A Structure-Based Platform for Predicting Chemical Reactivity

Frederik Sandfort,[a,‡] Felix Strieth-Kalthoff,[a,‡] Marius Kühnemund,[b,c,‡] Christian Beecks,[b] and Frank Glorius[a,*]

a) Organisch-Chemisches Institut, Westfälische Wilhelms-Universität Münster, Corrensstraße 40, 48149 Münster (Germany).

b) Institut für Informatik, Westfälische Wilhelms-Universität Münster, Einsteinstraße 62, 48149 Münster (Germany).

c) Institut für Wirtschaftsinformatik, Westfälische Wilhelms-Universität Münster, Leonardo-Campus 3, 48149 Münster (Germany).

*glorius@uni-muenster.de

‡ These authors contributed equally.

Despite their enormous potential, machine learning methods have only found limited application in predicting reaction outcomes, as current models are often highly complex and, most importantly, are not transferrable to different problem sets. Herein, we present the direct utilization of Lewis structures in a machine learning platform for diverse applications in organic chemistry. Therefore, an input based on multiple fingerprint features (MFF) as a universal molecular representation was developed and used for problem sets of increasing complexity: First, molecular properties across a diverse array of molecules could be predicted accurately. Next, reaction outcomes such as stereoselectivities and yields were predicted for experimental data sets that were previously evaluated using (complex) problem-oriented descriptor models. As a final application, a systematic high-throughput data set showed good correlation when using the MFF model, which suggests that this approach is general and ready for immediate adoption by chemists.

Introduction

While chemical intuition, based on experience, expertise and mechanistic understanding, has driven the discovery of new transformations in organic synthesis, the accurate prediction of the outcome of a single chemical reaction remains a major challenge for both, man's instinct and computer-based models.[1] In this regard, the optimization of an organic transformation therefore requires the collection of large amounts of empirical data. Even for well-established methodologies, experienced chemists frequently fail to predict whether a (complex) substrate might undergo the desired transformation, making the field of chemical synthesis highly challenging and laborious.[2,3] Whereas qualitative estimations based on mechanistic understanding can be accurate, the quantitative prediction of chemical reactivity is almost impossible with chemical intuition alone, mainly because of the enormously complicated correlation between structure and reactivity (Figure 1a).

Due to this complexity, chemists have tried to simplify the overall problem by correlating (known or easily accessible) molecular properties with a compound's reactivity. In this quantitative approach towards reactivity prediction, linear free energy relationships (LFER) are identified and solved using multivariate linear regression (MLR) models.[4,5] This statistical approach, which has been established by Sigman and co-workers, relies on physically meaningful parameters to represent the structures of interest.[6–8] These molecular descriptors are electronic or steric parameters, which can be determined by experimental or computational means (mainly DFT calculations). Parameter selection and model development are typically carried out in a lengthy iterative workflow until good correlation is achieved. The major strong point of this process is that, in many cases, valuable information on the underlying mechanism can be obtained.[9,10] Furthermore, after the successful identification of mechanistically relevant descriptors, the amount of datapoints required for a MLR prediction model can be comparatively small. However, this requires a representative set of training data, which has to be carefully considered.[6]

A more general approach to predict reactivity building on a given data set is machine learning. Due to their ability to recognize complex patterns, machine learning algorithms have been widely used in many scientific fields.[11] In the context of chemoinformatics, they were applied for use in drug discovery,[12–14] computer-aided synthesis planning[15–18] and the prediction of possible organic reaction products.[19–22] However, the quantitative prediction of chemical reactivity with machine learning algorithms has not been approached until very recently, primarily due to the lack of available data.[23] In general, the generation of datapoints in chemistry is traditionally rather expensive. While molecular properties can usually be obtained by density functional theory (DFT) calculations, there is no other cheap alternative for the generation of reaction-specific data better than physically going to the lab and running the experiment. Thus to generate this data, technical solutions for high-throughput experimentation under batch and flow conditions have been developed to carry out thousands of reactions in a short time using just a few milligrams of material.[24–26] These tools, combined with *in vitro* and *in silico* compound libraries, have recently opened the field of reaction development to machine learning models.[27–29] In pioneering work, Dreher, Doyle and co-workers were able to predict the reaction yields of C–N cross coupling reactions using a data set of more than 4000 reactions.[30] Furthermore, Denmark and co-workers could predict enantioselectivities using chiral phosphoric acid (CPA) catalysts based on a data set with more than 1000 experiments.[31] Similar to MLR studies, these methods rely on physically meaningful molecular descriptors (Figure 1b). In a thorough analysis, steric and electronic parameters are selected based on the underlying mechanism of the transformation and relevant properties of each reactant. In a next step, these parameters are determined by means of DFT calculations, as uniform sets of experimental values are not usually available for such a large number of molecules. However, while DFT calculation of properties is considered fast for a specific compound, the generation of multiple descriptors for a library of substrates can be time-consuming, especially so in cases where more complex structures are involved. Thus, DFT computations using established methods can easily become time-limiting, as these methods typically scale with a computation time complexity of O(N^4) depending on the size (N) of the system.[32]

Results and Discussion

The ultimate goal of machine learning in the context of (organic) chemistry is to "make computers develop a quantitative chemical intuition, based on a rationalization of all underlying fundamentals".[1,33–36] To achieve this goal, the full computational recognition of all involved compounds, i.e. a complete translation of the molecular structure into a machine-readable representation, is required.[37] In previous studies, this was approached by using an array of physicochemical parameters/descriptors (Figure 1b). However, (biased) manual selection of experimental or calculated descriptors can oversimplify the problem, which can introduce systematic errors in prediction through the loss of seemingly unimportant information.[38] Moreover, as every molecule and reaction is unique, the selection of a consistent and general set of physical properties as universal molecular descriptors is highly challenging.

