Predicting AMS Spectra using Cheminformatics and Machine Learning

James Allan & David Topping

University of Manchester & National Centre for Atmospheric Science

Or:

Reports of the Horse’s Death Have Been Greatly Exaggerated

James Allan & David Topping

University of Manchester & National Centre for Atmospheric Science

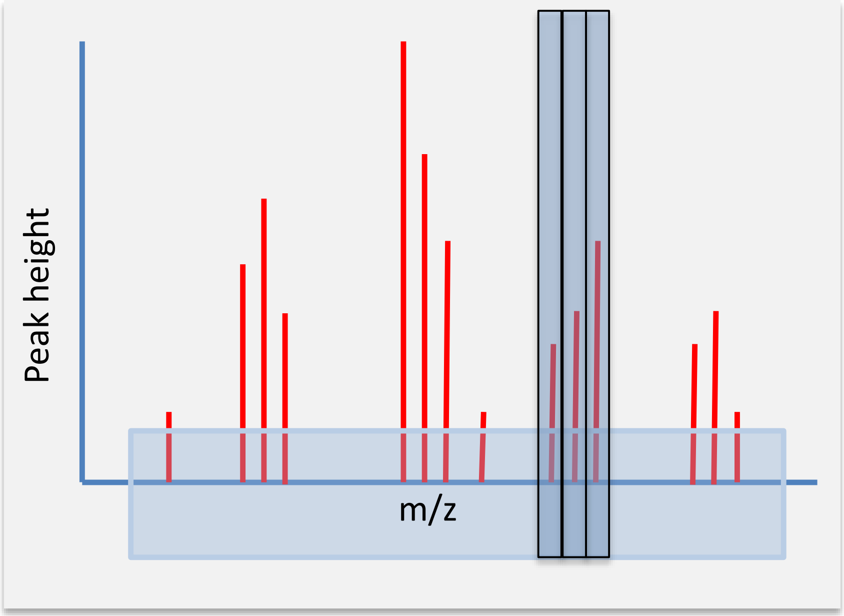
Predicting AMS Mass Spectra

- We have, by now, a large library of mass spectra for laboratory standards
- Behaviours in mass spectral peaks (m/z=44, 43, 57, etc.) have been quantitatively attributed to chemical functionalities (e.g. aliphatic chains, acids, carbonyls, etc.)
- Can we use this information such that a complete mass spectrum can be predicted based on any functionality?
- Can we arbitrarily predict what the mass spectrum of any molecule should look like?

