Navigating through the Maze of Homogeneous Catalyst Design with Machine Learning

Gabriel dos Passos Gomes,[§1,2] Robert Pollice[§1,2] and Alán Aspuru-Guzik*[1,2,3,4]

[1] Chemical Physics Theory Group, Department of Chemistry, University of Toronto, 80 St George St, Toronto, Ontario M5S 3H6, Canada.
[2] Department of Computer Science, University of Toronto, 214 College St., Toronto, Ontario M5T 3A1, Canada.
[3] Vector Institute for Artificial Intelligence, 661 University Ave Suite 710, Toronto, Ontario M5G 1M1, Canada.
[4] Lebovic Fellow, Canadian Institute for Advanced Research (CIFAR), 661 University Ave, Toronto, Ontario M5G 1M1, Canada.
§ *These authors contributed equally to this work.*
* *Corresponding author:* aspuru@utoronto.ca

Abstract

The ability to forge difficult chemical bonds through catalysis has transformed society on all fronts, from feeding our ever-growing populations to increasing our life-expectancies through the synthesis of new drugs. However, developing new chemical reactions and catalytic systems is a tedious task that requires tremendous discovery and optimization efforts. Over the past decade, advances in machine learning have revolutionized a whole new way to approach data-intensive problems, and many of these developments have started to enter chemistry. However, similar progress in the field of homogenous catalysis are only in their infancy. In this perspective, we want to outline our vision for the future of catalyst design and the role of machine learning to navigate this maze.

Introduction

The chemical industry accounts for about 10% of the global trade,[1] and about 85% of all industrial processes are catalytic.[2] About 25% of the global human energy consumption is used for producing chemicals,[2] and the chemical industry sector accounts for about 7% of the global anthropogenic greenhouse gas emissions.[3] To limit the global mean temperature rise to 2 °C above pre-industrial levels, a total reduction of absolute CO_2 emissions in the chemical industry of 30% by 2050 is necessary. This challenge is at odds with a projected increase in demand of 180% for the most energy-intensive chemicals in the same period,[3] mainly due to economic growth in developing countries. This *Gordian knot* can only be disentangled with the development of new catalysts for novel industrial processes.[1] However, the time from the initial idea to the discovery of new catalytic reactions can take several months to several years, and the subsequent process development cycle to deliver a commercial-scale plant adds several more years.[4] Hence, meeting these challenges demands for a significant speed-up in the discovery process of new catalysts, and large-scale application of machine learning promises to deliver such speed-ups.[5]

Catalysis is not only a means to make existing processes more efficient, but it also allows us to synthesize novel and unexplored molecules and materials, enabling the technologies of the future. Homogeneous catalysts, in particular, account for about 15% of all the catalytic

processes, and they are becoming increasingly crucial for specialty and fine chemicals, pharmaceuticals, and materials due to their typically higher selectivities compared to heterogeneous catalysts.[1,2,6] Some prominent exemplary processes include hydroformylation,[7] the single most important industrial process applying homogeneous catalysis, the Hoechst-Wacker process for the oxidation of ethene to acetaldehyde,[8] and the Suzuki-Miyaura cross-coupling,[9] which is especially attractive for fine chemicals.[6] Here, we aim to chart a course for the future of homogeneous catalyst design by the use of machine learning with the ambitious near-term goal to speed up the time from initial conception to experimental demonstration in homogeneous catalysis at least by a factor of two within the next ten years.

Basic Concepts and *Status Quo*

In recent decades, computational chemistry experienced a tremendous surge due to the increasing computational power, and the accompanying heightened practicality to simulate ever-larger ensembles of atoms. Accordingly, these significant advances shifted the focus of computational chemistry from developing methods to simulate matter and benchmarking the results against experiments to predicting the properties of unknown molecules and materials to define new targets for synthesis. This paradigm shift prompted one of us to formulate six grand challenges for the future of simulations summarizing central future research goals in the field of computations charting the way forward.[10] Similarly, computer-aided catalyst design experienced a strong surge and, until recently, one of the main approaches for computational catalyst design was based on *ab initio* simulations of chemical reactions and their potential energy surfaces to predict both thermodynamic and kinetic feasibility of specific transformations.[11–13] Powerful tools have been developed to automate this sometimes human-intensive and tedious process.[14]

In the past decade, homogeneous catalysis has seen a rise in new optimization strategies such as statistical modelling of experimental reactivity and selectivity data with chemical descriptors,[15–17] and the systematic application of machine learning is starting to become more and more common.[18] More specifically, the use of multivariate linear regression models for modelling experimental trends combined with the development of new computational descriptors has allowed for the rapid design of catalysts to improve yields, reaction rates and (enantio-)selectivities.[19–21] While these approaches have pushed the field forward, their limited ability to extrapolate structures beyond the training set hampers inverse-design.[15] More recent studies tried to improve upon the limitations of linear models and used a more sophisticated machine learning approaches like random forest successfully.[22,23] Also, classical high-throughput virtual screening has also been applied to the inverse design of enantioselective catalyst candidates.[24] More recently, a hybrid approach using both computational transition state modelling combined with machine learning has been shown to yield good accuracy for reproducing experimental Gibbs free energies of activation for nucleophilic aromatic substitution reactions, and this workflow is, in principle, also applicable to catalytic reactions.[25] Alternatively, in a series of recent papers, inspired by the classic Sabatier principle established in heterogeneous catalysis,[26,27] volcano plots were introduced as an effective tool to perform high-throughput virtual screening in homogeneous catalysis, and perform insightful result analysis at the same time, to find optimal catalysts.[28–31]

