What a Machine see?
— Exploring Galaxy Morphology with Unsupervised Machine Learning

Royal Astronomical Society

Ting-Yun Cheng, Marc Huertas-Company, Christopher Conselice, Alfonso Aragón-Salamanca, et al.

I. Quick Introduction

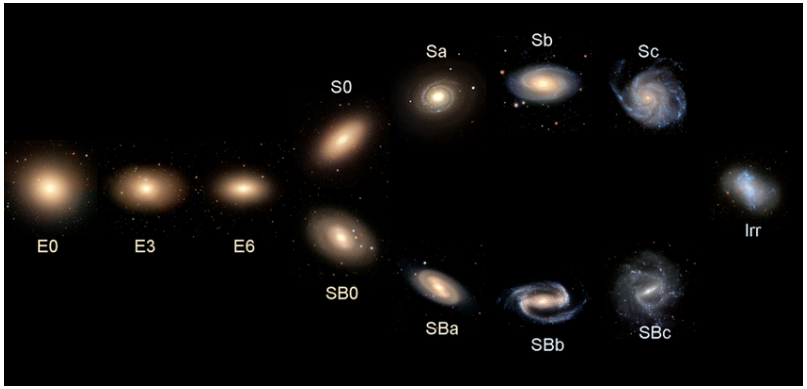


Fig.1 Hubble sequence classification scheme
(credit: Department of Physics and Astronomy, University of Iowa)

Galaxy morphology are strongly connected with the stellar properties and the formation history of galaxies. For a century, galaxy morphologies are categorised based on their visual appearance. However, this kind of visual classification systems such as Hubble types (Fig.1): ellipticals (E), lenticulars (S0), spirals (S), and irregulars (Irr) are intrinsically biased due to the subjective definition made by visual assessment.

In this work, we make machine "sensibly see" images by applying an unsupervised machine learning technique, which includes:

(1) **feature learning phase** by Vector-Quantised Variational AutoEncoder (VQ-VAE);

(2) **clustering phase** by Hierarchical Clustering (HC).

We test this unsupervised machine on the Sloan Digital Sky Survey (SDSS) imaging data to approach an objective morphological classification scheme without human involvement.

II. Methodology: Build a sensible machine to read galaxy morphology

Fig.2 shows the overview of the pipeline used in work. An autoencoder learns the distribution (representative features) of the input images by the encoder, and based on the learnt distribution (extracted features) it then reproduces the input images by the decoder. Each image is then represented by a set of extracted features. Hierarchical clustering then gradually merges two nearest datapoints (with similar features) into groups in the feature space. The vector quantisation process accelerates the feature learning phase from 4-5 days to a few hours on 100k images.

To make machine sensible, in terms of close to human opinions, we propose three strategies:

(1) to consider clustering performance simultaneously while learning representative features from images in the VQ-VAE;

(2) to use different distance thresholds in the HC depending on the complexity of galaxy images instead of single distance;

(3) to use the feature of galaxy orientation in the dataset into a distance cut to determine the optimal number of groups obtained from the HC.

