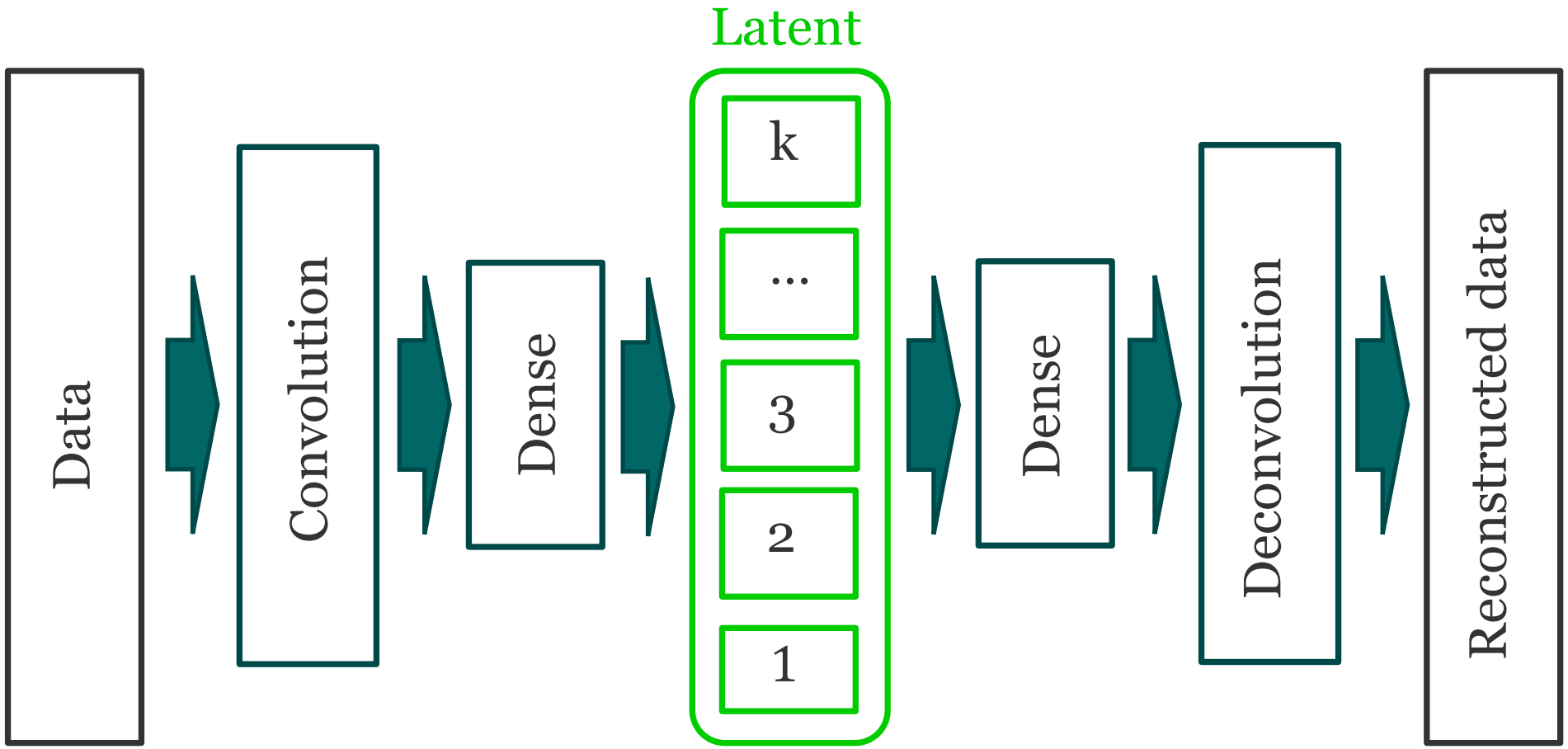
Lecture 32: Explainable ML

Instructor: Sergei V. Kalinin

Autoencoders



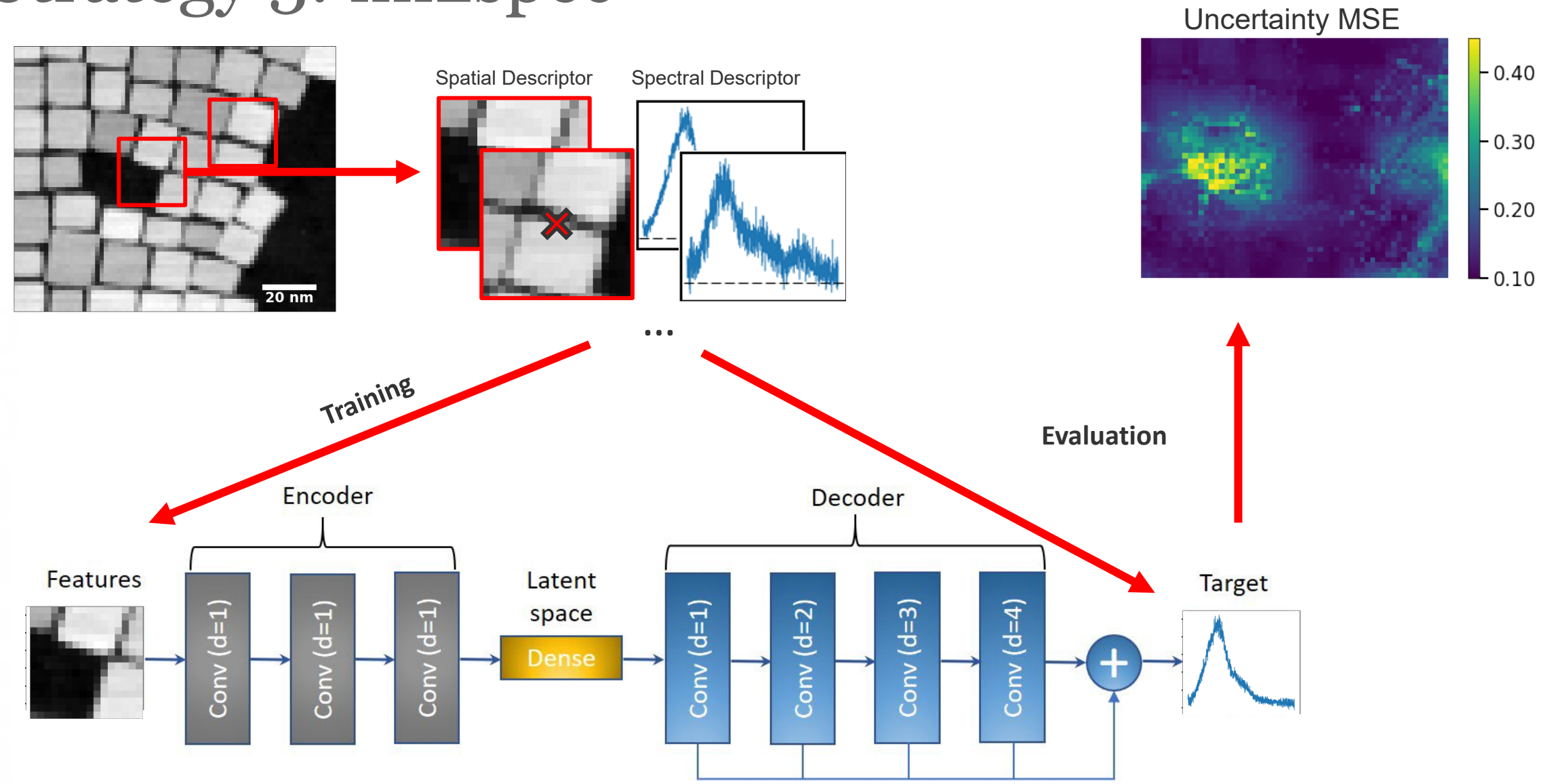
Loss: (some form of) reconstruction loss

Do we have to encode and decode the same type of objects?

