Contents

[Outline: 1](#_Toc164764232)

[Organised Outline 2](#_Toc164764233)

[Interesting Quotes and ideas from automation papers. 2](#_Toc164764234)

[Artificial Curiosity: A Search for Novel materials and Chemical Understanding. 5](#_Toc164764235)

[Main 5](#_Toc164764236)

[Combinatorial Chemistry 6](#_Toc164764237)

[Workflow 8](#_Toc164764238)

[Workflow Description 9](#_Toc164764239)

[References 12](#_Toc164764240)

# Outline:

* Hypothesis: it is possible to learn new chemistry from a robotic ML platform.
* Definition of chemistry and how it defines the search space.
* Removal of bias from sampling all combination reactions.
* Bias of decision maker for single, discrete, diamagnetic supramolecular structures.
* Add reagents used for generation of sample space.
* The choice of reagents were based on: bite angle, rigidity, aromaticity, chirality, sterics, metal valency, and metal size.
* Strength of using ML systems: analysis of huge amounts of data and come to unintuitive decisions.
* Supramolecular chemistry is chosen due to its combinatorics, complexity, easy reaction conidiations to automate, reactions being in equilibrium make single discrete species making it easier to be analysed by NMR.
* The advantage of automation comes with its organisation, (and a bit of saving human effort). Humans get tired and bored and will therefore mess up making up tens of different reactions. Robotic platforms do not have this issue and follow steps to the dot.
* The choice of features for model.
* What is SHAPS and how can it be used for inference (if smiles representations are used, SHAPS will highlight what part of the molecule has the greatest say in prediction).
* The physical setup of the platform.
* How the workflow operates.
* Discovery of Novel structures via a more reasoning approach. If this proof of concept works, the paper would not only have provided insights into the chemistry, but to the fact that ML can infact learn chemical rules. The ML can then be trained on a subset of the combination space, and then predict chemical structures for the remainder. This allows for the discovery of novel compounds without neccasrly having to carry out reactions.
* Robotic platforms allow for the standardisation of reactions, helping bridge the information gap present in todays research1.

# Organised Outline

1) Introduction:

* Hypothesis: it is possible to learn new chemistry from a robotic ML platform.
* Removal of bias from sampling all combination reactions.
* Strength of using ML systems: analysis of huge amounts of data and come to unintuitive decisions.
* Supramolecular chemistry is chosen due to its combinatorics, complexity, easy reaction conidiations to automate, reactions being in equilibrium make single discrete species making it easier to be analysed by NMR.
* The advantage of automation comes with its organisation, (and a bit of saving human effort). Humans get tired and bored and will therefore mess up making up tens of different reactions. Robotic platforms do not have this issue and follow steps to the dot.
* Discovery of Novel structures via a more reasoning approach. If this proof of concept works, the paper would not only have provided insights into the chemistry, but to the fact that ML can infact learn chemical rules. The ML can then be trained on a subset of the combination space, and then predict chemical structures for the remainder. This allows for the discovery of novel compounds without neccasrly having to carry out reactions.

2) Results and discussion:

* Definition of chemistry and how it defines the search space.
* Bias of decision maker for single, discrete, diamagnetic supramolecular structures.
* Add reagents used for generation of sample space.
* The choice of reagents were based on: bite angle, rigidity, aromaticity, chirality, sterics, metal valency, and metal size.
* The choice of features for model.
* What is SHAPS and how can it be used for inference (if smiles representations are used, SHAPS will highlight what part of the molecule has the greatest say in prediction).
* The physical setup of the platform.
* How the workflow operates.
* Robotic platforms allow for the standardisation of reactions, helping bridge the information gap present in todays research1.

3) Conclusions:

# Interesting Quotes and ideas from automation papers.

Allowing for the discovery of many materials and mechanisms used in the drugs used

1. Importance of research: chemical synthesis is the foundation of many aspects of society, from consumer healthcare and medicines, to transportation and communication.

Machine learning presents its self as a wonderfull tool to help tackle many chemical problems. Unfortunaly there is a lack of databases with real real world chemistry to play with.

Can be used to help filter chemical space.

Can be used to help understand chemistry.

