Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery

Dial-a-Molecule Grand Challenge Network

Directed Assembly Grand Challenge Network

AI^3 Science Discovery Network+ & Dial-a-Molecule Network & Directed Assembly Network: AI for Reaction Outcome & Synthetic Route Prediction Conference
09-11/03/2020
DeVere Tortworth Court Hotel
Tortworth, Wotton Under Edge, GL12 8HH

Dr. Wendy A. Warr
Wendy Warr & Associates

04/05/2020

AI3SD-Event-Series:Report-17

AI^3 Science Discovery Network+ & Dial-a-Molecule Network & Directed Assembly Network: AI for Reaction Outcome & Synthetic Route Prediction Conference
AI3SD-Event-Series:Report-17
04/05/2020

DOI: 10.5258/SOTON/P0021
Published by University of Southampton

Network: Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery
This Network+ is EPSRC Funded under Grant No: EP/S000356/1
Principal Investigator: *Professor Jeremy Frey*
Co-Investigator: *Professor Mahesan Niranjan*
Network+ Coordinator: *Dr Samantha Kanza*

Network: Dial-a-Molecule
This Network+ is EPSRC Funded under Grant No: EP/P007589/1
Principal Investigator: *Professor Richard Whitby*
Co-Investigator: *Dr Mimi Hii*
Network+ Coordinator: *Dr Gill Smith*

Network: Directed Assembly
This Network+ is EPSRC Funded under Grant No: EP/P007279/1
Principal Investigator: *Professor Harris Makatsoris*
Co-Investigator: *Professor Paul Raithby*
Network+ Coordinator: *Mr Martin Elliott*

Contents

A report by Wendy A. Warr (wendy@warr.com) on a three-day residential meeting organized by the Dial-a-Molecule, Directed Assembly and AI3SD Networks at the De Vere Tortworth on March 9-11, 2020.

1 Event Details

Title	AI for Reaction Outcome & Synthetic Route Prediction Conference
Organisers	AI^3 Science Discovery Network+, Dial-a-Molecule Grand Challenge Network & Directed Assembly Grand Challenge Network
Dates	09-11/03/2020
Programme	Programme
No. Participants	99
Location	Tortworth, Wotton Under Edge, GL12 8HH
Organisation Committee	Dr Gill Smith, Dial-a-Molecule Network, Dr Samantha Kanza, AI^3 Science Discovery Network+ & Mr Martin Elliott, Directed Assembly
Conference Chairs	Professor Richard Whitby, Dial-a-Molecule Network, Professor Jeremy Frey, AI^3 Science Discovery Network+ & Professor Harris Makatsoris, Directed Assembly

2 Welcome and Introduction

Professor Richard Whitby, University of Southampton

The meeting was organized by the Dial-a-Molecule,[1,2] Directed Assembly,[3] and AI^3 Science Discovery[4] Networks. Dial-a-Molecule's vision is that in 20-40 years, scientists will be able to deliver any desired molecule within a timeframe useful to the end-user, using safe, economically viable and sustainable processes. Predicting the outcome of unknown reactions is a key challenge, and a key problem is lack of data, particularly on "failed" reactions. Synthesis must become a data-driven discipline. Since 2011 the Dial-a-Molecule Network has organized many meetings around the themes of collecting better data and using automated reaction platforms for repeatable and captured procedures. A recent success has been the establishment of the center for Rapid Online Analysis of Reactions (ROAR) at Imperial College.[5]

3 Computer-assisted design of complex organic syntheses, 50 years on

Professor Peter Johnson, University of Leeds (talk sponsored by CAS)

The goal of Corey and Wipke's seminal work on computer-aided organic synthesis, the foundation for the Logic and Heuristics Applied to Synthetic Analysis (LHASA) program, was to *assist* chemists to find complete synthetic routes to target molecules.[6] Organic Chemical Synthesis Simulation (OCSS) preceded LHASA. It used a large PDP10 computer[7] that probably had a fraction of the computer power of today's smart cell phones.

In a typical retrosynthetic analysis,[8] level one precursors of a target molecule are found, and then precursors of those precursors are found, and so on, until an available starting material is identified, at which point the search terminates. In the absence of powerful strategies or frequent user pruning, the combinatorial nature of this approach prohibits very broad searches (many alternative reactions) as well as very deep searches (many steps between starting materials and products). For many systems the retrosynthetic analysis is driven by rules ("transforms") describing the scope, limitations, and structure changes associated with a reaction.

The reaction core provides a description of the essential structural features which must be present in the product for the transform to be applied; the extended reaction core contains additional information about individual atoms to make sure it is only applied in the correct situation. This may include the exclusion of structural features. In LHASA transforms were hand coded. Throughput was slow and newer advances in synthetic chemistry can invalidate older rules. Usage of LHASA declined when reaction databases became available.

In the 2000s, Pfizer approached Simbiosys, a Toronto based cheminformatics company, to develop a retrosynthetic tool "ARChem Route Designer" (later named ChemPlanner). The aim was to perform rule- and precedent-based retrosynthetic analysis of target molecules back to readily available materials, and where possible, provide automated generation of retrosynthetic reaction rules. Other requirements were to provide a comprehensive set of alternative routes to a given target; allow user guidance and control; and provide literature examples of the suggested transformations, and direct access to details of starting materials.

In ARChem, rules are extracted from a reaction database, and reaction cores identified from a reaction file with mapping of the atoms and bonds which were changed, made or broken in the reaction. The extracted core is extended to include all structural features *essential* for the reaction to occur, not those that are just "passengers". Literature examples of the reaction type, for example esterification, are clustered together and an esterification rule is generated. Generalized rules might involve a nucleofuge (NF), that is, a leaving group which carries away a bonding electron pair. In the completed reaction rule, the generalized NF group is replaced by the most common group, but it is recognized that many alternative nucleofuges could also work. This grouping together of similar reactions in a single rule is one of many heuristics used to reduce the combinatorial explosion.

Extended core perception requires chemistry judgment and knowledge of reaction mechanism. Examples are classified according to perceived mechanism type. Where different mechanistic types share a common reaction core, mechanistic type is used as the basis for splitting into separate rules and to determine which atoms to include in the extended core. An example is nucleophilic aromatic substitution where the addition-elimination mechanism requires a pi acceptor group in the *ortho* or *para* position, but the mechanism *via* an organometallic intermediate does not require an activating group in any position.

In contrast to LHASA, the reaction rules are *not* defined by chemists, but the rules for abstracting some of the transforms from reaction databases are. This requires a relatively small number of meta rules: many reactions, but fewer basic mechanistic types. Manual curation of rules is also undertaken. The hierarchical view of rules allows trivial variants to be identified and merged into fewer rules. Key reactions are identified and given higher priority (fewer examples are required) and minor variants of common reactions are identified and given lower priority (more examples required). In short, accurate rules require synthetic chemistry knowledge. The principle is to automate wherever possible and curate the rule set manually.

The current ARChem search engine uses an unsupervised heuristic search strategy. Search is exhaustive over a limited number of steps. Rules are invoked according to an "example count" parameter defined by the user. Only rules based on more examples than the value of the parameter are used in the search. Rules are weighted in accordance with importance: higher ratings need fewer examples, as do rules which generate relatively rare moieties. Heuristics to control the combinatorial explosion include amalgamation of similar reactions, synthetic depth limit, example count, user guidance, and search termination if a sufficiently cheap starting material is generated.

Interfering functionality is taken into consideration. Following rule abstraction, compatible functionality is detected by examining the examples. Moieties in the examples outside the extended core are listed as compatible. Functional groups not found in the examples will be identified as "possibly interfering". Possibly interfering functionality is penalized in scoring and highlighted to the user. This technology only works well if reactions differing by mechanism are in separate rules.

Another challenge for retrosynthetic analysis is efficient approaches to navigation of very large answer sets. Users need to see the best solutions first. A reaction scoring function takes into account intrinsic reaction score (some reactions are more highly regarded than others), and reduction of target complexity (the more bonds or rings broken in the retrosynthetic transformation the better). Disconnections which give larger fragments are preferred. Scoring minimizes wastage and maximizes starting material coverage. Thoroughly explored chemistry is given preference (based on example count). The cost of reactants is considered. Parameters for scoring are not embedded in the code but can be changed by an administrator. Users can mark bonds to be preserved or broken, or they can run a search oriented on starting material, by drawing a starting material and target, and mapping the atoms.

Presenting the results is also a challenge: the solutions space could be large, the balance between viewing full routes and individual steps needs consideration, and there is an enormous amount of information in various formats to be displayed. Peter showed some screenshots illustrating the principles of structure-driven viewing, with little text; hybrid step-by-step and full route display; and presentation of the evidence for the suggested transformation (literature precedent). In one example the number one ARChem predicted route was found before the synthesis was actually published.

Further general improvements have been made to the system including regioselectivity in aromatic electrophilic substitution reactions; sorting of examples based on similarity to the suggested transformation; route cost calculation; and links for starting materials. Vendor prices have hyperlinks to the chemical vendor's web pages, and starting materials found which are more expensive than a user set limit are not terminators and are subjected to further retrosynthetic analysis. In similarity sorting, similarity is measured by the presence of as many of the same functional groups as possible, and by structural similarity in the vicinity of the reaction core, using a technology more sophisticated than the shell-based similarity found in many systems.

In 2014, Wiley acquired Simbiosys with a view to integrating ARChem with the ChemInform database of about 2 million reactions. ARChem gained a new user interface and a new name, ChemPlanner. In 2017, CAS acquired the right to develop the system in order to integrate it with their reaction database of over 20 million reactions. In 2019, after major integration and development effort, the SciFinder^n retrosynthesis tool was released. Initially it produced just multistep "experimental" routes (all steps which exactly matched literature transformations). At the end of 2019, "predicted" routes (novel routes with a literature precedent for individual steps) were produced. SciFinder^n. Retrosynthesis is integrated with other features of SciFinder^n. Software engineers at CAS have extensive experience in handling big data, and that, coupled with CAS's very powerful hardware results in major gains in speed of certain operations.

Further enhancements are under development. Continuing improvement in automated rule generation and curation should lead to a huge reduction in the number of rules, and faster and deeper searches. Continuing evolution of the scoring function should bring better solutions to the top. Smarter searches will use only relevant rules and incorporate a variety of synthesis strategies. The user interface will be further improved. Stereochemistry rules to capture the huge advances made in enantioselective chemistry in the past few decades will be added soon; stereochemical match to both CAS reaction sequences and starting materials is already featured.

The stereochemical perception algorithm[9] is based on the analysis of atom coordinates and the presence of wedged and hashed bonds according to prescribed reference patterns. The pattern matching approach means that tetrahedral, olefinic, allenic, biaryl atropisomers and the stereochemistry of various coordination complexes can be located and described by a single, unified approach. The stereochemical requirements are set using a separate constraints graph which is solved simultaneously with the edge constraints of a regular graph matching algorithm. The stereochemical constraints graph can be set to solve for identical, mirror, epimer, or diastereoisomer match. The procedure is also used for stereochemical match of available starting materials.

The current approach which involves drawing stereochemical rules means that the algorithms used to perceive stereochemistry in target molecules can be equally applied to the drawn rules. Every alternative stereochemical consequence of a given reaction has its own rule. Even rules which represent stereochemically impossible reactions are included to catch errors in the reaction databases. In an automated procedure, each rule is matched to the whole original reaction database to extract examples matching the rule. About 30-35% of the most useful methods in the Corey-Kürti list[10] have been covered so far.

ARChem and its successors are heavily based on multiple heuristics derived from the experience of human practitioners of synthetic organic chemistry augmented by database searches. Some of the newer systems rely much less on such heuristics but instead apply machine learning techniques to the problem. It remains to be seen which approach yields the most useful results for users over the next decade.

4 Gathering molecules: representations and machine learning with minimal data

Professor Jonathan Goodman, University of Cambridge

Jonathan began by showing various representations of benzene (Figure 1). The name "benzene" is pronounceable. The 2D structure of benzene is appealing but has orientation issues. There are five CAS Registry Numbers (CAS RNs) for benzene, all proprietary. SMILES[11,12] representations can be constructed by humans, and they are somewhat human-readable, but they are not unique (although various algorithms have been written to make them unique). The IUPAC International Chemical Identifier (InChI)[13] is canonical (i.e., there is a one-to-one correspondence between structure and InChI string). It is constructed by computer and is readable by very dedicated humans. InChIKey is always the same length, and it is better than InChI for searching, but it is undecodable.

CAS RN: 71-43-2

SMILES: C1=CC=CC=C1; c1ccccc1

InChI=1S/C6H6/c1-2-4-6-5-3-1/h1-6H

InChIKey: UHOVQNZJYSORNB-UHFFFAOYSA-N

Figure 1: Benzene.

Another representation is the MDL molfile.[14,15] This is a connection table unambiguously describing the atoms and bonds in the molecule, but it contains much "unnecessary" information such as coordinates, bond localization, and atom numbering. The fragments present in a structure can be represented as a sequence of 0s and 1s, where 0 means that the fragment is not present in the structure and 1 means that it is present in the structure. Each 0 or 1 can be represented as a single bit in a bitstring. These bitstrings are called structure fingerprints. Fingerprints, of which there are many types,[16] are useful for machine learning, but you cannot go from a fingerprint back to a structure, two different molecules can have the same fingerprint, and some important features such as stereochemistry are omitted.

Jonathan showed various representations of remdesivir. In numbering the atoms in the structure for a canonical string such as an InChI it should not matter where you start with the first atom: there should still be only one InChI for one structure. There is only one InChI for remdesivir, but it is easy to make mistakes in InChI generation. If the structure on the left of Figure 2 is "stretched out" to give the structure on the right, a second InChI is generated because the ChemDraw to InChI conversion considers two centers to be undefined, and the second section of the InChI, which represents stereochemistry, differs. Only one of the two InChIs corresponds to remdesivir.

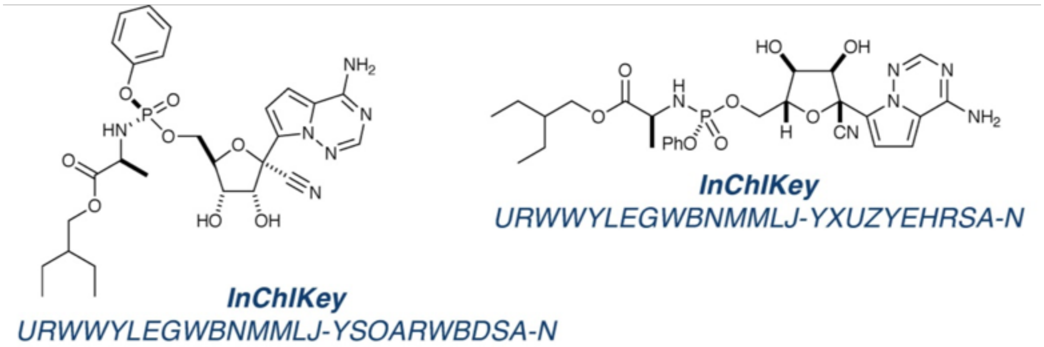
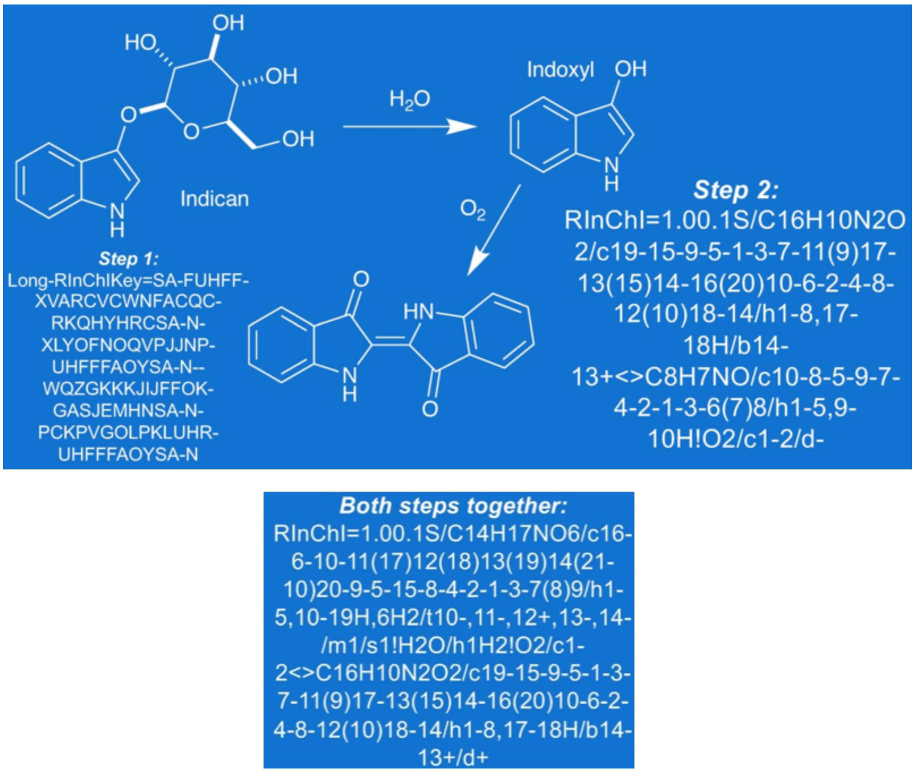


Figure 2: Remdesivir.

The basic molfile for remdesivir has only 2D coordinates. The RDKit default fingerprint is the same for both structure representations because it does not handle stereochemistry. The fingerprint has a great many bits set on but the lack of stereochemistry would be unfortunate if you were using machine learning to predict toxicity. Encoding structures is challenging but reactions are even harder to encode.

We are familiar with random pictures and recognizing images of kittens on the Internet; chemical space is more interesting than a random picture, but not as easy to recognize as a kitten. A molecule can be represented by a chemical name, a 2D or 3D structure, a molfile, an InChI, or a SMILES string, but each representation cannot necessarily be converted into another. A name can be converted to a CAS RN (at a cost), an InChI can be hashed into an InChIKey, and fingerprints can be made from a molfile or a structure, but an InChIKey cannot be converted back into an InChI, and fingerprints cannot be converted back into structures and molfiles. Conformations can be generated from 3D structures.

Reactions are hard to represent. How much detail should be included in a representation, and which is the more important: searchability or maximum information? Graphical representation standards for chemical reactions have been prepared for publication by Division VIII of IUPAC. The Reaction InChI (RInChI)[17] extends the idea of the InChI to reactions. There are also hashed representations (RInChIKeys) suitable for database and web operations. Jonathan used a reaction producing indigo as an example (Figure 3).



Step 1:
Long-RInChIKey=SA-FUHFF-
XVARCVCWNFACQC-
RKQHYHRCSA-N-
XLYOFNOQVPJJNP-
UHFFFAOYSA-N--
WQZGKKKJIJFFOK-
GASJEMHNSA-N-
PCKPVGOLPKLUHR-
UHFFFAOYSA-N

Step 2:
RInChI=1.00.1S/C16H10N2O
2/c19-15-9-5-1-3-7-11(9)17-
13(15)14-16(20)10-6-2-4-8-
12(10)18-14/h1-8,17-
18H/b14-
13+<>C8H7NO/c10-8-5-9-7-
4-2-1-3-6(7)8/h1-5,9-
10H!O2/c1-2/d-

Both steps together:
RInChI=1.00.1S/C14H17NO6/c16-
6-10-11(17)12(18)13(19)14(21-
10)20-9-5-15-8-4-2-1-3-7(8)9/h1-
5,10-19H,6H2/t10-,11-,12+,13-,14-
/m1/s1!H2O/h1H2!O2/c1-
2<>C16H10N2O2/c19-15-9-5-1-3-
7-11(9)17-13(15)14-16(20)10-6-2-
4-8-12(10)18-14/h1-8,17-18H/b14-
13+/d+

Figure 3: RInChI example.

The next step is to persuade everyone to use RInChIs; Jonathan's group has used them. The more information that is available, the easier it will be to use machine learning and AI. AI needs a consistent way of representing a reaction. Through synthesizing both candidate diastereoisomers of a model C1-C28 fragment of the potent cytotoxic marine polyketide hemicalide, Jonathan's team was able to assign the relative configuration between the C1-C15 and C16-C26 regions.[18] By detailed NMR comparisons with the natural product, the relative stereochemistry between these two 1,6-related stereoclusters is elucidated as 13,18-syn rather than the previously proposed 13,18-anti relationship. Hemicalide has 21 stereocenters plus cis/trans stereochemistry. The Morgan fingerprint may not work here because of the number of bonds involved.

NMR shielding tensors may be computed with the Gauge-Independent Atomic Orbital (GIAO) method. Jonathan's team has applied GIAO NMR shift calculation to the challenging task of reliably assigning stereochemistry with quantifiable confidence when only one set of experimental data is available.[19] They have compared several approaches for assigning a probability to each candidate structure and devised a new probability measure, termed DP4. Recently, Jonathan's team has reported an improved version of the procedure,[20] which can be downloaded as a Python script and which analyzes output from open-source molecular modeling programs and commercial packages. The new open-source workflow incorporates a method for the automatic generation of diastereoisomers using InChI strings. Jonathan thinks that we ought to be able to synthesize hemicalide.

More recently, Jonathan's team has also developed a system for automatic processing and assignment of raw 13C and 1H NMR data, DP4-AI, and has integrated it into the computational organic molecular structure elucidation workflow.[21] Starting from a molecular structure with undefined stereochemistry, the system allows for completely automated structure elucidation. It enables high-throughput analyses of databases and large sets of molecules, which were previously impossible, and paves the way for the discovery of new structural information through machine learning. This new functionality has been coupled with an intuitive user interface and is available as open-source software. Jonathan concluded that the application of AI in chemistry is still in the trough of disillusionment in terms of the Gartner hype cycle, but it will climb up the slope of enlightenment if we have more and better data.

5 Introduction to ML and structured matrix methods for learning outliers

Professor Mahesan Niranjan, University of Southampton

Niranjan first briefly discussed the foundations of machine learning. In an analogy with the elephant as seen by blind men, Niranjan said that machine learning[22,23] is seen in different ways by different disciplines: statistics, software, optimization, system identification, etc. He showed equations for function approximator, parameter estimation, prediction, regularization, modeling uncertainty, probabilistic inference and sequential estimation, expressing machine learning as data-driven modeling. Here we are trying to solve a classification or regression problem. Reinforcement learning[24,25] is one of three basic machine learning paradigms, alongside supervised learning and unsupervised learning. The Kalman filter is a widely applied concept in time series analysis. Particle filters are a set of Monte Carlo algorithms used to solve filtering problems arising in signal processing and Bayesian statistical inference.

Next, Niranjan looked at some recent developments. Data are not always in the form$\{x_n y_n\}_{n=1}^{N}$. They also come in the form of natural language text, sentences and documents, biological sequences, macromolecular structures, and small molecules as strings and graphs. In the past we used to extract features from them as a bag of words, fingerprint features, a bag of trigram (of amino acid) frequencies, and features derived from known properties. Now we learn distributed continuous representations in word2vec (word embedding in natural language modeling where discrete words are converted to vectors of real numbers which contain similarity information), dna2vec, prot2vec, and mol2vec. Vector algebra can be applied to natural language with word2vec. We set up self-supervised learning to predict a symbol from the context. A neural network predicts, for example, the probability of GHI in the context of ABC DEF *** JKL MNO, where the inputs and targets are trigrams of amino acids along a protein chain.

The advent of powerful and versatile deep learning frameworks in recent years has made it possible to implement convolution layers into a deep learning model easily. In a 2D convolution, you start with a kernel: a small matrix of weights. This kernel "slides" over the 2D input data, performing an elementwise multiplication with the part of the input it is currently on, and then summing up the results into a single output pixel. A filter is a collection of kernels. Each filter in a convolution layer produces one and only one output channel. Filters act as feature extractors. Depth helps: a series of convolutions with repeated subsampling of, for example, ImageNet,[26] with pooling for data reduction at each stage. Yet, depth makes error propagation difficult because of vanishing gradients in gradient descent.

Gradient descent methods are first-order, iterative, optimization methods. Each iteration updates an approximate solution to the optimization problem by taking a step in the direction of the negative of the gradient of the objective function. By choosing the step-size appropriately, such a method can be made to converge to a local minimum of the objective function. Gradient descent is used in machine learning by defining a loss function that reflects the error of the learner on the training set, and then minimizing that function.

One current community-wide trend is to build deeper and deeper networks. During training, these networks fall foul of an issue known as the vanishing gradient problem. The vanishing gradient problem manifests itself in these networks because the gradient-based weight updates derived through the chain rule for differentiation are the products of n small numbers, where n is the number of layers being backward-propagated through.

A residual neural network (ResNet) is an artificial neural network that uses skip connections, or shortcuts to jump over some layers. One motivation for skipping over layers is to avoid the problem of vanishing gradients, by reusing activations from a previous layer until the adjacent layer learns its weights. WaveNet[27] (from Google DeepMind) is an architecture for signal processing that does stacked dilated convolutions. A regression problem is thus changed into a classification problem which is much easier to solve. Niranjan noted that among the inference problems of classification, regression and estimation, classification is the easiest because all we need to define is a boundary whereas in regression problems, a real valued target has to be predicted. Density estimation is the hardest of the three because the so called curse of dimensionality stipulates that the amount of data required grows exponentially with dimensions to maintain the same accuracy of representing a density.

Another new (or rediscovered) topic Niranjan reviewed is capturing temporal information. The main difference between a convolutional neural network (CNN) and a recurrent neural network (RNN) is the ability of an RNN to process temporal information or data that comes in sequences, such as a sentence. Long short-term memory (LSTM) is an RNN architecture used in the field of deep learning. Unlike standard feedforward neural networks, LSTM has feedback connections. It can not only process single data points (such as images), but also entire sequences of data (such as speech or video).

Regularization is the process of adding information in order to solve an ill-posed problem or to prevent overfitting. Weight decay is perhaps the most widely-used technique for regularizing parametric machine learning models. Weight decay is defined as multiplying each weight in the gradient descent at each epoch by a factor, λ, smaller than one and greater than zero.

Multitask learning is a subfield of machine learning in which multiple learning tasks are solved at the same time, while exploiting commonalities and differences across tasks. This can result in improved learning efficiency and prediction accuracy for the task-specific models, when compared to training the models separately. Multitask learning works because regularization induced by requiring an algorithm to perform well on a related task can be superior to regularization that prevents overfitting by penalizing all complexity uniformly.

Early stopping is a form of regularization used to avoid overfitting when training a learner with an iterative method, such as gradient descent. Such methods update the learner so as to make it better fit the training data with each iteration. Up to a point, this improves the learner's performance on data outside of the training set. Past that point, however, improving the learner's fit to the training data comes at the expense of increased generalization error. Early stopping rules provide guidance as to how many iterations can be run before the learner begins to overfit.

Bootstrap aggregating, also called bagging (from **b**ootstrap **agg**regat**ing**), is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms. It also reduces variance and helps to avoid overfitting. Bagging is a special case of the model averaging approach. It is a method for generating multiple versions of a predictor and using these to get an aggregated predictor. The aggregation averages over the versions when predicting a numerical outcome and does a plurality vote when predicting a class. The multiple versions are formed by making bootstrap replicates of the learning set and using these as new learning sets.[28]

Not only is overfitting a serious problem in deep neural networks; large networks are also slow to use, making it difficult to deal with overfitting by combining the predictions of many different large neural nets at test time. Dropout (a regularization technique patented by Google) is a technique for addressing this problem. The key idea is to randomly drop units (along with their connections) from the neural network during training. This prevents units from co-adapting too much.

A topic particularly close to Niranjan's heart is cascade learning. His team have reported an approach, called deep cascade learning, for efficient training of deep neural networks in a bottom-up fashion using a layered structure.[29] Such training of deep networks in a cascade directly circumvents the vanishing gradient problem by ensuring that the output is always adjacent to the layer being trained. Niranjan's team found that that better, domain-specific, representations are learned in early layers when compared to what is learned in end-to-end training. They have also shown that such cascade training has significant computational and memory advantages over end-to-end training, and can be used as a pretraining algorithm to obtain a better performance.

Belilovsky *et al.* have produced an alternative to end-to-end training of CNNs that can scale to ImageNet.[30] They used 1-hidden layer learning problems to build deep networks sequentially, layer by layer, which can inherit properties from shallow networks. Extending this training methodology to construct individual layers by solving 2-and-3-hidden layer auxiliary problems, they obtained an 11-layer network that exceeded the performance of other methods on ImageNet.

Cascade learning is particularly suited to transfer learning, as learning is achieved in a layerwise fashion, enabling the transfer of selected layers to optimize the quality of transferred features. In the domain of human activity recognition, where the consideration of resource consumption is critical, cascade learning is of particular interest as it has demonstrated the ability to achieve significant reductions in computational and memory costs with negligible performance loss. Activity is learned from one setting, and the model is retrained in a different setting. Early layers learn coarse features; later layers specialize in the source problem. In a recent publication,[31] Niranjan's team has evaluated the use of cascade learning and compared it to end-to-end learning in various transfer learning experiments, all applied to human activity recognition. They found that cascade learning achieves state of the art performance for transfer learning in comparison to previously published work, improving F1 scores by over 15%. Cascade learning performs similarly to end-to-end learning considering F1 scores, with the additional advantage of requiring fewer parameters.

Another significant development Niranjan reviewed is the autoencoder. An autoencoder is a type of artificial neural network used to learn efficient data codings in an unsupervised manner. The aim of an autoencoder is to learn a representation for a set of data, typically for dimensionality reduction, by training the network to ignore signal noise. Along with the reduction side, a reconstructing side is learnt, where the autoencoder, trained using back-propagation tries to generate from the reduced encoding a representation as close as possible to its original input. A variational autoencoder (VAE) uses a probabilistic model in the latent space between encoder and decoder. Doersch has published a useful tutorial on the theory of VAEs.[32]

Another "trick" is knowledge distillation. Categorical data are variables that contain label values rather than numeric values. Many machine learning algorithms cannot operate on label data directly. They require all input variables and output variables to be numeric. This means that categorical data must be converted to a numerical form. One way of doing this involves one-hot encoding. One-hot is a group of bits among which the legal combinations of values are only those with a single high (1) bit and all the others low (0). A big model can learn, but a small model cannot, so a big model is trained, and outputs posterior probabilities, and then the small model is trained on the outputs of the big model.

Niranjan's final hot topic was attention. Recurrent neural networks, long short-term memory, and gated recurrent neural networks in particular, have been firmly established as state of the art approaches in sequence modeling and transduction problems such as language modeling and machine translation. The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. A new simple network architecture, the Transformer, based solely on attention mechanisms, dispenses with recurrence and convolutions entirely.[33]

Finally, Niranjan summarized his interest in outliers in a regression setting and outliers in a matrix approximation setting, in particular in analyzing biological data. Despite much dynamical cellular behavior being achieved by accurate regulation of protein concentrations, mRNA abundances, measured by microarray technology, and more recently by deep sequencing techniques, are widely used as proxies for protein measurements. Although for some species and under some conditions, there is good correlation between transcriptome and proteome level measurements, such correlation is by no means universal due to post-transcriptional and post-translational regulation, both of which are highly prevalent in cells.

Niranjan's team has developed a data-driven machine learning approach to bridging the gap between these two levels of high-throughput 'omic measurements, and has deployed the model in a novel way to uncover mRNA-protein pairs that are candidates for post-translational regulation.[34] The application of feature selection by sparsity inducing regression (l_1 norm regularization) leads to a stable set of features (i.e., mRNA, ribosomal occupancy, ribosome density, tRNA adaptation index, and codon bias) while achieving a feature reduction from 37 to 5. A linear predictor used with these features is capable of predicting protein concentrations fairly accurately. Proteins whose concentration cannot be predicted accurately, taken as outliers with respect to the predictor, were shown to have annotation evidence of post-translational modification, significantly more than random subsets of similar size. In a data mining sense, this work also shows a wider point that outliers with respect to a learning method can carry meaningful information about a problem domain.