Considering this challenge, we proposed that all physical parameters can eventually be traced back to a compound's (2D) Lewis structure, which contains the connectivity of all atoms in a simplified molecular topology. For example, DFT calculations require 3D structures as input, which are ultimately generated based on the topological connections given in the Lewis formula. Thus, Lewis structures can be considered to be the lowest common denominator in organic chemistry. In fact, human chemical intuition has often been based on understanding and rationalizing 2D connectivity. However, while human receptivity is limited, machine learning algorithms bear the potential of identifying new and unknown patterns within molecular structures.[11] Thus, we hypothesized that Lewis structures could serve as ideal inputs for the machine-learning-based prediction of chemical reactivity (Figure 1c).

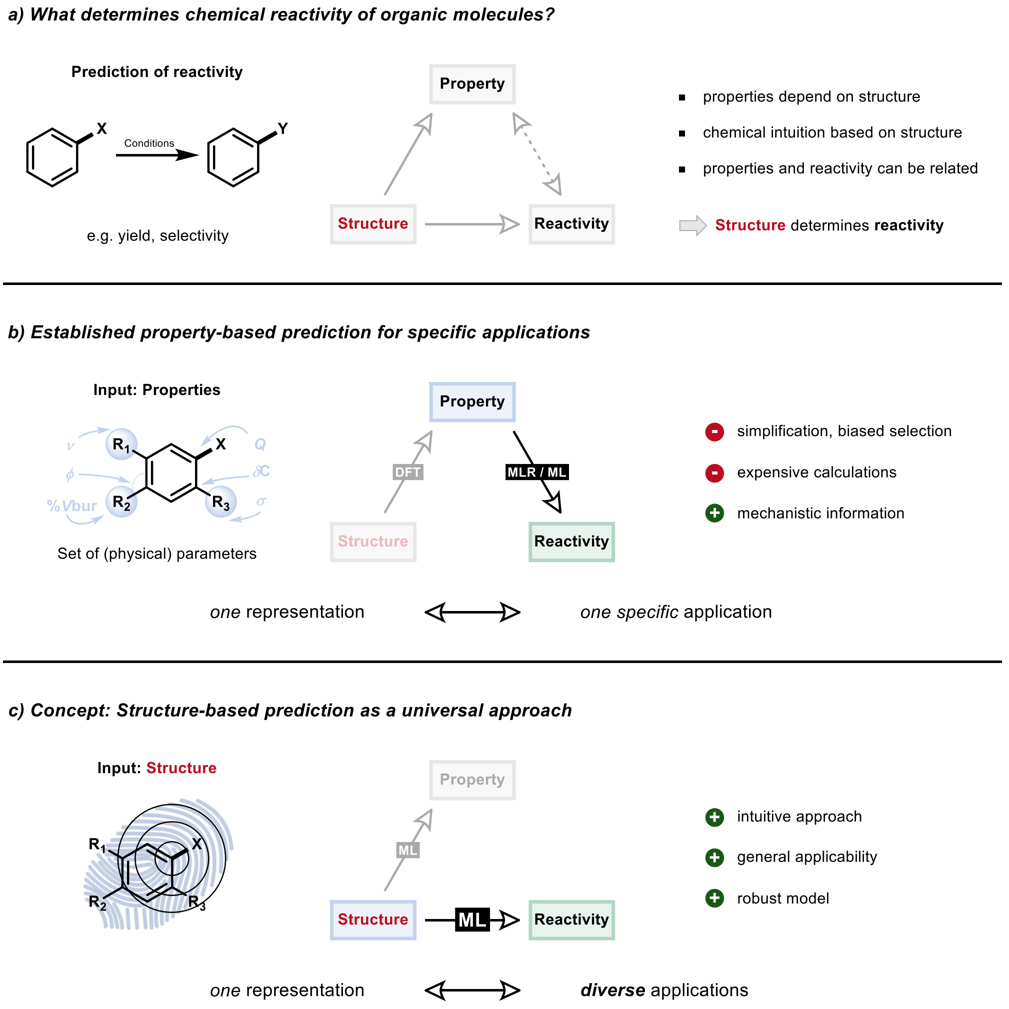


Figure 1. Conceptual sketch of a universal prediction platform for organic chemistry. a) General correlation of structure, property and reactivity of organic molecules. b) Property-based prediction for specific applications using physical parameters as molecular descriptors. c) Concept of structure-based prediction as a universal approach. One model can be used for several applications such as prediction of properties, selectivities and yields. ML: machine learning, MLR: multivariate linear regression, DFT: density functional theory.

As a first step towards realizing this goal, we sought to identify a suitable machine-readable representation of Lewis structures. Due to their simple and explicit description of a molecular structure, string representations (SMILES, InChI, etc.) are commonly used in chemoinformatics.[39,40] However, algorithm-based pattern recognition is challenging with such strings, as they do not have a fixed length or a defined starting point.[41] A straightforward solution to this problem is offered by molecular fingerprints.[42] These bit vectors have been designed for substructure and similarity searches, and have been successfully used in virtual screening for drug discovery.[43,44] Furthermore, they have been used in quantitative structure activity relationship (QSAR) models to correlate biological activity or properties with chemical structure.[45–47] A number of different fingerprints have been developed, which can all be computed efficiently on a subsecond timescale.[42] Moreover, independent of molecule size, molecular fingerprints are consistent in length. Thus, we considered that they would be well-suited as inputs for machine learning models.[48]

Herein, we present a structure-based machine learning platform for property and reactivity prediction in (organic) chemistry. This approach can be easily adopted and applied to existing problem sets, since it relies only on SMILES representations of all involved molecules as inputs, which are automatically converted into the corresponding molecular fingerprints. Specifically, for each molecule an array of 24 diversely configured fingerprints is generated using RDKit as a python package.[49] This multiple fingerprint feature (MFF) input, matched to the observed experimental data, is used to train a machine learning model, which is eventually capable of predicting properties or reactivity beyond the training data set (Figure 2).

