JOC The Journal of Organic Chemistry

pubs.acs.org/joc Editorial

Negative Data in Data Sets for Machine Learning Training

Cite This: J. Org. Chem. 2023, 88, 5239−5241 Read Online

ACCESS | Metrics & More | Article Recommendations

Data-driven chemistry has been described as the "future" of industrial organic synthesis that "will increasingly help guide synthetic chemists through the toughest synthesis problems", in a recent editorial in *The Journal of Organic Chemistry*.[1] Data-enabled machine learning (ML) methods have been shown to be equal and sometimes superior to human, intuition-driven approaches in common tasks in organic synthesis such as reaction optimization.[2] This is also reflected by the continued growth of publications discussing ML in organic synthesis over the past eight years (Figure 1).[3]

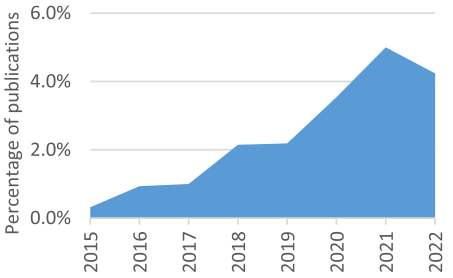

Figure 1. Trend of publications on machine learning and organic synthesis as portion of total publications, 2015−2022.

As in all ML applications, the performance of models for a given task relies heavily on the quality and scope of the training data that cover this task. Some ML models such as yield predictions need to distinguish between successful and unsuccessful reactions, so they require examples of high-, medium-, and low-yielding reactions in the training set. However, reactions with low or no yield (frequently referred to as "negative data") are rarely included in the published literature. This represents a significant limitation of the literature, which is therefore often insufficient for the purpose of training ML models for these purposes.[4] Literature-derived databases such as Reaxys or the data set extracted from the U.S. Patent and Trademark Office[5] suffer from the same selection bias in that often only successful and high-yielding reactions are reported. A recent editorial by Kozlowski touches on the reasons that negative data remain unpublished in the context of the importance of meaningful substrate scopes.[6] While the original data repositories such as experimental sections of Ph.D. theses or high-throughput experimentation (HTE) data sets do contain the full information on both high- and low-yielding reactions, they are usually not widely accessible or available in a computer-readable form. Electronic laboratory notebooks (ELNs) are an important step to address this problem and are widely used in industry. Their current inconsistent adoption in academia is likely to improve due to increasing requirements of publishers and funders for FAIR Data usage. Efforts to create open-access databases, such as the Open Reaction Database (ORD),[7] also try to address this problem by providing a data structure for reporting chemical reactions and by removing the distinction between "positive" and "negative" outcomes.

There has been much interest in minimum information standards in chemistry, including better research data management practices,[8] and *The Journal of Organic Chemistry* and *Organic Letters* have been active participants in this discussion.[9] In this Editorial, we argue that the efforts of including wider scope and yield ranges are necessary but not sufficient and need to be complemented by additional information that should be reported. There is much more information in low-yielding reactions than is commonly accepted, and simply stating that a reaction gave 0% yield is insufficient to learn from.

We consider the following scenarios:

1. **No remaining starting material and no product.** This result implies that the reaction produces a different product than originally intended. The barrier height for the intended reaction need not be prohibitively high, but side reactions have lower barriers. It would of course be desirable if the actual products were characterized and reported, but even without that additional effort, this scenario still needs to be identified.
2. **Most or all of the starting material remains.** In the case of low conversion, the reaction barrier is prohibitively high, or the reaction is thermodynamically unfavorable. In the case of catalytic reactions, this may also indicate that the catalyst has been deactivated. The latter effect is essentially impossible to distinguish in a single reaction, but it is straightforward to detect utilizing a competition reaction.
3. **The reaction was not performed as intended.** This could be the result of a variety of factors, including a reaction that was performed using unintended conditions (e.g., contaminations in substrates, reagents, or solvents), an error in physical manipulation (e.g., the reaction flask was dropped or product was lost due to a spill), the reaction was performed as a proof of concept

Published: April 26, 2023

and the product never quantified, or changing priorities caused a planned reaction to be abandoned after entry into the ELN but before execution.

When building an ML model for reaction outcome, the implications of the three scenarios would be quite different. In the absence of additional information, a model would not be able to distinguish them. Furthermore, it should be recognized that there are several different ways to report yields (crude, isolated, conversion, etc.). With these possibilities in mind, we propose the following standards for reporting data (negative or not) in ELNs, electronic databases, or traditional laboratory notebooks (LNs):

1. **Isolated and crude reaction yields** should be clearly denoted and reported separately, because problems in the workup or isolation procedure can lead to low reported yields for a high-yielding reaction. For example, crude yields are often preferable for the training of ML models. However, yields determined by chromatography should also provide information on reaction conversion, which can easily be determined by quantifying the amount of starting material by the same method.
2. **Other measures such as conversion are frequently used as proxies of yield,** especially in HTE. It is preferable that both conversion and yield are reported, but if only conversion is available, this should be clearly denoted and not reported as a yield. In this context, analytic problems such as overlapping peaks that can bias the yield readout should be flagged.
3. **A mandatory conclusion before closing the experiment,** which could take the form of a drop-down menu in an ELN, with the following options:
   A. Significant amount of product was detected (success)
   B. No significant product was detected, but starting material remains
   C. Neither starting material nor intended product was detected
   D. The reaction was not run as intended (incorrect setup, physical error, reaction canceled, other). In this case, a free-text comment describing the observation would be beneficial, but not essential.

