Product design / development / engineering | Healthcare / Pharmaceuticals | Defect detection / product inspection | Graph-based neural networks

AI-assisted plausible reaction product and impurity structure prediction and elucidation

AMGEN®

BUSINESS PROBLEM

Development of reaction schemes for small molecule synthesis is a key task of Process Development Team at Amgen. Most chemical reactions result in numerous by-products and side-products, apart from the intended major product(s). A priori prediction of all products is essential to ensure that the final drug substance is free from unintended impurities. While chemists can predict nearly all products of a single reaction step, tracking propagation of product/impurities along multi-step reactions becomes challenging. Conversely, identifying impurities post hoc from mass spectrometry data presents another significant challenge.

DATA SOURCES

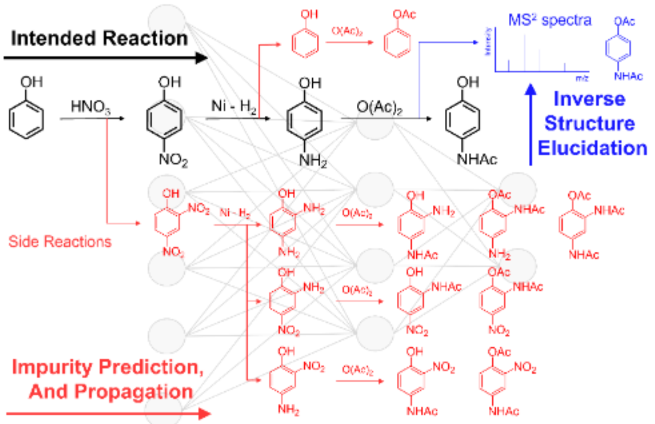
For impurity prediction, chemical reaction schemes for one or more Amgen projects in Process Development will be used to validate the tool. Benchmarking will be done by obtaining examples of selected named reactions from the literature. For inverse structure elucidation training and evaluation, MS2 data will be obtained from open-source databases, along with usage of proprietary Amgen data.

Data Types and Format

The reaction schemes will be received as ChemDraw files, with a document enumerating details for individual reaction steps. MS data will be received as m/z versus intensity, in a Microsoft Excel file.

APPROACH

Impurity prediction will be done by modifying ASKCOS, an AI-based reaction predictor. The predictor will be run iteratively to track impurity propagation in multi-step reactions. For inverse structure elucidation, MSNovelist, a pre-trained ML model, will be adapted to predict molecules from mass spectrometry data. A user interface will be developed to assist chemists interact with the program.



Author: Somesh Mohapatra

IMPACT

The solution primarily impacts (1) identification of impurities, (2) high-throughput reaction screening, and (3) raw materials risk assessment, with all the steps being absolutely core to the synthetic drug substance commercial process development business at Amgen. Impurity identification, both a priori and post synthesis, aids process development, with the former helping in the optimization of reactions, and the latter in the identification of possible impurities in the product mixture. For high-throughput reaction screening, the tool helps in narrowing down the chemical space of possible reactions, balancing the exploration of a wide range of reactions versus exploitation of selected reaction pathways, thereby accelerating experimental efforts for route selection and route optimization. Additionally, the solution helps in assessing the risk posed by low-level impurities in raw materials – as purchased, reaction intermediates, and API starting materials.

DRIVERS

We are constantly striving for smarter, faster, better commercial process development. This tool could: (1) give us higher quality data from high-throughput reaction screening experiments, ultimately facilitating better understanding and long-term development of reaction data sets, and (2) save us significant time in critical structure ID task that come up through development and tech transfer.

BARRIERS

The biggest bottleneck in the implementation of the project was access to computational resources at Amgen, such as, installation of relevant software, getting access to servers, and administrative privileges. Accessing experimental data, both chemical structure and spectra, for computational purpose; and ensuring that the ML model complemented experimental knowledge of process development without being redundant, were the other challenges.

ENABLERS

Amgen is one of the best places that encourages intrapreneurial thinking, providing strong support to the interested folks to disrupt the status quo. The desire to be at the forefront of technological advancements, by leveraging both external and internal research and development, propelled the project. Additionally, the MIT MLPDS consortium and existing academic collaborations provided necessary technical support in advancing the project.

ACTIONS

In addition to the technological development of the codebase, training/testing of ML models, and validation of unseen data, conversation with people at Amgen and building relationships with people, both internal and external, helped in implementing the solution. The conversations helped in narrowing down the scope, and identifying the specific requirements of scientists and engineers, thereby ensuring that the solution is going to be more useful.

INNOVATION

Combining AI-assisted impurity prediction to form a candidate set of possible impurities, with the inverse structure elucidation model to down select from that candidate set, presents a paradigm shift in process development. This innovative approach saves significant amount of time and resources as compared to the current practice of involving subject matter experts for manual identification, and/or using commercial packages, such as, Virscidian.

IMPROVEMENT

The solution for impurity prediction helps in iterative evaluation of potential impurities from sources other than reactants and reagents, such as low-level impurities in raw materials. The inverse structure elucidation tool is expected to save a lot of time in identifying impurity molecules from spectra. In combination, both the tools are expected to be of significant help for process chemists and engineers and accelerate their current workflow.

BEST PRACTICES

The solution will comprise of a codebase with relevant comments. Replicating the work using the codebase should be done by installing the exact libraries, as used in the development, and following the instructions in the documentation.

OTHER APPLICATIONS

The solution can be used both before and after pivotal commercial process development, i.e., in pre-pivotal process development and technology transfer. In the pre-pivotal step, it can help in narrowing down pathways of interest by avoiding reactions producing higher amounts of impurity. The inverse structure elucidation can help in identifying unknown impurities from spectra and address any concerns around potential genotoxicity or mutagenicity.