ACCOUNTS of chemical research

pubs.acs.org/accounts Article

Importance of Engineered and Learned Molecular Representations in Predicting Organic Reactivity, Selectivity, and Chemical Properties

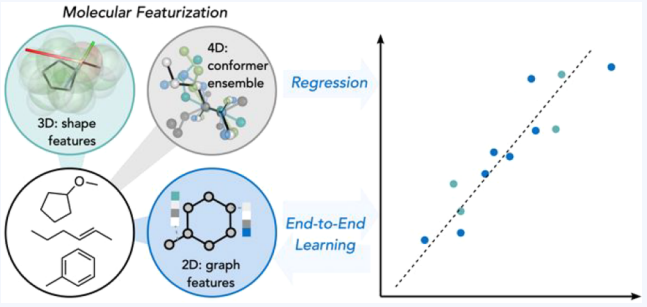
Published as part of the Accounts of Chemical Research special issue "Data Science Meets Chemistry".

Liliana C. Gallegos,[§] Guilian Luchini,[§] Peter C. St. John, Seonah Kim, and Robert S. Paton*

Cite This: *Acc. Chem. Res.* 2021, 54, 827–836 Read Online

ACCESS | Metrics & More | Article Recommendations

CONSPECTUS: Machine-readable chemical structure representations are foundational in all attempts to harness machine learning for the prediction of reactivities, selectivities, and chemical properties directly from molecular structure. The featurization of discrete chemical structures into a continuous vector space is a critical phase undertaken before model selection, and the development of new ways to quantitatively encode molecules is an active area of research. In this Account, we highlight the application and suitability of different representations, from expert-guided "engineered" descriptors to automatically "learned" features, in different prediction tasks relevant to organic and organometallic chemistry, where differing amounts of training data are available. These tasks include statistical models of stereo- and enantioselectivity, thermochemistry, and kinetics developed using experimental and quantum chemical data.



The use of expert-guided molecular descriptors provides an opportunity to incorporate chemical knowledge, domain expertise, and physical constraints into statistical modeling. In applications to stereoselective organic and organometallic catalysis, where data sets may be relatively small and 3D-geometries and conformations play an important role, mechanistically informed features can be used successfully to obtain predictive statistical models that are also chemically interpretable. We provide an overview of several recent applications of this approach to obtain quantitative models for reactivity and selectivity, where topological descriptors, quantum mechanical calculations of electronic and steric properties, along with conformational ensembles, all feature as essential ingredients of the molecular representations used.

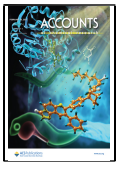
Alternatively, more flexible, general-purpose molecular representations such as attributed molecular graphs can be used with machine learning approaches to learn the complex relationship between a structure and prediction target. This approach has the potential to out-perform more traditional representation methods such as "hand-crafted" molecular descriptors, particularly as data set sizes grow. One area where this is particularly relevant is in the use of large sets of quantum mechanical data to train quantitative structure–property relationships. A general approach toward curating useful data sets and training highly accurate graph neural network models is discussed in the context of organic bond dissociation enthalpies, where this strategy outperforms regression using precomputed descriptors.

Finally, we describe how graph neural network predictions can be incorporated into mechanistically informed statistical models of chemical reactivity and selectivity. Once trained, this approach avoids the expensive computational overhead associated with quantum mechanical calculations, while maintaining chemical interpretability. We illustrate examples for which fast predictions of bond dissociation enthalpy and of the identities of radicals formed through cleavage of a molecule's weakest bond are used in simple physical models of site-selectivity and reactivity.

KEY REFERENCES

- Piou, T.; Romanov-Michailidis, F.; Romanova-Michaelides, M.; Jackson, K. E.; Semakul, N.; Taggart, T. D.; Newell, B. S.; Rithner, C. D.; Paton, R. S.; Rovis, T. Correlating Reactivity and Selectivity to Cyclopentadienyl Ligand Properties in Rh(III)-Catalyzed C-H Activation Reactions: An Experimental and Computa-

Received: November 9, 2020
Published: February 3, 2021



https://dx.doi.org/10.1021/acs.accounts.0c00745
Acc. Chem. Res. 2021, 54, 827–836

tional Study. *J. Am. Chem. Soc.* **2017**, *139*, 1296−1310.[1] *Experimental and computational catalyst descriptors are developed for cyclopentadienyl complexes and applied to predict reactivity, regioselectivity and diastereoselectivity in different catalytic reactions.*

- Brethomé, A. V.; Fletcher, S. P.; Paton, R. S. Conformational Effects on Physical-Organic Descriptors - the Case of Sterimol Steric Parameters. *ACS Catal.* **2019**, *9*, 2313−2323.[2] *An automated workflow software is developed to obtain 4D (conformer ensemble) steric parameters and applied to illustrate uncertainty in regression models.*
- St. John, P.; Guan, Y.; Kim, Y.; Kim, S.; Paton, R. S. Prediction of homolytic bond dissociation enthalpies for organic molecules at near chemical accuracy with sub-second computational cost. *Nat. Commun.* **2020**, *11*, 2328.[3] *A graph neural network approach for organic property prediction is developed, resulting in highly accurate BDE predictions that are also used as descriptors in mechanistic models for selectivity and reactivity.*

1. INTRODUCTION

Data-driven chemistry is propelled by innovations in the generation and curation of chemical data, the machine learning algorithms used for regression and classification, and how molecules are represented.[4] Machine-readable chemical structure representations were originally introduced to create the first searchable computational databases of molecules and reactions in the 1960s.[5] They are now a central element of chemical machine learning (ML). For decades, the development of predictive quantitative structure−activity and structure−property relationships (QSAR and QSPR) directly from chemical structure has been an area of active research in which the construction of expressive molecular *feature representations* that inform the physical nature of the input−output mapping is a central task.[6] *Feature vectors* encode information about molecular structure, in most cases, by combining a series of physically meaningful molecular descriptors that describe spatial, electronic, and energetic properties. The use of "expert crafted" descriptors provides an opportunity to incorporate chemical knowledge, domain expertise, and physical constraints into any given machine-learning approach while also potentially offering greater interpretability to the chemist as a result.[7]

The *featurization* or *embedding* of discrete molecular structures into a continuous vector space (i.e., as feature vectors) is a critical phase undertaken before model selection. Attempts to predict specific reaction outcomes such as reactivity or selectivity are routinely faced with small data sets on the order of tens to hundreds of examples. In these cases, manual approaches to feature engineering that rely upon specialized domain knowledge, such as a structural or mechanistic hypothesis, and physicochemical descriptors tend to achieve better results than more generalizable representations. Feature vectors derived from physical-organic parameters that describe a molecule's or substituent's electronic (e.g., HOMO/LUMO energies, atomic charges, Fukui[8] coefficients) and steric (e.g., Tolman cone angle,[9] Sterimol,[10] buried volume[11]) influence have been used in ML models to predict the yields[12] and diastereo-[1] and enantioselectivities[13] of organic and organometallic reactions by Sigman,[14] Doyle,[15] and others including ourselves. The continued development of physically motivated descriptors that succinctly and transparently capture the subtleties of molecular stereochemistry, conformation, and electronic effects is central to data-driven approaches for organic reaction prediction, as the ability to link a quantitative predictive model back to interpretable descriptors can be used to derive new understanding and mechanistic inferences.[16−18]

