kaggle.com

1. Data preprocessing

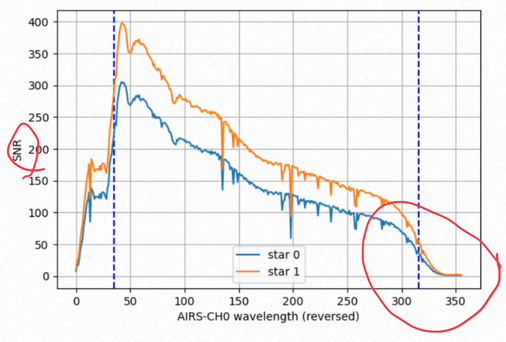
Calibration and spatial summation

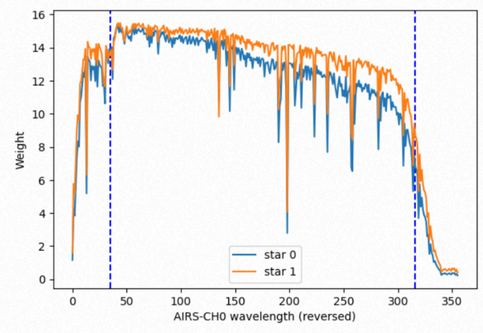
This step mainly reflects the calibration process proposed by the organizers (calibration notebook), with a few modifications aimed at enhancing the signal-to-noise ratio:

- Hot pixels are retained to prevent potential information loss, under the assumption that their signals remain valuable post dark correction.
- Only the top 50% of pixels with the highest intensity values are considered, which helped reduce noise by approximately 4%.
- Binning is reduced to 12 for AIRS-CH0 (and 144 for FGS1).

Weighting and dead pixel mitigation

The data is weighted to prioritize wavelengths with higher SNR during spectral axis averaging. For each wavelength, the weight is calculated as the ratio of the mean of the out-of-transit signal to its variance.





2. Extraction of raw spectrum values

This stage focuses on estimating the $(Rp/Rs)^2$ ratio for each wavelength from preprocessed data.

Detrending and estimation of the transit depth

The computation of the transit depth is based on a polynomial fitting of the temporal signal up to a degree 5, excluding ingress and egress transitions, with a multiplicative shift applied to the in-transit portion.

We use `scipy.optimize.curve_fit` with the following model function and a high uncertainty `sigma` on the transitions:

```
def fit_fn(ts, transit_depth, *trending_coeffs):
    return Polynomial(trending_coeffs)(ts)* (1. - transit_mask * transit_depth)
```

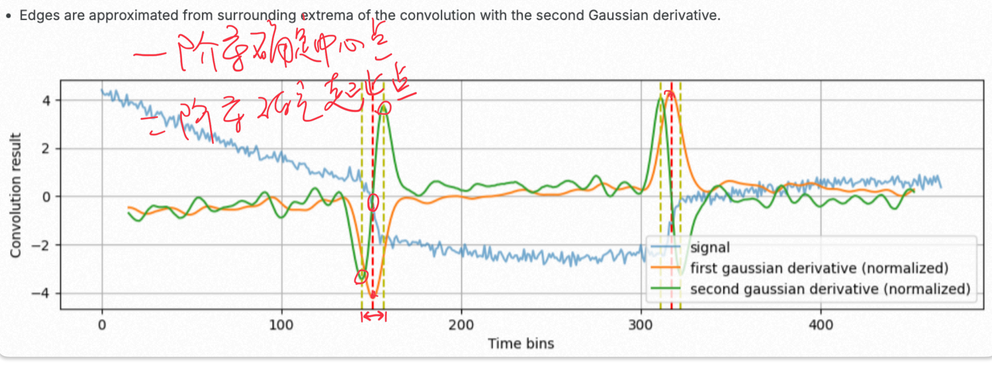
Averaging over wavelengths

To reduce noise, the optimization is applied to the average of the weighted signal over neighboring wavelengths [*k* - *N*, *k* + *N*] for AIRS (including the extra wavelengths out of the 282 targets). As a compromise between noise reduction and loss of precision due to spectral averaging, we selected *N* = 8 for the first 200 wavelengths, and *N* = 20 for the last wavelengths where SNR is very low and the spectrum dynamics apparently lower.

Determination of ingress and egress transitions

Transitions are detected via convolutions with the first two derivatives of a Gaussian:

- Centers are identified as extrema of the convolution with the first Gaussian derivative.
- Edges are approximated from surrounding extrema of the convolution with the second Gaussian derivative.



