Artificial Intelligence
Institute for Cosmic Origins
Funded by the U.S. National Science Foundation
and the Simons Foundation

A Rotation Invariant Galaxy Signature

and a fast nearest neighbor browser

Vectorized features for downstream ML
and interactive retrieval

Hamid Shafieasl[1] · Omkar Bait[2] · Jeff M. Phillips[1]

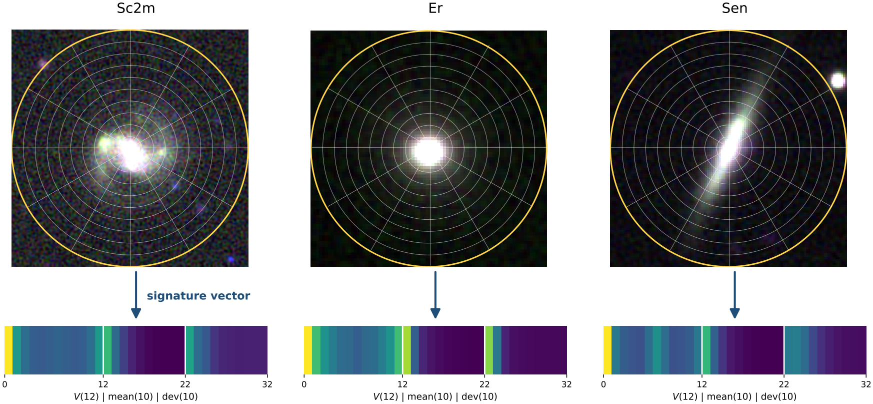
h.shfieasl@utah.edu · obait@nrao.edu · jeff.m.phillips@utah.edu

[1] University of Utah [2] NRAO

The challenge

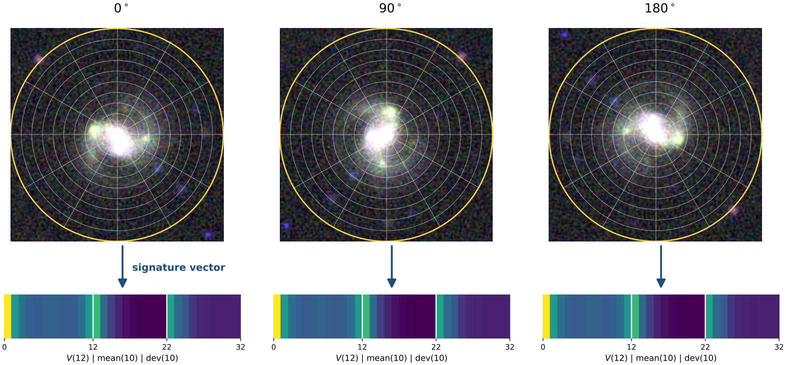
- Galaxy Zoo 2 (GZ2) gives us $\sim$**240,000** galaxies, each imaged at an arbitrary orientation in three SDSS bands (g, r, i).
- We want one representation that is
  - **fast to compute**,
  - **rotation invariant by construction**,
  - **ready to use** by any downstream model, from k-NN to a small neural network.
  - **unique up to rotation:** conjecture based on our vector sketch (arXiv:2605.27498).

One Galaxy, One Vector, Rotation Invariant



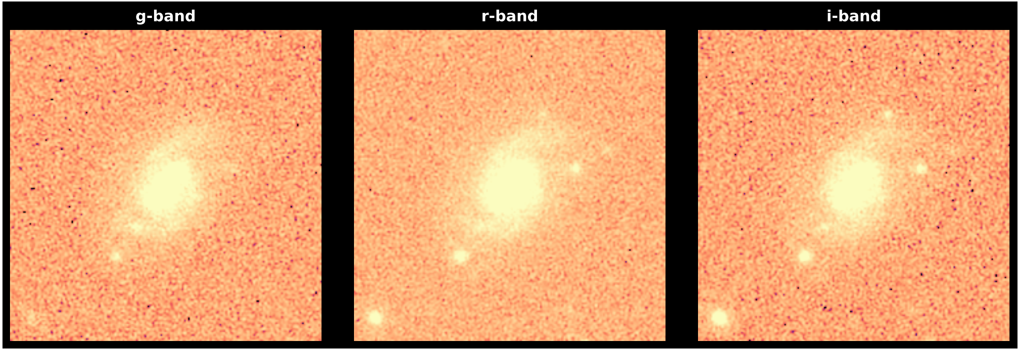
- Shown here at $N=12$, $K=10$ for visibility; the real experiment uses $N=K=100$, a 300 d signature.

