SEISMiQ: de novo impurity structure elucidation from tandem mass spectra boosts drug development.

Emilio Dorigatti[a, ORCID: 0000-0002-6829-7766], Jonathan Groß[b, ORCID: 0009-0004-2340-0629], Jonas Kühlborn[b], Robert Möckel[b, ORCID: 0000-0001-9883-9952], Frank Maier[a]* and Julian Keupp[b, ORCID: 0000-0001-8735-6883]*

[a] Development NCE, Analytical Development, Boehringer Ingelheim Pharma GmbH & Co. KG, D-55218, Ingelheim (Rhein), Germany

[b] Development NCE, Chemical Development, Boehringer Ingelheim Pharma GmbH & Co. KG, D-55218, Ingelheim (Rhein), Germany

* Corresponding Authors

Abstract

Liquid chromatography-tandem mass spectrometry (LC-MS/MS) is an essential analytical technique in the pharmaceutical industry, used particularly for elucidating the structure of unknown impurities in the synthesis of active pharmaceutical ingredients. However, the interpretation of mass spectra is challenging and time-consuming, requiring significant expertise. While recent computational tools aimed at automating this process have been developed, their accuracy in determining the chemical structure is limited. In this paper, we introduce a new method for elucidating unknown impurities from their MS/MS spectra. We are able to significantly improve elucidation accuracy by integrating domain experts' knowledge, specifically the impurity sum formula and known substructure, into the model's training and inference process. Further performance improvements can be achieved through transfer learning from simulated MS/MS spectra of impurities from an in-house database. Finally, the need for any experimental data collection for finetuning can be circumvented by simulating the entire drug substance synthesis process *in silico* via reaction templates.

Introduction

Structure elucidation of unknown impurities from high resolution LC-MS/MS is a crucial step in the pharmaceutical drug substance development.[1] Their characterization allows the assessment of toxicological implications, guides the development and optimization of the drug substance synthesis process and establishes quality control criteria employed during later lifecycle.[2] Despite its widespread adoption and essential contribution to drug substance development and many other endeavors, the interpretation of mass spectra remains challenging and requires hours of manual work of analytical experts who are specifically trained for this task.[3]

Several computational approaches have been developed to increase the speed and reliability of the MS/MS spectra interpretation workflow, with a particular focus on metabolomics.[4–7] Initial *in silico* solutions ranked molecules in a given list of candidates to surface molecules whose mass spectrum would be most similar to the given spectrum.[8–12] Such procedures were generally based on predicting relevant structural information from the MS/MS spectrum and matching these with the corresponding structural information computed from the candidate molecules. While this ranking approach could help practitioners in daily work, it is limited by its inability to propose novel structures not already in the initial list. The recent evolution of deep generative models removed the necessity of a pre-specified list of candidates and enabled *de novo* structural elucidation where the molecular structure is

https://doi.org/10.26434/chemrxiv-2025-6hhn1 **ORCID:** https://orcid.org/0000-0001-8735-6883 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

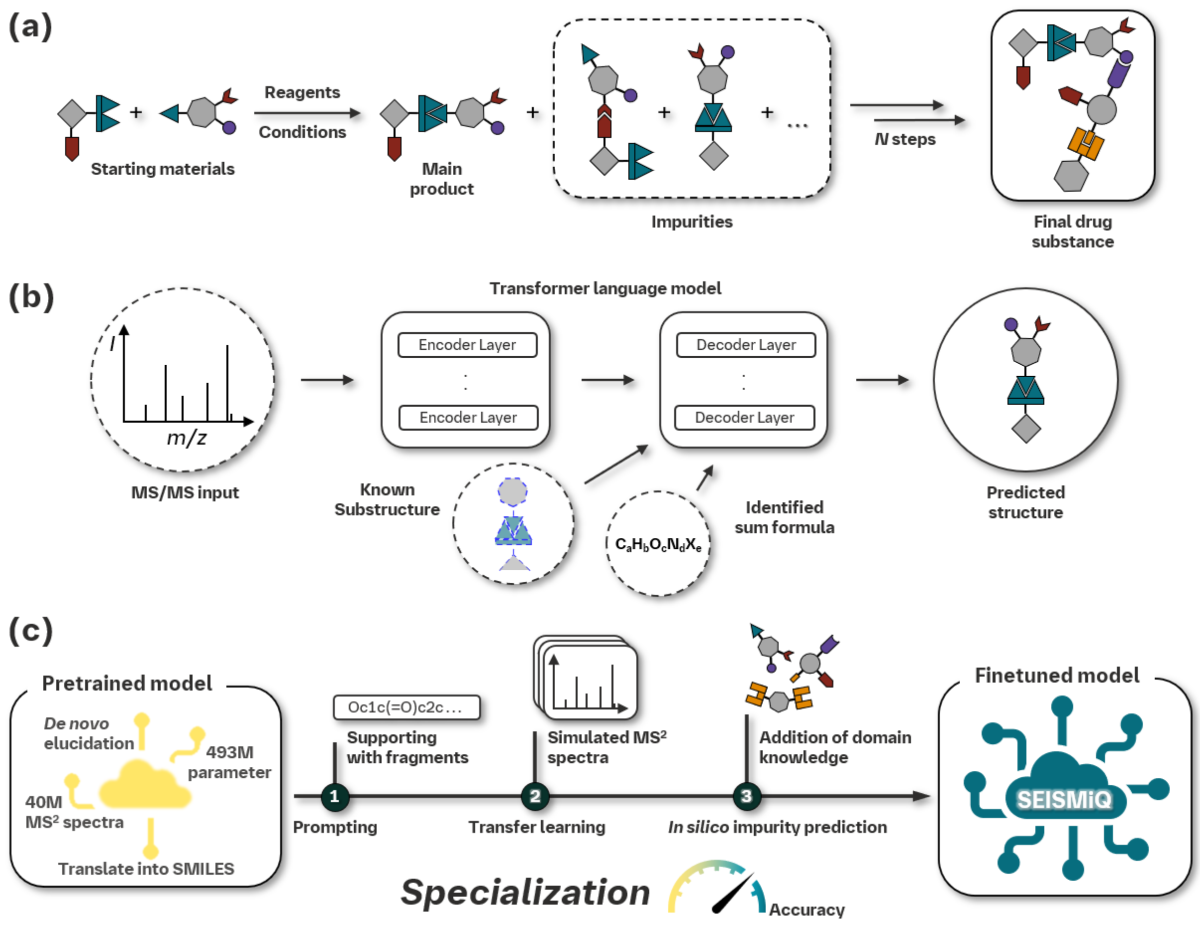

Figure 1: Motivation and approach of SEISMiQ. (a) Unwanted impurities are formed during the synthesis of a drug substance. Elucidating their structure from MS/MS spectra is however very time consuming. (b) We developed a transformer language model for molecular structure elucidation from MS/MS spectra that leverages expert domain knowledge through sum formula and prompting of known impurity fragments to provide more accurate structure proposals. (c) We specialized this model for impurity elucidation through transfer learning from simulated MS/MS spectra of related impurities, and finally simulated the entire synthetic route including impurity formation in silico to remove the need for experimental data.

