DPAC CU8
Astrophysical Parameters

Coryn Bailer-Jones, CU8 Scientific Coordinator
MPIA Heidelberg

on behalf of CU8

DPAC IR, ESOC, 27-28 Sept. 2011

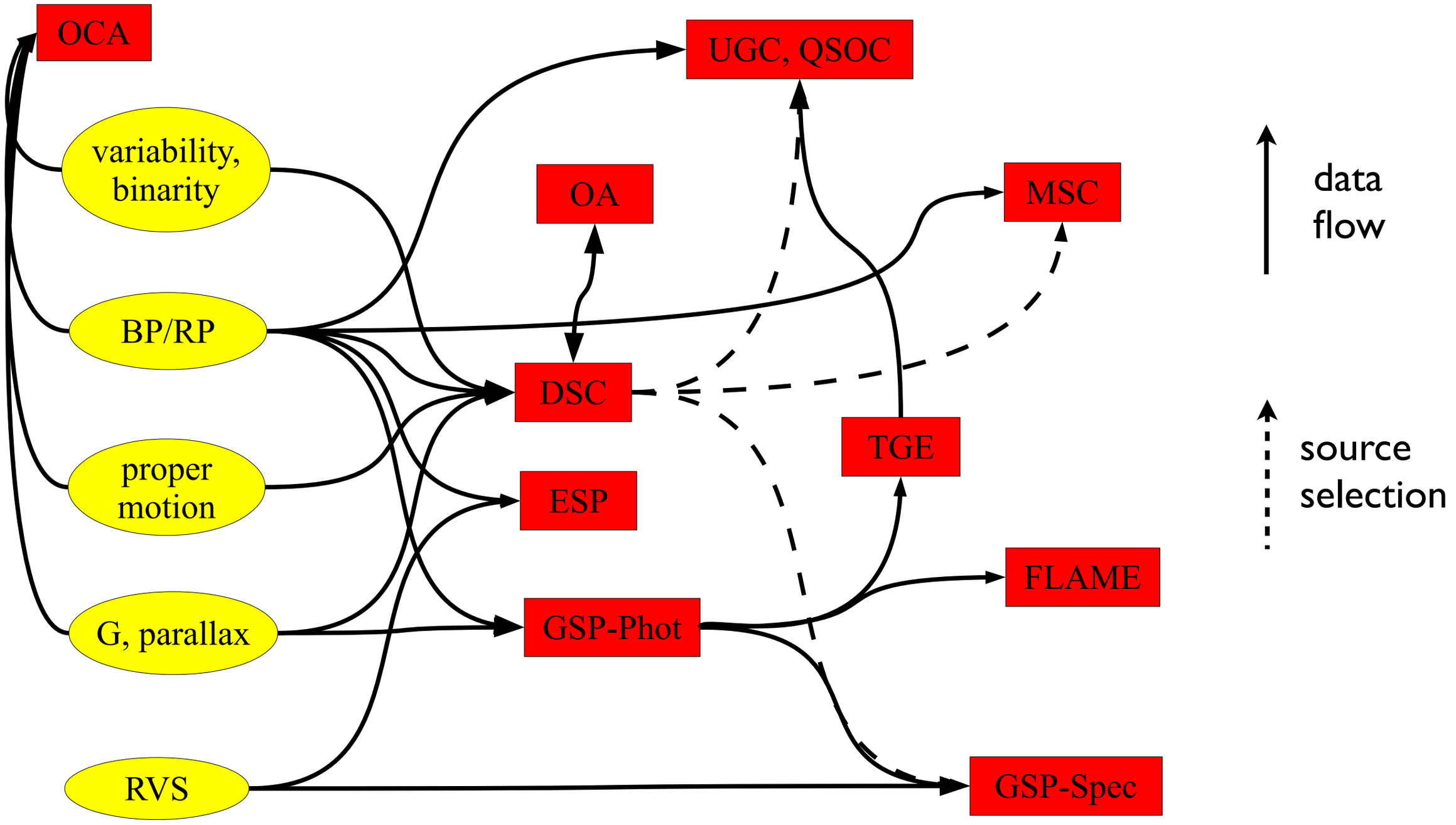
CU8 primary objectives

- Classify all unresolved objects as single star, galaxy, quasar, binary, unknown etc. (assign class probability)
  - ★ identify quasars for the astrometric reference frame
- Estimate main astrophysical parameters (APs) for single stars, T_{eff}, A_0, logg, [Fe/H]
  - ★ using BP/RP, RVS, astrometry
  - ★ provide suitable classifications/parameters to aid extraction and contamination removal in the RVS extraction
- Identify new or unusual types of objects (inevitable in a large, deep survey)

Principles

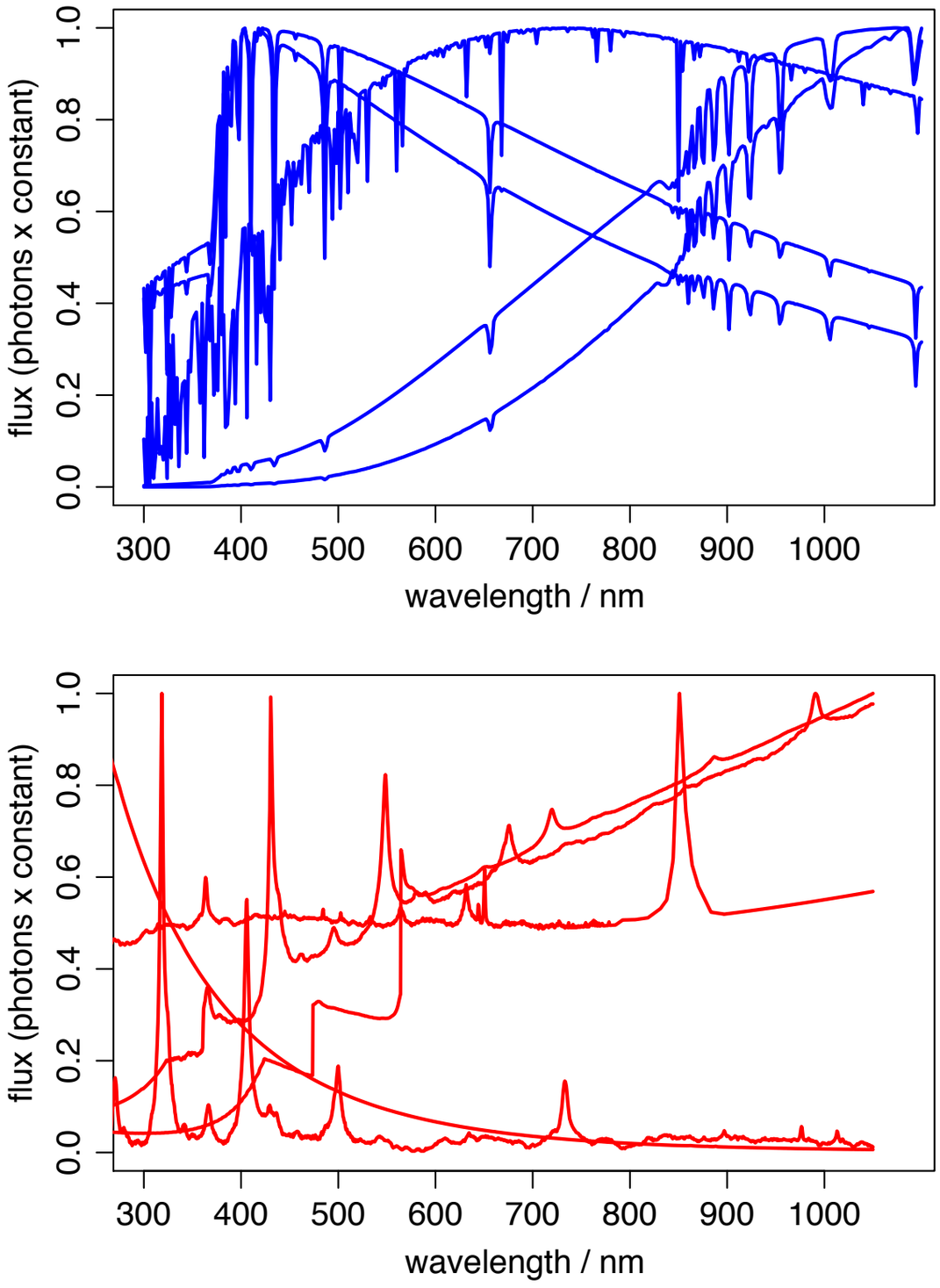
- Predominantly supervised learning methods
  - ★ astrophysical parameters are unavoidably tied to astrophysical models
  - ★ mostly synthetic spectra (also semi-empirical libraries and calibration)
- DSC and GSP-Phot as front-end "work horses"
  - ★ based on BP/RP (available for all objects)
  - ★ make pre-selection for other algorithms
- Other algorithms for RVS spectra, quasars and galaxies, better treatment of "extreme" stars, outlier detection
  - ★ multiple AP assignments for any one object in general

Astrophysical Parameters Inference System (Apsis)

