Unresolved Galaxy Classification and Parameterization for Gaia

(DPAC/CU8 GWP832-UGC)

Ioannis Bellas-Velidis [1]
Mary Kontizas [2]
Anastasios Dapergolas [1]
Petros Drazinos[2]
Evdokia Livanou [2]
Evangelos Kontizas [1]
Antonios Karampelas [2]

[1] *National Observatory Athens*
[2] *National Kapodistrian University Athens*

"UGC for Gaia", EGSG 2012 Workshop, OCA Nice, 10-11 September, 2012

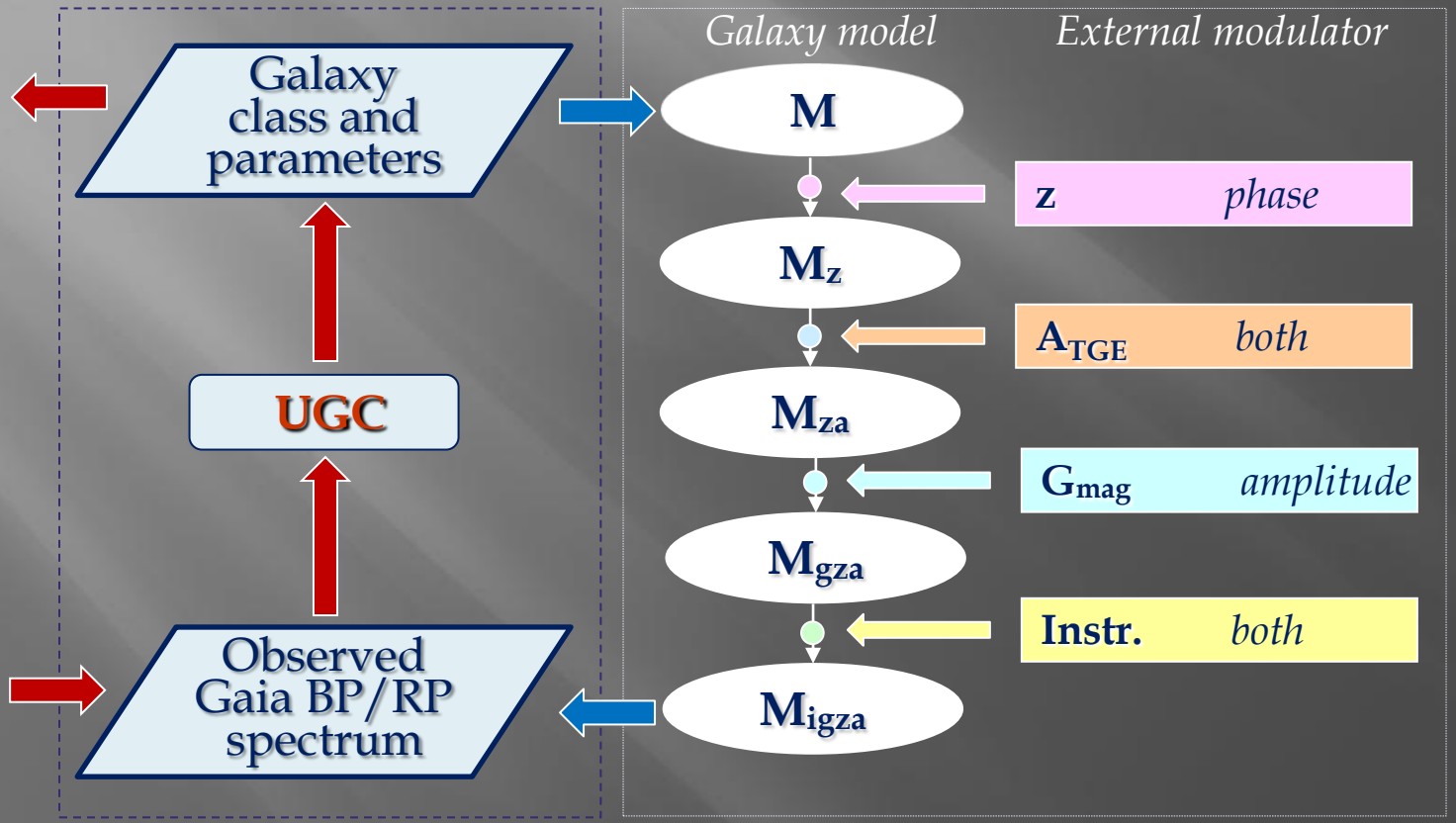
Contents

- UGC for Gaia - Overview
- Support Vector Machines (SVM) tests
- UGC Performance
- Artificial Neural Networks (ANN) tests

UGC: CU8/GWP-832 Overview

Objectives: unresolved galaxies classification and astrophysical parameters estimation .

Tasks: study, develop and test an algorithm, assuming that the observed objects are restricted to this type. Implement as software module of DPAC's scientific pipe-line.



Method:
an AI technique based on supervised learning via prototypes.

Prototypes:
libraries of galaxy spectra simulated for Gaia BP/RP.

UGC Overview: Synthetic Spectra

Pegase.2 galaxy spectra libraries:

Galaxy taxonomic class *galType*

- *Early*
- *Spiral*
- *Irregular*
- *Quenched StarBurst*

Star Formation parameters *galParams*

- *E* type: *E_p1* , *E_p2*
- *S* type: *S_p1* , *S_p2* , *S_infall*
- *I* type: *I_p1* , *I_p2* , *I_infall*
- *B* type: *B_p1* , *B_p2* , *B_p3* , *B_infall*

Star formation rate laws used:

Early type:

$$SFR(t) = (p2/p1) * e^{(-t/p1)}$$

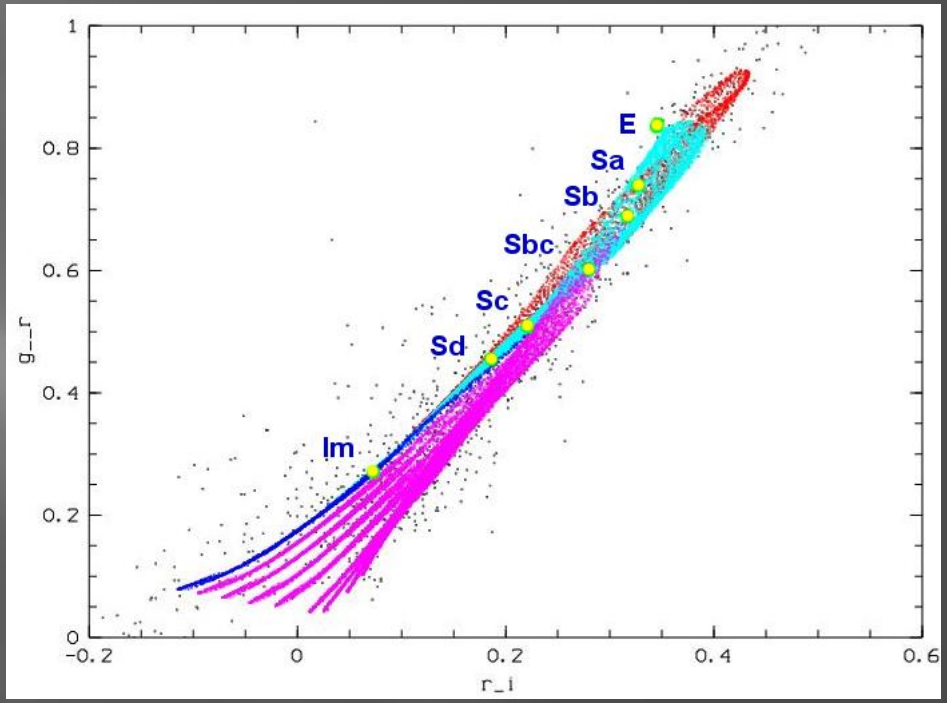
Spiral and Irregular type:

$$SFR(t) = M_{gas}(t)^{p1}/p2$$

Quenched StarBurst type:

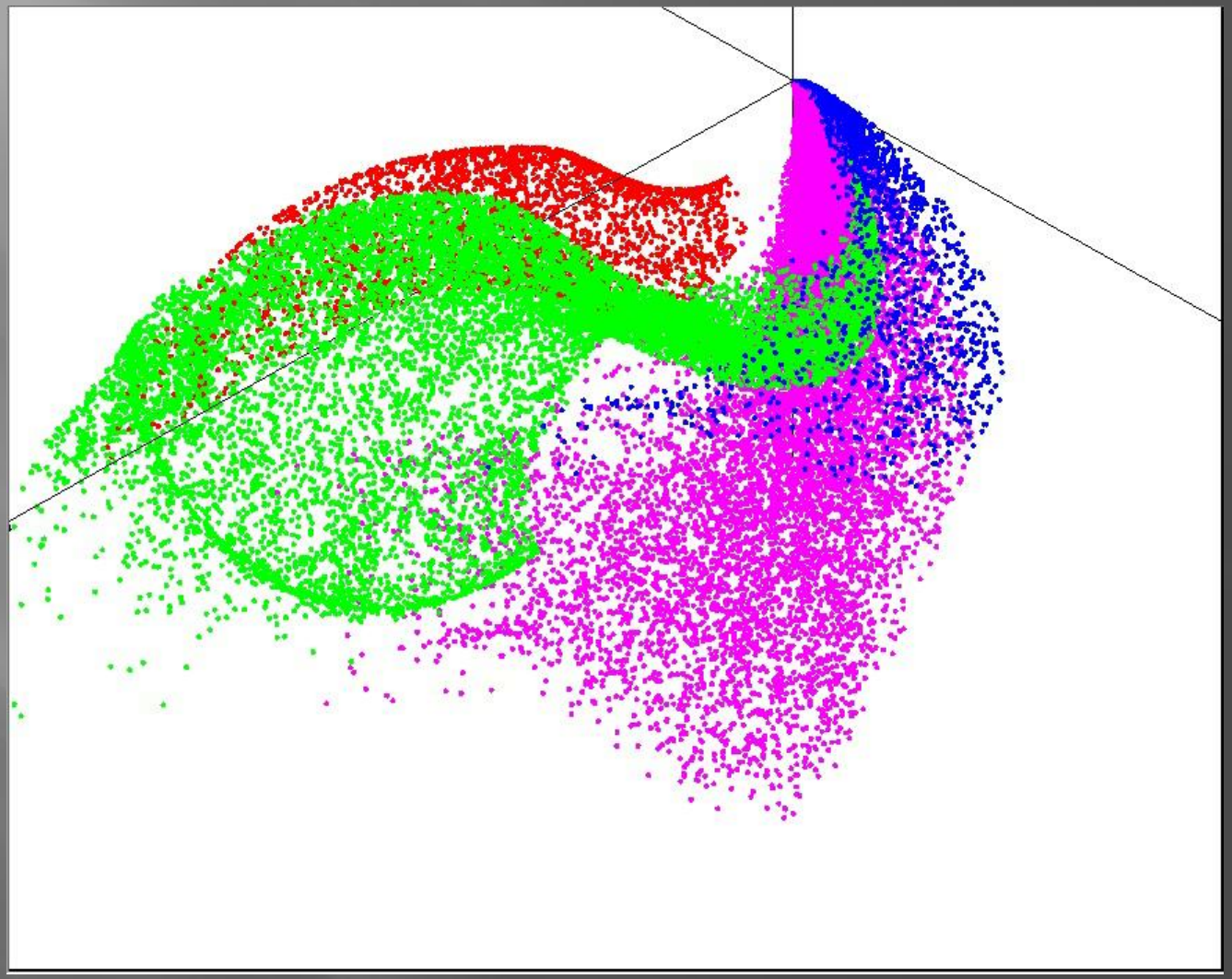
$$SFR(t) = M_{gas}(t)^{p1}/p2 \quad \text{for } t<Age-p3$$

$$SFR(t) = 0 \quad \text{for } t \geq Age-p3$$

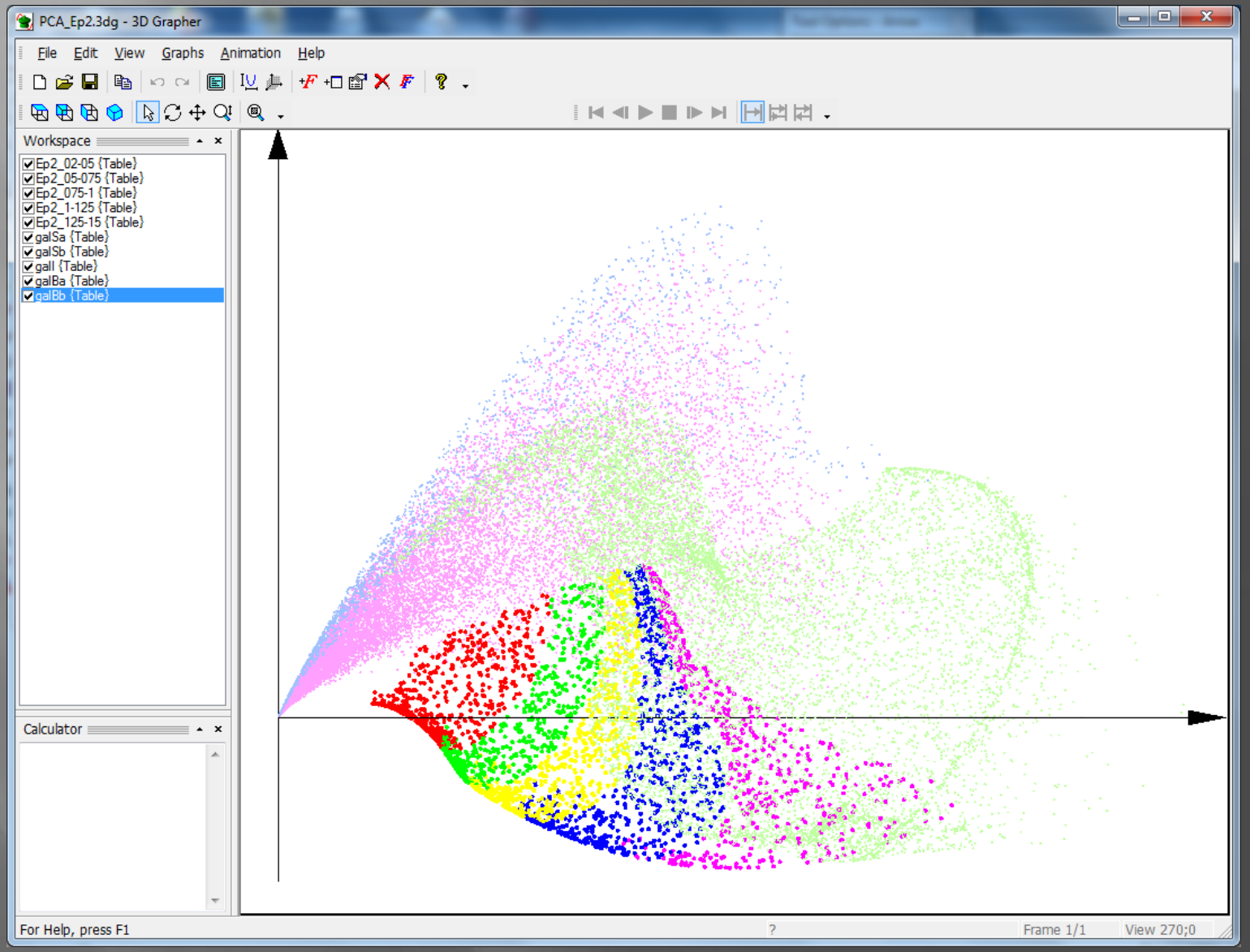


Color-color diagram of the SDSS galaxies with z<0.1 (black) and the Early (red), Spiral (cyan), Irregular (blue), and Quenched StarBurst (magenta) galaxy colors of the second library of synthetic galaxy spectra; yellow dots represent E, Sa, Sb, Sbc, Sc, Sd and Im typical galaxy colors derived from synthetic spectra.

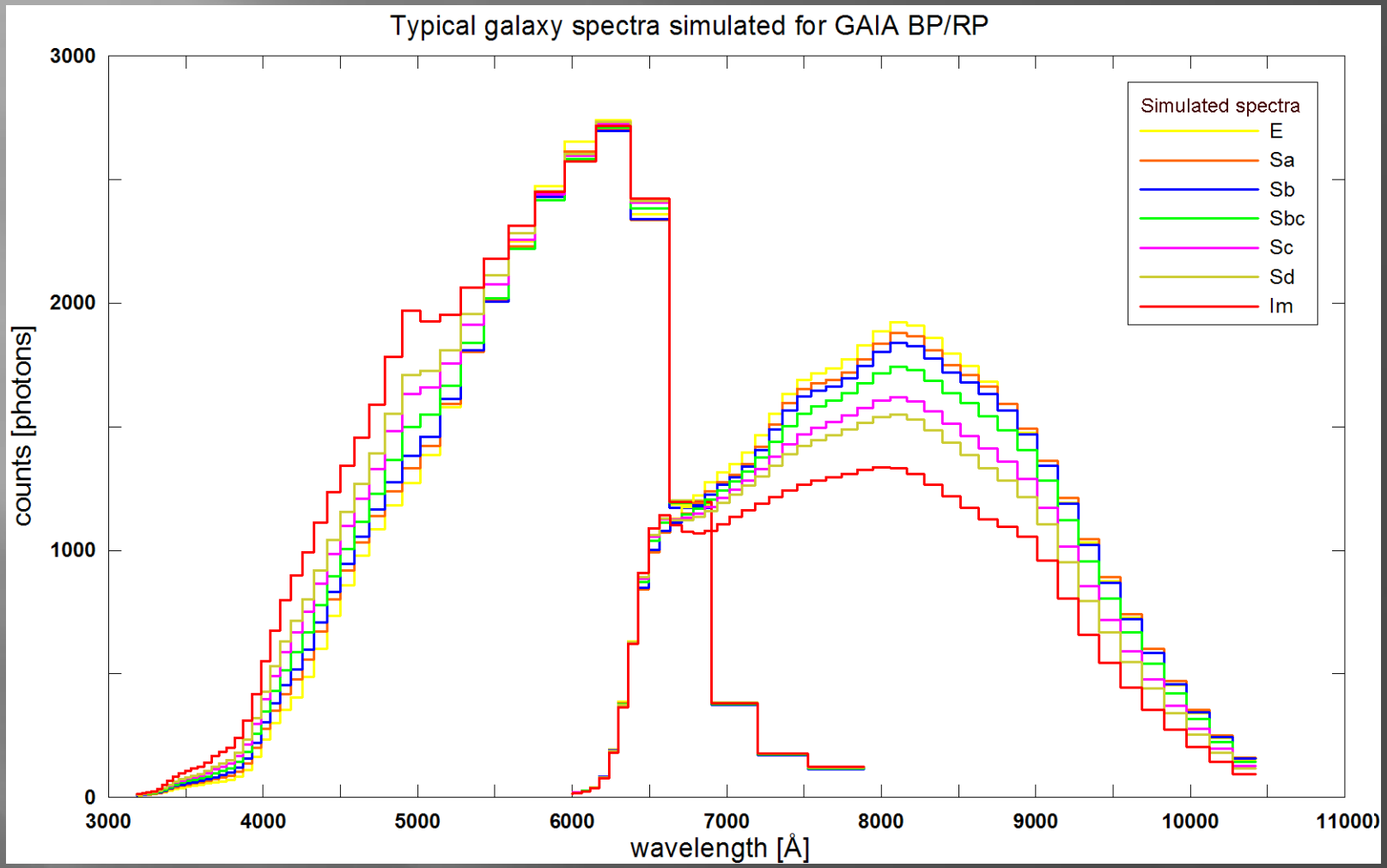
UGC Overview: Synthetic Spectra PCA



UGC Overview: Synthetic Spectra PCA

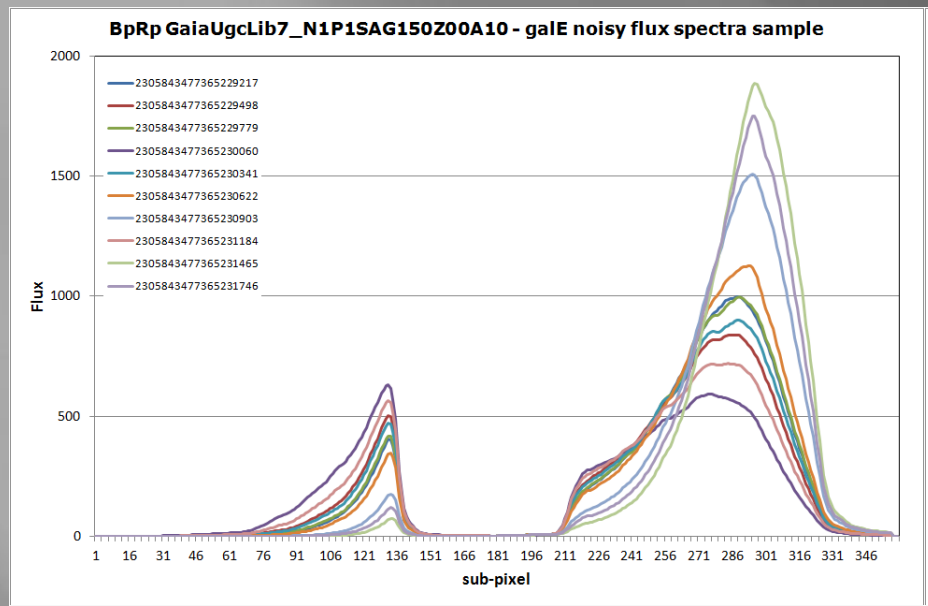


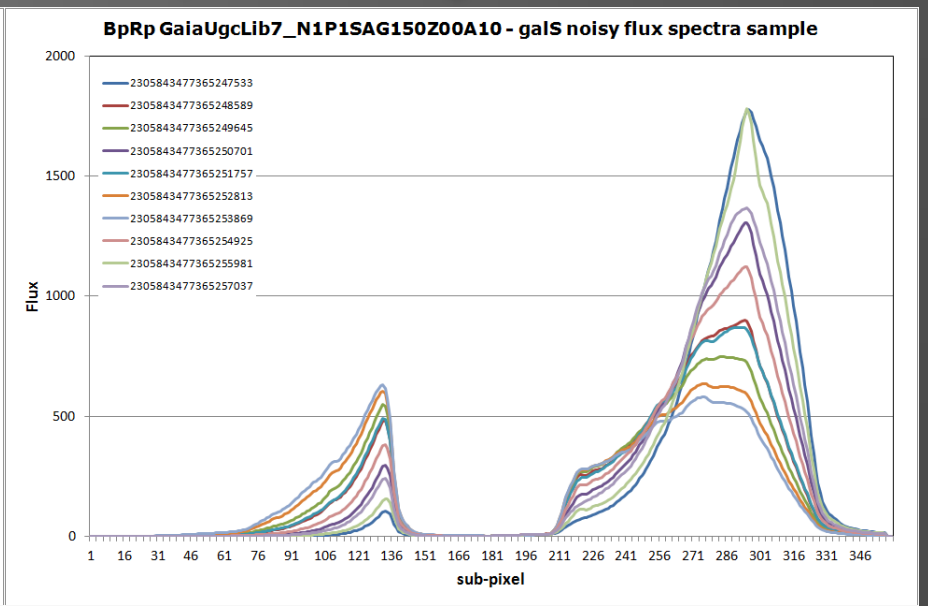
UGC: Simulated BP/RP Spectra

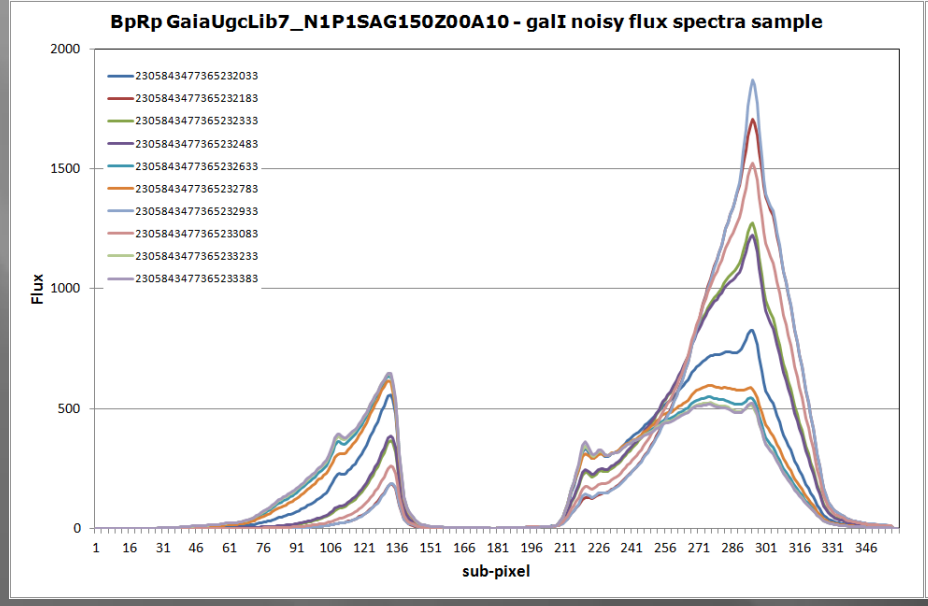


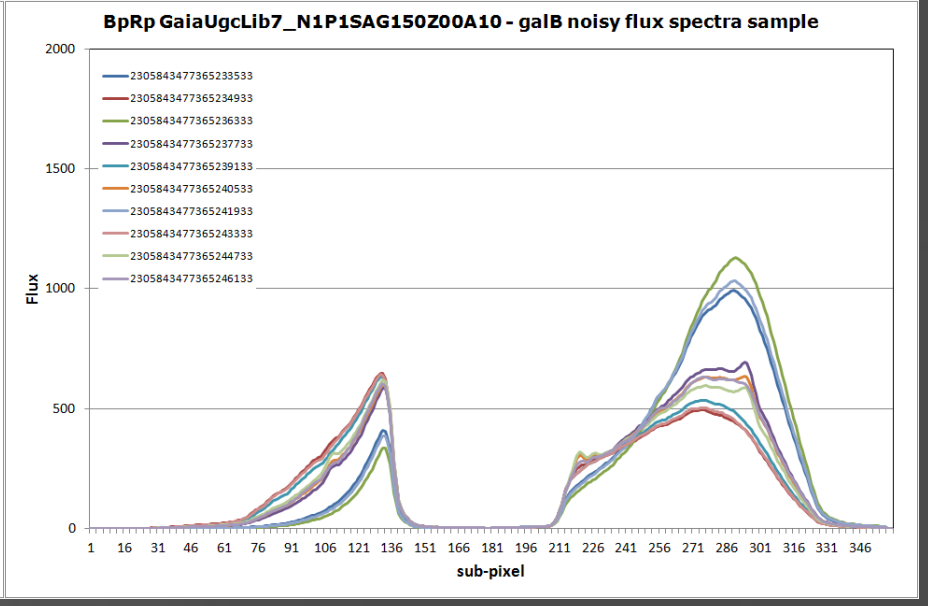
CU8/GWP-832 (UGC) => CU8/GWP-811 => CU2/DU3 => CU8/GWP-811 => CU8/GWP-832

UGC: Simulated BP/RP Spectra









UGC Overview: Software

- **CU8/WP832 Unresolved Galaxy Classifier**

  provide astrophysical information on Gaia observed unresolved galaxies
  - taxonomic classification: *galType* (E, S, I, B)
  - astrophysical parameters estimation: *galParams* (SF, z, A_o)

