Electronic Supplementary Material (ESI) for Digital Discovery.

Electronic Supplementary Information:

Semi-supervised learning of images with strong rotational disorder: assembling nanoparticle libraries

Maxim A. Ziatdinov[1,a] Muammer Yusuf Yaman,[2] Yongtao Liu,[3,]

David Ginger,[2,4] and Sergei V. Kalinin[1,4]

[1] Physical Sciences Division, Pacific Northwest National Laboratory, Richland, WA, 99354

[2] Department of Chemistry, University of Washington, Seattle, WA, 98195

[3] Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Oak Ridge, TN 37831

[4] Department of Materials Science and Engineering, University of Tennessee, Knoxville, TN 37996

The proliferation of optical, electron, and scanning probe microscopies gives rise to large volumes of imaging data of objects as diversified as cells, bacteria, pollen, to nanoparticles and atoms and molecules. In most cases, the experimental data streams contain the images having arbitrary rotations and translations within the image. At the same time, for many cases the small amounts of labeled data can be available, e.g. in the form of prior published results, image collections and catalogs, or even theoretical models. Here we develop the semi-supervised rotationally-invariant variational autoencoder (ss-rVAE) that allows to generalize from a small subset of labeled data with the weak orientational disorder to a large unlabelled data set with a much stronger orientational disorder. The performance of the ss-rVAE is illustrated using the synthetic data sets with the known factors of variation. We further demonstrate the application for the experimental data sets of the nanoparticles.

The jupyter notebooks were run on Google Colab. Nvidia Tesla T4 Processor, with 51 Gb RAM, was used for model training/testing. During the training process, we chose that *training batch size* is "16", the learning rate is "1*e*-4", training epoch is 1000, and task mode is "classification".

[a] maxim.ziatdinov@pnnl.gov

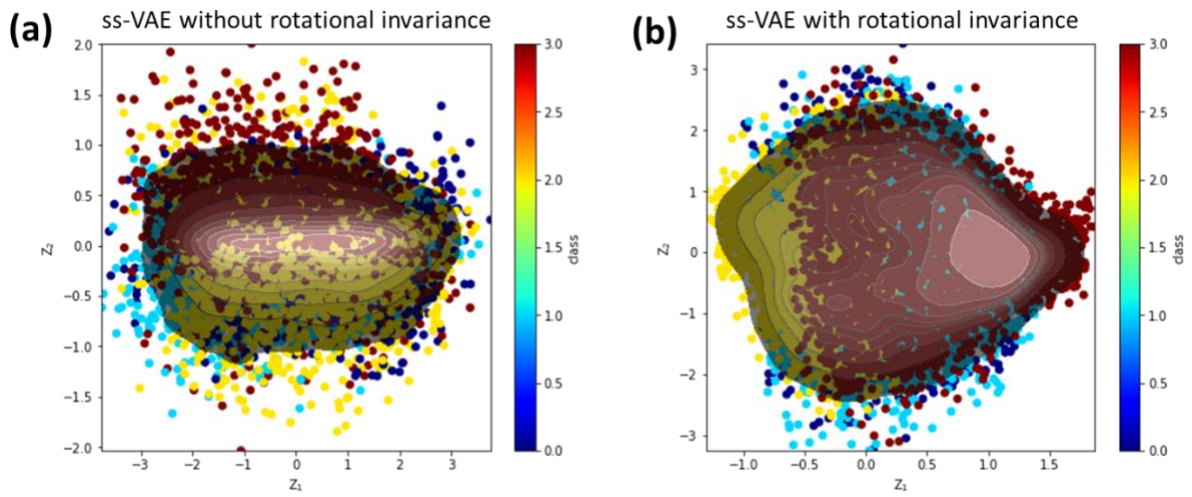

Fig. S1. Latent variables distribution with color corresponding to the class variable. (a) latent variables distribution of ss-VAE without rotational invariance. **(b)** latent variables distribution of ss-VAE with rotational invariance.

