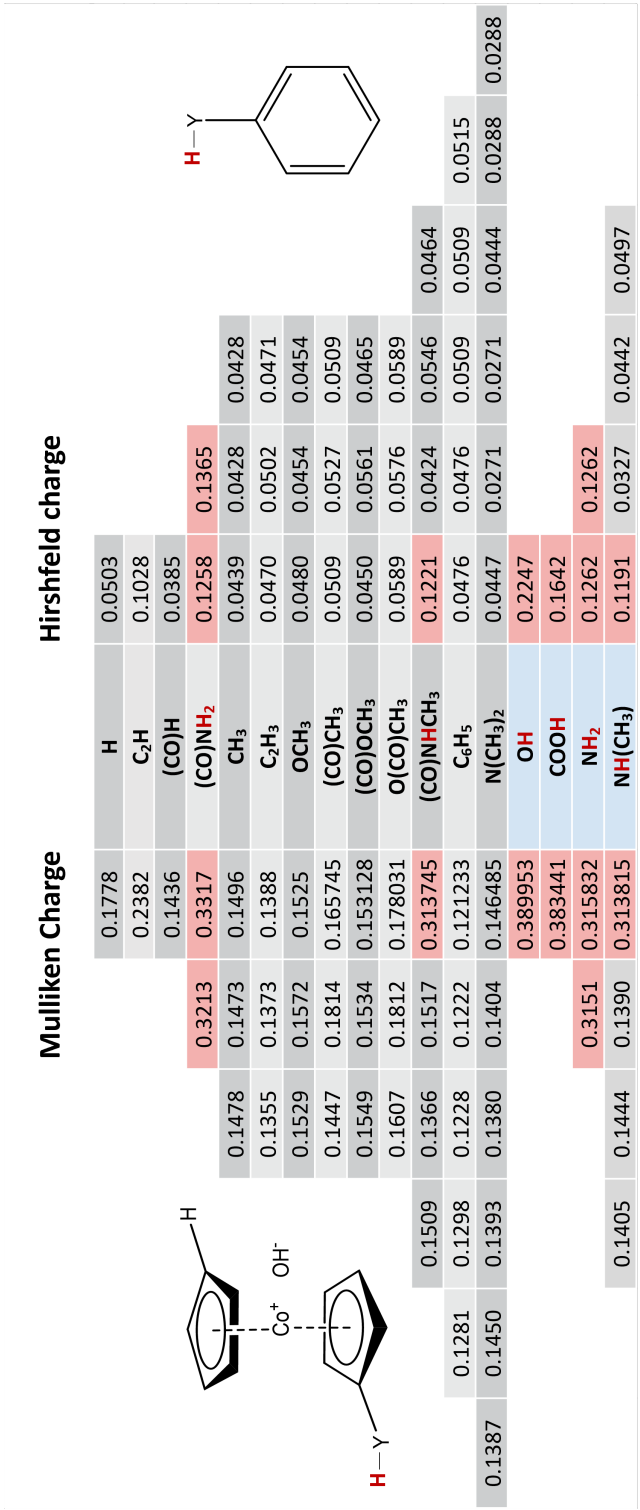


Figure S8: The Mulliken charge on H of the Y-group of [CoCpHCpY]OH, and the Hirshfeld charge on H on the Y-group of C_6H_5Y. The examined H-atom is highlighted with red in chemical structures. The values highlighted with red has the highest charge on H atoms which imply high acidity levels.