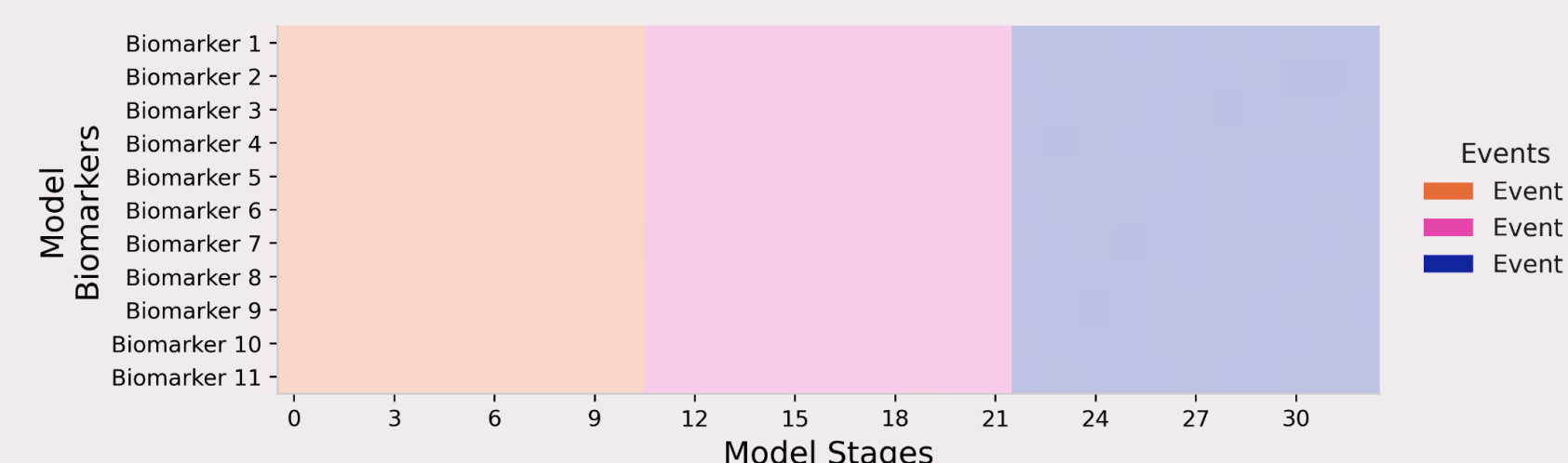


Background

- The Subtype and Stage Inference (SuStaln)¹ algorithm has been widely applied to different diseases on very diverse sets of cross-sectional data.
- SuStaln enables both the discovery of biomarker progression patterns, and the assignment of new patients to the discovered patterns.
- SuStaln forcibly assigns patients' data to one of the estimated trajectories, adding noise to the discovered patterns and hindering the reliability of out of sample assignments.
- Following robust clustering approaches (i.e. noise mixture modeling and sample trimming)² we aim to enhance SuStaln pattern estimation and outlier detection.

Materials & Methods

We propose the addition of the Null subtype – i.e. a uniform probability distribution of events along the sequence (given that the events happen in natural order) – to the SuStaln framework:



We simulate multiple scenarios where we generate synthetic data from two ground truth (GT) sequences and ten outlier sequences, we control for: tau rank correlation between GT sequences (–0.2 to 0.2), outlier sequences and GT sequences (–0.2 to 0.4), standard deviation of added noise to biomarker measurements (0.15 to 0.5), and outlier proportion (5 to 20%).

We discard the bottom 20% of the samples according to the ratio between the probability to one of the estimated sequences and the probability to the Null.

We evaluate the estimated model on the correlation between model likelihood and model probability, on the estimated sequences using the Hellinger distance (overlap between the estimate and the ground truth), and on the percentage of outliers detected.

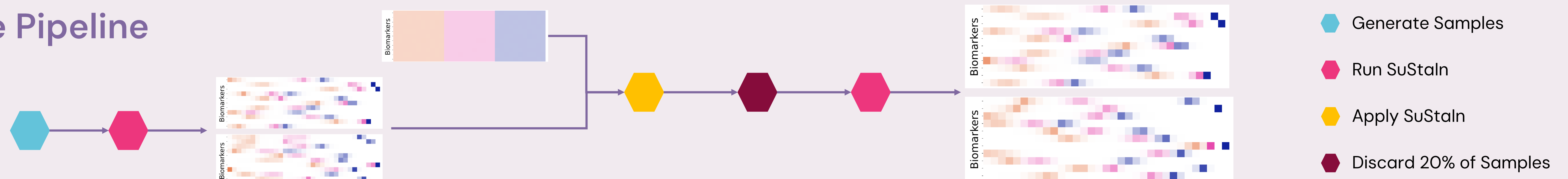
Conclusions

The Null subtype and the proposed pipeline shows:

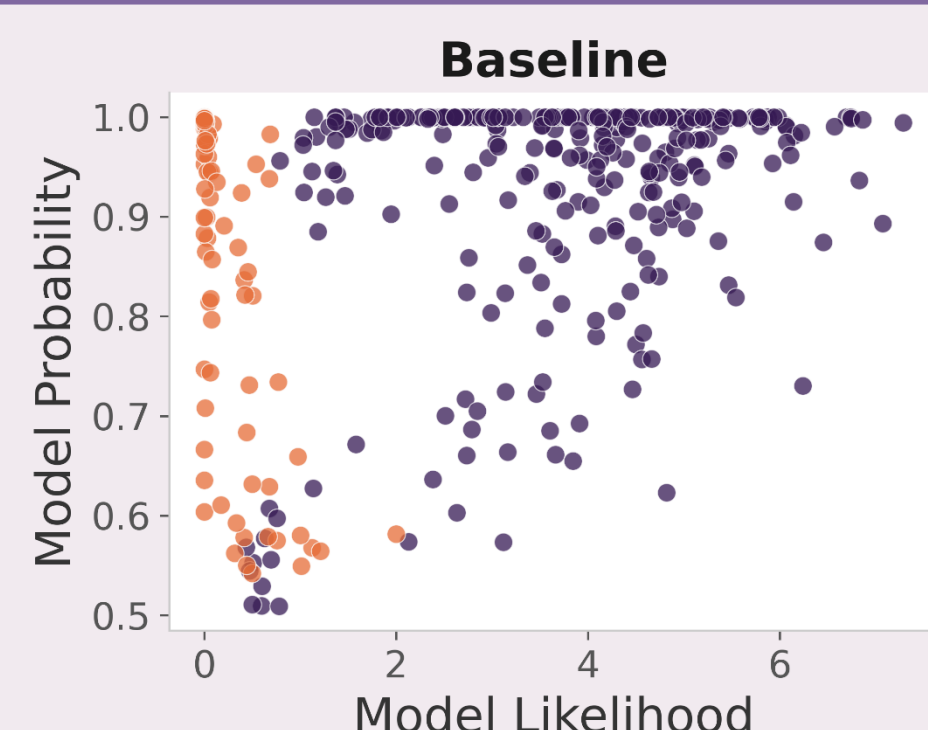
- ✓ Improved correlation between model probability and likelihood.
- ✓ Up to 0.15 better Hellinger distance between pattern estimates and ground truth sequences.
- ✓ More outliers detected.

The Null subtype represents a promising approach for improving **interpretability** and **robustness** of SuStaln.

