Searching for Dark Matter Subhalos in Astronomical Data using Deep Learning

2203.08161

Authors: Abdullah Bazarov, María Benito, Gert Hütsi, Rain Kipper, Joosep Pata, Sven Põder

Speaker: Sven Põder

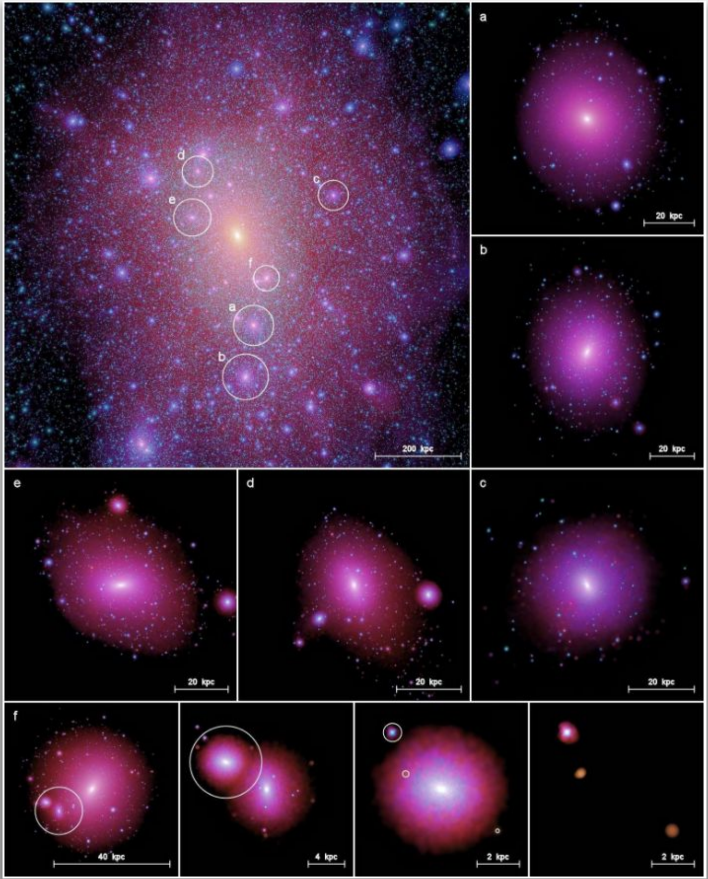
CERN Baltic Conference 2022

Outline

1. The Motivation
2. Galaxy Simulations
3. Synthetic Gaia Surveys
4. Looking for perturbations with ML
5. Training and Evaluation
6. Conclusion and next steps

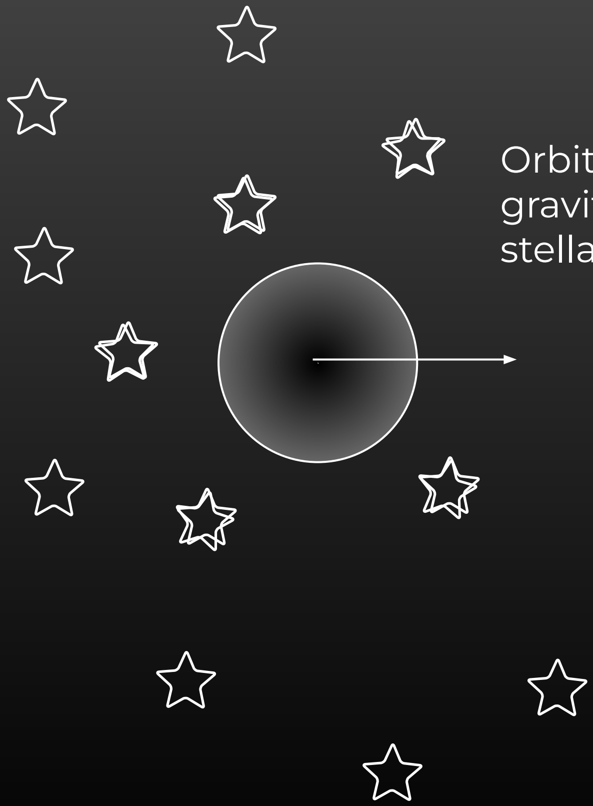
The Motivation

- The nature of dark matter can be explored by testing the prediction of subhalo abundance in LCDM
- CDM expected to form dark subhalos orders of magnitude below $10^8 M_\odot$ which remain dark
- The abundance of subhalos dependent on the particular DM model



