Referee response

Technical note: Towards atmospheric compound identification in chemical ionization mass spectrometry with machine learning

Federica Bortolussi[1], Hilda Sandström[2], Fariba Partovi[3,4], Joona Mikkilä[4], Patrick Rinke[2,5,6,7], and Matti Rissanen[1,3]

[1]Department of Chemistry, University of Helsinki, 00560 Helsinki, Finland
[2]Department of Applied Physics, Aalto University, Espoo, Finland
[3]Aerosol Physics Laboratory, Physics Unit, Tampere University, 33720 Tampere, Finland
[4]Karsa Ltd., A. I. Virtasen aukio 1, 00560 Helsinki, Finland
[5]Physics Department, TUM School of Natural Sciences, Technical University of Munich, Garching, Germany
[6]Atomistic Modelling Center, Munich Data Science Institute, Technical University of Munich, Garching, Germany
[7]Munich Center for Machine Learning (MCML)

Correspondence: Federica Bortolussi (federica.bortolussi@helsinki.fi)

We want to thank the Reviewers for the careful evaluation of this work and for providing comments and suggestions to improve the manuscript's quality. Below, we will address each comment in detail. The Reviewer's comments are displayed in black bold; the author's responses are shown in regular black, and manuscript revisions are highlighted in blue. Long sections of the text that remain unchanged between two updated paragraphs are indicated by "[...]".

Response to Reviewer #1

This paper uses analysis of pesticide standard materials by a variety of chemical ionization techniques to test the utility of machine learning methods in identifying a) whether a given molecule will be identifiable by a given CIMS ionization technique and b) how sensitive the CIMS technique will be to said detectable compound. Additional analysis investigates what structural characteristics (via elements of molecular descriptors) drive these capabilities. I congratulate the authors on the extremely impressive battery of instrumentation and analytical techniques used in this work and believe that it has the potential to significantly benefit the atmospheric chemistry community. In its current form however the manuscript has significant issues in framing that must be addressed before I can recommend it for publication.

We thank you for the positive feedback.

1. **First, the title and framing of this work must be updated to reflect that this work represents an analysis of pesticides, which are not representative of atmospheric organics.**

Our study utilizes CIMS, a technique widely used in atmospheric compound research, and our ultimate aim is to apply the methods developed here to atmospheric compounds. However, we currently lack access to sufficiently large reference datasets for a vast majority of atmospheric compounds causing secondary aerosol formation to achieve this goal directly. Additionally, many of the direct aerosol precursor structures, such as highly oxygenated organic molecules correspond to polyperoxide compounds that are almost certainly high-explosives in the condensed phase. Therefore, they are also unlikely to become available in the foreseeable future.

Given this limitation, we identified a dataset of pesticides—an accessible and annotated reference set—as a practical starting point for testing. While pesticides represent only a minor subset of atmospheric compounds (as noted in our manuscript with references to Brüggemann et al. (2024) and Houde et al. (2019)), they encompass a wide range of molecular sizes and functional groups, allowing us to explore the CIMS response across diverse ionization interactions. This structural diversity makes pesticides a suitable initial test case for developing and validating ML-based CIMS signal prediction methods. Additionally, pesticides are readily available as standard solutions from chemical suppliers at a manageable cost, which facilitates broader testing and replicability.

We note in our manuscript that pesticides only constitute a minor fraction of atmospheric compounds, and we highlight our choice of this dataset as a first step in developing ML-based tools for broader compound identification in atmospheric research.

The motivation for the use of pesticides instead of other atmospheric-related compounds has been updated in the Introduction section:

"In this work, we address the scarcity of atmospheric compound data standards by testing our methodology on a reference dataset of approximately 700 pesticides measured with CIMS. While pesticides represent only a small subset of atmospheric compounds (Brüggemann et al., 2024; Houde et al., 2019), they are chemically complex, with diverse molecular masses and functional groups that can interact in distinct ways with various reagent ions and that cover an extended range of detection with CIMS. This structural diversity provides a relevant test case that reaches and surpasses the complexity of many atmospheric compounds, allowing for an effective initial test of our methodology. Additionally, pesticides are readily available as standard chemicals from chemical suppliers at an accessible cost, and the dataset size is comparable to those used to establish early ML compound identification tools in metabolomics (Heinonen et al., 2012; Dührkop et al., 2015; Brouard et al., 2016; Nguyen et al., 2018, 2019). Thus, while limited to pesticides, this dataset offers a valuable preliminary benchmark for developing ML-based CIMS signal prediction. Once reference datasets for atmospheric compounds become available, this methodology can be directly applied or refined to encompass a broader range of atmospheric chemical analyses."

We have also updated the Title to reflect the application of our methodology to pesticides. The new Title is:

"Technical note: Towards atmospheric compound identification in chemical ionization mass spectrometry with pesticide standards and machine learning"

2. **Second, although this is a technical report, the authors must clearly articulate the goal and justification for the many methods applied.**

This is an important point. Our paper presents a proof-of-concept demonstration of machine learning (ML) predictions for CIMS sensitivities. The main advantage of our ML-based approach lies in its computational efficiency, especially in comparison to traditional quantum chemistry methods, and in its ability to interpolate to novel compounds without relying on extensive tabulated data once trained. We updated the Introduction section as follows:

"The advantage of an ML-based method is twofold: it is computationally inexpensive, especially when compared to quantum chemical calculations, and it can interpolate predictions to novel compounds without requiring extensive reference data once trained. This is essential for atmospheric chemistry, where thousands of large, highly oxidized organic compounds lack reference datasets. In the short term, our method could accelerate CIMS experimental optimization and aid in reagent ion selection. However, successful identification requires comprehensive collections of reference spectra, which are needed both for traditional spectral comparison and for training ML-based methods. Currently, a lack of data standards in atmospheric science hinders similar ML advancements for CIMS and fragmentation mass spectrometry (Sandström et al., 2024; Thoma et al., 2022)."

We have also clarified the utility of molecular representations and testing later in the introduction:

"The models are trained on molecular descriptors, which are mathematical transformations of the molecular structure that make it suitable for data-driven analysis. Different descriptors are tested and compared for both classification and regression, as each molecular representation encodes unique structural or chemical features and varies in complexity and interpretability. We tested five different representations: properties obtained from the pesticides' structure (RDKit-PROP), the topological fingerprint (TopFP) (James et al., 1995), the molecular access system keys (MACCS) (Durant et al., 2002), the Coulomb matrix (CM) (Rupp et al., 2012), and the many-body tensor representation (MBTR) (Huo and Rupp, 2022). Using this range of molecular representations and data from diverse ionization schemes, we evaluate the models' ability to predict CIMS detection and signal intensity of the compounds, providing insights into how structural characteristics influence CIMS sensitivity across different ionization methods."

We additionally included an explanation of the use of a classifier and a regression method in Section 4. Ideally, a larger dataset, with a more balanced distribution of signal intensities (including undetected signals), would allow a direct regres-

sion approach to predict both detection probability and signal intensity, accompanied by an in-depth chemical analysis. Due to the current dataset's limited size and imbalance of observations, we employed a classification approach to address the prevalence of undetected compounds. This combined classification-regression approach is necessary for a thorough investigation, as focusing on one alone would limit the study's depth. The paragraph in Section 4 is now:

"In this section, we briefly introduce the two ML methods that we use in this work. Figure 2c presented a potential problem for the direct training of a regression model: for individual ionization methods, the data is imbalanced with a relatively high amount of undetected pesticides. This imbalance suggests that there might not be enough instances to train a model able to generalize patterns and signals of the molecules, potentially leading to poor predictive performance. To tackle this problem, we decided to divide the CIMS signal prediction into a classification task and a regression task. To classify, if a pesticide is detectable or not with a specific ionization method, we will train a RF classifier. Subsequently, we will investigate, if we can predict the corresponding CIMS intensity with KRR."

3. **Third, references and comparisons to alternative methods, including fundamental chemistry related to proton affinity and other aspects of ionization chemistry that are typically used in identifying whether or not a given analyte is likely to be detectable by a given ionization method, are almost completely excluded. These comparisons form a critical foundation in establishing if and how these methods may be useful to atmospheric chemists and would be an invaluable sanity check on whether the machine learning methods used in this study are successfully identifying known reaction phenomena.**

We appreciate this insightful comment regarding alternative methodologies for characterizing ionization, such as proton affinity, a key parameter in proton transfer reaction mass spectrometry (PTRMS). Commonly volatile organic precursor molecules present at relatively high concentrations are routinely quantified utilizing PTRMS instruments. However, for direct highly functionalized aerosol precursors more selective and sensitive techniques, such as anion attachment applying NO_3^- and I^-, for example, have been generally applied. While some property databases offer information for standard compounds, our study focuses on developing ML methods to predict CIMS sensitivities across a range of reagent ions and ionization mechanisms. We aim to extend this capability to include non-standard compounds, such as highly oxygenated organic compounds, which lack tabulated properties and would otherwise require time-intensive quantum chemical property predictions. Therein lies the added value of our work for atmospheric chemists.

The goal of our study is to learn generalized patterns of detection across a molecular population, enabling us to interpolate to other molecules similar to those analysed when constructing our model. Molecular descriptors are ideal for this, as they efficiently capture a wide range of molecular properties without extensive computation. To test the ability of the ML methods applied, we included descriptors from RDKitPROP, which can be derived from SMILES representations, and offer indirect correlations to proton affinity. Examples include the number of hydrogen bond acceptors and the topo-

logical polar surface area, among others. Only a subset of these were critical to model performance, as highlighted in the main text. We added in the Introduction:

"In proton transfer reaction MS, properties like proton affinity are utilized to determine the detectability of compounds. Although this instrument is commonly used to quantify volatile organic precursor molecules at relatively high concentrations, more selective and sensitive techniques are typically required for analyzing highly functionalized aerosol precursors (e.g. NO_3^- or I^- (Lee et al., 2014; Rissanen et al., 2014))."

We also added comments in Sections 3.1 and 5.3, elaborating on our gained chemical insight regarding sensitivity prediction and comparison with other prediction methods. At the beginning of section 3.1, we added:

"RDKitPROP includes 43 properties computed from the molecular structure of the pesticides (represented by a SMILES string), by applying the function *rdkit.Chem.rdMolDescriptors.Properties* (Landrum, 2006). This descriptor was included in the analysis to evaluate the models' performance based on known properties that are computationally inexpensive to obtain."

