SOUMAP Clustering* *for* Large-Scale YSO Classification *in* SESNA

**and a lot of imputation preprocessing*

Josh Taylor & Stella Offner

CosmicAI

The University of Texas at Austin

ODEN INSTITUTE FOR COMPUTATIONAL ENGINEERING & SCIENCES

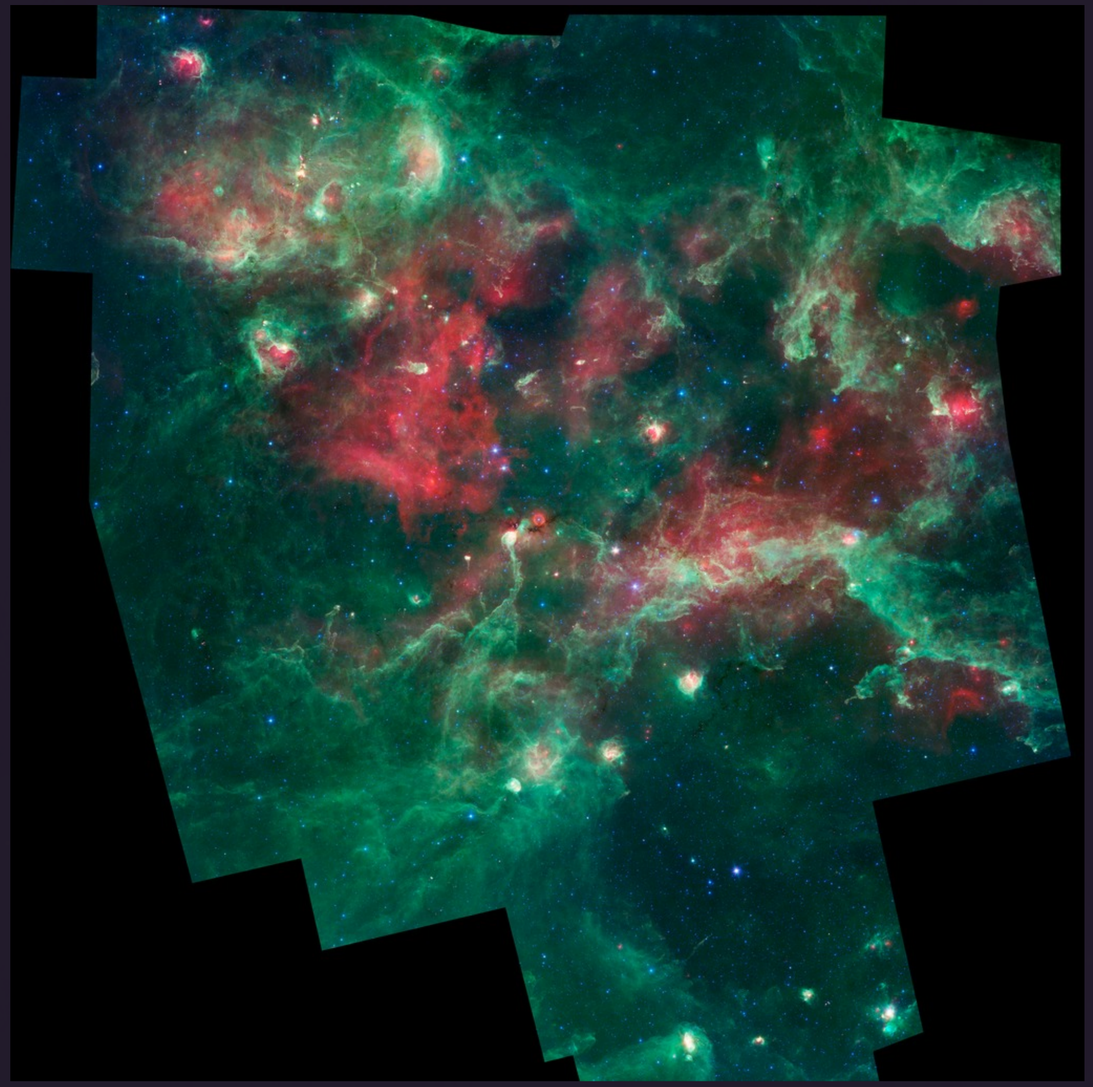
SESNA — The Spitzer Extended Solar Neighborhood Archive

R. A. Gutermuth et al 2009 ApJS 184 18

➡ Uniformly reprocessed **Spitzer mosaics** across nearby star-forming regions

➡ **Combines:**
2MASS (J, H, Ks)
+ IRAC (3.6–8.0 μm)
+ MIPS (24 μm)
➜ 8 photometric bands per source

➡ **This work:**
30 regions, ~8.4M sources

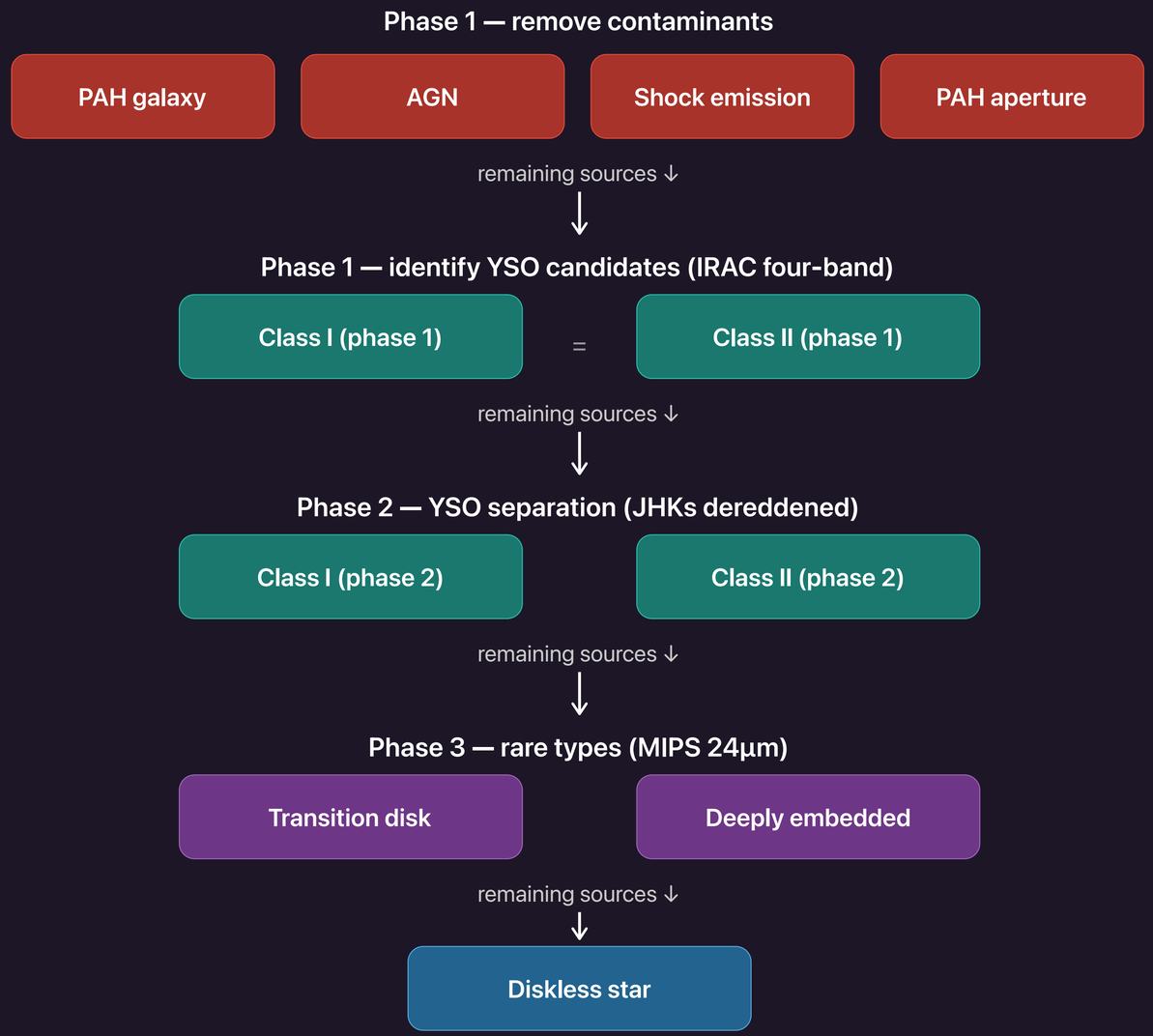


Cygnus X *NASA/JPL-Caltech/Harvard-Smithsonian CfA*

SESNA Color Cut Labeling

Labeling via sequential hard cuts
each source classified exactly once, where it first triggers a color-cut boundary:

