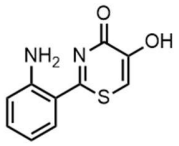
What I Learned from Analyzing Accurate Mass Data of 3000 SI Files

Mathias Christmann*

Institute of Chemistry and Biochemistry, Freie Universität Berlin, Takustr. 3, 14195 Berlin, Germany

Supporting Information Placeholder



Calculated for $[C_{10}H_8N_2O_2SNa]^+$ 243.2460 (243.0199),
found 243.0203 Mass Error: 928.7 ppm (1.6 ppm)
Alert! The molecular weight + 23.0000 has been used instead of the exact mass

ABSTRACT: A Python script for the systematic, high-throughput analysis of accurate mass data was developed and tested on over 3,000 Supporting Information (SI) PDFs from *Organic Letters*. For each SI file, quadruplets of molecular formula, measured ion, e.g. $[M+Na]^+$, reported calculated and found masses were extracted and analyzed. Interestingly, only one third of the files containing readable accurate mass data were found to be both internally consistent and in compliance with *The ACS Guide to Scholarly Communication*. The analysis revealed unexpected errors and provides actionable advice on how to improve data quality.

The rapid growth of scientific literature is driving the need for automated tools to efficiently extract, process, and analyze critical data. In chemistry, datasets documenting the synthesis of new chemical compounds typically consist of detailed preparation procedures, accompanied by characterization data to confirm the purity and structural integrity. Experimental sections have traditionally been written by humans, for humans, to facilitate replication and validation, as well as to allow verification of the work through visual inspection. In the age of digitization and automation, ongoing efforts are aimed at making natural language synthesis instructions machine-readable and -actionable,[1] leveraging robotic technologies[2,3] and enabling self-optimization.[4] In today's data-driven chemistry landscape, innovations that generate or curate high-quality, structured datasets are as essential as traditional experimental advancements.[5] Evaluating experimental data from research articles and supplemental information has become an increasingly time-consuming task for authors, reviewers, and editors alike. In 2004, Goodman et al. developed an applet to semi-automatically check various characterization data copied and pasted from manuscripts.[6] Detailed tests were conducted on ten papers and a further survey was conducted on a one hundred randomly selected data paragraphs of fifty papers. It was concluded that "preliminary tests with this program demonstrate that refereed and published experimental data is highly accurate, but errors are still occasionally perpetuated". While tools such as the experimental data checker can help improve data quality, their focus on individual errors limits the ability to gain broader insights into error patterns. The following research takes a closer look at the nature of errors in experimental chemistry papers using a single metric: accurate mass measurements using high-resolution mass spectrometry (HRMS). Accurate mass measurements support the proposed molecular formula and can be used to distinguish elemental compositions with similar nominal masses. Gratifyingly, the recorded data can also be easily checked for internal consistency. It was anticipated that a high throughput review of the accurate mass measurement data of over 3000 SI files might reveal patterns not visible through random sampling. To achieve this goal, a Python script was developed to perform a large-scale analysis of the data by systematically addressing the following tasks:

1) locate all PDFs within a given folder

2) locate and extract all accurate mass data from each PDF

3) for each measurement, recalculate the exact mass of the measured ion

4) calculate deviations of measured, calculated and recalculated masses (in ppm)

5) print a one-line analysis of each measurement highlighting unusual deviations

6) in cases of internal inconsistencies, provide a plausible explanation or, if possible, a solution to the problem