Examples of structure property relationships

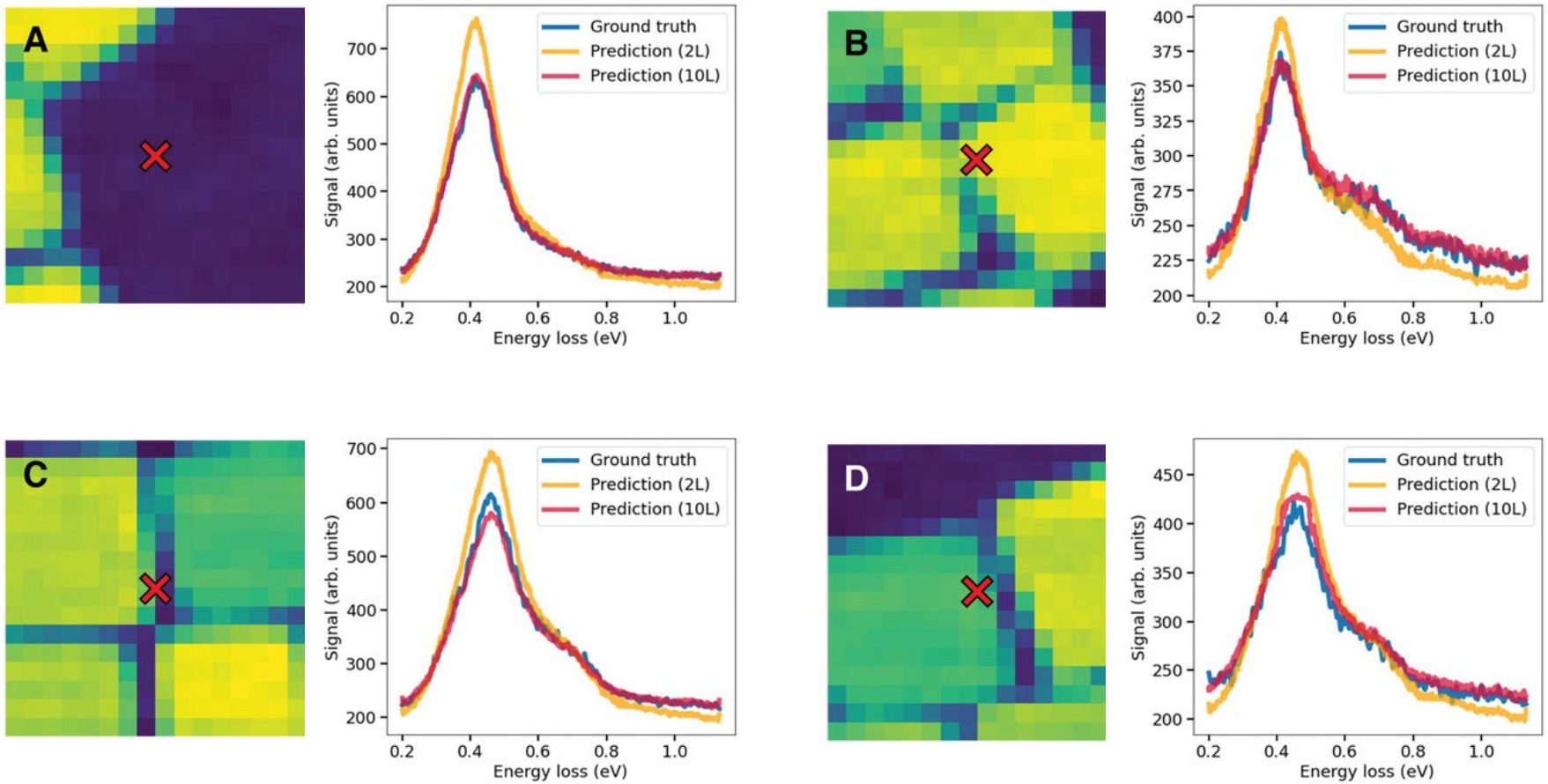
- Molecular structure – optical spectroscopy
- Atomic configurations in catalysts – catalytic activity
- Protein sequence – geometry
- Photonic structure – optical adsorption
- Materials microstructure – dielectric/conductive properties
- Composite structure – electrochemical properties
- Antenna shape – emission characteristics
- … and many more

Strategy 3: im2spec



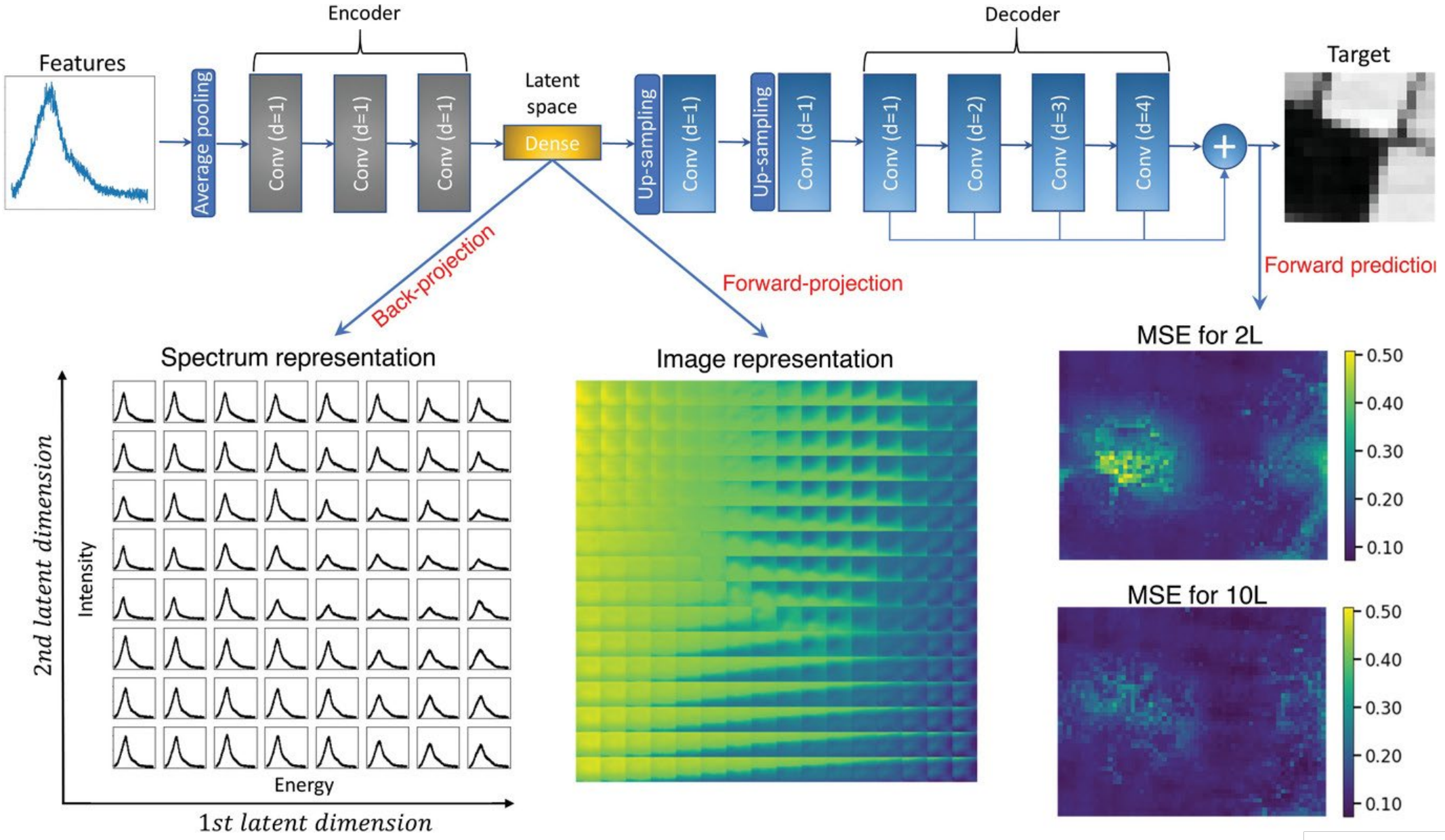
What if we create a network that encodes structure, and decodes spectra?