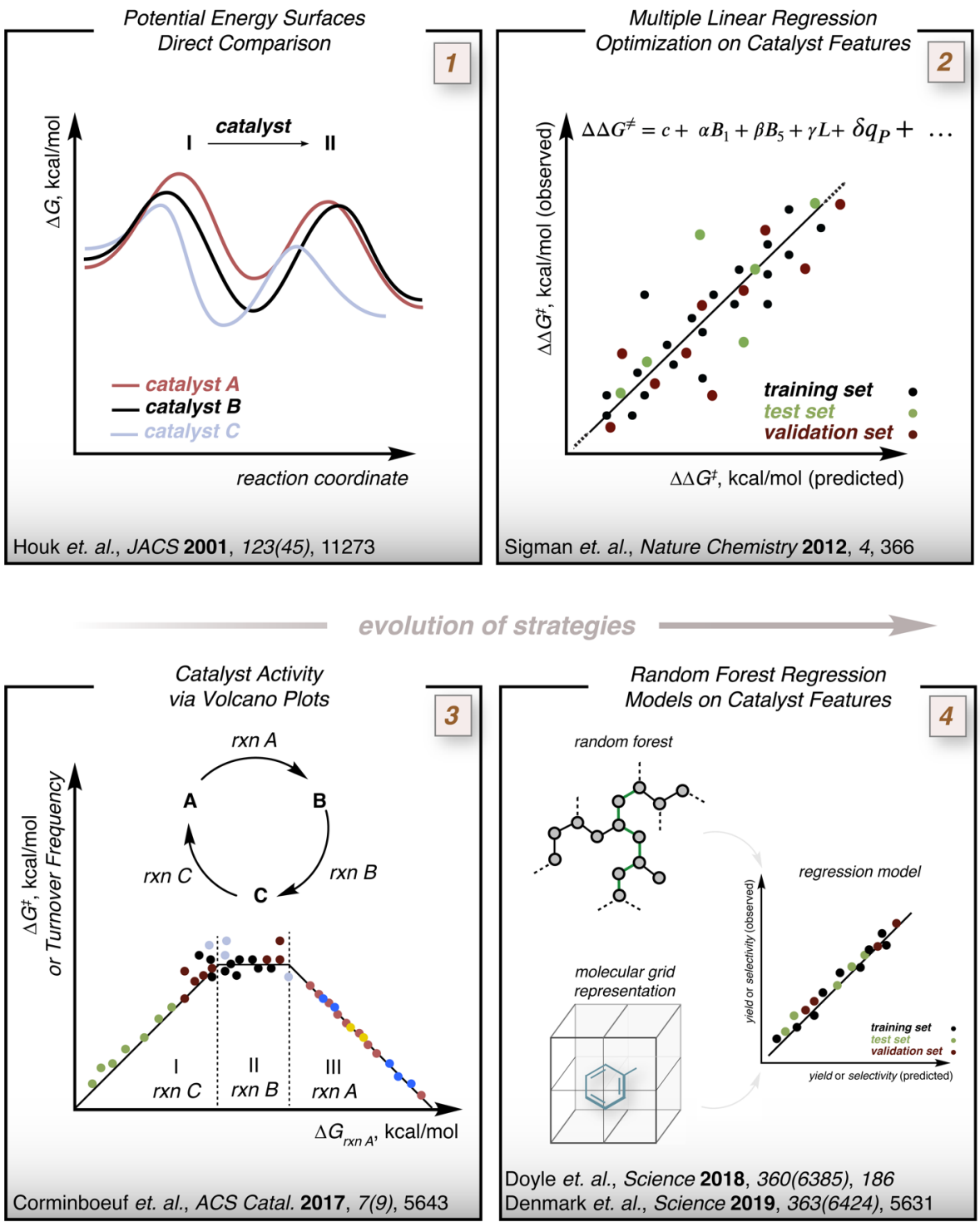

Figure 1. ***Status quo*** **and evolution of computational strategies for catalyst optimization. 1**. Direct comparison of potential energy surfaces for different catalysts, pioneered by the Houk group. **2**. Multiple linear regression optimization on catalyst features, with a particular focus on asymmetric catalysis, popularized by the Sigman group. **3**. The Corminboeuf group has shown how volcano plots can be harnessed for the optimization of catalyst activities. **4**. Random forest regression models have been employed by the Doyle group to predict reaction yields, and the Denmark group to predict enantioselectivity.

These approaches, summarized in Figure 1, outline the *status quo* of computer-aided homogeneous catalyst design and show advances for individual pieces of the overall design workflow puzzle.[32] However, none of them showcase a fully automated closed-loop catalyst design; for this challenge to be achieved, all pieces need to come together in an integrated fashion. This requires the use of automated planning and orchestration tools for both experiments and computations (*vide infra*), as well as algorithms or heuristics, to propose promising modifications to catalyst structures. Furthermore, full integration of experimentation into the design workflow, ideally using automation and robotics, will close the loop and blur the barriers between theory and experiment.

Catalyst Informatics and the Role of Representations

One cornerstone of machine learning is the distinct representations of data, which are used to train models. Most common in chemistry are descriptors, which are usually 1-dimensional real vectors providing information about a given (sub)structure, a classical example being the Hammett substituent constants.[33] In their early days, descriptors were derived by and large from experiments,[34] but in recent decades, the use of (cheap) computed descriptors has seen its heyday.[35] Together with the creation of experimental databases and the use of informatics methods, this led to the emergence of cheminformatics as a new field.[36] These approaches have been applied widely in quantitative structure-activity relationships (QSAR) for the prediction of basic physical properties of molecules, the study of reaction mechanisms, and drug design.[37] While bioinformatics had seen a similar rise already several decades ago,[38] materials informatics only started emerging in the past decade.[39,40] In classical homogeneous catalysis, ligand descriptors have received considerable attention due to their simplicity in classifying structures, rationalizing structure-property relations, and guiding catalyst optimizations.[41] However, only now, the field of catalyst informatics is emerging,[42] lagging far behind bioinformatics for biocatalysis.[38] We believe that the adoption of catalyst informatics will be widespread and rapid, and will have a lasting impact on catalysis in the near future.

From a fundamental point of view, one-dimensional descriptors are information about a molecule, which has many degrees of freedom, reduced in dimensionality almost to the extreme. Hence, most of the information is lost in the process, making them non-ideal when more informative representations are required. The rise of machine learning in chemistry resulted in the use of numerous alternative representations.[5] Typically, they fall into one of three categories, namely discrete, continuous, to which molecular descriptors belong, and graph-based representations.[5] The classic example of discrete representations is SMILES, a text-based encoding of the molecular graph.[43–45] Recently, our lab developed SELFIES as an extremely robust alternative to SMILES, especially for simple and straightforward use in arbitrary machine learning models.[46] Our lab also demonstrated the utility of SELFIES with genetic algorithms as a generative model to explore chemical space systematically and optimize target properties,[47]. We envision this type of approach to have a significant impact in catalysis as well. Another discrete representation that gained popularity for machine learning purposes are molecular fingerprints,[23] as they allow to correlate model predictions directly with structural features, as showcased in a

recent study from our lab on dihydrogen activation with Ir(I) complexes.[48] Alternatively, convolutional neural network models can be trained directly on molecular graphs enjoying the same advantage of interpretability.[49] A recent paper from our group extended these molecular graph-convolutional neural networks to the use of higher-order paths allowing us to account for molecular substructures and geometry.[50] We envision that graph convolutional neural networks will show significant adoption also for the design of new catalysts and will lead to further improvements in model performance in the field.

Data Swamps and Data Lakes

Another cornerstone of machine learning is data. Databases are commonplace in chemistry. Among the most important ones are structure and reaction databases,[51] the Cambridge Structural Database (CSD) for experimental crystal structures,[52] the Protein Data Bank (PBD) for 3D structures of proteins,[53] and numerous databases for both experimental and computed properties of materials.[54–59] However, to the best of our knowledge, databases focusing on homogeneous catalysis are nearly non-existent. Currently, the entirety of scientific and patent literature on homogeneous catalysis is a huge data swamp that is prohibitively tedious to mine as common standards for reporting results are not enforced. However, the need for a factual database of catalysts has been recognized decades ago.[60] We believe that a future database for homogeneous catalysis could be modeled around the standards defined for the Open Reaction Database (ORD),[61] and designing and implementing ontologies will be central.[62] Another database that could serve as a model for catalysis, but is still poorly used in the machine learning community, are the Active Thermochemical Tables (ATcT), which provide experimental thermodynamic data for an interconnected network of molecules.[63–66] Importantly, databases in chemistry, especially in catalysis, need to be expanded to host computational results as they are becoming increasingly important,[12] and will continue to rise in relevance in the future. The need for computational databases has been recognized before,[2] and realizations thereof are *Catalysis-Hub*, which focuses on surface reactions in heterogeneous catalysis,[67] ioChem-BD[68] and QCArchive,[69,70] both databases for storing and analyzing output files of *ab initio* computations. However, comprehensive integration across all sub-disciplines of catalysis is still lacking. Overall, it is abundantly clear that we need to convert the existing data swamps into data lakes for (homogeneous) catalysis, and we need user-friendly platforms to access them to facilitate data mining and enable interactive catalyst design. Consequently, we envision that a comprehensive data lake centered around catalysts and their properties will facilitate future catalyst designs tremendously, both classical human-driven and computer-guided design.

