Unlocking Gaia's potential with synthetic surveys

Robyn Sanderson
UPenn/Flatiron

Synthetic survey of a cosmo-hydro simulation
(Sanderson et al 2018)

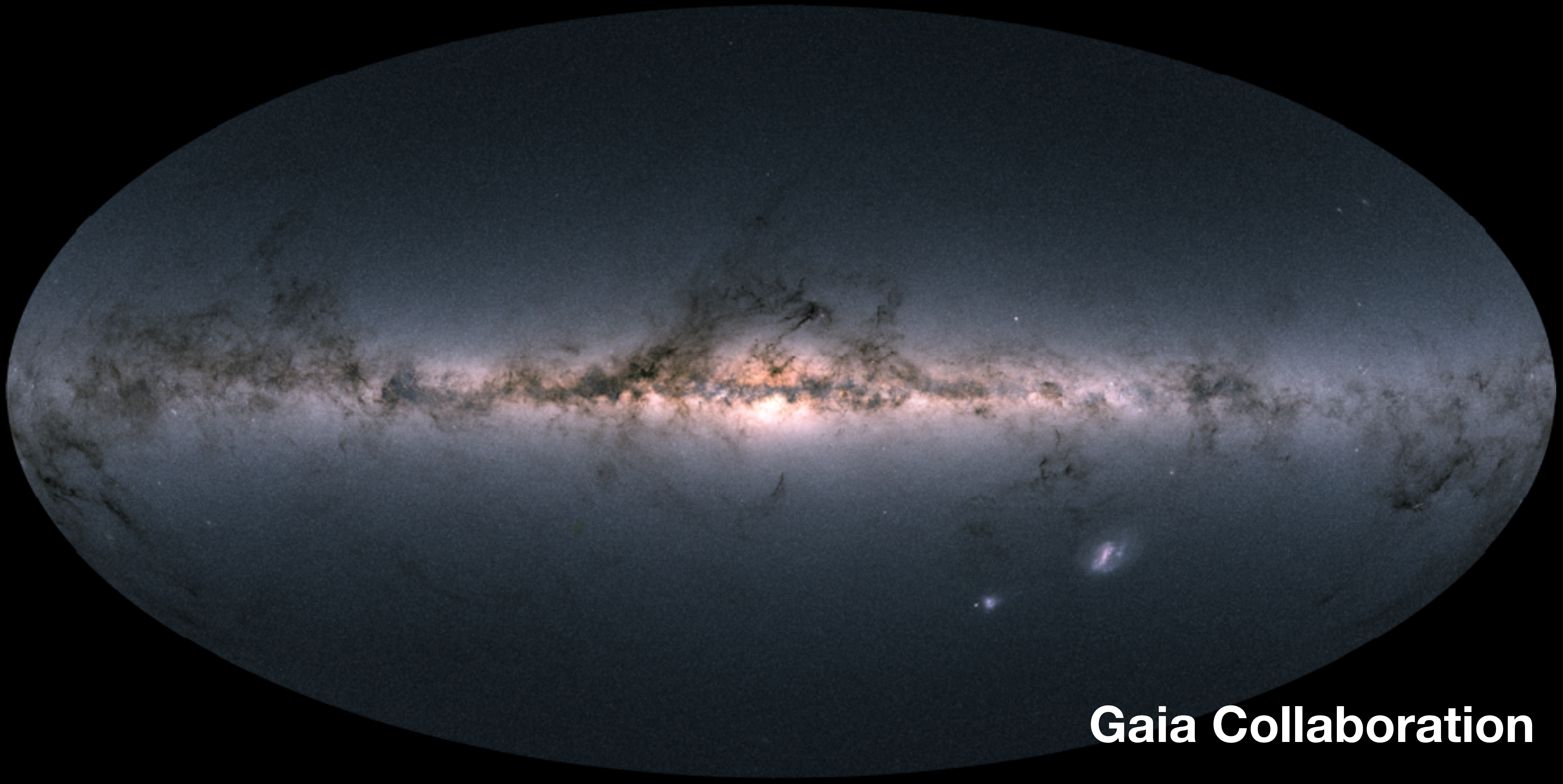
Unlocking Gaia’s potential with synthetic surveys

Robyn Sanderson
UPenn/Flatiron

Milky Way
(image credit:ESO)

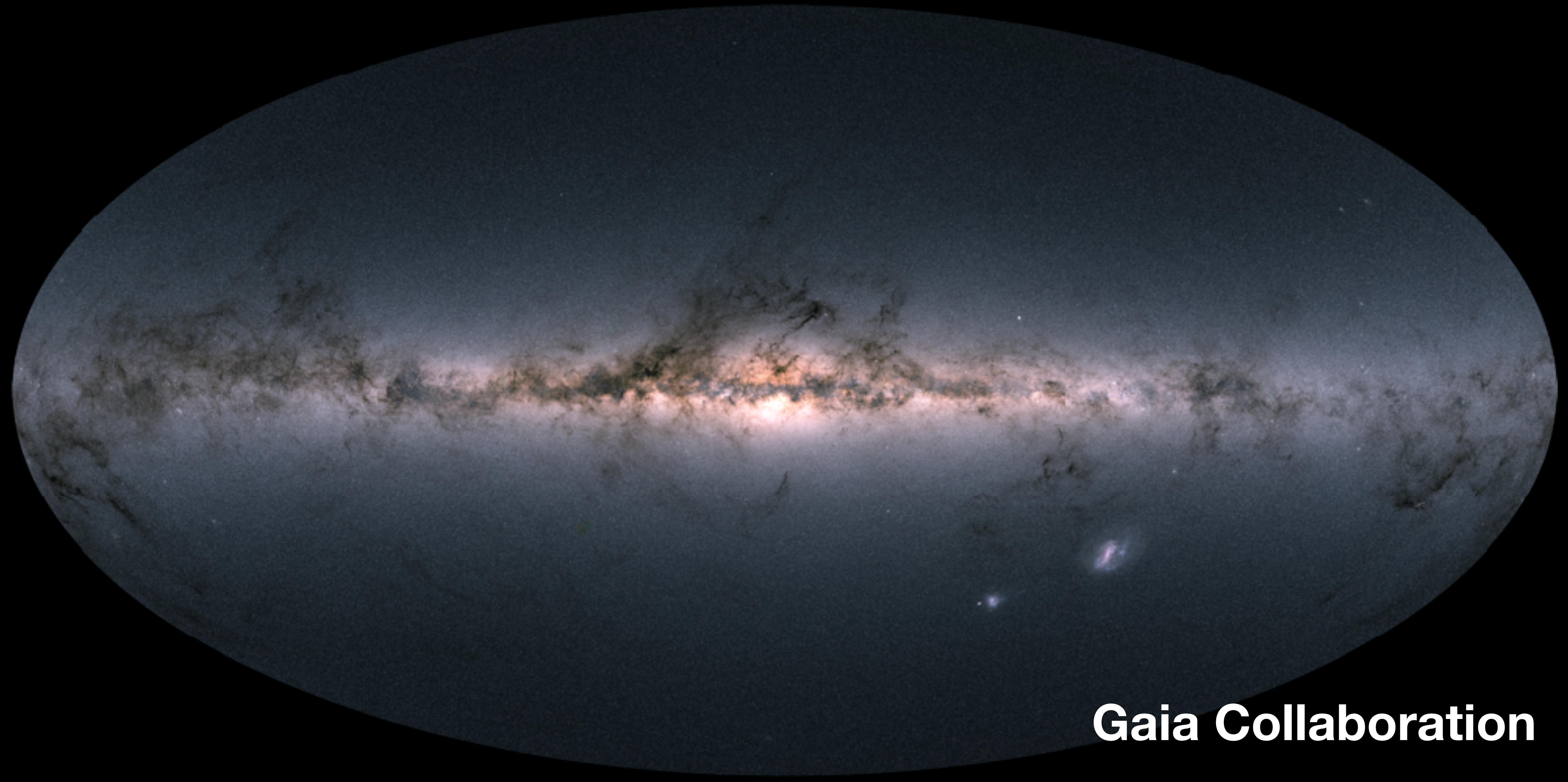
Wetzel et al. 2016, movie credit: Phil Hopkins

Wetzel et al. 2016, movie credit: Phil Hopkins

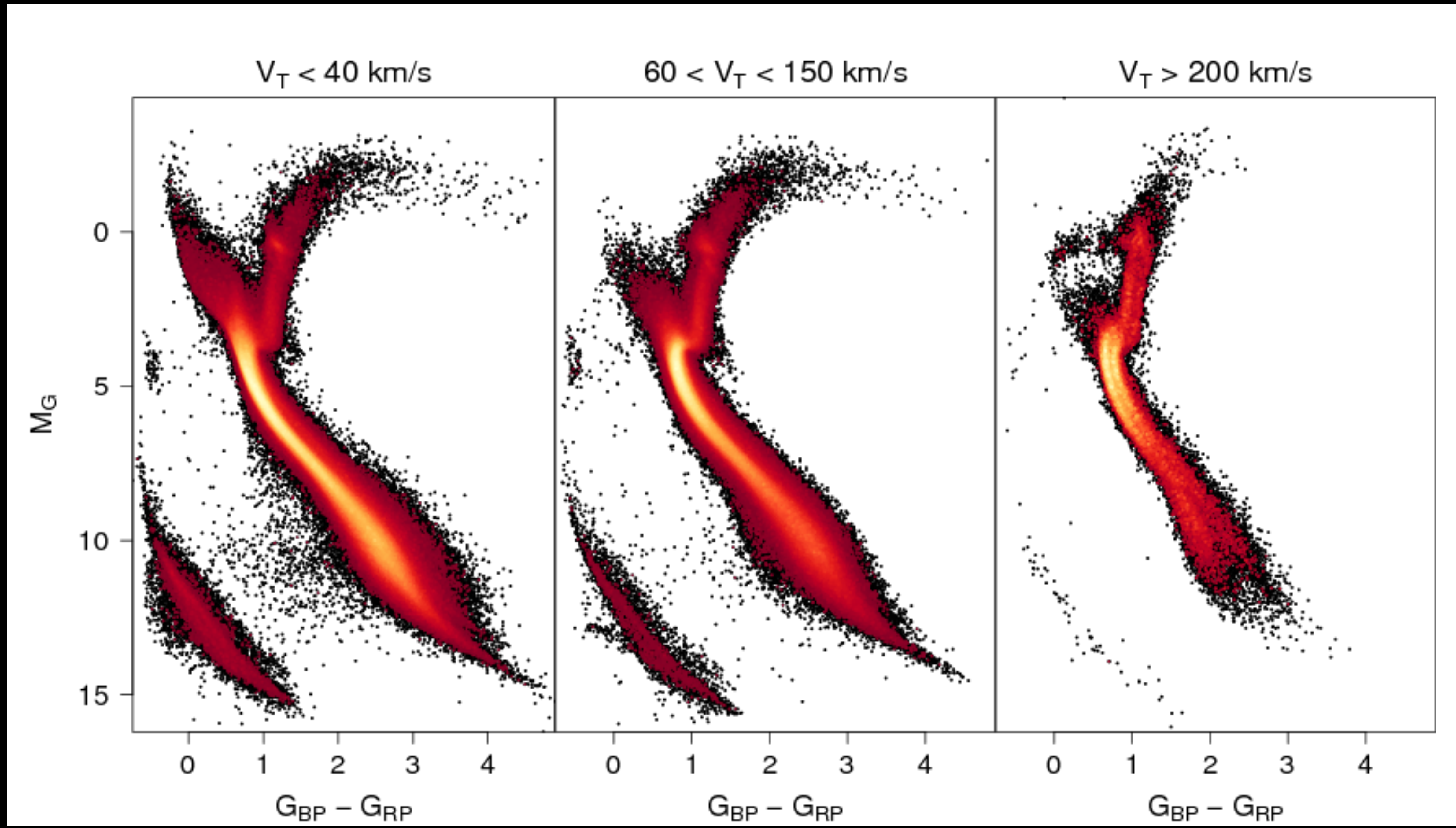


Gaia Collaboration

Wetzel et al. 2016, movie credit: Phil Hopkins



Gaia Collaboration



Wetzel et al. 2016, movie credit: Phil Hopkins

Babusiaux et al 2018

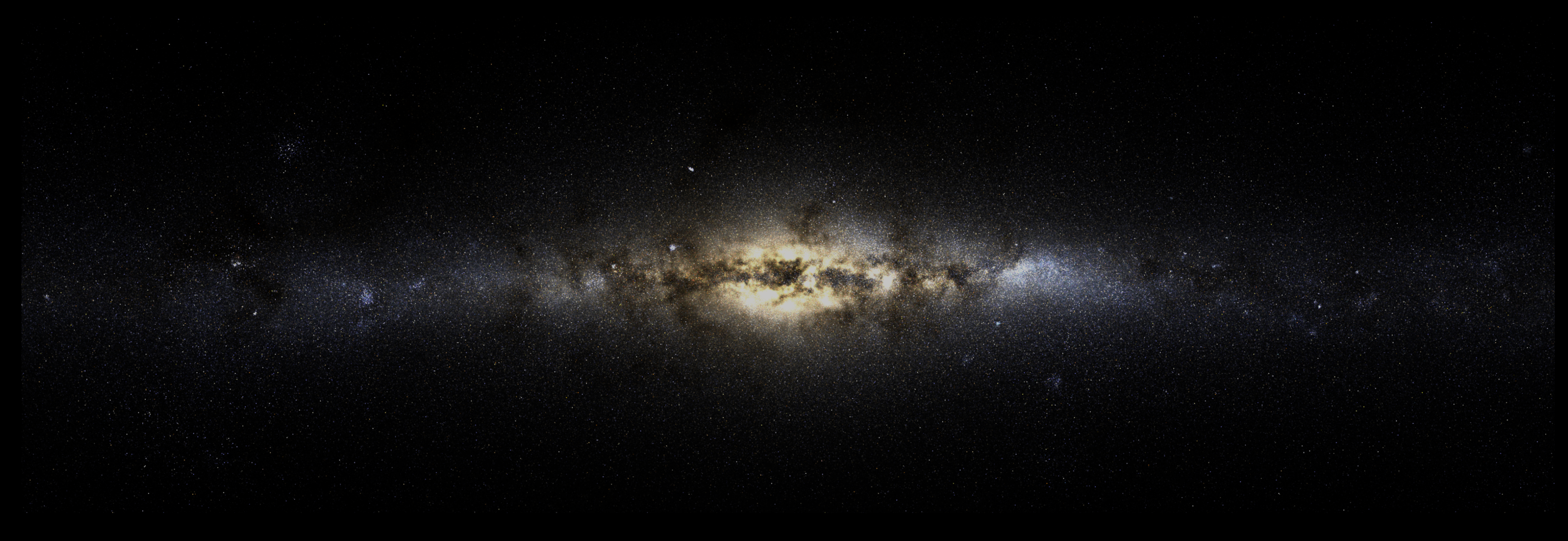
outline

- how to make a synthetic survey
- how to use synthetic surveys:
  - making forecasts & planning survey strategies
  - evaluating the “selection function” of a search
  - validating complex analysis methods
  - interpreting structures identified by data-driven models
  - training ML methods for application to Gaia

outline

- how to make a synthetic survey
- how to use synthetic surveys:
  - making forecasts & planning survey strategies
  - evaluating the “selection function” of a search
  - validating complex analysis methods
  - interpreting structures identified by data-driven models
  - training ML methods for application to Gaia

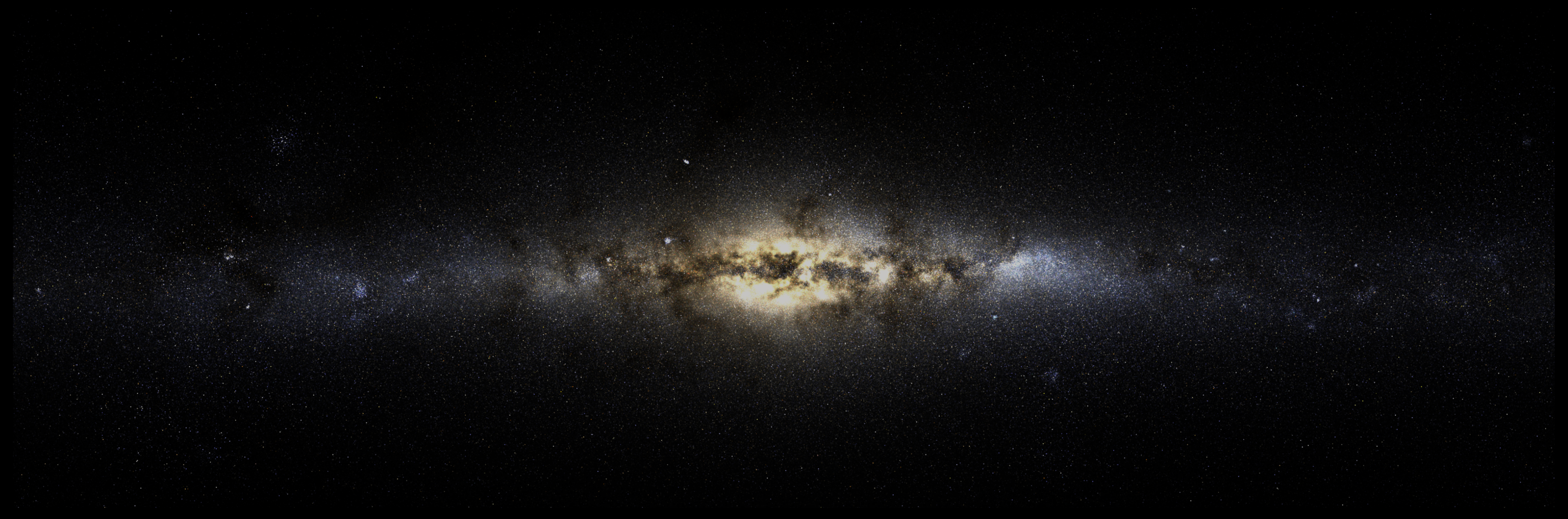
Making predictions for a 6+D galaxy



Making predictions for a 6+D galaxy

Galaxy Simulation
(cosmology, DM model, gravity, gas physics, star formation, stellar feedback, …)

One particle = many "stars"
…with same age, abundances



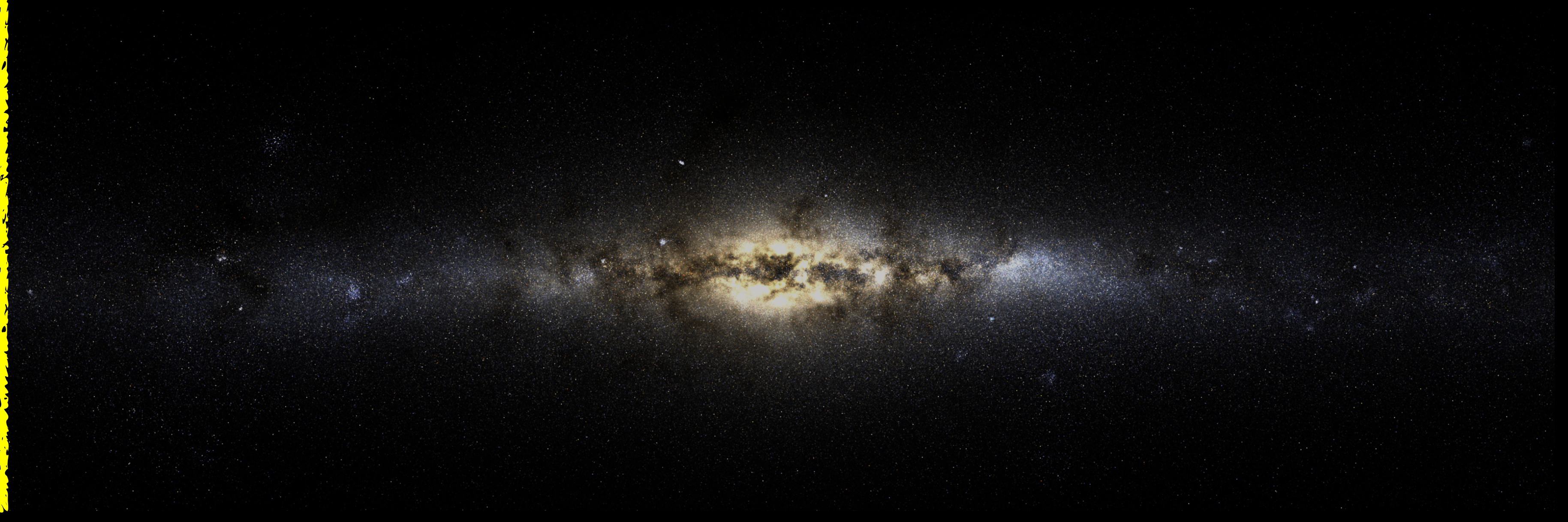
Making predictions for a 6+D galaxy

Galaxy Simulation
(cosmology, DM model, gravity, gas physics, star formation, stellar feedback, …)

Stellar Populations

(stellar structure, stellar evolution, convection models, isochrone mapping, IMF, …)

Phase-space density estimation
(kernel dimension, smoothing scales, ages, accretion history, …)