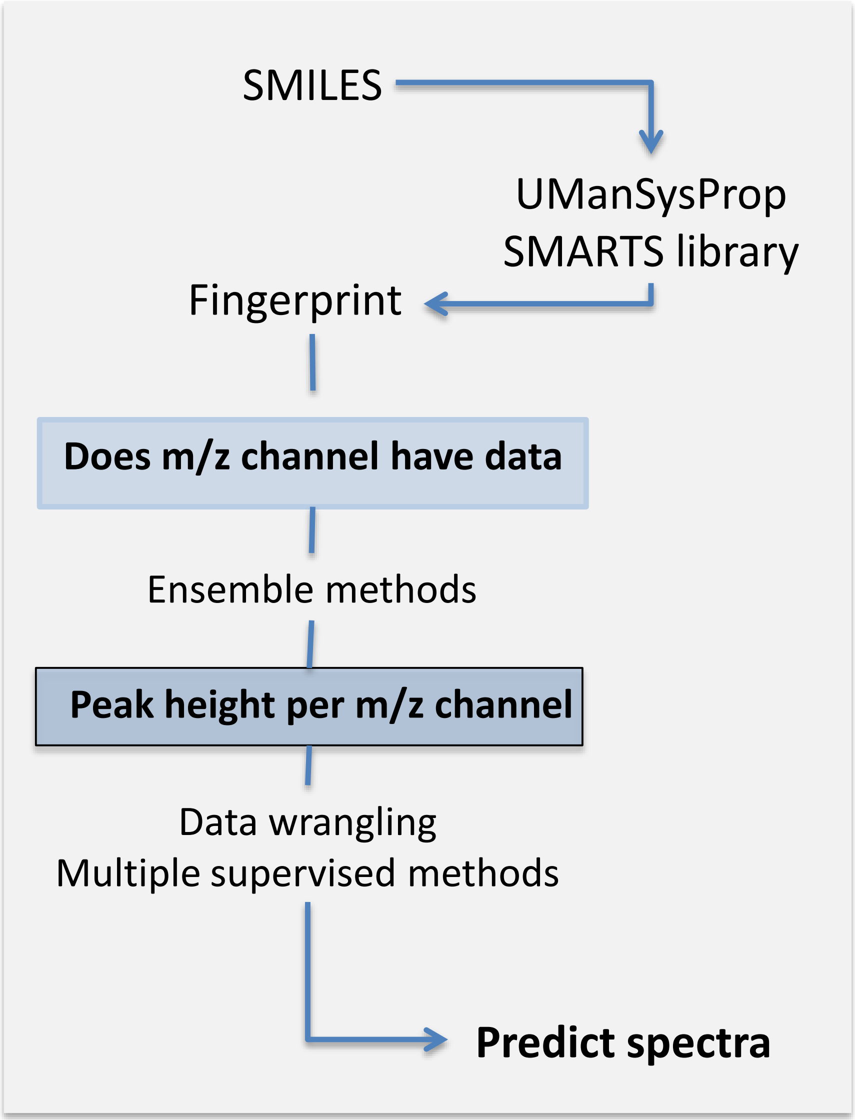
Cheminformatic Jargon

- Simplified Molecular-Input Line-entry System (SMILES): Method of representing molecular structures using ASCII strings
- Features: A property of a molecule based on functional groups and structure
  - e.g. “Alkyl group 3 carbons down from an alcohol group”, “group attached to a ring that has potential to change tautomeric form”, etc.
- SMiles ARbitrary Target Specification (SMARTS): A method of querying SMILES for features
- Fingerprints: A summary of the important features within a molecule
- These form the basis of the cheminformatic tools used in UManSysProp

