Multiwavelngth cross-match

M. Salvato (MPE)

Identifying correct counterparts to high-energy sources by "multiwavelength educated guesses" imbibed in a Bayesian statistic environment

a.k.a. NWAY (Salvato et al 2018, 2022)

M. Salvato (MPE)

Identifying correct counterparts to high-energy sources by "multiwavelength educated guesses" imbibed in a Bayesian statistic environment

a.k.a. NWAY (Salvato et al 2018, 2022)

M. Salvato (MPE)

P.S.: it works for any cross-match!

The talk

- **Motivations**
- **Issues**
- **A/The Solution**
- **Comparison between methods**
- **Identification & Classification**
- **Final remarks**

CTP means “counterpart” in this talk

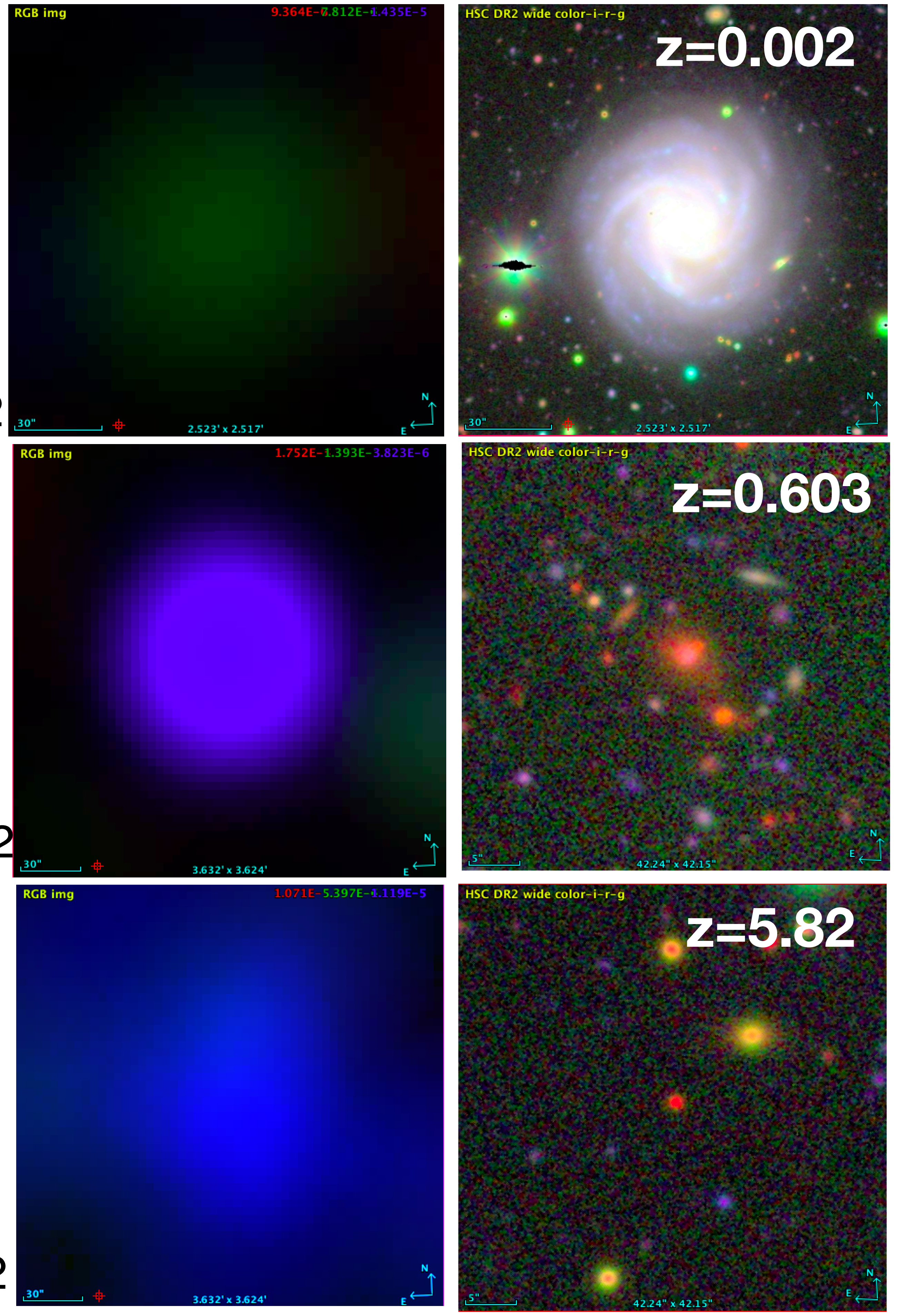
The challenge: find the right counterpart

eROSITA/eFEDS HSC

Vulic et al 2022

Brusa et al 2022

Wolf et al 2022



…but it could be also a Galactic source (e.g., Schneider et al 2022, Stelze et al 2022)
or an unresolved cluster (e.g., Bulbul et al 2022)

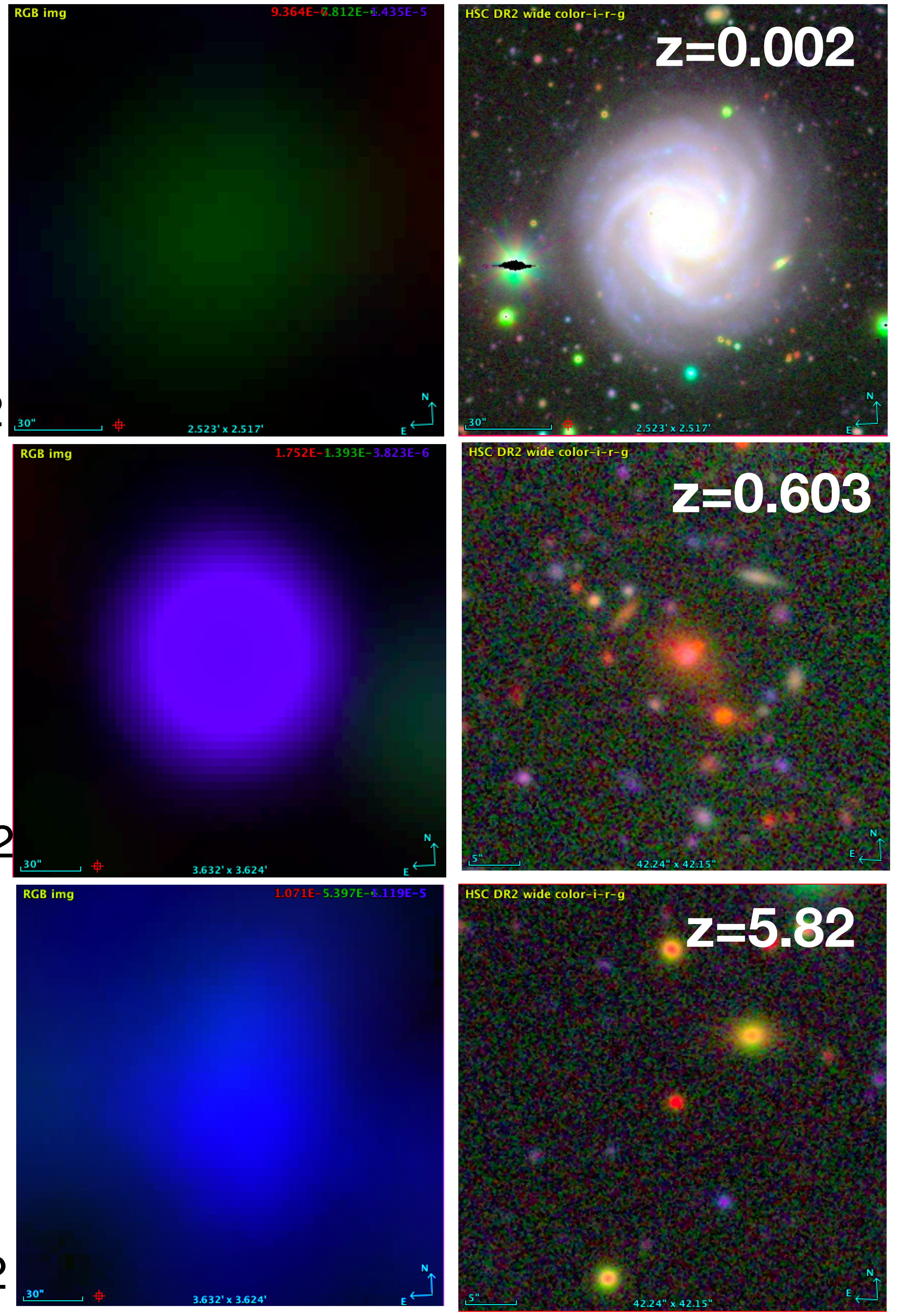
The challenge: find the right counterpart

eROSITA/eFEDS HSC

Vulic et al 2022

Brusa et al 2022

Wolf et al 2022



- The size of the area search depends on the positional error which correlates with the intensity of the source.

…but it could be also a Galactic source (e.g., Schneider et al 2022, Stelze et al 2022) or an unresolved cluster (e.g., Bulbul et al 2022)

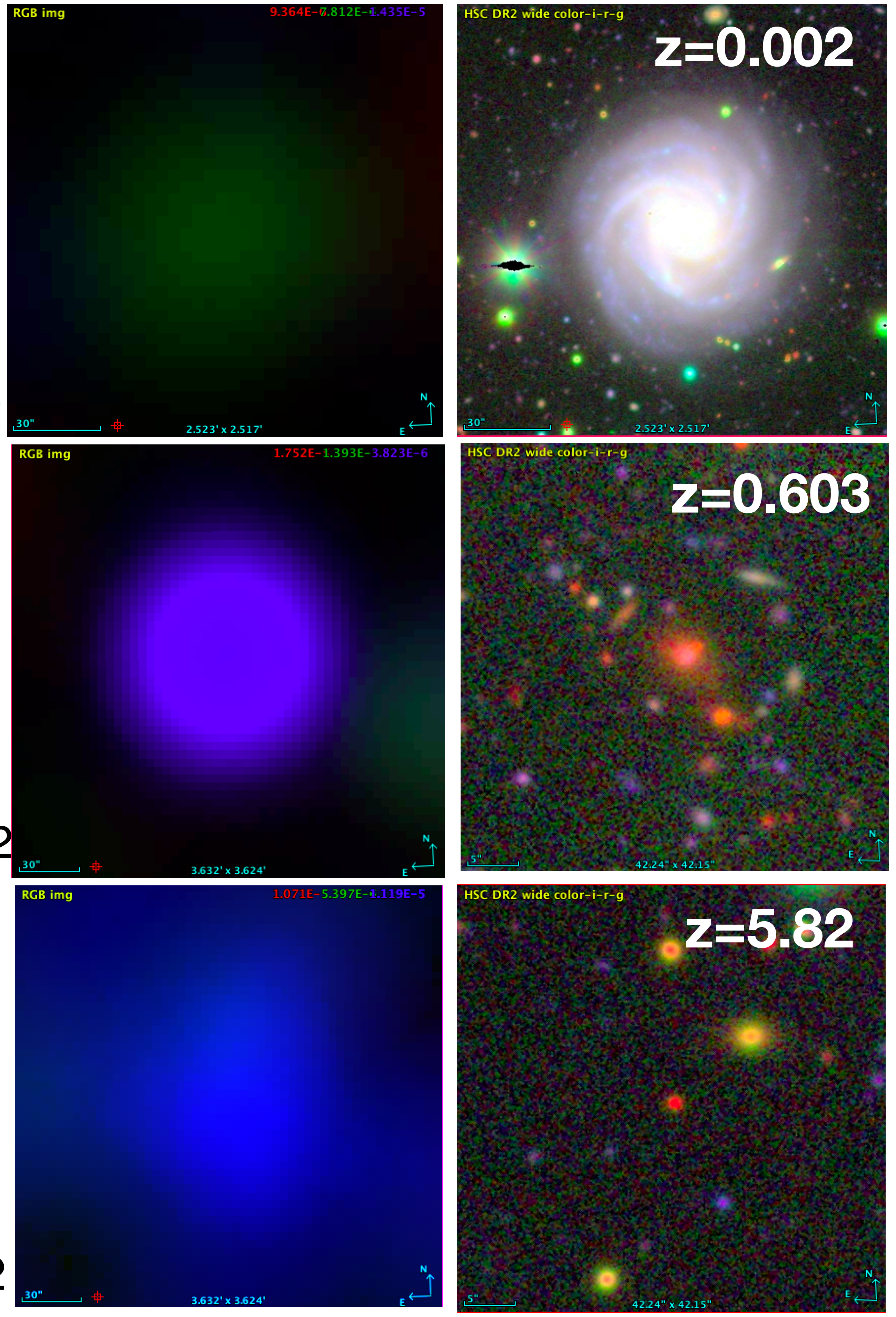
The challenge: find the right counterpart

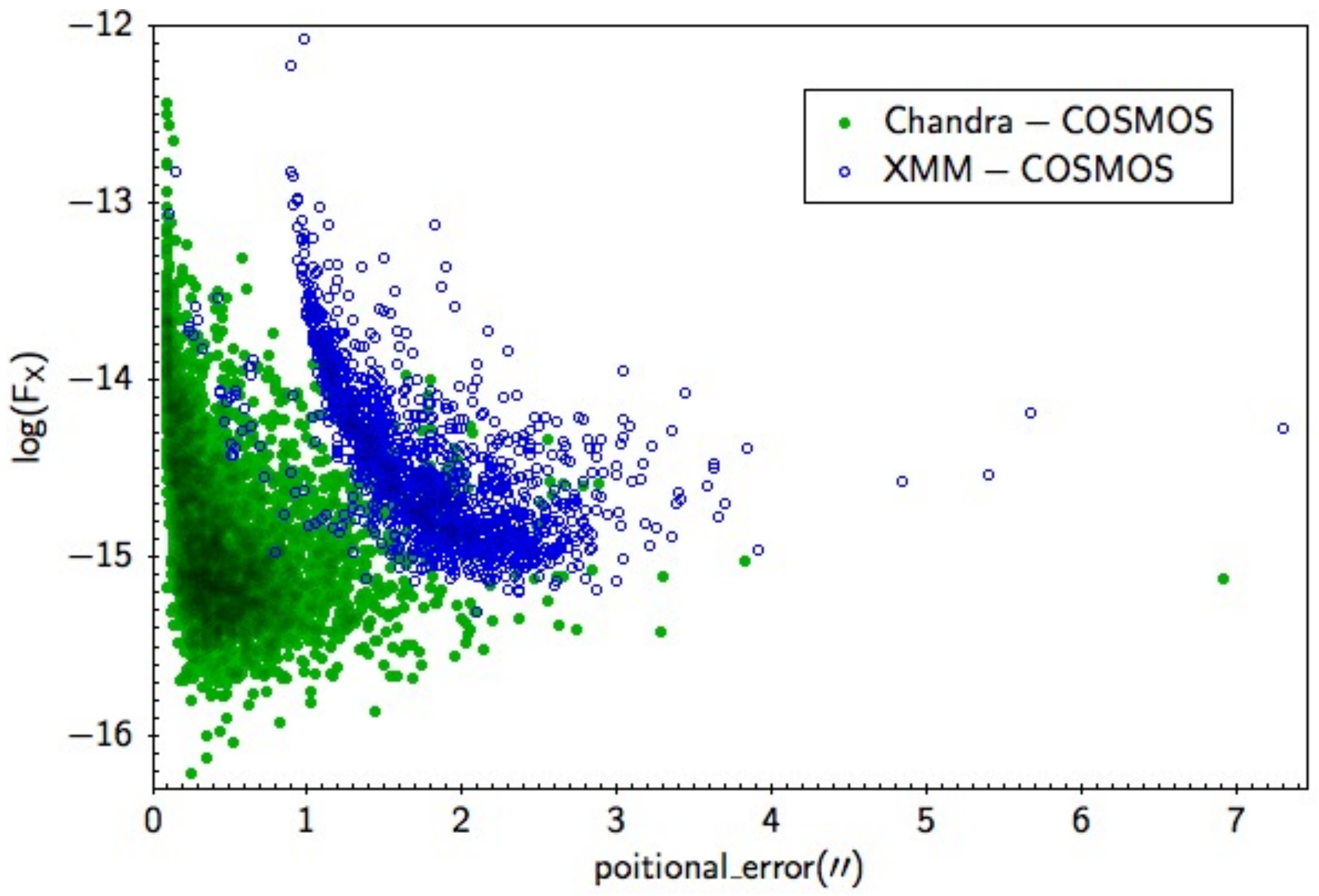
eROSITA/eFEDS HSC

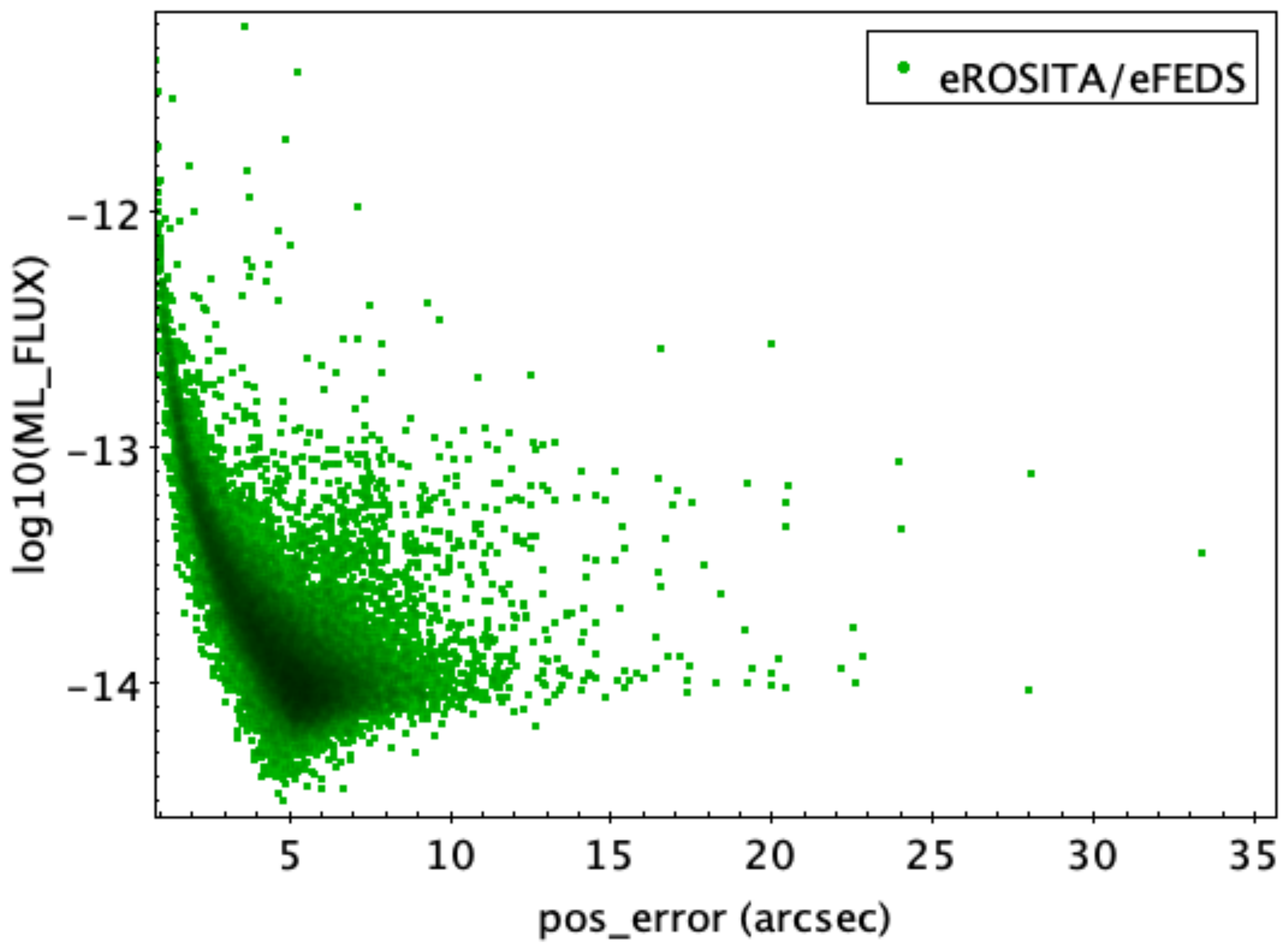
Vulic et al 2022

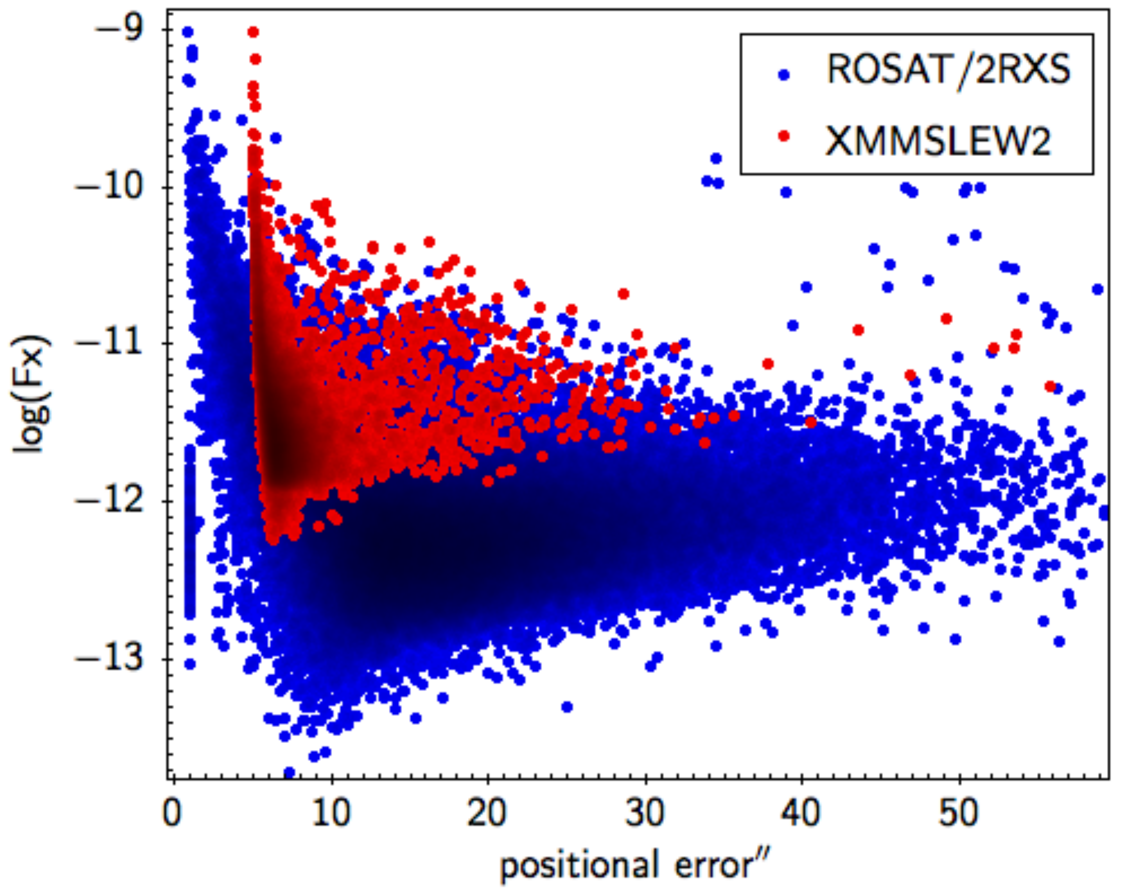
Brusa et al 2022

Wolf et al 2022







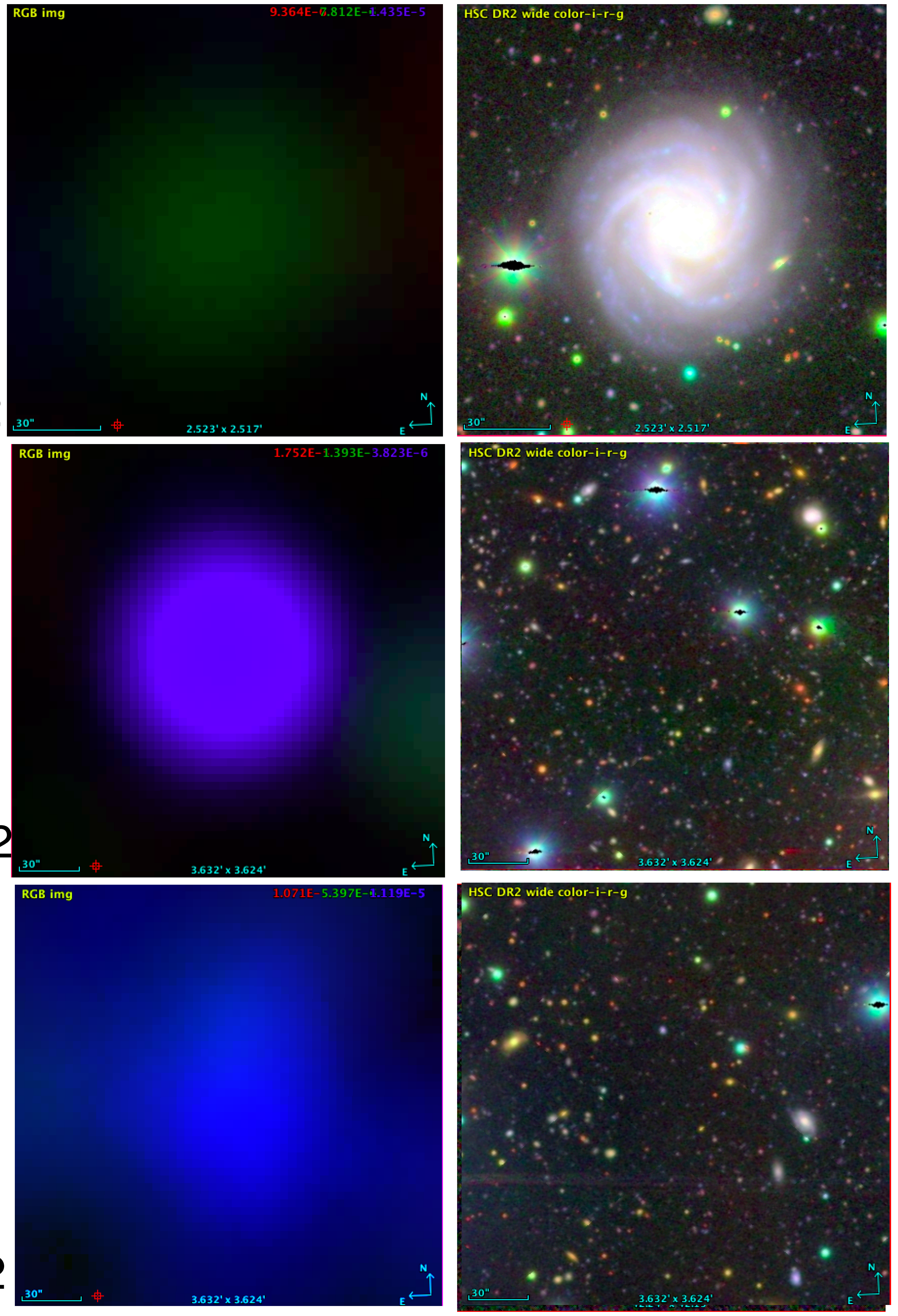


- The size of the area search depends on the positional error which correlates with the intensity of the source.

…but it could be also a Galactic source (e.g., Schneider et al 2022, Stelze et al 2022)
or an unresolved cluster (e.g., Bulbul et al 2022)

The challenge: find the right counterpart

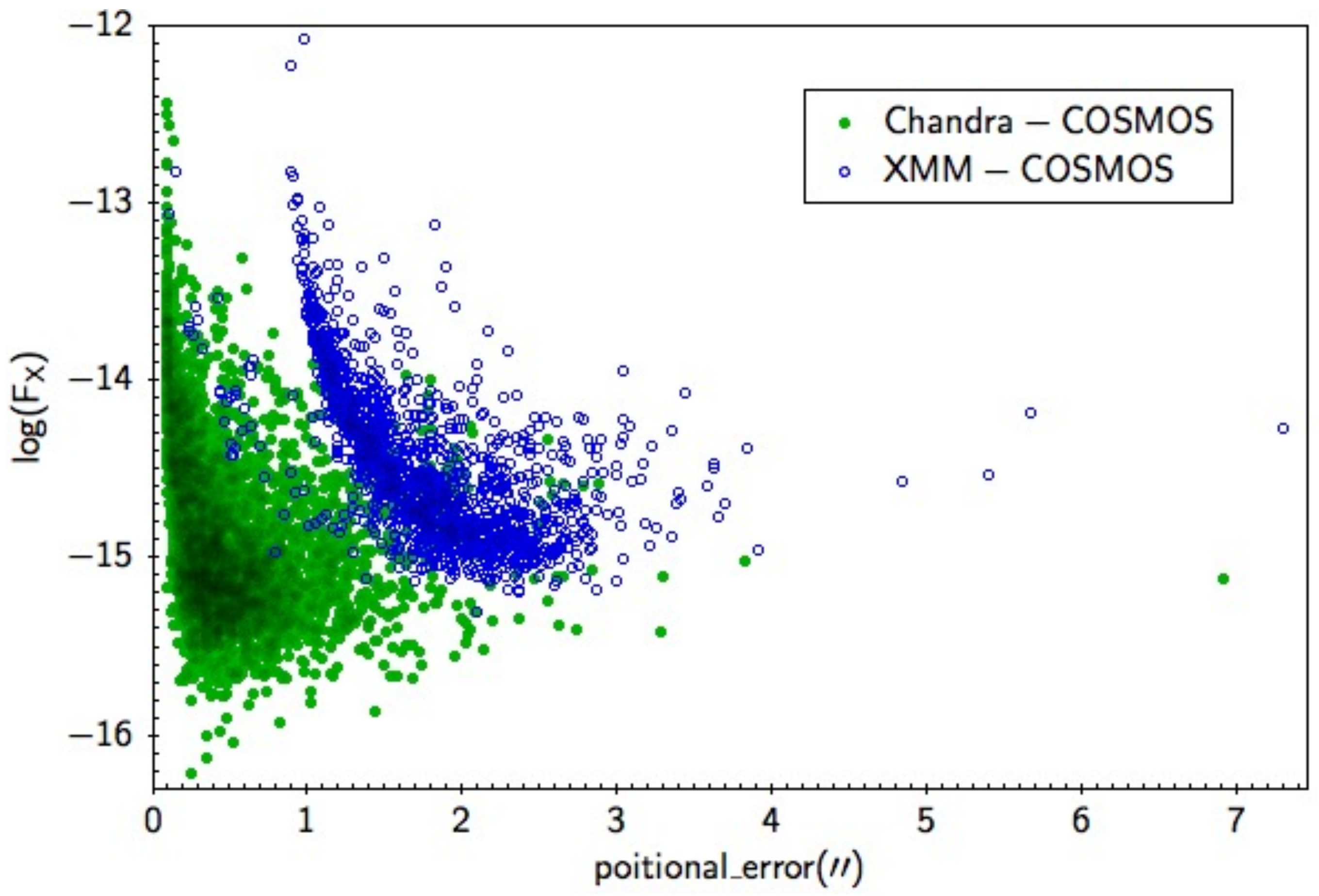
eROSITA/eFEDS HSC

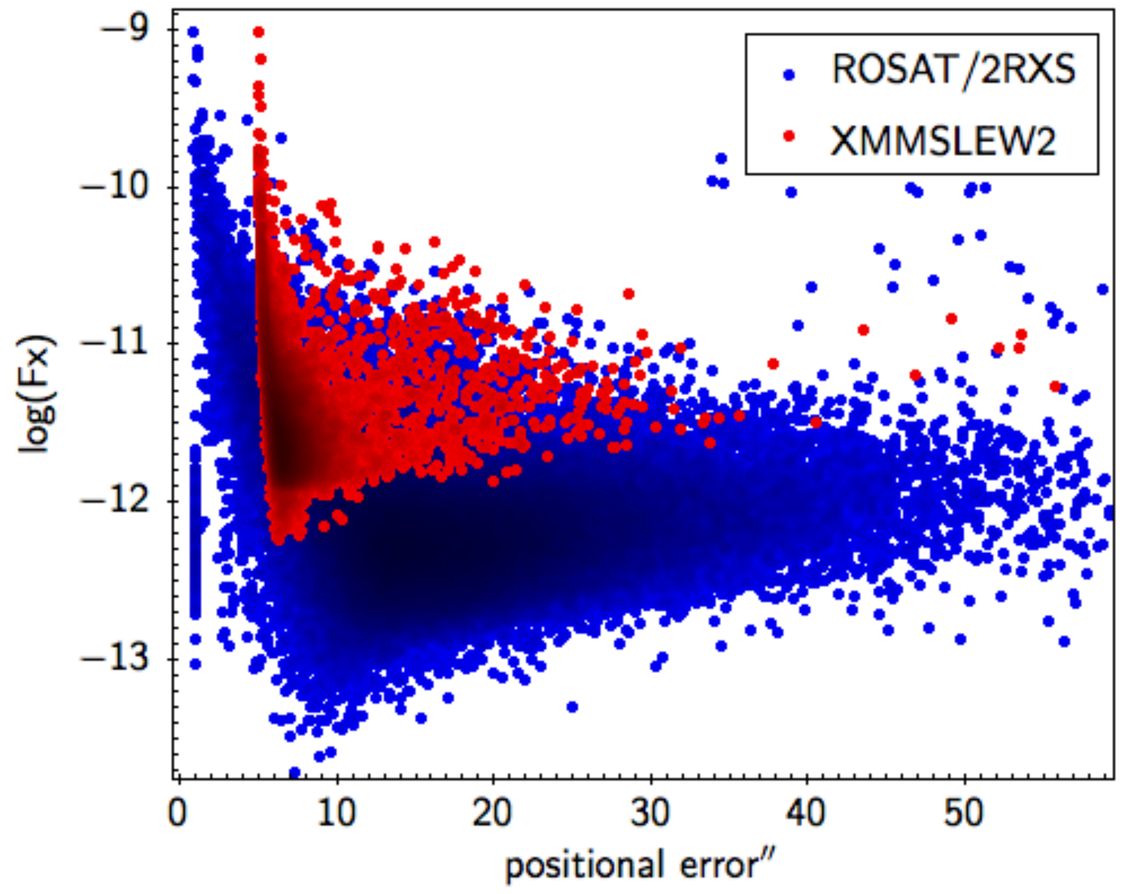


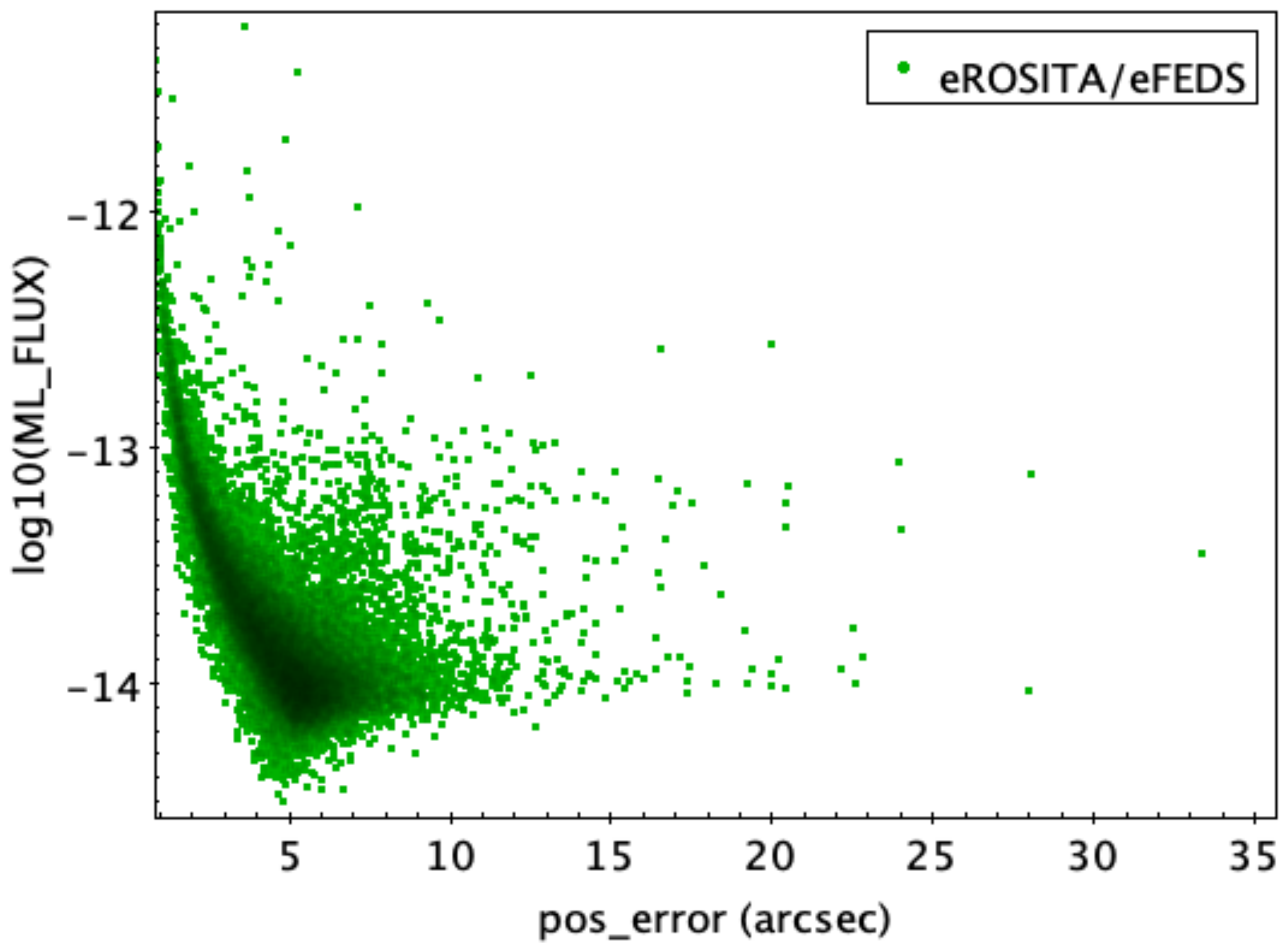
Vulic et al 2022

Brusa et al 2022

Wolf et al 2022







- The size of the area search depends on the positional error which correlates with the intensity of the source.

…but it could be also a Galactic source (e.g., Schneider et al 2022, Stelze et al 2022) or an unresolved cluster (e.g., Bulbul et al 2022)

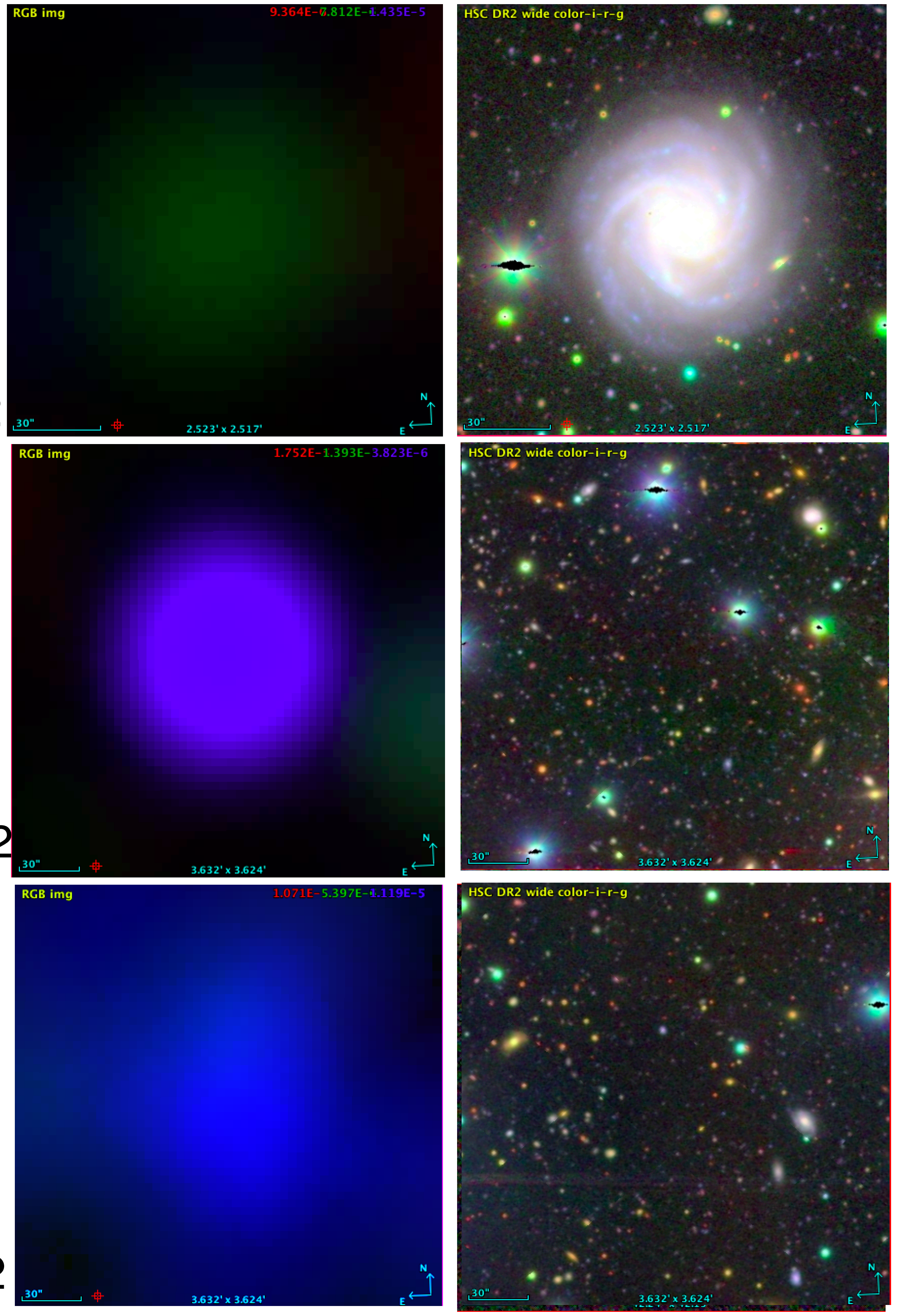
The challenge: find the right counterpart