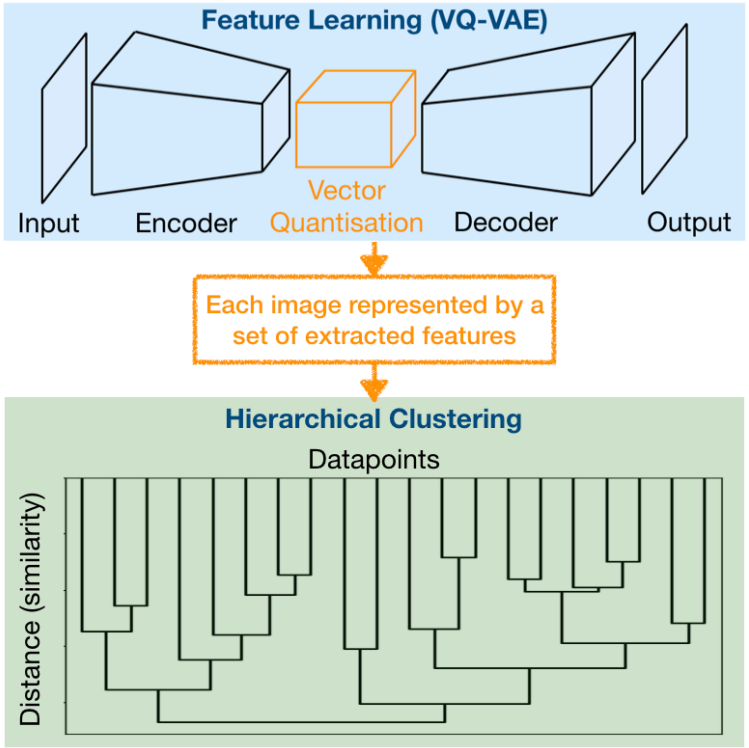


Fig.2 pipeline used in this work

II. Results: machine-defined