AI-assisted reaction impurity prediction and inverse structure elucidation

by

Somesh Mohapatra

Ph.D., Massachusetts Institute of Technology, 2022
B.Tech., Indian Institute of Technology Roorkee, 2018

Submitted to the MIT Sloan School of Management
in partial fulfillment of the requirements for the degree of

Master of Business Administration

in conjunction with the Leaders for Global Operations program

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2023

Author .....
MIT Sloan School of Management
November 18, 2022

Certified by .....
Rama Ramakrishnan, Thesis Supervisor
Professor of the Practice, Sloan School of Management

Certified by .....
Rafael Gómez-Bombarelli, Thesis Supervisor
Assistant Professor, Department of Materials Science and Engineering

Accepted by .....
Maura Herson
Assistant Dean, MBA Program, MIT Sloan School of Management

AI-assisted reaction impurity prediction and

inverse structure elucidation

by

Somesh Mohapatra

Submitted to the MIT Sloan School of Management
on November 18, 2022, in partial fulfillment of the
requirements for the degree of
Master of Business Administration

Abstract

Identification and control of impurities play a critical role in chemical process development for drug substance synthesis. Most chemical reactions result in a number of by-products and side-products, along with the intended major product. While chemists can predict many of the main process impurities, it remains a challenge to enumerate the possible minor impurities and even more of a challenge to track and propagate impurities derived from raw materials or from step to step. Further, in the absence of a systematic means for listing out possible-low-level impurities and performing impurity propagation, inverse structure elucidation – that is, identifying unknown impurities *post hoc* from analytical data, such as mass spectrometry data – presents a significant challenge.

In this work, impurity prediction was established by developing an AI-based reaction predictor that takes as input the main reactants, and reagents, solvents, and impurities in these materials. Further, the predictor was run iteratively to track impurity propagation in multi-step reactions. For inverse structure elucidation, a chemistry-informed language model was developed to translate mass spectrometry data to potential molecular structures, which can then be checked for matches against the predicted chemical reaction products. The impurity prediction tool was applied to synthesis of common small molecule drugs — paracetamol and ibuprofen, and the inverse structure elucidation tool was used for the identification of chemical structures from publicly available electrospray ionization mass spectrometry data, The models were applied to proprietary Amgen programs, both small molecule drugs and biologics, with significant results noted in both projects.

Thesis Supervisor: Rama Ramakrishnan
Title: Professor of the Practice, Sloan School of Management

Thesis Supervisor: Rafael Gómez-Bombarelli
Title: Assistant Professor, Department of Materials Science and Engineering

Acknowledgments

I would like to acknowledge the support of the MIT LGO program, advisors at MIT, Amgen, and my friends and family for helping me navigate this thesis.

I will always be grateful to Patricia Eames (Patty), Anna Voronova, Ted Equi, Thomas Roemer, Tamar Shulsinger, Amy Levin, and other staff, at the MIT LGO program. Without their help, I would not have been able to work through the thesis in the accelerated timeline that I have been able to. They have gone out of their way multiple times to accommodate my request to have an early submission and graduation, and I will always remain thankful to them for their advice and support.

As a part of this thesis, I am thankful to Prof Rafael Gomez-Bombarelli (Rafa), Engineering Advisor and Assistant Professor in the Department of Materials Science and Engineering, and Prof Rama Ramakrishnan, Management Advisor and Professor of Practice in Sloan School of Management, for their guidance as I worked on the thesis. Their suggestions have helped me shape the thesis to be more impactful, both in terms of the science and utility to Amgen. At times, their advice on using my time effectively to work on the more critical aspects of the thesis, without doing software engineering tasks, was helpful in allocating time appropriately to the project.

At Amgen, I would like to acknowledge the leadership support before and during the internship. Oliver Thiel, Seth Huggins, and Daniel Griffin (Dan), have been one of the most welcoming people, helping me shape the project, carry out the work, and provide opportunities to present it within and outside Amgen. I thank them wholeheartedly for believing in me, and supporting me during this period. Additionally, I would like to thank the LGOs at Amgen, Chris Garvin, Philipp Simons, Ketan Kumar amongst others, who extended their support, and helped me connect with a lot more people within Amgen.

I will always remain indebted to the LGO '23s, for being the most wonderful cohort, and providing a super-strong support system with whom I could interact, discuss common challenges faced at the internship, and come up with solutions.

Ultimately, I would like to thank my friends and family, for being supportive as I worked on during this thesis.

THIS PAGE INTENTIONALLY LEFT BLANK

Contents

List of Figures **9**

1 Background and Introduction **13**

1.1 Company Background - Amgen 13

1.2 Related Work 14

1.2.1 Chemical process development for synthetics 14

1.2.2 Structure elucidation and impurity identification 16

1.3 Problem Statement, and Thesis Overview 17

2 Approach **19**

2.1 Interviews with experimentalists 19

2.2 Development of impurity prediction pipeline 20

2.2.1 Reaction types 21

2.2.2 Input/output file system 22

2.2.3 Prediction model 24

2.2.4 Impurity prediction and propagation 25

2.3 Development of inverse structure elucidation pipeline 26

2.3.1 Input file system 26

2.3.2 Output file system and post-processing 27

2.3.3 Prediction model 28

3 Applications **29**

3.1 Impurity prediction for small molecule drugs 29

3.1.1 Paracetamol synthesis 29
3.1.2 Ibuprofen synthesis 33
3.1.3 Application at Amgen 35
3.2 Inverse structure elucidation from MS^2 data 35
3.2.1 Application at Amgen 36

4 Limitations, and Future Directions **37**
4.1 Limitations 37
4.2 Future Directions 38

5 Conclusion **41**

A List of Interviewees **49**

List of Figures

2-1 **Different reaction types for impurity prediction.** Both impurity prediction and propagation for the products from step 1 have been shown in the schematic, and only the step impurities and the major product of step 2 are shown. The primary reactant for step 1 is denoted as a white triangle, with the primary product denoted as a blue triangle. The primary product of step 1 acts as the primary reactant for step 2, and results in the primary product of step 2 - red triangle. The legend denotes the additional reagents and solvents as squares, with the impurities in them, noted as green squares. The numbers within these squares are for the specific step they are a part of. Reagents and solvents for each step are on the reaction arrow, with the subscript denoting the specific step. The different shapes - rhombus, pentagon, and hexagon - are for the different impurities arising from reactions other than the major reaction for that particular step. 21

2-2 **Input file for paracetamol synthesis. A.** Screenshot of the input schema is shown for different reactants, reagents, and products for the 3-step synthesis of paracetamol from phenol. **B.** Structures in the input schema have been shown after being processed through the ChemDraw add-in. 22

2-3 **Output file for paracetamol synthesis. A.** Screenshot of the summary output file, with all the structures, noted for step impurities. **B.** Screenshot of the enlisted output file explicitly showing the reactants and respective products. 23

2-4 **Overview of the impurity prediction model.** The input file is parsed into individual reactions, and the results are aggregated in the output file. The impurity prediction module in the ASKCOS software suite is used to predict the impurities for specific reactions. 24

2-5 **Input and output file for inverse structure elucidation pipeline. A.** Screenshot of the input file in the .mgf format is shown. **B.** Screenshot of the output file generated through the MSNovelist pipeline is shown. 27

2-6 **Overview of the inverse structure elucidation model.** The input is the MS^2 spectra and optionally the elemental composition in a .mgf file, while the output is a set of structures with the different chemical formulae. The prediction model is adapted from MSNovelist. 28

3-1 **Predicted step impurities and propagated impurities in paracetamol synthesis. A.** Reaction scheme for paracetamol synthesis. The specific steps have been marked as **R1**, **R2**, and **R3**, to specify the corresponding step and propagated impurities. Impurities from the reaction steps - **B. R1**, **C. R2**, and **E. R3**. **D.** Propagation of step impurities predicted in **R1** through **R2**. No propagation was predicted for step impurities in **R2**, and propagated impurities from **R1** were not propagated through into **R2**. . . . 30

3-2 **Additional impurities in paracetamol synthesis. A.** Impurities arising from reactions with known impurities, noted in green, along with the reagents. Only one impurity is identified for **R2**, across the three reaction steps. **B.** Splitting of **R2**, into **R2A** and **R2B** to force a reaction intermediate did not result in any additional impurity identification. 32

3-3 **Predicted step impurities and propagated impurities in ibuprofen synthesis. A.** Reaction scheme for ibuprofen synthesis. The specific steps have been marked as **R1**, **R2**, and **R3**, to specify the corresponding step and propagated impurities. Impurities from the reaction steps - **B. R1**, **C. R2**, and **E. R3**. **D.** Propagation of step impurities predicted in **R1** through **R2**. No propagation was predicted for step impurities in **R2**, and propagated impurities from **R1** were not propagated through into **R2**. 33

3-4 **Additional impurities in ibuprofen synthesis.** Impurities arising from reactions with known impurities, noted in green, along with the reagents, for each of the three reaction steps - **R1**, **R2**, and **R3**. 35

3-5 **Chemical structures predicted for two random spectra using the inverse structure elucidation pipeline.** Different structures corresponding to the elemental formula are noted for the two respective spectra in **A.** and **B.** . . 36

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 1

Background and Introduction

1.1 Company Background - Amgen

Amgen is one of the world's leading biotechnology companies, headquartered in Thousand Oaks, California [1]. Founded in 1980, the company is present in 100+ countries and regions, with its innovative medicines reaching millions of people across the world helping them fight against serious illnesses. The company focuses on six therapeutic areas: cardiovascular disease, oncology, bone health, neuroscience, nephrology, and inflammation. Apart from medicines, the company's corporate efforts fall into four categories - Healthy People, Healthy Society, Healthy Planet, and Healthy Amgen, showcasing the holistic efforts of the company to be a leader in science and society [2].

As a drug goes through five stages, namely, discovery and development, preclinical research, clinical research, US Food and Drug Administration (FDA) review and FDA post-market safety monitoring [3], a key component of the discovery and development stage is to develop a process to manufacture the drug at scale. At Amgen, this step is led by the Process Development organization, within Operations. In this organization, the small molecule or synthetics process development for the pivotal and commercial stage is led by Pivotal and Commercial Synthetics.

1.2 Related Work

Drug substance development necessitates a scalable and stable biological or chemical process to manufacture the drug before commercialization [4, 5]. For biologics, such as RNA, peptide, and antibody-based drugs, cell-based processes are used for manufacturing [6, 7]. Small molecule drugs are synthesized through a set of chemical reactions, where initial reactions focus on making the different motifs in the drug molecule, and the reaction products come together in the final steps to complete the intended molecule [8, 9]. As can be noted from the complexity of the reaction scheme in the small molecule synthesis, the by-products of individual reactions create impurities, and these impurities can propagate through, from step to step, to the final drug substance, if not identified and controlled via reaction optimization or purification operations [10, 11]. Thus, chemical process development, involving identification, development, optimization, and scale-up of chemical synthetic routes, is a major activity required in the commercialization of small molecule drug substances [12].

