Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Digital Discovery*, 2025, **4**, 2454

Enhancing automated drug substance impurity structure elucidation from tandem mass spectra through transfer learning and domain knowledge†

Emilio Dorigatti, [a] Jonathan Groß, [b] Jonas Kühlborn,[b] Robert Möckel, [b] Frank Maier*[a] and Julian Keupp *[b]

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 limits its use in practice. In this paper, we introduce a new method called SEISMiQ 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 using 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.

Received 21st March 2025
Accepted 17th July 2025

DOI: 10.1039/d5dd00115c

rsc.li/digitaldiscovery

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 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 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

[a]*Development NCE, Analytical Development, Boehringer Ingelheim Pharma GmbH & Co. KG, D-55218, Ingelheim (Rhein), Germany. E-mail: frank.maier@boehringer-ingelheim.com*

[b]*Development NCE, Chemical Development, Boehringer Ingelheim Pharma GmbH & Co. KG, D-55218, Ingelheim (Rhein), Germany. E-mail: julian.keupp@boehringer-ingelheim.com*

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d5dd00115c

starting materials and their impurities, the conditions under which reactions take place, possible unwanted side reactions, over reaction and others (Fig. 1a). This information provides substantial insights about the potential impurity structure, including for example 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, we introduce a novel method which we call SEISMiQ for elucidating 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 (Fig. 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 for finetuning (Fig. 1c). To facilitate future research in this area, we open source our implementation including training code, data, and pretrained checkpoints at the following link: **https://www.github.com/Boehringer-Ingelheim/seismiq**.

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[29] representation (Fig. 1b), similar to other works in the field.[13–15,22]

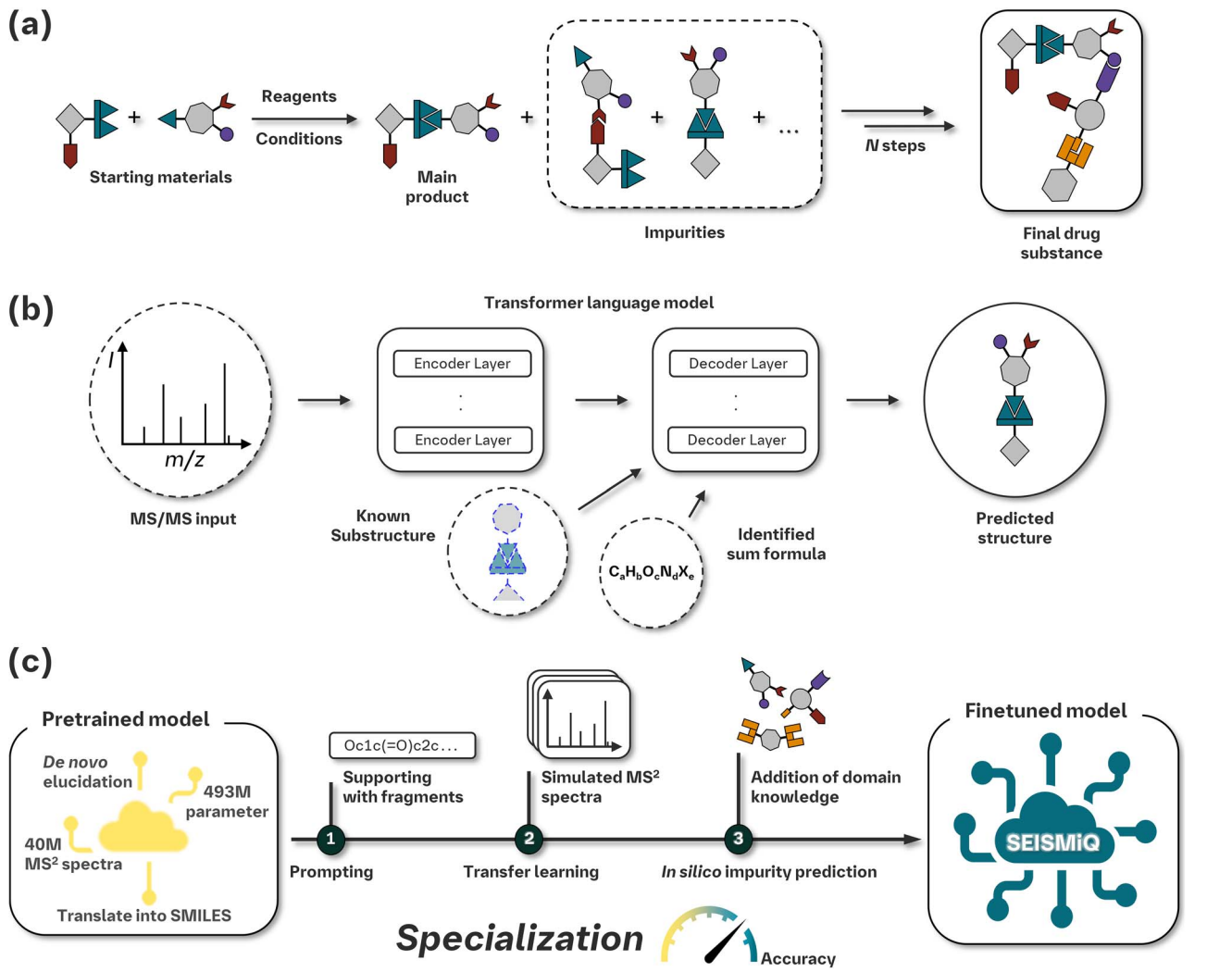

Fig. 1 Motivation and approach of SEISMiQ. (a) Unwanted impurities are formed during the synthesis of a drug substance. Elucidating their structure from MS/MS spectra is however very time consuming. (b) We developed a transformer language model for molecular structure elucidation from MS/MS spectra that 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.

