The 8th International Conference on Deep Learning in Computational Physics
June 19-21, 2024
SINP MSU, Moscow, Russia

Improving Representativity of Spectroscopic Data using Variational Autoencoders: Approaches and Problems*

Mushchina A.S.[1,2], Isaev I.V.[1], Sarmanova O.E.[1,2], Dolenko T.A.[1,2], Dolenko S.A.[1]

[1] D.V.Skobeltsyn Institute of Nuclear Physics, M.V.Lomonosov Moscow State University,
[2] Faculty of Physics, M.V.Lomonosov Moscow State University

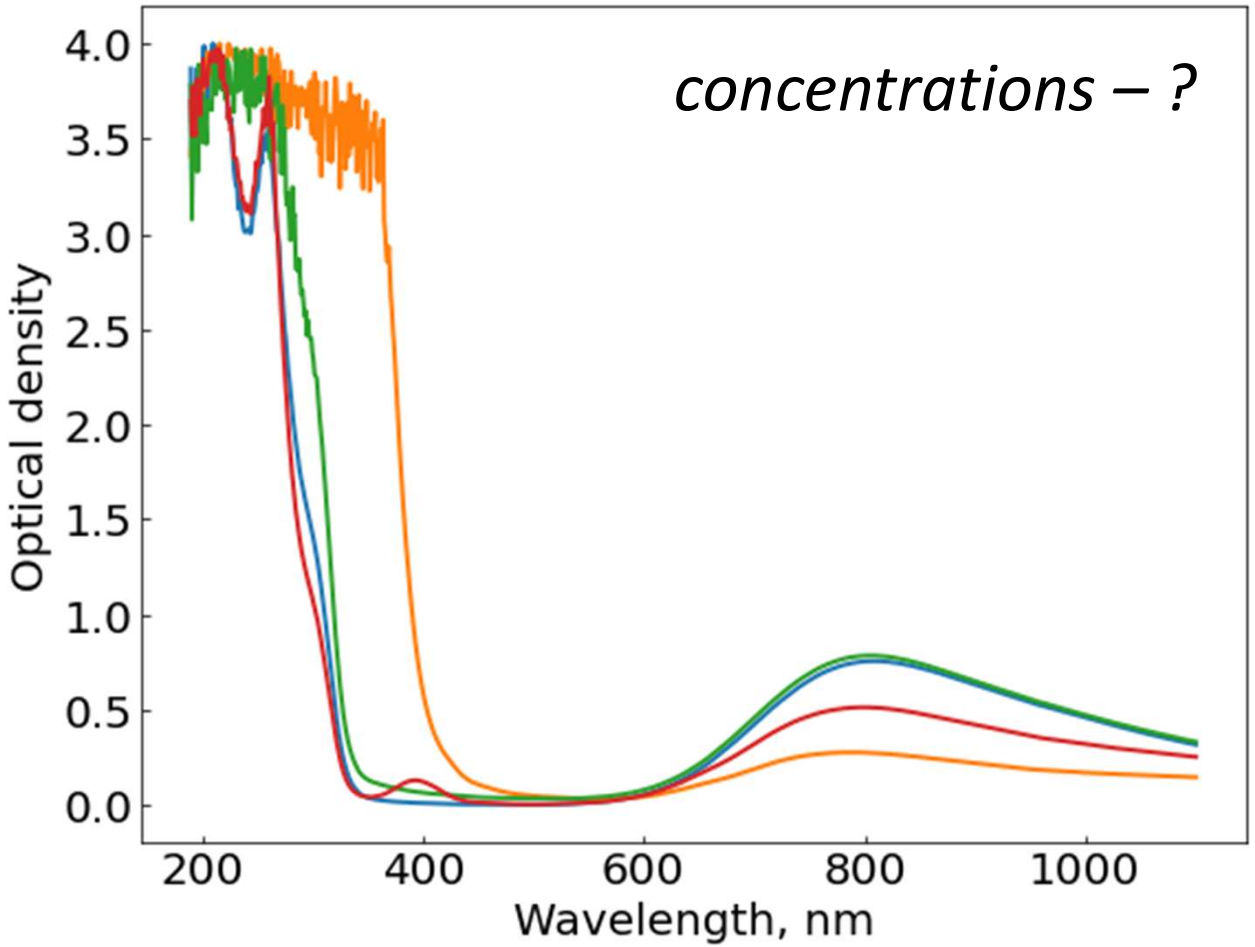
** The study was carried out at the expense of the grant No. 24-11-00266 from the Russian Science Foundation, https://rscf.ru/en/project/24-11-00266/.*

Background.
Inverse problem of spectroscopy

Optical absorption spectra of multicomponent aqueous solutions of salts.

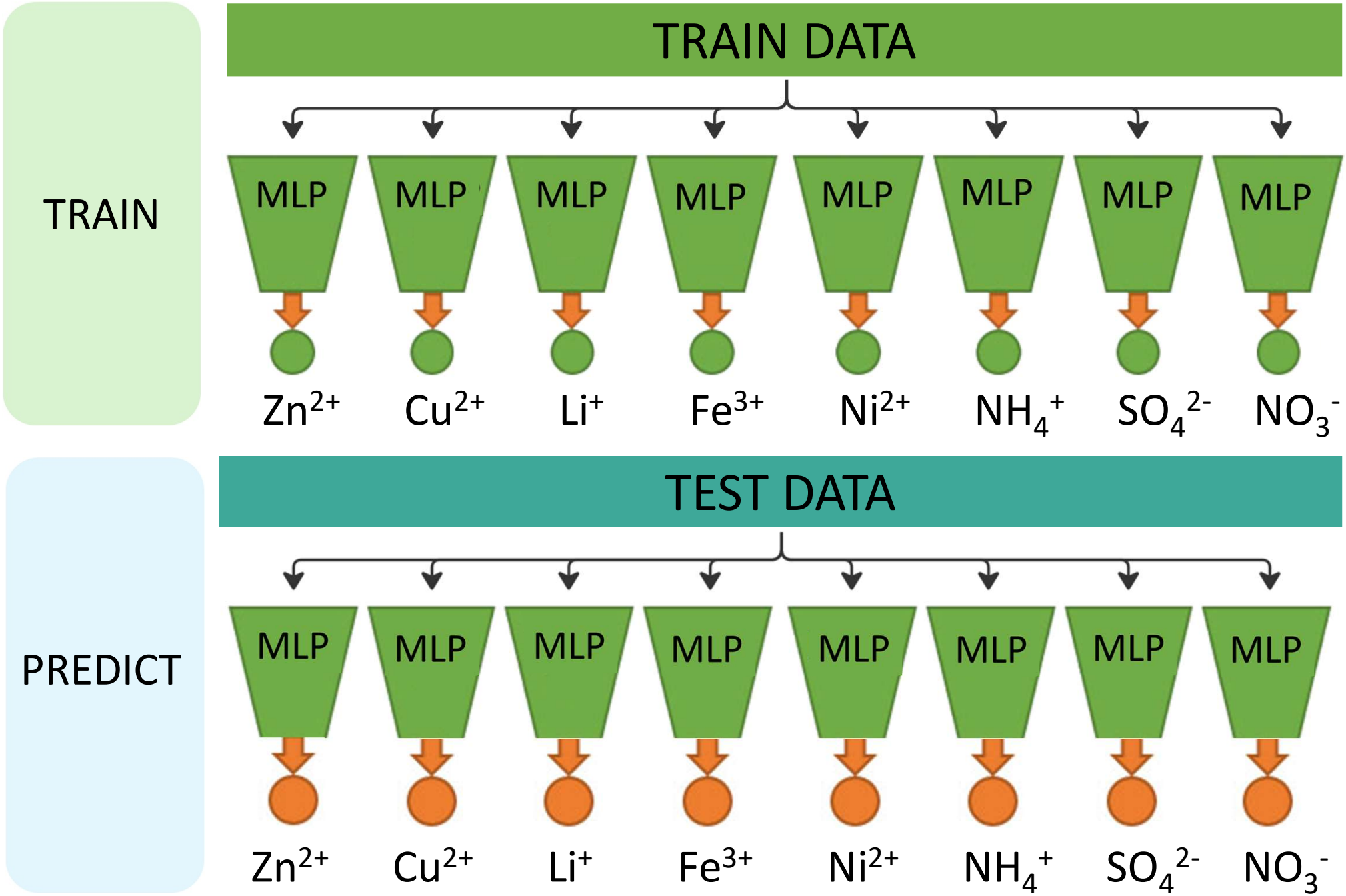
Ions: Zn^{2+}, Cu^{2+}, Li^{+}, Fe^{3+}, Ni^{2+}, NH_4^{+}, SO_4^{2-}, NO_3^{-}.

➢ 3744 spectra ➢ 911 channels



- Spectroscopic methods ensure remote and express sensing and provide a wide range of features that identify the constituents of water media
- High-dimensional and complex problem
- Requiring analysis of many spectral channels at once
- Such analysis may be performed using machine learning methods (ML), e.g. neural networks

Background. Baseline solution



Challenges in acquiring spectroscopic datasets for inverse problem solving with ML

- The experiments are **laborious, expensive and time-consuming**
- Spectra registration requires **complex equipment**
- Data interpretation and labeling requires **competent experts**
- Obtaining a representative dataset requires a **substantial amount of data**

To minimize the costs associated with increasing the data set, computer science community has developed a number of data augmentation techniques

Augmentation. Problems

- Adding noise
  - Results in an increased noise resilience of the model, but not in a decrease in the inverse problem solution error
- Using interpolation
  - The shape of the spectra is sensitive to the concentrations of ions, the dependence of spectral intensities on ion concentrations in multi-component solutions is complex and non-linear
- Using open resources
  - Lack of databases with various ion concentrations
  - Lack of datasets with specific components
  - Difficulties in adapting data patterns from different equipment to a single format

Motivation

- Challenges in acquiring spectroscopic data for ML
- Issues with standard augmentation methods

