09

Comparing Approaches to Improving Representativity of Spectroscopic Data using Variational Autoencoders

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

[1]Skobeltsyn Institute of Nuclear Physics,
Moscow State University,
119991 Moscow, Russia
[2]Department of Physics, Moscow State University,
119991 Moscow, Russia
e-mail: anastasemusa@gmail.com

Received February 7, 2025
Revised February 7, 2025
Accepted February 7, 2025.

Solving inverse problems in optical spectroscopy of multicomponent solutions to determine component concentrations is a complex task. One effective approach to solving this problem is the use of artificial neural networks. However, one of the major challenges in this approach is the limited representativity of experimental data, due to the complexity and high cost of large-scale physical experiments.

In this paper, algorithms for generating additional model data using variational autoencoders are explored and compared to enhance the representativity of the training dataset. The results show that the most promising approach is the use of a standard (unconditioned) variational autoencoder, generating patterns from the uniform distribution in the latent space. Further research should focus on identifying the optimal distribution in the latent space for generating patterns.

Keywords: data generation, inverse problem of spectroscopy, multicomponent solutions, artificial neural networks.

DOI: 10.61011/TP.2025.05.61133.15-25

Introduction

Currently, environmental pollution by heavy metals represents a global issue. Key sources of such pollution include the metallurgical industry, mining, improper waste disposal, transport, agriculture, and thermal power plants. Significant threats arise from aerial deposition from both stationary sources and vehicles; contamination of water resources due to industrial wastewater discharge into water bodies; sewage sludge; ash, slag, ore, and slurry waste dumps; as well as oil spills and brine leaks in oil extraction areas [1].

Pollution of water resources is rapidly increasing worldwide, making it a critical issue in many regions globally. Heavy metals in water primarily exist as ions and, unlike organic pollutants, do not undergo biodegradation. Once released into the environment, they can accumulate in water, soil, and living organisms.

Environmental scientists have identified a group of metals that are particularly hazardous to human and animal health, including cadmium, copper, arsenic, nickel, mercury, lead, zinc, and chromium. When heavy metals accumulate in organisms at high concentrations, they can disrupt nearly all biological systems, causing toxic, allergenic, carcinogenic, and gonadotropic effects [2].

Therefore, monitoring heavy metal ion concentrations in wastewater, natural water bodies, and groundwater is critically important.

Various methods exist for determining heavy metal concentrations in water, including chromatography [3], amperometry [4], potentiometry [5], flow injection analysis [6], voltammetry [7], capillary ion analysis [8], and others. However, no single method currently satisfies all the desired criteria of high accuracy, selectivity, rapid analysis, low cost, and minimal operator expertise requirements simultaneously. For instance, selectivity can be achieved by using target-specific reagents (e.g., in flow injection analysis), but this increases operational costs. Chromatography provides high analytical accuracy, but its application is limited by long analysis times, expensive equipment, and complex experimental procedures.

Optical spectroscopy methods [9] offer a balanced compromise, meeting key demands for speed, cost-effectiveness, simplicity, accuracy, and selectivity. Optical spectroscopy is widely used for ion concentration measurements in solutions. Its advantages include non-invasiveness, rapid analysis, and minimal sample preparation. Additionally, the method is cost-efficient, as major expenses are associated primarily with method development rather than routine implementation. Currently, the most widely used optical techniques are absorption spectroscopy and Raman spectroscopy.

However, when spectroscopy is used to determine the concentrations of solution components, it becomes necessary to solve an inverse problem (IP). Due to the high complexity of such problems — often ill-posed in nature, with a complex dependence of the spectral shape on physical parameters and the lack of an adequate physical model for analytical description — there exists no sufficiently accurate

mathematical model capable of predicting the spectrum of a solution with given component concentrations. Therefore, only approximation-based approaches are suitable for solving spectroscopic IPs, including machine learning methods such as artificial neural networks (ANNs).

The application of machine learning methods requires a high-quality, representative dataset composed of a large number of patterns. However, acquiring such datasets is often challenging in spectroscopic applications due to the cost-, labor-, and time-intensive nature of experiments. These experiments require complex instrumentation, and both sample preparation and data preprocessing demand the involvement of skilled personnel.