- **Phase 1** remove galaxy "contaminants", *then look for YSOs downstream*
- **Phase 1** requires all IRAC detections *locates candidate YSOs*
- **Phase 2** requires all 2MASS *Class I / II discriminator*
- **Phase 3** requires MIPS 24mu transitional / deeply embedded protostars
- **Phase 4** all remaining = mature stars

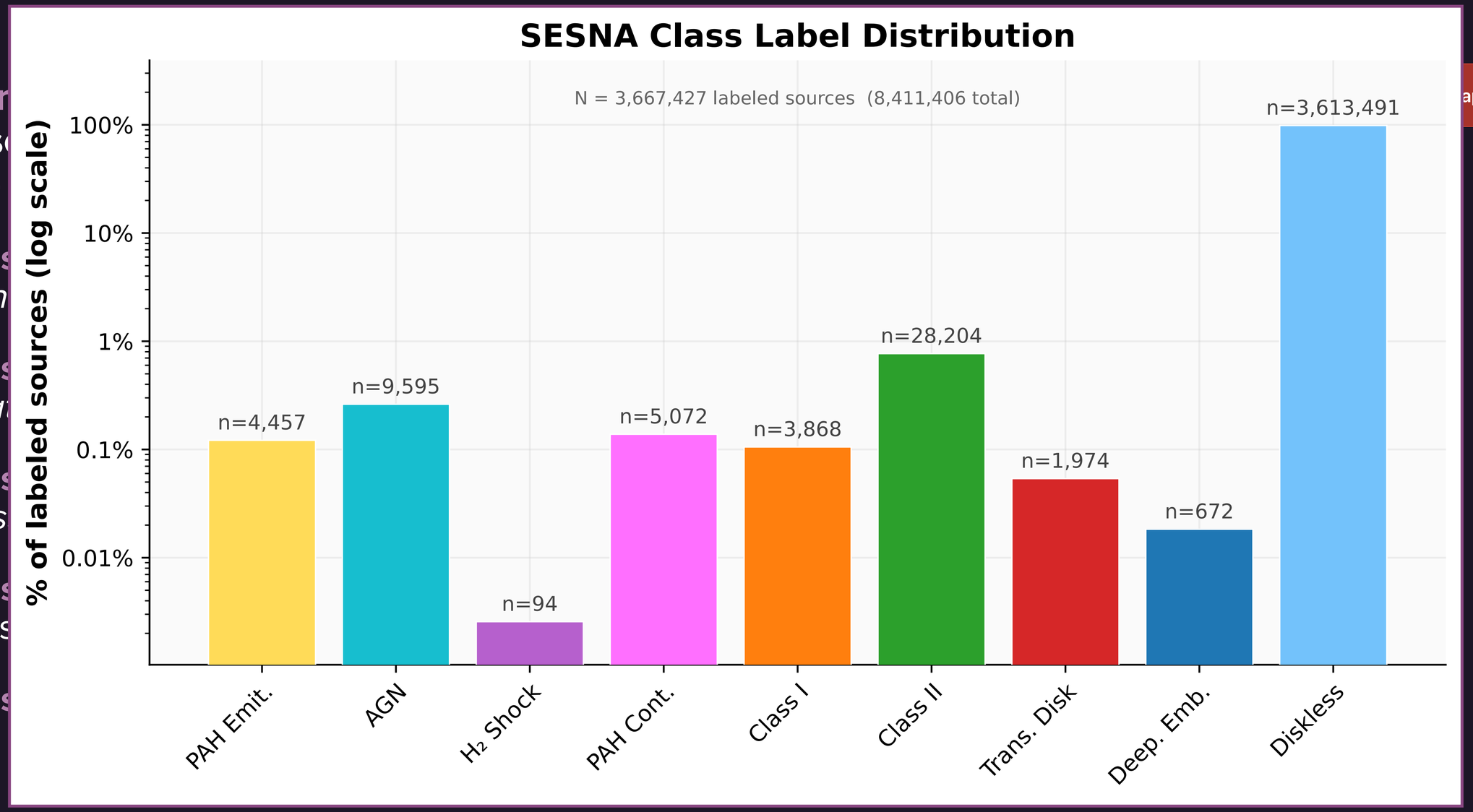


SESNA Color Cut Labeling

Labelin
each s
where

- **Phas** *then*
- **Phas** *loca*
- **Phas** *Clas*
- **Phas** trans
- **Phas**