Our model is a transformer language model trained on approximately 40 M MS/MS spectra for about 10 M molecules, most of which simulated *via* CFM-ID 4.0 (ref. 30) and FragGenie.[14] By design, our model requires a sum formula as input, since in the context of drug substance impurity elucidation it is already known, or easily predicted from the MS and MS/MS spectra.[31–34]

We evaluated our model on the Critical Assessment for Small Molecule Identification (CASMI) challenges,[35] as well as the newly released MassSpecGym benchmark[19] (Fig. 2). Our model achieved top-128 accuracies of 76.4%, 43.8% and 33.3% respectively for CASMI 2016, 2017 and 2022, and top-5 accuracies of 72.8%, 35.8% and 27.9% when the predictions were ranked using CSI:FingerID[36,37] scores. ESI S1† reports top-*k* performance for different values of *k* and ranking measures; for the remainder of this paper, we report top-128 performance of our model. 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 used for the evaluation only molecules that were not already in the model's training set. This resulted in 992 MS/MS spectra on which our model reached a top-128 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 (ESI 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[22] 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[38] 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.

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 (ESI S3† reports data collection standards). On this dataset, our model correctly elucidated only nine (5%) of the impurities, while providing predictions with Tanimoto (computed using the RDKit[39] fingerprint algorithm with default settings of 2048 bits, paths of length between one and seven bonds, excluding hydrogen atoms) 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. This knowledge can be used in conjunction with the predicted probabilities of the next SMILES token to navigate the model's predictions and exploring alternative possible structures (ESI S4†). Furthermore, in our experience, three quarters

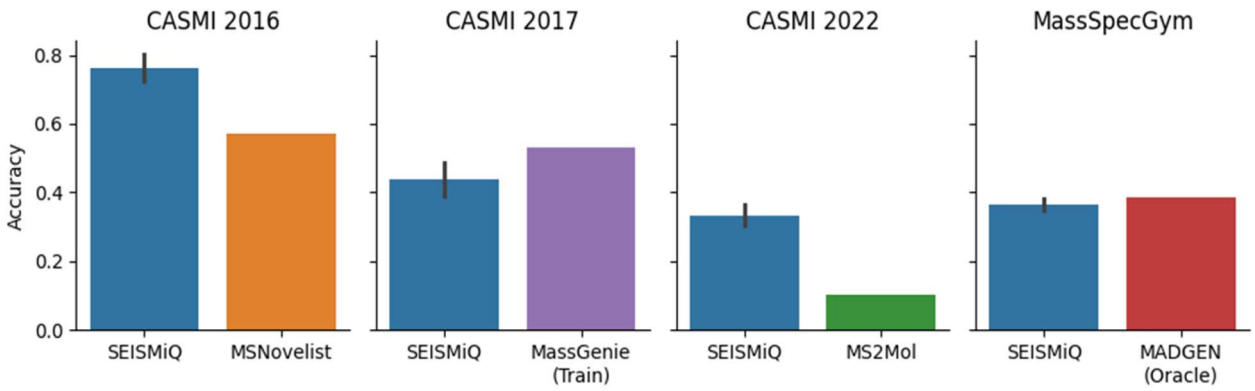

Fig. 2 *De novo* top-128 elucidation accuracy of SEISMiQ on CASMI 2016, 2017, 2022 and MassSpecGym compared to the published performance of other models. MADGEN (Oracle) was given the correct molecule scaffold to complete, and MassGenie was only evaluated on a subset of CASMI 2017 which was part of its own training set (see main text for differences in reporting standards). Error bars represent the standard error of the mean.

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 (Fig. 1b). To do this, we construct a SMILES string of the common fragment such that the last position in the string corresponds to the attachment point between the fragment and the impurity site of variation. We then let the model complete this SMILES string, thereby generating the remaining structure of the impurity conditioned on the known fragment and relative attachment point (Fig. 3a). 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.

On the public test datasets, this resulted in 48 628 fragments with an average of 27 and a maximum of 70 missing atoms (Fig. 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 (Fig. 3c). Nonetheless, for these fragments the average Tanimoto of the predicted molecules was 0.82 (Fig. 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] ESI S5† reports confidence intervals and standard errors for all accuracies.

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 (Fig. 4a). The performance of the latter model indicates to what extent it is possible to generalize from simulated to experimental MS/MS