To mitigate the difficulty of obtaining sufficient training data, standard data augmentation techniques have been proposed and successfully applied in many domains. These techniques are data-type dependent and include noise addition, geometric transformations, color correction, and scaling for images [10]; pitch and speed modifications for audio data; time stretching or compression [11]; and time shifts for time series data [12], among others.

Unfortunately, such standard augmentation techniques are not readily applicable to problems involving more complex and domain-specific data. In the case of spectroscopic data considered in this study, conventional augmentation methods fall short. Accurate interpolation is infeasible due to the highly nonlinear and intricate relationship between optical density and the concentrations of multiple components. Incorporating data from open-access sources or independent experiments is also problematic, as domain adaptation is required to account for variations arising from differences in experimental setups, instrumentation, and protocols — suitable external datasets are scarce. Adding noise to experimental spectra may improve model robustness to noise, but does not enhance the overall quality of the inverse problem solution [13].

In such cases, one can turn to methods that expand existing datasets with synthetic patterns generated by neural network-based generative models — most notably, variational autoencoders (VAEs) [14] and generative adversarial networks (GANs) [15].

While GANs have demonstrated impressive performance in many generative tasks, they are often less interpretable and more difficult to train due to issues such as vanishing gradients, training instability, and mode collapse [16].

The present study focuses on exploring the capabilities of variational autoencoders for generative modeling of spectroscopic data.

Variational autoencoders (VAEs) adopt a probabilistic approach to information encoding, enabling the construction of efficient latent representations of data. These representations can then be leveraged to transform the data in ways that are more suitable for downstream processing with artificial neural networks (ANNs).

Although synthetic data generated by a VAE do not introduce fundamentally new information — since the generative model is trained on the existing dataset — their use may still yield beneficial effects for several reasons. First, due to the architectural bottleneck inherent in VAEs, which compresses input data into a lower-dimensional latent space, the model can act as a denoiser, reducing the influence of noise in the training data. Second, generated patterns can help mitigate class imbalance and other distributional defects in the original dataset, which in turn may lead to improved performance during the training of machine learning models.

The goal of the present work is to test the hypothesis that expanding the training dataset with VAE-generated synthetic patterns can improve its representativity by reshaping the underlying data distribution, thereby reducing the error in solving the inverse problem.

Previous studies [17] have shown that even data standardization — which preserves the shape of the distribution but alters its statistical properties — can improve the stability of neural networks and enhance solution quality.

This study investigates several strategies for influencing the distribution of target parameters through generative modeling. One approach involves generating additional synthetic patterns from a normal distribution to reshape the combined dataset toward normality. This technique addresses the issue that the training set consists of spectra corresponding to solutions where each component concentration takes on only a few discrete values with fixed intervals — a parameter „grid“. A second strategy explored here involves generating additional patterns from a uniform distribution. This is intended to improve data coverage in underrepresented regions — particularly the „tails“ of the distribution — and thus enhance the model's generalization ability by mitigating imbalances in the original dataset.

It is important to note that these considerations apply to the distributions in the latent space rather than in the space of original input features or target variables. The autoencoder is not required to preserve the shape of data distributions when mapping between latent and observable spaces. Its role is to construct a compact representation that efficiently encodes information about the input data. Nevertheless, modifying the distribution of data in latent space can indirectly improve the structure and representativity of data in the original and target domains. Such improvements can positively impact the training of neural networks, potentially resulting in better overall performance on the inverse problem. Evaluating whether this effect can be achieved is the primary focus of the present work.

In this study, we address an eight-parameter inverse problem in spectroscopy, which involves the simultaneous determination of the concentrations of multiple ions in aqueous solutions of heavy metal salts. The target analytes include metal ions: Cu^{2+}, Ni^{2+}, Fe^{3+}, Zn^{2+}, Li^{+}, the NH_4^+ ion, and and acid residues SO_4^{2-} and NO_3^-, based on their optical absorption spectra. We explore methods for improving dataset representativity using variational autoencoders (VAEs).

