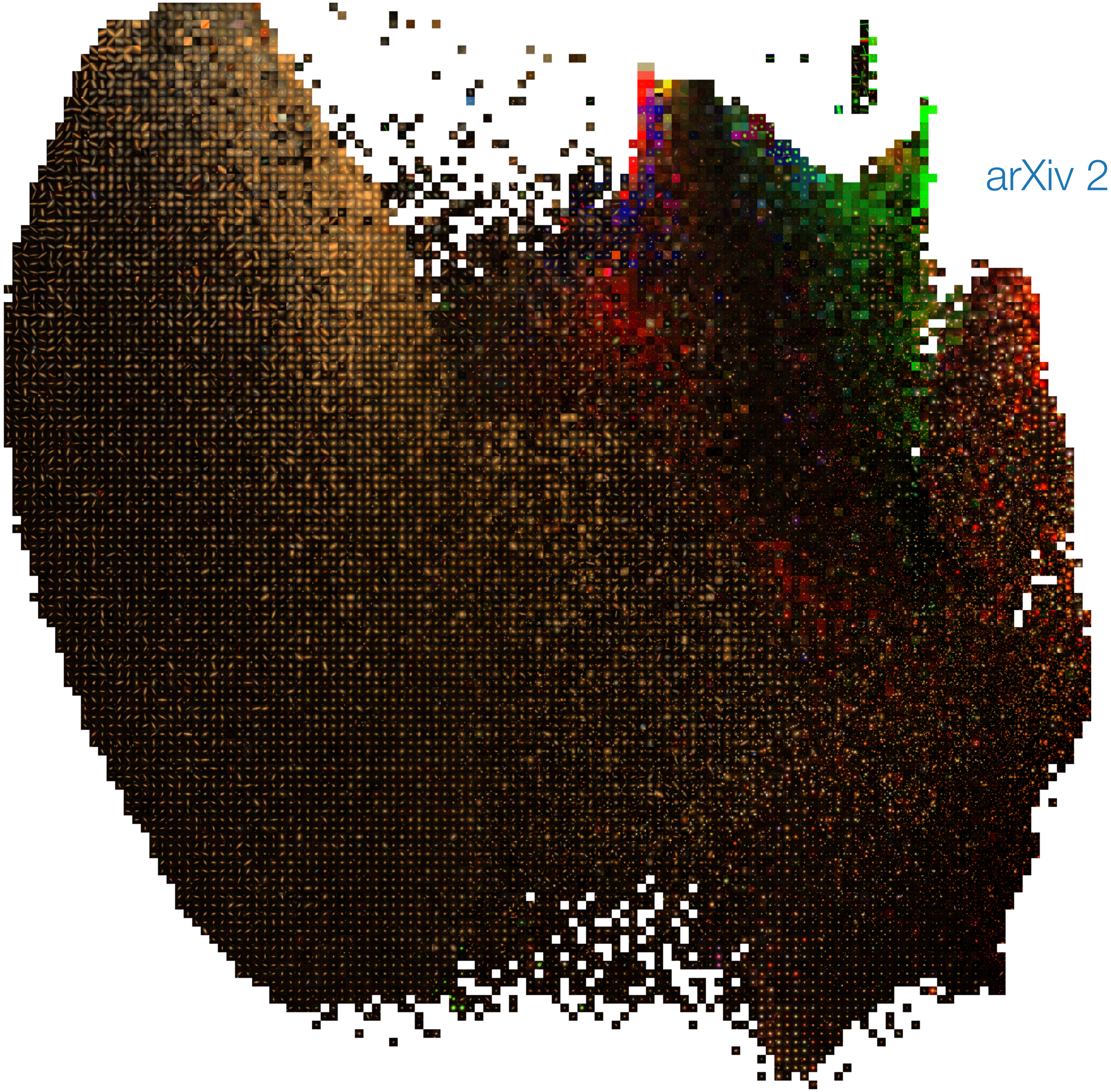
Self-Supervised Representation Learning for Astronomical Images

Md Abul Hayat mahayat@uark.edu
George Stein gstein@berkeley.edu
Peter Harrington pharrington@lbl.gov
Zarija Lukić zarija@lbl.gov
Mustafa Mustafa mmustafa@lbl.gov



*blue = clickable links throughout

arXiv 2012.13083

Sky Surveys

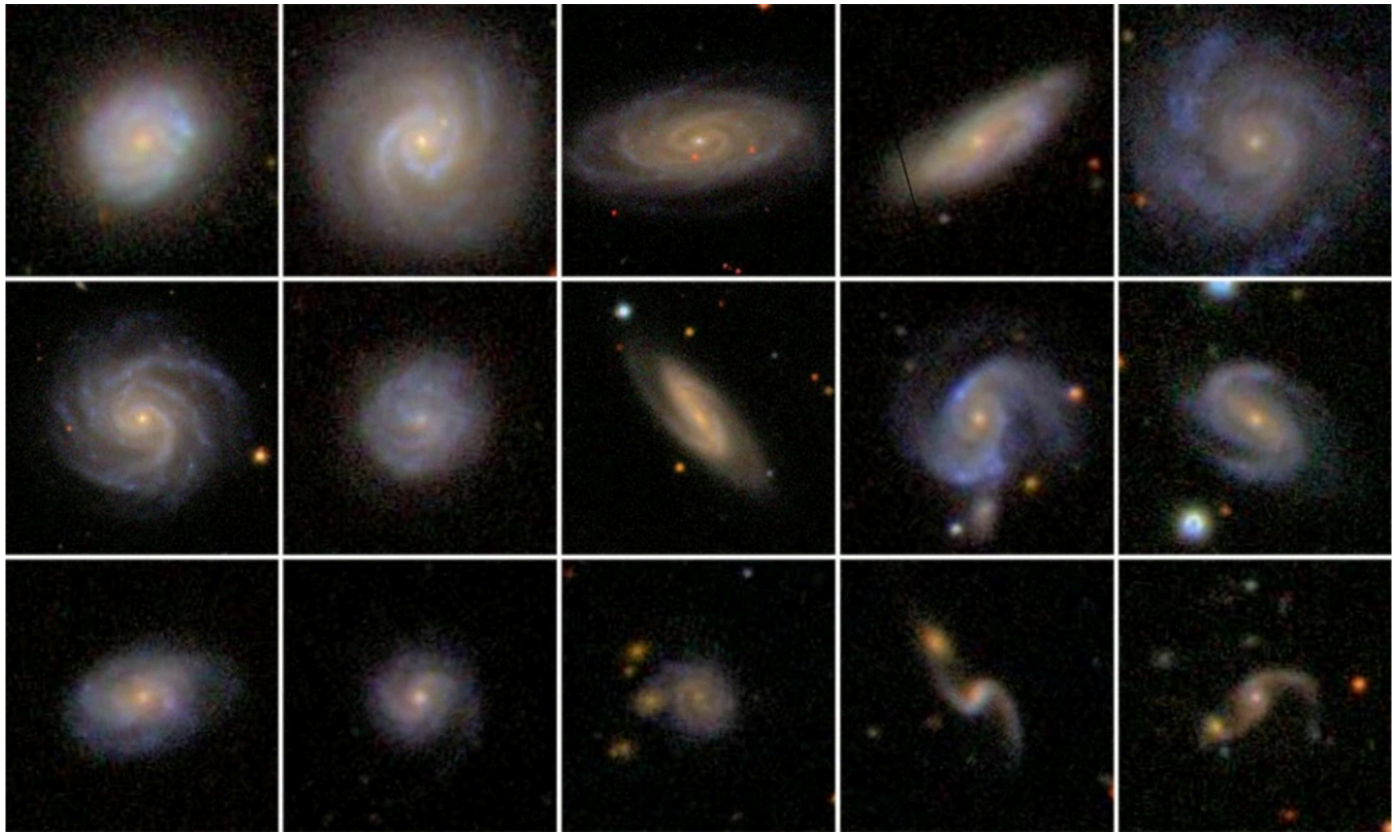
Sky surveys are massive data generators,
imaging ~10 billion galaxies in the near future

Organize
&
data discovery

Find rare objects
&
detect anomalies

Classification tasks
e.g.
Galaxy type

Regression tasks
e.g.
distance/redshift

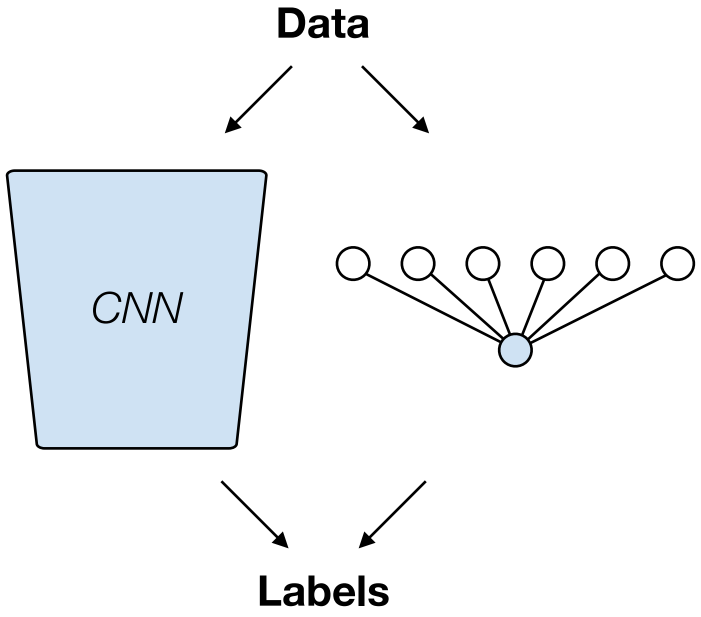


Galaxies from the Sloan Digital Sky Survey (SDSS)

Supervised learning

Target a specific task and collect labels

Organize & data discovery | Find rare objects & detect anomalies | Classification tasks e.g. galaxy type | Regression tasks e.g. distance/redshift



1. Design architecture
2. Split data into train/validate/test
3. Train model weights through back propagation
4. Deploy model, assuming unlabelled data comes from the same distribution

Supervised learning

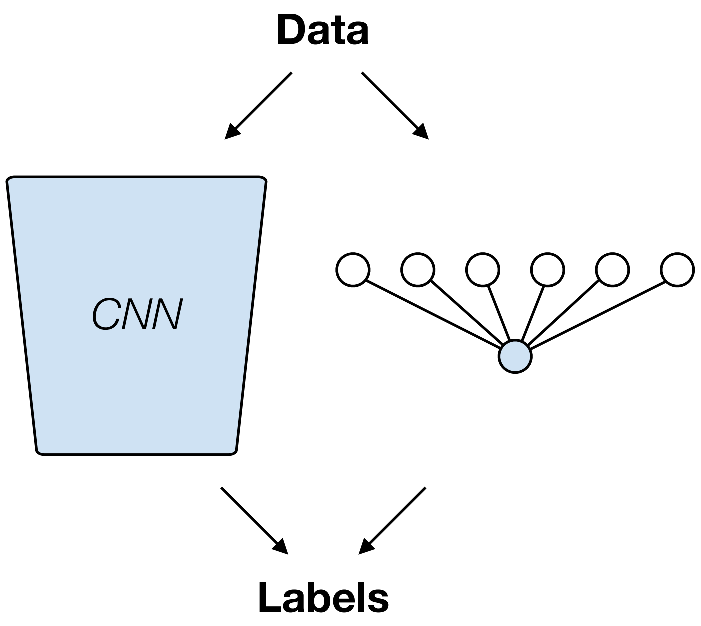
Target a specific task and collect labels

Organize
&
data discovery

Find rare objects
&
detect anomalies

Classification tasks
e.g.
galaxy type

Regression tasks
e.g.
distance/redshift



1. Design architecture
2. Split data into train/validate/test
3. Train model weights through back propagation
4. Deploy model, assuming unlabelled data comes from the same distribution

Unsupervised learning

Clustering, dimensionality reduction and feature selection, etc..

