JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Perspective

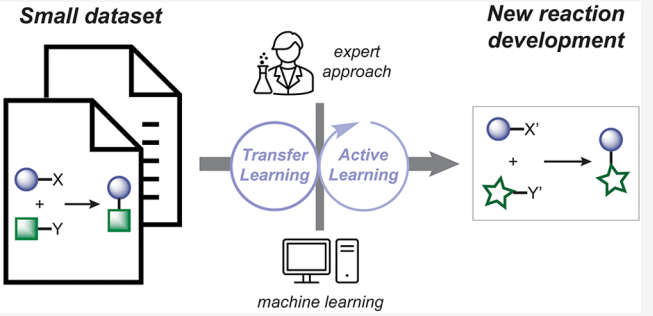
Machine Learning Strategies for Reaction Development: Toward the Low-Data Limit

Eunjae Shim, Ambuj Tewari, Tim Cernak, and Paul M. Zimmerman*

Cite This: J. Chem. Inf. Model. 2023, 63, 3659−3668 Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: Machine learning models are increasingly being utilized to predict outcomes of organic chemical reactions. A large amount of reaction data is used to train these models, which is in stark contrast to how expert chemists discover and develop new reactions by leveraging information from a small number of relevant transformations. Transfer learning and active learning are two strategies that can operate in low-data situations, which may help fill this gap and promote the use of machine learning for tackling real-world challenges in organic synthesis. This Perspective introduces active and transfer learning and connects these to potential opportunities and directions for further research, especially in the area of prospective development of chemical transformations.



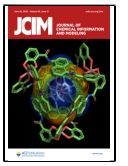
INTRODUCTION

The chemical space of molecules and reactions is rich and virtually unexplored. The number of drug-like molecules is thought to be on the order of 10^{60},[1,2] a scale greater than the mass of the universe in grams.[3] The number of plausible reaction conditions leading to these molecules also grows rapidly as key reaction components (for instance, catalysts, base, or oxidants) can be combined in various amounts. To reveal better performing molecules[4−6] and materials[7−16]—and more efficient ways of preparing them[17−23]—within this vast chemical universe, effective navigation is crucial. For ages, chemists have been constantly pushing the boundaries of chemical knowledge through hypothesis-driven experiments (Figure 1). Today, alternative strategies based on increasingly accessible data[24] and computational resources are emerging. This Perspective seeks the intersection between traditional expert strategies and computational means of exploring reaction spaces.

The incredible success of chemistry as a scientific field has been enabled by smart use of information from a variety of sources. Within the mind of a chemist, a new reaction is devised through creative processing of physically informed chemical principles and prior knowledge of related reactions. The chemist's intuition builds upon known reaction conditions derived from the literature to develop an initial set of experiments in the new space. As information from experiments becomes available, the expert's hypotheses are refined and the next set of experiments are planned. The scope and direction of the exploration, however, can be unintentionally bounded by the current body of chemical understanding, preventing consideration of potentially optimal solutions. Despite this shortcoming and the rather small data chemists are often equipped with—a few papers and manually conducted experimental data points—the traditional scientific process of chemistry undoubtably works. What can be done to bolster this traditional process and further facilitate the uncovering of physically sound, generalizable knowledge?

Machine learning has seen increased use and shown considerable promise for identifying useful chemistry. This makes sense because machine learning can perform a similar process as chemists—intake data, numerically transform it, and make predictions. This process also empowers the machine to effectively approximate problem domains. As a result, successfully trained models are capable of making quantitatively accurate chemical predictions such as reaction outcomes[25−36] or physical properties of molecules.[37−41] More importantly, when applied to iterative experimentation, this ability can quickly direct the exploration to better solutions and reduce experimentation timeframes.[7,42−45] Supplementing the conventional approach to chemistry with machine learning could therefore reinforce our ability to tackle ongoing chemical problems.

Received: April 17, 2023
Published: June 14, 2023



https://doi.org/10.1021/acs.jcim.3c00577
J. Chem. Inf. Model. 2023, 63, 3659−3668

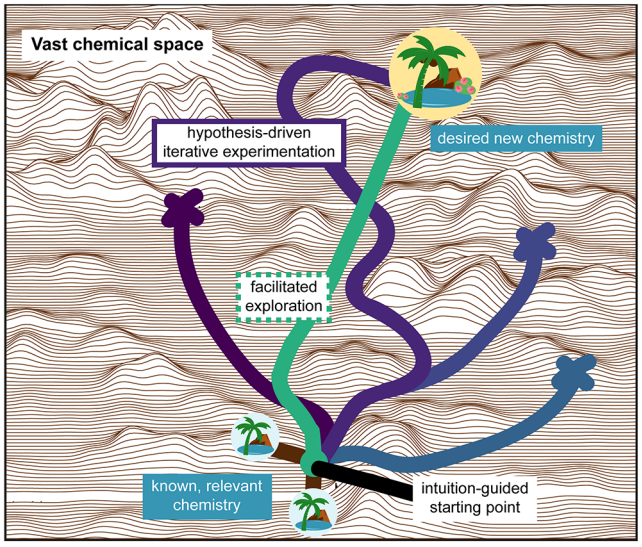


Figure 1. Schematic representation of chemical research. Within the vast chemical space, meaningful chemistry is sparse. To identify productive reaction routes, relevant prior results are processed by a chemist to determine the initial study. Experiments are then iteratively conducted, with results informing the subsequent exploration. While historically proven to be effective, the exploration process could be streamlined.

Better alignment between the requirements of machine learning and the reality of laboratory research would be helpful. The character of the data sets that are normally utilized clearly distinguishes the two. In particular, chemists work with a few data points that are most relevant to the problem. On the other hand, machines need orders of magnitude more data, covering a sizable fraction of the problem domain. While technologies like high-throughput experimentation[27,46−49] and flow chemistry[50] can help prepare such tailored data sets, they remain less broadly deployed compared to conventional experimentation. Another discrepancy stems from the involvement of scientific knowledge. For chemists, chemical principles are essential for making meaningful predictions in a new problem domain. On the other hand, big data sets digested by machine learning algorithms lack such generalizable concepts. In all, machine learning algorithms that operate on typical (small) chemical data sets and can integrate scientific knowledge would enhance their practicality for use in laboratories.

Incorporating the chemical intuition of expert chemists—which has long enabled successful research from small data—into machine learning algorithms presents opportunities for realizing such a goal. In this Perspective centered on organic synthesis, connections will be made between intuition- and hypothesis-driven chemical research and machine learning algorithms. An outlook on how these connections would impact prospective application is also presented.

TRANSFER LEARNING

General chemical principles combined with specific information from closely related research literature play a crucial role in elaborating early reaction exploration. For example, a previously reported reaction condition can be modified to accommodate a different functionality of a new substrate class. In this case, the new substrates are fixed, but the remaining, plausible components of the reaction are numerous. Even with a limited number of possibilities for each reaction component, the combinations grow quickly but only a limited number of reactions can be handled at once. Therefore, the possibilities must be narrowed down to a prioritized set of experiments to build capabilities in the new area.

Transfer learning[51,52] is a machine learning approach that aims to tackle such a process in a quantitative manner. In transfer learning, the aim is to use information extracted from a data set in hand (called the source domain) to achieve more efficient and effective modeling of the problem of interest (called the target domain). Ideally, the resulting model's predictions would provide a meaningful set of initial hypotheses to verify in a prioritized manner.

One popular transfer learning method is known as fine-tuning, where a model (usually deep learning model) trained on a large source data set, called a pretrained model, is refined on a smaller target data set. The potential of pretrained models is recognizable through Generative Pretrained Transformers (GPT), which are language models capable of doing a wide range of tasks through text production.[53] For synthetic chemistry, natural language processing models have been fine-tuned to specific reaction classes to predict stereospecific products[54] or yields.[55] In the former study, a transformer model[56] that predicts products, trained on approximately one million generic reactions,[57] was fine-tuned on a smaller carbohydrate chemistry data set of approximately 20,000 reactions. The resulting model's top-1 accuracy for predicting stereodefined carbohydrate products was 70%, which is an improvement of 27% and 40% from models that were trained only on the source and target data set, respectively. In addition, predictions with the highest confidence scores were mostly shown to be correct, which implies the possibility for prioritization of experiments in a prospective setting.

