ACD/LABS [ADVANCED CHEMISTRY DEVELOPMENT, INC.]

Facilitating AI/ML in Chemistry R&D

High Throughput Experimentation and Data Science

High throughput experimentation (HTE) facilitates the efficient exploration of reaction conditions through miniaturization and parallelization. It is also ideal to produce consistent, comprehensive datasets for machine learning (ML) data models, and can benefit greatly from Artificial Intelligence (AI)-assisted experiment design.

Pharmaceutical R&D organizations have invested in hardware to automate reaction execution, and in software to streamline these data-rich workflows. HTE teams at three of the world's leading pharmaceutical organizations have deployed Katalyst D2D software (designed to support HTE) to streamline parallel chemistry workflows, leverage data science, and produce data for data hungry AI/ML applications.

"We want to leverage AI/ML to help design experiments and give us answers more quickly. We're starting to pull data out of Katalyst as JSON files and push it into data visualization tools like Spotfire to understand trends and investigate data science applications."
Scientist Company #1

One organization is using the combination of HTE-generated datasets and ML-enabled experiment design to optimize reactions with 40-50% fewer reactions.

AI/ML Dataset Production with HTE

ML models that predict reaction outcomes are an excellent way to accelerate drug discovery. Robust predictive AI/ML models for the prediction of successful synthetic experiments require:

- High quality data
- Consistent acquisition and processing of data
- Negative and positive results
- Reaction conditions, yields, and side-product formation

"Katalyst ensures that all the criteria for building robust predictive AI/ML models are met for our HTE screens. Katalyst's ability to help scientists easily design large arrays of experiments and record data in a consistent fashion holds great promise for building efficient machine learning models."
Head of HTE, Company #2

Company #2 began to mine data from Katalyst for AI/ML applications within a year of deployment.

AI/ML Dataset Production at Company #3

Templated reaction designs coupled with automated analysis encouraged the team at Company #3 to think about Katalyst as more than just software for streamlining high throughout chemistry. They looked to Katalyst as an ideal way to produce large datasets to explore specific reactions and predict future outcomes.

Choosing a transformation used widely in medicinal and process chemistry and a number of reaction parameter variables (solvent, coupling reagents, and bases), the team set about investigating dataset generation using Katalyst. 3000 data points were generated over the course of the study that were quickly analyzed and exported to build a "proof of concept" machine learning model for reaction yield prediction (see Figure 1).

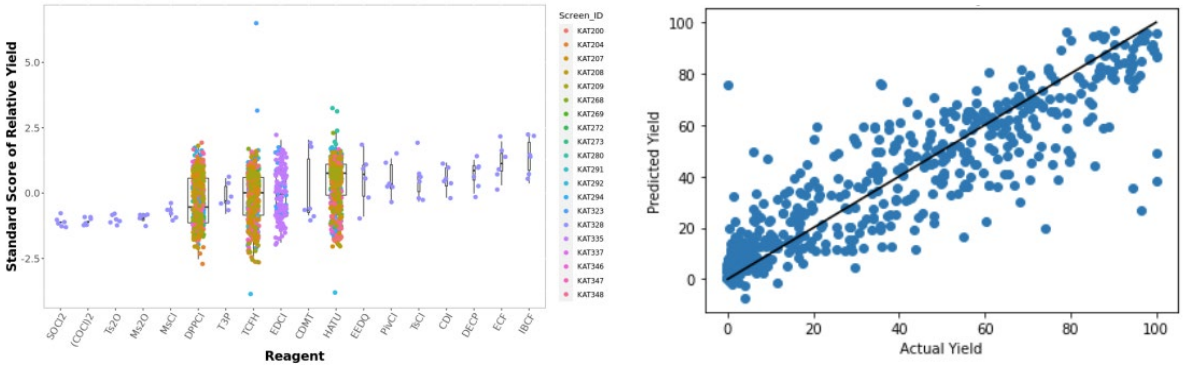

Figure 1. ~3000 data points generated for amide coupling (left) were used to create a preliminary machine learning model for predicting reaction yield (right scatter plot).

"Having a robust database and data architecture allows you to go from an experiment to a data science model very quickly, without the need to do very tedious data cleaning tasks. From a data science perspective, the data goes directly from running the experiment to being written to the Katalyst database. The data can be pulled from there into Python to then do your data science workflows, almost seamlessly."
Associate Scientific Director, Company #3

Leveraging ML in HTE Experiment Design

The space of chemical development can be extremely broad, sometimes requiring the investigation of hundreds or thousands of reactions. HTE may be used to run them all if the lab has the capacity, or scientists may employ methods to narrow the field. They may study the reaction mechanism to understand parameters that dictate reactivity. The process chemistry team at Company #3 turned their attention towards using experimental results to model reactions and reduce the number of experiments required for the identification of optimal conditions.