Robust Synthesis and Data-driven Experimentation

Ever since the rise of combinatorial chemistry,[71] high-throughput experimentation has become a standard tool of experimental homogeneous catalysis. Most of these studies rely on the design of experiments (DoE) to guide optimizations.[72] However, recent studies coming from our lab showcased the robustness and efficiency of Bayesian optimizers like Phoenics,[73] Chimera[74], and Gryffin[75] for real-life applications in chemistry. By using ChemOS as an experiment planning

platform,[76,77] both computational and experimental optimization problems have been tackled successfully.[78] These algorithmic developments need to be matched by technological advances, and full experimental workflow implementation will enable closed-loop optimization but is challenging to achieve.[79–81] Accordingly, autonomous closed-loop discovery is the ultimate dream for catalysis and science in general.[81]

It is abundantly clear that the success of data-driven catalyst optimization relies on generating significant amounts of high-quality experimental data providing both structural and quantitative information about the reaction substrates and products.[82] Most notable in that regard are recent developments in mass spectrometry (MS) methods enabling analysis times below 1s per sample allowing to screen a large number of samples essentially in parallel through imaging techniques, providing both structural and semi-quantitative information.[83–86] In that regard, scientists in catalysis should be inspired by the tremendous progress in the field of biochemistry, enabling directed evolution,[87,88] for instance, making use of microfluidics[89] like droplet sorting techniques.[90] Parallelization is another critical requirement in the experimental setup to be able to generate sufficient data in a reasonable timeframe with minimum effort. Digital microfluidic devices are tailor-made for performing a large number of experiments simultaneously.[91]
Furthermore, adjusting experiment design to extract both kinetic information (related to the catalyst turnover frequency) and yield information (related to the catalyst turnover number and provides information about catalyst stability), will also be paramount to gather most information with the least effort.

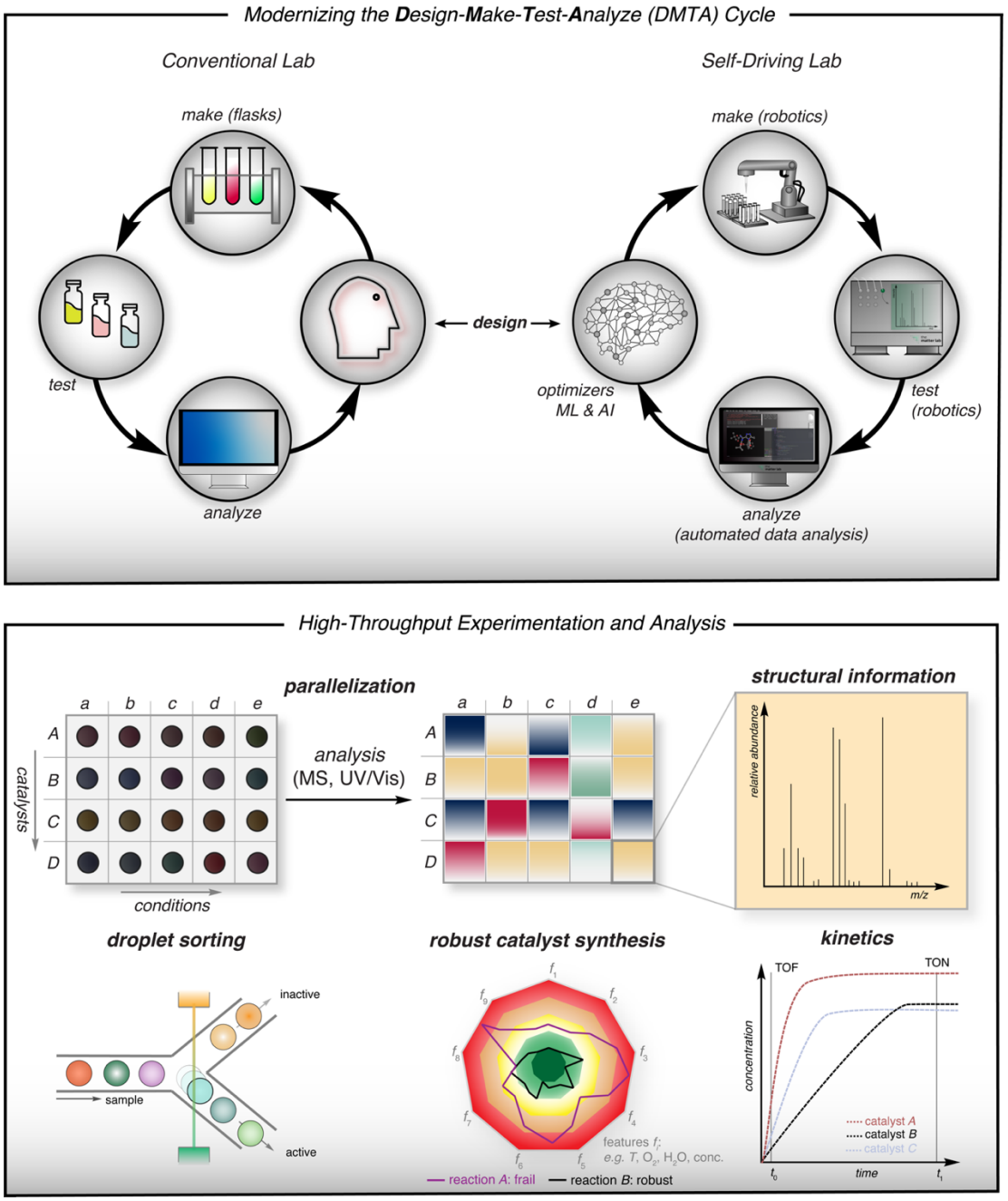

Figure 2: **Current and Future Paradigm in Catalysis Experimentation.** Top: Modernizing the Design-Make-Test-Analyze cycle. A comparison between conventional and self-driving labs. While both start at the catalyst/reaction design, the self-driving lab counts with robotics for its make and test steps. Additionally, data analysis processes are performed in an automated fashion. Powered by machine-learning algorithms, this approach tries to close the loop in autonomous experimentation. Bottom: High-Throughput Experimentation and Analysis in Catalysis. Modern catalyst optimizations need to rely on screenings parallelization and imaging. Analysis techniques for increasing throughput include droplet sorting. Robustness of catalyst synthesis can be assessed by quantification of sensitivity for perturbations of reaction conditions and for contaminations of ambient chemicals like oxygen or water. Kinetic performance metrics like turnover frequency (TOF) or turnover number (TON) need to be extracted from these analyses to facilitate the performance assessment of catalysts.

Nevertheless, the main bottleneck of catalyst optimization studies is usually, by a large margin, catalyst synthesis. Hence, not only is it necessary to streamline the catalytic experiments themselves and their analyses, but catalyst synthesis needs to be rethought from the ground up for efficient closed-loop optimization studies.[92] Typically, the catalysts that can be synthesized are limited by the scope and robustness of the synthetic procedures available.

