09

Сравнение подходов к повышению представительности спектроскопических данных с помощью вариационных автоэнкодеров

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

[1]Научно-исследовательский институт ядерной физики им. Д.В. Скобельцына
Московского государственного университета им. М.В. Ломоносова,
119991 Москва, Россия
[2]Московский государственный университет им. М.В. Ломоносова, физический факультет,
119991 Москва, Россия
e-mail: anastasemusa@gmail.com

Поступило в Редакцию 7 февраля 2025 г.
В окончательной редакции 7 февраля 2025 г.
Принято к публикации 7 февраля 2025 г.

Решение обратных задач оптической спектроскопии многокомпонентных растворов по определению концентраций компонентов представляет собой достаточно сложную проблему, один из эффективных способов решения которой состоит в использовании искусственных нейронных сетей. Однако одной из сложностей на этом пути может являться недостаточная представительность экспериментальных данных, связанная со сложностью и высокой стоимостью масштабного физического эксперимента. Рассмотрены и сравнены между собой алгоритмы генерации дополнительных модельных данных с помощью вариационных автоэнкодеров с целью повышения представительности обучающей выборки. Показано, что наиболее перспективным представляется использование обыкновенного (не обусловленного) вариационного автоэнкодера с генерацией примеров из равномерного распределения в латентном пространстве. Поиск оптимального распределения в латентном пространстве для генерации примеров должен стать предметом дальнейших исследований.

Ключевые слова: генерация данных, обратная задача спектроскопии, многокомпонентные растворы, искусственные нейронные сети.

DOI: 10.61011/JTF.2025.05.60292.15-25

Введение

В настоящее время загрязнение среды тяжелыми металлами является глобальной проблемой. Источниками такого загрязнения служат металлургическая промышленность, горнодобывающая промышленность, неправильная утилизация отходов, транспорт, сельское хозяйство, тепловые станции — представляют угрозу аэральные выпадения из стационарных источников и средств передвижения; загрязнение водных ресурсов от поступления промышленных сточных вод в водоемы; осадки сточных вод; отвалы золы, шлака, руд, шламов и т.п.; разливы нефти и солевых растворов в местах нефтедобычи [1].

В современном мире загрязнение водных ресурсов распространяется очень активно, и такая проблема актуальна для многих районов по всему миру. В воде тяжелые металлы находятся чаще всего в виде ионов, и в отличие от органических загрязнителей ионы тяжелых металлов не поддаются биодеградации, поэтому, попадая в окружающую среду, они могут накапливаться в воде, почве и попадать в живые организмы.

Специалистами по охране окружающей среды выделена группа металлов, наиболее опасных для здоровья человека и животных, в нее входят кадмий, медь, мышьяк, никель, ртуть, свинец, цинк, хром. Тяжелые металлы, накапливаясь в живых организмах в критических концентрациях, влияют практически на все системы организма, оказывая токсическое, аллергическое, канцерогенное, гонадотропное действие [2].

Таким образом, мониторинг концентраций ионов тяжелых металлов в сточных водах, природных водоемах и грунтовых водах является критически важным.

Существуют различные методы определения концентраций тяжелых металлов в воде, такие как: хроматография [3], амперометрия [4], потенциометрия [5], проточно-инжекционный анализ [6], вольтамперометрия [7], капиллярный ионный анализ [8] и др. При этом не существует методов, одновременно удовлетворяющих условиям хорошей точности, высокой селективности, быстроты, дешевизны и низким требованиям к квалификации персонала. Например, селективность может обеспечиваться за счет использования реагентов, специфических к определяемому веществу (проточно-инжекционный анализ), что повышает стоимость использования данного метода. Хроматография позволяет достичь высокой точности решения, однако ее применение ограничивается значительными временными затра-

тами, высокой стоимостью используемого оборудования и сложностью реализации эксперимента. Метод оптической спектроскопии представляет собой разумный компромисс между требованиями к скорости применения, экономичности, простоте, высокой точности и селективности.

Также метод оптической спектроскопии [9] используется для определения концентраций ионов в растворах. Преимущества метода в том, что его применение является неинвазивным, быстрым и не требует сложной предварительной подготовки образцов. Кроме того, применение метода не является дорогостоящим, так как основные финансовые затраты приходятся на этап разработки метода, а не на этап его использования. На сегодняшний день наиболее популярны методы спектроскопии поглощения и спектроскопии комбинационного рассеяния света.

Однако при использовании спектроскопии для определения концентраций компонентов раствора возникает необходимость в решении обратной задачи (ОЗ). Из-за высокой сложности такой задачи (часто задача плохо обусловлена; сложная зависимость формы спектра от физических параметров; сложность физической модели для аналитического описания) для описания формы спектра раствора с заданными концентрациями компонентов не существует адекватной математической модели. Поэтому для решения ОЗ спектроскопии подходят только аппроксимационные методы, в число которых входят методы машинного обучения, в частности искусственные нейронные сети (ИНС).

Использование методов машинного обучения требует наличия качественного репрезентативного набора данных, состоящего из большого количества примеров. Однако часто получение такого набора данных затруднительно, в частности, для задач спектроскопии, так как проведение экспериментов является трудным, дорогостоящим и времязатратным; кроме того, необходимо сложное оборудование, а для подготовки образцов и предобработки экспериментальных данных требуется привлечение компетентных специалистов.

Для преодоления трудностей с получением достаточного количества примеров для обучения уже были предложены стандартные технологии аугментации — расширения существующего набора данных, подходящие для многих задач. Эти технологии зависят от типа данных и включают такие техники как: наложение шума, геометрические преобразования, коррекция цвета, масштабирование — для изображений [10], изменения тона и скорости — для аудиоданных, растяжение или сжатие временной оси [11], сдвиг по времени — для временных рядов [12] и др.

Однако, к сожалению, такие методы не вполне применимы в задачах, имеющих дело с более сложными, специфическими данными. В случае со спектроскопическими данными, рассматриваемом в настоящей работе, стандартные техники аугментации не подходят: адекватную интерполяцию невозможно реализовать из-за того, что зависимость оптической плотности от концентраций нескольких компонентов является сложной и нелинейной; использование данных из открытых источников и от других экспериментов проблематично, так как требуется доменная адаптация (учитывающая различие в данных, возникающее из отличий в экспериментальном оборудовании, постановке эксперимента и пр.), и подходящих источников очень мало; наложение шума на экспериментальные спектры приводит к увеличению устойчивости модели к шумам, но не к повышению качества решения [13].