classes

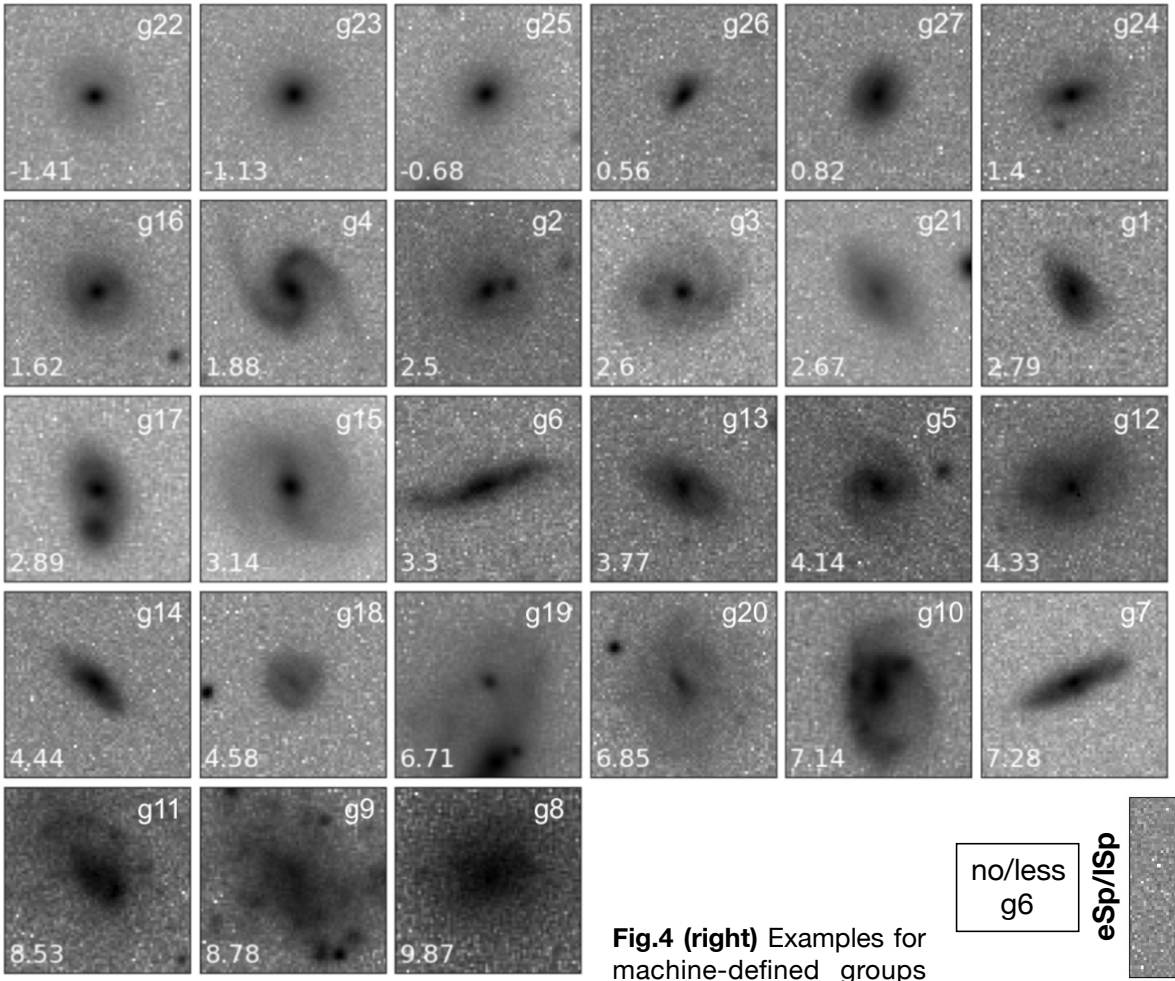
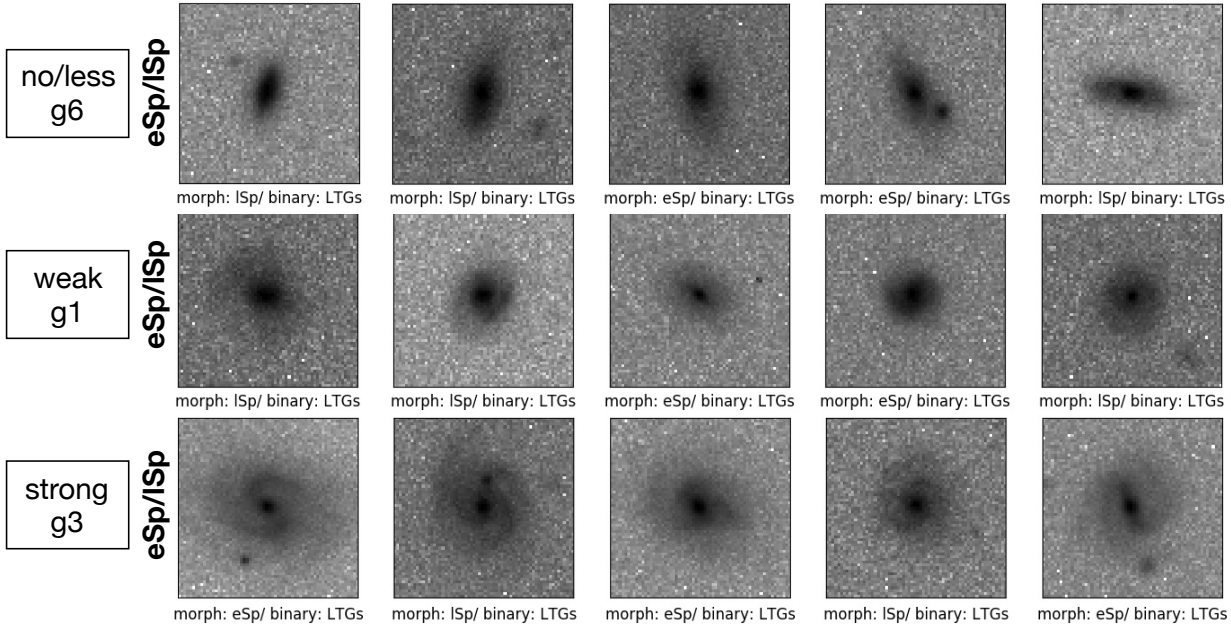