7) create a summary for all files investigated

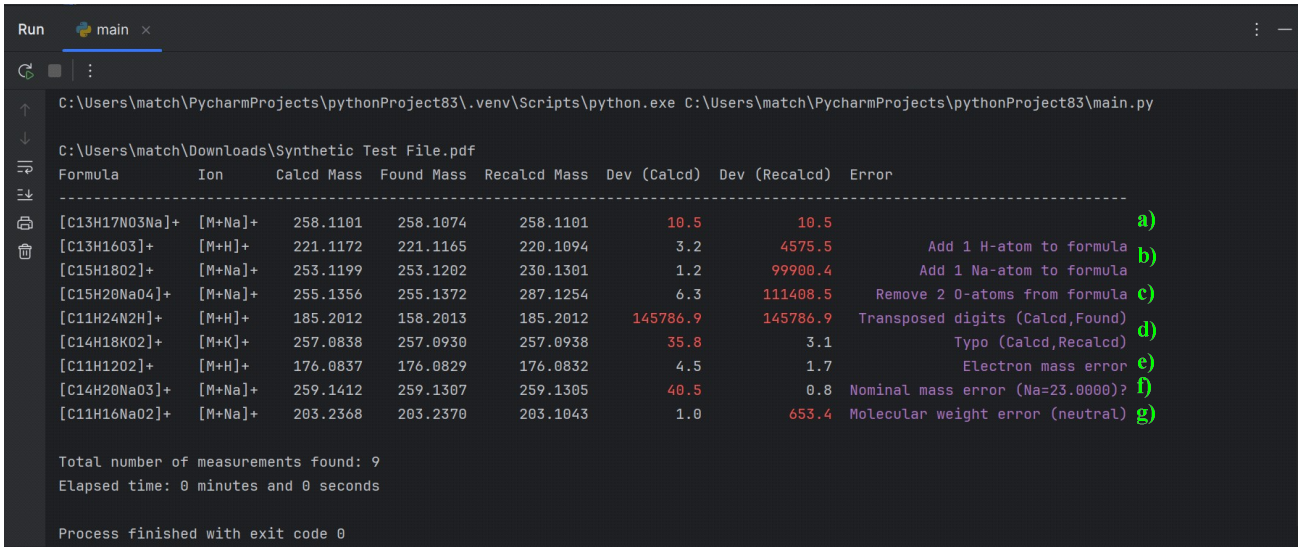


Figure 1. Automatic evaluation of a synthetic test PDF using the Python script in the PyCharm IDE: (a) ppm deviation is above the threshold; (b) added atoms ([M+H], [M+Na]) are not included in the formula; (c) incorrect formula; (d) typographical errors; (e) mass calculated for the neutral molecule; (f) nominal mass added; (g) molecular weight used instead of exact mass.

Before discussing the results of the automated screening, it is important to note that the Python script can only check for internal consistency, i.e. it is beyond the scope of this analysis to verify whether a molecular formula corresponds to the chemical structure. In addition, the conventions for reporting HRMS measurements in experimental sections need to be addressed. These vary from journal to journal in terms of acceptable deviations and the presentation of results. The *Journal of Organic Chemistry's* author guidelines state that for HRMS measurements,

"The reported molecular formulas and Calcd values should include any added atoms (usually H or Na). The ionization method and mass analyzer type (for example, Q-TOF, magnetic sector, or ion trap) should be reported. *The ACS Guide to Scholarly Communication* format for reporting accurate mass data is: HRMS (ESI/Q-TOF) m/z: $[M + Na]^+$ Calcd for $C_{13}H_{17}NO_3Na$ 258.1101; Found 258.1074."

To demonstrate how the Python script works, a synthetic SI PDF was generated and placed in the *Downloads* folder. Including the above paragraph within that file results in the output shown in Figure 1 (section a). The exact mass is followed by the recalculated value shown in parentheses, while the mass errors refer to the reported calculated and recalculated masses (in parentheses) in relation to the found mass. In the example above, the respective mass errors highlighted in red are inconsistent with an allowed arbitrary threshold of 10 ppm. Tolerated deviations can vary between journals.

A second type of internal inconsistency occurs when the formula of the measured ion does not match the calculated mass. Often, a discrepancy arises because added atoms, such as [M + H] or [M + Na], have not been included in the formula. While the root of this error is primarily a matter of convention (see author guidelines above), there are strong arguments for stating the formula of the actual ion being measured in HRMS. Reporting only the molecular formula in cases of [M + H] or [M + Na] measurements complicates verification and invites errors, as it necessitates additional steps like adding atoms (or worse: adding the mass of atoms) to obtain the displayed calculated mass. The Python script identifies missing atoms and suggests a formula that fits the calculated mass (Figure 1, section b).

Additionally, errors in the molecular formula may arise from workflows involving redundant human interventions. The accurate mass measurement itself includes an internal control mechanism: an incorrect molecular formula will not lead to a matching mass measurement. Errors occur if an incorrect formula is paired with the calculated and measured mass *after* the measurement. This can happen due to incorrect formula transfer (e.g., mistyping) or by pairing the numeric data with a newly generated formula. The script catches these mistakes and suggests a formula that does fit the calculated and measured masses by modulating the atomic composition of the given incorrect molecular formula (Figure 1, section c).[7]

Similar human-in-the-loop errors can occur when manually retyping numerical values from a printed report instead of directly transferring them into the SI. Common mistakes include transposition errors, where two adjacent digits are swapped, and substitution errors, such as typing "8" instead of "9" due to the proximity of keys on the keyboard (see Figure 1, section d).

