Developing sustainable chemistry with AI

Jiyizhe Zhang,[1,2] Polina Yaseneva[1,3] and Alexei Lapkin[*1,4,5]

[1] *Department of Chemical Engineering and Biotechnology, University of Cambridge, Cambridge CB3 0AS, United Kingdom*
[2] *Department of Chemical Engineering, University of Manchester, Manchester M13 9PL, United Kingdom*
[3] *Department of Chemical Engineering, University College London, London WC1E 6BT, United Kingdom*
[4] *Cambridge Centre for Advanced Research and Education in Singapore (CARES), 1 Create Way #05-05 Create Tower, Singapore 138602, Singapore*
[5] *Chemical Data Intelligence (CDI) Pte Ltd, 105 Cecil Street, The Octagon #22-25, Singapore 069534, Singapore*

* Corresponding Author: aal35@cam.ac.uk

Accounts Chemical Research

CONSPECTUS

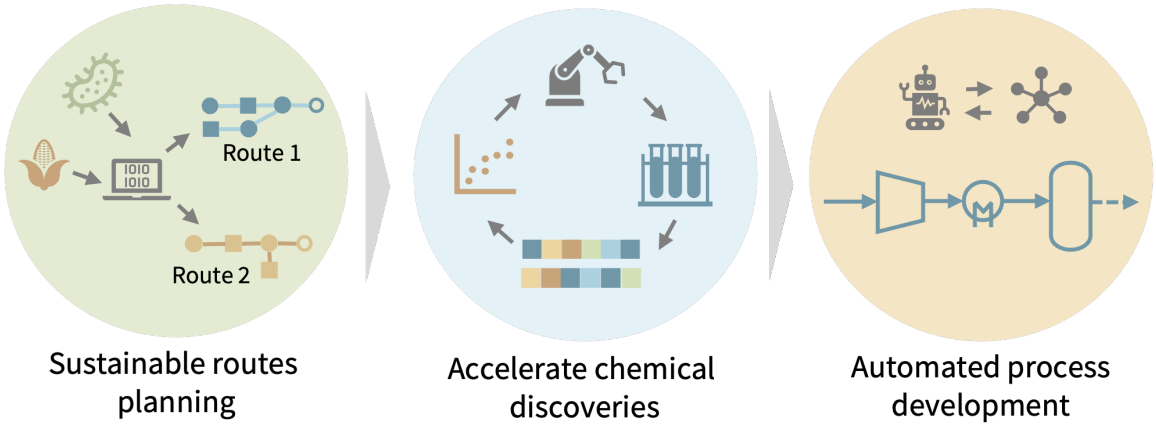
Today we are experiencing an unprecedently rapid transformation of science, driven by advances in computing and artificial intelligence (AI). At the same time, there are growing cumulative pressures of climate change impacts that not only affect economies but are creating significant societal challenges. The confluence of technology push and environmental-socio-economic pressures are promoting the development of *sustainable chemistry*. The link between transition to sustainability in chemistry and digital transformation of R&D and manufacture has been the long-standing focus in our research.[1] Sustainability is not a feature of a single chemical product or a process. It is incorrect to talk about 'sustainable polyethylene', as such narrow framing will miss many issues in the more holistic system that includes not only the cradle-to-grave life cycle emission impacts of polyethylene, but also considers all aspects of equity, impact on biodiversity, resilience of supply chains, etc. Sustainable chemistry only makes sense as a complex system. Analysis of complex systems requires large amounts of data and the use of appropriate analytical tools, such as advanced machine learning techniques.

Chemical industry has been an early adopter of digital technologies, motivated by the inherent complexity of chemical systems and the need for efficiency to stay economically competitive. As early as the 1970s, cheminformatics tools were introduced to support structure–property analysis. In the 1980s, we saw broad adoption of process modelling and simulation tools for

design, scale-up and optimisation of chemical processes, paralleled with the developments of the fields of process system engineering and control. Today's digital transition is much deeper; everything, from scientific discovery (molecules, materials and phenomena) to chemical process development, analytical methods and process analytical technologies, and plant operation is being reshaped. This deep transformation includes the development of a radically different approach to data in chemistry. Advances in self-driving laboratories, agentic AI workflows, digital factories, require access to structured, semantically rich data.

In this Account, we focus on three key areas of digitalisation and AI technologies advancing chemistry and chemical process development towards net-zero targets. First, we discuss how data mining tools are used to analyse the existing value chains and identify opportunities to integrate net-zero feedstocks. When introducing new molecules, data-driven route planning can be employed to design alternative synthetic pathways, which then need to be rigorously evaluated. Second, we focus on how to use AI to accelerate discoveries in chemistry, this includes innovative design of catalysts, rational selection of solvents and identification of optimal reaction conditions. By leveraging statistical modelling and data-driven optimisation, these traditionally time-consuming tasks can now be performed much faster. Lastly, we discuss how to bring these innovations from lab to manufacturing, which also requires innovation in process technologies. We show that by leveraging AI agents, process modelling and knowledge graphs, there is a potential to achieve automated process design and life cycle assessment, removing traditional choke points in scaling technologies, and providing timely support for decision-making in early-stage development. In the end, we discuss current challenges and future perspectives.

Abstract graphics (TOC)



Key references

- Fantke, P.; Cinquemani, C.; Yaseneva, P.; De Mello, J.; Schwabe, H.; Ebeling, B.; Lapkin, A. A. Transition to Sustainable Chemistry through Digitalization. *Chem.* **2021**, *7*, 2866–2882.[1] DOI: 10.1016/j.chempr.2021.09.012. *A perspective paper arguing for the critical role of digital infrastructure for development of sustainable chemical solutions.*

- Weber J. M.; Guo Z.; Lapkin A. A. Discovering circular process solutions through automated reaction network optimization. *ACS Eng. Au*. **2022**, 2(4), 333–349.[2] DOI: 10.1021/acsengineeringau.2c00002. *Introduce a novel automated framework that combines reaction network generation, kinetics, and optimisation to systematically identify circular and resource-efficient chemical processes, reducing the manual design efforts and accelerating discovery of sustainable processes*.

- Seidenberg J. R.; Khan A. A.; Lapkin A. A. Boosting autonomous process design and intensification with formalized domain knowledge, *Comp. Chem. Eng.*, **2023**, 169, 108097.[3] DOI: 10.1016/j.compchemeng.2022.108097. *The developed hierarchical reinforcement learning agent system interacts with an ontological knowledge base to discover the optimal intensified process designs, demonstrating great potential of using knowledge representation to empower the performance of intelligent agents*.

- Laub J.-F.; Zhang J.; Heyer M.; Lapkin A. A. Automated generation of mechanistic models for chemical process digital twins using reinforcement learning part II: Compartmentalization and learning-based recalibration. *Comp. Chem. Eng.*, **2026**, 204, 109384.[4] DOI: 10.1016/j.compchemeng.2025.109384. *Design a novel pipeline for automated generation of physical models for novel design of unit operations using coloured graphs and reinforcement learning. Embed human knowledge into the workflow by defining graph grammar and reward shaping*.

- Zhang, S.; Zhang, J.; Lapkin, A.A. A knowledge graph framework for digital twins of chemical processes. *Nature Chem. Eng*., **2026**, 3, 286–297[5] DOI: 10.1038/s44286-026-00392-1. *Here we describe the fundamental approach to handling chemical process knowledge through a knowledge graph; we introduce and expand the necessary ontologies and show how this technology is supporting multiple applications in process development.*

Introduction

Chemical industry is fundamental to modern civilisation; while many tasks in the 'service economy' depend largely on access to electricity, most industries depend on molecules and formulations of molecules and materials. We design molecules and materials that perform desired functions in a wide range of applications. We develop synthetic and purification procedures and combine them into manufacturing processes. Production, use and final disposal of functional molecules and materials come at the cost of used resources and of potentially harmful emissions.

Focusing on climate change, it is estimated that the chemical sector accounts for approximately 6% of global CO_2-equivalent emissions.[6] Apart from direct emission of CO_2 from chemical manufacturing, around one-third of total emissions comes from processing fossil feedstocks and end-of-life product treatment.[7] Tracing emissions and allocating them to individual stages within chemical value chains is a significant and urgent problem.[8] Having a clear picture of the sources of emissions is critical to our ability to construct more efficient value chains.[1]

Moving carbon from fossil stores to atmosphere through the current linear chemical value chain, and generation of further CO_2-eq. emissions in processing and end-of-life are the two core challenges that can be addressed through a holistic re-engineering of the whole chemical value chain. For this, the whole-system approach is required, that considers: (i) shifting from linear to circular supply chains (with bio-renewable, captured CO_2, and recycled net-zero feedstocks), (ii) designing novel synthetic procedures, (iii) linking chemical manufacture with non-fossil energy generation and storage, and (iv) introducing innovations in process technologies, that would drive down the global metrics of resource use, emissions and costs evaluated by cradle-to-grave life cycle assessment (LCA).[9] When we add other aspects of sustainability to the challenge of 'net-zero chemistry', our design vector for molecules and routes becomes even more complex, see Figure 1.