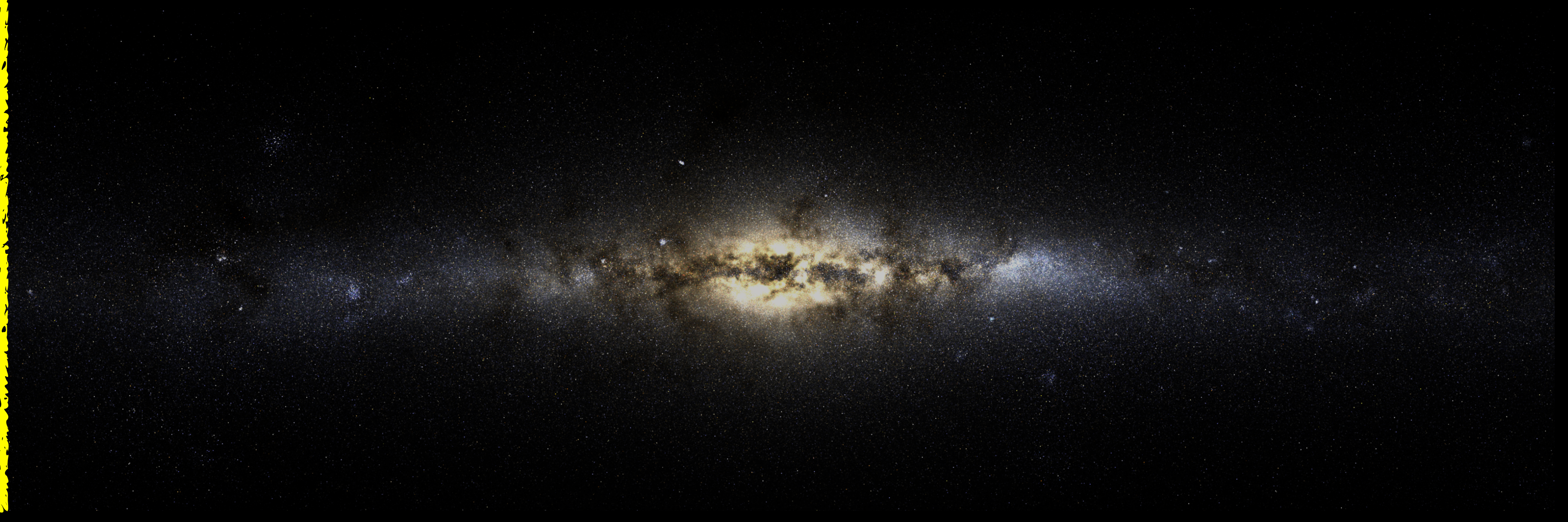


One particle = many “stars”
…with same age, abundances

Making predictions for a 6+D galaxy

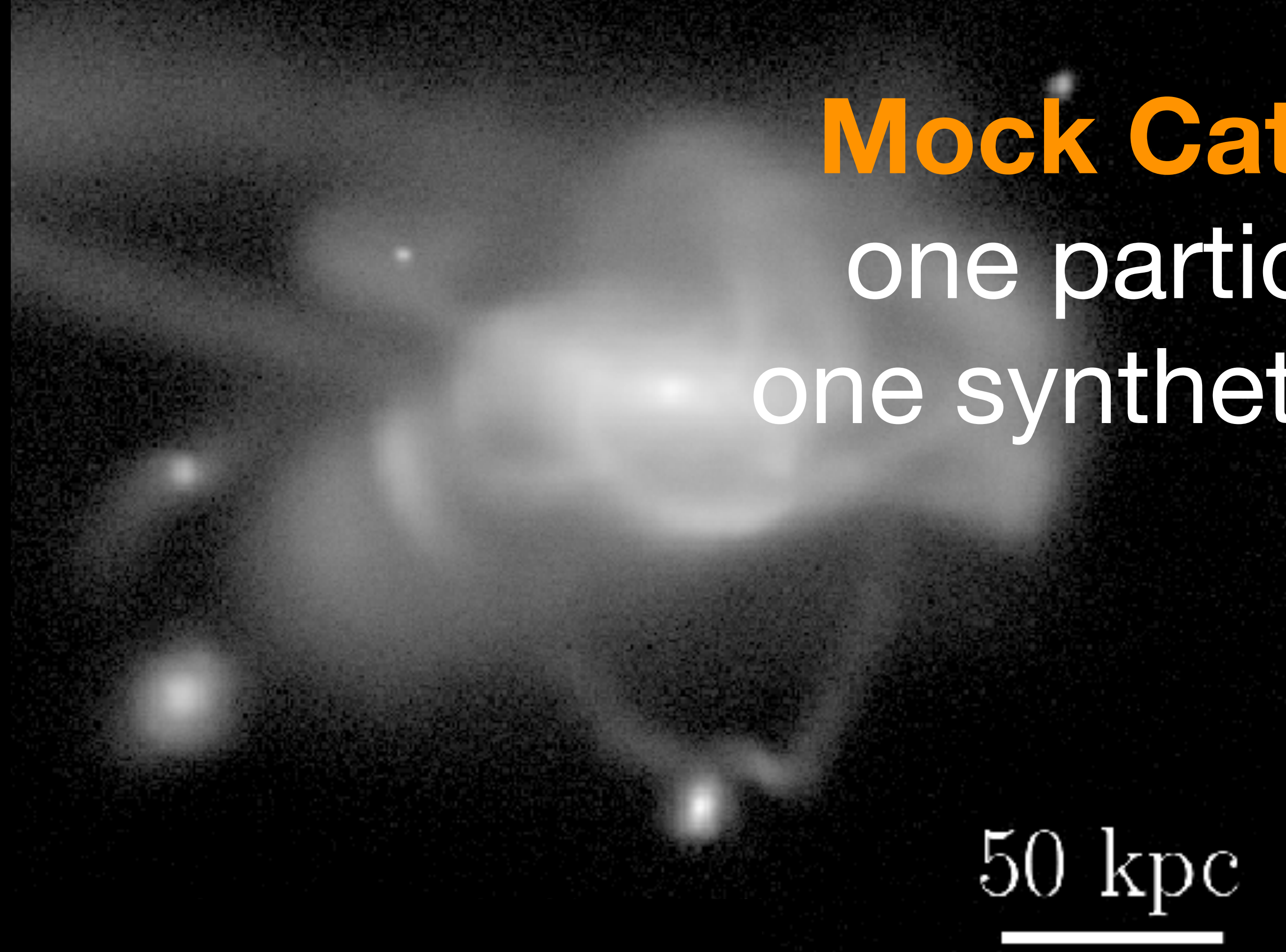
Galaxy Simulation
(cosmology, DM model, gravity, gas physics, star formation, stellar feedback, …)

Stellar Populations
(stellar structure, stellar evolution, convection models, isochrone mapping, IMF, …)



Phase-space density estimation
(kernel dimension, smoothing scales, ages, accretion history, …)

One particle = many “stars”
…with same age, abundances



Mock Catalog
one particle =
one synthetic star

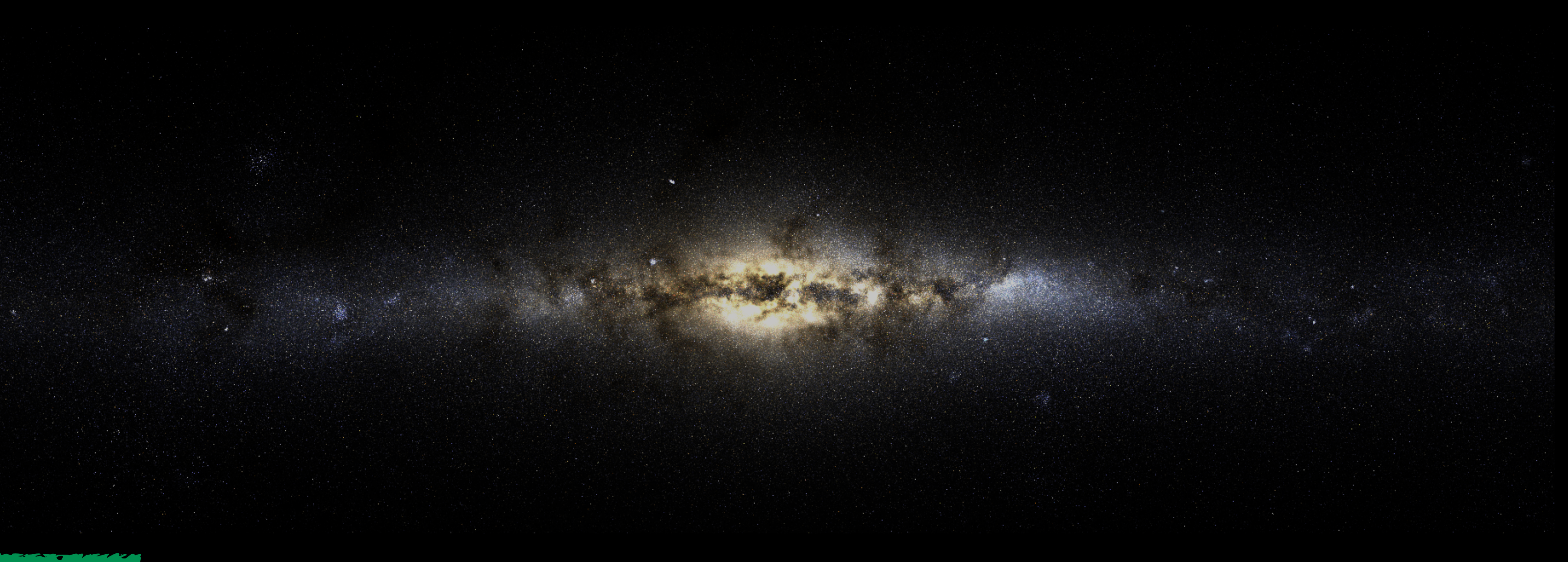
Making predictions for a 6+D galaxy

Galaxy Simulation
(cosmology, DM model, gravity, gas physics, star formation, stellar feedback, …)

Stellar Populations

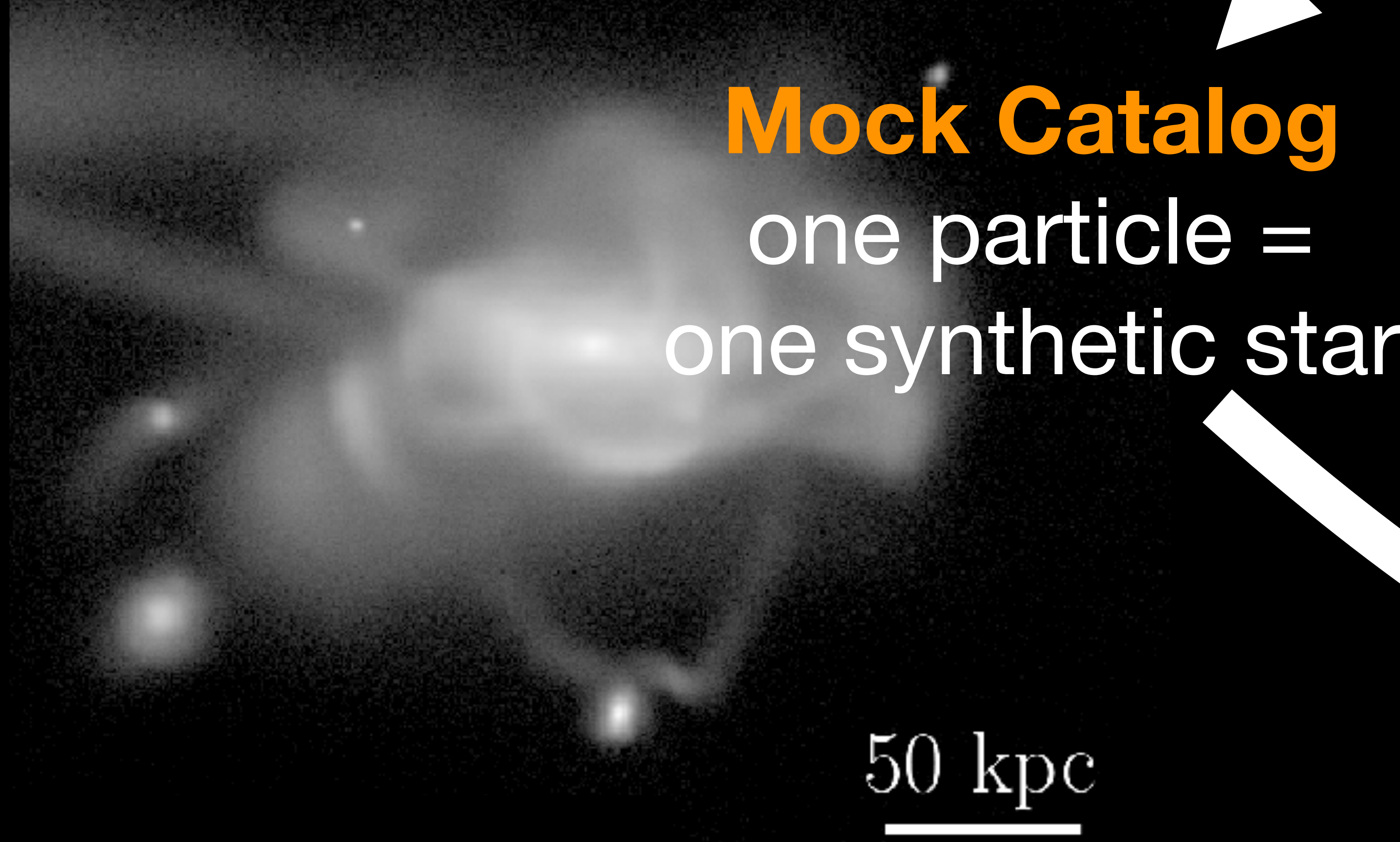
(stellar structure, stellar evolution, convection models, isochrone mapping, IMF, …)

Synthetic Survey
one particle = one “observed” star



Phase-space density estimation
(kernel dimension, smoothing scales, ages, accretion history, …)

One particle = many “stars”
…with same age, abundances



Mock Catalog
one particle =
one synthetic star

Survey description

(Magnitude/color limits, extinction/reddening, selection function, error model, instrument model, …)

Making predictions for a 6+D galaxy

Making predictions for a 6+D galaxy

- **Simple mock accreted halos**
  (e.g. Sanderson, Helmi, & Hogg 2015)
  - spherical analytic halo
  - building blocks matched to satellite mass function
  - single tracers ad hoc
    (e.g. K giants, RR Lyrae)

Making predictions for a 6+D galaxy

- **Simple mock accreted halos**
  (e.g. Sanderson, Helmi, & Hogg 2015)
  - spherical analytic halo
  - building blocks matched to satellite mass function
  - single tracers ad hoc
    (e.g. K giants, RR Lyrae)

- **Aquarius**
  (Cooper et al. 2010, Lowing et al 2012)
  - Resampled cosmological sim
  - DM-only + tagging
    (no disk)
  - 6D positions, velocities

Making predictions for a 6+D galaxy

- **Simple mock accreted halos**
  (e.g. Sanderson, Helmi, & Hogg 2015)
  - spherical analytic halo
  - building blocks matched to satellite mass function
  - single tracers ad hoc
    (e.g. K giants, RR Lyrae)

- **Galaxia, GUMS**
  (Sharma et al. 2011; Gaia DPAC)
  - semi-analytic accreted halo
    (Bullock & Johnston 2005)
  - empirical disk, bulge
    (Robin et al 2001)
  - complete stellar populations
  - 6D+Fe,“alpha”+age

- **Aquarius**
  (Cooper et al. 2010, Lowing et al 2012)
  - Resampled cosmological sim
  - DM-only + tagging
    (no disk)
  - 6D positions, velocities

Making predictions for a 6+D galaxy

- **Simple mock accreted halos**
  (e.g. Sanderson, Helmi, & Hogg 2015)
  - spherical analytic halo
  - building blocks matched to satellite mass function
  - single tracers ad hoc
    (e.g. K giants, RR Lyrae)

- **Galaxia, GUMS**
  (Sharma et al. 2011; Gaia DPAC)
  - semi-analytic accreted halo
    (Bullock & Johnston 2005)
  - empirical disk, bulge
    (Robin et al 2001)
  - complete stellar populations
  - 6D+Fe,“alpha”+age

- **Aquarius**
  (Cooper et al. 2010, Lowing et al 2012)
  - Resampled cosmological sim
  - DM-only + tagging
    (no disk)
  - 6D positions, velocities

- **Ananke, Aurigaia**
  (Sanderson et al 2018, Grand et al. 2018)
  - Cosmological sim with hydro —> realistic central MW
  - 6D + 10 abundances + ages + …
  - Complete stellar populations

Ananke

Sanderson et al. 2018,
arXiv:1806.10564

- Cosmological sim with hydro —> realistic central MW
- 6D + 10 abundances + ages + …
- Complete stellar populations
- 3 simulations x 3 observation volumes = 9 surveys

Andrew Wetzel

Sarah Loebman

Sanjib Sharma

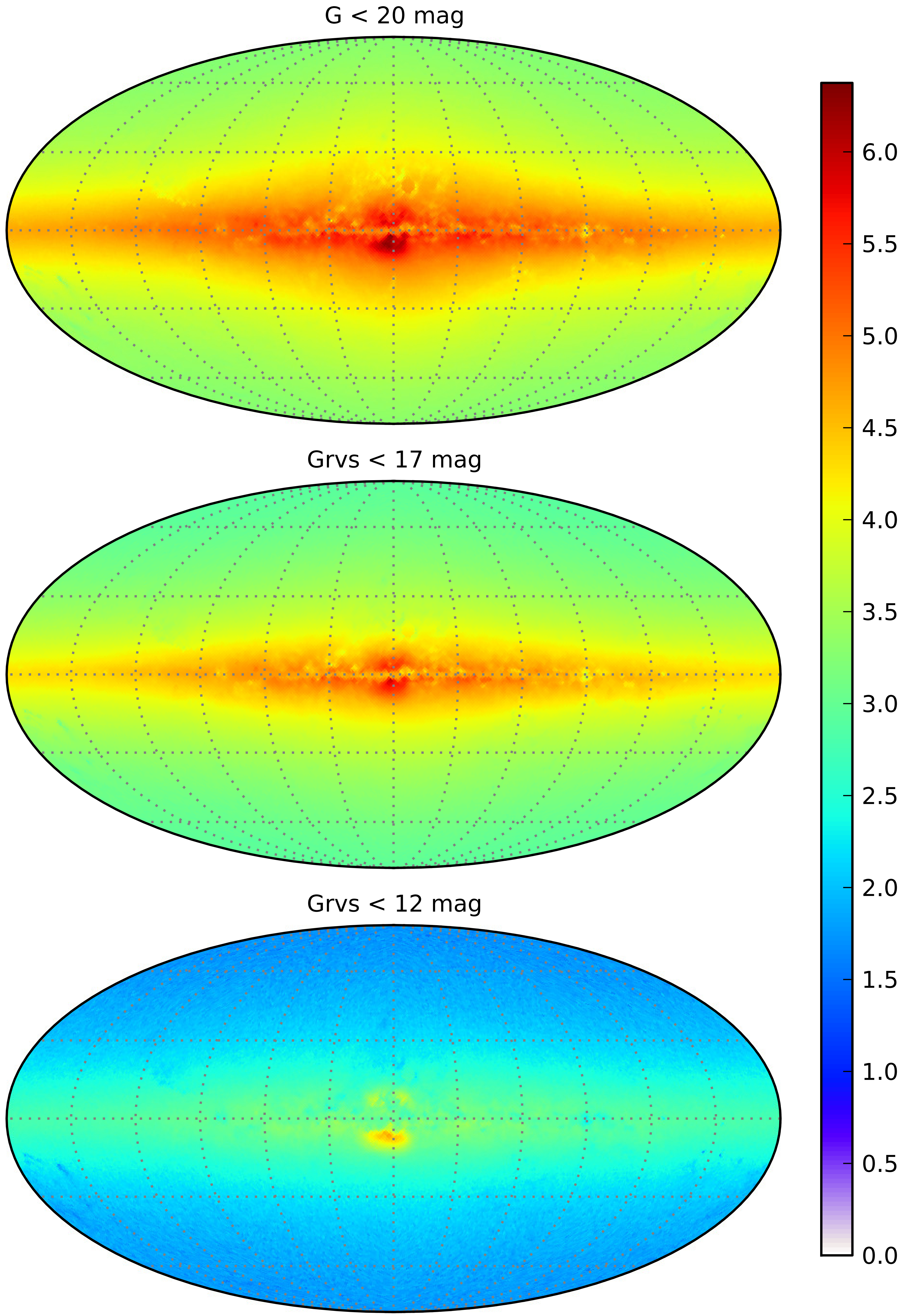
girder.hub.yt

Available for Gaia DR2 on:

outline

- how to make a synthetic survey
- how to use synthetic surveys:
  - making forecasts & planning survey strategies
  - evaluating the “selection function” of a search
  - validating complex analysis methods
  - interpreting structures identified by data-driven models
  - training ML methods for application to Gaia

- making forecasts & planning survey strategies
  - Gaia Universe Model Snapshot: Robin et al. 2012
    - multicomponent equilibrium model
    - tailored to MW
    - no cosmological accretion

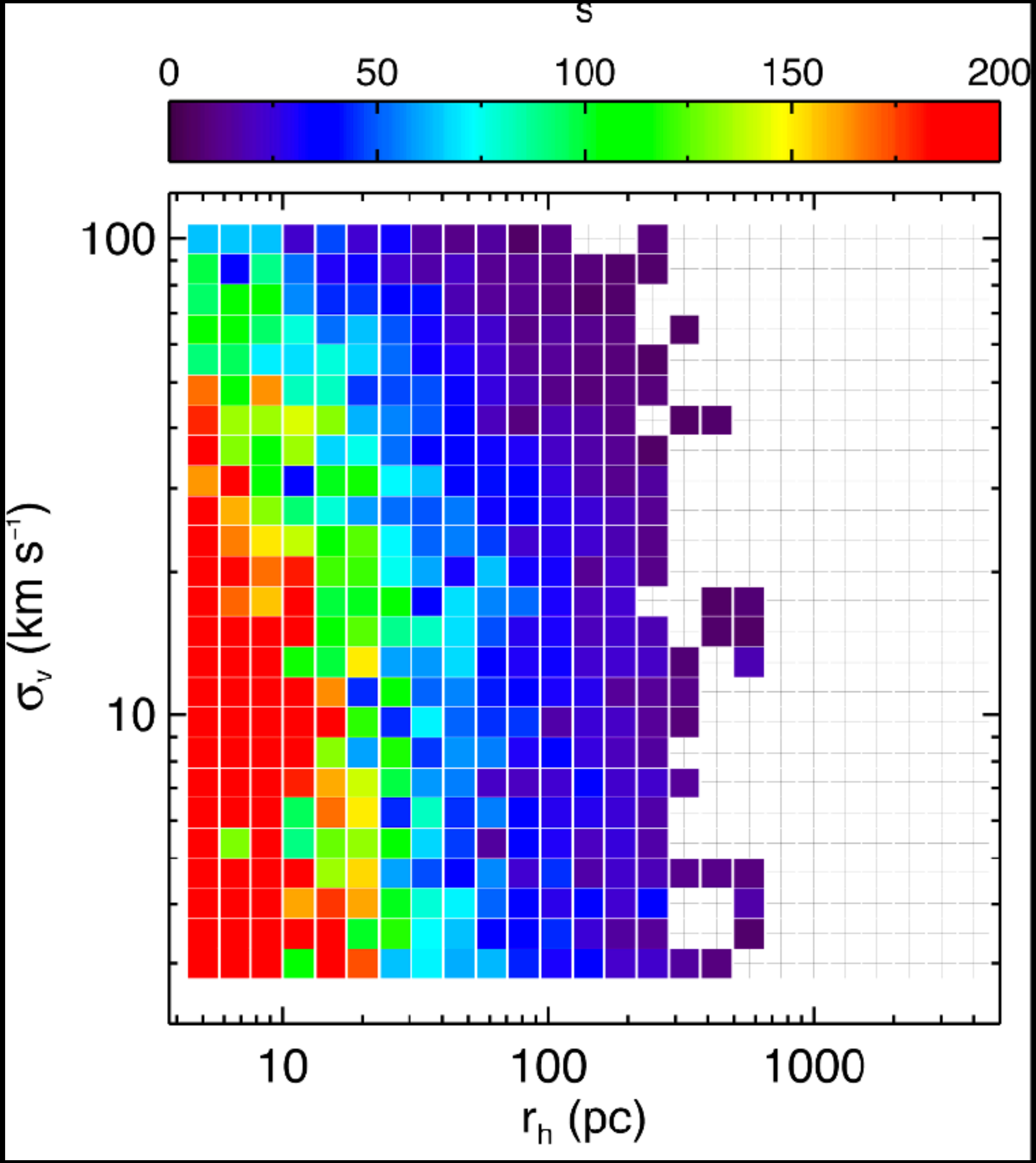


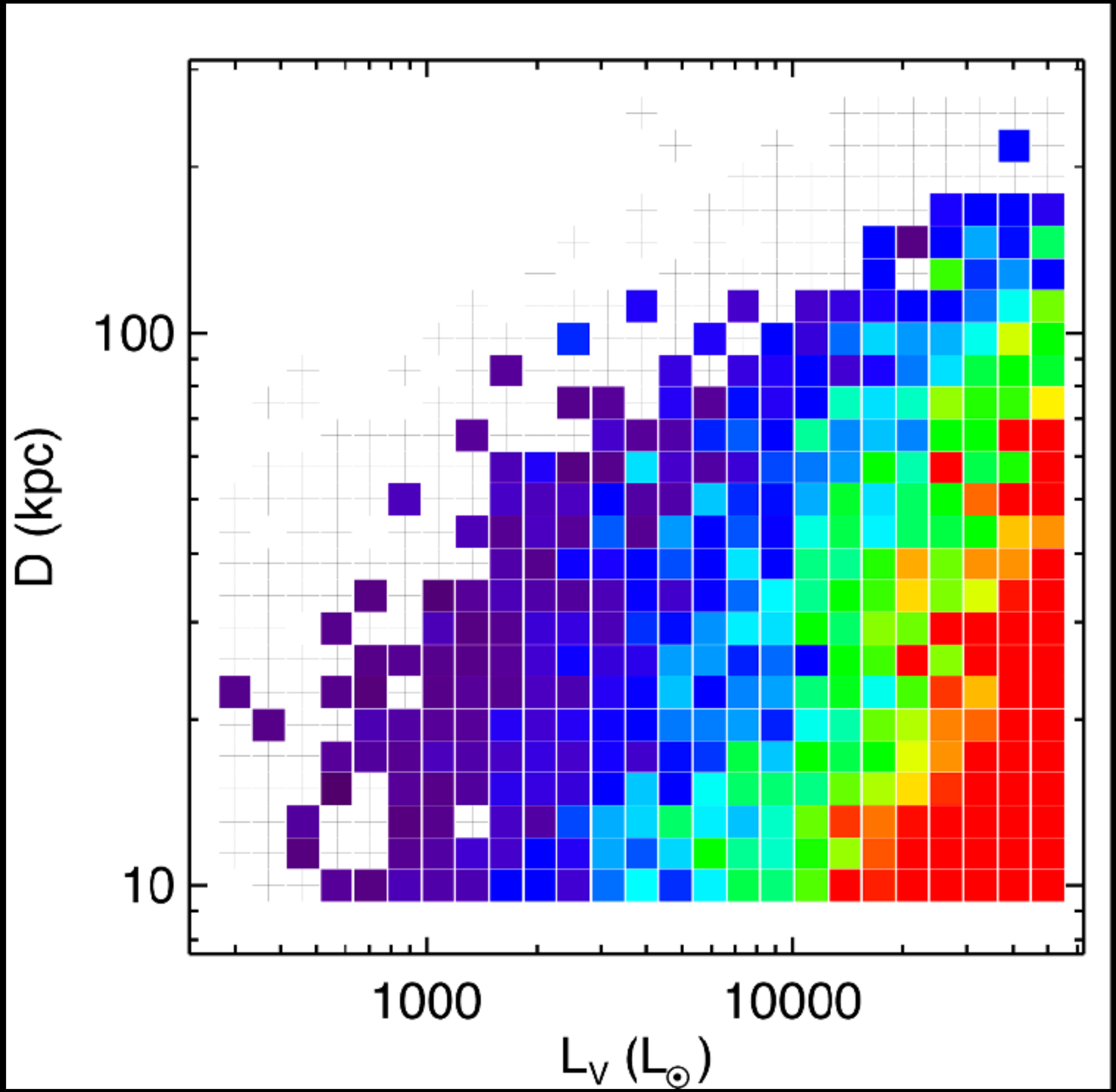
outline

- how to make a synthetic survey
- how to use synthetic surveys:
  - making forecasts & planning survey strategies
  - evaluating the “selection function” of a search
  - validating complex analysis methods
  - interpreting structures identified by data-driven models
  - training ML methods for application to Gaia