1.2.1 Chemical process development for synthetics

Chemical process development would significantly benefit from *a priori* prediction of possible impurities, thereby accelerating the route selection and optimization processes [13]. For instance, the information about impurities would help in avoiding problematic synthetic routes which might result in products that are potentially genotoxic or mutagenic [14, 15].

Impurity mapping (both hypothetical and known) for drug substance synthetic processes is an important exercise throughout the process development work stream; this can inform route selection early on and underpins the control strategy established in later development [16, 17]. The hypothetical impurity mapping exercise of listing out all possible impurities for a single reaction step, and brainstorming through different reaction intermediates, and how they would be impacting the resulting product, is a process that most organic chemists can do [18]. However, it is hard to do this systematically for even one step. The challenge explodes when taking numerous impurities through

the downstream reactions and tracking how their products react with other reagents and solvents. Thus, it becomes a Herculean task to establish the known impurity map empirically.

Computational impurity prediction methods fall under the broader umbrella of Computer-Aided Synthesis Planning or CASP, which focuses on forward reaction prediction, retro-synthesis, and reaction condition prediction [19, 20]. ASKCOS software suite, developed by Coley and co-workers at MIT [21], and the more recent Python-based database matching workflow [22], are some of the impurity predictors, built by repurposing the forward reaction models.

However, applying the impurity predictors, for application in synthetic process development workflows, highlights a few areas where further development is necessary, such as the current predictors focus on application with pure input materials, aim to predict the major products, and are designed for single-step reactions. The focus on application with perfectly pure reactants, reagents, and solvents, is the right focus when predicting impurities for reactions run in academic or development labs at a small scale in which highly pure material can be used, but leaves out real-world considerations in manufacturing where bulk materials are used and an understanding of what level of solvent, reagent, and starting material impurities can be tolerated is required. Similarly, these models have been trained on datasets, such as Reaxys [23] and Pistachio [24], which collect nearly all the reactions in the literature. The challenge here is that most people report the major products, while failed reactions, and minor products, are not as widely reported, thereby leading to a handicap for the impurity predictors not being aware of the complete spectrum of possible reaction products.

As a result of the current processes, the aforementioned approaches may miss complexities seen in the real world, such as impurities that propagate from prior steps, or ones that are a product of reactions with impurities in starting materials, which are rarely of 100% purity. Moreover, due to the implicit bias of the majority product in the training dataset, they might miss low-level impurities. Unfortunately, these low-level impurities are critical to the drug substance development process and need to be identified and controlled. Thus, it is important to develop computational methods that can address the

challenges around impurity prediction and propagation, and help in accelerating the chemical process development in the pharmaceutical industry.

1.2.2 Structure elucidation and impurity identification

Apart from the *a priori* impurity prediction from proposed reaction schemes, a major part of chemical process development involves the identification of impurities after the experiments. The identification involves analysis of reaction characterization data, such as nuclear magnetic resonance (NMR) [25, 26, 27, 28], liquid chromatography-mass spectrometry (LC-MS) [29, 30, 31, 32], and other analytical data streams. This *post hoc* analysis requires an intimate understanding of the specific analytical methods, thereby necessitating the involvement of subject matter experts from analytical chemistry, and other fields, to help in the elucidation of the impurity structures.

There have been recent computational approaches that have been developed to invert the analytical spectrometric data to identify molecules. Although, *de novo* generation of structure from NMR spectra has been attempted with reasonable success [33, 34, 35], the inversion of tandem mass spectrometric (MS^2) data remains a challenge. Spec2Mol and MassGenie made some advances in this space, by training on electron-ionization mass spectrometry data and generating simpler small molecule metabolites [36, 37], and MSNovelist worked on the problem using electrospray ionization mass spectrometry (ESI-MS) data [38]. A key bottleneck in off-the-shelf usage of these existing methods lies in the lack of fine-tuning for specific drug-like molecules. These models have been trained on datasets, such as those obtained from MassBank and NIST spectral libraries [39, 40], which cover a more general portion of the chemical space, rather than focusing on drug-like molecules.

Structure elucidation and impurity identification process can be significantly accelerated and even automated if we can predict a candidate set of impurities from which to select in structure elucidation from analytical data.

1.3 Problem Statement, and Thesis Overview

In this thesis, we propose a closed-loop AI-assisted method for impurity prediction and inverse structure elucidation from MS^2 data. We leverage current forward reaction predictors to predict a plausible set of impurities and provide the chemists an opportunity to do synthetic route selection and optimization. Additionally, we adapt an existing ESI-MS^2 to molecule predictor to invert spectra for drug-like molecules. We intend to re-frame the impurity prediction problem to predict an inclusive set which allows us to expand on what impurities process chemists can come up with, and then use the inverse structure elucidation pipeline as a filter to identify molecules that are in the set. In the upcoming chapters, we discuss our approaches (Chapter 2), applications of the methods to molecules in the public domain (Chapter 3), limitations and future directions for automated impurity identification (Chapter 4), and summarize our findings from this work (Chapter 5).

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 2

Approach

The thesis was developed by bringing together the technological requirements anticipated in the project, and interviews with potential end-users. The specifications in the project description were outlined to do automated impurity identification. To assess that the specifications aligned with the end-user needs and understand the current working practices, we interviewed more than 60 people across different organizations within Amgen in 100+ individual interactions (Appendix A).

2.1 Interviews with experimentalists

The first round of interviews was facilitated by leveraging the immediate network of my manager, Daniel Griffin, Principal Engineer, Pivotal and Commercial Synthetics. In these interviews, the basic outline of the project and the goals were discussed. These interviews served as the basis for the major connections and helped in building a base for test users and critics as the project progressed. A key outcome of the initial discussions was the realization of the discrete nature of impurity identification. Impurity prediction and inverse structure elucidation were identified as the two parts of the project. With impurity prediction being of stronger significance in route discovery and selection, the inverse structure elucidation was noted to be more significant for route and process optimization and control. Specifically, the prediction was noted to be needed to identify and eliminate risky routes and flag hard-to-remove impurities, and was used primarily

in the initial parts of the process development, while inverse structure elucidation was required for the identification of unknown, low-level, or non-trivial impurities observed throughout process development and characterization.

The two parts come together towards the end to close the loop and enable automated impurity identification. Impurity prediction provides a list of plausible and inclusive set of impurities. Similarly, inverse structure elucidation provides a list of potential molecules corresponding to mass spectrometry or other analytical data. The intersection of these two sets, where the predicted set is invariably a super-set of impurities, and the elucidated set of structures act as a filtering criterion or down-selection function to identify the specific impurities.

Further interviews were conducted by connecting with the suggested people from each of the interviews, with more people branching out from every single interview node. These interviews were conducted over a period of 4 months and involved sharing goals of the project, recent updates on the code, results using public data and from Amgen programs, and requesting suggestions on making the project more useful to folks at Amgen, and aligning it better to the needs.

The discussions across all the interviews provided a clear understanding of the necessity of both systems in enabling automated impurity identification. However, we noted that in regular usage, addressing the current needs of the work, these two parts may be used as stand-alone systems. To connect these independent parts, we realized that a significant amount of software engineering would be needed to build out interfaces, and make them available to the chemists.

2.2 Development of impurity prediction pipeline

The impurity prediction pipeline was designed on the basis of the different possible reaction types, with the reaction predictions done using ASKCOS [21]. At first, we enumerated all possible reaction types and used that as the basis to develop the codebase to do single-step impurity prediction. Extending this to multi-step reactions, or impurity propagation, was a combination of using a super-set constituents reaction, and following

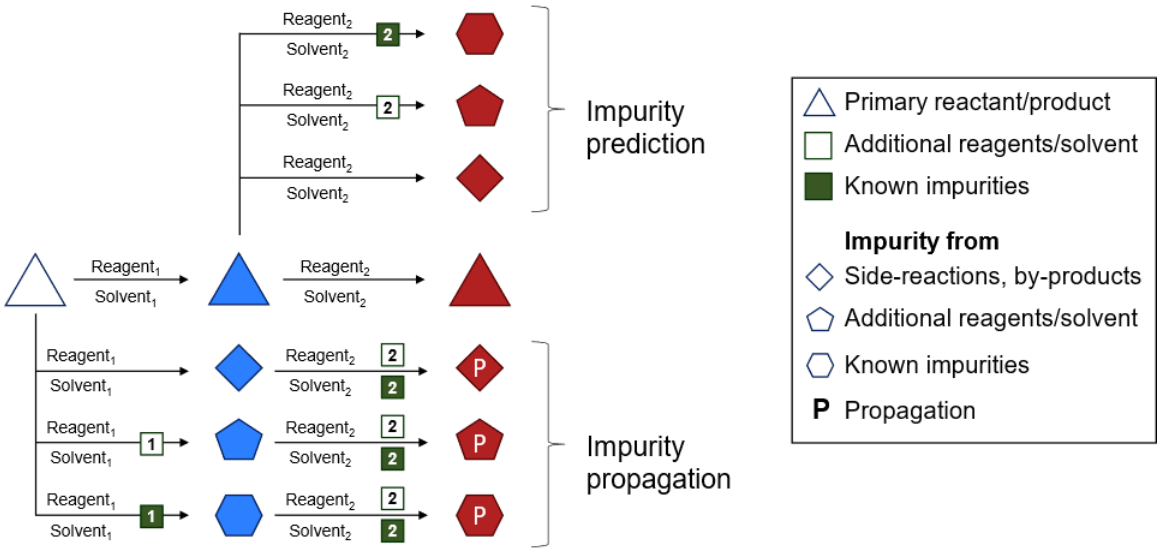

Figure 2-1: **Different reaction types for impurity prediction.** Both impurity prediction and propagation for the products from step 1 have been shown in the schematic, and only the step impurities and the major product of step 2 are shown. The primary reactant for step 1 is denoted as a white triangle, with the primary product denoted as a blue triangle. The primary product of step 1 acts as the primary reactant for step 2, and results in the primary product of step 2 - red triangle. The legend denotes the additional reagents and solvents as squares, with the impurities in them, noted as green squares. The numbers within these squares are for the specific step they are a part of. Reagents and solvents for each step are on the reaction arrow, with the subscript denoting the specific step. The different shapes - rhombus, pentagon, and hexagon - are for the different impurities arising from reactions other than the major reaction for that particular step.

all the impurities to n steps downstream from the point they were predicted. All impurity predictions were done using ASKCOS version 2022.07, with the application programming interface (API) version 2.