Source: Springel et al. (2008)

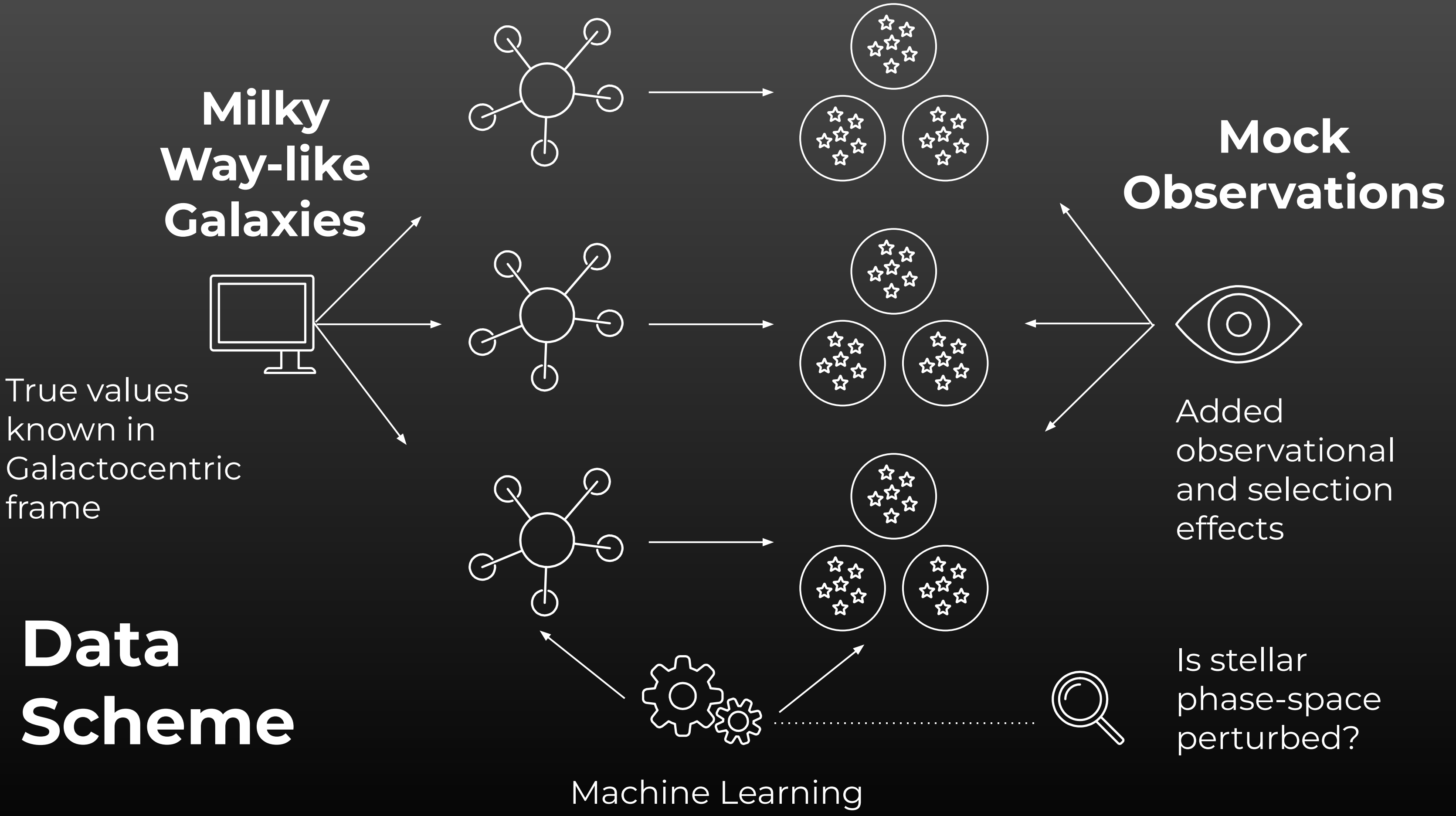
Gravitational Signatures in Stellar Phase-Space



