AI-assisted chemical reaction impurity prediction and propagation

Anonymous Author(s)
Affiliation
Address
email

Abstract

Most chemical reactions result in numerous by-products and side-products, apart from 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 systematically predict and track impurities derived from raw materials or those that have propagated from one synthetic step to the next. In this study, we developed an AI-assisted approach to predict and track impurities across multi-step reactions using the main reactants, and optionally reagents, solvents and impurities in these materials, as input. We demonstrated the utility of this tool for a simple case of synthesis of paracetamol from phenol, and provide a generalized framework that covers most chemical reactions. Our solution can be applied to enable (1) faster elucidation of impurities, (2) automated interpretation of data generated from high-throughput reaction screening, and (3) more thorough raw materials risk assessments, with each of these representing key workflows in small molecule drug substance commercial process development.

1 Introduction

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 substance [1]. In conducting chemical process development, *a priori* prediction of possible impurities would significantly accelerate the route selection and optimization processes [2]. For example, this knowledge would help in eliminating synthetic routes that might lead to the production of potential genotoxic and mutagenic impurities [3, 4].

However, computational impurity prediction done today, such as those presented by Coley and co-workers as a part of the ASKCOS software suite [5] and the Python-based workflow presented in Arun et al. [6], focus on primary reactants and aim to *selectively* predict likely impurities for a single reaction. As such, these approaches may miss some complexities seen in the real world, specifically around impurities propagated from multi-step syntheses or impurities derived from minor components in reagents and solvents, which are rarely perfectly pure. In an effort to be selective and only predict the most likely impurities to be observed, the aforementioned approaches may miss impurities that result from subsequent reactions or overreactions of formed impurities with other components of a complex reaction mixture. As a result, the predictions from existing impurity predictors frequently miss the low-level impurities that are critical to identify and control in drug substance development in the pharmaceutical sector.

On the other hand, current experimental approaches for impurity identification involves analysis of reaction characterization data, usually, liquid chromatography-mass spectrometry data (LC-MS) [7, 8, 9, 10]. The *post hoc* analysis necessitates the involvement of subject matter experts from process chemistry, mass spectrometry, and other fields to identify the impurities. This impurity

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

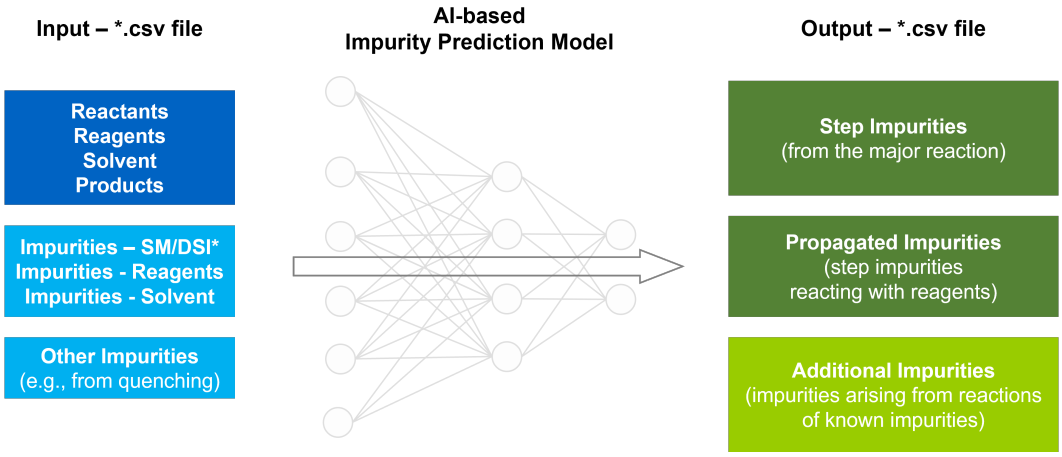

Figure 1: **Overview of the AI-assisted impurity prediction approach.** The tool takes in input in the form of a *.csv file, with the **reactants, reagents, solvent and products**, and **known impurities** in the aforementioned materials, processes the individual reactions using the AI model, and outputs a *.csv file with the **key impurities** and **additional impurities**. Abbreviations: SM – Starting Material, DSI – Drug Substance Impurity.