В таких случаях можно обратиться к методам дополнения существующих наборов данных искусственными примерами при помощи подходов, опирающихся на генеративные нейросетевые архитектуры — вариационные автоэнкодеры [14] и генеративно-состязательные нейросети [15].

Генеративно-состязательные нейросети менее прозрачны для понимания и интерпретации результатов и более сложны в обучении из-за ряда свойственных им проблем, таких как затухающие градиенты, нестабильность, коллапс моды [16].

Настоящее исследование сосредоточено на изучении возможностей вариационных автоэнкодеров в генеративном моделировании данных.

Вариационные автоэнкодеры используют вероятностный подход к кодированию информации, что позволяет строить эффективное латентное представление данных и затем с помощью этого представления преобразовывать данные для оптимальной работы с ИНС.

Несмотря на то что сгенерированные данные не несут новой информации об объекте, поскольку для обучения генеративной модели используется существующая выборка данных, положительный эффект от их использования может быть обусловлен следующими факторами. С одной стороны, в силу специфики своей архитектуры (наличия „узкого горла“, обеспечивающего понижение размерности данных при переходе в латентное пространство), автоэнкодеры могут снижать уровень шума в данных. С другой стороны, посредством сгенерированных данных возможно снизить несбалансированность и иные известные дефекты распределения исходных данных, что может положительно сказаться на процедуре обучения моделей машинного обучения. Целью настоящей работы является проверка гипотезы о том, что использование сгенерированных данных для расширения обучающей выборки способно повысить репрезентативность набора данных за счет воздействия на распределение данных и, как следствие, снизить погрешность решения обратной задачи.

Исследования [17] показывают, что даже стандартизация данных (которая не меняет форму распределения, но лишь его количественные характеристики) может улучшить стабильность нейронных сетей (НС) и повысить качество решения.

В рамках настоящего исследования рассматривались следующие подходы, позволяющие повлиять на распре-

деление определяемых параметров. Один из них — это генерация дополнительных примеров из нормального распределения и за счет этого приближение распределения объединенного набора данных к нормальному. Это может позволить справиться с негативным влиянием того обстоятельства, что тренировочный набор данных содержит спектры растворов, в которых концентрация каждого из компонентов принимает лишь несколько дискретных значений с определенным шагом („сетка“ параметров). Второй подход, исследуемый в настоящей работе — это генерация дополнительных примеров из равномерного распределения как метод выравнивания представительности для несбалансированных данных (повышения представительности на „крыльях“ нормального распределения), что должно способствовать улучшению обобщающей способности ИНС и более сбалансированному обучению.

Следует отметить, что сказанное относится к распределениям и к представительности данных в латентном пространстве, а не в пространствах исходных данных или целевых переменных — автоэнкодер не обязан сохранять форму распределения данных при переходе из латентного пространства в другие пространства. Его задача — построить компактное представление, которое наиболее эффективно кодирует информацию о данных. Однако при этом исправление распределения данных в латентном пространстве ведет к его изменению (в том числе к повышению представительности данных) и в других пространствах. Это может приводить к улучшению обучения НС на таких исправленных (дополненных синтетическими примерами) данных, что, в свою очередь, может приводить к повышению качества решения рассматриваемой задачи. Проверке того, возможно ли добиться такого эффекта, и посвящена настоящая работа.

В настоящей работе рассмотрено нейросетевое решение восьмипараметрической обратной задачи спектроскопии. Задача состоит в одновременном определении концентраций ионов металлов (Cu^{2+}, Ni^{2+}, Fe^{3+}, Zn^{2+}, Li^{+}), иона NH_4^+ и кислотных остатков (SO_4^{2-}, NO_3^-) многокомпонентных водных растворов солей тяжелых металлов по их спектрам поглощения. Исследованы методы улучшения репрезентативности набора спектроскопических данных, опирающиеся на вариационные автоэнкодеры (ВАЭ).

Настоящая работа является продолжением предыдущих исследований авторов. В работе [18] рассмотрено использование обусловленных ВАЭ для расширения обучающей выборки при решении задачи классификации. В работе [19] обусловленные ВАЭ используются в комбинации со специальной процедурой выборки из латентного пространства метода частичных наименьших квадратов. В настоящей работе производится сравнение подходов, связанных с использованием обусловленных ВАЭ и обыкновенных (не обусловленных) ВАЭ, с выборкой примеров из разных видов распределения в латентном пространстве ВАЭ.

1. Принцип работы ВАЭ

1.1. ВАЭ

Автоэнкодер представляет собой полносвязную НС, которая состоит из двух частей — кодера, который переводит информацию о данных из исходного пространства в пространство меньшей размерности, называемое латентным, и декодера, который восстанавливает сигнал из латентного пространства обратно в исходное. Эффективность такого преобразования определяется тем обстоятельством, что используемая при обучении функция потерь обеспечивает минимизацию отклонения между данными на входе и на выходе автоэнкодера.

Архитектура ВАЭ (рис. 1) практически такая же, с двумя отличиями: 1) специальная функция потерь, используемая при его обучении, позволяет преобразовывать информацию так, чтобы не только разница между представлениями данных на входе и на выходе ВАЭ была минимальной, но и чтобы при этом в латентном пространстве ВАЭ точки, представляющие исходные данные, имели распределение, близкое к нормальному распределению; 2) на выходе кодера формируется не представление данного конкретного примера, а параметры нормального распределения в латентном пространстве, наилучшим образом описывающего тренировочный набор данных.

Функция потерь для ВАЭ складывается из двух частей: первое слагаемое — средний квадрат отклонения между значениями на входе и на выходе ВАЭ — отвечает за воспроизведение входного сигнала на выходе ВАЭ, второе слагаемое — так называемая дивергенция Кульбака–Лейблера D_{KL} [20] — отвечает за схожесть распределения данных в латентном пространстве $\omega_D(h_1, h_2, \ldots)$ и многомерного нормального распределения $\omega_N(h_1, h_2, \ldots)$.

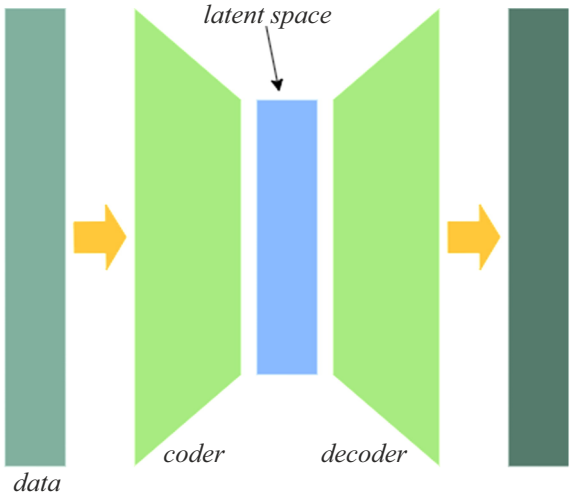

Рис. 1. ВАЭ.