6 Applying AI to retrosynthesis in the wilderness

Mikołaj Sacha, Molecule.one

Molecule.one first formulated their hypothesis in October 2018. They asked themselves how current AI could bring value to retrosynthesis in drug discovery pipelines. Mikołaj explained their motivation. Ilya Sutskever (co-inventor of the CNN AlexNet, and of AlphaGo and TensorFlow) said that if you use a lot of good and labeled training data and a big deep neural network, success is the only possible outcome. ImageNet[26] classification error (top 5) has gradually improved since 2011; ResNet exceeded human performance by 2015, and the CNN GoogLeNet version 4 was even better in 2016. In 2018, Geirhos *et al.* demonstrated that the CNN ResNet-50, having learned a texture-based representation on ImageNet, is able to learn a shape-based representation instead when trained on a stylized version of ImageNet.[35] This provides a much better fit for human behavioral performance and comes with a number of other unexpected benefits.

Deep neural networks can fail to generalize to out-of-distribution inputs.[36] Thus, a school bus is recognized as a garbage truck when viewed from underneath, as a punching bag when viewed toward its roof, and as a snowplow when on its side across a snowy road. Maybe this lack of deep understanding also applies to retrosynthesis,[37] but shallow understanding may be enough. Markovnikov's rule states that when an acid such as hydrogen bromide is added to an asymmetric alkene, the acidic hydrogen attaches itself to the carbon having a greater number of hydrogen substituents whereas the halide group attaches itself to the carbon atom which has a greater number of alkyl substituents. The rule was proposed in 1870, and carbocations were proposed as reaction intermediates in 1900, but it was not until about 1960 that the structure of carbocations was observed using NMR. Shallow understanding may be enough.

The Molecule.one team hypothesized that while the current incarnation of AI methods gains shallow understanding of chemistry and works well on some classes of reactions (e.g., popular reactions), strong, deep neural models that "know when they don't know" can provide great value to the industry. As an aside, the team asked themselves why reactions should not be encoded using rules as in Synthia.[38] One reason is that the rules get very complex.

Molecule.one began validation of their hypothesis in March 2019, examining whether they could expect their models to work well on some subsets of chemistry. To train and evaluate the models, they used 400,000 reactions from the set scraped by Lowe[39] from publicly available US patents as "true" reactions. They found about 1600 commonly occurring reaction templates in the dataset. They generated negative samples for each reaction by applying its template to all other existing matching places in substrates. The models were then evaluated by how well they discriminate between the two classes of reactions using the receiver operating characteristic (ROC) area under curve (AUC) metric. The researchers split the data into training and test sets by clustering reactions using chemical fingerprints of their substrates.

They tested a few neural network architectures: a model working on a reaction fingerprint, a Convolutional Neural Network and a few Graph Convolutional Networks (CGNs). In particular, they tested the Edge Attention Graph Neural Network (EAGCN),[40] which they adapted to chemical reactions. They also enhanced it with a multiheaded self-attention mechanism. The best model, EAGCN with two attention heads, reached an AUC score of 0.99 compared to 0.95 for the second-best model, EAGCN with a single attention head.

To understand better how well their models generalize, they compared them to a heuristic developed by a chemist specifically for a popular reaction type (aromatic nitration). The best model achieved a comparable performance to the expert heuristic (F1 score of 0.81, compared to 0.82 achieved by the heuristic), despite the fact that the team trained it on a whole dataset of reactions (including 10,000 nitration examples), while the heuristic was crafted specifically for this reaction type.[41] The expert heuristic was both very time-consuming to construct and limited in application to a narrow portion of the dataset. Learned models do not have these limitations.

The team had expected AI models to understand *some* chemistry but next they had to decide how to build an AI system that is useful and reliable. The difference in the Molecule.one approach is a state-of-the-art deep neural discriminator. A discriminator takes its input from a neural- or template-based generator. In the Molecule.one case, the discriminator (in-scope filter) distinguishes generator outputs from real reactions. It uses proper negative reactions, which means that high predictions correspond to high confidence, and *vice versa*. It helps users "know when they don't know".

For the graph attention network (GAT) in-scope filter, each reaction is represented as a graph (an imaginary transition structure),[42] where atoms are nodes, and bonds are represented as featurized adjacency matrices. The feature of a bond indicates its existence and type in substrates and in the product. The team has shown that choosing a proper reaction representation plays a crucial role in the task of predicting reaction feasibility. They discovered that the graph representation achieves by far the best results (Figure 4).

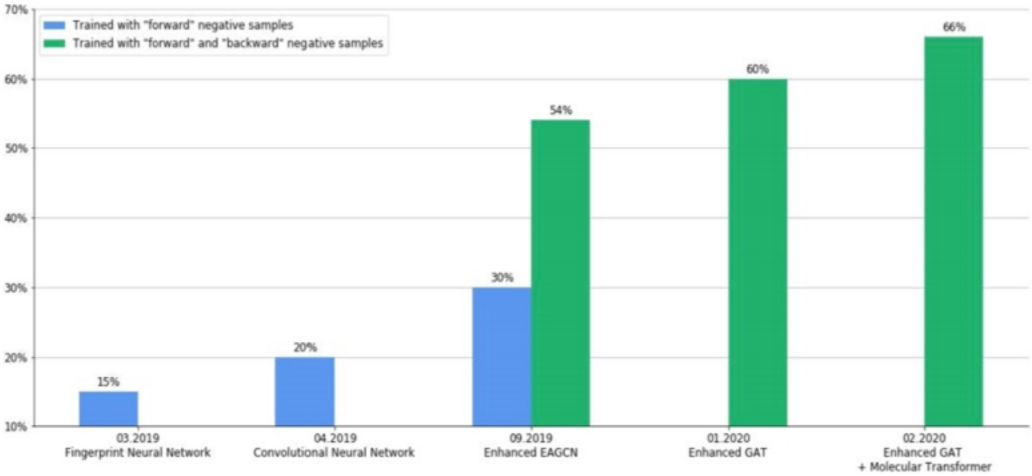


Figure 4: Accuracy in top 10 predictions for retrosynthesis.

A neural generator alone is not used because neural generators can make trivial mistakes. Li *et al.* found direct contradictions[43] in completing the last sentence in "I love baseball. It's awesome. I really dislike..." Some completions were "basketball", "it", "the", and "sports". Neural models for reaction generation, such as the Molecular Transformer,[44] can also make trivial mistakes. For instance, the Molecular Transformer often repeats commonly encountered patterns, such as carbon chains, or simply copies reaction substrates as proposed products.

In October 2019, Molecule.one launched its eponymous product and in 2020 the company is now working with its first industrial partners. One class of techniques of growing interest for early-stage drug discovery is *de novo* molecular generation and optimization. These techniques can suggest novel molecular structures, but ignorance of synthesizability is a problem.[45] Molecule.one can address this challenge. The RDKit[46,47] synthesizability score barely distinguishes targets that have a known synthesis path; Molecule.one's score gives high confidence to targets that have a known synthesis path.

Finally, Mikołaj summarized some ongoing work. Molecule Attention Transformer[48] (MAT). a model that Molecule.one collaborated on, augments the attention mechanism in Transformer[33] using interatomic distances and the molecular graph structure. Experiments show that MAT performs competitively on a diverse set of molecular prediction tasks. Most importantly, with a simple self-supervised pretraining, MAT requires tuning of only a few hyperparameter values to achieve state-of-the-art performance on downstream tasks. Attention weights learned by MAT are interpretable from the chemical point of view.

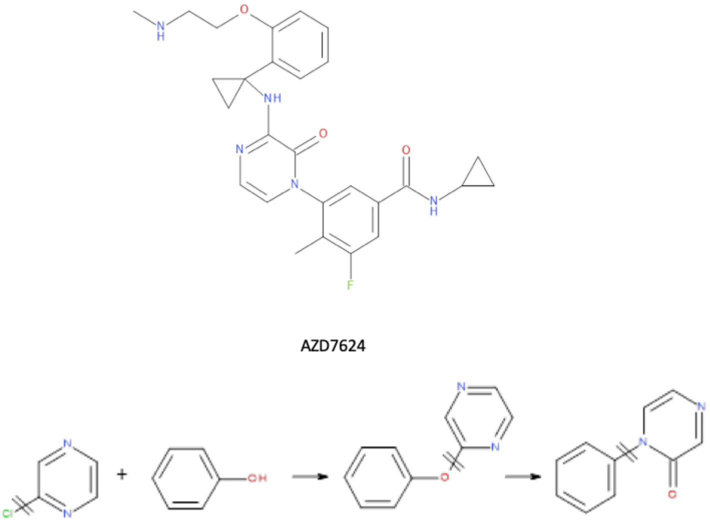
Neural models are often black boxes that cannot assess confidence of their steps. The Molecule.one team wants to make a model that explains the decisions that it makes. The Molecule Edit Graph Attention Network (MEGAN) encodes a reaction as a sequence of edits in the molecule graph. Its neural generator explicitly outputs molecule edits. Using the data preprocessed by Liu *et al.*[49] it outperforms Graph Logic Network,[50] NeuralSym,[51] Retrosim,[52] and Molecular Transformer.[44] We should not expect AI to fully understand chemistry, but we should expect AI to understand *some* chemistry. Molecule.one's approach uses state-of-the-art models trained to "know when they don't know". This approach enables Molecule.one to deliver real value to their partners.

7 Reproducibility in chemistry

Dr. Mark Warne, DeepMatter

Human interaction to produce chemical products introduces many opportunities for error. It is estimated that there are more than 500,000 chemists worldwide and 50% of their science is not reproducible.[53] This has a financial impact of $28 billion. Irreproducibility also has a reputational impact: the chemistry Nobel laureate Francis Arnold recently had to retract a paper in *Science* because the work was not reproducible. Machine learning and AI technologies are held out as a panacea for improving reaction outcomes, but can we rely on the underlying data if reproducibility is such a big problem? DeepMatter is addressing this issue with DigitalGlassware,[54] a software- and hardware-enabled repository of contextualized, primary, full time-course data for all reactions, whether successful or unsuccessful. In his talk, Mark concentrated on a pragmatic approach to improving synthesis prediction integrating IC*SYNTH*[55–57] and DigitalGlassware.[54] Primary laboratory data from DigitalGlassware can be used to enrich the historical data that were fed into the transform libraries in IC*SYNTH*.

Mark gave an example concerning a P38 inhibitor for prevention of exacerbations in chronic obstructive pulmonary disease, AZD7624 (see Figure 5). IC*SYNTH* retrosynthesis searching suggested a wide range of routes; one of which is based on a two-step transformation from commercially available starting materials involving a novel Chapman rearrangement. All precedents for the Chapman rearrangement sourced from InfoChem's SPRESI[58] database are fairly wide extrapolations but a wider literature search has revealed a heterocyclic Chapman rearrangement, providing an analogy much closer to the model N-phenylpyrazinone target. Step 1 was successful in producing an intermediate with a yield equivalent to that expected. A failed reaction was fed back into IC*SYNTH* for further optimization. Step 2 is still under investigation.



AZD7624

Figure 5: Retrosynthesis of AZD7624.

By using DigitalGlassware the researchers were able to study exactly what they had done in each run of step 1, in terms of tabulated scale, time, temperature, and yield figures, and visualizations of temperature and analytical parameters over time. The example illustrated how feeding your data back into DeepMatter's systems can improve your reaction outcomes.

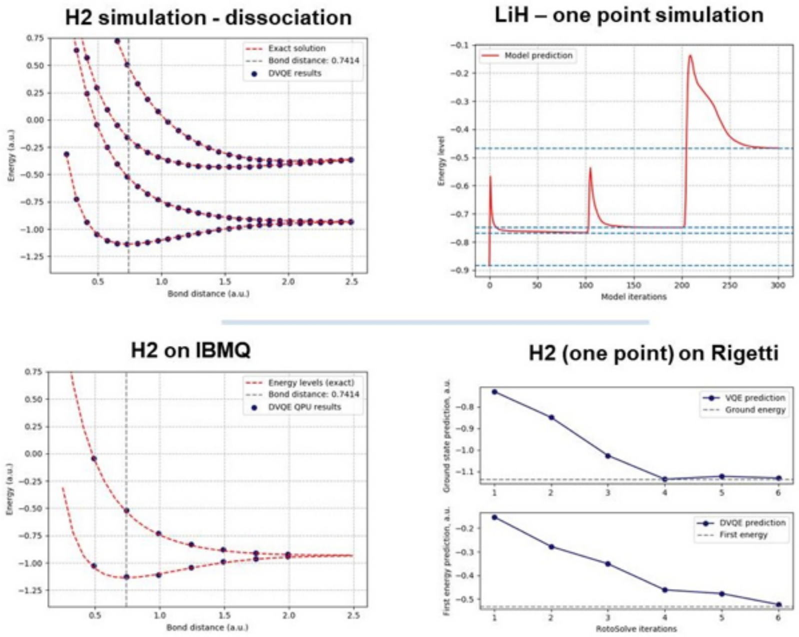
8 Accurate excited states calculations on near term quantum computers

Jules Tilly, Rahko

Quantum computing makes use of quantum-mechanical phenomena to perform computation. These computers use quantum logic gates, with quantum bits or qubits. They can model an exponentially growing complex system with a linear number of qubits by using quantum object properties and quantum information theory to reduce computational complexity. Quantum computers offer the potential to bring parallelism to calculations on a scale that cannot be matched by classical computers. They have the potential to improve modeling of chemicals within the next 5-10 years. In particular, calculating excited states is a nontrivial task with current computational chemistry methods, limiting the complexity of the molecules that can be studied or the accuracy of the results that can be obtained. Jules presented a method that could address this issue using quantum computing, and outlined that there are still many technical challenges to overcome.

He gave a brief introduction to quantum circuits,[59] starting with qubit gates and Pauli operators. When considering modeling of a molecular spectrum on a quantum computer, one should begin by encoding the molecular Hamiltonian in an appropriate format. The Jordan–Wigner transformation maps the fermionic creation and annihilation operators of the second quantized Hamiltonian onto spin operators which can be directly observed on a quantum computer. The ground state is computed using the variational quantum eigensolver (VQE).[60]

Jules and his co-workers have proposed a variational quantum machine learning based method to determine molecular excited states, aimed to be as resilient as possible to the defects of early noisy, intermediate-scale, quantum (NISQ) computers. They have demonstrated an implementation for H_2 on IBM's prototype quantum processor (IBMQ) and Rigetti's quantum processor. The method, named the Discriminative VQE, is inspired from generative machine learning and uses a combination of two parameterized quantum circuits, playing a guessing game to find iteratively the eigenstates of a molecular Hamiltonian.[61] Jules outlined his algorithm using a VQE to constrain the generator in his generator-discriminator quantum machine learning game. He presented results for H_2 on Rigetti and IBMQ computers, and also simulations for LiH (Figure 6).


H2 simulation - dissociation

LiH – one point simulation

H2 on IBMQ

H2 (one point) on Rigetti

Figure 6: Simulations, H_2 and LiH.

There are limitations remaining that need to be addressed. The largest QPUs use only about 50 qubits; error rates are too high for deep circuits; and QPUs remain very slow. As far as software is concerned, limitations include exploding terms in the qubit space Hamiltonian; variance of output due to the measurements limit; and depth of ansatz to model the molecule. Several other methods have been suggested for computation of excited states on QPUs but questions remain regarding their applicability to near term hardware, and scalability to larger systems. Other methods of interest include quantum equation of motion,[62] quantum subspace expansion,[63] imaginary time evolution,[64] overlap minimization,[65] and VQE with symmetry constraints.[66]

9 Making sense of predicted routes: the use of data as evidence for predictions in SciFinder

Paul Peters, CAS

Retrosynthesis in SciFinder[67] is based on ChemPlanner, the technology of which is described in Peter Johnson's earlier talk. In Phase 1 ("experimental") of the application (June 2019) the target molecule had to be in the literature for the planning to work. Phase 2 (December 2019) introduced "prediction" in which ChemPlanner also suggests reactions with no literature precedent, even if the target is novel.

According to user surveys, the main value of the application is that it leads to strategies and reaction steps users would not have otherwise considered, it exposes chemistry they were not familiar with, and it allows them to be more innovative. It supports tasks such as identifying alternatives to known routes, brain storming, finding synthetic routes for novel molecules, and validating developed routes.

The main requirements of an idea generator are quality of results, diversity of solutions, thorough coverage of known chemistry, tight linkage between predictions and experimental evidence, flexibility in guiding the search, options to filter and sort solutions, and ease of use. The quality of the results depends on rule generation, and advanced chemical perception, which facilitates generalizations.

The retrosynthesis plan provides links to starting materials which are commercially available; the target and precursors are given a letter (A, B, C, D...) and the synthetic availability is shown by an icon under the structure. Each subroute is designated by an Erlenmeyer flask, red in experimental planning and green in predictive planning. The user can select a subroute and work on it. In experimental planning, evidence is supplied in the form of reactions in the CAS database. Many filters can be applied. Nonparticipating functional groups is a recently added option. In predictive planning, evidence is in the form of rules.

Once a target has been entered, two sets of options are offered to the user. The first is synthetic depth (1-4), and the type of rules used; the second (optional) is selection of a bond that must be broken, or must remain unchanged. The rules used may be common, uncommon (including common), or rare (including common and uncommon).

Chemists are excited about the technology. Organic chemists are a critical but supportive audience and are very forthcoming in providing feedback and guidance. They appreciate when limitations of the technology (e.g., stereoselectivity and regioselectivity, chemical interference, and identifying the "best" solutions) are openly discussed. They are interested in more options to influence the search, better navigation and pruning options, and integration of proprietary data.

Retrosynthesis in SciFinder is comprehensive, rich and accurate; it uses the most current data; it is versatile and easy to use; it is a facet of a complete discovery experience; and it is continuously updated and improved.

10 What is the importance of false reactions for efficient data-driven retrosynthetic analysis?

Dr. Quentin Perron, Iktos

Iktos were inspired by the synthesis planning methodology reported by Marwin Segler and co-workers[68] (see Marwin's talk later in this report). Four steps are involved: identification of bond disconnections, application of the rules, reaction prediction, and Monte Carlo tree search (MCTS). Iktos concentrate on the third step: finding out if a proposed reaction will work. Since there is no database of failed reactions, a dataset of reactions which do not work has to be created. A "false reaction" has the same reagents as a "true" one but a different product (thanks to application of different templates). Multiple identical templates, regioselectivity, and incompatible functions can all cause a reaction to fail.

An in-scope filter can score the templates to help decide which of them should be applied. Unfortunately, the in-scope filter is not very efficient: it is built on very noisy data and false reactions can actually be true reactions. A probability threshold has to be selected to implement the filtration, and sometimes interesting disconnections can be discarded. Forward prediction[44] also cannot solve these problems.

Prediction is hard because of a lack of high quality data. Stoichiometry, reaction conditions, by-products, chirality, and other facts are not documented well enough. Failed reactions are not published. There are many "one reactant" reactions. Important parts of chemistry knowledge (e.g., functional group reactivity and incompatibility) are in books, not in reaction databases.

Building on the work of Waller's team,[68] Iktos scientists have built their own software, Spaya,[69] based on the Pistachio dataset[70] from NextMove Software, and Mcule compound sourcing[71] for commercial compounds. They have developed an efficient, proprietary, in-scope filter focusing on only very unlikely disconnections. Spaya is a free application online. Users enter a SMILES for a target and can select nonpreferred disconnections. Spaya displays a ranked list of synthetic routes: a user can pick the most relevant ones and analyze them step by step. Each retrosynthesis in Spaya ends with commercially available building blocks. The user can save the retrosynthesis results or place an order for the starting materials.

AI technology is highly valuable and powerful, but still far from solving all the challenges. At the end of the day what matters is the ranking. Iktos is counting on the chemistry community to help improve this ranking.

11 Combining artificial intelligence with structured high quality data in chemistry: delivering outstanding predictive chemistry applications

Dr. Abhinav Kumar, Elsevier

The drug design cycle leads to optimized candidates but necessitates both in-house and published knowledge and data. At each step of the cycle, predictive applications built on published and in-house data with machine-learning algorithms play a more and more important role. One such predictive application is the Reaxys-PAI Predictive Retrosynthesis tool,[72] developed in collaboration with Mark Waller and Marwin Segler,[68] which combines

Reaxys content with AI and machine learning technologies. This tool is extending syntheses of small organic molecules into predictive modeling of previously unpublished synthetic pathways and synthetic feasibility of virtual compounds. It is not reliant on manual encoding of rules. Users can integrate their own customized building blocks or proprietary reaction data. The system solves retrosyntheses for more than twice as many molecules, 30 times faster than traditional computer-aided methods.

The tool has been tested rigorously by the world's leading pharmaceutical and chemical companies and has been demonstrated to provide scientifically robust, diverse and innovative synthetic route suggestions. It is a valuable tool which is easy and intuitive to use and supports the needs of the business and researchers by being a very good assistant and idea generator. The predictive retrosynthesis solution has been trained on both positive and negative reactions data and solves synthesis design questions for novel molecules with direct links to experimental reactions available in Reaxys.

The development of this tool further underscores the importance of having good quality data and sufficient data to build predictive applications. As Kevin Poskitt from SAP[73] said, "Artificial intelligence without data intelligence is artificial". Researchers need high-quality published data; high-quality experimental in-house data (ideally with failed reactions); methods and tools to ingest, normalize, clean, merge and filter published data (e.g., Reaxys) and in-house data; and a scalable platform to process the data. Elsevier's answer to the challenge is the Entellect Reaction Workbench Platform.

12 Intelligence from data: towards prediction in organometallic catalysis

Dr. Natalie Fey, University of Bristol

The use of data for homogeneous catalysis opens up opportunities to select the "best" synthetic route, by making better use of resources; breaking "cultural attachments" to routes; finding new approaches; and knowing when to stop. Chemists can broaden their toolkit and explore new chemistry; deepen their understanding; and combine experiment and computation to work toward prediction. Using computers, chemists can optimize geometries, calculate energies, generate spectra, build and test models, test mechanistic ideas, analyze data, store results, repeat and automate processes, and inform synthesis. What they cannot do is consider everything that happens in a reaction vessel, prove a mechanism definitively, extrapolate, predict reliably (yet), truly design compounds, and replace synthesis.

Natalie has reviewed[74] how descriptors calculated from molecular structures have been used to map different areas of chemical space. She focused on organometallic catalysis, but also touched on other areas where similar approaches have been used, with a view to assessing the extent to which chemical space has been explored.

While truly rational design of new catalysts remains out of reach, detailed mechanistic information from both experiment and computation can be combined successfully with suitable parameters[75] characterizing catalysts and substrates to predict outcomes and guide screening. The computational inputs to this process rely on large databases of parameters characterizing ligand and complex properties in a range of different environments.[76–78] Such maps of catalyst space can be combined with experimental or calculated response data,[76] as well as large-scale data analysis.

Ligands stored in ligand knowledge bases (LKBs) can be used to tune complex properties. The ligands are characterized by size and electronics,[75,79] and their properties are mapped, in order to derive structure-property relationships. It has to be asked whether two dimensions are sufficient[80] and whether the knowledge derived from one ligand type is transferable to other ligand types. In her talk Natalie concentrated mainly on P-donor ligands. The use of individual descriptors of P-donor ligands received a considerable boost with the publication of Tolman's seminal review.[81] Further work was carried out by Fernandez,[80] and by Cundari, Rothenberg, Sigman, Paton, Jensen and others, summarized in a recent review.[75]

LKB-PP[77,82,83] was developed for P,P and P,N donor ligands. It uses 28 descriptors which are quite highly correlated. Projection methods reduce the number of variables (dimensions) by transformation of the original variables, and optimally represent distances between objects. Principal component analysis (PCA) describes the variation of the data in terms of uncorrelated variables. In the initial work[77,82] the P,P space was not sampled well.

Later, Jesús Jover, a postdoctoral research associate working with Natalie at the time, carried out a computational exploration of the effect of systematic variation of backbones and substituents on the properties of bidentate, cis-chelating P,P donor ligands.[84] The parameters used were the same as reported for LKB-PP but calculation protocols were streamlined. Analysis of the resulting LKB-PP$_{screen}$ database with PCA captured the effects of changing backbones and substituents on ligand properties and illustrated how these are complementary variables for these ligands. While backbone variation is routinely employed in ligand synthesis to modify catalyst properties, only a limited subset of substituents is commonly accessed. Jesús and Natalie highlighted substituents which are likely to generate new ligand properties, of interest for the design and improved sampling of bidentate ligands.

The ligand knowledge base for monodentate P-donor ligands (LKB-P) can be linked with the likely mechanism of reaction, in terms of the relative importance of steric and electronic effects.[74,76,85] Claire McMullin, a former Ph.D. student at Bristol, addressed this challenge using data published by Stauffer and Hartwig.[86] It was found that good amination catalysts may favor monoligated palladium. Bristol's LKB team, involving Natalie, Guy Orpen, Guy Lloyd-Jones, Jeremy Harvey and others, have found that there is a hotspot of large and electron-rich ligands (Figure 7, unpublished work). Dr. Ben Swallow, a former postdoctoral research assistant at Bristol, has tried more sophisticated models and different sampling sizes, but it proved easiest to predict mediocre catalysts.

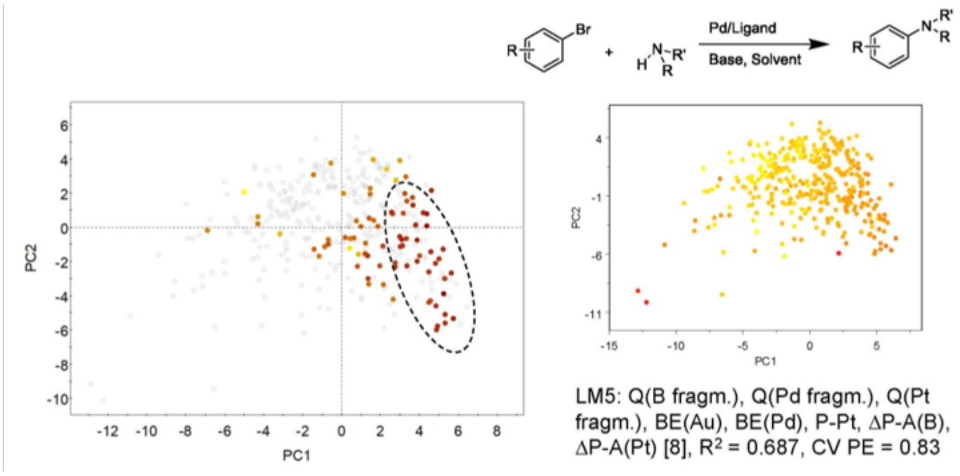


Figure 7: LKB-P. Hotspot of large and electron-rich ligands.

In summary, Natalie's descriptors show chemically intuitive relationships and can map chemical space computationally. Donors and backbones contribute to bidentate ligand properties. Good amination catalysts may favor monoligated palladium and large and electron-rich ligands, but it is easiest to predict mediocre catalysts. Thus, the descriptors can quantify similarity and can be used in experimental design. The variables can be combined for better tuning. Alternative catalysts can be suggested, but better data and suitable statistical models are needed if the catalysts are to be better than mediocre. So, Natalie's team continues to work on large-scale computational prediction of reactivity.

The Bristol Reactivity in Catalysis Knowledgebase (BRiCK) aims to discover new applications for known catalysts, by a systematic study of catalytic cycles. Phase 1 will be the study of substrate effects on barriers, phase 2 modification of the metal and ligands in the catalyst, and phase 3 experimental testing. After repetition of these activities, it is hoped that, eventually, reactivity descriptors will be identified.

Natalie outlined some work being carried out by Derek Durand (Figure 8).[87] Catalytic cycles are inherently multivariate. Mechanistic manifolds can be mapped out: Natalie showed the energy barriers for oxidative addition (two barriers), migratory insertion (MI) and reductive elimination for U = ethene and B3PW91-D3/6-31G(d,p)/LACV3P* / PCM (solvent = methanol).

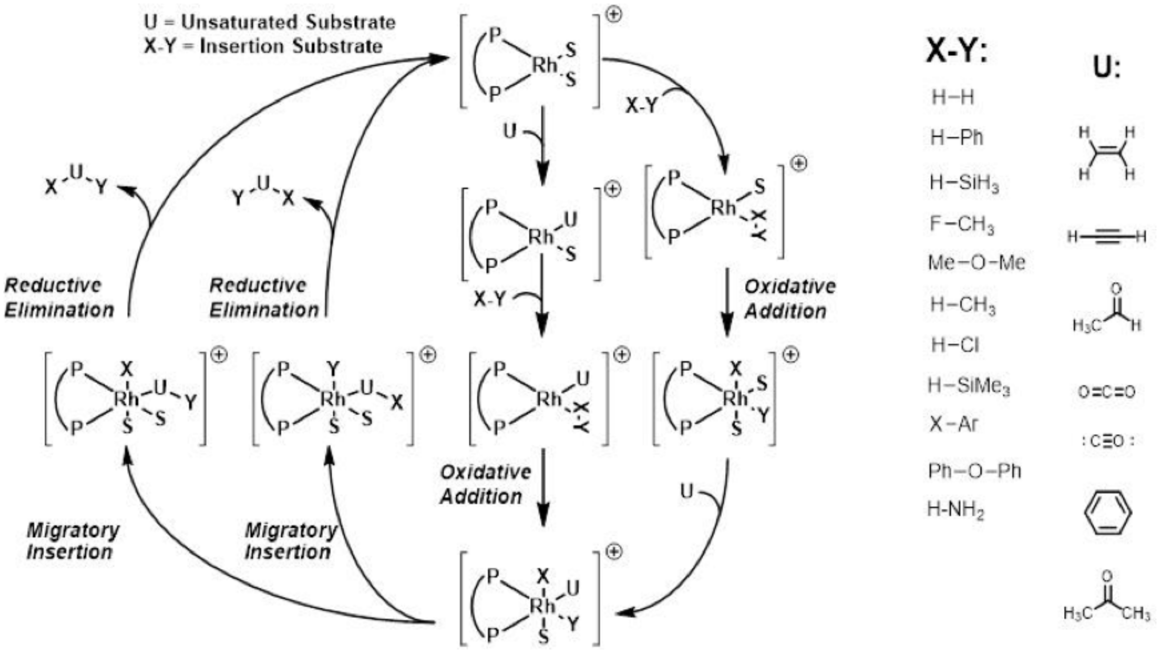


Figure 8: Catalytic cycle.

She then showed bar charts for ΔΔG data for each key step (oxidative addition, migratory insertion, reductive elimination) in the case of U = ethane and X-Y = H-H, H-SiH_3, H-CH_3, H-C_6H_5, F-CH_3, H-NH_2, and Me-OMe. These showed that oxidative addition can be a killer for electron-rich substrates (F-CH_3, H-NH_2, and Me-OMe); migratory insertion is not the biggest problem; and the elimination step may be a concern in some cases (e.g., Me-OMe). H_2 and H-SiR_3 look promising. The screen can be expanded to other options for U.

These virtual screening experiments show that reactivity descriptors based on such fundamental steps should be feasible. They can be mapped computationally ("computers do not get bored"). They are the foundation on which to explore substrate and catalyst modification. Since migratory insertion proceeds with reasonable barriers, it is not a problem, but it can be monitored. The elimination step may be a concern for some substrates; this problem must be balanced with optimizing oxidative addition. Experimental ("wet chemistry") work on H-SiR_3 compounds is starting soon. Varying the other substrate (U) presents opportunities for exploring relationships.

Rather than pursuing a purely computational solution of *in silico* catalyst design and evaluation, an iterative process of mechanistic study, data analysis, prediction and experimentation can accommodate complicated mechanistic manifolds and lead to useful predictions for the discovery and design of suitable catalysts. Natalie concluded by emphasizing the importance of data, both for use in maps, models and screening (design and optimization), and for use in mapping reactivity and moving toward prediction.

