nature portfolio

Peer Review File

Discovering Organic Reactions with a Machine-Learning-Powered Deciphering of Tera-Scale Mass Spectrometry Data

Corresponding Author: Professor Valentine Ananikov

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
I agree that it is useful (significant) to use existing HRMS data for prediction of reaction in a range of fields. The idea of this study is unique and very interesting.

But, I am not still sure how it works. First, how are the products estimated. It is understandable that the reactant information is retrieved from MS spectra. But, other information was not quite clear to me. It seems that the products information is also retrieved from the MS spectra (I guess the products should exist in the same MS spectra). The authors mentioned that the products information is limited to molecular formula. So, structural information is not available for the products and possibly other intermediates if any (in the same spectra). There must be some limitations, including reaction prediction as noted below. Furthermore, some extracted reactions were validated by DFT calculation. How can we calculate without structure data? (some more detailed descriptions with be necessary for the transition state calculation which is sometime challenging)

Regarding reaction prediction, some parts of structures including functional groups are used, as shown in Fig 3b, to construct the product structures. How can we obtain these functional structures? It must be very much depending on reactions? We may need to know some reactions, in advance, that happened in the system. I am very much confused because the title of this study is "reaction discovery". While the approach is unique and useful, but currently it is not clear to me what kind of information is necessary for this analysis.

Reviewer #2

(Remarks to the Author)
With data sharing policies across funding agencies now requiring making data publicly available, tools that can utilize these expected resources for reuse and discovery is prudent. Reaction Discovery with a Machine-Learning-Powered Deciphering of Tera-Scale Mass Spectrometry Data provides a data-search tool for re-mining HRMS spectral databases for characterizing and discovery of organic chemistry/synthesis reaction products. This tool demonstrated that known, unknown, and novel products could be identified, which were validated experimentally. My major and minor comments are below

Major comments
1. The manuscript should be limited to describe reaction discovery for organic chemistry/synthesis fields, although this isn't detailed in the title or abstract. I understood that there are no complex samples (blood, soil, tissue) in the model training or testing. Similarly, statements in the introduction are not accurate if the authors don't qualify that they are referring only to the field of organic chemistry/synthesis. Therefore the title, abstract, and manuscript need to reflect this. I suggest a title change to "Organic Chemistry Reaction Discovery…." or something similar, and adding this verbiage throughout. Examples of how the text is inaccurate if speaking about other fields is below:

A. Data re-use/re-mining/re-purposing is very common in other fields like metabolomics and proteomics. Therefore, "experimentation in the past" is not novel for these fields

B. Paragraph 2 page 4, "to date, search algorithms for complex (with more than one compound in the spectrum) mass spectrometry have been actively used mainly in proteomics studies". This is not accurate - it has been used routinely in metabolomics for over 10 years

C. Paragraph 2 page 4, “common disregard of isotopic distribution patterns” is also not accurate. From free packages like CAMERA and SIRIUS developed over 10 years ago, to all vendor software on the current market (e.g. Profinder by Agilent and Compound Discoverer by Thermo), all use isotope distribution for chemical annotation in metabolomics. It is a basic criteria for annotation and identification of compounds including tandem mass spectra, retention time, and m/z mass accuracy reported in all manuscripts.

2. The database for mining needs to be more described with much greater detail and the limitations of that set considered in determining whether this tool is fully developed.
All mass spectral data came from a single Institute. Were these data all collected on the same HRMS instrument using the same analytical method? Characteristics like Instrument, resolution, scan speed, instrument sensitivity, ion collection range need to be described as they can affect spectral quality. Was chromatography used? These details need to be summarized in the SI. If there is little variability in these parameters in the database, the findings are not generalizable, and the use of this tool for others is extremely limited—to perhaps single laboratories re-mining their own data or a core facility. The addition of spectral data from external sources—publicly available databases– are needed to provide evidence that this tool has broader application.

3. Figure 3A, t-SNE map needs to be compared to a reference database to describe the breadth of chemical space compared to the authors’ database. The author’s space should be overlayed with the chemical space described by a chemical library such as PubChem to see if it is representative.
minor: the colored dots in figure 3a legend are too small
Minor: a distribution of the spectra by user (A-Other) should be provided. It looks like user C provided > 60% of the data- all coming from a single instrument/method?

4. P7, paragraph 2: The authors admit that HRMS without fragmentation can provide information only about molecular formulas; thus structural isomers cannot be distinguished. The authors need to describe how many of the results in figure 3d are isomers as a limitation of this software.

5. P7 ,paragraph 3: “most of the search engine answers could not be validated due the lack of FAIR description data needed for recognition of the reaction mixture initial composition.” This needs to be quantified as it is a major limitation of the workflow. What FAIR data is minimally required for interpretability of the results of this tool? If these data are unavailable, does the tool still provide value?

6. How was the Mizoroki-Heck reaction chosen, and can the authors describe if there is any bias in this selection of this reaction based on the chemical space coverage of their database? The authors should also demonstrate rate of hits for several other reactions as well, even without going into a deep validation.

Minor comments

1.Abstract: “while obviating the need for conducting additional experiments.” I suggest “reducing” instead of obviating.

2. P 2. Paragraph 2. Suggested text edits for clarity, “in a typical organic synthesis workflow, chemists select particular experimental conditions for the optimization of a reaction to achieve maximal outcome of the desired product. Next, carrying out the reaction and sample preparation is followed by detection and characterization of the chemical compositions of the studied system using an appropriate analytical system (Figure 1b).”

3. I think Figure 1a is unnecessary, so the overall figure 1 is more simplified

4. Figure 2. The steps should be labeled 1-4 or A-D according to the text and in sequential order. The figure looks like steps 1 and 2 are occurring in parallel, which is not what is happening according to the text description. The authors can then describe which steps are course or fine searches. The shading of the squares for cosine distance from F1 and F2 should be differentiated by hash filled and non-hash filled so it can be seen easier. Figure 2 caption should be more descriptive.

5. P7 paragraph 3: “a search pipeline (Figure 2a) was run for each generated ion through the entire tera-scale HRMS database…”. Are the authors referring to the 520 ion formulas? This is unclear.
The computational run time for the query of the 520 ion formulas should be stated.

6. Page 8, Experimental validation: these sentences are unclear
“Formation of these catalyst transformation….” -which are these referring to?
“These reactions were previously described….” which ones and by who?
“....with the latter being described recently” which is latter?
“As previously found…” by whom?
“Apparently the process occurs via some kind of transfer hydrogenation reaction” why?

7. I had a hard time navigating the SI materials. The order in the main text start with section 3, then S4.1, then S4.2, but then the methods come in at S5 and refer forward and back. The SI data and figures needs to be ordered in a consequative manner to make it easier for the reader to naviagate through without going forward and backgward and all over.

8.Figure 4 and page 7, the collected results:. The text with the results numbered 1-6 need to be connected to the figure in 4. Using m/z may be the best way to make it easier to understand which result matches with which chemical by putting the m/z

in the text after the product name in paranthases as well as in the figure next to the structure. Also, where are all of the struures for each chemical? For A there are 4 compounds but only 1 spectra?

9.Page 9, the text should refer to the main text figure 5, but also all of the SI figures of the MSMS data that was collected. Figure 5d, should be re-named 5c, as this is the order of mention in the text.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
I do not have any further comments. Thanks for the authors revision and reply.

Reviewer #2

(Remarks to the Author)
The authors provide a novel machine learning tool for re-mining available HRMS data for hypothesis testing and new discovery of organic and catalytic reaction products, which can have a significant impact on this field and further promote data sharing.

The authors have responded comprehensively to all my questions. In particular, they have demonstrated applicability to an additional query, provided details to demonstrate the breadth of the test data, and improved formatting for readability. I have no further questions or comments.