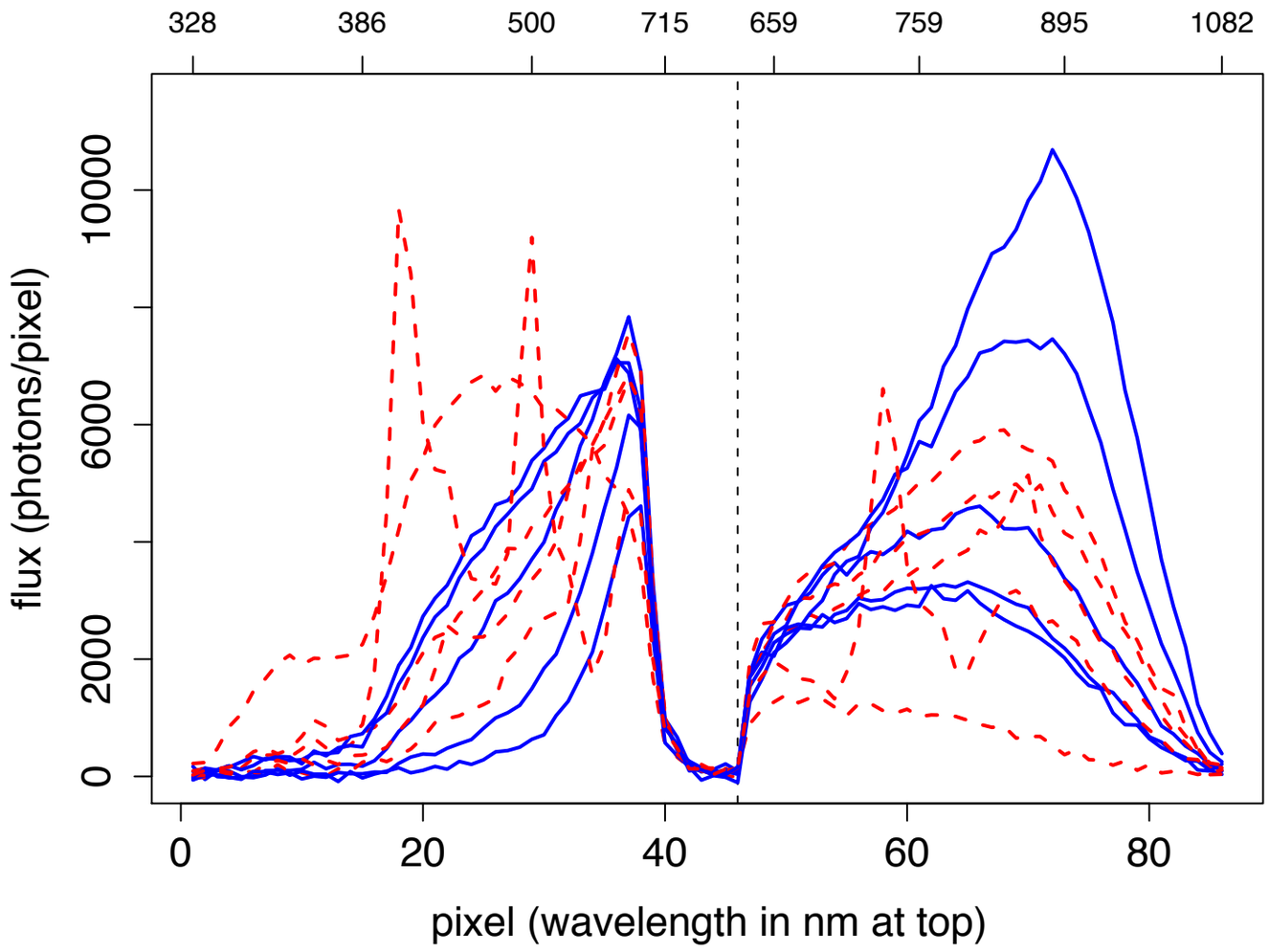


connectivity simplified

Input spectra



Gaia BP/RP spectra



blue = stars

red / dashed = quasars

Discrete Source Classifier (DSC)

- Probabilistic classifier, combining results from three subclassifiers
  - ★ Photometric (BP/RP) [support vector machine]
  - ★ Position–G magnitude [kernel density estimator]
  - ★ Astrometric (parallax and proper motion) [mixture model]
- Current classes:
  - ★ star, binary star, white dwarf, quasar, galaxy, unknown
- Refs: Smith et al. 2011 KS-019; CBJ et al. 2008, MNRAS 391, 1838

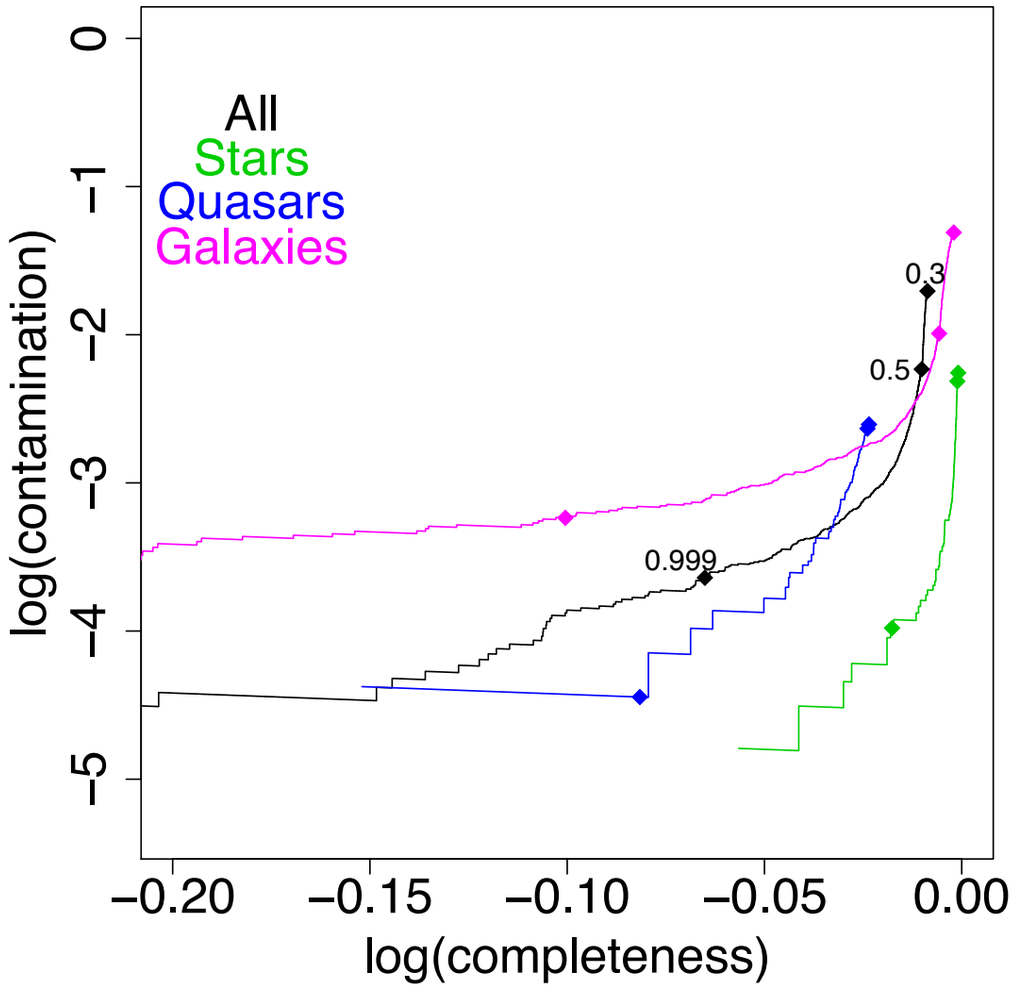
DSC confusion matrix (as percentages)

population spans G=6 to 20, end-of-mission SNR

Grid	N	star	white dwarf	binary	quasar	galaxy	unknown
APec	252	**96.03**			3.97	0.00	0.00
Fastrot	288	**98.26**			0.00	0.00	1.74
Phoenix N	45 610	**95.73**			0.02	0.54	3.70
Phoenix R	10 000	**99.04**			0.02	0.34	0.60
Stars SDSS	50 000	**99.89**			0.02	0.08	0.00
UCD Cond N	126	**29.37**			0.00	0.00	70.64
UCD Cond R	10 000	**78.06**			0.00	0.00	21.94
UCD Dust N	62	**74.19**			0.00	0.00	25.81
UCD Dust R	1 000	**98.60**			0.00	0.00	1.40
WR	43	**76.74**			0.00	0.00	23.26
Quasar SDSS	70 556	0.31			**94.74**	0.92	4.03
Galaxy SDSS	33 670	0.25			0.21	**98.73**	0.81

pass to Outlier Analysis package for further treatment

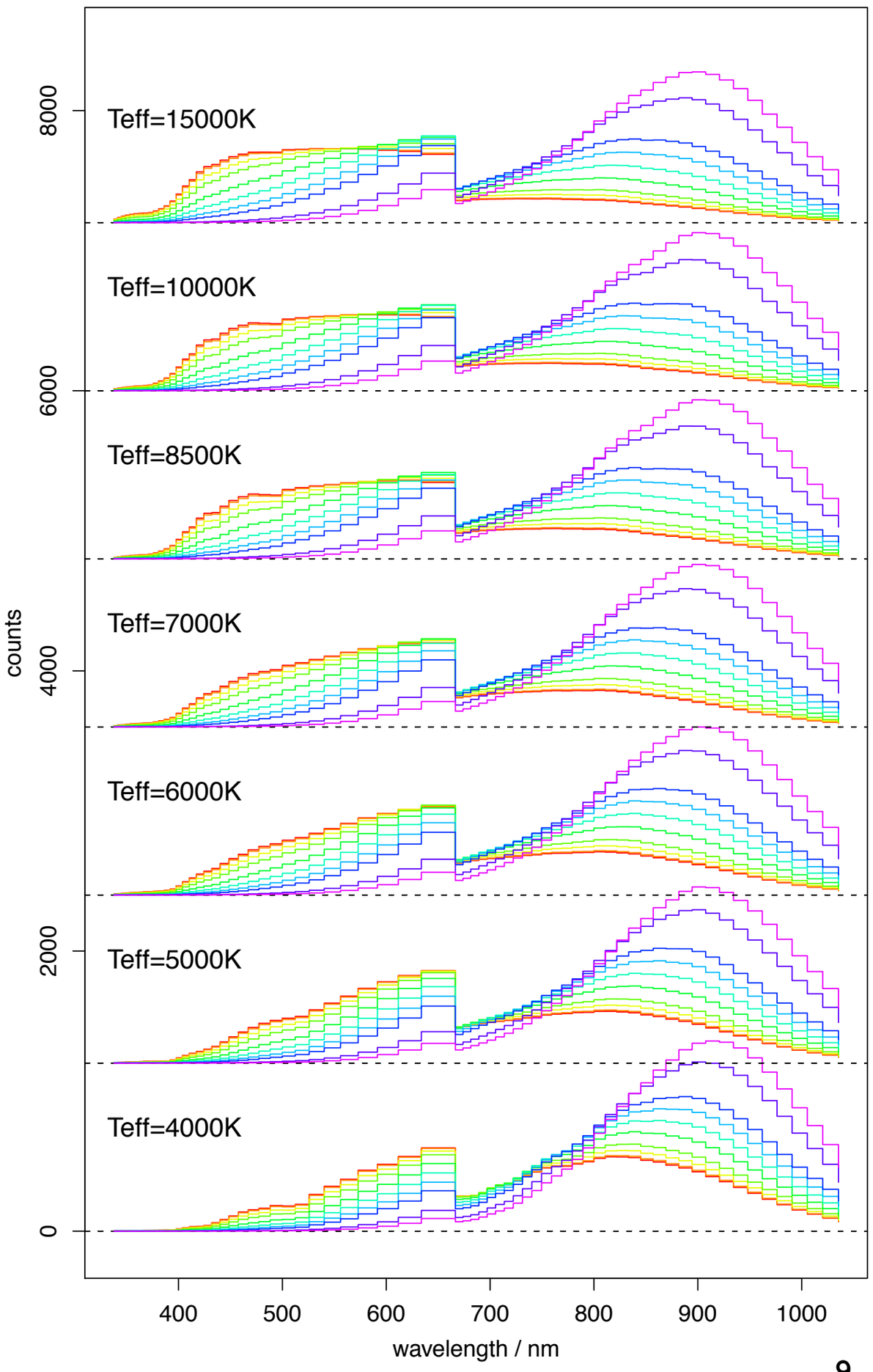
DSC: building samples using probabilities



trade off sample completeness with contamination by varying probability threshold

plot: equal class fractions in training and testing sets

- Quasar selection using method of modified priors to improve purity
  - ★ contamination <0.0025% with 65% (50%) completeness at G=18.5 (G=20)
  - ★ star sample 99% complete with 0.7% contamination

