# ReactionEye: Integrating GC–MS Data and Chemical Context for Multimodal Structure Elucidation in Reaction Screening

**Maik G. Niedziella** [1] **Philipp M. Pflüger** [1] **Frank Glorius** [1]

[1] *University of Münster, Institute of Organic Chemistry, Corrensstraße 40, 48149 Münster, Germany*. Correspondence to: Frank Glorius glorius@uni-muenster.de.

## 1. Introduction

Molecular structure elucidation is of major importance in organic chemistry and underpins the interpretation of analytical data in reaction development and discovery. In high-throughput experimentation (HTE), where large numbers of reactions are screened in parallel to efficiently explore broad chemical spaces, rapid and reliable analytical readout is essential. Consequently, liquid chromatography–electrospray ionization mass spectrometry (LC-ESI-MS) is most commonly employed in HTE screening campaigns due to its compatibility with large, polar, and non-volatile molecules and its straightforward coupling to automated workflows. For such compounds, the set of structures consistent with a given molecular mass is typically restricted by the reactants, and the probability of side products with identical mass is low, allowing confident assignment from accurate mass and chromatographic information alone. However, the structural information content of LC–ESI–MS spectra is often insufficient for unambiguous identification when screening reactions that yield low-molecular-weight or closely related by-products with overlapping or identical nominal masses. [1]

In these cases, gas chromatography–mass spectrometry (GC–MS) offers a decisive advantage by providing rich fragmentation patterns that encode detailed structural information, making it particularly well suited for the analysis of small molecules in reaction screening. However, this amount of information content comes at the cost of substantially more challenging data interpretation. Manual GC–MS structure elucidation is time-consuming and requires significant expertise, as chemists typically interpret fragmentation patterns in the context of known starting materials, comparing product spectra to those of reactants and reasoning about plausible structural transformations. As this contextual information is already available in reaction screening workflows, it naturally motivates a multimodal approach in which GC–MS spectra are interpreted jointly with chemical context, such as the Simplified Molecular Input Line Entry System (SMILES) representations of the reactants. Incorporating this complementary information has the potential to substantially reduce ambiguity in spectral interpretation and to alleviate structure elucidation as the dominant bottleneck in GC–MS-based screening. [2]

Modern HTE workflows, such as those enabled by pyGecko from our group, allow automated extraction and alignment of chromatographic peaks together with their associated mass spectra across large reaction sets. While this enables scalable and systematic analysis of GC–MS data, the assignment of chemical structures to detected peaks remains largely manual and limits the throughput of the overall workflow. [3]

## 2. Structure Elucidation Framework

To address this challenge, we present a transformer-based generative model for automated structure elucidation directly from GC–MS data. The model encodes mass spectral peak information and can optionally be conditioned on auxiliary chemical constraints, including molecular formulae and reactant context, to guide structure generation. Candidate molecular structures are generated autoregressively as SMILES strings, enabling prediction beyond fixed spectral libraries. In addition to full-structure prediction, the model assigns confidence scores to molecular substructures incorporated into the generated candidates, providing interpretable guidance that is particularly valuable for early-stage HTE screening. Applied to GC–MS data from HTE workflows, this approach enables scalable, library-independent structure elucidation and closes a critical gap between automated data acquisition and chemical decision-making.

Table 1: Results for the prediction of SMILES with given input features.

| | Spectrum + Reactants | Spectrum + Reactants + Sum Formula |
|---|---|---|
| *Real Spectra* / % | 77 | 86 |
| *Virtual Spectra* / % | 90 | 92 |

### 2.1 Related work

The prevailing strategy for compound identification from GC–MS data is library search, in which an experimental electron ionization (EI) spectrum is matched against large collections of reference spectra, such as those contained in the NIST mass spectral libraries. When reference spectra are available, this approach can yield highly reliable identifications. However, library-based methods fundamentally depend on database coverage and often fail when reaction screening produces novel compounds, regioisomers, or closely related structures that are absent or poorly represented in existing libraries. These limitations are particularly pronounced in reaction dis-

