北京大学
PEKING UNIVERSITY

天文信息学与虚拟天文台2025年学术年会，海口，2025/11/27-29

机器学习与类星体巡天

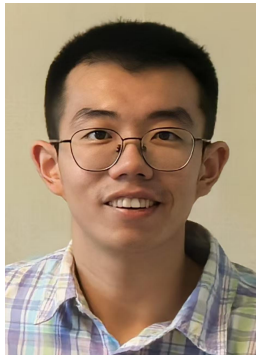
吴学兵、傅煜铭、庞宇萱、朱睿 等
（北京大学、莱顿大学、中国科学院大学）

2025-11-27

内容

1. 类星体巡天与机器学习
2. 迁移学习与银道面背景类星体巡天
3. XGBoost 与CatNorth/CatGlobe/QHSC 类星体星表
4. 机器学习与CSST类星体巡天
5. 总结


傅煜铭

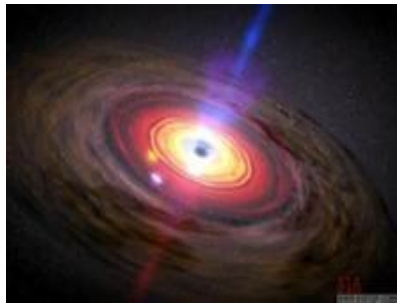

庞宇萱


朱睿

Fu, Wu et al. 2021, ApJS, 254, 6;
Fu, Wu et al. 2022, ApJS, 261, 32;
Fu, Wu et al. 2024, ApJS, 271, 54;
Fu, Wu et al. 2025, ApJS, 279, 54;

Pang, Wu et al. 2025, ApJ, 980, 223;
Pang, Wu et al. 2025, MNRAS, 540,2216
Zhu, Wu et al. 2025, ApJS, in press

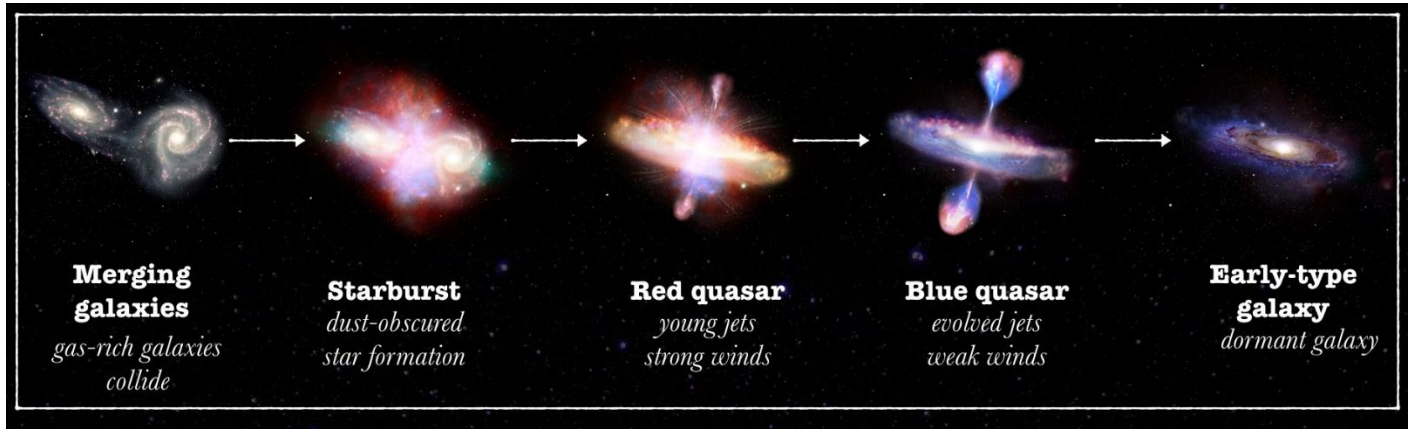
1. 类星体巡天与机器学习



北京大学 PEKING UNIVERSITY

星系与超大质量黑洞的形成与演化

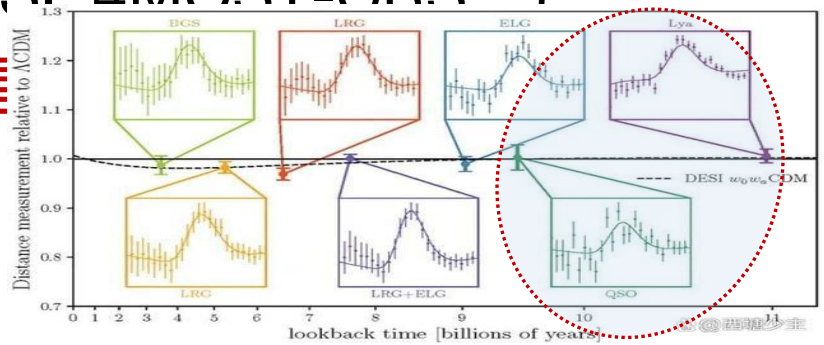
- 完备类星体样本对类星体光度函数、黑洞质量函数的精确测定至关重要
- 能谱较红的类星体: 完善星系到类星体演化历程的关键拼图 (Glickman et al. 2012; Klindt et al. 2019; Fujimoto et al. 2022)
- 射电强类星体、宽吸收线类星体：喷流、外流与AGN反馈



https://ras.ac.uk/sites/default/files/2019-

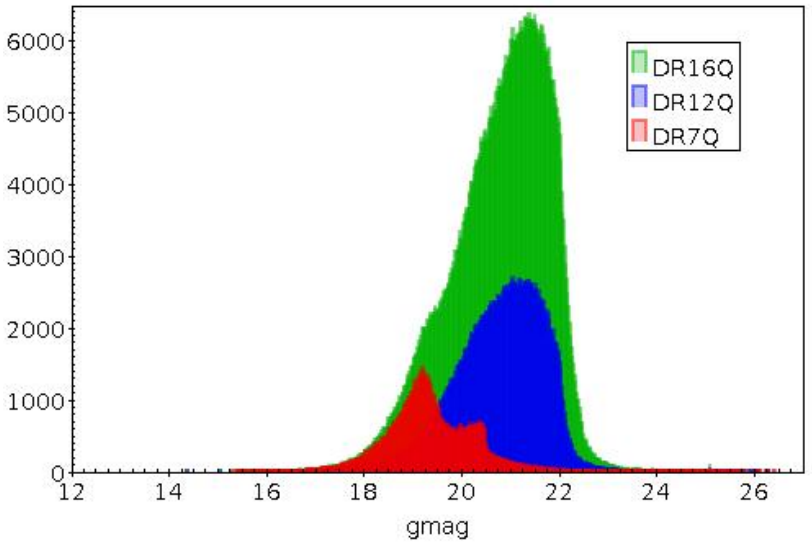
宇宙学研究

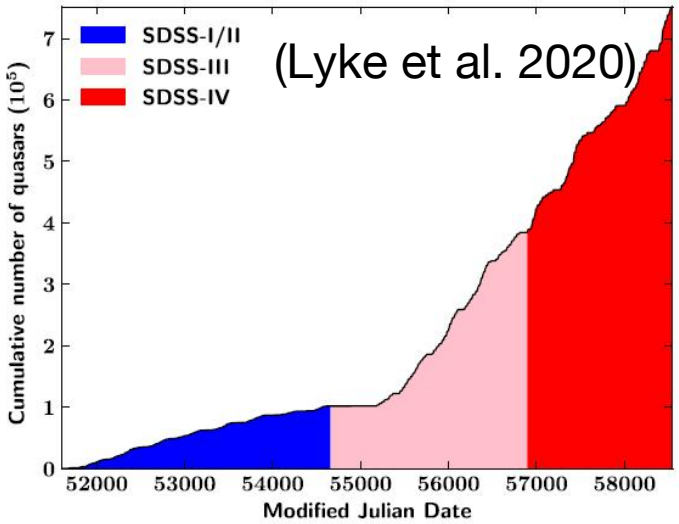
- 类星体作为高红移宇宙的探针 (Risaliti & Lusso 2015, 2019; Risaliti et al. 2023)
- 对类星体的 Sandage-Loeb 检验 (Sandage 1962; Loeb 1998) 直接测量哈勃常数 (QUBRICS; Cupani et al. 2022)
- 示踪暗物质和宇宙大尺度结构 (SDSS, DESI, 4MOST-CRS,)
- 视星



DESI之前的类星体巡天

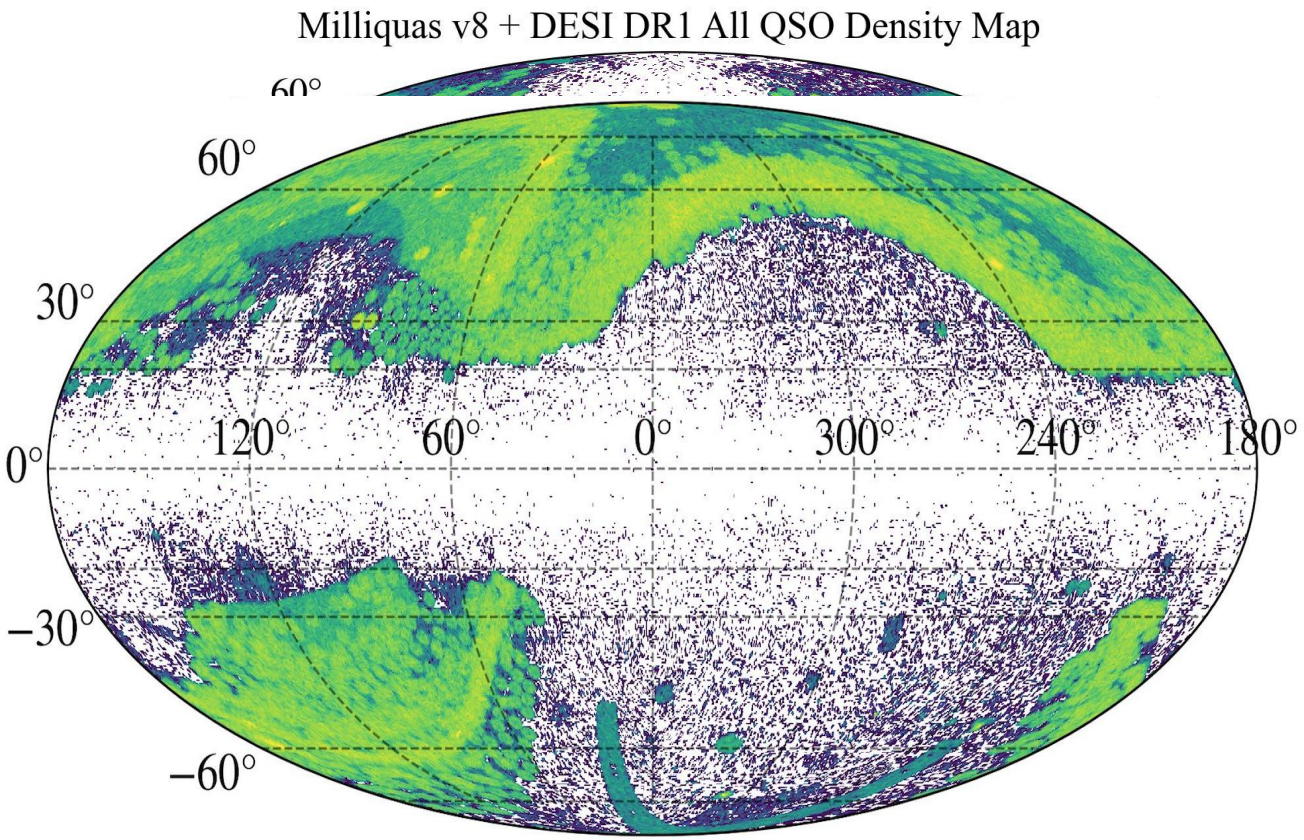
Representative surveys	Numbers
The Bright Quasar Survey (BQS; Schmidt & Green, 1983)	114
The Large Bright Quasar Survey (LBQS; Hewett, Foltz, & Chaffee, 1995)	1,000
The 2dF Quasar Redshift Survey (2QZ; Croom et al. 2004)	23,338
The 7th data release of the SDSS Quasar Catalog (DR7Q; Schneider et al. 2010)	105,783
SDSS DR12Q (Pâris et al. 2017)	297,301
SDSS DR14Q (Pâris et al. 2018)	526,356
SDSS DR16Q (Lyke et al. 2020) (contains candidates that are not visually inspected)	750,414
LAMOST Quasar Survey (Jin, Wu et al. 2023)	31,866





DESI 巡天后光谱证认的类星体数量新纪录

Milliquas v8 + DESI DR1 已有约200万个经光谱分类的类星体



光谱证认类星体的空间分布仍非常不均匀

- DESI, SDSS主要覆盖北天高银纬区域，**缺乏南天、低银纬覆盖**; DESI样本大且较深，但在亮端完备度偏低
- LAMOST 三期巡天新增了北天的明亮类星体，及银道面背景类星体候选体
- 南天以及全天范围仍缺少均匀性、一致性较好的类星体红移巡天样本

Gaia DR3 类星体候选体星表



北京大学 PEKING UNIVERSITY

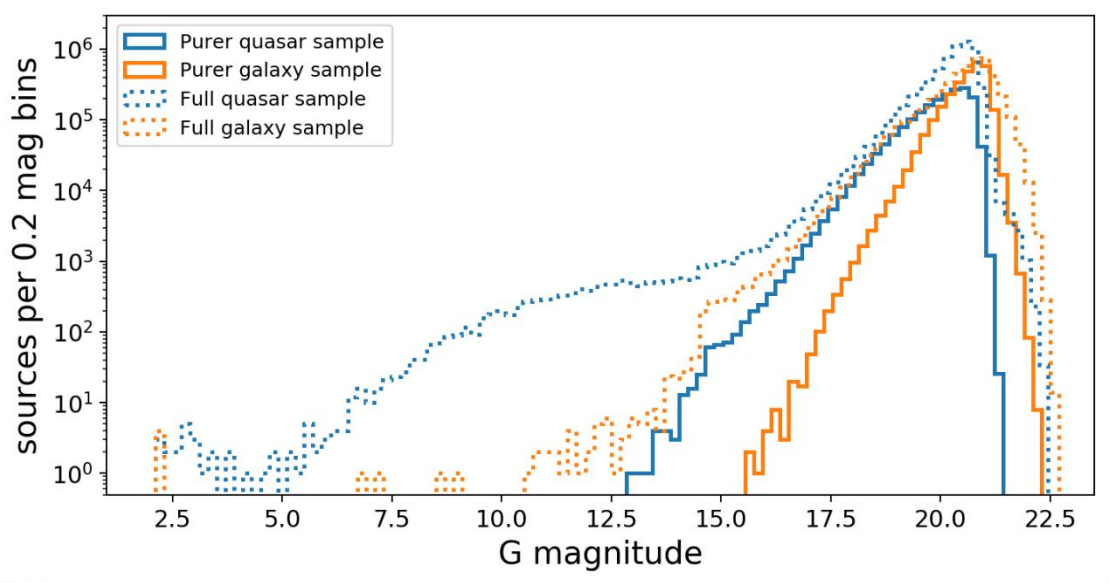
样本优势：

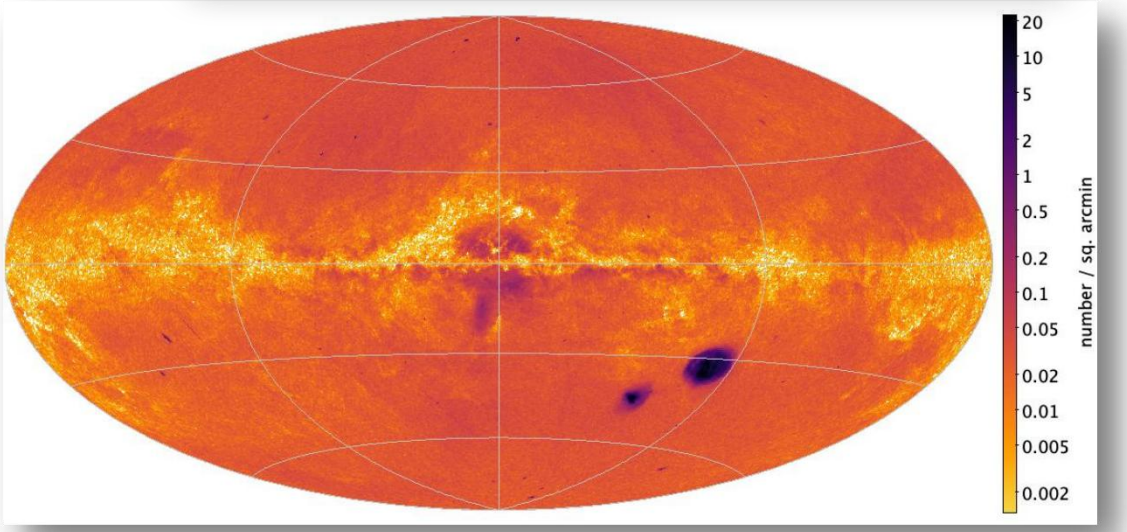
- **完备度高，样本量大，共计 660 万个源**
- 其中 162686 个源有Gaia 低分辨率光谱数据

样本不足：

- 纯度(真实类别为类星体的比例)低，据Gaia估计为约50%
- 红移测量准确度低（离群比例高）

Module	Selected sources	Featuring parameters
qso_candidates	6 649 162	
DSC	5 543 896	6 647 511
QSOC	1 834 118	6 375 063
Vari-Classification	1 035 207	1 122 361
Vari-AGN	872 228	872 228
Surface brightness	925 939	1 084 248
Gaia-CRF3	1 614 173	1 614 173
OA	N/A	2 803 225