13 Chemistry ontologies and artificial intelligence

Dr. Colin Batchelor, Royal Society of Chemistry

The Royal Society of Chemistry (RSC) was an early adopter of ontologies for annotating entities in text mining. The RSC text mining team were contributors to the open biomedical ontologies ChEBI,[88] Gene Ontology,[89] and Sequence Ontology.[90] They have developed ontologies and made them available under open licenses. They were involved in the Open PHACTS project,[91] and they have published papers on the specifics of text mining in chemistry.

Thomas Hofweber in the *Stanford Encyclopedia of Philosophy*[92] defines the larger discipline of ontology as having four parts, but for the purposes of this talk, an ontology is a machine-readable account of what there is in a given domain and how the things there relate to other things. The semantic web does not work for chemical reactions; it does not even work for most molecules. Web Ontology Language (OWL)[93] ontologies are based on description logics (a family of formal knowledge representation languages) and the open-world assumption. Specifically, the description logics used have the tree-model property, which means that they are underconstrained and give rise to unintended tree models. Ontologies like ChEBI therefore do not represent molecules in a way that can be robustly reasoned over.

Consider also reactions. There being a pyrrole ring on the right-hand side of an equation is necessary for the reaction to be a pyrrole synthesis, but not sufficient. We cannot just define a cyclization as a reaction where a cyclic compound is formed: the Friedel–Crafts acylation produces a cyclic compound but is not a cyclization. Without being able to specify the changes around individual atoms, we cannot robustly classify reactions in OWL.

Ontologies are useful because they provide stable identifiers that can be reused across applications. They capture tacit knowledge and what is obvious to human beings but not to computers. They are human-readable definitions in plain text and, for automatic classification, machine-readable ones. They offer typed relations for systematic correspondences (e.g., between methods and instruments, and between reactions and products). The Open Biological and Biomedical Ontology[94] (OBO) framework lets you use other ontologies to help build your own. Ontologies give specifications for synonyms (exact, broad, narrow, and related) for use in text mining.

The RXNO name reaction ontology[95] is a formal ontology of chemical named reactions. It was originally developed at the RSC and is associated with OBO. The RXNO ontology unifies several previous attempts to systematize chemical reactions including the Merck Index and the hierarchy of Carey *et al.*[96] The RSC team took three chemists (two organic, one theoretical) and 100 name reactions, and decided on a principal axis of classification, in this case the objective of the reaction. They developed an initial flowchart and refined it in batches of 100 reactions. RSC chose to use a representation of organic reactions based on the intent of the chemist because reaction mechanisms are difficult to determine and can depend on reaction conditions, and because there was no point in replicating what can be done with reaction fingerprints or embeddings (discussed later in this talk).

There are further relations: "protects" connects a protection reaction to a given group; "deprotects" connects a deprotection reaction to a given group; "has specified product", "has specified reactant", "has catalyst" and "has intermediate" connect reactions to their participants in different roles; and "achieves planned objective" connects a planned process to an objective specification. Another RSC reaction ontology is the molecular process ontology[95] which contains the underlying molecular processes, for example cyclization, methylation, and demethylation. There are over 500 classes in RXNO with full human-readable definitions and varying degrees of axiomatization. The ontology is displayed in an Infobox alongside the Wikipedia entry for RXNO. The SVG file can be downloaded and reused. RXNO is used in NextMove Software's NameRxn[97] which allows the recognition and categorization of reactions from their connection tables.

It has been claimed that ontologies capture tacit knowledge and things that are obvious to humans but not to computers. To explore this claim, Colin carried out an experiment with embeddings. An embedding is a mapping of an element of a textual space (a word, a phrase, a sentence, an entity, a relation, a predicate with or without arguments, an Resource Description Framework (RDF) triple, an image, etc.) into an element of a (frequently low dimensional) vectorial space. A word representation is a mathematical object associated with each word, often a vector. In distributional semantics the hypothesis is that the meaning of a word can be obtained "from the company it keeps". In neural networks in natural language processing, the size of semantic spaces can be reduced with embeddings.

Tamara Polajnar, who works with Colin, hypothesized that word embeddings trained on a large corpus will contain facts about chemistry that can be tested using an ontology as a source of truth. Potential tasks are to predict reactants and products, given a reaction; to make analogies (e.g., "Grignard is to magnesium as Suzuki is to. . . ") and, given a definition, to predict what it is the definition of.[98] The last was the task Colin's team pursued.

Colin cautioned that the following results are very preliminary work, with many adjustable parameters not yet investigated. The researchers used fastText[99] to train 100-dimensional embeddings on the entire PubMed Central corpus and the entire RSC journal corpus, with two settings: with and without subwords. Chemlistem[100] is an RSC application for chemical named entity recognition with deep neural networks. Colin showed some examples of source text with chemically aware tokenization.[101] He presented some results (Figure 9). The results without subwords are marginally worse.

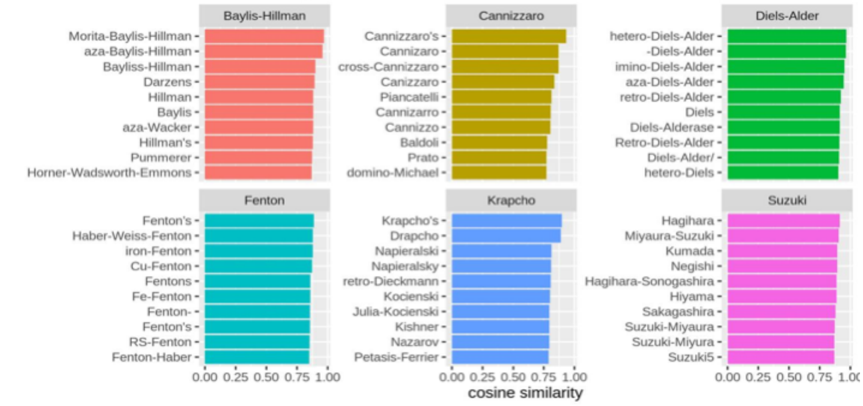


Figure 9: Cosine similarity, nearest neighbors with subwords.

Colin also showed some scatter plots (Figure 10) after dimensionality reduction with t-distributed Stochastic Neighbor Embedding.[102] The points are colored according to part-of-speech (code from J. Wijffels at GitHub[103] is used to assign parts of speech), and the proper names, which all came from name reactions, clustered together.

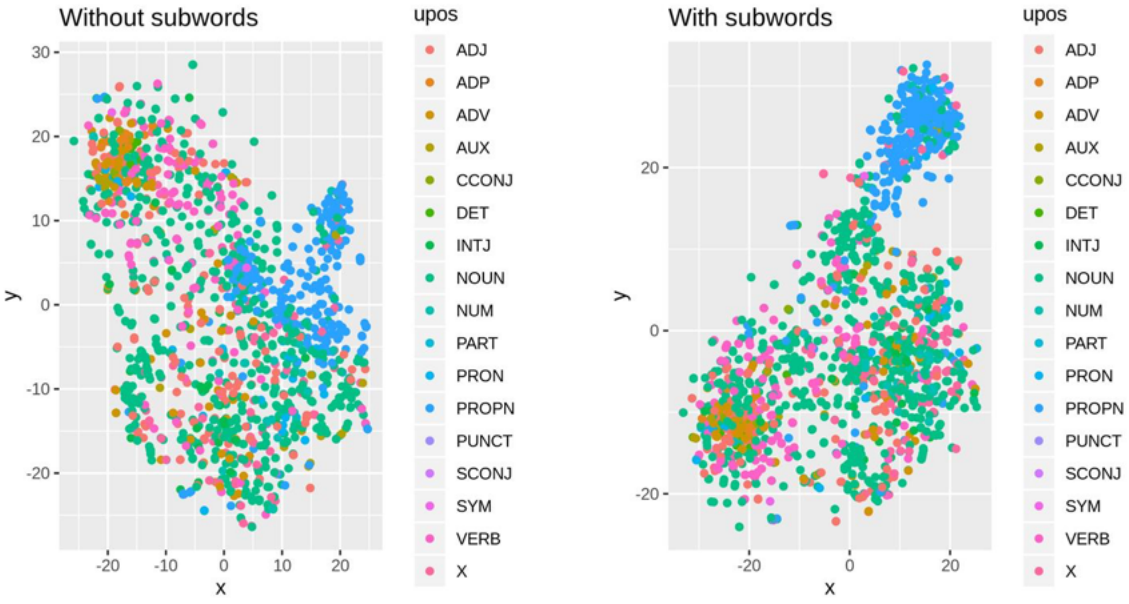


Figure 10: Clustering of names.

There are many ways of combining word embeddings to create embeddings for a whole phrase, some of them very simple.[104] In Colin's study he simply adds the embeddings together, and then ranks cosine similarity of the name embedding to the definition embedding. The dataset was RXNO ontology names and definitions, the development set 100 randomly-selected classes, and the test set 470 others. In the evaluation, AUC was 0.70 with subwords and 0.68 without subwords, compared with 0.53 for tf-idf-BoW. (Term frequency–inverse document frequency (tf–idf) is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. Bag of Words (BoW) is an algorithm that counts how many times a word appears in a document.)

The results suggest that embeddings do not capture everything: there is still a need for ontologies. Ontologies for chemical reactions are some way from being fully automated. They can serve as a gold standard for evaluating statistical methods in natural language processing. Colin has shown that there is a framework to assess quantitatively what ontologies have to offer.

14 UDM: a community-driven data format for the exchange of comprehensive reaction information

Dr. Jarek Tomczak, Pistoia Alliance

The mission of the Pistoia Alliance is to lower barriers to R&D innovation, enabling the development and incubation of integrated healthcare solutions. It is a not-for-profit organization established in 2008, founded by AstraZeneca, GSK, Novartis and Pfizer. It has more than 150 members, including 14 of the top 20 pharmaceutical companies, leading technology suppliers, and institutes, consultancies, publishers, and start-ups. The alliance currently has 10 ongoing projects and several communities. Jarek spoke about one of the alliance's projects.

The first reaction databases were developed in the early 1980s, and electronic laboratory notebooks have been in use in chemistry for almost 20 years, but we still do not have a well-defined way of capturing and exchanging information about chemical reactions, and we rely on imprecise or vendor-specific data formats. Without a common language and structure shared by all users to describe experiments or predictions, data integration is unnecessarily expensive, and large amounts of published data have not been readily available for processing or analysis. The Unified Data Model (UDM)[105] project delivers a solution to this problem.

The origin of the UDM was a Roche project (in 2012-2013) to integrate in-house chemistry data into Elsevier's Reaxys database. The project was further developed by Roche and Elsevier, with contributions from other pharmaceutical companies, as a data transfer format for chemical reactions from a variety of ELNs into Reaxys. The originators provided the UDM XML file format to the Pistoia Alliance and are committed to working together to make it more generic and to extend it to other experiment types. Founders were BIOVIA, Elsevier, GSK, Novartis, and Roche. Bristol-Myers Squibb joined in 2020.

The Pistoia Alliance UDM is a collective effort of vendors and life science organizations to create an open, extendable, and freely available reference model and data format for exchange of experimental information about compound synthesis and testing. In an ideal world, scientists would not have to interconvert files whenever one system has to talk to another one.

In the UDM, reactions are represented by the following entities:

- Reaction diagram (with optional atom-atom-mapping)
- Molecular properties
- Reactant, product, catalyst, solvent, reagent properties
- Conditions
- Analytical data
- Preparation section
- Scientists
- Literature and patent reference
- Reaction outcome
- Reaction scale

- Reaction classes
- Semantic annotations
- Comments
- Vendor data.

The RDfile[106] is a *de facto* standard data format for chemical reactions. RDfile implies 7-bit ASCII content whereas XML uses the 8-bit Unicode Transformation Format (UTF-8), a variable width character encoding capable of encoding all the valid code points in Unicode using one to four one-byte (8-bit) code units. Unfortunately UTF-8 is old-fashioned and not fully specified.

For UDM, RDfile is converted to XML. In the RDfile there is a tacit naming convention to represent a hierarchical data model ($DTYPE RXN:VARIATION(1):CONDITIONS(1): TEMP) and there are multiple, not necessarily compatible data hierarchies. An explicit data model is defined using an XML schema. In an RDfile there is no type control or validation of values of individual data field values ($DTYPE RXN:VARIATIONS(1):REACTANTS(1): EQUIVALENTS) whereas XML has strong typing and controlled vocabularies. An RDfile has a single representation of molecular structures (as molfiles) and reaction diagrams (rxnfiles), whereas in XML, multiple representations are allowed: molfile and rxnfile, InChI and RInChI, SMILES and Reaction SMILES, CDXML (the XML analogue of the binary CDX file type used by PerkinElmer's ChemDraw), and Wiswesser Line Notation. Validation, processing and conversion of RDfiles require dedicated tools or libraries whereas standard XML technologies provide a large part of the data processing operations: XML Schema (XSD) validation, XPath queries, and Extensible Stylesheet Language Transformations (XSLT).

Nevertheless, XML is not perfect. It has no understanding of chemistry; the XML schema has limitations; there are very few up-to-date implementations; and XML is too verbose. As an illustration Jarek showed how heavy MathML is when compared with TeX. He also showed a simplified UDM data model (Figure 11).

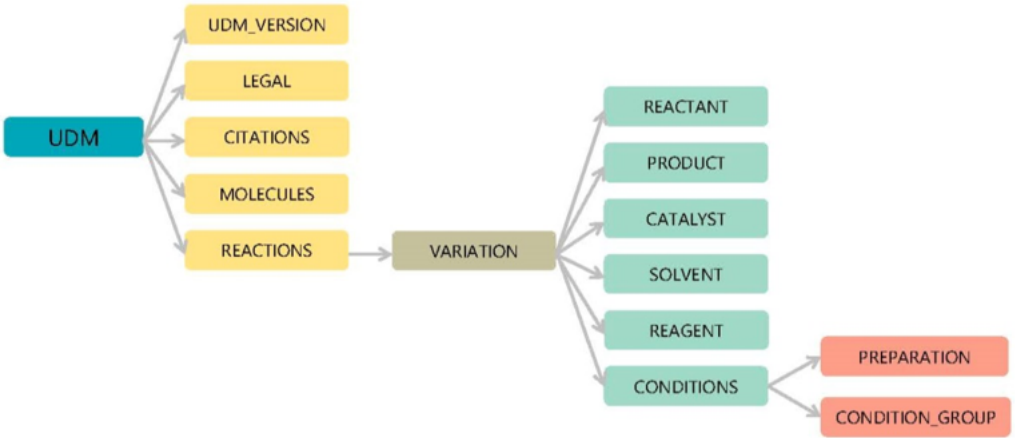


Figure 11: Simplified UDM data model.

UDM files can be downloaded from GitHub.[107] They comprise the UDM XML schema and vocabularies, sample datasets (currently 1000 reactions from Reaxys and 10,000 reactions from InfoChem's SPRESI), a license document, Python scripts to convert a SPRESI RDfile to the UDM, a glossary of key UDM terms, and a change log file. Version 6.0 was released in February 2020. Plans for further enhancements in 2020 include more sample datasets; reaction pathways; health and safety data; a possible ontology representation compatible with Basic

Formal Ontology (BFO),[108] and a Shapes Constraint Language (SHACL) model (a W3C language for validating RDF graphs against a set of conditions); and a UDM toolkit.

15 Retrosynthesis *via* machine learning

Dr. Marwin Segler, Benevolent.ai

Marwin listed three ways of designing new drug candidates: enumeration *via* building blocks and virtual reactions,[109] fragment spaces,[110] and *de novo* design.[111] Marwin uses *de novo* design and synthesis planning. *De novo* design can be thought of as the reverse of quantitative structure-activity relationships (QSAR) and virtual screening:[112] rather than analyze a set of compounds by means of its properties, you take a drug profile and predict a set of compounds to match it. You can learn to generate molecules with neural networks. Gomez-Bombarelli *et al.*[113] used Bayesian optimization and VAE (for VAE, see Niranjan's talk in this report). Marwin and his co-workers[114] used transfer learning and reinforcement learning. These approaches have had an impact in the pharmaceutical industry.[115–119]

After you have generated new molecules you want to know how to make them. One approach combines neural generative models and reaction prediction.[120] Here the generative model proposes a bag of initial reactants (selected from a pool of commercially-available molecules) and uses a reaction model to predict how they react together to generate new molecules. The generative model not only generates molecules, but also a synthesis route using available reactants. The reaction predictor used by Marwin and his co-workers is the Molecular Transformer[44] of Schwaller *et al.* Learning a way to decode to (and encode from) a bag of reactants, uses a parameterized encoder and decoder, and optimization in the latent space of a Wasserstein auto-encoder.

Another option for evaluating synthesizability is computer-aided synthesis planning (CASP).[121] Chemists are mainly using database search for CASP. This is the equivalent of using a hard copy road atlas instead of GPS navigation. Chemists could change their way of working in future. Synthesis planning is both an art and a science. It requires knowledge, experience, and creativity. Even experts are not perfect: Corey himself admitted this.[122] Moreover, not every molecule is synthesizable yet.

Games such as chess and Go are also an art. Recently, Silver *et al.*[123–125] showed that deep reinforcement learning coupled with Monte Carlo Tree Search could be used to train programs to learn how to play games by self-play, where the computer plays millions of games against itself in simulation. They only require specification of the game rules, which provide an exact simulator of the (game) environment. With this setup, they were able to create a program (AlphaZero) that won against the strongest grandmasters in Go, and beat the strongest known Go, Chess, and Shogi engines.

In most real-world domains, however, interactions with the environment are very expensive, and the rules of the environment are too complex to specify. Therefore you cannot learn and plan online. Robert Robinson solved the problem of synthesizing tropinone by recursively breaking the synthesis into simpler problems. A retrosynthesis reaction scheme, used in synthesis planning, is the reverse of a synthesis scheme. CASP can be measured and improved[51] by computational benchmarks, on which the chemist can give feedback, before conducting a synthesis in the laboratory. Computational benchmarks are fast to apply; synthesis in the laboratory is more accurate. If you can measure performance, you can improve performance, but benchmarks have been absent in the past. CASP challenges are to get the rules of chemistry into the machine; to prioritize the rules; to filter out infeasible reactions; and efficient search.

George Vléduts[126] was the first to suggest the representation of a reaction as a single graph; Corey and Wipke[6] later attempted to write down all chemical knowledge in a logical form. Unfortunately, in such systems rules have to be entered by hand;[127–129] there are reactivity conflicts; selectivity has to be explicitly encoded; purification, solubility, and stability are not taken into account; and there is no inherent ranking mechanism. It is not easy to decide on the correct rule to apply. Chemical knowledge grows exponentially, and doubles every decade. Manual coding of rules has been tried for 60 years and it does not work: expert systems do not work at scale.

So, Marwin and Mark Waller's team had the idea of combining rules with machine learning: learning transformation rules from data; learning to prioritize the rules; learning to predict reactions; and implementing modern, efficient search.[68,121,130] Machine learning provides a rigorous metrics framework. A seq2seq algorithm such as the Molecular Transformer[44] "translates" from products to reactants but does not generalize outside the training data, needs to learn from a small number of examples per rule, and needs to translate to the top k solutions correctly. Marwin compared rules-based methods with machine learning methods (Table 1).

	Data-efficient	Robust to noise	Generalizes
Purely Symbolic	Yes	No	Yes
Purely neural	No	Yes	No
Neural-symbolic	Yes	Yes	Yes

Table 1: Rules *versus* machine learning.

The 11 million reactions in Reaxys contain the data of our entire discipline. From these reactions Marwin and his colleagues learned 301,671 rules, with a minimum of three examples per rule. Successful reactions contain implicit knowledge. Only the reaction center is included in these rules. Other methods of automatic extraction of rules have been reported.[57,131,132]

In the retrosynthesis procedure used by Marwin's team,[51] the target molecule was described by ECFP4 descriptors which were input to a deep highway neural network.[133,134] The most probable rules were output, and they were applied to the retrosynthesis of the target in question. Machine learning not only predicts the transformation rules, but it also allows the tolerated molecular context to be learned. The model was trained on 3.5 million reactions. Reactions reported in 2014 or earlier were used to build the graph and reactions reported in 2015-2017 were used to test the model. This sort of time-splitting of training and test sets has been recommended by Sheridan.[135]

Reaction prediction can be used to filter out infeasible reactions. Marwin and his co-workers have proposed a model that mimics chemical reasoning, and formalizes reaction prediction as finding missing links in a knowledge graph.[136] Since there are so few failed reactions in the literature, the researchers have devised their own set of failed reactions.[136] The reactions are represented by reaction fingerprints: see also the work of Schneider's team on reaction fingerprints[137] and Gillet's team on reaction vectors.[138] Coley *et al.* have also devised a way of eliminating "false positives".[52] Marwin's in-scope filter had an ROC AUC of 0.986 for the neural network and a false positive rate of 1.5%. The score output correlates with LUMO energies and Hammett parameters. The neural symbolic procedure for rule selection and reaction prediction is summarized in (Figure 12).

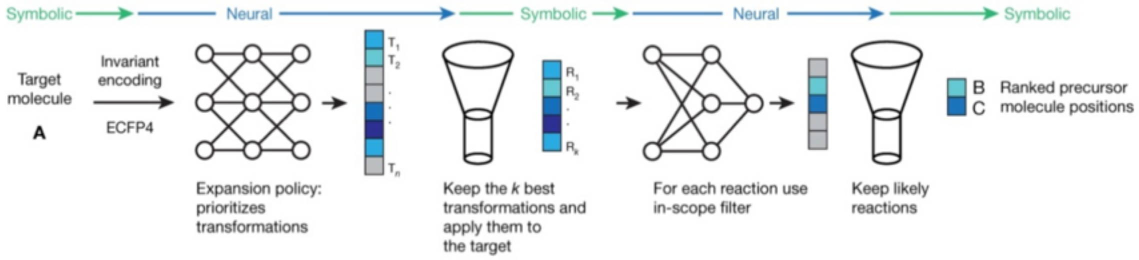

Symbolic
Neural
Symbolic
Neural
Symbolic
Target molecule
A
Invariant encoding
ECFP4
T_1
T_2
T_n
Expansion policy: prioritizes transformations
R_1
R_2
R_k
Keep the *k* best transformations and apply them to the target
For each reaction use in-scope filter
Keep likely reactions
B
C
Ranked precursor molecule positions

Figure 12: Rule selection and reaction prediction.

Finally, there is the challenge of efficient search. In heuristic best first search (BFS) a strong heuristic function is used to score the nodes in the retrosynthesis tree. Unfortunately, chemists disagree about good solutions, synthesis is solved only at the end, and molecular complexity needs to be tactically increased, for example by introducing protecting groups. In a search algorithm, we care mainly about the state we start with, synthesis is solved and able to be scored only at the end after decomposition into building blocks, and the value function needs to be computed online, since it is dependent on the building blocks (although it is also possible to learn a value function).

In MCTS,[139,140] approximate action values are calculated by random Monte Carlo simulation (an agent picks transforms randomly until the end of the synthesis) and these approximated values are used to build the search tree. The method is not dependent on a strong heuristic and can deal with very high branching factors. A quantitative analysis of three methods for 500 random molecules is presented in Table 2.

Method	Scoring	Percentage solved	Time per molecule
BFS	Heuristics[129]	56	422
BFS	Neural net	84	39
MCTS	Neural net	95	13

Table 2: Quantitative analysis of search methods.

For a qualitative analysis, a “chemical Turing test” was carried out by 45 Ph.D. chemists at Shanghai and Münster Universities (Figure 13).[68] In this double-blind A/B test, chemists on average considered the computer-generated routes as preferable to reported literature routes. (A/B testing is a way to compare two versions of a single variable, typically by testing a subject’s response to variant A against variant B, and determining which of the two variants is more effective.)

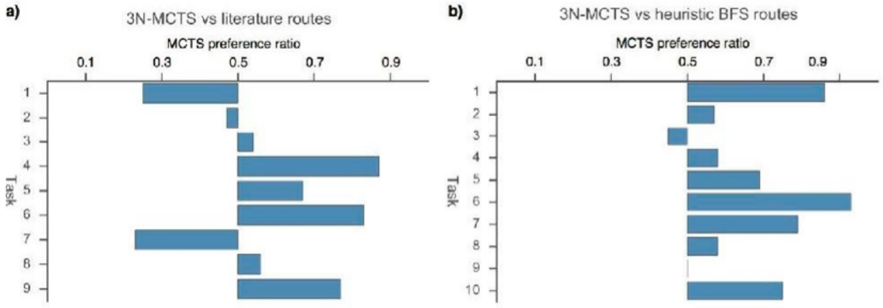


Figure 13: Double-blind AB testing of MCTS-derived routes against literature and BFS routes.

There are some unsolved challenges. It is unclear how to integrate conditions in the search. Chemists have not decided what they mean by failed reactions. There is no quantitative stereoselectivity and yield prediction yet. The system does not work well yet for natural products, the best models may not have been found yet, and the system cannot invent novel reactions, although that problem is semisolved.[136]

In future, the computer will only get better. Two new models have been published recently.[50,141] Chess engines have made chess players much stronger; CASP will make organic chemists much stronger. How will molecular design change if we can confidently order more uncommon molecules? Will we ever overcome trial and error?

16 From mechanisms to reaction selectivity

Professor Per-Ola Norrby, AstraZeneca

Forward reaction prediction has wide applications, from general scoring of retrosynthesis pathways to selection of appropriate reaction conditions. In drug discovery, many compounds are made, time is of the essence, yield is less important than in process development, impurities are removed by column, there is a low amount of waste, and reactions are safe on a small scale. Prediction needs to be fast and amenable to automation. In process chemistry, one compound is being made, it takes months or years to develop routes, the yield must be optimized, impurities are to be avoided, the chemistry must be sustainable, and safety is paramount. Therefore, prediction must be accurate, and must provide knowledge for optimizing the yield.

AstraZeneca is developing a platform for data mining coupled with machine learning and AI to exploit in developing robust, scalable chemical processes for drug development. It is believed that the integration of process-focused reactivity data and machine learning will drive the development of novel predictive process models.[142] Retrosynthesis by machine learning handles chemical logic and literature precedents; the reverse, forward prediction, answers questions such as how reaction conditions will affect chemoselectivity, regioselectivity, and enantioselectivity.

Stereogenic centers are ubiquitous in pharmaceutical compounds so asymmetric catalysis is an important subject. It is not easy to predict how a catalyst will influence selectivity. Optimizing an organometallic catalyst is an iterative process involving virtual screening of a chiral ligand library, reaction screening based on conditions and additives, and optimization of costs and properties.

Reaction selectivity prediction using quantum chemical methods is well established but the accuracy:cost ratio is not favorable, especially when multiple pathways must be considered. Per-Ola and his colleagues have therefore explored several alternatives with improved accuracy or lowered cost, still within a strong mechanistic framework and using density functional theory (DFT) as an essential component.

One way to model catalyst stereoselectivity is QSAR, or, in particular, quantitative structure selectivity relationships.[143] Multivariate linear regression methods are versatile, statistical tools for predicting and understanding the roles of catalysts and substrates and act as a useful complement to complex transition state calculations, with a substantially lower computational cost.

A second method is based on quantum mechanics (QM): DFT-based screening. The stereoselectivity of most organocatalytic reactions hinges on the balance of both favorable and unfavorable noncovalent interactions in the stereocontrolling transition state. Wheeler *et al.*[144] have reviewed attempts to understand the role of noncovalent interactions in organocatalyzed reactions and to develop new computational tools for organocatalyst design.

A third method is Quantum to Molecular Mechanics (Q2MM).[145–147] The accurate computational prediction of stereoselectivity in enantioselective catalysis requires adequate conformational sampling of the selectivity-determining transition state, but it has to be fast enough to compete with experimental screening techniques if it is to be useful for the synthetic chemist. Although electronic structure calculations are accurate and general, they are too slow to allow for sampling or fast screening of ligand libraries.

The combined requirements can be fulfilled by using appropriately fitted transition state force fields (TSFFs) that represent the transition state as a minimum and allow fast Monte Carlo conformational searches. Quantum-guided molecular mechanics (Q2MM) is an automated force field parametrization method that generates accurate, reaction-specific TSFFs by fitting the functional form of an arbitrary force field using only electronic structure calculations, by minimization of an objective function. After validation of the TSFF by comparison to electronic structure results for a test set and available experimental data, the stereoselectivity of a reaction can be calculated by summation over the Boltzmann-averaged relative energies of the conformations leading to the different stereoisomers.

Per-Ola's team has applied the Q2MM method to perform virtual ligand screens on a range of transition metal-catalyzed reactions Correlation coefficients of 0.8-0.9 between calculated and experimental enantioselectivity values are typical for a wide range of substrate-ligand combinations, and suitable ligands can be predicted for a given substrate with about 80% accuracy. Although the generation of a TSFF requires an initial effort and will therefore be most useful for widely used reactions that require frequent screening campaigns, the method allows for a rapid virtual screen of large ligand libraries to focus experimental efforts on the most promising substrate-ligand combinations.

The method has been implemented in a web-based virtual screening workflow at AstraZeneca. CatVS is an automated tool for the virtual screening of substrate and ligand libraries for asymmetric catalysis within hours. Per-Ola's team has demonstrated predictive computational ligand selection in the virtual ligand screen of a library of diphosphine ligands for the rhodium-catalyzed asymmetric hydrogenation of enamides. Subsequent experimental testing verified that the most selective substrate-ligand combinations are successfully identified by the virtual screen.[148]

Finally, Per-Ola discussed a hybrid machine learning and mechanistic model. Machine learning allows machines to learn from data, with continuous self-improvement, but "big data" is required. A mechanistic model uses a detailed reaction mechanism and direct calculation of selectivity, and no data may be needed. A hybrid model has a strong framework from theory, biased with knowledge, and trained with minimal data. Per-Ola's team is testing three different approaches: machine learning models using quantum mechanical (QM) descriptors; augmenting experimental data with QM-generated data; and a hybrid model to compensate, with machine learning, for low-level QM systematic errors (see the following talk by Kjell Jorner). Per-Ola gave a diagram (Figure 14) showing the pros and cons of the modeling approaches for selectivity.

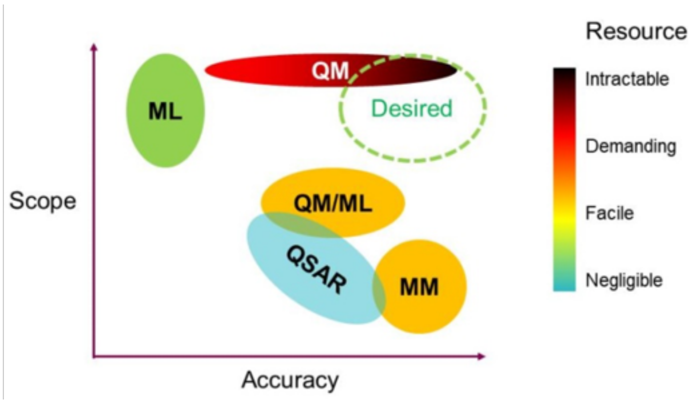


Figure 14: Modeling approaches compared.

Per-Ola's team have applied the hybrid machine learning and mechanistic approach to selectivity in C-H functionalization. In the early stages of the drug discovery process, thousands of compounds are synthesized. Recently, late stage functionalization has become increasingly prominent since these reactions selectively functionalize C-H bonds, allowing chemists to produce analogues quickly. Classical electrophilic aromatic halogenations are a powerful type of reaction in the late stage functionalization toolkit. The introduction of an electrophile in a regioselective manner on a druglike molecule is a challenging task. Per-Ola's team have reported a machine learning model able to predict the reactive site of an electrophilic aromatic substitution with an accuracy of 93%. The model takes as input a SMILES of a compound and uses six quantum mechanics descriptors to identify its reactive sites. On an external validation set, 90% of all molecules were correctly predicted.[149,150] These results are encouraging for the hybrid model.