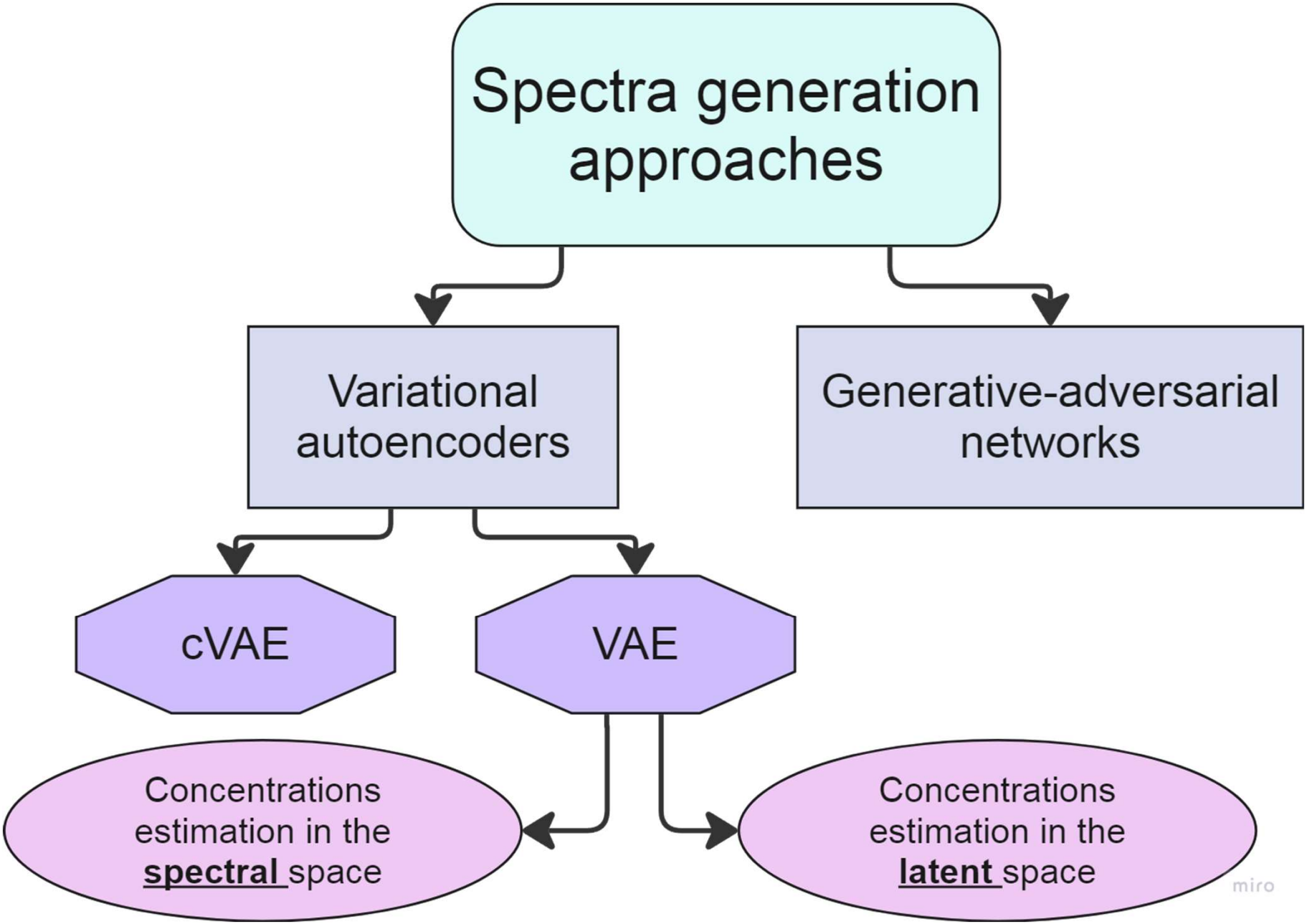
Objective

Increase the representativity

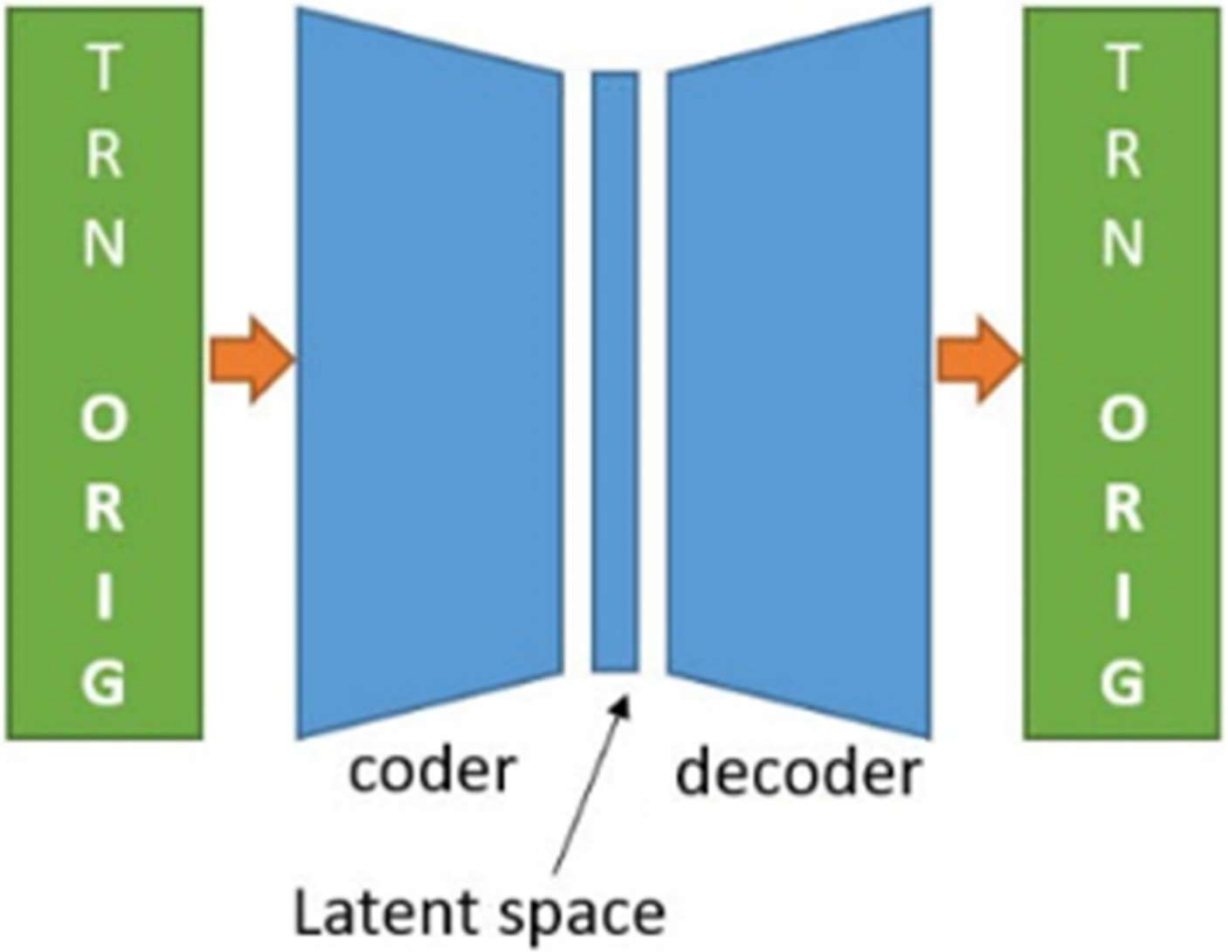
of optical spectroscopy datasets

via machine learning

ML approaches



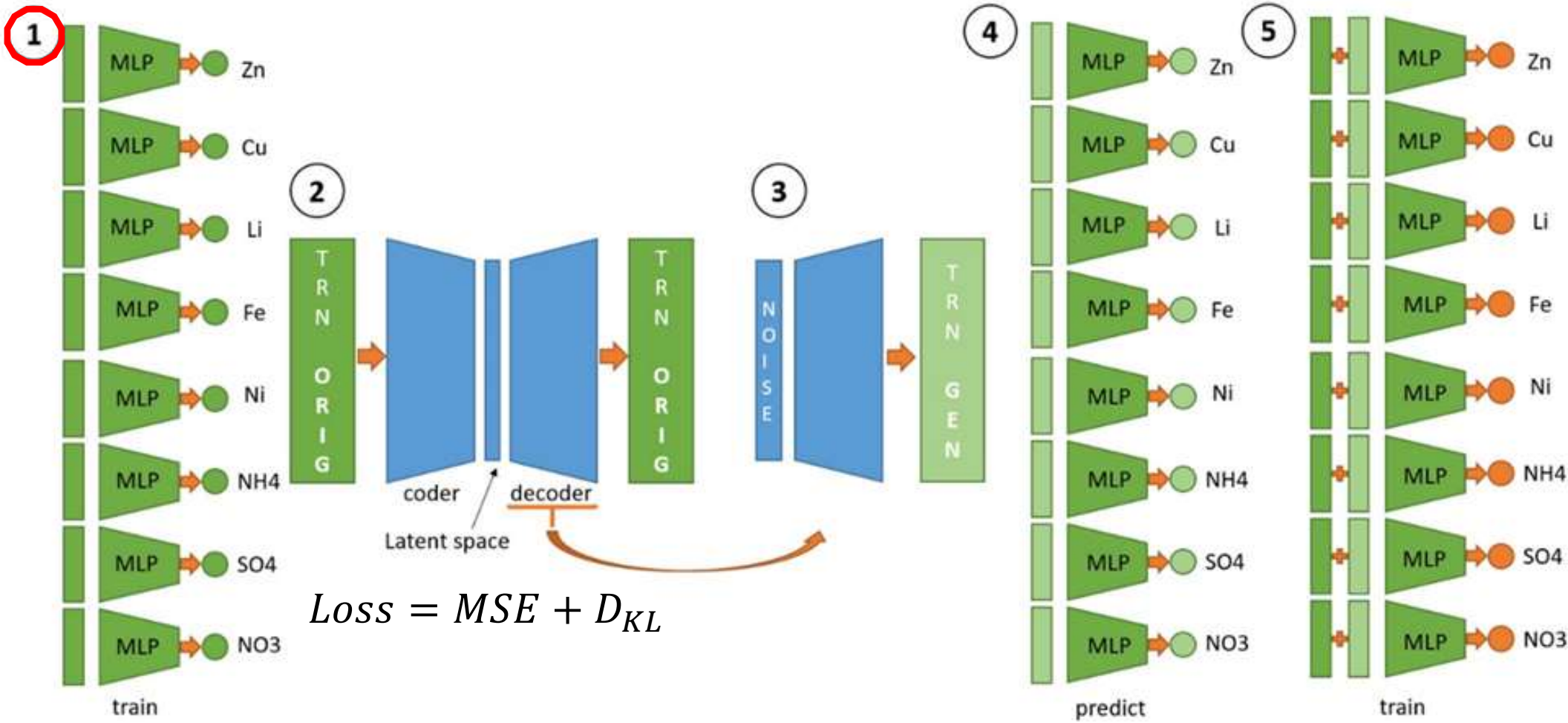
Variational Autoencoder (VAE)



$$Loss = MSE + D_{KL}$$

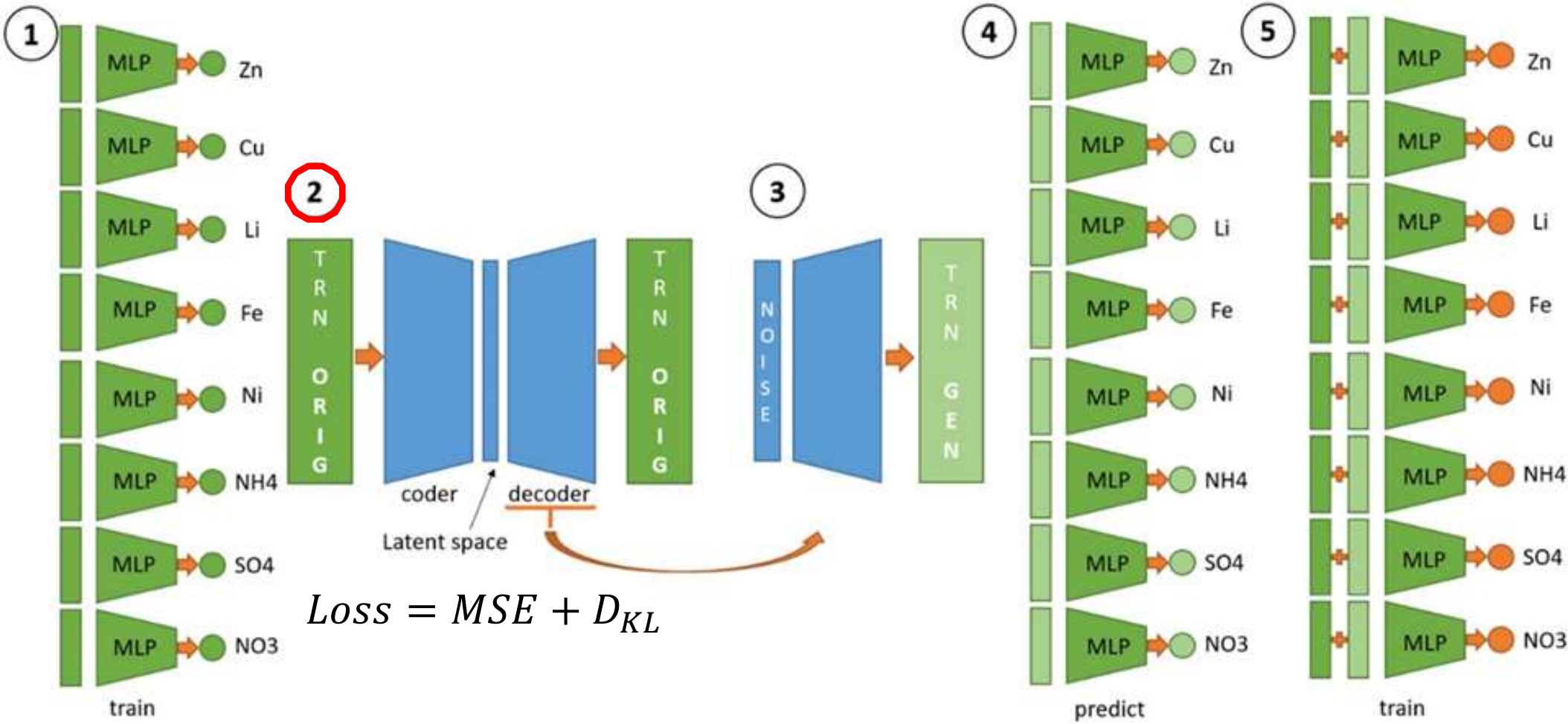
- Compresses the input signal into a signal of a smaller dimension in the latent space
- The information about the data in the latent space is represented as parameters of a certain multidimensional distribution
- Allows generating samples from random vectors drawn from this distribution using the decoder

