CURTAINS

Weakly Supervised Methods for new physics searches

CHIPP 2024 Annual Meeting, Geneva

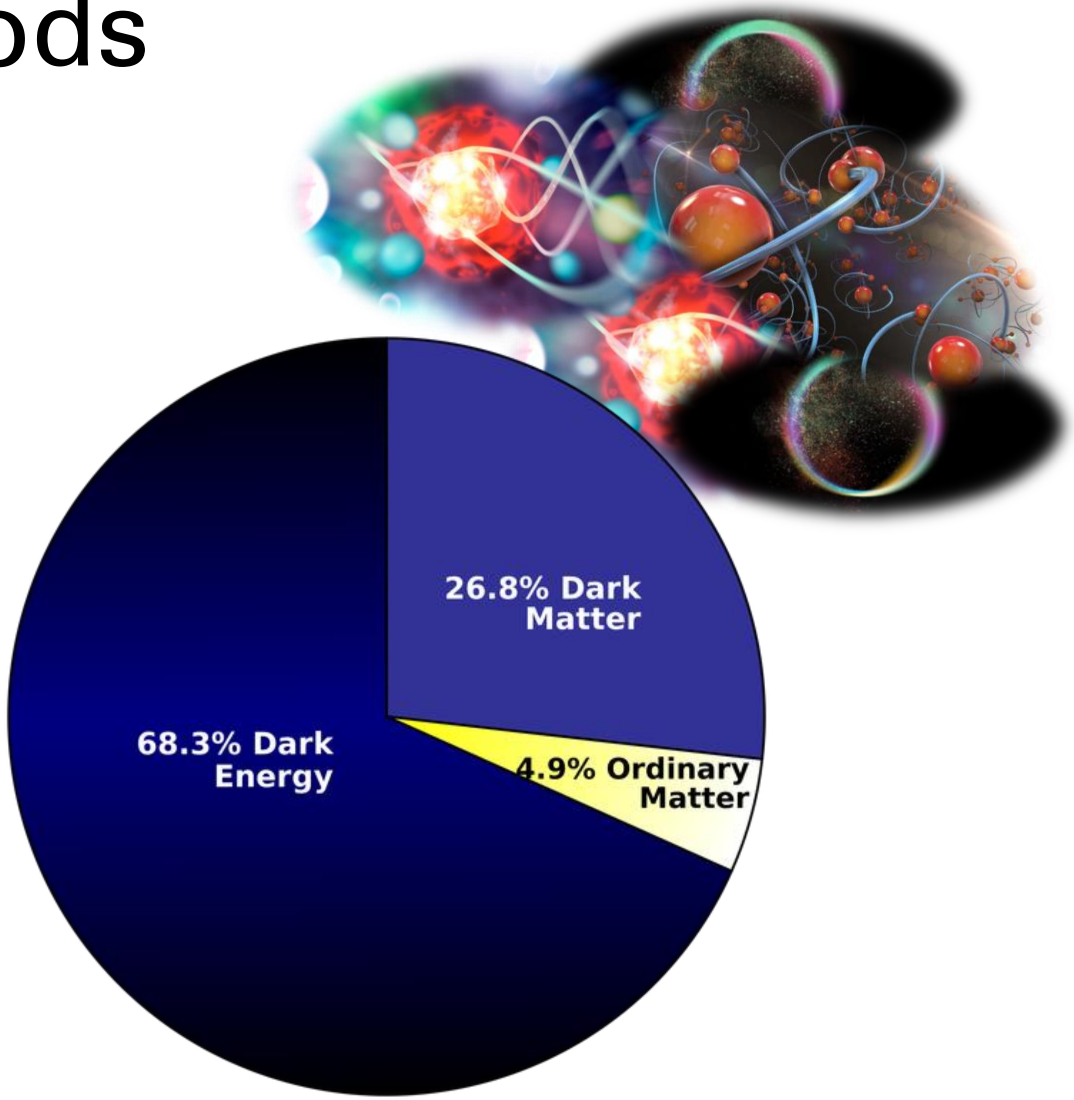
Deb, Sam Klein, Johnny Raine, Tobias Golling

Weakly Supervised Methods

~25 % of matter content is unknown ~ Many models

Supervised searches most optimal, **but not feasible**.

No labels in real life ~ but something close.

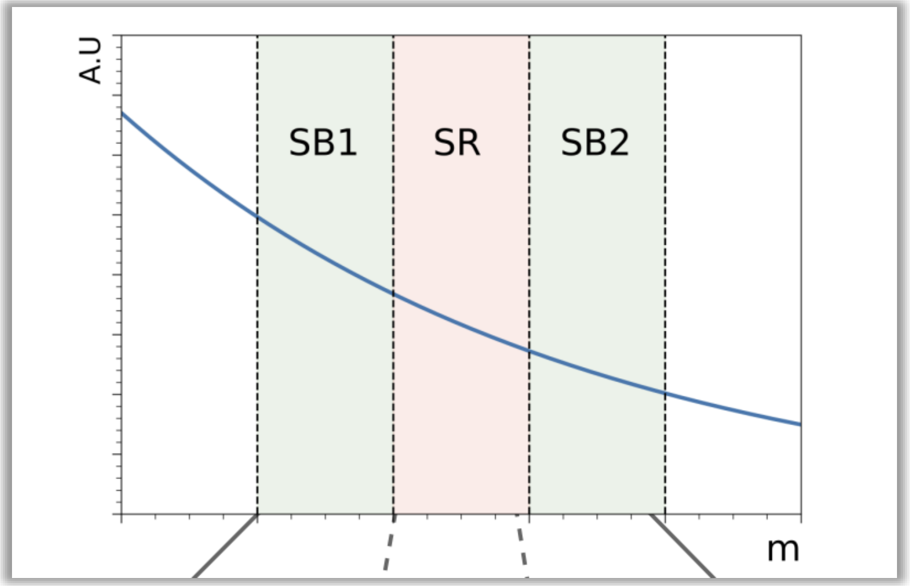


Composition of the universe, and MidJourney’s response to “What does new physics look like?”