	Source Dataset	*Transfer Strategy*	*Target Dataset*
Current status	• single enormous database	• fine-tuning models	• relatively large curated data
Opportunities	• small number of reactions • range of a single source dataset • use of multiple sources	• use of simple models • definition of similarity between source and target • avoiding negative transfer	• non-existent or all negatives

Figure 2. Overview of the components of transfer learning (top row). Features of each element in current fine-tuning work (middle row) are compared to the realistic reaction development setting (bottom row), presenting opportunities.

https://doi.org/10.1021/acs.jcim.3c00577
J. Chem. Inf. Model. 2023, 63, 3659−3668

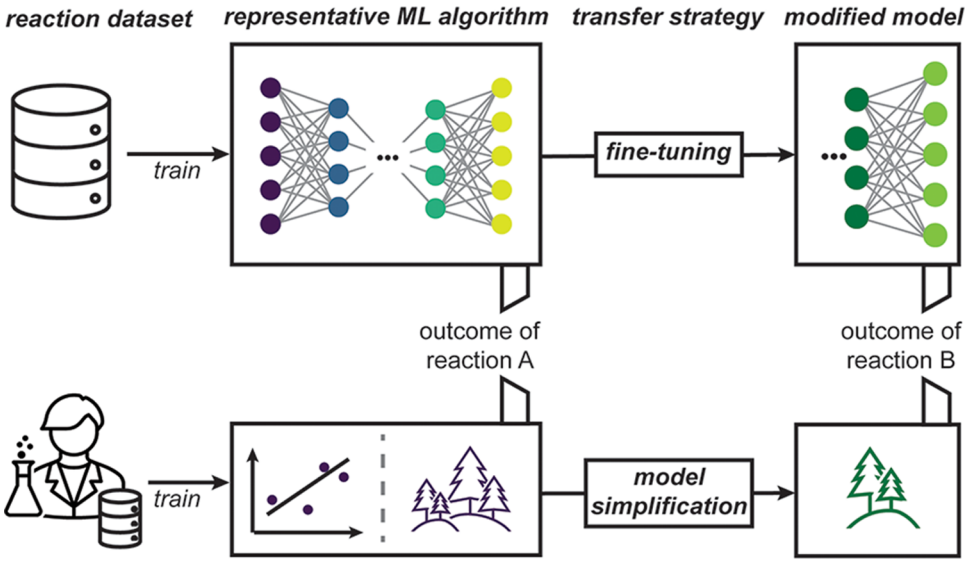


Figure 3. Mainstream machine learning algorithms often leverage large data sets (top). For the situation where data is scarce (bottom left), smaller models need to be considered. Transfer learning for these situations has not received as much attention.

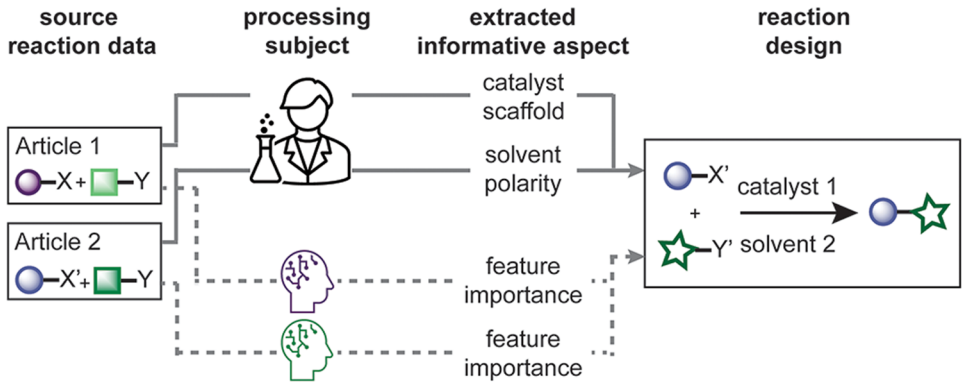


Figure 4. Comparison of multisource transfer learning by expert chemists and machine learning algorithms. Chemists often extract concise, qualitative chemical reasonings based on the literature and make appropriate adjustments toward the target reaction (along the solid gray arrows). In contrast, the black box nature of machine learning models makes it difficult to extract and modify features important to the target problem (along the dashed gray arrows).

With abundant source reaction data and a relatively smaller target data set, fine-tuning can achieve performance that seems impossible with only one of the two. Under this scenario, fine-tuning enables deep learning algorithms to provide powerful models that could not be created using the target data alone. Although public reaction databases[58,59] may serve as source data sets in a reaction development setting, in typical situations, target data does not exist (e.g., at the very beginning of the exploration) or comprises only negatives (i.e., all previous experiments were unsuccessful). In these cases, retraining the source model via small target data sets is not likely to be an effective means of updating the deep learning model. New concepts in the target reaction space may be required that are not part of the source model, limiting the utility of fine-tuning. To overcome these challenges, all components of transfer learning need to be addressed (Figure 2).

The source data set (Figure 2, left) is a crucial element of transfer learning, as it provides information that enables modeling of the target problem. For most machine learning exercises, the source data set is a large database of reactions (Figure 3, top left).[58,59] For chemists, source data sets can be conceived of at two levels. One encompasses the wide scope of literature used to formulate broad qualitative chemical concepts, for instance, reactivity trends and mechanisms. This knowledge informs each chemist's research strategy and also constitutes the core of chemical intuition. The second type of source data, which we will refer to as the source data set, is a small number of reactions—as low as a few dozen—that are selected specifically to be relevant to the reaction goal (Figure 3, bottom left). This latter situation has not received as much attention, as it does not fit into the standard practice of machine learning. For machine learning to mimic the expert chemist, new strategies for transfer learning that deal with such small, focused source data sets are needed.

Two approaches are plausible for specifying the source data set. One possibility is to combine all relevant data in hand as a single source data set, although the range of transformations to include can be ambiguous. In one example, stereoselectivity models of chiral phosphoric acid catalyzed nucleophilic additions to imines[60] were developed by combining reactions from multiple literature reports. The data set included

https://doi.org/10.1021/acs.jcim.3c00577
J. Chem. Inf. Model. 2023, 63, 3659−3668

nucleophiles that were diverse, spanning from transfer hydrogen sources to diazoacetamides. Models built with this process were able to predict stereoselectivity for three external test examples within 5% enantiomeric excess. To give another example, consider the situation where reaction data consists of a wide range of nucleophile classes (e.g., boron nucleophiles and organozinc reagents, etc.) undergoing nickel-catalyzed C−O activation reactions. For predicting yields of boron nucleophiles, one can consider either reactions in the same nucleophile class or the entire data set as the training data. In a recent retrospective study,[61] such comparisons were made from a data set compiled from the literature. The two training data schemes show comparable yield regression performance with R^2 values of 0.47 vs 0.45 and 0.54 vs 0.57 for boron and organozinc nucleophiles, respectively. The insignificant difference means that the models built from the nucleophiles of interest were not improved using the data of other nucleophiles. This showed that for reactions defined using domain specific knowledge, modest predictivity is plausible using relatively small data sets, in this case about 100 data points. On the other hand, a similar analysis on Buchwald−Hartwig C−N coupling reactions compiled from the literature suggests the opposite: using the entire data set improved the models.[62] At this point in time, recognizing *a priori* which approach is best is not a clearly defined task.