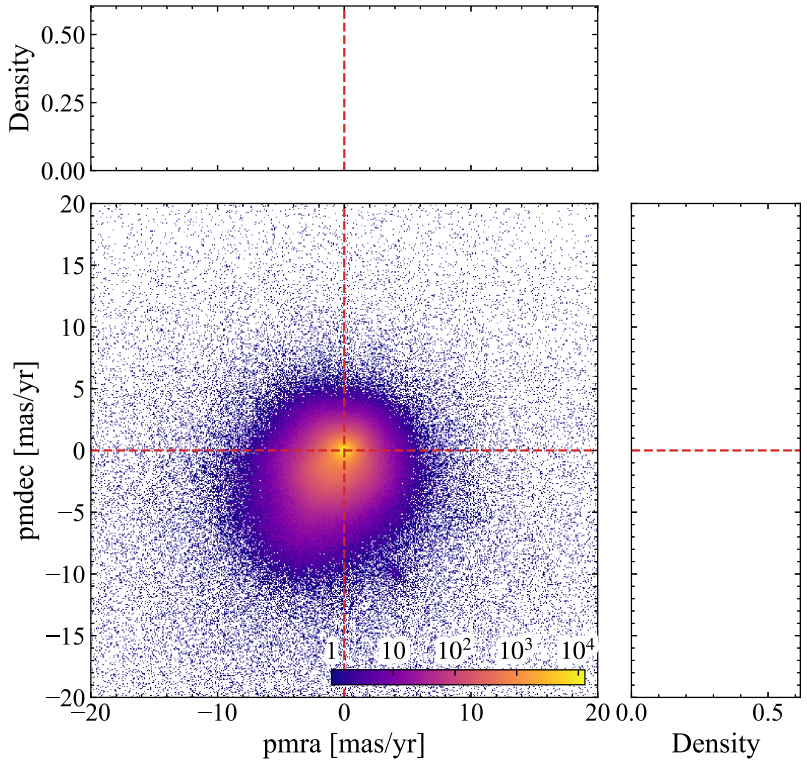


Gaia Collaboration, Bailer-Jones et al. 2022

Gaia DR3 类星体候选体自行分布

类星体距离遥远，在天空中的位置几乎不变，自行应接近于0；
Gaia类星体候选体的高自行也显示出**较高比例恒星污染**

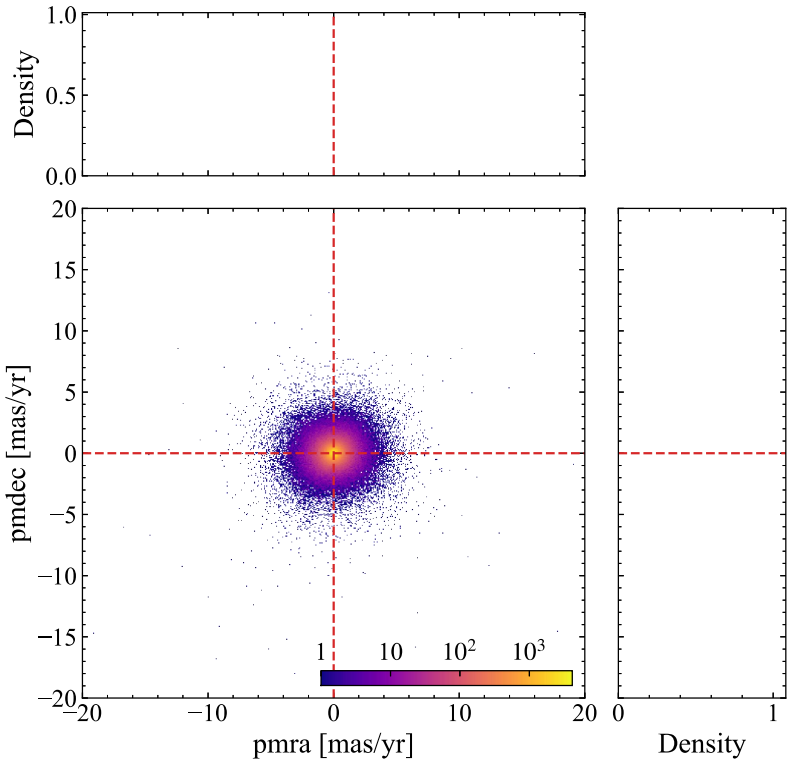
Gaia DR3
类星体候选体



赤纬方向自行

赤经方向自行

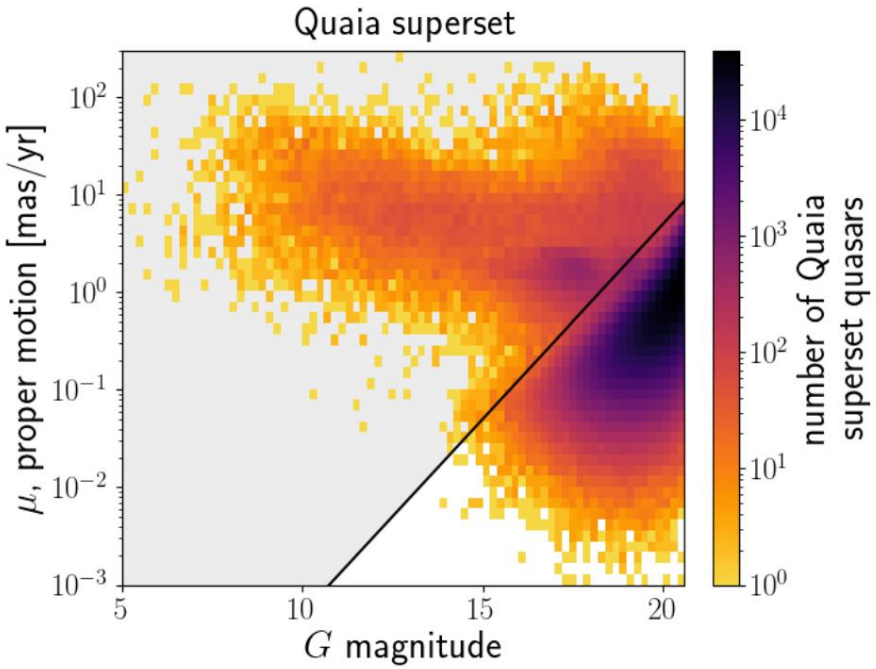
SDSS DR16Q
光谱证认的类星体



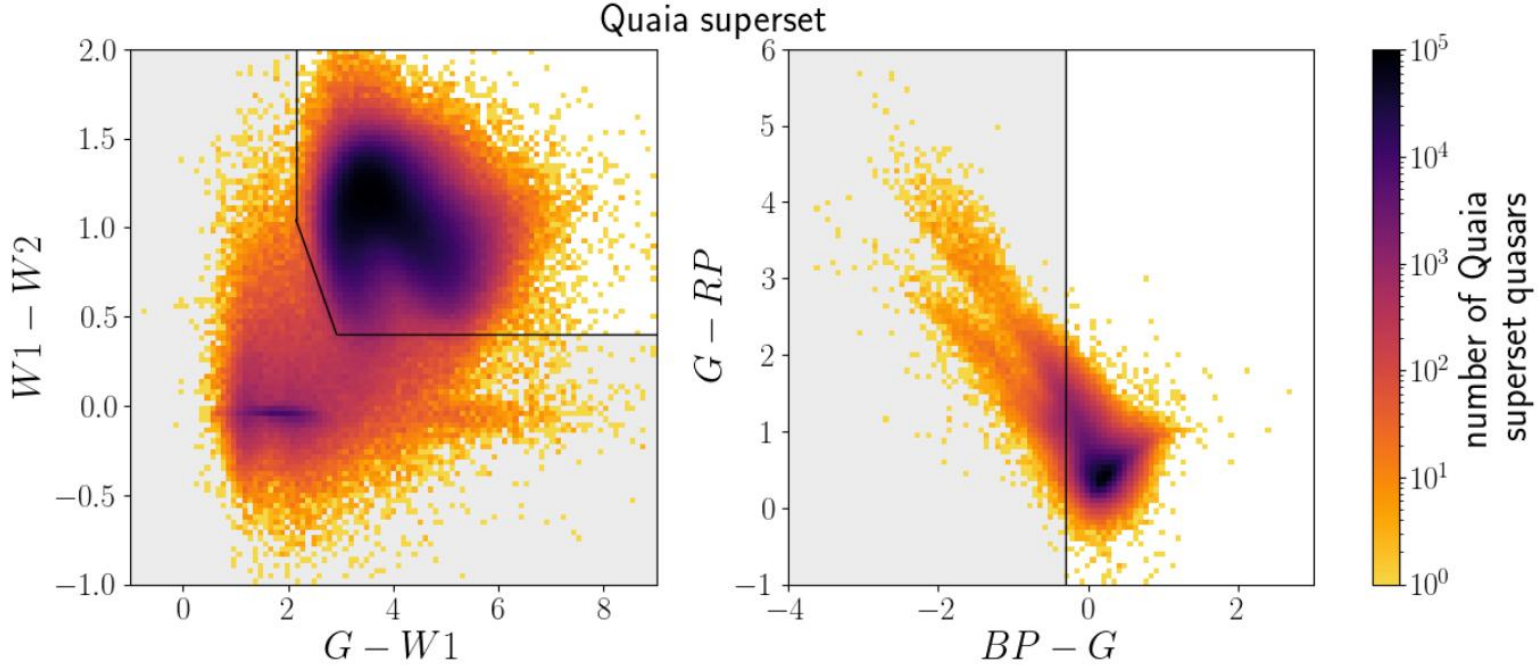
改进的使用简单特征截断的 Quaia 星表

Quaia比Gaia DR3 类星体候选体星表纯度更高，但降低完备度

Mag-proper motion cut



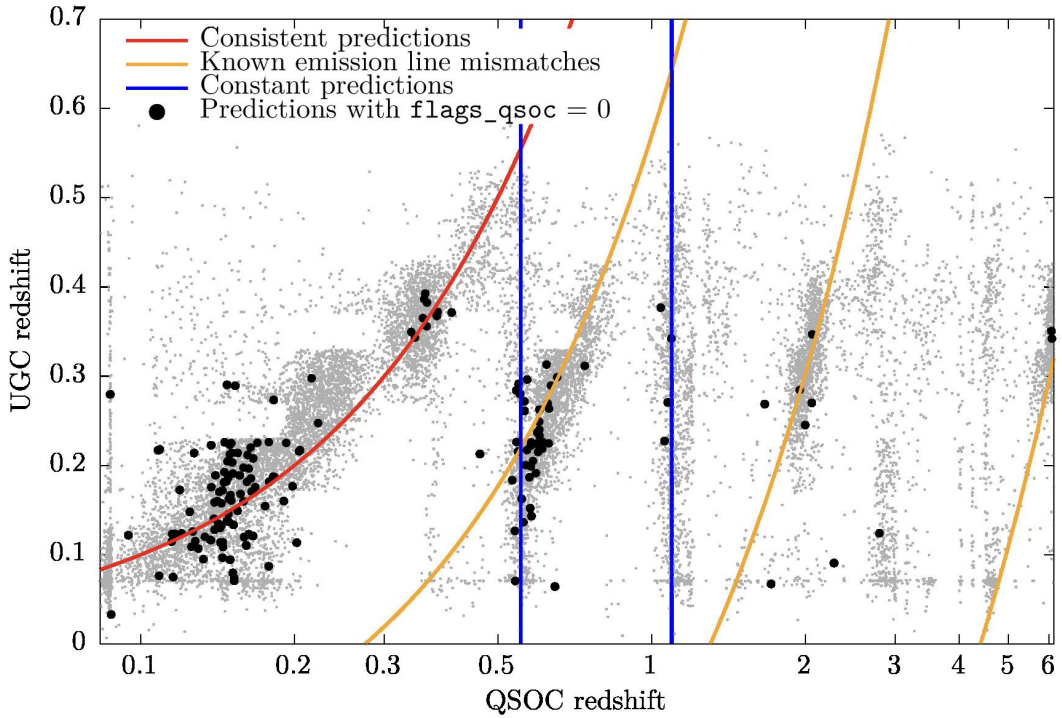
Color cuts



Storey-Fisher et al. 2023

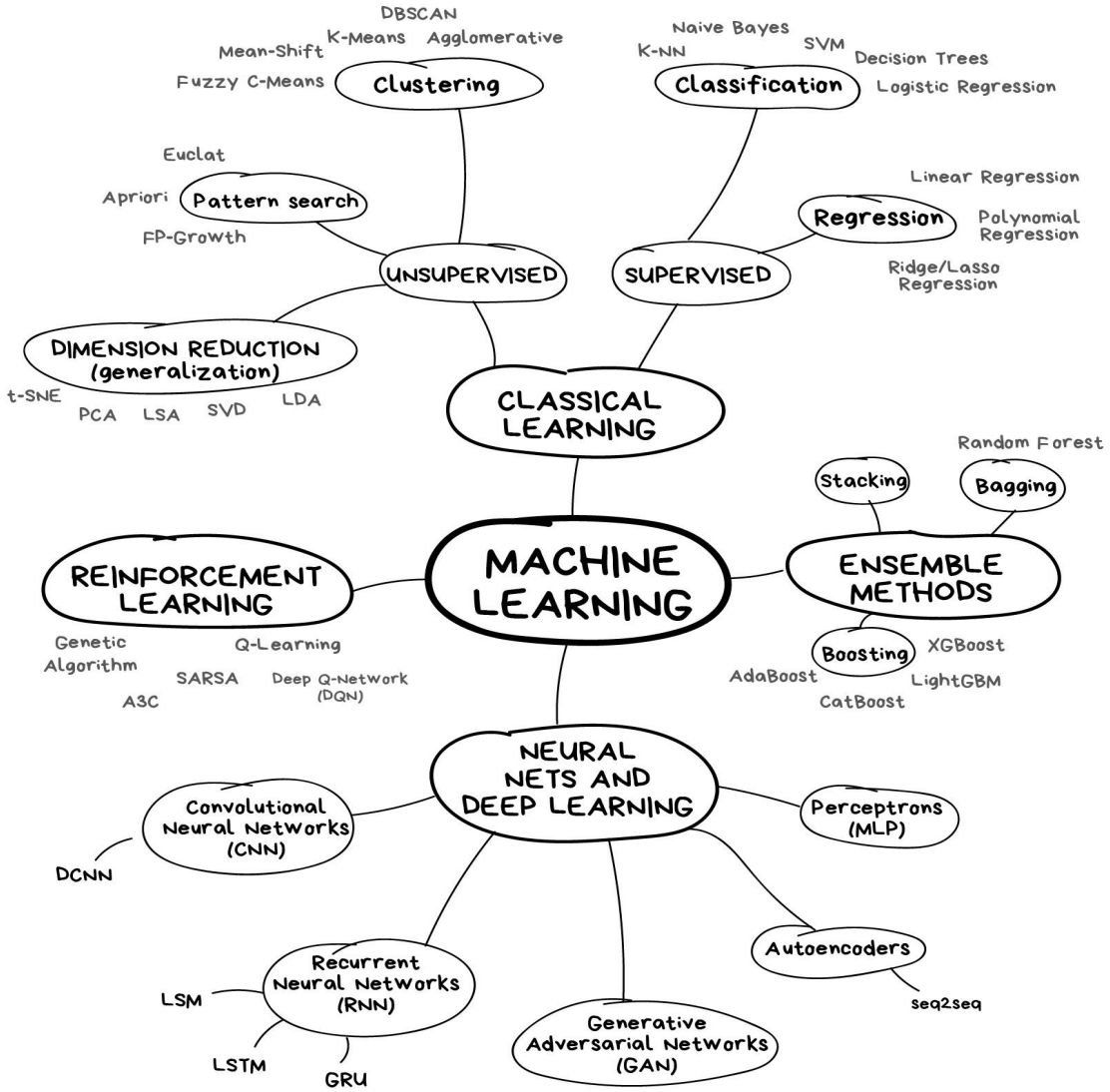
我们研究目标

- **基于机器学习利用测光和光谱巡天构建高纯度和高完备度的类星体样本**
  - 使用多波段数据进行机器学习选源
  - 应用Gaia自行数据进一步提升样本准确性 (Fu, Wu et al. 2021, ApJS)
- **更准确的测光 / 光谱红移估算**
  - 多波段数据+集成学习估算测光红移
  - 卷积神经网络(CNN)测量Gaia无缝光谱红移
- **为未来利用CSST测光和无缝光谱巡天构建类星体星表打下基础**



GDR3 QSOC redshift:
inconsistency between different models;
large scatters due to data from few passbands

机器学习



https://vas3k.com/blog/machine_learning/

监督学习（训练数据有标记）的经典算法：
k-NN、支持向量机、决策树等

无监督学习（训练数据未标记）的经典算法：
k-means聚类、主成分分析等

集成学习：通过使用bagging, boosting等集成方法，组合一组预测器的预测，比最好的单个预测器更好（“群体智慧”），如Random Forest、XGBoost、LightGBM等

决策树的集成学习方法

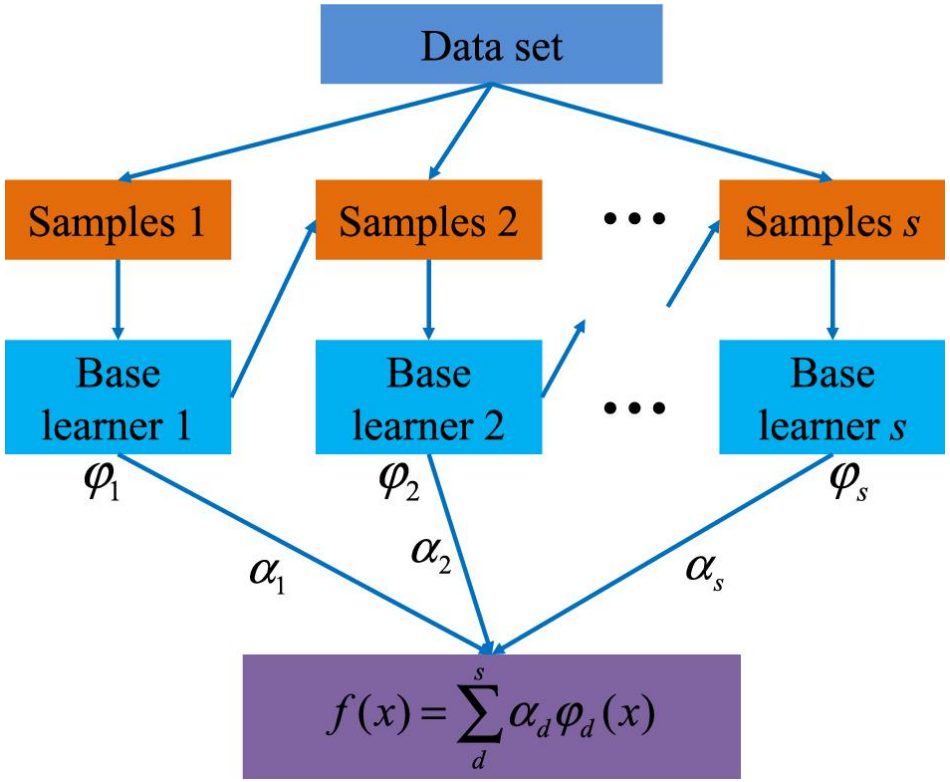
集成学习：用多个“弱”学习器相加成为强学习器

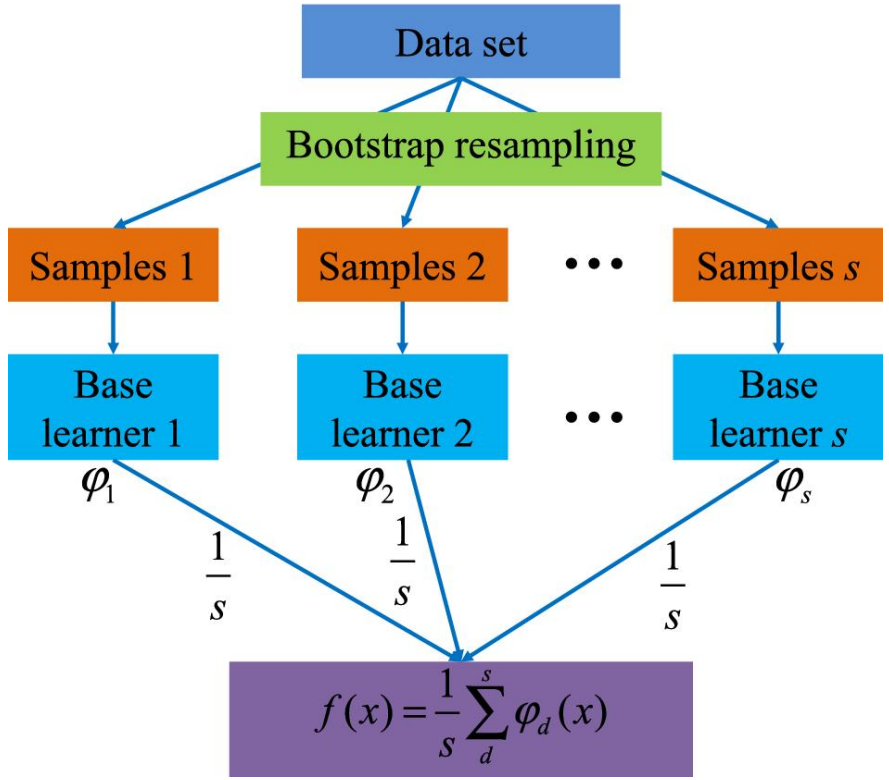
提升法（boosting）：学习器在前一轮训练基础上强化 (不断降低残差)，最后加权相加

例如：**XGBoost**, LightGBM, CatBoost等

装袋法（bagging）：弱学习器由有放回抽样训练得到

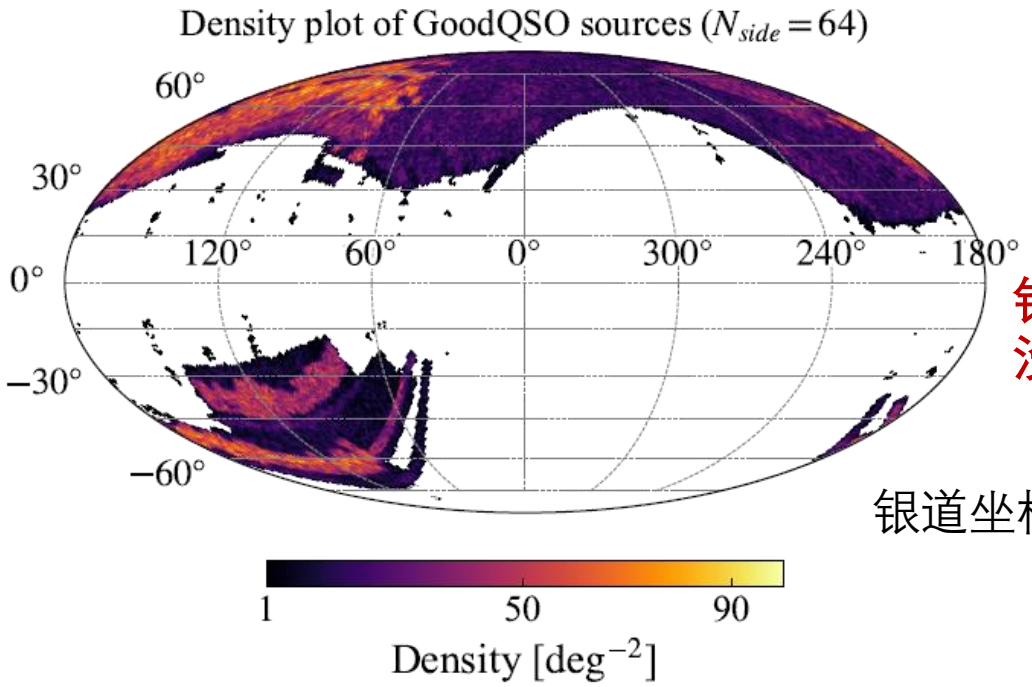
例如：Random Forest





示意图：Liang et al. 2020, Nat Hazards 104, 1923–1946.