In section 5.3 we added before discussing MACCS results, and as a final comparison and conclusion:

"We observe that several of the identified important features relate to proton affinity. The number of HBA (NumHBA, LipinskiHBA) is directly correlated to proton affinity as it calculates the number of sites available to accept a proton. The TPSA describes the molecule's polarity, and for certain molecules, a higher TPSA could correlate with a higher proton affinity. HallKierAlpha correlates as well since every atom's covalent radius is adjusted for its hybridization state and electronegativity, reflecting the likelihood of atoms within a molecule to donate electron density to a proton. FractionCSP3, while not correlating directly to proton affinity, might influence the overall basicity of the molecule (e.g. a higher amount of sp^3 carbons in the molecule potentially affects the electron density of heteroatoms indirectly).
We note that different reagent ions react with the analyte in distinct ways (see the Introduction). While our results support the predictive nature of properties like proton affinity, specifically for positive reagent ions, the ML model's advantage lies in its flexibility. As shown, this approach aligns well with established knowledge, yet the ML methodology combined with molecular representations can relate any reagent ion or ionization mechanism to the magnitude of CIMS signals using only the analyte molecular structure.

[...]

Overall, RDKitPROP and MACCS in combination with RF have given us valuable insights into CIMS ion-molecule interactions. In the case of positive polarity ionization methods, the results obtained with the chemical insight analysis

support known alternative methods of identifying whether a molecule can be detected, by highlighting a series of properties that can relate to proton affinity."

4. **In its current form, the manuscript's focus is so broad and technical and references to the potential use case of the different predictions are so incompletely addressed that the potential utility of the methods used is extremely difficult to identify. Overall, I would suggest that the authors consider a broader atmospheric chemistry audience in this work's structure, meaning that a justification of why each method is selected, what it is intended to predict, and how those predictions would be useful for a broader atmospheric chemistry community should be included at the beginning of each section. Finally, as this work exclusively operates in a forward direction from a known compound to predict its detection by CIMS, either additional analysis must be performed to evaluate how it might operate backwards from CIMS data to a prediction of identity, or the paper title must be re-framed away from a claim of providing compound identification. I again applaud the authors in this extremely impressive body of work and wish them success in re-framing the manuscript.**

Thank you again for the positive feedback on this work. Traditional methods for compound identification or annotation typically rely on comparing an unknown compound's spectrum to reference spectra in a database to find the closest match. Current state-of-the-art ML models for compound identification still require these reference databases. Our ML models are the first to introduce a data-driven approach for CIMS signal prediction in complex compounds, providing a groundwork for future tools in atmospheric compound identification. To enable robust compound identification from CIMS data in the future, a comprehensive and standardized dataset of atmospheric compounds would be required, ideally collected through international collaboration. We hope this work inspires further efforts towards such a dataset.

In the Conclusions, we have revised the text to clarify and outline the steps required for future applications in compound identification. Additionally, we have revised the title to reflect the forward prediction focus of this work more accurately. We believe these changes address the concerns about positioning the work accurately within the landscape of CIMS and ML based compound identification. The Conclusions were updated as follows:

"In summary, we developed a ML workflow for predicting the detection with CIMS (with a classification algorithm) and CIMS sensitivity to molecules (with a regression algorithm) to improve atmospheric compound identification. The goal is to evaluate if our ML model can accurately predict detections and signal intensities, thus offering a foundation to build a database of simulated compounds' signals for compound identification purposes with CIMS. Currently, compound identification is typically achieved by comparing an unknown compound's spectrum to a reference database. While this work does not provide direct identification of unknown compounds, it establishes a methodology for developing such a database, which could be expanded for broader use in atmospheric chemistry.

[...]

The results demonstrate that it is possible to extract predictive information even in small experimental datasets. However, more instances could help to generalize the structural features better and help prevent class imbalance problems. Currently, our ML models are directly applicable to predicting the detection and signal intensity of molecules with molecular structures similar to those in our dataset. For molecules with more diverse structures, transfer learning approaches could use these trained models as a baseline, updating learned parameters to accommodate the characteristics of new structures. Applying our approach directly to field measurements will require a comprehensive, standardized dataset of atmospheric compounds with a limited number of reagent ions for practical applications. Such a dataset could facilitate accurate mapping of ionization tendencies, potentially enabling compound identification directly from field CIMS measurements in the future."

Moreover, we agree that in its current form, the potential utility of the methods utilized is difficult to identify. In addition to the modified Conclusions, we updated the manuscript to consider a broader atmospheric chemistry audience.

Citing comment #2, we added in the Introduction:

"Using this range of molecular representations and data from diverse ionization schemes, we evaluate the models' ability to predict CIMS detection and signal intensity of the compounds, providing insights into how structural characteristics influence CIMS sensitivity across different ionization methods."

We added at the beginning of Section 3:

"A molecular representation is a transformation of a molecular structure that simplifies the structural information into a readable input for data-driven methods. Depending on the application, they can provide a valuable cost-efficient alternative to computationally expensive quantum chemical computations. These descriptors are numerical representations of atomistic systems that should fulfil certain requirements, such as being invariant to spatial and rotational transformations, invariant to permutation of atomic indices, unique, continuous, compact and computationally efficient (Himanen et al., 2020; Huo and Rupp, 2022; Rupp, 2015; Xue and J, 2020; Langer et al., 2022). Molecular descriptors may vary in complexity and interpretability; some reflect tangible properties that are easy for humans to understand, while others are calculated through mathematical means and may lack intuitive interpretation. However, a universal descriptor able to perform well for every chemical system and task does not exist. For this reason, being a first-of-a-kind study, we tested five different descriptors (Fig. 1a) for our classification task (prediction of the detection) and regression task (prediction of the CIMS signal intensity)."

Citing comment #2, we added at the beginning of Section 4:

"In this section, we briefly introduce the two ML methods that we use in this work. Figure 2c presented a potential problem for the direct training of a regression model: for individual ionization methods, the data is imbalanced with a relatively high amount of undetected pesticides. This imbalance suggests that there might not be enough instances to train a model able to generalize patterns and signals of the molecules, potentially leading to poor predictive performance. To tackle this problem, we decided to divide the CIMS signal prediction into a classification task and a regression task. To classify, if a pesticide is detectable or not with a specific ionization method, we will train a RF classifier. Subsequently, we will investigate, if we can predict the corresponding CIMS intensity with KRR."

We added at the beginning of Section 5:

"In this section, we present and evaluate the performance of our trained models. For the classification, we examine the ability of our RF models to predict the detected or undetected compounds in the test set. For the regression, we investigate whether our KRR models can accurately predict the CIMS sensitivity of the test set compounds. Furthermore, we analyze which descriptors most effectively enhance model performance and explore whether chemical qualitative insights can be derived from them."

5. **Pesticides do not present a representative sample of atmospheric compound composition; for example, they are significantly biased towards halogenated species and phosphates. I recommend that the title and focus of the manuscript be altered to reflect the pesticide focus area. The implications for predicting properties of atmospheric compounds more generally should be described in a conclusions or implications section.**

In addition to our motivation outlined in response to comment #1, we here cite the modified Title and the modified Introduction with the motivation behind working with the pesticides:

"Technical note: Towards atmospheric compound identification in chemical ionization mass spectrometry with pesticide standards and machine learning"

"In this work, we address the scarcity of atmospheric compound data standards by testing our methodology on a reference dataset of approximately 700 pesticides measured with CIMS. While pesticides represent only a small subset of atmospheric compounds (Brüggemann et al., 2024; Houde et al., 2019), they are chemically complex, with diverse molecular masses and functional groups that can interact in distinct ways with various reagent ions and that cover an extended range of detection with CIMS. This structural diversity provides a relevant test case that reaches and surpasses the complexity of many atmospheric compounds, allowing for an effective initial test of our methodology. Additionally, pesticides are readily available as standard chemicals from chemical suppliers at an accessible cost, and the dataset size

is comparable to those used to establish early ML compound identification tools in metabolomics (Heinonen et al., 2012; Dührkop et al., 2015; Brouard et al., 2016; Nguyen et al., 2018, 2019). Thus, while limited to pesticides, this dataset offers a valuable preliminary benchmark for developing ML-based CIMS signal prediction. Once reference datasets for atmospheric compounds become available, this methodology can be directly applied or refined to encompass a broader range of atmospheric chemical analyses."

Additionally, as discussed in comment #4, we added a paragraph in the conclusions for predicting properties of atmospheric compounds:

"The results demonstrate that it is possible to extract predictive information even in small experimental datasets. However, more instances could help to generalize the structural features better and help prevent class imbalance problems. Currently, our ML models are directly applicable to predicting the detection and signal intensity of molecules with molecular structures similar to those in our dataset. For molecules with more diverse structures, transfer learning approaches could use these trained models as a baseline, updating learned parameters to accommodate the characteristics of new structures.

Applying our approach directly to field measurements will require a comprehensive, standardized dataset of atmospheric compounds with a limited number of reagent ions for practical applications. Such a dataset could facilitate accurate mapping of ionization tendencies, potentially enabling compound identification directly from field CIMS measurements in the future."

6. **How was fragmentation accounted for when determining whether or not a species was detectable? Was only the parent ion included in the search?**

Thank you for this question. Fragmentation was not accounted for when determining species detectability. The analysis we performed was focused solely on the parent ions, as they are straightforward to identify and are more reliable indicators in complex, unknown samples as encountered in field deployments. In principle, some molecules might still be identifiable through characteristic fragments. However, a key advantage of atmospheric CIMS lies in its soft ionization, which inherently minimizes fragmentation. This allows for the stable detection of primary ions, even in complex environmental matrices, where a fragment could result from several similar species. For instance, organosulfates show characteristic fragments like SO_3^- and HSO_4^- across several mass spectrometry methods, yet HSO_4^- is also the ion from which atmospheric H_2SO_4 is quantified. We updated the main text in lines 72-73 as follows:

"The measurements from the two mixtures were combined into a single dataset for a total of 716 pesticide observations, where each observation corresponds to the parent ion's signal intensity. Due to CIMS' soft ionization, the parent ion

is expected to have the highest intensity, quantitatively, and qualitatively provides a one-to-one correspondence to the target compound."

7. **Line 114: please justify this assertion- could the differences in intensity instead reflect differences in the kinetics or thermodynamics of reactions leading to differing yields of charged parent ions?**

It is indeed possible that the observed differences in parent ion signal intensity stem not only from interaction at different sites on the analyte molecule but also from variations in the reagent ions' properties, which could influence the strength of interaction or the ionization kinetics. The relatively weak correlation observed between the ionization methods may suggest that these factors—different interaction sites or distinct ionization kinetics—contribute to the differences in signal intensity. This highlights qualitative differences in how each reagent ion interacts with the target molecules in this study. The text has been updated as follows:

"The general lack of correlation between opposite polarity ionization schemes indicates that different reagent ions interact with the target molecules in distinct ways, possibly engaging with different functional groups."

