Tolman Electronic Parameter Predictions from a Fast, Accurate, and Robust Machine Learning Model Provide Insight into Phosphine Ligand Electronic Effects

Harlan Stevens[a], Jeffrey Olsen[a], Justin K. Kirkland[a,*], and Daniel H. Ess[a,*]

[a]Department of Chemistry and Biochemistry, Brigham Young University, Provo, UT 84604 USA

Abstract

Phosphines are extremely important ligands in organometallic chemistry and their donor or acceptor ability can be measured through the Tolman electron parameter (TEP). Here we describe the development of a TEP machine learning model (called TEPid) that provides nearly instantaneous calculation of experimentally calibrated CO vibrational stretch frequencies for $(R)_3P\text{-}Ni^0(CO)_3$ complexes. This machine learning model with an error of less than 1 cm^{-1} was developed using >4,000 DFT calculated $(R)_3P\text{-}Ni^0(CO)_3$ TEP values and 19 key connectivity-based descriptors associated with SMILES strings. We also built a web-based interface to run the machine learning model where SMILES strings can be entered and TEP values returned. We applied this TEPid model to examine the donor and acceptor capability of phosphines in the large Kraken phosphine database. Surprisingly, this showed that the Kraken database is skewed towards donor phosphines. In the same spirit of the Kraken database, we generated tens of thousands of new experimentally based phosphines that when combined with Kraken phosphines provide a more electronically balanced ligand library.

Introduction

Phosphines represent a ubiquitous and extremely important class of ligands in organometallic chemistry, and their σ-donor properties greatly affect the coordination, thermodynamic stability, and reactivity of transition metal complexes. Phosphines have been shown to significantly modulate the reactivity of many organometallic catalyzed reactions, such as hydroformylation,[1-3] cross-coupling,[4-9] and hydrogenation.[10, 11] To develop new organometallic catalytic reactions where stability and reactivity is modulated by phosphines it would be desirable to have extremely fast methods to calculate and explore phosphine properties. In this general effort, Gensch, Sigman, and Aspuru-Guzik recently reported the Kraken database that consists of over 300,000 mostly hypothetical phosphine ligands created based on a combination of experimentally known phosphines.[12] Along with structures, Kraken contains several calculated steric and electronic descriptors. Fey has demonstrated and popularized the utility of ligand descriptors and properties to evaluate possible catalytic activity.[13-21]

One major chemical descriptor that has been used for several decades to compare the σ-donor capacity of phosphines is the Tolman electronic parameter (TEP). The TEP is a very sensitive measure of a ligand's electron donating or electron withdrawing capacity with only minor influence from steric and dispersion type effects.[22-24] Typically, a Tolman electronic parameter is measured by the A_1 symmetric carbonyl stretching frequency for an L-$Ni^0(CO)_3$ complex, where L is the ligand of interest. Figure 1 displays a 3-dimensional representation of $(CH_3)_3P$-$Ni^0(CO)_3$ with a depiction of the A_1 symmetric carbonyl stretch. If the L-type ligand is electron donating to the Ni metal center then there will be a propagated effect for donation to the CO ligands through π-backbonding, which results in weakening the CO bond strength and decreasing the TEP value. Conversely, if the L-type ligand is electron withdrawing to the Ni metal center then there will be less donation to the CO ligands through π-backbonding resulting in a larger TEP value.

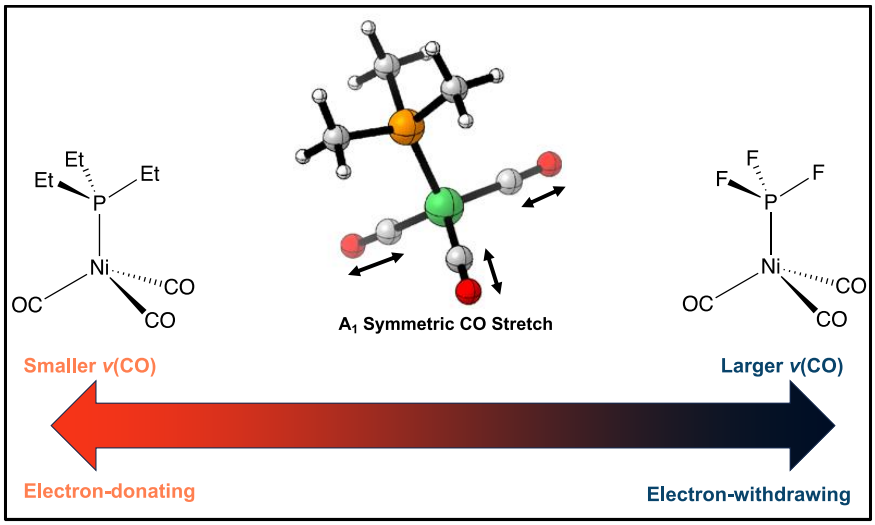

Figure 1. Representation of the A_1 symmetric CO stretch for the $(R)_3P\text{-}Ni^0(CO)_3$ system, which is the basis for TEP values.

While several TEP values have been experimentally determined, it would be ideal if calculations could be used to rapidly predict accurate values for new phosphines since this information could be used for interpreting experiment results or as a basis for selecting a phosphine in the design of a new metal complex, especially in the spirit of Fey's approach.[13-17] Density functional theory (DFT) calculations have been used to calculate TEP values.[25-27] As one representative example, Clot used B3PW91 DFT calculations to calculate 68 CO frequency values for $L\text{-}Ni(CO)_3$ complexes that included both phosphine and non-phosphine ligands.[28] Importantly, this work established the ability to directly correlate and correct DFT calculated values with experiment. This work also demonstrated the transferability of calculated TEP values to other types of complexes,[20] such as $CpMn(CO)_2L$.[22] Crabtree developed a bisphosphine scale that is related to TEP values.[29] TEP values have also been interpreted using natural orbitals for chemical valence calculations,[30] which allowed decomposition of donor and acceptor effects into specific types of orbital interactions. There have also been several studies showing fundamental property correlations with TEP values. For example, Koga used electrostatic potential values correlate with TEP values,[31] and this indicates that relatively simple descriptors can provide understanding of donor and acceptor effects of phosphines. Similar to phosphine TEP values, there are several instances where CO stretching calculations were used

to determine donor and acceptor effects for non-phosphine ligands. For example, Liu and Ke use DFT calculations to determine the steric and electronic parameters of N-heterocyclic boryl type ligands[32] and Kuzu examined carbodiphosphorane ligands.[33] It is useful to note that the use of TEP values has been slightly criticized by Cremer because they do not correlate with bond strength.[34, 35] However, regardless of this lack of correlation, TEP values allow comparison of phosphine electronic effects that can be potentially translated from the L-Ni$(CO)_3$ complex to new complexes, and they can be used as one of many chemical descriptors in catalyst design.

