Preview

Toward machine learning-enhanced high-throughput experimentation for chemistry

Sarah Callaghan[1],*
[1]Cell Press, 50 Hampshire St, Cambridge, MA, USA
*Correspondence: s.callaghan@cell.com

https://doi.org/10.1016/j.patter.2021.100221

High-throughput experimentation in chemistry allows for quick and automated exploration of chemical space to, for example, discover new drugs. Combining machine learning techniques with high-throughput experimentation has the potential to speed up and improve chemical space exploration and optimization.

The *Trends in Chemistry* February 2021 issue is a special issue on machine learning (ML) for molecules and materials. This special issue is understandably targeted toward the domain science of chemistry, rather than having the data science focus of *Patterns*, but it does highlight the important ways that machine learning informs, bridges, and aids aspects of the synthesis, discovery, and optimization cycle for new molecules and materials. Performing these tasks has historically been extremely difficult, costly, and/or labor intensive, so the application of machine learning to speed up this process has the potential to drive progress in this field. The guest editors of this special issue are Prof. Rafael Gómez-Bombarelli and Dr. Alexander B. Wiltschko.

The focus of this preview is the article ''Toward Machine Learning-Enhanced High-Throughput Experimentation'' by Eyke et al.[1] I chose it as a representative sample from the special issue, as it discusses many of the issues with data that are common across many fields, not just chemical discovery and synthesis.

High-throughput experimentation (HTE) allows many parallel chemistry experiments to be conducted simultaneously and more efficiently by using a variety of automated routine chemical workflows. The resulting experiments are conducted uniformly and more cheaply, and the analysis datasets are generated consistently. This allows the properties of large chemical libraries to be screened quickly and cost efficiently, helpful in a field where many experiments are required to make discoveries.

In the chemistry domain, much work has been done on ML-based experimental design tools and automated experimentation platforms. Combining these two methodologies has great potential to speed up and improve chemical space exploration and optimization. This combination also has the advantage of being self-reinforcing: the ML algorithms improve the efficiency with which the platforms can navigate chemical space, and the data that are collected on the platforms can be fed back into the ML models to improve their performance, although the most effective way of doing this combination is still up for debate.

The article describes the developments in ML for chemistry that facilitate data processing, experimental design for maximally efficient experimentation, and applications such as synthesis planning. The authors also describe the latest experimental platforms, including advances in platform-level control systems, hardware implementation, and comprehensive data capture and analytics.

Integration of automated analytical instruments that can generate a lot of information while preserving throughput, along with ML algorithms capable of automatic processing of the data, are a common theme in other physical science domains as well as chemistry. The authors point out that ''systems that automatically upload the data to reaction databases and/or export it into standardized formats that can be included in the supplementary information of publications to facilitate later extraction would help overcome the issues with existing data.'' My feeling is that this is a good first step, and the use of community standards and commonly used and trusted data repositories is essential. I would encourage the community to investigate data sharing and archiving systems in other physical science domains and also not to relegate the important information about the data to the supplementary information. Data are first-class research objects and are an essential part of ensuring scientific verifiability and reproducibility.

The discussion of automated HTE platforms acknowledges the fact that these platforms tend to be well suited to explore narrow chemical spaces, although efforts are ongoing to expand these spaces. Many powerful ML models have been reported in the literature for this domain also, but unsurprisingly, their accuracy and domain of applicability (DOA) is constrained by the available data.

A completely automated synthesis platform depends on access to a model that can readily predict the ''best'' route to a target compound (where ''best'' can depend on a wide range of, sometimes conflicting, factors). Existing datasets often suffer from missing information, or dataset imbalance, and many need substantial data cleaning and curation to be suitable to use with ML techniques. As a result of these issues, existing synthesis planning tools are generally capable of suggesting viable routes but are unable to fully specify synthesis recipes. The article gives specific examples of these and describes familiar results for researchers trying to create general use models, in that models trained on one dataset perform badly on others.