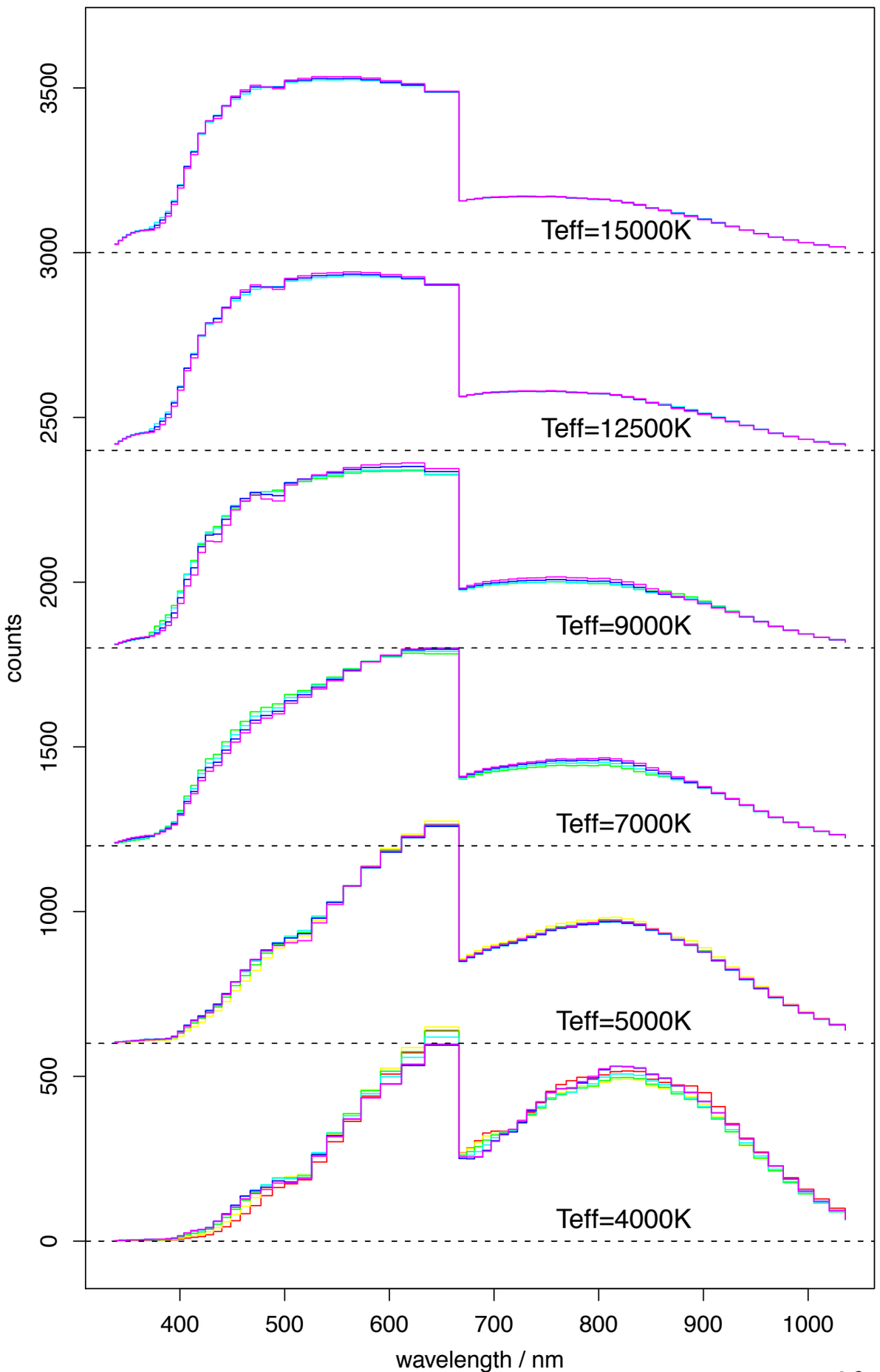


Impact of stellar parameters on BP/RP

T_{eff} and A_0 variation

A_0 = 0, 0.1, 0.5, 1, 2, 3, 4, 5, 8, 10

T_{eff} and A_0: "strong" APs



Impact of stellar parameters on BP/RP

T_{eff} and logg variation

logg = -0.5, 0.5, 2, 3, 4, 5
(A_0 = 0, [Fe/H] = 0)

logg (and [Fe/H]): “weak” APs

General Stellar Parametrizer (GSP-Phot)

- Three independent modules working in parallel
  - ★ Support Vector Machine (BP/RP only)
  - ★ ILIUM (BP/RP only). Also predicts covariance on APs
  - ★ q-method (BP/RP, parallax, apparent mag.). Gives a full PDF over APs
- Then infer bulk properties (mass, radius, luminosity, age) from atmospheric parameters (FLAME package)
- Refs: Liu 2011, CHL-005; CBJ 2010, MNRAS 403, 96; CBJ 2011, MNRAS 411, 425

GSP-Phot (SVM) performance

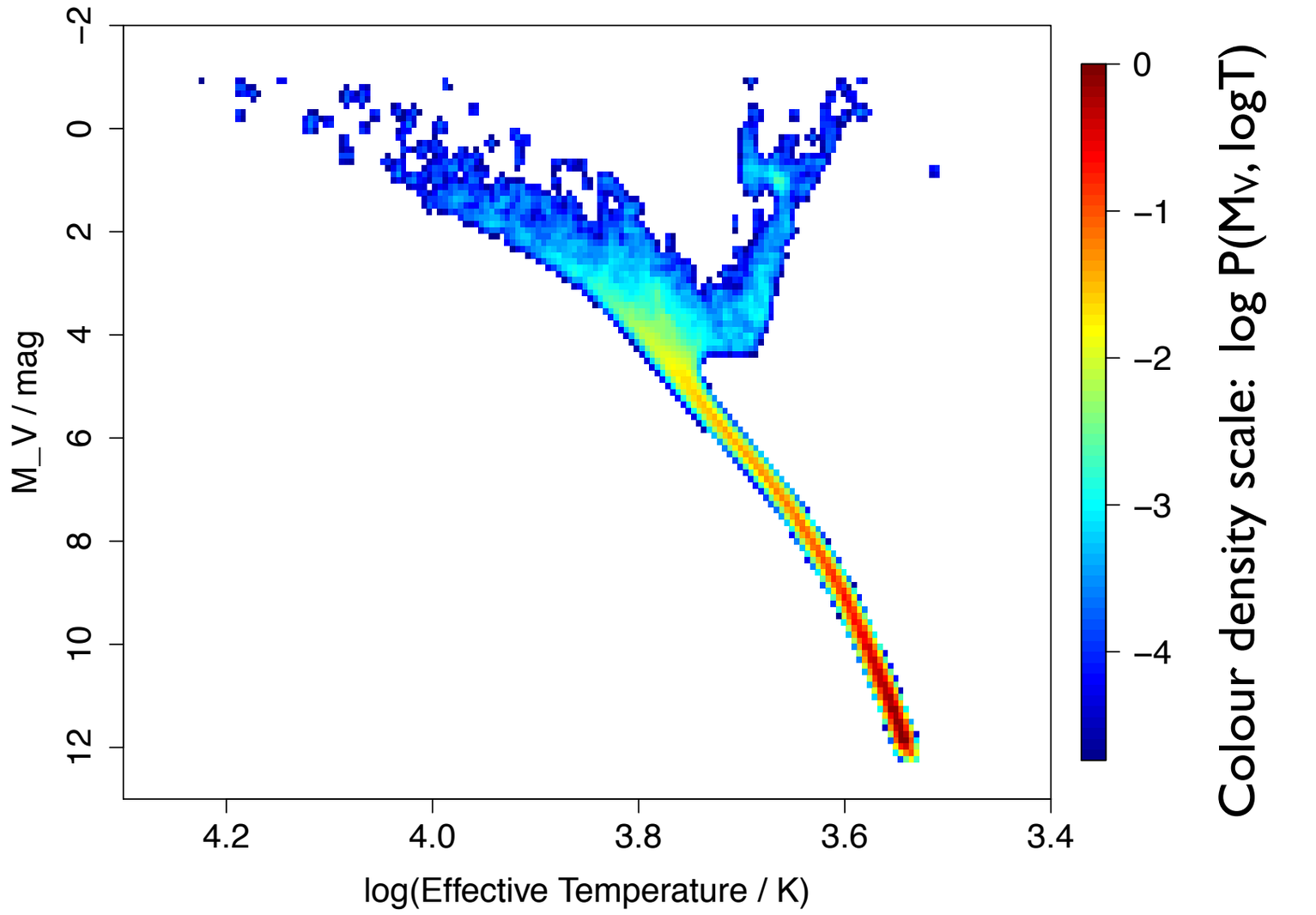
AP residual	All stars	A stars	F stars	G stars	K stars
G<16.5 mag					
$<\|dT_{\rm eff}\|>$(K)	160	564	253	89	84
$<\|dA_0\|>$(mag)	0.14	0.30	0.19	0.08	0.11
$<\|d[{\rm Fe/H}]\|>$(dex)	0.36	1.24	0.58	0.35	0.21
$<\|d{\rm logg}\|>$(dex)	0.16	0.16	0.20	0.19	0.13
G>16.5 mag					
$<\|dT_{\rm eff}\|>$(K)	364	939	606	297	249
$<\|dA_0\|>$(mag)	0.28	0.39	0.39	0.24	0.25
$<\|d[{\rm Fe/H}]\|>$(dex)	0.61	1.03	0.92	0.57	0.48
$<\|d{\rm logg}\|>$(dex)	0.19	0.30	0.26	0.23	0.13