aperture

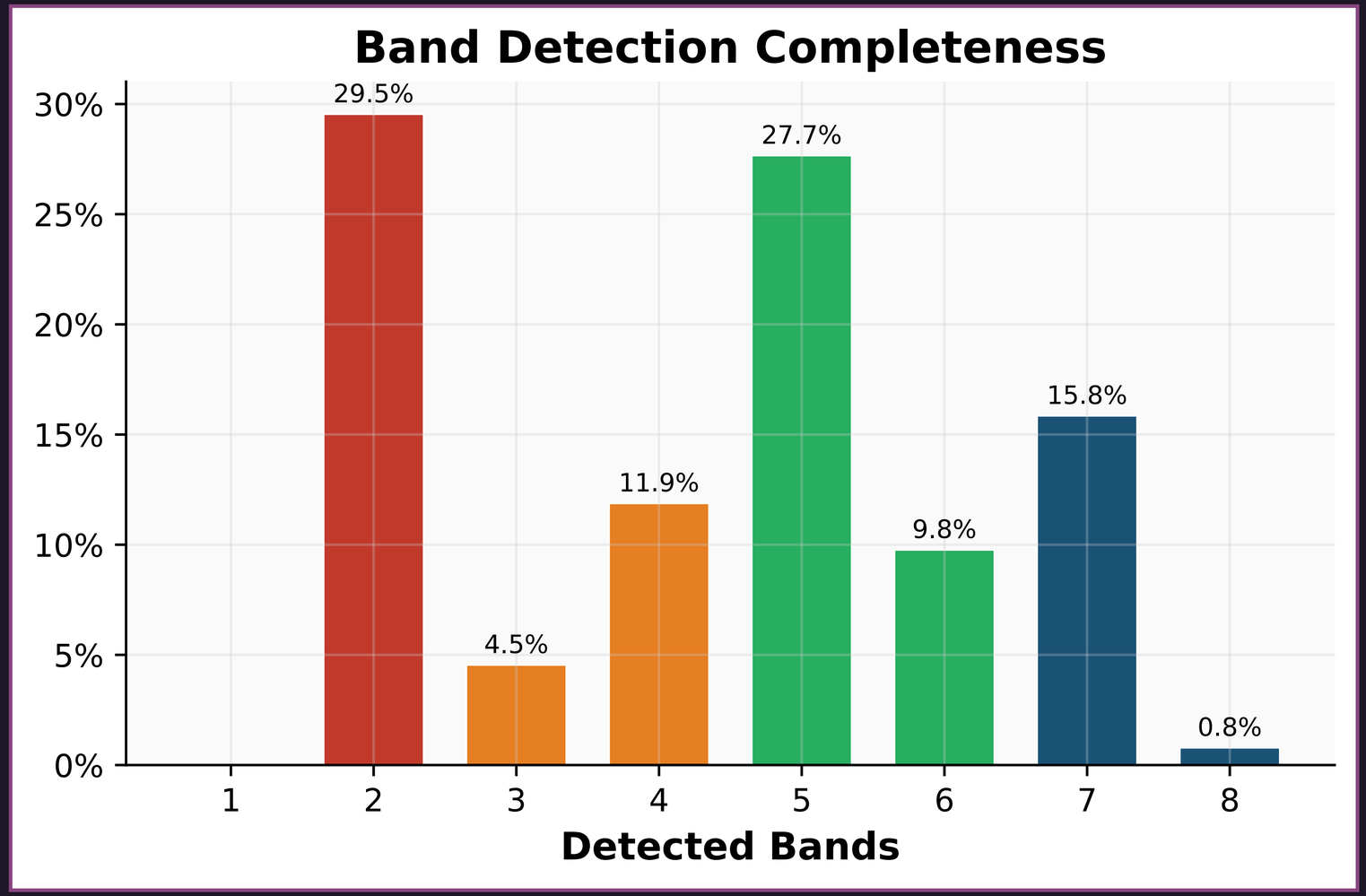


SESNA (In)Completeness

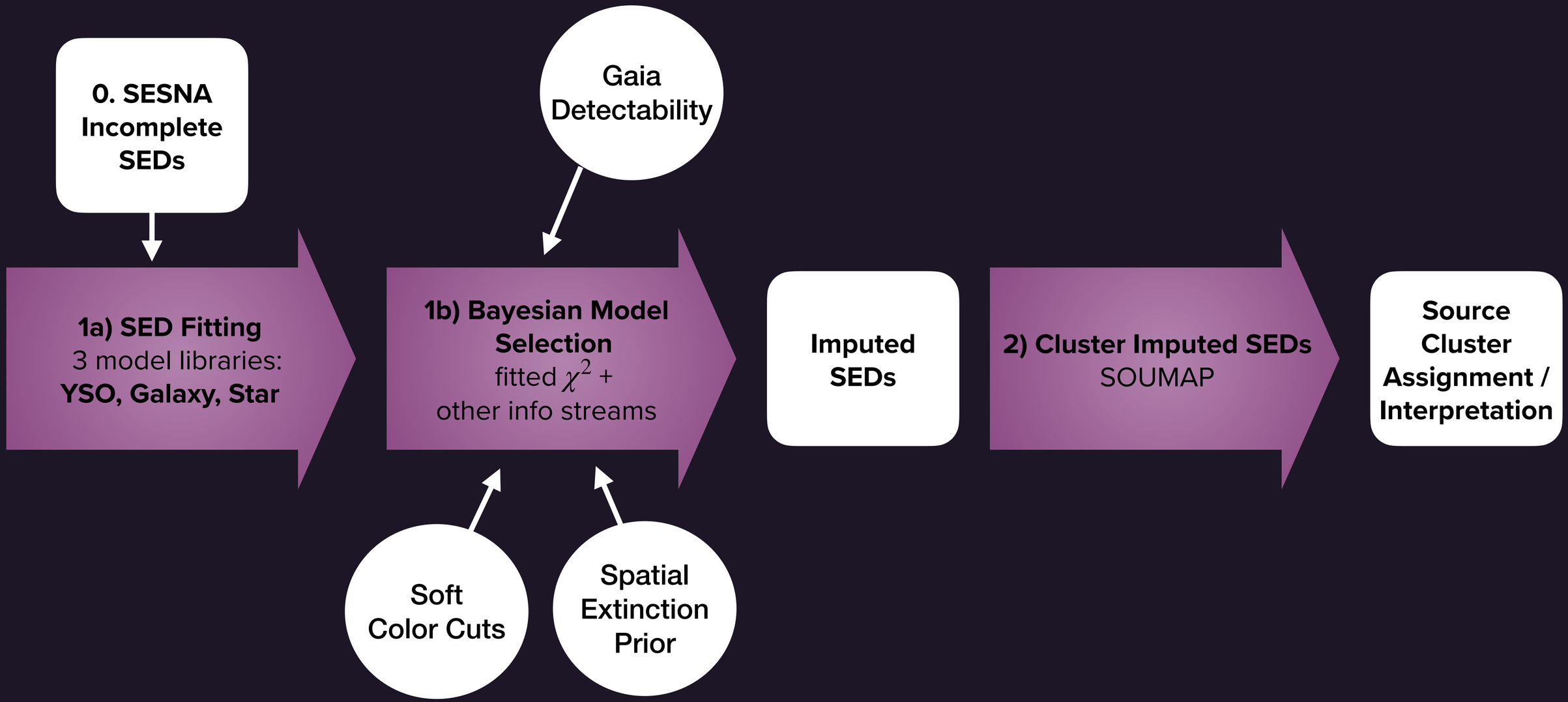
Color-cut classification requires specific band combinations

➡ ~5M sources can't be classified at all

➡ Complicates downstream ML / analysis



Plan: Complete SESNA, then Cluster



Part 1:
Model SED Imputation for SESNA

SED Fitting

SED fitting with 3 model libraries built from parameter grids:

- **YSO** (2M+ → 200k models)
  Richardson et al 2024 ApJ 961 188

- **Galaxy** (17k models, z- + luminosity)
  Berta et al. 2013 A&A 551 A100

- **Star** (3k models)
  Castelli & Kurucz 2003 IAU Symp. 210 A20

Top K = 5 fitted models + χ^2 stored from each model library

Bayesian Model Selection

Posterior class probability
YSO / GAL / Star

$$P(C \mid F_\nu, s) \propto P(C \mid s) \times \sum_{k=1}^{K=5} \mathscr{L}_{\text{fit}}(m_k, s)^{\alpha} \times \mathscr{L}_{\text{Gaia}}(m_k, s) \times \mathscr{L}_{\text{color}}(m_k, s)$$

Bayesian Model Selection

Posterior class probability
YSO / GAL / Star

$$P(C \mid F_\nu, s) \propto P(C \mid s) \times \sum_{k=1}^{K=5} \mathscr{L}_{\text{fit}}(m_k, s)^{\alpha} \times \mathscr{L}_{\text{Gaia}}(m_k, s) \times \mathscr{L}_{\text{color}}(m_k, s)$$

Prior class probability,
boosts YSO and suppresses galaxy prob.
in high-extinction sightlines

Bayesian Model Selection

Posterior class probability
YSO / GAL / Star

Converts model χ^2 into a likelihood, tempered by band completeness level α

$$P(C \mid F_\nu, s) \propto P(C \mid s) \times \sum_{k=1}^{K=5} \mathscr{L}_{\text{fit}}(m_k, s)^{\alpha} \times \mathscr{L}_{\text{Gaia}}(m_k, s) \times \mathscr{L}_{\text{color}}(m_k, s)$$

Prior class probability,
boosts YSO and suppresses galaxy prob. in high-extinction sightlines

Bayesian Model Selection

Posterior class probability
YSO / GAL / Star

Converts model χ^2 into a likelihood, tempered by band completeness level α

$$P(C \mid F_\nu, s) \propto P(C \mid s) \times \sum_{k=1}^{K=5} \mathscr{L}_{\text{fit}}(m_k, s)^{\alpha} \times \mathscr{L}_{\text{Gaia}}(m_k, s) \times \mathscr{L}_{\text{color}}(m_k, s)$$

Prior class probability,
boosts YSO and suppresses galaxy prob. in high-extinction sightlines

Rewards models whose optical (G-band) brightness matches source's Gaia detection status

Bayesian Model Selection

Posterior class probability ***YSO / GAL / Star***

Converts model χ^2 into a likelihood, tempered by band completeness level α

Soft distance to SESNA's color-cut boundaries

$$P(C \mid F_\nu, s) \propto P(C \mid s) \times \sum_{k=1}^{K=5} \mathscr{L}_{\text{fit}}(m_k, s)^{\alpha} \times \mathscr{L}_{\text{Gaia}}(m_k, s) \times \mathscr{L}_{\text{color}}(m_k, s)$$

Prior class probability, ***boosts YSO and suppresses galaxy prob. in high-extinction sightlines***

Rewards models whose optical (G-band) brightness matches source's Gaia detection status

