



CAS CUSTOM SERVICESSM

CUSTOM AI TRAINING DATASET ESTABLISHES NEW STANDARDS IN REGIOSELECTIVITY PREDICTION ACCURACY

Solution success story

The challenge: Open-access data gaps hinder prediction accuracy for proprietary AI tool

As an AI-driven contract research organization (CRO), ChemLex integrates artificial intelligence (AI) with high-throughput automated platforms to streamline chemical synthesis and deliver high-quality compounds. To further maximize production efficiency and minimize material waste, ChemLex sought to optimize synthesis planning for complex reaction pathways. However, planning feasible and efficient synthesis routes requires predicting regioselectivity—the preference of a reaction over competing reaction sites—with high accuracy.

ChemLex developed Regio-MPNN, an AI model designed to predict regioselectivity for metal-catalyzed cross-coupling reactions, which are widely used in medicinal chemistry. For its proprietary tool to deliver accurate predictions and optimize these reaction pathways, ChemLex needed reliable, comprehensive, and accurate data to train its AI algorithm. Yet, the ChemLex team quickly realized that available open-access data lacked the depth required to support robust model training and fuel reliable predictions.

A significant challenge in developing the Regio-MPNN model stemmed from the nature of open-source data, which is often unstructured, limited in volume, and requires extensive cleaning and curation. Additionally, certain metal-catalyzed cross-coupling reaction types operate under harsh conditions, making them less frequently recorded compared to others. This resulted in an imbalanced dataset in terms of reaction types, which further constrained the model's generalization ability across different reaction types.

Threatening its AI model prediction accuracy, open-access data gaps prompted ChemLex to collaborate with a trusted chemical data provider to access the necessary datasets and strengthen its much-needed tool.

The solution: Boosting prediction accuracy with a custom AI training solution

ChemLex and CAS Custom ServicesSM developed a tailored solution that boosts their AI model and improves regioselectivity prediction accuracy.

Working closely with ChemLex's AI department, CAS solution consultants gained a thorough understanding of the dataset requirements and format to meet the team's objectives. In just a few weeks, CAS data specialists scanned the human-curated CAS Content CollectionTM to identify, cross-check, and combine years of organic chemistry data. After a thorough quality assurance process performed by expert chemists, CAS Custom Services successfully delivered ChemLex a customized training dataset curated to Regio-MPNN-specific needs. To ensure seamless integration into its AI model, ChemLex further refined the dataset, collaborating with the CAS team for necessary adjustments.

Powered by the custom training dataset, Regio-MPNN's regioselectivity predictions are significantly improved, outperforming senior organic chemists and previous model architectures. This unmatched accuracy now supports effective synthesis planning, enabling ChemLex chemists to filter out potentially infeasible reactions and reduce material waste.

Working with CAS to optimize Regio-MPNN, ChemLex empowers its chemists to streamline synthesis planning. Through regular dataset updates, CAS Custom Services continues to support ChemLex and its proprietary AI tool to maintain unparalleled levels of prediction accuracy and simplify complex reaction pathways for all-level chemists.



"It is very challenging for humans to collect accurate data from a vast body of literature and transform it into a structured, AI-ready format. That's why we really enjoyed the collaboration, CAS did a great job supporting us."

Xiaoxue Wang, Co-founder and Vice President, ChemLex

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