8. **126: please justify why the ability of t-sne to cluster compounds based on which ionization chemistry they can be analyzed with is a justification of the ML approach- what underlying characteristics of the molecules does this reflect? Molecular characteristics rather than appropriate ionization chemistry for a given analyte is the stated methodological aim of this work and would be a more compelling target property for clustering or modelling.**

Thank you for the comment. We agree that the clustering of ionization signals alone does not justify the machine learning approach (see corrected sentence below) used in this study. Instead, we propose that the value of MION MS lies in its ability to capture complementary information across different reagent ions, each probing distinct compounds or structural features. Ideally, each reagent ion would yield unique signal intensities, providing a nuanced intensity profile that reflects molecular structure. The t-SNE visualization allows us to examine how frequently signals from different reagent ions qualitatively co-occur, offering insight into selectivity and interaction diversity among the ions.

This clustering does not convey direct molecular characteristics. In the main text, we initially stated:

"The presence of clear clusters suggests that the CIMS signals carry information that might allow ML to distinguish different pesticides."

which we have now corrected to read:

"The presence of clear clusters suggests that, collectively, the reagent ions have the potential to differentiate between molecular structures."

9. **Please be more explicit regarding what characteristics of the molecules each descriptor is intended to elucidate, why they are important for the atmospheric chemistry community, and how they are not currently adequately approximated by current CIMS analysis methods. There are so many methods applied that the goal is getting lost. The molecular descriptors section would be significantly improved by a clear explanation of each descriptor for which modelling was attempted and a justification for why being able to predict this property would be useful for the broader atmospheric chemistry community. This should come at the beginning of the molecular descriptors section (before section 3.1)**

Molecular descriptors are sets of features that describe the structure of a molecule, allowing ML models to predict the type of CIMS signal a molecule will produce during analysis. These descriptors encode different aspects of a molecule, enabling us to forecast its behaviour in specific situations, like in a CIMS measurement.

Different descriptors focus on various parts of a molecule. Some may analyze the types of atoms present, while others examine how those atoms are arranged or bonded, providing different levels of detail. By testing a range of descriptors, we can identify which ones best assist our models in predicting the CIMS signals of various pesticides and for different reagent ions.

Understanding these features and determining which descriptors perform effectively for CIMS spectrum prediction is crucial for the atmospheric chemistry community that employs CIMS to study diverse compounds. This knowledge can lead to enhanced methods for detecting and identifying significant compounds. In future work, we plan to further test these descriptors using datasets of atmospheric compounds and their corresponding CIMS signals as they become available. We have clarified these points by revising the text preceding Section 3.1 (mentioned as well in comment #4):

"A molecular representation is a transformation of a molecular structure that simplifies the structural information into a readable input for data-driven methods. Depending on the application, they can provide a valuable cost-efficient alternative to computationally expensive quantum chemical computations. These descriptors are numerical representations of atomistic systems that should fulfil certain requirements, such as being invariant to spatial and rotational transformations, invariant to permutation of atomic indices, unique, continuous, compact and computationally efficient (Himanen et al., 2020; Huo and Rupp, 2022; Rupp, 2015; Xue and J, 2020; Langer et al., 2022). Molecular descriptors may vary in complexity and interpretability; some reflect tangible properties that are easy for humans to understand, while others are calculated through mathematical means and may lack intuitive interpretation. However, a universal descriptor able to perform well for every chemical system and task does not exist. For this reason, being a first-of-a-kind study, we tested five different descriptors (Fig. 1a) for our classification task (prediction of the detection) and regression task (prediction

of the CIMS signal intensity)."

10. **Section 4.3: Can you please explain the performance metrics more clearly? Is each prediction actually a Boolean yes-no of whether the model correctly predicted the molecular descriptor? Is there no nuance on whether a prediction is "closer" or "farther off"? Can you please clarify how "undetected" compounds are being included in this analysis? Earlier the same wording was used to describe compounds that were excluded from analysis?**

Thank you for the comment. We first want to clarify that molecular descriptors are not the target of prediction in our study, but the detection and signal intensity of the molecules and the models use molecular representations to predict them. We are assuming that this comment refers to the classifier's performance metrics. In the case of the random forest classifier, the prediction is probabilistic. The model outputs probability scores for each class, and a threshold is set to determine the final label of the class. In this work, we kept the default threshold, which is 0.5 (which in a binary classification can range between 0 and 1). The accuracy metric is calculated by comparing the predicted class to the reference class. The receiver operating characteristic curve can visualize the relationship between true positives and false positives as the mentioned threshold varies. The classifier is trained to identify "detected" or "undetected" compounds; however, the latter are excluded from the regression task, which only predicts the signal intensity of detected compounds. These details were mentioned in Section 4.4 Computational details, but introduced in the explanation of the *area under the curve* in Section 4.3. To improve clarity, we updated Section 4.3:

"Different performance metrics will be adopted to evaluate the performance of the classifier and the regressor methods. For the classification task, the performance will be evaluated using two metrics: accuracy and the receiver operating characteristic (ROC) curve. The accuracy score is the fraction of correct predictions compared to the total number of observations present in a test set:

$$\text{accuracy}(y, \hat{y}) = \frac{1}{n} \sum_{i=0}^{n-1} 1(\hat{y}_i = y_i) \tag{1}$$

where $\hat{y}_i$ is the i-th predicted class, y_i is its reference class and n the number of samples in the test set.
In the case of the RF classifier, the model outputs probability scores for each class, and then a threshold is applied to determine the final class label, the predicted class. The ROC curve provides us with an additional performance assessment. The curve puts the correctly classified pesticides (true positive rate, vertical axis) in relation to the incorrectly classified ones (false positive rate, horizontal axis), across a range of different threshold levels. By varying the threshold used to convert probability scores into class labels, we can observe the model's performance across different operating points. The area under the curve (AUC) quantifies the overall ability of the classifier to distinguish between the two given classes (Géron, 2022). The more the curve shifts towards the top-left corner (with an AUC corresponding to 1), the better the

classification. A random classifier would correspond to a diagonal line with an AUC of 0.5."

11. **Section 5: Please provide some reference to previous methods and potential use cases when assessing performance- for example, what methods are currently used to identify whether a given molecule will be detectable by CIMS? Under what circumstances would this information be valuable?**

Currently, the only way to identify whether a given molecule will be detectable by CIMS is based on trial and error, by experimental means or by running quantum chemical computations. This information is needed during experiment design, and in our case, for building a library of CIMS signals. As mentioned above, this project aims to contribute to the compound identification of complex compounds which do not exist in current databases, by creating a model which can predict the CIMS signal based on the structure alone. Our model can predict the signal within an order of magnitude for a compound similar to our pesticide dataset.

In response to an earlier comment (#3), we have added two paragraphs in 5.3 before discussing MACCS results and as a final comparison:

"We observe that several of the identified important features relate to proton affinity. The number of HBA (NumHBA, LipinskiHBA) is directly correlated to proton affinity as it calculates the number of sites available to accept a proton. The TPSA describes the molecule's polarity, and for certain molecules, a higher TPSA could correlate with a higher proton affinity. HallKierAlpha correlates as well since every atom's covalent radius is adjusted for its hybridization state and electronegativity, reflecting the likelihood of atoms within a molecule to donate electron density to a proton. FractionCSP3, while not correlating directly to proton affinity, might influence the overall basicity of the molecule (e.g. a higher amount of sp^3 carbons in the molecule potentially affects the electron density of heteroatoms indirectly).
We note that different reagent ions react with the analyte in distinct ways (see the Introduction). While our results support the predictive nature of properties like proton affinity, specifically for positive reagent ions, the ML model's advantage lies in its flexibility. As shown, this approach aligns well with established knowledge, yet the ML methodology combined with molecular representations can relate any reagent ion or ionization mechanism to the magnitude of CIMS signals using only the analyte molecular structure.

[...]

Overall, RDKitPROP and MACCS in combination with RF have given us valuable insights into CIMS ion-molecule interactions. In the case of positive polarity ionization methods, the results obtained with the chemical insight analysis support known alternative methods of identifying whether a molecule can be detected, by highlighting a series of properties that can relate to proton affinity."

12. **3 chemical insight: Please clarify what chemical insight you are looking for here- I think chemical insight into why various molecules are or are not detectable by different ionization chemistries? This needs to be explicitly stated and explained and should be the primary focus of this section. Please also make at least some reference to the fundamental chemical mechanisms at play and how they do or do not agree with the features your model is identifying as important in predicting whether or not a given molecule will be detectable by CIMS.**

Thank you for this comment we agree that this needs to be explicitly stated and that we relate our findings to literature and knowledge of ionization mechanisms. The chemical insight we seek relates to understanding the interactions between reagent ions and analytes. Using our ML model, we can begin to identify which molecular features are influential in predicting whether a molecule will yield a detectable signal. Among the molecular descriptors used, RDKitPROP and MACCS are interpretable representations that allow us to draw connections between molecular structure and CIMS signal prediction.

We have updated the beginning of section 5.3 as follows:

"Next, we explore the chemical insight our ML classification models offer into ion-molecule interactions. As a proof of concept, we aim to relate the model's findings to established knowledge and note any unexpected influential features. We focus on the RF classifier model, as it enables straightforward identification of key molecular features associated with signal detectability. This is achieved by analyzing which features are most influential in the RF model's classification decisions. We will focus on the MACCS and RDKitPROP descriptors because they are the most interpretable. In the case of MACCS, the insight into the interaction can be extracted by analysing the occurrence of molecules detected and not detected for each feature, i.e. each MACCS key (sub-structure) of the molecular structure. In the case of RDKitPROP, the insight into the interaction can be formulated by analysing each feature, i.e. property. For each ionization method, we pick the largest training set size and then obtain the feature ranking and the corresponding coefficients (importance values) from the trained RF models for the five random seeds. We then average the importance values for each feature and rank again."