Im2spec prediction

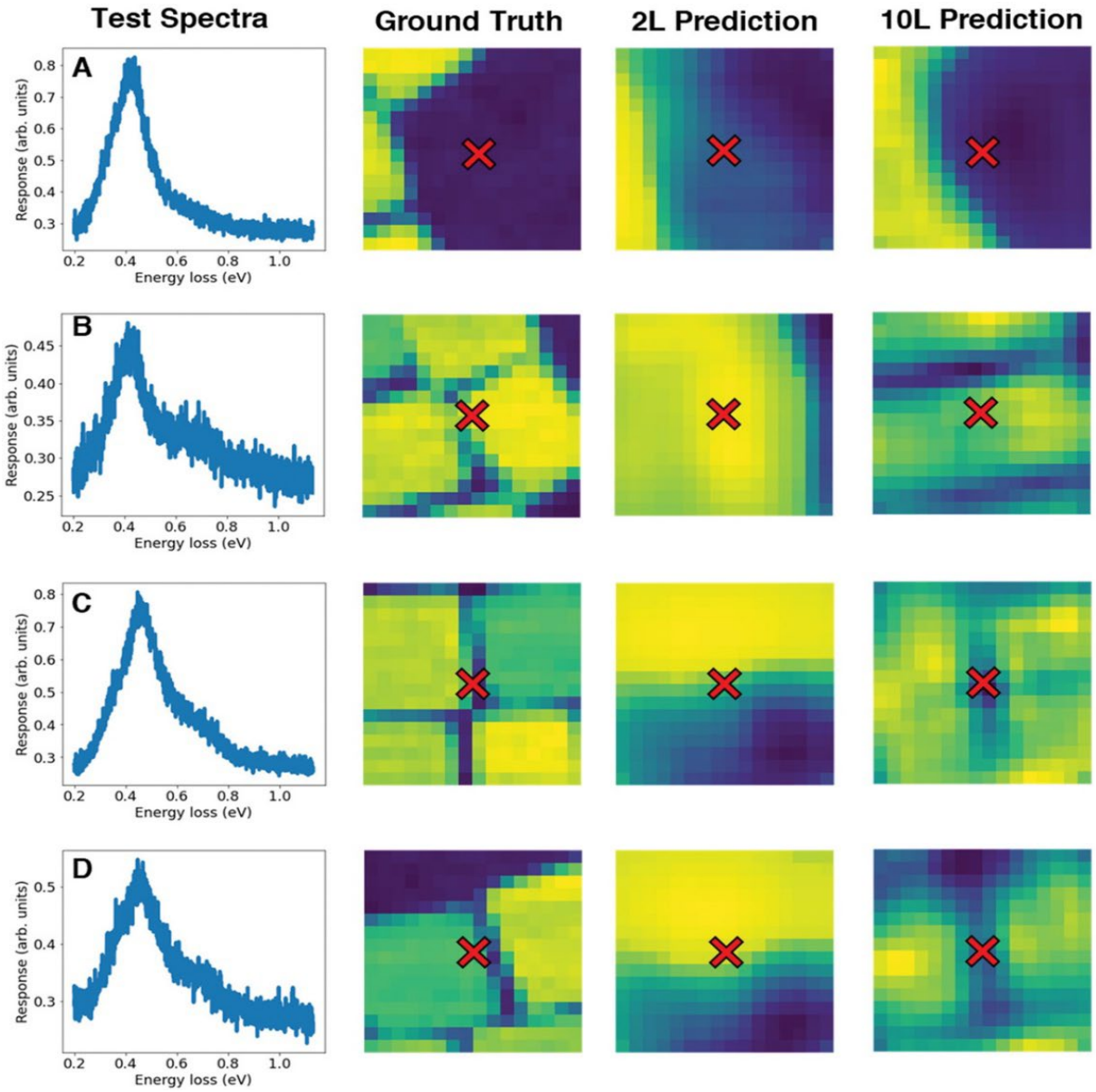


- After training, predict the spectral response of a geometric arrangement that the network has **never encountered**
- **Library** of geometric-plasmonic relationships
- Can be used for solution of inverse design in nanophotonics and other fields

Spec2im

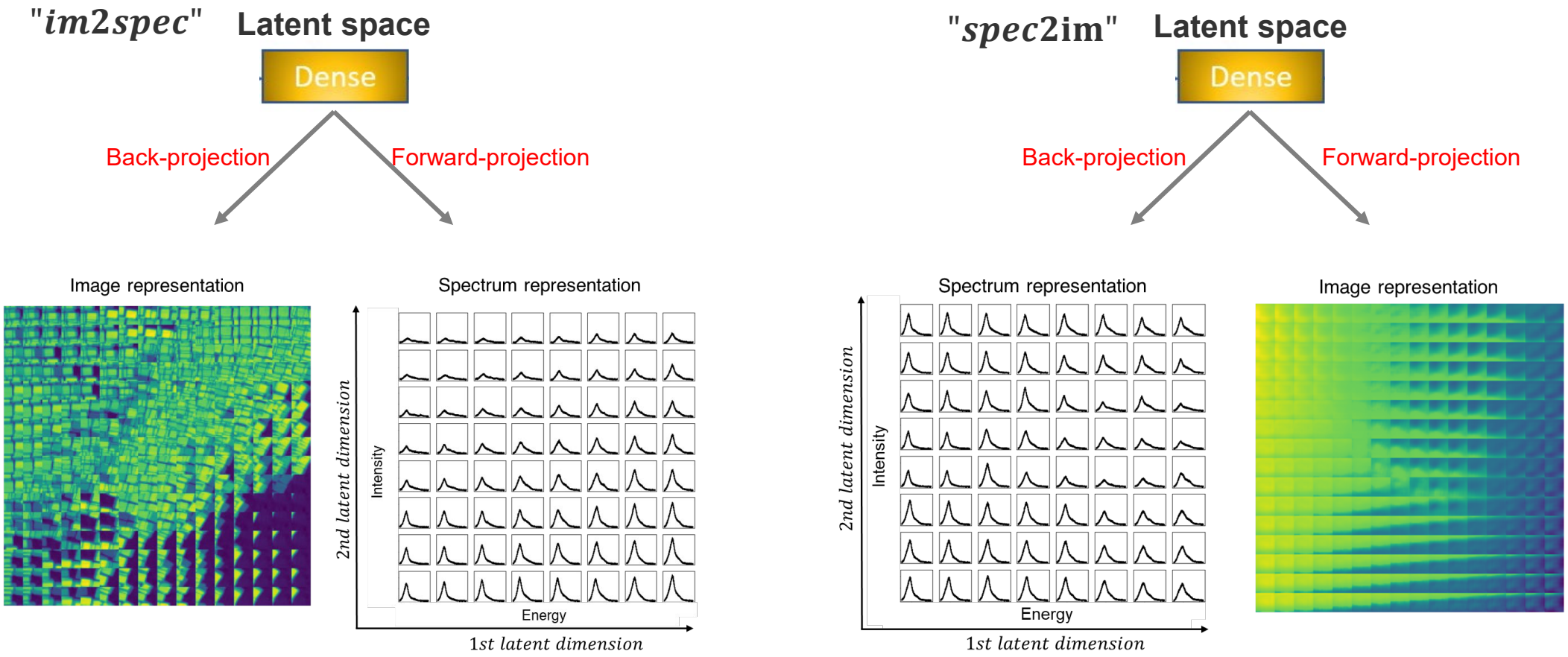

Encoder
Decoder
Features
Average pooling
Conv (d=1)
Conv (d=1)
Conv (d=1)
Latent space
Dense
Up-sampling
Conv (d=1)
Up-sampling
Conv (d=1)
Conv (d=1)
Conv (d=2)
Conv (d=3)
Conv (d=4)
Target
Back-projection
Forward-projection
Forward predictio
Spectrum representation
Image representation
MSE for 2L
MSE for 10L
2nd latent dimension
Intensity
Energy
1st latent dimension
0.50
0.40
0.30
0.20
0.10

Spec2im predictions



- After training, predict the spectral response of a geometric arrangement that the network has **never encountered**
- **2L** and **10L** refers to number of latent dimensions chosen