17 Reaction prediction in process chemistry with hybrid mechanistic and machine learning models

Dr. Kjell Jorner, AstraZeneca

Predicting the outcome of reactions in process chemistry with mechanistic models based on quantum chemistry has advantages over purely knowledge-based approaches. These advantages include quantitative measures of reactivity, and detailed understanding of the mechanism, the structural factors controlling reactivity, and the effect of catalysts or reagents. On the other hand, the applicability of current methods is limited by the sometimes poor accuracy of DFT and implicit solvent models, and by incomplete treatment of dynamic effects.

Consequently, chemical accuracy with errors below 1 kcal/mol can be reached only in fortuitous cases. One way forward could be correcting the computed values using machine learning with experimental data.

Data-driven methods for reaction prediction, such as IBM RXN,[151] learn from large quantities of data, and represent molecules as text strings or graphs. The chemistry is implicit and the method is general. Structure and property driven methods need fewer data and use chemical descriptors (atomic, molecular, and vibrational descriptors for the components of a palladium-catalyzed Buchwald-Hartwig cross-coupling in work reported by Ahneman *et al.*).[152] The chemistry is explicit and the method is specialized. Hybrid methods attempt to combine traditional DFT modeling with structure- and property methods.

To test the hybrid method, Kjell and his co-workers chose to study nucleophilic aromatic substitution (S_NAr). In this reaction, an anionic or neutral nucleophile displaces a leaving group in a concerted or stepwise mechanism. The reaction is influenced by acid-base catalysis and solvent effects. Nine percent of all reactions carried out in the pharmaceutical industry are nucleophilic aromatic substitutions.

Kjell and his co-workers used activation energies from kinetic data to predict absolute reactivity directly, and chemoselectivity and regioselectivity, from differences in activation energy, and the effects of a catalyst with explicit acid or base. The Predict S_NAr Python program automates reaction modeling. It extracts reactive species and atoms from a reaction SMILES, finds structures and energies, calculates quantum-mechanical descriptors, and stores the results in a database. Some hours are needed to run a reaction model. The problem of conformational flexibility has been solved by sampling; ionic reactions in the solvent have been addressed by an implicit and explicit solvation model; but the problem of acid and base catalysis has not yet been solved by explicit inclusion in the calculation.

Electronic descriptors include (*inter alia*) local electron attachment energy of the substrate (soft); average local ionization energy of the nucleophile (soft); surface electrostatic potential (hard); electrostatic potential at nuclei (hard); and global electrophilicity and nucleophilicity. Solvent-accessible surface area and the P_{int} descriptor of dispersion interaction potentials are steric descriptors. Solvent descriptors cover the environment. Activation energies, bond orders and charges, and other descriptors represent transition states.

Overall, Gaussian process regression (GPR) with radial basis function (RBF) kernel and support vector regression with RBF kernel outperformed linear regression and random forest. In test fold predictions of the energy barrier, for 10-fold cross-validation, GPR performed significantly better than DFT: mean absolute error (MAE) 0.8, root mean square error (RMSE) 1.4, and R^2 0.86, as opposed to MAE 2.9, RMSE 3.8, and R^2 0.11. Kjell next studied whether transition state descriptors help: MAE, RMSE, and R^2 were hardly affected by using no descriptors from the transition state calculations. A plot of MAE against the number of reactions showed that about 200 reactions were needed to achieve chemical accuracy (MAE $<$ 1 kcal/mol).

The proof-of-principle model of nucleophilic aromatic substitution can predict absolute reactivity, and chemoselectivity and regioselectivity, and the effect of solvent and catalysts. This information can be used by the synthetic chemist to make risk assessment of the steps in long synthetic routes. It also provides detailed understanding of the mechanism that can be used together with synthetic and mechanistic experiments to guide development of improved reaction conditions.

In future the team aims to finalize modeling and apply it to an external test set; to study the effects of acid and base catalysis; to extend the work to more reaction classes; and to gather more kinetic data and make databases from the literature, high-throughput experimentation, and different solvents and conditions.

18 Automated mining of a database of 9.3 million reactions from the patent literature, and its application to synthesis planning

Dr. Roger Sayle, NextMove Software

Pistachio is a reaction dataset interface providing loading, querying, and analytics of chemical reactions. It builds on and extends existing solutions from NextMove Software to enrich reaction data and provide query capabilities. The company uses text mining to extract a database of chemical reactions automatically from the patent literature. The reactions in the so-called USPTO dataset[39] were extracted using an enhanced version of Daniel Lowe's reaction extraction code[153] with NextMove's LeadMine being used for chemical entity recognition. "USPTO" is freely available.[39] Pistachio is not free; it is an improved version with extra features such as advanced entity recognition.

Elsevier's Reaxys, CAS SciFinder, and InfoChem's SPRESI are manually curated databases. USPTO and Pistachio are machine curated. Metrics for evaluating reaction databases include availability and price; coverage, size, and currency; and quality and annotation.

Roger showed a diagram of the workflow for the construction of Pistachio, and the NextMove Software involved (Figure 15). Currently Pistachio covers 9,320,005 patents (from 1976 to 2020) mined from the text of U.S. patent applications and grants, and European Patent Office (EPO) applications and grants, plus U.S. application and grant sketches. The number of unique patents (i.e., parents) is 2,945,919. At 5.3TB source, this is "big data".

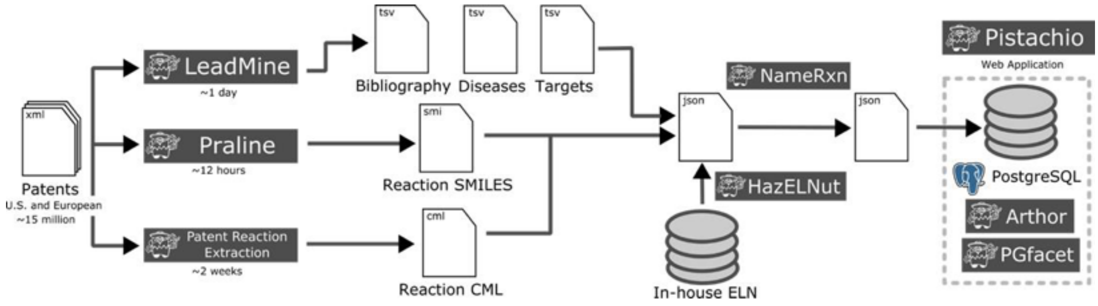


Figure 15: Pistachio workflow.

NextMove's CaffeineFix is used to match chemical names against a dictionary or grammar. It supports very large user dictionaries. NextMove handles an entity dictionary as a directed acyclic graph (dag). In mathematics, particularly graph theory, and computer science, a dag is a finite directed graph with no directed cycles. Deterministic finite automaton (DFA) minimization is the task of transforming a given DFA into an equivalent DFA that has a minimum number of states. Nitrogen containing heterocycles as minimal DFA include pyrrole, pyrazole, imidazole, pyridine, pyridazine, pyrimidine, and pyrazine. CaffeineFix is a chemical spell checker. For example, it corrects di-terf-butyl (4S)-/V-(fert-butoxycarbonyl)-4-{4-[3-

(tosyloxy)propyl]benzyl}-L-glutamate to di-tert-butyl (4S)-N-(tert-butoxycarbonyl)-4-{4-[3-(tosyloxy)propyl]benzyl}-L-glutamate.

The BioCreAtIvE V challenge[154] evaluated text-mining and extraction systems. Roger showed a plot of the web service response time to annotate an abstract evaluated for a chemical-disease relation task. NextMove's submission to the challenge was many times faster than its competitors. The company's efficient rule-based text-mining provides provenance for annotations and can mine the entire back-archive of U.S. patents in about 24 hours on a single machine.

Roger listed some advanced entity recognition features of Pistachio. These include improvements to Name2Structure, and to dictionaries and ontologies. Molecular formulas and line formulas (e.g., K_2CO_3, $PdCl_2(PPh_3)_2$, $Pd(P(o\text{-tolyl})_3)_2$, and $PdCl_2(dppf)\text{-}CH_2Cl_2$); inorganics, organometallics and salts; mixtures and formulations (e.g., 5% 2M methanolic ammonia/DCM, and 10% H_2O_2 in water); and apparatus (e.g., a 1L three-necked, round-bottomed flask) are all handled.

Pistachio re-interprets ChemDraw sketches, correcting systematic errors, extracting extra semantics, such as structure variation and reaction schemes, and categorizing sketches as interpretable or not. An example of ChemDraw sketch interpretation is given in Figure 16. Multiple single-step reaction equations are generated from the sketches of multistep reaction schemes.

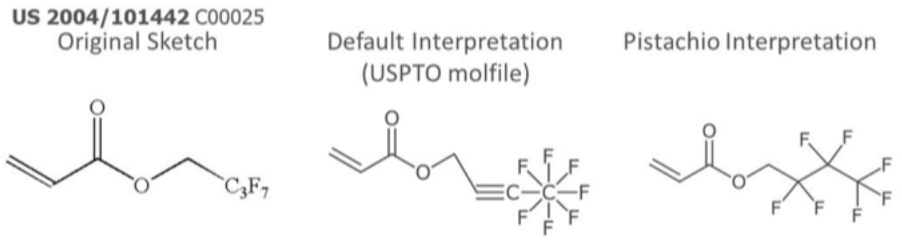


Figure 16: Interpretation of a ChemDraw sketch.

Where there is an identifier for a compound (e.g., "Compound 1", or "Reference compound 1", or "Example 1", or "cmpd 1" or "cpd. #1"), the structure and the identifier need to be paired up. The identifier may be defined multiple times (e.g., as a sketch and a chemical name). The identifier and the chemical structure may be in columns in a table. Single structures are enumerated from the rows in tables of generic ("Markush") structures.

Additional annotation includes a company ontology (so that Ciba-Geigy is recognized as Novartis); and calculated yields. Reaction steps and recipes are annotated in accordance with ANSI/ISA-88, a standard addressing batch process control, and a design philosophy for describing equipment and procedures.

Using published reaction classification categories,[96,155] the contents of one pharmaceutical ELN may be presented as a pie-chart (Figure 17) of the kinds of transformations it contains. Reactions in Pistachio are classified into a common subset of the classes defined by Carey *et al.*[96] and the RSC's RXNO ontology[95] There are 12 super-classes (e.g., code 3, C-C bond formation (RXNO:0000002)). These contain 84 categories (e.g., 3.5, Pd-catalyzed C-C bond formation (RXNO:0000316)), and the 84 categories contain about 300 named reaction types (e.g., 3.5.3, Negishi coupling (RXNO:0000088)). These require about 2490 SMIRKS-like transformations. Workers at NextMove and Novartis have published[137,156] on this classification work and its applications.

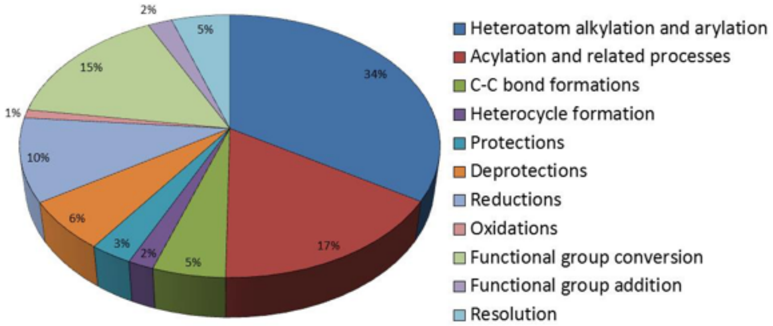
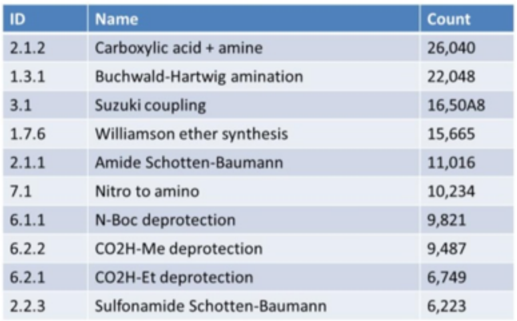


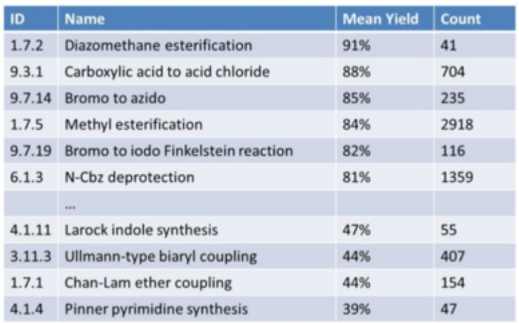
Figure 17: Reactions in a pharmaceutical industry ELN.

NextMove Software's NameRxn assigns reactions to over 1150 reaction categories using transformations. It can guarantee perfect atom-atom mapping. Mapping is an output not an input. Maximal common substructure mappers struggle with rearrangements. NameRxn expresses named reactions as SMIRKS strings. Writing SMIRKS is both an art and a science. The percentage of reactions with product atoms mapped is 71% for Marvin 6.0, 62% for ChemDraw 12, and nearly 80% by consensus. When those methods are combined with NameRxn, the numbers rise to over just over 80% for Marvin and ChemDraw, and nearly 90% by consensus. Roger listed Pistachio's 10 most popular reactions (Table 3), and the most and least successful (Table 4). Trends in named reaction and solvent use over the years can also be studied.



ID	Name	Count
2.1.2	Carboxylic acid + amine	26,040
1.3.1	Buchwald-Hartwig amination	22,048
3.1	Suzuki coupling	16,50A8
1.7.6	Williamson ether synthesis	15,665
2.1.1	Amide Schotten-Baumann	11,016
7.1	Nitro to amino	10,234
6.1.1	N-Boc deprotection	9,821
6.2.2	CO2H-Me deprotection	9,487
6.2.1	CO2H-Et deprotection	6,749
2.2.3	Sulfonamide Schotten-Baumann	6,223

Table 3: Pistachio's 10 most popular reactions.



ID	Name	Mean Yield	Count
1.7.2	Diazomethane esterification	91%	41
9.3.1	Carboxylic acid to acid chloride	88%	704
9.7.14	Bromo to azido	85%	235
1.7.5	Methyl esterification	84%	2918
9.7.19	Bromo to iodo Finkelstein reaction	82%	116
6.1.3	N-Cbz deprotection	81%	1359
	...		
4.1.11	Larock indole synthesis	47%	55
3.11.3	Ullmann-type biaryl coupling	44%	407
1.7.1	Chan-Lam ether coupling	44%	154
4.1.4	Pinner pyrimidine synthesis	39%	47

Table 4: Most and least successful reactions.

Are solvents getting greener? An analysis of the USPTO grants 1976 – 2013 dataset, shows that the top 10 solvents, water, ethanol, benzene, methanol, tetrahydrofuran, dichloromethane, dimethylformamide, acetic acid, chloroform, and acetone, in decreasing order of popularity, were used in 71% of reactions in 1976. In 2013, tetrahydrofuran, dichloromethane, water, dimethylformamide, methanol, ethyl acetate, ethanol, 1,4-dioxane, toluene, and acetonitrile were used in 82% of reactions.

Some rare named reactions are:

- Adams decarboxylation
- Angeli-Rimini reaction

- Aza-Baylis-Hillman reaction
- Boyer reaction
- Buchwald-Fischer indole synthesis
- Castro-Stephens coupling
- Chapman rearrangement
- Chugaev elimination
- Cook-Heilbron thiazole synthesis
- Fischer-Hepp rearrangement
- Fukuyama indole synthesis
- Gasman indole synthesis
- Imine Hosomi-Sakurai reaction
- Koch reaction
- Leuckart reaction
- Liebeskind-Srogl coupling
- Lossen rearrangement
- Ponzio reaction
- Prins reaction
- Reimer-Tiemann carboxylation.

Prediction is much harder than analysis. Consider hurricanes and tornadoes: it is much easier to follow the path of destruction by locating devastated neighborhoods, than to forecast the paths of such weather systems in advance. Prediction is also much harder than analysis for many chemical reactions. A tricky benchmark in regioselectivity is reactions of 2,4,5-trichloropyrimidine. The nature of pyrimidine makes the chloro at the 4-position more reactive than that at the 2-position which is more reactive than that at the 5-position. Simple quantum mechanical methods have difficulty discerning this order.

NameRxn can be applied in reaction planning. The top 10 most popular syntheses of cinnamic acid range from the bromo Heck reaction (272 examples), and the Horner-Wadsworth-Emmons reaction (268), down to the Stille reaction (2), and olefin metathesis (1). In nitration, (by refluxing with nitric acid and sulfuric acid), the appearance of one or more nitro groups indicates a nitration reaction, but predicting where on a nontrivial organic molecule this functional group appears is a much harder challenge. In this sense, reaction analysis is much simpler than (either forward or retrosynthetic) synthesis planning. *p*-Nitrotoluene has been prepared 96 times by nitration, once by a bromo Suzuki-type reaction, and once by a chloro Suzuki-type. On the other hand *p*-nitrobenzoic acid has been prepared 12 times by nitrile to carboxy, eight times by CO_2H-Me deprotection, five times by CO_2H-Et deprotection, once by ester hydrolysis, and only once by nitration.

Pitt *et al.*[157] have generated a complete list of 24,847 small aromatic ring systems. Using a machine learning approach, they estimated that the number of unpublished, but synthetically tractable, rings of this type could be over 3000, and the rate of publication of novel examples could be as low as 5-10 per year. This should provide fresh stimulus to creative organic chemists by highlighting a small set of apparently simple ring systems that are predicted to be tractable but, as yet, appear to be unconquered.

19 The semantic laboratory

Dr. Samantha Kanza, University of Southampton

The Semantic Web is the web of linked data. It is a way to bring context and meaning to data, and a set of common standards for data representation, integration, and search. The Resource Description Framework (RDF) is the Semantic Web's machine-readable linked data format. A semantic triple is the atomic data entity in the RDF data model. It is a set of three entities that codifies a statement about semantic data in the form of subject→predicate→object expression (e.g., "Bob knows John"). An ontology (see also the talk by Colin Batchelor) defines different concepts within a domain including hierarchies, relationships to other concepts, and terms used to refer to those concepts. Ontologies are written in OWL[93] or RDF Schema (RDF(S)).[158] SPARQL[159] is an RDF query language able to retrieve and manipulate data stored in RDF format. It facilitates search on concepts rather than text.

An affordance is any object that offers, or affords, its user the opportunity to perform an action. Affordances of the Semantic Web are defining common shared vocabularies for reuse; making data machine-readable; data linkages; inferencing; semantic search; and unlocking the potential of AI and machine learning. Common shared vocabularies enable a shared language to describe scientific data, concepts, and research. Scientific research can be significantly enhanced by linking datasets together to find undiscovered links, and answer questions that cannot be addressed with a single data source.[160] Description logic can be embedded into ontologies, enabling machines to "infer" additional information that is not explicitly defined in the data. For example, if Samantha is allergic to juniper, and gin has botanical juniper, then we can infer that Samantha is allergic to gin. Semantic search is searching on concepts across linked datasets. All this unlocks the potential of AI and machine learning because you do not waste the potential of your algorithms by feeding them garbage. High quality data avoid the "garbage in, garbage out" problem.

We need semantics at every stage of the scientific research process from planning through to publication. There is often a disconnection between a scientist's laboratory, laboratory notebooks, and electronic data. We need to link these areas together with interoperable data. Our scientific data and laboratories should be digital, consistent, understandable, searchable, and linkable. In short, our scientific data and laboratories should be semantic.[161]

Samantha outlined three current projects: Layered Semantic Laboratory Notebooks, Semantically Enhanced ELN Data (SEED, a Pistoia Alliance project), and Talk2Lab,[162] a connected "Lab of the Future". Layered Semantic Laboratory Notebooks was the subject of Samantha's Ph.D. thesis.[163] The work proposed an ELN environment using a three-layered approach: a notebook layer consisting of an existing cloud-based notebook;[164] a domain-specific layer with the appropriate knowledge; and a semantic layer that tags and marks up documents.

In Phase 1 of SEED, Pistoia Alliance members will reach a precompetitive agreement on the need for open standards and solutions to deliver semantic enrichment to ELNs, enabled by findable, accessible, interoperable, and reusable (FAIR) data. Overall, SEED aims to link ontologies into ELNs, to identify and use standard ontologies to describe different types of data, and annotate experiments semantically; to capture the provenance of data; to develop an effective application programming interface or open standard to connect to an ELN; and to link to the Lab of the Future.

Talk2Lab[162] is a University of Southampton project integrating smart devices in a laboratory environment. Users can "Talk to" Alexa to retrieve real time data from their laboratory; use sensors to monitor temperature, water flow and laser power in a laser laboratory; and get alerts and warnings through the collaboration hub Slack for out-of-specification readings.

Given the focus of this conference, and to bring home the usefulness of the Semantic Web, Samantha concluded by saying: "To the well-organized linked dataset, AI is but the next great adventure".

20 ASKCOS: data-driven chemical synthesis

Dr. Connor Coley, MIT

The hit-to-lead and lead optimization stages of the small molecule discovery and development pipeline involve a synthesize-design-test cycle in which hundreds or thousands of compounds are tested. Each cycle can take weeks or months. The procedures rely on labor-intensive planning, and synthesis.[165] Productivity could be improved by using information more efficiently (designing better compounds) or obtaining information more quickly (testing more compounds).[166,167] The discovery process involves a balance between the value of information and the cost of validation. Connor's talk concerned a reduction in the cost of validation.

In a typical retrosynthesis plan, the input is a product (the target compound), and the output is reactants (starting materials), intermediates, and conditions (without concentrations and reaction times). In Connor's vision for autonomous synthesis, the user inputs a command such as "make lidocaine for me", software converts this into instructions for a synthesis robot, and the output is a sample of lidocaine. How do we use 200 years' worth of historical reaction data (journal articles, patents, etc.) to help us rapidly synthesize new molecules? The team at MIT works on generalizing known chemistry to novel substrates using techniques in machine learning and artificial intelligence.

The earliest computer-aided synthesis (CASP) programs sought to codify expert chemist knowledge about what reactions are allowed.[168] As an alternative to manual encoding of allowable transformations, heuristics for algorithmic extraction have been developed.[56,57,68,132,136,169,170] These algorithms build generalized rules from known reaction examples. They identify the atoms that change connectivity as the reaction center. Different levels of generalization can then be used to extend that reaction center to include varying numbers of neighbors using either a fixed distance or heuristics that decide which neighboring atoms are relevant.

There are additional approaches to computer-aided retrosynthesis that avoid the need for reaction templates entirely. These include sequence-to-sequence models,[49,171] similarity-based methods,[52] and some graph-based methods that to date have been applied only to the problem of forward prediction.[172] Whether or not a chemical reaction can proceed is not defined by hard decision rules, and Connor's team has proposed a new approach to determining when reaction templates should be applied.[50]

Connor presented a workflow of algorithmic synthesis design (Figure 18).[173] There are many reasons why a reaction recommended by a computer might not succeed: lack of selectivity, the presence of competing reaction pathways, the reaction requiring a different context, the reaction being equilibrium-limited, and so on. The workflow of algorithmic synthesis design should thus include a loop to see if the reaction is feasible (see the top right loop of Figure 18).

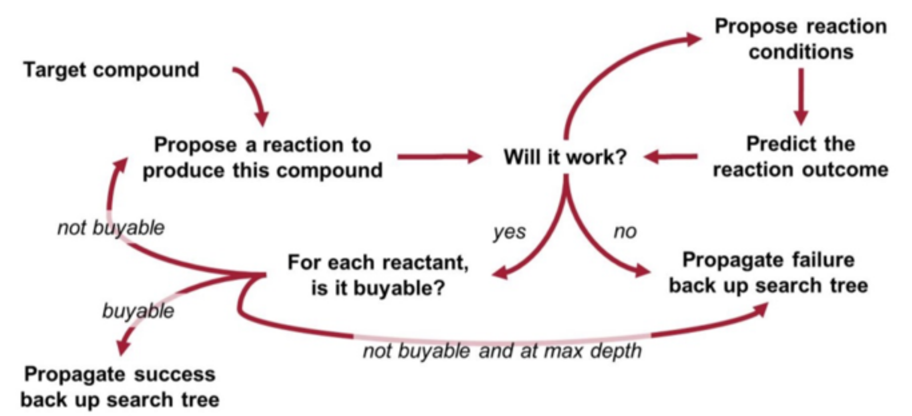

Target compound
Propose reaction conditions
Propose a reaction to produce this compound
Will it work?
Predict the reaction outcome
not buyable
yes
no
For each reactant, is it buyable?
Propagate failure back up search tree
buyable
not buyable and at max depth
Propagate success back up search tree

Figure 18: Workflow of algorithmic synthesis design.

An expert approach is to synthesis design is to write down all the rules. The Chematica[129] and Synthia programs (see the talk by Hugo Viana at the end of this report) have tens of thousands of templates manually encoded by expert chemists. Reactivity conflicts encoded by hand are not scalable and cannot generalize (e.g., "do not use this rule if a peroxide is present") outside of the rule set. Connor regards such crafted knowledge approaches as the "first wave of AI", while statistical learning approaches are the second wave.

There are three core tasks in the workflow of algorithmic synthesis design (Figure 18): proposing a reaction, proposing the reaction conditions, and predicting the reaction outcome. In the first step, the input is a product and reactants are output. In the second, reactants and product are input and conditions are the output. In the third, reactants and conditions are input and a product is the output.

In retrosynthetic search the size of the candidate tree grows exponentially with the number of reaction steps. The search can be guided along promising branches by value functions (to determine which chemical is the most simple) and action policies (to determine which template is most useful). Training can be a regression task or a classification task. Kishimoto *et al.* have used Monte Carlo tree search (also used by Segler *et al.*[68]) and depth-first proof-number (DFPN) search in retrosynthetic analysis.[174]

Connor and his co-workers have addressed the search problem using deep reinforcement learning[175] to identify policies that make (near) optimal reaction choices during each step of retrosynthetic planning according to a user-defined cost metric. Using a simulated experience, they train a neural network to estimate the expected synthesis cost or value of any given molecule based on a representation of its molecular structure. They have shown that learned policies based on this value network can outperform a heuristic approach that favors symmetric disconnections when synthesizing unfamiliar molecules from available starting materials using the fewest number of reactions. The reinforcement learning agent explores transformations that are "feasible" according to the environment it explores, namely, simple template rules. Action policy pretraining improves rare template usage.[176] Training an "applicability" network that predicts whether a template pattern will match improves the accuracy of the "relevance" network that also predicts strategy (especially for small datasets).

The second core task is condition recommendation. Reaction conditions are necessary for experimental implementation; they can change the outcome of the reaction, qualitatively and quantitatively. The reaction is input as a reaction fingerprint and variables are predicted: in this case a continuous variable, temperature, and three discrete variables (reagents, catalysts, and solvents). Full condition combinations are too sparse to train on, but relatively few species are actually used: restricting the system to 803 catalysts, 2247 reagents, and 232 solvents excludes only 5% of reactions. Nielsen *et al.* have reported an alternative approach to reaction condition recommendation, aiming to predict from a substrate to one of a few bases, one of a few reagents, and so on.[177]

Reaction conditions are highly intercorrelated. The correlation among solvents, catalysts, reagents, and temperature was captured through a classifier chain of five classification tasks and one regression task. Gao *et al.* reported a neural-network model to predict the chemical context and temperature most suitable for any particular organic reaction.[178] Trained on about 10 million examples from Reaxys, the model is able to propose conditions where a close match to the recorded catalyst, solvent, and reagent is found within the top-10 predictions 69.6% of the time, with top-10 accuracies for individual species reaching 80–90%. Moreover "wrong" is often reasonable.

While no explicit relationship between chemical species is included in the model, it implicitly learns the functional similarity of solvents and reagents through training, as it can suggest similar chemicals for the same reaction. Taking solvent as an example, the similarity information can be extracted from the neural network, specifically the weight matrix in the last hidden layer before the final likelihood prediction and softmax activation. If two rows in the weight matrix are similar, the model will tend to predict similar scores for the corresponding two solvents. In other words, each solvent can be represented by its corresponding row from the weight matrix. This representation contains information about how it is used in different reactions and can be used to characterize functional similarity, implicitly averaged over all training reactions. This is analogous to the word embedding in natural language modeling where discrete words are converted to vectors of real numbers which contain similarity information (word2vec).

The third core task is outcome prediction. In addition to evaluating the likelihood of a reaction's success during synthesis planning, trying to predict reaction outcomes can be useful in automating the analysis of high-throughput mass spectrometry data, and in assisting in impurity identification in combination with structural elucidation. The problem is that there are many input-output pairs, but it is not known how to write an exact function to relate them. Researchers have access only to what was published (i.e., positive examples), and there is little variation in reported reaction yields.

Connor and his colleagues initially took a reaction template-based approach to this problem.[169] They adopted a data augmentation strategy where the "true" recorded example is supplemented with many "false" alternatives. In the first stage, the researchers applied a library of forward synthetic templates to define which products could be produced based on the initial reactants. The recorded product of a reaction in the database is the "true" product that the model learns to predict, while the chemically plausible alternative products generated *via* templates are the "false" products which were not reported in the literature. A model can then be trained to identify the "true" product as a multiway classification or ranking problem. A neural network learns the relative likelihoods of the products. In a 5-fold cross-validation, the trained model assigns the major product rank 1 in 71.8% of cases, rank $\leqslant$3 in 86.7% of cases, and rank $\leqslant$5 in 90.8% of cases. The method is thus very promising, but an accuracy of about 70% leaves room for improvement; coverage is incomplete because predictions cannot be made outside of the template scope; and the method is not fast enough because template application is a CPU bottleneck.

So the problem was reformulated as predicting graph edits (which bonds are broken, which are formed).[172] By training on hundreds of thousands of reaction precedents covering a broad range of reaction types from the patent literature, a graph-convolutional neural model makes informed predictions of chemical reactivity. The model predicts the major product correctly over 85% of the time, requiring around 100 ms per example, a significantly higher accuracy than achieved by previous machine learning approaches.

Starting from an attributed graph representation of molecules, the researchers iteratively update feature vectors describing each atom by incorporating neighboring atoms' information. After multiple iterations of this embedding, a local feature vector is calculated for each atom, based on its updated representation and those of its neighbors. To account for the effects of disconnected atoms such as reagents, a global attention mechanism produces a context vector for each atom as a learned, weighted combination of all other atoms' local features. Finally, a combination of local features and context vectors is used to predict the likelihood of bond changes for each pair of atoms.

The new model achieves expert-level performance. Connor and his co-workers asked 11 human participants to write the likely major products for 80 reaction examples from the test set. The 80 questions were divided into eight categories of 10 randomly-selected questions, each based on the rarity of the reaction template that could have been used to recover the true outcome. Only one chemist achieved a significantly better performance than the model; two chemists were approximately as successful. Related work has been reported.[179–181]

In the graph convolution process, attention scores indicate how strongly an atom's perceived reactivity depends on every other atom. Connor presented some examples of reactions analyzed and visualized through the global attention mechanism. For each reaction example, he selected one atom, highlighted in green, and colored all other atoms by the attention score (highlighted in blue), where a darker color indicates a stronger influence. The true reaction partner has strongest attention.

Thomas Struble and others working with Connor have also tried to learn selectivity for aromatic C-H functionalization.[182] Given empirical data on reaction yields, they aimed to learn site selectivity without presupposing any one mechanism. Historically, prediction of site selectivity for aromatic C-H bonds has focused on electrophilic aromatic substitution reactions where the mechanism is known. Struble *et al.*[182] have reported a generalizable approach to prediction of site selectivity that is accomplished using a graph-convolutional neural network (a Weisfeiler-Lehman network) for the multitask prediction of 123 C-H functionalization tasks. In an 80:10:10 training:validation:testing pseudotime split of about 58,000 aromatic C-H functionalization reactions from the Reaxys database, the model achieves a mean reciprocal rank of 92%. Once trained, inference requires approximately 200 ms per compound to provide quantitative likelihood scores for each task. In comparison, the RegioSQM[183] approach achieves a mean reciprocal rank of about 89% and takes many days rather than seconds.

