The 9th International Conference in Deep Learning in Computational Physics

2-4 Июля, 2025

НИИЯФ МГУ, Москва, Россия

Сравнение методов генерации данных с использованием вариационных автоэнкодеров для спектрального анализа

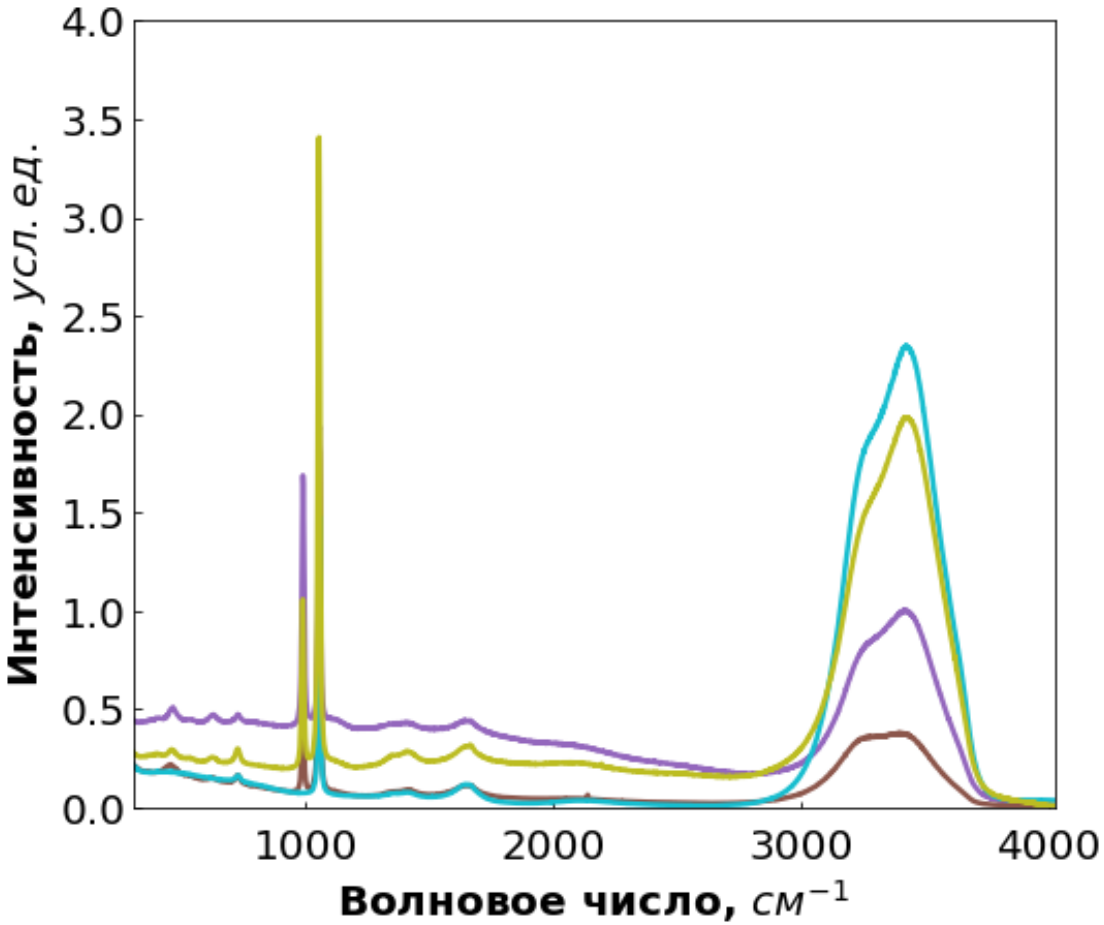
Мущина А.С.[1,2], Исаев И.В.[1], Сарманова О.Е.[1,2], Доленко Т.А.[1,2], Доленко С.А.[1]

[1] НИИ ядерной физики имени Д.В. Скобельцына МГУ имени М.В. Ломоносова,
[2] Физический факультет МГУ имени М.В. Ломоносова

** Исследование выполнено за счёт гранта Российского Научного фонда, проект: https://rscf.ru/en/project/24-11-00266/.*

Обратная задача спектроскопии

– определение концентраций компонентов раствора по спектру.



- Многокомпонентные водные растворы
- Спектры комбинационного рассеяния
- 3760 спектров
- 2598 каналов

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

Специфика ОЗ спектроскопии

- Спектроскопические методы обеспечивают **быстрый дистанционный анализ** и предоставляют **информативные данные, позволяющие идентифицировать состав водной среды**

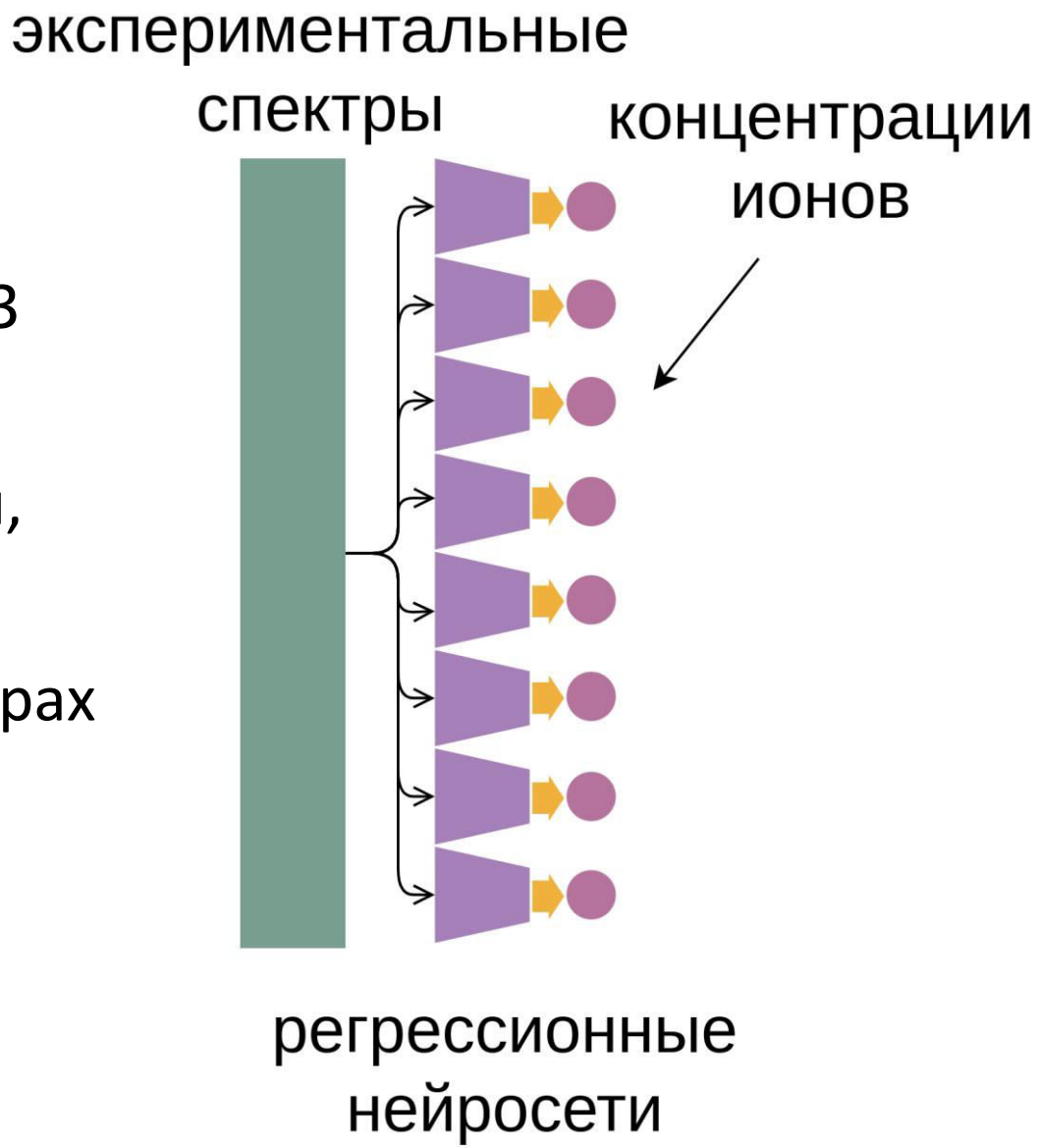
НО

- ОЗ спектроскопии отличается **высокой размерностью и сложностью**, поскольку требует одновременного анализа множества спектральных каналов

Подходящий анализ может быть выполнен с использованием методов машинного обучения (МО), например, нейронных сетей

Референсное решение

В качестве референсного решения ОЗ рассмотрим регрессионные нейросети, обученные на экспериментальных спектрах для каждого иона



Проблемы получения спектроскопических наборов данных для решения ОЗ с использованием МО

- Эксперименты **трудозатратные, дорогостоящие и занимают много времени**
- Требуется **специальное оборудование**
- Необходимы **квалифицированные специалисты**
- Для репрезентативности нужен **большой объём данных**

Кроме того:

Стандартные методы аугментации данных **не вполне соответствуют специфическим требованиям**, предъявляемым спектроскопическими данными

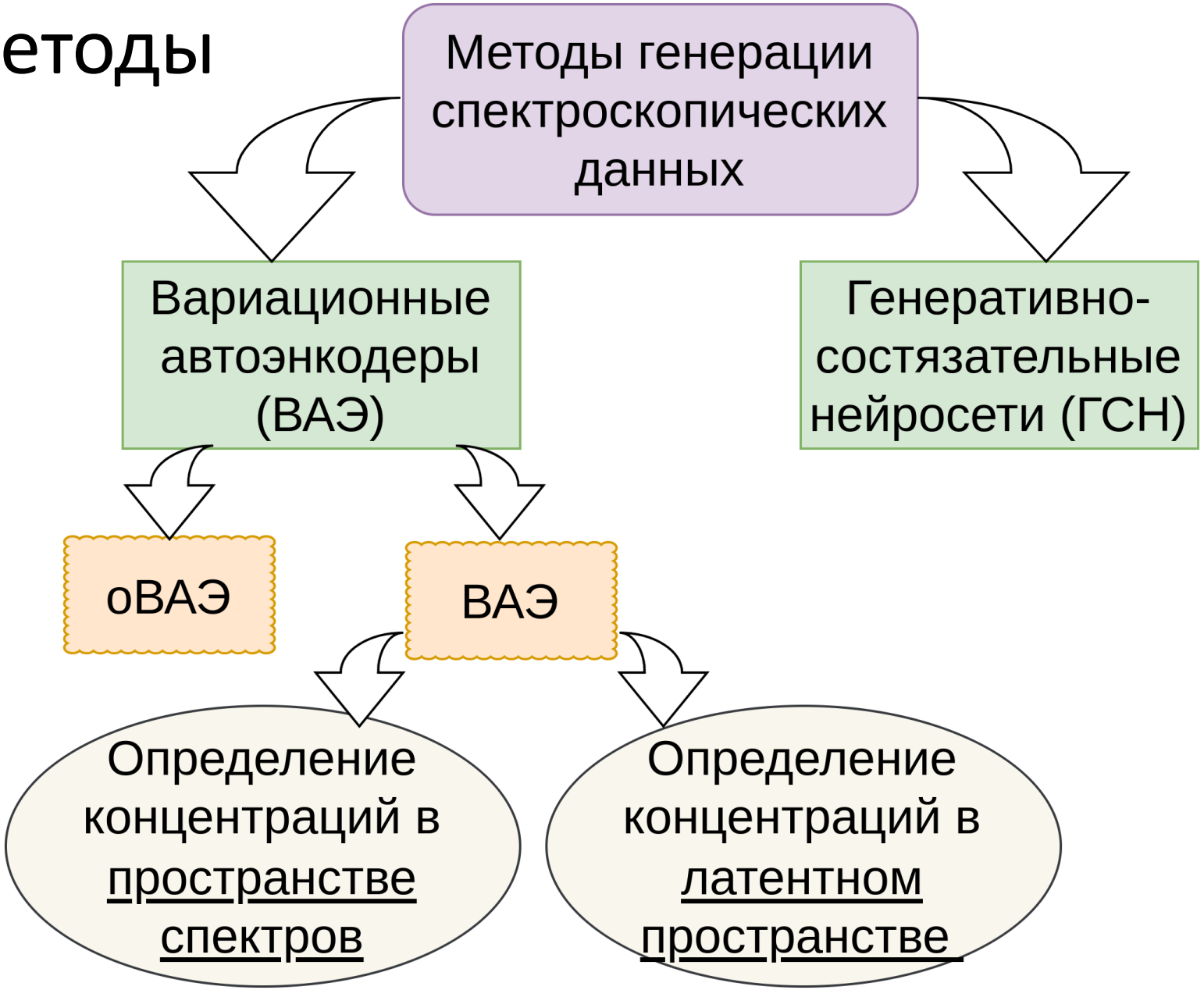
Мотивация

- Проблемы набора спектроскопических данных для методов МО
- Ограничения стандартных техник аугментации данных