We argue that the system-level challenge of developing sustainable chemistry requires digitalisation of chemistry knowledge and chemical manufacturing technologies, and the adoption of artificial intelligence (AI) analytical tools.[1] Chemical industry has been one of the pioneers of digitalisation and AI, specifically in the areas of cheminformatics for chemical synthesis and molecular property prediction,[11,1210,11] developing machine learning (ML) surrogates for process modelling,[12] and applying expert systems for process design.[13]

Increasing availability of data and computing power are beginning to translate into increased rate of scientific discovery; application of digital tools and AI to solving problems in science and engineering is becoming the new norm,[14] as highlighted by the Nobel prize of 2024. It is a good

time to consider how we could leverage these new capabilities to help us accelerate innovation in chemistry and chemical manufacturing for transition to sustainable chemical industry.

In this Account we explore this theme. We provide an overview of how AI can support chemistry research and process development towards sustainable future, by focusing on three aspects: design of sustainable value chains, acceleration of chemical discoveries, and development of chemical processes. Challenges and future directions are discussed in the end.

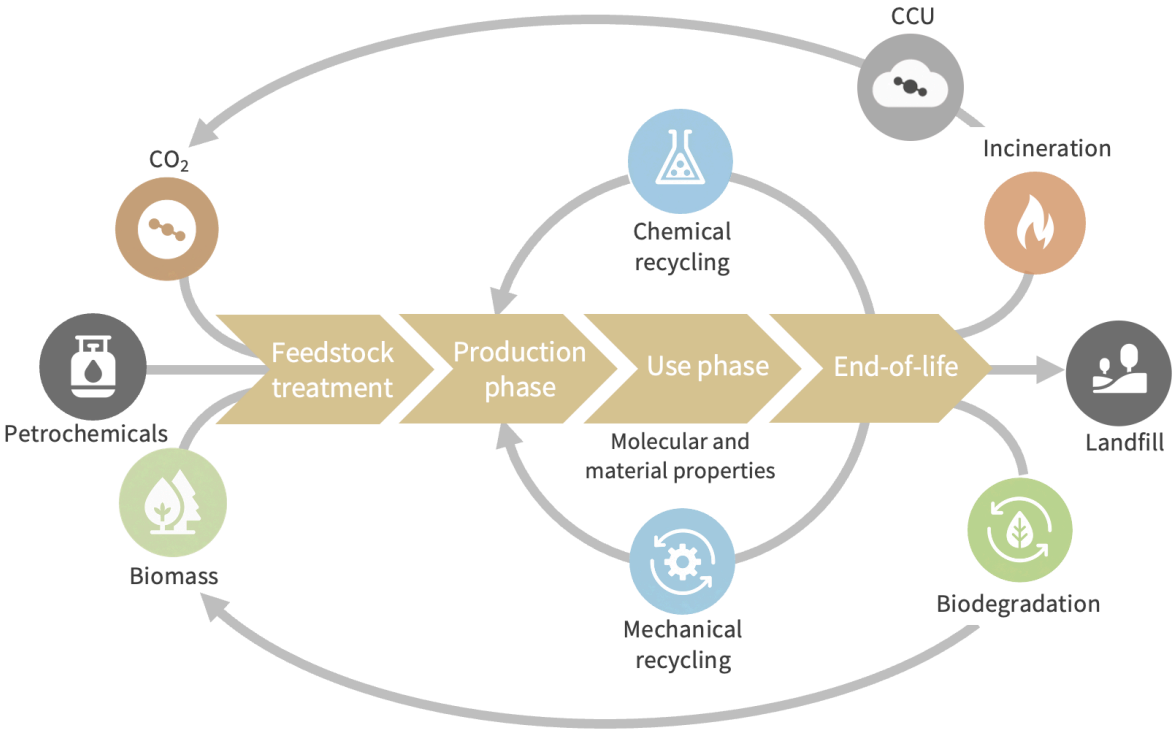


Figure 1. A schematic illustration of the emerging circular multi-feedstocks paradigm of sustainable chemistry.

1. Sustainable Chemical Value Chains

Today's chemical value chains begin primarily from fossil resources (oil, naphtha, coal, natural gas, shale gas and oil) as the main sources of carbon, heteroatoms and process energy. In the linear economy with many products having short usable lifetime, this leads to chemistry serving as a pump of fossil carbon into atmosphere, contributing to climate change. Long-lived chemicals are contributing to long-term pollution (plastics and 'forever chemicals' in the environment). As we strive to move chemistry to delivery of societally useful functions without contribution to climate change ('net-zero chemistry') and other environmental problems through adoption of new frameworks such as Safe and Sustainable by Design,[15] we are considering the nature of the feedstocks used, efficiencies of the procedures, and nature of energy and waste streams simultaneously.

A shift away from fossil feedstocks would require adoption of a diverse set of feedstocks, including fresh renewable biomass, bio-waste, captured and upgraded CO_2, and recycled carbon/heteroatoms from the end-of-life streams[16,17] as alternative sources of elements. The new set of raw materials lead to either the same (drop-in) or a new set of intermediates.[6] We found it productive to discuss intermediates in the context of their role within reaction networks as "effective connectors", thus defining them as "hub" or "strategic" molecules.[18]

The "drop-in" strategy is when net-zero feedstocks are converted to produce a "hub" molecule that is a part of an existing value chain. This approach benefits from the use existing manufacturing infrastructure. An example is the development of direct chemical transformations of lignocellulosic materials to produce platform molecules such as 5-hydroxymethylfurfural (HMF), or developing a route from terpenes to terephthalic acid, leading to a bio-based monomer;[19] these drop-ins can link seamlessly into the existing downstream processes.

The second approach is more disruptive and product-oriented, focusing on developing entirely new products and pathways, such as producing Cyrene from cellulose to replace the toxic solvent N-methyl-2-pyrrolidone.[20] This case is particularly interesting as Cyrene may not necessarily be used as a solvent, but as a new starting structure in the synthesis of more complex functional molecules, in other words, it may become a new "hub" molecule. This also applies to novel materials, that reduce the use of fossil resources and offer new functionalities, such as the emerging nanocrystalline cellulose-based products.[21] For both approaches, new reaction pathways need to be identified or developed to link with new feedstocks. We explored data mining and algorithmic methods to enable automated search for sustainable routes.

1.1 Reaction data mining and route planning

Any given value chain of molecules is a fragment of a much larger network of reactions. This earlier concept was revived by Grzybowski and co-workers,[22] demonstrating that a global view on reaction networks provides insights for reactivity analysis and for guiding synthesis planning. Within such networks there exist highly connected "hubs" which appear in most synthetic routes.[23] Here we are connecting the mathematical term "hub" from scale-free networks with a practical meaning in chemical synthesis. We first identified hub (strategic) molecules by defining their mathematical features and identifying them within a large network of molecules as outliers.[18] We further hypothesised that transition of chemical value chains away from fossil carbon sources will lead to the emergence of new strategic molecules,[24] the concept we are currently investigating.

Having access to the majority of the currently known chemical reactions, for example through exploring Reaxys™ as a source of data for reaction networks,[25] allows us to explore synthesis planning with the primary focus on sustainability: using bio-renewables as starting molecules,[26–28] generating multiple metrics to enable multi-criteria decision making in ranking promising routes,[29] optimising mass flow based on exergy through a reaction network, for example, by using Petri nets.[30] An automated pipeline has been developed to integrate data mining, network analysis, pathway planning and assessment,[2] which significantly reduces the effort of evaluating promising routes and reduces the barrier of adopting these technologies in early stage of process development, see Figure 2.

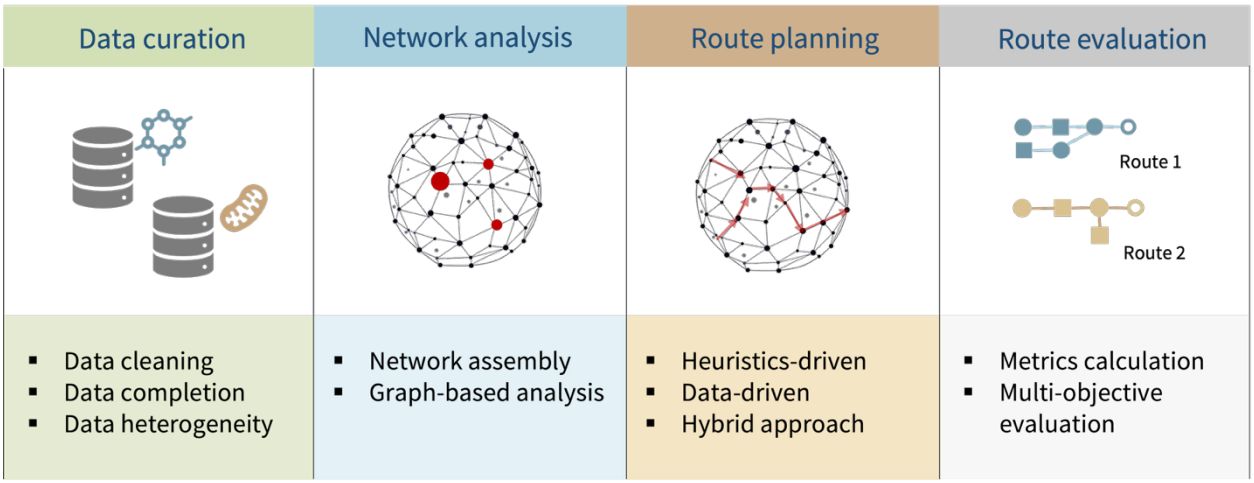


Figure 2. A schematic illustration of a workflow of reaction data mining and synthetic routes planning.

