PREDICTING NEW CHEMISTRY

Impact of high-quality training data on prediction of reaction outcomes

Miriam Wollenhaupt, Ph.D., computational chemist, Bayer AG
Martín Villalba, Ph.D., expert applied mathematics, Bayer AG
Orr Ravitz, Ph.D., synthesis planning solutions, CAS

BAYER
CAS
A division of the
American Chemical Society

In chemical synthesis planning applications, the goal is to generate sets of synthetic routes that are as diverse and accurate as possible to provide organic chemists with many plausible and distinct strategies to make their target molecules. However, data-driven computational applications can only be as good as the underpinning data. The quality of predicted results depends on the following main properties of the training data:

1. The **diversity** of the predictions is correlated to the breadth of the data source: how many reaction types are represented and how diverse the products and substrates are in each reaction.
2. The **accuracy** of the predictions depends on the quality and consistency of the data and its representation as well as its depth: the number of examples available for each reaction type and the spectrum of reactants, products, and reaction conditions are available.

In this study, we demonstrate the significant impact that even a moderately sized set of scientist-curated reactions from the CAS Content Collection™ can have on the predictive power of a synthesis planning tool.

A broad training set was enriched with examples targeting certain reaction types, which dramatically enhanced the predictive power of the machine learning models. This is a strong indication for the much greater potential for CAS content to drive AI applications in synthesis planning.

> **Enriching a training set with high quality, diverse CAS reactions had a significant impact on predictive power.**

Introduction

Computer-aided synthesis design (CASD) has been conceived more than a half a century ago, reflecting a convergence of two major paradigms—computer science and organic chemistry. Both fields have experienced dramatic advances during the preceding decade. In computer science, it has been the introduction of digital hardware, and the application and further growth of algorithmic approaches to solve complex mathematical problems. In organic chemistry, the synthesis of several natural products, such as vitamin A,[1] prostaglandin,[2] and vitamin B12,[3] contributes to the development of new methods, but more importantly led to a transition of organic synthesis from an art to a more scientific and methodic discipline. It was EJ Corey who first described a pseudo-algorithmic approach for synthesis planning in the form of retrosynthetic analysis,[4] and not surprisingly, he followed his theoretical framework with a first-of-its-kind implementation of the concept as a computer program, Organic Chemical Simulation of Synthesis (OCSS)[5], that evolved into Logic and Heuristics Applied to Synthesis Analysis (LHASA)[6] soon after.

LHASA and other systems from the foundational two decades of the field were expert systems, relying on manually coded knowledge bases to power the predictive capabilities.[7,8] Chemical reactions, reaction rules, synthetic strategies, and knowledge bases designed to address aspects like chemical interference, stability, and feasibility of structures, had to be coded by experts using technology that was specifically developed for the ever-growing complexity and diversity of chemistry. Not surprisingly, these approaches could not keep up with the progress in organic chemistry. The advent of chemical and reaction databases, as well as data mining tools during the 1980s and 1990s, caused CASD to fall out of favor.[9] However, computational chemistry and cheminformatics methods continued to advance, and along with tremendous improvements in computer hardware and steady growth in chemical databases, they paved the way for the re-emergence of the field in the past decade. The availability of data with the sophistication of processing methods contributed to a transition from manual approaches of capturing knowledge to algorithmic approaches that can capture the breadth of chemistry and can remain up-to-date.[10-15]

As with any other data-driven methodology, to reliably automate extraction of chemical knowledge from databases for applications in CASD, comprehensive, high-quality data sources are required.[16] To achieve good coverage of synthetic methods, the data source must cover a diverse set of reactions from a broad set of publications and a comprehensive set of examples from each class of reactions to capture the scope of the different methods. Quality of the source can be measured with various parameters, such as the accuracy of capturing the structures and the experimental details, the accuracy of atom mapping, which is essential for template- and rule-based approaches, and normalization of the data, including uniform representation of functional groups and handling of tautomerism and stereochemistry. This paper demonstrates the tight dependency between the quality and comprehensiveness of the training data and the predictive power of machine learning techniques. It lays out a case where machine learning models trained on a sizable reaction set provide poor predictive power for several classes of reactions, likely due to sparsely populated areas in the corresponding chemical space. By filling the space with a targeted reaction data set from CAS, the resulting models show a significant increase in predictive accuracy in the targeted chemical space without loss of accuracy overall.

The majority of current CASD systems use either a rule- or template-based approach. These approaches lend themselves most easily to linking between predictions and evidence from the training, making the results easy to interpret and rationalize. Template-free approaches that typically use a string-based description of chemical reactions and leverage techniques from language processing applications in AI for CASD have also gained attention in recent years and show some promise. Although this paper discusses the data dependence of structure-based approaches, many of the lessons would apply to template-free methods as well.