We also added a paragraph comparing our findings with existing knowledge on reagent ion-target compound interactions. We focused on whether the functional groups identified by our model align with known interaction patterns, highlighting both similarities and distinctions in computational approaches and associated costs. At the end of Section 5.3, we added the following paragraph:

"Overall, RDKitPROP and MACCS in combination with RF have given us valuable insights into CIMS ion-molecule interactions. In the case of positive polarity ionization methods, the results obtained with the chemical insight analysis support known alternative methods of identifying whether a molecule can be detected, by highlighting a series of proper-

ties that can relate to proton affinity. In the case of negative polarity ionization methods, a substantial comparison can be made with literature findings mainly based on detailed quantum chemical calculations. Based on RDKitPROP, the number of HBD in the molecule was attributed more than 10% of importance (by combining LipinskiHBD and NumHBD percentages); while based on MACCS, HBD groups such as OH and NH were found among the most important ones. These results agree with atmospheric chemistry studies such as Iyer et al. (2016) and Hyttinen et al. (2018), where quantum chemical calculations indicated that for organic vapours, OH is the primary functional group interacting with negative reagent ions. Similarly, Partovi et al. (2023), in a study of pesticide molecules, identified NH groups as significant in interactions with Br^- when OH groups were absent. Thus, our model supports these findings by identifying important features directly from data patterns without needing intensive quantum chemical methods.

While previous studies focused on single compound classes (e.g., homogenous sets of volatile organic compounds), or in a limited amount of complex compounds, our method utilizes a less homogeneous and larger dataset. The chemical insight analysis of our work provides a general profile of the interaction mechanism, supporting the findings from the literature but also highlighting other functionalities that might affect the signal due to their relation to the electronic structure.

This data-driven approach also required minimal computational resources due to the simplicity of the RDKitPROP and MACCS descriptors, contrasting with the higher demands of quantum chemical calculations. Although quantum chemical approaches remain essential for detailed, molecule-specific interactions, our ML model effectively reveals broader trends, distinguishing between detected and undetected molecules across the four studied ionization schemes."

13. **Section 6: This method works in a forward direction from a known compound to predict whether or not that compound will be detectable by CIMS, and if so how sensitive different CIMS mechanisms will be to said molecule. The conclusion states that this method will be useful in identifying atmospheric compounds, which inherently operate in a backwards direction from a detected CIMS exact mass and intensity to the identity, structure, and quantity of an unknown compound. Please explicitly state the suggested use case for this method, as it is not clear to me how the process could be reversed to assist in identification of atmospheric organics in a complicated ambient environment.**

We have clarified the usefulness of spectrum prediction for compound identification in response to comment #4. We will report the updated conclusions:

"In summary, we developed a ML workflow for predicting the detection with CIMS (with a classification algorithm) and CIMS sensitivity to molecules (with a regression algorithm) to improve atmospheric compound identification. The goal is to evaluate if our ML model can accurately predict detections and signal intensities, thus offering a foundation to build a database of simulated compounds' signals for compound identification purposes with CIMS. Currently, compound

identification is typically achieved by comparing an unknown compound's spectrum to a reference database. While this work does not provide direct identification of unknown compounds, it establishes a methodology for developing such a database, which could be expanded for broader use in atmospheric chemistry.

[...]

The results demonstrate that it is possible to extract predictive information even in small experimental datasets. However, more instances could help to generalize the structural features better and help prevent class imbalance problems. Currently, our ML models are directly applicable to predicting the detection and signal intensity of molecules with molecular structures similar to those in our dataset. For molecules with more diverse structures, transfer learning approaches could use these trained models as a baseline, updating learned parameters to accommodate the characteristics of new structures. Applying our approach directly to field measurements will require a comprehensive, standardized dataset of atmospheric compounds with a limited number of reagent ions for practical applications. Such a dataset could facilitate accurate mapping of ionization tendencies, potentially enabling compound identification directly from field CIMS measurements in the future."

14. **In the section describing the importance factors for the RF models, please provide some context on the fundamental underlying chemistry as to whether the machine learning model is identifying predictors as important that are known to be important drivers of CIMS ionization chemistry.**

Following comment #12, we have now provided the context at the end of Section 5.3. We will report again the changes:

"Overall, RDKitPROP and MACCS in combination with RF have given us valuable insights into CIMS ion-molecule interactions. In the case of positive polarity ionization methods, the results obtained with the chemical insight analysis support known alternative methods of identifying whether a molecule can be detected, by highlighting a series of properties that can relate to proton affinity. In the case of negative polarity ionization methods, a substantial comparison can be made with literature findings mainly based on detailed quantum chemical calculations. Based on RDKitPROP, the number of HBD in the molecule was attributed more than 10% of importance (by combining LipinskiHBD and NumHBD percentages); while based on MACCS, HBD groups such as OH and NH were found among the most important ones. These results agree with atmospheric chemistry studies such as Iyer et al. (2016) and Hyttinen et al. (2018), where quantum chemical calculations indicated that for organic vapours, OH is the primary functional group interacting with negative reagent ions. Similarly, Partovi et al. (2023), in a study of pesticide molecules, identified NH groups as significant in interactions with Br^- when OH groups were absent. Thus, our model supports these findings by identifying important features directly from data patterns without needing intensive quantum chemical methods.
While previous studies focused on single compound classes (e.g., homogenous sets of volatile organic compounds), or in a limited amount of complex compounds, our method utilizes a less homogeneous and larger dataset. The chemical

insight analysis of our work provides a general profile of the interaction mechanism, supporting the findings from the literature but also highlighting other functionalities that might affect the signal due to their relation to the electronic structure.

This data-driven approach also required minimal computational resources due to the simplicity of the RDKitPROP and MACCS descriptors, contrasting with the higher demands of quantum chemical calculations. Although quantum chemical approaches remain essential for detailed, molecule-specific interactions, our ML model effectively reveals broader trends, distinguishing between detected and undetected molecules across the four studied ionization schemes."

15. **Lines 34-42: previous work in characterization of atmospheric compound chemical properties via machine learning should be included here- (Sandstrom et al., 2024, Besel et al., 2024, others)**

Thank you for the suggestion, we modified the text accordingly:

"In this article, we explore if purely data-driven machine learning (ML) can facilitate CIMS compound identification. ML excels at pattern identification, data-driven classification and regression tasks. ML is proliferating in the natural sciences and has started to emerge in atmospheric science for, e.g., physicochemical property prediction and characterization of compounds (Lumiaro et al., 2021; Sandström et al., 2024; Besel et al., 2023, 2024; Hyttinen et al., 2022, 2024; Franklin et al., 2022), detection of new particle formation events (Su et al., 2022), boundary layer height estimation (Krishnamurthy et al., 2021), or aerosol classification (Siomos et al., 2020). In other chemical domains, e.g. metabolomics, ML has successfully enabled chemical compound identification from fragmentation mass spectrometry (Erban et al., 2019; Heinonen et al., 2012; Dührkop et al., 2015; Brouard et al., 2016; Nguyen et al., 2018, 2019)."

16. **Please address the treatment if isomers more completely. How many formulae corresponded to multiple isomers? How were the properties of the isomers amalgamated?**

A total of 38 molecular formulas in the dataset correspond to multiple isomers, yielding 81 isomers. Since excluding these would further reduce the dataset size, we retained them by assigning the same signal intensity to each isomer sharing the same formula. Each isomer is treated as a distinct data instance. However, this approach introduces a limitation, as the models cannot differentiate between these isomers due to identical signal intensities. Additional fragmentation mass spectrometry analyses are needed to distinguish isomers in a mixed standard solution. With larger datasets, removing indistinguishable isomers would be necessary and possible, for creating a model focused on compound identification at the structural level, which is our long-term goal.

We modified the text in Section 2 to read:

"Several isomers are present in the dataset (e.g., prometryn and terbutryn, or phoxim and quinalphos). Across 38 molecular formulas, there are 81 isomers in total. In CIMS, isomers produce peaks at the same mass-to-charge ratio and cannot be distinguished with a single ionization method, as they can in, e.g., fragmentation mass spectrometry. To retain dataset size, we included all isomers, assigning the same signal intensity to each if detected by an ionization method, and labeling all as undetected if no signal was present. This approach adds uncertainty to the ML model, can affect the evaluated model performance, depending on the structural difference of the isomers, and can limit the model validity for real-world applications requiring isomer distinction. This tradeoff allows for a larger dataset but reduces predictive accuracy at the structural level."

17. **Line 93: what characteristics did the undetectable pesticides share? How did these compare to the detectable species? What does this mean for the biases of this method when applied to atmospheric organics or pesticide mixtures? Can you please differentiate between the undetected and detected species as described in Figure 2? If panel c is illustrating which species are detectable by a single method rather than defining "undetectable" as meaning not detectable by any method, this should be more clearly explained and the color scheme should be different to reflect the different definition of "undetectable". Why does panel c appear to show an about equal number of undetected and detected species when the text states that the most species were detectable?**

Thank you for the comments. In this study, specifically in Figure 2, we provided a basic analysis of the molecular size and the elemental composition of both detectable and undetectable compounds. Panel (a), shows that the span of number of non-hydrogen atoms is relatively similar for the detected and undetected compounds. In Panel (b) tin was easily identified as present only in undetected molecules, but the other molecules do not have distinguishable characteristics, making it a complex dataset. No further analysis was done for the two groups, however, we analyzed the molecules detected and undetected individually for each ionization method in the results (Section 5.3), by analyzing the important features of the random forest model trained on MACCS. Table 4 presents, next to the importance value, the *proportion of presence* (reported as "PP%") and the *average group count per molecule* (reported as "Avg"). Similar results are reported for RDKitPROP predictor in the Supplement (Tables S16 and S17).

While pesticides are not representative of atmospheric compounds, both the descriptors discussed in Section 5.3 present either physical properties (such as the one related to the polarity of the molecule) or functional groups present also in atmospheric molecules (such as hydrogen bond donors groups) that might be useful for atmospheric organics.

We distinguish two cases for molecule detection in Figure 2. In the first case, a molecule is considered *detected* if at least one ionization method shows a non-zero CIMS sensitivity and *undetected* if none of the four methods yield a

non-zero signal. These are shown in Figures 2a and 2b, where the detected molecules outnumber the undetected ones. In the second case, individual ionization schemes are examined (Figure 2c): here, a molecule is labeled *detected* or *undetected* for each specific ionization method based on whether it produces a signal. To distinguish these cases clearly, we now refer to detected and undetected molecules in Panels a and b as "Detected by at least one ionization method" and "Undetected by all ionization methods." In Panel (c), detected is relabeled as "Detected by the specified ionization method" and undetected compounds as "Undetected by the specified ionization method".

The number of detected and undetected compounds differs in the panels of Figure 2. Panel (a) shows that there are 572 detected pesticides and 121 undetected ones (not detected by any ionization method. Panel (c) shows the detection counts by each ionization method. Contrary to Panel (a), where most molecules are detectable, Panel (c) shows that positive reagent ions detect more molecules than negative ions, with $AceH^+$ detecting the most and O_2^- the fewest. This highlights that negative ions are more selective for this dataset when detecting parent ions.