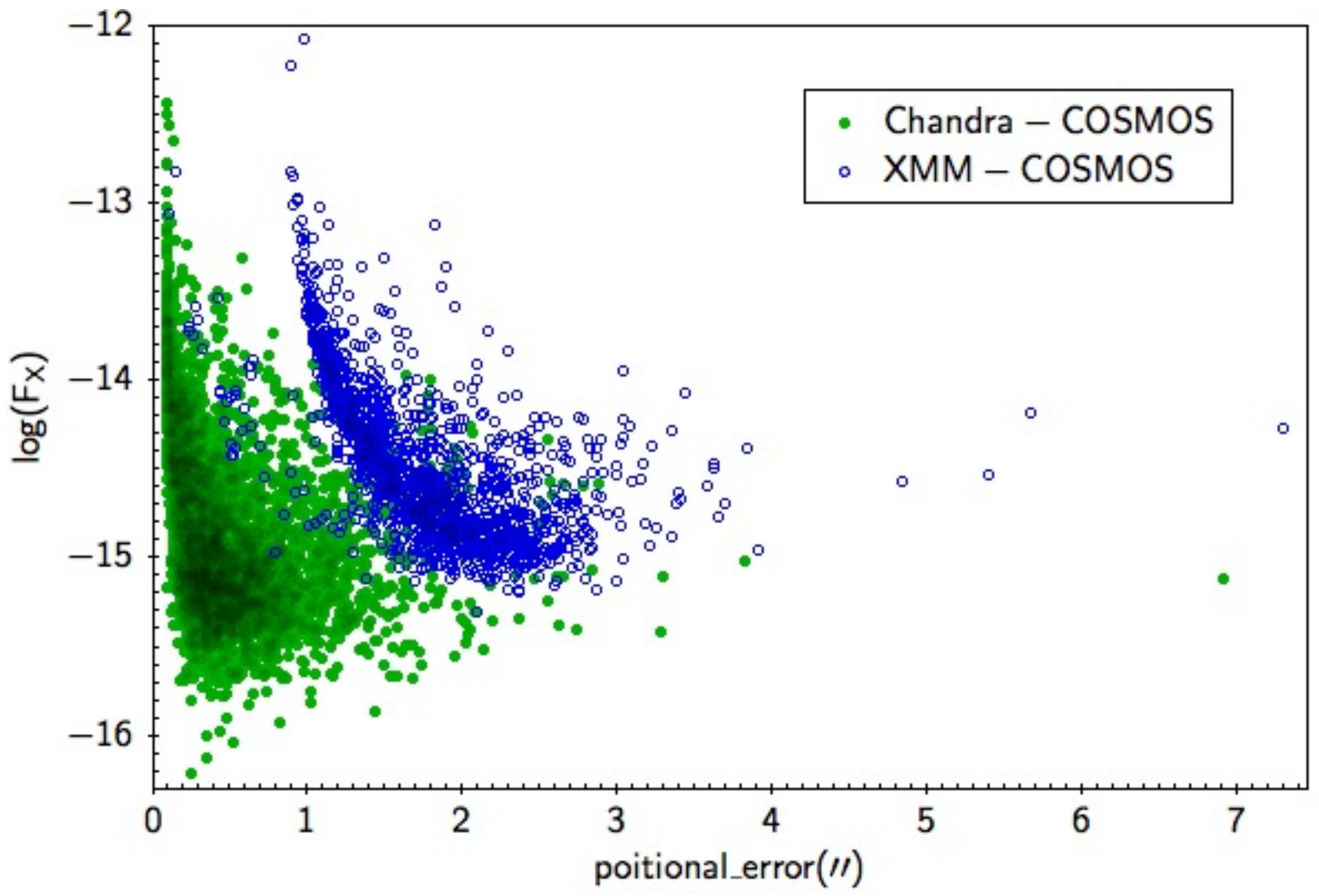
eROSITA/eFEDS HSC

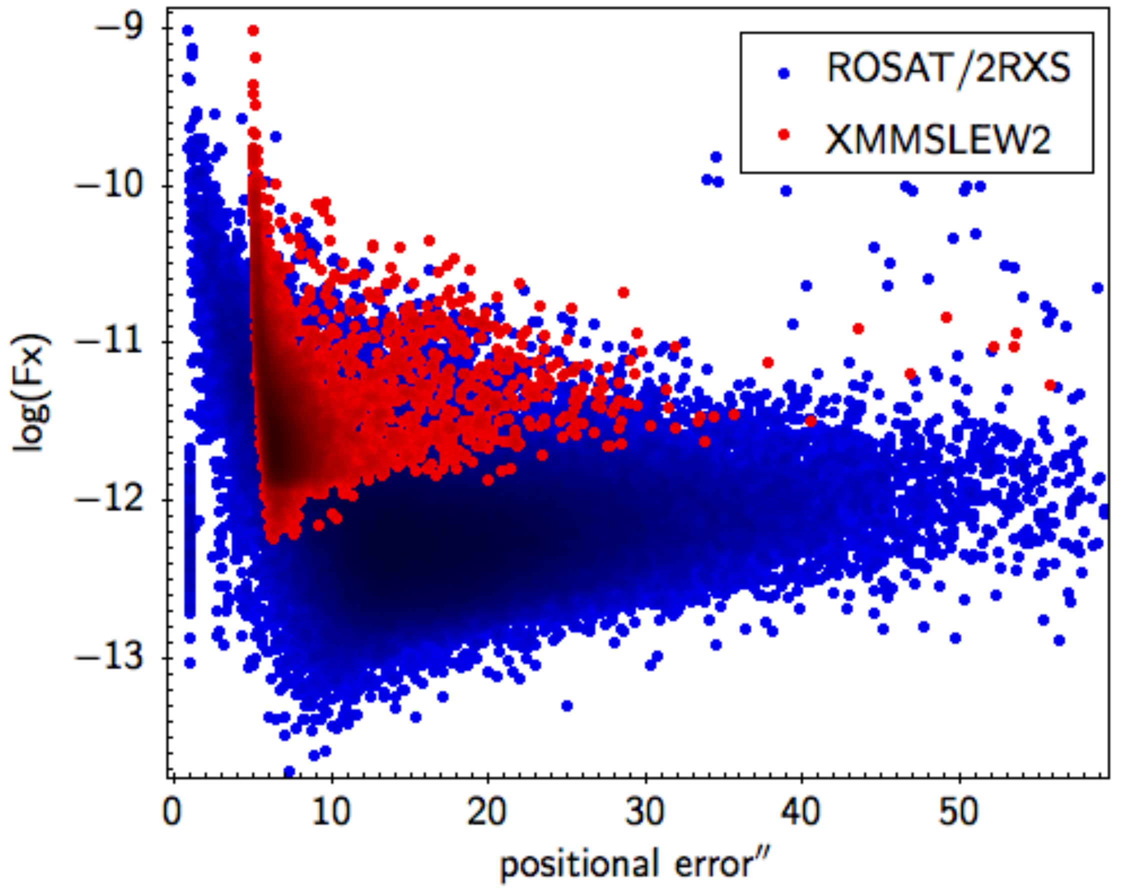
Vulic et al 2022

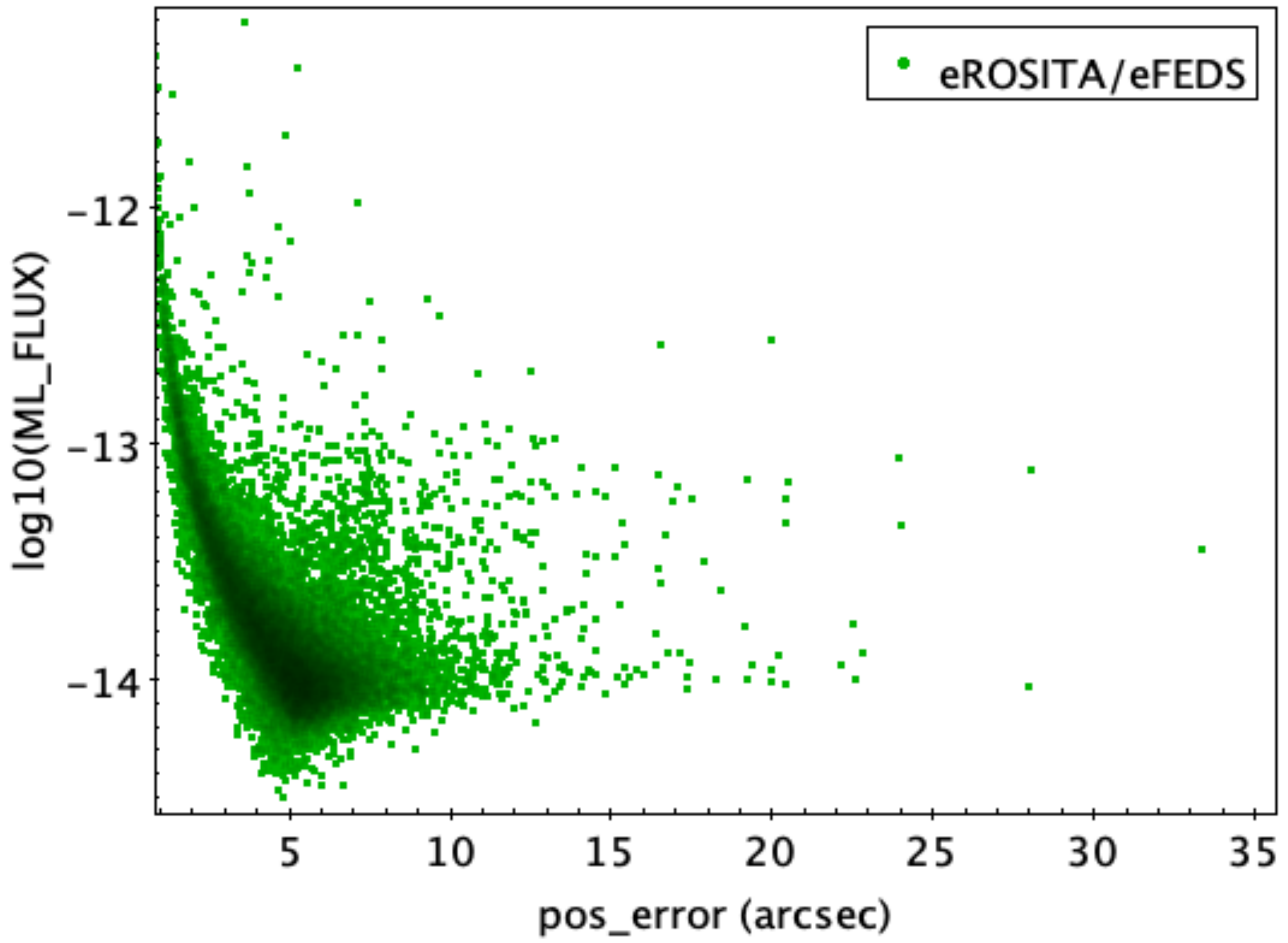
Brusa et al 2022

Wolf et al 2022









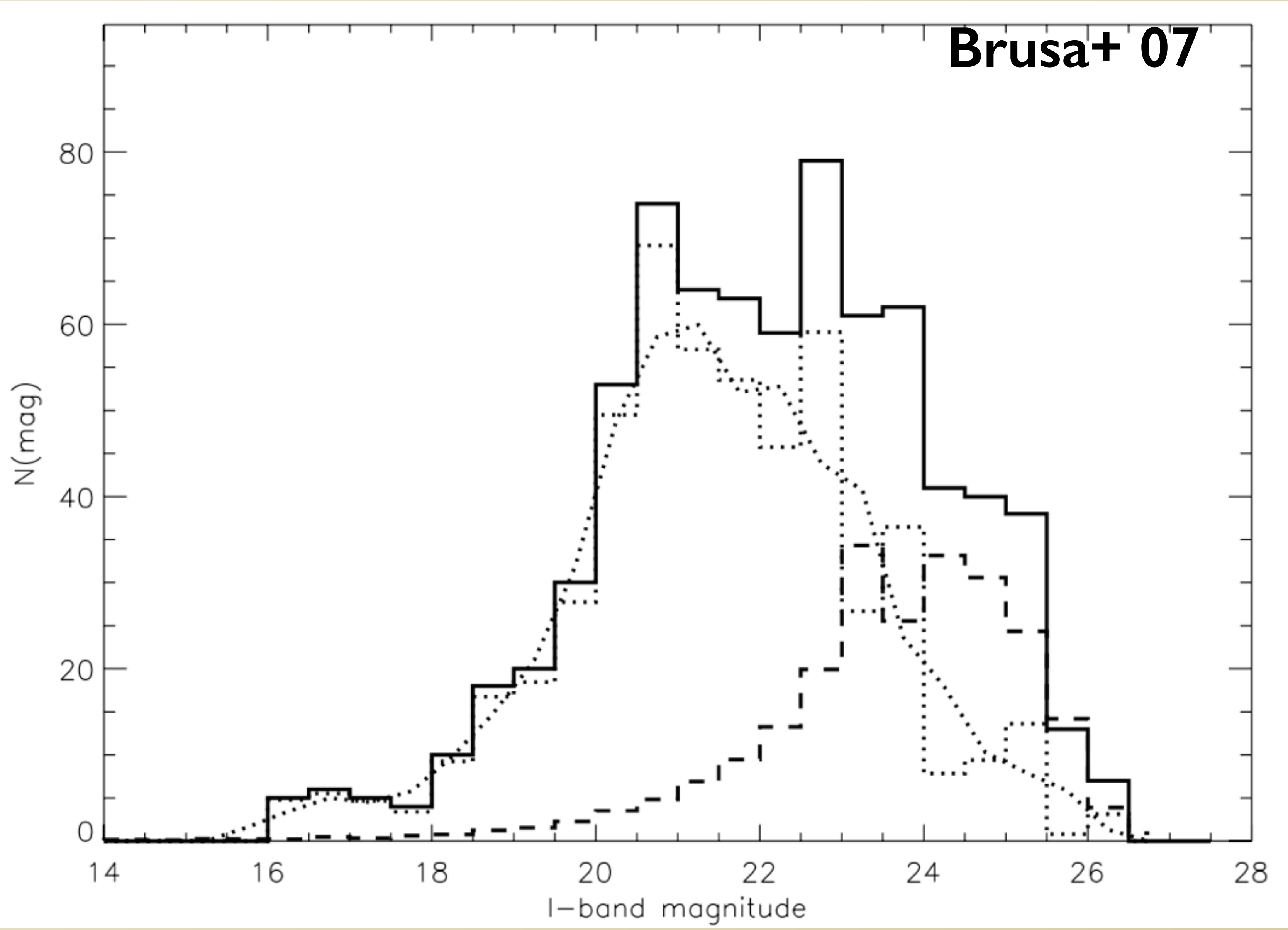
- The size of the area search depends on the positional error which correlates with the intensity of the source.

…but it could be also a Galactic source (e.g., Schneider et al 2022, Stelze et al 2022) or an unresolved cluster (e.g., Bulbul et al 2022)

- The X-ray coordinates and positional errors depend on how the X-ray data are treated (Hsu+2014)

Maximum Likelihood (ML) in a nutshell:

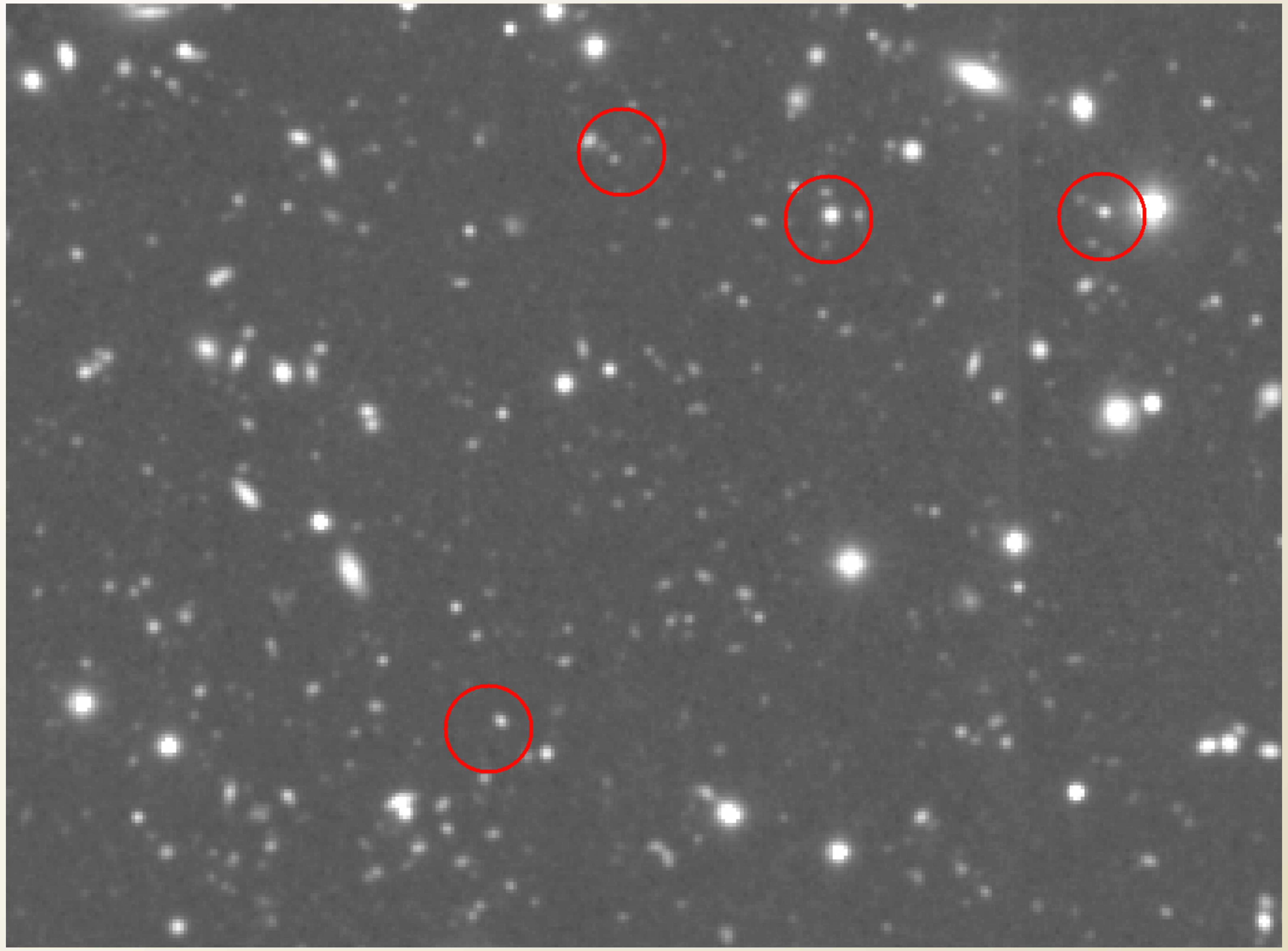
Sutherland&Saunders1992

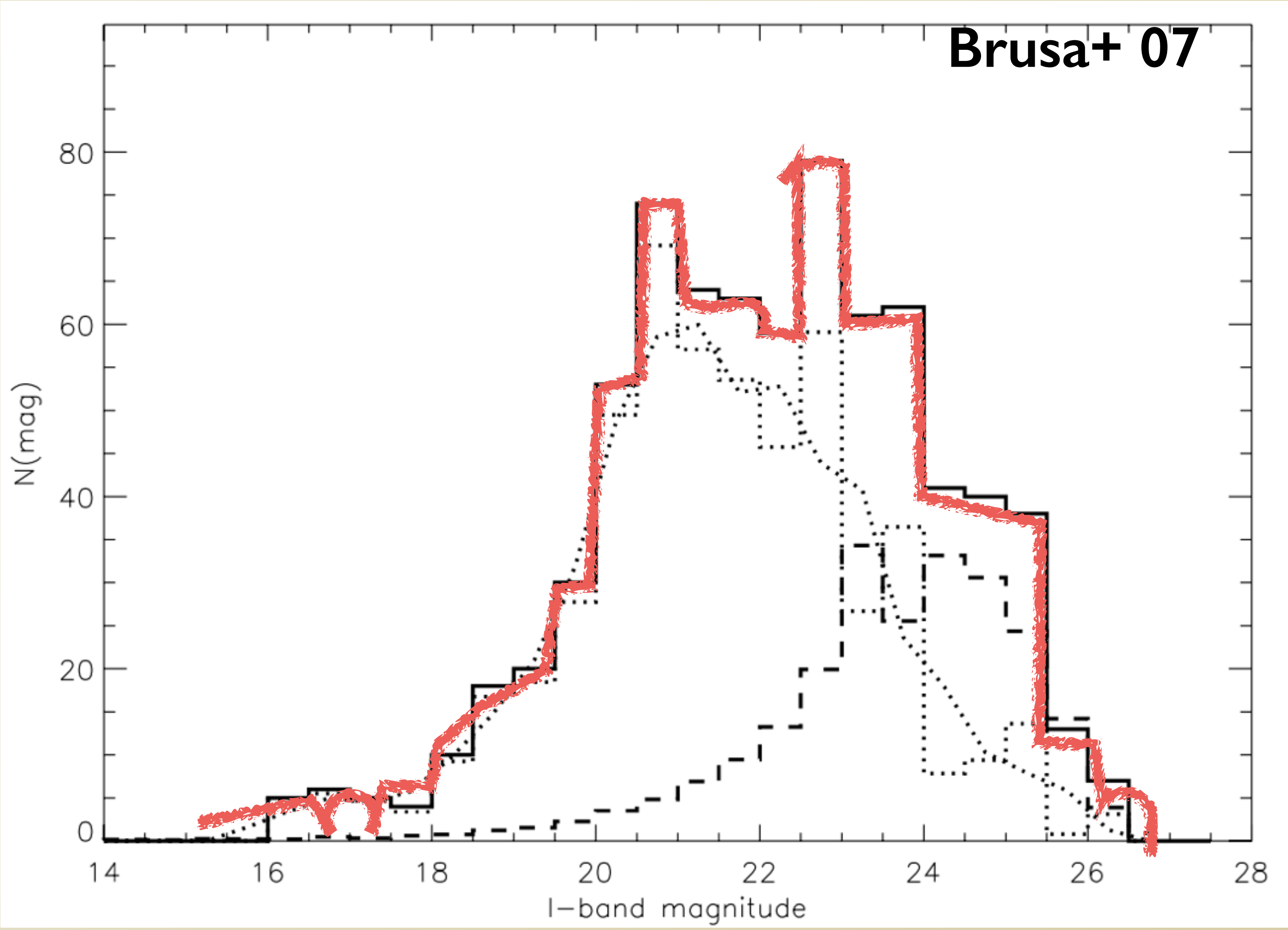


Naylor+13 for a review

Maximum Likelihood (ML) in a nutshell:

Sutherland&Saunders1992

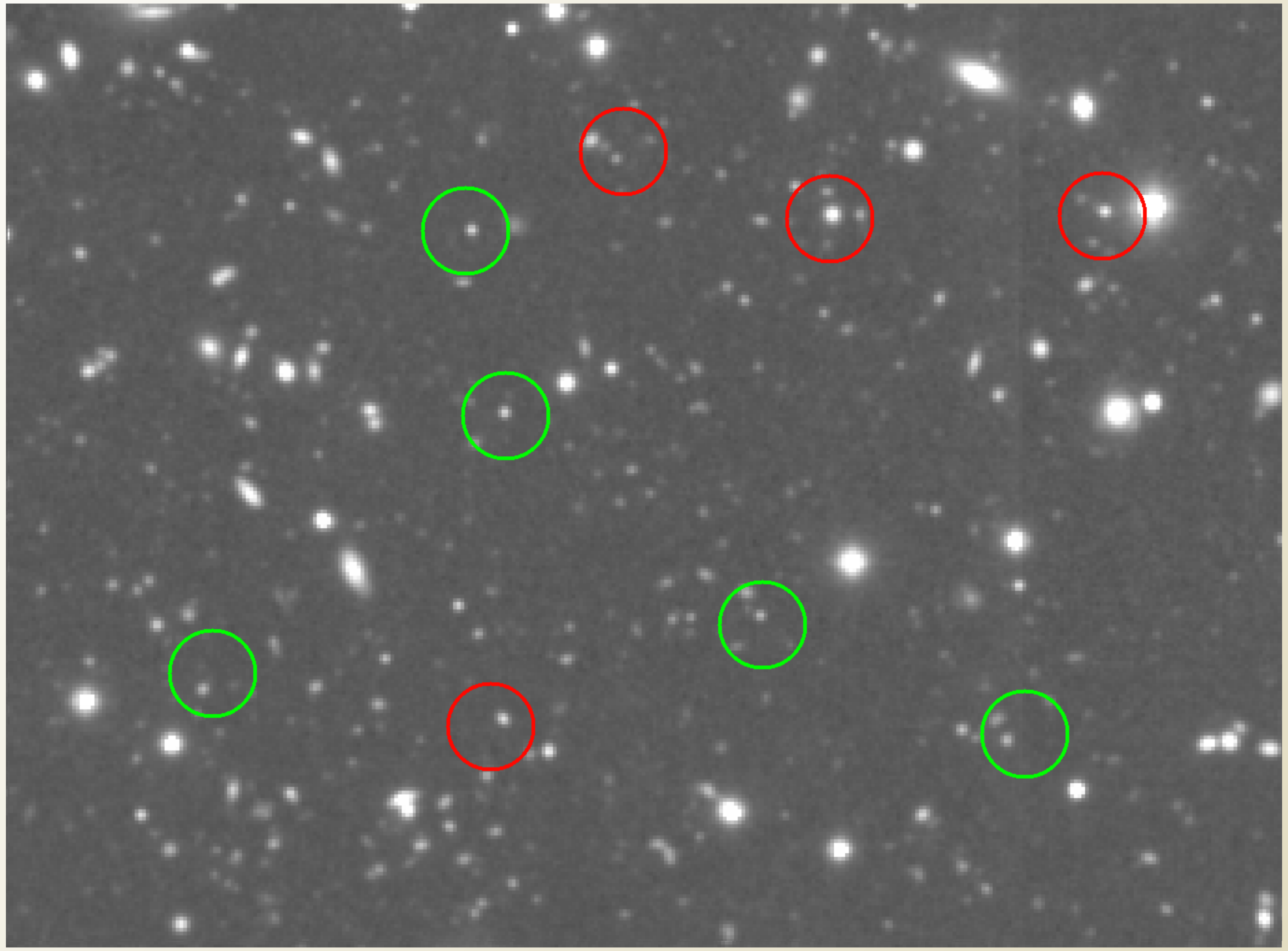


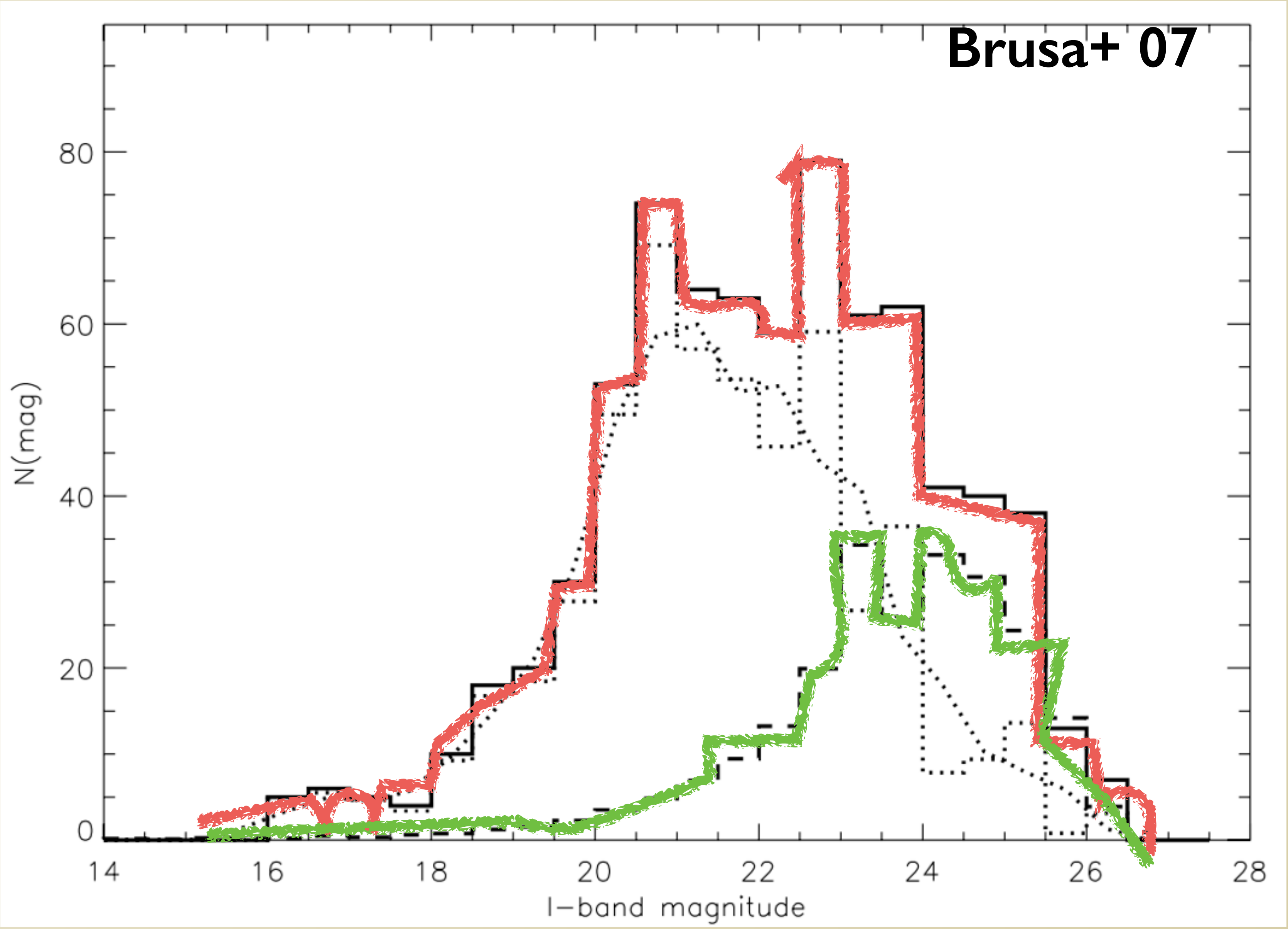


Naylor+13 for a review

Maximum Likelihood (ML) in a nutshell:

Sutherland&Saunders1992

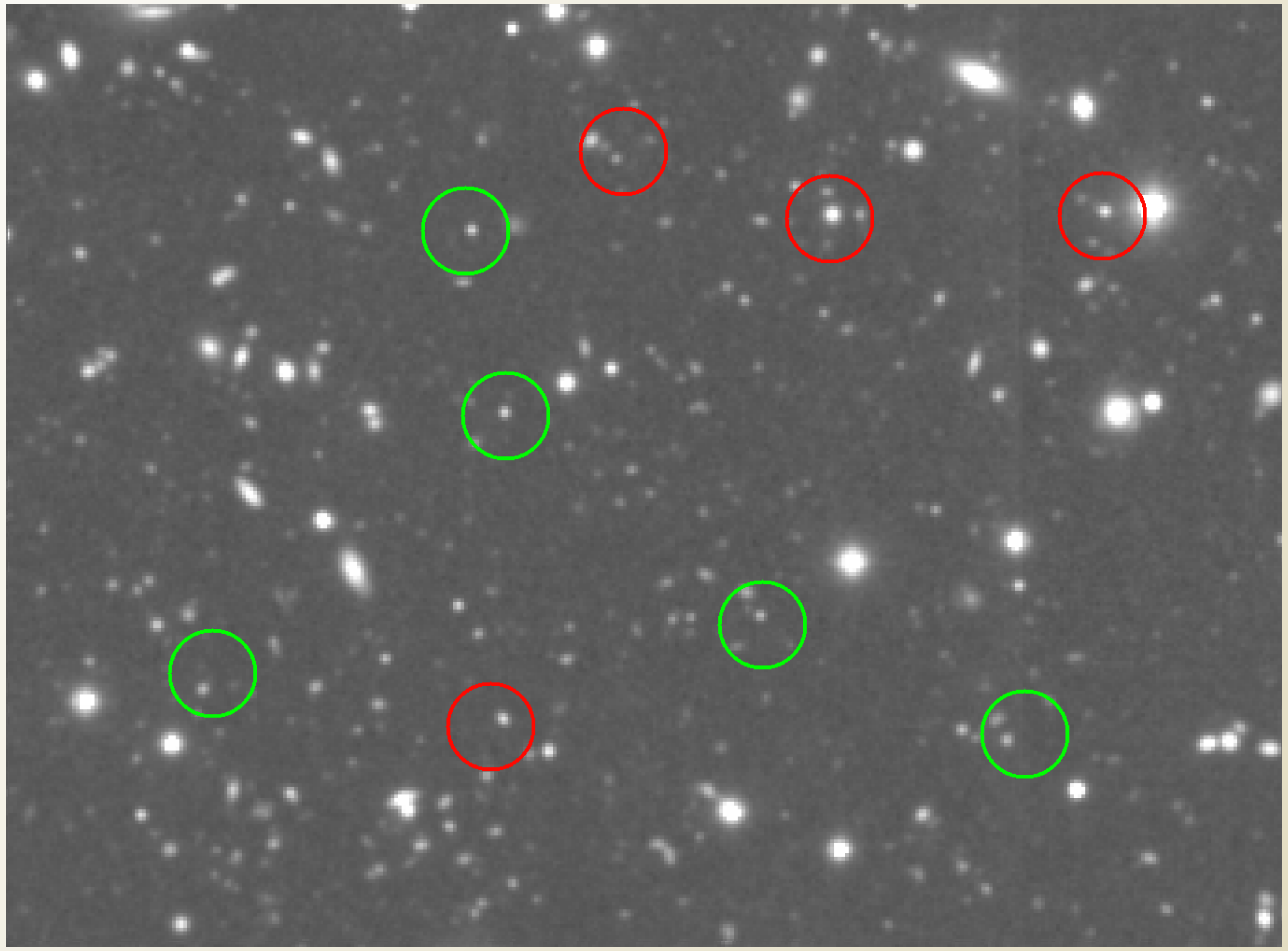


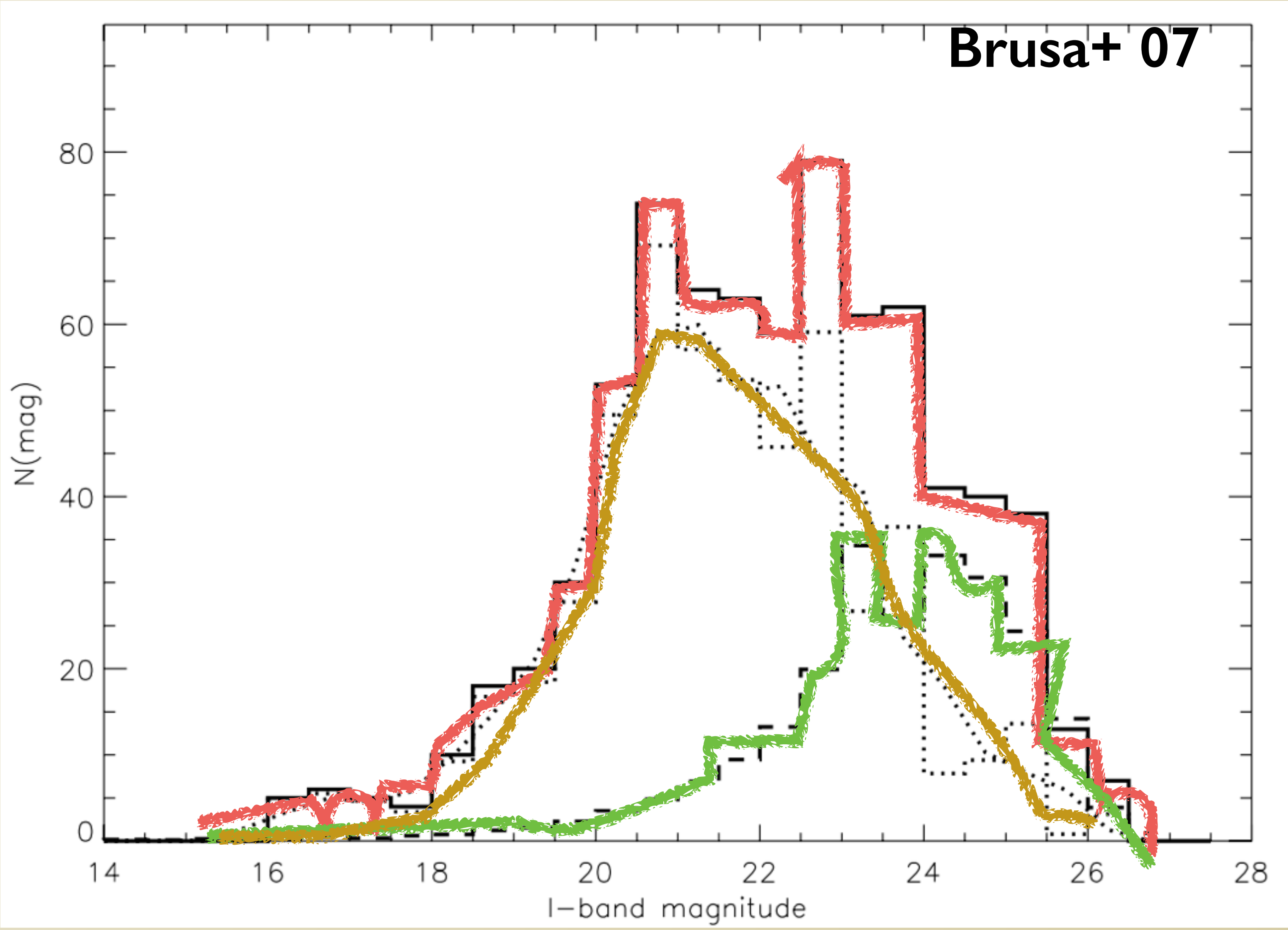


Naylor+13 for a review

Maximum Likelihood (ML) in a nutshell:

Sutherland&Saunders1992

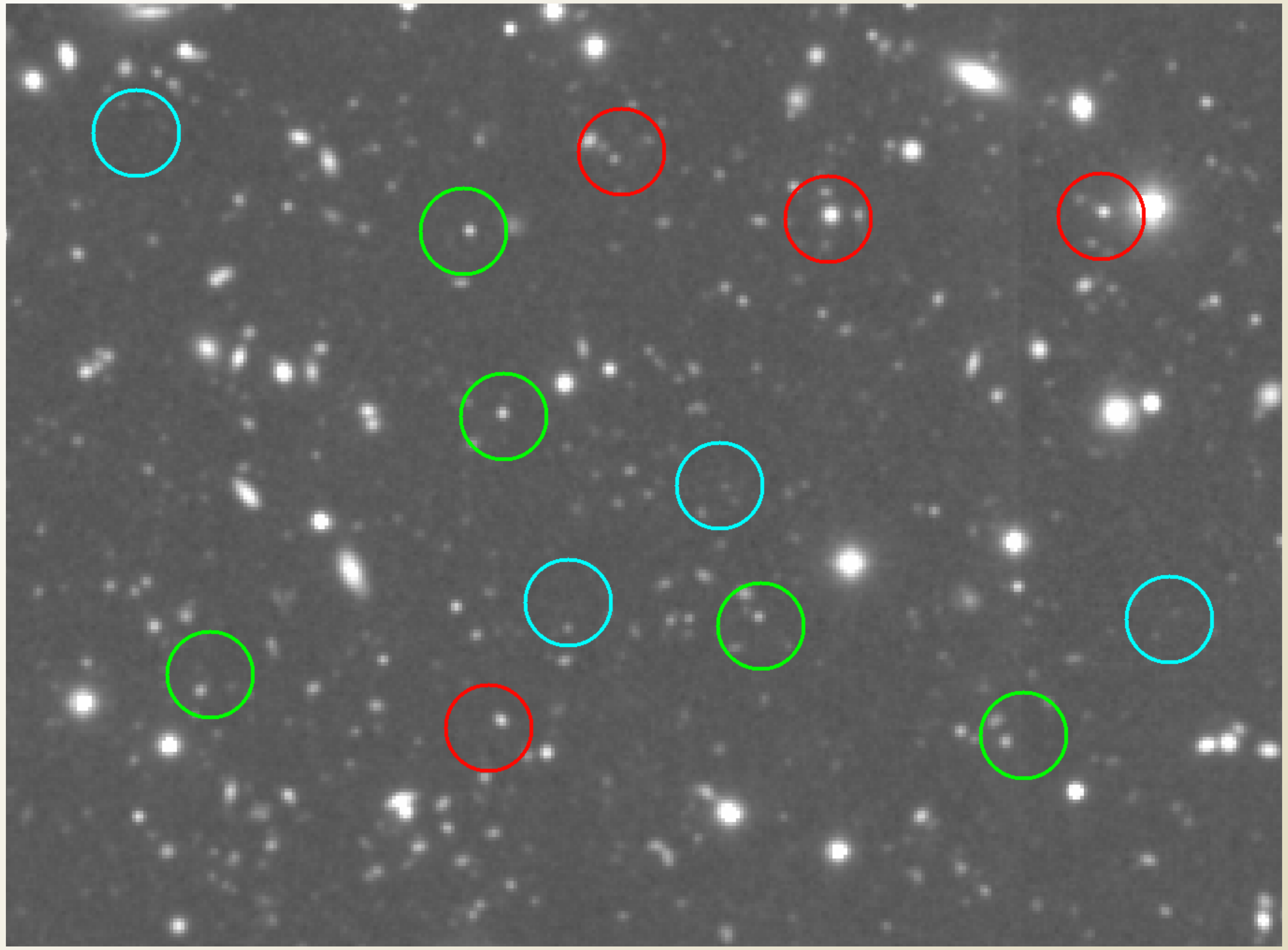


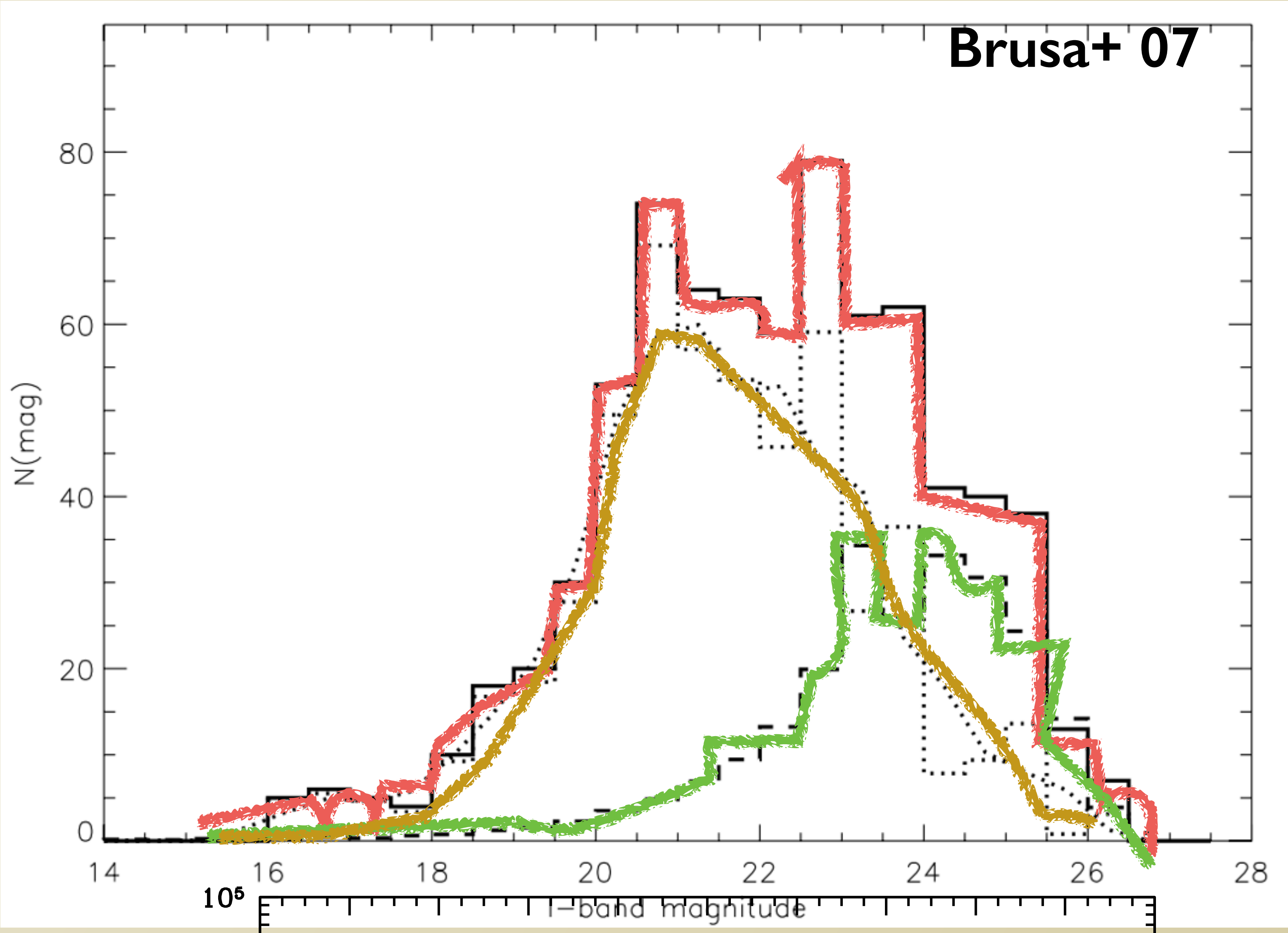


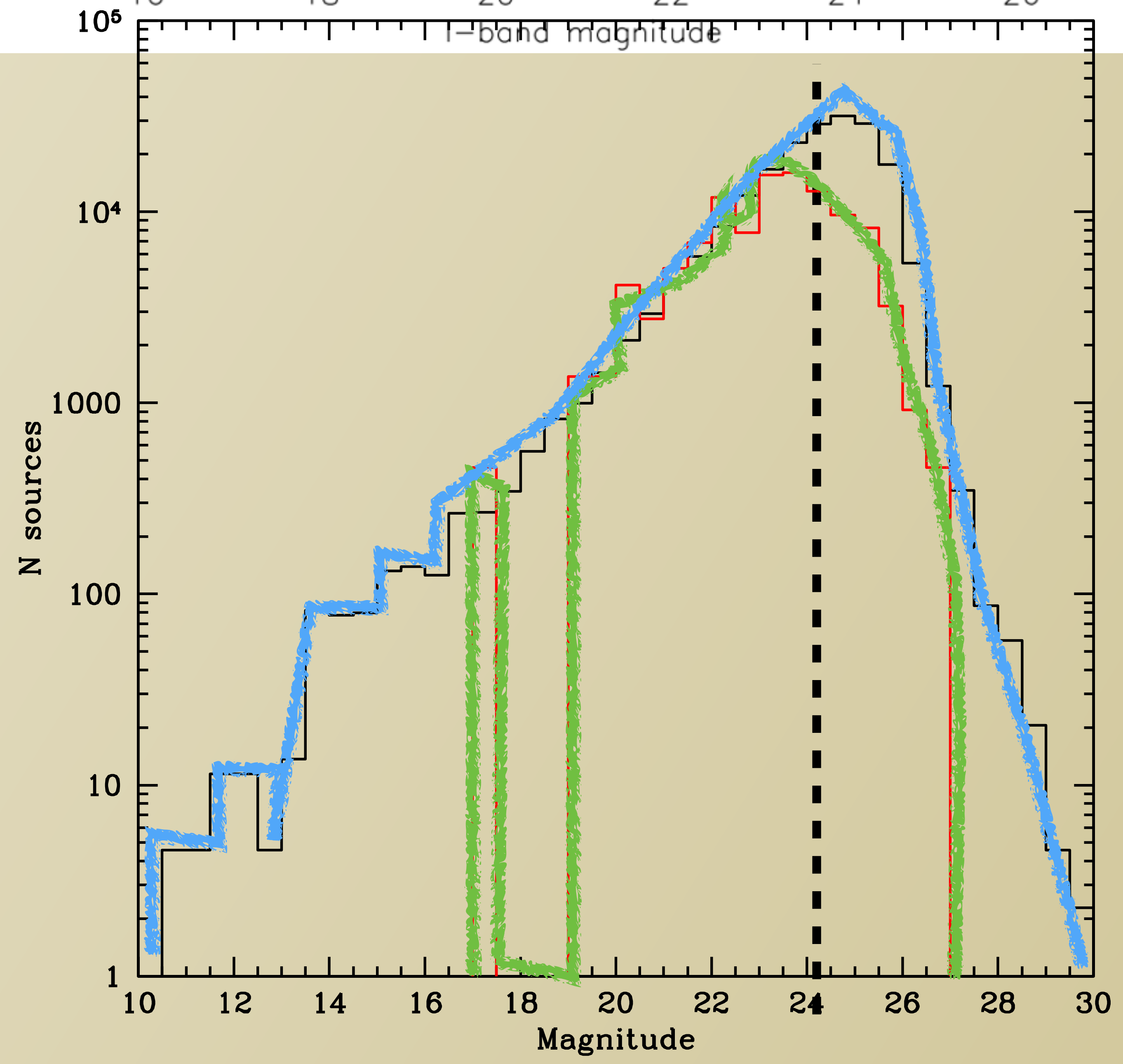
Naylor+13 for a review

Maximum Likelihood (ML) in a nutshell:

Sutherland&Saunders1992



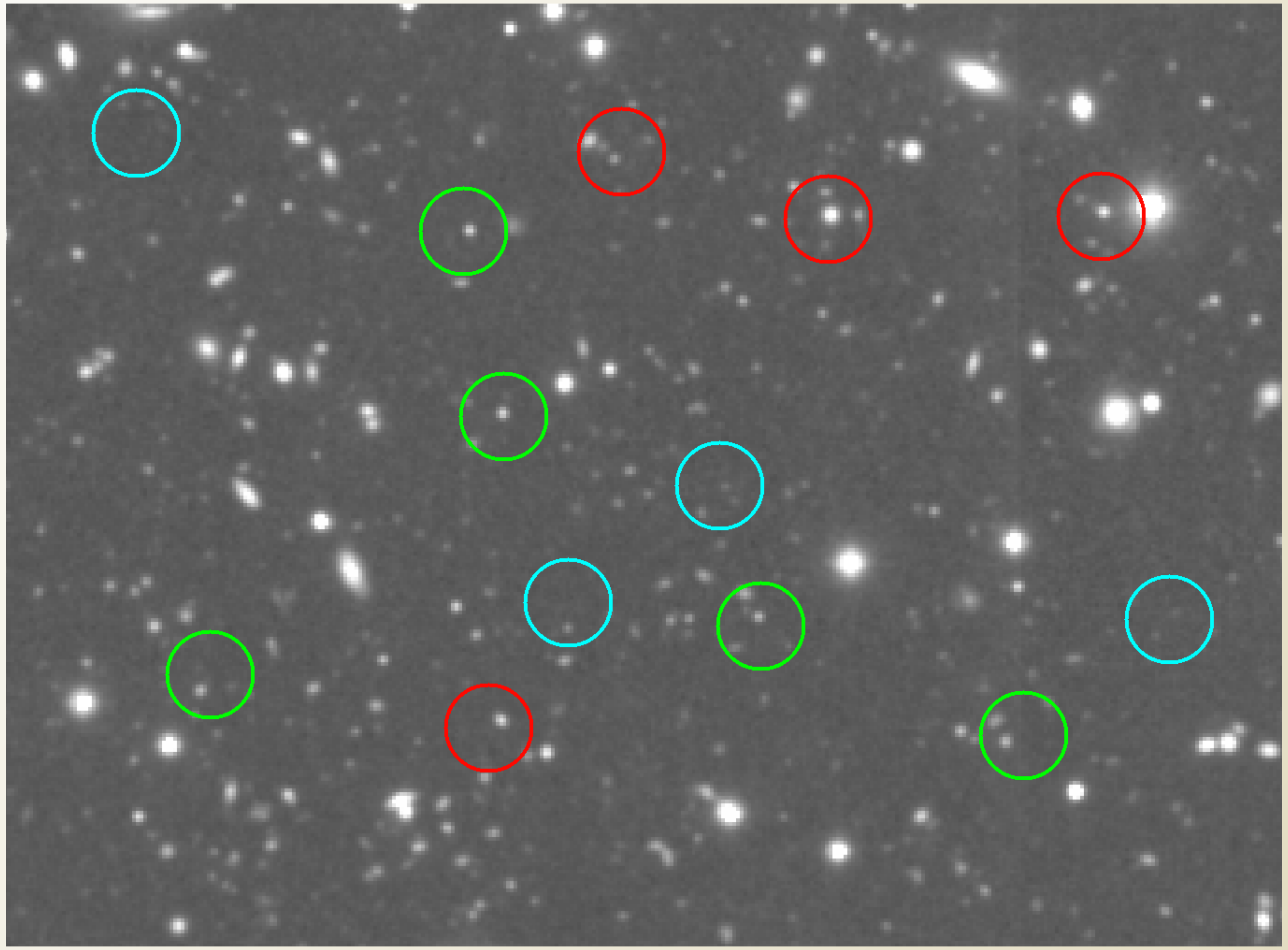


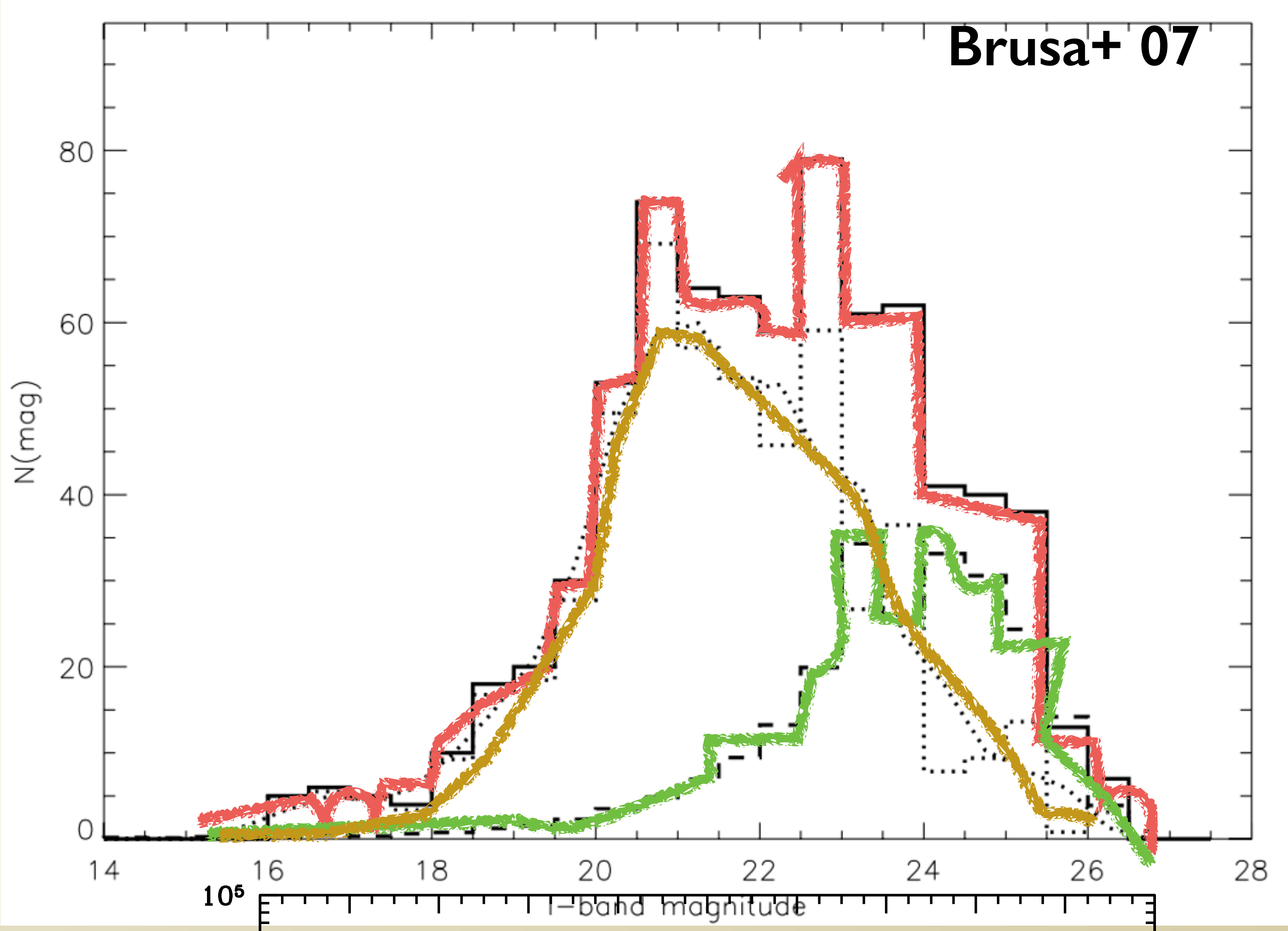


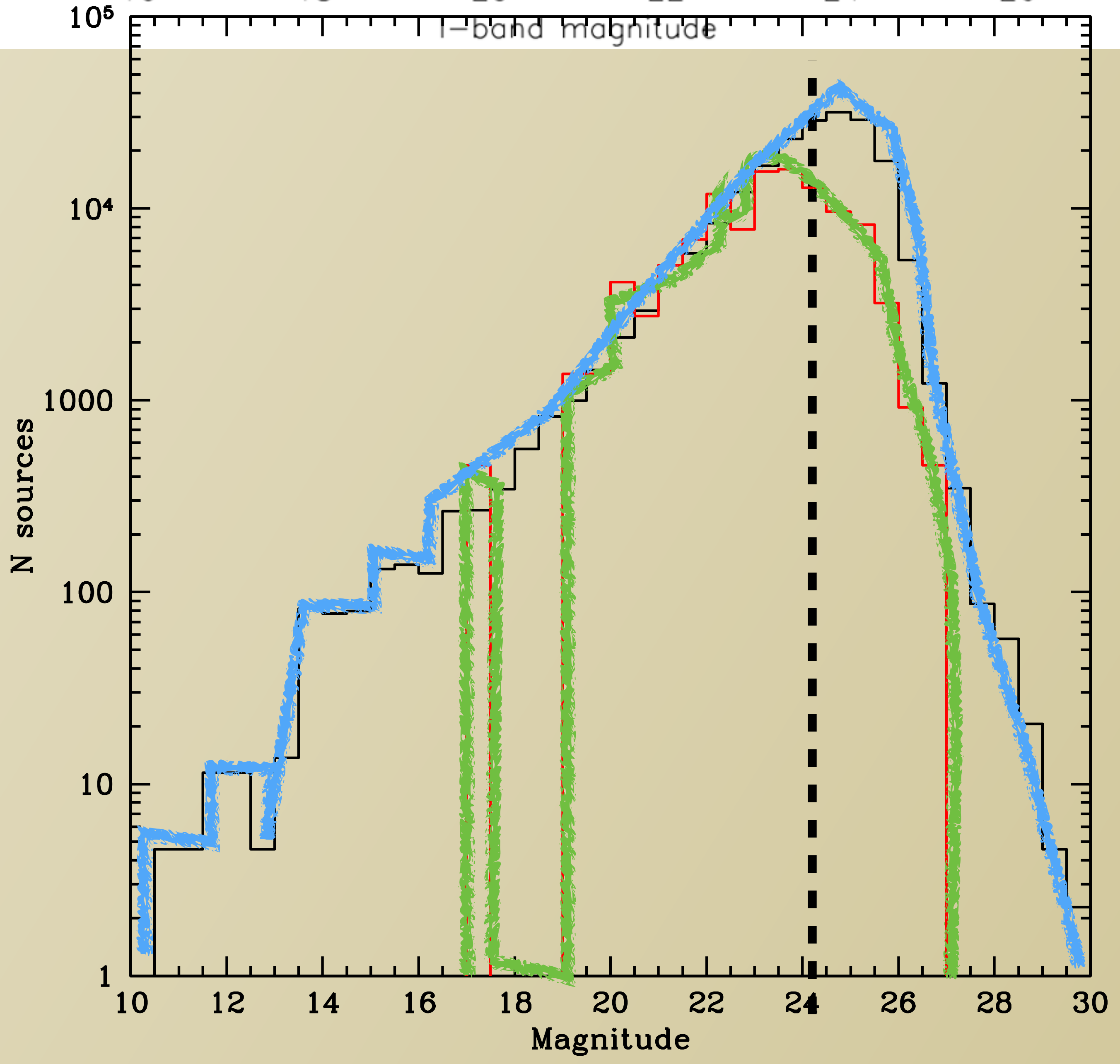
Naylor+13 for a review

Maximum Likelihood (ML) in a nutshell:

Sutherland&Saunders1992



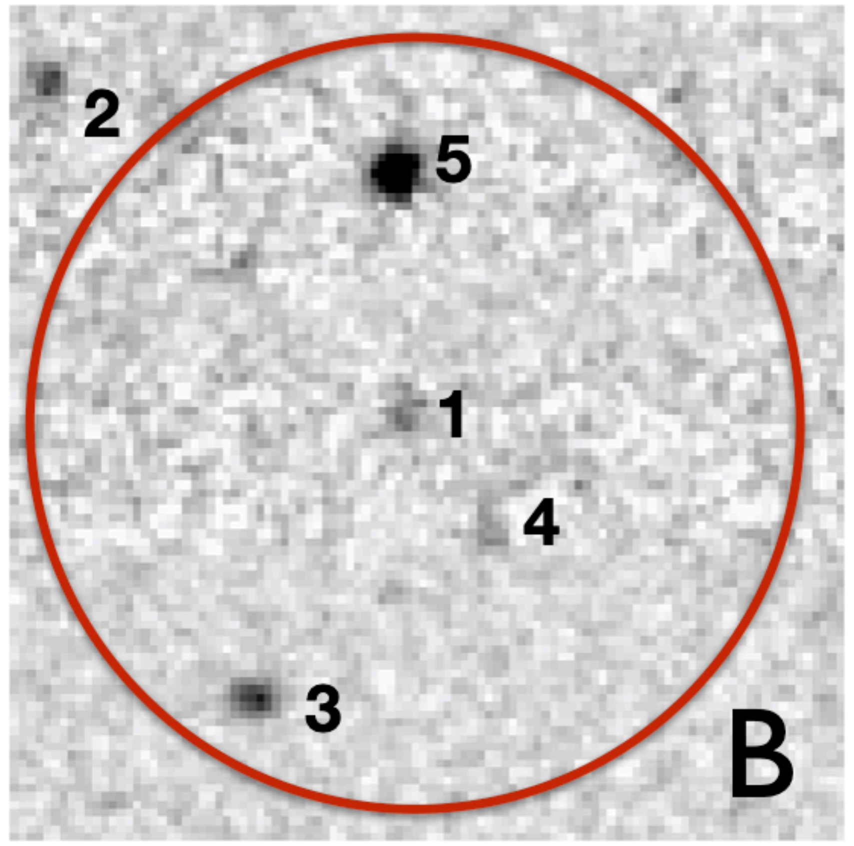




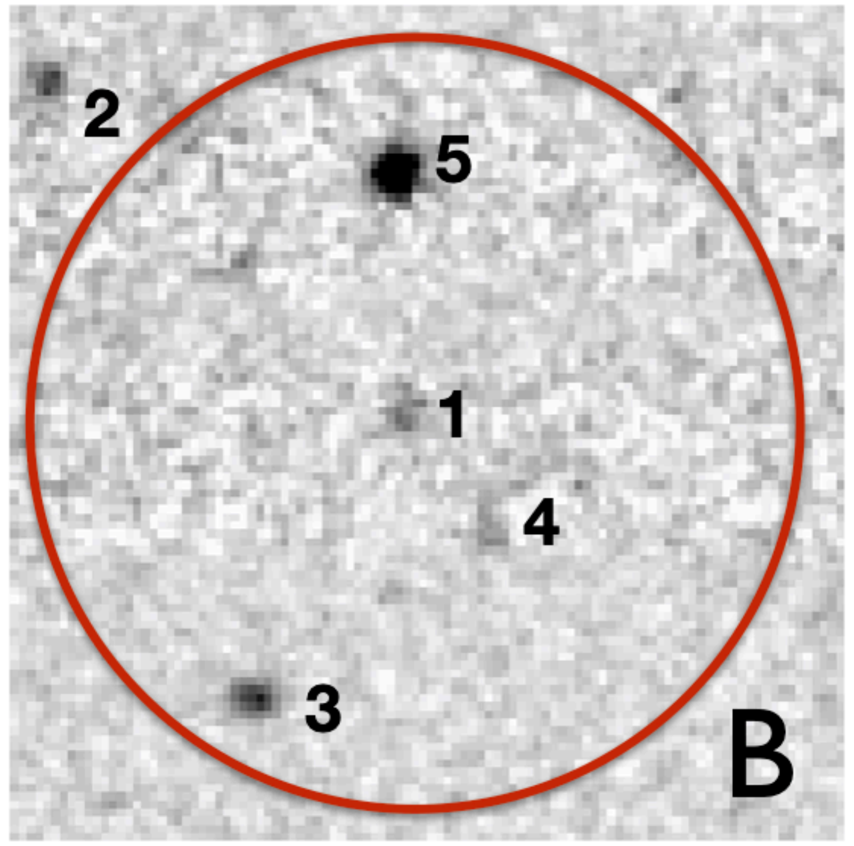
It is data driven:
problem for small data set

Naylor+13 for a review

issue: different bands provide different results

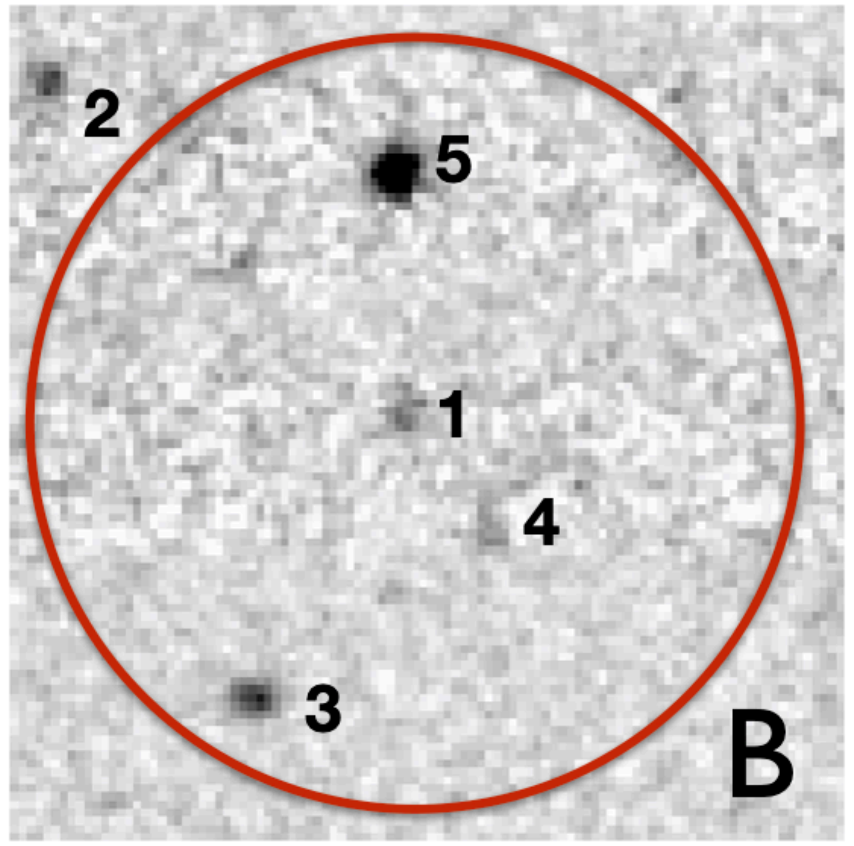


issue: different bands provide different results



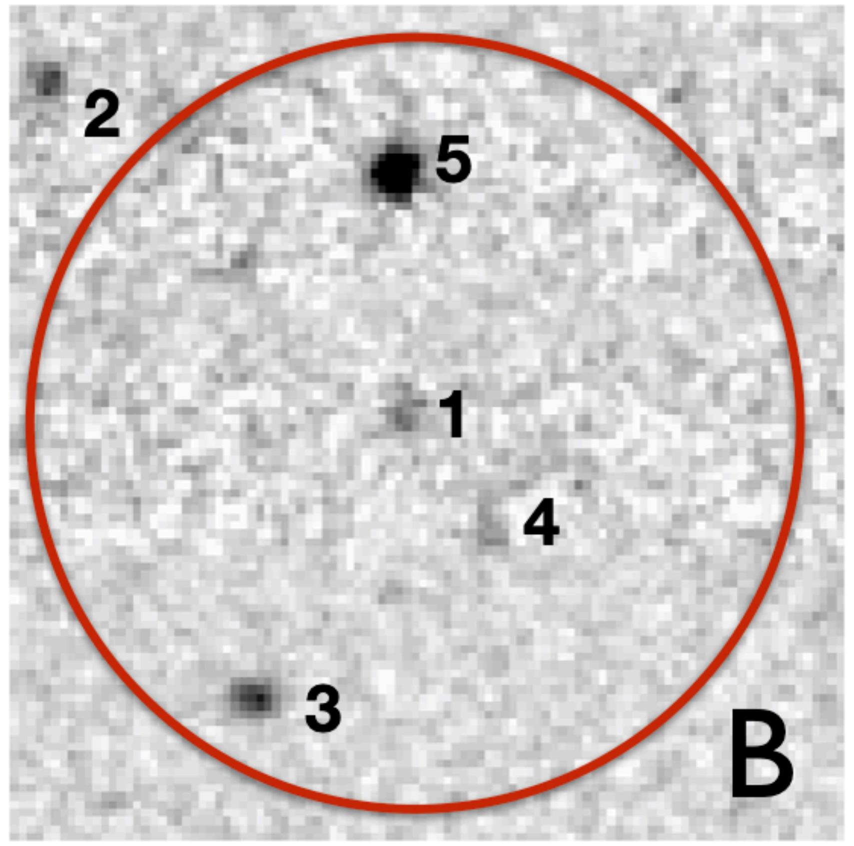
- when assigning a counterpart we consider:

issue: different bands provide different results



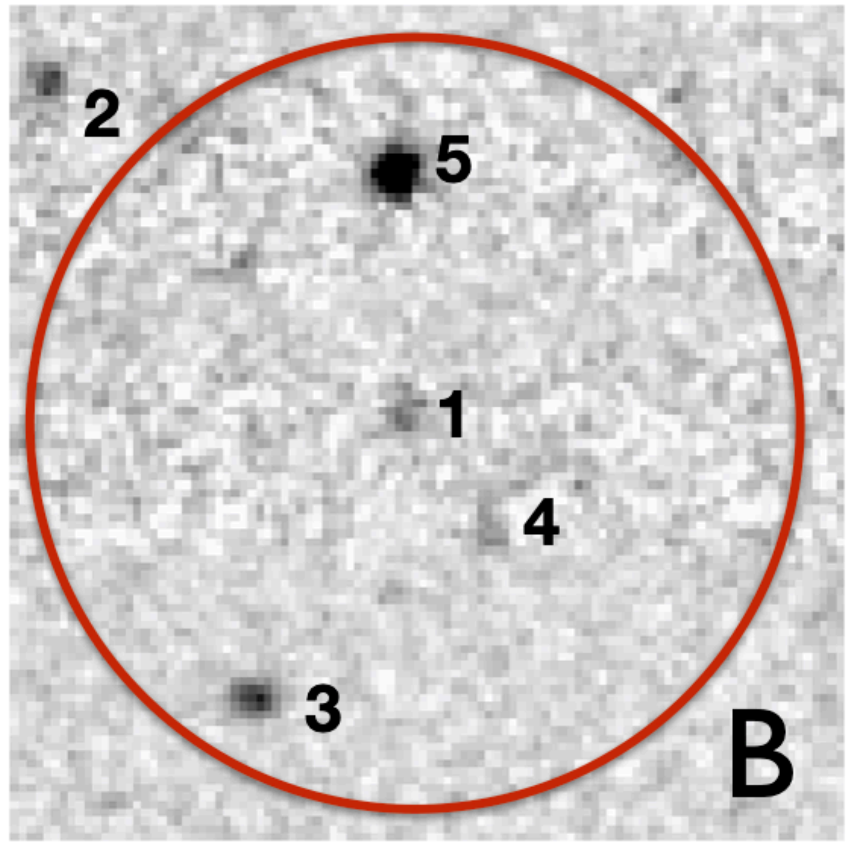
- when assigning a counterpart we consider:
  - the separation between the sources in the primary and secondary catalog

issue: different bands provide different results



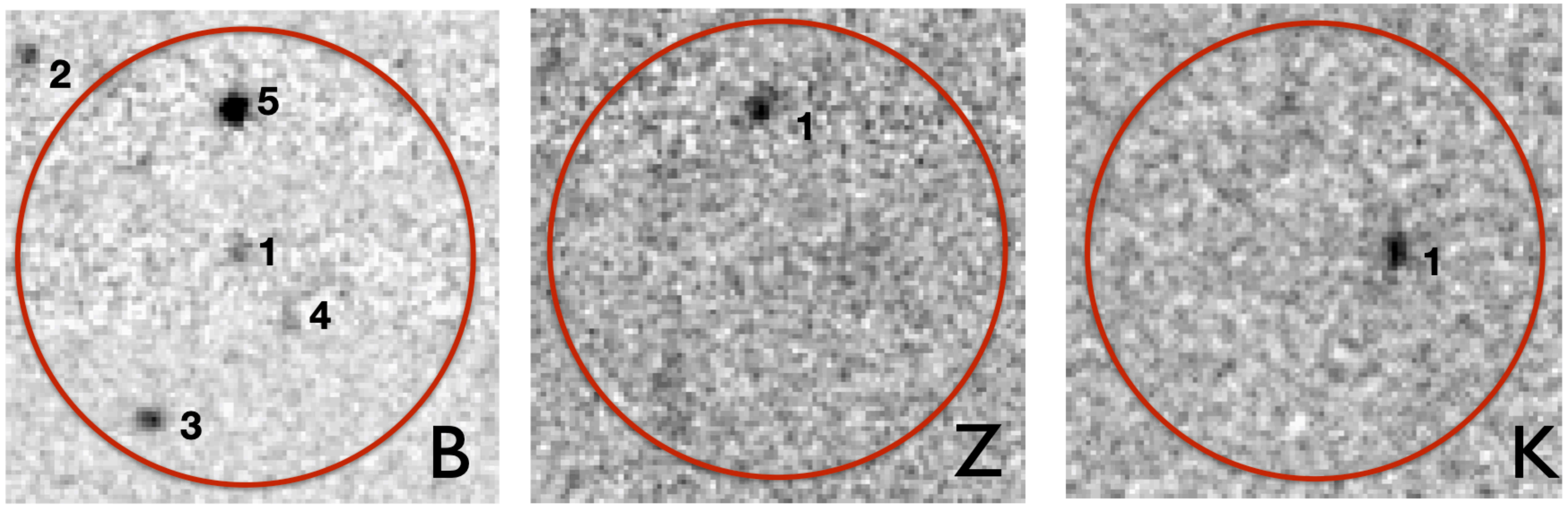
- when assigning a counterpart we consider:
    - the separation between the sources in the primary and secondary catalog
    - positional error for each source in the primary and secondary catalog

issue: different bands provide different results



- when assigning a counterpart we consider:
  - the separation between the sources in the primary and secondary catalog
  - positional error for each source in the primary and secondary catalog
  - source number density in the primary and secondary catalog

issue: different bands provide different results

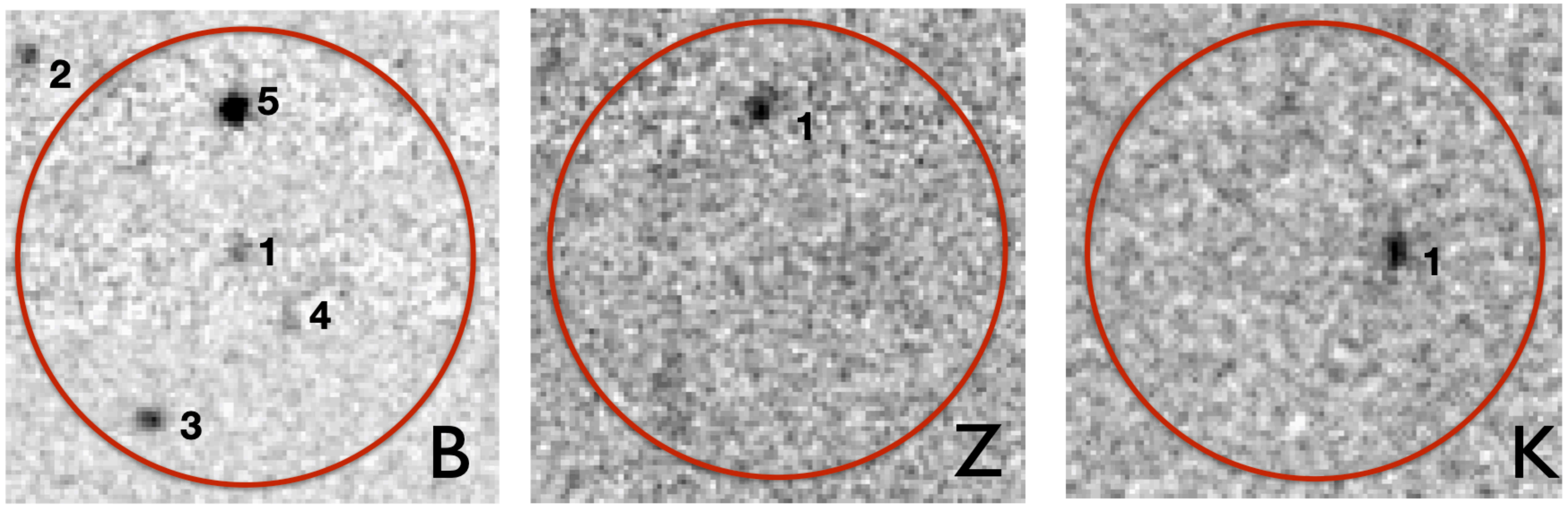


- when assigning a counterpart we consider:
    - the separation between the sources in the primary and secondary catalog
    - positional error for each source in the primary and secondary catalog
    - source number density in the primary and secondary catalog

- using only one band we do not account for the possibility that the actual counterpart is *NOT* detected in *THAT* band

- repeating the process with different bands and for each X-ray source selecting the counterpart from the band with the higher probability compensate only partially

issue: different bands provide different results



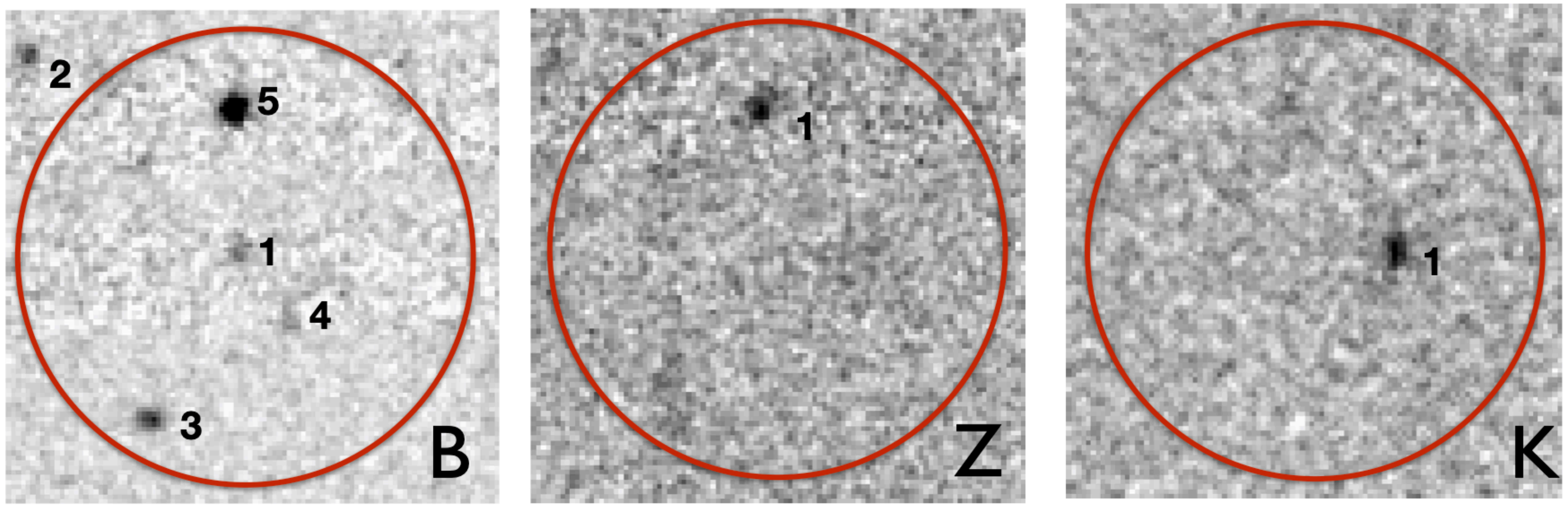
- when assigning a counterpart we consider:
    - the separation between the sources in the primary and secondary catalog
    - positional error for each source in the primary and secondary catalog
    - source number density in the primary and secondary catalog

- using only one band we do not account for the possibility that the actual counterpart is *NOT* detected in *THAT* band

- repeating the process with different bands and for each X-ray source selecting the counterpart from the band with the higher probability compensate only partially

go Bayesian: use ALL the bands at the same time
and combine the probabilities before assigning the CTP (Budavari & Szalay 2008)
But account also for missing data: Nway (Salvato+2018), Xmatch (Pineau2017)

issue: different bands provide different results



- when assigning a counterpart we consider:
  - the separation between the sources in the primary and secondary catalog
  - positional error for each source in the primary and secondary catalog
  - source number density in the primary and secondary catalog

- using only one band we do not account for the possibility that the actual counterpart is *NOT* detected in *THAT* band

- repeating the process with different bands and for each X-ray source selecting the counterpart from the band with the higher probability compensate only partially

X cat. entry	B cat. entry	Z cat. entry	K cat. entry	P (this is the correct ctp)
1	5	1	—	P1
1	—	—	1	P2
1	1	—	—	P3
1	3	—	—	...
1	...	...	...	...
2	...	...	...	...

go Bayesian: use ALL the bands at the same time
and combine the probabilities before assigning the CTP (Budavari & Szalay 2008)
But account also for missing data: Nway (Salvato+2018), Xmatch (Pineau2017)

benefit: different bands provide different information

prior
(e.g due to depth of data)

$$P(H|D) \propto P(H) \times P(D|H)$$

(posterior) prob.
of an association,
given the data

benefit: different bands provide different information

prior
(e.g due to depth of data)

$$P(H|D) \propto P(H) \times P(D|H).$$

(posterior) prob.
of an association,
given the data

likelihood

benefit: different bands provide different information

prior
(e.g due to depth of data)

$$P(H|D) \propto P(H) \times P(D|H)$$

(posterior) prob. of an association, given the data

likelihood

$$\begin{aligned} P(D|H) &= P(D_\phi|H) \times P(D_m|H) \\ &= P(D_\phi|H) \times \frac{\bar{q}(m)}{\bar{n}(m)}, \end{aligned}$$

benefit: different bands provide different information

prior
(e.g due to depth of data)

$$P(H|D) \propto P(H) \times P(D|H).$$

(posterior) prob.
of an association,
given the data

likelihood

$$\begin{aligned} P(D|H) &= P(D_\phi|H) \times P(D_m|H) \\ &= P(D_\phi|H) \times \frac{\bar{q}(m)}{\bar{n}(m)}, \end{aligned}$$

probability based purely
on spatial information

benefit: different bands provide different information

prior
(e.g due to depth of data)

$$P(H|D) \propto P(H) \times P(D|H).$$

(posterior) prob. of an association, given the data

likelihood

$$\begin{aligned} P(D|H) &= P(D_\phi|H) \times P(D_m|H) \\ &= P(D_\phi|H) \times \frac{\bar{q}(m)}{\bar{n}(m)}, \end{aligned}$$

probability based purely on spatial information

probability that a correct ctp to a X-ray source or a generic field source has a property m.

benefit: different bands provide different information

prior
(e.g due to depth of data)

$$P(H|D) \propto P(H) \times P(D|H).$$

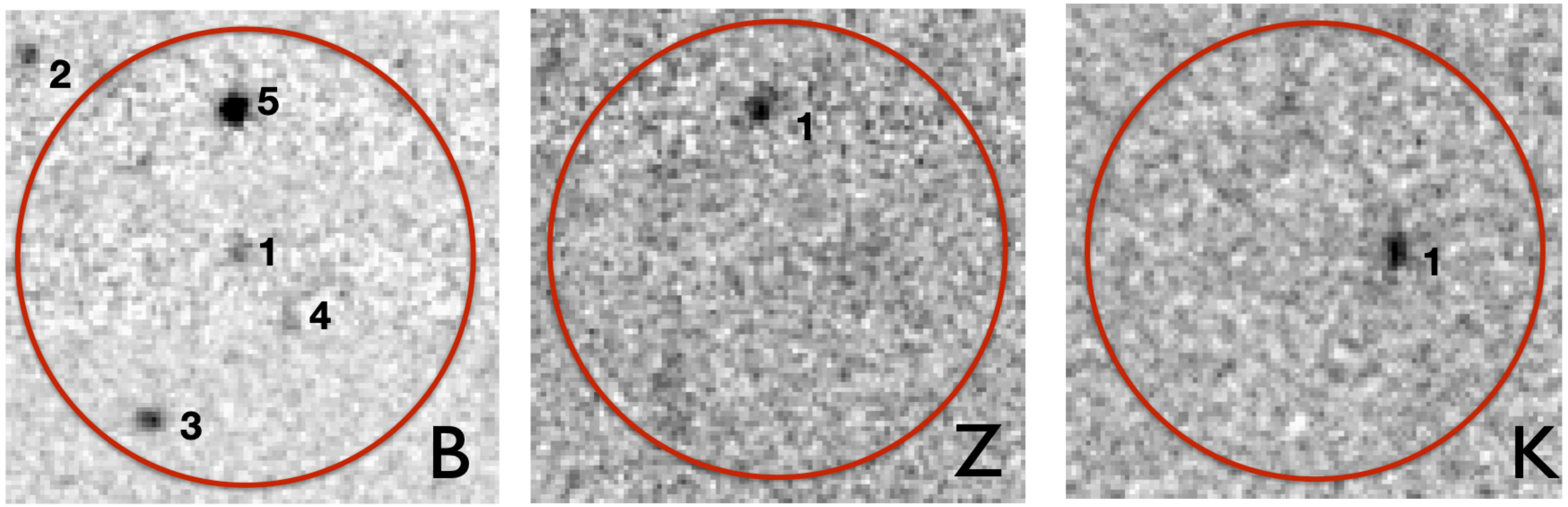
(posterior) prob. of an association, given the data

likelihood

$$\begin{aligned} P(D|H) &= P(D_\phi|H) \times P(D_m|H) \\ &= P(D_\phi|H) \prod \frac{\int_m \bar{q}(m)\, p(m|D_m)\, dm}{\int_m \bar{n}(m)\, p(m|D_m)\, dm} \end{aligned}$$

probability based purely on spatial information

probability that a correct ctp to a X-ray source or a generic field source has a property m.



benefit: different bands provide different information

prior
(e.g due to depth of data)

$$P(H|D) \propto P(H) \times P(D|H).$$

(posterior) prob. of an association, given the data

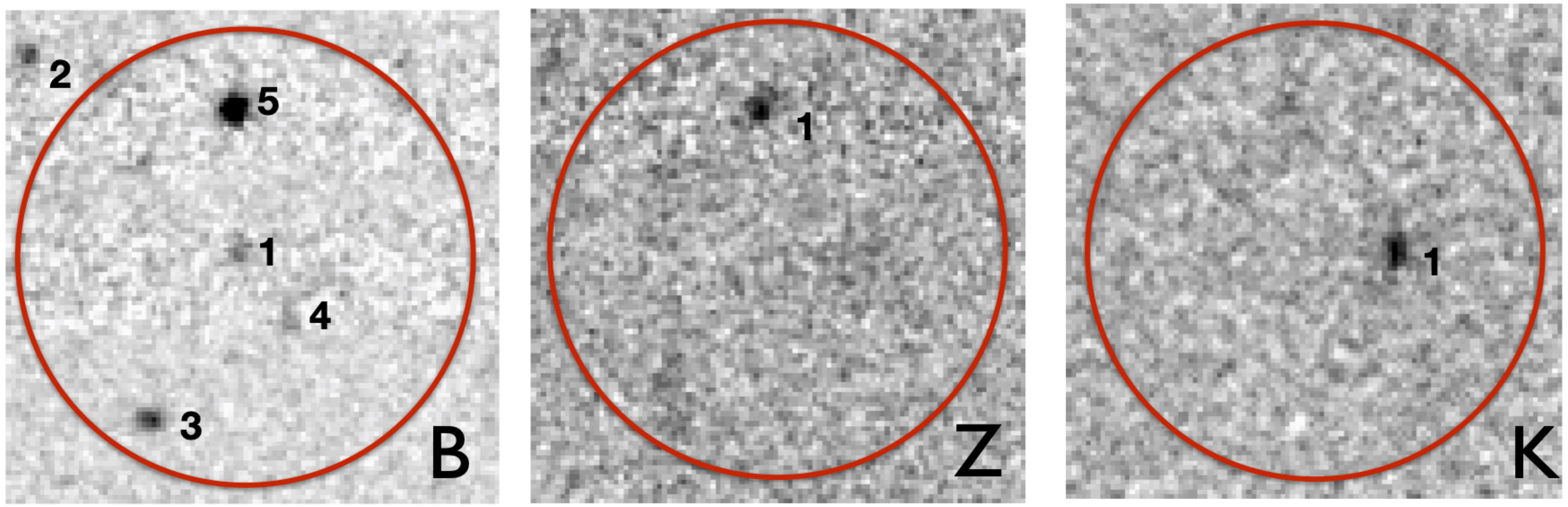
likelihood

$$\begin{aligned} P(D|H) &= P(D_\phi|H) \times P(D_m|H) \\ &= P(D_\phi|H) \prod \frac{\int_m \bar{q}(m)\, p(m|D_m)\, dm}{\int_m \bar{n}(m)\, p(m|D_m)\, dm} \end{aligned}$$

probability based purely on spatial information

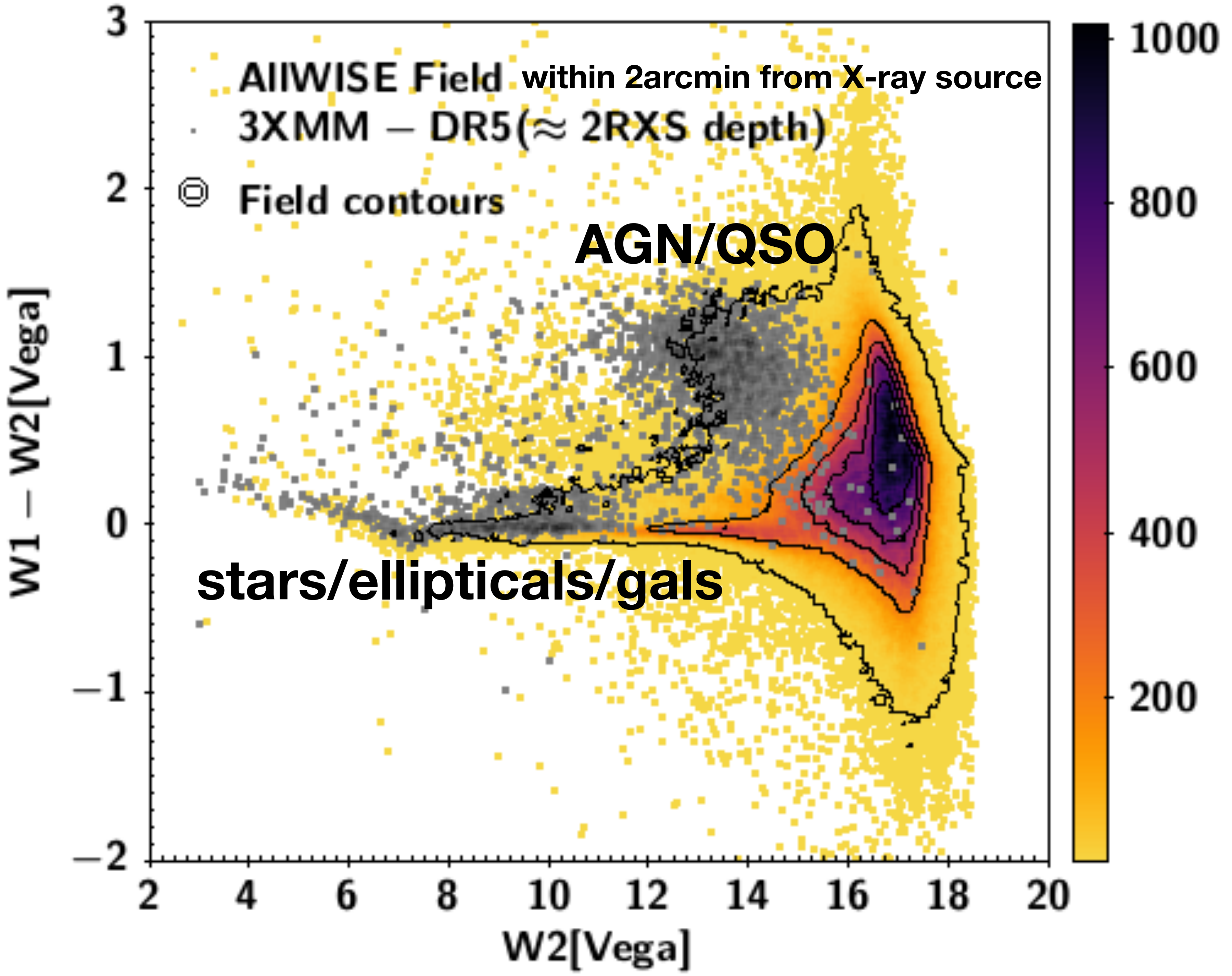
probability that a correct ctp to a X-ray source or a generic field source has a property m.

We KNOW the properties
(e.g., SEDs, variability, morphology) of X-ray emitters
thanks to 20 years of XMM and Chandra.
Let's use that!



goal: identify features that define X-ray emitters wrt field population

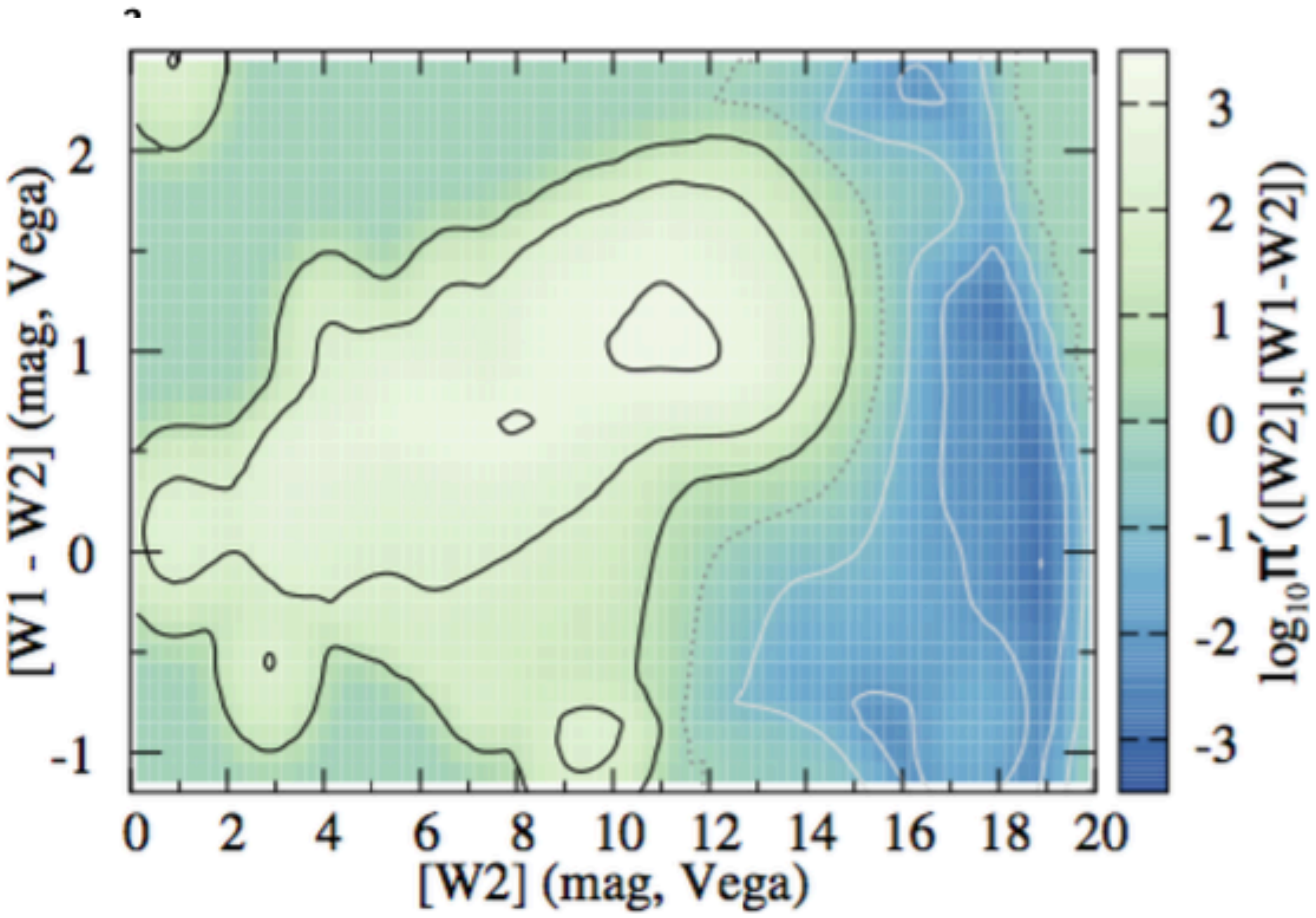
NWAY on ROSAT & XMMSLEW2 (Salvato+ 2018)



Validation: 1500 sources from 3XMM w/ secure ctp, shows 97% accuracy

goal: identify features that define X-ray emitters wrt field population

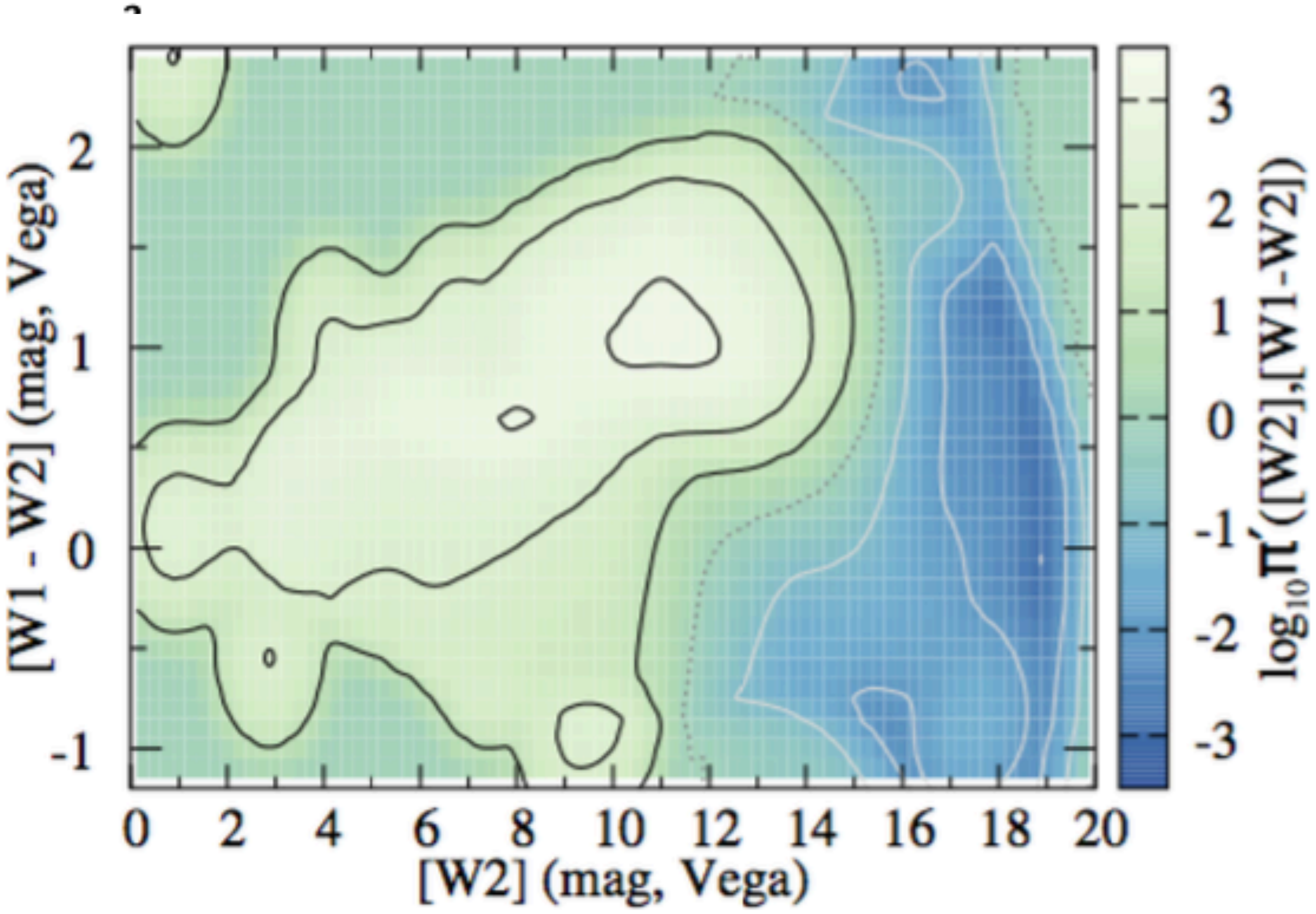
NWAY on ROSAT & XMMSLEW2 (Salvato+ 2018)



Validation: 1500 sources from 3XMM
w/ secure ctp, shows 97% accuracy

goal: identify features that define X-ray emitters wrt field population

NWAY on ROSAT & XMMSLEW2 (Salvato+ 2018)