Alternatively, multiple source data sets could be utilized for transfer learning, resembling the traditional approach where a chemist's source data contains just a handful of research articles. Each work informs different aspects of the reaction such as mechanistic concepts, viable catalysts and reagents, as well as conditions like temperature or concentration. How this information fuses into a unique design of a new reaction depends on the chemist's internal model, which is informed by years of training and expertise (Figure 4, solid lines). Similarly, multiple machine learning models can be trained on different source transformations (Figure 4, dashed lines). Each of these models would quantify how different aspects of reaction components impact the reaction outcome (e.g., if larger steric bulk of the substrate results in lower yield, a negative regression coefficient for a steric descriptor would appear). It is unlikely, however, that these parameters would quantitatively transfer toward a new target reaction. On the other hand, a subset of looser, qualitative features (e.g., the fact that the coefficient of substrate steric descriptor is negative, ignoring its magnitude) may still be useful hints for target reactivity. By extracting and combining the most relevant hints from the multiple source models, an effective consensus prediction can be made. While balancing multiple viewpoints into viable reaction conditions comes naturally to expert chemists, achieving such harmonization statistically is a challenging task that is currently being investigated by the machine learning community.[63]

Expert curation of source reactions does not guarantee that the source model will show statistical predictivity for the target problem. In fact, deteriorated performance is known to be a possible outcome of transfer learning and is coined as "negative transfer".[51] Given how resource-intense experimental campaigns are, being able to forecast negative transfer is critical. While various approaches have been proposed in the machine learning literature,[64] it remains an unsolved problem. As a source of negative transfer may be the involvement of uninformative data,[51] similarity metrics connecting source reactions and the transformation of interest could possibly be beneficial. Such metrics are available for structural similarity[65,66] but are not established for overall reaction similarity.

Forecasting negative transfer appears to be a tenuous task when the target domain contains little to no data. To investigate transferability in a realistic data set, our recent study evaluated the predictivity of models trained on a few dozen palladium-catalyzed cross-coupling reactions of a specific nucleophile (e.g., phenyl benzamide) toward reactions of other nucleophiles (e.g., pyrazole and pinacol boronates).[67] The ability to classify reaction outcomes was measured with the receiver operating characteristic−area under the curve (ROC-AUC) metric. ROC-AUC of 1.0 and 0.5 corresponds to perfect classification and random guessing, respectively. The predictions made by the phenyl benzamide model were excellent for reactions of pyrazole (ROC-AUC = 0.91) and related nucleophiles. In contrast, for pinacol boronate nucleophiles the quality of predictions was worse than random (ROC-AUC = 0.13), indicating negative transfer. From the model's perspective, this is surprising because the target reaction conditions were seen during training. However, this difference in transferability maps well onto expert classification of functional groups and their mechanistic knowledge, which would predict that the pinacol boronate nucleophiles are mechanistically distinct from nitrogen nucleophiles. Currently, identifying these mechanistic distinctions by experts seems to be the only means to avoid negative transfer.

In addition to the mechanistic relevance of the source to the target reaction domain, the information content of the transferred model influences the performance of transfer learning. Models trained from reaction data sets that span the full combinations of viable reagents could learn crucial interactions between reagents that are otherwise difficult to anticipate. However, there is a limitation to the number of reactions that can be conducted, so preparing complete data sets is impossible for all but the smallest experimental spaces. With realistic, sparse source data sets, predictions on target reaction condition candidates may be needed even for reagent combinations where no data exists. Manually, chemists have long made informed predictions on new conditions based on sparse reaction data, using chemical principles and intuition to bridge gaps in missing data. In a modeling context, this could be mimicked through regularization or model simplification.[68,69] Accordingly, we investigated the performance of simplified random forest models for predicting outcomes of target reactions involving reaction conditions that were unseen in the model's training. Simplified source models showed benefits to transfer for some pairs of source−target nucleophiles. Specifically, for a simplified benzamide source model and pyrazole target reaction pair, ROC-AUC of 0.65 was achieved, while a conventional cross-validated model had almost zero predictive ability, with an ROC-AUC of 0.52. Collectively, even without any target reaction data points, appropriately transferred models trained on reactions of a relevant substrate class can locate an effective starting point for exploring reaction conditions for the target substrate. However, definitive methods need to be developed for reliable prospective application of transfer learning, rather than the *ad hoc* approaches attempted so far.

In summary, transfer learning is a machine learning approach for making initial hypotheses in a new reaction space, maximizing use of prior reaction data from a nearby but indirectly related reaction space. Inspiration drawn from expert chemists' workflow can be used to better align all elements of

https://doi.org/10.1021/acs.jcim.3c00577
J. Chem. Inf. Model. 2023, 63, 3659−3668

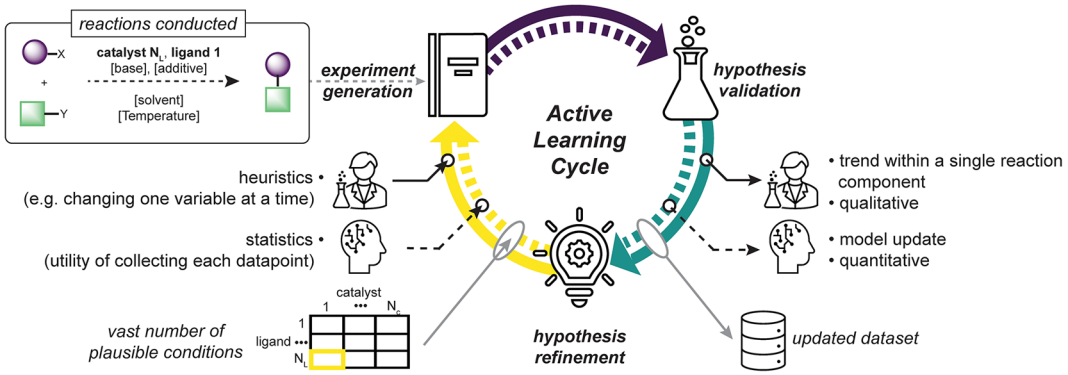


Figure 5. An analogy between hypothesis-driven iterative experimentation (solid lines) and active learning (dashed lines). The differences in operation between human and computational methods at each step are highlighted.

transfer learning—the source data set, transfer learning strategy, and target data set—with the reaction development setting (Figure 2). Smaller source data sets and models that can learn better from the limited data deserve more attention as large data sources are often unavailable for newer, and perhaps more interesting, transformations (Figure 3). Also, leveraging multiple source reaction data sets and their components toward reaction design could be an interesting transfer learning approach to reaction development (Figure 4). Such advances in transfer learning might lead to recommendations of more meaningful initial reaction conditions to explore.

■ ACTIVE LEARNING AND BAYESIAN OPTIMIZATION

Initial hypotheses for reaction conditions—coming from transfer learning or chemical intuition—must be updated as the results of initial experiments come in. In many situations, only one reaction component is updated and the variability with those components is evaluated. Improved reagents are merged into the current best-known set of reaction conditions, and the process is repeated until satisfactory yields are obtained. This method of validating, refining, and generating hypotheses (Figure 5) is a fundamental of data-driven science and has been a standard for development of new reactive chemistries for a long time.

When computers use statistical models to choose the data points to test, reaction condition development falls under the umbrella of active learning[68] (Figure 5). To initiate active learning, a model is trained to predict the utility of unlabeled data points (i.e., plausible reactions that are yet to be conducted). The model can either be exploitative and seek the best reaction outcome or be explorative and seek to reduce uncertainty in the model.[70,71] After conducting the most "useful" reactions, the model is updated[72] and then used to generate another set of experiments to validate, starting a new cycle. For organic reactions where the goal is quantitative (e.g., to obtain the highest yield or selectivity) optimization frameworks such as Bayesian optimization[73,74] are often preferred. As highlighted by Figure 5, active learning provides a computational analogue to the conventional hypothesis-driven experimentation described in the previous paragraph.