ML models require large amounts of data, and so researchers need to use pre-existing data. As is the case with so many experimental domains, the historical data available for chemical ML lack sufficient quality and/or relevance to fulfil objectives of interest. A strategy, common across domains, is to augment the available literature data to make them better suited to the task, which also means extracting data from the literature in a useful and standardized way. Quantity is not enough, however; data relevance and data quality are also vital aspects that need to be considered.

Sometimes the community has no choice but to generate higher-quality data. As we all know, brute force methods for data generation are inefficient as well as inelegant, and computational models are expensive to run, not only in terms of time but also in terms of carbon and electricity. Efficient experimental design tools, whether they're based on new or pre-existing data, navigate the chemical space and avoid the collection of redundant information. These tools narrow the experiments to be run from the set of all possible experiments in a domain to find the balance between those that are most informative (exploration) and/or most likely to be optimal (exploitation).

The focus on getting good quality data does have the benefit of an increasing community appreciation of the value of comprehensive data capture, aided by new initiatives such as the Open Reaction Database (https://docs.open-reaction-database.org), which aims not only to be a data repository but also to offer guidance on what kinds of data are useful to collect.

As well as a discussion of data collection and quality, the article also outlines methods of merging ML with traditional statistical methods of optimal experimental design for navigation of high-dimensional chemical space. These include the following:

- Traditional design of experiments (DOE) methods for reaction optimization tasks in a small design space involving a small number of primarily continuous variables.
- Bayesian optimization (BO) using a Gaussian process (GP)-based surrogate model to relate the input variables to the objective, although this does come with a computational expense associated with fitting GPs and optimizing the acquisition function in high dimensions. It is becoming common to perform GP-based BO in a dimensionality-reduced space defined using some sort of autoencoder such as a variational autoencoder (VAE) or more traditional dimensionality reduction algorithms like principal-component analysis (PCA), as this allows higher input dimensionality. The combination of BO with generative models is also a popular area of chemical research in recent years.
- Bayesian neural networks (BNNs) can also be used to construct the probabilistic surrogate model.
- Traditional neural networks (NNs) and random forests (RFs) can also be used as surrogate models and are therefore useful in large design spaces with high input dimensionality, even though they are not innately probabilistic. Strategies for uncertainty estimation for NNs and RFs exist, allowing exploration-exploitation experimental design schemes analogous to those deployed for BO.
- Other experimental design strategies mentioned include those based on reinforcement learning and divergence measures.

Critically, the information that is recorded during experimentation directly determines the types of chemical models that can be constructed from the data. Many current HTE platforms for reaction screening achieve increased throughput by initially restricting the analysis to a small set of low-cost observables, and the most promising or interesting results are subsequently investigated in greater detail offline. While this tiered approach has yielded promising results, the information derived from the initial, high-throughput phase lacks enough detail to be useful for most general modeling tasks, so there is a balance that must be found between the resources needed to comprehensively analyze a sample and the throughput needed to navigate large chemical spaces. A promising development for this problem is the use of automated, high-throughput, label-free techniques that can probe reaction chemistry in finer detail than targeted methods, automated at both the instrument and the data-processing levels.

Robust control software that is capable of translating model predictions into machine-executable tasks and workflows that provide comprehensive analysis of the molecules produced on these platforms are critical to provide information-rich datasets for ML efforts. Existing platform control networks are powerful but require specialized control-systems knowledge to implement and modify—knowledge that chemistry end-users do not typically have, making this a substantial barrier to entry. For ML-enhanced HTE platforms to be broadly accessible, there must be serious consideration of the operational design.

As the authors conclude: "The potential to quickly generate tailormade datasets with ML-enhanced HTE represents a promising path toward accurate models with broad capabilities that can be systematically created on demand." This work requires a close collaboration between domain researchers and data scientists but is an area that has a great deal of promise and potential.

WEB RESOURCES

Open Reaction Database, https://docs.open-reaction-database.org

REFERENCES

1. Eyke, N.S., Koscher, B.A., and Jensen, K.F. (2021). Toward Machine Learning-Enhanced High-Throughput Experimentation. Trends in Chemistry *3*, 120–132.