Using Machine Learning to Predict the pK_a of C–H Bonds. Relevance to Catalytic Methane Functionalization

Christopher X. Zhou, William M. Grumbles, Thomas R. Cundari*

Department of Chemistry, Center for Advanced Scientific Computing and Modeling (CASCaM), Denton, TX 76201

Abstract – Six machine learning models (random forest, neural network, support vector machine, k-nearest neighbors, Bayesian ridge regression, least squares linear regression) were trained on a dataset of 3d transition metal-methyl and -methane complexes to predict pK_a(C–H), a property demonstrated to be important in catalytic activity and selectivity. Results illustrate that the machine learning models are quite promising, with RMSE metrics ranging from 4.6 to 8.8 pK_a units, despite the relatively modest amount of data available to train on. Importantly, the machine learning models agreed that (a) conjugate base properties were more impactful than those of the corresponding conjugate acid, and (b) the energy of the highest occupied molecular orbital conjugate base was the most significant input feature in the prediction of pK_a(C–H). Furthermore, results from additional testing conducted using an external dataset of Sc-methyl complexes demonstrated the robustness of all models, with RMSE metrics ranging from 1.5 to 6.6 pK_a units. In all, this research demonstrates the potential of machine learning models in organometallic catalyst development.

Introduction

Methane is the primary component of natural gas, an abundant domestic energy resource. The process of catalytic methane functionalization has several significant industrial implications. First, the facile conversion of methane into methanol would provide a more efficient and cleaner source of energy than current methods of burning coal, petroleum or natural gas. Second, methanol – a liquid at ambient conditions – would be more convenient to transport than methane, as the latter it is quite costly to transport in its natural gaseous form. Third, methanol is itself a useful industrial chemical. Because of the thermodynamic strength and kinetic inertness of methane's C–H bonds, methane functionalization is presently infeasible without the aid of catalysts.

One area of research that has long focused on the potential applications of methane functionalization is the field of organometallic catalysis. While there have been many advances in developing such catalysts, there are still several challenges in being able to create catalysts that can selectively activate and functionalize methane.[1,2] With the many transition metals that can be used to define a catalyst's active site(s) and the near limitless combinations of supporting ligands, substituents, *etc*. that could be tested, it is not practical to evaluate each possibility, even with a search strategy largely or even completely limited to computational chemistry. Utilizing machine learning (ML) algorithms is, therefore, an attractive solution to the combinatorial problem of catalyst design. Machine learning is a broad field of techniques designed to discern relationships and patterns in data that are not immediately apparent via traditional analysis techniques. Ideally, ML allows one to produce predictions or decisions without humans providing explicit rules or instructions on how to do so. ML can also reveal important links between catalyst structure and catalyst performance, which itself is a great aid in the design scenario, indicating novel directions for both future computational and experimental research in catalysis.

Thus, it is of interest to investigate how potential catalyst candidates for methane activation can be identified using machine learning algorithms. In this study, several machine learning algorithms were investigated to determine the influence that various molecular properties of 3d metal ion complexes have on the calculated pK_a of the C–H bonds of coordinated methyl and methane ligands. Part and parcel of this study is the underlying hypothesis that the intrinsic acid/base properties of a hydrocarbon/hydrocarbyl C–H bond – and how these are controlled within the coordination sphere of a metal catalyst – are an important indicator of catalyst activity and/or selectivity. A recent study as well as several classical studies have indicated just such a link.[3,4,5] A recent DFT and coupled cluster study of methane adducts showed a direct connection between enhanced acidity of a methane C–H bond upon ligation to a 3d metal and a reduction in the subsequent barrier to C–H activation.[5] Fallah *et al.* likewise showed that deprotonation of methyl C–H bonds was a competing side-reaction to methyl–X functionalization, implying that the acid-base properties of hydrocarbyl C–H bonds impact catalyst selectivity.[6] More recent research by Grumbles and Cundari indicates that metal and supporting ligand effects on organometallic pK_a(C–H) values of methyl ligands are commensurate with, if not greater than, traditional inductive and resonance effects for organic acids.[7]

In all, six different machine learning models were tested in this research: neural network, support vector machine (SVM), k-nearest neighbors (kNN), Bayesian ridge regression, least squares linear regression, and random forest techniques. All models were trained on a data set of pK_a(C–H) values derived from density functional theory calculations on 91 transition metal methane/methyl structures, with a 75%/25% training-testing split. The data points were sequestered randomly into training and testing sets, with a set random seed for reproducibility. To facilitate comparison among the different ML techniques, twenty-one (21) input features were chosen to be utilized in the training of each model (**Figure 1**). These are typical and easily obtained atomic and molecular descriptors that one may assume to inform the resulting acid/base properties of a C–H bond within an organometallic environment.

