Science and Technology Facilities Council
Scientific Computing

Blueprinting AI for Science at Exascale (BASE-II)

Computing Insight UK 2025

December 5, 2025

Jaehoon Cha

Ai for Science
Rutherford Appleton Laboratory

Outline

1 Surrogate-modelling

2 Representation learning technique

Science and
Technology
Facilities Council
Scientific Computing

Science and Technology Facilities Council

Scientific Computing

Emulating CO Line Radiative Transfer with Machine Learning

Shiqi Su, Frederik De Ceuster, Jaehoon Cha, Mark I Wilkinson, Jeyan Thiyagalingam, Jeremy Yates, Yi-Hang Zhu, Jan Bolte

Science and Technology Facilities Council

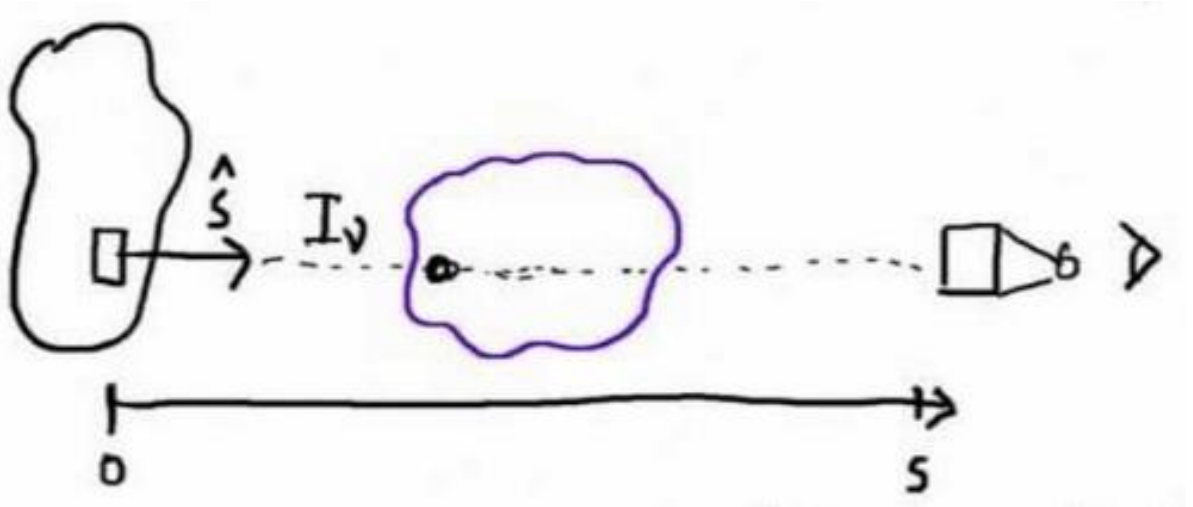
Scientific Computing

Emulating 3D Radiative Transfer Equation

- A linear partial integro-differential equation

$$\hat{n} \cdot \nabla I_\nu(\boldsymbol{x}, \hat{\boldsymbol{n}}) = \eta_\nu(\boldsymbol{x}) - \chi_\nu(\boldsymbol{x}) I_\nu(\boldsymbol{x}, \hat{\boldsymbol{n}}) + \oint d\Omega' \int_0^\infty \Phi_{\nu\nu'}(\boldsymbol{x}, \hat{\boldsymbol{n}}, \hat{\boldsymbol{n}}') I_\nu{}'(\boldsymbol{x}, \hat{\boldsymbol{n}}') d\nu'$$

- ❑ $\mathbf{x}$: spatial variable $(x, y, z) \in \mathbb{R}^3$
- ❑ $\hat{\boldsymbol{n}}$: direction of ray
- ❑ υ : frequency, $\frac{speed\ of\ wave}{wavelength} = \frac{c}{\lambda}$
- ❑ $I_\nu(x, \hat{n})$, radiative intensity
- ❑ $\eta_\nu(x)$, emission
- ❑ $\chi_\nu(x) I_\nu(\boldsymbol{x}, \hat{\boldsymbol{n}})$, absorption
- ❑ $\Phi(\cdot) I(\cdot)$, scattering