2、迁移学习与银道面背景类星体巡天



银道面缺少类星体！
没有训练样本怎么办？

银道坐标系

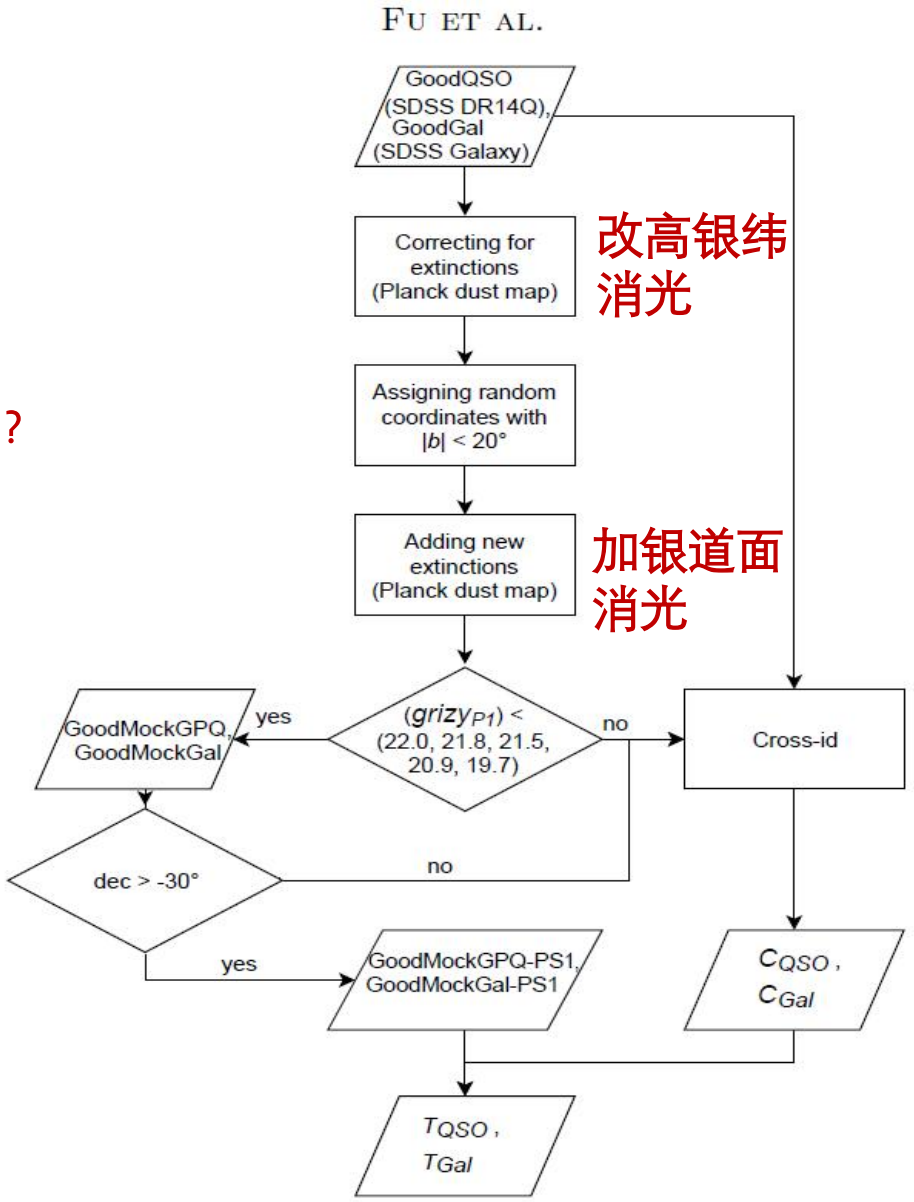
Input data:

高银纬：SDSS DR14Q (GoodQSO) and DR15 galaxies (GoodGal)

Output:

银道面模拟星表：GoodMockGPQ-PS1 and GoodMockGal-PS1

应用迁移学习方法　Fu, Wu, et al. 2021



类星体、星系、恒星两步分类法

Two-step classification using PS1 & ALLWISE data

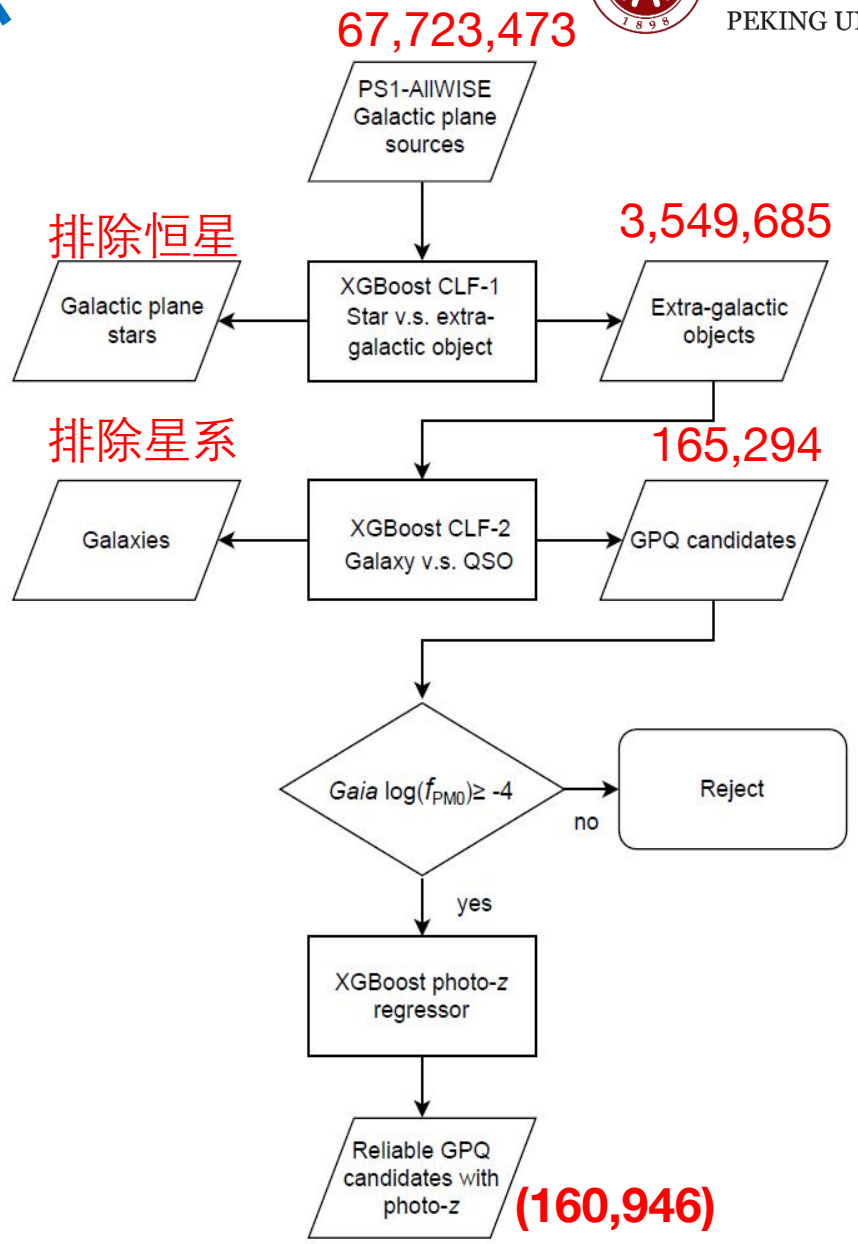
Feature selection:

g-r,r-i,i-z,z-y,g-W1,r-W1,i-W1,z-W1,y-W1,W1-W2,W2-W3,i-i_kron,z-z_kron.

Classification 1: Star vs. extragalactic (quasar + galaxy).

Classification 2: Galaxy vs. quasar.

Algorithm: XGBoost (Chen & Guestrin 2016). Hyper-parameter tuning tool: optuna (Akiba et al. 2019).



Fu, Wu, et al. 2021

测光红移
Estimation with XGBoost regression algorithm

Feature selection:

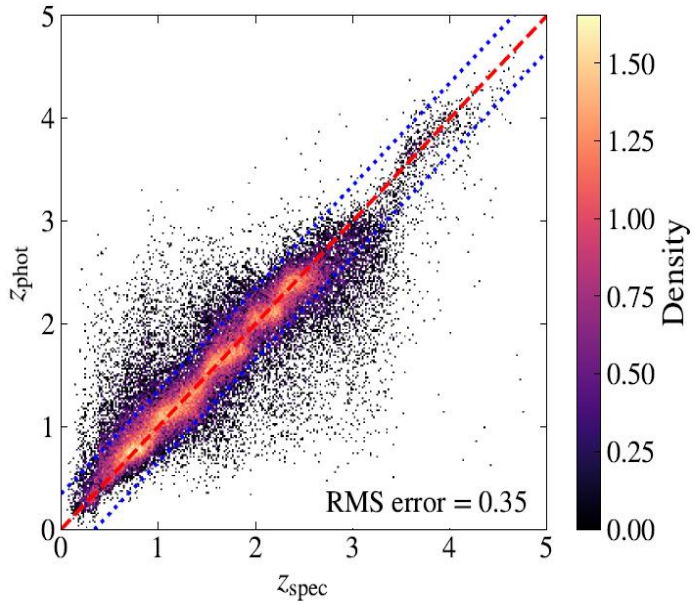
g-r, r-i, i-z, z-y, g-W1, r-W1, i-W1, z-W1, y-W1, W1-W2, W2-W3, i-i_kron, z-z_kron.

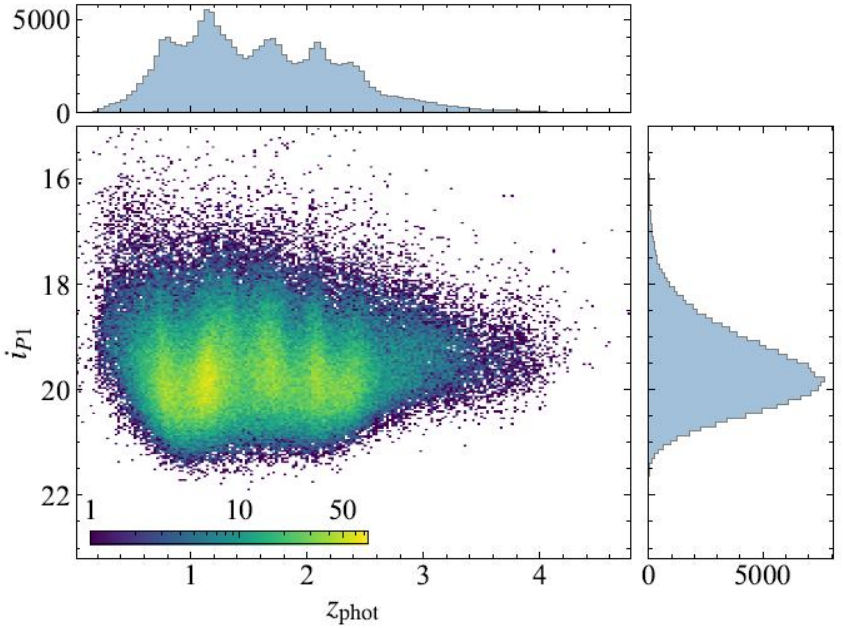
Root-mean-square error on the validation set: 0.35;

Photo-z accuracy R_0.1:

the fraction of quasars with $|\Delta z| \leq 0.1$,

where $|\Delta z| = |z_{\rm spec} - z_{\rm phot}|/(1 + z_{\rm spec})$

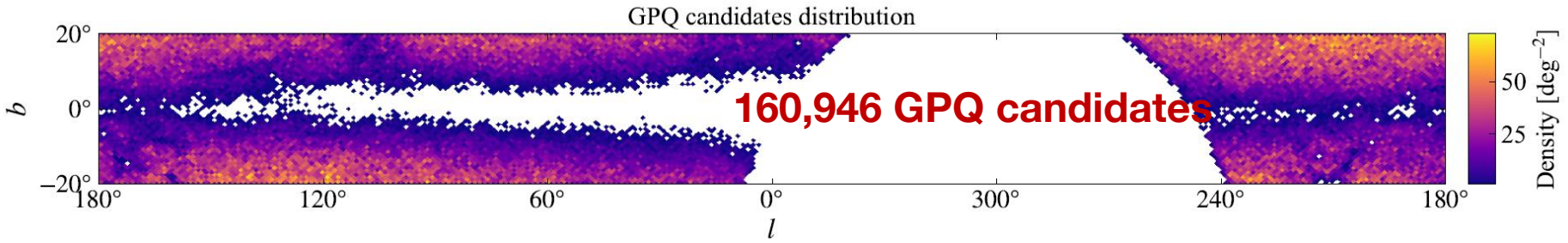


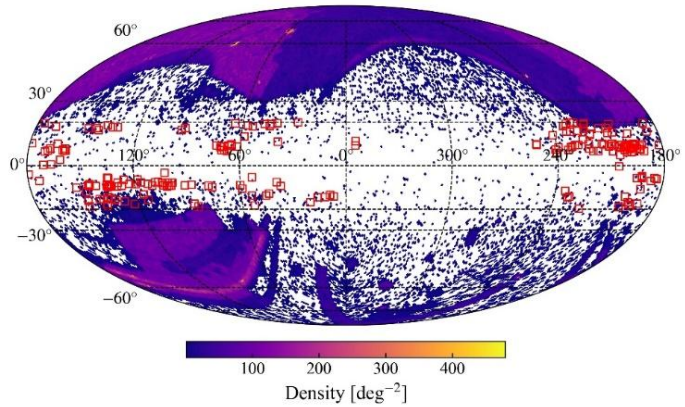


Source number	**160,946**
Photo-z range	[0.016, 4.777]
Expected redshift range	(0, 5)
Sky coverage	|b|<20° , δ>-30°

Fu, Wu, et al. 2021

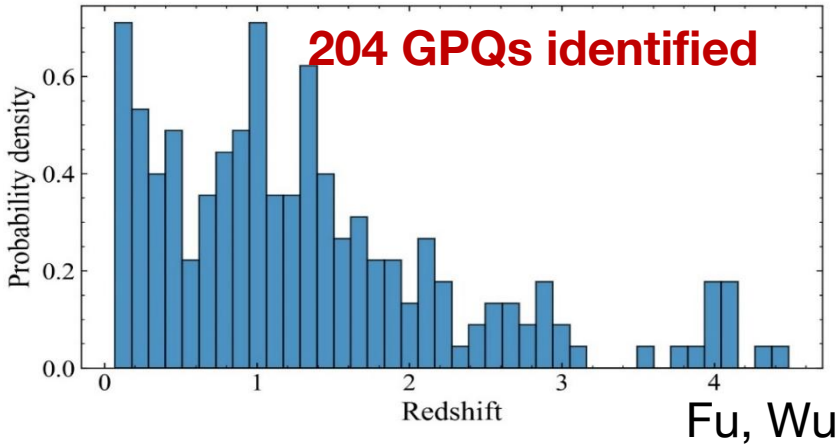
银道面背景类星体巡天(2018–)





Fu, Wu et al. 2021, ApJS

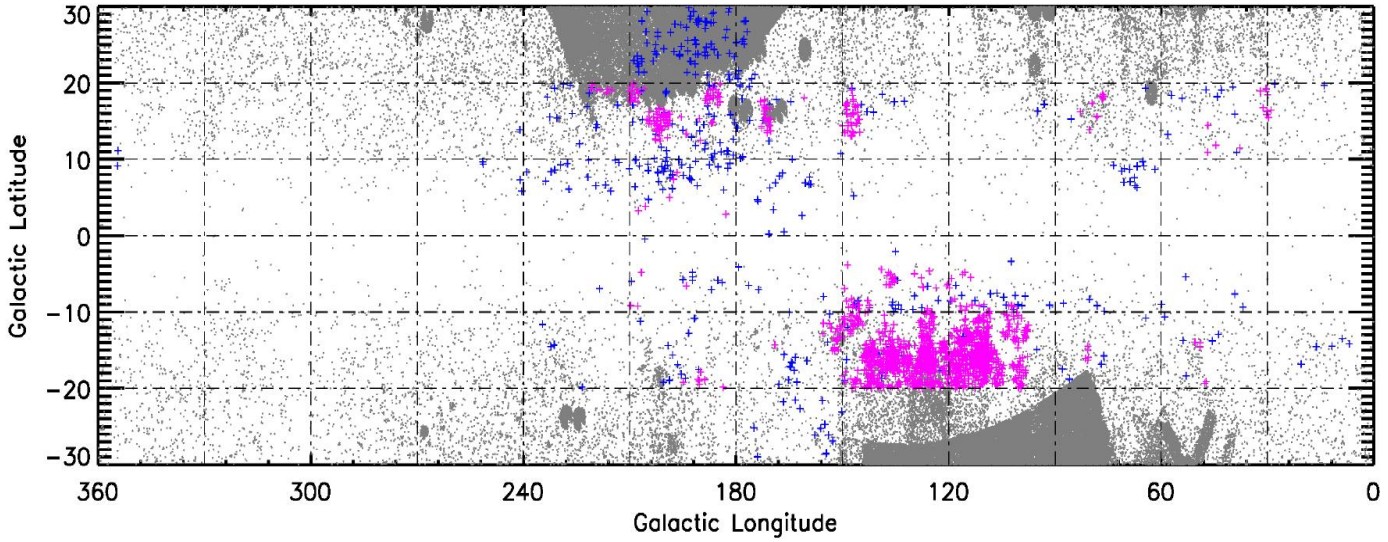
Telescope	$N_{candidate}$	N_{QSO}	N_{star}	N_{galaxy}	$N_{unknown}$	Min(z)	Max(z)	Min(i_{P1})③	Max(i_{P1})
LJT	83	76	2	1	4	0.099	3.744	15.69	21.00
XLT	76	61	4	1	10	0.071	3.541	15.51	18.68
MDM13	31	28	2	0	1	0.069	2.337	15.80	18.15
P200	40	26	10	0	4	0.135	4.487	17.59	19.9
ANU23	13	13	0	0	0	0.794	2.112	16.74	18.43
Total	243	204	18	2	19	0.069	4.487	15.51	21.00



Fu, Wu et al. 2022, ApJS

QUASARS BEHIND THE GALACTIC PLANE FROM LAMOST DR10

~ 1300 GPQs identified from LAMOST DR10!