Reaction Discovery with a Machine-Learning-Powered Deciphering of Tera-Scale Mass Spectrometry Data

Authors: Konstantin S. Kozlov†, Daniil A. Boiko †, Julia V. Burykina, Valentina V. Ilyushenkova, Alexander Yu. Kostyukovich, Ekaterina D. Patil, and Valentine P. Ananikov

Point-by-point responses to the comments and descriptions of the changes made in the revised manuscript

The authors are grateful to the Editor and Reviewers for their detailed manuscript reading and numerous important comments with deep insight into the subject of the manuscript.

Referee: 1

Comments to the Author

I agree that it is useful (significant) to use existing HRMS data for prediction of reaction in a range of fields. The idea of this study is unique and very interesting.

We are grateful to the Reviewer for the evaluation of our manuscript.

1. But, I am not still sure how it works. First, how are the products estimated. It is understandable that the reactant information is retrieved from MS spectra. But, other information was not quite clear to me. It seems that the products information is also retrieved from the MS spectra (I guess the products should exist in the same MS spectra). The authors mentioned that the products information is limited to molecular formula. So, structural information is not available for the products and possibly other intermediates if any (in the same spectra). There must be some limitations, including reaction prediction as noted below.

We thank the Reviewer for this important point. To clarify, the process of estimating products involves initially generating potential product structures, which are subsequently searched for within the mass spectra. This approach ensures that we can identify the products.

Initial assumptions for the product information may be indeed derived from MS spectra, and our methodology currently incorporates possibilities of determining and searching for molecular formulas on the basis of two types of hypotheses – human generated and automatically searched (individually or combining both).

In addition to human generation, in the revised article, we implemented the ways to generate these hypotheses automatically:

1. **Language model-based generation:** We identify bonds that are likely to break and can generate a large number of potential products, covering a complete set of variations within the considered chemical process. This helps in generating a diverse set of possible product hypotheses, thereby increasing the likelihood of identifying the correct hypotheses in the spectra.
2. **Decomposition of reactants:** We also apply the BRICS fragmentation method to break down reactants into potential constituents, which can then combinatorically predict products (complete chemical space evaluation).

It is important to mention that both approaches (#1 and #2 above) generate molecular formulas and chemical structures connected to these molecular formulas.

When the molecular formulas are detected in the mass spectra with the described approach for a particular chemical space, the plausible chemical structures can be estimated as well. Further clarifications of the chemical structures can be made by conducting tandem mass spectrometry experiments to obtain structural information or use readily available orthogonal methods (like, NMR, etc.).

We added a discussion on automated hypothesis generation in the Supporting Information (Section S5), with reference in the main text (page 5, last paragraph).

In the revised article, these changes were reflected in Figures 1 (section b in the revised manuscript) and 2 (block G) in the manuscript text (page 5, third paragraph in the revised manuscript). Information about structure confirmation was added to the first paragraph of page 3 in the revised manuscript.

2. Furthermore, some extracted reactions were validated by DFT calculation. How can we calculate without structure data? (some more detailed descriptions with be necessary for the transition state calculation which is sometime challenging)

Thank you for the comment; we appreciate critical evaluation, which helped us improve the manuscript. The answer to the above question clarifies the way structural information can be obtained.

The corresponding DFT calculations were performed on the basis of hypotheses about the reaction products derived from our method. Our approach encourages users to form hypotheses, which are then validated via previously collected experimental data or by DFT calculations.

In the current revision, we provided ways to perform automated hypothesis generation (see the answer above). As the next step, for testing the hypothesis and elucidating further structural information, we show that the users may perform DFT calculations to verify promising experimental findings from the MS data and use additional experiments to collect structural information if it was not collected before.

3. Regarding reaction prediction, some parts of structures including functional groups are used, as shown in Fig 3b, to construct the product structures. How can we obtain these functional structures? It must be very much depending on reactions? We may need to know some reactions, in advance, that happened in the system. I am very much confused because the title of this study is "reaction discovery". While the approach is unique and useful, but currently it is not clear to me what kind of information is necessary for this analysis.

This is indeed a correct comment in stating that deriving the functional structures involves some understanding of the reactions taking place. In the context of product structure construction, functional groups and substructures are derived on the basis of user hypotheses and are not directly retrieved from the data.

To clarify, the term "reaction discovery" in the title refers to the process of discovering and validating new chemical transformations leveraging mass spectrometric data. The reaction information is derived from the reagents and possible guessed products – the information typically presented in the description of the MS sample and stored together with the spectrum. I.e., when the users gave samples for MS analysis, the chemistry expected is provided as sample

information (as an example of such a sample description, see, Figure S1 in the Supporting Information).

In this revision, we have aimed to lessen the dependency on prior knowledge of reactions and demonstrate that manually derived hypotheses, while beneficial, are not strictly necessary. For example, the introduction of the BRICS fragmentation method facilitates systematic decomposition of reactants into possible products, thus providing exhaustive information to construct product hypotheses. Important to mention, that automatic hypothesis variation scans a large chemical space and can find new products that would never be expected by a human researcher.

We thank you again for this detailed question and hope this explanation clarifies our approach, especially the integral role of user hypotheses, automatic hypotheses and mass spectrometric evaluation, and data-driven insights in the analysis.

We thank the Reviewer for the evaluation of our article and for helpful comments! The revised manuscript was improved!

Referee: 2

Comments to the Author

With data sharing policies across funding agencies now requiring making data publicly available, tools that can utilize these expected resources for reuse and discovery is prudent. Reaction Discovery with a Machine-Learning-Powered Deciphering of Tera-Scale Mass Spectrometry Data provides a data-search tool for re-mining HRMS spectral databases for characterizing and discovery of organic chemistry/synthesis reaction products. This tool demonstrated that known, unknown, and novel products could be identified, which were validated experimentally. My major and minor comments are below.

We are grateful to the Reviewer for the evaluation of our manuscript.

1. The manuscript should be limited to describe reaction discovery for organic chemistry/synthesis fields, although this isn't detailed in the title or abstract. I understood that there are no complex samples (blood, soil, tissue) in the model training or testing. Similarly,

statements in the introduction are not accurate if the authors don't qualify that they are referring only to the field of organic chemistry/synthesis. Therefore the title, abstract, and manuscript need to reflect this. I suggest a title change to "Organic Chemistry Reaction Discovery…." or something similar, and adding this verbiage throughout. Examples of how the text is inaccurate if speaking about other fields is below:

Thank you for this comment. We agree with the importance of precision when describing the research scope and replaced the title with "Discovering Organic Reactions with Machine-Learning-Powered Deciphering of Tera-Scale Mass Spectrometry Data".

In fact, the subject coverage includes organic and catalytic reactions (which are sometimes named separately in the literature), and the title can include "Discovering Organic and Catalytic Reactions ..". Currently, we keep a shorter version and will be grateful for a comment if a longer version is preferred.

A. Data re-use/re-mining/re-purposing is very common in other fields like metabolomics and proteomics. Therefore, "experimentation in the past" is not novel for these fields

Thank you for this comment. We added a clarifying statement, focusing the scope of this work to organic chemistry (see page 4, last paragraph in the revised manuscript). We also reference the corresponding discussion on related topic ref. 63.

B. Paragraph 2 page 4, "to date, search algorithms for complex (with more than one compound in the spectrum) mass spectrometry have been actively used mainly in proteomics studies". This is not accurate - it has been used routinely in metabolomics for over 10 years

We apologize for the oversight. We added this information to the metabolomics section (page 4, first paragraph, refs. 40–42).

C. Paragraph 2 page 4, "common disregard of isotopic distribution patterns" is also not accurate. From free packages like CAMERA and SIRIUS developed over 10 years ago, to all vendor software on the current market (e.g. Profinder by Agilent and Compound Discoverer by Thermo), all use isotope distribution for chemical annotation in metabolomics. It is a basic

criteria for annotation and identification of compounds including tandem mass spectra, retention time, and m/z mass accuracy reported in all manuscripts.

Thank you for this correction! We rewrote this sentence (also page 4, first paragraph in the revised manuscript, refs. 57–58) to highlight the importance of isotopic distribution patterns. We would also like to point out that the developed tool aims to merge hypothesis generation methods (as added in this revision, in addition to the manual approach) and search large databases of experimental mass spectra as quickly as possible.