After addressing formula errors and typos, we now turn to inconsistencies resulting from miscalculations. A common, albeit subtle, numerical discrepancy (low ppm range) is observed when the exact mass is calculated for a neutral molecule, while the measured mass corresponds to a charged species – typically a cation (Figure 1, section e). Significantly larger mass

errors in adducts ([M+H], [M+Na]) arise if the nominal mass[8] of the added atom (1.0000 for H or 23.0000 for Na) is used instead of its precise isotopic mass (1.0078 for H or 22.9898 for Na) (Figure 1, section f). Unlike elements such as sodium (Na) and fluorine (F), which are monoisotopic, most elements, like carbon (C) and hydrogen (H), are polyisotopic. In compounds containing these polyisotopic elements, confusion between molecular weight (MW) and monoisotopic mass usually results in significant errors (Figure 1, section g). This example also illustrates a situation where an apparent miscalculation is misleadingly validated by matching measurements. In rare instances, certain isotopic compositions result in the molecular weight and exact mass being very similar or identical. For example, the exact mass of $C_{11}H_{22}BN_2^+$ is 265.1871, while the molecular weight of $C_{11}H_{22}BN_2$ is 265.1870.

Efficient extraction and processing of large datasets can enable meta-analyses that reveal hidden patterns and trends. Recognizing that *Organic Letters* is committed to delivering high-quality Supporting Information,[9] we initiated an analysis of over 3,000 SI PDFs from the journal. To foster a discussion on how to further improve data quality, Table 1 summarizes a screening of all SI PDFs from 2023 and 2024 (issues 1–36), comprising 3,028 files and totaling 26.3 GB of data.

Table 1. Screening of Supporting Information of *Organic Letters* (2023-present)

	Org. Lett. **2023,** 1-51	*Org. Lett.* **2024**, 1-36
Number of SI PDF files	1,677 (14.5 GB)	1,351 (11.8 GB)
Files with HRMS data	1,618 (96%)	1,294 (96%)
HRMS measurements	56,134	45,749
Files without errors	422 (26%)	340 (26%)
Molecular weight errors	17	25
Nominal mass errors	91	147
Electron mass errors	10,000+	8074
Transposed digits	9	6
Typographical errors	53	47
1 H-atom added	1,617	1362
1 Na-atom added	679	393
1 O-Atom added	21	23
1 C-Atom added	10	8
2 H-atoms added	7	8
2 O-atoms added	7	6
1 H-atom removed	154	241
1 Na-atom removed	9	8
1 CH_2-group removed	3	4

All calculations were performed on a personal computer, with no need for any data to leave the device. On a laptop computer (EliteBook 840 G8, Intel i5 @ 2.4 GHz), scanning a single SI PDF takes less than a second and checking a whole volume of *Organic Letters* SI PDFs takes about 15 min. The script demonstrated a high accuracy rate (>99%), successfully identifying HRMS data in over 95% of the analyzed files. The remaining <5% largely comprised files that lacked HRMS data altogether (e.g., those related to computational studies).

Among the files with HRMS data, only about one-third fully adhered to journal guidelines. The most common minor deviation observed was the calculation of the exact mass for the neutral molecule, rather than for the charged species. The second most frequent error involved the omission of added atoms (e.g., [M+H], [M+Na]) in the molecular formula. Cases where formulas included incorrect or missing atoms (such as O, F, Cl) or groups (e.g., CH_2) were swiftly detected and corrected. Similarly, simple typographical errors were easily identified and addressed.

Beyond these minor oversights, the script unexpectedly uncovered 280 significant errors (molecular weight errors and nominal weight errors), which were often rooted in fundamental misunderstandings of how to correctly calculate exact masses. In some cases, unusual discrepancies between measured and miscalculated masses were simply overlooked, while in others, the measurements appeared to validate the incorrect calculations. It is hoped that this script will save authors and reviewers considerable time and effort in identifying and correcting such errors before publication.

What can be learned from this? To enhance data quality, it is key to implement automated protocols that take the human out of the loop in data handling post-measurement, thus reducing the risk of manually introducing errors. Additionally, adhering to a journal's conventions for presenting data is essential to reduce ambiguity and facilitate verification. The problem can be approached from both ends by both putting a focus on machine-readability[10] and by devising tools[11] that can help to translate chemical language and representations such as chemical drawings.

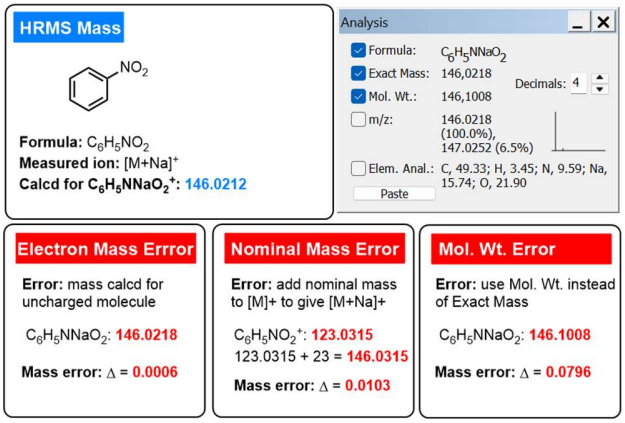


Figure 2. Examples of calculating and miscalculating the exact mass using ChemDraw®.

