The Winning Solution for the Mars Spectrometry: Detect Evidence for Past Habitability Challenge

Dmytro Poplavskiy

[redacted]

Abstract

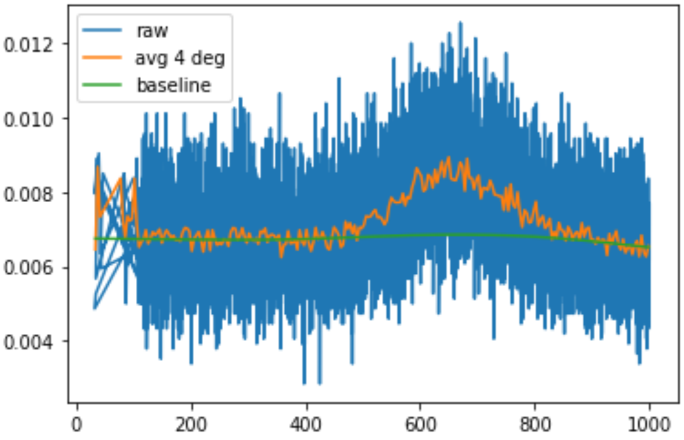
The solution is based on the ensemble of diverse CNN, RNN and transformer-based deep models over 2d representations (m/z over temperature) of the mass spectrometry data. Heavy data augmentation is applied during training and the simple ensemble used, averaging the results of the different architecture models, trained on the differently preprocessed data.

Overall the solution does not rely as much on pre and postprocessing of the SAM testbed as on augmentation of the commercial data.

Data Preprocessing

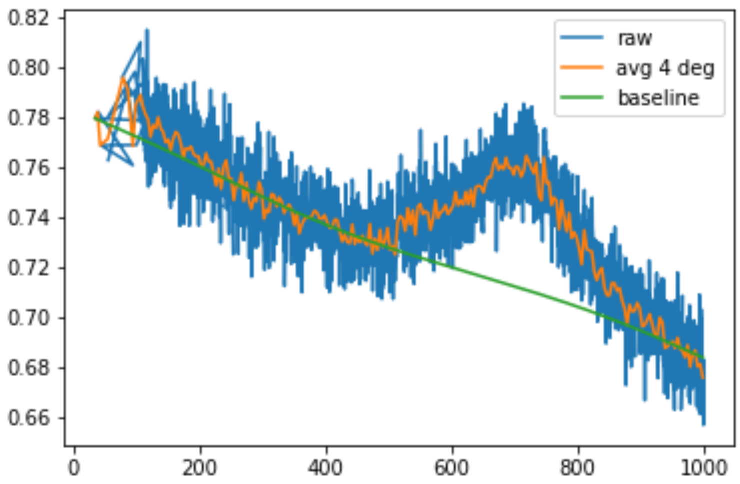
The mass spectrometry data used for training and prediction is restricted to the most common subset: temperature range of 28-1000 degrees and m/z range of 0-100.

SAM testbed preprocessing



The next step is applied only to the SAM testbed data: the data looks much noisier, with a larger number of samples, so the SAM testbed samples are averaged over the 4 degrees bins.

All the next steps are common for both commercial and SAM testbed samples: background subtraction, data augmentation, prediction on 2d data representation and the simple ensembling as an average of models predictions.



Data augmentation

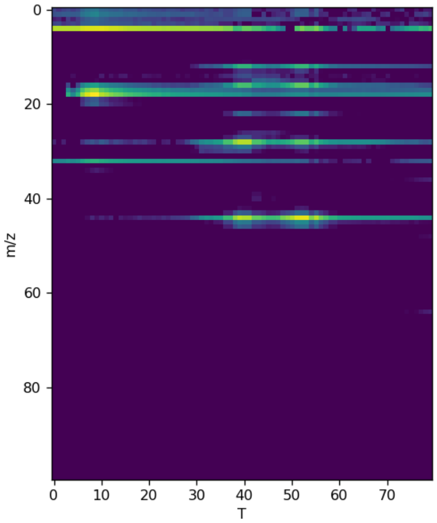
To avoid overfitting to the training dataset and make the models more generalisable to the SAM testbed samples, the following augmentations have been applied during training and prediction:

- Randomly choose to correct for background, either subtract the estimated baseline or simply a minimum value. In both cases, the random combination of 0.1%, 5%, 10% and 20% percentiles is subtracted.
- Random query of the abundance values from the overlapping temperature bands when converting to 2D image.
- Every m/z plot is randomly scaled by $2^{\text{random\_normal(0 mean, sigma)}}$ with 1.0 or 0.5 sigma values.
- For some models, instead of normalizing data by maximum value over all measurements, m/z plots are normalized to 1 individually. Surprisingly, the models' performance was similar to the maximum scaling. So some models in the ensemble were trained with the max normalization and some with per m/z band normalization.
- Mixing two samples. 2D representations of two samples are mixed together, either as an average or maximum value with the label combined as maximum. Mixing a few available SAM testbed samples with commercial samples may help to generalize to other SAM testbed samples.

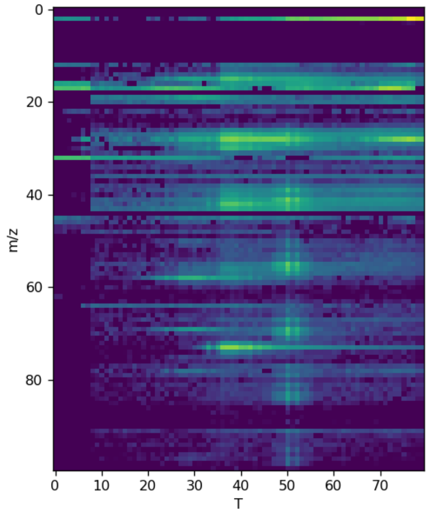
2D representations

Unlike the baseline solution, with abundance values averaged over 100℃ bands, I decided to use the smaller 12℃ step with only one abundance value randomly sampled from the overlapping 32℃ bands. This allows to better describe the peak shape while the weighted random sampling acts like an extra augmentation. Data is normalized and converted to log space.

Example of 2D mass spectrogram from the commercial sample:



2D spectrogram representation from the SAM testbed sample:



Models

All models used the 80x100 2D mass spectrogram as input and produced 10 logits for individual labels. The performance for all listed models has been comparable. Models include the CNN or visual transformer-based models over 2d data and RNN and 1D based transformers with the temperature used as a sequence dimension. Simple models worked well, while ensembling allowed to significantly improve the score.

Suggested improvements

With only 12 SAM testbed samples available for training and validation, and given such a drastic difference between the commercial and the SAM testbed data, I don’t really expect the model to perform very well. Collecting more training samples is the most obvious and straightforward way to improve the model performance, but since the number of the SAM testbed samples is limited, just adding them to the training set is not the most efficient way. It may work better to use the extra samples and understanding of processes that caused the difference with the commercial data, to design the better augmentations processes (for example to generate the random background, similar to possible SAM testbed cases). Mixing the small number of the SAM testbed samples with a large number of available commercial data should also work as a very good augmentation technique.