Methods

In this study, we evaluated results of two of the three neural networks described by Segler et al.,[11] specifically the policy expansion network and the in-scope network. The policy expansion neural network receives a target molecule as input and selects the most promising templates that can be used as transforms. Applying the templates in the retrosynthetic direction generates sets of precursors, or the educts that are necessary for synthesizing the target. The in-scope neural network checks if each reaction is feasible. In this system it is called the viability filter, as its goal is to filter out transformations that are unlikely to work. These two networks are therefore vital in generating the qualified components of the retrosynthesis pipeline. The mechanism by which synthetic plans are created is beyond the scope of this paper, but the synthetic accuracy and diversity is by and large determined by the above two networks. The results below focus on the performance of the in-scope filter.

Data

Our base training set includes 8 millon reactions from a commercially available database. These reactions are a subset of 17.5 million reactions, restricted to machine-readable reactions with at most two reactants and one product. A template extraction process was applied that identified about 10K templates with at least 25 examples. Frequency of the templates varied strongly in the reactions set and the number of reactions available for training the viability filter for each template varied accordingly. Additionally, the number of available example reactions varied significantly from class to class. The Pareto plot in Figure 1 shows statistics on the size and distribution of all available transformation classes.

Additionally, 24 million implied negative reactions were created using the reaction set and the extracted templates in a procedure akin to the one described in Segler et al. These reactions, together with the 8 million positive reactions, made up the V1-base training set.

Additional data sets were used to quantify the predictive capability of the viability filter. A second training data set (V2-cas) was created by expanding V1-base with 14.5K positive CAS reactions. These reactions were custom curated to provide additional examples to templates with relatively few reactions associated. The curated CAS reactions for specific templates comprise only 0.05% of the total reactions in the V2-cas training set.

A third training data set (V3-neg) was created by expanding V1-base with 135K negative reactions extracted from Bayer's electronic lab notebooks (ELNs). These negative reactions were obtained from real-life negative experiments and were randomly sampled to be a representative subset of Bayer's ELN library. This data set was chosen to measure the effectiveness of both filters when dealing with unseen chemistry, as neither filter has seen real negative reactions before.

We trained and tested the models using the data set combinations summarized in Table 1. The data sets were all split to training, validation, and testing sets - V1-base is in a 70/10/20 split, while both V2-cas and V3-neg data sets are in a more traditional 80/10/10 split.

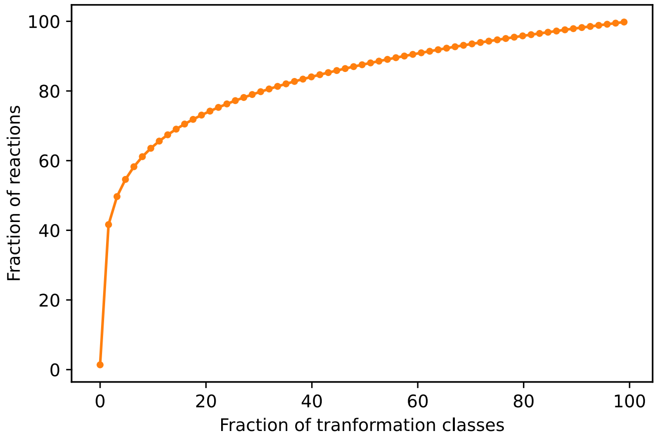

Figure 1. Size and fraction of reaction classes present in our dataset

Training set	Make-up
V1-base	Commercially available positive data and synthetic negative data (32 million)
V2-cas	V1-base reactions over 32 million curated CAS reactions for specific templates (14.5K)
V3-neg	V1-base reactions (32 million) plus negative reactions from Bayer's ELN library (135K)

Table 1: Data sets used in this study as different combinations of available data sources

CAS reaction collection

The vast and diverse CAS reaction collection is updated daily by scientists from a wide variety of patents, journals, and other reference works from 1840 to the present. Reactions link directly to the CAS REGISTRY® substance database to identify reaction participants. Structure representations for reactions match directly to the structure representations in CAS REGISTRY. Accurate data collection reflects the specific policies used to define the collection and recording of reaction data. For example, careful delineation of which reaction participants contribute carbons to reaction products and which do not, make the data particularly useful for further analysis, as in the case of the work described in this paper. Analysts also carefully differentiate between catalysts, which are present in sub stoichiometric amounts or are common reaction catalysts, and reagents used in stoichiometric amounts. Reaction indexing includes data for yields, times, temperatures, pressures, and pH when authors provide this information. Added tags, such as stereoselective, highlight key aspects of reactions. Mapping of atoms from reactants to products serves as a quality check for the indexing of molecules participating in the reactions. CAS audits reactions to additionally check accuracy and ensure high quality.