This work builds on our previous studies. In [18], we investigated the use of conditioned VAEs for training set expansion in a classification problem. In [19], conditioned VAEs were employed in conjunction with a latent-space sampling procedure inspired by the partial least squares (PLS) method. In the present study, we compare strategies based on both conditioned and standard (non- conditioned) VAEs, using latent-space sampling from different prior distributions.

1. Working Principle of Variational Autoencoders

1.1. VAE

An autoencoder is a fully connected neural network comprising two main components: an encoder, which maps information from the original data space into a lower-dimensional space called the latent space, and a decoder, which reconstructs the data from the latent space back to its original form. The effectiveness of this transformation is ensured by the loss function used during training, which minimizes the discrepancy between the input and output of the autoencoder.

The architecture of a variational autoencoder (VAE) (Fig. 1) is similar, with two key distinctions: (1) a specialized loss function allows the model not only to minimize the reconstruction error between the input and output, but also to ensure that the latent representations of the data follow a distribution close to a multivariate normal distribution; (2) instead of producing a direct representation of a specific pattern, the encoder outputs the parameters of a normal distribution in the latent space that best describes the training data.

The loss function for a VAE consists of two terms: the first term, the mean squared error between the input and reconstructed output, enforces data reconstruction; the second term, the Kullback-Leibler divergence D_{KL} [20], penalizes deviations between the distribution of latent variables ω_D $(h_1, h_2, \ldots)$ and a multivariate normal distribution ω_N $(h_1, h_2, \ldots)$.

Thus, the total loss function L for these networks is defined as:

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

where $\{x_i\}$ denotes the input vectors, $\{y_i\}$ — the reconstructed output vectors, D_{KL} is the Kullback-Leibler divergence, ω_N represents the multivariate normal prior distribution, ω_D is the distribution of the latent representations, and N is the number of patterns in the dataset.

It is important to note that training a network to solve an inverse problem requires knowledge of the target variable values corresponding to each input spectrum in the training set. For the original experimental spectra, this information is known a priori — for example, if the solutions were specifically prepared and the true component concentrations are known, or if the concentrations were determined by an independent analytical method. However, for spectra generated using a VAE, the target variables are not inherently available. Therefore, an additional method is required to assign appropriate target values to each generated spectrum.

1.2. cVAE

In a conditioned variational autoencoder (cVAE), the decoder receives not only the latent representation of the data but also the corresponding target variables of the inverse problem — i. e., the component concentrations that the generated spectrum is expected to represent (Fig. 2).

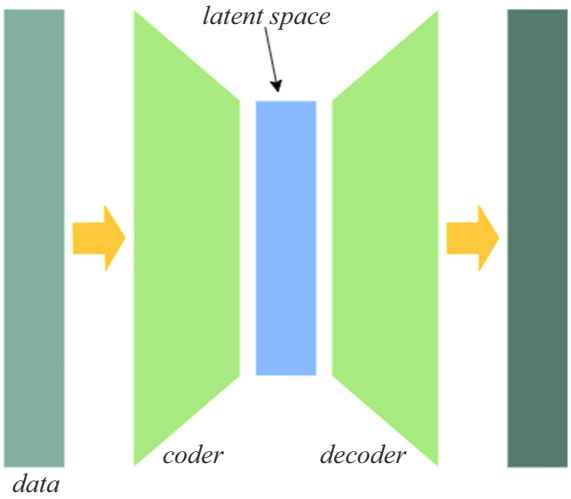

Figure 1. VAE.