Self-supervised representation learning

Without any labels, learn low-dimensional representations of data which preserve semantic information

Then, use representations for “downstream tasks” (regression, classification, etc…)

Self-supervised representation learning

Without any labels, learn low-dimensional representations of data which preserve semantic information

Then, use representations for “downstream tasks” (regression, classification, etc…)

Self-supervised Learning: Generative or Contrastive

Generative

- Autoencoder, VAE
- Flow-based
- …

Self-supervised representation learning

Without any labels, learn low-dimensional representations of data which preserve semantic information

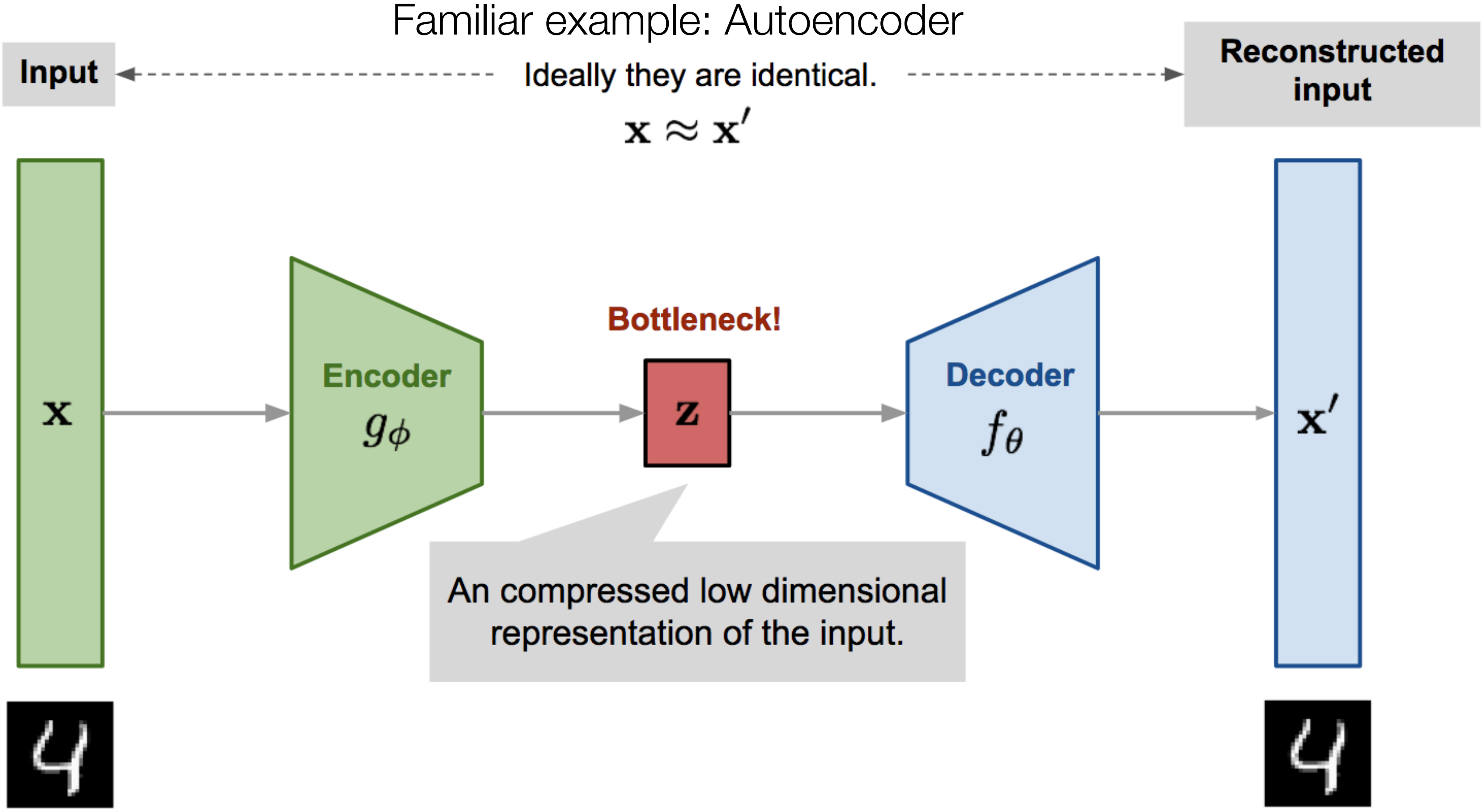
Then, use representations for "downstream tasks" (regression, classification, etc…)

Self-supervised Learning: Generative or Contrastive

Generative

- Autoencoder, VAE
- Flow-based
- …

Familiar example: Autoencoder


Input
Ideally they are identical.
$\mathbf{x} \approx \mathbf{x}'$
Reconstructed input
$\mathbf{x}$
Encoder
g_ϕ
Bottleneck!
$\mathbf{z}$
Decoder
f_θ
$\mathbf{x}'$
An compressed low dimensional representation of the input.

Train using MSE loss or similar

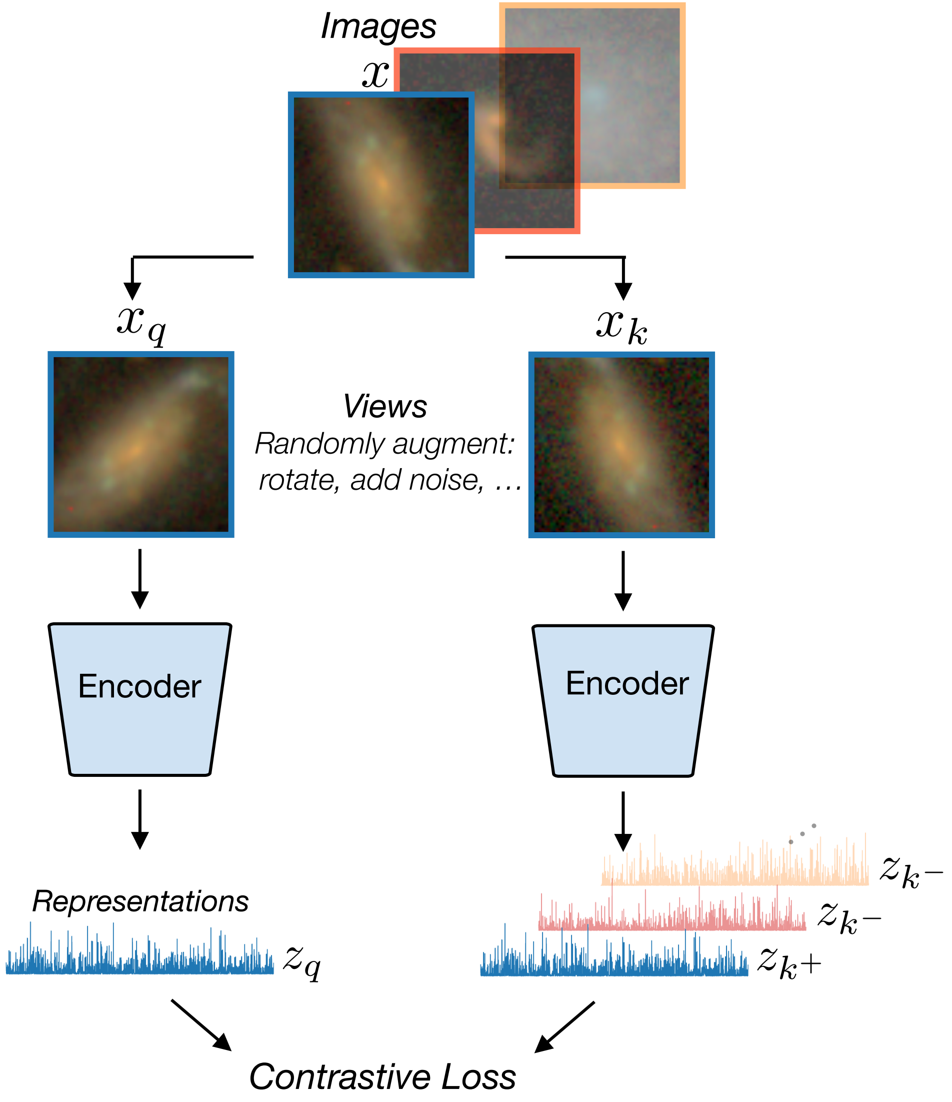
nothing to enforce that small variations in the input image result in similar representations

Contrastive learning

Learn representations that are invariant to desired augmentations

1. Self-supervised contrastive representation learning

Learn representations in an unsupervised manner



Task-agnostic augmentations for galaxy surveys:

- Rotation
- Jitter
- Gaussian noise
- Galactic extinction
- Point spread function (blur)
- …

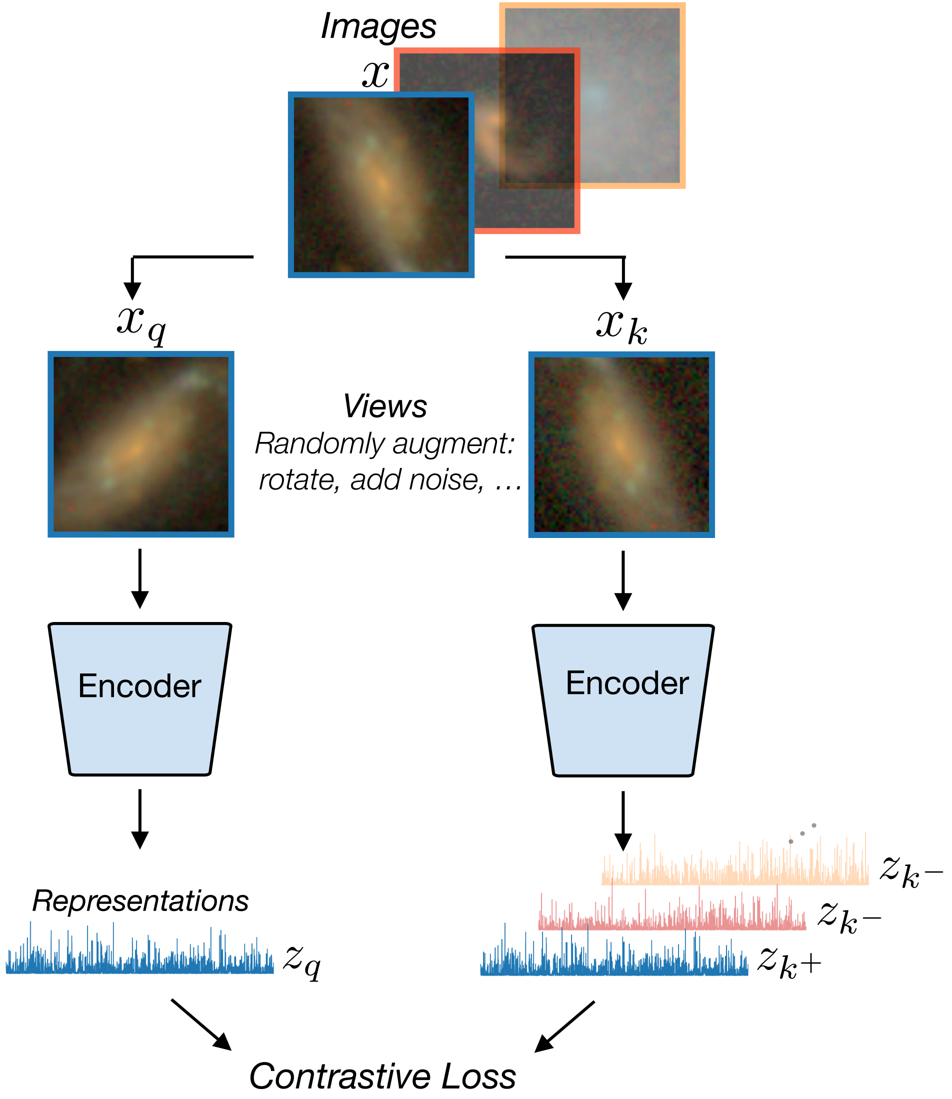
For other data/applications, choose your own using domain knowledge