Experimental pipeline: Variational Autoencoder (VAE)



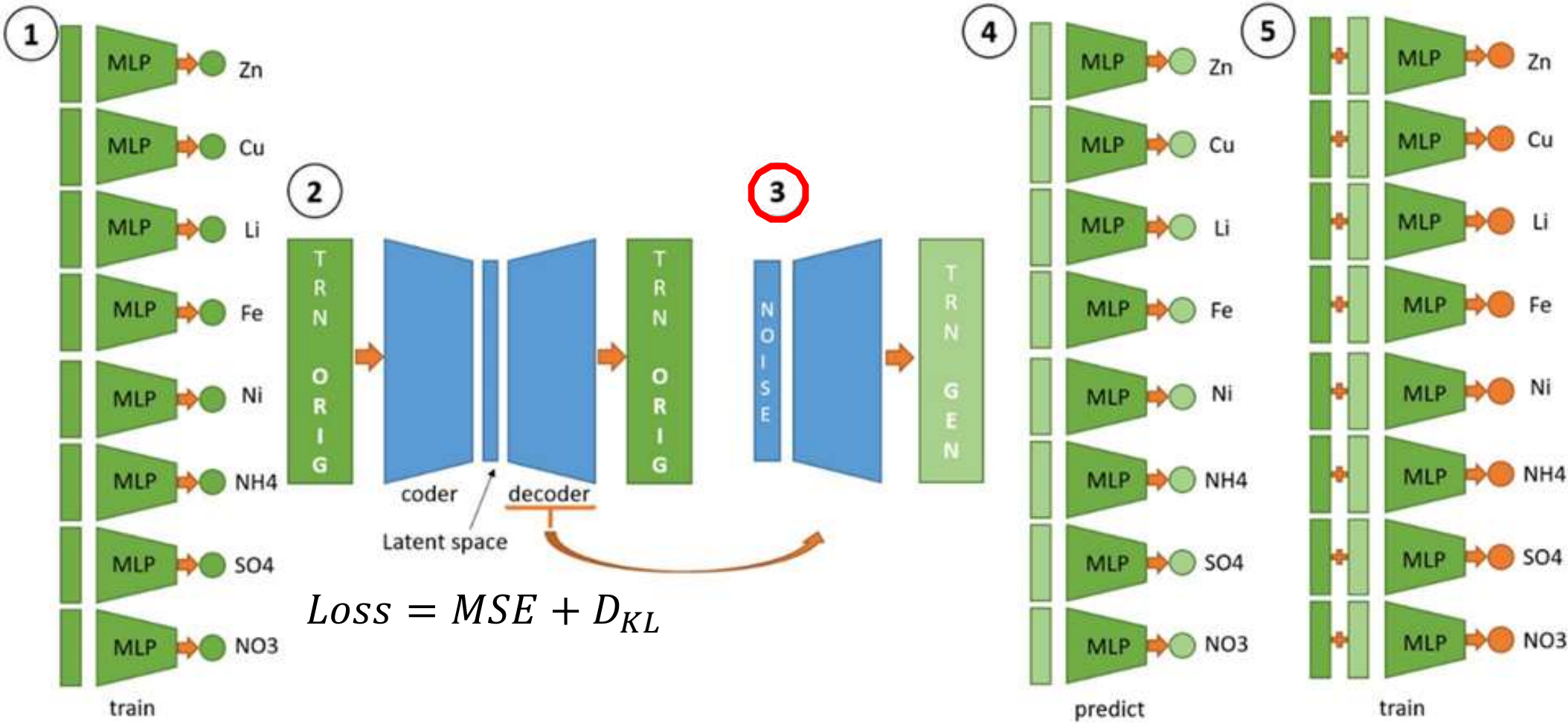
1) Training regression networks for each ion on experimental data.

Experimental pipeline: Variational Autoencoder (VAE)



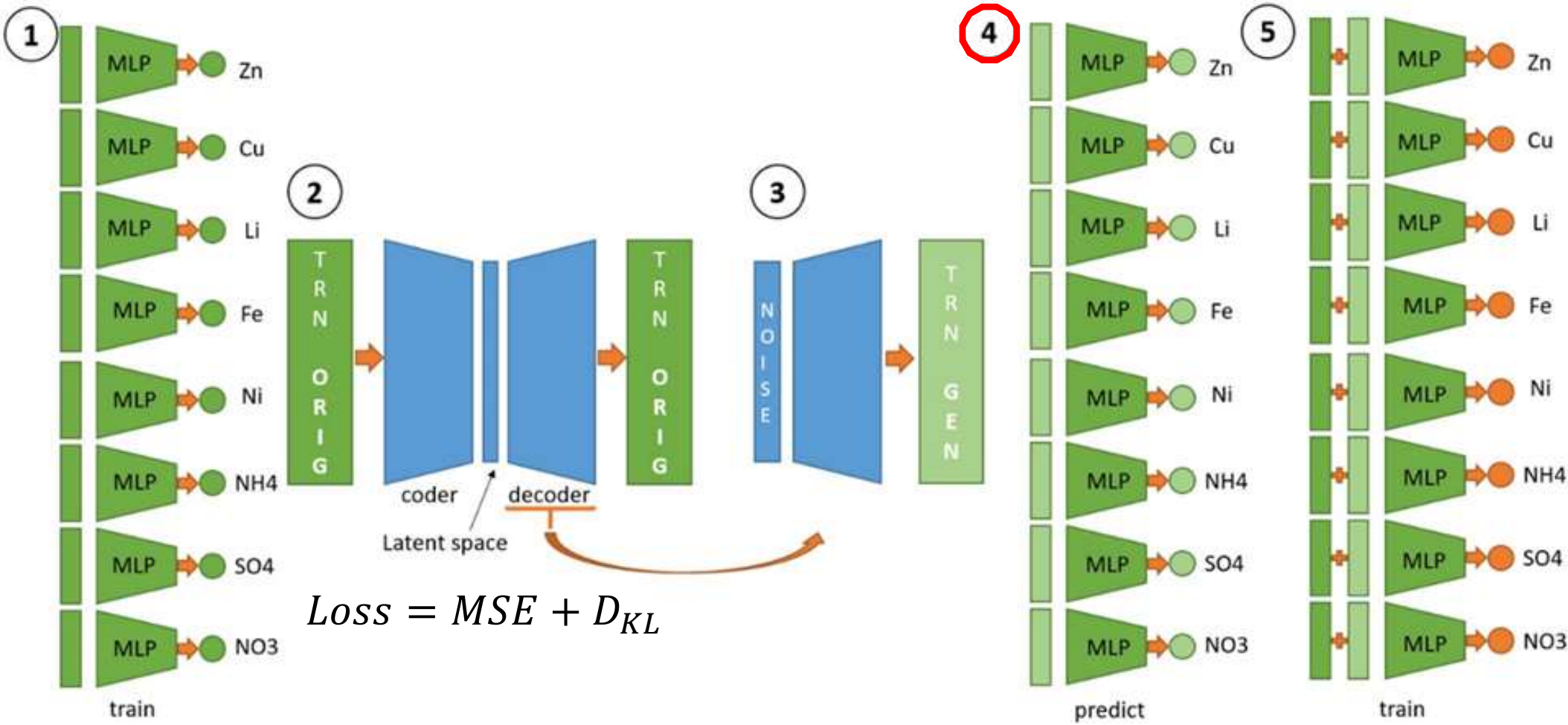
2) Training VAE

Experimental pipeline: Variational Autoencoder (VAE)



3) Generating patterns using VAE to double the size of the training dataset

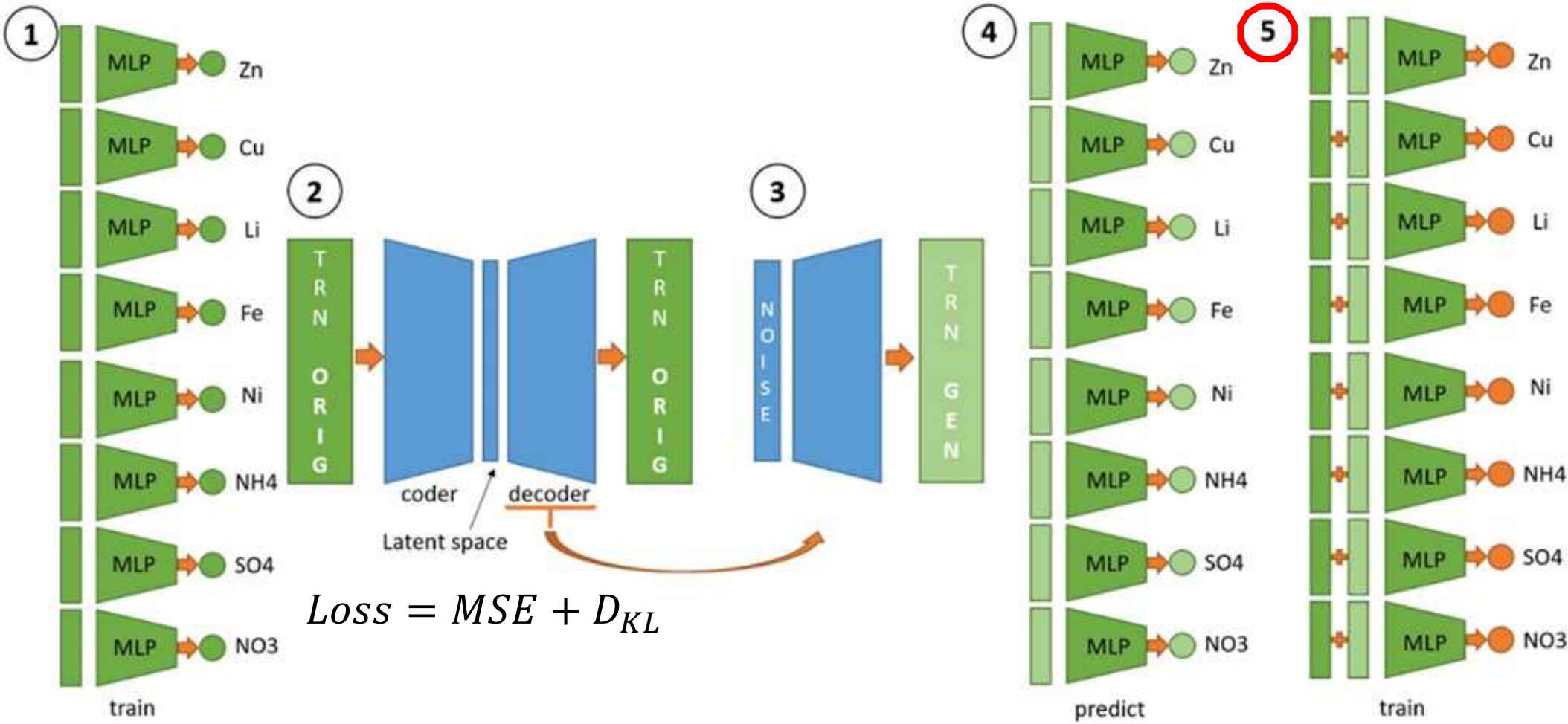
Experimental pipeline: Variational Autoencoder (VAE)