Orbiting subhalo imprints a gravitational signature in the stellar phase-space

We hope to quantify this disturbance from the data

Data Scheme



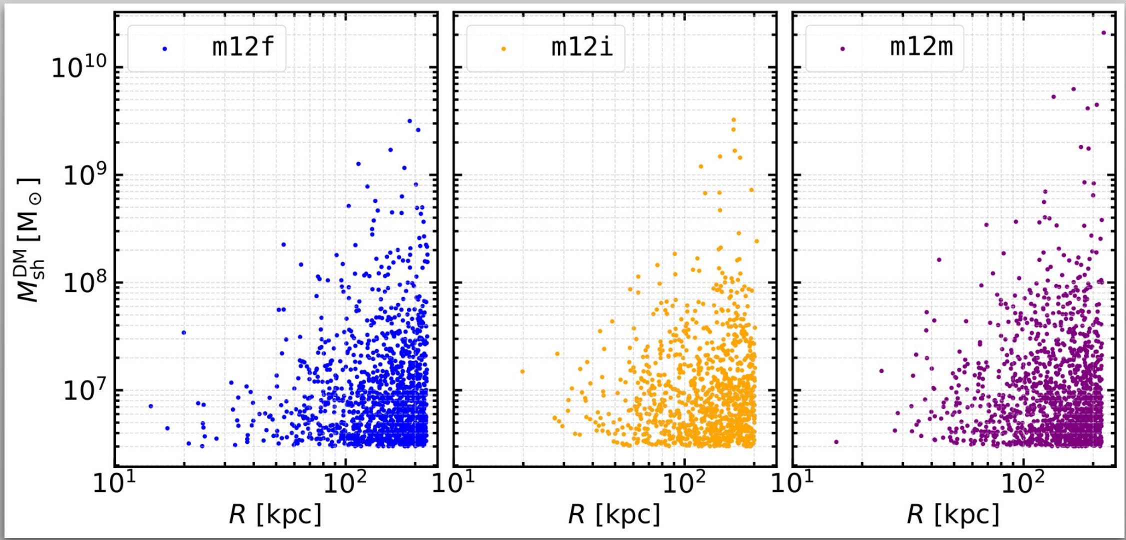
Milky Way-like Galaxies
True values known in Galactocentric frame
Mock Observations
Added observational and selection effects
Is stellar phase-space perturbed?
Machine Learning

Milky Way-like Galaxy Simulations

Each galaxy approx. 2 TB in size

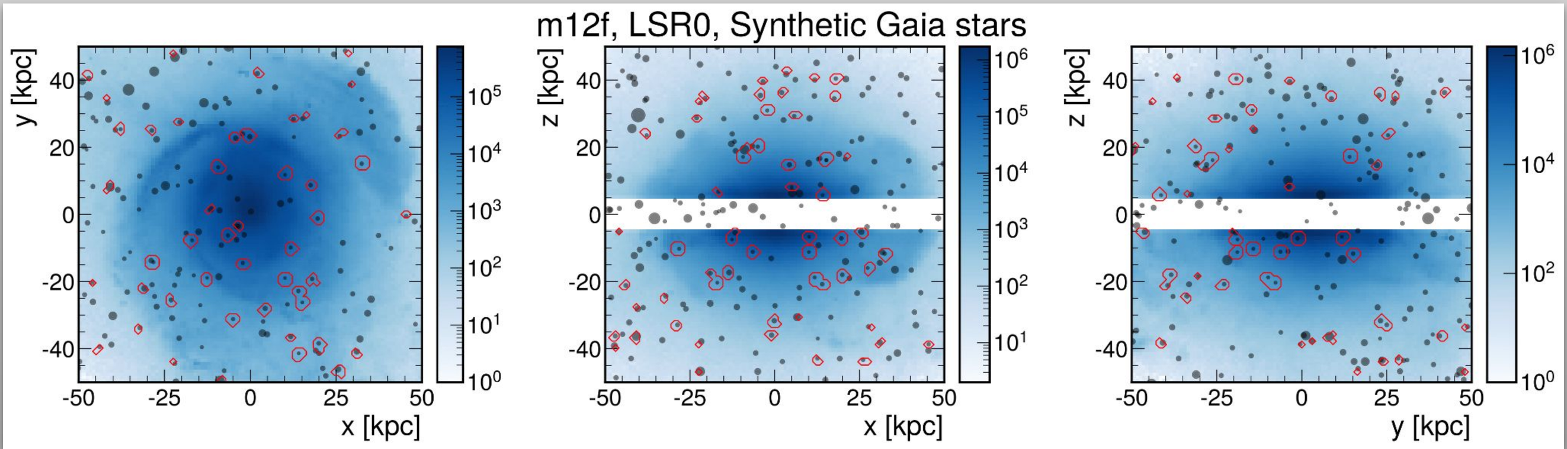
- We used three Milky Way-like galaxies from the Latte suite of FIRE-2 simulations: **m12f**, **m12i** and **m12m** (Wetzel et al., 2016 [1602.05957])
- Subhalos are identified using the Amiga Halo Finder Code (AHF) (Knollmann and Kebe, 2009 [0904.3662])