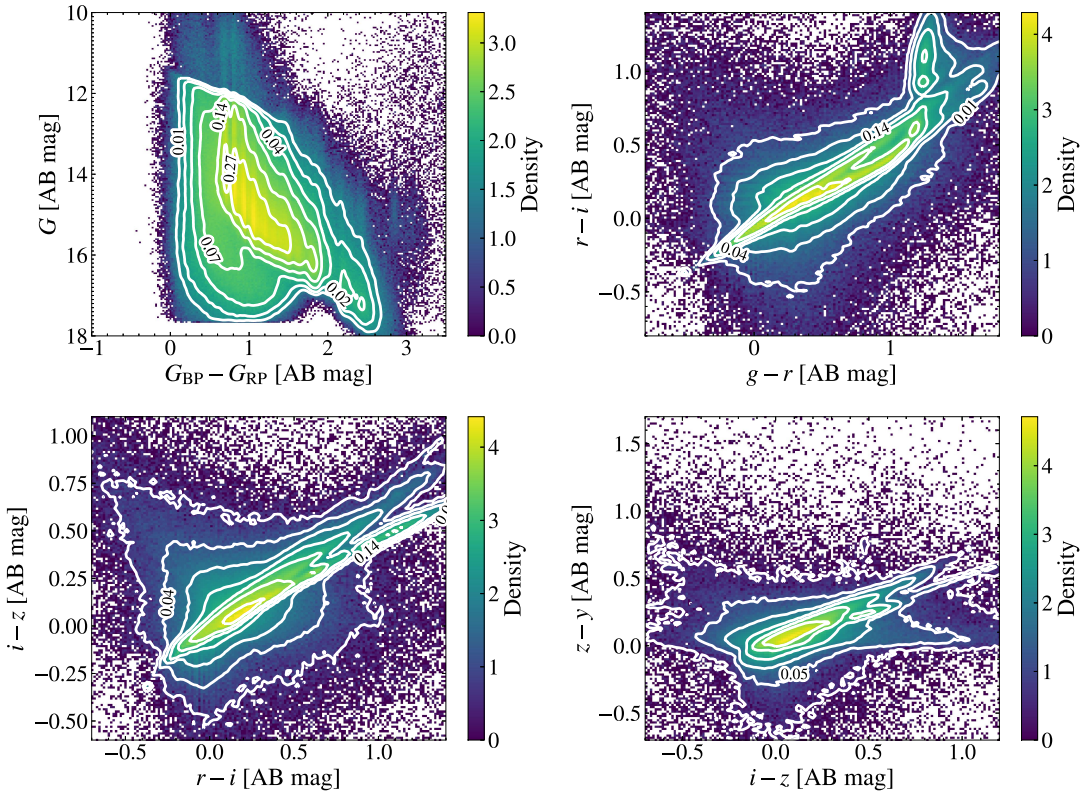


Huo, Fu, et al. 2025, ApJS

3. XGBoost与CatNorth/CatGlobe/QHSC类星体星表

类星体选源的机器学习分类

- 对原有低纯度 Gaia DR3 类星体候选体源表进行**三分类：恒星、星系、类星体**
- 算法: XGBoost (Chen & Guestrin 2016)，极限梯度提升决策树模型
- 从现有观测数据构建**类别标注准确、多样性高**的训练集/验证集:
  - LAMOST 增值星表的恒星样本
  - Gaia DR3 的优质恒星样本（优于“Golden sample”）
  - 具有足够代表性的污染源样本：低质量恒星（包括褐矮星）、白矮星、碳星
  - SDSS DR16Q 类星体样本、SDSS DR17 星系样本



Gaia DR3 “Golden” 恒星样本颜色空间分布存在难以区分的多个成分，不适于用作训练集的恒星样本

XGBoost 集成分类模型训练

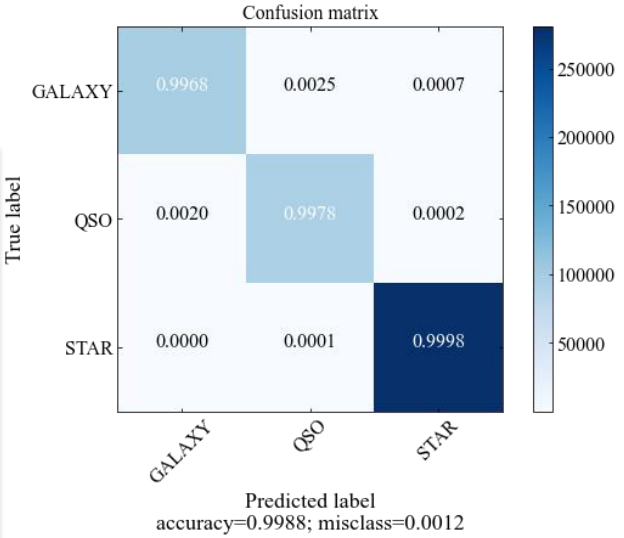
- **数据来源**（光学+红外）：Gaia DR3, Pan-STARRS DR1, & CatWISE2020
- **特征**: *g*-*r*, *r*-*i*, *i*-*z*, *z*-*y*, *g*-W1, *r*-W1, *i*-W1, *z*-W1, *y*-W1, W1-W2, G_{BP}-G_{RP}, G_{BP}-*G*, *G*-G_{RP}, corrected BP/RP excess factor (形态)
- **分类器 CLF_LVAC 使用样本**: LAMOST VAC+恒星样本、SDSS类星体和星系样本
- **分类器 CLF_GDR3 使用样本**: Gaia DR3+恒星样本、SDSS类星体和星系样本
- 在验证集上，两个分类器均得到0.9988的**准确度**
- **使用两分类器平均结果作为最终结果，pQSO_mean>0.95 的源判定为类星体候选体**

超参数调节（使用optuna）

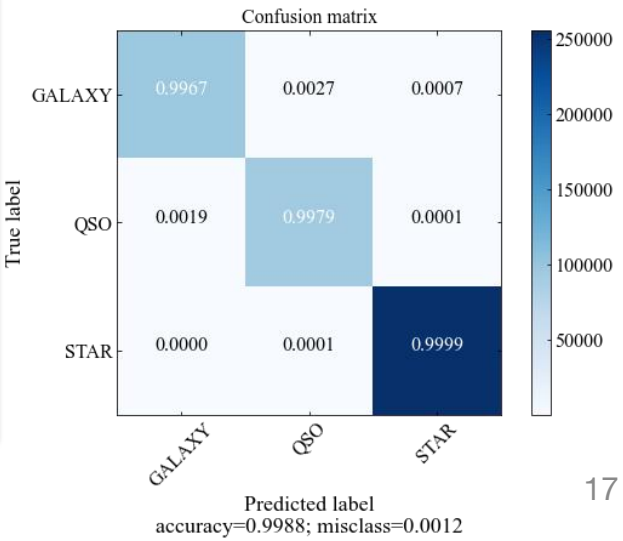
Table 2. Default and Optimal Hyperparameter Settings for CLF_LVAC and CLF_GDR3 `eta` = 0.1, `num_boost_round` = 100

Hyperparameter	CLF_LVAC		CLF_GDR3	
	Default	Optimal	Default	Optimal
`lambda`	1	1.18	1	1.32
`alpha`	0	1.61	0	0.33
`max_depth`	6	9	6	9
`gamma`	0	0.71	0	0.18
`grow_policy`	`depthwise`	`lossguide`	`depthwise`	`lossguide`
`min_child_weight`	1	3	1	4
`subsample`	1	0.87	1	0.70
`colsample_bytree`	1	0.61	1	0.74
`max_delta_step`	0	5	0	8
Balanced accuracy	0.9972	0.9977	0.9973	0.9979
Precision (weighted)	0.9981	0.9985	0.9982	0.9985
Recall (weighted)	0.9981	0.9985	0.9982	0.9985
F_1 (weighted)	0.9981	0.9985	0.9982	0.9985
MCC	0.9967	0.9973	0.9968	0.9975

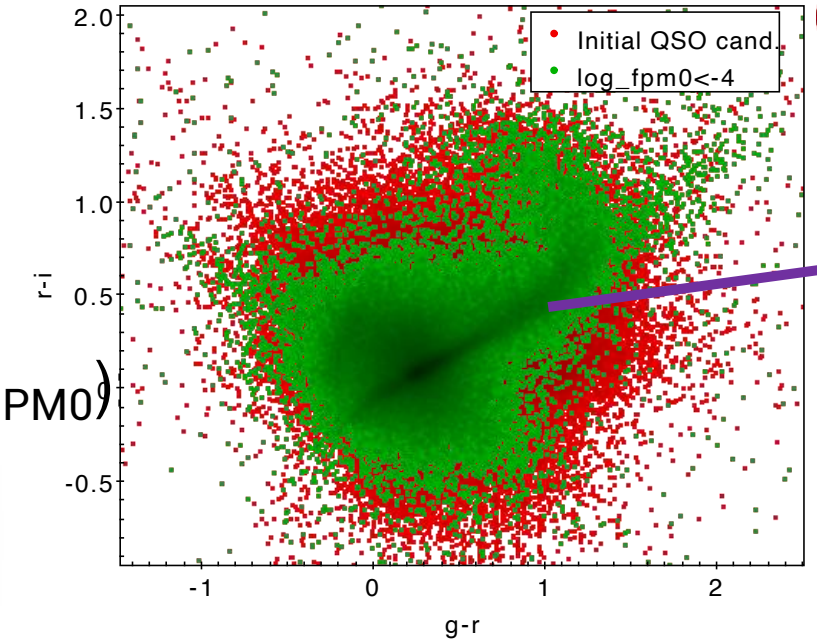
CLF_LVAC



CLF_GDR3



零自行概率密度辅助判据：去除恒星污染源



Stellar locus revealed
(被我们自行判据排除)

Probability density of zero proper motion (f_{PM0})

$$f_{\rm PM0} = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x}{\sigma_x}\right)^2 - \frac{2\rho xy}{\sigma_x\sigma_y} + \left(\frac{y}{\sigma_y}\right)^2\right]\right\}$$

x = pmra, y = pmdec

σ_x = pmra_error

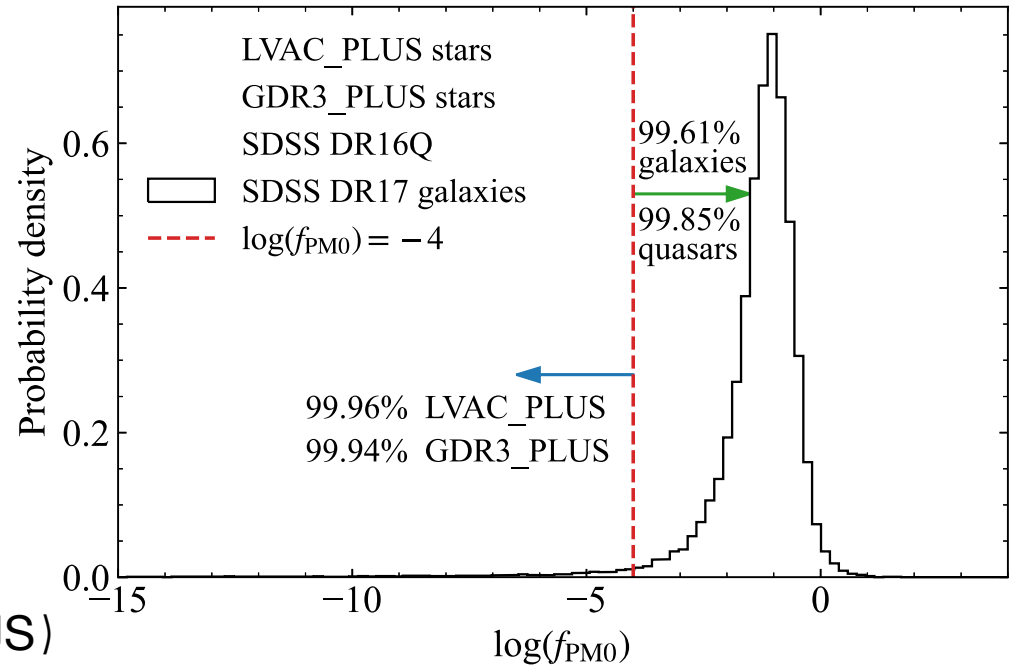
σ_y = pmdec_error

ρ = pmra_pmdec_corr

Apply cut **$\log(f_{PM0}) \geq -4$**

to select GPQ candidates which can

- exclude >99.9% of the stars
- retain >99.8% of the quasars

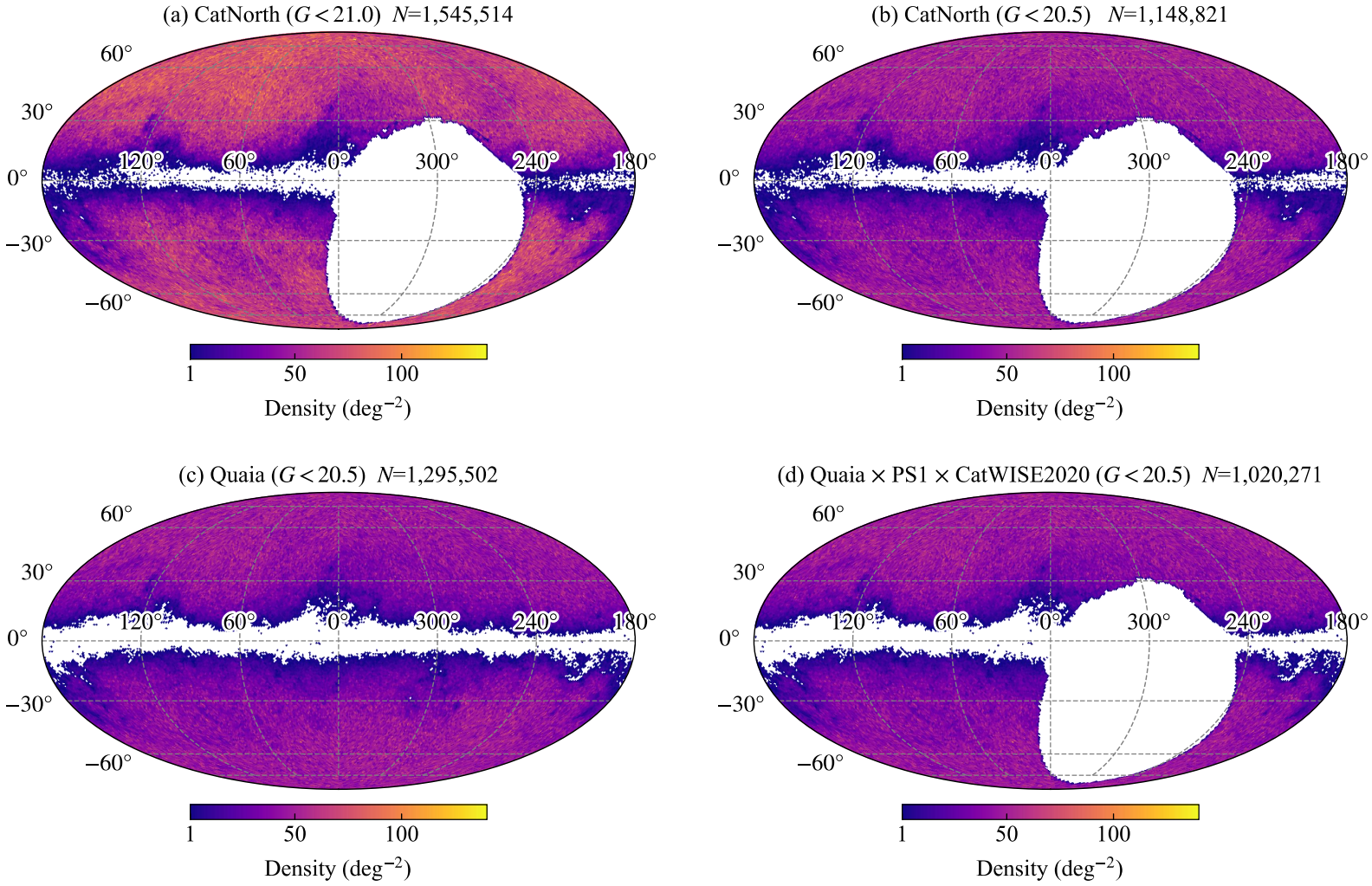


（Fu, Wu et al. 2021, ApJS）

选源结果

CatNorth 百万类星体候选体星表

- 3π天区（受Pan-STARRS覆盖天区限制）
- 共**154万**多个源
- 同天区和星等范围内比Quaia源数目更多



红移估算

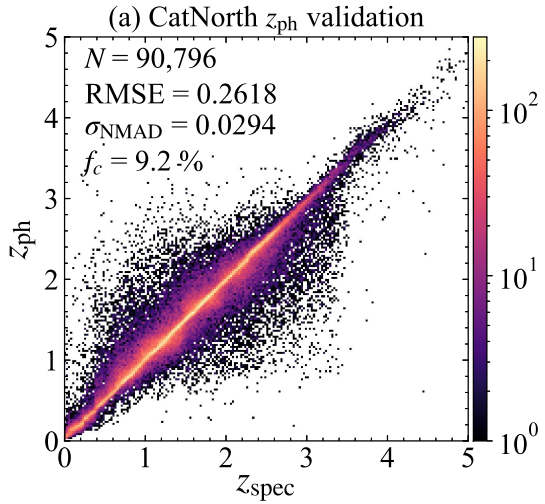
测光红移的集成学习模型 + 低分辨率无缝光谱红移的CNN模型

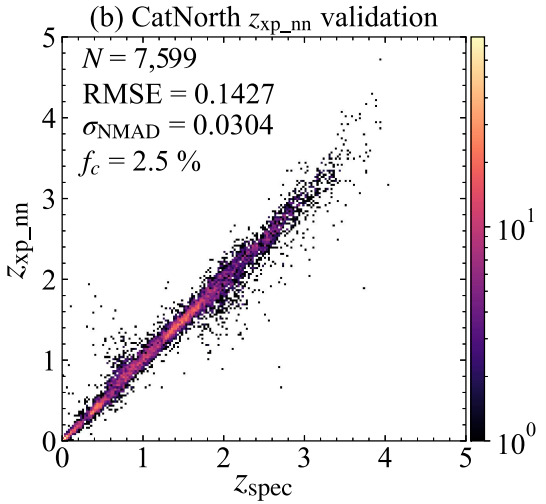
- **测光红移：**
  - CatNorth全部154万多个源
  - 基础模型包括XGBoost, TabNet (深度学习模型), FT-Transformer (深度学习模型)
  - 平均集成模型有最好的综合表现
- **CNN无缝光谱红移：**
  - 仅89100个具有公开Gaia BP/RP光谱的源
  - 波长范围4000-10000A，采样间隔20A，每条光谱3[illegible]
  - 自 $\mathrm{MNSE} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{z_i - \hat{z}_i}{1 + z_i}\right)^2.$

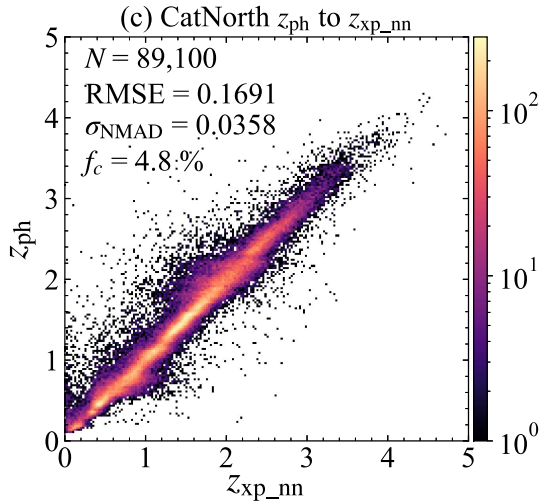
模型表现（验证集）

Table 3
Scores of All Photometric Redshift Regression Models (XGBoost, TabNet, FT-Transformer, and the Ensemble Model), and the Spectroscopic Redshift Regression Model (RegNet) on the Validation Sets

	Photo-z Models				Gaia BP/RP Spec-z Model
Model	XGBoost	TabNet	FT-Transformer	Ensemble	RegNet
Metric	RMSE	Smooth L1	Smooth L1		MNSE
Loss					
RMSE	0.2734	0.2685	0.2723	0.2618	0.1427
σ_{NMAD}	0.0351	0.0303	0.0307	0.0294	0.0304
f_c	10.65%	9.04%	9.21%	9.16%	2.46%