Weakly Supervised Methods

- Supervised searches most optimal, **but not feasible**.
- No labels in real life ~ but something close.
- Signal Regions (SR), Sidebands (SB) ~ different fraction of signals.
- CWoLa principle – Optimal classifier for two different admixtures is the optimal classifier between the two classes.

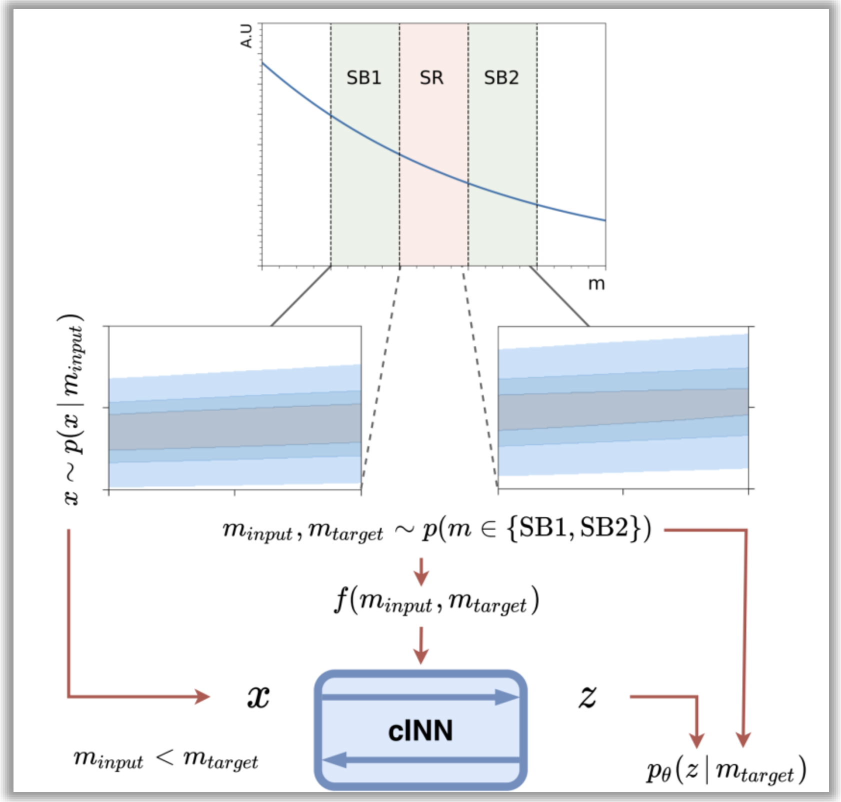
Strategy:
1. Construct background enriched templates in SR
2. CWoLa



Signal region (SR) and Sidebands (SB) around a hypothetical resonance in m.

CURTAINS 2203.09470 2305.04646

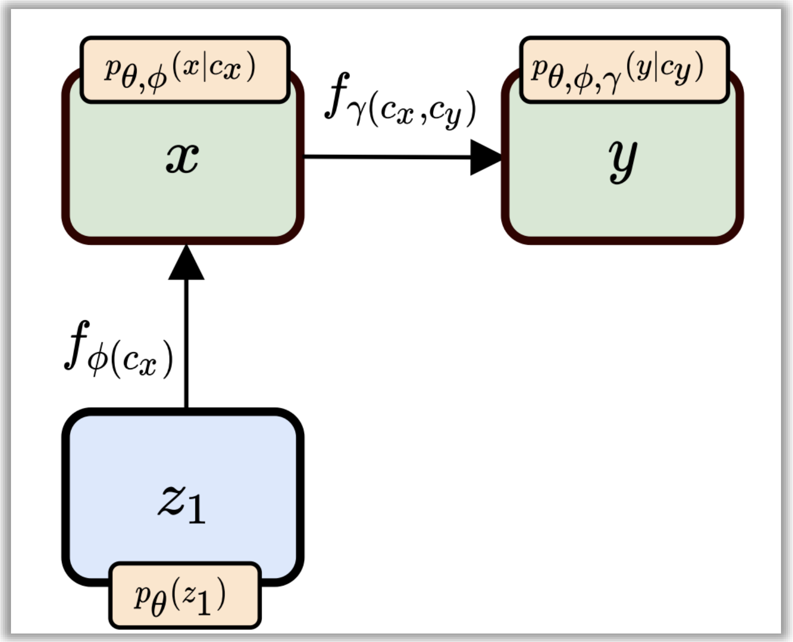
- Transform SB data to SR under exact likelihood
  - Using Flows for Flows
- Background enriched template in the SR



Schema for curtains, data from sidebands are passed through cINN in a forward or inverse pass depending on input and target m

FLOWS FOR FLOWS

- Transform SB data to SR under exact likelihood
- Train a normalizing flow SB⇔SB
  - $p_{\theta\phi\gamma}(y \mid c_x, c_y) = p_{\theta\phi}(x|c_x, c_y) \cdot \det\left| J_{f_\gamma}(x|c_x, c_y) \right|$ ➔ Top Flow
  - $p_{\theta\phi}(x|c_x, c_y) = \pi_\theta(z) \cdot det\left| J_{f_\phi}(x|c_x) \right|$ ➔ Base Flow
- Template → Sample masses from SR and transform SB into SR.