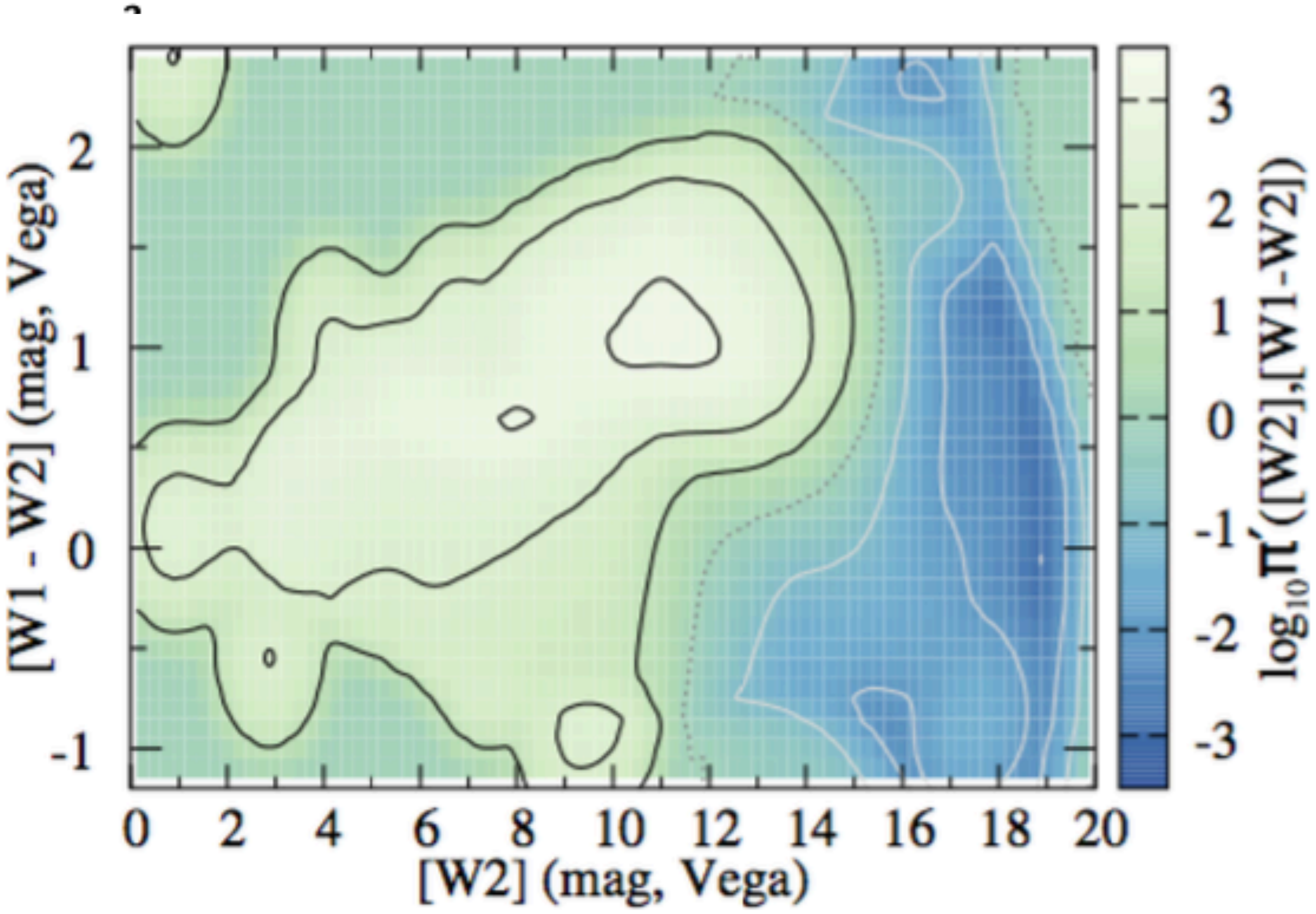


Validation: 1500 sources from 3XMM
w/ secure ctp, shows 97% accuracy

XMMSL-2RXS Separation arcsec	Sources in common N	Identical AllWISE ctp. %
Sep. ≤5	1111	98.5
Sep. ≤10	3448	98.7
Sep. *le*30	7834	96.1
Sep. *le*60	8768	93.0

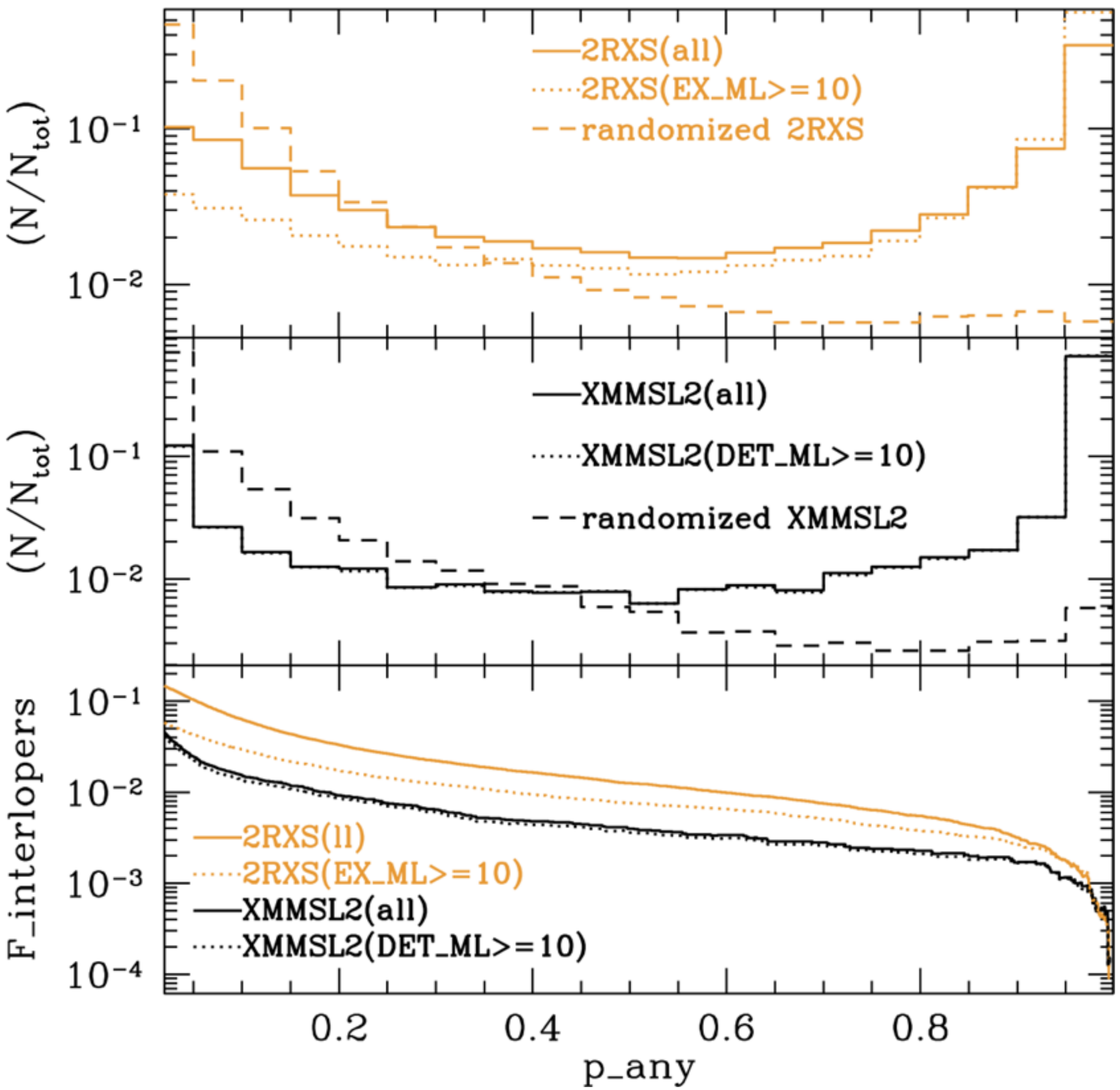
goal: identify features that define X-ray emitters wrt field population

NWAY on ROSAT & XMMSLEW2 (Salvato+ 2018)



Validation: 1500 sources from 3XMM
w/ secure ctp, shows 97% accuracy

XMMSL-2RXS Separation arcsec	Sources in common N	Identical AllWISE ctp. %
Sep. ≤5	1111	98.5
Sep. ≤10	3448	98.7
Sep. *le*30	7834	96.1
Sep. *le*60	8768	93.0

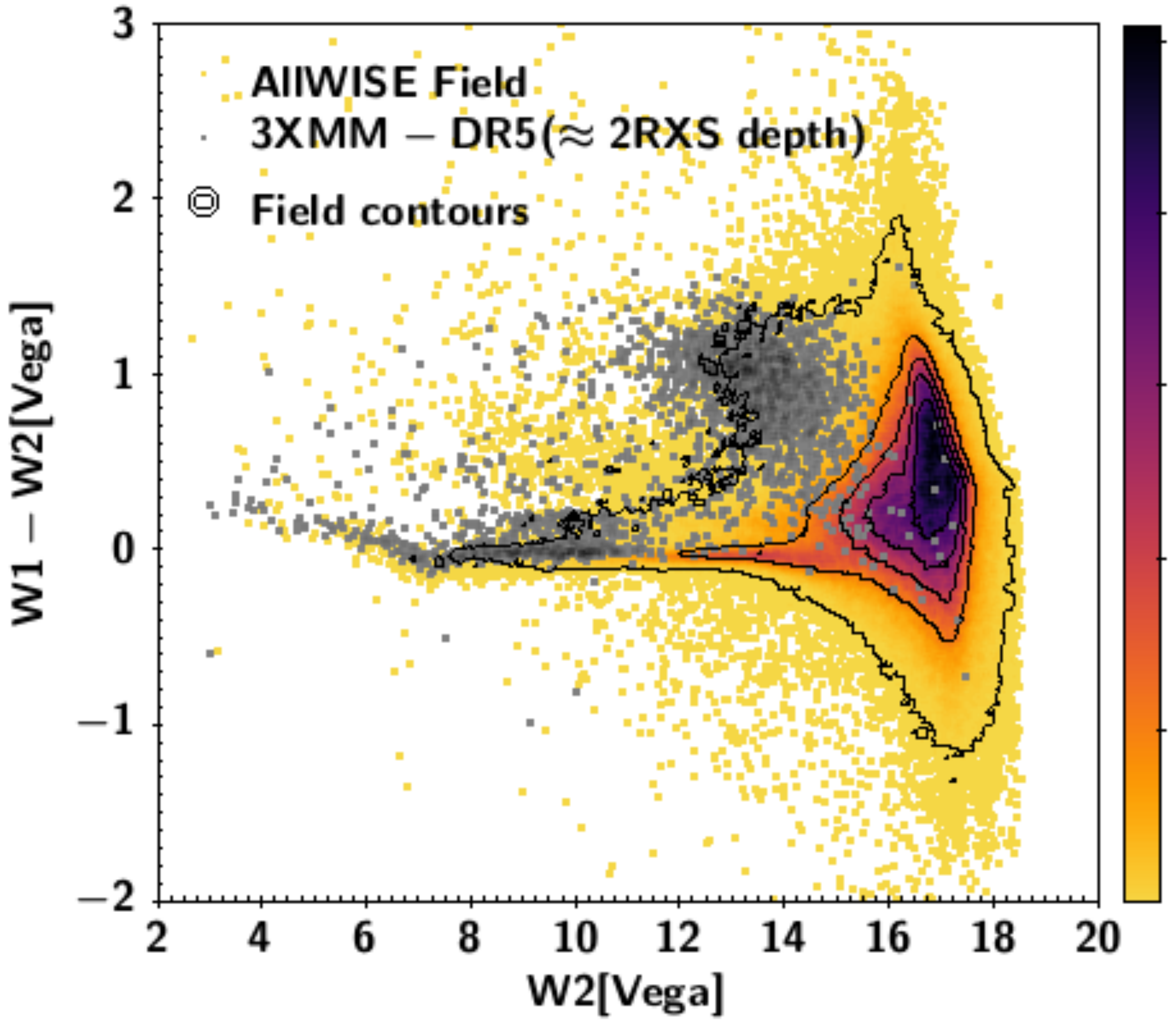


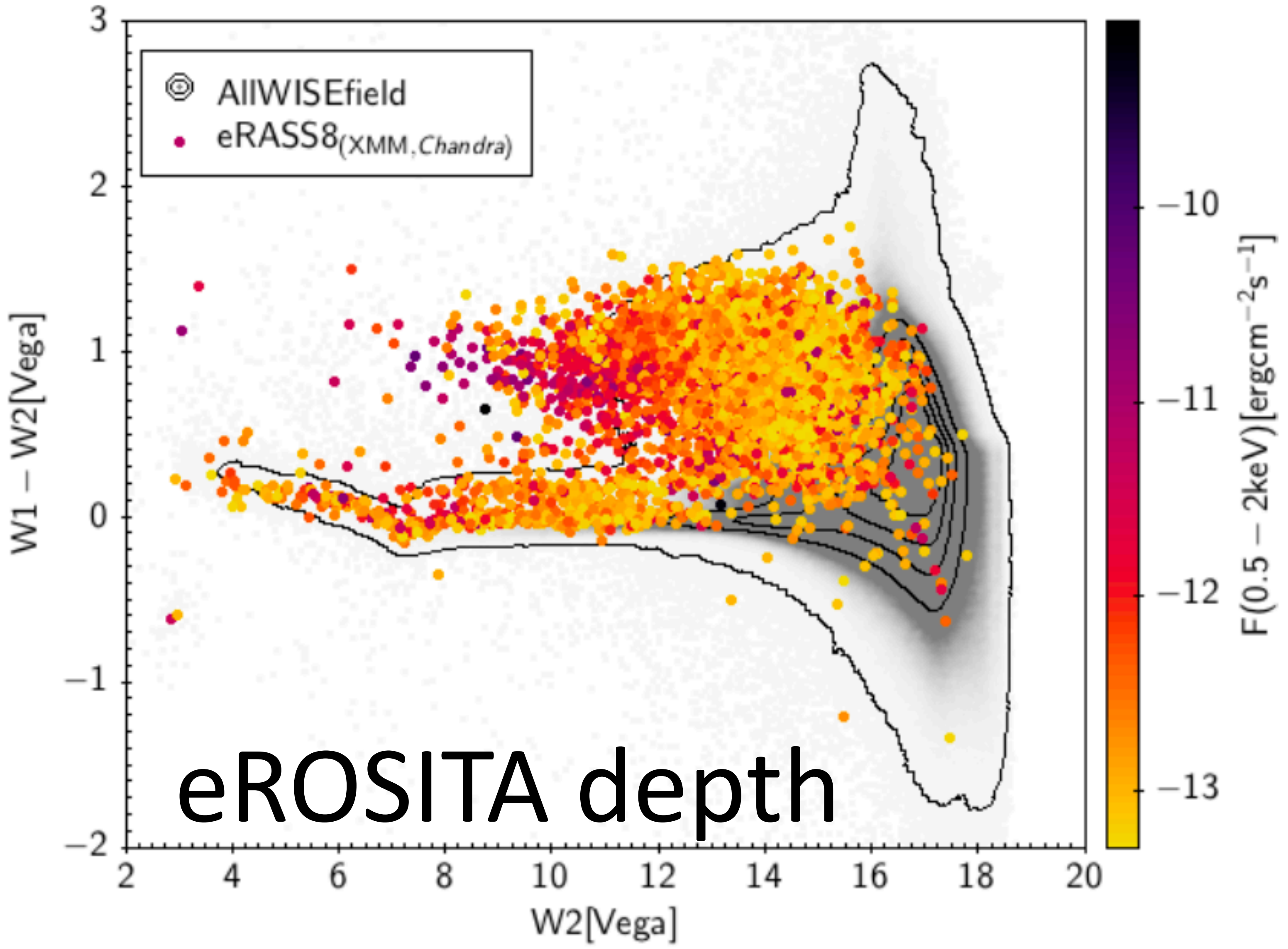
For each problem we need to find the right solution!

W2 vs W1-W2 is NOT ALWAYS the solution!
the parameter space may provide no information

For each problem we need to find the right solution!

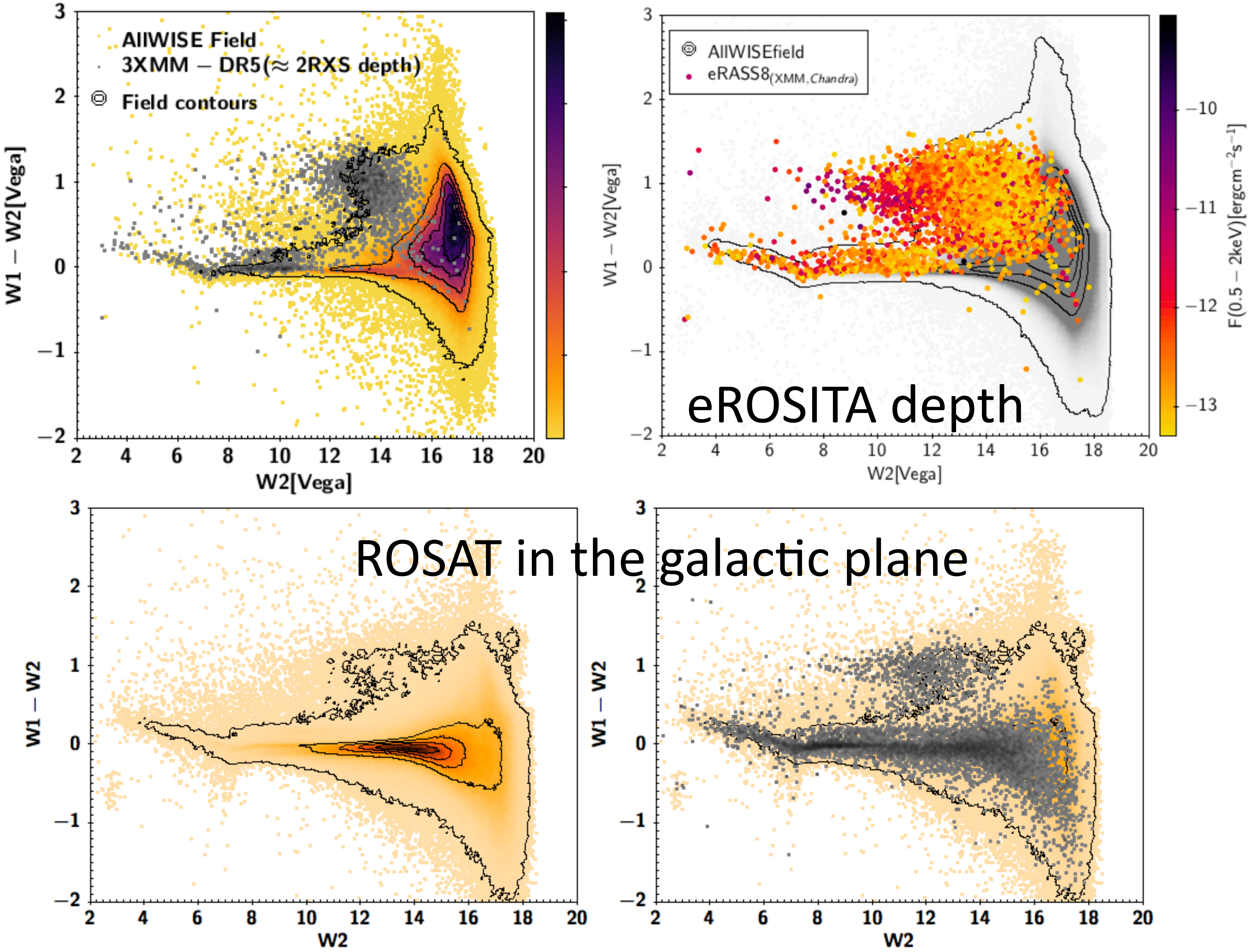
W2 vs W1-W2 is NOT ALWAYS the solution!
the parameter space may provide no information





For each problem we need to find the right solution!

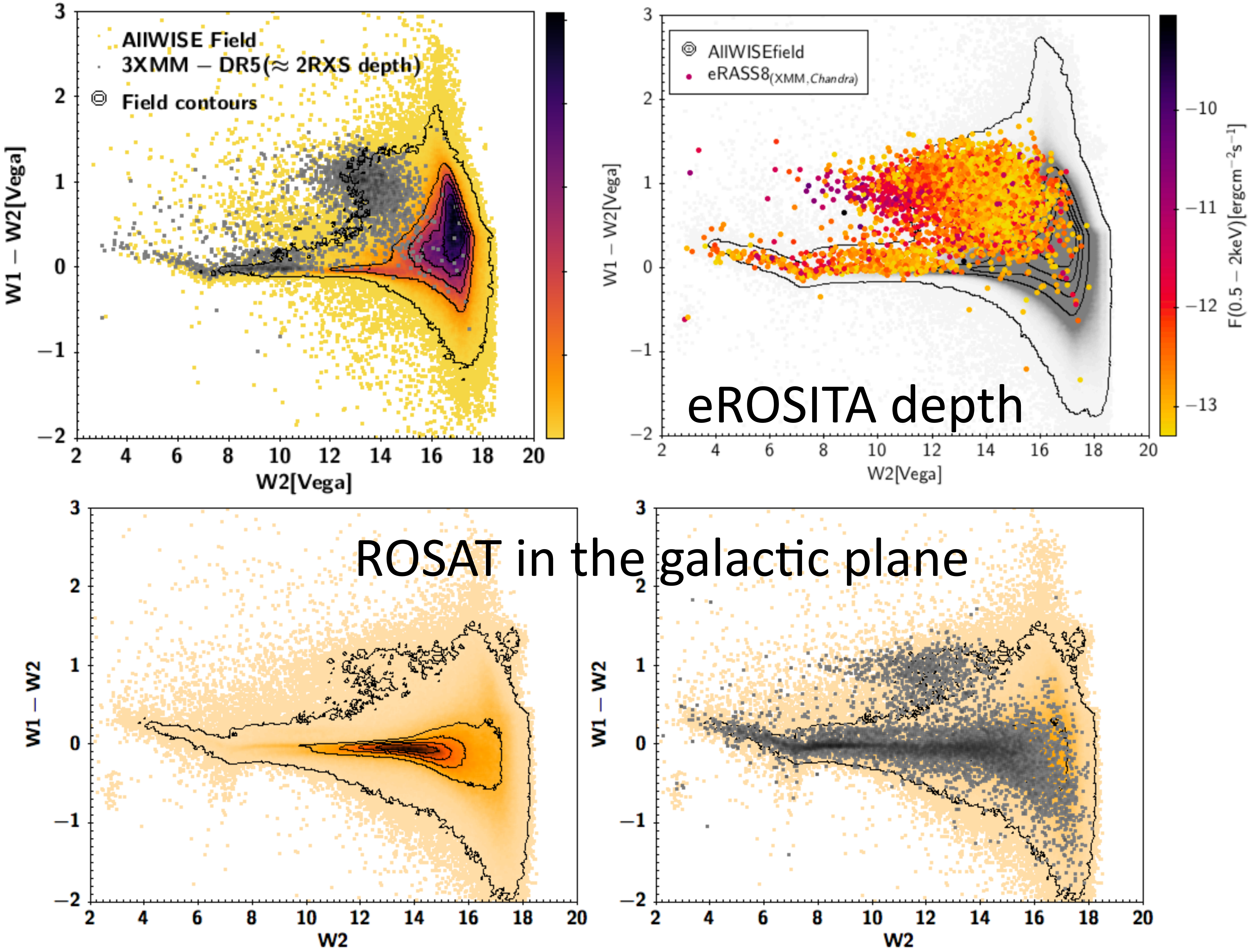
W2 vs W1-W2 is NOT ALWAYS the solution!
the parameter space may provide no information



For each problem we need to find the right solution!

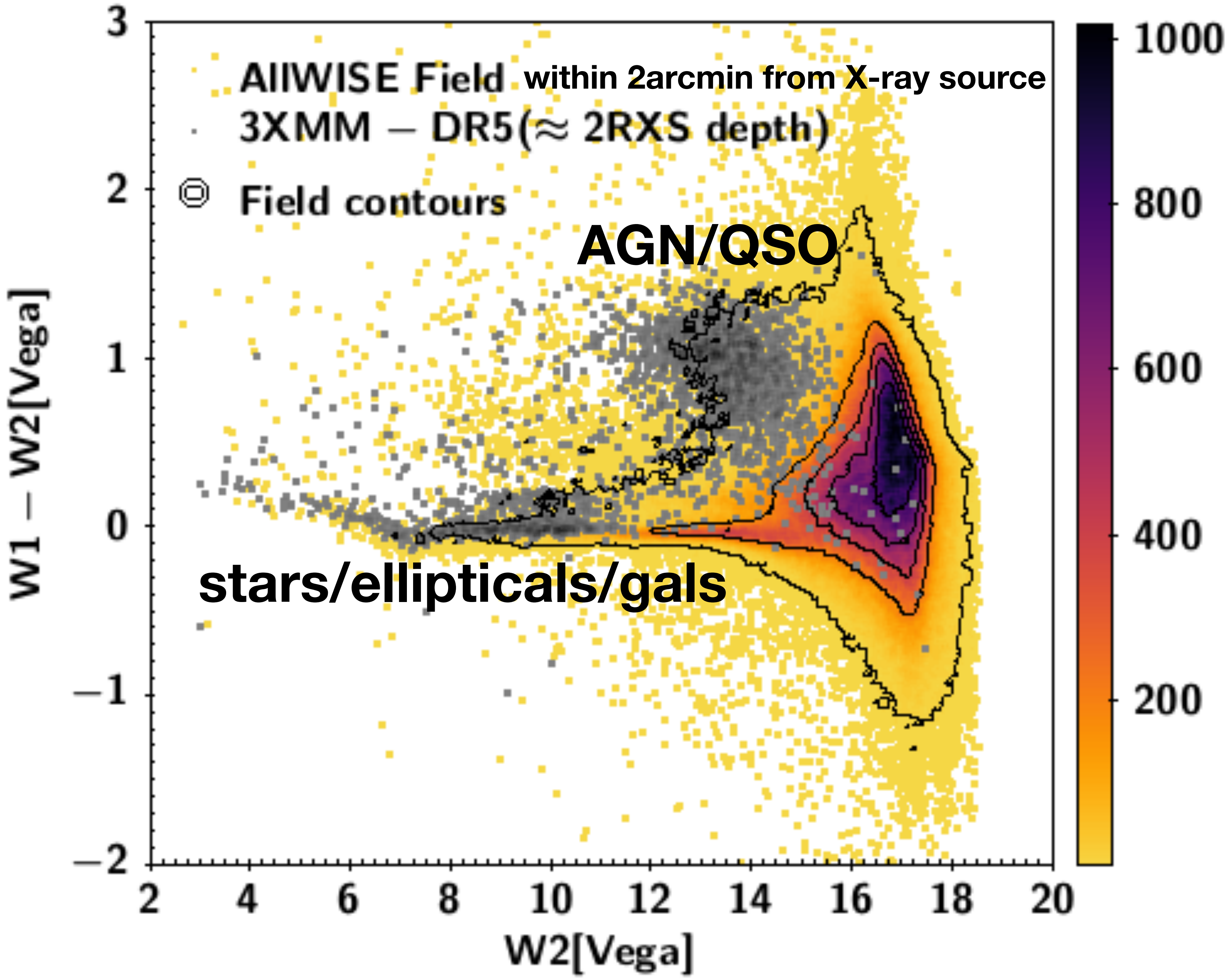
W2 vs W1-W2 is NOT ALWAYS the solution!
the parameter space may provide no information

In each case, we need to find the
way to disentangle the actual CTP from the field



From Brain Learning to Machine Learning

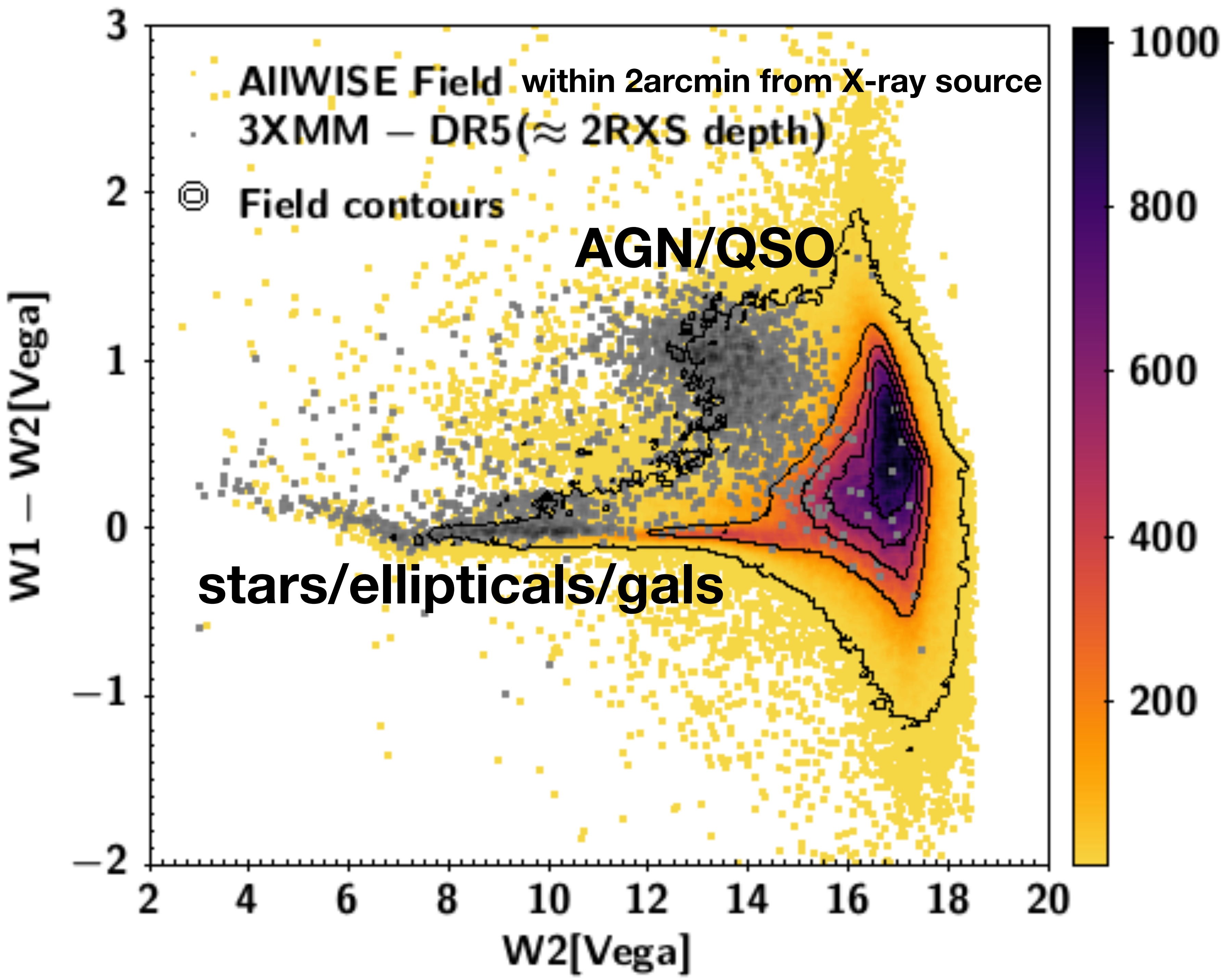
NWAY on ROSAT & XMMSLEW2 (Salvato+ 2018)



NWAY on eROSITA/eFEDS (Salvato+ 2022)

From Brain Learning to Machine Learning

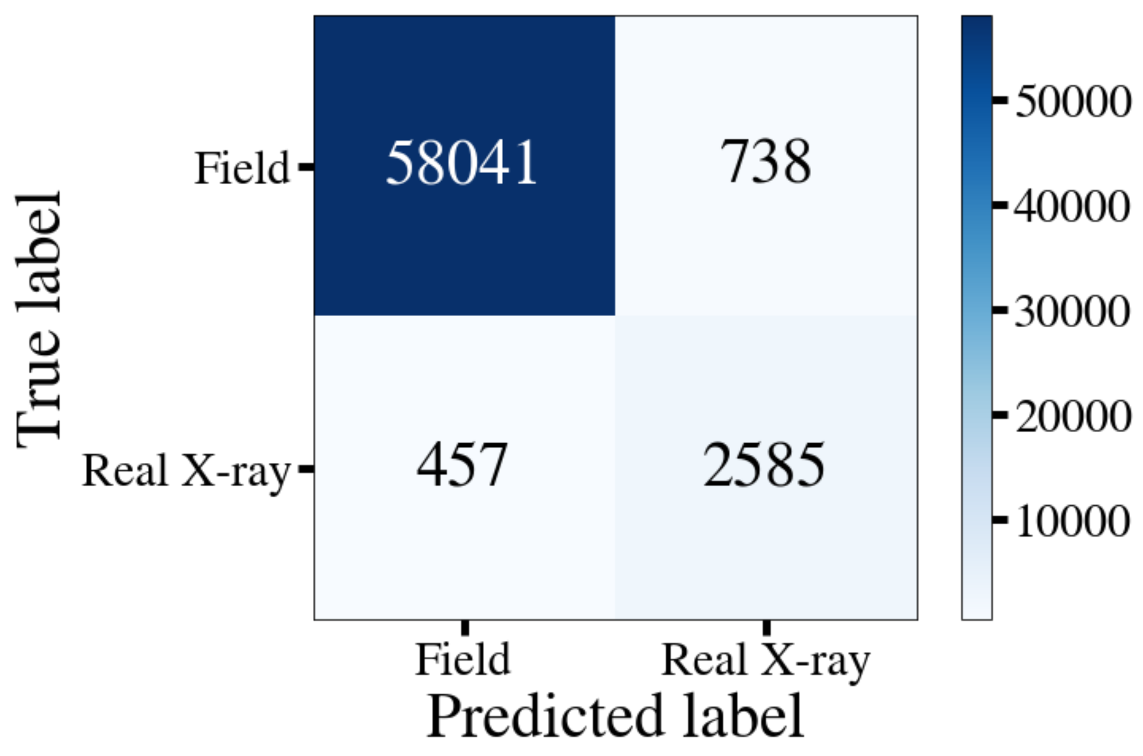
NWAY on ROSAT & XMMSLEW2 (Salvato+ 2018)



NWAY on eROSITA/eFEDS (Salvato+ 2022)

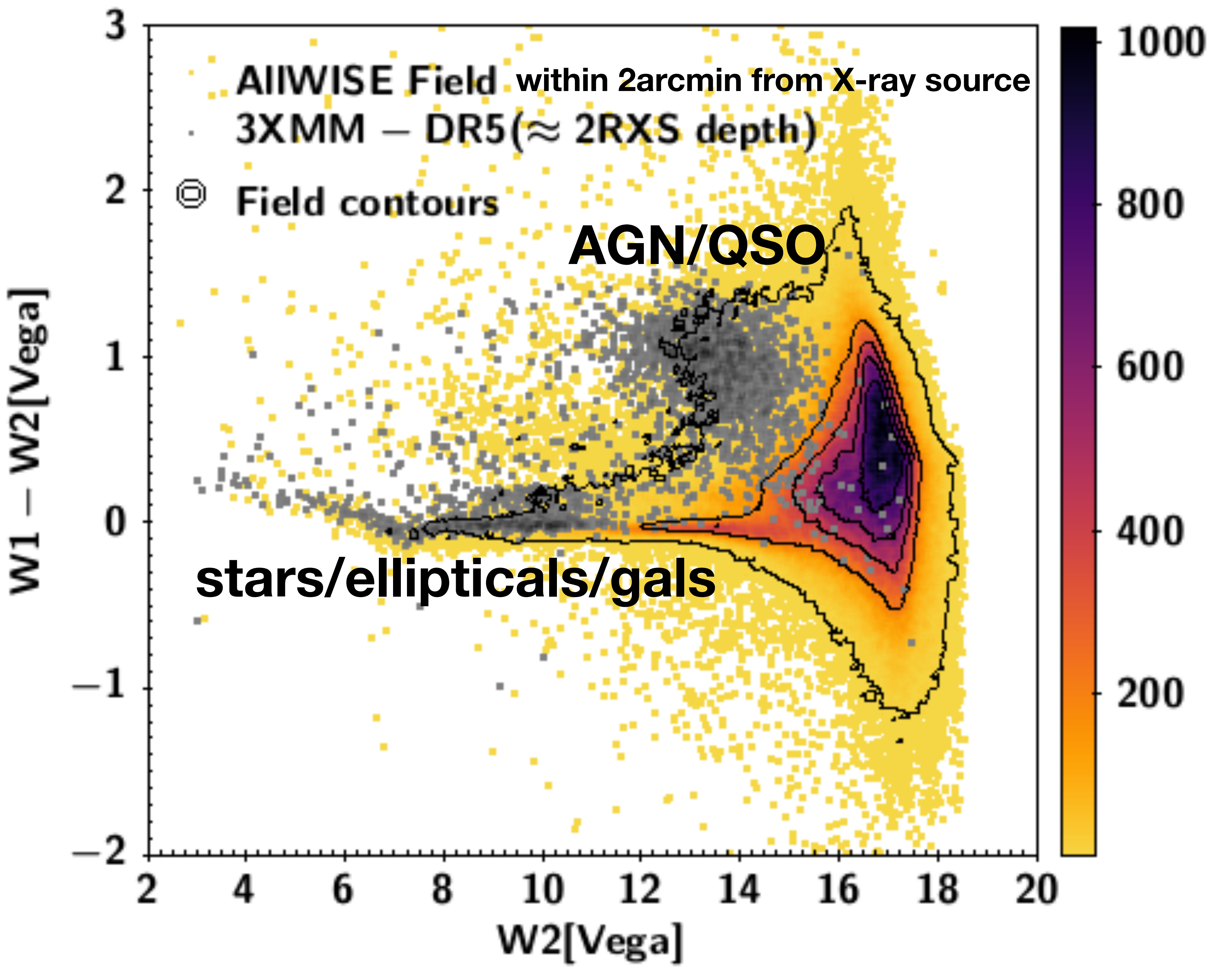
- Random Forest on :
  - training sample: 23K XMM sources with depth comparable to eFEDS and w/ secure CTP in Legacy Survey DR8 (Dey+2019).
  - control sample: the rest of the sources within 30” from each X-ray position

Feature	Description
flux_*/mw_transmission_*	deredenned flux in $g,r,z,W1,W2$
gaia_phot_*_mean_mag	original GAIA phot. in G, G_{bp}, G_{rp}
snr_*	S/N for $g,r,z,W1,W2,G,G_{bp},G_{rp}$
$\sqrt{pmra^2 + pmdec^2}$	Gaia proper motion
parallax	Gaia paralllax
g-r, r-z, z-$W1$, r-$W2$	dereddened colors



From Brain Learning to Machine Learning

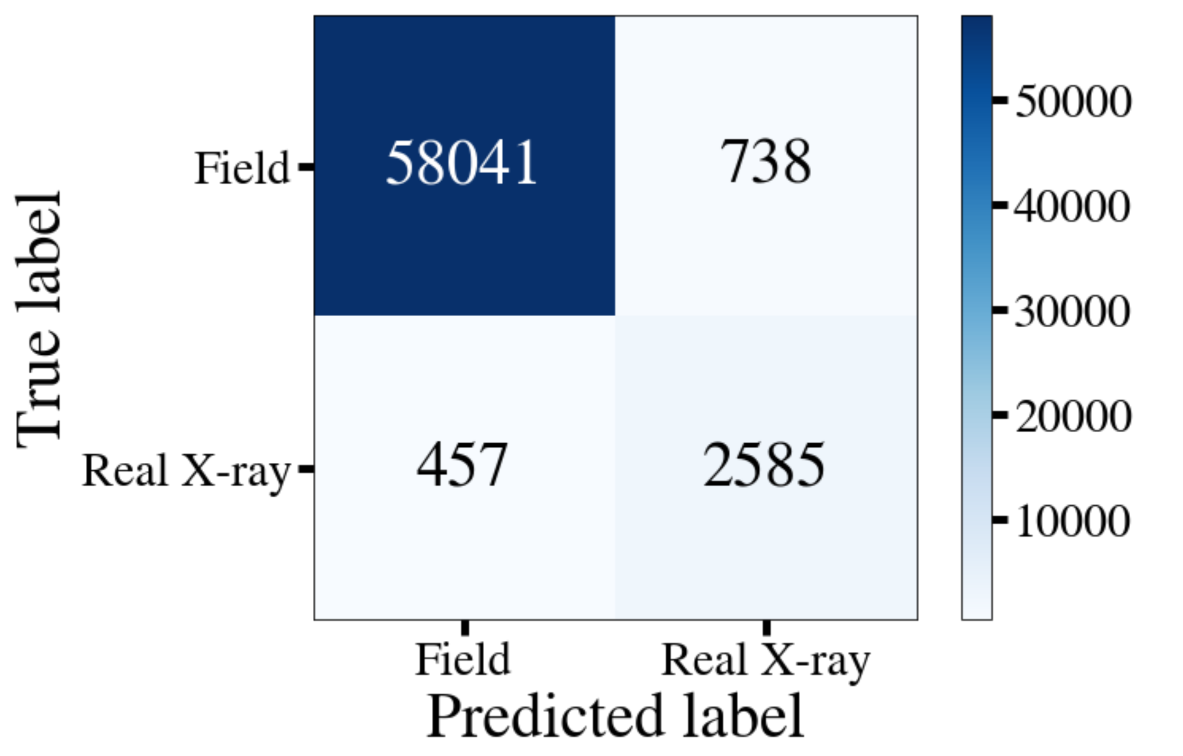
NWAY on ROSAT & XMMSLEW2 (Salvato+ 2018)



NWAY on eROSITA/eFEDS (Salvato+ 2022)

- Random Forest on :
  - training sample: 23K XMM sources with depth comparable to eFEDS and w/ secure CTP in Legacy Survey DR8 (Dey+2019).
  - control sample: the rest of the sources within 30” from each X-ray position

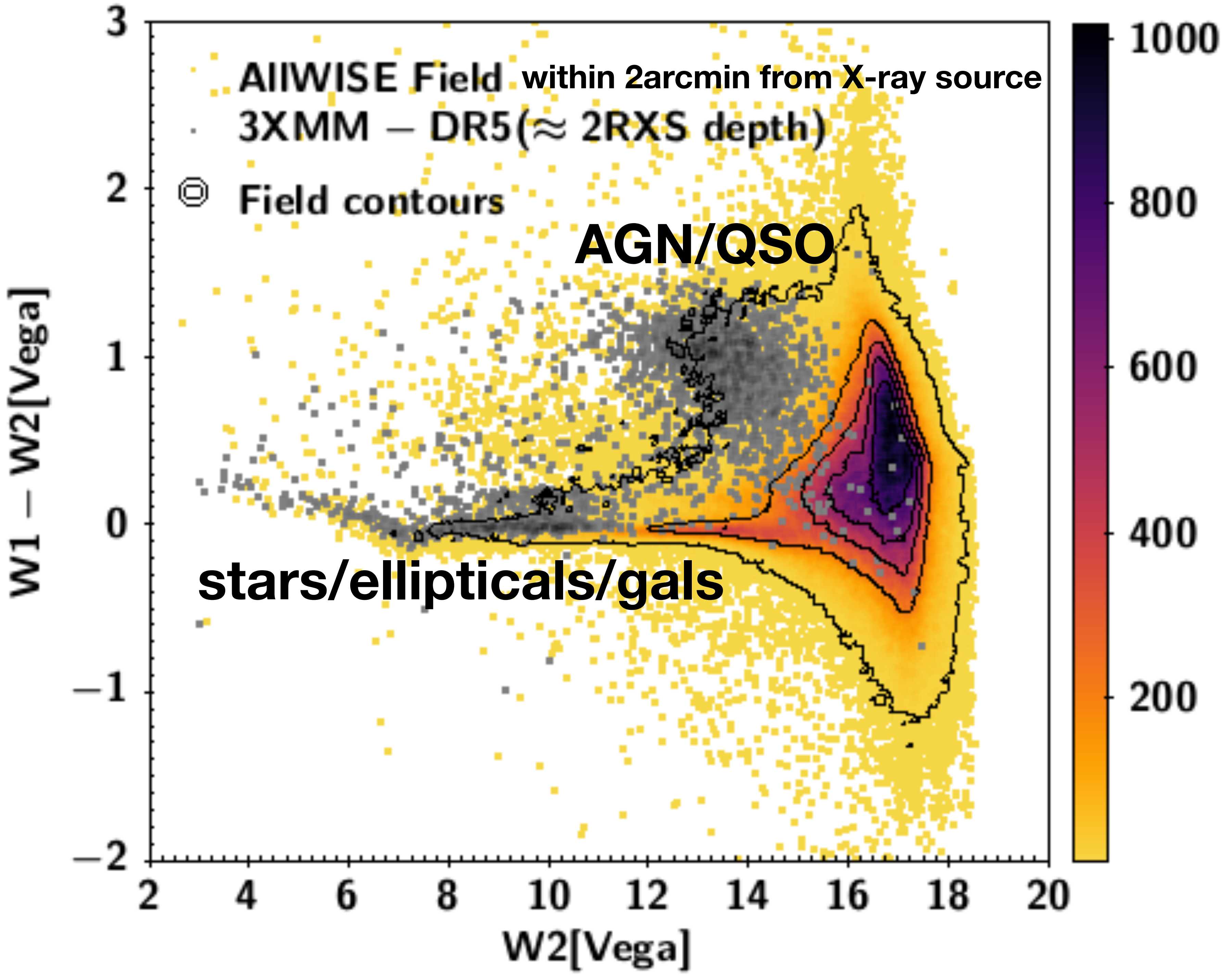
Feature	Description
flux_*/mw_transmission_*	deredenned flux in $g,r,z,W1,W2$
gaia_phot_*_mean_mag	original GAIA phot. in G, G_{bp}, G_{rp}
snr_*	S/N for $g,r,z,W1,W2,G,G_{bp},G_{rp}$
$\sqrt{pmra^2 + pmdec^2}$	Gaia proper motion
parallax	Gaia paralllax
g-r, r-z, z-$W1$, r-$W2$	dereddened colors



- validation sample: 3500 Chandra sources with depth comparable to eFEDS, **made eROSITA-like**

From Brain Learning to Machine Learning

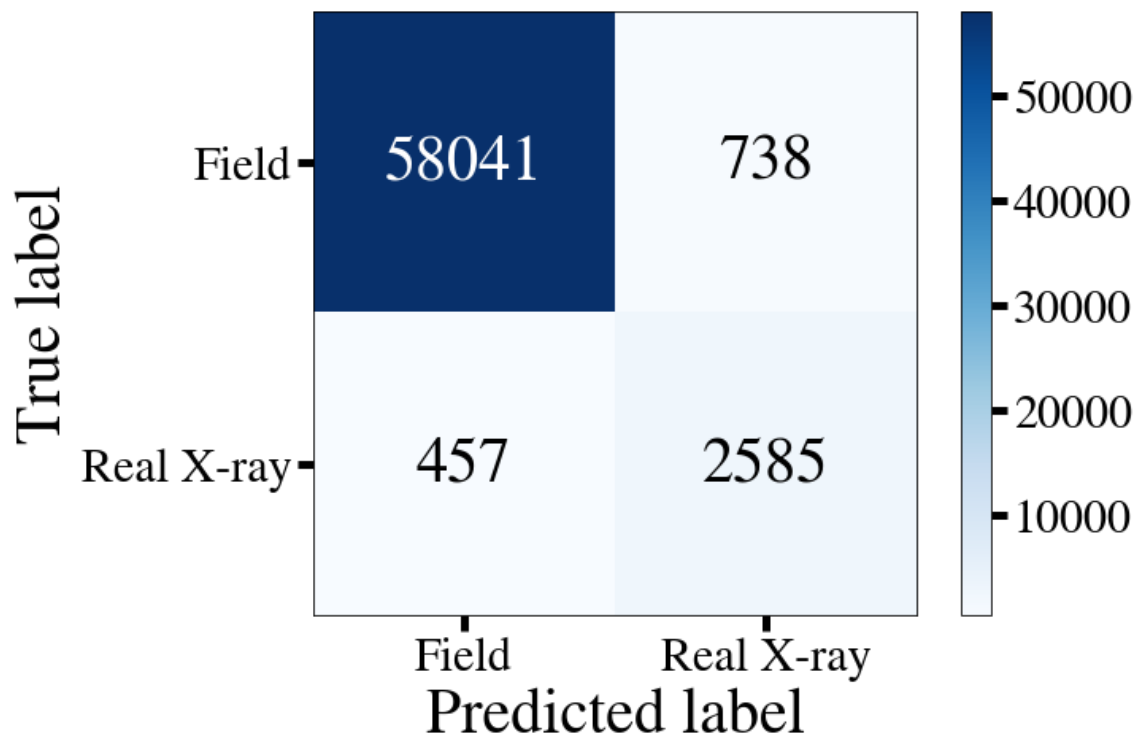
NWAY on ROSAT & XMMSLEW2 (Salvato+ 2018)