Therefore, our major goal was to develop a machine learning model from descriptors derived from simplified molecular-input line-entry system (SMILES) representations of phosphines that can then be used to instantaneously (at least compared to DFT) calculate accurate TEP values. Figure 2a describes the approach of how this machine learning model, called TEPid, was developed. Using a combination of automated building and calculation tools we curated >4,000 DFT calculated $(R)_3P\text{-}Ni^0(CO)_3$ TEP values. We then extracted dozens of connectivity-based descriptors associated with SMILES strings from the $(R)_3P\text{-}Ni^0(CO)_3$ complexes. We then optimized regressor-type machine learning models and pruned the descriptors to the 19 key model features. This TEPid machine learning model is extremely fast and provides an error of less than 1 cm^{-1}. Additionally, we built a web-based interface to run the machine learning model where SMILES strings can be entered and TEP values returned. We then applied this TEPid model to examine the donor and acceptor capability of phosphines in the large Kraken phosphine database. This revealed that the Kraken database is somewhat skewed towards donor phosphines. In the same spirit of Kraken, we generated new phosphines that when used with Kraken provide a ligand library that is electronically more balanced.

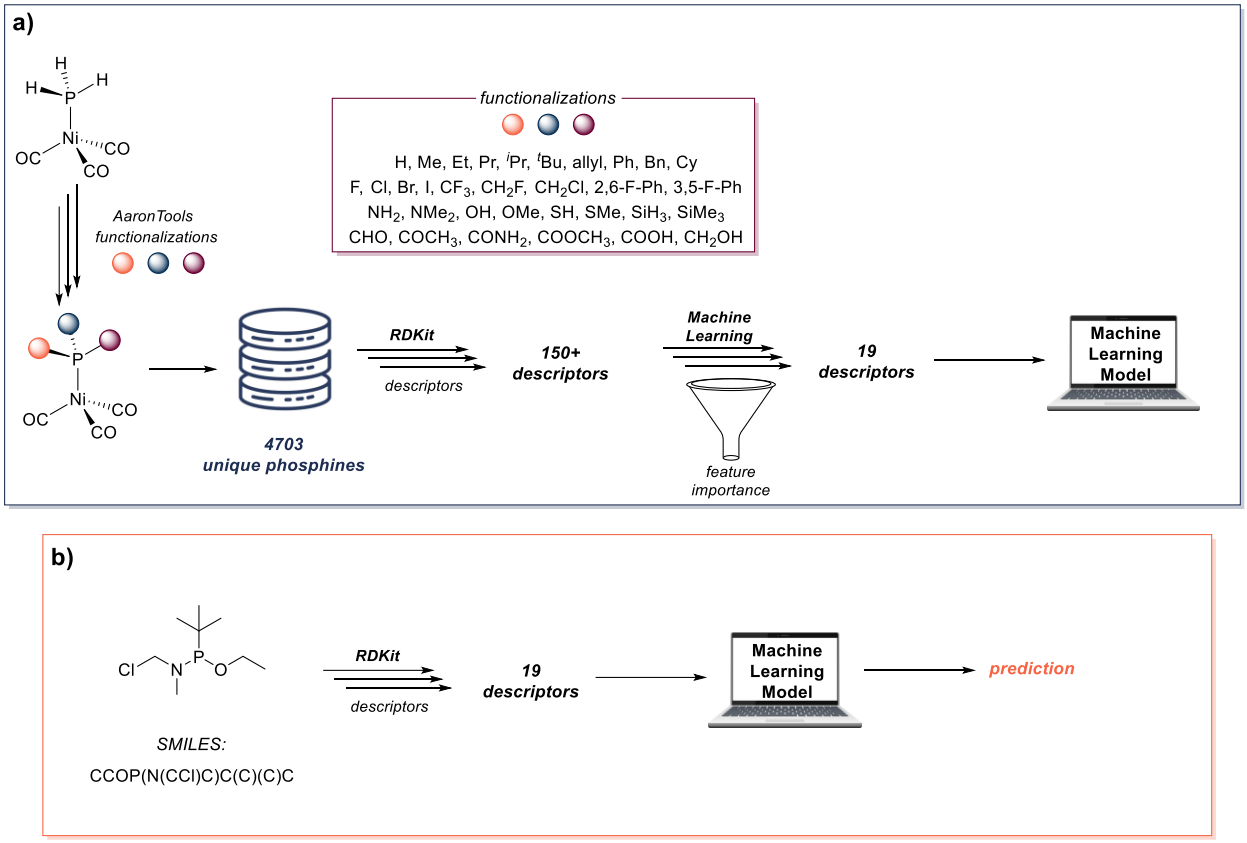

Figure 2. a) The general flowchart for the curation of a database consisting of >4000 calculated TEP values, extraction of connectivity-based chemical descriptors, and creation of a regressor machine learning model. b) Outline of how the machine learning model can be used to predict a TEP value for a new phosphine.