Early examples of autonomous iterative experimentation focused on using flow reactors to understand the kinetics of specific chemical reactions[75,76] and optimizing their outcomes.[77] Models such as linear or quadratic fitting are built with available data to inform the next experiment to conduct. These strategies are useful in understanding, improving, and scaling up chemical processes.[78] More recently, machine learning algorithms have taken prominence to replace the prior modeling strategies, though predicting quantitative reaction outcomes remains the goal.

The black-box nature of machine learning algorithms, however, merits further comparisons between human- and machine-driven experimentation, and shows key differences and limitations. First, the quality of the information in the reaction data and how it is used is different between the two types of experimentation (Figure 5, green arrow). The common expert approach of manually changing one variable at a time allows the extraction of qualitative yet interpretable reactivity trends from reaction data. In contrast, machine learning models are trained through numerical processes, which as a result makes them less interpretable.[79] More importantly, the process of selecting a small set of experiments to conduct from the vast space of plausible reaction condition differs significantly (Figure 5, yellow arrow). In the traditional approach, the best performing reagent identified from the previous iteration is fixed and the next single reaction component to screen is selected with heuristics based on intuition. While this approach can improve yield, it is a search in a narrow region of space and may lead to identification of local maxima. Active learning, on the other hand, can evaluate the whole candidate space defined by the chemist with statistical scores. Whether active learning searches a narrow or wide space depends on its objective function and underlying models, which in principle could achieve either limit or some goal between the two. Therefore, even a theoretically "perfect" active learning method would require guidance from a chemist to choose its objectives.

Regardless of potential limitations, active learning and Bayesian optimization continue to be demonstrated in a handful of reaction optimization campaigns in batch experimentation settings.[42,80−83] One of the earlier studies used active learning with random forest regressors that predict yields to identify optimal numerical variables such as stoichiometry, temperature, and time. Reaction conditions for a variety of transformations were optimized within a few iterations. For example, for an *O*-glycosylation of tyrosine, an increase in yield from 40% to 68% was achieved within seven iterations (sampling approximately 0.1% of the total conceivable reaction conditions, Figure 6A).[42] More recently,

https://doi.org/10.1021/acs.jcim.3c00577
J. Chem. Inf. Model. 2023, 63, 3659−3668

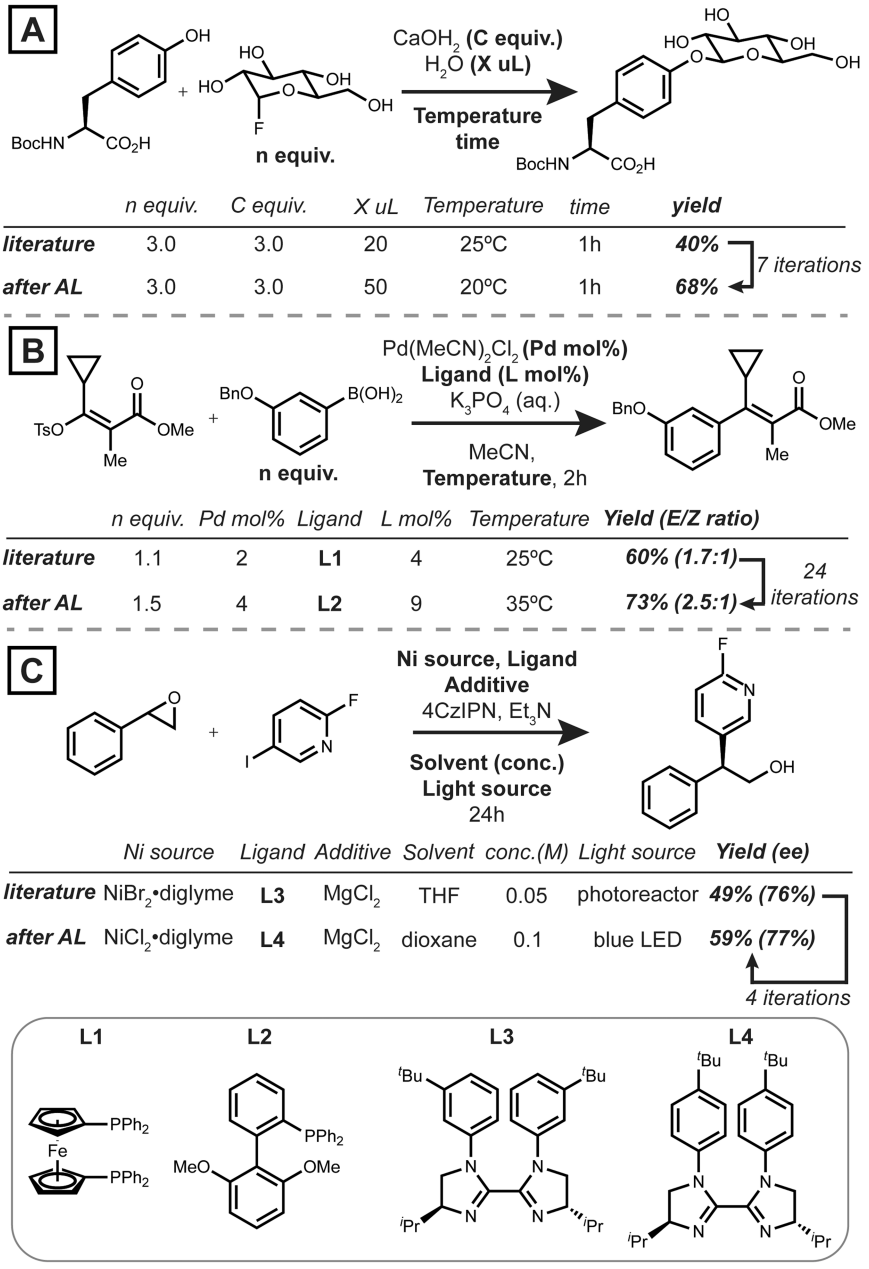


Figure 6. Representative examples of active learning[42] and Bayesian optimization studies.[80,81]

optimization of a stereospecific Suzuki−Miyaura coupling has been achieved using a robotic platform connected to inline analytic instrumentation and multiobjective Bayesian optimization frameworks. The yield and stereoselectivity were increased from 60% yield and E/Z = 1.7:1 to 73% yield and E/Z = 2.5:1 (Figure 6B).[80] In another work, a multiobjective BO tool was developed and used to improve nickel/photoredox-catalyzed enantioselective couplings between styrene oxide and two aryl iodides. Yields increased by 10% while retaining stereoselectivity (one of the two examples shown in Figure 6C).[81]

Going beyond the optimization of a reaction between a specific substrate pair, another Bayesian optimization approach optimizes for substrate generality (i.e., across multiple substrate pairs). By conducting five rounds of experimental optimization consisting of 300 automated Suzuki-coupling reactions using 11 pairs of heteroaryl bromides and *N*-methyliminodiacetic acid boronates, three improved reaction conditions were identified. Compared to a benchmark literature condition,[84] two showed better yields for 20 diverse and challenging Suzuki couplings.[85] In another recent report, active learning was conducted on a wide array of substrate classes for the Buchwald−Hartwig reaction with a goal to develop a general yield prediction model. After conducting experiments with more than 130 reactant pairs, each under 24 reaction conditions, the resulting model could prioritize high-yielding reaction conditions for new substrates.[86]

The above examples show that active learning and Bayesian optimization have made inroads toward reaction optimization under realistic experimentation settings. As described, the models must be trained before their first-round hypotheses can be made so selection of initial reaction data is critical to their success. For instance, random selection of reaction conditions offers a set of reaction conditions that is unbiased,[87] but likely

https://doi.org/10.1021/acs.jcim.3c00577
J. Chem. Inf. Model. 2023, 63, 3659−3668

suboptimal.[42,80] More sophisticated methods select data points that span the feature space as widely as possible, providing better initial models for active learning.[81] When multiple substrates are under consideration, selecting centroids of clusters based on structural similarity[85] or physical properties[86] can be used. For reaction development, however, it is important to have information on positive yielding reactions that would lead to the desired product. Being able to utilize effective chemical knowledge at the start of an active learning experimentation therefore could be key to developing reaction conditions as quickly as possible.