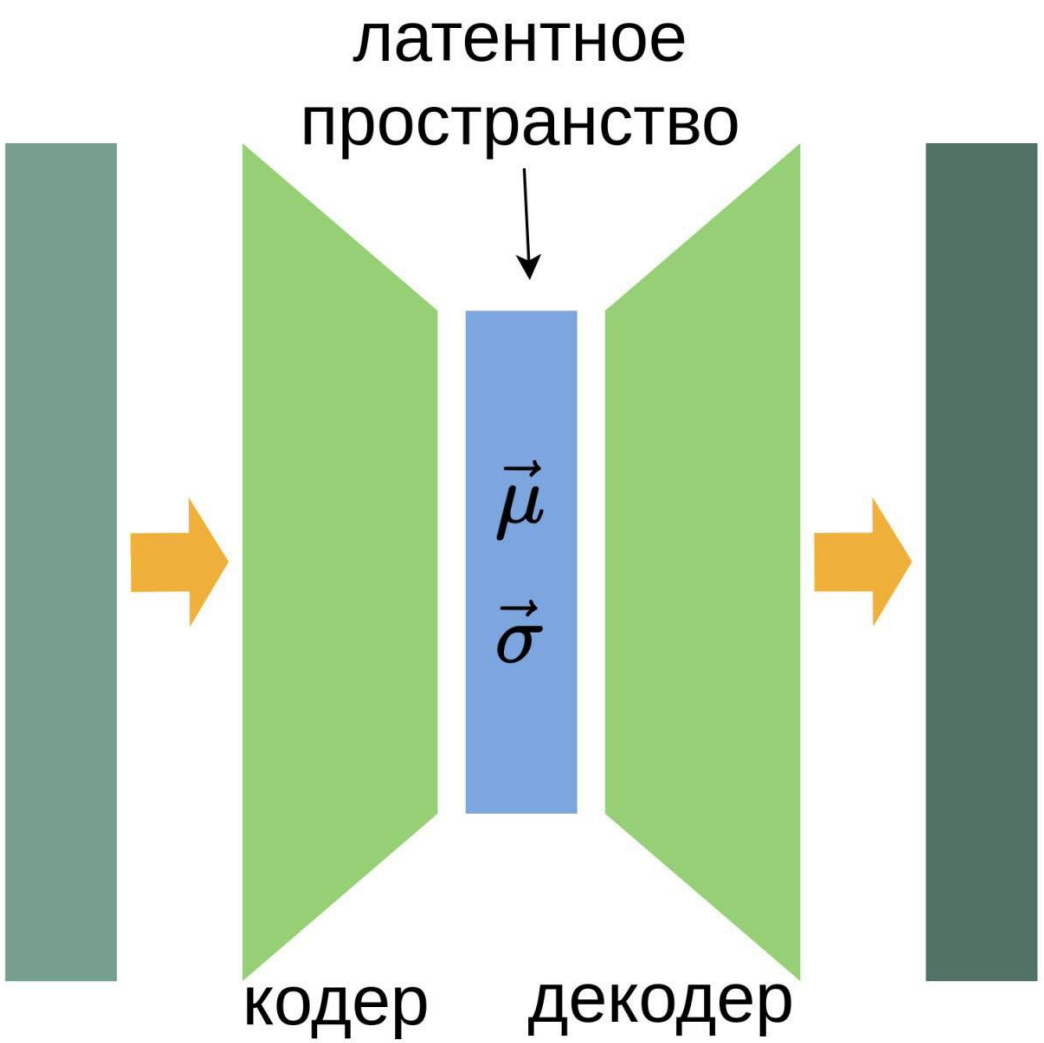
Цель

Улучшить репрезентативность
спектроскопического набора данных
с использованием
генеративных нейросетевых подходов

Генеративные AI-методы

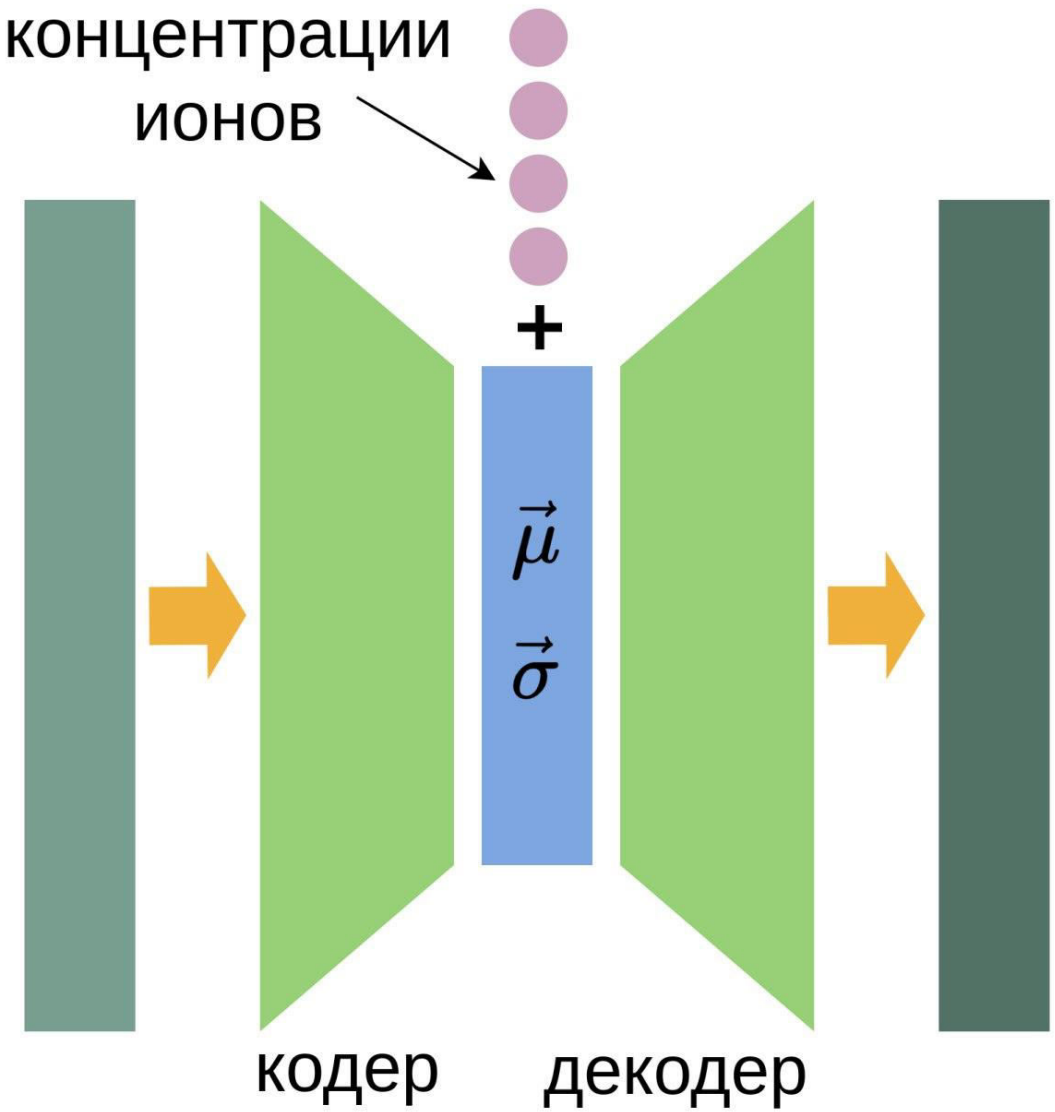


Вариационный автоэнкодер (ВАЭ)



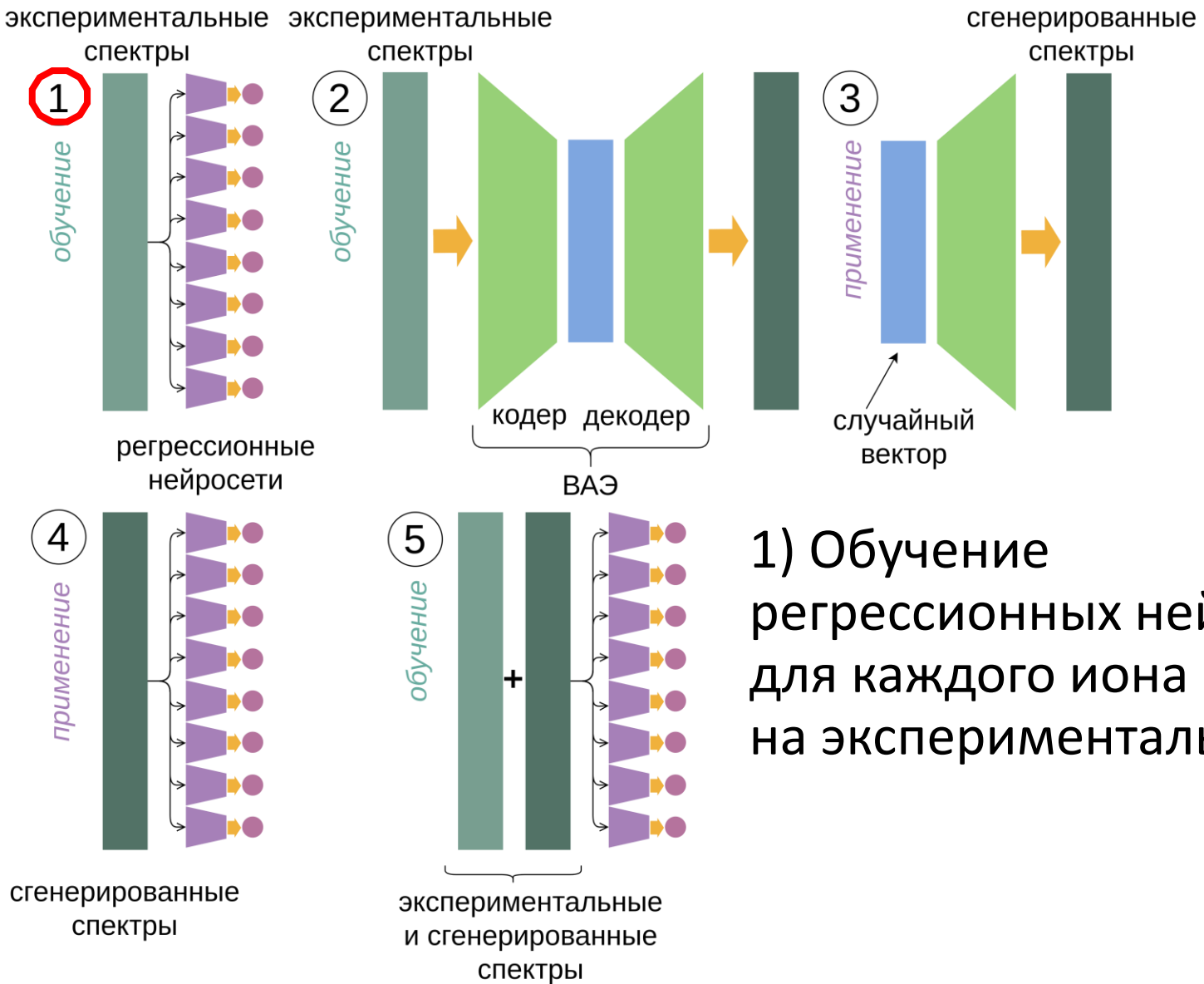
- Сжимает входной сигнал в представление меньшей размерности в латентном пространстве
- Информация о данных в латентном пространстве представляется в виде параметров многомерного распределения, чаще всего нормального
- Позволяет генерировать примеры путём декодирования случайных векторов, сэмплированных из этого распределения
- $Loss = MSE + D_{KL}$

Обусловленный вариационный автоэнкодер (оВАЭ)