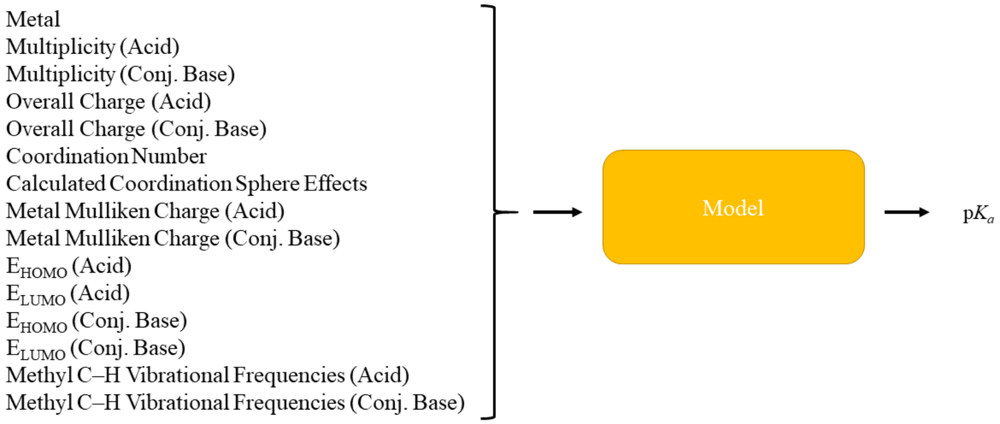

Metal
Multiplicity (Acid)
Multiplicity (Conj. Base)
Overall Charge (Acid)
Overall Charge (Conj. Base)
Coordination Number
Calculated Coordination Sphere Effects
Metal Mulliken Charge (Acid)
Metal Mulliken Charge (Conj. Base)
E_{HOMO} (Acid)
E_{LUMO} (Acid)
E_{HOMO} (Conj. Base)
E_{LUMO} (Conj. Base)
Methyl C–H Vibrational Frequencies (Acid)
Methyl C–H Vibrational Frequencies (Conj. Base)
Model
pK_a

Figure 1. Representative diagram of a ML model generating organometallic pK_a(C–H) predictions by using twenty-one features as inputs.

To calculate the pK_a(C–H) values, all structures were modeled in the context of a Brønsted-Lowry acid-base reaction, in which a 3d transition metal methane adduct (or methyl complex) is deprotonated by a DMSO solvent molecule, resulting in the formation of a methyl (or methylidene) conjugate base and the conjugate acid of DMSO (**Figure 2**).

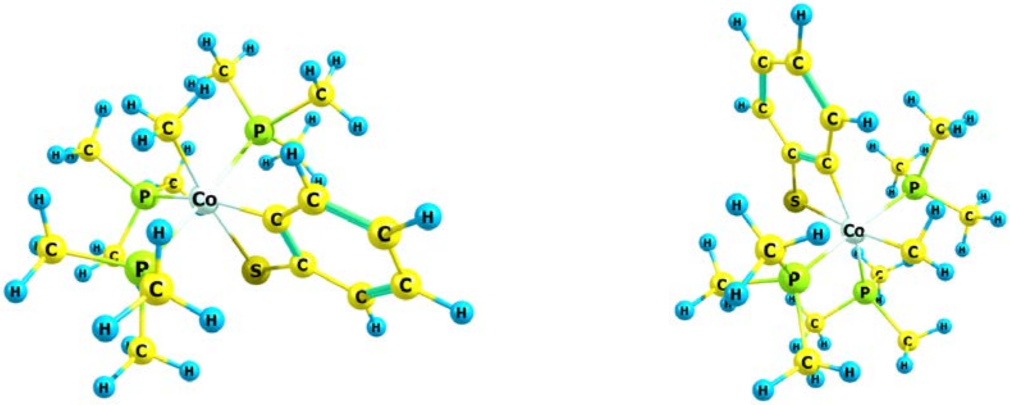


Figure 2. A representative Co-methyl "acid" modeled in this study (CSD refcode: AGODAC[8]), left. Its conjugate base is the corresponding anionic methylidene complex, right.

