Weighted Reaction Fingerprints for Visualizing Reactivity Cliffs and Generality

Babak Mahjour[1], Daniel Schorin[2], Tim Cernak[*1,3]

[1]Department of Medicinal Chemistry, University of Michigan
[2]Department of Information, University of Michigan
[3]Department of Chemistry, University of Michigan

ABSTRACT: Visualization of reaction space is a critical step in improving human understanding of bulk chemical reaction data. We present weighted reaction fingerprinting, a simple method to rapidly analyze and evaluate the results of massive reaction corpuses. Weighted reaction fingerprints can be utilized to rapidly identify successful and failing conditions and systems for chemical transformations in addition to giving the user the ability to divide, parse, and query reactions with specific components. Reactions are encoded into a standardized template formed with reagent classes such as nucleophile, electrophile, catalyst, ligand, and solvent. Each reaction is converted into a typical fingerprint matrix and multiplied by a weight vector to generate the weighted reaction fingerprint. These fingerprints are fed into dimensionality reduction algorithms such as principal component analysis (PCA) or t-stochastic neighbor embedding (t-SNE) to create visualizable 2-D manifolds that reveal reaction context. We demonstrate how weighted reaction fingerprinting can identify reactivity cliffs, reveal reaction conditions with high generality, and generate regions of underrepresented reaction space in the analysis of high-throughput experimentation (HTE) campaigns and provide an online interface to create weighted reaction fingerprint manifolds directly from standardized reaction datasets.

INTRODUCTION

HTE has emerged as a valuable method to create reaction datasets that can be analyzed using statistical modeling.[1-14] The development of software tools that allow chemists to rapidly understand the results of their arrays are necessary for bridging the gap between high throughput data collection and black box data-driven models. Additionally, open-source Python scripts and web interfaces are expected to facilitate broader adoption of these tools, and HTE technology in general, by the research community. The primary goals of HTE analysis include statistical data profiling, allowing chemists to rapidly identify best, worst, and average performing reaction conditions, to understand which reaction conditions work best for certain substrate pairs, and to catalyze the generation of ideas for new experimental space to explore. Such reaction informatics provide human-interpretable analyses compared to opaque machine learning or artificial intelligence algorithms.

Indeed, machine-readable molecular representations[15-20] are critical in developing robust predictive models of chemical reactivity.[21-24] Graph representations[25,26] and molecular fingerprints[27-29] have been used for the prediction of chemical properties,[30-32] similarity searching,[33] and structure optimization. In a reaction context, embeddings of the molecules from chemoinformatic, chemometric and quantum descriptors in conjunction with reaction outcomes are used to train models to predict reactivity and elucidate mechanisms. In particular, fingerprinting methods, such as the Morgan instantiation[34] of extended connectivity fingerprints (ECFP),[29] provide a fast and computationally non-intensive method to analyzing chemical data and the influx of reported reaction data in a standardized format. In recent literature, fingerprinting methods for reactions have included the concatenation of reagent fingerprints,[35,36] feature binning fingerprints,[37] and reaction difference fingerprints[38] – all of which have been used successfully in reaction prediction tasks.

As reaction data[8,39] is deposited into centralized databases,[40] techniques to parse and interpret large reaction corpuses are being developed to allow chemists to decipher patterns of reactivity at scale. These data-driven chemical models[23,24,41-50] help chemists with common tasks such as methodology discovery and optimization,[35,51-60] reagent design,[61-64] mechanistic analysis,[54,65-79] retrosynthesis,[80-92] computer aided synthesis planning,[93-104] and

reaction prediction.[22,36,105-132] This style of chemical research represents a paradigm shift from the traditional method of browsing reaction data in a manual and ad-hoc fashion using tools such as Reaxys or Scifinder.

Herein, we demonstrate the utility of weighted reaction fingerprinting (Figure 1) – a simple abstractable method applicable to any large reaction corpus stored in a standardized format. Reaction discovery campaigns are analyzed to evaluate the results of multiplexed reaction arrays and large reaction datasets. This algorithm can be used and explored with preloaded datasets at https://fingerprints.cernaklab.com.

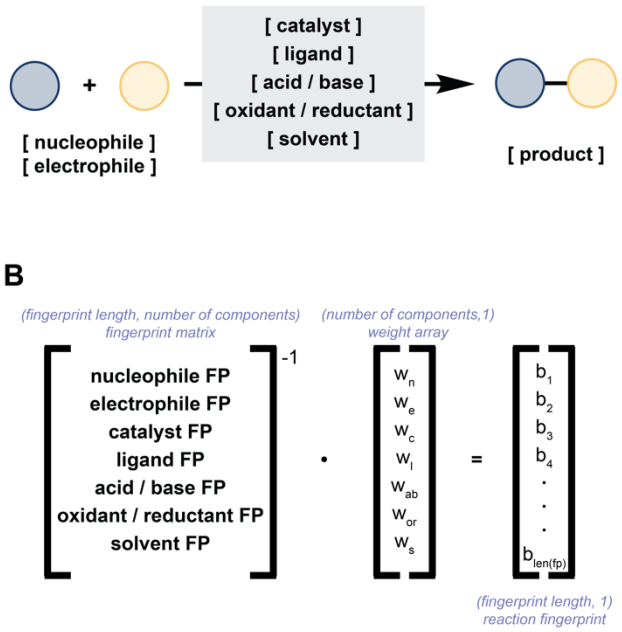

Figure 1. A) Reactions are defined by a template. In this schema, each reaction contains a nucleophile and/or an electrophile, a product, and optionally a catalyst, ligand, acid/base, oxidant/reductant, or solvent. B) The reaction fingerprint can be calculated by taking the product between a matrix of component fingerprints and a vector of weight arrays. Any fingerprint or feature vector can be utilized as long as the vector lengths for each reaction components are equal. Reaction component weights (w) are initialized at 1 and can be set to any value by the user.

Methods

We introduce reaction array fingerprinting with a dataset consisting of 2,786 Suzuki reactions. Each reaction is composed of an electrophile, nucleophile, catalyst, base, and solvent. In Figure 2, the Morgan Fingerprint with radius 4 and 2,048 bits was calculated for each substance, then summed such that the five 2,048-bit binary fingerprints for each reaction become a single 2,048-bit count reaction array fingerprint - a one-dimensional representation of the coupling. This reaction fingerprinting method contrasts with currently published methods that include reagent concatenation,[35,36] reaction feature binning,[37] and reaction difference fingerprints.[38] This array of 2,786 reaction fingerprints were fed into a t-SNE algorithm with no hyperparameter tuning to generate Figure 2A, where reaction points are colored by the product yield for the reaction – high, medium, low, and 0% - based on product/internal standard integrations. The observed clusters are chemically distinguishable, with each cluster composed of similar or identical topological fragments in the reaction mixture. Results of different dimensionality reduction algorithms such as PCA, UMAP, and SOMs as well as different fingerprint representations are shown in the Supporting Information. The six plots of Figure 2B display the same embedding shown in Figure 2A with six alternative color scales representing different features. The first plot's points are colored by the exact product/internal standard value for each reaction as calculated in the dataset. The remaining five plots are each colored by reagent per specific reagent class as defined in the template (this reaction dataset consists entirely of electrophile, nucleophile, catalyst, base, and solvent components.)