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

We propose a closed-loop approach combining AI-assisted plausible impurity prediction with inverse structure elucidation from analytical data for automated impurity identification. Leveraging current forward reaction predictors, we aim to obtain an *inclusive* set of impurities that might result from a set of reactions, involving the primary reactants, reagents, solvent, and known impurities therein. The inverse structure elucidation model will help in identifying structures from the analytical data, for instance, identifying molecules from MS^2 spectra [11]. Re-framing the impurity prediction problem with the goal of predicting an inclusive set allows us to expand the inputs and relax the selectivity in predicting low-level impurities, such as those from subsequent reactions of primary process impurities. This approach comes at the cost of producing a large set of potential impurities, many of which will not be observed in reality. Therefore, the impurity prediction strategy needs to be coupled with an inverse structure elucidation model as a selection function to down-select molecules in the large inclusive set.

In this work, we focus on the first part, and discuss the prediction of a plausible or inclusive set of impurities for multi-step reactions, from both primary reactants and known impurities. We have used phenol to paracetamol synthesis as a use-case to demonstrate the utility of our tool.

2 Results and Discussion

2.1 Input/output file system

The input file requires the user to specify the reactants and products, and optionally, the reagents, solvent, and other known impurities. To keep track of the impurities being predicted further, the input impurities have been classified into impurities from starting materials, drug substance, reagents and solvent.

The output file notes the process or *step* impurities, propagated impurities, and additional raw-material-derived impurities. Step impurities are by-products and side-products of the intended reaction. Propagated impurities include the impurities derived from impurities formed in previous steps of the process that persist through purification. Additional impurities correspond to impurities formed by reactions of known impurities in the input raw materials (reagents, solvents).

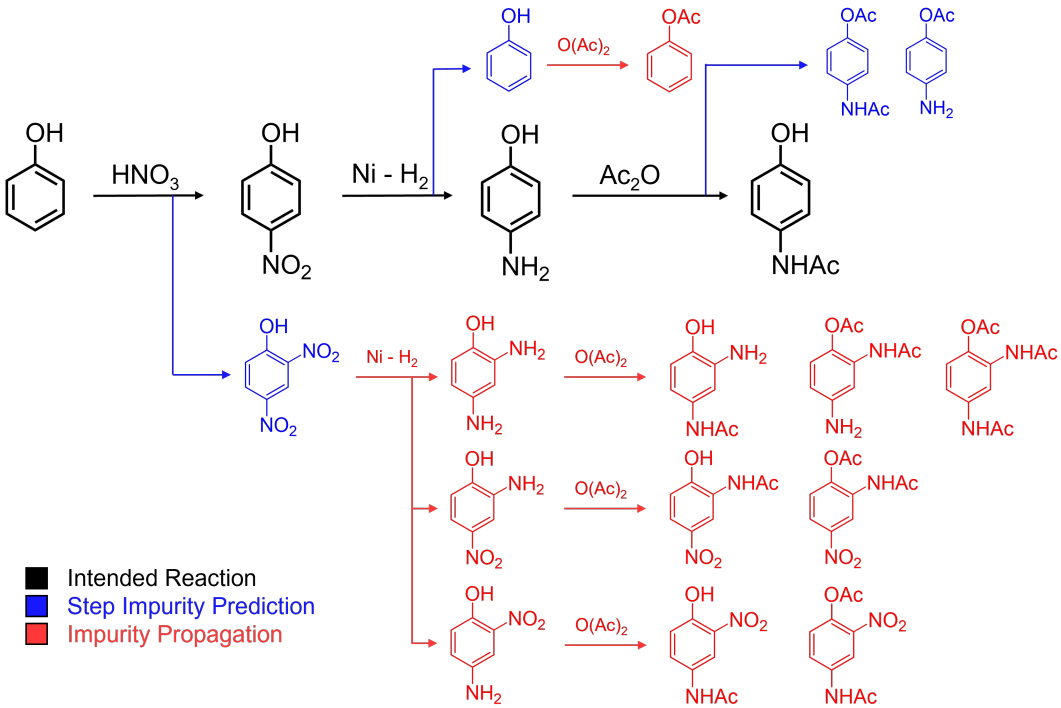

Figure 2: **Impurity prediction in phenol to paracetamol chemical synthesis.** The intended reaction from phenol to paracetamol with **major products**, **step impurities** and **propagated impurities** is shown. For each reaction step, the major step impurity and few propagated impurities are shown.

2.2 Forward prediction model

We used the reaction product prediction model present in the ASKCOS software suite, specifically the model trained on Pistachio dataset, to develop our impurity predictor [5, 12]. To obtain an inclusive set of products - major product and plausible impurities, we increased the top-k predictions to 10 from the default 3 and decreased the threshold probability from the default 0.1 to 0.01.

The ASKCOS suite was hosted on an AWS EC2 instance with default configuration, and was queried through API from a 4-core Intel i7 system. For the demonstrated 3-steps reaction, we clocked 4 min of wall time and 719 ms of CPU time to obtain the results.

