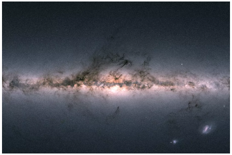
Seeing the Milky Way halo through Gaia's eyes

with machine learning

Marat Freytsis
Rutgers, NHETC (/UC, Berkeley/LBNL)

KMI, Nagoya University, Machine Learning for the LHC
February 6, 2020

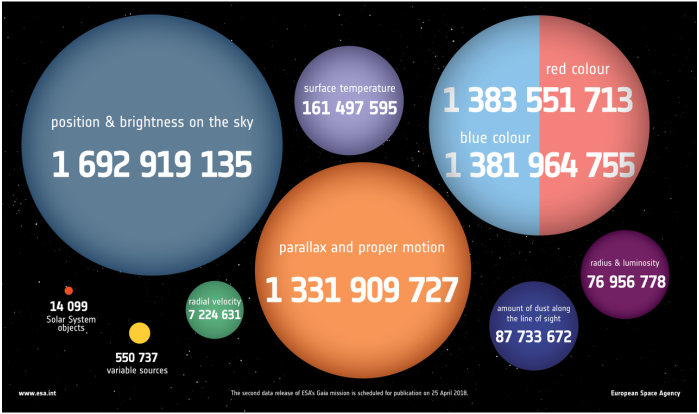
Tim Cohen, MF, Mariangela Lisanti, Lina Necib, Bryan Ostdiek and FIRE collaboration members: Shea Garrison-Kimmel, Andrew Wetzel, Robyn Sanderson, Philip F. Hopkins
[arXiv:1907.06652, 07190, 07681]



The Gaia mission

- A space-based celestial object observatory
  - Launched 2013
  - Data taking since 2014
  - On a Lissajous orbit around Earth's L2 point
- Mission until 2022 (maybe 2024+)
- Data release schedule
  - DR1: 14 September 2016
  - DR2: 25 April 2018
  - EDR3: Q3 2020
  - DR3: Q3/Q4 2021
  - FR: ????
- Position, velocity and spectrophotometry for resolved objects

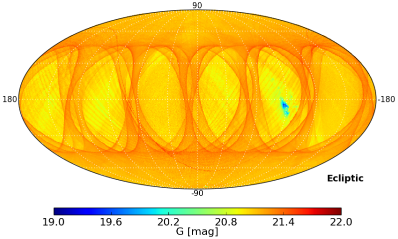
Gaia in numbers

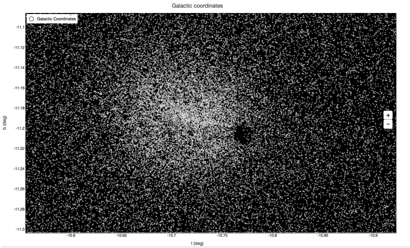


Next best:

- Proper positions/proper motions: $\sim$ 58 million (UCAC2)
- With radial measurement: $\sim 120,000$ (Hipparcos)

Data taking

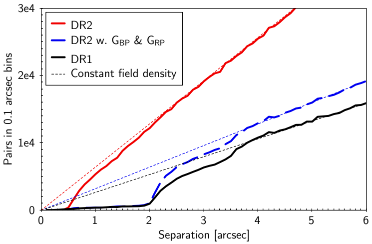




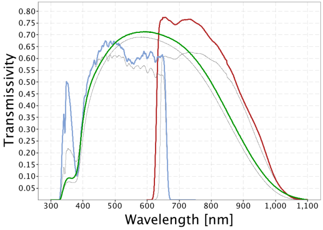
- Continuous scanning (0.75 deg^2 active area)
  - $\sim$ 70 million resolved objects/day
  - $\sim$ 600 million measurements/day
- Sensitivity limited by scan trajectory/local density
- Objects with velocities $\gtrsim$ 1 arcsec/yr hard to identify

Angular resolution

- DR2 angular resolution
  - ~ 0.4 arcsec
  - Will improve by 10^2–10^3
  - Systematic errors $\lesssim 0.1$ mas
  - Significantly better than existing surveys
- Currently no overlap processing
  - Galactic center/binaries
  - Will be included in future DRs

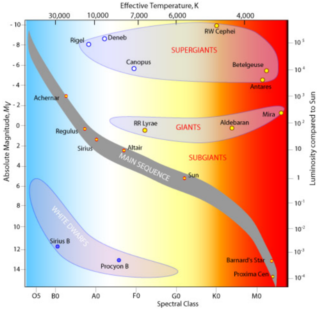


Spectroscopy



- Blue, red, and visible bands
  - $\sigma_G \sim 0.3$–10 mmag
  - $\sigma_{G_{R,B}} \sim 2$– 200 mmag