It is worth noting that such additional information would be among the first questions raised in discussing low-yielding reactions. The same should be expected when publishing a result as part of a substrate scope or control experiment; however, this is currently not included in commonly used data sources. The proposed standard would help to categorize (E)LN entries as well as further expand the information content of reported negative results by giving the necessary context of what did occur in the reaction. This information is of no extra cost or effort to the chemist, and these data would be invaluable to ML models that are being trained on (E)LN data.

The inclusion of these data would be beneficial in training ML models to predict reaction yields and conditions, which present a long-standing challenge in the application of predictive methods. While reactions that were not run as intended can simply be discarded, the other classifications can be informative for model building and could be used as additional features. We therefore encourage all experimental chemists and authors to implement procedures that provide the crucial information on "what happened in a low-yielding reaction" in a way that is both easy and standardized according to FAIR Data principles increasingly required by reviewers, editors, and funding agencies.

Michael P. Maloney orcid.org/0009-0001-3385-7567
Connor W. Coley orcid.org/0000-0002-8271-8723
Samuel Genheden orcid.org/0000-0002-7624-7363
Nessa Carson orcid.org/0000-0002-2769-1775
Paul Helquist orcid.org/0000-0003-4380-9566
Per-Ola Norrby orcid.org/0000-0002-2419-0705
Olaf Wiest orcid.org/0000-0001-9316-7720

AUTHOR INFORMATION

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.joc.3c00844

Notes

Views expressed in this editorial are those of the authors and not necessarily the views of the ACS.
This editorial is jointly published by *Organic Letters* and *The Journal of Organic Chemistry*.

Biographies

Connor W. Coley builds machine learning tools for chemistry and is one of the cofounders of the Open Reaction Database.

Samuel Genheden, **Nessa Carson**, and **Per-Ola Norrby** are chemists at AstraZeneca working at the interface of data science, high-throughput experimentation, and computational chemistry in an industrial setting.

Michael P. Maloney, **Paul Helquist**, and **Olaf Wiest** apply computational methods to mechanistic and synthetic organic chemistry. O.W. is director of the NSF Center for Computer Assisted Chemistry and a former associate editor of *J. Org. Chem.*

ACKNOWLEDGMENTS

C.W.C. and O.W. gratefully acknowledge financial support of our work in this area through the NSF Center for Computer Assisted Synthesis (C-CAS) through CHE−2202693.

REFERENCES

(1) Ragan, J. A.; Dreher, S. D. Excellence in Industrial Organic Synthesis 2019: The Past, Present, and Future. *J. Org. Chem.* **2019**, *84*, 4577−4579.
(2) Shields, B. J.; Stevens, J.; Li, J.; Parasram, M.; Damani, F.; Alvarado, J. I. M.; Janey, J. M.; Adams, R. P.; Doyle, A. G. Bayesian reaction optimization as a tool for chemical synthesis. *Nature* **2021**, *590*, 89−96.
(3) PubMed search, "machine learning" and "organic" and "synthesis", 2015−2022 as part of total number of publications in PubMed, performed on 3/11/2023. Note that PubMed has a time lag for inclusion, so 2022 publications will not be complete. Search performed using: Sperr E. PubMed by Year. 2016. Available from http://esperr.github.io/pubmed-by-year/
(4) Strieth-Kalthoff, F.; Sandfort, F.; Kühnemund, M.; Schäfer, F. R.; Kuchen, H.; Glorius, F. Machine learning for chemical reactivity: The importance of failed experiments. *Angew. Chem., Int. Ed.* **2022**, *61*, No. e202204647.
(5) Lowe, D. Chemical reactions from US patents (1976−Sep2016). *Figshare*. 2017. Data set. DOI: 10.6084/m9.figshare.5104873.v1.
(6) Kozlowski, M. C. On the Topic of Substrate Scope. *Org. Lett.* **2022**, *24*, 7247−7249.
(7) Kearnes, S. M.; Maser, M. R.; Wleklinski, M.; Kast, A.; Doyle, A. G.; Dreher, S. D.; Hawkins, J. M.; Jensen, K. F.; Coley, C. W. The open reaction database. *J. Am. Chem. Soc.* **2021**, *143*, 18820−18826.

(8) Herres-Pawlis, S.; Bach, F.; Bruno, I. J.; Chalk, S. J.; Jung, N.; Liermann, J. C.; McEwen, L. R.; Neumann, S.; Steinbeck, C.; Razum, M.; Koepler, O. Minimum Information Standards in Chemistry: A Call for Better Research Data Management Practices. *Angew. Chem., Int. Ed.* **2022**, *61*, No. e202203038.
(9) Hunter, A. M.; Carreira, E. M.; Miller, S. J. Encouraging Submission of FAIR Data at *The Journal of Organic Chemistry* and *Organic Letters*. *J. Org. Chem.* **2020**, *85*, 1773−1774.

https://doi.org/10.1021/acs.joc.3c00844
J. Org. Chem. 2023, 88, 5239−5241