predicted from scratch rather than by searching a known pool of molecules.[13–18] The major challenge in the field is the inherent ambiguity of MS/MS spectra and the relative scarcity of open datasets and benchmarks, with, to our knowledge, the largest available covering only about 29,000 different molecules.[19] The difficulty of obtaining high quality expert annotations of MS/MS spectra is likely to prevent the growth of such available data to the amounts used to train the latest molecular generative models[20,21] Common workarounds for this issue include using pretrained models[15,16,18] and augmenting the training set with large numbers of simulated MS/MS spectra,[14,22] approaches which have seen notable innovations recently.[7,23–28]

While these developments have greatly raised *de novo* elucidation accuracy, the performance of these models is not yet at the level desired by practitioners to enhance their productivity in drug substance impurity elucidation. In order to correctly elucidate impurities from MS/MS spectra, analytical chemists leverage a wide range of domain knowledge regarding the synthetic route that generated the impurity, including starting materials and their impurities, the conditions under which reactions take place, possible unwanted side reactions, over reaction and others (Figure 1a). This information suggests a great deal about the impurity structure, including identical fragments shared with the main compound and sites of variation. By focusing on a purely *de novo* setting, current models for structural elucidation remain unable to leverage this knowledge and as a result do not achieve the desired accuracy level while at the same time making easily avoidable elucidation mistakes. Motivated by this,

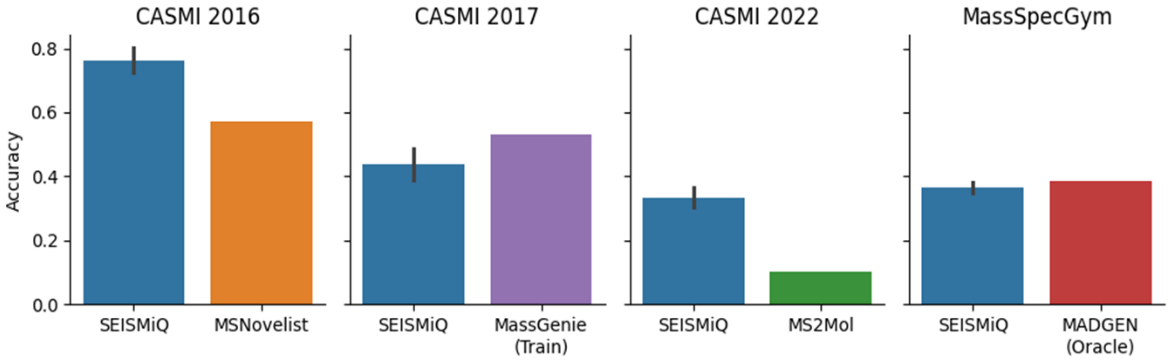

Figure 2: De novo elucidation accuracy of SEISMiQ on CASMI 2016, 201, 2022 and MassSpecGym compared to the reported performance of other published models. MADGEN (Oracle) was given the correct molecule scaffold to complete, and MassGenie was only evaluated on a subset of CASMI 2017 which was part of its own training set. Error bars represent the standard error of the mean.

we introduce a novel method for *de novo* elucidation of small molecules from MS/MS spectra and specifically apply it to the problem of elucidation of unknown structures in the synthesis process of a drug substance. We demonstrate how to integrate the knowledge of domain experts into the training and inference process of the model to improve elucidation accuracy (Figure 1b). By finetuning it on simulated MS/MS spectra of related impurities, we further enhance the model's performance, showing for the first time the potential of transfer learning from simulations. Lastly, we simulate the entire synthetic route *in silico*, including impurity formation events, removing the need for any experimental data collection (Figure 1c).

Results

We tackle the problem of structural elucidation as an automated machine translation problem based on language modeling, where a model is trained to "translate" the MS/MS spectrum into the corresponding molecule as a SMILES representation (Figure 1b), similar to other works in the field.[13–15,22] Our model is a transformer language model trained on 40M MS/MS spectra (10M molecules), most of which simulated via CFM-ID[29] and FragGenie.[14] By design, our model requires a sum formula as input, since it is either already known, or being easily accessible through several tools that can accurately predict it from MS/MS spectra.[35–38]

We evaluated our model on the Critical Assessment for Small Molecule Identification (CASMI) challenges,[39] as well as the newly released MassSpecGym benchmark[19] (Figure 2). Our model achieved top-128 accuracies of 76.4%, 43.8% and 33.3% respectively for CASMI 2016, 2017 and 2022 (top-*k* performance for different values of k in Supplementary Information S1). All CASMI molecules and their spectra were removed from the model's training and validation sets to ensure an unbiased evaluation. As the MassSpecGym benchmark was published after our model was trained, we only used for the evaluation molecules that were not in the model's training set. This resulted in 992 MS/MS spectra on which our model reached an accuracy of 36.5%. On this benchmark, we did not find a difference in performance between different instrument types, and a slight decrease for [M+Na]+ adducts (Supplementary Information S2).

MSNovelist[15] reached a top-128 accuracy of 57% and top-1 of 26% on CASMI 2016 when using the sum formula predicted by SIRIUS, which was correct in 93.8% of the cases. MS2Mol[40] does not require a sum formula as input and reached a top-25 accuracy of 9% on CASMI 2022. MassGenie[14] reported an accuracy of 53% on a subset of 93 challenges of CASMI 2017 with small molecular weight that were also used to train their model, while Mass2SMILES[13] correctly elucidated 2/236 challenges of CASMI

2022. Spec2mol[18] could not be evaluated on the CASMI challenges, as their model requires four input spectra, combining positive and negative ionization with high and low collision energy, to elucidate a molecular structure. MADGEN[41] is a scaffold completion model that obtained a top-10 accuracy of 1.6% on MassSpecGym when choosing a scaffold from a list of 256 options, and 38.6% when given the true scaffold of the molecule. The improvement in performance of our model can be attributed to the larger and more diverse training dataset, the data augmentation protocol employed during training, the larger model size, and the fact that the correct sum formula is given as input. As most of these models lack public and freely usable implementations, we limit ourselves to reporting their performance as originally stated in the respective publications, and to facilitate future research in this area, we open source our implementation including training code, data, and pretrained checkpoints upon peer reviewed publication at the following link at https://github.com/Boehringer-Ingelheim/seismiq.