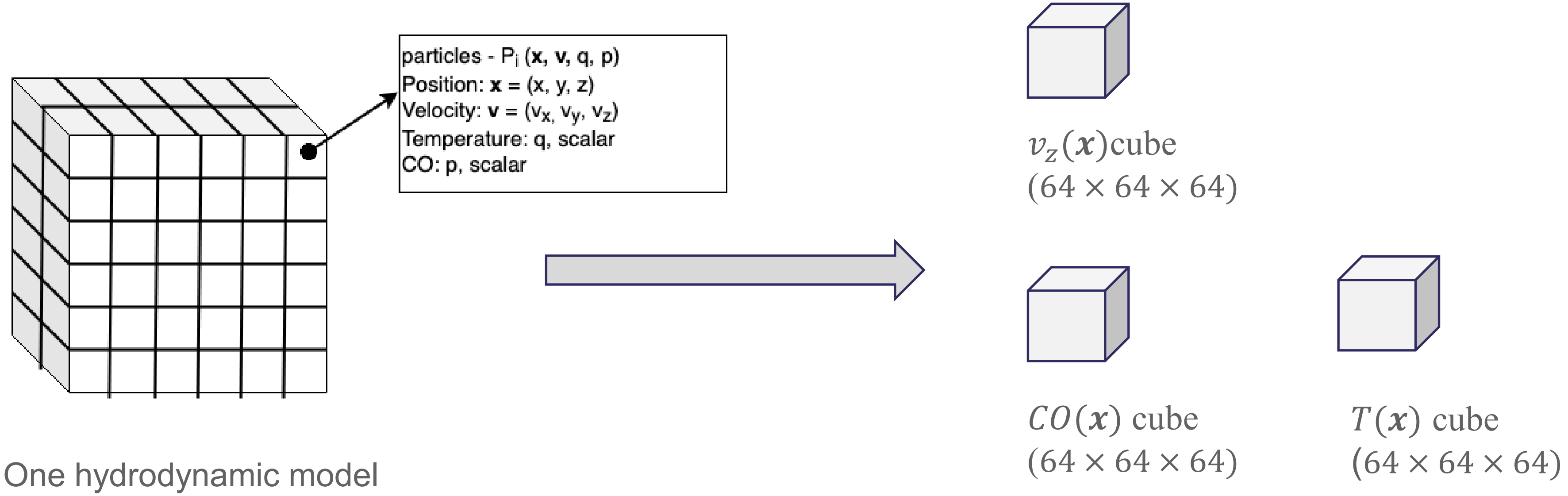


Science and Technology Facilities Council
Scientific Computing

Shiqi Su., et al. (2025). Emulating CO line radiative transfer with deep learning. RAS Techniques and Instruments