Computational and Data Science Methods

All density functional theory (DFT) calculations were performed with the Gaussian 16 software using the B3LYP density functional[36] with the def2-SVP basis set.[37, 38] Dispersion corrections were applied using Grimme's D3 dispersion method with the Becke-Johnson damping function.[39, 40] Figure 2a shows that a core comprised of a $Ni(CO)_3$ and PH_3 phosphine system was first optimized and then AaronTools was used to functionalize the phosphine with the following groups: Me, Et, Pr, ^{i}Pr, ^{t}Bu, allyl, Ph, Bn, Cy, F, Cl, Br, I, CF_3, CH_2F, CH_2Cl, 2,6-F-Ph, 3,5-F-Ph, NH_2, NMe_2, OH, OMe, SH, SMe, SiH_3, $SiMe_3$, CHO, $COCH_3$, $CONH_2$, $COOCH_3$, COOH, CH_2OH. All 4703 unique $(R)_3P\text{-}Ni^0(CO)_3$ structures were then optimized in the gas-phase with B3LYP-D3(BJ)/def2-SVP. Using the SMILES string representations for each of the

phosphines, more than 100 cheminformatic features were calculated using the RDKit software package.[41] Permutation feature importance and SHAP feature values were used to condense the model down to 19 descriptors.

The LightGBM Regressor was used for the machine learning model with hyperparameters optimized using the Optuna algorithm.[42, 43] Optuna performs automated searches for hyperparameters based on a chosen function to either minimize or maximize. In the case of this application, the 5-fold cross-validation average root-mean-square error across the training set was chosen to be minimized. The 4073 phosphines were split into a training and testing sets with a 70:30 split with 3292 training structures and 1411 testing structures.

To generate a Kraken-style phosphine ligand library, we used a combination of our program Polyjuice and OpenBabel and our ReaLigands ligand library.[44] ReaLigands was curated by detaching and classifying all ligands from all mononuclear transition metal complexes in the Cambridge Structure Database (CSD). For this work, we used 1,078 unique monodentate phosphine ligands. On these 1,078 phosphine ligands we used our Polyjuice program, which is a modified depth-first searching algorithm, to detach all organic groups connected to each phosphine. To assemble the large phosphine ligand library, we used the same approach as Gensch et al. and built phosphines with two identical organic groups and one different organic group (i.e., R'R_2P). To electronically balance this ligand set, we added select phosphines with three different organic groups as well as phosphines that have two covalent connections to a single organic group and a third independent organic group. OpenBabel was used for all structure conversions.

Results and Discussion

To evaluate the performance of B3LYP-D3(BJ)/def2-SVP to calculate TEP values, we used 14 different experimental values that provide a range of donating and accepting phosphines. Table 1 gives the experimental symmetric CO stretch values and the DFT calculated values. This comparison shows that the

DFT calculated values, as expected based on the work of Clot, have a systematic error. Therefore, a simple correction factor was employed to shift the DFT values to align with the experimental values. The correction for the DFT values was determined through creating a linear regression between the uncorrected DFT TEP values and the experimental values for these seven cases. This regression had a high degree of correlation with a $R^2 = 0.960$.

Table 1. B3LYP-D3(BJ)/def2-SVP calculated TEP values with and without the correction factor. [a]Experimental values obtained from reference 24.

Phosphine	Uncorrected CO A_1 Stretch (cm^{-1})	Corrected CO A_1 Stretch (cm^{-1})	Experimental CO A_1 Stretch (cm^{-1})[a]	Error (cm^{-1})
PH_3	2185.6	2084.5	2084.1	+0.4
PF_3	2210.0	2110.8	2110.8	-2.8
PCl_3	2202.5	2100.8	2097.0	+3.8
PMe_3	2167.6	2067.2	2064.1	+3.1
PPh_3	2185.6	2084.5	2083.2	+1.3
$P(NMe_2)_3$	2165.0	2064.7	2061.9	+2.8
PCl_2Ph	2189.9	2088.7	2092.1	-3.4
$P(OMe)_3$	2180.8	2079.9	2079.5	+0.4
$PClPh_2$	2179.2	2078.4	2080.7	-2.3
PEt_3	2210.0	2062.9	2061.7	+1.2
PH_2Ph	2179.8	2079.0	2077.0	+2.0
PMe_2CF_3	2180.9	2080.0	2080.9	-0.9
$PPh_2(CH{=}CH_2)$	2169.4	2068.9	2069.3	-0.4
$PPhBn_2$	2161.4	2061.3	2067.6	-6.3

The corrected value for the CO A_1 symmetric stretch frequency can be determined using Equation 1 where v_{corr} is the corrected value of the Tolman electronic parameter and v_{uncorr} is the uncorrected value. The corrected DFT values have an RMSE of 2.7 cm^{-1} from the experimental values.

$$v_{corr} = 0.9623 * v_{uncorr} - 18.671\ cm^{-1} \qquad \text{Eq. (1)}$$

Using the DFT optimized structures and vibrational frequencies, Equation 1 was used to calculate the TEP values for the 4703 phosphines. Figure 3 displays the distribution of DFT calculated TEP values. The distribution is relatively uniform with a mean of 2078.1 cm^{-1} and a standard deviation of 9.9 cm^{-1}, showing a good spread of the overall dataset.

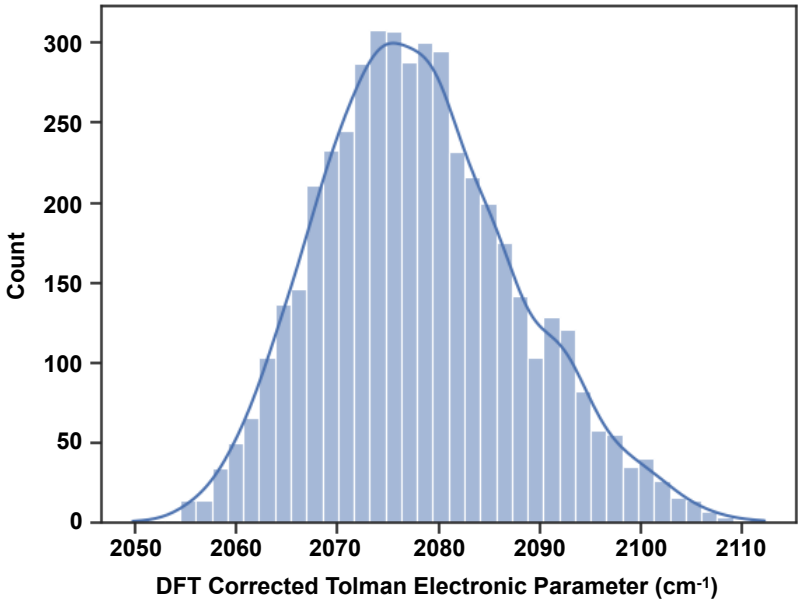

Figure 3. Distribution of B3LYP calculated and corrected TEP values across 4703 phosphines (in cm^{-1}).