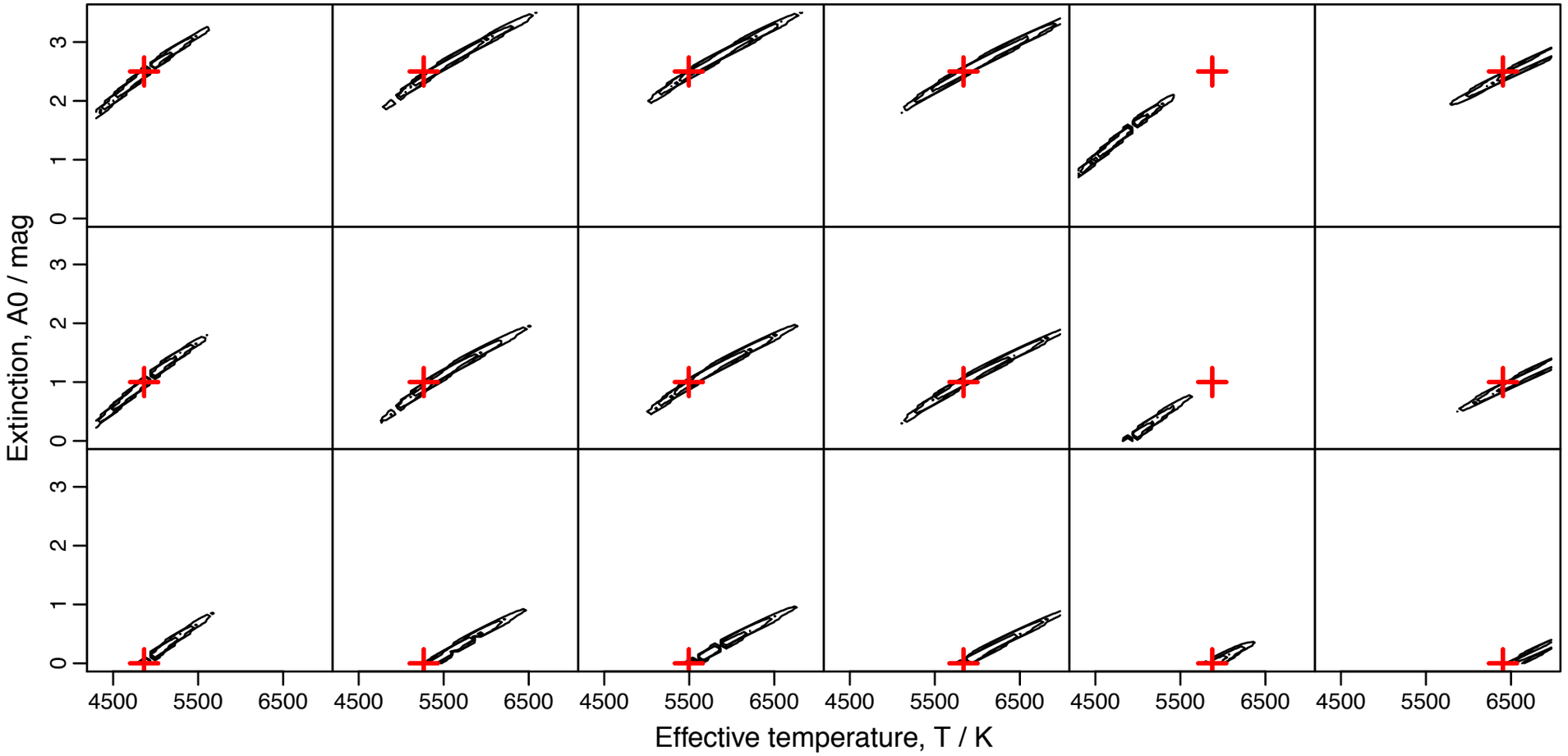
mean absolute residuals
averaged over a very broad stellar parameter space (esp. 0-10 mag. extinction)

GSP-Phot: Bayesian inference

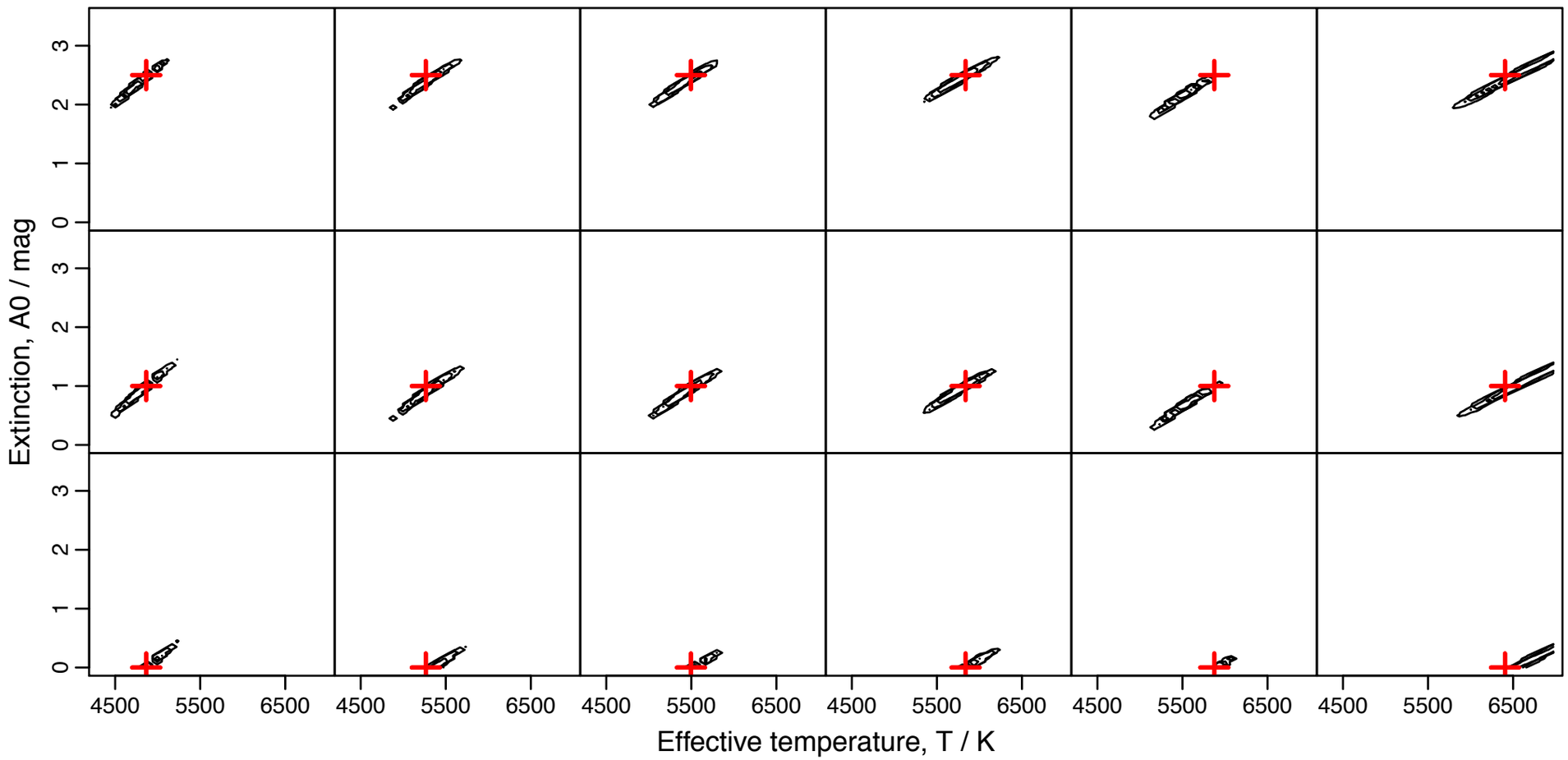
- Spectrum: p constrains T_{eff} and A_V — $P(p \mid A_V, T_{eff})$
- Parallax, apparent mag.: q constrains M_V + A_V — $P(q \mid M_V, A_V, T_{eff})$
- HRD ("prior") constrains M_V and T_{eff} — $P(M_V, T_{eff})$

$$q \equiv V + 5\log\varpi$$
$$= M_V + A_V - 5$$





BVJHK
only

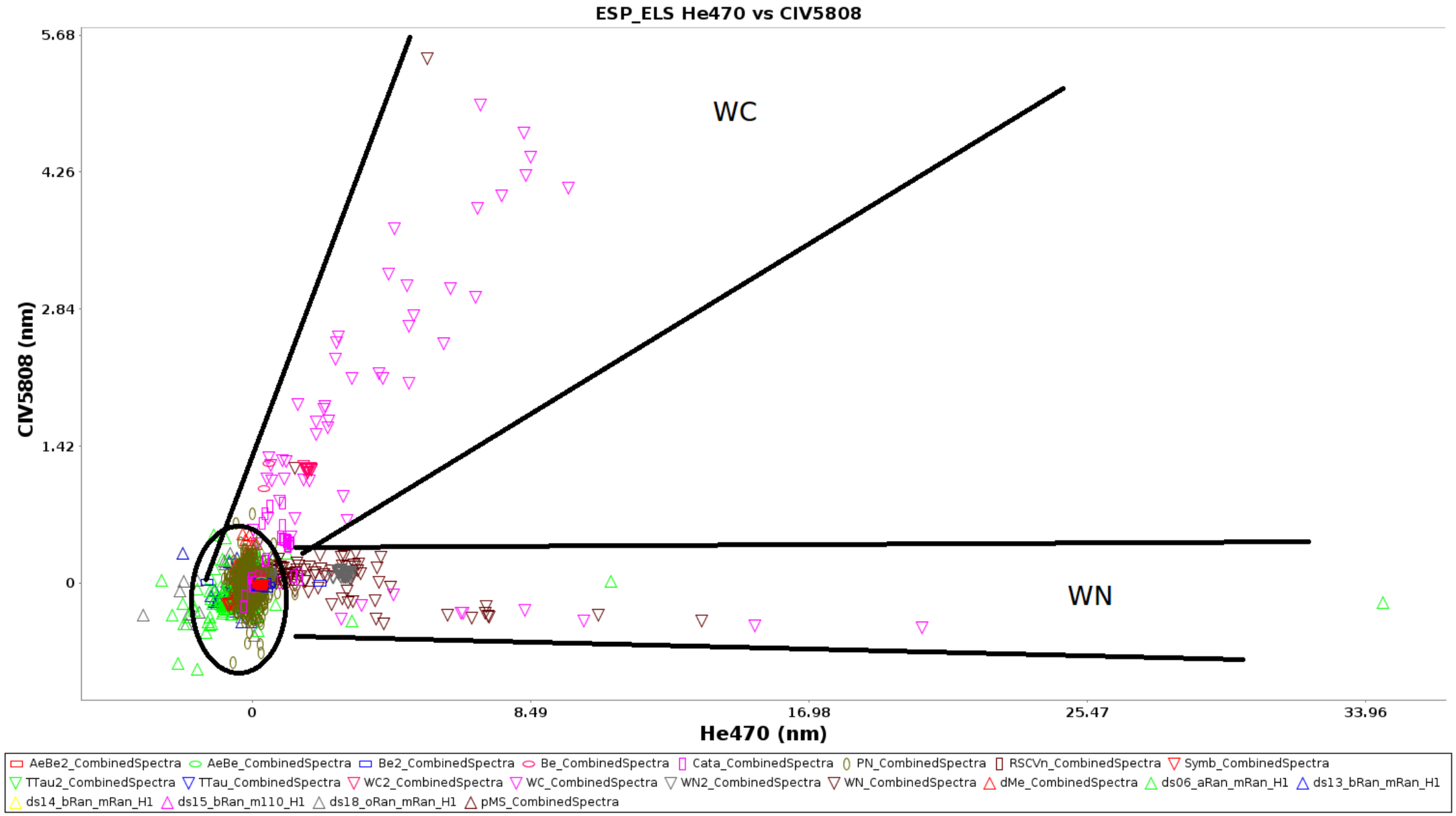


BVJHK
+ q, HRD

Extended Stellar Parametrizer (ESP)