Surrogate Modelling

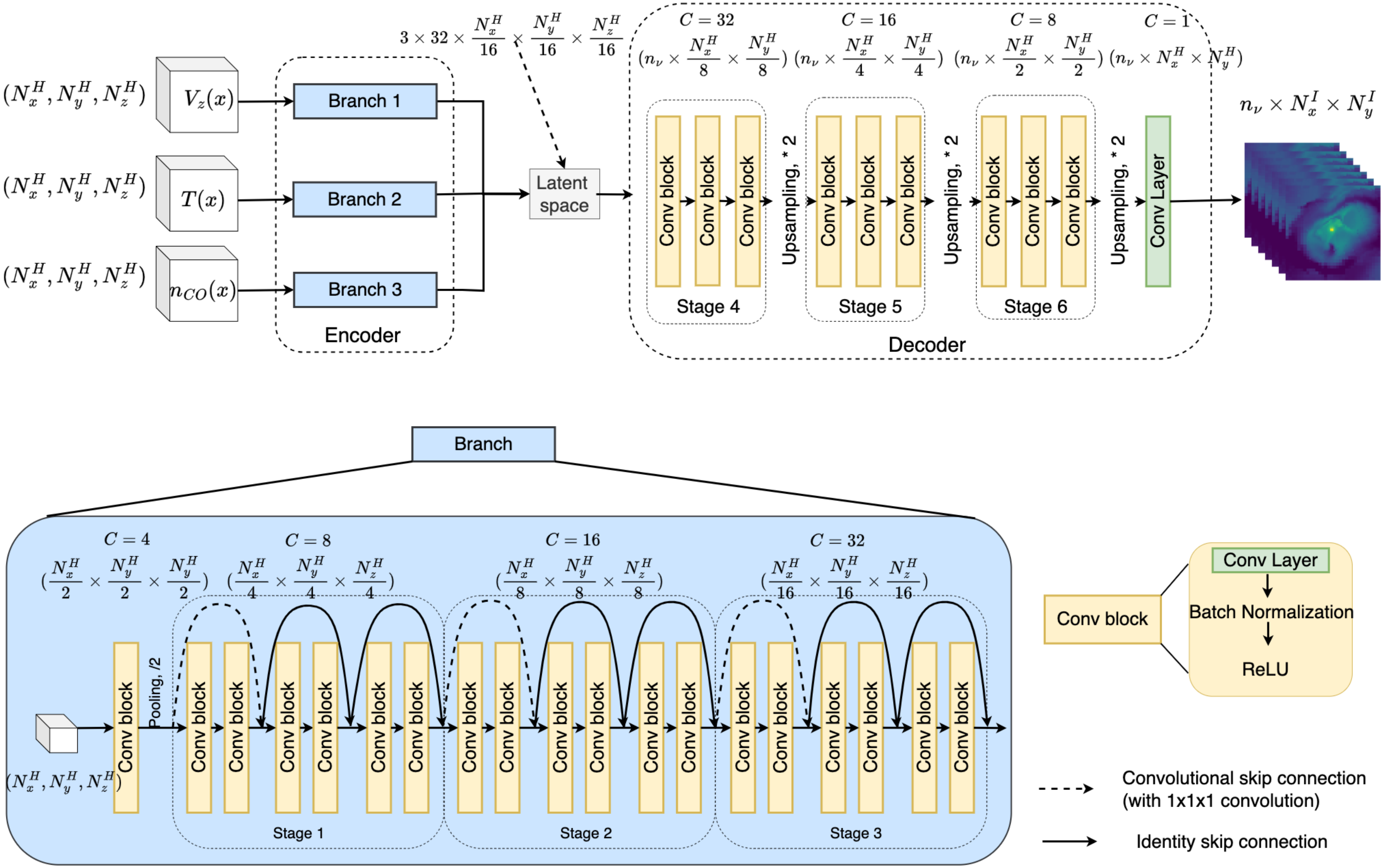
- The input is a hydrodynamic model, a mathematical framework describing the motion and behaviour of fluids, with a total size of around 7 TB.
- Under the assumption of local thermodynamic equilibrium (LTE), the spectral line model is fully determined by a few parameters.
- The model includes velocity along the z-axis, Carbon monoxide (CO) density, and temperature.



One hydrodynamic model

Shiqi Su., et al. (2025). Emulating CO line radiative transfer with deep learning. RAS Techniques and Instruments