2.2.1 Reaction types

To predict a comprehensive impurity map, we outlined different possible reactions, other than the intended major reaction (Figure 2-1). First, we noted impurities arising from possible side reactions with the same set of reactants, reagents, and solvent, but resulting in different products [41, 11]. Second, we captured the impurities when additional reagents or solvents are added, as a part of the work-up or isolation processes, such as crystallization [42, 43]. Lastly, we outlined the impurities arising from reactions with

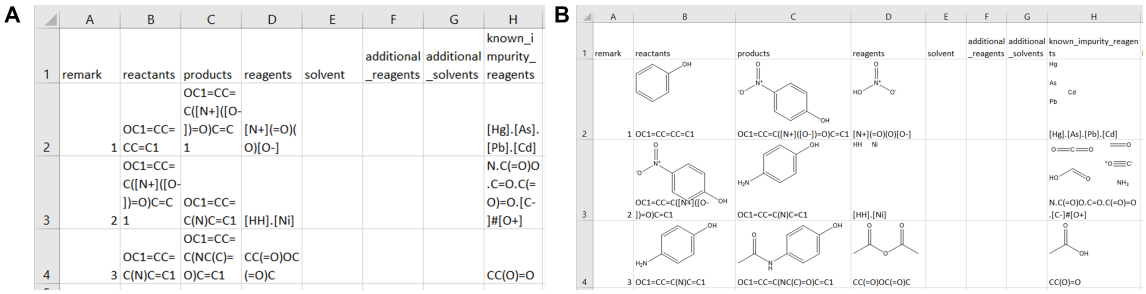

Figure 2-2: **Input file for paracetamol synthesis. A.** Screenshot of the input schema is shown for different reactants, reagents, and products for the 3-step synthesis of paracetamol from phenol. **B.** Structures in the input schema have been shown after being processed through the ChemDraw add-in.

known impurities in the starting materials [44, 45]. Overall, we enumerated three different types of possible reactions, other than the major reaction, for a single step.

For impurity propagation, we propagated individual impurity products predicted as a part of the previous step reactions, with all the reagents, solvent, additional reagents, and solvent, and known impurities [46, 47]. Bringing all of the non-reactants together, instead of predicting four different reactions as earlier, was a decision to make the prediction process computationally tractable, and also not to complicate the readability aspect.

2.2.2 Input/output file system

To make the impurity prediction pipeline easily accessible to chemists, we developed an input/output file system based on Microsoft Excel with ChemDraw version 21.0 add-ins [48] (Figures 2-2, 2-3). The Microsoft Excel tabular format made it easy to specify the different inputs for the reaction, and also analyze the outputs in a similar manner. Given that Microsoft Excel and Python does not natively support chemical structures, we used simplified molecular input line entry system (SMILES) to depict the chemical structures as strings [49]. Using the ChemDraw add-in, we then converted the Excel spreadsheet into a ChemDraw worksheet and showed the chemical structures along with the SMILES in the same cell or box in the spreadsheet [50]. We have used the input and output files for impurity prediction in the case of synthesis of paracetamol from phenol to visualize how the files look like.

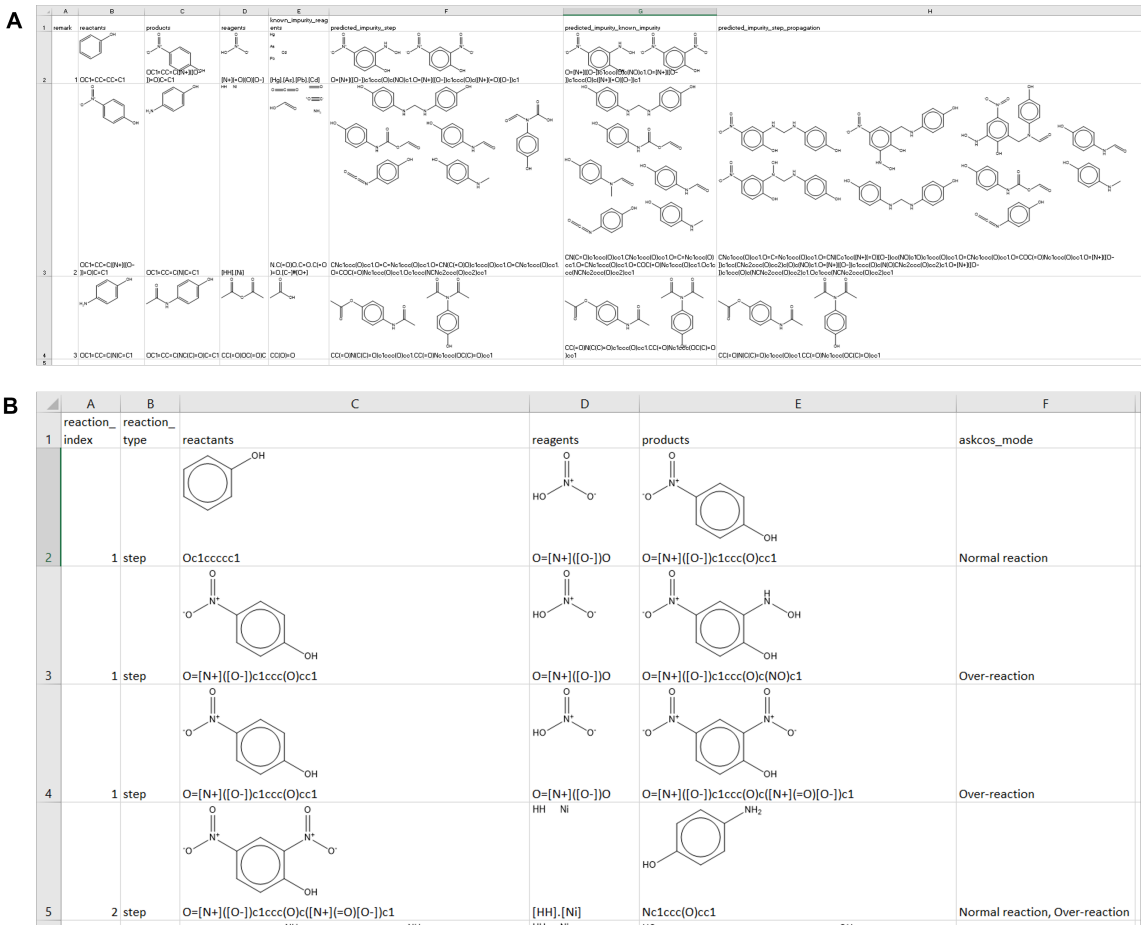

Figure 2-3: **Output file for paracetamol synthesis. A.** Screenshot of the summary output file, with all the structures, noted for step impurities. **B.** Screenshot of the enlisted output file explicitly showing the reactants and respective products.

The input file has a set of columns to specify the primary reactants, reagents, solvent, additional reagents and solvent, and known impurities in the reagents and solvent (Figure 2-2). Across these 6 columns, only primary reactants and products are mandatory, while others remain optional. Individual reaction steps are entered in separate rows and are treated independently of one another, except in the case of prediction for impurity propagation.

We developed Microsoft Excel-based output file systems, in line with the tabular format of the input file. To help with the level of abstraction that end-users need, as per the discussions with the experimentalists (Section 2.1), we developed two formats for the output - *summary* and *enlisted* files. The summary file extends the format of the input file,

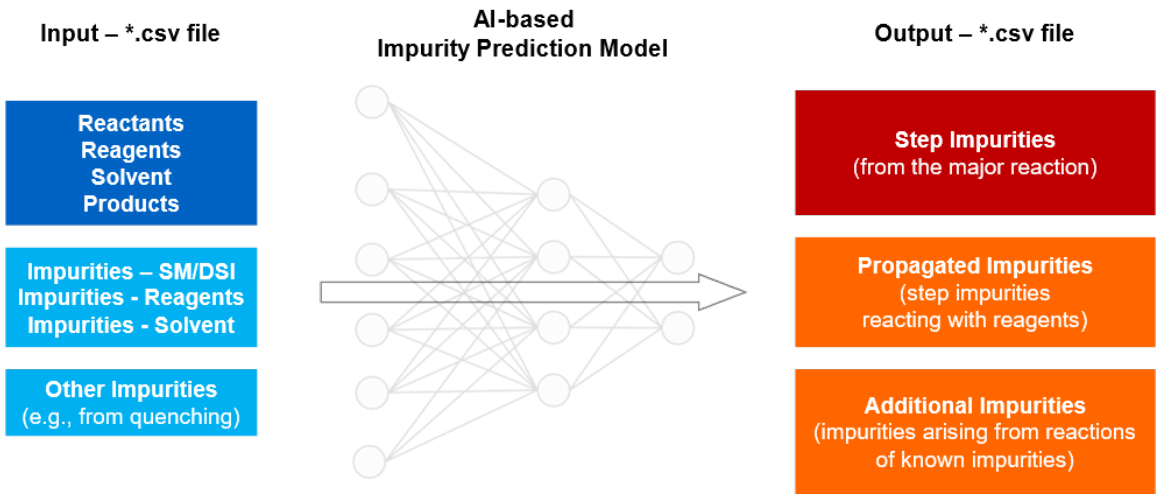

Figure 2-4: **Overview of the impurity prediction model.** The input file is parsed into individual reactions, and the results are aggregated in the output file. The impurity prediction module in the ASKCOS software suite is used to predict the impurities for specific reactions.

and adds the impurities in additional columns in the same rows (Figure 2-3A). The output columns are listed as the impurities arising from the side reactions, additional reagents and solvents, known impurities in the starting materials, and propagated impurities from each of the aforementioned impurity types. In this format, the impurities are all noted together, which made it difficult for the experimentalists to assess how each of the molecules might have come into being. This challenge led us to develop the *enlisted* output file format (Figure 2-3B). In the enlisted format, we showed the specific reactants, reagents, and solvents that led to the impurity molecule. The different levels of abstraction has been shown to have helped the experimentalists identify the source of non-trivial impurities.

2.2.3 Prediction model

For impurity prediction, we used the impurity prediction module in the ASKCOS software suite (Figure 2-4). Specifically, we used the latest model trained on the larger Pistachio dataset [40], as compared to the Reaxys [23] dataset. A key difference between these models lies in the curated data in Pistachio, along with the coverage of more reactions reported in the US Patents and Trademark Office data releases [51]. Since the impurity

prediction module, by default, is meant to produce the top impurities in the dataset, and we were more interested in getting access to the larger set of low-level impurities, we changed some of the default parameters. After a set of carefully designed manual iterations, we settled on increasing the top-k predictions to 10 from the default 3, and decreasing the threshold probability from the default 0.1 to 0.01. The top-k predictions is a criterion to filter the output predictions based on their similarity to the major product, while the threshold probability is a way to calculate the probability of the existence of the product as a result of the reaction under consideration. Here, we would like to clarify that the probability is an implicit function of the training dataset, and the lack of reported impurities in the dataset affects the probability. Also, the probability does not necessarily reflect the chemical feasibility of the impurity product, nor does it reflects the kinetic barriers for a particular product being generated. We have discussed how these chemistry-informed concepts can be incorporated into the impurity prediction pipeline in Future Directions, Chapter 4.

Computational sustainability was a key element in the design of the impurity prediction pipeline, apart from the ease of usage by experimentalists. To ensure that the pipeline remained computationally feasible, within reasonable costs for Amgen, while also providing results that did not take longer times to compute, we tested a range of computing configurations on Amazon Web Services [52]. Ultimately, we settled on the ASKCOS software suite being hosted on an AWS EC2 instance with the default configuration. For the impurity prediction, we queried the ASKCOS instance within the enterprise firewall. All the tests were made on a 4-core Intel i7 system. For the demonstrated 3-steps reaction of paracetamol synthesis, we clocked 4 min of wall time and 719 ms of CPU time to obtain the results.