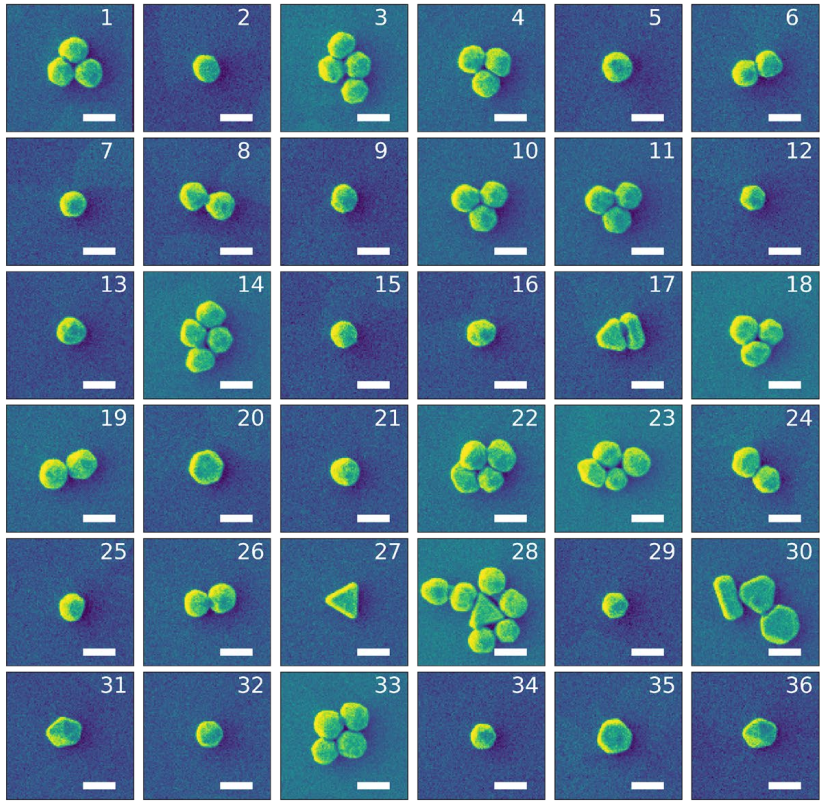
Spec2im and im2spec



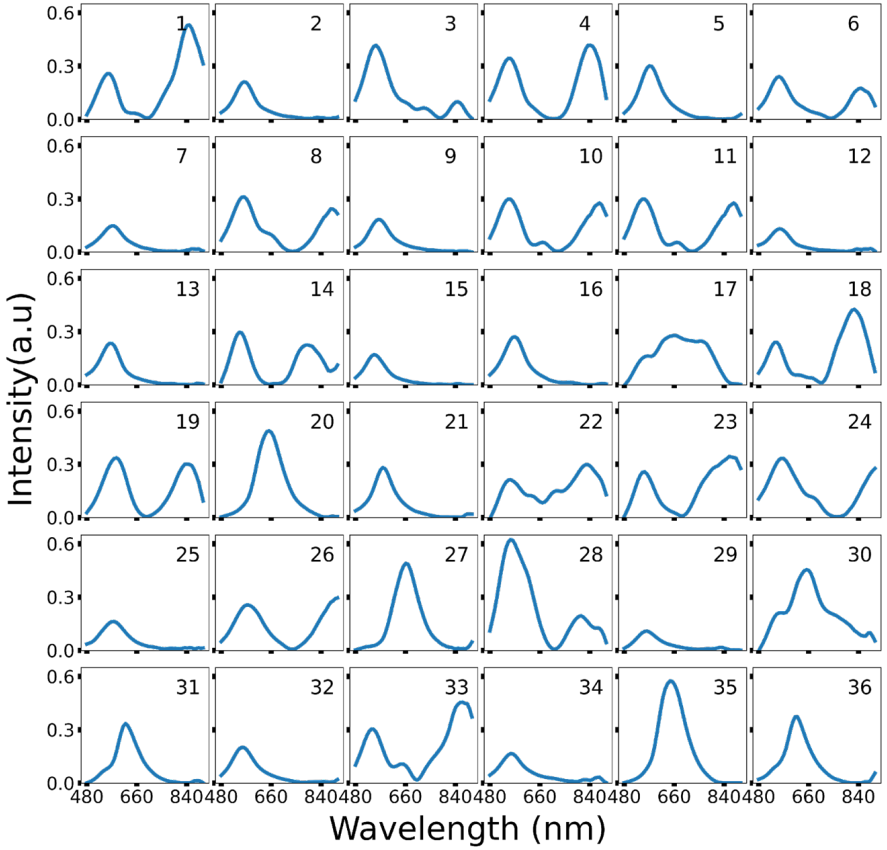
- Problem with im2spec and spec2im: they are generative only wrt. one transformation

Dual VAE: structure-property relationships

SEM images: “Structure Information”

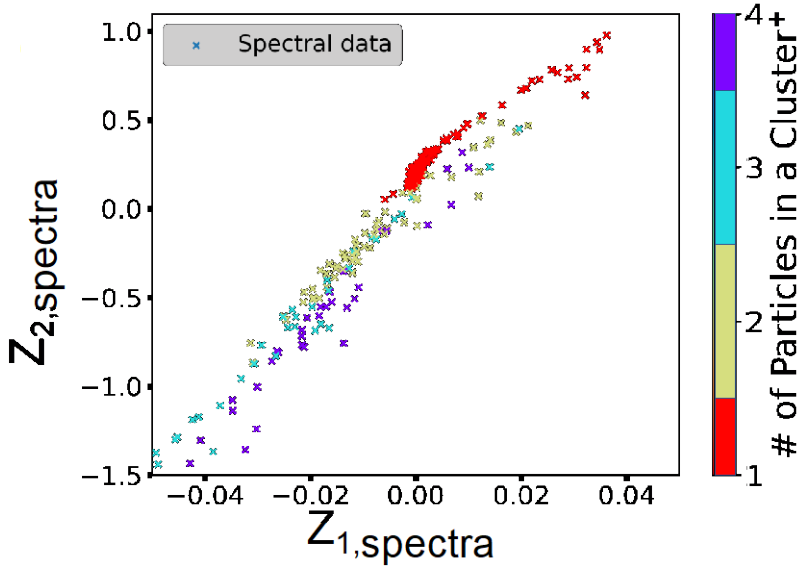


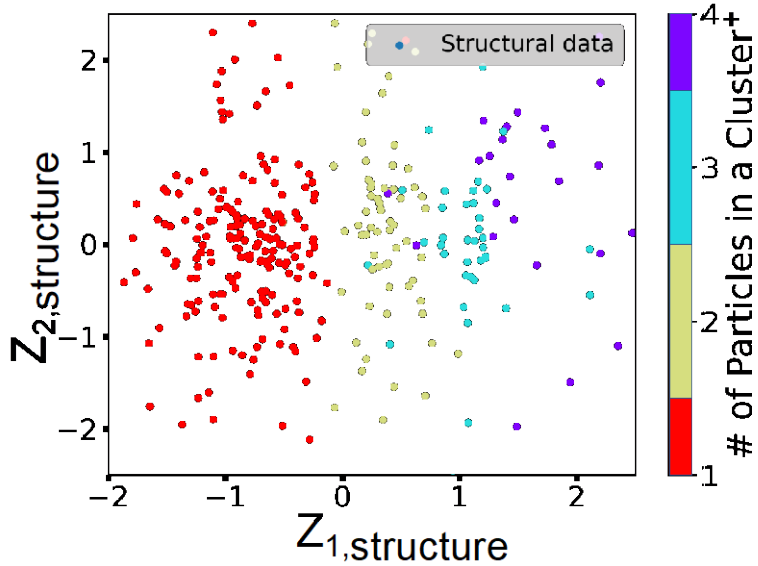
Hyperspectral microscope: “Property Information”

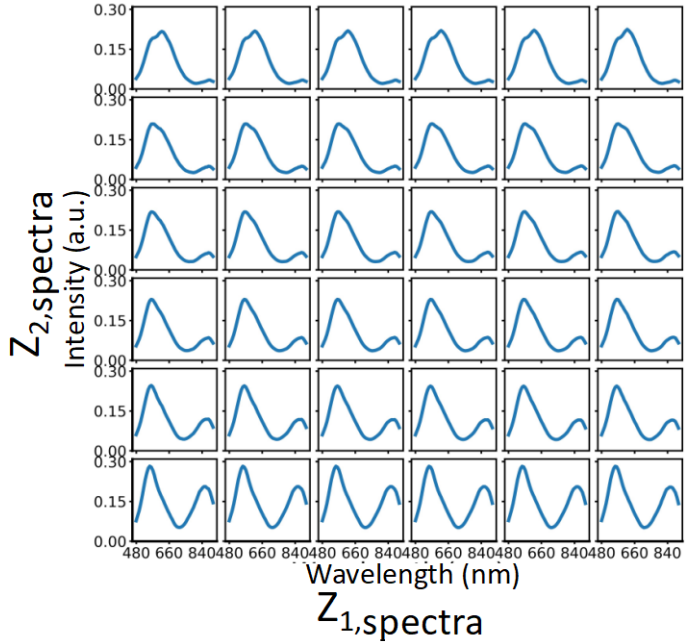


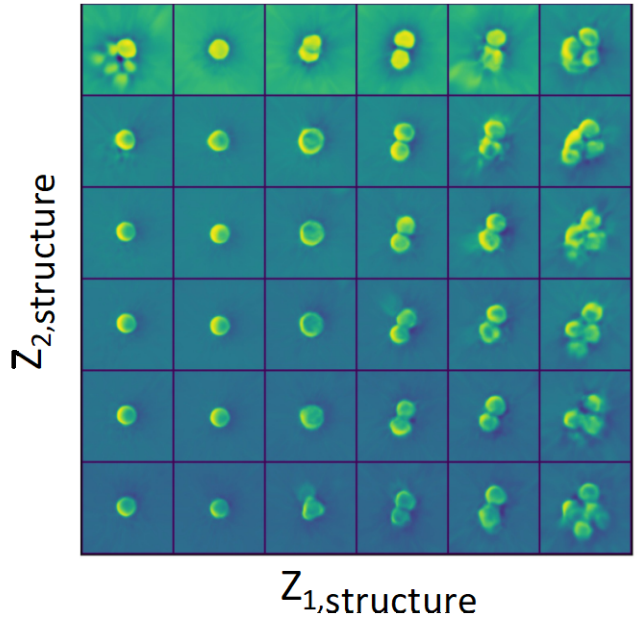
- Far field optical spectroscopy: images and spectra
- Here, we also have simple labels (number of clusters)