We modified the paragraph addressing Figure 2, the plot of Panel (c) and the figure's caption as follows:

"Figure 2 presents basic dataset statistics (molecular size, element composition and detection by ionization method). In Panel (a) and (b), we distinguish between *detected*, when a molecule presents a signal with at least one ionization method, and *undetected* otherwise. In Panel (c),*detected* refers to when a molecule presents a signal for a specified ionization method and *undetected* otherwise.

[...]

In Figure 2c the total count of detected and undetected pesticides for each ionization method is shown. Differing from Panel (a), where a large number of pesticides appear to be detected, Panel (c) reveals that in contrast to the positive reagent ions, the two negative reagent ions exhibit a higher number of undetected molecules than detected ones. Most pesticides are detected with $AceH^+$ and fewest with O_2^-. The figure highlights that, for this specific dataset, negative reagent ions are more selective than positive ones for the detection of parent ions."

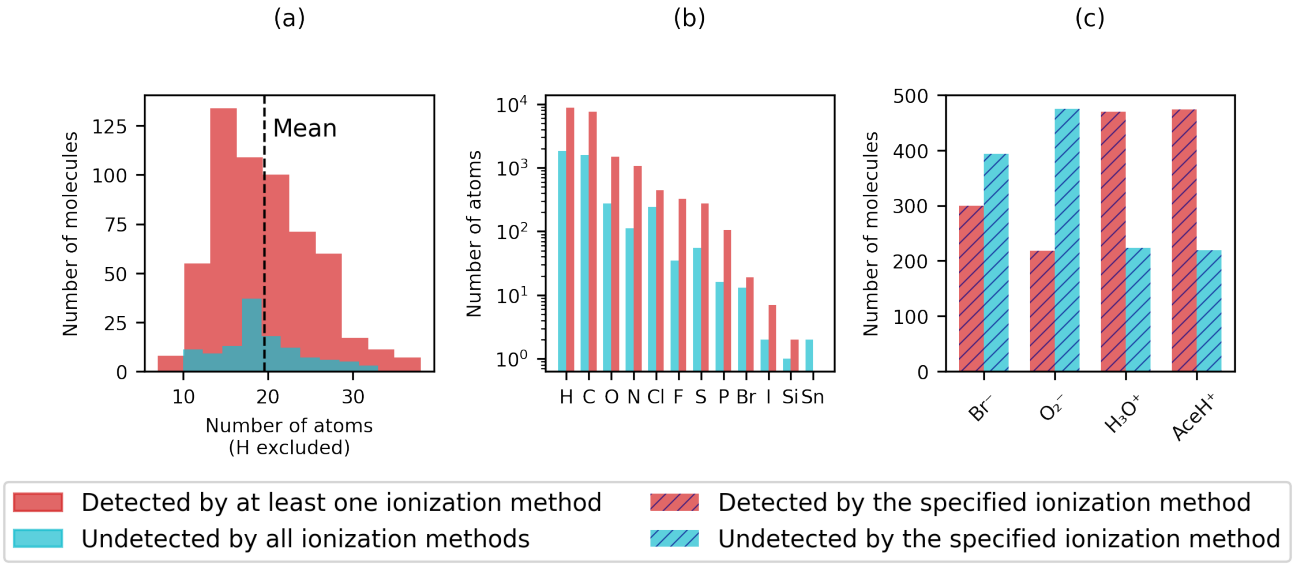

Figure 2. Distribution of (a) heavy atoms, (b) element types in logarithmic scale and (c) detection rate of the four reagent ions (Br^-, O_2^-, H_3O^+ and $AceH^+$). Detected pesticides are shown in red, undetected pesticides in light blue. In Panel (a) and (b), a molecule is considered detected if at least one ionization method presents a signal (full color). In Panel (c) the detection status is determined per ionization method individually (striped color).

18. **Please more clearly explain the purpose of figure 3 panel a- I am not clear on what this is illustrating**

Figure 3a illustrates the distribution of logarithmic signal intensities for each reagent ion, showing how signal values are spread across different ionization methods. This visualization is essential, as it provides insight into the data distribution that the model will learn from, particularly highlighting the intensity range and variance across ions. The caption has been revised to:

"

Figure 3. (a) Distribution of logarithmic signal intensities for molecules detected by each of the four ionization methods, and (b) scatter matrix of logarithmic signal intensities for molecules detected by all reagent ions, illustrating correlations between different ionization signals.

"

19. **Line 285- I think there are some typos in this sentence and I am not sure what it is meant to say**

We modified the sentence "We ascribe that the variance across the learning curves is moreless similar to the fact that our datasets are extremely small." to:

"We attribute the variability in learning curves to the small size of the training dataset. When subsets of data are selected using different random seeds, the limited number of observations leads to inconsistent sampling of data patterns, which prevents the model from stabilizing, particularly against outliers. As a result, the variance across learning curves as the training set size grows remains high due to the small sample size."

20. **Figure 7: can MAE be appropriately calculated and compared on log-normal data? Please justify**

Yes MAE can be appropriately calculated and compared for models predicting log-transformed data, as it reflects error on the same scale. Before training, we log-transformed non-zero CIMS intensities to create a normal-like distribution, minimizing the impact of outliers and stabilizing variance. In Figure 7, we present MAE based on these log-transformed values, indicating the model's typical error in terms of order of magnitude. We specified in section 4.4 Computational details, the mentioned reasons. The text was changed to:

"The regressors were trained on the logarithmic CIMS intensity of the detected pesticides. Before training, we log-transformed the non-zero CIMS intensities to create a normal-like distribution, reducing outliers' impact and stabilizing variance. MAE is then reported on this log scale, showing the model's error in terms of order-of-magnitude accuracy."

Response to Reviewer #2

Bortolussi and coauthors present machine learning methods that will be of great utility to users of chemical ionization mass spectrometry (CIMS) for atmospheric chemical detection. The methods outlined here are described well and the applications will allow atmospheric scientists to calculate the feasibility of a CIMS technique in detecting a potential analyte without the need for a detailed understanding of ion-molecule interactions or quantum mechanical calculations for estimating binding enthalpies, proton affinities, or electron affinities.

We thank you for the positive feedback.

1. **One major comment is in regards to general applicability for the community. As written, it is not clear how a field atmospheric chemist can use these outlined methods with their own instrument without first calibrating for a massive suite of compounds. There is only brief mention that these methods can be used "prior to deployment". Is the idea that eventually a large library can be generated using a growing training set (beyond pesticides) and then one can generally search if a reagent ion can detect an analyte with a considerable enough relative signal intensity? I also understand intensities vary across instruments which I address in the specific comments. I have other specific comments below, mainly regarding clarification on ion chemistry, fragmentation, superiority over previous methods and linking to prior research, and general applicability that should be addressed before publication. Again, I think the methods are well-outlined and this is a helpful guide for understanding reagent ion selectivity and CIMS detection sensitivity in a more applied way. It would be great to eventually see this extended to other popular CIMS reagent ions like I-, NH4+, and benzene cluster cations in other work.**

   We envision the applicability of our methods in both the short and long term. This work demonstrates the potential for predicting CIMS sensitivities, offers error estimates, and establishes the reference dataset size needed for accurate predictions. By providing a deeper understanding of ionization mechanisms, it establishes a framework that field atmospheric chemists can use for benchmarking with CIMS—an established tool in atmospheric studies.

   In the short term, our study provides immediate value by delivering chemical insights and a benchmark for reagent ion-pesticide sensitivity, setting a practical precedent for atmospheric chemists aiming to replicate this with other compounds. While our study focuses on pesticides, it demonstrates a data-driven approach that can be extended for identifying compounds using MION-CIMS in atmospheric research. For example, predictions of CIMS sensitivities for various compounds can, in principle, help determine parent ion mass and suggest structural features, including functional groups. With models like those developed here, sensitivities for thousands of atmospheric compounds could be predicted without the high computational costs of e.g., quantum chemistry calculations, paving way for simulated mass spectral datasets that field chemists could compare to their measurements for compound identification.

However, given that our study uses a pesticide dataset, applying this model to atmospheric compounds would first require standardized testing on a similarly sized dataset of atmospheric compounds. Our model's accuracy is currently validated only for pesticide-like compounds, though we show that an effective atmospheric dataset would require fewer than 1,000 compounds, collected collaboratively. With this dataset, our method could be applied through transfer learning to adapt sensitivity predictions for atmospheric compounds. In future work, we aim to develop models tailored to other popular CIMS reagent ions, such as I^-, NH_4^+, and benzene cluster cations, thereby broadening the method's scope and utility across the atmospheric chemistry community.

2. **Lines 40-41: There are quite a few CIMS datasets now, and each set generates a massive amount of data. The issue is more that they are not publicly available and there are not many detailed standard data sets. I would remove the word "data" and keep data standards. Further, you mention fragmentation but there is no method to address fragmentation patterns in this manuscript.**

Thank you for mentioning this. We have removed "data" from the sentence. Regarding fragmentation, no method is addressed because we misused the word "fragments" when instead we meant a part of the molecular structure. To clarify our intended meaning, we changed "fragments" to "sub-structure".

3. **Lines 45-46: You should elaborate on why you chose pesticides rather than more ubiquitous atmospheric compounds that represent more of the reactive gas abundance. Are pesticides diverse and represent a range of functional groups that can be detected with CIMS? Are they obscure and thus a good training set for difficult to identify compounds?**

Thank you for your comment. We received a similar comment from Reviewer #1 (comment #1). We will report the response and changes here as well.

Our study utilizes CIMS, a technique widely used in atmospheric compound research, and our ultimate aim is to apply the methods developed here to atmospheric compounds. However, we currently lack access to sufficiently large reference datasets for a vast majority of atmospheric compounds causing secondary aerosol formation to achieve this goal directly. Additionally, many of the direct aerosol precursor structures, such as highly oxygenated organic molecules correspond to polyperoxide compounds that are almost certainly high-explosives in the condensed phase. Therefore, they are also unlikely to become available in the foreseeable future.

Given this limitation, we identified a dataset of pesticides—an accessible and annotated reference set—as a practical starting point for testing. While pesticides represent only a minor subset of atmospheric compounds (as noted in our manuscript with references to Brüggemann et al. (2024) and Houde et al. (2019)), they encompass a wide range of

molecular sizes and functional groups, allowing us to explore the CIMS response across diverse ionization interactions. This structural diversity makes pesticides a suitable initial test case for developing and validating ML-based CIMS signal prediction methods. Additionally, pesticides are readily available as standard solutions from chemical suppliers at a manageable cost, which facilitates broader testing and replicability.