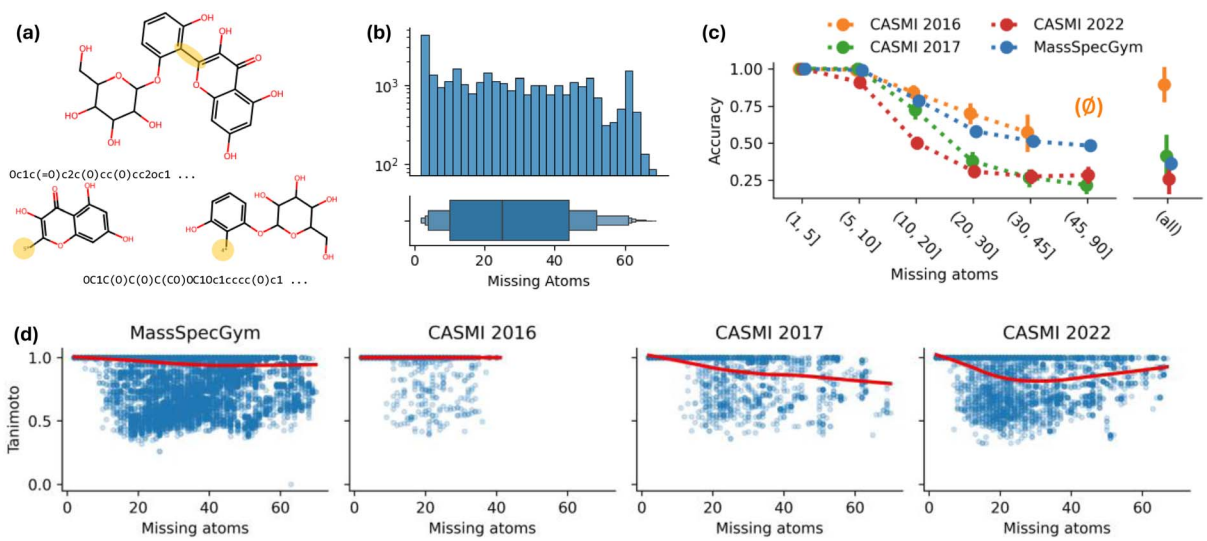

Fig. 3 Prompting the model with impurity fragments to boost elucidation accuracy. (a) Breaking a molecule (top) on the highlighted bond generates two fragments (bottom), with their SMILES representation created so that the attachment atom to the other fragment (highlighted) is in last position of the string. This constitutes the prompt to the model, which completes it into a fully-formed SMILES string. (b) Distribution of the number of missing atoms for each fragment. Error bars represent the standard error of the mean estimator. (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.

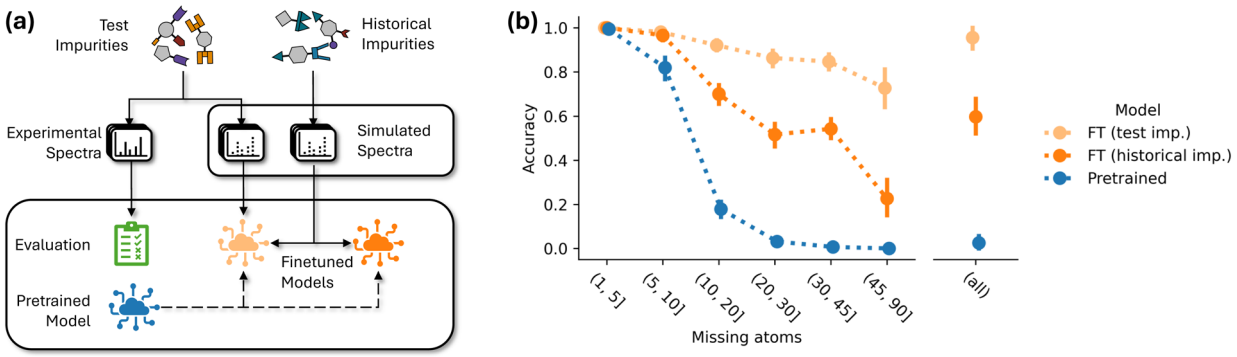

Fig. 4 Transfer learning ability from simulated to experimental spectra. (a) Experimental protocol: we used simulated spectra to create two finetuned models, one using the historical impurities and one using historical and test impurities and evaluated all models on the experimental spectra of the test 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). Error bars represent the standard error of the mean estimator.

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.

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 (Fig. 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. In ESI S6† we analyze training and validation curves for pretrained and finetuned models, including a comparison between the pretrained model and a model finetuned on CASMI itself. In ESI S7† we perform a quantitative evaluation of the similarity between simulated and experimental spectra.

These results 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 for 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.

We developed an impurity predictor based on SMARTS reaction templates,[40] describing how the products in a chemical reaction are formed by combining fragments of the starting materials (Fig. 5a and b). We integrated data from an internal electronic laboratory notebook reporting performed reactions with the corresponding starting materials and analytically detected impurities (Fig. 5c) allowing us to cover both the desired reactions forming the main compound and additional processes that generated the detected impurities. For this approach, only the starting materials and respective product(s) are needed, while reagents as well as reaction conditions can be neglected. After performing a sanitization check with RDKit, the template was extracted using the RDChiral package.[41] We did not filter templates by score since the formation of impurities in production batches can sometimes be difficult to explain using traditional organic transformation rules and knowledge. Using these reaction templates, we could reproduce the synthesis route of an asset by iteratively applying all templates to the starting materials of each chemical step as well as all products resulting from the previous steps. Known downstream impurities that were not predicted can be covered by manually adding the structure of interest to the inputs of the respective step. This procedure resulted in a dataset of impurities for the complete manufacturing process of an asset that is entirely simulated from first principles only based on the knowledge of the synthesis route (Fig. 5d, sample SMARTS templates as well as