2.3 Impurity prediction and propagation

Impurity prediction was done per reaction step in a sequential manner, starting with the intended reaction, followed by the reaction with the reagents, solvent, and impurities from preceding steps. For each step, the model was queried with specific reactants, reagents, solvent and products as input, and all products, excluding the major product of the reaction, were noted as impurities. To categorize the impurities resulting from the intended reaction, we denoted them as step impurities (Figure 2, structures in **blue**).

Impurity propagation was tracked by predicting the products for the reaction of step impurities with the reagents and solvent of the succeeding step(s). To avoid a combinatorial explosion, we limited the number of propagating steps to 1, by default, and extendable to n or all subsequent steps. These impurities were denoted as propagated impurities (Figure 2, structures in **red**).

To improve on results from existing single-step predictors, we added the functionality to include known impurities from primary starting materials in the prediction workflow. This functionality helps in assessing the risk posed by low-level impurities in raw materials—as purchased, reaction intermediates, and API starting materials. Apart from the aforementioned impurities, there is added functionality to include impurities from individual unit operations in the chemical process, such as those arising from addition of anti-solvent in the crystallization step. In the paracetamol synthesis

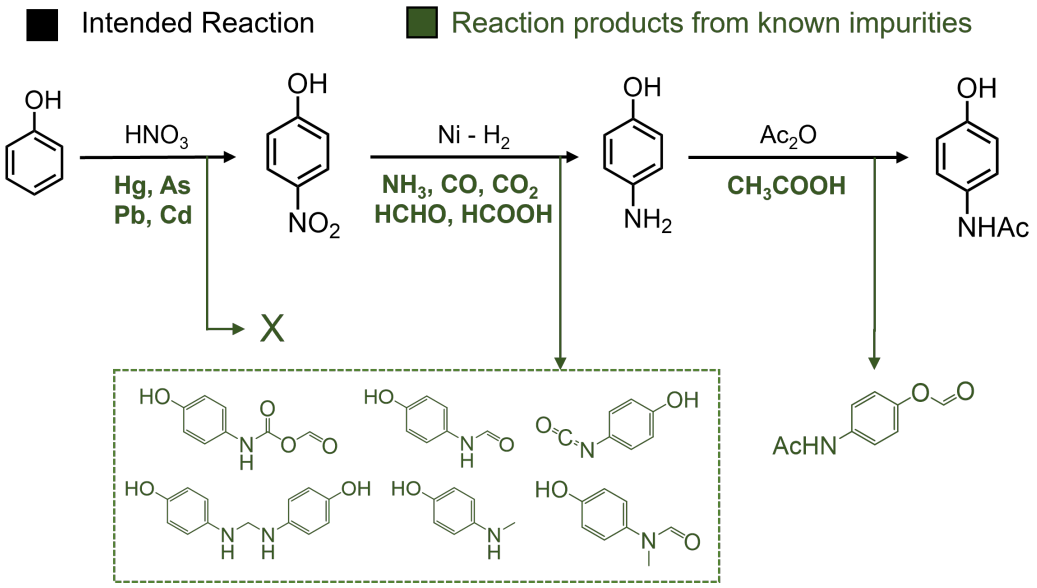

Figure 3: **Reaction products predicted by including known impurities in reagents in phenol to paracetamol chemical synthesis.** The intended reaction from phenol to paracetamol with **major products**, and **additional impurities** is shown. **Impurities in reagents** for the concerned steps are noted below the reagents.

example, we observed no new impurities coming from reactions with trace metal impurities (Hg, As, Pb, Cd) in nitric acid (HNO_3), multiple impurities 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 the lone impurity formed by reaction with acetic acid (CH_3COOH) impurity in acetic anhydride (Ac_2O) (Figure 3, structures in **green**). Known impurities in the reagents were obtained from the literature [13, 14, 15].

3 Limitations, Future Work and Conclusion

The limitations of our impurity prediction approach are intrinsically linked to two key factors - predictive accuracy of the forward prediction model, and known impurities in the starting materials. In the future, we intend to improve the impurity prediction by using a model ensemble approach. Specifically, we plan on including predictions from forward reaction predictors with different model architectures, such as language-based models in IBM-RXN, along with the current graph-based model in ASKCOS. We believe that this approach will help expand the chemical space of impurities, while ensuring the prediction of plausible impurities. To help in predicting low-level impurities and perform raw materials risk assessments, we plan on building a database of known impurities in reagents, solvents and starting materials, and adding the functionality to automatically include these impurities to the prediction workflow.