Fig.3 (top) Examples of a randomly picked galaxy from each machine-defined group in the order of the average values of T-Types[1].

Fig.4 (right) Examples for machine-defined groups with different fractions of barred galaxies



The methodology applied in this work provides **27 classes** for galaxy morphology (Examples in Fig.3). These machine-defined classes are separated based on galaxy shape and structure.

Barred galaxies are distinctive and important visual classes in galaxy morphology (bottom inFig.1, e.g., SBa) as well as in galaxy evolution and formation. This structural feature can also be distinguished by our unsupervised machine. In the 27 machine-defined classes, some of them are significantly dominated by barred galaxies (examples: bottom in Fig.4), or not (examples: top in Fig.4).

Cheat sheet: abbreviation used in this work				
E	**S0**	**eSp**	**lSp**	**Irr**
Ellipticals	Lenticulars	early spirals	late spirals	Irregulars

*1: T-Types is a type of galaxy morphological classification system using continuous values. The definition used in this work: -3 for ellipticals (E), -2 for lenticular at early stage (S0-), -1 for lenticular at intermediate to late stages (S0), 0 for the transition from lenticular to early spirals (S0/a). Positive values represent in general spiral galaxies from early to late spiral, except for 10 which represents irregular galaxies (Irr).

II. Results: machine classification versus human visual classification

To further analyse the visual features recognised by our unsupervised machine, we associated the 27 machine-defined classes with the visual morphology types such as Hubble types (Fig.5). No clean cluster is dominated by only E and eSp in Fig.5 due to a great similarity shared between E, S0, and eSp in structure. Additionally, most groups have a mixture of different Hubble types within them which indicates galaxies with similar features in appearance can be visually classifying into a variety of morphology types (example in Fig.6). This result reveals an intrinsic vagueness of the visual classification systems such that they are not always accurately defined.

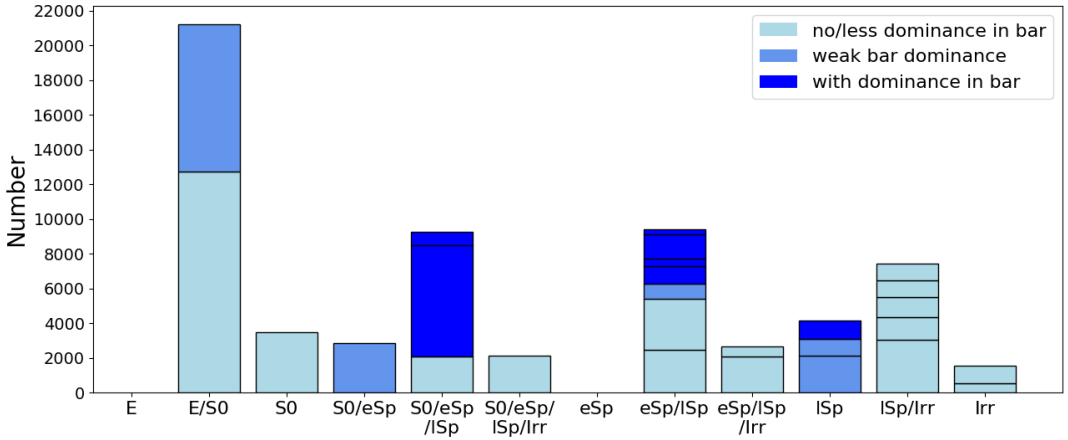


Fig.5 The accumulated distribution of the machine-defined groups compared with Hubble visual morphological types. The x-axis shows one or a mix of visual types which dominates the groups. All 27 clusters are plotted here, and each coloured bar represents one cluster. The different colours of the bars show different dominance levels of barred galaxies in the cluster, such that from deep to light blue represent more barred galaxies to no/fewer barred galaxies in the cluster.

Cheat sheet: abbreviation used in this work				
E	**S0**	**eSp**	**lSp**	**Irr**
Ellipticals	Lenticulars	early spirals	late spirals	Irregulars