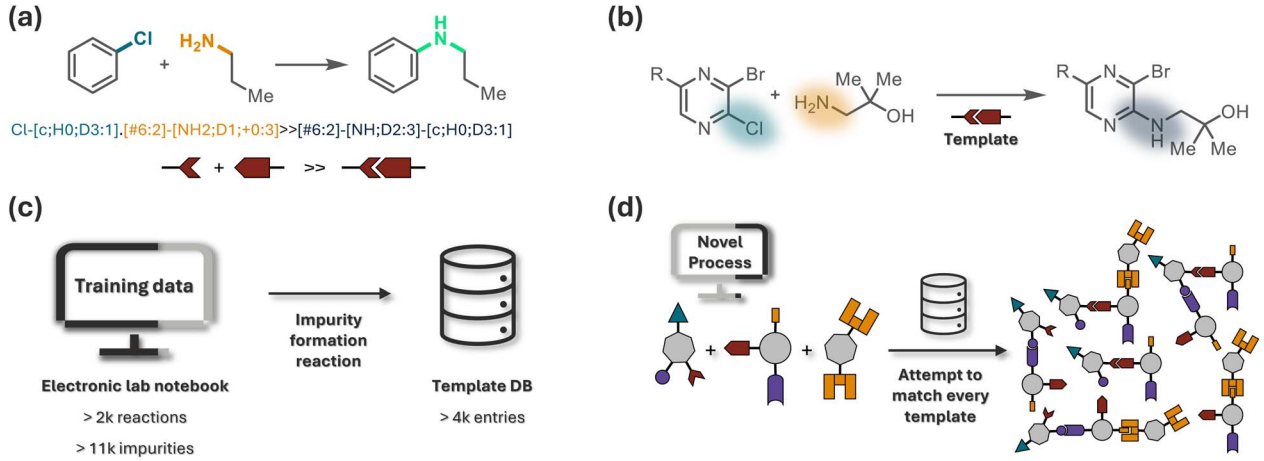

Fig. 5 *In silico* prediction of impurities generated by a reaction. (a) SMARTS templates are transformation rules that describe how the reactive sites in the starting materials are combined to form the main product. These templates can be extracted automatically given an example reaction. (b) The same template can be applied to a variety of starting materials that present the required reactive sites. (c) We extracted over 4000 reaction templates from our internal electronic lab notebook containing data for more than 2000 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.

synthetic impurity datasets based on public templates can be found in ESI S7 and S8†).

We focused on the synthesis route of an internal asset from Boehringer Ingelheim's development pipeline. This molecule consists of 41 non-hydrogen atoms, possesses a molar weight of *ca.* 600 g mol^{-1} 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. While we cannot disclose the chemical structure of this API, we believe that it constitutes a challenging test case for our methodology that is representative of the complexity and variety of real world new chemical entities under active development in the pharmaceutical industry. We excluded from the template extraction procedure all reactions reported in the electronic laboratory notebook that were part of the test asset synthesis route, leaving us with 4446 templates in total (summary statistics can be found in ESI S10†). Their application resulted in 20 813 simulated impurities with mass below 1200 Da, and 154 756 corresponding simulated mass spectra. Our test dataset contained 61 experimentally-detected impurities related to this asset, and the impurity generation procedure correctly predicted 27 of these 61 impurities. In general, the chemical space covered by the simulated impurities included close matches for all experimental impurities (Fig. 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.

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 (Fig. 6b). Nonetheless, when averaging the model's performance across fragments of all sizes, the model finetuned on simulated data had 5.6% higher accuracy (88.7% *vs.* 83.1%).

The results in this section show that an entirely *in silico* simulation approach of process impurities and their MS/MS spectra can compensate, without loss of accuracy, the absence of relevant experimental data for finetuning, and result in a model with significantly higher performance compared to a model pretrained only on public data.

Discussion

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

We explored three ways of dealing with this challenge. First, we employed data augmentation at training time teaching the model to complete a user-provided 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 considerable gains in elucidation accuracy. Second, we finetuned the model using an internal dataset of historical, experimentally detected impurities. We showed that our model can successfully transfer knowledge from the corresponding simulated MS/MS spectra further boosting elucidation accuracy both in a *de novo* and in a fragment

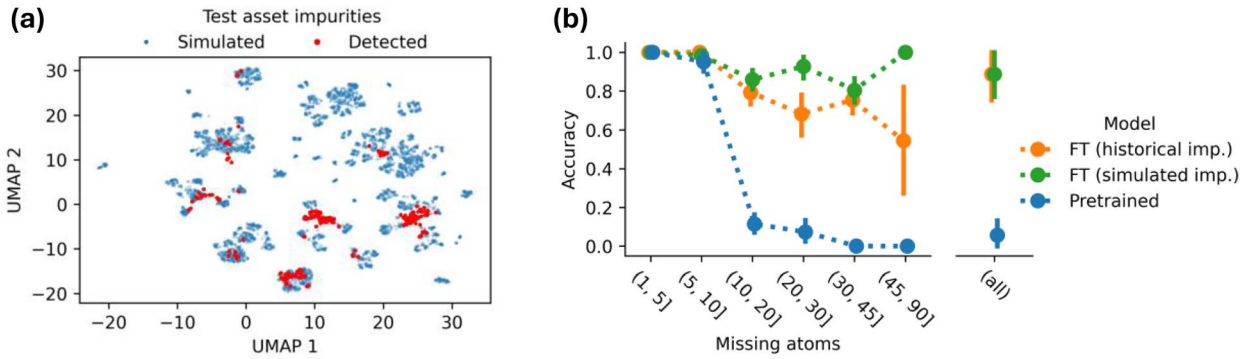

Fig. 6 Results of impurity simulation and model finetuning. (a) UMAP[42] visualization of the simulated (blue) and experimentally detected (red) test asset impurities. (b) Accuracy (*y* axis) on the fragment completion task for the test asset impurities as a function of missing atoms to be completed (*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). Error bars represent the standard error of the mean estimator.

completion setting. Third, considering the significant time and monetary investments required to obtain such a dataset, we simulated the entire synthesis process of an asset and predicted the impurities that are likely to be generated in the real world. We found that finetuning exclusively on these simulated impurities and their simulated MS/MS spectra resulted in a model that is slightly more accurate than a model finetuned on experimental data, thus enabling accurate elucidation of impurities without the necessity of any prior experimental measurement.