Same Galaxy, Rotated: Same Vector



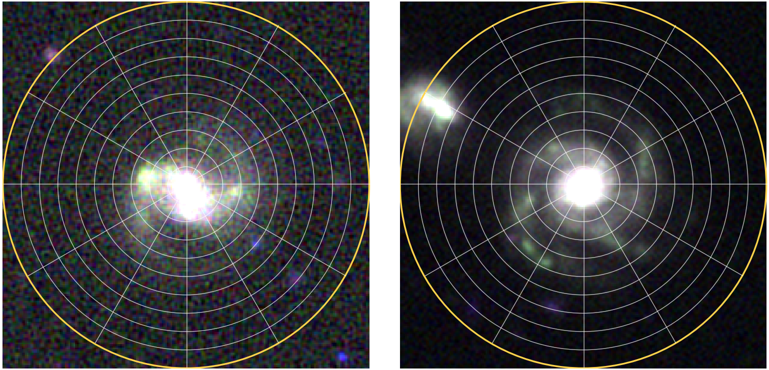
- A rotation only permutes the polar sectors.
- The three strips are *identical*: exact invariance, not approximate.

Encoding the Three SDSS Bands



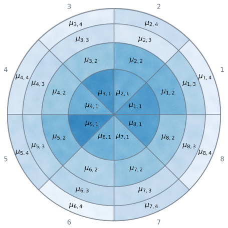
The same galaxy in the SDSS g, r, i bands: brightness and structure differ markedly between bands. We build the signature *jointly* across the three bands, so per-band structure and cross-band information are encoded together.

Construction via the Polar Grid



- Per band, pixels are binned into a polar grid of N=100 angular sectors and K=100 radial annuli.

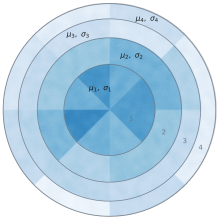
Statistics on the Polar Grid



polar grid: $N = 8$ sectors, $K = 4$ annuli

$$\Longrightarrow \quad M = \begin{bmatrix} \mu_{1,1} & \mu_{1,2} & \mu_{1,3} & \mu_{1,4} \\ \mu_{2,1} & \mu_{2,2} & \mu_{2,3} & \mu_{2,4} \\ \mu_{3,1} & \mu_{3,2} & \mu_{3,3} & \mu_{3,4} \\ \mu_{4,1} & \mu_{4,2} & \mu_{4,3} & \mu_{4,4} \\ \mu_{5,1} & \mu_{5,2} & \mu_{5,3} & \mu_{5,4} \\ \mu_{6,1} & \mu_{6,2} & \mu_{6,3} & \mu_{6,4} \\ \mu_{7,1} & \mu_{7,2} & \mu_{7,3} & \mu_{7,4} \\ \mu_{8,1} & \mu_{8,2} & \mu_{8,3} & \mu_{8,4} \end{bmatrix}$$

rows n: angular sectors   columns k: annuli



same grid: statistics pooled per annulus $(K = 4)$

radial mean and deviation profiles

$$\Longrightarrow \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{bmatrix} \qquad \sigma = \begin{bmatrix} \sigma_1 \\ \sigma_2 \\ \sigma_3 \\ \sigma_4 \end{bmatrix}$$

$\mu_k,\ \sigma_k$: over all pixels of annulus k

- Per band b, per cell: mean intensity $\to$ one polar matrix $M^{(b)} \in \mathbb{R}^{N\times K}$.
- Per band b, per annulus: mean and standard deviation $\to$ two radial profiles $\mu^{(b)}, \sigma^{(b)} \in \mathbb{R}^{K}$.

Exact Rotation Invariance via Sketching

For the three-band polar tensor $\mathcal{M} \in \mathbb{R}^{N\times K\times 3}$, define the **sketch**

$$\boxed{V_j \;=\; \sum_{i=1}^{N} \exp\!\left(-\frac{\|\mathcal{M}_i - \mathrm{shift}_j(\mathcal{M}_i)\|}{W}\right)}$$

- $V = \{V_j\}_{j=1}^{N}$ is unchanged by every cyclic shift of $\mathcal{M}$: exact for whole-sector rotations; analytic, not learned.
- **The sketch does the real work:** it captures the detailed angular structure that simple radial statistics miss, and removes *only* the orientation.
- Averaging over angles is the trivial way to be rotation invariant, and it loses the angular structure.
- **The sketch is invariant the hard way:** it removes only the orientation and keeps everything else.

Final 300d signature: $\left[\, V(100) \,\|\, \bar{\mu}(100) \,\|\, \bar{\sigma}(100) \,\right] \in \mathbb{R}^{300}$.

