ML and HTE: ACD/Labs

Scientific software firm ACD/Labs, whose solutions encompass analytical data management, processing and analysis, earlier this year announced the latest version of its Spectrus Software for analytical and chemical knowledge management, supporting instrument techniques such as MS, chromatography and spectroscopy. The new version began shipping in September.

Among the new features in the latest release are capabilities for Spectrus' Katalyst D2D (Design to Decide) high-throughput experimentation (HTE) module for experiment planning, execution and analysis, available through one interface. Chemical HTE refers to a large number of parallel experiments designed to test and optimize chemical reactions for applications such as organic synthesis. ACD/Labs' Katalyst D2D supports HTE for several fields ranging from drug discovery to materials science. Katalyst D2D also enables the preparation of ML training sets that can be used to train clients' ML models.

An ML-based technique employed in Katalyst D2D to make HTE design, planning and execution more efficient is Experimental Design Bayesian Optimization (EDBO). "Bayesian statistical methods provide a sound mathematical framework to combine prior knowledge about controllable variables of importance in a process, if available, with the actual data," as defined by Enrique del Castill and Marco S. Reis in their paper, "Bayesian Predictive Optimization of Multiple and Profile Response Systems in the Process Industry: A Review and Extensions," published in 2020 in the journal *Chemometrics and Intelligent Laboratory Systems*.

> *"So even though this is a relatively new option for Katalyst customers, our collaborative efforts have demonstrated significant reduction in the number of physical optimization reaction trials required to optimize complex reactions/processes."*

EDBO enables learning from each experiment in order to better design the next version of that experiment. Andrew Anderson, Vice President, Innovation and Informatics Strategy at ACD/Labs, gave an example. "The full factorial, the full combination of experiment variables, would be 10,000 individual reactions that you would have had to run. But when you're doing an iterative process with Bayesian, it is an inference-based process," he told ***IBO***. "The whole point of Bayesian is you start with an initial cherry picked subset of those variables to actually run. You check the results of those, and then you feed that back to the algorithm to say, 'These variables have these effects on the outcome…[and the software then] recommends another set of iterations to run, ultimately arriving at an optimized set of variables."

Mr. Anderson described in detail the Katalyst D2D workflow for experimental design. "Katalyst users have the ability to define their critical process parameters and process objectives. In the case of chemical synthesis optimization, these are defined as reaction conditions and materials as features and variables; [and] reaction yield and output material quality as optimization objectives. The system automatically generates a 'full factorial' experiment scope," he explained. "This information is sent via Katalyst Web Service to a variety of customer-preferred 'Optimizer' services—the output is automatically consumed by Katalyst for facile experiment execution and analysis. Analysis results are automatically submitted (as 'responses') to the Optimizer for the next round of experiment trials."

Sanji Bhal, Director of Marketing and Communications, described to ***IBO*** ACD/Labs' work with Bristol Myers Squibb utilizing EMBO in Katalyst D2D. "For the reaction they were looking to optimize, there were 1,296 possible experiments, as indicated by the full factorial design. Instead, using the Bayesian optimizer allowed for four rounds of six experiments (just 2% of the full factorial design)—24 experiments to identify optimal conditions." As Mr. Anderson further explained, "This release is based on a multi-year collaboration. So even though this is a relatively new option for Katalyst customers, our collaborative efforts have demonstrated significant reduction in the number of physical optimization reaction trials required to optimize complex reactions/processes."

The latest version of Katalyst includes a dedicated Bayesian module that enables clients to use their own Bayesian models. "It has been developed in a 'model-agnostic' fashion. While the module includes an 'off-the-shelf' option, the system was developed to offer clients to integrate to their preferred (open source or commercial) model," noted Mr. Anderson.

Katalyst D2D's ML-based also capabilities include digital twin enablement. Specifically, Katalyst automatically prepares digital representations to make comparative analyses easier and reduce the number of physical experiments required to achieve process development and/or material output objectives. "Foundationally, one of our key Spectrus Platform–wide capabilities is to help our clients enable the creation of ML-friendly Substance and Process digital twins," explained Mr. Anderson. "By augmenting your experts with this type of ML and AI capability, you have additional security and comfort that ML-type applications may reveal insights that you as a scientist may not readily observe."

Among the capabilities of Katalyst D2D digital twins Mr. Anderson cited were "automated 'assembly' of digital representations of Experiment + Sample + Analysis data." He also highlighted other capabilities, such as "human access (via JS [Java Script] 'family' [of programming languages] and machine access (via API and/or automated processing) to these assembled data packages," which make it easier for scientists to directly review the data and train their ML models using it.

Also designed to improve clients' ML training using Spectrus data are new features in Katalyst D2D for "automated data marshaling and conversion enhancements—supporting an even wider range of instrument-generated 'composition and identify-indicating JSON Objects [a data format of the JSON open standard file format and data interchange language],'" stated Mr. Anderson. JSON, he noted, "allows for easy digestion into a ML-type application."

Another improvement enabling better training of ML models by scientists is described by Mr. Anderson as "facile data science." A common strategy for ML enablement is the use of logical views in relational databases. "In the 2024 release, data scientists can, 'self-configure' their own logical data views for establishing reliable and robust data pipelines for model development and validation." Explaining self-configuring, he told ***IBO***, "Data scientists may not want to have to do all of the data preparation to get a portable training data set right for ML. So what we have in our interface is the ability, with one-button click, to extract relevant data or configure persistent data pipelines to get the data out really fast. This avoids having to call a database administrator to work on applicable SQL queries, which would get the data out in the format that data scientists need."

Another ML-enabled feature of Katalyst is the AutoArray Module. "One of the greatest challenges in HTE is to convert 'conceptual' experiment designs (coming from DoE, Bayesian, etc.) to 'physical' experiment designs (usually plate maps with specific material locations/amounts and corresponding dispense operations," observed Mr. Anderson. "We collaborated with an existing client to develop an ML-based automated parallel/HTE reaction array mapping for Katalyst D2D."

Mr. Anderson also described the extent of other ML tools the company is implementing. "Our focus is to provide 'on-demand' ML model development and validation focused on our key client application areas—process/control methods development, material 'grouping'—based on the breadth of molecular, spectral and chromatographic material attribute assemblies."

This includes ML-based prediction tools. "ACD/Labs currently offers a variety of prediction models supporting neural networks-based spectral prediction, chromatographic retention time prediction, physicochemical property prediction [and] ADME and toxicology prediction," noted Mr. Anderson. "Finally, I should add that, like most software firms, our R&D efforts include evaluations of various large language models and related UI integration activities."

The company's ML activities also include partnerships, such as a recent agreement with AI firm Atinary Technologies (Atinary) for integration of Atinary's SDLabs (Self-Driving Labs) system for optimizing workflows. Mr. Anderson told ***IBO***, "Our collaboration with Atinary covers a wider range of ML application areas. Current efforts are focused on optimization (Atinary offers a number of unique ML capabilities)—we'll be announcing details regarding our first integration before the end of the calendar year."