Although initially highly promising, combinatorial screening strategies did not make a major breakthrough in the fields of energy materials or small molecules, due to the exponential growth of the number of required experiments with every added experimental variable.2

The vast size and high dimensionality (dimension refers here to a continuous or discrete experimental variable) of the chemical design spaces that need to be experimentally explored require new integrated strategies to accelerate the discovery of new molecules and advanced functional materials, as well as to find sustainable ways for their scaled-up synthesis and manufacturing.2

Importantly, the use of SDLs can avail a substantial amount of the researcher’s time to focus on new conceptual or intellectual challenges, rather than on time-consuming repetitive tasks in the lab.2

In contrast to the frequently misinterpreted purpose of SDLs (replacing highly trained scientists in research settings), intelligent robotic assistants are meant to accelerate discovery and avail the time of chemists and materials scientists to high-level scientific questions.2

The main impact of SDLs is the ‘research acceleration’ to generate new knowledge that leads to the discovery of novel compounds or manufacturing routes of the best-performing materials 10–1,000 times faster than by utilizing one-at-a-time variable exploration or combinatorial experiments. The acceleration factor directly translates into a substantial reduction in research time, cost, resources, waste and carbon footprint in academia and industry. This critical adaptive aspect of autonomous experimentation leverages the uncertainty quantification of data-driven ML models to overcome the limitations of non-adaptive combinatorial screening techniques. In general, SDLs improve the experimental data reproducibility through digitization, enhanced accuracy, transferrable knowledge and minimization of the impact of human errors. 2 We encourage the ML community in chemical and materials sciences to focus their future efforts on the facile benchmarking of application-specific algorithms[45](https://www.nature.com/articles/s44160-022-00231-0#ref-CR45), expanding open-access databases and making the design space exploration and/or exploitation software user-friendly.2.

It can be argued that we have now identified the majority of reactions in these easy to reach places, and consequently, we must now consider strategies to search for new areas for discovery3.

To accelerate the discovery of new or improved homogeneous catalysts, research groups in industry and academia have embraced high throughput experimentation (HTE)4. Homogeneous metal catalysts are very well-suited for HTE, since in many cases, they can be prepared by simply mixing a metal precursor and a ligand. However, an HTE program requires a large set of chemically diverse ligands, i.e. a ligand library. The discovery of new or improved homogeneous catalysts is usually achieved via an experimental program where various amounts of candidates are tested until the desired catalytic performance (activity, selectivity, longevity…) is reached. Indeed, it is rather safe to state that no recent publication in the field of homogeneous catalysis started with a theoretical discussion from which one single catalyst was proposed, prepared in the laboratory and successfully tested to give the predicted activity/selectivity. The chemical space is simply far too big to be explored randomly. As an illustration for this, it is estimated that the total number of organic molecules with a mass of <500 Da is around 1060,[1](https://pubs.rsc.org/en/content/articlelanding/2018/cs/c7cs00844a/unauth#cit1) and that there would not be enough atoms on Earth to synthesize them all! 4

“The curse of dimensionality” is a common term in data science that is used to describe the exponential increase in a parameter space size as the dimensionality of a problem increases[13](https://www.nature.com/articles/s41467-023-37139-y#ref-CR13). This issue is prominent in multi-step decision-making processes, including multi-step syntheses, ubiquitous in chemistry and materials science, which exhibit large parameter space complexity after only a few decision steps.5

However, an increase of reaction throughput does not automatically lead to the serendipitous discovery of entirely new transformations while, on the other hand, the discovery of new reaction pathways from first principles (i.e., *in silico*, based on quantum mechanics) is hard due to both the combinatorial explosion of potassible reaction pathways and the computational cost of accurate modeling of the energy hypersurface.6

The idea behind the algorithm is to train the model on a small fraction of the chemical space, explored at random, and use the knowledge acquired to predict the reactivity of the remaining possible combinations. The reactant combinations would then be sorted by predicted reactivity and the best candidates reacted in the platform. After each reaction the model is retrained using the newly obtained reactivity information. By guiding the robot with such a reactivity-driven algorithm it will be possible to perform the reactive combinations first, meaning that only a fraction of the chemical space will need to be explored. It is therefore possible to update and use the model on any organic reaction involving three reagents, meaning that this method is easily scalable to vast chemical spaces with a large number of starting materials. We imagine that by training the model on bigger data sets the scope of the predictions will also expand6.

Using combinatorial chemistry has the disadvantage of the combinatorial explosion. TO mitigate this a space can be learnt from ML and then explored via extrapolation.