Contrastive learning

Learn representations that are invariant to desired augmentations

1. Self-supervised contrastive representation learning

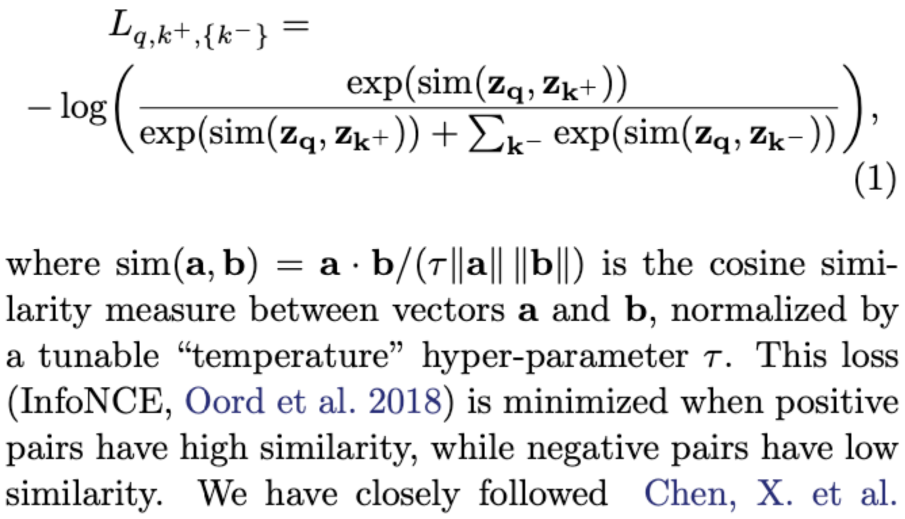
Learn representations in an unsupervised manner



Task-agnostic augmentations for galaxy surveys:

- Rotation
- Jitter
- Gaussian noise
- Galactic extinction
- Point spread function (blur)
- …

For other data/applications, choose your own using domain knowledge



$$L_{q,k^+,\{k^-\}} = -\log\left(\frac{\exp(\mathrm{sim}(\mathbf{z_q},\mathbf{z_{k^+}}))}{\exp(\mathrm{sim}(\mathbf{z_q},\mathbf{z_{k^+}})) + \sum_{\mathbf{k^-}}\exp(\mathrm{sim}(\mathbf{z_q},\mathbf{z_{k^-}}))}\right), \quad (1)$$

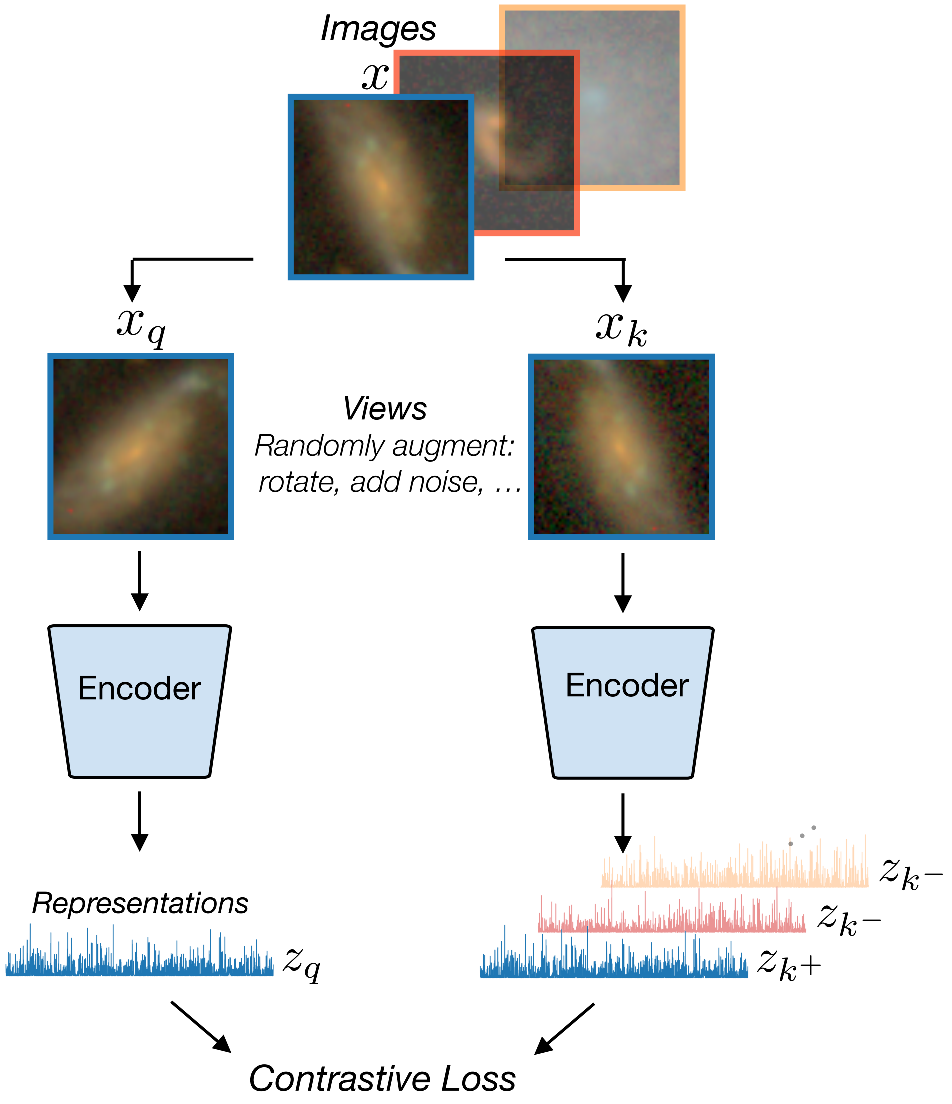
where $\mathrm{sim}(\mathbf{a},\mathbf{b}) = \mathbf{a}\cdot\mathbf{b}/(\tau\|\mathbf{a}\|\,\|\mathbf{b}\|)$ is the cosine similarity measure between vectors $\mathbf{a}$ and $\mathbf{b}$, normalized by a tunable "temperature" hyper-parameter τ. This loss (InfoNCE, Oord et al. 2018) is minimized when positive pairs have high similarity, while negative pairs have low similarity. We have closely followed Chen, X. et al.

Contrastive learning

Learn representations that are invariant to desired augmentations

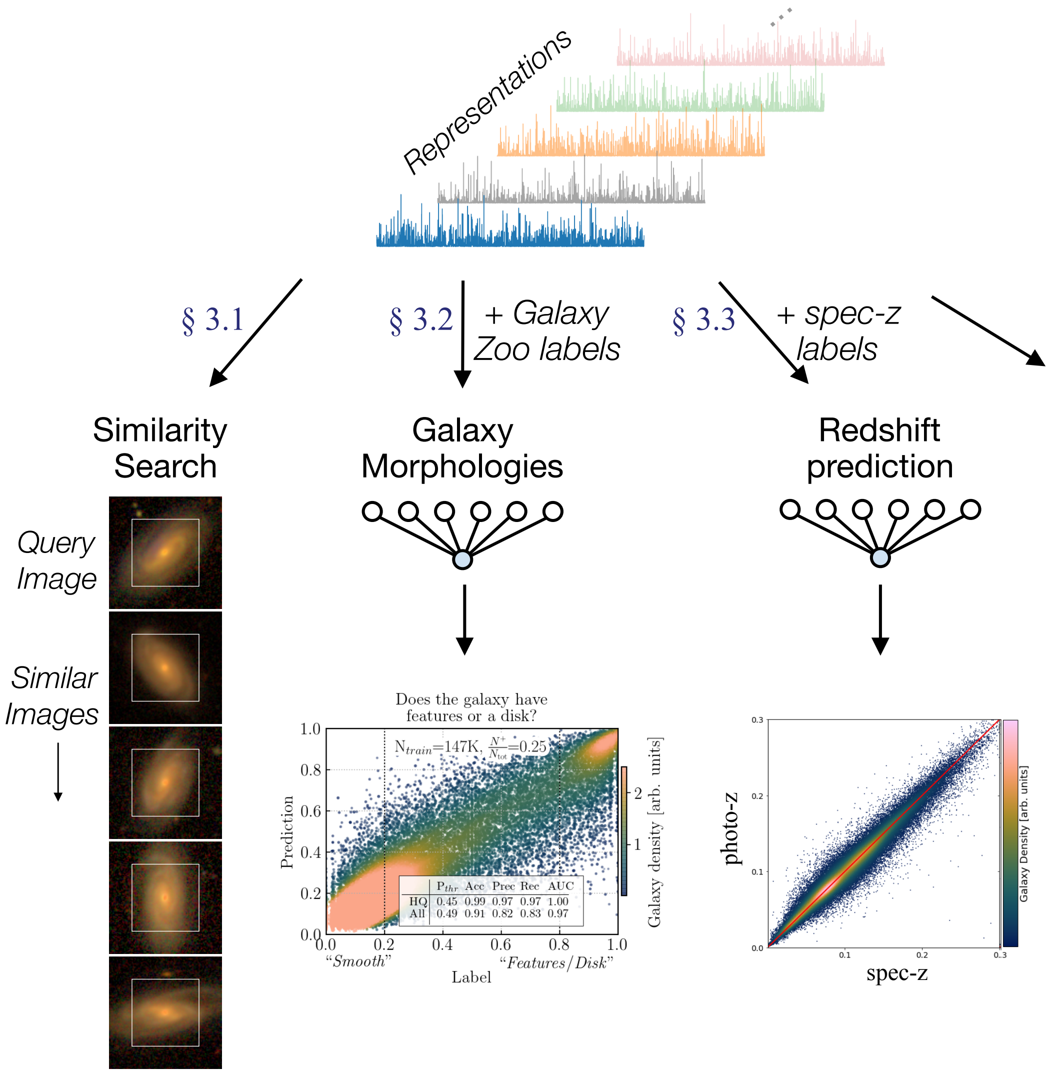
1. Self-supervised contrastive representation learning

Learn representations in an unsupervised manner



2. Downstream tasks

Use representations for a variety of applications



	P_{thr}	Acc	Prec	Rec	AUC
HQ	0.45	0.99	0.97	0.97	1.00
All	0.49	0.91	0.82	0.83	0.97

- Anomaly detection
- Strong lens finding
- Low brightness gals.
- etc…

arxiv: 2012.13083

Dataset

SDSS 5 band images (ugriz)

1.3 million images total

500k have redshift labels from spectroscopic followup

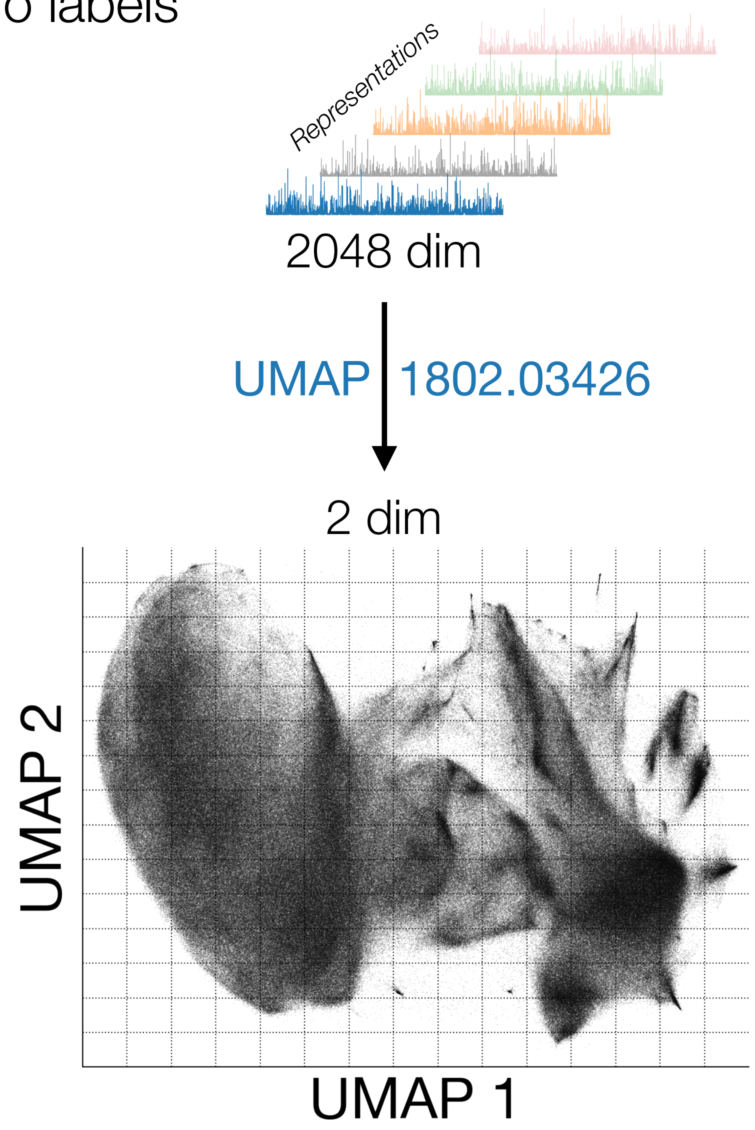
180k have crowd sourced morphology classifications