We extracted more than 100 cheminformatic features for each phosphine using the RDKit[41] software package and used these feature values to build a regressor machine learning model. Importantly, we systematically decreased the number of model features to only 19 cheminformatic descriptors while keeping the same model performance as including all descriptors. SHAP and permutation feature importances were used to determine the top 20 most important features to the model, and then each

combination of the 20 were employed until it was determined that only 19 were necessary for best model performance. The hyperparameters for the LightGBM model used for predictions can be found in the Supplementary Information (SI). The 5-fold cross-validation on the training set provided average training and validation RMSE values of 0.989 ± 0.007 cm^{-1} and 2.041 ± 0.002 cm^{-1}, respectively. Therefore, on average, the model has an error of approximately 2 cm^{-1} when predicting structures outside of the training set. Figure 4 displays the correlation between the DFT calculated TEP values and the LightGBM model values. The predictions for both the training set and the testing set show good agreement with the true values, having R^2 values of 0.992 and 0.960, respectively.

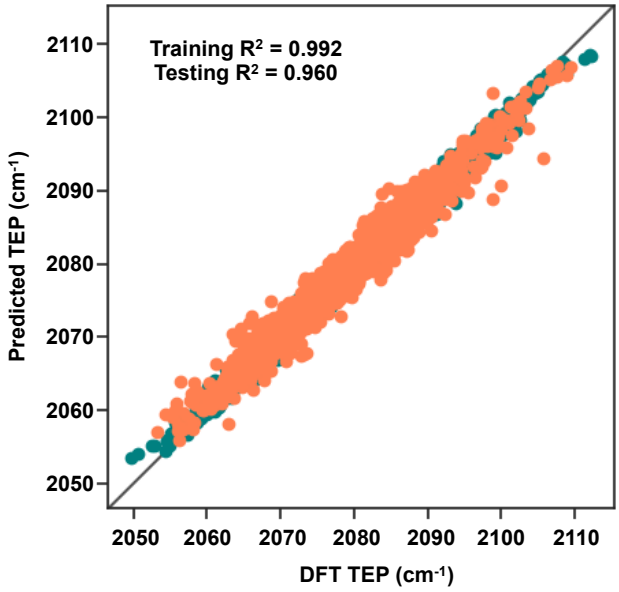

Figure 4. Plot of the machine learning predicted TEP values versus the M06/def2-SVP calculated values. The data points for the training set are plotted in teal while the data points for the testing set are plotted in orange.

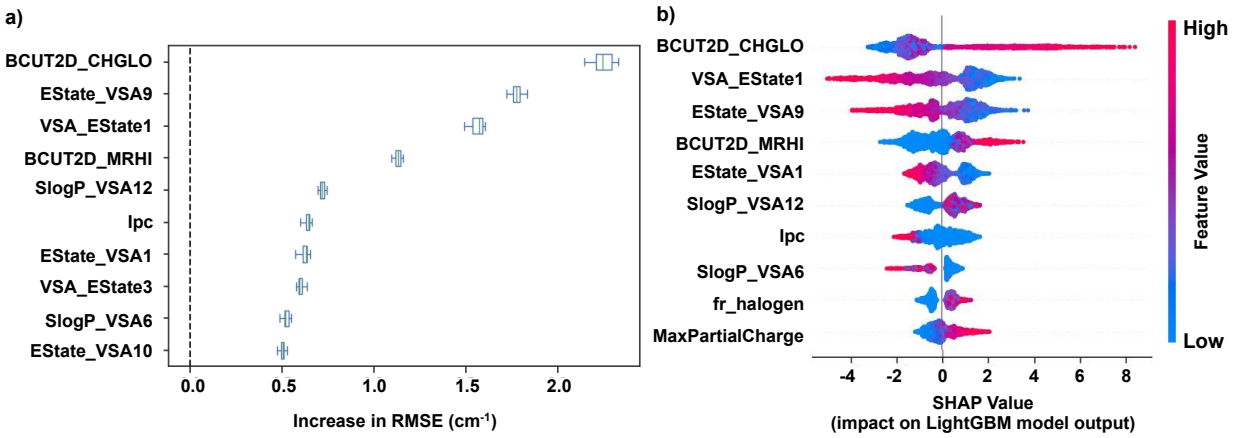

Figure 5. Plots of feature importances from a) a permutation feature method and b) the SHapley Additive exPlanations (SHAP) method for the training data.

Figure 5 displays two different methods for calculating the importance of features for the LightGBM machine learning model. Figure 5a shows permutation feature importance wherein each individual descriptor has its values shuffled among the different data points to test how much the error is introduced by the shuffling. This process was performed 20 times for each feature and is displayed as a box-and-whisker with regards to the increase in the RMSE. Figure 5b shows the feature importance through the SHapley Additive exPlanations (SHAP) method which uses a game theoretic approach to explain each descriptor's effect on the model output for each individual datapoint.[45] A negative SHAP value corresponds to a descriptor lowering the model's predicted output value for that particular data point and vice versa for a positive SHAP value. The color gradient (red versus blue) describes the value of that feature, that is, a pure red data point equates to a larger more positive value for that descriptor associated with a data point. Taking the VSA_EState1 feature as an example, the red data points mainly have negative SHAP values while the blue data points mainly have positive SHAP values, which means that having a larger more positive value for the VSA_EState1 feature will result in a lower predicted value for the TEP value.

There are eight features out of ten that overlap from each importance methodology. The top feature in both sets is the BCUT2D_CHGLO. The BCUT2D descriptors represent eigenvalues of a connectivity

matrix with a specific property along the diagonal with the off-diagonal elements depending on the bond order.[46, 47] The CHGLO tag refers to the lowest eigenvalue of the connectivity matrix with the diagonal values being the Gasteiger charges showing the electronic a direct correlation between overall charge distribution across the molecule with its TEP value, and this is an expected key feature for electronic effects. Among the top 10 features are electrotopological state (EState) index features which represent the combination of the electronic character and topological environment of each atom in a molecule.[48-50] Most of the top ten features represent graph-based descriptors that focus on the electron distribution and polarity of molecules. Perhaps as expected, but nonetheless interesting, common descriptors that provide evaluation of size and steric of atoms and molecules were not found to be important, which is consistent with TEP values describing electronic effects without significant steric effects.