In the realm of chemical education, it's important to emphasize the distinctions between nominal mass, exact mass, molecular weight, and the mass differences between charged and uncharged species ([M] vs. [M]+). Figure 2 shows the magnitude of the error relative to the mass of the cation [M+Na]+ depending on how the exact mass was miscalculated.

Although this author had no prior coding experience with Python, this script was developed in a relatively short timeframe by following a 4-hour Python tutorial, leveraging large language models (LLMs) such as ChatGPT, Gemini, and Claude for code generation and utilizing existing Python libraries like Molmass.[12] By making this script freely available, the author hopes to contribute to improving the quality and reliability of scientific data and inspire other data-driven approaches.

ASSOCIATED CONTENT

Data Availability Statement

The Python script and the test file can be accessed at:
https://github.com/match22lab/HRMS-Checker-2.0

AUTHOR INFORMATION

Corresponding Author

* E-mail: mathias.christmann@fu-berlin.de

ORCID

Mathias Christmann: 0000-0001-9313-2392

Notes

The author declares no financial interest.

REFERENCES

(1) Steiner, S.; Wolf, J.; Glatzel, S.; Andreou, A.; Granda, J.; Keenan, G.; Hinkley, T.; Aragon-Camarasa, G.; Kitson, P. J.; Angelone, D.; Cronin, L. Organic Synthesis in a Modular Robotic System Driven by a Chemical Programming Language. *Science* **2019**, *363*, 144-152. DOI: 10.1126/science.aav2211

(2) Hahm, H. S.; Schlegel, M. K.; Hurevich, M.; Eller, S.; Schuhmacher, F.; Hofmann, J.; Pagel, K.; Seeberger, P. H. Automated Glycan Assembly Using the Glyconeer 2.1 Synthesizer. *Proc. Natl. Acad. Sci. U.S.A.* **2017**, *114*, E3385-E3389. DOI: 10.1073/pnas.1700141114

(3) Blair, D. J.; Chitti, S.; Trobe, M.; Pupo, G.; Doney, A. C.; Hartwig, J. F.; Burke, M. D. Automated Iterative Csp3–C Bond Formation. *Nature* **2022**, *604*, 92–97. DOI: 10.1038/s41586-022-04491-w

(4) Slattery, A.; Wen, Z.; Tenblad, P.; Sanjosé-Orduna, J.; Pintossi, D.; den Hartog, T.; Noël, T. Automated Self-Optimization, Intensification, and Scale-Up of Photocatalysis in Flow. *Science* **2024**, *385*, 647-651. DOI: 10.1126/science.adj1817

(5) Ai, Q.; Meng, F.; Shi, J.; Pelkie, B.; Coley, C. W. Extracting Structured Data from Organic Synthesis Procedures Using a Fine-Tuned Large Language Model. *Dig. Discov.* **2024,** *3* (9), 1822–1831. DOI: 10.1039/D4DD00091A

(6) Adams, S. E.; Goodman, J. M.; Kidd, R. J.; McNaught, A. D.; Murray-Rust, P.; Norton, F. R.; Townsend, J. A.; Waudby, C. A. Experimental Data Checker: Better Information for Organic Chemists. *Org, Biomol. Chem.* **2004**, *2*, 3067-3070. DOI: 10.1039/b411699m

(7) The script can handle isotopes when designated. Determining isotopic compositions, e.g. isotopic composition of a dibromide, is beyond the scope of this script.

(8) Attygalle, A.B., Pavlov, J. Nominal Mass? *J. Am. Soc. Mass Spectrom.* **2017**, *28*, 1737–1738. DOI: 10.1007/s13361-017-1647-6

(9) Hunter, A. M.; Smith, A. B. Review of Supporting Information at Organic Letters. *Org. Lett.* **2015**, *17*, 3182-3185.
DOI: 10.1021/acs.orglett.5b01700

(10) Jablonka, K. M.; Patiny, L.; Smit, B. Making the Collective Knowledge of Chemistry Open and Machine Actionable. *Nat. Chem.* **2022**, *14*, 365-376. DOI: 10.1038/s41557-022-00910-7.

(11) Rajan, K.; Brinkhaus, H. O.; Agea, M. I.; Zielesny, A.; Steinbeck, C. DECIMER.ai: An Open Platform for Automated Optical Chemical Structure Identification, Segmentation and Recognition in Scientific Publications. *Nat. Commun.* **2023**, *14*, 5045. DOI: 10.1038/s41467-023-40782-0.

(12) Gohlke, C. *cgohlke/molmass: v2024.10.25*; Zenodo, **2024**. https://doi.org/10.5281/zenodo.13992852.