COMBINING TRANSFER LEARNING AND ACTIVE LEARNING

Using prior art to design plausible reactions is at the core of chemist's process of reaction development. In the machine learning analogy, transfer learning can leverage prior art to help direct active learning toward a potentially fruitful space. In addition to providing an initialization scheme, the information within the source model could also guide subsequent cycles of exploration. To gain such benefits, transfer learning needs to be seamlessly incorporated into active learning or Bayesian optimization, which requires two aspects to be considered. Most importantly, transfer learning must boost the performance of active learning. In other words, negative transfer must be avoided, which is a significant problem on its own, as described above. If this condition is satisfied, the role of transferred source reaction information needs to be carefully considered. For example, as target reaction data is collected from different areas of the reaction space, only a subset of the source information might be useful for training a model, or the source data might be reweighted according to its utility in the target region. Although challenging, an effective design combining transfer learning and active learning or Bayesian optimization may enable performance that cannot be reached with either approach alone.

For challenging cases where transferred models are incapable of making effective predictions in the target reaction space, the combination of active learning with transfer learning can be particularly useful.[67] In our previous report, random forest models trained on coupling reactions involving benzamide were unable to predict reaction conditions for pyrazole (the rate of identifying positive reactions was narrowly better than random selection). When the models were updated by retraining on source and target data simultaneously, only a slight improvement was seen in predictive ability in the target space. Better performance was achieved by training a target model separately from the source model and using the two models in tandem. The transferred model and active learning acted hand in hand, the former providing knowledge of catalysts and the latter supplementing information on other reaction components (base and solvent) which helped the source model to adapt in the target space.[67]

REFLECTIONS AND OUTLOOK

More and more examples have been showing that data science can facilitate the development of new chemistries.[88−92] Transfer and active learning have gained attention due to their ability to empower predictivity in low-data situations. The examples mentioned above show how reaction outcome prediction and reaction condition prioritization have been approached through transfer and active learning, respectively. There is no formal restriction, however, to applying the concepts of transfer and active learning to broader classes of problems. The general applicability of machine learning methods therefore deserves considerably more study. Demonstration of these tools to less-explored reactions with diverse above-the-arrow conditions is needed to advance the field. This is further necessary when heuristics are involved (for instance, selecting initial experiments), since their generalization to new reaction types becomes less certain. For these tools to gain a higher reputation for reliability across a broader community of organic chemists, discussions on how careful studies can address common concerns (such as scope of applicability and cost) are warranted.

How do we demonstrate that data science tools have utility for prospective applications in reaction development? One way is to share all attempts at prospective application regardless of their success.[24] Because a wide range of heuristics and models are involved, both positive and negative results need to be shared to assess the strengths and weaknesses of various data-driven strategies. Moreover, unsuccessful scenarios could not only be the start of an improved algorithm, but also be an indicator of challenging chemistry where new thinking is needed to make progress.

The practical cost of machine learning strategies in reaction development is a central factor that can limit their widespread application. While it appears accepted that machine learning can model chemical properties and transformations, the amount of data required to do so is often much more than required by an expert chemist to make predictions in the same space. Therefore, cost analyses would be a strong support for their adoption, and we suggest three here. The extent of improvement in reaction outcome such as yield or selectivity is a crucial factor. Although comparison to algorithmic baselines is useful for this metric, comparison to human chemists is relatively less explored but can be insightful.[42,82,93] Next, the cost of experiments needs to be considered. How can machine learning approaches—which usually demand a large amount of data—demonstrate reductions in cost, time, and human resources to accomplish a challenging reaction development task? Lastly, it is important to note that chemical understanding leads to generalization of reactions, which is a highly valuable outcome of the reaction development process. In contrast, the complexity of machine learning algorithms makes it difficult to extract reactivity principles. The machine learning community is well aware of this issue, and there are ongoing efforts to build interpretable models.[94] Moreover, the statistical nature of machine learning-guided explorations may illuminate underexplored chemical reactivities. Although metrics for these three costs are largely nonobvious and intertwined, they will be helpful to support the utility of prospective application of machine learning approaches.[78−83]

CONCLUSION

Conventional and high-throughput experiments, mechanistic studies, and modeling approaches such as quantum chemistry work together to continuously develop new catalysts, reactions, and synthetic routes to challenging chemical targets. Machine learning can inject statistical backing to these chemical studies and provide practical value, although currently the "when and if" this will hold true is difficult to tell from existing studies. How much data is needed? Which machine learning approaches are most tolerant to realistic, low-data scenarios? Will useful chemical concepts be transparently uncovered? We

https://doi.org/10.1021/acs.jcim.3c00577
J. Chem. Inf. Model. 2023, 63, 3659−3668

hope that the high level of discussions[77−80] and increasing appearances of systematic, prospective studies[95,96] help answer these questions. As the community accumulates experience, the science of machine learning explorations through reaction space—especially in the low-data limit—may soon become an indispensable toolbox for chemists. Ultimately, this will let experts focus on the real challenge of designing and discovering important transformations and functional molecules.

AUTHOR INFORMATION

Corresponding Author

Paul M. Zimmerman − *Department of Chemistry, University of Michigan, Ann Arbor, Michigan 48109, United States;* orcid.org/0000-0002-7444-1314; Email: paulzim@umich.edu

Authors

Eunjae Shim − *Department of Chemistry, University of Michigan, Ann Arbor, Michigan 48109, United States;* orcid.org/0000-0002-4085-9659

Ambuj Tewari − *Department of Statistics and Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, Michigan 48109, United States;* orcid.org/0000-0001-6969-7844

Tim Cernak − *Department of Chemistry, University of Michigan, Ann Arbor, Michigan 48109, United States; Department of Medicinal Chemistry, University of Michigan, Ann Arbor, Michigan 48109, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.3c00577

Notes

The authors declare the following competing financial interest(s): The Cernak Lab has received research funding or in-kind donations from MilliporeSigma, Relay Therapeutics, Janssen Therapeutics, SPT Labtech, and Merck & Co., Inc. T.C. holds equity in Scorpion Therapeutics and is a cofounder of and equity holder in Entos, Inc. Other authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank the NIH for support through R35GM128830 (E.S. and P.M.Z.) and the NSF through CHE-2236215 (T.C.) and IIS-2007055 (A.T.).

REFERENCES

(1) Bohacek, R. S.; McMartin, C.; Guida, W. C. The Art and Practice of Structure-Based Drug Design: A Molecular Modeling Perspective. *Med. Res. Rev.* **1996**, *16*, 3−50.

(2) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.

(3) NASA. *What is the Universe Made Of?* https://map.gsfc.nasa.gov/universe/uni_matter.html (accessed 2022-11−23).

(4) Li, X.; Xu, Y.; Yao, H.; Lin, K. Chemical Space Exploration Based on Recurrent Neural Networks: Applications in Discovering Kinase Inhibitors. *J. Cheminform.* **2020**, *12*, 42.

(5) Zhavoronkov, A.; Ivanenkov, Y. A.; Aliper, A.; Veselov, M. S.; Aladinskiy, V. A.; Aladinskaya, A. V.; Terentiev, V. A.; Polykovskiy, D. A.; Kuznetsov, M. D.; Asadulaev, A.; Volkov, Y.; Zholus, A.; Shayakhmetov, R. R.; Zhebrak, A.; Minaeva, L. I.; Zagribelnyy, B. A.; Lee, L. H.; Soll, R.; Madge, D.; Xing, L.; Guo, T.; Aspuru-Guzik, A. Deep Learning Enables Rapid Identification of Potent DDR1 Kinase Inhibitors. *Nat. Biotechnol.* **2019**, *37*, 1038−1040.