While manually engineered features may focus on describing a specific type of molecule or reaction, more flexible, general-purpose molecular representations such as attributed molecular graphs[19] can be used in combination with ML approaches to learn the complex relationship between a structure and prediction target. Deep learning approaches have proven to be particularly well-suited to the representation of organic structures, automatically learning "rich" features and improving the accuracy of chemical property and reactivity prediction over traditional hand-coded or molecular fingerprint representations.[20] In particular, the rise of graph neural networks (GNNs)[21] in modeling chemical properties has enabled "end-to-end" learning on molecular structure: an ML strategy where traditional feature engineering is replaced by a learned molecular representation derived from an attributed molecular graph. These approaches have led to best-in-class prediction accuracies on a range of applications from total energies, interatomic forces,[22] and bond strengths,[3] especially as the amount of available training data grows.[23,24]

In this Account, we present an overview of how distinct featurization strategies can be applied to organic and organometallic chemistry, to predict reactivity, stereoselectivity, and chemical properties. In the case of small (<100) reaction data sets, hand-crafted physicochemical descriptors are shown to yield interpretable models such as multivariate linear regressions (MLR) of catalytic enantio- and diastereoselectivities. With larger data sets, such as those obtained from high-throughput quantum chemical data sets, learned representations with flexible GNNs can be trained to produce excellent quantitative predictions of atomic or molecular properties at low computational cost. This is illustrated for organic bond dissociation enthalpies. Finally, we describe how GNN predictions can be incorporated into mechanistically informed statistical models of chemical reactivity and selectivity. Once trained, this approach avoids the expensive computational overhead associated with QM calculations and maintains chemical interpretability.

2. MOLECULAR REPRESENTATIONS: FROM ONE TO FOUR DIMENSIONS

While QSAR/QSPR models emerged largely in the context of medicinal chemistry and drug discovery, attempts to relate structure with catalytic activity and selectivity have emerged more recently. Traditional cheminformatics representations largely focus on 2D or *topological* molecular representations that define the connectivity and bonding types of atoms in a molecule. 2D molecular descriptors, such as topological fingerprints, are simple to define and can be obtained without geometry optimization. However, a number of features of mechanistic relevance to reactivity and selectivity depend upon a molecule's 3D structure and conformation, including electronic properties such as atomic and molecular charge distributions, as well as other structure-dependent features that capture a molecule's steric influence, chirality, volume or surface area. Quantum mechanically (QM) optimized molecular coordinates (e.g., using density functional theory,

https://dx.doi.org/10.1021/acs.accounts.0c00745
Acc. Chem. Res. 2021, 54, 827−836

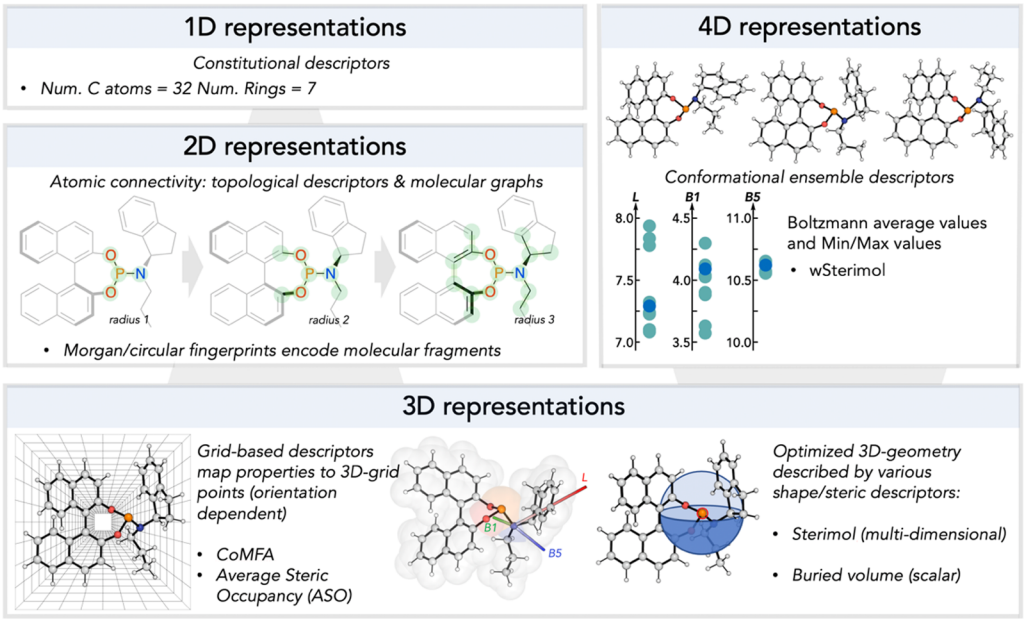


Figure 1. Hierarchy of molecular representations used to encode organic and organometallic structures.

DFT) can be used to obtain such descriptors, in addition to other QM-computed properties such as thermochemical values, molecular orbital energies, vibrational frequencies, and noncovalent interaction energies. In QSAR vernacular, 3D-descriptors generally refer to those that map molecular interactions to a prealigned grid of points, as in Comparative Molecular Field Analysis[25] (COMFA) and the more recently developed Average Steric Occupancy (ASO) from Denmark.[26] More generally, properties dependent upon 3D-structure, and especially spatial/steric occupancies, have been described by scalar parameters such as buried volume and higher dimensional objects such as topological maps or multidimensional Sterimol parameters, approaches pioneered by the Cavallo[11] and Sigman[27] groups, respectively. This hierarchy of molecular representations used across organic chemistry is shown in Figure 1.

The dependence of 3D or DFT-derived descriptors upon the molecular geometry means that unlike 2D representations, conformational dynamics are important to consider. Indeed, the presence of multiple conformations may itself be an important descriptor relating to catalytic proficiency.[28] Thus, in addition to featurization of the most stable conformer, consideration of the full conformational ensemble may be necessary to quantitatively encode macroscopic behavior. Ensemble approaches have been referred to as 4D-QSAR.[29] Spatial parameters may be sensitive to conformational behavior and to the level of theory employed to generate an ensemble, which led us to develop a software tool, wSterimol, to automate conformer-sampling and featurization.[2] With this, we have been able to include estimates of parameter uncertainty into regression models of stereoselectivity.

3. PERFORMANCE IN REGRESSION MODELS OF REACTIVITY AND STEREOSELECTIVITY

Aspuru-Guzik, Balcells et al. have demonstrated the application of topological descriptors to predict DFT-computed activation barriers with high chemical accuracy.[30] Around 2500 QM transition structures (TSs) were obtained for the oxidative addition of dihydrogen to varying IrL_3X complexes with varying ligand types (Figure 2).[31] Ir-complexes were encoded by full autocorrelation (FA) functions (using Kulik's MolSimplify[32]) concatenated with Morgan and RDKit fingerprints, which encode the presence or absence of particular molecular substructures.[33,34] Regression with a Gaussian process (GP) model gave excellent quantitative performance (MAE = 0.6 kcal/mol, R^2 = 0.95) while also identifying important substructural motifs based on feature importance that can be interpreted chemically. The application of topological descriptors has also been applied to predict relative reactivity of alkanes toward C−H abstraction by metal carbene electrophiles.[35]

