National Institutes of Health (NIH) Workshop on Reaction Informatics

Wendy A. Warr

Wendy Warr & Associates, 6 Berwick Court, Holmes Chapel, Cheshire, CW4 7HZ, United Kingdom.

Email: wendy@warr.com

Introduction

The virtual workshop took place on May 18-20, 2021. It was a follow-up from the December 2020 NIH Workshop on Ultra Large Chemistry Databases.[1] The organizers were:

- Marc Nicklaus, Head, Computer-Aided Drug Design Group, Center for Cancer Research, National Cancer Institute
- Gergely Zahoranszky-Kohalmi, Informatics Lead, ASPIRE, National Center for Advancing Translational Sciences, National Institutes of Health
- Eric Stahlberg, Director, Biomedical Informatics and Data Science, Frederick National Laboratory, National Cancer Institute
- G. Sitta Sittampalam, Senior Advisor to the Director, National Center for Advancing Translational Sciences, National Institutes of Health
- Janelle Cortner, Director, Data Management Program, National Cancer Institute

A major theme emerging from the December 2020 workshop was the fact that all the databases of a billion or more structures are virtual. For each virtual molecule the question then arises of whether, or how, it can be synthesized. The organizers therefore assembled speakers to give presentations about how reaction-related data are represented, captured, managed in databases, analyzed, used for drug design, applied in robotics, and exchanged locally as well as globally.

There were about 500 registered "attendees" from about 30 different countries. Maximum live attendance at any one time was about 220. This report summarizes talks from 27 practitioners in the reaction informatics field. The aim is to represent as accurately as possible the information that was delivered by the speakers; the report does not seek to be evaluative. The themes, in the order used for this report, were reaction representations, file formats, and standards; sources of reaction data; AI and machine learning applications of reaction-related data in *de novo* drug design, synthetic accessibility, synthesis planning, reaction prediction etc.; and automation and progression toward autonomous synthesis.

The CHMTRN and PATRAN languages for representing chemical reactions

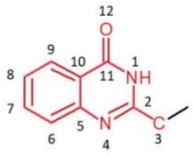
Philip Judson, Lhasa Limited, Leeds, United Kingdom

The CHeMistry TRaNslator and PAttern TRANslator (CHMTRN/PATRAN) languages were developed in the Logic and Heuristics Applied to Synthetic Analysis (LHASA) project at Harvard University, a pioneering project in artificial intelligence led by E. J. Corey. The project began more than 50 years ago as Organic Chemical Simulation of Synthesis (OCSS). Corey's team aimed to make a computer think as a chemist would. The work led to the retrosynthetic approach to synthesis planning, and a Nobel Prize to Corey himself.

By 1972, OCSS had become LHASA. The aim now was to make the computer both think and communicate as a chemist.[2] The graphical end-user input and output was unprecedented at that time. The architecture consisted of an interface, a reasoning (inference) engine, and knowledge base. The knowledge base is independent of the software, another novel idea for 1972. Data such as the reaction of ethanol with chloroethane to form diethyl ether, and similar reactions, can be converted to knowledge in the form of a generic reaction: alcohols plus alkyl halides can react to form ethers. In the retrosynthetic version, an ether is made from alcohol and alkyl halide precursors.

In practice, generic reactions do not work in all cases (e.g., an alkyl aryl ether cannot be made from an alcohol and a phenyl halide). A good knowledge base needs to understand such facts and express them as a chemist would. This was done with CHMTRN (pronounced “chemtran”), in which transforms are keyed by functional group(s) at the reaction center. The transformation to the precursors and qualifying statements about features favoring or disfavoring the retroreaction are described using an English-like language.

There are limitations to the use of functional groups. Chemists give names to only a relatively small number of reactive structural features, and LHASA, being constrained by the capabilities of computers at the time, could only include a maximum of 64 functional groups. Moreover, many reactions involve multiple atoms and bonds that are not normally thought of as functional groups. The solution was to use PATRAN (Figure 1). In practical terms CHMTRN and PATRAN are used as a single language in the knowledge base for LHASA, PATRAN being embedded in retroreaction descriptions written in CHMTRN.

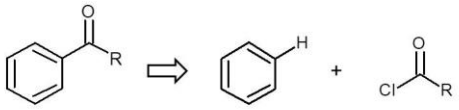


N-C(-C)=N-C%C%C%C%C%C(%@5)-C(=O)-@1

- Alternative atoms and bonds can be specified:
  -C,N,S- C=,%C etc.
- Properties can be attached to atoms and bonds:
  N[HS=1] C[ARYL=YES] C[ARYL=EITHER]
  C[FGS=LEAVING;FGNOT=BROMIDE,CHLORIDE]
  -[FUSION=DIALKYL] etc.

Figure 1. PATRAN example.

Judson showed an example of some CHMTRN statements (Figure 2). Originally, “rating” statements on the “quality” of a reaction were given as percentages. Nowadays, qualitative rating statements are written (e.g., TYPICAL*YIELD, EXCELLENT; RELIABILITY, GOOD; REPUTATION, GOOD; HOMOSELECTIVITY, FAIR; and HETEROSELECTIVITY, FAIR).



C[HETS= 1](=O)-C%C%C%C%C%C%@1

KILL IF ANYWHERE OFFPATH THERE IS A GOOD LEAVING GROUP
SUBTRACT 20 IF THERE IS A BULKY GROUP ALPHA TO ATOM*4 OR ALPHA TO ATOM*8

...
...
CONDITIONS AlCl3/80
...
BREAK BOND*2
INTRODUCE A CHLORINE ATOM ON ATOM*1

Figure 2. Some CHMTRN statements for a Friedel-Crafts reaction.

The Synthetically Accessible Virtual Inventory (SAVI)[3] is a database of over 1 billion compounds predicted to be easily synthesizable. The compounds have been created by a set of transforms based on an adaptation and extension of CHMTRN)/PATRAN. SAVI uses qualitative rating increments and decrements. For example, original statements said:

SUBTRACT 30 IF THERE IS A WITHDRAWING GROUP ON ATOM*5

IF ATOM*3 IS IN A RING OF SIZE 6 THEN ADD 15

but the qualitative statements are:

LOWER*RATING STRONGLY IF THERE IS A WITHDRAWING GROUP ON ATOM*5

IF ATOM*3 IS IN A RING OF SIZE 6 THEN RAISE*RATING SLIGHTLY.

The functional group properties of atoms ([FGS=...]; [FGNOT=...]) are currently limited in SAVI to the 64 functional groups declared in CHMTRN. It would make the language more powerful if more subpatterns could be used. Functional groups are already defined using PATRAN patterns in the knowledge base, so there is not an inherent language limitation.

Lhasa Limited's end-user interface for knowledge bases has always been graphical. Graphical alternatives have been used for knowledge base editing but textual knowledge base source code has the advantages of portability, long term stability, and the requirement for only a standard, simple, text editor. A dual scheme is an option. For example, the rule base for Harmoneus (a set of decision trees for chemical hazard classification) can be edited using a graphical interface but it can also be written to, and read from, a source text file.

The CHMTRN/PATRAN combination was the earliest language to be developed for chemical reaction knowledge bases. It was designed to be English-like language. It was not widely adopted outside the LHASA project and almost fell out of use. It is still the most powerful and flexible textual language for describing chemical reactions. The time is right to make new use of CHMTRN/PATRAN either by refining and modernizing it, or as the basis and inspiration for something new, or maybe for both options.

LHASA revival for forward-synthetic evaluation in CACTVS[4]

Wolf-Dietrich Ihlenfeldt, Xemistry, Glashütten, Germany

LHASA was originally designed to work retrosynthetically. The SAVI[3] team have re-implemented CHMTRN/PATRAN to use both original LHASA knowledge base entries, and encode new reaction knowledge developed after the 1980s. The software is operated in a scripting environment (Python or Tcl) and can be adapted to different tasks. SAVI uses the engine for forward reaction prediction to generate an ultralarge database of reaction products which are synthesizable in a single step (currently) from commercially available starting materials. Starting material scans, forward reactions, retrosynthetic scoring, and report generation are derived from a LHASA transform. Ihlenfeldt demonstrated these steps for a simple reaction: the Williamson ether synthesis (Figure 3).

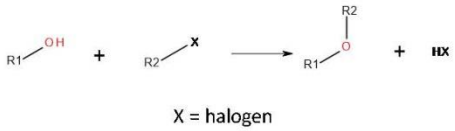


Figure 3. Williamson ether synthesis.

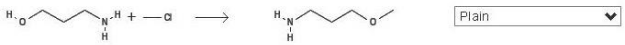
Ihlenfeldt showed methoxyethane as rendered (forwardly) by a naïve transform-based enumerator from SMIRKS.[5] He then showed the results of loading a byte-compiled version of the Williamson ether synthesis rule, and querying a few attributes. Reacting a starting material ensemble in a forward direction gives a duplicate-filtered list of products which can be formed according to the simple reaction pattern in the transform code. In this case, there is only a single product: methoxyethane, looking very much the same as the SMIRKS approach.

The intelligent part, however, is in the retrosynthetic scoring of the forward products in the LHASA approach. There are multiple retrosynthetic candidate retroreactions, and each one has its own score. The overall score is the best score of any of those. The candidate reactions in this case are the six combinations of methanol or ethanol with chloromethane, or bromomethane or iodomethane, but which one of the scored alternatives corresponds to the exact reaction used in the forward reaction? (Incidentally, the simple algorithm only works only for standard cases such as this ether synthesis. Some transforms are more complex, perhaps retrosynthetically attaching “superatoms” (e.g., “Alk” for “alkyl”), and then a more advanced substructure matching procedure is needed to match the retro reaction and forward reaction.) The reaction carries detailed information about how the score was derived. The code is not really human-readable but it can be visualized (Figure 4).

Transform	6039	Williamson Ether Synthesis		
Wikipedia				(no entry)
Literature (generic)				March (4th ed.), p. 386 Buehler, C.A. and Pearson, D.E., "Survey of Organic Syntheses" (Wiley-Interscience, New York, NY, 1970), p. 286 Patai (Ed.), Chemistry of the Ether Linkage, pp. 446-450; pp. 460-468
Literature (specific)				
Conditions				NaOR
Warnings				(no warnings)
Score	83			(Base Score 68)
[ID]/Side	**Delta**	**Base Operation**	**Condition**	**Rule Source Line**
[1] Product	+15	Add 15	specifiedatom*1 is/has hydrogenligs = 3	103 (library #1)

Figure 4. Visualization of a LHASA transform.

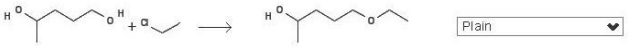
(Scoring for symmetrical ethers is slightly more complex, involving some filtering and automatic deduplication.) Nonparticipating functional groups can have an effect on retrosynthetic scores. Amines, for example, lower the score considerably (Figure 5).



Transform	6039	Williamson Ether Synthesis		
Wikipedia				(no entry)
Literature (generic)				March (4th ed.), p. 386 Buehler, C.A. and Pearson, D.E., "Survey of Organic Syntheses" (Wiley-Interscience, New York, NY, 1970), p. 286 Patai (Ed.), Chemistry of the Ether Linkage, pp. 446-450; pp. 460-468
Literature (specific)				
Conditions				NaOR
Warnings				(no warnings)
Score	43			(Base Score 68)
[ID]/Side	**Delta**	**Base Operation**	**Condition**	**Rule Source Line**
[1] Product	-40	Subtract 40	anywhere amine*1	65
[2] Product	-4	Lower*rating 4	anywhere amine*1	66
[3] Product	+15	Add 15	specifiedatom*1 is/has hydrogenligs = 3	103 (library #1)

Figure 5. Effect of amines on ether synthesis.

Phenols are more acidic than alkyl alcohols and allow more gentle reaction conditions: “pH9:10” is displayed under “Conditions” as a milder alternative (which indirectly has an effect on the compatibility evaluation of nonreacting functional groups). If there is another hydroxy group in the starting alcohol, and a nonsymmetrical ether is made, the reaction gets seriously downgraded (Figure 6).



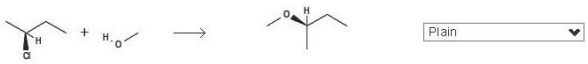
Transform	6039		Williamson Ether Synthesis	
Wikipedia				(no entry)
Literature (generic)				March (4th ed.), p. 386 Buehler, C.A. and Pearson, D.E., "Survey of Organic Syntheses" (Wiley-Interscience, New York, NY, 1970), p. 286 Patai (Ed.), Chemistry of the Ether Linkage, pp. 446-450; pp. 460-468
Literature (specific)				
Conditions				NaOR
Warnings				(no warnings)
Score	33			(Base Score 68)
[ID]/Side	**Delta**	**Base Operation**	**Condition**	**Rule Source Line**
[1] Product	-40	Subtract 40	specifiedatom*1 is/has not appendages	56
[2] Product	-4	Lower*rating 4	specifiedatom*1 is/has not appendages	60
[3] Product	+5	Add 5	specifiedatom*1 is/has hydrogenligs = 2	100 (library #1)

Figure 6. Effect of additional hydroxy group.

The transform is stereospecific, and if there is stereochemistry at the reaction center, it gets inverted (Figure 7).

```
print(l.ruleflags)
l.score('C[C@H](CC)OC')
display(HTML(filename=l.scoredreactions[3].X_LHASA_REPORT))
```

```
frozenset({'inverts*stereo', 'stereoselective', 'student'})
```



Transform	6039		Williamson Ether Synthesis	
Wikipedia				(no entry)
Literature (generic)				March (4th ed.), p. 386 Buehler, C.A. and Pearson, D.E., "Survey of Organic Syntheses" (Wiley-Interscience, New York, NY, 1970), p. 286 Patai (Ed.), Chemistry of the Ether Linkage, pp. 446-450; pp. 460-468
Literature (specific)				
Conditions				NaOR
Warnings				(no warnings)
Score	43			(Base Score 68)
[ID]/Side	**Delta**	**Base Operation**	**Condition**	**Rule Source Line**
[1] Product	-25	Subtract 25	specifiedatom*1 is/has hydrogenligs = 1	36 (library #1)

Figure 7. Stereo inversion in ether synthesis.

The above are some highlights in the top-level transform code, but there are also functions implemented in the various auxiliary library files. The library part is a generic inspection engine for nucleophilic substitution (S_N2) reactions which can detect what can go awry with these reactions. It is reusable for all reactions of that class. Some parts of the library code are never executed in this transform because of an implicit hard, basic reaction condition setting, as derived from the transform context. For the generic S_N2, there are about 280 lines of code, more than for the actual top-level Williamson transform. Ihlenfeldt presented some examples. The score for vinyl alcohol and

chloromethane, for cyclohexanol and chloromethane, and for *tert*-butanol and chloromethane is 83. Further examples are shown in Figure 8.

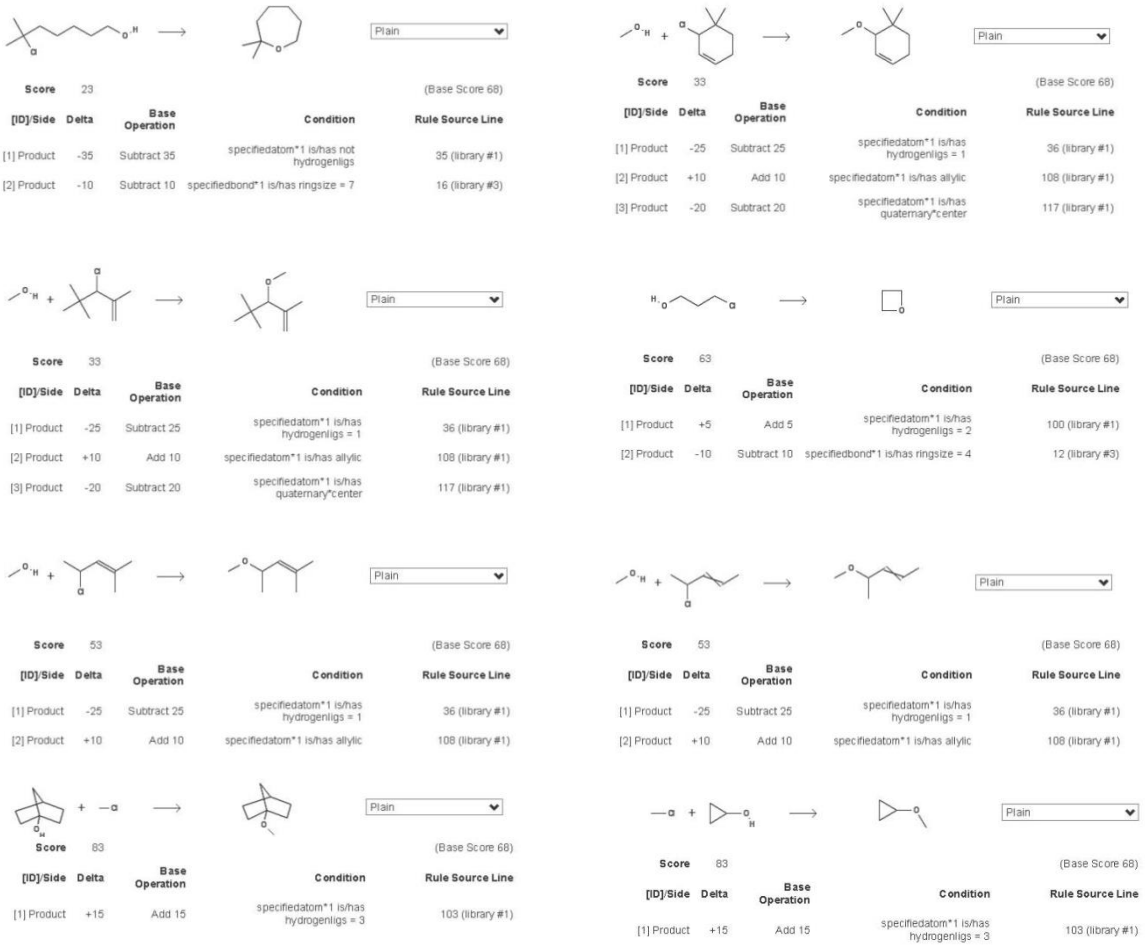


Figure 8. Further examples from auxiliary library files.

Examples cover cases of ring-specific compilations (bridgeheads, small ring formation, preferred elimination on 6-membered rings, and blocked inversion on small rings), steric hindrance, topological symmetry *versus* asymmetry, alternative S_N2' reaction mechanisms with allylic configurations and their topological equivalence or inequivalence with the S_N2 mechanism, interfering groups and their effect on selecting suitable starting materials (e.g., forcing use of I or Br if Cl is present and to be preserved), and dozens of other factors which are impossible to catch with a simple reaction template application mechanism.

These were the basics of how to run forward reactions, how to score those reactions, and how to check that the scored set both actually contains the forward reaction, and has an acceptable score. It is also possible to find suitable starting materials for these transforms to batch-generate libraries of synthesizable molecules. The LHASA object supports the generation of queries for reagents from the patterns for intermolecular or intramolecular reactions, and with constraints for the multiplicity of potentially reacting groups. Ihlenfeldt has set up an in-memory sample dataset for demonstration. In the standard case, these queries would be run on an SDfile,[6] or in the Xemistry[4] database cartridge.

In summary, Ihlenfeldt illustrated the basics of how to run forward reactions, how to score those reactions, and how to check that the scored set both actually contains the forward reaction, and has an acceptable score. He demonstrated these steps on a very simple reaction, and highlighted why these are much more powerful than the run-of-the-mill transform pattern application in standard library enumerators, especially with respect to the reliability of the proposed reactions.

Using different reaction formalisms to manage reactions and explore chemical space

Victorien Delannée, Hitesh Patel and Marc C. Nicklaus, National Cancer Institute (NCI) Computer-Aided Drug Design (CADD) Group, Frederick, Maryland, United States

Over the past 50 years, different reaction formalisms have been created for different specific purposes: unique identification, data exchange, classification, writing transforms, and models and predictions. Probably the first reaction format was used in Corey and Wipke's OCSS[7] in 1969. OCSS was based on rules to generate new molecules. Since then many different reaction formats have been created and are emerging to meet different needs. Delannée and co-workers have created two new open-source formats: ReactionCode,[8,9] and Smarts and Logic In ChEmistry (SLICE).[10]

ReactionCode is a new machine-readable format for reaction searching, analysis, classification, and transforms, and for encoding and decoding. The researchers considered the use of Reaxys BINCODE[11] and ClassCode[12] to retrieve reactions in a reaction database, "visualize" them, and make inferences. These representations are helpful for similarity search and classification, and they are fast and can be automated, but they are overly general (e.g., elements are grouped by atoms, such as group 1 instead of chlorine) and they cannot recover the exact reaction. Thus there was a need for ReactionCode, a new machine-readable format for reaction searching, analysis, classification, transforms, and encoding and decoding.[8,9] ReactionCode is a multilayer, machine-readable code, which aggregates reactants and products into a Condensed Graph of Reaction (CGR, see the following presentation by Alexandre Varnek). The pseudomolecule is encoded by layers, starting from the reaction center and moving to the extremities, and it is organized in three blocks (Figure 9).

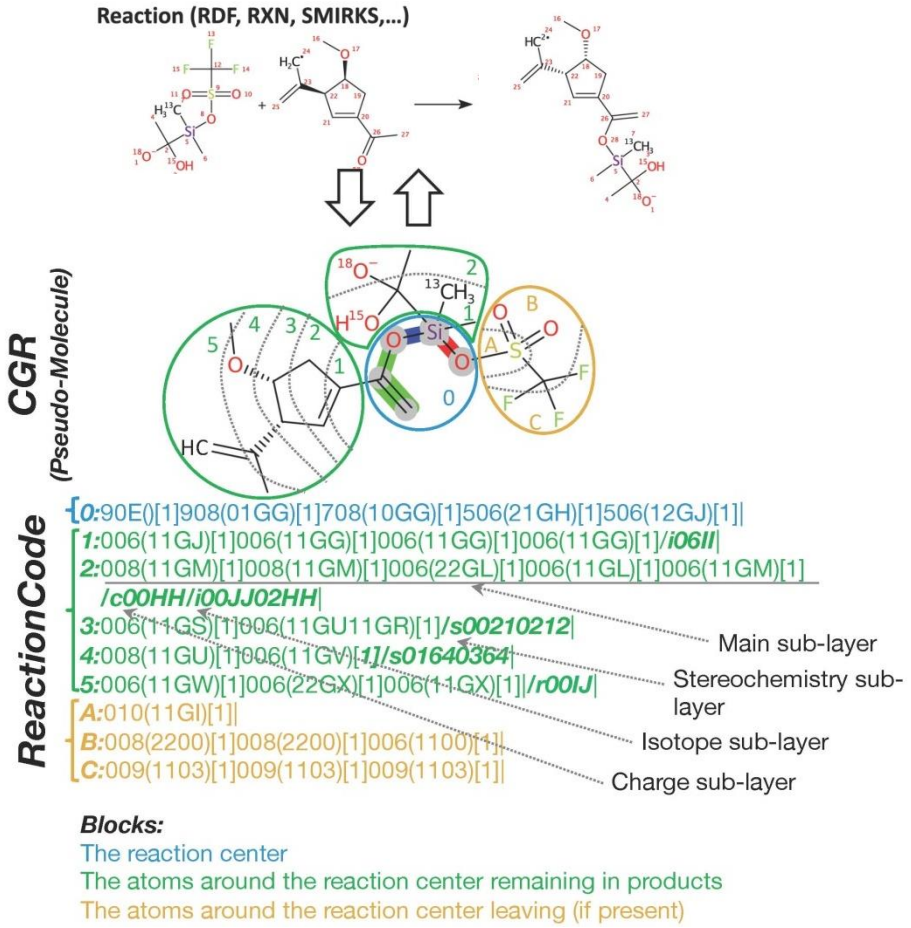


Figure 9. ReactionCode example.

ReactionCode is used for indexing, compression, and fast searching for similar reactions based on the reaction center and neighborhood similarities in a database context. It is helpful in graph databases. The team has used it in a diversity analysis of the 479,035 reactions in the USPTO database.[13] ReactionCode identified 9532 different reaction centers. In other words, the UPSTO dataset contains 9532 reaction types. The 10 most-represented reaction types are found in 203,776 (42.5%) of the reactions, and 90% of the USPTO dataset is covered by only 400 reaction types, which corresponds to 4.2% of all reaction types identified in this dataset. The NCI team also found that 4607 reaction types (48.3%) are only represented by one single reaction in the USPTO dataset.

ReactionCode is a versatile, open source[9] reaction transform language which can be used in reaction searching and identification, in classification and diversity analysis, and in machine learning. It can also be used to identify misannotated reactions and to correct unbalanced reactions.

In future, sublayers will be replaced by two auxiliary layers, one for charge, isotope and stereochemistry information, and the other for SMIRKS capability for properties, and adding flexibility. Next there will be a new stereochemistry system (with shorthand and ordered indexes based on the Cahn, Ingold, Prelog (CIP) rules) and bonds will be encoded using the number of electrons involved in the bond. Half-bonds, ionic bonds and any noncovalent sigma, pi, delta, single, double, triple, and quadruple bonds will be determined *a posteriori*. One- and three-electron bonds

found in radical compounds will be covered and so will three-center, two-electron or three-center, four-electron bond systems. ReactionCode will be integrated in SAVI.[3]

The second new open-source formats devised by Delannée and his colleagues is SLICE,[10] a new formalism for encoding chemistries to explore chemical space and move toward target-specific library generation. The team wanted to devise an open source and upgradable format to improve the speed of compound generation for both coders and noncoders. SLICE is based on CHMTRN/PATRAN, discussed earlier in this report.

A unique strength of CHMTRN is the use of a logic validated by chemists to predict both possible failure and success of reactions. It is a reasoning language combining FORTRAN-like syntax and English "buzz words" to describe chemical synthesis knowledge. Its capabilities include conditional statements (IF … THEN), control statements (FOR EACH), a scoring system, reusability (with no need for hardcoded, in-house solutions), and published documentation.

Unfortunately CHMTRN also has some drawbacks. It is an old, unstructured, and nonstandardized language (e.g., "KILL IF THERE IS AN AMINE*3" and "IF THERE IS AN AMINE*3 THEN KILL" express the same statement). It is slow: about 2 million CPU hours were needed to generate 1.75 billion products. It is also limited. There is no graphical PATRAN editor. CHMTRN is complex (a programming background is needed to use it) and there are no writing standards. It works retrosynthetically, which makes a forward-synthetic use more complicated, and is currently implemented only in the cheminformatics toolkit CACTVS[4] after the effective demise of the LHASA program.

The NCI team set the following specifications for the new language, SLICE. Firstly, it must be simple. It must be usable by novices without a programming background. It must have a GUI that is easy to learn and read. It must need minimal code, and use a structured and controlled language. Secondly, it must be powerful. It must be fast, bidirectional (forward and retrosynthetic), have advanced usage, and contain all CHMTRN/PATRAN functionalities, and more (e.g., variables, functions, external libraries, and operations). Thirdly, it must be open: open source, upgradable, compatible with multiple programming languages, and interoperable, allowing maximal CHMTRN to SLICE compatibility.

Writing a transform requires a chemistry description; a chemical pattern description language (SMARTS, SMIRKS, and ReactionCode); a reasoning language to encode chemistry rules; and a development environment requiring minimal code with a text editor using a template. The easiest way of writing a transform would be with a GUI and there is ongoing work on a new GUI based on JChemPaint.[14] Delannée presented a number of screen shots from this. A logic assistant (Figure 10) for writing statements (without the user having to know how to code) in the reasoning language is built around a simple structure inspired by "If This Then That" (IFTTT) and powered by Google Blockly.[15] The transform information is held in an XML file.

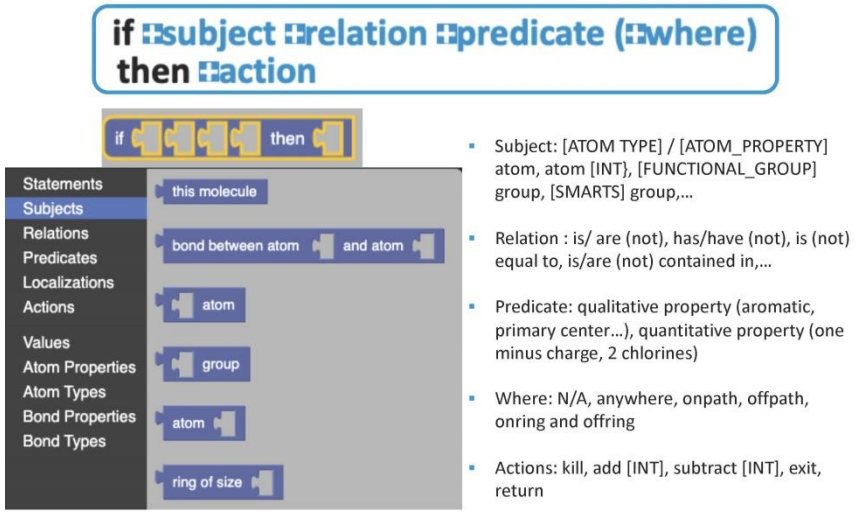


Figure 10. SLICE logic assistant.

In conclusion, SLICE is a language dedicated to chemistry. It is fast: 2000 times faster than the CHMTRN/PATRAN implementation. It can be used for reaction predictions and for bringing logic to SMARTS. It has a unique coding structure for novice users, with no compilation errors thanks to preconfigured blocks. Minimal code is needed. SLICE is open source, upgradable, and interoperable.

Condensed Graph of Reaction: an efficient approach to reactions mining

Alexandre Varnek[1,2] and Timur Madzhidov[3]; [1]University of Strasbourg, France; [2]ICReDD, Hokkaido University, Japan; [3]Kazan Federal University, Russia

Representing chemical reactions is much more complex than representing molecules: there are multiple species (principally reagents and products), reactions can be multistep, and the yield depends on experimental conditions. Condensed Graph of Reaction (CGR) is a pseudomolecule representing a given reaction (Figure 11). The concept of a superimposed reaction skeleton graph was introduced by Yuri Kiho[16] and George Vladutz[17] and was later reinvented by Shinsaku Fujita as an "imaginary transition structure".[18] Gérard Kaufmann and co-workers[19-21] at the University of Strasbourg called it CGR and used it in reaction classification, reaction rules, and synthesis design. Varnek later used CGR to describe reactions in descriptor-based chemical space and to develop machine-learning models.[22] A software package (CGRtools)[23] is publicly available for CGR manipulations such as canonicalization and standardization, signature calculation, substructure isomorphism, extraction of transformation rules, and reaction enumeration.

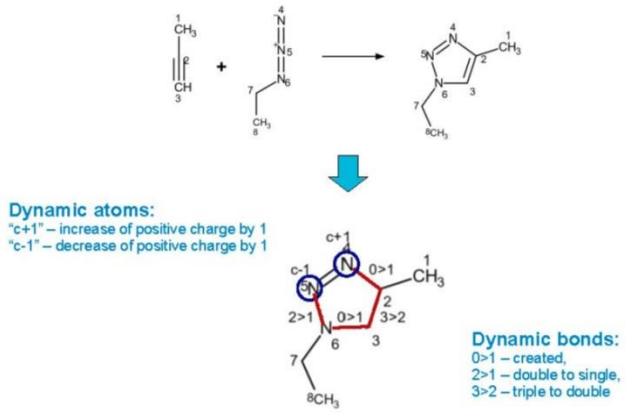


Figure 11. Condensed Graph of Reaction.

.

Reactions can be encoded as CGR/SMILES,[24] as CGR molfiles (CGR/MOL), as CGR *in silico* design and data analysis (ISIDA) substructural fragments for machine learning applications,[25,26] or as CGR hashcodes. For CGR/MOL, the standard fields (atom block and bond block) of an MDL CTfile are parsed and visualized by ChemAxon software[27] and a CGR block is added.

Varnek's team has published a workflow for standardizing, curating, and cleaning reaction data, followed by curation of transformations, and curation of reaction conditions and endpoints.[28] CGR is used in atom-to-atom mapping, reaction role assignment, removal of duplicates, and reaction balancing.[24] The team has trained a sequence-to-sequence autoencoder with bidirectional Long Short-Term Memory (LSTM) layers on SMILES/CGR strings, encoding reactions of the USPTO database[13] and has enumerated novel chemical reactions that are stoichiometrically coherent (balanced).[24]

Madzhidov, Varnek and co-workers[28] have reported a reaction standardization protocol followed by a comparison of some popular atom-to-atom mapping tools (ChemAxon's,[29] EPAM's Indigo tool,[30] RDTool,[31] NextMove Software's NameRxn[32] and RXNMapper[33] from IBM) and some consensus atom-to-atom mapping strategies. For this purpose, a dataset of 1851 manually curated and mapped reactions was prepared and used as a reference set. Success rate varies from 34% (for NameRxn) to 84% (for RXNMapper). Despite the fact that it has some disadvantages, RXNMapper was selected as the best tool, and it was applied to map the USPTO[13] dataset. Heuristic rules were used to correct erroneous mapping.[34]

In a method similar to that of InfoChem's CLASSIFY,[12] the teams of Varnek and Madzhidov consider the structural environment in spheres around the reacting centers. Broad classification uses reaction centers alone, medium classification uses reaction centers plus alpha atoms (excluding hydrogens), and narrow classification uses reaction centers plus alpha atoms (excluding hydrogens and consecutive sp^3 atoms). In the USPTO database, the researchers found 219,000 CGR motifs of "reaction centers plus alpha atoms" in 1.36 million reactions; 1063 motifs occur in ≥100 reactions.

Rakhimbekova *et al.* have studied the application domain[35] and cross-validation[36] of quantitative structure-property relationship models based on CGR descriptors. In work soon to appear in *Mendeleev Communications*, they have compared the use of ISIDA fragments and difference RDKit fingerprints[37] as descriptors in models for the reaction rate constant of Diels-Alder, S_N2, and E2

reactions and a tautomeric equilibrium constant; CGR descriptors were used in the top-ranked models.

In earlier work, a Ph.D. student in cosupervised by Varnek and Madzhidov built a structure-reactivity database of 10,000 manually prepared records for chemical reactions with annotated reactivity data and reaction conditions from Viktor Palm's handbook[38] and from Ph.D. theses. Madzhidov and Varnek's group have used this in the first direct QSPR modeling of equilibrium constants of tautomeric transformations in different solvents and at different temperatures, which do not require intermediate assessment of acidity (basicity) constants for all tautomeric forms.[39] A support vector regression method (SVM) was used to build the models with CGR descriptors. The training set consisted of 785 transformations belonging to 11 types of tautomeric reactions. The models obtained performed well both in cross-validation and on two external test sets. Benchmarking studies demonstrated that the models outperformed results obtained with a density functional theory (DFT) method and with ChemAxon's tautomerizer[40] applicable only in water at room temperature.

The teams have also used generative topographic mapping (GTM, a probabilistic extension of self–organizing maps)[41] to visualize, analyze and model the equilibrium constants of tautomeric transformations as a function of both structure and conditions.[42] The modeling set contained 695 entries corresponding to 350 unique transformations of 10 tautomeric types, for which equilibrium constant values were measured in different solvents and at different temperatures. The cross-validated balanced accuracy was close to 1 and the clusters, assembling equilibrium of particular classes, were well separated in 2D GTM latent space. Data points corresponding to similar transformations measured under different experimental conditions were well separated on the maps. SVM methods were compared.