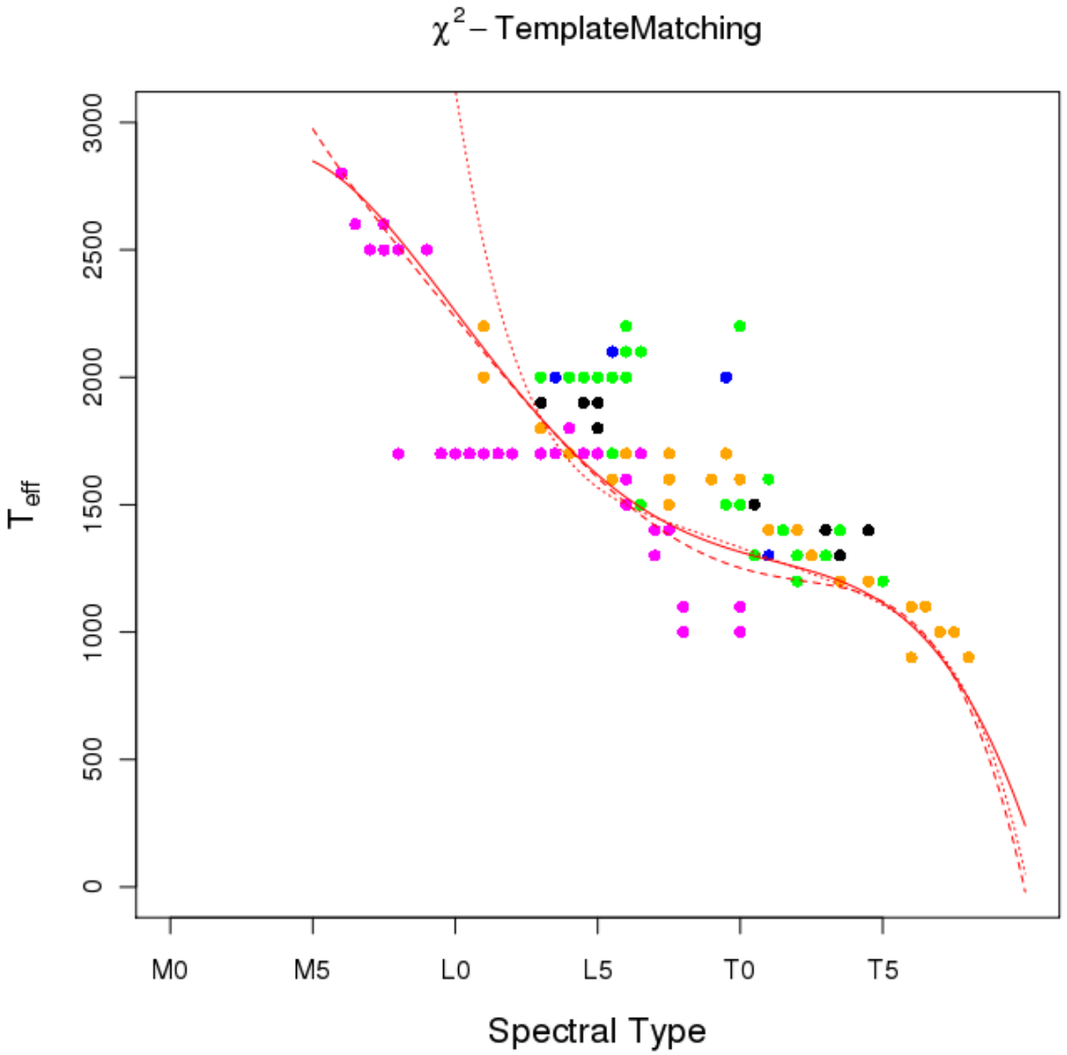
- Four modules to deal with "extreme" types of stars
  - ★ emission line stars
  - ★ OB stars
  - ★ chromospherically active stars and M stars
  - ★ ultra cool dwarfs
- Improved treatment over GSP-Phot and GSP-Spec
- Combining BP/RP, RVS and parallax
- Refs: SDD CHL-004-4; Fremat et al. 2011 YF-005

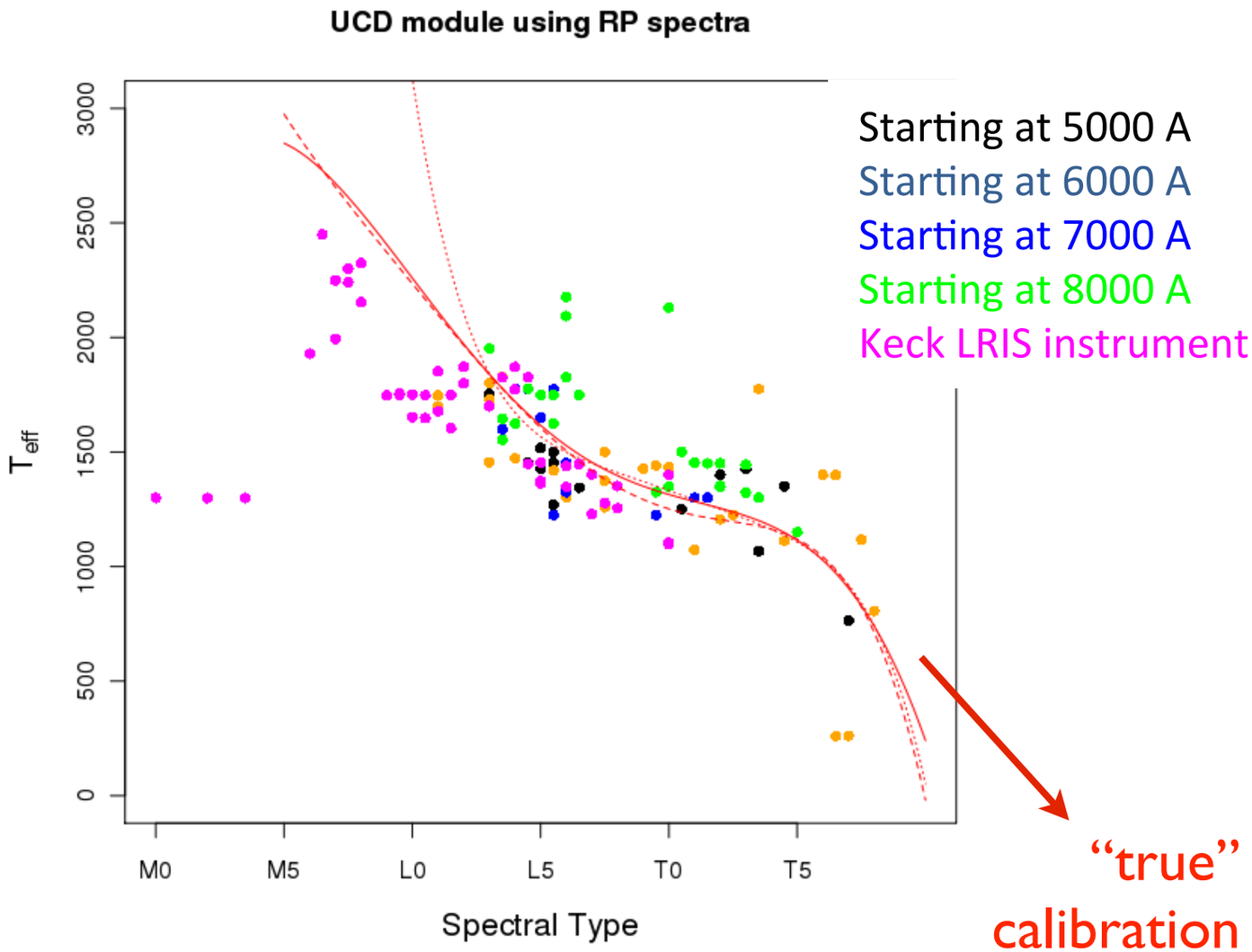
ESP on Emission Line Stars



- Selection of specific features in RVS
  - ★ here C IV and He II for WC and WN Wolf Rayet stars