2. The database for mining needs to be more described with much greater detail and the limitations of that set considered in determining whether this tool is fully developed.
All mass spectral data came from a single Institute. Were these data all collected on the same HRMS instrument using the same analytical method? Characteristics like Instrument, resolution, scan speed, instrument sensitivity, ion collection range need to be described as they can affect spectral quality. Was chromatography used? These details need to be summarized in the SI. If there is little variability in these parameters in the database, the findings are not generalizable, and the use of this tool for others is extremely limited—to perhaps single laboratories re-mining their own data or a core facility. The addition of spectral data from external sources—publicly available databases– are needed to provide evidence that this tool has broader application.

Thank you for this question! The data were collected on a single instrument over more than 10 years at a very broad focus on organic chemistry research. For data collection, we used a Bruker maXis Q-TOF mass spectrometer, which is used to collect most HRMS data at the institute. Currently, the instrument has a resolution of up to 30,000.

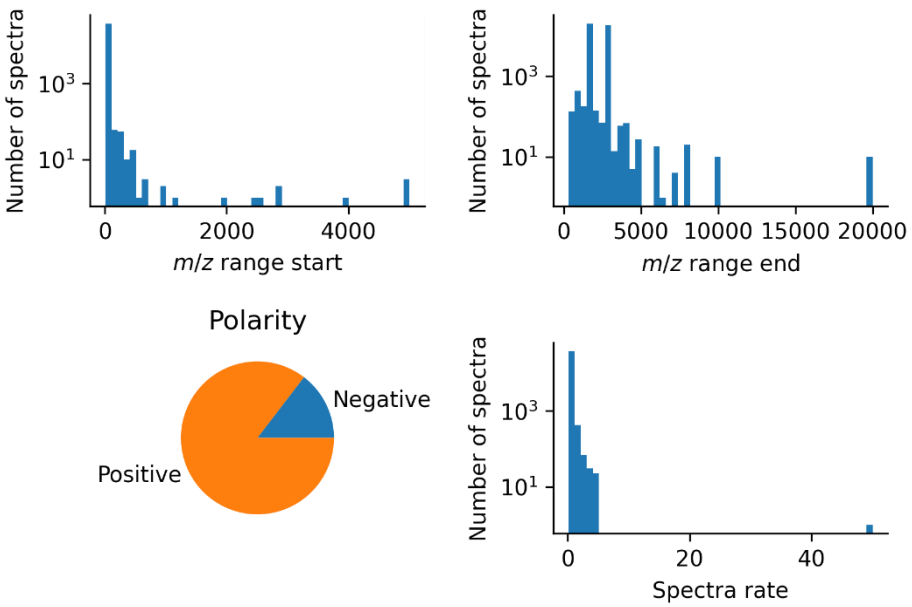


Figure R1. Statistics of the data used in the study (Figure S20 in the Supporting Information).

Statistics on other quantities, such as polarity, mass range, and the spectra collection rate, are provided in **Figure R1**. File size (which allows inference of the analysis method: direct infusion, LC, or reaction monitoring) statistics are provided in the Supporting Information, **Figure S19**. The majority of spectra are collected via direct infusion, with smaller numbers being obtained via LC–MS runs or reaction monitoring. Overall, there is high variability in these quantities, even though only one instrument was used. We added details about the database to the Supporting Information (Section S9).

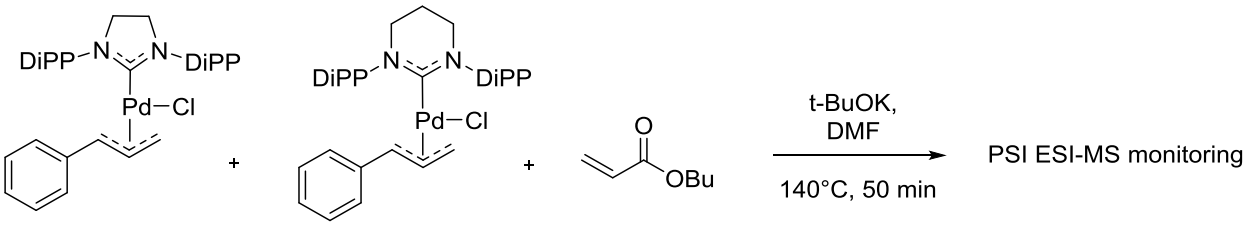
To further explore the applicability of the tool, we should consider potential failure points and determine if they are covered during our analysis. To successfully perform the search, the following conditions should be met:

1. A compound should ionize first. Most of our observations are made for the direct infusion case, which is the most challenging because of matrix effects. We can assume that in the case of LC analysis, the probability of observing the compound is greater (see extended analysis below).
2. The mass spectrometer should be sensitive enough to observe the ions. Modern research-grade instruments, on average, have higher sensitivity than our instrument, which is almost two decades old.
3. Our algorithm should be able to identify the ion. The core algorithm used for the matching process was previously described in our prior work (10.1021/jacs.2c03631)

and was shown to be applicable for ultrahigh-resolution data from an FT-ICR mass spectrometer. By demonstrating that it could be a part of a larger system for Q-TOF data, we can assume that it works at resolutions of 10,000 and higher.

To further confirm that ions can be observed across multiple modalities of reaction data collection, we provide pressure sample infusion ESI reaction monitoring for the discussed vinyl–NHC coupling process. The spectra were acquired in positive ion mode, and the formation of the vinyl-NHC coupling products was observed after 10 minutes. See **Scheme R1** and **Figure R2** for details. The similar shape of the EIC chromatograms is indicative of the similar reactivities of the studied complexes.

After the competition of the reaction, we performed liquid chromatography with mass spectrometry on Bruker maXis Q-TOF mass spectrometer equipped with electrospray ionization source and chromatograph Agilent 1200 equipped with a chromatographic column ZORBAX Eclipse Plus C18 (5 µm, 4.6x150 mm). The formation of new vinyl-NHC coupling products was confirmed according to the HPLC–MS data (see **Figure R3** for details).



Scheme R1. Pressure sample infusion (PSI) MS reaction monitoring (Scheme S3 in the Supporting Information).