Improvements on spectrum extraction

- Degree selection:

  We successively evaluate degrees 2, 4, and 5, selecting the one with the best RMS error penalized with the square of the degree. A `savgol` filter is applied to each segment outside transitions to ensure representative RMS differences.

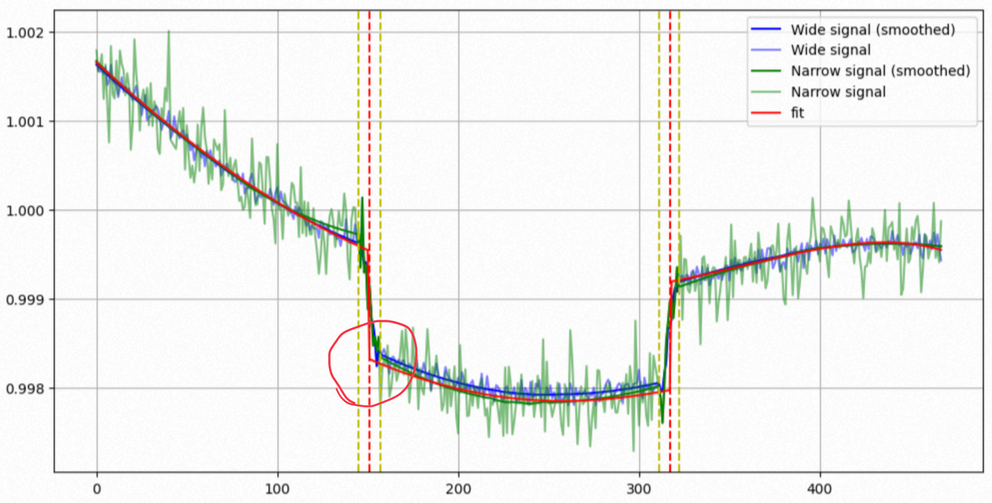
  This degree selection reduced the risk of overfitting noise in the transit area, that would degrade transit depth estimation.

  In addition, it proved useful to cap the maximum degree to the one selected for a first overall signal evaluation.

- Global detrending and estimation on a narrow wavelength range:

  A final boost in precision and score, allowing us to go past 0.700 on the leaderboard (LB) on the last few days, consisted on first evaluating a more reliable detrending polynomial on a wider range of wavelengths (up to 100 on each side), and keep this polynomial in a subsequent optimization on the initial range with only 3 parameters: an offset and a magnitude factor and the researched transit depth.

  A narrower range of *N* = 5 instead of *N* = 8 is also considered around wavelength 183 for stars 0 and 1 (CH4 peak).



3. Spectrum postprocessing and sigma estimation

This stage involves experimental heuristics and rule-based adjustments.

Spectrum dynamics consideration

The main feature that enabled us to reach LB 0.65 was to segregate the post-processing and sigma estimation of a spectrum in function of its dynamics.

- For low dynamics (75% of transits), an average prediction line is applied.
- For high dynamics, the raw spectrum is retained but adapted.

We initially based the determination of the dynamics on the extent of the $(Rp/Rs)^2$ ratio. We then switched to an evaluation of the best correlation with the training set labels. In our final submissions, we included additional Taurex-generated samples with random chemistries (+0.005 on private LB).

Another modification we performed to increase the score was to slightly distort the average line in the direction of the raw spectrum (proportional to the degree of correlation).

Offset correction

We then reached LB 0.68 by slightly increasing the spectrum values in function of the average Rp/Rs ratio. We first assumed this was necessary to counter the effect of limb darkening, but this is most likely needed to counter not so representative starts and ends of transitions (with a detection step not capturing the early changes of slope when the planet starts to enter or exit the limb of the star).

[edit: it might indeed be linked to the added foreground, as presented by @cnumber in https://www.kaggle.com/competitions/ariel-data-challenge-2024/discussion/543853#3034316]

Spectrum adaptations

After application of a `savgol` filter and offset correction, additional tweaks are performed:

- Clipping enveloppe, e.g., -1. * std in the middle of the spectrum.
- Replacing the final portion with a linear ramp except for star 0.

Sigma estimation

Sigma varies with wavelength. Is is empirically constructed, considering higher sigma for deeper transits and for values that are farther from the mean. Two different sets of parameters are used based on the identified spectrum dynamics (low / high).

4. Final spectra refinement using PCA

Once the per-planet processing is complete, principal component analysis (PCA) is applied to refine the spectra, removing residual noise and artifacts. This step, applied per star keeping the first 5 components for reprojection, reduces RMSE from 4.3e-5 to 3.3e-5 on the training set and increases LB score by 0.015.