1.2. Introduction of net-zero feedstocks

There is a particular challenge in designing sustainable routes towards molecules containing heteroatoms. Some bio-renewable feedstocks (e.g., lignocellulose) contain carbon-, hydrogen- and oxygen-rich molecules, whereas nitrogen-containing building blocks are found in very different renewable feedstocks (chitin or proteins), or through potential net-zero ammonia routes. As a result, synthesis planning may require access to multiple feedstocks. In a recent study, we used a deep data mining approach based on similarity-based heuristics and in-depth molecular structure analysis to identify promising synthetic pathways for the production of pyrrole-2-carboxylic acid using both cellulose and chitin as raw materials.[26] Several promising synthetic routes were identified and subsequently verified experimentally. In the broader literature, data-driven approaches complement graph methods and are being actively developed for route design; a recent review on this topic can be found elsewhere.[31]

Advanced route planning tools allow exploration of the impact of novel synthetic methods on the overall synthesis planning. Developments of new synthetic methods such as electro-, photo-, or biocatalytic reactions are enabling novel bond disconnections and introduce new options for route design. In a recent study, we have demonstrated design of synthesis routes that include both, chemical and bio-chemical reactions.[32] For this we built a database of feasible bio- and chemical transformations and used reinforcement learning for optimising synthetic paths. Such an approach allows to bring novel synthetic methods into synthesis planning.

2. Acceleration of Discoveries in Chemistry and Chemical Processes

Innovation in chemistry has been a protracted and costly process, with discoveries often made through serendipity, while development timelines often stretching over decades.[33] Today, our aspiration is to significantly reduce the chemical development timeframes through a combination of several key technologies: (i) AI tools for molecular/materials design, (ii) automation and AI enabling faster data generation in chemical reactions development and in process conditions exploration/optimisation, and (iii) adoption of new synthetic methods. We are particularly interested in technologies that allow significant acceleration of the development timelines for chemical reactions: forward and inverse design, automated optimisation and incorporating prior knowledge through chemical descriptors.

2.1. Forward and inverse design

Forward design in the context of chemical reactions takes the form of trial-and-error testing of molecules, reaction conditions, catalysts, materials that allow to obtain the target result. For example, identification of an optimal catalyst for a specific reaction is typically a screening of available options. This approach to computational catalysts design is through forward enumeration – exploring the potentially feasible catalytic structures and then screening for reactivity and selectivity in the desired transformations. This can be done for both homogeneous and heterogeneous catalytic reactions.[34–36] Here we can make use of the growing number of datasets of materials structures and DFT-computed energies.[37,38] The accuracy of such predictions and the size of the dataset are conflicting demands, as accurate calculations of equilibrium geometries and of activation energies by an appropriate level of DFT theory are highly expensive calculations; hence, we note that currently only limited types of catalysts are considered in each individual study. A rapidly developing field of machine learned force fields is promising to attain the required level of accuracy of energy prediction to be able to substitute expensive DFT calculations with the much faster force fields methods, which will revolutionise the field of catalyst design.[39]

Inverse design aims to identify the solution directly from the set of target properties: e.g., identifying the optimal catalyst structure directly from the specification of the target reaction. For

the inverse design of catalysts, we may imagine a workflow when we start by identifying plausible energy diagrams, or reaction coordinates, that enable a selective transformation from starting molecules to the desired product(s). Identification of potential feasible energy diagrams can be formulated an optimisation problem that has multiple solutions. Once we have a set of possible potential energy diagrams, we would want to identify what catalysts may enable such reaction coordinates. The next step, the inverse materials design challenge is, however, much harder. For the inverse design methodology to be developed, we need to link fundamental theoretical approaches allowing forward predictions, with machine learning models built on experimental or computational observations. As an example, a library of water-gas shift catalysts and their experimental studies was assembled; a neural network was trained on the dataset which then was used to design a new heterogeneous catalyst.[40] The use of statistical machine learning is a very fruitful direction for materials and formulations design – areas where large datasets of high-quality experimental data may exist. Such models are limited in their predictive capacity, since the models usually do not allow to predict materials that are not present in the dataset. However, some extrapolation may be feasible if catalysts/materials are parameterised and machine learning models are trained on "descriptors", rather than the instances of specific materials.[41] A recent review describes state-of-the-art and potential directions for future development of computational methods of catalysts design.[42]

2.2. Reaction development with ML and self-driving laboratories

Reaction systems involve combinations of catalysts, solvents, and operating conditions (e.g., temperature, pressure, concentration), creating a high-dimensional space, where identifying optimal conditions for later scale-up can be challenging. This is effectively addressed by multi-dimensional optimisation in closed-loop experiments (Figure 3(a)). Two key techniques are combined: a surrogate function that captures the complexity of a multivariate problem, and efficient sampling methods that reduces the number of observations (experiments), and hence the experimental effort, required to identify optimal conditions. When an efficient sampling algorithm is combined with an automated experimental system we obtain a closed-loop or a "self-driving" experiment.[43,44]

The algorithmic side of "self-driving" laboratories has seen rapid and widespread development, see Table 1 for an overview of available packages. Initially we worked with algorithms in which multiple objectives (e.g., reaction yield and cost) were scalarised to a single objective.[45] Later we moved to multi-objective Bayesian optimisation algorithms.[46] Notably, these algorithms can be used in manual[47] and semi-automated[48] workflows, as well as within fully automated "self-driving" laboratories. A comprehensive review of the development of self-driving labs can be found elsewhere.[49]

Table 1. A non-comprehensive list of open-source python packages for reaction optimisation (SOBO: single-objective Bayesian optimisation, MOBO: multi-objective Bayesian optimisation).

Package	Description	Link
Gryffin[50]	SOBO for both continuous and categorical variables	https://github.com/aspuru-guzik-group/gryffin
NE Torch[51]	MOBO for continuous variables	https://github.com/VlachosGroup/nextorch
Summit[52]	SOBO and MOBO for both continuous and categorical variables	https://github.com/sustainable-processes/summit
BayBe[53]	SOBO and MOBO for both continuous and categorical variables	https://github.com/emdgroup/baybe
BoFire[54]	SOBO and MOBO for both continuous and categorical variables	https://github.com/experimental-design/bofire

2.3. Chemical knowledge representation via descriptors

A specific sub-problem in the field of reaction development is the challenge of chemical knowledge representation.[55] Here, chemical descriptors is a way of enriching the data available for training machine learning models, by encoding knowledge of physical chemistry, molecular/materials structures, interactions of different length scales, etc in specific *descriptors*. Because the databases of known chemical facts are very small and rarely explore facts systematically, it is challenging to rely on ML to learn chemistry from observations. Addition of chemical descriptors in many cases allows to improve accuracy of ML predictive models.

A particular success story with the use of molecular descriptors is the prediction of the role of solvents in reactions. Solvents are ubiquitous for chemical transformations, providing the environment in which chemical reactions occur, as well as serving useful process functions, such as heat transfer. Solvents significantly influence reaction rates and selectivity, either enhancing or reducing yields.[56] Selection of a good solvent is often guided by experts' experience or past literature. Computer aided approach based on Density Functional Theory (DFT) calculation leverages quantum chemistry to predict solvent properties,[56] which allows for screening a set of solvent candidates. However, these methods are computationally expensive.

Amar and Zhang developed automated workflows for rational selection of solvents for a catalytic reaction systems.[57,58] Machine learning surrogates were trained based on molecular descriptors from literature as well as descriptors based on physical knowledge from COSMO-RS simulations.

These surrogates were then used in Bayesian optimisation to successfully identify the best solvents to meet two targets simultaneously, namely high conversion and high diastereomeric excess (Figure 3(b)), or the more efficient Mitsunobu reaction conditions.