To aid in letting the chemist focus more on the intellectual challenges we decided to create a programme that lets the chemist define such chemistries. This is done via x y and z.

There is a lack of data for the experimentation of new ML systems, the generation of more databases allows not just for the chemical exploration but algorthimic exploration.

# Things to add

How to standardise dataset: <https://pubs.acs.org/doi/10.1021/jacs.1c09820>

Some approaches, such as high-throughput screening (HTS), are experimentally driven strategies for catalyst searches. In HTS, a large selection of catalysts, reactants, and solvents are screened in an automated fashion (often carried out on a highly robotized synthesis platform) to find a better-suited catalyst or the best-performing reaction conditions. Because testing all possible combinations of a relatively small set of reactants, solvents and catalysts leads to an exponential increase in the complexity and numbers of experiments, these campaigns are typically limited to only a few tens of catalyst candidates. Other approaches, such as machine learning, are becoming indispensable tools for a large number of in silico tasks, such as molecular design,[6](https://pubs.rsc.org/en/content/articlelanding/2023/dd/d2dd00125j#cit6) virtual screening. The lack of organometallic molecules in training data for existing pre-trained models[6,23,24](https://pubs.rsc.org/en/content/articlelanding/2023/dd/d2dd00125j#cit6) and the use of inconsistent representations of catalysts are the primary causes of inadequate chemical space coverage in generative models for the design of catalysts7.

# Artificial Curiosity: A Search for Novel materials and Chemical Understanding.

Random discoveries play a major role in Chemistry which have revolutionised different industries. Unfortunately, this makes finding novel materials and compounds rare and unpredictable. Here we present a machine learning database that demonstrates capabilities of discovering novel chemistry while providing the researcher with a chemical explanation. Such a database is acquired by a robotic platform capable of carrying out hundreds of reactions. To show the capabilities of such a tool, the paper explores imine-based metal organic complexes with 720 different amine, imine, transition metal combinations. Preliminary X-ray results led to the discovery of a novel supramolecular structure.

## Main

The chemical space is huge. It is thought that the total space of organic small molecules (<500 daltons) is in the order of 1060 compounds4. This is close to the total number of possible ways to stack a 52-card deck. Its massive!

Currently, most methods of discovering novel compounds and chemistry is via serendipity. Such random molecular and material discoveries have aided in tacking challenges in healthcare, materials, energy storage, and sustainability. Just to name a few Nylon, ferrocene, fullerene, electrically conducting polymers, and Wittig olefination were all discovery by chance8. Rarely, have new materials been developed from theory, one must just look at the field of catalysis4. It has been shown that chemist have a bias towards starting materials with historically similar compositions9. Therefore, if such vast chemical space can be explored more quickly in an unbiased manner or the chemistry better understood, novel reactions and materials with socio-economic impact may be discovered and tested at a faster rate.

A hot method to quickly identify if a chemical space may be interesting is via self-driving labs (SDLs)2. These are robotic platforms that carry out reactions in a chemical space, directed by computers namely, machine learning (ML) algorithms. Thus, with the advent of ML and robotics, Macmillan’s accelerated serendipity reaches a new era: hyper acceleration. A more traditional method of searching and understanding the chemical space is via combinatorial chemistry, where a library of compounds is reacted and the reactions analysed.

It would be interesting to combine both combinatorial chemistry and SDLs for both space exploration and chemical inference. Such a two-pronged approach allows the chemist to find novel molecules and possibly design them from a better theoretical frame point. Additionally, getting an understanding of the chemistry improves the trust between both the chemist and ML algorithms. Analysis of real reaction out competes any ab initio explorations and screening due to the latter’s associated computational costs6.

The advantage of using combinatorial chemistry with ML, is its ability to use a subset of the combination space to train the model to then extrapolate reactivities to the remainder of the space - allowing for a rapid screening of novel molecules.

Furthermore, exploring a combinatorial space via a robotic system introduces many advantages. The first being the generation of a dataset which may be used to further develop alternative ML tools. Secondly robots allow the chemist to focus on the intellectual challenges. Likewise, by improving reproductivity and accuracy of results, robotic platform reduces human error2. Lastly, exploring chemistry as combinations help tackle some biased data (its limitations are further discussed in the paper), a major problem with current digital chemistry10,11.