$$Loss = MSE + D_{KL}$$

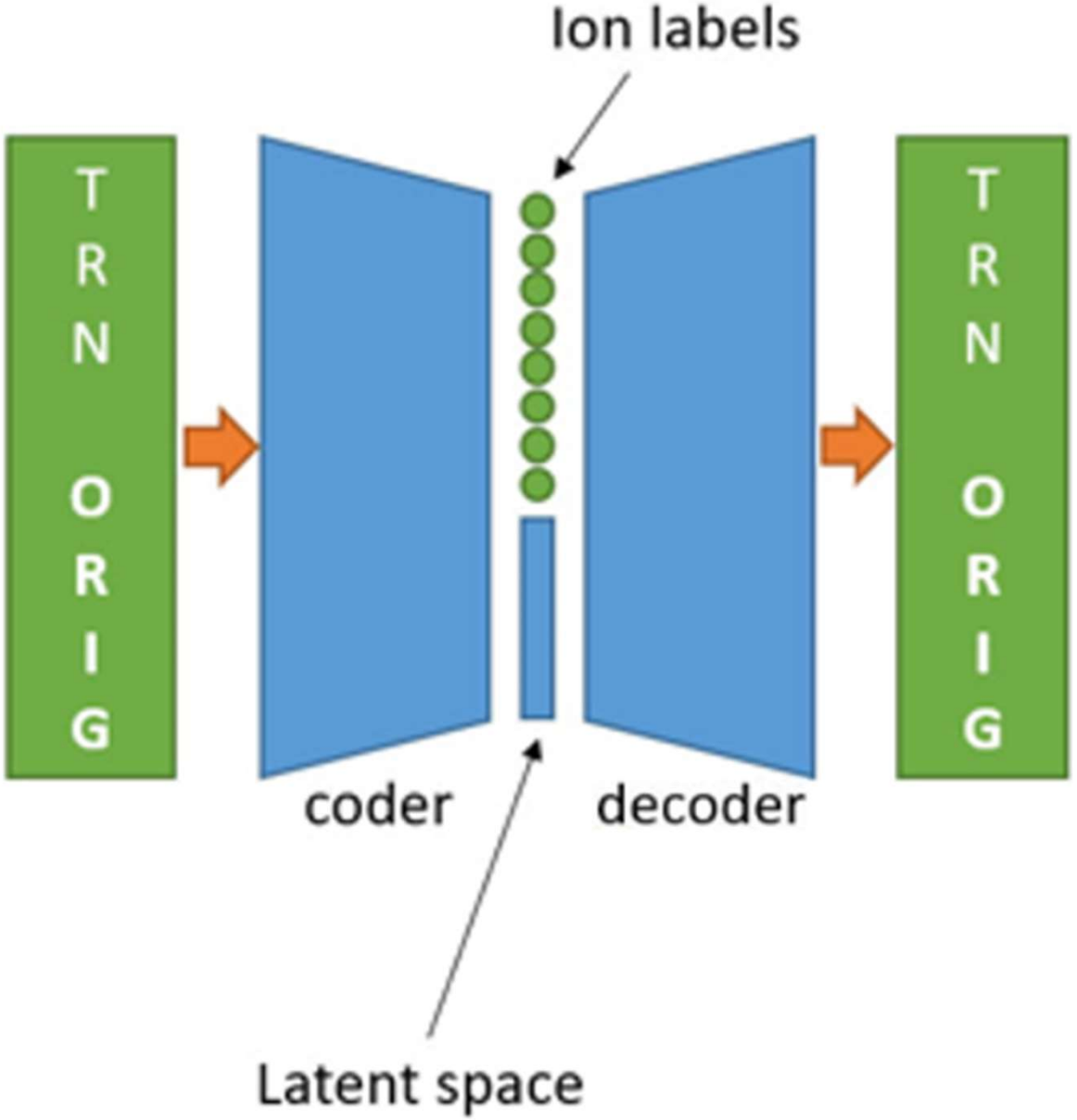
4) Determination of the ions concentrations in generated patterns using the trained regression models

Experimental pipeline: Variational Autoencoder (VAE)



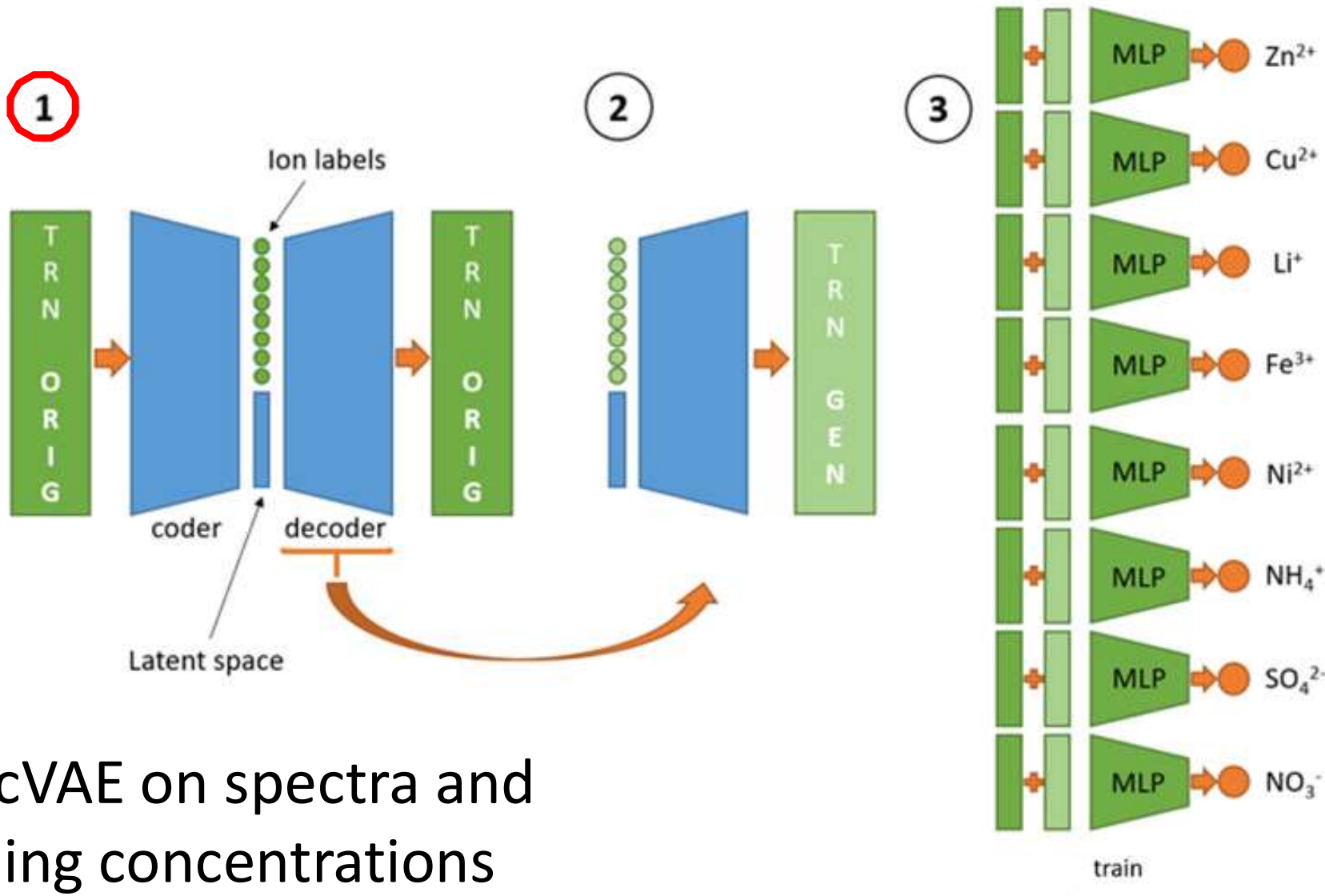
5) Training regression models on an extended training set of optical absorption spectra

Conditional Variational Autoencoder (cVAE)



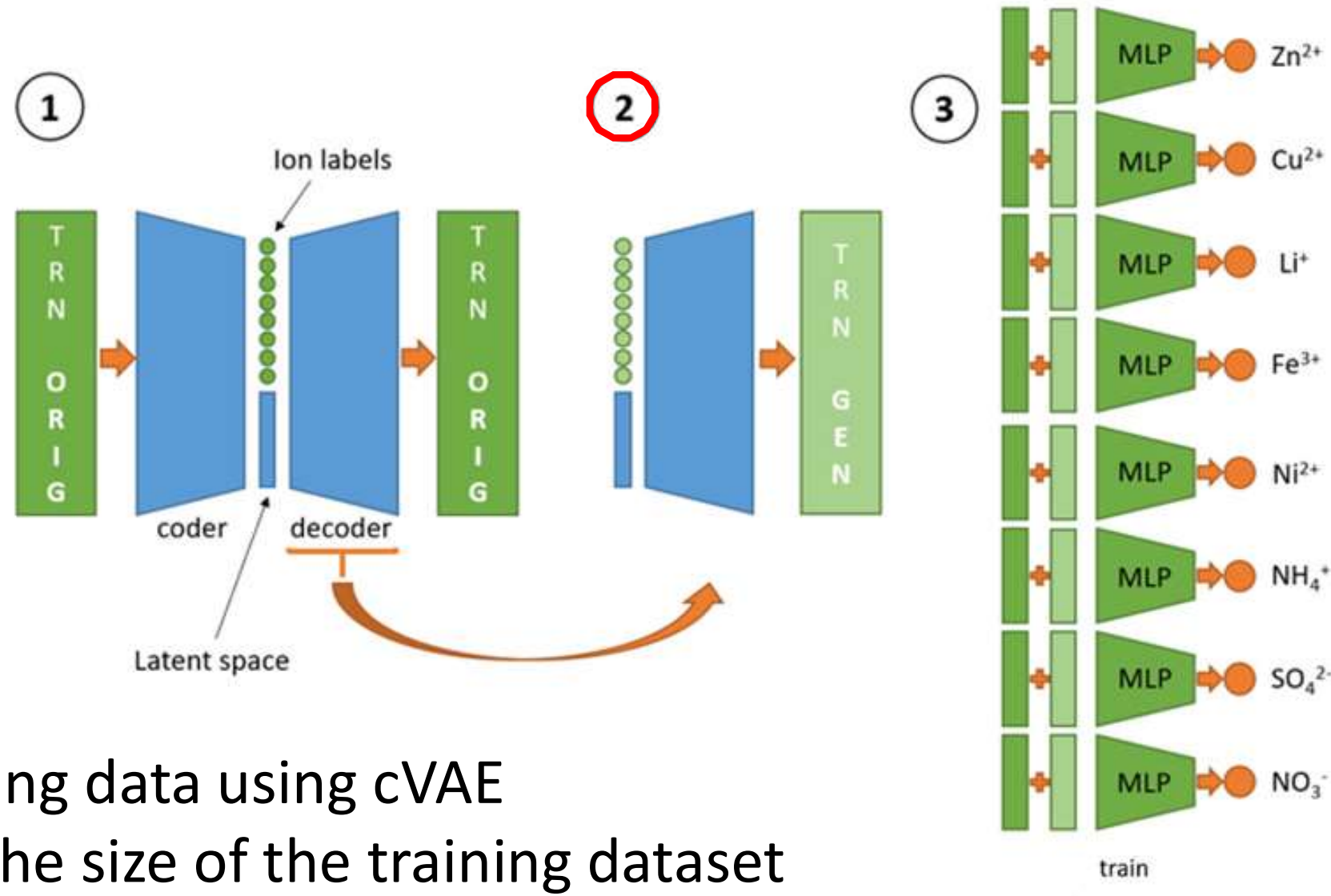
- Very similar to VAE
- The key difference: decoder receives both the spectral information and the corresponding sets of concentrations as inputs
- Allows generating spectra with specific desired concentration values