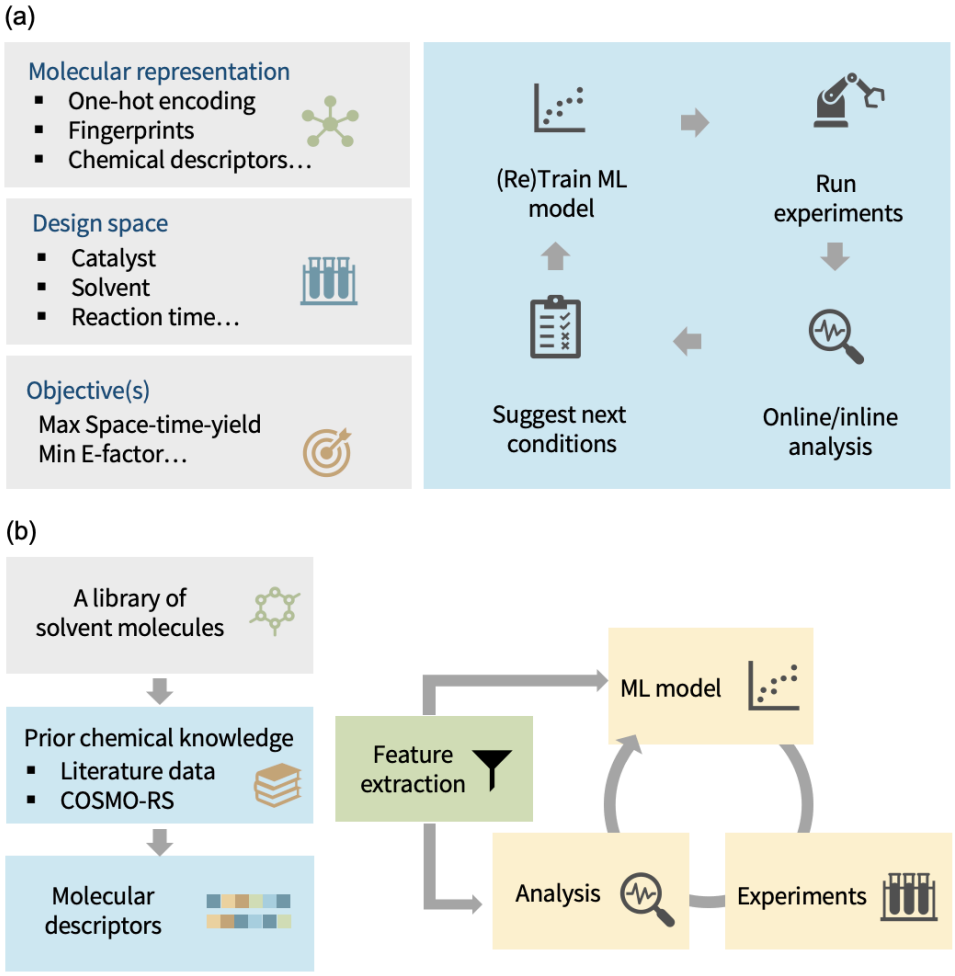


Figure 3. Applications of using AI for development of chemical transformation. (a) Closed-loop reaction optimisation assisted by Bayesian optimisation and (b) Use of chemical descriptors to assist solvent selection.

3. From lab to manufacture

For many chemical processes, transition from lab discoveries to manufacture is very challenging. This is true for scaling manufacture of small molecules, of biologics, of nanomaterials, or any processes involving multiple phases. To successfully scale a multiphase process, it is critical to understand the variation in process *context* and to track the changes that must be implemented to attain the desired product quality and process key performance indicators (KPIs). The concept of *context* is becoming particularly important in both chemistry and in AI. In chemistry, *context* is critical in developing correct first-principles models and for scaling.[59] In AI, context has different meanings, from a collection of true facts, to the assembly of text chunks exposed to a language

model. Transition to digital R&D and digital manufacture is promising to make the challenge of scaling processes more tractable, and especially with the advances in agentic AI frameworks and the digital twins technology (Figure 4).

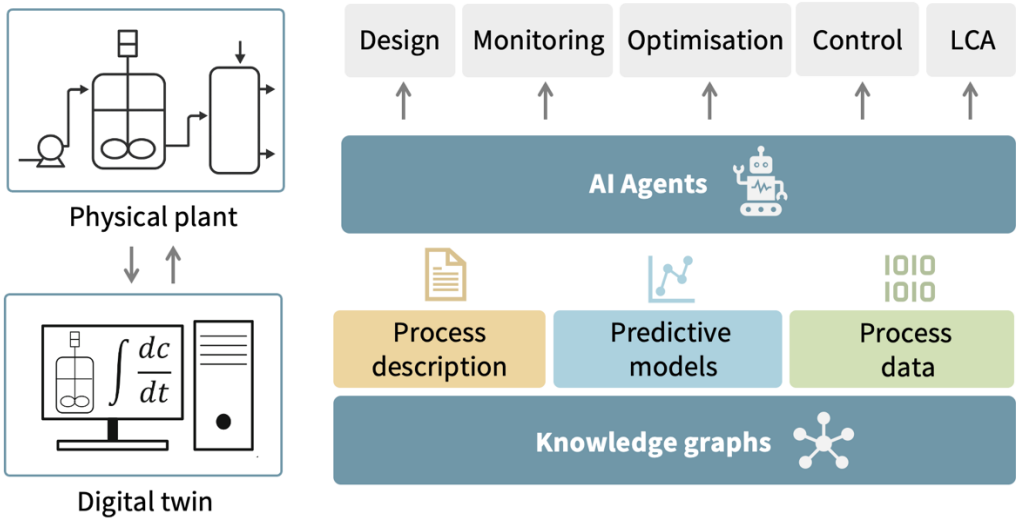


Figure 4. An illustration of the concept of digital twin and AI agentic system.

3.1. *In silico* and automated process design

The combination of predictive models and knowledge graphs is highly promising for creating "digital twins" of chemical processes and enabling *in silico* process design, which can also be extended to automated process design and full plant operation in the agentic framework.[60] For the case of finding an optimal flowsheet composed of intensified process technologies, a generalised methodology was proposed by Sundmacher and Freund,[61] which is based on optimising a path through a thermodynamic state space from starting conditions to the desired products, and can be framed as a flux optimisation problem.[62] Identifying such optimal path and assembly of process options can be addressed through hierarchical reinforcement learning (RL), one of AI methods which aims to find a successful strategy of winning in a game given a set of rules.[63] Recently we developed a hierarchical RL methodology for automated process design.[3,64] To provide information to RL on feasible steps to take, we give it access to a knowledge store in the form of a *knowledge graph* of fundamental process phenomena.

A knowledge graph is an assembly of facts and their relationships. In our case, we store knowledge of possible molecular and energy fluxes relevant to specific chemical tasks, such as synthesis, separation, etc. This knowledge base is used to assemble potential designs within an RL environment. The environment provides data on efficiency of the created design. RL process has access to information about all past trials – model evaluations – and guides design towards the most rewarding strategy. Ultimately this approach leads to a ranked set of potential flowsheet designs (Figure 5(a)).

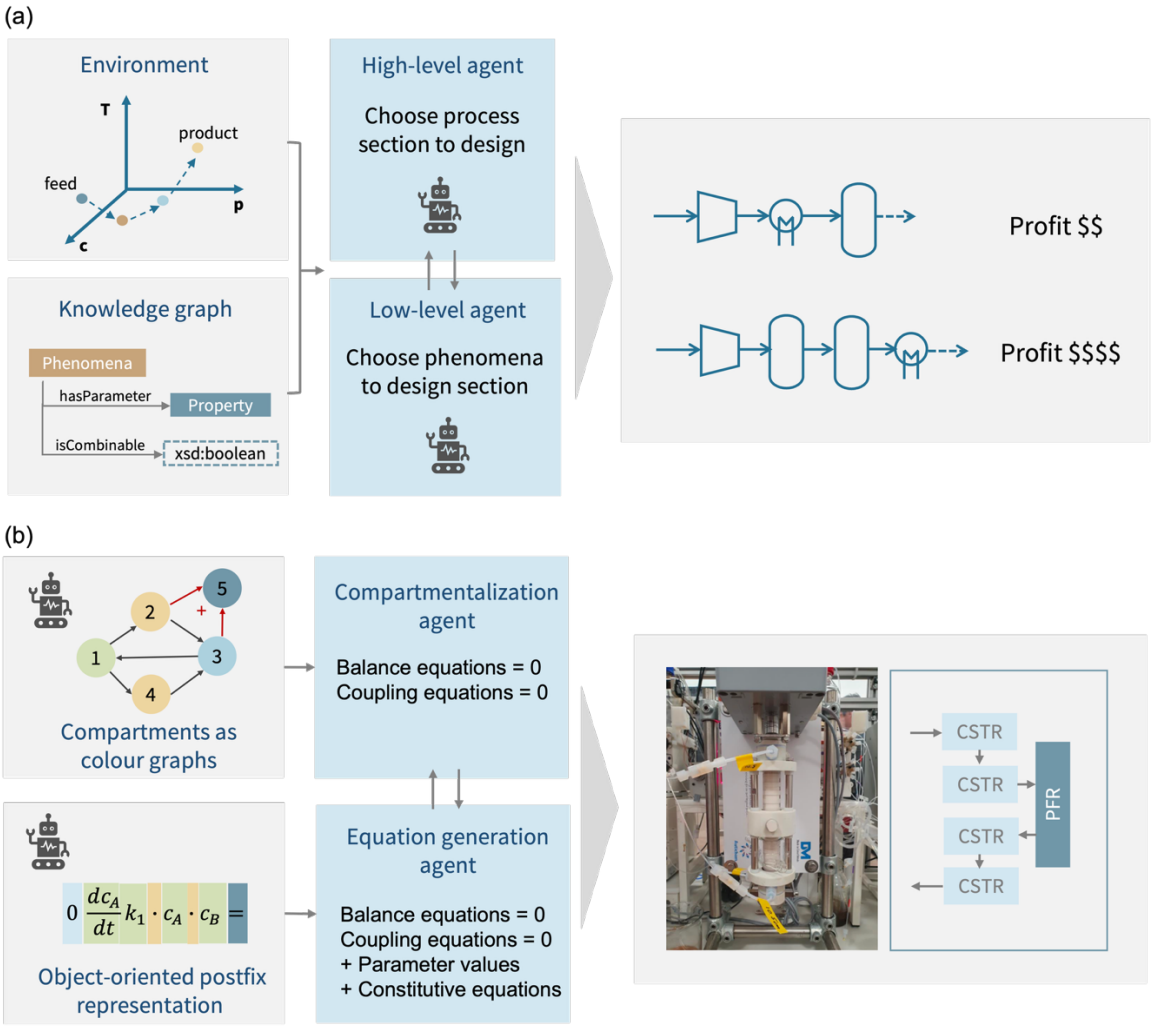


Figure 5. Reinforcement learning agents for (a) automated process design and (b) modelling of novel reactors.