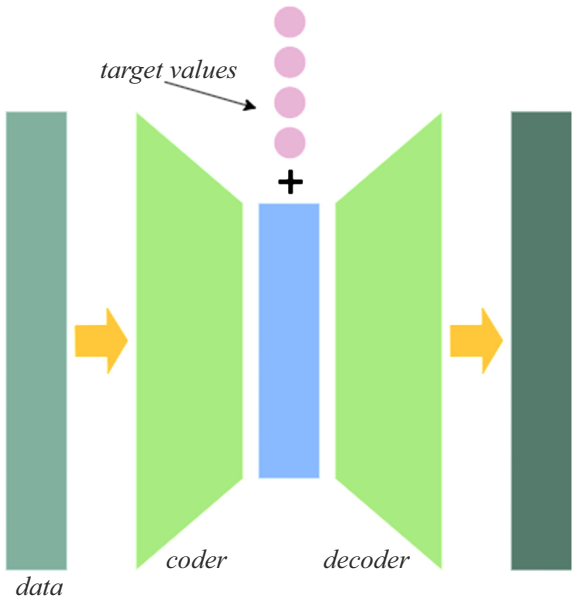

Figure 2. cVAE

This conditioning mechanism allows the cVAE to generate spectra associated with predefined target values [21], which is particularly valuable when using the generated data to train models for solving inverse problems. As a result, the challenge inherent to standard (unconditioned) VAEs — assigning target values to generated spectra — is no longer present.

Furthermore, conditioning enables the generation of patterns with target variable values chosen to shape the distribution of the dataset directly in the target space, rather than in the latent space. Provided the decoder operates reliably, this allows for more precise control over the representativity and balance of the training data in terms of the desired physical parameters.

2. Physical experiment

The physical experiment conducted to obtain the data for this study involved measuring the optical absorption spectra of aqueous solutions containing various combinations and concentrations of the following 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_4NO_3.

2.1. Spectrophotometry of multicomponent aqueous solutions of inorganic salts. Registration of optical absorption spectra

The spectra were recorded using a Shimadzu UV-1800 spectrophotometer in the spectral range of 190–1000 nm with a 1 nm step size and a slit width of 1 nm. Measurements were performed in a thin cuvette (optical path length: 1 mm) using distilled water as the reference. Thus, each sample — represented by an optical absorption spectrum — was characterized by 911 features corresponding to the absorbance values at specific wavelengths (spectral channels).

Figure 3 shows the optical absorption spectra of selected aqueous solutions.

2.2. Dataset

A total of 3.744 aqueous solutions of inorganic salts were prepared using distilled water. These solutions contained various concentrations of the following ions: Zn^{2+}, Cu^{2+}, Li^+, Fe^{3+}, Ni^{2+}, NH_4^-, SO_4^{2-}, and NO_3^-. The concentrations of each ion ranged from 0 up to a maximum value specified in Table 1, with varying step sizes. The selected concentration ranges reflect the typical ion concentration intervals found in industrial water systems used in non-ferrous metallurgy facilities.

Table 2 presents information on the number of one-, two-, three-, and so on, component solutions, i. e., solutions containing the specified number of components.

Thus, the experimental dataset consists of 3744 optical absorption spectra of multicomponent aqueous solutions containing various combinations and concentrations of the following ions: Zn^{2+}, Cu^{2+}, Li^+, Fe^{3+}, Ni^{2+}, NH_4^+, SO_4^{2-}, and NO_3^-. Each pattern is described by 911 features.

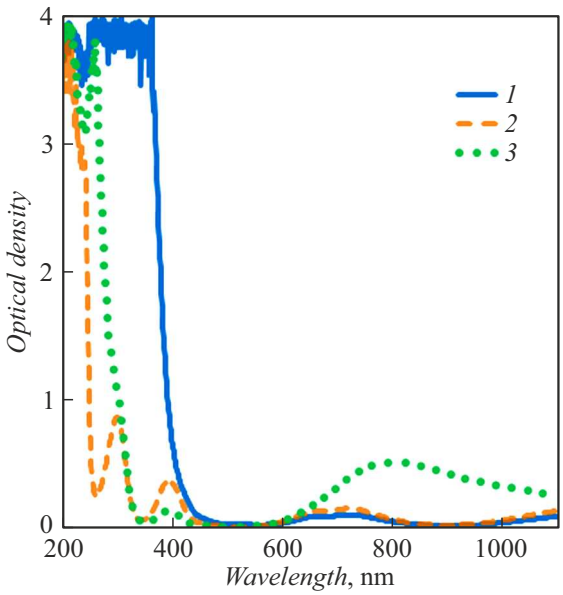

Figure 3. Examples of optical absorption spectra of the studied aqueous solutions. The presented spectra correspond to solutions with the following ion concentrations in mol/L (M): Spectrum 1: Zn^{2+} – 0.23, Cu^{2+} – 0, Li^+ – 0.12, Fe^{3+} – 0.22, Ni^{2+} – 0.49, NH_4^+ – 0, SO_4^{2-} – 0.49, NO_3^- – 1.23. Spectrum 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. Spectrum 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.