An example of a reaction cluster identified from the manifold is shown in Figure 2C. All reactions in the dataset using bromide electrophile **1** and boronate nucleophile **2** exist within this cluster. It is rapidly identified that this substrate pair was tested with 16 different catalysts and six different bases. Figure 2D showcases a simple extension of the fingerprinting algorithm using the data from this cluster. When summing the fingerprints of the individual components, a weight factor can be multiplied into a reagent's fingerprint to influence the clustering within the manifold. The 96 reactions between **1** and **2** were encoded as reaction fingerprints in two different formats. Once where the catalyst fingerprint was weighed by a factor of three, and again where the base fingerprint was weighed by a factor of three. These two datasets were then fed into the t-SNE reduction algorithm, and three plots colored by

product/internal standard integrations, catalyst, and base for each of the two datasets are displayed in Figure 2D. As revealed by the color encoded reagents, elevating the catalyst weight produces manifolds with catalyst clusters, and similar behavior is seen with base clusters when elevating the base weight.

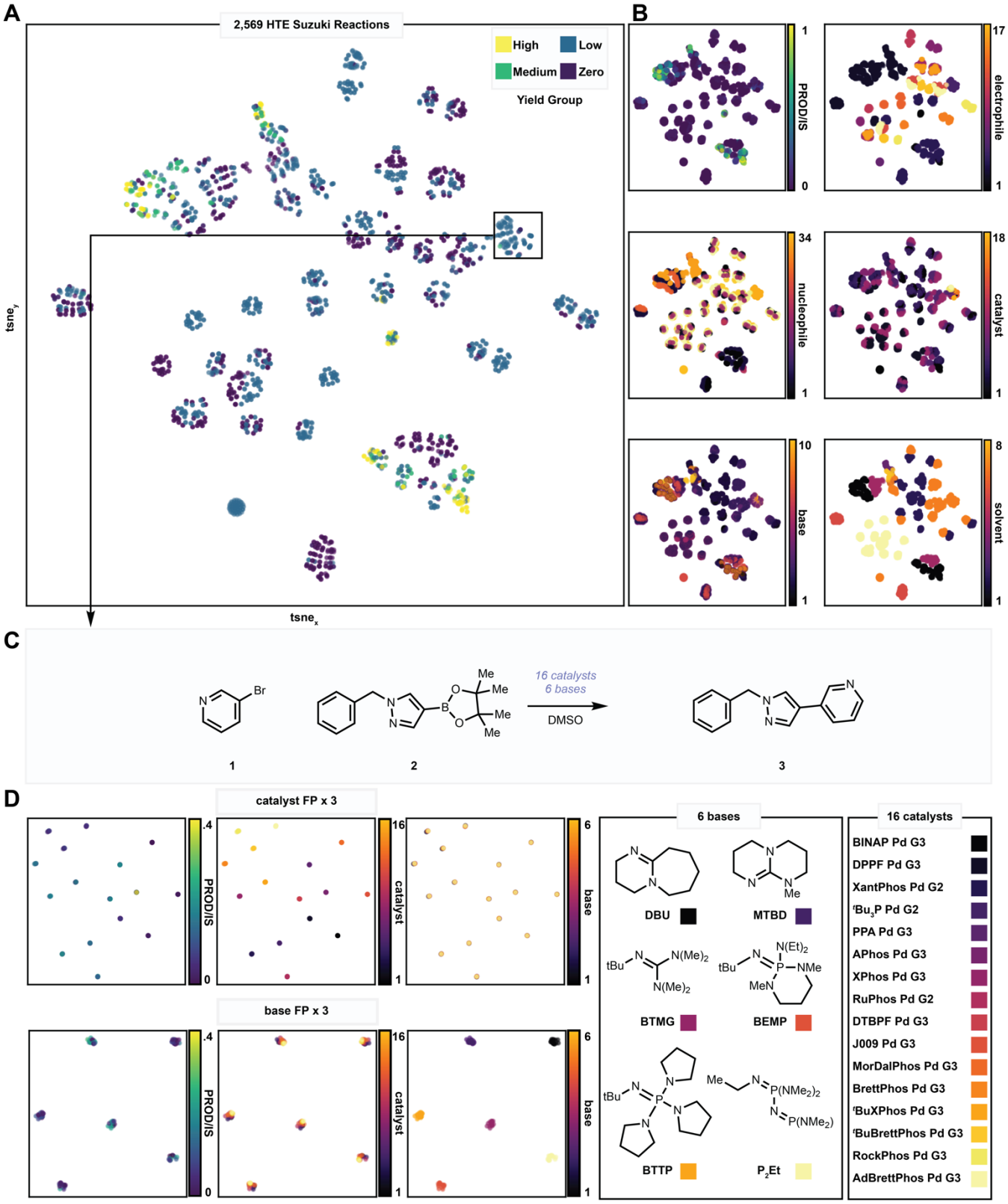

Figure 2. 2,786 Suzuki reactions performed in a high-throughput format plotted using a t-SNE trained on *reaction fingerprints*, each of which formed through the sum or concatenation of a reactant fingerprint for all components involved a particular reaction. (A) The 2,048-bit radius four Morgan Fingerprint of each reaction's electrophile, nucleophile, base, catalyst, and solvent were summed, plotted, and colored by yield group. (B) The same manifold colored by exact product/internal standard values and reagent classes (electrophile, nucleophile, catalyst, base, solvent.) (C) A specific reaction cluster containing all reactions between substrates **1** and **2**. (D) t-SNE of the reactions in the selected cluster when the catalyst fingerprint is multiplied by three (top) and when the base fingerprint is multiplied by three (bottom). The three manifolds for each result are colored by product/internal standard values, catalyst, and base.

Results and Discussion

Figure 3 demonstrates the ability to control the reaction landscape by showcasing a study where the line of chemist reasoning is visualized from the campaign discovery of a sp^3–sp^3 deaminative–decarboxylative carbon–carbon cross coupling reaction. In this study, the fingerprints of all acid electrophiles are multiplied by 3 (w_e = 3) before summing with the fingerprints of the other reagents. This trivial modification will result in the dimensionality reduction algorithm making clusters of reactions containing the substance with the elevated fingerprint. As shown in Figure 3, with the acid electrophiles having an elevated weight, the clusters are cleanly divided into reactions with different acids components. This creates a distinct landscape that separates periods of methodology optimization and substrate scope exploration. In this case, NHPI activated N-Boc proline **5** was used as a model substrate to develop a sp^3–sp^3 deaminative–decarboxylative carbon–carbon cross coupling reaction. Initially the free acid was used to develop the reactivity. For this reason, most of the reactions in this dataset fall within the NHPI activated (**5**) and free acid (**4**) N-Boc proline clusters colored in black and yellow respectively. In the PROD/IS manifold shown in Figure 3B, a direct path can be followed as experiments drive the product output from 0% to close to 100% within the NHPI activated N-Boc proline cluster (circled and in yellow). The discovery campaign began with using the free acid **4** as the model substrate (Figure 3B - location 1) but after a limit to the reactivity was realized, efforts moved to the NHPI activated acid **5** (location 2). Once ideal conditions were developed for N-Boc proline, this reaction system was tested with a variety of other acid electrophile substrates (**6**-**26**, location 3). The efficacy of this system on these substrates is shown in the t-SNE with clusters of various acid electrophile substrates forming in different locations, each with their own PROD/IS distributions. Average yields of each acid electrophile tested are shown below the plots in Figure 3 as well as the number of reactions they were tested in.