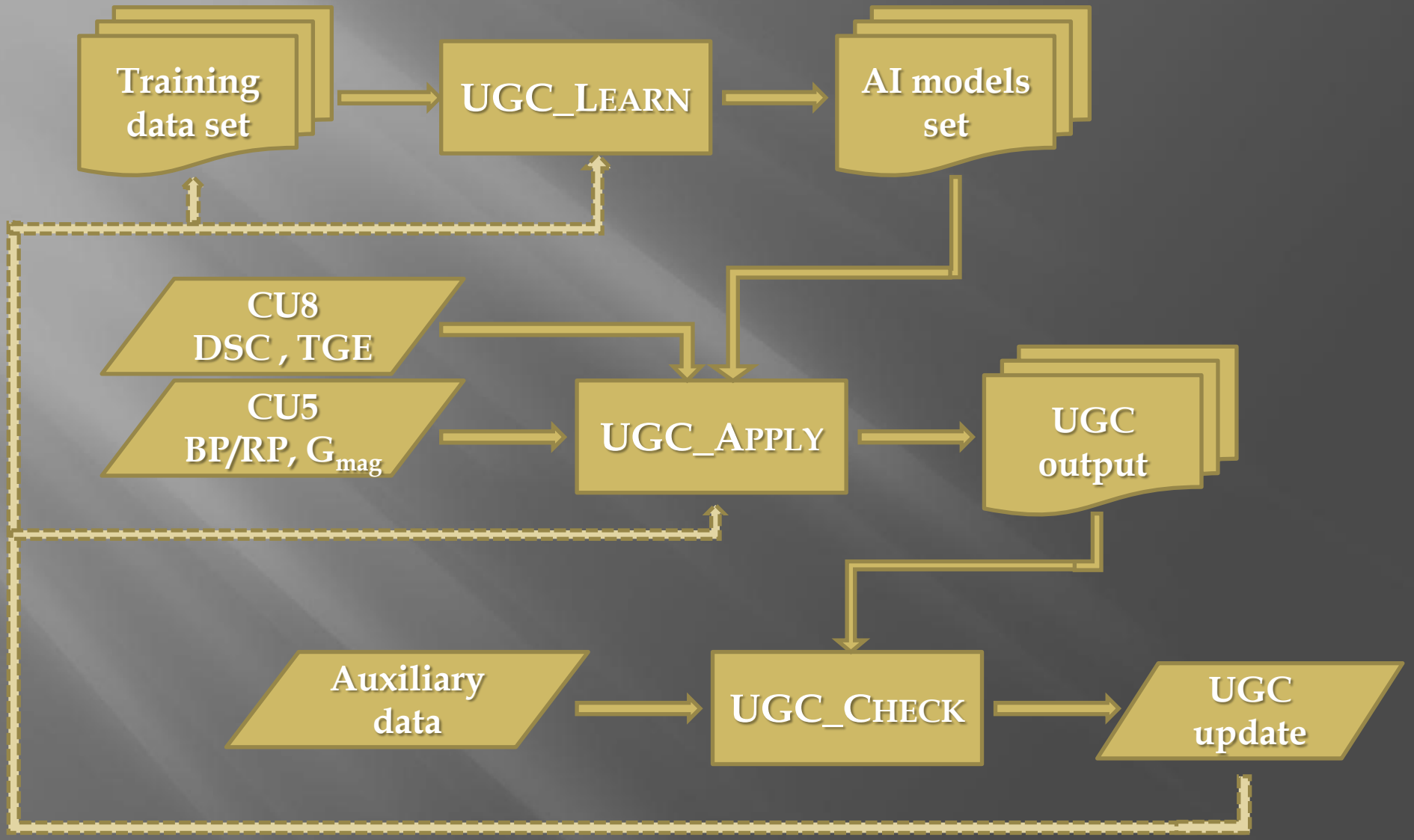
- **External Input for a Source**
  - CU5 - mean (ME and EOM) BP/RP spectra and G_{mag}
  - CU3 - astrometry
  - CU8/DSC – Class Probability vector
  - CU8/TGE – A_{TGE} for healpix

- **Software Implementation**
  - Supervised training based algorithm – Support Vector Machines (SVM)
  - Gaia MDB & DPC DataModel, common GaiaTools , DPCC and CU8 Tools
  - Offline learning and analysis modules
  - Online application module – part of DPAC ground-based science pipeline

UGC Overwiew: Functional approach

- **UGC_LEARN**
  - offline package based on AI supervised training technique
  - input simulated spectra library (SEDs and parameters)
  - tune, train and test AI-models (one at a time)
  - output trained AI-models set
- **UGC_APPLY**
  - ground-based pipeline package
  - external data input (CU5 G_{mag} and BP/RP, CU8 DSC and TGE)
  - auxiliary input - set of trained AI-models
  - processing algorithm based on AI application
  - output (***galType***, ***galParams***)
- **UGC_CHECK**
  - offline analysis package
  - decision on training set, learning and/or application updates

UGC Overview: Functional approach



UGC Overview: Performance

- **Classification – *galType* spectral types**

  galE, galS, galI, galB

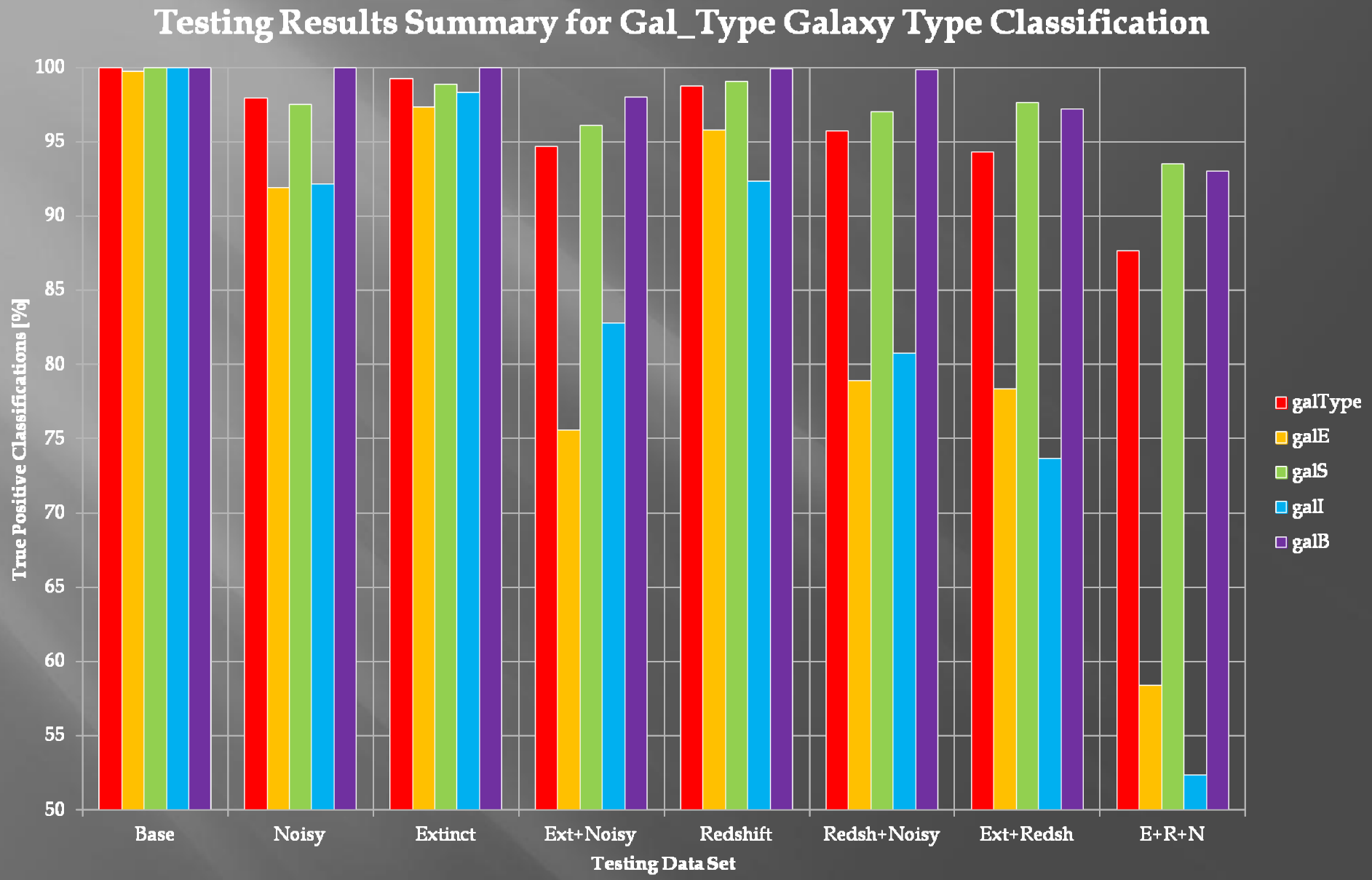
  performance = TruePositive Percentage
  [% of sources with *galaxyProb*(real galType) > 0.5]

- **Regression – *galParams* parameters**

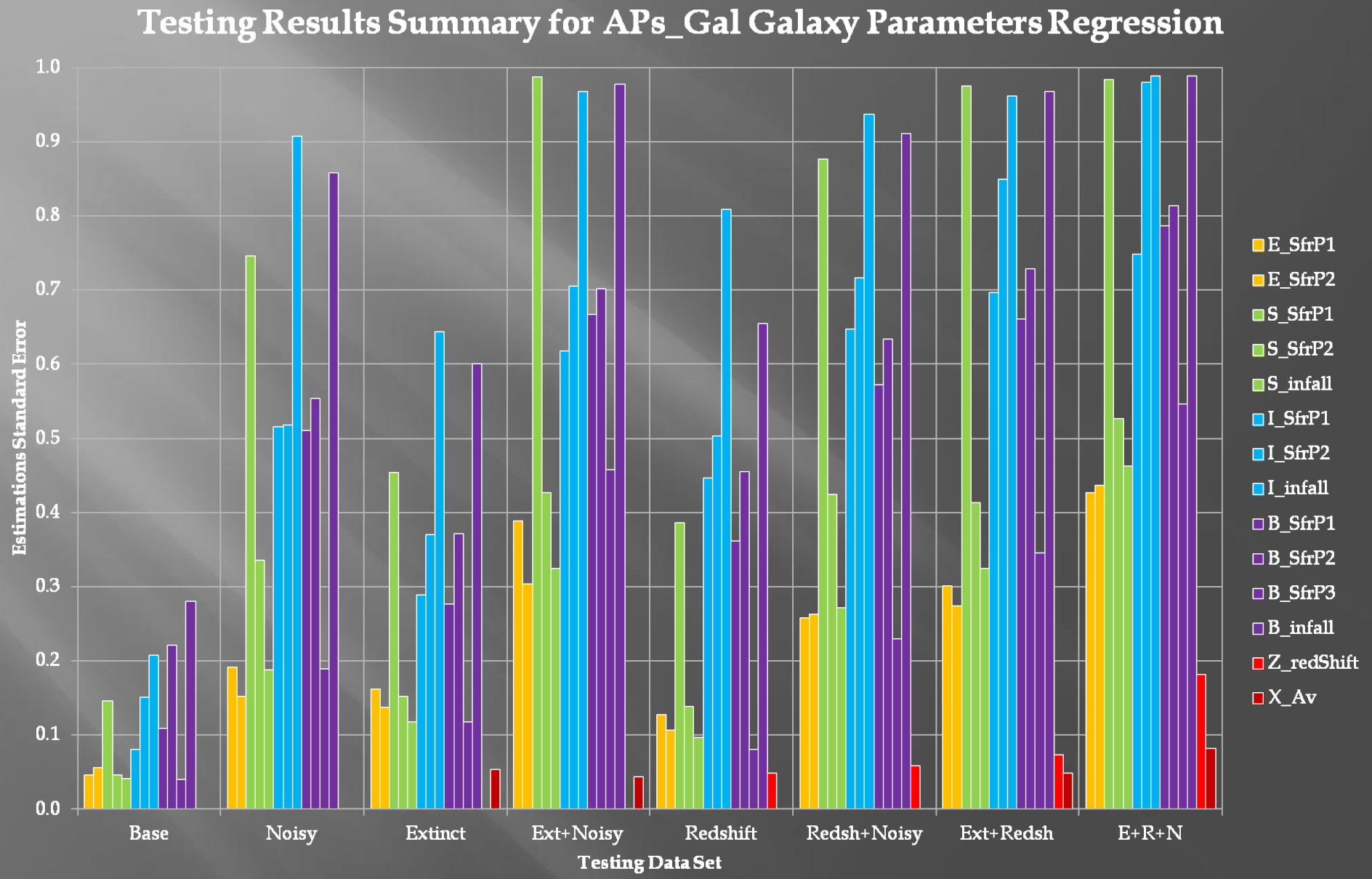
Z_redShift	*X_A$_0$*		
E_sfrP1	*E_sfrP2*		
S_sfrP1	*S_sfrP2*	*S_infall*	
I_sfrP1	*I_sfrP2*	*I_infall*	
B_sfrP1	*B_sfrP2*	*B_sfrP3*	*B_infall*

  performance = RMSD
  [estimated vs. real parameter value of the sources]