Apart from the improvements to impurity prediction, we aim to extend our approach to include more types of chemical reactions. In the current state, our approach is suitable only for process chemistry with conventional chemicals. We would like to build out functionality to cater to chemical synthesis involving biological steps, such as enzyme mediated reactions.

In conclusion, our method demonstrates how an AI-assisted approach can enable impurity prediction and propagation for multi-step chemical reactions. Our tool provides the functionality to include impurities from reagents, solvent and starting materials. We envision that AI-assisted impurity prediction in conjunction with inverse structure elucidation from analytical data will help in faster impurity identification, and ultimately accelerate chemical process development.

References

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

[2] Sheun Oshinbolu, Louisa J Wilson, Will Lewis, Rachana Shah, and Daniel G Bracewell. Measurement of impurities to support process development and manufacture of biopharmaceuticals. *TrAC Trends in Analytical Chemistry*, 101:120–128, 2018.

[3] David J Snodin. A primer for pharmaceutical process development chemists and analysts in relation to impurities perceived to be mutagenic or "genotoxic". *Organic Process Research & Development*, 24(11):2407–2427, 2020.

[4] Duane A Pierson, Bernard A Olsen, David K Robbins, Keith M DeVries, and David L Varie. Approaches to assessment, testing decisions, and analytical determination of genotoxic impurities in drug substances. *Organic Process Research & Development*, 13(2):285–291, 2009.

[5] Connor W Coley, Wengong Jin, Luke Rogers, Timothy F Jamison, Tommi S Jaakkola, William H Green, Regina Barzilay, and Klavs F Jensen. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chemical science*, 10(2):370–377, 2019.

[6] Adarsh Arun, Zhen Guo, Simon Sung, and Alexei Lapkin. Reaction impurity prediction using a data mining approach. *ChemRxiv*, 2022.

[7] Jaan A Pesti, Thomas LaPorte, John E Thornton, Lori Spangler, Frederic Buono, Gerard Crispino, Frank Gibson, Paul Lobben, and Christos G Papaioannou. Commercial synthesis of a pyrrolotriazine–fluoroindole intermediate to brivanib alaninate: Process development directed toward impurity control. *Organic Process Research & Development*, 18(1):89–102, 2014.

[8] Eric A Standley, Dustin A Bringley, Selcuk Calimsiz, Jeffrey D Ng, Keshab Sarma, Jinyu Shen, David A Siler, Andrea Ambrosi, Wen-Tau T Chang, Anna Chiu, et al. Synthesis of rovafovir etalafenamide (part i): Active pharmaceutical ingredient process development, scale-up, and impurity control strategy. *Organic Process Research & Development*, 25(5):1215–1236, 2021.

[9] Peter K Dornan, Travis Anthoine, Matthew G Beaver, Guilong Charles Cheng, Dawn E Cohen, Sheng Cui, William E Lake, Neil F Langille, Susan P Lucas, Jenil Patel, 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. *Organic Process Research & Development*, 24(4):481–489, 2020.

[10] Christopher J Borths, Mark D Argentine, John Donaubauer, Eric L Elliott, Jared Evans, Timothy Talbot Kramer, Heewon Lee, Rodney Parsons, Jeffrey C Roberts, Gregory W Sluggett, et al. Control of mutagenic impurities: Survey of pharmaceutical company practices and a proposed framework for industry alignment. *Organic Process Research & Development*, 25(4):831–837, 2021.

[11] Michael A Stravs, Kai Dührkop, Sebastian Böcker, and Nicola Zamboni. Msnovelist: De novo structure generation from mass spectra. *Nature Methods*, pages 1–6, 2022.

[12] Amol Thakkar, Thierry Kogej, Jean-Louis Reymond, Ola Engkvist, and Esben Jannik Bjerrum. Datasets and their influence on the development of computer assisted synthesis planning tools in the pharmaceutical domain. *Chemical science*, 11(1):154–168, 2020.

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

[14] WS Calcott, FL English, and OC Wilbur. Analysis of acetic anhydride. *Industrial & Engineering Chemistry*, 17(9):942–944, 1925.

[15] Claire Beurey, Bruno Gozlan, Martine Carré, Thomas Bacquart, Abigail Morris, Niamh Moore, Karine Arrhenius, Heleen Meuzelaar, Stefan Persijn, Andrés Rojo, et al. Review and survey of methods for analysis of impurities in hydrogen for fuel cell vehicles according to iso 14687: 2019. *Frontiers in Energy Research*, 8:615149, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] Code is proprietary. Data and instructions have been provided to reproduce the work.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [No] Code is proprietary.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]