ESP on Ultra Cool Dwarfs

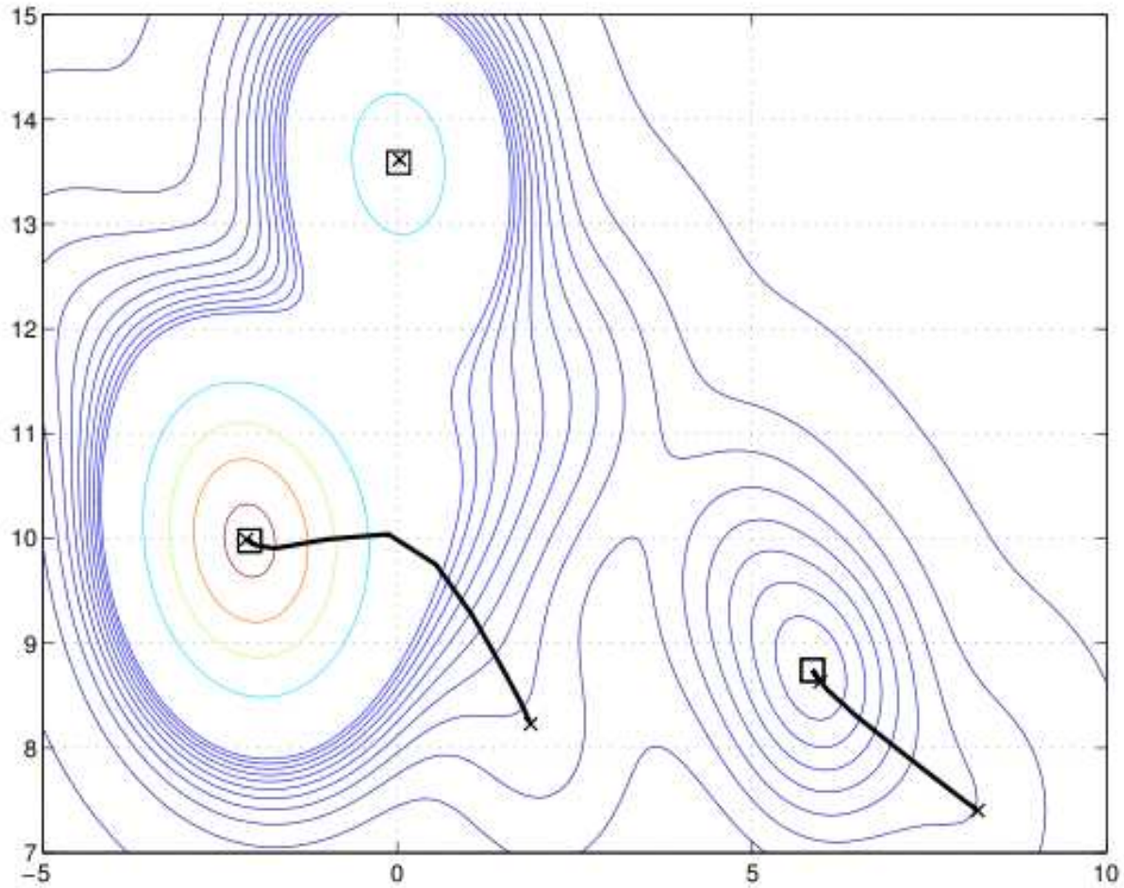




- Inferring Teff for stars cooler than 3000K (various techniques)
  - ★ Right: RP simulations of real spectra/truncated spectra (used on left)

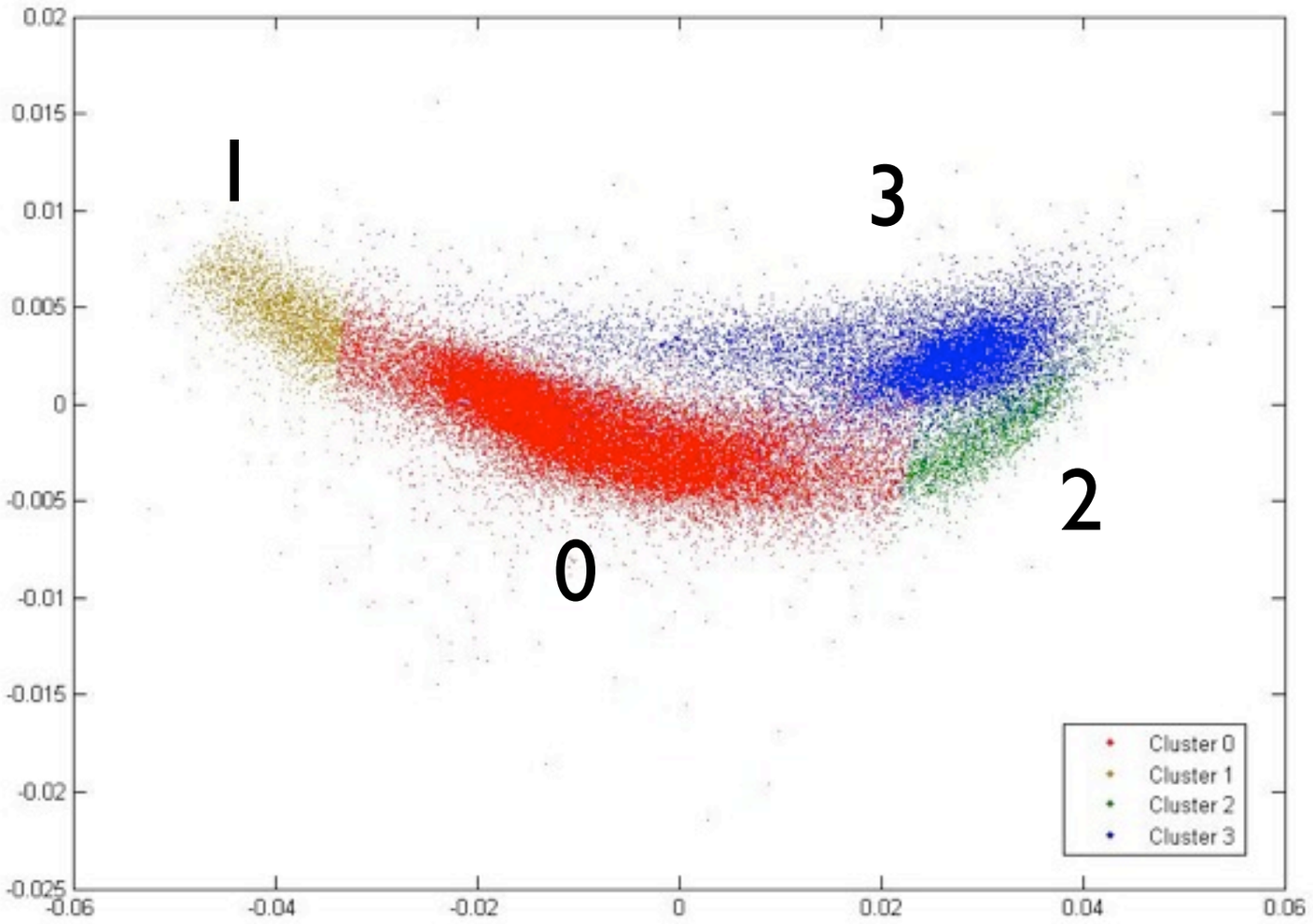
Object Clustering Analysis (OCA)

- Supervised analysis
  - ★ correct information only on known (= modelled) objects
- Unsupervised analysis
  - ★ discover "natural" classes/ parameters by clustering, variance analysis etc.
- Clustering algorithms
  - ★ K-means, mode-association clustering (MAC)
- References: SDD; `OCA Gaia Wiki page` (Garcia Torres, Sarro)



The MAC algorithm uses EM to ascend modes

Object Clustering Analysis (OCA)



- 10^5 stars and galaxies from BP/RP semi-empirical SDSS libraries
- Dimensionality reduction to the first four statistical moments
- Six clusters found (two virtually empty)

cluster id	--	#stars	--	#galaxies	--	#sources
0	--	33656	--	18	--	33674
1	--	2496	--	1	--	2497
2	--	2397	--	0	--	2397
3	--	10	--	13306	--	13316

CU8 operations phase

- Algorithm/software development continues throughout mission
  - ★ accuracy of AP estimation depends critically on quality/nature of data
  - ★ learn about calibration
  - ★ iterative improvement of libraries/training sets (e.g. DSC-OA link)
- Algorithms retrained according to latest spectral libraries
  - ★ maintenance of a simulator until final catalogue delivery is essential

Where do we go from here?

- "Robustification"

  ... of algorithms to imperfect data

- Calibration

  ... of algorithms to accommodate the synthetic/real data mismatch

  ... use of on-sky "labelled" objects (from GBOG, Gaia-ESO, ...)

- Integration testing

  ... of Apsis, and of Apsis with rest of DPAC (e2eS3)

Where do we go from here?

- Develop analysis packages

  ... for automated and manual validation and verification of results

- Improve algorithm performance

  ... in terms of estimation precision/accuracy

  ... in terms of computational resources (mostly run times)

- Extend algorithm capabilities

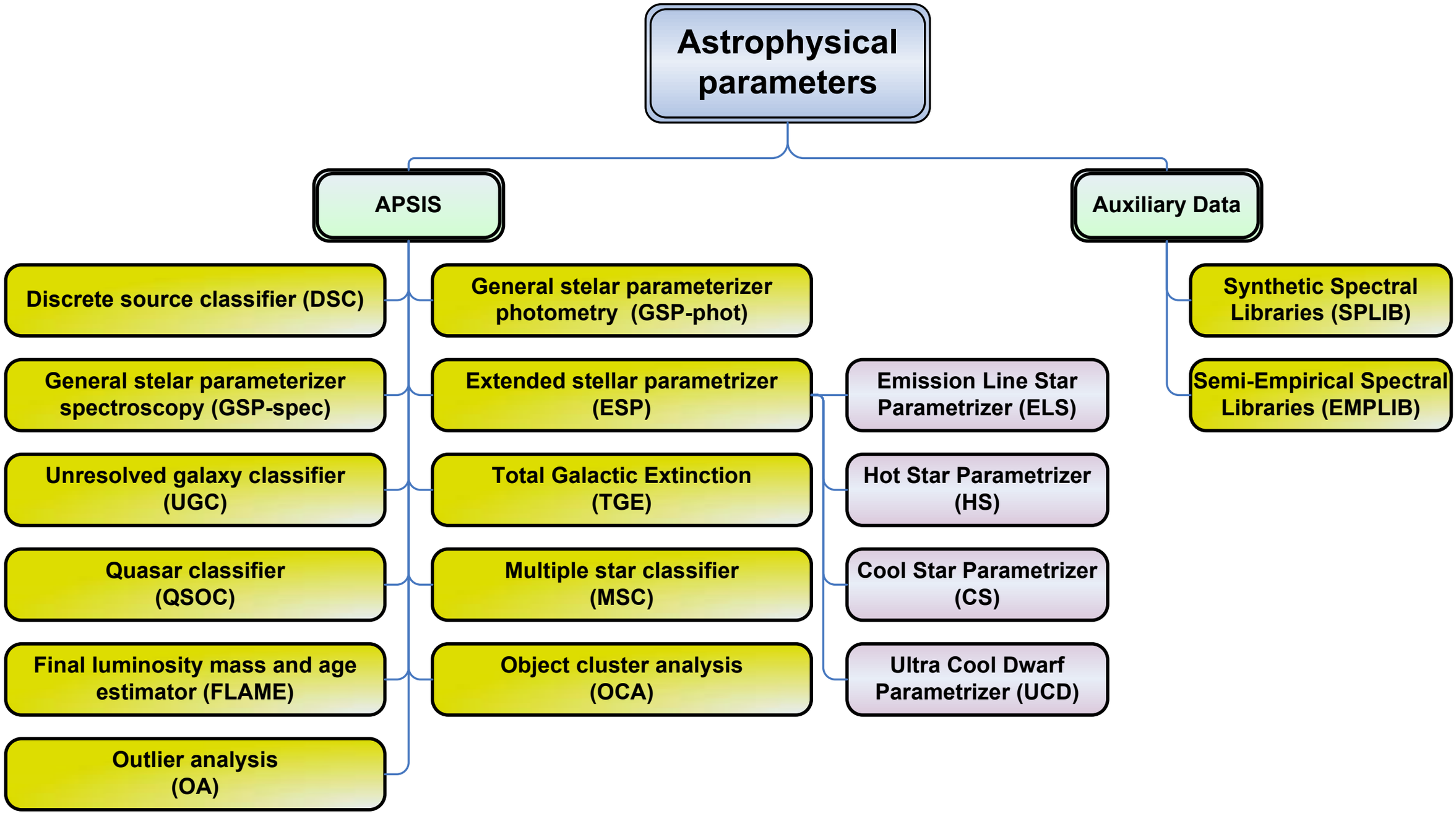
  ... e.g. include variability information in DSC