Model Imputed Flux

$$P(C \mid F_\nu, s) \propto P(C \mid s) \times \sum_{k=1}^{K=5} \mathscr{L}_{\text{fit}}(m_k, s)^{\alpha} \times \mathscr{L}_{\text{Gaia}}(m_k, s) \times \mathscr{L}_{\text{color}}(m_k, s)$$

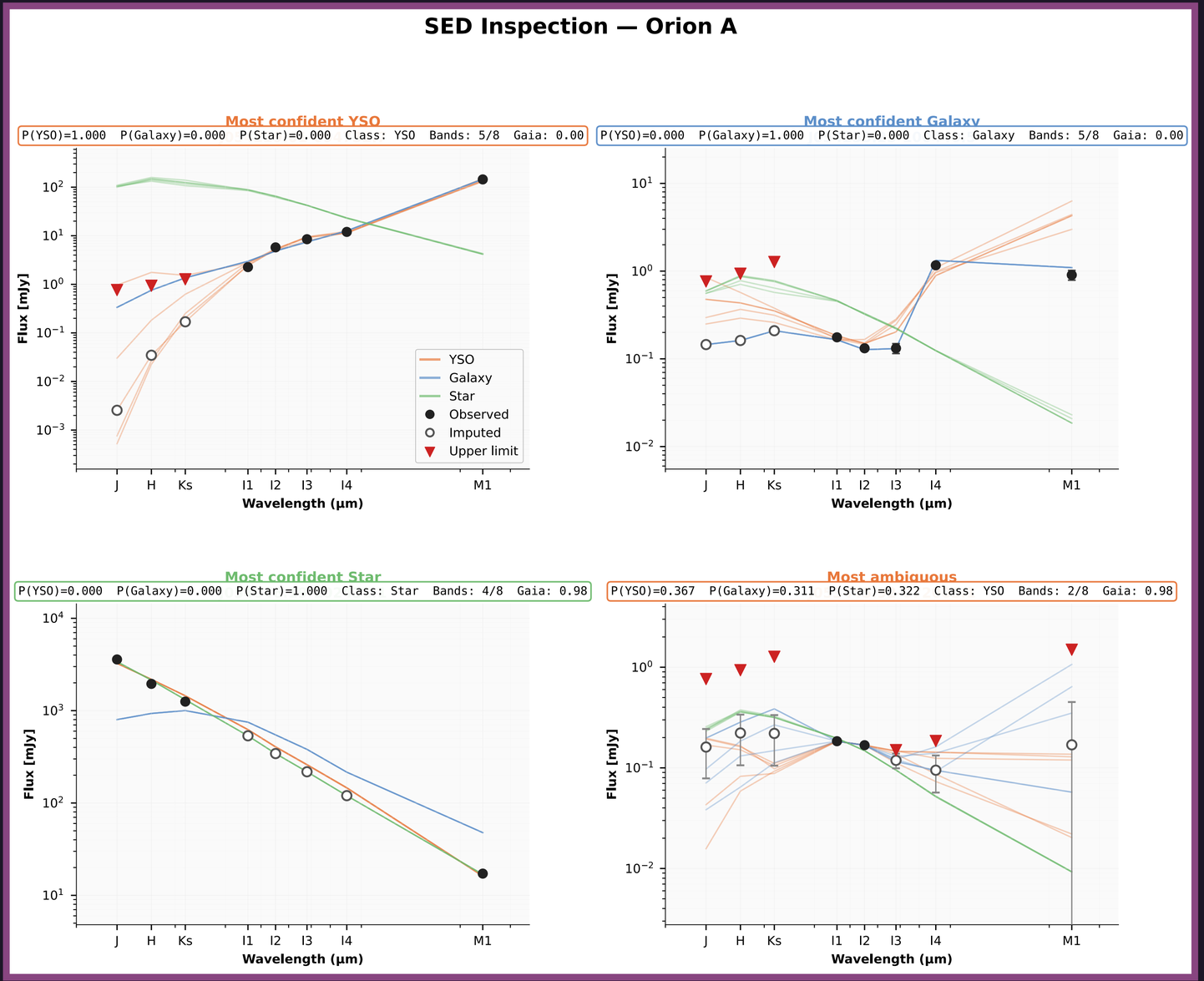
Impute missing flux + uncertainties from posterior-weighted model average

Observed flux is preserved

Imputation Diagnostics

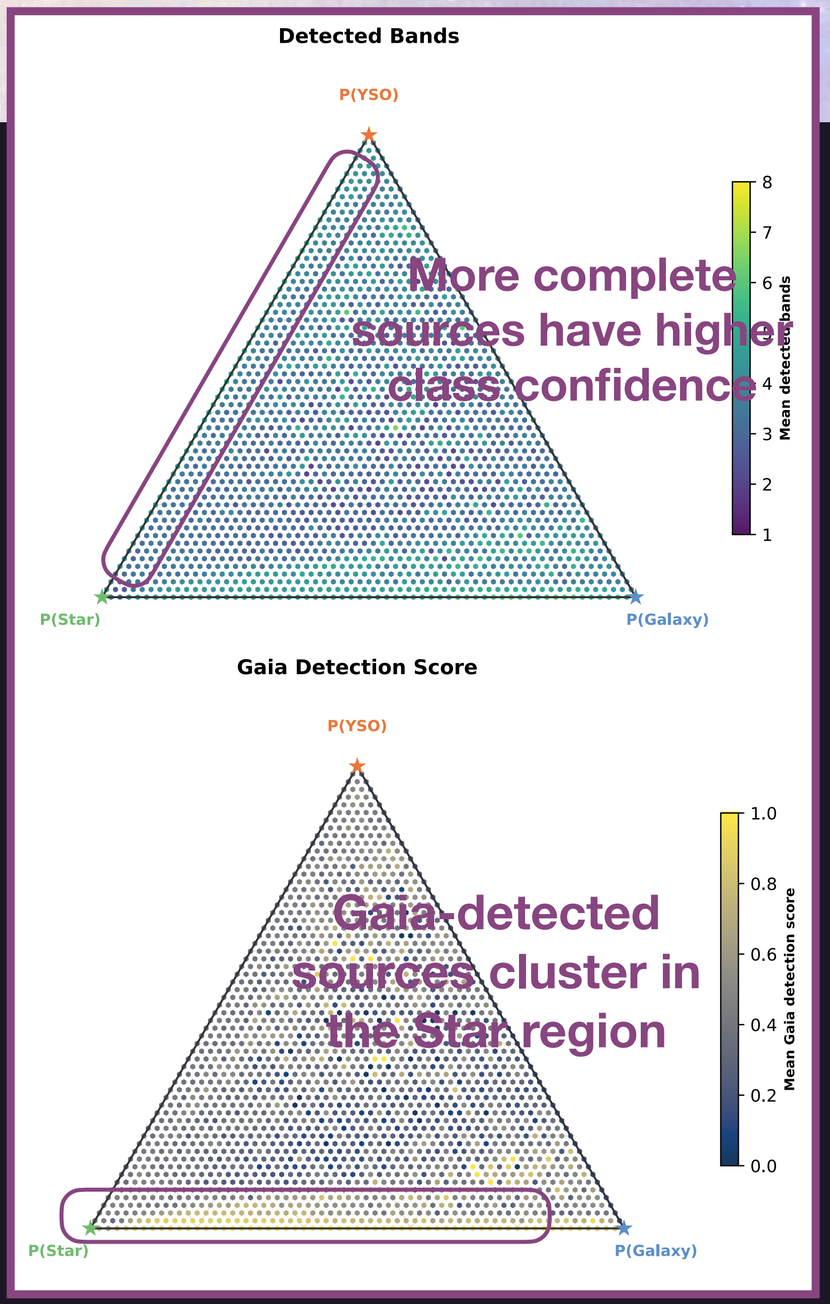
Observed flux preserved, missing flux filled from posterior-weighted model averages