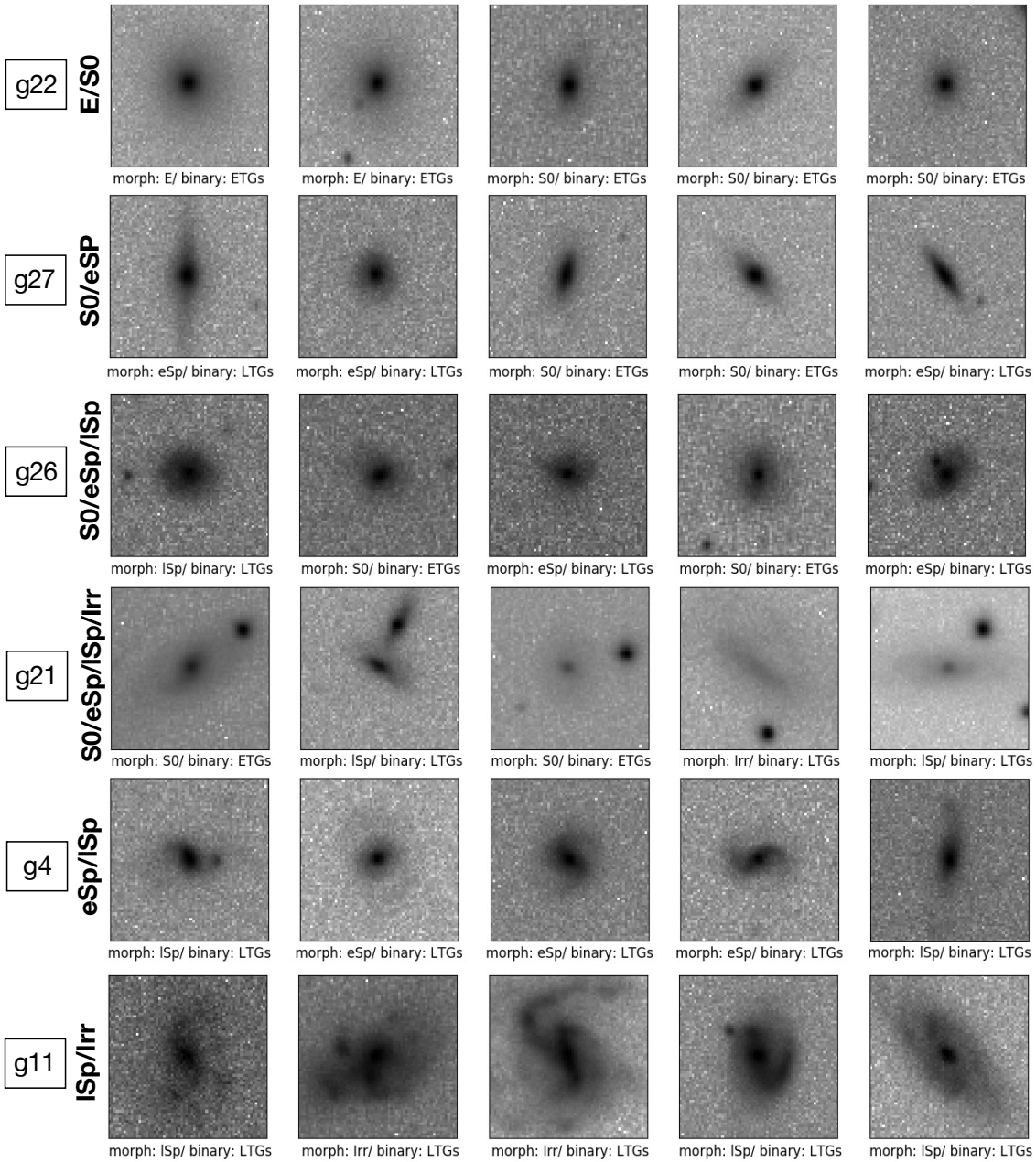


Fig.6 Examples of galaxies categorised with similar visual and structural features by our machine but shown to have a mix of many visual morphology types.

