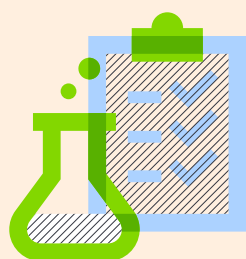


Unlocking greenness latent in chemical data

Reaxys® data are the raw material of a new generation of chemists looking to radically change how we make things



Data networks are digital laboratories for a generation of chemical engineers mastering a new way of tackling chemistry problems. They explore novel reaction spaces and greener ways to make value-added compounds by predicting the unknown from patterns in molecular interactions. Chonghuan Zhang and Adarsh Arun, two young researchers working with Professor Alexei Lapkin (shown above) at the University of Cambridge, share how they use Reaxys data to improve manufacturing processes.



Sustainable
Reaction
Engineering

The Sustainable Reaction Engineering group at the University of Cambridge is a highly motivated team of young researchers and engineers led by Professor Alexei Lapkin. The group develops clean chemistries through modeling, process improvement and life cycle optimization. As part of their work, they construct and tap into reaction networks using data-driven analytics to create sustainable synthesis routes. To this end, the group has successfully queried and used data from Reaxys for years.

The work is challenging. The reactions they seek are complex. They must consider multiple parameters: yield, reaction speed, byproducts, process economics and environmental impact. Plus, for each reaction they must define an optimal combination of several elements: feedstock, reactants, solvent, catalyst, conditions and so on. Yet, the team sees tremendous power in using modern methods like machine learning to reinvent conventional chemical synthesis and design cleaner manufacturing processes.

Working with data means knowing data

The networks built by students in Professor Lapkin's group consist of millions of reactions. Each node of the network is a molecule characterized by a range of classification and property data. Those nodes are connected by edges representing reactions, each with a cataloguing and descriptive dataset. Gleaning insights from these networks requires a deep understanding of the data they contain and gaining that understanding is a learning process.

Chonghuan Zhang and Adarsh Arun are students in the Lapkin group at bookends of doctoral training. Chonghuan is in his final year, finishing up his project and writing up his findings. Originally from China, he has a degree in Chemical Engineering from the University of Sydney. Adarsh has started his second year and is defining his research question and strategy. He joined the group after studying at the National University of Singapore and completing a master's degree in Chemical Engineering at Cambridge. Both students shared insights into their work and the role of Reaxys data in advancing their research projects and supporting their training.



Merging cheminformatics and bioinformatics

Industrial and pharmaceutical chemistry traditionally operate with a limited set of tools. Organic synthesis has been the workhorse of manufacturing but represents only a fraction of the complete landscape of available reactions. So, can we leverage our knowledge of conventional reactions to expand the toolset with commercially feasible alternatives from the remaining space of reactions?

To answer that question, Chonghuan examines the potential of enzymatic transformations in solving chemical engineering problems and producing functional molecules. “I’m developing a method to hybridize the emerging field of synthetic biology with conventional organic synthesis,” explains Chonghuan. “I merge conventional organic chemistry and enzymatic chemistry into a hybrid reaction network. I then examine that network for economical production routes of complex molecules and assess how much added value synthetic biology reactions create.”

Chonghuan’s research project expands the degrees of freedom in synthetic route prediction by bringing together cheminformatic data from Reaxys and metabolic reaction data from the KEGG (Kyoto Encyclopedia of Genes and Genomes) database. “I want to find alternative opportunities in the production of chemicals using key synthetic biology reactions — that is, enzymatic or metabolic reactions — to replace uneconomical synthesis steps based on harsh catalytic reactions.”

At this stage, Chonghuan is assessing the feasibility and value of his method. “We set up a few rules to define the goodness and greenness of reactions. We then use reinforcement learning to create synthetic routes based just on organic reaction data, just on biosynthesis data and based on the hybrid data. Then we can see how good the hybrid synthesis is compared with organic and synthetic biology alone.” As a next step, the predicted hybrid synthesis routes would be assessed in solving real-world manufacturing problems. Chonghuan is confident that the routes are feasible. “Our reaction network builds on historical, literature-excerpted reactions of Reaxys and other sources. We believe the predictions we produce are doable.”