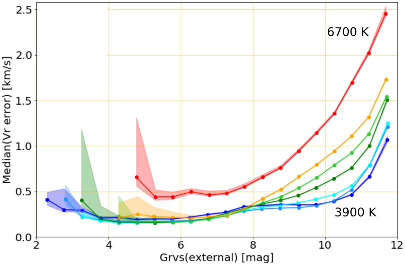
- Can locate stars on H–R diagram for classification:



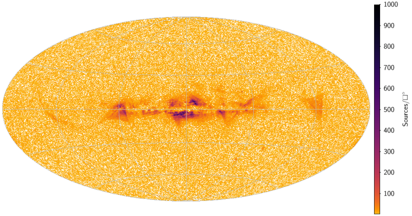
Hertzsprung-Russell Diagram

Radial measurements

Radial velocity precision

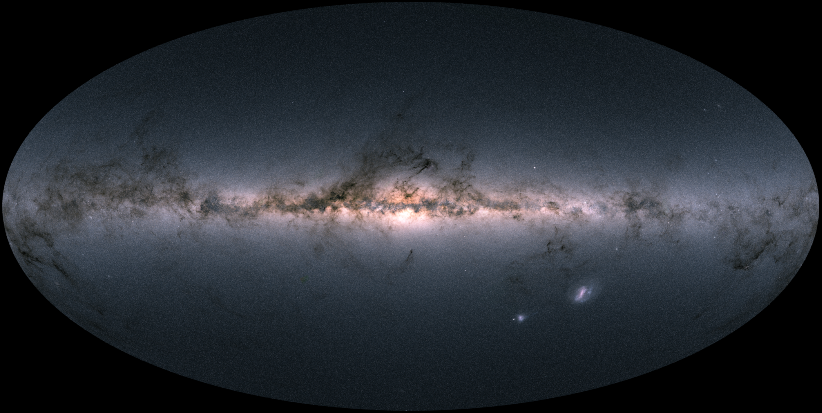


- Dedicated v_r spectrometer
- Systematics $\lesssim 0.25$ km/s
- Require longer exposure
  - Only ~ 7 million measurements



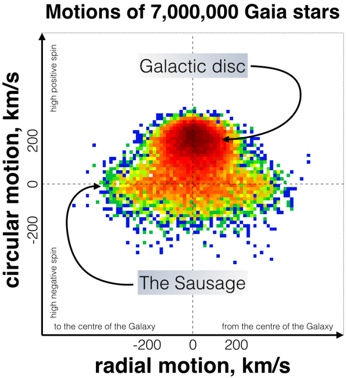
- HQ parallaxes ($\gtrsim 10$ mas) comparable to v_r number
- Converted to radial distance
  - Augmented with variable star calibration in future

The result



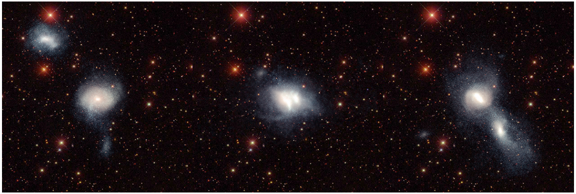
Insights and structures in astrophysics

e.g., the Gaia sausage



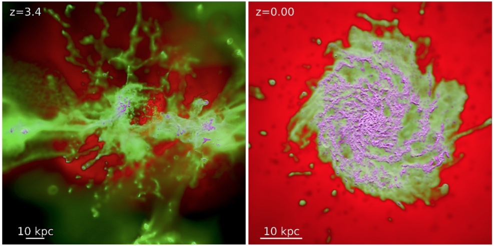
- a.k.a. Gaia–Enceladus
- 3D v measurements reveals a MW–large dwarf merger $\sim$ 10 Gyr ago
- Smeared in position, velocity crucial

[arXiv:1805.00453]



