CAS CUSTOM SERVICES℠

CUSTOM-CURATED MACHINE LEARNING TRAINING DATASET ACCELERATES OPTIMIZATION OF ORGANIC SYNTHESIS WORKFLOW

Solution success story

THE CHALLENGE: Disparate and inconsistent open-access data slows machine learning model development

As one of Europe's largest preclinical contract research organizations (CRO), Selvita continuously refines its organic synthesis workflows to deliver faster solutions to its pharmaceutical partners. To fully harness the power of robotics and automation, the company's chemists sought to cluster compounds from similar synthetic reactions to optimize parallel processing and boost production throughput. However, performing reactions in parallel without compromising yields requires precise control over conditions—such as temperature, reagents, and pressure.

Recognizing the importance of Suzuki and amide coupling reactions in their internal workflows and the pharmaceutical industry, Selvita prioritized optimizing these reactions. To avoid costly trial-and-error processes, the CRO's computational chemists developed a machine learning model to predict reaction yields and identify the best conditions for grouping compounds. Yet, achieving reliable predictions to support informed decision-making required a robust machine learning training dataset.

Selvita's team needed to extract and combine relevant information to build a dataset tailored to their specific requirements, but access to reliable databases was limited. The team had to rely on unverified and unstructured open-access sources, which would have required months of data validation and cleaning. Realizing the need for high-quality and reliable data to optimize their automated workflows, Selvita turned to a trusted ally to unlock the full potential of its model.

CAS
A division of the
American Chemical Society

> **"We worked together regularly throughout the process; it was a collaboration that made the project a success."**
>
> Fabrizio Ambrogi,
> Senior Machine Learning Specialist,
> Selvita

THE SOLUTION: A custom machine learning training dataset built in days with the CAS Content Collection™

Selvita collaborated with CAS Custom Services to develop a tailored solution for training and validating its machine learning model.

Working alongside Selvita's computational chemists, CAS Solution Consultants defined the dataset requirements and optimal format to align with the team's objectives. CAS data specialists then meticulously searched the human-curated CAS Content Collection to identify and extract relevant information for the target reactions, reducing the burden of heavy data-cleaning efforts. To ensure a robust dataset, CAS Content Consultants and expert chemists performed quality assurance on each entry using advanced automated tools. Within days, CAS delivered a verified dataset containing 400,000 entries for Suzuki reactions and 200,000 for amide coupling reactions. Using these curated training datasets, Selvita applied filters to categorize and rank common reagents, solvents, and catalysts to group reactions for parallel synthesis.

The model was first validated using reserved test sets of reactions. In addition, Selvita's team went the extra mile to ensure prediction accuracy by validating its model in the real world. Pitted against one of their senior chemists, the model was tasked with predicting optimal conditions using a set of reactions never performed before. Within just a few hours, the model identified conditions that significantly outperformed those chosen by the chemist, who took more than a week to select and cluster them. These promising results confirmed the model's ability to predict optimal reaction conditions for efficient synthesis.

Working with CAS, Selvita avoided months of manual data work and trial-and-error processes, empowering the team to streamline essential organic synthesis workflows and offer faster, more efficient drug discovery solutions.

Find out how CAS Custom Services can help you transform scientific data into actionable, evidence-based insights that maximize investment and fuel success. Learn more at **cas.org.**

CAS
A division of the
American Chemical Society

SVCGENENGWHP102319260101