A small subset of the CAS reaction collection was utilized for this project. The team at Bayer supplied example reactions for a range of chemical transformations to the CAS Custom Services^{SM} team and asked for related reactions from the CAS Content Collection. Analysts utilized SMARTS strings for the transformations of interest to search, combine, and deduplicate results. CAS information scientists with synthetic organic chemistry backgrounds performed a final quality assurance step. As a result, 14.5K reactions were identified that were highly underrepresented in the initial V1-base data set and were used to bolster reaction examples across the targeted six reaction types.

Model training

All models share the same architecture, a fully Connected Network with ReLu activation and Sigmoid output. The input to the network is the concatenation of two ECFP fingerprints (radius=2, features=8192), giving a final input dimension of 2 x 8192 = 16384. Each ECFP fingerprint is the embedding of a product and a reactant. The single hidden layer's dimension is 32, and its kernel is regularized with both L1 and L2 penalties (1e-5 and 1e-4 respectively). The output is a single sigmoid neuron, where a reaction is only considered likely to succeed if its value is >= 0.6. The model was trained using Adam as optimizer (learning rate = 1e-4) and it was trained until convergence. A successful prediction has been defined as the ability to predict that the fingerprints of a given product and reactant lead to a successful reaction.

Results and discussion

To quantify the predictive capability of the viability filter with different data sources, supplemental data added to the V1-base training set were analyzed independently. Accuracy was measured as the percentage of cases where the neural network would correctly determine whether a reaction would be successful or not.

The reaction classes that initially had limited examples available were specifically tested, as these may indicate areas of chemistry that are not as frequently explored. Predictions in these rare classes may reveal new chemistry and increase the number of plausible strategies uncovered to make target molecules. Training on the V1-base data set and testing against the selected reaction classes shows a poor predictive power of 16% (Table 2). By adding CAS reactions, the accuracy in rare reaction classes increases to 48% a boost of 32 percentage points.

There is no statistically significant loss of accuracy when testing over the negative data and the strong regularization in our current architecture precludes the presence of an overfitting effect.

Table 3 shows that the smaller datasets that supplemented the V1-base training set have virtually no effect on the overall predictive accuracy of the model in the context of the full set. This is expected as the supplemental data sets are an order of magnitude smaller than the V1-base set. Additionally, the majority of reaction classes in the test set are unrelated to the classes represented in the CAS reaction set.

The increase in prediction accuracy of rare reaction classes seen with CAS reaction data has no statistically significant impact on the predictive accuracy across all reaction classes as observed when testing with the full set.

Training set Test set	V1-base	V2-cas
V1-base	97%	97%
V2-cas	97%	97%
V3-neg	94%	94%

Table 3: Accuracy for entire datasets

Conclusion

A well-documented challenge in our fast-filter architecture is that the model is not great at generalizing and has trouble identifying chemistry that it has not seen in training. This limitation is inherent to template-based approaches, especially when the templates are not designed to capture the essence of the reactions and ignore the structural and functional context. Rule-based and template-free approaches offer better generalization, but they, as well, depend on the breadth and accuracy of their training sets. By measuring how different types of data affect the training, we can make stronger assertions regarding the novelty of the chemistry included in each type of data set.

This study shows that supplementing a large data source with comprehensive sets of reactions in specific domains may enhance the predictive power in those domains without impacting the overall predictive performance. By utilizing data not typically found in data sources for these domains, understanding into useful, new chemistry is expanded. This enhanced predictive power in 'rare' categories may open previously difficult areas of science.

These findings illustrate that the performance of neural networks depends on the amount, quality, and diversity of the training set. In this case, the effect was seen on a small class of reactions. However, it can be expected that predictive gains would increase relative to the number of templates with underrepresented reaction examples that are reinforced. More so, should the base training data begin with strong amount, quality, and diversity across all templates, the gains in predictive power would be even greater.