The screenshot shows the Reaxys search interface. The search criteria are set to 'Green chemistry'. The results show 98,413 reactions. Two reaction schemes are visible: one for the hydrogenation of methyl acrylate to methyl propylamine using a nickel catalyst, and another for the hydrogenation of methyl acrylate to methyl propylamine using a nickel catalyst. The search results table includes columns for 'Yield' and 'Reference'.

Reaxys search results of reactions for 'green chemistry'

Predicting byproducts and impurities

Adarsh uses networks and machine learning to develop sustainable manufacturing paths that integrate biomass or biowaste into chemical reactions that generate value-added chemicals. For one aspect of his work, Adarsh looks at byproducts of reactions. “The efficient and responsible manufacture of high-value chemicals starts with design. Any synthetic plan will invariably have impurities. If you can predict at an early stage where impurities arise, you can design an optimal process to separate them. If they’re toxic, you can develop greener routes. But to do that, you need to see the whole picture — not just the main product but also the byproducts. There are a lot of data in Reaxys that can be leveraged for these data-driven predictions.”

These impurity prediction tools are still a work in progress, but Adarsh anticipates that they will streamline experimentation and optimization during process development. “I did several case studies examining common drug molecules and process impurities mentioned in the literature. Authors had used HPLC (high-performance, or high-pressure, liquid chromatography) methods to elucidate these impurities. With this prediction tool, you can anticipate those beforehand and support early-stage decisions to better target experiments.”

The screenshot shows the Reaxys search interface with search criteria set to 'Alexei Lapkin'. The results list several documents, including 'Transformation of Corn Lignin into Sun Cream Ingredients', 'Searching for optimal process routes: A reinforcement learning approach', 'Introduction to green chemistry and reaction engineering', 'CONSTANT SHEAR CONTINUOUS REACTOR DEVICE', 'CONSTANT SHEAR CONTINUOUS REACTOR DEVICE', 'Continuous synthesis of doped layered double hydroxides in a meso-scale flow reactor', 'Dynamic Optimization and Non-linear Model Predictive Control to Achieve Targeted Particle Morphologies', and 'Biosynthesis of spathulenol and camphor stand as a competitive route to artemisinin production as revealed by a new chemometric convergence approach based on nine locations' field-grown Artemisia annua L.

Reaxys search results for documents authored by 'Alexei Lapkin'

For Adarsh, a key aspect of conceptualizing and building these impurity prediction tools is gaining experience with the nature and structure of the data flowing into his reaction networks. For him, Reaxys is a training ground. “Another motivation for this work is to get acquainted with Reaxys data as it will fit into my PhD work; learn how to leverage it, understand its breadth, depth and limitations,” he says. Through his exploration of Reaxys data, Adarsh is laying out the roadmap for a broader project on biomass. “I want to unlock the potential of biomass and biowaste and sustainably derive value from them. A reaction network fits this problem really well and Reaxys has the bits and pieces to my problem. I just need to pull them together. I’m looking forward to it.”

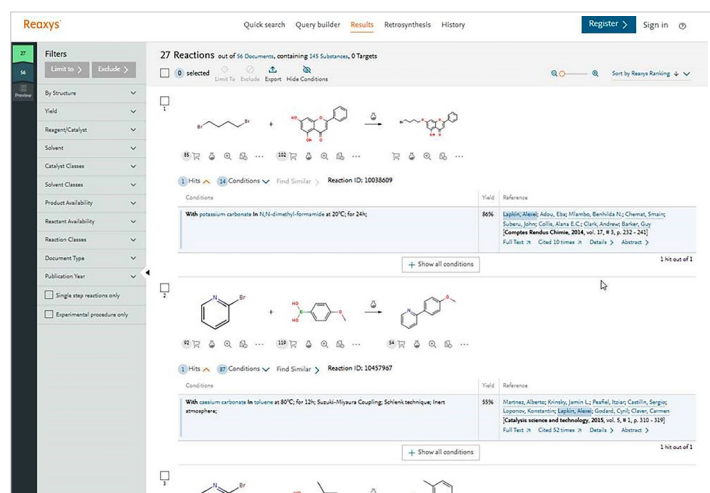
Benefits of Reaxys: number and diversity of reactions