Таким образом, функция потерь L для этих сетей задается выражением:

$$L = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2 + D_{KL}(\omega_N, \omega_D),$$

где $\{x_i\}$ — входной вектор, $\{y_i\}$ — выходной вектор, D_{KL} — дивергенция Кульбака–Лейблера, ω_N — нормальное распределение, ω_D — распределение данных в латентном пространстве, N — количество примеров в наборе данных.

Следует обратить внимание на то обстоятельство, что для обучения сети, решающей ОЗ, требуется информация о значениях целевых переменных, соответствующих каждому спектру (примеру обучающей выборки). Для исходных экспериментальных спектров эта информация известна *a priori*, например, если растворы, спектры которых используются, изготавливались специально, и реальные значения концентраций компонент для них известны, или если они измерены альтернативным способом. Для спектров, сгенерированных с помощью ВАЭ, эта информация отсутствует, и дополнительно необходим метод определения значений целевых переменных, соответствующих каждому сгенерированному спектру.

1.2. Обусловленный ВАЭ

В обусловленном ВАЭ (оВАЭ), кроме информации о данных в латентном пространстве, декодер во время обучения получает на вход информацию о значениях целевых переменных решения ОЗ (концентраций компонентов, которым должен соответствовать сгенерированный декодером спектр (рис. 2). За счет этого при помощи оВАЭ становится возможным генерировать примеры, соответствующие заранее задаваемым значениям целевых переменных [21], что оказывается ценным при последующем использовании сгенерированных данных для обучения сетей, решающих ОЗ — описанная выше для не оВАЭ проблема определения значений целевых переменных для сгенерированного спектра здесь отсутствует. Более того, можно задавать для генерируемых примеров такие значения целевых переменных, которые должны в случае адекватной работы декодера обеспечивать желаемое распределение данных непосредственно в пространстве целевых переменных, а не в латентном пространстве.

2. Физический эксперимент

Физический эксперимент по получению данных для исследования представлял собой измерение спектров оптического поглощения водных растворов солей $Zn(NO_3)_2$, $ZnSO_4$, $Cu(NO_3)_2$, $CuSO_4$, $LiNO_3$, $Fe(NO_3)_3$, $NiSO_4$, $Ni(NO_3)_2$, $(NH_4)_2SO_4$, NH_4NO_3 с различными комбинациями ионов и их концентраций.

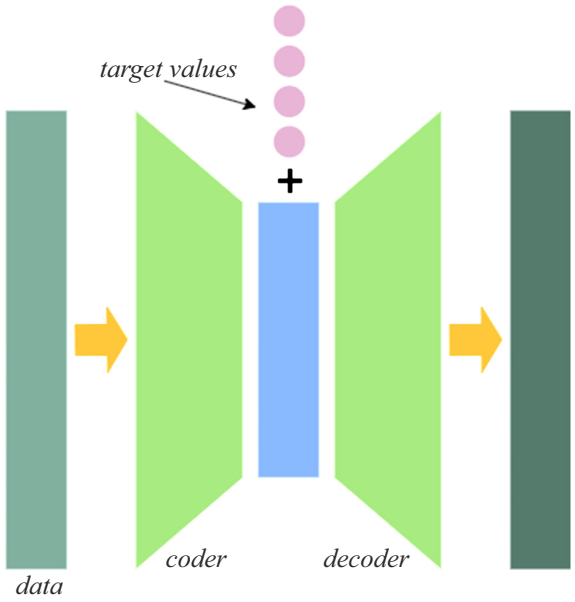

Рис. 2. оВАЭ.

2.1. Спектрофотометрия многокомпонентных водных растворов неорганических солей. Регистрация оптических спектров поглощения

Спектры регистрировались на спектрофотометре Shimadzu UV-1800 в спектральной области 190–1000 nm с шагом 1 nm, ширина щели — 1 μm. Регистрация проводилась в тонкой кювете (длина оптического пути 1 mm) относительно дистиллята. Таким образом, каждый пример — спектр оптического поглощения — охарактеризован 911 признаками — значениями оптической плотности для определенных длин волн, соответствующих каналам спектра.

На рис. 3 приведены спектры оптического поглощения некоторых водных растворов.

2.2. Набор данных

Всего было приготовлено 3744 раствора неорганических солей в дистиллированной воде, содержащих ионы Zn^{2+}, Cu^{2+}, Li^{+}, Fe^{3+}, Ni^{2+}, NH_4^{+}, SO_4^{2-}, NO_3^{-} в различных концентрациях. Концентрации указанных ионов варьировались от 0 до максимального значения, указанного в табл. 1, с разным шагом по концентрациям. Выбранные диапазоны концентраций отражают диапазоны изменения концентрации указанных ионов в технологических водных средах, используемых на предприятиях цветной металлургии.

В табл. 2 приведена информация о количестве одно-, двух-, трех- и т. д. компонентных растворов, т. е. растворов, в которых присутствует указанное количество компонентов.

Таким образом, исходный набор данных состоит из 3744 спектров оптического поглощения многокомпо-