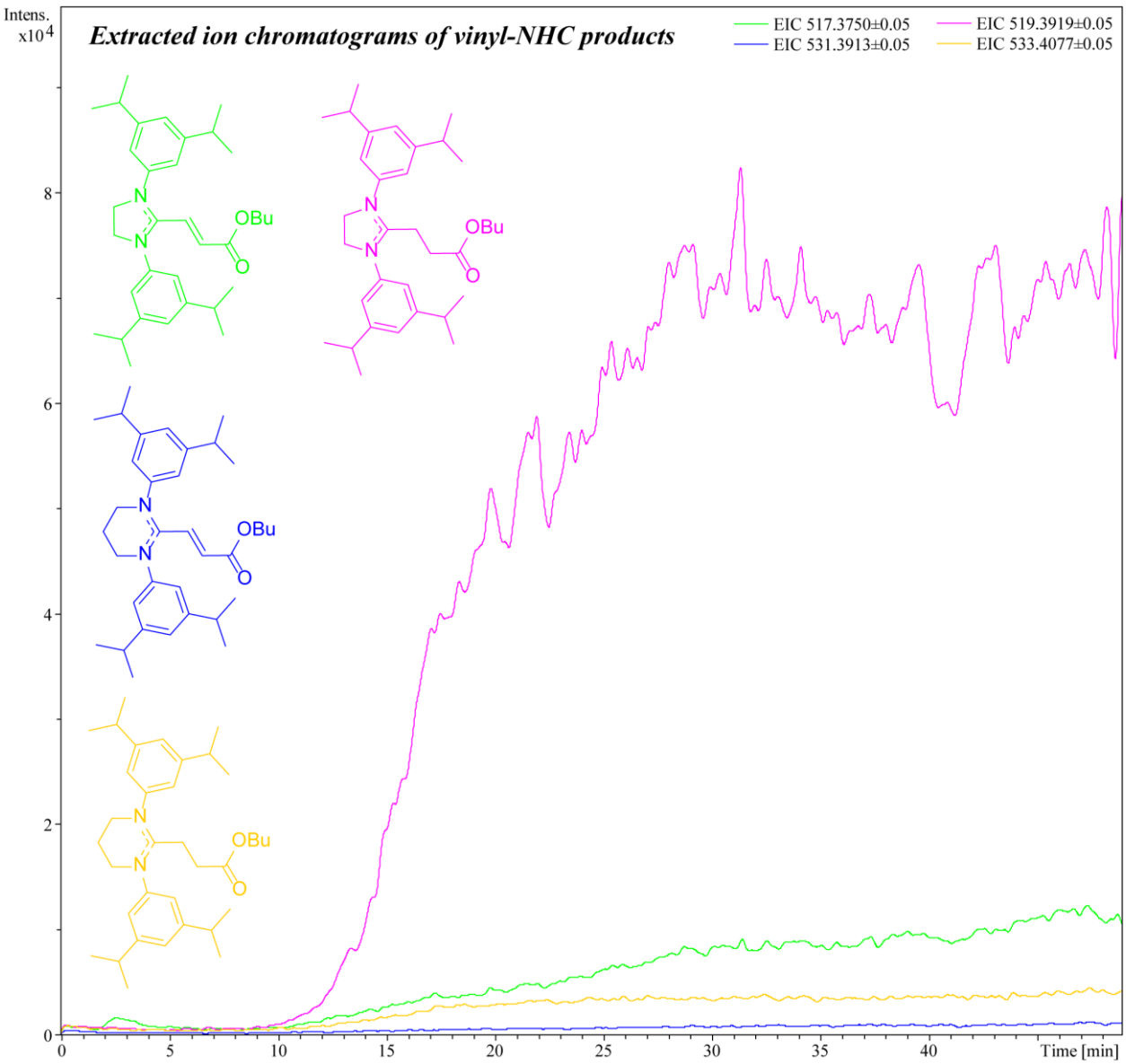


Figure R2. Gauss-smoothed curves (3 sec, 5 cycles) for the real-time abundances of the vinyl-NHC products. The colors of the structures correspond to the colors of the extracted ion chromatograms (Figure S41 in the Supporting Information).

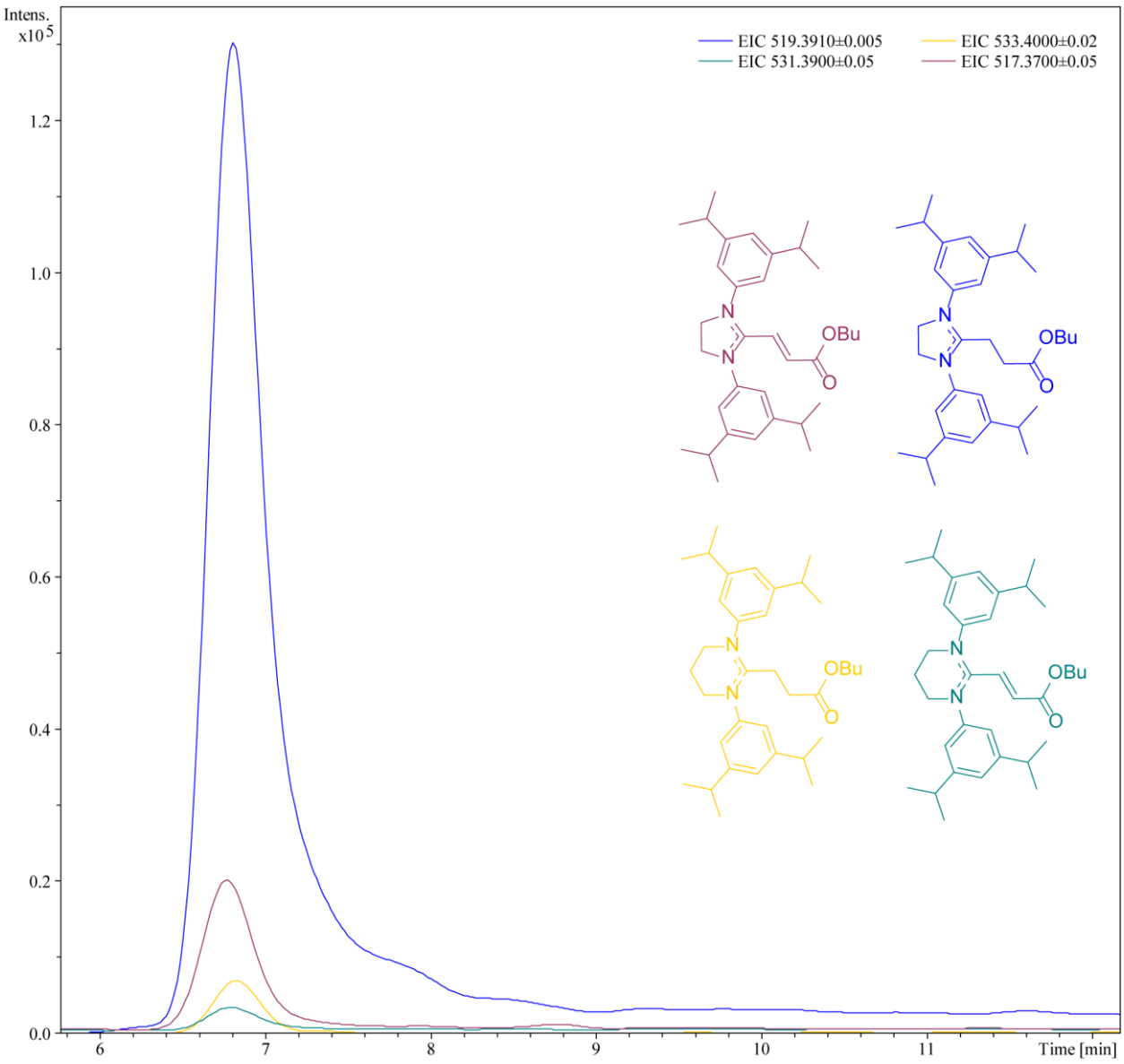


Figure R3. HPLC–MS analysis of vinyl-NHC products. The colors of the structures correspond to the colors of the extracted ion chromatograms (Figure S42 in the Supporting Information).

In all these cases, our algorithm successfully detected the presence of the ions in question. The corresponding discussion was added to the Supporting Information (Section S12.2). Overall, we believe that the described method is robust enough across the scope of common mass spectrometry experiments in organic chemistry. Unfortunately, unlike fields such as metabolomics, organic chemistry does not have a strong practice of data sharing. However, we hope to change this by further promoting and developing this project, initially demonstrating its value with this publication.

3. Figure 3A, t-SNE map needs to be compared to a reference database to describe the breadth of chemical space compared to the authors' database. The author's space should be overlayed with the chemical space described by a chemical library such as PubChem to see if it is representative.

We agree that comparing the t-SNE map to a database such as PubChem would provide insights into the breadth of the chemical space covered by our dataset. To address this, we manually parsed 125 forms that researchers use to submit samples, and we show where they fall in the chemical space. This was used to construct the plot below (**Figure R4**) and in **Figure 3a**.

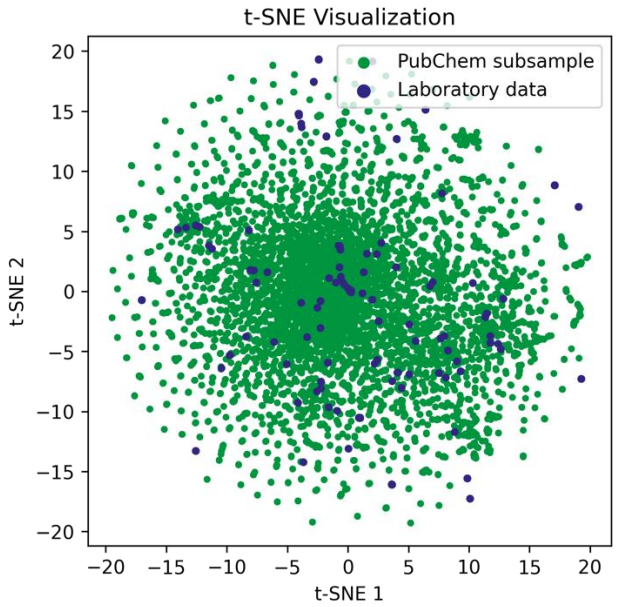


Figure R4. t-SNE plot of the PubChem chemical space with overlaid forms results (Figure 3 in article).

minor: the colored dots in figure 3a legend are too small

Thank you for bringing this up! In our revised figure, we have increased their size.