Figure 6 shows the learning curve for the training set using the LightGBM model. The dataset is further split into a set used to train the model and a separate validation set to test the model upon. At each training set size, 5-fold cross-validation was used to generate an average RMSE value and standard deviation across all five folds. At around 500 structures the model begins to learn the TEP values consistently with a 5-fold RMSE of approximately 4 cm^{-1} and a relatively small standard deviation. As more structures are added there is a slow but persistent decrease in the validation RMSE.

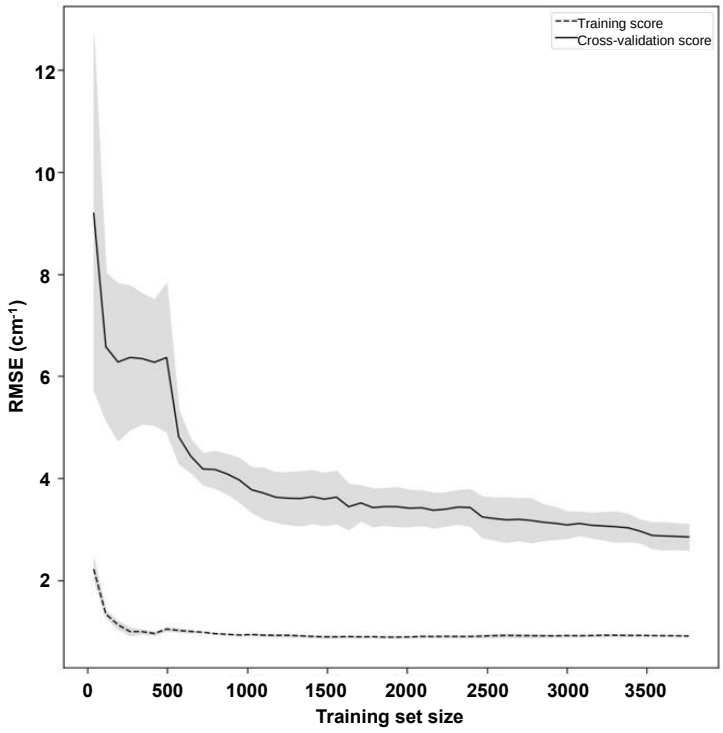

Figure 6. Learning curve of the LightGBM machine learning model's 5-fold CV averaged root-mean-square error (in cm^{-1}) versus the number of systems in the training set. The light grey shading represents the standard deviation of the 5-fold CV RMSE values (in cm^{-1}).

With the development of the TEPid machine learning model, we were then able to analyze the Kraken phosphine ligand library to determine the distribution of electronic effects. We then calculated the TEP values for 294,860 phosphine structures retrieved from the Kraken database's application programming interface (API). A few of the phosphines could not be calculated with TEPid because they had null values for some of the 19 descriptors. Figure 7 plots the distribution of Kraken phosphine TEP values. Comparison of Figure 7 with Figure 3 shows that the Kraken phosphines have a significant shift to lower TEP values than our DFT calculated set of phosphines with a mean of 2066.8 cm^{-1} and a standard deviation of 5.7 cm^{-1} compared to the DFT calculated set of phosphines with a mean of 2078.1 cm^{-1} and a standard deviation of 9.9 cm^{-1}. This suggests that the structures in the Kraken database tend to more electron-donating phosphine complexes that transfer more electron density to the nickel.

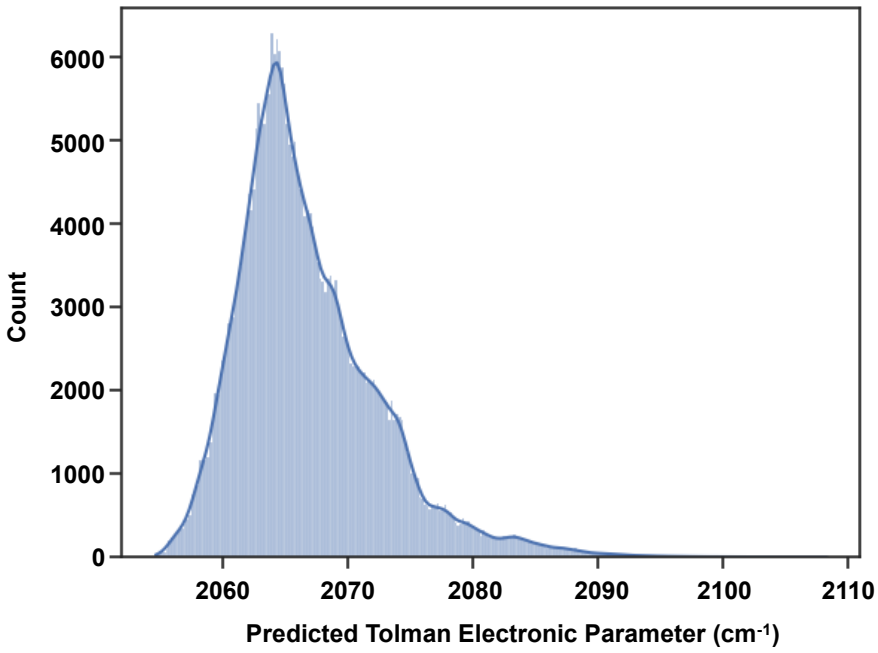

Figure 7. Distribution of the machine learning calculated TEP values (in cm^{-1}) for 294,860 phosphine structures from the Kraken database.