The 2D-descriptors can be complemented by DFT-derived features that may be absent from the simpler representation, such as atomic partial charges; the enantioselectivity of a key synthetic step used to prepare an approved antiviral, letermovir, was optimized with this featurization approach and MLR by Sigman et al.[36] The key asymmetric aza-Michael addition step (Figure 3) is promoted by chiral triflamide catalysts, in which varying aromatic substituents were studied (initial data set of around 30 structures). One particular 2D-descriptor (FX1sp3CX2sp207) was found to have a high univariate correlation (R^2 of 0.57) with enantioselectivity: this is a substructural parameter between fluorine atoms seven bonds away from an sp^2-hybridized carbon atom, that encodes steric effects. Combining this feature with those obtained from electronic structure calculations, including Natural Population Analysis (NPA) charges at C1 and C2 positions and isotropic polarizability ($polar_i$) of the aryl-ligand, led to a multivariate regression with excellent fit for training (R^2 of 0.90) and validation (R^2 of 0.77) data sets. Mechanistic studies suggest that two H-bonds are formed between Michael-acceptor carbonyl and catalyst N−H groups in the key TS (Figure 3).[37] The statistical model indicates the significance of fluorine-substituents, which lock the catalyst in a stable conformation and form stabilizing interactions with the substrate's aromatic ring. Consequently, the optimal enantioselectivity (90.4 %ee) was obtained by incorporating 2-fluoro,4-trimethylsilyl aryl groups into the catalyst, structures

https://dx.doi.org/10.1021/acs.accounts.0c00745
Acc. Chem. Res. 2021, 54, 827−836

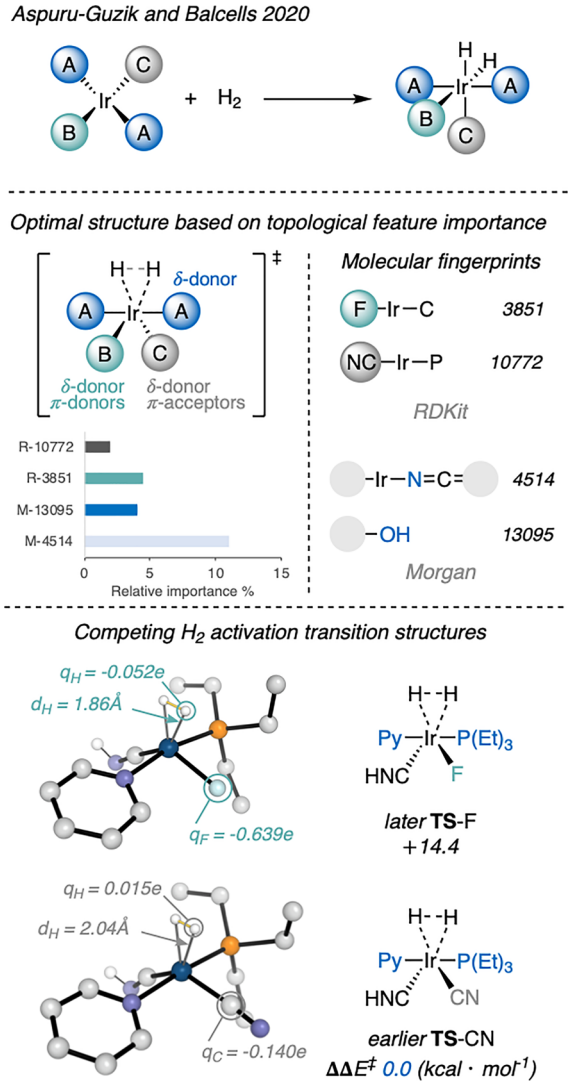


Figure 2. Application of topological fingerprints to predict activation barrier heights.

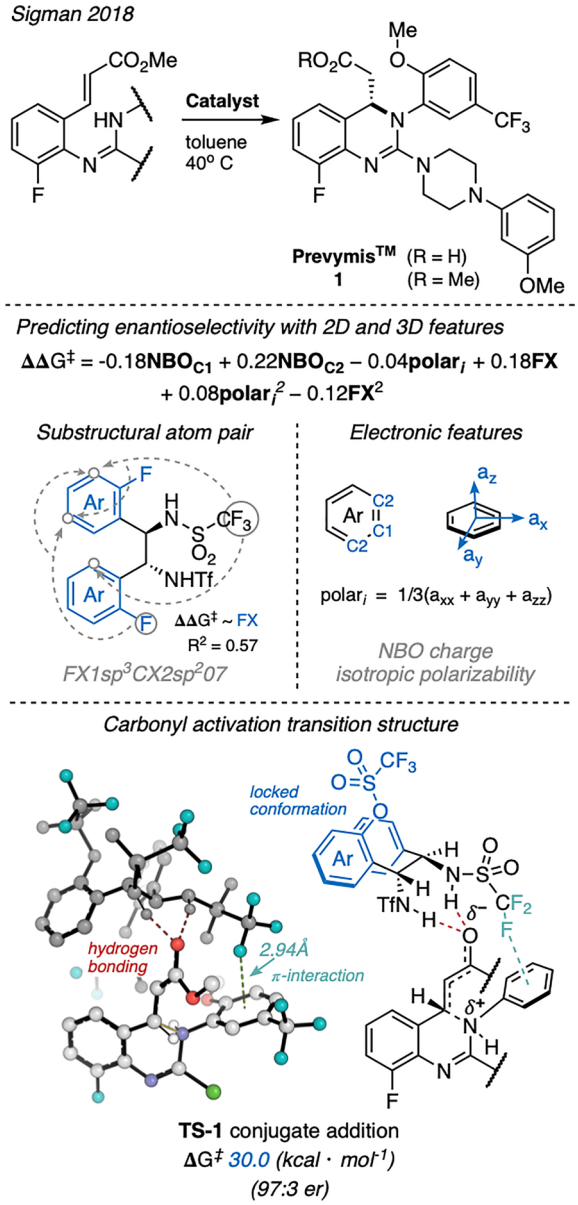


Figure 3. Combined 2D-topological and DFT-derived electronic features in enantioselectivity prediction.

unlikely to be generated during traditional intuition-driven screening.

DFT-derived descriptors are more expensive than topological features, however, mechanistic knowledge can be used to target the collection of chemically relevant features, as in Jensen et al.'s seminal QSAR study of ligand effects in olefin metathesis.[38] The use of QM-derived descriptors in ML models has proven to be a powerful quantitative approach to catalyst and ligand design along with the extraction of mechanistic insight,[39−42] and is also complemented by transition state studies.[43] We have found this approach to be particularly useful in the development of predictive models for asymmetric catalysis with chiral phosphoramidite ligands. Mechanistic studies of Rh- and Cu-phosphoramidite complexes performed in collaboration with the Fletcher and Anderson groups (Figure 4) have led us to conclude that the amine substituents of these ligands are able to interact directly with transition metal via metal-arene interactions,[44] imparting both steric and electronic effects upon levels of enantioselectivity. Catalyst descriptors derived from QM calculations have proven necessary to capture these nuanced effects in applying quantitative structure-selectivity relationships (QSSR) to experimental phosphoramidite libraries comprising 30−40 ligand structures in several copper-catalyzed asymmetric conjugate additions.[45−47]

In Cu-catalyzed asymmetric conjugate additions to β-substituted cyclopentenones, an iterative workflow was pursued in which new phosphoramidite ligands were prepared in response to statistical analysis.[46] While the ligand's BINOL backbone controls the sense of enantioinduction in this, and related transformations, the level of selectivity can be fine-tuned by altering the amino-substituents. Based on preliminary observations that relatively electron-rich substituents gave more promising enantioselectivities, both electronic and steric features were investigated in developing a statistical model (multivariate linear regression, MLR) with an experimental data set containing around 20 ligands. DFT calculations were used to obtain highest occupied molecular orbital (HOMO) energies of amino-substituents, features that are more directly relevant to metal-arene interactions than atomic charges, along

https://dx.doi.org/10.1021/acs.accounts.0c00745
Acc. Chem. Res. 2021, 54, 827−836