Another case study is similarity-based assessment of optimal reaction conditions.[43] It is assumed that similar reactions proceed under similar conditions. For a given query, a tool searches the most similar reactions in a database and retrieves their reaction conditions (catalyst, solvent, temperature, etc.). The similarity is assessed using the Tanimoto coefficient for bitstrings computed for CGRs. CGR-based, in-house tools were used to process data for 142,111 catalytic hydrogenation reactions extracted from the Reaxys database.[44] These were filtered to 72,000 reactions with four functional groups and 67 protective groups. The protective groups were classified as "cleaved" or "remaining", depending on the reaction conditions. The models developed in the study showed high accuracy (~90%) for predicting optimal experimental conditions of protective group deprotection. Comparison of the results with Greene's manually prepared reactivity charts[45] showed that reactivity assignments for some protective groups in Greene's charts are erroneous or statistically poorly supported.

CGRs have also been used in *de novo* design of novel chemical transformations. Varnek's team has used GTM to explore the latent space of the SMILES-based autoencoders and generate focused molecule libraries.[46] They built a sequence-to-sequence neural network with bidirectional LSTM layers and trained it on the SMILES strings from ChEMBL23.[47] Very high (>98%) reconstruction rates of the test set molecules were achieved. Using GTM, the researchers visualized the autoencoder latent space on the two-dimensional topographic map. Targeted map zones can be used for generating novel molecule structures by sampling associated latent space points and decoding them

to SMILES. In later work,[24] novel latent space points were sampled around a map area populated by Suzuki reactions and decoded to corresponding reactions. Thirteen Suzuki-like reactions new with respect to the training data were detected and five of them have been found in recent publications.

Varnek concludes that CGR is an elegant way of reducing reaction complexity. It can be used efficiently in various cheminformatics applications such as encoding, building databases, reaction data curation and visualization, reaction classification, machine-learning and modeling, and *de novo* design of new transformations.

CSRML, an XML based transform reaction language

Tomasz Magdziarz, MN-AM, Nürnberg, Germany

Chemical Subgraphs and Reaction Markup Language (CSRML)[48] is an XML-based representation with a well-defined object model built around the concept of a chemotype. Chemotypes allow users to encode not only connectivity and topology but also various properties of atoms, bonds, electron systems, and whole molecules. Annotations and queries can be easily combined with Boolean operators. Hydrogen atoms are explicit. CSRML has built-in validation and rich metadata.

Metadata include XML comments; label, title, comment, description, revision, timestamp and author of the CSRML document; and label, title, comment, and description of other elements, including subgraphs, molecules, atoms, bonds, and reaction rules. The multiple hierarchy organization of subgraphs and reaction rules is shown in Figure 12.

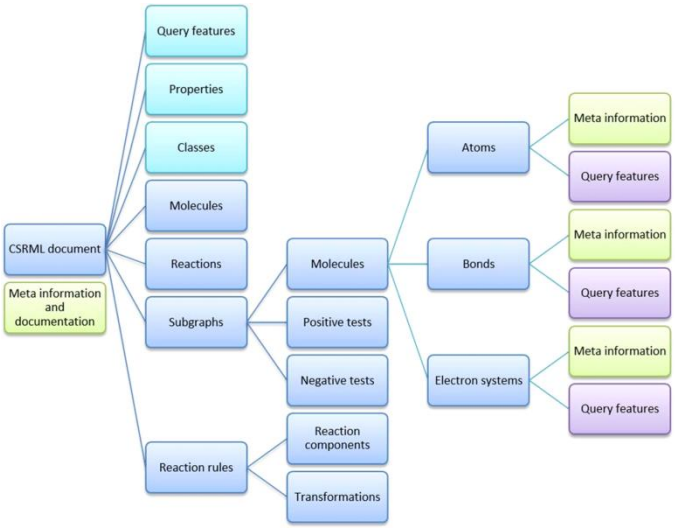


Figure 12. Object model of CSRML.

Matching and testing elements are included to assist in the design, development and validation of new chemotypes. In addition to conventional representations based on valence shell electron-pair repulsion theory, CSRML also provides a novel approach for querying electron systems for cases in which the actual order of the underlying bonds is irrelevant and what matters is the type of the bonding electron system (e.g., σ or π) and the number of electrons. Magdziarz gave an example of a nitro group attached to an aromatic atom, as in nitrobenzene where a maximum of 10 π electrons is allowed. A similar electron rule applied in SMARTS may fail or be more complex in cases such as *m*-dinitrobenzene where the 14 π electrons cause a mismatch.

CSRML reaction rules reuse subgraphs with all annotations, including physicochemical properties. There is an extensive set of transformation changes. Precise and accurate matchings and product generation are featured. The reaction rules have a hierarchy of reaction parts (roles, stoichiometry, atom mapping, and subgraphs (including validation)); transformations (bond types and order, atom property changes, and absolute or relative changes; and reaction examples (illustration and validation). Subgraphs can be defined in place or referenced. Magdziarz presented as an example the Diels-Alder cycloaddition reaction of a diene and a dienophile: a pericyclic reaction of six carbon atoms and six π electrons. In this reaction activated dienophiles react in relatively mild conditions, dienes need to adapt to a *s-cis* conformation, and electron-rich dienes are favored. Electronic effects lower the activation energy: electron-withdrawing substituents on the dienophile and electron-donating groups on the diene. Steric hindrance is another issue, involving bulky substituents on the dienophile, bulky substituents on the termini of the diene, and intramolecular repulsions in the diene.

CSRML takes a two-pronged approach to steric hindrance at the diene (Figure 13). In the subgraph approach only one of R or R' can be a small alkyl (CH_3), and one of R1 or R2 can be an even larger substituent such as *tert*-butyl. The second approach uses an atom hindrance property: a measure of the sum of volumes of neighboring atoms in a function of inverse squared distance. Termini of the diene cannot be sterically hindered and C-2 and C-3 cannot be simultaneously hindered.

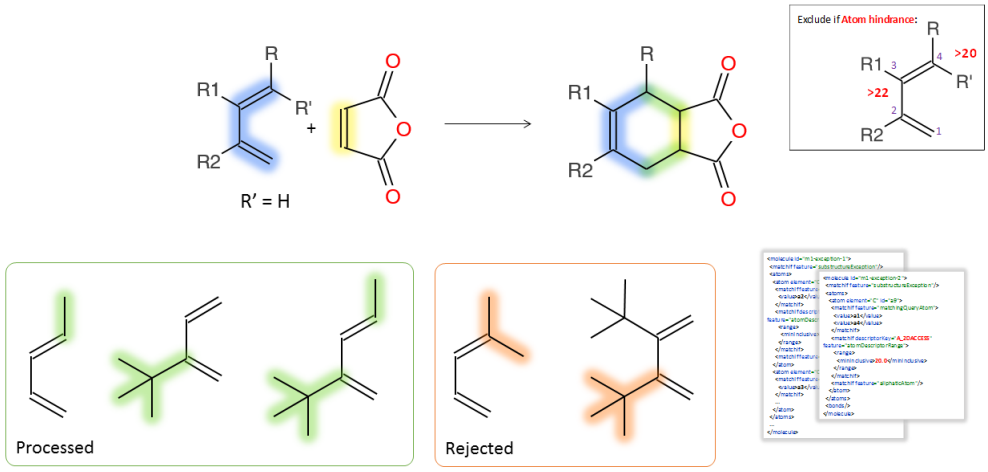


Figure 13. CSRML rule with exclusion of steric hindrance at diene.

Reaction rules can be used to generate reactions for matching reactants to predict, for example, the metabolites of a compound. The ChemTunes[49] database and knowledge base for safety evaluation and risk assessment is a knowledge base of *in vitro* and *in vivo* toxicity information. It comprises multiple components and workflows to support the safety and risk assessment of chemical compounds, including a database with expert quality control; the mechanism of action based ToxGPS prediction system for a series of human health and regulatory-relevant toxicity endpoints; and the Liver BioPath, a tool for human metabolism prediction. It is a metabolism prediction web service based on CSRML rules for phase I/II metabolism.

Magdziarz presented the example of safrole: a weak genotoxic hepato-carcinogen in mice and rats, exerting toxicity through metabolic bio-activation.[50] ChemTunes Liver BioPath predicted eight metabolites of safrole, the first six of which are reported metabolites. The CSRML rules applied are shown in Figure 14. Metabolites reported or not reported by Ioannides[50] are also indicated in the figure.

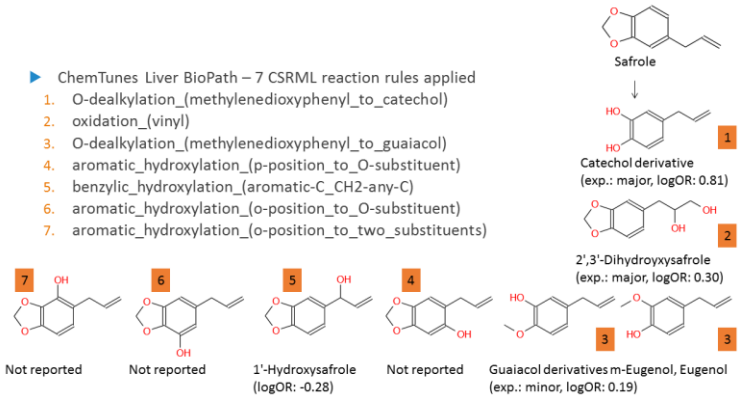


Figure 14. CSRML reaction rules applied to safrole metabolites.

CSRML is also used in ToxPrint,[51,52] a publicly available invariant reference set of structural features targeted to cover chemical structures from large toxicity databases and regulatory inventories. The ToxPrint chemotype library can be applied by using the ChemoTyper program.[53,54] CSRML also has an extensive set of literature-based rules covering all major tautomerism types.

In summary, the proper use of the connectivity and topology in combination with physicochemical properties, and the ability to define electron system queries can significantly elevate the accuracy of CSRML defined subgraphs. All defined subgraphs can be readily reused in CSRML reaction rules. Subgraphs are used to represent and match the chemical species, reactants or products, participating in a reaction. Both sides of the reaction equation are thus represented with chemotypes and together with a set of transformations allow users to generate accurate reactions for matching substrates and products. CSRML reaction rules have been used in metabolism prediction. CSRML also has applications in tautomerization, toxicity fingerprints, and many other fields.

Documentation and publication of reactions with Chemotion ELN and Repository

Nicole Jung, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

To facilitate and improve academic work on research projects, an electronic laboratory notebook (ELN) and a repository for research data have been developed as open-source software at the Karlsruhe Institute of Technology (KIT). The two systems can be used in combination or independently to plan, record, store, and disclose experiments or research data in chemistry.

ELNs are a key prerequisite to comprehensive documentation of research processes, the digital storage of research data, and their reuse. ELNs allow faster research processes and faster access to information. They enable researchers to store data in a standardized way and to manage research data based on automatically generated identifiers and descriptors. This is important for the single

bench chemist but it is even more important for the scientific community which can benefit from this resource of well-organized research data, if mechanisms for their disclosure are established.

Researchers do not supply the information to the community. Chemotion[55] addresses the challenge of digitization of chemistry:. There are 1800 reactions in the repository at KIT and they relate to 25,000 reactions in the ELN. In the Chemotion ELN, users can save reactions and samples, make calculations, and generate reports. Reaction-related data include free text entries as well as standardized descriptors. Literature references are also stored. Data can be imported with ChemScanner which extracts chemical information from ChemDraw files (.cdx, .cdxml, or cdx(ml) files containing .doc and .docx files). ChemSpectra allows users to view, edit and export spectra from mzML XML-based files for proteomics mass spectrometric data[56] and Joint Committee on Atomic and Molecular Physical Data (JCAMP-DX) files.[57] The spectral viewer function does not require any other software to be installed. Jung showed a movie of IR data analysis and uploading of IR data to the ELN. During the process, ontologies are applied and metadata are added.

Reactions and data can be exported and reports can be produced. Microsoft Excel lists based on SMILES identifiers (and soon, the IUPAC International Chemical Identifier, InChI)[58] and reaction information can be produced. Collections can be exported as zip files (reaction information as JavaScript Object Notation (JSON) and associated files). Whole Chemotion data collections can be exported as SDfiles,[6] .xlsx, .docx, and JSON. Videos of Chemotion functionality can be seen on the Chemotion website.[55]

The configuration of different report functions allows users to produce publications including chemical structures and a DOI link to the repository. Some parts of the ELN are machine readable and some taxonomies are included but others are neither standardized nor machine readable. There is a need to extend report functions for reactions and to support different standards (e.g., EnzymeML,[59] and the Unified Data Model (UDM),[60] described later in this report). Users can publish chemical structures, and attach characterization data and make them citable by DOI using the Chemotion repository[61] for molecules, reactions, and research data. Registration with a few scientific data providers is automated. Data from the repository are checked for input to PubChem.[62]

The next steps for machine readability include concepts to gain machine readable reaction descriptions and intuitive use for scientists, to enable transfer into protocols for machines, and to establish efficient workflows to mirror reactions to automated processes. KIT is investing about €4 million (euros) plus personnel through 2021 to 2022 to establish an automated platform for chemical synthesis. Partners and collaborators are welcomed; BASF is already a participant. Open hardware and software concepts are preferred.

Germany is funding its National Research Data (“Forschungsdaten”) Infrastructure (NFDI)[63] with over 85 million euros for the next 10 years. This covers research data management for all areas of science, represented by 30 consortia. The NFDI recognizes that digital data storage is an indispensable prerequisite for treating new research issues, generating findings and making innovations. NFDI4Chem[64] is the chemistry consortium in the NFDI. It is an initiative to build an open and findable, accessible, interoperable, and reusable (FAIR)[65] infrastructure for research data management in chemistry. Free facilities are now available. There are no excuses for not sharing data.

RXNO, the name reaction ontology

Colin Batchelor, Royal Society of Chemistry, Cambridge, United Kingdom

Thomas Hofweber in the *Stanford Encyclopedia of Philosophy* defines the larger discipline of ontology as having four definitions[66] but for the purposes of this presentation, an ontology is a machine-readable account of what things exist in a domain and the relations that necessarily hold between those things. Ontologies tell you what holds by definition and they help constrain what you can say about what is happening in particular. Good ontologies are tautologous, pedantic, trivial, and obvious. You can use the language of ontologies to label things in a database. An example is given in Figure 15.

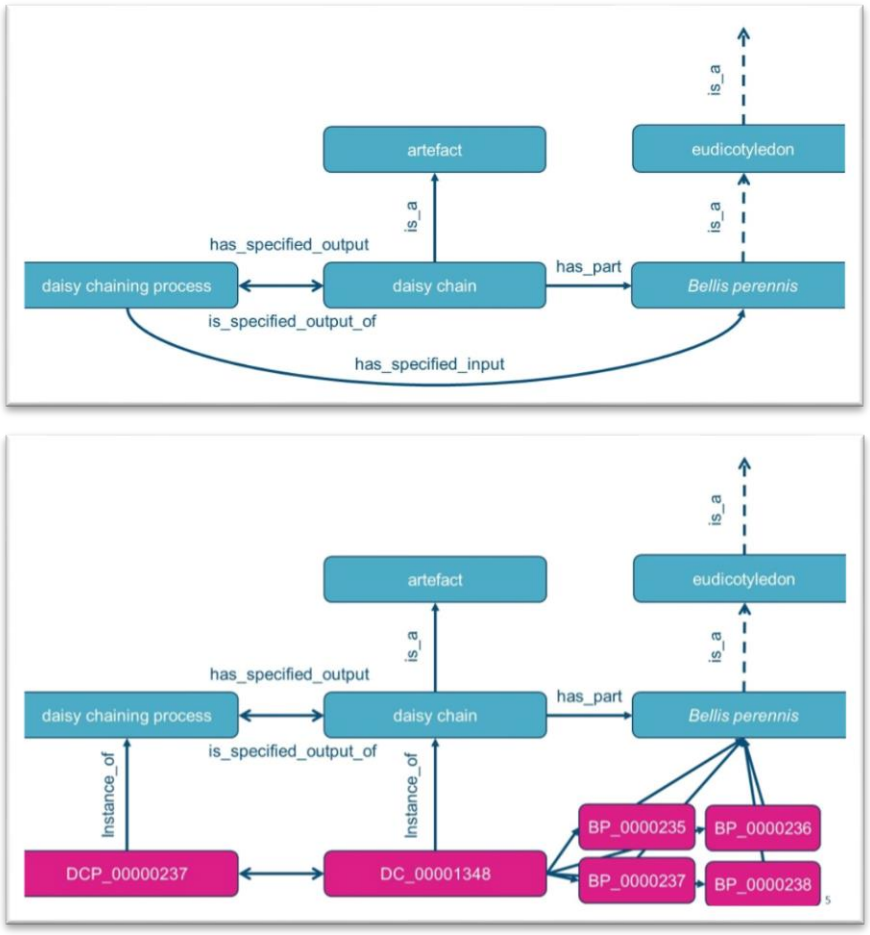


Figure 15. Example of an ontology.

Ontologies are useful in making data ready for use in AI applications. 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 (Figure 16). They offer typed relations for systematic correspondences (e.g., between methods and instruments, and between reactions and products). The Open Biological and Biomedical Ontology (OBO)[67] framework lets you use other ontologies to help build your own. It is "lighter weight" than Web Ontology Language (OWL).[68]. Figure 16 shows the application of OBO to the RXNO name reaction ontology,[69] a formal ontology of chemical named reactions. Ontologies give specifications for synonyms (exact, broad, narrow, and related) for use in

text mining. They are an important part of linked data projects and graph databases in the life sciences.

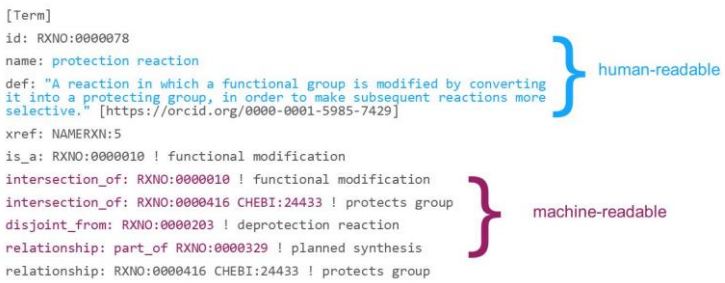


Figure 16. Reaction ontology example.

The Royal Society of Chemistry (RSC) was an early adopter of ontologies for annotating entities in text mining. RSC's Project Prospect[70] ran from 2006 to 2010, aiming at semantic annotation of chemistry articles, using manual annotation and text mining to output enhanced HTML, RSS feeds, and open-source ontologies. There are very well-established and understood methods for indexing and searching chemical structures, providing many chemists with the chief route into the literature. Chemical structures can be extracted at least semi-automatically from molecule names[71] and author-supplied graphics, in particular PerkinElmer's .cdx files, but chemistry papers talk about more than just chemical structures.

What ontologies might be used for chemical reactions? The IUPAC "color books", the world's authoritative resource for chemical nomenclature, terminology and symbols, might be considered:

- Green Book: Quantities, Units and Symbols in Physical Chemistry
- Red Book: Nomenclature of Inorganic Chemistry
- Blue Book: Nomenclature of Organic Chemistry
- Purple Book: Compendium of Polymer Terminology and Nomenclature
- Orange Book: Analytical Nomenclature
- Silver Book: Compendium of Terminology and Nomenclature of Properties Clinical Laboratory Sciences
- White Book: Biochemical Nomenclature
- Gold Book: Chemical Terminology

The Red Book does not give definitions but it does tell you how to name compounds. The Gold Book[72] is almost an ontology. It has definitions in the right form but the text is written for human beings rather than for machines. It is also available in XML and JSON. ChEBI,[73] a chemical ontology, was developed at the European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI) as a spinoff from the Gene Ontology.[74] It covers small molecules, subatomic particles, parts of a small molecule, biological roles, and applications. The RSC has developed three ontologies:[75] the chemical methods ontology (CHMO), describing about 3000 classes of experiments and instruments; the name reactions ontology (RXNO), describing about 600 classes of organic syntheses; and molecular processes (MOP), describing molecular processes in general (partly based on ChEBI).

The RXNO ontology unifies several previous attempts to systematize chemical reactions including the Merck Index and the hierarchy of Carey *et al.*[76] To devise RXNO, three chemists (two organic and one theoretical) took 100 name reactions (from personal knowledge and several compendia of named organic reactions) and decided on the principal axis of classification: the objective of the reaction. They developed an initial flowchart (Figure 17) and refined the classification in batches of 100 reactions. The objective of the reaction was chosen for classification because mechanisms are difficult to determine and can depend on reaction conditions. There is also no point in replicating what can be done with reaction fingerprints or reaction embeddings. Therefore organic reactions are represented based on the intent of the chemist, that is, what the chemist was trying to achieve with the reaction.

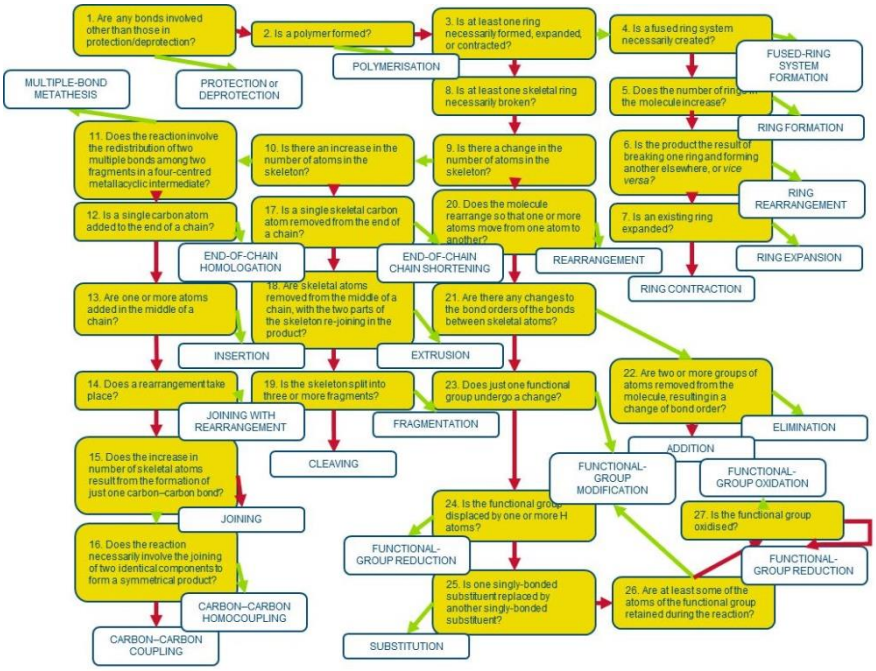


Figure 17. Development of RXNO.

Some further relations were used: “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. The ontology for a Diels-Alder reaction is shown in Figure 18.

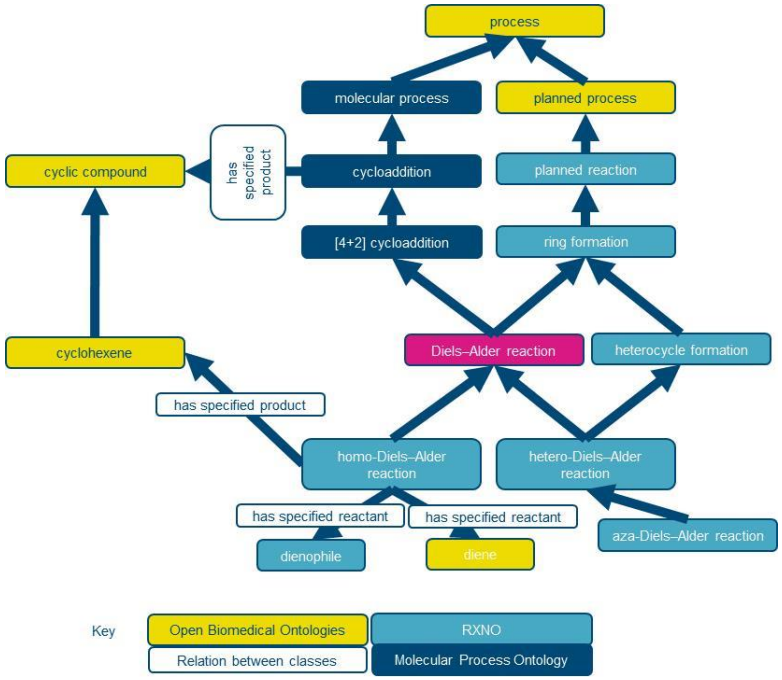


Figure 18. RXNO for the Diels-Alder reaction.

SMIRKS and generic Reaction IUPAC International Identifiers (RInChIs)[77] are not included in RXNO because the RSC found it very hard to get them right and the data structures implied by OWL were not a good fit.

RXNO is an open-source ontology, available in OWL and OBO formats, under a CC BY 4.0 license.[75] It has about 600 classes with full human-readable definitions and varying degrees of axiomatization. RXNO is used in the Unified Data Model[60] and in NextMove Software's NameRxn.[32] NFDI4Chem[64] have been using the ontology and have been improving the documentation and submitting new classes. There is also anecdotal evidence of use of RXNO inside big pharma.

Tracking reactions with the Reaction InChI (RInChI)

Jonathan Goodman, Cambridge University, Cambridge, United Kingdom

Goodman's research group studies reactions using experimental methods and computational techniques. The first approach gives rich data, rather slowly. DFT methods are accurate but also rather slow. The team uses them both to get quantitative information about reactions and also to develop qualitative models to help people understand reactions: what picture can you draw on your fume-cupboard to tell you which catalyst to use?

1,1'-Bi-2-naphthol (BINOL) is an organic compound that is often used as a ligand for transition-metal catalyzed asymmetric synthesis. A related reaction[78] is shown in Figure 19. BINOL has axial chirality and the two enantiomers can be readily separated and are stable toward racemization. Lou et al.[79] reported the asymmetric allylboration of ketones using 3,3'-Br_2-BINOL; the reaction products were obtained in good yields and high enantiomeric ratios. Goodman's team has used DFT calculations[78] to establish the identity of the reacting chiral species. The results show that a cyclic Lewis acid-activated boronate is the most reactive species on the basis of calculated energy barriers, and it is

only this species that leads to the correct enantiomer. The stereoinduction can be rationalized in terms of the competing chair-like transition structures.

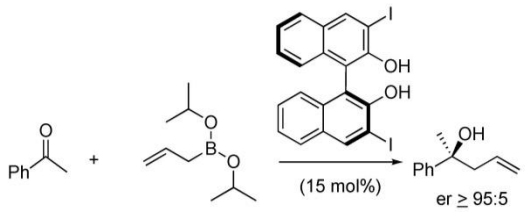


Figure 19. Asymmetric allylboration of ketones.

Chiral phosphoric acids have become powerful catalysts for the stereocontrolled synthesis of a diverse array of organic compounds.[80] Catalysts based on the BINOL-derived phosphoric acid scaffold show a large structural diversity, especially in the 3,3' substituents, and little is known about the molecular requirements for high selectivity. As a result, selection of the best catalyst for a particular transformation requires a trial and error screening process, as the size of the 3,3' substituents is not simply related to their efficacy: the right choice is neither too large nor too small. Goodman's team has developed computational approaches to identify and quantify structural features on the catalyst that determine selectivity.[80,81]

A large number of organic reactions feature post-transition-state bifurcations.[82] Selectivities in such reactions are difficult to analyze because they cannot be determined by comparing the energies of competing transition states. Molecular dynamics approaches can provide answers but are computationally very expensive. Goodman's team has reported an algorithm that predicts the major products in bifurcating organic reactions with negligible computational cost. It requires two calculated transition states, two product geometries, and no additional information The algorithm is quick and simple to run and, except for two reactions with long alkyl chains, calculates selectivity more accurately than transition state theory alone.[83]

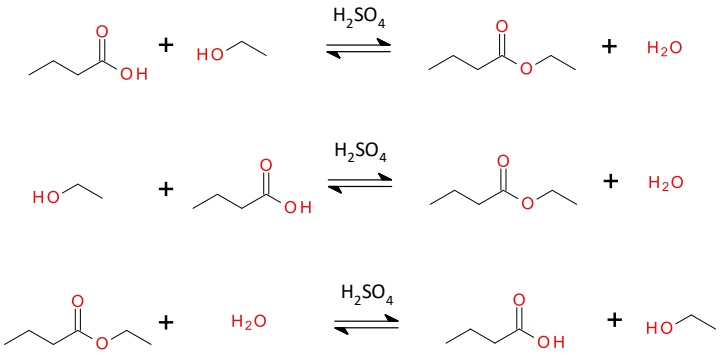
Goodman's team has also reported a system for automatic processing and assignment of raw ^{13}C and ^{1}H NMR data. The system, DP4-AI,[84] has been integrated into a computational organic molecular structure elucidation workflow. It maintains the same high rate of correct structure elucidation as DP4 using NMR descriptions written by an expert chemist but has achieved a 60-fold increase in processing speed, and near-elimination of the need for scientist time, when rigorously evaluated using a challenging test set of molecules.[85]

All these examples illustrate that the more we learn about reactions, the more we learn about their complexity. To understand them, we need as many data on reactions as possible but the data are useful only if we can analyze them using the resources that we have available. RInChI is an important tool for addressing this challenge.[77] InChI[58] is an identifier for molecules; RInChI is an identifier for reactions. RInChI enables users to make connections between reactions from different data sources, even when the number of reactions being considered is very large. It was designed to be easy to construct and to use, to be based on InChI, to be canonical, and to be based only on the relevant reaction, not on a central authority.

Like InChI, it has a layered structure (Figure 20):

- Layer 1: RInChI version, and underlying InChI version
- Layer 2 and 3: starting materials and products
- Layer 4: solvents, catalysts, and other “stuff” which survives the reaction
- Layer 5: direction: d+, d-, d=, (dx, for failed), or unspecified
- Layer 6: count of no-structure materials.

Nothing else is included. For comparison, reaction SMILES handles only reactants, reagents, and products, and is not canonical.



RInChI=1.00.1S/C2H6O/c1-2-3/h3H,2H2,1H3!C4H8O2/c1-2-3-4(5)6/h2-3H2,1H3,(H,5,6)<>C6H12O2/c1-3-5-6(7)8-4-2/h3-5H2,1-2H3!H2O/h1H2<>H2O4S/c1-5(2,3)4/h(H2,1,2,3,4)/d=

Layer 1, Layer 2, Layer 3, Layer 4, Layer 5
Layer six is omitted, because all the molecules involved have structures

Figure 20. RInChI example: three different representations of the same equilibrium all generate the same RInChI.

Hashed representations of RInChIs (RInChIKeys) are suitable for database and web operations. The current version of the RInChI code provides three options containing successively less information: Long-RInChIKeys, Short-RInChIKeys and Web-RInChIKeys. Web-RInChIKeys deduplicate InChIs over all groups and hash all major and minor InChI layers into a fixed length string ignoring the specific role of the reaction components.

Goodman presented the results of preliminary tests on RInChI using the SAVI database of more than a billion reactions.[3] He downloaded 1,748,464,003 SAVI-generated products and reactions on April 6, 2021. He found 1,094,782,440 Web-RInChIKeys of which 1,050,824,321 were different. Hash collisions are possible, but unlikely in a database of this size. Most of the duplicates are present as a pair but one Web-RInChIKey (IPWKGWBMOUVREXHYK-NUHFFFADPSCTJSA) is present 12 times: there are 12 different SMILES for the product (Figure 21) and its tautomers. He calculated that there are, at most, 1,094,782,429 reactions in SAVI.

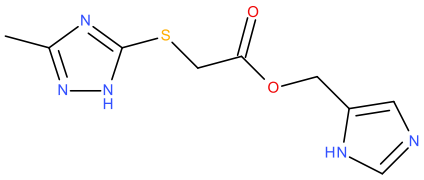


Figure 21. Web-RInChIKey: IPWKGWBMOUVREXHYK-NUHFFFADPSCTJSA.

Goodman concludes that RInChI is complicated enough to be useful. It is good at linking reactions and good at differentiating reactions. RInChIs are indeed easy to use, based on InChI, canonical, and based only on the reaction, not on a central authority.

Papers do not usually report negative results, even though these can be the key to understanding reactivity and the limits of the applications of new synthetic methods. Goodman is working with Prof. Simon Woodward at the University of Nottingham to develop a standard set of conditions which will be applicable to as many reactions as possible. These will not always be successful, of course, but researchers may be more ready to report a reaction which failed because of the constraints on the conditions than one which failed despite attempts to optimize the conditions, and robots will record the outcome of the standard conditions before optimization processes. It is hoped that this will lead to more reporting of reactions which did not deliver the anticipated outcomes.

New RInChI features are being developed, including ways of recording generic reactions and atom mapping, which should lead to the easier grouping of reactions into related classes using the RInChI. The RInChI should be an effective tool in promoting the use and reuse of open data in the discovery of new reactions.

Structure transformations with Ambit-SMIRKS

Nikolay Kochev[1,2], Svetlana Avramova[2], Nina Jeliazkova[2]; [1]University of Plovdiv, Department of Analytical Chemistry and Computer Chemistry, Plovdiv, Bulgaria, [2]Ideaconsult Ltd, Sofia, Bulgaria

AMBIT[86] began in 2005 as part of the European Chemical Industry Council (CEFIC) Long Range Initiative. It is built on top of the Chemistry Development Kit (CDK).[87] It has many cheminformatics modules but the current presentation concentrated on Ambit-SMARTS[88] and Ambit-SMIRKS.[89] SMiles ARbitrary Target Specification (SMARTS) is a language used for describing molecular patterns and properties. Its rules are straightforward extensions of SMILES. The language SMIRKS is defined for generic reactions. It is a hybrid of SMILES and SMARTS and a restricted version of reaction SMARTS involving changes in atom-bond patterns.

Ambit-SMIRKS has functionality for parsing of SMIRKS linear notations into internal reaction representations based on the CDK objects; application of the stored reactions against target (reactant) molecules for actual transformation of the target chemical objects; reaction searching; stereo information handling; product postprocessing, etc. The transformations can be applied on various sites of the reactant molecule in several modes: single, nonoverlapping, nonidentical, nonhomomorphic or externally specified list of sites using an efficient substructure searching algorithm. Ambit-SMIRKS handles the molecules' stereo information and supports basic chemical stereo elements implemented in the CDK library. The full SMARTS logical expressions syntax for

reactions specification is supported, including recursive SMARTS expressions and additional syntax extensions. The architecture is shown in Figure 22.

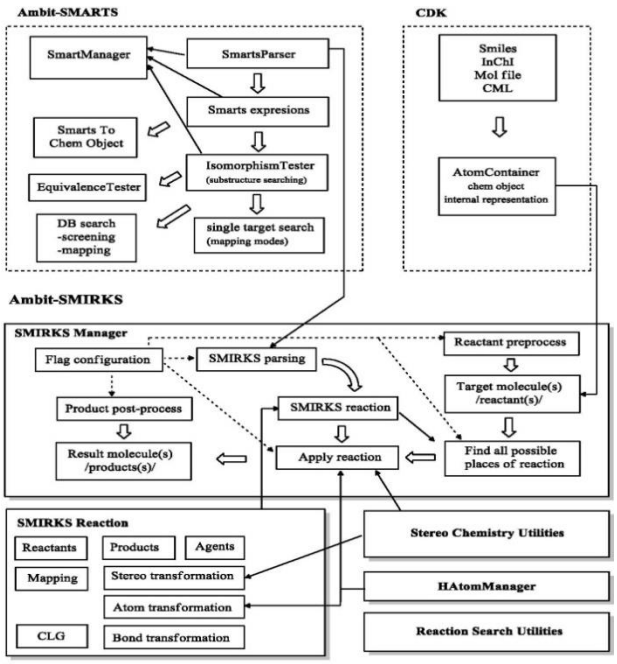


Figure 22. Ambit-SMIRKS architecture.

Ambit-SMIRKS is distributed as a Java library under a Lesser General Public License (LGPL).[90] SMIRKSManager and SMIRKSReaction Java modules can be fine-tuned with a set of flags to carry out various functions some of which Kochev discussed.