Separated VAE

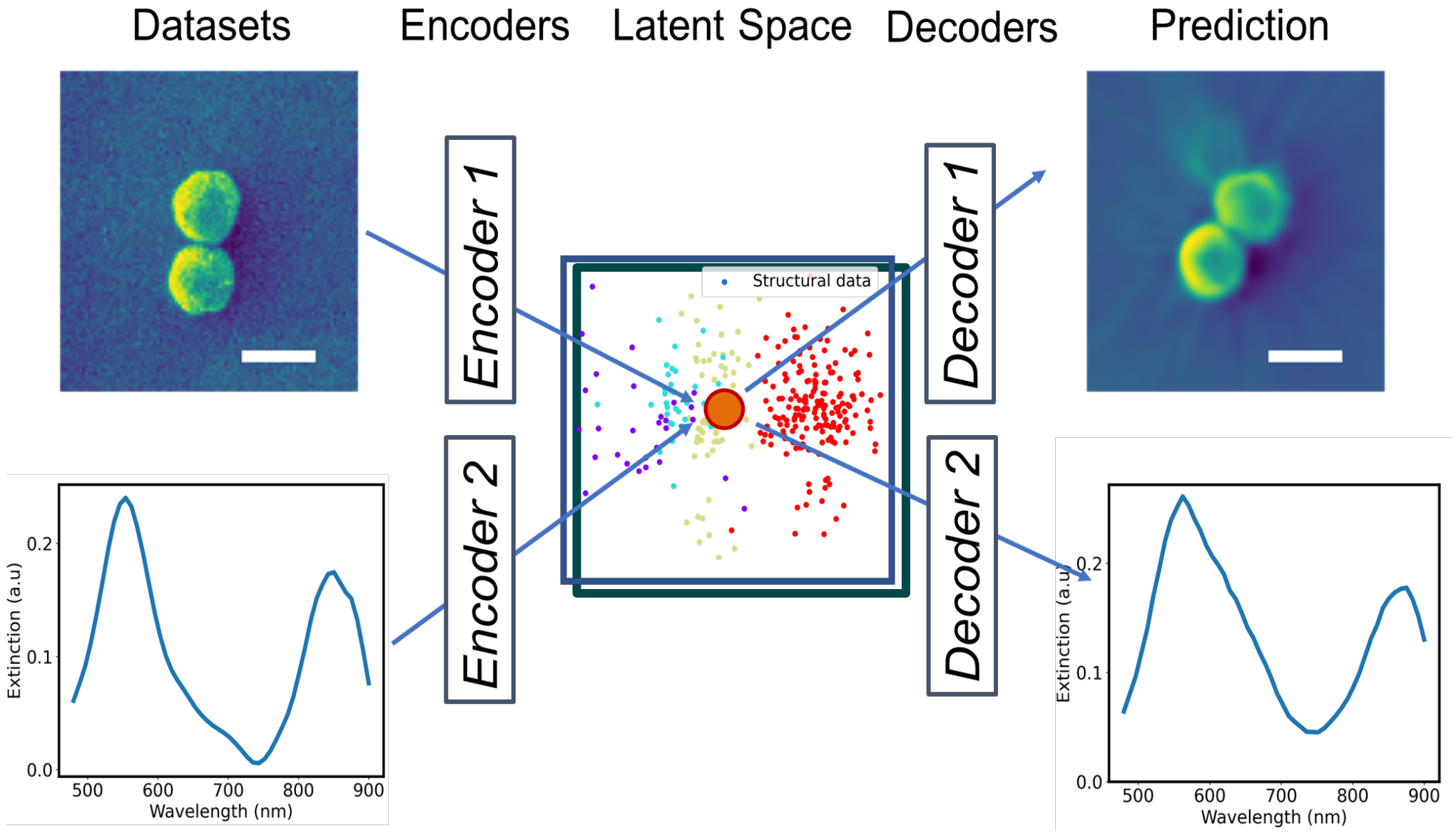








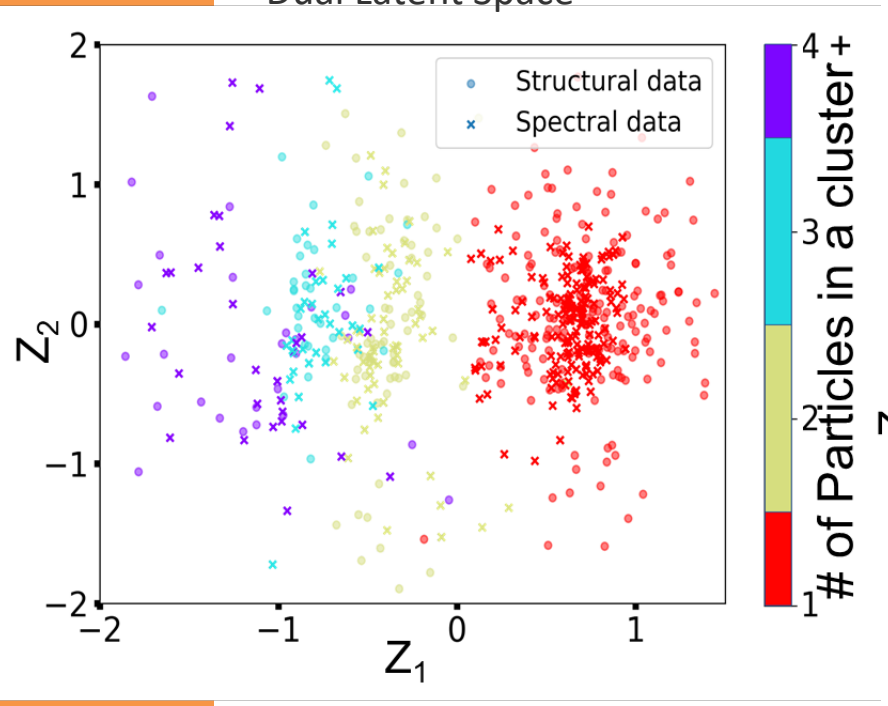
Dual VAE


Datasets
Encoders
Latent Space
Decoders
Prediction
Encoder 1
Encoder 2
Decoder 1
Decoder 2
Structural data
Extinction (a.u)
Wavelength (nm)
0.0
0.1
0.2
500
600
700
800
900

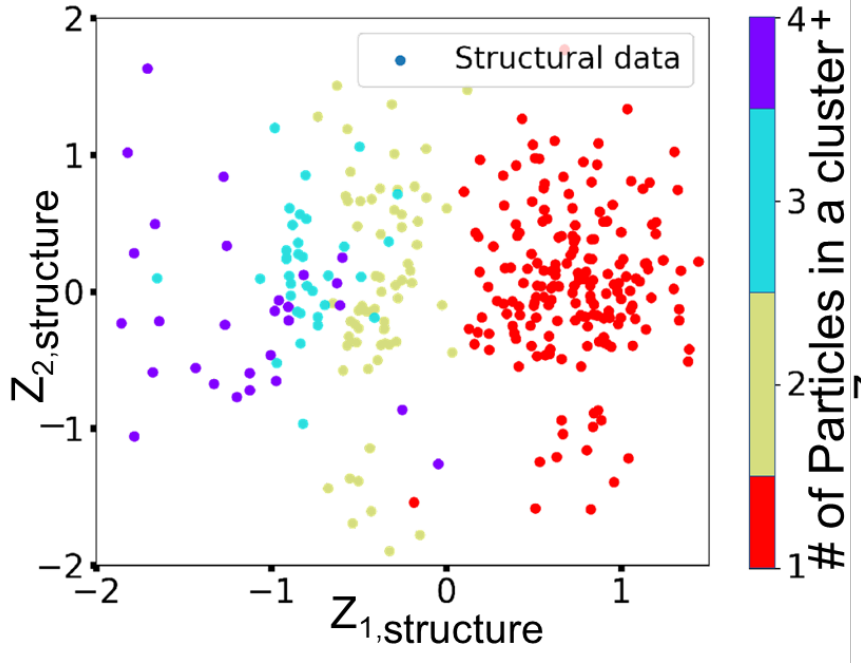
Dual VAE: Latent Distributions

Latent Space Representation

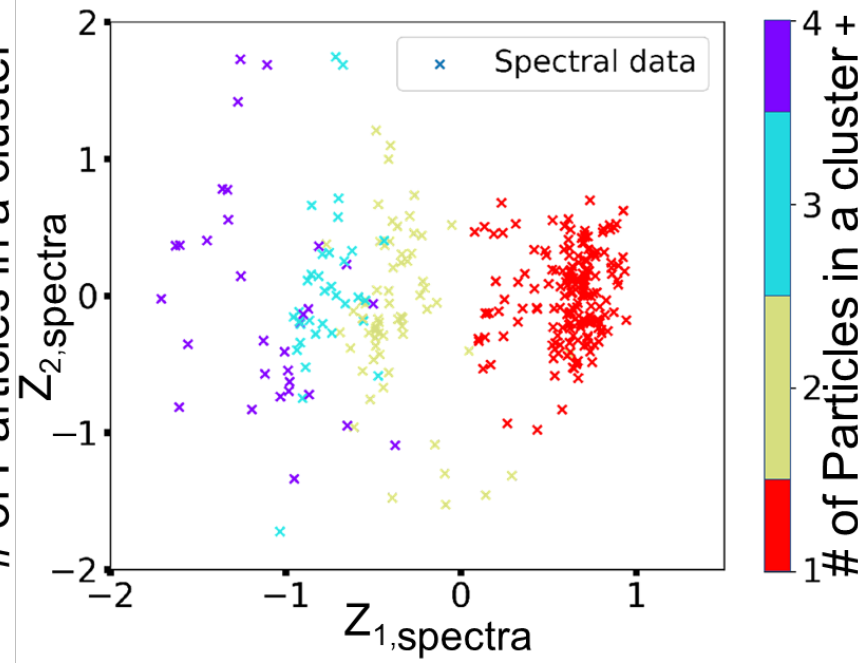
"Dual Latent Space"