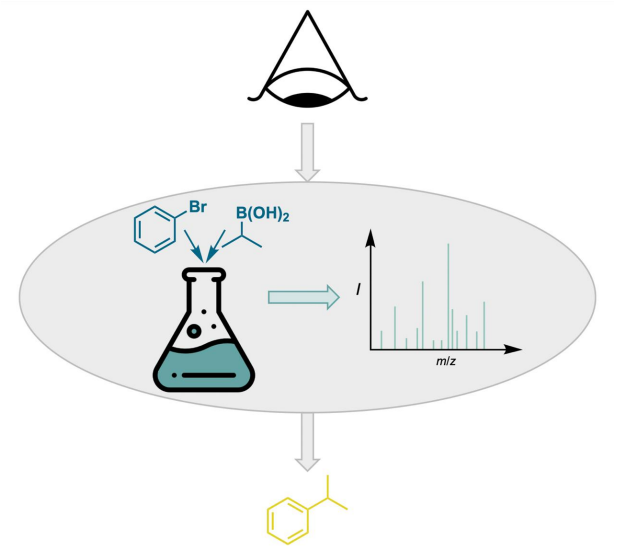


Fig. 1: Structure elucidation from reactants and GC-MS spectrum

covery settings, where structural novelty is expected rather than exceptional. [4]

To extend identification beyond experimentally measured reference spectra, several approaches rely on *in-silico* spectral libraries, in which EI mass spectra are predicted for large collections of candidate molecules. Models such as RASSP enable rapid prediction of EI fragmentation patterns and thereby allow matching against extensive, computationally generated libraries. By greatly expanding spectral coverage, these approaches mitigate some limitations of classical library search. However, they remain inherently constrained by the predefined candidate space and depend on accurate forward modeling of fragmentation. In reaction screening settings, where the identity and structural class of products may not be known *a priori*, defining a suitable candidate set remains a significant challenge. [5]

Moving beyond retrieval, recent work has explored generative approaches that aim to infer molecular structures directly from mass spectral data. SpecTUS represents a notable example of this direction, framing structure elucidation as a sequence generation problem in which molecular structures are generated from spectral inputs. These models offer a path toward truly library-independent structure elucidation and are particularly attractive for novel compounds. However, spectrum-only generation remains intrinsically ambiguous for GC–MS EI data of small molecules, where multiple distinct structures can give rise to similar fragmentation patterns. [6]

ReactionEye builds on these developments by targeting structure elucidation in reaction screening contexts, where additional chemical information is already available but often underutilized. Rather than relying solely on spectral evidence—whether through library matching or predicted spectra—ReactionEye adopts a multimodal formulation that combines GC–MS data with chemical context derived from the reaction, such as reactant SMILES representations. This design mirrors expert practice in manual GC–MS interpretation while enabling scalable, automated, and library-independent structure elucidation.

**Acknowledgments**

The Innovationsförderagentur NRW is gratefully acknowledged for generous financial support of this work.

**References**

[1] Joseph Mason, Harry Wilders, David J. Fallon, Ross P. Thomas, Jacob T. Bush, Nicholas C. O. Tomkinson, and Francesco Rianjongdee. Automated lc-ms analysis and data extraction for high-throughput chemistry. *Digital Discovery*, 2:1894–1899, 2023.

[2] Emma L. Schymanski, Christine M. J. Gallampois, Martin Krauss, Markus Meringer, Steffen Neumann, Tobias Schulze, Sebastian Wolf, and Werner Brack. Consensus structure elucidation combining gc/ei-ms, structure generation, and calculated properties. *Analytical Chemistry*, 84(7):3287–3295, 2012. PMID: 22414024.

[3] Felix Katzenburg, Florian Boser, Felix R. Schäfer, Philipp M. Pflüger, and Frank Glorius. Calibration-free quantification and automated data analysis for high-throughput reaction screening. *Digital Discovery*, 4:384–392, 2025.

[4] William E. Wallace and Arun S. Moorthy. Nist mass spectrometry data center standard reference libraries and software tools: Application to seized drug analysis. *Journal of Forensic Sciences*, 68(5):1484–1493, 2023.

[5] Richard Licheng Zhu and Eric Jonas. Rapid approximate subset-based spectra prediction for electron ionization–mass spectrometry. *Analytical Chemistry*, 95(5):2653–2663, 2023. PMID: 36695638.

[6] Adam Hájek, Helge Hecht, Elliott J. Price, and Aleš Křenek. Spectus: Spectral translator for unknown structures annotation from ei-ms spectra. *arXiv:2502.05114*, 2025.