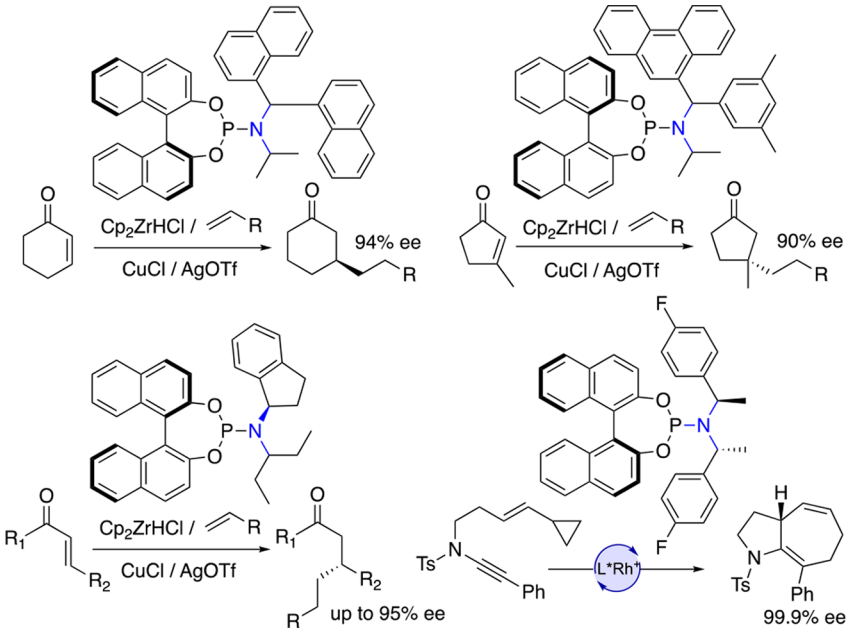


Figure 4. Computational and statistical modeling studies have aided the optimization of chiral phosphoramidite complexes used in asymmetric transformations.

with multidimensional Sterimol parameters for these groups. A positive correlation between $\Delta\Delta G^{\ddagger}$ and the substituents' HOMO energy and width was observed, which led to the preparation of a second generation of phosphoramidite ligands with extended and electron-rich π-systems as amino-substituents. Accordingly, a phenanthrene-substituted ligand gave the highest enantioselectivity of up to 92% ee (Figure 4). Interpretation of this model was supported by analysis of competing TSs, in which favorable coordination of the Cu-center by an aromatic amino-substituent occurs more easily when the enone substrate is oriented to form the major enantiomer (Figure 5).

To correlate the effect of cyclopentadienyl ligand structure with diastereoselectivity in Rh(III)-catalyzed cyclopropane insertions developed by Rovis et al. (Figure 6), we combined experimental macroscopic parameters (NMR chemical shifts and coupling constants, redox potentials) with DFT-derived catalyst descriptors.[1] Python tools were developed to compute cone angles and Sterimol parameters with different sets of van der Waals radii that were used to capture steric demands across a data set of around 20 ligands.[48] We recorded minimum and maximum values of cone angles to capture the unsymmetrical nature of most of the cyclopentadienyl ligands, similar to the way in which B1 and B5 Sterimol parameters bracket anisotropic steric demands.

MLR models were obtained in which ligand sterics were described by either the Sterimol B1 value or the minimum cone angle value—these models performed similarly, consistent with the substrate approaching the catalyst from its less hindered side. Classically, kinks or breaks in univariate correlations (e.g., Hammett plots) are indicative of a change in mechanism for certain data set members. In this multivariate correlation, the observation that the only indenyl ligand was an obvious outlier prompted us to investigate competing transition structures of the selectivity-determining step with this catalyst (Figure 6).[49] Slippage of the ligand binding mode, from η^5 toward η^3 coordination, occurs predominantly in the favored TS to accommodate the cyclopropane substrate and contributing to an enhancement in selectivity.

4. LARGE QUANTUM CHEMISTRY DATA SETS IN MACHINE LEARNING

End-to-end learning frameworks tend to outperform traditional representation methods (e.g., hand-crafted features discussed above) in the limit of large data sizes, such as those containing the results of 10^5 or more QM calculations.[50] The emergence of these large data sets has been aided by the creation of public computational results databases, such as QM9, ioChem-BD, and the MolSSI QCArchive.[51] The creation of these data sets requires large-scale, high-throughput approaches to performing QM calculations in which the (computational) chemist's traditional intuition and validity checks need to be codified and embedded in the computational pipeline (Figure 7).

In creating a data set of around 290,000 bond dissociation enthalpy (BDE) values[52] we built a computational pipeline for high-throughput DFT calculations that leverages a shared database managed through PostgreSQL. The database is initialized with realistic organic compounds from *PubChem* which were automatically and exhaustively fragmented homolytically. SMILES strings are used to index molecules when the 3D structure is not critical, but care has to be taken to ensure a consistent SMILES canonicalization. With a completed, deduplicated database of closed-shell molecules and radicals, DFT calculations were conducted by distributing calculations across a range of compute processes. A key benefit of a transactional database in managing calculation results is that parallel read and write operations can be serialized, such that rows containing molecule calculations that are in progress are locked until the calculation is finished. Various convergence checks (Figure 7) were automated to filter erroneous calculations and structures.

https://dx.doi.org/10.1021/acs.accounts.0c00745
Acc. Chem. Res. 2021, 54, 827−836

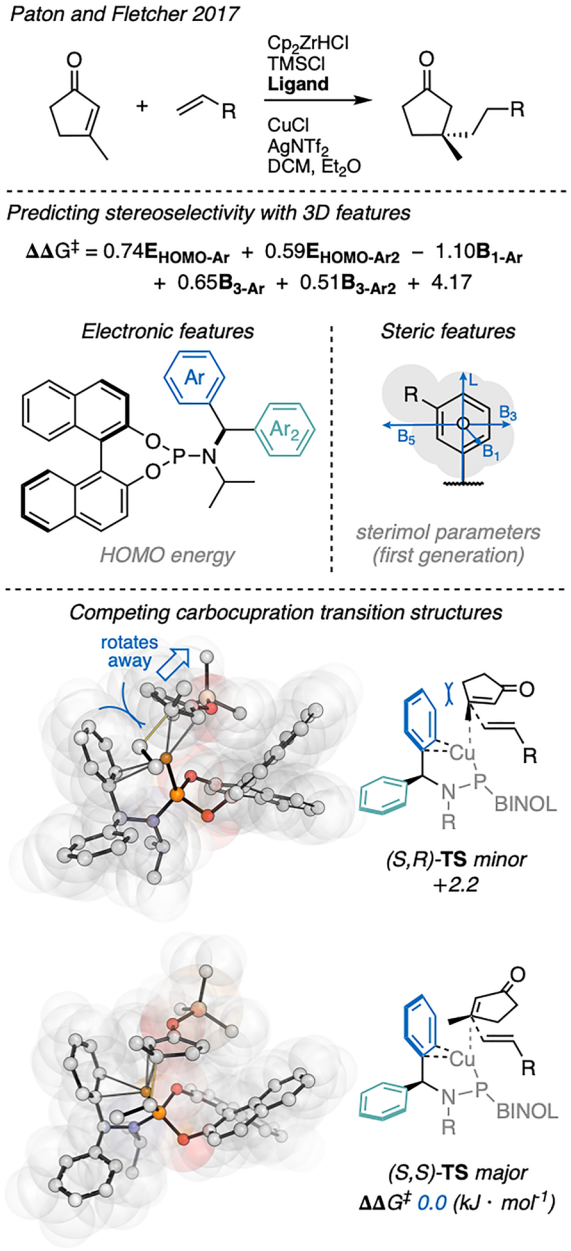


Figure 5. Chiral phosphoramidite optimization in Cu-catalyzed asymmetric conjugate addition aided by DFT-derived featurization and statistical modeling.