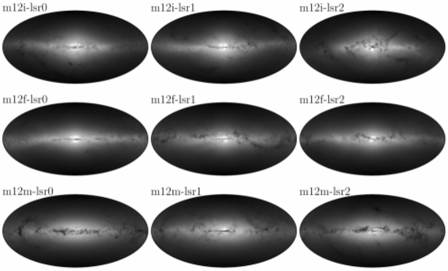
Mock catalogs
FIRE simulations

- Need MW-like data with known structure to train/validate
- FIRE project Latte simulation suite
  - 3-component galactic formation from $z = 100$ to present
  - DM, star, and gas particles
  - Star formation occurs in gas
  - Feedback from radiation pressure, supernovae blowout, stellar mass loss, photoionization, and photoelectric heating
- Trace stars and DM through galaxy formation via clustering



Mock catalogs
Gaia on FIRE

- 3 MW-like simulations
- 3 viewpoints/galaxy
- All $R_\odot$ from center
- Gas extinction and measurement uncertainty effects
- Format like Gaia DRs



Gaia and dark matter

- Gaia: the largest 5D/6D catalog of local astronomical objects ever
- Can it teach us about the dark matter halo of the Milky Way?
- Why improve our halo models?
  - Astronomers: Learn galactic formation histories
  - Particle physicists: Halo feeds into detection rates
- Older stars act as tracers for (some) dark matter
- The challenge: identifying old stars with Gaia only

Plan

- Gaia and DM
- Halo models and stellar tracers
  - Toy models & merger histories
  - Finding visible tracers of DM
- Machine learning with Gaia through FIRE
  - General methods
  - Validating performance
- A first look in the full Gaia DR2

Toy models of Milky Way

visible galaxy

central bulge +
(thin & thick) disk

us: $\sim 8\,\mathrm{kpc}$ out

$M_{\mathrm{stellar}} \approx 5 \times 10^{10} M_{\odot}$

$$z_{\mathrm{disk}} \approx 0.6(3)\,\mathrm{kpc}$$
$$R_{\mathrm{disk}} \approx 15\,\mathrm{kpc}$$
$$R_{\mathrm{bulge}} \approx 4\,\mathrm{kpc}$$

Toy models of the Milky Way

DM halo

rotation curves ($v_c(r) = \sqrt{\frac{GM}{r}}$) $\Longrightarrow$ visible galaxy inside DM halo

$R_{\rm halo} \sim 100\,{\rm kpc}$, $M_{\rm halo} \sim 10^{12} M_\odot$

flat $v_c(r) \Longrightarrow M(r) \propto r$

$\rho(r) \propto r^{-2}$

$v_c(R_{\rm halo}) \sim 200\,{\rm km/\,sec}$

- collisionless
- nonrelativistic
- self-gravitating
- isotropic/isothermal

Toy models of the Milky Way

DM halo

rotation curves ($v_c(r) = \sqrt{\frac{GM}{r}}$) $\Longrightarrow$ visible galaxy inside DM halo

$R_{\text{halo}} \sim 100\,\text{kpc}$, $M_{\text{halo}} \sim 10^{12} M_\odot$

flat $v_c(r) \Longrightarrow M(r) \propto r$

$\rho(r) \propto r^{-2}$

$v_c(R_{\text{halo}}) \sim 200\,\text{km/sec}$

- collisionless
- nonrelativistic
- self-gravitating
- isotropic/isothermal

Toy models of the Milky Way
DM halo

rotation curves ($v_c(r) = \sqrt{\frac{GM}{r}}$) $\Longrightarrow$ visible galaxy inside DM halo

$R_{\text{halo}} \sim 100\,\text{kpc}$, $M_{\text{halo}} \sim 10^{12} M_\odot$
flat $v_c(r) \Longrightarrow M(r) \propto r$

$$\rho(r) \propto r^{-2}$$
$$v_c(R_{\text{halo}}) \sim 200\,\text{km}/\sec$$

- collisionless
- nonrelativistic
- self-gravitating
- isotropic/isothermal

Hierarchical merger model

Where did all this come from?

1. Density fluctuations after big bang lead to protogalactic fragments of $O(10^6$–$10^8 M_\odot)$
2. Fragments evolve in isolation creating stars/globular clusters
3. Collisions and tidal disruptions lead to distribution of halo (stars and DM)
4. Gas in the mergers interacts and collapses to disk
5. Young, metal rich stars produced in the disk

The last major merger occurred $\sim 10\,\mathrm{Gyr}$ ago
Minor mergers still happening

Hierarchical merger model

Where did all this come from?