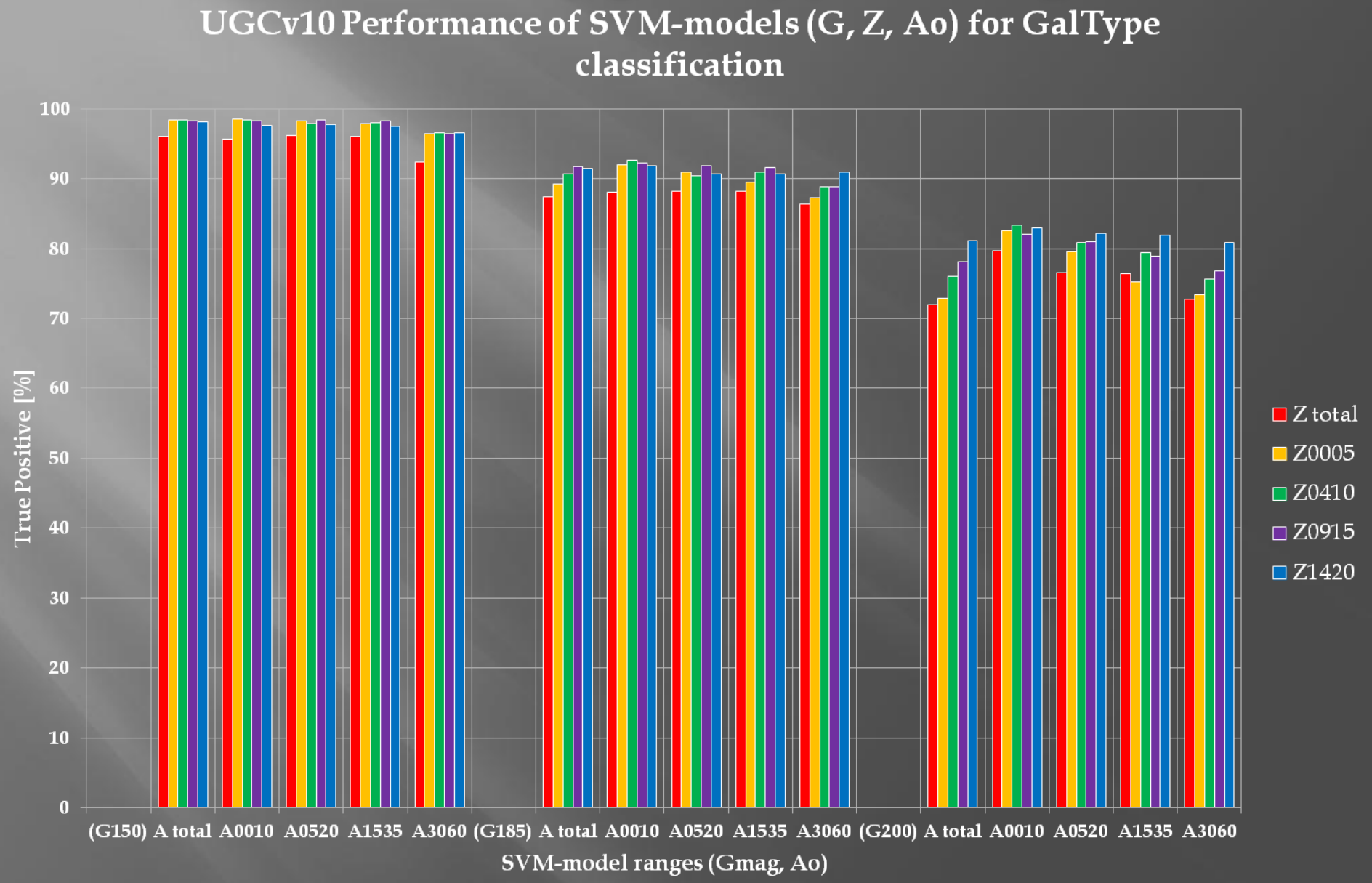
SVM tests: *galType* classification



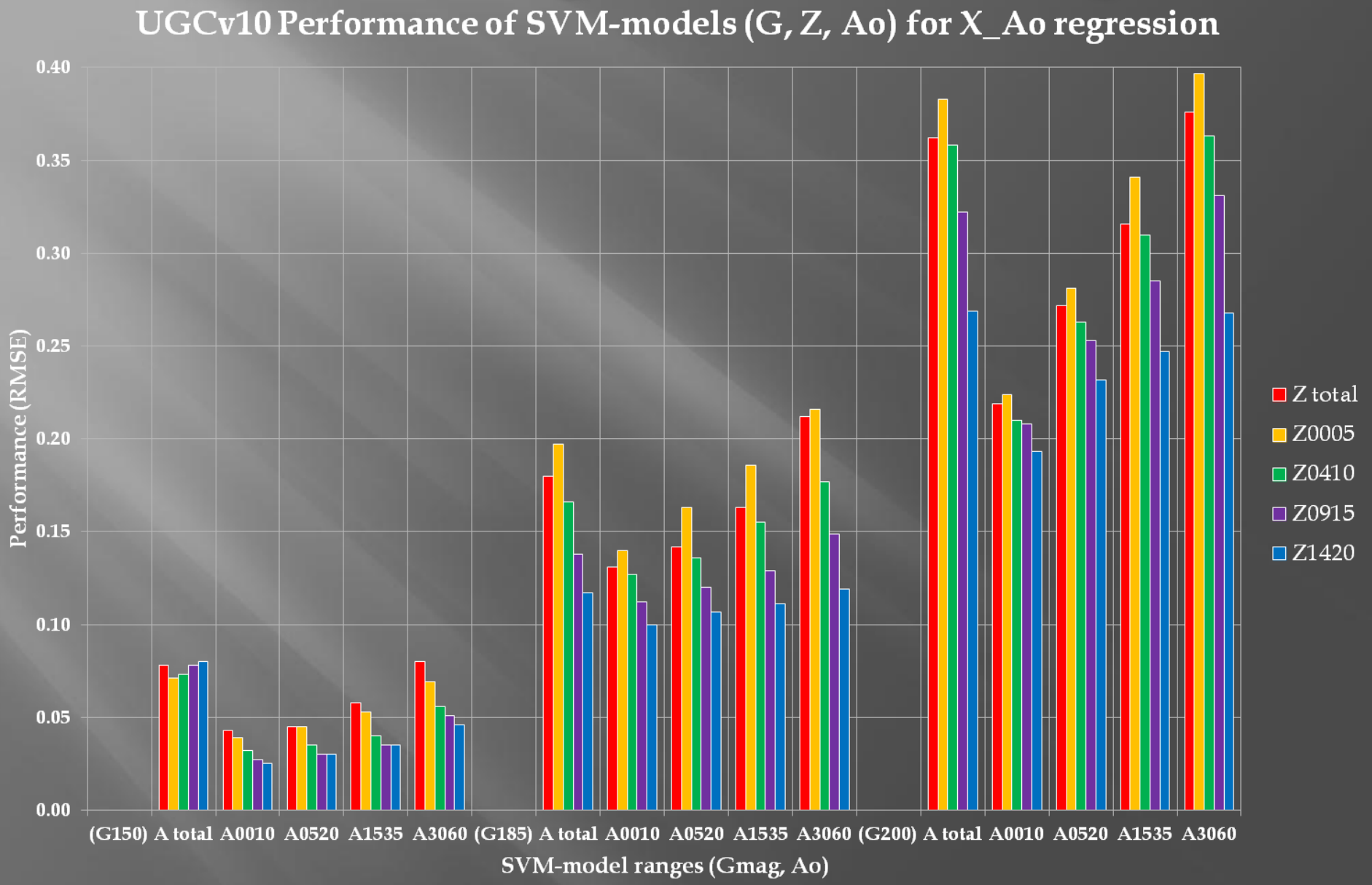
SVM tests: *galParams* regression



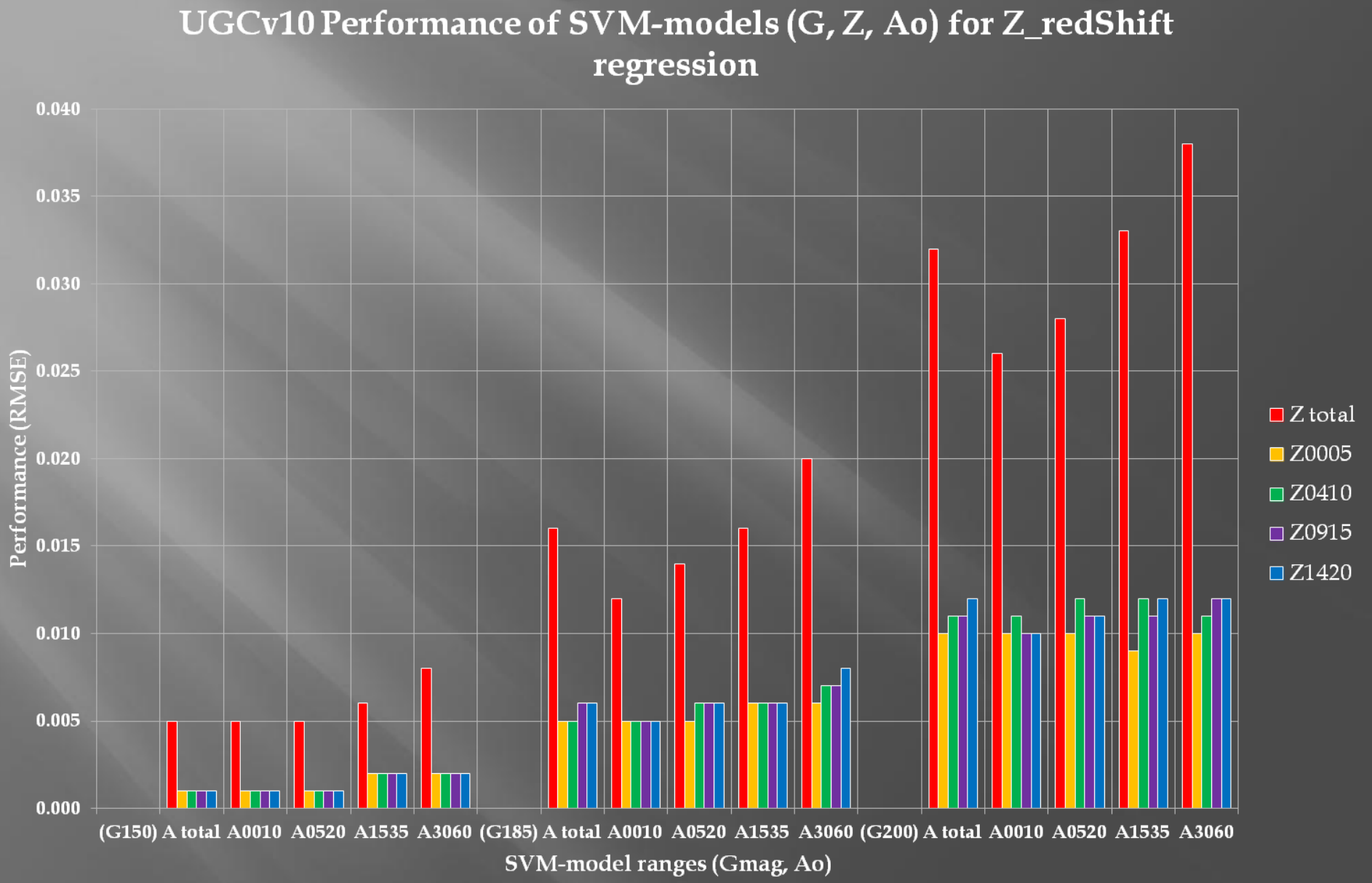
SVM tests: *galType* ranges



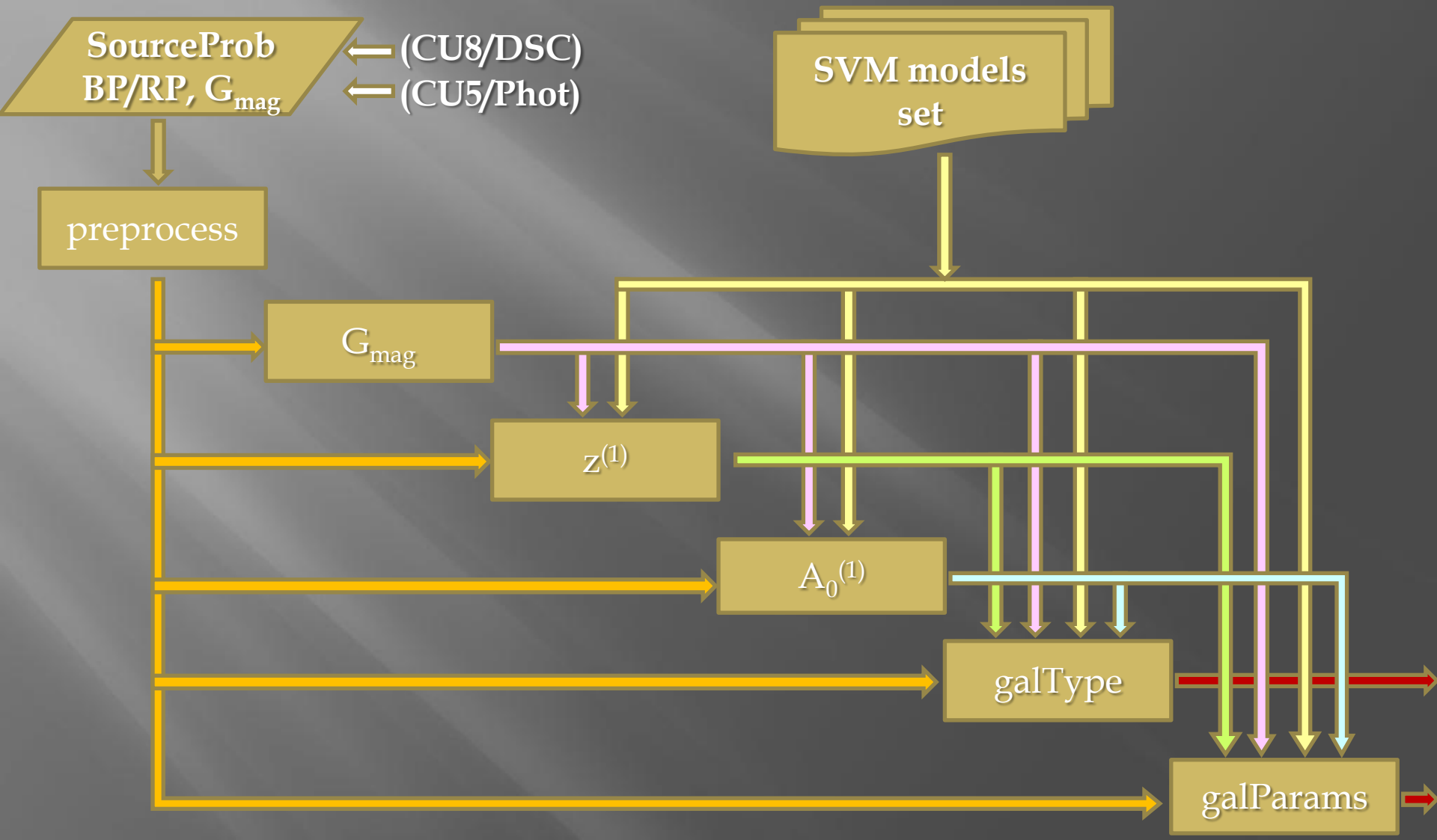
SVM tests: *galParams ranges*



SVM tests: *galParams ranges*



UGC : Application module

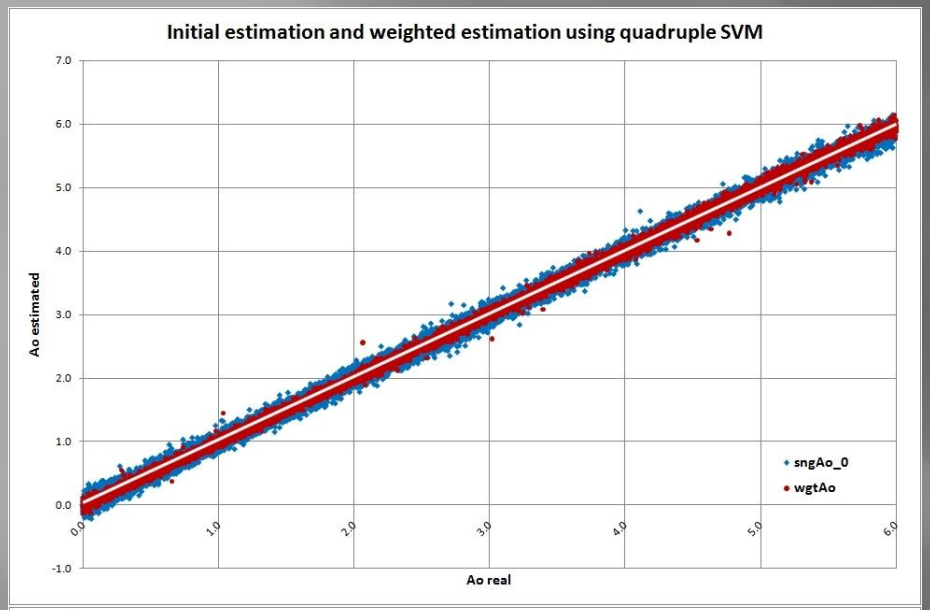


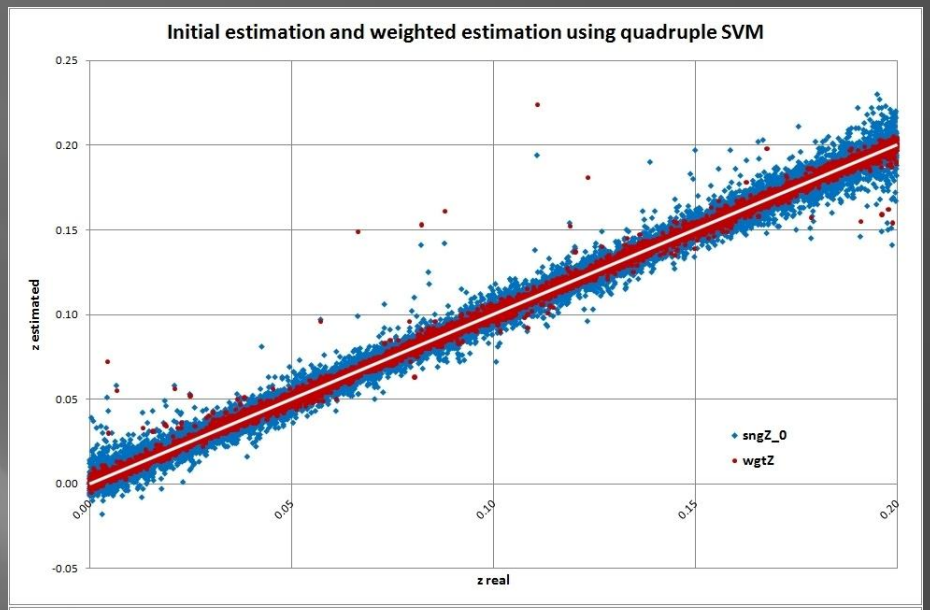
UGC Performance: *Ao* & *z* parameters

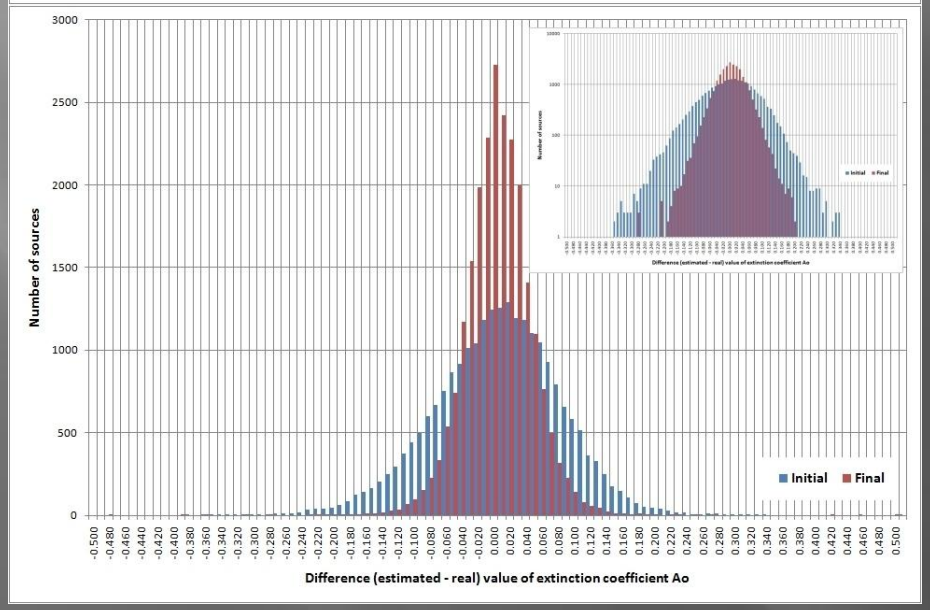
Galaxy model external parameter		UGCv12 regression performance					
		G_{mag}=15.0		G_{mag}=18.5		G_{mag}=20.0	
APs_gal	Range	initial	final	initial	final	initial	final
z	0.00 - 0.20	0.005	**0.002**	0.015	**0.011**	0.031	**0.028**
Ao	0.0 – 6.0	0.08	**0.04**	0.18	**0.15**	0.36	**0.35**

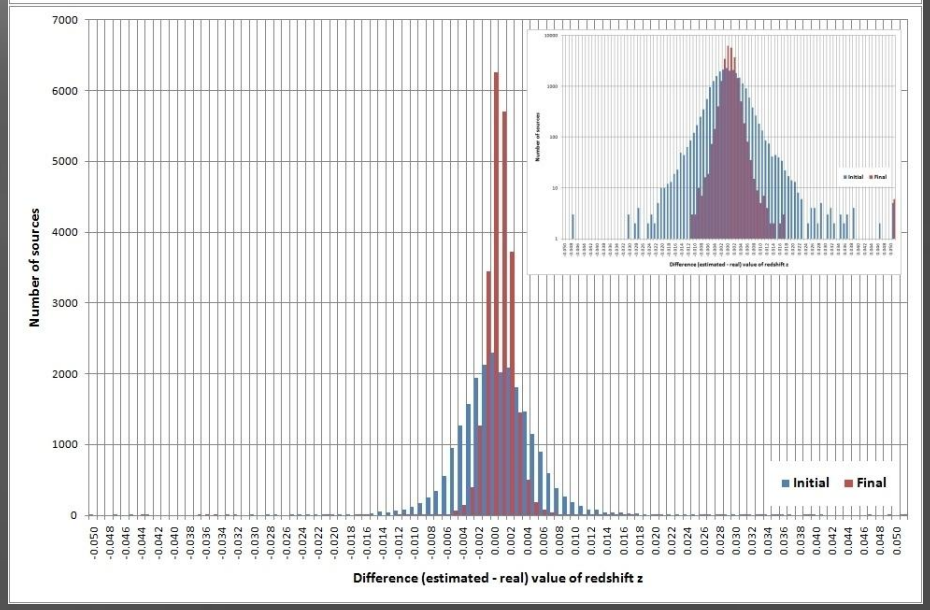
(The performance is given as RMSD for the initial and for the final estimation)

UGC Performance: *Ao* & *z* (G150)

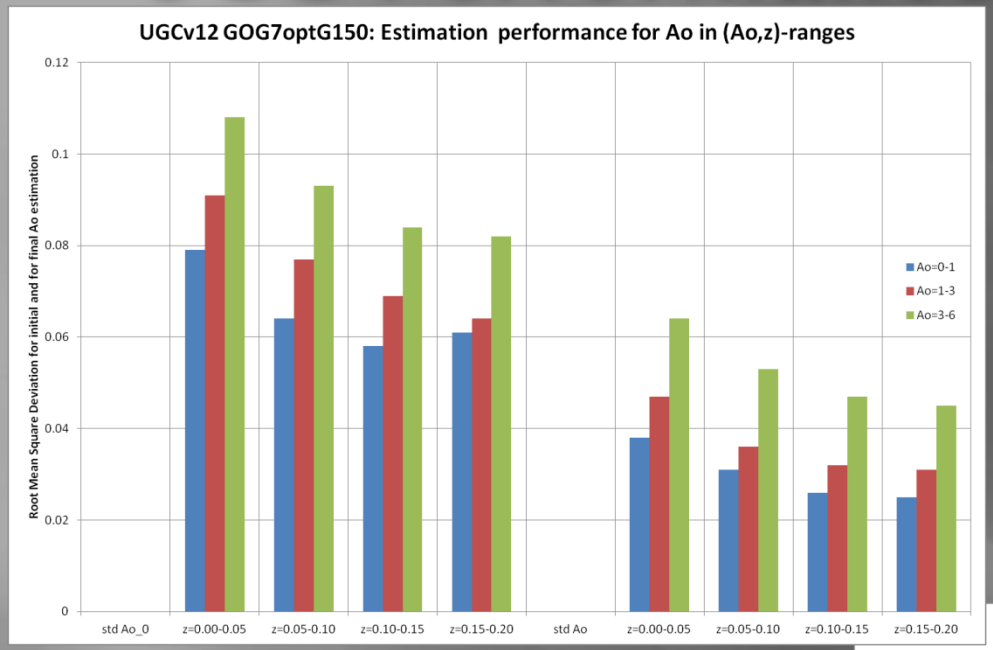


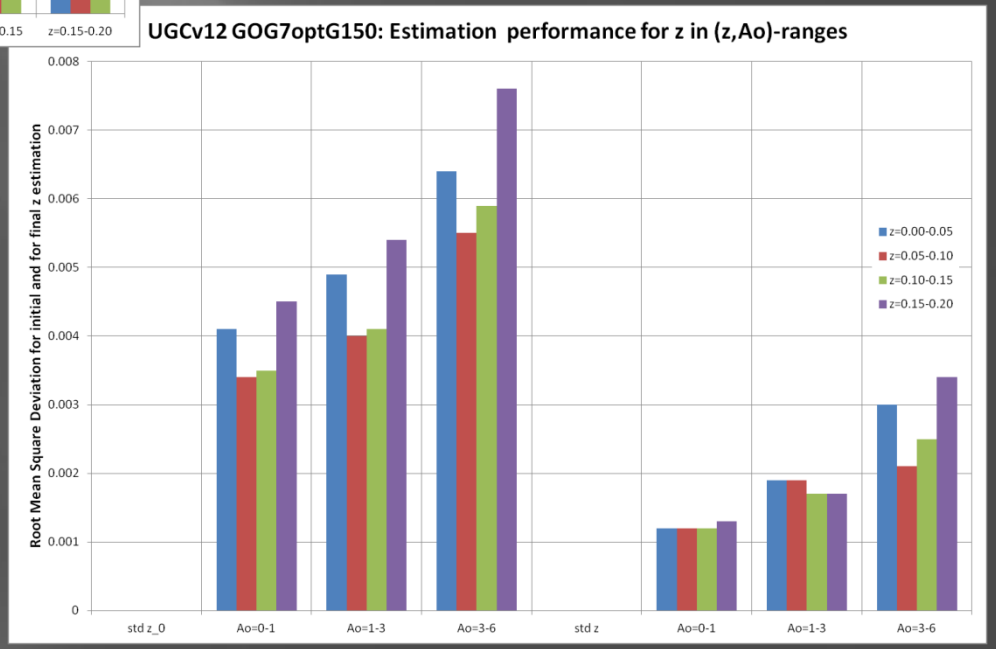




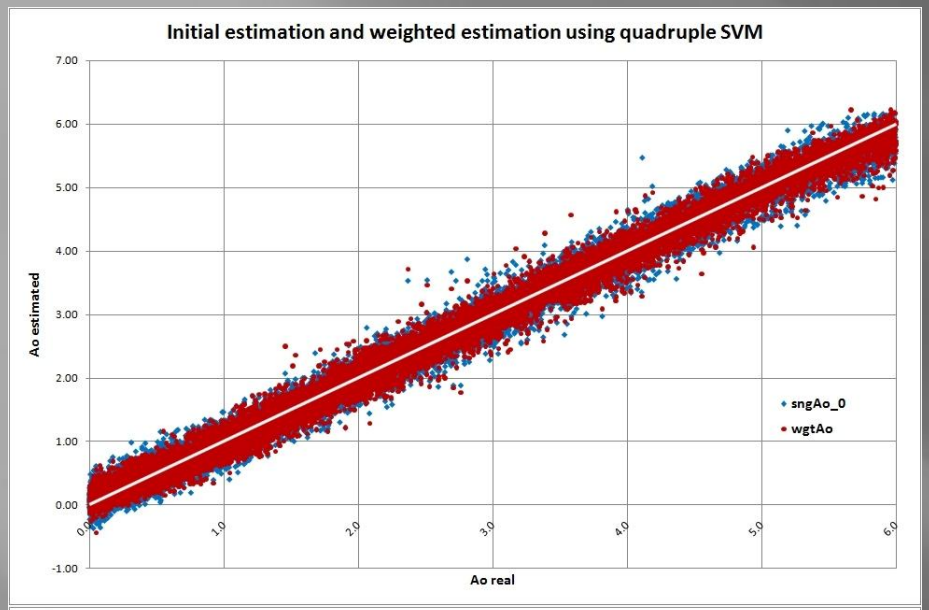


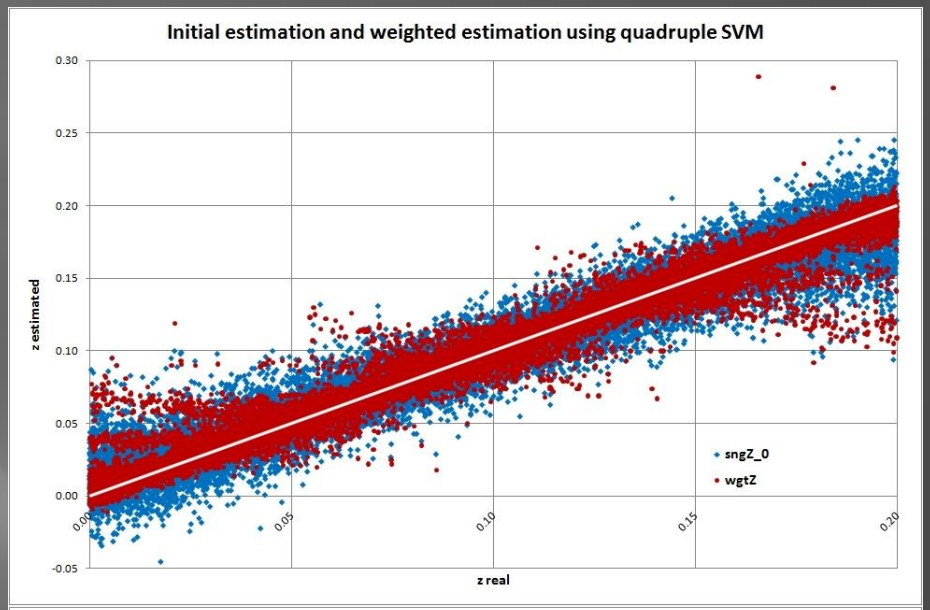
UGC Performance: *Ao & z* (G150)

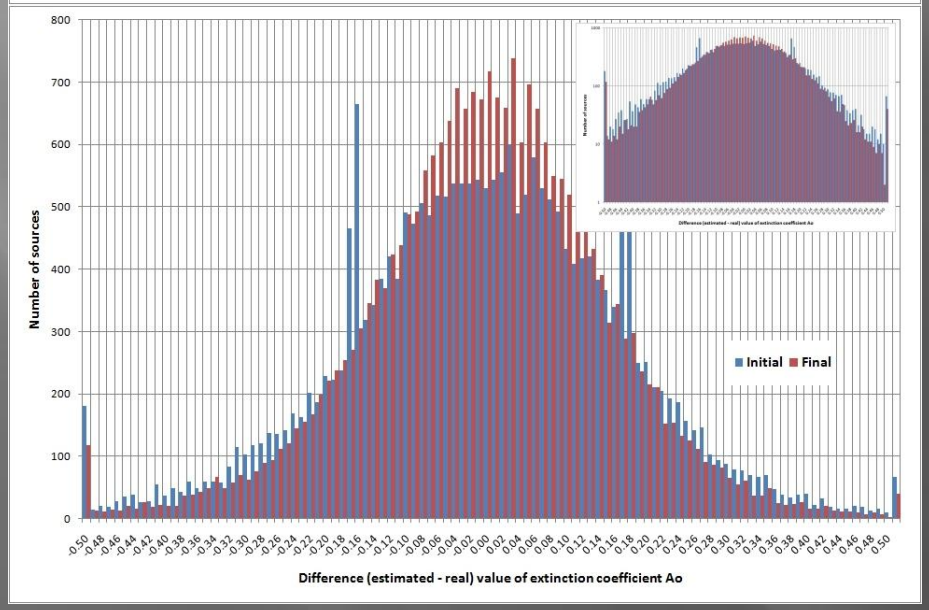


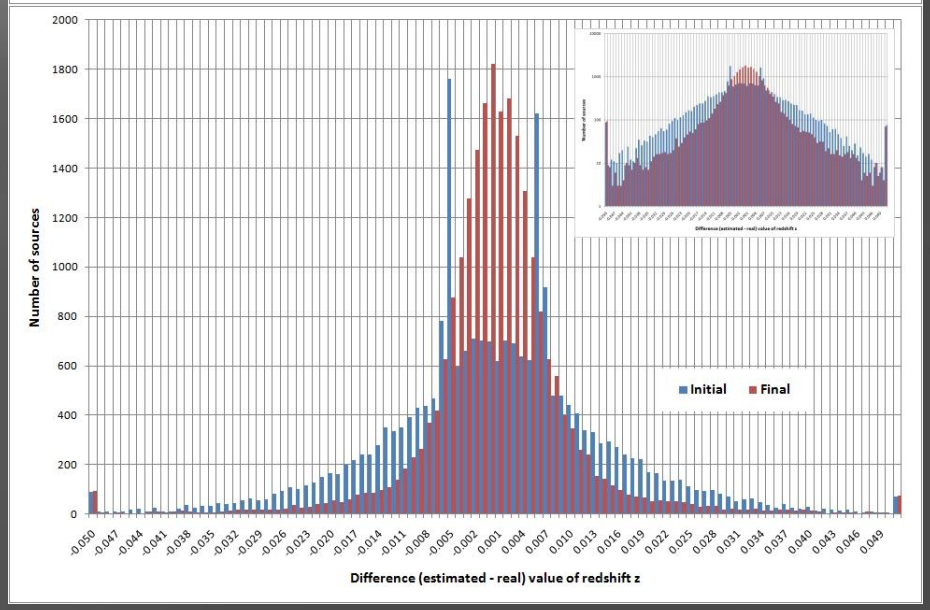


UGC Performance: *Ao* & *z* (G185)

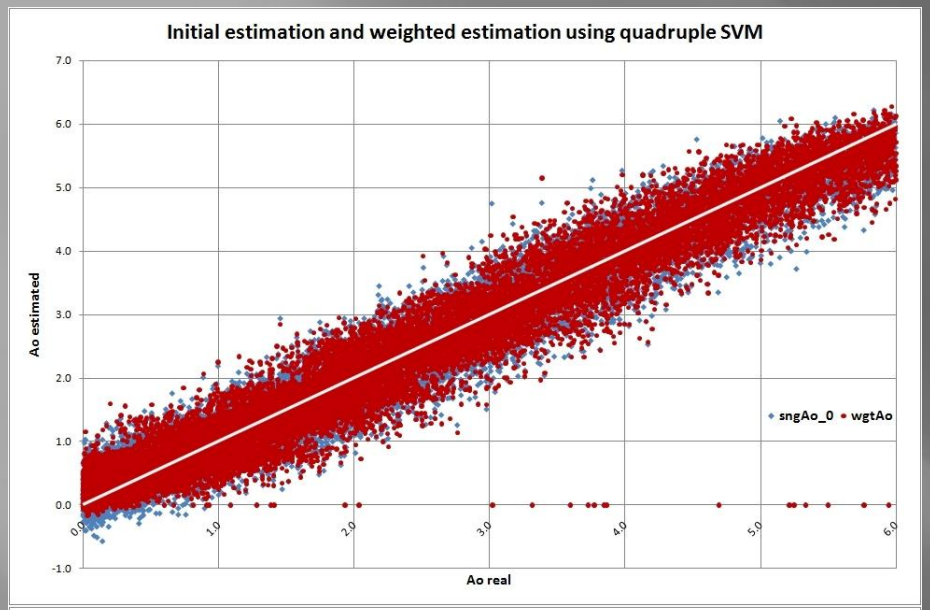


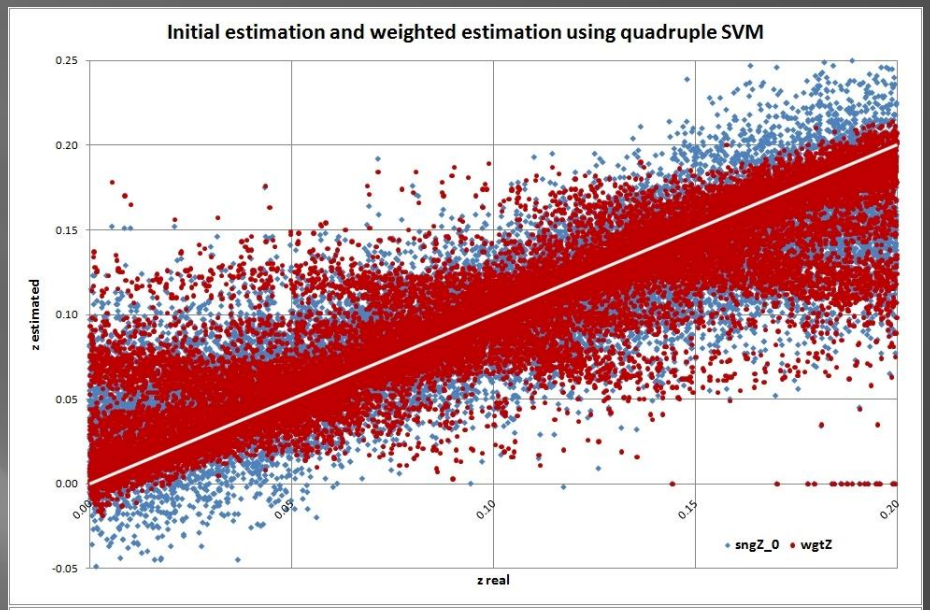


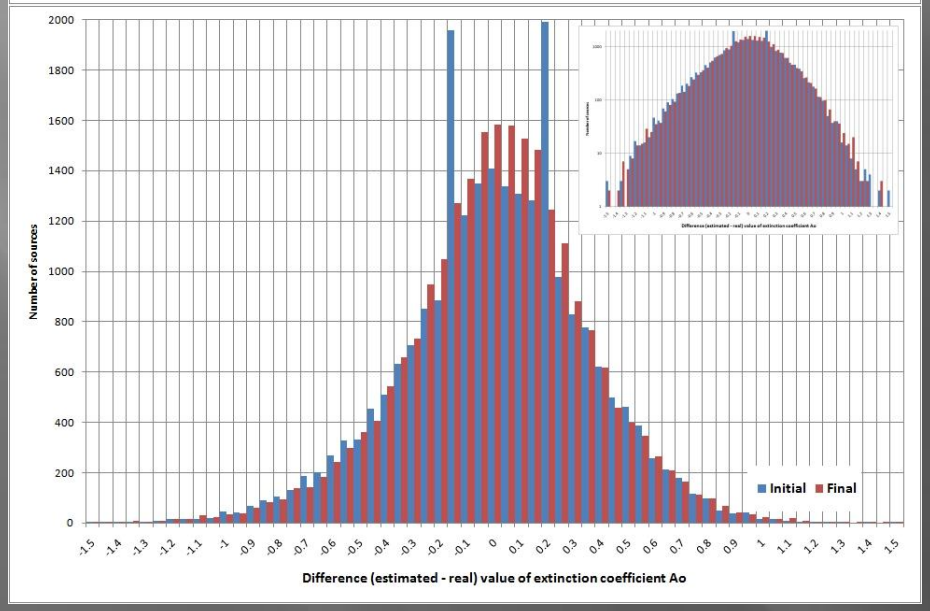


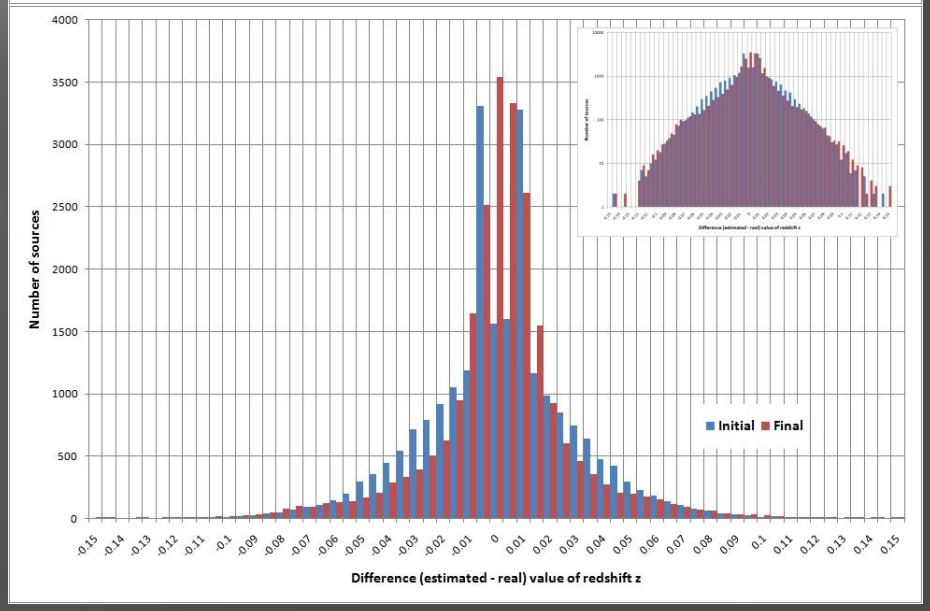


UGC Performance: *Ao* & *z* (G200)









UGC performance: *galType* classification

(prob>0.5)	*UGC gal_E*	*UGC gal_S*	*UGC gal_I*	*UGC gal_B*	*Unknown*
Real gal_E	**93%**	7%	0%	0%	1%
Real gal_S	1%	**98%**	0%	0%	0%
Real gal_I	0%	5%	**89%**	3%	3%
Real gal_B	0%	0%	0%	**99%**	0%