The reactant part of the SMIRKS linear notation is used as a definition of a SMARTS substructure search query, where the mapping indices are ignored. Ambit-SMIRKS uses the substructure search implementation of Ambit-SMARTS to find the reaction sites. The substructure searching can be performed in several modes: single, non-overlapping, non-identical, nonhomomorphic, or externally specified list of sites (Figure 23). A reaction transformation according to the substructure match modes is shown in Figure 24.

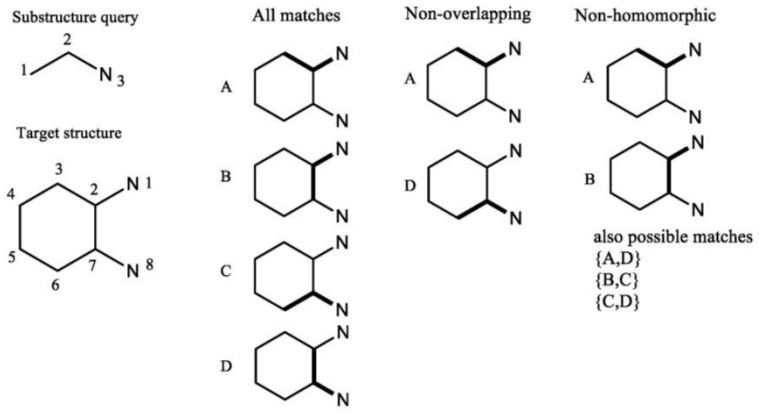


Figure 23. Substructure search match in various modes for cyclohexane-1,2-diamine.

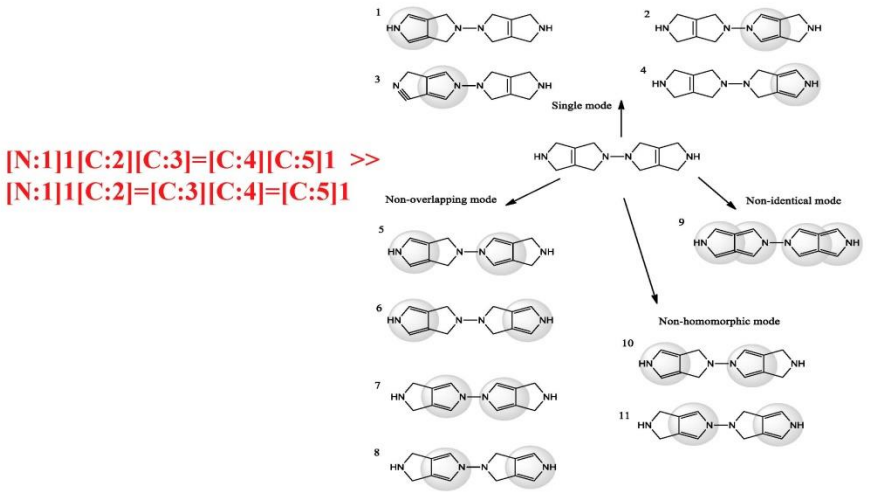


Figure 24. Modes of reaction application.

The SMIRKS linear notation supports atom mapping. Unmapped atoms on the reactant side of the SMIRKS, and all bonds incident to unmapped atoms are removed from the resulting products. Unmapped atoms on the product side of SMIRKS are created and added to the resulting products and the corresponding new bonds (from the unmapped atom to other atoms) are created as well. In all other cases where atoms are "rearranged" by changing, adding or removing bonds, obligatory usage of mapped atoms is considered a good practice. An incorrect usage of unmapped atoms leads to side effects and "strange" or incorrect application of the reactions SMIRKS. Even if specifying syntactically correct SMIRKS, the chemical logic when using unmapped atoms is different and Ambit-SMIRKS will follow exactly the transformation logic. For normal chemical transformations, SMIRKS atom mapping is needed and within the Ambit-SMIRKS module it is considered as a good practice.[89]

An exception to the recommended practice of fully using mapped atoms is the case of explicit H atoms. The majority of cheminformatics software systems, as well as Ambit-SMIRKS, handle the hydrogen atoms in two ways: as implicit H atoms described as attributes to other heavy atoms and as explicit H atoms which are treated as normal heavy atoms. Usually, the implicit hydrogen atoms approach is preferred but using explicit H atoms in SMIRKS transformations allows robust and more precise description of the chemical reaction logic. The three main scenarios of H atom treatment within Ambit-SMIRKS software are handling H atoms explicitly, or handling H atoms implicitly or explicitly with an H atom transformation option, or handling H atoms "automatically".[89] Handling H atoms automatically is not recommended but there are ways of doing it, with postprocessing clean-up.

Cheminformatics systems handle aromaticity in two major ways: by Kekulé resonance structure representations and by delocalized aromatic systems, typically represented by aromaticity flags of atoms and bonds. The aromaticity information within SMIRKS is primarily used to define the substructure searching queries for identification of reaction transformation sites. Making use of such information (particularly within the product side of the SMIRKS) to define aromatic system transformations is quite challenging. Within Ambit-SMIRKS, it is considered a good practice to handle aromatic transformation as Kekulé structures, since in this way all bonds orders are defined explicitly and the SMIRKS transformation of the bonds is clearly defined as well. After applying a

reaction rule, Ambit-SMIRKS performs a postprocessing aromaticity detection algorithm and if aromatic systems are formed due to the bond changes, the aromatic atom and bond flags are assigned accordingly. The resulting molecules could be represented in aromatic form or stay in a Kekulé form.

Some may consider the need to rely on particular aromaticity detection algorithm a disadvantage. This is only a reasonable point when the cheminformatics system lacks a good aromaticity detector. Ambit-SMIRKS relies on the CDK aromaticity detector which has been significantly improved in the latest releases of CDK. When users prefer their own aromaticity detector, Ambit-SMIRKS offers another option.

The stereochemistry in cheminformatics systems is represented in two main ways. In one, the absolute stereochemistry approach describes the elements of the stereo group by ordering the stereo elements on the basis of absolute chemical logic that does not depend on the atom numbering. For example, the CIP priority rules are used in some cases of computer representation and handling of molecular stereo information. The widely used approach for stereo handling on a topological level is the so called relative stereo representation. The relative stereo approach is used for the CDK-based internal stereo representation of a molecule, as well as for the molecule SMILES.

The SMILES, SMARTS and SMIRKS notations are based on the relative stereo approach, which is used to describe the stereo configurations in molecules, search queries and reactions. The stereo element priorities within relative approaches depend on the atom numbering and thus influence the algorithms of atom iteration used to define the sets of stereo elements. The priority of the stereo elements in the case of SMIRKS, SMARTS or SMILES is defined by the order of appearance in the linear notation, which is equivalent to usage of random atom numbering.

Ambit-SMIRKS stereo handling is based on the relative approach for stereo information representation, as both the SMIRKS linear notation and the internal CDK objects are based on it. The major types of stereo elements supported by CDK library are tetrahedral chiral atoms, *cis*/*trans* double bond configuration, and allene atom chirality. Ambit-SMIRKS stereo transformation functionality depends on whether or not the stereo transformation is directly specified by SMIRKS. If the stereo information is not directly specified by SMIRKS, Ambit-SMIRKS handles three cases: stereo element preservation, stereo element change of ligand, and stereo element removal. If the stereo information *is* directly specified by SMIRKS, Ambit-SMIRKS handles three cases: creating a new stereo element, stereo element update or change, and stereo element removal.[89]

Since its initial development for the purpose of metabolite generation within Toxtree (a method for toxic hazard estimation using a decision tree approach),[91] the Ambit-SMIRKS module has been used in various cheminformatics projects, both developed by the authors of the package and by external teams. These include enviPath,[92] a database and prediction system for the microbial biotransformation of organic environmental contaminants; BioTransformer,[93] a software tool that predicts small molecule metabolism in mammals, their gut microbiota, and the soil and aquatic microbiota; ExCAPE-DB,[94] a chemogenomics resource of over 70 million structure-activity relationship data points from PubChem[62] and ChEMBL, [47] and GLORY, a generator of the structures of cytochrome P450 metabolites.[95] The Ambit-SMIRKS GUI[90] and the Ambitcli[90,96] command line Java application for processing chemical files, structure standardization, import into an AMBIT database and processing AMBIT database entries are publicly available.

Reaction SPL documents

Gunther Schadow, Pragmatic Data, Indianapolis, IN, USA

The purpose of a reaction data format is to share reaction knowledge and data between different people and systems. Many people and systems have created reaction data formats; each having their own point of view and specific purpose and limitations. The idea of standardization is that *many* points of view and purposes are represented and can relate to each other, without an insistence on details some just do not care about, and without inhibiting others to express all their details. We should want a format that support the full life cycle without barriers among R&D experiment, publications, patents, documentation and control of the production process, regulatory applications and monitoring, and trade and logistics.

Data are important but language has infinite expressiveness. Humans like documents with their freedom of unconstrained expression. Data are best carried in the documents where they originate because information extraction and data mining are hard and error prone; it is a chore for people to have to enter data into other systems; and data entry to other systems divorces data from the original source. Databases with "comment" fields are not as useful because the original train of thought is butchered. Rich text document support is important for expressivity. Data should not be divorced from text. If you have a downstream system that requires data alone to be shared, and cannot handle user text input, the text can simply be excluded for that system, but the users should not be deprived of a place to express themselves completely.

Structured Product Labeling (SPL)[97] is a robust yet fairly light XML document standard. It uses a highly generic but usefully refinable data schema, which is, like a language, highly expressive. It originated in medicine but almost all use cases have been examined in great detail: people, organizations, products, and devices; science and measurements, including complex data, waveforms, and imaging; missing data and uncertainty; workflows, protocols, and processes; and scale from geography down to organization, building, devices, substances, molecules, and even subatomic structures, if need be.

Schadow has worked in health and life science data interoperability for over 30 years and created convenient and useful solutions, including major parts of the HL7 standard[98] which comes from the medical domain but which includes all scientific data types. The Unified Code for Units of Measure (UCUM), adopted by HL7, has been the recognized, *de facto* standard for units for over a decade. Schadow has also created a comprehensive standard of essential real world responsive data types that was ratified by the International Standards Organization (ISO), and a generic, powerful "language"-based HL7 Reference Information Mode (RIM). He has applied all these tools to the SPL standard for the U.S. Food and Drugs Administration (FDA), for regulated products and substances, for over a decade. The FDA Pharmaceutical Quality/Chemistry, Manufacturing and Controls (PQ/CMC) terminology is ready for use.

Many models and schemas for reaction representation (RInChI, CSRML, the UDM, and the Open Reaction Database (ORD)) are discussed elsewhere in this report. Chemical Markup Language (CML)[99] and the reaction molfile[6] are not. CHMTRN/PATRAN (discussed earlier in this report) is a domain-specific programming language. Software systems have been written by StructurePendium (related to RInChI and UDM, also discussed in this report), KIT (see Chemotion above) and NextMove Software (see the discussion about Pistachio below).

UDM, CSRML, and ORD extract data on the data package; reactions; molecules; analysis; and basic data types, including physical quantities, and units (scalar, vector, matrix, etc., and string, Boolean, text, date and time stamps, and ID). The data package records author, version, bibliography, literature references, and attached documents. Reaction data for UDM, CSRML, and ORD include a representation; conditions (in the case of UDM and ORD); properties perhaps; set-up and work-up in the case of ORD; and not usually atom-mapping. (Note that CGR carries atom-mapping.) UDM' reaction representation encapsulates RXNfile[6] or RInChI.[77] CSRML models reaction rules, and goes from generic entities to specific experiments and processes. UDM represents reactants, products, reagents, catalysts, solvents, etc. UDM and ORD encapsulate various molecule formats as a string; CSRML has substructures, references, atom, bond, and electron system. Analytical data are carried in UDM as name-value pairs for "observations, property"; in ORD, they are embedded in "Outcomes: Analysis, ProductMeasurement". No system handles unintended reactions and impurities well.

Schadow gave his own impressions from analysis of UDM, ORD, CSRML. All of them have some aspects of documents with author, version, etc. but none has free, mixed, content-rich text (except CSRML to some extent, perhaps). In UDM and ORD, chemical details are encapsulated within only higher level structures; CSRML has a detailed chemical model for molecule, substructure, atom, bond, and even electron system. ORD takes a very experimental point of view with out*comes*, not out*puts*, and the identity of the output molecules is only the result of analytic observations (*a posteriori*). UDM takes a more *a priori* approach with chemical structure assertions of inputs and outputs, whereas CSRML is very oriented toward chemical structure *a priori*. A universal standard should accommodate all these aspects as desired by the author of any particular document.

An ontology describes what an entity *is* (and does), whereas an epistemology concentrates on how we know what it is. Epistemology deals with observable analysis of something, for example, the inferred molecule; or change as seen in observable properties, for example, the inferred mechanistic model. We need to consider a thing *versus* the property of a thing *versus* observations and measurements about the thing.

In Schadow's philosophy, a *substance* (chemical entity) can be thought of as a molecule or a material. A molecule is an abstract concept, only known in quantities, moiety, atoms, bonds, and electrons. A material exists on a macroscopic scale: we can experience instances of materials (e.g., this bag of NaOH pellets) not just abstract kinds. Materials are substance plus form, some forms can be counted (e.g., tablets) other measured in mass or volume (e.g., powder, liquid). Materials are created by a maker, through some process, in a certain quality or purity (or impurity). *Reactions* are change, or (inter)action among substances. On a molecular scale, reactions have a theoretical, physicochemical mechanism; on a macroscopic scale, they are performed as an experiment or manufacturing process. Process *versus* process step, and reaction *versus* reaction step are considered. There is also a modality (or "mood") dimension of reactions and processes as defined, planned, executed, observed, or hypothesized. Reactants, products (including reversibility), solvents, catalysts, etc. are all "interactors" which participate in the reaction. They are called "participations" (some people call them "roles"). Other participants are macroscopic vessels and instruments (devices). Specification, analytics, and properties of things are considered as opposed to reactions, conditions, and control parameters.

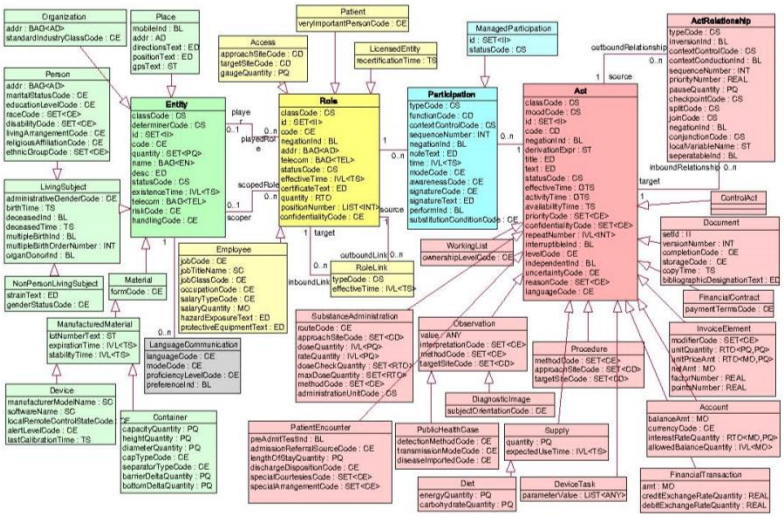


Figure 25. HL7 Reference Information Mode (RIM).

Schadow showed a typical HL7 RIM (Figure 25) and a detailed schema for how the concept might be applied to a chemical process (Figure 26). He demonstrated the XML document package, showed rendering and validation, and ran through the XML in some detail to indicate how practical and applicable this solution is.

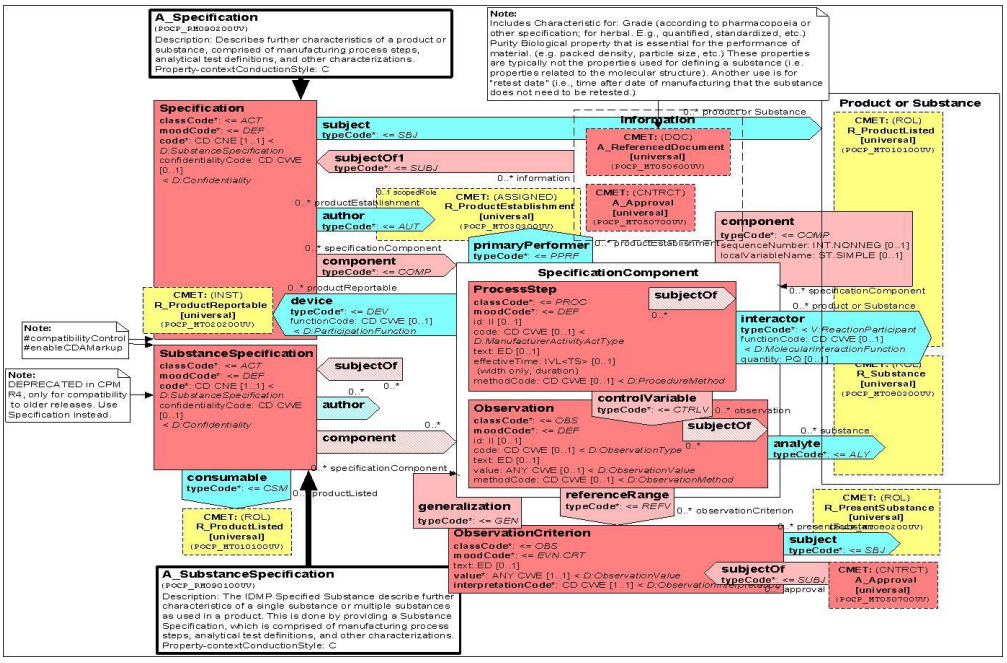


Figure 26. SPL schema for a chemical process.

SPL is practical. The drug industry has experience with it. SPL has proven ability to clean up databases. It has tools: XForms, W3C XML Schema and Schematron, rendering stylesheet (XSLT), and database import. A draft specification for PQ/CMC was developed for the FDA in 2017 and is ready for use when somebody wants an actual solution. Schadow has been working on standardization for 30 years, and on SPL since 2003; Pragmatic Data LLC has existed since 2008 and they have implemented dozens of business cases taking only weeks and months, not years, thanks to SPL.

XML is well-known in the publishing industry and is similar to HTML. It is ideal for text and data, and has proven tools for processing: XPath to navigate and validate (Schematron); XSLT as a sound and fast, data-oriented programming language; HTML as the *lingua franca* for rendering; CSS flexible styling; and HTML5, CANVAS, and SVG, for multimedia, and animation, without limits. XML is preferable to JSON. JSON is generally used for smaller data packages not documents; XML is easier to read for text than JSON and is also not hard to parse in JavaScript. For 30 years, Schadow has seen silver bullet fads come and go; for 15 years, he has worked with XML tools and nothing has surpassed them. His solution can transform to and from any syntax; what matters is the conceptual model.

There are other document standards. XML DOCBOOK is very powerful, but perhaps overkill. DOCX should be supremely powerful, and most people use Word, but it is very hard to work with DOCX. HTML is perhaps too simple and is screen-oriented. T_EX is much more mature and widespread, and produces scientific documents perfectly laid out for printers and publishers. Schadow would favor it but would still transform the data in XML and XSLT.

His next steps are to publish the specification, an implementation guide, and detailed validation procedures. He will create examples and conversion tools; extend the SPL stylesheet for rendering reaction details; and provide an online resource for preview, display, and rendering, validation, and some XForms-based authoring. He will test the system with interested parties, and he welcomes input and participation.

The Unified Data Model (UDM)

Elena Herzog[1], Frederik van den Broek[2], **Gerd Blanke**[3], Jarek Thomczak[4], Markus Fischer[1]; [1]Elsevier Information Systems, Frankfurt, Germany; [2]Elsevier, Amsterdam, The Netherlands; [3]StructurePendium Technologies, Essen Germany; [4]Informatics Unlimited, Cambridge, United Kingdom

Most of the reaction-centric cheminformatics systems (for reaction searching and navigation, reaction similarity and classification, automatic determination of reaction mapping, and mechanism elucidation) were developed before the mid-1990s, but there has been a renewed interest for the last seven years. Much recent research has been carried out on improved reaction prediction by using machine learning; improved reaction mechanism recognition by machine learning; synthetic feasibility; retrosynthesis (design, planning, and optimization); and reaction outcome prediction (prediction of products, yield, specificity, and safety). The ultimate goal is fully automated synthesis machineries.

Researchers want to have quick and easy access to all necessary data. In the ideal world everything would be seamlessly integrated: integration of data would allow search across different domains. In the real world, data generation, exchange, integration, management, and use by the pharmaceutical industry is much less straightforward. In-house, chemists collect data from instruments, and store and use them in ELNs and laboratory information management systems (LIMS), and in drug design and modeling. They want to exchange data with contract research organizations (CROs) and commercial suppliers. The data generation challenge involves dozens of different file formats. The data then have to be integrated in warehouses, lakes, and data marts which are used for data mining, analysis, and reporting, and for preparation of patents, and so on. The problem is that there

is no common language to exchange data among most of these tools. This makes data exchange and integration unnecessarily expensive.

The UDM[60] aims at improving data exchange and integration of chemical reaction data for efficient search, visualization, analysis and predictive modeling. The UDM project is a collective effort of vendors and life science organizations to create an open, extendable, and freely available reference model and data format for the exchange of experimental information about compound synthesis and testing.

The project carried out nine interviews concerning data use cases in the pharmaceutical industry and found that the vast majority of users wanted data integration from various ELNs; data exchange between pharma and CROs; data preparation for AI and machine learning; data exchange between various applications (e.g., LIMS, molecule design tools, and visualization tools); data comparison (e.g., predicted and published synthesis routes, routes predicted from different AI and machine learning models); data preparation for ingestion into in-house repositories (warehouses, lakes, and data marts); and data transfer from and to laboratory devices. An additional opportunity is to make data FAIR.[65]

The data integration task for the UDM project is to collect and integrate existing in-house and publicly available reaction data; to use machine learning, AI, and similarity methods to improve and optimize the prediction of chemical reactions; to integrate synthesis machineries into the workflow, and to become FAIR.

The origin of the UDM was a Roche project (in 2012-2013) to integrate Roche in-house chemistry data into Elsevier's Reaxys[44] database system. It 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.

To unify the different data models into one common system (UDM), several public reaction databases licensed by Roche had to be integrated. Fortunately, most of them largely shared the FIZ Chemie ChemInform data model for reactions. Additionally, the UDM had to integrate the model of the Synopsis/Accelrys Protecting Groups database and the MDL Metabolite database which contributed about one third of all data field definitions. Reactions from the Roche electronic lab journal, with a mostly "flat" approach to store reaction and reaction data, had no significant influence on the UDM development.

Version 1.0 of UDM was developed to define the transfer model for data export into Reaxys. It was written for the MDL RDfile format,[6] with BIOVIA's Pipeline Pilot[100] as the transfer framework. The internal ELN data were exported using NextMove Software's HazELNut[101] into Pipeline Pilot; the other databases were exported from MDL ISIS Direct into Pipeline Pilot.

UDM version 2.0 moved from an RDfile-based format to an XML-based format with significant advantages. RDfile implies 7-bit ASCII content whereas XML uses the 8-bit Unicode Transformation Format (UTF-8). The RDfile format uses naming conventions to represent a hierarchical data model with potentially multiple data hierarchies. In general, the actual format of RDfiles depends on the data model of the database to which it is linked so that RDfiles from different sources may not be compatible. In comparison, UDM 2.0 uses one explicit data model for its XML schema. In an RDfile

there is no type control or validation of values of individual data field values whereas XML has strong typing and controlled vocabularies. RDfiles represent chemical reactions in the RXN format that internally uses the molfile format for each reaction component, whereas in the UDM, multiple representations are allowed (e. g. explicit molfiles for each reaction component, RXN files, Reaction InChIs, etc.). Validation, processing and conversion of RDfiles require dedicated tools or libraries whereas standard XML technologies provide a large part of the data processing operations.

Elsevier took over the rights for the further development of UDM at the end of 2013. It has become the major exchange format for Reaxys data. In 2017, Elsevier donated the UDM to the Pistoia Alliance[102] for further development. The founders of the Pistoia working group were BIOVIA, Elsevier, GSK, Novartis, and Roche. The current UDM team includes AstraZeneca, Bionocvision, BIOVIA, Bristol-Myers Squibb, CAS, ChemAxon, Discovery Information, Elsevier, GSK, IDBS, Ideayabio, InfoChem, Informatics Unlimited, KIT, NextMove Software, PerkinElmer, Roche, and StructurePendium.

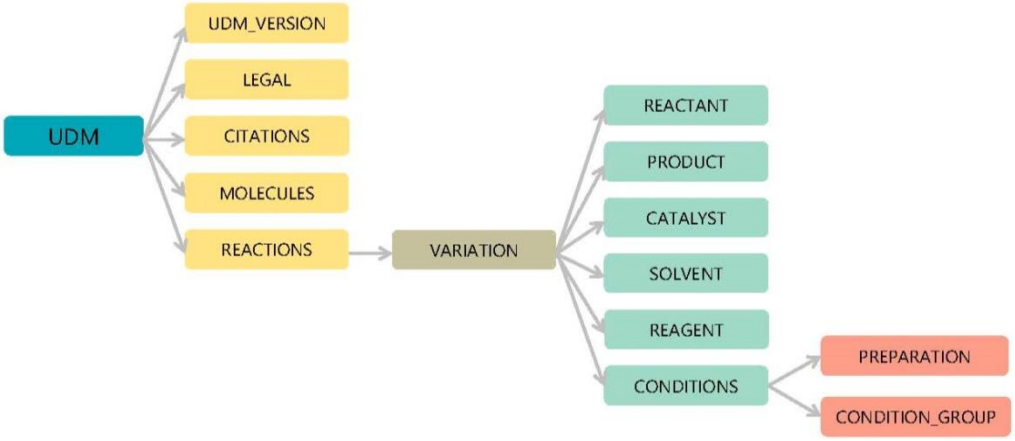


Figure 27. Simplified UDM data model.

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.

A simplified data model is shown in Figure 27. A history of UDM releases is shown in Figure 28. BIOVIA provides a UDM node with the latest (2021) version of Pipeline Pilot.

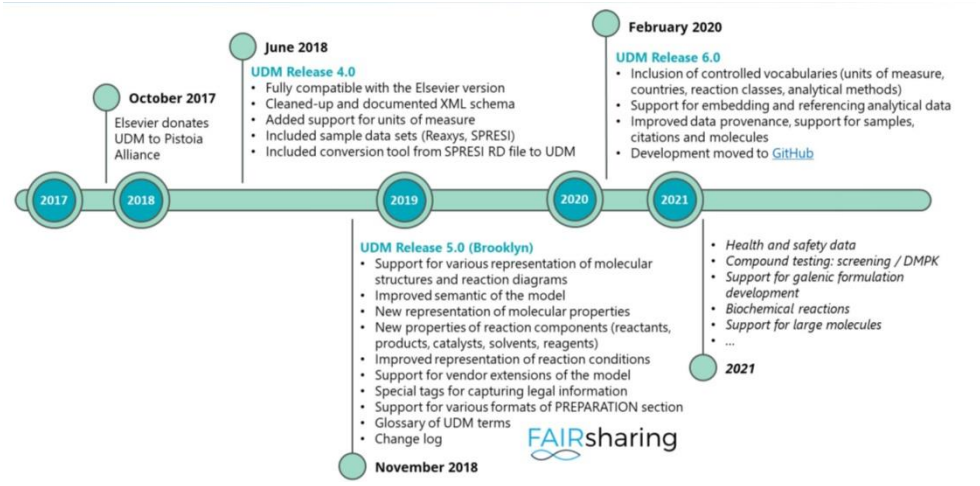


Figure 28. UDM releases.

Under investigation for future enhancement are improved documentation of XML tools to work with UDM and conversion and validation tools; fuller support for multistep reactions; better support for the results of reaction predictions; extensions for biological data, materials and predictive data; and extensions of supported chemical entity types (mixtures and formulations and chemical entities that are described by recipes). In perspective is a move from the Unified Data Model to the Universal Data Model.

The Pistoia Alliance closed the UDM as a Pistoia project with the successful delivery of version 6.0 and the UDM is being transferred into a community-led development. The UDM project can be downloaded from GitHub.[60] The transition conditions for the UDM into an open source project have been negotiated. A new organization to handle the financial issues and to define the governance for the open source development is being sought. Interviews with the stakeholders are underway to investigate the current usage of the UDM and ask for potential financial support. The UDM is now available under an MIT license.

Pistachio

John Mayfield, Ingvar Lagerstedt, and Roger Sayle, NextMove Software, Cambridge, United Kingdom

Pistachio[103] is a document-centric database of 13.3 million reactions automatically extracted from U.S. Patent Office (USPTO), European Patent Office (EPO), and World Intellectual Property Organization (WIPO) patents. JSON and SMILES are provided for bulk analysis and model building. There is a containerized web application for exploring and querying the data. The aim is to extract reactions as described in the original document, whatever faults they may have. This may include mistakes in the original description such as incorrect IUPAC names.

The first patent extraction process[104] from NextMove Software was based on work published in Daniel Lowe's Ph.D. thesis,[105] but used NextMove's LeadMine[106] instead of Open-Source Chemistry Analysis Routines (OSCAR4).[107] LeadMine improves chemical entity and physical quantity recognition, and corrects spelling. The Pistachio data have had a significant impact.[33,108-114] Note that

although there are 13.3 million reactions in Pistachio, many are identical by connection table, or almost identical, as evidenced by the fact that there are only 4.2 million RInChIs (without any role normalization).

The USPTO CC-Zero reaction data up to September 2016 are freely available.[13] Pistachio is not free. It uses improved extraction methods and is updated quarterly with the latest patent data. In addition to the U.S. patents, Pistachio includes reactions extracted from EPO and WIPO text, as well as USPTO sketches. NameRxn[32] classification and atom-to-atom mapping are provided for 71.5% of the reaction citations. Example and step labels, solvent mixtures, solvent associations, document assignees, and targets and diseases are also now captured. Due to the improved extraction methods, and tweaks to filtering, Pistachio is a not quite a "superset" of USPTO but the majority of the reactions in USPTO CC-Zero are included.

Mayfield outlined the sectioning, tagging and tokenization, parsing, action phrases, and reaction assembly stages of the text mining operation.[115] Since 2001, the USPTO has redrawn chemical sketches with ChemDraw for all patent applications and grants. This dataset provides a large collection in which to identify conventions and rectify systematic problems. By identifying and resolving these problems, NextMove Software has extracted a high quality collection of fragments, molecules, reactions, and schemes using their Praline tool.[116] Pistachio includes the reaction SMILES extracted using this tool.

Reactions are filtered and roles are re-assigned before atom-to-atom mapping; if a reagent contributes any atom to the product it is considered a reactant. NameRxn was originally written as a classification tool; mapping is a by-product. It has 1543 rule-based classes, so it is easy to update a mapping disagreement. It has higher precision but lower recall than some other mapping algorithms as it can only those reactions it can recognize. Typically, on an in-house ELN the coverage is 70-80% but it is easy to construct a benchmark on rarely used reactions that NameRxn will not recognize. Mapping is fast and takes only a few hours to remap the entire database; remapping is carried out before every release. In Lin *et al.*'s benchmarking study,[28] NextMove software demonstrated the lowest number of correct and incorrect atom-to-atom mapping results compared with the other tools because it adopts the principle that no answer is better than a wrong answer. There are also cases where a product atom is unmapped because the software did not know where a group came from, a group or catalyst was missing, or there was a stoichiometry problem (multiple groups from one reactant).

Indigo[30] and RXNMapper[33] fail in a reaction example containing the faulty name 8-(3,5-Bis-trifluoromethyl-benzoyl)-3-furan-2-yl-methyl-1-o-tolyl-1,3,8-triaza-spiro[4.5]decane-2,4-dione. NextMove software changes the name to 8-(3,5-Bis-trifluoromethyl-benzoyl)-3-furan-2-ylmethyl-1-o-tolyl-1,3,8-triaza-spiro[4.5]decane-2,4-dione and maps the reaction correctly. In a case study of 23 reactions from the patent US 2020/0087299 A1, NameRxn succeeded in over 80% of mappings and Indigo in under 10%.

NextMove Software appreciates feedback and will work on correcting any errors reported. Plans for the future include identifying compound numbers that appear only in reaction schemes, better indication of quality (integration of RXNMapper, mapping bench indicators, and boot-strapping reaction sequences to resolve ambiguous chemical names), handling reactions from patents not

written in English, and addressing the fact that general procedures and example references currently resolve only compounds.

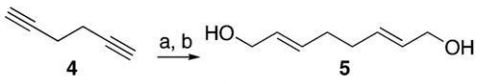
Machine learning with CAS reaction data

David Dastrup, CAS, Columbus, OH, USA

CAS has a comprehensive collection of connected science: over 250 million substances and more than 130 million single and multistep reactions have been gathered from over 50,000 scientific journals and documents. Fifty languages are translated, and patents from over 64 patent offices worldwide are indexed. CAS scientists curate, connect, and analyze scientific knowledge. Curation gives meaning to data, information across disciplines is connected, and analysis reveals insights.

Substances in the CAS Reactions collection are linked to CAS REGISTRY[117] which contains 182 million organic and inorganic substances and about 70 million protein and nucleic acid sequences published since the early 1800s. Reactions published since 1840 are indexed with yield data, detailed reaction conditions (time, temperature, pressure, and pH), and reaction descriptions such as "stereoselective".

Points for consideration when curating data for machine learning are reaction representation, data capture, reaction selection based on project, and data exchange. Published reactions are not applicable for all objectives. What is disclosed and what can be directly implied from the literature has to be represented. An example of CAS reaction capture is shown in Figure 29: the literature item is shown at the top, and at the bottom is part of a CAS SciFinder^n record for the first reaction captured by CAS.



(a) *n*-BuLi, $(HCHO)_n$, 71%. (b) Na, NH_3, reflux, 76%. The *trans,trans*-diene-diol **5** was prepared using Rosenblum's procedure in 51% yield.[11]

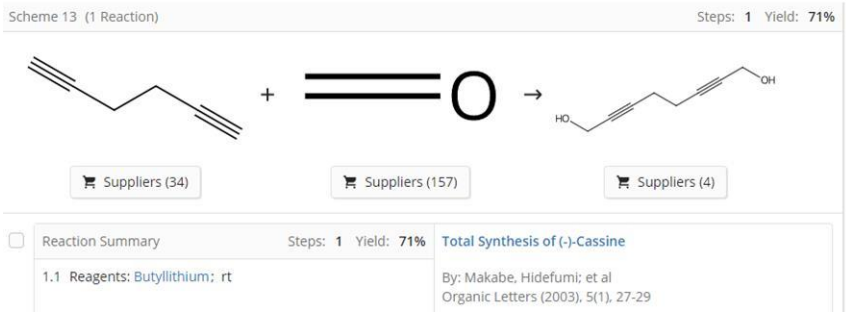


Figure 29. CAS data capture (bottom) for a literature reaction (top).

In collaboration with a customer, CAS selected a subset of reactions for a particular transformation with yields evenly distributed across selected content, ensuring a variety of reactant and catalyst structural features. Failed reactions are not often reported but low-yield reactions are included in

this subset of 10,000 reactions which was supplied to a CAS customer for machine learning purposes. Reaction data are exchanged as RDfiles (with atom-mapped reactants and products);[6] as XML files with reactants, products, reagents, solvents, catalysts, conditions and notes; and as SDfiles[6] for substance representation.

Dastrup presented a case study in which Bayer sought to predict the viability of reactions for synthetic planning, with an emphasis on novel areas of science.[118] Lack of training data diversity had impacted the performance of a synthesis planning application: testing against selected reaction classes initially showed only 16% predictive power. The diversity of the predictions is correlated with the breadth of the data source: how many reaction types are represented and how diverse the reactants and product are in each reaction. The accuracy of the predictions depends on the quality and consistency of the data and their representation: the number of examples available for each reaction type and the spectrum of reactants, products, and reaction conditions available. Bayer's data diversity was strengthened with custom-curated CAS reactions to provide additional examples to underrepresented templates.