2.2.4 Impurity prediction and propagation

All the possible reactions arising from the input file system were split into different reaction types, and thus individual single-step reactions. These individual reactions were then queried as a batch for a single primary reaction step. We noted all but the major product as the impurities from the different reactions, and aggregated them in the

summary format, and enumerated them as individual reactions for the *enlisted* format.

For impurity propagation, we batched all the impurities from the previous step, irrespective of their origins, such as side reactions, from additional reagents and solvents, or reactions with known impurities. We propagated these impurities by treating them as primary reactants and considering all possible reagents, solvents, and known impurities in the succeeding step(s). In order to avoid a combinatorial explosion and also understand that isolation (crystallization, followed by filtering, washing, and drying) operations lead to the elimination of most impurities, we kept the number of propagating steps to 1, by default. The codebase has in-built functionality to extend this to n steps, as desired by the end user.

2.3 Development of inverse structure elucidation pipeline

The inverse structure elucidation model was largely adapted from the MSNovelist work, with additional improvements to the input and output file systems [38].

2.3.1 Input file system

We developed a parser to convert the MS^2 data in the tabular format of the ratio of mass to charge, and intensity values, to the Mascot Generic Format (MGF) file system (Figure 2-5A). In the file format, we varied the charge on the molecule, and the chemical formula to bias the predictor in giving a varied set of results.

By varying the charge from -2 to +2, at integer values, we saw a difference in the predicted molecules. Similarly, the presence of the chemical formula biased the model in predicting molecules for the specific formula alone. This aspect was useful if a single formula or multiple formulae could be determined with high accuracy from the high-resolution mass spectrometry instrument processing software. Depending on the accuracy of the mass and the sensitivity of the instrument, a number of formulae could be determined and thus used to get specific results. In the absence of a chemical formula,

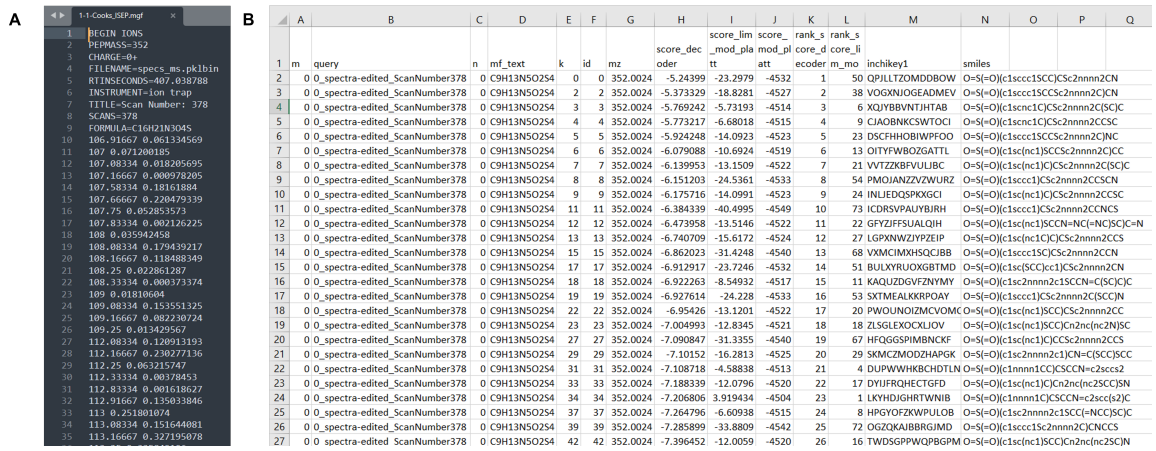

Figure 2-5: **Input and output file for inverse structure elucidation pipeline. A.** Screenshot of the input file in the .mgf format is shown. **B.** Screenshot of the output file generated through the MSNovelist pipeline is shown.

the MSNovelist pipeline used an in-built elemental composition determination model based on SIRIUS, and predicted structures for a wide range of chemical formulae with similar mass [53, 54].

2.3.2 Output file system and post-processing

The output file obtained from the MSNovelist pipeline was processed to remove duplicates and chemically infeasible molecules (Figure 2-5B). Duplicates were removed by converting all the predicted chemical structures in the SMILES format to canonical SMILES, and then using the *set* function in Python.

Canonical SMILES are analogous to the naming convention of a chemical structure, where any of the functional groups can, in principle, be placed at any position, or we could start from any of the branches in the structure, but the International Union of Pure and Applied Chemistry (IUPAC) convention dictates a certain naming convention [55]. Similarly, canonical SMILES are considered to be the IUPAC analogs for SMILES, since in this case also the atoms could be enumerated starting with different branches [56]. Canonical SMILES were generated using RDKit [57].

To filter the dataset and remove the infeasible molecules, we used RDKit to *sanitize* the molecules. Sanitization of molecules ensures that the molecules can be represented as octet-complete Lewis dot structures. Molecules that could not be *sanitized*, or denoted

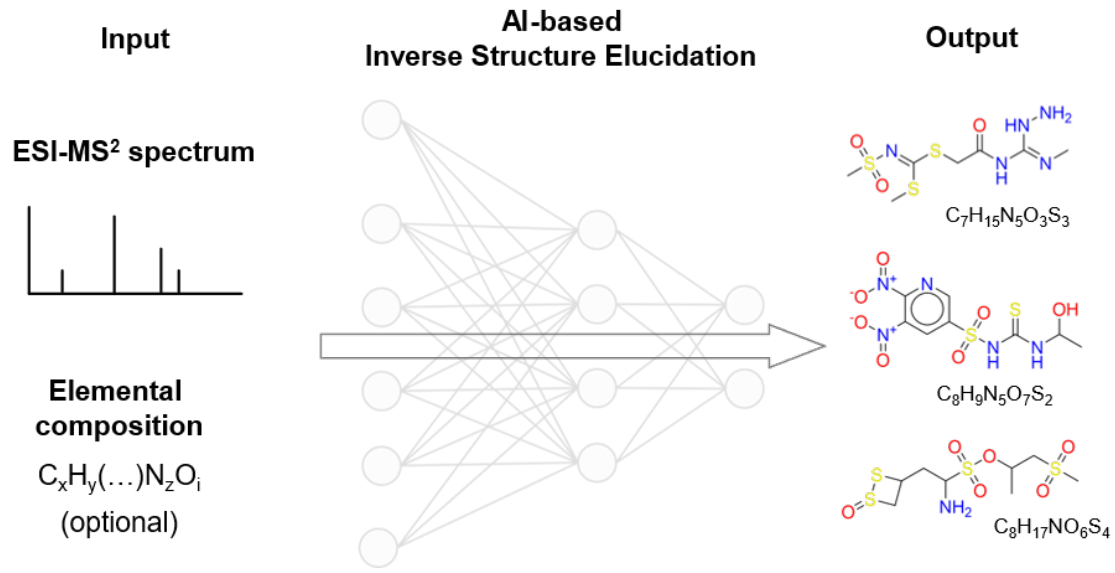

Figure 2-6: **Overview of the inverse structure elucidation model.** The input is the MS^2 spectra and optionally the elemental composition in a .mgf file, while the output is a set of structures with the different chemical formulae. The prediction model is adapted from MSNovelist.

as infeasible by RDKit, were removed from the final output.

2.3.3 Prediction model

Our prediction pipeline is based on the MSNovelist model, using the intermediate fingerprint generation, and the language model for translating into SMILES (Figure 2-6). The MSNovelist model architecture processes the spectra into CSI: FingerID representation. This representation is a common format for converting spectra into a more informative fingerprint based on potential substructures in the spectra [58]. This approach has been used to search for the specific structure, based on the spectra, through large mass spectrometry databases, with varying degrees of success [59, 60, 61].

In the MSNovelist study, Zamboni and co-workers use a long short-term memory (LSTM) language model to translate the spectra to molecules directly, with a computational complexity of $\mathcal{O}(1)$, instead of the database searching at complexity of $\mathcal{O}(n)$ [62]. With LSTM models having been used to reliably write sequences for several natural language processing, chemistry, and biology tasks [63, 64], and the training of the model using MassBank, a dataset of biomolecules, we used the model as-is.

Chapter 3

Applications

3.1 Impurity prediction for small molecule drugs

We evaluated the impurity prediction pipeline for two commonly used drugs - paracetamol and ibuprofen [65, 66]. N-(4-hydroxyphenyl)acetamide, paracetamol or acetaminophen is one of the most widely used over-the-counter drugs in the world, with kilotons being manufactured every year [67]. (RS)-2-(4-(2-Methylpropyl)phenyl)propanoic acid, Advil or ibuprofen is another commonly used over-the-counter drug, listed as one of the essential medicines on the World Health Organization list [68, 69, 70]. These examples were chosen based on their simplicity, and being publishable as a part of the thesis, without any legal constraints.

3.1.1 Paracetamol synthesis

We used the synthetic route of paracetamol from phenol, as outlined in [66]. The reaction has 3 distinct steps -

- **R1:** Nitration of phenol to para-nitrophenol using HNO_3
- **R2:** H_2 reduction of para-nitrophenol into para-aminophenol
- **R3:** Acetylation of para-aminophenol into paracetamol

Main process impurities

The impurity prediction pipeline was evaluated for paracetamol synthesis using the default parameters, as discussed in Section 2.2.3 (Figure 3-1A). Step impurities were predicted for **R1**, **R2**, and **R3**. These impurities *mostly* align with the impurities that are expected as per heuristically determined impurities in these reactions (Figure 3-1B, C, E).