The focus of this paper is on the creation of a workflow that generates a combinatorial space for a ML algorithm to then learn and interpret the reactivity of a combination, therefore showing it can be used to learn a subset of a larger combination space and extrapolate reactivity. The ML algorithm is then used to generate new chemical inferences showing that it truly understood the chemistry it explored.

Robotic platforms for use in chemistry have limited capabilities due to the complexity of chemistry. To add perspective, the platform must be able of handling: different reagent solubilities, heterogenous mixtures, stirring, heating, quantitative workup procedures, solid dispensing, purification, side reactions, and unexpected crashes. Simple chemistry to perform via robots is therefore imperative.

The paper explores imine-based metal organic complexes due to its chemical reversibility, complexity of generated structures, combinatorial accessibility and reaction simplicity.

Chemical reversibility ensures a single discrete thermodynamic product is formed, making analysis a simple task. Typically, when high through put (HTE) is used to discovery novel reactivity and materials, reagents require some form of tags or substituent modification for analysis12. If a single product is formed, a clean 1HNMR spectra with slightly different peaks from the starting materials is expected. Hence, identification of reactivity without tags is straightforward.

Reaction simplicity cuts down the number of challenging steps the platform must be capable of handling. The ChemSpeed used for this research, is only capable of heating, stirring, and transferring volumes. The spontaneity of imine formation and coordination means the reaction may be carried out at room temperature with the presence of the metals shifting the reaction towards the products13. Additionally, the solubility of many metals, amines, and aldehydes in acetonitrile allows for the ChemSpeed to handle reagents in the form of stock solutions.

Combinatorial accessibility allows for the introduction of multiple metals, imines and aldehydes to have an effect on the structure of the product, in other words, with the same reagents multiple combinatorial spaces which still have reactivity may be generated, there are just a greater number of participating reagents in a reaction. Plus, the diversity of commercially availabile aldehydes and amines does not make reagent selection a limiting factor in this study13.

Lastly, the diversity and complexity of product structures makes it a challenge for both the chemist and ML to predict reactivity. The complexity of structures are often unintuitive to the chemist as they can be massive and convoluted, the ML in this case has the advantage of quick prediction with minimal computation and the ability to explore complex and unintuitive data14. The discovery of new metal containing compounds has always been shadowed by more traditional organic chemicals, therefore it remains an underexplored area – a perfect test case for the system9.

## Combinatorial Chemistry

To guide the reactions the robotic platform must take, combinatorial like chemistry is defined. A reaction is first broken down into reagents with similar functional groups and chemistries. The paper focuses on imine-based metal organic complexes and therefore a reaction can be subdivided into amines, aldehydes, metals (Figure 1a). A chemist then procures a library of reagents for each subgroup (Figure 1b). It is possible to have only one reagent in a library for a subgroup.