Workflow

1. Train Resnet-50 encoder through self-supervised contrastive representation learning on all 1.3M images using *MoCo v2* framework (see *SimCLR* for alternative)
2. Add in labels, and use for downstream tasks

Visualize learned representations

Self-supervised = no labels



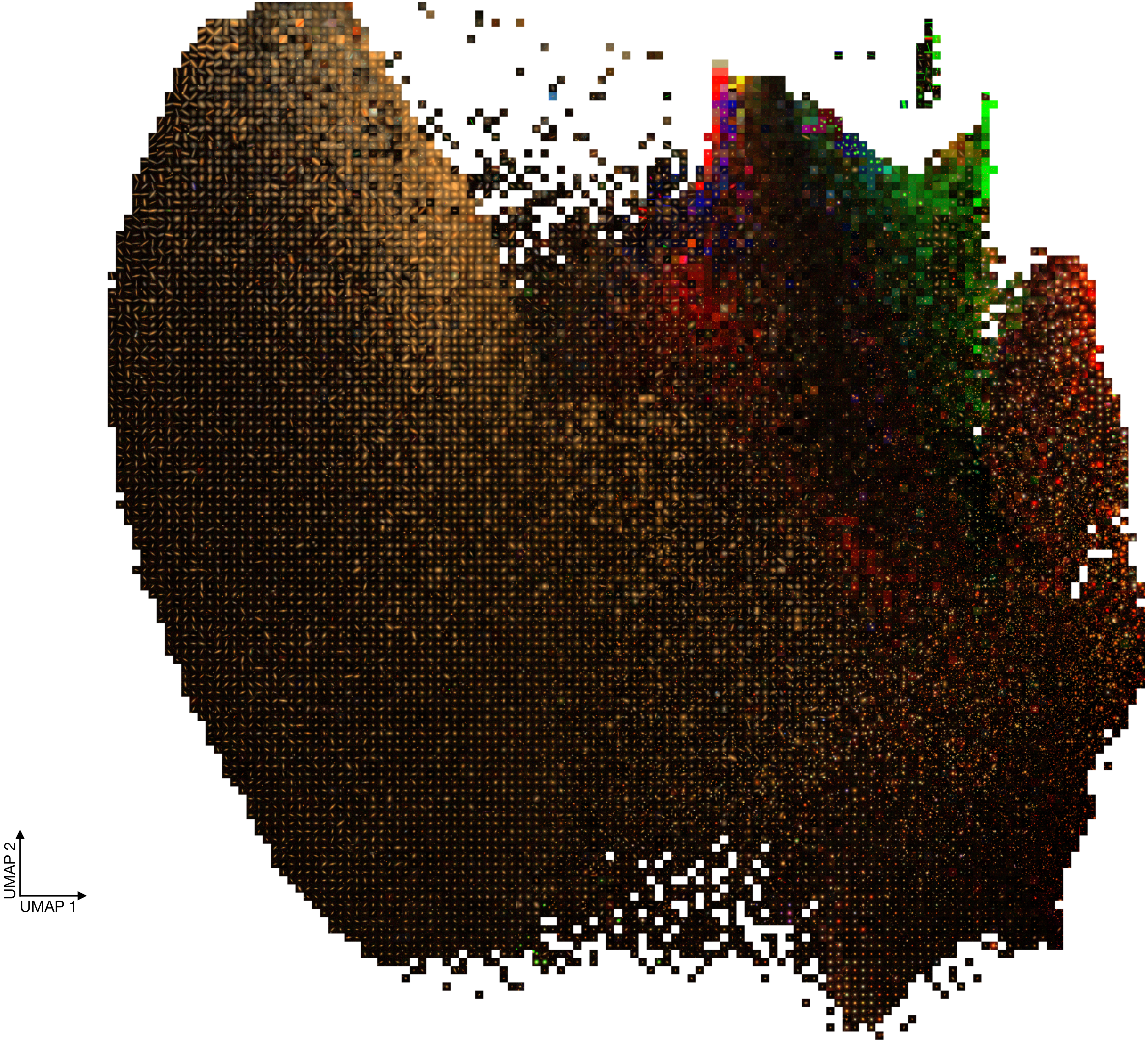
2048 dim

UMAP 1802.03426

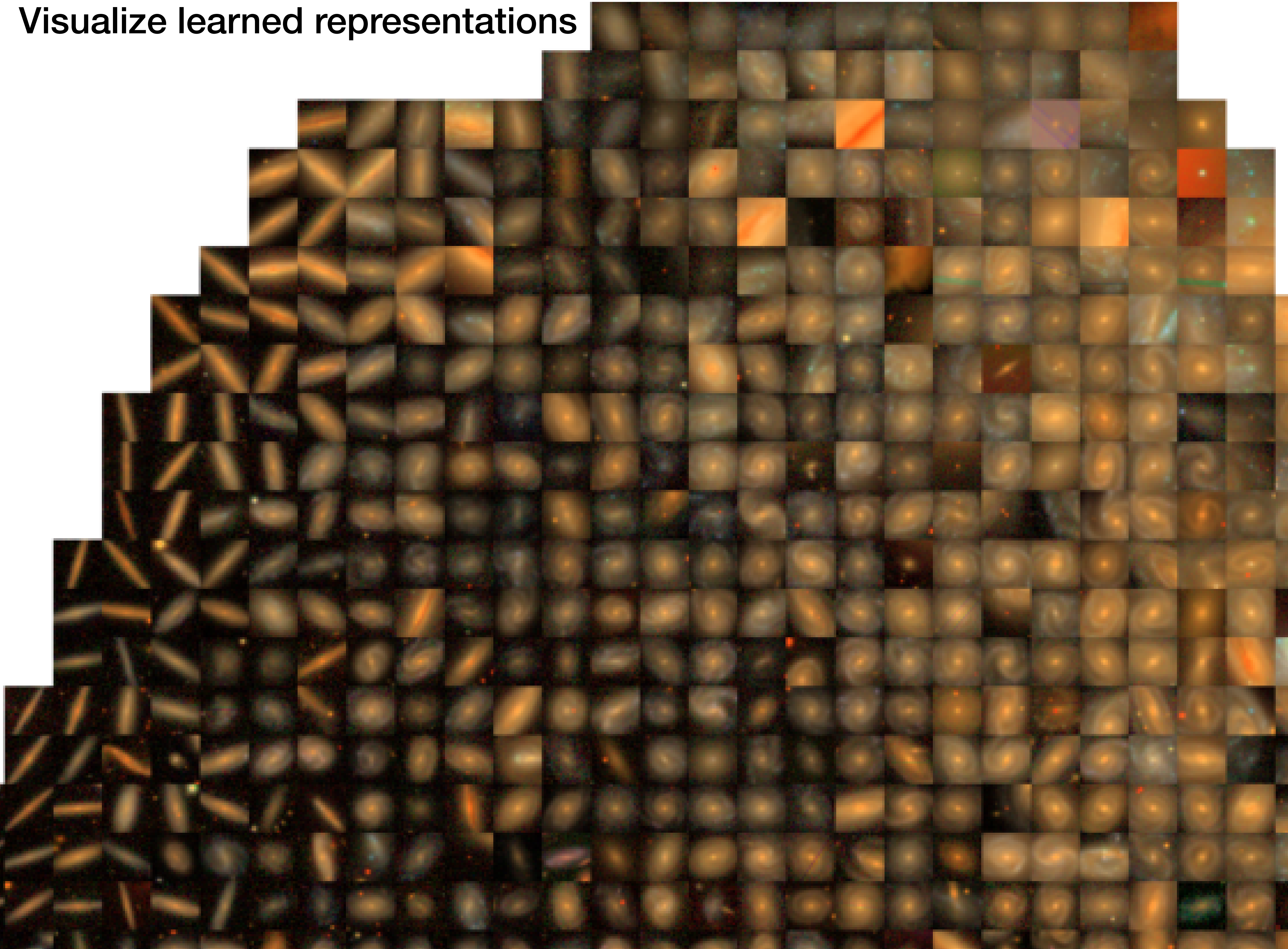
2 dim

Every data point represents a 5-band galaxy image

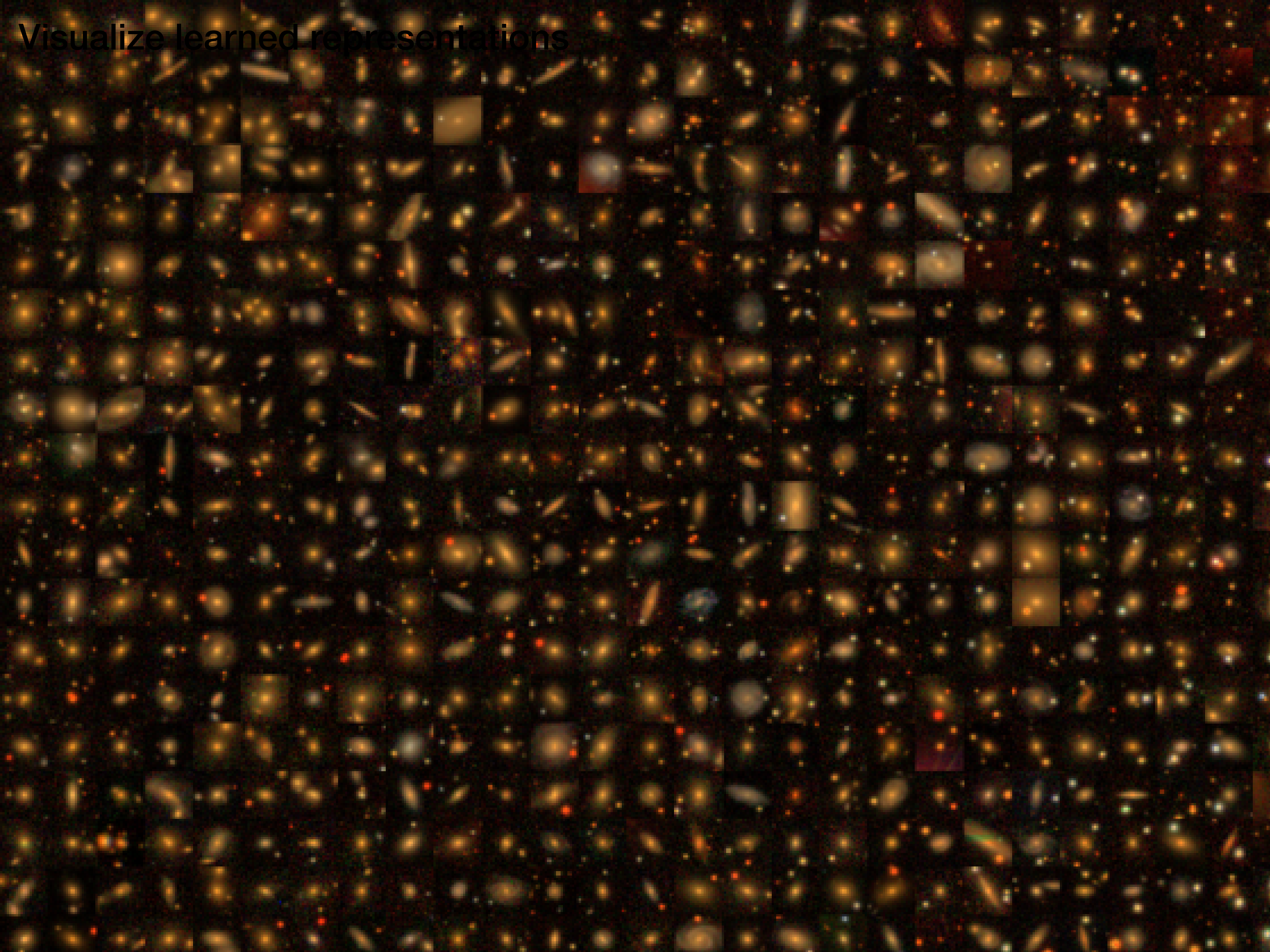
Visualize learned representations



Visualize learned representations

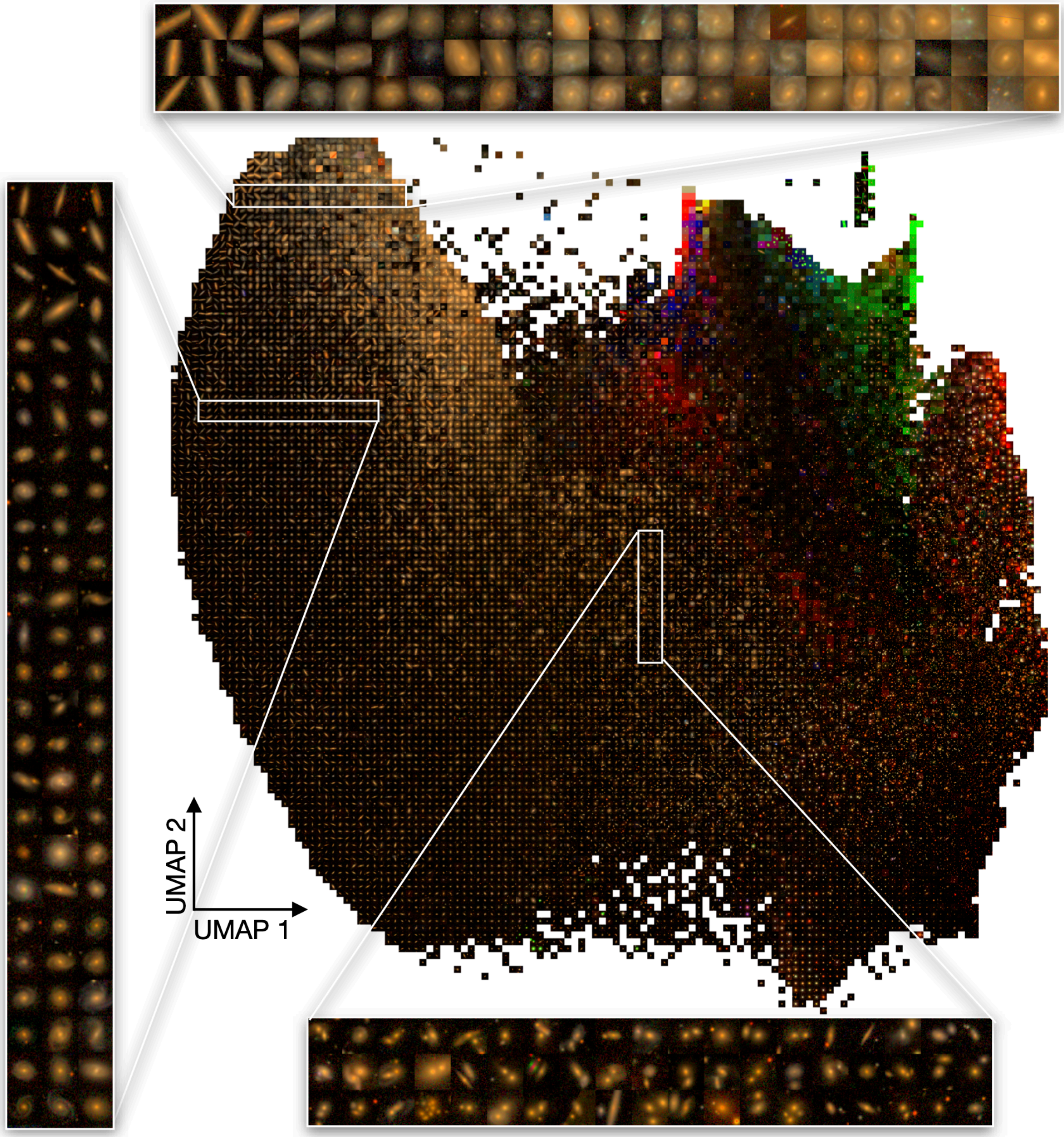


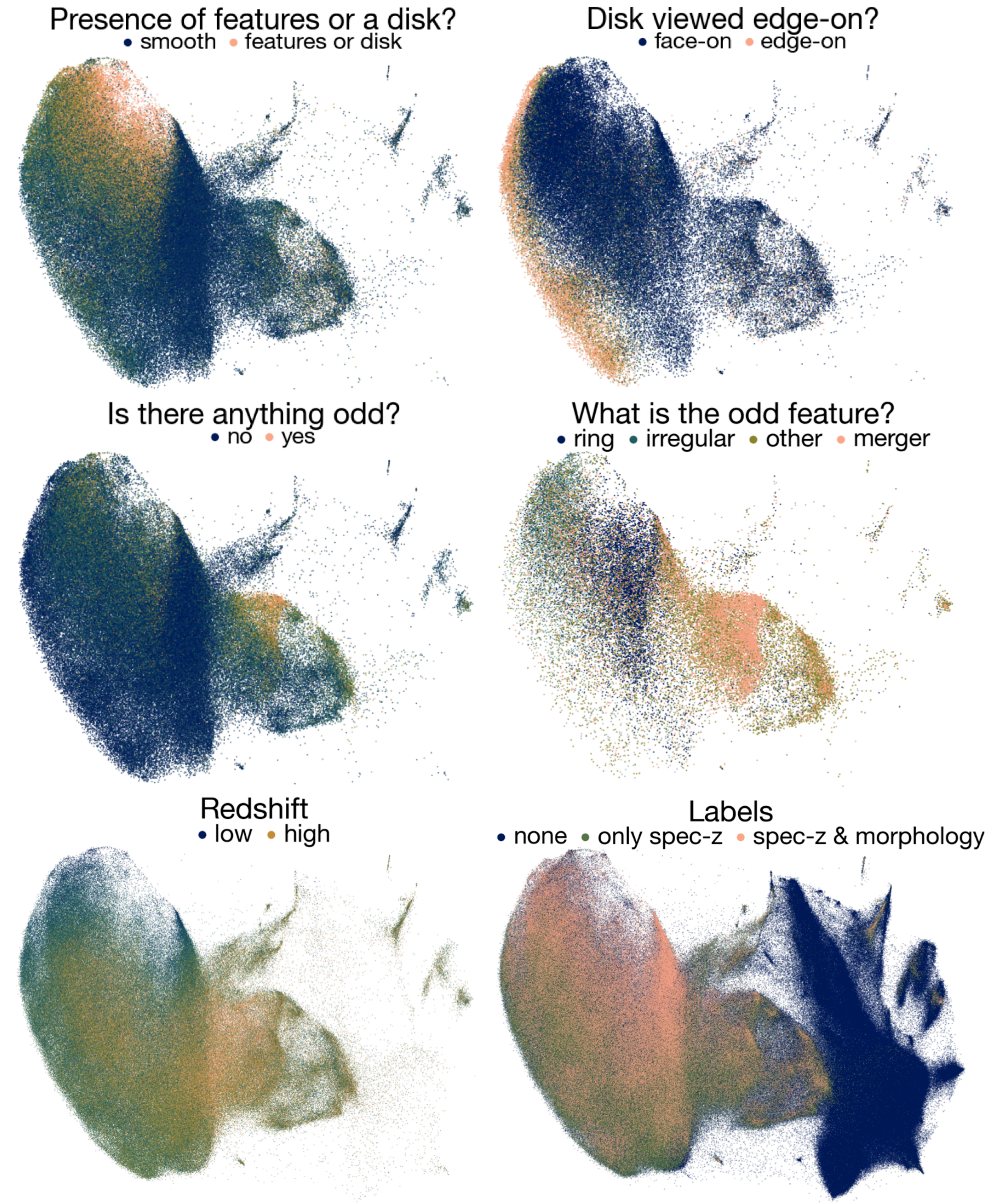
Visualize learned representations



Visualize learned representations

In context of labels





Labels have not been used during training, yet decision boundaries can almost be drawn by eye

Downstream task #1

Similarity search for data discovery

Goal: Given an image, find other “similar” images in the dataset

Approach:

1. Select desired query image