A similar RL methodology can be adopted to automatically assemble first-principles models of reactors or devices, such as new mixing designs, compartmental membrane reactors, multiphase catalytic systems. The modelling of such systems can similarly be formulated as a hierarchical RL problem (Figure 5(b)). We have recently proposed an overall pipeline for automated model generation and demonstrated its performance on creating a hydrodynamic model of a Taylor-Couette flow reactor.[4,65] The pipeline starts from a library of process phenomena and produces executable process models. We explored different ways of formalising human knowledge into the RL agentic workflow by defining graph grammars or reward shaping. We also observed that the modelling ability learned by the agent is transferable from one scenario to another, showing the potential of reuse and generalisation of the strategy.

3.2. Automated life cycle assessment (LCA)

In developing novel technologies, a key focus is on improvements in environmental impacts of novel technologies over the industry benchmarks in delivery of similar functions. Evaluation of environmental impacts is best done through life cycle assessment (LCA), a comprehensive and systematic evaluation of environmental impacts over the full cradle-to-grave life cycle of a product/process against a defined function. While we are now able to automatically construct individual models and even optimal flowsheets, the next challenge is to be able to perform automated LCA of complete process sequences, to support process design and decision-making.

Our early example of integrating LCA with process models for automated process optimisation was shown in a study by Helmdach, Yaseneva et al.[66] Here a two-step chemical conversion-separation sequence of terpenes into useful intermediates was optimised by a multi-objective Bayesian optimisation algorithm which had access to a rigorous process model to estimate economic costs and an LCA software for evaluation of environmental impacts.

A challenge to fully automate LCA is the generation of reliable life cycle inventory data, which is limited by the availability of high-quality, publicly accessible data. One potential solution is to use rigorous process simulation[67] to estimate missing inventory data, which could be further supported by automated process design and model generation workflows. Another approach demonstrated in the literature is to use a decision tree model to estimate the missing data in life cycle inventories.[68] Such a model was trained on gate-to-gate inventory data from of 409 organic chemical processes. This method is based on stoichiometric reactions data, which simplifies the calculation of the required amounts of raw materials. The authors demonstrate improved accuracy of predictions for some impacts, but note that uncertainty of predictions is still very high. Another option is to train statistical ML models based on existing databases to directly link molecular structure with its life cycle environmental impacts.[69,70] This approach generally leads to very large uncertainties due to very small available training datasets and inability of capture process variability by models trained on so little data.

4. Challenges and outlook

Rapid evolution of computing and AI is reshaping the way we design chemistry and chemical processes, and will be affecting manufacture. Several challenges and future development directions are discussed below.

4.1. Data

Data remains a major bottleneck in chemistry R&D. This concerns availability of historical data and generation of new high-quality data.

For the historical data many challenges are well publicised: data may exist in diverse formats, have arbitrary coverage (e.g., unequal exploration of reactivity patterns of molecules with different substitution patterns, or lack of studies of the effects of solvents in reactions), be poorly labelled or mislabelled (when a catalyst appears as a solvent in a record, etc), or only exist behind paywalls or inhouse company repositories. The issue of data quality in historical records is multi-faceted. Apart from data gaps there exist the issue of lack of standardisation in how some data are collected, or as a minimum, lack of metadata on how experiments were conducted, which would allow some normalisation between data collected by different group. There is an issue of data quality as a function of the instrumentation used. Thus, early data on electronic microscopy may only be partially useful, as newer instruments provide significantly better resolution, and, hopefully, also help to reduce the operator-dependency in collecting high-quality images. Because of problems with access and quality of historical data, many groups favour generation of new data.

Advances and broad adoption of lab automation, high-throughput experimentation and process analytical technologies (PAT) offer opportunities to generate data more efficiently. Combination of lab automation with intelligent sampling further improves the quality of collected data for specific applications. Advanced computational techniques, such as transfer learning reduces the need for new data generation.[71] In some cases multidimensional data are needed, such as both temporal and spatial data to reveal details of transport phenomena in multiphase processes. Advanced sensing techniques such as hyperspectral imaging could potentially be useful to efficiently generate experimental multi-dimensional datasets.[72,73]

Lastly, while this review primarily focuses on chemical reactions, both upstream and downstream process steps, such as generation of renewable feedstocks or product separation and chemical products end-of-life conversion, are as important for the overall process design, and data efforts should extend to cover these fields as well.

4.2. Models

As famously noted, “all models are wrong, but some are useful”; models are never perfect. This goes for both, the first principles and the data driven models. First-principles models reflect our understanding of physical reality. They are developed through reductionist approach – generating data that allow to study individual, or few, phenomena. Data driven models and research methods allow to reveal interactions in multi-dimensional spaces. They can enhance and sometimes replace the reductionist methodologies in the study of difficult problems, often promising a significant benefit in terms of the lower experimental budgets, for example through use of active learning. A combination of first-principles models with ML methods, either in the form of physics informed neural networks, or by using physical models as priors in Bayesian methods, allows to

solve inverse problems and provide interpretability to ML. New machine learning methods are also actively being developed to enhance the process of discovery of the underlying first-principles models from data, for example through symbolic regression.[74]

4.3. Infrastructure

The increasing adoption of AI in research workflows requires better infrastructure for storing and using research data. This requires commitment to standardised data management aligned with FAIR (Findable, Accessible, Interoperable, Reusable) principles. Good practices of data management in structural biology enabled breakthroughs such as AlphaFold.[75] Recent efforts in chemistry, such as the Open Reaction Database[76] and materials ontologies,[77] are actively advancing standardisation for chemistry data. This will need to extend to process chemistry and chemical process design. Standardisation does not necessarily mean one standard. Several standards may co-exist, as long as their schemas are published. Here research community is re-discovering the results of computer science from 1980-90s. For example, the issue of interoperability of ontologies has been studied well before ontologies become an important topic.[78] For chemistry data, recording the method of data generation and creating semantically enriched data 'packets', such as RO Crates,[79] seem to be reasonably straightforward to implement. Data could be retrieved from RO Crates and prepared into specific data products to match most suitable data format to the model being trained. Standardisation is also critical for communication between hardware (e.g., between lab automation devices[80]) and software (e.g., interoperability between modelling platforms,[81] between AI agents[82]). Researchers are encouraged to adopt and expand existing standards where possible.

4.4. Training power and investments

While digitalisation and AI offer exciting opportunities for chemistry and engineering, it should be noted that adopting AI itself should be "sustainable" as well.[83] Training large AI models demands substantial energy, often from non-renewable sources. For example, training a 175-billion-parameter model can consume 1,287 million kWh – enough to power a town of 4,000 households for a day.[84] Principles of smart computing – using appropriate tools for the problems at hand – should be taught, popularised and encouraged, alongside other principles of responsible research and innovation frameworks. At the same time, we note that established research institutions are poorly prepared to fully digital R&D. From buildings designed for manual laboratories, to ineffective IT management, outdated data centres and computational resources, to very slow timelines for funding decisions – all require radical rethink in the very near time. Recent public release of a substantial cache of research tools by Google (Gemini for Science) and Anthropic (Claude Science), is indicative of another trend in science. Large corporations have built infrastructure for conducting research using AI agents and are banking on selling access to their

LLMs or agentic environments. Tech companies will be making it as easy as possible to “do research” using such frameworks, which will have several implications.

On a purely technical/commercial side, there is a risk of being locked into a specific provider of ”science services”. “Own your AI”, “own your knowledge graph” are repeated warnings in tech blogs. On the positive site, digital infrastructure for science should lead to lowering the barrier to entry and reaching to a much larger talent pull, likely significantly disrupting the status quo of few “power houses” of science. AI-assisted education is also developing very rapidly and should further support the development of a larger talent base. There is emerging evidence of enhancement in the rate of hypotheses evaluation and generation in the labs that actively use digital tools (AI, automation, data integration).

On the risk side, we are entering the period of proliferation of “junk science”, much the same as we observe with AI-generated content in other fields. The sheer volume of mundane research outputs will require the development of new methods of scoring and filtering. Benchmarking of outputs and proof of ownership of data and ideas become critical to sieve out true talent from routine, and to help promote genuine innovation. Finally, science may not be immune from the risk of over-reliance on AI for routine technical tasks, which are nevertheless required for skills training.

Author contributions

CRediT: J.Z. conceptualization, investigation, visualization, writing - original draft, writing - review & editing; P.Y. investigation, writing - original draft, writing - review & editing; A.A.L. conceptualization, original draft, writing - review & editing, supervision, funding acquisition.

Notes

The authors declare no competing financial interest.

Biographies

Jiyizhe Zhang was a Senior Research Associate working with Prof. Alexei Lapkin at University of Cambridge between 2022-2025. Her research interests focus on AI for chemical process development and chemical separation. She received her Ph.D. from Tsinghua University in 2022, and a jointly awarded Ph.D. from University of Melbourne. She is now a Lecturer (Assistant Professor) in the Department of Chemical Engineering, University of Manchester.

Polina Yaseneva is an Assistant Research Professor at the University of Cambridge, and teaches Sustainability at University College London. Her research is on decision support methods in scaling sustainable technologies, with recent examples in scaling manufacture of novel bio-based nano-materials.