selection functions

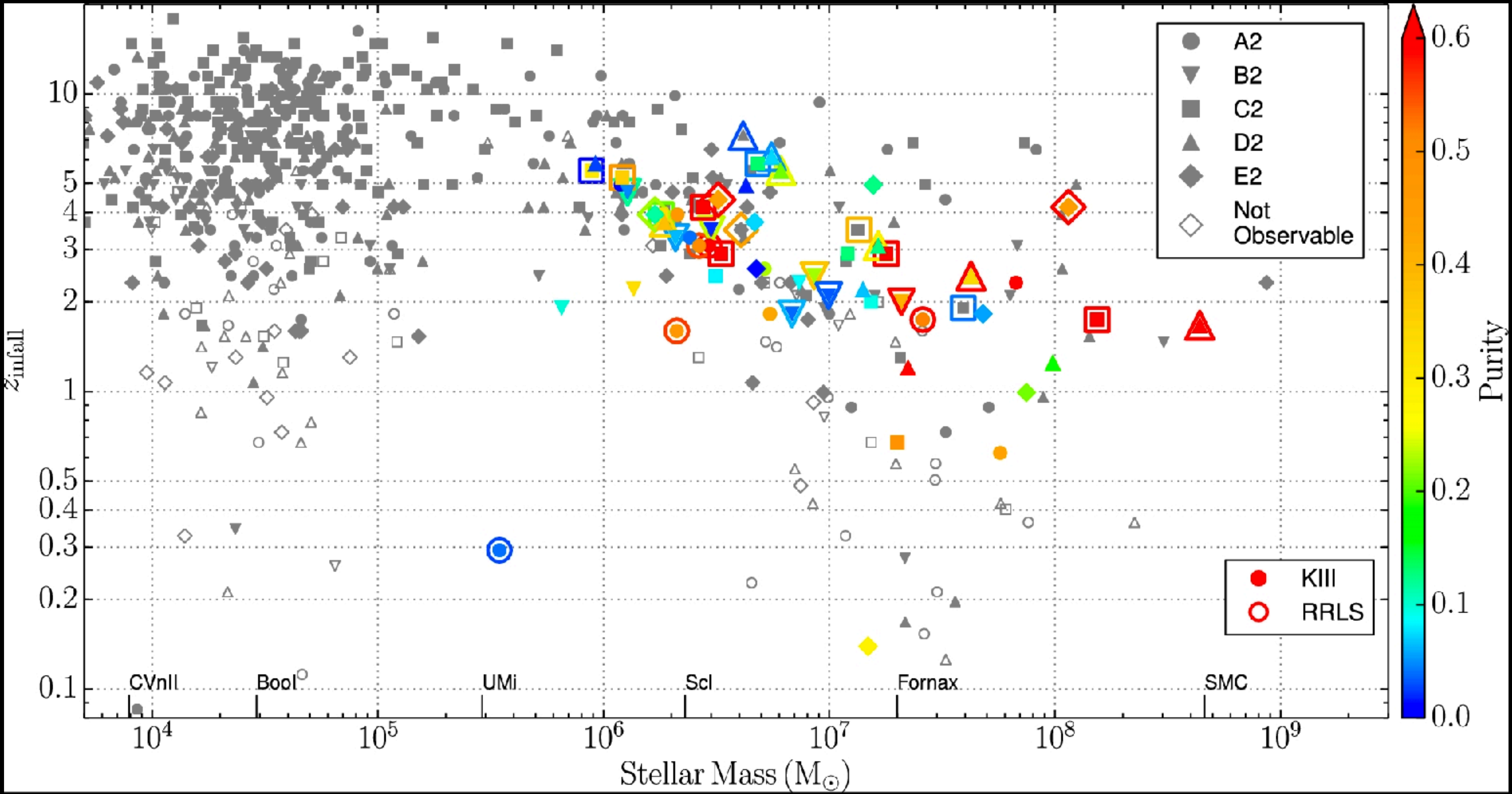
Antoja et al .2015 for Gaia dwarfs





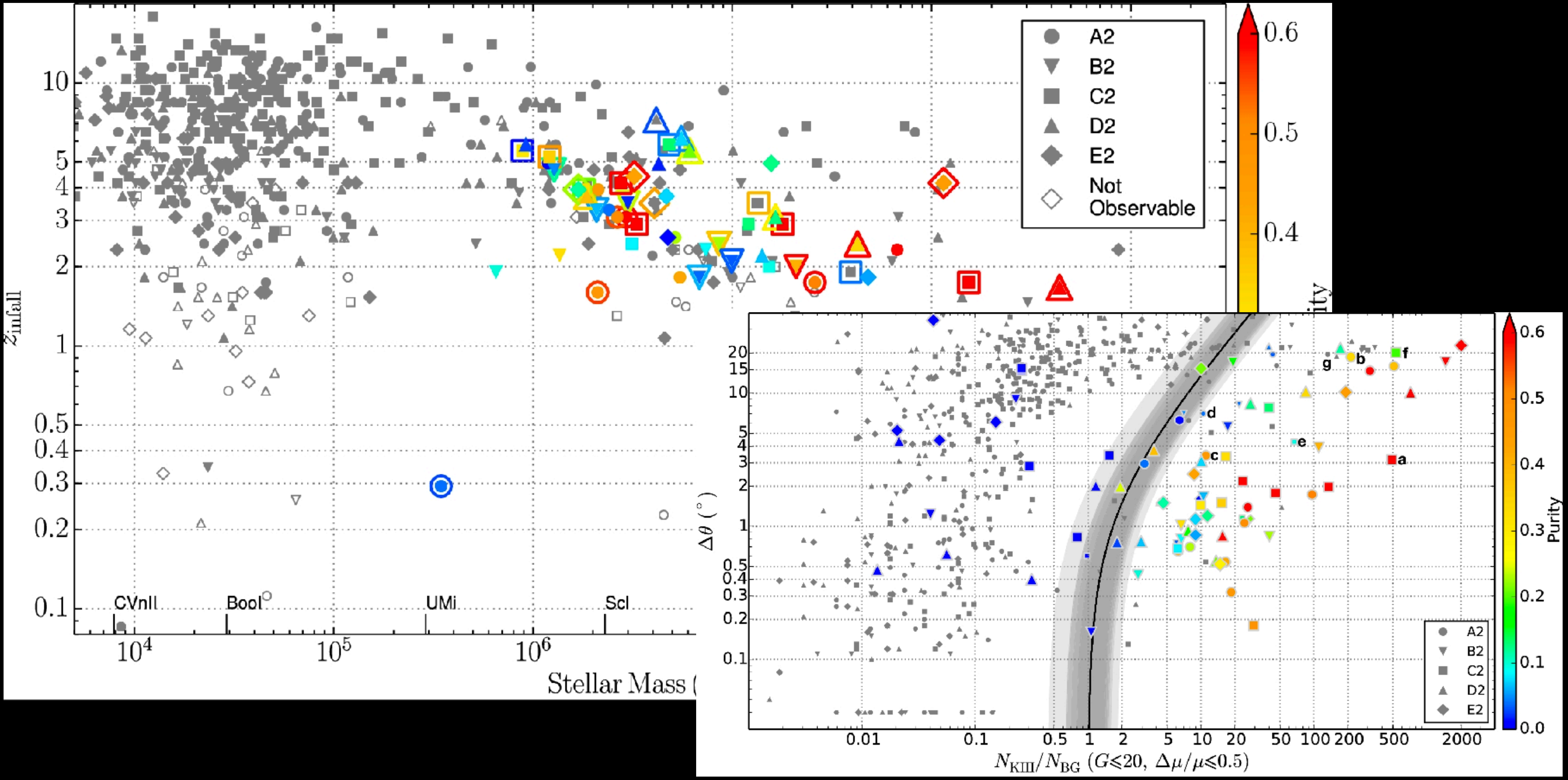
selection functions

Mateu et al 2017 for Gaia streams



selection functions

Mateu et al 2017 for Gaia streams

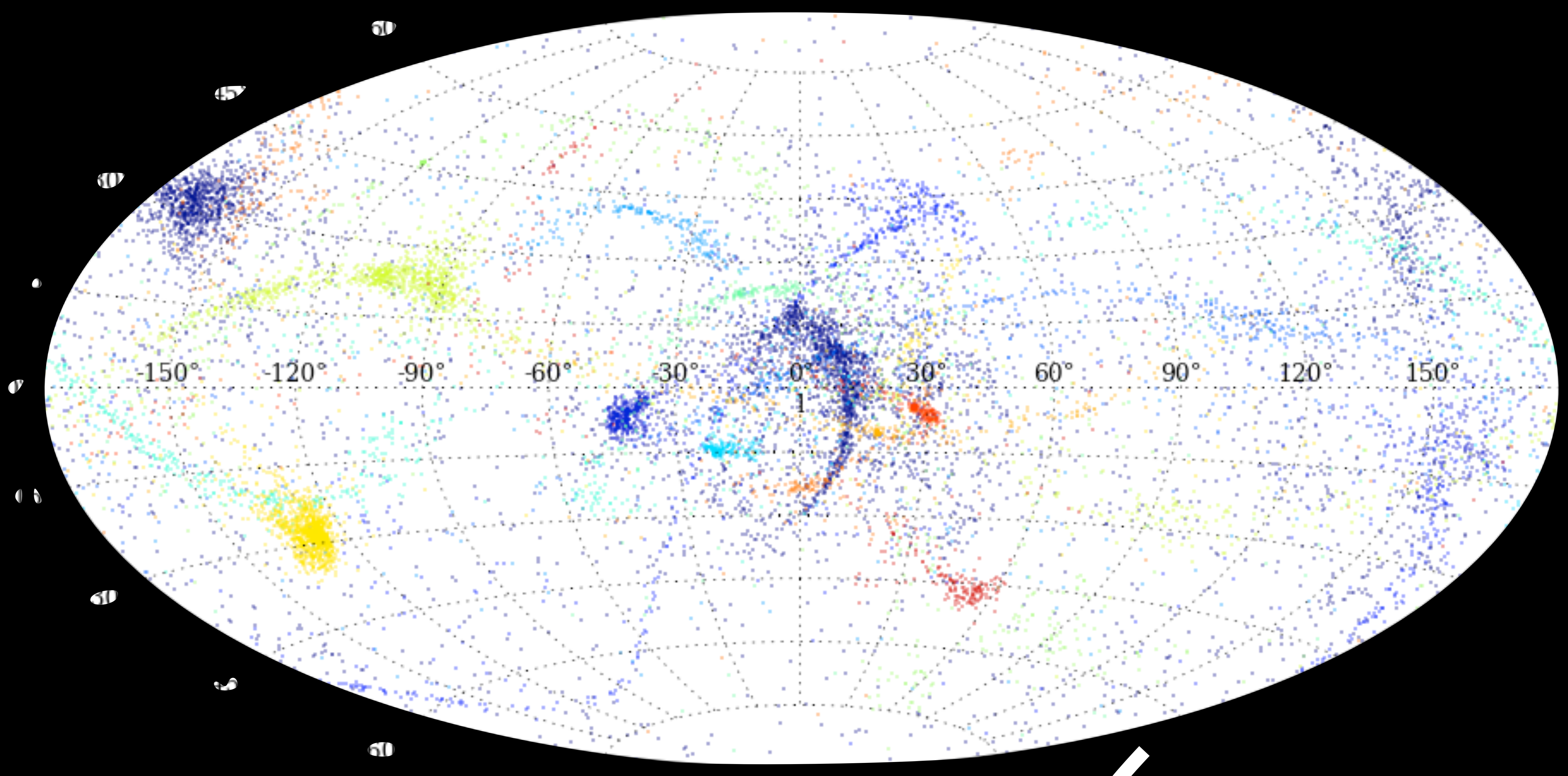


outline

- how to make a synthetic survey
- how to use synthetic surveys:
  - making forecasts & planning survey strategies
  - evaluating the “selection function” of a search
  - validating complex analysis methods
  - interpreting structures identified by data-driven models
  - training ML methods for application to Gaia