Flows for Flows schematic

CURTAINS FOR DIJETS

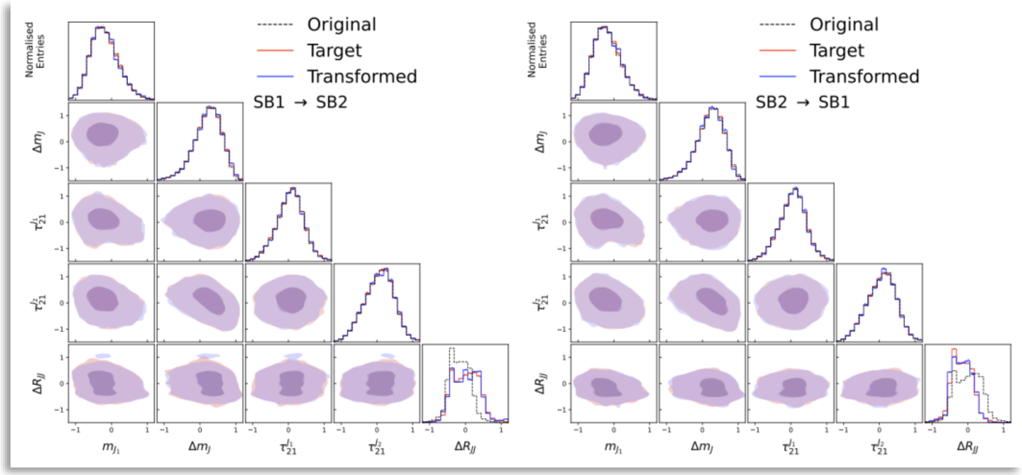
- Public benchmark dataset: LHCO dijet (Zenodo)
- Background: QCD dijets ~ 1M
- Signal: W' (3.5TeV)→ X (.5TeV) (qq) Y (.1TeV) (qq) ~ 100k
- R=1.0 Jets, pT > 1.2 TeV
- Features used:
  - Mj1, Mj1-Mj2, τ21, τ32, dR

TEMPLATE FIDELITY

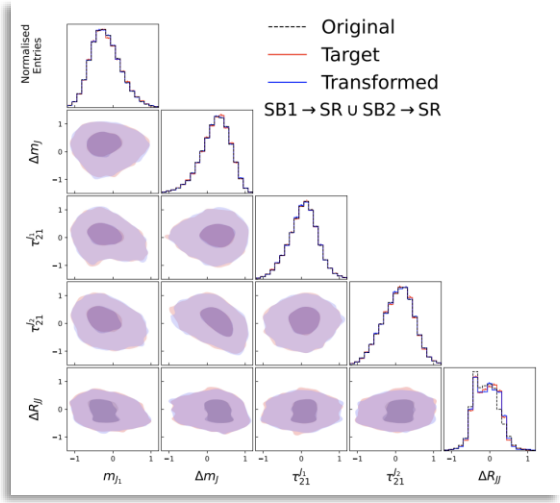
- Train a classifier between generated template and Data.
- If template is good → Classifier should have a nearly random selection ➔ AUC 0.5
  - SB1, SB2 templates look nearly perfect with AUC 0.51
  - SR template AUC of 0.50

- A good template is crucial for downstream anomaly detection!
  - Bad template might kill sensitivity to weak signals.

AUC: 0.51

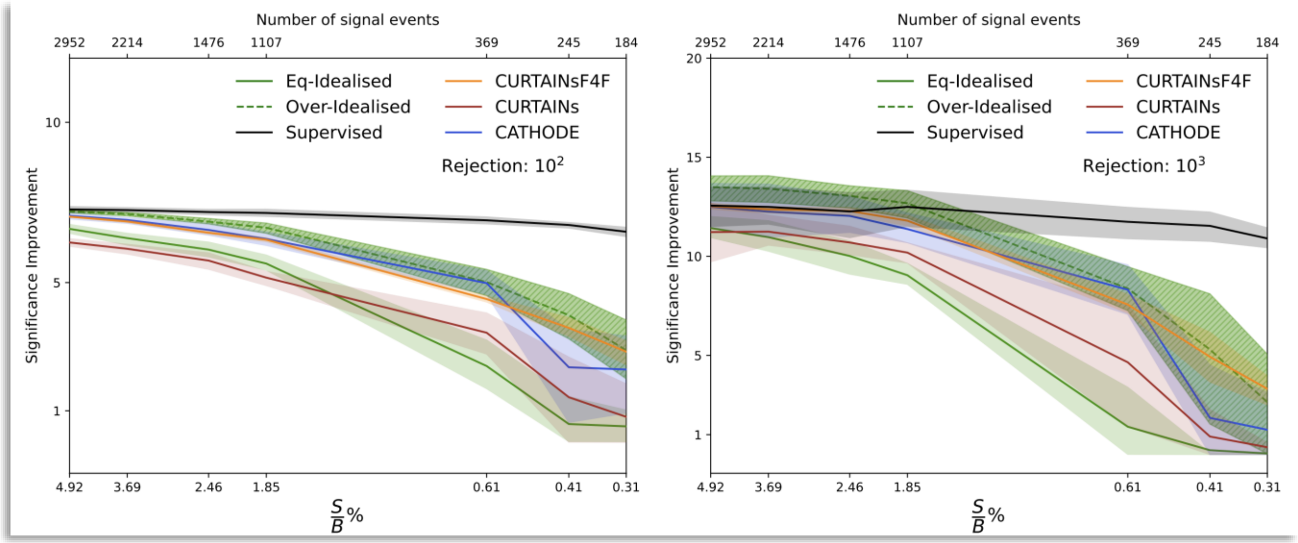


AUC: 0.501



SIGNAL SENSITIVITY

- Dope the data with n signal events
    - $\mathrm{SIC} = \frac{\epsilon_S}{\sqrt{(\epsilon_B)}}$
    - CURTAINsF4F (orange) sensitive to signal even at quite low signal presence.
    - Can find evidence of a signal when initial $\frac{s}{\sqrt{b}} = 0.7$
    - Idealised and Supervised lines for comparison.