We note in our manuscript that pesticides only constitute a minor fraction of atmospheric compounds, and we highlight our choice of this dataset as a first step in developing ML-based tools for broader compound identification in atmospheric research.

The motivation for the use of pesticides instead of other atmospheric-related compounds has been updated in the Introduction section:

"In this work, we address the scarcity of atmospheric compound data standards by testing our methodology on a reference dataset of approximately 700 pesticides measured with CIMS. While pesticides represent only a small subset of atmospheric compounds (Brüggemann et al., 2024; Houde et al., 2019), they are chemically complex, with diverse molecular masses and functional groups that can interact in distinct ways with various reagent ions and that cover an extended range of detection with CIMS. This structural diversity provides a relevant test case that reaches and surpasses the complexity of many atmospheric compounds, allowing for an effective initial test of our methodology. Additionally, pesticides are readily available as standard chemicals from chemical suppliers at an accessible cost, and the dataset size is comparable to those used to establish early ML compound identification tools in metabolomics (Heinonen et al., 2012; Dührkop et al., 2015; Brouard et al., 2016; Nguyen et al., 2018, 2019). Thus, while limited to pesticides, this dataset offers a valuable preliminary benchmark for developing ML-based CIMS signal prediction. Once reference datasets for atmospheric compounds become available, this methodology can be directly applied or refined to encompass a broader range of atmospheric chemical analyses."

4. **Lines 74-76: Why are the routes of ionization specified for H3O+ and O2- but not Br- and AceH+? Also, can you list somewhere in the manuscript or supplement the experimental conditions for your ion-molecule reactions in the front end of your instrument, i.e. the temperature and pressure of your reaction chamber? Many species detected with O2- are generated as secondary compounds with signal that scales with front end pressure (and frequency of collisions) so this would be a helpful reference.**

Thank you for bringing this to our attention. In the case of Br^- and $AceH^+$, the neutral reagents are injected into the system and ionized via an X-ray source to generate the desired reagent ions. The experimental conditions align with those reported in Partovi et al. (2023). Our TD-MION inlet operates under atmospheric pressure conditions, and the sample is introduced on a custom-made thermal desorption filter developed by Karsa Oy (Karsa). The temperature range in the desorber varies from 30°C to 250°C, allowing for the gradual evaporation of pesticides based on their volatility. We have

revised the dataset's introduction to specify the experimental conditions and ion generation routes for each reagent ion as follows:

"The CIMS experiments were conducted at Karsa Oy laboratory (Karsa) with a TD-MION inlet operating at atmospheric pressure coupled to a linear trap quadrupole orbitrap mass spectrometer. A sample was placed on a custom-made filter (Karsa) and heated in the desorber from 30°C to 250°C; different pesticides evaporate from the filter at various temperatures. A schematic of the instrument and sampling methodology are presented in Partovi et al. (2023, 2024).

[...]

Each pesticide was measured with the following ionization schemes: bromide (Br^-) ionization (produced from dibromomethane, CH_2Br_2); protonated acetone ($(CH_3)_2COH^+$, $AceH^+$) ionization (produced from acetone, $(CH_3)_2CO$); proton-transfer (H^+) ionization by hydronium ions (H_3O^+, produced from trace amounts of water, H_2O^+); and electron transfer (-) ionization by dioxide ($O_2{}^-$). The first two ions were obtained by feeding the neutral reagents into the ion source, while the two latter ions were obtained by feeding dopant-free air instead. The pesticides were detected as protonated ions ($AceH^+$, H_3O^+), as deprotonated ions ($O_2{}^-$), or as adduct ions (Br^-)."

5. **Lines 85-86: The number of isomers should be listed. If there are many isomers then wouldn't keeping isomers in the dataset introduce considerable uncertainty for your performance metrics and chemical insights section?**

This is a good observation. We detailed the effects and tradeoff that led us to include the isomers in response to Reviewer 1's comment (#16):

A total of 38 molecular formulas in the dataset correspond to multiple isomers, yielding 81 isomers. Since excluding these would further reduce the dataset size, we retained them by assigning the same signal intensity to each isomer sharing the same formula. Each isomer is treated as a distinct data instance. However, this approach introduces a limitation, as the models cannot differentiate between these isomers due to identical signal intensities. Additional fragmentation mass spectrometry analyses are needed to distinguish isomers in a mixed standard solution. With larger datasets, removing indistinguishable isomers would be necessary and possible, for creating a model focused on compound identification at the structural level, which is our long-term goal.

We modified the text in Section 2 to read:

"Several isomers are present in the dataset (e.g., prometryn and terbutryn, or phoxim and quinalphos). Across 38 molecular formulas, there are 81 isomers in total. In CIMS, isomers produce peaks at the same mass-to-charge ratio and cannot be distinguished with a single ionization method, as they can in, e.g., fragmentation mass spectrometry. To retain dataset

size, we included all isomers, assigning the same signal intensity to each if detected by an ionization method, and labeling all as undetected if no signal was present. This approach adds uncertainty to the ML model, can affect the evaluated model performance, depending on the structural difference of the isomers, and can limit the model validity for real-world applications requiring isomer distinction. This tradeoff allows for a larger dataset but reduces predictive accuracy at the structural level."

6. **Lines 88-89: Was there any evidence that these species underwent fragmentation for the four ionization methods? A strong application of the work in this manuscript could be to identify fragments that may contribute signal artifacts to other quantified species at presumed parent masses. Using this method as is for predicting intensities at a parent ion mass and applying to the field can get complicated when there are contributing fragments. Your manuscript title asserting identification is not entirely true without considering potential fragments.**

Thank you for this insightful comment. Our analysis focuses on the parent ion's signal; we did not take into account the possible fragmentation of the target molecules but assumed that they mostly remained intact during our analyses. This assumption is supported by data collected using atmospheric chemical ionization with MION inlet, indicating no remarkable fragmentation under our operating conditions. However, we recognize that fragmentation inherently occurs in mass spectrometers, including orbitrap instruments. For instance, too loosely bound Br^- adducts fragment back into the Br^- reagent ion and the neutral target molecule, and will not appear in the final spectrum. Protonated and deprotonated species can also fragment in ways that yield charged fragments detectable in the mass spectrum. Currently, the orbitrap employed in this work applies mild trapping to suppress fragmentation, therefore, a low level of fragmentation is expected.

While our dataset consists solely of signals from parent ions, we agree that identifying potential fragments could provide additional spectral details that enhance compound identification through spectral comparison. Our current focus is to develop models for predicting parent ion signals across various reagent ions, using these values for spectral comparison instead. While this approach overlooks fragments, we believe their impact on our findings is likely minor, given that previous studies in mass spectrometry have successfully employed machine learning models that map fragmentation spectra to molecular properties based solely on sets of the strongest consistent signals (Franklin et al., 2022, AMT 15, 3779–3803). We assume that the strongest signals in our spectra correspond to the parent ions.

In field deployments, concentrating on parent ions is essential due to the complexity and unknown nature of the samples we encounter. Although certain molecules could be identified from characteristic fragments, the strength of atmospheric CIMS lies in its soft ionization, which minimizes fragmentation. We believe that primary ions remain detectable even in complex environmental matrices, while fragment ions may arise from various similar species, complicating their interpretation. For example, organosulfates exhibit characteristic fragments such as SO_3^- and HSO_4^- in mass spectra; however, HSO_4^- is also the ion used for quantifying atmospheric H_2SO_4.

We updated the main text in lines 72-73 as follows:

"The measurements from the two mixtures were combined into a single dataset for a total of 716 pesticide observations, where each observation corresponds to the parent ion's signal intensity. Due to CIMS' soft ionization, the parent ion is expected to have the highest intensity, quantitatively, and qualitatively provides a one-to-one correspondence to the target compound."

7. **Line 101-102: You should specify that the negative ionization methods are more selective than positive ones for detecting parent masses of this specific suite of molecules.**

Thank you for pointing this out. We modified the sentence accordingly:

"The figure highlights that, for this specific dataset, negative reagent ions are more selective than positive ones."

8. **Line 114-115: Can you please specify what you mean by "different parts of the target molecules"?**

Thank you for pointing this out, that sentence was not clear. Given the different correlation of intensities between the negative polarity ionization methods with the positive ones, we can conclude that there is indeed a different interaction between the reagent ion and the target molecule and the difference is pronounced when comparing the intensities of opposite polarities. This is a known observation, however, it can hint that different functional groups might be involved in the interaction leading to the detection and the intensity of the signal. We updated the sentence in lines 114-115 to:

"The general lack of correlation between opposite polarity ionization schemes indicates that different reagent ions interact with the target molecules in distinct ways, possibly engaging with different functional groups."

9. **Line 123-124: This seems consistent with intuition for CIMS reagent ion selectivity. You should tie in previous work using Br- and I- reagent ions for detecting these types of compounds to support this and explain why this method for selecting a reagent ion is superior to calculating proton affinities, electron affinities, or binding enthalpies. One may prefer to learn Gaussian and run a few relatively quick simulations rather than collect an extensive calibration suite and apply these methods. I understand the ML methods presented here can offer more detailed insight than compound to compound QM simulations but since this is a technical note presenting developments in compound detection with CIMS, a technical comparison is warranted. Is this method more**

straightforward to implement AND more accurate than previous methods? This could be tied into your performance metrics section since the accuracy looks pretty good.

The exploratory analysis discussed in lines 123-124 does not take into consideration any molecular structure, it is a data-driven clustering technique calculated directly from signal intensities. Moreover, as clustering technique t-SNE is broadly used only for visualization of the data, due to a meaningless attribution to the distance between each cluster (t-SNE visualizes high-dimensional data in lower dimensions preserving the local similarity of data points). We do not believe that t-SNE, as it is used in this work, is superior to calculating proton affinities, electron affinities or binding enthalpies, we just believe that since there is a clear division of the clusters, from a data science point of view there are higher chances to train successfully a model. Reviewer #1 had a similar comment on t-SNE (comment #8). We discussed that the t-SNE visualization allows us to examine how frequently signals from different reagent ions qualitatively co-occur, offering insight into selectivity and interaction diversity among the ions. In the main text, we initially stated: "The presence of clear clusters suggests that the CIMS signals carry information that might allow ML to distinguish different pesticides." which we have now corrected to read:

"The presence of clear clusters suggests that, collectively, the reagent ions have the potential to differentiate between molecular structures."