1. Density fluctuations after big bang lead to protogalactic fragments of $O(10^6\text{–}10^8 M_\odot)$
2. Fragments evolve in isolation creating stars/globular clusters
3. Collisions and tidal disruptions lead to distribution of halo (stars and DM)
4. Gas in the mergers interacts and collapses to disk
5. Young, metal rich stars produced in the disk

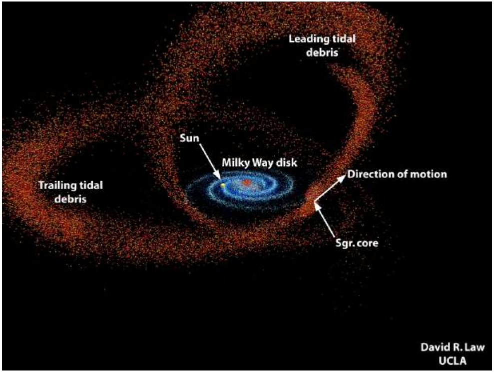
The last major merger occurred $\sim 10\,\mathrm{Gyr}$ ago
Minor mergers still happening

Old stars as tracers

Local halo imprinted with merger history

Stars and DM interact (almost) only through gravity

To find DM, find stars from early mergers



Tracing DM

How to detect the oldest stars?

Early merger $\longrightarrow$ old star $\longrightarrow$ low metallicity

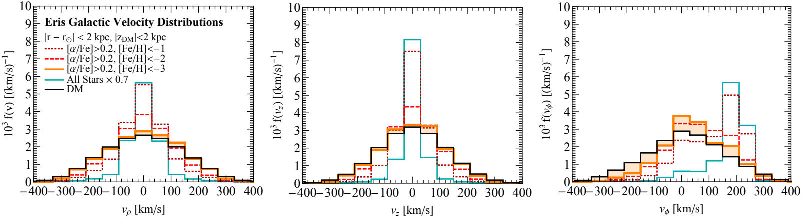
$$[\mathbf{Fe}/\mathbf{H}] = \log_{10}\left(\frac{N_{\mathrm{Fe}}}{N_{\mathrm{H}}}\right) - \log_{10}\left(\frac{N_{\mathrm{Fe}}}{N_{\mathrm{H}}}\right)_{\odot} < C$$

Also helps not to look directly in the disk

$$|z| > z_{\mathrm{cut}}$$

Tracing DM
results in simulation

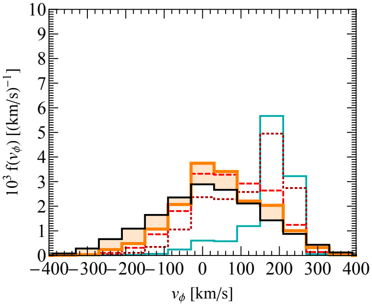
Does this work?

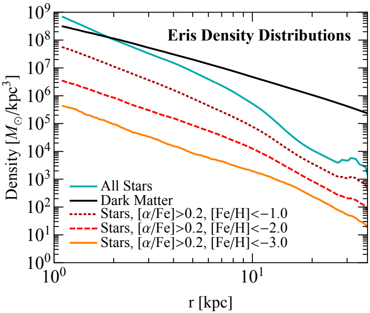


Herzog-Arbeitman, Lisanti, Madau, Necib [arXiv:1704.04499]

Old stars and DM share the same velocity distributions!

Tracing DM
results in simulation





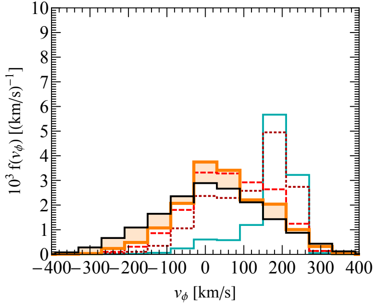
Herzog-Arbeitman, Lisanti, Madau, Necib [arXiv:1704.04499]

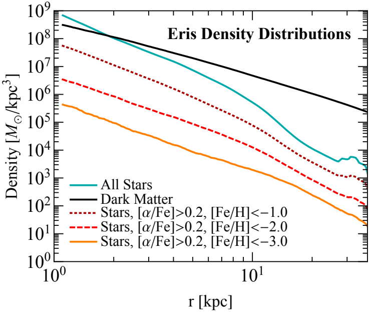
Old stars and DM share the same density profile!

Can stellar tracers of virialized DM be isolated in practice?

Tracing DM

results in simulation





Herzog-Arbeitman, Lisanti, Madau, Necib [arXiv:1704.04499]

Old stars and DM share the same density profile!

Can stellar tracers of virialized DM be isolated in practice?