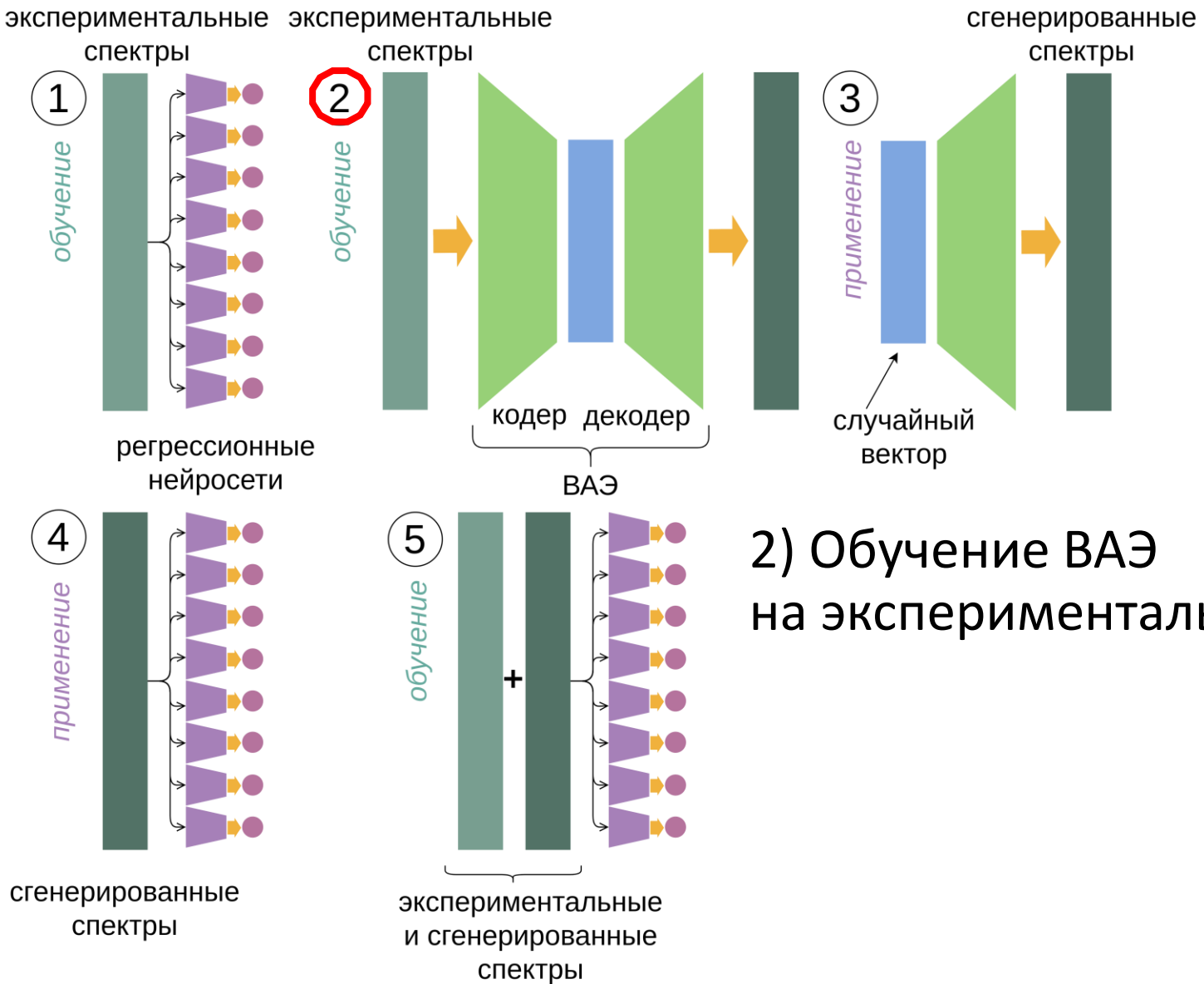
- Ключевое отличие от ВАЭ: декодер получает на вход как спектральную информацию, так и соответствующие наборы концентраций
- Позволяет генерировать спектры с заданными целевыми значениями концентраций

Вычислительный эксперимент: Вариационный автоэнкодер (ВАЭ)



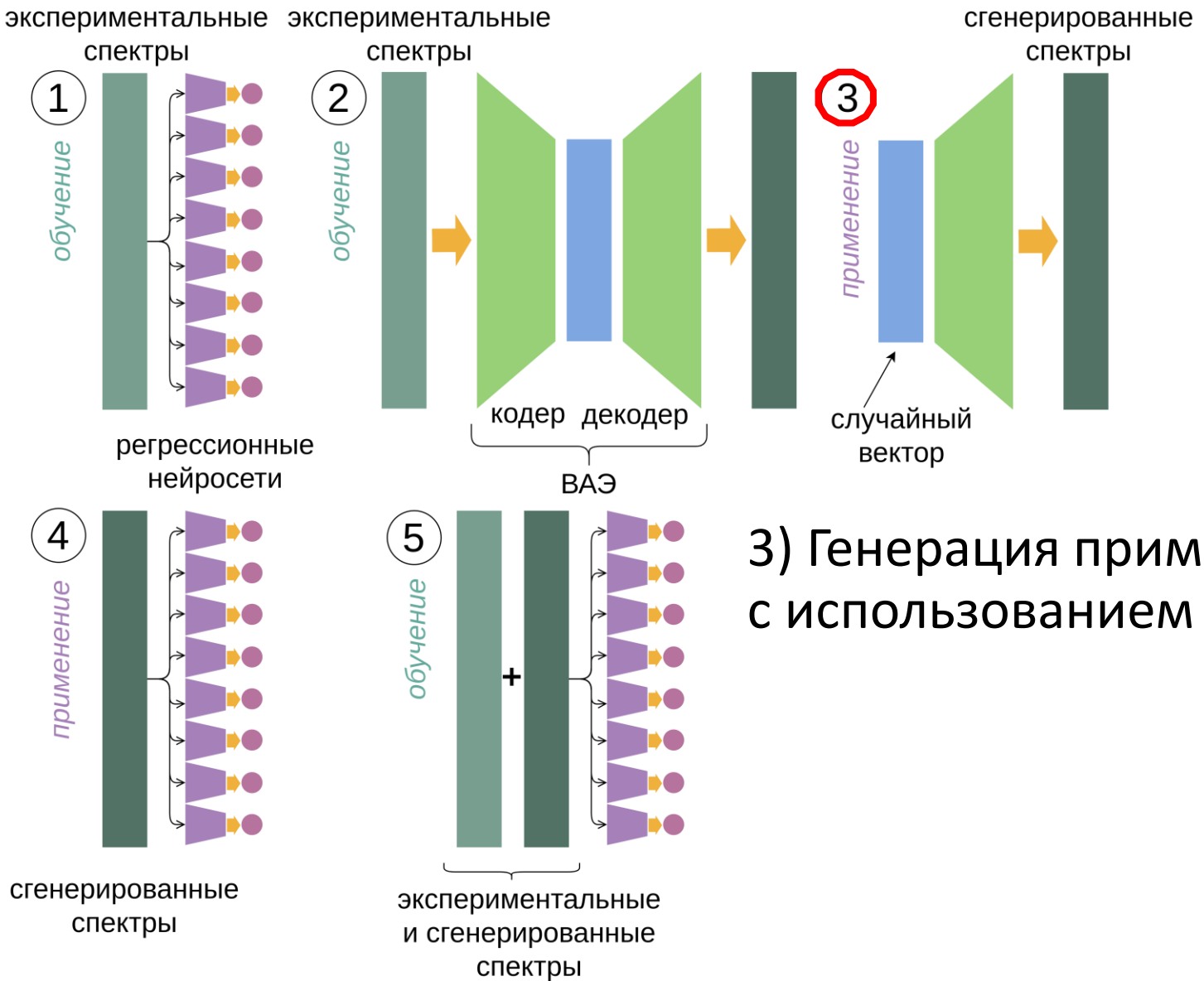
1) Обучение регрессионных нейросетей для каждого иона на экспериментальных данных

Вычислительный эксперимент: Вариационный автоэнкодер (ВАЭ)



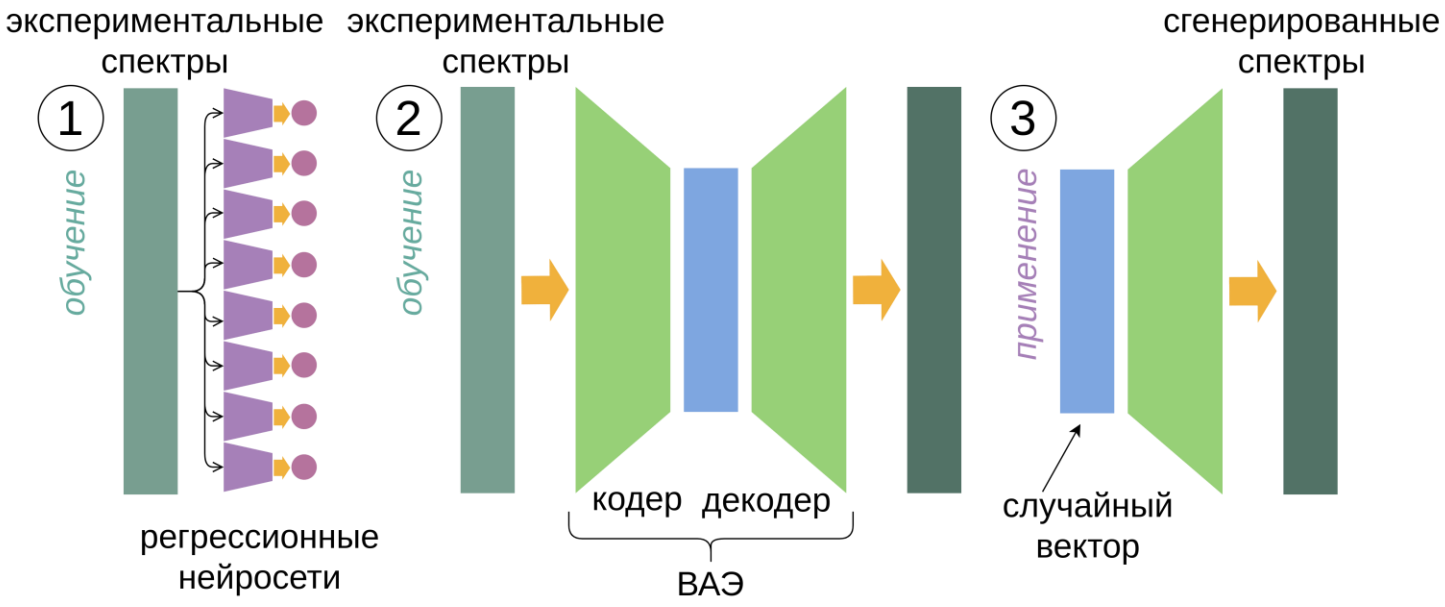
2) Обучение ВАЭ
на экспериментальных данных

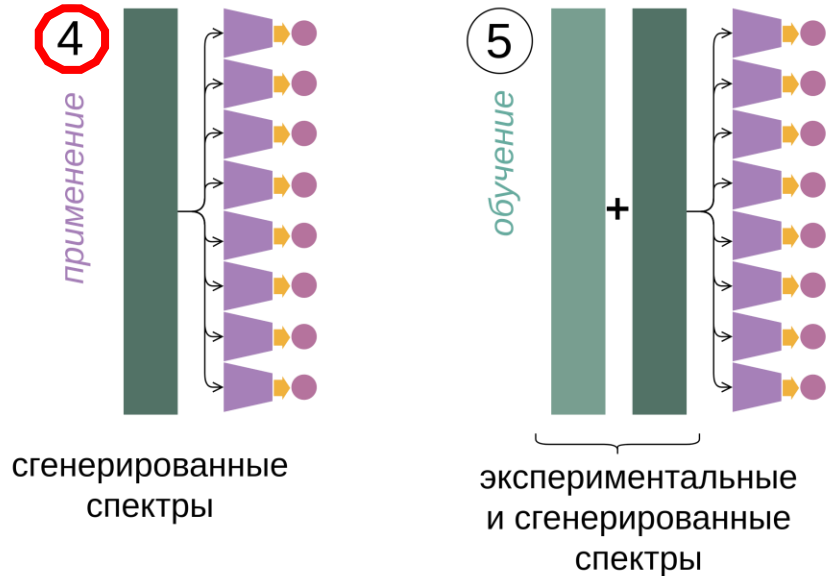
Вычислительный эксперимент: Вариационный автоэнкодер (ВАЭ)



3) Генерация примеров с использованием декодера ВАЭ

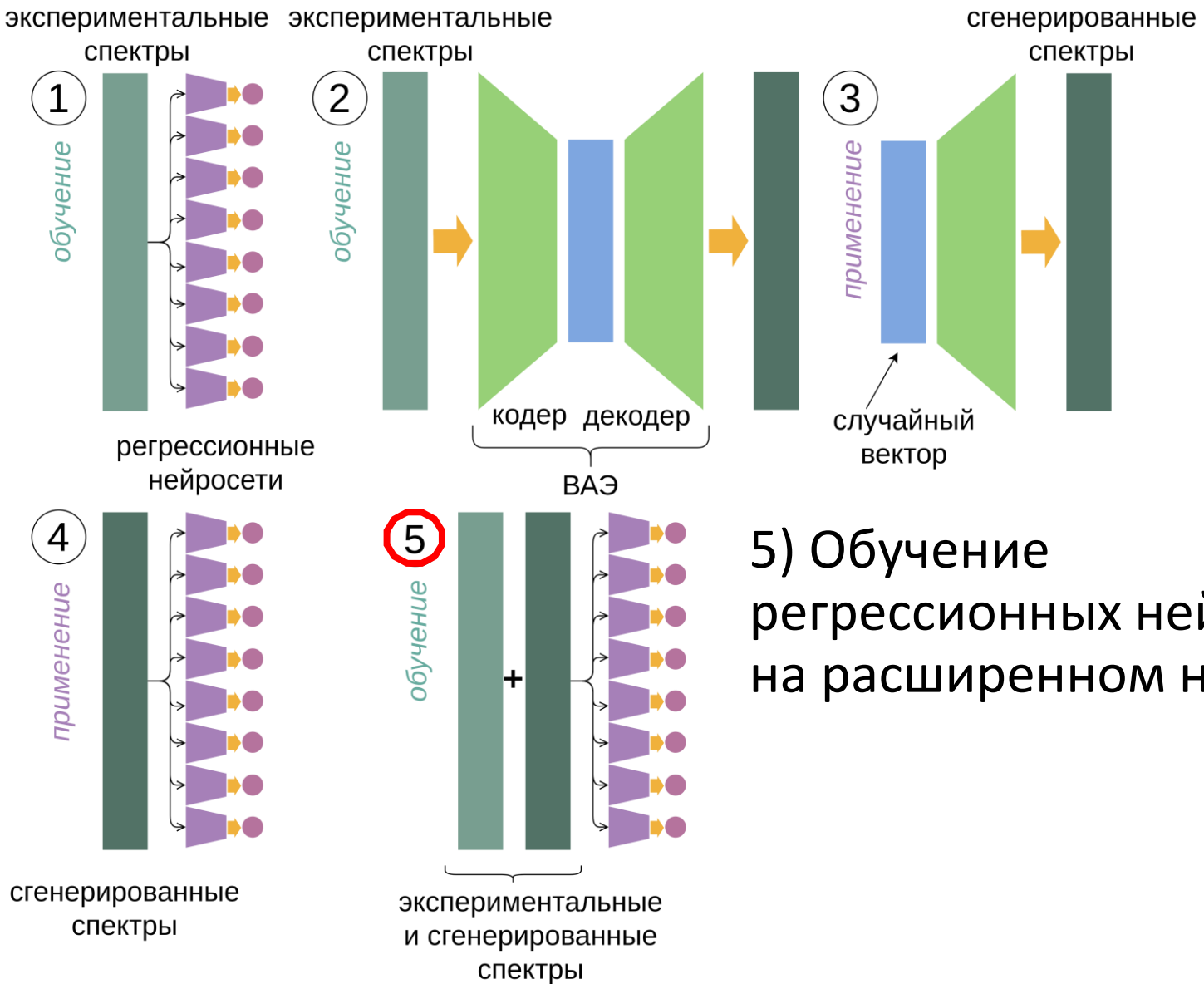
Вычислительный эксперимент: Вариационный автоэнкодер (ВАЭ)