Computational Methods

All structures studied in this investigation were optimized with the Gaussian 09/16 software packages,[9,10] using the BMK/6-31+G(d) level of theory. The pK_a of methyl[5] and methane[7] C–H bonds for each structure was calculated using the following equations:

$$\mathrm{p}K_a = \frac{\Delta \mathrm{G}}{2.303 \cdot \mathrm{RT}}$$

$$\mathrm{p}K_{a_{corr}} = 1.0308 \cdot \mathrm{p}K_a - 12.146$$

with a linear correction factor derived from the work of Nazemi and Cundari.[11]

While many strategies for feature selection have been reported in the literature, in the present work a simple strategy was selected. A Python program was developed to automate the extraction of these features from the Gaussian output of the DFT-optimized structures. All metal-methyl complexes submitted to quantum calculations were found in the Cambridge Structural Database,[8] while all metal-methane complexes used for this study were sourced from a previous study.[5] However, one of the features, termed CCSE, was not directly extracted from the DFT-optimized structures. It was devised to account for electrostatic effects that different modes of coordination have on the pK_a(C–H). This feature was calculated with the following equation, using the computed Mulliken charges extracted from the DFT-optimized acid structures:

$$CCSE = \sum_i \frac{M_{C_{metal}} \cdot M_{C_i}}{r_i^2}$$

$M_{C_{metal}}$ denotes the Mulliken charge on the metal center of the complex, while M_{C_i} denotes the Mulliken charge on the i^{th} atom in the complex, which is distance r_i away from the metal center.

Results and Discussion

In **Figure 3**, calibration plots for all six tested ML models are displayed. Observing the R^2 values of each plot, all models show quite promising predictive performance. Using the root mean square error (RMSE) between DFT-calculated and ML-predicted pK_a(C–H) values as a metric, the following values (in pK_a units) are obtained: RF (4.6), 2LNN (7.2), SVM (5.6), KNN (8.8), BRR (5.3), LSLR (5.4), **Figure 3**. The training and evaluation of each model was also performed in Python, using scikit-learn machine learning model implementations.[12]