This is a call for action to not only develop new catalytic reactions but also make the experimental protocols robust and highly reproducible. One step towards that goal is the widespread use of previously proposed standardized robustness screens to establish a common baseline for comparing methodologies, which are only rarely applied.[93–96] We believe that the use of a common baseline will lead to improved fidelity in the assessment of reaction robustness and, ultimately, facilitate the design of enhanced catalysts. Improved catalytic reactions will streamline the synthesis of new catalysts making novel developments possible.[97] In that sense, catalysis is autocatalytic for the development of new catalysts.

New Ideas and Paradigms

At present, progress in machine learning and artificial intelligence comes at an astonishing pace, and it usually takes time for the most recent developments to enter other fields. One of the outstanding challenges is realizing explainable artificial intelligence, also referred to as the interpretability problem.[98–100] The current black-box nature of many machine learning approaches is unsatisfying, as Eugene Wigner said:[101] "*It is nice to know that the computer understands the problem. But I would like to understand it too.*" Accordingly, the importance of the interpretability for machine learning models in chemistry has been outlined before.[102] One pathway towards inherently explainable artificial intelligence could be the re-emergence of symbolic artificial intelligence.[103] In chemistry, interpretability goes hand-in-hand with the representation of a given problem.[104] Hence, the path towards explainable artificial intelligence naturally leads to rethinking representations used in the models. In that regard, the extensive use of descriptors is somewhat unsatisfying as it can lead to the exploitation of hidden correlations. From a quantum mechanical perspective, the direct use of molecular wavefunctions or molecular electron densities, or systematic simplifications thereof, would be most appealing. In that framework, machine learning models represent the operators acting on the wavefunctions, *i.e.*, the molecular representations, to deliver the corresponding observables.[105–107] Overall, we believe that the development of explainable artificial intelligence in chemistry will lead to enhanced human understanding, inspire hybrid human- and computer-guided catalyst design, and ultimately lead to improved machine learning models.

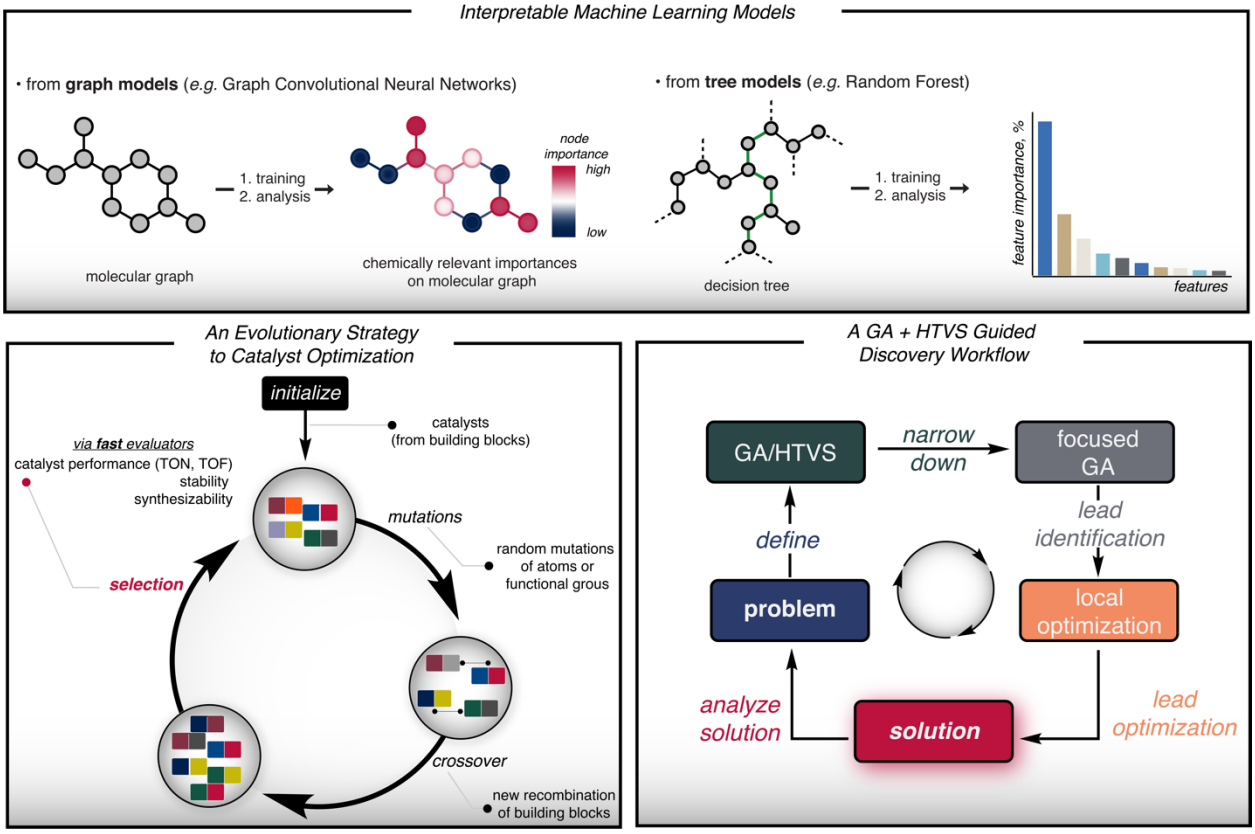

Figure 3: **New strategies for catalyst optimizations with machine learning**. Top: interpretable machine learning models enable chemists to rapidly make predictions of new catalysts. Additionally, it remediates the reasonable dissatisfaction with lack of understanding behind black-box models. Bottom left: Evolving catalysts with a genetic algorithm (GA) counts on random mutations and crossover on molecular structure and diversity. Appropriate catalysts can be selected by their performance and stability. Bottom right: A combined workflow of a generative model (GA) and high-throughput virtual screening (HTVS) for systematic exploration of chemical space and thorough optimization of catalysts of interest.

Other notable developments in machine learning that are progressively entering the center stage in chemistry are generative models.[108–110] In the context of deep learning, generative models have entered the public discussion via so-called "deep fakes," which are images, sound, or other media that are entirely computer-generated but appear realistic because of their far-reaching implications in society.[111] In chemistry, the idea of generative models is to propose new molecular structures with specific target properties. This is perfectly suited for exploratory optimization problems without restricting the molecular space. Recently, our lab developed a workflow using a genetic algorithm as the generative model for the systematic exploration of chemical space looking for molecules with specific target properties (see Figure 3 for a schematic overview of this approach).[47] We believe that this kind of workflow has enormous potential when applied systematically to catalyst design, with significant advances in that area shortly. Moreover, another powerful concept for generative models is deep reinforcement learning,[112] and this has been demonstrated recently by the tremendous advances of *AlphaZero*

to master the games Go, shogi, and chess.[113] The idea is to formulate the molecular design problem in terms of a Markov decision process,[114] *i.e.*, every step of this process begins with a starting molecule and chooses from several allowed structural modifications to obtain a new molecule. The new molecule is evaluated in its performance, providing a reward for the decision taken. This step is repeated until a molecule with the desired properties is obtained, or the maximum number of steps is reached. Notably, this algorithm closely resembles the human approach to molecular design. Recently, this kind of workflow has been demonstrated in chemistry in the molecular design task and tested in benchmark molecular optimization tasks.[115] Alternative successful implementations of reinforcement learning include optimization of reaction conditions for product yields[116] and for polymer molecular weight distributions.[117] To the best of our knowledge, deep reinforcement learning has not been applied to catalyst design workflows, and we foresee a vast number of potential applications in the coming decade.