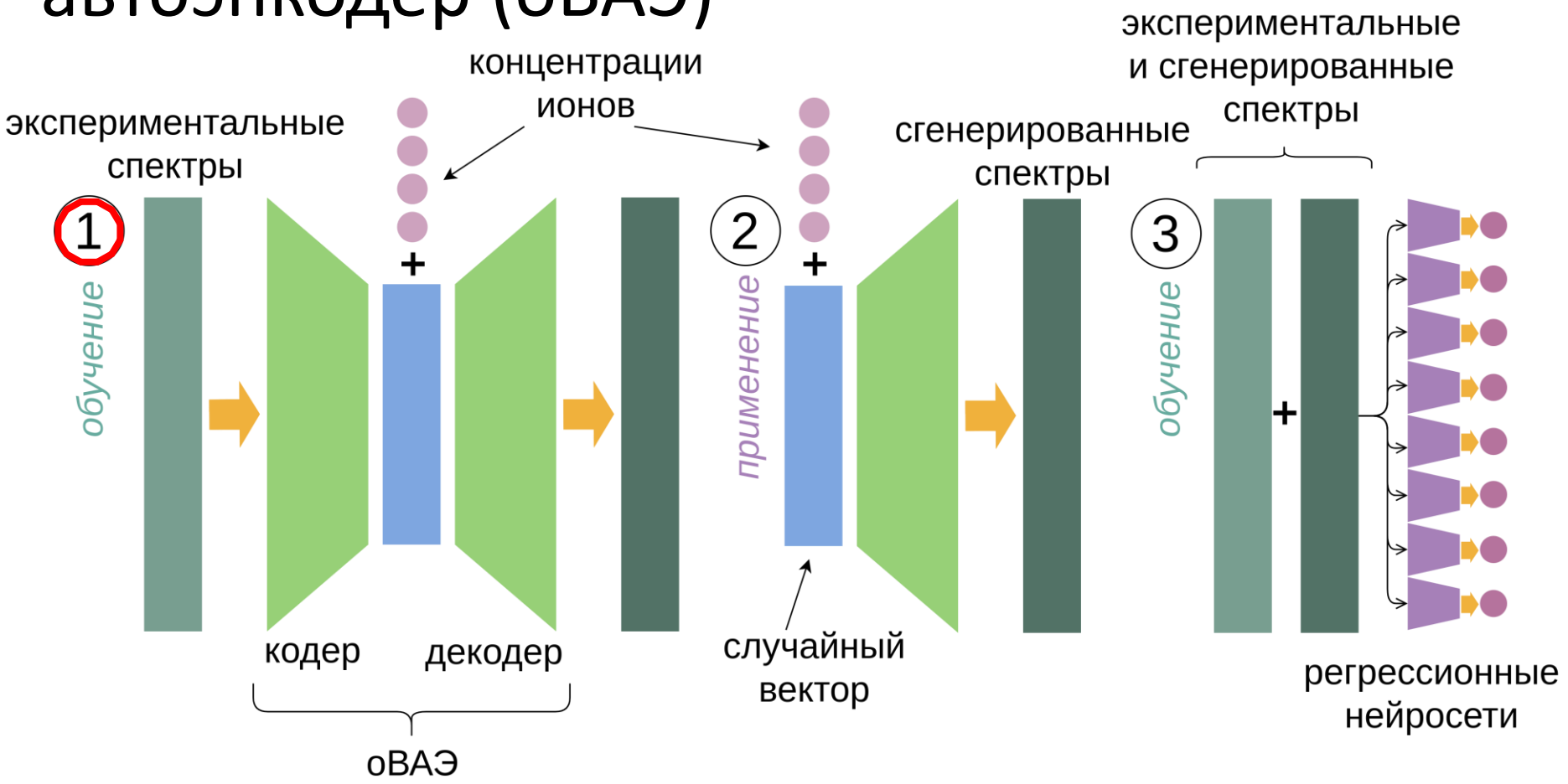
4) Определение концентраций ионов сгенерированных спектров с помощью регрессионных нейросетей, обученных на Шаге 1

Вычислительный эксперимент: Вариационный автоэнкодер (ВАЭ)



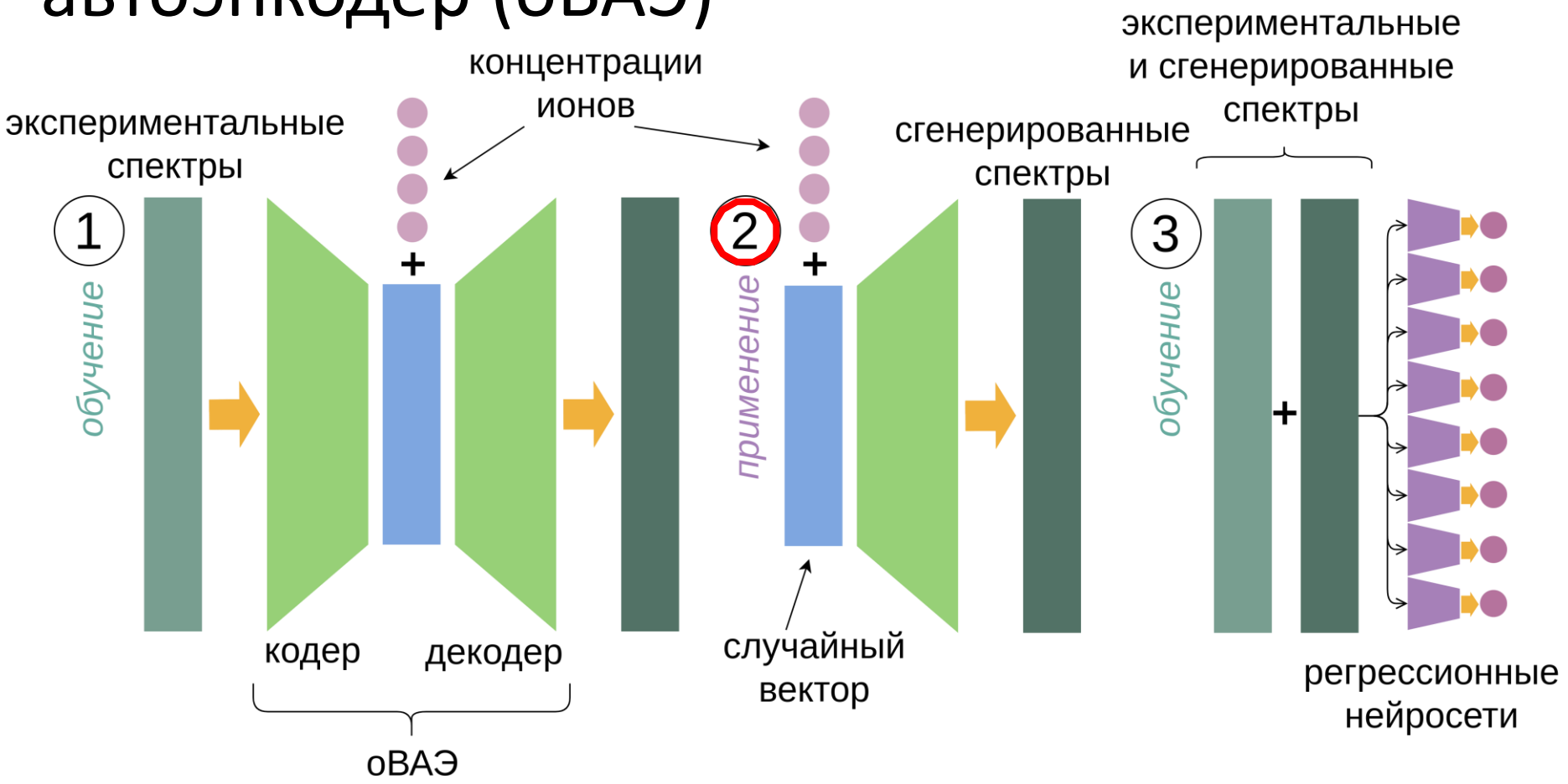
5) Обучение регрессионных нейросетей на расширенном наборе данных

Вычислительный эксперимент: Обусловленный вариационный автоэнкодер (оВАЭ)



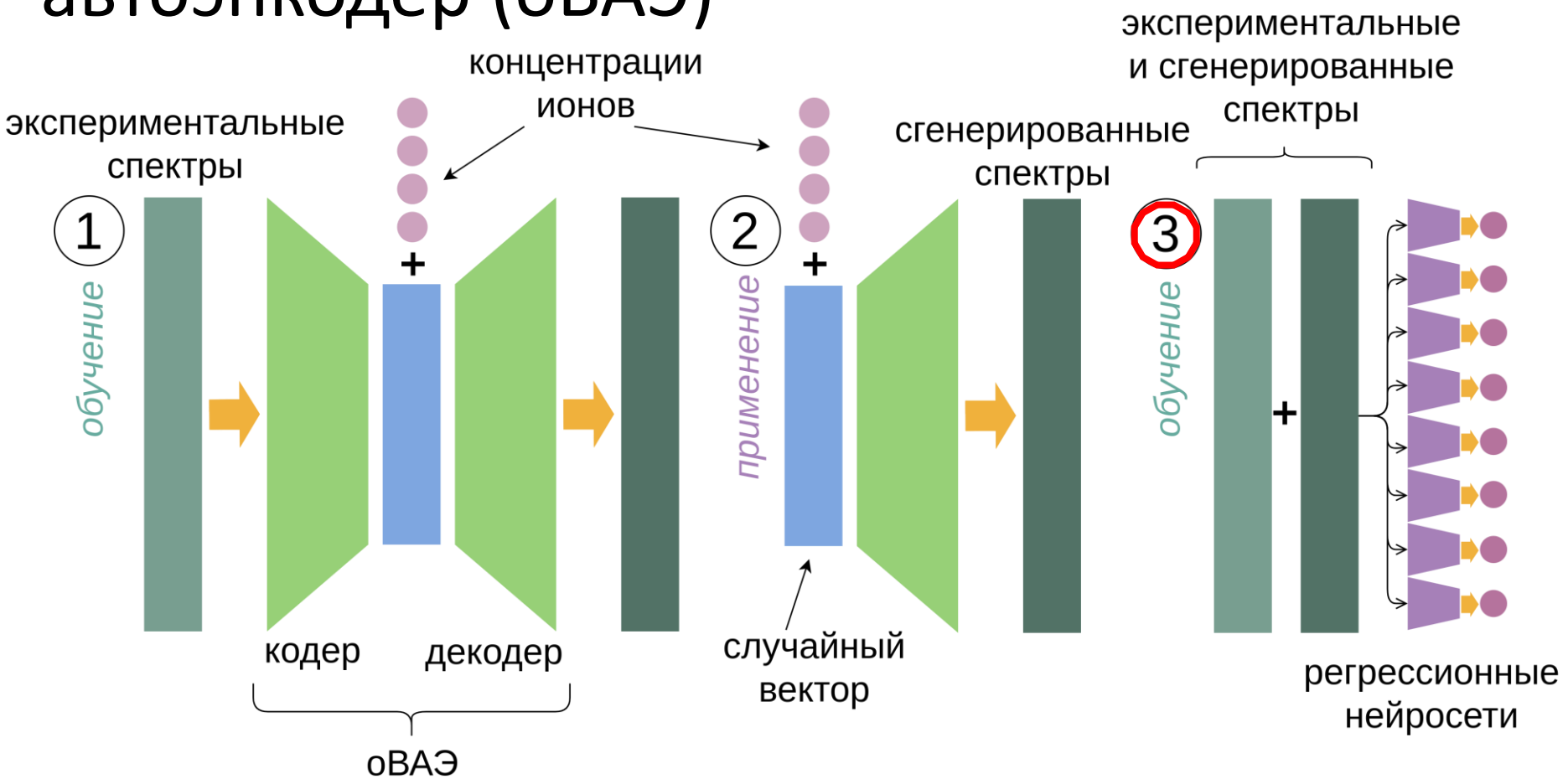
1) Обучение оВАЭ на экспериментальных спектрах и соответствующих наборах концентраций

Вычислительный эксперимент: Обусловленный вариационный автоэнкодер (оВАЭ)