COEmuNet



Science and Technology Facilities Council

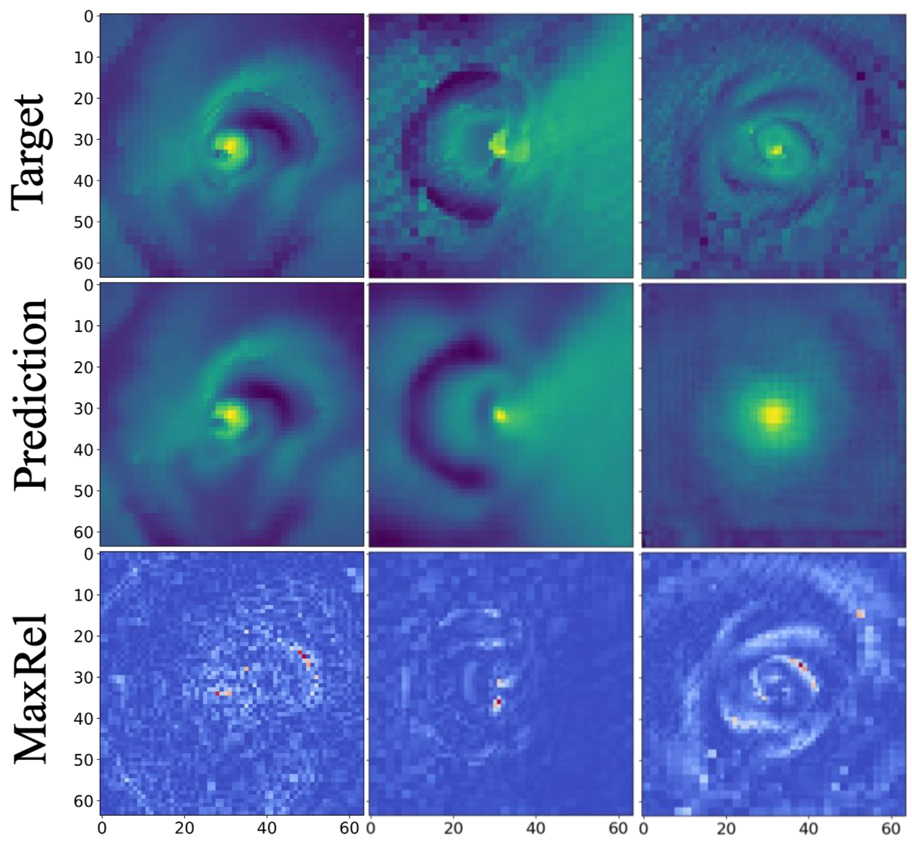
Scientific Computing

Shiqi Su., et al. (2025). Emulating CO line radiative transfer with deep learning. RAS Techniques and Instruments

Results

- We trained the model to output the middle seven frequencies and randomly rotated the input 100 times.
- Total data size is about 7TB.
- The model contains 143,303,809 parameters, making a multi-GPU approach using data-distributed parallelism (DDP) necessary.
- We use four A100-40GB GPUs, and it takes one hour per epoch.

Inference time (sec)	
Numerical solver	Surrogate model
2.67601	0.01181



Shiqi Su., et al. (2025). Emulating CO line radiative transfer with deep learning. RAS Techniques and Instruments

Science and Technology Facilities Council

Scientific Computing

Discovering Interpretable Representations in Scientific Data

Jaehoon Cha, Jinhae Park, Samuel Pinilla, Kyle L Morris, Christopher S Allen, Mark I Wilkinson, Jeyan Thiyagalingam

Science and Technology Facilities Council

Scientific Computing

Why Learning Interpretable Representations Matters

Scientific data may look complex, but a few key factors often explain most of it.

- A 1D spectrum may have many wavelengths, but only a few peaks matter.
- A 2D image can be understood through shape, position, or orientation.