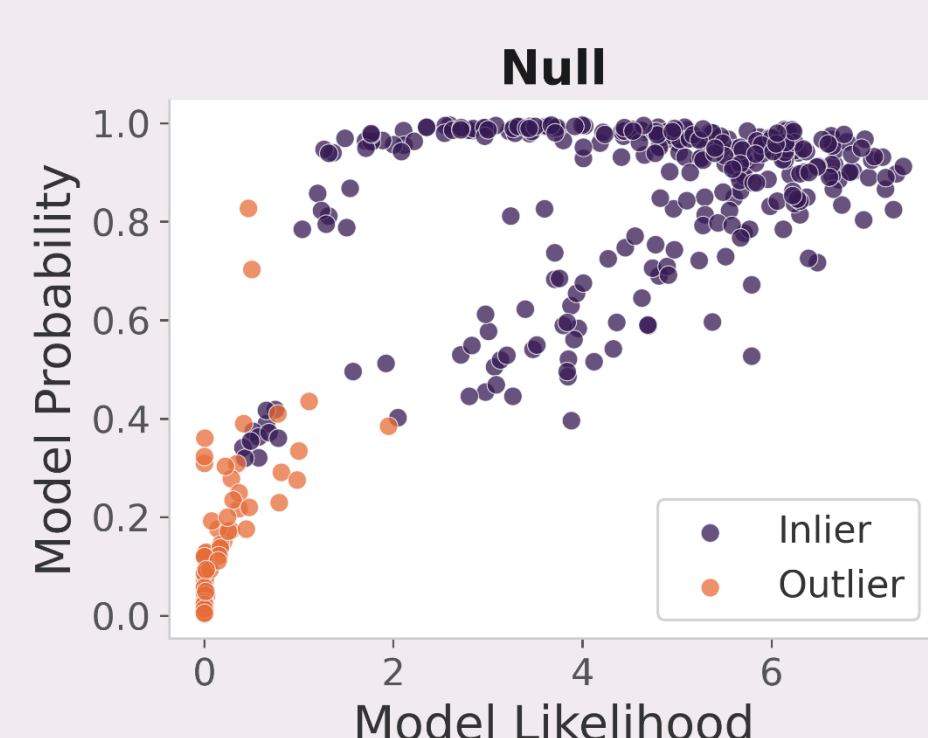
Null Subtype Pipeline



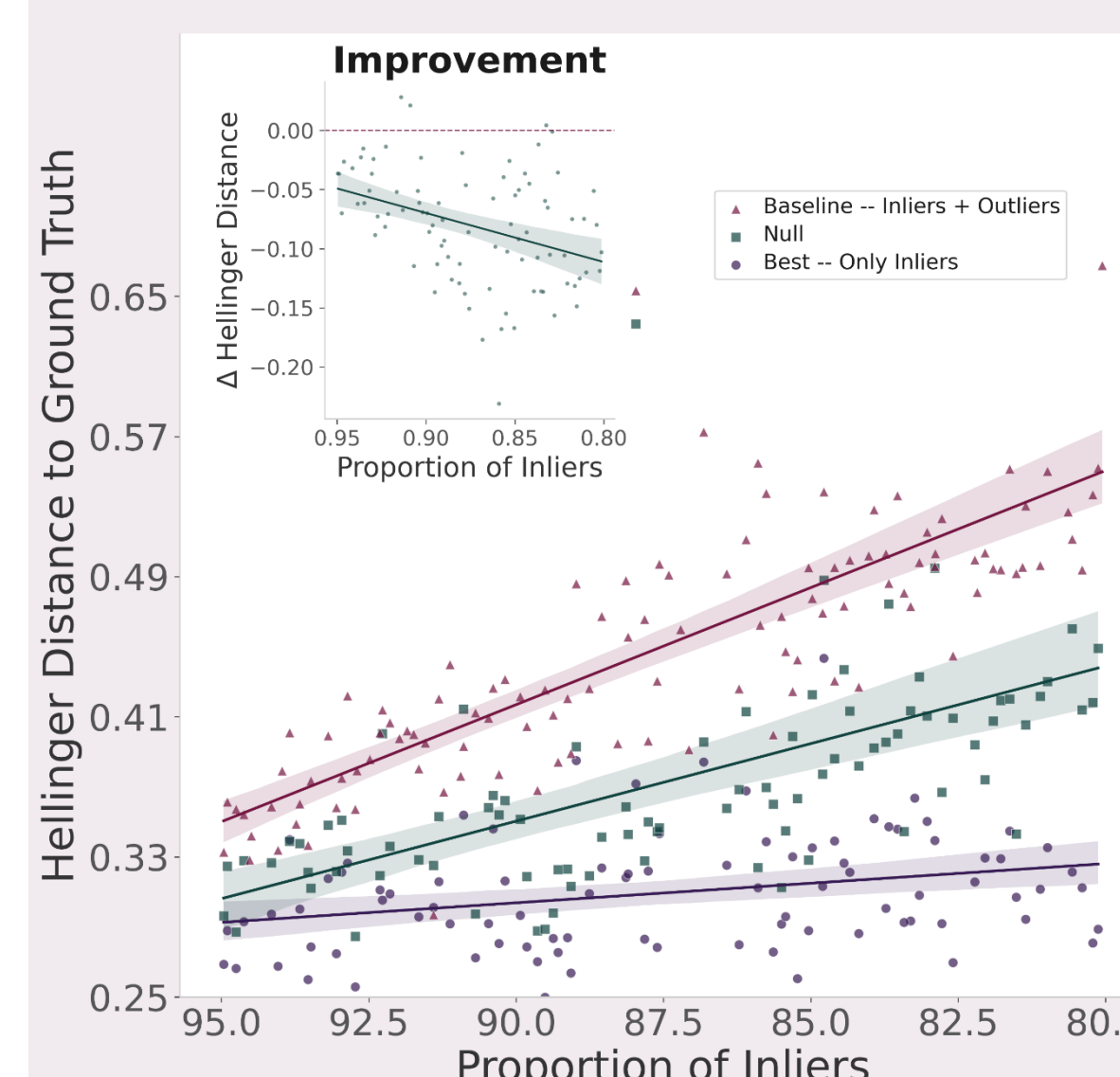
Results



When SuStaln is forced to assign data samples to the discovered subtypes, their subtype probability and subtype likelihood is uncorrelated.