2. Compute similarity metric of query representation with all others
3. Sort similarity by decreasing order and return images

2101.04293

Downstream task #1

Similarity search for data discovery

Goal: Given an image, find other “similar” images in the dataset

Approach:

1. Select desired query image
2. Compute similarity metric of query representation with all others
3. Sort similarity by decreasing order and return images

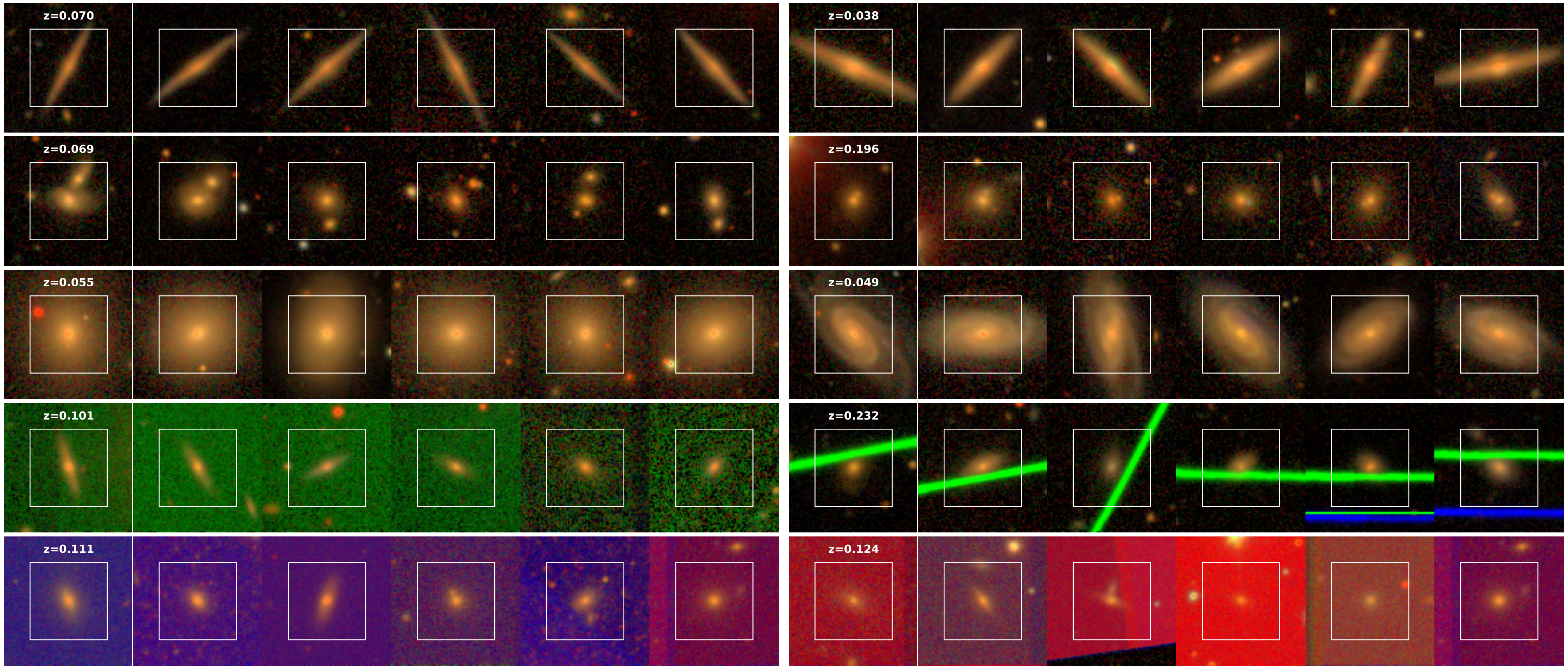


Figure 1: Reference SDSS galaxies from the validation set (leftmost panels with redshift labels) and the most similar galaxies from the training set (following 5 panels) identified through a self-supervised similarity search. White squares outline the 64^2 pixels that are “seen” by the network. 2101.04293

As the contrastive self-supervised loss was similarity-based,
the corresponding representations are by construction organized by their visual similarity

Downstream task #2

Morphology classification

Goal: predict Galaxy Zoo 2 crowd sourced morphological labels directly from representations.