G_{mag} = 15.0

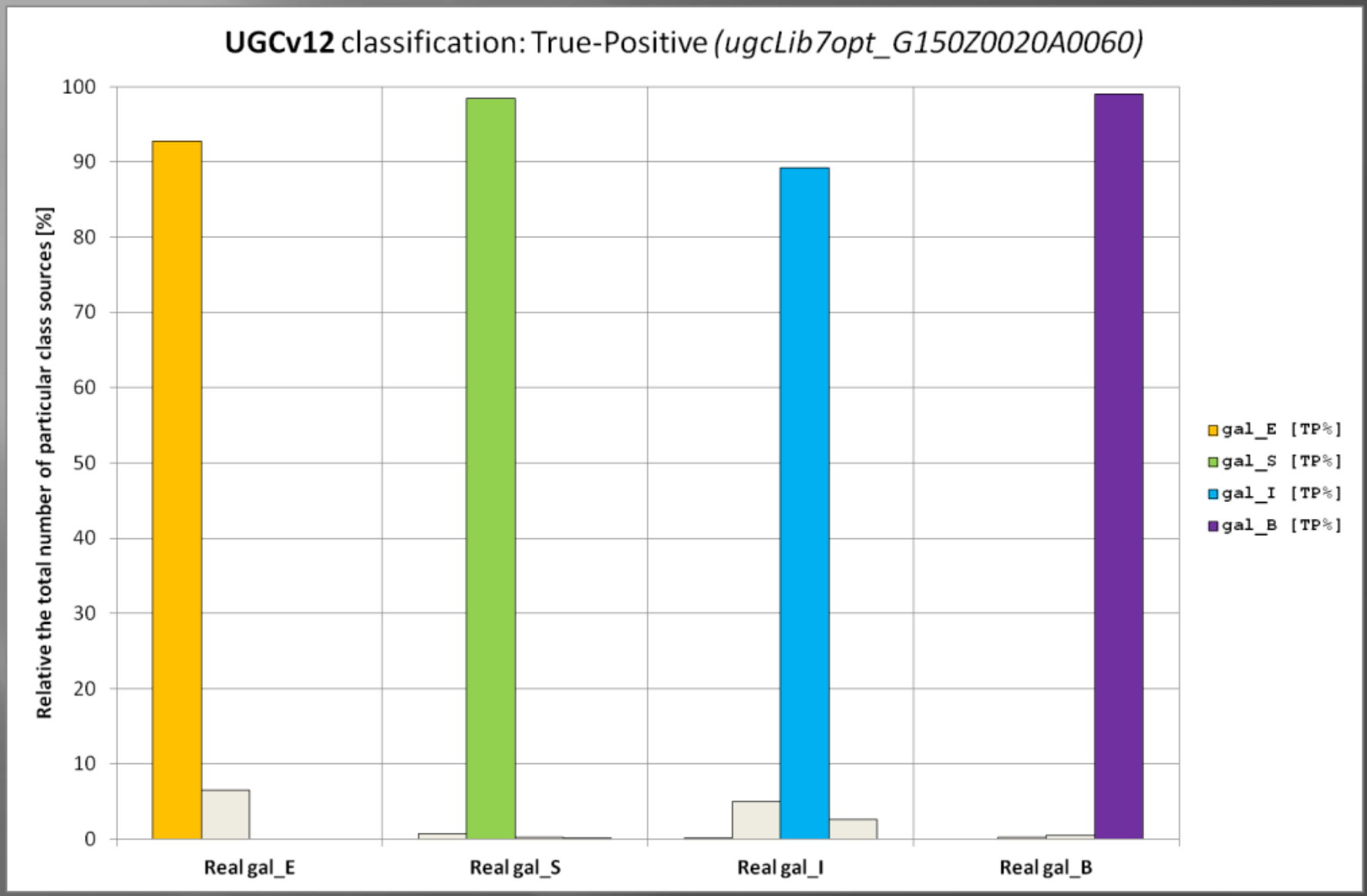
(prob>0.5)	*UGC gal_E*	*UGC gal_S*	*UGC gal_I*	*UGC gal_B*	*Unknown*
Real gal_E	**78%**	19%	0%	0%	2%
Real gal_S	4%	**90%**	0%	3%	3%
Real gal_I	0%	7%	**51%**	28%	13%
Real gal_B	0%	2%	2%	**94%**	2%

G_{mag} = 18.5

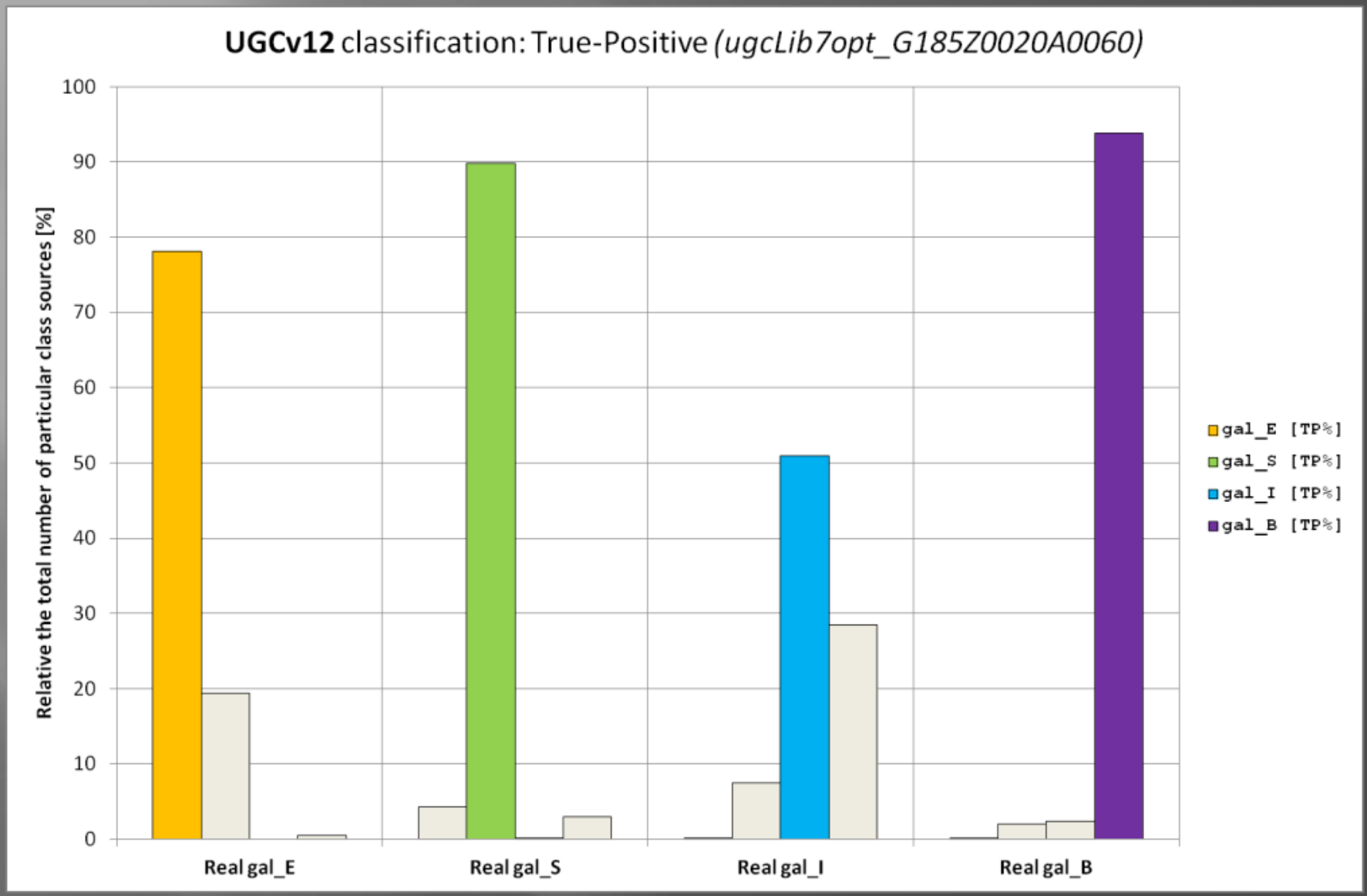
(prob>0.5)	*UGC gal_E*	*UGC gal_S*	*UGC gal_I*	*UGC gal_B*	*Unknown*
Real gal_E	**44%**	34%	0%	7%	14%
Real gal_S	6%	**64%**	0%	18%	12%
Real gal_I	0%	11%	**28%**	45%	17%
Real gal_B	0%	7%	2%	**83%**	7%

G_{mag} = 20.0

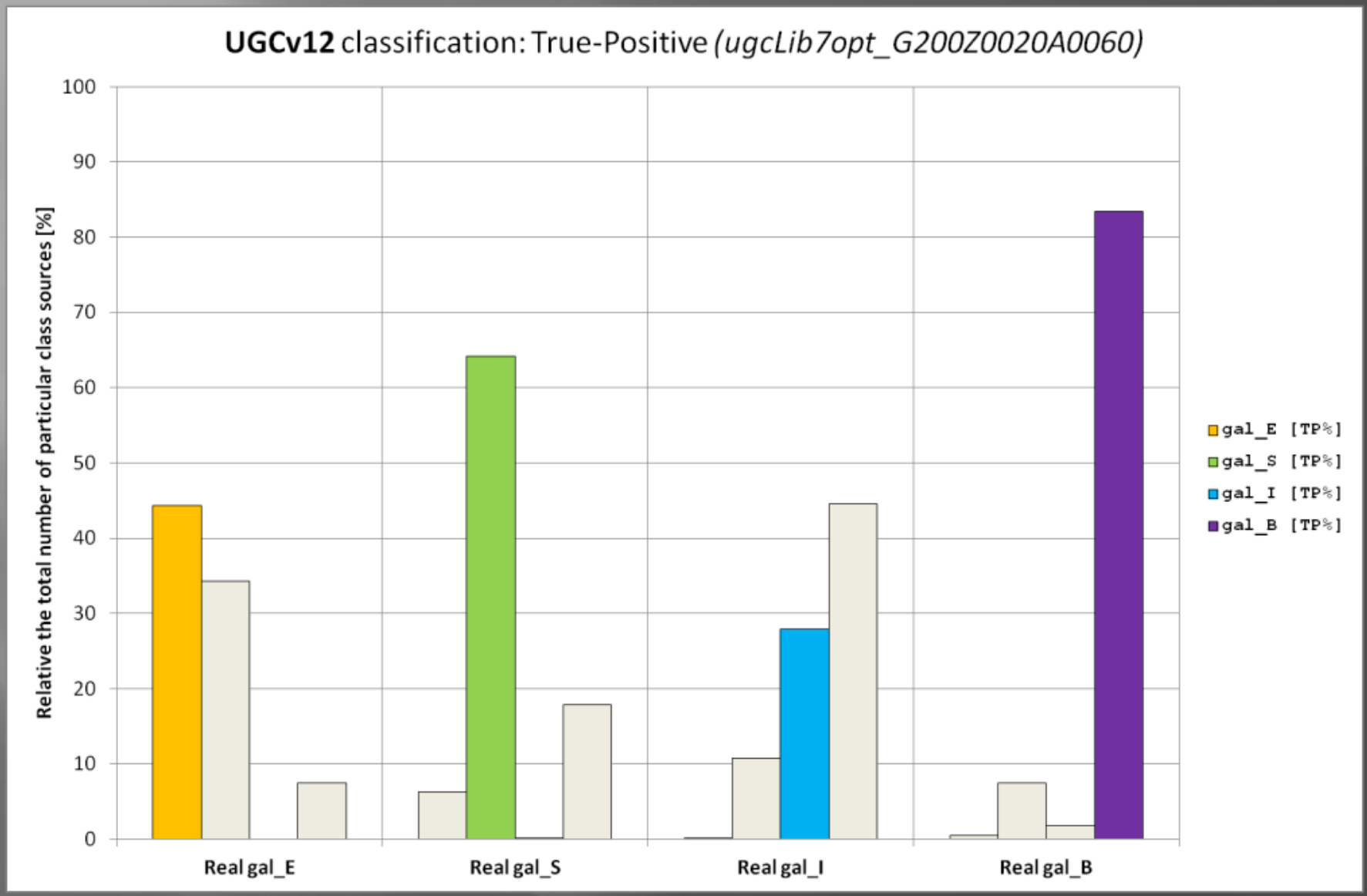
UGC performance: *galType (G150)*



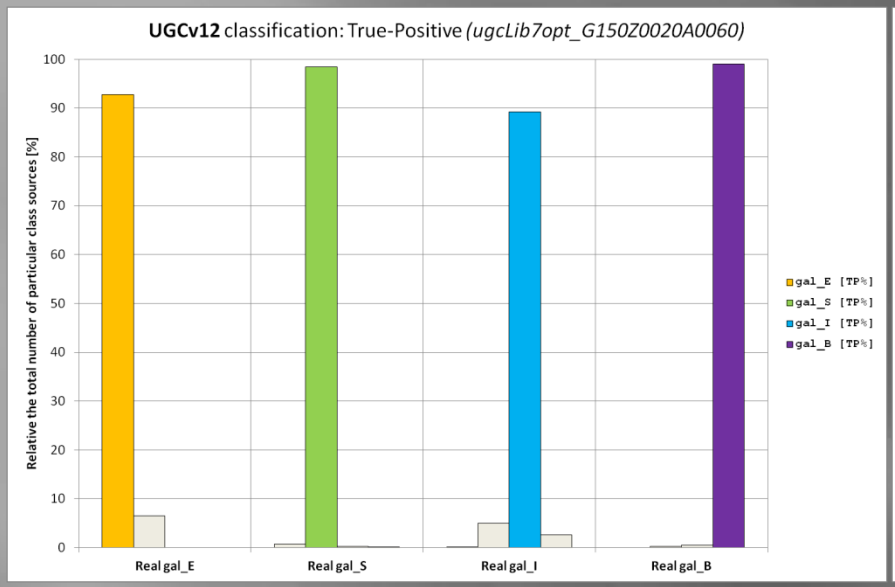
UGC performance: *galType (G185)*

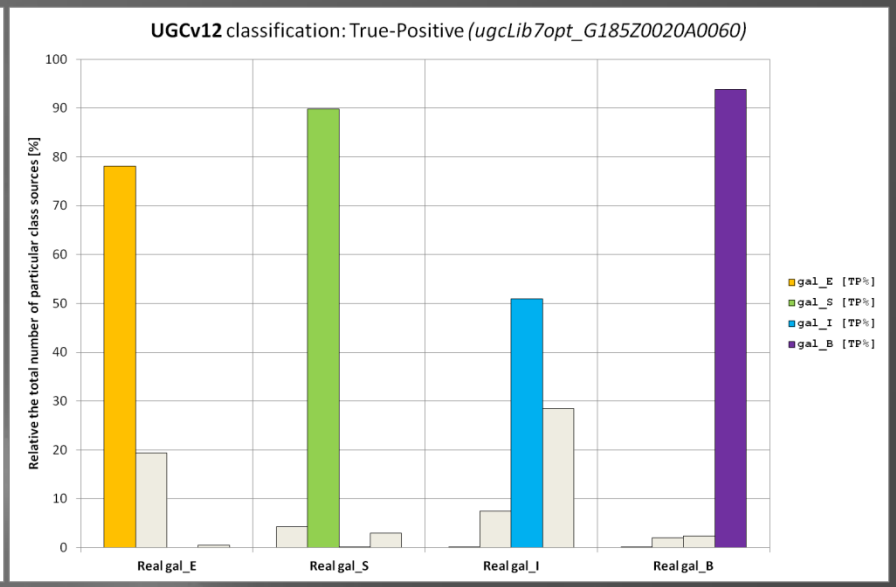


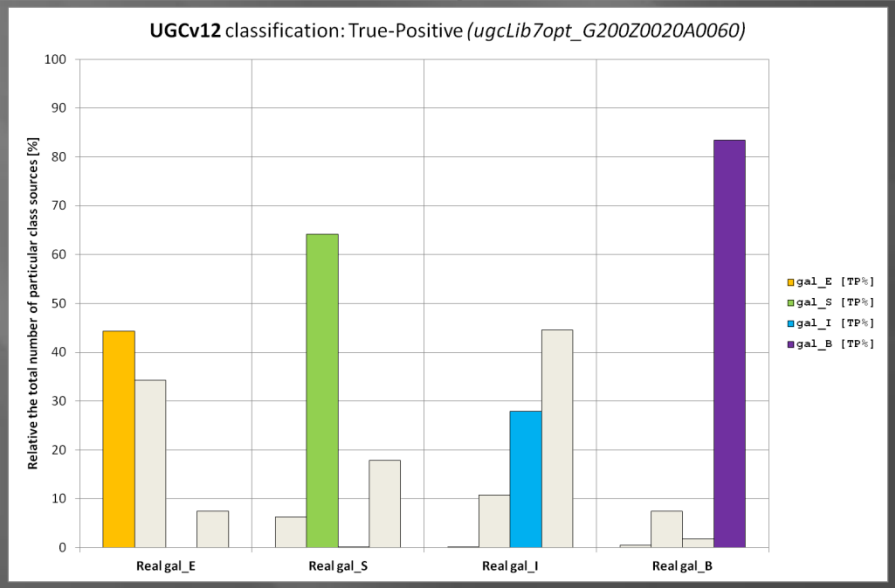
UGC performance: *galType (G200)*



UGC performance: *galType* classification







UGC performance: *galType*

(prob>0.5)	*UGC gal_E*	*UGC gal_S*	*UGC gal_I*	*UGC gal_B*	*Unknown*
Real gal_E	**91%**	7%	0%	2%	1%
Real gal_S	1%	**95%**	0%	2%	1%
Real gal_I	0%	5%	**86%**	5%	4%
Real gal_B	1%	0%	0%	**98%**	0%

G_{mag} = (13.0 - 16.0)

(prob>0.5)	*UGC gal_E*	*UGC gal_S*	*UGC gal_I*	*UGC gal_B*	*Unknown*
Real gal_E	**79%**	18%	0%	1%	2%
Real gal_S	4%	**92%**	0%	3%	3%
Real gal_I	0%	8%	**54%**	23%	14%
Real gal_B	1%	2%	2%	**94%**	1%

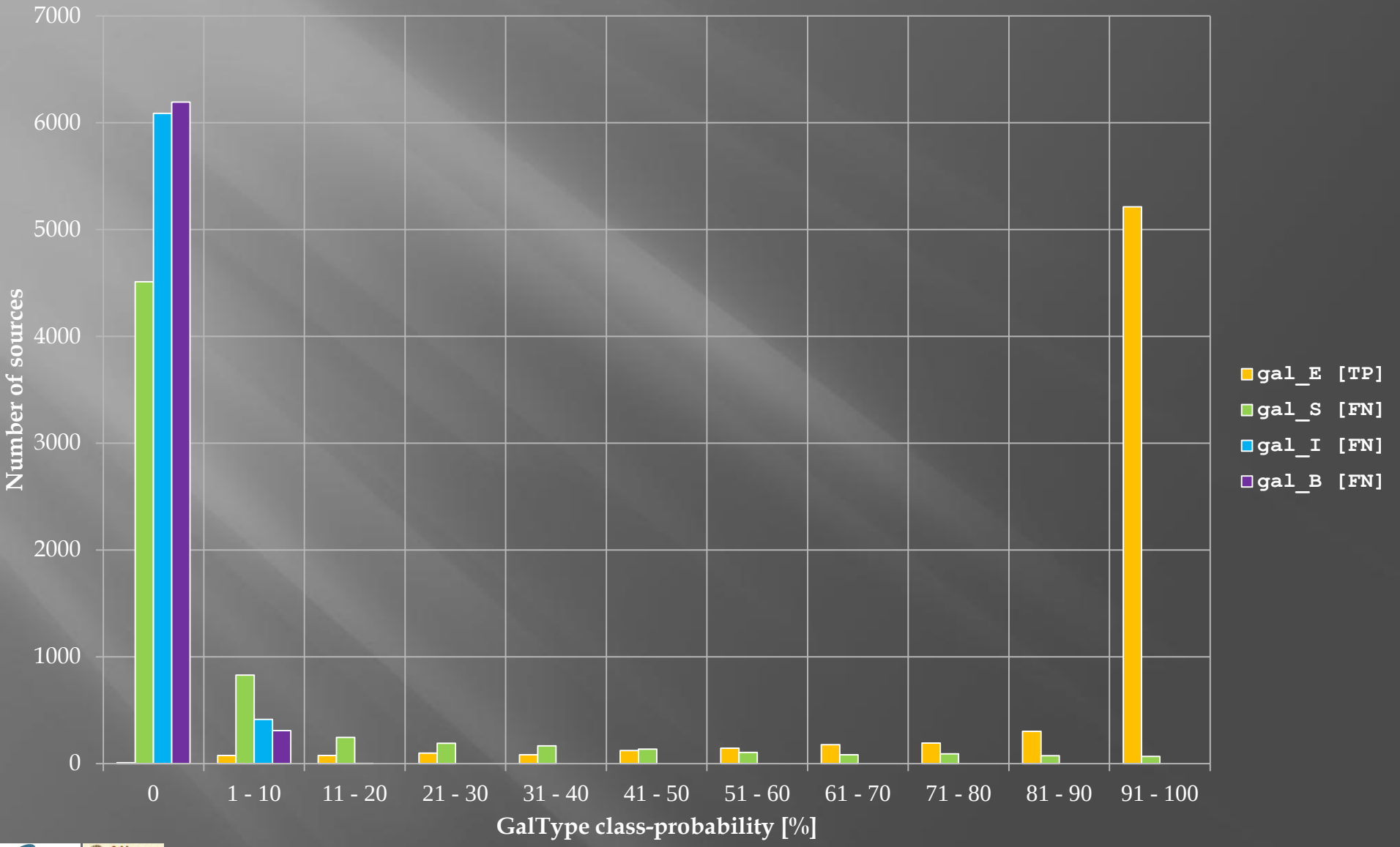
G_{mag} = (16.0 - 19.0)

(prob>0.5)	*UGC gal_E*	*UGC gal_S*	*UGC gal_I*	*UGC gal_B*	*Unknown*
Real gal_E	**50%**	33%	0%	3%	13%
Real gal_S	7%	**73%**	0%	11%	9%
Real gal_I	1%	12%	**26%**	41%	20%
Real gal_B	2%	7%	1%	**83%**	6%

G_{mag} = (19.0 - 20.0)

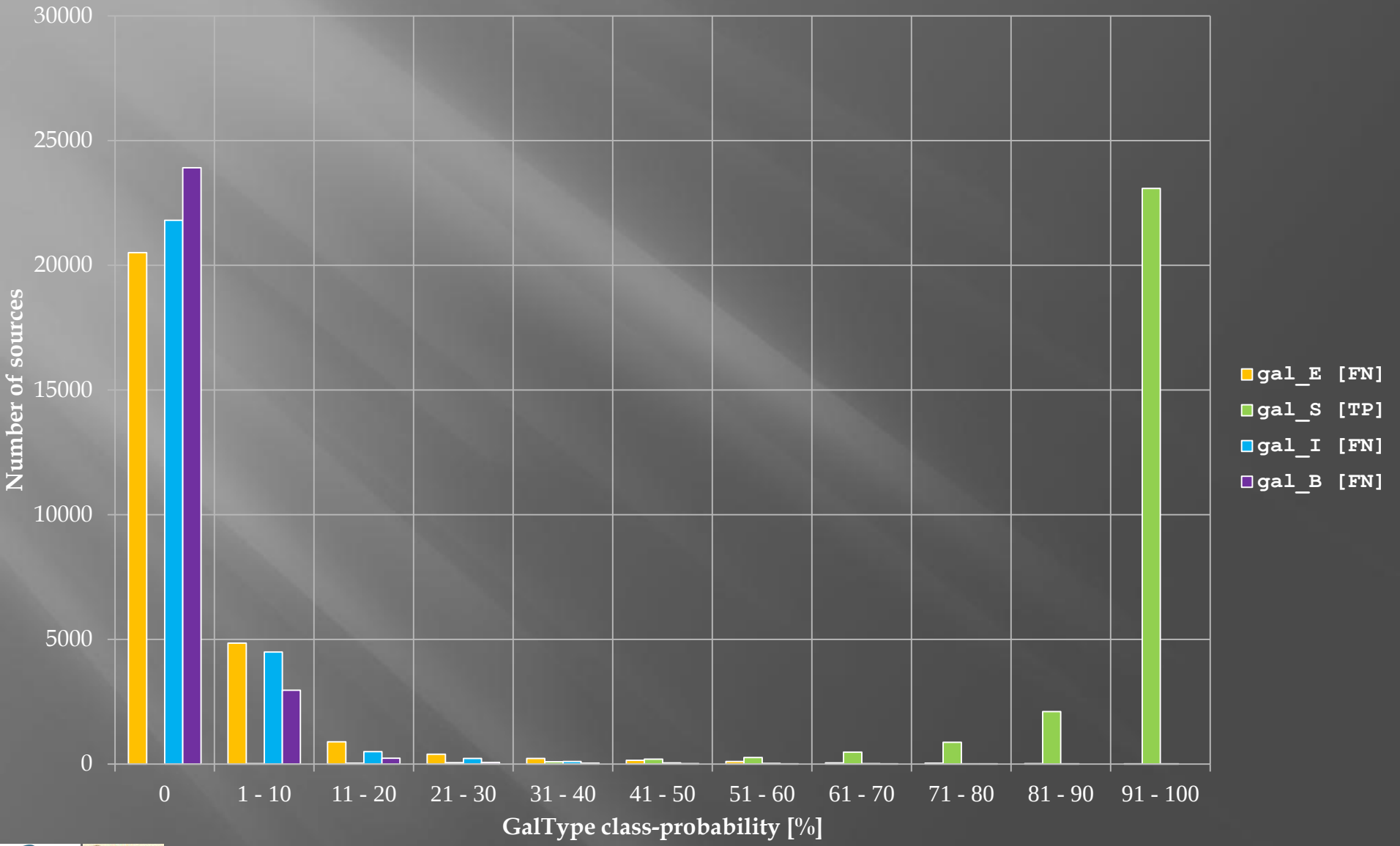
UGC performance: *gal_E* class (G150)

UGCv12 classification: ***gal_E*** type spectra *(ugcLib7opt_G150Z0020A0060)*



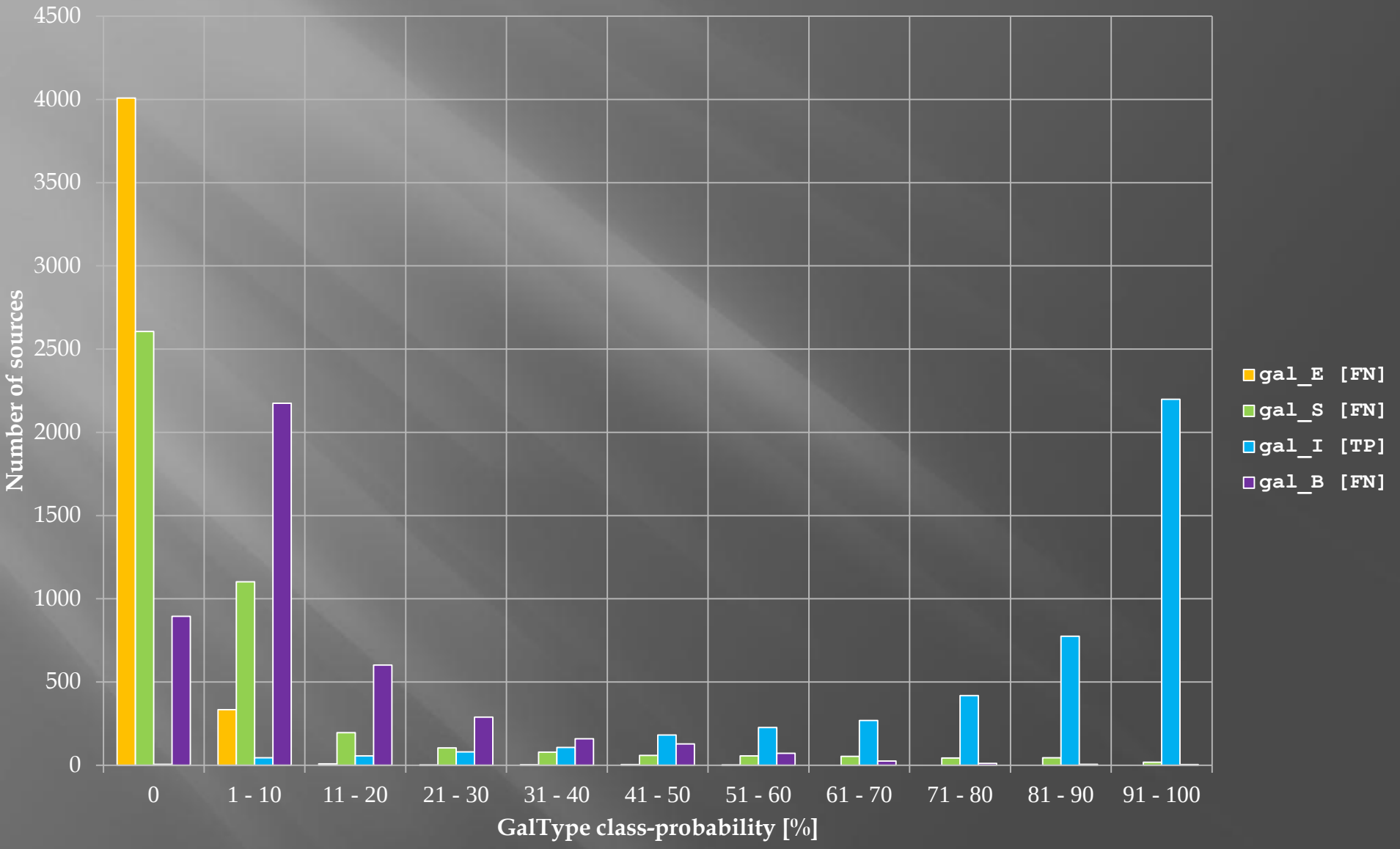
UGC performance: *gal_S* class (G150)