III. Results: machine classification versus physical properties of galaxies

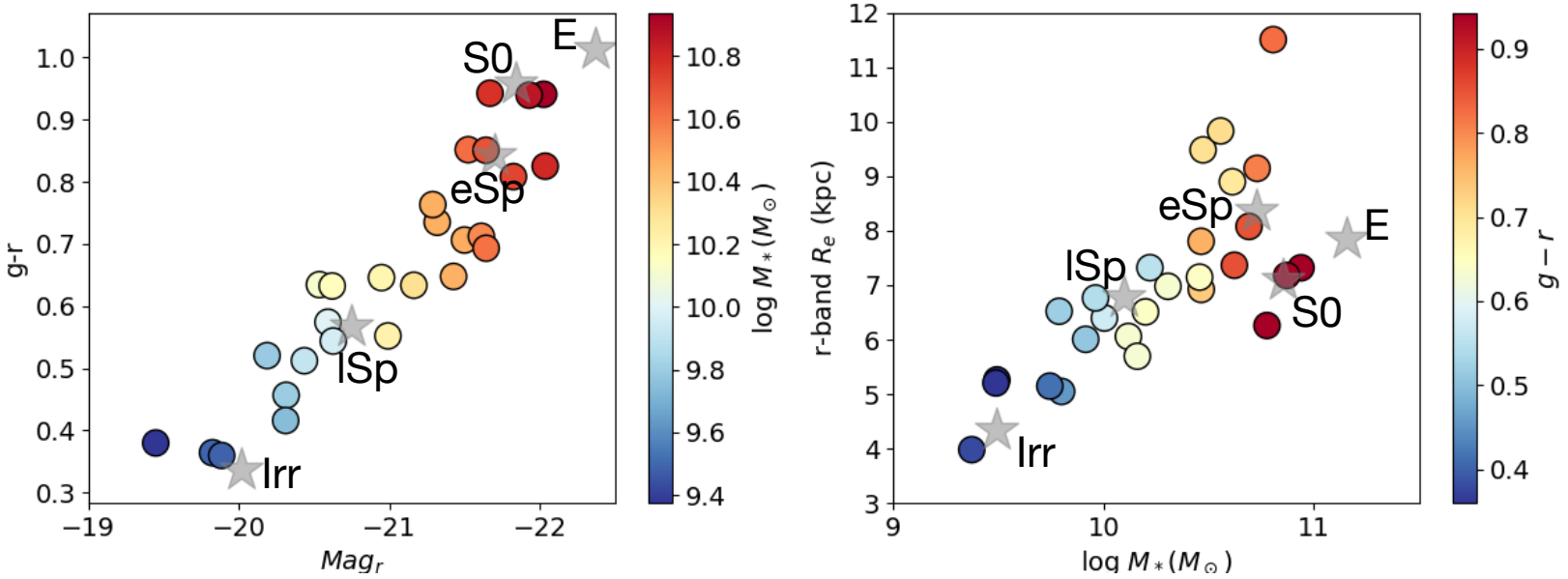


Fig.7 *Left:* the average values of color-magnitude diagram for each classification group, where one circle is one group coloured by stellar mass of galaxies, *r*-band absolute magnitude (Mag_r) in *x*-axis and colour (g-r) in *y*-axis.

Right: the average values of mass-size relation of the given clusters, where the x-axis and y-axis is the stellar mass (M_*) and galaxy physical sizes (R_e, kpc), and circles are coloured by galaxy colour (g-r).

Each star in both plots shows the average value of the data with a certain visual morphology type (written in black) for comparison.

On the mass-size diagram (right in Fig.7) that the five orange clusters above the eSp star label are dominated by barred galaxies, in particular, the top cluster with the largest average size has ~80% barred galaxies in the cluster. Galaxies in this cluster have larger sizes, larger stellar masses, and are redder in colour than other clusters with a mix of typical spiral galaxies.

Each galaxy cluster as defined by the machine has distinctive physical properties in galaxy colour, absolute magnitude, stellar mass, and physical size (Fig.7). This indicates that the machine-defined morphological classes show a strong connection with galaxy evolution. Additionally, our machine classes show a clear transition on the colour-magnitude diagram and mass-size relation between galaxy morphology and galaxy properties. They as well fill in the gap on the diagrams along with the Hubble types (stars in Fig.7). This indicates that the machine classification scheme can complete the missing morphologies in the visual classification systems without involving human potential bias.

IV. Summary

In this poster, we present a machine-defined morphological classification scheme for galaxy. With the strategies proposed in this work, our machine categorises galaxies into 27 classification clusters based on galaxy shape and structure.

The machine-defined classes show a more “accurate” separation in structure and visual features than visual classification systems such as Hubble types. We then reveal an intrinsic uncertainty existed in visual classification scheme in precisely classifying galaxies. Additioanlly, the machine classes show a clear transition between galaxy morphology and galaxy properties. This indicates a strong connection of our machine-defined morphology with the galaxy evolution. With a novel classification scheme proposed by machine learning, we can re-approach studies of galaxy evolution and formation in a different perspective.