Alexei Lapkin is a Professor of Sustainable Reaction Engineering and the University of Cambridge. Previously he was a Professor of Engineering at the University of Warwick and Reader at the University of Bath. He obtained his PhD from the University of Bath in 2000. He combines academic research with translation to practice through Innovation Centre in Digital Molecular Technologies and Cambridge Centre for Advanced Research and Education in Singapore.

Acknowledgements

This work was in part supported by EPSRC EP/W031019/1 "Bio-derived and Bio-inspired Advanced Materials for Sustainable Industries (VALUED)" and by the project "Sustainable Manufacture of Molecules and Materials" (SM_3) funded by National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme.

References

(1) Fantke, P.; Cinquemani, C.; Yaseneva, P.; De Mello, J.; Schwabe, H.; Ebeling, B.; Lapkin, A. A. Transition to Sustainable Chemistry through Digitalization. *Chem.* **2021**, *7*, 2866–2882. https://doi.org/10.1016/j.chempr.2021.09.012

(2) Weber, J. M.; Guo, Z.; Lapkin, A. A. Discovering Circular Process Solutions through Automated Reaction Network Optimization. *ACS Eng. Au* **2022**, *2* (4), 333–349. https://doi.org/10.1021/acsengineeringau.2c00002

(3) Seidenberg, J. R.; Khan, A. A.; Lapkin, A. A. Boosting Autonomous Process Design and Intensification with Formalized Domain Knowledge. *Computers & Chemical Engineering* **2023**, *169*, 108097.

(4) Laub, J.-F.; Zhang, J.; Heyer, M.; Lapkin, A. A. Automated Generation of Mechanistic Models for Chemical Process Digital Twins Using Reinforcement Learning Part II: Compartmentalization and Learning-Based Recalibration. *Computers & Chemical Engineering* **2026**, *204*, 109384. https://doi.org/10.1016/j.compchemeng.2025.109384

(5) Zhang, S.; Zhang, J.; Lapkin, A. A. A Knowledge Graph Framework for Digital Twins of Chemical Processes. *Nature Chemical Engineering* **2026**, *3*, 286–297. https://doi.org/10.1038/s44286-026-00392-1

(6) *Defossilising the Chemical Industry*; Royal Society. https://royalsociety.org/news-resources/projects/defossilising-chemicals/ (accessed 2025-04-30)

(7) Bauer, F.; Kulionis, V.; Oberschelp, C.; Pfister, S.; Tilsted, J. P.; Finkill, G. D.; Fjäll, S. *Petrochemicals and Climate Change : Tracing Globally Growing Emissions and Key Blind Spots in a Fossil-Based Industry*; Lund University, 2022.

(8) Visinovschi, I.; Mosnier, F. *Lessons in Chemistry. Climate Action Giants*; Planet Tracker, 2025. https://planet-tracker.org/wp-content/uploads/2025/05/Lessons-in-Chemistry.pdf (accessed 2025-11-05)
(9) A. Lapkin; P. Yaseneva. Life Cycle Assessment of Flow Chemistry Processes. In *Sustainable Flow Chemistry - Methods and Applications*; L. Vaccaro, Ed.; Wiley-VCH, 2017; pp 249–276.
(10) Bajorath, J. Milestones in Chemoinformatics: Global View of the Field. *Journal of Cheminformatics* **2024**, *16* (1), 124. https://doi.org/10.1186/s13321-024-00922-0
(11) Ugi, I.; Bauer, J.; Bley, K.; Dengler, A.; Dietz, A.; Fontain, E.; Gruber, B.; Herges, R.; Knauer, M.; Reitsam, K.; Stein, N. Computer-Assisted Solution of Chemical Problems—The Historical Development and the Present State of the Art of a New Discipline of Chemistry. *Angewandte Chemie International Edition in English* **1993**, *32* (2), 201–227. https://doi.org/10.1002/anie.199302011
(12) Venkatasubramanian, V. The Promise of Artificial Intelligence in Chemical Engineering: Is It Here, Finally? *AIChE Journal* **2019**, *65* (2), 466–478. https://doi.org/10.1002/aic.16489
(13) Peremezhney, N.; Jacob, P.-M.; Lapkin, A. Alternative Methods of Processing Bio-Feedstocks in Formulated Consumer Product Design. *Front. Chem.* **2014**, *2*. https://doi.org/10.3389/fchem.2014.00026
(14) Wang, H.; Fu, T.; Du, Y.; Gao, W.; Huang, K.; Liu, Z.; Chandak, P.; Liu, S.; Van Katwyk, P.; Deac, A.; Anandkumar, A.; Bergen, K.; Gomes, C. P.; Ho, S.; Kohli, P.; Lasenby, J.; Leskovec, J.; Liu, T.-Y.; Manrai, A.; Marks, D.; Ramsundar, B.; Song, L.; Sun, J.; Tang, J.; Veličković, P.; Welling, M.; Zhang, L.; Coley, C. W.; Bengio, Y.; Zitnik, M. Scientific Discovery in the Age of Artificial Intelligence. *Nature* **2023**, *620* (7972), 47–60. https://doi.org/10.1038/s41586-023-06221-2
(15) E. Abbate; I. Garmendia Aguirre; G. Bracalente; L. Mancini; D. Tosches; K. Rasmussen; M.J. Bennett; S. Sala. Safe and Sustainable by Design Chemicals and Materials - Methodological Guidance, 2024. https://publications.jrc.ec.europa.eu/repository/handle/JRC138035
(16) Think Feedstocks First. *Nature Chemical Engineering* **2024**, *1*, 261.
(17) Guo, Z.; Yan, N.; Lapkin, A. A. Towards Circular Economy: Integration of Bio-Waste into Chemical Supply Chain. *Current Opinion in Chemical Engineering* **2019**, *26*, 148–156. https://doi.org/10.1016/j.coche.2019.09.010
(18) Weber, J. M.; Lió, P.; Lapkin, A. A. Identification of Strategic Molecules for Future Circular Supply Chains Using Large Reaction Networks. *React. Chem. Eng.* **2019**, *4* (11), 1969–1981. https://doi.org/10.1039/C9RE00213H
(19) Tibbetts, J. D.; Russo, D.; Lapkin, A. A.; Bull, S. D. Efficient Syntheses of Biobased Terephthalic Acid, p-Toluic Acid, and p-Methylacetophenone via One-Pot Catalytic Aerobic Oxidation of Monoterpene Derived Bio-p-Cymene. *ACS Sustain. Chem. Eng.* **2021**, *9* (25), 8642–8652. https://doi.org/10.1021/acssuschemeng.1c02605
(20) Kong, D.; Dolzhenko, A. V. Cyrene: A Bio-Based Sustainable Solvent for Organic Synthesis. *Sustainable Chemistry and Pharmacy* **2022**, *25*, 100591. https://doi.org/10.1016/j.scp.2021.100591
(21) Hao, Z.; Hamad, W. Y.; Yaseneva, P. Understanding the Environmental Impacts of Large-Scale Cellulose Nanocrystals Production: Case Studies in Regions Dependent on Renewable and Fossil Fuel Energy Sources. *Chemical Engineering Journal* **2023**, *478*, 147160. https://doi.org/10.1016/j.cej.2023.147160