Our work is not free from limitations. First, our finetuning protocol introduced some overfitting to the dataset used for finetuning, despite our use of common overfitting mitigation strategies[43,44] including dropout, weight decay, and data augmentation. The danger of overfitting to small datasets is to some extent unavoidable[45–47] and in the context of a transfer learning setup like ours still constitutes a fertile ground for current research[48–54] with no accepted best practices.[45,55] For our model, the risk of overfitting can be reduced by generating more relevant synthetic data for finetuning, for example by leveraging our impurity simulation approach and by increasing the variety of MS/MS spectra predictors. Furthermore, the danger of unreliable predictions could be identified at inference time by comparing the input MS/MS spectrum with the finetuning dataset for example *via* CSI:FingerID[36,37] fingerprints, using the pretrained model as a fallback, or by employing uncertainty quantification techniques making the model more robust to out-of-distribution inputs.[48–52,56,57] Second, our method in its current form assumes a single attachment point when completing fragments; it is not uncommon however for impurities to differ from the main compound in more than one location. Such cases cannot be encoded into a single prompt that allows our model, as presently trained, to predict all attached fragments in a single shot. This task in its general form is known as scaffold completion, and several recent works explore adapting chemical language models like ours for this kind of predictions.[58–60] At present, this limitation can be circumvented simply by filtering out the model's predictions that do not conform to the known fragment or *via* a custom fragment-aware beam search sampling procedure. Third, the data used to train our model was composed of metabolites and small drug-like molecules, making it inappropriate to elucidate larger molecules such as peptides. Furthermore, the accuracy of the MS/MS simulation approaches employed to generate the training spectra could also limit our model's performance on some compound classes and certain adduct types.[5] Nonetheless, we did not find significant differences between the accuracy of the MS/MS predictions for our internal test dataset and the CASMI challenges (ESI S5†), suggesting that the models we employed are equally applicable to metabolites and drug substance impurities.

In conclusion, we achieved substantial advancements in the *de novo* elucidation of impurities from MS/MS spectra by considering the unique aspects of impurity generation and embedding them into the model's training and inference procedures. In the pharmaceutical industry impurity characterization is essential for optimizing manufacturing processes, understanding degradation pathways, ensuring drug substance stability, maintaining quality control, and achieving regulatory compliance. This work represents a significant step forward as we show how to increase the accuracy of a weak baseline into a model that is practically useful to assist analytical chemists in daily production workflows. By incorporating impurity elucidation earlier in the process, we strive to alleviate the workload of analytical chemists and facilitate semi-automated elucidation workflows, ultimately enhancing the efficiency and mitigating the cost of drug substance development, a process that typically spans five to ten years and costs upwards of 1 billion USD.[61,62]

Methods

Dataset

We used publicly available positive ion mode mass spectra from the GNPS,[63] MassBank[64,65] and MatchMS[66,67] online libraries, including spectra from HMDB,[68] SUMNER,[69] and MoNA.[70] 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,[71] PubChem[72] and ZINC[73] using the CFM-ID[30] 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 41,058,643 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. 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 37 (2187) 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[74] and DeepSMILES,[75] 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 time, and a randomized atom order is used in the other 75% of the time. Regardless of the order, 50% of the time we kekulized the SMILES, and removed all stereochemistry information. The SMILES tokens were encoded with a learnable embedding in addition to a sinusoidal encoding for the token position. In addition, 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]

Model

The model was a transformer[76] with 16 encoder and 16 decoder layers, each with 16 heads, with hidden dimension of 1024 and feed-forward dimension of 4096. Multilayer perceptrons (MLPs) were used to transform the peak masses, the SMILES tokens concatenated with the number of remaining atoms to be generated and to predict the following tokens. All these MLPs used one hidden layer of 2048 units and the rectifier activation function (ReLU[77]). This architecture resulted in 493 M trainable parameters.

The model was trained using the cross-entropy loss with a label smoothing of 0.1, and re-weighted samples to correct for over- and under-represented molecules in the training dataset. We used the AdamW optimizer[78] with learning rate of 3×10^{-5}, linearly annealed from 6×10^{-7} over the first 1000 training steps. A dropout of 0.2 and weight decay of 1×10^{-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 23 M 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 prespecified 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.[79] Our approach follows template-based rules in the SMARTS format (SMILES arbitrary target specification)[40] for describing chemical transformations. Compared to publicly available templates, *e.g.* the widely used USPTO dataset,[80] we envision that the utilization of our company internal impurity formation knowledge should boost the impurity prediction capabilities specifically for our research area.[81]

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.[82] Subsequently, we extracted a reaction template for this hypothetical reaction leading to an observed impurity applying the RDChiral library.[41] Overall, this approach resulted in 4446 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[39] over two rounds of generation.

Data availability

The published code contains facilities to download and process the public datasets used to train and evaluate the model. A snapshot of the code and data at time of publication is also archived on Zenodo at **https://doi.org/10.5281/zenodo.15790301**. The internal test data will not be published as it contains proprietary assets under active development and impurities thereof.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