Minor: a distribution of the spectra by user (A-Other) should be provided. It looks like user C provided > 60% of the data- all coming from a single instrument/method

Thank you for this comment. We acknowledge that operator C provided a significant portion of the data. To provide a clearer picture of the distribution, we added a detailed breakdown of the

spectra by the user to the Supporting Information in **Figure S20** and **Figure R5**. Although such a situation is not ideal, this operator has a limited influence on the diversity of the analyzed samples, so it should not be compromised.

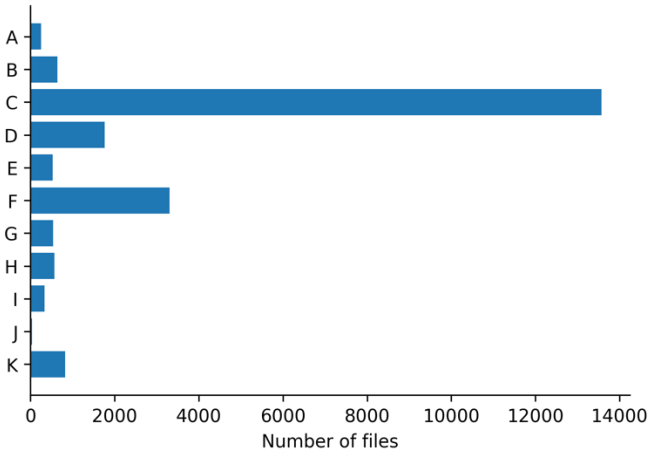


Figure R5. Number of samples recorded by each user.

4. P7, paragraph 2: The authors admit that HRMS without fragmentation can provide information only about molecular formulas; thus structural isomers cannot be distinguished. The authors need to describe how many of the results in figure 3d are isomers as a limitation of this software.

Thank you for your question! In **Figure 3d**, within each group and between them, all the formulas are unique and have different exact masses. Moreover, even the unit masses are different. For the full set of formulas, there were only 400 with unique masses. A corresponding note was added to the paper (page 7, second paragraph; page 8, first paragraph).

5. P7 , paragraph 3: “most of the search engine answers could not be validated due the lack of FAIR description data needed for recognition of the reaction mixture initial composition.” This needs to be quantified as it is a major limitation of the workflow. What FAIR data is minimally required for interpretability of the results of this tool? If these data are unavailable, does the tool still provide value?

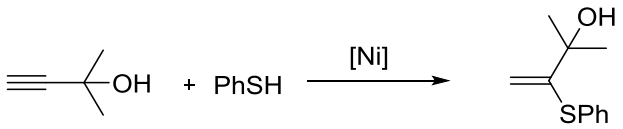
The minimal FAIR data required for the interpretability of our tool include information on the reactants that were mixed. Nevertheless, even without accurate initial composition details, our tool still holds value in such scenarios through the potential formulation of complex queries.

For example, if a user hypothesizes that substances A and B form product C, they could search for spectra where all three substances are mentioned, thus reducing false positives. We discuss this methodology further in the manuscript (page 13, third paragraph), enhancing the utility of our tool even when certain aspects of FAIR data are lacking.

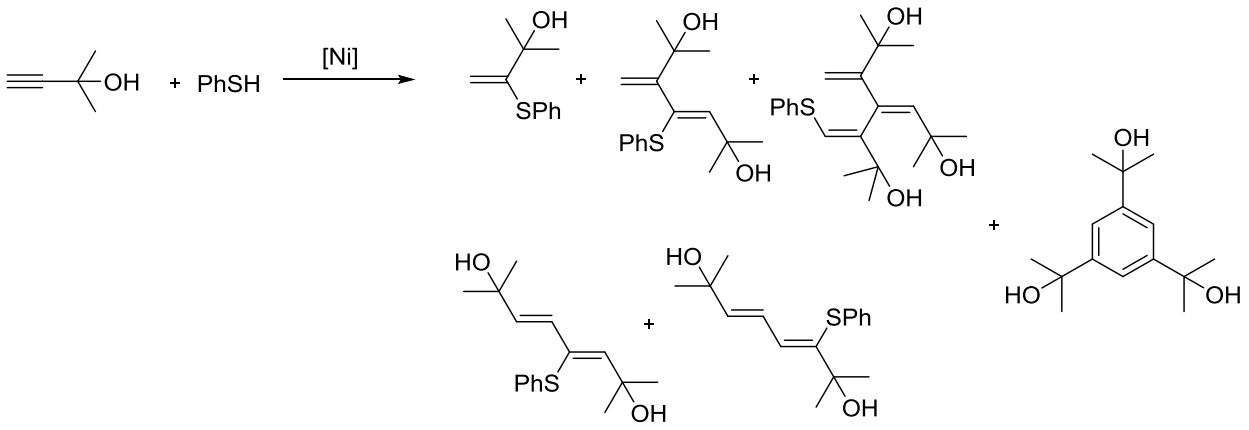
6. How was the Mizoroki-Heck reaction chosen, and can the authors describe if there is any bias in this selection of this reaction based on the chemical space coverage of their database? The authors should also demonstrate rate of hits for several other reactions as well, even without going into a deep validation.

Thank you for this comment. We selected the reactions that are of key importance to synthetic practice and we demonstrated that unusual findings can take place even in very well-known chemistry if a massive (very detailed for a given chemical space) scan of hypotheses is performed. With respect to our background, we have enough expertise to interpret the results of the described organic and catalytic reactions. The database used consists of data collected from different projects (see the answer above with PubChem space overlap).

To further demonstrate an extended set of reactions, we provide the following additional example for the hydrothiolation reaction, which was executed multiple times to optimize the formation of the Markovnikov-type vinyl sulfide product (**Scheme R2,** Scheme S1 in the Supporting Information). However, the overall reactivity network appears to be much more diverse, and the chemical space was underestimated (**Scheme R3,** Scheme S2 in the Supporting Information**)**.



Scheme R2. Nickel-catalyzed hydrothiolation reaction executed to synthesize Markovnikov-type addition product (Scheme S1 in the Supporting Information).



Scheme R3. Chemical space potentially detectable upon analysis of the recorded spectra of the nickel-catalyzed hydrothiolation reaction (Scheme S2 in the Supporting Information).

All these ions were found in historical data (see **Figure R6** for examples, Figure S28 in the Supporting Information). Therefore, we could proceed with experimental validation.