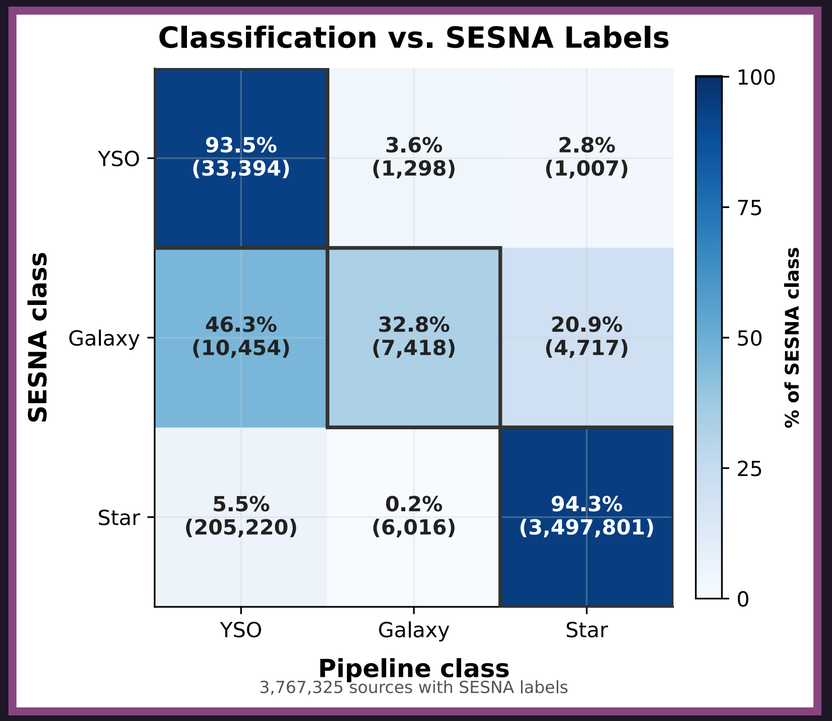
bottom right:
2 bands observed — ambiguous posterior



Imputation Diagnostics

Ternary plots show class posteriors track physical expectations





Posterior analysis shows agreement with SESNA labels for YSOs and Stars

Part 2:
Self-Organizing UMAP Clustering

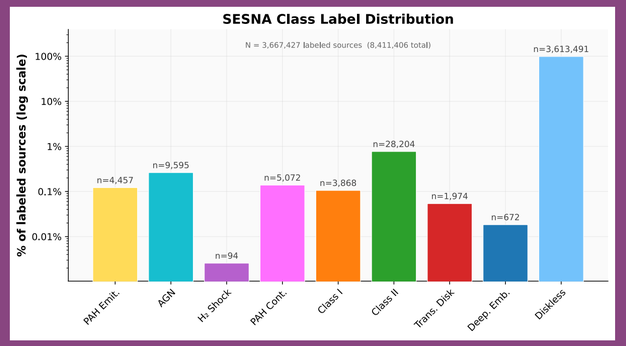
Taylor, J. & Offner, S. 2024, WSOM+LVQ, Springer Nature

SOUMAP Clustering

- ➡ Learn prototype representations of the data in multiple views (*high-D + latent)*,
- ➡ fuse the view,
- ➡ cluster the fused view

Preliminary Clustering - Gating Stars (stargating?)

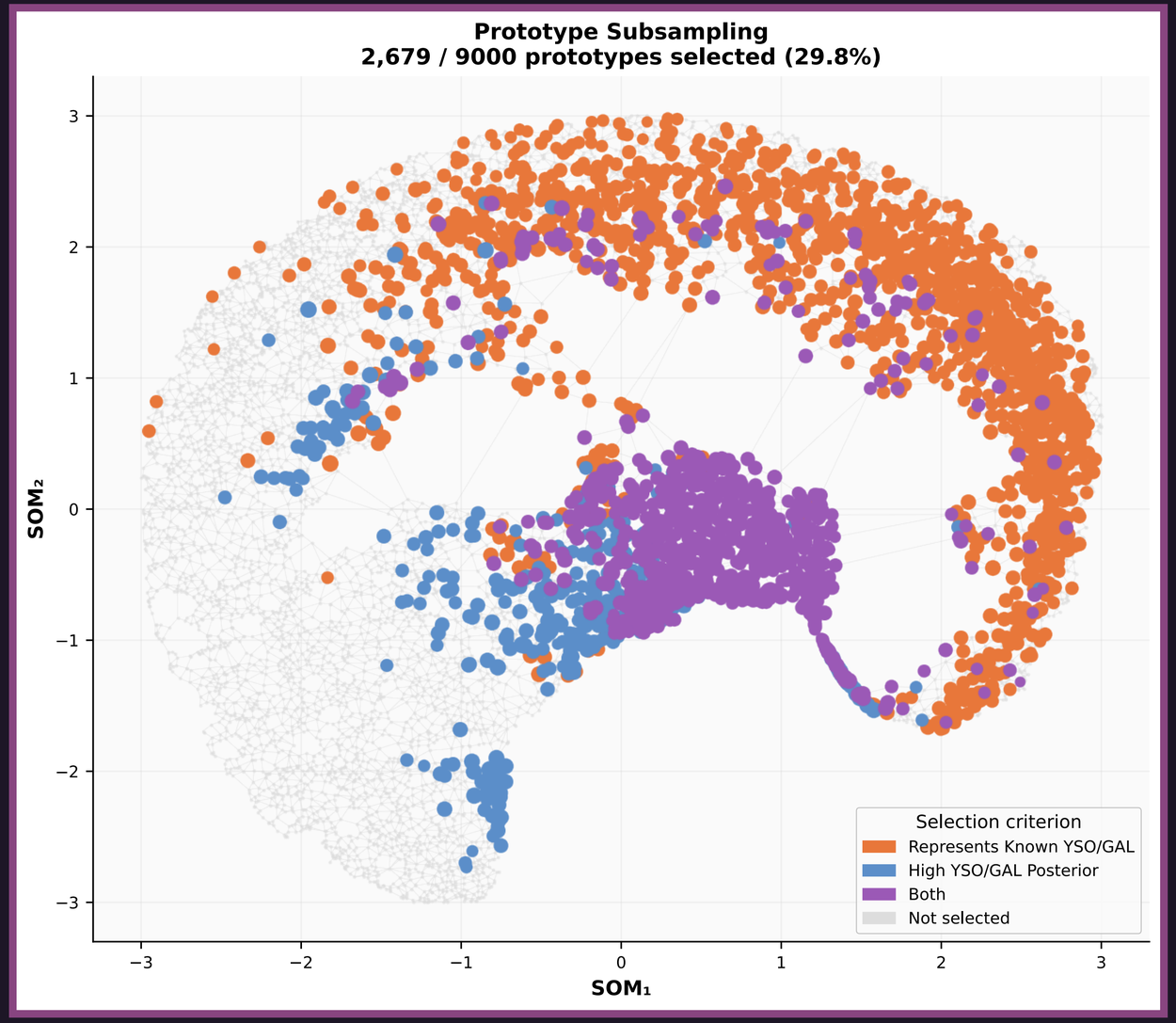
Stars overwhelm



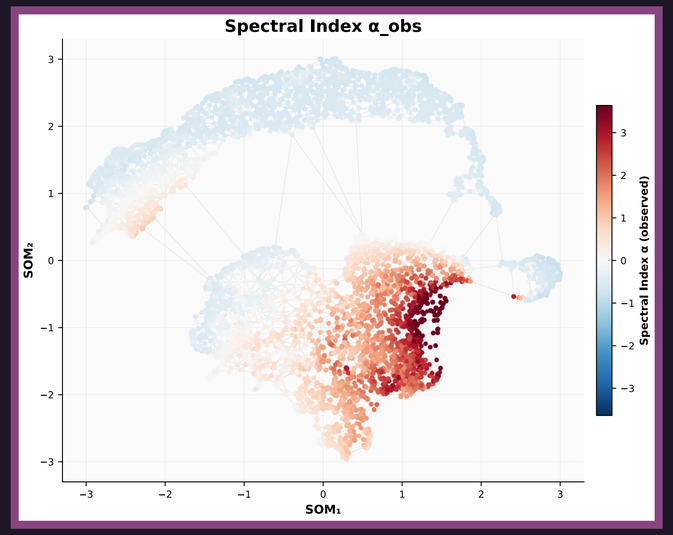
Initial SOUMAP embedding identifies **data subset most likely to contain YSOs**