Approximately 10^3 subhalos in each MW-like which have > 10^6 $M_\odot$

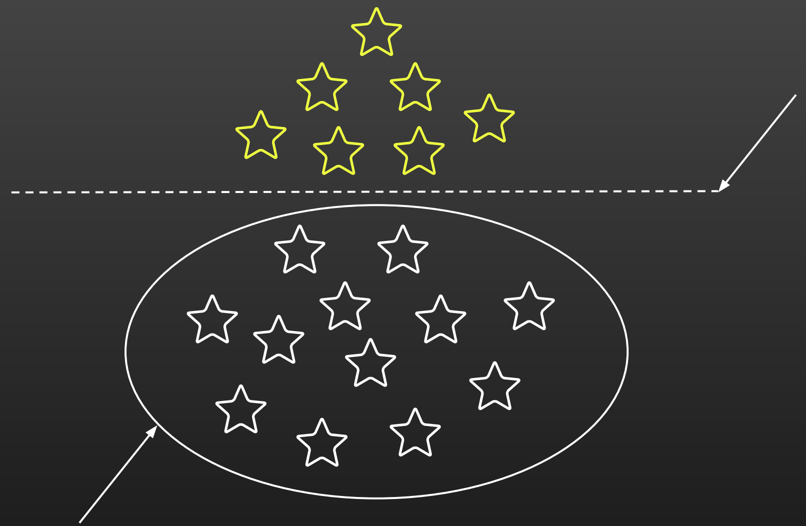


Synthetic Gaia surveys

- Three Gaia DR2-like synthetic surveys per simulated galaxy (Sanderson et al., 2020)
  - Approximately 10^9 mock stellar observations per survey (total of 9 surveys)
  - After removing disk - 1.5 billion observations across the galaxies
  - Stars correlated with potentially observable DM subhalo locations



Signal stars



Binary Classifier (Supervised)

Signal model is used explicitly in optimization

Anomaly Detection (Unsupervised)

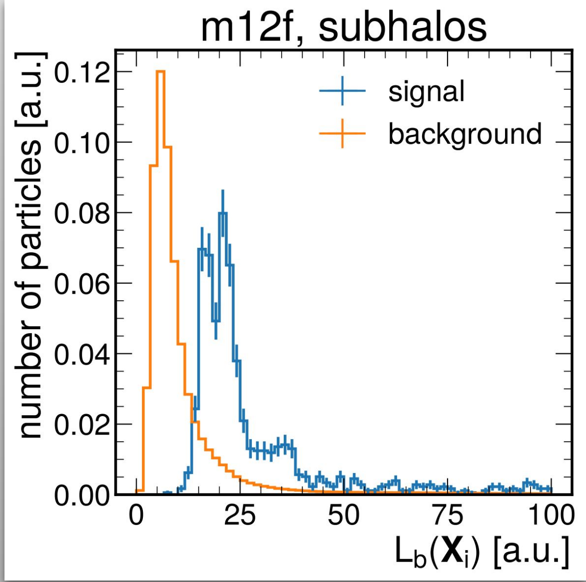
Background samples are used in optimization

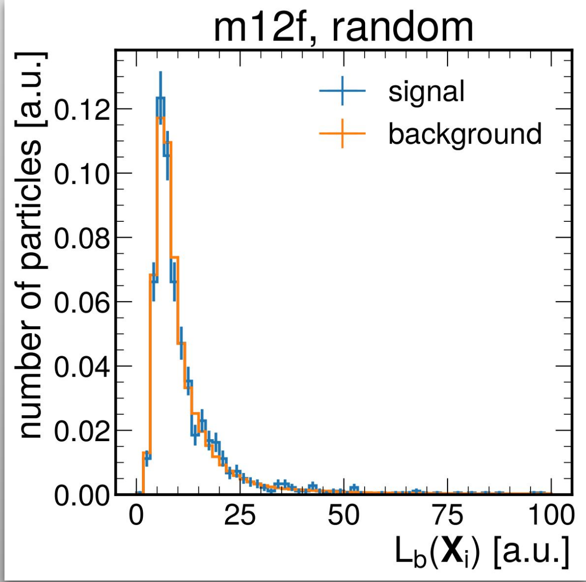
Background stars

Anomaly Detection Approach

Detectability in Ideal Conditions

- Model optimized and validated on m12m and m12i respectively
- Testing done on m12f which is never involved in training
- We check our approach by defining a fake signal