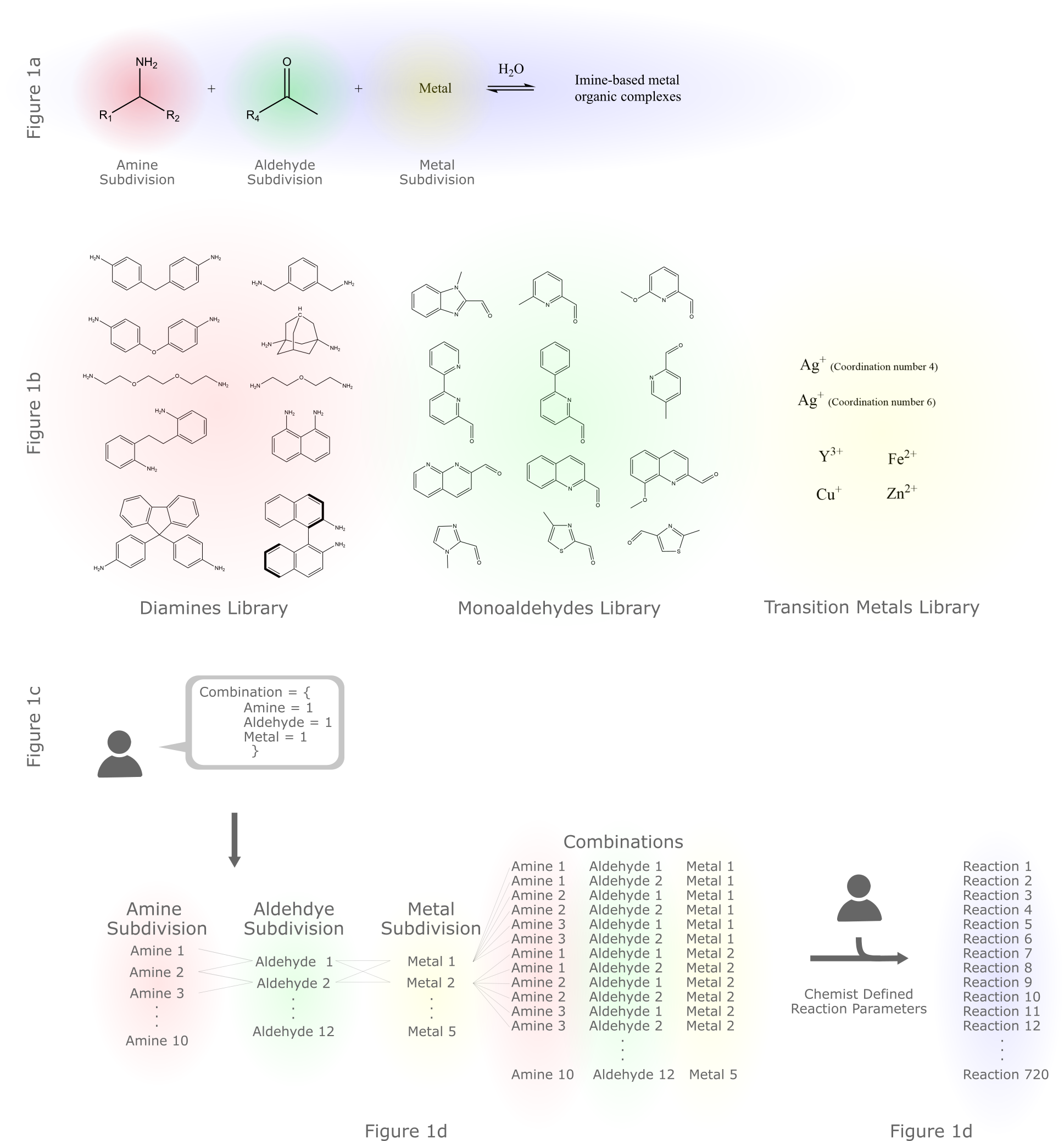


Figure : Generating the combinatorial space, where red, green and yellow represent subgroups while blue represents reactions. 1a) The subdivision of the chemistry explored in the paper. 1b) The reagent libraries used in the paper. There are three libraries for each subgroup. 1c) Stating how many reagents of a subgroup should appear in a combination. The paper uses one of each. 1d) The generation of combinations using the previously defined libraries, subgroups and subgroup quantities. 1e) Translation of combinations to reactions based on several different chemist defined parameters.

The paper focused on diamines to form linkers, monoaldehydes to form imines and metals for the imine to coordinate to. More specifically, different reagents where chosen on the basis of encapsulating functional diversity: bite angle, rigidity, aromaticity, chirality, sterics, electronegativity, metal valency, and metal size (Figure 1b). Selection of varied reagents improves the generalisability of the model and leads deeper insights into, potentially, both imine and organisational chemistries.

The chemist then defines how many reagents of each subgroup appears in a combination (Figure 1c). This can be any positive integer starting from 0, where 0 implies the subgroup should not appear in the combination, and two implies two reagents of the same subgroup should appear in a combination. The paper uses one reagent of each subgroup in a combination (Figure 1c) once.

Combinations are then generated based on libraries and number of subgroups in a combination (Figure 1d). This paper used 10 different diamines, 12 different monoaldehydes, and 5 different transition metals (of which silver was considered to have coordination numbers of 4 and 6) for a total space size of 720 combinations. Here the power of combinatorial chemistry may be noted: both reactions that work and reactions that wont work will be tried and analysed reducing bias in the dataset. The robotic platform then ensures the reactions are standardised and replicable, further helping bridge the information gap present in literature1. It is important to note that the current robotic platform does create bias towards nongaseous, stable, soluble reagent libraries.

The chemist then defines reaction and reagent parameters to translate combinations to reactions (Figure 1e). Parameters include: concentration of a subgroup in a combination, chemical weights of reagents, solubility of reagents, price of reagents, coordination number, and transfer volumes. This paper takes combinations, heats them at 60oC for 40h to create reactions.