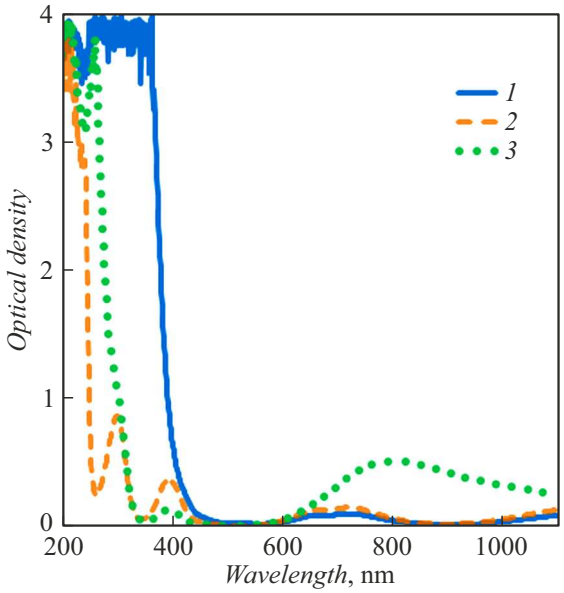

Рис. 3. Примеры спектров оптического поглощения рассматриваемых водных растворов. Представленные спектры соответствуют растворам со следующими концентрациями ионов в mol/l (M). Спектр 1: Zn^{2+} — 0.23, Cu^{2+} — 0, Li^{+} — 0.12, Fe^{3+} — 0.22, Ni^{2+} — 0.49, NH_4^{+} — 0, SO_4^{2-} — 0.49, NO_3^{-} — 1.23. Спектр 2: Zn^{2+} — 0.47, Cu^{2+} — 0, Li^{+} — 0.35, Fe^{3+} — 0, Ni^{2+} — 0.73, NH_4^{+} — 0.2, SO_4^{2-} — 0.83, NO_3^{-} — 1.29. Спектр 3: Zn^{2+} — 0, Cu^{2+} — 0.44, Li^{+} — 0.23, Fe^{3+} — 0, Ni^{2+} — 0.24, NH_4^{+} — 0.8, SO_4^{2-} — 0.64, NO_3^{-} — 1.12.

Таблица 1. Характеристики набора данных

Ион	Максимальная концентрация, M	Число образцов с ненулевой концентрацией иона
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

Таблица 2. Число растворов с одним, двумя, тремя и т.д. компонентами

Число компонентов	Число приготовленных растворов
1	24
2	240
3	1080
4	1590
5	726
6	84

нентных водных растворов солей с различными комбинациями ионов Zn^{2+}, Cu^{2+}, Li^{+}, Fe^{3+}, Ni^{2+}, NH_4^{+}, SO_4^{2-}, NO_3^{-} и их концентраций. Каждый пример описывается 911 признаками.

Таблица 3. Размеры наборов данных для обучения

Набор данных	Количество примеров
Тренировочный	2620
Валидационный	749
Тестовый	375

3. Вычислительный эксперимент

3.1. Кросс-валидация

Исходный набор данных был разделен на тренировочный, валидационный и тестовый в соотношении 7:2:1, таким образом, были получены размеры наборов данных для обучения, представленные в табл. 3.

Для корректной оценки качества решения ОЗ во всех экспериментах было проведено восьмикратное случайное повторение разбиения для перекрестной проверки — кросс-валидация Монте-Карло [22].

Во всех экспериментах по генерации дополнительных данных количество генерируемых спектров выбрано так, чтобы удваивать объем экспериментальной тренировочной выборки, т.е. равно 2620.

3.2. Предобработка входных данных

Для эффективного обучения регрессионных НС (ускорения сходимости во время обучения) в качестве предварительной обработки полученных спектров была проведена процедура нормировки значений оптической плотности в каждом канале в диапазон от 0 до 1. Выбор границы диапазона в нуле физически обусловлен тем, что значения оптической плотности определены в неотрицательной области, а отсутствие поглощения на определенной длине волны должно соответствовать нулевому значению в соответствующем канале спектра.

Для каждого признака x на полученных наборах данных определялись минимальные $x_{\min}$ и максимальные значения $x_{\max}$, затем каждый признак преобразовывался согласно уравнению:

$$x_{\text{norm}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}.$$

Значения концентраций ионов для всех примеров были оставлены без изменений.

3.3. Снижение выходной размерности задачи

В настоящей работе для решения ОЗ спектроскопии как для регрессионной задачи применены ИНС — многослойные персептроны, выполняющие функцию универсального аппроксиматора. Использовался стандарт-

ный подход к снижению выходной размерности многопараметрических задач — автономное определение [23], где для каждого определяемого параметра обучается отдельная модель с одним выходом. Таким образом, всего рассматривается 8 регрессионных НС — по количеству ионов в задаче.

3.4. Увеличение репрезентативности набора данных за счет генерации примеров

В настоящей работе для увеличения качества решения ОЗ спектроскопии с помощью ИНС используются методы повышения представительности данных путем расширения обучающей выборки, опирающиеся на ВАЭ.

При помощи ВАЭ генерируются искусственные спектры, которые затем добавляются к экспериментальной тренировочной выборке. На расширенной тренировочной выборке обучаются регрессионные нейросети.

3.4.1. Референсное решение ОЗ спектроскопии — обучение регрессионных НС только на экспериментальных данных

Для каждого из восьми компонентов Zn^{2+}, Cu^{2+}, Li^{+}, Fe^{3+}, Ni^{2+}, NH_4^{+}, SO_4^{2-}, NO_3^{-} обучены регрессионные нейросети, представляющие собой многослойные персептроны.

3.4.2. Расширение набора данных с помощью ВАЭ

В рамках первого подхода производилось обучение регрессионных ИНС на наборе данных, расширенном с помощью обыкновенного (не обусловленного) ВАЭ.

Алгоритм расширения набора данных с помощью ВАЭ состоит из следующих шагов:

Шаг 1. Обучение референсных регрессионных НС в автономном режиме для прогнозирования концентрации ионов на исходном обучающем наборе экспериментальных спектров оптического поглощения без его расширения.

Шаг 2. Обучение ВАЭ на том же исходном обучающем наборе экспериментальных спектров.

Шаг 3. Генерация примеров с помощью декодера ВАЭ из шума — случайных точек латентного пространства, выбранных в соответствии с параметрами многомерного нормального распределения данных в латентном пространстве, полученного при обучении кодера ВАЭ — для расширения обучающего набора с использованием ВАЭ. Если значение оптической плотности в канале полученного спектра было отрицательным, то оно полагалось равным нулю; если больше единицы, то оно полагалось равным единице.

Шаг 4. Определение концентраций ионов, соответствующих сгенерированным спектрам, с использованием референсных регрессионных НС, обученных на Шаге 1. Если полученное значение концентрации хотя бы одного иона для данного примера выходило за диапазон допустимых значений, пример не включался в выборку.

Шаг 5. Обучение регрессионных НС в автономном режиме для прогнозирования концентрации ионов на расширенном обучающем наборе спектров оптического поглощения двойного размера, включавшем в себя половину экспериментальных спектров и такое же количество спектров, сгенерированных с помощью ВАЭ с применением описанной процедуры.

Кроме того, в рамках описанного подхода также испробован вариант генерации спектров на Шаге 3 из векторов шума, имеющих не нормальное, а равномерное распределение в латентном пространстве.

Схема вычислительного эксперимента по расширению набора данных с помощью ВАЭ представлена на рис. 4.