Training data



Model development



SMILES

UManSysProp SMARTS library

Fingerprint

Does m/z channel have data

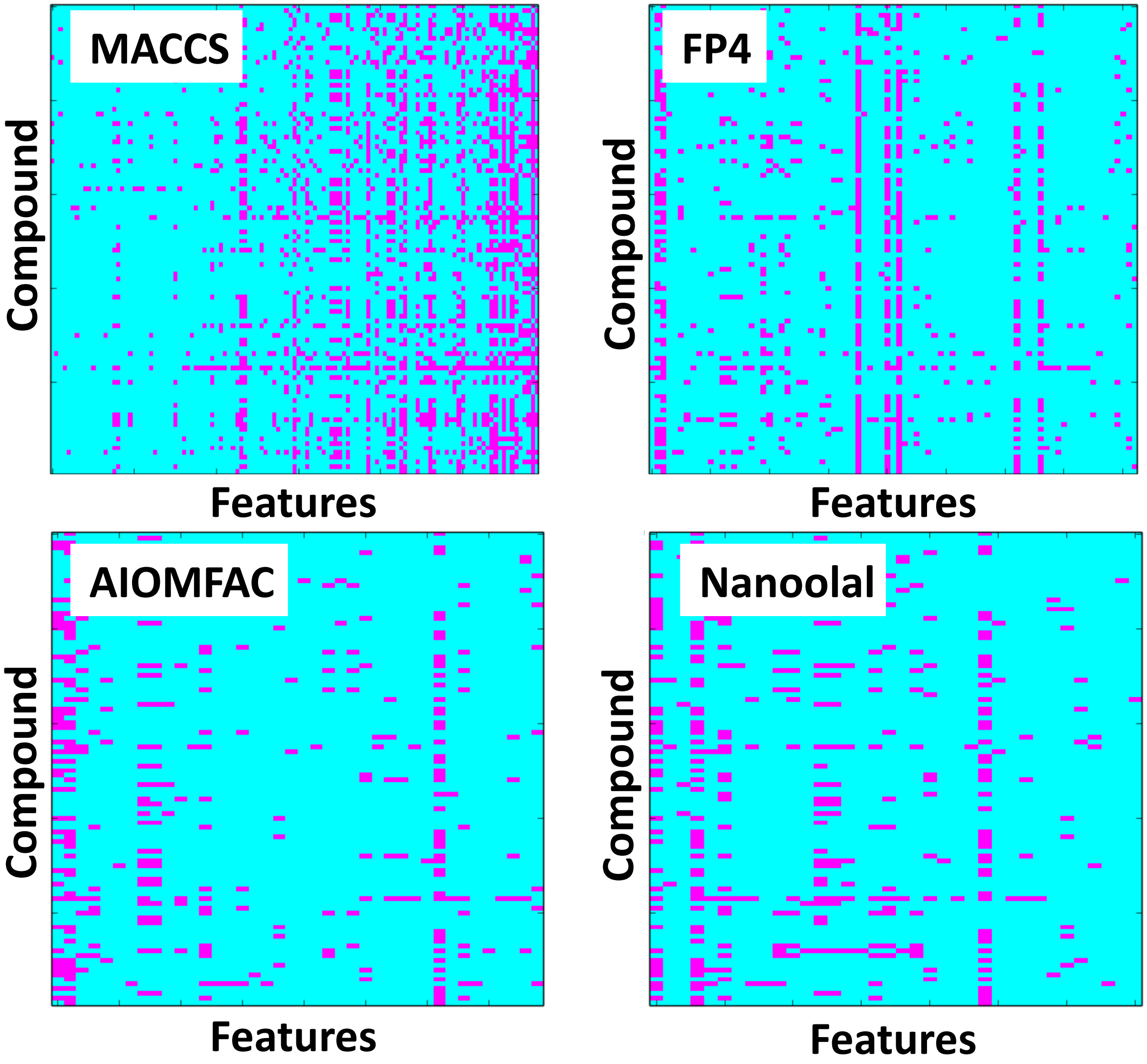
Ensemble methods

Peak height per m/z channel

Data wrangling
Multiple supervised methods

Predict spectra

Fingerprinting



- Different fingerprinting methods were tested:
  - MACCS and FP4 were developed for generic applications
  - AIOMFAC and Nanoolal were developed specifically for activity and vapour pressure estimation
- Each magenta box represents a feature identified for a given compound according to a different SMARTS library
- Max number of unique features that could be extracted:
  - MACCS – 162
  - FP4 – 320
  - AIOMFAC – 82
  - Nanoolal – 76

Learning algorithms

When simply evaluating predicted spectra against spectral library, choice of fingerprint affects performance. However, choice of supervised method more important if we only use these values