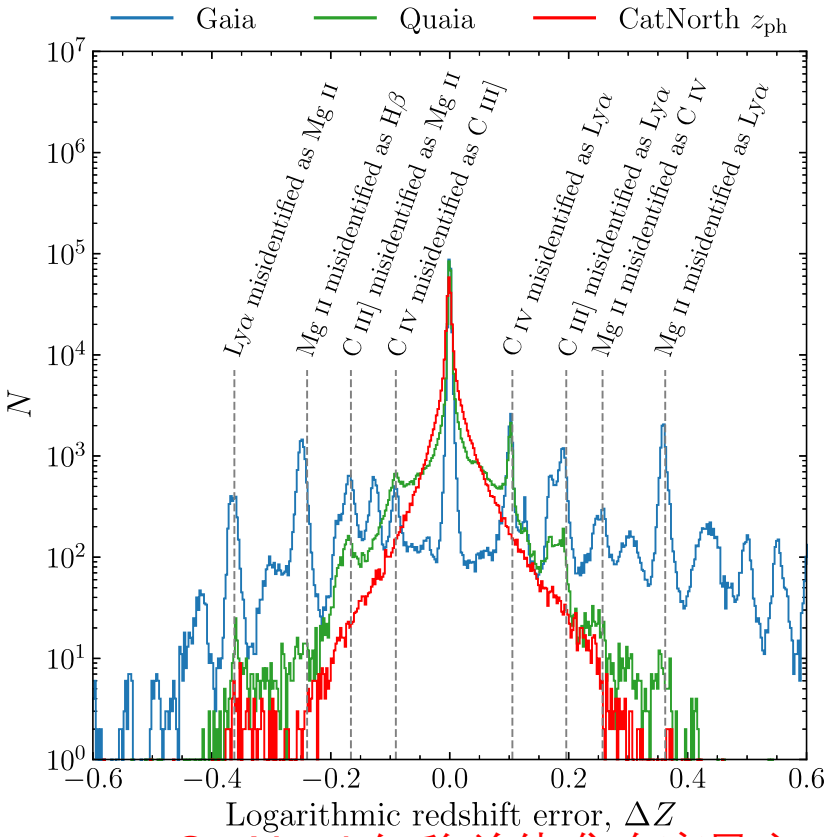




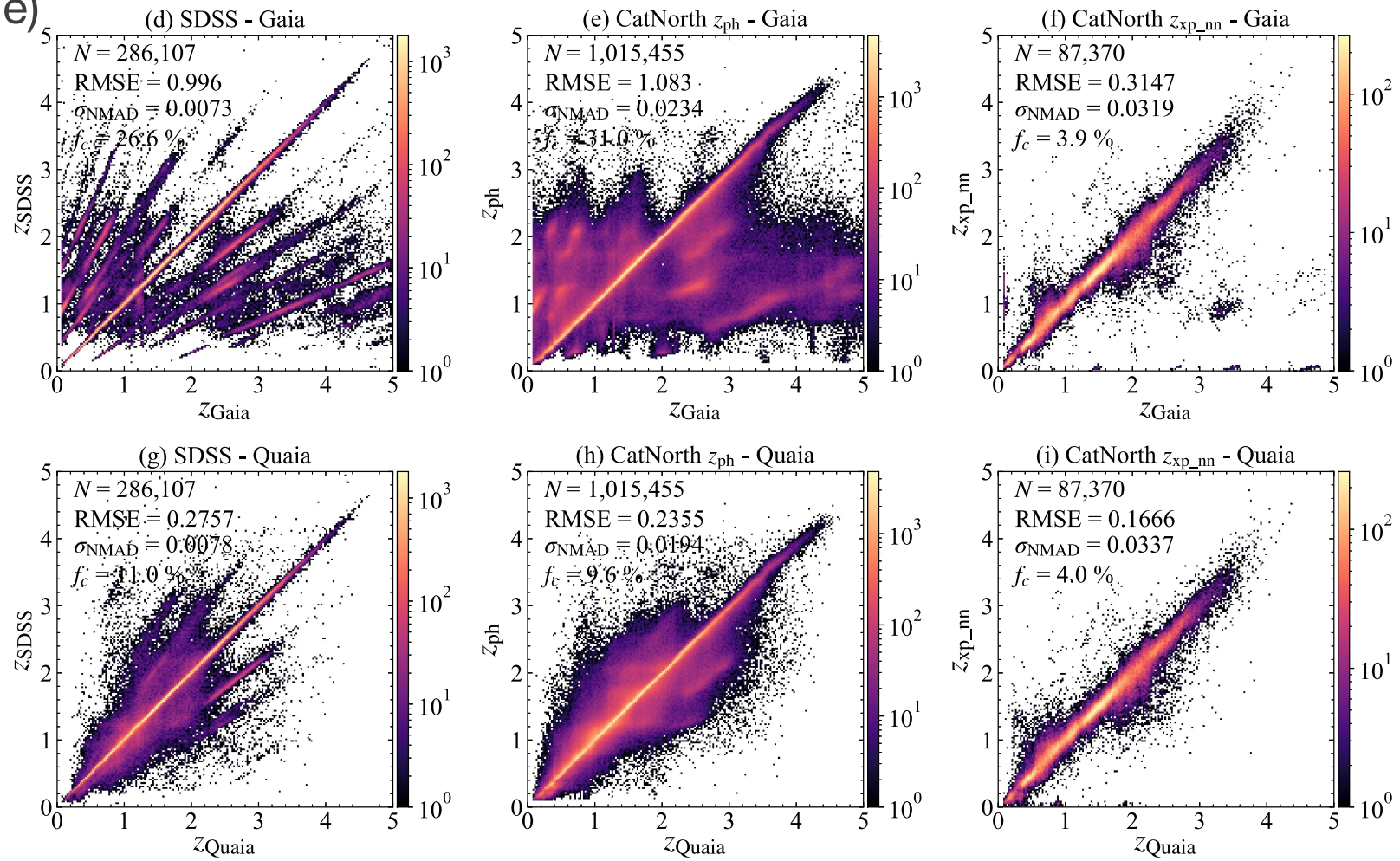


CatNorth 红移与 Gaia/Quaia 红移对比：CatNorth 显著降低谱线误认比例，总体准确度最高

- ΔZ = log(1 + zpred) – log(1 + ztrue)
- ΔZ = log λtrue – log λfalse



CatNorth红移总体准确度最高

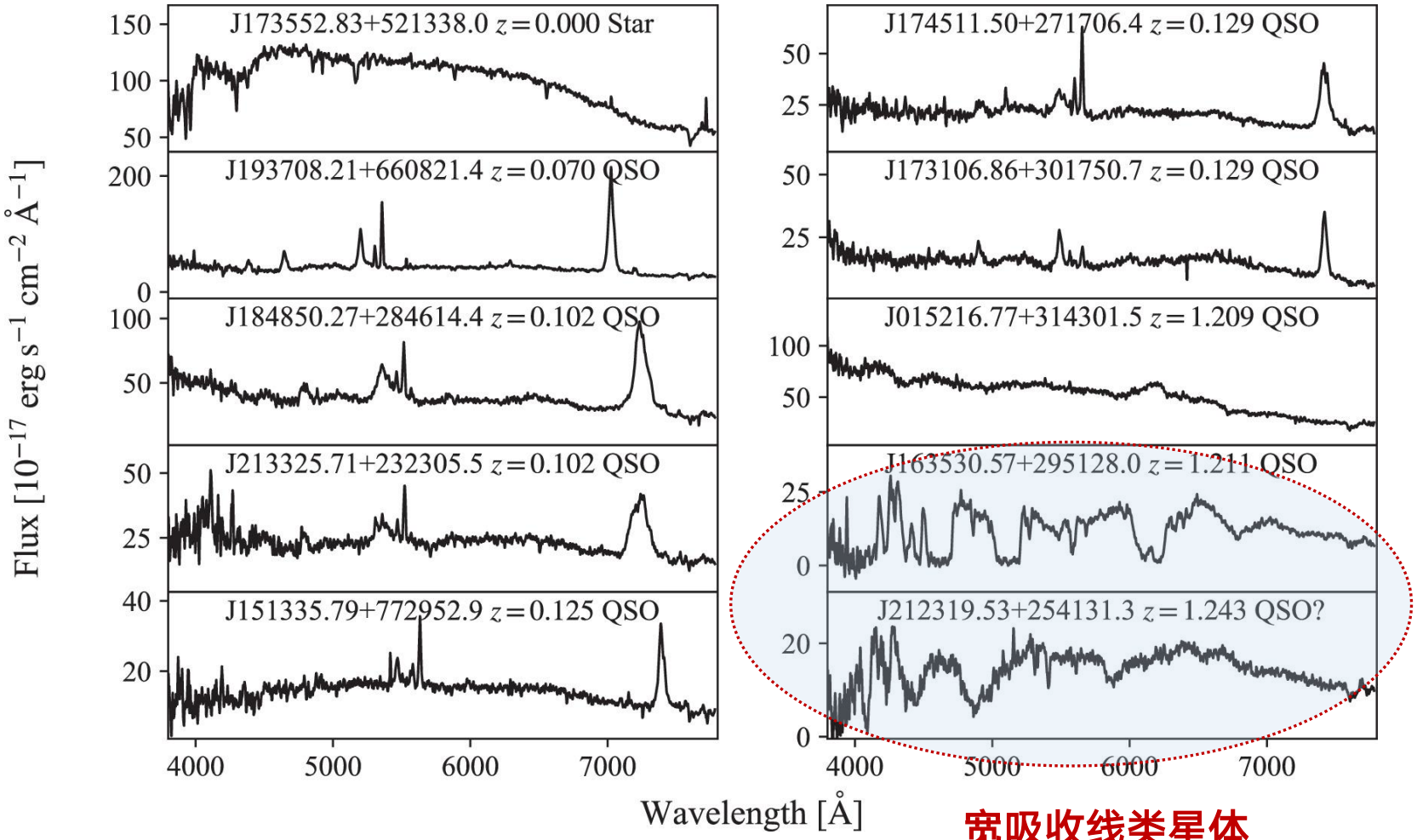


CatNorth观测证认：高纯度和完备性

印度HCT 2米望远镜**随机**光谱证认 (2023-08-16):

- **G<20.5**
- **在 CatNorth中 但不在Quaia 中**
- **新候选体**

Similar accuracy with Quaia，
Much higher completeness!

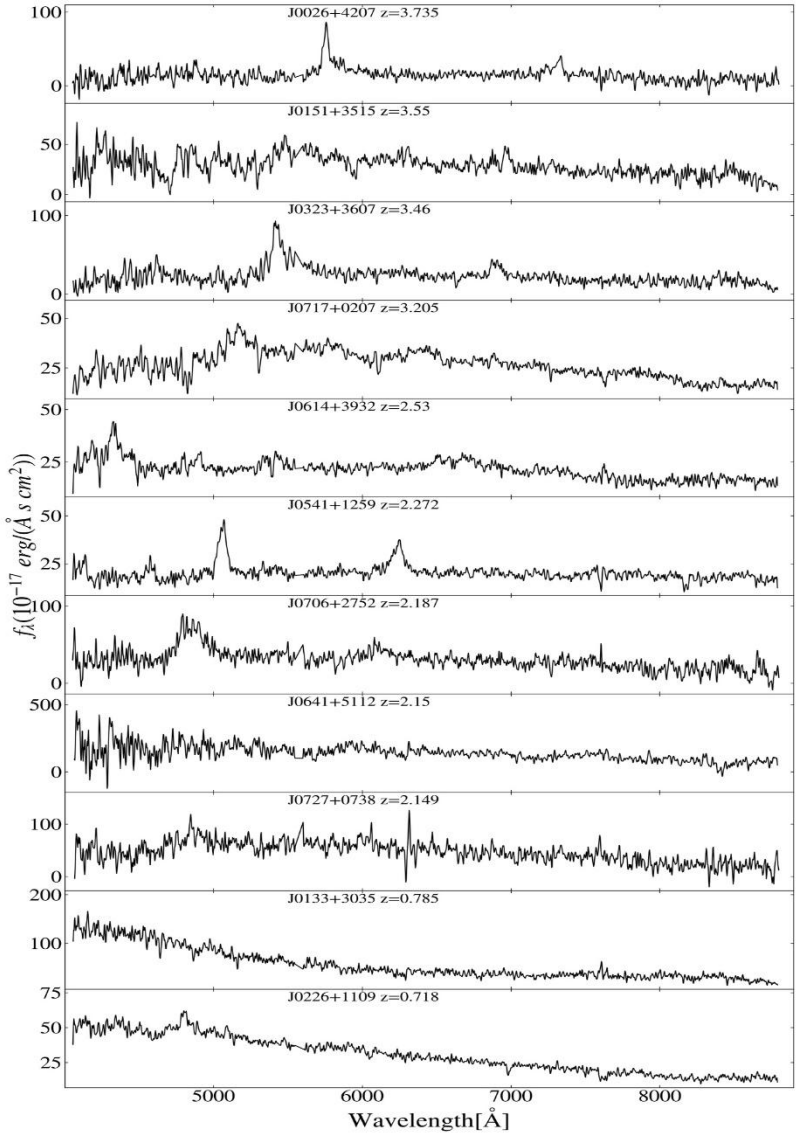


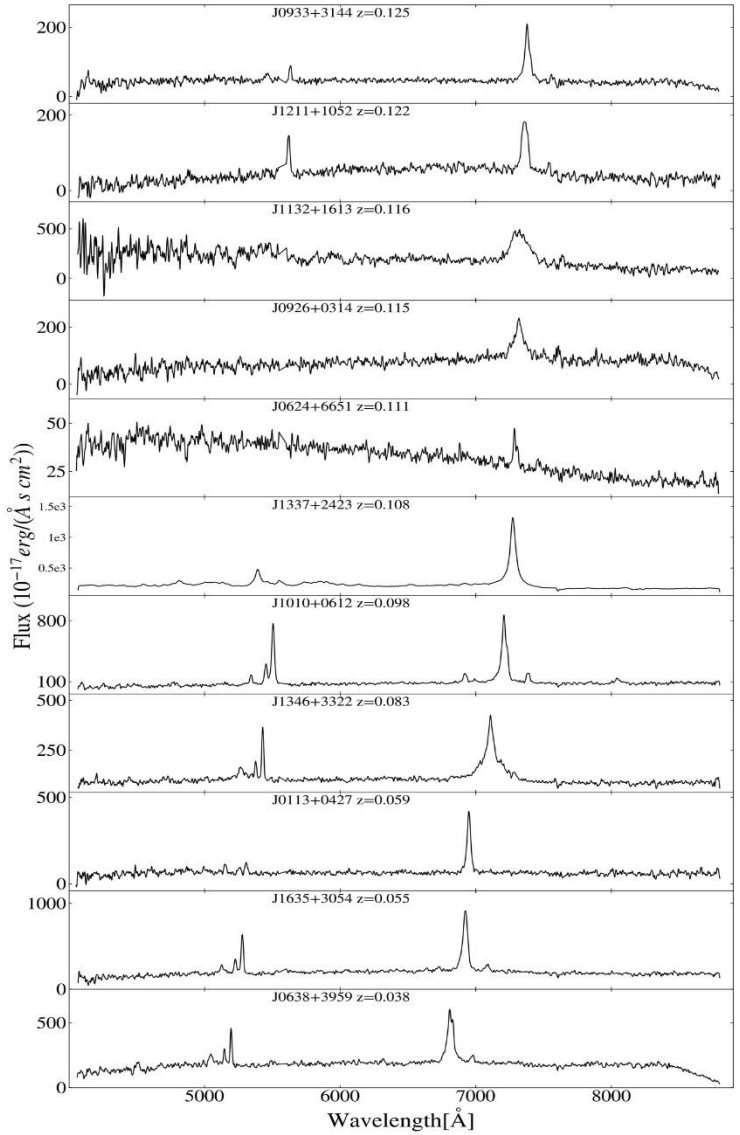
CatNorth观测证认：高纯度和完备性

兴隆 2.16米和丽江2.4米望远镜光谱证认结果 (2023-2024, ongoing):

27个源：22个类星体，5个非类星体

类星体证认成功率是81.5%

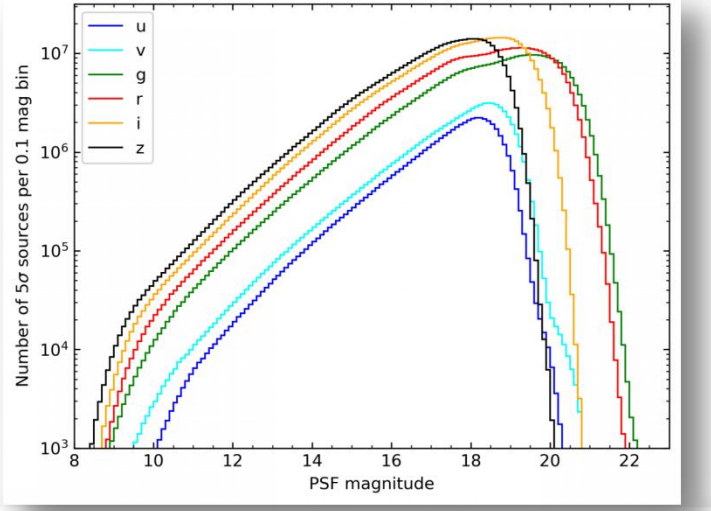




南天测光星表

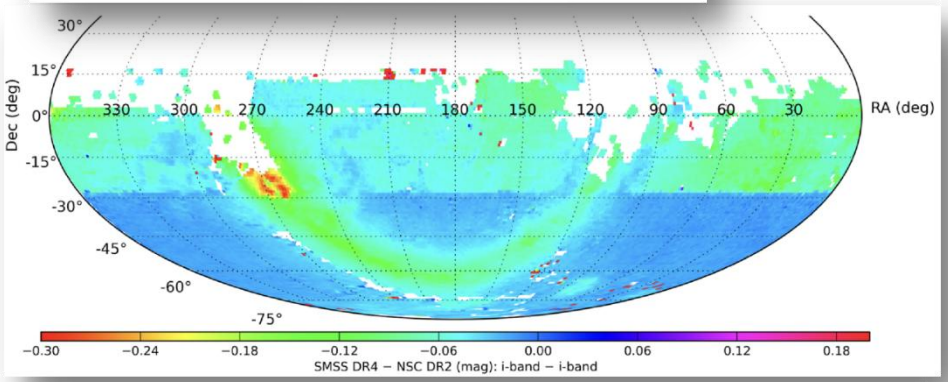
整合多个巡天，实现最佳天区覆盖：

- Gaia DR3, CatWISE2020 + AllWISE
- 可见光：
  - **SkyMapper Southern Sky Survey (SMSS) DR4**；主要覆盖南天（δ<0°），部分覆盖至δ<16°；选取*griz*波段
  - 使用 NSC (NOIRLab Source Catalog) DR2替换缺失的*griz*数据（无转换）
- 近红外：
  - **VISTA** 巡天(VHS, VIKING, VMC, VVV)；选取J和Ks波段
  - 使用2MASS替换缺失数据（经过2MASS到VISTA的星等转换）