We also assessed our model on an internal dataset composed of 174 experimentally detected impurities of several small molecule drug substances collected during routine operations in analytical development (Supplementary Information S3). On this dataset, our model correctly elucidated only nine (5%) of the impurities, while providing predictions with Tanimoto of at least 0.8 for 49 (30%) impurities, highlighting the challenge posed by the lack of representative training data for reliably elucidating impurities. This problem is exacerbated in a pharmaceutical setting, where substrates change significantly from one drug substance project to the next, posing considerable challenges for creating a truly representative training set.

Progress by Prompting: Enhancing Correct Elucidation Rates by Leveraging Main Compound-Impurity Similarity

While our model obtained very low accuracy on our internal test set, analytical chemists were able to elucidate these structures. They could do so by leveraging a wide range of additional information that is known based on the expected main component, the chemical process used to synthesize it, and additional knowledge gathered over time by working on the project. All this information can suggest where the impurity differs from the main compound, and sometimes even in which way. In our experience, three quarters of the impurities considered in small molecule analytical development share between 50% and 80% of the molecular structure with the main compound. This common substructure provides a crucial starting point to allow analytical chemists to manually elucidate impurities from MS/MS spectra.

Based on these considerations, we specifically selected a model architecture that can take this known common substructure as expert-provided input and complete it into a fully formed molecule (Figure 1b). To do this, we prompt the language model with a SMILES string constructed in such a way that the attachment point between the common fragment and the impurity site of variation is in the last position of the string. We quantitatively validated the ability of our model to elucidate the molecular structure when prompted in this way by simulating different known fragments from the test datasets. Specifically, we generated fragments by breaking all single bonds of each molecule in the dataset, prompted the model with the SMILES of each of the two fragments in turn and evaluated how close the model's predictions were to the whole molecule (Figure 3a).

On the public test datasets, this resulted in 48,628 fragments with an average of 27 and a maximum of 70 missing atoms (Figure 3b). On such fragments, the model obtained an accuracy of 96.3% when it was tasked to complete fragments missing up to 10 atoms, 71.5% for fragments missing up to 30 atoms, and 35.4% for fragments beyond 30 (Figure 3c). Nonetheless, for these fragments the average

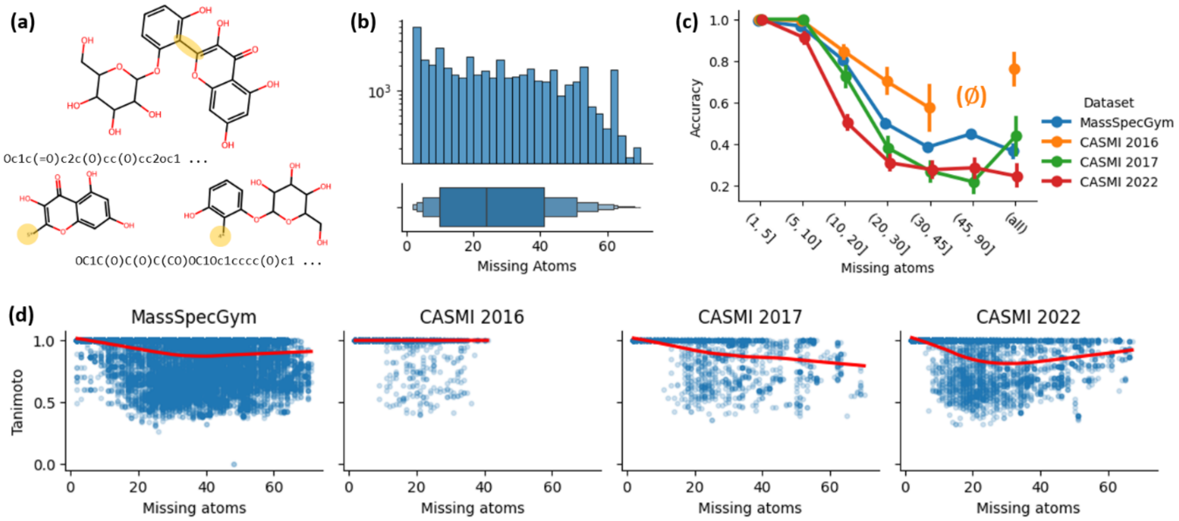

Figure 3: Prompting the model with impurity fragments to boost elucidation accuracy. (a) Breaking a molecule (top) on the highlighted bond generates two fragments (bottom), with their SMILES representation created so that the attachment atom to the other fragment (highlighted) is in last position of the string. (b) Distribution of the number of missing atoms for each fragment. (c) Model accuracy on four public datasets by number of atoms missing from the fragment (no fragments missing more than 45 atoms for CASMI 2016, all means that no prompt was given). Error bars represent the standard error of the mean. (d) Tanimoto scores (y-axis) for each fragment by number of missing atoms (x-axis) for the four public test datasets. Red lines given by locally weighted linear regression fits.

Tanimoto of the predicted molecules was 0.82 (Figure 3d) and in 73.0% of the cases the Tanimoto similarity was at least 0.675, indicating a close agreement to the ground truth.[22]

Despite these encouraging results, the molecular structures under consideration were entirely new for the model and never seen during training. Drug substance impurities however tend to be structurally similar among each other; during the development and optimization process of the synthesis pathway a significant number of related impurities are characterized, and several distinct projects make use of similar reactions to synthesize the respective main compounds. These considerations motivated us to make use of this historical data and investigate ways to incorporate this implicit process knowledge into the model.

Transfer Learning Triumph: Improved Elucidation Accuracy by Fine-Tuning on Predicted Spectra of Historical Impurities

Only 2% of the MS/MS spectra in the training dataset were collected experimentally, with the rest being predicted *in silico*. Considering the promising model performance, we wanted to quantify how well information from simulated data is transferable to experimental spectra. To evaluate this, we collected all impurities from the internal company database, which amounted to 22,353 molecules and originated 109,343 simulated MS/MS spectra, and we finetuned the model on this dataset after removing all impurities that were already in the test set. We also created a second model that was finetuned on the simulated spectra of the test impurities, in addition to the historical impurities. We then evaluated both models on the experimental spectra of the test impurities (Figure 4a). The performance of the latter model indicates to what extent it is possible to generalize from simulated to experimental MS/MS spectra of the same molecule, while the former model allows us to evaluate the model's ability to generalize from structurally related molecules, for example by recognizing common structural motifs.

