Primary Architecture

The best performing architecture was a neural network with 2 Conv1d modules, operating over temperature, followed by a Linear layer across m/z channels and then a multi-target classifier.

The network inputs are the first 100 m/z channels, with temperature binned into 25 degree increments. Temperatures included are roughly 100 to 1100 degrees, with random noise applied to produce a diverse set of models.

All Conv1d modules are kernel 3, stride 2, and the first layer involves dilations of 1, 2, 5, 10, 20, 50, and 100, with zero padding. All blocks use Group Normalization, Dropout (~0.1), and Random ReLU (~1/3 to ~1/8 negative slope), with final layer dropout of 0.5.

Three different input scalings are used:
- divide by sum of all m/z values at a given temperature (best)
- divide by sum of all temperature values for a given m/z taken to the 0.5 - 1.0 power (second best)
- divide by the maximum value across the entire array (third best)

Models include either the first, first and second, or all three channels, with substantial gains from ensembling all three approaches.

All models were trained for up to 100 epochs with SGDP *without* momentum at a learning rate of 0.1 and a batch size of 12.

Data Cleaning:
All m/z channels are min-max scaled to their ~0.01 quantile value and to the maximum value present in any spectrum, excluding helium.

GBT Features:
Features include area under the curve, peak value, peak-to-average, peak temperatures, width of peak, jitter, and then various statistics on the top three peaks. Similar to the neural networks, the most important features in terms of performance gain are prevalence of each ion as a percent of the total signal observed at each temperature.

Cross-Validation:
All folds are selected through Multi-Label Stratification to ensure similar target frequencies across all folds.

LightGBM Training:

All parameters are selected through nested cross-validation within each fold, along with some light Gaussian noise, feature dropout, row dropout, and other forms of augmentation.

Stacking:
All models are then stacked with a custom classifier, which minimizes log-loss while constraining weights to be positive and summing to no more than one. These typically involve fairly light L1+L2 regularization.

Performance Improvements:
~200 basis points from adding prevalence features (relative signal of each ion compared to all others, for each temperature interval) to initial LightGBM model

~100 basis points from experiments on epoch count, m/z channel count and dropout, lighter weight decay, random temperature offsets, and a floating logit sum in stacking

~100 basis points from adding tunable dropout to every block of the model, randomly dropping out entire input channels, and then running many more epochs with lighter weight decay on these more robust models

~200 basis points from access to validation data--note: this indicates the model would be substantially more performant with additional data

Improvements:
All models were trained without SAM data, and all SAM predictions were a simple average of models. Cross-validation performance was 0.1001 overall, and 0.1287 on SAM samples, which also happened to have *much* lower predictions than commercial samples.

The vast gap between first place and our model indicates a fundamentally different set of techniques. (All other top teams are at a reasonable gap that exhibits one or two missing architectural or dropout/augmentation/regularization ideas.)

Replacing the Linear layer aggregating across m/z with a Transformer would be a good first place to start, as there are clear relationships between ions. Improving performance on SAM samples is also possible with a CycleGAN model that translates between SAM spectra and commercial spectra (these were distinguishable with 100% accuracy).