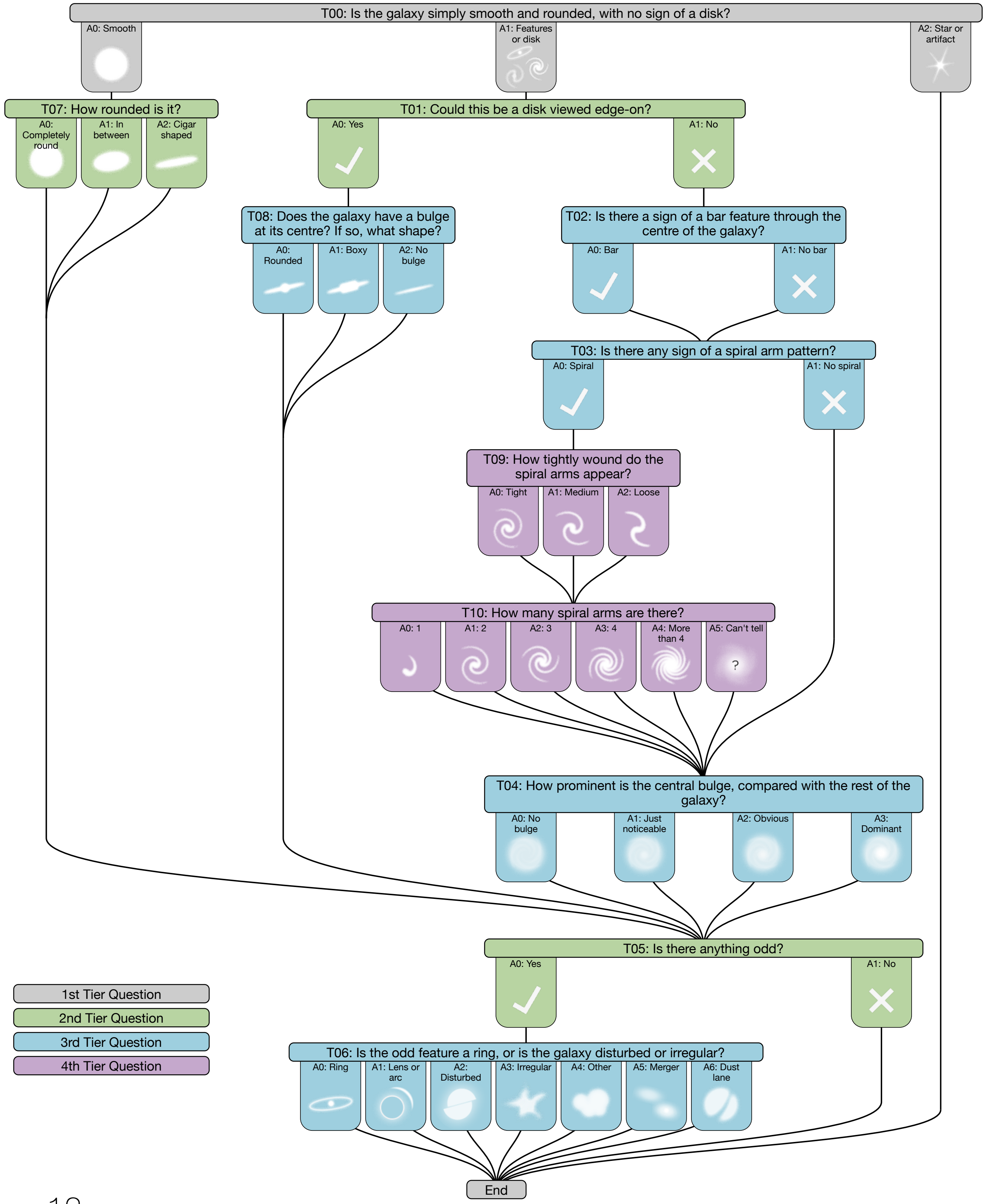
Q: “Does the galaxy have features or a disk?”
A: “smooth” or “features/disk”

Q: “Could this be a disk viewed edge-on?”
A: “Face-on” or “Edge-on”

Approach:

Linear layer from 2048 dimensions to 1,
followed by sigmoid activation.
Trains in 0.5-10 seconds on a GPU

Complex decision tree



T00: Is the galaxy simply smooth and rounded, with no sign of a disk?
A0: Smooth
A1: Features or disk
A2: Star or artifact

T07: How rounded is it?
A0: Completely round
A1: In between
A2: Cigar shaped

T01: Could this be a disk viewed edge-on?
A0: Yes
A1: No

T08: Does the galaxy have a bulge at its centre? If so, what shape?
A0: Rounded
A1: Boxy
A2: No bulge

T02: Is there a sign of a bar feature through the centre of the galaxy?
A0: Bar
A1: No bar

T03: Is there any sign of a spiral arm pattern?
A0: Spiral
A1: No spiral

T09: How tightly wound do the spiral arms appear?
A0: Tight
A1: Medium
A2: Loose

T10: How many spiral arms are there?
A0: 1
A1: 2
A2: 3
A3: 4
A4: More than 4
A5: Can't tell

T04: How prominent is the central bulge, compared with the rest of the galaxy?
A0: No bulge
A1: Just noticeable
A2: Obvious
A3: Dominant

T05: Is there anything odd?
A0: Yes
A1: No

T06: Is the odd feature a ring, or is the galaxy disturbed or irregular?
A0: Ring
A1: Lens or arc
A2: Disturbed
A3: Irregular
A4: Other
A5: Merger
A6: Dust lane

End

1st Tier Question
2nd Tier Question
3rd Tier Question
4th Tier Question

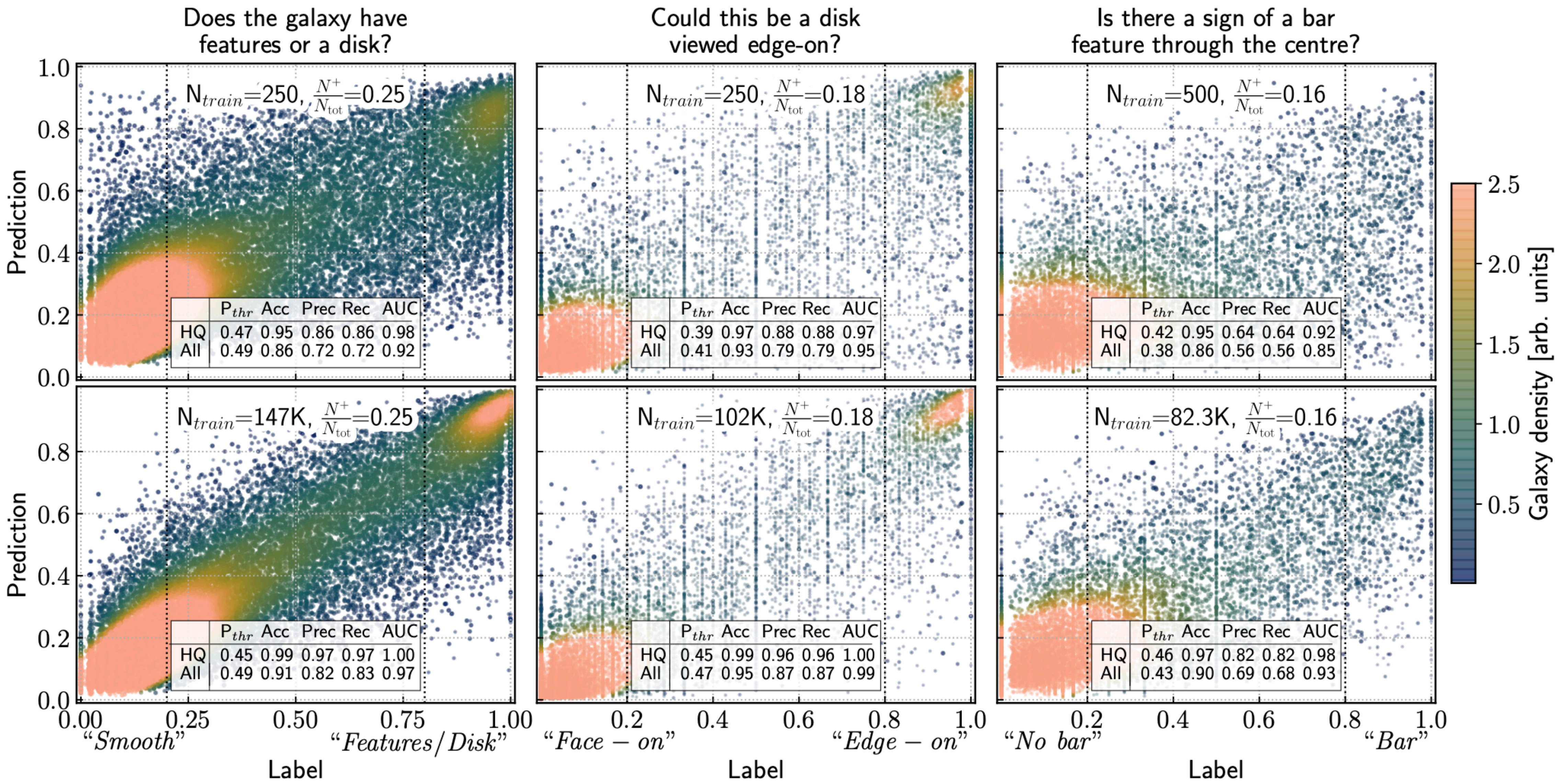
arxiv: 2012.13083

Downstream task #2

Morphological classification

Goal: predict Galaxy Zoo 2 crowd sourced morphological labels directly from representations.