SIC at two different working points as a function of doping

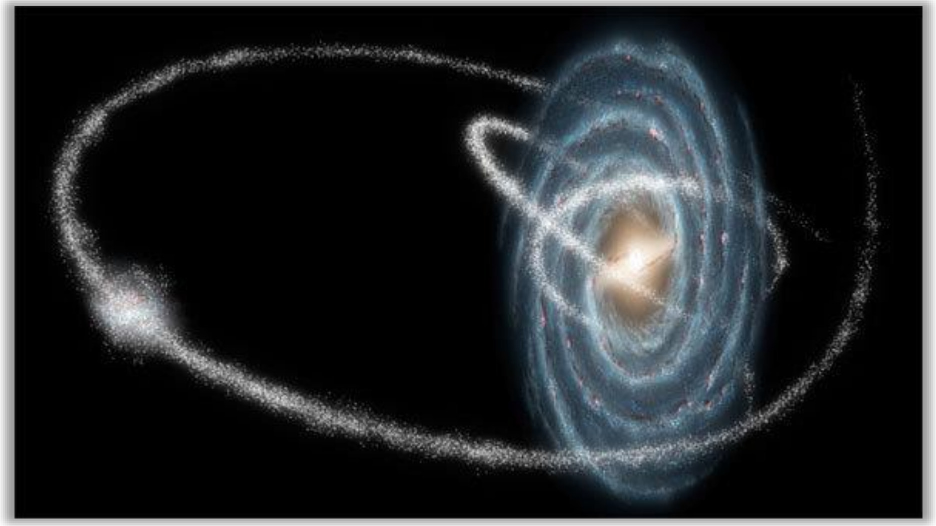
CURTAINS IN EXPERIMENTAL HEP

- Data driven method ~ few assumptions about potential signal.
    - Resonant in some feature ~invariant mass
- Deployed in ATLAS for a model agnostic search – ongoing.

CURTAINS IN THE SKY

STELLAR STREAMS

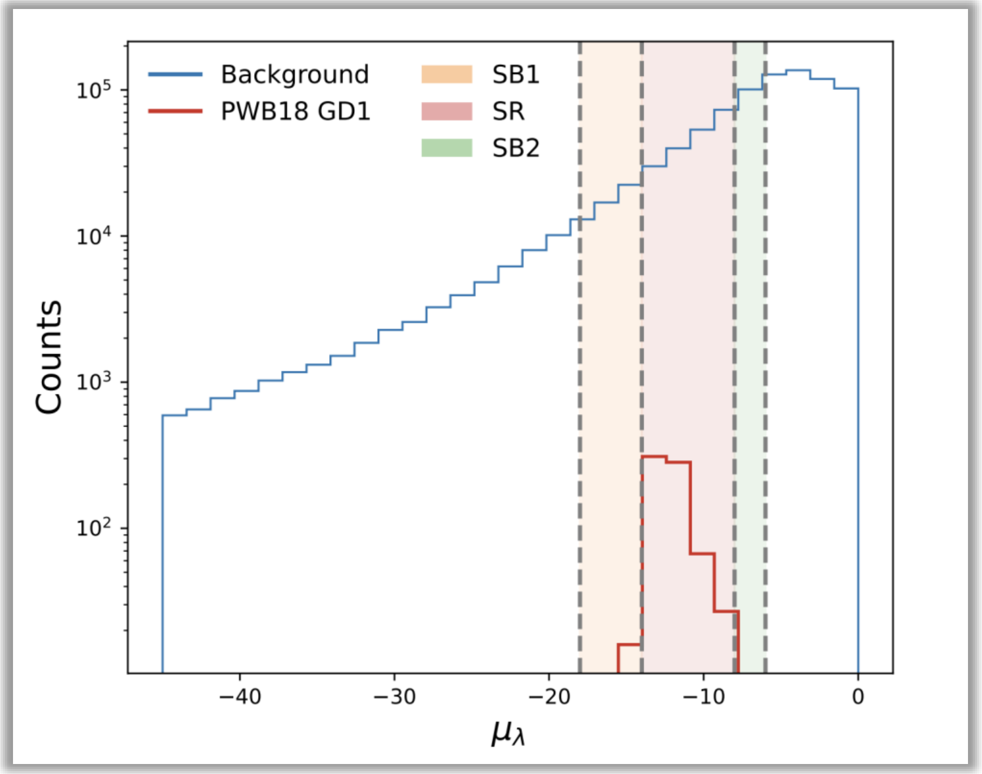
- Tidally stripped remnants of a dwarf galaxy
- Excellent probe for:
  - Galaxy merger history, formation
  - Galaxy mass, potential
  - **Dark Matter Subhalos Mass Function**
    - Density perturbations along the stream indicate subhalo flyby
    - Nature of DM ~compact, self interacting?
- Typically O(100)-O(1000) stars in a stream
- Know about 40* streams in the Milky Way
  - Expensive searches
  - Mostly model dependent ~ Assume MW potential, Chemical composition
  - Can we do this in a model agnostic way?



Artist rendition of stellar streams overserved from the Milky Way Galaxy

PROBLEM PARALLELS

- Stellar Stream search is essentially an overdensity search in a feature space
- Group of stars moving congruently ~ produce overdensities in proper motion
- Features:
    - Proper motion in the sky
    - Position in the sky
    - Color of the star
    - Luminosity of the star
- Gaia Survey: > 1.8 billion sources catalogued in the Milky Way Galaxy
- Benchmark CURTAINs!