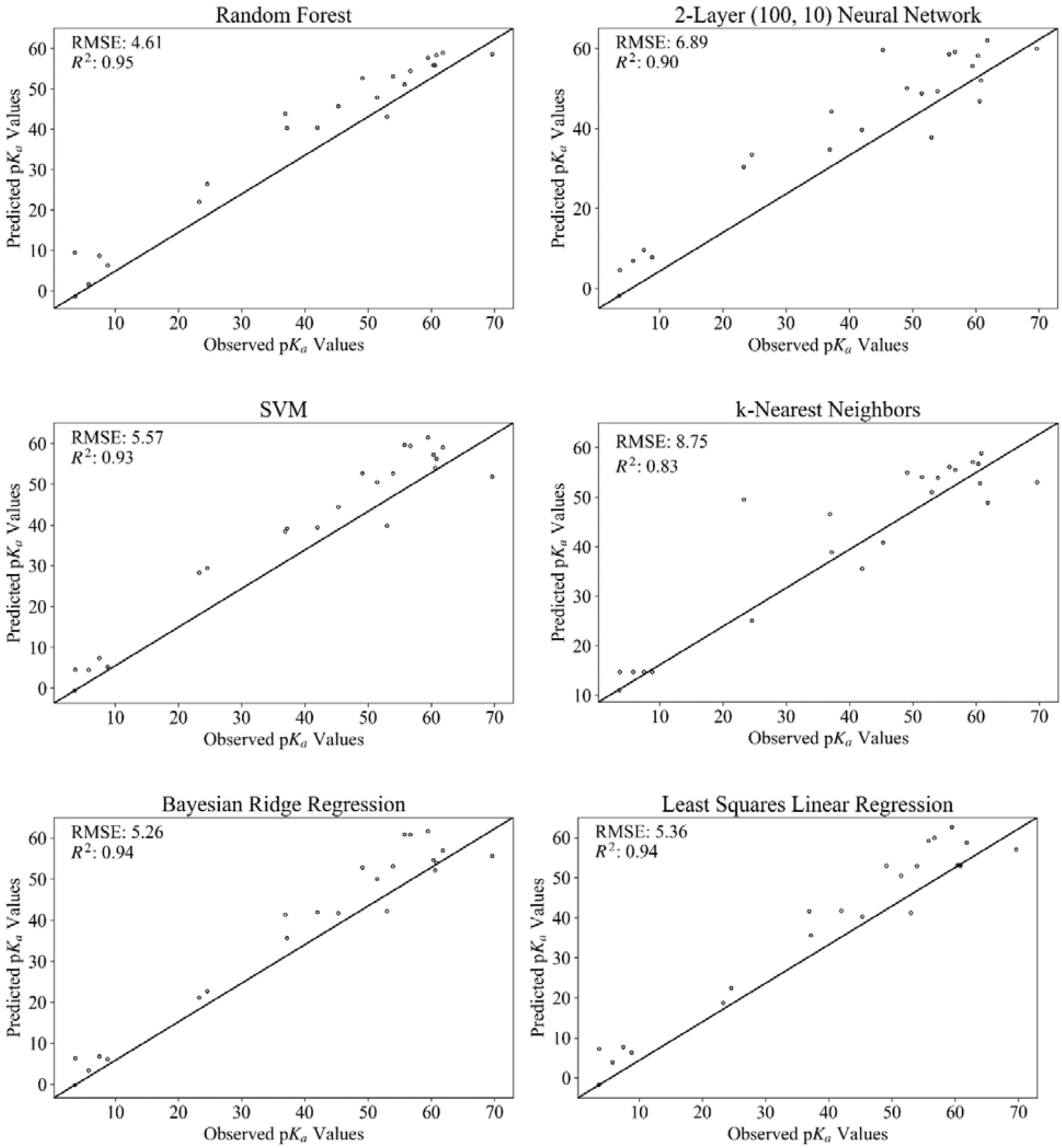


Figure 3. Calibration plots of six studied machine learning models for pK_a(C–H). RMSE = root mean square error .

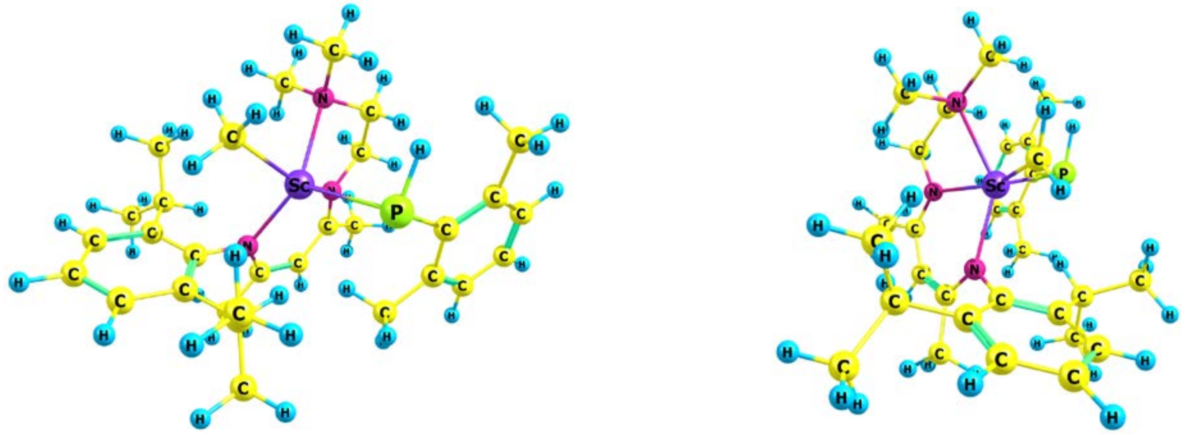
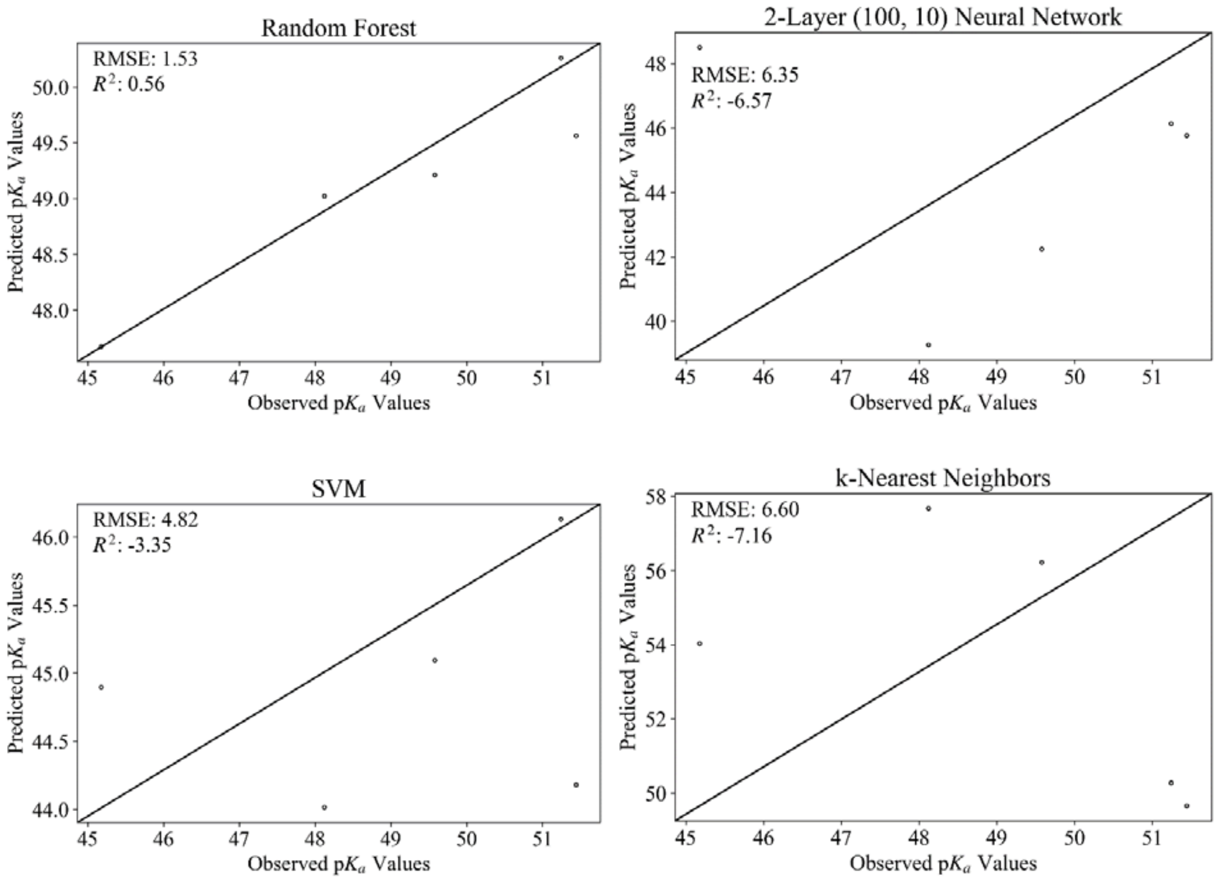