Catalogs of real data

Phase space

- Gaia DR1 (2-D location for 1.1 billion stars)
  - Crossmatched with Hipparcos Tycho-2 catalog (2 million stars)
- Gaia DR2 (5-D PS for 1.3 billion stars)

RAVE-TGAS (255,922 stars)

Spectroscopy + v_r

- RAdial Velocity Experiment
- Sloan Digital Sky Survey

Gaia-SDSS (193,162 stars)

Catalogs of real data

Phase space

- Gaia DR1 (2-D location for 1.1 billion stars)
  - Crossmatched with Hipparcos Tycho-2 catalog (2 million stars)
- Gaia DR2 (5-D PS for 1.3 billion stars)

RAVE-TGAS (255,922 stars)

Spectroscopy + v_r

- RAdial Velocity Experiment
- Sloan Digital Sky Survey

Gaia-SDSS (193,162 stars)

Catalogs of real data

Phase space

- Gaia DR1 (2-D location for 1.1 billion stars)
  - Crossmatched with Hipparcos Tycho-2 catalog (2 million stars)
- Gaia DR2 (5-D PS for 1.3 billion stars)

RAVE-TGAS (255,922 stars)

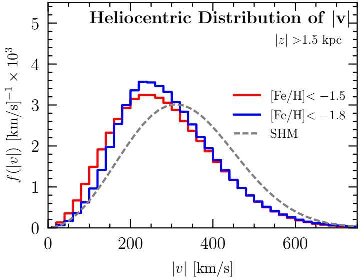
Spectroscopy + v_r

- RAdial Velocity Experiment
- Sloan Digital Sky Survey

Gaia-SDSS (193,162 stars)

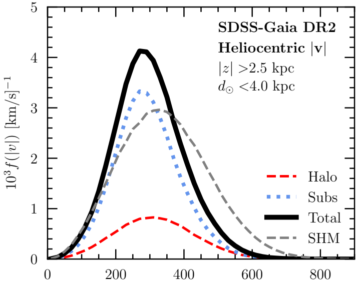
...and real-world results

RAVE-TGAS



[arXiv:1708.03635]

Gaia-SDSS



[arXiv:1807.02519]

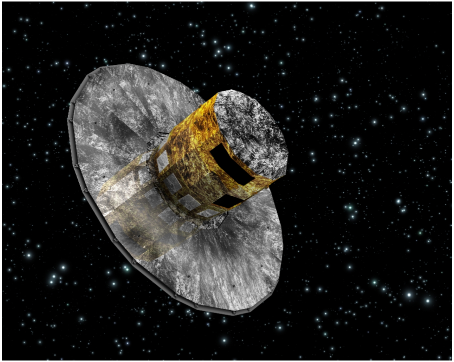
virialized DM velocities smaller than standard halo model
$\Longrightarrow$ potential implications for DM direct detection

But accuracy limited by cross-correlating data

Plan

- Gaia and DM
- Halo models and stellar tracers
  - Toy models & merger histories
  - Finding visible tracers of DM
- Machine learning with Gaia through FIRE
  - General methods
  - Validating performance
- A first look in the full Gaia DR2

Letting Gaia see on its own



DR2: 5-D kinematics and 2-band spectroscopy on 1.3 billion stars

Not enough information to extract metallicity conventionally
Idea: Use neural network classifier as old star distribution fitter

Gaia data format details

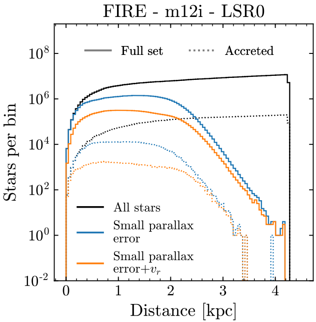
Stellar information provided

- Galactic longitude and latitude (ℓ, b)
- Proper motion in right ascension and declination $(\mu_{\alpha,\delta})$
- Parallax
- Blue- and red-band magnitude $(G_{\mathrm{BP,RP}})$

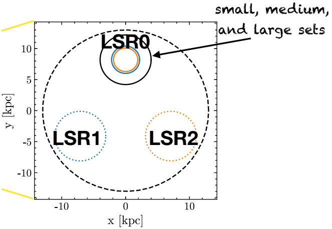
Provides 5D phase-space information (radial v missing)
Complementary information to parallax in G
if neural network can learn distance–luminosity function
Residual information about metallicity also in G?