Our Vision as a Maze

The path towards autonomous catalyst discovery is far from linear, as many designs and implementation choices remain to be decided. Accordingly, we view the future of homogeneous catalyst design as a maze (Figure 4). We incorporated what we envision to be important milestones as forks along the path. However, while Figure 4 depicts only one path towards the center of the maze, as the proverb goes, "All roads lead to Rome," we believe that there are many viable paths ahead towards this goal.

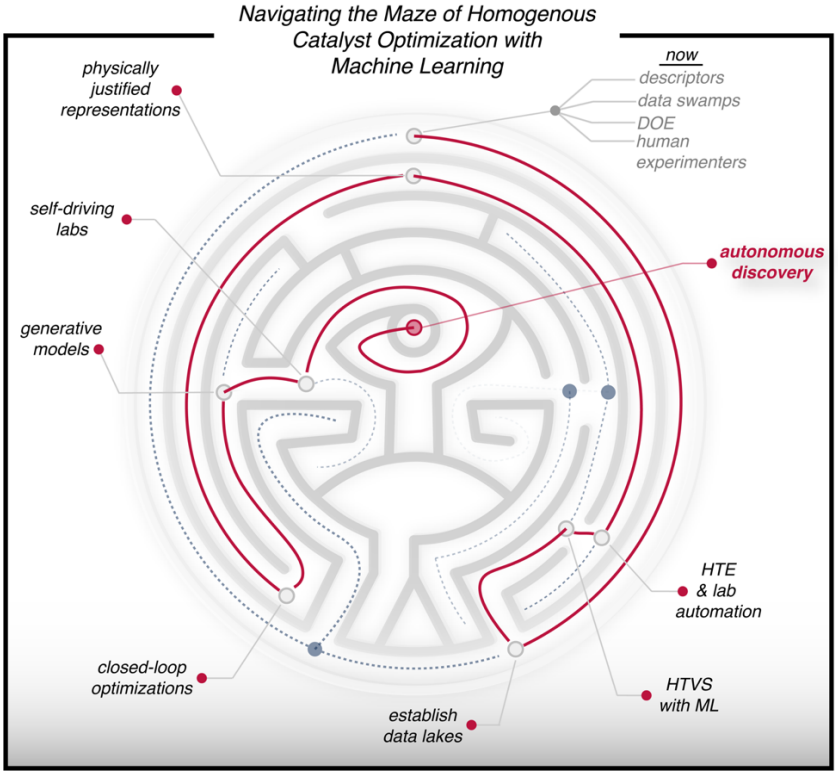

Figure 4: **Navigating the maze of homogeneous catalyst optimization with machine learning.** From the data swamps and descriptor-based models to closed-loop optimizations and, ultimately, the autonomous discovery of catalysts.

Achieving this goal requires the close collaboration of scientists across many disciplines and will only be accomplishable as an interdisciplinary endeavor. However, establishing autonomous catalyst discovery is not a self-serving research goal. Ultimately, it needs to deliver the catalysts of the future and solve some of the greatest challenges that humanity is facing in the coming decades, including climate catastrophe and environmental pollution. We believe that machine learning is the most promising way forward to explore the chemical space at an unprecedented pace and increase the rate of discovery significantly.
We are looking forward to all the exciting advances the field will experience in the years to come.

Acknowledgments
G. P. G gratefully acknowledges the Natural Sciences and Engineering Research Council of Canada (NSERC) for the Banting Postdoctoral Fellowship. We acknowledge the Defense Advanced Research Projects Agency (DARPA) under the Accelerated Molecular Discovery Program under Cooperative Agreement No. HR00111920027, dated August 1, 2019. The content of the information presented in this work does not necessarily reflect the position or the policy of the Government. A. A.-G. thanks Anders G. Frøseth for his generous support. A. A.-G. also acknowledges the generous support of Natural Resources Canada and the Canada 150 Research Chairs program. We also acknowledge the Department of Navy award (N00014-19-1-2134) issued by the Office of Naval Research. The United States Government has a royalty-free license throughout the world in all copyrightable material contained herein. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Office of Naval Research.

Disclaimer Statement
A.A.-G. is a co-founder and the Chief Visionary Officer of Kebotix, Inc.

References and Notes
(1) Chemical Industry and Homogeneous Catalysis. In *Homogeneous Catalysis*; John Wiley & Sons, Ltd, 2014; pp 1–21.
(2) Thomas, J. M. Summarizing Comments on the Discussion and a Prospectus for Urgent Future Action. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **2016**, *374*, 20150226.
(3) Levi, P. G.; Cullen, J. M. Mapping Global Flows of Chemicals: From Fossil Fuel Feedstocks to Chemical Products. *Environ. Sci. Technol.* **2018**, *52*, 1725–1734.
(4) Council, N. R. *Impact of Advances in Computing and Communications Technologies on Chemical Science and Technology: Report of a Workshop*; 1999.
(5) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse Molecular Design Using Machine Learning: Generative Models for Matter Engineering. *Science* **2018**, *361*, 360–365.
(6) Homogeneously Catalyzed Industrial Processes. In *Industrial Catalysis*; John Wiley & Sons, Ltd, 2015; pp 47–80.
(7) Franke, R.; Selent, D.; Börner, A. Applied Hydroformylation. *Chem. Rev.* **2012**, *112*, 5675–5732.
(8) Keith, J. A.; Henry, P. M. The Mechanism of the Wacker Reaction: A Tale of Two Hydroxypalladations. *Angewandte Chemie International Edition* **2009**, *48*, 9038–9049.