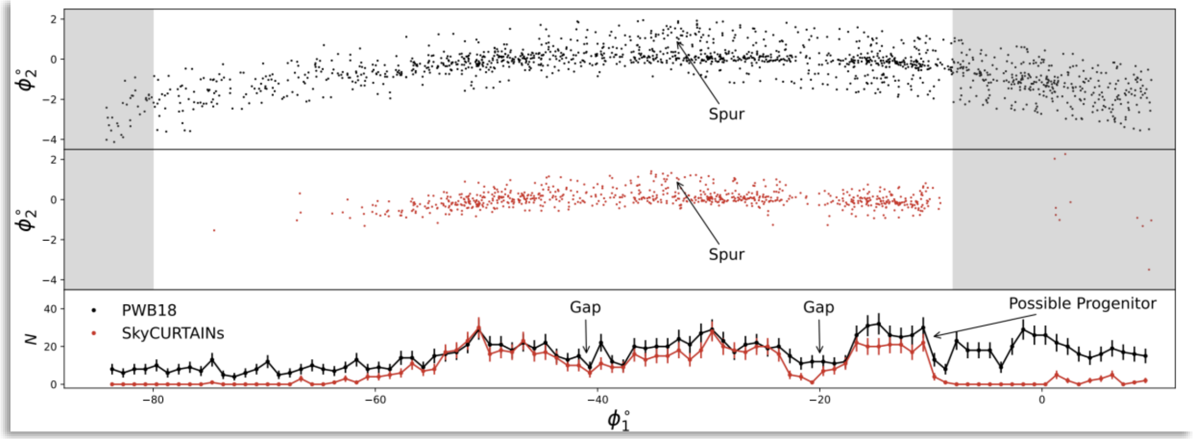


Distribution of proper motion of bg sources and members of a known stream GD-1

THE GD-1 STREAM

- Well known, narrow, large stream in the galaxy.
- CURTAINs finds GD1 in different patches of the sky with a very high purity.
- Recovers the density perturbations within the stream.
  - Without any prior model dependence!

SkyC Recovered GD1 stream in stream aligned coordinates



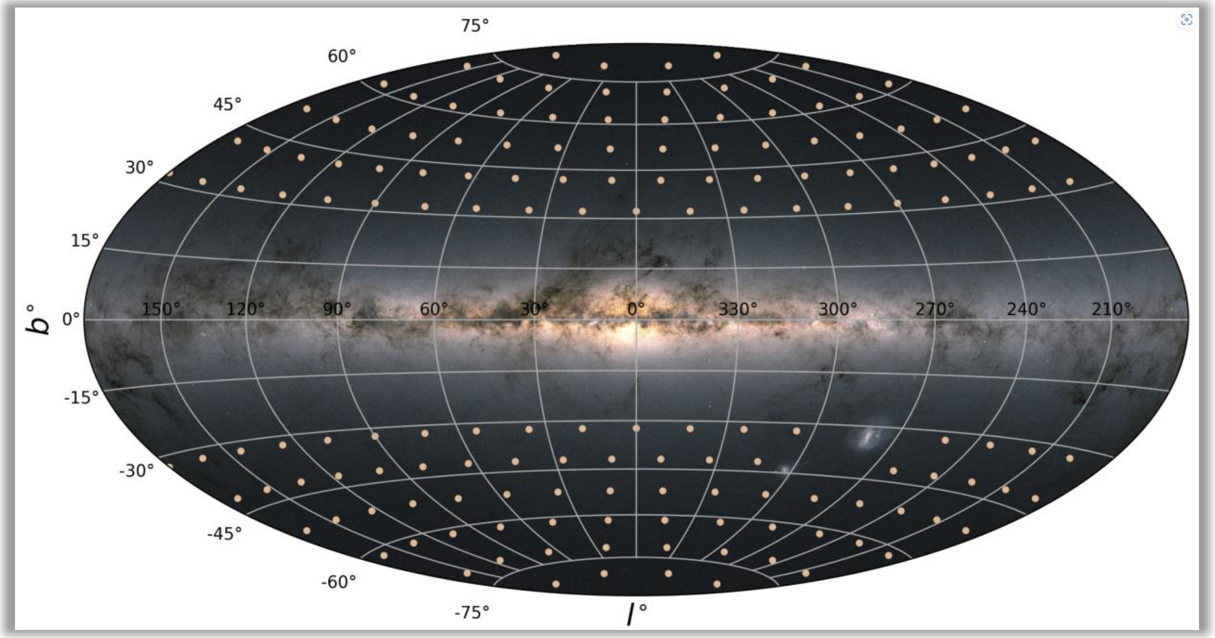
Patch (α, δ)	p SkyCURTAINs
(128.4°, 28.8°)	**82.99**
(132.6°, 16.9°)	**78.05**
(136.5°, 36.1°)	**90.56**
(138.8°, 25.1°)	**90.79**
(142.7°, 14.5°)	**86.79**
(146.9°, 35.6°)	**91.79**
(148.6°, 24.2°)	**94.9**
(148.6°, 47.0°)	**93.15**
(156.2°, 57.5°)	**70.14**
(156.9°, 34.1°)	**88.17**
(160.5°, 45.5°)	**87.43**
(171.4°, 43.0°)	**89.52**
(171.8°, 54.7°)	**89.66**
(174.3°, 65.1°)	**64.94**
(185.4°, 50.0°)	**84.0**
(192.0°, 58.7°)	**83.87**
(138.1°, 5.7°)	0.0
(203.7°, 49.1°)	**0.13**
(212.7°, 55.2°)	0.0
(224.7°, 60.6°)	**2.58**
(202.4°, 66.5°)	0.0

Purity of GD1 recovered per patch

2405.12131

A FULL SKY SCAN

- Currently being deployed on GDR3 data of more than 1.8 billion sources.
  - Each dot represents the center of a 15-degree circular patch ~ 154 patches
- Expect to recover known streams
- Discover and catalogue new streams