Figure 4. A representative Sc-methyl “acid” from the external dataset (CSD refcode: XUKWUX[8]), left. The corresponding conjugate base is also displayed, right.



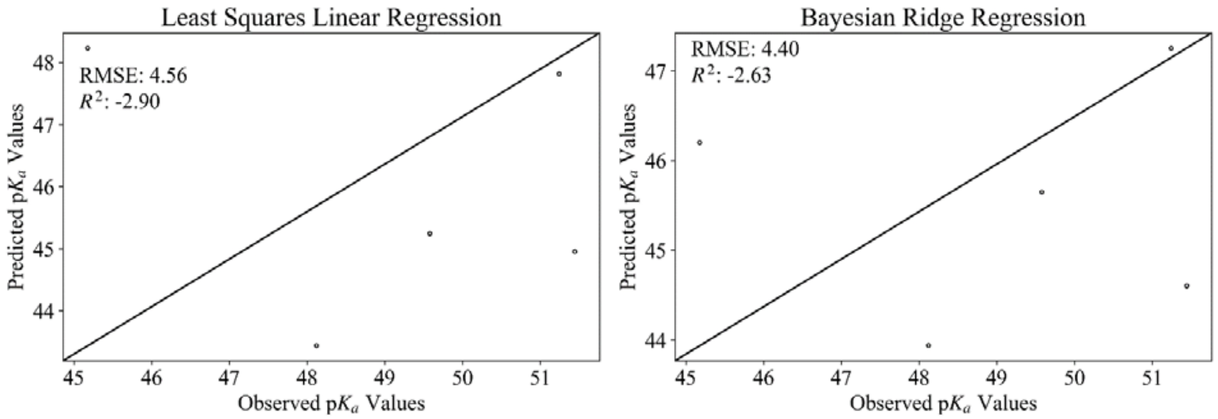


Figure 5. Calibration plots of six studied machine learning models, tested on an additional external dataset of six scandium-methyl complexes. As the points are in most cases randomly scattered, many of the R^2 values are negative and quite large in magnitude; this seems to be a clear indication that the dataset is sparse, rather than a sign of poor model performance, since measured RMSE values are still quite reasonable.

Using RMSE as a performance metric, the RF technique had the best performance in terms of generalization, *i.e.*, the ability to reproduce DFT-calculated pK_a(C–H) values of testing set methane/methyl complexes. It is worth noting that even with modern multi-core, high performance computing resources typical DFT geometry optimization/frequency protocols – assuming no complications from spurious imaginary frequencies, SCF convergence difficulties, *etc.* – take roughly a day per calculation. A typical training/testing simulation – after identification of appropriate ML architectures and protocols – takes on the order of a few seconds per simulation.

As an additional test, the six ML models investigated in this study were evaluated on a separate data set of Sc-methyl complexes to gauge the generalization power of each, given that there were no scandium-methyl complexes in the original training/testing sets. All complexes were found in the Cambridge Structural Database (**Figure 4**) and their pK_a(C–H) values calculated using the same procedures summarized in Computational Methods.[8] While sparse, the external test set illustrates that all models perform reasonably well on previously unseen data (**Figure 5**). However, the random forest model performs exceedingly well on this external test set with an RMSE of 1½ pK_a units, **Figure 5**, roughly one-third of any other method tested herein.