Network and training procedure

- 5-layer MLP classifier
  - 7 inputs à la Gaia
  - 3 hidden layers of 100 nodes each
  - binary cross-entropy loss
  - star classified as accreted or not
- Label from FIRE merger history
  - Remove metallicity middleman
- 600 million stars per viewpoint
- Include measurement uncertainty by resampling each star within its errors 20 times

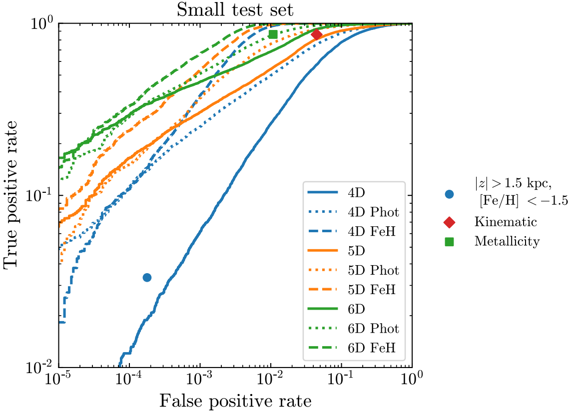


Crosschecks and transfer learning



- Maybe just learn particular local distribution/merger history?
  - Compare different observations
  - Compare different simulations
- Systematic errors in FIRE mocks?
- Compensate via transfer learning
  - Lower NN layers learn simple cuts
  - High-level observables in top layer
  - Train full network on a dataset
  - Reset *top layer only* and retrain *only that layer* on new data
  - Requires much less data in 2nd set
  - Reduce sensitivity to complex features in original training set

Classifying close stars

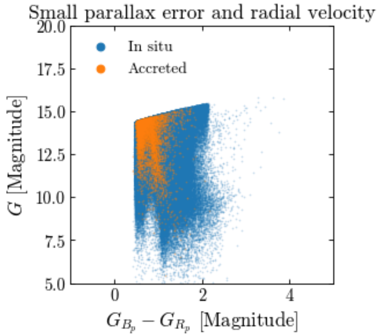


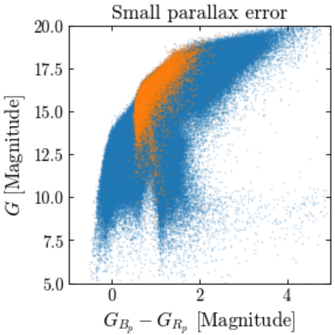
Close stars have multiple parallax measurements

$\longrightarrow$ radial velocity recovered, full 6-D PS information available

Photometric data help when only reduced PS information exists

A closer look at photometric data



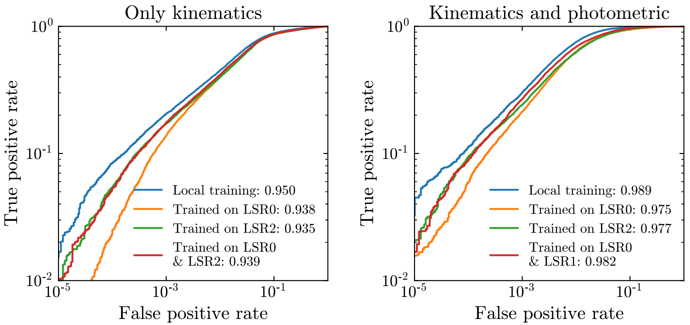


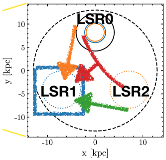
At smaller distances, training data doesn't cover full HR diagram
Luminocity-distance relations not fully learned

must be careful training set goes out as far as real data with photometry

Comparing viewpoints

Testing on LSR1



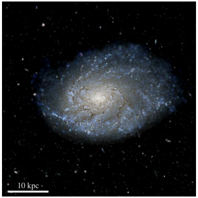


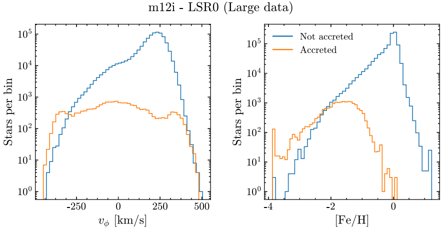
training on multiple viewpoints
$\Longrightarrow$ improved generalization