"Structure Information"

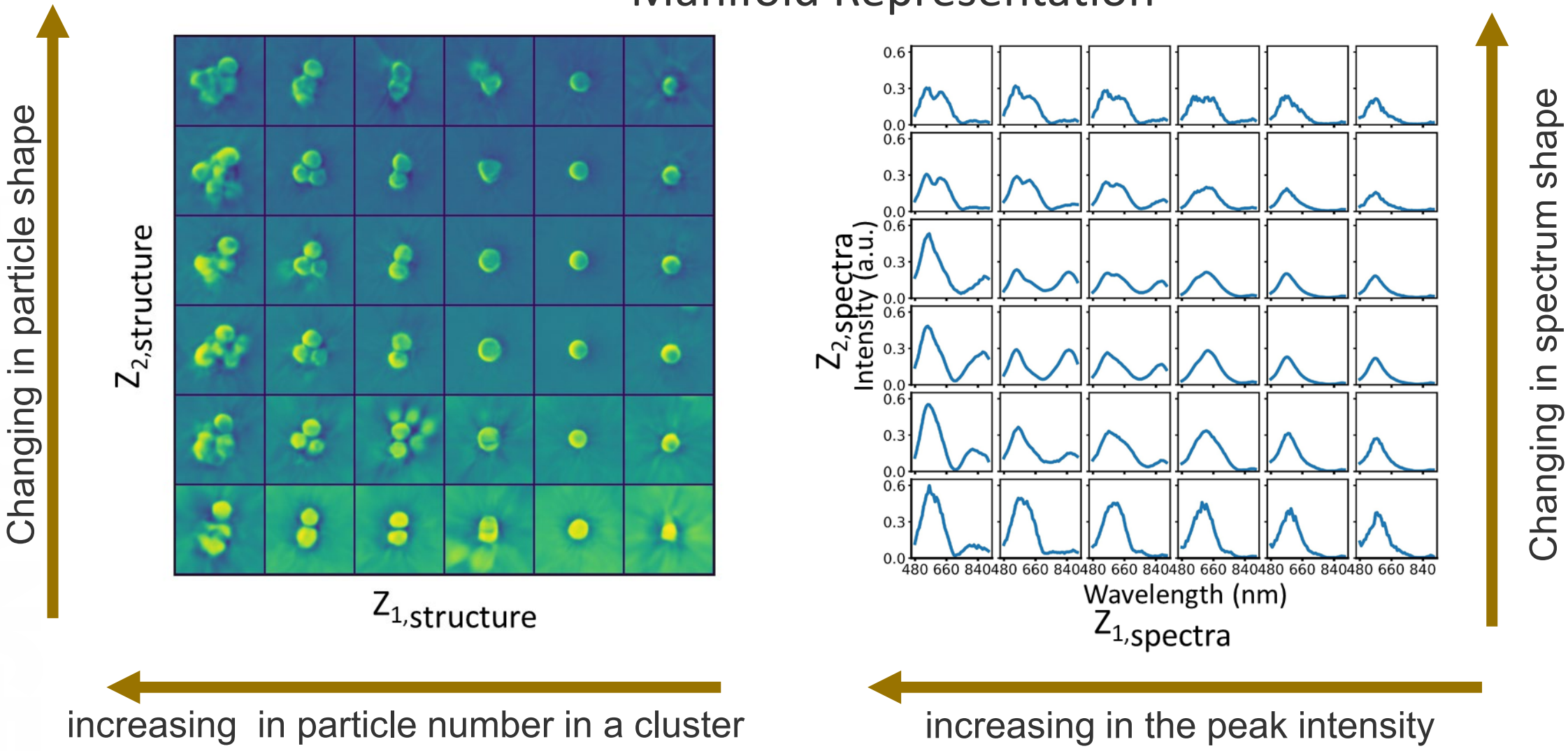


"Property Information"