UGCv12 classification: ***gal_S*** type spectra *(ugcLib7opt_G150Z0020A0060)*



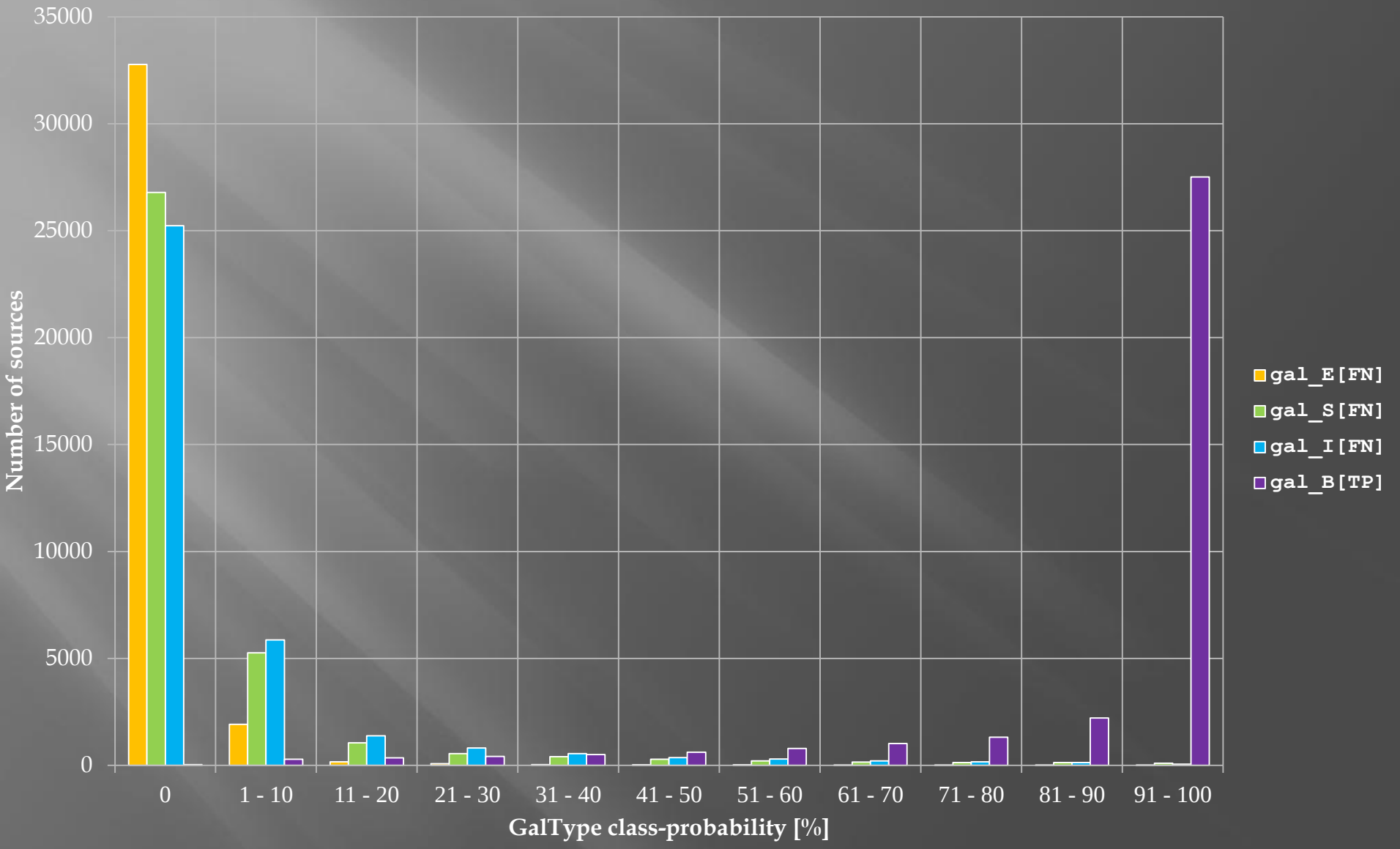
UGC performance: *gal_I* class (G150)

UGCv12 classification: ***gal_I*** type spectra (*ugcLib7opt_G150Z0020A0060*)



UGC performance: *gal_B* class (G150)

UGCv12 classification: gal_B type spectra (*ugcLib7opt_G185Z0020A0060*)

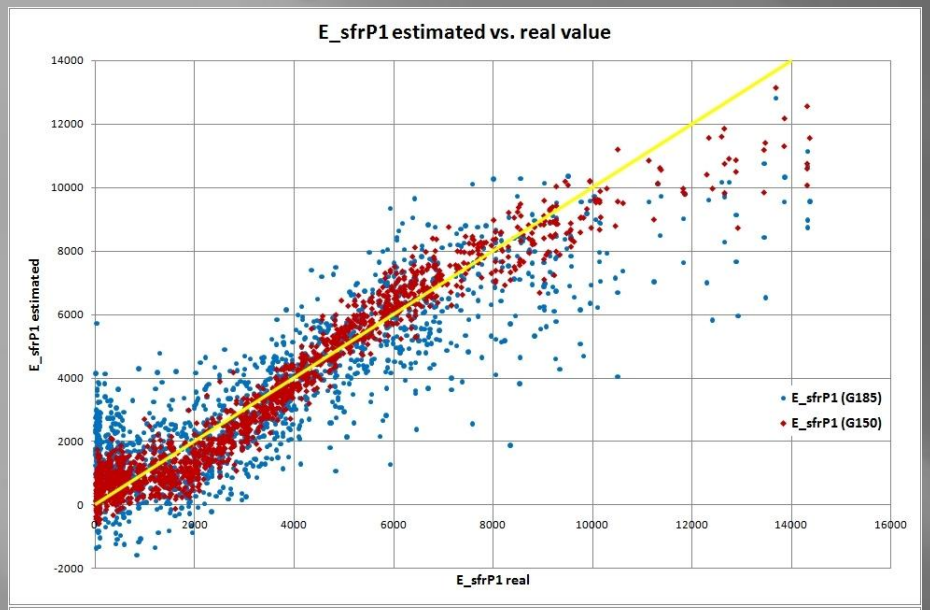


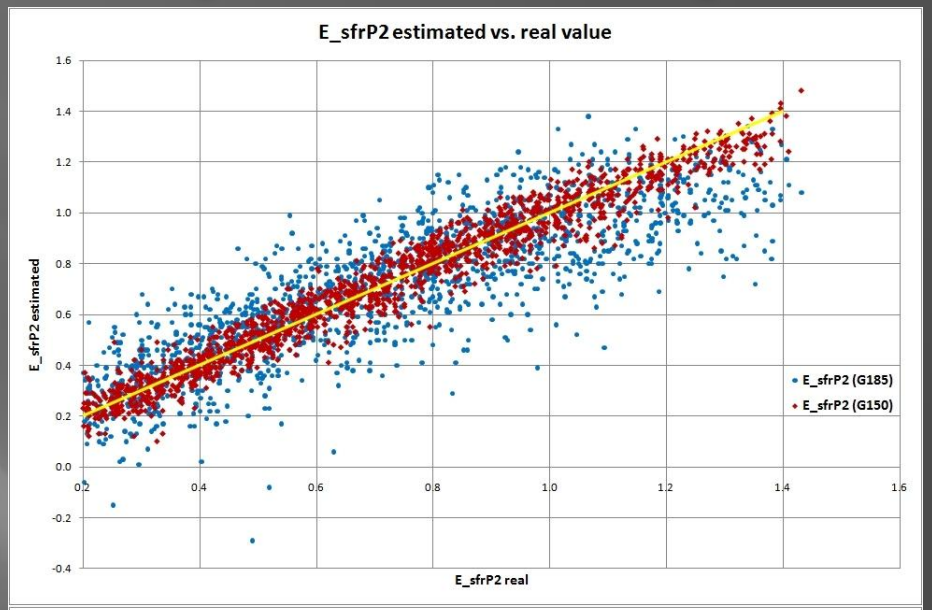
UGC performance: *SF-parameters*

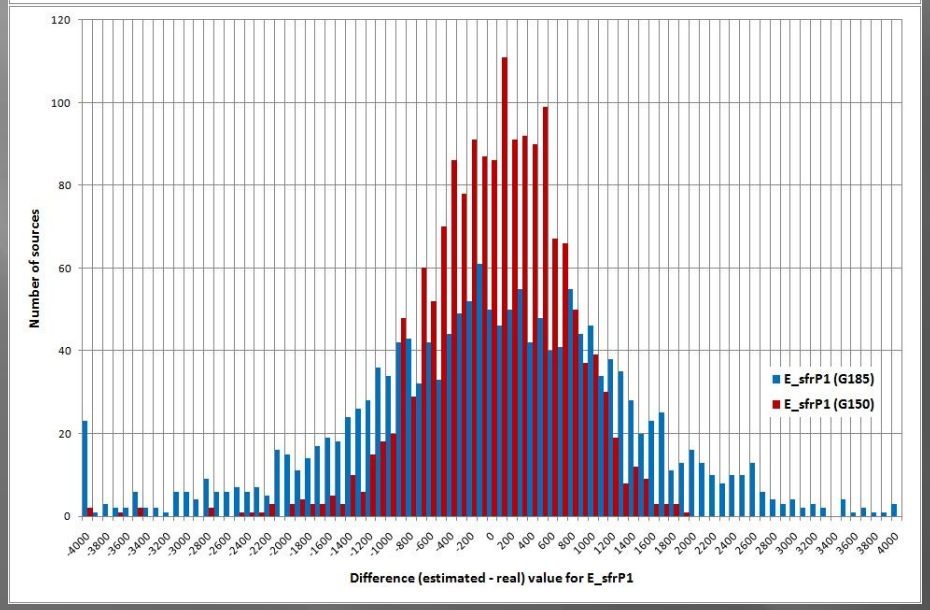
Galaxy model				UGCv12 regression performance			
galType	*galParam*	Range	Units	G_{mag}=15.0		G_{mag}=18.5	
gal_E	*E_sfrP1*	10 - 14400	*My*	**696**	(0.048)	**1459**	(0.101)
	E_sfrP2	0.20 - 1.45	$M_{\odot}$	**0.06**	(0.048)	**0.16**	(0.128)
gal_S	*S_sfrP1*	0.30 - 2.40		**0.56**	* (0.267)	**0.60**	* (0.286)
	S_sfrP2	5 - 30000	$My/M_{\odot}$	**3521**	(0.117)	**4979**	(0.166)
	S_infall	5 - 16000	*My*	**1336**	(0.084)	**2714**	(0.170)
gal_I	*I_sfrP1*	0.60 - 3.90		**0.61**	(0.185)	**0.68**	(0.206)
	I_sfrP2	4000 - 70000	$My/M_{\odot}$	**14400**	(0.218)	**18048**	* (0.273)
	I_infall	5000 - 30000	*My*	**5759**	* (0.230)	**6143**	* (0.246)
gal_B	*B_sfrP1*	0.60 - 3.90		**0.56**	(0.170)	**0.74**	(0.224)
	B_sfrP2	4000 - 70000	$My/M_{\odot}$	**11881**	(0.180)	**17250**	* (0.261)
	B_sfrP3	1 - 150	*My*	**15.03**	(0.101)	**29.6**	(0.199)
	B_infall	5000 - 30000	*My*	**5828**	* (0.233)	**6113**	* (0.245)

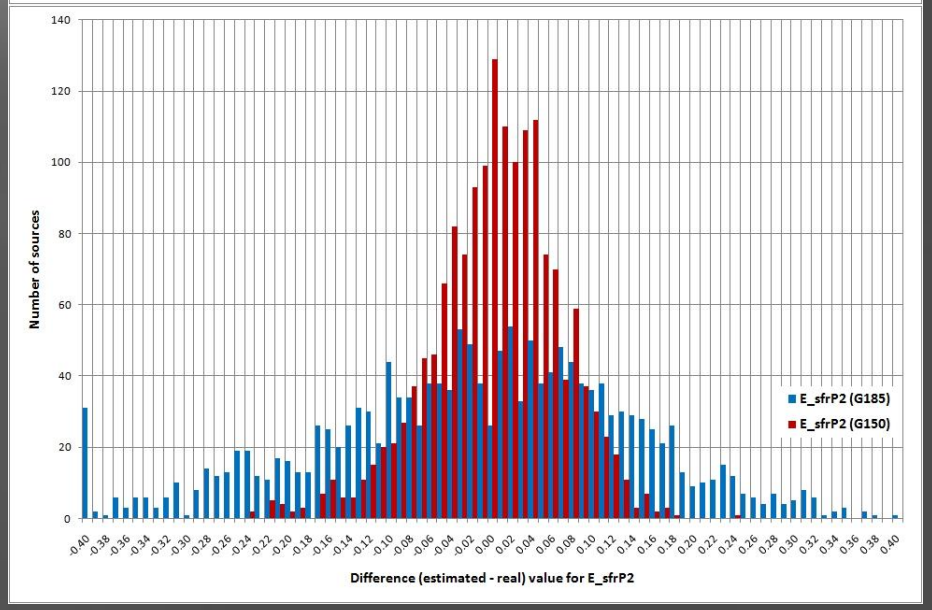
(The performance is given as RMSD and in NRMSD in parentheses)