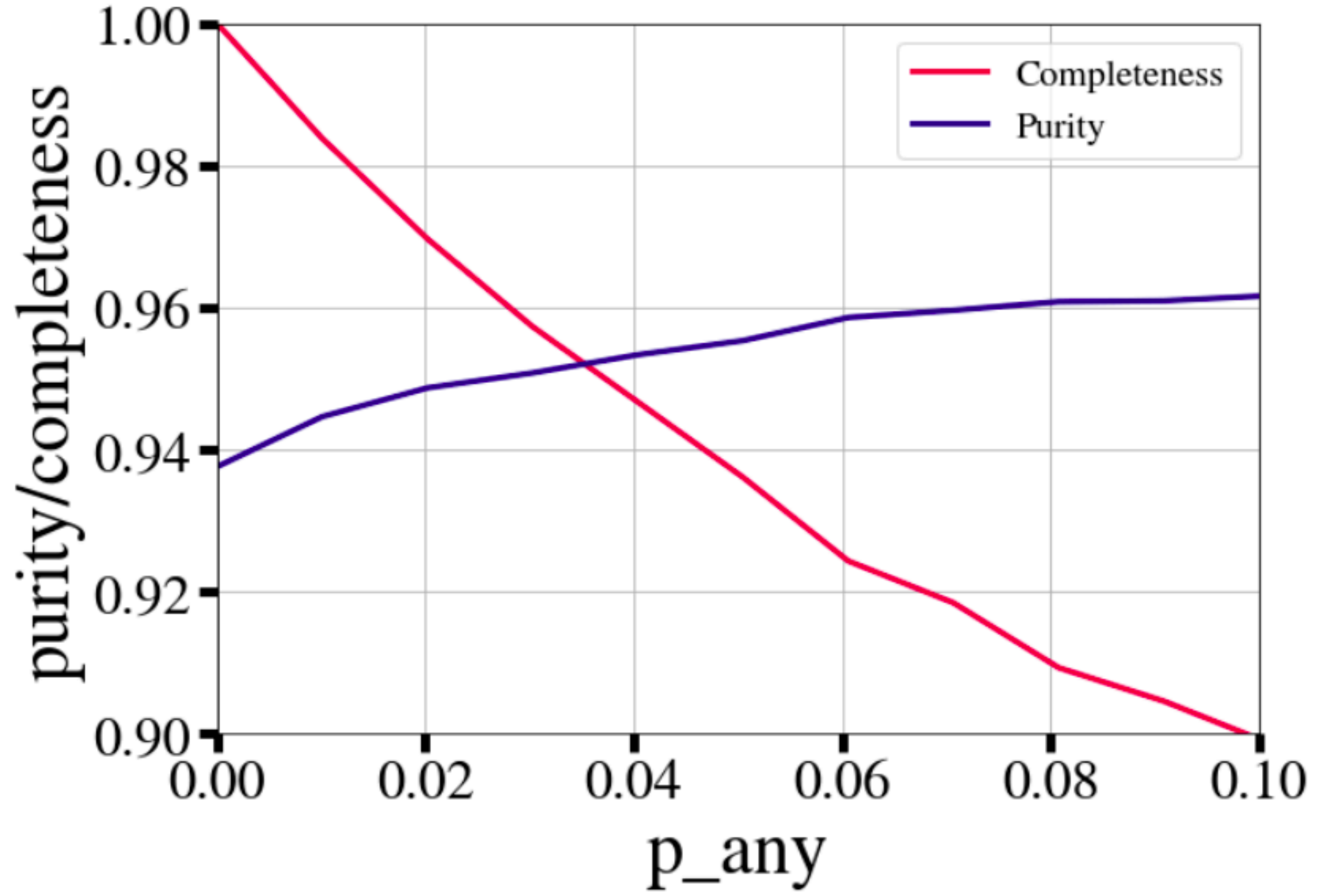
NWAY on eROSITA/eFEDS (Salvato+ 2022)

- Random Forest on :
  - training sample: 23K XMM sources with depth comparable to eFEDS and w/ secure CTP in Legacy Survey DR8 (Dey+2019).
  - control sample: the rest of the sources within 30” from each X-ray position

Feature	Description
flux_*/mw_transmission_*	deredenned flux in $g,r,z,W1,W2$
gaia_phot_*_mean_mag	original GAIA phot. in G, G_{bp}, G_{rp}
snr_*	S/N for $g,r,z,W1,W2,G,G_{bp},G_{rp}$
$\sqrt{pmra^2 + pmdec^2}$	Gaia proper motion
parallax	Gaia paralllax
g-r, r-z, z-$W1$, r-$W2$	dereddened colors



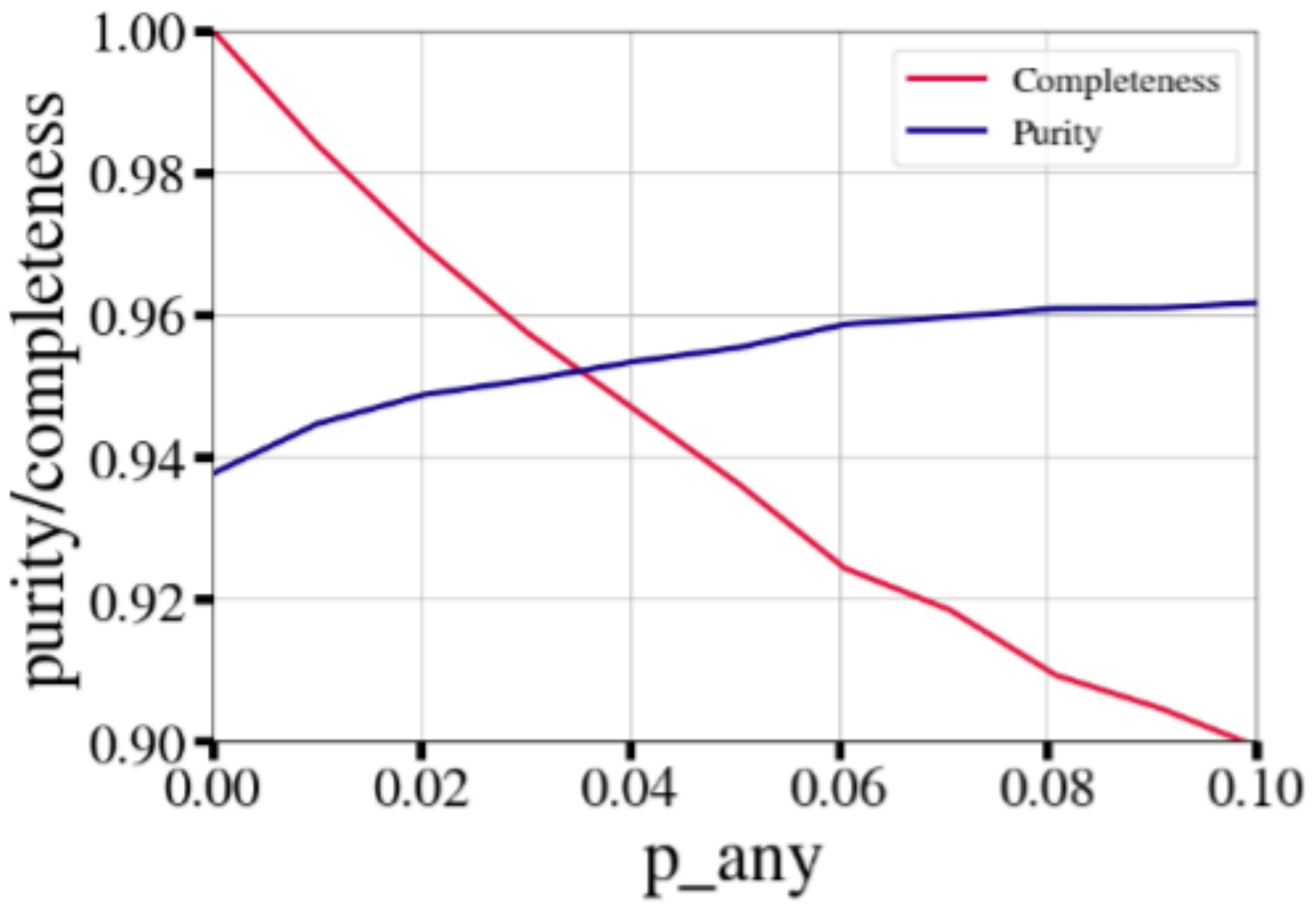
- validation sample: 3500 Chandra sources with depth comparable to eFEDS, **made eROSITA-like**



Comparison between methods

NWAY

Feature	Description
flux_*/mw_transmission_*	deredenned flux in $g,r,z,W1,W2$
gaia_phot_*_mean_mag	original GAIA phot. in G, G_{bp}, G_{rp}
snr_*	S/N for $g,r,z,W1,W2,G,G_{bp},G_{rp}$
$\sqrt{pmra^2 + pmdec^2}$	Gaia proper motion
parallax	Gaia paralllax
g-r, r-z, z-$W1$, r-$W2$	dereddened colors



95% correct ctp

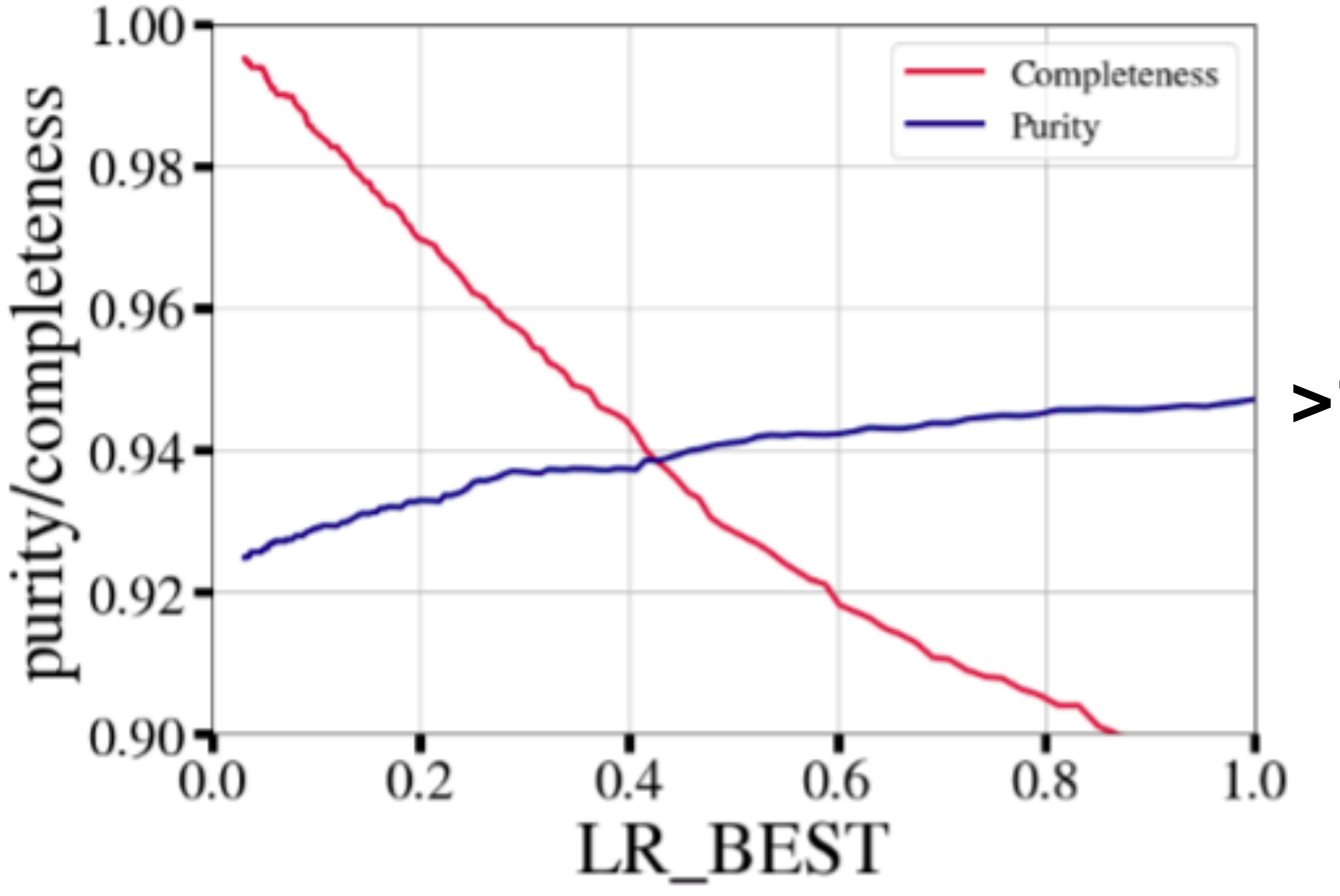
2% w/ second possible CTP

MLR in Astromatch (Ruiz+2018)

Using the same training, control and validation samples

- 3-D distribution W2, W1-W2, TYPE
- 3-D distribution r-W2, g, TYPE
- g band

then select the CTP with higher LR from one of the 3 methods



89% correct ctp

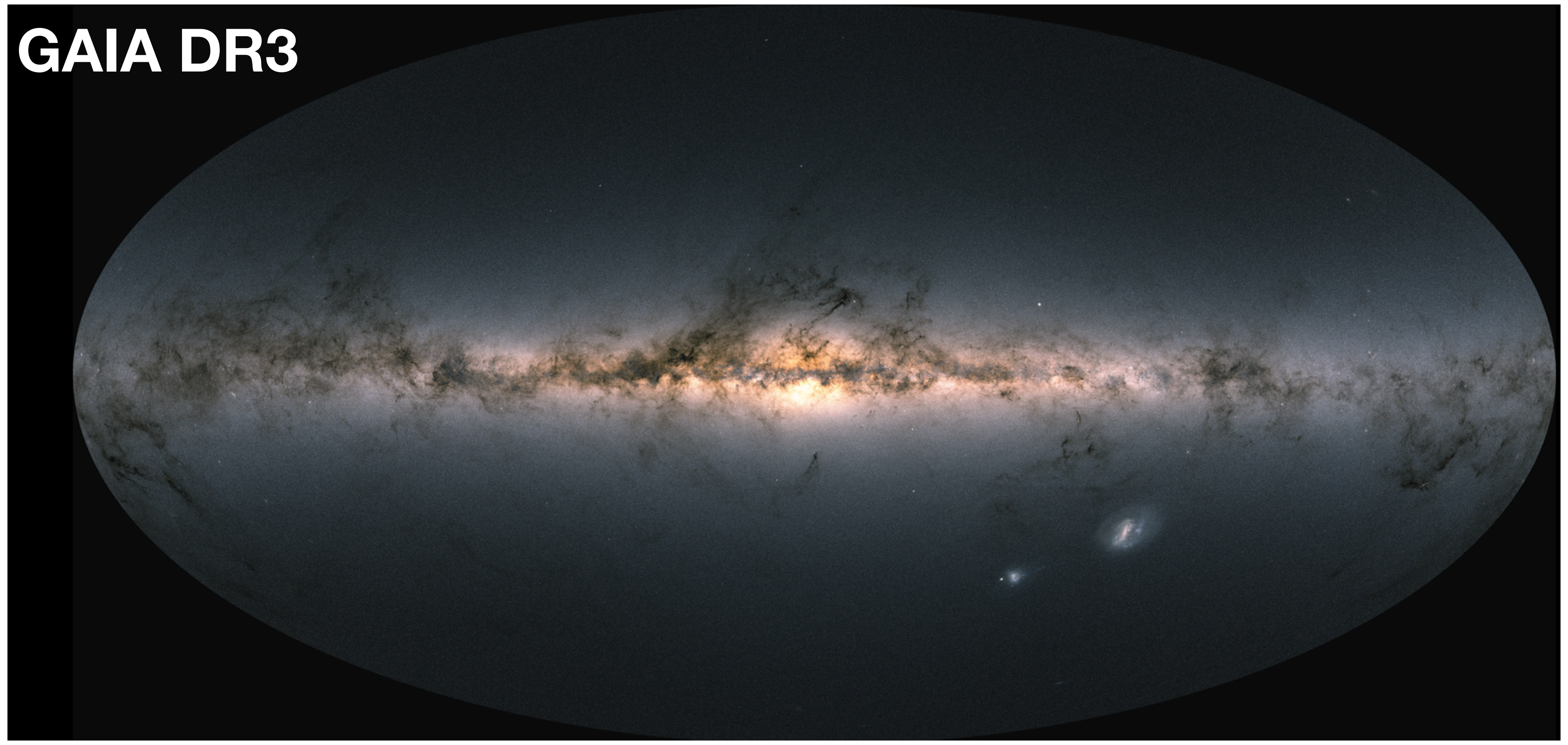
>10% w/ second possible CTP

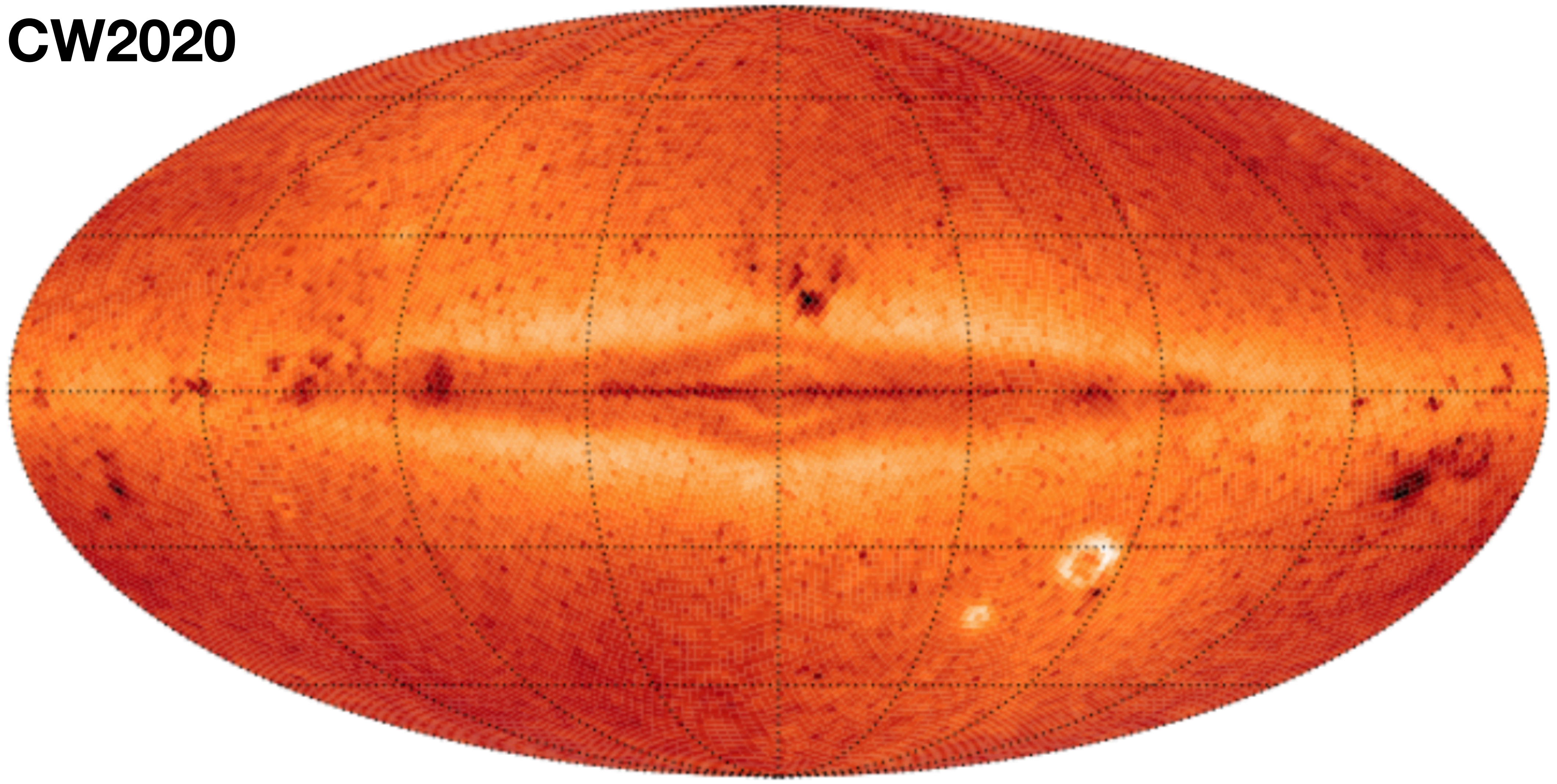
88% agreement, with fraction of disagreement increasing with the positional error

(details in Salvato+2022)

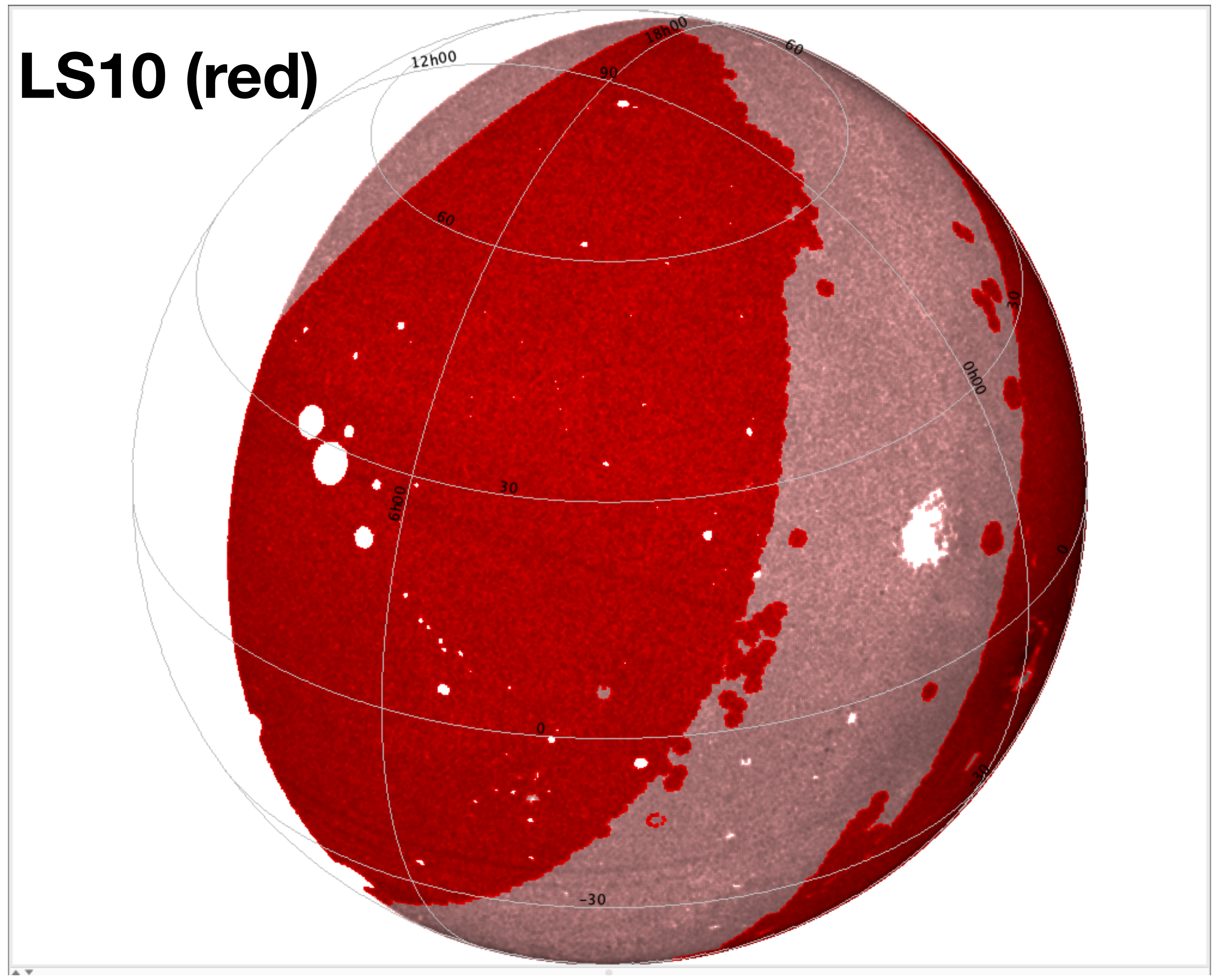
For all-sky, we are depending on the availability of suitable ancillary data

(details in Salvato+2024 in prep)





Survey	Depth	Bands	Coverage	targets
Gaia	20	Gr,Gb,G	all-sky	stars, compact objects,qso
CW2020	20.4, 20.8	W1,W2	all-sky	cold stars, QSO, AGN
LS10	~24	griz(W1-4)	14k sq.deg	all, including clusters



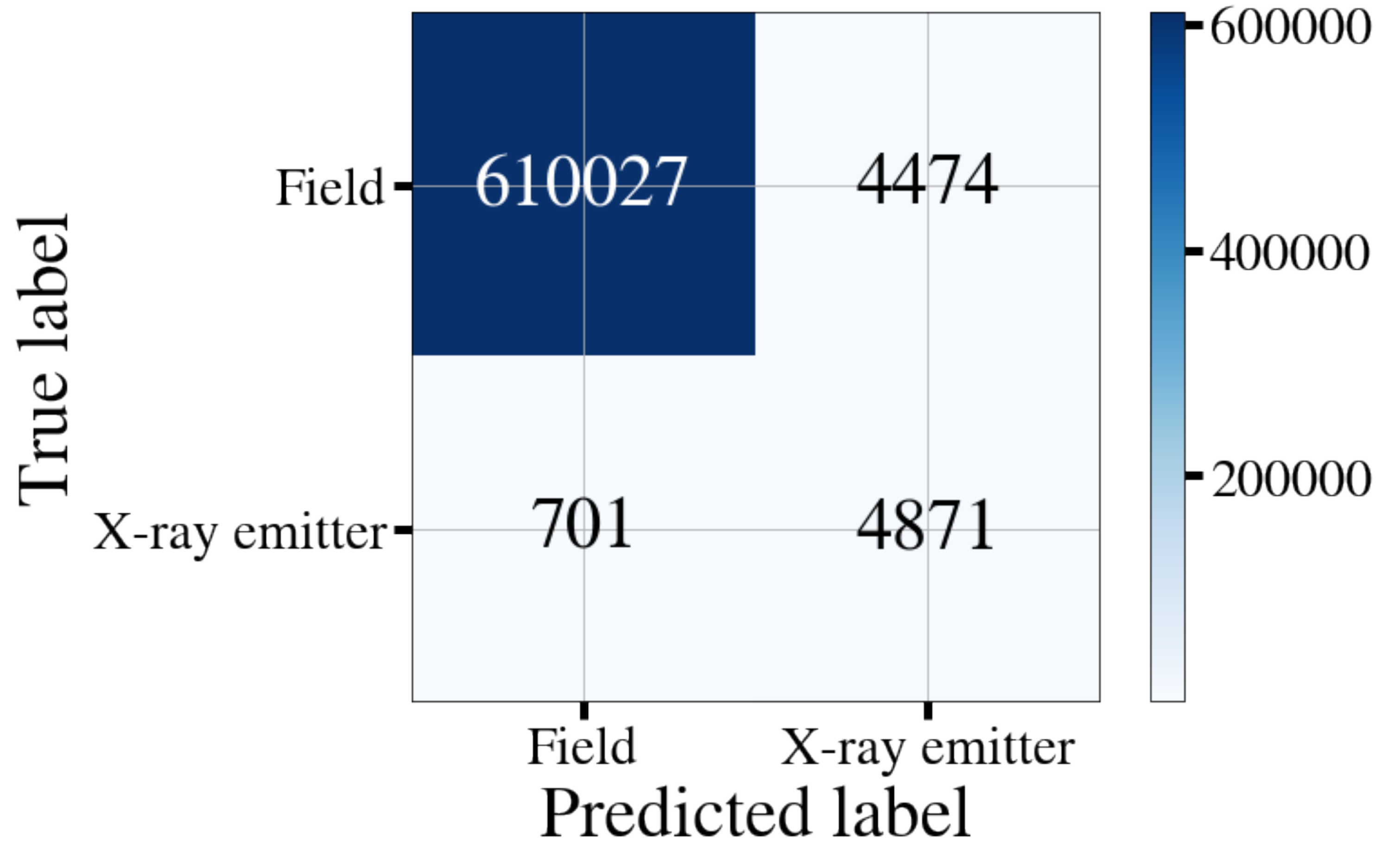
NWAy_RF: the features used as priors

(details in Salvato+2024 in prep)

- LS10

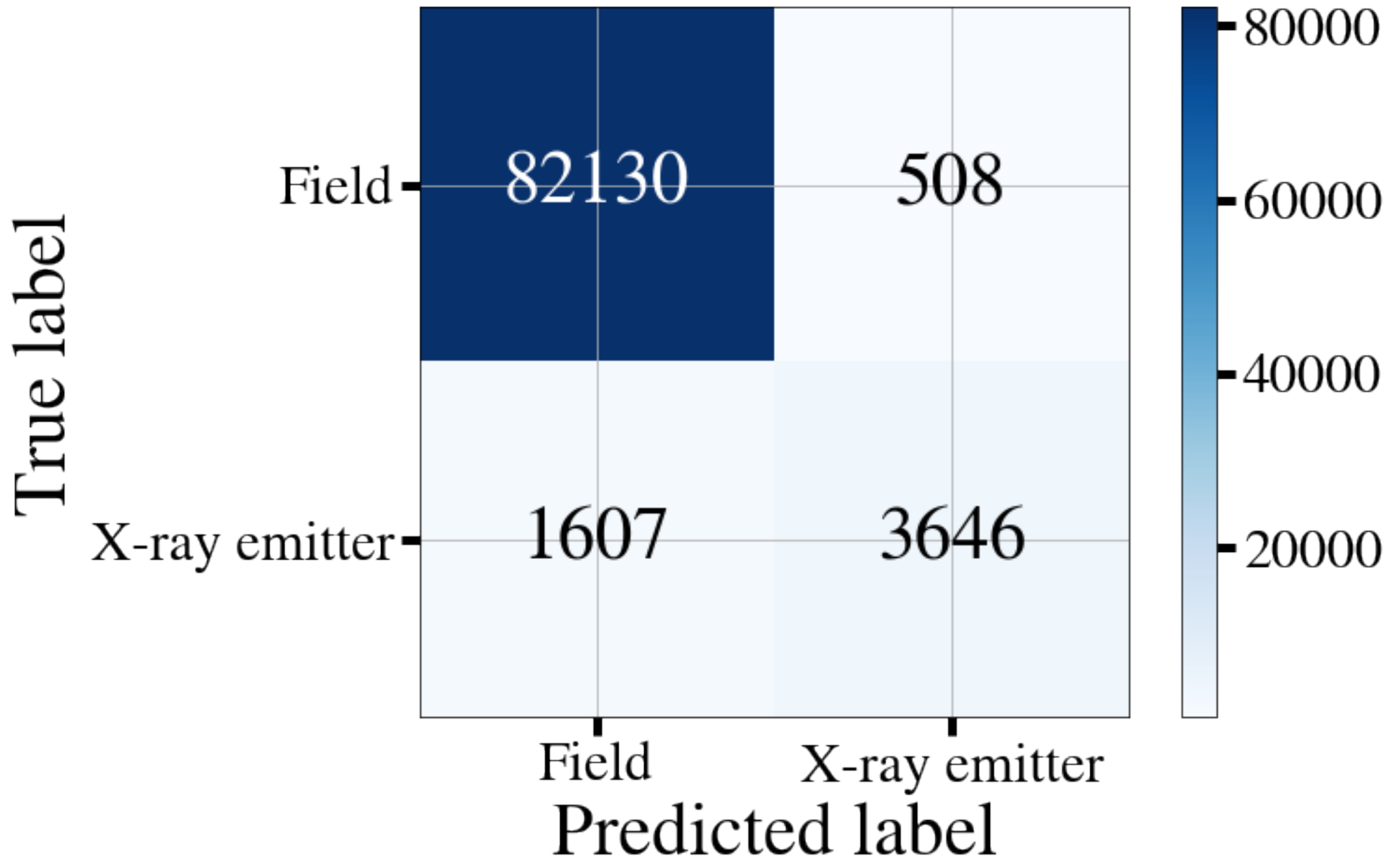
```
extinction-corrected fluxes
flux errors
flux ratios (colors)
shape_r','shape_e1','shape_e2','sersic',
Gaia photometry
Gaia S/N
Gaia proper motion and error
Gaia parallax and error
```

- 87.5% recall fraction
- 0.7% leakage



- Gaia EDR3

```
Gaia photometry
Gaia S/N
Gaia proper motion and error
Gaia parallax and error
Gaia astrometric_excess_noise
```

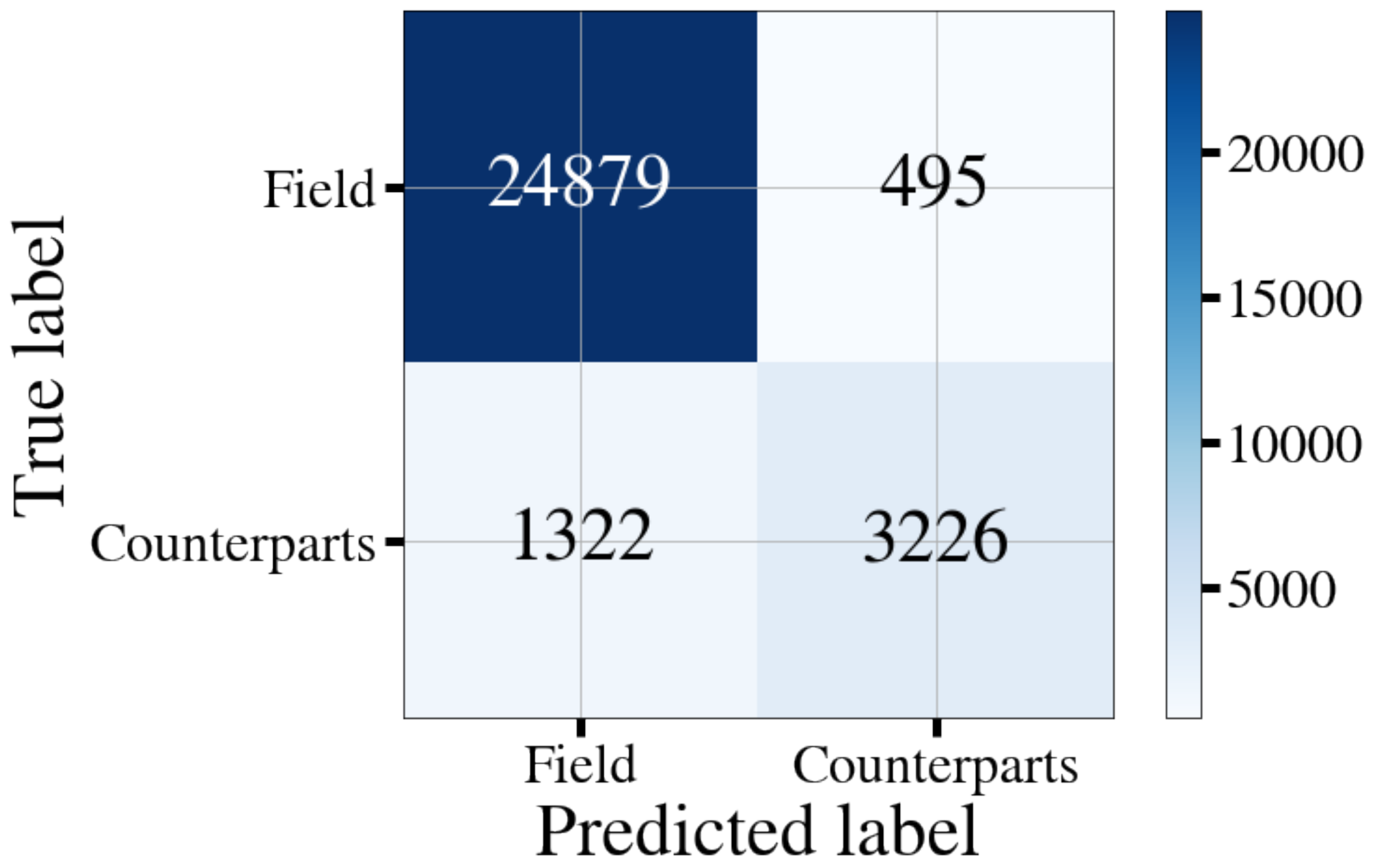


- 69% recall fraction
- 0.7% leakage

- CatWISE 2020

```
W1 and W2 fluxes
W1 an W2 aperture photometry
W1-W2 color
Flux errors
proper motion and error
```

- 71% recall fraction
- 2% leakage



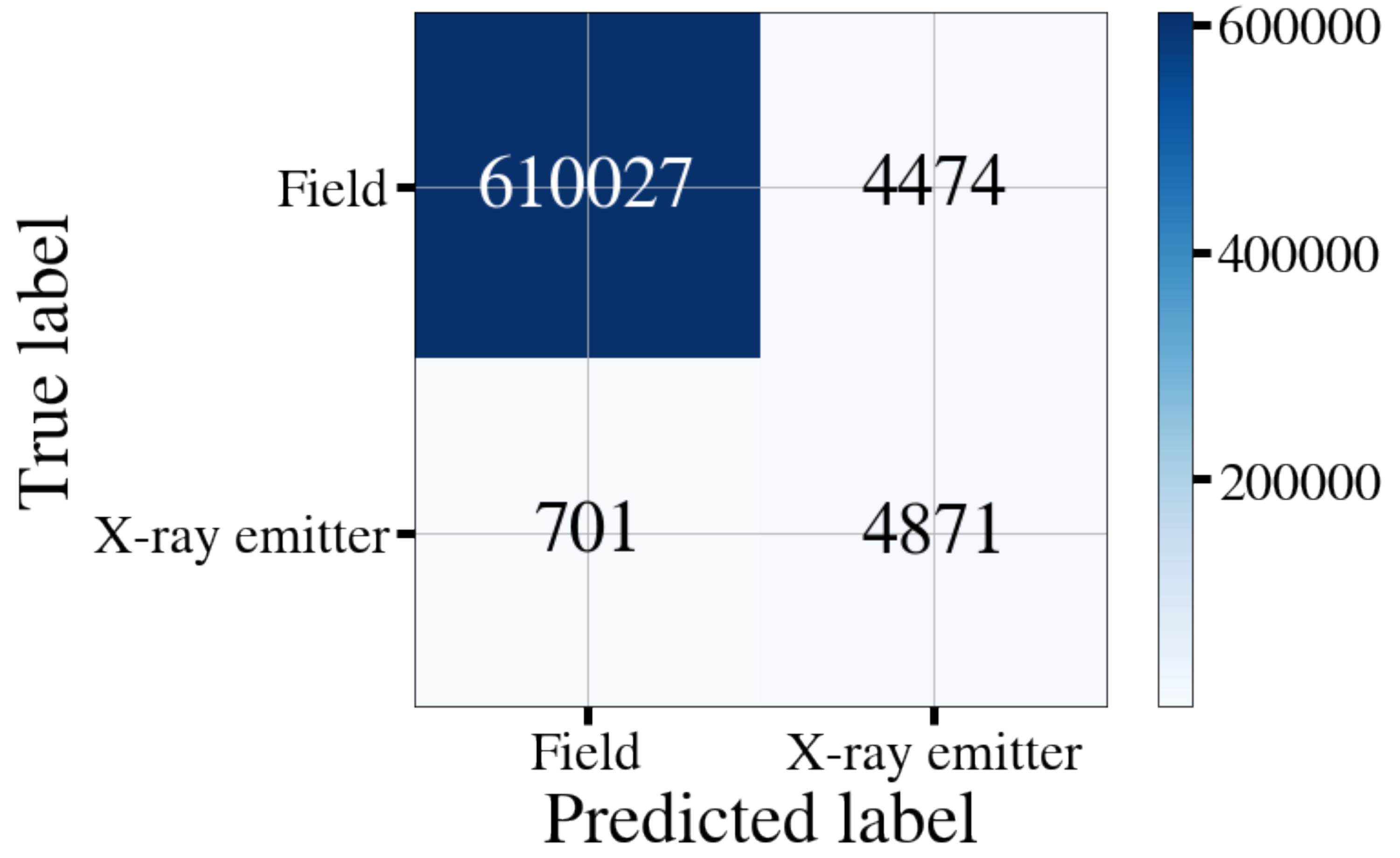
NWAy_RF: the features used as priors

(details in Salvato+2024 in prep)

- LS10

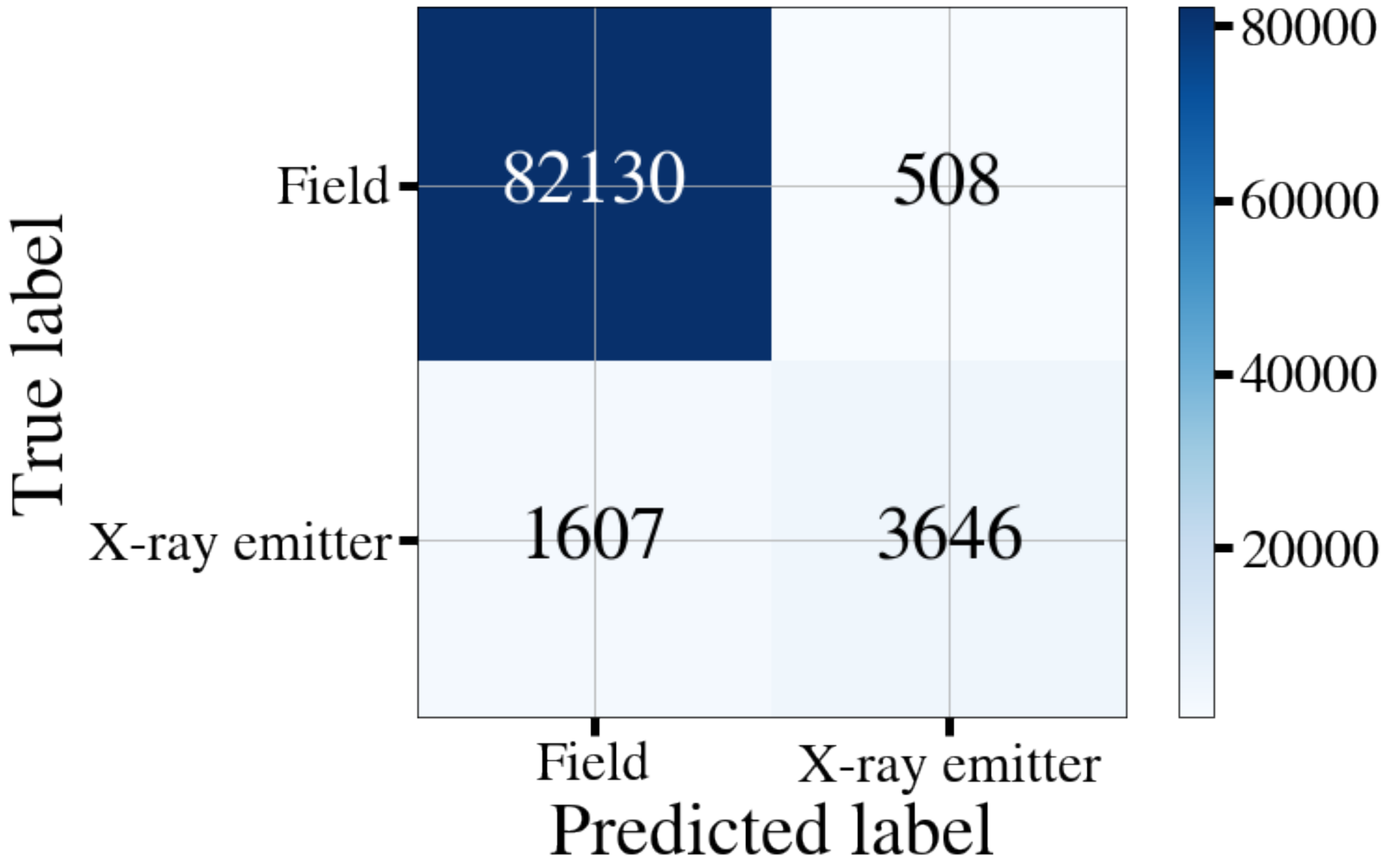
```
extinction-corrected fluxes
flux errors
flux ratios (colors)
shape_r','shape_e1','shape_e2','sersic',
Gaia photometry
Gaia S/N
Gaia proper motion and error
Gaia parallax and error
```

- 87.5% recall fraction
- 0.7% leakage



- Gaia EDR3

```
Gaia photometry
Gaia S/N
Gaia proper motion and error
Gaia parallax and error
Gaia astrometric_excess_noise
Gaia Colors
Correction for extinction
```

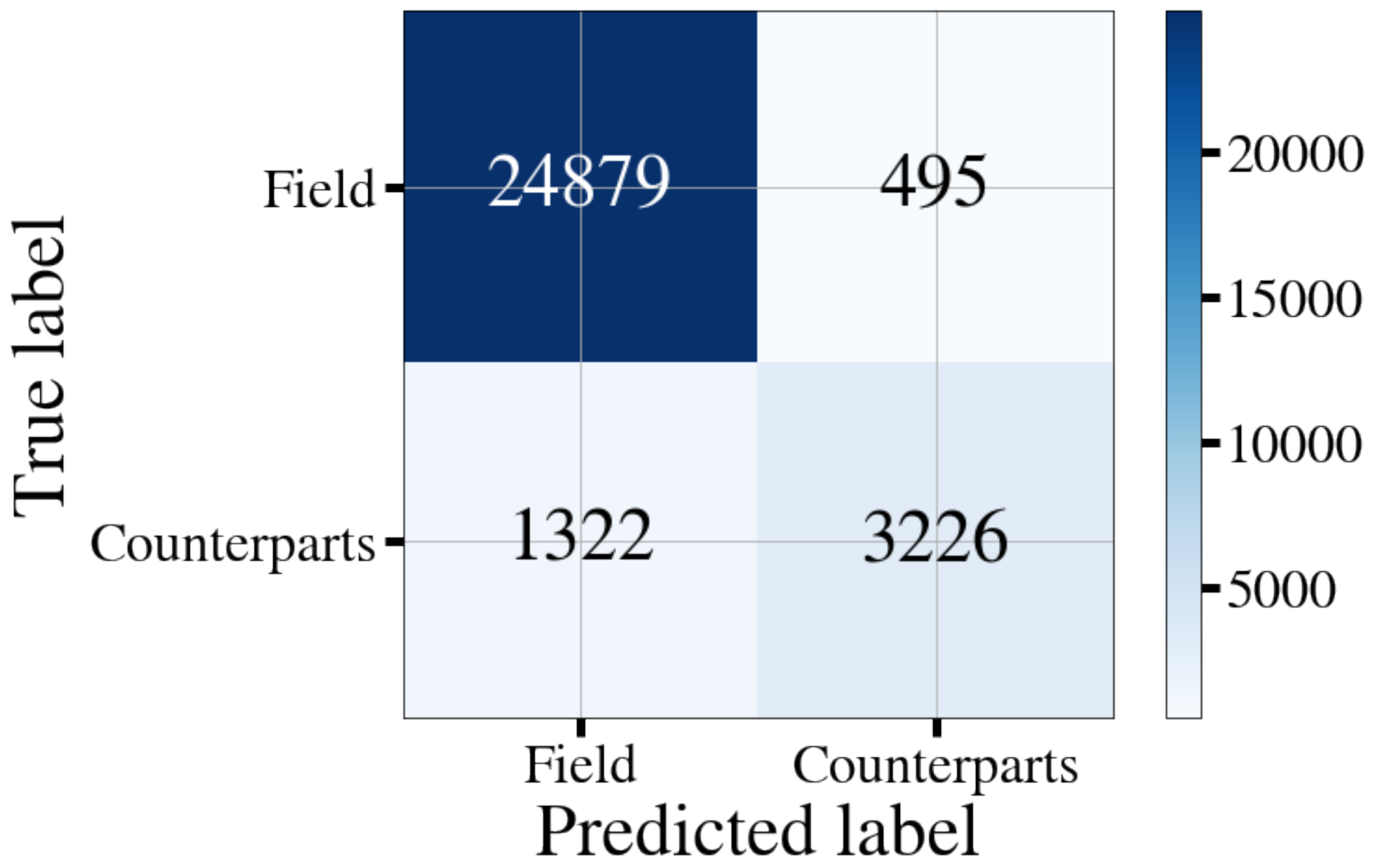


- 69% recall fraction
- 0.7% leakage

- CatWISE 2020

```
W1 and W2 fluxes
W1 an W2 aperture photometry
W1-W2 color
Flux errors
proper motion and error
```