(9) Miyaura, Norio.; Suzuki, Akira. Palladium-Catalyzed Cross-Coupling Reactions of Organoboron Compounds. *Chem. Rev.* **1995**, *95*, 2457–2483.
(10) Aspuru-Guzik, A.; Lindh, R.; Reiher, M. The Matter Simulation (R)Evolution. *ACS Cent. Sci.* **2018**, *4*, 144–152.
(11) Houk, K. N.; Liu, F. Holy Grails for Computational Organic Chemistry and Biochemistry. *Acc. Chem. Res.* **2017**, *50*, 539–543.
(12) Houk, K. N.; Cheong, P. H.-Y. Computational Prediction of Small-Molecule Catalysts. *Nature* **2008**, *455*, 309–313.
(13) Bahmanyar, S.; Houk, K. N. Transition States of Amine-Catalyzed Aldol Reactions Involving Enamine Intermediates: Theoretical Studies of Mechanism, Reactivity, and Stereoselectivity. *J. Am. Chem. Soc.* **2001**, *123*, 11273–11283.
(14) Guan, Y.; Ingman, V. M.; Rooks, B. J.; Wheeler, S. E. AARON: An Automated Reaction Optimizer for New Catalysts. *J. Chem. Theory Comput.* **2018**, *14*, 5249–5261.
(15) Sigman, M. S.; Harper, K. C.; Bess, E. N.; Milo, A. The Development of Multidimensional Analysis Tools for Asymmetric Catalysis and Beyond. *Acc. Chem. Res.* **2016**, *49*, 1292–1301.
(16) Santiago, C. B.; Guo, J.-Y.; Sigman, M. S. Predictive and Mechanistic Multivariate Linear Regression Models for Reaction Development. *Chem. Sci.* **2018**, *9*, 2398–2412.
(17) Reid, J. P.; Sigman, M. S. Comparing Quantitative Prediction Methods for the Discovery of Small-Molecule Chiral Catalysts. *Nature Reviews Chemistry* **2018**, *2*, 290–305.
(18) Strieth-Kalthoff, F.; Sandfort, F.; Segler, M. H. S.; Glorius, F. Machine Learning the Ropes: Principles, Applications and Directions in Synthetic Chemistry. *Chem. Soc. Rev.* **2020**.
(19) Harper, K. C.; Bess, E. N.; Sigman, M. S. Multidimensional Steric Parameters in the Analysis of Asymmetric Catalytic Reactions. *Nature Chemistry* **2012**, *4*, 366–374.
(20) Milo, A.; Bess, E. N.; Sigman, M. S. Interrogating Selectivity in Catalysis Using Molecular Vibrations. *Nature* **2014**, *507*, 210–214.
(21) Orlandi, M.; Coelho, J. A. S.; Hilton, M. J.; Toste, F. D.; Sigman, M. S. Parametrization of Non-Covalent Interactions for Transition State Interrogation Applied to Asymmetric Catalysis. *J. Am. Chem. Soc.* **2017**, *139*, 6803–6806.
(22) 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.
(23) 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*, eaau5631.
(24) Rosales, A. R.; Wahlers, J.; Limé, E.; Meadows, R. E.; Leslie, K. W.; Savin, R.; Bell, F.; Hansen, E.; Helquist, P.; Munday, R. H.; et al. Rapid Virtual Screening of Enantioselective Catalysts Using CatVS. *Nature Catalysis* **2019**, *2*, 41–45.
(25) Jorner, K.; Brinck, T.; Norrby, P.-O.; Buttar, D. Machine Learning Meets Mechanistic Modelling for Accurate Prediction of Experimental Activation Energies. **2020**.
(26) Sabatier, P. Hydrogénations et Déshydrogénations Par Catalyse. *Berichte der deutschen chemischen Gesellschaft* **1911**, *44*, 1984–2001.
(27) Balandin, A. A. Modern State of the Multiplet Theory of Heterogeneous Catalysis. In *Advances in Catalysis*; Eley, D. D., Pines, H., Weisz, P. B., Eds.; Academic Press, 1969; Vol. 19, pp 1–210.
(28) Busch, M.; Wodrich, M. D.; Corminboeuf, C. A Generalized Picture of C–C Cross-Coupling. *ACS Catal.* **2017**, *7*, 5643–5653.
(29) Wodrich, M. D.; Sawatlon, B.; Busch, M.; Corminboeuf, C. On the Generality of Molecular Volcano Plots. *ChemCatChem* **2018**, *10*, 1586–1591.
(30) Meyer, B.; Sawatlon, B.; Heinen, S.; Lilienfeld, O. A. von; Corminboeuf, C. Machine Learning Meets Volcano Plots: Computational Discovery of Cross-Coupling Catalysts. *Chem. Sci.* **2018**, *9*, 7069–7077.
(31) Wodrich, M. D.; Sawatlon, B.; Solel, E.; Kozuch, S.; Corminboeuf, C. Activity-Based Screening of Homogeneous Catalysts through the Rapid Assessment of Theoretically Derived Turnover Frequencies. *ACS Catal.* **2019**, *9*, 5716–5725.
(32) Foscato, M.; Jensen, V. R. Automated in Silico Design of Homogeneous Catalysts. *ACS Catal.*

2020, *10*, 2354–2377.
(33) Hammett, L. P. The Effect of Structure upon the Reactions of Organic Compounds. Benzene Derivatives. *J. Am. Chem. Soc.* **1937**, *59*, 96–103.
(34) Hansch, Corwin.; Leo, A.; Taft, R. W. A Survey of Hammett Substituent Constants and Resonance and Field Parameters. *Chem. Rev.* **1991**, *91*, 165–195.
(35) *Handbook of Molecular Descriptors*, 1st ed.; John Wiley & Sons, Ltd, 2000.
(36) Engel, T. Basic Overview of Chemoinformatics. *J. Chem. Inf. Model.* **2006**, *46*, 2267–2277.
(37) 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.; et al. QSAR without Borders. *Chem. Soc. Rev.* **2020**, *49*, 3525–3564.
(38) Ouzounis, C. A.; Valencia, A. Early Bioinformatics: The Birth of a Discipline—a Personal View. *Bioinformatics* **2003**, *19*, 2176–2190.
(39) Agrawal, A.; Choudhary, A. Perspective: Materials Informatics and Big Data: Realization of the "Fourth Paradigm" of Science in Materials Science. *APL Materials* **2016**, *4*, 053208.
(40) Takahashi, K.; Tanaka, Y. Materials Informatics: A Journey towards Material Design and Synthesis. *Dalton Trans.* **2016**, *45*, 10497–10499.
(41) Durand, D. J.; Fey, N. Computational Ligand Descriptors for Catalyst Design. *Chem. Rev.* **2019**, *119*, 6561–6594.
(42) Takahashi, K.; Takahashi, L.; Miyazato, I.; Fujima, J.; Tanaka, Y.; Uno, T.; Satoh, H.; Ohno, K.; Nishida, M.; Hirai, K.; et al. The Rise of Catalyst Informatics: Towards Catalyst Genomics. *ChemCatChem* **2019**, *11*, 1146–1152.
(43) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.
(44) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97–101.
(45) Weininger, D. SMILES. 3. DEPICT. Graphical Depiction of Chemical Structures. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 237–243.
(46) Krenn, M.; Häse, F.; Nigam, A.; Friederich, P.; Aspuru-Guzik, A. Self-Referencing Embedded Strings (SELFIES): A 100% Robust Molecular String Representation. *arXiv:1905.13741 [physics, physics:quant-ph, stat]* **2020**.
(47) Nigam, A.; Friederich, P.; Krenn, M.; Aspuru-Guzik, A. Augmenting Genetic Algorithms with Deep Neural Networks for Exploring the Chemical Space. *arXiv:1909.11655 [physics]* **2020**.
(48) Friederich, P.; Gomes, G. dos P.; Bin, R. D.; Aspuru-Guzik, A.; Balcells, D. Machine Learning Dihydrogen Activation in the Chemical Space Surrounding Vaska's Complex. *Chem. Sci.* **2020**, *11*, 4584–4601.
(49) Duvenaud, D.; Maclaurin, D.; Aguilera-Iparraguirre, J.; Gómez-Bombarelli, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. *arXiv:1509.09292 [cs, stat]* **2015**.
(50) Flam-Shepherd, D.; Wu, T.; Friederich, P.; Aspuru-Guzik, A. Neural Message Passing on High Order Paths. *arXiv:2002.10413 [cs, stat]* **2020**.
(51) Papadakis, E.; Anantpinijwatna, A.; Woodley, J. M.; Gani, R. A Reaction Database for Small Molecule Pharmaceutical Processes Integrated with Process Information. *Processes* **2017**, *5*, 58.
(52) Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge Structural Database. *Acta Cryst B* **2016**, *72*, 171–179.
(53) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res* **2000**, *28*, 235–242.
(54) MaterialsGenome, Inc http://www.materialsgenome.com/ (accessed Aug 6, 2020).
(55) Liu, Z. Perspective on Materials Genome®. *Chin. Sci. Bull.* **2014**, *59*, 1619–1623.
(56) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; et al. Commentary: The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation. *APL Materials* **2013**, *1*, 011002.
(57) Saal, J. E.; Kirklin, S.; Aykol, M.; Meredig, B.; Wolverton, C. Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD).