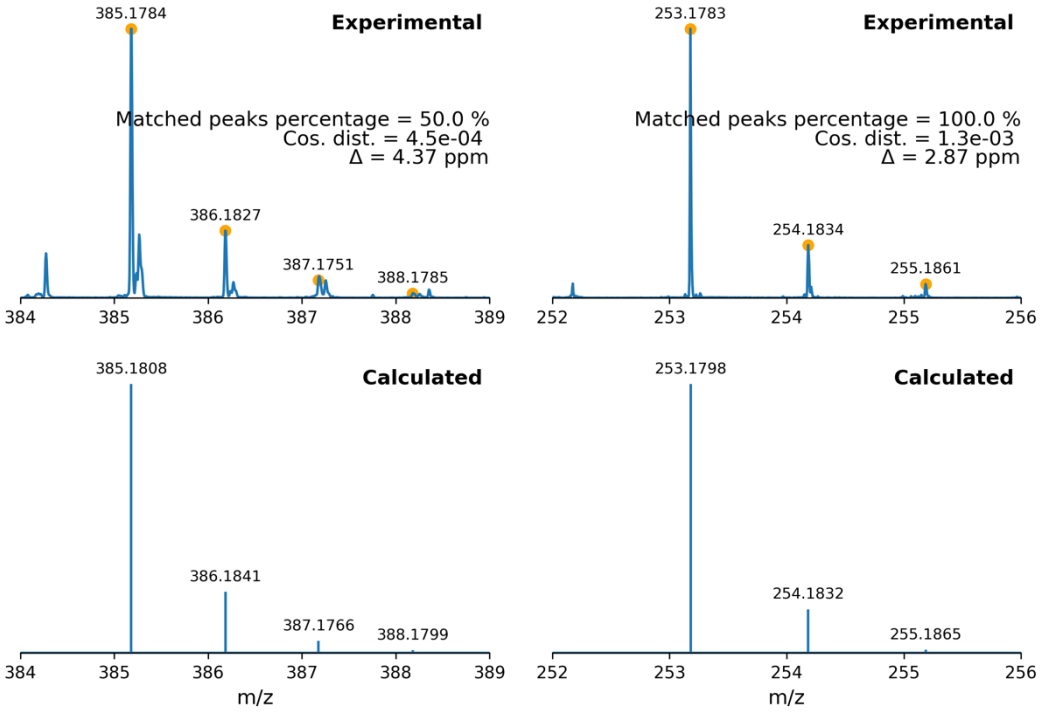


Figure R6. Examples of ions found in the dataset: left — triene (September 2015), right — trimerization product (February 2020) (Figure S28 in the Supporting Information).

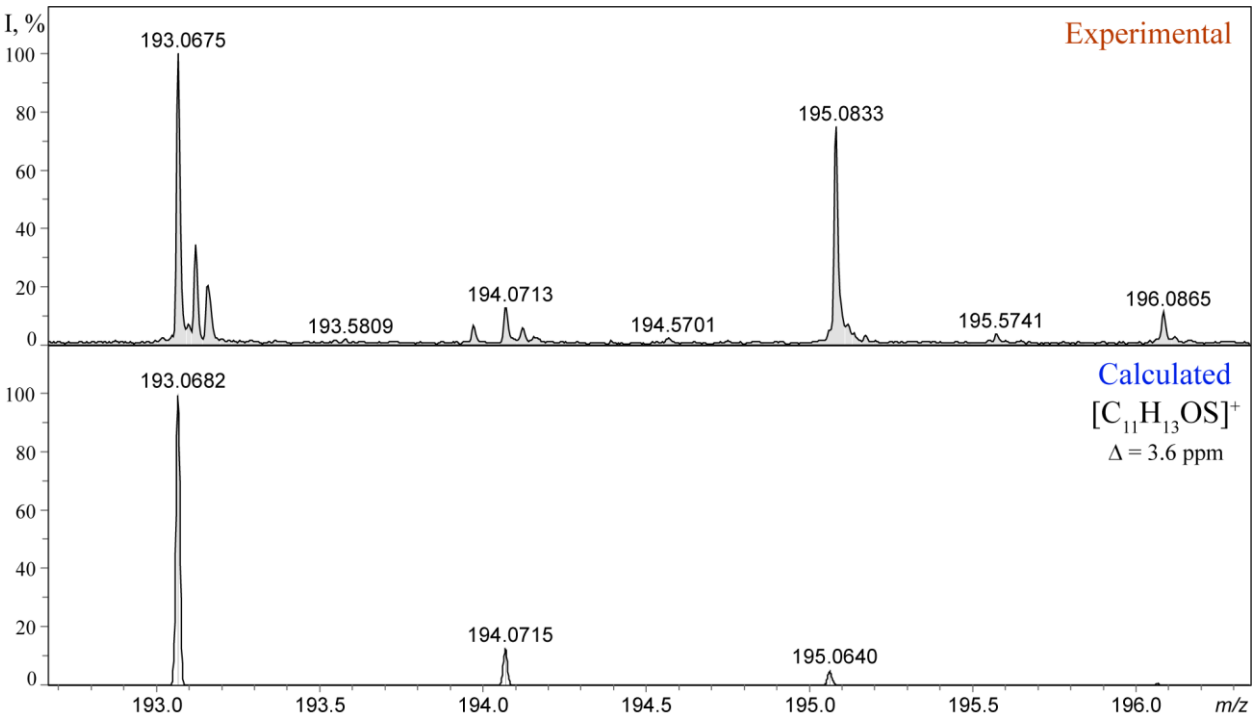


Figure R7. Experimental (top) and calculated (bottom) ESI-(+)MS spectrum of $C_{11}H_{13}OS^+$ ion; experimental peak $[M-H]^+$ = 193.0675 Da, calculated for $C_{11}H_{13}OS$ = 193.0682 Da, Δ = 3.6 ppm (Figure S29 in the Supporting Information).

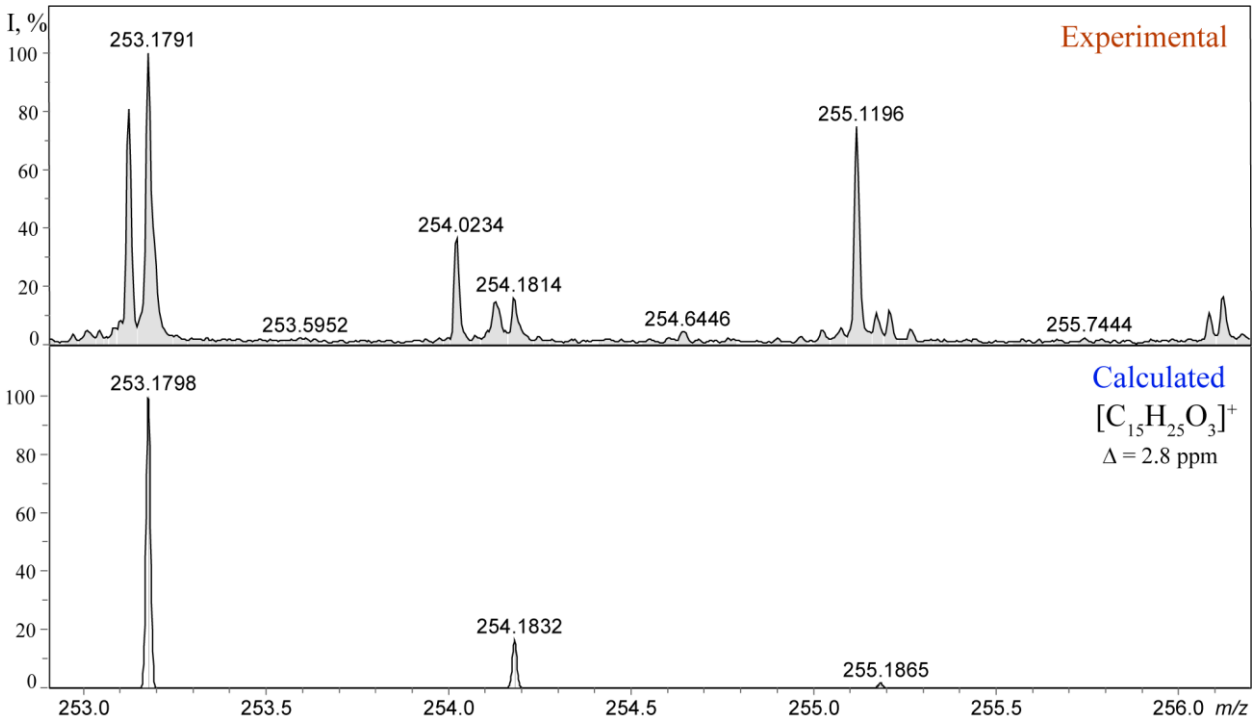


Figure R8. Experimental (top) and calculated (bottom) ESI-(+)MS spectrum of $C_{15}H_{25}O_3^+$ ion; experimental peak $[M+H]^+$ = 253.1791 Da, calculated for $C_{15}H_{25}O_3$ = 253.1798 Da, Δ = 2.8 ppm (Figure S30 in the Supporting Information).