- 71% recall fraction
- 2% leakage



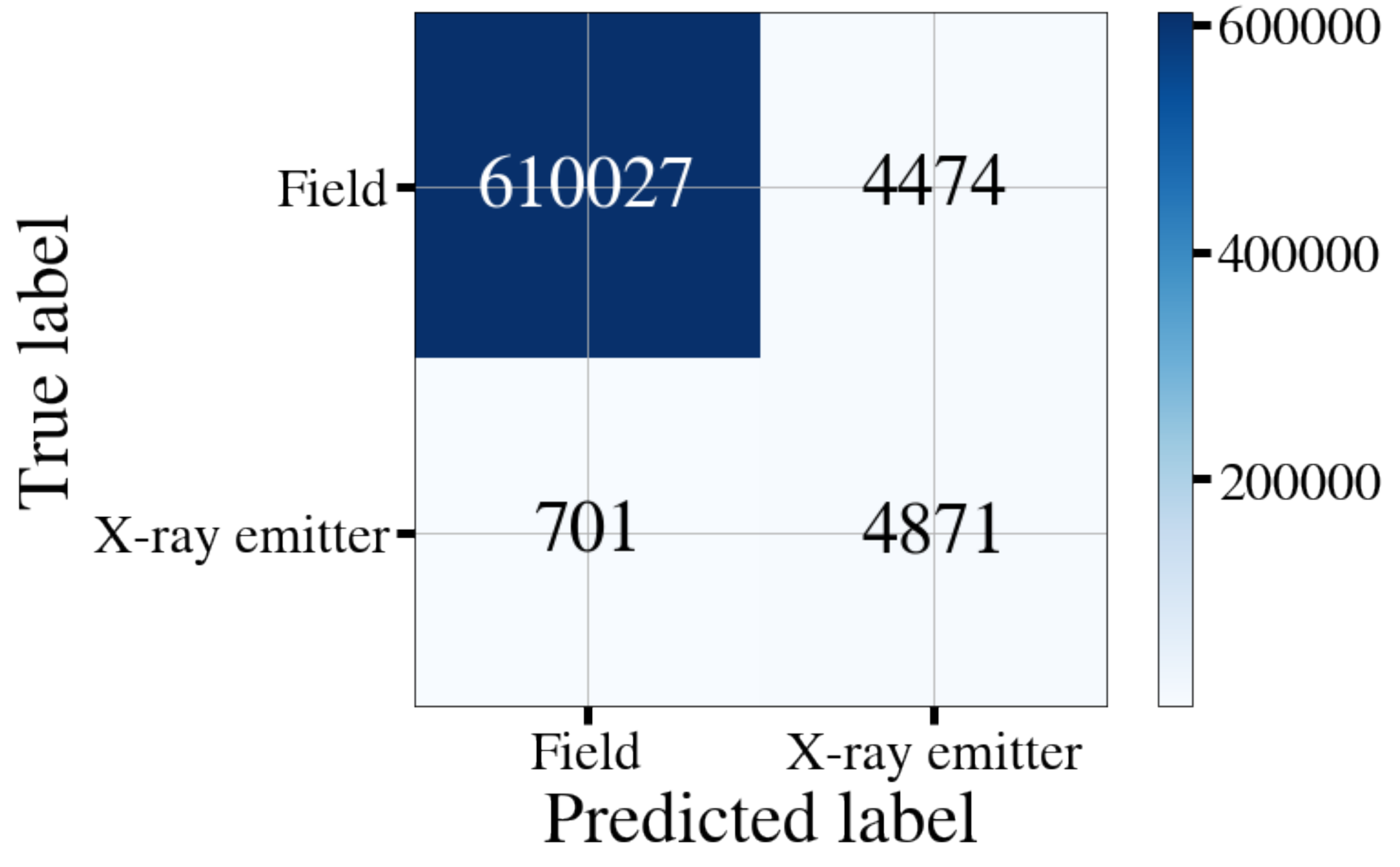
NWAy_RF: the features used as priors

(details in Salvato+2024 in prep)

- LS10

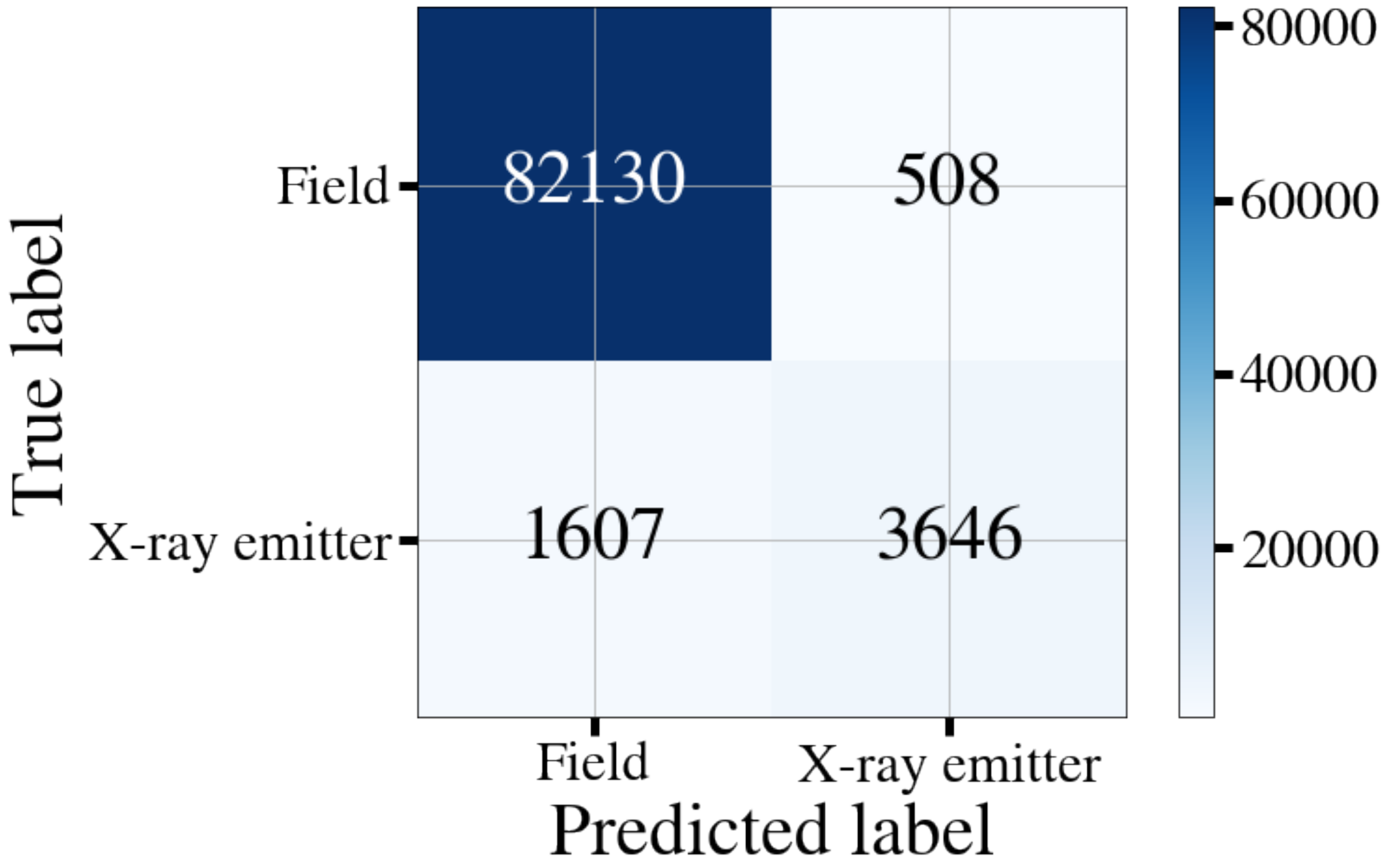
extinction-corrected fluxes
flux errors
flux ratios (colors)
shape_r','shape_e1','shape_e2','sersic',
Gaia photometry
Gaia S/N
Gaia proper motion and error
Gaia parallax and error

- 87.5% recall fraction
- 0.7% leakage



- Gaia EDR3

Gaia photometry
Gaia S/N
Gaia proper motion and error
Gaia parallax and error
Gaia astrometric_excess_noise
Gaia Colors
Correction for extinction

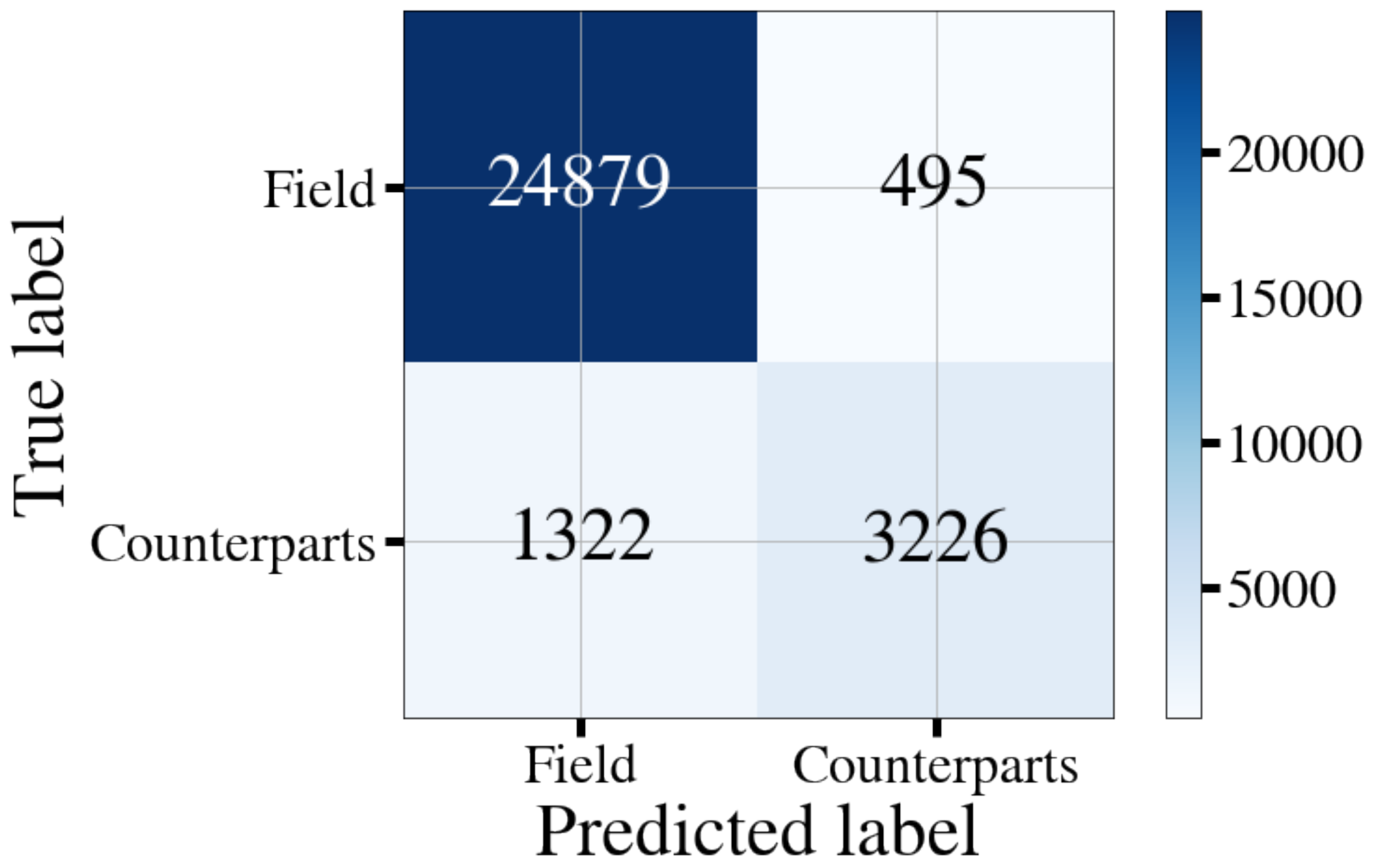


- 72% recall fraction
- 0.7% leakage

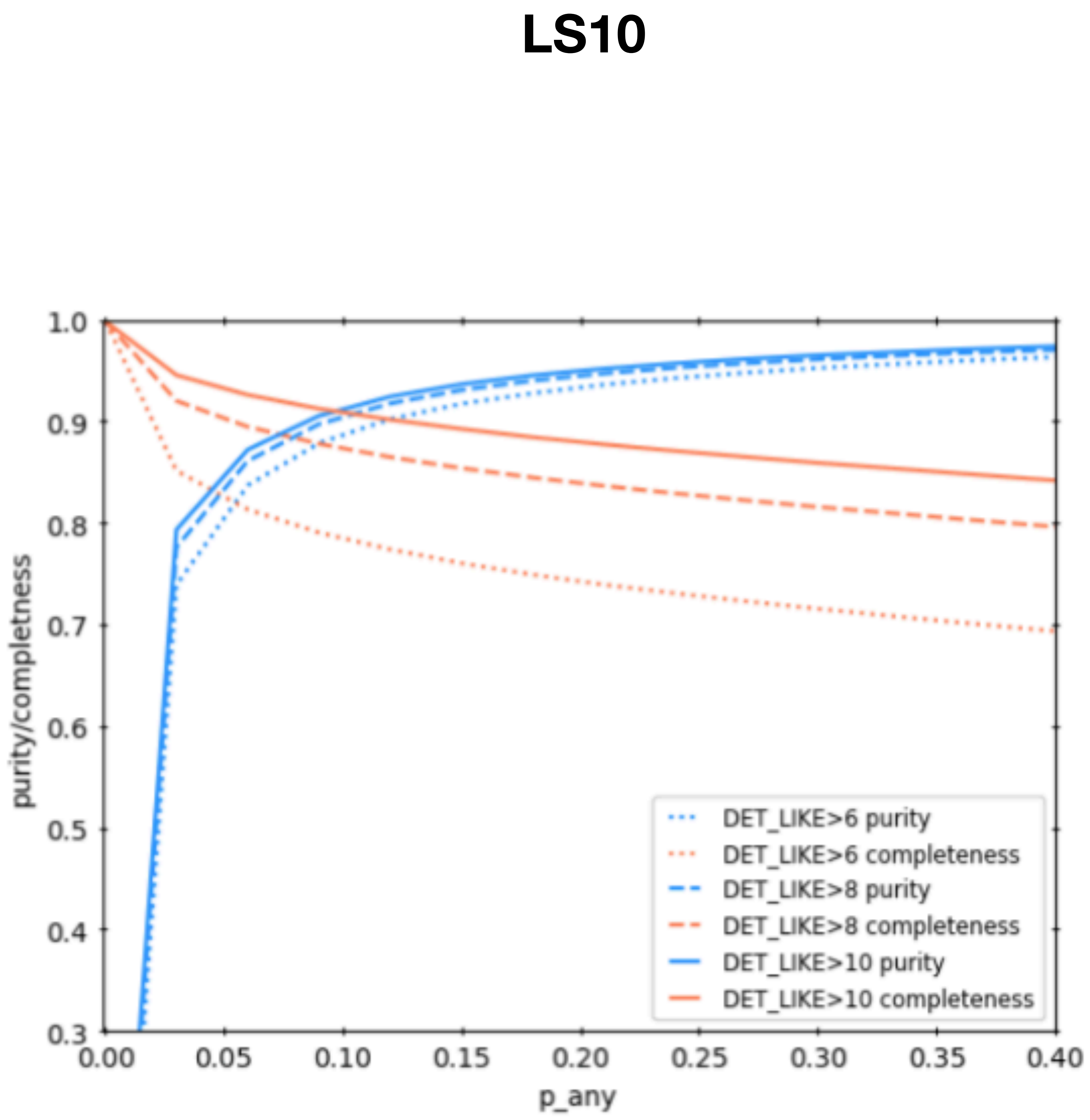
- CatWISE 2020

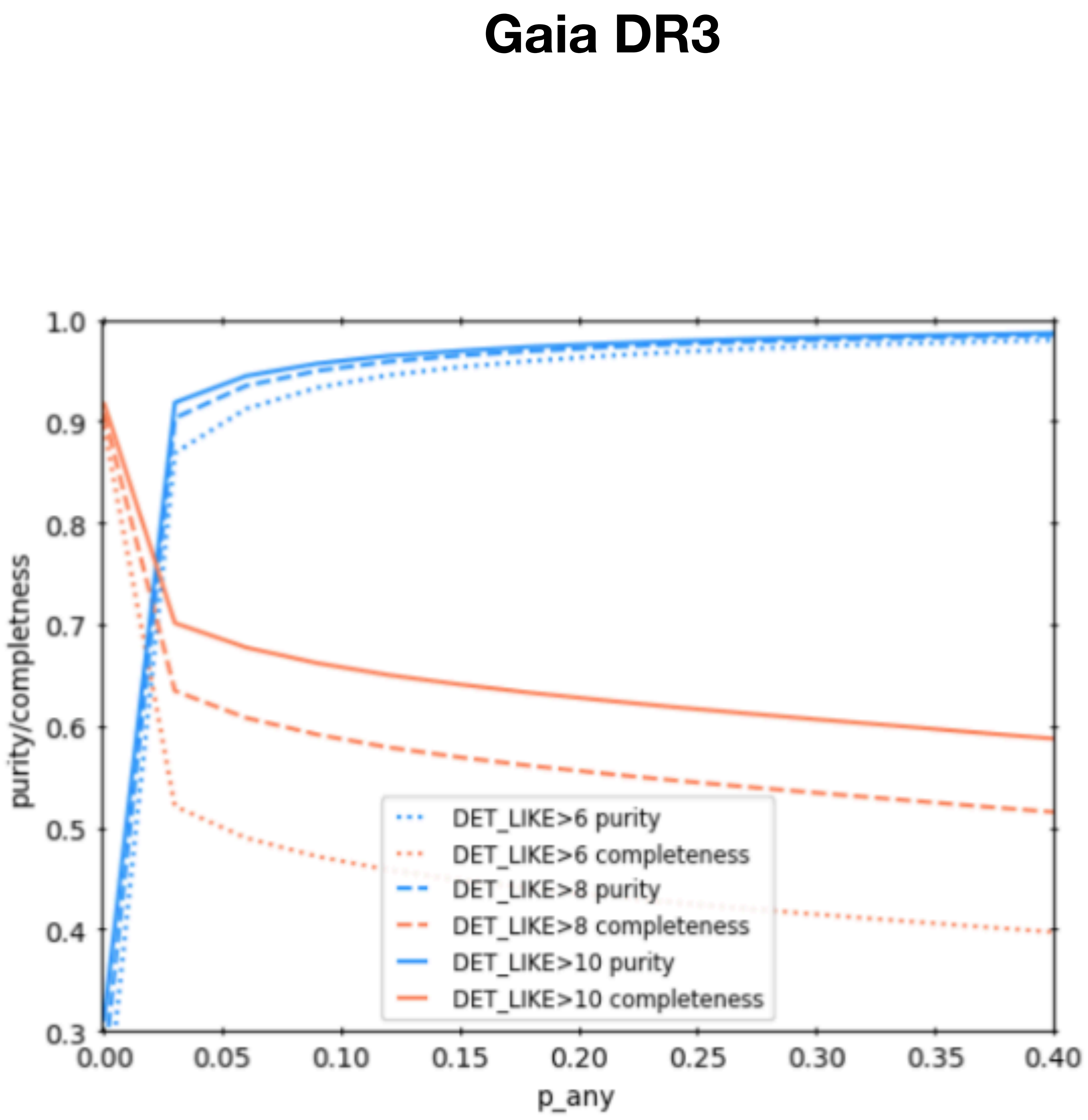
W1 and W2 fluxes
W1 an W2 aperture photometry
W1-W2 color
Flux errors
proper motion and error

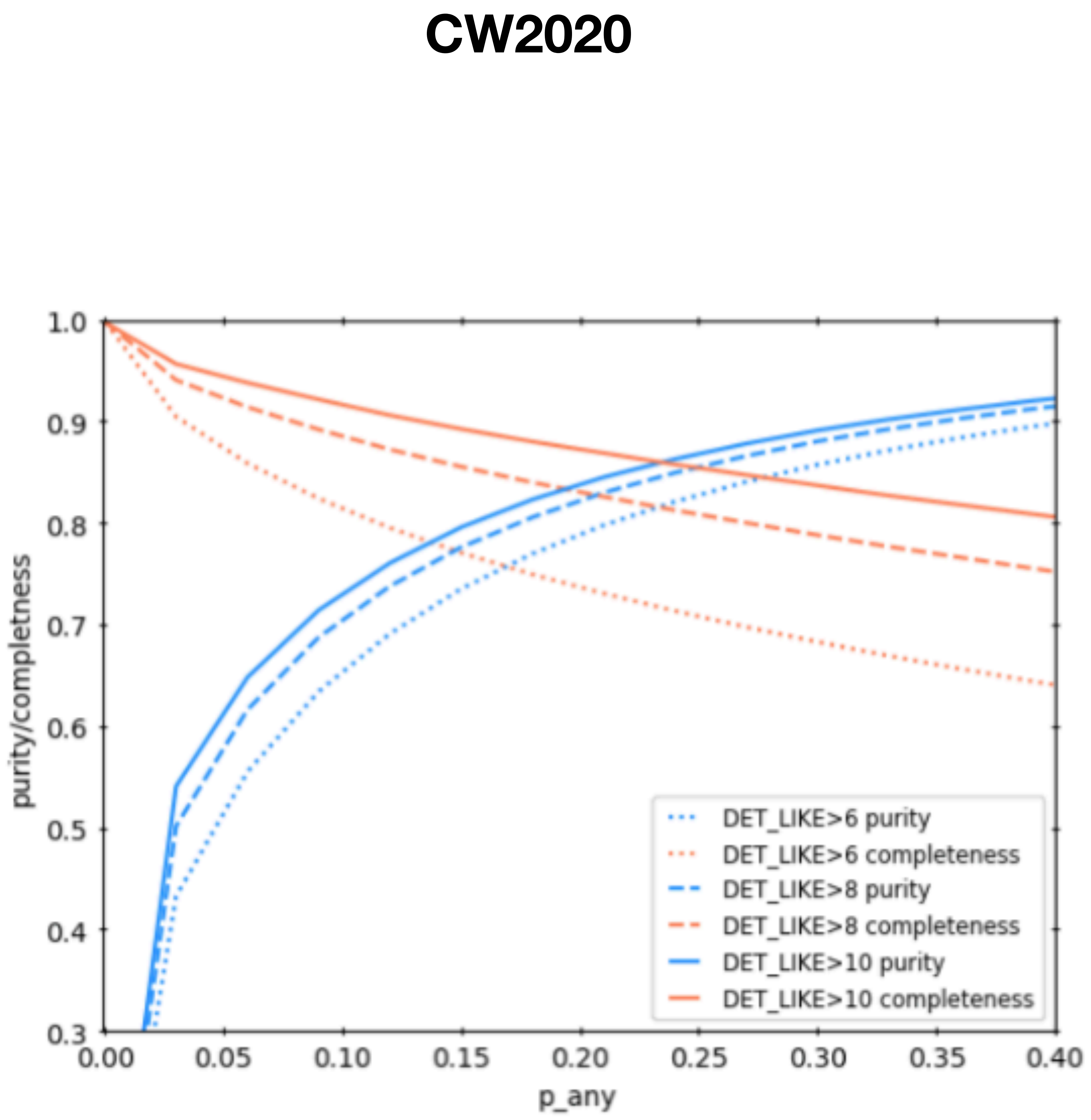
- 71% recall fraction
- 2% leakage



Averaged purity/completeness over the entire sky look good but …

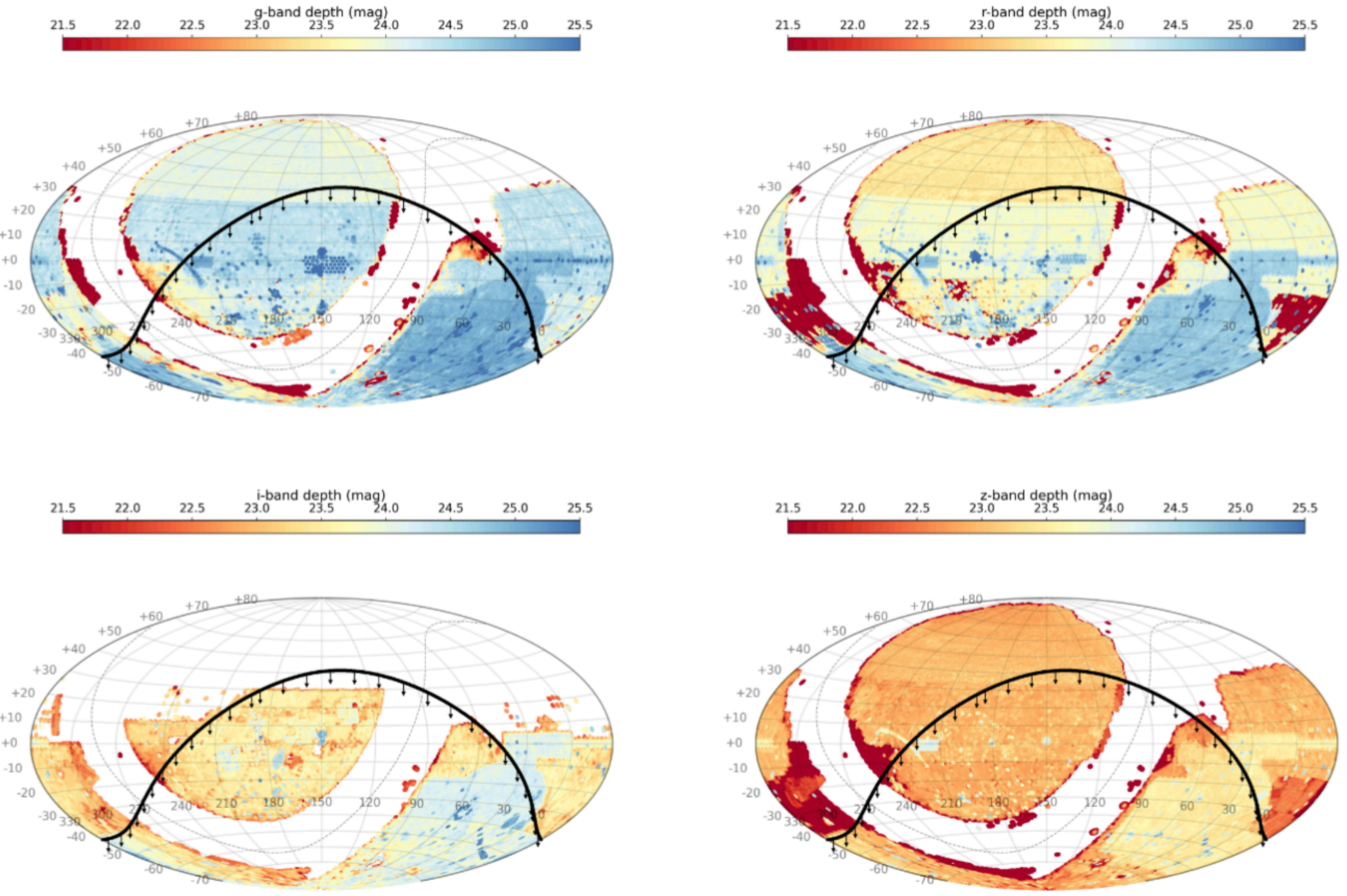


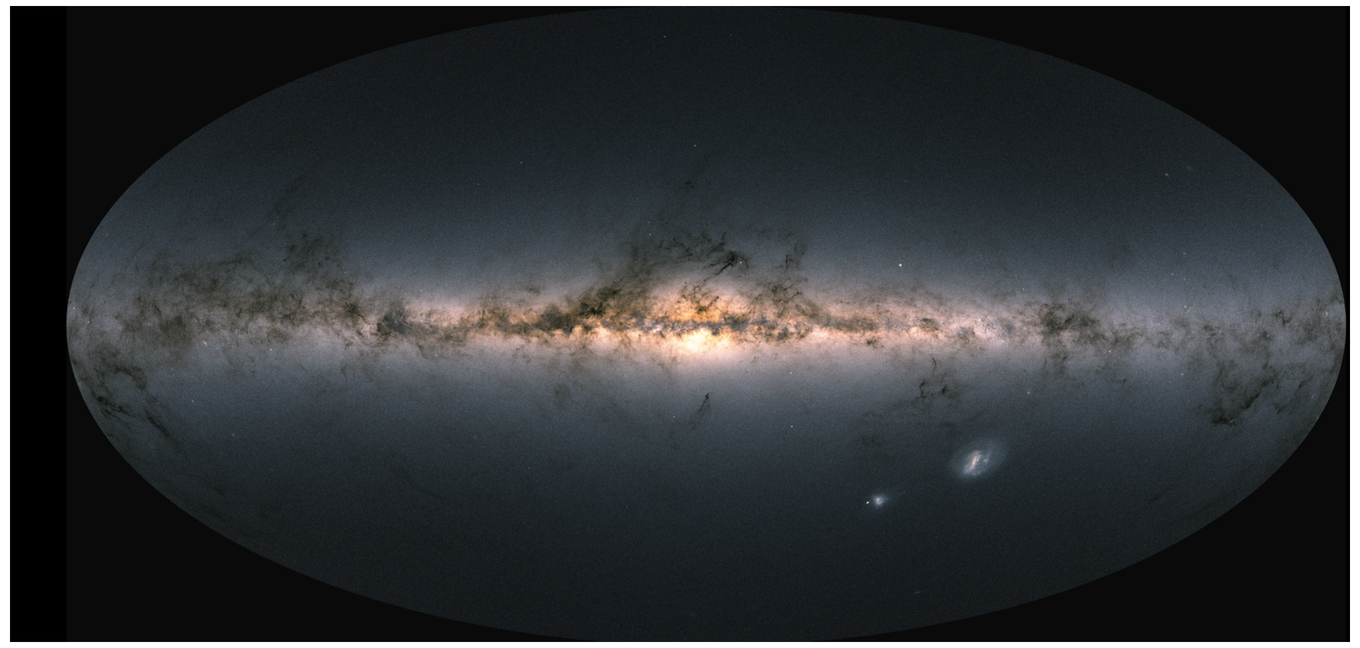


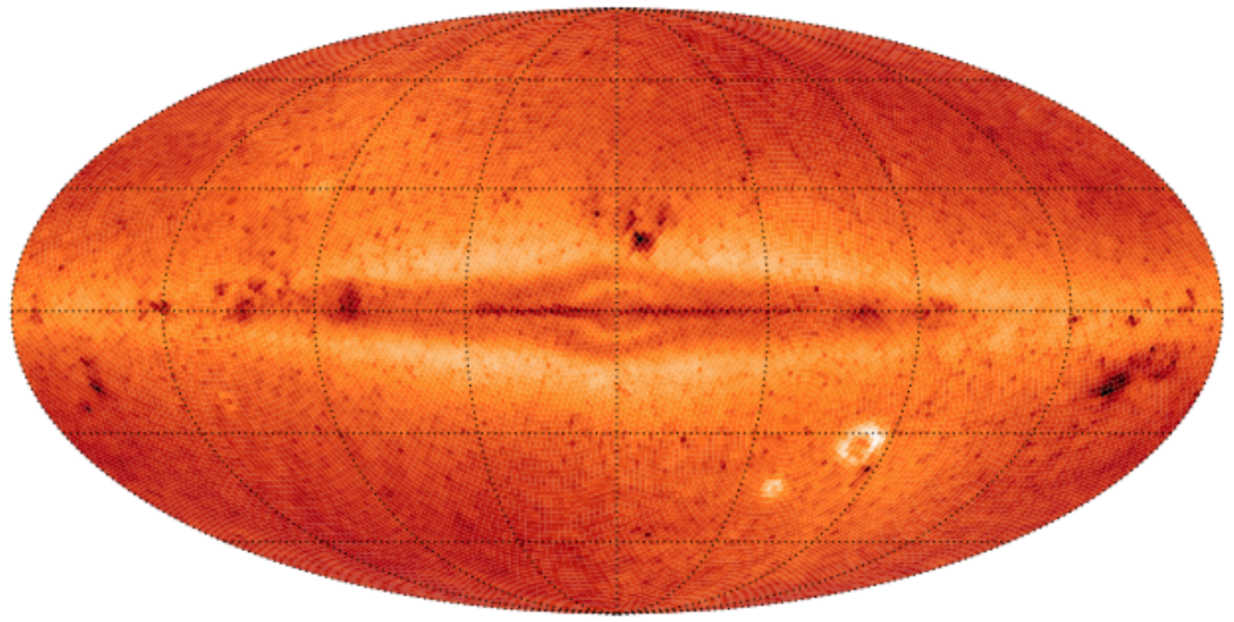


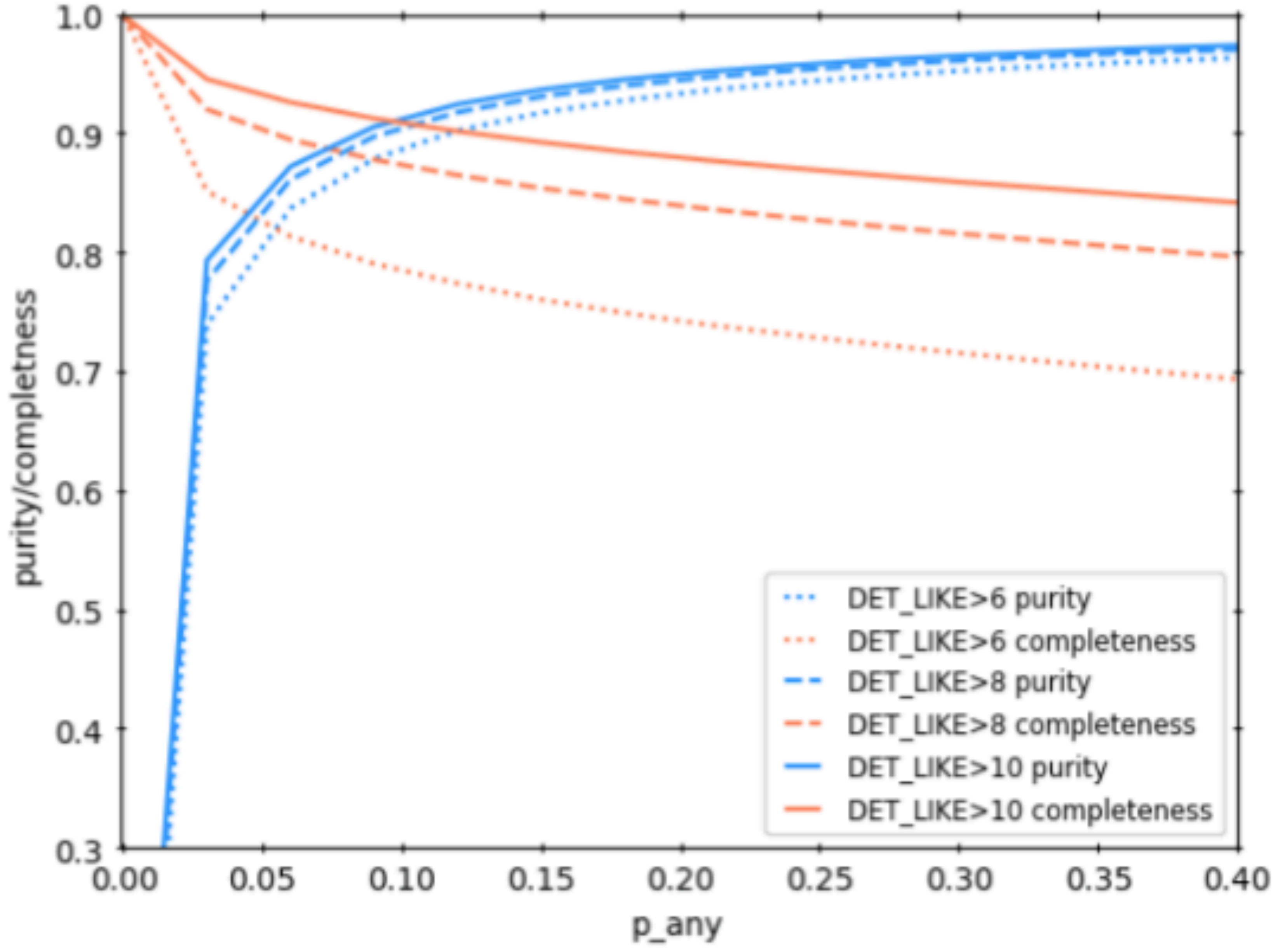
(details in Salvato+2024 in prep)

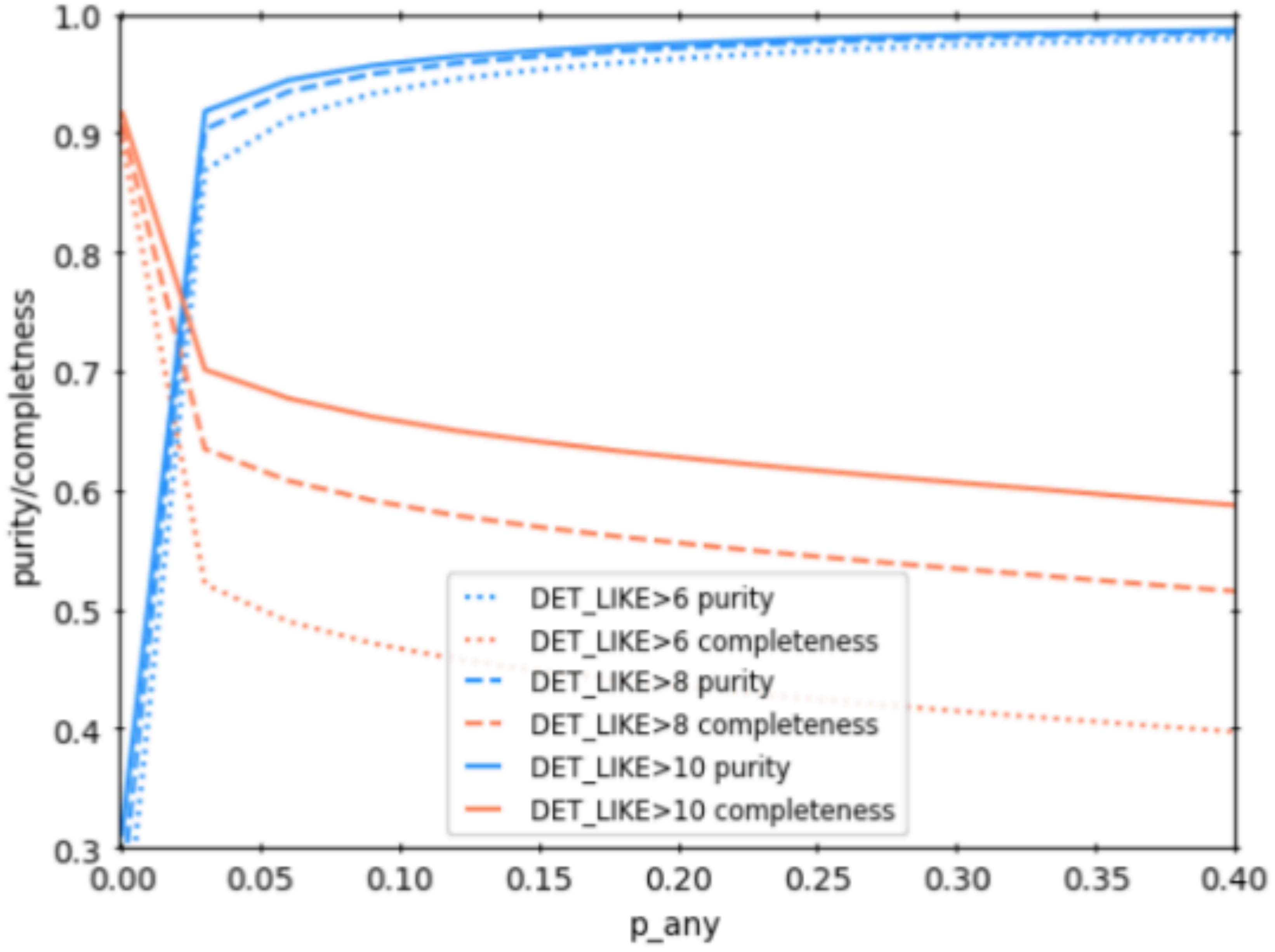
Averaged purity/completeness over the entire sky look good but …

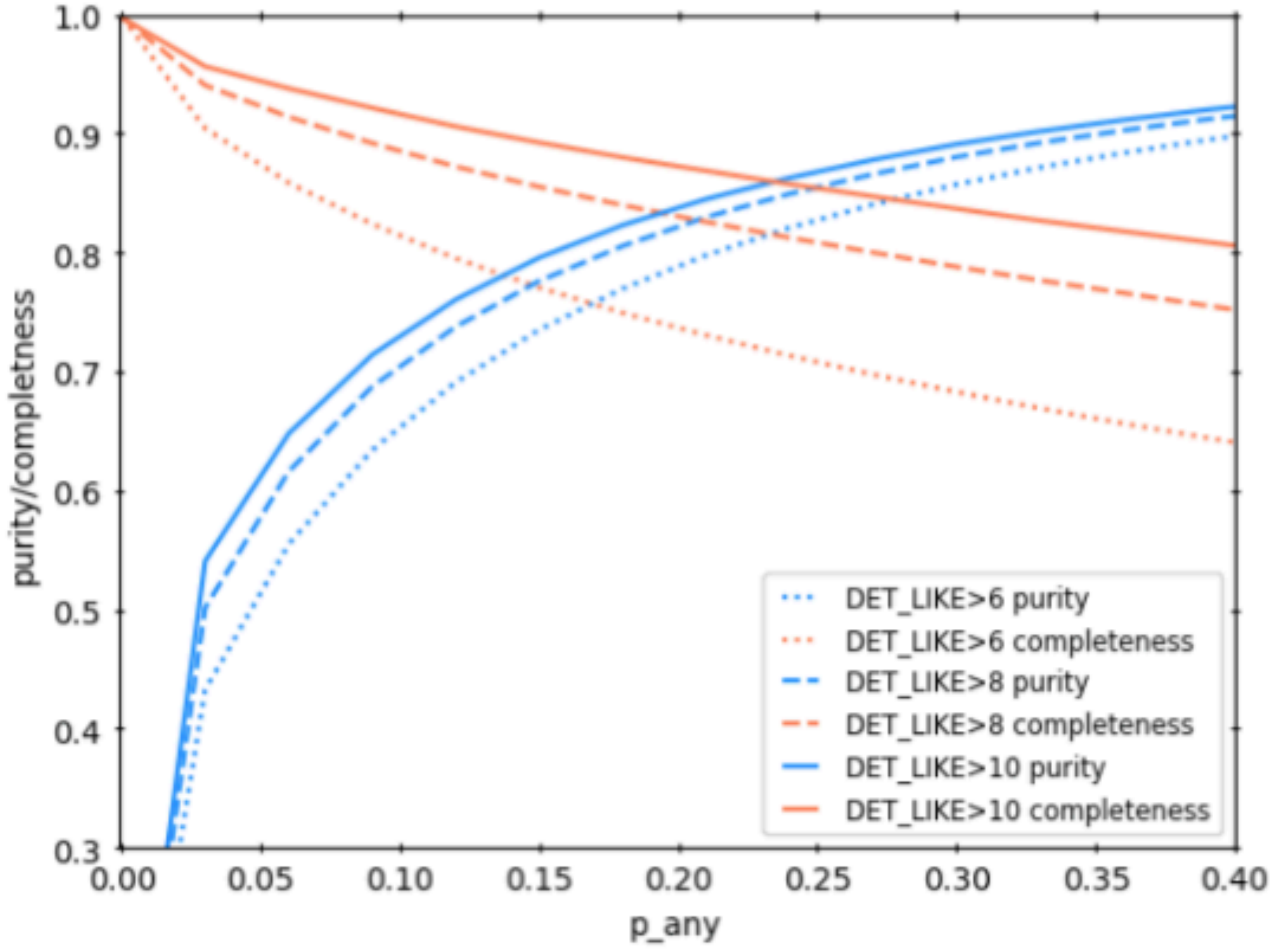






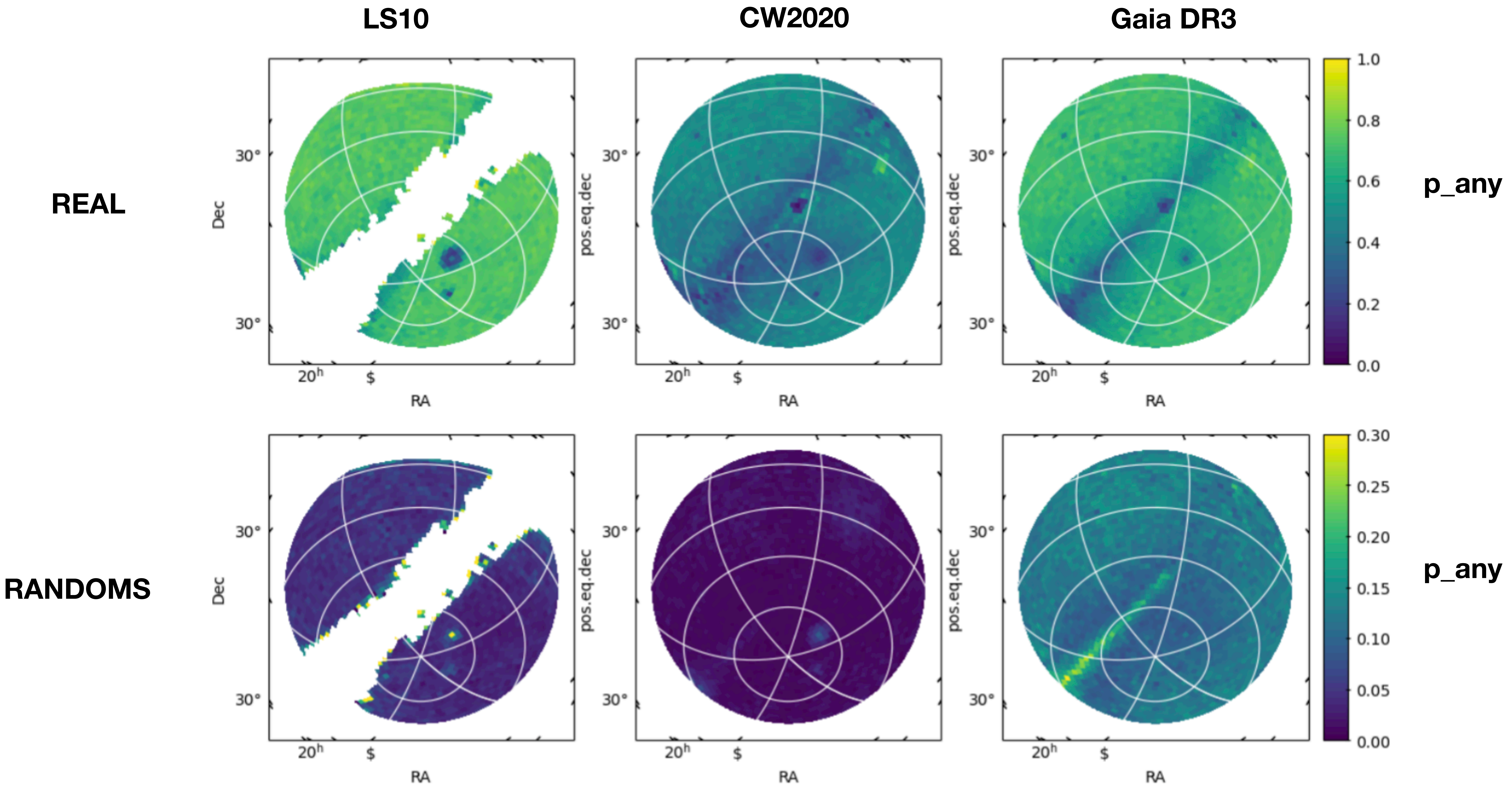






(details in Salvato+2024 in prep)

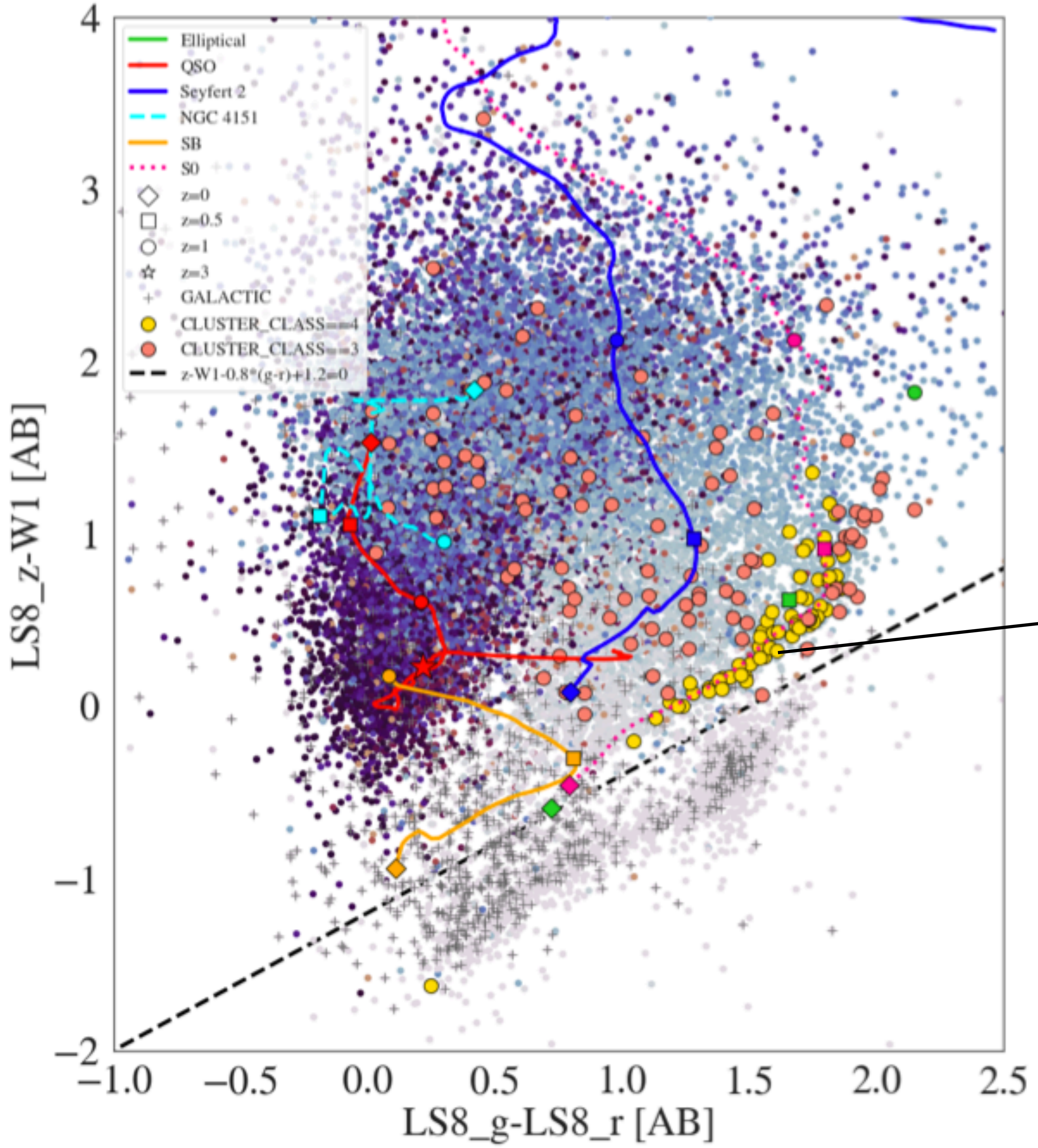
completeness and purity are coordinates depending

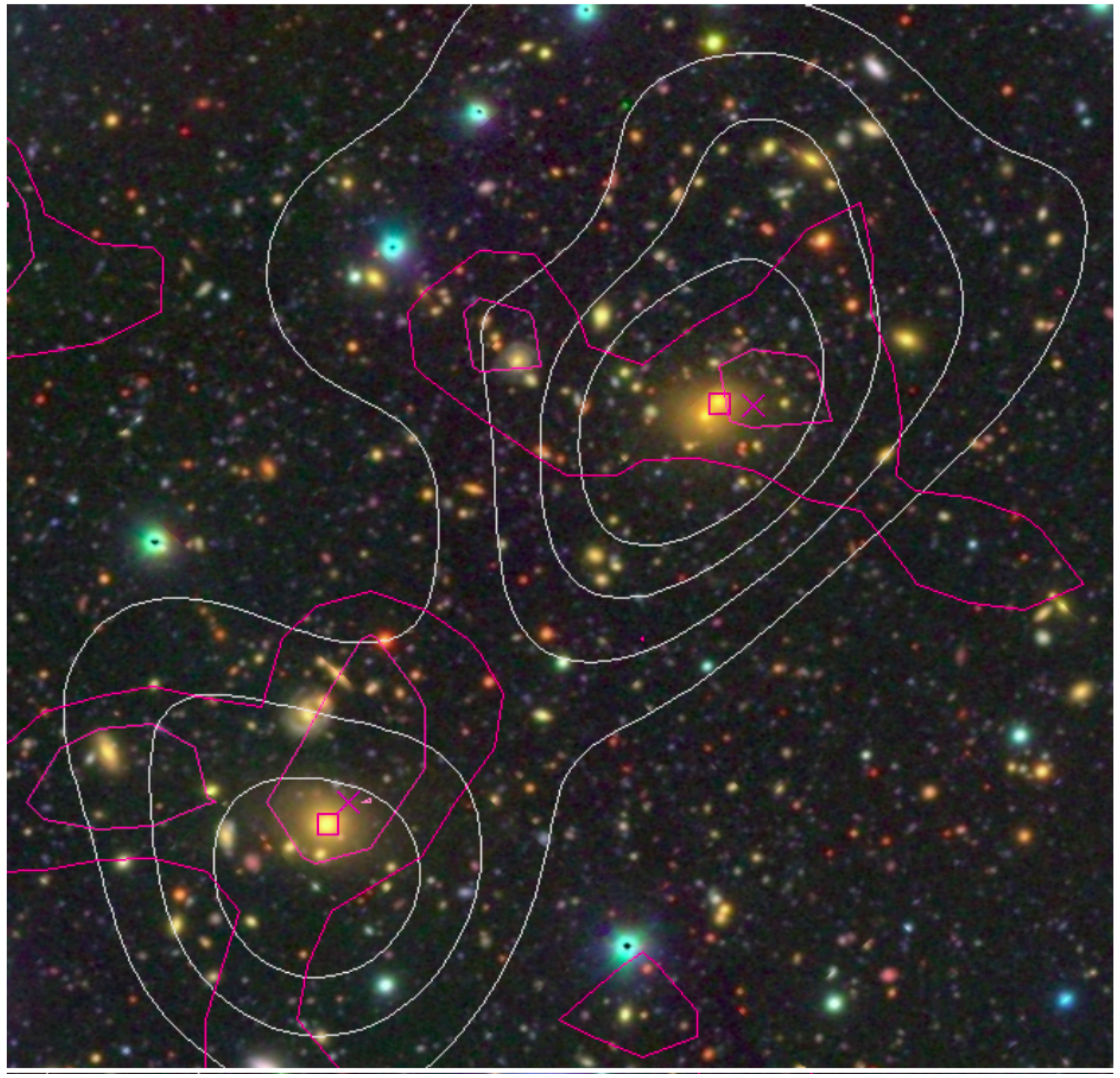


Salvato et al 2024

The probabilities actually mean something!