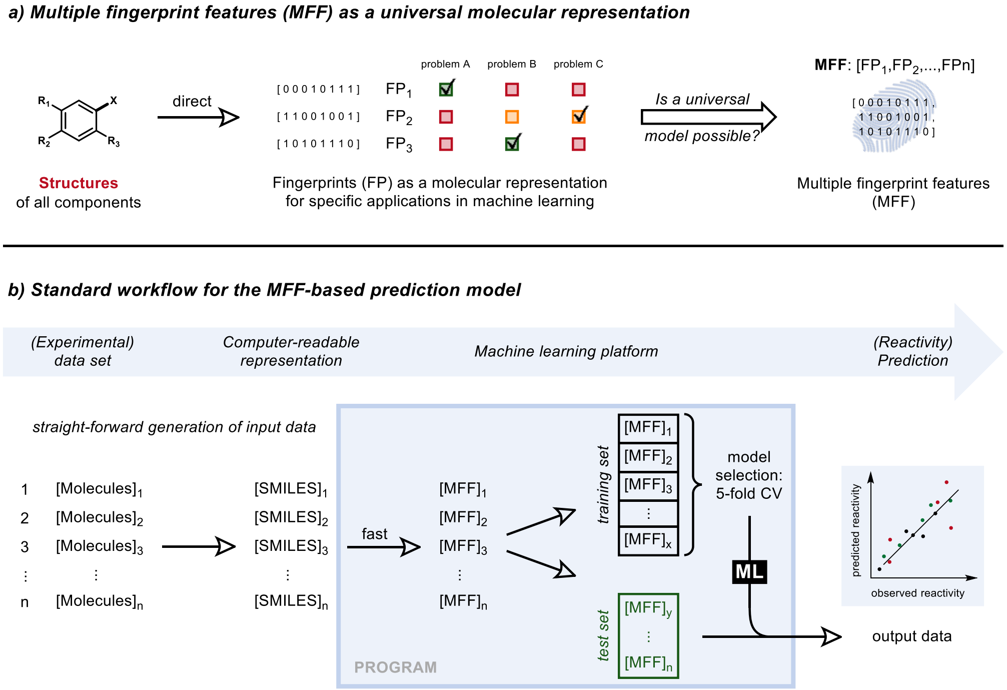


Figure 2. The multiple fingerprint feature (MFF) model. a) Use of fingerprints as a molecular representation for machine learning models. The multiple fingerprint feature can be applied to diverse problem sets. b) Standard workflow for the prediction model based on multiple fingerprint features (MFF). Only features are depicted, and targets are left out for clarity. Model selection is performed by 5-fold nested cross validation (CV). ML: machine learning.

A vast number of machine learning algorithms have been developed over the last decades, which can be loosely categorized into distance-based and non-distance-based methods.[50] Distance-based approaches build on the assumption that similar input generates similar output, and *vice versa*. However, in organic chemistry, structural similarity does not necessarily correlate with similar reactivity. Thus, we assumed that non-distance-based algorithms, such as random forests or neural networks that rely on (complex) decision trees, were best suited for our predictive tool. We selected a multi-parameter random forest algorithm, as it offers many advantages such as high robustness and low computational cost. However, it should be noted that its extrapolative properties beyond the trained target space are inherently limited. We implemented a variational hyperparameter optimization for model selection based solely on the training data (nested cross-validation, Figure 2b).[51] The implementation was conducted using the Python package Scikit-learn.[52]

At the outset of our investigations, we observed that every single fingerprint performed differently for each problem set and that no universally applicable fingerprint existed (Figure 2a). Even within a single specific data set, the best-performing fingerprint changed for different train/test splittings. Furthermore, the selection of a best single fingerprint for the prediction of an unknown (out-of-sample) test set, solely based on the training data, could not be achieved in all cases.[53] To circumvent this problem, we sought to use an array of multiple fingerprints in order to generate a more general and robust input that would provide the best possible representation of the Lewis structure.

Current descriptor-based prediction models typically require molecules that have at least one structural motif, atom or functional group – a reactive center – in common. Thus, a model which was originally designed to predict yields for reactions of aryl halides, will – without major modifications – not be able to assess e.g. chiral catalyst systems (and *vice versa*). Our fingerprint-based model, however, was developed to be applicable to a variety – ultimately any – organic chemical prediction problem. Therefore, as a first step, its applicability to a series of structurally different molecules was demonstrated. We decided to investigate the HOMO (highest occupied molecular orbital) – LUMO (lowest unoccupied molecular orbital) gap, as obtained from DFT calculations (Figure 3a, see SI for computational details). For this purpose, we selected a database of more than 2900 small organic molecules from our group's chemical inventory. Here, we were pleased to find that the MFF model was capable of predicting HOMO–LUMO gaps with high accuracy: Using a random 70/30 split of the abovementioned dataset, very good correlation between observed and predicted HOMO–LUMO gaps was obtained, as indicated by an average R^2 of 0.89 over ten random cross-validation steps (Figure 3b).