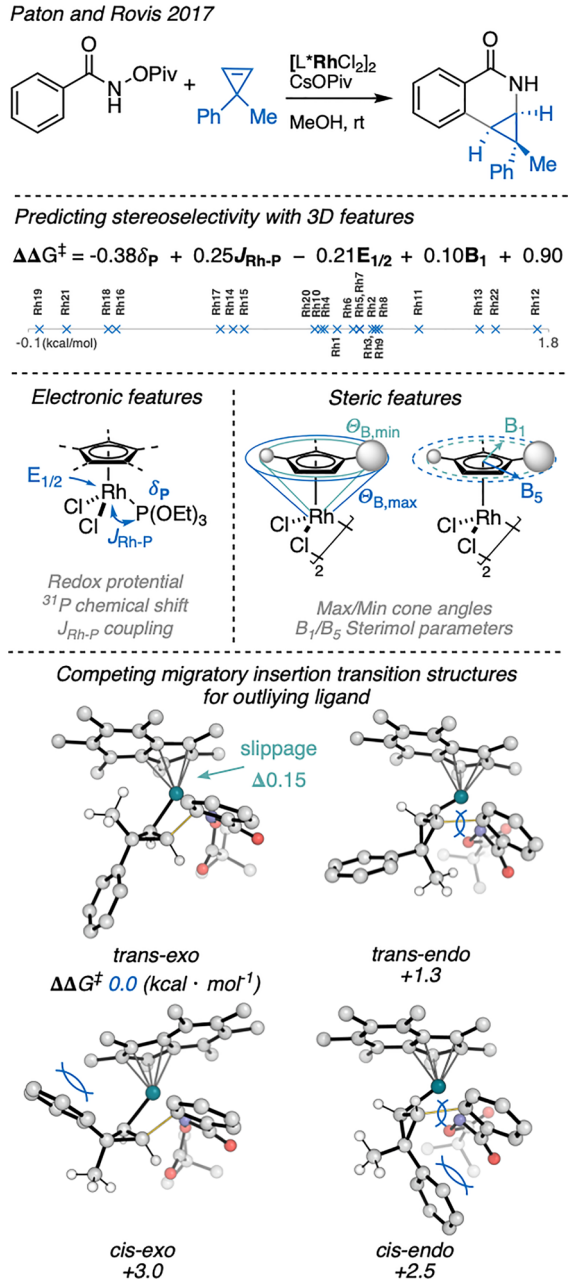


Figure 6. Correlating diastereoselectivity to cyclopentadienyl Ligand Properties in Rh(III)-catalyzed cyclopropane insertion.

5. GRAPH NEURAL NETWORKS FOR PROPERTY PREDICTION

Features used to encode structural elements in GNN approaches are themselves optimized alongside the prediction model. Initial atom (e.g., element, charge) and bond-level (e.g., bond order, element types of participating atoms) features are assigned to an initial embedding vector that is iteratively updated over a sequence of message-passing steps, where information is exchanged between neighboring nodes and edges. The model's expressive power is similar to atom or bond-centered fingerprints with a radius equal to the number of message updates. In practice, the models often identify the correct amount of information to share between nodes. Using our data set of 290,000 QM-computed BDE values, we were able to train a GNN with a mean absolute error (MAE) of 0.58 kcal mol^{-1} (vs M06-2X/def2-TZVP) for BDEs of unseen organic molecules (Figure 8). An ML model using fixed descriptors trained against 12,000 DFT values previously recorded an MAE of 3.4 kcal/mol.[53] Since the initial atom and bond attributes can be supplied directly from the molecular graph (e.g., with *rdkit*), predictions can be made from a SMILES query in less than a second, including via a Web server.[54] This approach works clearly well for applications where large amounts of data can be generated: the learning curve in Figure 8 shows the improvement in predictive accuracy as the training data is increased. For smaller data sets, this approach may not improve on traditional embedding approaches.

GNN-derived property predictions can themselves be used in place of expensive QM calculations to obtain high-quality features for statistical modeling. For instance, C−H BDEs are

https://dx.doi.org/10.1021/acs.accounts.0c00745
Acc. Chem. Res. 2021, 54, 827−836

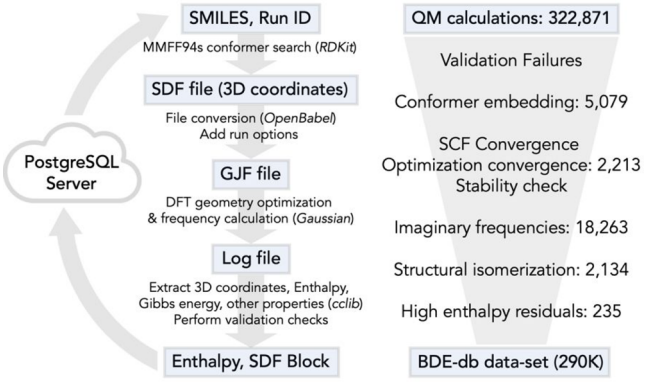


Figure 7. ComputationalPipeline and data proceeding involved in creating a large QM data set.

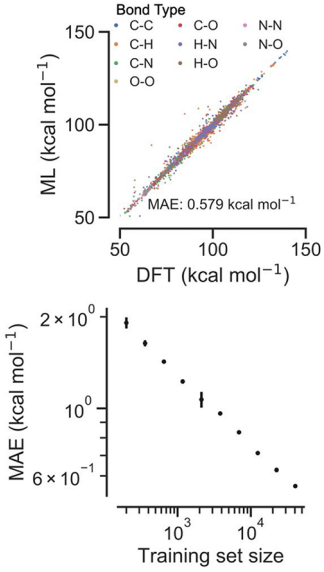
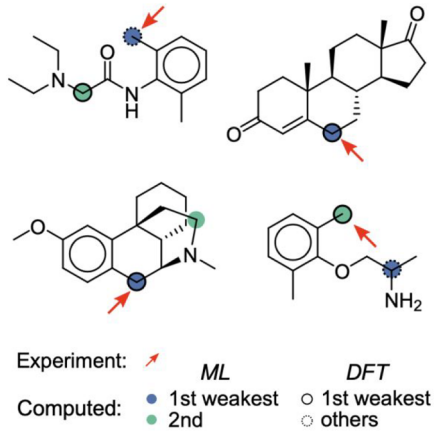


Figure 8. BDE prediction from a GNN approach (above); validation error as a function of training set size (below).

an indispensable component in building reliable models of first-pass metabolism of small molecules by cytochrome P450 enzymes.[55] A classifier was built using the relative strengths of C−H BDEs (as obtained from the GNN model) to predict the site of oxidative degradation, which classifies all bonds within a certain energy tolerance from a molecule's weakest C−H bond as potential reactive sites (Figure 9). The true positive versus false positive rates of this classifier were tabulated as the tolerance energy was increased, known as a receiver operating characteristic (ROC) curve. Our ML model gave nearly identical ROC curves to full DFT calculations, with an area under the ROC curve of 0.87 for ML and 0.86 for DFT (higher is better).[1]

We have also applied ML BDE predictions to construct a mechanistically inspired model of soot formation during the combustion or organic compounds (Figure 10). The weakest bond in each molecule identified by the GNN was used to predict the identities of the two radicals formed upon thermally induced homolysis, since the reactivity and stability of these radicals is a major determinant of sooting tendency. A QSPR model was developed for a set of 217 different fuel molecules with measured yield sooting index (YSI) values. Each molecule was represented by only two parameters: one



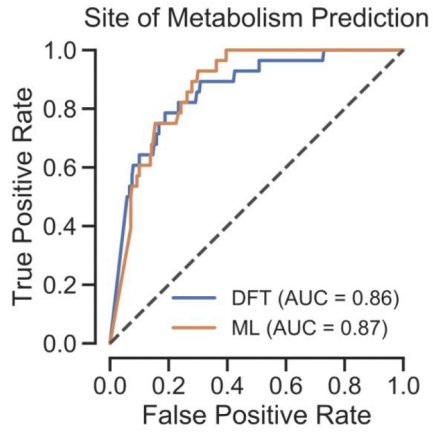


Figure 9. Regioselectivity of C−H oxidation predicted by GNN-derived BDE values.