The intersection of artificial intelligence methods and chemistry was initially explored decades ago, but practical implementation was hindered by limited access to sufficient data. The data-intensive infrastructure needed to feed models and machine learning algorithms simply did not



exist. Today, the Lapkin group downloads millions of reaction entries and condition data from Reaxys via API queries. These data are then used and augmented by students as needed for specific projects. In a way, Reaxys data are an extensive and diverse scaffold for their reaction networks.

For Chonghuan, the broad range of reaction and molecule types was instrumental in his work. “With Reaxys, we can explore a large portion of the reaction space,” he says. Additionally, he relied on extracted property, condition and reaction component data. “To analyze the hybrid reaction network, we looked not only at reactions and molecules themselves, but also at related attributes like reaction conditions, catalyst, ligand, solvents and more. Reaxys has a lot of these data types.”



The screenshot shows the Reaxys search results page. It features a sidebar with filters, a main search area with a query builder, and a list of search results. Two reaction examples are highlighted. The first reaction is: CCCCC1=CC=CC=C1 + C1=CC=C(C=C1)C(=O)O >> CCCC1=CC=CC=C1C(=O)O. The conditions are: **With potassium carbonate in N,N-dimethyl-formamide at 20°C for 24h.** The second reaction is: C1=CC=C(C=C1)C(=O)O + C1=CC=C(C=C1)C(=O)O >> C1=CC=C(C=C1)C(=O)O. The conditions are: **With potassium carbonate in toluene at 80°C for 12h; Substrate Mixture Grouping; Solvent technique; then evaporate.**

Reaxys search results of reactions contained within Alexei Lapkin publications

Adarsh’s exploration of Reaxys data so far has led him to similar conclusions. “Reaxys has millions of reactions and molecules. In terms of coverage, it has a broader variety of reaction classes than a lot of other databases. And in my particular work, classifications like reactants, products, reagents, solvents, catalysts, as well as reaction conditions like temperature, pressure and reaction time are very useful. Reaxys captures that spectrum of information.”

Powering predictive augmentation

The utility of Reaxys data as a scaffold for reaction networks reflects the need for high granularity when exploring uncharted reaction spaces. Given the vast amount of data that goes into building reaction networks, the group invests time into understanding data quality to maximize the value they get from Reaxys. Where possible, they compare, complement and process Reaxys data to integrate information sourced from other databases. For example, Chonghuan was able to hybridize KEGG and Reaxys data by constructing an in-house database of canonical SMILES (simplified molecular-input line-entry system) strings to map metabolic pathways to reactions.

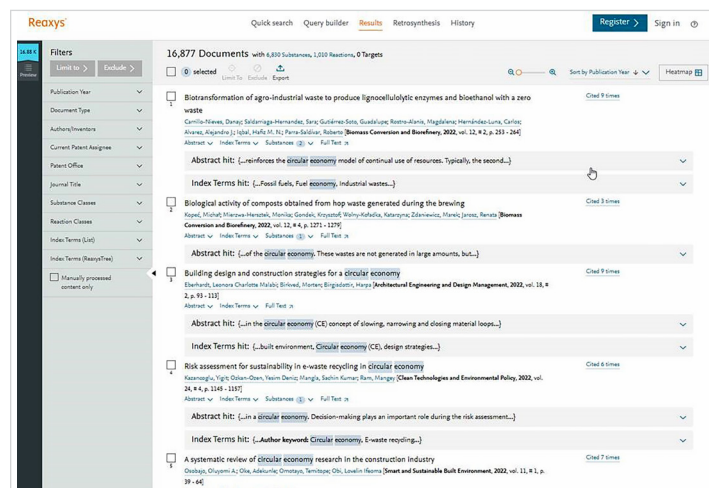
Intriguingly, the group also enriches their networks with predictions based on a priori knowledge from the full body of reactions in Reaxys. Take Chonghuan’s work: compared to organic synthesis, reactions for enzymatic synthesis are sparse. So, he uses insights about known organic and enzymatic reactions and molecule fragments to expand the enzymatic options at each step of a synthesis route. “We want to