2) Генерация примеров
с использованием декодера оВАЭ

Вычислительный эксперимент: Обусловленный вариационный автоэнкодер (оВАЭ)



3) Обучение регрессионных нейросетей на расширенном наборе данных

Параметры экспериментов

- Данные
  - 3760 спектров
  - 2598 каналов
  - 8 ионов
- Эксперименты
  - Кросс-валидация
  - Adam, lr=0.001
- Нейронные сети
  - ❑Регрессионные нейросети для каждого иона
    - 3 скрытых слоя (64, 32 и 16 нейронов)
    - 1 выход

❑ВАЭ

Кодер: MLP

2598 нейронов во входном слое

256 нейронов в скрытом слое

2*128 нейронов в выходном слое

Декодер: MLP

128 нейронов во входном слое

256 нейронов в скрытом слое

2598 нейронов в выходном слое

❑оВАЭ

Кодер: MLP

2598 нейронов во входном слое

256 нейронов в скрытом слое

2*128 нейронов в выходном слое

Декодер: MLP

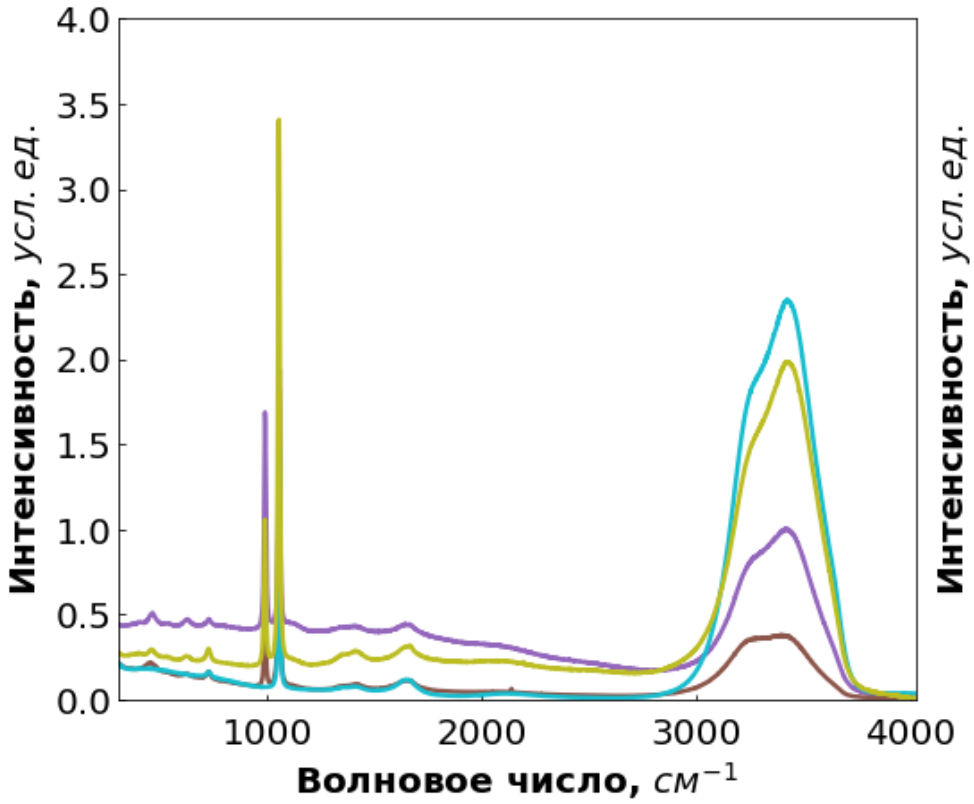