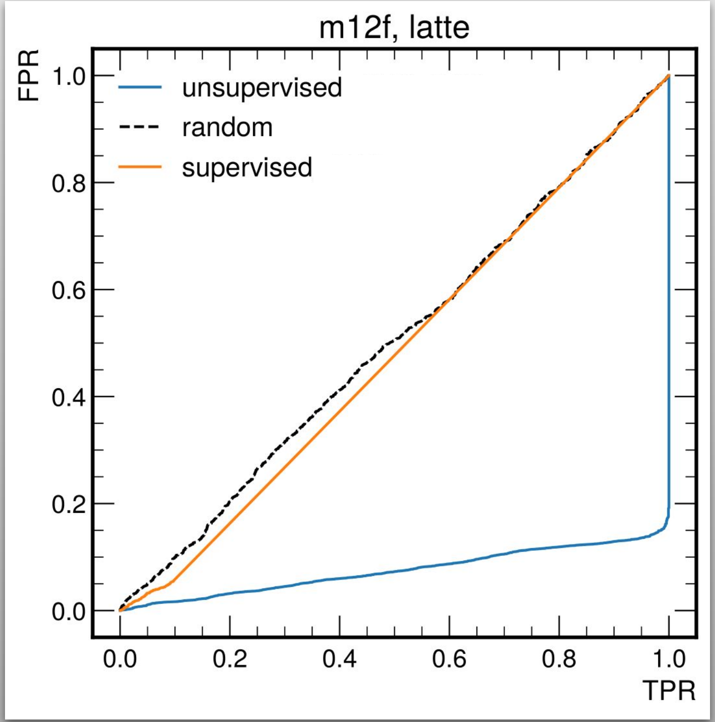




Performance in Ideal Conditions

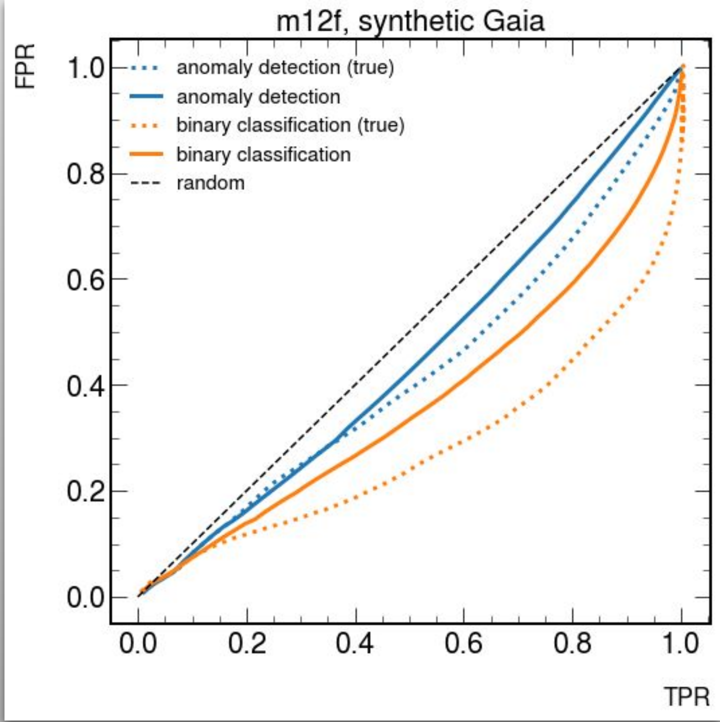
- We quantify the performance of the anomaly detection model in terms of true positive and false positive rates

- We observe that halo-associated star particles in m12f have a distinguishable distribution in the 6D phase-space



Detectability in synthetic Gaia

- We check the performance when selection and experimental effects are taken into account
- Training and testing is done on all three LSRs simultaneously
  - A total of 1.5 billion stars used
- Sensitivity for both anomaly detection and binary classification is nonzero
  - Able to quantify detectability in a given DM model scenario and under particular experimental conditions



Conclusion & Future work

- In our work, we investigated the gravitational imprint of subhalos in stellar kinematics
    - We treated this as a big data problem and used simple ML algorithms to gauge the signal
- We found that despite the limited signal statistics, we are able to obtain non-trivial sensitivity when differentiating between signal and background stars
- We plan to work towards developing a method for setting observational limits on DM scenarios
    - Extending simulations across a range of DM scenarios
    - Going from point anomalies to group anomalies
    - Extending to real observational data

Thank you!