Properties

Vectorized	one $\mathbb{R}^{300}$ array per galaxy
Rotation invariant	analytic, not learned
No information lost	up to discretization/normalization (conjecture: arXiv:2605.27498)
Band aware	joint over the SDSS g, r, i bands
Compact	$\sim$ 1.2 KB per galaxy
Fast to index	vector database over all 240k galaxies in 3.4 s
Fast to query	0.2 ms per query, any galaxy, any K
No retraining	a new galaxy is simply inserted into the index
Interpretable	radial profile plus angular self similarity
Ready to use for ML	feeds into kNN, SVM, random forests, MLPs, ...
Browser ready	served as a public demo

Fast: a 240k-Galaxy Vector Database

0.2 ms

one query against all 240k galaxies (FAISS, CPU)

3.4 s

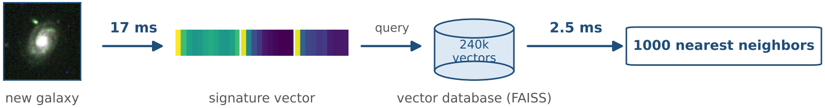
building the index for the entire survey

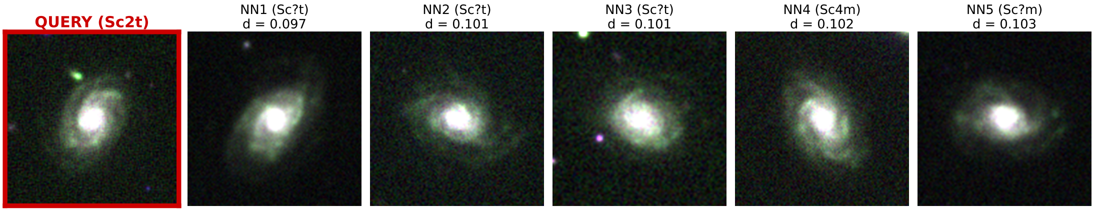
450 MB

the whole database in memory: vectors plus index

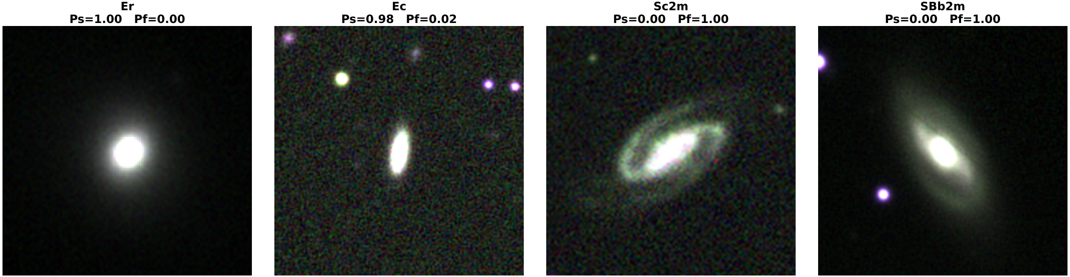
Query a New Galaxy: Vector-Database Speed

- A **brand-new** galaxy, not yet in the vector database.
- Signature in **17 ms**; its 1000 nearest neighbors from the index in **2.5 ms**. Under **20 ms** end to end.





Labels in Galaxy Zoo 2



Er round elliptical, **Ei** intermediate elliptical, **Sc2m** unbarred 2 arm spiral, **SBb2m** barred 2 arm spiral.

Two Classification Results

Binary classification with logistic regression, no fine tuning of the representation.

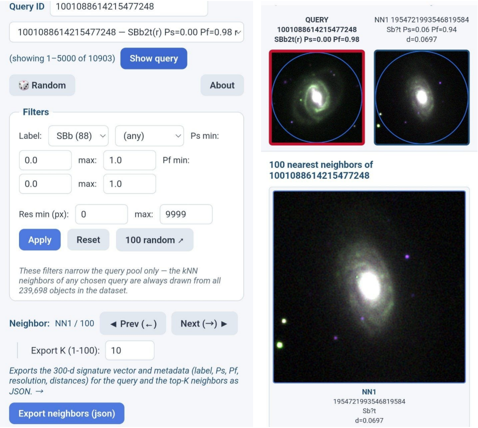
task	galaxies	accuracy
Smooth (Er, Ei) vs. Featured	2,819	**0.964**
Sc unbarred vs. SB barred spiral	1,960	**0.806**

- **Learned CNNs:** 95–99% on similar GZ2 tasks.
- **Their cost:** hours of training on rotation-augmented images.
- **Ours is untuned:** no optimization of the representation yet, so we expect these numbers to improve.

- Dieleman, Willett, Dambre: Rotation-invariant convolutional neural networks for galaxy morphology prediction.
- Zhu et al.: Galaxy morphology classification with deep convolutional neural networks.
- Abraham et al.: Detection of bars in galaxies using a deep convolutional neural network.

Try it: an interactive browser

- Pick any of the ∼240k galaxies and step through its 100 nearest neighbors.
- Filter by morphology label, smooth or featured probability, and image resolution.
- Export top K neighbors with their 300 d signatures and metadata as JSON.



Summary

- A 300d **vectorized** signature for every galaxy.
- **Rotation invariant** by construction, **band aware** across the g, r, i bands.
- Built in seconds; any new query answered in milliseconds.
- Drop-in for nearest neighbors, classical ML, and **neural networks**.
- **Public Demo**: **https://hamidmath.github.io/Galaxies/**
- Collaborate with us! What should we optimize/search/train?