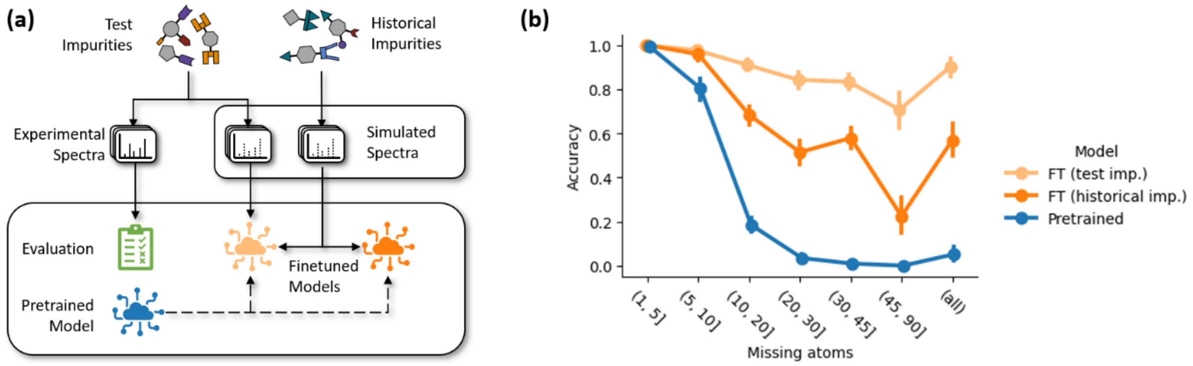

Figure 4: Transfer learning ability from simulated to experimental spectra. (a) Experimental protocol: we used simulated spectra to create two finetuned models, one using the historical impurities and one using historical and test impurities and evaluated all models on the experimental spectra of the test impurities. (b) Accuracy (y axis) on the fragment completion task on the test impurities as a function of the number of atoms to be completed (x axis) for the pretrained model (blue), a model finetuned (FT) on the simulated spectra of historical impurities excluding the test impurities (dark orange), a model finetuned on the simulated spectra of the test impurities (light orange)

While there was no difference in performance when completing fragments missing up to five atoms, the positive effect of finetuning is apparent starting from ten missing atoms (Figure 4b). Between ten and twenty missing atoms, the pretrained model obtained an accuracy of only 18.5%, while the model finetuned on historical impurities excluding the test ones obtained 68.4% correct predictions. Finetuning on simulated spectra of the test impurities further raised accuracy to 91.1%. The gap between pretrained and finetuned models further widens for *de novo* elucidation, where the accuracy of 5.2% of the pretrained model was improved to 58.9% by using historical data and 90% when using simulated spectra of test impurities. Our finetuning protocol caused, however, detrimental effects on the performance on molecules that were not related to the finetuning set. In the CASMI challenges, for example, accuracies decreased by 32, 31 and 22 percentage points for the years 2016, 2017 and 2022 respectively. Supplementary Information S4 shows the performance on the CASMI challenges when finetuning the model on their simulated spectra.

These results convincingly show that it is possible to considerably boost elucidation accuracy by fine-tuning the model on simulated MS/MS spectra of structurally related or even identical molecules, at a certain price on unrelated molecules. Obtaining such a dataset is however extremely time consuming, as it requires substantial efforts to manually collect and elucidate hundreds or thousands of impurities.

In-silico Synthetic Solution: Removing the Need of Historical Data by Simulating Impurities for a Synthesis Route

When a new synthesis project is started, there is not yet enough historical data regarding the impurities for that specific drug substance, thus the organic chemistry knowledge of experts plays a central role in enabling structural elucidation from MS/MS. This also poses significant challenges in finetuning our model, as the number of available impurity examples would be too low to allow reliable and generalizable training. To alleviate this “cold start” issue, we attempted to simulate the entire synthesis process of an asset *in silico*, including impurity formation events. As most impurities arise from known chemical processes, including for example incorrect selectivity and overreactions, we reasoned that it should be possible to reproduce this process given all the starting materials and impurities thereof.

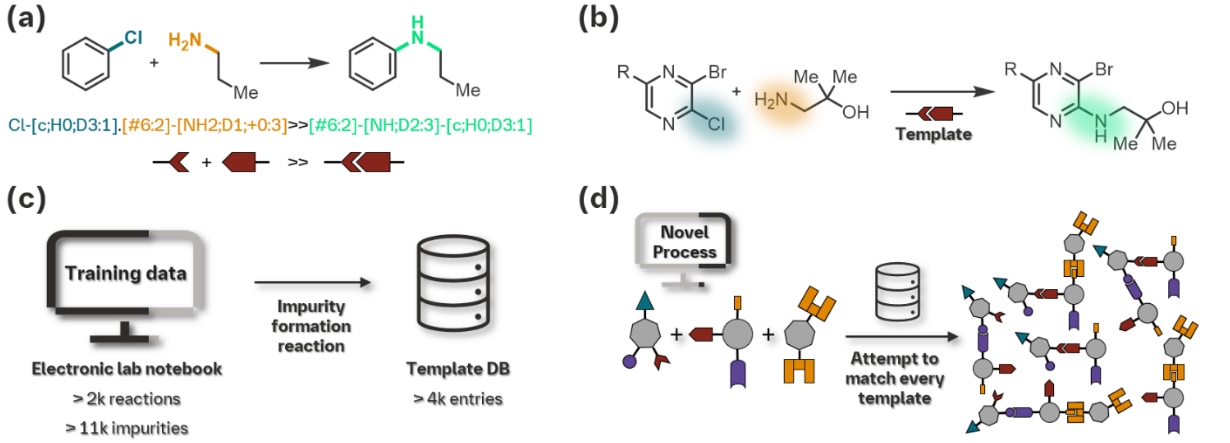

Figure 5: In silico prediction of impurities generated by a reaction. (a) SMARTS templates are transformation rules that describe how the reactive sites in the starting materials are combined to form the main product. These templates can be extracted automatically given an example reaction. (b) The same template can be applied to a variety of starting materials that present the required reactive sites. (c) We extracted over 4,000 reaction templates from our internal electronic lab notebook containing data for more than 2,000 reactions and 11,000 resulting impurities. (d) We simulated the synthesis route of an internal asset using the extracted reaction templates, excluding all templates extracted from reactions belonging to the test asset project.

We developed an impurity predictor based on SMARTS reaction templates,[42] describing how the products in a chemical reaction are formed by combining fragments of the starting materials (Figure 5a, b). We integrated data from an internal electronic laboratory notebook reporting performed reactions with the corresponding starting materials and analytically detected impurities (Figure 5c) allowing us to cover both the desired reactions forming the main compound and additional processes that generated detected impurities. By codifying an asset's synthesis route as a sequence of template applications we were able to generate a dataset of impurities that is entirely simulated from first principles only based on the knowledge of the synthesis route (Figure 5d).