128 +8 нейронов во входном слое

256 нейронов в скрытом слое

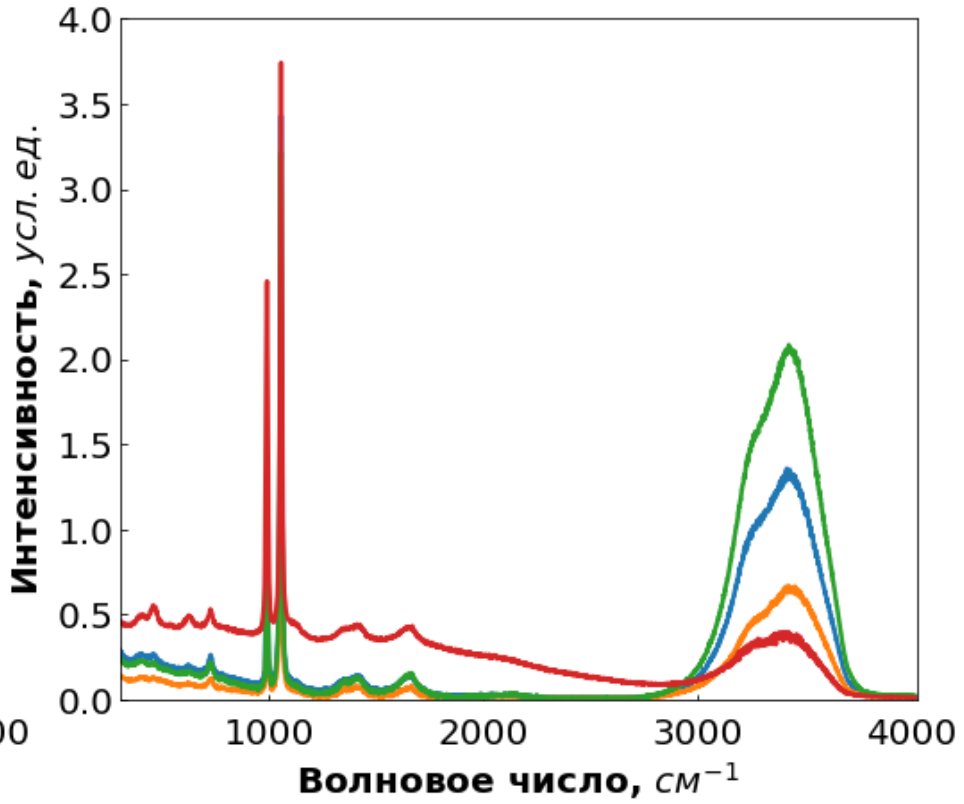
2598 нейронов в выходном слое

ВАЭ-сгенерированные спектры. Анализ

Примеры **экспериментальных** спектров

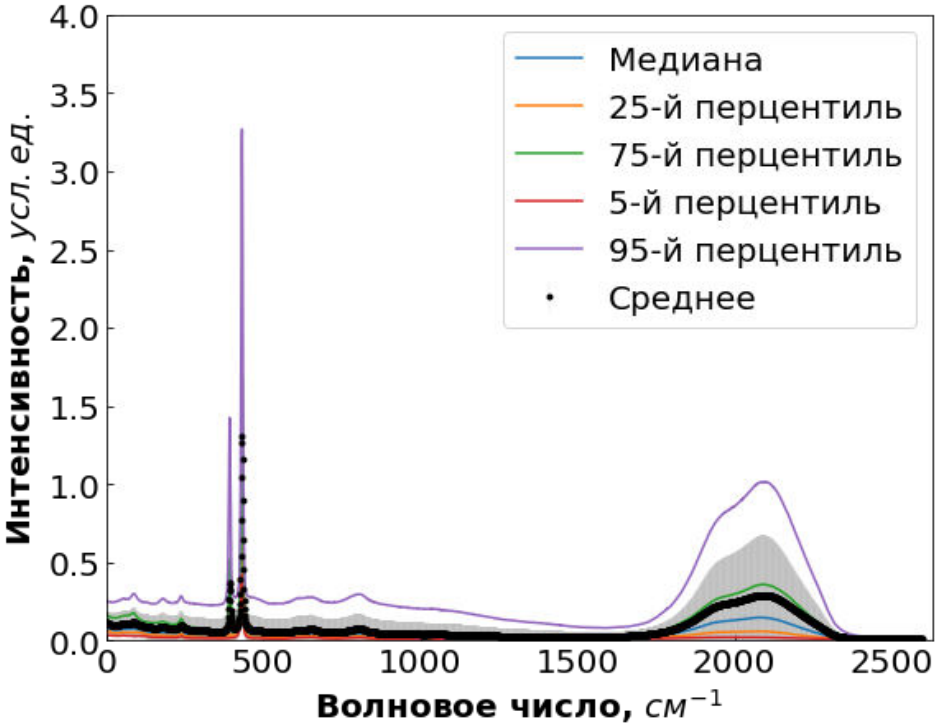


Примеры **ВАЭ-сгенерированных** спектров

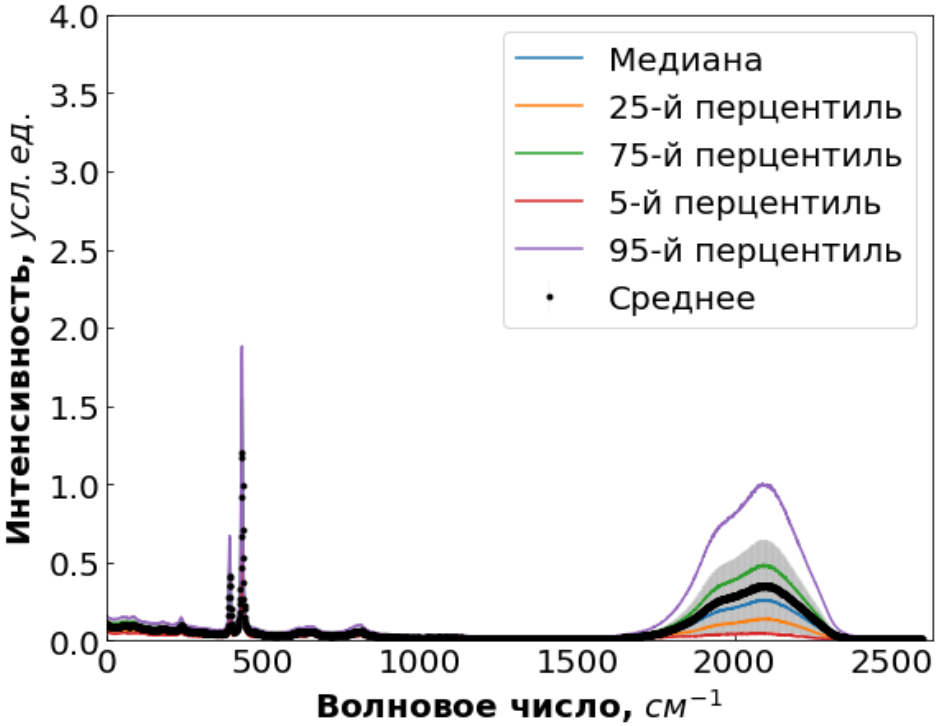


Поканальные статистики. ВАЭ

Экспериментальный
набор данных

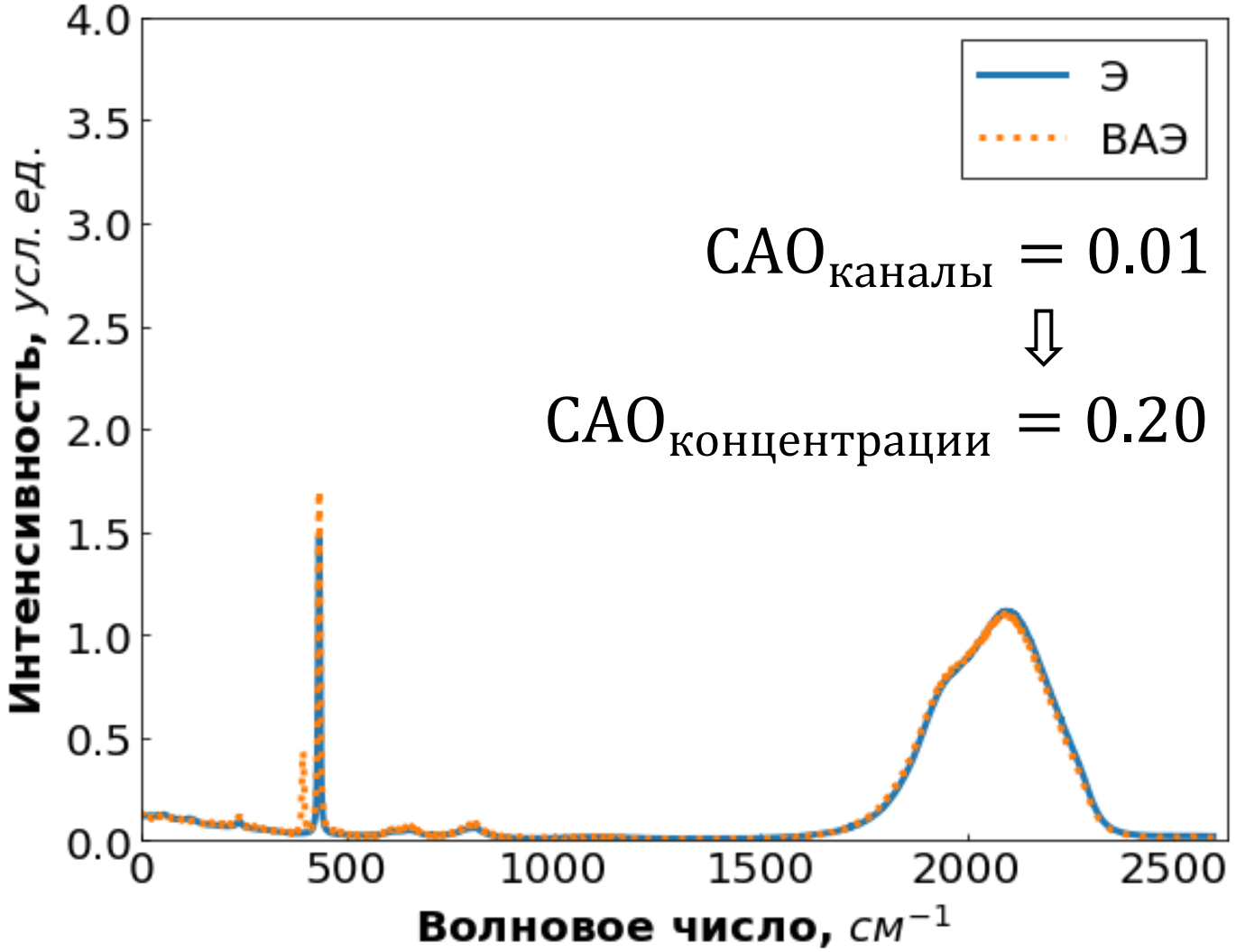


ВАЭ-сгенерированный
набор данных



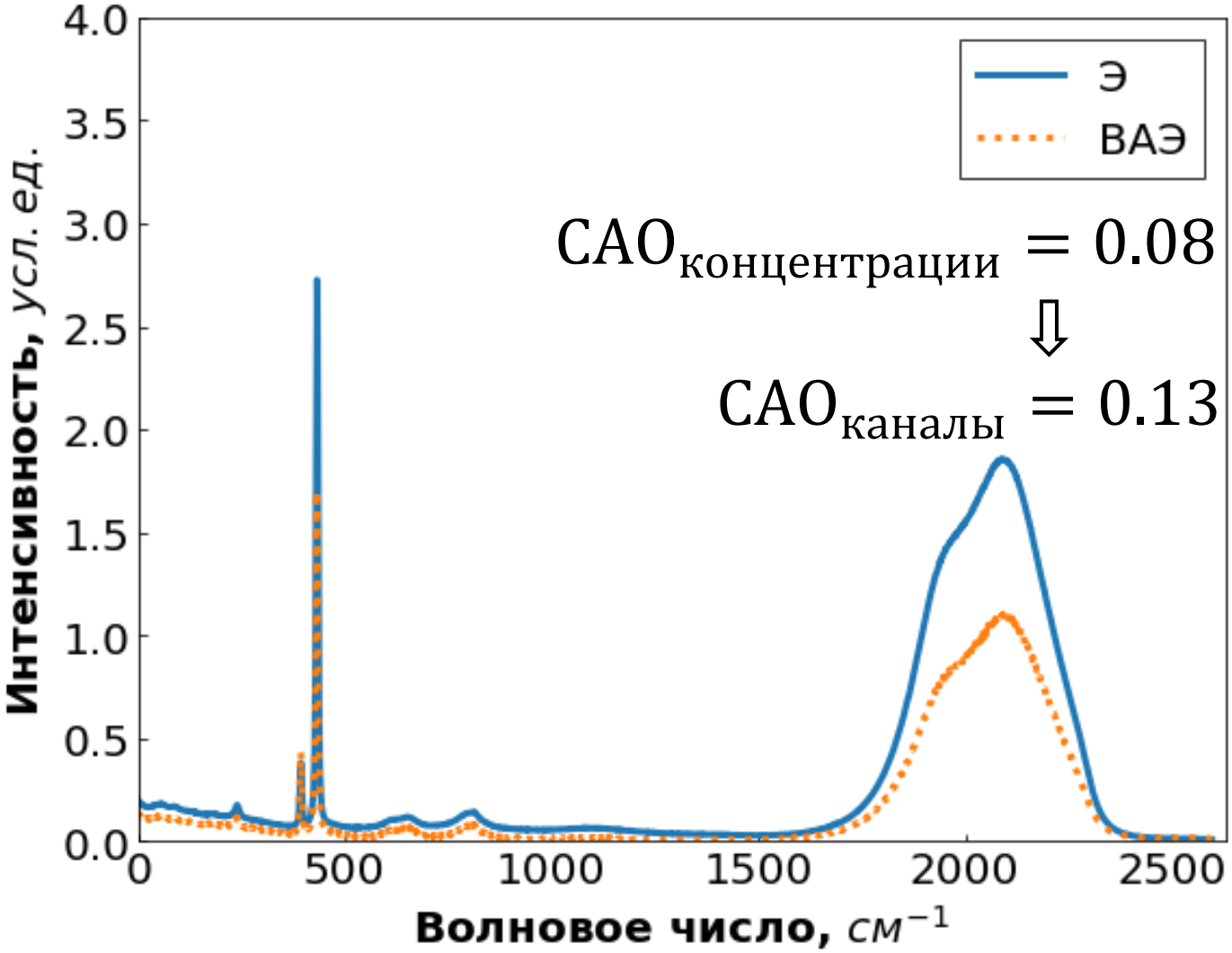
Анализ схожести спектров. ВАЭ

Сравнение экспериментальных и ВАЭ-сгенерированных спектров с близкими значениями в пространстве спектров



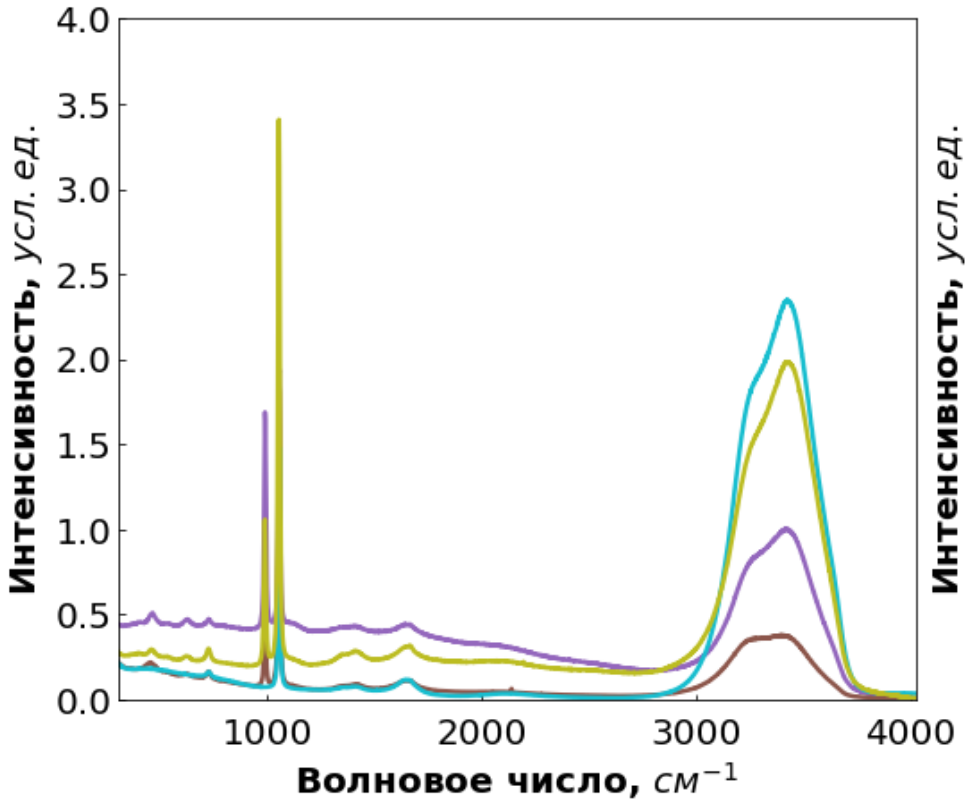
Анализ схожести спектров. ВАЭ

Сравнение экспериментальных и ВАЭ-сгенерированных спектров с похожими наборами концентраций