Contact

E-mail: sven.poder@kbfi.ee
LinkedIn: linkedin.com/in/sven-põder/

Anomaly detection approach

- Quantifying the difference between halo-associated and background stars
- We use an autoencoder neural network
  - Each star is characterised by a feature vector **X**

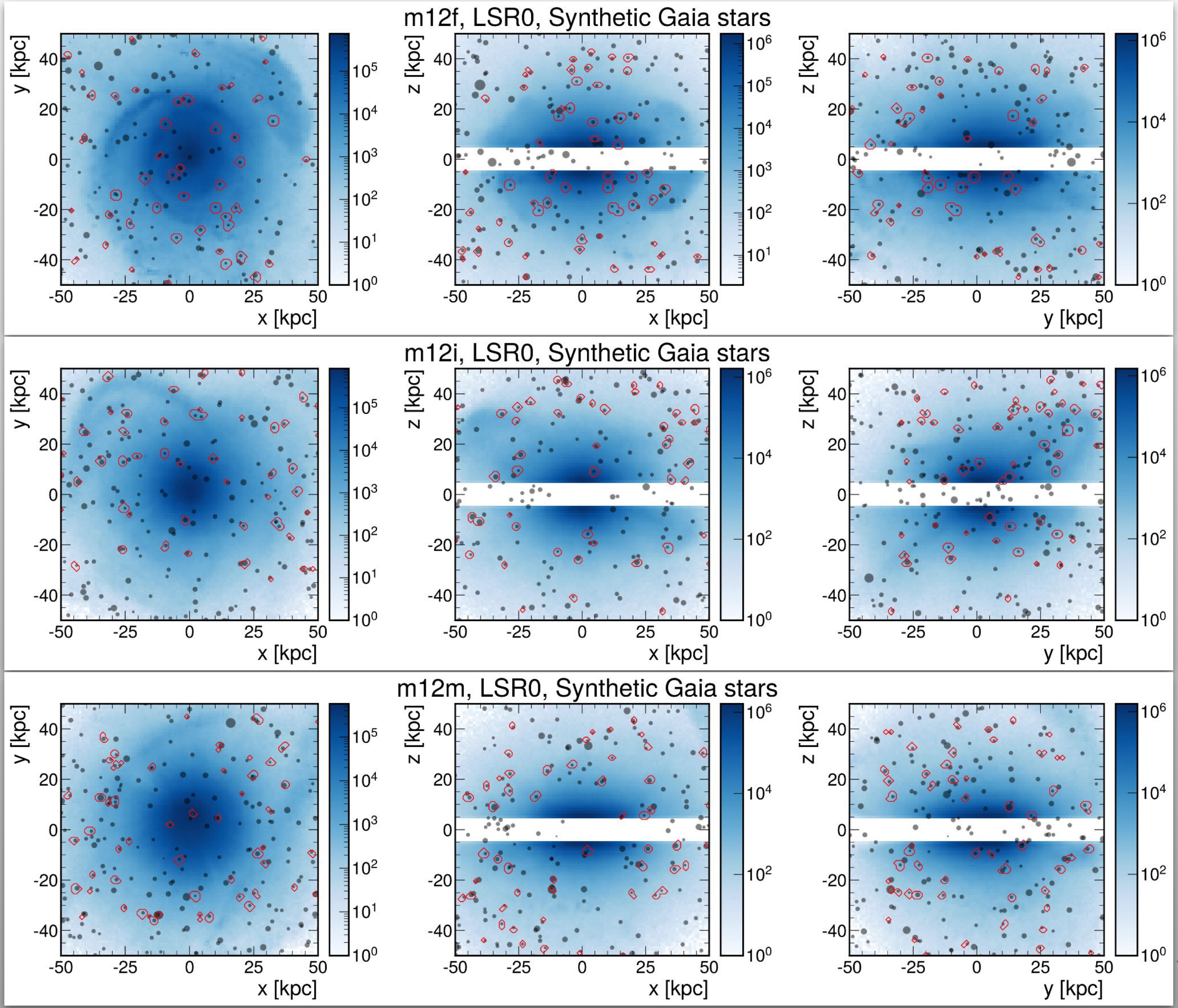
$$E(\mathbf{X}) \rightarrow \mathbf{z} \in \mathbb{R}^D$$