The work of Connor's team and other researchers is enabling rapid ideation of full synthetic pathways. Software and hardware can now be brought together.[184] Millions of previously published reactions inform the computational design of synthetic routes, and expert-refined chemical recipe files (CRFs) are run on a robotic flow chemistry platform for scalable, reproducible synthesis. Suggested routes partially populate CRFs, which require additional details from chemist users to define residence times, stoichiometries, and concentrations that are compatible with continuous flow. To execute these syntheses, a robotic arm assembles modular process units (reactors and separators) into a continuous flow path according to the desired process configuration defined in the CRF.

This paradigm of flow chemistry development was demonstrated for a suite of 15 medicinally relevant small molecules. In another example, PRN1008 was a first-time disclosure from the ACS national meeting in spring 2018; a 13-step synthesis of this was produced in 60 seconds. This was not necessarily the shortest, cheapest, or highest-yielding pathway but was just one of many possible pathways for a chemist to review. A pathway was found in 6 seconds for the new disclosure BMS-986195, and in 15 seconds for LY3104607.

Connor's work is relevant to molecular generation. There is a trend to move away from precomputing numerical representations of molecules in order to predict physicochemical properties, and to move toward learning the properties[113] from structures. One class of techniques of growing interest for early-stage drug discovery is *de novo* molecular generation and optimization. These techniques can suggest novel molecular structures intended to maximize a multiobjective function, for example, suitability as a therapeutic against a particular target. It is now possible to check if the new molecules are synthesizable before passing the suggestions on to a synthetic chemist.[45,120,185]

The MIT team has hosted a number of computational tools to assist in synthetic planning and other aspects of organic chemistry in the Automated System for Knowledge-based, Continuous Organic Synthesis (ASKCOS).[186] Models can be built to describe patterns of chemical reactivity with relevance to synthesis planning and it is possible to learn to define areas of synthetically accessible chemical space. Chemists are actually using these tools and find value, but access to data is needed to drive algorithm development, and many, many challenges still remain.

21 Integrating AI with robust automated chemistry: AI-driven route design and automated reaction and route validation

Dr. Mario Latendresse, SRI Biosciences

SynFini is an automated, multistep synthesis platform (Figure 19) from virtual molecule to product. It allows rapid route design, validation and optimization in hours or days rather than weeks or months. Its high-fidelity digital synthesis drives efficiency, reproducibility, and transferability. Integrated "design, make, test" cycles bring AI and automation into one platform.

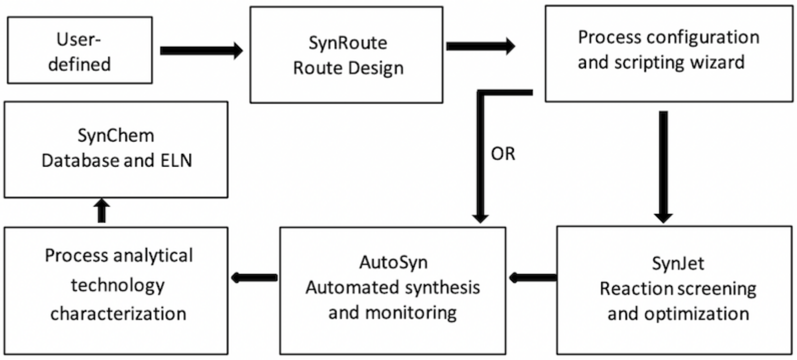

User-defined
SynRoute Route Design
Process configuration and scripting wizard
SynChem Database and ELN
OR
Process analytical technology characterization
AutoSyn Automated synthesis and monitoring
SynJet Reaction screening and optimization

Figure 19: SynFini web portal.

SynRoute for retrospective route design is the most complete component in the SynFini web portal (Figure 19). It uses a database of 17 million known reactions chosen from Reaxys, and proposes new reactions based on robust transformations. It designs multistep routes in minutes. The concept behind it was, firstly, to mimic the design process of a synthetic chemist to generate synthetic strategies using literature chemistry and reaction transformations, by combining AI (new reactions from known transformations) with results reported in reaction databases to produce synthetic routes. Secondly, the system had to prioritize the "best" synthetic strategies based on cost, likelihood of success, and ease of implementation.

To incorporate machine learning into route discovery, Mario and his colleagues created and trained classifiers for each transformation from the medicinal chemist's toolbox (MCT).[155] The goal was to predict whether a computer-generated reaction is workable. During a route search, all applicable MCT transformations are used to create new reactions. The ML classifier is applied to each computer-generated reaction and only reactions classified as "workable" are used in the search. The machine learning classifier reduces the exponential complexity of the search challenge and produces routes that have higher confidence overall.

Positive examples are defined as all reactions from Reaxys with a yield of greater than 20%. Negative examples are defined as reactions from Reaxys, with a yield of greater than 20%, where the reactants are applicable, but the reported product is from a different reaction type. The performance of a multilayer perceptron (MLP) was compared with that for random forest machine learning classifiers. One criterion was accuracy. Accuracy = (TP + TN) / (TP+FP+FN+TN), where TP = true positive, TN = true negative, FP = false positive, and FN = false negative. The average accuracy for MLP is 94.5% and for random forest is 89.5%.

Generating reactions and fast searching of routes uses a vectorized Dijkstra-like algorithm, as follows. Starting from the target compound, apply reaction transformations to generate reactions. Keep only the generated reactions with predicted likelihood of good "yield", based on machine learning classifiers. Repeat a maximum of t times until a connection to known compounds in Reaxys is found. Search for multiple k diversified routes based on optimality conditions and other restrictions.

The searches incorporate minimization of the combined costs of starting materials, solvents, reagents, and reaction implementation, while maximizing the overall yield of products to find routes that are efficient and cost-effective. All compounds and reactions connected to the target compound are tagged. A priority queue is then initiated with reactions that can proceed from feedstock: reactions can only proceed if all reactants are feedstock, and the reaction cost is calculated as the sum of feedstock divided by yield (when available). The reaction with minimum cost and process is considered the active reaction, and the next reaction is then processed: it is added to the priority queue and the procedure is repeated for each following reaction in turn, until the target is reached. The speed of the algorithm does not directly depend on the length of the route but on the number of reactions visited. The performance is almost linear on k, the number of top optimal cost routes found.

As an example, Mario showed the top route returned for imatinib, using 50,000 new reactions and the reaction classifiers: a three-step synthesis that involves one step from Reaxys and two generated new reactions. SynRoute recommends conditions for computer-generated reactions based on statistical analysis of reaction data.

Mario and his co-workers also downloaded all drugs from the FDA website and filtered them to get only single ingredient drugs, with a molecular weight of less than 1,500 Da, excluding gases and salts. They then searched for routes for all these drugs using, firstly Reaxys reactions alone and secondly Reaxys reactions plus computer-generated reactions. Ninety-three percent of the compounds were synthesizable in 5 steps or less. The majority of compounds with no routes found even with MCTs were large natural product compounds. The average search times for routes using MCTs were from 30.6 seconds for three-step syntheses up to 66.9 seconds for routes with 15 steps.

SynJet (Figure 20), for reaction screening and optimization, performs synthesis on a μg to mg scale. A customized inkjet dispenses about 1rxn/sec for a 10 μL reaction. There is highly parallel screening for varied conditions including continuous variables: (temperature, time, stoichiometry, pressure, and pH) and categoric variables (reagent, solvent, and substrate). In reaction building block screening, online analysis with direct analysis in real time mass spectrometry (DART MS) takes 5 seconds per reaction. Offline analysis with HPLC/MS takes 120 seconds per reaction. SynJet is coupled with SynRoute: results from multistep reactions augment the chemical database. It maps well to flow processes. A standard design-of-experiment (DoE) driven optimization process would normally take 1-2 weeks; with SynJet it takes less than a day.

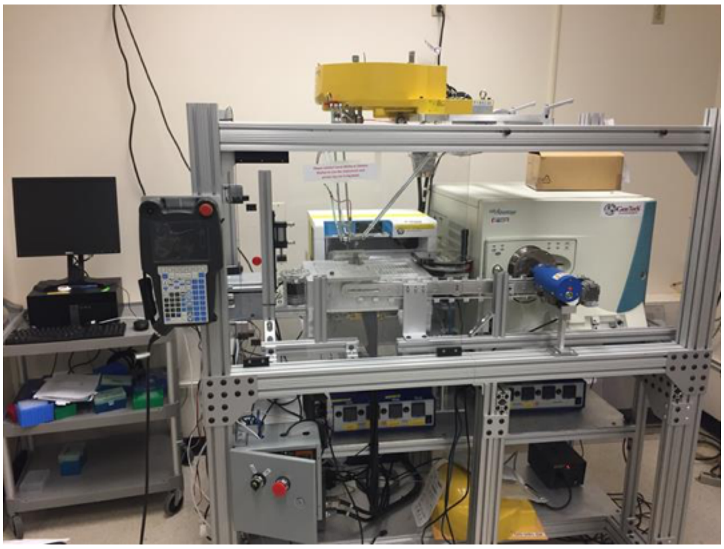


Figure 20: SynJet.

The AutoSyn automated bench chemistry platform (Figure 21) is a miniaturized, flow chemistry plant with integrated analytics. The “cityscape” in the figure is the miniaturized chemical plant. Automated synthesis includes start-up, operation, and shut-down. The apparatus carries out multistep synthesis on a mg-gm scale. It features the ability to switch between two targets in less than 2 hours, using valves to select the flow path. Characterization is both in-line and on-line. A “subway map” on the cityscape maps synthetic routes on the baseline configuration. The number of pathways possible is 3888 including different residence times.

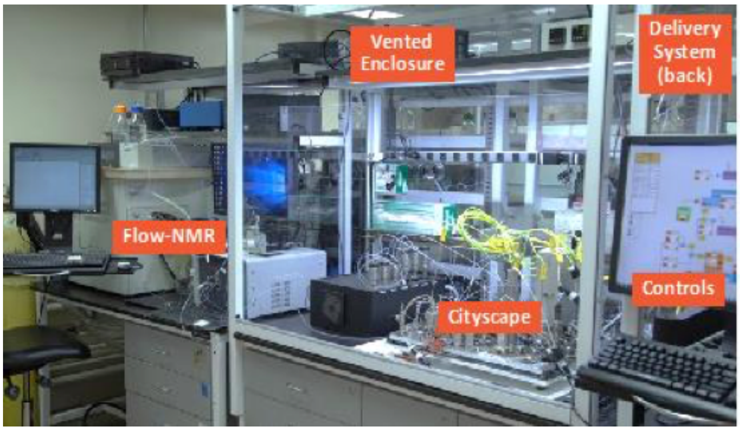


Figure 21: AutoSyn.

The SynChem database stores digitally captured and highly reproducible transformations. Twenty-two of the MCT transformations have validated examples on AutoSyn. The database includes successes and failures, and optimization data. Reactions are added to the SynRoute reaction database for more efficient route design.

SynFini was validated by multistep synthesis of target compounds on AutoSyn, with integrated separation and analysis, producing new synthetic routes to 22 known drugs. There were from two to five steps per route. (The current capacity of the cityscape configuration is about seven steps.) It typically takes less than two weeks to validate a preliminary route. Increasingly complex, chiral routes, longer syntheses, and convergent methods are then carried out. Crude purity is greater than 80%. Purified purity is greater than 95% across all routes. Examples of routes from SynRoute that have been performed on automated chemistry platforms demonstrate the potential for a continuously improving, data-driven synthetic planning platform.

22 A nondeterministic Chemputer for running chemical programs

Professor Lee Cronin, University of Glasgow

Research in the Cronin Group[187] is motivated by the fascination for complex chemical systems, and the desire to construct complex functional molecular architectures that are not based on biologically derived building blocks. Lee showed a short video of some of the work done by his team ("Making Stuff with a Matter Operating System") and went on to describe the history and context of the "Chemputer".

What does chemical space look like? How can we define and then search the space? How does complexity arise? *De novo* target identification and studying complexity depend on defining and searching the space. The labor involved in synthesizing big molecules creates a bottleneck that the Chemputer can remove.[188] The aim is to automate boring, routine processes and improve search efficiency. Lee's group are trying to navigate through the vastness of chemical space using the Chemputer, and are studying how this technique can be used not only to find promising cures for diseases, but also to understand the transition from chemistry to biology, and how to create artificial life in the lab.

Applying robotic technologies along with artificial intelligence-based approaches to chemistry requires a holistic approach to chemical synthesis design and execution. Lee's team has outlined a universal approach to this problem, chemputing, which exploits the concept of the Chemputer.[189] The Chemputer refers to a modular system for chemistry that can understand the representation of a practice of chemical synthesis that then informs the programming and automation required for its practical realization. Using this foundation to construct closed-loop robotic chemical search engines, they can generate new discoveries that may be verified, optimized, and repeated entirely automatically. The robots can perform chemical reactions and analyses much faster than can be done manually. This leads to a road map whereby molecules can be discovered, optimized, and made on demand from a digital code.

A few well-defined areas such as polypeptide and oligonucleotide chemistry have automated chemical synthesis but laboratory-scale and discovery-scale syntheses have remained predominantly a manual process. Recent advances in areas such as flow chemistry, oligosaccharide synthesis, and iterative cross-coupling are expanding the number of compounds synthesized by automated methods, but there is no universal and interoperable standard that allows the automation of chemical synthesis more generally.

In developing the Chemputer platform,[190] Lee's team wanted to build on 200 years of chemical literature, and the experience of thousands of bench chemists, in a way that would lead to a standardization embodied in a codified standard recipe for molecular synthesis. For this to be possible, it was essential that the approach mirror the way the bench chemist works. Lee's team therefore used the round-bottomed flask as the primary reactor for batch synthesis. A relatively small array of equipment was assembled to accomplish a wide variety of different syntheses, and the abstraction of chemical synthesis encompasses the four stages: reaction, workup, isolation, and purification (see a video[188] on the Glasgow website). With these four modules, it was possible to automate the synthesis of the pharmaceutical compounds diphenhydramine hydrochloride, rufinamide, and sildenafil without human interaction, in yields comparable to those achieved in traditional manual syntheses.

The standardized format for reporting a chemical synthesis procedure, coupled with an abstraction and formalism linking the synthesis to physical operations of an automated robotic platform, yields a universal approach to a chemical programming language. The architecture and abstraction form the Chemputer. Chemify[191] is a project dedicated to this new era of chemical synthesis driven using a universal language developed to make molecules more accessible, cheaply and safely, as well as reducing labor and expanding chemical space in terms of the number of molecules that can be made. Instructions on how to build a Chemputer and code for syntheses that have been successfully performed on Chemify platforms are available on the website. A video of the equipment[191] is also available. A 12-step convergent synthesis in a Chemputer has been demonstrated, with automated diazirine coupling, cleavage, precipitation, and filtration.

In addition to its synthesis modules, the Chemputer has sensors for real time feedback for pH, spectral data, conductivity, computer vision, chromatography, and thermal and acoustic imaging. It is currently possible to operate pumps and valves live using the application; other device controllers have been written, but need testing and debugging. Remote control is particularly useful for hazardous procedures. The sensor system connected to the Chemputer Schlenk line control system is controlled by ethernet cable which means that the system can also do optimization of the chemputing process.

Research into the Chemputer has received funding of about 20 million pounds sterling from the Defense Advanced Research Projects Agency (DARPA), the United Kingdom and the

European Union, over five years, and more than 50 organizations are interested in it. Beta versions have been built outside of Glasgow, for example by GSK and BAM[192] in Berlin. Using the Chemputer (at Glasgow University), and exploring chemical space using statistical methods (at Arizona State University), two teams have won a challenge prize for innovative solutions to pain, and opioid use disorder and overdose.[193] The aim of the Chemputer paradigm is not to replace the chemist but to do as much manual bench chemistry as possible robotically. This will allow teams to share code, discovery and optimization data, and give teams access to new molecules quickly ensuring quality and reproducibility. Lee therefore now wants feedback on the 100 top molecules that are really annoying to make. He plans to get people to work together to make the Chemputer code for those molecules and validate them. This project is called Chemify100.

The Chemputer paradigm and chemputing is now opening a new window on drug discovery. Only 200 million molecules are known, but 10^{100} could possibly be synthesized in up to 10 steps. Discovery is limited by the number of steps; machine learning, and using artificial intelligence systems in general, is hard without a standard way of generating the data, and this is one of the key things that Chemputers can do. Intelligent automated chemistry platforms for discovery orientated tasks need to be able to cope with the unknown, which is a profoundly hard problem. Henson *et al.* have described how recent advances in the design and application of algorithms, coupled with the increased amount of chemical data available, and automation and control systems may allow more productive chemical research and the development of chemical robots able to target discovery.[194]

Lee's team have reported an organic synthesis robot that can perform chemical reactions and analysis faster than they can be performed manually, as well as predict the reactivity of possible reagent combinations after conducting a small number of experiments, thus effectively navigating chemical reaction space.[195] By using machine learning for decision making, enabled by binary encoding of the chemical inputs, the reactions can be assessed in real time using NMR and IR spectroscopy. The machine learning system was able to predict the reactivity of about 1000 reaction combinations with accuracy greater than 80 percent after considering the outcomes of slightly over 10 percent of the dataset. This approach was also used to calculate the reactivity of published datasets. Further, by using real-time data from the robot, these predictions were followed up manually by a chemist, leading to the discovery of four reactions.

Lee presented the autonomous "Chemputer Discovery Machine": closed loop reactivity discovery with a process language. Liquid handling and automated analysis connect to continuous reactivity assessment for a reagent dataset, after which chemical space modeling can be carried out, before the loop is closed and liquid handling and automated analysis are repeated. Lee thinks that a good use of convolutional neural networks is to estimate reactivity. The chemical code, or Chemical Markup Language (XDL), can be used to constrain inputs, and the loop can be closed. This is reaction discovery beyond using rules found in the organic chemistry literature. This is because new heuristics are discovered, and rules can be applied in the conventional way.

Finally Lee outlined a chemical search engine for the origin of life (OOL): a transition from chemistry to biology, to create artificial life in the laboratory (see Figure 22). A new approach to finding and targeting molecules is the "messy (primordial) soups" bottom-up approach, as opposed to the targeted synthesis top-down approach. Organic chemists are "creationists", but LIFE is an evolutionary approach.

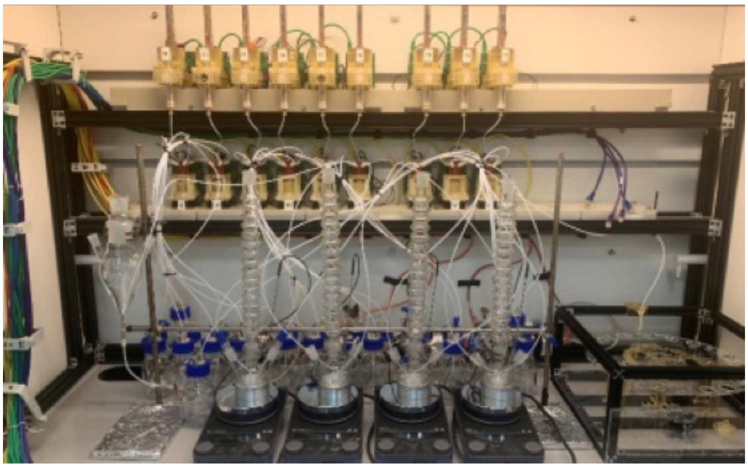


Figure 22: A photograph of the OOL-Chemputer aimed at searching chemical soups for complex molecules that can go on to establish an evolvable chemical system.

In object assembly theory, complex molecules are impossible without LIFE. With object assembly theory, molecules are identified as biosignatures. Lee's team has reported a new type of complexity measure, called "Assembly Complexity", (and as applied to chemistry, "Molecular Assembly Complexity"), that allows them not only to find a threshold to the abiotic-biotic divide, but also to demonstrate a probabilistic approach based on object abundance and complexity which can be used to unambiguously assign complex objects as biosignatures.[196,197] They hope that this approach will not only open up the search for biosignatures beyond the Earth, but also allow them to explore the Earth for new types of biology, and to determine when a complex chemical system discovered in the laboratory could be considered alive.

23 Data-driven exploration of the catalytic reductive amination reaction

Dr. Benjamin Deadman, Imperial College – ROAR

Over a century of synthesis has yielded a vast library of reactions but there is a pressing need for more comprehensive datasets which include negative results, multiple time-point reaction data, and interoperable synthesis procedures. The center for Rapid Online Analysis of Reactions (ROAR) at Imperial College[5] is a new facility which brings together high-throughput (HT) batch and flow reactor platforms, *in situ* analytic technologies, and automation expertise to enable data-centric research in synthesis. ROAR is a facility not a research unit; it exists to support the research projects of others. Usage of the facility is currently subsidized for academic research on the condition that the results are made available on an open access data repository after an embargo period. Some commercial usage of the facility, without a requirement to make the data open access, is possible for a fee.

Ben presented an exploration of the catalytic reductive amination reaction using the ROAR facility. Reductive amination is an important reaction for many industries.[198–201] It is a simple reaction with multiple variables which can be optimized. In the ROAR approach, HT robotic batch reactor platforms are used to screen a range of heterogeneous catalysts for activity and selectivity.

There is a need to perform kinetic studies of pressurized reactions but a technical challenge is sampling pressurized reactions in a controlled manner without venting the reactor. The solution is the Unchained Labs Optimization Sampling Reactor (OSR) unit which has antechamber system for sampling under pressure. The eight-reactor OSR unit enables scientists to screen a broad experimental space with precise and independent pressure and temperature control for each reactor.

In a time-points study of the reaction in Figure 23, it was found that conversion and yield of the target secondary amine reached a maximum after about 30 minutes, and then the yield fell away over time. Temperature optimization of the first step showed that this is a fast reaction (greater than 95% conversion in 5 minutes), and low temperatures (25 °C to 40 °C) hinder product decomposition.

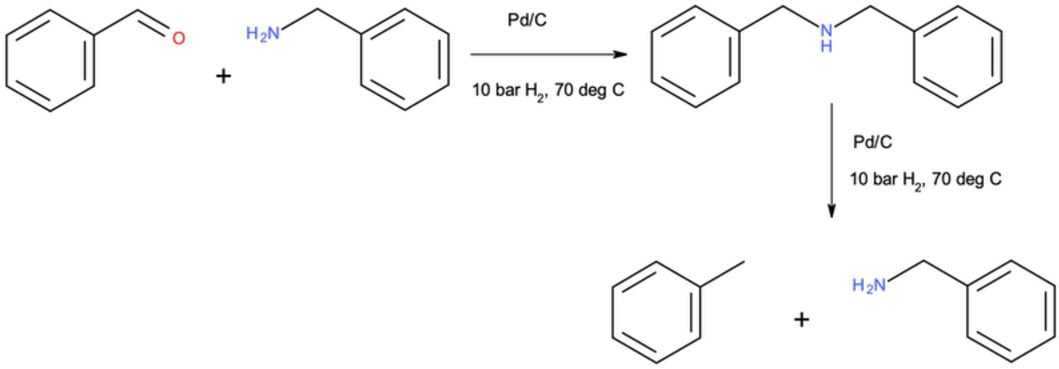


Figure 23: Reductive amination.

Substrate screening (varying the aldehyde or ketone and the amine) was carried out on an Unchained Labs Deck Screening Pressure Reactor (DSPR). The DSPR allows up to 48 reactions to be performed in parallel with a common pressurized gas in the headspace. Ben presented the results of the substrate screening array, with color coding to categorize the combinations of carbonyl compounds and amines as low-, medium-, and high-yielding for the target amine. The team are continuing to develop this screening technique as a tool for mapping out the reaction space for this particular class of reductive aminations.

Ben discussed some of the aspects of planning HT reactions which ROAR have found to be particularly important. In their typical planning for a HT process they will always consider the analysis first, since a robust analysis method is essential to obtaining quality data. What is the minimum sampling volume required for the intended analysis method(s)? What state (liquid, biphasic etc.) will the reaction be in when sampling takes place? Another concern is the chemical sources: in which order to add them, and in which state, their solubility, and the preparation of any stock solutions. After that, what is the minimum dispense volume and weight? Is the reaction reproducible at the intended scales, and across the plate?

The reductive amination reaction is the first system to be studied comprehensively in ROAR, but these techniques will be applied more widely in the ambition to develop quality datasets for other chemical transformations, including negative results, multiple time-point reaction data, and interoperable synthesis procedures.

24 Machine-assisted flow chemistry for organic synthesis

Dr. Christopher A. Hone, Research Center Pharmaceutical Engineering (RCPE)

Chris works within the Center for Continuous Flow Synthesis and Processing (CC FLOW) at RCPE. RCPE is a nonprofit spin-out company from the universities in Graz, Austria. In his talk, Chris summarized efforts in his laboratory to use automation and computational methods for the development of flow chemistry processes. RCPE's modular flow platform for real-time control of the synthesis of active pharmaceutical ingredients using model-based strategies is shown in Figure 24.

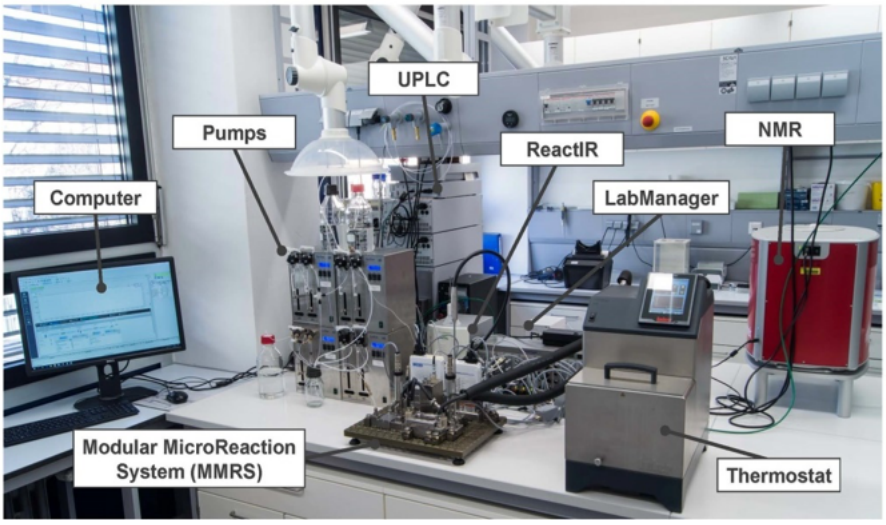


Figure 24: Modular flow platform.

RCPE used the Ehrfeld Mikrotechnik Modular MicroReaction System and the Lonza FlowPlate microreactor. In particular, Chris discussed the coupling of the modular flow platform with real-time analysis by IR and NMR, and online UPLC, for the efficient optimization of a multistep organometallic transformation without the need for human intervention.[202] The case study concerned alpha- functionalization of carbonyl compounds *via* lithium enolate intermediates Figure 25. This is hazardous chemistry, involving an exothermic reaction. It is also challenging because of solid formation and sensitivity to moisture.

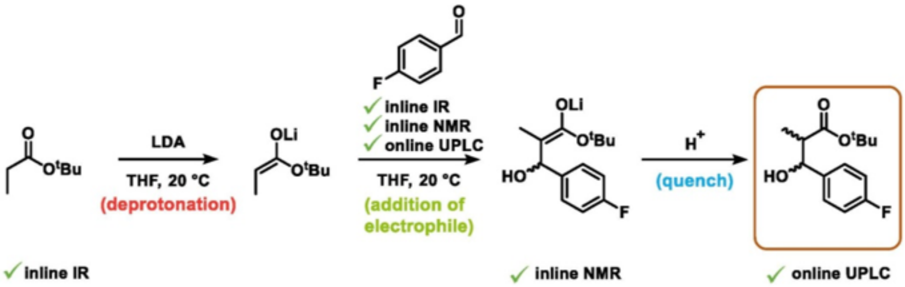


Figure 25: Analytical strategy for lithium enolate reaction.

The team rapidly generated experimental data (17 iterations in under 2 hours) to access information on the different chemical species at multiple points within the reactor and to generate process understanding. The optimized continuous flow conditions were demonstrated in a scale-out experiment with in-process monitoring to afford the desired product in 70% isolated yield with a throughput of 4.2 g h^{-1}.

Chris presented a further case study involving the generation of process models for an aerobic oxidation operating within a segmented flow regime.[203] Continuous flow reactors have facilitated safer and more efficient utilization of O_2, whilst enabling protocols to be scalable. Chris studied the aerobic oxidation of diphenyl sulfide to diphenyl sulfoxide. The model-based experimentation process involved the following steps: perform experiments; perform experiments and simulations important for scalability (pulse experiments, mass transfer, calorimetry); recreate the experiment *in silico*; estimate kinetic parameters; simulate, optimize and design *in silico* batch and continuous processing; validate the experiment; optimize and scale up. The process models generated were underpinned with a residence time distribution (RTD) study and computational fluid dynamics (CFD) simulation. gPROMS FormulatedProducts[204] software was used for the *in silico* modeling to identify the optimal operating conditions.

25 Encoding solvents and product outcomes to improve reaction prediction systems

Dr. Ella M. Gale, University of Bristol

Ella is in the technology-enhanced chemical synthesis center for doctoral training at Bristol (TECS-CDT). Students are taught Python, supervised and nonsupervised learning, etc. and then they have access to a ChemSpeed automated synthesis robot to optimize reactions. Synthetic outcomes are hugely dependent on the conditions used, like the choice of solvent or temperature, but most retrosynthetic algorithms do not encode this information into the input data. The output data are usually encoded in a binary way, as to whether a particular desired output chemical is present or not.

Ella has worked on expanding the complexity of both input and output data to make these algorithms more practically useful. Categorical data are variables that contain label values rather than numeric values; for example a solvent label may be “water” or “ethanol”. Categorical data can be converted into numerical data by integer encoding (each label is assigned a number) and one-hot encoding. One-hot encoding is a vector representation where all the elements of the vector are 0 except one, which has 1 as its value. Ella inputs the solvents with a one-hot code referring to a database of around 600 popular and commercially relevant solvents, and temperature is input as being within ranges. The reaction output products are coded in a trinary way: the value 2 if the product is present, and is either the top-most reported product or in a high enough yield to be synthetically useful, 1 if the product is present but in a low yield, and 0 if the product is naively possible but not seen in the laboratory.

Other types of solvent data include a formula (SMILES in this case), and “bag-of-words’ style nonhierarchical descriptors of functional groups and molecule types (from Marcus Johns). This new chemical group database has 167 parameters, for example, alkane, ketone, sulfide, and aromatic. There are also physicochemical descriptors: out of 634 solvents,[2] 286 were used, with more than 24 parameters, including boiling point, molar volume, surface tension, and Catalan,[205] Laurence[206] and Hansen’s solvent parameters.[207] Finally there is principal components analysis (PCA) of physicochemical values.[2] Ella has done previous work[208] searching such databases of solvent properties.

In earlier work, Gao *et al.*[178] have developed a neural-network model to predict the catalyst(s), solvent(s), and reagent(s), and the temperature most suitable for any particular organic reaction. The model is able to propose conditions where a close match to the recorded catalyst, solvent, and reagent is found within the top-10 predictions 69.6% of the time, with top-10 accuracies for individual species reaching 80–90%. The researchers also demonstrated that the model implicitly learns a continuous numerical embedding of solvent and reagent species that captures their functional similarity.

While no explicit relationship between chemical species is included in the model, it implicitly learns the functional similarity of solvents and reagents through training, as it can suggest similar chemicals for the same reaction. Taking solvent as an example, the similarity information can be extracted from the neural network, specifically the weight matrix in the last hidden layer before the final likelihood prediction and softmax activation. If two rows in the weight matrix are similar, the model will tend to predict similar scores for the corresponding two solvents. In other words, each solvent can be represented by its corresponding row from the weight matrix. Gao and his co-workers[178] refer to this as “solvent embedding”. To visualize the embedding of solvents, the top 50 solvents with the highest frequency in the dataset of 232 were selected, and labeled manually into four types: nonpolar, polar nonprotic, protic, and halogenated.