For this experiment, we focused on the synthesis route of an internal asset whose main compound has 41 non-hydrogen atoms, a molar weight of ca. 600 g/mol and is comprised of various functional groups, multiple annulated rings as well as a chiral spiro carbon. The synthesis route for this asset spans seven distinct steps involving four different starting materials and covering multiple reaction types like condensation, oxidation, or reductive amination reactions. We excluded from the template extraction procedure all reactions reported in the electronic laboratory notebook as making part of the test asset synthesis route, leaving us with 4,446 templates in total. Their application resulted in 20,813 simulated impurities with mass below 1,200 Da, and 154,756 corresponding simulated mass spectra. Of these impurities, 27 were also in our test dataset, which contained in total 61 impurities for this asset. In general, the chemical space covered by the simulated impurities included close matches for all experimental impurities (Figure 6a) and revealed additional impurity clusters that were not detected, possibly because the reaction conditions did not allow for such impurities to form in sufficient quantities, or they were not stable or isolated under the given work-up conditions.

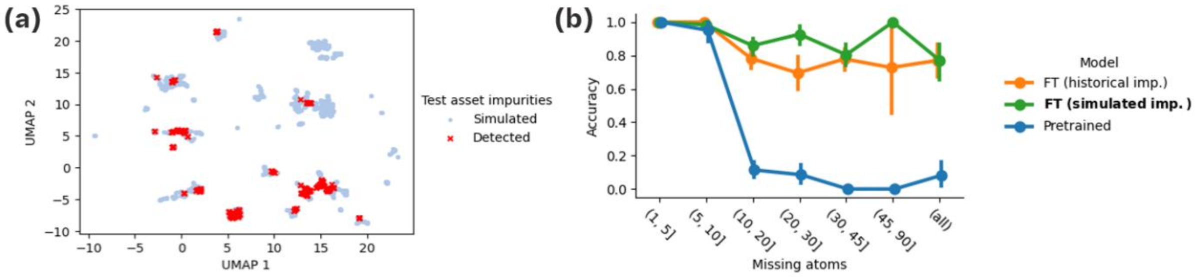

Figure 6: Results of impurity simulation and model finetuning. (a) UMAP[53] visualization of the simulated (light blue dots) and experimentally detected (red crosses) test asset impurities. (b) Accuracy (y axis) on the fragment completion task for the test asset impurities as a function of missing atoms to be completed (x axis) for the pretrained model (blue), a model finetuned on the simulated spectra of historical in-house impurities (orange), and a model finetuned on the simulated spectra of the simulated impurities (green).

We finetuned and evaluated our model following the same protocols as before, excluding from the finetuning set the 61 experimentally detected test asset impurities, and we compared this model with the model finetuned on historical in-house impurities described in the previous section. Both models correctly predicted the same 46 (78.0%) impurities in a *de novo* setting. For fragment completion, finetuning on simulated data appeared to result in better performance, although the small dataset size of only 61 impurities caused some fluctuations (Figure 6b). Nonetheless, when considering all fragments altogether the model finetuned on simulated data had 5.6% higher accuracy (88.7% vs 83.1%). This shows that an entirely *in silico* impurity simulation approach of process impurities and their MS/MS spectra can be used to significantly boost elucidation accuracy in absence of relevant experimental data for finetuning.

Discussion

In this paper we tackled the problem of elucidating unknown impurities generated during the synthesis of drug substances from MS/MS spectra. Our approach compared favorably with contemporary tools for *de novo* elucidation on public benchmarks, however its performance on an internal impurity dataset was insufficient to provide useful insights.

We explored three ways of dealing with this challenge. First, we leveraged data augmentation at training time to teach the model how to complete a given molecule fragment. Analytical chemists are able to identify which parts of the impurity are identical to the main compound and providing this fragment to the model resulted in significant gains in accuracy. Second, we finetuned the model on simulated MS/MS spectra of structurally related molecules and showed that this approach significantly improved elucidation accuracy. Third, as obtaining such datasets for finetuning is extremely time-consuming, we attempted to simulate the synthesis process of an asset *in silico* to predict the impurities that are likely to be generated. We showed that finetuning only on these simulated impurities resulted in a model that is slightly more accurate than one finetuned on historical impurities only, thus enabling elucidation of impurities without the necessity of any prior experimental measurement.

Our approach is not free from limitations. First, we did not make use of any ranking method for the predictions. Several methods for this have been presented in the literature, including for example the modified Platt score[15] based on the predicted CSI:FingerID[43,44] fingerprints of the input MS/MS spectrum, a dedicated reranking model,[22] ranking based on the predicted spectra of the predicted molecules,[14] or simply ranking by perplexity. All these methods are compatible and directly applicable to our model as well. Second, our finetuning protocol introduced some overfitting to the dataset used for finetuning. The danger of reduced accuracy could be identified at inference time by comparing the input MS/MS spectrum with the finetuning dataset for example via CSI:FingerID fingerprints, and the pretrained model could be used as a fallback. Other finetuning protocols that reduce the risk of overfitting could also be employed. Third, our method is limited to completing fragments with a single extension point. It is not uncommon however for impurities to differ from the main compound in two or more locations, rendering our model unable to directly deal with such cases.

In conclusion, by considering the unique aspects of impurity generation we made significant breakthroughs in the *de novo* elucidation of impurities from MS/MS spectra. This advancement holds particular relevance in the pharmaceutical industry, where impurity characterization is essential for optimizing manufacturing processes, understanding degradation pathways, ensuring drug substance stability, maintaining quality control, and achieving regulatory compliance. Our model aims to enhance the efficiency of drug substance development—a process that typically spans five to ten years—and mitigate its cost. By incorporating impurity elucidation earlier in the process, we aim to alleviate the workload of analytical chemists and facilitate semi-automated elucidation workflows.

Methods

Dataset

We used publicly available positive ion mode mass spectra from the GNPS, MassBank and MatchMS online libraries. We shifted the *m/z* peaks of each spectrum to remove the adduct effect and discarded all spectra which resulted in peaks with negative mass or mass larger than the monoisotopic mass of the corresponding molecule, as well as all spectra with less than five peaks. We augmented this dataset with simulated mass spectra for a large set of molecules from ChEMBL,[32] PubChem[33] and ZINC[34] using the CFM-ID[29] and FragGenie[14] predictors simulating the mass spectra at three different collision energies (10, 20 and 40 eV for CFM-ID, and building fragmentation trees with depths from one to three for FragGenie). All molecules were cleaned by removing chiral information and any charges. In total, we collected 10,750,283 molecules and 41,058643 spectra. Of these, 1,090,317 spectra were measured experimentally, corresponding to 51,417 molecules. We reserved 1% of the data (418,587 spectra / 107,398 molecules) for validation, such that the training set did not contain any spectrum of any molecule in the validation set.