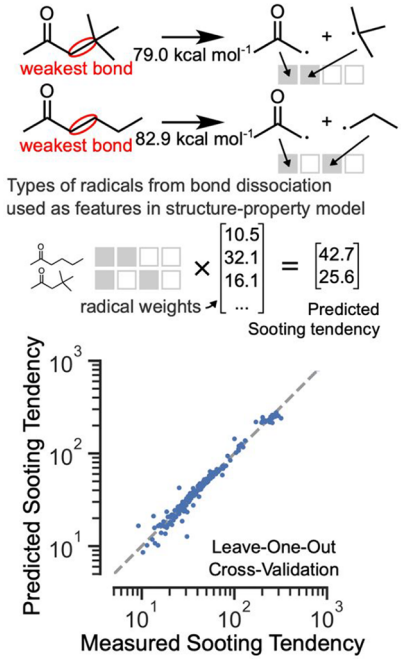


Figure 10. Mechanistically derived sooting QSPR model for combustion chemistry based on GNN-predicted BDE values.

for each of the two radicals formed by cleavage of the weakest bond. The cross-validated predictive accuracy of the new

https://dx.doi.org/10.1021/acs.accounts.0c00745
Acc. Chem. Res. 2021, 54, 827−836

model achieved a weighted least-squares loss less than half that of a recently developed group-contribution model on the same data set.[1]

6. CONCLUSIONS AND FUTURE PERSPECTIVES

Diverse structural representations spanning the hierarchy of dimensions have been instrumental in deriving statistical models for organic reactivity, selectivity, and molecular properties. Particularly in the limit of small data set sizes, expert-guided descriptors encode the nuances of organic and organometallic structures in a predictively useful and interpretable way. As data set size grows, the flexibility of more general representations such as graph neural networks can be leveraged to obtain excellent predictive performance and to rapidly generate descriptors for mechanistically informed models. The development of new descriptors, their adoption by the broader community, and evaluation for different ML prediction tasks will be enhanced by the sustained development of open source tools and libraries.

AUTHOR INFORMATION

Corresponding Author

Robert S. Paton − *Department of Chemistry, Colorado State University, Fort Collins, Colorado 80523, United States;* orcid.org/0000-0002-0104-4166; Email: robert.paton@colostate.edu

Authors

Liliana C. Gallegos − *Department of Chemistry, Colorado State University, Fort Collins, Colorado 80523, United States*

Guilian Luchini − *Department of Chemistry, Colorado State University, Fort Collins, Colorado 80523, United States*

Peter C. St. John − *Biosciences Center, National Renewable Energy Laboratory, Golden, Colorado 80401, United States*

Seonah Kim − *Biosciences Center, National Renewable Energy Laboratory, Golden, Colorado 80401, United States;* orcid.org/0000-0001-9846-7140

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.accounts.0c00745

Author Contributions

[§]L.C.G. and G.L. contributed equally. The manuscript was written through contributions of all authors.

Funding

R.S.P. thanks the NSF under the CCI Center for Computer Assisted Synthesis (CHE-1925607) for support.

Notes

The authors declare no competing financial interest.

Biographies

Liliana C. Gallegos received her B.S. degree in Biochemistry from the University of Texas at Austin. She began graduate research in organic chemistry in 2014 with Prof. Jeremy May at the University of Houston on the synthesis of bridged bicyclic natural products. In 2018, she continued her Ph.D. studies with Prof. Robert Paton at Colorado State University and is a member of the NSF Center for Computer-Assisted Synthesis (C-CAS). Her current research is focused on the interface of organic and computational chemistry for mechanistic insight and predictive modeling of catalytic reactions.

Guilian Luchini received his B.S. in chemistry from the University of Portland in 2018. He is currently a Ph.D. candidate at Colorado State University and C-CAS member. His research focuses include automating the analysis of density functional theory calculations and developing novel molecular representations and computational tools.

Peter C. St. John is from Uxbridge, MA, and received his undergraduate degree in Chemical Engineering from Tufts University in 2010. He received his Ph.D. in Chemical Engineering in 2015 from the University of California, Santa Barbara. With his thesis advisor Francis J. Doyle III, Peter used mathematical modeling to understand the gene regulatory networks underlying mammalian circadian rhythms. Peter started as a postdoctoral researcher at the National Renewable Energy Laboratory (NREL) in 2015, where he is now a staff scientist. Peter's research utilizes of machine learning techniques and high-through quantum mechanics calculations to increase the throughput of molecular design efforts and strengthen our understanding of the relationship between molecular structure and function.

Seonah Kim completed an M.S. in Computer Science at the University of Houston with B. Montgomery Pettitt and obtained her Ph.D. with Adrian Roitberg at the University of Florida in protein simulation. In 2008, she joined the group of K. N. Houk at UCLA as a postdoctoral associate working on *de novo* enzyme design. Seonah joined NREL, where she remains as a senior scientist in the Biosciences Center. Her group is interested in quantum chemical analyses of reaction mechanisms related to combustion, autoignition, and soot formation, and the development of ML fuel property prediction tools.

Robert S. Paton is from Stockport, UK. He received his undergraduate degree in Natural Sciences (2004) and Ph.D. (2008) at the University of Cambridge, working with Jonathan Goodman. After a research stay at ICIQ with Feliu Maseras, Rob moved to UCLA as a Fulbright postdoctoral fellow with Ken Houk in 2009. From 2010 to 2017 ,he was a professor at the University of Oxford, moving to Colorado State University in 2018 where he is currently an Associate Professor. The Paton group are invested in developing quantitative, data-driven approaches that make synthetic chemistry more predictable.

ACKNOWLEDGMENTS

The authors are grateful to current and past members of the Paton group and for our experimental collaborators, the Fletcher laboratory (University of Oxford), the Rovis laboratory (Columbia University), and the founding members of the NSF Center for Computer-Assisted Synthesis (Chawla, Doyle, Sarpong, Sigman, and Wiest groups) for their contribution to the work detailed in this Account.

REFERENCES

(1) Piou, T.; Romanov-Michailidis, F.; Romanova-Michaelides, M.; Jackson, K. E.; Semakul, N.; Taggart, T. D.; Newell, B. S.; Rithner, C. D.; Paton, R. S.; Rovis, T. Correlating Reactivity and Selectivity to Cyclopentadienyl Ligand Properties in Rh(III)-Catalyzed C-H Activation Reactions: An Experimental and Computational Study. *J. Am. Chem. Soc.* **2017**, *139*, 1296−1310.