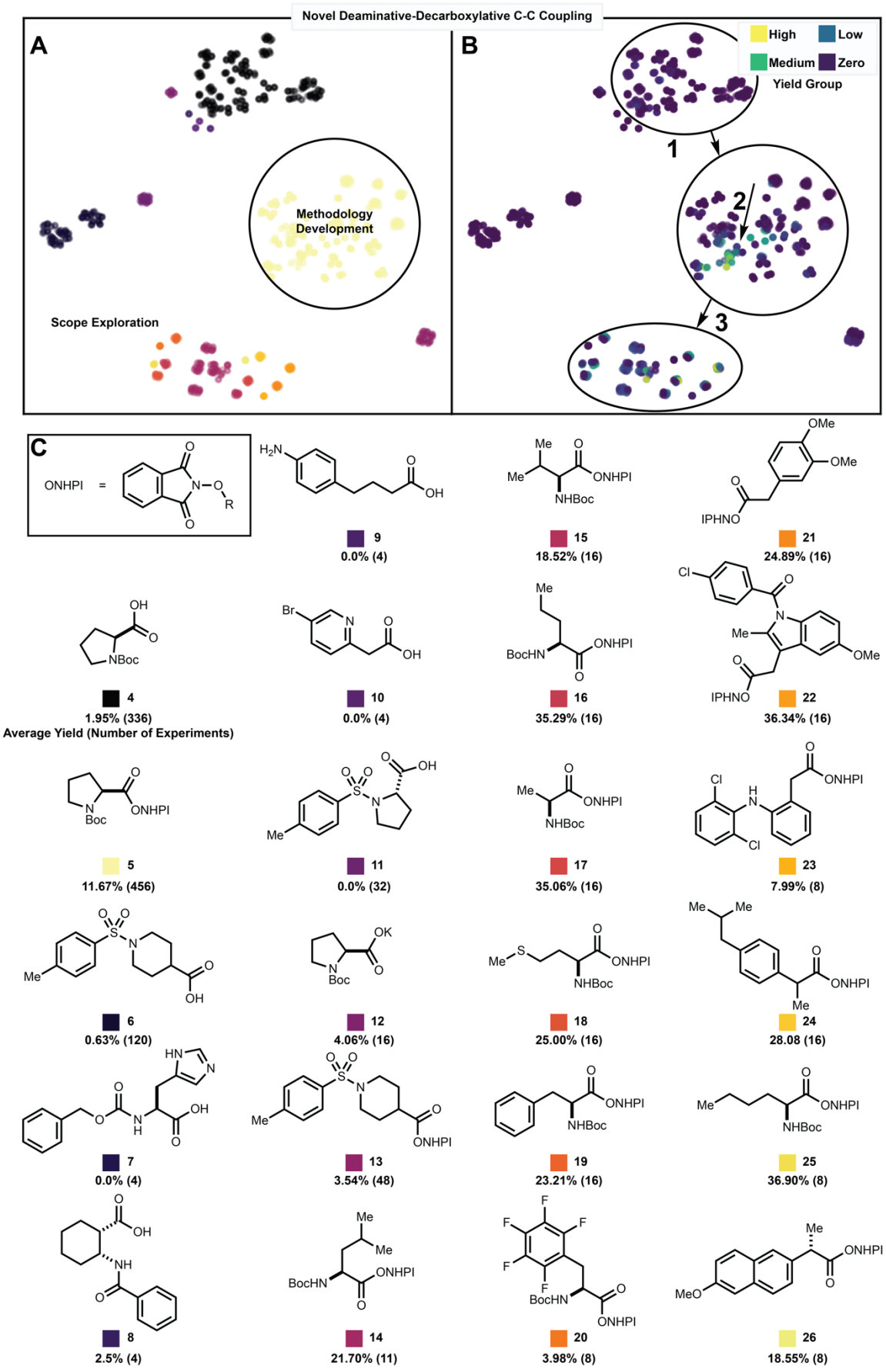

Figure 3. 1,296 reactions performed in an HTE format towards the development of a novel sp^3-sp^3 deaminative-decarboxylative carbon-carbon cross coupling. When elevating the weight of a substrate,

the manifold can be cleanly split between areas of initial methodology development and substrate scope exploration (left). In this case, the activated acid electrophile NHPI ester N-Boc-proline was used as the model substrate for the reaction, resulting in a large cluster with few substrates and many conditions. Once ideal conditions were found, they were tested on a variety of other acids, creating a tight cluster of many electrophile acids and few conditions.

In Figure 4, a case study analyzing the Suzuki dataset reveals *reactivity cliffs* – clusters that contain some reactions that fail and some that are successful, indicating minor changes to the system that cause the reaction to "flip" on or off.[133] An electrophile nucleophile substrate pair is identified to have a reactivity cliff in Figure 4A. The plots in Figure 4A are reaction array fingerprints only containing the fingerprints for the electrophile and nucleophile to emphasis focus on the substrate flexibility of the Suzuki reaction as opposed to the specifics of the reagents. On further analysis of this cluster with reactions that use electrophile **27** and nucleophile **28** to form **29**, it is revealed that this substrate pair was tested with four different solvents, six bases, and eight catalysts. In Figure 4B, we plot the reaction array fingerprints for these reactions, producing clean clusters that separate all components and producing a humanly interpretable explanation of the behavior behind the reactivity. Since experimental chemistry is rife with reactivity cliffs, oftentimes as subtle as a switch in solvent, atmosphere or even order of reagent addition, it is critical to be able to visualize and interpret this behavior. From the color-coded solvent plot, it is clear that the reactivity of the substrate pair **28** and **29** is controlled by the solvent system used. The manifold directly identifies failing and working solvents when traversing the space from 1:9 ^{t}AmOH:NMP to 1:3 water:NMP. The addition of water in the solvent system is found to be critical in achieving desired reactivity as shown by another reactivity cliff between 1:3 water:NMP and pure NMP. A chi-squared analysis is reported in the Supporting Information to validate this finding. Thus, it can be rapidly identified which solvent systems poison the reaction even though a variety of reagents are being changed.