Since the HOMO–LUMO gap is a property of the overall molecule, the multiple fingerprint feature model seems to not only represent and compare local substructures of a molecule, but rather global molecular characteristics. This result served to reinforce our original hypothesis, that (computed) molecular properties can eventually be traced back to patterns in the (2D) Lewis structure.

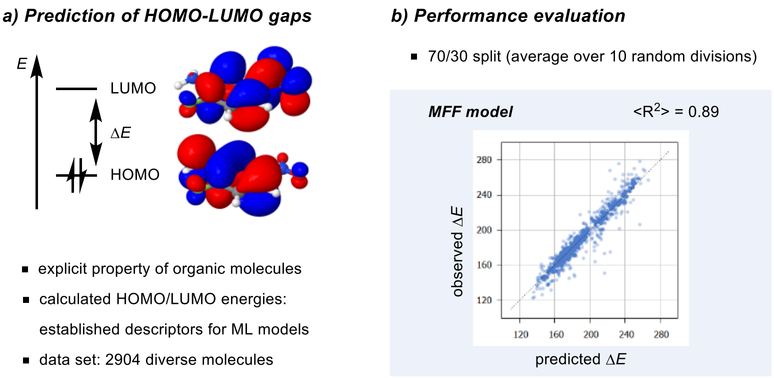
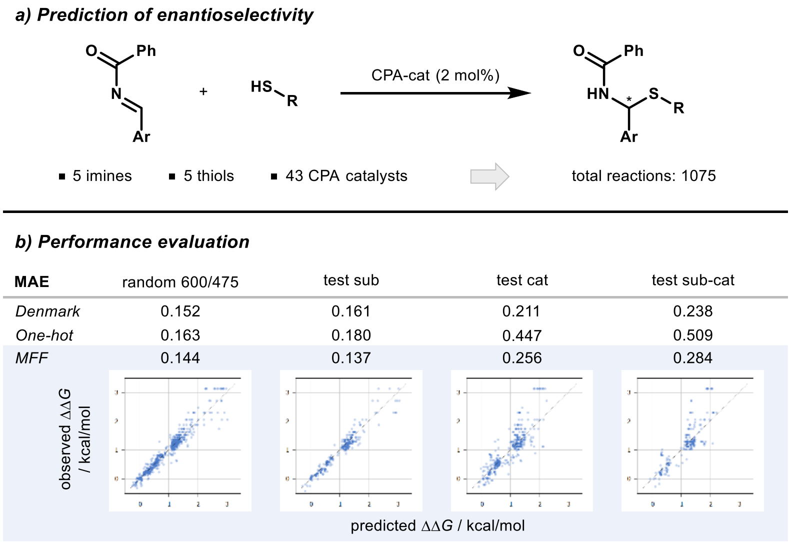


Figure 3. Prediction of orbital energies. a) Calculated HOMO–LUMO gaps as an explicit molecular property. Calculated HOMO and LUMO geometries shown for 2-fluoro-5-nitroaniline. b) Performance evaluation of the MFF model. Plot given for one explicit division.

It should also be noted that orbital energies have commonly been used as descriptors in MLR and machine learning models.[6,30] This indicates that our model should also be applicable to more complex reactivity-based data sets. More precisely, we aimed to demonstrate this generality by testing multiple data sets of increasing complexity available in the literature, that had successfully been applied in machine learning studies. It should be emphasized that we do not intend to question the relevance and importance of the reported models, but rather wish to present an alternative approach in order to unify organic chemical prediction models.

The prediction of stereoselectivities of catalytic reactions has been of great interest to the chemical community and a major focus of many MLR models.[4–10] Recently, Denmark and co-workers described a machine-learning based approach for the prediction of enantioselectivity using chiral phosphoric acid (CPA) catalysts.[31] The possibility to predict (parts) of this data using MLR models trained with other nucleophiles was later demonstrated by Sigman and co-workers.[54] In the initial work, Denmark *et al*. chose an asymmetric *N*,*S*-acetal formation as a model reaction. The training set included combinatorial variations of 43 CPA catalysts, five *N*-acyl imines and five thiols, resulting in a total of 1075 reactions (Figure 4a). A new steric parameter, the average steric occupancy (ASO), based on DFT-computed 3D representations of multiple conformers was developed to represent the catalysts. Weighted grid point occupancies in combination with calculated electronic parameters were used to train a machine learning model in order to predict enantioselectivity ($\Delta\Delta G$ in kcal/mol). A distance-based support vector machine algorithm was found to perform best on a random 600/475 split of training and test data (<MAE> = 0.152 kcal/mol, average over ten random divisions).[31] We found that our model performed with slightly higher accuracy (<MAE> = 0.144 kcal/mol, average over ten random divisions) using a random forest algorithm, while a one-hot encoded model as statistical probe resulted in lower correlation (<MAE> = 0.163 kcal/mol, average over ten random divisions) (Figure 4b). In their original work, the authors divided the data for out-of-sample prediction into a common training set, a test set for substrates (sub), a test set for catalysts (cat) and one for both (sub-cat). The same division was analyzed using our MFF model, which showed good accuracy in all three test sets. In particular, the performance for the most challenging catalyst out-of-sample predictions (cat, sub-cat) stands out. While a one-hot encoded model resulted in very low correlation measures, the simple MFF model performed nearly as well as the original complex descriptor model. These results further demonstrate the validity of this multiple fingerprint feature approach and its transferability to different chemical problems.