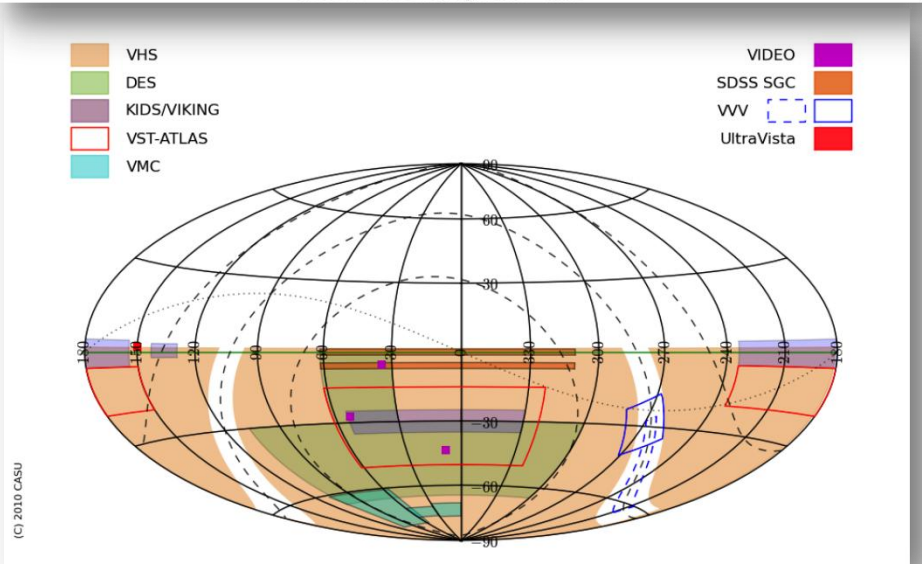


SMSS DR4星等分布

Onken et al. 2024, PASA, 41, e061



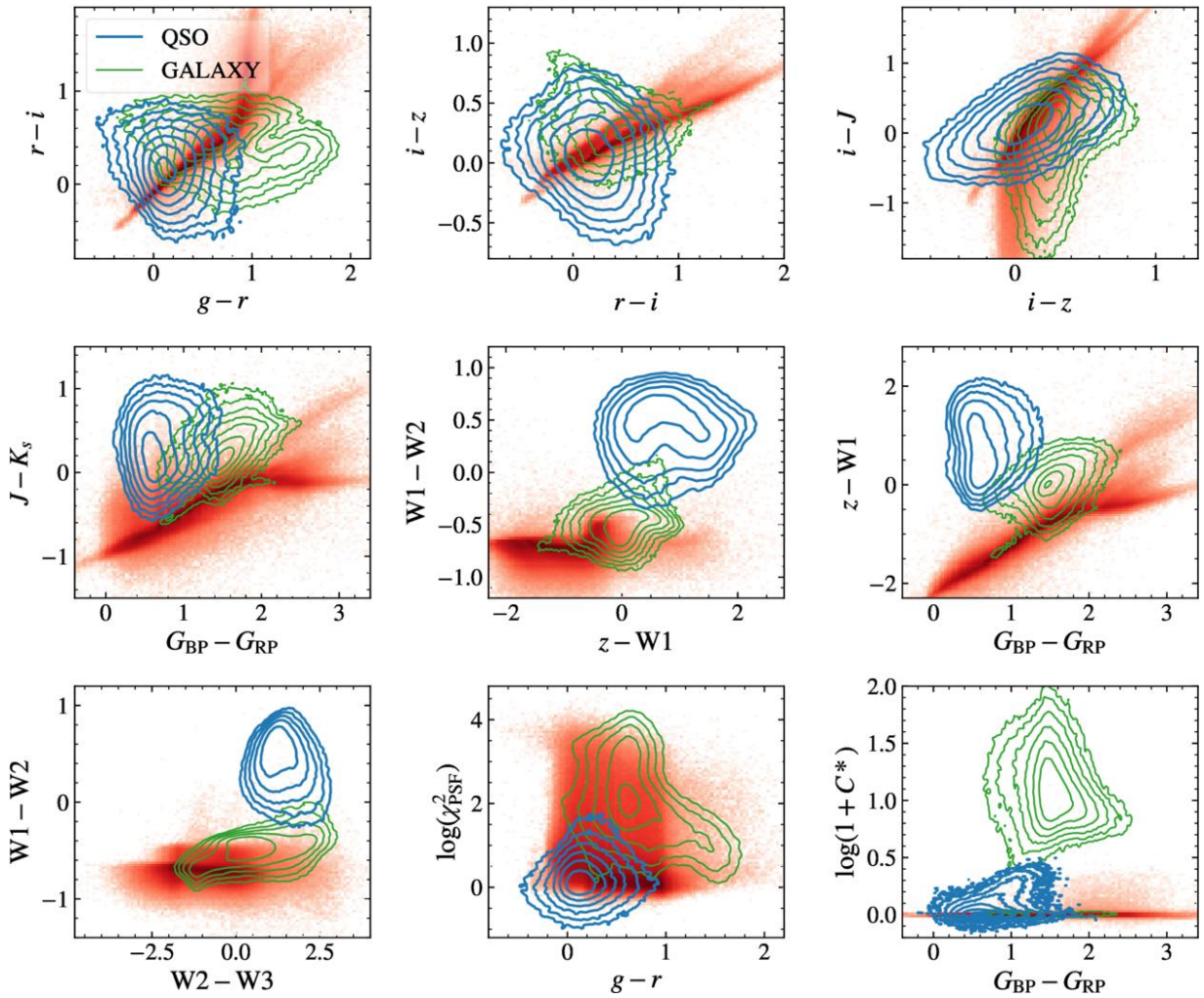
SMSS DR4巡天覆盖天区



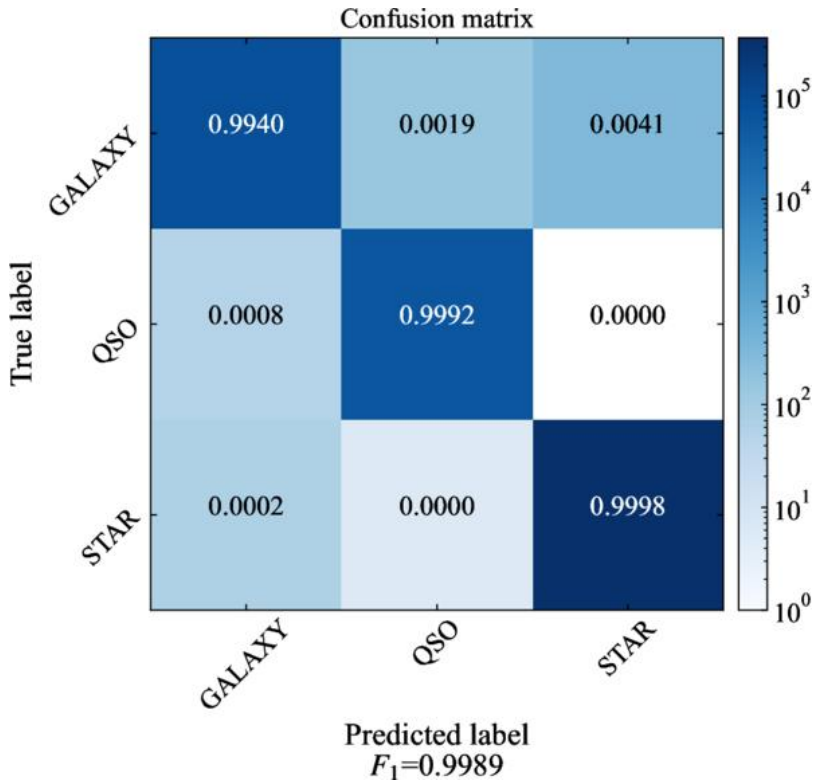
VISTA 巡天覆盖范围（2010版；VHS wiki）

XGBoost 分类模型训练

输入特征：W1, $g-r$, $r-i$, $i-z$, $i-J$, $J-K_s$, $i-$W1, $z-$W1, $J-$W1,
W1 – W2, W2 – W3, $G_{BP}-G_{RP}$, $G_{BP}-G$, $G-G_{RP}$, **$\log(\chi^2_{PSF})$, & $\log(1+C^*)$ (最后2个为形态参数)**



混淆矩阵

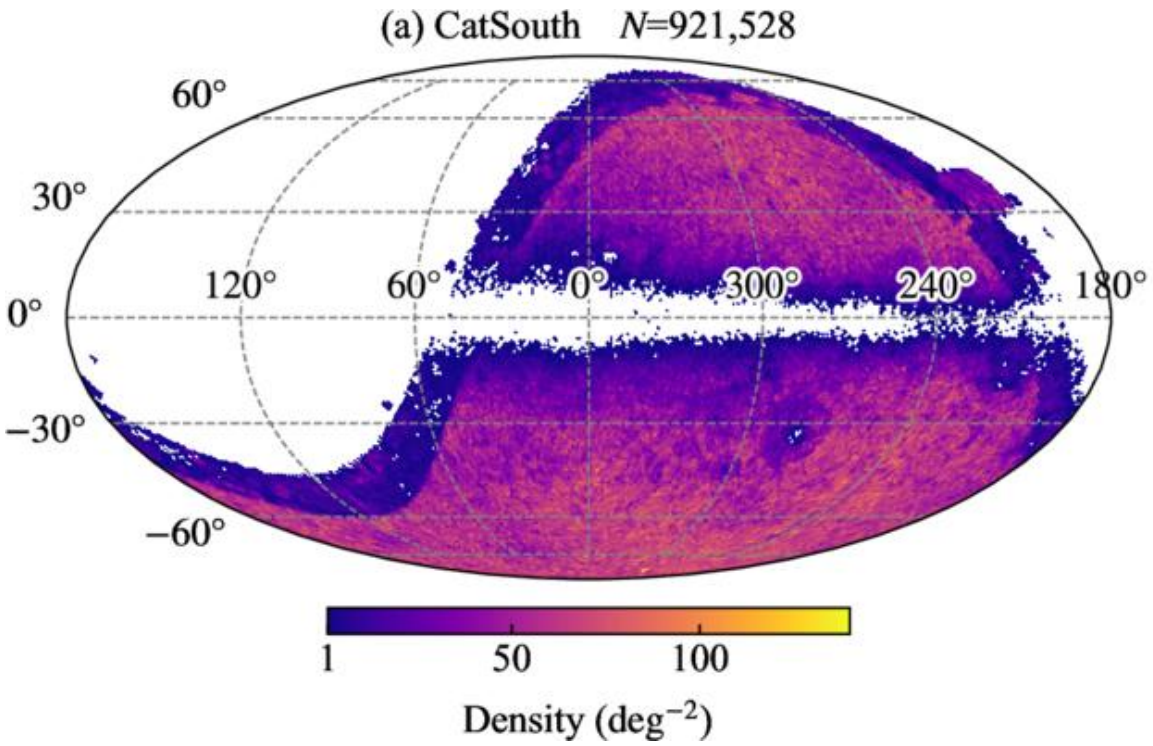


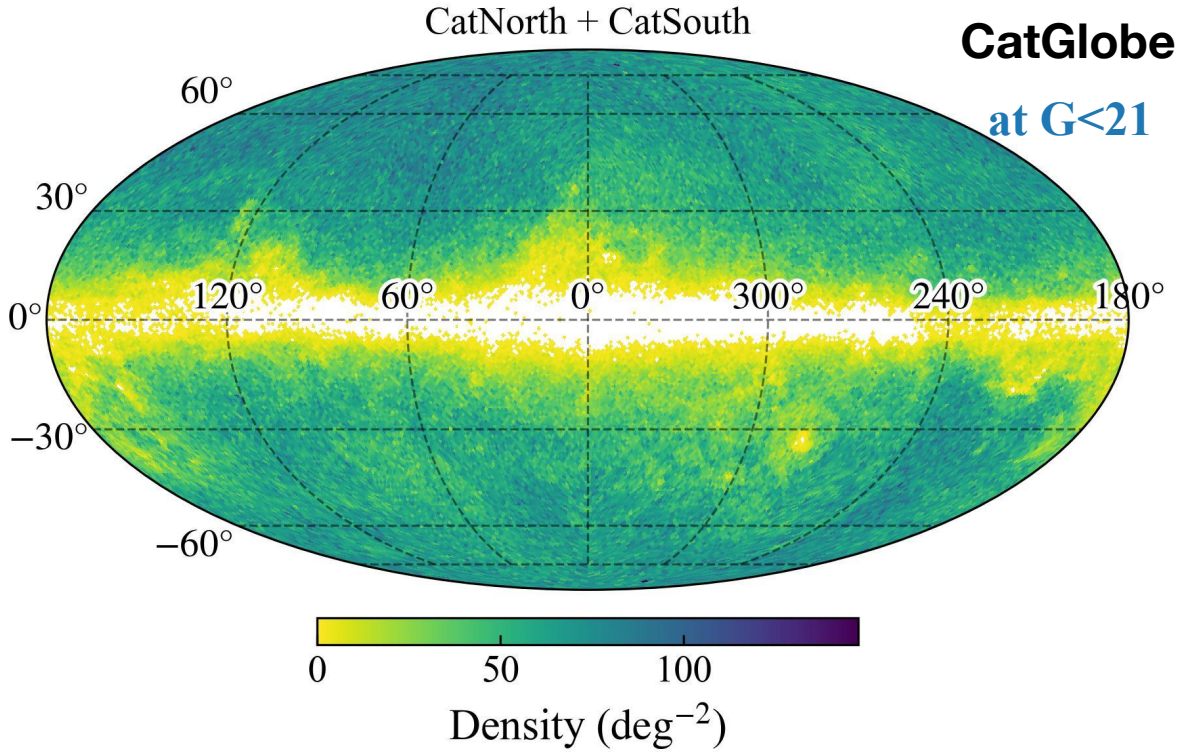
左图中红色密度分布为恒星，绿色为星系
蓝色为类星体

CatSouth/CatGlobe 类星体星表

(Fu, Wu +2025 ApJS, arXiv:2503.14141)