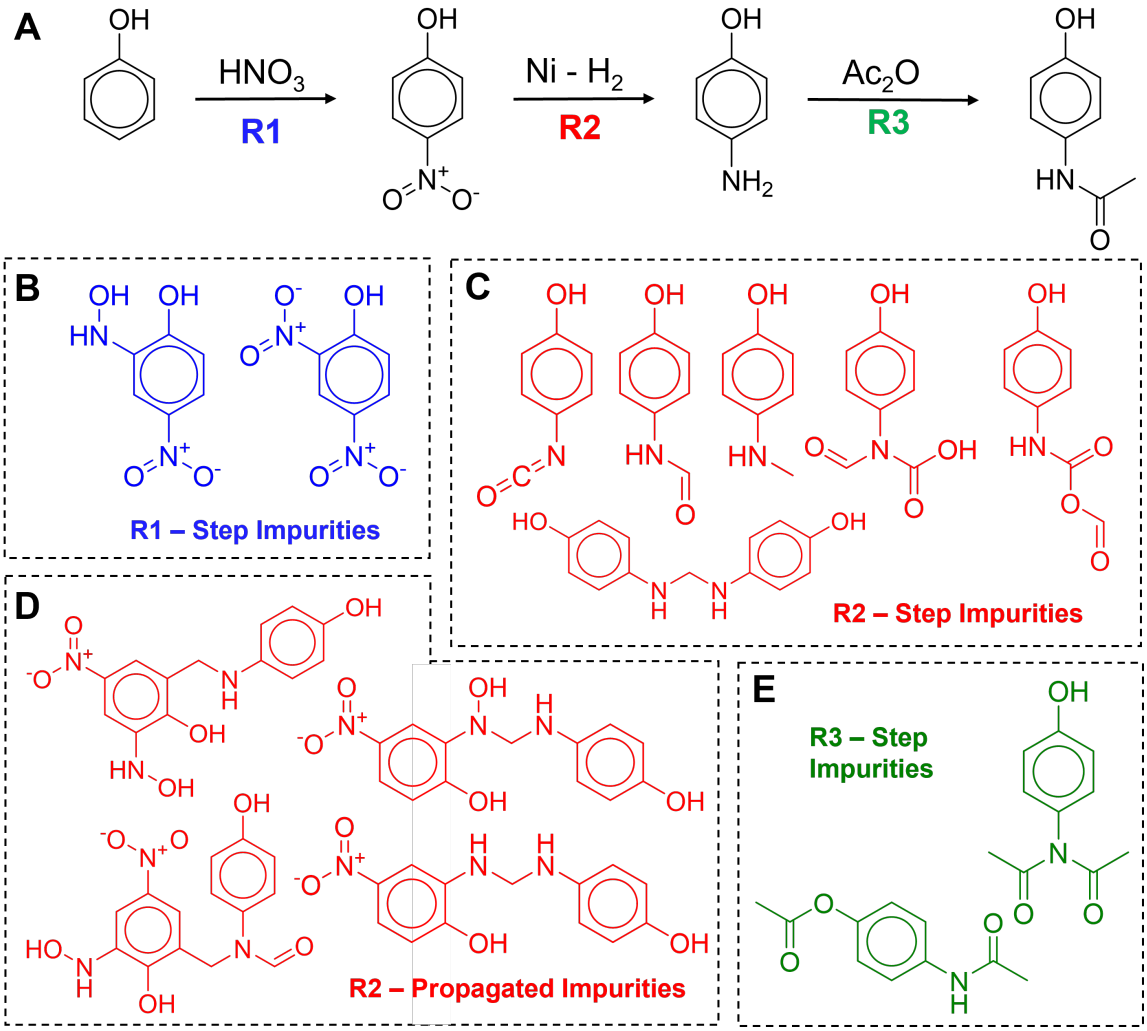

Figure 3-1: **Predicted step impurities and propagated impurities in paracetamol synthesis. A.** Reaction scheme for paracetamol synthesis. The specific steps have been marked as **R1**, **R2**, and **R3**, to specify the corresponding step and propagated impurities. Impurities from the reaction steps - **B. R1**, **C. R2**, and **E. R3**. **D.** Propagation of step impurities predicted in **R1** through **R2**. No propagation was predicted for step impurities in **R2**, and propagated impurities from **R1** were not propagated through into **R2**.

A key observation was the missing ortho-nitrophenol in **R1** step impurities. The pres-

ence of this impurity has been reported across multiple works [66, 71, 72]. It is surprising that the ASKCOS impurity module could not predict this impurity. We attributed this absence to the training dataset bias, with most reactions in the literature reporting the majority products in the chemical scheme, while impurities and side-products are moved to the text, thus not forming a part of the model training.

Impurity propagation was done for a single step, i.e. step impurities, arising from side reactions and otherwise, were propagated to the next step only, as per the default parameters. In the case of paracetamol synthesis, we only observed the propagation of impurities from the first step, **R1**, under the reaction conditions of **R2** (Figure 3-1D). It can be observed that the step impurities have a number of combinatorial choices as they move through the H_2 reduction, followed by dimerization, thereby resulting in a large number of dimers in the propagated impurities.

In addition to the impurity prediction, we only visualized unique impurities. In the order of precedence, we followed the step, propagated, and then impurities arising from reactions with known impurities. In this manner, the impurities that are already present as impurities from side reactions, but also generated as a result of impurity propagation, were not shown. The *summary* and *enlisted* files, however, show the complete list without any deduplication to provide the experimentalists with a complete picture of impurities and their respective sources.

Impurities arising from known impurities in starting materials

We added known impurities in the starting materials to see if there were any new impurities that were being produced. Using a similar impurity deduplication step, as earlier, we found one new impurity that was being produced in **R2** (Figure 3-2A). We observed no new impurities coming from reactions with trace metal impurities (Hg, As, Pb, Cd) in nitric acid (HNO_3), a single impurity from reactions with ammonia (NH_3), formaldehyde (HCHO), formic acid (HCOOH), carbon monoxide (CO), and carbon dioxide (CO_2) in hydrogen gas (H_2), and no new impurity formed by reaction with acetic acid (CH_3COOH) impurity in acetic anhydride (Ac_2O). The impurities in the specific reagents were obtained from the literature [73, 74, 75]. In this example, no additional impurities, such as those

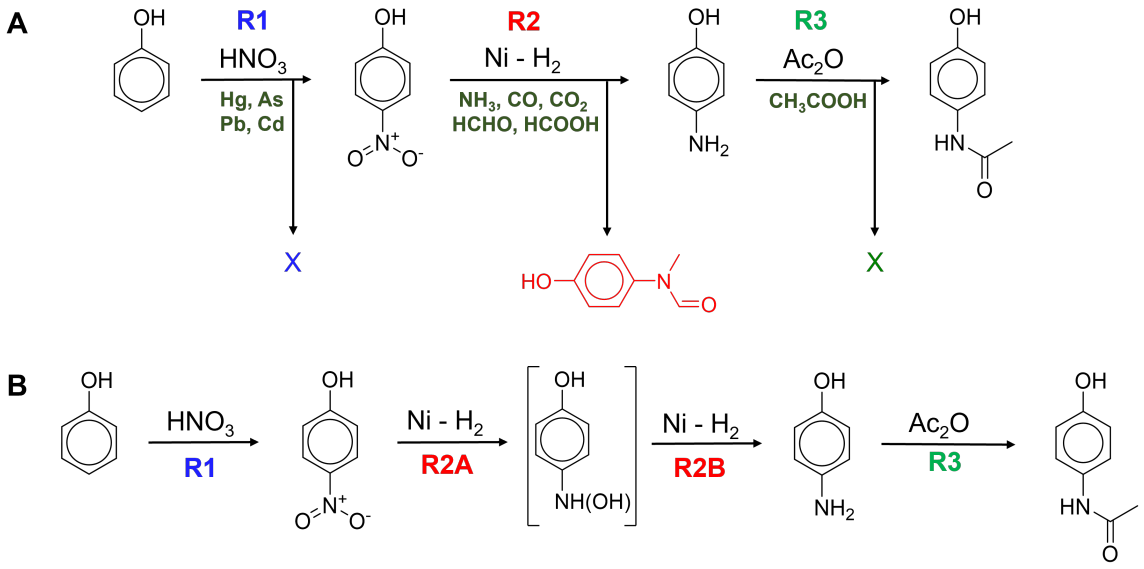

Figure 3-2: **Additional impurities in paracetamol synthesis. A.** Impurities arising from reactions with known impurities, noted in green, along with the reagents. Only one impurity is identified for **R2**, across the three reaction steps. **B.** Splitting of **R2**, into **R2A** and **R2B** to force a reaction intermediate did not result in any additional impurity identification.

arising from different work-up steps were added.

Impurities arising from reaction intermediates

In order to extend the capabilities of the impurity prediction and bias the predictor to find impurities from reaction intermediates, we introduced a reaction intermediate in the paracetamol synthesis route. Specifically, we split **R2**, H_2 reduction process, into **R2A** and **R2B**. **R2A** results in the production of para-hydroxoaminophenol from para-nitrophenol, and **R2B** completes the reaction as earlier leading to the formation of para-aminophenol. However, we did not observe the production of any new impurities due to the reaction intermediate. We believe that this might be related to the reaction intermediate case not being a part of the training datasets, and thus, the predictor not having any explicit idea to predict the impurities. However, it is supposed that the introduction of intermediates in different reaction routes might result in interesting results.

3.1.2 Ibuprofen synthesis

The 3-step route, outlined in [76], was used for the impurity prediction of ibuprofen synthesis (Figure 3-3A) -

- **R1:** Acetylation, in the presence of HF, of isobutylbenzene to 1-(4-isobutylphenyl) ethanone
- **R2:** H_2 reduction of 1-(4-isobutylphenyl)ethanone into 1-(4-isobutylphenyl) ethanol
- **R3:** Carbonylation of 1-(4-isobutylphenyl) ethanol into ibuprofen

Main process impurities

Several step impurities were identified for **R1**, **R2**, and **R3** (Figures 3-3B, C, E). Six impurities were noted in **R1**, with only three-step impurities each being predicted for **R2** and **R3**.

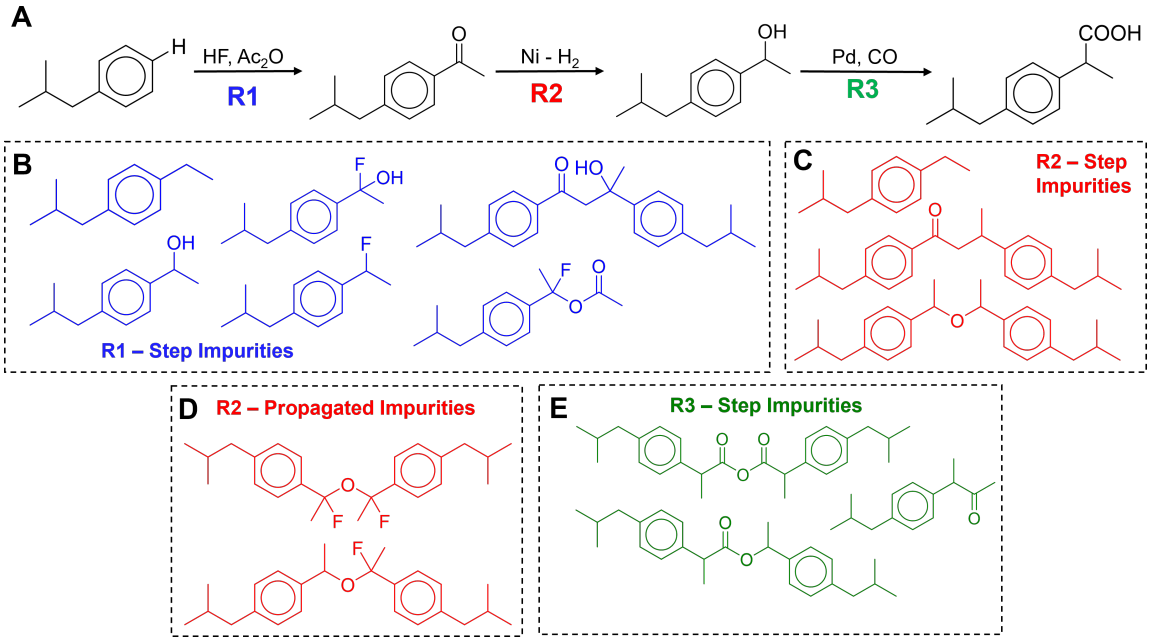

Figure 3-3: **Predicted step impurities and propagated impurities in ibuprofen synthesis. A.** Reaction scheme for ibuprofen synthesis. The specific steps have been marked as **R1**, **R2**, and **R3**, to specify the corresponding step and propagated impurities. Impurities from the reaction steps - **B. R1**, **C. R2**, and **E. R3**. **D.** Propagation of step impurities predicted in **R1** through **R2**. No propagation was predicted for step impurities in **R2**, and propagated impurities from **R1** were not propagated through into **R2**.