JOM **2013**, *65*, 1501–1509.
(58) Kirklin, S.; Saal, J. E.; Meredig, B.; Thompson, A.; Doak, J. W.; Aykol, M.; Rühl, S.; Wolverton, C. The Open Quantum Materials Database (OQMD): Assessing the Accuracy of DFT Formation Energies. *npj Computational Materials* **2015**, *1*, 1–15.
(59) Curtarolo, S.; Setyawan, W.; Hart, G. L. W.; Jahnatek, M.; Chepulskii, R. V.; Taylor, R. H.; Wang, S.; Xue, J.; Yang, K.; Levy, O.; et al. AFLOW: An Automatic Framework for High-Throughput Materials Discovery. *Computational Materials Science* **2012**, *58*, 218–226.
(60) Ito, T.; Hamada, H.; Kintaichi, Y.; Sasaki, M. Database for Catalysis Design. *Catalysis Today* **1991**, *10*, 223–232.
(61) Overview — Open Reaction Database documentation https://ord-schema.readthedocs.io/en/latest/overview.html (accessed Jul 3, 2020).
(62) Takahashi, L.; Miyazato, I.; Takahashi, K. Redesigning the Materials and Catalysts Database Construction Process Using Ontologies. *J. Chem. Inf. Model.* **2018**, *58*, 1742–1754.
(63) Active Thermochemical Tables - Home https://atct.anl.gov/ (accessed Jul 3, 2020).
(64) Ruscic, B.; Pinzon, R. E.; Laszewski, G. von; Kodeboyina, D.; Burcat, A.; Leahy, D.; Montoy, D.; Wagner, A. F. Active Thermochemical Tables: Thermochemistry for the 21st Century. *J. Phys.: Conf. Ser.* **2005**, *16*, 561–570.
(65) Ruscic, B.; Pinzon, R. E.; Morton, M. L.; von Laszevski, G.; Bittner, S. J.; Nijsure, S. G.; Amin, K. A.; Minkoff, M.; Wagner, A. F. Introduction to Active Thermochemical Tables: Several "Key" Enthalpies of Formation Revisited. *J. Phys. Chem. A* **2004**, *108*, 9979–9997.
(66) Ruscic, B. Uncertainty Quantification in Thermochemistry, Benchmarking Electronic Structure Computations, and Active Thermochemical Tables. *International Journal of Quantum Chemistry* **2014**, *114*, 1097–1101.
(67) Winther, K. T.; Hoffmann, M. J.; Boes, J. R.; Mamun, O.; Bajdich, M.; Bligaard, T. Catalysis-Hub.Org, an Open Electronic Structure Database for Surface Reactions. *Scientific Data* **2019**, *6*, 75.
(68) Álvarez-Moreno, M.; de Graaf, C.; López, N.; Maseras, F.; Poblet, J. M.; Bo, C. Managing the Computational Chemistry Big Data Problem: The IoChem-BD Platform. *J. Chem. Inf. Model.* **2015**, *55*, 95–103.
(69) Smith, D. G. A.; Altarawy, D.; Burns, L. A.; Welborn, M.; Naden, L. N.; Ward, L.; Ellis, S.; Pritchard, B. P.; Crawford, T. D. The MolSSI QCArchive Project: An Open-Source Platform to Compute, Organize, and Share Quantum Chemistry Data. *WIREs Computational Molecular Science* e1491.
(70) The MolSSI QCArchive https://qcarchive.molssi.org/about/ (accessed Aug 10, 2020).
(71) Szostak, J. W. Introduction: Combinatorial Chemistry. *Chem. Rev.* **1997**, *97*, 347–348.
(72) Weissman, S. A.; Anderson, N. G. Design of Experiments (DoE) and Process Optimization. A Review of Recent Publications. *Org. Process Res. Dev.* **2015**, *19*, 1605–1633.
(73) Häse, F.; Roch, L. M.; Kreisbeck, C.; Aspuru-Guzik, A. Phoenics: A Bayesian Optimizer for Chemistry. *ACS Cent. Sci.* **2018**, *4*, 1134–1145.
(74) Häse, F.; Roch, L. M.; Aspuru-Guzik, A. Chimera: Enabling Hierarchy Based Multi-Objective Optimization for Self-Driving Laboratories. *Chem. Sci.* **2018**, *9*, 7642–7655.
(75) Häse, F.; Roch, L. M.; Aspuru-Guzik, A. Gryffin: An Algorithm for Bayesian Optimization for Categorical Variables Informed by Physical Intuition with Applications to Chemistry. *arXiv:2003.12127 [physics, stat]* **2020**.
(76) Roch, L. M.; Häse, F.; Kreisbeck, C.; Tamayo-Mendoza, T.; Yunker, L. P. E.; Hein, J. E.; Aspuru-Guzik, A. ChemOS: Orchestrating Autonomous Experimentation. *Science Robotics* **2018**, *3*, eaat5559.
(77) Roch, L. M.; Häse, F.; Kreisbeck, C.; Tamayo-Mendoza, T.; Yunker, L. P. E.; Hein, J. E.; Aspuru-Guzik, A. ChemOS: An Orchestration Software to Democratize Autonomous Discovery. *PLOS ONE* **2020**, *15*, e0229862.
(78) MacLeod, B. P.; Parlane, F. G. L.; Morrissey, T. D.; Häse, F.; Roch, L. M.; Dettelbach, K. E.; Moreira, R.; Yunker, L. P. E.; Rooney, M. B.; Deeth, J. R.; et al. Self-Driving Laboratory for Accelerated Discovery of Thin-Film Materials. *Science Advances* **2020**, *6*, eaaz8867.
(79) Häse, F.; Roch, L. M.; Aspuru-Guzik, A. Next-Generation Experimentation with Self-Driving