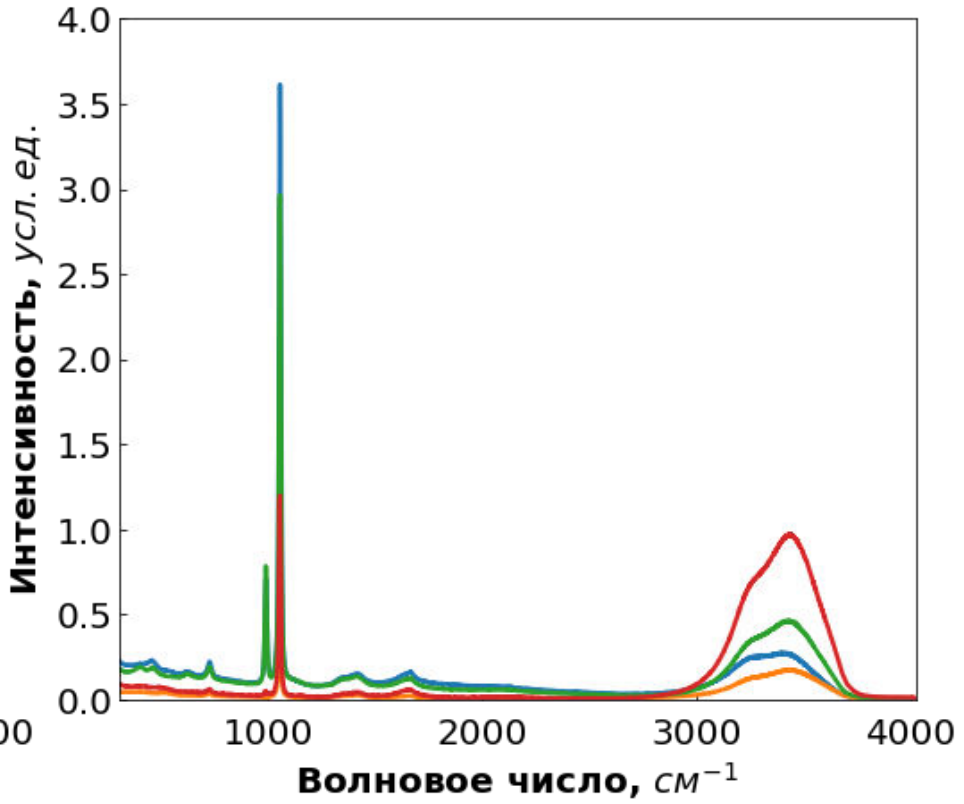


оВАЭ-сгенерированные спектры. Анализ

Примеры **экспериментальных** спектров

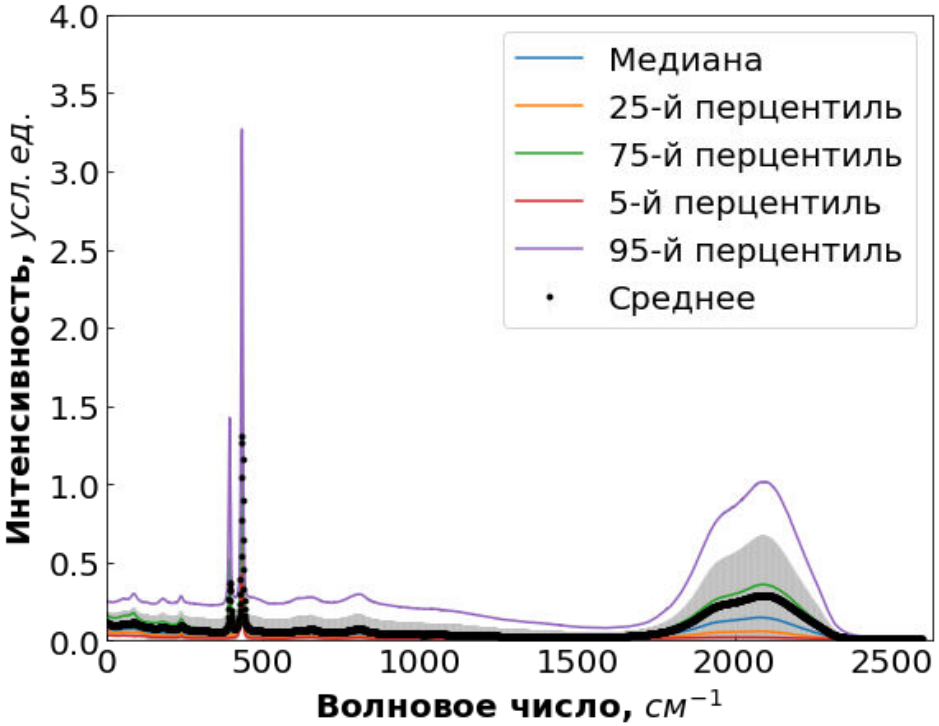


Примеры **оВАЭ-сгенерированных** спектров

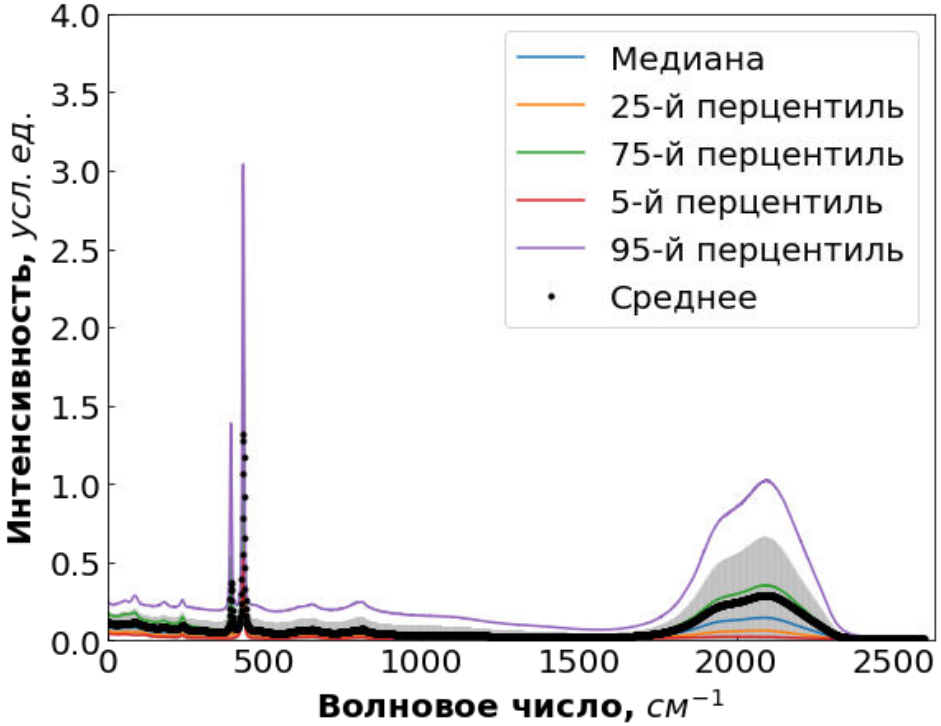


Поканальные статистики. оВАЭ

Экспериментальный
набор данных

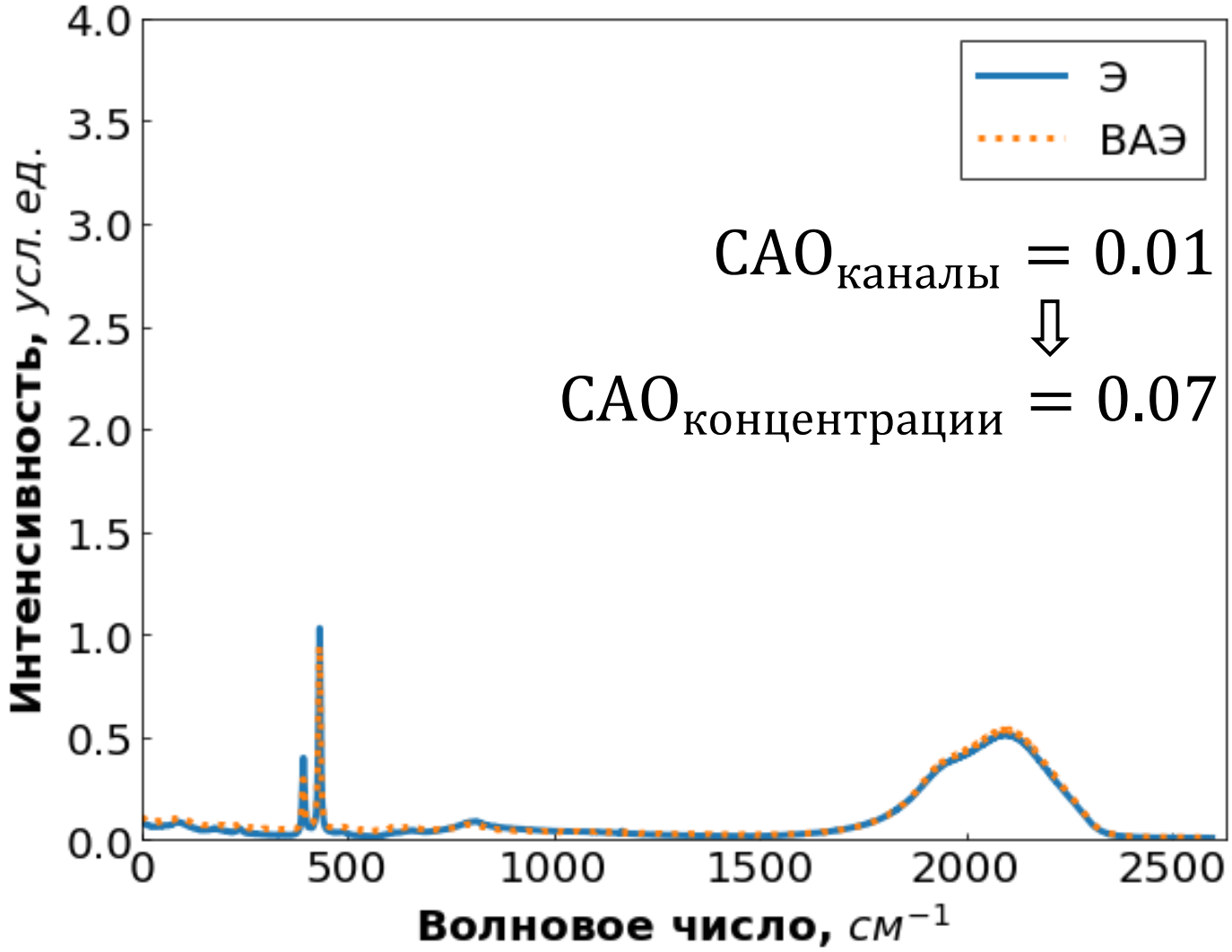


оВАЭ-сгенерированный
набор данных



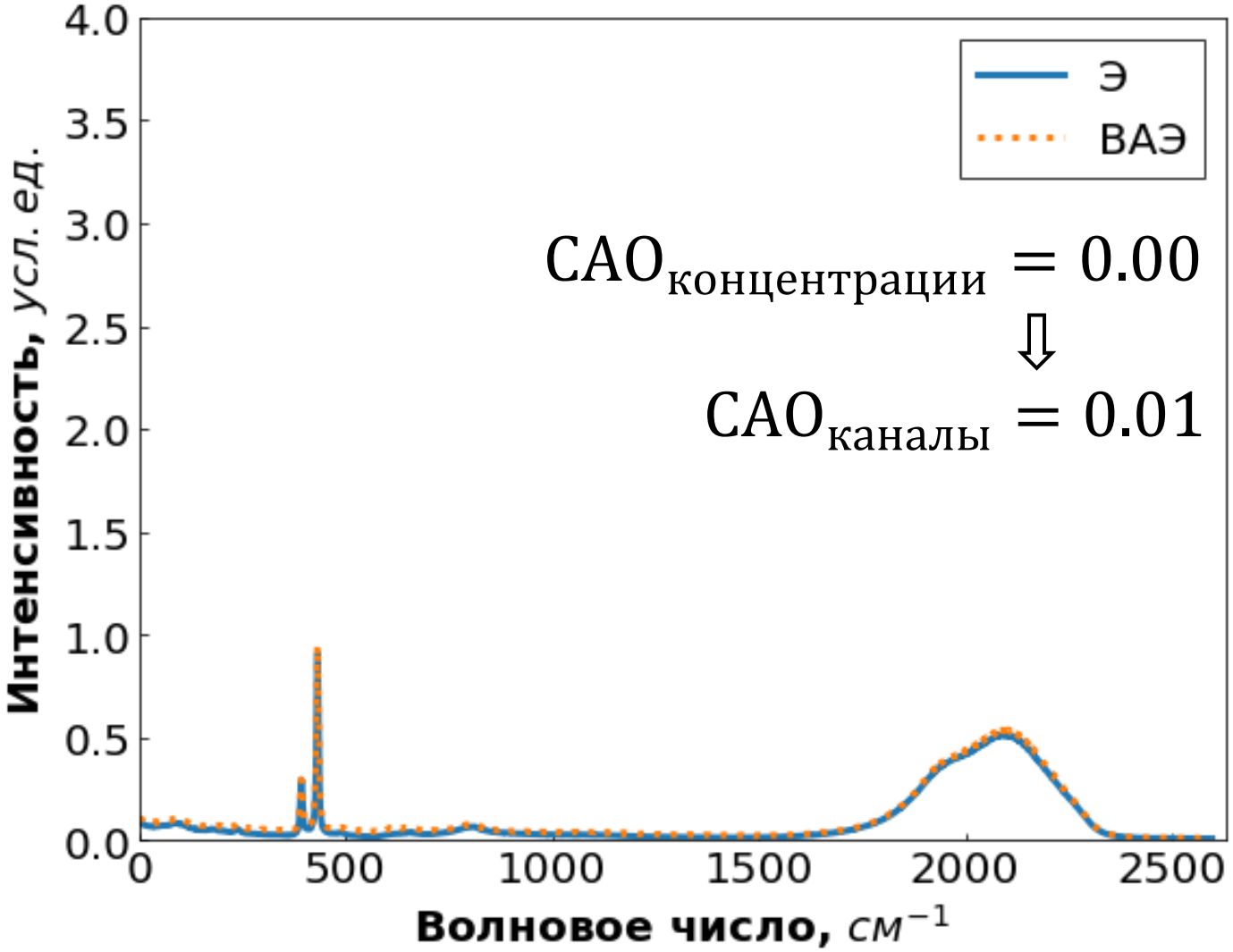
Анализ схожести спектров. оВАЭ

Сравнение экспериментальных и оВАЭ-сгенерированных спектров с близкими значениями в пространстве спектров



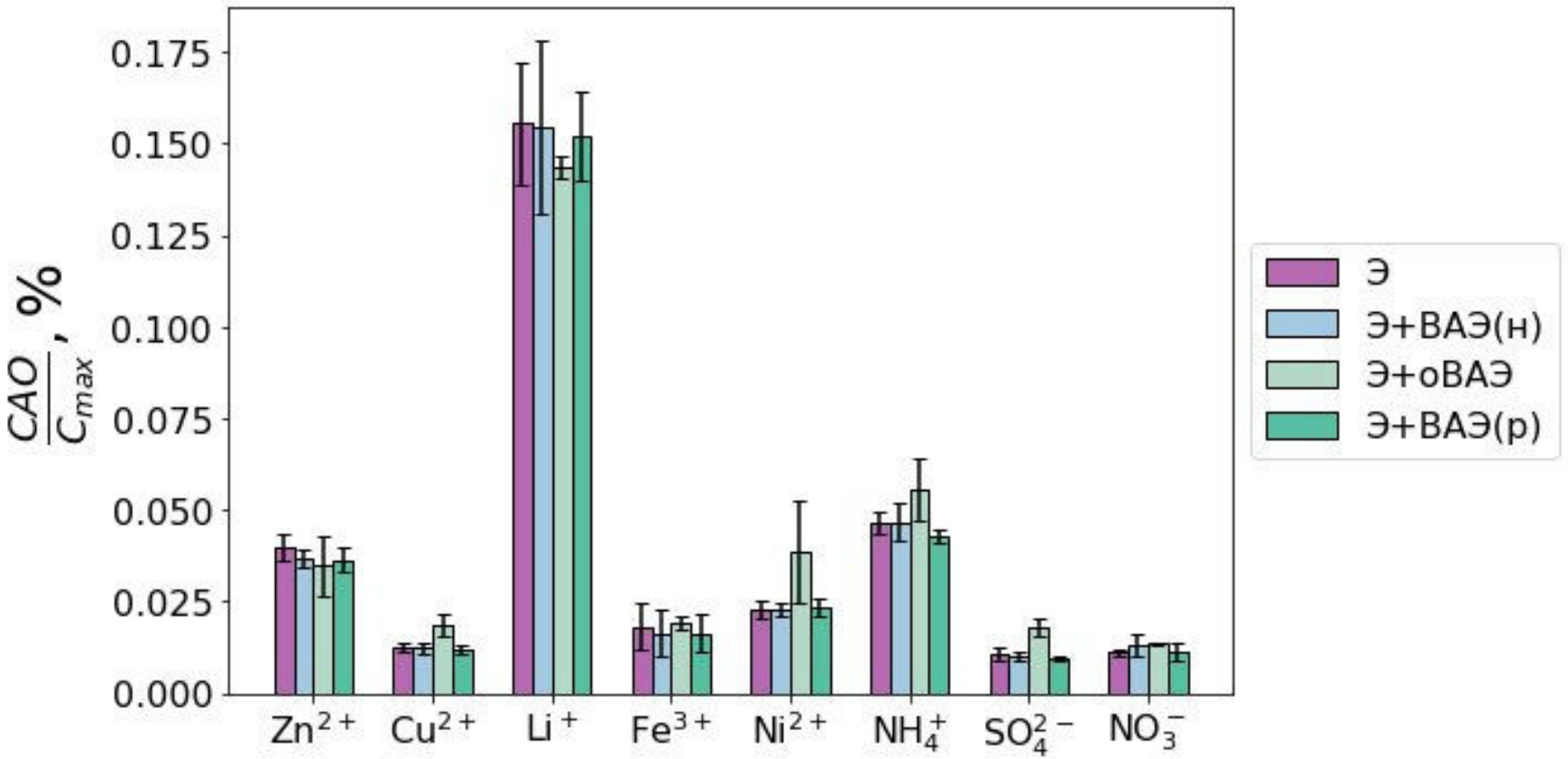
Анализ схожести спектров. оВАЭ

Сравнение экспериментальных и оВАЭ-сгенерированных спектров с одинаковыми наборами концентраций