The training set initially used was commercially available positive data (8 million reactions) and implied synthetic negative data (24 million). Then 14,500 curated CAS reactions for specific templates were added to the training set and accuracy for the test set of selected reactions increased to 48%. This enhanced predictive power within "rare" reaction categories now contributes new, useful results and opens up previously difficult areas of science. CAS Reactions are available through the CAS SciFinder Discovery Platform[119] and STN IP Protection Suite.[120] Customers can tailor their own data and delivery format through CAS Custom Services.[121]

The Open Reaction Database

Connor Coley, Massachusetts Institute of Technology (MIT), Cambridge, MA, USA and **Steven Kearnes**, Relay Therapeutics, Cambridge, MA, USA

The Open Reaction Database (ORD)[122] is designed to support machine learning and related efforts in reaction prediction, chemical synthesis planning, and experiment design. Its goals are to provide a structured data format for chemical reaction data; to provide an interface for easy browsing and downloading of data; to make reaction data freely and publicly available for anyone to use; and to encourage sharing of precompetitive proprietary data, especially high throughput experimentation (HTE) data. ORD's initial goals are not to capture reaction processes as action sequences for robotic execution, or to store processed, structured, analytical data or other inputs that are not directly related to the machine learning efforts, or to integrate model building or other use of the data as part of the database.

Members of the governing committee are: Connor Coley (MIT), Abby Doyle (Princeton), Spencer Dreher (Merck), Joel Hawkins (Pfizer), Klavs Jensen (MIT), and Steven Kearnes (Relay Therapeutics). Advisory Board members are: Juan Alvarez (Merck), Alán Aspuru-Guzik (Toronto, and MADNESS), Tim Cernak (Michigan and Entos), Lucy Colwell (Cambridge, SynTech, and Google), Werngard Czechtizky (AstraZeneca), Matthew Gaunt (Cambridge and SynTech), Mimi Hii (Imperial, and ROAR), Greg Landrum (T5 Informatics), Fabio Lima (Novartis), Christos A. Nicolaou (Lilly), Sarah Reisman (Caltech), Matthew Sigman (Utah), Jay Stevens (BMS), Sarah Trice (Entos), and Matt Tudge (GSK).

The ORD schema captures the most important aspects of reactions in a structured format since structured data enable downstream machine learning applications. Guided by a recent survey, the focus is on single-step batch reactions. Additional details are in a flexible, unstructured format. Chemists' expectations around structure and nomenclature are matched. ORD records what physically occurred in a chemical reaction and de-emphasizes recording of a chemist's intent. For example, ORD records the actual masses and volumes that were used to create a stock solution, not the target concentration.

Primary uses of the ORD are in synthetic organic chemistry. For high-throughput experimentation, data are recorded in spreadsheet formats including only varied parameters; one template reaction is defined to specify all aspects held constant; and the dataset is defined by iterating over the spreadsheet and creating one reaction entry per experimental condition. For traditional bench chemistry, a chemist uses a graphical web form to define the settings and outcomes of all reactions used within a paper or project; the structured dataset is saved, and uploaded to the ORD and used as part of the supporting information; and a list of reactions is exported from the dataset in a text format like that of supporting information in a journal article.

The schema is shown in Figure 30. The protocol[123] is not unlike XML but is smaller, faster, and simpler. In the readable definitions, each field gets a name and a tag number. Storage formats are text and binary. The code compiles to language-specific classes.

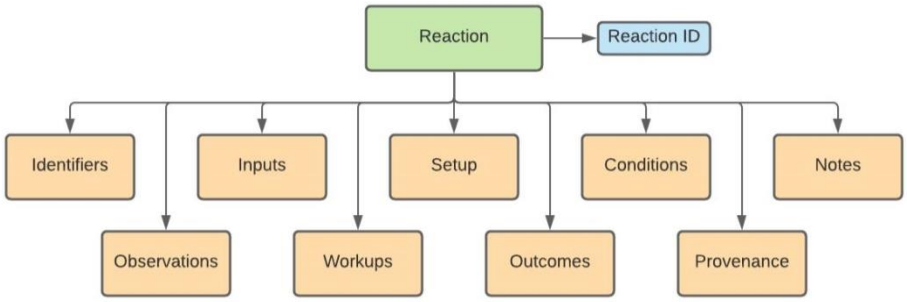


Figure 30. ORD schema.

Kearnes showed a reaction example and its inputs (Figure 31), its setup and conditions (Figure 32), and its outcomes (Figure 33). There is an interactive editor.[124]

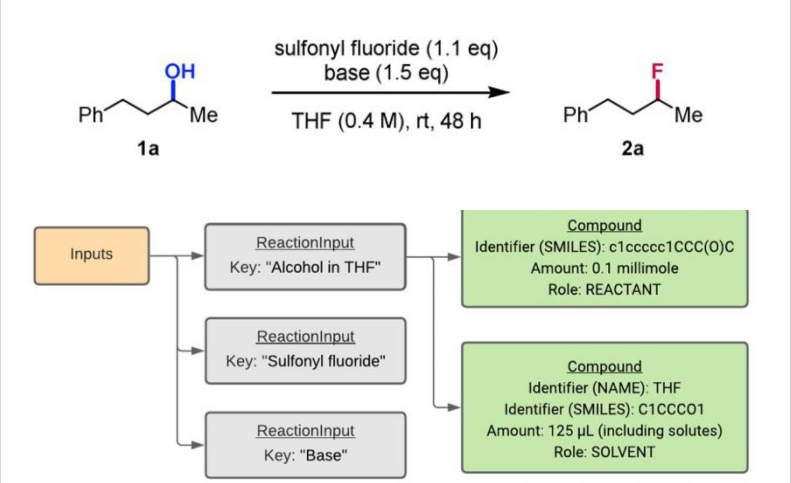


Figure 31. Example of ORD reaction input.

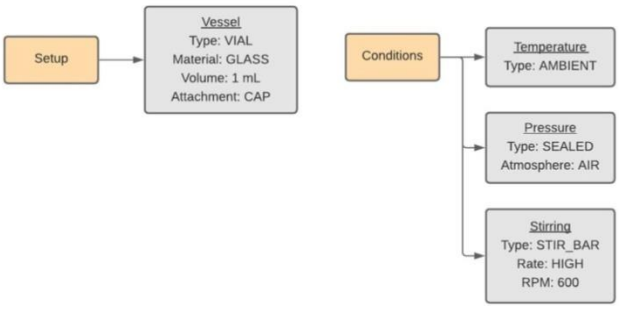


Figure 32. Example of ORD reaction setup and conditions.

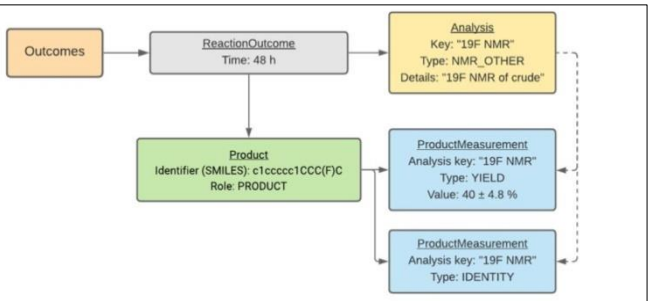


Figure 33. Example of ORD reaction outcomes.

Users can search the database by exact match or substructure for multiple inputs or outputs, and look up reactions by ID, DOI, or reaction SMARTS. A Python client with equivalent functionality is also provided.

The data are available on GitHub[125] under a CC-BY-SA 4.0 license. Daniel Lowe's USPTO grants dataset[13] is also available, converted from CML.[99] Schema, code, and web interfaces are available on GitHub under an Apache 2.0 license.[125] Examples and tutorials are available on GitHub and YouTube. The terms of use have been drafted in kind by Google lawyers. Alpha testing began in September 2020, beta testing and prelaunch expansion in November 2020. Public launch is planned for mid-2021. ORD will be open to all contributors for submissions; specific contributions from industry and

academia will be invited; and downstream use in machine learning and other applications will be solicited.

Using reaction data for generative molecular design

Valerie Gillet, University of Sheffield, Sheffield, United Kingdom

The many different approaches to *de novo* drug design[126] first used in the 1990s had restricted sampling of chemical space. Scoring tended to be 3D-based (e.g., it predicted binding) and the methods were agnostic of chemical synthesis. Reaction-based design of new molecules uses transformation rules extracted from collections of known reactions and the aim is to restrict the enumerated chemical space to a manageable number of synthetically accessible structures. Generative model approaches typically use joint distributions to generate new molecules (usually as SMILES strings) with characteristics similar to the training data. They take no direct account of synthetic accessibility.

In reaction-based approaches, a typical approach to account for synthetic accessibility has been to encode a fixed set of known transformations which can then be applied to starting molecules. For example, the Flux program[127] is a fragment-based *de novo* design tool in which a simple retrosynthesis method is used to fragment molecules into building blocks. The building blocks can then be used to construct new molecules.

The reaction-based *de novo* design research of Gillet's team has focused on using reaction vectors (RVs)[128] derived from large collections of known reactions. RVs represent the structural changes that take place at the reaction center along with the environment in which the reaction occurs. They are counts of atom pair descriptors that change during the reaction: negative ones for those removed from reactant(s) and positive ones for those gained in the product(s). A database of reactions is converted into a database of RVs with literature references. A structure generation algorithm applies RVs to previously unseen starting materials in order to suggest novel syntheses.[128] The reaction center is described by AP2 atom pairs and the center is extended through the use of AP3s (Figure 34). The fragmentation and reconstruction algorithm applied during RV database generation enables fast structure generation: hundreds of structure generation operations per second.

AP2: X1(*n*,*p*,*r*)-2(*BO*)-X2(*n*,*p*,*r*)
AP3: X1(*n*,*p*,*r*)-3-X2(*n*,*p*,*r*)

X = element type
n = number of bonds to heavy atoms
p = number of π bonds
r = number of ring memberships
BO = bond order

Figure 34. Reaction vector atom pairs.

In the first part of the procedure, reactions have to be cleaned. All species not involved in reaction mapping (e.g., catalysts and reagents) are removed. Reaction balancing is then carried out, adding in species not included in the reaction and separation of reactions. AP2s and AP3s are then calculated and the RV is validated for structure generation. Duplicates are finally removed. Gillet's team processed 1.8 million reactions from the USPTO database.[13] Sixty seven percent of the reactions in

the database had RVs with redundant AP2s and AP3s; 18% failed the cleaning stage; and 9% gave unique unclassified RVs. Following cleaning, calculation of RVs, and removal of duplicates, the RVs corresponding to a reaction which had been annotated by class were validated. Of the 115,602 unique classified RVs, 92,530 were validated.

To test whether the method can reproduce known drugs, the team used an approach similar to the pseudoretrosynthetic design protocol of Flux [127] (Figure 35) but with fragment combinations driven by RVs. The ligands fragmented were 73 of the top 200 prescribed drugs in the United States in 2017. Similar fragments were retrieved from a database of 750,000 reagents in the Enamine database.[129] The 92,530 validated RVs were used in *de novo* molecule generation. Similarity calculations for the Enamine searches and for similarity of the output molecules to a known drug were based on RDKit Morgan fingerprints.[130] The top scoring compound generated for 70% of the drugs had more than 0.5 similarity to the parent drug.

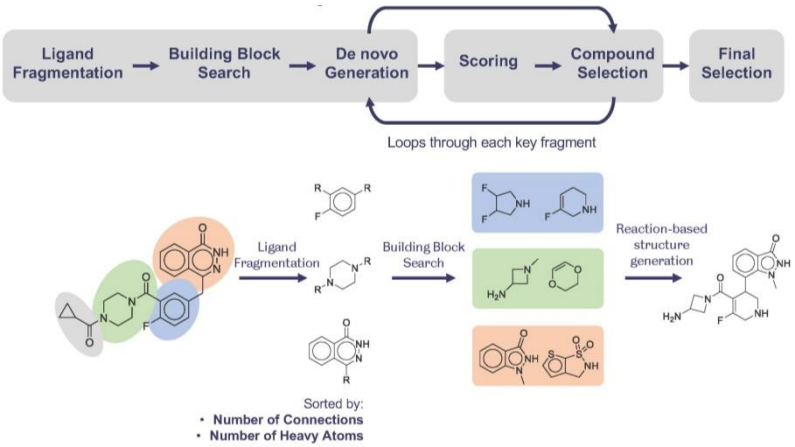


Figure 35. Flux flowchart.

Another validation was also devised to test whether products from the RV method are synthetically accessible. The team used 15,000 starting materials (fragments, leads, and drugs) selected from ZINC[131] and 10,000 reagents also selected from ZINC. A set of 26,757 single-step reactions was extracted from *Journal of Medicinal Chemistry* publications covering the period January–September 2018, using Reaxys,[44] and was cleaned. Using 4500 RVs from this dataset ("JMC"), 8.91 million products were generated, 8.2 million of which were unique, leading to 8.24 million unique readable products after invalid SMILES had been removed. Of these 8.24 million products, 88.98% were synthetically accessible according to RAscore.[132]

In a prospective design study, the aim was to generate compounds with poly [ADP-ribose] polymerase 1 (PARP1) activity, maintaining PARP1 affinity, reducing affinity for antitargets (P-glycoprotein (Pgp) and breast cancer resistance protein (BCRP)), and improving blood brain barrier penetration. Fragments from PARP1 ligands were used to search 750,000 reagents in the Enamine database, and the 93,000 RVs from the USPTO database were used in *de novo* generation (see the flowchart at the top of Figure 35). For scoring, a QSAR regression model was built from 1800 compounds in ChEMBL.[47] Classification models were built from 500 Pgp and 250 BCRP (substrates and nonsubstrates, respectively) and from 2000 positive and negative blood brain penetrators, extracted from the literature. In the final selection, functional group transformations were applied to

compounds identified as reactive using Hann definitions.[133] Unsuitable compounds located by searching for various SMARTS patterns were filtered out and the remaining compounds were docked to the Protein Data Bank (PDB)[134] structure 4RGE using GOLD.[135]

Following manual inspection of the docking poses, 20 compounds were selected for synthesis. All were predicted as PARP1 active, and nonsubstrates of Pgp and BCRP, and able to penetrate the blood brain barrier. None was available for purchase or was in ChEMBL[47] or PubChem.[62] Eight of the 20 compounds were submitted for synthesis by medicinal chemists at Evotec. The proposed synthetic routes according to RVs were inspected and adjusted according to reagent availability (cost, delivery times, etc.), additional steps required (e.g., protection chemistry) and successful conditions. The compounds are currently being tested for PARP1 activity. Three are "scaffold hops" based on known scaffolds, the others are novel.

The reaction databases used in the work above were relatively small. Commercial reaction databases can contain millions of reactions and a mechanism to control the combinatorial explosion achievable with *de novo* design is required, especially when multistep reactions are considered. As modern drug discovery is a multiobjective optimization problem, further work has focused on embedding the structure generation algorithm within an iterative loop to provide populations of molecules that satisfy multiple objectives while also having a high degree of confidence that they are synthetically accessible.

One approach is to use a recommender system[136] to filter the RVs by recommended reaction classes. Gillet's team has used two related machine learning models: a reaction classification model[137] and a reaction class recommender.[138] The former is a multiclass model which takes in a set of reaction vectors labeled by reaction class and predicts the reaction class of a previously unseen reaction. The recommender is a multilabel model which takes in a set of starting materials represented as molecular descriptors and labeled by reaction classes, and predicts a list of recommended reaction classes for a previously unseen starting material.

The training set for the reaction classification model was 111,000 unique RVs, represented as AP2s and AP3s, from granted patents in the USPTO database. Random forest (RF) was selected as the best machine learning method. For internal validation, 336 classes with >30 RVs were selected from the whole dataset and divided 40:60 into a training set and a test set. For external validation, 25,000 unseen RVs were extracted from USPTO patent applications (i.e., RVs from USPTO grants were not included). The model performance was similar to that of the model of Schneider *et al.*[139] but the dataset was extended from 50 to 336 classes. The model was combined with a conformal prediction approach to return a confidence in the prediction.

A limitation of RVs is that they account for structural changes that occur at the core of a reaction only, and they do not consider the presence of competing functionalities that can compromise the reaction outcome. The reaction class recommender enhances the reaction vector framework to address this issue. A training set of starting materials was extracted from 1.1 million classified reactions. Starting materials with identical descriptors were merged and the reaction class labels were appended. The data were split into 80% used for training and 20% for testing. The external set used in validation was JMC. Extensive experiments were carried out for optimal training of the recommender, combining two sets of classification labels with nine molecular descriptors, and four multilabel approaches. Multilabel classification problems are generally addressed using two

alternative methods: problem transformation (PT), where the multilabel problem is transformed to be compatible with traditional classifiers such as random forests (RF) or SVM; and algorithm adaptation (AA) methods (e.g. multilabel *k*-nearest neighbors). In preliminary work, the PT method gave better performance than AA so Gillet's team focused on PT methods which can be divided into binary relevance, classifier chain (CC), and label powerset. Avalon and FeatMorgan fingerprints gave the best performance followed by MACCS fingerprints.[138] Some results are shown in Figure 36.

Test set

Fingerprint	Setup	"Micro" recall	"Micro"precision	"Micro" F1-score
Avalon 1024-bit	CC-RF	0.29	0.76	0.42
MACCS	CC-RF	0.25	0.69	0.37

External set

Fingerprint	Correct	Wrong	No recommendation
Avalon 1024-bit	33.5	21.9	44.6
MACCS	37.1	23.1	39.8

Figure 36. Reaction class recommender: validation (1).

`

The recommender was further validated in a procedure similar to the one described earlier for the retrospective design of known drugs (73 of the top 200 prescribed drugs using 750,000 Enamine reagents and 93,000 RVs). Results (Figure 37) suggest that the use of the recommender drastically reduces the number of solutions explored by the algorithm while preserving the chance of finding relevant solutions and increasing the global synthetic accessibility of the designed molecules.

		Number of products generated		Enumeration times (hours)	
Drug	Steps	Without Recommender	With Recommender	Without Recommender	With Recommender
Brimonidine	1	333,361	97,842 (-71%)	3.0	1.2 (-60%)
Glipizide	2	732,705	251,821 (-66%)	5.2	2.5 (-52%)
Glyburide	2	1,317,776	1,016,319 (-23%)	7.5	6.3 (-16%)
Levofloxacin	1	732,285	135,084 (-82%)	4.1	1.5 (-63%)
Naproxen	1	425,693	113,726 (-73%)	3.1	1.3 (-58%)
Rivaroxaban	3	1,282,308	536,212 (-58%)	7.7	3.9 (-49%)

Figure 37. Reaction class recommender: validation (2).

Gillet's team is currently developing a reaction-based *de novo* design algorithm based on Monte Carlo tree search (Figure 38). A simulation step is used in this approach whereby an intermediate molecule is scored on the basis of the structures that can be generated from it.

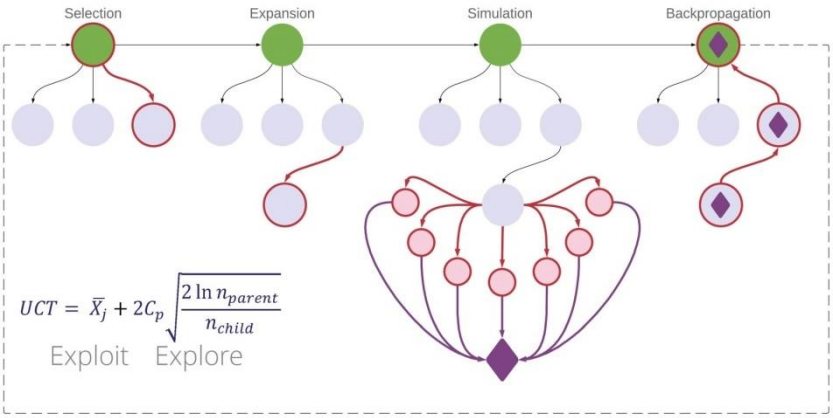


Figure 38. Monte Carlo tree search.

SynSpace: multistep forward-synthesis for scaffold hopping and generative design in synthetically feasible chemical space

Greg Makara, ChemPass, Budapest, Hungary

Historically, the lead discovery process at a big pharmaceutical company takes 2.5-4.0 years, consuming the resource of 250 full-time equivalent staff. The design-make-test-analyze (DMTA) cycle time and the number of cycles are critical factors in getting a drug to market earlier. Chemical synthesis is by far the longest and most expensive stage, taking 0.5 –4 months per cycle. The data of "stragglers" miss out on several cycles; compelling design ideas are hard to dismiss even where there are perceived synthetic challenges; and it is hard to drop a compound once synthetic effort is underway as "there is always another route". Off-the-shelf and virtual commercial catalogs will have little impact after lead finding.

Consider a lead optimization workflow in which 300,000 building blocks are reduced to 110,000 dockable compounds by various filters. About 30,000 of the docking hits are submitted to machine learning and free energy perturbation (FEP) procedures, 5000 are selected, and after further refinement, about 35 are selected. Retrosynthesis can be carried out at the end, after all the expensive virtual screening and filtering processes, or a lot of money could be spent on running retrosynthesis on 100,000 compounds, or millions of compounds, most of which get thrown out in the docking and FEP stages (though perhaps RAscore[132] can help).

It would be much better if the entire idea space were synthetically feasible, so that costly processes could be carried out on relevant, synthetically feasible chemical space. All designed structures would be worthy of evaluation and there would be no wasted resources and cost, assuming that the designed space were as good as that offered by deep generative design or simple enumeration.

Therefore ChemPass design technology looks at what can be synthesized from starting materials, intermediates, or lead structures. Technology development was required to solve rule-based AI for forward *in silico* synthesis, molecule design based on multistep *in silico* synthesis, and control of the combinatorial explosion. SynSpace[140] offers customized ideation in synthetically feasible space, with about 300 transformations in the current version. It is a user-friendly computational tool that can bridge the gap between medicinal and computational chemists, making preclinical research more

efficient: computational chemists do not need synthesis knowledge and medicinal chemists do not need cheminformatics knowledge.

There are SynSpace modules for multiple design tasks:

- reaction-based design for library design (including DNA-encoded libraries)
- starting material base design for side-chain analogue design
- scaffold hopping and scaffold analogue design
    - 1-step process, 1-Click design
    - 2-step general scaffold design
- multistep and multisite library enumeration
- retrosynthesis
- derivatization design (DD) for automated generation
- automated three-step generative design (reactant-based design).

SynSpace 1-Click scaffold design (for forward synthesis and 3D overlap) is a simple tool that requires no cheminformatics skill set. The design outcome is influenced by simple user settings for H-bonding features, aromaticity, and ring size. An intelligent ring closing method is included so that bicyclic derivatives of monocyclic leads can be easily explored. New scaffolds with properties, synthetic information, and novelty assessment can be displayed in a spreadsheet format.

There are two proprietary generative design tools: reagent-based generative design and derivatization design.[141] The latter offers simple control on the number and type of variations, the depth of modification at each site (exploration or exploitation of scaffold hopping and variational analogues), similarity, and desired set size. Simple user inputs drive the fully automated process. All the designed molecules possess vital synthesis, reagent, and vendor data.

Makara presented a DD example (Figure 39). There were 5214 products when the similarity thresholds were set as in Figure 39A. Positional (or reaction step) contribution to the total result set depends on the reagent (positional) similarity range set by user (the primary driver); commercial (or custom uploaded) reagent diversity; and reaction type (bimolecular better than monomolecular.) Figure 39B shows the contributions to diversity for Figure 39A. Figure 39C shows the contributions to diversity for the case when all similarities are set to 0.7 and there are 9264 products. The positional (or reaction step) contribution to the total result set has now become well distributed.

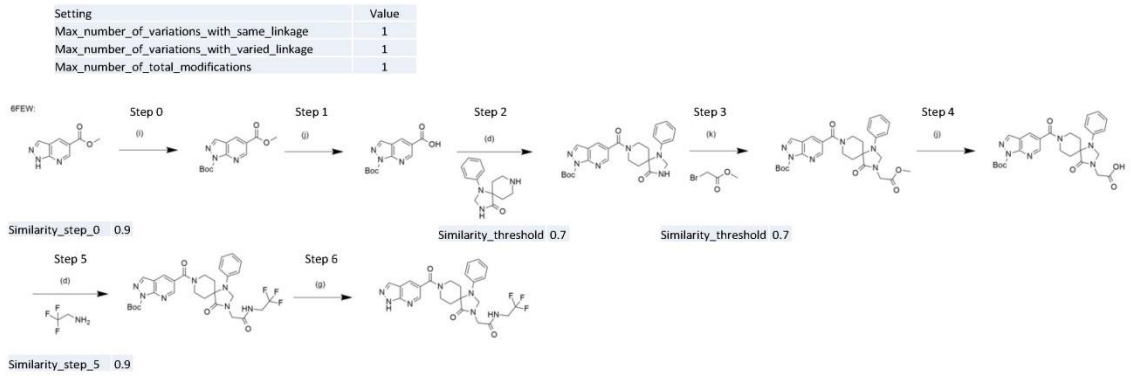


Figure 39A

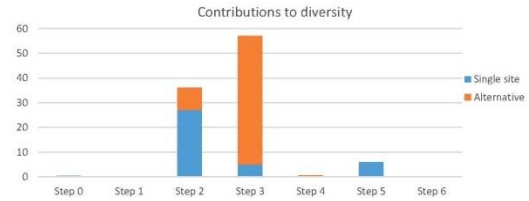


Figure 39B

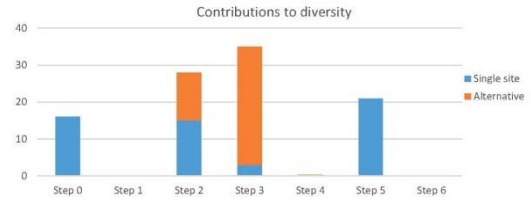


Figure 39C

Figure 39. Derivatization design example.

The ChemPass team have reported a case study with the discoidin domain receptor 1 (DDR1), comparing derivative design (DD) to generative design by generative adversarial network (GAN).[142] SynSpace DD is superior to the industry standard deep design[142] in this study: it has a much higher hit rate, better docking scores, and more new motifs. It can be used in early lead optimization: it needs no large training set and a user can start from one hit or lead. It can effectively optimize a lead, or create new lead classes, or do both at the same time. It produces synthetically feasible compounds: users have had an 85% wet lab success rate for schemes with 3-6 steps.

Makara summarized lead generation *via* scaffold hopping, derivatization design, and docking pose analyses in the DDR1 use case. Scaffold 1 with a new DDR1 motif from the derivatization design case study had a GlideScore of -13.1 and docking ligand efficiency (DLE) 0.62. (Ponatinib has a score of -15.4, DLE 0.54, and IC_{50} 9 nM.) ChemPass has a tool for automated analysis of docking results (see below) which flags structural irregularities, leading to false positives (inactive structures). Of course, experimental spot checking should be done to confirm this. Scaffold 1 had flagged poses.

When docking was carried out in 20,000 analogue space, all the docked molecules were flagged. The best GlideScore was -16.7 with DLE 0.65. Exploration with 1-click scaffold design led to 39 relevant scaffold analogues with desired features. Scaffold 2 had Glide SP score -12.5 with DLE 0.55, and no flags; Scaffold 1 control had Glide SP score -14.0 with DLE 0.60, flagged. Scaffold 2 was exploited by docking in 30,000 analogue space. Many docked analogues had no flags. The best GlideScore was -16.9 with DLE 0.68, and no flags. Two compounds were selected for synthesis and testing. One related to Scaffold 1 had GlideScore 15.5, flagged. IC_{50} was >10uM. One compound related to Scaffold 2 had GlideScore -14.9 and no flags. It had an IC_{50} around 100nM offering an attractive entry point for further optimization. Exploration and exploitation in synthetically feasible lead analogue

space have become simple, rapid, and cost-effective processes, running virtual cycles, as seen in the example, or traditional DMTA cycles.

Techniques to lower the false positive rate in docking have been reported.[143,144] One technique[144] is somewhat similar to one of the decoy finding tools ChemPass have developed and used in the example above for flagging docked structures. ChemPass's experimental data for the flagged structure with IC_{50} >10uM confirm the tool's red flagging the compound, while the second, with IC_{50} 100nM and no flag, is experimentally confirmed to be active.

Additional in-house tools in AI-assisted discovery include active learning, desirability scoring, machine learning and deep learning models, and automated analysis of docking results. Active learning enhances throughput and speed. Active learning selection and desirability scoring outperform GANs and medicinal chemists.

Deep learning models to predict reaction conditions

Matt Clark, Elsevier, Philadelphia, PA, USA

There are many reasons for wanting to find the best conditions for a given reaction. Augmenting computer retrosynthesis to help evaluate paths based on conditions can help to identify reactions with conditions suitable for automated systems. It can also provide options for greener conditions by predicting when greener solvents and reagents will work; reduce costs by selecting less expensive reagents and solvents; and circumvent patented processes with alternative conditions.

Chemical reactions can run under many conditions. Process chemists like reactions such as Suzuki coupling which give a high yield under many conditions. Clark displayed a reaction which runs to 100% yield with many conditions. This reaction has high "condition diversity": a measure of the diversity of conditions possible to carry out a reaction. It has many solvents, reagents, and catalysts nearly all of which work to provide high yield. The way that a deep learning model is trained can be customized to emphasize which factor makes one set of conditions more desirable. Many reactions have only one set of conditions but it could be that other, untried conditions would work even better. The Reaxys[44] reaction with the most sets of conditions has 2,839 unique sets of conditions and the reagents are diverse.

Clark's aim, given a reaction, expressed in reaction SMILES, was to suggest the best reagents, catalysts, time, temperature, and associated yield using an AI model. He created a neural network model based on vectorized reactions and conditions. He trained it on Reaxys data and all reagents and conditions reported for each reaction. Organic reactions alone were selected. The reagents, solvents, and catalysts were modeled as a "one-hot" bitmap, one for each reagent (Figure 40). Since the same reaction can be carried out with different reagent sets, each row is really a "variation". The network looks at all the variations of all reactions in the training. The model can then suggest the prioritized best reagent set for a test reaction.

Reaxys Reaction ID	triethylamine	potassium carbonate	hydrogenchloride	hydrogen	acetic acid	pyridine	water	N-ethyl-N,N-diisopropylamine	sodium hydroxide	trifluoroacetic acid	dmap	sulfuric acid	caesium carbonate	triphenylphosphine	copper(I) iodide	sodium carbonate	palladium diacetate	sodium hydroxide	sodium hydride	sodium tetrahydroborate	potassium hydroxide	toluene-4-sulfonic acid	tetrakis(triphenylphosphine) palladium(0)	oxygen	sodium hydrogencarbonate	...
32604																			1							
29198	1																									
482988		1							1																	

Figure 40. One-hot encoding.

The reaction descriptors are Morgan 1024 bit fingerprints: reactant(s) fingerprint + product(s) fingerprint + Xor(reactants, products), where Xor encodes changes that took place between reactants and products. The descriptor thus has 1024*3 bits.

Conditions are grouped to predict a “class” of conditions more reliably. Ranges are selected by evaluating the histogram of observed values in Reaxys. The prediction therefore gives the ranges of time categorized as 1, 6, 12, and 24 hours; temperature categorized as –78, 0, 20, 50, and 100 degrees, and yield categorized as 25, 50, 75, and 100%. The network (a Keras Tensorflow system) is trained to predict a reagent, solvent, and a time, temperature, and yield set based on reaction fingerprints. It is a “classification” of which reagents and conditions are appropriate. Of all the networks tested, the one in Figure 41 worked best. The model was trained on 90% of 8,137,207 organic reactions. The internal test set was the 10% that constituted the most recent reactions.

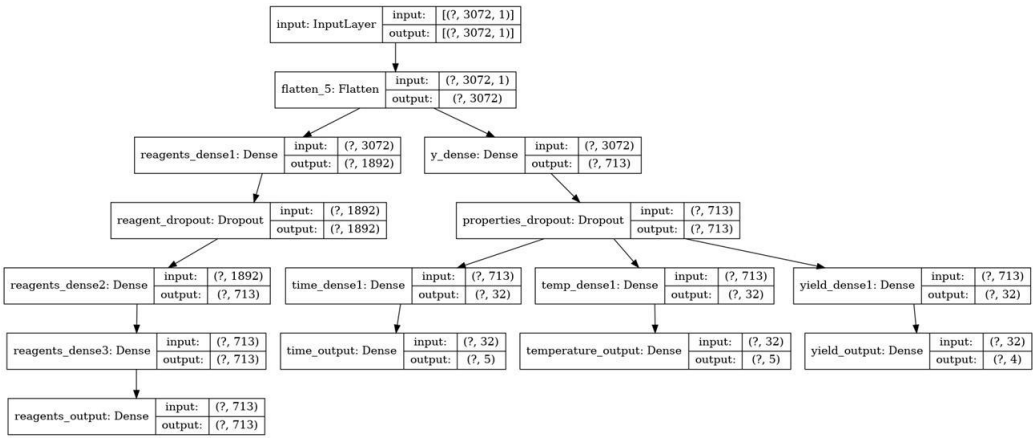


Figure 41. Best neural network.

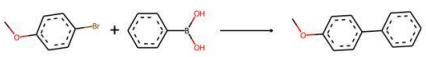
For the Reaxys reaction with the most reported conditions (2,839 unique sets of conditions), the predicted conditions are among the many 100% yield conditions (Figure 42).

reagents	oxygen (0.91) dihydrogen peroxide (0.43) *tert*-butylhydroperoxide (0.20) 2,2,6,6-Tetramethyl-1-piperidinyloxy free radical (0.13) potassium bromide (0.09)
solvents	water (1.00) acetonitrile (0.56) dichloromethane (0.32) toluene (0.13) sulfuric acid (0.12)
temperature	100C (0.33) 20C (0.33) 50C (0.18)
time	1h (0.28) 6h (0.18) 24h (0.18)
yield	100% (0.39) 25% (0.32) 75% (0.21)
sensitivity	8.803

Figure 42. Example of prediction results.

The predictions for the most recorded Suzuki reaction (Figure 43) are good even though the model was trained on all sorts of different reactions.

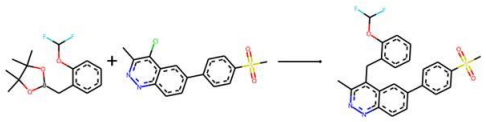
reagents	potassium carbonate (0.85) sodium carbonate (0.23) potassium phosphate (0.14) palladium diacetate (0.11) tetrakis(triphenylphosphine) palladium(0) (0.09) tetrabutylammomium bromide (0.08)
solvents	water (1.00) ethanol (0.44) N,N-dimethyl-formamide (0.18) 1,4-dioxane (0.12)
temperature	100C (0.39) 50C (0.19) 20C (0.19)
time	1h (0.39) 0h (0.19) 6h (0.14)
yield	100% (0.74) 25% (0.15) 75% (0.11)
sensitivity	4.432



Reported Reagents	
Name	Count
potassium carbonate	859
sodium carbonate	144
potassium phosphate	135
cesium carbonate	126
palladium diacetate	115
tetrabutylammonium bromide	85
potassium hydroxide	62
palladium dichloride	47
sodium hydroxide	31
potassium *tert*-butylate	23

Figure 43. Suzuki reaction.

Clark also tried an example from a recent publication on reaction prediction.[145] His method predicted the reagents, solvent, and temperature correctly; the time and yield predictions were not so good (Figure 44).