(2) Brethomé, A. V.; Fletcher, S. P.; Paton, R. S. Conformational Effects on Physical-Organic Descriptors - the Case of Sterimol Steric Parameters. *ACS Catal.* **2019**, *9*, 2313−2323.

(3) St. John, P.; Guan, Y.; Kim, Y.; Kim, S.; Paton, R. S. Prediction of homolytic bond dissociation enthalpies for organic molecules at near chemical accuracy with sub-second computational cost. *Nat. Commun.* **2020**, *11*, 2328.

(4) Haghighatlari, M.; Li, J.; Heidar-Zadeh, F.; Liu, Y.; Guan, X.; Head-Gordon, T. Learning to Make Chemical Predictions: The

Interplay of Feature Representation, Data, and Machine Learning Methods. *Chem.* **2020**, *6*, 1527−1542.
(5) Chen, W. L. Chemoinformatics: past, present, and future. *J. Chem. Inf. Model.* **2006**, *46*, 2230−2255.
(6) Dudek, A. Z.; Arodz, T.; Gálvez, J. Computational Methods in Developing Quantitative Structure-Activity Relationships (QSAR): A Review. *Comb. Chem. High Throughput Screening* **2006**, *9*, 213−228.
(7) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, Germany, 2000.
(8) Fukui, K.; et al. Molecular Orbital Theory of Orientation in Aromatic, Heteroaromatic, and Other Conjugated Molecules. *J. Chem. Phys.* **1954**, *22*, 1433−1442.
(9) Tolman, C. A. Steric effects of phosphorus ligands in organometallic chemistry and homogeneous catalysis. *Chem. Rev.* **1977**, *77*, 313−348.
(10) Verloop, A. *Drug Design*. Ariens, E. J., Ed.; Academic Press: New York, 1976; Vol. *III*.
(11) (a) Poater, A.; Cosenza, B.; Correa, A.; Giudice, S.; Ragone, F.; Scarano, V.; Cavallo, L. SambVca: A Web Application for the Calculation of the Buried Volume of N-Heterocyclic Carbene Ligands. *Eur. J. Inorg. Chem.* **2009**, *2009*, 1759−1766. (b) Falivene, L.; Credendino, R.; Poater, A.; Petta, A.; Serra, L.; Oliva, R.; Scarano, V.; Cavallo, L. SambVca 2. A Web Tool for Analyzing Catalytic Pockets with Topographic Steric Maps. *Organometallics* **2016**, *35*, 2286−2293.
(12) Yada, A.; Nagata, K.; Ando, Y.; Matsumura, T.; Ichinoseki, S.; Sato, K. Machine Learning Approach for Prediction of Reaction Yield with Simulated Catalyst Parameters. *Chem. Lett.* **2018**, *47*, 284−287.
(13) Reid, J. P.; Sigman, M. S. Holistic prediction of enantioselectivity in asymmetric catalysis. *Nature* **2019**, *571*, 343−348.
(14) Santiago, C. B.; Guo, J.-Y.; Sigman, M. S. Predictive and mechanistic multivariate linear regression models for reaction development. *Chem. Sci.* **2018**, *9*, 2398−2412.
(15) Nielsen, M. K.; Ahneman, D. T.; Riera, O.; Doyle, A. G. Deoxyfluorination with Sulfonyl Fluorides: Navigating Reaction Space with Machine Learning. *J. Am. Chem. Soc.* **2018**, *140*, 5004−5008.
(16) Niemeyer, Z. L.; Milo, A. D.; Hickey, P.; Sigman, M. S. Parameterization of phosphine ligands reveals mechanistic pathways and predicts reaction outcomes. *Nat. Chem.* **2016**, *8*, 610−617.
(17) Wu, K.; Doyle, A. G. Parameterization of phosphine ligands demonstrates enhancement of nickel catalysis via remote steric effects. *Nat. Chem.* **2017**, *9*, 779−784.
(18) Amar, Y.; Schweidtmann, A. M.; Deutsch, P.; Cao, L.; Lapkin, A. Machine learning and molecular descriptors enable rational solvent selection in asymmetric catalysis. *Chem. Sci.* **2019**, *10*, 6697−6706.
(19) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59*, 3370−3388.
(20) Mater, A. C.; Coote, M. L. Deep Learning in Chemistry. *J. Chem. Inf. Model.* **2019**, *59*, 2545−2559.
(21) Battaglia, P. W.; Hamrick, J. B.; Bapst, V.; Sanchez-Gonzalez, A.; Zambaldi, V.; Malinowski, M.; Tacchetti, A.; Raposo, D.; Santoro, A.; Faulkner, R.; Gulcehre, C.; Song, F.; Ballard, A.; Gilmer, J.; Dahl, G.; Vaswani, A.; Allen, K.; Nash, C.; Langston, V.; Dyer, C.; Heess, N.; Wierstra, D.; Kohli, P.; Botvinick, M.; Vinyals, O.; Li, Y.; Pascanu, R. Relational inductive biases, deep learning and graph networks. *arXiv (Machine Learning)*, October 17, 2018, 1806.01261, ver. 3.
(22) Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K. R. SchNet - A deep learning architecture for molecules and materials. *J. Chem. Phys.* **2018**, *148*, 241722.
(23) (a) Faber, F. A.; Hutchison, L.; Huang, B.; Gilmer, J.; Schoenholz, S. S.; Dahl, G. E.; Vinyals, O.; Kearnes, S.; Riley, P. F.; von Lilienfeld, O. A. Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error. *J. Chem. Theory Comput.* **2017**, *13*, 5255. (b) Feinberg, E. N.; Sheridan, R.; Joshi, E.; Pande, V. S.; Cheng, A. C. Step Change Improvement in ADMET Prediction with PotentialNet Deep Featurization. *arXiv (Machine Learning)*, March 28, 2019, 1903.11789, ver. 1.
(24) Coley, C. W.; Barzilay, R.; Green, W. H.; Jaakkola, T. S.; Jensen, K. F. Convolutional Embedding of Attributed Molecular Graphs for Physical Property Prediction. *J. Chem. Inf. Model.* **2017**, *57*, 1757−1772.
(25) (a) Lipkowitz, K. B.; Pradhan, M. Computational studies of chiral catalysts: a comparative molecular field analysis of an asymmetric Diels-Alder reaction with catalysts containing bisoxazoline or phosphinooxazoline ligands. *J. Org. Chem.* **2003**, *68*, 4648−4656. (b) Ianni, J. C.; Annamalai, V.; Phuan, P.-W.; Panda, M.; Kozlowski, M. A Priori Theoretical Prediction of Selectivity in Asymmetric Catalysis: Design of Chiral Catalysts by Using Quantum Molecular Interaction Fields. *Angew. Chem.* **2006**, *118*, 5628−5631.
(26) Zahrt, A. F.; Athavale, S. V.; Denmark, S. E. Quantitative Structure-Selectivity Relationships in Enantioselective Catalysis: Past, Present, and Future. *Chem. Rev.* **2020**, *120*, 1620−1689.
(27) Harper, K. C.; Bess, E. N.; Sigman, M. S. Multidimensional steric parameters in the analysis of asymmetric catalytic reactions. *Nat. Chem.* **2012**, *4*, 366−374.
(28) Crawford, J. M.; Sigman, M. S. Conformational Dynamics in Asymmetric Catalysis: Is Catalyst Flexibility a Design Element? *Synthesis* **2019**, *51*, 1021−1036.
(29) Andrade, C. H.; Pasqualoto, K. F.; Ferreira, E. I.; Hopfinger, A. J. 4D-QSAR: perspectives in drug design. *Molecules* **2010**, *15*, 3281−3294.
(30) Friederich, P.; Dos Passos Gomes, G.; De Bin, R.; Aspuru-Guzik, A.; Balcells, D. Machine learning dihydrogen activation in the chemical space surrounding Vaska's complex. *Chem. Sci.* **2020**, *11*, 4584−4601.
(31) Vaska, L.; DiLuzio, J. W. Activation of Hydrogen by a Transition Metal Complex at Normal Conditions Leading to a Stable Molecular Dihydride. *J. Am. Chem. Soc.* **1962**, *84*, 679−680.
(32) Ioannidis, E. I.; Gani, T. Z. H.; Kulik, H. J. molSimplify: A toolkit for automating discovery in inorganic chemistry. *J. Comput. Chem.* **2016**, *37*, 2106−2017.
(33) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(34) *RDKit: Open-Source Cheminformatics and Machine Learning*. https://www.rdkit.org/docs/source/rdkit.Chem.Fingerprints.html.
(35) Besora, M.; Olmos, A.; Gava, R.; Noverges, B.; Asensio, G.; Caballero, A.; Maseras, F.; Pérez, P. J. A Quantitative Model for Alkane Nucleophilicity Based on C?H Bond Structural/Topological Descriptors. *Angew. Chem., Int. Ed.* **2020**, *59*, 3112−3116.
(36) Metsanen, T. T.; Lexa, K. W.; Santiago, C. B.; Chung, C. K.; Xu, Y.; Liu, Z.; Humphrey, G. R.; Ruck, R. T.; Sherer, E. C.; Sigman, M. S. Combining traditional 2D and modern physical organic-derived descriptors to predict enhanced enantioselectivity for the key aza-Michael conjugate addition in the synthesis of Prevymis (letermovir). *Chem. Sci.* **2018**, *9*, 6922−6927.
(37) Chung, C. K.; Liu, Z.; Lexa, K. W.; Andreani, T.; Xu, Y.; Ji, Y.; DiRocco, D. A.; Humphrey, G. R.; Ruck, R. T. Asymmetric Hydrogen Bonding Catalysis for the Synthesis of Dihydroquinazoline-Containing Antiviral, Letermovir. *J. Am. Chem. Soc.* **2017**, *139*, 10637−10640.
(38) Occhipinti, G.; Bjørsvik, H. R.; Jensen, V. R. Quantitative Structure—Activity Relationships of Ruthenium Catalysts for Olefin Metathesis. *J. Am. Chem. Soc.* **2006**, *128*, 6952−6964.
(39) Orlandi, M.; Coelho, J. A. S.; Hilton, M. J.; Toste, F. D.; Sigman, M. S. Parameterization of non-covalent interactions for transition state interrogation applied to asymmetric catalysis. *J. Am. Chem. Soc.* **2017**, *139*, 6803−6806.
(40) Milo, A.; Neel, A. J.; Toste, F. D.; Sigman, M. S. A data-intensive approach to mechanistic elucidation applied to chiral anion catalysis. *Science* **2015**, *347*, 737−743.
(41) Maley, S. M.; Kwon, D.-H.; Rollins, N.; Stanley, J. C.; Sydora, O. L.; Bischof, S. M.; Ess, D. H. Quantum-Mechanical Transition-State Model Combined with Machine Learning Provides Catalyst Design Features for Selective Cr Olefin Oligomerization. *Chem. Sci.* **2020**, *11*, 9665−9674.