3.4.3. Расширение набора данных с помощью оВАЭ

Алгоритм расширения набора данных с использованием оВАЭ состоит из следующих шагов:

Шаг 1. Обучение оВАЭ на исходном обучающем наборе экспериментальных спектров оптического поглощения без его расширения. При обучении используются соответствующие этим спектрам известные значения целевых переменных (концентраций ионов).

Шаг 2. Генерация примеров с помощью декодера оВАЭ из шума — случайных точек латентного пространства, выбранных в соответствии с параметрами многомерного нормального распределения данных в латентном пространстве, полученного при обучении кодера оВАЭ — для расширения обучающего набора с использованием оВАЭ. Для генерации дополнительных примеров были выбраны наборы концентраций, присутствующие в тренировочной выборке экспериментального набора данных. Если значение оптической плотности в канале полученного спектра было отрицательным, то оно полагалось равным нулю; если больше единицы, то оно полагалось равным единице.

Шаг 3. Обучение регрессионных НС в автономном режиме для прогнозирования концентрации ионов на расширенном обучающем наборе спектров оптического поглощения двойного размера, включавшем в себя половину экспериментальных спектров и такое же количество спектров, сгенерированных с помощью оВАЭ с применением описанной процедуры.

Схема вычислительного эксперимента по расширению набора данных с использованием оВАЭ представлена на рис. 5.

3.5. Параметры моделей

3.5.1. Архитектура регрессионных НС

В качестве регрессионных НС были использованы многослойные персептроны. Архитектура каждого персептрона включает два скрытых слоя, содержащих 64 и 16 нейронов соответственно, и выходной слой размером

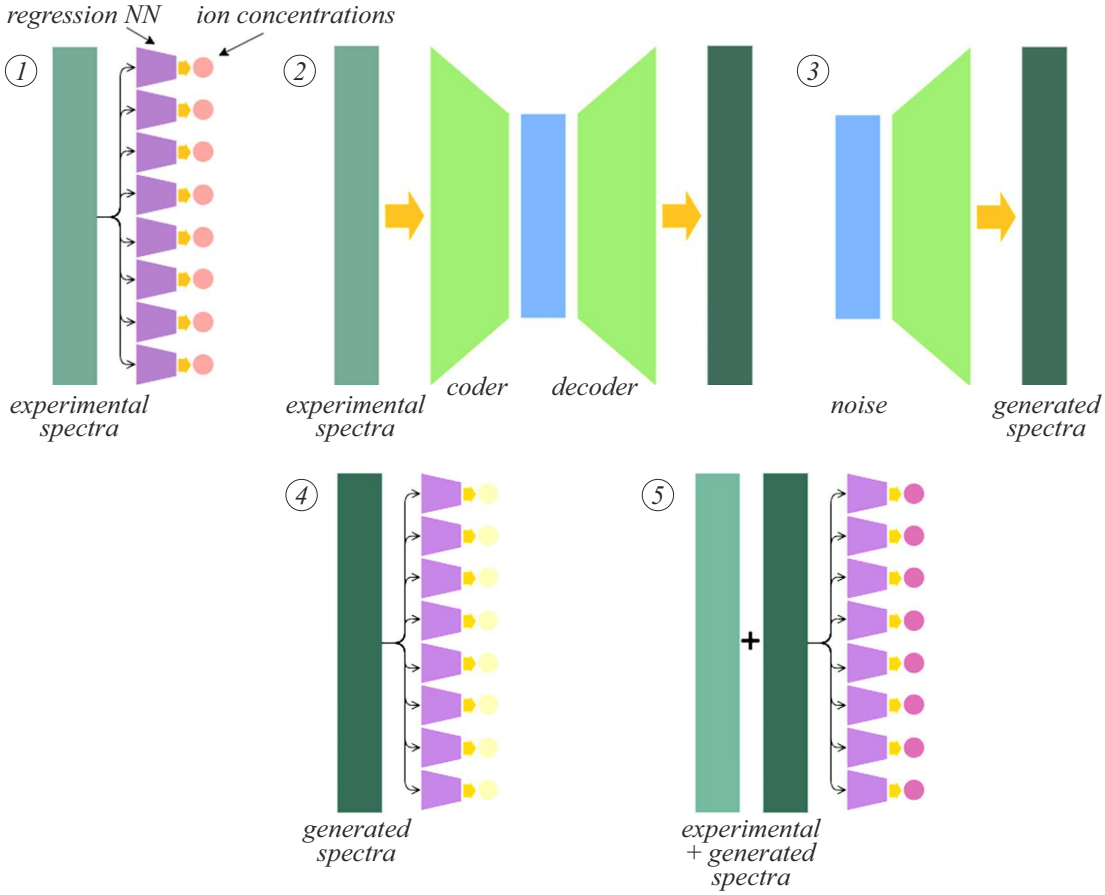

Рис. 4. Схема вычислительного эксперимента по расширению набора данных с помощью ВАЭ: *1* — обучение референсных регрессионных НС на наборе экспериментальных спектров; *2* — обучение ВАЭ на наборе экспериментальных спектров; *3* — использование декодера ВАЭ для генерации модельных спектров из точек с требуемым распределением данных в латентном пространстве (нормальным или равномерным); *4* — применение референсных НС, обученных на Шаге 1, для получения значений целевых переменных, соответствующих сгенерированным спектрам; *5* — обучение регрессионных НС на наборе экспериментальных спектров, расширенном вдвое модельными спектрами, сгенерированными на Шаге 3.

в 1 нейрон. В скрытых слоях в качестве функции активации использовалась сигмоидальная, а в выходном слое — линейная функция.

3.5.2. Архитектура ВАЭ

Кодер — многослойный персептрон с одним скрытым слоем. В скрытом слое 256 нейронов и сигмоидальная функция активации; в выходном слое $2 \cdot LS = 2 \cdot 91$ нейронов и линейная функция активации. Здесь LS — размерность латентного пространства. На выходе у кодера двумерный вектор, длина которого равна размерности латентного пространства $LS = 91$. Одна компонента этого вектора представляет собой значение математического ожидания μ, другая — значение дисперсии σ^2 нормального распределения по соответствующей координате латентного пространства.

Декодер — многослойный персептрон с одним скрытым слоем. Во входном слое $LS = 91$ нейрон и сигмоидальная функция активации; в скрытом слое 256 нейронов и сигмоидальная функция активации; в выходном слое 911 нейронов и линейная функция активации. Здесь LS — размерность латентного пространства. На входе у декодера одномерный вектор $h = v^g_{noise} \times \sigma + \mu$, длина которого равна размерности скрытого пространства $LS = 91$, где v^g_{noise} — случайный вектор размерности $LS = 91$.