Molleweide projection of the Milky Way Galaxy and the patch centers for the search

OUTLOOK

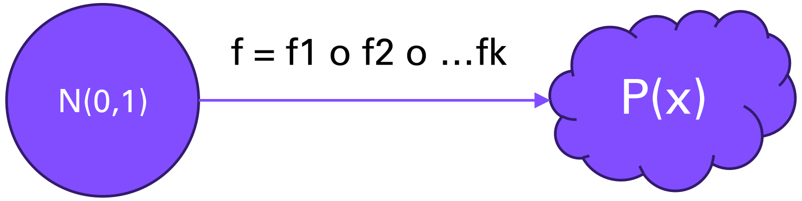
- Data driven methods powerful, versatile ~ Applicable across different domains
- Abstracting problems ~ Your problem may have been solved two doors down the office!

BACKUP

FLOWS FOR FLOWS

NORMALISING FLOWS

- Map a known distribution to an arbitrary distribution through change of variables
- $p_{\theta\phi}(x|c_x, c_y) = \pi_\theta(z) \cdot det\left|J_{f_\phi(x|c_x)}\right|$



FLOWS FOR FLOWS

- To map between two arbitrary distributions under exact likelihood – employ the change of variables formula.
  - $\mathrm{p}_{\theta\phi\gamma}(\mathrm{y} \mid \mathrm{c_x},\ \mathrm{c_y}) = \mathrm{p}_{\theta\phi}(x|c_x, c_y) \cdot \det\left| J_{f_\gamma}(x|c_x, c_y) \right|$
- But we do not know p(x) → learn it with another flow!

CURTAINS DIJET

TRAINING

- Mix SB1,SB2 – conditionally learn to transform random pairs of data.
  - Condition on some function of mjj1, mjj2.
- First train BaseFlow for ~100 epochs.
- Freeze BaseFlow, train TopFlow~10 epochs.
- *EfficientMode :* When scanning multiple SR – train one BaseFlow, use it everywhere.

† Timing is for the nominal side-bands, this would vary as the signal region changes due to total number of training events.

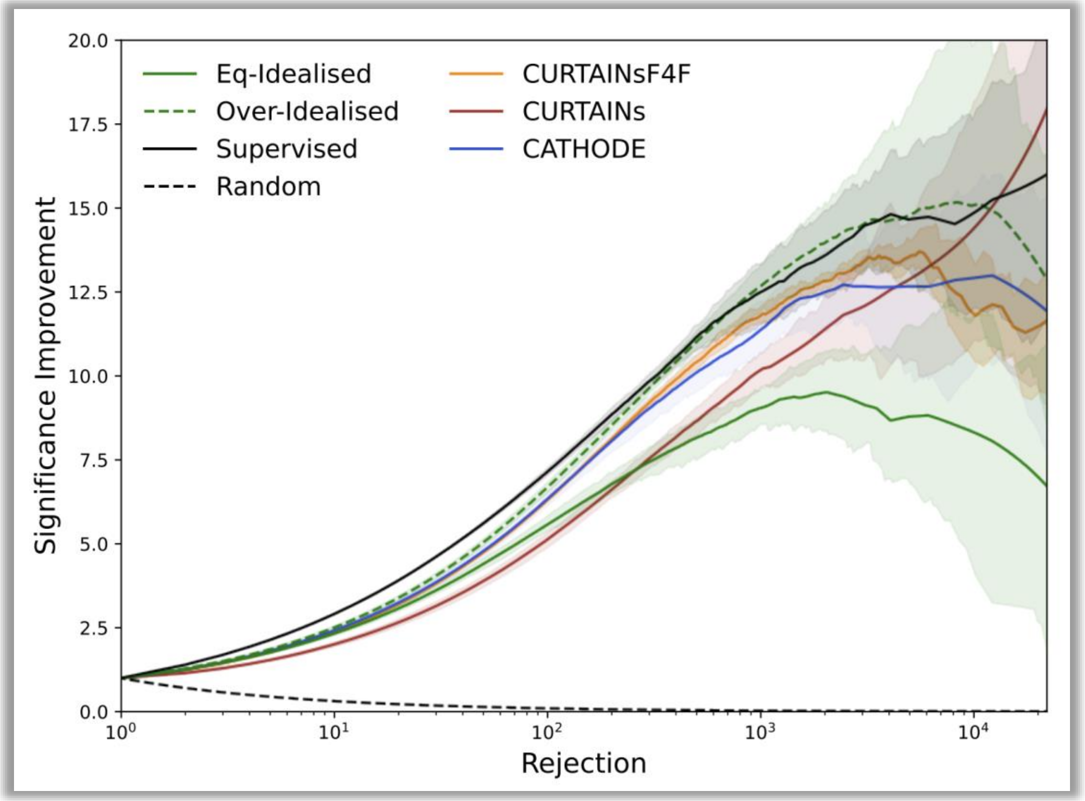
	Time / epoch [s]	N epochs	Total time [min]
	Default		
Base	32.4†	100	54
Top flow	31.5†	100	53
		One Signal Region	107
		(Extrapolated†) Ten Signal Region	1070
	Efficient		
Base	104.2	100	174
Top flow	21.3†	20	7
		One Signal Region	181
		(Extrapolated†) Ten Signal Region	244
		(Extrapolated†) 125 Signal Region	1049

TEMPLATE GENERATION

- Context generation in SR is done with fitting a PDF in SB and then sampling in SR.
    - ATLAS 3 parameter function $f(x) = p_1(1-x)^{p_2}(x)^{p_3}$
- Can generate a template in any Window, provided context can be provided.