increase the density of the bioinformatic space,” he says. “We don’t have the ambition of creating as big a space as the chemical one, but we want to find more reaction opportunities from either reaction templates, from machine learning methods like graph neural networks, or via mechanistic predictions. We seek the right enzyme for a given step and evaluate if the predicted reaction is going to work.” By further enriching his hybrid reaction network, Chonghuan will also be able to assess the merger of biosynthesis and organic reactions against more sophisticated criteria. He strongly believes that the expanded knowledge will further strengthen the demonstrated value of this hybrid approach.

In essence, Chonghuan is building a new layer of information into his reaction network based on patterns that emerge from all reactions in Reaxys and KEGG. Adarsh does so as well in predicting byproducts and impurities. He evaluates how functional groups of reactants interact across all Reaxys reactions to generate a list of potential products for individual reactions sorted by relevance. “I basically look at molecular fragments based on the functional groups present and find all reactions in Reaxys that contain those fragments. I then determine how those fragments interact and create a reaction template. Finally, I apply that template back to the original reaction to suggest new byproducts and impurities,” says Adarsh. And here, coverage of Reaxys is instrumental. “In a lot of databases, certain niche classes of reactions aren’t well represented; but in Reaxys, they do exist. That coverage allows me to capture most of the ways in which these functional groups can interact.”

A roadmap to sustainability

Both Chonghuan and Adarsh emphasize that their work is only part of a much larger picture. Talking about the group’s ambitions, Chonghuan says, “The whole idea is to create a circular economy and green synthesis of molecules. Instead of the linear processing of fossil fuels to target molecules with certain byproducts, we want to use byproducts from one reaction as feedstock to another and replace fossil fuels with biomass.” Adarsh adds, “In the future, I envision an extensive knowledge graph where you can start from anywhere — a country, biomass or target chemical — and navigate a path to a valuable biomass or an optimal synthesis route. I find that suggesting paths for discovery, novel synthesis options and optimal reaction routes can save time and open new opportunities. That’s a win. And I think that what we are doing today could lead the way.”



The screenshot shows the Reaxys search results page for the query 'circular economy'. It displays 16,877 documents. Several results are highlighted with their abstracts and index terms. For example, one result is: **Bioremediation of agro-industrial waste to produce lignocellulosic enzymes and bioethanol with a zero waste**. The abstract is: **Abstract hit: [...reinforces the circular economy model of continual use of resources. Typically, the second...]**. Another result is: **Building design and construction strategies for a circular economy**. The abstract is: **Abstract hit: [...of the circular economy. These wastes are not generated in large amounts, but...]**.

Reaxys search results of documents for ‘circular economy’



Reaxys speaks the language of chemistry. Reaxys is a highly-curated, easy-to-use chemical information solution built on validated data. It harnesses the power of machine learning to help researchers, teachers and students to find, connect and utilize chemistry literature, property and reaction data, patents and experimental procedures.

For more information about Reaxys, visit:
www.elsevier.com/solutions/reaxys

Elsevier offices

ASIA AND AUSTRALIA

Tel: + 65 6349 0222

JAPAN

Tel: + 81 3 5561 5034

KOREA AND TAIWAN

Tel: +82 2 6714 3000

EUROPE, MIDDLE EAST AND AFRICA

Tel: +31 20 485 3767

NORTH AMERICA, CENTRAL AMERICA AND CANADA

Tel: +1 888 615 4500

SOUTH AMERICA

Tel: +55 21 3970 9300

CHINA

Tel: +86 1085 2087 65

For a complete list of Elsevier offices,
please visit elsevier.com/about/locations.



CC BY-NC-SA 4.0 International:
creativecommons.org/licenses/by-nc-sa/4.0/

Reaxys is a service mark of Elsevier B.V.
July 2022

ELSEVIER