Table 1. Characteristics of datasets.

Ion	Maximal concentration, M	Number of samples 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

3. Computational Experiment

3.1. Cross-validation

The original dataset was split into training, validation, and test subsets in a 7:2:1 ratio, resulting in the set sizes presented in Table 3.

Table 2. The number of one-, two-, three-, and so on, -component solutions.

The number of components	The number of solutions
1	24
2	240
3	1080
4	1590
5	726
6	84

Table 3. Sizes of the training, validation, and test datasets.

Dataset	The number of patterns
training	2620
validation	749
test	375

To ensure a robust evaluation of the inverse problem solution quality, all experiments employed eight-fold random resampling for cross-validation, following the Monte Carlo approach [22].

In all experiments involving the generation of additional data, the number of generated spectra was chosen so as to double the size of the experimental training dataset, i. e., 2,620.

3.2. Preprocessing of Input Data

To facilitate effective training of the regression neural networks and to accelerate convergence during optimization, a normalization procedure was applied to the optical absorption spectra as a preprocessing step. Specifically, the absorbance values in each spectral channel were rescaled to the range [0,1] [26].

The lower bound of zero is physically justified by the fact that absorbance values are inherently non-negative; the absence of absorption at a given wavelength corresponds to a zero value in the corresponding spectral channel.

For each feature x in the dataset, the minimum $x_{\min}$ and maximum $x_{\max}$ values were computed. Each feature was then transformed according to the following equation:

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

The concentration values of the ions were retained in their original form for all patterns.

3.3. Output Dimensionality Reduction

In this study, the inverse spectroscopy problem was formulated as a regression task and solved using artificial neural networks — specifically, multilayer perceptrons acting as universal function approximators. A standard approach for reducing output dimensionality in multi-parameter regression tasks was adopted: autonomous determination [23], in which a separate model is trained for each target variable. Consequently, eight regression neural networks were trained in total — one for each ion considered in the problem.

3.4. Increasing Dataset Representativity via Pattern Generation

In this study, to improve the performance of inverse spectroscopy problem solutions using neural networks, we apply data representativity enhancement techniques based on training set expansion through variational autoencoders (VAEs).

Artificial spectra are generated using a VAE and subsequently added to the experimental training set. Regression neural networks are then trained on the expanded dataset.

3.4.1. Reference IP Solution — Training Regression Neural Networks on Experimental Data Only

For each of the eight components — Zn^{2+}, Cu^{2+}, Li^{+}, Fe^{3+}, Ni^{2+}, NH_4^{+}, SO_4^{2-}, NO_3^{-} — a regression neural network in the form of a multilayer perceptron was trained using only the experimental data.

3.4.2. Expanding Dataset with VAE

The first approach involved training regression neural networks (NNs) on a dataset expanded using a standard (non-conditioned) variational autoencoder (VAE).

The dataset expansion algorithm using a VAE consists of the following steps:

Step 1. Training reference regression NNs autonomously to estimate ion concentrations on the original training set of experimental optical absorption spectra.

Step 2. Training the VAE on the original training set of experimental spectra.

Step 3. Generating synthetic patterns using the VAE decoder from random vectors in the latent space, sampled according to the parameters of the multivariate normal distribution learned by the VAE encoder. If the normalized optical density in any spectral channel was negative, it was set to zero; if it exceeded 1, it was clamped to 1.

Step 4. Determining ion concentrations corresponding to the generated spectra using the reference regression NNs trained in Step 1. If the predicted concentration of any ion fell outside the permissible range, the pattern was discarded. The minimum allowed concentration was zero, and the maximum was the highest observed concentration of that ion in the entire experimental dataset.