	Key:			
Method	MACCS keys	FP4	AIOM	Nan
SVM RBF	0.71	0.67	0.66	0.68
SVM Poly	0.60	0.63	0.62	0.62
SVM Lin	0.56	0.65	0.68	0.66
BRR	**0.91**	**0.87**	**0.87**	**0.85**
OLS	**1.00**	**0.95**	**0.92**	**0.91**
SGDR	**0.80**	0.72	0.71	0.69
Tree	**1.00**	**0.98**	**0.98**	**0.98**
Forest	**1.00**	**1.00**	**1.00**	**1.00**

	MACCS keys			
Method	Full	Var Select	Subset	Var Select / Subset
SVM RBF	0.71	0.69	0.71	0.71
SVM Poly	0.60	0.66	0.62	0.66
SVM Lin	0.56	0.65	0.71	0.69
BRR	**0.91**	**0.87**	**0.89**	**0.88**
OLS	**1.00**	**0.94**	**0.97**	**0.93**
SGDR	**0.80**	0.79	**0.80**	0.77
Tree	**1.00**	**0.98**	**0.98**	**0.97**
Forest	**1.00**	**0.99**	**1.00**	**0.95**

Cosine angle statistics

Bold values all above 0.8

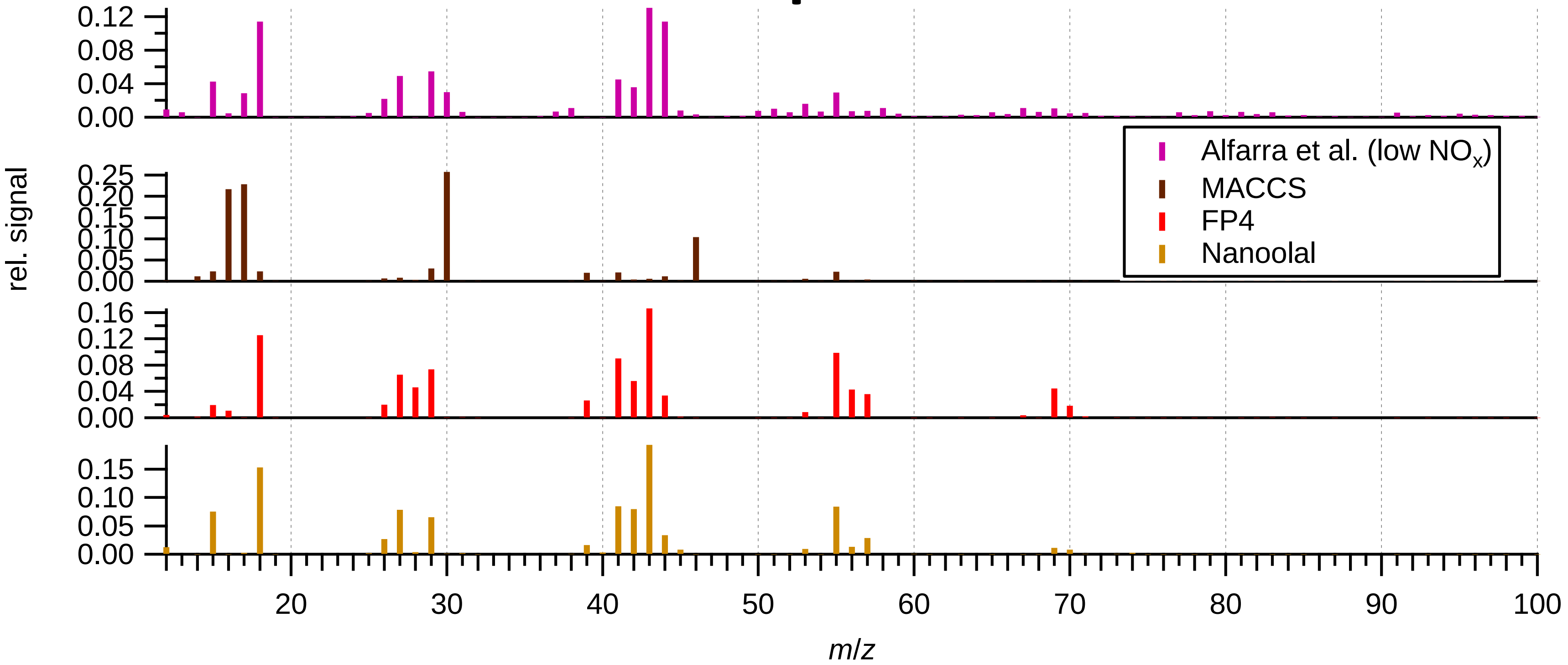
Training to a subset reveals more interesting dependencies, the same supervised methods still dominating performance.