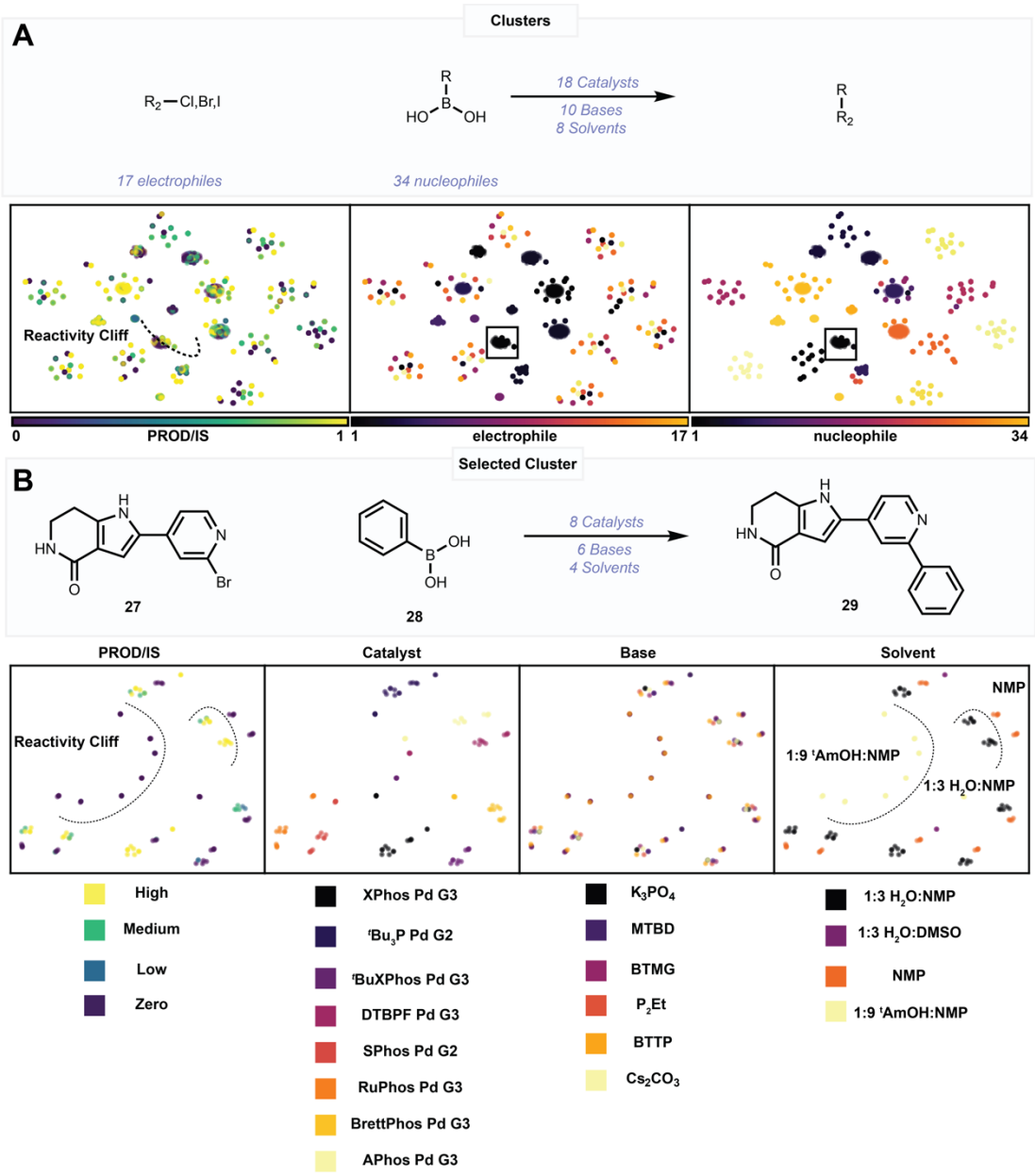

Figure 4. Reactivity cliffs are identified when focusing on specific substrate pairs. (A) Compounds **27** and **28** cluster together with high and poor performing reactions. (B) Repeating the analysis on the cluster that form **29** from **27** and **28** reveals reagents that cause this reactivity to flip on. Reactivity cliffs are readily identified between solvent regimes. The 1:3 water:NMP regime sits between two reactivity cliffs, separating it from the two failing solvent systems 1:9 ^{t}AmOH:NMP and pure NMP.

The creation of reaction array fingerprint manifolds is automatable. When using a HTE management system such as phactor™ or obtaining machine readable reaction data in a standardized format, reaction arrays can be rapidly analyzed using this method. In Figure 5, we show the automated creation of six reaction fingerprint t-SNEs utilizing output files procedurally generated from phactor™. Hyperparameters including t-SNE perplexity and reagent weights (w_x) were optimized to best illuminate reactivity trends (see Supporting Information). In all cases, specific reagent classes were clustered. For instance, there are four clusters in Figure 5A representing the four ligands used in the reaction array which couples **30** and **14** to form **31**. These clusters are split into two subclusters, each representing one of the two reductants used. These clusters are then further split into three column-shaped clusters

that represent the three catalysts and each of these columns consist of four points each corresponding to one of the four additives in the reaction array. Four clusters were formed in Figure 5B coupled 12 electrophile acids and 8 nucleophile amines, producing reaction hit that uses **32** and **5** to form **33**. Each of the 12 acids formed a cluster in the manifold. In Figure 5C eight ligands and 12 catalysts were used to make **36** from **34** and **35**. The t-SNE clustered each reaction by ligand, resulting in eight clusters. In substrate scope ultraHTE experiments shown in Figures 5D and 5E, the t-SNEs are clustered by the number of diazoniums used in the screen, eight and four respectively. In Figure 5D the reaction where **39** was formed from **37** and **38** is highlighted in the manifold, and the reaction where ester **42** is formed from **40** and **41** is highlighted in Figure 5E. Finally, in the ultraHTE direct-to-biology assay shown in Figure 5F, a cluster is formed for each of the 80 amines used in the reaction array. The reaction which amide inhibitor **45** is generated from the coupling of **43** and **44** is identified in the t-SNE.

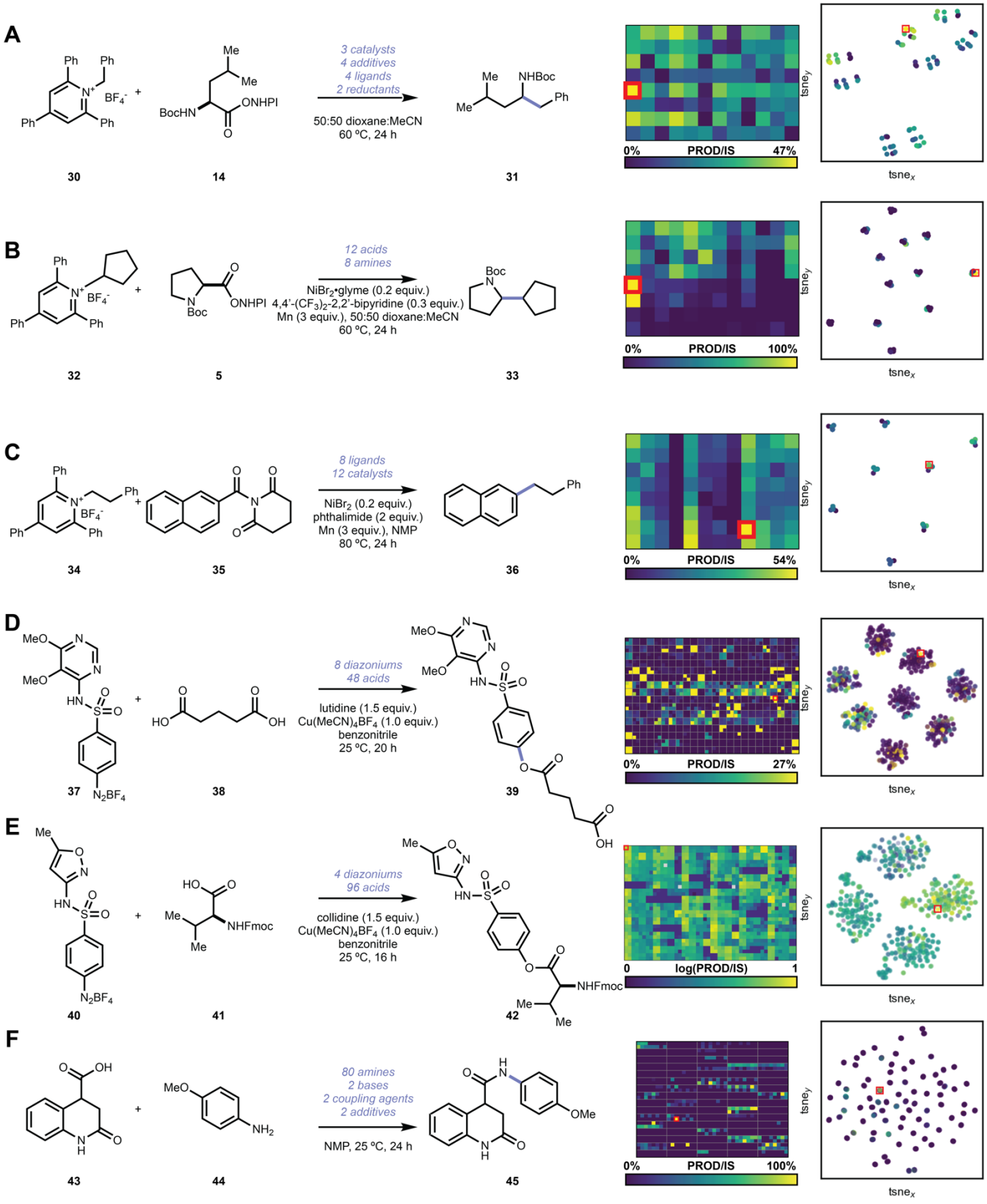

Figure 5. Reaction array manifolds can be procedurally generated from the output files produced by phactor™. (A)-(F) reaction array results as reported in ref. 8 and their corresponding manifolds colored by output value. Perplexity and weights were modified as described in the Supporting Information to optimize the latent space for visualization.