(6) Lu, C.; Liu, S.; Shi, W.; Yu, J.; Zhou, Z.; Zhang, X.; Lu, X.; Cai, F.; Xia, N.; Wang, Y. Systemic Evolutionary Chemical Space Exploration for Drug Discovery. *J. Cheminform.* **2022**, *14*, 19.

(7) Burger, B.; Maffettone, P. M.; Gusev, V. V.; Aitchison, C. M.; Bai, Y.; Wang, X.; Li, X.; Alston, B. M.; Li, B.; Clowes, R.; Rankin, N.; Harris, B.; Sprick, R. S.; Cooper, A. I. A Mobile Robotic Chemist. *Nature* **2020**, *583*, 237−241.

(8) 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* **2016**, *533*, 73−76.

(9) Xu, S.; Li, J.; Cai, P.; Liu, X.; Liu, B.; Wang, X. Self-Improving Photosensitizer Discovery System via Bayesian Search with First-Principle Simulations. *J. Am. Chem. Soc.* **2021**, *143*, 19769.

(10) Nandy, A.; Duan, C.; Kulik, H. J. Using Machine Learning and Data Mining to Leverage Community Knowledge for the Engineering of Stable Metal-Organic Frameworks. *J. Am. Chem. Soc.* **2021**, *143*, 17535−17547.

(11) Sumita, M.; Yang, X.; Ishihara, S.; Tamura, R.; Tsuda, K. Hunting for Organic Molecules with Artificial Intelligence: Molecules Optimized for Desired Excitation Energies. *ACS Cent. Sci.* **2018**, *4*, 1126−1133.

(12) Cheng, C. Y.; Campbell, J. E.; Day, G. M. Evolutionary Chemical Space Exploration for Functional Materials: Computational Organic Semiconductor Discovery. *Chem. Sci.* **2020**, *11*, 4922−4933.

(13) del Cueto, M.; Troisi, A. Determining Usefulness of Machine Learning in Materials Discovery Using Simulated Research Landscapes. *Phys. Chem. Chem. Phys.* **2021**, *23*, 14156−14163.

(14) Gómez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Duvenaud, D.; Maclaurin, D.; Blood-Forsythe, M. A.; Chae, H. S.; Einzinger, M.; Ha, D.-G.; Wu, T.; Markopoulos, G.; Jeon, S.; Kang, H.; Miyazaki, H.; Numata, M.; Kim, S.; Huang, W.; Hong, S. I.; Baldo, M.; Adams, R. P.; Aspuru-Guzik, A. Design of Efficient Molecular Organic Light-Emitting Diodes by a High-Throughput Virtual Screening and Experimental Approach. *Nat. Mater.* **2016**, *15*, 1120−1127.

(15) Jablonka, K. M.; Jothiappan, G. M.; Wang, S.; Smit, B.; Yoo, B. Bias Free Multiobjective Active Learning for Materials Design and Discovery. *Nat. Commun.* **2021**, *12*, 2312.

(16) Janet, J. P.; Ramesh, S.; Duan, C.; Kulik, H. J. Accurate Multiobjective Design in a Space of Millions of Transition Metal Complexes with Neural-Network-Driven Efficient Global Optimization. *ACS Cent. Sci.* **2020**, *6*, 513−524.

(17) Molga, K.; Szymkuć, S.; Grzybowski, B. A. Chemist Ex Machina: Advanced Synthesis Planning by Computers. *Acc. Chem. Res.* **2021**, *54*, 1094−1106.

(18) Wołos, A.; Koszelewski, D.; Roszak, R.; Szymkuć, S.; Moskal, M.; Ostaszewski, R.; Herrera, B. T.; Maier, J. M.; Brezicki, G.; Samuel, J.; Lummiss, J. A. M.; McQuade, D. T.; Rogers, L.; Grzybowski, B. A. Computer-Designed Repurposing of Chemical Wastes into Drugs. *Nature* **2022**, *604*, 668−676.

(19) Coley, C. W.; Green, W. H.; Jensen, K. F. Machine Learning in Computer-Aided Synthesis Planning. *Acc. Chem. Res.* **2018**, *51*, 1281−1289.

(20) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nature* **2018**, *555*, 604−610.

(21) Caramelli, D.; Granda, J. M.; Mehr, S. H. M.; Cambié, D.; Henson, A. B.; Cronin, L. Discovering New Chemistry with an Autonomous Robotic Platform Driven by a Reactivity-Seeking Neural Network. *ACS Cent. Sci.* **2021**, *7*, 1821−1830.

(22) 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* **2018**, *559*, 377−381.

(23) Gromski, P. S.; Henson, A. B.; Granda, J. M.; Cronin, L. How to Explore Chemical Space Using Algorithms and Automation. *Nat. Rev. Chem.* **2019**, *3*, 119−128.