Interestingly, the substrate out-of-sample test set could also be predicted with high accuracy using a one-hot encoded model, which only contains the information whether a compound is included in the reaction. This demonstrates that such models can be powerful for pattern recognition in complex statistically correlated data sets, which has also been utilized by Cronin and co-workers.[29]

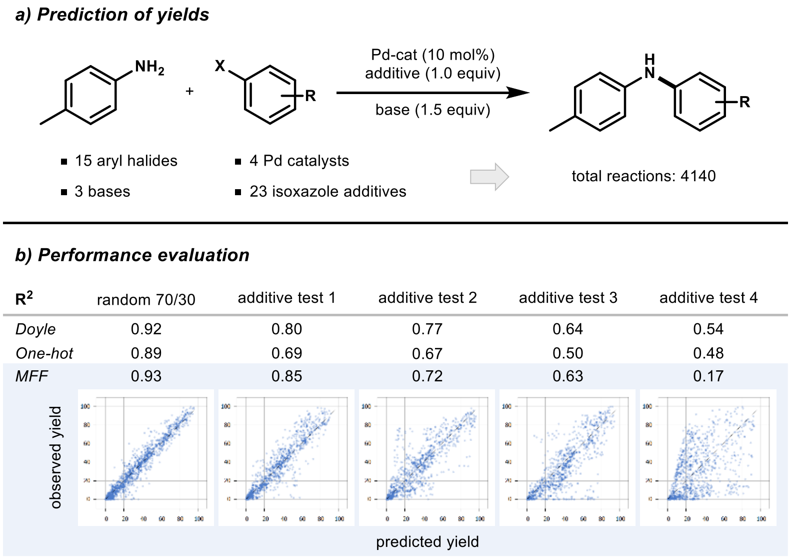


MAE	random 600/475	test sub	test cat	test sub-cat
Denmark	0.152	0.161	0.211	0.238
One-hot	0.163	0.180	0.447	0.509
MFF	0.144	0.137	0.256	0.284

Figure 4. Prediction of enantioselectivities. a) Asymmetric *N,S*-acetal formation using chiral phosphoric acid (CPA) catalysts by Denmark *et al.*[31] b) Comparison of the original model, a one-hot encoded model as statistical probe and the MFF model (mean absolute error (MAE) in kcal/mol given as correlation measure). Plots for the MFF model given.

In comparison to stereoselectivities, the quantitative prediction of yields can be even more demanding, since they are influenced by many parameters and not only rely on one elementary step. In a recent report, Dreher, Doyle and co-workers described a machine learning approach to predict reaction performance in C–N cross coupling reactions.[30] The training data, including combinatorial variation of four reaction components applying an additive-based approach,[55] was collected using high-throughput experimentation. All possible combinations of 15 aryl halides, four ligands, three bases and 23 isoxazole additives were evaluated in a total of 4140 reactions (Figure 5a). The molecules were represented by electronic, atomic and vibrational descriptors that were extracted from DFT calculations. A variety of regression models was subjected to a random 70/30 split into training and test data, and a random forest model was found to show the best performance in predicting product yields ($R^2 = 0.92$). Pleasingly, we found that our simplified model based on a universal input strategy resulted in comparable correlation (Figure 5b, $\langle R^2 \rangle = 0.93$, average over ten random divisions).

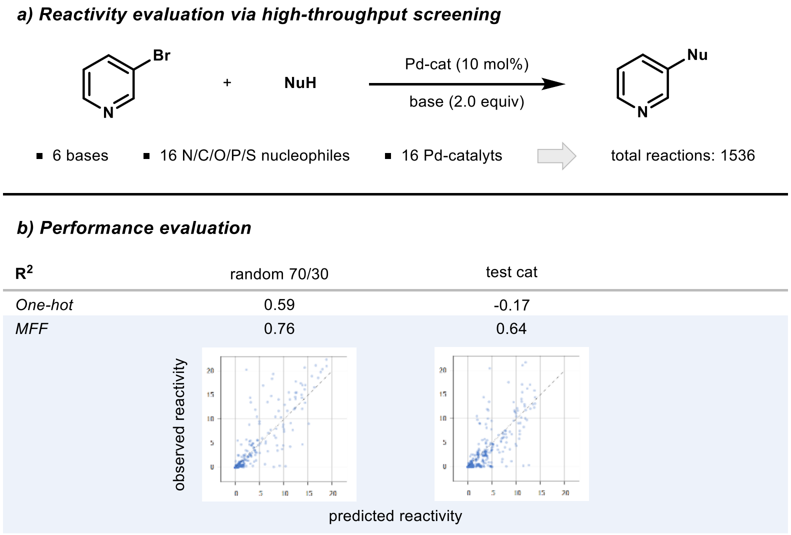
However, such a random 70/30 split of the entire combinatorial data, consisting of only 46 different molecules, results in a training set which contains all molecules at least once. Consequently, a one-hot encoded model as statistical probe showed slightly lower but still very good performance ($\langle R^2 \rangle = 0.89$, average over ten random divisions).[56] The appropriate test to prove the relevance of chemical features in such models is out-of-sample prediction, i.e. the prediction of reactivity for molecules which were not included in the training data set. Thus, the authors split the isoxazole additives into a variety of representative training and test sets and could prove good performance of the chemical feature model in these cases.[57] The same division for out-of-sample prediction using multiple fingerprints features (MFF) as input, showed comparable correlation in three of four test sets.[58] It should be emphasized that the one-hot encoded models performed significantly worse (Figure 5b).