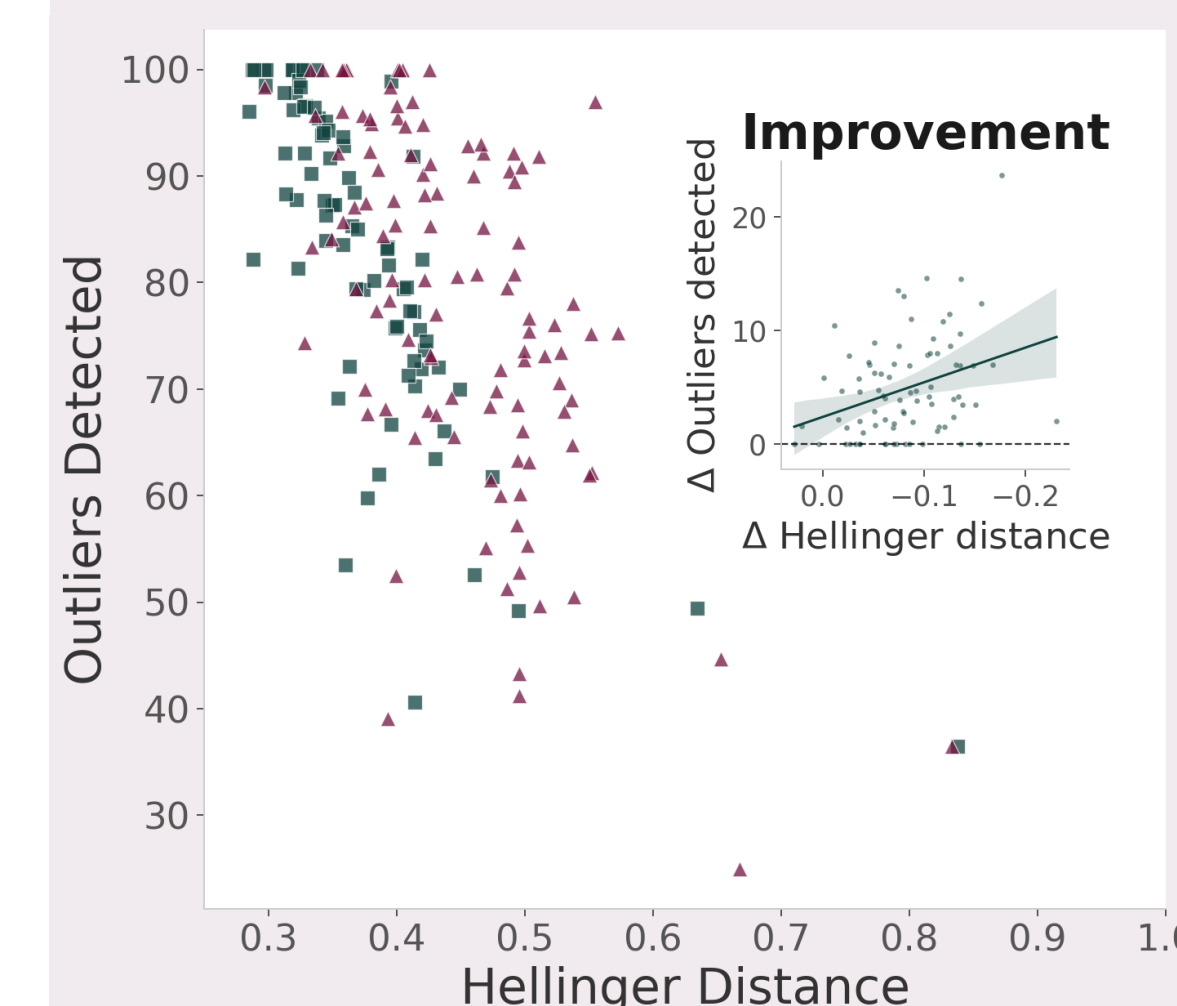


Adding the Null subtype makes SuStaln output more **interpretable**; now higher model probability implies higher model likelihood.



In the presence of outliers in the training sample, the estimated sequences diverge significantly from the best-case scenario (no outliers).

Adding the Null subtype, allows to discard weak assignments, improving the estimates after re-training SuStaln.



When evaluating out of distribution samples, adding the Null subtype helps SuStaln to detect outliers.

The better the estimates, the more outliers are detected, making SuStaln more **robust**.

References

- [1] A. L. Young, *et al.* Nat Commun 2018
- [2] H.H. Bock, Statistics in Transition 2002

Contacts

Gonzalo Castro Leal; gonzalo.leal.23@ucl.ac.uk
PhD Candidate
Hawkes Institute and Department of Computer Science, University College of London, UK.