Trying a new galaxy

old galaxy

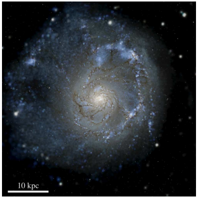
m12i

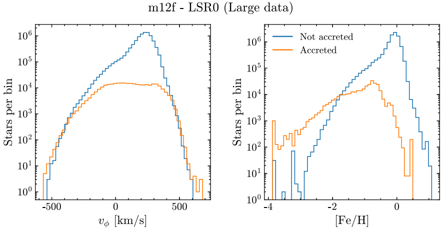




new galaxy

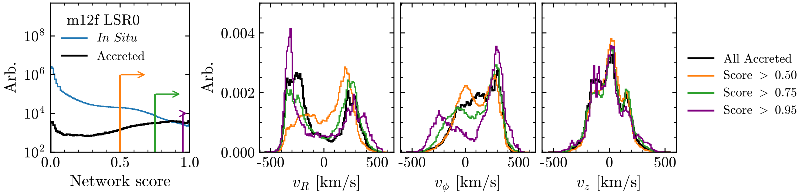
m12f





Different merger history indicated by v_φ distribution

Reconstruction of underlying kinematics



Too loose $\Longrightarrow$ *in situ* contamination
Too strong $\Longrightarrow$ distortion due to aggressive cuts on disk-like stars

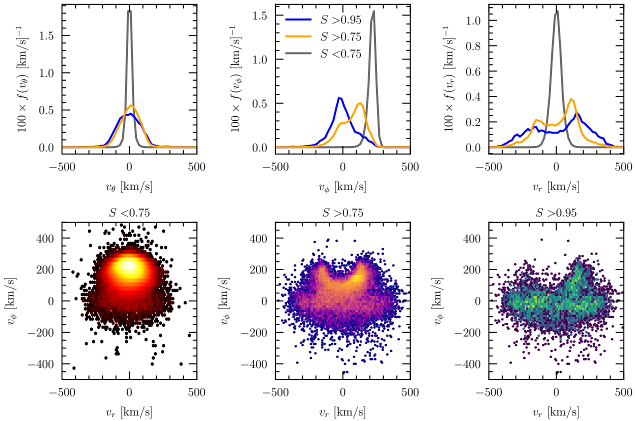
Plan

- Gaia and DM
- Halo models and stellar tracers
  - Toy models & merger histories
  - Finding visible tracers of DM
- Machine learning with Gaia through FIRE
  - General methods
  - Validating performance
- A first look in the full Gaia DR2

First look at Gaia DR2

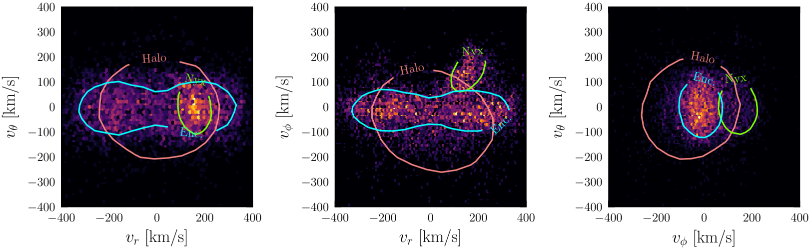
Have an (expected) 60% pure accreted Gaia DR2 dataset