R^2	random 70/30	additive test 1	additive test 2	additive test 3	additive test 4
Doyle	0.92	0.80	0.77	0.64	0.54
One-hot	0.89	0.69	0.67	0.50	0.48
MFF	0.93	0.85	0.72	0.63	0.17

Figure 5. Prediction of yields. a) C–N cross coupling reactions of 4-methylaniline with various aryl halides by Dreher, Doyle *et al.*[30] b) Comparison of the original model, a one-hot encoded model as statistical probe and the MFF model (R^2 given as correlation measure). Plots for the MFF model given.

As a last – and most demanding – application, we aimed to use an experimental data set which had neither been specifically designed nor used for machine learning. In 2015, Cernak, Dreher *et al.* performed an automated high-throughput screening on a nanomole scale in order to find suitable coupling conditions for C–heteroatom bond forming reactions.[24] In a Pd-catalyzed reaction, coupling of one electrophile, 3-bromopyridine, with 16 different nitrogen, oxygen, carbon, phosphorus and sulfur nucleophiles was evaluated (Figure 6a). For this, 16 catalysts and six bases were investigated, giving a total of 1536 reactions which were carried out on nanomole scale using around 0.2 mg of material per reaction in less than one day. The relative conversion, determined by liquid chromatography–mass spectrometry (LC–MS) analysis, was used for quantification. This exemplifies a "real-world problem", as for unknown complex molecules an accurate yield determination can hardly be carried out before the synthesis of the compound. Thus, we aimed to directly predict the relative conversion using our MFF model in a similar manner to the previously reported yield prediction. Encouragingly, a random 70/30 split of the data set resulted in good correlation for reactivity prediction (Figure 6b, $\langle R^2 \rangle = 0.76$, average over ten random divisions). In contrast, a one-hot encoded model as the statistical probe showed significantly lower performance ($\langle R^2 \rangle = 0.59$, average over ten random divisions), demonstrating the generalizability of the MFF approach. Moreover, out-of-sample prediction for catalysts could be achieved. Therefore, the data of twelve catalysts was used to predict the reaction outcomes of the remaining four catalysts (test cat). While a one-hot encoded model gave no correlation at all, the MFF model showed satisfactory performance (R^2 = 0.64), further underlining its abilities to learn chemical structures and potential to predict chemical reactivity.



R^2	random 70/30	test cat
One-hot	0.59	-0.17
MFF	0.76	0.64

Figure 6. Prediction of reactivity. a) Nanomole-scale reactivity evaluation of C–heteroatom coupling reactions by Cernak, Dreher *et al.*[24] b) Comparison of a one-hot encoded model as statistical probe and the MFF model (R^2 given as correlation measure). Plots for the MFF model given.

Conclusion

In summary, we report a machine-learning-based prediction platform for diverse applications in organic chemistry. A numerical representation of 2D Lewis structures, the multiple fingerprint feature (MFF), was developed, which can be computed efficiently in seconds for a large number of molecules. Since this model is based on the assumption that reactivity can be directly derived from molecular structures, it should be applicable to any problem related to (small) organic molecules. Its generalizability was demonstrated on four examples of increasing complexity using a random forest algorithm. First, the ability of the MFF input to represent and compare diverse molecular structures was demonstrated by the prediction of HOMO–LUMO gaps as an explicit molecular property. Furthermore, prediction of reaction performance, enantioselectivities and yields could be achieved with similar accuracy to established descriptor-based models, which rely on problem-oriented parameter selection. It should be noted that the generation of the simple and intuitive structure-based model is several orders of magnitude faster. Finally, the general applicability of this approach was shown on a data set (obtained from high-throughput experimentation) which had not been used for machine learning before. To aid the rapid uptake of this approach, we provide a readily applicable software tool, and the development of an extended software package is ongoing in our group. In the light of rapid development of improved machine learning algorithms and new molecular representations such as graph-convolutional networks, we believe that this generalized structure-based approach will help to rapidly accelerate the adoption of machine learning based prediction models in molecular chemistry.

References

1. Davies, I. W. The digitization of organic synthesis. *Nature* **570**, 175–181 (2019).
2. Markó, I. E. The Art of Total Synthesis. *Science* **294**, 1842–1843 (2001).
3. Wender, P. A. & Miller, B. L. Synthesis at the molecular frontier. *Nature* **460**, 197–201 (2009).
4. Sigman, M. S., Harper, K. C., Bess, E. N. & Milo, A. The Development of Multidimensional Analysis Tools for Asymmetric Catalysis and Beyond. *Acc. Chem. Res.* **49**, 1292−1301 (2016).
5. Denmark, S. E., Gould, N. D. & Wolf, L. M. A Systematic Investigation of Quaternary Ammonium Ions as Asymmetric Phase-Transfer Catalysts. Application of Quantitative Structure Activity/Selectivity Relationships. *J. Org. Chem.* **76**, 4337–4357 (2011).