UGC performance: *E_sfrP1 & E_sfrP2*

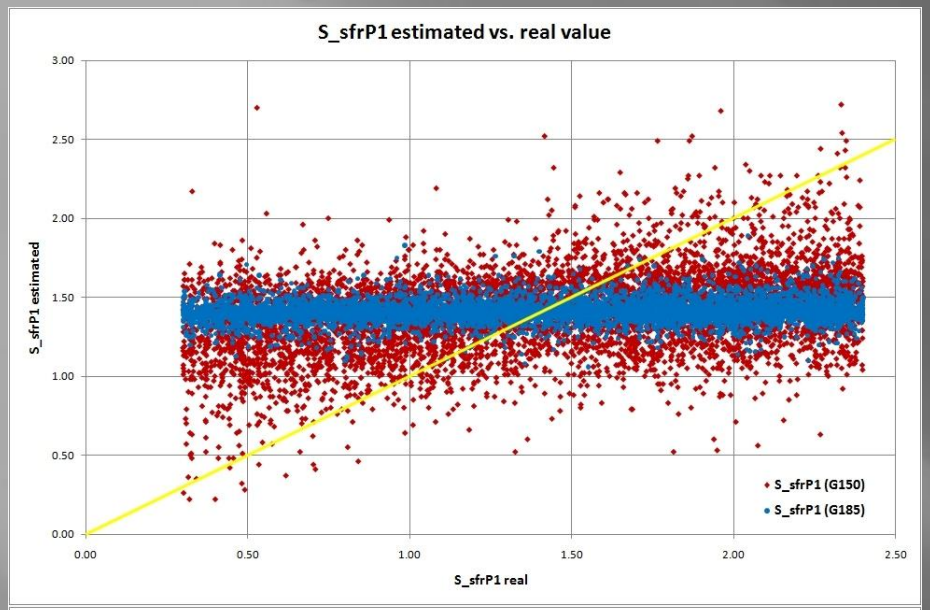


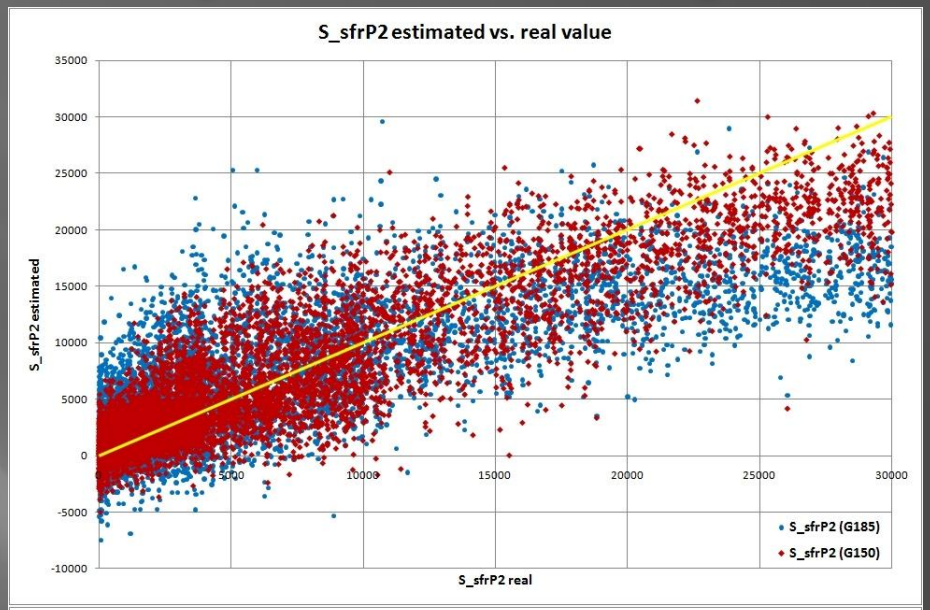


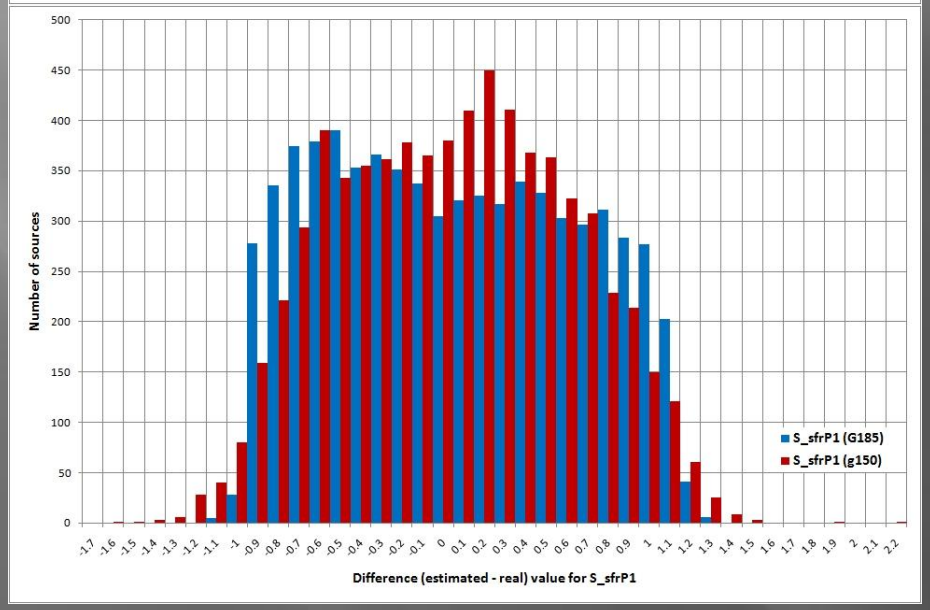


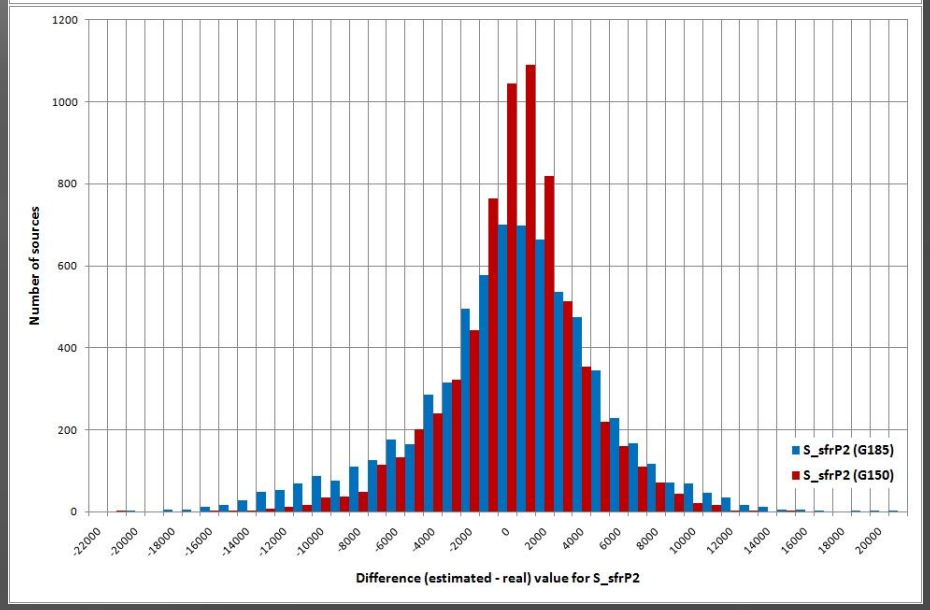


UGC performance: *S_sfrP1 & S_sfrP2*

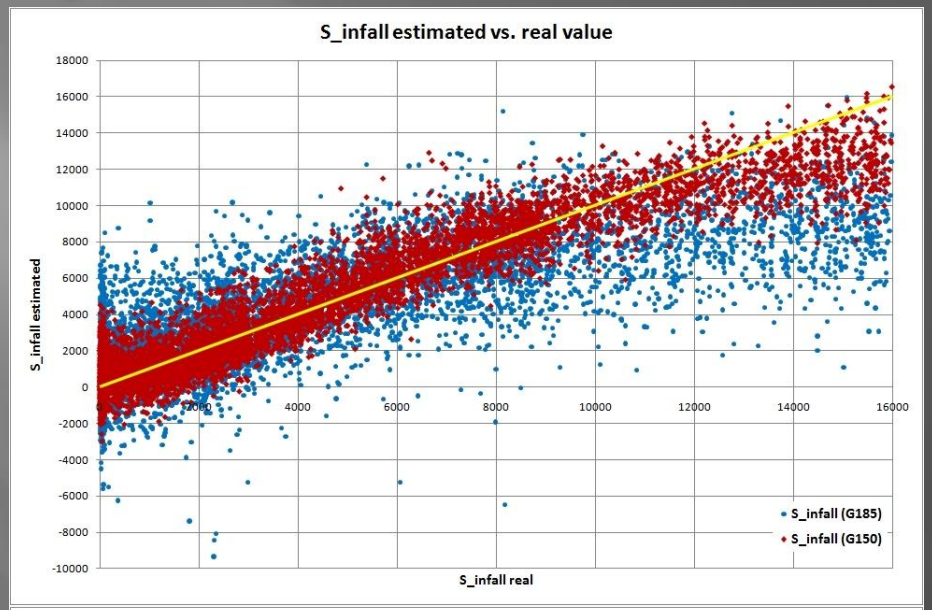


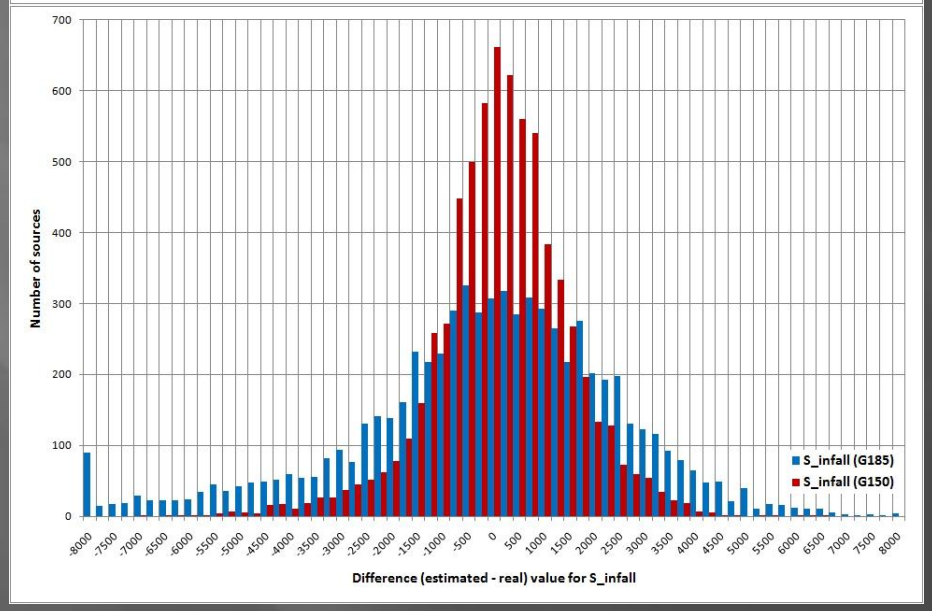




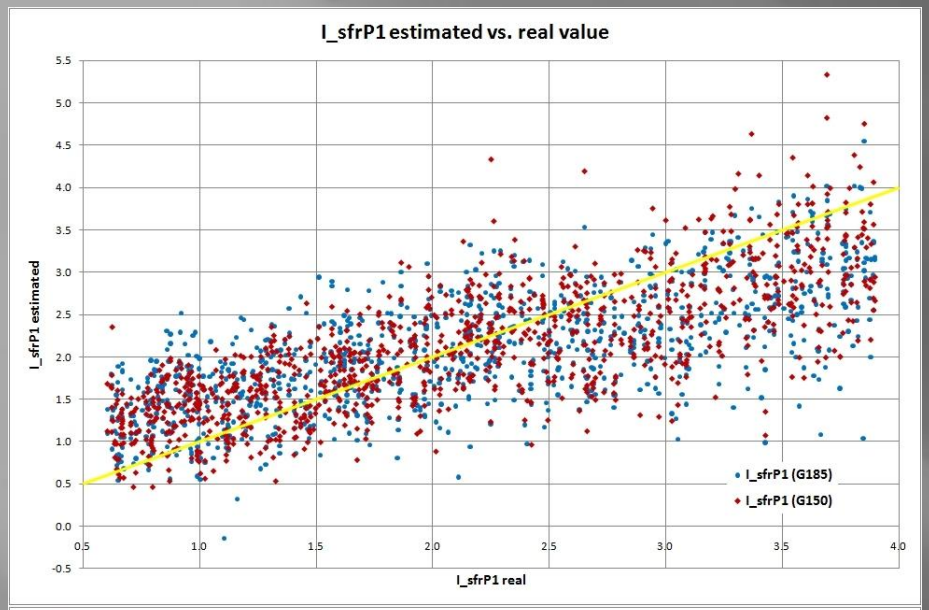


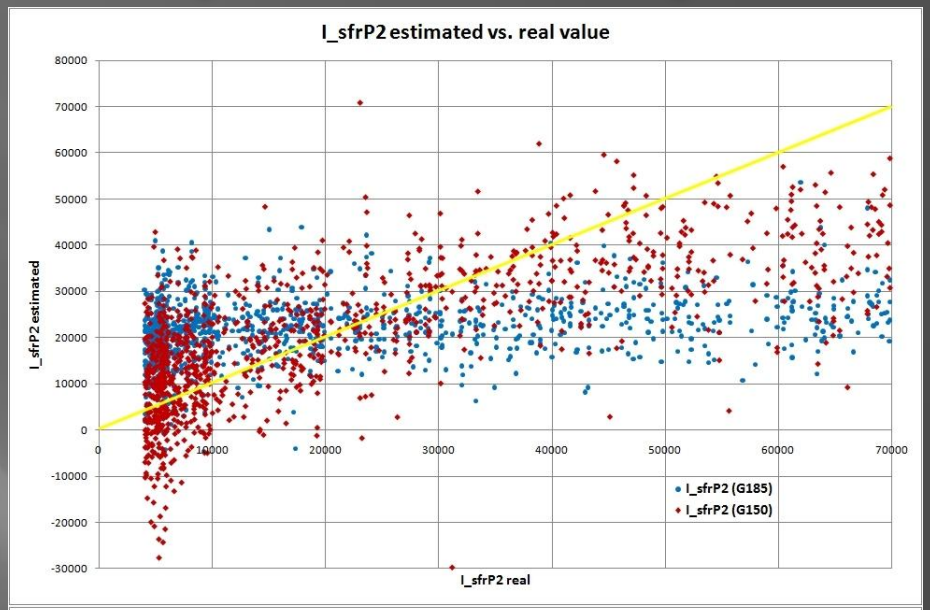
UGC performance: *S_infall*

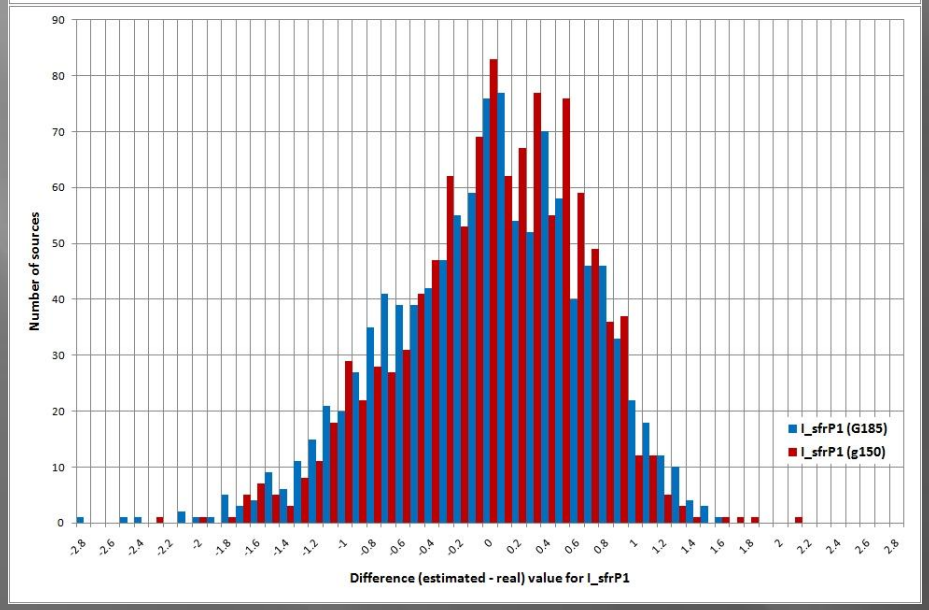


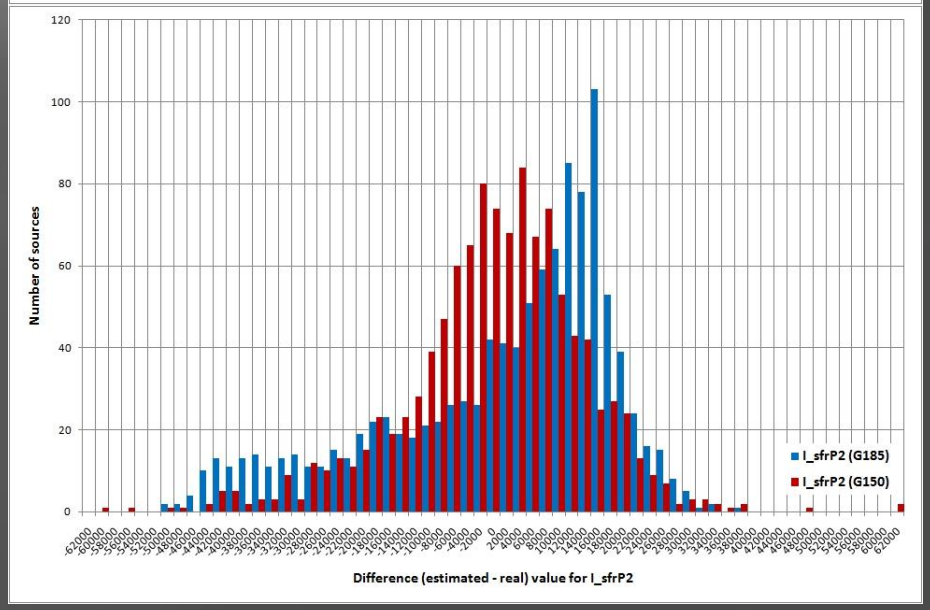


UGC performance: *l_sfrP1 & l_sfrP2*

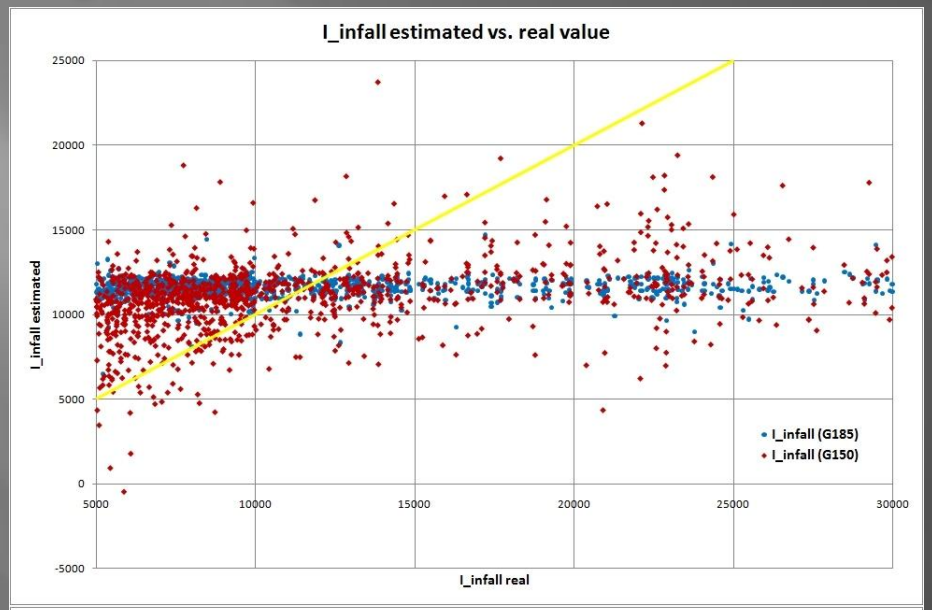


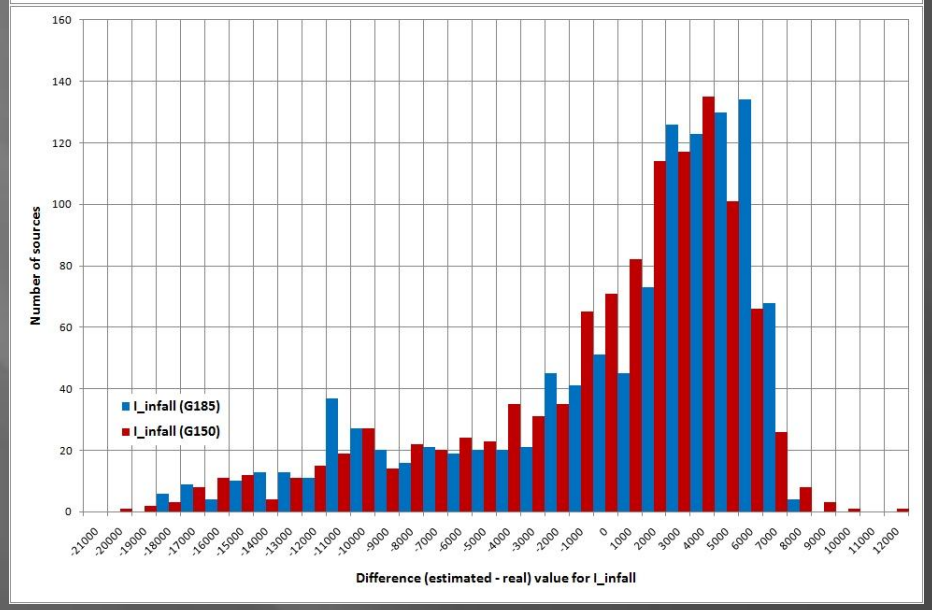




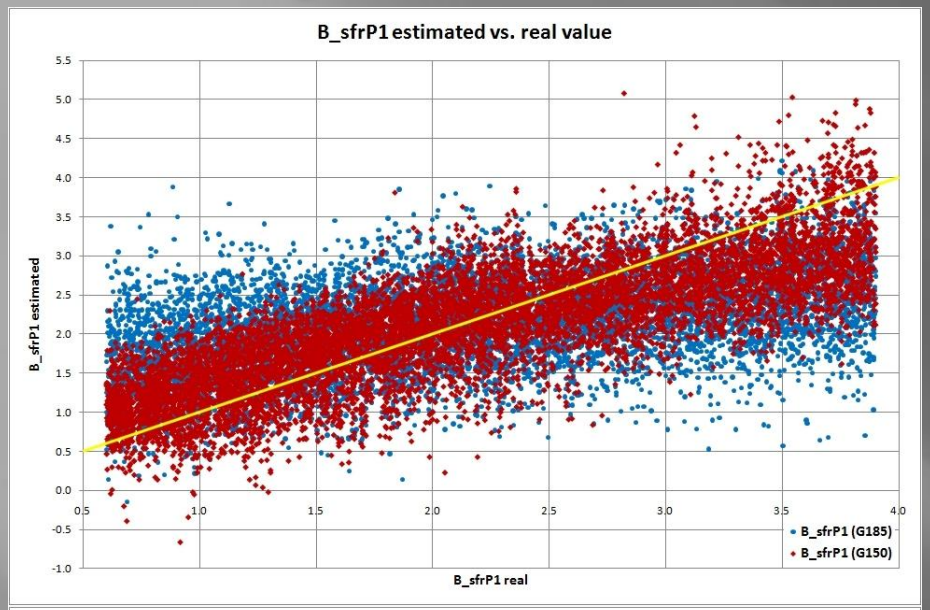


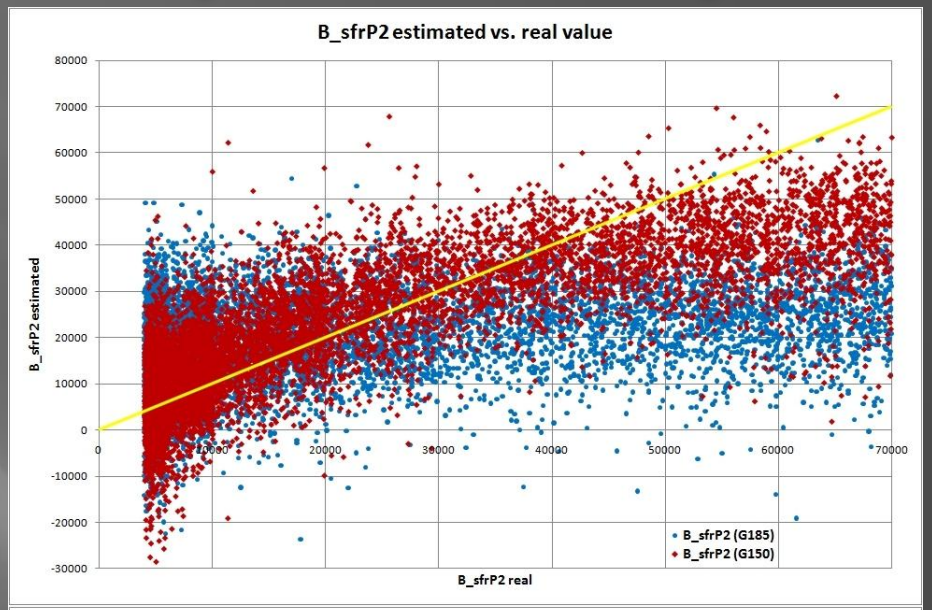
UGC performance: *I_infall*

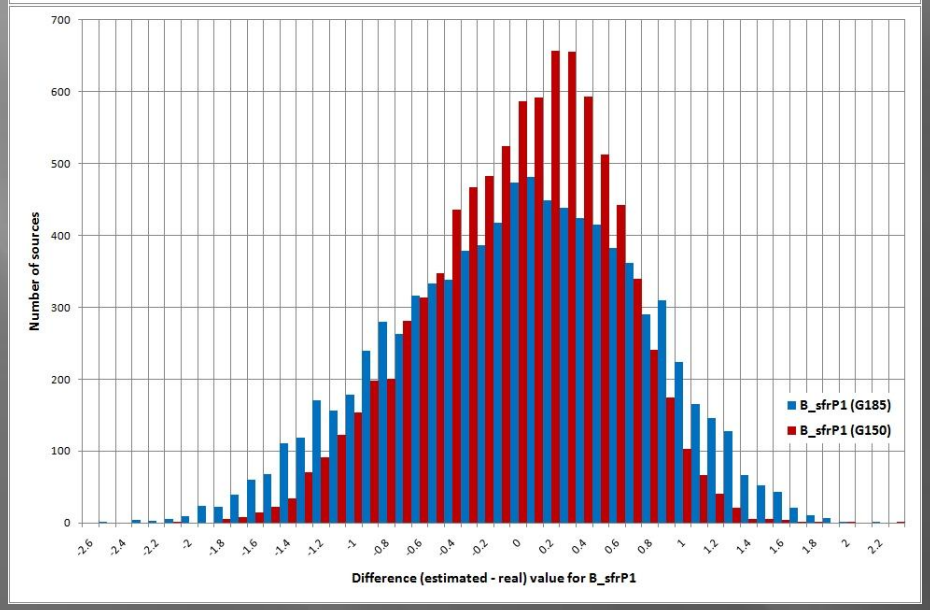


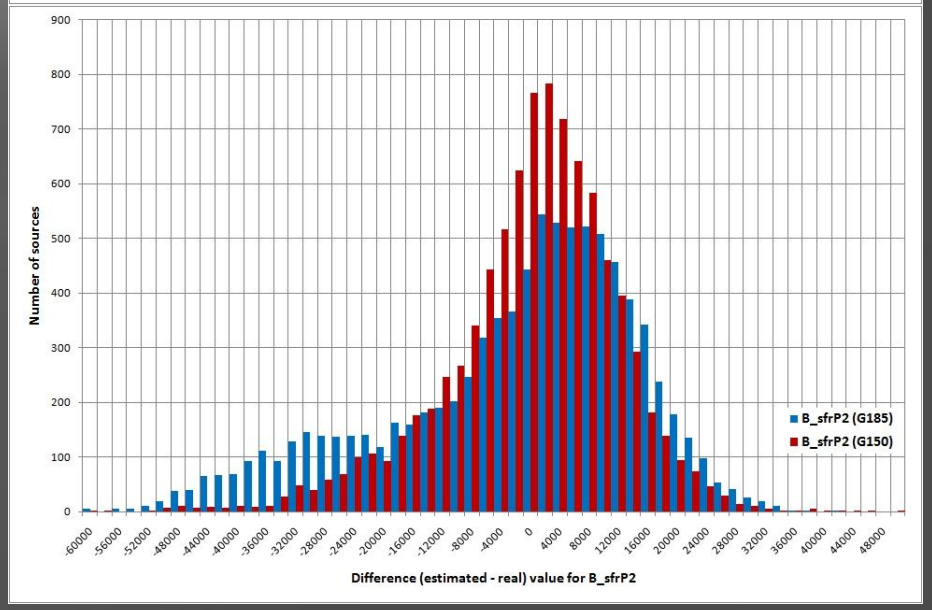


UGC performance: *B_sfrP1 & B_sfrP2*

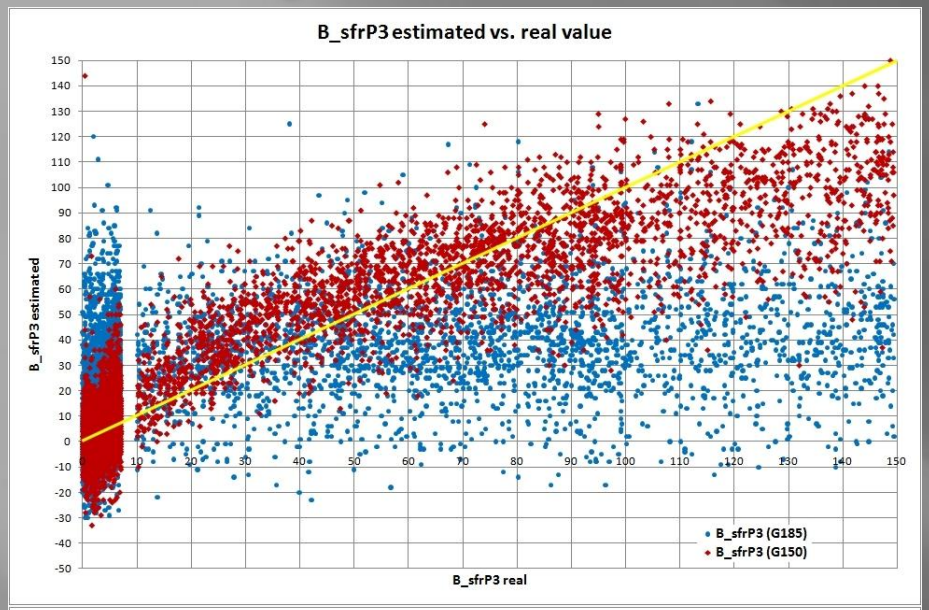


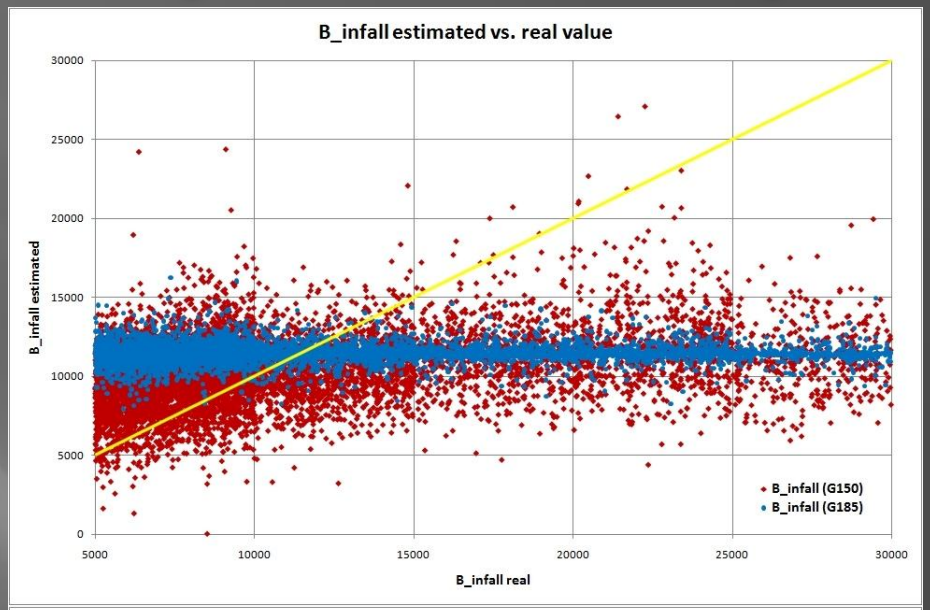


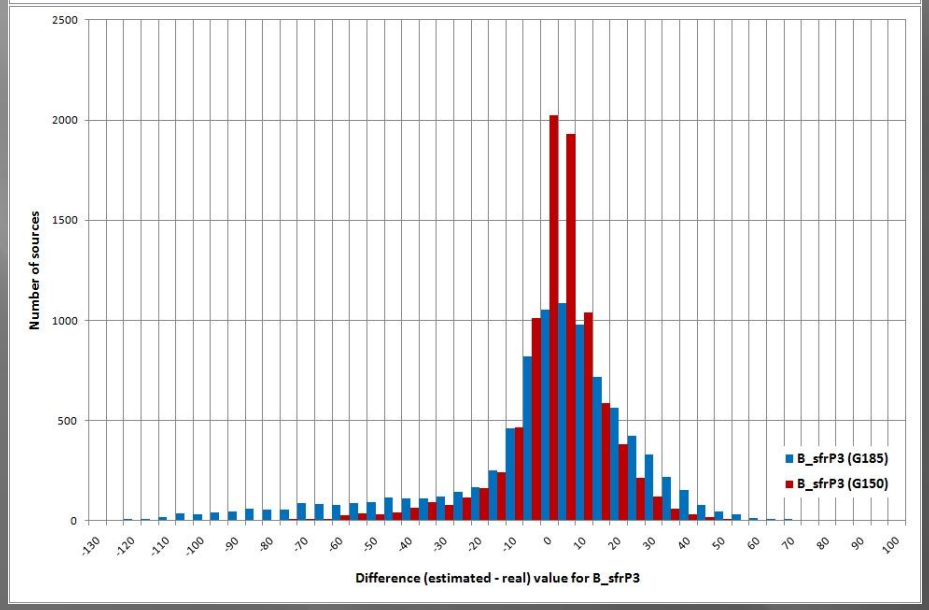


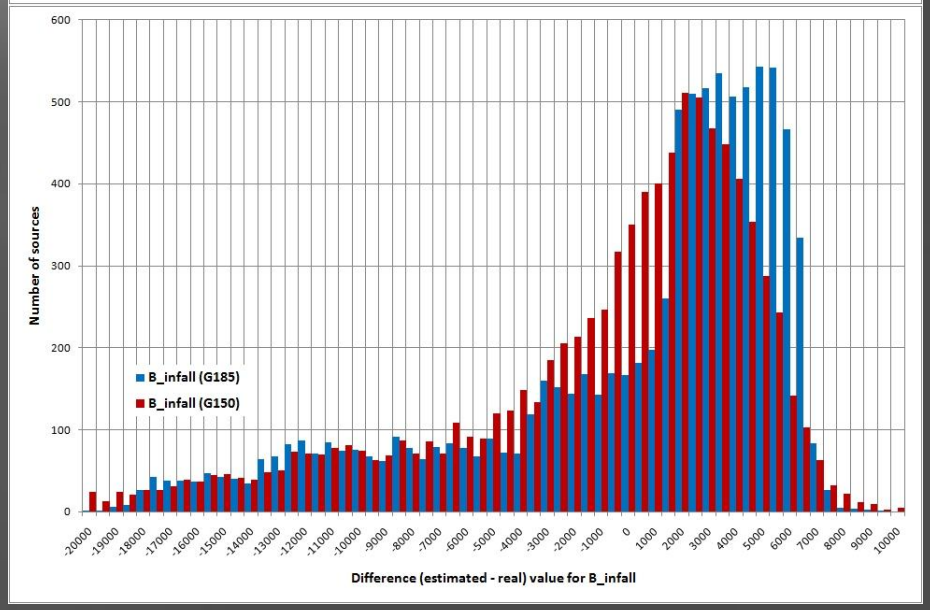


UGC performance: *B_sfrP3 & B_infall*









ANN approach:

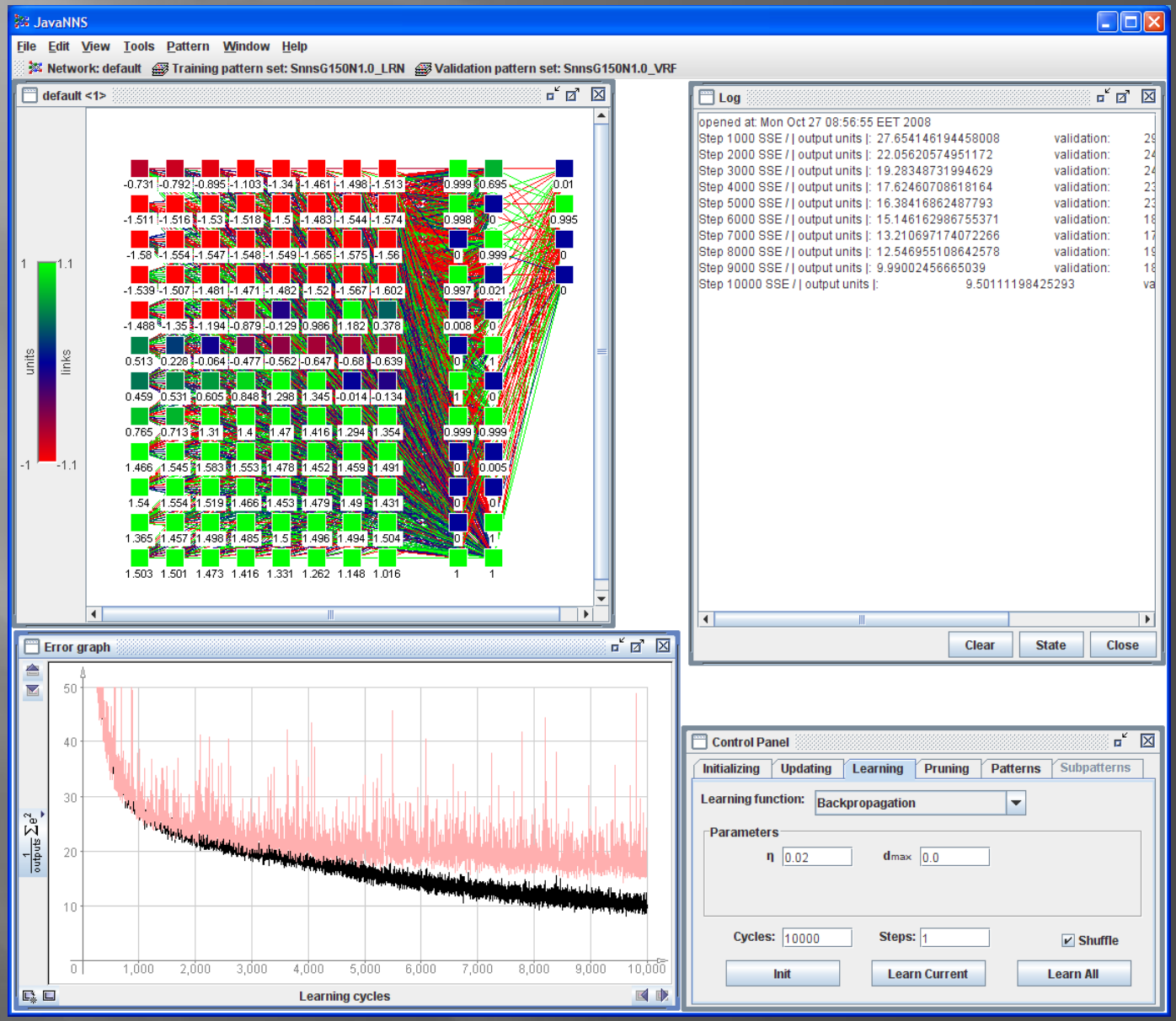
- **Why?**

  could provide supplementary or alternative algorithm to the SVM approach
  - quite different learning procedure
  - an ANN-model can be trained to estimate more than one parameters at once
  - possibly to learn the interdependence of outputs (SF parameters of a galType)
  - smaller number of required ANN-models