(24) Kearnes, S. M.; Maser, M. R.; Wleklinski, M.; Kast, A.; Doyle, A. G.; Dreher, S. D.; Hawkins, J. M.; Jensen, K. F.; Coley, C. W. The Open Reaction Database. *J. Am. Chem. Soc.* **2021**, *143*, 18820−18826.
(25) Sandfort, F.; Strieth-Kalthoff, F.; Kühnemund, M.; Beecks, C.; Glorius, F. A Structure-Based Platform for Predicting Chemical Reactivity. *Chem.* **2020**, *6*, 1379−1390.
(26) 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.
(27) 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* **2018**, *360*, 186−190.
(28) Schwaller, P.; Vaucher, A. C.; Laino, T.; Reymond, J.-L. Prediction of Chemical Reaction Yields Using Deep Learning. *Mach. Learn. Sci. Technol.* **2021**, *2*, 015016.
(29) Segler, M. H. S.; Waller, M. P. Modelling Chemical Reasoning to Predict and Invent Reactions. *Chem.—Eur. J.* **2017**, *23*, 6118−6128.
(30) Struble, T. J.; Coley, C. W.; Jensen, K. F. Multitask Prediction of Site Selectivity in Aromatic C-H Functionalization Reactions. *React. Chem. Eng.* **2020**, *5*, 896−902.
(31) Kreutter, D.; Schwaller, P.; Reymond, J.-L. Predicting Enzymatic Reactions with a Molecular Transformer. *Chem. Sci.* **2021**, *12*, 8648−8659.
(32) 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. *Angew. Chem., Int. Ed.* **2019**, *58*, 4515−4519.
(33) Schwaller, P.; Gaudin, T.; Lányi, D.; Bekas, C.; Laino, T. Found in Translation": Predicting Outcomes of Complex Organic Chemistry Reactions Using Neural Sequence-to-Sequence Models. *Chem. Sci.* **2018**, *9*, 6091−6098.
(34) Coley, C. W.; Jin, W.; Rogers, L.; Jamison, T. F.; Jaakkola, T. S.; Green, W. H.; Barzilay, R.; Jensen, K. F. A Graph-Convolutional Neural Network Model for the Prediction of Chemical Reactivity. *Chem. Sci.* **2019**, *10*, 370−377.
(35) 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* **2019**, *363*, No. eaau5631.
(36) Seidl, P.; Renz, P.; Dyubankova, N.; Neves, P.; Verhoeven, J.; Wegner, J. K.; Segler, M.; Hochreiter, S.; Klambauer, G. Improving Few- and Zero-Shot Reaction Template Prediction Using Modern Hopfield Networks. *J. Chem. Inf. Model.* **2022**, *62*, 2111−2120.
(37) Jiang, J.; Wang, R.; Wang, M.; Gao, K.; Nguyen, D. D.; Wei, G.-W. Boosting Tree-Assisted Multitask Deep Learning for Small Scientific Datasets. *J. Chem. Inf. Model.* **2020**, *60*, 1235−1244.
(38) Yamada, H.; Liu, C.; Wu, S.; Koyama, Y.; Ju, S.; Shiomi, J.; Morikawa, J.; Yoshida, R. Predicting Materials Properties with Little Data Using Shotgun Transfer Learning. *ACS Cent. Sci.* **2019**, *5*, 1717−1730.
(39) St. John, P. C.; Guan, Y.; Kim, Y.; Kim, S.; Paton, R. S. Prediction of Organic Homolytic Bond Dissociation Enthalpies at Near Chemical Accuracy with Sub-Second Computational Cost. *Nat. Commun.* **2020**, *11*, 2328.
(40) Muratov, E. N.; Bajorath, J.; Sheridan, R. P.; Tetko, I. V.; Filimonov, D.; Poroikov, V.; Oprea, T. I.; Baskin, I. I.; Varnek, A.; Roitberg, A.; Isayev, O.; Curtalolo, S.; Fourches, D.; Cohen, Y.; Aspuru-Guzik, A.; Winkler, D. A.; Agrafiotis, D.; Cherkasov, A.; Tropsha, A. QSAR Without Borders. *Chem. Soc. Rev.* **2020**, *49*, 3525−3564.
(41) Roszak, R.; Beker, W.; Molga, K.; Grzybowski, B. A. Rapid and Accurate Prediction of pKa Values of C-H Acids Using Graph Convolutional Neural Networks. *J. Am. Chem. Soc.* **2019**, *141*, 17142−17149.
(42) Reker, D.; Hoyt, E. A.; Bernardes, G. J. L.; Rodrigues, T. Adaptive Optimization of Chemical Reactions with Minimal Experimental Information. *Cell Reports Phys. Sci.* **2020**, *1*, 100247.
(43) Almeida, A. F.; Ataíde, F. A. P.; Loureiro, R. M. S.; Moreira, R.; Rodrigues, T. Augmenting Adaptive Machine Learning with Kinetic Modeling for Reaction Optimization. *J. Org. Chem.* **2021**, *86*, 14192−14198.
(44) Li, X.; Maffettone, P. M.; Che, Y.; Liu, T.; Chen, L.; Cooper, A. I. Combining Machine Learning and High-Throughput Experimentation to Discover Photocatalytically Active Organic Molecules. *Chem. Sci.* **2021**, *12*, 10742−10754.
(45) Reker, D.; Schneider, P.; Schneider, G. Multi-Objective Active Machine Learning Rapidly Improves Structure-Activity Models and Reveals New Protein-Protein Interaction Inhibitors. *Chem. Sci.* **2016**, *7*, 3919−3927.
(46) Lin, S.; Dikler, S.; Blincoe, W. D.; Ferguson, R. D.; Sheridan, R. P.; Peng, Z.; Conway, D. V.; Zawatzky, K.; Wang, H.; Cernak, T.; Davies, I. W.; DiRocco, D. A.; Sheng, H.; Welch, C. J.; Dreher, S. D. Mapping the Dark Space of Chemical Reactions with Extended Nanomole Synthesis and MALDI-TOF MS. *Science* **2018**, *361*, No. eaar6236.
(47) Gesmundo, N. J.; Sauvagnat, B.; Curran, P. J.; Richards, M. P.; Andrews, C. L.; Dandliker, P. J.; Cernak, T. Nanoscale Synthesis and Affinity Ranking. *Nature* **2018**, *557*, 228−232.
(48) Buitrago Santanilla, A.; Regalado, E. L.; Pereira, T.; Shevlin, M.; Bateman, K.; Campeau, L.-C.; Schneeweis, J.; Berritt, S.; Shi, Z.-C.; 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* **2015**, *347*, 49−53.
(49) Mahjour, B.; Shen, Y.; Cernak, T. Ultrahigh-Throughput Experimentation for Information-Rich Chemical Synthesis. *Acc. Chem. Res.* **2021**, *54*, 2337−2346.
(50) 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* **2018**, *359*, 429−434.
(51) Pan, S. J.; Yang, Q. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 1345−1359.
(52) Cai, C.; Wang, S.; Xu, Y.; Zhang, W.; Tang, K.; Ouyang, Q.; Lai, L.; Pei, J. Transfer Learning for Drug Discovery. *J. Med. Chem.* **2020**, *63*, 8683−8694.
(53) Open AI. GPT-4 Technical Report. *arXiv* [cs.CL]; 2303.08774; **2023**. DOI: 10.48550/arxiv.2303.08774.
(54) Pesciullesi, G.; Schwaller, P.; Laino, T.; Reymond, J.-L. Transfer Learning Enables the Molecular Transformer to Predict Regio- and Stereoselective Reactions on Carbohydrates. *Nat. Commun.* **2020**, *11*, 4874.
(55) Singh, S.; Sunoj, R. B. A Transfer Learning Protocol for Chemical Catalysis Using a Recurrent Neural Network Adapted from Natural Language Processing. *Digital Discovery* **2022**, *1*, 303−312.
(56) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent. Sci.* **2019**, *5*, 1572−1583.
(57) Lowe, D. M. *Chemical Reactions from US Patents (1976-Sep2016)*, 2017. https://figshare.com/articles/dataset/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873.
(58) Reaxys database. https://www.reaxys.com.
(59) Lowe, D. M. *Extraction of Chemical Structures and Reactions from the Literature*; University of Cambridge, 2012.
(60) Reid, J. P.; Sigman, M. S. Holistic Prediction of Enantioselectivity in Asymmetric Catalysis. *Nature* **2019**, *571*, 343−348.
(61) Schleinitz, J.; Langevin, M.; Smail, Y.; Wehnert, B.; Grimaud, L.; Vuilleumier, R. Machine Learning Yield Prediction from NiCOlit, a Small-Size Literature Data Set of Nickel Catalyzed C-O Couplings. *J. Am. Chem. Soc.* **2022**, *144*, 14722−14730.
(62) Fitzner, M.; Wuitschik, G.; Koller, R.; Adam, J.-M.; Schindler, T. Machine Learning C-N Couplings: Obstacles for a General-Purpose Reaction Yield Prediction. *ACS Omega* **2023**, *8*, 3017−3025.