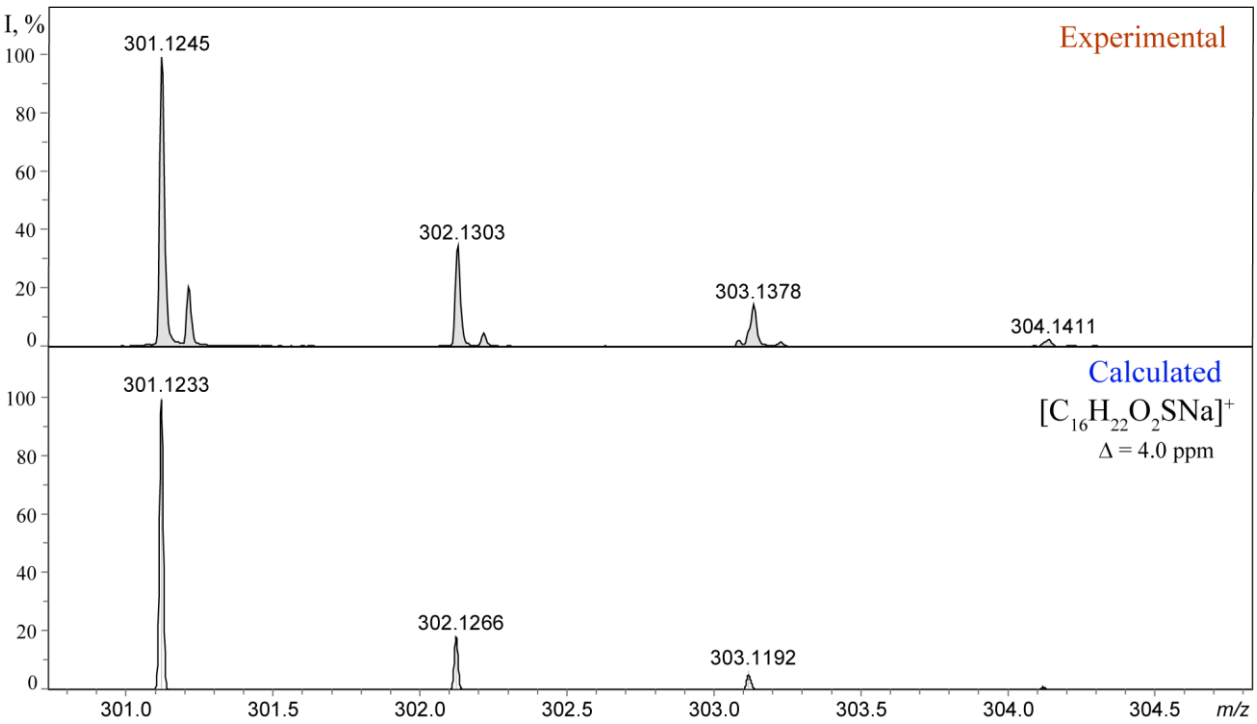


Figure R9. Experimental (top) and calculated (bottom) ESI-(+)MS spectrum of $C_{16}H_{22}O_2SNa+$ ion; experimental peak $[M+Na]^+$ = 301.1245 Da, calculated for $C_{16}H_{22}O_2SNa$ = 301.1233 Da, Δ = 4.0 ppm (Figure S31 in the Supporting Information).

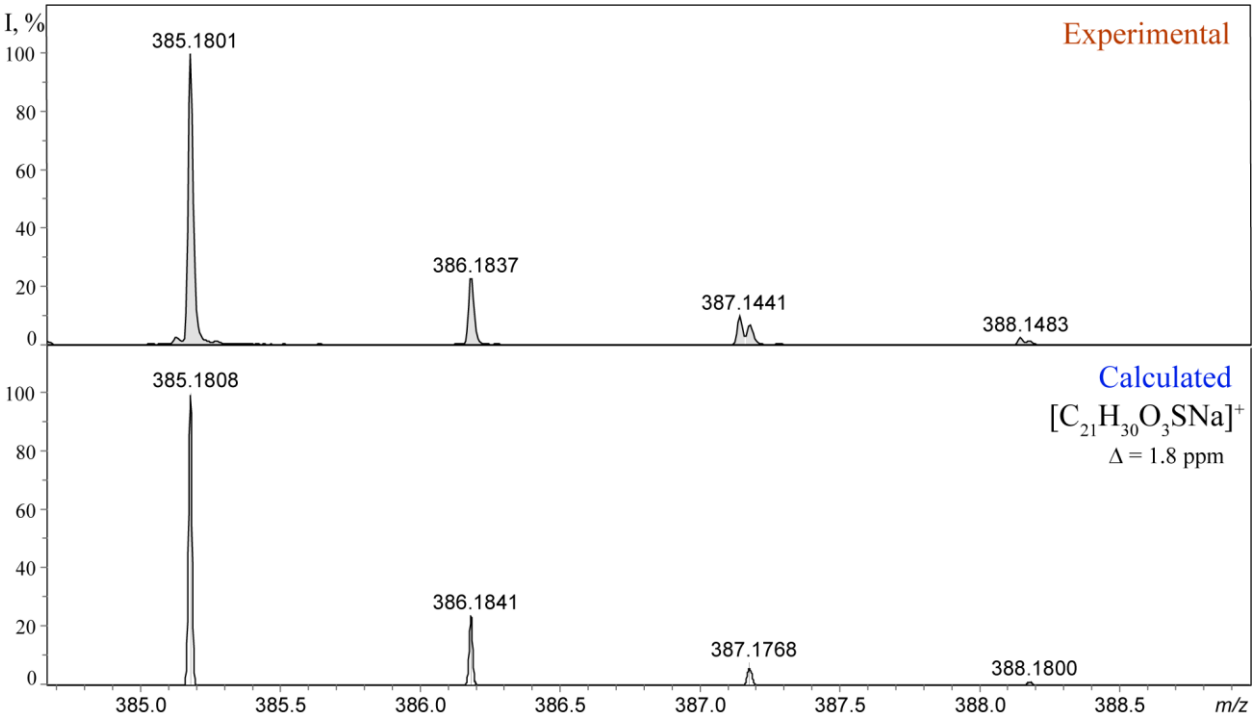


Figure R10. Experimental (top) and calculated (bottom) ESI-(+)MS spectrum of C21H30O3SNa+ ion; experimental peak [M+Na]+ = 385.1801 Da, calculated for C21H30O3SNa = 385.1808 Da, Δ = 1.8 ppm (Figure S32 in the Supporting Information).

To “rediscover” this reaction, we chose nickel-catalyzed (2 mol.% $Ni(acac)_2$) addition of thiophenol to the alkyne at 40 °C for 40 min in the absence of a solvent. The reactions with a PhSH:alkyne ratio of 1:2 yielded a number of byproducts, and the major product was vinylsulfide. After the competition of the reaction, the reaction mixture was independently studied by ESI–MS and NMR spectroscopy. In the mass spectrum, we observed signals corresponding to all types of products. Vinylsulfide –*m/z* 193.0682, dienes with *m/z* 301.1233, triene with *m/z* 385.1808 and the product of alkyne trimerization with *m/z* 253.1798 (**Figure R7-10,** Figure S29-S32 in the Supporting Information).

Now we can investigate this reaction in more details and develop other synthetic methods based on the reactions of alkynes and thiols (will be reported elsewhere).

Importantly, all the compounds shown on **Scheme R3** are easy to generate with combinatorial hypothesis analysis in an automatic manner since the products follow a general formula $(alkyne)_n(thiol)_m$ (where n = 1 .. 3; m = 0 .. 1). Then, after hypothesis generation, the MS search locates the formulas in the mass spectra.