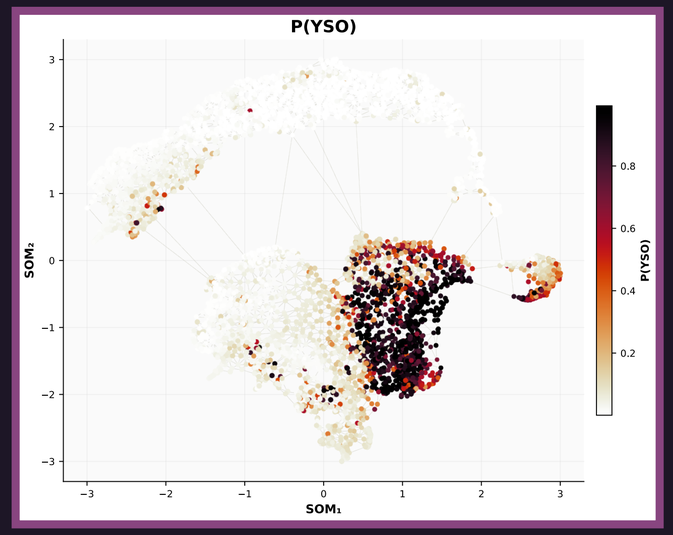
Extract subset for formal clustering

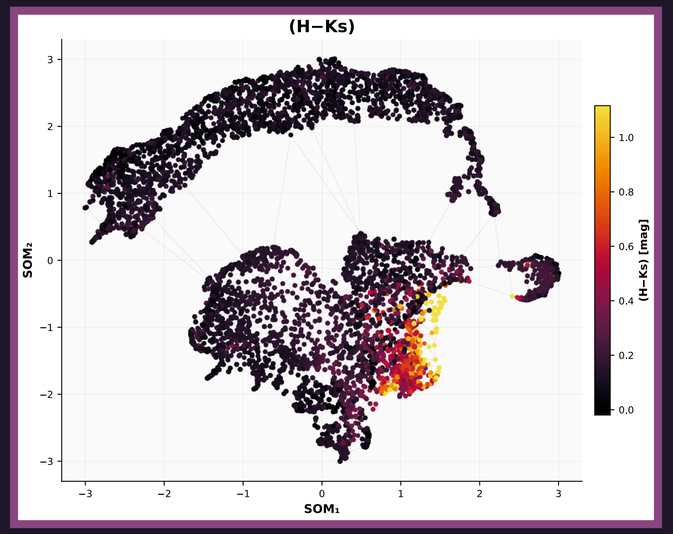
N = 8.4M ➡ N = 2.5M

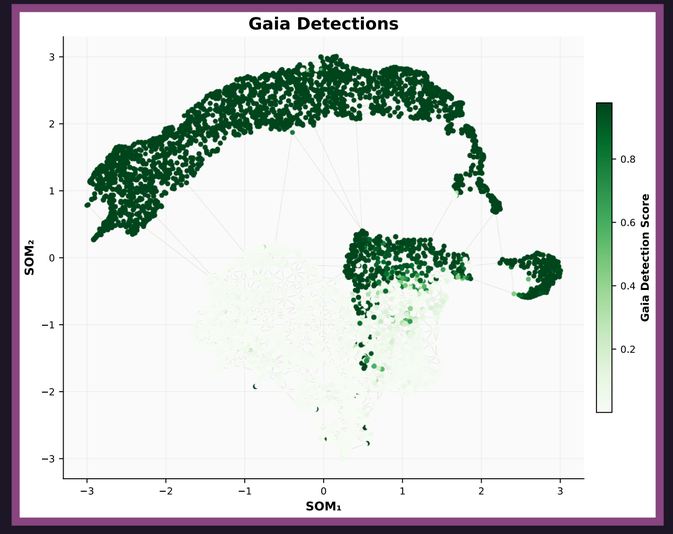


SOUMAP Embedding

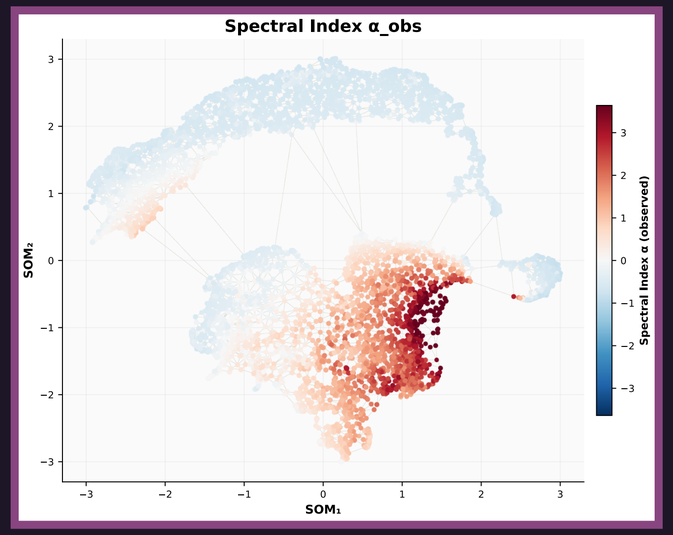


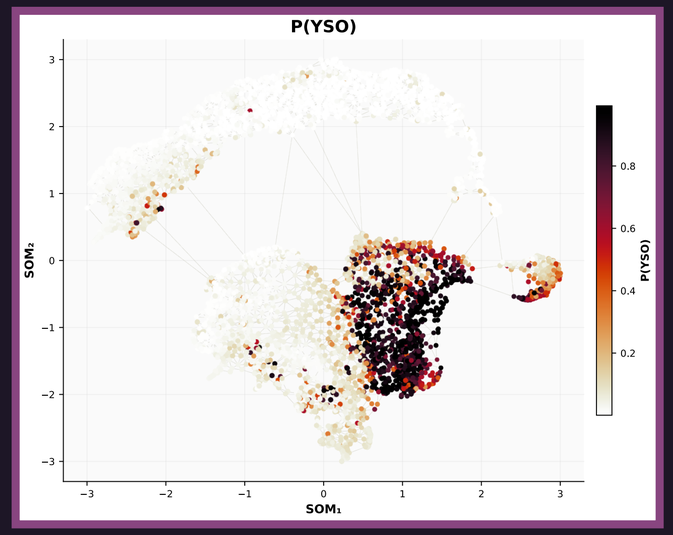


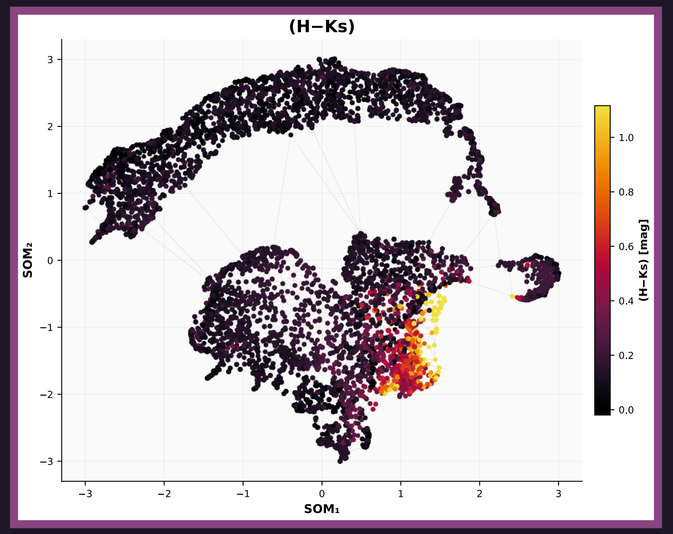


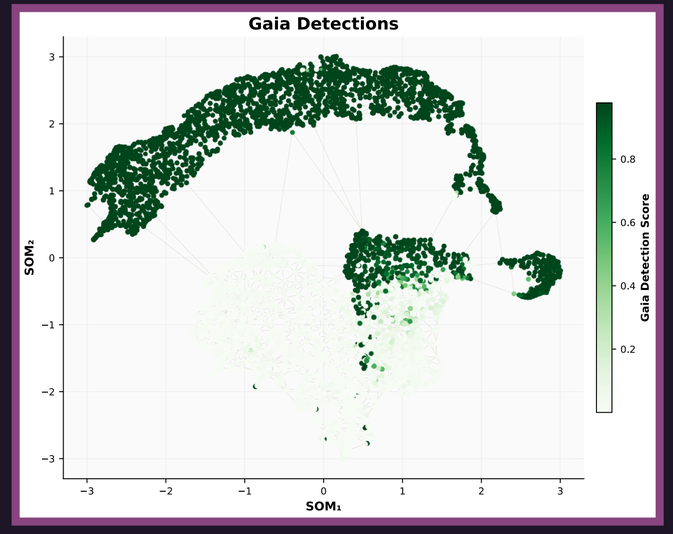


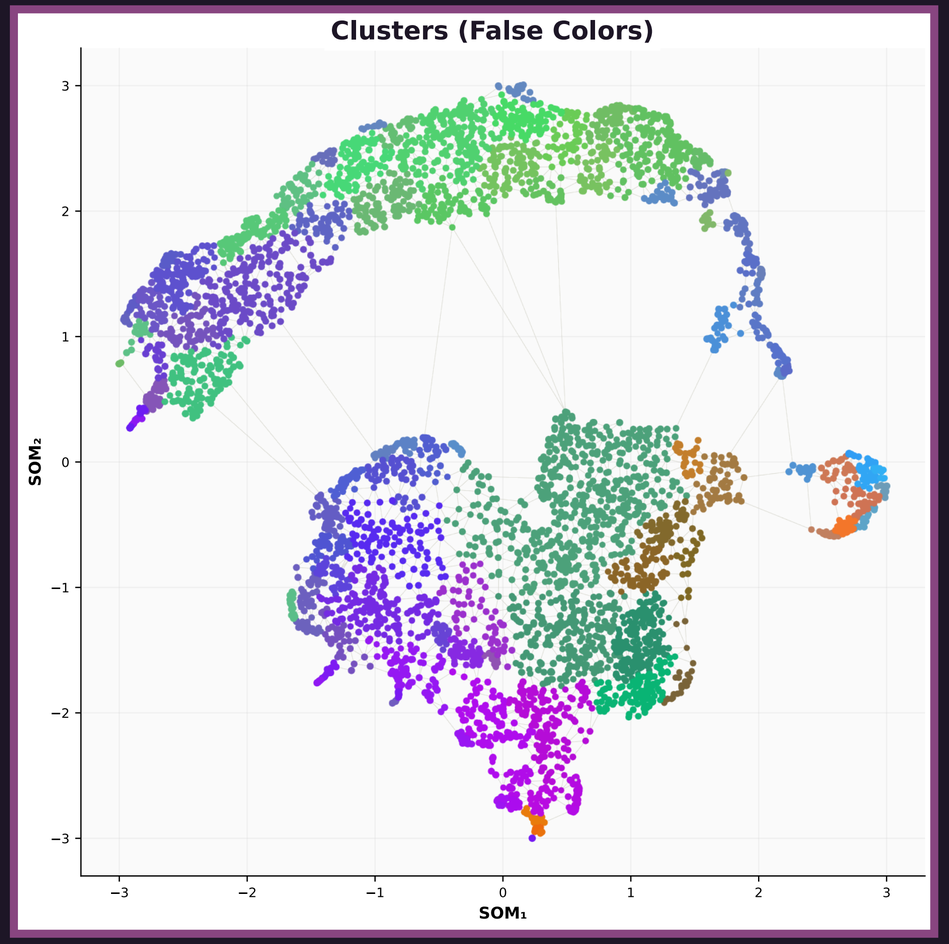
SOUMAP Embedding + Clustering





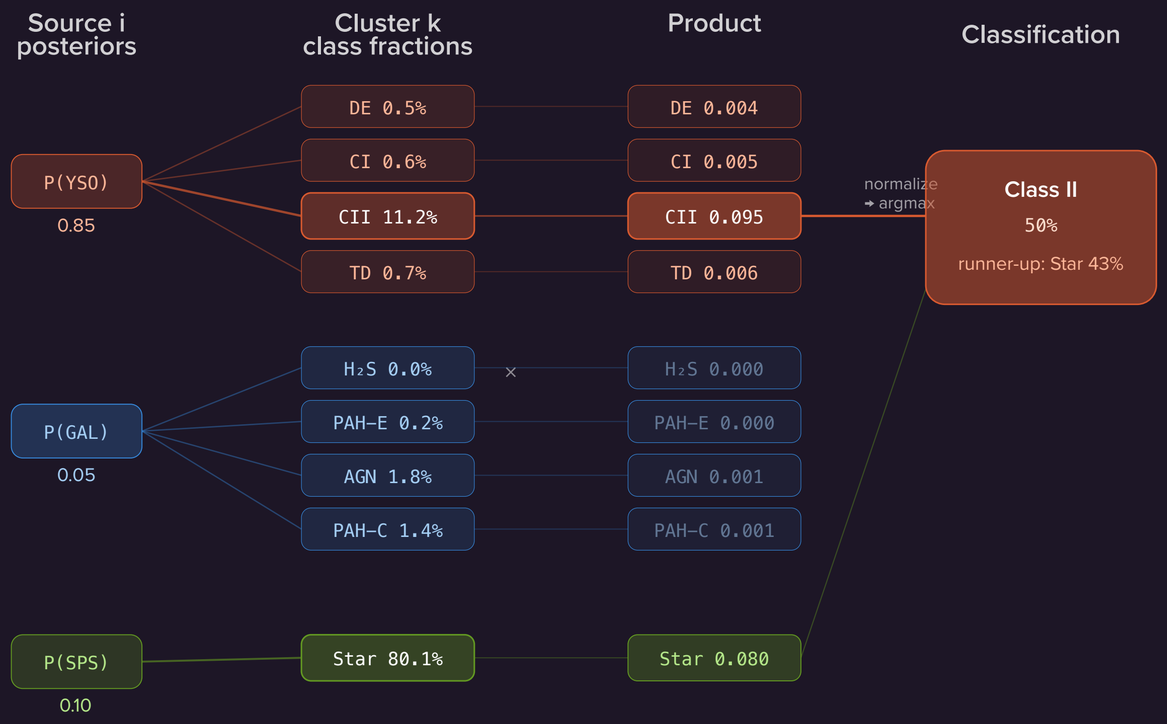






SOUMAP Cluster-Driven Classification

Classification Logic

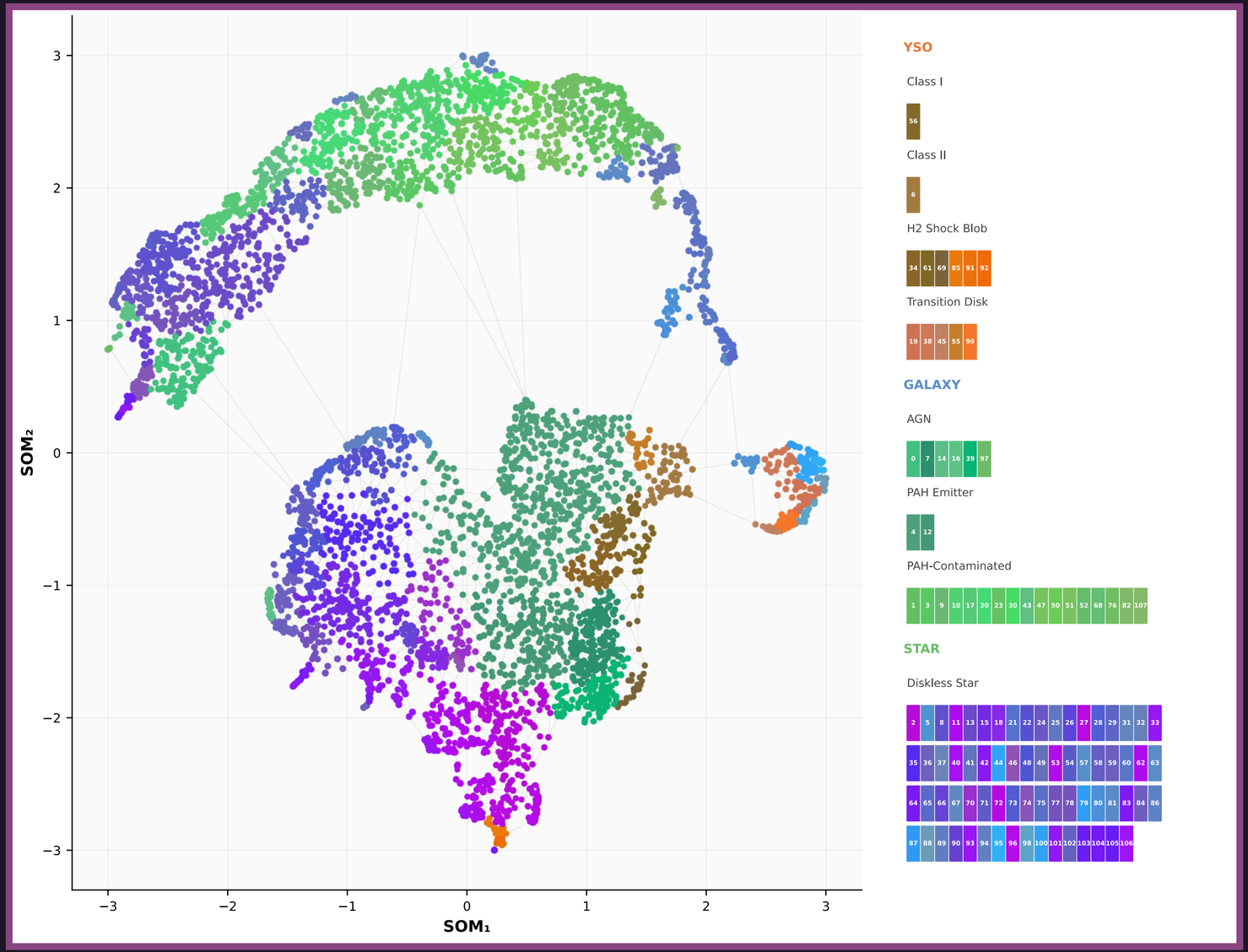


P(YSO) = 0.85 boosts all YSO subclass fractions