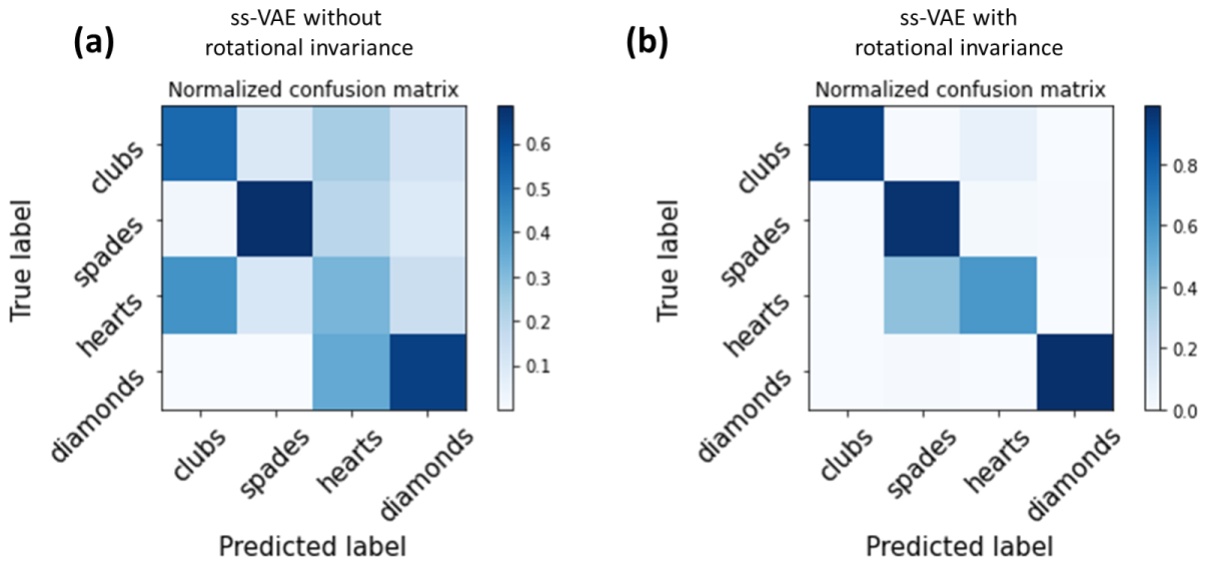

Figure S2. Confusion matrices of ss-VAE analyses on validation data. (a) confusion matrices of ss-VAE without rotational invariance. **(b)** confusion matrices of ss-VAE with rotational invariance.

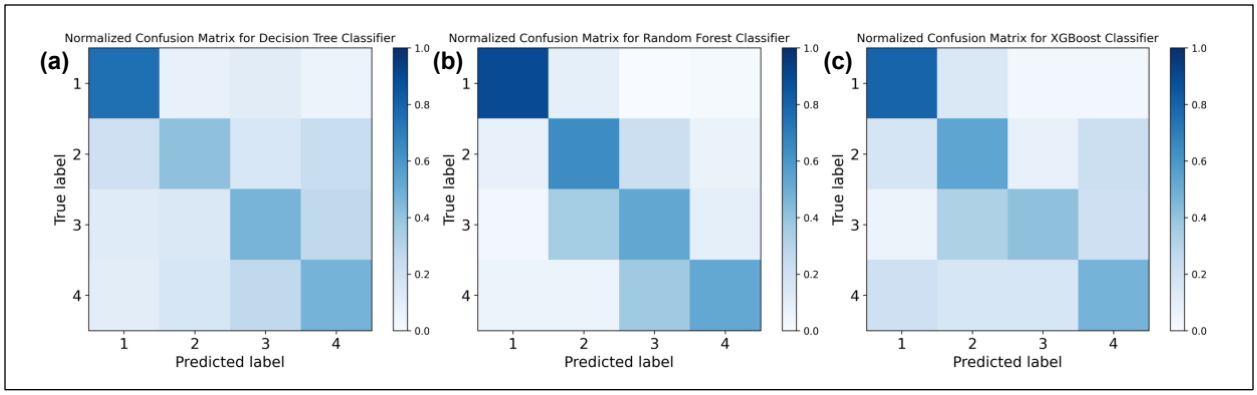

Figure S3. Confusion matrices of existing machine learning on the nanoparticle dataset. (a) Normalized confusion matrices of (a) decision tree classifier, **(b)** random forest classifier and **(c)** XGBoost classifier. The color bars show the model accuracy in each class.

The highest accuracy of decision tree classifier, random forest classifier and XGboost classifier for class one is 0.75, 0.89 and 0.8, respectively. The accuracy for other classes is around 0.4-0.5.

Supplementary Video 1: Training process of ss-VAE without rotational invariance:
https://drive.google.com/file/d/1H2fcOGl2ctHP-Z8KYMmoJ3zRXyKNsTzD/view?usp=sharing

Supplementary Video 2: Training process of ss-VAE with rotational invariance:
https://drive.google.com/file/d/16Qdisn7I8uwftJEkRs-LJIOLl8Fs9Nqo/view?usp=sharing

Supplementary Video 3: The evolution of ss-rVAE latent space with increasing the ratio of supervised data:
https://drive.google.com/file/d/14_5FClVtsGjJypNWLYDq3O3uFhKw996e/view?usp=sharing