Ella investigated this solvent embedding reaction space to see which solvent data embedding it is closest too (because if she found which data the neural network is learning from reaction space, she could just give that data to the network). She investigated the latent space encoding of data (Figure 26).

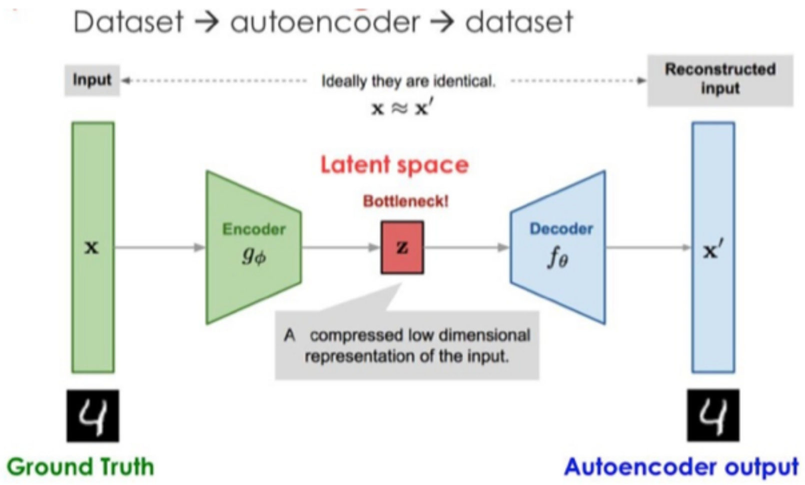
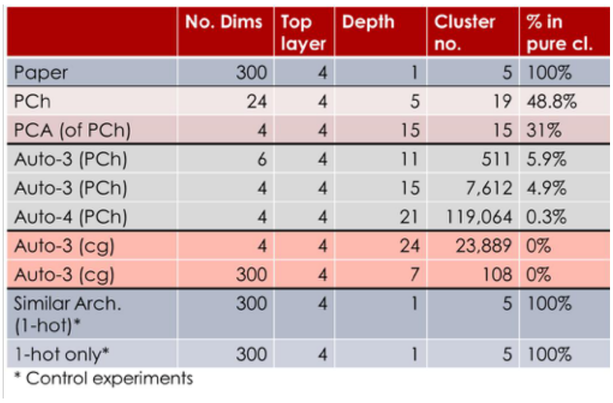

Dataset → autoencoder → dataset
Input
Ideally they are identical.
$\mathbf{x} \approx \mathbf{x}'$
Reconstructed input
Latent space
Bottleneck!
x
Encoder
g_ϕ
z
Decoder
f_θ
x′
A compressed low dimensional representation of the input.
Ground Truth
Autoencoder output

Figure 26: Latent space encoding of data.

The data types input were physicochemical data (pCh), comprising 24 parameters for 287 solvents, PCA of pCh, and one-hot (control). Derived latent space encoding types were PCA-auto-PCA, pCh-auto-pCh, fg-auto-fg, and 1-hot-auto-1-hot (control). These abbreviations refer to the data put into the autoencoder and the output. Thus “PCA-auto-PCA” means PCA input and output. “Fg” is “functional group” (from the dataset mentioned earlier). The abbreviation “1-hot-auto-1-hot” means running an autoencoder on 1-hot data. This is the control as it contains no information about the solvents. The four chemical type labels of Gao *et al.* were used but were not used in the autoencoder.

Now Ella looked for clusters in 2D projections. In the first experiment, *t*-distributed stochastic neighbor embedding (t-SNE), multi-dimensional scaling (MDS), and isomap (ISO) were used for dimensionality reduction. With t-SNE every scatter plot looked much the same. MDS and ISO are preferred to t-SNE as they try to preserve a real distance. MDS tries to preserve the distance, and ISO tries to preserve the geodesic distance between the points, and so Ella used them to check that there was not a real cluster in the data.

In a second experiment, Ella used a new ClusterFlow algorithm, which carries out semisupervised clustering of multidimensional data into hypercuboids, using labels. She clustered the 50 solvents and the four labels of Gao *et al.* in 300-dimensional space, with the results shown in Table 5.



	No. Dims	Top layer	Depth	Cluster no.	% in pure cl.
Paper	300	4	1	5	100%
PCh	24	4	5	19	48.8%
PCA (of PCh)	4	4	15	15	31%
Auto-3 (PCh)	6	4	11	511	5.9%
Auto-3 (PCh)	4	4	15	7,612	4.9%
Auto-4 (PCh)	4	4	21	119,064	0.3%
Auto-3 (cg)	4	4	24	23,889	0%
Auto-3 (cg)	300	4	7	108	0%
Similar Arch. (1-hot)*	300	4	1	5	100%
1-hot only*	300	4	1	5	100%

* Control experiments

Table 5: Clustering of solvents.

Key to Table 5:

- No. Dims = number of dimensions of the latent variables (i.e., the lengths of the latent variable vectors)
- Top layer = the number of clusters at the top layer of the hierarchical cluster
- Depth = depth of the hierarchical cluster
- Cluster no. = the number of clusters in the hierarchical cluster. (The clustering algorithm makes clusters of clusters until it has assigned all the points to a correctly labeled cluster.)
- % in pure cluster = the number of points that were assigned to a cluster which contained points with only one label
- Paper = the results from the actual data in Gao *et al.*'s paper
- PCh = clustering just the original physicochemical properties
- PCA of PCh = clustering the first four principal components of a principal component transform of the physicochemical data
- Auto-3 PCh = clustering of the latent variables of a three-layer autoencoder trained on the physicochemical properties
- Auto-4 PCh =clustering of the latent variables of a four-layer autoencoder trained on the physicochemical properties
- Auto-3 cg = clustering of the latent variables of a three-layer autoencoder trained on the functional group labels
- Similar arch (1-hot) = clustering of the latent variables trained on 1-hot using an architecture of the same size and shape as in Gao *et al.* paper (control as 1-hot is completely random)
- 1-hot only = clustering of 1-hot labels (a control as 1-hot is completely random, so any clusters seen are illusionary)

As the control experiment clusters matched those seen in the Gao paper, it seems that Gao *et al.* did not really find clustering in their solvent embedding. The human eye is easily tricked and sees patterns everywhere, which is why it is important to do the control experiments. From all this work Ella's team finds that the solvent properties are not really needed for the retrosynthesis algorithms.

Thus, Ella concluded that specific solvents do not matter for retrosynthesis (with the proviso that the solvent must dissolve the reactants and products). Solvent selection and optimization is best done as a separate task, so it will be taught to the students as a separate step. The output from Ella's expanded retrosynthesis algorithms can be used as a guide to the inputs to DoE programs on the ChemSpeed robot, allowing for the fast optimization of reaction conditions. Moreover, since the physicochemical properties of the solvents in the database are known, this information can be used to search for greener and cheaper alternatives to test.

26 Evolutionary computing strategies and feedback control for directed execution and optimization of chemical reactions

Professor Harris Makatsoris, King's College, London

The evolution of a chemical system towards a desired property within a fitness landscape is a very attractive strategy for reaction optimization experimentally. It allows efficient exploration within complex parameter spaces that may contain multiple maxima or minima but without any detailed knowledge of the structure of the space. Unlike other approaches, it does not impose a requirement to collect information necessary to calculate gradients. Experimental design approaches that have been recently reported have demonstrated good performance, but they employ search strategies along a single steepest ascent or descent pathway with some cases requiring gradient calculation. This prevents them from discovering better designs within complex spaces since they get trapped very quickly within a particular region of the space, because they explore around only a single extremum. In contrast, evolutionary approaches avoid this as they sample points from across the whole search space. Furthermore, evolutionary strategies are robust and resilient to experimental and measurement errors and can be applied in manual or fully automated experimental scenarios.

A different search space requires rapid and efficient execution, and directed control. Centillion[209–213] is an innovative approach that meets these requirements. It is a new flow chemistry system of precision-engineered modules that when connected, enable chemical, biochemical, and bioprocessing companies to create productive and sustainable continuous processes to occur in the laboratory and at production scale. With these, the control of fluids in innovative ways is possible allowing repeatable, productive, and sustainable product manufacture. The mission of Centillion Technology Ltd. is to digitize this experience, liberate creativity, and change how industry develops and scales materials.

Harris and his co-workers have developed a thermoacoustic heat engine (TAHE),[214] a type of prime mover that converts thermal power to acoustic power. Although the geometry of the TAHE is simple, the behavior of the engine is complex with over 30 design parameters that affect the performance of the device; therefore, designing such a device remains a significant challenge. The researchers have reported a methodology using reinforcement learning (RL) for the design and optimization of a TAHE.[215] It is eventually hoped that with increased understanding of the design problem, in terms of the RL framework, it will be possible to ultimately create an autonomous RL agent for the design and optimization of complex TAHEs with minimal predefined conditions and restrictions. Harris envisions this as a move toward totally automated optimization (Figure 27).

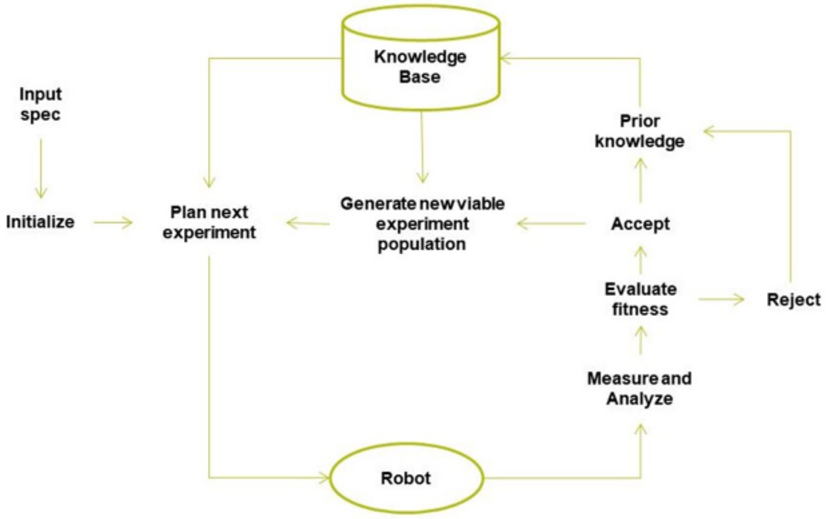


Figure 27: Toward totally automated optimization.

Flow chemistry is inherently scalable, in terms of intensification, and precise control of the factors that influence a process, allowing repeatable and operator-independent production. There is a minimal mixture inventory and equipment footprint. Flow chemistry is scalable, repeatable and transferable with model support.

DoE and response surface methodology (RSM) are popular techniques developed by the automotive industry. RSM leads to a model that can be used with optimization approaches,[216] but it is based on Taylor Series approximation of a function in a local region, and needs screening experiments first to identify the region. It applies local search.

Harris took as an example the synthesis of [Fe(Htrz)2(trz)](BF4): a precipitation reaction with particle engineering opportunities.[217] Batch synthesis of this compound is by combination of two solutions $Fe(BF_4)_2$ and triazole (Htrz) at room temperature. This results in a range of particle sizes. Synthesis in flow format yielded a narrow particle size distribution compared to batch. Mixing and RTD control lead to different morphologies. Control of the process is critical for obtaining particles with narrow particle size distribution, especially as scale increases. Harris showed scanning electron microscope results for targeted particle property optimization and scale-up with a model based approach, proving the advantages of Centillion.

For local optimization, the Nelder-Mead[218] method is a heuristic technique, which uses a "geometrical shape" called a simplex as a template that proceeds within a region, descending or ascending towards a local best fit. At every iteration, it proceeds to reshape or move this simplex, one vertex at a time. It aims to improve the best value by adjusting the template. Initial screening and defining the template is the starting point. The approach entails changing over time one or more "traits" (e.g. process parameter space) in a population of options by mutating and reproducing those available and selecting those that not only enhance the fitness of the population by some selected measure ("dialed" property) but also remain in successive generations (survive changes).

Evolutionary algorithms have many advantages. They are capable of operating in unstructured spaces with large or even undefined number of parameters; are easy to develop; are robust and suitable for hardware in the loop; and are stable in the presence of noise in a fitness value. Global search avoids getting trapped in local regions.

Harris's team is combining approaches. In the top layer, an evolutionary algorithm samples the space, assesses fitness, and controls selection. In the bottom Layer, heuristic local search explores the space in detail by anticipating the next best move. Automated technique platforms rely on feedback mechanisms that require the integration of process analytical tools and methodologies to preprocess the data from observables before determining the next experimental design of the next iteration. Harris demonstrated the application of these techniques with the use of a fully automated flow system.

As an example, he took the oxidation of epinephrine (Figure 28). His team optimized the metastable trihydroxyindole or the final product, (oxidized trihydroxyindole), with synthesis on demand using evolutionary algorithms. Epinephrine and the final oxidized trihydroxyindole product are nonfluorescent. The trihydroxyindole intermediate is fluorescent. The team maximized fluorescence to obtain the intermediate or eliminated fluorescence to obtain the final product, using self-optimization within Centillion to obtain the conditions.

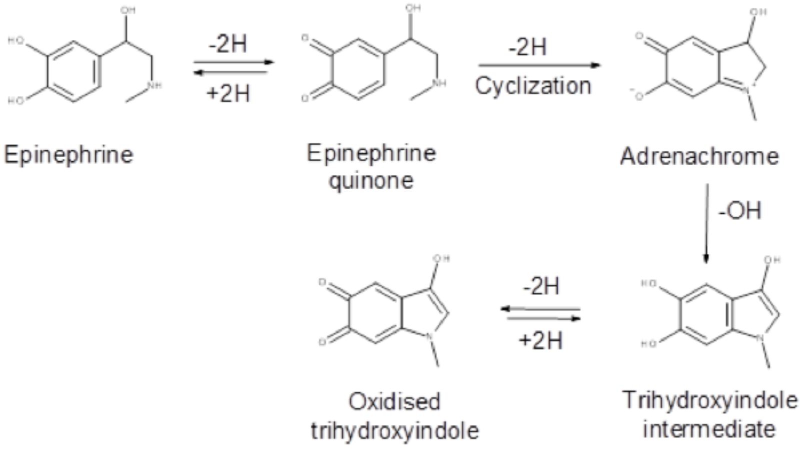


Figure 28: Oxidation of epinephrine.

Another example concerned gold nanoparticles in segmented flow. Harris outlined the spectroscopic regression model used for spectral feature extraction in online monitoring. A key recent application is the development of smart microreactor technology for enzymatic reactions for the future manufacture of vaccines. Harris's team is involved in the Engineering and Physical Sciences Research Council (EPSRC) Vaccines Manufacturing Hub ("Factory in a Box"), working on rapid process design and scale-up for RNA vaccines.

27 Computational design *via* metal-driven self-assembly: from molecular building blocks to emerging functional materials

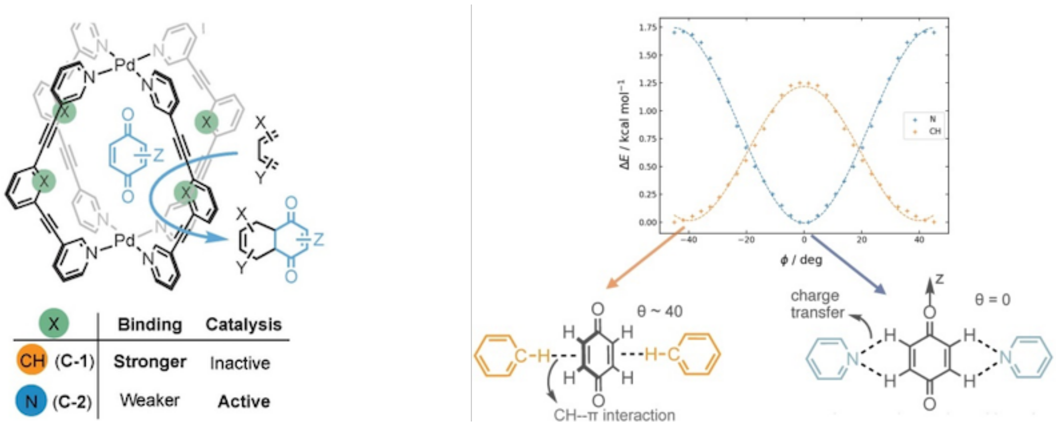
Professor Fernanda Duarte, University of Oxford

Current computational methods enable chemists to interrogate chemical processes at the molecular level. Despite these advances, several challenges remain when exploring unusual reactivity or targeting novel catalysts. Among them are the accurate description of both electronic and energetic properties; the efficient modeling of structurally dynamic systems; and the efficient evaluation of novel catalysts.

Self-assembled coordination cages based on organic ligands and a variety of transition metal cations have emerged as promising supramolecular materials thanks to their applicability in fields such as drug delivery, molecular recognition and catalysis. Guest molecules can reside within these cavities and much of the interest in these systems is derived from these fascinating host-guest interactions.[219–223] Back in 1998, Sanders[224] wondered why among the outpouring of new supramolecular arrays there are so few effective catalysts. The question could still be asked today. Examples include cyclodextrins,[225] a resorcin[4]arene-based capsule,[226] tetrahedral assemblies of stoichiometry M_4L_6 such as $[Ga_4L_6]12-$ capsules,[227–229] and Pd_2L_4 metallocages.[230]

Computational modeling of metal-driven self-assembly has been carried out with both quantum mechanics (DFT) and molecular mechanics (molecular dynamics, MD). Fernanda believes that to get meaningful answers you must use both. Fernanda has worked with Paul Lusby's group in Edinburgh on Pd_2L_4 catalysis.[230–232]

The Diels-Alder reaction is a cornerstone of synthesis, yet nature does not use catalysts for intermolecular [4+2] cycloadditions. Attempts to create artificial "Diels-Alderases" have also met with limited success, plagued by product inhibition. Using a simple Pd_2L_4 capsule (Figure 29A), Marti-Centelles *et al.*[231] have shown Diels-Alder catalysis that combines efficient turnover alongside enzyme-like hallmarks. This includes excellent activity ($k_{cat}/k_{uncat} > 103$), selective transition-state stabilization comparable to the most proficient Diels-Alder catalytic antibodies, and control over regioselectivity and chemoselectivity that would otherwise be difficult to achieve using small-molecule catalysts. Unlike other catalytic approaches that use synthetic capsules, this method is not defined by entropic effects; instead multiple H-bonding interactions modulate reactivity, reminiscent of enzymatic action.



A

B

Figure 29: Pd_2L_4 capsule binding.

Three approaches have been taken to force field parameters for molecular dynamics simulations: the soft-sphere model,[233–235] the bonded model,[236–238] and the dummy model (Figure 30).[239–242] The cationic dummy atom approach provides a powerful nonbonded description for a range of alkaline-earth and transition-metal centers, capturing both structural and electrostatic effects. Duarte *et al.*[239] have refined existing literature parameters for some metals and shown that they are easily transferable to any force field that describes nonbonded interactions using Coulomb and Lennard-Jones potentials. They provide a valuable resource for the molecular simulation community, as they extend the range of metal ions that can be studied using classical approaches, while also providing a starting point for subsequent parametrization of new metal centers.

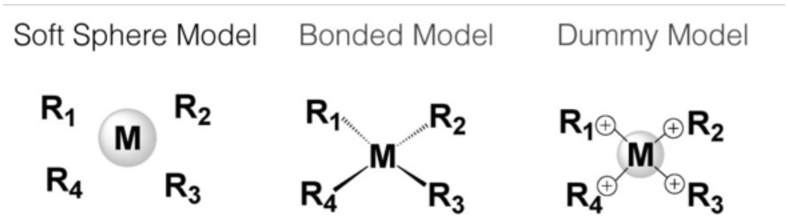


Figure 30: Modeling metal-ligand interactions.

Employing the dummy model approach[239,243] in combination with classical MD techniques, Fernanda's team studied the flexibility of the Pd_2L_4 cage. They showed that a simple and efficient DFT-based methodology, informed by explicitly solvated molecular dynamics and coupled cluster calculations, is sufficient to reproduce experimental guest binding affinities accurately (Figure 29B) and identify the catalytic Diels-Alder proficiencies (>80% accuracy) of two homologous Pd_2L_4 metallocages with a variety of substrates.[230] This analysis revealed how subtle structural differences in the cage framework affect binding and catalysis. These effects manifest in a smaller distortion and more favorable interaction energy for the catalytic cage compared to the inactive structure.

In summary, in this study of the effect of noncovalent interactions and flexibility on biomimetic catalysis, electronic activation is observed for both cages, but significant distortion energy hinders catalysis. Note that Sanders[224] concluded: "I have suggested that the fear of entropy has taken supramolecular chemists too far in the direction of rigidity and preorganization, and that the future may lie in more flexible systems that rely on noncovalent interactions to impose order on three-dimensional structure."

The Duarte group has applied their self-assembly design approach in an open source tool,[244] cgbind,[245] to generate and analyze metallocage structures. While related tools exist for COF and MOFs,[246–249] no similar tool is currently available for metallocages. The cgbind tool is still under development and is currently employed for the discovery of new catalytic palladium metallocages.

Finally, the group has also developed a computational tool to automatize the exploration of reaction profiles, autodE.[250] In contrast to available open-ended search approaches,[251–256] autodE combines graph theory and chemical knowledge in order to reduce the size of the chemical space required for sampling. Currently, it allows the full generation of reaction energy profiles from 2D representation of the reactants and products only.

AutodE has been applied to explore basic reactions in chemistry, including substitution, addition, elimination and cyclisation. To applicability of autodE to handle more complex scenarios has been demonstrated by exploring the reaction mechanism leading to the formation of the metabolite (+)-brevianamide. This molecule has been for the first time synthesized by chemical means by Lawrence and coworkers.[257] AutodE predicted the correct selectivity employing only 2D information extracted from ChemDraw.

Fernanda's group aims to combine these tools for the development of inverse design approaches in organic and supramolecular catalysis. Fernanda concluded that there is much to be learned from small systems, including increasing complexity, datasets, and transferability.

28 Predictive models for assessing conditions of hydrogenation reactions

Dr. Timur Madzhidov, Kazan Federal University

Retrosynthesis of a target molecule depends on finding a strategy (a pathway of synthetic procedures leading to the compound), finding suitable conditions to perform every reaction, and determining the reliability of the pathway, in terms of yield, purity etc.[258] One of the key challenges in the development of a synthesis strategy is the selection of optimal reaction conditions that provide the necessary regioselectivity and stereoselectivity, together with a high yield of the target reagent.[177,178,259,260] Prediction of optimal conditions using machine learning is a difficult task which is complicated by the absence of negative results, a possibility that the reaction can be carried out under several sets of conditions, and the false negatives uncertainty (if a predicted condition does not coincide with an experimental one it does not generally mean that the prediction is wrong). These problems are solved by ranking the predicted conditions.

Timur and his co-workers focused on hydrogenation reactions which are widely used in synthetic chemistry. The reaction yield varies as a function of the catalyst, solvent, pressure, and temperature. About 234,000 hydrogenation reactions were downloaded from Reaxys *via* the Reaxys API. Structures and conditions were then standardized. Out of 143,000 catalyst, reagent, and solvent names, 9024 of the most popular ones were standardized to 2371 standard names which fully covered 84% of reactions. The 233,905 reactions were reduced to 50,916 with known temperature and pressure, and 38,739 of those had a catalyst involved. The test set included 3692 reactions with reactants and products unseen in the training set.

Conditions were represented by a 40-bit vector: three bits for low, medium, or high temperature, three bits for low, medium, or high pressure; three bits for presence of acid, base, or catalytic poison; and 31 bits for the presence of catalyst. Each reaction was represented by a condensed graph or reaction (CGR)[261] handled by CGRtools,[262] a Python library for processing molecules, reactions, and CGRs.[263] A reaction can be encoded by a descriptor vector which can be used in data analysis or in structure-reactivity modeling. The descriptors are *in silico* design and data analysis (ISIDA) substructural fragments.[264]

Timur and his co-workers developed two neural network models. The first one, the likelihood ranking model, used ISIDA descriptors as input to the network, and ranked the top k predicted reaction conditions according to their applicability to a particular transformation. For comparison, the researchers also used a nearest neighbors approach and a null model (Figure 31). The precision of the likelihood ranking model for $k = 10$ was 85%, compared with 81% for the nearest neighbors approach, and 68% for the null model.

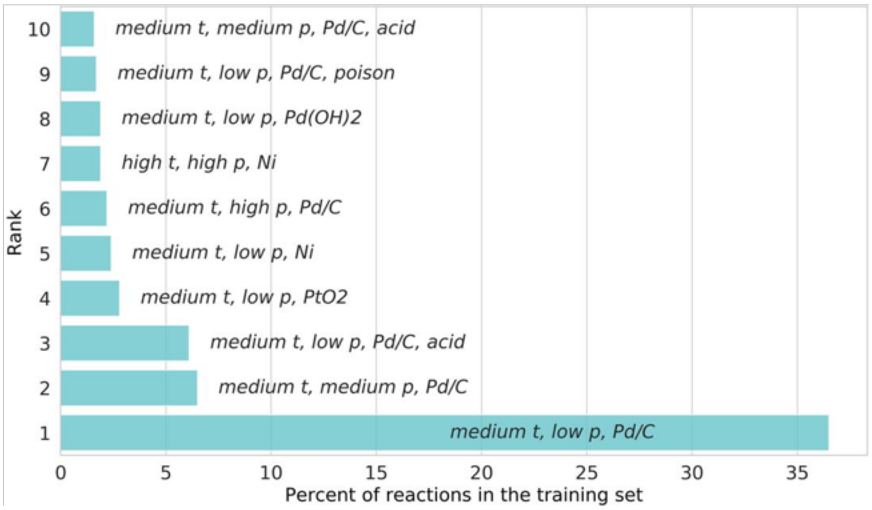


Figure 31: Null model.

A comparison with the model of Gao *et al.*[178] was also carried out (Figure 32). The comparison is not entirely fair for a number of reasons. The proportion of conditions without catalyst in the training dataset from the paper by Gao *et al.* is 87.3%. The model from that paper tends to predict the absence of any catalyst (33.5% of predictions). There is no standardization, so the names of catalysts are not unique. Only 18 conditions are considered, not all possible combinations. To make the comparison fairer, Timur and his co-workers tried a model much modified from that of Gao *et al.*,[178] using a neural network for each of the four sets of descriptors in turn. This recurrent prediction model is much better but is no improvement on the likelihood ranking model.

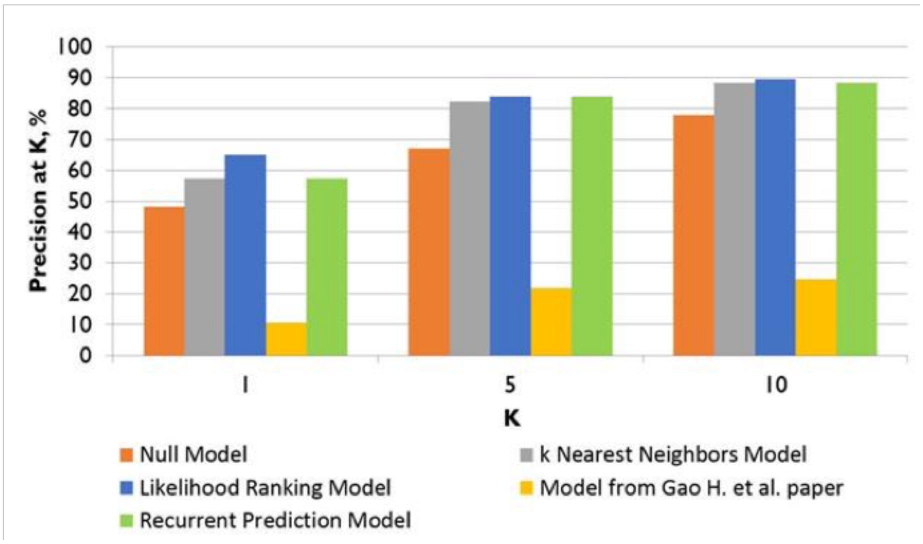


Figure 32: Test set predictions, with pressure not taken into account.

The likelihood ranking model was validated experimentally. The reactant **1** in Figure 33 could give rise to compounds **2** and **3** among many others. The predicted conditions "t = 30°C; p = 1 atm; Pt/C" gave compound **2** in 100% yield; "t = 30°C; p = 1 atm; Pt/C + acid" gave compound **2** in 42% yield and compound **3** in 58% yield. Two more concurrent reactions were also used for external validation.

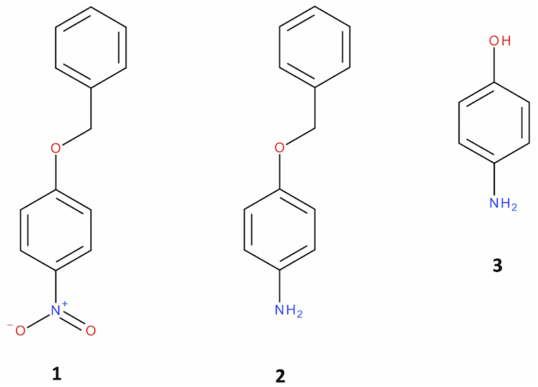


Figure 33: Compounds in experimental validation.

Timur drew the following conclusions. Condition prediction is a ranking problem. Data quality and problem-adapted modeling perform better than wide-scope (at least on this example). Simple models behave no worse than a more complex (and slower) one. Experimental validation of the likelihood ranking model on flow reactors supported the validity of the model. This model performs best but it could be applied only if the set of possible conditions is finite.

29 Retrosynthetic software for practicing chemists: novel and efficient *in silico* pathway design validated at the bench

Dr. Hugo Viana, Merck Millipore

Synthia[265] (developed under the name "Chematica") is retrosynthesis software that augments the chemist's expertise. It combines the powers of network theory, advanced computing, and chemical knowledge to help chemists identify viable pathways, and successfully synthesize target molecules, based on a chemist's unique style, quickly and economically. The development of Chematica[37,129,266,267] was led by Bartosz Grzybowski. In 2017, the software and database were purchased by Merck KGaA.

The manual process of deciding how to make a compound is a bottleneck in drug discovery. There are many challenges. Humans cannot remember every reaction. Established routes may not be reproducible. The chemist may not be able to design a path. Failure can happen at any step. Finding common building blocks may fail.

Synthia saves money because all pathways lead back to available building blocks. It essentially produces a tree with terminal nodes that are chemicals that are available in the laboratory or that can be bought. Imagine if the tree has 1000 reactions at each stage in a five-step reaction scheme. No chemist could cope with this challenge manually. There are millions of reactions in SciFinder and Reaxys: all this information is available, but it is not easy to get in the right format.

Synthia is unprecedented in that it takes into account complete stereochemistry, regiochemistry, protection chemistry, and potential reactivity conflicts.[38] Instead of relying on chemical reactions extracted by machine from existing literature precedents, which is both limiting and chemically faulty, Synthia uses 90,000 reaction rules coded by experts, carefully considering the indirect effects of the chemical environment (i.e., all the atoms and substituents present in the molecules which do not participate in the reaction directly), and their decisive influence on the applicability of a synthetic move.

Simple machine extraction from the literature does not work: it does not consider the entire context of the reaction core. Hugo showed examples of the pitfalls of machine extraction of reactions from the literature, one involving a stereodirecting methoxy group, another involving acidic protons which will not succeed under basic conditions, and a third involving incorrect stereochemistry. The Chematica team hand-coded all its rules as SMARTS. One rule might take several weeks to encode. Hugo showed a rule for the stereoselective condensation of esters with aldehydes.