‘True’ model performance

Test run on modelled data

- The AMS mass spectrum simulator was run on the model outputs of an explicit GECKO-A simulation of α-pinene oxidation
  - Valorso et al., doi: 10.5194/acp-11-6895-2011
  - This simulation produced a plausible mass concentration of SOA, albeit sensitive to the partitioning model
  - GECKO-A was used instead of the MCM because it uses predicted rather than prescribed reactions and can thus generate data on exotic molecules likely to be present in SOA
    - This feature is coming in MCM v4
- Data on ~55,000 particle-phase molecules were generated
- Predictions of AMS data were generated from a mass-weighted average of predictions and compared with previously published smog chamber spectra
  - Chhabra et al., doi: 10.5194/acp-11-8827-2011
  - Alfarra et al., doi:10.5194/acp-13-11769-2013

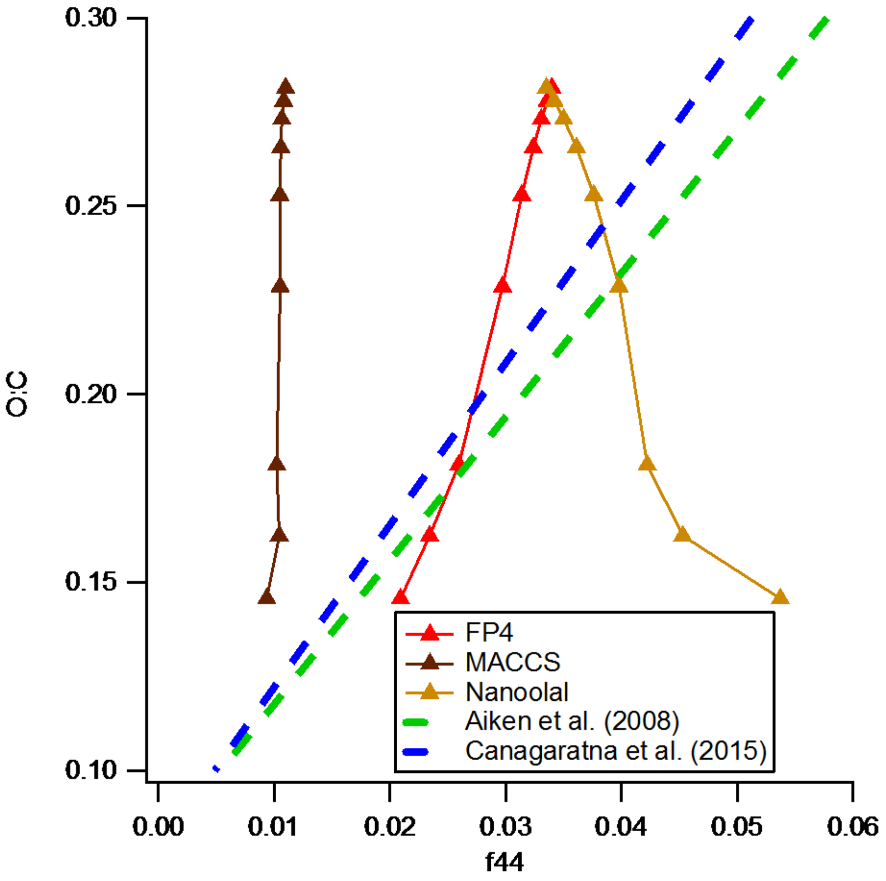
Mass Spectra



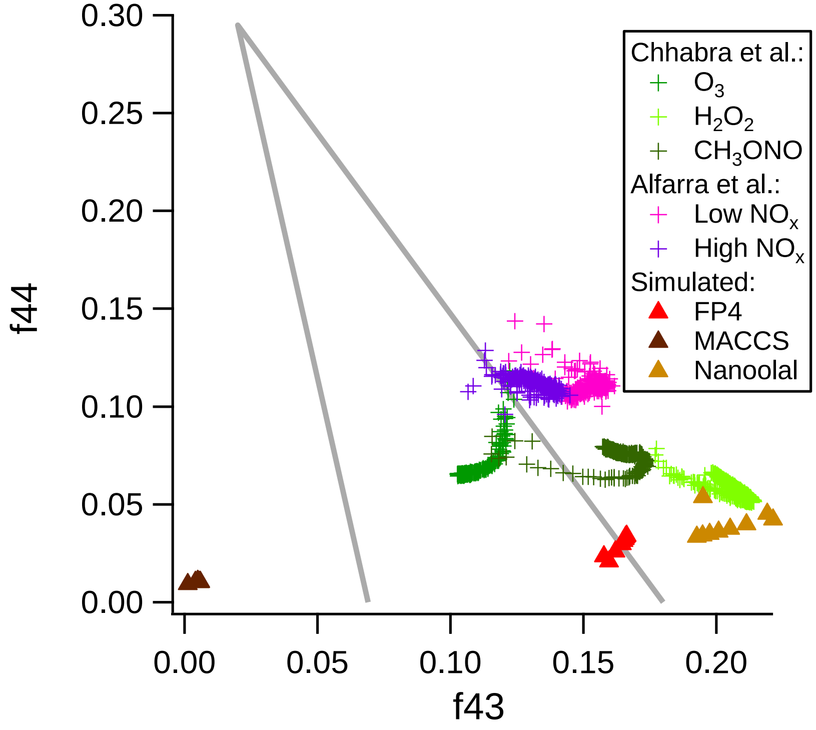
- Major peaks (41, 43, 55) predicted well by FP4 and Nanoolal – some differences in minor peaks
- MACCS completely off and looks more like ammonium nitrate – possibly over-trained?

O:C ratio vs f44

- GECKO-A predicts a monotonic increase in O:C over time
  - Values are low compared to typical atmospheric LV-OOA
- FP4 and Nanoolal give absolute f44s that compare well with published calibrations relative to O:C
  - The trend in f44 is reversed for Nanoolal, although the values are within the spread of calibration values used in the papers, so could still be plausible



f44 vs f43



- f43 values for FP4 and Nanoolal plausible compared to published studies
- f44 systematically low for all fingerprints, however this may be due to a lack of mechanisms such as autooxidation in the model
  - This is included in a newer version of GECKO-A (McVay et al. doi:10.5194/acp-16-2785-2016)
- Note the trajectories are complex and not monotonic for either the experimental or simulated data

Possible applications

- Enhance measurement-model comparisons beyond simple metrics such as mass concentration and O:C
- Assist with the development of explicit models of chemistry and partitioning
  - These can in turn inform parametric models such as VBS
- Allow predictions to be made when testing hypotheses, facilitating experiment design
- Testing the plausibility of proposed mechanisms and molecules when explaining observations
  - Note: Not a substitute for actual experimental evidence!

Further Work

- Publication of methodology (probably in GMD, which entails release of code)
- More training data (i.e. more analysis of standards)
- More testing of fingerprinting and training methods
- Application to HR data
- Looking at other modelled systems
  - Change precursors (e.g. anthropogenic)
  - Add/remove mechanisms, as per McVay et al. (2016)
  - Try with different models (e.g. MCM, different partitioning schemes)
- Comparing Lagrangian models with field data
- Inclusion into UManSysProp
  - http://umansysprop.seaes.manchester.ac.uk/

Questions

- James Allan

james.allan@manchester.ac.uk

- David Topping

david.topping@manchester.ac.uk

- James Brooks

james.brooks-2@manchester.ac.uk