In **R1**, fluorination of the impurities due to the reaction of the produced impurities with HF, probably as a result of over-reaction, was identified as a key process for the generation of multiple impurities. Interestingly, 1-(4-isobutylphenyl) ethanol, which is the intended major product of the following step, **R2**, was noted as one of the impurities of **R1**. Also, only one dimerized impurity was observed in the first reaction step.

For **R2**, amongst the impurities, we noted a recurrence of 1-ethyl-4-isobutylbenzene, which was also noted as an impurity in **R1**. Such impurities, occurring in multiple steps suggest the need for the development of purification catered to similar or same impurities in each step and underscore the need for unit operation in each step where the impurities occur. **R3** produced a major impurity - 3-(4-isobutylphenyl)butan-2-one, and its dimerized impurities.

Propagation of impurities from **R1** to **R2** resulted in several dimerized and fluorinated impurities. However, no impurities were found when we attempted to propagate the **R2** impurities to **R3**.

The occurrence of such a low number of impurities in the ibuprofen synthesis shows the amount of optimization that has gone into this synthetic route, from the original synthesis route developed in the 1960s by the Boots group [69, 77, 76].

Impurities arising from known impurities in starting materials

With the addition of known impurities in the starting materials, several previously unobserved impurities were predicted for ibuprofen synthesis (Figure 3-4). In **R1**, CH_3COOH and H_2O are known to be present in acetic anhydride, and hydrofluoric acid, respectively [74, 78]. Their presence resulted in two new impurities. For **R2**, we noted multiple carbonylated and oxidized impurities with the presence of CO, CO_2, HCHO, and HCOOH in the H_2 gas. In R**3**, palladium (Pd) and CO contained several trace metal impurities and several other common gaseous impurities, leading to the production of several low-level impurities [79].

The multiple impurities noted in ibuprofen synthesis are substantially more than those noted in the case of paracetamol synthesis, despite having similar reagents. Our observation further highlights the need for raw materials risk assessments in the context

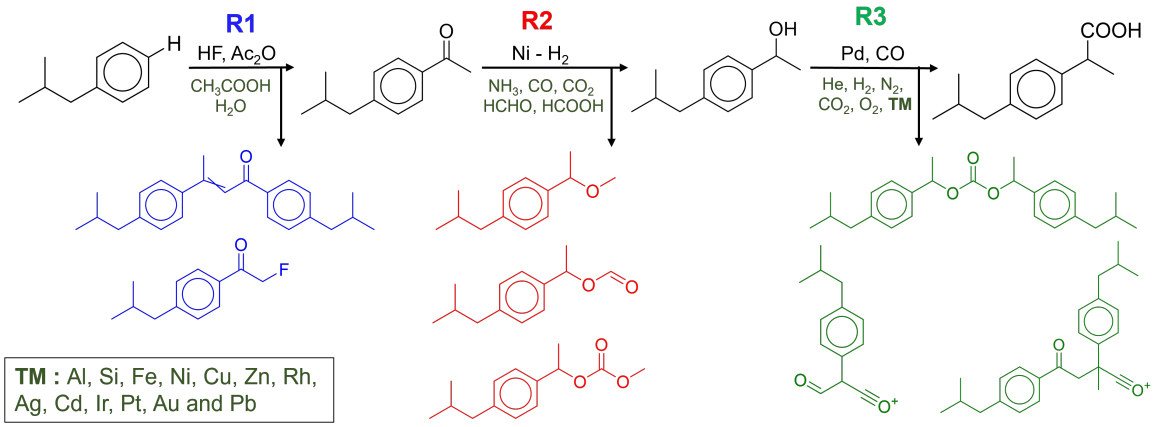

Figure 3-4: **Additional impurities in ibuprofen synthesis.** Impurities arising from reactions with known impurities, noted in green, along with the reagents, for each of the three reaction steps - **R1**, **R2**, and **R3**.

of individual reactions, and evaluation of what reaction products can result from known impurities in starting materials.

3.1.3 Application at Amgen

We used the impurity prediction pipeline to obtain an impurity map, with *summary* and *enlisted* modes, for a critical Amgen small molecule. Based on the conversations with experimentalists leading the project, the model was able to identify key impurities in different steps.

3.2 Inverse structure elucidation from MS^2 data

We used two random spectra from the public dataset released by Graham Cooks lab to evaluate our inverse structure elucidation pipeline (Figure 3-5) [80]. In these two examples, we noted several structures, with different elemental compositions but similar molecular weight, being predicted using the inverse structure elucidation pipeline, as described in Section 2.3. When the spectra were run in the mode without providing any elemental composition, it was noted that the model was biased toward finding structures of lower molecular masses. The above examples demonstrated the utility of the tool, providing evidence of how an unknown electrospray ionization (ESI) mass spectrometry

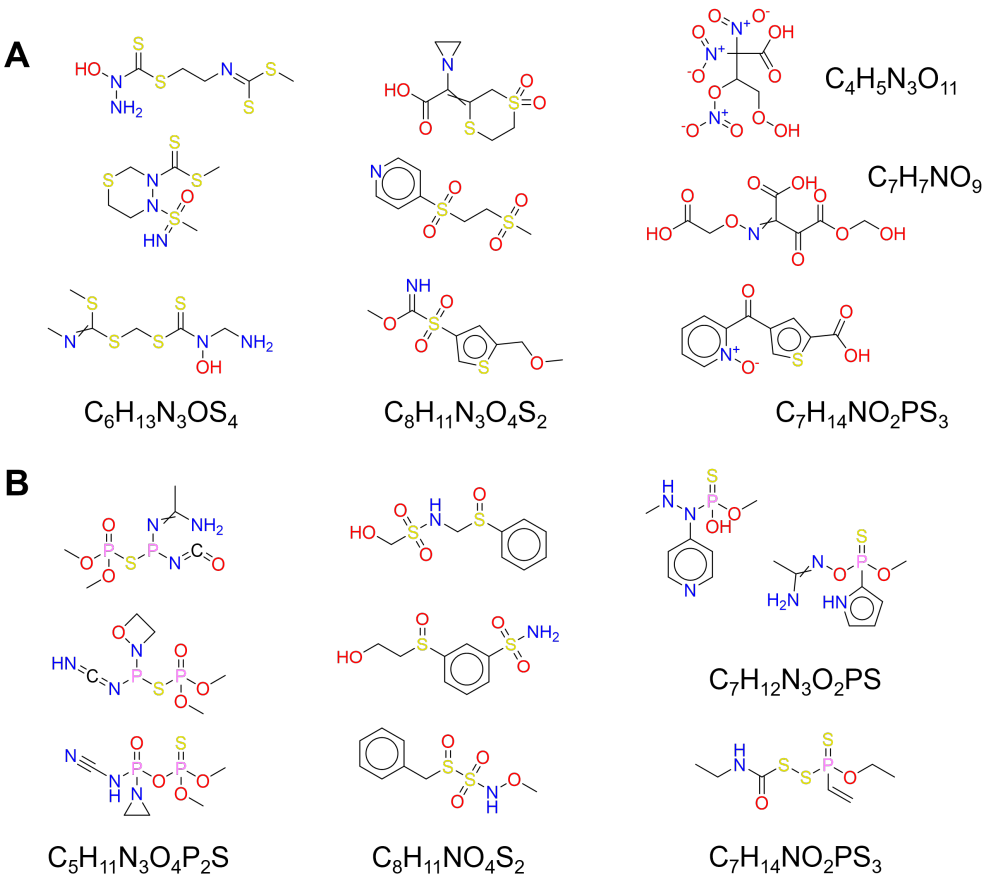

Figure 3-5: **Chemical structures predicted for two random spectra using the inverse structure elucidation pipeline.** Different structures corresponding to the elemental formula are noted for the two respective spectra in **A.** and **B.**

data can be translated into chemical structures of varying elemental composition.

3.2.1 Application at Amgen

The tool was applied to two current, and real-world, unknown impurity identification tasks at Amgen. At the time of this thesis, the impurities had not been yet been positively identified. However, the tool proposed novel chemical structures that, by the review of subject matter experts, were judged to be more likely than the structures proposed earlier by chemists on the project and thus provided directions to pursue the impurity identification tasks.

Chapter 4

Limitations, and Future Directions

4.1 Limitations

We have noted several limitations for our impurity prediction, and inverse structure elucidation pipelines, and have discussed them here. The limitations of our impurity prediction approach are intrinsically linked to the training datasets, the accuracy of the forward prediction model, and known impurities in the starting materials. Similar to the impurity prediction pipeline, the inverse structure elucidation pipeline is also limited by the training data and the coverage of chemical space.

The training datasets act as a limitation in the prediction pipelines. For impurity prediction, the datasets, both from Reaxys [23] and Pistachio [40], mostly comprise reactions with major products. In these datasets, side products, and minor impurities, are not mentioned as a part of the scheme. This information is occasionally discussed within the manuscript text or moved to supplementary information. Thus, obtaining a training dataset focused on additional products remains a significant challenge, limiting the prediction of impurities for new reactions. Similarly, the aggregated datasets of electrospray ionization mass spectrometry do not have quality constraints, and have been collected from multiple sources for different biological molecules [38, 39]. Although they are supposed to cover a wide range of chemical space, the similarity of metabolites or biological molecules to specific drug-like compounds needs to be assessed on a case-by-case basis. Thus, the performance of the pipeline, as-is and with our improvements, is

limited by how the initial models were trained.

For impurity predictions, we are limited by the accuracy of the forward reaction prediction model being used. In our case, using ASKCOS, we are implicitly limited by how well the prediction model functions [81]. Since ASKCOS is focused on predicting major products alone, we tried to extend the predictions by increasing probability thresholds and top-n predictions. However, these changes result in filtering out predicted molecules, showing a larger number of molecules, but not changing the predictions themselves. Thus, if ASKCOS was not able to predict an impurity amongst the tens of molecules, then it would not be able to do so unless the training dataset is improved. Further, the ranking of products for the impurity module in ASKCOS is based on structural similarity to the major product. This metric might not be an accurate depiction of a lot of impurities. A number of hard-to-identify impurities arise from reactions with impurities in the starting materials. With limited information about the low-level impurities in the starting materials, it is nearly impossible to predict such impurities.

4.2 Future Directions

In the future, the impurity prediction model may be improved using a model ensemble approach. Different model architectures, such as the transformer-based models in IBM-RXN [82], where the model learns to map different character tokens for reaction prediction, can help in obtaining a larger and potentially more diverse set of impurities than the graph-based approach in ASKCOS. In addition to that, novel architectures that combine the attention mechanism of transformer models with the inherent graph representations of molecules, such as those presented by Mao and co-workers [83], and Tu and Coley [84], can be adapted for impurity prediction. The key consideration here is using a similar training dataset, assuming that a large-scale dataset is infeasible in the short term, and leveraging different model architectures to identify different products.

Rank ordering of impurities, other than by similarity to the major product, needs to be addressed in future works. Prediction of reaction kinetics and/or activation energy barriers might be potential directions [85, 86, 87, 88]. These approaches will provide

quantitative insights on the likelihood of the presence of the impurities, and help in filtering down by chemistry-informed methods.