> **Enhanced predictive power in 'rare' categories may open previously difficult areas of science.**

References

1. Über die Ester und Äther des synthetischen Vitamins A. **Isler, O., Ronco, A., Guex, W., Hindley, N.C., Huber, W., Dialer, K. and Kofler, M.** 2, 1949, Helv. Chim. Acta, Vol. 32, pp. 489-505.
2. Stereo-controlled synthesis of dl-prostaglandins F2α and E2. Corey, E.J., Schaaf T.K., Huber, W., Koelliker, U. and Weinshenker, N.M. 2, 1969, J. Am. Chem. Soc., Vol. 91, pp. 5675-5677.
3. Recent advances in the chemistry of natural products. **Woodward, R.B.** 3-4, 1968, Pure Appl. Chem., Vol. 17, pp. 519-547.
4. General methods for the construction of complex molecules. **Corey, E.J.** 1, 1967, Pure Appl. Chem., Vol. 14, pp. 19-38.
5. Computer-assisted design of complex organic syntheses. **Corey, E.J. and Wipke, W.T.** 3902, 1969, Science, Vol. 166, pp. 178-192.
6. Computer-assisted analysis in organic synthesis. **Corey, E.J., Long, A.K. and Rubenstein, S.D.** 4698, 1985, Science, Vol. 228, pp. 408-418.
7. Simulation and evaluation of chemical synthesis—SECS: An application of artificial intelligence techniques. **Wipke, W.T., Ouchi, G.I. and Krishnan,** S. 1-2, 1978, Artificial Intelligence, Vol. 11, pp. 173-193.
8. Empirical Explorations of SYNCHEM. **Gelernter, H.L., Sanders, A.F., Larsen, D.L., Agarwal, K.K., Boivie, R.H., Spritzer, G.A. and Searleman, E.J.** 4308, 1977, Science, Vol. 197, pp. 1041-1049.
9. Computer-assisted planning of organic syntheses: the second generation of programs. **Ihlenfeldt, W. and Gasteiger**, J. 23-24, 1996, Angew. Chem., Int. Ed. Engl., Vol. 34, pp. 2613-2633.
10. Route Designer: A Retrosynthetic Analysis Tool Utilizing Automated Retrosynthetic Rule Generation. **Law, J., Zsoldos, Z., Simon, A., Reid, D., Liu, Y., Khew, S.Y., Johnson, A.P., Major, S., Wade, R.A. and Ando, H.Y.** 3, 2009, J. Chem. Inf. Model., Vol. 49, pp. 593-602.
11. Planning chemical syntheses with deep neural networks and symbolic AI. **Segler, M., Preuss, M. and Waller, M.** 2018, Nature, Vol. 555, pp. 604-610.
12. Predicting retrosynthetic pathways using a combined linguistic model and hyper-graph exploration strategy. **Schwaller, P., Petraglia, R., Zullo, V., Nair, V.H., Haeuselmann, R.A., Pisoni, R., Bekas, C., Iuliano, A. and Laino, T.** Chem. Sci., 2020, Vol. 11, pp. 3316-3325
13. Context Aware Data-Driven Retrosynthetic Analysis. **Nicolaou, C.A., Watson, I.A., LeMasters, M., Masquelin, T. and Wang, J.** 6, 2020, J. Chem. Inf. Model., Vol. 60, pp. 2728–2738.
14. AiZynthFinder: A Fast Robust and Flexible Open-Source Software for Retrosynthetic Planning. **Genheden, S., Thakkar, A., Chadimova, V., Reymond, J.L., Engkvist, O. and Bjerrum, E.J.** 70, 2020, J. Chemin., Vol. 12.
15. Robotic platform for flow synthesis of organic compounds informed by AI planning. **Coley, C.W., Thomas, D.A., III, Lummiss, J.M., Jaworski, J.N., Breen, C.P., Schultz, V., Hart, T., Fishman, J.S., Rogers, L. and Gao, H. A.** 6453, 2019, Science, Vol. 365.
16. Datasets and their influence on the development of computer assisted synthesis planning tools in the pharmaceutical domain. **Thakkar, A., Kogej, T., Reymond, J.L. and Ola Engkvista, O.** 154, 2020, Chem. Sci., Vol. 11.

About Bayer

Bayer is a global enterprise with core competencies in the life science fields of health care and nutrition. Its products and services are designed to benefit people by supporting efforts to overcome the major challenges presented by a growing and aging global population. At the same time, the Group aims to increase its earning power and create value through innovation and growth. Bayer is committed to the principles of sustainable development, and the Bayer brand stands for trust, reliability, and quality throughout the world.

For more information, visit bayer.com

About CAS

CAS is a leader in scientific information solutions, partnering with innovators around the world to accelerate scientific breakthroughs. CAS employs over 1,400 experts who curate, connect, and analyze scientific knowledge to reveal unseen connections. For over 100 years, scientists, patent professionals, and business leaders have relied on CAS solutions and expertise to provide the hindsight, insight, and foresight they need so they can build upon the learnings of the past to discover a better future. CAS is a division of the American Chemical Society.

Connect with us at cas.org

BAYER

CAS
A division of the
American Chemical Society