A programme developed by the group is responsible for storing the reagent library, and for the generation, and translation of the combinatorial space. Digitally representing such chemical spaces makes it easy to implement in a workflow.

## Workflow

Now with a list of reactions for a ML to explore, data must be generated for it to learn. The process of carrying out and analysis of chemical space is handled by a Chemspeed Swing platform, a Bruker benchtop NMR machine, and a Waters Acquity LCMS system (Figure 2).

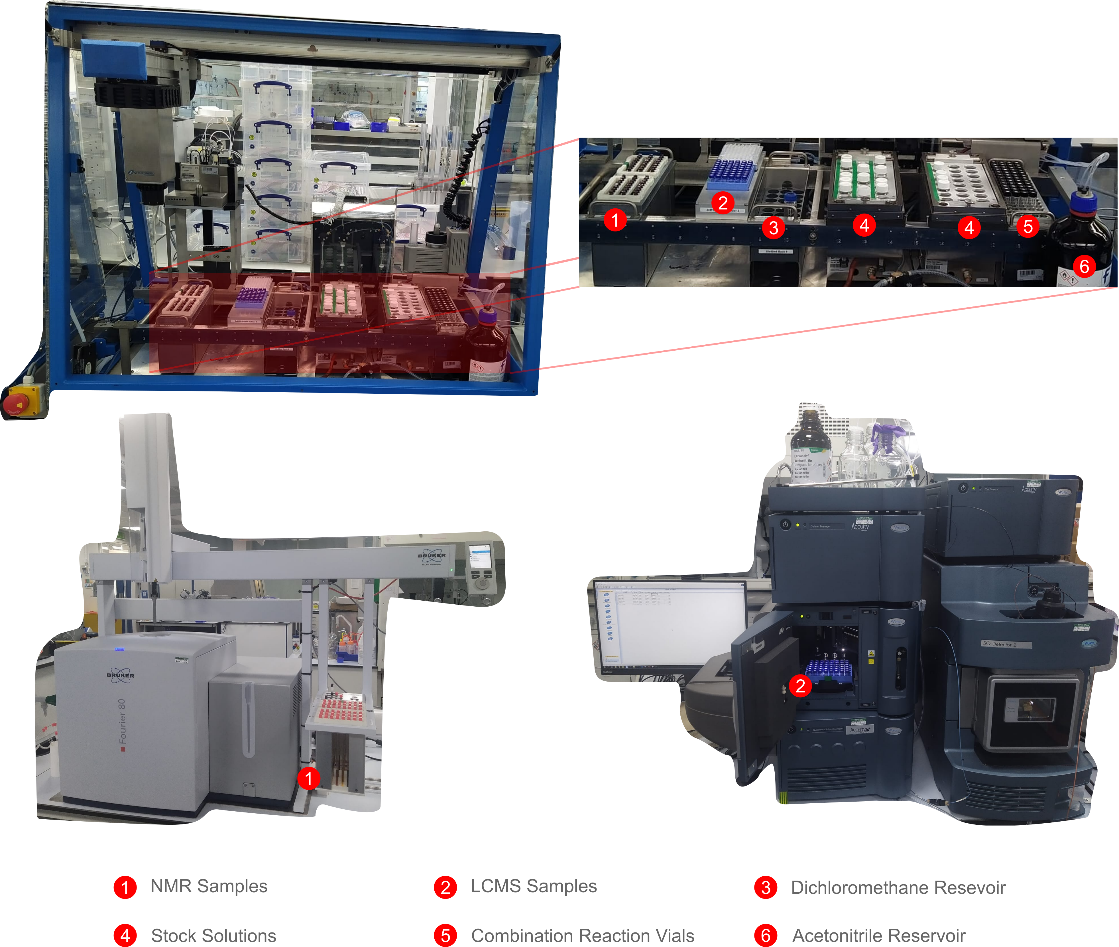


Figure : The three platforms used in the workflow. At the top is the Chemspeed platform, responsible for making up stock solutions, reactions scheduled by a programme and taking NMR and LCMS samples for each combination. At the bottom left is the benchtop NMR used in the workflow along with the in house build NMR sample rack. At the bottom right is the LCMS machine used in the workflow, along with reaction samples.

Since platforms are independent of each other and are limited to 48 samples, a programme must schedule and communicate samples to make up between platforms. The workflow the programme follows and its capabilities are shown in Figure 3.