Experimental pipeline: Conditional Variational Autoencoder (cVAE)



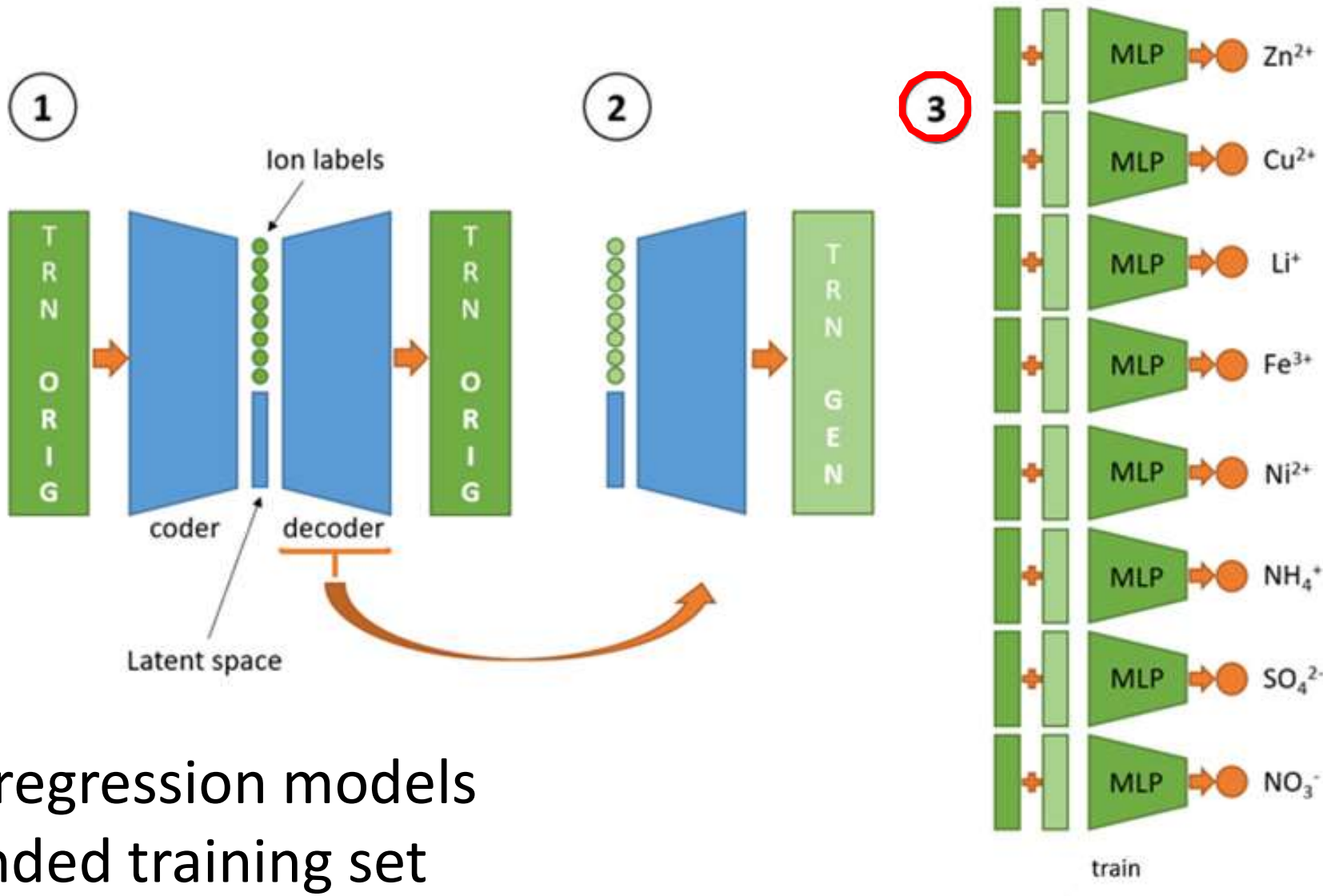
1) Training cVAE on spectra and corresponding concentrations

Experimental pipeline: Conditional Variational Autoencoder (cVAE)



2) Generating data using cVAE
to double the size of the training dataset

Experimental pipeline: Conditional Variational Autoencoder (cVAE)



3) Training regression models on an extended training set of optical absorption spectra

Parameters of experiments

- Data
  - 3744 spectra
  - 911 channels
  - 8 ions
- Experiments
  - 8-fold cross-validation
  - Adam, lr=0.001, 100 epochs
- Neural Networks
  - ❑Regression Neural Network for each ion
    - 2 hidden layers (64 and 16 neurons)
    - 1 output

❑VAE

Coder: MLP
- 911 neurons in the input layer
- 256 neurons in the hidden layer
- 2*91 neurons in the output layer

Decoder: MLP
- 91 neurons in the input layer
- 256 neurons in the hidden layer
- 911 neurons in the output layer

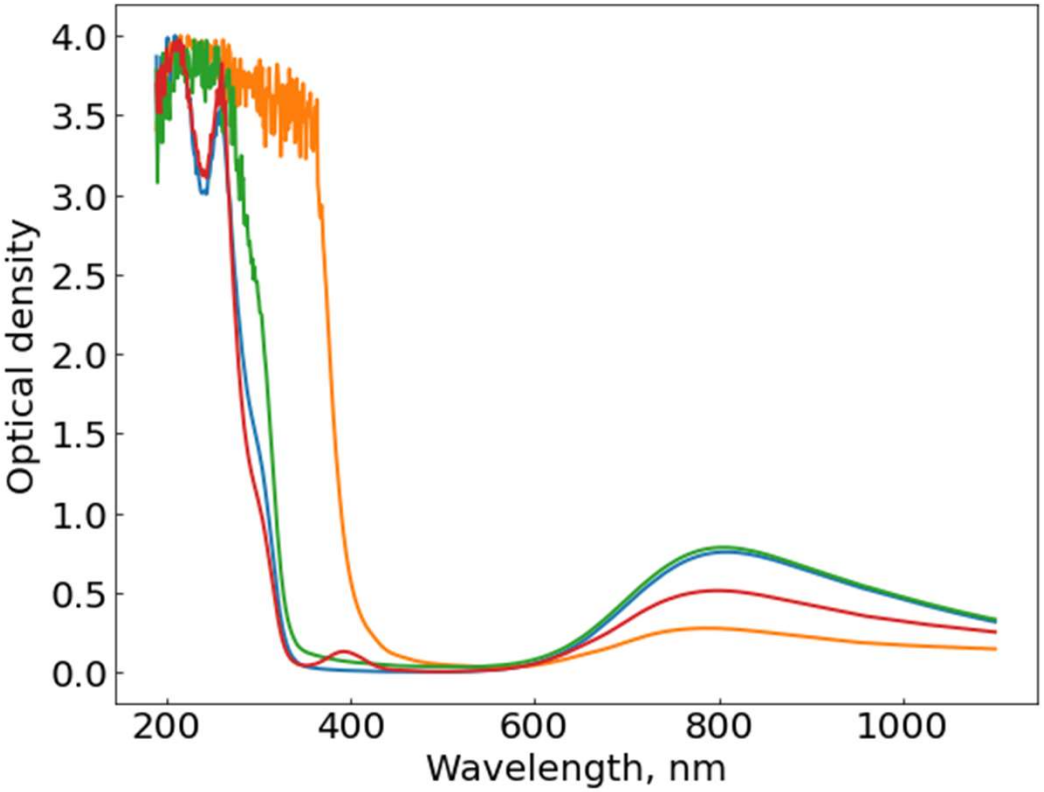
❑cVAE

Coder: MLP
- 911 neurons in the input layer
- 256 neurons in the hidden layer
- 2*91 neurons in the output layer

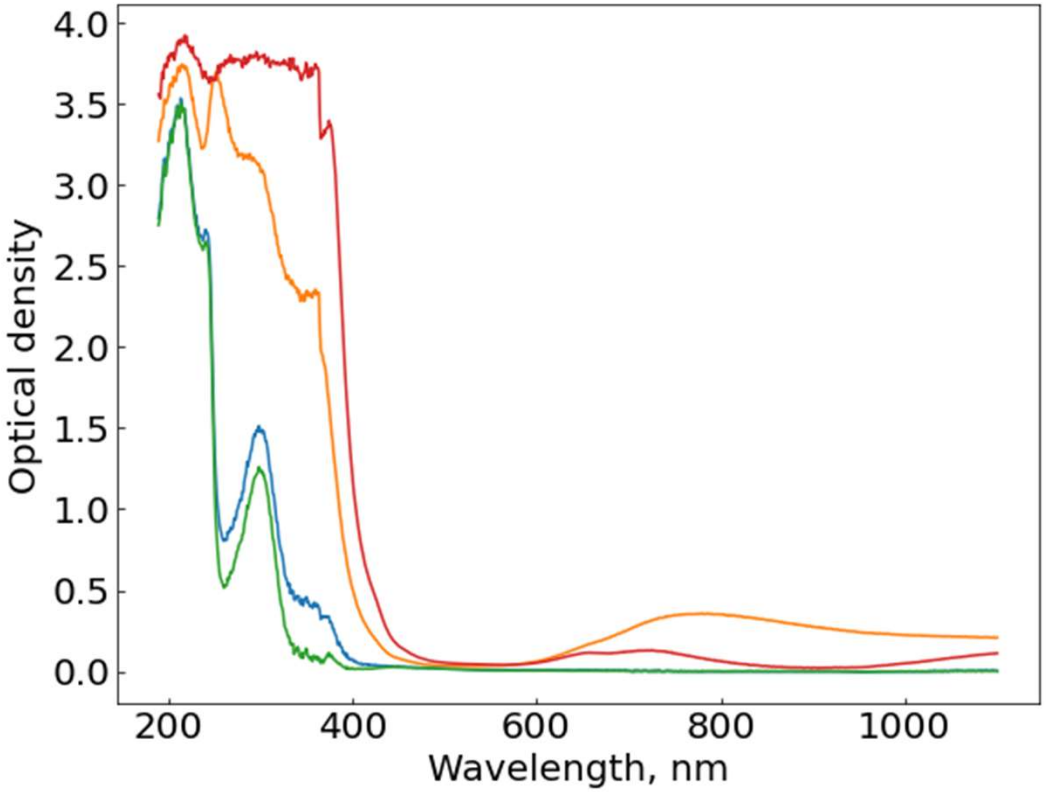
Decoder: MLP
- 91+8 neurons in the input layer
- 256 neurons in the hidden layer
- 911 neurons in the output layer