(63) Shu, Y.; Kou, Z.; Cao, Z.; Wang, J.; Long, M. Zoo-Tuning: Adaptive Transfer from a Zoo of Models. *Int. Conf. Mach. Learn.* **2021**, *38*, 9626−9637.
(64) Zhang, W.; Deng, L.; Zhang, L.; Wu, D. A Survey on Negative Transfer. *IEEE/CAA J. Autom. Sinica* **2023**, *10*, 305.
(65) Maggiora, G.; Vogt, M.; Stumpfe, D.; Bajorath, J. Molecular Similarity in Medicinal Chemistry. *J. Med. Chem.* **2014**, *57*, 3186−3204.
(66) Lin, Y.; Zhang, R.; Wang, D.; Cernak, T. Computer-Aided Key Step Generation in Alkaloid Total Synthesis. *Science* **2023**, *379*, 453−457.
(67) Shim, E.; Kammeraad, J. A.; Xu, Z.; Tewari, A.; Cernak, T.; Zimmerman, P. M. Predicting Reaction Conditions from Limited Data Through Active Transfer Learning. *Chem. Sci.* **2022**, *13*, 6655−6668.
(68) BRESLOW, L. A.; AHA, D. W. Simplifying Decision Trees: A Survey. *Knowledge Engineering Review* **1997**, *12*, 1−40.
(69) Rapp, M.; Mencía, E. L.; Fürnkranz, J. Simplifying Random Forests: On the Trade-Off Between Interpretability and Accuracy. *arXiv [cs.LG]* **2019**, 1911.04393 DOI: 10.48550/arxiv.1911.04393.
(70) Settles, B. Active Learning Literature Survey. *Computer Sciences Technical Report*, **2009**.
(71) Ren, P.; Xiao, Y.; Chang, X.; Huang, P.-Y.; Li, Z.; Gupta, B. B.; Chen, X.; Wang, X. A Survey of Deep Active Learning. *ACM Comput. Surv.* **2022**, *54*, 1.
(72) Chen, K.; Chen, G.; Li, J.; Huang, Y.; Wang, E.; Hou, T.; Heng, P.-A. MetaRF: Attention-Based Random Forest for Reaction Yield Prediction with a Few Trails. *J. Cheminform.* **2023**, *15*, 43.
(73) Shahriari, B.; Swersky, K.; Wang, Z.; Adams, R. P.; de Freitas, N. Taking the Human Out of the Loop: A Review of Bayesian Optimization. *Proc. IEEE* **2016**, *104*, 148−175.
(74) Garnett, R. *Bayesian Optimization*; Cambridge University Press, 2023. https://bayesoptbook.com/.
(75) McMullen, J. P.; Jensen, K. F. Rapid Determination of Reaction Kinetics with an Automated Microfluidic System. *Org. Process Res. Dev.* **2011**, *15*, 398−407.
(76) Taylor, C. J.; Booth, M.; Manson, J. A.; Willis, M. J.; Clemens, G.; Taylor, B. A.; Chamberlain, T. W.; Bourne, R. A. Rapid, Automated Determination of Reaction Models and Kinetic Parameters. *Chem. Eng. J.* **2021**, *413*, 127017.
(77) Reizman, B. J.; Jensen, K. F. Feedback in Flow for Accelerated Reaction Development. *Acc. Chem. Res.* **2016**, *49*, 1786−1796.
(78) Taylor, C. J.; Pomberger, A.; Felton, K. C.; Grainger, R.; Barecka, M.; Chamberlain, T. W.; Bourne, R. A.; Johnson, C. N.; Lapkin, A. A. A Brief Introduction to Chemical Reaction Optimization. *Chem. Rev.* **2023**, *123*, 3089−3126.
(79) Kammeraad, J. A.; Goetz, J.; Walker, E. A.; Tewari, A.; Zimmerman, P. M. What Does the Machine Learn? Knowledge Representations of Chemical Reactivity. *J. Chem. Inf. Model.* **2020**, *60*, 1290−1301.
(80) Christensen, M.; Yunker, L. P. E.; Adedeji, F.; Hase, F.; Roch, L. M.; Gensch, T.; dos Passos Gomes, G.; Zepel, T.; Sigman, M. S.; Aspuru-Guzik, A.; Hein, J. E. Data-Science Driven Autonomous Process Optimization. *Commun. Chem.* **2021**, *4*, 112.
(81) Torres, J. A. G.; Lau, S. H.; Anchuri, P.; Stevens, J. M.; Tabora, J. E.; Li, J.; Borovika, A.; Adams, R. P.; Doyle, A. G. A Multi-Objective Active Learning Platform and Web App for Reaction Optimization. *J. Am. Chem. Soc.* **2022**, *144*, 19999−20007.
(82) Shields, B. J.; Stevens, J.; Li, J.; Parasram, M.; Damani, F.; Alvarado, J. I. M.; Janey, J. M.; Adams, R. P.; Doyle, A. G. Bayesian Reaction Optimization as a Tool for Chemical Synthesis. *Nature* **2021**, *590*, 89−96.
(83) Eyke, N. S.; Green, W. H.; Jensen, K. F. Iterative Experimental Design Based on Active Machine Learning Reduces the Experimental Burden Associated with Reaction Screening. *React. Chem. Eng.* **2020**, *5*, 1963−1972.
(84) Knapp, D. M.; Gillis, E. P.; Burke, M. D. A General Solution for Unstable Boronic Acids: Slow-Release Cross-Coupling from Air-Stable MIDA Boronates. *J. Am. Chem. Soc.* **2009**, *131*, 6961−6963.
(85) Angello, N. H.; Rathore, V.; Beker, W.; Wołos, A.; Jira, E. R.; Roszak, R.; Wu, T. C.; Schroeder, C. M.; Aspuru-Guzik, A.; Grzybowski, B. A.; Burke, M. D. Closed-Loop Optimization of General Reaction Conditions for Heteroaryl Suzuki-Miyaura Coupling. *Science* **2022**, *378*, 399−405.
(86) Rinehart, N. I.; Saunthwal, R. K.; Wellauer, J.; Zahrt, A. F.; Schlemper, L.; Shved, A. S.; Bigler, R.; Fantasia, S.; Denmark, S. E. Development and Validation of a Chemoinformatic Workflow for Predicting Reaction Yield for Pd-Catalyzed C-N Couplings with Substrate Generalizability. *ChemRxiv* **2022**, DOI: 10.26434/chemrxiv-2022-hspwv.
(87) Jia, X.; Lynch, A.; Huang, Y.; Danielson, M.; Lang'at, I.; Milder, A.; Ruby, A. E.; Wang, H.; Friedler, S. A.; Norquist, A. J.; Schrier, J. Anthropogenic Biases in Chemical Reaction Data Hinder Exploratory Inorganic Synthesis. *Nature* **2019**, *573*, 251−255.
(88) Rose, B. T.; Timmerman, J. C.; Bawel, S. A.; Chin, S.; Zhang, H.; Denmark, S. E. High-Level Data Fusion Enables the Chemoinformatically Guided Discovery of Chiral Disulfonimide Catalysts for Atropselective Iodination of 2-Amino-6-arylpyridines. *J. Am. Chem. Soc.* **2022**, *144*, 22950−22964.
(89) Liles, J. P.; Rouget-Virbel, C.; Wahlman, J. L. H.; Rahimoff, R.; Crawford, J. M.; Medlin, A.; O'Connor, V. S.; Li, J.; Roytman, V. A.; Toste, F. D.; Sigman, M. S. Data Science Enables the Development of a New Class of Chiral Phosphoric Acid Catalysts. *Chem.* **2023**, DOI: 10.1016/j.chempr.2023.02.020.
(90) Goebel, J. F.; Löffler, J.; Zeng, Z.; Handelmann, J.; Hermann, A.; Rodstein, I.; Gensch, T.; Gessner, V. H.; Gooßen, L. J. Computer-Driven Development of Ylide Functionalized Phosphines for Palladium-Catalyzed Hiyama Couplings. *Angew. Chem., Int. Ed.* **2023**, *62*, No. e202216160.
(91) Tu, Z.; Stuyver, T.; Coley, C. W. Predictive Chemistry: Machine Learning for Reaction Deployment, Reaction Development, and Reaction Discovery. *Chem. Sci.* **2023**, *14*, 226−244.
(92) Su, A.; Wang, X.; Wang, L.; Zhang, C.; Wu, Y.; Wu, X.; Zhao, Q.; Duan, H. Reproducing the Invention of a Named Reaction: Zero-Shot Prediction of Unseen Chemical Reactions. *Phys. Chem. Chem. Phys.* **2022**, *24*, 10280−10291.
(93) Walker, E.; Kammeraad, J.; Goetz, J.; Robo, M. T.; Tewari, A.; Zimmerman, P. M. Learning to Predict Reaction Conditions: Relationships between Solvent, Molecular Structure, and Catalyst. *J. Chem. Inf. Model.* **2019**, *59*, 3645−3654.
(94) Molnar, C. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*; 2022. https://christophm.github.io/interpretable-ml-book/.
(95) Zahrt, A. F.; Mo, Y.; Nandiwale, K. Y.; Shprints, R.; Heid, E.; Jensen, K. F. Machine-Learning-Guided Discovery of Electrochemical Reactions. *J. Am. Chem. Soc.* **2022**, *144*, 22599−22610.
(96) Seumer, J.; Hansen, J. K. S.; Nielsen, M. B.; Jensen, J. H. Computational Evolution Of New Catalysts For The Morita-Baylis-Hillman Reaction. *Angew. Chem., Int. Ed.* **2023**, *62*, No. e202218565.

https://doi.org/10.1021/acs.jcim.3c00577
J. Chem. Inf. Model. 2023, 63, 3659−3668