- **How?**

  offline training using the JavaNNS, implementation of ANN-application module
  - tests – ANN architecture, activation functions, learning modes, normalization
  - training data – like in UGCv12; currently without subranges in *(z,Ao)* space
  - preliminary tests - results analysis

ANN approach: JavaNNS training



ANN approach: *galType* performance

ANN (G_{mag}=15.0; z and A_0 tota, Lib7 3,3)

(prob>0.5)	*ANN gal_E*	*ANN gal_S*	*ANN gal_I*	*ANN gal_B*	*Unknown*
Real gal_E	**87**	13	0	0	0%
Real gal_S	0	**99**	0	0	0%
Real gal_I	0	11	**76**	13	0%
Real gal_B	0	2	2	**96**	0%

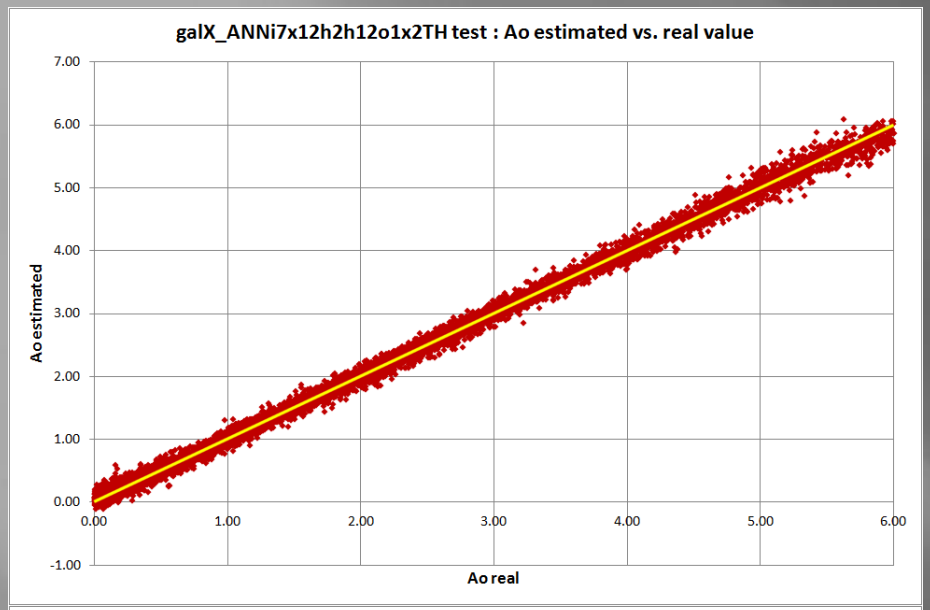
UGC/SVM (G_{mag}=15.0; z and A_0 ranges, Lib7)

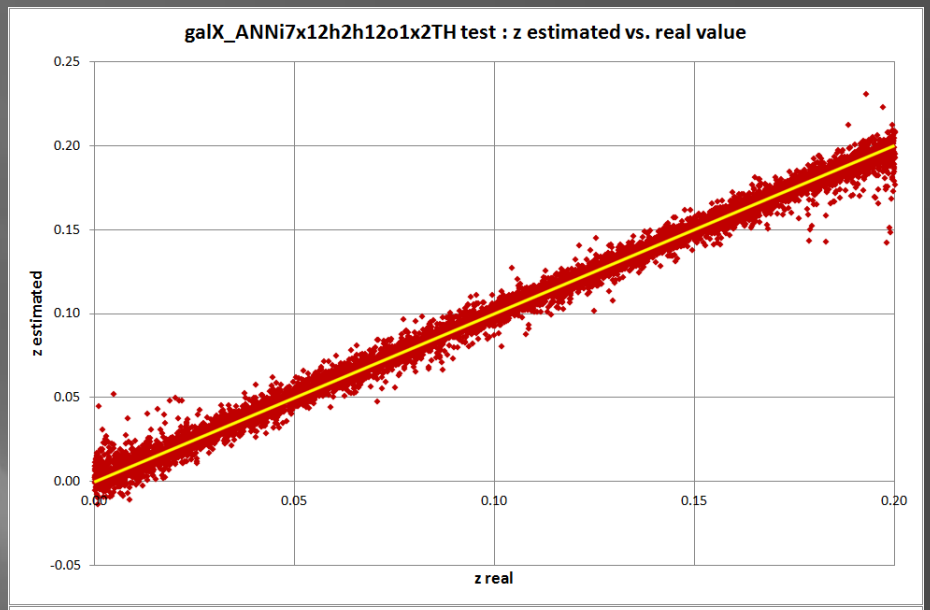
(prob>0.5)	*UGC gal_E*	*UGC gal_S*	*UGC gal_I*	*UGC gal_B*	*Unknown*
Real gal_E	**91**	7	0	2	0%
Real gal_S	1	**95**	0	2	1%
Real gal_I	0	5	**86**	5	4%
Real gal_B	1	0	0	**98**	0%

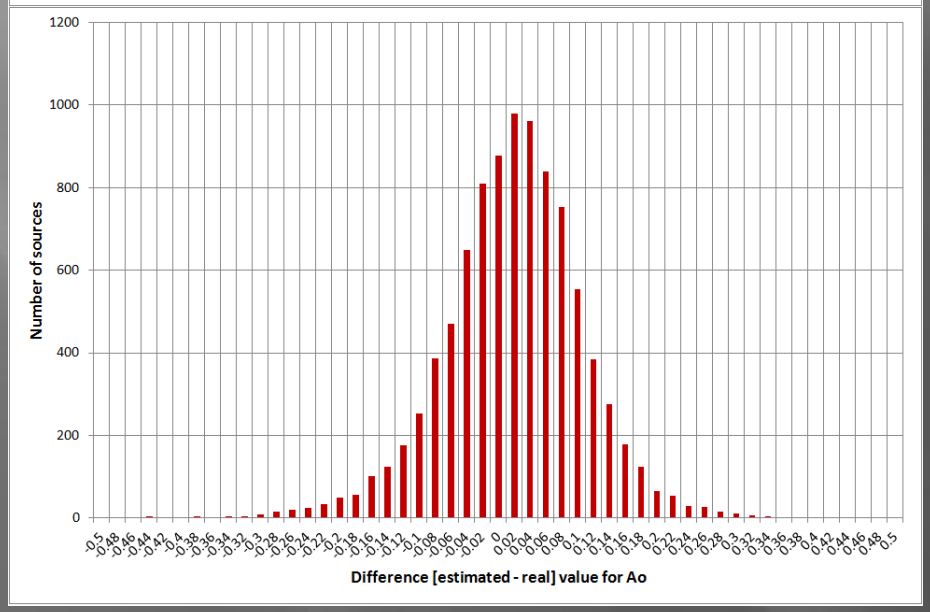
ANN approach: *galParams* performance

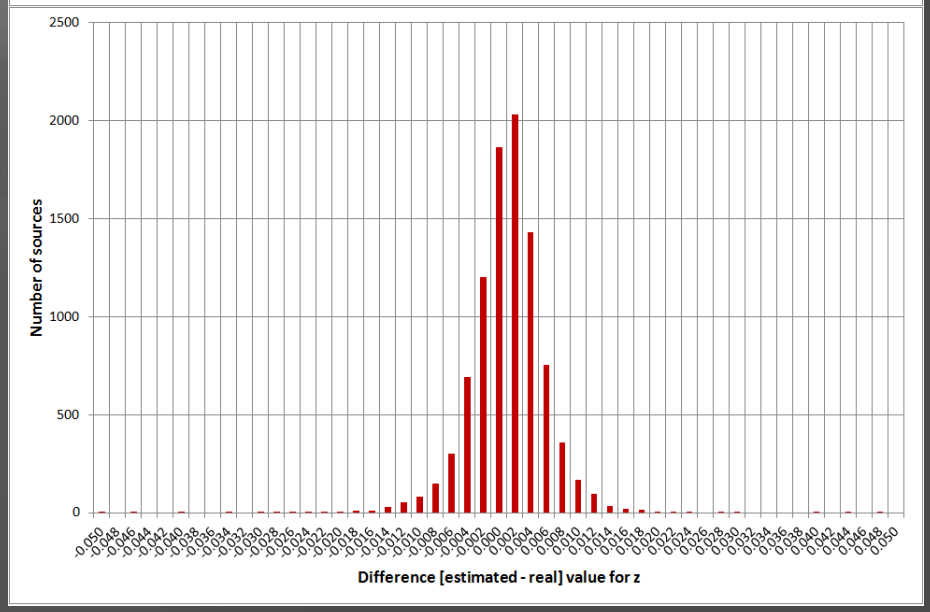
Galaxy model				G_{mag}=15.0 regression performance			
galType	*galParam*	Range	Units	SVM (ranges)		ANN (total)	
(all)	*z*	0.00 – 0.20		0.002		0.005	Lib7 3,3
	Ao	0.0 – 6.0		0.04		0.09	Lib7 3,3
gal_E	*E_sfrP1*	10 - 14400	*My*	696		1017	Lib7 3,3
	E_sfrP2	0.20 - 1.45	$M_{\odot}$	0.06		0.10	Lib7 3,3
gal_S	*S_sfrP1*	0.30 - 2.40		0.56	*	0.61	* Lib7 3,3
	S_sfrP2	5 - 30000	$My/M_{\odot}$	3521		4192	Lib7 3,3
	S_infall	5 - 16000	*My*	1336		1893	Lib7 3,3
gal_I	*I_sfrP1*	0.60 - 3.90		0.61		0.74	* Lib7 3,3
	I_sfrP2	4000 - 70000	$My/M_{\odot}$	14400		19789	* Lib7 3,3
	I_infall	5000 - 30000	*My*	5759	*	6813	* Lib7 3,3
gal_B	*B_sfrP1*	0.60 - 3.90		0.56		0.62	Lib7 3,3
	B_sfrP2	4000 - 70000	$My/M_{\odot}$	11881		13752	Lib7 3,3
	B_sfrP3	1 - 150	*My*	15.03		20.5	Lib7 3,3
	B_infall	5000 - 30000	*My*	5828	*	6000	* Lib7 3,3

ANN tests: *Ao* & *z* performance (G150)

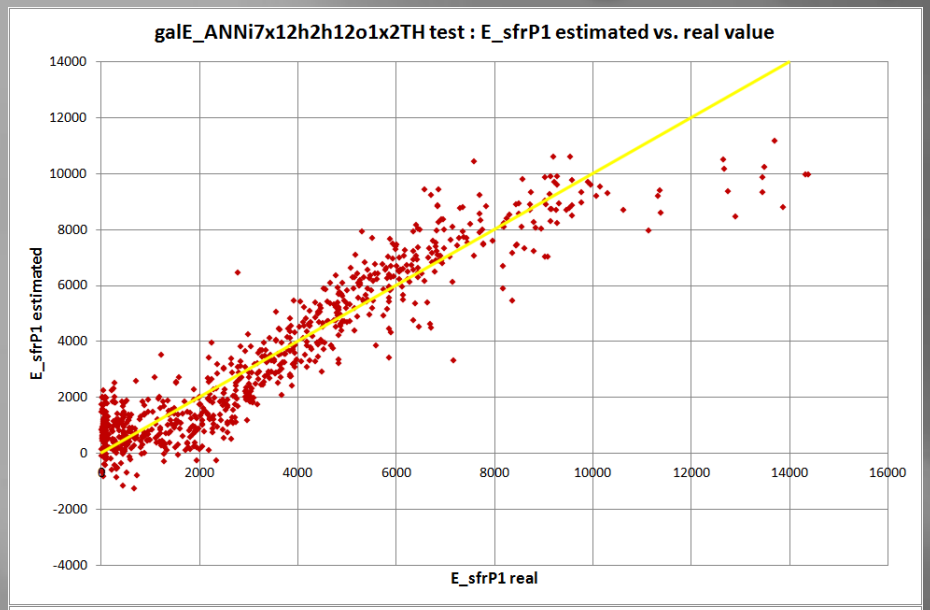


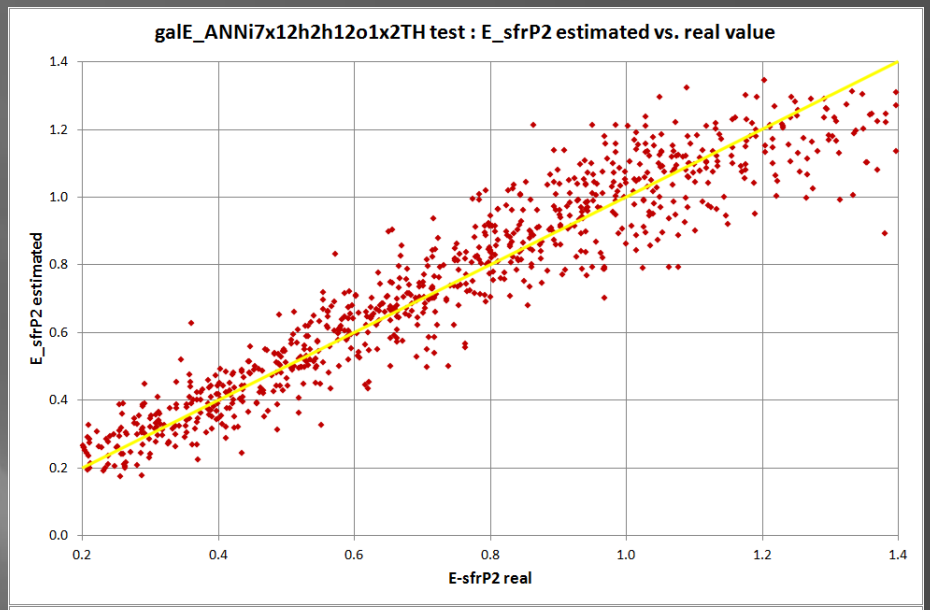


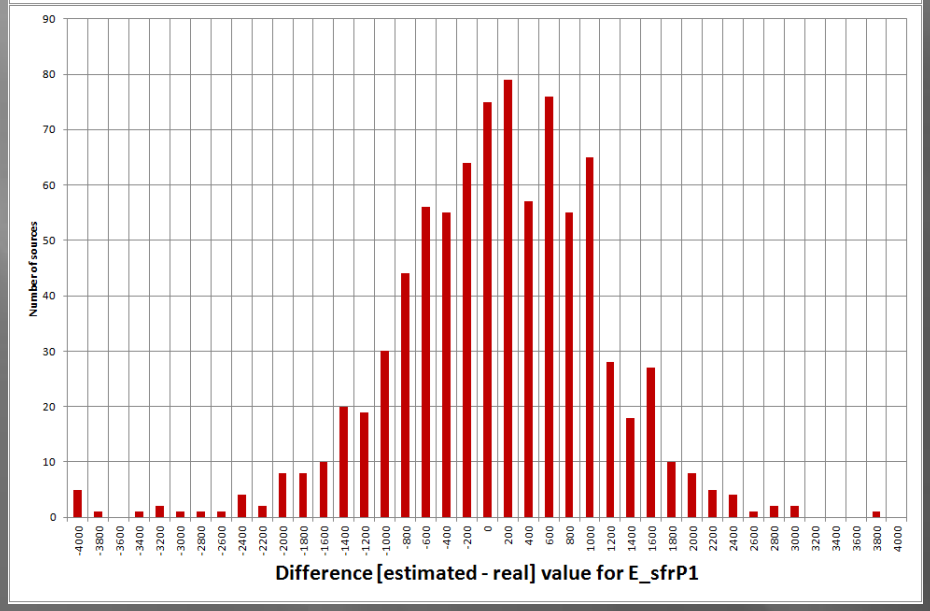


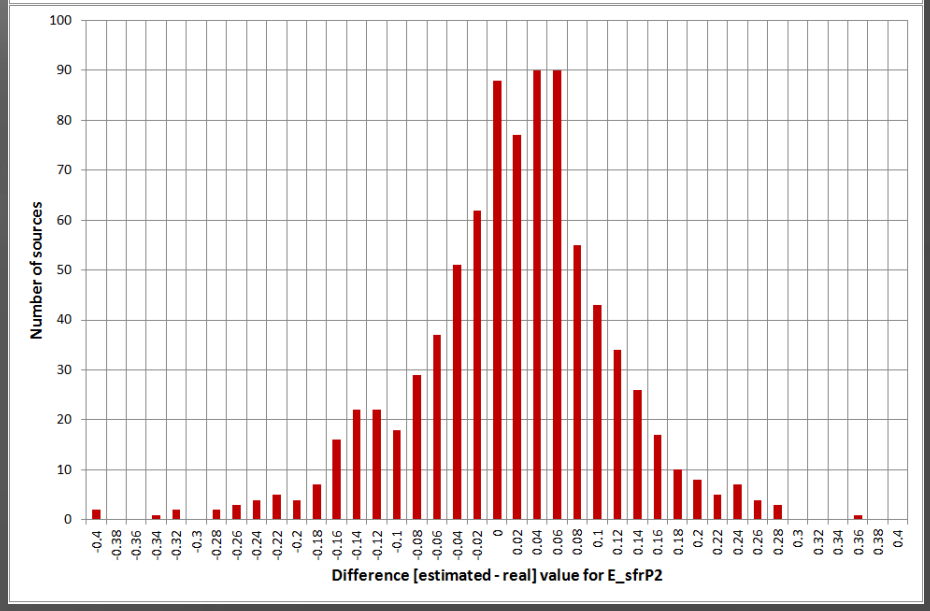


ANN tests: *gal_E* performance (G150)

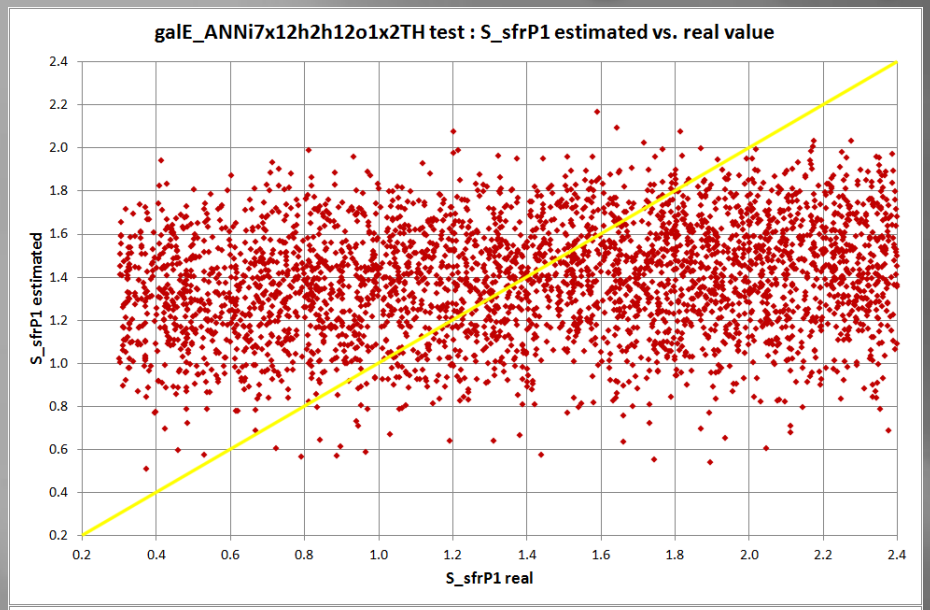


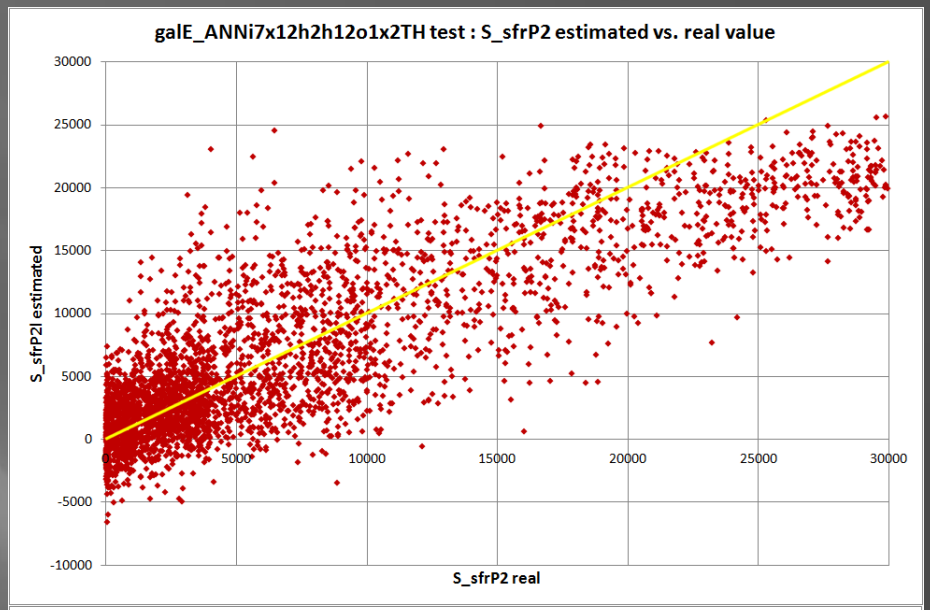


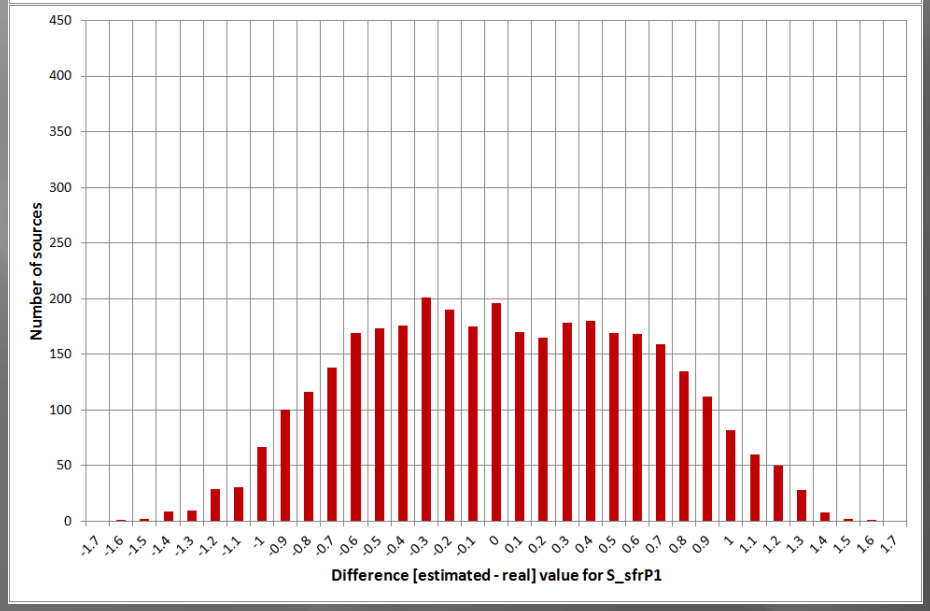


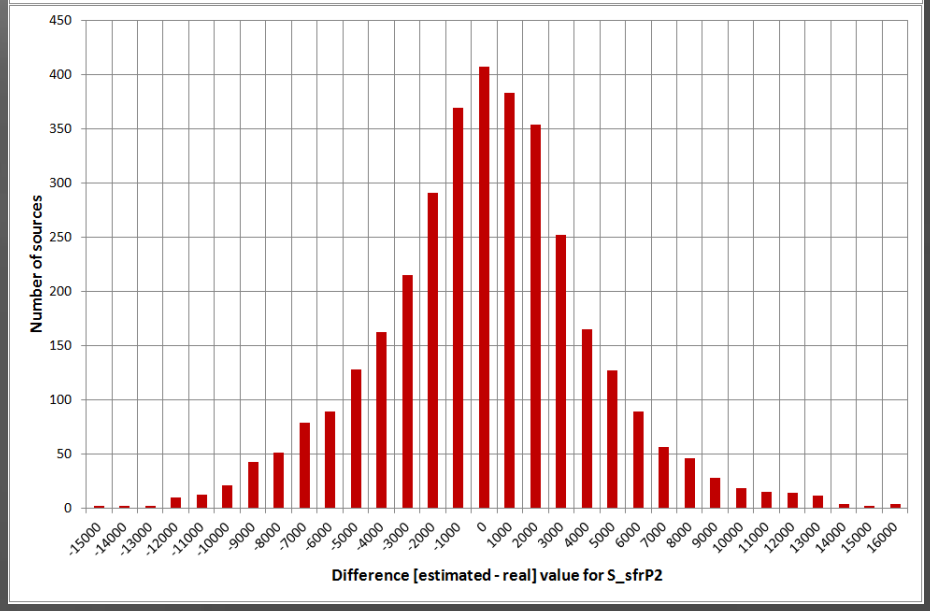


ANN tests: *gal_S* performance (G150)