However, we agree with your comment: a technical comparison is warranted and our previous version of the manuscript did not mention any qualitative comparison between our results and the literature. A similar comment was made by the Reviewer #1. As mentioned, we added a full paragraph at the end of section 5.3 that does a comparison with other articles treating CIMS' interaction mechanism between reagent ion and target molecule. Due to the differences in methodologies, it is not fully possible to compare literature results from quantum chemical calculations (compound-specific precise results) with our machine learning models based on statistical connections and usually generalizing the features of a population. So we focused on the comparison of whether the functional groups that are found interacting with the reagent ion correspond to what we have found and on the comparison of computational costs. Here is a new paragraph at the end of section 5.3:

"Overall, RDKitPROP and MACCS in combination with RF have given us valuable insights into CIMS ion-molecule interactions. In the case of positive polarity ionization methods, the results obtained with the chemical insight analysis support known alternative methods of identifying whether a molecule can be detected, by highlighting a series of properties that can relate to proton affinity. In the case of negative polarity ionization methods, a substantial comparison can be made with literature findings mainly based on detailed quantum chemical calculations. Based on RDKitPROP, the number of HBD in the molecule was attributed more than 10% of importance (by combining LipinskiHBD and NumHBD

percentages); while based on MACCS, HBD groups such as OH and NH were found among the most important ones. These results agree with atmospheric chemistry studies such as Iyer et al. (2016) and Hyttinen et al. (2018), where quantum chemical calculations indicated that for organic vapours, OH is the primary functional group interacting with negative reagent ions. Similarly, Partovi et al. (2023), in a study of pesticide molecules, identified NH groups as significant in interactions with Br^- when OH groups were absent. Thus, our model supports these findings by identifying important features directly from data patterns without needing intensive quantum chemical methods.

While previous studies focused on single compound classes (e.g., homogenous sets of volatile organic compounds), or in a limited amount of complex compounds, our method utilizes a less homogeneous and larger dataset. The chemical insight analysis of our work provides a general profile of the interaction mechanism, supporting the findings from the literature but also highlighting other functionalities that might affect the signal due to their relation to the electronic structure.

This data-driven approach also required minimal computational resources due to the simplicity of the RDKitPROP and MACCS descriptors, contrasting with the higher demands of quantum chemical calculations. Although quantum chemical approaches remain essential for detailed, molecule-specific interactions, our ML model effectively reveals broader trends, distinguishing between detected and undetected molecules across the four studied ionization schemes."

10. **Line 293-295: CIMS signal intensity should only correlate to binding strength of an analyte and reagent ion for reagent ions that undergo adduct formation. In this case, that is only true for Br- and not applicable to the other reagent ions. This sentence should either be revised to apply to only Br-, offer another explanation for why all descriptors perform similarly, or remove this statement.**

Thank you for pointing this out. Indeed, only Br^- undergoes appreciable adduct formation, which makes our previous comment on binding strength incorrect for the other reagent ions. Regarding the comment on line 293 about the similar performance of models trained with molecular descriptors that provide different structural details, we would like to offer another explanation as to why all descriptors perform similarly. The comparable performance of different descriptors suggests that they all capture the relationship between CIMS signal and molecular structure at a similar level of detail. However, the fact that more information-rich descriptors, such as MBTR, do not outperform simpler ones may be attributed to the nature of the dataset. The noise and variability in the data could obscure the potential advantages of these more complex descriptors. Moreover, since we employed Kernel Ridge Regression as our model, its ability to learn intricate patterns might be limited by the inherent challenges present in the data.

The modified paragraph beginning at line 293 now reads:

"We believe this behaviour stems from the inherent characteristics of the dataset. The noise and variability in the data could obscure the potential advantages of these more complex descriptors. Moreover, since we employed KRR as our

regression model, its ability to learn intricate patterns might be limited by the inherent challenges present in the data."

11. **Line 344-345: I am a little confused by the repeated mention of binding strength for all reagent ions, since only signals from Br- ionization should be dependent on binding enthalpies (like the Iyer et al. 2016 paper for I- ionization cited in this manuscript). For "binding", are you more referring to an orientation that increases the likelihood of ion-molecule collisions, resulting in ionized analyte? In this example you mention van der Waals binding for positive ion species, but is there a role for the higher molecular weights to correlate with proton affinity due to an increased availability of electron density across a larger molecular structure for molecules with similar functional groups? Although this gets complicated when considering fragmentation for larger molecules. In general, I think you need to clarify throughout what is meant by "binding" vs. actual ionization mechanisms.**

We appreciate this comment regarding our use of the term "binding." To clarify our terminology, we have replaced "binding" with "tendency of ionization" throughout the manuscript. This change better reflects the concept of the reagent ion's ability to effectively ionize the target molecule.

In our revised text on lines 344-345, we highlight the significance of molecular weight and the number of atoms (NumAtoms) as key features in our model for distinguishing between detected and undetected molecules. We believe that this importance is due to the positive correlation between molecular size and the number of potential interaction sites with the reagent ion. A larger molecular size not only increases the available functional groups but also enhances the collision cross-section, thereby raising the likelihood of successful ion-molecule collisions for ionization to occur.

Accordingly, we have updated the text on pages 344-345 to read:

"Similar to TPSA, CrippenClogP emphasizes the role of hydrophilicity in our interaction analysis. The importance of molecular weight and NumAtoms indicates that larger molecular size correlates with detectability, as it provides more functional groups, and a greater collision cross-section, thereby possibly increasing the likelihood of interactions with the reagent ion."

12. **I am not familiar with orbitrap mass spectrometers as routine field instruments. Do you expect applicability to change across CIMS instruments, particularly for routinely used, but lower signal and resolution field deployable time-of-flight mass spectrometers? You do a good job of keeping intensities and absolute differences in log space as a way of normalizing signal, but would considerably lowering the resolution narrow your training data set and limit predictions? In other words, do you need the resolution of an orbitrap for your compounds? Some mention of general applicability to other instruments would be helpful.**

This is a very important point. MION inlets are already being used combined with Time-of-flight (TOF) detectors. Thus, similar datasets could be obtained with field-deployable TOFs. However, analyzing complex samples containing hundreds of compounds requires high mass resolution—typically on the order of 100,000 to accurately separate and identify compounds. We tested our pesticide standards on a Tofwerk TOF instrument and found that the resolution was insufficient to distinguish many signals, which limits its applicability for comprehensive analysis.

In recent years, advancements in orbitrap technology have made these instruments more accessible for complex sample analysis. For example, the modern Exploris series orbitraps offer reliability, lower cost and excellent mass calibration, making them suitable for studying complex samples. In the context of field measurements, orbitraps are now more robust for varied environments: they feature temperature control mechanisms that stabilize performance across temperature fluctuations, and the actual orbitrap unit is small and inside the unit. Recent studies, such as that by Keating et al. (Anal. Chem. 2024, 96, 21, 8234–8242), have demonstrated that orbitrap instruments can be deployed in field-like environments with stable signal and mass accuracy. Thus, we anticipate that orbitraps will increasingly be used in field studies as portable high-resolution alternatives to TOFs, particularly for complex environmental samples.

However, we want to highlight that, as a proof of concept, the methodology developed in this work can be applied to any lower-resolution data as well. We added in the Conclusions:

"Moreover, while this workflow was developed using high-resolution orbitrap data, it can also be utilized with lower-resolution data, though this may introduce greater uncertainties."

13. **In general, did you test your ML methods for compounds not identified in your training set? You generate predictors and accuracy metrics for already identified species that you train with. What happens if you predict the signal intensity for a given concentration of a compound not in your set? I think that would be the eventual, broader application of these methods. It would also assert the superiority of the presented methods over previous ones, in addition to showing how it could advance our understanding of ion-molecule interactions in CIMS.**

That is a good observation. Our objective is indeed to train ML models applicable to compounds not seen in the training set. To assess our models' capability to generalize to unseen compounds, we report their performance on an independent test set. This test data, representing approximately 20% of the dataset, is withheld during training and acts as an out-of-sample subset, enabling us to evaluate model generalizability beyond the training compounds.

While our current work focuses on pesticides, testing on atmospheric compounds would be a logical next step. This would require a standardized set of atmospheric compound measurements from MION MS measurements, which we currently do not have access to. However, for compounds structurally similar to those in the training set, the model can be used without further modification, providing predictions within the error range presented here. If new compounds have significantly different structural features, the model may require retraining or updating with additional molecular

classes, so-called transfer learning. Transfer learning could allow us to adapt a pre-trained model by updating it with data for additional compounds.

To clarify that the models were tested on compounds outside the training set, we modified the following part in the section "4.4 Computational details":

"We train a separate classification and regression model for each ionization method. The datasets were randomly split into test (20%) and training (80%) sets, ensuring that the trained model's performance is evaluated with an out-of-sample subset of data."

and added the following part in the Conclusions:

"Currently, our ML models are directly applicable to predicting the detection and signal intensity of molecules with molecular structures similar to those in our dataset. For molecules with more diverse structures, transfer learning approaches could use these trained models as a baseline, updating learned parameters to accommodate the characteristics of new structures."

14. **Line 25-26: The sentence about MIONs seems out of place in that paragraph. It should be merged more with CIMS benefits rather than come right after "but compound identification remains challenging".**

The sentence was revised as suggested.

15. **Line 171: You should specify that the Coulomb matrix is M.**

Thank you for noticing this. We revised the sentence as suggested:

"The Coulomb matrix (M, Fig. 5c) encodes [...]"

16. **Line 336-337: Please explain a little more what you mean by "... if the polarity is right...".**

Thank you for pointing this out. We have realized that the sentence contains a typo. We revised it for clarity, now it states:

"The high importance of TPSA highlights the significance of the molecular polar surface in the ionization mechanism. The polarity of the target molecule can increase the chances of interacting with the reagent ion, therefore increasing the resulting signal intensity."

17. **Line 284-285: This last sentence needs to be rewritten. I am assuming you mean that you are attributing the variance in the learning curves to the small size of the datasets?**

Thank you for the observation. We agree and have revised the sentence for clarity, addressing both your comment and Reviewer #1's feedback (Comment #20). The updated sentence now reads:

"We attribute the variability in learning curves to the small size of the training dataset. When subsets of data are selected using different random seeds, the limited number of observations leads to inconsistent sampling of data patterns, which prevents the model from stabilizing, particularly against outliers. As a result, the variance across learning curves as the training set size grows remains high due to the small sample size."