Scientists at Company #3 collaborated with the Doyle lab at UCLA to create a Bayesian Optimization algorithm that could be applied to exploit known reaction information in the design of high throughput experiments. The resulting Experiment Design Bayesian Optimizer algorithm (EDBO+) was integrated into Katalyst software.

Applying Machine Learning to Accelerate Reaction Optimization

The EDBO+ machine learning algorithm was applied to the design of an amide coupling reaction (Figure 2).

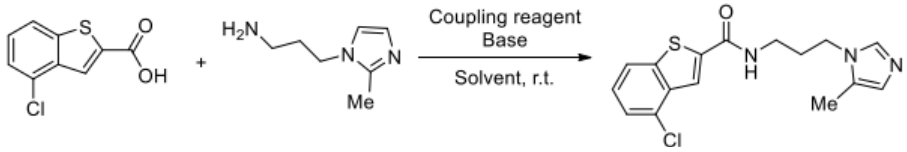

Figure 2. Amide coupling reaction used to test ML-enabled reaction design in Katalyst D2D software.

The study investigated 20 reaction variables—three solvents, two coupling reagents at three different equivalents, two bases, two reaction concentrations, and three different equivalents of coupling reagents, nucleophiles, and bases.

The study represents a total design space of 1296 reactions. Investigating these experimentally would require 14 rounds of experiments in 96-well plates, which is both time-consuming and requires a significant amount of material.

Leveraging ML-Powered Experiment Design

The team set up the variables and objectives of the study in Katalyst. The software suggested the first round of six reactions based on the input data. Upon completion these results are used by the EDBO+ algorithm to suggest the next round of six reactions. This step is repeated in an iterative process until the scientist obtains optimal results.

Reaction Optimization with 50% Fewer Experiments

AI/ML Models Require "Bad Data" Too

Synthetic chemistry research is heavily biased towards successful experiments. Process chemists from experience may recognize parameter combinations suggested by the software that are not ideal. The EDBO+ algorithm looks for the fewest combinations of variables and learns from "poor" outcomes. Ideally, AI/ML algorithms need "good data" (e.g., reactions that produce the desired product) and "bad data" (e.g., failed experiments). Running "non-ideal" reactions, therefore, is part of the machine learning process.

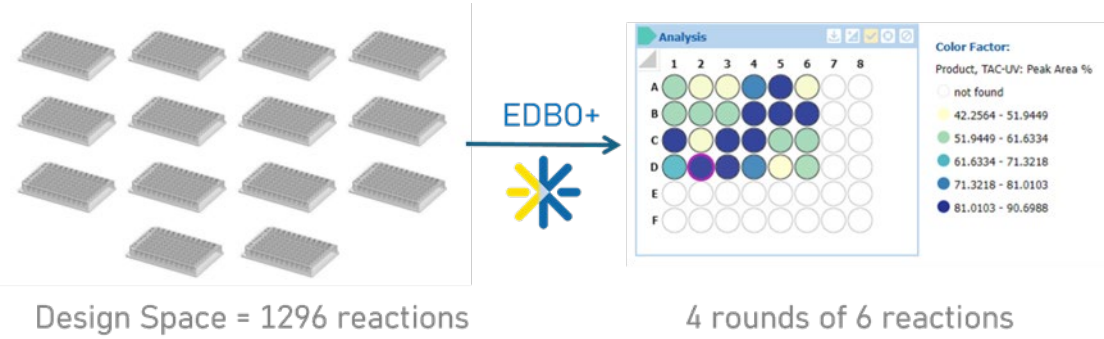

Figure 3. The EDBO+ algorithm integrated in Katalyst enabled optimal reaction conditions to be identified in 24 reactions instead of the full factorial of 1296 reactions.

Using the EDBO+ machine learning algorithm, the group identified optimal conditions for the amidation reaction in 24 total reactions (four rounds of six), executing only a fraction (2%) of the full factorial 1296 reactions to identify optimal experimental conditions (Figure 3). This was a significant savings of time, effort, and consumables.

"We were able to identify the highest yielding reaction conditions relatively quickly with a 3% variance between predicted results in Katalyst and experimental screening data."
Scientist, Company #3

"We're finding experimental optima by running 40-50% fewer experiments using the Bayesian Optimization, 30-40 experiments instead of 60-70."
Associate Scientific Director, Company #3

AI/ML and HTE—A Symbiotic Relationship

HT experimentation lends itself to consistent and efficient dataset production for AI/ML. In turn, integration of these technologies into HT workflows delivers measurable results. Research teams are streamlining HTE workflows and improving the efficiency of chemistry R&D by applying AI/ML technologies in the lab. The ultimate goal being to increase productivity and accelerate time-to-market.

More Resources

Webinar: Enabling AI and Dataset Production across the Chemistry Enterprise at BMS