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

References

1 N. Rahman, S. N. H. Azmi and H.-F. Wu, *Accredit. Qual. Assur.*, 2006, **11**, 69–74.
2 F. Qiu and D. L. Norwood, *J. Liq. Chromatogr. Relat. Technol.*, 2007, **30**, 877–935.
3 T. D. Vijlder, D. Valkenborg, F. Lemiére, E. P. Romijn, K. Laukens and F. Cuyckens, *Mass Spectrom. Rev.*, 2018, **37**, 607–629.
4 A. G. Beck, M. Muhoberac, C. E. Randolph, C. H. Beveridge, P. R. Wijewardhane, H. I. Kenttämaa and G. Chopra, *ACS Meas. Sci. Au*, 2024, **4**, 233–246.
5 Y. Liu, T. D. Vijlder, W. Bittremieux, K. Laukens and W. Heyndrickx, *Rapid Commun. Mass Spectrom.*, 2021, e9120.
6 X.-Y. Lu, H.-P. Wu, H. Ma, H. Li, J. Li, Y.-T. Liu, Z.-Y. Pan, Y. Xie, L. Wang, B. Ren and G.-K. Liu, *Anal. Chem.*, 2024, **96**, 7959–7975.
7 J. Nguyen, R. Overstreet, E. King and D. Ciesielski, *J. Am. Soc. Mass Spectrom.*, 2024, **35**, 2256–2266.
8 M. Ludwig, K. Dührkop and S. Böcker, *Bioinformatics*, 2018, **34**, i333–i340.
9 K. Dührkop, H. Shen, M. Meusel, J. Rousu and S. Böcker, *Proc. Natl. Acad. Sci. U. S. A.*, 2015, **112**, 12580–12585.
10 S. Böcker and F. Rasche, *Bioinformatics*, 2008, **24**, i49–i55.
11 M. Heinonen, H. Shen, N. Zamboni and J. Rousu, *Bioinformatics*, 2012, **28**, 2333–2341.
12 Y. Yu and M. Li, *Nat. Mach. Intell.*, 2025, **7**, 85–95.
13 D. Elser, F. Huber and E. Gaquerel, *bioRxiv*, 2023, preprint, DOI: **10.1101/2023.07.06.547963**.
14 A. D. Shrivastava, N. Swainston, S. Samanta, I. Roberts, M. W. Muelas and D. B. Kell, *Biomolecules*, 2021, **11**, 1793.
15 M. A. Stravs, K. Dührkop, S. Böcker and N. Zamboni, *Nat. Methods*, 2022, **19**, 865–870.
16 G. Asher, M. Cadosh Delmar, J. M. Campbell, J. Geremia and T. Kassis, *ChemRxiv*, 2024, preprint, DOI: **10.26434/chemrxiv-2024-k06gb-v3**.
17 R. Bushuiev, A. Bushuiev, R. Samusevich, C. Brungs, J. Sivic and T. Pluskal, *Nat. Biotechnol.*, 2025.
18 E. E. Litsa, V. Chenthamarakshan, P. Das and L. E. Kavraki, *Commun. Chem.*, 2023, **6**, 132.
19 R. Bushuiev, A. Bushuiev, N. F. de Jonge, A. Young, F. Kretschmer, R. Samusevich, J. Heirman, F. Wang, L. Zhang, K. Dührkop, M. Ludwig, N. A. Haupt, A. Kalia, C. Brungs, R. Schmid, R. Greiner, B. Wang, D. S. Wishart, L.-P. Liu, J. Rousu, W. Bittremieux, H. Rost, T. D. Mak, S. Hassoun, F. Huber, J. J. van der Hooft, M. A. Stravs, S. Böcker, J. Sivic and T. Pluskal, *Adv. Neural Inf. Process Syst.*, 2024, 110010–110027.
20 V. Bagal, R. Aggarwal, P. K. Vinod and U. D. Priyakumar, *J. Chem. Inf. Model.*, 2022, **62**, 2064–2076.
21 Y. Fang, N. Zhang, Z. Chen, X. Fan, H. Chen, *The Twelfth International Conference on Learning Representations*, 2024, **https://openreview.net/forum?id=9rPyHyjfwP**.
22 T. Butler, A. Frandsen, R. Lightheart, B. Bargh, T. Bollerman, T. Kerby, K. West, G. Voronov, K. Moon, T. Kind, P. Dorrestein, A. Allen, V. Colluru and D. Healey, *ChemRxiv*, 2023, DOI: **10.26434/chemrxiv-2023-vsmpx**.
23 A. Young, H. Röst and B. Wang, *Nat. Mach. Intell.*, 2024, 1–13.
24 S. Goldman, J. Bradshaw, J. Xin and C. Coley, *Adv. Neural Inf. Process Syst.*, 2023, 48548–48572.
25 S. Goldman, J. Li and C. W. Coley, *Anal. Chem.*, 2024, **96**, 3419–3428.
26 A. Young, F. Wang, D. Wishart, B. Wang, H. Röst and R. Greiner, *arXiv*, 2024, DOI: **10.48550/arXiv.2404.02360**.
27 M. Murphy, S. Jegelka, E. Fraenkel, T. Kind, D. Healey and T. Butler, *Proceedings of the 40th International Conference on Machine Learning*, 2023.
28 J. N. Wei, D. Belanger, R. P. Adams and D. Sculley, *ACS Cent. Sci.*, 2019, **5**, 700–708.
29 D. Weininger, *J. Chem. Inf. Comput. Sci.*, 1988, **28**, 31–36.
30 F. Wang, J. Liigand, S. Tian, D. Arndt, R. Greiner and D. S. Wishart, *Anal. Chem.*, 2021, **93**, 11692–11700.
31 S. Xing, S. Shen, B. Xu, X. Li and T. Huan, *Nat. Methods*, 2023, **20**, 881–890.
32 K. Dührkop, M. Fleischauer, M. Ludwig, A. A. Aksenov, A. V. Melnik, M. Meusel, P. C. Dorrestein, J. Rousu and S. Böcker, *Nat. Methods*, 2019, **16**, 299–302.
33 T. Pluskal, T. Uehara and M. Yanagida, *Anal. Chem.*, 2012, **84**, 4396–4403.
34 T. Kind and O. Fiehn, *BMC Bioinf.*, 2007, **8**, 105.
35 Critical Assessment of Small Molecule Identification, 2022, **http://www.casmi-contest.org/2022/index.shtml**.
36 K. Dührkop, H. Shen, M. Meusel, J. Rousu and S. Böcker, *Proc. Natl. Acad. Sci. U. S. A.*, 2015, **112**, 12580–12585.