Качество решения ОЗ

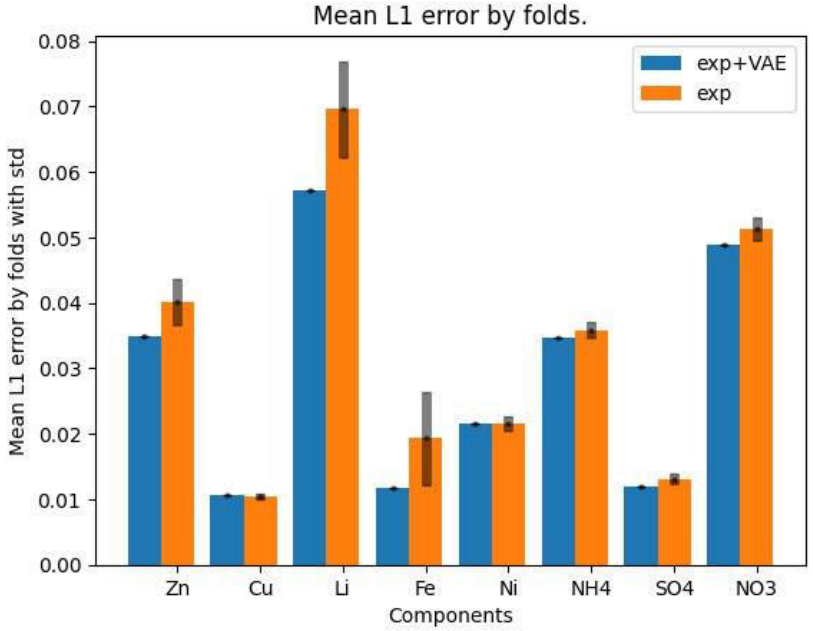
- Расширение обучающего набора ВАЭ-сгенерированными спектрами не привело к снижению качества решения ОЗ
- Сильная зависимость от разбиения набора данных

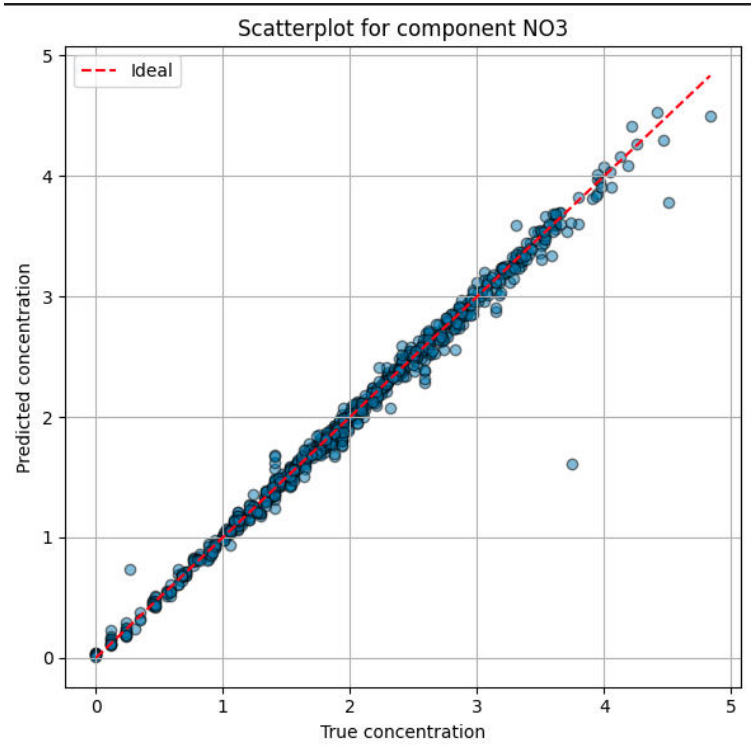


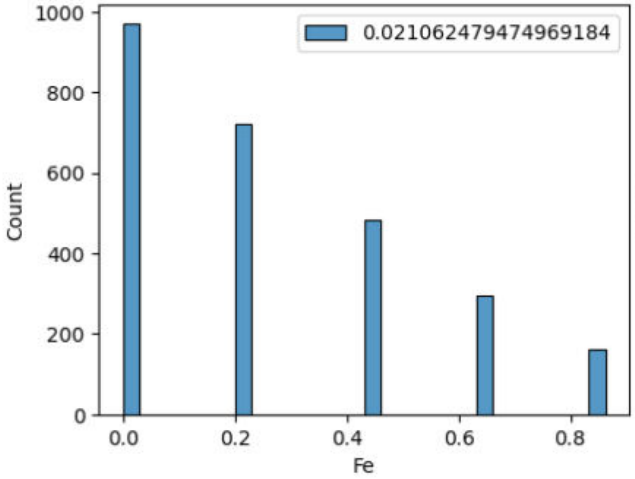
Выводы. Обсуждение

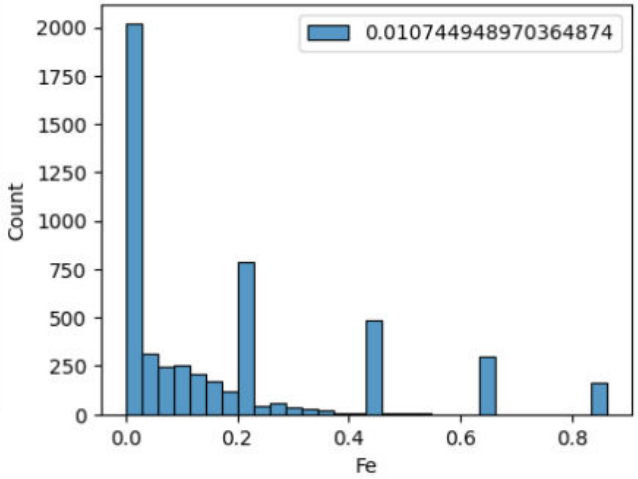
- С использованием ВАЭ возможно генерировать примеры, эффективно имитирующие экспериментальные спектры, но при этом отличающиеся от них
- Требуется исследование возможных стратегий генерации и стратегий задания наборов концентраций
- Для отдельных разбиений набора данных получено незначительное снижение ошибок
- Наблюдается сильный разброс величин ошибок в зависимости от разбиения

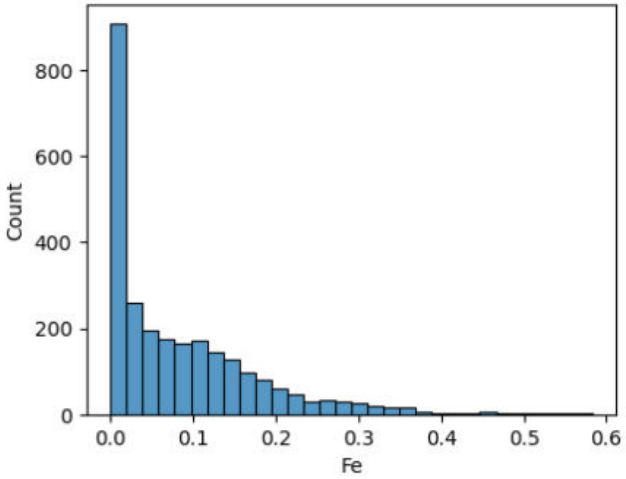
Спасибо за внимание


Mean L1 error by folds.
exp+VAE
exp
Mean L1 error by folds with std
Components
Zn
Cu
Li
Fe
Ni
NH4
SO4
NO3


Scatterplot for component NO3
Ideal
Predicted concentration
True concentration


0.021062479474969184
Count
Fe


0.010744948970364874
Count
Fe


Count
Fe

- Возможное обоснование работы метода
  - Снижение шума за счёт понижения размерности в латентном пространстве
  - Выравнивание распределения

Будущие исследования

- Подбор оптимальных параметров для решения ОЗ
  - Размер латентного пространства ВАЭ
  - Соотношение exp/gen
  - Вид распределения для генерации
  - Стратегия задания значений концентраций при генерации с оВАЭ
- Изучение представления данных в пространстве АЭ, ВАЭ, оВАЭ
- Изучение влияния ВАЭ на шум в данных

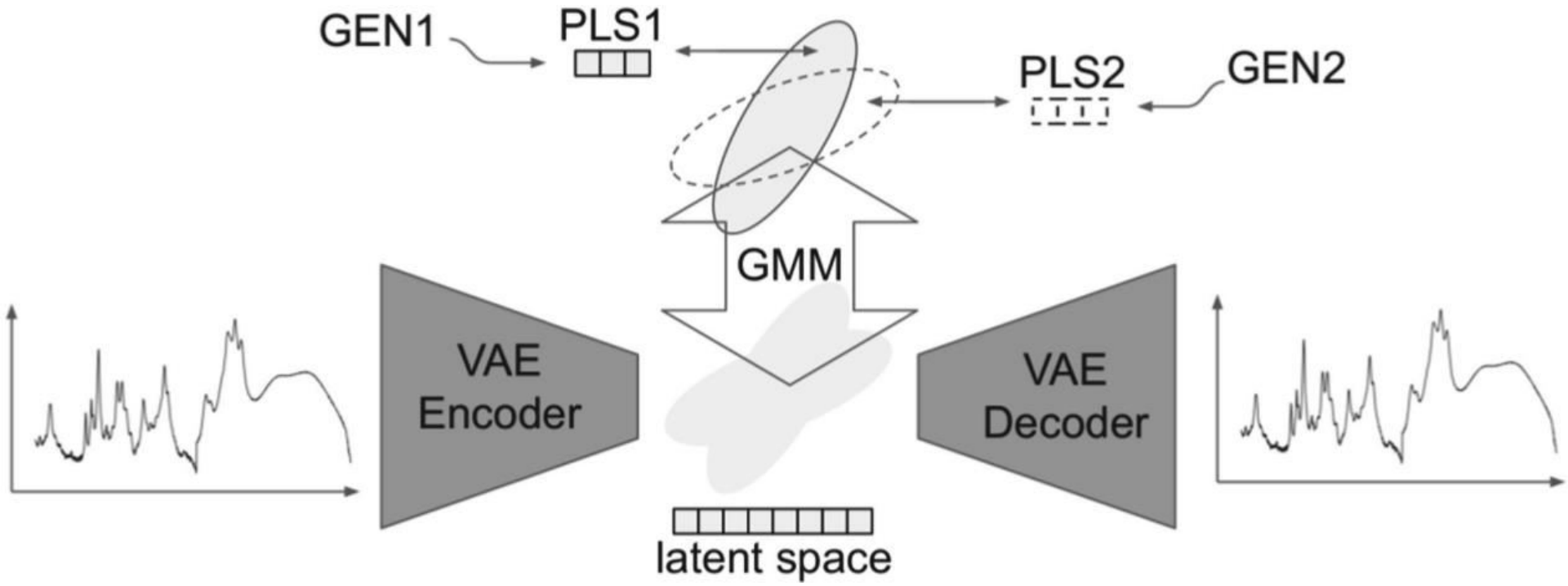
- «Сравнение подходов к повышению представительности спектроскопических данных с помощью вариационных автоэнкодеров» *Журнал технической физики*, 2025, том 95, вып. 5, А.С. Мущина, И.В. Исаев, О.Э. Сарманова, С.А. Буриков, Т.А. Доленко, С.А. Доленко
- Вклад в проект Российского Научного фонда: https://rscf.ru/en/project/24-11-00266/.
- International Conference on Deep Learning in Computational Physics-2024
- Физика плазмы в Солнечной системе 2025
- International Conference «Neuroinformatics-2024»
- ANNUAL INTERNATIONAL CONFERENCE SARATOV FALL MEETING2024
- Научно-исследовательский семинар лаборатории ЛАМОД НИИЯФ МГУ
- Конференция «Ломоносов» 2025

Initial data set consisted of 3744 optical absorption spectra of multicomponent aqueous solutions of salts $Zn(NO_3)_2$, $ZnSO_4$, $Cu(NO_3)_2$, $CuSO_4$, $LiNO_3$, $Fe(NO_3)_3$, $NiSO_4$, $Ni(NO_3)_2$, $(NH_4)_2SO_4$, $NH_4(NO_3)$.

The selected range represents the range of ion concentration changes in technological aqueous media used in non-ferrous metal production facilities.

Ion	Maximum concentration, M	Number of patterns with non-zero ion concentration
Zn^{2+}	1.089	2373
Cu^{2+}	0.955	2373
Li^{+}	0.466	2373
Fe^{3+}	0.862	2373
Ni^{2+}	0.972	2373
NH_4^{+}	0.801	2373
SO_4^{2-}	1.373	3361
NO_3^{-}	4.906	3740

Generating in the latent space



Проблемы получения спектроскопических наборов данных для решения ОЗ с использованием МО

- Эксперименты **трудозатратны, дорогостоящи и занимают много времени**
- Требуется **специальное оборудование**
- Необходимы **квалифицированные специалисты**
- Для репрезентативности нужен **большой объём данных**

Для минимизации затрат, связанных с расширением существующего набора данных, в научном сообществе были разработаны различные методы аугментации данных

Проблемы аугментации данных

- **Добавление шума**
  - Приводит к повышению устойчивости модели к шуму, но не снижает ошибку решения обратной задачи
- **Интерполяция**
  - Форма спектров чувствительна к концентрациям ионов, зависимость оптической плотности от концентраций в многокомпонентных растворах является сложной
- **Использование данных из других источников**
  - Недостаток подходящих баз данных
  - Недостаток наборов данных с нужными компонентами
  - Требуется доменная адаптация