Reaction array fingerprint manifolds can be created in the context of existing inventories containing reagents libraries. This allows for the expansion of the reaction space into hypothetical unperformed reactions. Performed reaction reagents are extracted and enumerated against the library to generate the new reaction space overlaid with previously performed reactions. In Figure 6, the HTE dataset consisting of sp^3–sp^3 deaminative-decarboxylative carbon–carbon cross coupling reaction points are injected with a library of carboxylic acids. In this case, a principal component analysis was used to embed the reactions to maximize topological relevance between points. Reaction conditions are enumerated against the new acids to create a manifold of reactions with new untested substrates. This embedding can be sampled using a variety of acquisition functions to rapidly design reaction arrays. This method can be utilized to explore substrates and reagents to expand scope or optimize reaction conditions, respectively.

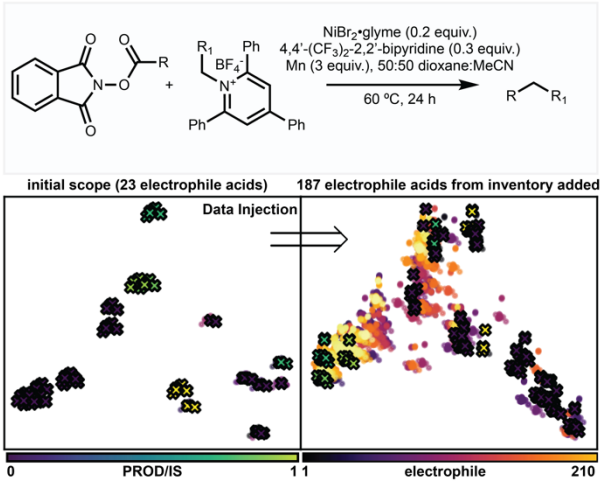

Figure 6. Full enumerated scope of substrates (all acids versus all amines) when considering acids and amines used in the CC study (left). Most combinations were performed in the study, but few were not (circles without an 'x'). Addition of our chemical inventory consisting of 187 acids not used in the study into the model expands the projection of the substrate scope. These manipulatable manifolds containing both enumerated hypothetical and performed reactions lay the groundwork for novel sampling algorithms to enhance both drug development and reaction methodology optimization. Points are colored by the respective acid used in the actual or hypothetical reaction.

Conclusion

Weighted reaction fingerprinting is a powerful and easy-to-perform method for the analysis of massive reaction datasets. The algorithm is chemically interpretable and allows chemists to rapidly understand and navigate through large collections of rapid data. Manifolds can be easily optimized and modified by changing the embedding algorithm, its hyperparameters, or the reagent weights. The algorithm is comparable with files and datasets automatically generated by procedural workflow managers such as phactor™ or from reaction databases such as the ORD. Datasets from different reaction arrays can be merged and used in the analysis and chemical inventories can be incorporated to generate experimental space. A web interface is provided to facilitate the adoption of this technology by the community and to assist in the analysis of bulk reaction data.

Supporting Information

Corresponding Author

* Tim Cernak – Department of Medicinal Chemistry, College of Pharmacy, University of Michigan, Ann Arbor, MI, USA 48104; orcid.org/0000-0001-5407-0643.

Email: tcernak@med.umich.edu

Author Contributions

Funding Sources

Notes

T.C. holds equity in Scorpion Therapeutics, and is a co-Founder and equity holder of Entos, Inc.

Acknowledgements

References

1 Cernak, T. *et al.* Microscale high-throughput experimentation as an enabling technology in drug discovery: Application in the discovery of (Piperidinyl) pyridinyl-1 H-benzimidazole diacylglycerol acyltransferase 1 inhibitors. *Journal of medicinal chemistry* **60**, 3594-3605 (2017).

2 Douthwaite, J. L. *et al.* The Formal Cross-Coupling of Amines and

Carboxylic Acids to Form sp3–sp2 Carbon–Carbon Bonds. (2022).
3 Gesmundo, N. *et al.* Miniaturization of Popular Reactions from the Medicinal Chemists' Toolbox for Ultrahigh-Throughput Experimentation. (2022).
4 Kutchukian, P. S. *et al.* Chemistry informer libraries: a chemoinformatics enabled approach to evaluate and advance synthetic methods. *Chemical science* **7**, 2604-2613 (2016).
5 Lin, S. *et al.* Mapping the dark space of chemical reactions with extended nanomole synthesis and MALDI-TOF MS. *Science* **361**, eaar6236 (2018).
6 Mahjour, B., Shen, Y., Liu, W. & Cernak, T. A map of the amine–carboxylic acid coupling system. *Nature* **580**, 71-75 (2020). https://doi.org:10.1038/s41586-020-2142-y
7 Mahjour, B., Shen, Y. & Cernak, T. Ultrahigh-throughput experimentation for information-rich chemical synthesis. *Accounts of Chemical Research* **54**, 2337-2346 (2021).
8 Mahjour, B. *et al.* Rapid Planning and Analysis of High-Throughput Experiment Arrays for Reaction Discovery. (2022).
9 McGrath, A., Zhang, R., Shafiq, K. & Cernak, T. Repurposing amine and carboxylic acid building blocks with an automatable esterification reaction. (2022).
10 Buitrago Santanilla, A. *et al.* Nanomole-scale high-throughput chemistry for the synthesis of complex molecules. *Science* **347**, 49 (2015). https://doi.org:10.1126/science.1259203
11 Shen, Y., Mahjour, B. & Cernak, T. Development of copper-catalyzed deaminative esterification using high-throughput experimentation. *Communications Chemistry* **5**, 83 (2022). https://doi.org:10.1038/s42004-022-00698-0
12 Uehling, M. R., King, R. P., Krska, S. W., Cernak, T. & Buchwald, S. L. Pharmaceutical diversification via palladium oxidative addition complexes. *Science* **363**, 405 (2019). https://doi.org:10.1126/science.aac6153
13 Zhang, Z. & Cernak, T. The Formal Cross-Coupling of Amines and Carboxylic Acids to Form sp3–sp3 Carbon–Carbon Bonds. *Angewandte Chemie International Edition* **60**, 27293-27298 (2021). https://doi.org:https://doi.org/10.1002/anie.202112454
14 Gesmundo, N. J. *et al.* Nanoscale synthesis and affinity ranking. *Nature* **557**, 228-232 (2018). https://doi.org:10.1038/s41586-018-0056-8
15 Krenn, M., Häse, F., Nigam, A., Friederich, P. & Aspuru-Guzik, A. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Machine Learning: Science and Technology* **1**, 045024 (2020). https://doi.org:10.1088/2632-2153/aba947
16 Mahjour, B., Bench, J., Zhang, R., Frazier, J. & Cernak, T. Molecular sonification for molecule to music information transfer. *Available at SSRN 4066810* (2022).
17 O'Boyle, N. & Dalke, A. DeepSMILES: an adaptation of SMILES for use in machine-learning of chemical structures. (2018).
18 Senese, C. L., Duca, J., Pan, D., Hopfinger, A. J. & Tseng, Y. J. 4D-fingerprints, universal QSAR and QSPR descriptors. *Journal of chemical information and computer sciences* **44**, 1526-1539 (2004).
19 Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Modeling* **28**, 31-36 (1988). https://doi.org:10.1021/ci00057a005
20 Xu, L.-C. *et al.* A Molecular Stereostructure Descriptor Based On Spherical Projection. *Synlett* **32**, 1837-1842 (2021).
21 Chen, Y. *et al.* Electro-Descriptors for the Performance Prediction of Electro-Organic Synthesis. *Angewandte Chemie International Edition* **60**, 4199-4207 (2021).
22 Beker, W., Gajewska, E. P., Badowski, T. & Grzybowski, B. A. Prediction of Major Regio-, Site-, and Diastereoisomers in Diels–Alder Reactions by Using Machine-Learning: The Importance of Physically Meaningful Descriptors. *Angewandte*