37 K. Dührkop, M. Fleischauer, M. Ludwig, A. A. Aksenov, A. V. Melnik, M. Meusel, P. C. Dorrestein, J. Rousu and S. Böcker, *Nat. Methods*, 2019, **16**, 299–302.
38 Y. Wang, X. Chen, L. Liu and S. Hassoun, *The Thirteenth International Conference on Learning Representations*, 2025.
39 G. Landrum, P. Tosco, B. Kelley, Ric, D. Cosgrove, sriniker, R. Vianello, gedeck, NadineSchneider, G. Jones, E. Kawashima, D. N. A. Dalke, B. Cole, M. Swain, S. Turk, A. Savelev, A. Vaucher, M. Wójcikowski, I. Take, V. F. Scalfani, D. Probst, K. Ujihara, G. godin, A. Pahl, R. Walker, J. Lehtivarjo and F. Berenger, strets123 and jasondbiggs, rdkit/rdkit: Release 2023.09.5, 2024, **https://zenodo.org/doi/10.5281/zenodo.10633624**.
40 Daylight Theory: SMARTS - A Language for Describing Molecular Patterns, **https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html**.
41 C. W. Coley, W. H. Green and K. F. Jensen, *J. Chem. Inf. Model.*, 2019, **59**, 2529–2537.
42 L. McInnes, J. Healy and J. Melville, *arXiv*, 2018.
43 R. Moradi, R. Berangi and B. Minaei, *Artif. Intell. Rev.*, 2020, **53**, 3947–3986.
44 Y. Tian and Y. Zhang, *Information Fusion*, 2022, **80**, 146–166.
45 B. Zhang, Z. Liu, C. Cherry and O. Firat, *arXiv*, 2024, DOI: **10.48550/arxiv.2402.17193**.
46 C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li and P. J. Liu, *J. Mach. Learn. Res.*, 2020, **21**, 5485–5551.
47 D. Hernandez, J. Kaplan, T. Henighan and S. McCandlish, *arXiv*, 2021, DOI: **10.48550/arxiv.2102.01293**.
48 Y. Lin, X. Ma, X. Chu, Y. Jin, Z. Yang, Y. Wang and H. Mei, *arXiv*, 2024, DOI: **10.48550/arxiv.2404.09610**.
49 X.-K. Wu, M. Chen, W. Li, R. Wang, L. Lu, J. Liu, K. Hwang, Y. Hao, Y. Pan, Q. Meng, K. Huang, L. Hu, M. Guizani, N. Chao, G. Fortino, F. Lin, Y. Tian, D. Niyato and F.-Y. Wang, *Big Data Cogn. Comput.*, 2025, **9**, 87.
50 W. Zhang and I. Vaidya, *arXiv*, 2021, DOI: **10.48550/arxiv.2102.11402**.
51 S. Huang, D. Xu, I. E. H. Yen, Y. Wang, S.-e. Chang, B. Li, S. Chen, M. Xie, S. Rajasekaran, H. Liu and C. Ding, *arXiv*, 2021, DOI: **10.48550/arxiv.2110.08190**.
52 K. You, Z. Kou, M. Long and J. Wang, *Adv. Neural Inf. Process Syst.*, 2020, 17236–17246.
53 X. Chen, S. Wang, B. Fu, M. Long and J. Wang, *Adv. Neural Inf. Process Syst.*, 2019, 1906–1916.
54 G. Vrbančič and V. Podgorelec, *IEEE Access*, 2020, **8**, 196197–196211.
55 W. M. Kouw and M. Loog, *arXiv*, 2018, DOI: **10.48550/arxiv.1812.11806**.
56 X. Li, Y. Dai, Y. Ge, J. Liu, Y. Shan and L.-Y. Duan, *arXiv*, 2022, DOI: **10.48550/arxiv.2202.03958**.
57 M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, V. Makarenkov and S. Nahavandi, *Information Fusion*, 2021, **76**, 243–297.
58 J. Arús-Pous, A. Patronov, E. J. Bjerrum, C. Tyrchan, J.-L. Reymond, H. Chen and O. Engkvist, *J. Cheminf.*, 2020, **12**, 38.
59 M. Thomas, M. Ahmad, G. Tresadern and G. d. Fabritiis, *J. Cheminf.*, 2024, **16**, 77.
60 E. Noutahi, C. Gabellini, M. Craig, J. S. C. Lim and P. Tossou, *Digital Discovery*, 2024, **3**, 796–a04.
61 M. Schlander, K. Hernandez-Villafuerte, C.-Y. Cheng, J. Mestre-Ferrandiz and M. Baumann, *PharmacoEconomics*, 2021, **39**, 1243–1269.
62 S. Simoens and I. Huys, *Front Med.*, 2021, **8**, 760762.
63 M. Wang, J. J. Carver, V. V. Phelan, L. M. Sanchez, N. Garg, Y. Peng, D. D. Nguyen, J. Watrous, C. A. Kapono, T. Luzzatto-Knaan, C. Porto, A. Bouslimani, A. V. Melnik, M. J. Meehan, W.-T. Liu, M. Crüsemann, P. D. Boudreau, E. Esquenazi, M. Sandoval-Calderón, R. D. Kersten, L. A. Pace, R. A. Quinn, K. R. Duncan, C.-C. Hsu, D. J. Floros, R. G. Gavilan, K. Kleigrewe, T. Northen, R. J. Dutton, D. Parrot, E. E. Carlson, B. Aigle, C. F. Michelsen, L. Jelsbak, C. Sohlenkamp, P. Pevzner, A. Edlund, J. McLean, J. Piel, B. T. Murphy, L. Gerwick, C.-C. Liaw, Y.-L. Yang, H.-U. Humpf, M. Maansson, R. A. Keyzers, A. C. Sims, A. R. Johnson, A. M. Sidebottom, B. E. Sedio, A. Klitgaard, C. B. Larson, C. A. Boya P, D. Torres-Mendoza, D. J. Gonzalez, D. B. Silva, L. M. Marques, D. P. Demarque, E. Pociute, E. C. O'Neill, E. Briand, E. J. N. Helfrich, E. A. Granatosky, E. Glukhov, F. Ryffel, H. Houson, H. Mohimani, J. J. Kharbush, Y. Zeng, J. A. Vorholt, K. L. Kurita, P. Charusanti, K. L. McPhail, K. F. Nielsen, L. Vuong, M. Elfeki, M. F. Traxler, N. Engene, N. Koyama, O. B. Vining, R. Baric, R. R. Silva, S. J. Mascuch, S. Tomasi, S. Jenkins, V. Macherla, T. Hoffman, V. Agarwal, P. G. Williams, J. Dai, R. Neupane, J. Gurr, A. M. C. Rodríguez, A. Lamsa, C. Zhang, K. Dorrestein, B. M. Duggan, J. Almaliti, P.-M. Allard, P. Phapale, L.-F. Nothias, T. Alexandrov, M. Litaudon, J.-L. Wolfender, J. E. Kyle, T. O. Metz, T. Peryea, D.-T. Nguyen, D. VanLeer, P. Shinn, A. Jadhav, R. Müller, K. M. Waters, W. Shi, X. Liu, L. Zhang, R. Knight, P. R. Jensen, B. Ø. Palsson, K. Pogliano, R. G. Linington, M. Gutiérrez, N. P. Lopes, W. H. Gerwick, B. S. Moore, P. C. Dorrestein and N. Bandeira, *Nat. Biotechnol.*, 2016, **34**, 828–837.
64 H. Horai, M. Arita, S. Kanaya, Y. Nihei, T. Ikeda, K. Suwa, Y. Ojima, K. Tanaka, S. Tanaka, K. Aoshima, Y. Oda, Y. Kakazu, M. Kusano, T. Tohge, F. Matsuda, Y. Sawada, M. Y. Hirai, H. Nakanishi, K. Ikeda, N. Akimoto, T. Maoka, H. Takahashi, T. Ara, N. Sakurai, H. Suzuki, D. Shibata, S. Neumann, T. Iida, K. Tanaka, K. Funatsu, F. Matsuura, T. Soga, R. Taguchi, K. Saito and T. Nishioka, *J. Mass Spectrom.*, 2010, **45**, 703–714.
65 MassBank consortium and its contributors, *MassBank/MassBank-data: Release version 2023.11*, 2023, **https://zenodo.org/doi/10.5281/zenodo.10213786**.
66 F. Huber, S. Verhoeven, C. Meijer, H. Spreeuw, E. Castilla, C. Geng, J. van der Hooft, S. Rogers, A. Belloum, F. Diblen and J. Spaaks, *J. Open Source Softw.*, 2020, **5**, 2411.
67 N. F. de Jonge, H. Hecht, M. Strobel, M. Wang, J. J. J. van der Hooft and F. Huber, *J. Cheminf.*, 2024, **16**, 88.