Laboratories. *Trends in Chemistry* **2019**, *1*, 282–291.
(80) Burger, B.; Maffettone, P. M.; Gusev, V. V.; Aitchison, C. M.; Bai, Y.; Wang, X.; Li, X.; Alston, B. M.; Li, B.; Clowes, R.; et al. A Mobile Robotic Chemist. *Nature* **2020**, *583*, 237–241.
(81) Flores-Leonar, M. M.; Mejía-Mendoza, L. M.; Aguilar-Granda, A.; Sanchez-Lengeling, B.; Tribukait, H.; Amador-Bedolla, C.; Aspuru-Guzik, A. Materials Acceleration Platforms: On the Way to Autonomous Experimentation. *Current Opinion in Green and Sustainable Chemistry* **2020**, 100370.
(82) Shevlin, M. Practical High-Throughput Experimentation for Chemists. *ACS Med. Chem. Lett.* **2017**, *8*, 601–607.
(83) Wang, Y.; Shaabani, S.; Ahmadianmoghaddam, M.; Gao, L.; Xu, R.; Kurpiewska, K.; Kalinowska-Tluscik, J.; Olechno, J.; Ellson, R.; Kossenjans, M.; et al. Acoustic Droplet Ejection Enabled Automated Reaction Scouting. *ACS Cent. Sci.* **2019**, *5*, 451–457.
(84) DiRico, K. J.; Hua, W.; Liu, C.; Tucker, J. W.; Ratnayake, A. S.; Flanagan, M. E.; Troutman, M. D.; Noe, M. C.; Zhang, H. Ultra-High-Throughput Acoustic Droplet Ejection-Open Port Interface-Mass Spectrometry for Parallel Medicinal Chemistry. *ACS Med. Chem. Lett.* **2020**.
(85) Wleklinski, M.; Loren, B. P.; Ferreira, C. R.; Jaman, Z.; Avramova, L.; Sobreira, T. J. P.; Thompson, D. H.; Cooks, R. G. High Throughput Reaction Screening Using Desorption Electrospray Ionization Mass Spectrometry. *Chem. Sci.* **2018**, *9*, 1647–1653.
(86) Lin, S.; Dikler, S.; Blincoe, W. D.; Ferguson, R. D.; Sheridan, R. P.; Peng, Z.; Conway, D. V.; Zawatzky, K.; Wang, H.; Cernak, T.; et al. Mapping the Dark Space of Chemical Reactions with Extended Nanomole Synthesis and MALDI-TOF MS. *Science* **2018**, *361*.
(87) Arnold, F. H. Directed Evolution: Bringing New Chemistry to Life. *Angewandte Chemie International Edition* **2018**, *57*, 4143–4148.
(88) Arnold, F. H. Innovation by Evolution: Bringing New Chemistry to Life (Nobel Lecture). *Angewandte Chemie International Edition* **2019**, *58*, 14420–14426.
(89) Whitesides, G. M. The Origins and the Future of Microfluidics. *Nature* **2006**, *442*, 368–373.
(90) Chiu, F. W. Y.; Stavrakis, S. High-Throughput Droplet-Based Microfluidics for Directed Evolution of Enzymes. *ELECTROPHORESIS* **2019**, *40*, 2860–2872.
(91) Torabinia, M.; Asgari, P.; Dakarapu, U. S.; Jeon, J.; Moon, H. On-Chip Organic Synthesis Enabled Using an Engine-and-Cargo System in an Electrowetting-on-Dielectric Digital Microfluidic Device. *Lab Chip* **2019**, *19*, 3054–3064.
(92) Renom-Carrasco, M.; Lefort, L. Ligand Libraries for High Throughput Screening of Homogeneous Catalysts. *Chem. Soc. Rev.* **2018**, *47*, 5038–5060.
(93) Collins, K. D.; Glorius, F. A Robustness Screen for the Rapid Assessment of Chemical Reactions. *Nature Chemistry* **2013**, *5*, 597–601.
(94) Collins, K. D.; Rühling, A.; Glorius, F. Application of a Robustness Screen for the Evaluation of Synthetic Organic Methodology. *Nature Protocols* **2014**, *9*, 1348–1353.
(95) Gensch, T.; Teders, M.; Glorius, F. Approach to Comparing the Functional Group Tolerance of Reactions. *J. Org. Chem.* **2017**, *82*, 9154–9159.
(96) Pitzer, L.; Schäfers, F.; Glorius, F. Rapid Assessment of the Reaction-Condition-Based Sensitivity of Chemical Transformations. *Angewandte Chemie International Edition* **2019**, *58*, 8572–8576.
(97) Schultz, D.; Campeau, L.-C. Harder, Better, Faster. *Nature Chemistry* **2020**, 1–4.
(98) Rai, A. Explainable AI: From Black Box to Glass Box. *J. of the Acad. Mark. Sci.* **2020**, *48*, 137–141.
(99) Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Information Fusion* **2020**, *58*, 82–115.
(100) Carvalho, D. V.; Pereira, E. M.; Cardoso, J. S. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics* **2019**, *8*, 832.
(101) Heller, E. J.; Tomsovic, S. Postmodern Quantum Mechanics. *Physics Today* **1993**, *46*, 38.
(102) Haghighatlari, M.; Li, J.; Heidar-Zadeh, F.; Liu, Y.; Guan, X.; Head-Gordon, T. Learning to Make Chemical Predictions: The Interplay of Feature Representation, Data, and Machine Learning Methods. *Chem* **2020**, *6*, 1527–1542.
(103) Garnelo, M.; Shanahan, M. Reconciling Deep Learning with Symbolic Artificial Intelligence:

Representing Objects and Relations. *Current Opinion in Behavioral Sciences* **2019**, *29*, 17–23.
(104) Lambard, G.; Gracheva, E. SMILES-X: Autonomous Molecular Compounds Characterization for Small Datasets without Descriptors. *Mach. Learn.: Sci. Technol.* **2020**, *1*, 025004.
(105) Ramakrishnan, R.; von Lilienfeld, O. A. Many Molecular Properties from One Kernel in Chemical Space. *CHIMIA International Journal for Chemistry* **2015**, *69*, 182–186.
(106) Ramakrishnan, R.; Lilienfeld, O. A. von. Machine Learning, Quantum Chemistry, and Chemical Space. In *Reviews in Computational Chemistry*; John Wiley & Sons, Ltd, 2017; pp 225–256.
(107) Christensen, A. S.; Faber, F. A.; von Lilienfeld, O. A. Operators in Quantum Machine Learning: Response Properties in Chemical Space. *J. Chem. Phys.* **2019**, *150*, 064105.
(108) Lim, J.; Hwang, S.-Y.; Moon, S.; Kim, S.; Kim, W. Y. Scaffold-Based Molecular Design with a Graph Generative Model. *Chem. Sci.* **2020**, *11*, 1153–1164.
(109) Maziarka, Ł.; Pocha, A.; Kaczmarczyk, J.; Rataj, K.; Danel, T.; Warchoł, M. Mol-CycleGAN: A Generative Model for Molecular Optimization. *Journal of Cheminformatics* **2020**, *12*, 2.
(110) Schwalbe-Koda, D.; Gómez-Bombarelli, R. Generative Models for Automatic Chemical Design. In *Machine Learning Meets Quantum Physics*; Schütt, K. T., Chmiela, S., von Lilienfeld, O. A., Tkatchenko, A., Tsuda, K., Müller, K.-R., Eds.; Lecture Notes in Physics; Springer International Publishing: Cham, 2020; pp 445–467.
(111) Westerlund, M. The Emergence of Deepfake Technology: A Review. *Technology Innovation Management Review* **2019**, *9*, 40–53.
(112) Li, Y. Deep Reinforcement Learning. *arXiv:1810.06339 [cs, stat]* **2018**.
(113) Silver, D.; Hubert, T.; Schrittwieser, J.; Antonoglou, I.; Lai, M.; Guez, A.; Lanctot, M.; Sifre, L.; Kumaran, D.; Graepel, T.; et al. A General Reinforcement Learning Algorithm That Masters Chess, Shogi, and Go through Self-Play. *Science* **2018**, *362*, 1140–1144.
(114) Bellman, R. A Markovian Decision Process. *Journal of Mathematics and Mechanics* **1957**, *6*, 679–684.
(115) Zhou, Z.; Kearnes, S.; Li, L.; Zare, R. N.; Riley, P. Optimization of Molecules via Deep Reinforcement Learning. *Scientific Reports* **2019**, *9*, 10752.
(116) Zhou, Z.; Li, X.; Zare, R. N. Optimizing Chemical Reactions with Deep Reinforcement Learning. *ACS Cent. Sci.* **2017**, *3*, 1337–1344.
(117) Li, H.; R. Collins, C.; G. Ribelli, T.; Matyjaszewski, K.; J. Gordon, G.; Kowalewski, T.; J. Yaron, D. Tuning the Molecular Weight Distribution from Atom Transfer Radical Polymerization Using Deep Reinforcement Learning. *Molecular Systems Design & Engineering* **2018**, *3*, 496–508.