Chemie International Edition **58**, 4515-4519 (2019).
23 Schneider, N., Lowe, D. M., Sayle, R. A. & Landrum, G. A. Development of a Novel Fingerprint for Chemical Reactions and Its Application to Large-Scale Reaction Classification and Similarity. *Journal of Chemical Information and Modeling* **55**, 39-53 (2015). https://doi.org:10.1021/ci5006614
24 Xie, L., Xu, L., Kong, R., Chang, S. & Xu, X. Improvement of prediction performance with conjoint molecular fingerprint in deep learning. *Frontiers in pharmacology* **11**, 606668 (2020).
25 Duvenaud, D. K. *et al.* Convolutional networks on graphs for learning molecular fingerprints. *Advances in neural information processing systems* **28** (2015).
26 Zhang, R., Mahjour, B. & Cernak, T. Exploring the Combinatorial Explosion of Amine–Acid Reaction Space via Graph Editing. (2022).
27 Graziano, G. Fingerprints of molecular reactivity. *Nature Reviews Chemistry* **4**, 227-227 (2020).
28 Pattanaik, L. & Coley, C. W. Molecular Representation: Going Long on Fingerprints. *Chem* **6**, 1204-1207 (2020). https://doi.org:https://doi.org/10.1016/j.chempr.2020.05.002
29 Rogers, D. & Hahn, M. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* **50**, 742-754 (2010). https://doi.org:10.1021/ci100050t
30 Liu, R. & Zhou, D. Using molecular fingerprint as descriptors in the QSPR study of lipophilicity. *Journal of chemical information and modeling* **48**, 542-549 (2008).
31 Myint, K.-Z., Wang, L., Tong, Q. & Xie, X.-Q. Molecular fingerprint-based artificial neural networks QSAR for ligand biological activity predictions. *Molecular pharmaceutics* **9**, 2912-2923 (2012).
32 Yang, K. *et al.* Analyzing Learned Molecular Representations for Property Prediction. *Journal of Chemical Information and Modeling* **59**, 3370-3388 (2019). https://doi.org:10.1021/acs.jcim.9b00237
33 Cereto-Massagué, A. *et al.* Molecular fingerprint similarity search in virtual screening. *Methods* **71**, 58-63 (2015).
34 Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *Journal of Chemical Documentation* **5**, 107-113 (1965). https://doi.org:10.1021/c160017a018
35 Shields, B. J. *et al.* Bayesian reaction optimization as a tool for chemical synthesis. *Nature* **590**, 89-96 (2021).
36 Sandfort, F., Strieth-Kalthoff, F., Kühnemund, M., Beecks, C. & Glorius, F. A structure-based platform for predicting chemical reactivity. *Chem* **6**, 1379-1390 (2020).
37 ChemAxon. *Reaction fingerprints*, <https://docs.chemaxon.com/display/docs/reaction-fingerprint-rf.md> (2022).
38 Probst, D., Schwaller, P. & Reymond, J.-L. Reaction classification and yield prediction using the differential reaction fingerprint DRFP. *Digital discovery* **1**, 91-97 (2022).
39 Roch, L. M. *et al.* ChemOS: An orchestration software to democratize autonomous discovery. *PLOS ONE* **15**, e0229862 (2020). https://doi.org:10.1371/journal.pone.0229862
40 Kearnes, S. M. *et al.* The open reaction database. *Journal of the American Chemical Society* **143**, 18820-18826 (2021).
41 Guo, J. *et al.* Automated chemical reaction extraction from scientific literature. *Journal of Chemical Information and Modeling* **62**, 2035-2045 (2021).
42 Madzhidov, T. I. *et al.* Machine learning modelling of chemical reaction characteristics: yesterday, today, tomorrow. *Mendeleev Communications* **31**, 769-780 (2021).
43 Mitchell, J. B. O. Machine learning methods in chemoinformatics. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **4**, 468-481 (2014).
44 Pflüger, P. M. & Glorius, F. Molecular machine learning: the future of synthetic chemistry? *Angewandte Chemie*

International Edition **59**, 18860-18865 (2020).

45 Pinheiro, G. A. *et al.* Machine learning prediction of nine molecular properties based on the SMILES representation of the QM9 quantum-chemistry dataset. *The Journal of Physical Chemistry A* **124**, 9854-9866 (2020).

46 Strieth-Kalthoff, F., Sandfort, F., Segler, M. H. S. & Glorius, F. Machine learning the ropes: principles, applications and directions in synthetic chemistry. *Chemical Society Reviews* **49**, 6154-6168 (2020).

47 Strieth-Kalthoff, F. *et al.* Machine learning for chemical reactivity: The importance of failed experiments. *Angewandte Chemie International Edition* **61**, e202204647 (2022).

48 Toniato, A., Schwaller, P., Cardinale, A., Geluykens, J. & Laino, T. Unassisted noise reduction of chemical reaction datasets. *Nature Machine Intelligence* **3**, 485-494 (2021).

49 Williams, W. L. *et al.* The evolution of data-driven modeling in organic chemistry. *ACS central science* **7**, 1622-1637 (2021).

50 Weber, J. M., Guo, Z., Zhang, C., Schweidtmann, A. M. & Lapkin, A. A. Chemical data intelligence for sustainable chemistry. *Chemical Society Reviews* (2021).

51 Torres, J. A. G. *et al.* A Multi-Objective Active Learning Platform and Web App for Reaction Optimization. *Journal of the American Chemical Society* **144**, 19999-20007 (2022).

52 Baumgartner, L. M., Coley, C. W., Reizman, B. J., Gao, K. W. & Jensen, K. F. Optimum catalyst selection over continuous and discrete process variables with a single droplet microfluidic reaction platform. *Reaction Chemistry & Engineering* **3**, 301-311 (2018).

53 Christensen, M. *et al.* Data-science driven autonomous process optimization. *Communications Chemistry* **4**, 1-12 (2021).

54 De Jesus Silva, J. *et al.* Development and Molecular Understanding of a Pd-Catalyzed Cyanation of Aryl Boronic Acids Enabled by High-Throughput Experimentation and Data Analysis. *Helvetica Chimica Acta* **104**, e2100200 (2021).

55 Dotson, J. *et al.* Data-driven multi-objective optimization tactics for catalytic asymmetric reactions. (2022).

56 Kariofillis, S. K. *et al.* Using Data Science To Guide Aryl Bromide Substrate Scope Analysis in a Ni/Photoredox-Catalyzed Cross-Coupling with Acetals as Alcohol-Derived Radical Sources. *Journal of the American Chemical Society* **144**, 1045-1055 (2022).