The accreted stellar halo is clumpy in constants-of-motion space

Galactic coordinates



One particle = many stars

Base simulation: Cooper et al. 2010

Extremely nonlinear transformation

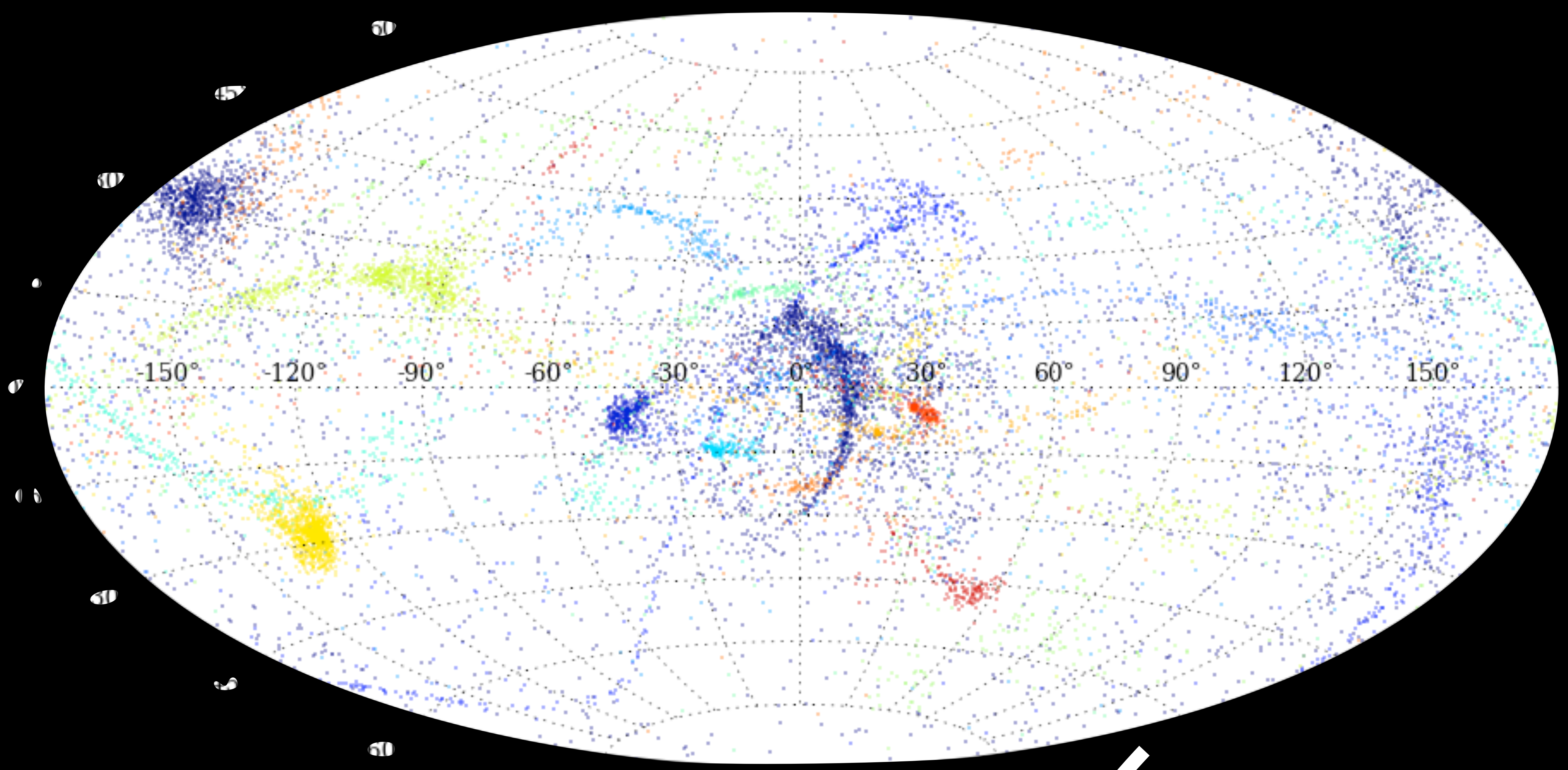
+ lots of assumptions

Sanderson, Helmi, & Hogg 2015

Sanderson et al. 2017a

The accreted stellar halo is clumpy in constants-of-motion space

Galactic coordinates

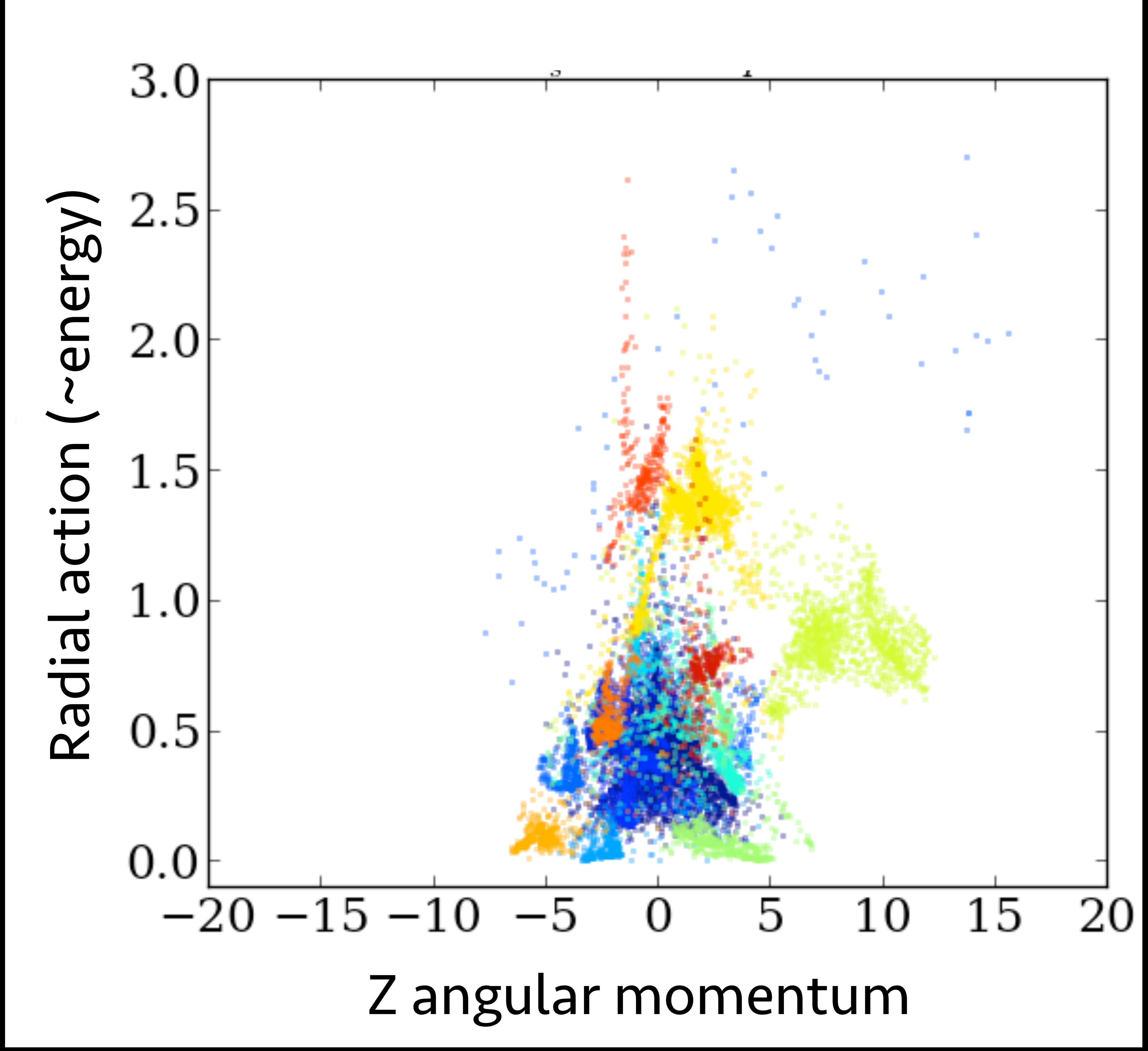


One particle = many stars

Base simulation: Cooper et al. 2010

Extremely nonlinear transformation

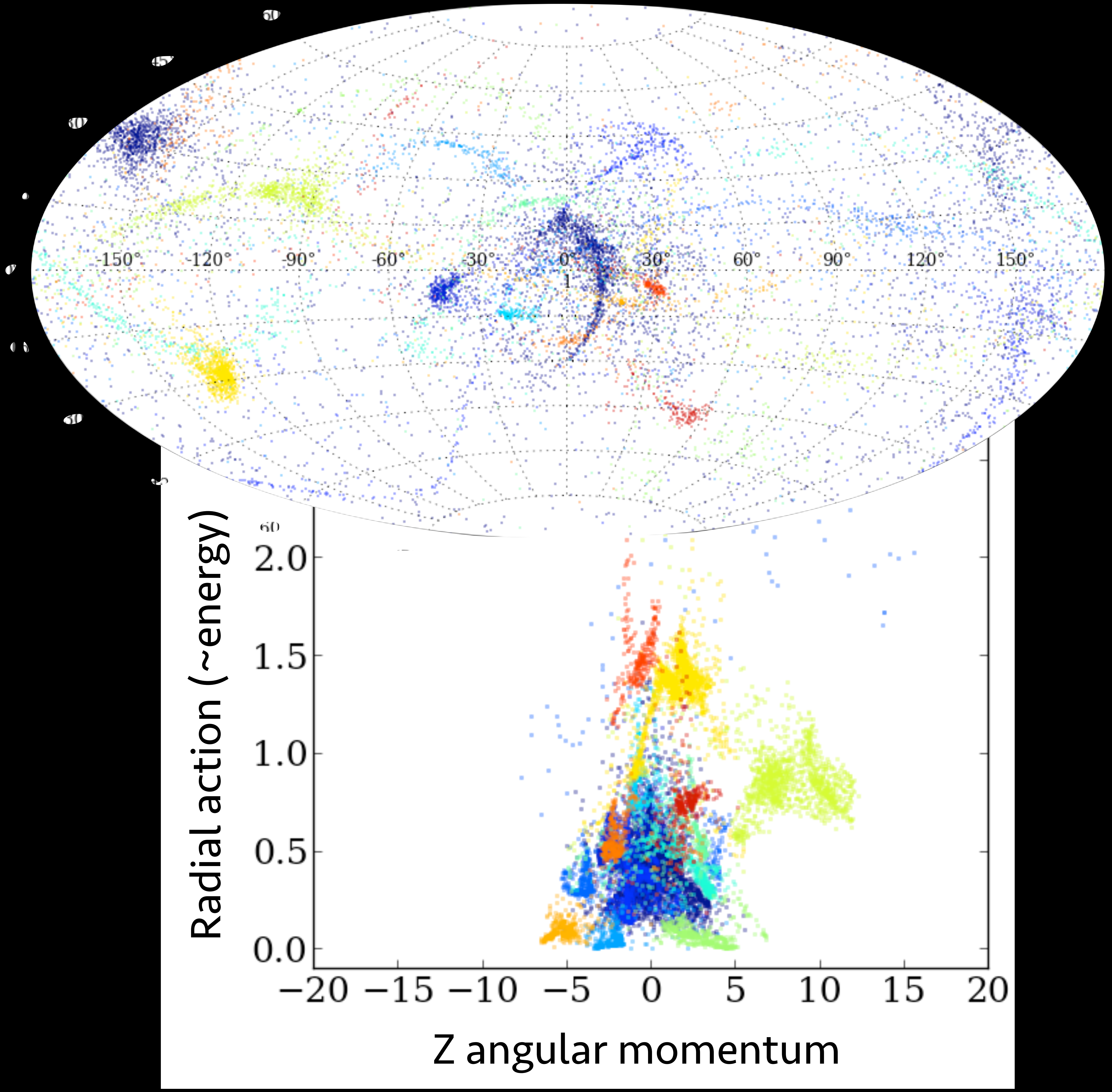
+ lots of assumptions



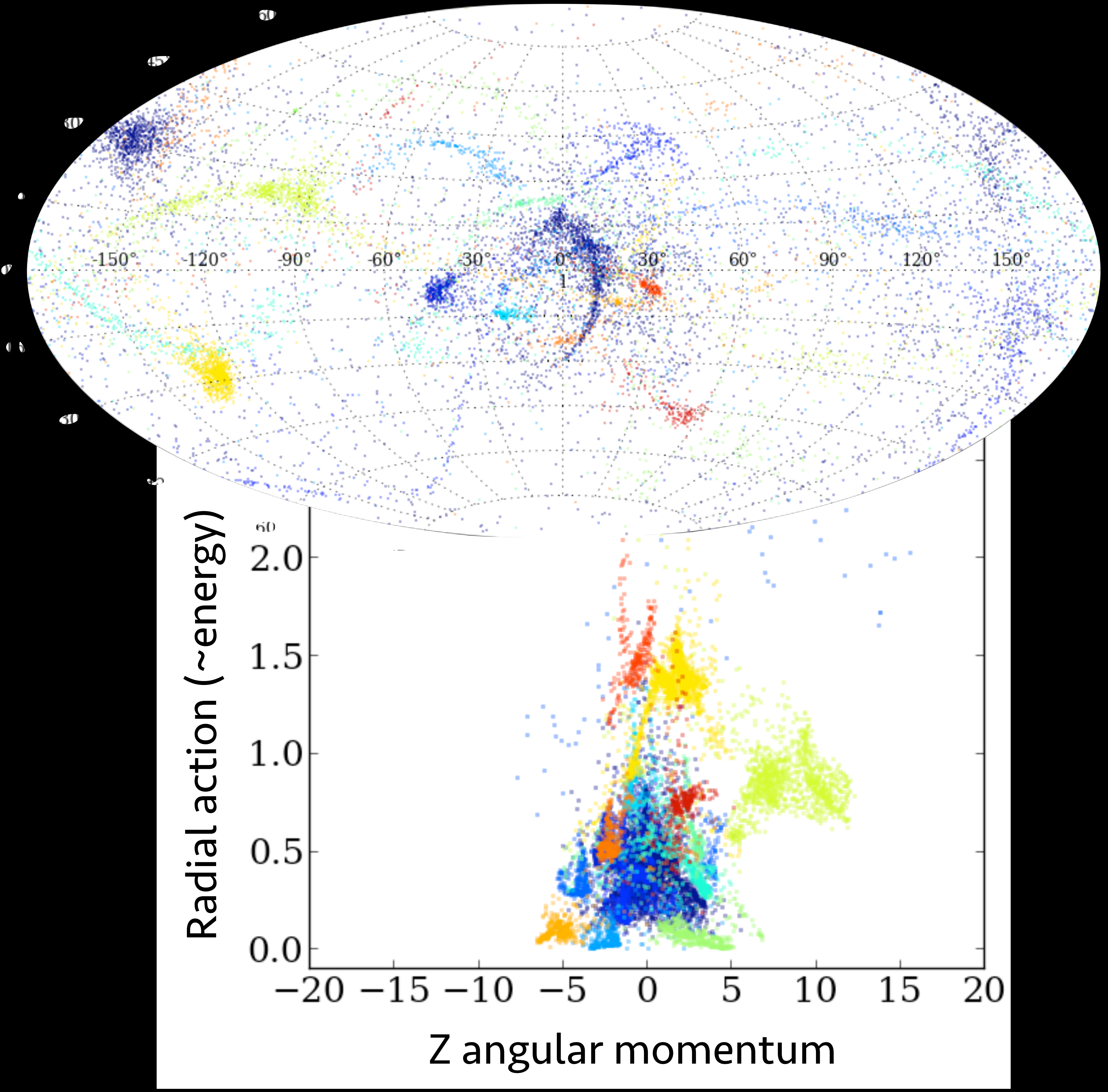
Constants of motion
(using best-fit mass model for host galaxy)

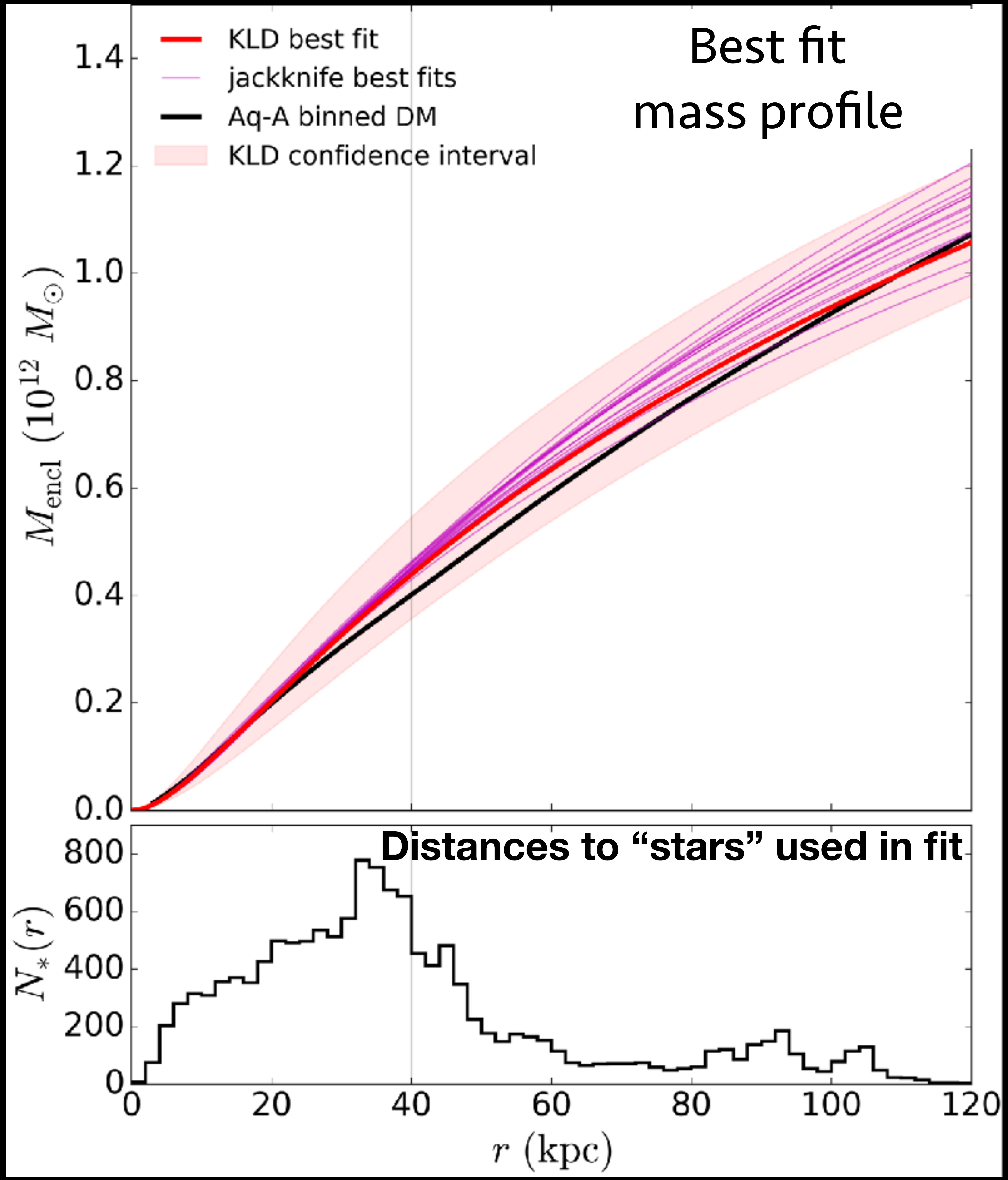
Sanderson, Helmi, & Hogg 2015
Sanderson et al. 2017a

The stellar halo constrains the MW's gravitational potential



The stellar halo constrains the MW's gravitational potential



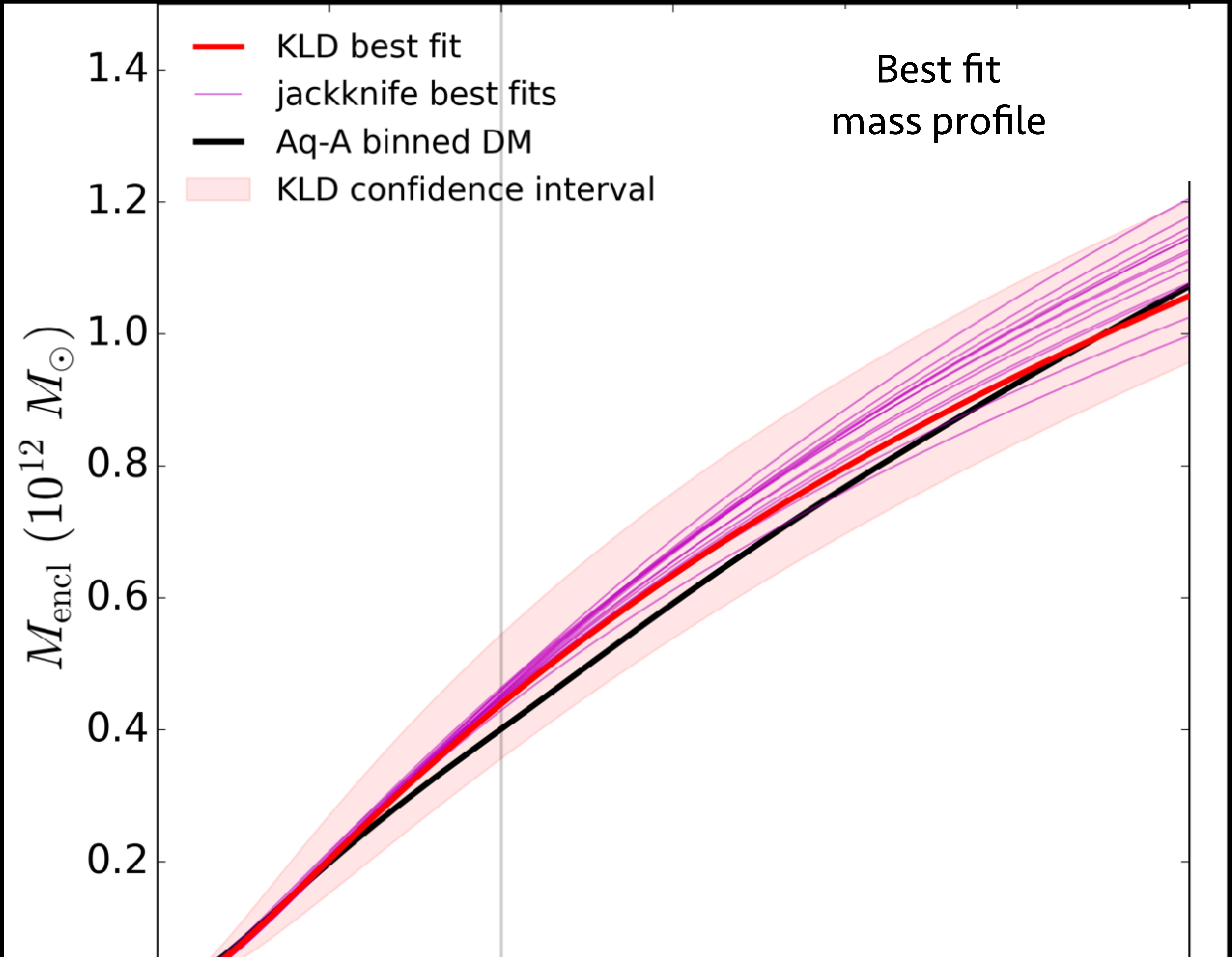


Sanderson et al. 2017a

With synthetic surveys, can test effect of the accretion history/selection of data

Variance when
leaving out one
stream at a time
<->
uncertainty on
best fit

With synthetic surveys, can test effect of the accretion history/selection of data

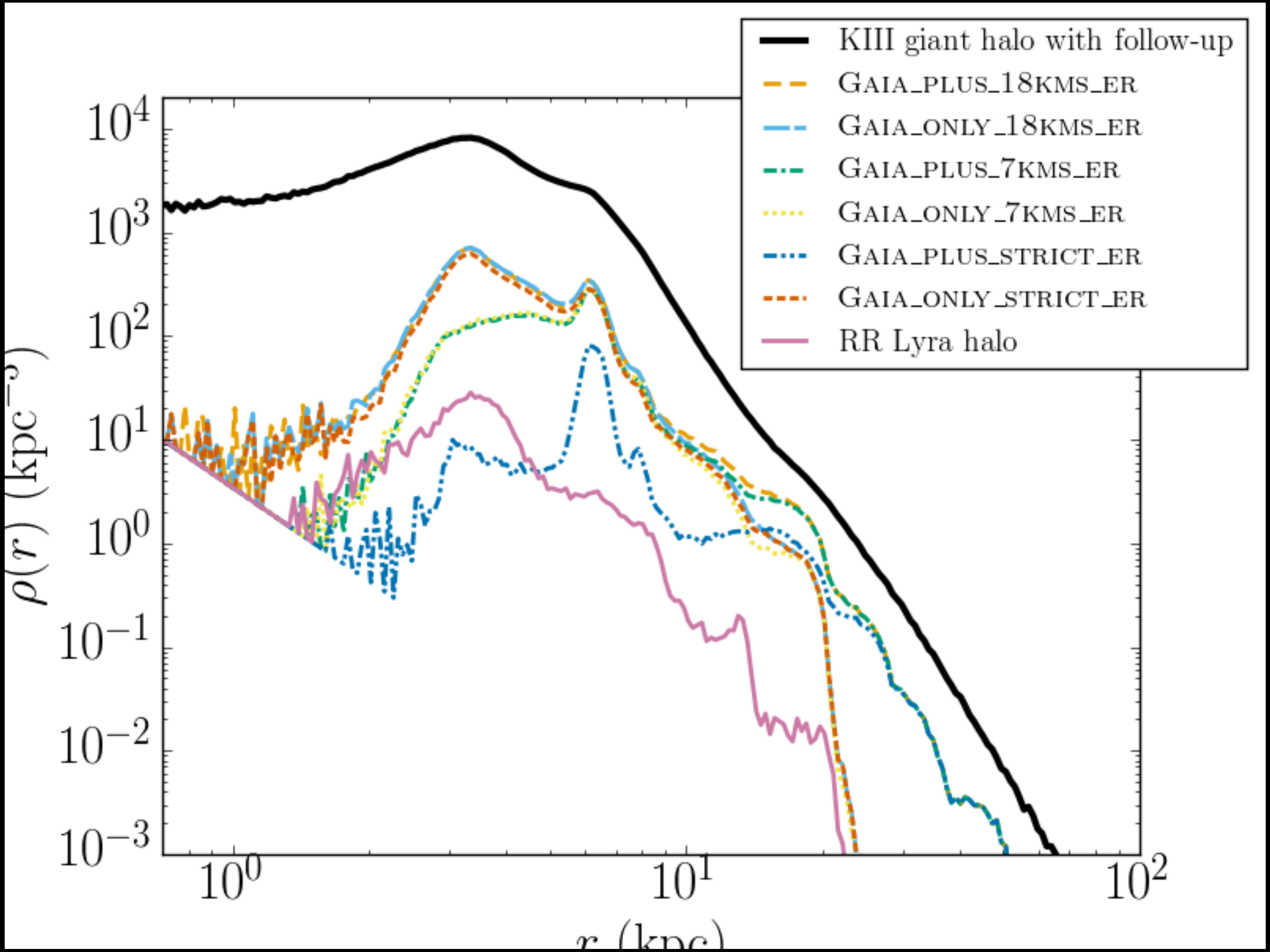


Variance when leaving out one stream at a time

<->

uncertainty on best fit

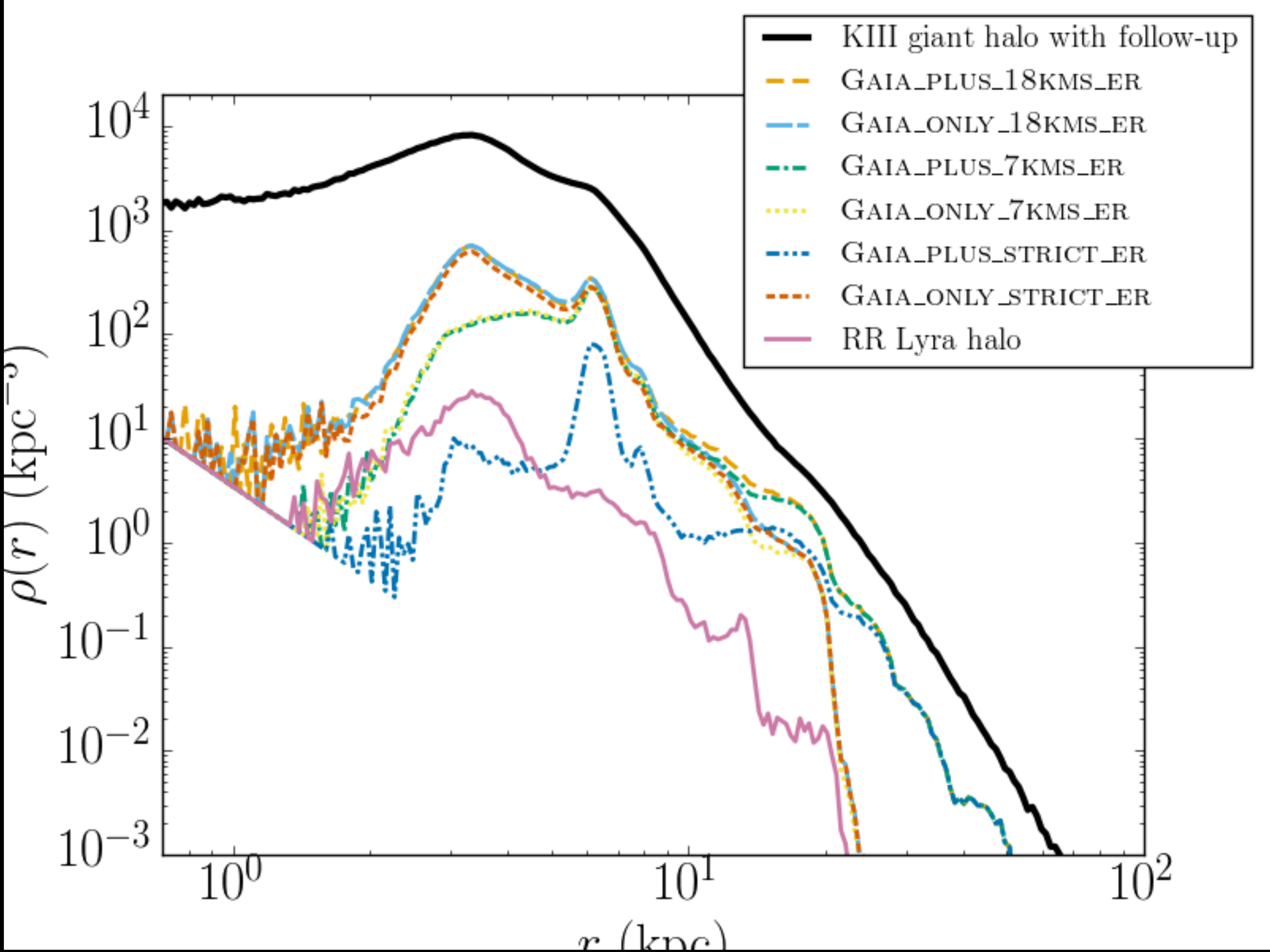
and how much observational systematics affect results