密度中值为60/deg^2，最大值接近150/deg^2

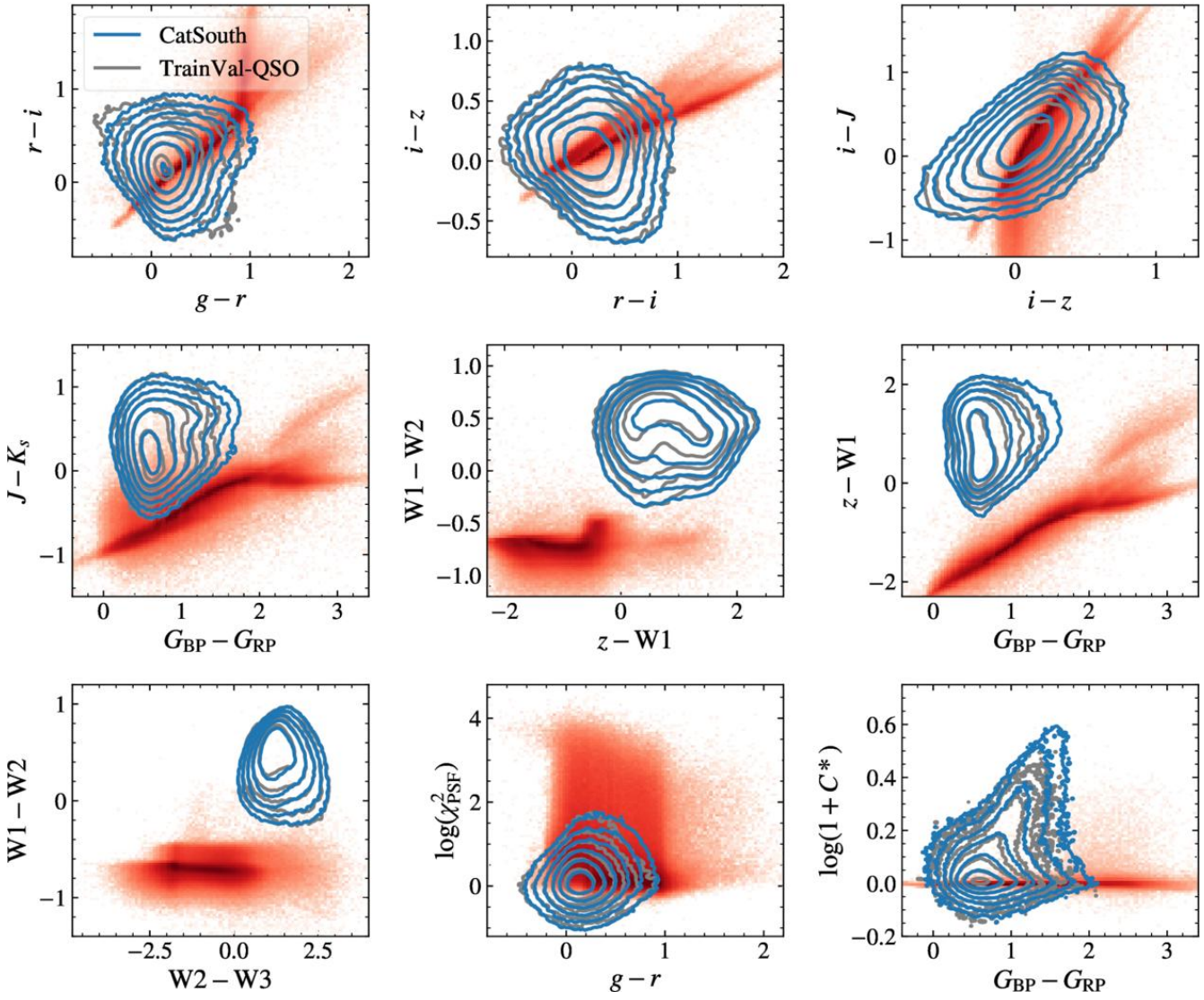


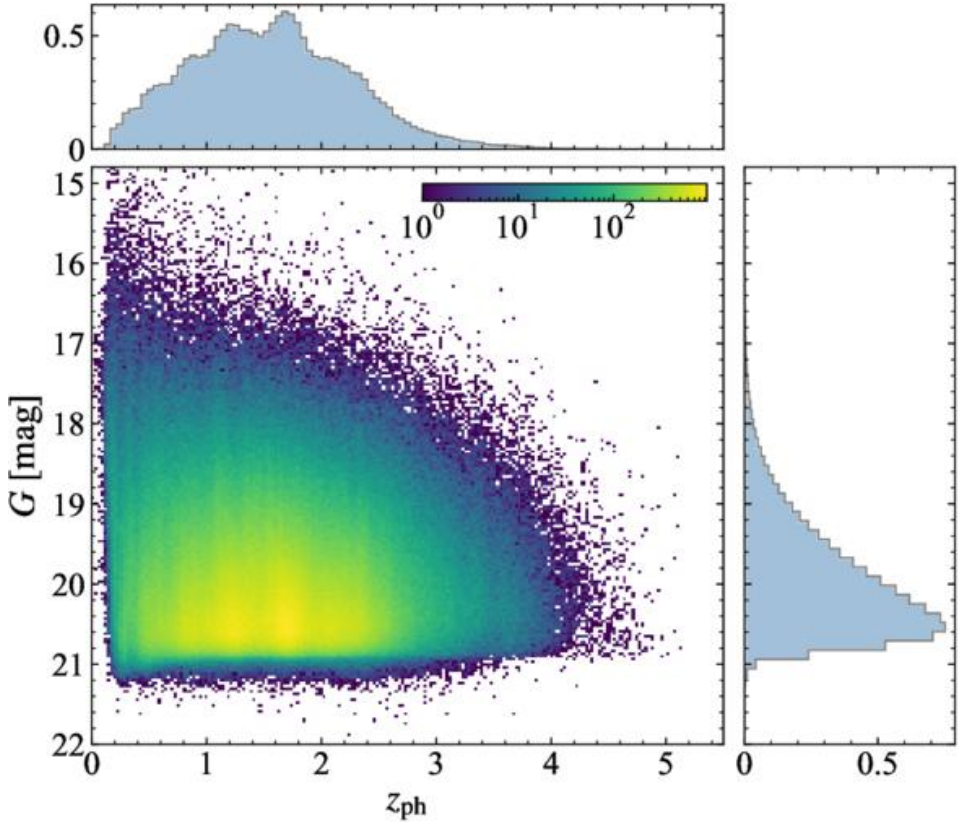


左： δ≲16° 的CatSouth类星体表（基于Gaia, WISE, Skymapper, VISTA），包含92万个源

右：**合并CatNorth与CatSouth得到的CatGlobe源表，包含约188万个源**

CatSouth 颜色和红移-星等分布

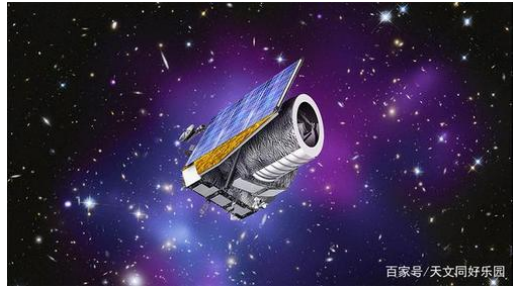


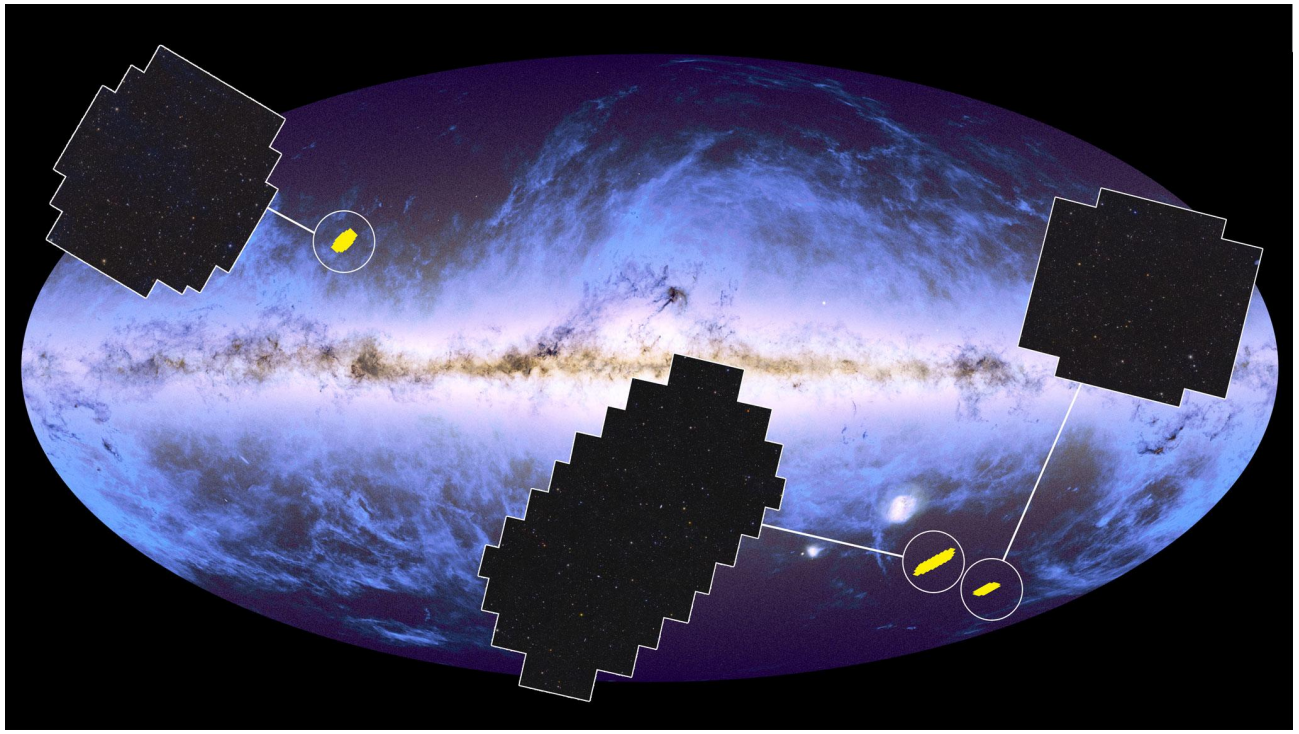


左图：

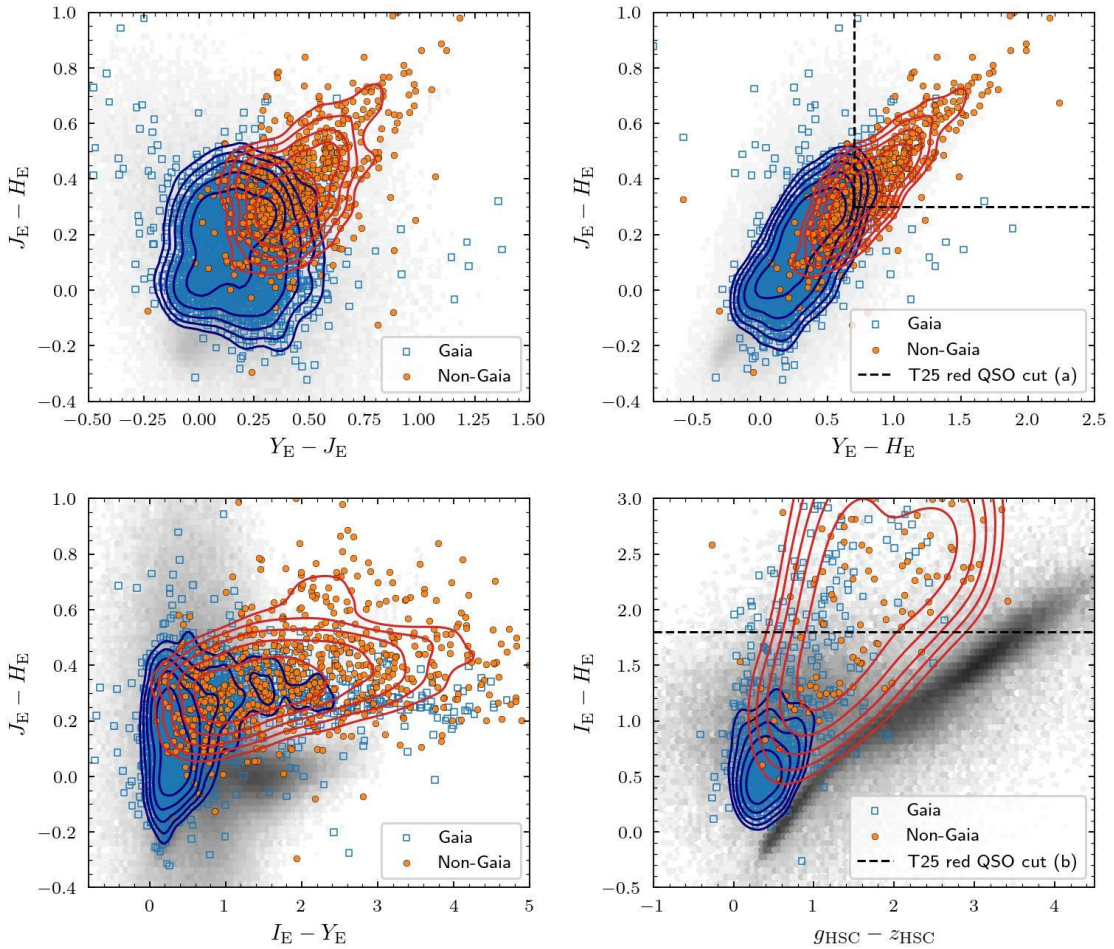
- 蓝色等密度线：CatSouth 类星体候选体
- 灰色等密度线：训练与验证集的类星体
- 红色密度图：恒星

CatGlobe 初步应用：
Euclid Q1光谱证认类星体的输入样本





Euclid Q1数据发布的63平方度天区示意图
Credits: Esa/Euclid/Euclid Consortium/Nasa; Esa/Gaia/Dpac; Esa/Planck Collaboration



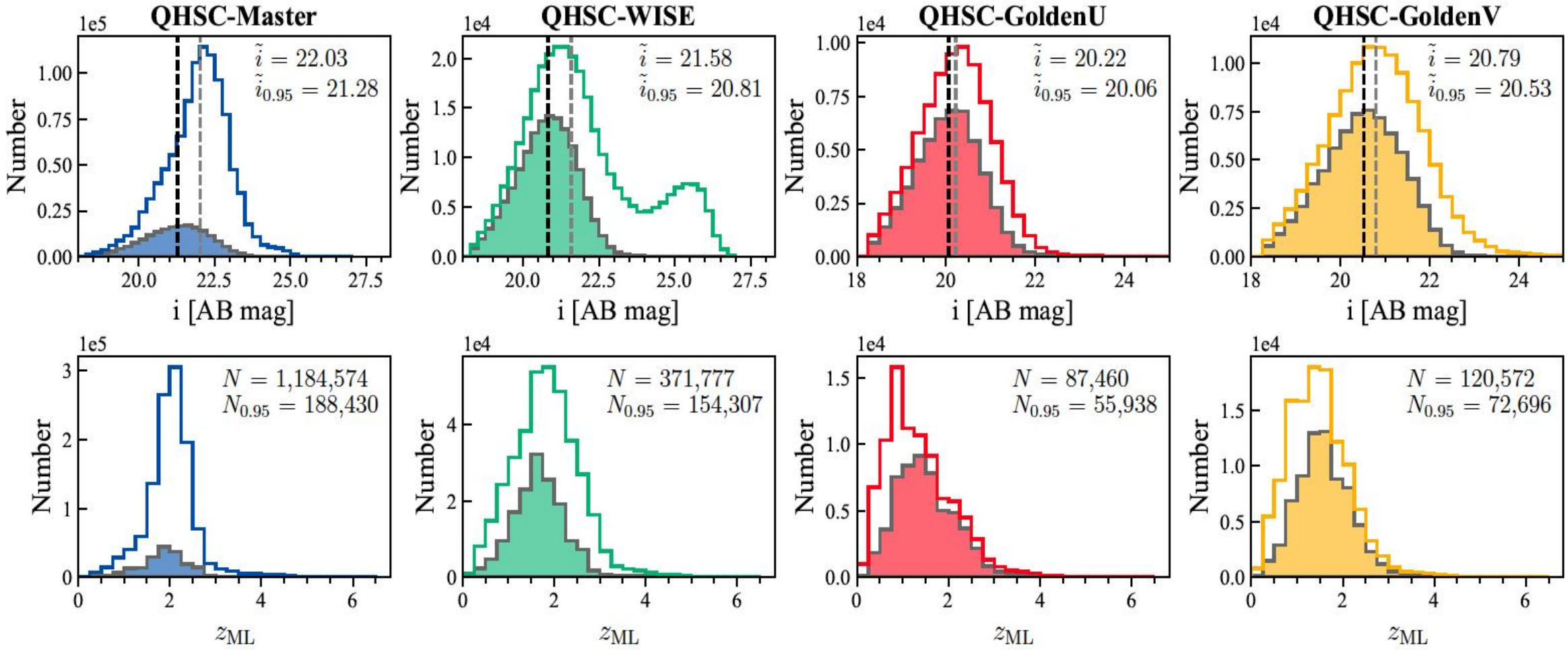
约3500个Euclid Q1光谱证认类星体的颜色分布，其中“Gaia”样本主要来自CatGlobe (Euclid Collaboration: Fu et al. 2025, to be submitted)

QHSC: Subaru/HSC 百万类星体候选体

Zhu, Wu, et al. 2025, ApJS, in press

北京大学 PEKING UNIVERSITY

(QHSC Paper) (QHSC Catalog)



4. 机器学习与CSST类星体巡天

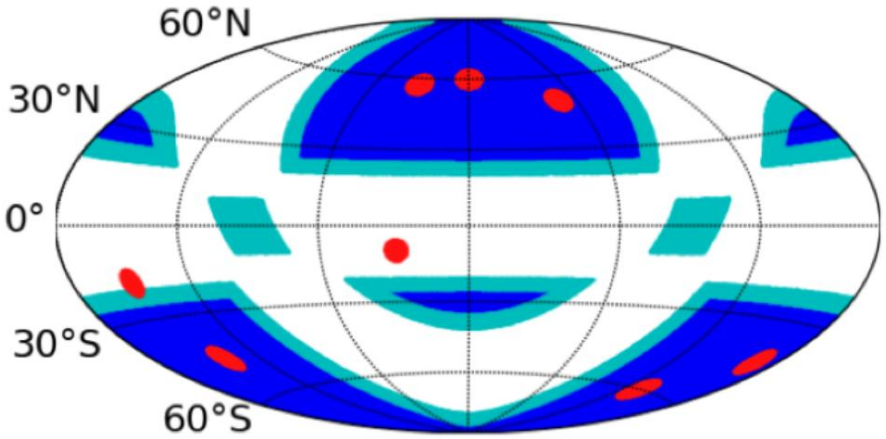


北京大学 PEKING UNIVERSITY

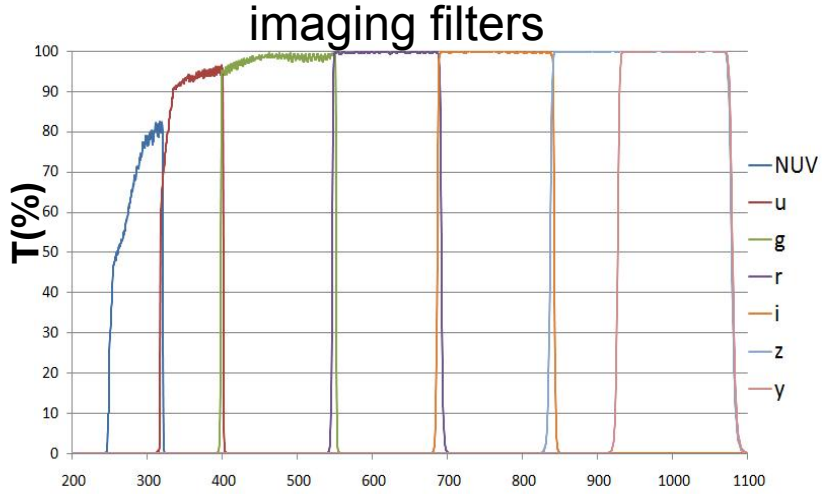
CSST：中国空间站巡天望远镜

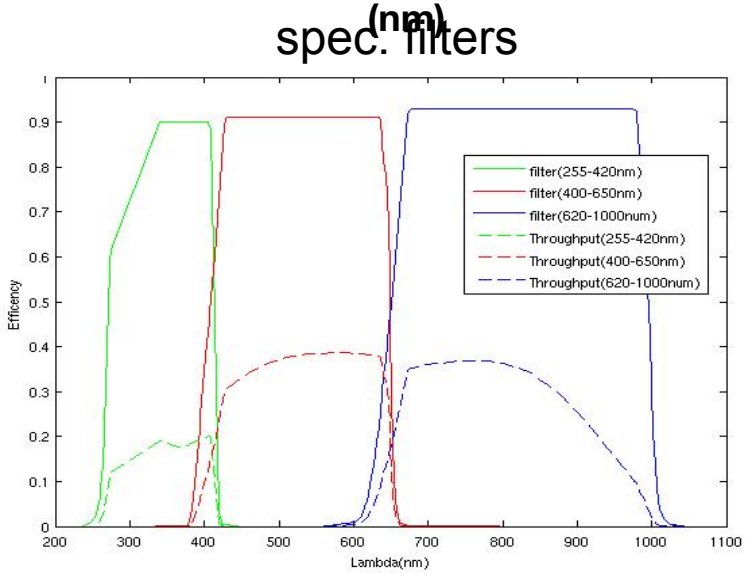
CSST 巡天计划

- 巡天区域：
  - Wide field survey: 17,500 deg² (~26等)
  - Deep field survey: 400 deg²(~27等)
  - Ultra-deep field survey (TBD): 9 deg²



- **CSST类星体巡天将产生上千万可靠类星体候选体和超过百万无缝光谱证认的类星体！**



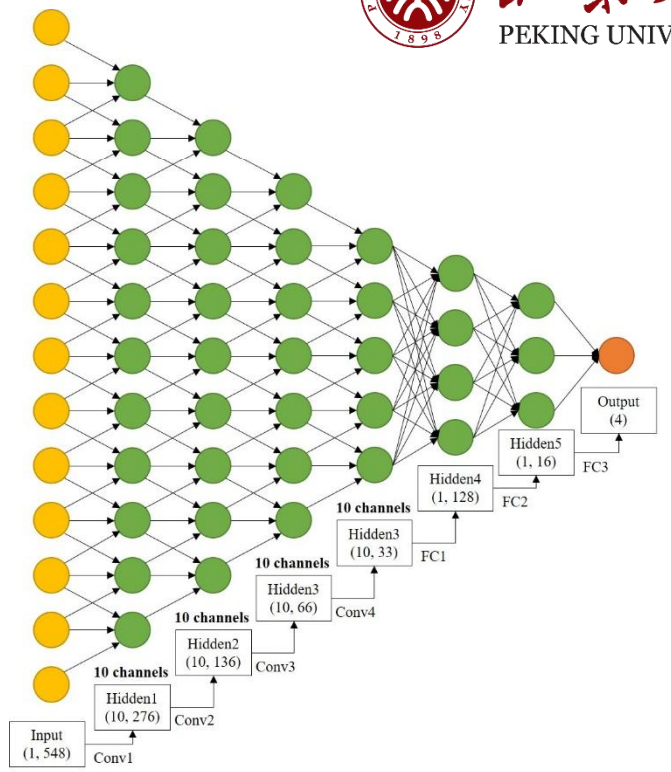


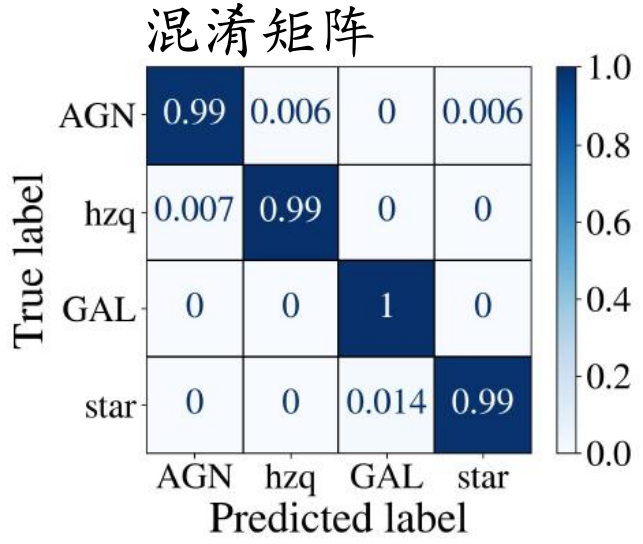
类星体模拟选源

- 光谱模拟数据处理:
  - 数据清洗（去除离群值, 并将光谱流量正则化）
  - 将光谱统一重采样到同一个波长序列

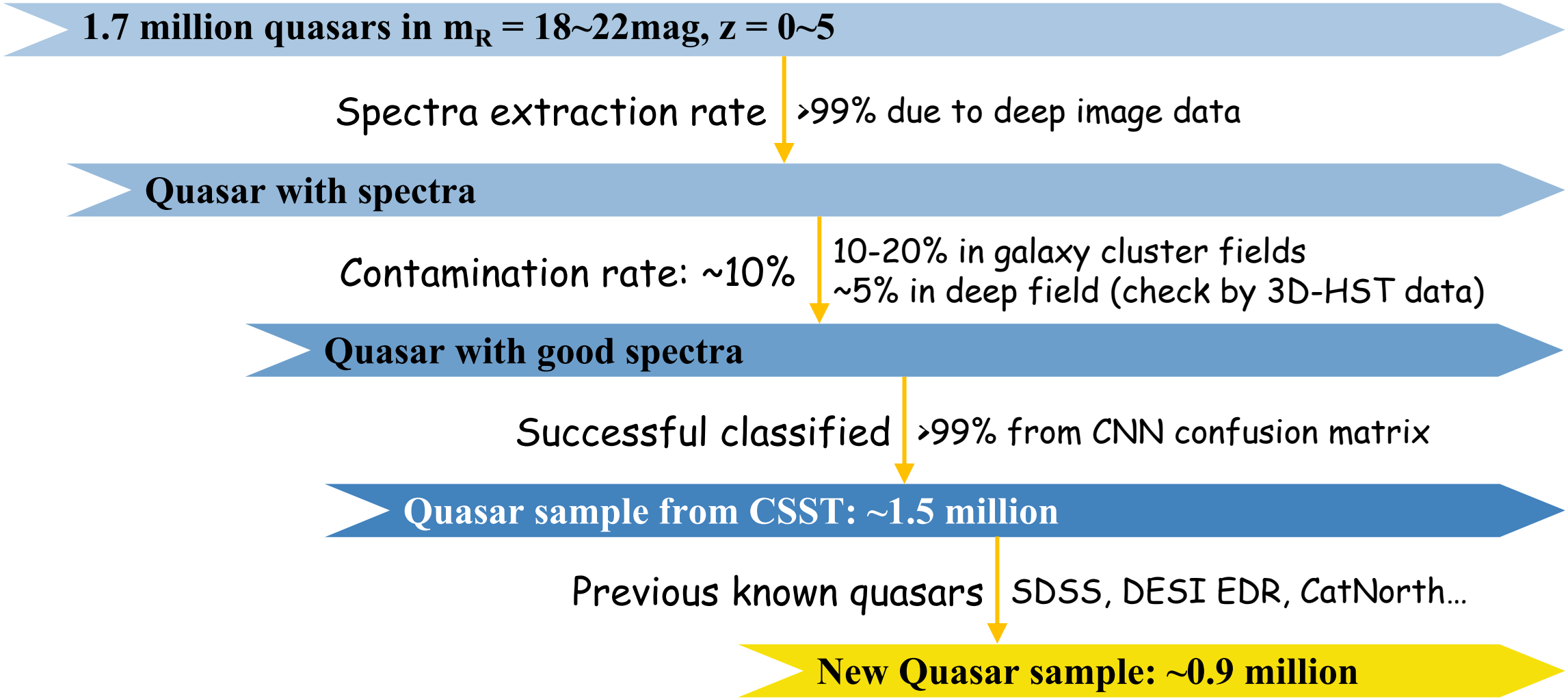
- 分类器：卷积神经网络（CNN）
  - 分类: 低红移类星体, 高红移类星体(Lyα 线进入GU, z~2)，　　星系, 恒星
  - 结构: 4 卷积层 + 3 全连接层