Hand-coding allows for very unusual reactions: so-called "black swans". Hugo gave an example of a cycloaddition that had appeared in only 128 literature references. He also showed the facile preparation of *syn*-1,3-amino alcohols. Synthia concentrates on hand-coded rules that can account for the entire context of a molecule, including incompatible functional groups, protecting groups and stereochemistry and regiochemistry, all the while eliminating unfeasible chemistry. It explores novel and known solutions, eliminates nonviable options, and presents the user with the most promising pathways to explore. It stops at available starting materials, and can deal with novel molecules, not just those known in the literature.

Hugo presented a flowchart of how the software works. Synthia takes in SMILES and SMARTS, first ignores the reaction core, and then looks at the protecting groups needed, and uses a list of incompatible functional groups. Because the number of choices at each retrosynthetic step is about 100, the number of possibilities within n steps scales as 100^n. To search such an enormous synthetic space, intelligent algorithms are needed to truncate, and revert from unpromising "branches", and channel the searches toward the most efficient and elegant sequences of steps. Synthia avoids unpromising routes by using numerous heuristics prohibiting unlikely structural motifs, penalizing reactions that are nonselective, or those that would have to proceed through very strained intermediates. The searches are then guided toward the most feasible solutions by scoring functions. The scoring function is driven by the chemist's preferences, for example, the number of steps, the cost of starting materials, protecting group requirements, commercial or known inventory, nontoxic intermediates, catalysts, and exclusion of certain chemical classes.

It is claimed that Synthia is the first tool to be validated successfully in the laboratory. In a validation experiment,[38] Synthia was used to plan synthetic pathways of eight, structurally diverse, bioactive and natural products. To investigate whether retrosynthetic software such as Synthia could "empower" less experienced chemists to perform synthetic work that would typically be carried out in classic synthetic laboratories, the first four targets were made by MilliporeSigma chemists, whereas the last four were synthesized by students not experienced in multistep organic synthesis. Chemists were free to customize search criteria according to their own "synthetic style" and choose their preferred route.

Hugo presented just two of the syntheses as examples, although all of the eight syntheses were successful. There was no literature precedent for the synthesis of the second compound, α-hydroxyetizolam. In the eight-step route proposed by Synthia (overall yield 3.2%), the key hydroxyethyl side chain was installed at an early stage providing confidence in the route. The software correctly identified a side product, and the need for protection at one stage was indicated in the software's plan (by a blue halo). The time and cost savings were huge (40% cost savings) compared to the originally proposed route that was considered too risky to execute. MilliporeSigma has added the finished product to the catalog, and has successfully used the Synthia route to establish analogues. The objective for Synthia with the sixth compound ($5\beta/6\beta$-hydroxylurasidone) was to design an efficient pathway that would avoid the only known, but patented route. Using Synthia, MilliporeSigma developed an alternative patent-free pathway, and added the finished product to the catalog. Here the program made a choice that a chemist might consider "'risky". All in all, the route proved easily scalable to multigram quantities, and gave an overall yield of about 55%, that is, two and a half times higher than reported in the patented route.

Synthia is experiencing rapid and diverse global adoption. It saves chemists time, and augments their expertise. It has become the ally of bench chemists by "learning" chemistry much like chemists would themselves, and suggesting diverse pathways towards their targets, thus generating ideas and providing cost-effective routes based on each user's unique needs. Synthia will not replace the chemists; it will support them. It is becoming the chemist's preferred starting point and it is continually improving with feedback from users.

30 Conclusion

It was remarkable that this meeting took place and was so well attended: the coronavirus pandemic was already having a major impact worldwide, but this conference was scheduled immediately before "lockdown" in the United Kingdom. In the end, only three or four speakers were forced to withdraw (including, sadly, one of the three keynote speakers) and two speakers presented remotely. Readers who would like to see the full planned program, exhibitors, abstracts, biographies, posters, and some of the slides presented can find AI React 2020 on the web.[268]

I would like to thank the organizers most sincerely for inviting me. All of them deserve a big round of applause for the smooth running of the meeting, and the excellent opportunities for networking and exchange of opinions. I hate to single anyone out, but I do have to thank Samantha Kanza in particular for her amazing efficiency and coolness, faced with a mass of equipment in both Apple and Microsoft environments, coping with sound systems and remote presentations, and with constant changes to the agenda and slides.

It was a pleasure to attend this meeting and write the official "proceedings". The breadth of subject matter was quite fascinating: cheminformatics, retrosynthesis, chemical engineering, robotics, catalysis, computer science, quantum computers, and much more. This reflects the wide impact that AI could have on chemistry, "the central science". The reader may still ask "How and when?" but the future is clearly exciting.

References

(1) Dial-a-Molecule. `http://generic.wordpress.soton.ac.uk/dial-a-molecule/` (accessed March 23, 2020).

(2) Murray, P. M.; Bellany, F.; Benhamou, L.; Bučar, D.-K.; Tabor, A. B.; Sheppard, T. D. The application of design of experiments (DoE) reaction optimisation and solvent selection in the development of new synthetic chemistry. *Org. Biomol. Chem.* **2016**, *14* (8), 2373–2384.

(3) Directed Assembly Network. `http://directedassembly.org/` (accessed March 23, 2020).

(4) AI3SD Network. `http://www.ai3sd.org/home` (accessed March 23, 2020).

(5) Center for Rapid Online Analysis of Reactions (ROAR). `http://www.imperial.ac.uk/rapid-online-analysis-of-reactions/` (accessed March 23, 2020).

(6) Corey, E.; Wipke, W. T. Computer-assisted design of complex organic syntheses. *Science* **1969**, *166* (3902), 178–192.

(7) Wipke, W. T., Evolution of molecular graphics. In *Graphics for Chemical Structures. Integration with Text and Data*; Warr, W. A., Ed.; American Chemical Society: Washington, DC, 1987; Vol. ACS Symposium Series 341, pp 1–8.

(8) Cook, A.; Johnson, A. P.; Law, J.; Mirzazadeh, M.; Ravitz, O.; Simon, A. Computer-aided synthesis design: 40 years on. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2012**, *2* (1), 79–107.

(9) Cook, A. P. F. A treatment of stereochemistry in computer aided organic synthesis. Ph.D. Thesis, University of Leeds, 2015.

(10) Corey, E. J.; Kurti, L., *Enantioselective chemical synthesis: methods, logic, and practice*; Wiley-VCH: Weinheim, Germany, 2013.

(11) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28* (1), 31–36.

(12) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29* (2), 97–101.

(13) Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. InChI, the IUPAC international chemical identifier. *J. Cheminf.* **2015**, *7*, 23.

(14) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *J. Chem. Inf. Comput. Sci.* **1992**, *32* (3), 244–255.

(15) Chen, W. L.; Leland, B. A.; Durant, J. L.; Grier, D. L.; Christie, B. D.; Nourse, J. G.; Taylor, K. T. Self-contained sequence representation: bridging the gap between bioinformatics and cheminformatics. *J. Chem. Inf. Model.* **2011**, *51* (9), 2186–2208.

(16) RDKit fingerprints. http://www.rdkit.org/docs/GettingStartedInPython.html#list-of-available-fingerprints (accessed March 26, 2020).

(17) Grethe, G.; Blanke, G.; Kraut, H.; Goodman, J. M. International chemical identifier for reactions (RInChI). *J. Cheminf.* **2018**, *10*, 22.

(18) Han, B. Y.; Lam, N. Y.; MacGregor, C. I.; Goodman, J. M.; Paterson, I. A synthesis-enabled relative stereochemical assignment of the C1–C28 region of hemicalide. *Chem. Commun.* **2018**, *54* (26), 3247–3250.

(19) Smith, S. G.; Goodman, J. M. Assigning stereochemistry to single diastereoisomers by GIAO NMR calculation: The DP4 probability. *J. Am. Chem. Soc.* **2010**, *132* (37), 12946–12959.

(20) Ermanis, K.; Parkes, K. E.; Agback, T.; Goodman, J. M. Expanding DP4: application to drug compounds and automation. *Org. Biomol. Chem.* **2016**, *14* (16), 3943–3949.

(21) Howarth, A.; Ermanis, K.; Goodman, J. M. DP4-AI automated NMR data analysis: straight from spectrometer to structure. *Chem. Sci.* **2020**, *11* (17), 4351–4359.

(22) Machine Learning. http://en.wikipedia.org/wiki/Machine_learning (accessed March 30, 2020).

(23) Burkov, A. *The hundred-page machine learning book*; Andriy Burkov Publishing: Quebec, Canada, 2019.

(24) Modeling reinforcement learning (Part I): Defining and simulating RL models. http://speekenbrink-lab.github.io/modelling/2019/02/28/fit_kf_rl_1.html (accessed March 30, 2020).

(25) Modeling reinforcement learning (Part II): Maximum likelihood estimation. http://speekenbrink-lab.github.io/modelling/2019/08/29/fit_kf_rl_2.html (accessed March 30, 2020).

(26) ImageNet. http://www.image-net.org/ (accessed March 27, 2020).

(27) van den Oord, A.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; Kavukcuoglu, K. WaveNet: A generative model for raw audio. 2016, arXiv e-Print archive. http://arxiv.org/abs/1609.03499 (accessed March 28, 2020).

(28) Breiman, L. Bagging predictors. *Machine Learning* **1996**, *24* (2), 123–140.

(29) Marquez, E. S.; Hare, J. S.; Niranjan, M. Deep cascade learning. *IEEE transactions on neural networks and learning systems* **2018**, *29* (11), 5475–5485.

(30) Belilovsky, E.; Eickenberg, M.; Oyallon, E., Greedy layerwise learning can scale to ImageNet. 2018, arXiv e-Print archive. http://arxiv.org/abs/1812.11446 (accessed March 28, 2020).

(31) Du, X.; Farrahi, K.; Niranjan, M. Transfer learning across human activities using a cascade neural network architecture. In *Proceedings of the 23rd International Symposium on Wearable Computers (ISWC '19)*; Association for Computing Machinery: New York, NY, 2019, pp 35–44.

(32) Doersch, C., Tutorial on variational autoencoders. 2016, arXiv e-Print archive. http://arxiv.org/abs/1606.05908 (accessed March 29, 2020).

(33) Vaswami, A.; Shazeer, N.; Parmar, N.; Uszkorelt, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I., Attention is all you need. 2017, arXiv.org e-Print archive. http://arxiv.org/pdf/1706.03762.pdf (accessed March 30, 2020).

(34) Gunawardana, Y.; Niranjan, M. Bridging the gap between transcriptome and proteome measurements identifies post-translationally regulated genes. *Bioinformatics* **2013**, *29* (23), 3060–3066.

(35) Geirhos, R.; Rubisch, P.; Michaelis, C.; Bethge, M.; Wichmann, F. A.; Brendel, W., ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. 2018, arXiv e-Print archive. http://arxiv.org/abs/1811.12231 (accessed March 30, 2020).

(36) Alcorn, M. A.; Li, Q.; Gong, Z.; Wang, C.; Mai, L.; Ku, W.-S.; Nguyen, A., Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. 2018, arXive e-Print archive. http://arxiv.org/abs/1811.11553 (accessed March 30, 2020).

(37) Molga, K.; Gajewska, E. P.; Szymkuć, S.; Grzybowski, B. A. The logic of translating chemical knowledge into machine-processable forms: a modern playground for physical-organic chemistry. *React. Chem. Eng.* **2019**, *4* (9), 1506–1521.

(38) Klucznik, T.; Mikulak-Klucznik, B.; McCormack, M. P.; Lima, H.; Szymkuć, S.; Bhowmick, M.; Molga, K.; Zhou, Y.; Rickershauser, L.; Gajewska, E. P.; Toutchkine, A.; Dittwald, P.; Startek, M. P.; Kirkovits, G. J.; Roszak, R.; Adamski, A.; Sieredzińska, B.; Mrksich, M.; Trice, S. L. J.; Grzybowski, B. A. Efficient syntheses of diverse, medicinally relevant targets planned by computer and executed in the laboratory. *Chem* **2018**, *4* (3), 522–532.

(39) Lowe, D. M., Chemical reactions from US patents (1976-Sep2016). http://figshare.com/articles/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873 (accessed March 20, 2020).

(40) Shang, C.; Liu, Q.; Chen, K.-S.; Sun, J.; Lu, J.; Yi, J.; Bi, J. Edge attention-based multi-relational graph convolutional networks. 2018, arXiv e-Print archive. http://arxiv.org/abs/1802.04944 (accessed March 31, 2020).

(41) Sacha, M.; Błaż, M.; Byrski, P.; Włodarczyk-Pruszyński, P.; Jastrzębski, S. Well-designed chemical deep learning models match performance of expert heuristics and can generalize to new reactions. http://molecule.one/storage/app/media/poster.pdf (accessed March 31, 2020).

(42) Fujita, S. Description of organic reactions based on imaginary transition structures. 1. Introduction of new concepts. *J. Chem. Inf. Comput. Sci.* **1986**, *26* (4), 205–212.

(43) Li, M.; Roller, S.; Kulikov, I.; Welleck, S.; Boureau, Y.-L.; Cho, K.; Weston, J. Don't say that! Making inconsistent dialogue unlikely with unlikelihood training. 2019, arXiv e-Print archive. http://arxiv.org/abs/1911.03860 (accessed March 30, 2020).

(44) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS Cent. Sci.* **2019**, *5* (9), 1572–1583.

(45) Gao, W.; Coley, C. W. The synthesizability of molecules proposed by generative models. 2020, arXiv.org e-print archive. http://arxiv.org/abs/2002.07007 (accessed March 19, 2020).

(46) RDKit: open-source cheminformatics. http://www.rdkit.org (accessed March 30, 2020).

(47) Ertl, P.; Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminf.* **2009**, *1*, 8.

(48) Maziarka, Ł.; Danel, T.; Mucha, S.; Rataj, K.; Tabor, J.; Jastrzębski, S. Molecule attention transformer. 2020, arXiv e-Print archive. http://arxiv.org/abs/2002.08264v1 (accessed March 30, 2020).

(49) Liu, B.; Ramsundar, B.; Kawthekar, P.; Shi, J.; Gomes, J.; Luu–Nguyen, Q.; Ho, S.; Sloane, J.; Wender, P.; Pande, V. Retrosynthetic reaction prediction using neural sequence-to-sequence models. *ACS Cent. Sci.* **2017**, *3* (10), 1103–1113.

(50) Dai, H.; Li, C.; Coley, C. W.; Dai, B.; Song, L. Retrosynthesis prediction with conditional graph logic network. 2020, arXiv.org e-Print archive. http://arxiv.org/pdf/2001.01408.pdf (accessed March 19, 2020).

(51) Segler, M. H.; Waller, M. P. Neural-symbolic machine learning for retrosynthesis and reaction prediction. *Chem. - Eur. J.* **2017**, *23* (25), 5966–5971.

(52) Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F. Computer-assisted retrosynthesis based on molecular similarity. *ACS Cent. Sci.* **2017**, *3* (12), 1237–1245.

(53) Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature* **2016**, *533* (7604), 452–454.

(54) DigitalGlassware. http://www.deepmatter.io/blog/welcome-to-digitalglassware (accessed March 18, 2020).

(55) ICSYNTH. http://www.infochem.de/synthesis/ic-synth (accessed March 18, 2020).

(56) Kraut, H.; Eiblmaier, J.; Grethe, G.; Löw, P.; Matuszczyk, H.; Saller, H. Algorithm for reaction classification. *J. Chem. Inf. Model.* **2013**, *53* (11), 2884–2895.

(57) Bøgevig, A.; Federsel, H.-J.; Huerta, F.; Hutchings, M. G.; Kraut, H.; Langer, T.; Löw, P.; Oppawsky, C.; Rein, T.; Saller, H. Route design in the 21st century: The IC SYNTH software tool as an idea generator for synthesis prediction. *Org. Process Res. Dev.* **2015**, *19* (2), 357–368.

(58) SPRESI database. http://www.spresi.com/ (accessed March 18, 2020).

(59) Quantum gates. http://www.codeproject.com/Articles/5160469/Quantum-Computation-Primer-Part-2 (accessed April 1, 2020).

(60) Peruzzo, A.; McClean, J.; Shadbolt, P.; Yung, M.-H.; Zhou, X.-Q.; Love, P. J.; Aspuru-Guzik, A.; O'brien, J. L. A variational eigenvalue solver on a photonic quantum processor. *Nat. Commun.* **2014**, *5*, 4213.

(61) Tilly, J.; Jones, G.; Chen, H.; Wossnig, L.; Grant, E. Computation of molecular excited states on IBMQ using a discriminative variational quantum eigensolver. 2020, arXiv e-Print archive. http://arxiv.org/abs/2001.04941 (accessed March 31, 2020).

(62) Ollitrault, P. J.; Kandala, A.; Chen, C.-F.; Barkoutsos, P. K.; Mezzacapo, A.; Pistoia, M.; Sheldon, S.; Woerner, S.; Gambetta, J.; Tavernelli, I. Quantum equation of motion for computing molecular excitation energies on a noisy quantum processor. 2019, arXiv e-Print archive. http://arxiv.org/abs/1910.12890 (accessed March 31, 2020).

(63) Colless, J. I.; Ramasesh, V. V.; Dahlen, D.; Blok, M. S.; Kimchi-Schwartz, M.; McClean, J.; Carter, J.; De Jong, W.; Siddiqi, I. Computation of molecular spectra on a quantum processor with an error-resilient algorithm. *Phys. Rev. X* **2018**, *8* (1), 011021.

(64) McArdle, S.; Jones, T.; Endo, S.; Li, Y.; Benjamin, S. C.; Yuan, X. Variational ansatz-based quantum simulation of imaginary time evolution. *npj Quantum Information* **2019**, *5*, 75.

(65) Higgott, O.; Wang, D.; Brierley, S. Variational quantum computation of excited states. *Quantum* **2019**, *3*, 156.

(66) Greene-Diniz, G.; Ramo, D. M. Generalized unitary coupled cluster excitations for multireference molecular states optimized by the variational quantum eigensolver, 2019, arXiv e-Print archive. http://arxiv.org/abs/1910.05168 (accessed March 31, 2020).

(67) Synthesis planning in SciFinder. http://www.cas.org/products/scifinder/retrosynthesis-planning (accessed April 2, 2020).

(68) Segler, M. H.; Preuss, M.; Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **2018**, *555* (7698), 604–610.

(69) Spaya. http://beta.spaya.ai/ (accessed April 2, 2020).

(70) Pistachio dataset. http://www.nextmovesoftware.com/pistachio.html (accessed April 2, 2020).

(71) Mcule compound sourcing. http://mcule.com/compound-sourcing/ (accessed April 2, 2020).

(72) Kumar, A.; Swienty-Busch, J.; Bottjer, T.; Krstic, I.; Bali, S.; Waller, M. P.; Segler, M., Reaxys Predictive Retrosynthesis (PAI): rewiring chemistry and redesigning synthetic routes. http://drive.google.com/file/d/1haN_zmjT54_xJ6rJ8ORDetq9Wb-6M22w/view (accessed March 22, 2020).

(73) Artificial intelligence without data intelligence is artificial. http://www.digitalistmag.com/cio-knowledge/2019/05/23/artificial-intelligence-without-data-intelligence-is-artificial-06198533 (accessed April 6, 2020).

(74) Fey, N. Lost in chemical space? Maps to support organometallic catalysis. *Chem. Cent. J.* **2015**, *9*, 38.

(75) Durand, D. J.; Fey, N. Computational ligand descriptors for catalyst design. *Chem. Rev.* **2019**, *119* (11), 6561–6594.

(76) Jover, J.; Fey, N.; Harvey, J. N.; Lloyd-Jones, G. C.; Orpen, A. G.; Owen-Smith, G. J.; Murray, P.; Hose, D. R.; Osborne, R.; Purdie, M. Expansion of the ligand knowledge base for monodentate P-donor ligands (LKB-P). *Organometallics* **2010**, *29* (23), 6245–6258.

(77) Jover, J.; Fey, N.; Harvey, J. N.; Lloyd-Jones, G. C.; Orpen, A. G.; Owen-Smith, G. J.; Murray, P.; Hose, D. R.; Osborne, R.; Purdie, M. Expansion of the ligand knowledge base for chelating P, P-donor ligands (LKB-PP). *Organometallics* **2012**, *31* (15), 5302–5306.

(78) Pickup, O. J.; Khazal, I.; Smith, E. J.; Whitwood, A. C.; Lynam, J. M.; Bolaky, K.; King, T. C.; Rawe, B. W.; Fey, N. Computational discovery of stable transition-metal vinylidene complexes. *Organometallics* **2014**, *33* (7), 1751–1761.

(79) Fey, N.; Orpen, A. G.; Harvey, J. N. Building ligand knowledge bases for organometallic chemistry: Computational description of phosphorus (III)-donor ligands and the metal–phosphorus bond. *Coord. Chem. Rev.* **2009**, *253* (5-6), 704–722.

(80) Fernandez, A.; Reyes, C.; Wilson, M. R.; Woska, D. C.; Prock, A.; Giering, W. P. Examination of drago's EB and CB parameters for phosphines through the quantitative analysis of ligand effects (QALE). *Organometallics* **1997**, *16* (3), 342–348.

(81) Tolman, C. A. Steric effects of phosphorus ligands in organometallic chemistry and homogeneous catalysis. *Chem. Rev.* **1977**, *77* (3), 313–348.

(82) Fey, N.; Harvey, J. N.; Lloyd-Jones, G. C.; Murray, P.; Orpen, A. G.; Osborne, R.; Purdie, M. Computational descriptors for chelating P,P- and P,N-donor ligands. *Organometallics* **2008**, *27* (7), 1372–1383.

(83) Newland, R. J.; Smith, A.; Smith, D. M.; Fey, N.; Hanton, M. J.; Mansell, S. M. Accessing alkyl-and alkenylcyclopentanes from Cr-catalyzed ethylene oligomerization using 2-phosphinophosphinine ligands. *Organometallics* **2018**, *37* (6), 1062–1073.

(84) Jover, J.; Fey, N. Screening substituent and backbone effects on the properties of bidentate P, P-donor ligands (LKB-PP screen). *Dalton Trans.* **2013**, *42* (1), 172–181.

(85) Fey, N.; Tsipis, A. C.; Harris, S. E.; Harvey, J. N.; Orpen, A. G.; Mansson, R. A. Development of a ligand knowledge base, part 1: computational descriptors for phosphorus donor ligands. *Chem. - Eur. J.* **2005**, *12* (1), 291–302.

(86) Stauffer, S. R.; Hartwig, J. F. Fluorescence resonance energy transfer (FRET) as a high-throughput assay for coupling reactions. Arylation of amines as a case study. *J. Am. Chem. Soc.* **2003**, *125* (23), 6977–6985.

(87) Durand, D. J.; Fey, N.; Pringle, P. G.; Lynam, J. M.; Webster, R. L.; Cresswell, A. J. Catalyst seeking substrate: computational prediction in homogeneous organometallic catalysis. http://drive.google.com/file/d/1HHO-EBcYWi2oZH6gp9psmE1bLdYzcXSf/view (accessed April 4, 2020).

(88) Chemical entities of biological interest (ChEBI). http://www.ebi.ac.uk/chebi/ (accessed April 5, 2020).

(89) Gene ontology (GO). http://geneontology.org/ (accessed April 5, 2020).

(90) Sequence ontology. http://www.sequenceontology.org/ (accessed April 5, 2020).

(91) Open PHACTS. http://www.openphactsfoundation.org/ (accessed April 5, 2020).

(92) Logic and ontology. http://plato.stanford.edu/entries/logic-ontology/ (accessed April 5, 2020).

(93) Web Ontology Language (OWL). http://www.w3.org/OWL/ (accessed April 5, 2020).

(94) Open Biological and Biomedical Ontology (OBO). http://www.obofoundry.org/ (accessed April 5, 2020).

(95) Reaction ontology, RXNO. http://github.com/rsc-ontologies/rxno (accessed April 5, 2020).

(96) Carey, J. S.; Laffan, D.; Thomson, C.; Williams, M. T. Analysis of the reactions used for the preparation of drug candidate molecules. *Org. Biomol. Chem* **2006**, *4* (12), 2337–2347.

(97) NameRxn software from NextMove Software. http://www.nextmovesoftware.com/namerxn.html (accessed April 5, 2020).

(98) Hill, F.; Cho, K.; Korhonen, A.; Bengio, Y. Learning to understand phrases by embedding the dictionary. 2015, arXiv e-Print archive. http://arxiv.org/abs/1504.00548 (accessed April 5, 2020).

(99) fastText. http://fasttext.cc/ (accessed April 5, 2020).

(100) Corbett, P.; Boyle, J. Chemlistem: chemical named entity recognition using recurrent neural networks. *J. Cheminf.* **2018**, *10*, 59.

(101) Chemlistem chemically aware tokenization. http://bitbucket.org/rscapplications/chemlistem/src/master/chemtok/ (accessed April 5, 2020).

(102) An R wrapper around the fast T-distributed Stochastic Neighbor Embedding implementation by Van der Maaten. http://cran.r-project.org/web/packages/Rtsne/index.html (accessed April 14, 2020).

(103) Wijffels, J. R package containing Universal Dependencies 2.4 Models for UDPipe. http://github.com/jwijffels/udpipe.models.ud.2.4 (accessed April 14, 2020).

(104) Rimell, L.; Maillard, J.; Polajnar, T.; Clark, S. RELPRON: A relative clause evaluation data set for compositional distributional semantics. *Computational Linguistics* **2016**, *42* (4), 661–701.

(105) Pistoia Alliance UDM project. http://www.pistoiaalliance.org/projects/udm/ (accessed April 6, 2020).

(106) BIOVIA CTfile formats. http://www.3dsbiovia.com/products/collaborative-science/biovia-draw/ctfile-no-fee.html (accessed April 6, 2020).

(107) Pistoia Alliance UDM distribution. http://github.com/PistoiaAlliance/UDM (accessed April 6, 2020).

(108) Basic formal ontology (BFO). http://basic-formal-ontology.org/ (accessed April 6, 2020).

(109) Van Hilten, N.; Chevillard, F.; Kolb, P. Virtual compound libraries in computer-assisted drug discovery. *J. Chem. Inf. Model.* **2019**, *59* (2), 644–651.

(110) Rarey, M.; Stahl, M. Similarity searching in large combinatorial chemistry spaces. *J. Comput.-Aided Mol. Des.* **2001**, *15* (6), 497–520.

(111) Hartenfeller, M.; Schneider, G. Enabling future drug discovery by de novo design. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011**, *1* (5), 742–759.

(112) Baskin, I.; Gordeeva, E.; Devdariani, R.; Zefirov, N.; Paliulin, V.; Stankevich, M. Solving the inverse problem of structure-property relations for the case of topological indexes. *Dokl. Akad. Nauk SSSR* **1989**, *307* (3), 613–617.

(113) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* **2018**, *4* (2), 268–276.

(114) Segler, M. H. S.; Kogej, T.; Tyrchan, C.; Waller, M. P. Generating focussed molecule libraries for drug discovery with recurrent neural networks, 2017, arXiv e-Print archive. http://arxiv.org/abs/1701.01329 (accessed April 8, 2020).

(115) Olivecrona, M.; Blaschke, T.; Engkvist, O.; Chen, H. Molecular de-novo design through deep reinforcement learning. *J. Cheminf.* **2017**, *9*, 48.

(116) Ertl, P.; Lewis, R.; Martin, E.; Polyakov, V. In silico generation of novel, drug-like chemical matter using the LSTM neural network. 2017, arXiv e-Print archive. http://arxiv.org/abs/1712.07449 (accessed April 8, 2020).

(117) Pogány, P.; Arad, N.; Genway, S.; Pickett, S. D. De novo molecule design by translating from reduced graphs to SMILES. *J. Chem. Inf. Model.* **2019**, *59* (3), 1136–1146.

(118) Winter, R.; Montanari, F.; Noé, F.; Clevert, D.-A. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chem. Sci.* **2019**, *10* (6), 1692–1701.

(119) Engkvist, O. Big data and AI in drug design, molecular de novo generation, synthesis and chemogenomics predictions. Talk given at Big Ideas in Big Data in Drug Discovery, April 12, 2019. http://bigdatamgms2019.wordpress.com/talk4/ (accessed April 8, 2020).

(120) Bradshaw, J.; Paige, B.; Kusner, M. J.; Segler, M. H. S.; Hernández-Lobato, J. M. A model to search for synthesizable molecules. 2019, arXiv.org e-Print archive. http://arxiv.org/abs/1906.05221 (accessed March 23, 2020).

(121) Segler, M.; Preuss, M.; Waller, M. P. Towards “AlphaChem”: Chemical synthesis planning with tree search and deep neural network policies. 2017, arXiv e-Print archive. http://arxiv.org/pdf/1702.00020.pdf (accessed April 8, 2020).

(122) Corey, E. J.; Cheng, X.-M. *The logic of chemical synthesis*; Wiley: New York, NY, 1995.

(123) Silver, D.; Huang, A.; Maddison, C. J.; Guez, A.; Sifre, L.; Van Den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; Dieleman, S.; Grewe, D.; Nham, J.; Kalchbrenner, N.; Sutskever, I.; Lillicrap, T.; Leach, M.; Kavukcuoglu, K.; Graepel, T.; Hassabis, D. Mastering the game of go with deep neural networks and tree search. *Nature* **2016**, *529* (7587), 484–489.

(124) Silver, D.; Schrittwieser, J.; Simonyan, K.; Antonoglou, I.; Huang, A.; Guez, A.; Hubert, T.; Baker, L.; Lai, M.; Bolton, A.; Chen, Y.; Lillicrap, T.; Hui, F.; Sifre, L.; van den Driessche, G.; Graepel, T.; Hassabis, D. Mastering the game of go without human knowledge. *Nature* **2017**, *550* (7676), 354–359.

(125) Silver, D.; Hubert, T.; Schrittwieser, J.; Antonoglou, I.; Lai, M.; Guez, A.; Lanctot, M.; Sifre, L.; Kumaran, D.; Graepel, T.; Lillicrap, T.; Simonyan, K.; Hassabis, D. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science* **2018**, *362* (6419), 1140–1144.

(126) Vlédut︡s, G. É. Concerning one system of classification and codification of organic reactions. *Information Storage and Retrieval* **1963**, *1* (2-3), 117–146.

(127) Kayala, M. A.; Azencott, C.-A.; Chen, J. H.; Baldi, P. Learning to predict chemical reactions. *J. Chem. Inf. Model.* **2011**, *51* (9), 2209–2222.

(128) Kayala, M. A.; Baldi, P. ReactionPredictor: prediction of complex chemical reactions at the mechanistic level using machine learning. *J. Chem. Inf. Model.* **2012**, *52* (10), 2526–2540.

(129) Szymkuć, S.; Gajewska, E. P.; Klucznik, T.; Molga, K.; Dittwald, P.; Startek, M.; Bajczyk, M.; Grzybowski, B. A. Computer-assisted synthetic planning: The end of the beginning. *Angew. Chem., Int. Ed.* **2016**, *55* (20), 5904–5937.

(130) Segler, M. World programs for model-based learning and planning in compositional state and action spaces. 2019, arXiv e-Print archive. http://arxiv.org/pdf/1912.13007.pdf (accessed April 8, 2020).

(131) Christ, C. D.; Zentgraf, M.; Kriegl, J. M. Mining electronic laboratory notebooks: analysis, retrosynthesis, and reaction based enumeration. *J. Chem. Inf. Model.* **2012**, *52* (7), 1745–1756.

(132) Law, J.; Zsoldos, Z.; Simon, A.; Reid, D.; Liu, Y.; Khew, S. Y.; Johnson, A. P.; Major, S.; Wade, R. A.; Ando, H. Y. Route designer: a retrosynthetic analysis tool utilizing automated retrosynthetic rule generation. *J. Chem. Inf. Model.* **2009**, *49* (3), 593–602.