3.5.3. Архитектура оВАЭ

Кодер — многослойный персептрон с одним скрытым слоем. В скрытом слое 256 нейронов и сигмоидальная функция активации; в выходном слое $2 \cdot LS = 2 \cdot 91$ нейронов и линейная функция активации. Здесь LS — размерность латентного пространства. На выходе у кодера двумерный вектор, длина которого равна размерности латентного пространства $LS = 91$. Одна компонента этого вектора представляет собой значение математического ожидания μ, другая — значение дисперсии σ^2 нормального распределения по соответствующей координате латентного пространства.

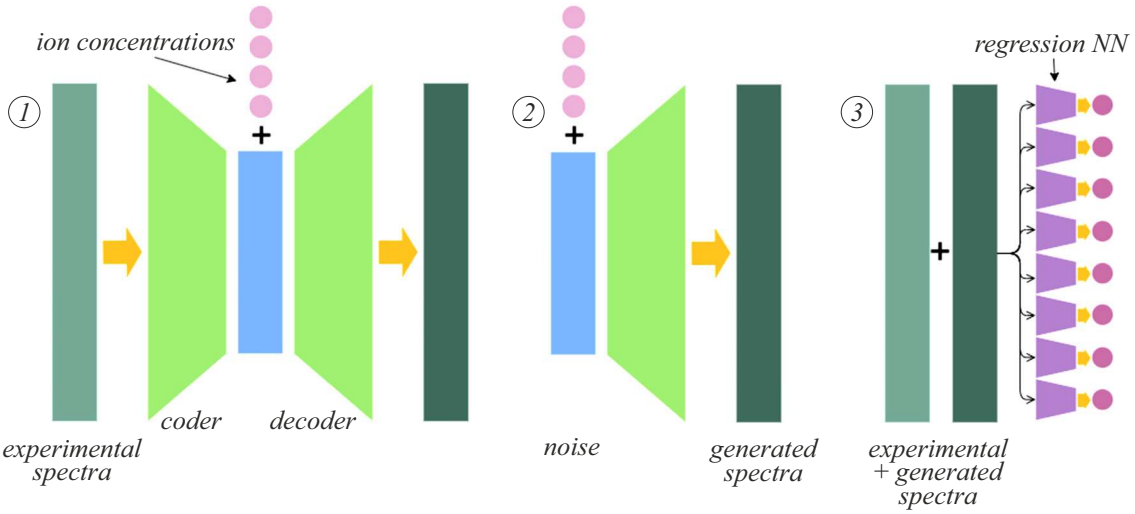

Рис. 5. Схема вычислительного эксперимента по расширению набора данных с помощью обусловленного ВАЭ: *1* — обучение оВАЭ на наборе экспериментальных спектров; *2* — использование декодера оВАЭ для генерации модельных спектров из точек с требуемым распределением данных в латентном пространстве (нормальным или равномерным), соответствующих требуемым наборам концентраций; *3* — обучение регрессионных НС на наборе экспериментальных спектров, расширенном вдвое модельными спектрами, сгенерированными на Шаге 2.

Декодер — многослойный персептрон с одним скрытым слоем. Во входном слое $LS + 8 = 91 + 8$ нейронов и сигмоидальная функция активации; в скрытом слое 256 нейронов и сигмоидальная функция активации; в выходном слое 911 нейронов и линейная функция активации. Здесь LS — размерность латентного пространства, 8 — количество целевых переменных (количество определяемых ионов). На входе у декодера одномерный вектор $h = v^g_{noise} \cdot \sigma + \mu$, длина которого равна размерности скрытого пространства $LS = 91$, где v^g_{noise} — случайный вектор размерности $LS = 91$, и вектор длиной 8, содержащий заданный набор концентраций.

3.5.4. Параметры обучения моделей

Во всех экспериментах по обучению НС критерием остановки обучения НС являлось достижение числа эпох 250 после последнего падения функции потерь на валидационном наборе данных. В качестве алгоритма оптимизации использовалась модификация градиентного спуска Adam [24], скорость обучения составляла 0.001, размер пакета (батча) составлял 64 примера.

Функция потерь для регрессионных НС — среднеквадратичное отклонение MSE:

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i^{true} - y_i^{pred})^2,$$

где y_i^{true} — истинное значение для i-го примера; y_i^{pred} — ответ сети для i-го примера; N — размер выборки (количество примеров в наборе данных).

Функция потерь для вариационных автоэнкодеров — сумма MSE и дивергенции Кульбака–Лейблера D_{KL}:

$$L = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2 + D_{KL}(\omega_N, \omega_D),$$

где $\{x_i\}$ — входной вектор, $\{y_i\}$ — выходной вектор, D_{KL} — дивергенция Кульбака–Лейблера, ω_N — нормальное распределение, ω_D — распределение данных в латентном пространстве, N — количество примеров в наборе данных.

4. Результаты

Для оценки качества решения ОЗ спектроскопии использовано среднее абсолютное отклонение (САО) (Mean Absolute Error,MAE), вычисляющее среднюю абсолютную разность между прогнозируемыми значениями целевой переменной (ответами сети) и истинными значениями. MAE определяется уравнением:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i^{true} - y_i^{pred}|,$$

где y_i^{true} — истинное значение для i-го примера; y_i^{pred} — ответ сети для i-го примера; N — размер выборки (количество примеров в наборе данных).

4.1. Сгенерированные спектры

Спектры, сгенерированные с помощью ВАЭ и оВАЭ, визуально схожи с экспериментальными спектрами (рис. 6). Кроме того, в коротковолновой области у