To improve the generalizability of our model and reduce the risk of overfitting, we leverage several techniques to further augment this dataset during training, in addition to the traditional regularization techniques of dropout and weight decay typically employed in deep learning. First, we chose a random number of peaks between 5 and 50, randomly sampled from the entire spectrum with probability proportional to their intensity. Further, the *m/z* value of each peak was slightly perturbed by a random amount of uniform noise with magnitude 0.02, thus making the model resilient to measurement noise. Each peak was paired with the corresponding neutral loss, and both were encoded with a sinusoidal position encoding to a dimensionality of 512. The frequencies for the sinusoidal encoding included the atomic masses of H, C, N, O, Cl, S, P, K, F, and Br, and 246 additional frequencies between 3^{-5} (0.0041) and 3^{7} (2,187) evenly distributed in base three log-space.

The model takes as input molecules encoded as SMILES strings. Initial experiments on a small development dataset revealed that this encoding performs on par or slightly better than SELFIES[45] and DeepSMILES,[46] while having favorable computational requirements due to their shorter lengths. SMILES strings were tokenized with one token per atom, so that C and Cl were mapped to different tokens, resulting in 305 tokens in total. The model was presented with the SMILES string in canonical order 25% of the times, and a randomized atom order is used in the other 75% of the times. Regardless of the order, 50% of the time we kekulized the SMILES.

In addition to the SMILES tokens, the model received as input an encoding of the remaining heavy atoms to be generated to complete the molecule, computed based on the sum formula of the molecule and updated with every generated token as done in MSNovelist.[15] We also included a sinusoidal positional encoding for the SMILES tokens.

Model

The model was a transformer with 16 encoder and 16 decoder layers, each with 16 heads, with hidden dimension of 1,024 and feed-forward dimension of 4,096. Multilayer perceptrons (MLPs) were used to transform the peak masses, the SMILES tokens joined with the remaining atoms to be generated and predict the following tokens. All these MLPs used one hidden layer of 2,048 units and the rectifier activation function (ReLU). This architecture resulted in 493M trainable parameters.

The model was trained using the cross-entropy loss with a label smoothing 0.1, and re-weighted samples to correct for over- and under-represented molecules in the training dataset. We used the AdamW optimizer with learning rate of $3e^{-5}$, linearly annealed from $6e^{-7}$ over the first 1,000 training steps. A dropout of 0.2 and weight decay of $1e^{-2}$ was applied. The model was trained concurrently on four NVIDIA A100 GPUs each using batch size of 64 and mixed 16-bit precision. Based on the validation metrics, the model did not exhibit signs of overfitting, and we stopped training shortly after 22 epochs, or 3.5 million training steps, taking a total of 23 days.

Model fine-tuning was performed by freezing the transformer and tuning the MLPs, which had overall 23M trainable parameters. We used the AdamW optimizer with weight decay 10^{-4} and initial learning rate of 10^{-4}, exponentially decayed with a factor of 0.995 over 250 epochs. No early stopping was performed.

Inference

Structural elucidation at inference time is performed autoregressively, whereby the tokens constituting the SMILES string sampled one at a time conditioned on the mass spectrum and the preceding tokens already predicted. We used beam search for sampling, which provides higher likelihood molecules at the expense of slightly increased computational requirements compared to greedy sampling. Beam search maintains a pre-specified number *k* of *beams*, each corresponding to a different sequence being generated, as well as its predicted probability. At each step, the model is queried for the probability of each token following each beam, and the top-*k* probability sequences are kept. This procedure is iterated for a given maximum number of steps. Whenever the model predicts a stop-token, the resulting sequence is stored for later analysis. Unless otherwise specified, we report results sampled via *k*=128 beams.

Impurity Simulation

Published AI-driven retrosynthetic prediction tools focus on the prediction of the major product of a reaction (sequence), while we are more interested in a large number of structures of possible impurities.[47] Our approach follows template-based rules in the SMARTS format (SMILES arbitrary target specification)[42] for describing chemical transformations. Compared to publicly available templates, e.g. the widely used USPTO dataset,[48] we envision that the utilization of our company internal impurity formation knowledge should boost the impurity prediction capabilities specifically for our research area.[49]

In practice, we followed a data mining approach by combining internal data of process development and analytically detected impurities, excluding every reaction that is related to the test asset from this approach. For a performed experiment in our electronic lab notebook for the development of new

APIs, we combined the reagents with all detected impurities that were detected. From this *in silico* constructed reactions, we performed in the first step an atom-mapping based on RXNMapper.[50] Subsequently, we extracted a reaction template for this hypothetical reaction leading to an observed impurity applying the RDChiral library.[51] Overall, this approach resulted in 4,446 templates.

With the internal template-set present, we subjected each possible bimolecular combination of reactants for a given experiment of asset to RDKit to generate possible impurity structures[52] over two rounds of generation.

Acknowledgment

We are grateful to Vladimir Lekic, Florian Rottach, Jonas Kühlborn, Marcel Reimann, Niklas Adebar, Jens Smiatek, Andreas Ding, Thorsten Noack, Ulrich Scholz, Gerd Benirschke for their support, encouragement, insightful comments, and discussions.

Supplementary Information

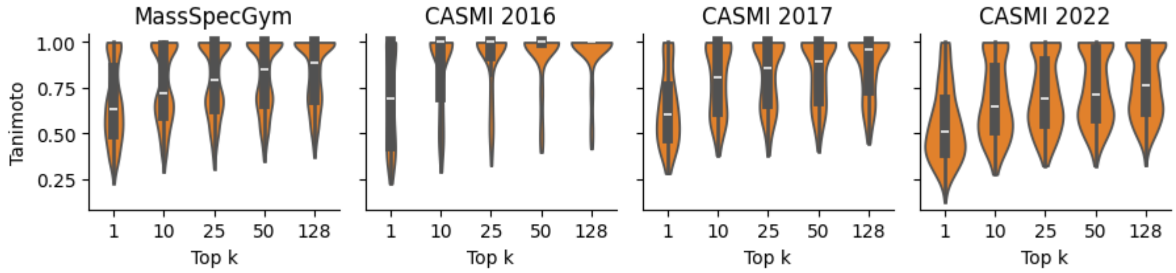

Figure S 1: Tanimoto of the top-k predictions with lowest perplexity of the pretrained model on the public test datasets.

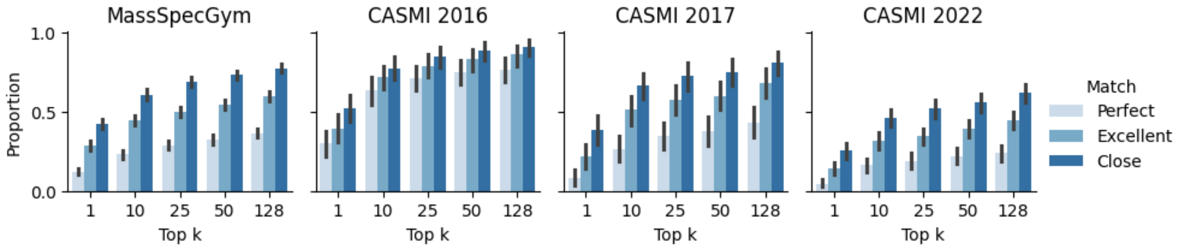

Figure S 2: Proportion of predictions amongst the top-k with lowest perplexity with a given match quality to the true model for the pretrained model on the public test datasets. Perfect, excellent, and close matches correspond to a Tanimoto score of at least 1.0, 0.8, and 0.675 respectively.

