Exploring ionizability and adduct formation of pharmaceutical compounds in MALDI and MALDI-2 Mass Spectrometry Imaging with machine learning

Krischan Koerfer[1], Andrew Palmer[2], Klaus Dreisewerd[1], Jens Soltwisch[1], Peter Marshall[2]

[1]Institute of Hygiene, University of Münster, Münster, Germany.

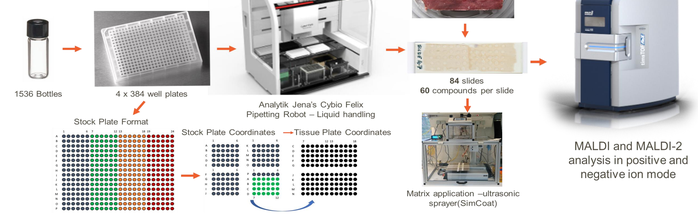
[2]Research & Development, GSK.

Introduction

In this work, we leverage a MALDI-MS imaging screen of 1,200 drug-like compounds under “simulated real life imaging” conditions to determine to what extent the ability of a compound to be ionised and which adducts are likely to be observed can be predicted based on a 2D chemical structure and its physicochemical properties.
Our goal in this study is to predict which ion types would be expected for a given chemical compound based on its 2D chemical structure and explore whether we can link ionisation behaviour to specific parts of the chemical structure.

Methods

Sample Preparation

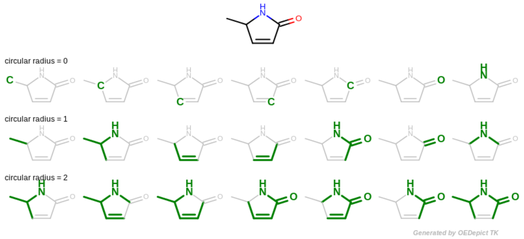


Data Processing

The presence of a compound-ion was determined in MALDI and MALDI-2 for common ion types, ([M•]⁻, [M-H]⁻, [M•]⁺, [M+H]⁺) using isotope pattern matching, yielding 8 prediction endpoints.

Machine Learning

Two-dimensional chemical structures were standardized by de-salting and charge neutralization and were digitally vectorized using Morgan fingerprints. These fingerprints record the presence or absence of specific substructures (e.g. functional groups), in a fixed-length binary vector.



Example of enumerating circular fragments with various radii: *https://docs.eyesopen.com/toolkits/cpp/graphsimtk/fingerprint.html*

In this study, fingerprints were generated with a 3-atom radius and were represented as 128-bit hashes. Compounds were randomly assigned to train/test/validate groups and used to train the (hyper-) parameters of a machine learning model (Ensemble of Classifier Chains (ECC) based on Support Vector Machines (SVMs)).

Chemical Interpretability

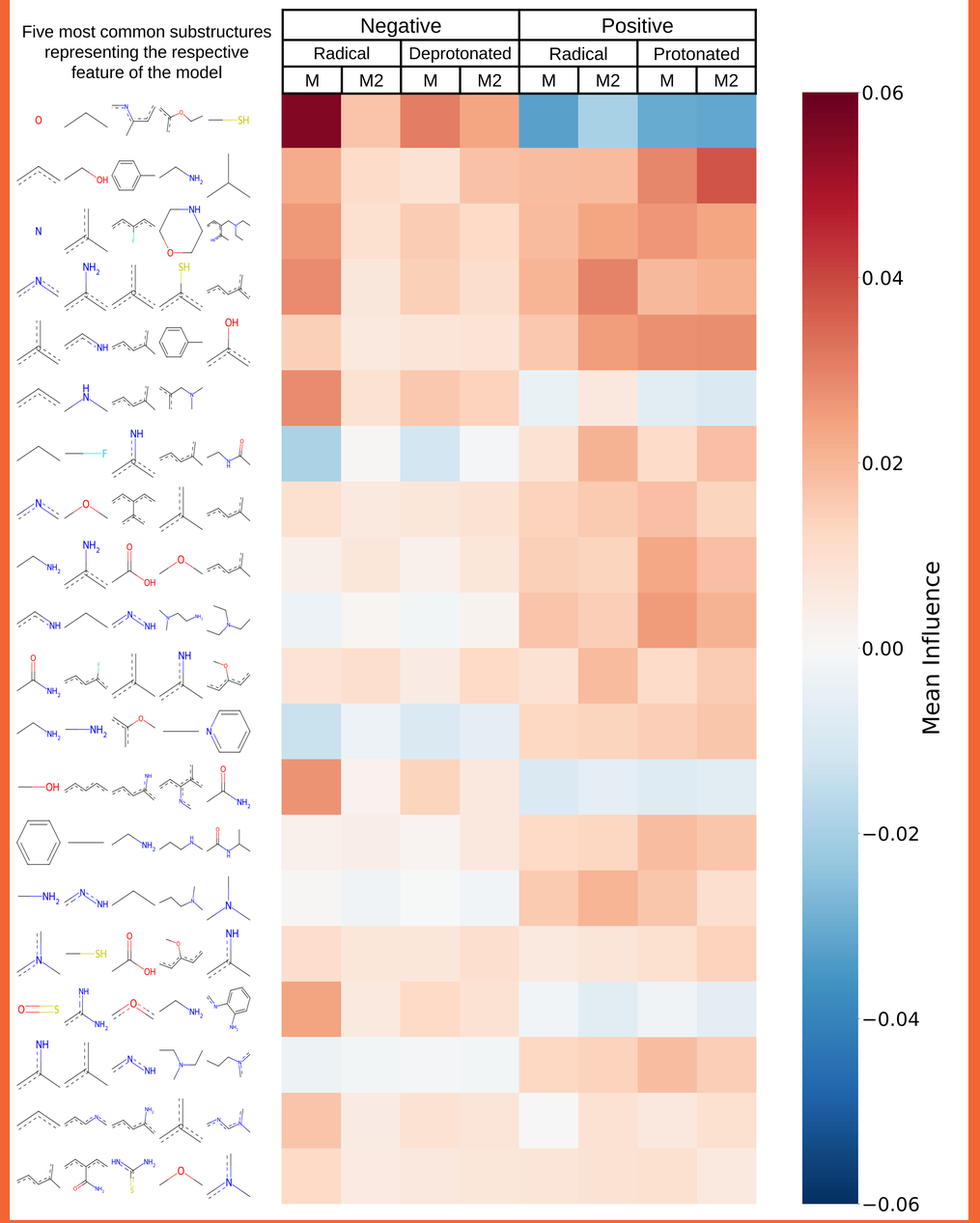
Restricting ourselves to linear models provides increased interpretability through inspection of the structural motifs most important for predicting ionisation behaviour. We used the Shapely Additive Parameters framework to understand which structural motifs are most influential on the model predations and compare these between adducts and after post-ionisation to generate clues to the mechanistic behaviour of MALDI & MALDI-2 ionisation.