Interpretable representations help us

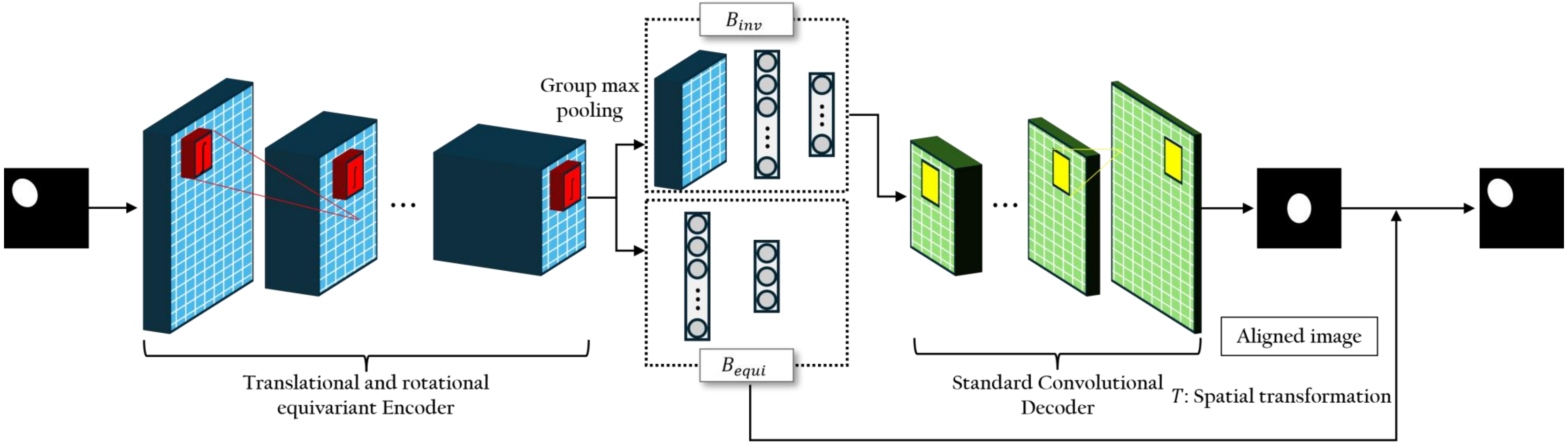
- Understand how our data varies.
- Cluster data in a meaningful way.
- Support discovery, and data collection.

Disentangled Representation

- A disentangled representation is a representation that separates the underlying factors of variation so each can be controlled independently.

Translational and rotational equivariant Encoder

- The encoder has stacks of rotated kernels to learn different orientations of objects.
- It enables learning both centroids and orientations of objects.
- However, this makes the encoder bigger.

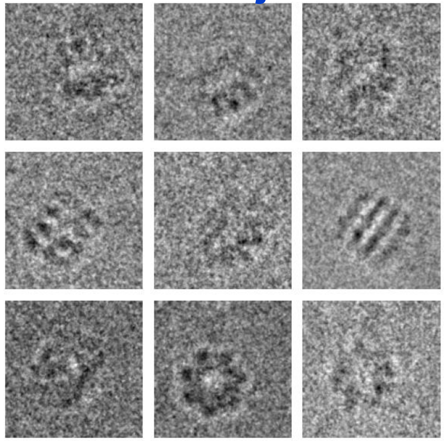


Cha, J., et al. (2025). Discovering fully semantic representations via centroid- and orientation-aware feature learning. Nature Machine Intelligence