VAE-generated spectra. Analysis

Experimental spectra

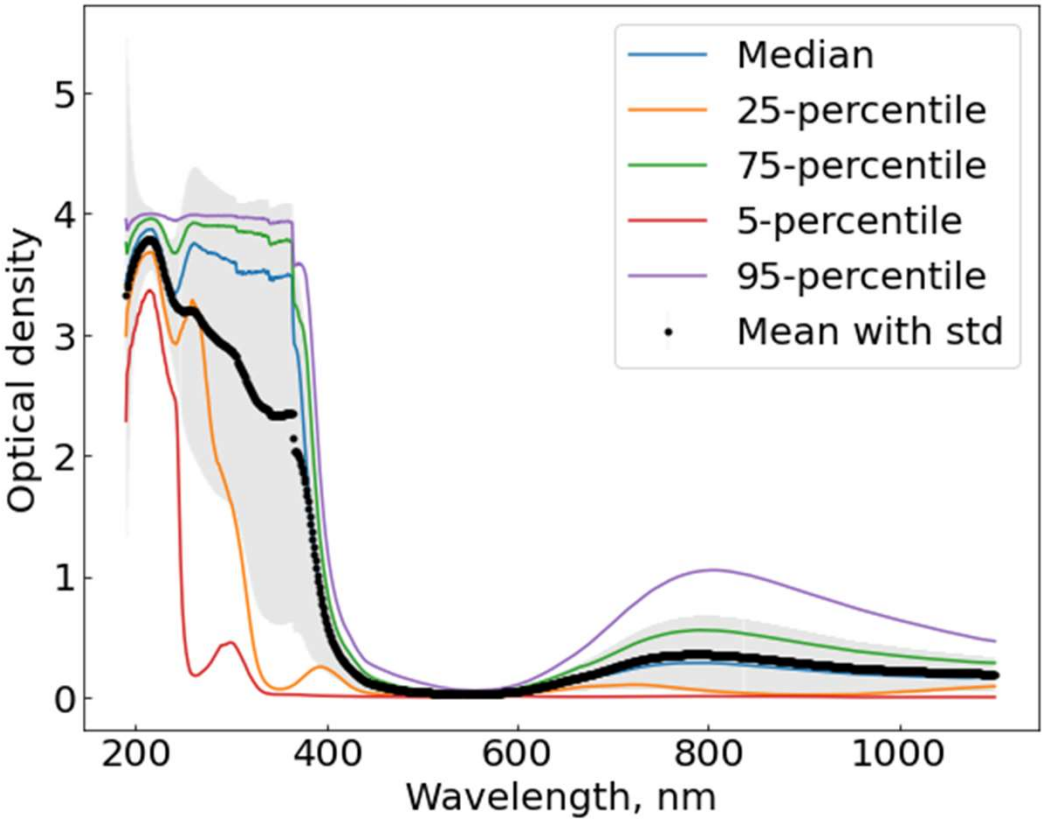


VAE-generated spectra

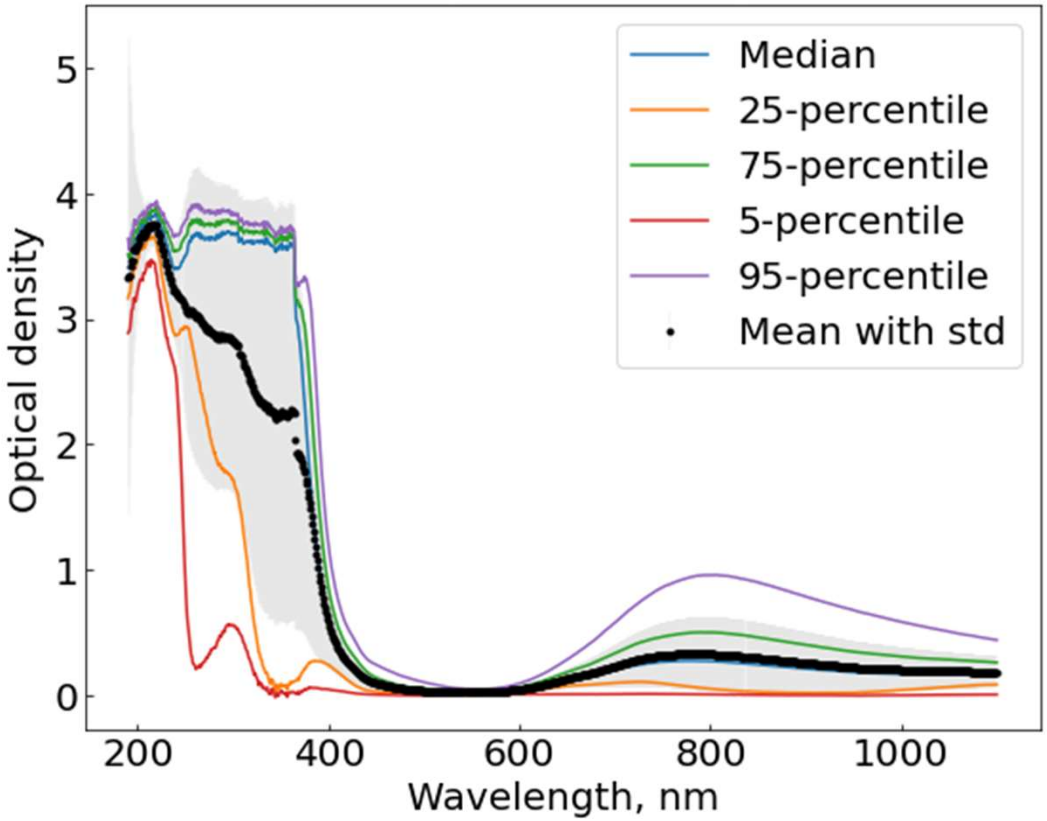


Statistics by channels. VAE

Experimental spectra

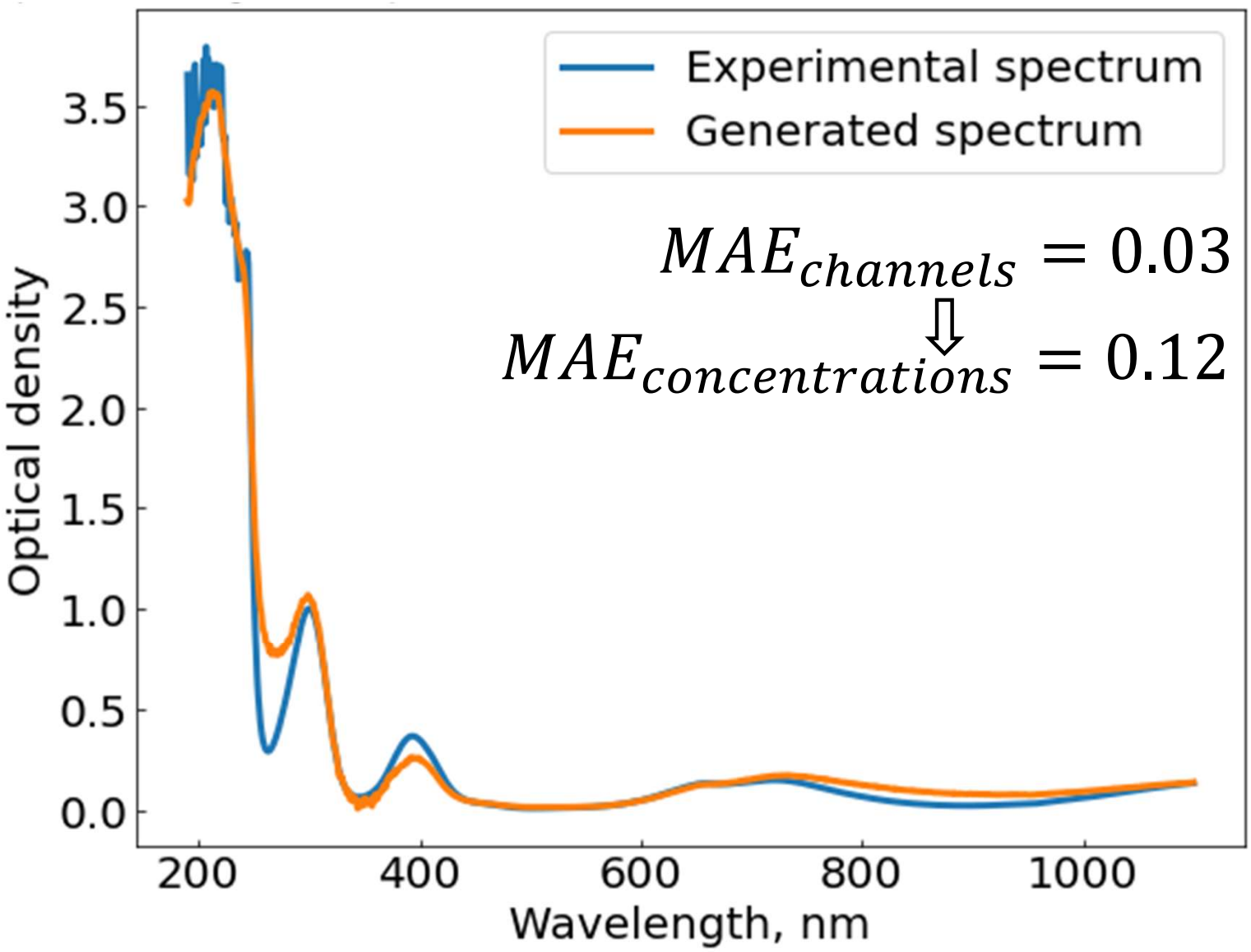


VAE-generated spectra



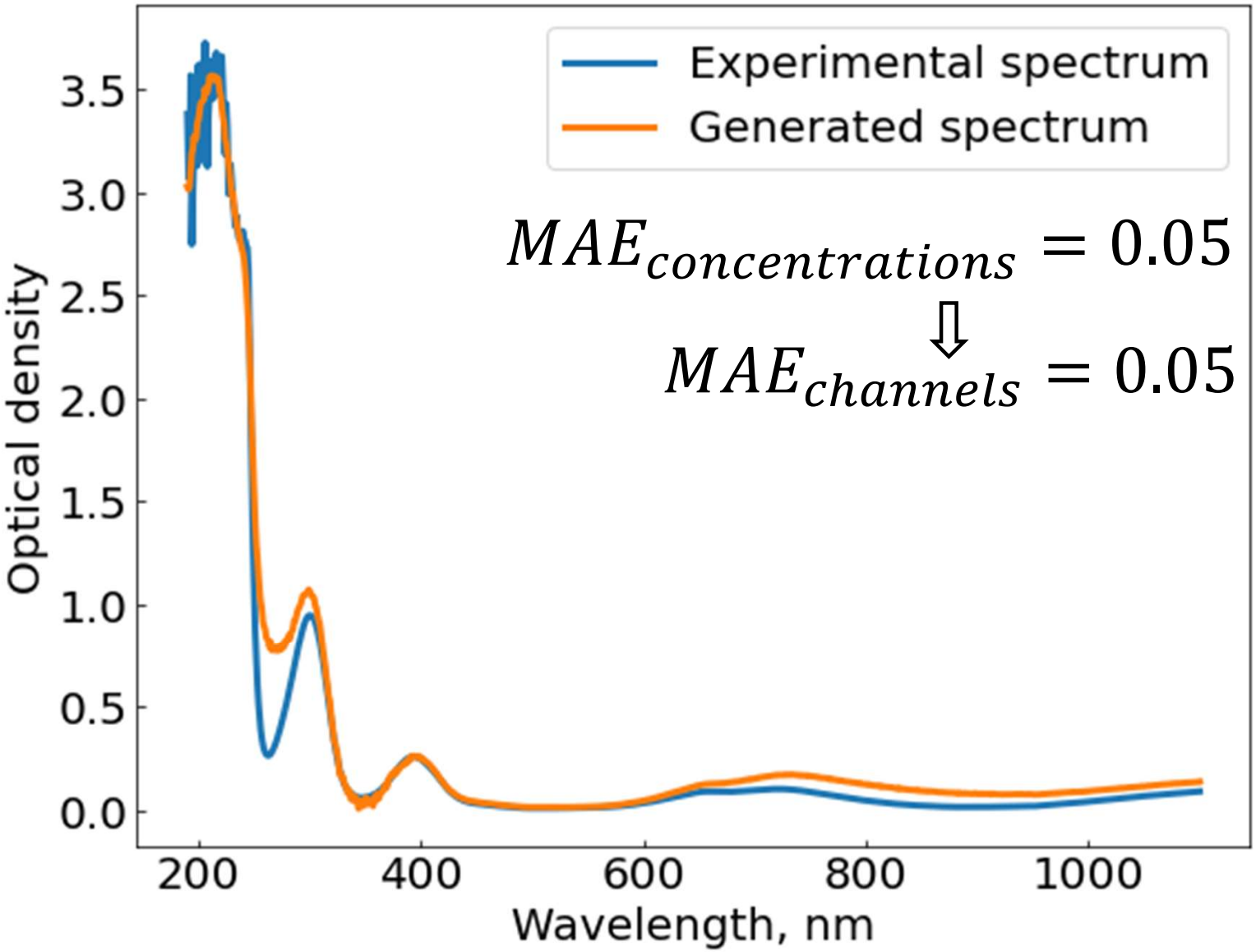
Similarity analysis. VAE

Comparison of experimental and VAE-generated spectra, similar in the channel space