(133) Srivastava, R. K.; Greff, K.; Schmidhuber, J. Training very deep networks. Advances in neural information processing systems 28 (NIPS 2015). http://arxiv.org/abs/1507.06228 (accessed January 29, 2018).

(134) Clevert, D.-A.; Unterthiner, T.; Hochreiter, S. Fast and accurate deep network learning by exponential linear units (ELUs). 2015, arXiv e-Print archive. http://arxiv.org/abs/1511.07289 (accessed April 8, 2020).

(135) Sheridan, R. P. *J. Chem. Inf. Model.* **2013**, *53* (4), 783–790.

(136) Segler, M. H.; Waller, M. P. Modelling chemical reasoning to predict and invent reactions. *Chem. - Eur. J.* **2017**, *23* (25), 6118–6128.

(137) Schneider, N.; Lowe, D. M.; Sayle, R. A.; Landrum, G. A. Development of a novel fingerprint for chemical reactions and its application to large-scale reaction classification and similarity. *J. Chem. Inf. Model.* **2015**, *55* (1), 39–53.

(138) Patel, H.; Bodkin, M. J.; Chen, B.; Gillet, V. J. Knowledge-based approach to de novo design using reaction vectors. *J. Chem. Inf. Model.* **2009**, *49* (5), 1163–1184.

(139) Coulom, R. Efficient selectivity and backup operators in monte-carlo tree search. 5th International Conference on Computer and Games, May 2006, Turin, Italy. 2006, HAL-inria e-Print archive. http://hal.inria.fr/inria-00116992 (accessed April 8, 2020).

(140) Kocsis, L.; Szepesvári, C. Bandit Based Monte-Carlo Planning. Bandit based monte-carlo planning. In *Machine Learning: ECML 2006*; Fürnkranz, J., Scheffer, T., Spiliopoulou, M., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2006, pp 282–293.

(141) Thakkar, A.; Selmi, N.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E. J. “Ring breaker”: Neural network driven synthesis prediction of the ring system chemical space. http://chemrxiv.org/articles/_Ring_Breaker_Assessing_Synthetic_Accessibility_of_the_Ring_System_Chemical_Space/9938969 (accessed April 25, 2020).

(142) Buttar, D.; Jorner, K.; Norrby, P.-O. AI/machine learning for chemical development. http://drive.google.com/file/d/1otVjd_gRPau8UNXbmqhlbdKctzdLOTYo/view (accessed April 11, 2020).

(143) Reid, J. P.; Sigman, M. S. Comparing quantitative prediction methods for the discovery of small-molecule chiral catalysts. *Nat. Rev. Chem.* **2018**, *2* (10), 290–305.

(144) Wheeler, S. E.; Seguin, T. J.; Guan, Y.; Doney, A. C. Noncovalent interactions in organocatalysis and the prospect of computational catalyst design. *Acc. Chem. Res.* **2016**, *49* (5), 1061–1069.

(145) Hansen, E.; Rosales, A. R.; Tutkowski, B.; Norrby, P.-O.; Wiest, O. Prediction of stereochemistry using Q2MM. *Acc. Chem. Res.* **2016**, *49* (5), 996–1005.

(146) Tomberg, A.; Muratore, M. É.; Johansson, M. J.; Terstiege, I.; Sköld, C.; Norrby, P.-O. Relative strength of common directing groups in palladium-catalyzed aromatic C- H activation. *iScience* **2019**, *20*, 373–391.

(147) Rosales, A. R.; Wahlers, J.; Limé, E.; Meadows, R. E.; Leslie, K. W.; Savin, R.; Bell, F.; Hansen, E.; Helquist, P.; Munday, R. H.; Wiest, O.; Norrby, P.-O. Rapid virtual screening of enantioselective catalysts using CatVS. *Nat. Catal.* **2019**, *2* (1), 41–45.

(148) Rosales, A. R.; Quinn, T. R.; Wahlers, J.; Tomberg, A.; Zhang, X.; Helquist, P.; Wiest, O.; Norrby, P.-O. Application of Q2MM to predictions in stereoselective synthesis. *Chem. Commun.* **2018**, *54* (60), 8294–8311.

(149) Chansen, E., Quantum to molecular mechanics (Q2MM). http://github.com/q2mm (accessed April 11, 2020).

(150) Tomberg, A.; Johansson, M. J.; Norrby, P.-O. A predictive tool for electrophilic aromatic substitutions using machine learning. *J. Org. Chem.* **2019**, *84* (8), 4695–4703.

(151) IBM RXN. http://rxn.res.ibm.com/ (accessed March 23, 2020).

(152) Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. Predicting reaction performance in C-N cross-coupling using machine learning. *Science* **2018**, *360* (6385), 186–190.

(153) Lowe, D. M. Extraction of chemical structures and reactions from the literature, Ph. D. Thesis, University of Cambridge, Cambridge U.K., June 2012. http://www.repository.cam.ac.uk/bitstream/handle/1810/244727/lowethesis.pdf?sequence=1&isAllowed=y (accessed March 20, 2020).

(154) Wei, C.-H.; Peng, Y.; Leaman, R.; Davis, A. P.; Mattingly, C. J.; Li, J.; Wiegers, T. C.; Lu, Z. Assessing the state of the art in biomedical relation extraction: overview of the BioCreative V chemical-disease relation (CDR) task. *Database* **2016**, *baw032*, 1–8.

(155) Roughley, S. D.; Jordan, A. M. The medicinal chemist's toolbox: An analysis of reactions used in the pursuit of drug candidates. *J. Med. Chem.* **2011**, *54* (10), 3451–3479.

(156) Schneider, N.; Lowe, D. M.; Sayle, R. A.; Tarselli, M. A.; Landrum, G. A. Big data from pharmaceutical patents: a computational analysis of medicinal chemists' bread and butter. *J. Med. Chem.* **2016**, *59* (9), 4385–4402.

(157) Pitt, W. R.; Parry, D. M.; Perry, B. G.; Groom, C. R. Heteroaromatic rings of the future. *J. Med. Chem.* **2009**, *52* (9), 2952–2963.

(158) RDF Schema. http://www.w3.org/TR/rdf-schema/ (accessed April 15, 2020).

(159) SPARQL query language for RDF. http://www.w3.org/TR/rdf-sparql-query/ (accessed April 15, 2020).

(160) Kanza, S.; Frey, J. G. A new wave of innovation in Semantic web tools for drug discovery. *Expert Opin. Drug Discovery* **2019**, *14* (5), 433–444.

(161) Kanza, S.; Gibbins, N.; Frey, J. G. Too many tags spoil the metadata: investigating the knowledge management of scientific research with semantic web technologies. *J. Cheminf.* **2019**, *11*, 23.

(162) Knight, N. J.; Kanza, S.; Cruickshank, D.; Brocklesby, W. S.; and Frey, J. G. Talk2Lab: The smart lab of the future. 2020, Southampton eprints archive. http://eprints.soton.ac.uk/441022/1/Talk2Lab_The_Smart_Lab_of_the_Future.pdf (accessed June 14, 2020).

(163) Kanza, S. What influence would a cloud based semantic laboratory notebook have on the digitisation and management of scientific research? Ph.D. Thesis, University of Southampton, Southampton, U.K., April 2018. http://eprints.soton.ac.uk/421045/ (accessed April 15, 2020).

(164) Kanza, S.; Willoughby, C.; Gibbins, N.; Whitby, R.; Frey, J. G.; Erjavec, J.; Zupančič, K.; Hren, M.; Kovač, K. Electronic lab notebooks: can they replace paper? *J. Cheminf.* **2017**, *9*, 31.

(165) Paul, S. M.; Mytelka, D. S.; Dunwiddie, C. T.; Persinger, C. C.; Munos, B. H.; Lindborg, S. R.; Schacht, A. L. How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discovery* **2010**, *9* (3), 203–214.

(166) Coley, C. W.; Eyke, N. S.; Jensen, K. F. Autonomous discovery in the chemical sciences part II: Outlook. *Angew. Chem., Int. Ed.* **2019**, Ahead of print.

(167) Jensen, K. F.; Coley, C. W.; Eyke, N. S. Autonomous discovery in the chemical sciences part I: Progress. *Angew. Chem., Int. Ed.* **2019**, Ahead of print.

(168) Corey, E.; Jorgensen, W. L. Computer-assisted synthetic analysis. Synthetic strategies based on appendages and the use of reconnective transforms. *J. Am. Chem. Soc.* **1976**, *98* (1), 189–203.

(169) Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; Jensen, K. F. Prediction of organic reaction outcomes using machine learning. *ACS Cent. Sci.* **2017**, *3* (5), 434–443.

(170) Coley, C. W.; Green, W. H.; Jensen, K. F. RDChiral: An RDKit wrapper for handling stereochemistry in retrosynthetic template extraction and application. *J. Chem. Inf. Model.* **2019**, *59* (6), 2529–2537.

(171) Chen, B.; Shen, T.; Jaakkola, T. S.; Barzilay, R. Learning to make generalizable and diverse predictions for retrosynthesis. 2019, arXiv.org e-Print archive. http://arxiv.org/abs/1910.09688 (accessed March 19, 2020).

(172) Coley, C. W.; Jin, W.; Rogers, L.; Jamison, T. F.; Jaakkola, T. S.; Green, W. H.; Barzilay, R.; Jensen, K. F. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chem. Sci.* **2019**, *10* (2), 370–377.

(173) Coley, C. W.; Green, W. H.; Jensen, K. F. Machine learning in computer-aided synthesis planning. *Acc. Chem. Res.* **2018**, *51* (5), 1281–1289.

(174) Kishimoto, A.; Buesser, B.; Chen, B.; Botea, A. Depth-first proof-number search with heuristic edge cost and application to chemical synthesis planning. Advances in Neural Information Processing Systems (NIPS 32), 2019. http://papers.nips.cc/paper/8943-depth-first-proof-number-search-with-heuristic-edge-cost-and-application-to-chemical-synthesis-planning.pdf (accessed March 19, 2020).

(175) Schreck, J. S.; Coley, C. W.; Bishop, K. J. Learning retrosynthetic planning through simulated experience. *ACS Cent. Sci.* **2019**, *5* (6), 970–981.

(176) Fortunato, M.; Coley, C. W.; Barnes, B.; Jensen, K. F. Data augmentation and pretraining for template-based retrosynthetic prediction in computer-aided synthesis planning. 2020, ChemRxiv e-print archive. http://chemrxiv.org/articles/Data_Augmentation_and_Pretraining_for_Template-Based_Retrosynthetic_Prediction_in_Computer-Aided_Synthesis_Planning/11811564 (accessed March 19, 2020).

(177) Nielsen, M. K.; Ahneman, D. T.; Riera, O.; Doyle, A. G. Deoxyfluorination with sulfonyl fluorides: navigating reaction space with machine learning. *J. Am. Chem. Soc.* **2018**, *140* (15), 5004–5008.

(178) Gao, H.; Struble, T. J.; Coley, C. W.; Wang, Y.; Green, W. H.; Jensen, K. F. Using machine learning to predict suitable conditions for organic reactions. *ACS Cent. Sci.* **2018**, *4* (11), 1465–1476.

(179) Jin, W.; Coley, C. W.; Barzilay, R.; Jaakkola, T. S. Predicting organic reaction outcomes with weisfeiler-lehman network. 2017, arXiv.org e-Print archive. http://arxiv.org/pdf/1709.04555.pdf (accessed March 19, 2020).

(180) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Bekas, C.; Lee, A. A. Molecular transformer for chemical reaction prediction and uncertainty estimation. 2018, arXiv.org e-Print archive. http://arxiv.org/abs/1811.02633 (accessed March 19, 2020).

(181) Qian, W. W.; Russell, N.; Simons, C. L. W.; Luo, Y.; Burke, M. D.; Peng, J. Integrating deep neural networks and symbolic inference for organic reactivity prediction. http://chemrxiv.org/articles/Integrating_Deep_Neural_Networks_and_Symbolic_Inference_for_Organic_Reactivity_Prediction/11659563 (accessed June 11, 2020).

(182) Struble, T. J.; Coley, C. W.; Jensen, K. F. Multitask prediction of site selectivity in aromatic C-H functionalization reactions. http://chemrxiv.org/articles/Multitask_Prediction_of_Site_Selectivity_in_Aromatic_C-H_Functionalization_Reactions/9735599 (accessed March 19, 2020).

(183) Kromann, J. C.; Jensen, J. H.; Kruszyk, M.; Jessing, M.; Jørgensen, M. Fast and accurate prediction of the regioselectivity of electrophilic aromatic substitution reactions. *Chem. Sci.* **2018**, *9* (3), 660–665.

(184) Coley, C. W.; Thomas, D. A.; Lummiss, J. A.; Jaworski, J. N.; Breen, C. P.; Schultz, V.; Hart, T.; Fishman, J. S.; Rogers, L.; Gao, H.; Hicklin, R. W.; Plehiers, P. P.; Byington, J.; Piotti, J. S.; Green, W. H.; Hart, A. J.; Jamison, T. F.; Jensen, K. A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science* **2019**, *365* (6453), eaax1566.

(185) Korovina, K.; Xu, S.; Kandasamy, K.; Neiswanger, W.; Poczos, B.; Schneider, J.; Xing, E. P. ChemBO: Bayesian optimization of small organic molecules with synthesizable recommendations. 2019, arXiv.org e-Print archive. http://arxiv.org/abs/1908.01425 (accessed March 21, 2020).

(186) Automated system for knowledge-based, continuous organic synthesis (ASKCOS). http://askcos.mit.edu (accessed March 21, 2020).

(187) Cronin group, Glasgow. http://www.chem.gla.ac.uk/cronin/ (accessed March 23, 2020).

(188) Chemputer video. http://www.gla.ac.uk/news/archiveofnews/2018/november/headline_624196_en.html (accessed March 23, 2020).

(189) Gromski, P. S.; Granda, J. M.; Cronin, L. Universal chemical synthesis and discovery with 'The Chemputer'. *Trends Chem.* **2020**, *2* (1), 4–12.

(190) Steiner, S.; Wolf, J.; Glatzel, S.; Andreou, A.; Granda, J. M.; Keenan, G.; Hinkley, T.; Aragon-Camarasa, G.; Kitson, P. J.; Angelone, D.; Cronin, L. Organic synthesis in a modular robotic system driven by a chemical programming language. *Science* **2019**, *363* (6423), eaav2211.

(191) Chemify. http://www.chem.gla.ac.uk/cronin/chemify/ (accessed March 23, 2020).

(192) BAM. https://www.bam.de/Navigation/EN/Home/home.html (accessed March 23, 2020).

(193) NCATS challenge. http://ncats.nih.gov/aspire/2018ChallengeWinners#c5 (accessed March 23, 2020).

(194) Henson, A. B.; Gromski, P. S.; Cronin, L. Designing algorithms to aid discovery by chemical robots. *ACS Cent. Sci.* **2018**, *4* (7), 793–804.

(195) Granda, J. M.; Donina, L.; Dragone, V.; Long, D.-L.; Cronin, L. Controlling an organic synthesis robot with machine learning to search for new reactivity. *Nature* **2018**, *559* (7714), 377–381.

(196) Marshall, S. M.; Murray, A. R.; Cronin, L. A probabilistic framework for identifying biosignatures using Pathway Complexity. *Philos. Trans. A Math. Phys. Eng. Sci.* **2017**, *375* (2109).

(197) Marshall, S. M.; Moore, D.; Murray, A. R. G.; Walker, S. I.; Cronin, L. Quantifying the pathways to life using assembly spaces. 2019, arXiv e-Print archive. http://arxiv.org/abs/1907.04649 (accessed March 30, 2020).

(198) Afanasyev, O. I.; Kuchuk, E.; Usanov, D. L.; Chusov, D. Reductive amination in the synthesis of pharmaceuticals. *Chem. Rev.* **2019**, *119* (23), 11857–11911.

(199) Bagal, D. B.; Watile, R. A.; Khedkar, M. V.; Dhake, K. P.; Bhanage, B. M. PS-Pd–NHC: an efficient and heterogeneous recyclable catalyst for direct reductive amination of carbonyl compounds with primary/secondary amines in aqueous medium. *Catal. Sci. Technol.* **2012**, *2* (2), 354–358.

(200) Pelckmans, M.; Renders, T.; Van de Vyver, S.; Sels, B. Bio-based amines through sustainable heterogeneous catalysis. *Green Chem.* **2017**, *19* (22), 5303–5331.

(201) Yang, H.; Cui, X.; Deng, Y.; Shi, F. Reductive amination of aldehydes and amines with an efficient Pd/NiO catalyst. *Synth. Commun.* **2014**, *44* (9), 1314–1322.

(202) Sagmeister, P.; Williams, J. D.; Hone, C. A.; Kappe, C. O. Laboratory of the future: a modular flow platform with multiple integrated PAT tools for multistep reactions. *React. Chem. Eng.* **2019**, *4* (9), 1571–1578.

(203) Hone, C. A.; Kappe, C. O. The use of molecular oxygen for liquid phase aerobic oxidations in continuous flow. *Top. Curr. Chem.* **2019**, *377* (1), 2.

(204) gPROMS FormulatedProducts. http://www.psenterprise.com/products/gproms/formulatedproducts (accessed April 19, 2020).

(205) Catalán, J. Toward a generalized treatment of the solvent effect based on four empirical scales: dipolarity (SdP, a new scale), polarizability (SP), acidity (SA), and basicity (SB) of the medium. *J. Phys. Chem. B* **2009**, *113* (17), 5951–5960.

(206) Laurence, C.; Legros, J.; Chantzis, A.; Planchat, A.; Jacquemin, D. A database of dispersion-induction DI, electrostatic ES, and hydrogen bonding $\alpha 1$ and $\beta 1$ solvent parameters and some applications to the multiparameter correlation analysis of solvent effects. *J. Phys. Chem. B* **2015**, *119* (7), 3174–3184.

(207) *Hansen solubility parameters: a user's handbook*, Second ed; Hansen, C. M., Ed.; CRC Press LLC: Boca Raton, FL: 2007.

(208) Gale, E.; Wirawan, R. H.; Silveira, R. L.; Pereira, C. S.; Johns, M. A.; Skaf, M. S.; Scott, J. L. Directed discovery of greener cosolvents: New cosolvents for use in ionic liquid based organic electrolyte solutions for cellulose dissolution. *ACS Sustainable Chem. Eng.* **2016**, *4* (11), 6200–6207.

(209) Centillion. http://www.centillion-tech.com/ (accessed April 21, 2020).

(210) Makatsoris, C. Modular fluid flow reactor. WO2019193346A1, 2019.

(211) Makatsoris, C. Modular fluid flow reactor. GB2572589A, 2019.

(212) Makatsoris, C.; Paramonov, L.; Alsharif, R. A modular flow reactor. WO2013050764A1, 2013.

(213) Makatsoris, C.; Paramonov, L.; Alsharif, R. Modular flow reactor. US20150010445A1, 2015.

(214) Mumith, J.-A.; Makatsoris, C.; Karayiannis, T. Design of a thermoacoustic heat engine for low temperature waste heat recovery in food manufacturing: A thermoacoustic device for heat recovery. *Appl. Therm. Eng.* **2014**, *65* (1), 588–596.

(215) Mumith, J.-A.; Karayiannis, T.; Makatsoris, C. Design and optimization of a thermoacoustic heat engine using reinforcement learning. *Int. J. Low-Carbon Technol.* **2016**, *11* (3), 431–439.

(216) Witek-Krowiak, A.; Chojnacka, K.; Podstawczyk, D.; Dawiec, A.; Pokomeda, K. Application of response surface methodology and artificial neural network methods in modelling and optimization of biosorption process. *Bioresour. Technol.* **2014**, *160*, 150–160.

(217) Urakawa, A.; Van Beek, W.; Monrabal-Capilla, M.; Galán-Mascarós, J. R.; Palin, L.; Milanesio, M. Combined, modulation enhanced X-ray powder diffraction and raman spectroscopic study of structural transitions in the spin crossover material [Fe (Htrz) 2 (trz)](BF4). *J. Phys. Chem. C* **2011**, *115* (4), 1323–1329.

(218) Nelder-Mead optimization. http://codesachin.wordpress.com/2016/01/16/nelder-mead-optimization/ (accessed April 21, 2020).

(219) Cook, T. R.; Zheng, Y.-R.; Stang, P. J. Metal–organic frameworks and self-assembled supramolecular coordination complexes: comparing and contrasting the design, synthesis, and functionality of metal–organic materials. *Chem. Rev.* **2013**, *113* (1), 734–777.

(220) Fujita, D.; Ueda, Y.; Sato, S.; Mizuno, N.; Kumasaka, T.; Fujita, M. Self-assembly of tetravalent Goldberg polyhedra from 144 small components. *Nature* **2016**, *540* (7634), 563–566.

(221) Vasdev, R. A.; Preston, D.; Crowley, J. D. Multicavity metallosupramolecular architectures. *Chem. - Asian J.* **2017**, *12* (19), 2513–2523.

(222) Saha, S.; Regeni, I.; Clever, G. H. Structure relationships between bis-monodentate ligands and coordination driven self-assemblies. *Coord. Chem. Rev.* **2018**, *374*, 1–14.

(223) Fang, Y.; Powell, J. A.; Li, E.; Wang, Q.; Perry, Z.; Kirchon, A.; Yang, X.; Xiao, Z.; Zhu, C.; Zhang, L.; Huang, F.; Zhou, H.-C. Catalytic reactions within the cavity of coordination cages. *Chem. Soc. Rev.* **2019**, *48* (17), 4707–4730.

(224) Sanders, J. K. Supramolecular catalysis in transition. *Chem. - Eur. J.* **1998**, *4* (8), 1378–1383.

(225) Breslow, R.; Dong, S. D. Biomimetic reactions catalyzed by cyclodextrins and their derivatives. *Chem. Rev.* **1998**, *98* (5), 1997–2011.

(226) Pahima, E.; Zhang, Q.; Tiefenbacher, K.; Major, D. T. Discovering monoterpene catalysis inside nanocapsules with multiscale modeling and experiments. *J. Am. Chem. Soc.* **2019**, *141* (15), 6234–6246.

(227) Vaissier Welborn, V.; Head-Gordon, T. Electrostatics generated by a supramolecular capsule stabilizes the transition state for carbon–carbon reductive elimination from gold (III) complex. *J. Phys. Chem. Lett.* **2018**, *9* (14), 3814–3818.

(228) Vaissier Welborn, V.; Li, W.-L.; Head-Gordon, T. Interplay of water and a supramolecular capsule for catalysis of reductive elimination reaction from gold. *Nat. Commun.* **2020**, *11* (1), 415.

(229) Norjmaa, G.; Maréchal, J.-D.; Ujaque, G. Microsolvation and encapsulation effects on supramolecular catalysis: C–C reductive elimination inside $[Ga_4L_6]^{12-}$ metallocage. *J. Am. Chem. Soc.* **2019**, *141* (33), 13114–13123.

(230) Young, T. A.; Martí-Centelles, V.; Wang, J.; Lusby, P. J.; Duarte, F. Rationalizing the acctivity of an “artificial Diels-Alderase”: Establishing efficient and accurate protocols for calculating supramolecular catalysis. *J. Am. Chem. Soc.* **2019**, *142* (3), 1300–1310.

(231) Martí-Centelles, V.; Lawrence, A. L.; Lusby, P. J. High activity and efficient turnover by a simple, self-assembled “artificial Diels–Alderase”. *J. Am. Chem. Soc.* **2018**, *140* (8), 2862–2868.

(232) Spicer, R. L.; Stergiou, A.; Young, T. A.; Duarte, F.; Symes, M. D.; Lusby, P. J. Host-guest induced electron transfer triggers radical-cation catalysis. *J. Am. Chem. Soc.* **2020**, *142* (5), 2134–2139.

(233) Åqvist, J. Modelling of ion-ligand interactions in solutions and biomolecules. *J. Mol. Struct.: THEOCHEM* **1992**, *88*, 135–152.

(234) Stote, R. H.; Karplus, M. Zinc binding in proteins and solution: a simple but accurate nonbonded representation. *Proteins: Struct., Funct., Genet.* **1995**, *23* (1), 12–31.

(235) Joung, I. S.; Cheatham III, T. E. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J. Phys. Chem. B* **2008**, *112*, 9020–9041.

(236) Vedani, A.; Huhta, D. W. A new force field for modeling metalloproteins. *J. Am. Chem. Soc.* **1990**, *112* (12), 4759–4967.

(237) Lin, F.; Wang, R. Systematic derivation of AMBER force field parameters applicable to zinc-containing systems. *J. Chem. Theory Comput.* **2010**, *6* (6), 1852–1870.

(238) Neves, R. P.; Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. Parameters for molecular dynamics simulations of manganese-containing metalloproteins. *J. Chem. Theory Comput.* **2013**, *9* (6), 2718–2732.

(239) Duarte, F.; Bauer, P.; Barrozo, A.; Amrein, B. A.; Purg, M.; Åqvist, J.; Kamerlin, S. C. L. Force field independent metal parameters using a nonbonded dummy model. *J. Phys. Chem. B* **2014**, *118* (16), 4351–4362.

(240) Aaqvist, J.; Warshel, A. Free energy relationships in metalloenzyme-catalyzed reactions. Calculations of the effects of metal ion substitutions in staphylococcal nuclease. *J. Am. Chem. Soc.* **1990**, *112* (8), 2860–2868.

(241) Pang, Y.-P.; Xu, K.; El Yazal, J.; Prendergast, F. G. Successful molecular dynamics simulation of the zinc-bound farnesyltransferase using the cationic dummy atom approach. *Protein Sci.* **2000**, *9* (10), 1857–1865.

(242) Oelschlaeger, P.; Klahn, M.; Beard, W. A.; Wilson, S. H.; Warshel, A. Magnesium-cationic dummy atom molecules enhance representation of DNA polymerase β in molecular dynamics simulations: Improved accuracy in studies of structural features and mutational effects. *J. Mol. Biol.* **2007**, *366* (2), 687–701.

(243) Yoneya, M.; Yamaguchi, T.; Sato, S.; Fujita, M. Simulation of metal–ligand self-assembly into spherical complex M6L8. *J. Am. Chem. Soc.* **2012**, *134* (35), 14401–14407.

(244) Cgbind GUI. http://cgbind.chem.ox.ac.uk/ (accessed April 22, 2020).

(245) Cgbind code. http://github.com/duartegroup/cgbind (accessed April 22, 2020).

(246) Willems, T. F.; Rycroft, C. H.; Kazi, M.; Meza, J. C.; Haranczyk, M. Algorithms and tools for high-throughput geometry-based analysis of crystalline porous materials. *Microporous Mesoporous Mater.* **2012**, *149* (1), 134–141.

(247) Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; Ceder, G. Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **2013**, *68*, 314–319.

(248) Greenaway, R.; Santolini, V.; Bennison, M.; Alston, B.; Pugh, C.; Little, M.; Miklitz, M.; Eden-Rump, E.; Clowes, R.; Shakil, A.; Cuthbertson, H. J.; Armstrong, H.; Briggs, M. E.; Cooper, A. I.; Santolini, V.; Miklitz, M.; Jelfs, K. E. High-throughput discovery of organic cages and catenanes using computational screening fused with robotic synthesis. *Nat. Commun.* **2018**, *9*, 1–11.

(249) Rosen, A. S.; Notestein, J. M.; Snurr, R. Q. Identifying promising metal–organic frameworks for heterogeneous catalysis via high-throughput periodic density functional theory. *J. Comput. Chem.* **2019**, *40* (12), 1305–1318.

(250) Automated reaction profile generation. http://github.com/duartegroup/autodE (accessed April 28, 2020).

(251) Martínez-Núñez, E. An automated transition state search using classical trajectories initialized at multiple minima. *Phys. Chem. Chem. Phys.* **2015**, *17* (22), 14912–14921.

(252) Habershon, S. Automated prediction of catalytic mechanism and rate law using graph-based reaction path sampling. *J. Chem. Theory Comput.* **2016**, *12* (4), 1786–1798.

(253) Jacobson, L. D.; Bochevarov, A. D.; Watson, M. A.; Hughes, T. F.; Rinaldo, D.; Ehrlich, S.; Steinbrecher, T. B.; Vaitheeswaran, S.; Philipp, D. M.; Halls, M. D.; Friesner, R. A. Automated transition state search and its application to diverse types of organic reactions. *J. Chem. Theory Comput.* **2017**, *13* (11), 5780–5797.

(254) Dewyer, A. L.; Argüelles, A. J.; Zimmerman, P. M. Methods for exploring reaction space in molecular systems. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2018**, *8* (2), e1354.

(255) Simm, G. N.; Vaucher, A. C.; Reiher, M. Exploration of reaction pathways and chemical transformation networks. *J. Phys. Chem. A* **2018**, *123* (2), 385–399.

(256) Guan, Y.; Ingman, V. M.; Rooks, B. J.; Wheeler, S. E. AARON: an automated reaction optimizer for new catalysts. *J. Chem. Theory Comput.* **2018**, *14* (10), 5249–5261.

(257) Godfrey, R.; Green, N.; Nichol, G.; Lawrence, A. Chemical synthesis of (+)-brevianamide A supports a Diels-Alderase-free biosynthesis. http://chemrxiv.org/articles/Chemical_Synthesis_of_-Brevianamide_a_Supports_a_Diels_Alderase-Free_Biosynthesis/8224148 (accessed June 11, 2020).

(258) Baskin, I. I.; Madzhidov, T. I.; Antipin, I. S.; Varnek, A. A. Artificial intelligence in synthetic chemistry: achievements and prospects. *Russ. Chem. Rev.* **2017**, *86* (11), 1127–1156.

(259) Marcou, G.; Aires de Sousa, J.; Latino, D. A.; De Luca, A.; Horvath, D.; Rietsch, V.; Varnek, A. Expert system for predicting reaction conditions: the Michael reaction case. *J. Chem. Inf. Model.* **2015**, *55* (2), 239–250.

(260) Lin, A. I.; Madzhidov, T. I.; Klimchuk, O.; Nugmanov, R. I.; Antipin, I. S.; Varnek, A. Automatized assessment of protective group reactivity: a step toward big reaction data analysis. *J. Chem. Inf. Model.* **2016**, *56* (11), 2140–2148.

(261) Varnek, A.; Fourches, D.; Hoonakker, F.; Solov'ev, V. P. Substructural fragments: an universal language to encode reactions, molecular and supramolecular structures. *J. Comput.-Aided Mol. Des.* **2005**, *19* (9/10), 693–703.

(262) CGRtools. http://github.com/cimm-kzn/CGRtools (accessed April 23, 2020).

(263) Nugmanov, R. I.; Mukhametgaleev, R. N.; Akhmetshin, T.; Gimadiev, T. R.; Afonina, V. A.; Madzhidov, T. I.; Varnek, A. CGRtools: Python library for molecule, reaction, and condensed graph of reaction processing. *J. Chem. Inf. Model.* **2019**, *59* (6), 2516–2521.

(264) Varnek, A. Fragment descriptors in structure–property modeling and virtual screening. In *Chemoinformatics and Computational Chemical Biology*; Bajorath, J., Ed.; Humana Press: New York, NY, 2010, pp 213–243.

(265) Synthia. http://www.sigmaaldrich.com/chemistry/chemical-synthesis/synthesis-software.html (accessed April 24, 2020).

(266) Fialkowski, M.; Bishop, K. J.; Chubukov, V. A.; Campbell, C. J.; Grzybowski, B. A. Architecture and evolution of organic chemistry. *Angew. Chem., Int. Ed.* **2005**, *44* (44), 7263–7269.

(267) Kowalik, M.; Gothard, C. M.; Drews, A. M.; Gothard, N. A.; Weckiewicz, A.; Fuller, P. E.; Grzybowski, B. A.; Bishop, K. J. Parallel optimization of synthetic pathways within the network of organic chemistry. *Angew. Chem., Int. Ed.* **2012**, *51* (32), 7928–7932.

(268) AI React 2020. http://www.ai3sd.org/aireact2020 (accessed April 29, 2020).