CryoEM single particle analysis

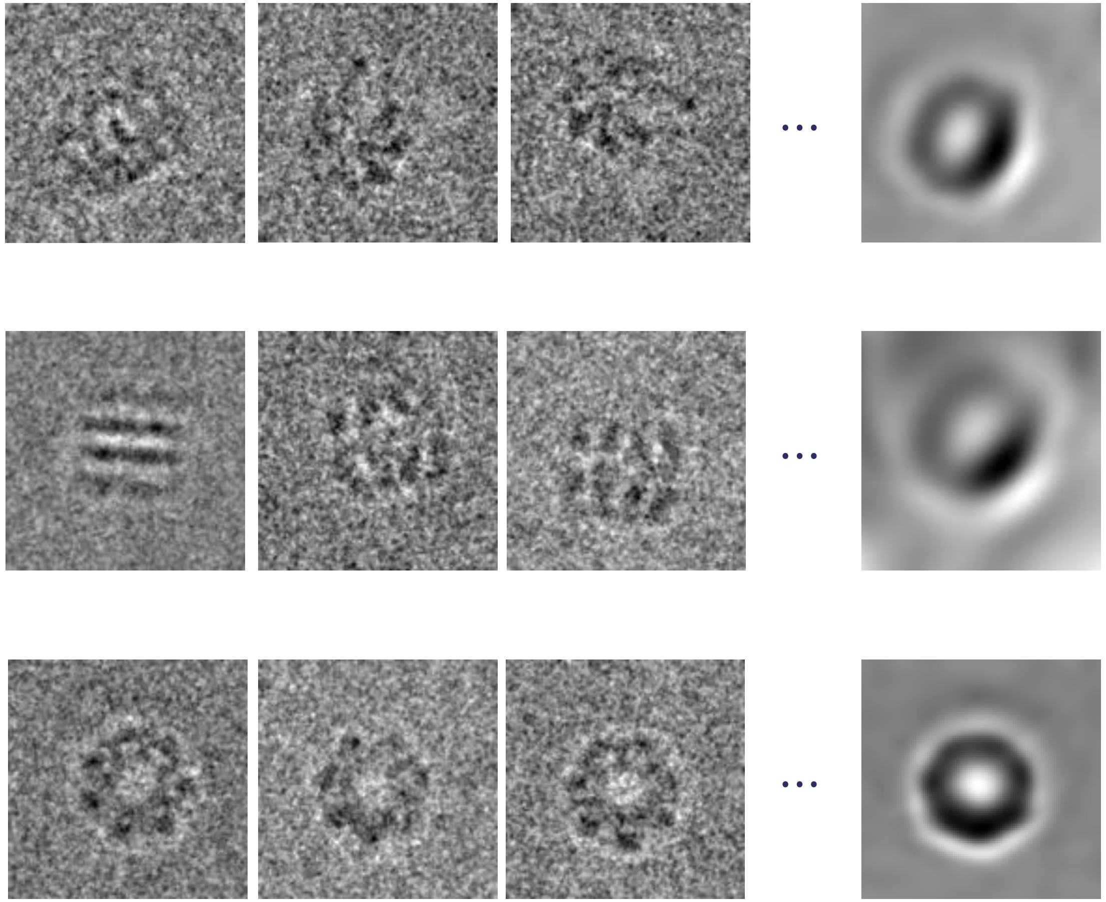
- The cryoEM single particle analysis method generates 2D projection images with low signal-to-noise ratios.
- Grouping 2D projections of the molecule captured from similar viewing angles (or object poses) and aligning them using in-plane rotations and translations improves the signal-to-noise ratio, enabling more effective 2D image analysis.

Examples of 2D Protein Projections



2D projections

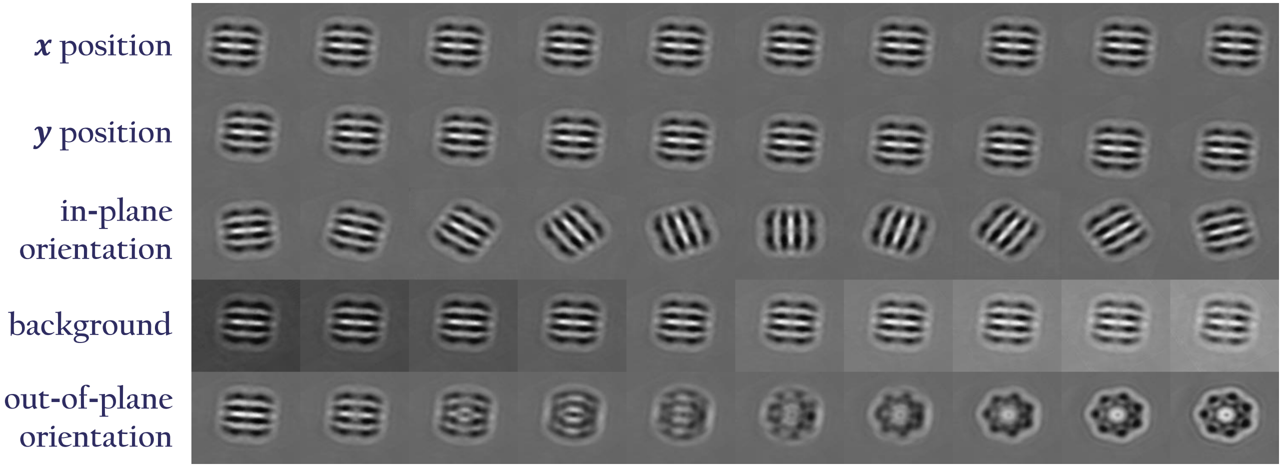
Average of projections



Cha, J., et al. (2025). Discovering fully semantic representations via centroid- and orientation-aware feature learning. Nature Machine Intelligence