P(GAL) = 0.05 suppresses all galaxy fractions
The SED decides — same cluster, different source → different classification

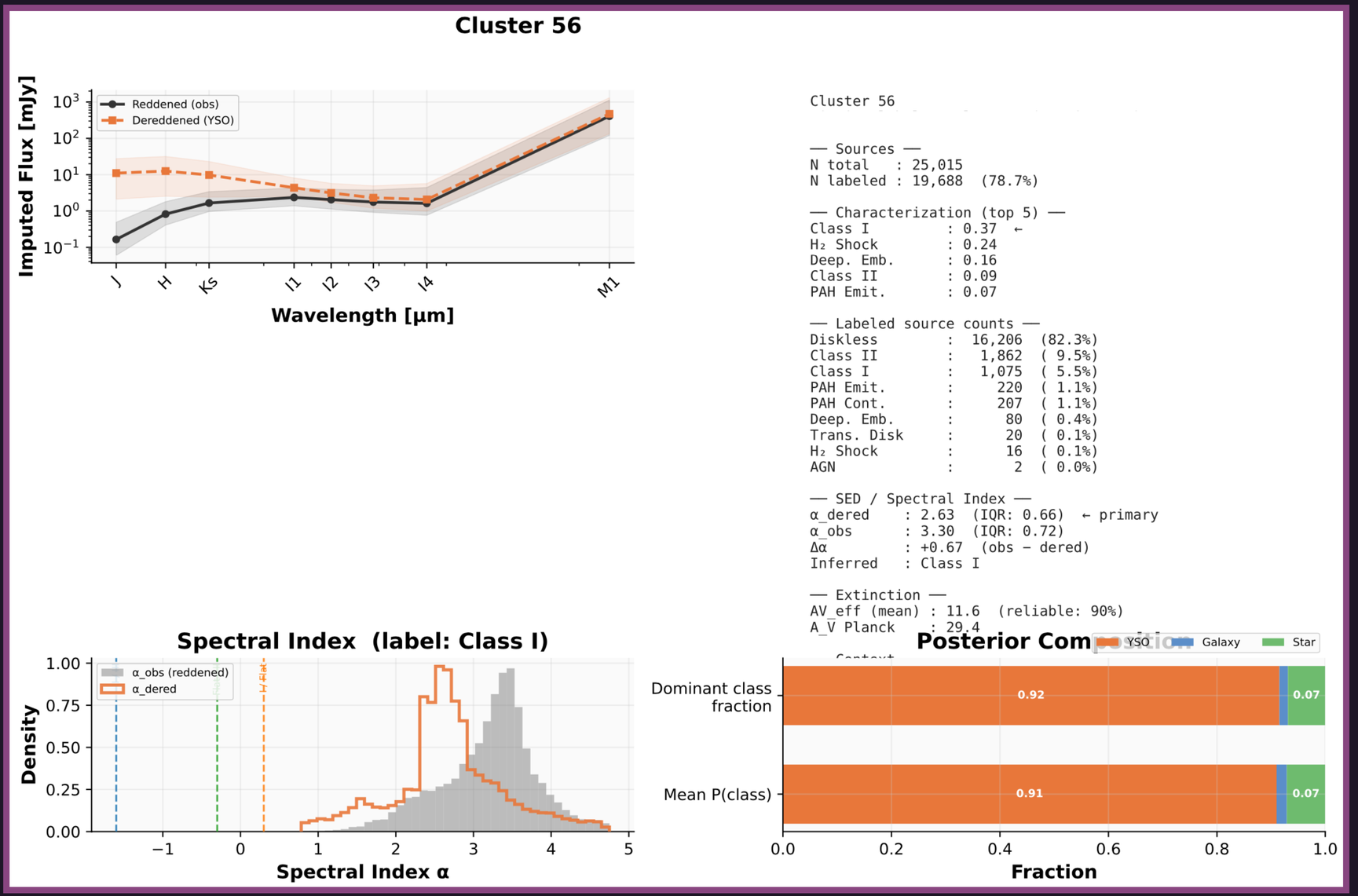
Labeling

	SESNA Class		SOUMAP Cluster Derived Class	
Class	**Count**	**%**	**Count**	**%**
Deep Embed.	672	0.01%	17,584	0.21%
Class I	3,874	0.05%	56,884	0.68%
Class II	28,216	0.34%	107,544	1.28%
Transition Disk	1,976	0.02%	4,904	0.06%
H2 Shock Blob	94	0.00%	355	0.00%
PAH Emitter	4,457	0.05%	6,233	0.07%
AGN	9,595	0.11%	26,787	0.32%
PAH Contam.	5,076	0.06%	5,158	0.06%
Diskless Star	3,614,305	42.97%	8,186,813	97.32%
Unclassified	**4,743,997**	**56.4%**	**0**	**0.0%**

Enrichment-Based Cluster Labels



Cluster Passport Inspection



Conclusions

- **Bayesian model imputation completes 8M+ SEDs**
  SED fitting + Gaia detectability + color-cut diagnostics fills ~95% missing photometry
- **SOUMAP embedding reveals physical organization**
  spectral / stellar properties emerge naturally in the learned manifold
- **Two-stage clustering overcomes extreme class imbalance**
  prototype-level extraction gates the 98% stellar population, enables YSO subclass resolution
- **Every SESNA source now has a probabilistic classification**
  ~180k new YSO candidates identified beyond Gutermuth's original catalog