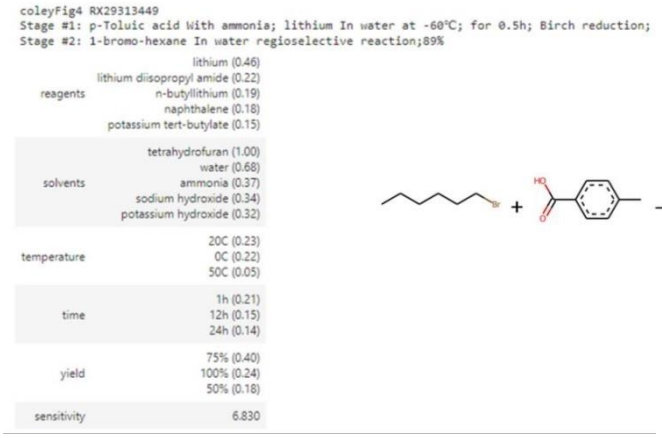


Reisman A RX40306344
With dichloro(1,1'-bis(diphenylphosphanyl)ferrocene)palladium(II)*CH2Cl2; potassium carbonate In 1,4-dioxane; water at 100°C; for 27h; Inert atmosphere; Sealed tube; 15% yield

reagents	(1,1'-bis(diphenylphosphino)ferrocene)palladium(II) dichloride (0.47) potassium phosphate (0.45) dichloro(1,1'-bis(diphenylphosphanyl)ferrocene)palladium(II) *CH2Cl2 (0.43) potassium acetate (0.31) cesium carbonate (0.23)
solvents	1,4-dioxane (1.00) water (0.96) potassium carbonate (0.38) N,N-dimethyl-formamide (0.11) 1,2-dimethoxyethane (0.10)
temperature	100C (0.48) 20C (0.17) 50C (0.10)
time	1h (0.20) 12h (0.19) 24h (0.13)
yield	25% (0.43) 50% (0.27) 75% (0.22)
sensitivity	6.521

Figure 44. Test on example from recent publication on reaction prediction.

A team at MIT has reported four “worst case” examples.[146] Clark tried them in his own system. The first (Figure 45) is multistage; modeling does not suggest a number of stages. The use of ammonia and lithium are suggested. The reported water as a solvent is interesting in connection with lithium. The yield is predicted well but time and temperature are not.

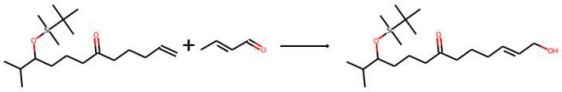


coleyFig4 RX29313449
Stage #1: p-Toluic acid With ammonia; lithium In water at -60°C; for 0.5h; Birch reduction;
Stage #2: 1-bromo-hexane In water regioselective reaction;89%

reagents	lithium (0.46) lithium diisopropyl amide (0.22) n-butyllithium (0.19) naphthalene (0.18) potassium tert-butylate (0.15)
solvents	tetrahydrofuran (1.00) water (0.68) ammonia (0.37) sodium hydroxide (0.34) potassium hydroxide (0.32)
temperature	20C (0.23) 0C (0.22) 50C (0.05)
time	1h (0.21) 12h (0.15) 24h (0.14)
yield	75% (0.40) 100% (0.24) 50% (0.18)
sensitivity	6.830

Figure 45. “Worst case” number one.

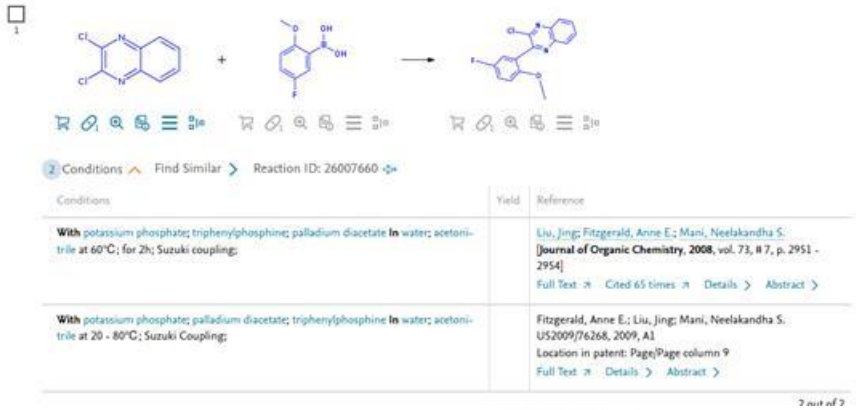
In the second example (Figure 46), ruthenium and Grubbs catalysts are predicted correctly, and predicted time and yield are generally correct.



coleyFig4 RX36087581 Stage #1: 10-(tert-butyldimethylsilyloxy)-11-methyldodec-1-en-6-one; crotonaldehyde With tricyclohexylphosphine[1,3-bis(2,4,6-trimethylphenyl)-4,5-dihydroimidazol-2-ylidine][benzylidene]ruthenium(II) dichloride In dichloromethane for 2h; Grubbs Olefin Metathesis; Inert atmosphere; Reflux; Stage #2: With sodium tris(acetoxy)borohydride In benzene for 16h; Inert atmosphere; Reflux;86%	reagents	tricyclohexylphosphine[1,3-bis(2,4,6-trimethylphenyl)-4,5-dihydroimidazol-2-ylidine][benzylidene]ruthenium(II) dichloride (0.47) Hoveyda-Grubbs catalyst second generation (0.29) Grubbs catalyst first generation (0.18) osmium(VIII) oxide (0.12) 3-chloro-benzenecarboperoxoic acid (0.12)
	solvents	dichloromethane (1.00) tetrahydrofuran (0.22) toluene (0.15) water (0.13) benzene (0.12)
	temperature	20C (0.39) 0C (0.13) 50C (0.12)
	time	24h (0.22) 1h (0.17) 12h (0.16)
	yield	75% (0.47) 100% (0.24) 50% (0.18)
	sensitivity	5.016

Figure 46. “Worst case” number two.

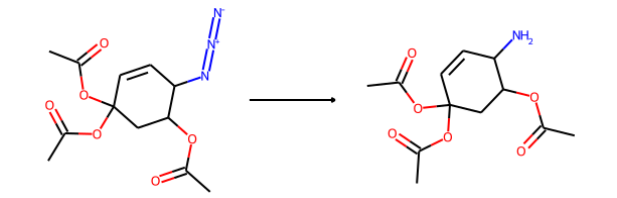
In the third example (Figure 47), the predicted reagents are generally correct, and time is in a generally correct category, but the two reports in the literature demonstrate different temperatures.



reagents		potassium carbonate (0.88) tetrakis(triphenylphosphine) palladium(0) (0.78) sodium carbonate (0.70) potassium phosphate (0.41) palladium diacetate (0.35) bis-triphenylphosphine-palladium(II) chloride (0.25)
solvents		water (1.00) 1,4-dioxane (0.63) toluene (0.39) ethanol (0.25)
temperature		100C (0.62) 20C (0.07) 50C (0.05)
time		12h (0.20) 24h (0.17) 0h (0.12)
yield		75% (0.49) 50% (0.24) 100% (0.18)
sensitivity		8.164

Figure 47. “Worst case” number three.

In the fourth example (Figure 48), different reagents are suggested but they are plausible (e.g., Staudinger reaction and catalyzed hydrogen). The predicted temperature and time are generally correct.



coleyFig4 RX8648319
With diothiothreitol In pyridine; water at 20°C; for 1h;
Reduction;

reagents	hydrogen (1.00) palladium on activated charcoal (0.86) triphenylphosphine (0.75) trimethylphosphane (0.20) zinc (0.12) lithium aluminium tetrahydride (0.08) tin(II) chloride (0.07)
solvents	tetrahydrofuran (0.81) water (0.75) methanol (0.74)
temperature	20C (0.37) 0C (0.05) 50C (0.04)
time	1h (0.25) 24h (0.16) 12h (0.13)
yield	100% (0.45) 75% (0.33) 25% (0.13)
sensitivity	7.197

Figure 48. “Worst case” number four.

Clark also tried a very recent process patent from Reaxys, not in the training set. The recorded conditions were “boron trifluoride diethyl etherate at 30 ^{o}C for 2 hours”. Boron trifluoride diethyl etherate was the top reagent predicted by Clark’s system. Other reagents that might work were also suggested.

Finally, Clark has plotted a graph of all reported synthetic paths to sorafenib back to industrial chemicals and determined the minimum number of reagents, plus condition changes, to carry out the synthesis. He is currently looking at reaction paths and evaluating pathways.

The “ground truth” for assessing accuracy of reaction predication is not straightforward; assessment of retrosynthesis has similar issues. Some reactions have hundreds of conditions that will produce 100% yield: the concept of condition diversity is worth exploring. Several areas of application are enabled by the current level of prediction of conditions. Finding alternative reagents and solvents can be a part of patent circumvention with retrosynthesis. Another use is determining which reactions could take place with a given set of reagents.

Reaction Predictor: reaction prediction at the mechanistic level using deep learning

Pierre Baldi, University of California Irvine, Irvine, CA, USA

Reaction prediction is a complex, multifaceted, problem, in both forward and backwards directions. Reaction conditions and rates, different applications (e.g., retrosynthesis, or “brewing”), and global reactions *versus* mechanistic reactions are some of the aspects. Rule-based reaction prediction systems are based on the manual implementation of a set of rules. They are fast at production time but are limited by the coverage of the rules and hard to maintain. Systems based on quantum mechanics are, in principle, the most accurate and satisfying but they are computationally very expensive and not scalable. Machine learning based systems learn from data. They are fast at

production time, scalable, and relatively accurate. The main problem is that appropriate data may not be available. Hybrid systems are also possible.

2D chemical structure data may be represented as labeled graphs, strings, fingerprint vectors, or lists of atom coordinates. There exist deep learning approaches that can handle each type of representation. If the inputs are vectors of fixed size, as for example in computer vision, a typical architecture is a feedforward neural network. If the inputs are structured objects (e.g., graphs, molecules, or sequences) of variable size, a typical architecture is a recursive neural network (RNN). There are two kinds of approaches for designing an RNN: the inner approach and the outer approach. The inner approach requires that the data and the approach be represented by a directed acyclic graph (DAG). This approach uses RNNs to "crawl" the edges inside the DAG. The outer approach does not require a DAG. It uses RNNs in a direction that is orthogonal to (or outside) the data graph to "fold" the graph. Since molecules are typically described by undirected cyclic graphs, in order to use an inner approach, Baldi's team has developed methods to address the discrepancy, essentially by considering an ensemble of recursive neural networks associated with all possible vertex-centered acyclic orientations of the molecular graph.[147]

Baldi's team has worked on reaction prediction at the mechanistic level using deep learning. Having a mechanistic prediction partially addresses issues of causality and interpretability (and debugging). On the other hand, elementary steps must be chained to form global reactions; this means that such a system may be slower at production time. The main obstacle, however, is the lack of data. In 2009 there was no database of "arrow-pushing" mechanisms, so Baldi's student Jonathan Chen built a set of 1800 SMIRKS-based transformation rules in the Reaction Explorer system.[148,149] Reaction Explorer covers the undergraduate organic chemistry curriculum in an interactive educational system now licensed by Wiley and distributed worldwide. Unfortunately the rules are tedious to update and they cover only a fraction of known chemical reactions. The system is not scalable.

So, Baldi's team started to develop Reaction Predictor, a system based on machine learning.[150,151] The system describes single mechanistic reactions as interactions between coarse approximations of molecular orbitals (MOs) and uses topological and physicochemical attributes as descriptors. Using the existing Reaction Explorer rule-based system, Baldi's team derived a dataset consisting of 1630 full multistep reactions with 2358 distinct starting materials and intermediates, associated with 2989 productive mechanistic steps and 6.14 million unproductive mechanistic steps.

Reaction Predictor learns a ranking model over potential filled-to-unfilled MO interactions such that the top-ranked mechanistic steps yield the major products. The machine learning implementation follows a two-stage approach. The first is atomic prediction of potential source and sink sites: finding the MOs on each atom and predicting atoms with the electron donor (source) and electron acceptor (sink) MOs. Atom level reactivity filters are trained to prune 94% of nonproductive reactions. The second step is ranking the predicted pairs: forming the reactions by letting the source and sink MOs interact, ranking the outcomes, and pruning. An ensemble of ranking models perfectly ranks the productive mechanism at the top 89.05% of the time, rising to 99.86% of the time when the top four are considered.[150] Polar, pericyclic, and radical reaction type ranking models have been successfully developed.[151] Feedforward and convolutional neural networks, RNNs (plus LSTMs),[152] graph neural networks (GNNs), and transformers have been used.

Global reactions can be identified by chaining together the elementary reaction predictions. Baldi's team curated a dataset of over 11,000 elementary reactions. Using these for training, they demonstrated an 80% top-five recovery rate on a separate, challenging benchmark set of reactions. A fundamental problem of synthetic chemistry is the identification of unknown products observed *via* mass spectrometry. Reaction Predictor includes a pathway search feature that can help identify such products through multitarget mass search.[152]

In the final part of his presentation, Baldi discussed the core problem of the data. There is a need for a public, free, downloadable, comprehensive database of reactions. Some seeds exist, for example, the USPTO dataset and datasets in some academic laboratories. The problem with the lack of a publicly available reaction database is somewhat reminiscent of the situation for small molecules about 20 years ago, when the American Chemical Society (ACS) raised opposition to the building of the PubChem[62] database. There is no evidence that PubChem impacted ACS negatively.

Two decades ago, Baldi's team built ChemDB, a public database and web server of small molecules and related cheminformatics resources.[153] It supports multiple molecular formats and is periodically updated, automatically whenever possible. The database includes a user-friendly graphical interface on the web,[154] and chemical reactions capabilities, as well as unique search capabilities.

Nevertheless, Baldi's vision of a long term stable solution to the data problem cannot depend on individual academic laboratories or for-profit corporations. A public database of chemical reactions should ideally be developed and managed by an international consortium (e.g., something like the Conseil Européen pour la Recherche Nucléaire, CERN). Failing that, and more realistically, it could be developed and managed by an agency of the federal government such as the NIH, the U.S. National Science Foundation, or the U.S. Department of Energy or some combination of those. A multipronged approach is needed to populate and sustain such a database, involving negotiations with commercial database companies, aggregation of existing datasets plus crowd sourcing, automatic extraction from the literature and the web using AI approaches, and legislation (e.g., requiring chemical vendors to provide chemical reactions).

Integrating synthetic accessibility with AI-based generative drug design

Brian Atwood, Iktos, Paris, France

De novo molecule generation and optimization can suggest novel molecular structures suitable as therapeutics against a particular target but the utility of these approaches is hindered by ignorance of synthesizability. To highlight the severity of this issue, Gao *et al.* have used a data-driven computer-aided synthesis planning program to quantify how often molecules proposed by state-of-the-art generative models cannot be readily synthesized. The analysis demonstrates that there are several tasks for which these models generate unrealistic molecular structures despite performing well on popular quantitative benchmarks.[155]

Known heuristics methods for scoring synthesizability are fast but fail to distinguish the synthesizability of compounds perfectly. They include SMILES, SAscore,[156] SCScore,[157] and RAscore.[132] The first is a simple heuristic which considers the length of the SMILES. SAscore uses the complexity and frequency of known fragments, on a scale of 1 to 10, the lower the better. SCScore uses a model which encodes the increasing complexity of reactions' sequences (from 1 to 5, the lower the better).

RAscore uses a classification model which predicts the feasibility of a molecule according to AiZynthFinder (from 0 to 1, the higher the better).

Synthetic accessibility prediction is a highly nonlinear task, with steep “feasibility cliffs”: one of two very similar molecules may be very easy to make while the other is very hard to make. Depending on the database of starting materials used, the prediction can go from simple to complex: for example, elvitegravir can be made in one step from an available starting material according to one system but requires 10 steps before an available starting material is found according to another system. We need more than a heuristic.

Synthetic feasibility is also difficult to define, even for expert chemists. Process chemists and medicinal chemists look at synthetic accessibility from a different viewpoint. Iktos asked four chemists to score 100 molecules from a generative AI output on a scale of 1 (hard) to 4 (easy). Only 18% were in complete agreement; 56% agreed within one on the scale; 25% disagreed by two points.

The Spaya[158] program from Iktos employs a data-driven AI approach to discover retrosynthetic routes. An iterative exploration of all possible routes is performed until commercially available starting materials are identified. The reaction and starting materials databases are customizable. An easy-to-use online platform enables chemists to generate and explore retrosynthetic routes. Once users choose a route among the ones found by Spaya, they can easily navigate the retrosynthetic tree. They can also further expand the tree by breaking up the starting materials.

Users take the API and get a retro score (RScore) on a scale of 0 to 1 (the higher the better), after a few seconds or up to one minute per molecule per CPU. The Spaya API scales in the cloud. RScore is defined for a given route and is dependent on the probability of the model, its applicability domain, the number of steps in the route, and the convergence of the route ((except for RScore = 1 which means an exact literature match). Correlation between the RScore and the number of reaction steps has been demonstrated on a sample of molecules from ChEMBL.[47] If RScore is considered as a ground truth, SAscore is the best known metric to assess synthetic feasibility compared to SCScore or SMILES length (Figure 49).

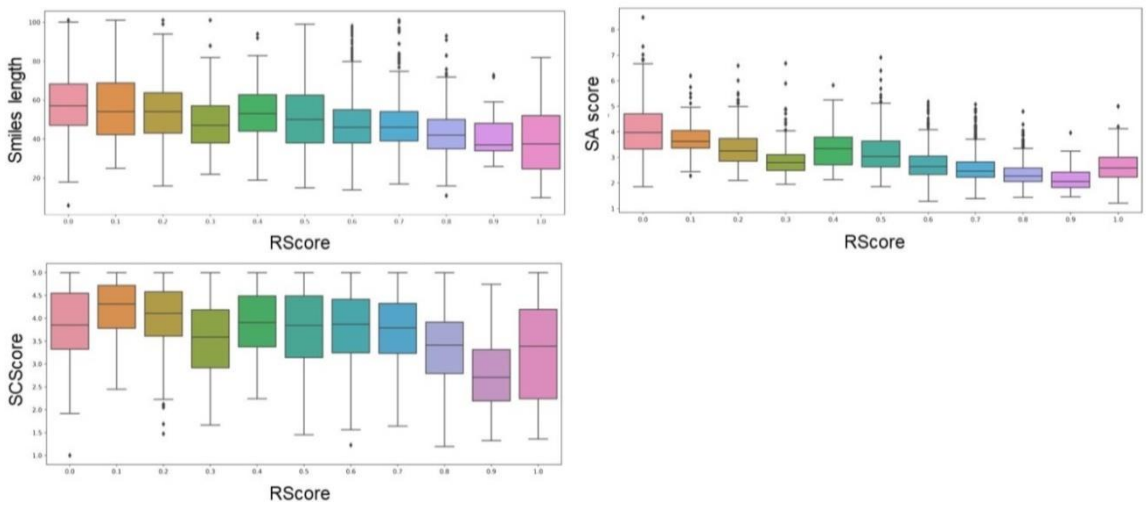


Figure 49. Comparison of RScore with other scores.

Since SAscore is the metric that correlates the best with RScore, Iktos decided to compare the scores on different tasks from the GuacaMol[159] benchmark. There was no major difference between SAscore and RScore on the output in simple tasks: the percentage of molecules with RScore > 0.5 and the average reward on the top 100 molecules.

The team also studied RScore *versus* known scores during generation on more complex tasks. Phosphoinositide 3-kinase (PI3K) pathway is a potential target for cancer chemotherapy and inhibitors of other nodes in the pathway such as mammalian target of rapamycin (mTOR) are also significant. Iktos workers used PI3K and mTOR datasets and evaluated generation against two QSAR kinase activity models with two metrics, including similarity to the initial dataset, and QED, a quantification[160] of "chemical beauty". The PI3K/mTOR dataset serves as a proxy for a real-life lead optimization task, where the more stringent constraints of the QSAR activity models often result in the molecular generator proposing nonsynthesizable molecules.[161] The Iktos team studied the impact of the different scores during the generation, using RScore in postprocessing as a guide (Figure 50). Classic generation and RAscore produce many molecules but most of them are difficult or infeasible to make.

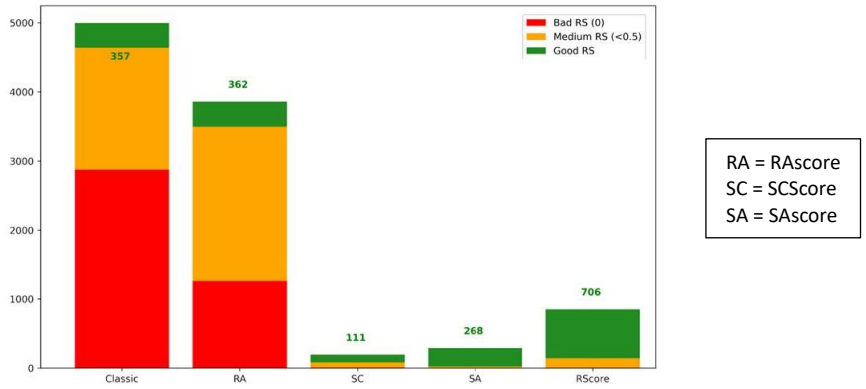


Figure 50. Number of molecules fulfilling the two objectives (PI3K mTOR).

RScore produces more molecules fitting the objectives, with a better diversity, and with a very high proportion of easy-to-make molecules compared to other scores (Table 1). It is interesting to observe that a simple metric like SAscore produces good results in this case study.

Table 1. RScore *versus* known scores during generation on more complex tasks

	blueprint	in GT	Coverage	Simi	Standard murcko	Generic murcko	average RScore	feasible	Good RScore
classic	5005	2	56	0.67	1230	97	0.08	1959	282
RA	3574	2	54	0.69	775	112	0.11	2660	360
SC	211	1	23	0.68	33	19	0.35	202	127
SA	311	2	35	0.76	48	36	0.56	311	286
RS	850	2	51	0.71	102	60	0.49	843	706

Contrary to other scores, Spaya RScore is able to distinguish between complex molecules which are nevertheless easy to make and molecules containing irrelevant moieties: molecules which are obviously not synthesizable and sometimes unstable (Figure 51).

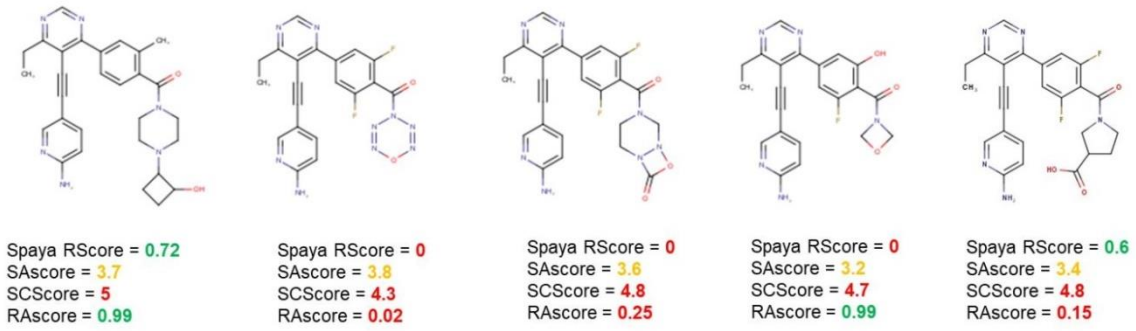
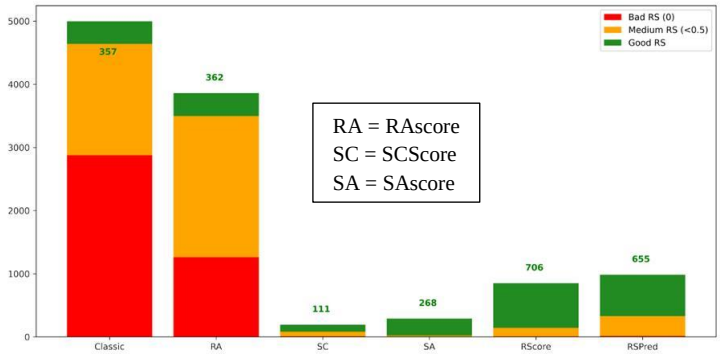


Figure 51. Some examples of generative model output.

Contrary to heuristics or a predictive model, RScore can be easily customized. Iktos studied the impact of the database of commercially available starting materials on a benchmark of known molecules. RScore can be tuned for synthetic feasibility or for synthetic complexity assessment. The chemistry can also be changed, focusing on classic reactions such as amide coupling or cross-coupling.

RScore is highly accurate and produces very good results, but it takes a bit more time to compute than Iktos would have liked, so they developed RSPred, based on a neural network trained on 230,000 molecules retrosynthesized by Spaya API. The area under the receiver operating characteristic curve (ROC AUC) was 0.96 for a neural network model based on Morgan fingerprints to predict RScore > 0.4. RScore and RSpred produce comparable results (Figure 52) but the time needed is reduced from seconds to 10 milliseconds.



	blueprint	in GT	Coverage	Simi	Standard murcko	Generic murcko	average RScore	feasible	Good RScore
classic	5005	2	56	0.67	1230	97	0.08	1959	282
RA	3574	2	54	0.69	775	112	0.11	2660	360
SC	211	1	23	0.68	33	19	0.35	202	127
SA	311	2	35	0.76	48	36	0.56	311	286
RS	850	2	51	0.71	102	60	0.49	843	706
RSPred	985	2	54	0.73	167	90	0.46	971	655

Figure 52. RScore and RSPred.

Iktos has used RSpred in a complex client project recently, concerning AI generation with eight objectives (activity, selectivity, and absorption, distribution, metabolism, and excretion (ADME)). RSpred appears to outperform SA score by a very large margin in this complex multiparameter optimization project.

In summary, using generative AI without a synthetic accessibility heuristic leads to ugly molecules. Among the known heuristics, SAscore appears to be the best one to generate appealing molecules. RScore appears to be better than SAscore, producing more, easy-to-make, diverse molecules. This may be due to the fact that it runs a full retrosynthesis, requiring that the input molecule can be made from an available starting material. Contrary to other scores, RScore can be easily customized either by modifying the scope of the chemistry or the catalog of starting materials. Despite the quality of the output of RScore, computing this score is time-consuming and not really suitable for generative AI. RSpred appears to reach a similar performance to that of RScore but it is much faster, at a much lower cost, which makes it a very attractive tool for generative AI. RScore is accurate and very precise so there is great value in its score for postprocessing and prioritization of designs, whereas RSpred is very efficient for generative AI.

Data-driven synthesis planning beyond retrosynthesis

Connor Coley, Massachusetts Institute of Technology (MIT), Cambridge, MA, USA

Classically, the discovery of physical matter, such as in lead optimization for drug discovery, is divided into stages of design, make, and test. An analogous cycle for searching hypothesis space could be described as hypothesize, validate, and revise beliefs. This is exemplified by preclinical drug discovery, where discovering a new drug candidate can take over two years and cost more than $500 million. The need is to use information more efficiently (make better compounds) or to obtain information more quickly (test more compounds), or both. Central to the scientific process is this inherent trade-off between the value of information, and the cost required to get that information. We constantly make these decisions when we do research, whether it is conscious or not. In order to tip the scales, Coley's team want to make it easier to test the performance of candidate compounds, and have focused on compound synthesis as a significant rate- and cost-determining step.

Coley has a vision for autonomous synthesis[162,163] where the user inputs a command such as "make ibuprofen", software converts this into instructions for a synthesis robot, and the output is a sample of ibuprofen. In data-driven synthesis planning we move from the concept of "known compounds made through known routes incorporating known reaction types" to the concept of "new compounds made through new routes incorporating known reaction types". Even in current manual workflows, computer-aided synthesis planning (CASP) can alleviate some manual tasks, and ideally, bring greater objectivity to route design.

Assuming retrosynthesis rules are not written manually, template-based methods for retrosynthesis involve algorithmic extraction of templates (rules), and determining the relevance and applicability of a template. In the extraction procedure,[164] the core of the transformation is found, generalized neighbors are added, the template is extended to known functional groups, and SMARTS patterns from atom-mapped reaction SMILES strings are canonicalized and recorded. A neural network is trained to predict the most probable rules to apply to a particular reaction,[165] and augmentation and pretraining[166] teach the neural network an increased set of templates that could theoretically lead to successful reactions for a given target.

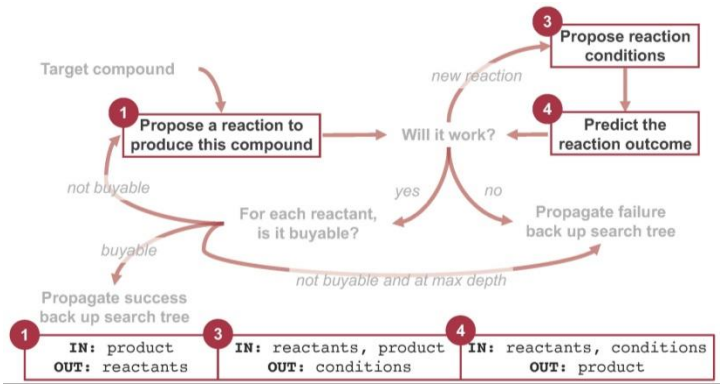


Figure 53. Workflow for algorithmic synthesis design.

The workflow[167] is shown in Figure 53. The CASP steps are: (1) one-step retrosynthesis, (2) multistep planning, (3) condition recommendation, and (4) reaction outcome prediction. Given a product molecule, what reactants could produce it? Given a product molecule and a one-step retrosynthesis model, which multistep pathway starts from commercially available starting materials? Given a product molecule and reactants that could potentially produce it, what are the actual conditions that should be used for the experiment? If I run this experiment as planned, what products do I expect to make?

The third task is condition recommendation. Full condition combinations are too sparse in the datasets for training to predict reaction conditions. Each (catalyst, reagent(s), solvent(s)) set is unlikely to have been seen many times, and relatively few species are actually used. Restricting the answer set[146] to 803 catalysts, 2247 reagents, and 232 solvents excludes only 5% of reactions in Reaxys.[44]

Gao *et al.* developed a neural-network model to predict the chemical conditions most suitable for any particular organic reaction.[146] The task of condition prediction can be divided into two parts: chemical context prediction (catalysts, solvents, reagents) is treated as a set of multiclass classification problems (i.e., choosing chemical species from a fixed list), while temperature prediction is treated as a regression problem. Trained on about 10 million examples from Reaxys,[44] the network 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.

The fourth task is prediction of reaction outcomes. The issue is that there are many input-output pairs, but it is not known how to write an exact function to relate them. In one approach, Coley and his co-workers have used template-free prediction as a sequence of graph edits.[168] They have reported a supervised learning approach to predict the products of organic reactions given their reactants, reagents, and solvent(s). By training a graph convolutional neural network model on hundreds of thousands of reaction precedents from the patent literature, the neural model makes informed predictions of chemical reactivity. The overall model structure is designed to reflect how expert chemists approach the task (Figure 54).

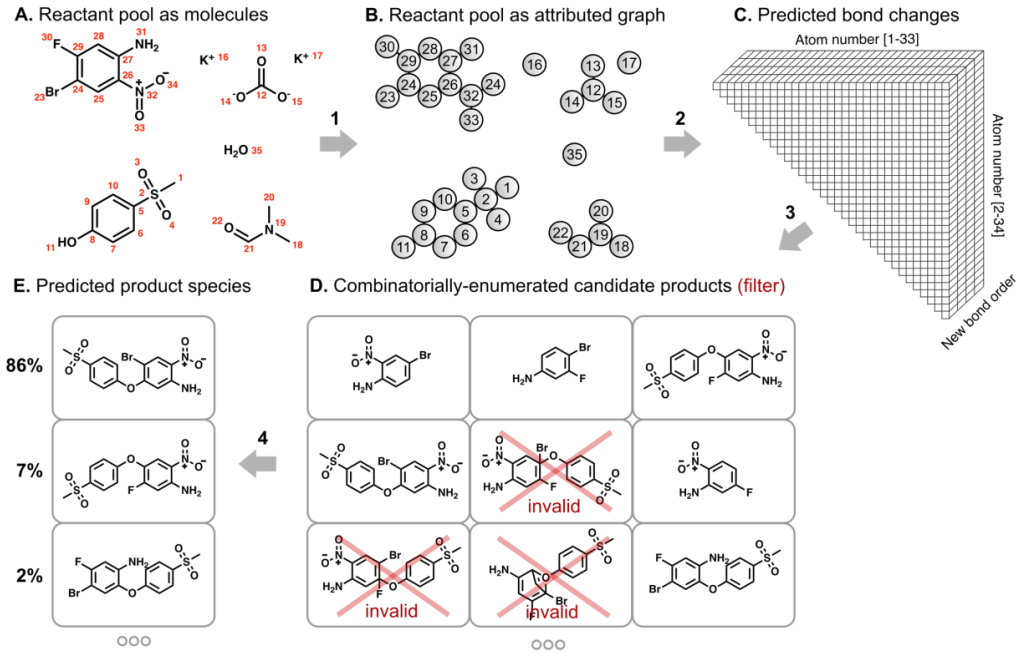


Figure 54. Prediction of reaction outcomes.

First, the system learns to identify reactive sites that are most likely to undergo a change in connectivity (arrow 2 above). This parallels the identification of reactive functional groups and consideration of how they might react, but without codifying rigid rules about functional group decomposition. Next, the software performs a focused enumeration of products that could result from those interactions subject to chemical valence rules (arrow 3). It then learns to rank those candidates (determining what modes of reactivity are most likely, as would a chemist) to produce the final prediction of major products (arrow 4).

One of the things that all this research is enabling is rapid ideation of full synthetic pathways. Longer, more complex pathways are possible. Chemists will be able to study many possible pathways to a target, not the shortest, cheapest, highest-yielding, "best" pathway. Application to semi-automated synthesis is also possible. The ideal automated synthesis platform would be capable of planning its own synthetic routes and executing them under conditions that facilitate scale-up to production goals. Individual elements of the chemical development process (design, route development, experimental configuration, and execution) have been streamlined in previous studies[169-172] but no one has presented a path toward integration of CASP, expert-refined chemical recipe generation, and robotically executed chemical synthesis. Recently Coley and his colleagues have reported a proof of concept toward fully autonomous synthesis.[173]

If reduced to liquid handling steps, automation is "easy". "Automated synthesis" platforms all require some expert chemist guidance when asked to make a new molecule (e.g., selecting concentrations, vessel types, or reaction times). If we manage to get a starting point of > 0% yield, empirical reaction optimization is possible.

There have been many recent advances in retrosynthesis. Rule-based, expert systems (since LHASA) include Chematica (SYNTHIA),[174,175] IC*SYNTH*,[176] SciFinder,[119] and Reaxys.[44] Rule-based, learned approaches have been reported by Segler and Waller,[177] Coley *et al.*(ASKCOS),[164,173] Genheden *et al.*

(AiZynthFinder),[178] Dai *et al.*,[179] and molecule.one,[180] Galixir, Iktos (Spaya),[158] and PostEra.[181] Learned approaches have used nearest-neighbor,[112] neural translation,[110,182,183] and graph translations.[184-186]

Data-driven methods work best on molecules similar to training molecules. Standard small, druglike molecules are handled well by all of these approaches. Expert rules can be better at extrapolating from reactions with few precedents where evaluating the appropriateness of generality or specificity in algorithmic extraction is challenging. Factors such as stereoselectivity or regioselectivity are hard to predict by all data-driven methods, especially when controlled by sterics or distant directing groups.