(22) Grzybowski, B. A.; Bishop, K. J. M.; Kowalczyk, B.; Wilmer, C. E. The “wired” Universe of Organic Chemistry. *Nature Chem.* **2009**, *1* (1), 31–36.
(23) Jacob, P.-M.; Lapkin, A. Statistics of the Network of Organic Chemistry. *Reaction Chemistry & Engineering* **2018**, *3* (1), 102–118. https://doi.org/10.1039/C7RE00129K
(24) Weber, J. M.; Guo, Z.; Zhang, C.; Schweidtmann, A. M.; Lapkin, A. A. Chemical Data Intelligence for Sustainable Chemistry. *Chemical Society Reviews* **2021**, *50*, 12013–12036. https://doi.org/10.1039/D1CS00477H
(25) Reaxys. *Elsevier R&D Solutions, Reaxys Fact Sheet (2018)*. https://supportcontent.elsevier.com/RightNow%20Next%20Gen/Reaxys/New_RX_FactSheet_Jul_2018.pdf (accessed 2025-05-29)
(26) Pham, T. T.; Guo, Z.; Li, B.; Lapkin, A. A.; Yan, N. Synthesis of Pyrrole-2-Carboxylic Acid from Cellulose- and Chitin-Based Feedstocks Discovered by the Automated Route Search. *ChemSusChem* **2024**, *17* (3), e202300538. https://doi.org/10.1002/cssc.202300538
(27) A. Lapkin, A.; K. Heer, P.; Jacob, P.-M.; Hutchby, M.; Cunningham, W.; D. Bull, S.; G. Davidson, M. Automation of Route Identification and Optimisation Based on Data-Mining and Chemical Intuition. *Faraday Discussions* **2017**, *202* (0), 483–496. https://doi.org/10.1039/C7FD00073A
(28) Guo, Z.; Yan, N.; Lapkin, A. A. Towards Circular Economy: Integration of Bio-Waste into Chemical Supply Chain. *Current Opinion in Chemical Engineering* **2019**, *26*, 148–156. https://doi.org/https://doi.org/10.1016/j.coche.2019.09.010
(29) Jacob, P.-M.; Yamin, P.; Perez-Storey, C.; Hopgood, M.; Lapkin, A. A. Towards Automation of Chemical Process Route Selection Based on Data Mining. *Green Chem.* **2017**, *19*, 140–152. https://doi.org/10.1039/c6gc02482c
(30) Weber, J. M.; Schweidtmann, A. M.; Nolasco, E.; Lapkin, A. A. Modelling Circular Structures in Reaction Networks: Petri Nets and Reaction Network Flux Analysis. In *Computer Aided Chemical Engineering*; Pierucci, S., Manenti, F., Bozzano, G. L., Manca, D., Eds.; 30 European Symposium on Computer Aided Process Engineering; Elsevier, 2020; Vol. 48, pp 1843–1848. https://doi.org/10.1016/B978-0-12-823377-1.50308-6
(31) Coley, C. W.; Green, W. H.; Jensen, K. F. Machine Learning in Computer-Aided Synthesis Planning. *Acc. Chem. Res.* **2018**, *51* (5), 1281–1289. https://doi.org/10.1021/acs.accounts.8b00087
(32) Zhang, C.; Lapkin, A. A. Reinforcement Learning Optimization of Reaction Routes on the Basis of Large, Hybrid Organic Chemistry–Synthetic Biological, Reaction Network Data. *React. Chem. Eng.* **2023**, *8* (10), 2491–2504. https://doi.org/10.1039/D2RE00406B
(33) Cheetham, A. K.; Seshadri, R.; Wudl, F. Chemical Synthesis and Materials Discovery. *Nat. Synth* **2022**, *1* (7), 514–520. https://doi.org/10.1038/s44160-022-00096-3
(34) Shambhawi, S.; Csányi, G.; Lapkin, A. A. Active Learning Training Strategy for Predicting O Adsorption Free Energy on Perovskite Catalysts Using Inexpensive Catalyst Features. *Chemistry–Methods* **2021**, *1* (10), 444–450. https://doi.org/10.1002/cmtd.202100035
(35) Hashemi, A.; Bougueroua, S.; Gaigeot, M.-P.; Pidko, E. A. HiREX: High-Throughput Reactivity Exploration for Extended Databases of Transition-Metal Catalysts. *J. Chem. Inf. Model.* **2023**, *63* (19), 6081–6094. https://doi.org/10.1021/acs.jcim.3c00660
(36) Assante, M.; Pahlen, S.; Pijper, B.; Cao, L.; Specht, O.; Simeth, N.; Lapkin, A.; Green, C.; Engkvist, O.; Bergonzini, G.; Kabeshov, M. Automated DFT-Machine Learning Integration

Enables Data-Efficient and Generalizable Feasibility Predictions in Metallaphotoredox Sp2–Sp3 Cross-Coupling Reactions. *ACS Catal.* **2026**, *16* (4), 3547–3560. https://doi.org/10.1021/acscatal.5c07857

(37) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A. Commentary: The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation. *APL Materials* **2013**, *1* (1), 011002. https://doi.org/10.1063/1.4812323

(38) Takahashi, L.; Miyazato, I.; Takahashi, K. Redesigning the Materials and Catalysts Database Construction Process Using Ontologies. *J. Chem. Inf. Model.* **2018**, *58* (9), 1742–1754. https://doi.org/10.1021/acs.jcim.8b00165

(39) Schaaf, L. L.; Fako, E.; De, S.; Schäfer, A.; Csányi, G. Accurate Energy Barriers for Catalytic Reaction Pathways: An Automatic Training Protocol for Machine Learning Force Fields. *npj Comput Mater* **2023**, *9* (1), 1–10. https://doi.org/10.1038/s41524-023-01124-2

(40) Chattoraj, J.; Hamadicharef, B.; Syadzali, Y. N. A.; Limantara, G. O.; Zeng, Y.; Poh, C. K.; Chen, L.; Tan, T. L. Preparation of a Water–Gas Shift Database to Evaluate the Performance of Noble Metal Catalysts Using Theory-Guided Machine Learning. *ACS Catal.* **2023**, *13* (21), 14334–14345. https://doi.org/10.1021/acscatal.3c04467

(41) Mok, D. H.; Li, H.; Zhang, G.; Lee, C.; Jiang, K.; Back, S. Data-Driven Discovery of Electrocatalysts for CO2 Reduction Using Active Motifs-Based Machine Learning. *Nat Commun* **2023**, *14* (1), 7303. https://doi.org/10.1038/s41467-023-43118-0

(42) Shambhawi; Mohan, O.; Choksi, T. S.; Lapkin, A. A. The Design and Optimization of Heterogeneous Catalysts Using Computational Methods. *Catal. Sci. Technol.* **2024**, *14* (3), 515–532. https://doi.org/10.1039/D3CY01160G

(43) Houben, C.; Lapkin, A. A. Automatic Discovery and Optimization of Chemical Processes. **2015**. https://doi.org/10.1016/j.coche.2015.07.001

(44) 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* (6), 3089–3126. https://doi.org/10.1021/acs.chemrev.2c00798

(45) Peremezhney, N.; Hines, E.; Lapkin, A.; Connaughton, C. Combining Gaussian Processes, Mutual Information and a Genetic Algorithm for Multi-Target Optimization of Expensiveto-Evaluate Functions. *Engng. Optim.* **2014**, *46*, 1593–1607. https://doi.org/10.1080/0305215X.2014.881997

(46) Bradford, E.; Schweidtmann, A. M.; Lapkin, A. Efficient Multiobjective Optimization Employing Gaussian Processes, Spectral Sampling and a Genetic Algorithm. *J. Global Optim.* **2018**, *71*, 407–438. https://doi.org/10.1007/s10898-018-0609-2

(47) Jorayev, P.; Russo, D.; Tibbetts, J. D.; Schweidtmann, A. M.; Deutsch, P.; Bull, S. D.; Lapkin, A. A. Multi-Objective Bayesian Optimisation of a Two-Step Synthesis of p-Cymene from Crude Sulphate Turpentine. *Chemical Engineering Science* **2022**, *247*, 116938. https://doi.org/10.1016/j.ces.2021.116938

(48) Jorayev, P.; Soritz, S.; Sung, S.; Jeraal, M. I.; Russo, D.; Barthelme, A.; Toussaint, F. C.; Gaunt, M. J.; Lapkin, A. A. Machine Learning-Driven Optimization of Continuous-Flow Photoredox Amine Synthesis. *Org. Process Res. Dev.* **2025**, *29* (6), 1411–1422. https://doi.org/10.1021/acs.oprd.4c00533

(49) Tom, G.; Schmid, S. P.; Baird, S. G.; Cao, Y.; Darvish, K.; Hao, H.; Lo, S.; Pablo-García, S.; Rajaonson, E. M.; Skreta, M.; Yoshikawa, N.; Corapi, S.; Akkoc, G. D.; Strieth-Kalthoff, F.; Seifrid, M.; Aspuru-Guzik, A. Self-Driving Laboratories for Chemistry and Materials Science. *Chem. Rev.* **2024**, *124* (16), 9633–9732. https://doi.org/10.1021/acs.chemrev.4c00055
(50) Häse, F.; Aldeghi, M.; Hickman, R. J.; Roch, L. M.; Aspuru-Guzik, A. GRYFFIN: An Algorithm for Bayesian Optimization of Categorical Variables Informed by Expert Knowledge. *Appl. Phys. Rev.* **2021**, *8* (3), 031406. https://doi.org/10.1063/5.0048164
(51) Wang, Y.; Chen, T.-Y.; Vlachos, D. G. NEXTorch: A Design and Bayesian Optimization Toolkit for Chemical Sciences and Engineering. *J. Chem. Inf. Model.* **2021**, *61* (11), 5312–5319. https://doi.org/10.1021/acs.jcim.1c00637
(52) Felton, K. C.; Rittig, J. G.; Lapkin, A. A. Summit: Benchmarking Machine Learning Methods for Reaction Optimisation. *Chem.-Methods* **2021**, *1* (2), 116–122. https://doi.org/10.1002/cmtd.202000051
(53) Fitzner, M.; Šošić, A.; Hopp, A. V.; Müller, M.; Rihana, R.; Hrovatin, K.; Liebig, F.; Winkel, M.; Halter, W.; Brandenburg, J. G. BayBE: A Bayesian Back End for Experimental Planning in the Low-to-No-Data Regime. *Digital Discovery* **2025**, *4* (8), 1991–2000. https://doi.org/10.1039/D5DD00050E
(54) Dürholt, J. P.; Asche, T. S.; Kleinekorte, J.; Mancino-Ball, G.; Schiller, B.; Sung, S.; Keupp, J.; Osburg, A.; Boyne, T.; Misener, R.; Eldred, R.; Costa, W. S.; Kappatou, C.; Lee, R. M.; Linzner, D.; Walz, D.; Wulkow, N.; Shafei, B. BoFire: Bayesian Optimization Framework Intended for Real Experiments. *Journal of Machine Learning Research* **2025**, *26*, 1–7.
(55) Wigh, D. S.; Goodman, J. M.; Lapkin, A. A. A Review of Molecular Representation in the Age of Machine Learning. *WIREs computational Molecular Science* **2022**, *n/a* (n/a), e1603. https://doi.org/https://doi.org/10.1002/wcms.1603
(56) Struebing, H.; Ganase, Z.; Karamertzanis, P. G.; Siougkrou, E.; Haycock, P.; Piccione, P. M.; Armstrong, A.; Galindo, A.; Adjiman, C. S. Computer-Aided Molecular Design of Solvents for Accelerated Reaction Kinetics. *Nature Chem* **2013**, *5* (11), 952–957. https://doi.org/10.1038/nchem.1755
(57) Amar, Y.; Schweidtmann, A. M.; Deutsch, P.; Cao, L.; Lapkin, A. Machine Learning and Molecular Descriptors Enable Rational Solvent Selection in Asymmetric Catalysis. *Chem. Sci.* **2019**, *10* (27), 6697–6706. https://doi.org/10.1039/C9SC01844A
(58) Zhang, C.; Amar, Y.; Cao, L.; Lapkin, A. A. Solvent Selection for Mitsunobu Reaction Driven by an Active Learning Surrogate Model. *Org. Process Res. Dev.* **2020**, *24*, 2864–2873. https://doi.org/10.1021/acs.oprd.0c00376
(59) Lapkin, A.; Voutchkova, A.; Anastas, P. A Conceptual Framework for Description of Complexity in Intensive Chemical Processes. *Chem. Engng. Processing: Process Intensification* **2011**, *50*, 1027–1034.
(60) Zhang, S.; Tan, Y. R.; Nguyen, C. M.; Karan, D.; Ganguly, S.; Jose, N.; Lim, M. Q.; Guo, S.; Jiang, L. L.; Kraft, M.; Lapkin, A. A. A Semantic Framework for Chemical Process Digitalisation Using Ontologies. *Chemical Engineering Journal* **2026**, 174361. https://doi.org/10.1016/j.cej.2026.174361
(61) Freund, H.; Sundmacher, K. Towards a Methodology for the Systematic Analysis and Design of Efficient Chemical Processes. *Chemical Engineering and Processing: Process Intensification* **2008**, *47* (12), 2051–2060. https://doi.org/10.1016/j.cep.2008.07.011