While an ML model that can both memorize (training) and generalize (testing) the subject database is valuable for understanding trends in a given data set, it is as important in a catalyst design scenario that chemical insight be extracted from such simulations. Therefore, each ML model was further evaluated to determine the features most important to the prediction process. Feature importance was assessed by iteratively removing each of the 21 input features, **Figure 1**, shuffling the order of the remaining features, and retraining the model and calculating the percent increase in the root mean square error (RMSE). This error metric was obtained by performing a leave-one-out (LOO) cross validation on each model. Examining these evaluations (**Figure 6**), results show that nearly all models agreed that the $\mathbf{E}_{HOMO}$ energy of the conjugate base was the most important feature in determining pK_a(C–H). Interestingly, after the evaluation process, the two lowest performing models, NN and kNN, both indicated a number of redundant features with much higher percent decreases in RMSE as compared to other models. Another point of interest is that both the LSLR and BRR models demonstrate striking similarities in their ranking of features' importance; this is likely due to the similar nature of these linear models.

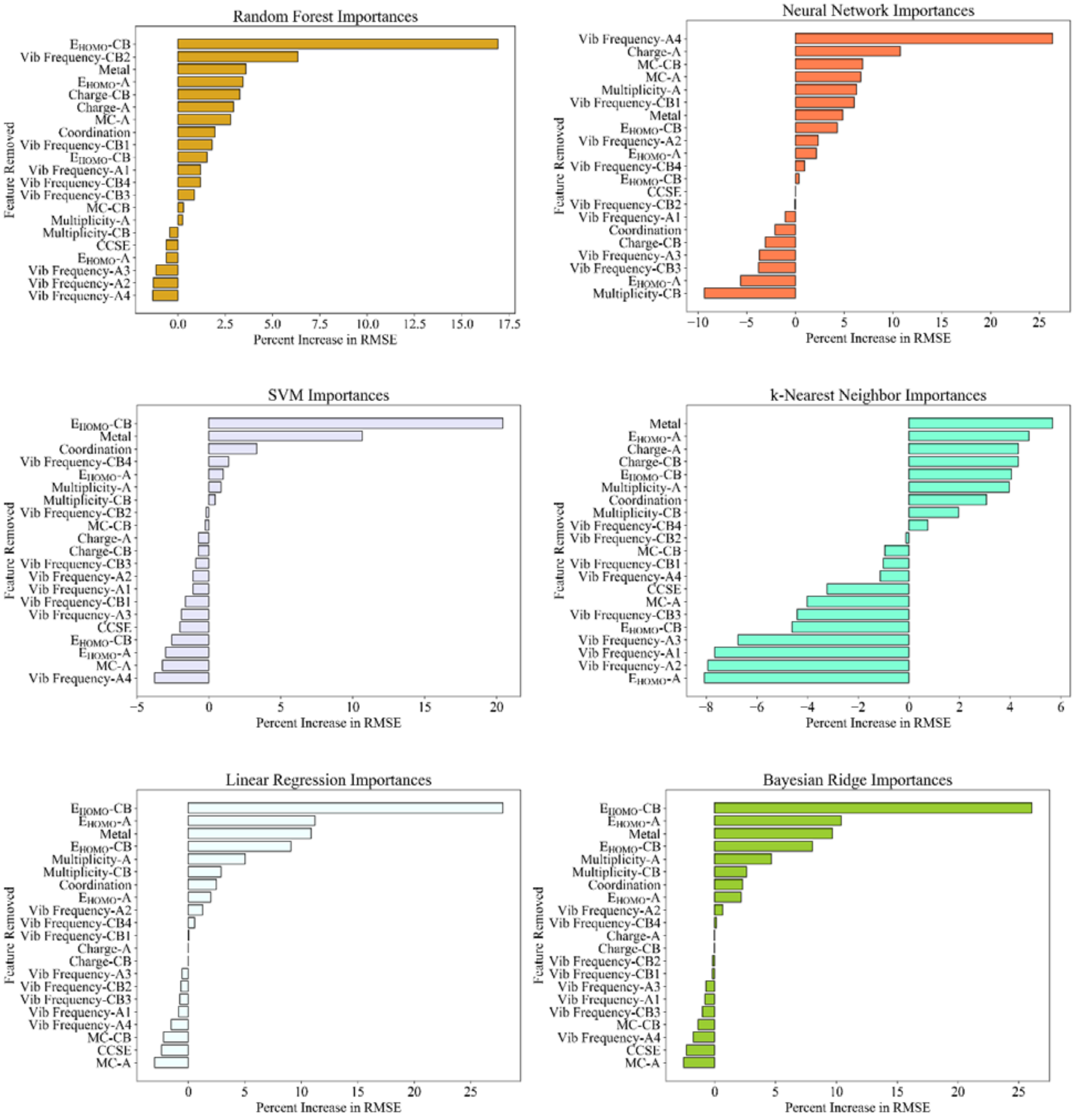


Figure 6. Importance of features determined by each of the six tested ML models. Features suffixed with "A" represent features extracted from the DFT-optimized acid structure, and likewise those suffixed with "CB" refer to features extracted from the DFT-optimized conjugate base structure. "Vib Frequency" features refer to the calculated IR stretching frequencies for each structure's target methane/methyl/methylidene ligand. Each "Vib Frequency" feature includes a suffixed number as well, which represents the original order of the frequencies in the DFT optimization (lowest to highest). "MC" is the abbreviation for the DFT-calculated Mulliken charge on the complex's metal center. Metal refers to the atomic number of each complex's metal center. "CCSE" is an abbreviation of calculated coordination sphere effects; its calculation can be referenced in Computational Methods.

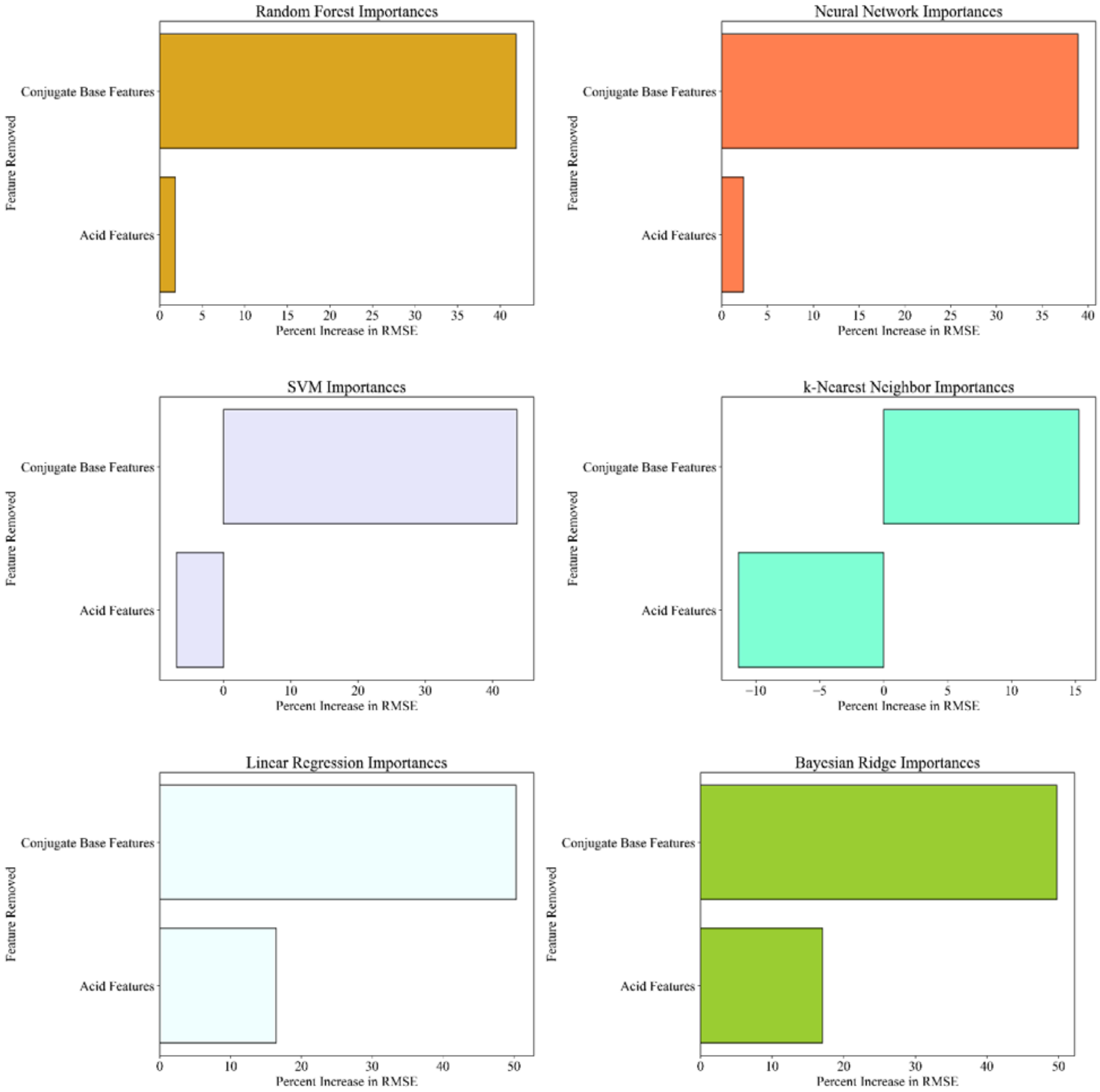


Figure 7. Importance of acid vs. conjugate base features for all six machine learning models.