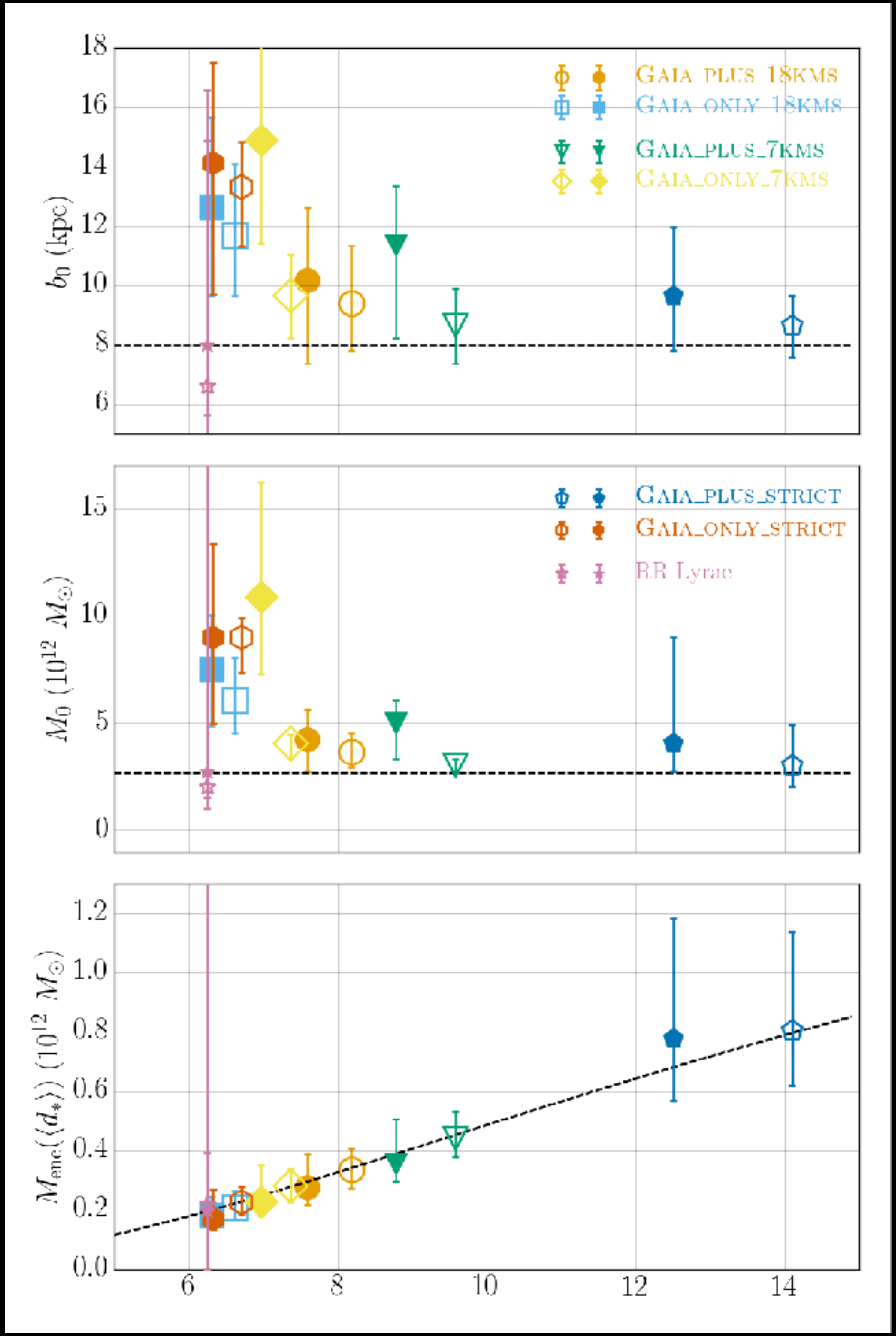


Sanderson 2017

Base simulation: Sanderson, Helmi, & Hogg 2015

and how much observational systematics affect results

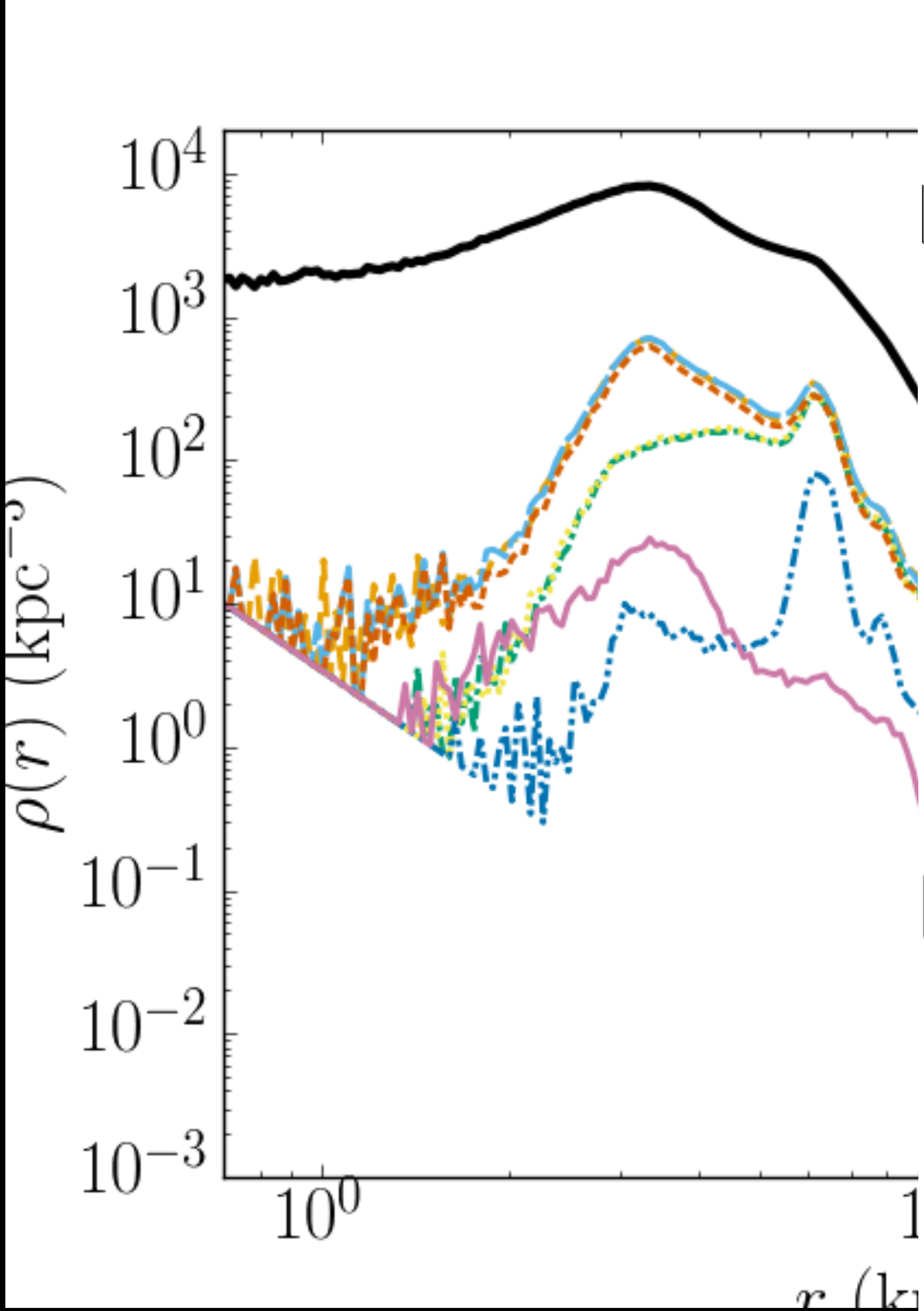


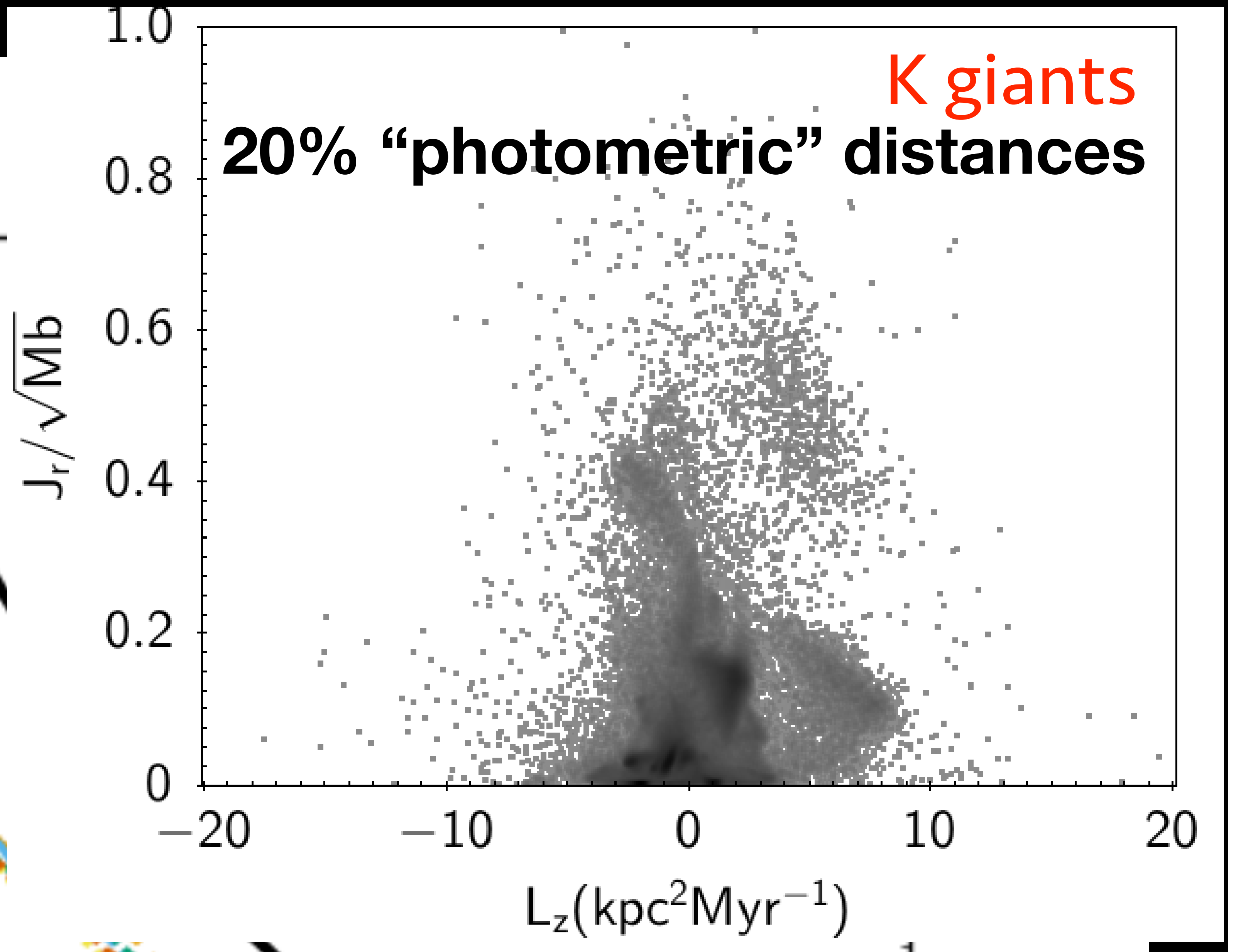


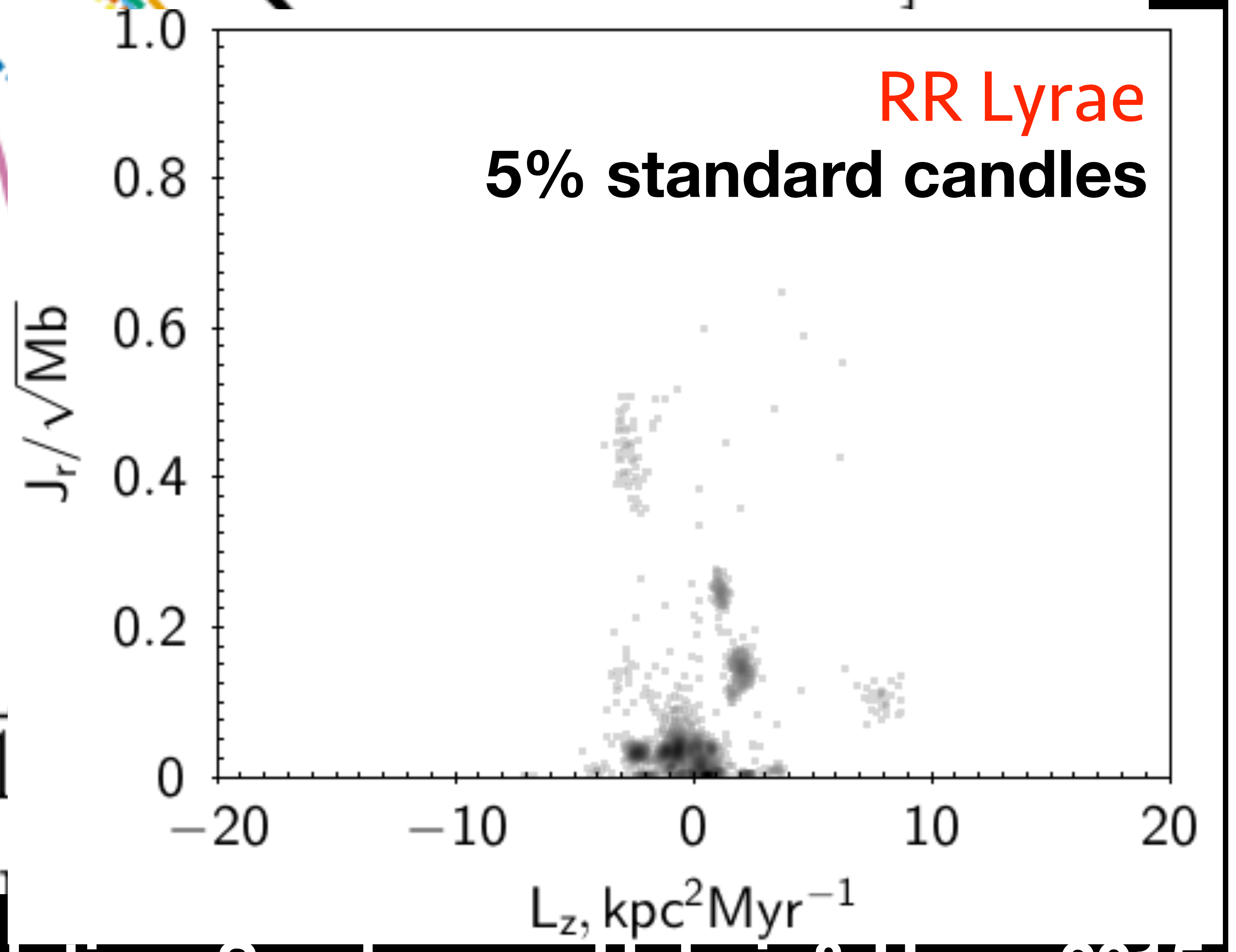
Sanderson 2017

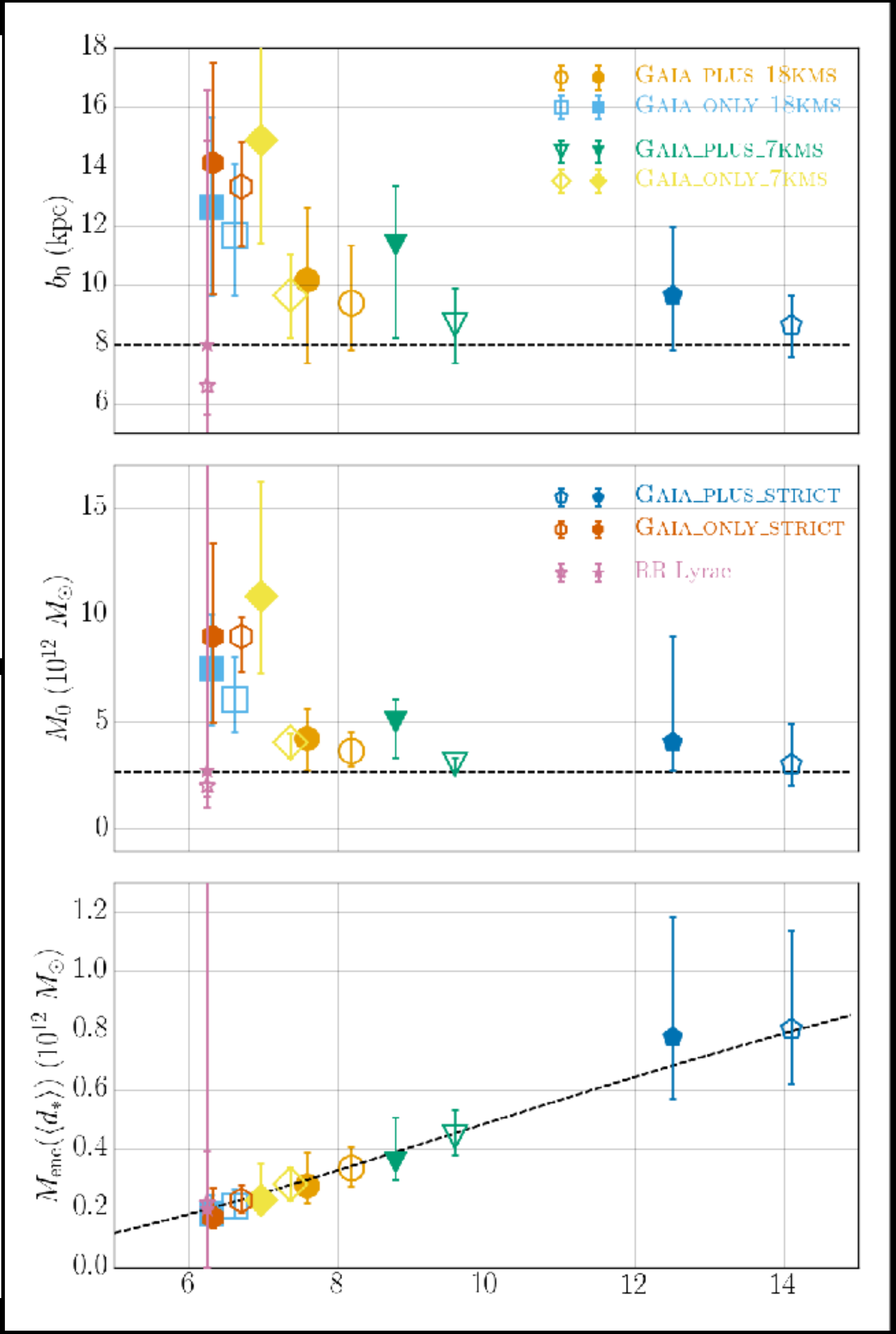
Base simulation: Sanderson, Helmi, & Hogg 2015

and how much observational systematics affect results









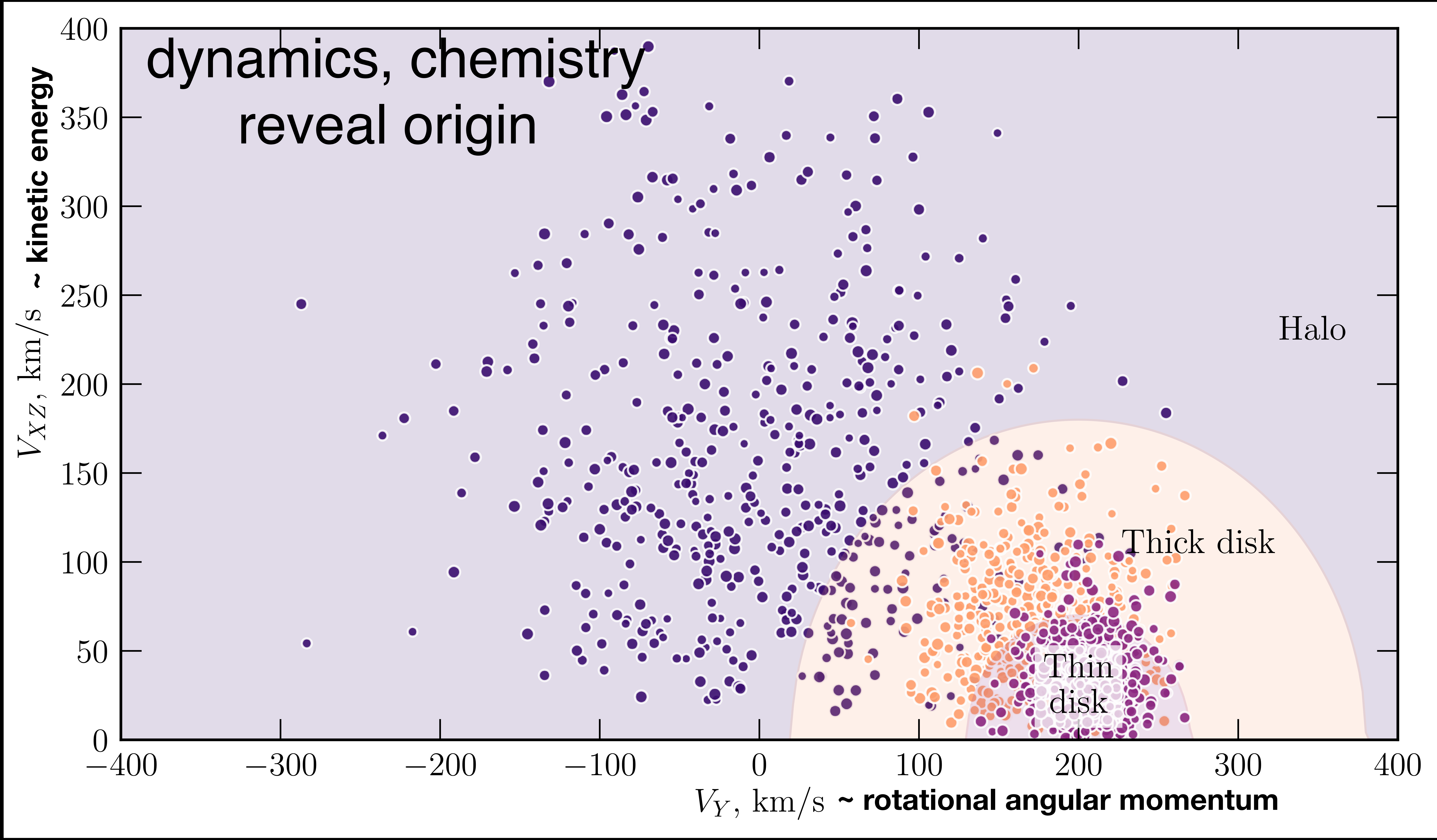
Sanderson 2017

Base simulation: Sanderson, Helmi, & Hogg 2015

outline

- how to make a synthetic survey
- how to use synthetic surveys:
  - making forecasts & planning survey strategies
  - evaluating the “selection function” of a search
  - validating complex analysis methods
  - interpreting structures identified by data-driven models
  - training ML methods for application to Gaia

Classical picture of the Solar neighborhood is informed by galaxy formation theory



Sanderson, Nikakhtar, Bonaca et al. in prep

But what does the data tell us directly about the number of distinct populations?