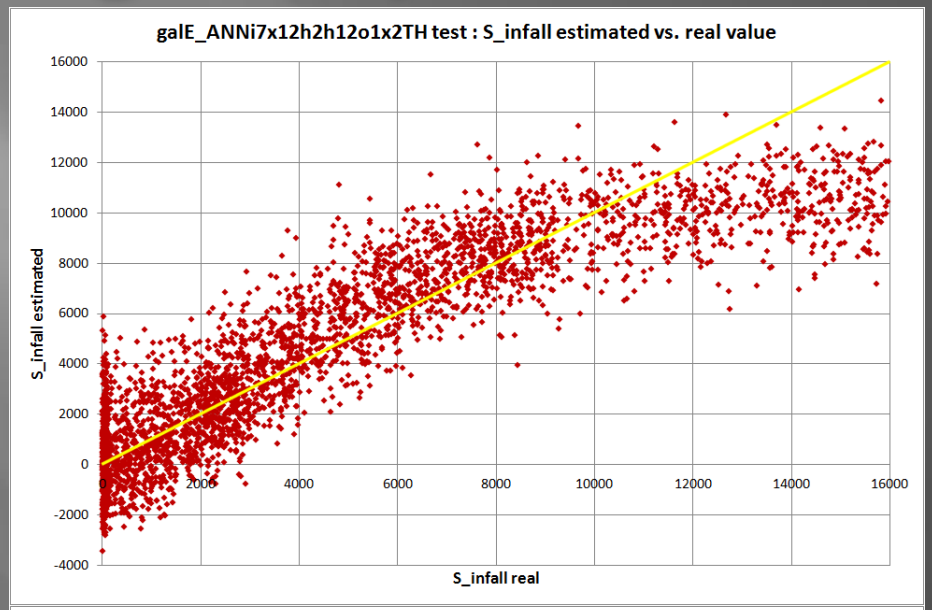


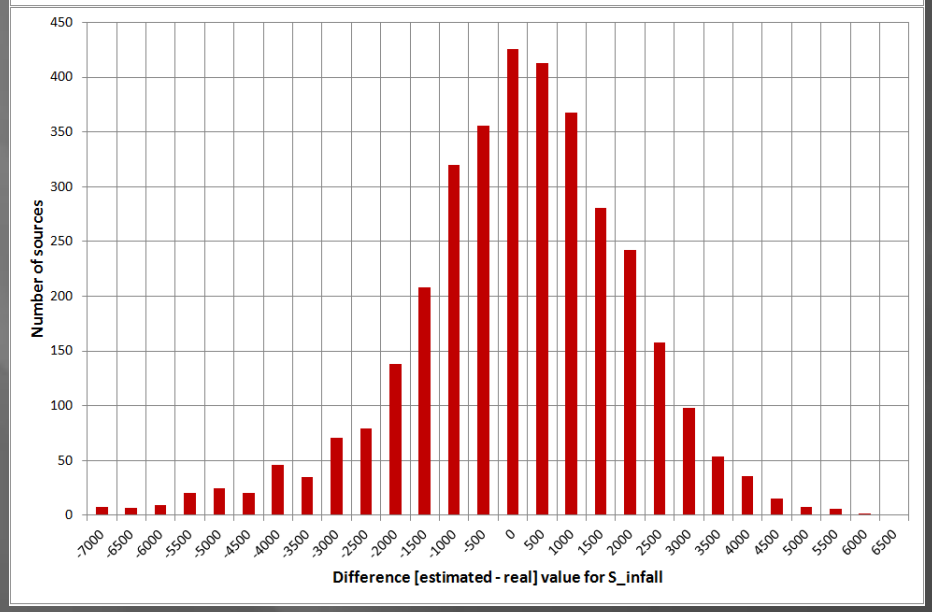




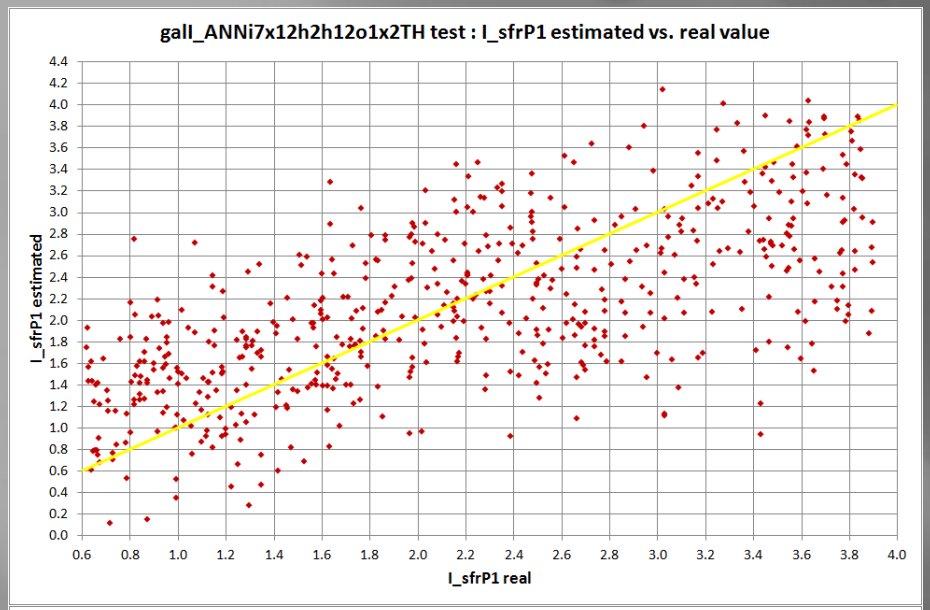


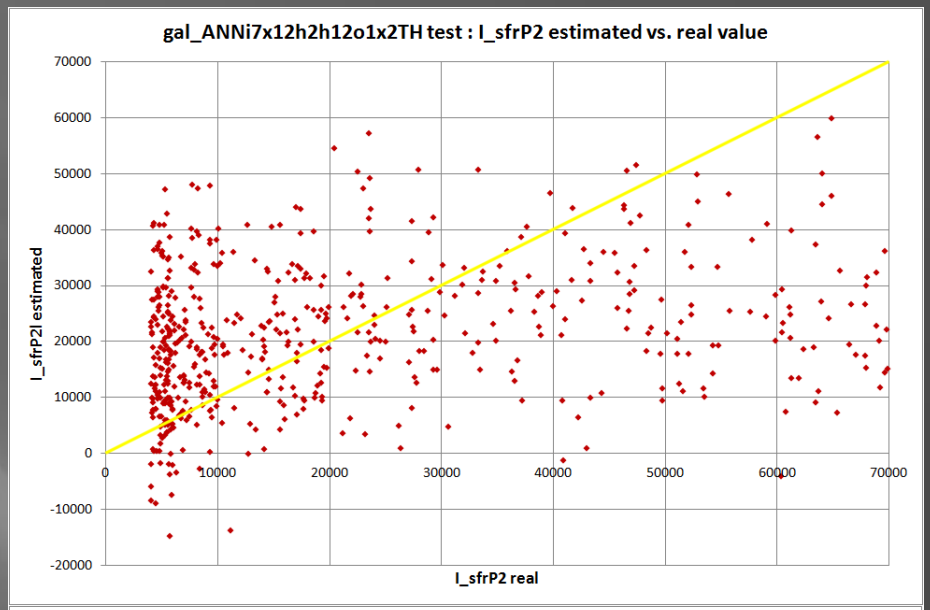
ANN tests: *gal_S* performance (G150)

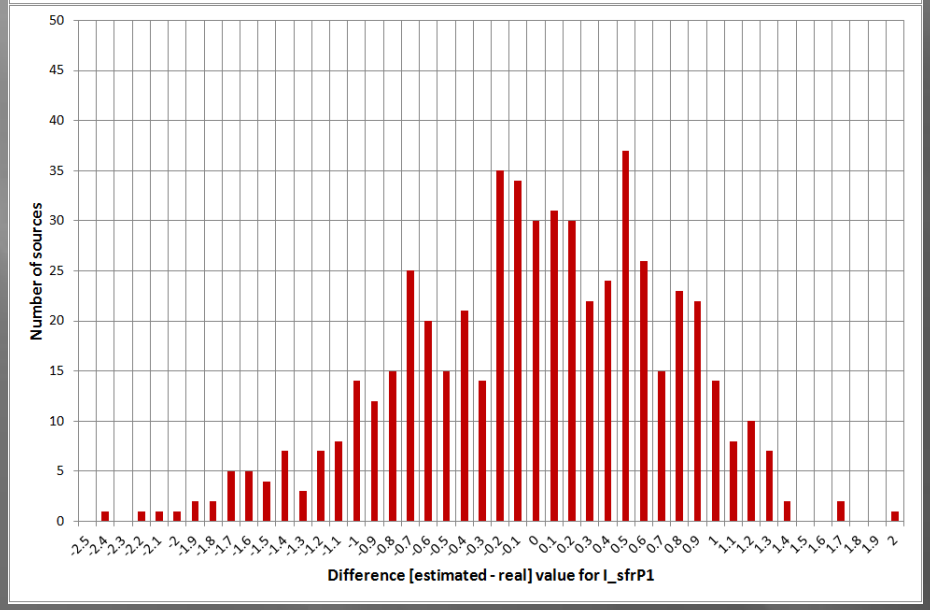


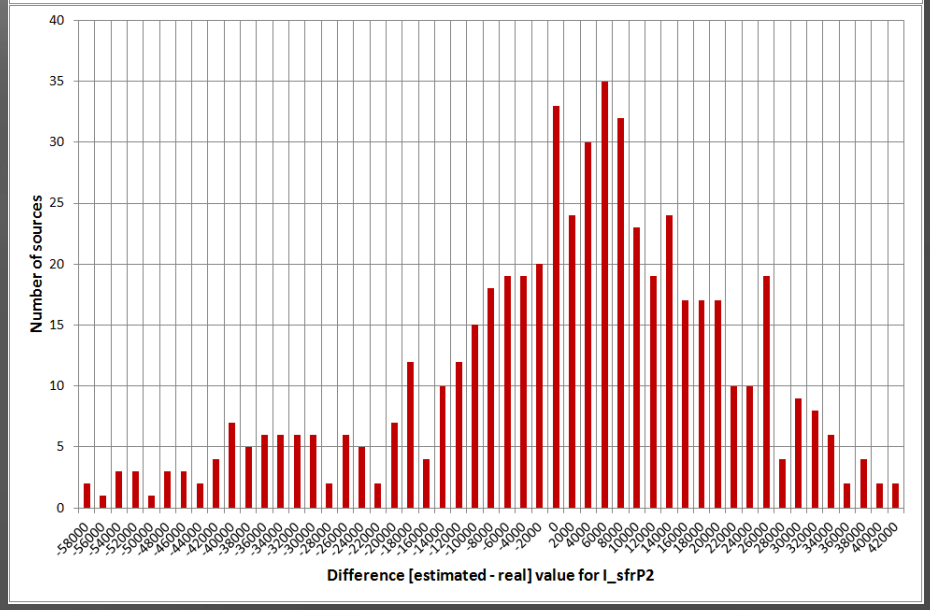


ANN tests: *gal_I* performance (G150)

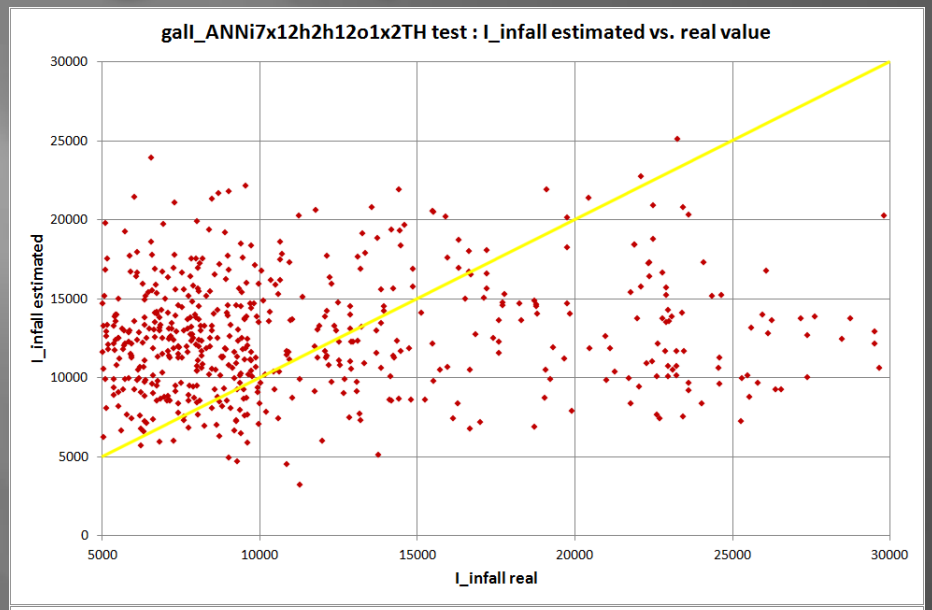


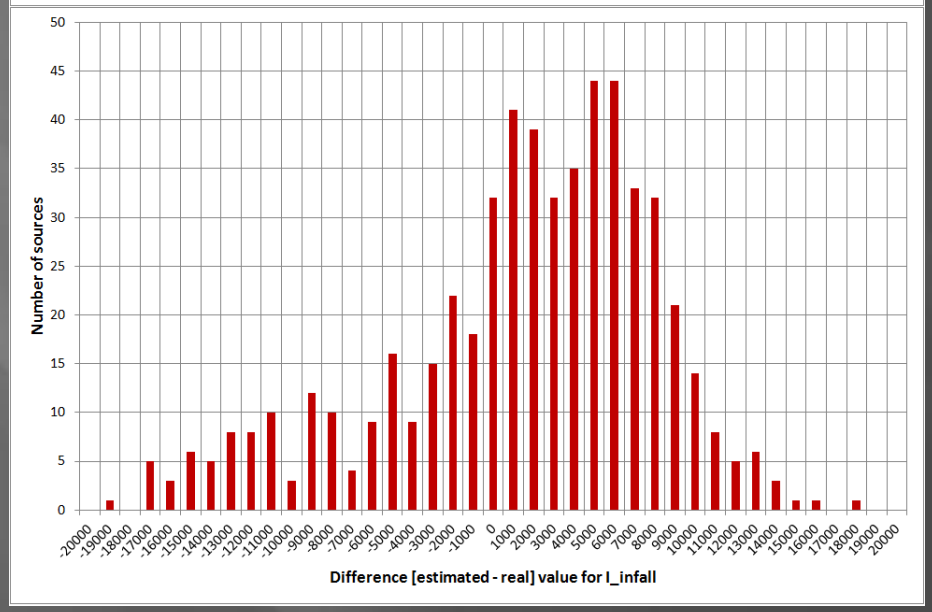




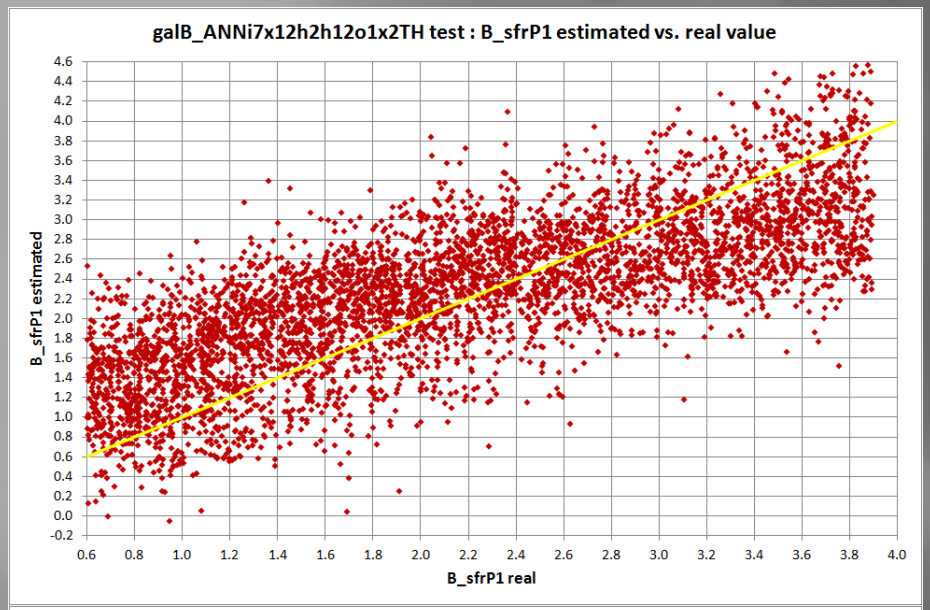


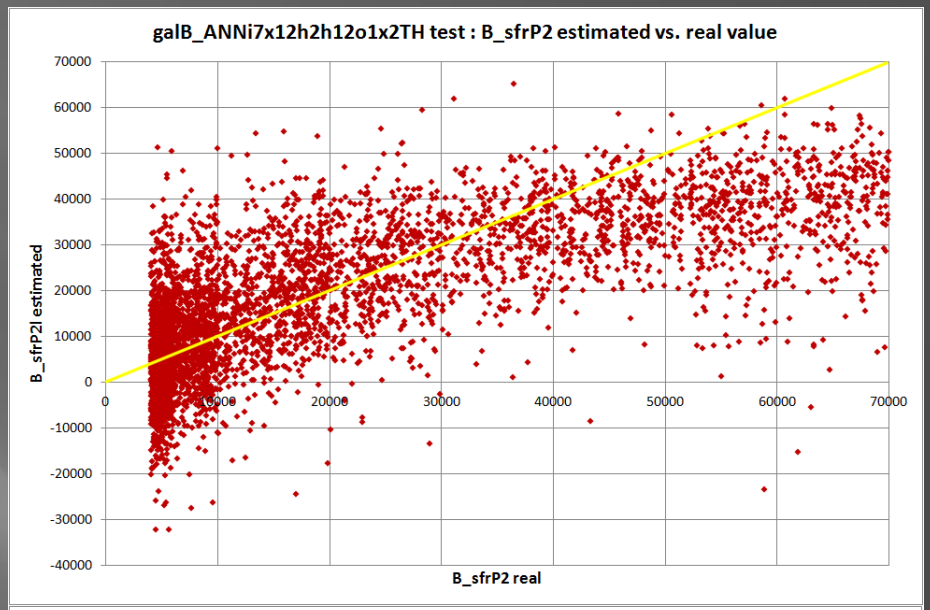
ANN tests: *gal_I* performance (G150)

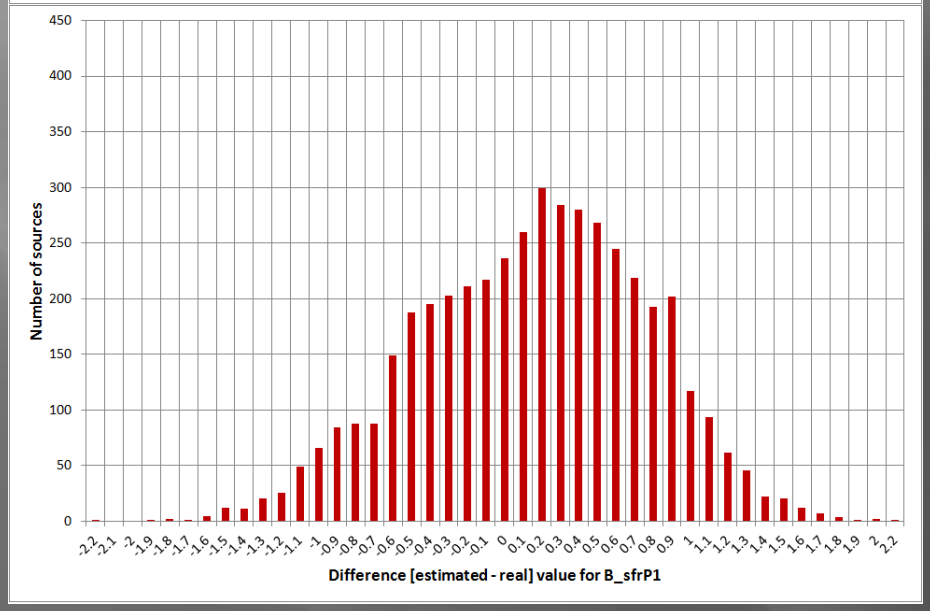


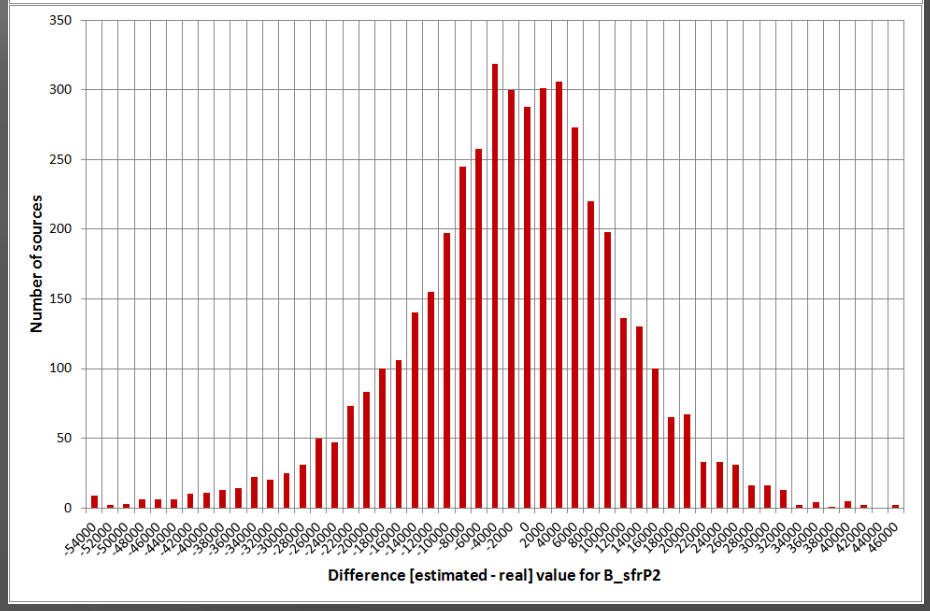


ANN tests: *gal_B* performance (G150)

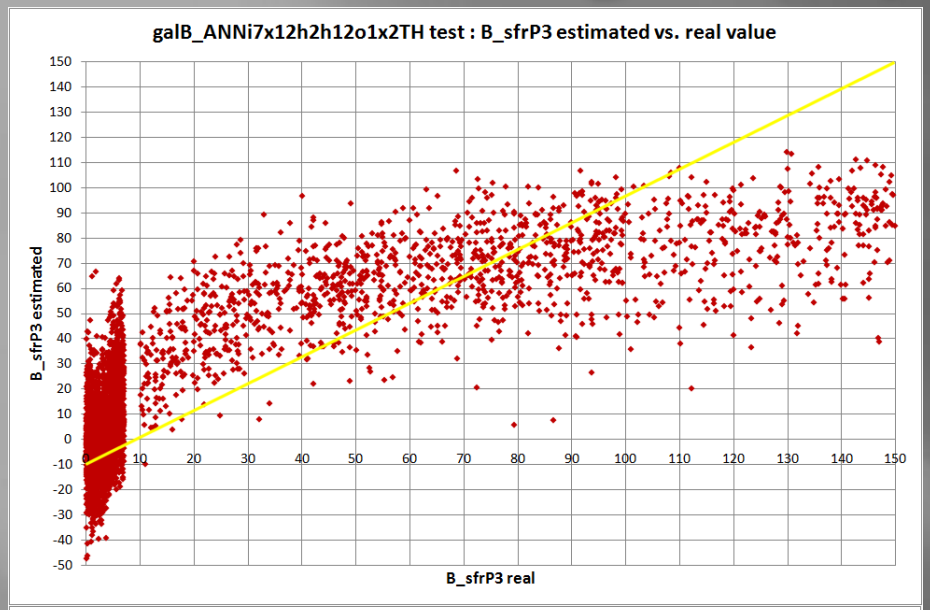


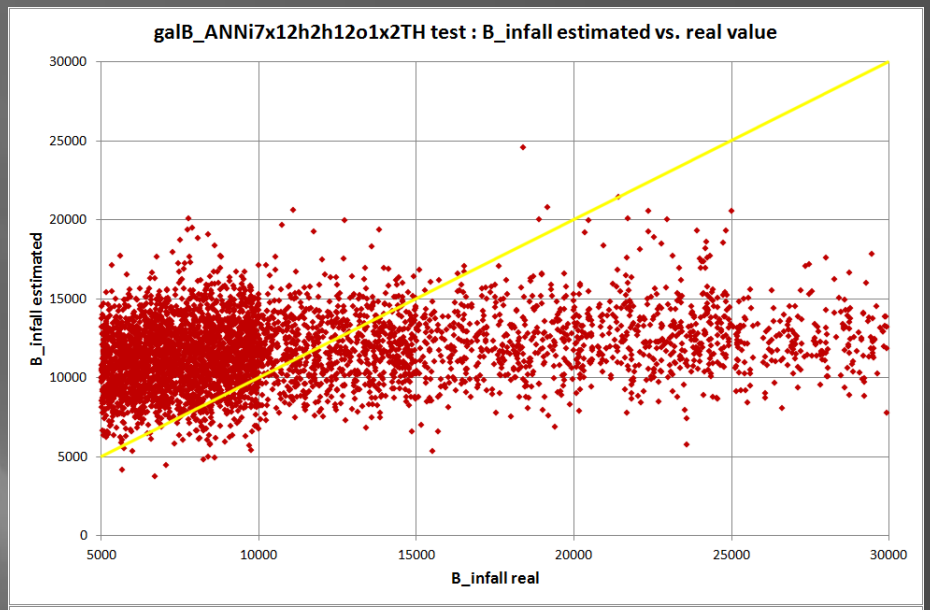


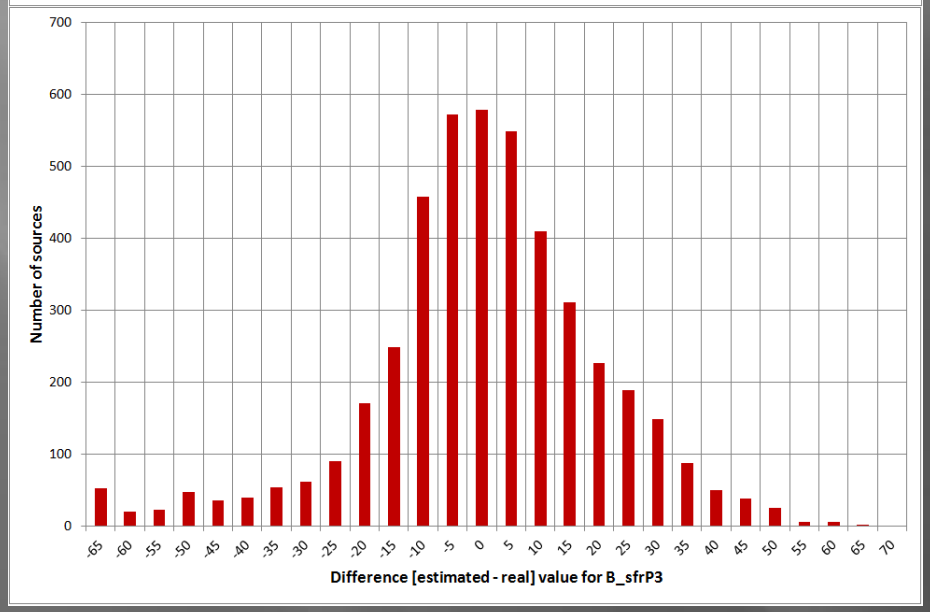


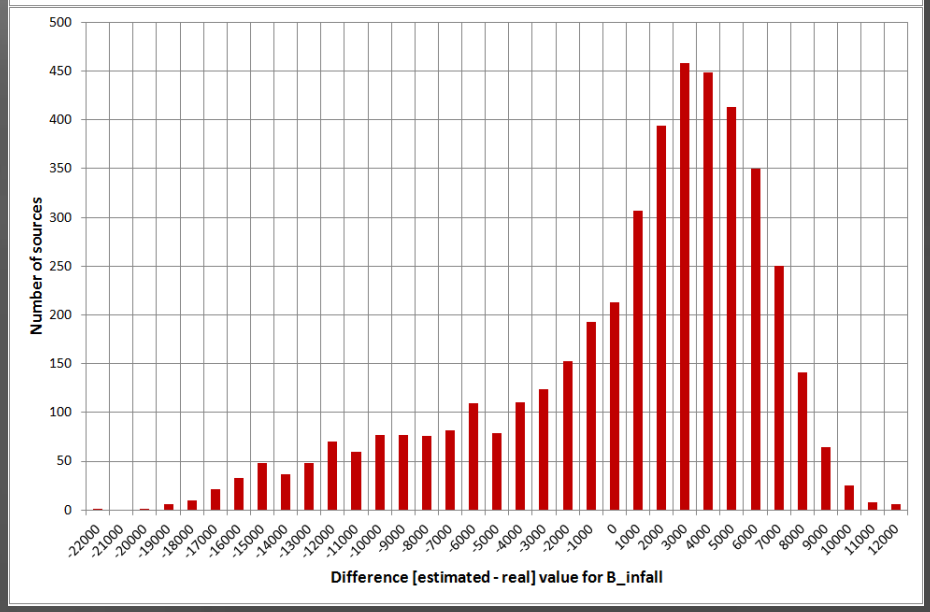


ANN tests: *gal_B* performance (G150)









UGC for Gaia – development

- **UPDATED GALAXY MODEL**
  - Evolutionary correction
  - Geometry and inclination parameters
  - LRG galaxy type (new)

- **SVM VS ANN APPROACH**
  - Compare classification performance
  - Compare regression performance on external (z and Ao) parameters
  - Compare regresson performance on SF parameters
  - Implementation of SVM and/or ANN models

- **ALGORITHM UPDATES**
  - Optimal subranges in Ao (main range 0.0-0.5)
  - G_{mag} vs S/N ranges
  - Combine B-type and I-type galParams (common SVM-models)
  - Parameters nonlinear preprocessing
  - Derived parameters

Acknowledgments

The UGC is supported by the ESA/PRODEX programme for Greece (2012-2014).

The synthetic spectra are produced using Pegase.2 spectral evolution model of galaxies developed by Brigitte Rocca-Volmerange and Michel Fioc (IAP).

Simulated data provided by the Simulation Unit (CU2) of the Gaia Data Processing Analysis Consortium (DPAC) have been used to complete this work. The simulations have been done in the supercomputer MareNostrum at Barcelona Supercomputing Center - Centro Nacional de Supercomputacion (The Spanish National Supercomputing Center). They are gratefully acknowledged for this contribution.

We thank Antonella Vallenari and Rosanna Sordo, members of the CU8 of DPAC, for their work in organization and preparation the data for simulation.

The South-Eastern Europe Grid infrastructure (SEE-GRID), part of the European Grid Infrastructure, has been used to complete the tuning, training and testing of the large number of SVM-models.

Thank you!