6. Santiago, C. B., Guo, J.-Y. & Sigman, M. S. Predictive and mechanistic multivariate linear regression models for reaction development. *Chem. Sci.* **9**, 2398–2412 (2018).
7. Milo, A., Bess, E. N. & Sigman, M. S. Interrogating selectivity in catalysis using molecular vibrations. *Nature* **507**, 210–214 (2014).
8. Harper, K. C. & Sigman, M. S. Three-Dimensional Correlation of Steric and Electronic Free Energy Relationships Guides Asymmetric Propargylation. *Science* **333**, 1875–1878 (2011).
9. Milo, A., Neel, A. J., Toste, F. D. & Sigman, M. S. A data-intensive approach to mechanistic elucidation applied to chiral anion catalysis. *Science* **347**, 737–743 (2015).
10. Bess, E. N. Bischoff, A. J. & Sigman, M. S. Designer substrate library for quantitative, predictive modeling of reaction performance. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 14698–14703 (2014).
11. Jordan, M. I. & Mitchell, T. M. Machine learning: Trends, perspectives, and prospects. *Science* **349**, 255–260 (2015).
12. Lavecchia, A. Machine-learning approaches in drug discovery: methods and applications. *Drug Discov. Today* **20**, 318–331 (2015).
13. Chen, H., Engkvist, O., Wang, Y., Olivecrona, M. & Blaschke, T. The rise of deep learning in drug discovery. *Drug Discov. Today* **23**, 1241–1250 (2018).
14. Ma, J., Sheridan, R. P., Liaw, A., Dahl, G. E. & Svetnik, V. Deep Neural Nets as a Method for Quantitative Structure−Activity Relationships. *J. Chem. Inf. Model.* **55**, 263−274 (2015).
15. Coley, C. W., Green, W. H. & Jensen, K. F. Machine Learning in Computer-Aided Synthesis Planning. *Acc. Chem. Res.* **51**, 1281−1289 (2018).
16. Segler, M. H. S., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604–610 (2018).
17. Zhou, Z., Li, X. & Zare, R. N. Optimizing Chemical Reactions with Deep Reinforcement Learning. *ACS Cent. Sci.* **3**, 1337−1344 (2017).
18. Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernandez-Lobato, J. M., Sanchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P. & Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **4**, 268−276 (2018).
19. Kayala, M. A., Azencott, C.-A., Chen, J. H. & Baldi, P. Learning to Predict Chemical Reactions. *J. Chem. Inf. Model.* **51**, 2209–2222 (2011).
20. Wei, J. N., Duvenaud, D. & Aspuru-Guzik, A. Neural Networks for the Prediction of Organic Chemistry Reactions. *ACS Cent. Sci.* **2**, 725−732 (2016).
21. Coley, C. W., Barzilay, R., Jaakkola, T. S., Green, W. H. & Jensen, K. F. Prediction of Organic Reaction Outcomes Using Machine Learning. *ACS Cent. Sci.* **3**, 434−443 (2017).
22. Liu, B., Ramsundar, B., Kawthekar, P., Shi, J., Gomes, J., Luu Nguyen, Q., Ho, S., Sloane, J., Wender, P. & Pande, V. Retrosynthetic Reaction Prediction Using Neural Sequence-to-Sequence Models. *ACS Cent. Sci.* **3**, 1103–1113 (2017).
23. Raccuglia, P., Elbert, K. C., Adler, P. D. F., Falk, C., Wenny, M. B., Mollo, A., Zeller, M., Friedler, S. A., Schrier, J. & Norquist, A. J. Machine-learning-assisted materials discovery using failed experiments. *Nature* **533**, 73–76 (2016).
24. Buitrago Santanilla, A., Regalado, E. L., Pereira, T., Shevlin, M., Bateman, K., Campeau, L., Schneeweis, J., Berritt, S., Shi, Z., Nantermet, P. Liu, Y., Helmy, R., Welch, C. J., Vachal, P., Davies, I. W., Cernak, T. & Dreher, S. D. Nanomole-scale high-throughput chemistry for the synthesis of complex molecules. *Science* **347**, 49–53 (2015).
25. Bédard, A.-C., Adamo, A., Aroh, K. C., Russell, M. G., Bedermann, A. A., Torosian, J., Yue, B., Jensen, K. F. & Jamison, T. F. Reconfigurable system for automated optimization of diverse chemical reactions. *Science* **361**, 1220–1225 (2018).