Test directly for number of components in Toomre+FeH space

...using a Gaussian mixture model

$$\mathcal{L}(V_{XZ}, V_Y, [\mathrm{Fe/H}] | \vec{\tau}, \{\vec{\mu}\}, \{\mathbf{\Sigma}\}) = \sum_{i=1}^{n_c} \tau_i \mathcal{N}(V_{XZ}, V_Y, [\mathrm{Fe/H}] | \vec{\mu}_i, \mathbf{\Sigma}_i)$$

Test directly for number of components in Toomre+FeH space

…using a Gaussian mixture model

Weights

$$\mathcal{L}(V_{XZ}, V_Y, [\mathrm{Fe/H}] | \vec{\tau}, \{\vec{\mu}\}, \{\boldsymbol{\Sigma}\}) = \sum_{i=1}^{n_c} \boxed{\tau_i} \mathcal{N}(V_{XZ}, V_Y, [\mathrm{Fe/H}] | \vec{\mu}_i, \boldsymbol{\Sigma}_i)$$

Test directly for number of components in Toomre+FeH space

...using a Gaussian mixture model

Weights

Means

$$\mathcal{L}(V_{XZ}, V_Y, [\mathrm{Fe/H}] | \vec{\tau}, \{\vec{\mu}\}, \{\mathbf{\Sigma}\}) = \sum_{i=1}^{n_c} \boxed{\tau_i} \mathcal{N}(V_{XZ}, V_Y, [\mathrm{Fe/H}] | \boxed{\vec{\mu}_i}, \mathbf{\Sigma}_i)$$

Test directly for number of components in Toomre+FeH space

...using a Gaussian mixture model

Weights

Means

Covariances

$$\mathcal{L}(V_{XZ}, V_Y, [\mathrm{Fe/H}] | \vec{\tau}, \{\vec{\mu}\}, \{\boldsymbol{\Sigma}\}) = \sum_{i=1}^{n_c} \tau_i \mathcal{N}(V_{XZ}, V_Y, [\mathrm{Fe/H}] | \vec{\mu}_i, \boldsymbol{\Sigma}_i)$$

Test directly for number of components in Toomre+FeH space

...using a Gaussian mixture model

Weights

Means

Covariances

Number of Components

$$\mathcal{L}(V_{XZ}, V_Y, [\mathrm{Fe/H}] | \vec{\tau}, \{\vec{\mu}\}, \{\mathbf{\Sigma}\}) = \sum_{i=1}^{\boxed{n_c}} \boxed{\tau_i} \mathcal{N}(V_{XZ}, V_Y, [\mathrm{Fe/H}] | \boxed{\vec{\mu}_i}, \boxed{\mathbf{\Sigma}_i})$$

Test directly for number of components in Toomre+FeH space

...using a Gaussian mixture model

Weights

Means

Covariances

Number of Components

$$\mathcal{L}(V_{XZ}, V_Y, [\mathrm{Fe/H}] | \vec{\tau}, \{\vec{\mu}\}, \{\mathbf{\Sigma}\}) = \sum_{i=1}^{\boxed{n_c}} \boxed{\tau_i} \mathcal{N}(V_{XZ}, V_Y, [\mathrm{Fe/H}] | \boxed{\vec{\mu}_i}, \boxed{\mathbf{\Sigma}_i})$$

Penalty for adding more components

Likelihood of best-fit model with k params

$$\textbf{BIC} = k \ln N_* - 2 \ln \hat{\mathcal{L}}$$

$$k = \frac{\boxed{n_c}(n_f^2 + 3n_f + 2)}{2} - 1$$

dimensionality of data (in this case 3)

Test directly for number of components in Toomre+FeH space

...using a Gaussian mixture model

Weights

Means

Covariances

Number of Components

$$\mathcal{L}(V_{XZ}, V_Y, [\mathrm{Fe/H}] | \vec{\tau}, \{\vec{\mu}\}, \{\Sigma\}) = \sum_{i=1}^{n_c} \tau_i \mathcal{N}(V_{XZ}, V_Y, [\mathrm{Fe/H}] | \vec{\mu}_i, \Sigma_i)$$

Penalty for adding more components

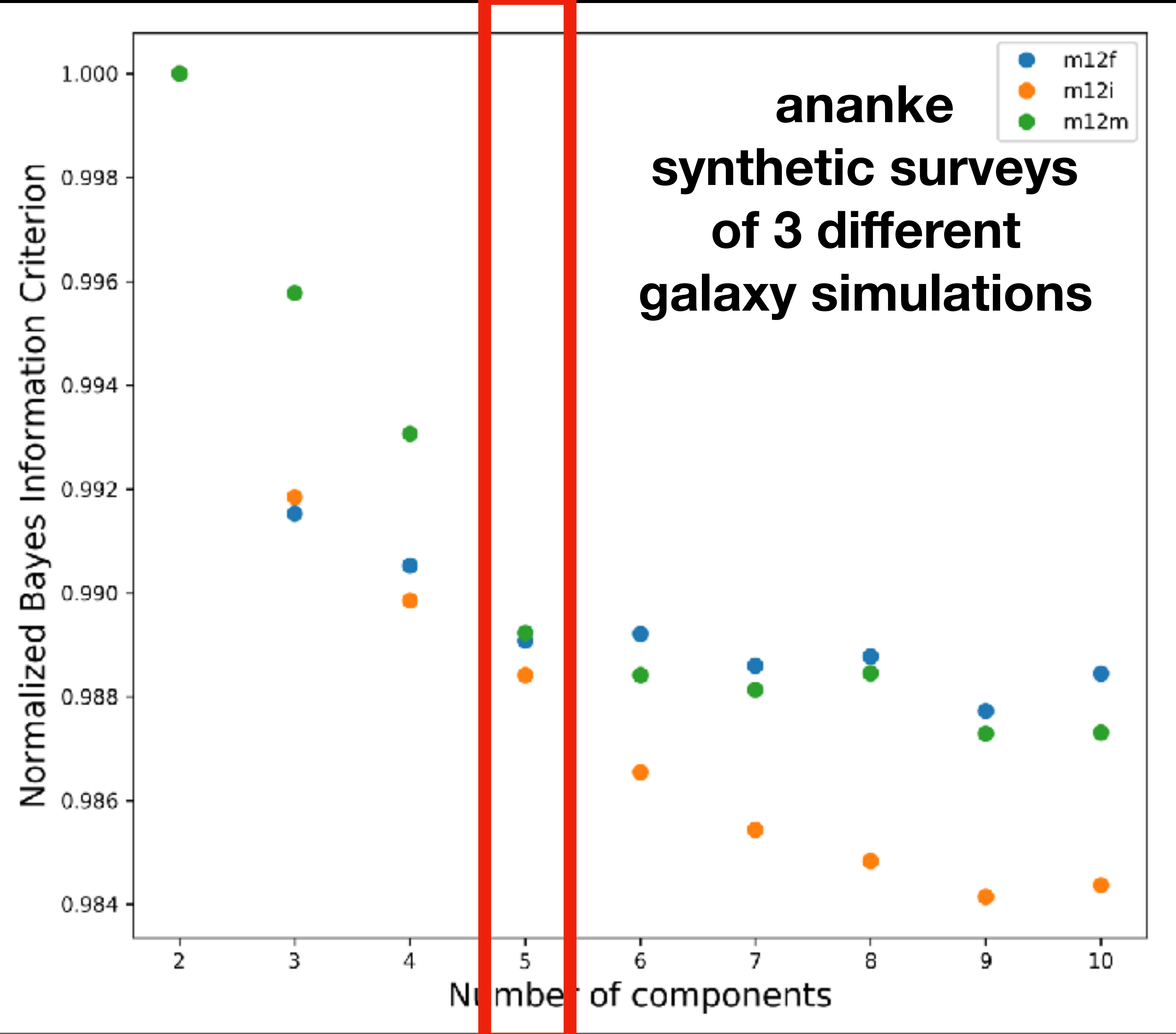
Likelihood of best-fit model with k params

$$\textbf{BIC} = k \ln N_* - 2 \ln \hat{\mathcal{L}}$$

$$k = \frac{n_c (n_f^2 + 3 n_f + 2)}{2} - 1$$

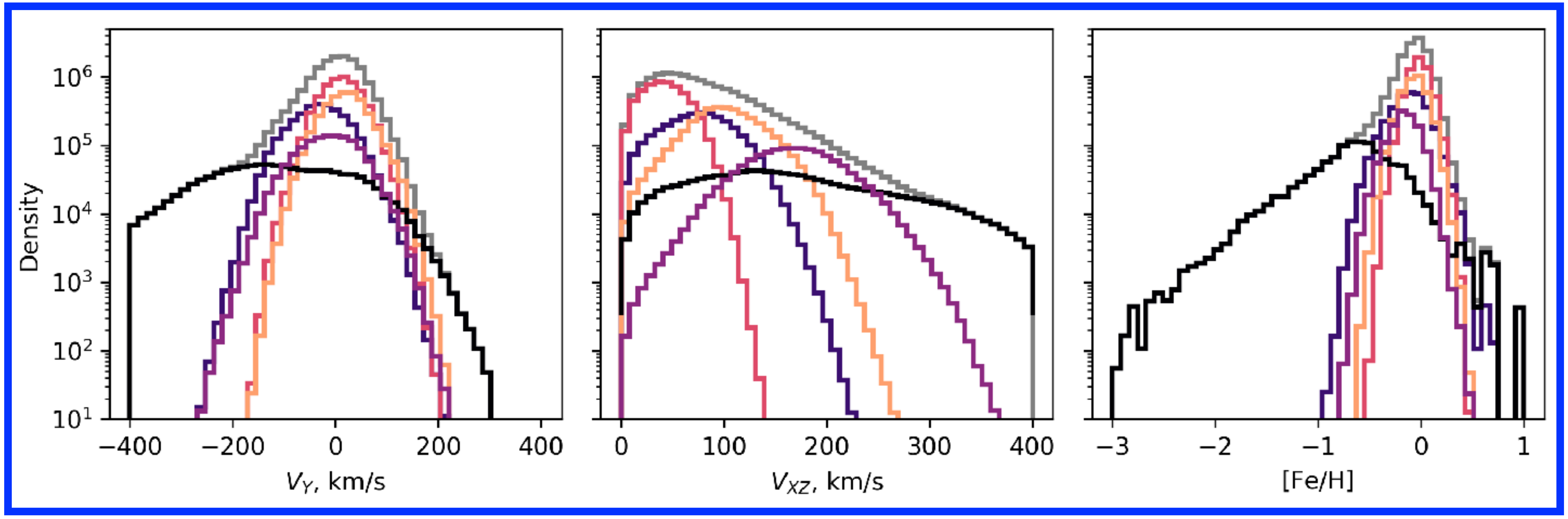
dimensionality of data (in this case 3)

Lower is Better



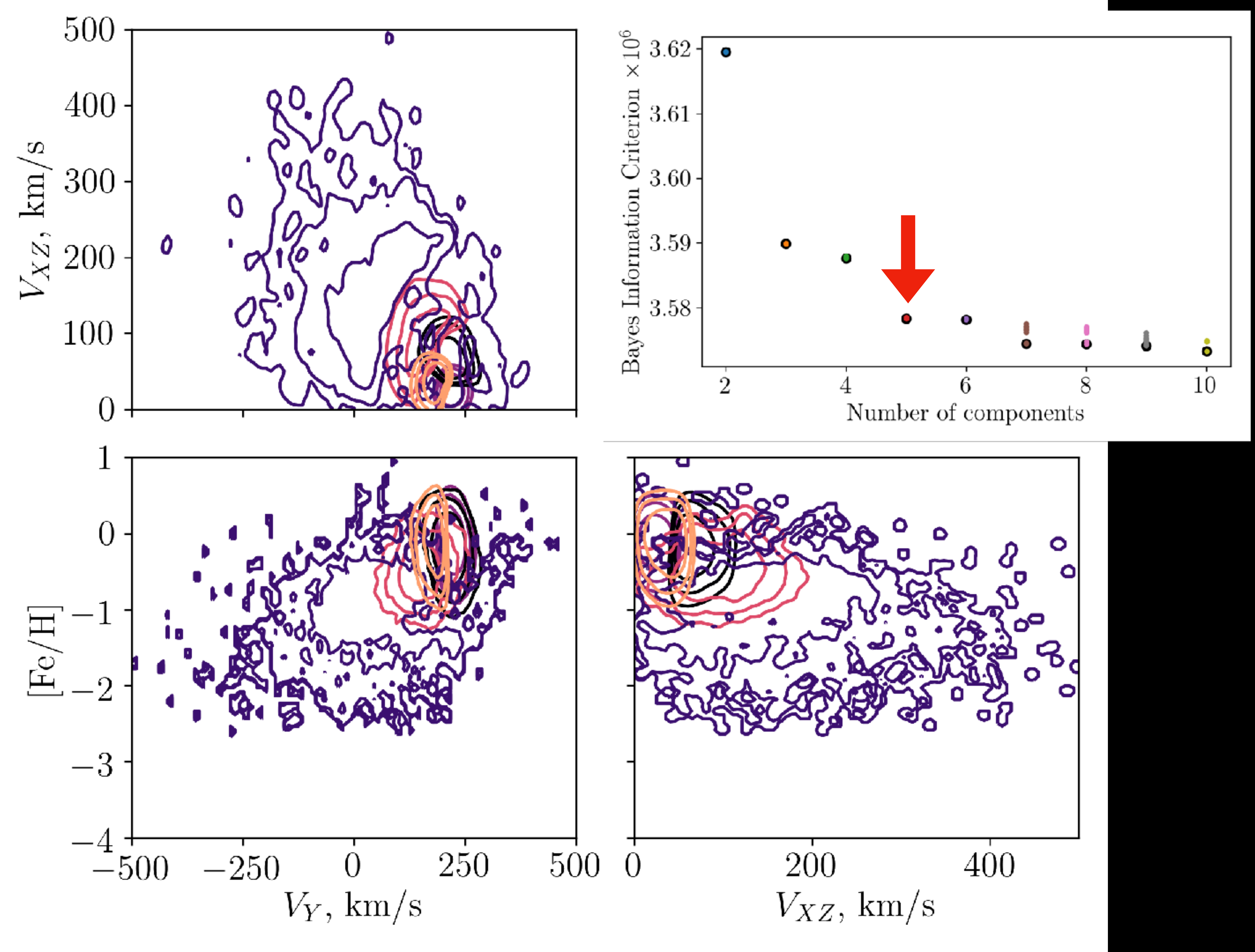
Components are recognizable in Toomre space

And correspond to notions of galaxy-formation theory



And correspond to notions of galaxy-formation theory

Mock data helps interpret the real data (in this case, RAVE-TGAS)

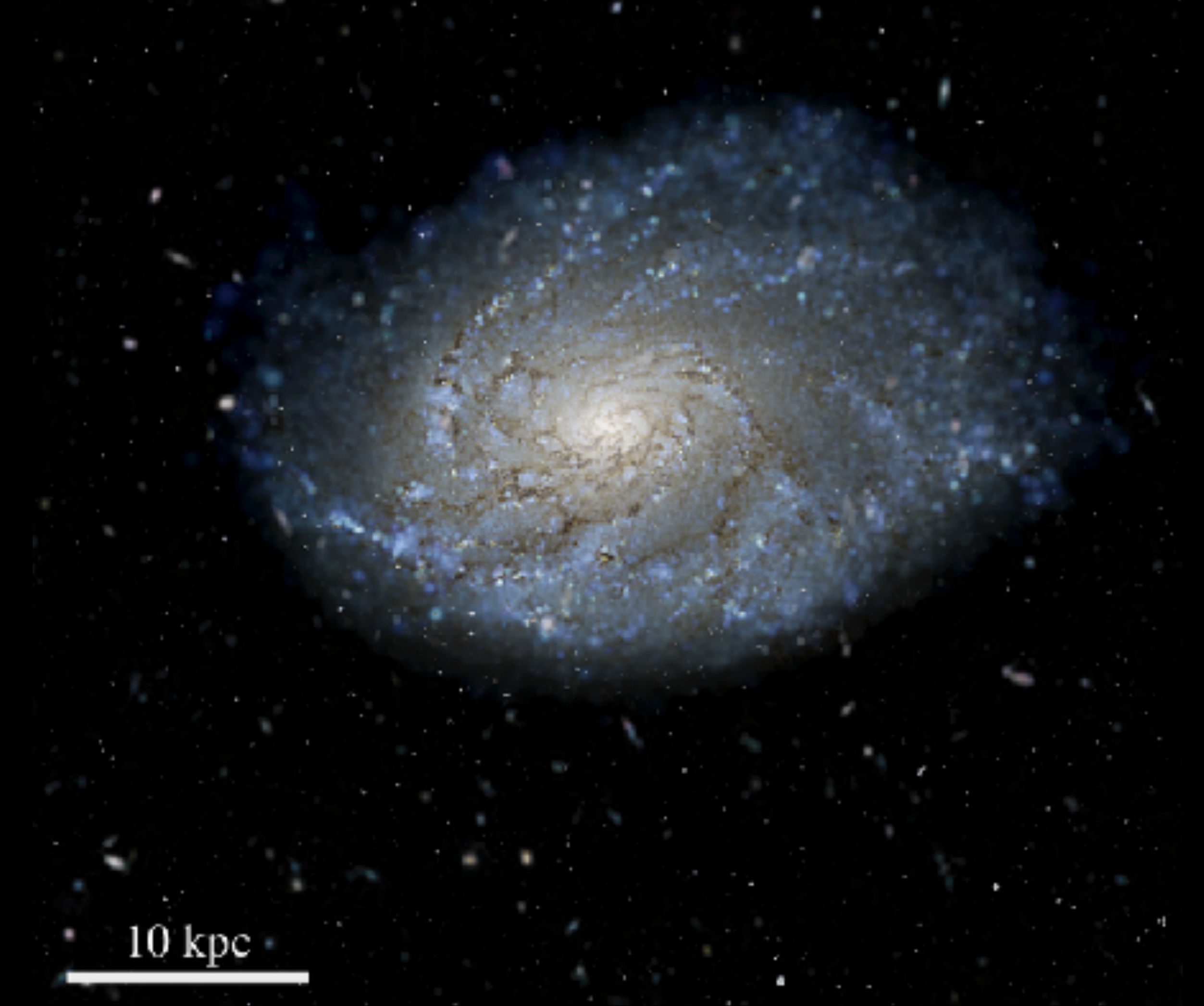


outline

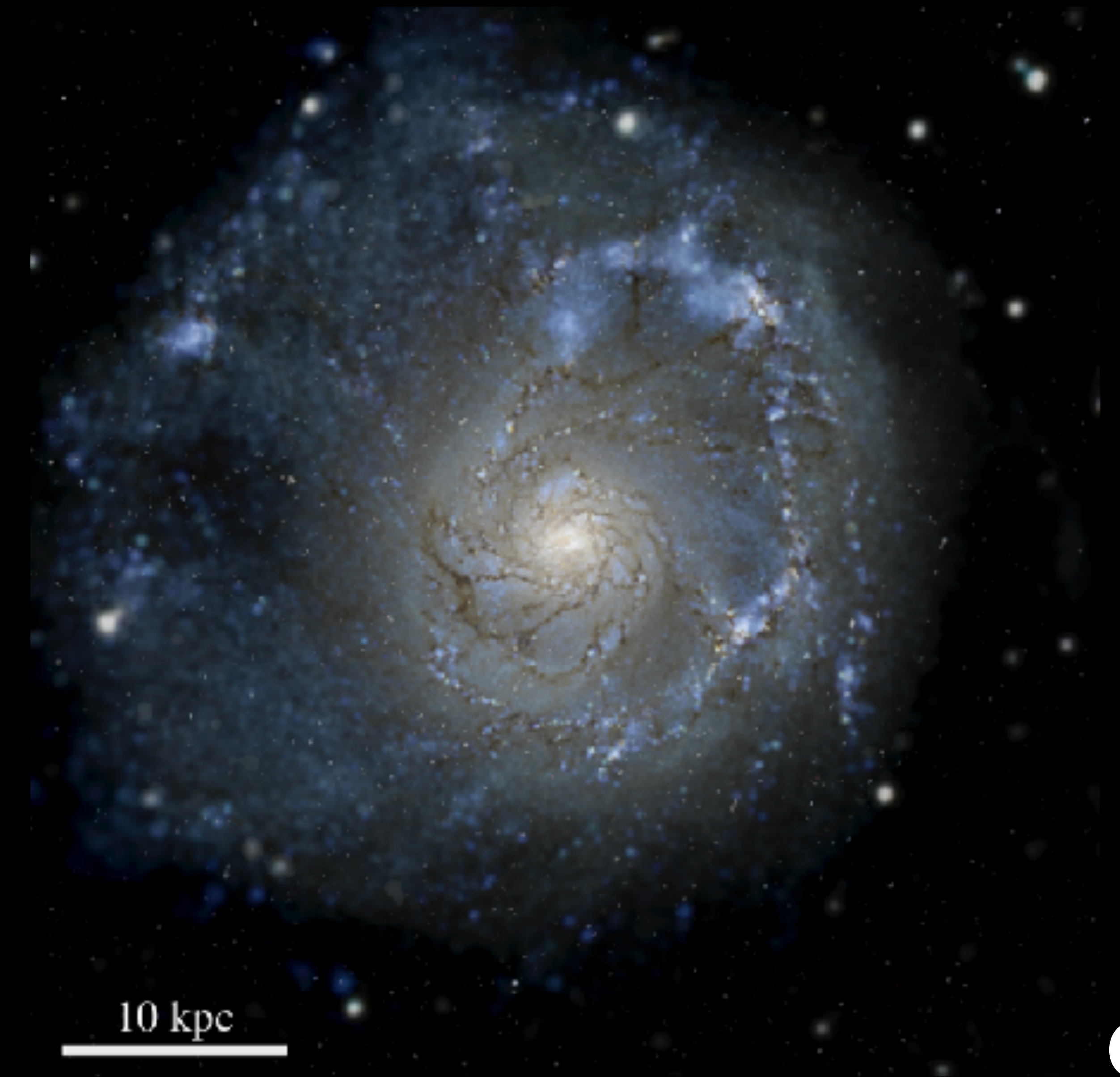
- how to make a synthetic survey
- how to use synthetic surveys:
  - making forecasts & planning survey strategies
  - evaluating the “selection function” of a search
  - validating complex analysis methods
  - interpreting structures identified by data-driven models
  - training ML methods for application to Gaia

Can we tell accreted stars from those formed in situ?