Many approaches to condition recommendation have been reported. Reaxys-trained classifiers have been used in a "global" approach (with "any" reaction as input).[145,146,187] "Local" approaches to condition prediction treat a specific reaction family.[188-190] Global approaches to product and yield prediction have addressed outcome prediction[111,112,168,191-193] or selectivity.[194-198] Local approaches to product and yield prediction have addressed high throughput experimentation of a single type[199,200] and catalyst screening.[201,202] Local data-driven models do not generalize between reaction family HTE campaigns. None of the local or global methods is useful for discovering *new* synthetic transformations; at best, they can discover novel pairs of known electrophiles and nucleophiles.

Chemists are actually starting to use these data-driven synthesis tools routinely for route scouting: discovery chemists are using routes as proposed and process chemists are using them for idea generation. Data-driven methods can be retrained easily on the most recent reaction data. These tools can help accelerate chemical development but they are not providing reliable suggestions that are immediately actionable (e.g., using robotics), or discovering new synthetic methods. Better data will lead to better methods, and combining computation with lab automation will enable "exploration". The tools are not expanding synthetically accessible chemical space; removing the need for expert chemist expertise; or helping in low data environments.

In retrosynthetic planning, data-driven tools are not helping with complex natural product synthesis (better data, from CAS, for example, and better methods are needed), and they are not perfectly generalizing from rare reactions. In condition recommendation, data-driven tools are not proposing catalysts and ligands to enable fundamentally new transformations. In reaction outcome prediction, data-driven tools are not operating at the mechanistic level or extrapolating to new reaction types. Is there a compromise with Baldi's ReactionPredictor (see above)? Work is ongoing to ground models in physical organic chemistry.

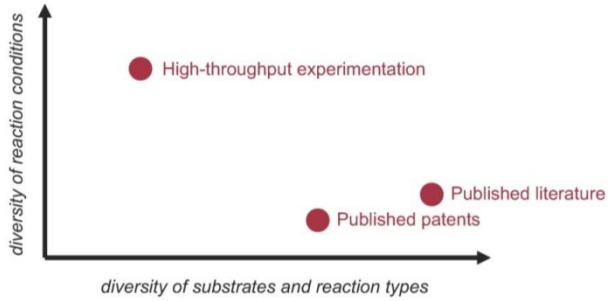


Figure 55. Information sources affect method development.

Information sources affect method development (Figure 55). Currently absent from databases are order of addition, addition speed, ambient temperature and humidity, reagent purity, chemical vendor, and other useful information.

Learning the language of chemical reactions using transformer models

Philippe Schwaller, IBM Research Europe, Rüschlikon, Switzerland

Lowe *et al.* have text-mined U.S. patent data to derive reactions and reaction SMILES.[13,105] Lowe admitted that, while typically correct, the atom-to-atom maps are wrong in many cases, and hence should not be entirely relied upon. So, the IBM team have mainly developed approaches that are independent of atom-mapping, although they recently developed their own atom-to-atom mapping program, RXNMapper.[33]

Earlier, Schwaller and his colleagues had used the updated version of the USPTO reaction dataset and treated reaction prediction as a data-driven, machine translation problem between SMILES strings of reactants plus reagents, and the products.[111] To map the sequence of the reactants and reagents to the sequence of the products, they used an attention-based model borrowed from human language translation. In standard seq2seq, two distinct recurrent neural networks (RNN) work together: an encoder that processes the input sequence and emits its context vector *C*, and a decoder that uses this representation to output a probability over a prediction. In seq2seq with attention there is one state per input, and attention provides the ability to concentrate selectively on one aspect of context. The Molecular Transformer[111] architecture uses multihead context attention:[203] no recurrent neural networks are needed, but stacks of attention layers.

An enhanced version of the Molecular Transformer predicts regioselective and stereoselective carbohydrate reactions using transfer learning.[197] It was experimentally validated on a 14-step synthesis of a lipid-linked oligosaccharide but the transfer learning approach should be applicable to any reaction class of interest.

A retrosynthetic version of the transformer has been developed in conjunction with the University of Pisa.[183] The transformer also predicts reagents, catalysts, and solvents and it is not dependent on atom-by-atom mapping. Since no chemical knowledge is embedded other than the information learnt from reaction data, the quality of the datasets plays a crucial role in the performance of the prediction models built by Schwaller and his co-workers. Toniato *et al.*[114] have proposed a machine learning based, unassisted approach to remove chemically wrong entries from chemical reaction collections. The results show an improved prediction quality for models trained on the cleaned and balanced USPTO and Pistachio[103] datasets.

To make its AI model accessible, IBM has made IBM RXN for Chemistry freely available.[204] Users draw reactants and run a prediction. They get back the product and a confidence score. There is also a retrosynthesis feature. IBM RoboRXN for Chemistry[205] is a pioneer project combining AI, automation and cloud to accelerate material discovery. Vaucher *et al.* devised a method to convert unstructured experimental procedures written in English to structured synthetic steps (action sequences) reflecting all the operations needed to conduct the corresponding chemical reactions.[206,207] Chemical recipes are thus converted to machine-readable instructions. The team generated a dataset of 693,517 chemical equations and associated action sequences by extracting and processing text from patents. They used the dataset to train three different models: a nearest-

neighbor model based on recently introduced reaction fingerprints, and two deep-learning sequence-to-sequence models based on the transformer, and on bidirectional and auto-regressive transformer (BART)[208] architectures. An analysis by a trained chemist revealed that the predicted action sequences are adequate for execution without human intervention in more than 50% of the cases.

Reaction transformer models may transform from a sequence to a sequence or from a sequence to a single value or label. In the latter case no decoder is needed. Schwaller and his colleagues have shown that such transformer-based models can infer reaction classes from text-based representations of chemical reactions.[209] The best model reaches a classification accuracy of 98.2%. The team also showed that the learned representations can be used as reaction fingerprints that capture fine-grained differences between reaction classes better than traditional reaction fingerprints. The insights into chemical reaction space enabled by the learned fingerprints have been illustrated by an interactive reaction atlas providing visual clustering and similarity searching. Schwaller and his colleagues have also predicted reaction yields, given a text-based representation of the reaction, using an encoder transformer model combined with a regression layer, and have demonstrated outstanding prediction performance on two sets of high-throughput experiment reactions.[210]

Bidirectional encoder representations from transformers (BERT)[211] and a “lite” version, ALBERT,[212] are designed to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. The resulting pretrained BERT or ALBERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks.

Schwaller and his co-workers have developed a BERT reaction classification model (Figure 56) with an encoder made of stacks of self-attention layers. All self-attention layers consist of multiple attention heads. Using a classifier head, the model was applied to a chemical reaction classification task. The encoding of the [CLS] token can also be used as a reaction fingerprint (RXNFP).

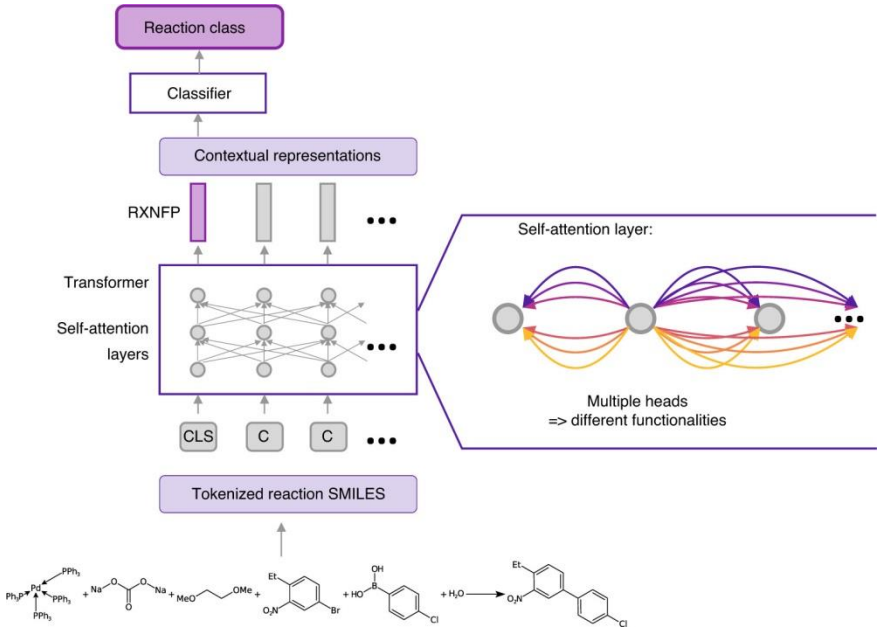


Figure 56. BERT reaction classification.

The pretraining stage is unsupervised; training data are unlimited. The researchers pretrained the models using masked language modeling loss (Figure 57) on the chemical reactions. The task of the model in masked language modeling consists of predicting individual tokens of the input sequence that have been masked with a probability of 0.15. A special class token [CLS] was prepended to the tokenized reaction SMILES. The [CLS] token was never masked during the self-supervised training. The embeddings of the [CLS] token were taken as input for the classifier head. The next stage is supervised training with fine-tuning for two tasks: BERT plus classification head for reaction classification, and BERT plus regression head for reaction yield prediction.

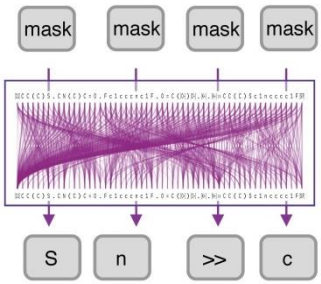


Figure 57. Self-supervised training.

In another work, Schwaller and colleagues solely pretrained BERT/ALBERT models on unlabeled reactions and asked what knowledge the models had extracted from the reaction data during pretraining. It was visual inspection of the attention heads that led the IBM team to realize that the system had learned atom-by-atom mapping: unsupervised transformers capture the hidden grammar of chemical reactions. The team used this signal to build an atom-mapping tool called RXNMapper.[33]

In the RXNMapper algorithm, attention weights learned by a transformer, without atom-mapping supervision or human labeling, encode atom rearrangement information between products and reactants. A neighbor multiplier algorithm gets attention weights from the model and iterates through product atoms, mapping the most likely pair and increasing attention from the product neighbors to the corresponding reactant neighbors. The mapper shows a remarkable performance in terms of accuracy and speed (7 seconds for 1000 imbalanced reactions). On 49,000 strongly unbalanced patent reactions,[109] RXNMapper predicted 99.4% of atom mappings correctly (up from about 96% without the neighbor multiplier). Transformers capture the grammar of chemical reactions and could be used in synthesis planning, quantum mechanical simulations, and reaction accessibility and interpretability studies.

Template-based approaches to reaction planning are dependent on atom mapping in template building;[112,173,177,213] GNN-based approaches need atom maps for graph edits.[168] SMILES-to-SMILES approaches are independent of atom mapping.[111,183] All systems can benefit from better atom mapping. A demonstration of RXNMapper[214] and open source code[215] are available on the web.

Structuring time course sensor data for improved chemical outcomes and reproducibility

Chris Smith, DeepMatter, Glasgow, Scotland

DeepMatter digitizes chemistry, developing products that combine laboratory hardware and state-of-the-art software to enable improved reproducibility, predictability, and speed in scientific outcomes. The company's products span fields such as pharmaceutical research and development, fine chemicals, scientific publications, and education. DigitalGlassware[216] simplifies data collection and data structuring, bringing real-time sensor data from the lab to a browser, and producing a minable data repository of high quality data. IC*SYNTH*[176] is a computer-aided synthesis design and retrosynthesis tool that supplies novel synthetic routes to a target compounds, runs fast calculations against high quality curated data, and can be integrated with customers' proprietary data.

The company o2h discovery has an integrated drug discovery platform operating from a state-of-the-art research center in India. Late in 2020, DeepMatter carried out some chemistry experiments with o2h to improve reproducibility and improve yield using a more structured approach to the chemistry using time-course data. They then analyzed the data using machine learning.

A Buchwald-Hartwig coupling reaction was used: a palladium-catalyzed "one-pot" amination which is a reliable metal-catalyzed amine-based coupling. There were four steps in the experimental protocol. Chemist A used a published paper to execute a chemical synthesis in triplicate in the traditional way, analyzing the reaction with high-performance liquid chromatography (HPLC) over time. Chemist B encoded and refined the protocol as a digital recipe, using the recipe builder in DigitalGlassware. Chemist A, using the recipe developed by Chemist B, performed the reaction in triplicate, analyzing the reaction with HPLC over time. The whole process allowed for data analysis using machine learning on subsequent work.

DeepMatter codify their chemistry using recipes: digital protocols as a basis for structured information, stored in the cloud. These are easy to create, share and recreate. The gap between human- and machine-readable procedures is bridged in this migration from "sketch and text" to readable, writable and controllable XML. DigitalGlassware provides intent and context when the

chemist ultimately considers outcomes such as the expected volume of addition set against the actual volume of addition.

The reaction in the o2h use case was analyzed every 30 minutes using HPLC since DeepMatter wanted time series comparison with the data captured by DeviceX, the self-contained multisensor device, from DigitalGlassware, that autonomously monitors chemical reactions. A significant quantity of time series data across a portion of the reaction space was captured by the digital controller (Figure 58). Note the increase of about 50% in yield and the 80% reduction in error.

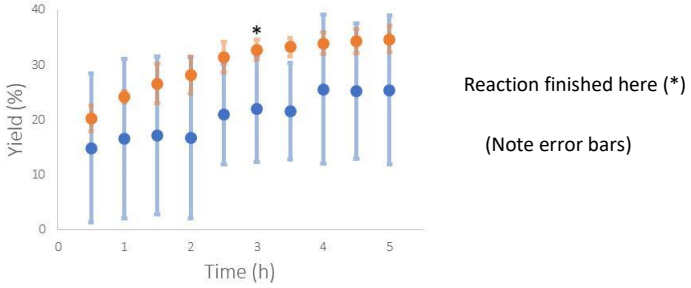
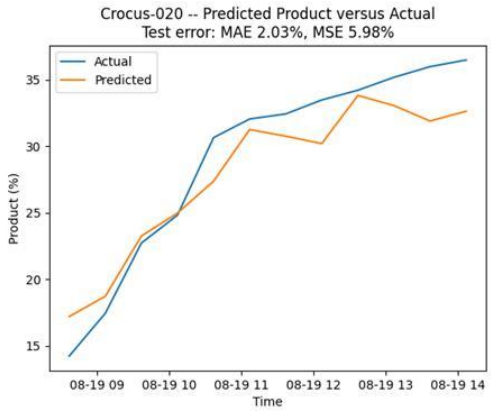


Figure 58. Yield assessed by HPLC, triplicate measurements.

Next, DeepMatter used data science to explore the data further and model them. The first step was to try and visualize the data, in this case, the aggregated sensor data with respect to HPLC outcomes. In scatter plots where each point represented data associated with an HPLC sample, a pattern existed indicating that there was indeed a correlation between sensor data and outcome over time, albeit noisy in some parts. A time series among individual runs and recipe versions could be observed indicating that modeling could go ahead using the derived features.

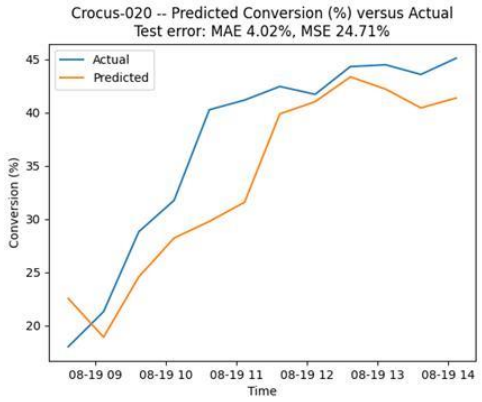
The first analysis was yield prediction. The objective was to predict product formation over time, given sensor and recipe data as input. In process-style chemistry, where repeated runs are similar, predicting yield over time enables the chemist to detect failing chemistry or anomalous progress, and to decide whether to continue chemistry. Predicting yield over time allows data scientists to start to model reaction space (optimize yield, lower cost etc.), and reduces demand on the HPLC machine. The modeling used a variety of approaches and dataset permutations but gradient-boosted regression trees performed the best. The results (Figure 59) indicated that modeling yield is possible; some runs had a very low error rate (<3% mean absolute error (MAE)). The indications of which features were most important were noisy, but there was no “killer feature”. Hold over temperature was highly ranked and UV seemed to be a consistent indicator of progression. The system allows real-time, *in situ* prediction of yield.



MAE = mean absolute error
MSE = mean squared error

Figure 59. Yield prediction.

The objective of predicting reaction completion was similar to that of yield prediction, but was focused on conversion instead. The objectives were to predict what percentage of the original starting material has been converted into another product and predict whether the reaction had finished rather than the yield at which it finished. The input data features and modeling approaches investigated were the same as for yield prediction. Yield and conversion correlated very closely, but not 1:1. The results (Figure 60) were slightly better than yield prediction in some cases. There were overestimates and underestimates in some cases, but the same pattern often showed. Some chemists may care more about whether a reaction has completed than about the amount of product formed.



MAE = mean absolute error
MSE = mean squared error

Figure 60. Predicting reaction completion.

Also analyzed were sensor swimlanes. A swimlane diagram is a type of flowchart that delineates who does what in a process. It provides clarity and accountability by placing process steps within the horizontal or vertical “swimlanes” of a particular employee, work group or department. It shows connections, communication, and handoffs among these lanes, and it can serve to highlight waste, redundancy and inefficiency in a process.

In the present example, the objective is to visualize adherence to or deviation from previous runs of the same reaction. It could be used for exploring both options. One objective is quickly to identify deviations from the norm in behavior of target variables, and so reduce wastage. The analysis could also be used as a means of iteratively refining processes to get marginal gains, and thus increase

profit. The analysis used only original sensor data, on a per sensor stream basis. It modeled the signals from historic runs to obtain the “average” signal over time. It required mapping of signal features between runs: even repeated chemistry is never exactly replicated in terms of time and duration. It was shown that chemists can now, at run-time, be alerted to deviations detected in behavior. The analysis is complementary to the yield and conversion models.

Another use case was work performed with a DeepMatter industry partner on some novel chemistry previously thought to take 20 hours. The effects of color change and pH were studied. Using a digitized recipe and sensor data, DeepMatter discovered that the process is actually complete in 20 minutes. Three correlations for reaction completion were found, including one previously unknown correlation.

Chemistry as an observation-based science can lead to poor reproducibility and knowledge loss. Human interaction to make chemical products introduces many opportunities for error. Labs of the future are modernizing data capture and dissemination through digitization; data transfer and sharing are at the heart of this revolution. DeepMatter are committed to open data standards with the Practical Chemistry Markup Language (PCML), Practical Chemical Runtime Record (PCRR) XML standards, and the XDL format (see the talk by Cronin below). Aggregated, structured, time-course datasets improve real-time and post-run analytics, leading to better productivity and discovery. Industry 4.0 will move toward full AI integration in the chemistry pipeline (robotics, automated synthesis, reaction optimization etc.), for which strong commercial and public datasets are required.

A universal approach to reaction informatics

Leroy Cronin, University of Glasgow, Glasgow, Scotland, United Kingdom

Dedicated automated synthesis instruments have been constructed for peptide and oligonucleotide synthesis, flow chemistry, sugar chemistry, cross coupling, and radiosynthesis. The Eli Lilly synthesis laboratory is more widely applicable in that it covers an entire laboratory but their system is not universal: the code used in there is limited to the precise hardware and software built for the Lilly system and hence it is a one-off system. The reactionware,[217] chemputer, chemputation, and chemical synthesis language devised by Cronin’s team[172] are, in contrast, universally applicable in terms of their architecture, language and hardware requirements.[218,219]

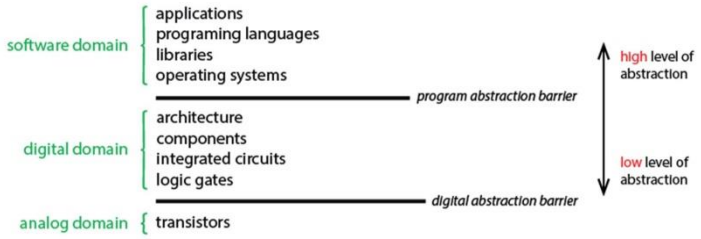


Figure 61. Levels of abstraction.

Universality in chemistry is vital and requires a well-defined set of abstracts, as was required for digital computing (Figure 61). Currently, chemistry lacks rigorous mappings and abstractions. The chemistry literature contains more than a century’s worth of instructions for making molecules, all written by and for humans, not computers, and hence is ambiguous because a lot of tacit knowledge

is needed, but practical chemistry can be encoded from abstractions (Figure 62). The XDL and Chemify[220] project is dedicated to a 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.

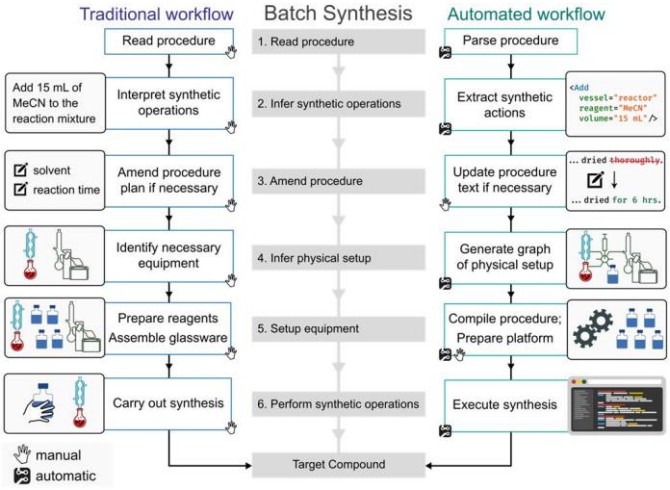


Figure 62. Encoding practical chemistry from abstractions.

Cronin's team has developed the first chemical state machine (the architecture of which was formerly called a "chemputer"): an autonomous compiler and robotic laboratory platform to synthesize organic compounds on the basis of a chemical programming language that removes ambiguity and defines a new standard system for a chemical programming language, XDL (pronounced "chi-DL").[172] XDL is a "Turing-complete" programming language doing for chemical reactions because the language abstraction can represent any possible function to make any compound. This language for expressing chemistry and materials science has been released free under an MIT or Apache 2.0 license.[221] It is compatible with *any* robot system and will link with the Open Reaction Database, Standardization in Lab Automation (SiLA 2)[222], optimization and machine learning routines, and RXN files[6] etc.

Cronin hypothesized that a 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, would lead to a new era of reproducibility and reliable discovery as well as the ability to collaborate and scale reactions. Chemputation is the process of running XDL code reliably on *any* compatible hardware (named by analogy with computation, the Turing-complete running of programs on a digital computer). He calls this architecture and abstraction the Chemical Processing Unit or ChemPU. The chemical synthesis state machine is universal because the abstraction of chemical assembly leads to a state machine that can make any molecule or material on any machine or robot. Inputs are digital and physical; outputs are physical. XDL is human- and machine-readable and is verified for reproducibility and security.[218]

The requirements placed on software and hardware for the full digitization of chemistry are shown in Figure 63: (a) the basic requirements and scope of a programming language for chemistry and (b) the executable procedures. The API should contain operations that are easily recognized by chemists, and the syntax should be as straightforward as possible.

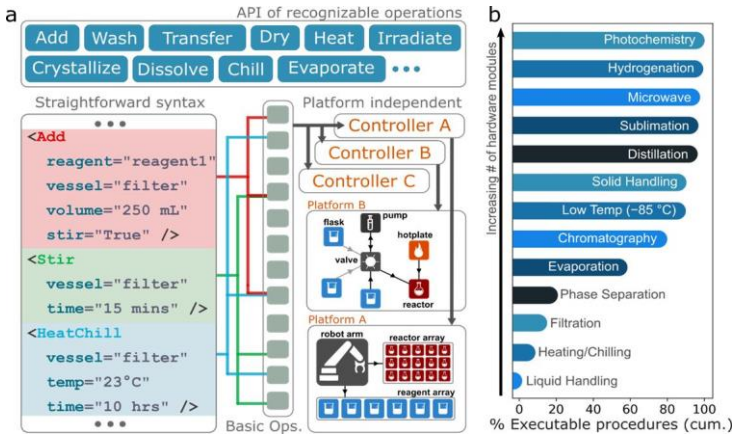


Figure 63. Software and hardware requirements.

Cronin demonstrated the procedure: copy the text and translate it to XDL (Figure 64), fix any translation errors, give the project a title, generate a graph of the physical setup required, compile the XDL and graph together to produce an XDL execution file, simulate the execution file to check for runtime errors, set up a physical platform to match the graph, and run the execution file.

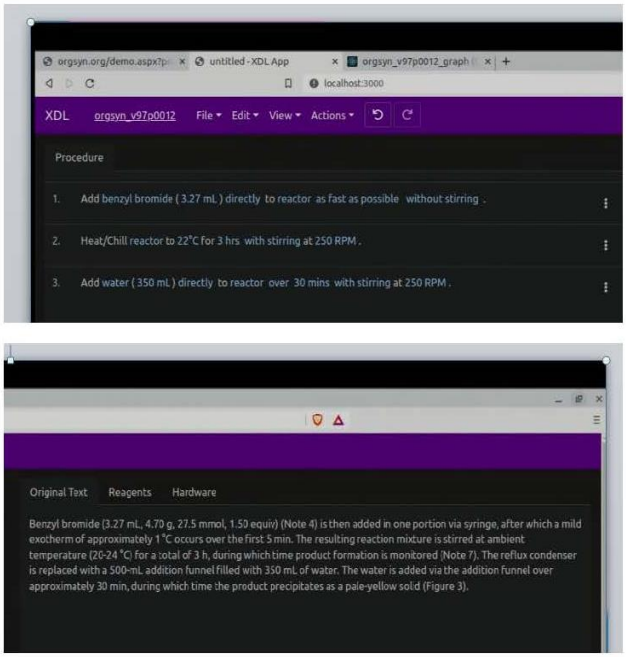


Figure 64. XDL screen shot (split down center).

XDL is moving toward XDL 2 which will allow parallel reactions, more complex reaction planning, and integration of medium and high throughput systems (see Figure 65 which shows a new setup for stacking and scheduling reactions). Since XDL is fully formed language, it will be much more reliable than a scripting system which is important for security, verification and reliability.

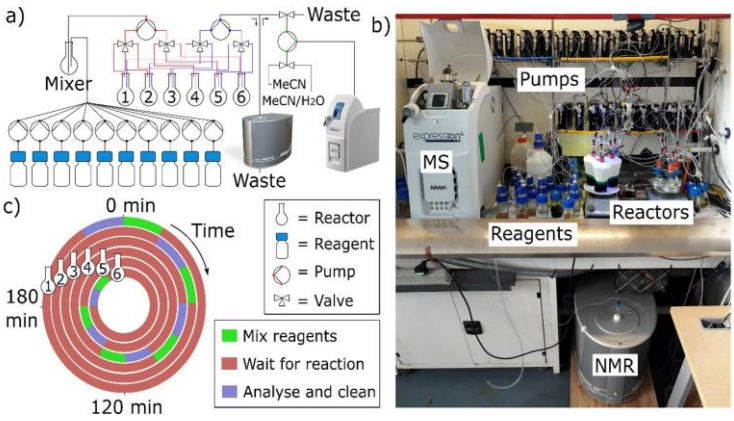


Figure 65. Setup for stacking reactions.

The integration of sensor systems for monitoring and controlling the state of the chemical synthesis machine, and the integration of high resolution spectroscopic tools are vital if these systems are to facilitate closed-loop autonomous experiments which then can update the XDL in real time. Systems that not only make molecules and materials, but also optimize their function, and use algorithms to assist with the development of new synthetic pathways and process optimization become possible.[219]

Digital-chemical robot systems need to integrate feedback from simple sensors and online analytics in order to navigate process space autonomously. This will open the door to accessing known molecules (synthesis), exploring whether known compounds or reactions are possible under new conditions (optimization), and searching chemical space for unknown new molecules, reactions, and modes of reactivity (discovery). Cronin's team is currently working on parallelization, and concurrent synthetic threads, for higher throughput. The overall vision is that following from the first Chemputer robot, the system has developed into a universal architecture that can run on any qualified hardware. The process of running the XDL chemical programs is Chemputation and the process of checking if XDLs can run on any given configuration safely and reliably will need the Chemputability of the procedures to be confirmed thus ushering in the era of digital chemistry.

A map of the amine-acid coupling system

Tim Cernak, University of Michigan, Ann Arbor, MI, USA

Analysis of the role of synthetic organic chemistry in hit and lead optimization efforts suggests that only a few reactions dominate. Thus, the uptake of new synthetic methodologies in drug discovery is limited.[223] Amide coupling (Figure 66) is the single most-used reaction. It works well; it is robust and trusted. Chemical transformations determine the structure of a product, and therefore its properties, which in turn affect complex macroscopic functions such as the metabolic stability of pharmaceuticals or the volatility of perfumes. Therefore, reaction selection can influence the success or failure of a candidate molecule to meet a functional objective. Amide coupling is popular but there are many other ways to connect an amine with a carboxylic acid. Cernak's team have shown computationally that amines and acids can couple *via* hundreds of hypothetical yet plausible transformations, and they have demonstrated experimentally the application of a dozen such reactions.[224]

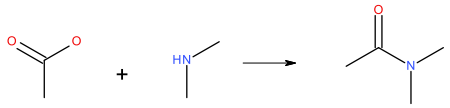


Figure 66. Amine-acid coupling.

To study the contribution of chemical transformations to properties, the team developed a string-based notation and used an enumerative combinatorics approach (Figure 67) to produce a map of conceivable amine-acid coupling transformations, which could be charted using cheminformatics techniques.

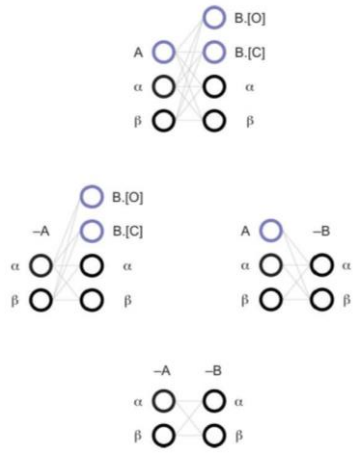


Figure 67. Enumerative combinatorics.

Graph enumeration in a simple focused set of amines and acids suggested different products which can be produced using different catalysts (Figure 68). Each had a different property set depending, for example, on the number of hydrogen bond donors and acceptors in the acid molecule. In one case, log*P* changed by two orders of magnitude when the catalyst was changed.

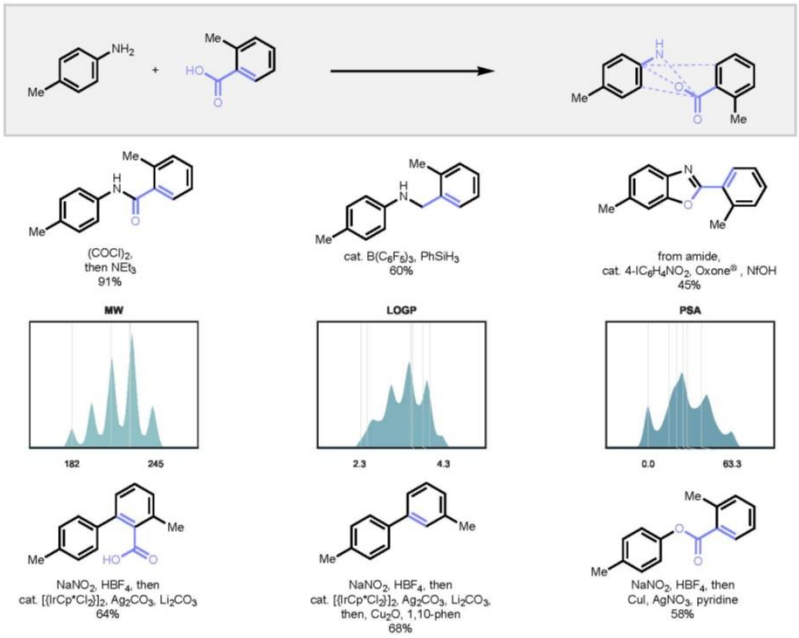


Figure 68. Results of amine-acid enumeration.

Cernak showed a map of the amine-acid-coupling system (Figure 69) with molecules from DrugBank,[225] showing many connections between theoretical compounds and known drugs. The high degree of connectivity between hypothetical reactions and drugs is simply observed through the presence of many connecting (purple) lines. The researchers conclude from this that most, if not all, of the amine-acid coupling reactions in the map could find use in drug discovery and synthesis, since the reactions produce common drug substructures.

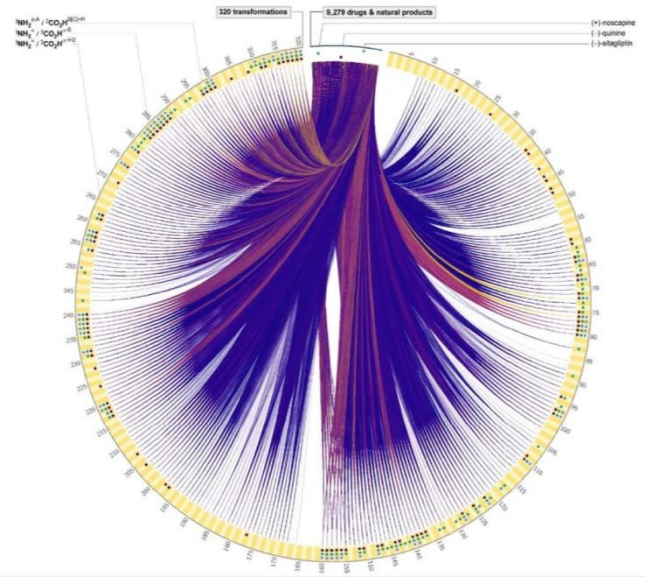


Figure 69. Map of the amine-acid coupling system.

Data mining the amine-acid coupling system produced should enable reaction discovery: after learning from purple lines in the map, new coupling reactions can be tried with HTE. Cernak's team demonstrated this by developing an esterification reaction (Figure 70) found within the mapped

space. Complex molecules with distinct property profiles can also be discovered within the amine-acid coupling system, impacting the late-stage diversification of drugs and natural products.

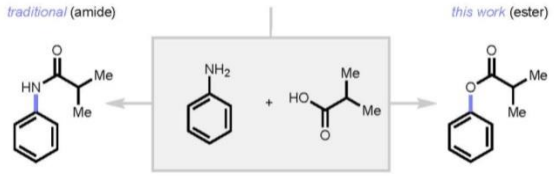


Figure 70. Amine-acid esterification.

HTE has evolved over the past few decades as a tool for experimental reaction development. The beauty of HTE is that reactions are run in a systematic format, so data points are internally consistent, the reaction data are reported whether the desired product is observed or not, and automation may reduce the occurrence of false positive or negative data points. Also, experimental workflows for HTE lead to datasets with reaction metadata that are captured in a machine-readable format. Cernak's team has reported technologies, with case studies, for running synthetic reactions in parallel from the milligram to microgram scale in glass vials and plastic well plates.[226]

In ultrahigh throughput experimentation (ultraHTE), reactions are run in ~1 µL droplets inside of 1536-well microtiter plates to minimize the use of starting materials while maximizing the output of experimental information. The performance of ultraHTE in 1536-well microtiter plates has led to an explosion of available reaction data, which have been used to identify specific substrate-catalyst pairs for maximal efficiency in novel cross-coupling reactions.[227]

Cernak's team have developed software, called phactor, to facilitate the performance and analysis of HTE in a chemical laboratory.[228] It allows experimentalists to design arrays of chemical reactions in 24-, 96-, 384-, or 1536-well plates. Users can access online reagent data, such as a lab inventory, to populate wells, and produce instructions to perform the screen manually, or with the assistance of a liquid handling robot. After completion of the screen, analytical results can be uploaded for evaluation, and to guide the next series of experiments. All chemical data, metadata, and results are stored in a machine-readable format.