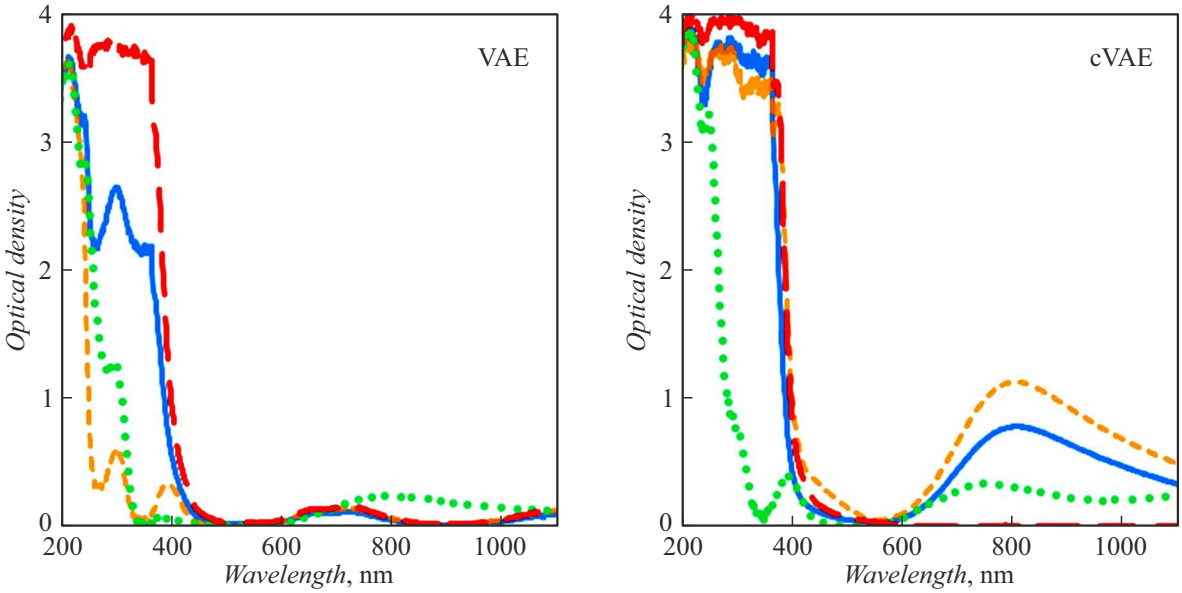

Рис. 6. Примеры спектров, сгенерированных с помощью ВАЭ и оВАЭ.

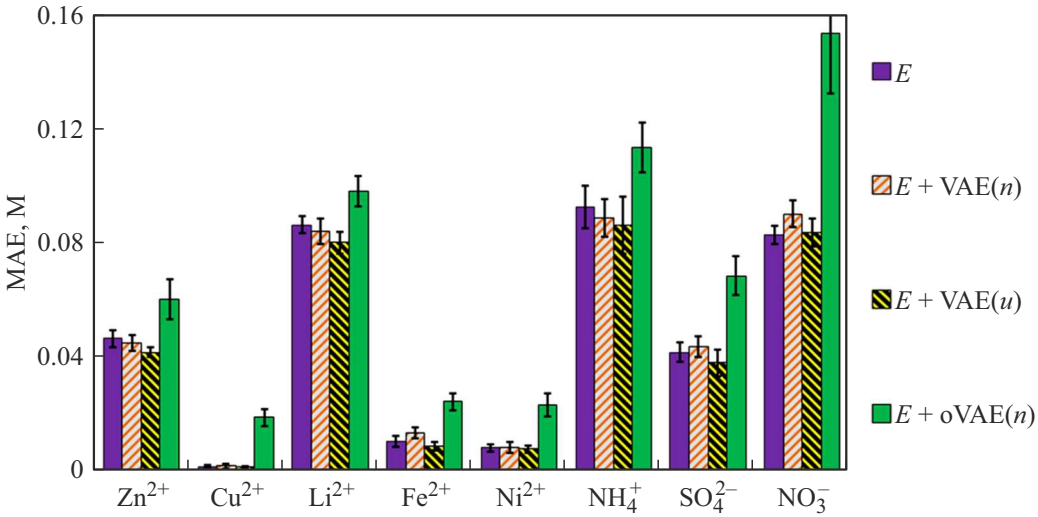

Рис. 7. Метрики ошибок, полученные при решении ОЗ спектроскопии референсным методом (без расширения выборки, E) и с расширением выборки модельными спектрами при использовании ВАЭ с выборкой из нормального распределения в латентном пространстве $E + \mathrm{VAE}(n)$, ВАЭ с выборкой из равномерного распределения в латентном пространстве $E + \mathrm{VAE}(u)$ и оВАЭ с выборкой из нормального распределения $E + \mathrm{oVAE}(n)$.

сгенерированных спектров наблюдается снижение уровня шума по сравнению с экспериментальными.

4.2. Решение регрессионной задачи

Метрики ошибок, полученные при решении ОЗ спектроскопии референсным методом (без расширения выборки) и с расширением модельными спектрами при использовании ВАЭ и оВАЭ, представлены на рис. 7.

Данные результаты говорят о том, что качество решения регрессионных НС в методе с использованием ВАЭ с генерацией модельных спектров из нормального распределения в латентном пространстве практически не изменилось относительно референсного решения в пределах погрешности вычислительного эксперимента. От добавления искусственных примеров в тренировочную выборку статистически значимого ухудшения или улучшения качества решения ОЗ не наступило.

При генерации модельных спектров из равномерного распределения в латентном пространстве получаемые результаты оказываются в среднем лучше, чем при генерации из нормального распределения. Для ионов Zn^{2+} и Li^+ показано статистически значимое улучшение по сравнению с референсными моделями. Это

свидетельствует о том, что в целом рассматриваемый подход, связанный с генерацией модельных спектров с „исправленным“ распределением, позволяет повысить представительность данных и является перспективным.

При этом с использованием оВАЭ ошибки на тестовой выборке возрастают, для некоторых компонентов весьма существенно. Возможно, такое поведение удастся исправить, управляя стратегией выбора наборов концентраций для генерации.

Заключение

В работе рассмотрено решение обратной задачи спектроскопии многокомпонентных водных растворов солей с содержанием восьми ионов: Zn^{2+}, Cu^{2+}, Li^{+}, Fe^{3+}, Ni^{2+}, NH_4^{+}, SO_4^{2-}, NO_3^{-}, при помощи ИНС. Исследована возможность расширения обучающей выборки экспериментального спектроскопического набора данных с использованием ВАЭ. Цель работы — получить улучшение метрик качества решения ОЗ по сравнению с референсным решением, не использующим расширение тренировочной выборки.

Предложены и описаны алгоритмы для расширения экспериментального набора данных, использующие ВАЭ и оВАЭ. С помощью обученных ВАЭ сгенерированы модельные спектры, которые затем были использованы наряду с экспериментальными во время обучения регрессионных НС, решающих обратную задачу.

Для проверки гипотезы о способности предложенных методов влиять на репрезентативность данных за счет воздействия на их распределение проведены эксперименты по обучению регрессионных НС на разных тренировочных выборках: экспериментальной, расширенной с помощью оВАЭ и расширенной с помощью обыкновенного (не обусловленного) ВАЭ с использованием генерации из нормального и равномерного распределений. Получены метрики качества решения ОЗ.