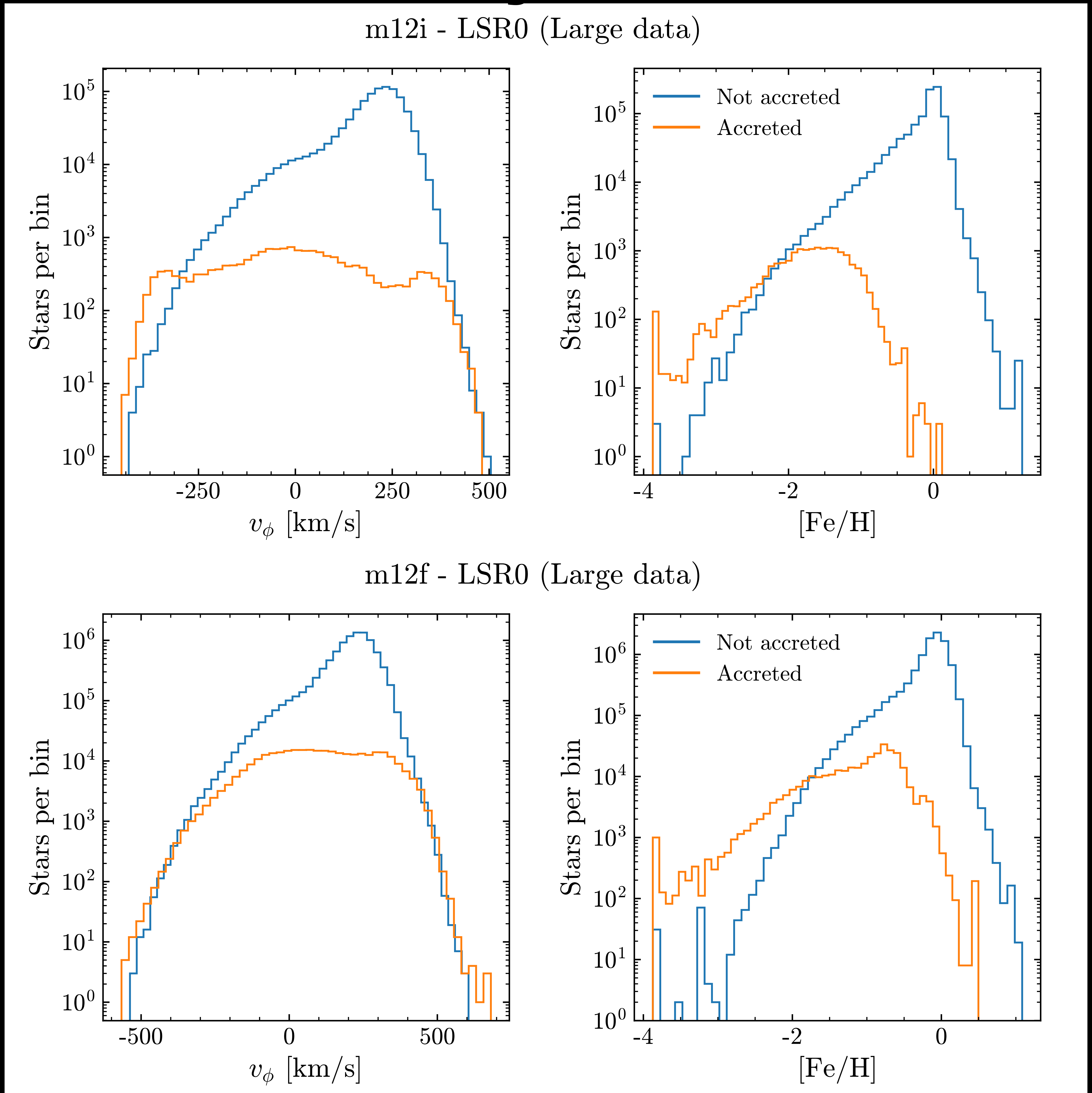
simulated galaxy m12i



simulated galaxy m12f



Distributions differ in many dimensions



Gaia does not measure either quantity for all stars in its range

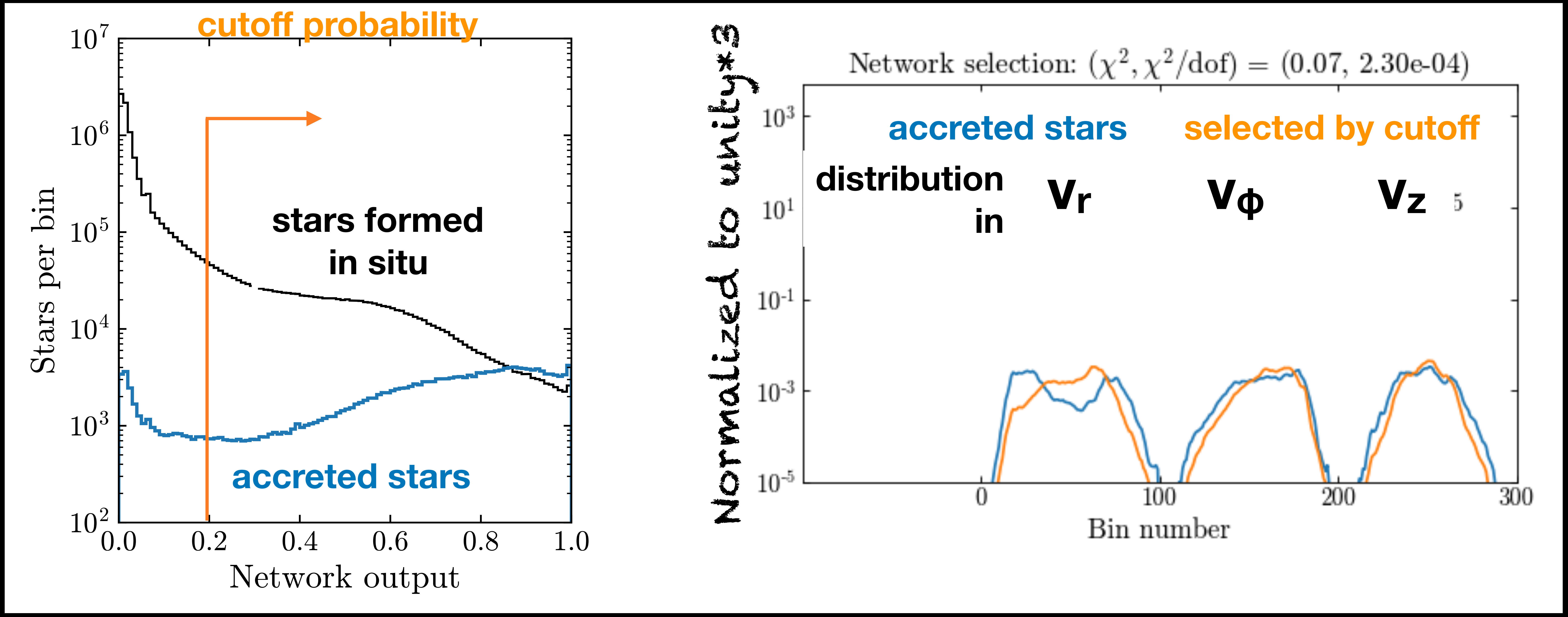
Project led by:

Bryan Ostdiek, Oregon -> Harvard

Lina Necib, Caltech

Train a machine learning algorithm to find accreted stars in Gaia

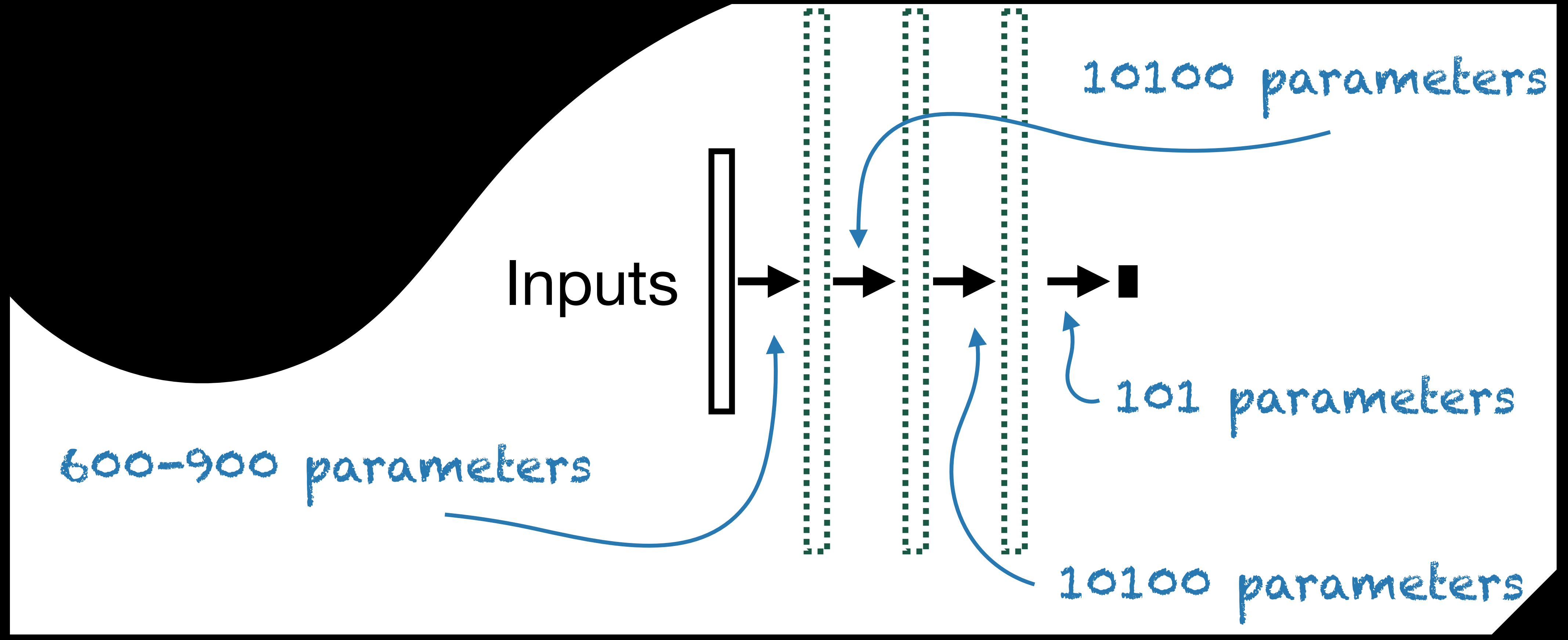
(see also earlier work by Veljanoski, Breddels, & Helmi on gradient boosted trees)



Use one synthetic survey to choose a cutoff value that recovers the velocity distributions of the accreted component

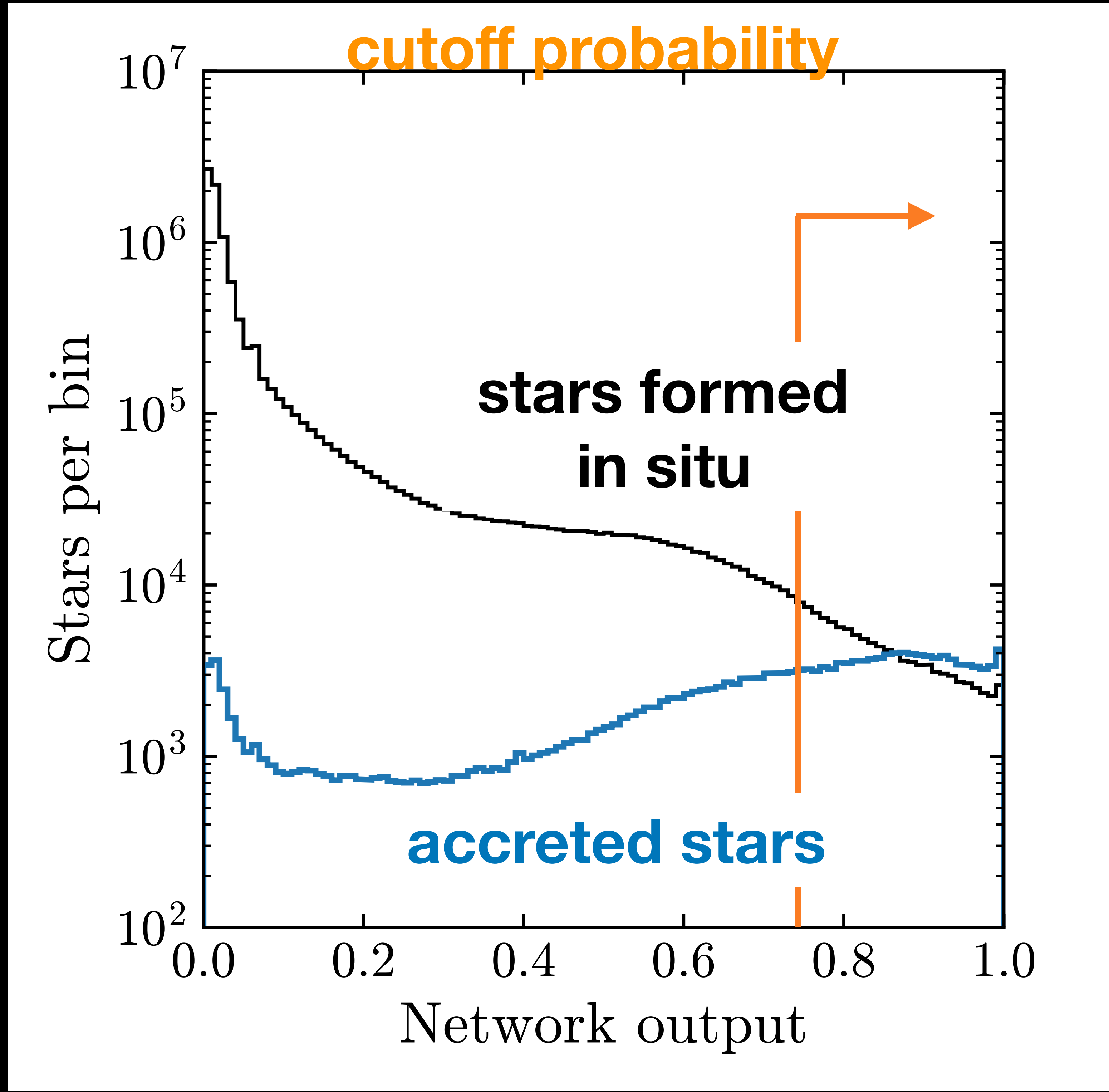
The ML classifier assigns a probability that each star is accreted.

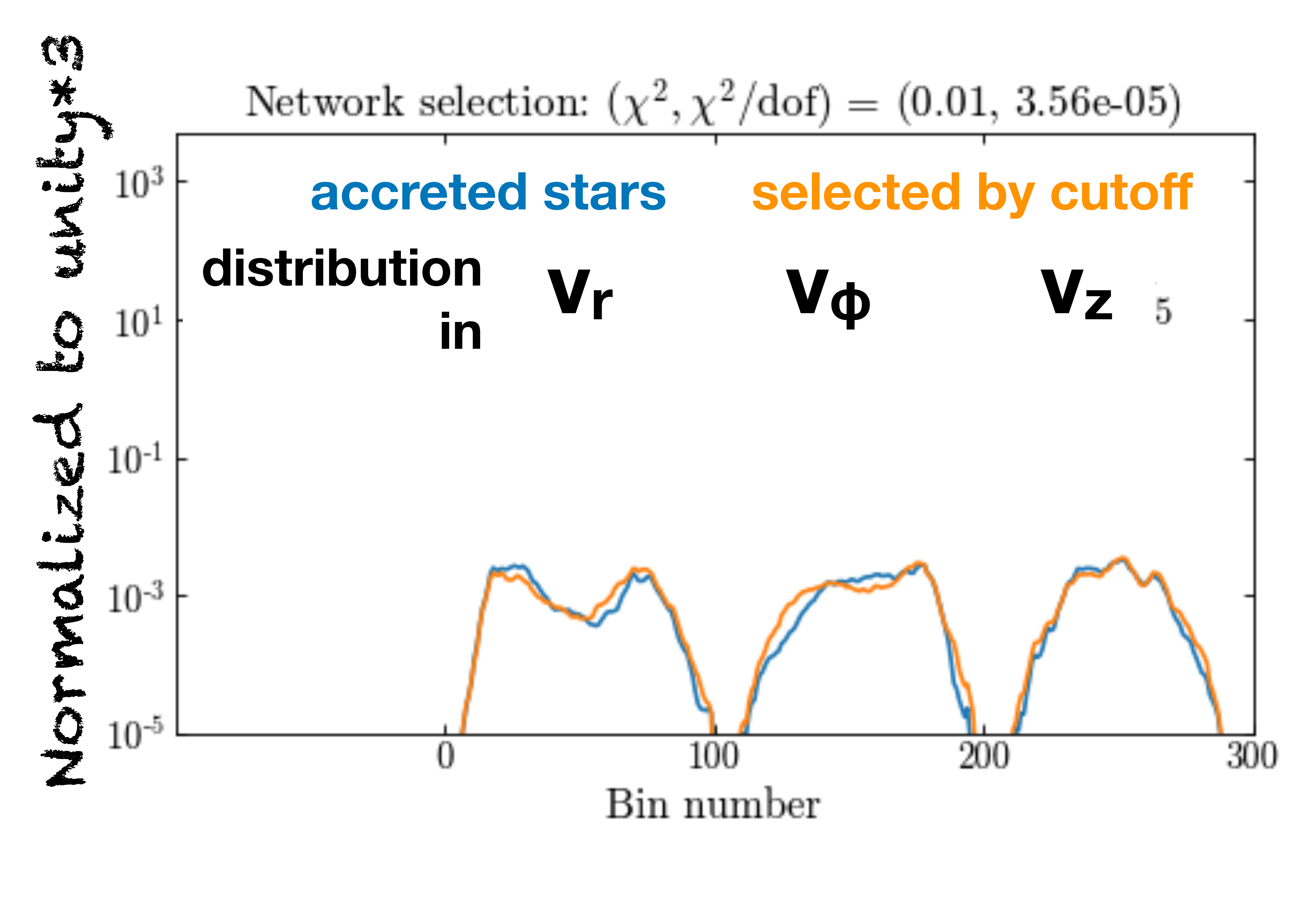
Inputs are 5-dimensional Gaia-like kinematics (parallax quality cut, no RVs)



Train a machine learning algorithm to find accreted stars in Gaia

(see also earlier work by Veljanoski, Breddels, & Helmi on gradient boosted trees)

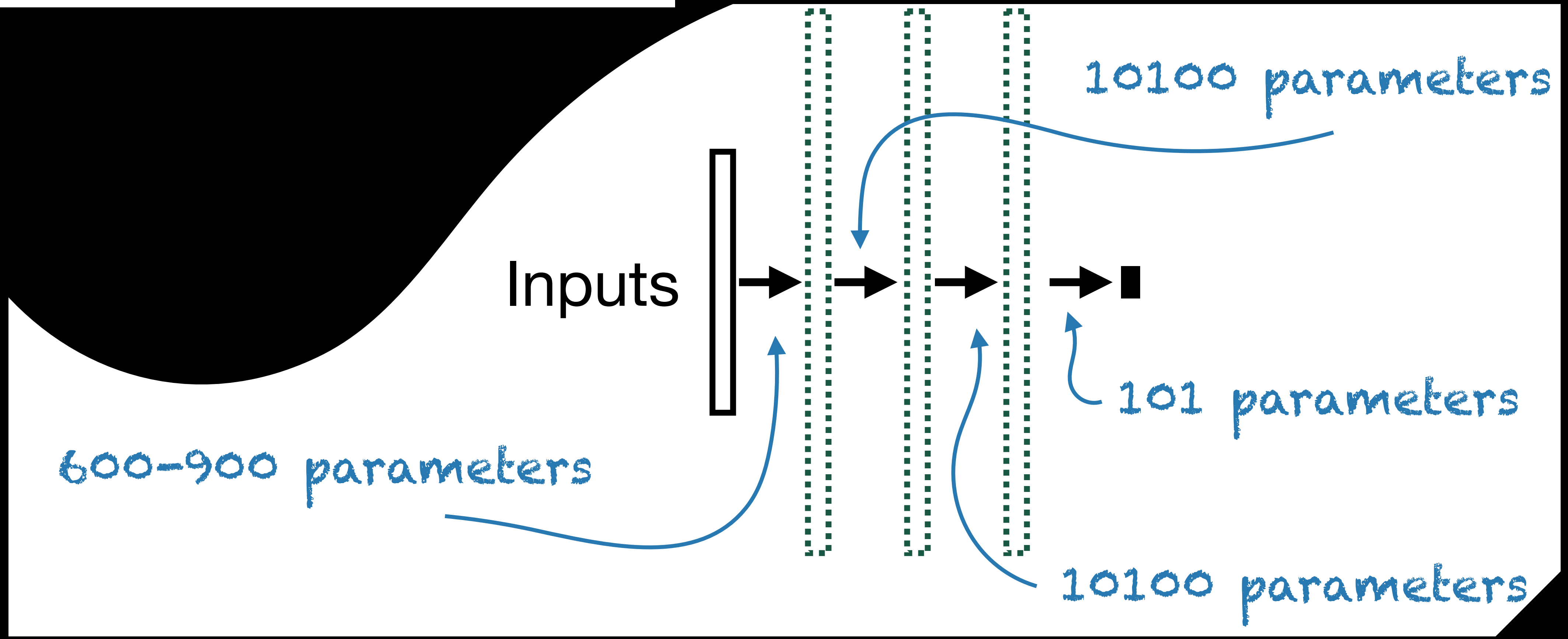




Use one synthetic survey to choose a cutoff value that recovers the velocity distributions of the accreted component

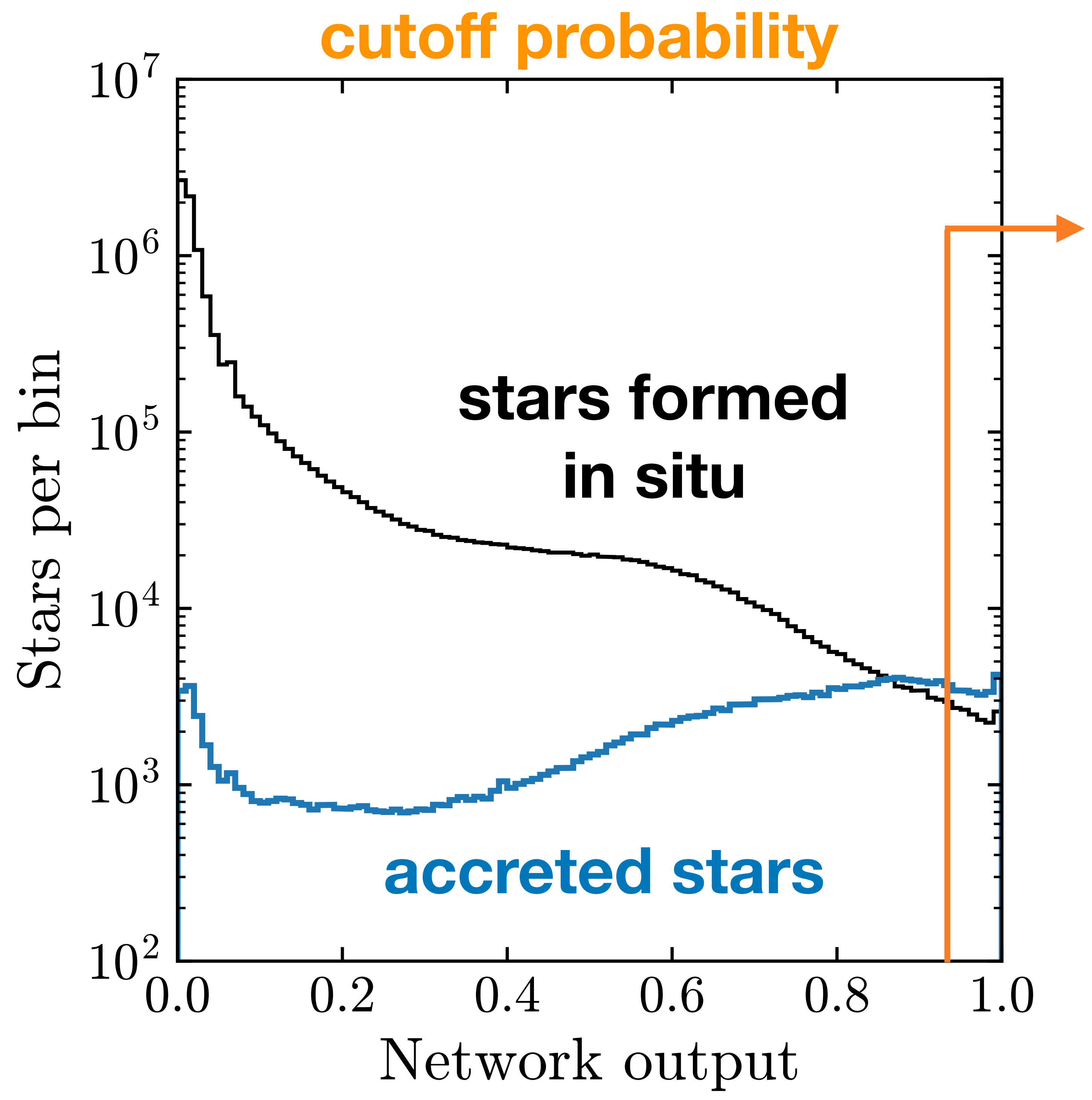
The ML classifier assigns a probability that each star is accreted.