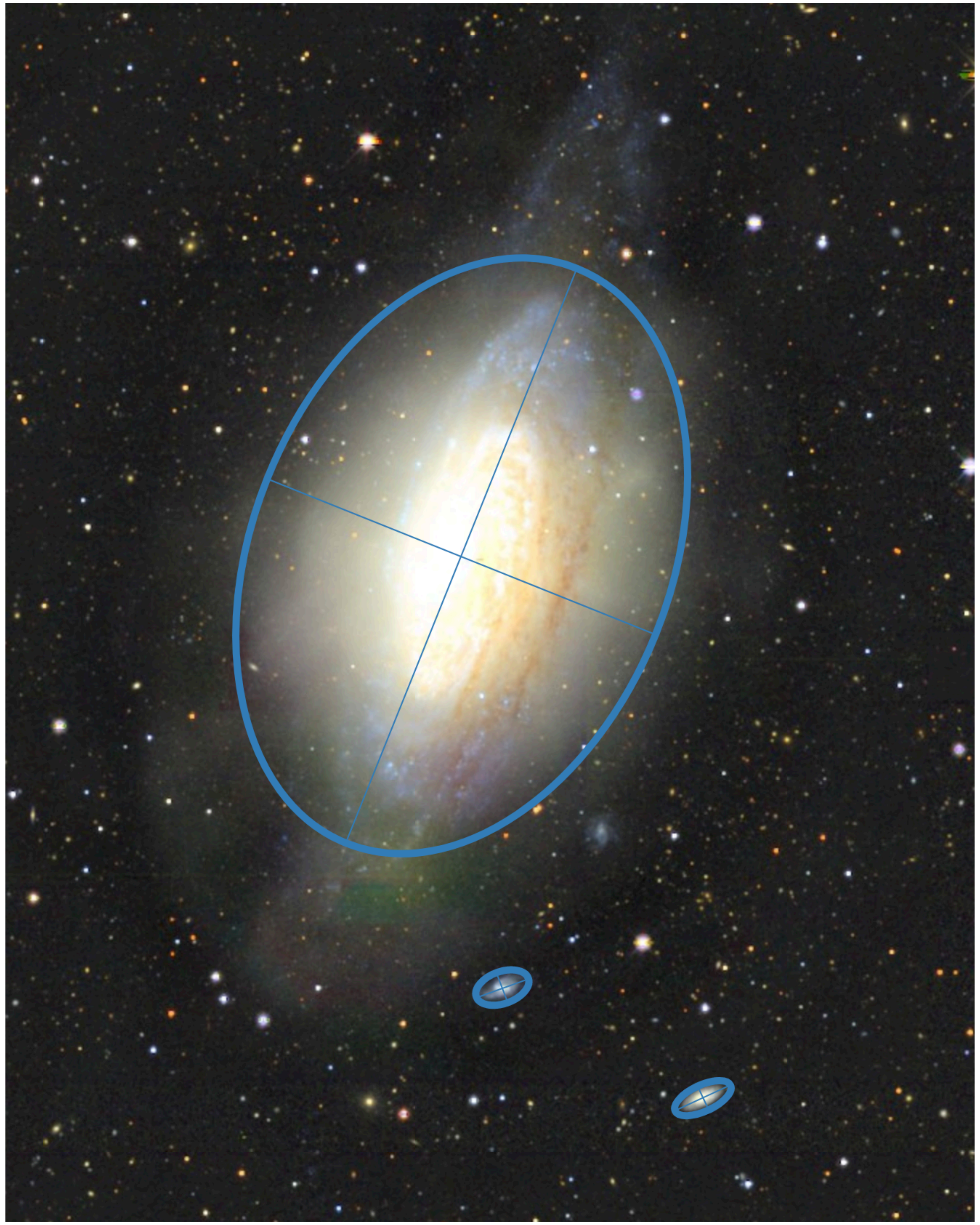
- Making first the CTPs identification and then their classification, allow pinpointing interesting populations (e.g., unresolved clusters)

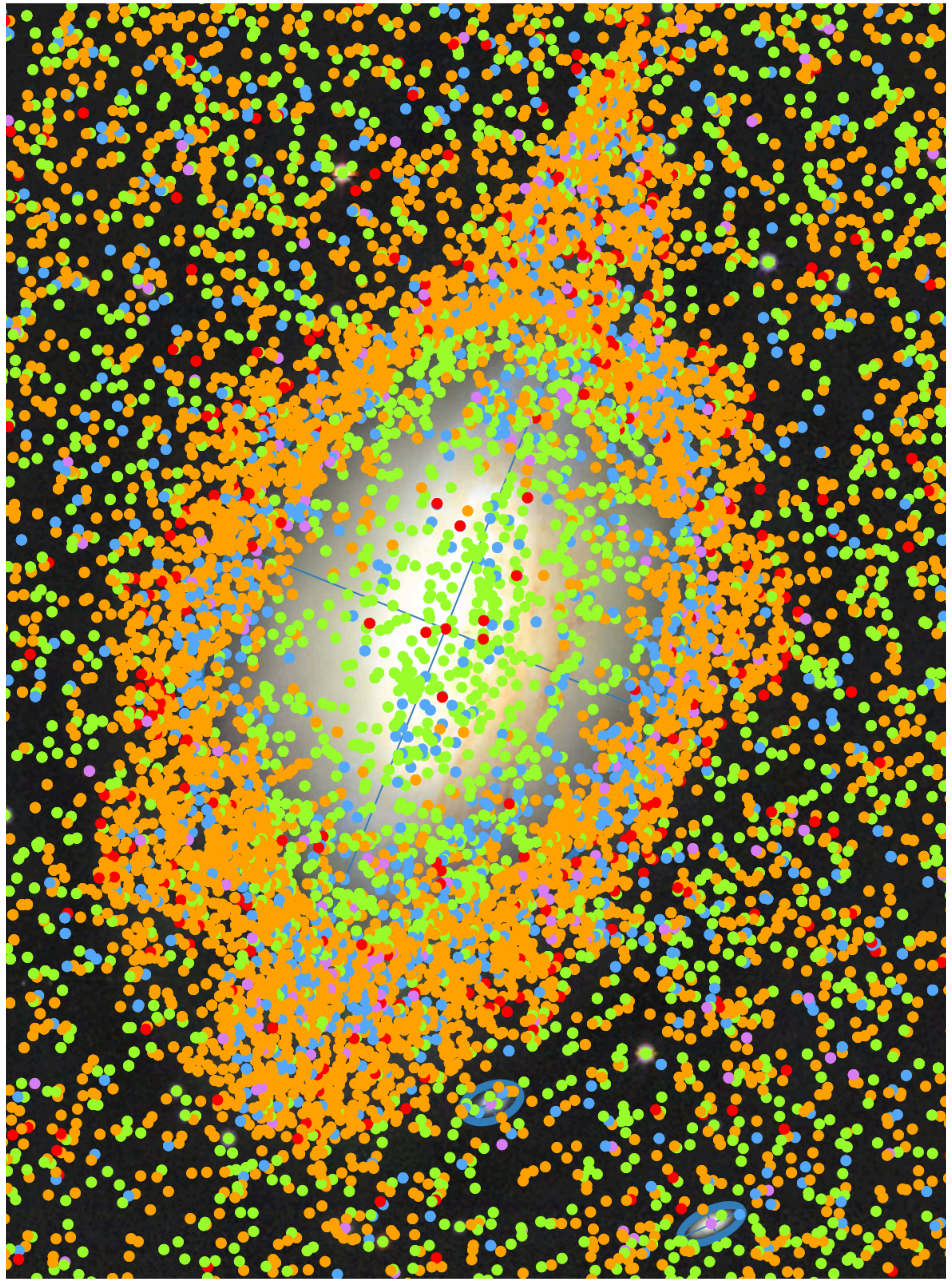




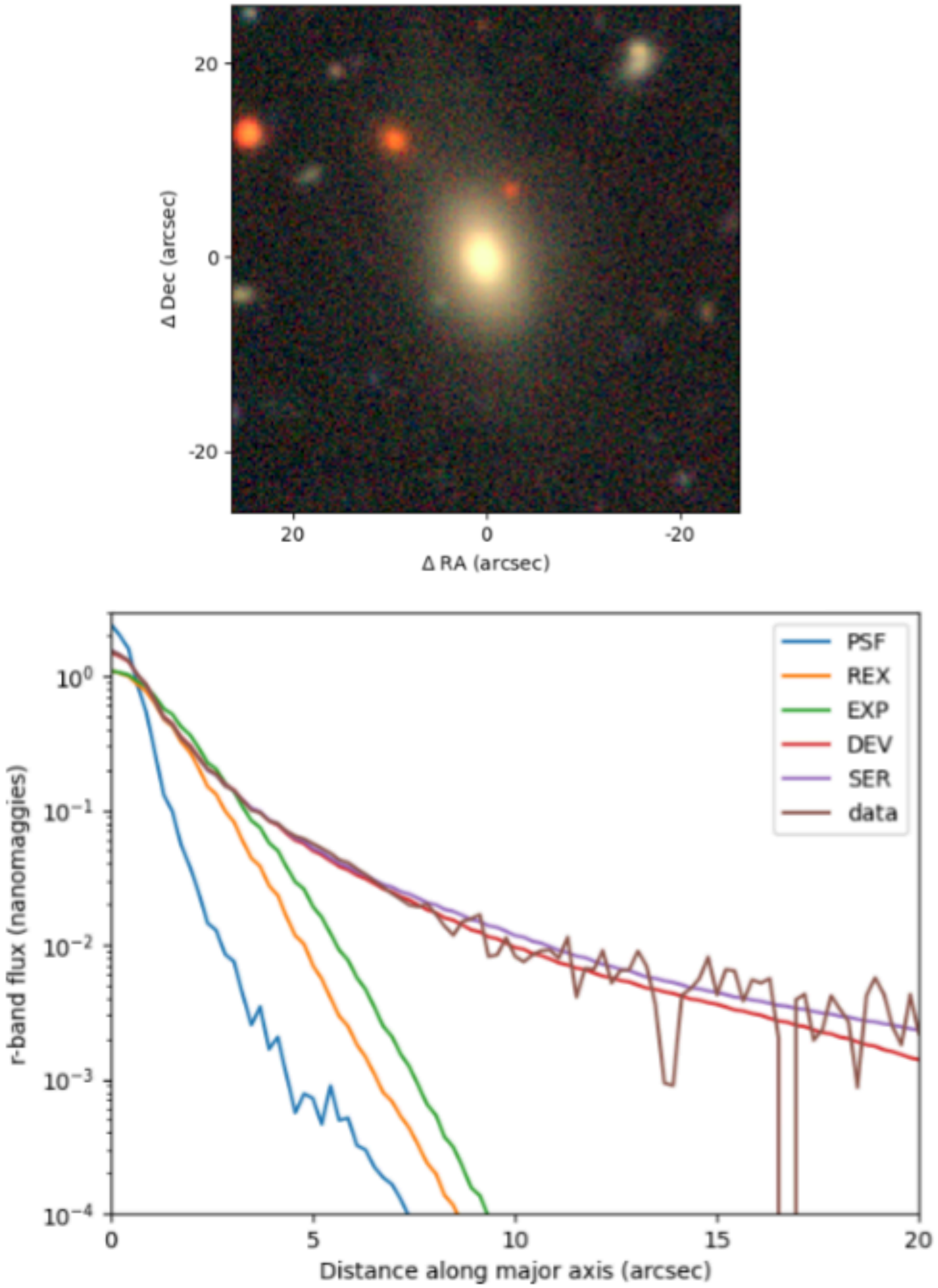
Salvato et al 2022, Bulbul et al 2022, Balzer et al 2024

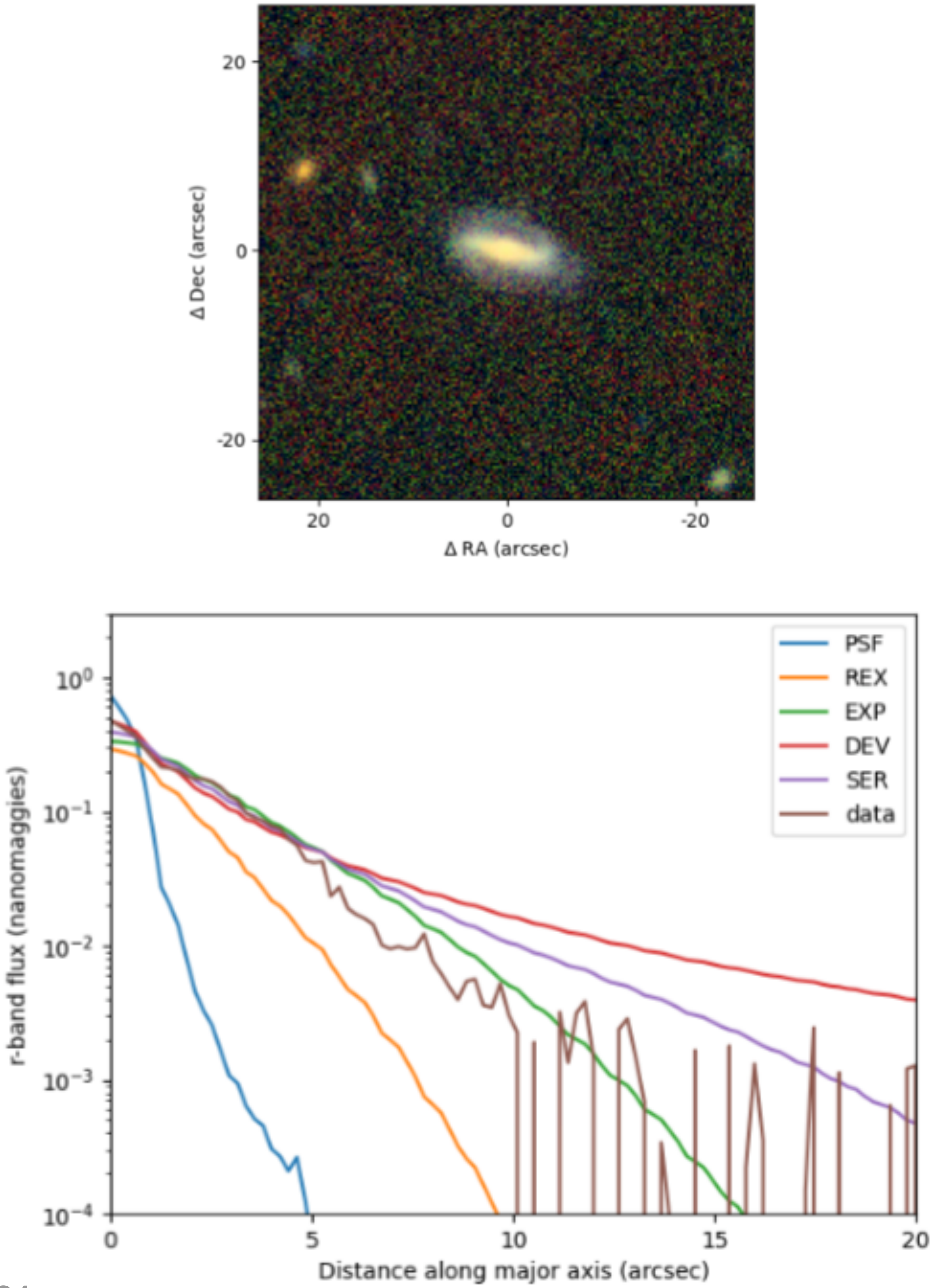
The ancillary data that we are using were not created with our needs in mind !



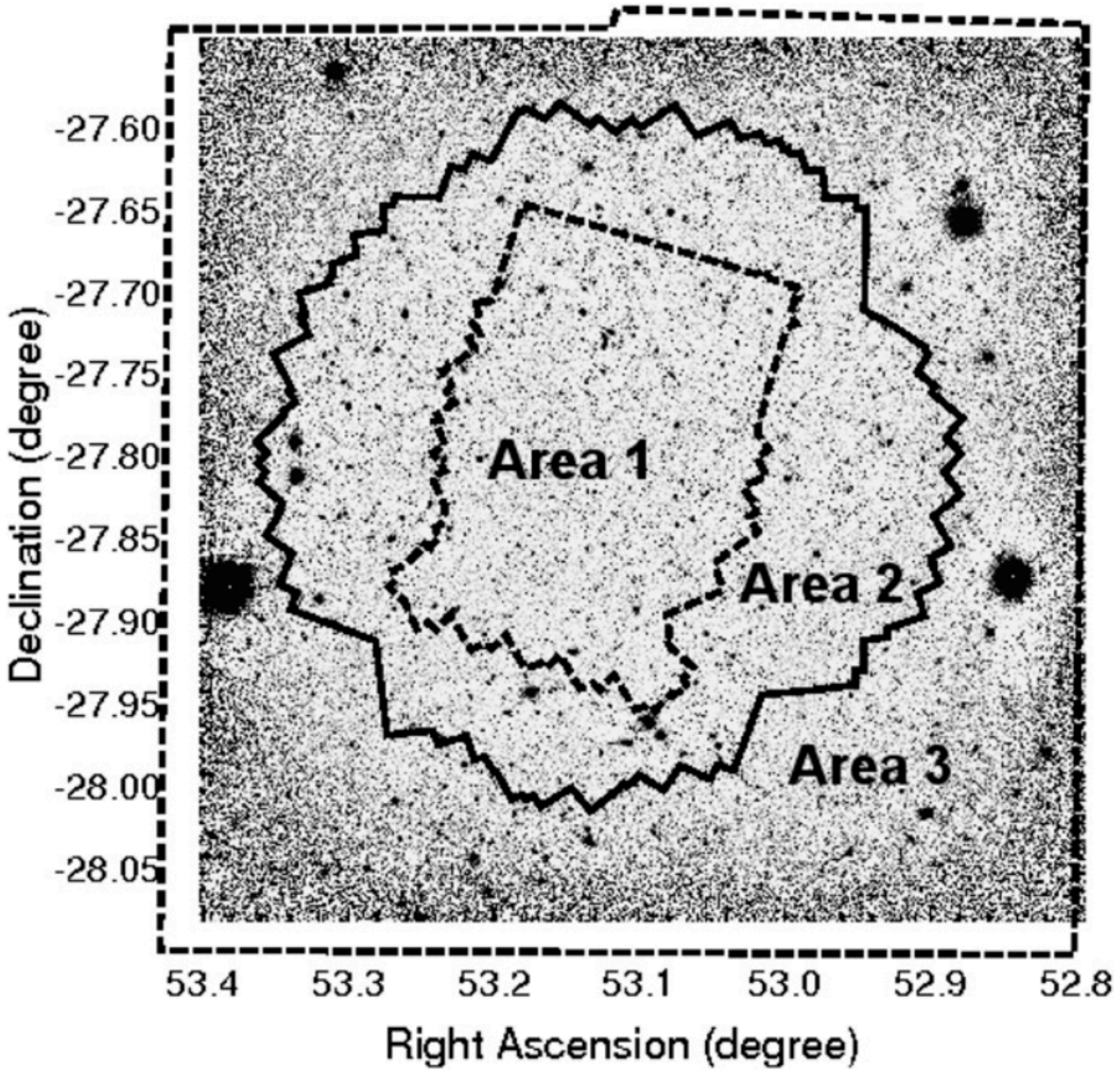


The ancillary data that we are using were not created with our needs in mind : Model_flux does not work for AGN





Comparison between assumptions



Hsu et al 2014

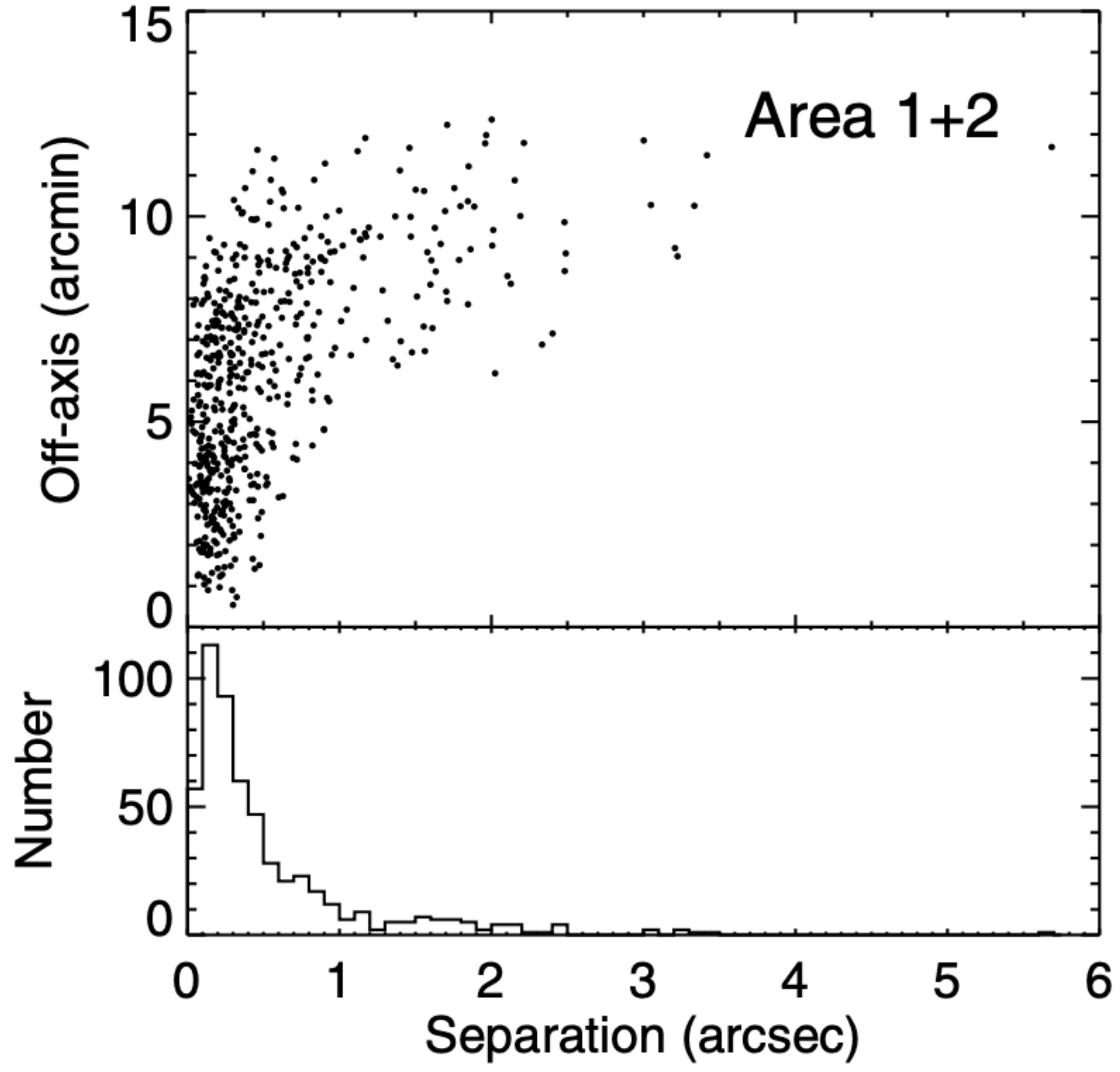


Figure 3. Coordinate differences between the X11 and R13 X-ray catalogs. The lower panel shows a histogram of offsets for the 545 sources that Areas 1 and 2 have in common in the two catalogs. The upper panel shows the off-axis angle from the *Chandra* aim point as a function of the angular offset.

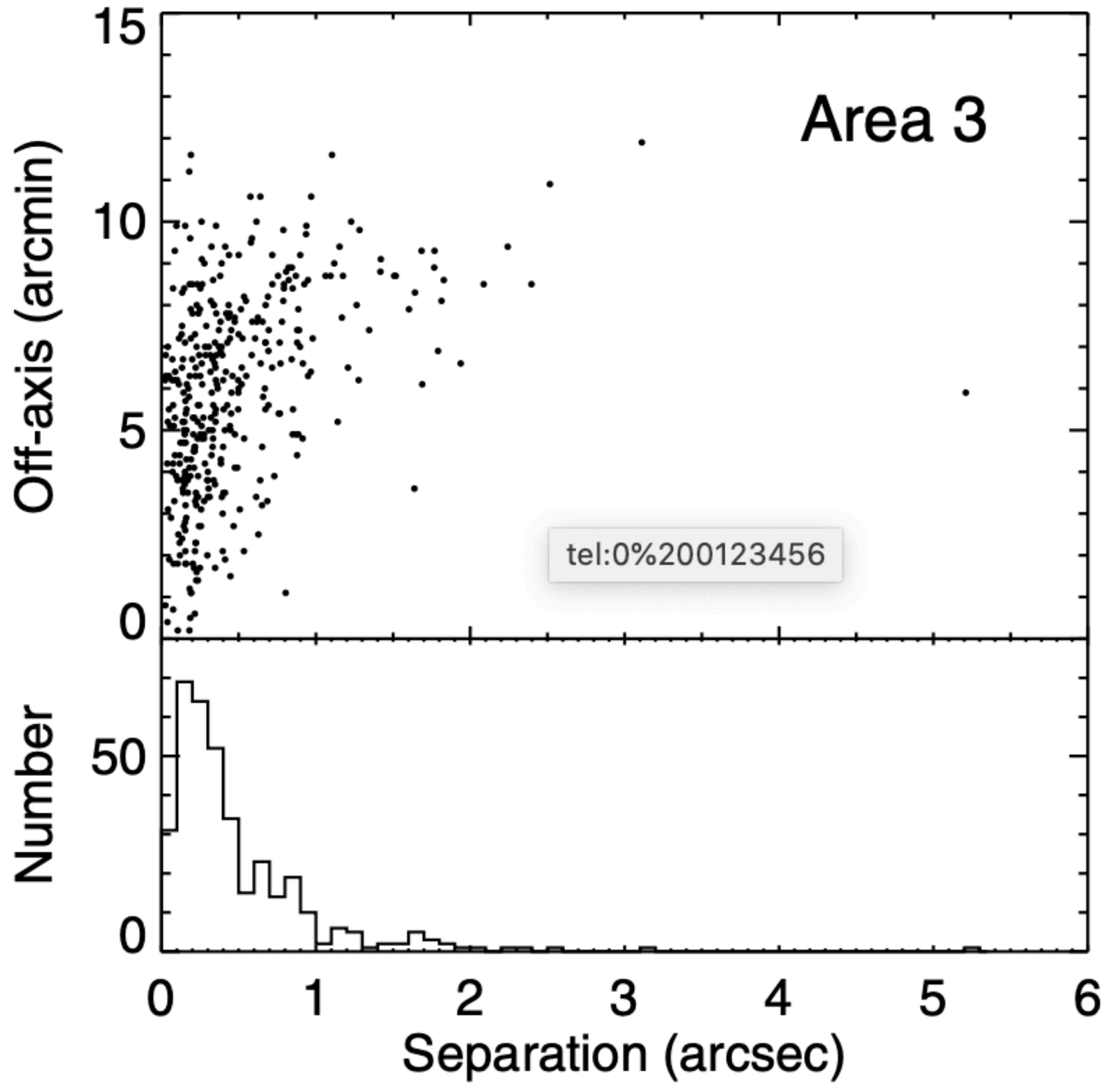
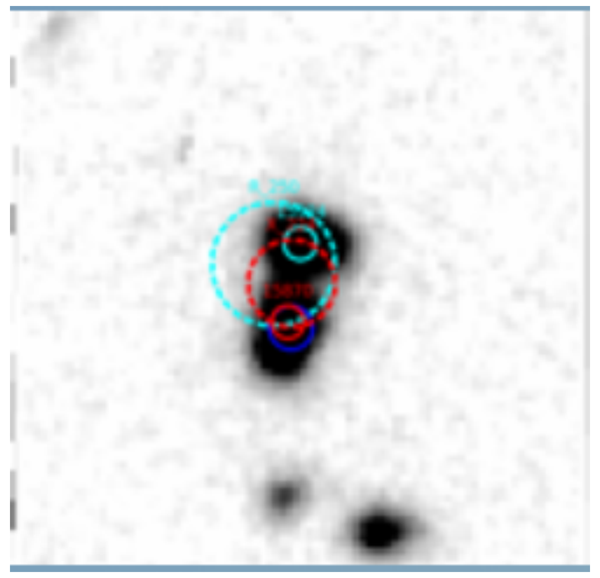
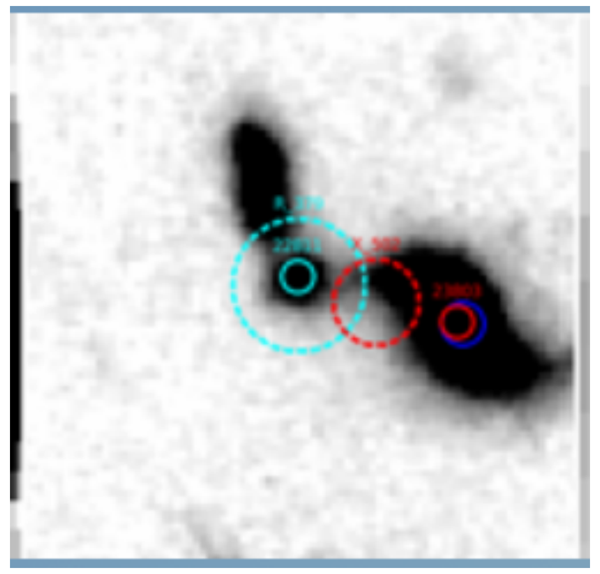


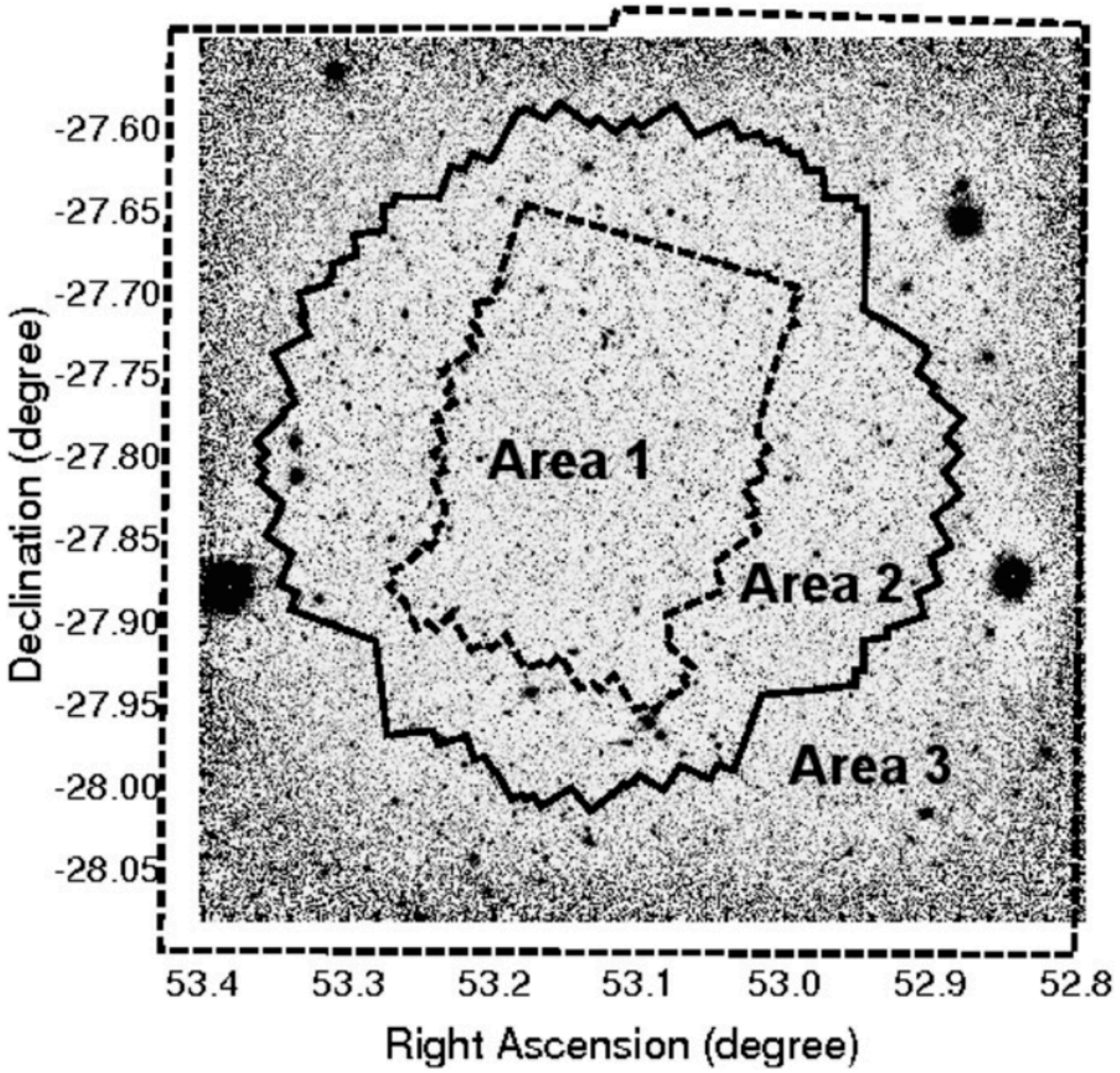
Figure 4. Coordinate differences between L05 and V06 X-ray catalogs. The lower panel shows a histogram of offsets for the 495 sources in Area 3 that are in common in the two catalogs. The upper panel shows the off-axis angle from the *Chandra* aim point as a function of the angular offset.





Comparison between assumptions

- The X-ray coordinates and positional errors depend on how the X-ray data are treated.



Hsu et al 2014

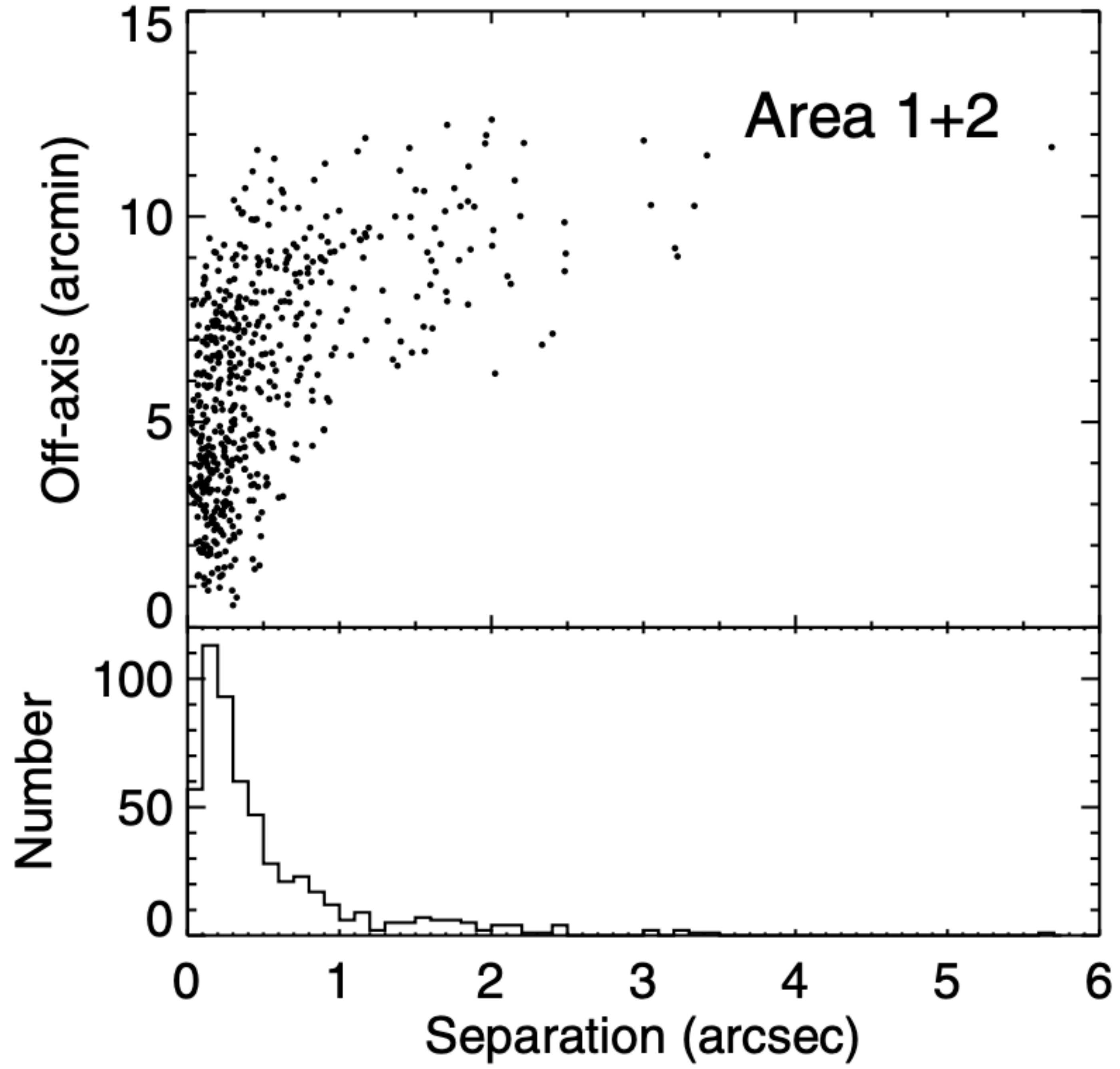


Figure 3. Coordinate differences between the X11 and R13 X-ray catalogs. The lower panel shows a histogram of offsets for the 545 sources that Areas 1 and 2 have in common in the two catalogs. The upper panel shows the off-axis angle from the *Chandra* aim point as a function of the angular offset.