- Contains 21 304 stars with full 6D information



Gaia-Enceladus clearly visible

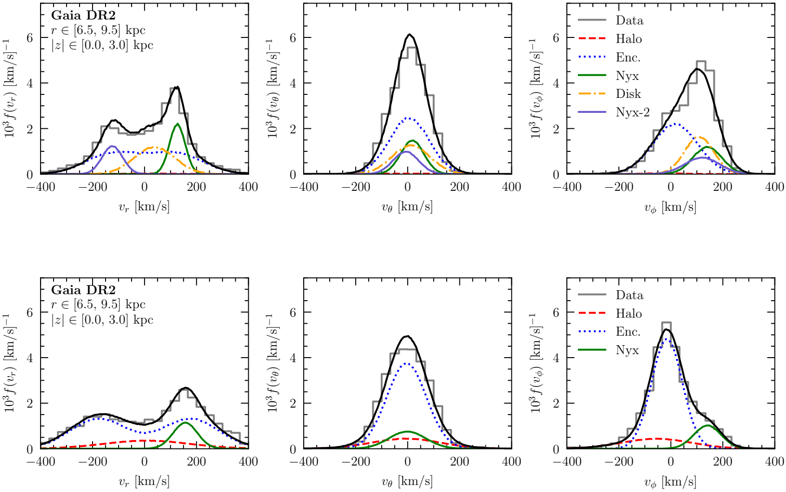
Gaussian component analysis



All stars within $r \in [6.5, 9.5]$kpc and $|z| < 3$ kpc

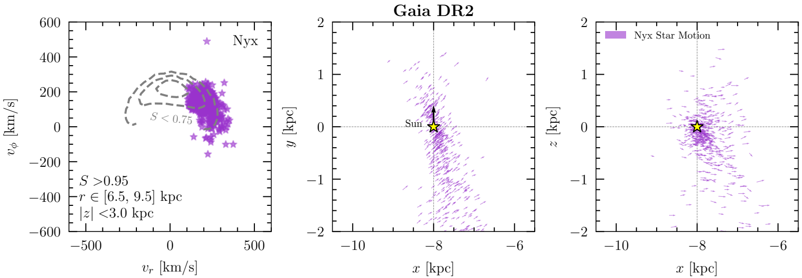
Well described by a multi-component Gaussian analysis
if 3rd r-asym. component is added to halo and Enceladus

More components and looser cuts



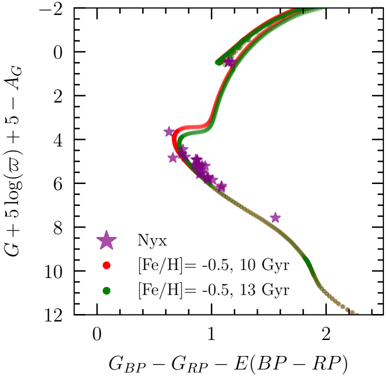
Looser cuts give evidence second (possibly related) new component

Nyx alone



Nyx
$S < 0.75$
$S > 0.95$
$r \in [6.5, 9.5]$ kpc
$|z| < 3.0$ kpc
v_ϕ [km/s]
v_r [km/s]
Gaia DR2
Sun
y [kpc]
x [kpc]
Nyx Star Motion
z [kpc]
x [kpc]

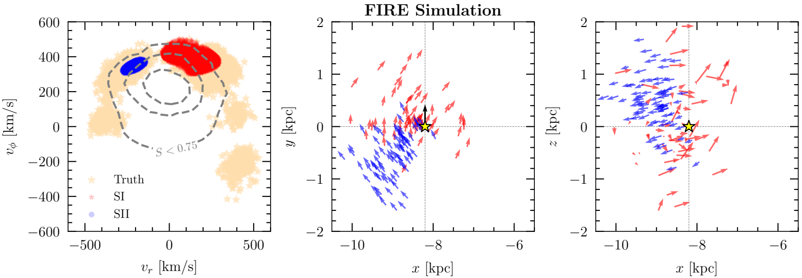
Dating Nyx



Only a 27 stars cross-correlated with spectroscopic surveys

- weak evidence of old isochrone consistency
- follow-up surveys will make a big difference

Why should you believe us?



Identical analysis performed on FIRE `m12f` simulation

Correctly identifies most dense regions of 2 largest streams

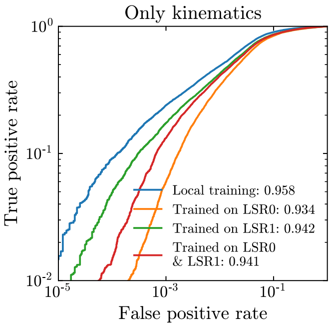
Conclusions

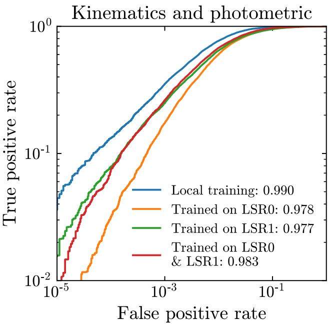
- Hierarchical mergers imply old stars are efficient DM tracers
  - metallicity and kinematics serve as efficient selection criteria
  - Gaia has no access to metallicity; cut-based analyses insufficient
- ML allows the full resolving power of the Gaia dataset to be brought on the problem
  - Kinematic and spectral information can be as powerful
  - Training must be performed carefully to avoid sample bias
  - Transfer learning techniques help control systematics
- ML gives a path to unlocking the full potential of the Gaia
  - Accreted catalog publicly available for other analyses
- Analysis of stars with only 5D PS in the near future?
- Can be say anything about unvirialized/unresolved DM?

Thank you!

Comparing viewpoints

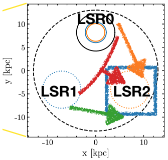
Testing on LSR2





details depend on local kinematics

seemingly more stable generalization with $G_{\mathrm{BP,RP}}$



DBSCAN locations of known streams