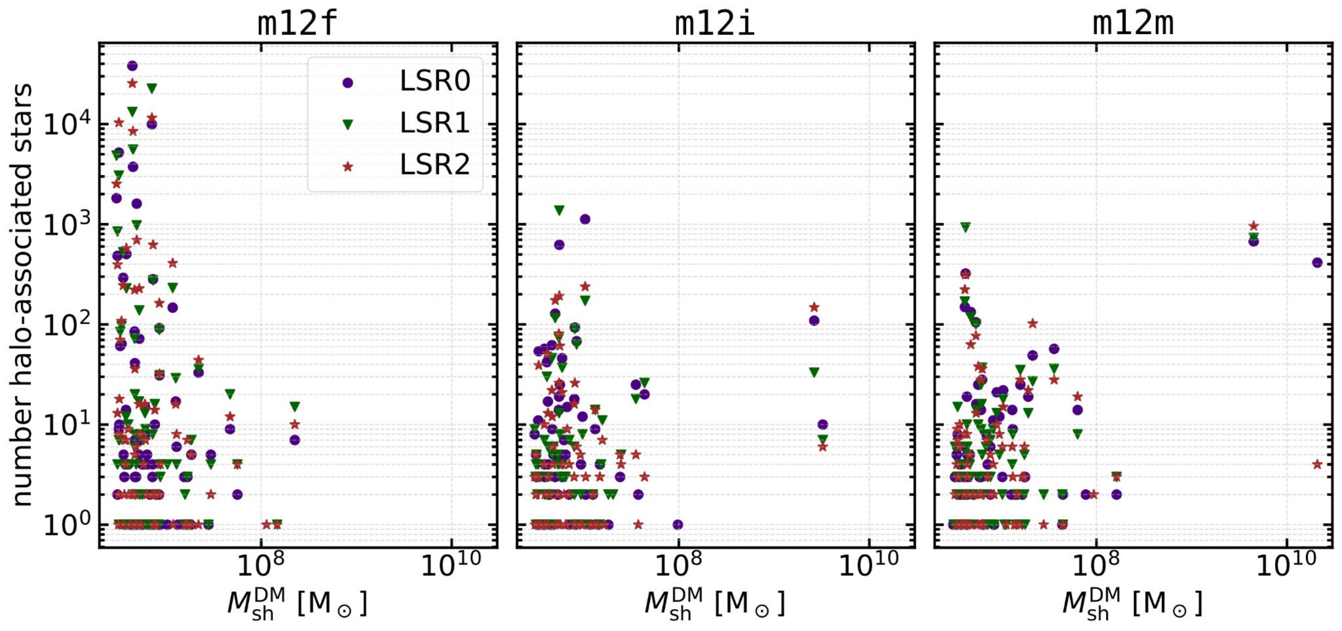
$$D(\mathbf{z}) \rightarrow \mathbf{X}' \in \mathbb{R}^6$$

- The distribution of the reconstruction loss is used as an empirical discriminator between background and halo-associated stars

$$L_b(\mathbf{X}_i) = \|\mathbf{X}_i - D(E(\mathbf{X}_i))\|$$

Projected stellar number densities in the synthetic Gaia datasets for LSR0 in all three galaxies.





The total number of stars associated to a subhalo as a function of the subhalo's mass.

Summary statistics of synthetic Gaia DR2

		stars with $|z| > 5$ kpc	halo-associated stars [%]	with v_r [%]	subhalos w/ halo-associated stars
m12f	LSR0	216,446,024	0.0291%	0.35%	73
	LSR1	182,538,592	0.0291%	0.32%	76
	LSR2	204,017,261	0.0306%	0.35%	71
m12i	LSR0	139,167,343	0.0019%	0.41%	63
	LSR1	132,655,442	0.0017%	0.41%	61
	LSR2	131,474,668	0.0010%	0.23%	67
m12m	LSR0	170,255,144	0.0013%	0.09%	67
	LSR1	156,093,757	0.0016%	0.12%	71
	LSR2	161,369,511	0.0013%	0.19%	68

Training Figures