References

Besel, V., Todorović, M., Kurtén, T., Rinke, P., and Vehkamäki, H.: Atomic structures, conformers and thermodynamic properties of 32k atmospheric molecules, Scientific data, 10, 450, 2023.

Besel, V., Todorović, M., Kurtén, T., Vehkamäki, H., and Rinke, P.: The search for sparse data in molecular datasets: Application of active learning to identify extremely low volatile organic compounds, Journal of Aerosol Science, 179, 106 375, https://doi.org/10.1016/J.JAEROSCI.2024.106375, 2024.

Brouard, C., Shen, H., Dührkop, K., D'Alché-Buc, F., Böcker, S., and Rousu, J.: Fast metabolite identification with Input Output Kernel Regression, Bioinformatics, 32, i28–i36, https://doi.org/10.1093/BIOINFORMATICS/BTW246, 2016.

Brüggemann, M., Mayer, S., Brown, D., Terry, A., Rüdiger, J., and Hoffmann, T.: Measuring pesticides in the atmosphere: current status, emerging trends, and future perspectives, Environmental Sciences Europe, 36, https://doi.org/10.1186/s12302-024-00870-4, 2024.

Durant, J. L., Leland, B. A., Henry, D. R., and Nourse, J. G.: Reoptimization of MDL Keys for Use in Drug Discovery, Journal of Chemical Information and Computer Sciences, 42, 1273–1280, 2002.

Dührkop, K., Shen, H., Meusel, M., Rousu, J., and Böcker, S.: Searching molecular structure databases with tandem mass spectra using CSI:FingerID, Proceedings of the National Academy of Sciences of the United States of America, 112, 12 580–12 585, https://doi.org/10.1073/PNAS.1509788112/SUPPL_FILE/PNAS.201509788SI.PDF, 2015.

Erban, A., Fehrle, I., Martinez-Seidel, F., Brigante, F., Más, A. L., Baroni, V., Wunderlin, D., and Kopka, J.: Discovery of food identity markers by metabolomics and machine learning technology, Scientific Reports, 9, https://doi.org/10.1038/s41598-019-46113-y, 2019.

Franklin, E. B., Yee, L. D., Aumont, B., Weber, R. J., Grigas, P., and Goldstein, A. H.: Ch3MS-RF: A random forest model for chemical characterization and improved quantification of unidentified atmospheric organics detected by chromatography-mass spectrometry techniques, Atmospheric Measurement Techniques, 15, 3779–3803, https://doi.org/10.5194/AMT-15-3779-2022, 2022.

Géron, A.: Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, 1 edn., 2022.

Heinonen, M., Shen, H., Zamboni, N., and Rousu, J.: Metabolite identification and molecular fingerprint prediction through machine learning, Bioinformatics, 28, 2333–2341, https://doi.org/10.1093/BIOINFORMATICS/BTS437, 2012.

Himanen, L., Jäger, M. O., Morooka, E. V., Canova, F. F., Ranawat, Y. S., Gao, D. Z., Rinke, P., and Foster, A. S.: DScribe: Library of descriptors for machine learning in materials science, Computer Physics Communications, 247, 106 949, https://doi.org/10.1016/j.cpc.2019.106949, 2020.

Houde, M., Wang, X., Colson, T.-L. L., Gagnon, P., Ferguson, S. H., Ikonomou, M. G., Dubetz, C., Addison, R. F., and Muir, D. C. G.: Trends of persistent organic pollutants in ringed seals (Phoca hispida) from the Canadian Arctic, Science of The Total Environment, 665, 1135–1146, https://doi.org/10.1016/j.scitotenv.2019.02.138, 2019.

Huo, H. and Rupp, M.: Unified representation of molecules and crystals for machine learning, Machine Learning: Science and Technology, 3, https://doi.org/10.1088/2632-2153/aca005, 2022.

Hyttinen, N., Otkjær, R. V., Iyer, S., Kjaergaard, H. G., Rissanen, M. P., Wennberg, P. O., and Kurtén, T.: Computational Comparison of Different Reagent Ions in the Chemical Ionization of Oxidized Multifunctional Compounds, Journal of Physical Chemistry A, 122, 269–279, https://doi.org/10.1021/acs.jpca.7b10015, 2018.

Hyttinen, N., Pihlajamäki, A., and Häkkinen, H.: Machine Learning for Predicting Chemical Potentials of Multifunctional Organic Compounds in Atmospherically Relevant Solutions, Journal of Physical Chemistry Letters, 13, 9928–9933, https://doi.org/10.1021/ACS.JPCLETT.2C02612/ASSET/IMAGES/LARGE/JZ2C02612_0003.JPEG, 2022.

Hyttinen, N., Li, L., Hallquist, M., and Wu, C.: Machine Learning Model to Predict Saturation Vapor Pressures of Atmospheric Aerosol Constituents, ACS EST Air, 1, 1156–1163, https://doi.org/10.1021/ACSESTAIR.4C00113, 2024.

Iyer, S., Lopez-Hilfiker, F., Lee, B. H., Thornton, J. A., and Kurtén, T.: Modeling the Detection of Organic and Inorganic Compounds Using Iodide-Based Chemical Ionization, Journal of Physical Chemistry A, 120, https://doi.org/10.1021/acs.jpca.5b09837, 2016.

James, C., Weininger, D., and Delany, J.: Daylight Theory Manual. Daylight Chemical Information Systems, 1995.

Karsa: Karsa Oy, https://karsa.fi/, accessed: 2024-06-04.

Krishnamurthy, R., Newsom, R. K., Berg, L. K., Xiao, H., Ma, P.-L., and Turner, D. D.: On the estimation of boundary layer heights: a machine learning approach, Atmospheric Measurement Techniques, 14, 4403–4424, https://doi.org/10.5194/amt-14-4403-2021, 2021.

Landrum, G.: RDKit: Open-Source Cheminformatics Software, http://www.rdkit.org, accessed: 2024-06-04, 2006.

Langer, M. F., Goeßmann, A., and Rupp, M.: Representations of molecules and materials for interpolation of quantum-mechanical simulations via machine learning, npj Comput Mater, 8, https://doi.org/10.1038/s41524-022-00721-x, 2022.

Lee, B. H., Lopez-Hilfiker, F. D., Mohr, C., Kurtén, T., Worsnop, D. R., and Thornton, J. A.: An Iodide-Adduct High-Resolution Time-of-Flight Chemical-Ionization Mass Spectrometer: Application to Atmospheric Inorganic and Organic Compounds, Environmental Science Technology, 48, 6309–6317, https://doi.org/10.1021/es500362a, 2014.

Lumiaro, E., Todorović, M., Kurten, T., Vehkamäki, H., and Rinke, P.: Predicting gas–particle partitioning coefficients of atmospheric molecules with machine learning, Atmos. Chem. Phys., 21, https://doi.org/10.5194/acp-21-13227-2021, 2021.

Nguyen, D. H., Nguyen, C. H., and Mamitsuka, H.: SIMPLE: Sparse Interaction Model over Peaks of moLEcules for fast, interpretable metabolite identification from tandem mass spectra, Bioinformatics, 34, i323–i332, https://doi.org/10.1093/BIOINFORMATICS/BTY252, 2018.

Nguyen, D. H., Nguyen, C. H., and Mamitsuka, H.: Recent advances and prospects of computational methods for metabolite identification: a review with emphasis on machine learning approaches, Briefings in Bioinformatics, 20, 2028–2043, https://doi.org/10.1093/BIB/BBY066, 2019.

Partovi, F., Mikkilä, J., Iyer, S., Mikkilä, J., Kontro, J., Ojanperä, S., Juuti, P., Kangasluoma, J., Shcherbinin, A., and Rissanen, M.: Pesticide Residue Fast Screening Using Thermal Desorption Multi-Scheme Chemical Ionization Mass Spectrometry (TD-MION MS) with Selective Chemical Ionization, ACS Omega, 8, 25 749–25 757, 2023.

Partovi, F., Mikkilä, J., Iyer, S., Mikkilä, J., Kontro, J., Ojanperä, S., Shcherbinin, A., and Rissanen, M.: Multi-Scheme Chemical Ionization for Pesticide Detection: A MION-Orbitrap Mass Spectrometry Study, Manuscript in preparation, 2024.

Rissanen, M. P., Kurtén, T., Sipilä, M., Thornton, J. A., Kangasluoma, J., Sarnela, N., Junninen, H., Jørgensen, S., Schallhart, S., Kajos, M. K., Taipale, R., Springer, M., Mentel, T. F., Ruuskanen, T., Petäjä, T., Worsnop, D. R., Kjaergaard, H. G., and Ehn, M.: The Formation of Highly Oxidized Multifunctional Products in the Ozonolysis of Cyclohexene, Journal of the American Chemical Society, 136, 15 596–15 606, https://doi.org/10.1021/ja507146s, 2014.

Rupp, M.: Machine Learning for Quantum Mechanics in a Nutshell, International Journal of Quantum Chemistry, 115, 1058–1073, https://doi.org/10.1002/qua.24954, 2015.

Rupp, M., Tkatchenko, A., Müller, K.-R., and von Lilienfeld, O. A.: Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning, Physical Review Letters, 108, 1–5, 2012.

Sandström, H., Rissanen, M., Rousu, J., and Rinke, P.: Data-Driven Compound Identification in Atmospheric Mass Spectrometry, Advanced Science, 11, 2024.

Siomos, N., Fountoulakis, I., Natsis, A., Drosoglou, T., and Bais, A.: Automated Aerosol Classification from Spectral UV Measurements Using Machine Learning Clustering, Remote Sensing, 12, 965, https://doi.org/10.3390/rs12060965, 2020.

Su, P., Joutsensaari, J., Dada, L., Zaidan, M. A., Nieminen, T., Li, X., Wu, Y., Decesari, S., Tarkoma, S., Petäjä, T., Kulmala, M., and Pellikka, P.: New particle formation event detection with Mask R-CNN, Atmospheric Chemistry and Physics, 22, 1293–1309, https://doi.org/10.5194/acp-22-1293-2022, 2022.

Thoma, M., Bachmeier, F., Gottwald, F. L., Simon, M., and Vogel, A. L.: Mass spectrometry-based *Aerosolomics*: a new approach to resolve sources, composition, and partitioning of secondary organic aerosol, Atmospheric Measurement Techniques, 15, 7137–7154, https://doi.org/10.5194/amt-15-7137-2022, 2022.

Xue, L. and J, J. B.: Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening, Comb Chem High Throughput Screen, 8, https://doi.org/10.2174/1386207003331454, 2020.