Therefore, analysis of already recorded mass spectra coupled with testing multiple products opportunity is a powerful approach to discover new compounds and then develop new reactions. Of course, a varying performance may be observed depending on the reaction systems, but the approach described in the present study is worth applying, taking into account the existence on many mass-spectral studies.

We added this discussion to the Supporting Information (Section S10.2 in Supporting Information).

Minor comments

1.Abstract: “while obviating the need for conducting additional experiments.” I suggest “reducing” instead of obviating.

Thank you for this comment, correction was made.

2. P 2. Paragraph 2. Suggested text edits for clarity, “in a typical organic synthesis workflow, chemists select particular experimental conditions for the optimization of a reaction to achieve maximal outcome of the desired product. Next, carrying out the reaction and sample preparation is followed by detection and characterization of the chemical compositions of the studied system using an appropriate analytical system (Figure 1b).”

Thank you for this suggestion; this has been corrected.

3. I think Figure 1a is unnecessary, so the overall figure 1 is more simplified.

We agree and have moved panel “a” to the Supporting Information (Section S1 “Reaction data example”).

4. Figure 2. The steps should be labeled 1-4 or A-D according to the text and in sequential order. The figure looks like steps 1 and 2 are occurring in parallel, which is not what is happening according to the text description. The authors can then describe which steps are course or fine searches. The shading of the squares for cosine distance from F1 and F2 should be differentiated by hash filled and non-hash filled so it can be seen easier. Figure 2 caption should be more descriptive.

Thank you for the comment. We have renamed the steps, corrected the shading of the squares, and revised the C2 panel to better reflect the text. The figure caption has also been updated.

5. P7 paragraph 3: “a search pipeline (Figure 2a) was run for each generated ion through the entire tera-scale HRMS database…”. Are the authors referring to the 520 ion formulas? This is unclear. The computational run time for the query of the 520 ion formulas should be stated.

Thank you for the clarifying question. A comprehensive search was conducted for each of the 520 ions generated, with the total run time requiring approximately 3–4 days. We reformulated the text and added information about the search run time in Section «Practical application for the discovery of novel transformation pathways».

6. Page 8, Experimental validation: these sentences are unclear “Formation of these catalyst

transformation….” -which are these referring to? “These reactions were previously described….” which ones and by who? “….with the latter being described recently” which is latter? “As previously found…” by whom? “Apparently the process occurs via some kind of transfer hydrogenation reaction” why?

We thank the Reviewer for this question and have corrected the text on page 8 (page 9 in the revised manuscript). The mechanisms of M/NHC catalysis can be described in two main types (**Figure R11**). The first mechanism involves NHC-connected mechanisms with a metal–CNHC sigma bond; the second mechanism involves NHC-disconnected mechanisms without metal–CNHC sigma bonds (10.1039/d0sc02629h).

The evaluation of Pd/NHC complexes and M-NHC bond cleavage is a convenient subject for MS investigations and has allowed us to discover new types of reactions. When combined with a large database of spectra covering various kinds of chemistry, it becomes a valuable task for evaluation, not limiting the overall applicability of the tools but allowing us to validate the results experimentally.

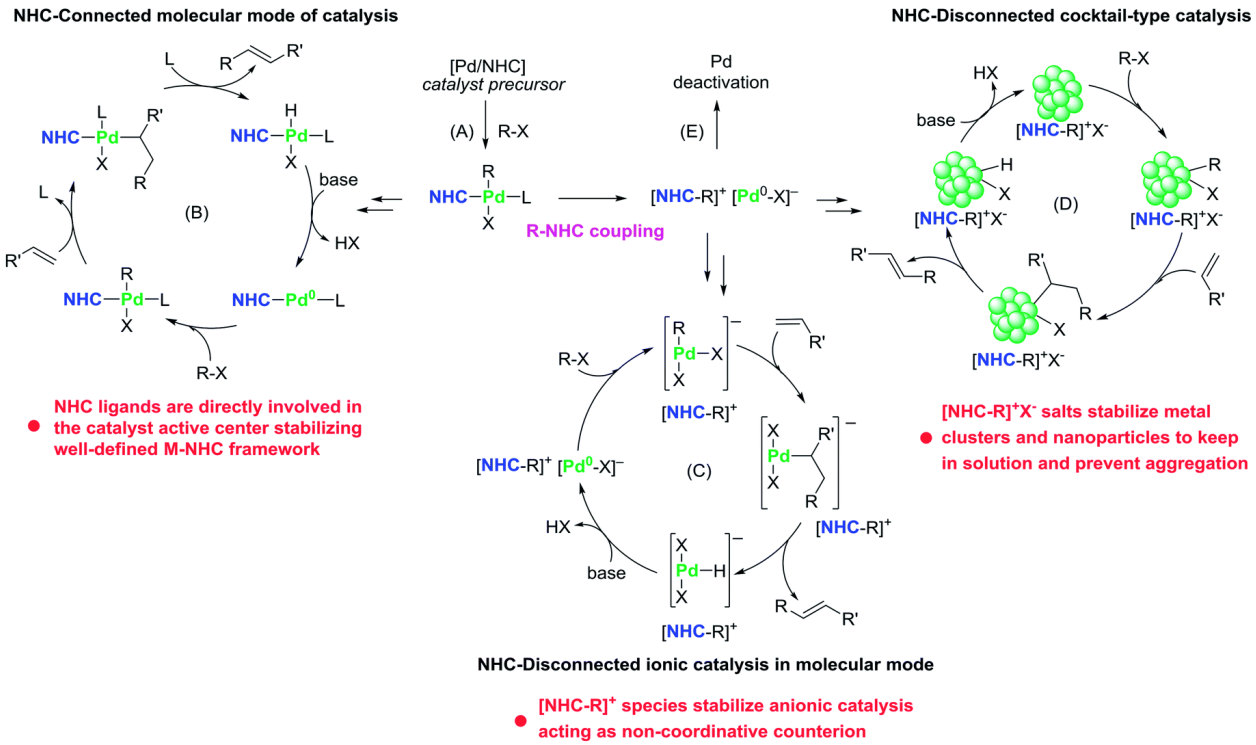


Figure R11. Different modes of operation of Pd/NHC complexes are shown using the Mizoroki–Heck reaction as an example: (A) – initial stage of precatalyst activation; (B) – NHC-connected molecular catalysis; (C) – NHC-disconnected molecular catalysis; (D) – NHC-disconnected cocktail-type catalysis

involving nanoparticles; (E) – catalyst degradation after R–NHC coupling (other pathways of catalyst degradation also exist in each of the catalytic cycles; not shown here to avoid picture overload). (10.1039/d0sc02629h)

7. I had a hard time navigating the SI materials. The order in the main text start with section 3, then S4.1, then S4.2, but then the methods come in at S5 and refer forward and back. The SI data and figures needs to be ordered in a consequative manner to make it easier for the reader to naviagate through without going forward and backgward and all over.

Thank you for the suggestion. We have completely restructured the Supporting Information contents to make navigation easier. We have attempted to organize the references to the Supporting Information sections in a consecutive manner in the main text.

8.Figure 4 and page 7, the collected results:. The text with the results numbered 1-6 need to be connected to the figure in 4. Using m/z may be the best way to make it easier to understand which result matches with which chemical by putting the m/z in the text after the product name in paranthases as well as in the figure next to the structure. Also, where are all of the struures for each chemical? For A there are 4 compounds but only 1 spectra?

Thank you for your comment and suggestion. We corrected the figure and referenced the compounds according to their m/z values. The spectra for known compounds are provided in the Supporting Information (see Section S10) to avoid overloading the figure with data. We also added a reference to this Supporting Information section in the main text (page 7, last paragraph).

9.Page 9, the text should refer to the main text figure 5, but also all of the SI figures of the MSMS data that was collected. Figure 5d, should be re-named 5c, as this is the order of mention in the text.

Thank you for your comment. We have updated the figure and added references to the Supporting Information figures (see page 7, last paragraph in the revised manuscript) in the main text.

We thank the Reviewer for the evaluation of our article and for helpful comments! The revised manuscript was improved!