Метод с оВАЭ показал ухудшение качества решения по сравнению с референсным методом (без расширения обучающей выборки). Метод с использованием ВАЭ и генерацией примеров из нормального распределения в латентном пространстве продемонстрировал качество решения, сопоставимое с референсным, тогда как использование ВАЭ с генерацией примеров из равномерного распределения в латентном пространстве привело к небольшому снижению ошибок на тестовой выборке, статистически значимому для ионов Zn^{2+} и Li^{+} большинства компонентов метрики качества нейронных сетей, обученных на расширенном наборе данных, все еще пересекаются в пределах погрешностей с референсными метриками.

По результатам работы можно сделать вывод о том, что подход является перспективным, но подбор оптимальных условий и параметров вычислительного эксперимента требует дальнейшего исследования. В частности, следует провести исследования по поиску оптимального распределения примеров в латентном пространстве, которое должно использоваться для генерации модельных спектров и расширения обучающей выборки.

Финансирование работы

Исследование выполнено за счет гранта Российского научного фонда № 24-11-00266, https://rscf.ru/project/24-11-00266/

Конфликт интересов

Авторы заявляют, что у них нет конфликта интересов.

Список литературы

[1] Ю.Н. Водяницкий, Д.В. Ладонин, А.Т. Савичев. *Загрязнение почв тяжелыми металлами* (Типография Россельхозакадемии, М., 2012)
[2] Г.А. Теплая. Астраханский вестник экологического образования, **1** (23), 182 (2013).
[3] Y. Fa, Y. Yu, F. Li, F. Du, X. Liang, X. Liu. J. Chromatography A, **1554**, 123 (2018). https://doi.org/10.1016/j.chroma.2018.04.017
[4] N.G. Carpenter, D. Pletcher. Anal. Chim. Acta, **317**, 287 (1995). https://doi.org/10.1016/0003-2670(95)00384-3
[5] C. Pasquini, I.B.S. Cunha. Analyst, **120** (11), 2763 (1995). https://doi.org/10.1039/AN9952002763
[6] N. Porter, B.T. Hart, R. Morrison, I.C. Hamilton. Anal. Chim. Acta, **308**, 313 (1995). https://doi.org/10.1016/0003-2670(94)00330-O
[7] C. Neuhold, K. Kalcher, W. Diewald, X. Cai, G. Raber. Electroanalysis, **6**, 227 (1994). https://doi.org/10.1002/elan.1140060309
[8] B. Saad, F.W. Pok, A.N.A. Sujari, M.I. Saleh. Food Chem., **61** (1-2), 249 (1998). https://doi.org/10.1016/S0308-8146(97)00024-1
[9] V.K. Maurya, R.P. Singh, L.B. Prasad. Orient. J. Chem., **34** (1), 100 (2018). http://dx.doi.org/10.13005/ojc/340111
[10] C. Shorten, T.M. Khoshgoftaar. J. Big. Data, **6**, 60 (2019). https://doi.org/10.1186/s40537-019-0197-0
[11] L. Nanni, G. Maguolo, M. Paci. Ecolog. Inform., **57**, 101084 (2020). https://doi.org/10.1016/j.ecoinf.2020.101084
[12] Q. Wen, L. Sun, F. Yang, X. Song, J. Gao, X. Wang, H. Xu. Comp. Sci. Machine Learning, (2020). https://doi.org/10.48550/arXiv.2002.12478
[13] И.В. Исаев, С.А. Буриков, Т.А. Доленко, К.А. Лаптинский, С.А. Доленко. *Повышение устойчивости нейросетевого решения обратной задачи спектроскопии комбинационного рассеяния света к искажениям, обусловленным смещением частоты каналов спектра.* IV Междунар. конф. и молодежная школа „Информационные технологии и нанотехнологии“ (ИТНТ-2018) (Новая техника, Самара, 2018), с. 2710–2715. [I.V. Isaev, S.A. Burikov, T.A. Dolenko, K.A. Laptinskiy, S.A. Dolenko. *Improving the resilience of neural network solution of inverse problems in Raman spectroscopy to the distortions caused by frequency shift of the spectral channels* (Samara, 2018), p. 2710–2715.]
[14] D.P. Kingma, M. Welling. Foundations and Trends in Machine Learning, **12** (4), 307 (2019). http://dx.doi.org/10.1561/2200000056

[15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio. Commun. ACM, **63** (11), 139 (2020). https://doi.org/10.1145/3422622
[16] J. Li, A. Madry, J. Peebles, L. Schmidt. *On the Limitations of First-Order Approximation in GAN Dynamics*, Proceed. 35th Intern. Conf. Machine Learning, PMLR, **80**, 3005-3013 (2018).
[17] M. Shanker, M.Y. Hu, M.S. Hung. Omega, **24** (4), 385 (1996). https://doi.org/10.1016/0305-0483(96)00010-2
[18] A. Efitorov, T. Dolenko, K. Laptinskiy, S. Burikov, S. Dolenko. Proceed. Sci., **410**, art. 013 (2021). https://doi.org/10.22323/1.410.0013
[19] A. Efitorov, S. Burikov, T. Dolenko, S. Dolenko. Studies in Computational Intelligence, **1064**, 557 (2023). https://doi.org/10.1007/978-3-031-19032-2_56
[20] S. Kullback, R.A. Leibler. Annals Mathemat. Statistics, **22**, 79 (1951).
[21] K. Sohn, X. Yan, H. Lee. *Learning structured output representation using deep conditional generative models Advances in Neural Information Processing Systems*, ed. C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, R. Garnett (Curran Associates, Inc. **28**, 2015)
[22] Qing-Song Xu, Yi-Zeng Liang. Chemometrics and Intelligent Laboratory Systems, **56** (1), 1 (2001). https://doi.org/10.1016/S0169-7439(00)00122-2
[23] I. Isaev, S. Burikov, T. Dolenko, K. Laptinskiy, A. Vervald, S. Dolenko. (2018). *Joint Application of Group Determination of Parameters and of Training with Noise Addition to Improve the Resilience of the Neural Network Solution of the Inverse Problem in Spectroscopy to Noise in Data. In: V. Kůrková, Y. Manolopoulos, B. Hammer, L. Iliadis, I. Maglogiannis. (eds) Artificial Neural Networks and Machine Learning — ICANN* 2018. (Lecture Notes in Computer Science, **11139**, 435 (2018) Springer, Cham.)
[24] D.P. Kingma. (2014). *Adam: A method for stochastic optimization* arXiv preprint. https://doi.org/10.48550/arXiv.1412.6980