In a similar fashion, the models were also evaluated to determine whether the acid or conjugate base species was more important to generating predictions. Interestingly, from the data in **Figure 7**, in all cases, conjugate base input features were overwhelmingly more important than acid descriptors.

Summary and Conclusions

In conclusion, we have observed that several common machine learning models and one more unconventional (random forest) technique all have extraordinary predictive power in approximating the pK_a(C–H) values of organometallic methane and methyl complexes. This is clearly evidenced by the low RMSE metrics and high positive R^2 correlations obtained, despite the modest amount of data available to train on. Moreover, it is interesting to note how accurate both linear models (least squares linear regression and Bayesian ridge regression) are compared to the other more complex ML techniques. Furthermore, in performing extra evaluations on the models, it was found that not only do all have promising generalization power – displaying the ability to predict a novel selection of scandium-methyl complexes, with the RF model being clearly the best. All models also agreed that conjugate base input features are more significant than those of the conjugate acid in the pK_a(C–H) prediction process. From a chemical perspective, this may hint, for example, that the stability of the metal-methylidene moiety is most critical in determining the acidity of its metal-methyl conjugate acid. A final point of interest lies in the fact that feature importance evaluations revealed that overall, $\mathbf{E}_{HOMO}$ energy of the conjugate base was the single most important feature.

The present study has focused on uncovering how effective machine learning algorithms are in predicting acid-base properties of coordinated methane and methyl. While these models clearly hold great potential, much remains to be uncovered, in particular, searching for more features that can be extracted directly from unoptimized structures, *e.g.*, from a chemical graph or easy-to-derive one- and two-dimensional chemical descriptors.

Supporting Information

Cartesian coordinates of all BMK/6-31+G(d) optimized geometries (XYZ). Extracted features from Gaussian outputs (CSV). Extraction program developed to parse Gaussian outputs into a CSV file (PY).

Acknowledgements

The authors thank the National Science Foundation for their generous support through grant CHE-1464943, and their support of the UNT CASCaM HPC cluster via grant CHE-1531468.

AUTHOR INFORMATION

Corresponding Author

*E-mail: t@unt.edu

Notes

The authors declare no competing financial interests.

References

(1) Cundari, T. R. Methane Manifesto: A Theorist's Perspective on Catalytic Light Alkane Functionalization. *Comments Inorg. Chem.* **2016**, 37 (5), 219–237.

(2) Caballero, A.; Pérez, P. J. Methane as Raw Material in Synthetic Chemistry: the Final Frontier. *Chem. Soc. Rev.* **2013**, 42 (23), 8809.

(3) Olah, G. A.; Schlosberg, R. H. Chemistry in Super Acids. I. Hydrogen Exchange and Polycondensation of Methane and Alkanes in FSO_3H-SbF_5 ("Magic Acid") Solution. Protonation of Alkanes and the Intermediacy of CH_5 and Related Hydrocarbon Ions. The High Chemical Reactivity of "Paraffins" in Ionic Solution Reactions. *J. Am. Chem. Soc.* **1968**, 90, 2726−2727.

(4) Streitwieser, A.; Taylor, D. R. Kinetic Acidity of Methane. *J. Chem. Soc. D* **1970**, 19, 1248.

(5) Zhou, C. X.; Cundari, T. R. Computational Study of 3d Metals and Their Influence on the Acidity of Methane C–H Bonds. *ACS Omega* **2019**, 4 (23), 20159–20163.

(6) Fallah, H.; Horng, F.; Cundari, T. R. Theoretical Study of Two Possible Side Reactions for Reductive Functionalization of 3d Metal−Methyl Complexes by Hydroxide Ion: Deprotonation and Metal−Methyl Bond Dissociation. *Organometallics* **2016**, 35, 950−958.

(7) Grumbles, W. M.; Cundari, T. R. Computational Determination of pK_a(C–H) in 3d Transition Metal-Methyl Complexes. *Organometallics* – accepted.

(8) Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge Structural Database. *Acta Crystallogr., Sect. B: Struct. Sci., Cryst. Eng. Mater.* **2016**, 72 (2), 171–179.

(9) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, V.; Barone, G.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Keith, T.;

Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, Revision E.01; Gaussian, Inc.: Wallingford CT, 2013.

(10) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, V.; Barone, G.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Keith, T.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 16*, Revision A.03; Gaussian, Inc.: Wallingford CT, 2016.

(11) Nazemi, A.; Cundari, T. R. Control of C-H Bond Activation by Mo-Oxo Complexes: pK_a or Bond Dissociation Free Energy (BDFE)? *Inorg. Chem.* **2017**, 56, 12319−12327.

(12) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research* **2011** 12, 2825-2830.