Summary

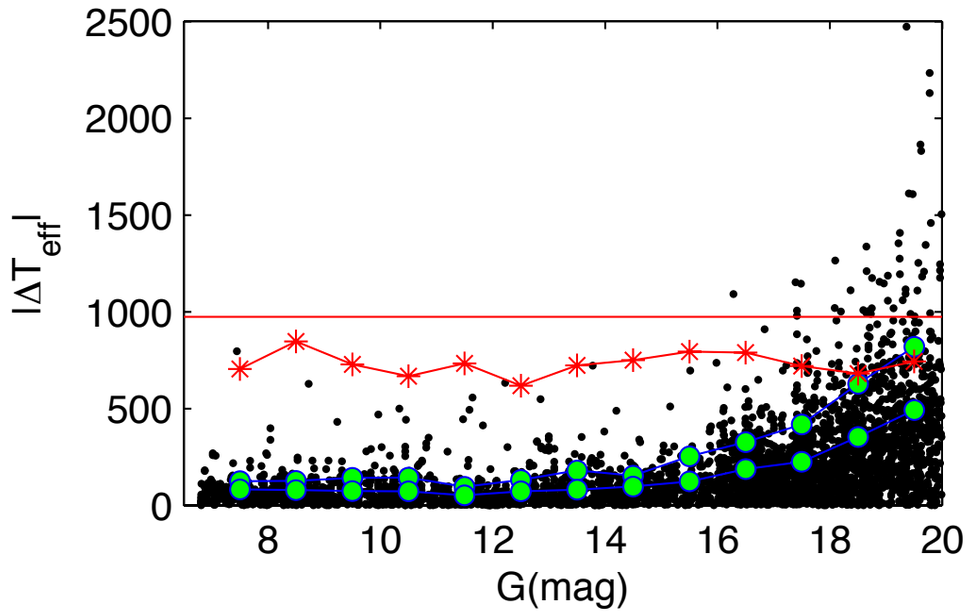
- classification and parameter estimation of stars, galaxies and quasars using BP/RP, G, astrometry and RVS
  - ★ physical parameters unavoidably tied to physical models
- mixture of general purpose and specific algorithms
- give probabilities or probability density functions where possible
  - ★ permits user-defined sample selection
  - ★ characterizes uncertainties and degeneracies
- potentially multiple, model-dependent parameters per source

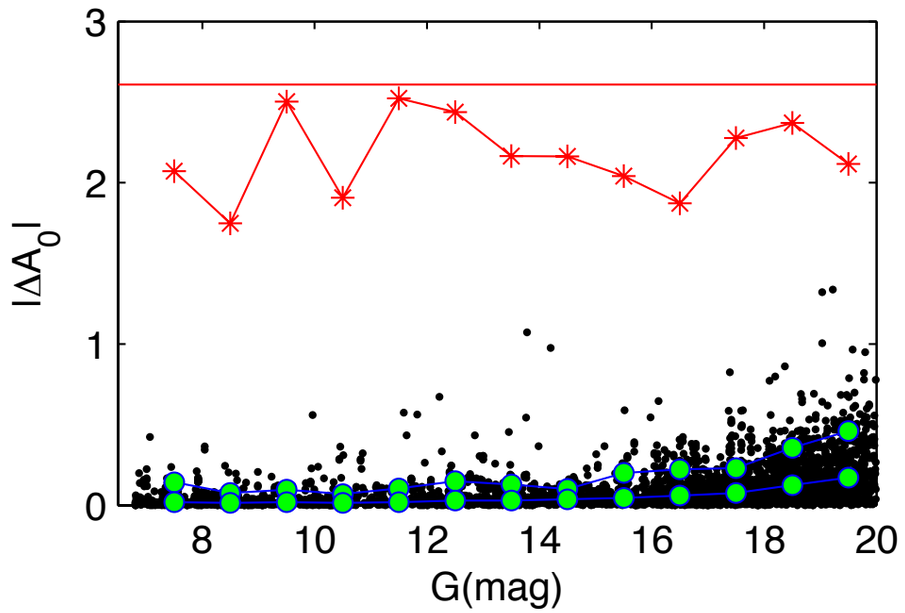
Extra information

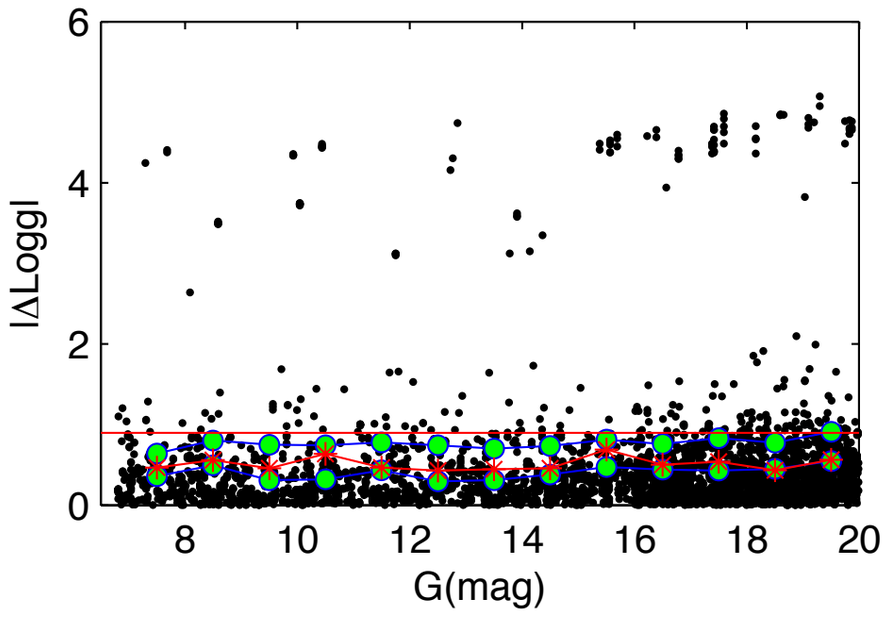
CU8 products

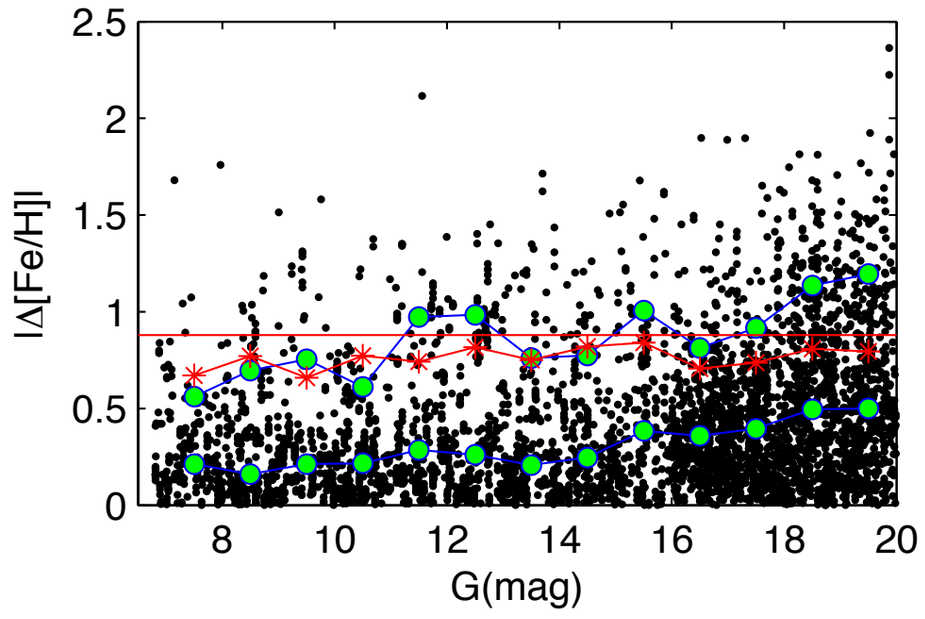


Performance for GSP-Phot (SVM)



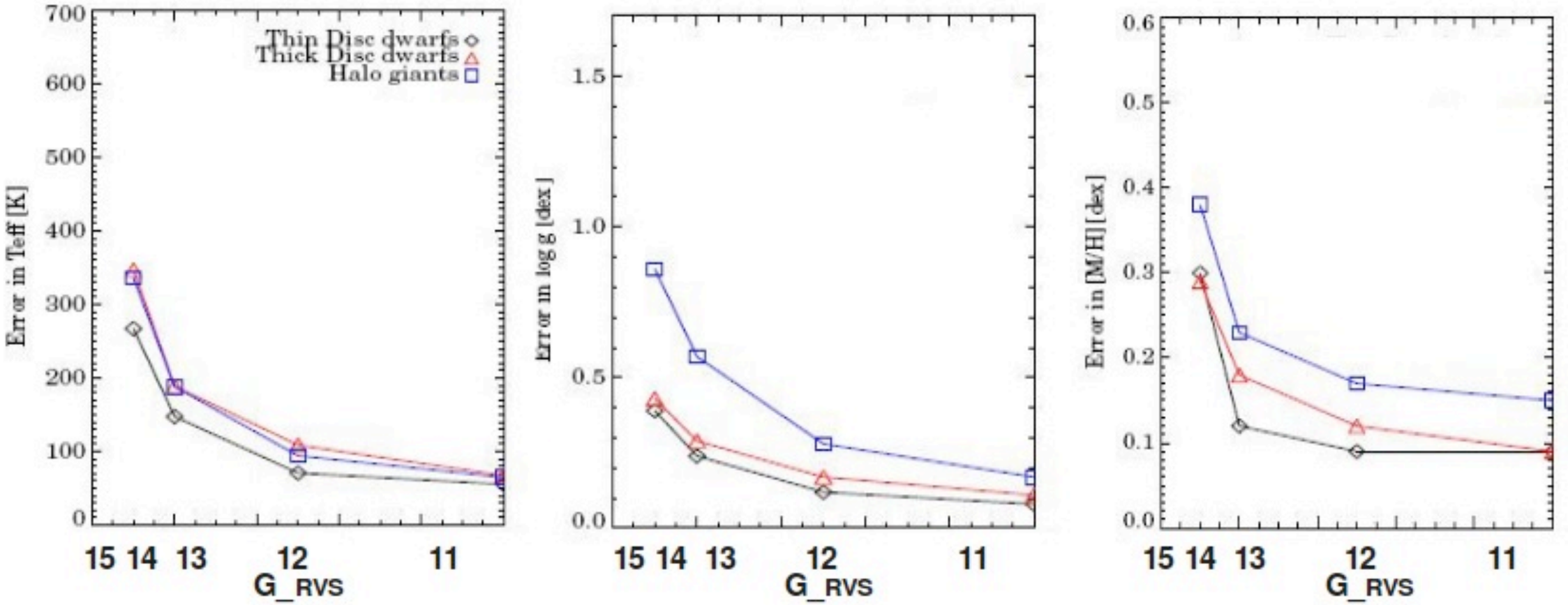






Black points: mean absolute residuals for individual stars Green points: 50% and 90% quantiles
Red points: mean absolute residuals for a random classifier