In an effort to have access to more electron-withdrawing phosphines and an electronically balanced library, we decided to build our own phosphine ligand library beginning with our ReaLigands[44] library that contains all ligands detached and classified from all transition metal complexes in the CSD. In our ReaLigands library there are 1,078 phosphines. The machine learning calculated TEP values for these experimental phosphines ranges from 2053.1 to 2107.7 cm^{-1} with a mean of 2068.1 cm^{-1} (standard deviation of 7.7 cm^{-1}). We dismantled these 1,078 phosphines using our program Polyjuice[51] to obtain a set of 205 organic groups that were then used to assemble new phosphines. Like Kraken, we first generated phosphines with two identical organic groups and one different organic group (labeled as RR'$_2$P). This generated 42,025 unique phosphines with a mean calculated TEP value of 2067.8 cm^{-1} with a standard deviation of 7.3 cm^{-1}, which is a very similar mean and distribution to the Kraken phosphine library. To increase the number of electron-withdrawing phosphines we identified the 76 organic groups that provided TEP values of 2070 cm^{-1} or greater and then used these groups to generate 76,076 new phosphines where all three groups on the phosphine are different (labeled as RR'R''P). This set of phosphines showed the

expected higher mean value of 2076.3 cm^{-1} with a standard deviation of 7.1 cm^{-1}. To this library, we also added phosphines that have two covalent connections to a single organic group and a third independent organic group (labeled as RR'P). This was done by using Polyjuice to identify 201 unique organic groups from the ReaLigand's phosphines that provide two covalent bonds to phosphine. We then combined these 201 organic groups with the previously mentioned 205 single connection organic groups to yield 31,666 more phosphine structures. Figure 8a plots the TEP values for all the created phosphine ligands. The complete set of phosphines is available for download from GitHub. Figure 8b plots the TEP values for the categories for phosphines, such as RR'_2P, RR'R''P, and RR'P.

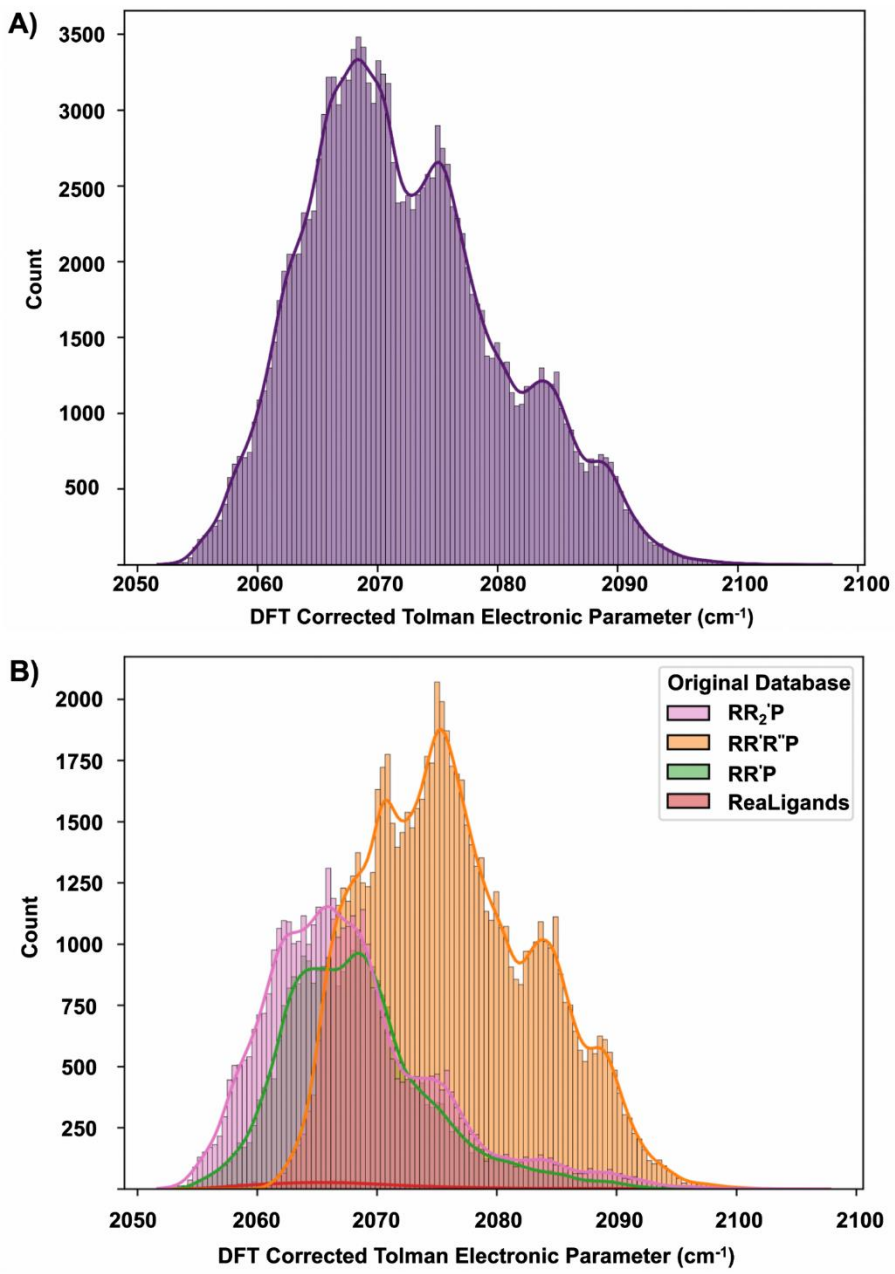

Figure 8. a) Distribution of machine learning calculated TEP values (cm^{-1}) for all phosphines created by using the organic groups detached from phosphines in the ReaLigand library. b) Color-coded plot of machine learning calculated TEP values (cm^{-1}) for ReaLigand phosphines (red) and newly created phosphines with the structures RR'_2P (pink), RR'R''P (orange), and RR'P (green).

Conclusion

Because phosphines play a central role in organometallic chemistry, we developed TEPid, which is a machine learning model that provides extremely fast evaluation of TEP values that provide a measure of phosphine electron donating versus electronic withdrawing capability. >4,000 DFT calculated $(R)_3P$-$Ni^0(CO)_3$ TEP values with an experimental correction factor were used as the training data for the machine learning model. With only a few connectivity-based descriptors the TEPid model provides accuracy within 1 cm^{-1} compared to DFT calculated values. We applied TEPid to analyze the Karken phosphine database, and this showed that the database contains more electron donating phosphines than electron withdrawing phosphines. To begin to have an electronically balanced phosphine ligand library we started with the phosphines from our ReaLigands library, disassembled the organic groups from the phosphine center and then reassembled many new combinations. This resulted in a new library that is balanced between electron donating phosphines and electron withdrawing phosphines.