57 Nielsen, M. K., Ahneman, D. T., Riera, O. & Doyle, A. G. Deoxyfluorination with sulfonyl fluorides: navigating reaction space with machine learning. *Journal of the American Chemical Society* **140**, 5004-5008 (2018).

58 Shim, E. *et al.* Predicting reaction conditions from limited data through active transfer learning. *Chemical science* **13**, 6655-6668 (2022).

59 Tu, Z., Stuyver, T. & Coley, C. W. Predictive chemistry: Machine learning for reaction deployment, reaction development, and reaction discovery. *Chemical Science* (2022).

60 Xu, L. C. *et al.* Towards Data-Driven Design of Asymmetric Hydrogenation of Olefins: Database and Hierarchical Learning. *Angewandte Chemie* **133**, 22986-22993 (2021).

61 Gensch, T. *et al.* A comprehensive discovery platform for organophosphorus ligands for catalysis. *Journal of the American Chemical Society* **144**, 1205-1217 (2022).

62 Kulik, H. J. & Sigman, M. S. Vol. 54 2335-2336 (ACS Publications, 2021).

63 Reid, J. P. & Sigman, M. S. Comparing quantitative prediction methods for the discovery of small-molecule chiral catalysts. *Nature Reviews Chemistry* **2**, 290-305 (2018).

64 Gómez-Bombarelli, R. *et al.* Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent Sci* **4**, 268-276 (2018). https://doi.org:10.1021/acscentsci.7b00572

65 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).
66 Akita, S., Guo, J.-Y., Seidel, F. W., Sigman, M. S. & Nozaki, K. Statistical Analysis of Catalytic Performance in Ethylene/Methyl Acrylate Copolymerization Using Palladium/Phosphine-Sulfonate Catalysts. *Organometallics* (2022).
67 Dotson, J. J., Anslyn, E. V. & Sigman, M. S. A Data-Driven Approach to the Development and Understanding of Chiroptical Sensors for Alcohols with Remote γ-Stereocenters. *Journal of the American Chemical Society* **143**, 19187-19198 (2021).
68 Engl, P. S. *et al.* Exploiting and understanding the selectivity of Ru-N-heterocyclic carbene metathesis catalysts for the ethenolysis of cyclic olefins to α, ω-dienes. *Journal of the American Chemical Society* **139**, 13117-13125 (2017).
69 Fitzner, M. *et al.* What can reaction databases teach us about Buchwald–Hartwig cross-couplings? *Chemical science* **11**, 13085-13093 (2020).
70 Heid, E. & Green, W. H. Machine learning of reaction properties via learned representations of the condensed graph of reaction. *Journal of chemical information and modeling* (2021).
71 Heinen, S., von Rudorff, G. F. & von Lilienfeld, O. A. Toward the design of chemical reactions: Machine learning barriers of competing mechanisms in reactant space. *The Journal of Chemical Physics* **155**, 064105 (2021).
72 Jorner, K., Tomberg, A., Bauer, C., Sköld, C. & Norrby, P.-O. Organic reactivity from mechanism to machine learning. *Nature Reviews Chemistry* **5**, 240-255 (2021).
73 Samha, M. H. *et al.* Exploring Structure–Function Relationships of Aryl Pyrrolidine-Based Hydrogen-Bond Donors in Asymmetric Catalysis Using Data-Driven Techniques. *ACS Catalysis* **12**, 14836-14845 (2022).
74 Silva, J. D. J., Ferreira, M. A. B., Fedorov, A., Sigman, M. S. & Copéret, C. Molecular-level insight in supported olefin metathesis catalysts by combining surface organometallic chemistry, high throughput experimentation, and data analysis. *Chemical Science* **11**, 6717-6723 (2020).
75 Tang, T. *et al.* Investigating Oxidative Addition Mechanisms of Allylic Electrophiles with Low-Valent Ni/Co Catalysts Using Electroanalytical and Data Science Techniques. *Journal of the American Chemical Society* **144**, 20056-20066 (2022).
76 Yada, A. *et al.* Ensemble Learning Approach with LASSO for Predicting Catalytic Reaction Rates. *Synlett* **32**, 1843-1848 (2021).
77 Yang, L.-C., Li, X., Zhang, S.-Q. & Hong, X. Machine learning prediction of hydrogen atom transfer reactivity in photoredox-mediated C–H functionalization. *Organic Chemistry Frontiers* **8**, 6187-6195 (2021).
78 Yang, L. C., Zhu, L. J., Zhang, S. Q. & Hong, X. Machine Learning Prediction of Structure-Performance Relationship in Organic Synthesis. *Chinese Journal of Chemistry* **40**, 2106-2117 (2022).
79 Crawford, J. M., Kingston, C., Toste, F. D. & Sigman, M. S. Data science meets physical organic chemistry. *Accounts of Chemical Research* **54**, 3136-3148 (2021).
80 Jiang, Y. *et al.* Artificial Intelligence for Retrosynthesis Prediction. *Engineering* (2022).
81 Lin, K., Xu, Y., Pei, J. & Lai, L. Automatic retrosynthetic route planning using template-free models. *Chemical science* **11**, 3355-3364 (2020).
82 Gao, H. *et al.* Combining retrosynthesis and mixed-integer optimization for minimizing the chemical inventory needed to realize a WHO essential medicines list. *Reaction Chemistry & Engineering* **5**, 367-376 (2020).
83 Coley, C. W., Rogers, L., Green, W. H. & Jensen, K. F. Computer-assisted retrosynthesis based on molecular similarity. *ACS central science* **3**, 1237-1245 (2017).
84 Dai, H., Li, C., Coley, C., Dai, B. & Song, L. Retrosynthesis prediction with conditional graph logic network. *Advances in Neural Information Processing Systems* **32** (2019).
85 Fortunato, M., Coley, C., Barnes, B. & Jensen, K. Data Augmentation and Pre-training for Template-Based