Stellar parametrization with RVS: GSP-Spec

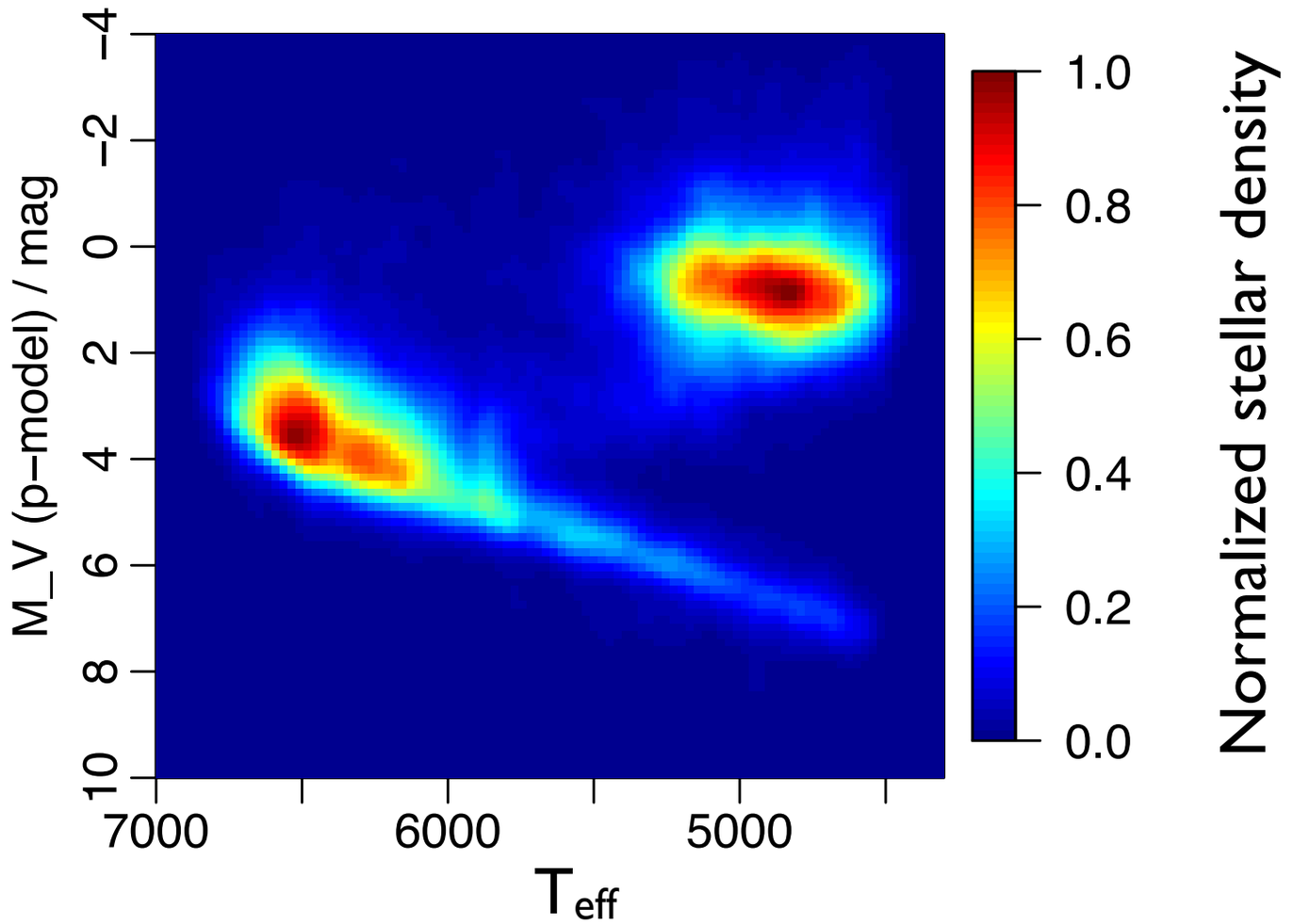


Average residuals from simulated RVS spectra for stars selected from specific stellar populations (SNR for end of mission)

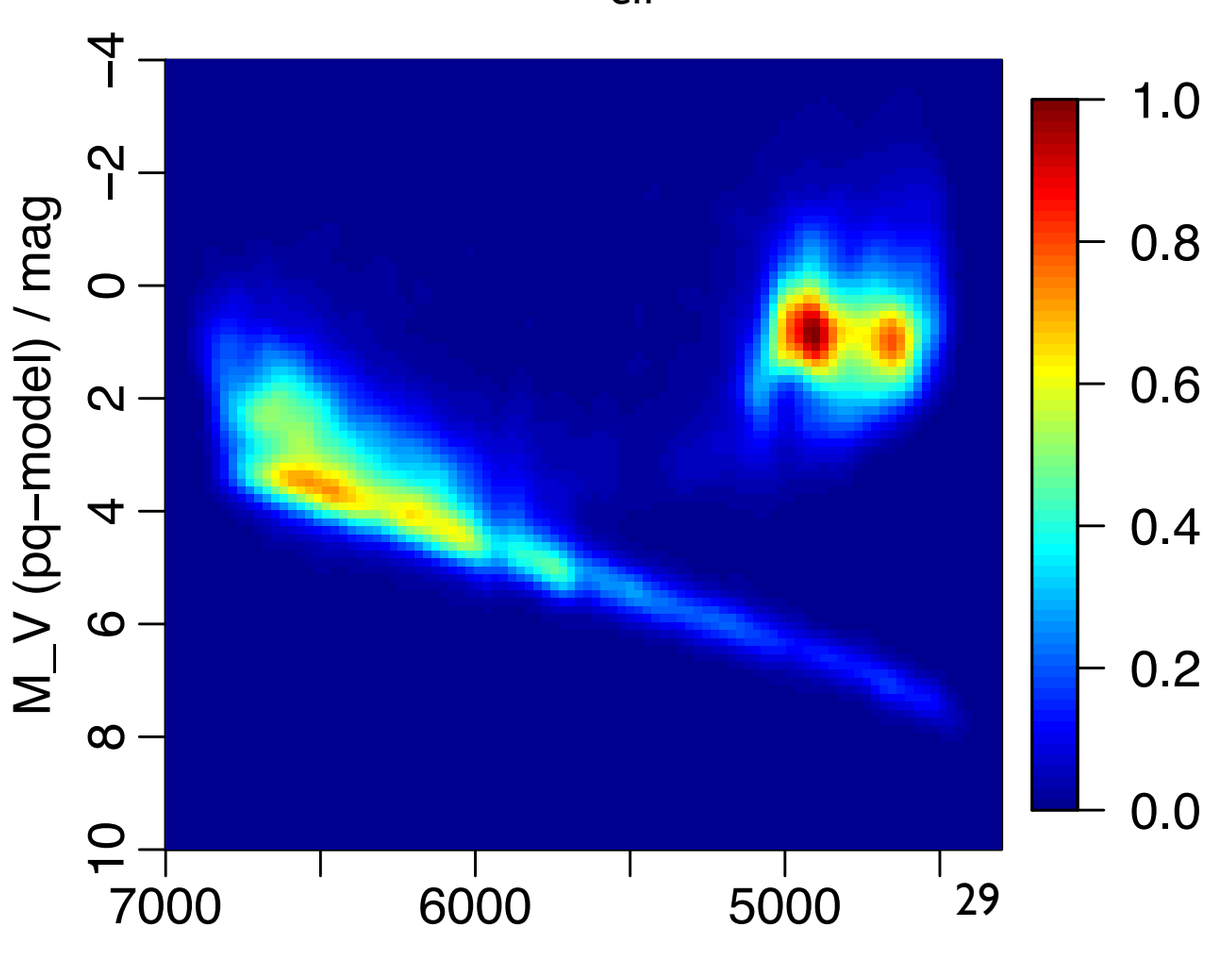
Algorithm speed

Algorithm	MegaFLOP counts (1 source) mean (min:max)	Duration in sec @1GFlop/sec (1 source) mean (min:max)
DSC	64 (20:130)	0.064 (0.02:0.13)
GSP_PHOT SVM	20	0.020
GSP_PHOT ILIUM	1300	1.3
GSP_PHOT Q-METHOD	12250 (4400:17000)	12.25 (4.4:17)
GSP_SPEC	7 (5:40)	0.007 (0.005:0.04)
ESP ELS	25 (10:440)	0.025 (0.01:0.4)
ESP HS (BPRP)	340 (10:740)	0.34 (0.01:0.74)
ESP HS (BPRP+RVS)	8134 (2200:18100)	8.134 (2.2:18)
ESP CS	80	0.08
ESP UCD	8	0.008
UGC	50 (25:115)	0.05 (0.025:0.115)
MSC	150	0.15
QSOC	3271	3.271
TGE	62 (20:100)	0.062 (0.02:0.1)

From the ApsisOpsDv1 pre-integration campaign (see Wiki)



BVJHK
only



BVJHK
+ q, HRD

Outlier Analysis (OA)

- **Unsupervised analysis of "rare" objects. Those not classified by both OCA and DSC**
- **1º Stage: Clustering using several data representations.**
- **2º Stage: Cluster labelling**
- **3º Stage: Feedback to DSC**