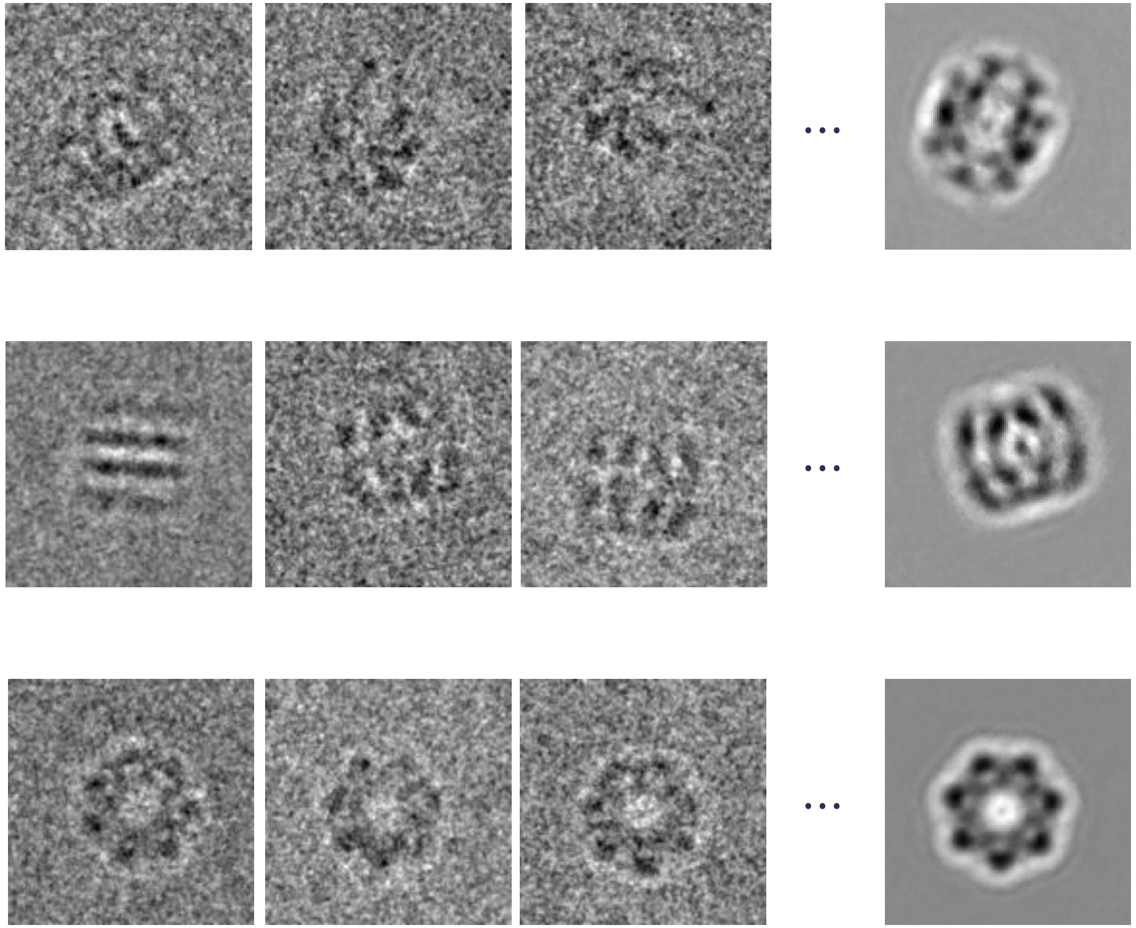
Extracting 2D particles from Cryo-EM

Latent traversals



2D projections

Average of projections after alignment



Cha, J., et al. (2025). Discovering fully semantic representations via centroid- and orientation-aware feature learning. Nature Machine Intelligence