Ultimately, to capture the low-level impurities, a dataset of safety data sheets for known starting materials needs to be built [73]. This dataset will help in listing out known impurities, and identify reaction products *a priori* saving significant time and resources that might have to be invested in impurity identification.

For the inverse structure elucidation pipeline, datasets focused on drug-like molecules can be built, or aggregated from existing experimental sources, to improve the quality of the predictions. The model architecture can be improved by using more recent translation model architectures, such as transformers [89].

Apart from the improvements to the models themselves, it is important to focus on the user experience as a future direction. The cycle time to incorporate the changes to the different modules in the software suite, and the inclusion of different types of chemical reactions, such as enzyme-mediated reactions, need to be considered to increase the usability of the models in the industry setting.

THIS PAGE INTENTIONALLY LEFT BLANK

Chapter 5

Conclusion

In this thesis, we developed tools to automate impurity identification, by predicting a plausible set of impurities using an impurity predictor, and then proposing to use an inverse structure elucidation function on analytical data from experiments to down-select impurities.

Impurity prediction was done by developing an AI-based reaction predictor, that takes as input the main reactants but also reagents, solvents, and impurities in these materials. Further, the predictor can be run iteratively to track impurity propagation in multi-step reactions. For inverse structure elucidation, a chemistry-informed language model was developed to translate mass spectrometry data to potential molecular structures, which can then be checked for matches against the predicted chemical reaction products.

The impurity prediction tool was applied to the synthesis of common small molecule drugs – paracetamol and ibuprofen, and the inverse structure elucidation tool was used for the identification of chemical structures from publicly available electrospray ionization mass spectrometry data, The models were applied to proprietary Amgen programs, both small molecule drugs, and biologics, with significant results noted in both projects.

The tool can have an impact on a number of important activities in chemical process development for synthetic drug substances, including (1) identification of impurities, (2) high-throughput reaction screening and detailed reaction kinetic analysis, and (3) raw materials risk assessments. 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 and detailed reaction kinetic analysis, the tool provides a route to automate impurity identification 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.

Bibliography

[1] *About | Amgen*. https://www.amgen.com/about. (Accessed on 10/24/2022).

[2] *Responsibility | Amgen*. https://www.amgen.com/responsibility. (Accessed on 10/24/2022).

[3] *The Drug Development Process | FDA*. https://www.fda.gov/patients/learn-about-drug-and-device-approvals/drug-development-process. (Accessed on 10/24/2022).

[4] Noah J Wichrowski et al. "An overview of drug substance manufacturing processes". In: *AAPS PharmSciTech* 21.7 (2020), pp. 1–6.

[5] Miriam Sarkis et al. "Emerging challenges and opportunities in pharmaceutical manufacturing and distribution". In: *Processes* 9.3 (2021), p. 457.

[6] W Price, II Nicholson, and Arti K Rai. "Manufacturing barriers to biologics competition and innovation". In: *Iowa L. Rev.* 101 (2015), p. 1023.

[7] Jared J Nathan, Monica Ramchandani, and Primal Kaur. "Manufacturing of biologics". In: *Biologic and Systemic Agents in Dermatology*. Springer, 2018, pp. 101–110.

[8] Christopher L Burcham, Alastair J Florence, and Martin D Johnson. "Continuous manufacturing in pharmaceutical process development and manufacturing". In: *Annual review of chemical and biomolecular engineering* 9 (2018), pp. 253–281.

[9] Arno Behr et al. "New developments in chemical engineering for the production of drug substances". In: *Engineering in Life Sciences* 4.1 (2004), pp. 15–24.

[10] Madhav Shelke, Shirish Shriram Deshpande, and Sanjay Sharma. "Quinquennial review of progress in degradation studies and impurity profiling: an instrumental perspective statistics". In: *Critical Reviews in Analytical Chemistry* 50.3 (2020), pp. 226–253.

[11] Kelly Zhang et al. "Reactive impurities in large and small molecule pharmaceutical excipients–A review". In: *TrAC Trends in Analytical Chemistry* 101 (2018), pp. 34–42.

[12] Robin Smith. *Chemical process: design and integration*. John Wiley & Sons, 2005.

[13] Sheun Oshinbolu et al. "Measurement of impurities to support process development and manufacture of biopharmaceuticals". In: *TrAC Trends in Analytical Chemistry* 101 (2018), pp. 120–128.

[14] David J Snodin. “A primer for pharmaceutical process development chemists and analysts in relation to impurities perceived to be mutagenic or “genotoxic””. In: *Organic Process Research & Development* 24.11 (2020), pp. 2407–2427.

[15] Duane A Pierson et al. “Approaches to assessment, testing decisions, and analytical determination of genotoxic impurities in drug substances”. In: *Organic Process Research & Development* 13.2 (2009), pp. 285–291.

[16] P Venkatesan and K Valliappan. “Impurity profiling: theory and practice”. In: *Journal of Pharmaceutical Sciences and Research* 6.7 (2014), p. 254.

[17] Sándor Görög. “Chemical and analytical characterization of related organic impurities in drugs”. In: *Analytical and bioanalytical chemistry* 377.5 (2003), pp. 852–862.

[18] Fenghe Qiu and Daniel L Norwood. “Identification of pharmaceutical impurities”. In: *Journal of liquid chromatography & related technologies* 30.5-7 (2007), pp. 877–935.

[19] Sara Szymkuć et al. “Computer-assisted synthetic planning: the end of the beginning”. In: *Angewandte Chemie International Edition* 55.20 (2016), pp. 5904–5937.

[20] Yuning Shen et al. “Automation and computer-assisted planning for chemical synthesis”. In: *Nature Reviews Methods Primers* 1.1 (2021), pp. 1–23.

[21] Connor W Coley et al. “A graph-convolutional neural network model for the prediction of chemical reactivity”. In: *Chemical science* 10.2 (2019), pp. 370–377.

[22] Adarsh Arun et al. “Reaction impurity prediction using a data mining approach”. In: *ChemRxiv* (2022).

[23] *Reaxys - An expert-curated chemistry database*. `https://www.elsevier.com/solutions/reaxys`. (Accessed on 10/26/2022).

[24] *NextMove Software | Pistachio*. `https://www.nextmovesoftware.com/pistachio.html`. (Accessed on 10/26/2022).

[25] Rubén M Maggio et al. “Pharmaceutical impurities and degradation products: Uses and applications of NMR techniques”. In: *Journal of pharmaceutical and biomedical analysis* 101 (2014), pp. 102–122.

[26] Karen M Alsante et al. “Pharmaceutical impurity identification: a case study using a multidisciplinary approach”. In: *Journal of Pharmaceutical Sciences* 93.9 (2004), pp. 2296–2309.

[27] Martin Sandvoss et al. “HPLC–SPE–NMR in pharmaceutical development: capabilities and applications”. In: *Magnetic Resonance in Chemistry* 43.9 (2005), pp. 762–770.

[28] András Darcsi, Ákos Rácz, and Szabolcs Béni. “Identification and characterization of a new dapoxetine impurity by NMR: Transformation of N-oxide by Cope elimination”. In: *Journal of Pharmaceutical and Biomedical Analysis* 134 (2017), pp. 187–194.

[29] Jaan A Pesti et al. “Commercial Synthesis of a Pyrrolotriazine–Fluoroindole Intermediate to Brivanib Alaninate: Process Development Directed toward Impurity Control”. In: *Organic Process Research & Development* 18.1 (2014), pp. 89–102.

[30] Eric A Standley et al. “Synthesis of Rovafovir Etalafenamide (Part I): Active Pharmaceutical Ingredient Process Development, Scale-Up, and Impurity Control Strategy”. In: *Organic Process Research & Development* 25.5 (2021), pp. 1215–1236.

[31] Peter K Dornan et al. “Continuous process improvement in the manufacture of carfilzomib, Part 1: Process understanding and improvements in the commercial route to prepare the epoxyketone warhead”. In: *Organic Process Research & Development* 24.4 (2020), pp. 481–489.

[32] Christopher J Borths et al. “Control of mutagenic impurities: Survey of pharmaceutical company practices and a proposed framework for industry alignment”. In: *Organic Process Research & Development* 25.4 (2021), pp. 831–837.

[33] Youngchun Kwon et al. “Molecular search by NMR spectrum based on evaluation of matching between spectrum and molecule”. In: *Scientific Reports* 11.1 (2021), pp. 1–9.

[34] Bhuvanesh Sridharan et al. “Spectra to Structure: Deep Reinforcement Learning for Molecular Inverse Problem”. In: (2021).

[35] Jinzhe Zhang et al. “NMR-TS: de novo molecule identification from NMR spectra”. In: *Science and technology of advanced materials* 21.1 (2020), pp. 552–561.

[36] Eleni Litsa et al. “Spec2Mol: An end-to-end deep learning framework for translating MS/MS Spectra to de-novo molecules”. In: (2021).

[37] Aditya Divyakant Shrivastava et al. “MassGenie: A transformer-based deep learning method for identifying small molecules from their mass spectra”. In: *Biomolecules* 11.12 (2021), p. 1793.

[38] Michael A Stravs et al. “MSNovelist: De novo structure generation from mass spectra”. In: *Nature Methods* (2022), pp. 1–6.

[39] Hisayuki Horai et al. “MassBank: a public repository for sharing mass spectral data for life sciences”. In: *Journal of mass spectrometry* 45.7 (2010), pp. 703–714.

[40] *chemdata:start []*. `https://chemdata.nist.gov/dokuwiki/doku.php?id=chemdata:start`. (Accessed on 10/26/2022).

[41] Florencio Zaragoza Dörwald. *Side reactions in Organic Synthesis: a guide to successful synthesis design*. John Wiley & Sons, 2006.

[42] Stephanie J Urwin et al. “A structured approach to cope with impurities during industrial crystallization development”. In: *Organic process research & development* 24.8 (2020), pp. 1443–1456.

[43] Paul A Meenan, Stephen R Anderson, and Diana L Klug. “The influence of impurities and solvents on crystallization”. In: *Handbook of industrial crystallization* (2002), pp. 67–100.

[44] Mark D Argentine, Paul K Owens, and Bernard A Olsen. “Strategies for the investigation and control of process-related impurities in drug substances”. In: *Advanced drug delivery reviews* 59.1 (2007), pp. 12–28.

[45] Brandon J Reizman et al. “Data-driven prediction of risk in drug substance starting materials”. In: *Organic Process Research & Development* 23.7 (2019), pp. 1429–1441.

[46] Amandine Dispas et al. “‘Quality by Design’approach for the analysis of impurities in pharmaceutical drug products and drug substances”. In: *TrAC Trends in Analytical Chemistry* 101 (2018), pp. 24–33.

[47] Thomas Zahel et al. “Integrated process modeling—a process validation life cycle companion”. In: *Bioengineering* 4.4 (2017), p. 86.

[48] *ChemDraw/ChemOffice v21.0 Release – PerkinElmer*. `https://informatics-support.perkinelmer.com/hc/en-us/articles/4586496010004-ChemDraw-ChemOffice-v21-0-Release`. (Accessed on 10/28/2022).