Supplementary Information S1 – Ranking Results

The model's predictions were ranked by perplexity. The distribution of Tanimoto scores for the top-*k* predictions is shown in Figure S 1, while the proportion of predictions with a given match quality to the true molecule is shown in Figure S 2.

Supplementary Information S2 – Performance by adduct and instrument

On the MassSpecGym benchmark, the accuracy of SEISMiQ on [M+Na]+ adducts (*n=431)* was 10.0 percentage points lower and statistically significantly worse than [M+H]+ adducts (n=561, *p=0.001,* permutation test). This difference can be attributed to the fact that our training dataset contained very few adducts other than [M+H]+, and even though we corrected for mass differences in the peaks caused by different adducts our model could not generalize to different fragmentation patterns. Additionally, the model performed slightly better on Orbitrap measurements (*n=274*) compared to QTOF devices (*n=450*), however the difference in accuracy of 6.6% was not statistically significant (*p = 0.087*, permutation test). The instrument type was not reported for 268 measurements, on which the performance difference was not found to be statistically significant compared to the two instruments (*p=0.26* and *p=0.73* respectively).

Supplementary Information S3 – Information on internal test impurities

The synthesized impurities were characterized by NMR, MS/MS, and IR. The high-resolution MS/MS data was recorded on Agilent Technologies, Waters and Thermo Scientific devices and includes both TOF and Orbitrap MS/MS data. Various functional groups like aldehydes, aryl amides, benzylamines, prolines, esters, acids, substituted aromatic rings, heterocycles and others were present in the impurities, with molecular masses between 136 and 1260 Da (median: 434) and between 10 and 91 atoms other than hydrogen (median: 30), including Sulfur, Silicon, Bromine, and Phosphorous (19, 8, 7 and 2 impurities respectively).

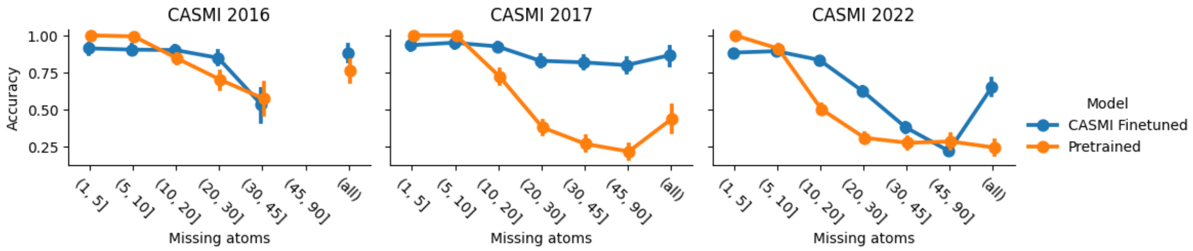

Figure S 3: Accuracy of a model finetuned on simulated spectra of the CASMI challenges and evaluated on the respective experimental spectra. No fragment in CASMI 2016 was missing more than 45 atoms.

Supplementary Information S4 – Finetuning on CASMI challenges

We finetuned the model on the simulated spectra of the CASMI challenges and evaluated it on their experimental spectra following the same protocols for finetuning and evaluation as for the main results (Figure S 3). The *de novo* elucidation accuracy without any given molecule fragment increased by 11.8, 42.9 and 40.8 percentage points respectively for the years 2016, 2017 and 2022.

References

1. Rahman, N., Azmi, S. N. H. & Wu, H.-F. The importance of impurity analysis in pharmaceutical products: an integrated approach. *Accreditation Qual. Assur.* **11**, 69–74 (2006).

2. Qiu, F. & Norwood, D. L. Identification of Pharmaceutical Impurities. *J. Liq. Chromatogr. Relat. Technol.* **30**, 877–935 (2007).

3. Vijlder, T. D. *et al.* A tutorial in small molecule identification via electrospray ionization-mass spectrometry: The practical art of structural elucidation. *Mass Spectrom. Rev.* **37**, 607–629 (2018).

4. Beck, A. G. *et al.* Recent Developments in Machine Learning for Mass Spectrometry. *ACS Meas. Sci. Au* **4**, 233–246 (2024).

5. Liu, Y., Vijlder, T. D., Bittremieux, W., Laukens, K. & Heyndrickx, W. Current and future deep learning algorithms for tandem mass spectrometry (MS/MS)-based small molecule structure elucidation. *Rapid Commun. Mass Spectrom.* e9120 (2021) doi:10.1002/rcm.9120.

6. Lu, X.-Y. *et al.* Deep Learning-Assisted Spectrum–Structure Correlation: State-of-the-Art and Perspectives. *Anal. Chem.* (2024) doi:10.1021/acs.analchem.4c01639.

7. Nguyen, J., Overstreet, R., King, E. & Ciesielski, D. Advancing the Prediction of MS/MS Spectra Using Machine Learning. *J. Am. Soc. Mass Spectrom.* **35**, 2256–2266 (2024).

8. Ludwig, M., Dührkop, K. & Böcker, S. Bayesian networks for mass spectrometric metabolite identification via molecular fingerprints. *Bioinformatics* **34**, i333–i340 (2018).

9. Dührkop, K., Shen, H., Meusel, M., Rousu, J. & Böcker, S. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proc. Natl. Acad. Sci.* **112**, 12580–12585 (2015).

10. Böcker, S. & Rasche, F. Towards de novo identification of metabolites by analyzing tandem mass spectra. *Bioinformatics* **24**, i49–i55 (2008).

11. Heinonen, M., Shen, H., Zamboni, N. & Rousu, J. Metabolite identification and molecular fingerprint prediction through machine learning. *Bioinformatics* **28**, 2333–2341 (2012).

12. Yu, Y. & Li, M. Towards highly sensitive deep learning-based end-to-end database search for tandem mass spectrometry. *Nat. Mach. Intell.* **7**, 85–95 (2025).

13. Elser, D., Huber, F. & Gaquerel, E. Mass2SMILES: deep learning based fast prediction of structures and functional groups directly from high-resolution MS/MS spectra. *bioRxiv* 2023.07.06.547963 (2023) doi:10.1101/2023.07.06.547963.

14. Shrivastava, A. D. *et al.* MassGenie: A Transformer-Based Deep Learning Method for Identifying Small Molecules from Their Mass Spectra. *Biomolecules* **11**, 1793 (2021).