In as yet unpublished work, Cernak's team have studied the amine-acid esterification shown in Figure 71. Sixteen catalysts were tried with 6 ligands, and 8 bases with 12 ligands. The results are shown in Table 2.

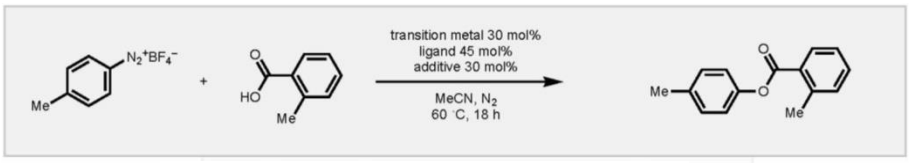


Figure 71. HTE: amine-acid esterification.

Table 2. Amine-acid Esterification, Results

Entry	Catalyst	Catalyst mol %	Base	Solvents	NMR Yield (%)
1	CuBr	30	2,4,6-collidine	MeCN	44
2	$Cu(MeCN)_4BF_4$	30	2,4,6-collidine	MeCN	91(89)
3	$Cu(MeCN)_4PF_6$	30	2,4,6-collidine	MeCN	54
4	$(CuOTf)_2PhH$	15	2,4,6-collidine	MeCN	81
5	CuI	30	2,4,6-collidine	MeCN	17
6	$(CuOTf)_2PhH$	20	2,4,6-collidine	MeCN	92
7	$Cu(MeCN)_4BF_4$	40	2,4,6-collidine	MeCN	94
8	$Cu(MeCN)_4BF_4$	30	K_2CO_3	MeCN	8
9	$Cu(MeCN)_4BF_4$	30	TMEDA	MeCN	5
10	$Cu(MeCN)_4BF_4$	30	LiOH	MeCN	18
11	$Cu(MeCN)_4BF_4$	30	None	MeCN	0
12	None	N.A.	2,4,6-collidine	MeCN	0
13	$Cu(MeCN)_4BF_4$	40	2,4,6-collidine	BuCN	37
14	$Cu(MeCN)_4BF_4$	40	2,4,6-collidine	PhCN	27
15	$Cu(MeCN)_4BF_4$	40	2,4,6-collidine	THF	0
16	$Cu(MeCN)_4BF_4$	40	2,4,6-collidine	DMAc	0

The team has had success with the SPT Labtech mosquito in liquid handling for nanomole-scale high-throughput chemistry.[229-231]

The mosquito robot was used to interrogate many different activated amine and acid substrate combinations using ultraHTE. The results were run in quadruplicate to ensure reproducibility from the nanoscale result. Most substrate pairs gave the product and the reproducibility of the esterification reaction across four reaction replicates was very good. Reaction "hits" from the 1536-well experiments were repeated on a traditional synthesis scale of ~25-50 mg to obtain isolated yields, which were generally good and in agreement with the nanoscale result. This technique was used in the late-stage diversification of complex pharmaceuticals, such as sulfadoxin.

More broadly, Cernak proposed that new amine-acid coupling reactions will be of high value in late stage diversification, and demonstrated the application to a variety of drugs that carried amine or acid functional handles. Thus, from a single substrate pair, but varying the amine-acid transformation through the judicious selection of reaction conditions, they were able to obtain diverse analogues with different physicochemical properties such as pK_a, log$P$, and HBD.

Orchestrating automated synthesis: designing actionable routes

Christos A. Nicolaou and Todd de Collo, Eli Lilly & Company, Indianapolis, IN, USA

In the drug discovery process, hypotheses are tested in a DMTA cycle. Advances have been made in design, make and test stages. *In silico* structure design is now commonplace. Aiding the "make" stage are automated synthesis laboratories, automated purification laboratories and CASP. In the "test" stage are automation and robotics for screening. There has been an explosion in the number of publications on AI in cheminformatics and computer-aided structure design.

The Proximal Lilly Collection (PLC),[108] aims firstly, to define the chemical space of small, druglike compounds that could be synthesized using in-house resources and secondly, to facilitate access to compounds in this space for the purposes of drug discovery. Through library design and virtual screening, promising compounds are identified, filtered computationally, and scored, prioritized and selected for synthesis and testing.

DNA-encoded library (DEL) technology is a novel ligand identification strategy that allows the synthesis and screening of unprecedented chemical diversity more efficiently than conventional methods. Nicolaou and his colleagues have systematically studied how to increase the diversity of DELs and improve the molecular property space that can be covered. They have developed and applied eDESIGNER, an algorithm that comprehensively generates all possible library designs, enumerates and profiles samples from each library and evaluates them to select the libraries to be synthesized.[232] This tool uses suitable on-DNA chemistries and available building blocks to design and identify libraries with a predefined molecular weight distribution and maximal diversity compared with compound collections from other sources.

Highly integrated functions and processes are needed from start to finish in the DMTA cycle. With this in mind, Lilly created an integrated, globally accessible, automated chemical synthesis laboratory (ASL)[169] and an automated purification laboratory. Building on this, Lilly created the Lilly Life Science Studio in San Diego to enable a computationally driven approach to the DMTA cycle, physically integrating several areas of the drug discovery process. This automated, cloud-based platform consists of 16 autonomous, yet interconnected, automated workstations for functions such as compound and reagent management, synthesis, and purification; and analytical, biological and biophysical testing. The concept behind the current laboratory automation paradigm is “idea to molecule”. The paradigm behind the Lilly Life Sciences Studio next generation drug discovery platform is “idea to data”.[233]

There has been an explosion in the number of publications on AI in cheminformatics and CASP (many of them cited in this report). The Lilly “ChemoPrint” software[234] is CASP in practice. It is a chemical context aware, data-driven method built upon millions of available reactions, with attractive run-time characteristics, to recommend synthetic routes matching a precedent-derived template. Coupled with modern automated synthesis platforms and available building block collections, the method enables drug discovery researchers to identify routes for target compounds which are easy to interpret and implement.

Hurdles and challenges in closing the drug discovery loop are determining if a compound is really synthesizable, and estimating the real cost and the opportunity cost. Reducing everything to practice involves route instantiation: selecting the best route, going from theoretical routes to actionable recipes, and getting inspiration from observation (i.e., learning lessons from history) by taking advantage of known reaction procedures, and condition recommendation.

Assessing the synthetic feasibility of a structural hypothesis involves making (and testing) the compound in question. Selecting the optimal route in not straightforward: there are too many options, some better than others, and the reactants identified may give rise to more than one reaction and product. Expert rule scoring can be used in assessing routes. These are heuristic rules, defined by expert chemists, concerning reactant availability, and reaction ease, robustness, and simplicity, and amenability to automation.

Another approach is attempting the forward reaction of generated routes. Reactants for each matching route and a reaction template are input and all potential (predicted) products are output, with a flag for any reactant pair generating more than one product. This is a brute force approach and it is slow. The MIT-led Machine Learning for Pharmaceutical Discovery and Synthesis (MLPDS)

consortium[235] is working on forward prediction models and impurity prediction, scoring the “best” route.

The major missing link in closing the drug discovery loop is moving from a theoretical synthesis route to an actionable synthesis execution recipe: building block IDs, and the source; reaction conditions; detailed synthesis instructions (i.e., a procedure or recipe); and a complete execution order for automated systems (for expert review, not execution). Factors to consider are the procedure (from an ELN, for example), reaction conditions, and route instantiation (translation to a machine-ready synthesis workflow). Lilly validated five chemistries (amide coupling, Buchwald-Hartwig coupling, Suzuki coupling, sulfonamide formation, and urea formation) in 2019 and 15 more in 2020.

One example of inspiration from observation is the work of Schneider *et al.* on unraveling the content of the medicinal chemist’s toolbox. The researchers used a sophisticated text-mining pipeline to extract 1.15 million unique whole reaction schemes, including reaction roles and yields, from pharmaceutical patents. The reactions were assigned to well-known reaction types using an expert system, and the evolution of reaction types over time was analyzed.

Lilly’s virtual synthesis engine uses an annotated reaction repository for which a reaction database and an ontology were developed using NextMove Software’s NameRxn.[32] About 2 million reactions were classified into more than 700 reaction types using the ontology implementation. More than 60,000 chemical reactions have been executed in the ASL system by more than 220 researchers.[108] Lilly’s reactions are expert-reviewed rather than expert-defined. The system capitalizes on the availability of clean, robust reaction data in Lilly’s synthetic history (Synthory) database which is continuously updated, with adaptive learning.

Lilly are carrying out ongoing work on identifying actionable routes. Lilly’s “eLN” containing millions of reactions has been mined, cleaned, standardized, and characterized. A subset containing associated automated synthesis workflows has been made. This will allow search for a proposed route: finding a similar automated reaction, and identifying the synthesis workflow. The system will recommend an actionable route, extract an automated workflow reaction template, and instantiate it with the desired reaction reactants, agents, etc. Results so far are promising.

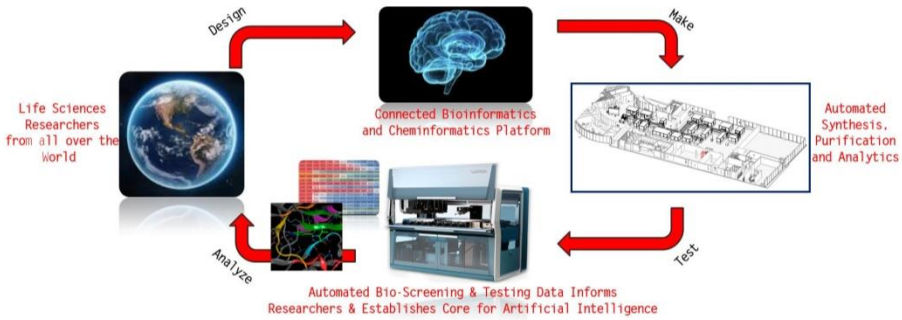


Figure 72. The Idea2Data vision.

We are getting close to the possibility of autonomous discovery (Figure 72) but the missing link is still instantiation of the route in a machine language.

The SynRoute computational synthetic planning tool: lessons learned

Mario Latendresse, James Herson, Markus Krummenacker, Peter Madrid, Jeremy Malerich, and Nathan Collins, SRI International, Menlo Park, CA, USA

SRI Biosciences has developed SynFini, a suite of tools that makes use of AI to automate the translation of ideas into testable physical molecules.[171] The SynFini platform helps scientists maximize time spent on what to make, rather than how to make it, by accelerating chemistry design, development and synthesis. Deep Adaptive Semantic Logic (DASL)[236] is a framework for automating the generation of deep neural networks that incorporates user-provided formal knowledge to improve learning from data. SynRoute is a CASP tool that rapidly designs multistep routes. SynJet, for fast reaction screening and optimization, performs synthesis on a µg to mg scale. It is coupled with SynRoute. The AutoSyn automated bench chemistry platform is a miniaturized, flow chemistry plant with integrated analytics. Finally, data are generated from rapid *in vitro* bioassays and ADMET tests.

AutoSyn incorporates a "cityscape": a miniaturized chemical plant.[171] 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 two 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.

SynJet features 24 reagent dispensers, a transfer arm, six heater blocks, and a vial hopper. A customized Inkjet dispenses around one reaction a second for a 10 µL reaction. Reaction processing is highly parallel with screening for independently varied conditions including time, temperature, solvent, reagents, stoichiometry and catalysts. Chemical analysis is by HPLC/MS (at 120 seconds per reaction).

SynRoute needs to work with both AutoSyn and SynJet. It has a web interface designed for use by chemists who enter a target compound as SMILES, InChI, InChIKey, or common name, or by drawing a structure, or inputting a structure as a file. They enter constraints (e.g., the maximum number of steps or maximum cost). Recently saved routes based on target compounds are displayed. After a route search, the results are displayed as strategies (e.g., 2-5 steps are summarized so the user can see the complexity of the route). Users then view strategy 3, route 1, say, and can refine it by avoiding or keeping certain compounds or reactions. They can curtail a route by selecting feedstock. A bill of materials can be prepared to help the chemist to order materials quickly. The route can be saved and put into the SynFini platform for creation of a process from a route. Pumps and a reactor are chosen and the operator of the synthesis system can load the pumps and choose solvents and reagents to be used.

Fifty-nine reaction transformations from the medicinal chemist's toolbox (MCT)[237] have been incorporated in SynRoute. The route searching algorithm (Figure 73) generates reactions that connect to feedstocks or synthesizable compounds. A subnetwork is selected according to constraints such as route length. A diversified *k*-best route is applied based on cost and length of routes: shorter routes are better. Multiple routes are found, then clustered based on strategies. A typical search time is 20 to 90 seconds.

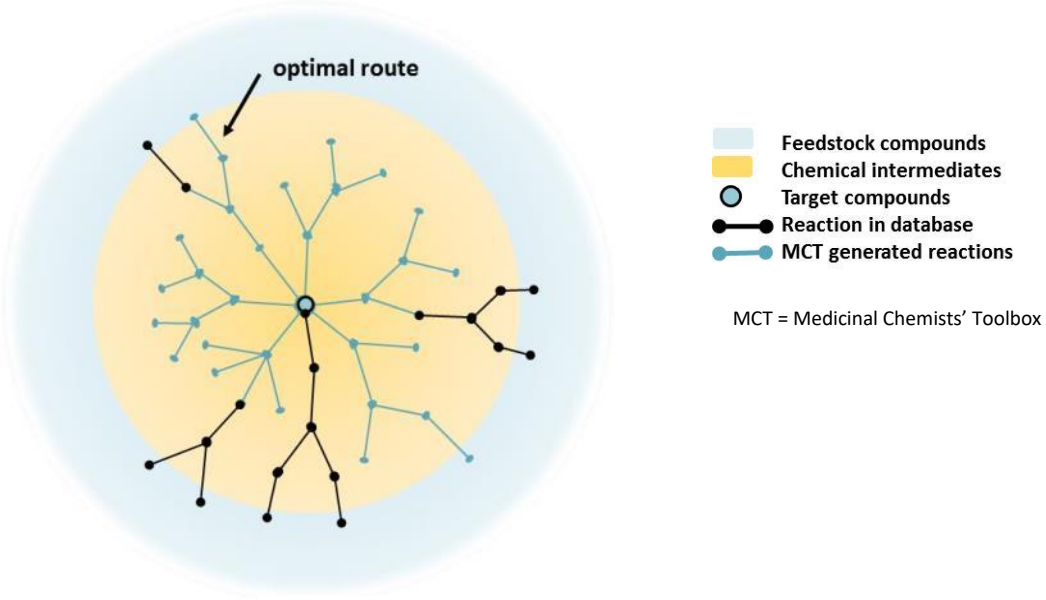


Figure 73. SynRoute search algorithm.

Latendresse and his colleagues created and trained classifiers for each transformation from the MCT. The goal is to predict whether a computer-generated reaction is workable. During a route search, all applicable MCT transformations are used to create new reactions. The machine learning 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.[238]

Routes are "diversified" with different strategies: variations on their first retrosynthetic reaction. Finding multiple diversified routes is essential for the chemists to get a broader view of what is possible. This is technically done in the searching algorithm by spreading the routes found on each strategy. SynRoute uses a curated version of the Pistachio database[103] from NextMove Software. Reactions that could not be classified are not used. As purchasable building blocks, SynRoute uses tier 1 and tier 2 of the eMolecules catalog.[239] (Tier 3 molecules cannot be readily purchased.)

Latendresse tested the performance of the classifier algorithm. Classifiers were created and trained for each MCT transformation. The encoding of reactions uses a sparse vector of atom classes. An atom class is defined as an atom species, its properties, and its direct neighbor atom species and properties with their bond types. Considering all atoms in the Reaxys database,[44] 27,429 classes were extracted. The classifiers have an average accuracy of 90% based on cross-validation studies with literature data (Figure 74). The overall performance of SynRoute is shown in Table 3.

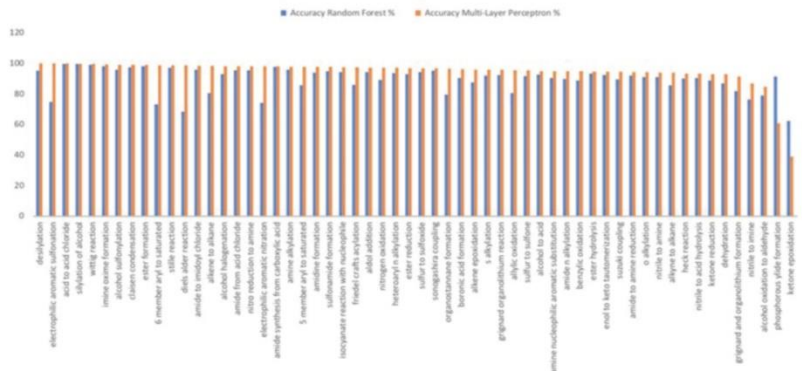


Figure 74. Accuracy of MCT classifiers.

Table 3. Overall Performance of SynRoute

Reaction Sources	FDA Approved Drugs	ChEMBL Med Chem Literature
Only MCTs	560/1359 (41.2%)	127/504 (25.2%)
Only Reaxys	668/1359 (49.1%)	178/504 (35.3%)
Reaxys + MCTs	1078/1359 (79.3%)	311/504 (61.7%)

Reaction templates (i.e., transformation rules) can be generated automatically from a set of reactions. Typically, reaction SMIRKS/SMARTS are used to represent the templates as retrosynthetic transformations, for example, ester formation: [C:5][C:2](=[O:1])[O:3][C,c:4]>>[C:5][C:2](=[O:1])[OH].[OH:3][C,c:4]. The published techniques create about 100,000 specific templates to cover the given reaction set. These techniques have a major drawback: it is difficult to find the applicable and appropriate templates given a molecule. One possible solution is to reduce the number of templates by generalizing them.

Unfortunately, the SMIRKS representation does not have well-defined semantics. For example, the well-known RDKit[240] implementation does not implement the “or” syntax as expected in [NX3;!$(NC=O):1][CX4:2]>>[N:1].[F,Cl,Br,I][CX4:2]. It does not produce four different molecules, as expected from the or notation “[F,Cl,Br,I]”, but only one. There is a need for a more well-defined representation with semantics that do not depend on a unique implementation (e.g., RDKit).

Innovative graph technologies for synthesis route design in ASPIRE.

Gergely Zahoránszky-Kőhalmi,[1] Mark Backus,[1] Wendy M. Charles,[2] Busola Grillo,[1] Manideep Gurumurthy,[1] Tyson S. Henry,[2] Brian D. Jackson,[2] Robert C. Lubeck,[2] Nikita Lysov,[1] Biju Mathew,[1] Dimitrios Metaxotos,[1] Byron N. Nash,[2] Frank J. Ricotta,[2] Gianna Ricotta,[2] Rafat Sarosh,[1] James A. Scarpella,[2] Nick Schaub,[1] Ke Wang,[1] Alexander G. Godfrey;[1] [1]National Center for Advancing Translational Sciences (NCATS/NIH), Rockville, MD, USA; [2]BurstIQ Inc., Denver, CO, USA

NCATS proposes to transform chemistry from an individualized craft to a modern, information-based science through “A Specialized Platform for Innovative Research Exploration” (ASPIRE).[241] By addressing long-standing challenges in the field of chemistry, including lack of standardization, low reproducibility, and an inability to predict how new chemicals will behave, ASPIRE is designed to bring novel, safe and effective treatments to more patients more quickly at lower cost. ASPIRE builds on the power of recent and emerging technological innovations such as chemical laboratory

automation, microfluidic flow chemistry, high-throughput screening, and machine learning. This convergence of technologies, together with innovative cross-disciplinary engineering, provides a new opportunity to break translational bottlenecks in chemistry and benefit science and health. This initiative promotes multidisciplinary collaborations among government, academic and pharmaceutical researchers; funders; professional societies; scientific publishers; and other stakeholders.

Challenges in the collaborative research environment include sharing data with third parties, and accessing the data needed for evidence-based synthesis route design, and for building AI and machine learning models. There is a need to keep track of the chain of custody. Secure reaction information storage also presents challenges: database access allows structures to be revealed, the risks of reverse mapping public InChIKeys to potentially novel reactions must be reduced, and substructure and similarity search need to be carried out in a protected database.

Precomputed reaction graphs can be stored in a Neo4j[242] multipartite graph database. Substances are encoded as InChIKeys; reactant, reagent, and product information are treated as edge labels; and "metanodes" represent annotation according to the RXNO name ontology.[69] In a PostgreSQL[243] database with the RDKit cartridge,[240] SMILES are stored *only* as encrypted strings and substructure search uses on-the-fly decryption (in memory). These databases can be used in evidence-based synthesis route design.

NCATS is collaborating with BurstIQ[244] in a blockchain pilot project to reduce the cost and risk of collaborative chemical reaction research. The advantages of blockchain are its tamper-resistant and tamper-evident data history, tracking of data access (chain-of-custody), consent contracts (granting and revoking permissions), scalability, and strong encryption, all with minimal manual administration.

Reaction protections and collaborations are enhanced with blockchain. Access to reactions by reaction ID or InChIKey is controlled using cryptographic ownership and control, governance, and consent management. The chemical reactions cannot be reconstituted from the blockchain-based encryption. Access to blockchain assets can be controlled with granular, dynamic consent. Only the asset owner(s) and controller(s) can grant, modify, or revoke permissions for individuals or groups (projects) to access reactions at a very granular level, from a petabyte of data to a single data object. The blockchain serves as a single source of truth with a full audit trail. This facilitates data sharing with both trusted and untrusted partners. This approach also ensures that researchers are able to share and contribute to each other's research without losing control of valuable intellectual property.

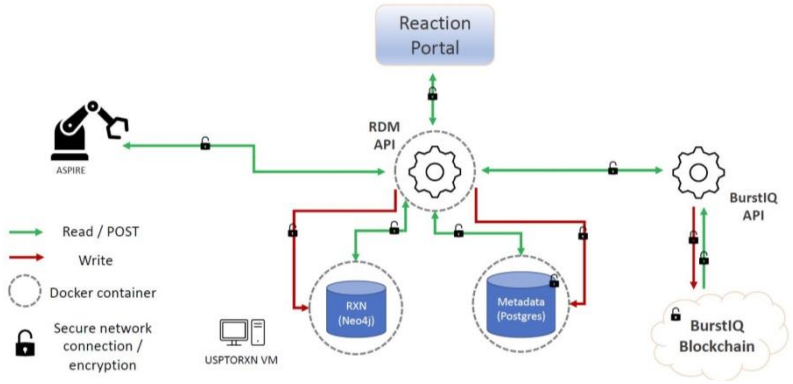


Figure 75. ASPIRE reaction data management system.

The ASPIRE reaction data management (RDM) system (Figure 75) has an API to:

- find a reaction by reaction ID
- find a reaction route by a target molecule's InChIKey
- find similar reactions (based on reaction fingerprints)
- find reactions with a reactant, reagent, or product similar to a query structure
- find reactions where a reactant, reagent or product contains the query substructure
- find reactions associated with an RXNO reaction name (class)
- depict a molecule (as a structure in SVG format).

A prototype reaction portal has been built. Next steps are integration of blockchain functionality into all RDM API endpoints; integration of a consensus-based secure data sharing mechanism into the reaction portal; extending synthesis route design features; and exploring further utilization of the BurstIQ platform in knowledge base design.

Conclusion

While the event on ultralarge databases dealt largely with virtual libraries, the current one is more concerned with realities. There are useful large collections of curated reaction data such as Reaxys, CAS Reactions, and Pistachio but they are commercial. The most commonly used open source collection is the one loosely called "USPTO" and it covers only reactions published in patents. Other issues are the problems of bias, noise, and missing data fields. There is a particular problem with missing and erroneous data on reaction yields. The Open Reaction Database described herein aims to make reaction data freely and publicly available in a structured data format but the project is only in its infancy.

The reaction representations discussed here start with the oldest, CHMTRN/PATRAN and run through *de facto* standards such as RDfiles and RXN files, and SMIRKS and SMARTS, and through CGR, right up to recent XML options, RInChI, and UDM. A new format, Reaction SPL, based on an existing standard, has also been proposed. There is a reaction ontology (RXNO). Software developments in connection with the SAVI and Pistachio databases and the open source Chemotion ELN and Repository are also noteworthy.

There is much interest in transformer models for reactions: a number of sections herein discuss such systems. Atom-to atom mapping is an essential for some reaction informatics software applications,

but not for others. The IBM RXN for Chemistry team have mainly developed approaches that are independent of atom-mapping, but they recently developed their own atom-to-atom mapping program, RXNMapper. Reaction informatics software applications discussed include *de novo* drug design, predicting reaction conditions, retrosynthesis, reaction outcome prediction, and synthetic accessibility (including various scoring systems).

Access to synthetically accessible molecules in virtual libraries was a critical component of the previous NIH conference on ultralarge chemistry databases. Designing a new molecule is of little practical interest if it cannot be realized synthetically. Hence there is an interest in software to predict the feasibility of synthesis, and automation to streamline the actual synthesis. The fourth, and final, theme in this report is progress toward autonomous synthesis. Lilly's Idea2Data vision, SRI's SynFini, IBM RoboRXN, Cronin's "Chemputation" and Cernak's ultra-high throughput experimentation are reported herein. Standard formats are needed to link synthetic routes to robots. This is another topic for this report. The universality of XDL is one of the subjects explored. DeepMatter also report adherence to standards in their work on time-course sensor data. Finally, the ASPIRE project reports an interesting experiment with blockchain.

As a result of significance of the subject matter of the two NIH workshops, and the interesting discussions that have taken place, the *Journal of Chemical Information and Modeling* has issued a call for papers for a special issue[245] on reaction informatics and chemical space, edited by Matthias Rarey, Marc Nicklaus, and Wendy Warr, to be published early in 2022. This will highlight the achievements of recent years, showcase current research, and motivate scientists in academia and industry alike to explore the opportunities in navigating chemical space.

Acknowledgments

I am grateful to the organizers for all their support in helping me to prepare this report, and especially to Marc Nicklaus and Janelle Cortner of NIH. Every speaker was invited to check and correct the text of his or her individual presentation. I am grateful to all the speakers for their helpfulness and cooperation. Most of the presentations and recordings are currently available on the NIH website.[246]

References

(1) Warr, W. A. Report on an NIH Workshop on Ultralarge Chemistry Databases. http://chemrxiv.org/engage/chemrxiv/article-details/60c75883bdbb89984ea3ada5 (accessed August 2, 2021).
(2) Corey, E. J.; Wipke, W. T.; Cramer, R. D., III; Howe, W. J. Computer-assisted synthetic analysis. Facile man-machine communication of chemical structure by interactive computer graphics. *J. Am. Chem. Soc.* **1972,** *94* (2), 421-430.
(3) Patel, H.; Ihlenfeldt, W.-D.; Judson, P. N.; Moroz, Y. S.; Pevzner, Y.; Peach, M. L.; Delannée, V.; Tarasova, N. I.; Nicklaus, M. C. SAVI, in silico generation of billions of easily synthesizable compounds through expert-system type rules. *Sci. Data* **2020,** *7* (1), 384.
(4) CACTVS toolkit from Xemistry. http://xemistry.com/ (accessed June 29, 2021).
(5) SMIRKS: a reaction transform language http://www.daylight.com/dayhtml/doc/theory/theory.smirks.html (accessed July 7, 2021).
(6) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; 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.

(7) Corey, E. J.; Wipke, W. T. Computer-assisted design of complex organic syntheses. *Science* **1969,** *166* (3902), 178-192.
(8) Delannee, V.; Nicklaus, M. C. ReactionCode: format for reaction searching, analysis, classification, transform, and encoding/decoding. *J. Cheminf.* **2020,** *12* (1), 72.
(9) ReactionCode. http://cactus.nci.nih.gov/reactioncode/ (accessed June 29, 2021).
(10) Judson, P. N.; Ihlenfeldt, W.-D.; Patel, H.; Delannee, V.; Tarasova, N.; Nicklaus, M. C. Adapting CHMTRN (CHeMistry TRaNslator) for a New Use. *J. Chem. Inf. Model.* **2020,** *60* (7), 3336-3341.
(11) Swienty Busch, J. What are the next steps in your synthesis? The Reaxys experience. (CINF 48. Transcript available from wendy@warr.com). In *250th National Meeting of the American Chemical Society, Boston, MA, August 16-20, 2015*.
(12) Kraut, H.; Eiblmaier, J.; Grethe, G.; Loew, P.; Matuszczyk, H.; Saller, H. Algorithm for reaction classification. *J. Chem. Inf. Model.* **2013,** *53* (11), 2884-2895.
(13) Lowe, D. M. Chemical reactions from US patents (1976-Sep2016). http://figshare.com/articles/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873 (accessed June 21, 2021).
(14) Krause, S.; Willighagen, E.; Steinbeck, C. JChemPaint - using the collaborative forces of the Internet to develop a free editor of 2D chemical structures. *Molecules* **2000,** *5* (1), 93-98.
(15) Google Blockly. http://developers.google.com/blockly/ (accessed June 29, 2021).
(16) Kiho, Y. K. The formal definition of some concepts of quantitative organic chemistry. *Org. React. (Tartu)* **1972,** *8* (2), (In Russian).
(17) Vladutz, G.; Gould, S. R. Joint compound/reaction storage and retrieval and possibilities of a hyperstructure-based solution. In *Chemical Structures: The International Language of Chemistry;* Warr, W. A., Ed.; Springer-Verlag: Berlin/Heidelberg, Germany, 1988; pp 371-384.
(18) 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.
(19) Jauffret, P.; Hanser, T.; Tonnelier, C.; Kaufmann, G. Machine learning of generic reactions. 1. Scope of the project; the GRAMS program. *Tetrahedron Comput. Methodol.* **1990,** *3* (6A), 323.
(20) Jauffret, P.; Tonelier, C.; Hanser, T.; Kaufmann, G.; Wolff, R. Machine learning of generic reactions. 2. Toward an advanced computer representation of chemical reactions. *Tetrahedron Comput. Methodol.* **1990,** *3* (6A), 335.
(21) Tonnelier, C.; Jauffret, P.; Hanser, T.; Kaufmann, G. Machine learning of generic reactions. 3. An efficient algorithm for maximal common substructure determination. *Tetrahedron Comput. Methodol.* **1990,** *3* (6A), 351.
(22) 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.
(23) 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.
(24) Bort, W.; Baskin, I. I.; Gimadiev, T.; Mukanov, A.; Nugmanov, R.; Sidorov, P.; Marcou, G.; Horvath, D.; Klimchuk, O.; Madzhidov, T.; Varnek, A. Discovery of novel chemical reactions by deep generative recurrent neural network. *Sci. Rep.* **2021,** *11* (1), 3178.
(25) Varnek, A.; Fourches, D.; Horvath, D.; Klimchuk, O.; Gaudin, C.; Vayer, P.; Solov'ev, V.; Hoonakker, F.; Tetko, I. V.; Marcou, G. ISIDA - platform for virtual screening based on fragment and pharmacophoric descriptors. *Curr. Comput.-Aided Drug Des.* **2008,** *4* (3), 191-198.
(26) 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.
(27) ChemAxon software solutions. http://chemaxon.com/ (accessed June 19, 2021).

(28) Lin, A.; Dyubankova, N.; Madzhidov, T.; Nugmanov, R.; Rakhimbekova, A.; Ibragimova, Z.; Akhmetshin, T.; Gimadiev, T. R.; Suleymanov, R.; Verhoeven, J.; Wegner, J. K.; Ceulemans, H.; Varnek, A. Atom-to-Atom Mapping: A Benchmarking Study of Popular Mapping Algorithms and Consensus Strategies. http://chemrxiv.org/engage/chemrxiv/article-details/60c7505aee301c33b8c7a85e (accessed June 18, 2021).
(29) ChemAxon atom mapping. http://docs.chemaxon.com/display/docs/atom-mapping.md (accessed June 21, 2021).
(30) EPAM Indigo atom-to-atom mapping tool http://lifescience.opensource.epam.com/indigo/api/index.html#reaction-atom-to-atom-mapping (accessed June 21, 2021).
(31) Rahman, S. A.; Torrance, G.; Baldacci, L.; Cuesta, S. M.; Fenninger, F.; Gopal, N.; Choudhary, S.; May, J. W.; Holliday, G. L.; Steinbeck, C.; Thornton, J. M. Reaction Decoder Tool (RDT): extracting features from chemical reactions. *Bioinformatics* **2016,** *32* (13), 2065-2066.
(32) NameRxn software from NextMove Software. http://www.nextmovesoftware.com/namerxn.html (accessed June 25, 2021).
(33) Schwaller, P.; Hoover, B.; Reymond, J.-L.; Strobelt, H.; Laino, T. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. *Sci. Adv.* **2021,** *7* (15), eabe4166.
(34) Laboratoire de Cheminformatique, reation data cleaning. http://github.com/Laboratoire-de-Chemoinformatique/Reaction_Data_Cleaning (accessed June 21, 2021).
(35) Rakhimbekova, A.; Madzhidov, T. I.; Nugmanov, R. I.; Gimadiev, T. R.; Baskin, I. I.; Varnek, A. Comprehensive analysis of applicability domains of QSPR models for chemical reactions. *Int. J. Mol. Sci.* **2020,** *21* (15), 5542.
(36) Rakhimbekova, A.; Akhmetshin, T. N.; Minibaeva, G. I.; Nugmanov, R. I.; Gimadiev, T. R.; Madzhidov, T. I.; Baskin, I. I.; Varnek, A. Cross-validation strategies in QSPR modelling of chemical reactions. *SAR QSAR Environ. Res.* **2021,** *32* (3), 207-219.
(37) RDKit reaction fingerprints. http://www.rdkit.org/docs/cppapi/structRDKit_1_1ReactionFingerprintParams.html (accessed June 21, 2021).
(38) *Tables of Rate and Equilibrium Constants of Heterolytic Organic Reactions,* Palm, V., Ed.; VINITI: Moscow, Russia, 1977.
(39) Gimadiev, T. R.; Madzhidov, T. I.; Nugmanov, R. I.; Baskin, I. I.; Antipin, I. S.; Varnek, A. Assessment of tautomer distribution using the condensed reaction graph approach. *J. Comput.-Aided Mol. Des.* **2018,** *32* (3), 401-414.
(40) ChemAxon tautomer generation. http://docs.chemaxon.com/display/docs/tautomer-generation-plugin.md (accessed June 21, 2021).
(41) Kireeva, N.; Baskin, I. I.; Gaspar, H. A.; Horvath, D.; Marcou, G.; Varnek, A. Generative Topographic Mapping (GTM): Universal Tool for Data Visualization, Structure-Activity Modeling and Dataset Comparison. *Mol. Inf.* **2012,** *31* (3-4), 301-312.
(42) Glavatskikh, M.; Madzhidov, T.; Baskin, I. I.; Horvath, D.; Nugmanov, R.; Gimadiev, T.; Marcou, G.; Varnek, A. Visualization and Analysis of Complex Reaction Data: the Case of Tautomeric Equilibria. *Mol. Inf.* **2018,** *37*, 1800056.
(43) 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.
(44) Reaxys. http://www.elsevier.com/en-gb/solutions/reaxys (accessed June 21, 2021).
(45) *Greene's Protective Groups in Organic Synthesis,* 5th ed.; Wuts, P. G. M., Ed.; Wiley: New York, 2014.
(46) Sattarov, B.; Baskin, I. I.; Horvath, D.; Marcou, G.; Bjerrum, E. J.; Varnek, A. De Novo Molecular Design by Combining Deep Autoencoder Recurrent Neural Networks with Generative Topographic Mapping. *J. Chem. Inf. Model.* **2019,** *59* (3), 1182-1196.