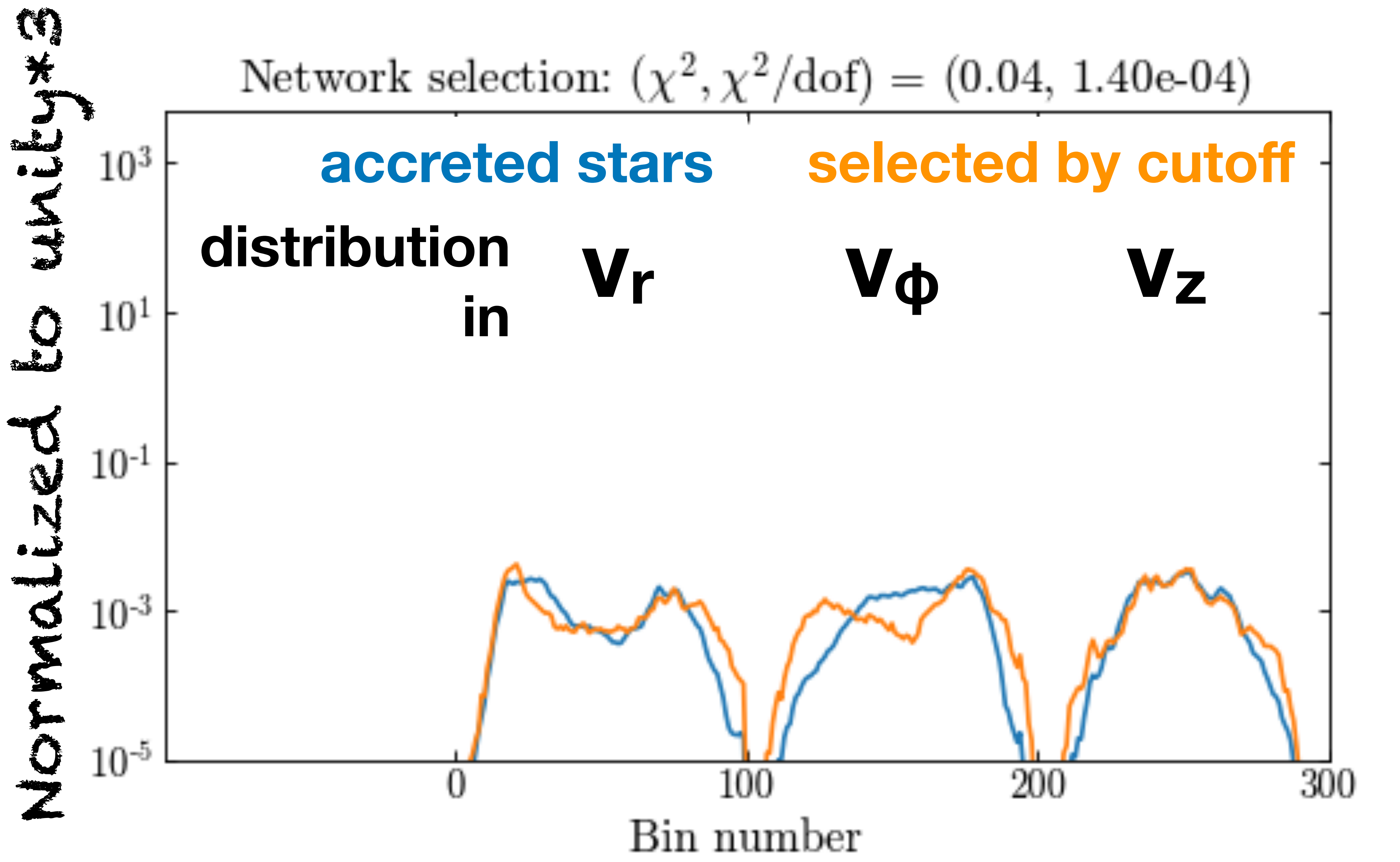
Inputs are 5-dimensional Gaia-like kinematics (parallax quality cut, no RVs)



Train a machine learning algorithm to find accreted stars in Gaia

(see also earlier work by Veljanoski, Breddels, & Helmi on gradient boosted trees)

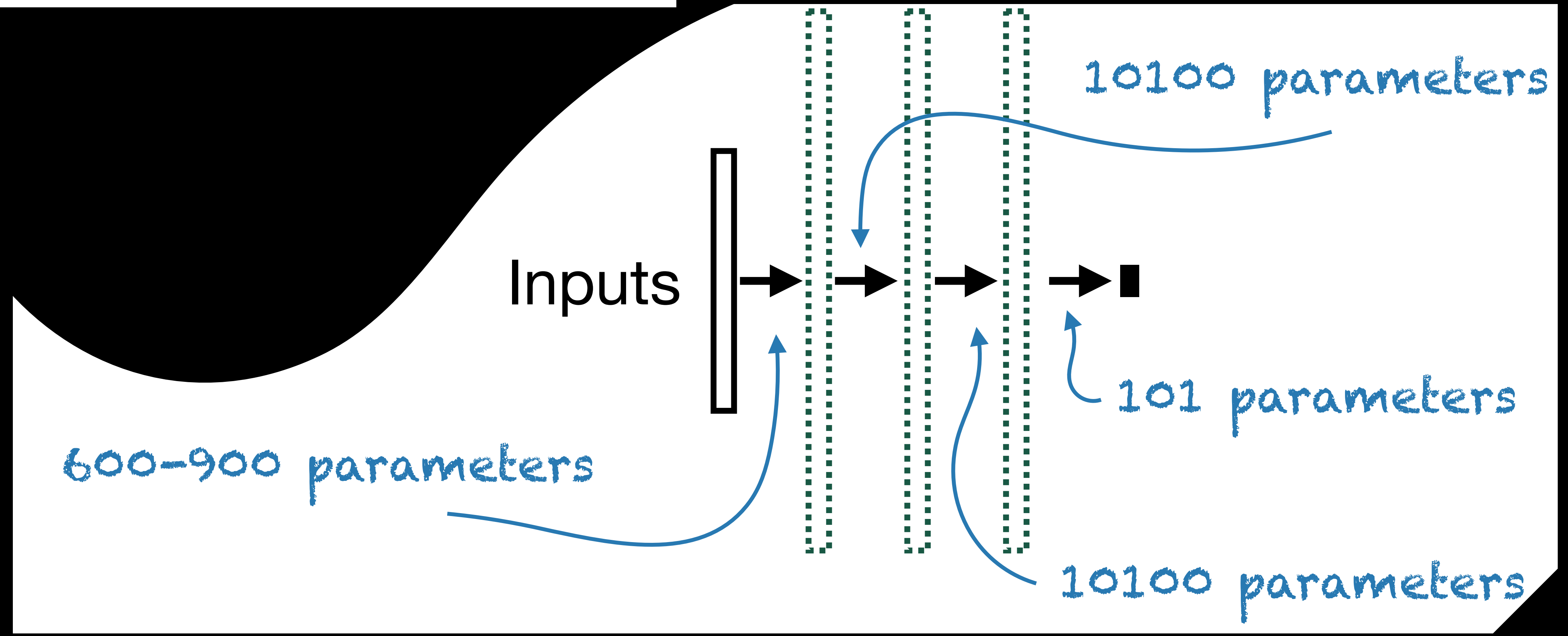




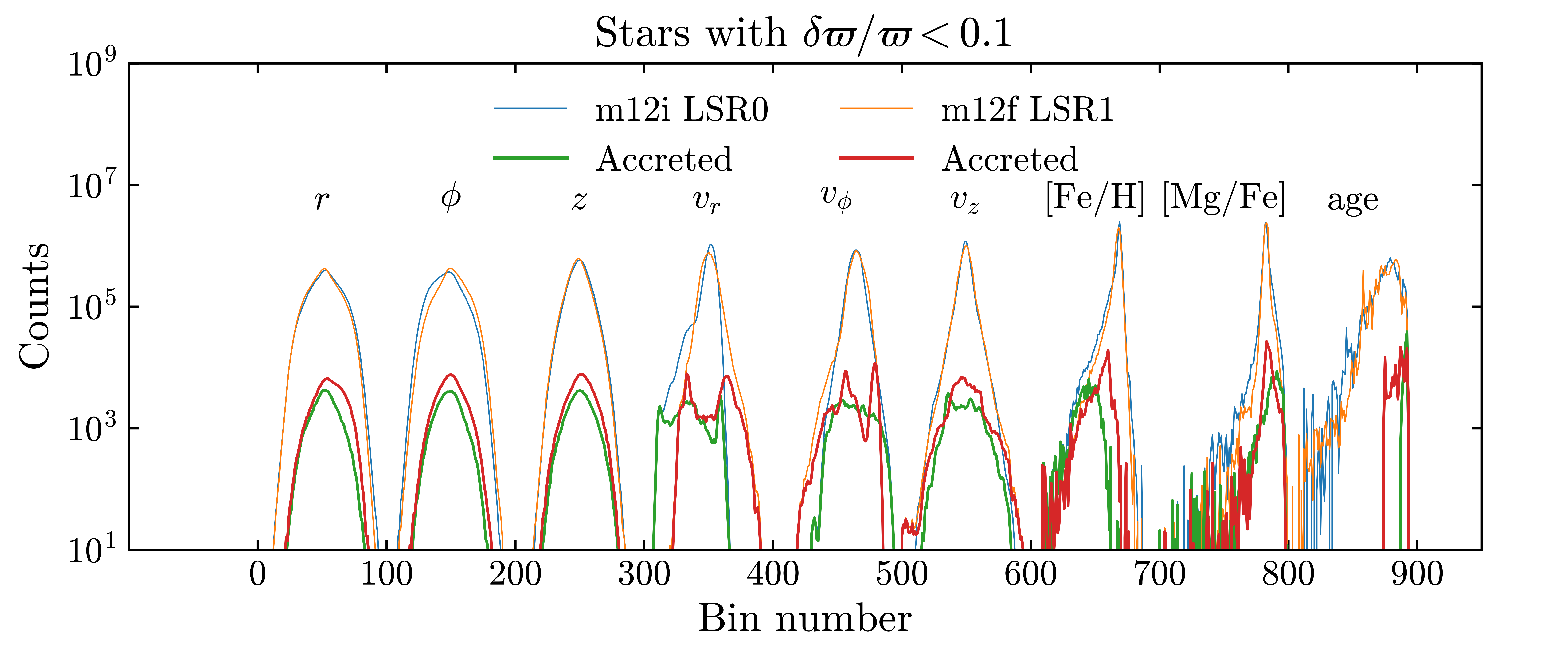
Use one synthetic survey to choose a cutoff value that recovers the velocity distributions of the accreted component

The ML classifier assigns a probability that each star is accreted.

Inputs are 5-dimensional Gaia-like kinematics (parallax quality cut, no RVs)



It’s not obvious that the calibration will apply to another galaxy



Trained on this one

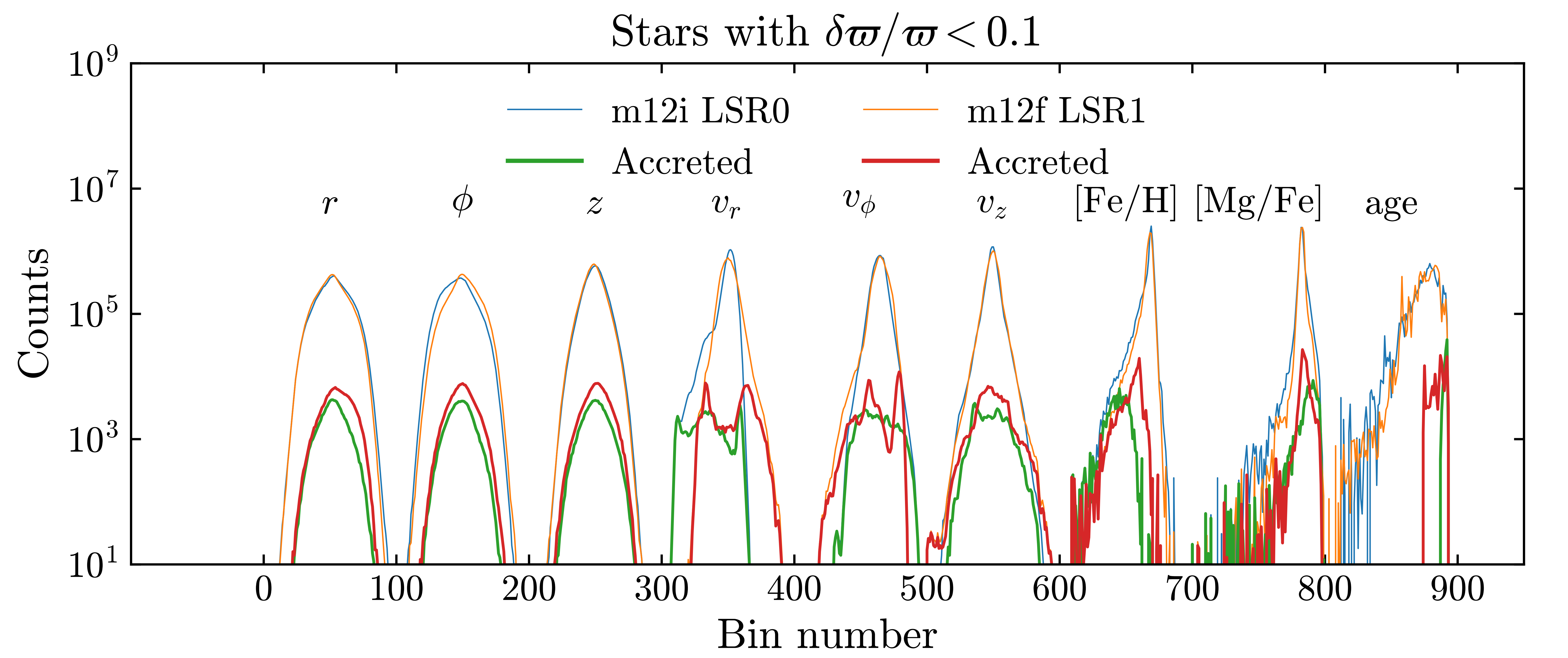
Test on this one

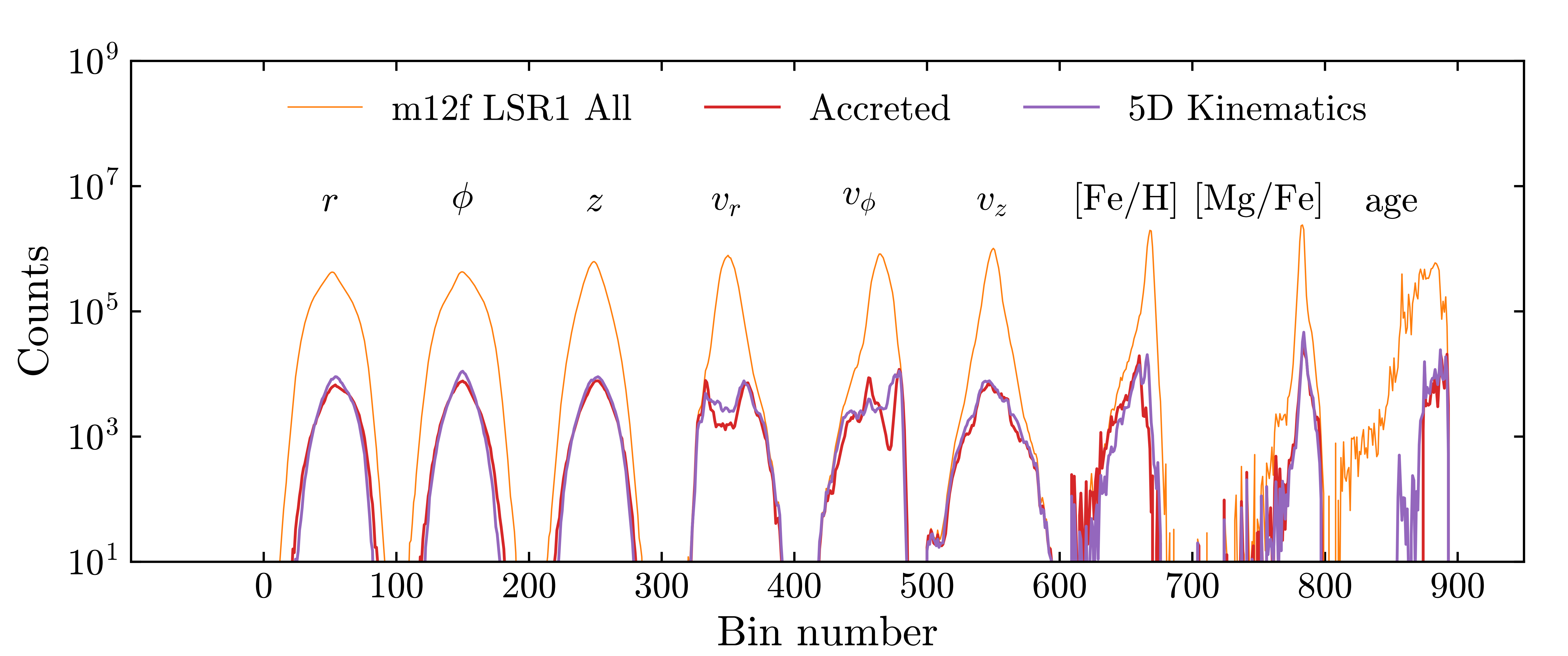
So far so good

Trained on this one

Test on this one

Retrain last layer of network using m12f LSR1 - **0.4%** of the total parameters are allowed to change

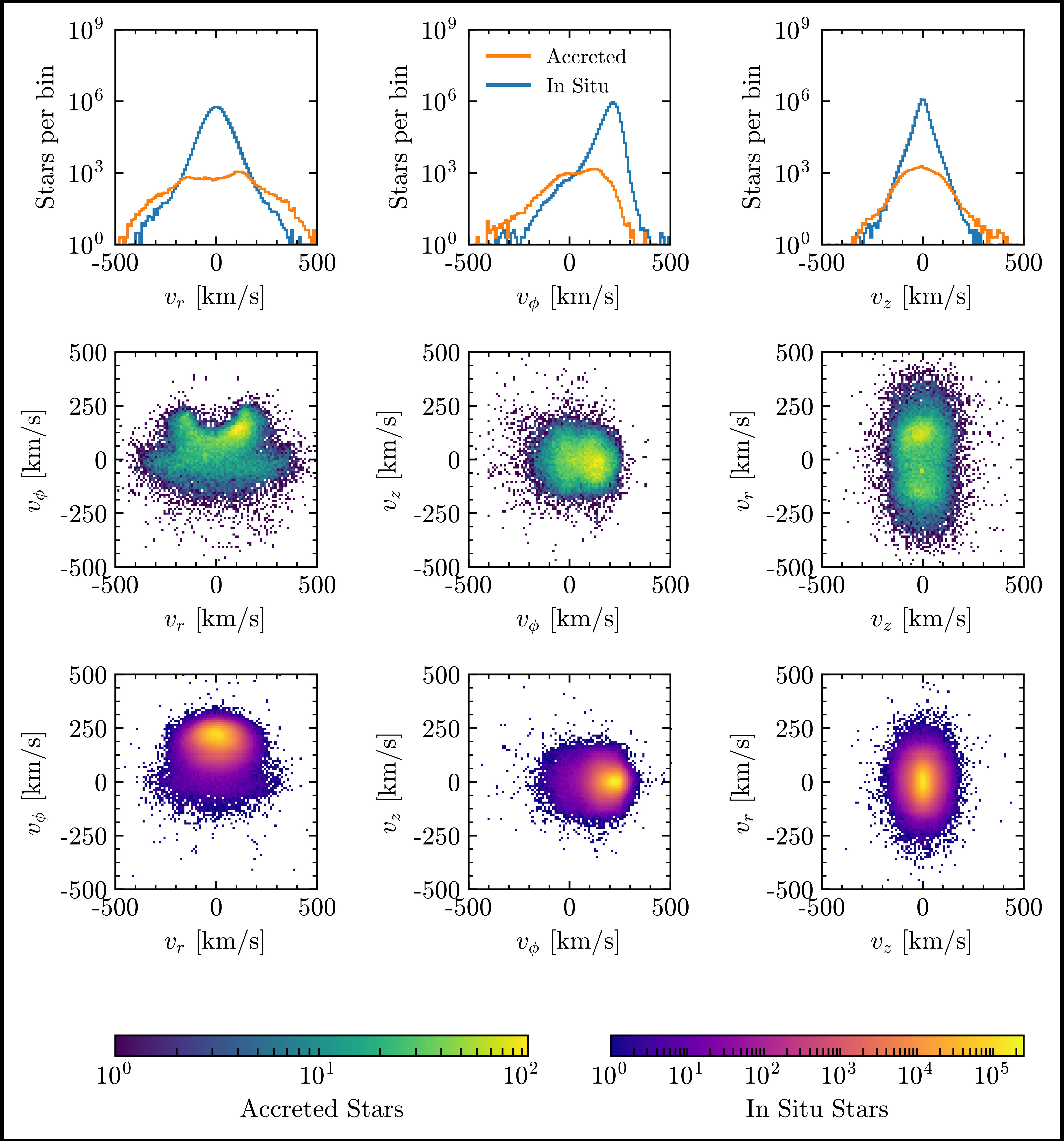




Stars classified using 5D kinematics

Not perfect, but does not imprint features from one galaxy on another

Checks with RAVE DR5 x Gaia DR2



Huge contrast in number of stars in each component

Structure apparent in velocity space for accreted component

In situ component is much smoother at this scale

Application to DR2, $\varpi/\sigma_\varpi \geq 10$

Ostdiek, Necib, ..., Sanderson et al in prep

Takeaway points

- ananke is sufficiently realistic to be used as a training set (with retraining on a small amount of real data)
- Many synthetic surveys were needed for data diversity during training and testing: at least 5 of the 9 available (all 3 viewpoints, 2/3 simulations)
- photometry was harder to transfer than kinematics
- parallax quality was important
- data-driven classification opens up new discovery spaces

outline

- how to make a synthetic survey
- how to use synthetic surveys:
    - making forecasts & planning survey strategies
    - evaluating the “selection function” of a search
    - validating complex analysis methods
    - interpreting structures identified by data-driven models
    - training ML methods for application to Gaia