Corresponding Author Information

***Justin K. Kirkland** – *Department of Chemistry and Biochemistry, Brigham Young University, Provo, Utah 84604, United States*; Email: justin.kirkland.phd@gmail.com

***Daniel H. Ess** – *Department of Chemistry and Biochemistry, Brigham Young University, Provo, Utah 84604, United States*; Email: dhe@chem.byu.edu

Acknowledgements

We thank Brigham Young University and the Fulton Supercomputing Lab for computational resources.

Supporting Information

The TEPid code, Polyjuice code, phosphine ligand libraries, DFT xyz coordinates are available from the following GitHub repository: https://github.com/DanielEss-lab/. The web interface for using TEPid is at https://tepid.chem.byu.edu.

Notes

The authors declare no conflict of interest.

References

1. Jiao, Y.; Torne, M. S.; Gracia, J.; Niemantsverdriet, J. W.; van Leeuwen, P. W. N. M., Ligand effects in rhodium-catalyzed hydroformylation with bisphosphines: steric or electronic? *Catalysis Science & Technology* **2017,** *7* (6), 1404-1414.

2. Franke, R.; Selent, D.; Börner, A., Applied Hydroformylation. *Chemical Reviews* **2012,** *112* (11), 5675-5732.

3. Schmitz, C.; Holthusen, K.; Leitner, W.; Franciò, G., Highly Regio- and Enantioselective Hydroformylation of Vinyl Esters Using Bidentate Phosphine,P-Chiral Phosphorodiamidite Ligands. *ACS Catalysis* **2016,** *6* (3), 1584-1589.

4. Lundgren, R. J.; Stradiotto, M., Addressing Challenges in Palladium-Catalyzed Cross-Coupling Reactions Through Ligand Design. *Chemistry – A European Journal* **2012,** *18* (32), 9758-9769.

5. Martin, R.; Buchwald, S. L., Palladium-Catalyzed Suzuki−Miyaura Cross-Coupling Reactions Employing Dialkylbiaryl Phosphine Ligands. *Accounts of Chemical Research* **2008,** *41* (11), 1461-1473.

6. Ruiz-Castillo, P.; Buchwald, S. L., Applications of Palladium-Catalyzed C–N Cross-Coupling Reactions. *Chemical Reviews* **2016,** *116* (19), 12564-12649.

7. Burello, E.; Farrusseng, D.; Rothenberg, G., Combinatorial Explosion in Homogeneous Catalysis: Screening 60,000 Cross-Coupling Reactions. *Adv Synth Catal* **2004,** *346* (13-15), 1844-1853.

8. Fleckenstein, C. A.; Plenio, H., Sterically demanding trialkylphosphines for palladium-catalyzed cross coupling reactions—alternatives to PtBu3. *Chemical Society Reviews* **2010,** *39* (2), 694-711.

9. Bariwal, J.; Van der Eycken, E., C–N bond forming cross-coupling reactions: an overview. *Chemical Society Reviews* **2013,** *42* (24), 9283-9303.

10. Erre, G.; Enthaler, S.; Junge, K.; Gladiali, S.; Beller, M., Synthesis and application of chiral monodentate phosphines in asymmetric hydrogenation. *Coordination Chemistry Reviews* **2008,** *252* (5), 471-491.

11. Ager, D. J.; de Vries, A. H. M.; de Vries, J. G., Asymmetric homogeneous hydrogenations at scale. *Chemical Society Reviews* **2012,** *41* (8), 3340-3380.

12. Gensch, T.; dos Passos Gomes, G.; Friederich, P.; Peters, E.; Gaudin, T.; Pollice, R.; Jorner, K.; Nigam, A.; Lindner-D'Addario, M.; Sigman, M. S.; Aspuru-Guzik, A., A Comprehensive Discovery Platform for Organophosphorus Ligands for Catalysis. *J Am Chem Soc* **2022,** *144* (3), 1205-1217.

13. Durand, D. J.; Fey, N., Computational Ligand Descriptors for Catalyst Design. *Chemical Reviews* **2019,** *119* (11), 6561-6594.

14. Jover, J.; Fey, N., The Computational Road to Better Catalysts. *Chemistry – An Asian Journal* **2014,** *9* (7), 1714-1723.

15. Fey, N.; Orpen, A. G.; Harvey, J. N., Building ligand knowledge bases for organometallic chemistry: Computational description of phosphorus(III)-donor ligands and the metal–phosphorus bond. *Coordination Chemistry Reviews* **2009,** *253* (5), 704-722.

16. Fey, N., Lost in chemical space? Maps to support organometallic catalysis. *Chemistry Central Journal* **2015,** *9* (1), 38.

17. Fey, N., The contribution of computational studies to organometallic catalysis: descriptors, mechanisms and models. *Dalton Transactions* **2010,** *39* (2), 296-310.

18. Sigman, M. S.; Harper, K. C.; Bess, E. N.; Milo, A., The Development of Multidimensional Analysis Tools for Asymmetric Catalysis and Beyond. *Accounts of Chemical Research* **2016,** *49* (6), 1292-1301.

19. Santiago, C. B.; Guo, J.-Y.; Sigman, M. S., Predictive and mechanistic multivariate linear regression models for reaction development. *Chemical Science* **2018,** *9* (9), 2398-2412.

20. Kühl, O., Predicting the net donating ability of phosphines—do we need sophisticated theoretical methods? *Coordination Chemistry Reviews* **2005,** *249* (5), 693-704.

21. Murray, P. M.; Tyler, S. N. G.; Moseley, J. D., Beyond the Numbers: Charting Chemical Reaction Space. *Organic Process Research & Development* **2013,** *17* (1), 40-46.

22. Strohmeier, W.; Müller, F.-J., Klassifizierung phosphorhaltiger Liganden in Metallcarbonyl-Derivaten nach der π-Acceptorstärke. *Chemische Berichte* **1967,** *100* (9), 2812-2821.