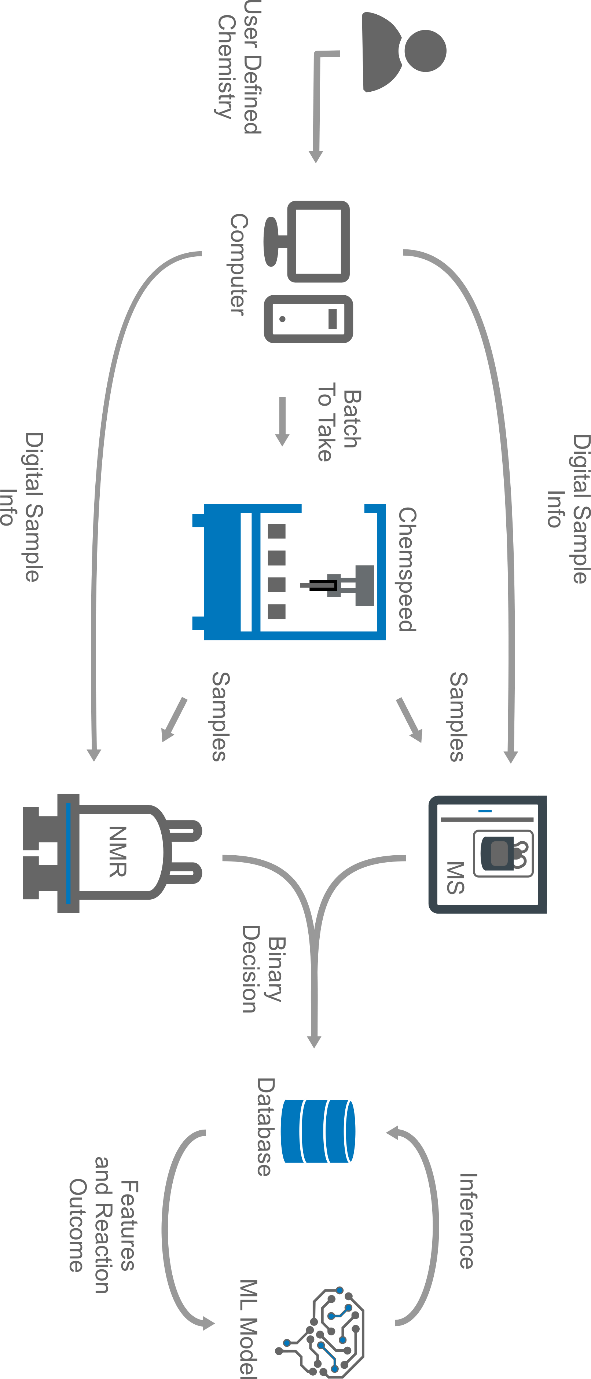


Figure : The overall workflow shown as a sequence of steps. The logical order shown in the diagram is the same as the one taken out physically. There are four main pieces of equipment: a centralised computer to store data and manage tasks, a ChemSpeed platform to carry out reactions, an NMR and LCMS machine to analyse data.

Once the chemist adds the subgroup library into the programme, a chemical space (all combinations) to explore is generated. The space is then filtered by price and reagents. Early stages of workflows are prone to errors, hence starting with the cheapest reagents reduces project expenses. Reagent masses are measured manually, hence minimising the number of reagents to measure in a batch reduces labour.

After sorting the chemical space, standard reactions are added to assess the quality of a batch. Only then are batches sent as digital samples to the robotic, NMR and LCMS platforms. Next, the robotic platform makes up the reactions and creates samples for the chemist to then transport to the NMR and LCMS machines, which thanks to the digital sample knows what sample corresponds to what combination. With analytical data acquired, an algorithm developed by the group uses chemical shifts and peak positions from 1HNMR data to determine if a reaction has occurred, it then uses m/z ratios from mass spec (MS) data to determine the size of the supramolecular structure (it is possible that only imines are formed and not a complex, MS data helps identify such combinations). The final output of the algorithm is a binary classification (0,1) of what combinations yield supramolecular structures.

The analytical classification algorithm, unfortunately, introduces bias to the dataset. It is only able to classify discrete, paramagnetic species. High spin iron (II) complexes, and oligomers will not be accounted for, although preliminary manually curated results demonstrated such products due exist in the paper’s chemical space.