Retrosynthetic Prediction. *Bulletin of the American Physical Society* **65** (2020).
86 Lin, M. H., Tu, Z. & Coley, C. W. Improving the performance of models for one-step retrosynthesis through re-ranking. *Journal of cheminformatics* **14**, 1-13 (2022).
87 Liu, B. *et al.* Retrosynthetic reaction prediction using neural sequence-to-sequence models. *ACS central science* **3**, 1103-1113 (2017).
88 Mo, Y. *et al.* Evaluating and clustering retrosynthesis pathways with learned strategy. *Chemical science* **12**, 1469-1478 (2021).
89 Schreck, J. S., Coley, C. W. & Bishop, K. J. M. Learning retrosynthetic planning through simulated experience. *ACS central science* **5**, 970-981 (2019).
90 Schwaller, P. *et al.* Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chemical science* **11**, 3316-3325 (2020).
91 Somnath, V. R., Bunne, C., Coley, C., Krause, A. & Barzilay, R. Learning graph models for retrosynthesis prediction. *Advances in Neural Information Processing Systems* **34**, 9405-9415 (2021).
92 Tu, Z. & Coley, C. W. Permutation invariant graph-to-sequence model for template-free retrosynthesis and reaction prediction. *Journal of chemical information and modeling* **62**, 3503-3513 (2022).
93 Shen, Y. *et al.* Automation and computer-assisted planning for chemical synthesis. *Nature Reviews Methods Primers* **1**, 1-23 (2021).
94 Gao, W., Raghavan, P. & Coley, C. W. Autonomous platforms for data-driven organic synthesis. *Nature Communications* **13**, 1-4 (2022).
95 Coley, C. W., Green, W. H. & Jensen, K. F. Machine Learning in Computer-Aided Synthesis Planning. *Accounts of Chemical Research* **51**, 1281-1289 (2018). https://doi.org:10.1021/acs.accounts.8b00087
96 Coley, C. W. *et al.* A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science* **365**, eaax1566 (2019). https://doi.org:10.1126/science.aax1566
97 Gong, Y., Xue, D., Chuai, G., Yu, J. & Liu, Q. DeepReac+: deep active learning for quantitative modeling of organic chemical reactions. *Chemical science* **12**, 14459-14472 (2021).
98 Kelly, S. P. *et al.* Data Science-Driven Analysis of Substrate-Permissive Diketopiperazine Reverse Prenyltransferase NotF: Applications in Protein Engineering and Cascade Biocatalytic Synthesis of (−)-Eurotiumin A. *Journal of the American Chemical Society* **144**, 19326-19336 (2022).
99 Satoh, H. & Funatsu, K. SOPHIA, a Knowledge Base-Guided Reaction Prediction System - Utilization of a Knowledge Base Derived from a Reaction Database. *Journal of Chemical Information and Computer Sciences* **35**, 34-44 (1995). https://doi.org:10.1021/ci00023a005
100 Schwaller, P. *et al.* Mapping the space of chemical reactions using attention-based neural networks. *Nature Machine Intelligence* **3**, 144-152 (2021).
101 Segler, M. H. S. & Waller, M. P. Modelling chemical reasoning to predict and invent reactions. *Chemistry–A European Journal* **23**, 6118-6128 (2017).
102 Segler, M. H. S., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604-610 (2018).
103 Szymkuć, S. *et al.* Computer-assisted synthetic planning: the end of the beginning. *Angewandte Chemie International Edition* **55**, 5904-5937 (2016).
104 Coley, C. W. Defining and Exploring Chemical Spaces. *Trends in Chemistry* **3**, 133-145 (2021). https://doi.org:https://doi.org/10.1016/j.trechm.2020.11.004
105 Coley, C. W. *et al.* A graph-convolutional neural network model for the prediction of chemical reactivity. *Chemical science* **10**, 370-377 (2019).
106 Yarish, D. *et al.* Advancing molecular graphs with descriptors for the prediction of chemical reaction yields. *Journal of Computational Chemistry* (2022).
107 Ertl, P. *et al.* Chemical Reactivity Prediction: Current Methods and

Different Application Areas. *Molecular informatics* **41**, 2100277 (2022).
108 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).
109 Bess, E. N., Bischoff, A. J. & Sigman, M. S. Designer substrate library for quantitative, predictive modeling of reaction performance. *Proceedings of the National Academy of Sciences* **111**, 14698-14703 (2014).
110 Boni, Y. T., Cammarota, R. C., Liao, K., Sigman, M. S. & Davies, H. M. L. Leveraging Regio-and Stereoselective C (sp3)–H Functionalization of Silyl Ethers to Train a Logistic Regression Classification Model for Predicting Site-Selectivity Bias. *Journal of the American Chemical Society* **144**, 15549-15561 (2022).
111 Coley, C. W., Barzilay, R., Jaakkola, T. S., Green, W. H. & Jensen, K. F. Prediction of organic reaction outcomes using machine learning. *ACS central science* **3**, 434-443 (2017).
112 Coley, C. W., Rogers, L., Green, W. H. & Jensen, K. F. SCScore: Synthetic Complexity Learned from a Reaction Corpus. *Journal of Chemical Information and Modeling* **58**, 252-261 (2018). https://doi.org:10.1021/acs.jcim.7b00622
113 Gallarati, S. *et al.* Reaction-based machine learning representations for predicting the enantioselectivity of organocatalysts. *Chemical science* **12**, 6879-6889 (2021).
114 Gao, H. *et al.* Using machine learning to predict suitable conditions for organic reactions. *ACS central science* **4**, 1465-1476 (2018).
115 Gao, W. & Coley, C. W. The synthesizability of molecules proposed by generative models. *Journal of chemical information and modeling* **60**, 5714-5723 (2020).
116 Guan, Y. *et al.* Regio-selectivity prediction with a machine-learned reaction representation and on-the-fly quantum mechanical descriptors. *Chemical science* **12**, 2198-2208 (2021).
117 Haas, B. C., Goetz, A. E., Bahamonde, A., McWilliams, J. C. & Sigman, M. S. Predicting relative efficiency of amide bond formation using multivariate linear regression. *Proceedings of the National Academy of Sciences* **119**, e2118451119 (2022).
118 Häse, F., Roch, L. M., Kreisbeck, C. & Aspuru-Guzik, A. Supplementary Information Phoenics: A Bayesian Optimizer for Chemistry.
119 Haywood, A. L. *et al.* Kernel methods for predicting yields of chemical reactions. *Journal of Chemical Information and Modeling* (2021).
120 Jin, W., Coley, C., Barzilay, R. & Jaakkola, T. Predicting organic reaction outcomes with weisfeiler-lehman network. *Advances in neural information processing systems* **30** (2017).
121 Kayala, M. A., Azencott, C.-A., Chen, J. H. & Baldi, P. Learning to predict chemical reactions. *Journal of chemical information and modeling* **51**, 2209-2222 (2011).
122 Kwon, Y., Lee, D., Choi, Y.-S. & Kang, S. Uncertainty-aware prediction of chemical reaction yields with graph neural networks. *Journal of Cheminformatics* **14**, 1-10 (2022).
123 Newman-Stonebraker, S. H. *et al.* Univariate classification of phosphine ligation state and reactivity in cross-coupling catalysis. *Science* **374**, 301-308 (2021).
124 Rakhimbekova, A. *et al.* Cross-validation strategies in QSPR modelling of chemical reactions. *SAR and QSAR in Environmental Research* **32**, 207-219 (2021).
125 Reid, J. P. & Sigman, M. S. Holistic prediction of enantioselectivity in asymmetric catalysis. *Nature* **571**, 343-348 (2019).
126 Sato, A., Miyao, T. & Funatsu, K. Prediction of Reaction Yield for Buchwald-Hartwig Cross-coupling Reactions Using Deep Learning. *Molecular Informatics* **41**, 2100156 (2022).
127 Schwaller, P. *et al.* Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Central Science* **5**, 1572-1583 (2019). https://doi.org:10.1021/acscentsci.9b00576

128 Schwaller, P., Vaucher, A. C., Laino, T. & Reymond, J.-L. Prediction of chemical reaction yields using deep learning. *Machine learning: science and technology* **2**, 015016 (2021).
129 Struble, T. J., Coley, C. W. & Jensen, K. F. Multitask prediction of site selectivity in aromatic C–H functionalization reactions. *Reaction Chemistry & Engineering* **5**, 896-902 (2020).
130 Stuyver, T. & Coley, C. W. Quantum chemistry-augmented neural networks for reactivity prediction: Performance, generalizability, and explainability. *The Journal of Chemical Physics* **156**, 084104 (2022).
131 Zhu, X. Y. *et al.* Prediction of Multicomponent Reaction Yields Using Machine Learning. *Chinese Journal of Chemistry* **39**, 3231-3237 (2021).
132 Zuranski, A. M., Martinez Alvarado, J. I., Shields, B. J. & Doyle, A. G. Predicting reaction yields via supervised learning. *Accounts of chemical research* **54**, 1856-1865 (2021).
133 Stumpfe, D. & Bajorath, J. r. Exploring activity cliffs in medicinal chemistry: miniperspective. *Journal of medicinal chemistry* **55**, 2932-2942 (2012).