26. Perera, D., Tucker, J. W., Brahmbhatt, S., Helal, C. J., Chong, A., Farrell, W., Richardson, P. & Sach, N. W. A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow. *Science* **359**, 429–434 (2018).
27. Macarron, R., Banks, M. N., Bojanic, D., Burns, D. J., Cirovic, D. A., Garyantes, T., Green, D. V. S., Hertzberg, R. P., Janzen, W. P., Paslay, J. W., Schopfer, U. & Sittampalam, G. S. Impact of high-throughput screening in biomedical research. *Nat. Rev. Drug Discov.* **10**, 188–195 (2011).
28. Awale, M., Sirockin, F., Stiefl, N. & Reymond, J.-L. Medicinal Chemistry Database GDBMedChem (2019). doi:10.26434/chemrxiv.7770809.v1
29. Granda, J. M., Donina, L., Dragone, V., Long, D.-L. & Cronin, L. Controlling an organic synthesis robot with machine learning to search for new reactivity. *Nature* **559**, 377–381 (2018).
30. Ahneman, D. T., Estrada, J. G., Lin, S., Dreher, S. D. & Doyle, A. G. Predicting reaction performance in C–N cross-coupling using machine learning. *Science* **360**, 186–190 (2018).
31. Zahrt, A. F., Henle, J. J., Rose, B. T., Wang, Y., Darrow, W. T. & Denmark, S. E. Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning. *Science* **363**, eaau5631 (2019).
32. Jensen, F. *Introduction to Computational Chemistry* (Wiley, 2017).
33. Engel, T. Basic Overview of Chemoinformatics. *J. Chem. Inf. Model.* **46**, 2267–2277 (2006).
34. Agrafiotis, D. K., Bandyopadhyay, D., Wegner, J. K. & van Vlijmen, H. Recent Advances in Chemoinformatics. *J. Chem. Inf. Model.* **47**, 1279–1293 (2007).
35. Willett, P. Chemoinformatics: a history. *Wiley Interdiscip. Rev.-Comput. Mol. Sci.* **1**, 46–56 (2011).
36. Senese, C. L., Duca, J., Pan, D., Hopfinger, A. J. & Tseng, Y. J. 4D-Fingerprints, Universal QSAR and QSPR Descriptors. *J. Chem. Inf. Comput. Sci.* **44**, 1526–1539 (2004).
37. Shahlaei, M. Descriptor Selection Methods in Quantitative Structure−Activity Relationship Studies: A Review Study. *Chem. Rev.* **113**, 8093−8103 (2013).
38. Skoraczyński, G. Predicting the outcomes of organic reactions via machine learning: are current descriptors sufficient? *Sci. Rep.* **7**, 3582 (2017).
39. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **28**, 31–36 (1988).
40. Heller, S., McNaught, A., Stein, S., Tchekhovskoi, D. & Pletnev, I. InChI - the worldwide chemical structure identifier standard. *J. Cheminformatics* **5**, 7 (2013).
41. O'Boyle, N. & Dalke, A. DeepSMILES: An Adaptation of SMILES for Use in Machine-Learning of Chemical Structures (2018). doi:10.26434/chemrxiv.7097960.v1
42. Cereto-Massagué, A., Ojeda, M. J., Valls, C., Mulero, M., Garcia-Vallvé, S. & Pujadas, G. Molecular fingerprint similarity search in virtual screening. *Methods* **71**, 58–63 (2015).
43. Melville, J. L., Burke, E. K. & Hirst, J. D. Machine Learning in Virtual Screening. *Comb. Chem. High Throughput Screen.* **12**, 332–343 (2009).
44. Venkatraman, V., Pérez-Nueno, V. I., Mavridis, L. & Ritchie, D. W. Comprehensive Comparison of Ligand-Based Virtual Screening Tools Against the DUD Data set Reveals Limitations of Current 3D Methods. *J. Chem. Inf. Model.* **50**, 2079–2093 (2010).
45. Myint, K.-Z., Wang, L., Tong, Q. & Xie, X.-Q. Molecular Fingerprint-Based Artificial Neural Networks QSAR for Ligand Biological Activity Predictions. *Mol. Pharmaceutics* **9**, 2912–2923 (2012).
46. Liu, R. & Zhou, D. Using Molecular Fingerprint as Descriptors in the QSPR Study of Lipophilicity. *J. Chem. Inf. Model.* **48**, 542–549 (2008).
47. Lo, Y.-C., Rensi, S. E., Torng, W. & Altman, R. B. Machine learning in chemoinformatics and drug discovery. *Drug Discov. Today* **23**, 1538–1546 (2018).
48. Elton, D. C., Boukouvalas, Z., Butrico, M. S., Fuge, M. D. & Chung, P. W. Applying machine learning techniques to predict the properties of energetic materials. *Sci. Rep.* **8**, 9059 (2018).

49. RDKit: Open-source chemoinformatics and machine learning; http://www.rdkit.org (03/26/2019).
50. Mitchell, J. B. O. Machine learning methods in chemoinformatics. *Wiley Interdiscip. Rev.-Comput. Mol. Sci.* **4**, 468–481 (2014).
51. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
52. 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. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
53. See the supporting information for exemplified studies using single fingerprint features.
54. Reid, J. P. & Sigman, M. S. Holistic prediction of enantioselectivity in asymmetric catalysis. *Nature* **571**, 343–348 (2019).
55. Collins, K. D. & Glorius, F. A robustness screen for the rapid assessment of chemical reactions. *Nat. Chem.* **5**, 597–601 (2013).
56. Chuang, K. V. & Keiser, M. J. Comment on “Predicting reaction performance in C–N cross-coupling using machine learning”. *Science* **362**, eaat8603 (2018).
57. Estrada, J. G., Ahneman, D. T., Sheridan, R. P., Dreher, S. D. & Doyle, A. G. Response to Comment on “Predicting reaction performance in C–N cross-coupling using machine learning”. *Science* **362**, eaat8763 (2018).
58. The authors performed the selection of the additive test sets after thorough analysis of the results and learning about the most influential parameters for the descriptor-based model.[30,57] Thus, these splits are, to some extent, unbalanced and the comparable performance of the unbiased MFF model on most of these test sets is remarkable.

Acknowledgements

Financial support by Fonds der Chemischen Industrie (Fellowship to F.S.), Deutsche Forschungsgemeinschaft / SPP2102 (F.S-K.) is gratefully acknowledged. The authors thank Tiffany Paulisch, Philipp Pflüger, Dr. Eloisa Serrano, Dr. Michael Teders, Dr. Michael J. James and Prof. Dr. Herbert Kuchen (all WWU Münster) for helpful discussions.

Author contributions

The underlying concept was developed by all authors. The software package was developed by M.K. and C.B. and coded by M.K. All data sets were prepared and evaluated by F.S., F.S.-K. and M.K. The final manuscript was prepared by all authors.

Author Information

The authors declare no competing financial interest. Correspondence and requests for materials should be addressed to F.G.

Supplementary Information

Detailed information on the structure of program package, the utilized data sets and all methods are provided in the supporting information (SI). All data sets (SMILES features + targets), including different splittings, and the software package are given as ZIP-files (Input_Data_All.zip; MFF_Tool.zip).