Once all reactions have been carried out and classified, it is stored in a local database to then be used for ML model training and SHAP analysis. Simply put, SHAP analysis allows the chemist to understand what parameters were the most relevant for a ML model to predict reactivity. Due to SHAP’s game theory approach, SHAP analysis lies in local explanations but may be used for global explanations. Local explanations look at the chemistry of a single reaction, while global explanations look at the general chemical trend, both provide powerful insights of the explored space.

Defining chemistry in such a way comes with the following advantages:  
the addition of the ketene

#have a section describing the chemist time vs reaction time (i.e. robotic platform v human chemist time)

## References

(1) Schrader, M. L.; Schäfer, F. R.; Schäfers, F.; Glorius, F. Bridging the Information Gap in Organic Chemical Reactions. *Nat. Chem.* **2024**, *16* (4), 491–498. https://doi.org/10.1038/s41557-024-01470-8.

(2) Abolhasani, M.; Kumacheva, E. The Rise of Self-Driving Labs in Chemical and Materials Sciences. *Nat. Synth.* **2023**, *2* (6), 483–492. https://doi.org/10.1038/s44160-022-00231-0.

(3) Collins, K. D.; Gensch, T.; Glorius, F. Contemporary Screening Approaches to Reaction Discovery and Development. *Nat. Chem.* **2014**, *6* (10), 859–871. https://doi.org/10.1038/nchem.2062.

(4) Renom-Carrasco, M.; Lefort, L. Ligand Libraries for High Throughput Screening of Homogeneous Catalysts. *Chem. Soc. Rev.* **2018**, *47* (13), 5038–5060. https://doi.org/10.1039/C7CS00844A.

(5) Volk, A. A.; Epps, R. W.; Yonemoto, D. T.; Masters, B. S.; Castellano, F. N.; Reyes, K. G.; Abolhasani, M. AlphaFlow: Autonomous Discovery and Optimization of Multi-Step Chemistry Using a Self-Driven Fluidic Lab Guided by Reinforcement Learning. *Nat. Commun.* **2023**, *14* (1), 1403. https://doi.org/10.1038/s41467-023-37139-y.

(6) 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* (11), 1821–1830. https://doi.org/10.1021/acscentsci.1c00435.

(7) Schilter, O.; Vaucher, A.; Schwaller, P.; Laino, T. Designing Catalysts with Deep Generative Models and Computational Data. A Case Study for Suzuki Cross Coupling Reactions. *Digit. Discov.* **2023**, *2* (3), 728–735. https://doi.org/10.1039/D2DD00125J.

(8) Rulev, A. Y. Serendipity or the Art of Making Discoveries. *New J. Chem.* **2017**, *41* (11), 4262–4268. https://doi.org/10.1039/C7NJ00182G.

(9) *Exploration of the chemical space and its three historical regimes*. https://doi.org/10.1073/pnas.1816039116.

(10) Beker, W.; Roszak, R.; Wołos, A.; Angello, N. H.; Rathore, V.; Burke, M. D.; Grzybowski, B. A. Machine Learning May Sometimes Simply Capture Literature Popularity Trends: A Case Study of Heterocyclic Suzuki–Miyaura Coupling. *J. Am. Chem. Soc.* **2022**, *144* (11), 4819–4827. https://doi.org/10.1021/jacs.1c12005.

(11) Artrith, N.; Butler, K.; Coudert, F.-X.; Han, S.; Isayev, O.; Jain, A.; Walsh, A. Best Practices in Machine Learning for Chemistry. *Nat. Chem.* **2021**, *13*, 505–508. https://doi.org/10.1038/s41557-021-00716-z.

(12) Troshin, K.; Hartwig, J. F. Snap Deconvolution: An Informatics Approach to High-Throughput Discovery of Catalytic Reactions. *Science* **2017**, *357* (6347), 175–181. https://doi.org/10.1126/science.aan1568.

(13) van Dam, A.; van Schendel, R.; Gangarapu, S.; Zuilhof, H.; Smulders, M. M. J. DFT Study of Imine-Exchange Reactions in Iron(II)-Coordinated Pincers. *Chem. – Eur. J.* **2023**, *29* (61), e202301795. https://doi.org/10.1002/chem.202301795.

(14) Jelfs, K. E. Computational Modeling to Assist in the Discovery of Supramolecular Materials. *Ann. N. Y. Acad. Sci.* **2022**, *1518* (1), 106–119. https://doi.org/10.1111/nyas.14913.