23. Tolman, C. A.; Seidel, W. C.; Gosser, L. W., Formation of three-coordinate nickel(0) complexes by phosphorus ligand dissociation from NiL4. *J Am Chem Soc* **1974,** *96* (1), 53-60.

24. Tolman, C. A., Steric effects of phosphorus ligands in organometallic chemistry and homogeneous catalysis. *Chemical Reviews* **1977,** *77* (3), 313-348.

25. Ciancaleoni, G.; Scafuri, N.; Bistoni, G.; Macchioni, A.; Tarantelli, F.; Zuccaccia, D.; Belpassi, L., When the Tolman Electronic Parameter Fails: A Comparative DFT and Charge Displacement Study of [(L)Ni(CO)3]0/– and [(L)Au(CO)]0/+. *Inorganic Chemistry* **2014,** *53* (18), 9907-9916.

26. Kégl, T. R.; Pálinkás, N.; Kollár, L.; Kégl, T., Computational Characterization of Bidentate P-Donor Ligands: Direct Comparison to Tolman's Electronic Parameters. *Molecules* **2018,** *23* (12), 3176.

27. Shanahan, J. P.; Szymczak, N. K., Lewis Acid Effects on Calculated Ligand Electronic Parameters. *Organometallics* **2020,** *39* (23), 4297-4306.

28. Perrin, L.; Clot, E.; Eisenstein, O.; Loch, J.; Crabtree, R. H., Computed Ligand Electronic Parameters from Quantum Chemistry and Their Relation to Tolman Parameters, Lever Parameters, and Hammett Constants. *Inorganic Chemistry* **2001,** *40* (23), 5806-5811.

29. Anton, D. R.; Crabtree, R. H., Dibenzo[a,e]cyclooctatetraene in a proposed test for heterogeneity in catalysts formed from soluble platinum-group metal complexes. *Organometallics* **1983,** *2* (7), 855-859.

30. Ardizzoia, G. A.; Brenna, S., Interpretation of Tolman electronic parameters in the light of natural orbitals for chemical valence. *Physical Chemistry Chemical Physics* **2017,** *19* (8), 5971-5978.

31. Suresh, C. H.; Koga, N., Quantifying the Electronic Effect of Substituted Phosphine Ligands via Molecular Electrostatic Potential. *Inorganic Chemistry* **2002,** *41* (6), 1573-1578.

32. Jing, Y.; Jiang, J.; Liu, Y.; Ke, Z., Electronic and Steric Properties of N-Heterocyclic Boryl Ligands. *Organometallics* **2022,** *41* (5), 627-633.

33. Flosdorf, K.; Jiang, D.; Zhao, L.; Neumüller, B.; Frenking, G.; Kuzu, I., An Experimental and Theoretical Study of the Structures and Properties of [CDPMe-Ni(CO)3] and [Ni2(CO)4(μ2-CO)(μ2-CDPMe)]. *European Journal of Inorganic Chemistry* **2019,** *2019* (42), 4546-4554.

34. Setiawan, D.; Kalescky, R.; Kraka, E.; Cremer, D., Direct Measure of Metal–Ligand Bonding Replacing the Tolman Electronic Parameter. *Inorganic Chemistry* **2016,** *55* (5), 2332-2344.

35. Kalescky, R.; Kraka, E.; Cremer, D., New Approach to Tolman's Electronic Parameter Based on Local Vibrational Modes. *Inorganic Chemistry* **2014,** *53* (1), 478-495.

36. Becke, A. D., Density-functional thermochemistry. III. The role of exact exchange. *The Journal of Chemical Physics* **1993,** *98* (7), 5648-5652.

37. Weigend, F., Accurate Coulomb-fitting basis sets for H to Rn. *Physical Chemistry Chemical Physics* **2006,** *8* (9), 1057-1065.

38. Weigend, F.; Ahlrichs, R., Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Physical Chemistry Chemical Physics* **2005,** *7* (18), 3297-3305.

39. Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H., A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *The Journal of Chemical Physics* **2010,** *132* (15).

40. Grimme, S.; Ehrlich, S.; Goerigk, L., Effect of the damping function in dispersion corrected density functional theory. *Journal of Computational Chemistry* **2011,** *32* (7), 1456-1465.

41. RDKit: Open-source cheminformatics.

42. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M., Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International*

Conference on Knowledge Discovery & Data Mining, Association for Computing Machinery: Anchorage, AK, USA, 2019; pp 2623–2631.

43. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y., LightGBM: a highly efficient gradient boosting decision tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Curran Associates Inc.: Long Beach, California, USA, 2017; pp 3149–3157.

44. Chen, S-S.; Meyer, Z.; Jensen, B.; Kraus, A.; Lambert, A.; Ess, D. H. Ess, ReaLigands: A Ligand Library Cultivated from Experiment and Intended for Molecular Computational Catalyst Design. *ChemRxiv* **2023** https://chemrxiv.org/engage/chemrxiv/article-details/64dcf7454a3f7d0c0d45ca3c.

45. Lundberg, S. M.; Lee, S.-I., A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Curran Associates Inc.: Long Beach, California, USA, 2017; pp 4768–4777.

46. Bayada, D. M.; Hamersma, H.; van Geerestein, V. J., Molecular Diversity and Representativity in Chemical Databases. *Journal of Chemical Information and Computer Sciences* **1999,** *39* (1), 1-10.

47. Burden, F. R., Molecular identification number for substructure searches. *Journal of Chemical Information and Computer Sciences* **1989,** *29* (3), 225-227.

48. Hall, L. H.; Mohney, B.; Kier, L. B., The Electrotopological State: An Atom Index for QSAR. *Quantitative Structure-Activity Relationships* **1991,** *10* (1), 43-51.

49. Hall, L. H.; Kier, L. B., Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. *Journal of Chemical Information and Computer Sciences* **1995,** *35* (6), 1039-1045.

50. Hall, L. H.; Kier, L. B., The E-State as the Basis for Molecular Structure Space Definition and Structure Similarity. *Journal of Chemical Information and Computer Sciences* **2000,** *40* (3), 784-791.

51. https://github.com/DanielEss-lab/PolyjuicePotion

TOC graphic