“Soft” labels, with high degree of label uncertainty and mislabelling



Does the galaxy have features or a disk? (N_{train}=250)

	P_{thr}	Acc	Prec	Rec	AUC
HQ	0.47	0.95	0.86	0.86	0.98
All	0.49	0.86	0.72	0.72	0.92

Could this be a disk viewed edge-on? (N_{train}=250)

	P_{thr}	Acc	Prec	Rec	AUC
HQ	0.39	0.97	0.88	0.88	0.97
All	0.41	0.93	0.79	0.79	0.95

Is there a sign of a bar feature through the centre? (N_{train}=500)

	P_{thr}	Acc	Prec	Rec	AUC
HQ	0.42	0.95	0.64	0.64	0.92
All	0.38	0.86	0.56	0.56	0.85

Does the galaxy have features or a disk? (N_{train}=147K)

	P_{thr}	Acc	Prec	Rec	AUC
HQ	0.45	0.99	0.97	0.97	1.00
All	0.49	0.91	0.82	0.83	0.97

Could this be a disk viewed edge-on? (N_{train}=102K)

	P_{thr}	Acc	Prec	Rec	AUC
HQ	0.45	0.99	0.96	0.96	1.00
All	0.47	0.95	0.87	0.87	0.99

Is there a sign of a bar feature through the centre? (N_{train}=82.3K)

	P_{thr}	Acc	Prec	Rec	AUC
HQ	0.46	0.97	0.82	0.82	0.98
All	0.43	0.90	0.69	0.68	0.93

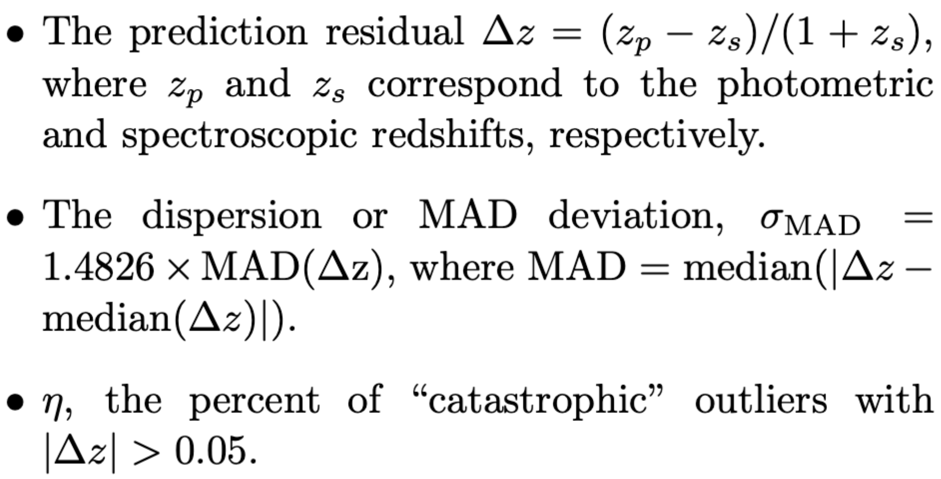
With highly limited number of training samples achieve high classification performance
Using full set of labels achieve state-of-the-art
No need to train separate networks for each classification task

Downstream task #3

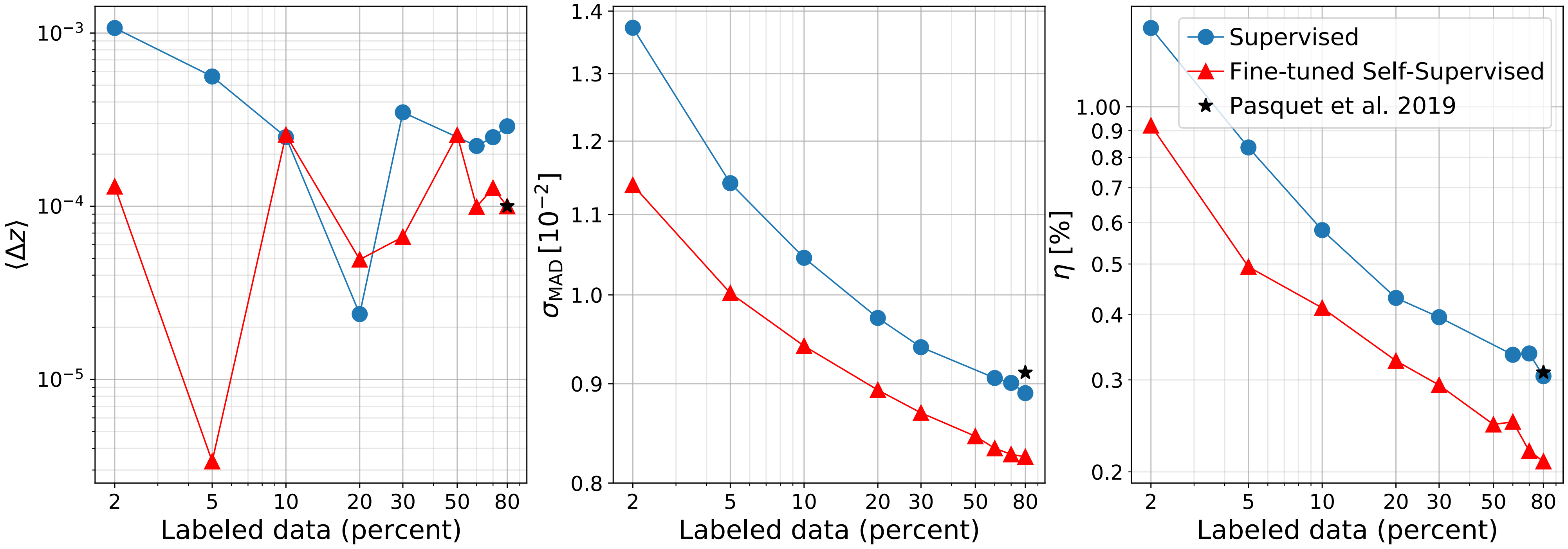
Photometric redshift prediction

Goal: Given an image, predict the redshift of the galaxy. ~500,000 labels from spectroscopic followup

Approach: Linear layer, trained as a classifier over a discrete set of 180 redshift bins spanning 0 < z < 0.4
Fine-tune the encoder during training, using a small learning rate



- The prediction residual $\Delta z = (z_p - z_s)/(1 + z_s)$, where z_p and z_s correspond to the photometric and spectroscopic redshifts, respectively.
- The dispersion or MAD deviation, $\sigma_{\mathrm{MAD}} = 1.4826 \times \mathrm{MAD}(\Delta \mathrm{z})$, where $\mathrm{MAD} = \mathrm{median}(|\Delta z - \mathrm{median}(\Delta z)|)$.
- η, the percent of "catastrophic" outliers with $|\Delta z| > 0.05$.



Self-supervised pre-training results in accuracy equivalent to 2-4x more labelled samples over supervised learning

Summary

Self-supervised representation learning:

- yields notable performance gains over supervised learning for multiple tasks, using the same network
- allows for accurate classifications with highly-limited number of training samples
- representations provide a rich avenue for data discovery
- representations allow for similarity search to pull out similar objects (one-shot anomaly detection)

Future avenues

- Automated anomaly detection, and how this can be done on the representations. Density based models?
- Robustness quantification. Can we use this to determine in-distribution/out-of-distribution for unlabelled samples?

  Using Self-Supervised Learning Can Improve Model Robustness and Uncertainty
- Larger model trained on much more data, and serve to the community, much like the operation of existing state-of-the-art language models

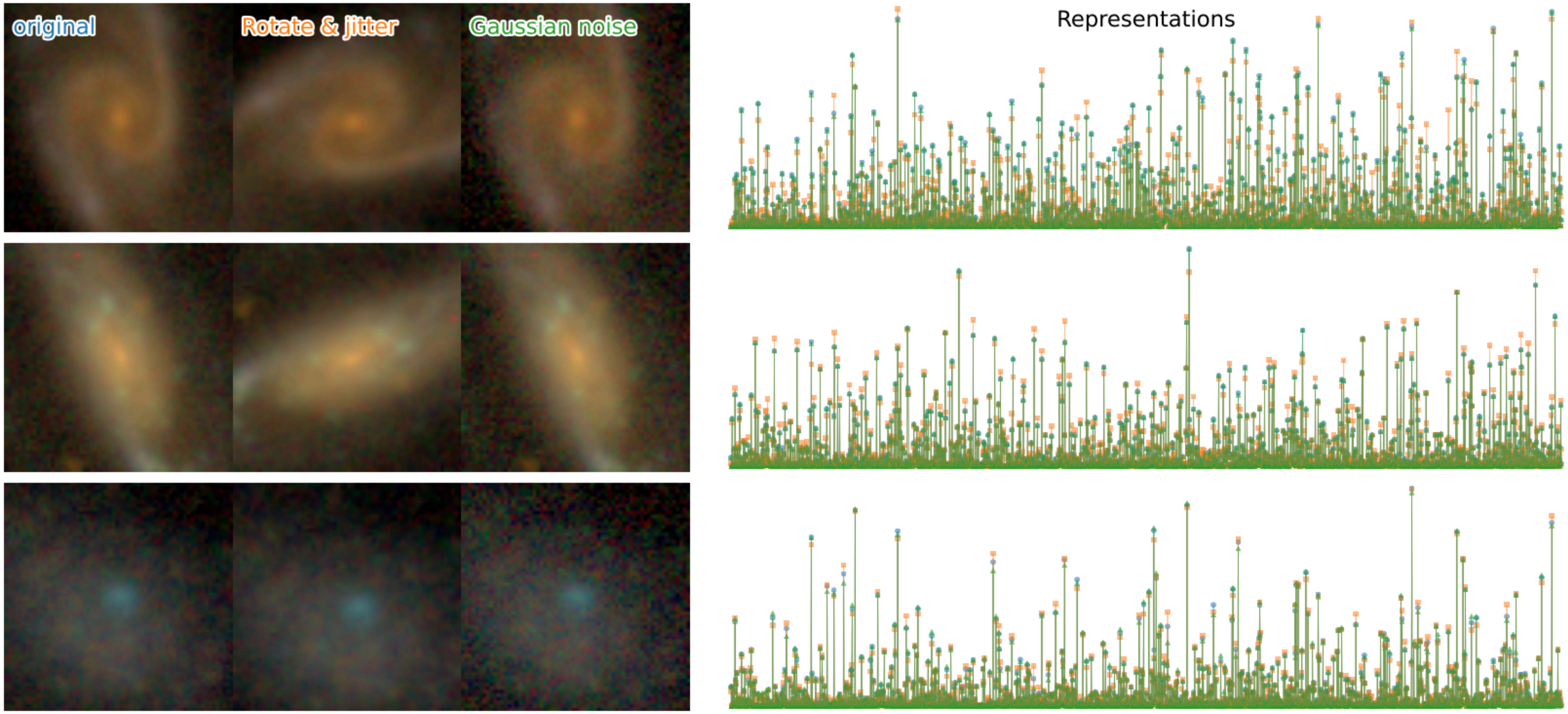
Sketch for self-supervised learning in other fields

1. For given experiment, take all data with or without labels. Can be 1D, 2D, 3D, …
2. Construct data augmentations that reflect changes in the data you want network to be agnostic to. (don't choose ones that effect semantic information of data for desired downstream tasks)
    - Rotations about certain axes
    - Jitter
    - Masking
    - Smoothing
    - Various types of noise
    - Scaling
3. Learn representations through self-supervised contrastive framework
4. Use representations for downstream tasks
    - Data discovery
    - Anomaly detection/similarity search
    - Classification tasks
    - Regression tasks
5. Improve upon what is capable in a supervised framework

examples:

Big Self-Supervised Models Advance Medical Image Classification

Representations after augmentations



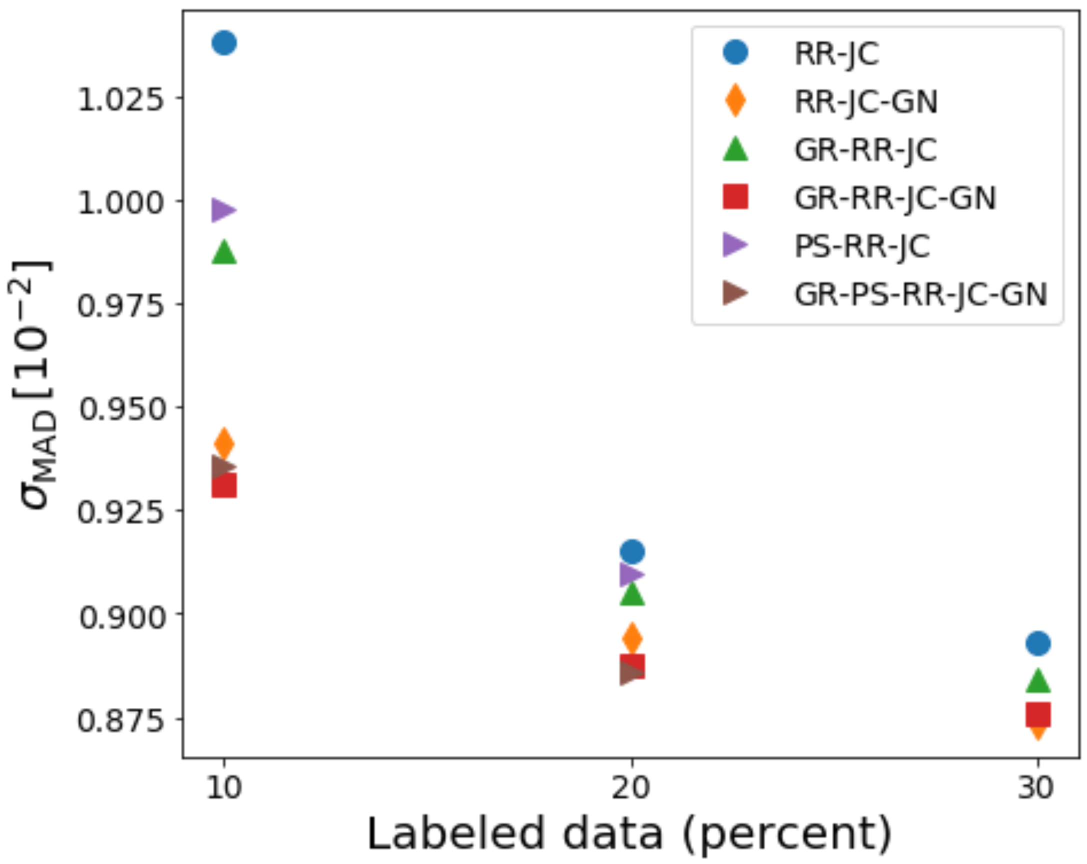
Representations close to invariant under different augmentations

Extra Slides

Ablation study

Which augmentations are most powerful?