- 结果: 类星体挑选>99%成功率





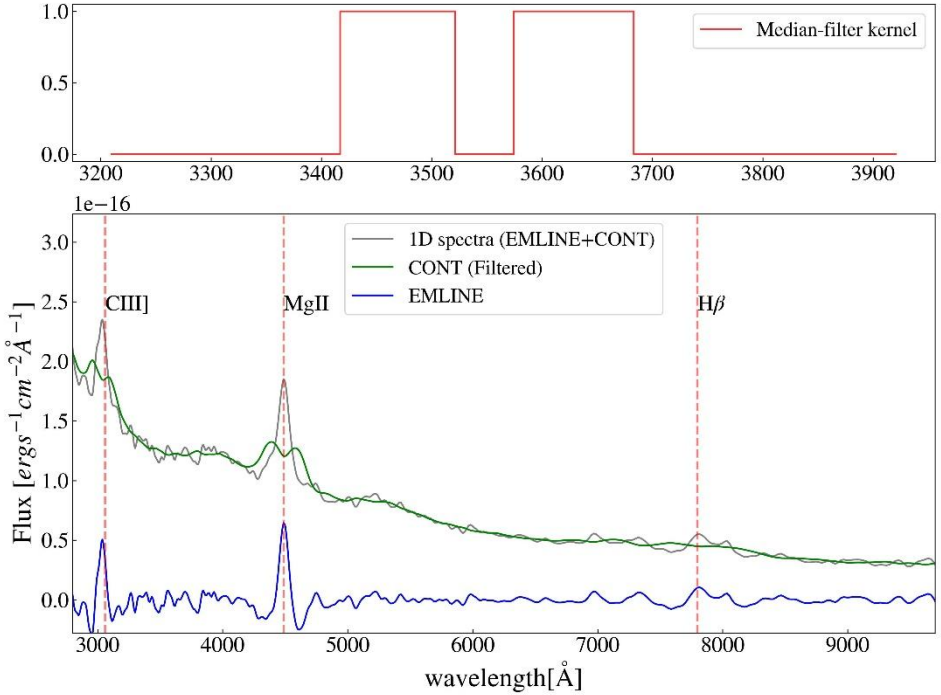
类星体数目估计

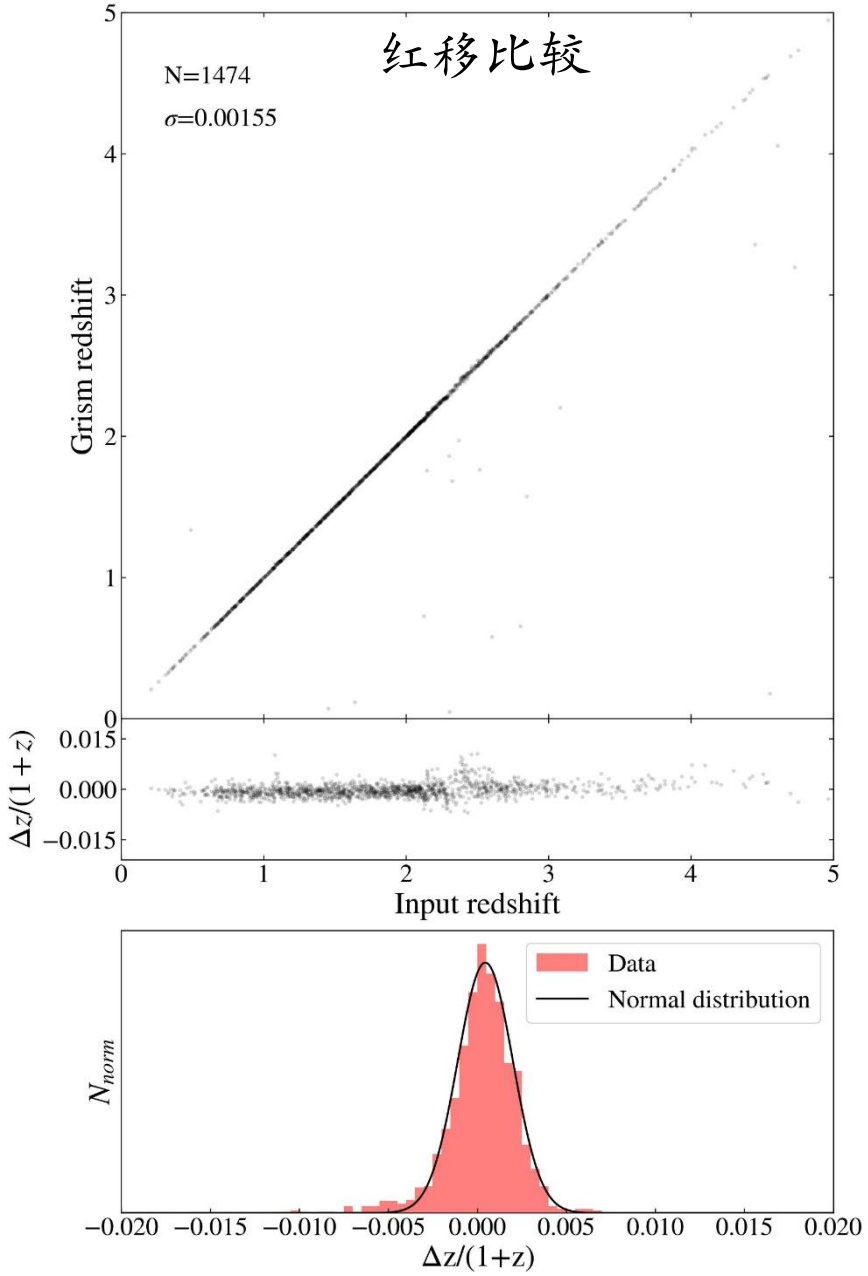


Pang, Wu et al. 2025, ApJ, 980, 223

类星体红移测量

- 方法：
  - 中值滤波识别发射线中心波长
  - 类星体强发射线匹配
- 结果：
  - $\sigma_{NMAD} = 0.0016$ (略优于HST grism)
  - $f_c = 1.5\%$ $(\Delta z/(1+z) > 0.01)$





Select quasars from HST grism data

3D-HST Catalog

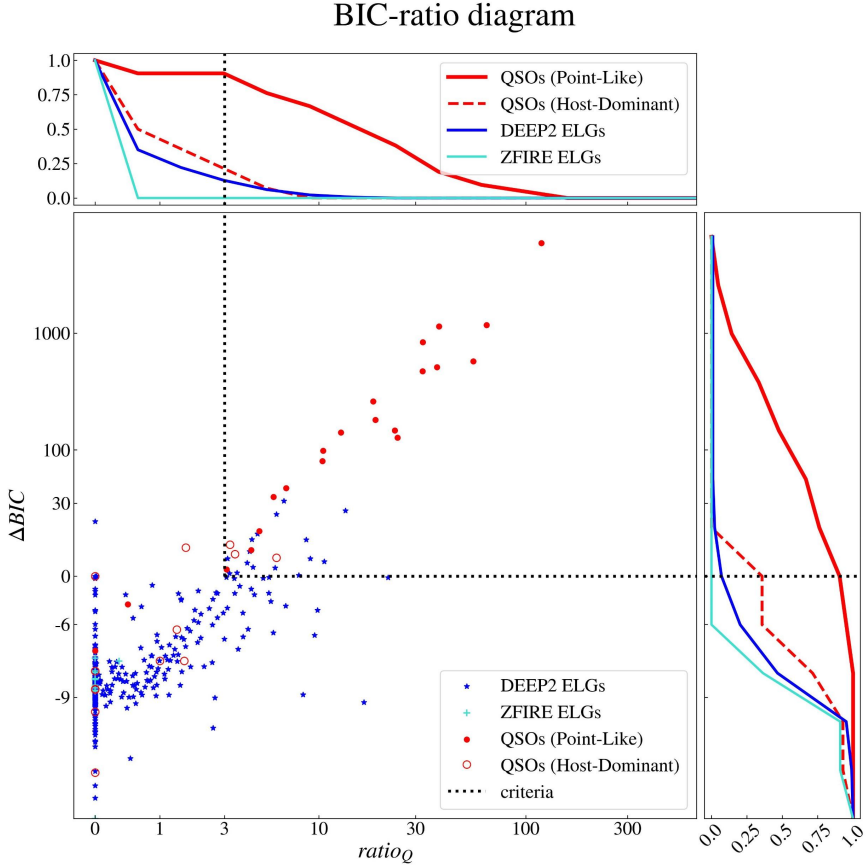
↓

"Broad" Emission Line Sample

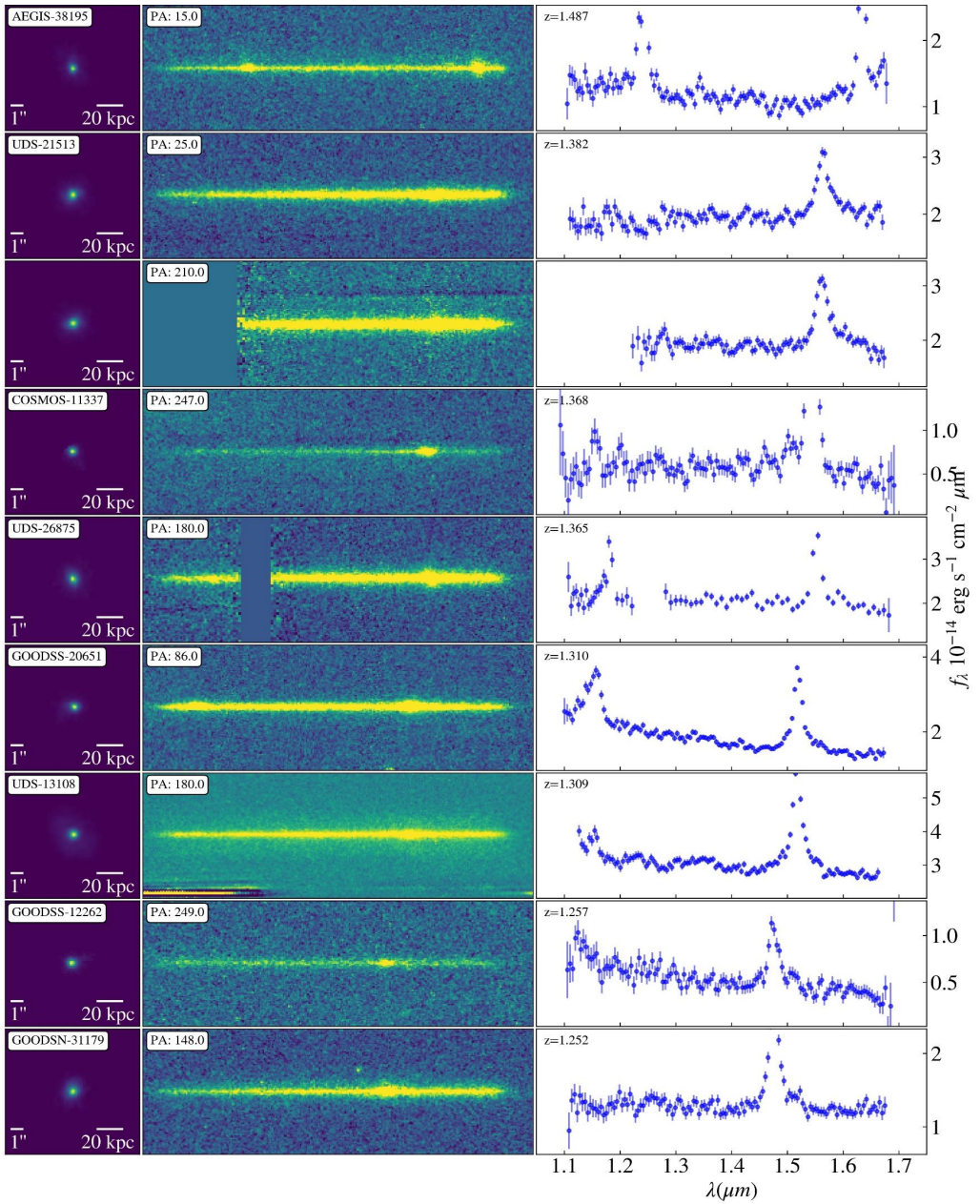
↓

QSO Candidates

BIC: Bayesian Information Criterion

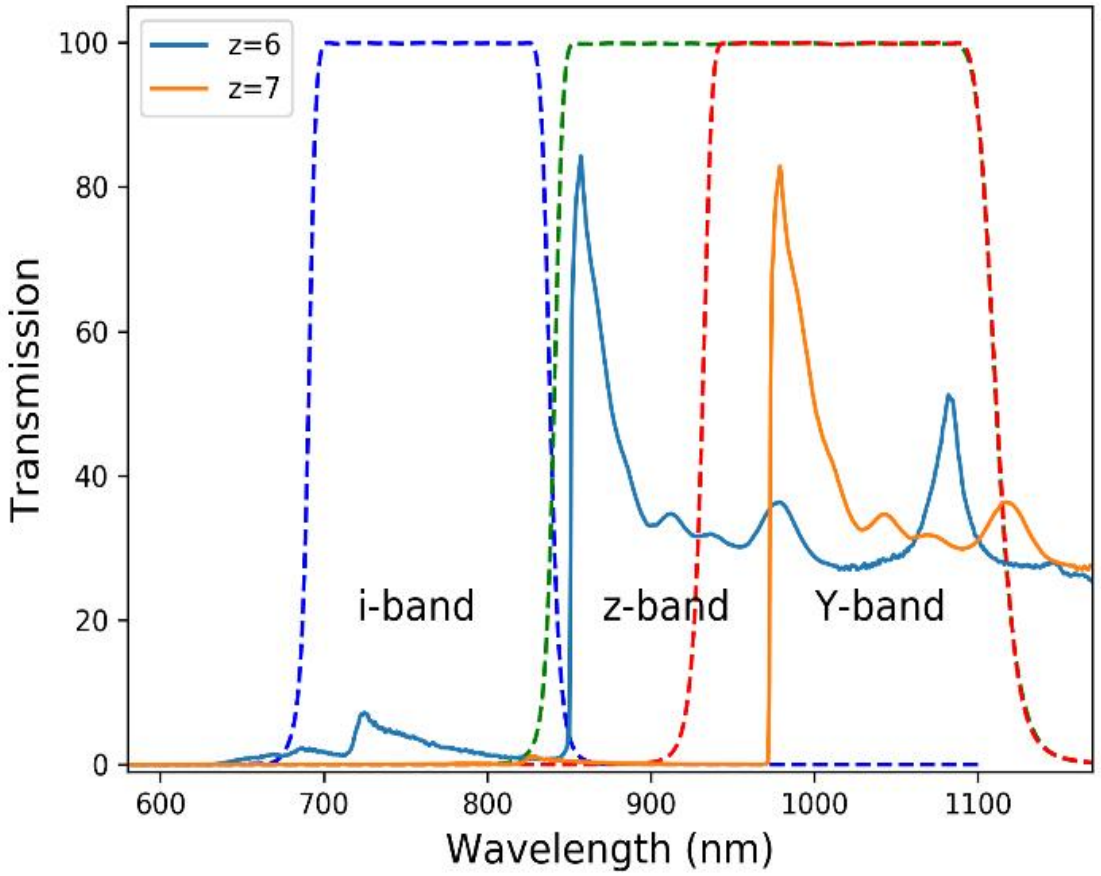


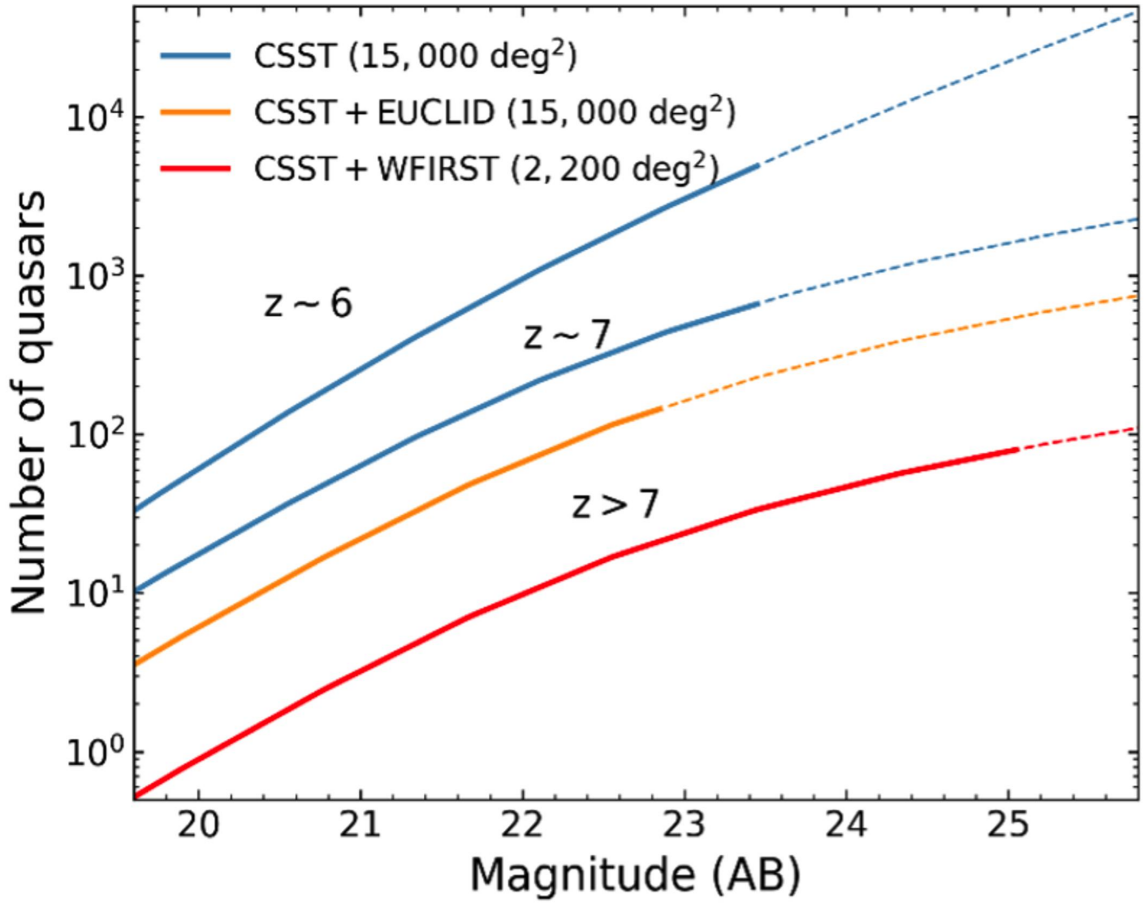
quasar templates improve the fitting



Pang, Wu, Fu et al. 2025, MNRAS

High-z quasars in CSST (a few thousands z>6 quasars)





5. 总结

- 类星体巡天是发展迅速、应用广泛领域，机器学习对类星体巡天非常重要，可为研究类星体及其演化提供宝贵样本
- 使用迁移学习构建训练样本，开展银道面背景类星体选源，选出16万多个候选体，已证认1500个银道面背景类星体
- 使用机器学习XGBoost算法和测光巡天数据集成分类模型对Gaia DR3类星体候选体源表进行了重新分类，结合Gaia自行数据，构建了154万源、有高纯度(>80%)和高完备度的CatNorth和188万源CatGlobe类星体候选体星表。利用Subaru数据构建了110万源的QHSC候选体星表
- CSST测光巡天将获得千万类星体候选体，使用卷积神经网络模型可高效从CSST模拟无缝光谱数据中筛选类星体，预计CSST可光谱证认百万类星体，并获得物理参数