15. Stravs, M. A., Dührkop, K., Böcker, S. & Zamboni, N. MSNovelist: de novo structure generation from mass spectra. *Nat. Methods* **19**, 865–870 (2022).

16. Asher, G., Delmar, M. C., Campbell, J. M., Geremia, J. & Kassis, T. LSM1-MS2: A Foundation Model for MS/MS, Encompassing Chemical Property Predictions, Search and de novo Generation. (2024) doi:10.26434/chemrxiv-2024-k06gb-v3.

17. Bushuiev, R., Bushuiev, A., Samusevich, R., Šivic, J. & Pluskal, T. Emergence of molecular structures from self-supervised learning on mass spectra. (2023) doi:10.26434/chemrxiv-2023-kss3r.

18. Litsa, E. E., Chenthamarakshan, V., Das, P. & Kavraki, L. E. An end-to-end deep learning framework for translating mass spectra to de-novo molecules. *Commun. Chem.* **6**, 132 (2023).

19. Bushuiev, R. *et al.* MassSpecGym: A benchmark for the discovery and identification of molecules. *arXiv* (2024) doi:10.48550/arxiv.2410.23326.

20. Bagal, V., Aggarwal, R., Vinod, P. K. & Priyakumar, U. D. MolGPT: Molecular Generation Using a Transformer-Decoder Model. *J. Chem. Inf. Model.* **62**, 2064–2076 (2022).

21. Fang, Y. *et al.* Domain-Agnostic Molecular Generation with Chemical Feedback. *arXiv* (2023) doi:10.48550/arxiv.2301.11259.

22. Butler, T. *et al.* MS2Mol: A transformer model for illuminating dark chemical space from mass spectra. (2023) doi:10.26434/chemrxiv-2023-vsmpx-v4.

23. Young, A., Röst, H. & Wang, B. Tandem mass spectrum prediction for small molecules using graph transformers. *Nat. Mach. Intell.* 1–13 (2024) doi:10.1038/s42256-024-00816-8.

24. Goldman, S., Bradshaw, J., Xin, J. & Coley, C. W. Prefix-Tree Decoding for Predicting Mass Spectra from Molecules. *arXiv* (2023) doi:10.48550/arxiv.2303.06470.

25. Goldman, S., Li, J. & Coley, C. W. Generating Molecular Fragmentation Graphs with Autoregressive Neural Networks. *Anal. Chem.* **96**, 3419–3428 (2024).

26. Young, A. *et al.* FraGNNet: A Deep Probabilistic Model for Mass Spectrum Prediction. *arXiv* (2024).

27. Murphy, M. *et al.* Efficiently predicting high resolution mass spectra with graph neural networks. *arXiv* (2023) doi:10.48550/arxiv.2301.11419.

28. Wei, J. N., Belanger, D., Adams, R. P. & Sculley, D. Rapid Prediction of Electron–Ionization Mass Spectrometry Using Neural Networks. *ACS Cent. Sci.* **5**, 700–708 (2019).

29. Wang, F. *et al.* CFM-ID 4.0: More Accurate ESI-MS/MS Spectral Prediction and Compound Identification. *Anal. Chem.* **93**, 11692–11700 (2021).

30. Wang, M. *et al.* Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* **34**, 828–837 (2016).

31. Horai, H. *et al.* MassBank: a public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.* **45**, 703–714 (2010).

32. Gaulton, A. *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40**, D1100–D1107 (2012).

33. Kim, S. *et al.* PubChem 2025 update. *Nucleic Acids Res.* **53**, D1516–D1525 (2024).

34. Irwin, J. J. & Shoichet, B. K. ZINC − A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **45**, 177–182 (2005).

35. Xing, S., Shen, S., Xu, B., Li, X. & Huan, T. BUDDY: molecular formula discovery via bottom-up MS/MS interrogation. *Nat. Methods* **20**, 881–890 (2023).

36. Dührkop, K. *et al.* SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat. Methods* **16**, 299–302 (2019).

37. Pluskal, T., Uehara, T. & Yanagida, M. Highly Accurate Chemical Formula Prediction Tool Utilizing High-Resolution Mass Spectra, MS/MS Fragmentation, Heuristic Rules, and Isotope Pattern Matching. *Anal. Chem.* **84**, 4396–4403 (2012).

38. Kind, T. & Fiehn, O. Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinform.* **8**, 105 (2007).

39. Critical Assessment of Small Molecule Identification. http://www.casmi-contest.org/2022/index.shtml (2022).

40. MS2Mol: A transformer model for illuminating dark chemical space from mass spectra. doi:10.26434/chemrxiv-2023-vsmpx.

41. Wang, Y., Chen, X., Liu, L. & Hassoun, S. MADGEN: Mass-Spec attends to De Novo Molecular generation. *arXiv* (2025) doi:10.48550/arxiv.2501.01950.

42. Daylight Theory: SMARTS - A Language for Describing Molecular Patterns. https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html.

43. Dührkop, K., Shen, H., Meusel, M., Rousu, J. & Böcker, S. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proc. Natl. Acad. Sci.* **112**, 12580–12585 (2015).

44. Dührkop, K. *et al.* SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat. Methods* **16**, 299–302 (2019).

45. Krenn, M., Hse, F., Nigam, A., Friederich, P. & Aspuru-Guzik, A. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Mach. Learn.: Sci. Technol.* **1**, 045024 (2020).

46. O'Boyle, N. & Dalke, A. DeepSMILES: An Adaptation of SMILES for Use in Machine-Learning of Chemical Structures. (2018) doi:10.26434/chemrxiv.7097960.v1.

47. Long, L., Li, R. & Zhang, J. Artificial Intelligence in Retrosynthesis Prediction and its Applications in Medicinal Chemistry. *J. Med. Chem.* (2025) doi:10.1021/acs.jmedchem.4c02749.

48. Lowe, D. Chemical reactions from US patents. https://figshare.com/articles/dataset/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873/1 (2017).

49. Lowe. Extraction of chemical structures and reactions from the literature. https://doi.org/10.17863/CAM.16293 (2012).

50. Schwaller, P., Vaucher, A. C., Laino, T. & Reymond, J.-L. Prediction of chemical reaction yields using deep learning. *Mach. Learn.: Sci. Technol.* **2**, 015016 (2021).

51. Coley, C. W., Green, W. H. & Jensen, K. F. RDChiral: An RDKit Wrapper for Handling Stereochemistry in Retrosynthetic Template Extraction and Application. *J. Chem. Inf. Model.* **59**, 2529–2537 (2019).

52. RDKit: Open-source cheminformatics. https://www.rdkit.org.

53. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv* (2018) doi:10.48550/arxiv.1802.03426.