Ethics statement

All animal studies were ethically reviewed and carried out in accordance with Animals (Scientific procedures) Act 1986 and the GSK Policy on the Care, Welfare and Treatment of Laboratory Animals.

Acknowledgements

We thank Bruker and Analytik Jena for their kind support.

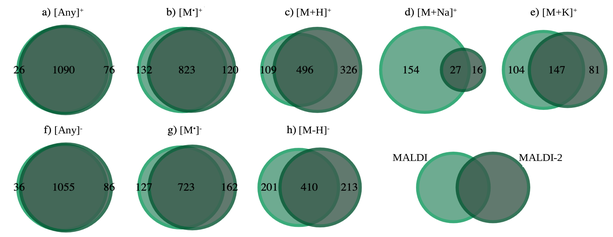


Molecular sub-structures were found to promote/inhibit the detection of ions

Influence measures the effect that the presence of a substructure has on the probability of detection.

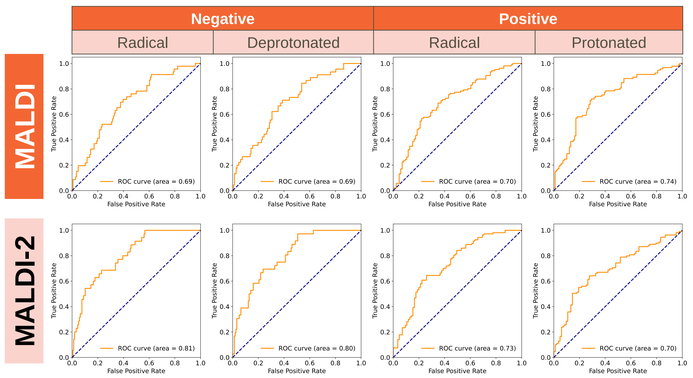
Results

Measured Detectability



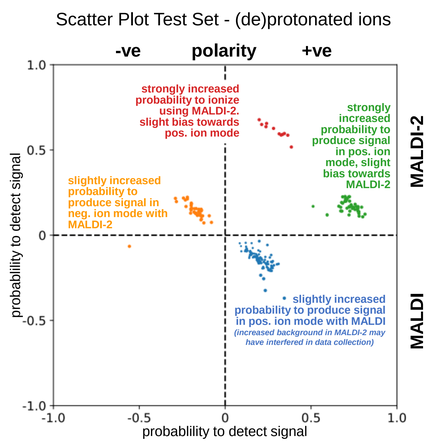
Overlap in annotated compounds between MALDI and MALDI-2 for the ion types considered if only isotope pattern matching but no S/N threshold is considered. “Any" indicates the overlap in annotation for compounds detected as at least one of the ion types considered in positive ion mode

Model Accuracy



Eight models were trained and evaluated. An average ROC-AUC of >0.7 was achieved for the ion types considered, providing a chemo-informatics model with satisfactory predictive performance. This indicates the model has effectively captured underlying relationships between the 2D chemical structures and ionisation.

Prediction of most suitable mass spectrometry conditions



Molecules plotted by predicted ionisation behaviour show distinct behavioural groupings which can be used to inform the experimentalist which mass spectrometry conditions to use. Each point is a chemical structure grouped across polarities by (de-) protonated ions. Distance from the origin show the specificity of molecules to ionisation polarity and use of post-ionisation respectively. Molecules around zero are equally likely to be detected under any conditions. Marker size indicates the overall likelihood of being detected. Clusters were assigned based on a k-means clustering.

Conclusions

Using Machine Learning we have been able to identify chemical substructures which if present in a molecule will promote or hinder its detection. Additionally, the prediction ionisation behaviour can be used to determine the most suitable mass spectrometry conditions to use and if post-ionisation is advisable.