(47) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R. The ChEMBL database in 2017. *Nucleic Acids Res.* **2016,** *45* (D1), D945-D954.
(48) Yang, C.; Tarkhov, A.; Marusczyk, J.; Bienfait, B.; Gasteiger, J.; Kleinoeder, T.; Magdziarz, T.; Sacher, O.; Schwab, C. H.; Schwoebel, J.; Terfloth, L.; Arvidson, K.; Richard, A.; Worth, A.; Rathman, J. New Publicly Available Chemical Query Language, CSRML, To Support Chemotype Representations for Application to Data Mining and Modeling. *J. Chem. Inf. Model.* **2015,** *55* (3), 510-528.
(49) ChemTunes ToxGPS. http://www.mn-am.com/products/chemtunestoxgps (accessed June 22, 2021).
(50) Ioannides, C.; Delaforge, M.; Parke, D. V. Safrole: its metabolism, carcinogenicity, and interactions with cytochrome P-450. *Food Cosmet. Toxicol.* **1981,** *19* (5), 657.
(51) Toxprint. http://toxprint.org (accessed June 28. 2021).
(52) MN-AM ToxPrint. http://www.mn-am.com/products/toxprint (accessed June 22, 2021).
(53) ChemoTyper. http://chemotyper.org (accessed June 28, 2021).
(54) MN-AM ChemoTyper. http://www.mn-am.com/products/chemotyper (accessed June 22, 2021).
(55) Chemotion Electronic Laboratory Notebook (ELN) & Repository for Research Data. http://www.chemotion.net/chemotionsaurus/index.html (accessed June 24, 2021).
(56) Turewicz, M.; Deutsch, E. W. Spectra, chromatograms, metadata: mzML-the standard data format for mass spectrometer output. *Methods Mol. Biol. (New York, NY)* **2011,** *696*, 179-203.
(57) Joint Committee on Atomic and Molecular Physical Data (JCAMP-DX) http://sourceforge.net/projects/jcamp-dx/ (accessed June 24, 2021).
(58) Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *J. Cheminf.* **2015,** *7*, 23.
(59) EnzymeML. http://enzymeml.org/ (accessed June 24, 2021).
(60) United Data Model (UDM) distribution. http://github.com/PistoiaAlliance/UDM (accessed June 30,2021).
(61) Chemotion repository. http://www.chemotion-repository.net/welcome (accessed June 24, 2021).
(62) PubChem. http://pubchem.ncbi.nlm.nih.gov/ (accessed June 27, 2021).
(63) Germany's National Research Data Infrastructure, NFDI. http://www.nfdi.de/en-gb/ (accessed June 24, 2021).
(64) NFDI4Chem, the chemistry consortium in Germany's National Research Data Infrastructure, NFDI. http://nfdi4chem.de/ (accessed June 24, 2021).
(65) Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; Bouwman, J.; Brookes, A. J.; Clark, T.; Crosas, M.; Dillo, I.; Dumon, O.; Edmunds, S.; Evelo, C. T.; Finkers, R.; Gonzalez-Beltran, A.; Gray, A. J. G.; Groth, P.; Goble, C.; Grethe, J. S.; Heringa, J.; t Hoen, P. A. C.; Hooft, R.; Kuhn, T.; Kok, R.; Kok, J.; Lusher, S. J.; Martone, M. E.; Mons, A.; Packer, A. L.; Persson, B.; Rocca-Serra, P.; Roos, M.; van Schaik, R.; Sansone, S.-A.; Schultes, E.; Sengstag, T.; Slater, T.; Strawn, G.; Swertz, M. A.; Thompson, M.; van der Lei, J.; van Mulligen, E.; Velterop, J.; Waagmeester, A.; Wittenburg, P.; Wolstencroft, K.; Zhao, J.; Mons, B. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016,** *3*, 160018.
(66) Logic and ontology. http://plato.stanford.edu/entries/logic-ontology/ (accessed April 5, 2020).
(67) Open Biological and Biomedical Ontology (OBO). http://www.obofoundry.org/ (accessed June 25, 2021).
(68) Web Ontology Language (OWL). http://www.w3.org/OWL/ (accessed June 25, 2021).
(69) Reaction ontology, RXNO. http://github.com/rsc-ontologies/rxno (accessed June 25, 2021).

(70) Andrews, D. M.; Broad, L. M.; Edwards, P. J.; Fox, D. N. A.; Gallagher, T.; Garland, S. L.; Kidd, R.; Sweeney, J. B. The creation and characterisation of a National Compound Collection: the Royal Society of Chemistry pilot. *Chem. Sci.* **2016,** *7* (6), 3869-3878.
(71) Lowe, D. M.; Corbett, P. T.; Murray-Rust, P.; Glen, R. C. Chemical name to structure: OPSIN, an open source solution. *J. Chem. Inf. Model.* **2011,** *51* (3), 739-753.
(72) *Compendium of Chemical Terminology. The Gold Book,* 2nd ed.; McNaught, A. D.; Wilkinson, A., Eds.; Blackwell Science: Oxford, U.K., 1997.
(73) Chemical entities of biological interest (ChEBI). http://www.ebi.ac.uk/chebi/ (accessed June 25, 2021).
(74) Gene Ontology (GO). http://geneontology.org/ (accessed June 25, 2021).
(75) RSC ontologies. http://github.com/rsc-ontologies (accessed June 25, 2021).
(76) 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.
(77) Grethe, G.; Blanke, G.; Kraut, H.; Goodman, J. M. International chemical identifier for reactions (RInChI). *J. Cheminf.* **2018,** *10*, 22.
(78) Paton, R. S.; Goodman, J. M.; Pellegrinet, S. C. Mechanistic Insights into the Catalytic Asymmetric Allylboration of Ketones: Bronsted or Lewis Acid Activation? *Org. Lett.* **2009,** *11* (1), 37-40.
(79) Lou, S.; Moquist, P. N.; Schaus, S. E. Asymmetric allylboration of ketones catalyzed by chiral diols. *J. Am. Chem. Soc.* **2006,** *128* (39), 12660-12661.
(80) Reid, J. P.; Simon, L.; Goodman, J. M. A Practical Guide for Predicting the Stereochemistry of Bifunctional Phosphoric Acid Catalyzed Reactions of Imines. *Acc. Chem. Res.* **2016,** *49* (5), 1029-1041.
(81) Reid, J. P.; Goodman, J. M. Goldilocks Catalysts: Computational Insights into the Role of the 3,3' Substituents on the Selectivity of BINOL-Derived Phosphoric Acid Catalysts. *J. Am. Chem. Soc.* **2016,** *138* (25), 7910-7917.
(82) Lee, S.; Goodman, J. M. Rapid Route-Finding for Bifurcating Organic Reactions. *J. Am. Chem. Soc.* **2020,** *142* (20), 9210-9219.
(83) Lee, S.; Goodman, J. M. VRAI-selectivity: calculation of selectivity beyond transition state theory. *Org. Biomol. Chem.* **2021,** *19* (17), 3940-3947.
(84) DP4-AI. http://github.com/Goodman-lab/DP4-AI (accessed June 26, 2021).
(85) 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.
(86) AMBIT. http://ambit.sourceforge.net/ (accessed June 26, 2021).
(87) Chemical Develpment Kit (CDK). http://cdk.github.io/ (accessed June 24, 2020).
(88) Jeliazkova, N.; Kochev, N. AMBIT-SMARTS: Efficient Searching of Chemical Structures and Fragments. *Molecular Informatics* **2011,** *30* (8), 707-720.
(89) Kochev, N.; Avramova, S.; Jeliazkova, N. Ambit-SMIRKS: a software module for reaction representation, reaction search and structure transformation. *J. Cheminf.* **2018,** *10*, 42.
(90) Ambit-SMIRKS. http://ambit.sourceforge.net/smirks.html (accessed June 27, 2021).
(91) Jeliazkova, N.; Jeliazkov, V. AMBIT RESTful web services: an implementation of the OpenTox Application Programming Interface. *J. Cheminf.* **2011,** *3*, 18.
(92) EnviPath. http://envipath.org/ (accessed June 27, 2021).
(93) BioTransformer. http://biotransformer.ca/ (accessed June 27, 2021).
(94) Sun, J.; Jeliazkova, N.; Chupakhin, V.; Golib-Dzib, J.-F.; Engkvist, O.; Carlsson, L.; Wegner, J.; Ceulemans, H.; Georgiev, I.; Jeliazkov, V.; Kochev, N.; Ashby, T. J.; Chen, H. ExCAPE-DB: an integrated large scale dataset facilitating Big Data analysis in chemogenomics. *J. Cheminf.* **2017,** *9*, 17.
(95) Kops, C. d. B.; Stork, C.; Sicho, M.; Kochev, N.; Svozil, D.; Jeliazkova, N.; Kirchmair, J. GLORY: generator of the structures of likely cytochrome P450 metabolites based on predicted sites of metabolism. *Front. Chem. (Lausanne, Switz.)* **2019,** *7*, 402.
(96) Ambitcli. http://ambit.sourceforge.net/ambitcli_standardisation.html (accessed June 26. 2021).

(97) HL7 Version 3 Standard: Structured Product Labeling, Release 7 (SPL R7). (HL7 SPL). http://www.hl7.org/implement/standards/product_brief.cfm?product_id=440 (accessed June 27, 2021).
(98) Health Level 7. http://en.wikipedia.org/wiki/Health_Level_7 (accessed June 28, 2021).
(99) Murray-Rust, P.; Rzepa, H. S. CML: Evolution and design. *J. Cheminf.* **2011,** *3* (1), 44.
(100) Warr, W. A. Scientific workflow systems: Pipeline Pilot and KNIME. *J. Comput.-Aided Mol. Des.* **2012,** *26* (7), 801-804.
(101) Next Move Software's HazELNut http://www.nextmovesoftware.com/hazelnut.html (accessed March 12, 2021).
(102) Pistoia Alliance. http://www.pistoiaalliance.org/ (accessed July 1, 2021).
(103) Pistachio. http://www.nextmovesoftware.com/pistachio.html (accessed July 1, 2021).
(104) Patent reaction extractor. http://github.com/dan2097/patent-reaction-extraction (accessed July 1, 2021).
(105) 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 July 1, 2021).
(106) NextMove Software LeadMine. http://www.nextmovesoftware.com/leadmine.html (accessed July 1, 2021).
(107) Jessop, D.; Adams, S.; Willighagen, E.; Hawizy, L.; Murray-Rust, P. OSCAR4: a flexible architecture for chemical text-mining. *J. Cheminf.* **2011,** *3* (1), 41.
(108) Nicolaou, C. A.; Watson, I. A.; Hu, H.; Wang, J. The Proximal Lilly Collection: Mapping, Exploring and Exploiting Feasible Chemical Space. *J. Chem. Inf. Model.* **2016,** *56* (7), 1253-1266.
(109) 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.
(110) 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.
(111) 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.
(112) 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.
(113) Thakkar, A.; Johansson, S.; Jorner, K.; Buttar, D.; Reymond, J.-L.; Engkvist, O. Artificial intelligence and automation in computer aided synthesis planning. *React. Chem. Eng.* **2021,** *6* (1), 27-51.
(114) Toniato, A.; Schwaller, P.; Cardinale, A.; Geluykens, J.; Laino, T. Unassisted noise reduction of chemical reaction datasets. *Nat. Mach. Intell.* **2021,** *3* (6), 485-494.
(115) Extraction of reactions from patents using grammars. http://ceur-ws.org/Vol-2696/paper_221.pdf (accessed July 2, 2021).
(116) Sketchy Sketches: Hiding Chemistry in Plain Sight. http://www.nextmovesoftware.com/talks/Lowe_SketchySketches_ACS_201608.pdf (accessed July 2, 2021).
(117) CAS REGISTRY. http://www.cas.org/cas-data/cas-registry (accessed July 3, 2021).
(118) Predicting New Chemistry: Impact of High-Quality Training Data on Prediction of Reaction Outcomes. http://www.cas.org/resources/whitepapers/predicting-new-chemistry (accessed July 3, 2021).
(119) CAS SciFinder Discovery Platform. http://www.cas.org/solutions/cas-scifinder-discovery-platform (accessed July 3, 2021).

(120) STN IP Protection Suite. http://www.cas.org/solutions/stn-ip-protection-suite (accessed July 3, 2021).
(121) CAS Custom Services. http://www.cas.org/solutions/cas-custom-services (accessed July 3, 2021).
(122) Open Reaction Database. http://docs.open-reaction-database.org/ (accessed July 3, 2021).
(123) Google protocol buffers. http://developers.google.com/protocol-buffers (accessed July 8, 2021).
(124) Open Reaction Database Interactive Editor. http://editor.open-reaction-database.org/datasets (accessed July 3, 2021).
(125) Open Reaction Database: GitHub. http://github.com/Open-Reaction-Database (accessed July 3, 2021).
(126) Hartenfeller, M.; Schneider, G. Enabling future drug discovery by de novo design. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2011,** *1* (5), 742-759.
(127) Fechner, U.; Schneider, G. Flux (1): A Virtual Synthesis Scheme for Fragment-Based de Novo Design. *J. Chem. Inf. Model.* **2006,** *46* (2), 699-707.
(128) 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.
(129) Enamine. http://enamine.net/ (accessed March 3, 2021).
(130) RDKit fingerprints. http://www.rdkit.org/docs/GettingStartedInPython.html#list-of-available-fingerprints (accessed July 4, 2021).
(131) ZINC database. http://zinc.docking.org/ (accessed July 5, 2021).
(132) Thakkar, A.; Chadimová, V.; Bjerrum, E. J.; Engkvist, O.; Reymond, J.-L. Retrosynthetic accessibility score (RAscore) – rapid machine learned synthesizability classification from AI driven retrosynthetic planning. *Chem. Sci.* **2021,** *12*, 3339-3349.
(133) *Modelling Molecular Structure and Reactivity in Biological Systems,* Naidoo, K.; Brady, J.; Field, M. J.; Gao, J.; Hann, M., Eds.; Royal Society of Chemistry: Cambridge, United Kingdom, 2006.
(134) Protein Data Bank (PDB). http://www.rcsb.org/pages/about-us/index (accessed July 5, 2021).
(135) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997,** *267* (3), 727-748.
(136) *Recommender Systems Handbook,* Ricci, F.; Rokach, L.; Shapira, B.; Kantor, P. B., Eds.; Springer: Boston, MA, 2011.
(137) Ghiandoni, G. M.; Bodkin, M. J.; Chen, B.; Hristozov, D.; Wallace, J. E. A.; Webster, J.; Gillet, V. J. Development and Application of a Data-Driven Reaction Classification Model: Comparison of an Electronic Lab Notebook and Medicinal Chemistry Literature. *J. Chem. Inf. Model.* **2019,** *59* (10), 4167-4187.
(138) Ghiandoni, G. M.; Bodkin, M. J.; Chen, B.; Hristozov, D.; Wallace, J. E. A.; Webster, J.; Gillet, V. J. Enhancing reaction-based de novo design using a multi-label reaction class recommender. *J. Comput.-Aided Mol. Des.* **2020,** *34* (7), 783-803.
(139) 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.
(140) ChemPass SynSpace http://chempassltd.com/synspace/ (accessed July 7, 2021).
(141) Makara, G. M.; Kovacs, L.; Szabo, I.; Pocze, G. Derivatization Design of Synthetically Accessible Space for Optimization: In Silico Synthesis vs Deep Generative Design. *ACS Med. Chem. Lett.* **2021,** *12* (2), 185-194.
(142) Zhavoronkov, A.; Ivanenkov, Y. A.; Aliper, A.; Veselov, M. S.; Aladinskiy, V. A.; Aladinskaya, A. V.; Terentiev, V. A.; Polykovskiy, D. A.; Kuznetsov, M. D.; Asadulaev, A.; Volkov, Y.; Zholus, A.; Shayakhmetov, R. R.; Zhebrak, A.; Minaeva, L. I.; Zagribelnyy, B. A.; Lee, L. H.; Soll, R.; Madge, D.; Xing, L.; Guo, T.; Aspuru-Guzik, A. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat. Biotechnol.* **2019,** *37* (9), 1038-1040.

(143) Perola, E. Minimizing false positives in kinase virtual screens. *Proteins: Struct., Funct., Bioinf.* **2006,** *64* (2), 422-435.
(144) Gu, S.; Smith, M. S.; Yang, Y.; Irwin, J. J.; Shoichet, B. K. Ligand Strain Energy in Large Library Docking. http://www.biorxiv.org/content/10.1101/2021.04.06.438722v1.full (accessed July 9, 2021).
(145) Maser, M. R.; Cui, A. Y.; Ryou, S.; DeLano, T. J.; Yue, Y.; Reisman, S. E. Multilabel Classification Models for the Prediction of Cross-Coupling Reaction Conditions. *J. Chem. Inf. Model.* **2021,** *61* (1), 156-166.
(146) 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.
(147) Lusci, A.; Pollastri, G.; Baldi, P. Deep Architectures and Deep Learning in Chemoinformatics: The Prediction of Aqueous Solubility for Drug-Like Molecules. *J. Chem. Inf. Model.* **2013,** *53* (7), 1563-1575.
(148) Chen, J. H.; Baldi, P. No electron left behind: A rule-based expert system to predict chemical reactions and reaction mechanisms. *J. Chem. Inf. Model.* **2009,** *49* (9), 2034-2043.
(149) Chen, J. H.; Baldi, P. Synthesis Explorer: A chemical reaction tutorial system for organic synthesis design and mechanism prediction. *J. Chem. Educ.* **2008,** *85* (12), 1699-1703.
(150) 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.
(151) 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.
(152) Fooshee, D.; Mood, A.; Gutman, E.; Tavakoli, M.; Urban, G.; Liu, F.; Huynh, N.; Van Vranken, D.; Baldi, P. Deep learning for chemical reaction prediction. *Mol. Syst. Des. Eng.* **2018,** *3* (3), 442-452.
(153) Chen, J.; Swamidass, S. J.; Dou, Y.; Bruand, J.; Baldi, P. ChemDB: a public database of small molecules and related chemoinformatics resources. *Bioinformatics* **2005,** *21* (22), 4133-4139.
(154) ChemDB. http://chemdb.ics.uci.edu/ (accessed July 9, 2021).
(155) Gao, W.; Coley, C. W. The Synthesizability of Molecules Proposed by Generative Models. *J. Chem. Inf. Model.* **2020,** *60* (12), 5714-5723.
(156) 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.
(157) Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F. SCScore: Synthetic Complexity Learned from a Reaction Corpus. *J. Chem. Inf. Model.* **2018,** *58* (2), 252-261.
(158) Spaya. http://spaya.ai/ (accessed July 13, 2021).
(159) Brown, N.; Fiscato, M.; Segler, M. H. S.; Vaucher, A. C. GuacaMol: Benchmarking Models for de Novo Molecular Design. *J. Chem. Inf. Model.* **2019,** *59* (3), 1096-1108.
(160) Bickerton, G. R.; Paolini, G. V.; Besnard, J.; Muresan, S.; Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nat. Chem.* **2012,** *4* (2), 90-98.
(161) Renz, P.; Van Rompaey, D.; Wegner, J. K.; Hochreiter, S.; Klambauer, G. On failure modes in molecule generation and optimization. *Drug Discovery Today: Technologies* **2019,** *32-33*, 55-63.
(162) Coley, C. W.; Eyke, N. S.; Jensen, K. F. Autonomous Discovery in the Chemical Sciences Part I: Progress. *Angew. Chem., Int. Ed.* **2020,** *59* (51), 22858-22893.
(163) Coley, C. W.; Eyke, N. S.; Jensen, K. F. Autonomous Discovery in the Chemical Sciences Part II: Outlook. *Angew. Chem., Int. Ed.* **2020,** *59* (52), 23414-23436.
(164) 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.
(165) Segler, M. H. S.; Waller, M. P. Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction. *Chem. - Eur. J.* **2017,** *23* (25), 5966-5971.
(166) Fortunato, M. E.; Coley, C. W.; Barnes, B. C.; Jensen, K. F. Data Augmentation and Pretraining for Template-Based Retrosynthetic Prediction in Computer-Aided Synthesis Planning. *J. Chem. Inf. Model.* **2020,** *60* (7), 3398-3407.

(167) Coley, C. W.; Green, W. H.; Jensen, K. F. Machine Learning in Computer-Aided Synthesis Planning. *Acc. Chem. Res.* **2018,** *51* (5), 1281-1289.
(168) 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.
(169) Godfrey, A. G.; Masquelin, T.; Hemmerle, H. A remote-controlled adaptive medchem lab: an innovative approach to enable drug discovery in the 21st Century. *Drug Discovery Today* **2013,** *18* (17–18), 795-802.
(170) Baranczak, A.; Tu, N. P.; Marjanovic, J.; Searle, P. A.; Vasudevan, A.; Djuric, S. W. Integrated Platform for Expedited Synthesis-Purification-Testing of Small Molecule Libraries. *ACS Med. Chem. Lett.* **2017,** *8* (4), 461-465.
(171) Collins, N.; Stout, D.; Lim, J.-P.; Malerich, J. P.; White, J. D.; Madrid, P. B.; Latendresse, M.; Krieger, D.; Szeto, J.; Vu, V.-A.; Rucker, K.; Deleo, M.; Gorfu, Y.; Krummenacker, M.; Hokama, L. A.; Karp, P.; Mallya, S. Fully Automated Chemical Synthesis: Toward the Universal Synthesizer. *Org. Process Res. Dev.* **2020,** *24* (10), 2064-2077.
(172) 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.
(173) Coley, C. W.; Thomas, D. A.; Lummiss, J. A. M.; 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. F. A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science* **2019,** *365* (6453), eaax1566.
(174) Szymkuc, 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.
(175) Mikulak-Klucznik, B.; Golebiowska, P.; Bayly, A. A.; Popik, O.; Klucznik, T.; Szymkuc, S.; Gajewska, E. P.; Dittwald, P.; Staszewska-Krajewska, O.; Beker, W.; Badowski, T.; Scheidt, K. A.; Molga, K.; Mlynarski, J.; Mrksich, M.; Grzybowski, B. A. Computational planning of the synthesis of complex natural products. *Nature* **2020,** *588* (7836), 83-88.
(176) 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 ICSYNTH Software Tool as an Idea Generator for Synthesis Prediction. *Org. Process Res. Dev.* **2015,** *19* (2), 357-368.
(177) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **2018,** *555* (7698), 604-610.
(178) Genheden, S.; Thakkar, A.; Chadimová, V.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E. AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning. *J. Cheminf.* **2020,** *12* (1), 70.
(179) Dai, H.; Li, C.; Coley, C. W.; Dai, B.; Song, L. Retrosynthesis Prediction with Conditional Graph Logic Network. arXiv.org e-Print archive. http://arxiv.org/pdf/2001.01408.pdf (accessed March 19, 2020).
(180) 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).
(181) PostEra. http://postera.ai/ (accessed July 14, 2021).
(182) Lin, K.; Xu, Y.; Pei, J.; Lai, L. Automatic retrosynthetic route planning using template-free models. *Chem. Sci.* **2020,** *11* (12), 3355-3364.
(183) Schwaller, P.; Petraglia, R.; Zullo, V.; Nair, V. H.; Haeuselmann, R. A.; Pisoni, R.; Bekas, C.; Iuliano, A.; Laino, T. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chem. Sci.* **2020,** *11* (12), 3316-3325.
(184) Shi, C.; Xu, M.; Guo, H.; Zhang, M.; Tang, J. A Graph to Graphs Framework for Retrosynthesis Prediction. arXiv e-Print archive. http://arxiv.org/pdf/2003.12725.pdf (accessed February 2. 2021).

(185) Somnath, V. R.; Bunne, C.; Coley, C. W.; Krause, A.; Barzilay, R. Learning Graph Models for Template-Free Retrosynthesis. http://arxiv.org/pdf/2006.07038.pdf (accessed February 2, 2021).
(186) Sacha, M.; Błaż, M.; Byrski, P.; Włodarczyk-Pruszyński, P.; Jastrzębski, S. Molecule Edit Graph Attention Network: Modeling Chemical Reactions as Sequences of Graph Edits. http://arxiv.org/pdf/2006.15426.pdf (accessed July 14, 2021).
(187) Ryou, S.; Maser, M. R.; Cui, A. Y.; DeLano, T. J.; Yue, Y.; Reisman, S. E. Graph Neural Networks for the Prediction of Substrate-Specific Organic Reaction Conditions. arXiv e-Print archive. http://arxiv.org/pdf/2007.04275.pdf (accessed February 4, 2021).
(188) 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.
(189) Li, J.; Eastgate, M. D. Making better decisions during synthetic route design: leveraging prediction to achieve greenness-by-design. *React. Chem. Eng.* **2019,** *4* (9), 1595-1607.
(190) Fitzner, M.; Wuitschik, G.; Koller, R. J.; Adam, J.-M.; Schindler, T.; Reymond, J.-L. What can reaction databases teach us about Buchwald-Hartwig cross-couplings? *Chem. Sci.* **2020,** *11* (48), 13085-13093.
(191) Jin, W.; Coley, C. W.; Barzilay, R.; Jaakkola, T. S. Predicting Organic Reaction Outcomes with Weisfeiler-Lehman Network. http://arxiv.org/pdf/1709.04555.pdf (accessed July 14. 2021).
(192) Schwaller, P.; Gaudin, T.; Lanyi, D.; Bekas, C.; Laino, T. "Found in Translation": predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chem. Sci.* **2018,** *9* (28), 6091-6098.
(193) 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).
(194) 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.
(195) Beker, W.; Gajewska, E. P.; Badowski, T.; Grzybowski, B. A. Prediction of Major Regio-, Site-, and Diastereoisomers in Diels-Alder Reactions by Using Machine-Learning: The Importance of Physically Meaningful Descriptors. *Angew. Chem., Int. Ed.* **2019,** *58* (14), 4515-4519.
(196) Struble, T. J.; Coley, C. W.; Jensen, K. F. Multitask prediction of site selectivity in aromatic C-H functionalization reactions. *React. Chem. Eng.* **2020,** *5* (5), 896-902.
(197) Pesciullesi, G.; Schwaller, P.; Laino, T.; Reymond, J.-L. Transfer learning enables the molecular transformer to predict regio- and stereoselective reactions on carbohydrates. *Nat. Commun.* **2020,** *11* (1), 4874.
(198) Guan, Y.; Coley, C. W.; Wu, H.; Ranasinghe, D.; Heid, E.; Struble, T. J.; Pattanaik, L.; Green, W. H.; Jensen, K. F. Regio-selectivity prediction with a machine-learned reaction representation and on-the-fly quantum mechanical descriptors. *Chem. Sci.* **2021,** *12* (6), 2198-2208.
(199) 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.
(200) 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.
(201) Reid, J. P.; Sigman, M. S. Holistic prediction of enantioselectivity in asymmetric catalysis. *Nature* **2019,** *571* (7765), 343-348.
(202) Zahrt, A. F.; Henle, J. J.; Rose, B. T.; Wang, Y.; Darrow, W. T.; Denmark, S. E. Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning. *Science* **2019,** *363* (6424), eaau5631.
(203) Vaswami, A.; Shazeer, N.; Parmar, N.; Uszkorelt, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention is all you need. http://arxiv.org/pdf/1706.03762.pdf (accessed July 24, 2019).
(204) IBM RXN for Chemistry. http://rxn.res.ibm.com/ (accessed July 12, 2021).

(205) IBM RoboRXN for Chemistry http://rxn.res.ibm.com/rxn/robo-rxn/welcome (accessed July 11, 2021).
(206) Vaucher, A. C.; Zipoli, F.; Geluykens, J.; Nair, V. H.; Schwaller, P.; Laino, T. Automated extraction of chemical synthesis actions from experimental procedures. *Nat. Commun.* **2020,** *11*, 3601.
(207) Vaucher, A. C.; Schwaller, P.; Geluykens, J.; Nair, V. H.; Iuliano, A.; Laino, T. Inferring experimental procedures from text-based representations of chemical reactions. *Nat. Commun.* **2021,** *12* (1), 2573.
(208) Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; Zettlemoyer, L. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. http://arxiv.org/pdf/1910.13461.pdf (accessed July 11, 2021).
(209) Schwaller, P.; Probst, D.; Vaucher, A. C.; Nair, V. H.; Kreutter, D.; Laino, T.; Reymond, J.-L. Mapping the space of chemical reactions using attention-based neural networks. *Nat. Mach. Intell.* **2021,** *3*, 144–152.
(210) Schwaller, P.; Vaucher, A. C.; Laino, T.; Reymond, J.-L. Prediction of chemical reaction yields using deep learning. *Mach. Learn.: Sci. Technol.* **2021,** *2*, 015016.
(211) Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers);* ACL, Ed.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp 4171–4186.
(212) Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. http://arxiv.org/pdf/1909.11942v1.pdf (accessed July 12, 2021).
(213) Thakkar, A.; Kogej, T.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E. J. Datasets and their influence on the development of computer assisted synthesis planning tools in the pharmaceutical domain. *Chem. Sci.* **2020,** *11* (1), 154-168.
(214) RXNMapper. http://rxnmapper.ai (accessed July 12, 2021).
(215) RXNMapper source code. http://github.com/rxn4chemistry/rxnmapper (accessed July 12, 2021).
(216) DigitalGlassware. http://www.deepmatter.io/products/digitalglassware/ (accessed July 12, 2021).
(217) Kitson, P. J.; Marie, G.; Francoia, J.-P.; Zalesskiy, S. S.; Sigerson, R. C.; Mathieson, J. S.; Cronin, L. Digitization of multistep organic synthesis in reactionware for on-demand pharmaceuticals. *Science* **2018,** *359* (6373), 314-319.
(218) Mehr, S. H. M.; Craven, M.; Leonov, A. I.; Keenan, G.; Cronin, L. A universal system for digitization and automatic execution of the chemical synthesis literature. *Science* **2020,** *370* (6512), 101-108.
(219) Wilbraham, L.; Mehr, S. H. M.; Cronin, L. Digitizing Chemistry Using the Chemical Processing Unit: From Synthesis to Discovery. *Acc. Chem. Res.* **2021,** *54* (2), 253-262.
(220) Chemify. http://www.chem.gla.ac.uk/cronin/chemify/ (accessed March 23, 2020).
(221) XDL standard. http://croningroup.gitlab.io/chemputer/xdl/standard/index.html (accessed July 27, 2021).
(222) SiLA 2. http://sila-standard.com/standards/ (accessed July 15, 2021).
(223) Bostroem, J.; Brown, D. G.; Young, R. J.; Keseru, G. M. Expanding the medicinal chemistry synthetic toolbox. *Nat. Rev. Drug Discovery* **2018,** *17* (10), 709-727.
(224) Mahjour, B.; Shen, Y.; Liu, W.; Cernak, T. A map of the amine-carboxylic acid coupling system. *Nature* **2020,** *580* (7801), 71-75.

(225) Wishart, D. S.; Knox, C.; Guo, A. C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* **2008,** *36* (Database Issue), D901-D906.
(226) Wong, H.; Cernak, T. Reaction miniaturization in eco-friendly solvents. *Curr. Opin. Green Sustainable Chem.* **2018,** *11*, 91-98.
(227) Mahjour, B.; Shen, Y.; Cernak, T. Ultrahigh-Throughput Experimentation for Information-Rich Chemical Synthesis. *Acc. Chem. Res.* **2021,** *54* (10), 2337-2346.
(228) Mahjour, B.; Cernak, T. phactor - a high throughput experimentation management system. http://chemrxiv.org/engage/chemrxiv/article-details/60c75166702a9bcd6918bf39 (accessed August 3, 2021).
(229) Buitrago Santanilla, A.; Regalado, E. L.; Pereira, T.; Shevlin, M.; Bateman, K.; Campeau, L.-C.; Schneeweis, J.; Berritt, S.; Shi, Z.-C.; Nantermet, P.; Liu, Y.; Helmy, R.; Welch, C. J.; Vachal, P.; Davies, I. W.; Cernak, T.; Dreher, S. D. Nanomole-scale high-throughput chemistry for the synthesis of complex molecules. *Science* **2015,** *347* (6217), 49-53.
(230) Gesmundo, N. J.; Sauvagnat, B.; Curran, P. J.; Richards, M. P.; Andrews, C. L.; Dandliker, P. J.; Cernak, T. Nanoscale synthesis and affinity ranking. *Nature* **2018,** *557* (7704), 228-232.
(231) Lin, S.; Dikler, S.; Blincoe, W. D.; Ferguson, R. D.; Sheridan, R. P.; Peng, Z.; Conway, D. V.; Zawatzky, K.; Wang, H.; Cernak, T.; Davies, I. W.; DiRocco, D. A.; Sheng, H.; Welch, C. J.; Dreher, S. D. Mapping the dark space of chemical reactions with extended nanomole synthesis and MALDI-TOF MS. *Science* **2018,** *361* (6402), eaar6236.
(232) Martin, A.; Nicolaou, C. A.; Toledo, M. A. Navigating the DNA encoded libraries chemical space. *Commun. Chem.* **2020,** *3* (1), 127.
(233) Masquelin, T.; Kaerner, A.; Bernhardt, R. J.; Wang, J.; Nicolaou, C. A. Automated Synthesis. In *Burger's Medicinal Chemistry, Drug Discovery and Development*; John Wiley & Sons: Hoboken, NJ, 2021; pp 1-37.
(234) Nicolaou, C. A.; Watson, I. A.; LeMasters, M.; Masquelin, T.; Wang, J. Context Aware Data-Driven Retrosynthetic Analysis. *J. Chem. Inf. Model.* **2020,** *60* (6), 2728-2738.
(235) Struble, T. J.; Alvarez, J. C.; Brown, S. P.; Chytil, M.; Cisar, J.; DesJarlais, R. L.; Engkvist, O.; Frank, S. A.; Greve, D. R.; Griffin, D. J.; Hou, X.; Johannes, J. W.; Kreatsoulas, C.; Lahue, B.; Mathea, M.; Mogk, G.; Nicolaou, C. A.; Palmer, A. D.; Price, D. J.; Robinson, R. I.; Salentin, S.; Xing, L.; Jaakkola, T.; Green, W. H.; Barzilay, R.; Coley, C. W.; Jensen, K. F. Current and Future Roles of Artificial Intelligence in Medicinal Chemistry Synthesis. *J. Med. Chem.* **2020,** *63* (16), 8667-8682.
(236) Sikka, K.; Silberfarb, A.; Byrnes, J.; Sur, I.; Chow, E.; Divakaran, A.; Rohwer, R. Deep Adaptive Semantic Logic (DASL). http://www.sri.com/wp-content/uploads/2020/07/2003.07344.pdf (accessed July 19, 2021).
(237) 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.
(238) Warr, W.; Kanza, S.; Frey, J. G.; Whitby, R. J. AI3SD, Dial-a-Molecule & Directed Assembly: AI for Reaction Outcome and Synthetic Route Prediction Conference Report 2020. http://eprints.soton.ac.uk/441628/ (accessed July 19, 2021).
(239) eMolecules. http://www.emolecules.com/ (accessed July 19, 2021).
(240) RDKit: open-source cheminformatics. http://www.rdkit.org (accessed March 30, 2020).
(241) A Specialized Platform for Innovative Research Exploration (ASPIRE). http://ncats.nih.gov/aspire (accessed July 19., 2021).
(242) http://neo4j.com/ (accessed July 19. 2021).
(243) PostgreSQL. http://www.postgresql.org/ (accessed July 19, 2021).
(244) BurstIQ. http://www.burstiq.com/ (accessed July 19, 2021).
(245) Rarey, M.; Nicklaus, M. C.; Warr, W. Call for Papers for the Special Issue: From Reaction Informatics to Chemical Space. *J. Chem. Inf. Model.* **2021,** *61* (4), 1531-1532.

(246) NIH Virtual Workshop on Reaction Informatics, May 18-20, 2021. http://cactus.nci.nih.gov/presentations/NIHReactInf_2021-05/NIHReactInf.html (accessed August 2, 2021).