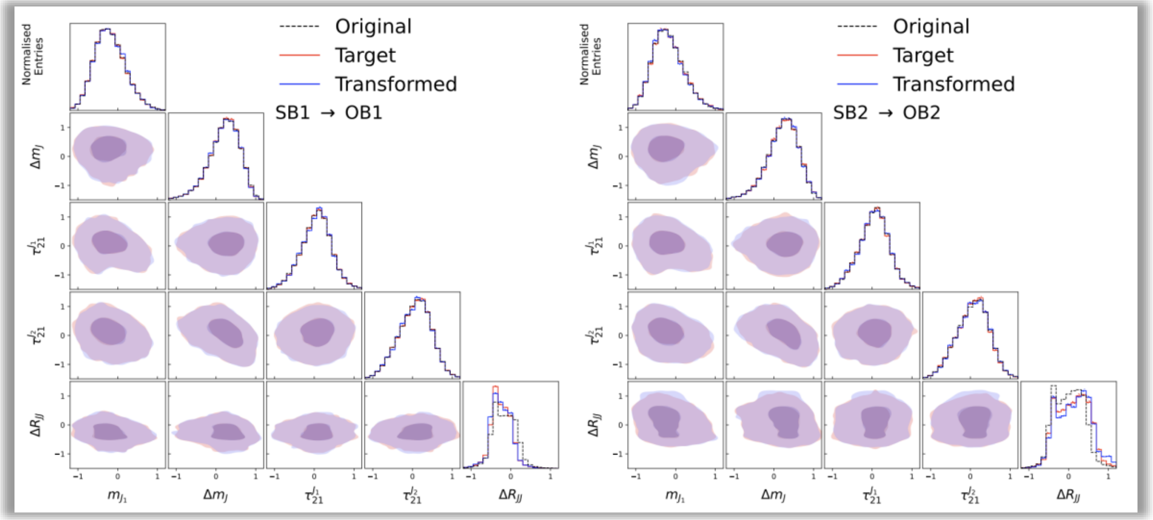
CWOLA

- Template vs SR Data in 5-Fold setup.
- MLP 32x3 for 20 epochs (overfits otherwise).
    - Tuned to get 0.5 AUC in zero doping case, good separability in 3000 doping case.



VALIDATION

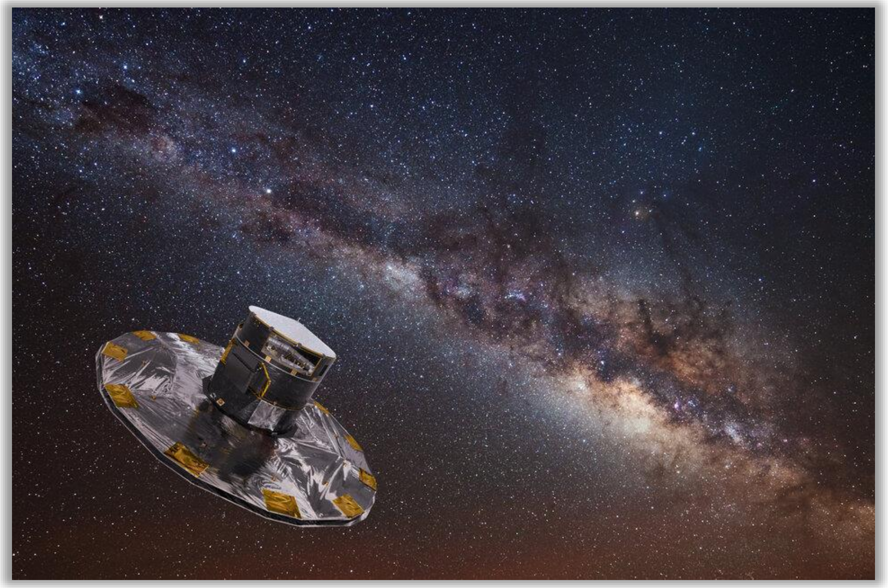
- SR blinded, check performance in other control regions.
- AUC for OB1, OB2 <= 0.52



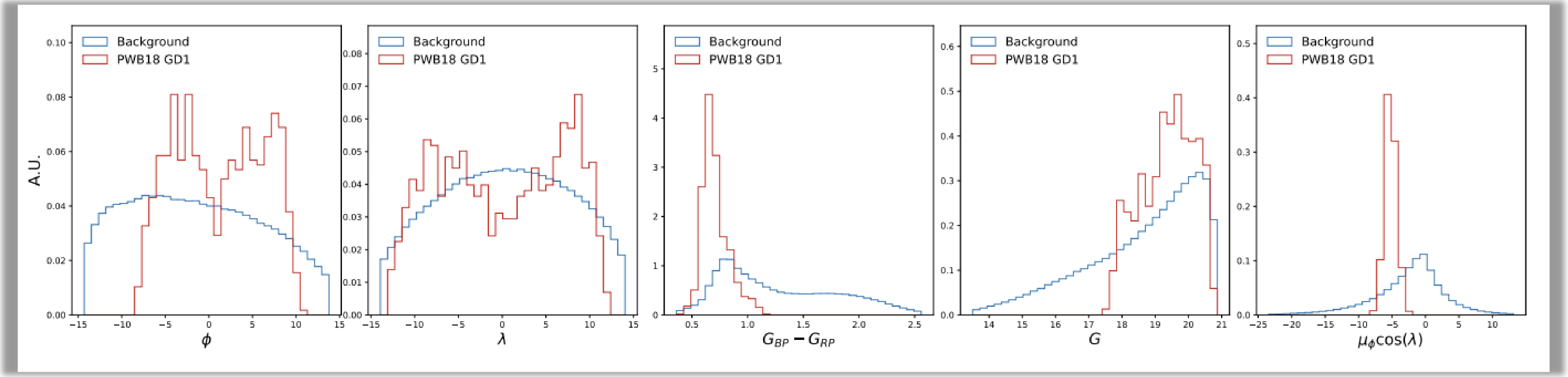
SKYCURTAINS

GAIA SURVEY

- Most precise 3D map of the Milky Way Galaxy from the L2 point.
- Astro and photometric observations of nearly 2 billion sources.
- Positions of objects as faint as magnitude 20, and those < 15, accuracy upto 24 microarcsec.
- Expected to function until 2025