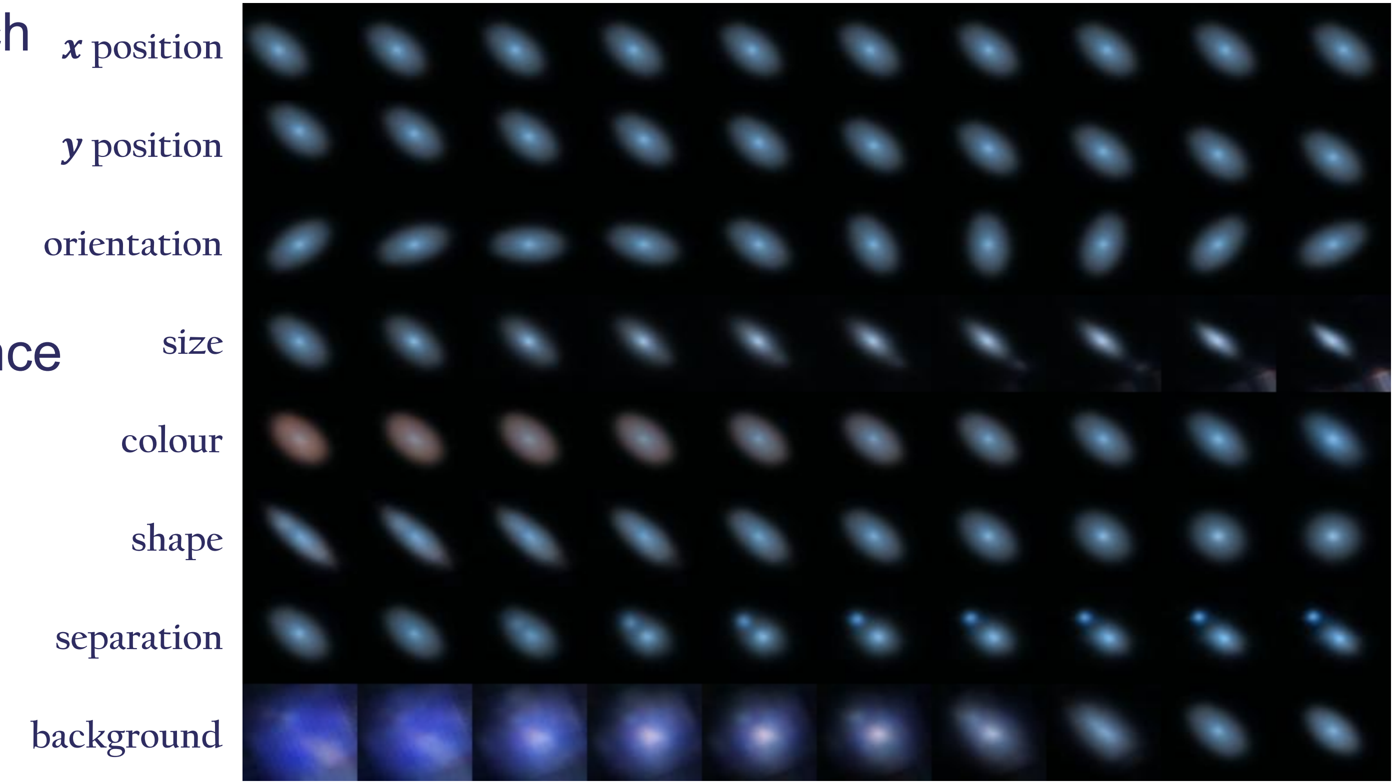
Galaxy Images

- The pose of galaxy images does not affect their intrinsic properties, highlighting the importance of an unsupervised approach to learn semantic representations of galaxies while capturing their pose information.

- In this study, we evaluate the performance of a disentanglement model using the Galaxy-Zoo dataset.

- The results demonstrate that the model effectively identifies key features of the dataset, including pose, size, colours, shape, separation, and background.

Galaxy-Zoo from astronomy with DiRAC (Mark Wilkinson)



Cha, J., et al. (2025). Discovering fully semantic representations via centroid- and orientation-aware feature learning. Nature Machine Intelligence

Science and Technology Facilities Council

Scientific Computing

Thank you

scd.stfc.ac.uk

@SciComp_STFC

Science and Technology Facilities Council

Ada Lovelace Centre