https://dx.doi.org/10.1021/acs.accounts.0c00745
Acc. Chem. Res. 2021, 54, 827−836

(42) Durand, D.; Fey, N. Computational Ligand Descriptors for Catalyst Design. *Chem. Rev.* **2019**, *119*, 6561−6594.
(43) Reid, J. P.; Sigman, M. S. Comparing quantitative prediction methods for the discovery of small-molecule chiral catalysts. *Nat. Rev. Chem.* **2018**, *2*, 290−305.
(44) Straker, R. N.; Peng, Q.; Mekareeya, A.; Paton, R. S.; Anderson, E. A. Computational ligand design in enantio- and diastereoselective ynamide [5 + 2] cycloisomerization. *Nat. Commun.* **2016**, *7*, 10109.
(45) Ardkhean, R.; Roth, P. M. C.; Maksymowicz, R. M.; Curran, A.; Peng, Q.; Paton, R. S.; Fletcher, S. P. Enantioselective conjugate addition catalyzed by a copper-phosphoramidite complex: Computational and experimental exploration of asymmetric induction. *ACS Catal.* **2017**, *7*, 6729−6737.
(46) Ardkhean, R.; Mortimore, M.; Paton, R. S.; Fletcher, S. P. Formation of quaternary centres by copper catalysed asymmetric conjugate addition to -substituted cyclopentenones with the aid of a quantitative structure-selectivity relationship. *Chem. Sci.* **2018**, *9*, 2628−2632.
(47) Brethomé, A. V.; Paton, R. S.; Fletcher, S. P. Retooling Asymmetric Conjugate Additions for Sterically Demanding Substrates with an Iterative Data-Driven Approach. *ACS Catal.* **2019**, *9*, 7179−7187.
(48) (a) Jackson, K. E.; Paton, R. S. Sterimol. *Zenodo*, 2021, DOI: 10.5281/zenodo.1320768 (accessed Jan 25, 2021). (b) Brethomé, A. V.; Fletcher, S. P.; Paton, R. S. wSterimol. *Zenodo*, 2021, DOI: 10.5281/zenodo.1434440 (accessed Jan 25, 2021).
(49) Semakul, N.; Jackson, K. E.; Paton, R. S.; Rovis, T. Heptamethyl Indenyl (Ind*) Enables Diastereoselective Benzamidation of Cyclopropenes via Rh(III)-Catalyzed C-H Activation. *Chem. Sci.* **2017**, *8*, 1015−1020.
(50) Faber, F. A.; Hutchison, L.; Huang, B.; Gilmer, J.; Schoenholz, S. S.; Dahl, G. E.; Vinyals, O.; Kearnes, S.; Riley, P. F.; von Lilienfeld, O. A. Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error. *J. Chem. Theory Comput.* **2017**, *13*, 5255−5264.
(51) Bo, C.; Maseras, F.; López, N. The role of computational results databases in accelerating the discovery of catalysts. *Nat. Catal.* **2018**, *1*, 809−810.
(52) St John, P.; Guan, Y.; Kim, Y.; Etz, B. D.; Kim, S.; Paton, R. S. Quantum chemical calculations for over 200,000 organic radical species and 40,000 associated closed-shell molecules. *Sci. Data* **2020**, *7*, 244.
(53) Qu, X.; Latino, D. A.; Aires-De-Sousa, J. A big data approach to the ultra-fast prediction of DFT-calculated bond energies. *J. Cheminf.* **2013**, *5*, 1−13.
(54) ALFABET: A Machine Learning derived, Fast, Accurate BdE Tool. *NREL*. https://bde.ml.nrel.gov (accessed 2020-11-3).
(55) Drew, K. L. M.; Reynisson, J. The impact of carbon-hydrogen bond dissociation energies on the prediction of the cytochrome P450 mediated major metabolic site of drug-like compounds. *Eur. J. Med. Chem.* **2012**, *56*, 48−55.

https://dx.doi.org/10.1021/acs.accounts.0c00745
Acc. Chem. Res. 2021, 54, 827−836