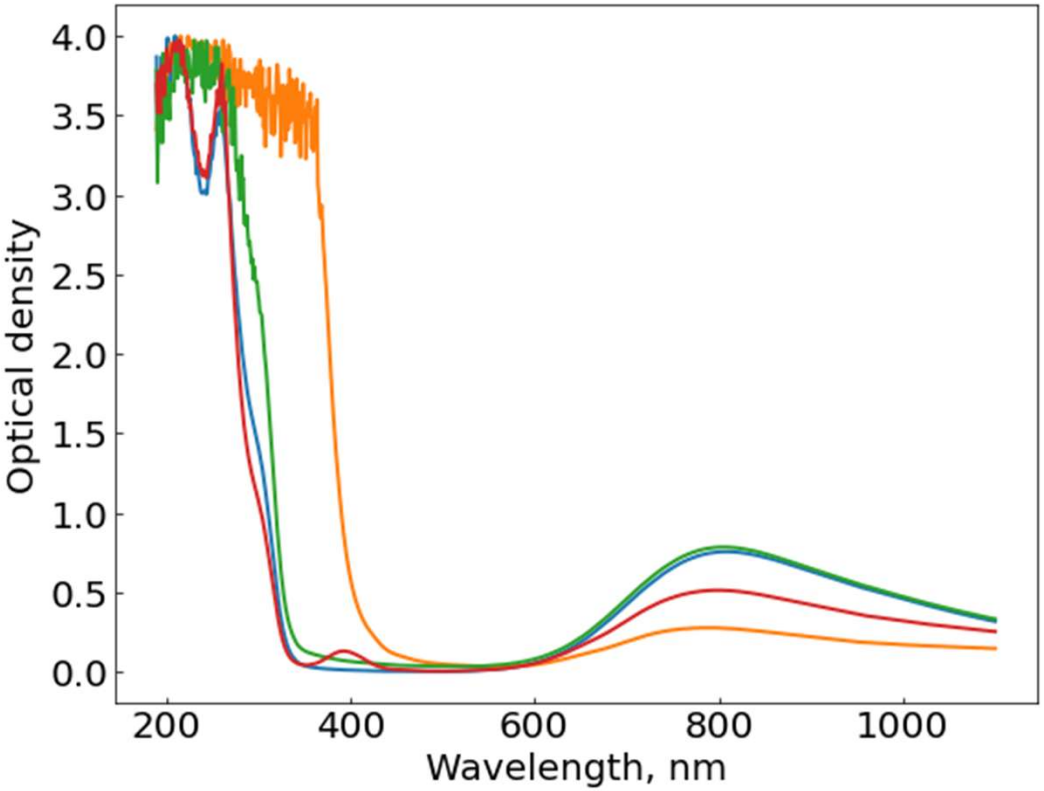
Similarity analysis. VAE

Comparison of experimental and VAE-generated spectra with similar concentrations

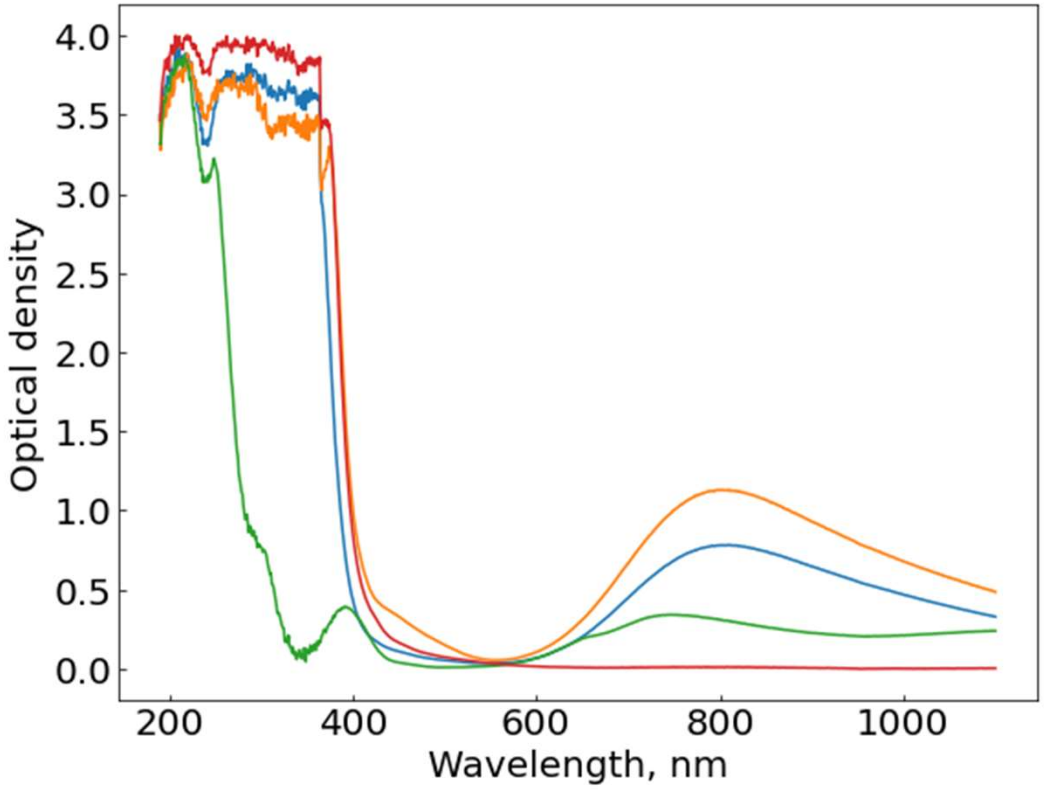


cVAE-generated spectra. Analysis

Experimental spectra

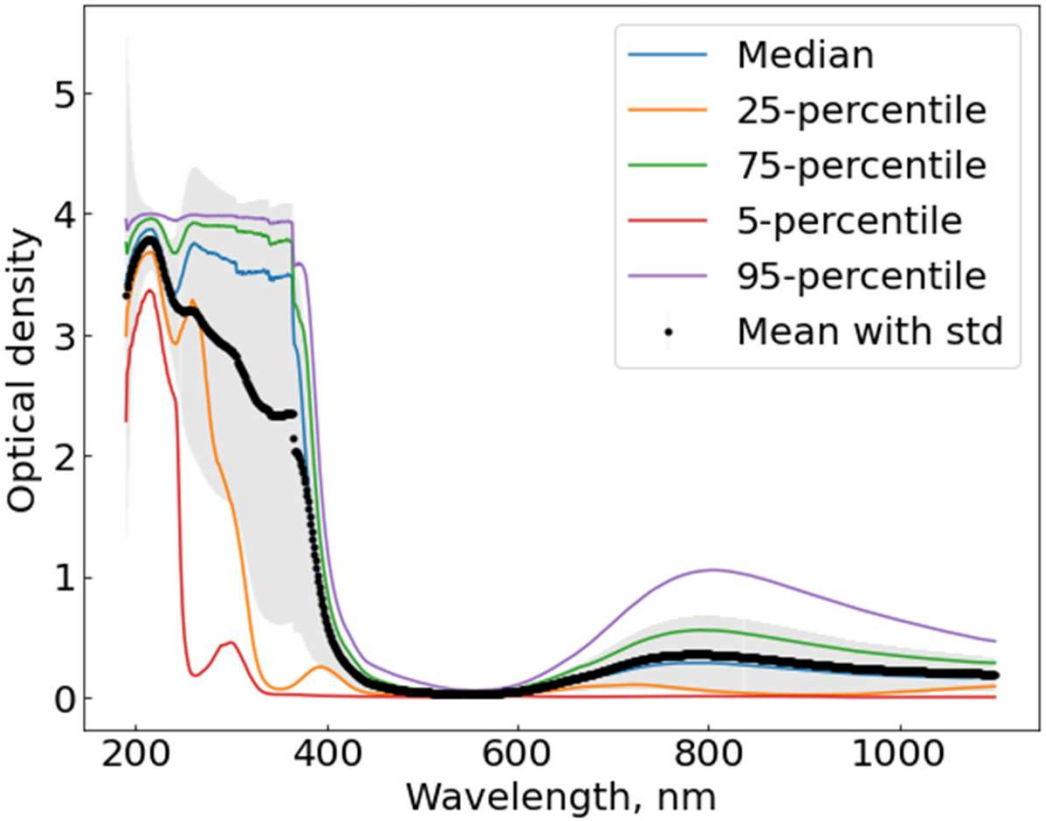


cVAE-generated spectra

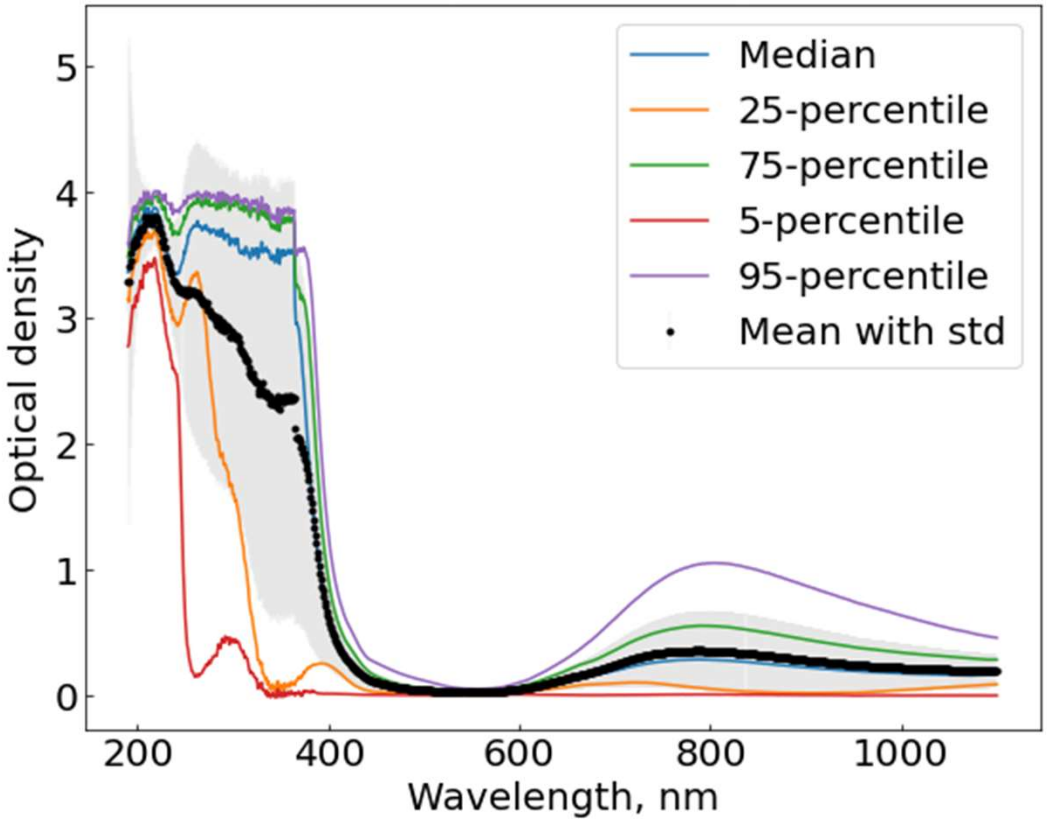


Statistics by channels. cVAE

Experimental spectra

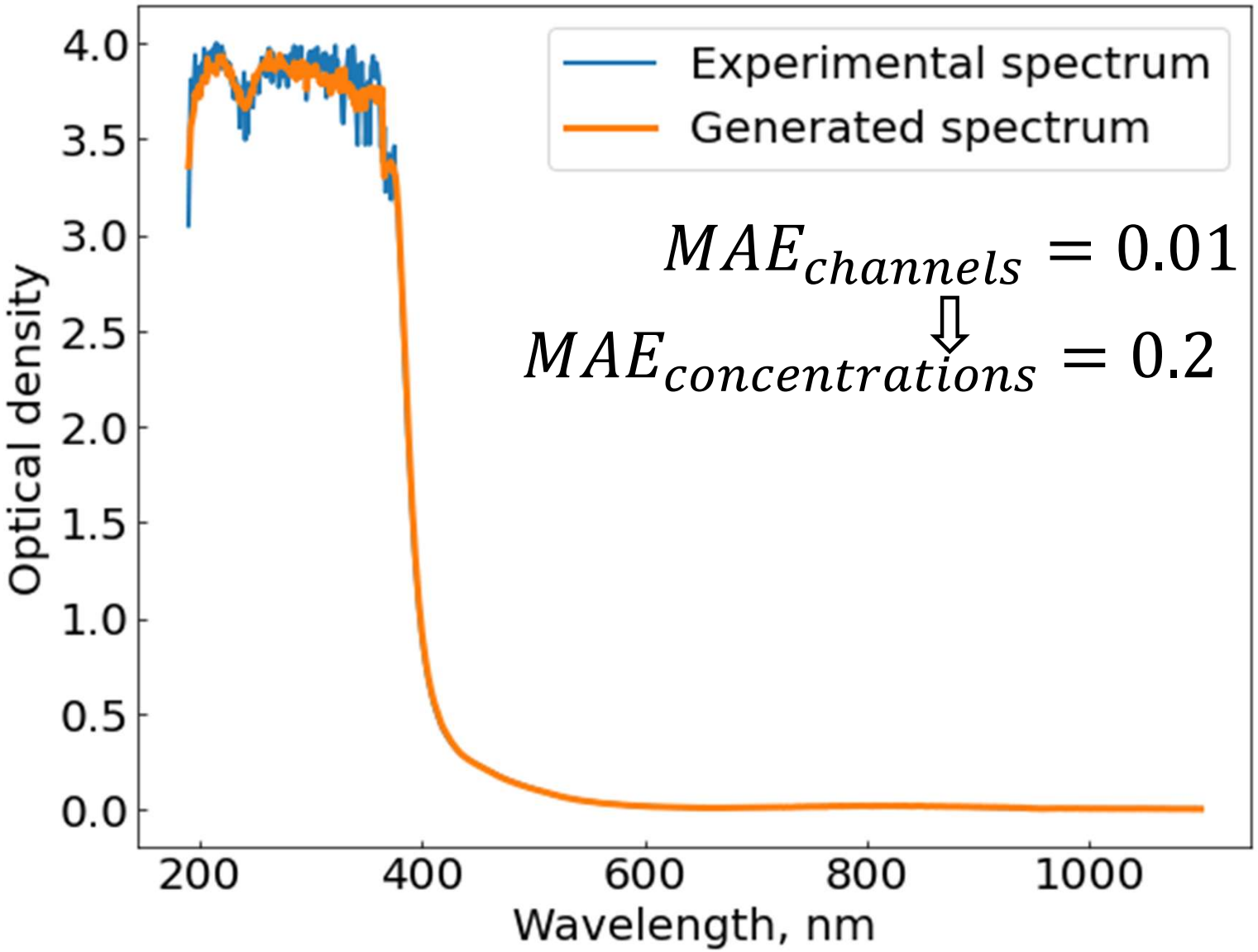


cVAE-generated spectra



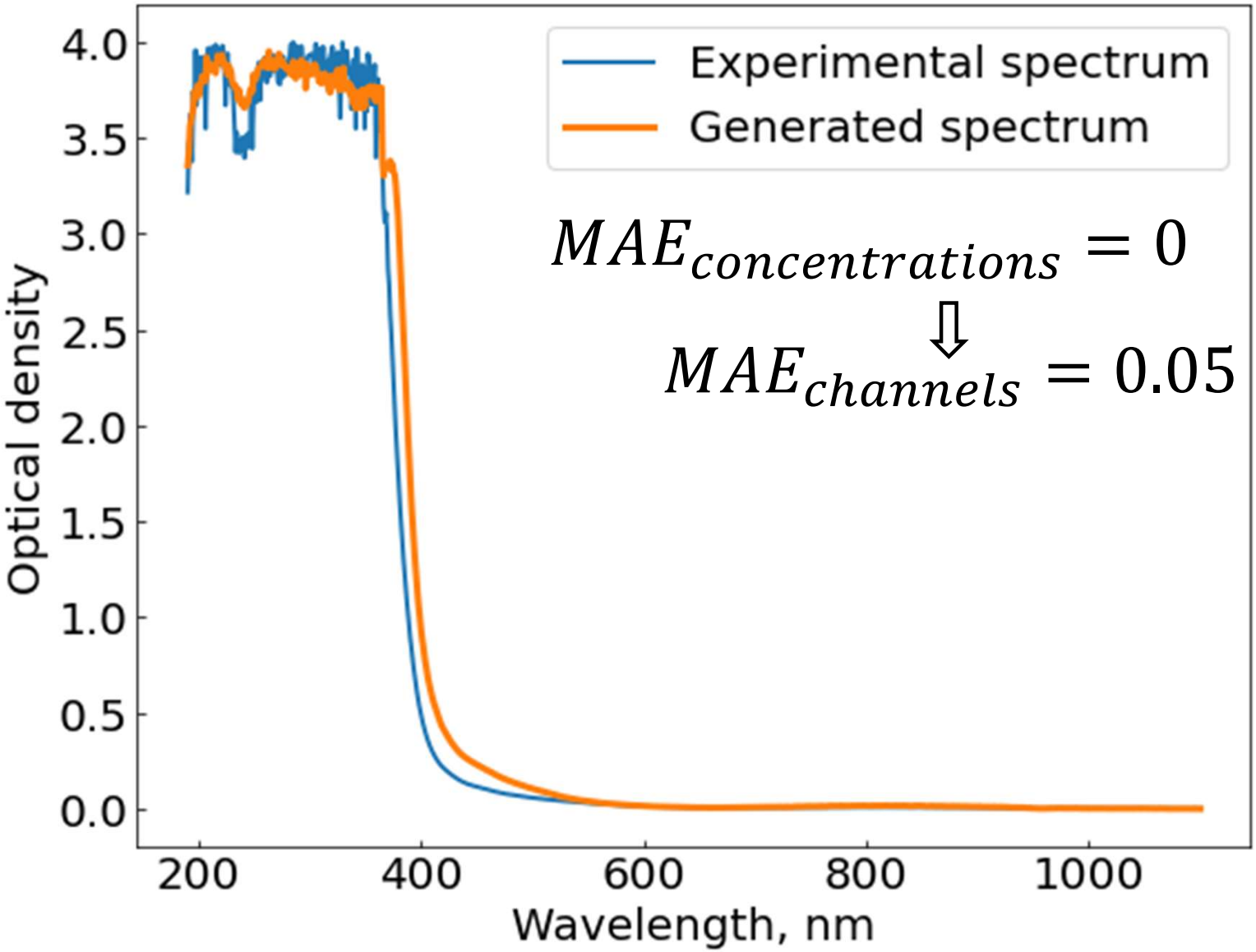
Similarity analysis. cVAE

Comparison of experimental and cVAE-generated spectra, similar in the channel space



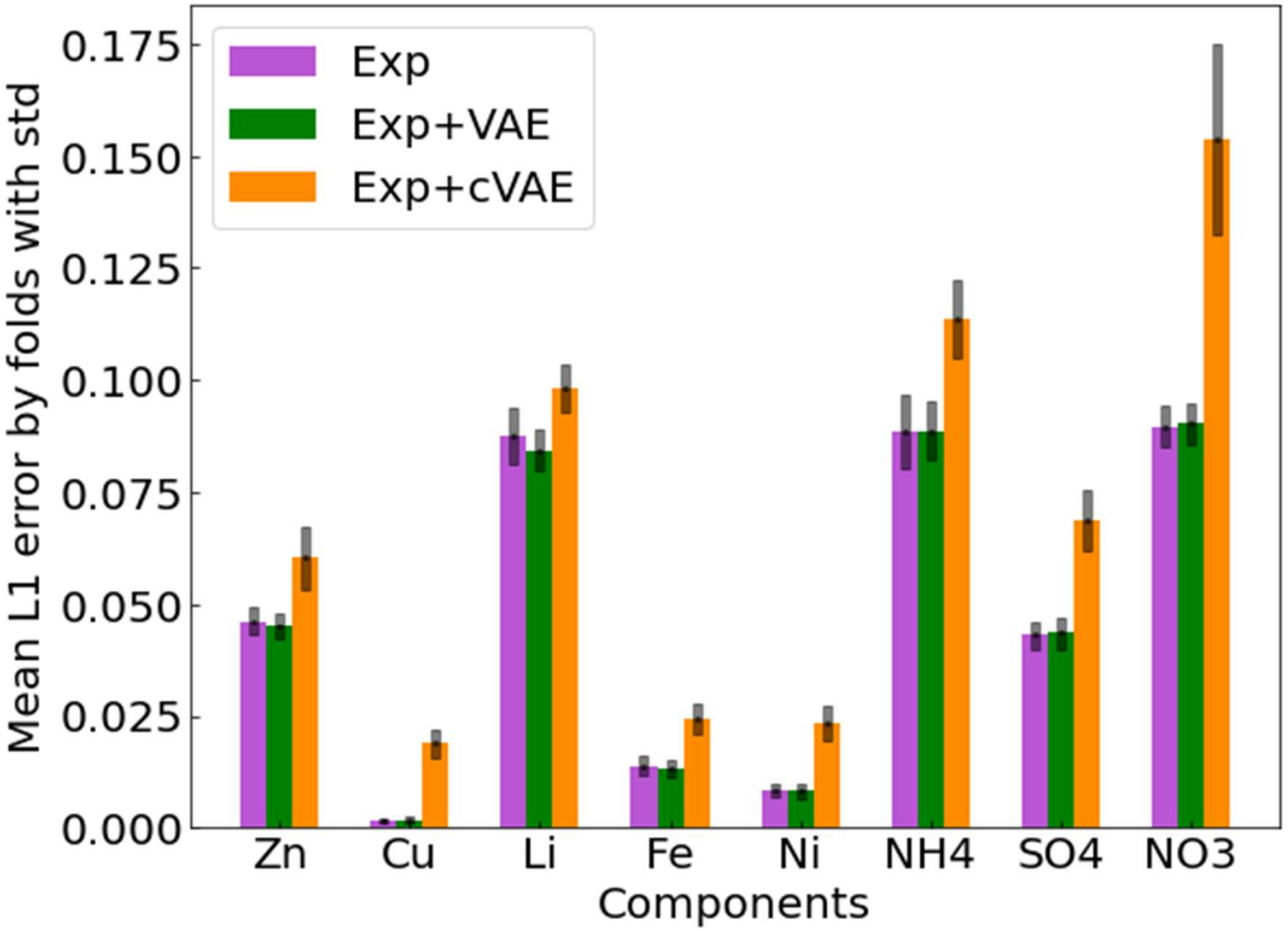
Similarity analysis. cVAE

Comparison of experimental and cVAE-generated spectra with similar concentrations



Performance

- Extension of the training set with patterns generated by Variational Autoencoder (VAE), did not result in a decline in solution quality
- Using the cVAE for the same purpose increases errors



Conclusions. Discussion

- Conclusions
  - With VAE it is possible to generate patterns that effectively mimic experimental spectra while still differing from them
  - cVAE fails to generate valid samples, which may be due to the strategy for selecting concentration sets.

- Possible approaches in generation
  - Uniform distribution in latent space
  - Normal distribution in latent space
    - Generation from the same distribution
    - Generation from inverted distribution (equalization)
- Possible reasons for the hypothetical improvement
  - Noise reduction due to reduction of data dimensionality in latent space
  - Distribution equalization

Thank you for your attention

Generating in the latent space