68 D. S. Wishart, A. Guo, E. Oler, F. Wang, A. Anjum, H. Peters, R. Dizon, Z. Sayeeda, S. Tian, B. L. Lee, *et al.*, *Nucleic Acids Res.*, 2022, **50**, D622–D631.
69 Z. Lei, L. Jing, F. Qiu, H. Zhang, D. Huhman, Z. Zhou and L. W. Sumner, *Anal. Chem.*, 2015, **87**, 7373–7381.
70 A. Vaniya, S. Mehta, G. Wohlgemuth and O. Fiehn, *Berichte aus dem* Julius Kühn-Institut, 2019.
71 A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani and J. P. Overington, *Nucleic Acids Res.*, 2012, **40**, D1100–D1107.
72 S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang and E. E. Bolton, *Nucleic Acids Res.*, 2024, **53**, D1516–D1525.
73 J. J. Irwin and B. K. Shoichet, *J. Chem. Inf. Model.*, 2005, **45**, 177–182.
74 M. Krenn, F. Hse, A. Nigam, P. Friederich and A. Aspuru-Guzik, *Machine Learning: Science and Technology*, 2020, **1**, 045024.
75 N. O'Boyle and A. Dalke, ChemRXiv, 2018, preprint, DOI: **10.26434/chemrxiv.7097960.v1**.
76 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser and I. Polosukhin, *Adv. Neural Inf. Process Syst.*, 2017, 5998–6008.
77 V. Nair and G. E. Hinton, *Proceedings of the 27th international conference on machine learning*, ICML-10, 2010, pp. 807–814.
78 I. Loshchilov and F. Hutter, *International Conference on Learning Representations*, 2019.
79 L. Long, R. Li and J. Zhang, *J. Med. Chem.*, 2025, 2333–2355.
80 D. Lowe, Chemical reactions from US patents, 2017, **https://figshare.com/articles/dataset/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873/1**.
81 Lowe, *Extraction of chemical structures and reactions from the literature*, 2012, DOI: **10.17863/CAM.16293**.
82 P. Schwaller, A. C. Vaucher, T. Laino and J.-L. Reymond, *Machine Learning: Science and Technology*, 2021, **2**, 015016.