[49] David Weininger. “SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules”. In: *Journal of chemical information and computer sciences* 28.1 (1988), pp. 31–36.

[50] Kimberley R Cousins. *ChemDraw 6.0 Ultra CambridgeSoft Corporation, 100 Cambridge Park Drive, Cambridge, MA 02140. http://www. camsoft. com. Commercial Price: 1395.AcademicPrice: 699.* 2000.

[51] Amol Thakkar et al. “Datasets and their influence on the development of computer assisted synthesis planning tools in the pharmaceutical domain”. In: *Chemical science* 11.1 (2020), pp. 154–168.

[52] *Cloud Computing Services - Amazon Web Services (AWS)*. `https://aws.amazon.com/`. (Accessed on 10/28/2022).

[53] Kai Dührkop et al. “SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information”. In: *Nature methods* 16.4 (2019), pp. 299–302.

[54] Marcus Ludwig et al. “De novo molecular formula annotation and structure elucidation using SIRIUS 4”. In: *Computational Methods and Data Analysis for Metabolomics* (2020), pp. 185–207.

[55] Stanislaw Skonieczny. “The IUPAC rules for naming organic molecules”. In: *Journal of chemical education* 83.11 (2006), p. 1633.

[56] David Weininger, Arthur Weininger, and Joseph L Weininger. “SMILES. 2. Algorithm for generation of unique SMILES notation”. In: *Journal of chemical information and computer sciences* 29.2 (1989), pp. 97–101.

[57] Greg Landrum. “Rdkit documentation”. In: *Release* 1.1-79 (2013), p. 4.

[58] Kai Dührkop et al. “Searching molecular structure databases with tandem mass spectra using CSI: FingerID”. In: *Proceedings of the National Academy of Sciences* 112.41 (2015), pp. 12580–12585.

[59] Yuanyue Li et al. “Identification of metabolites from tandem mass spectra with a machine learning approach utilizing structural features”. In: *Bioinformatics* 36.4 (2020), pp. 1213–1218.

[60] Marcus Ludwig, Kai Dührkop, and Sebastian Böcker. “Bayesian networks for mass spectrometric metabolite identification via molecular fingerprints”. In: *Bioinformatics* 34.13 (2018), pp. i333–i340.

[61] Sebastian Böcker. “Searching molecular structure databases using tandem MS data: are we there yet?” In: *Current Opinion in Chemical Biology* 36 (2017), pp. 1–6.

[62] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. “Learning to forget: Continual prediction with LSTM”. In: *Neural computation* 12.10 (2000), pp. 2451–2471.

[63] Yong Yu et al. “A review of recurrent neural networks: LSTM cells and network architectures”. In: *Neural computation* 31.7 (2019), pp. 1235–1270.

[64] Carly K Schissel et al. “Deep learning to design nuclear-targeting abiotic miniproteins”. In: *Nature chemistry* 13.10 (2021), pp. 992–1000.

[65] Peter William Atkins. *Atkins' molecules*. Cambridge University Press, 2003.

[66] Bryan G Reuben and Harold A Wittcoff. *Pharmaceutical chemicals in perspective*. Wiley-Interscience, 1989.

[67] Jonas W Wastesson et al. “Trends in use of paracetamol in the Nordic countries”. In: *Basic & Clinical Pharmacology & Toxicology* 123.3 (2018), pp. 301–307.

[68] THOMAS G KANTOR. “Ibuprofen”. In: *Annals of Internal Medicine* 91.6 (1979), pp. 877–882.

[69] KD Rainsford. “Ibuprofen: pharmacology, efficacy and safety”. In: *Inflammopharmacology* 17.6 (2009), pp. 275–342.

[70] World Health Organization et al. “WHO model list of essential medicines for children: 6th list (March 2017, amended August 2017)”. In: (2017).

[71] Roxan Joncour et al. “Amidation of phenol derivatives: a direct synthesis of paracetamol (acetaminophen) from hydroquinone”. In: *Green chemistry* 16.6 (2014), pp. 2997–3002.

[72] Irshad Maajid Taily, Debarshi Saha, and Prabal Banerjee. “Direct Synthesis of Paracetamol via Site-Selective Electrochemical Ritter-type C–H Amination of Phenol”. In: *Organic Letters* 24.12 (2022), pp. 2310–2314.

[73] Robert E Lenga. *The Sigma-Aldrich library of chemical safety data*. Sigma-Aldrich Corp., 1988.

[74] WS Calcott, FL English, and OC Wilbur. “Analysis of Acetic Anhydride.” In: *Industrial & Engineering Chemistry* 17.9 (1925), pp. 942–944.

[75] Claire Beurey et al. “Review and survey of methods for analysis of impurities in hydrogen for fuel cell vehicles according to ISO 14687: 2019”. In: *Frontiers in Energy Research* 8 (2021), p. 615149.

[76] MC Cann and ME Connelly. "The BHC Company Synthesis of Ibuprofen, a Greener Synthesis of Ibuprofen Which Creates Less Waste and Fewer Byproducts, chapter from "Real World Cases in Green Chemistry"". In: *Publisher American Chemical Society, Washington DC* (2000), pp. 19–24.

[77] S Bashyal. "Ibuprofen and its different analytical and manufacturing methods: a review". In: *Asian J. Pharm. Clin. Res* 11 (2018), pp. 25–29.

[78] Sigma-Aldrich Safety Data Sheet. *Hydrofluoric acid*. 2020.

[79] Jonathan Edward McIntyre, Kofi Nhyira Colecraft, and Yuhan Yang. "Palladium On Carbon Catalyst". In: (2018).

[80] MyPhuong T Le et al. "Fragmentation of polyfunctional compounds recorded using automated high-throughput desorption electrospray ionization". In: *Journal of the American Society for Mass Spectrometry* 32.8 (2021), pp. 2261–2273.

[81] *ASKCOS Homepage*. `https://askcos.mit.edu/`. (Accessed on 11/09/2022).

[82] Philippe Schwaller et al. "Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction". In: *ACS central science* 5.9 (2019), pp. 1572–1583.

[83] Yu Rong et al. "Self-supervised graph transformer on large-scale molecular data". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 12559–12571.

[84] Zhengkai Tu and Connor W Coley. "Permutation invariant graph-to-sequence model for template-free retrosynthesis and reaction prediction". In: *Journal of chemical information and modeling* 62.15 (2022), pp. 3503–3513.

[85] Charles T Campbell and Zhongtian Mao. "Analysis and prediction of reaction kinetics using the degree of rate control". In: *Journal of Catalysis* 404 (2021), pp. 647–660.

[86] Kevin A Spiekermann, Lagnajit Pattanaik, and William H Green. "Fast Predictions of Reaction Barrier Heights: Toward Coupled-Cluster Accuracy". In: *The Journal of Physical Chemistry A* 126.25 (2022), pp. 3976–3986.

[87] Kevin Spiekermann, Lagnajit Pattanaik, and William H Green. "High accuracy barrier heights, enthalpies, and rate coefficients for chemical reactions". In: *Scientific Data* 9.1 (2022), pp. 1–12.

[88] Colin A Grambow, Lagnajit Pattanaik, and William H Green. "Deep learning of activation energies". In: *The journal of physical chemistry letters* 11.8 (2020), pp. 2992–2997.

[89] Tianyang Lin et al. "A survey of transformers". In: *AI Open* (2022).

Appendix A

List of Interviewees

Table A.1: List of people at Amgen organizations with whom the project and approach was discussed. Abbreviations: DS - Drug Substance; PCS - Pivotal and Commercial Synthetics; ODSC - Operations Digital Strategy and Capabilities; DIPT - Digital Integration and Predictive Technologies; IS - Information Systems; Tech - Technology; DP - Drug Product.

Sl No	Name	Organization	Role
1	Rahul Sangodakar	DS PCS	Principal Engineer
2	Ning Yang	Attribute Sciences	Sr Scientist
3	Nandini Sarkar	Attribute Sciences	Scientist
4	Andrew Cosbie	DS PCS	Principal Engineer
5	Erin Shen	DS PCS	Data Scientist
6	Carolyn Wei	DS PCS	Principal Scientist
7	Chris Garvin	ODSC	Director
8	Fabrice Schlegel	DIPT	Director
9	Pablo Rolandi	DIPT	Executive Director
10	Neil Langille	DS PCS	Principal Scientist
11	Heather Nunn	DS Biologics	Director
12	Matthew Beaver	DS PCS	Principal Scientist
13	James Murray	DS PCS	Sr Scientist
14	Liang Zhang	DS PCS	Sr Scientist
15	Andrew Parsons	DS PCS	Principal Scientist
16	Matt Porter-Racine	DIPT	Sr Manager
17	JoaoBerto	IS	SWE
18	Sean Cole	DS Biologics	Sr Scientist
19	Tom Mistretta	DS Tech	Director
20	James Fisher	DS Tech	Director

Sl No	Name	Organization	Role
21	Barry Pearson	Commercial DP	Director
22	Simone Spada	DS PCS	Principal Engineer
23	Sinem Oruklu	ODSC	Sr Data Scientist
24	Amlan Das	DS Biologics	Director
25	Jessica Virani	ODSC	Data Scientist
26	Muhammad Irfan	DS PCS	Engineer
27	Prashant Agrawal	Drug Product	Sr Ass Data Scientist
28	Neeraj Agarawal	Attribute Sciences	Director
29	Kathleen Rand	ODSC	Sr Data Scientist
30	Tim Lauer	Attribute Sciences	Sr Data Scientist
31	Susan Burke	Attribute Sciences	Director
32	Yong Xie	Attribute Sciences	Director
33	Zhongqi Zhang	Attribute Sciences	Principal Scientist
34	Natalia Gomez	DS	Director
35	Andy Clyne	Manufacturing	Sr Manager
36	Seb Caille	DS PCS	Director
37	Mike Lovette	DS	Sr Manager
38	Darren Reid	Drug Product	Director
39	Bharath Venkataram	DS PCS	Sr Engineer
40	Elcin Icten	DIPT	Sr Manager
41	Cenk Undey	ODSC	Executive Director
42	Ketan Kumar	Attribute Sciences	Sr Manager
43	Aik Jun Tan	Manufacturing	Project Manager
44	Seyma Bayrak	Quality	Director
45	Pavel Bondarenko	Attribute Sciences	Director
46	Saleh Alkhalifa	ODSC	Sr Data Scientist
47	Fides Lay	DS	Sr Scientist
48	Jasmine Tat	DS	Sr Ass Data Scientist
49	Dan Gschwend	ODSC	Director
50	Dan Greene	Drug Product	Scientist
51	Jian Wu	DS Biologics	Sr Principal Scientist
52	Michaela Murr	ODSC	Co-op
53	Aditya Tulsyan	ODSC	Sr Manager
54	Tony Wang	ODSC	Sr Manager
55	Laura Blue	Attribute Sciences	Director
56	Jonathan Truong	DS PCS	Associate Scientist
57	Adam Simon	Research	Sr Scientist
58	Dan Fallon	Analytics	Data Scientist
59	James Rider	ODSC	Executive Director
60	Philipp Simons	ODSC	Manager