FEATURE SPACE



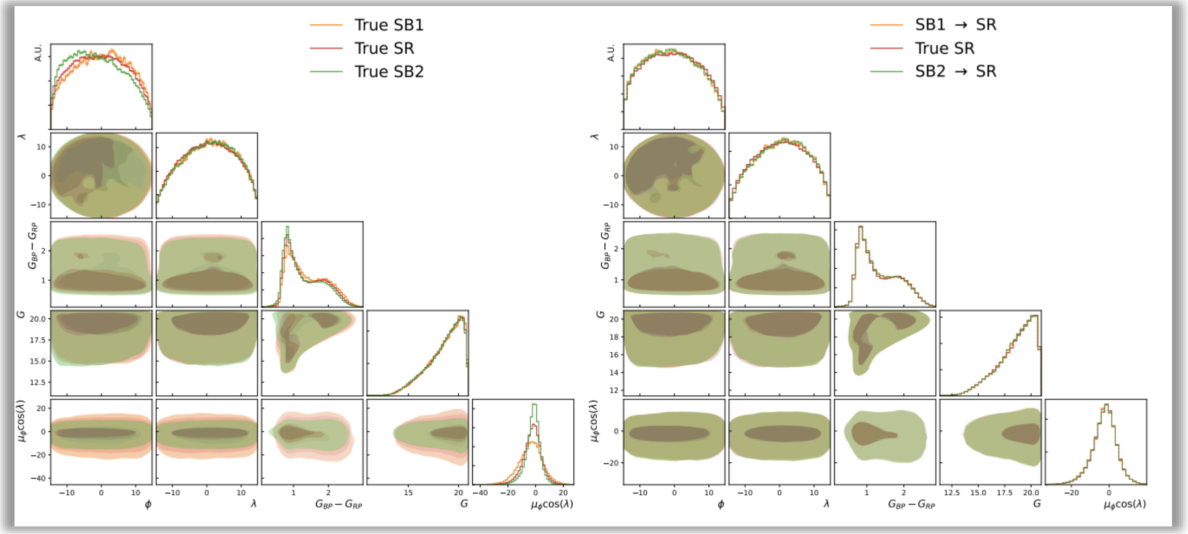
- Longitude, latitude local to patch
- Color = GBP – GRP
- Magnitude = G
- Proper motion across the sky

FIDUCIAL CUTS

- Proper motion > 2 mas/yr – reject too distant stars
- Magnitude < 20.2 – Gaia has a non-completeness fainter than this
- Color in (0.5, 1.2) – select out stars with similar metallicity.

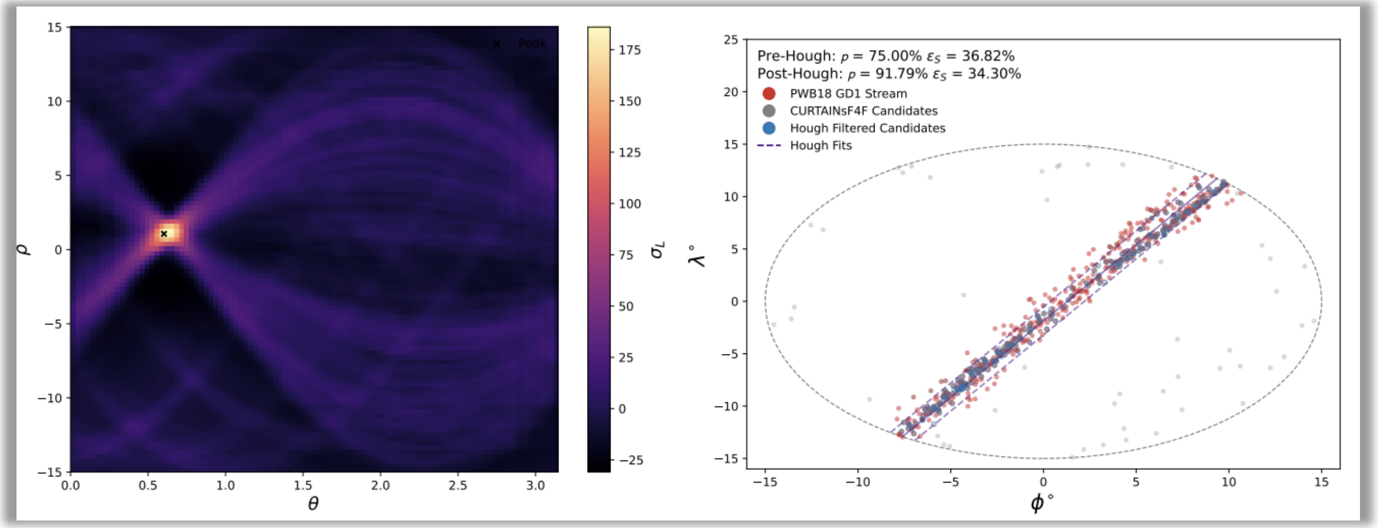
TEMPLATE FIDELITY

- AUC ~ 0.51
- Decent for downstream CWoLa



HOUGH FILTERS

- $\rho = \phi' \cos\theta - \lambda' \sin\theta$
- Line candidates in image space ~ points in parameter space.
- Finding line = finding overdensity in parameter space.



Hough space for curtains candidates (left), hough filtered candidates in blue (right)