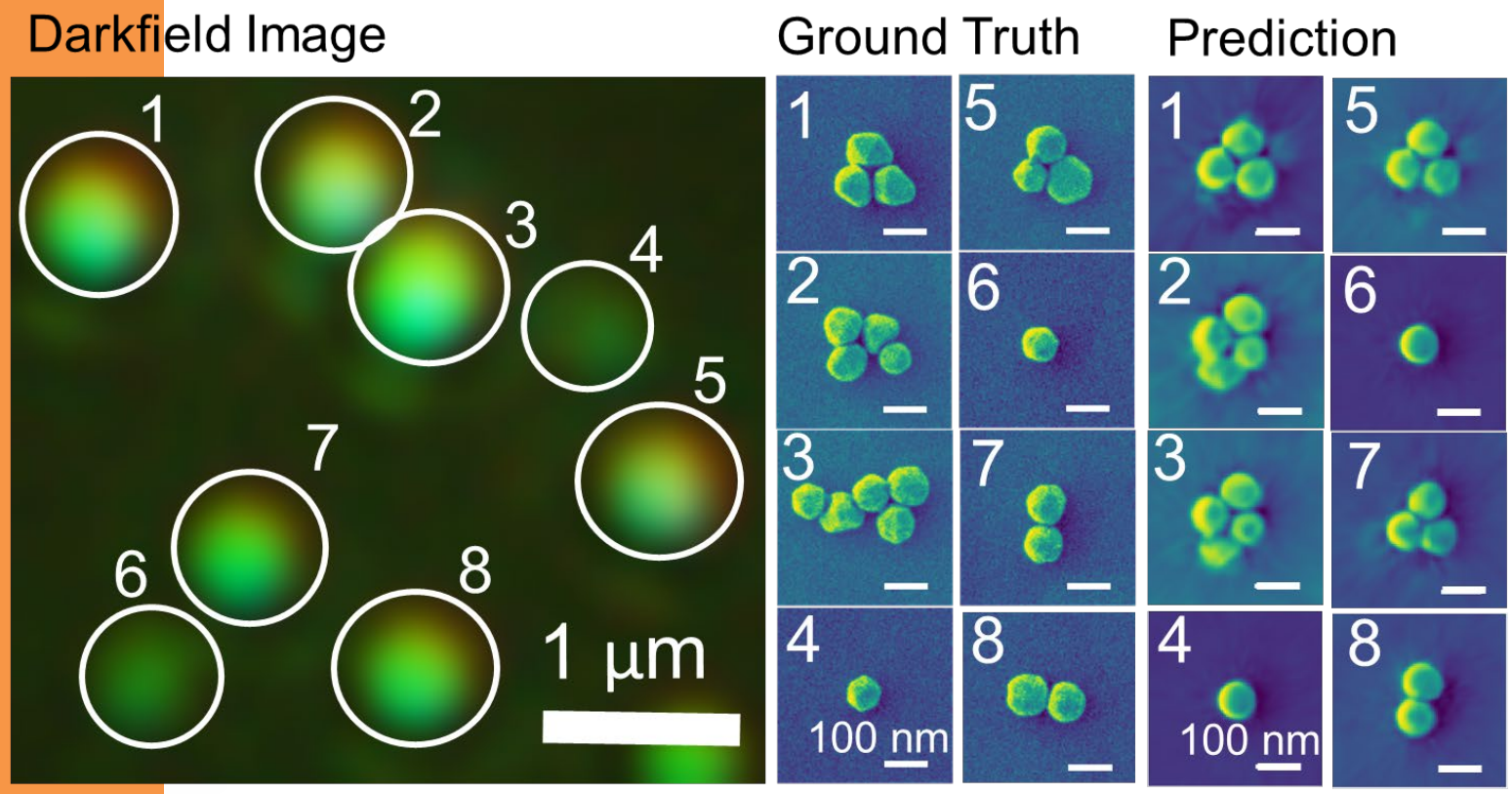
Dual VAE: Latent Representations

Manifold Representation

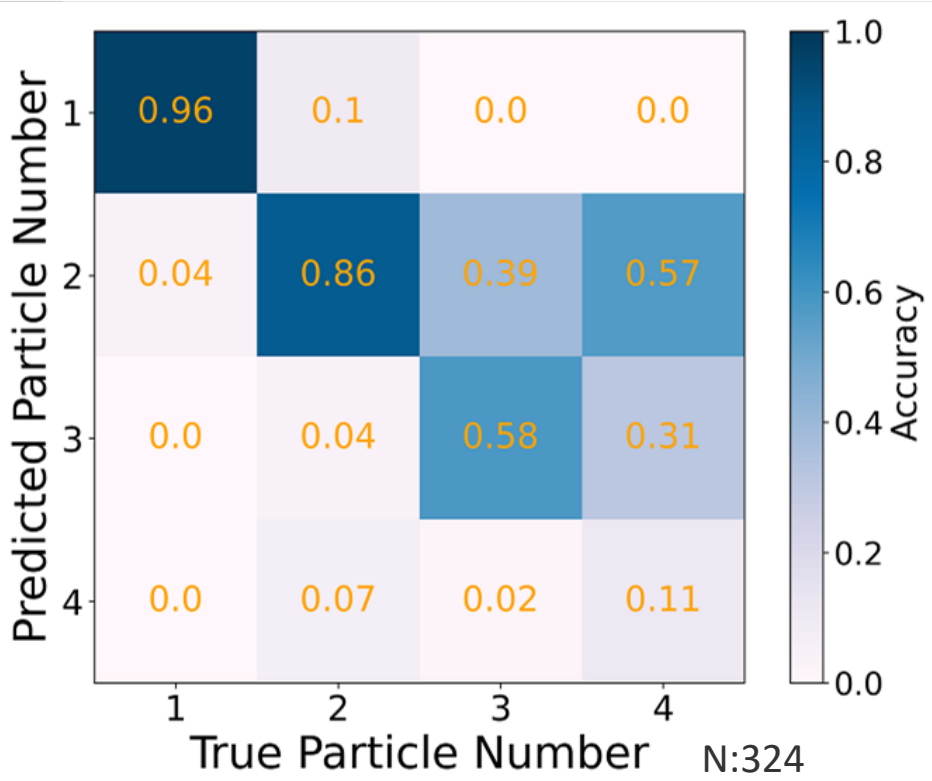


Dual VAE: Predictions

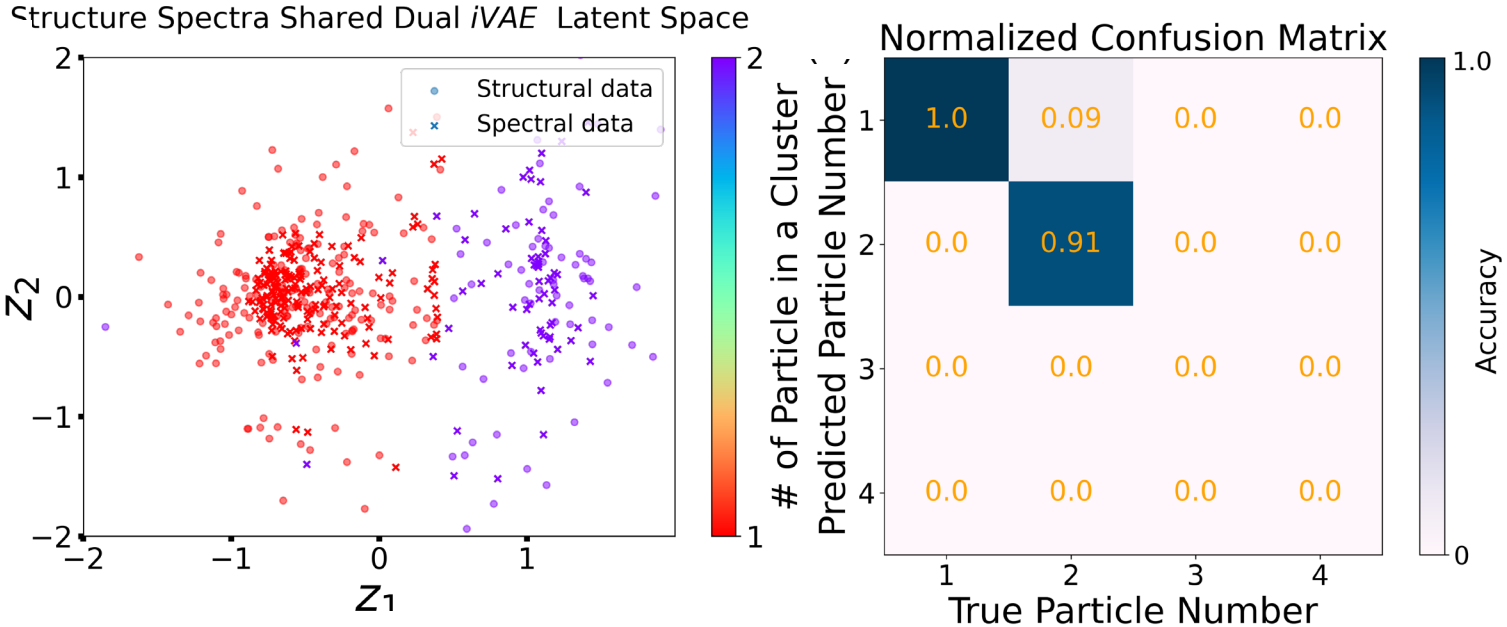
Example

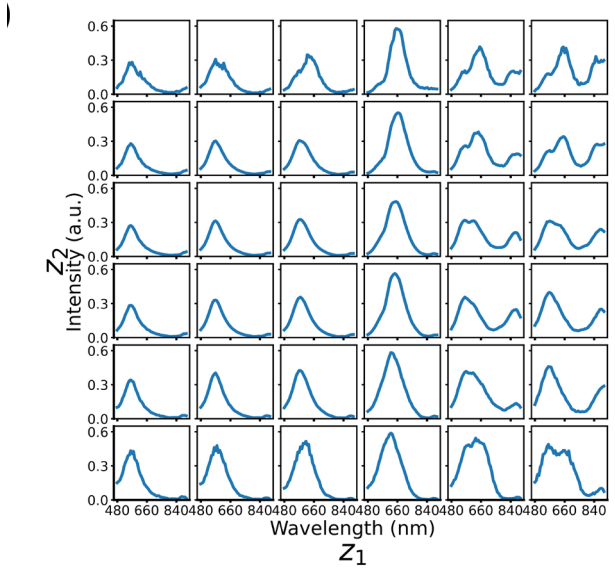


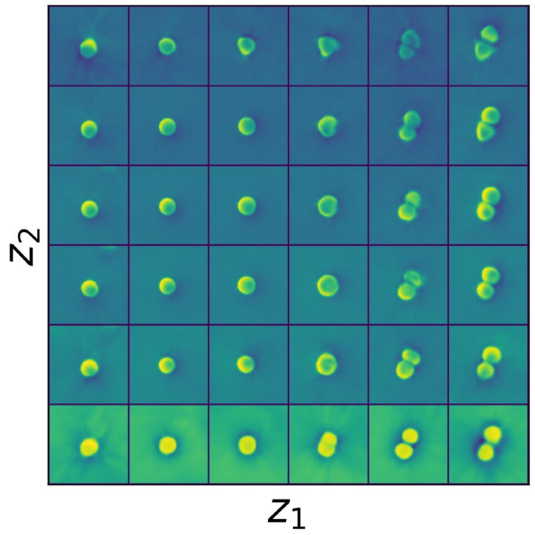
Overall Particles



Dual VAE: Predictions for only 1 and 2 mers







Questions to ask when using ML

Data science:

- What is the dimensionality of feature and target spaces
- How variable are the features in these spaces (VAE can help)
- How much data do I have?

Domain knowledge:

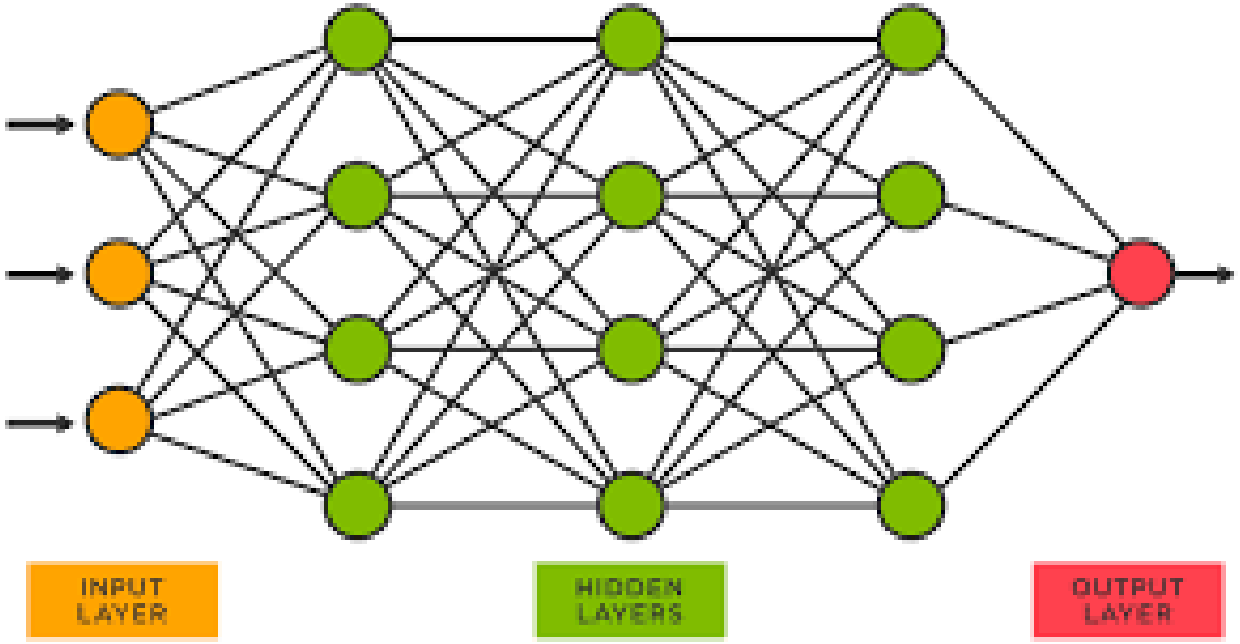
- How predictive do I expect this relationship to be
- What other factors matter from materials side
- Can measurements introduce additional factors of variance?

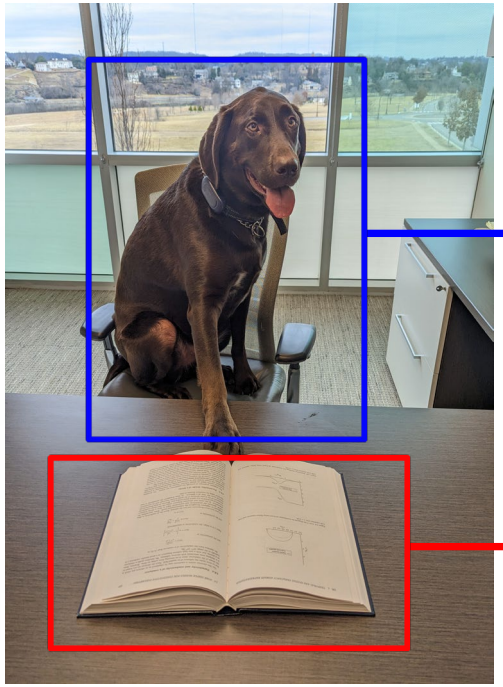
Experiment planning:

- How much data can I have?
- Is it a static learning problem, or am I interested in ML-assisted experiment?

Supervised Machine Learning

- Regression
- Classification
- Semantic segmentation
- Instance segmentation
- …

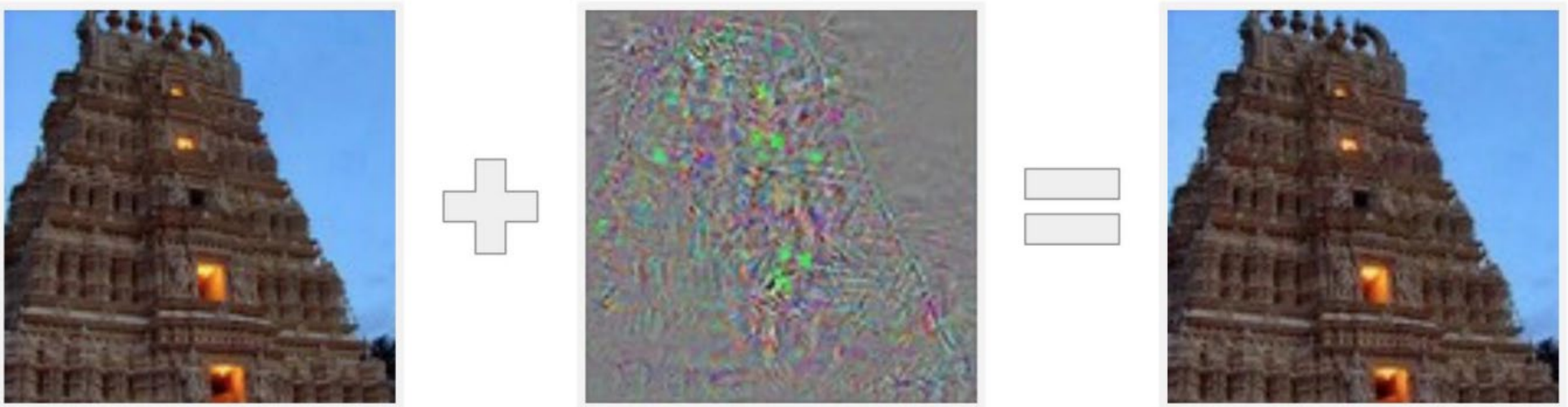




Dog

Book

Adversarial attacks



https://course.ccs.neu.edu/ds4420sp20/slides/18-ds4420-interpretability.pdf

What are the problems with ML models?

- We don’t trust the models
- We don’t know what happens in extreme cases
- Mistakes can be expensive / harmful
- Does the model make similar mistakes as humans ?
- How to change model when things go wrong ?

What do we want to get?

- Interactive feedback - can model learn from human actions in online setting ? (Can you tell a model to not repeat a specific mistake ?)
- Recourse – Can a model tell us what actions we can take to change its output ? (For example, what can you do to improve your credit score?)

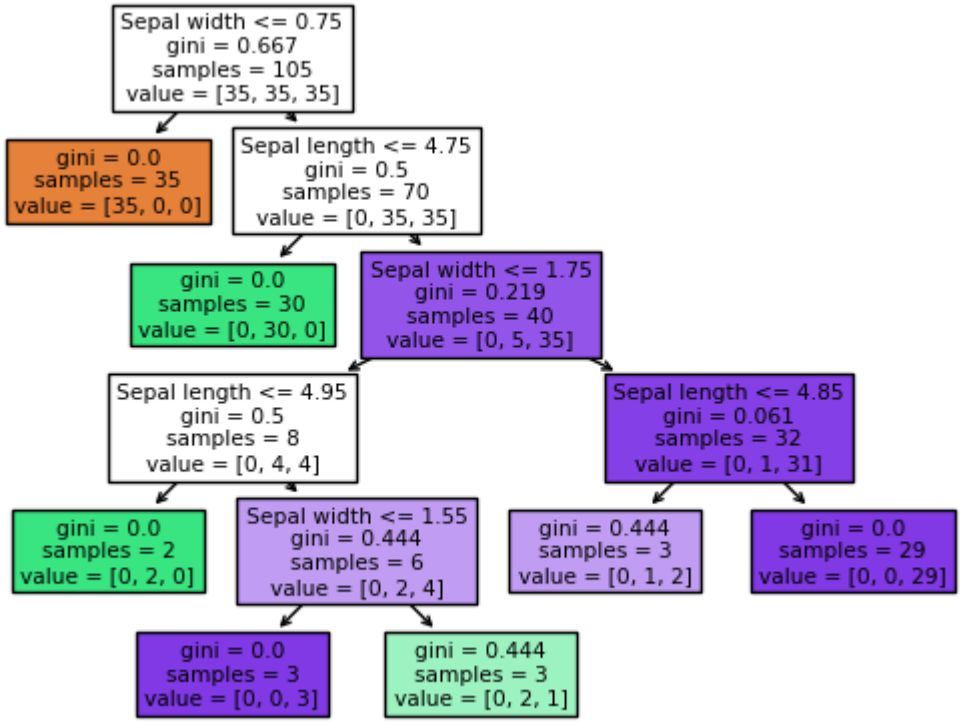
Models: Explainable and Not

Some models are explainable:

1. Linear or physics-defined function
2. Decision trees

But what about:

1. Image segmentation
2. Natural language processing
3. Classification
4. …



What is explainability?

- **Faithfulness:** how to provide explanations that accurately represent the true reasoning behind the model's final decision.
- **Plausibility:** Is the explanation correct or something we can believe is true, given our current knowledge of the problem
- **Understandable:** Can I put it in terms that end user without in-depth knowledge of the system can understand ?
- **Stability:** Does similar instances have similar interpretations ?

What do we expect from explainer?

1. Interpretable: It should provide a qualitative understanding between the input variables and the response. It should be easy to understand.

2. Local Fidelity: It might not be possible for an explanation to be completely faithful unless it is the complete description of the model itself. Having said that it should be at least locally faithful, i.e it must replicate the model's behavior in the vicinity of the instance being predicted.

3. Model Agnostic: The explainer should be able to explain any model and should not make any assumptions about the model while providing explanations.

4. Global perspective:The explainer should explain a representative set to the user so that the user has a global intuition of the model

https://towardsdatascience.com/decrypting-your-machine-learning-model-using-lime-5adc035109b5