Photo-z application

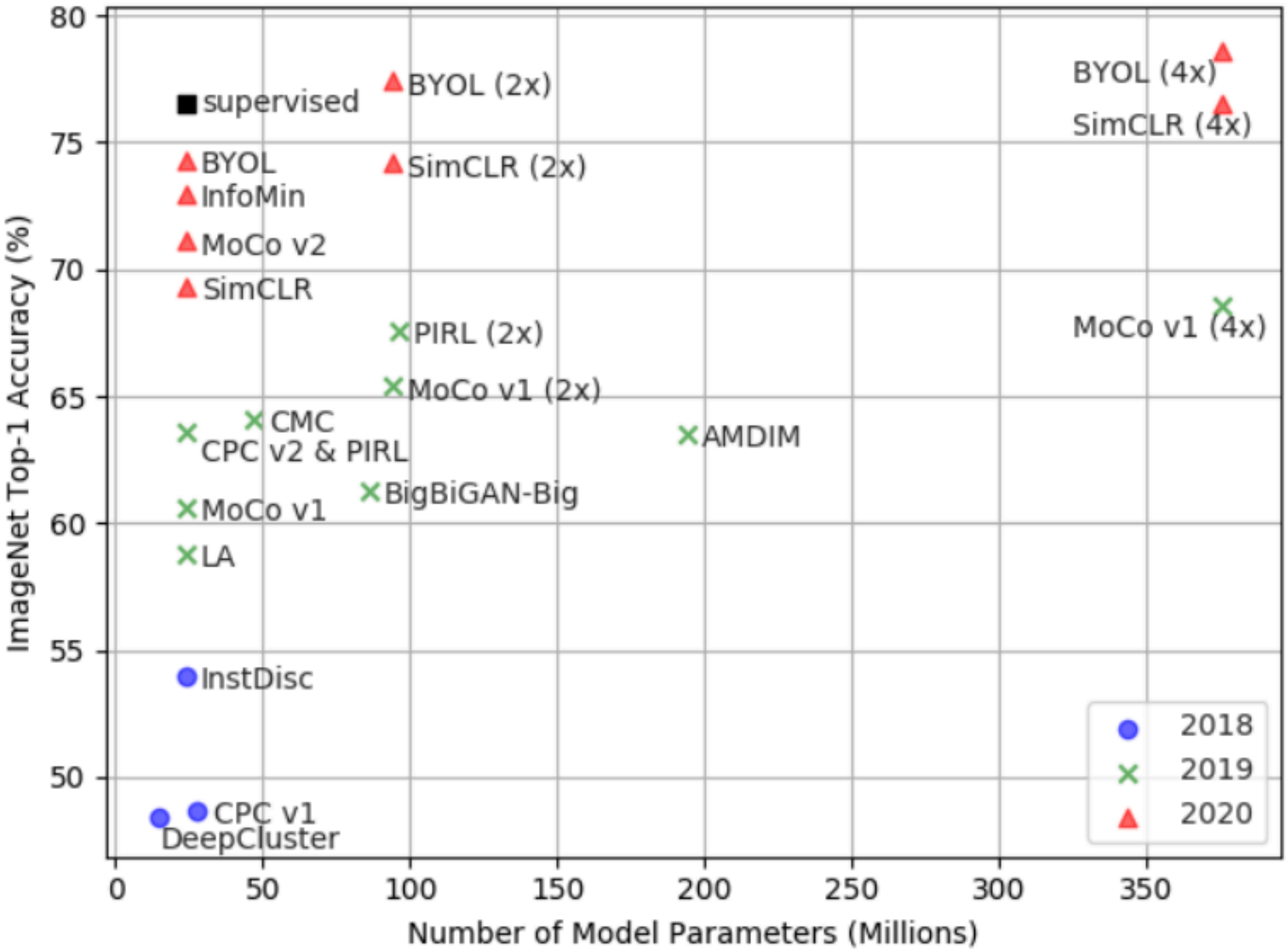


RR: Random rotate
JC: Jitter-crop
GN: Gaussian Noise
GR: Galactic reddening
PS: PSF smoothing

Extra Slides

Self-supervised learning frameworks

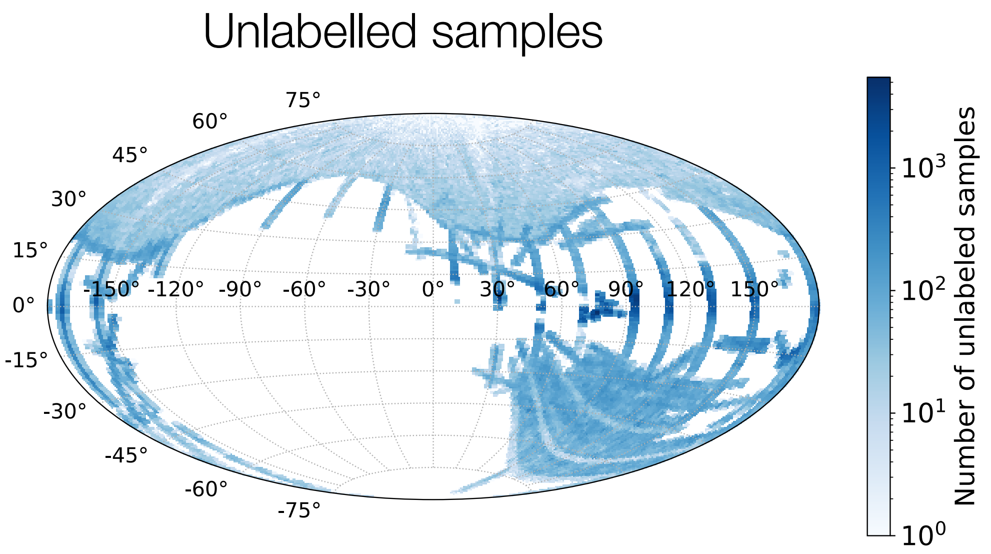
Self-supervised representation learning performance on ImageNet top-1 accuracy in June, 2020, under linear classification protocol.

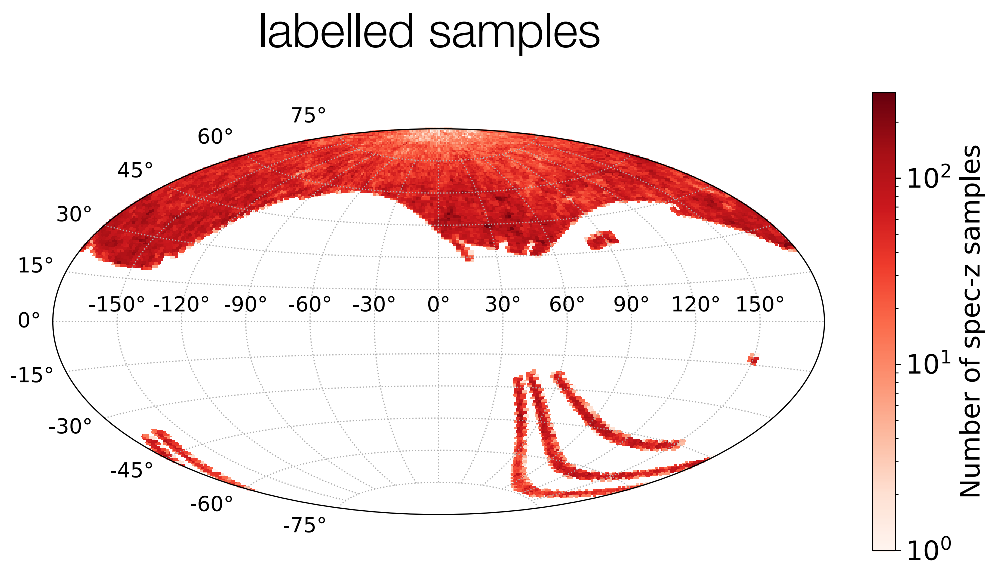


from *Self-supervised Learning: Generative or Contrastive*

Extra Slides

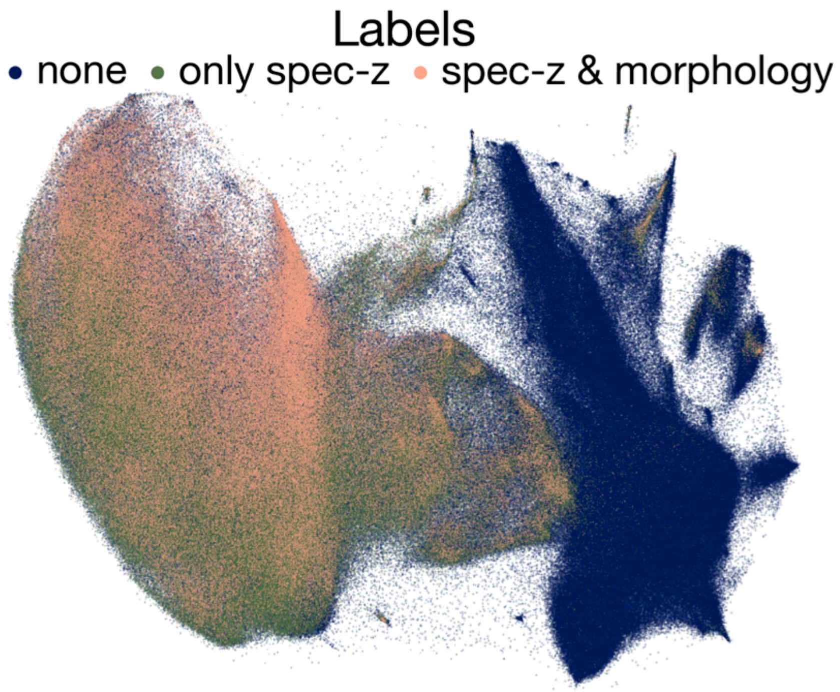
Robustness





A large number of unlabelled samples are from different parts of the sky than those with labels

By design, as (redshift) labels were obtained only for the cleanest images in sky surveys



This is apparent in the learned representation space, i.e. there are unlabelled samples that are not “near” any with labels

(note there are still tonnes of blue dots hidden under green and pink ones)

-> Use for robustness metric of inference on samples that do not have labels:
Unlabelled samples “near" ones with labels = trust inference
Unlabelled samples far from any with labels = don’t trust inference

“Near” can be linear classification, density based, iD/OoD, …