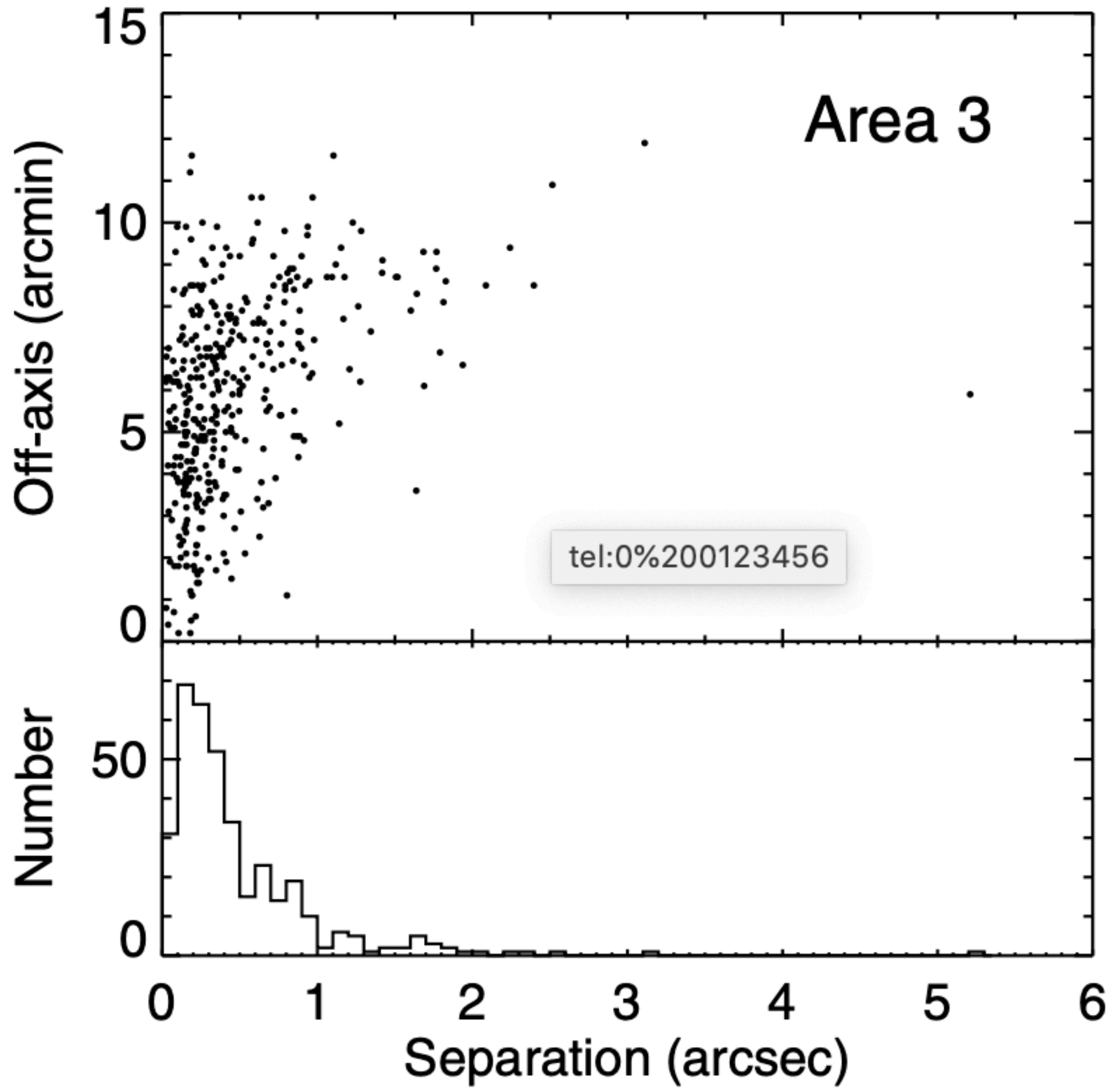
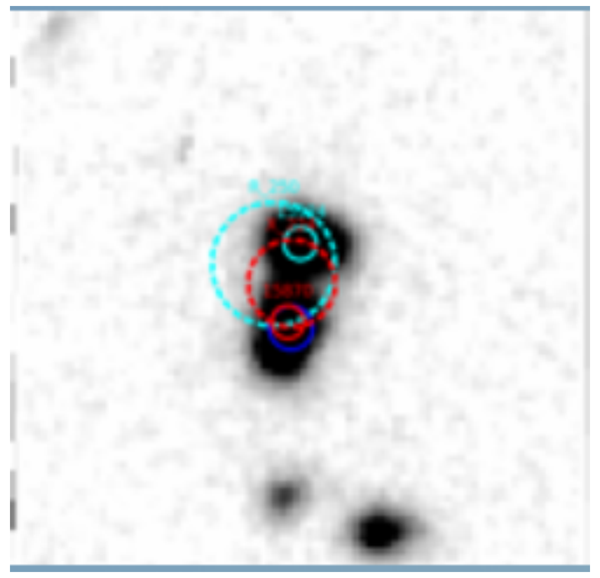
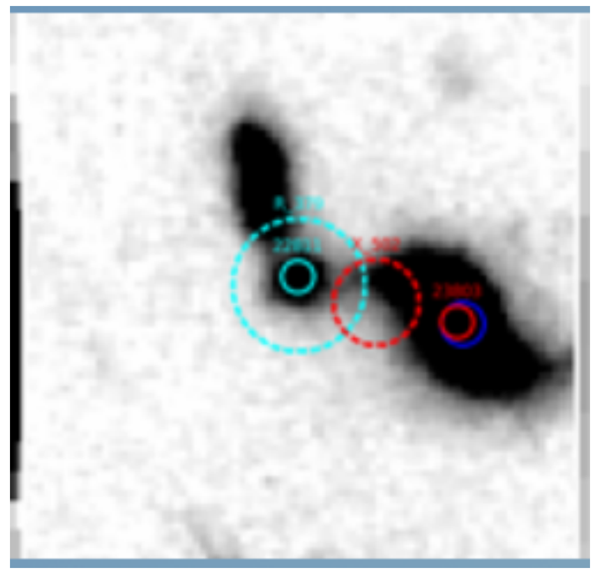
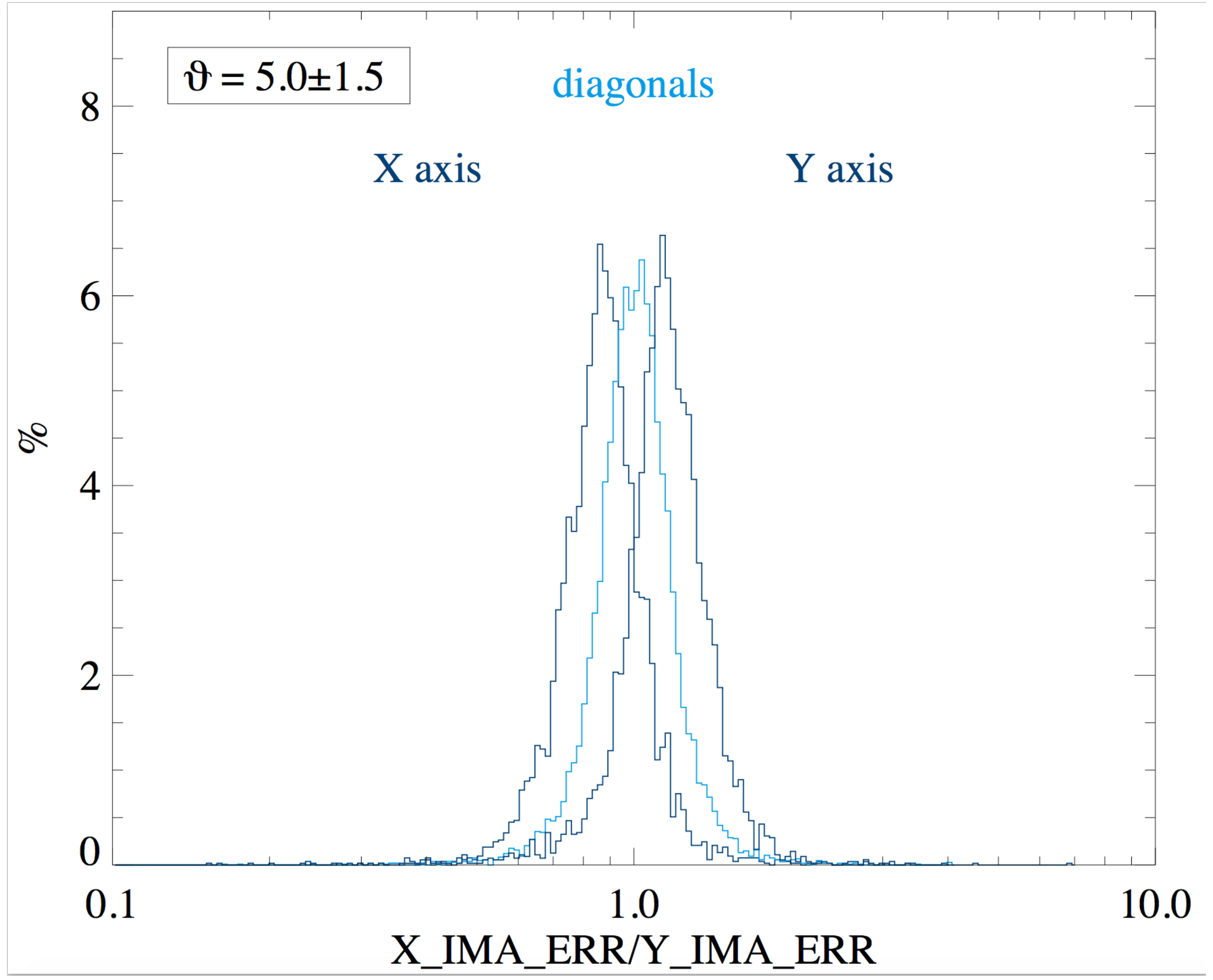


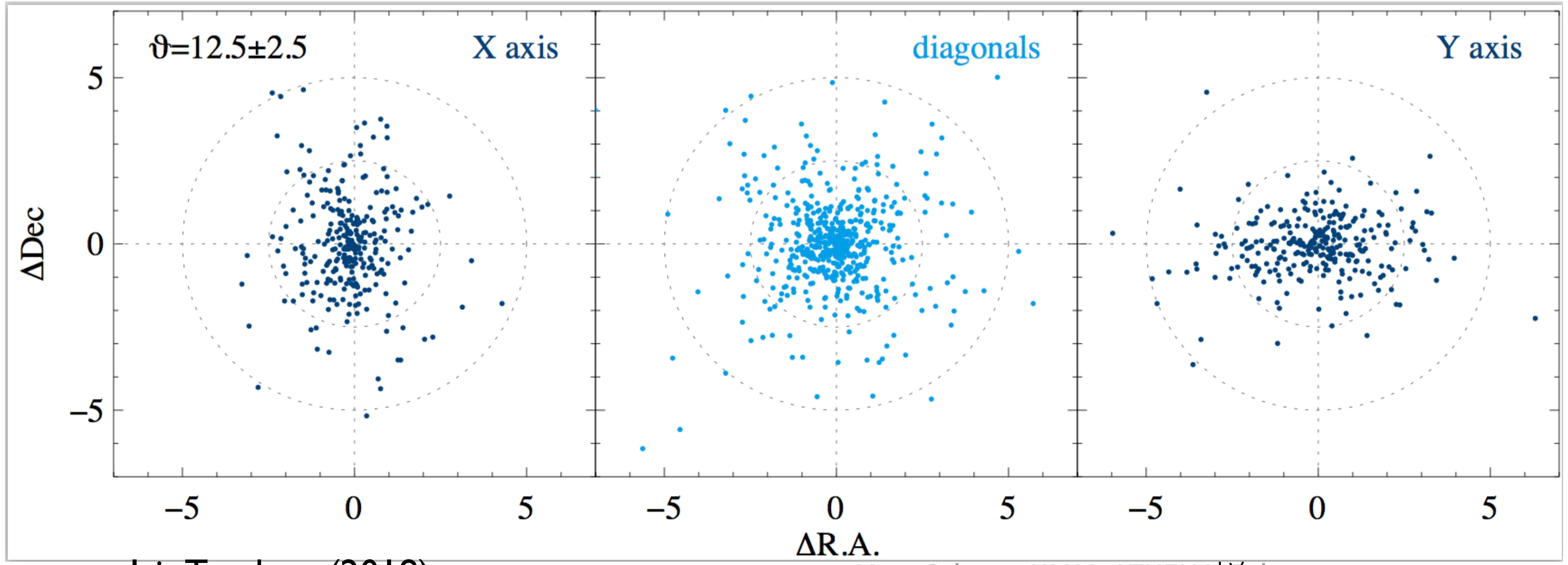
Figure 4. Coordinate differences between L05 and V06 X-ray catalogs. The lower panel shows a histogram of offsets for the 495 sources in Area 3 that are in common in the two catalogs. The upper panel shows the off-axis angle from the *Chandra* aim point as a function of the angular offset.





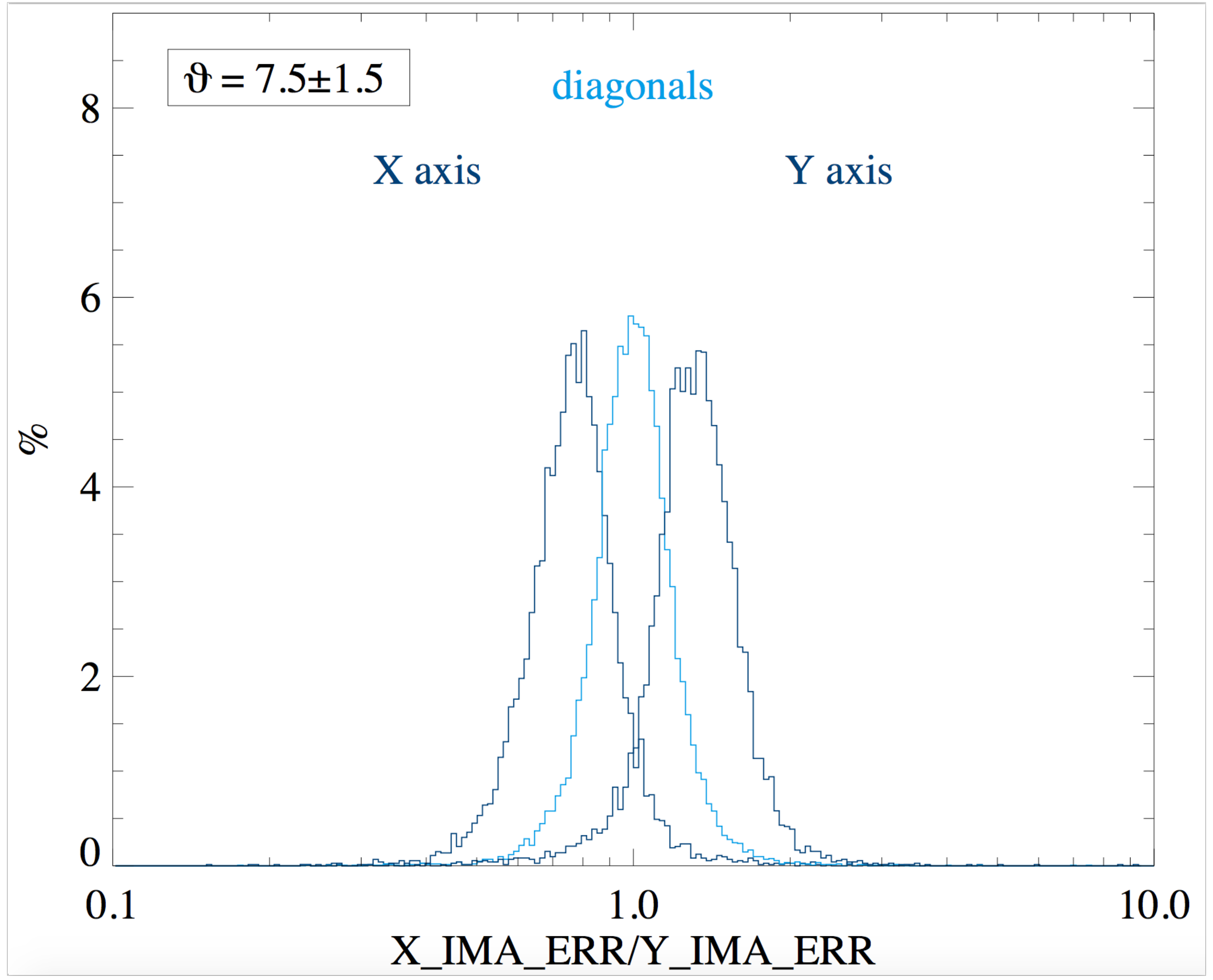
Don’t look for the CTP where it is not

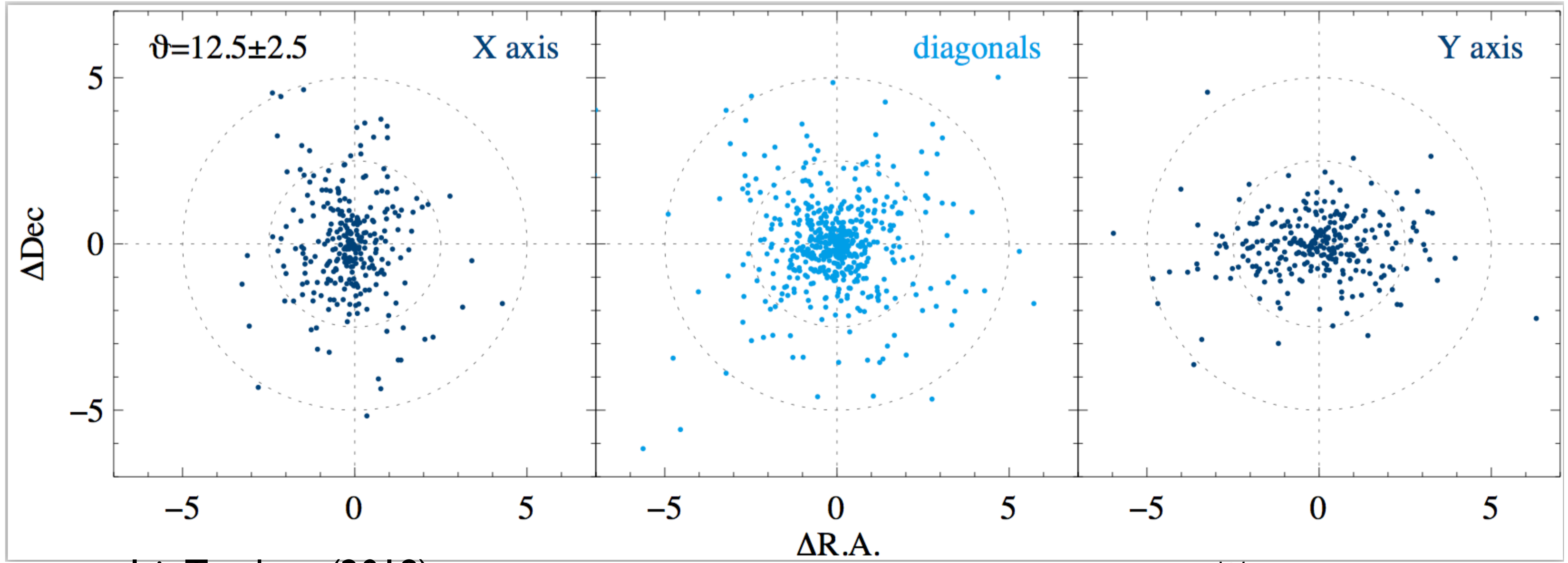




Courtesy: Iris Traulsen (2018)

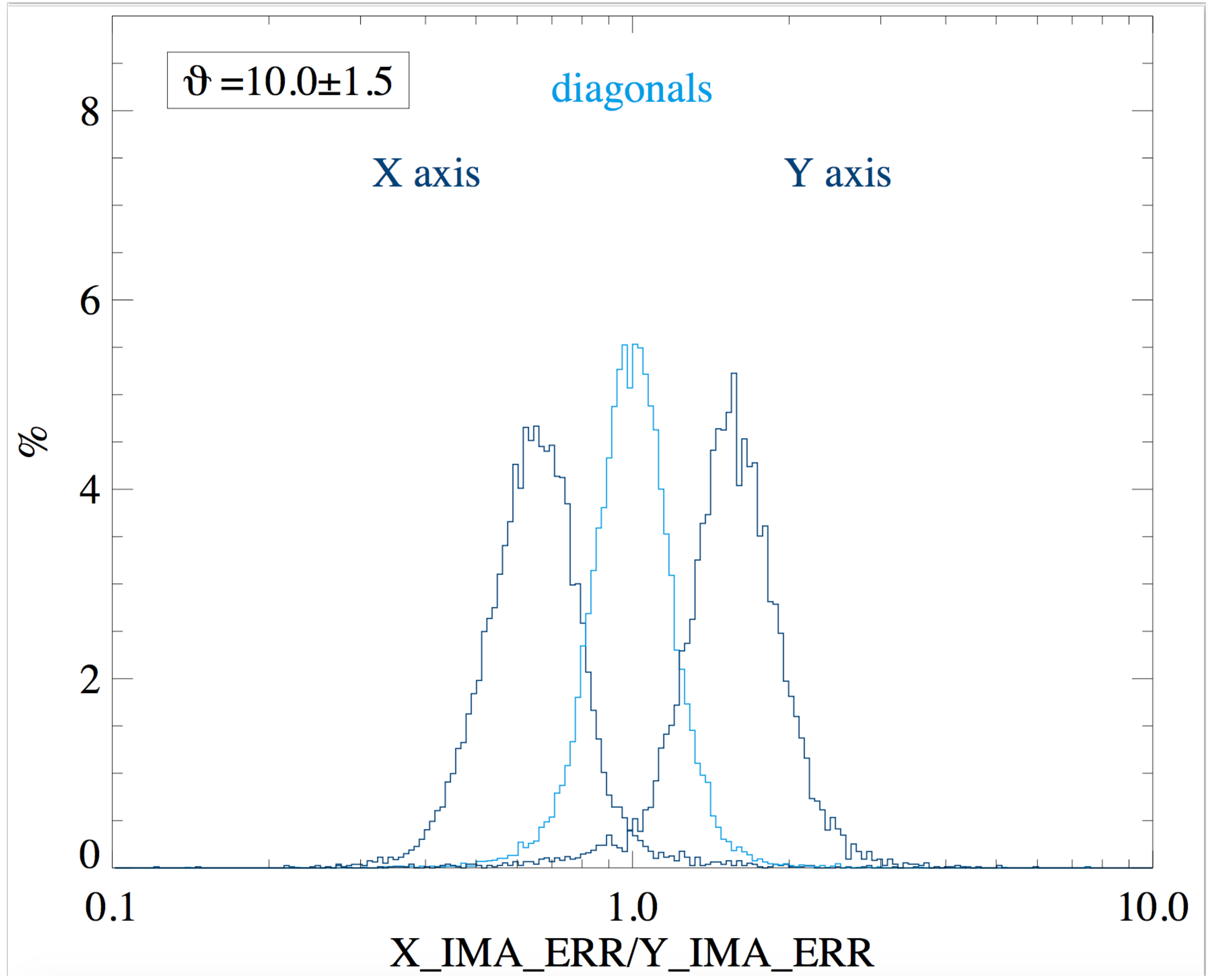
Don’t look for the CTP where it is not

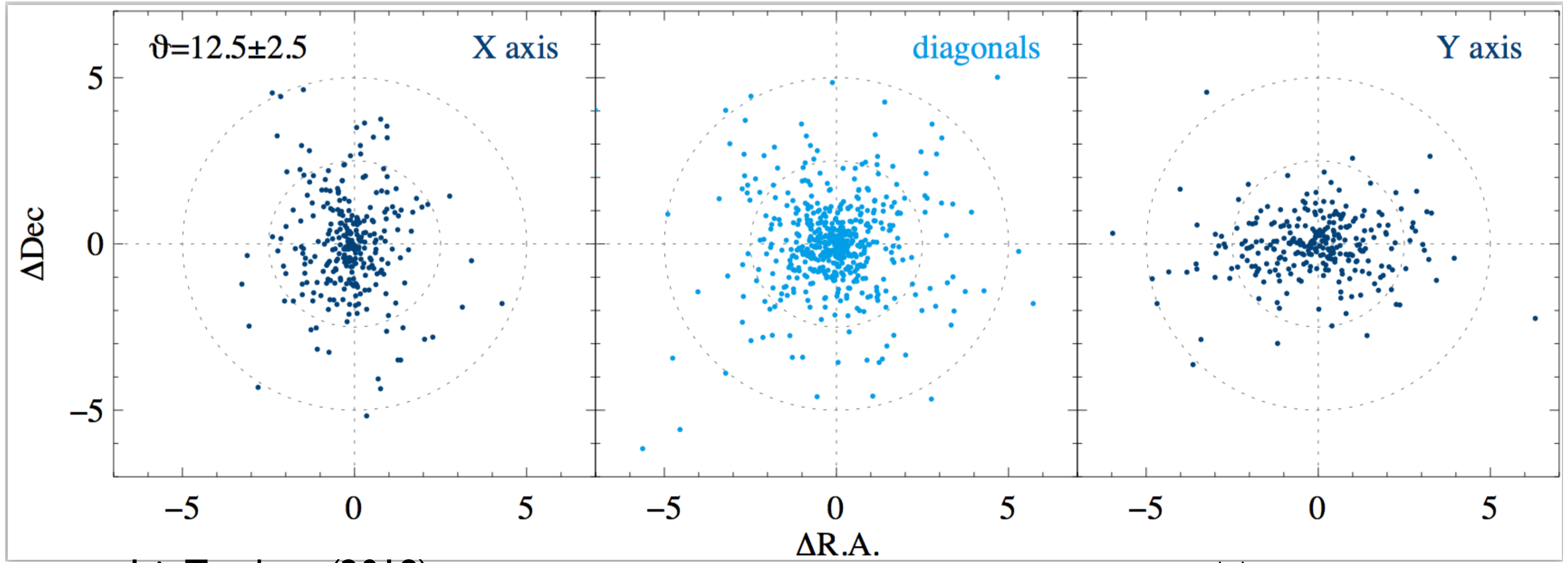




Courtesy: Iris Traulsen (2018)

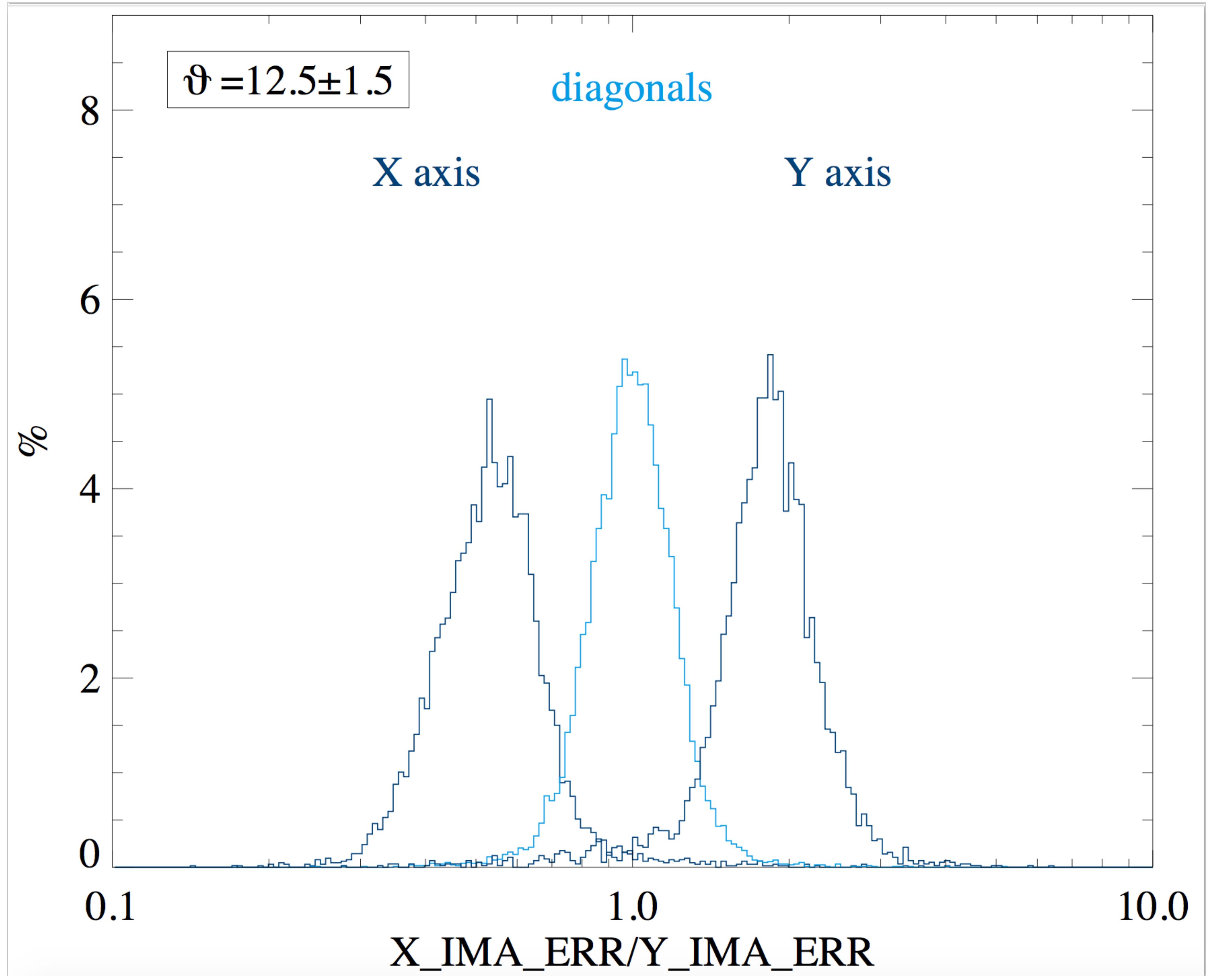
Don’t look for the CTP where it is not

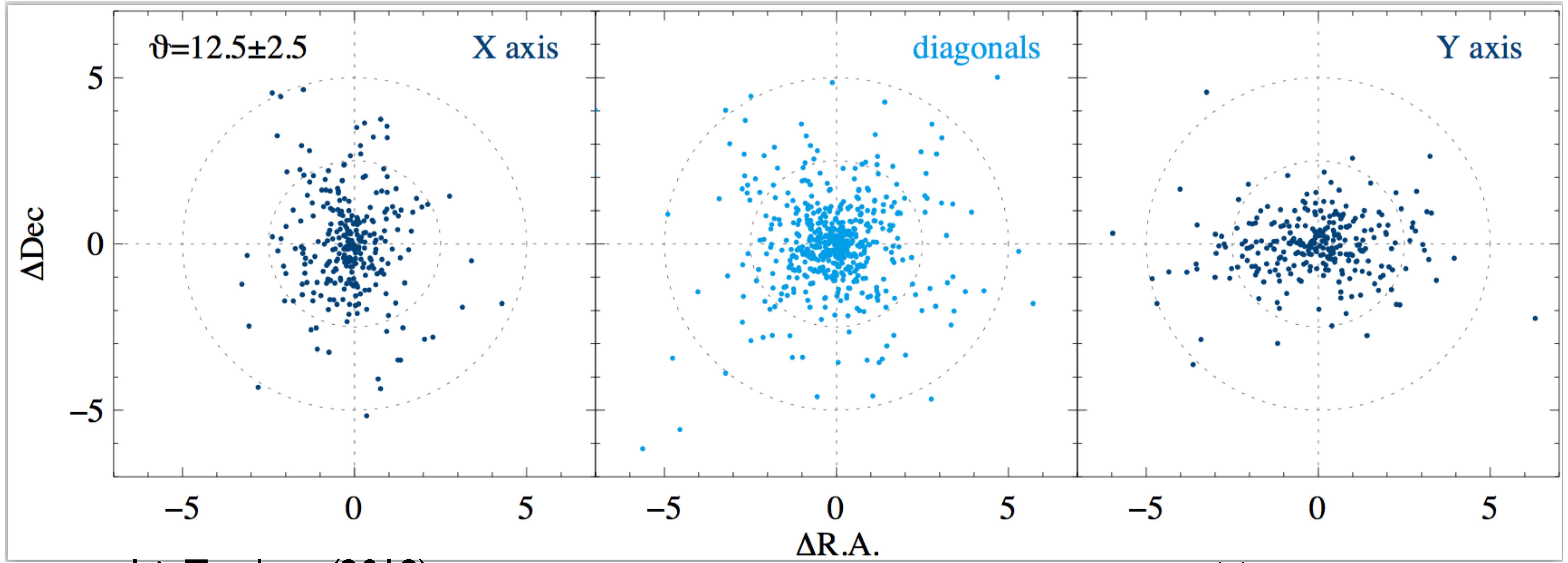




Courtesy: Iris Traulsen (2018)

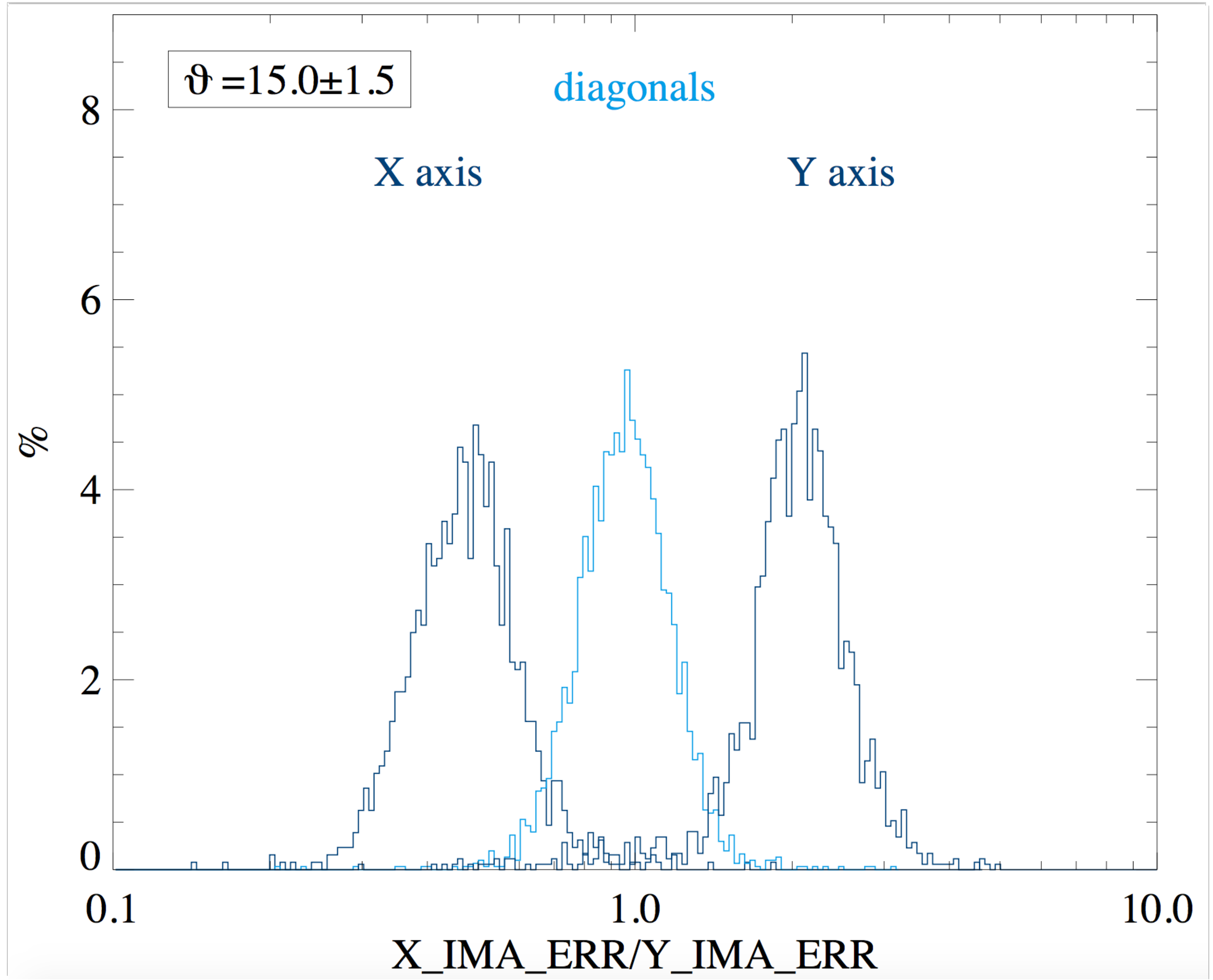
Don't look for the CTP where it is not

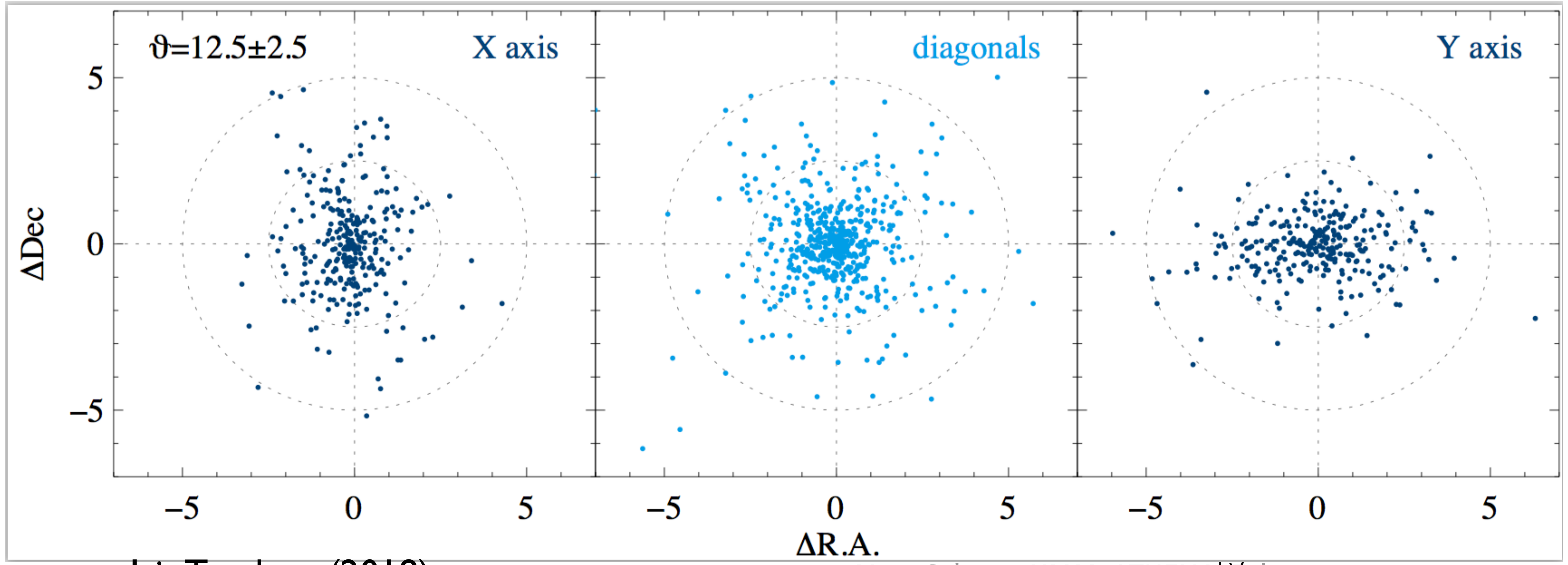




Courtesy: Iris Traulsen (2018)

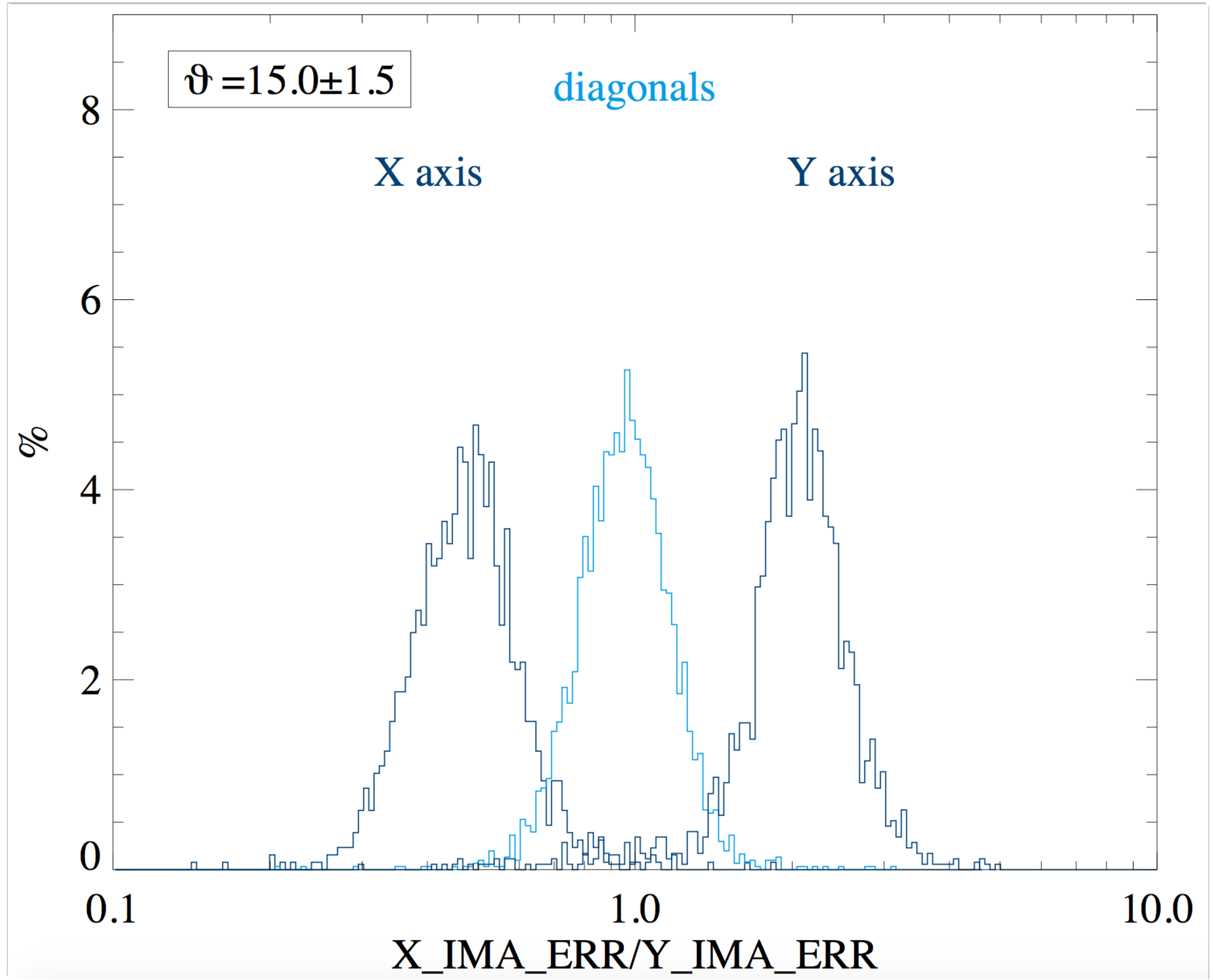
Don’t look for the CTP where it is not





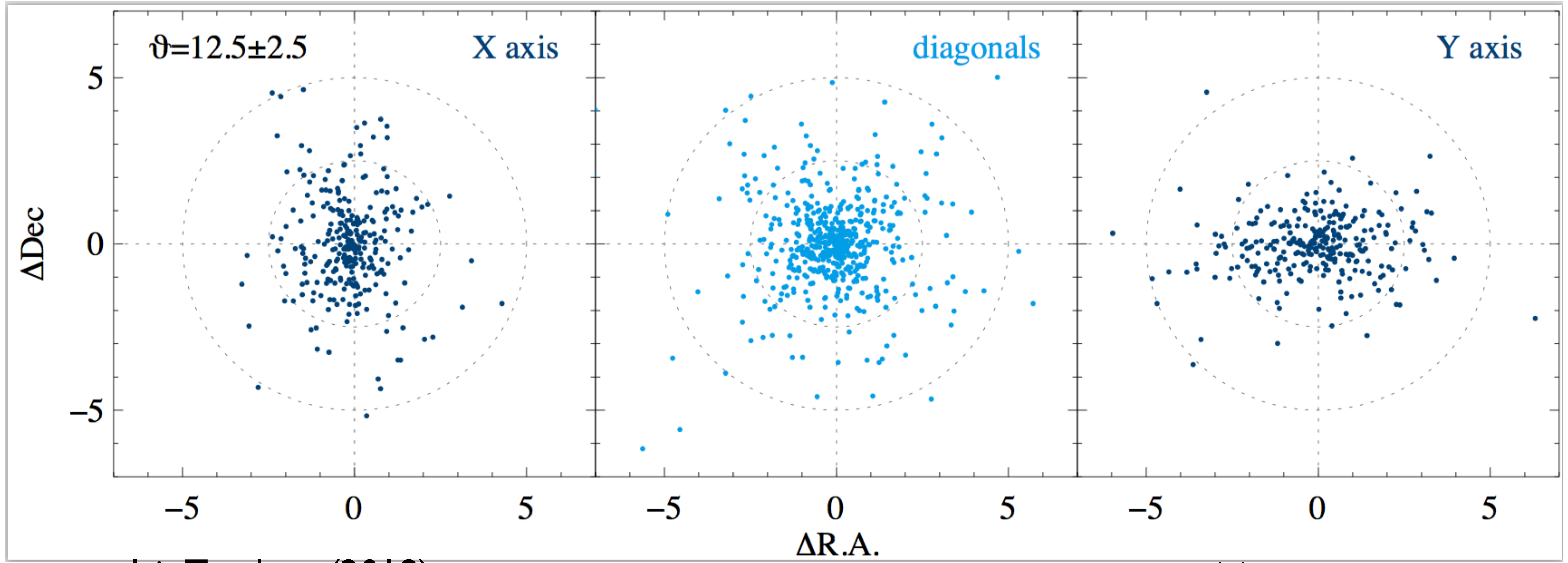
Courtesy: Iris Traulsen (2018)

Don’t look for the CTP where it is not



- The actual search of a ctp should be within an ellipse along the scanning direction

- Nway is already set for that

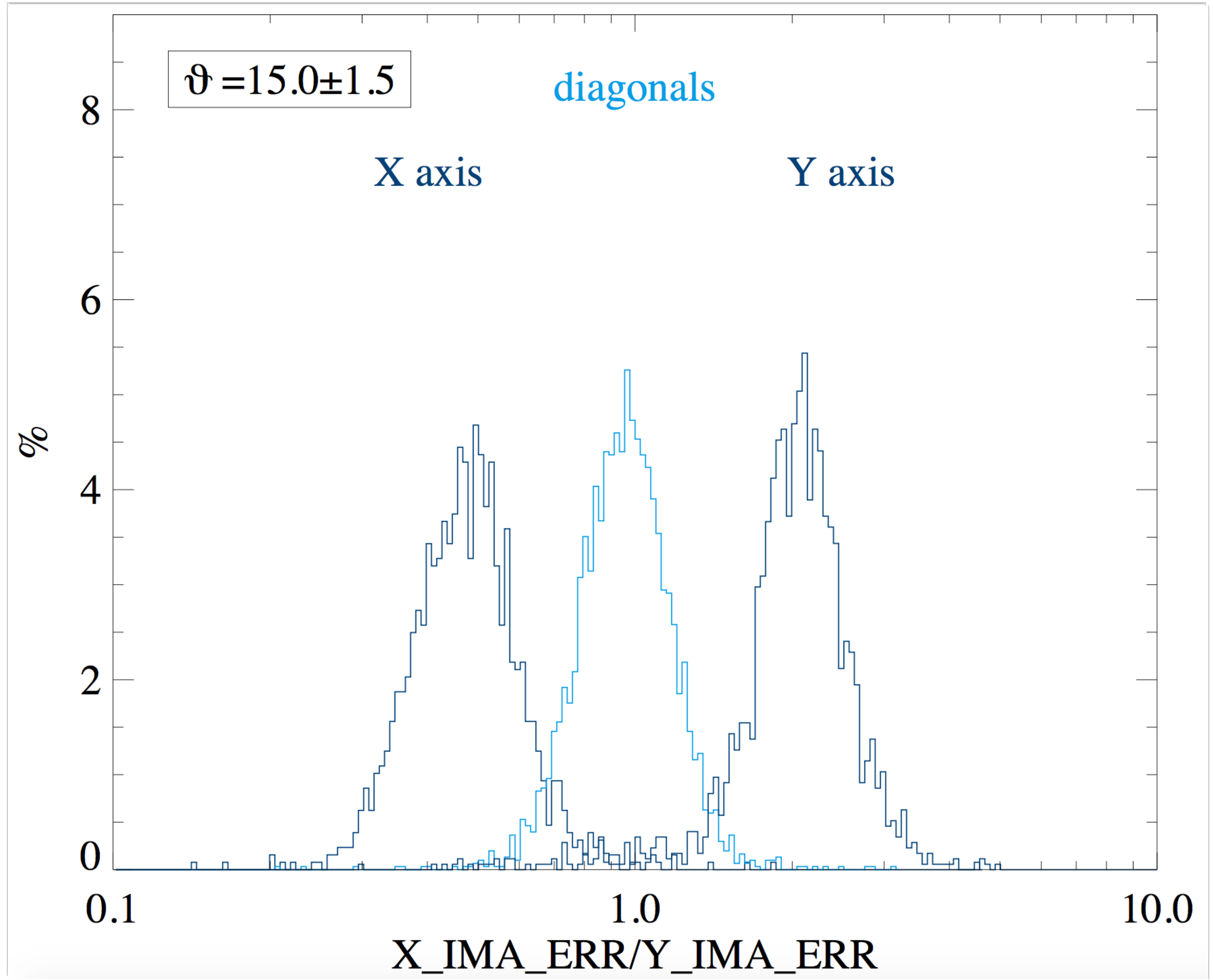


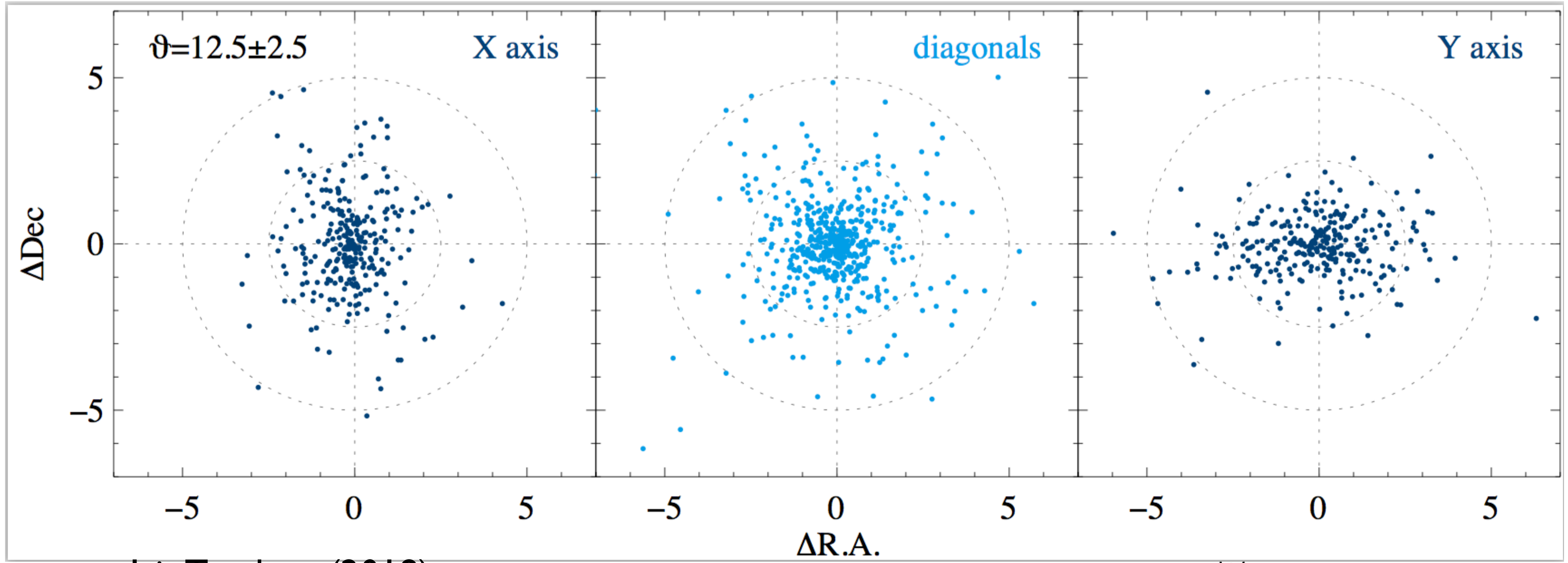
Courtesy: Iris Traulsen (2018)

Don't look for the CTP where it is not

- The actual search of a ctp
should be within an ellipse along the scanning direction

- Nway is already set for that





Courtesy: Iris Traulsen (2018)

Final Remarks

- Cross-matches are rarely trivial (and I did not mention blending, variability, proper motion, etc)

- Using a single magnitude/band we need to account for the possibility that the CTP is absent in that specific band.
  Better to use all bands at once (a.k.a. the SED), accounting for missing data

- We should not spend time searching for the ctp in the wrong place (use elliptical errors when possible)

- When using prior knowledge, the model must be built using training/control/validating samples that are representative of the survey that we interested on

- Our work is catalogs-based, and the catalogs were not built for us. We are not spending enough time understanding the caveats accompanying the catalogs (purpose of the catalog, depth, flag system).

- Specific to X-ray surveys: emitters are Galactic sources, resolved sources in nearby galaxies, AGN, QSO, and unresolved clusters:
  It is dangerous to focus only on a specific source type from the beginning.

- Making first the CTPs identification and then their classification allows to pinpoint interesting populations (e.g., unresolved clusters)

NWAY is a beautiful tool, but it still needs a brain to be used properly