(62) Schack, D.; Liesche, G.; Sundmacher, K. The FluxMax Approach: Simultaneous Flux Optimization and Heat Integration by Discretization of Thermodynamic State Space Illustrated on Methanol Synthesis Process. *Chemical Engineering Science* **2020**, *215*, 115382. https://doi.org/10.1016/j.ces.2019.115382
(63) Sutton, R. S.; Barto, A. *Reinforcement Learning: An Introduction*, Second edition.; Adaptive computation and machine learning; The MIT Press: Cambridge, Massachusetts London, England, 2020.
(64) Khan, A. A.; Lapkin, A. A. Designing the Process Designer: Hierarchical Reinforcement Learning for Optimisation-Based Process Design. *Chem. Eng. Process. - Process. Intensif.* **2022**, *180*, 108885. https://doi.org/10.1016/j.cep.2022.108885
(65) Heyer, M.; Zhang, J.; Sugisawa, N.; Laub, J.-F.; Lapkin, A. A. Automated Generation of Mechanistic Models for Chemical Process Digital Twins Using Reinforcement Learning Part I: Conceptual Framework and Equation Generation. *Computers & Chemical Engineering* **2025**, *202*, 109281. https://doi.org/10.1016/j.compchemeng.2025.109281
(66) Helmdach, D.; Yaseneva, P.; Heer, P. K.; Schweidtmann, A. M.; Lapkin, A. A. A Multiobjective Optimization Including Results of Life Cycle Assessment in Developing Biorenewables-Based Processes. *ChemSusChem* **2017**, *10* (18), 3632–3643. https://doi.org/10.1002/cssc.201700927
(67) Hallamasek, S.; Ubbenjans, V.; Lapkin, A. A. Life Cycle Assessment of a Process for Paracetamol Flow Synthesis from Bio-Waste Derived β-Pinene. *Sustainable Chemistry and Pharmacy* **2024**, *40*, 101629. https://doi.org/10.1016/j.scp.2024.101629
(68) Langhorst, T.; Winter, B.; Tuchschmid, M.; Roskosch, D.; Bardow, A. From Reaction Stoichiometry to Life Cycle Assessment: Decision Tree-Based Estimation Tool. *ACS Environ. Au* **2025**, *5* (6), 550–560. https://doi.org/10.1021/acsenvironau.4c00065
(69) Kleinekorte, J.; Kleppich, J.; Fleitmann, L.; Beckert, V.; Blodau, L.; Bardow, A. APPROPRIATE Life Cycle Assessment: A PROcess-Specific, PRedictive Impact AssessmenT Method for Emerging Chemical Processes. *ACS Sustainable Chem. Eng.* **2023**, *11* (25), 9303–9319. https://doi.org/10.1021/acssuschemeng.2c07682
(70) Blanco, C. F.; Pauliks, N.; Donati, F.; Engberg, N.; Weber, J. Machine Learning to Support Prospective Life Cycle Assessment of Emerging Chemical Technologies. *Current Opinion in Green and Sustainable Chemistry* **2024**, *50*, 100979. https://doi.org/10.1016/j.cogsc.2024.100979
(71) Taylor, C. J.; Felton, K. C.; Wigh, D.; Jeraal, M. I.; Grainger, R.; Chessari, G.; Johnson, C. N.; Lapkin, A. A. Accelerated Chemical Reaction Optimization Using Multi-Task Learning. *ACS Cent. Sci.* **2023**. https://doi.org/10.1021/acscentsci.3c00050
(72) Zavala, V. M. Outlook: How I Learned to Love Machine Learning (A Personal Perspective on Machine Learning in Process Systems Engineering). *Ind. Eng. Chem. Res.* **2023**, *62* (23), 8995–9005. https://doi.org/10.1021/acs.iecr.3c01565
(73) Chemla, Y.; Levin, I.; Fan, Y.; Johnson, A. A.; Coley, C. W.; Voigt, C. A. Hyperspectral Reporters for Long-Distance and Wide-Area Detection of Gene Expression in Living Bacteria. *Nat Biotechnol* **2025**, 1–11. https://doi.org/10.1038/s41587-025-02622-y
(74) Neumann, P.; Cao, L.; Russo, D.; Vassiliadis, V. S.; Lapkin, A. A. A New Formulation for Symbolic Regression to Identify Physico-Chemical Laws from Experimental Data. *Chem. Eng. J.* **2020**, *387*, 123412. https://doi.org/https://doi.org/10.1016/j.cej.2019.123412

(75) Sauro, H. M.; Bergmann, F. T. Standards and Ontologies in Computational Systems Biology. *Essays in Biochemistry* **2008**, *45*, 211–222. https://doi.org/10.1042/bse0450211
(76) 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* (45), 18820–18826. https://doi.org/10.1021/jacs.1c09820
(77) Rajamohan, B. P.; Bradley, A. C. H.; Tran, V. D.; Gordon, J. E.; Caldwell, H. W.; Mehdi, R.; Ponon, G.; Tran, Q. D.; Dernek, O.; Kaltenbaugh, J.; Pierce, B. G.; Wieser, R.; Yue, W.; Lin, K.; Kambo, J.; Lopez, C.; Nihar, A.; Savage, D. J.; Brown, D. W.; Sharma, H.; Giera, B.; Tripathi, P. K.; Wu, Y.; Li, M.; Davis, K. O.; Bruckman, L. S.; Barcelos, E. I.; French, R. H. Materials Data Science Ontology(MDS-Onto): Unifying Domain Knowledge in Materials and Applied Data Science. *Sci Data* **2025**, *12* (1), 628. https://doi.org/10.1038/s41597-025-04938-5
(78) Gruber, T. R. A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition* **1993**, *5* (2), 199–220. https://doi.org/10.1006/knac.1993.1008
(79) Soiland-Reyes, S.; Sefton, P.; Crosas, M.; Castro, L. J.; Coppens, F.; Fernández, J. M.; Garijo, D.; Grüning, B.; La Rosa, M.; Leo, S.; Ó Carragáin, E.; Portier, M.; Trisovic, A.; RO-Crate Community; Groth, P.; Goble, C. Packaging Research Artefacts with RO-Crate. *Data Science* **2022**, *5* (2), 97–138. https://doi.org/10.3233/DS-210053
(80) Bär, H.; Hochstrasser, R.; Papenfuß, B. SiLA: Basic Standards for Rapid Integration in Laboratory Automation. *SLAS Technology* **2012**, *17* (2), 86–95. https://doi.org/10.1177/2211068211424550
(81) van Baten, J.; Pons, M. CAPE-OPEN: Interoperability in Industrial Flowsheet Simulation Software. *Chemie Ingenieur Technik* **2014**, *86* (7), 1052–1064. https://doi.org/10.1002/cite.201400009
(82) *AI Agent Standards Initiative*. NIST. https://www.nist.gov/artificial-intelligence/ai-agent-standards-initiative (accessed 2026-07-09)
(83) *Engineering Responsible AI: Foundations for Environmentally Sustainable AI*; National Engineering Policy Centre, 2025. https://nepc.raeng.org.uk/media/2aggau2j/foundations-for-sustainable-ai-nepc-report.pdf
(84) Luccioni, A. S.; Viguier, S.; Ligozat, A.-L. Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model. arXiv November 3, 2022. https://doi.org/10.48550/arXiv.2211.02001