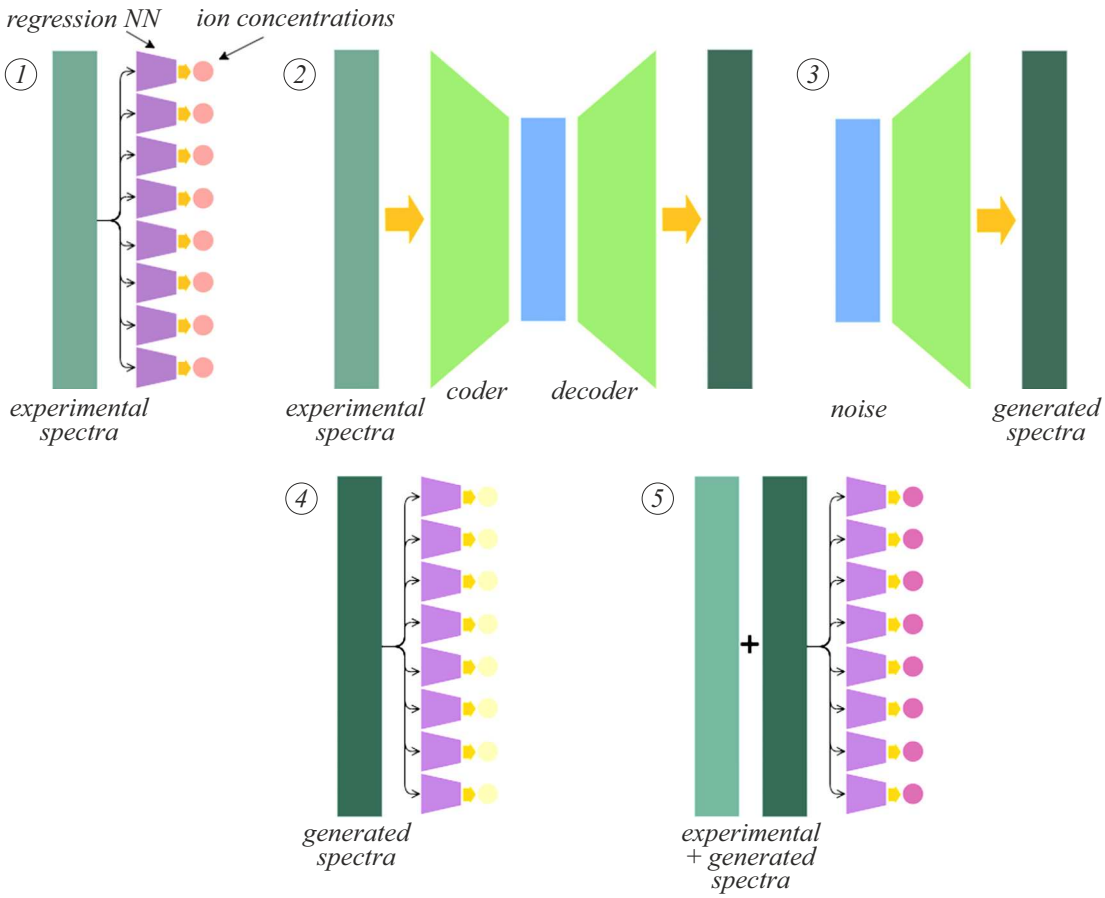

Figure 4. Computational workflow for dataset expansion using VAE: *1*. Training reference regression NNs on experimental spectra; *2*. Training VAE on experimental spectra; *3*. Using the VAE decoder to generate synthetic spectra from latent vectors sampled from the target distribution; *4*. Applying reference NNs (trained in Step 1) to estimate target parameters for generated spectra; *5*. Training regression NNs on the expanded dataset (doubled in size by adding synthetic spectra generated in Step *3* to the experimental spectra).

Step 5. Training regression NNs autonomously on the expanded training set (double the original size), consisting of half of experimental spectra and an equal number of VAE-generated spectra following the described procedure.

Additionally, an alternative version of Step 3 was tested, where latent vectors were sampled from a uniform distribution instead of a normal distribution.

The computational workflow for dataset expansion using the VAE is illustrated in Fig. 4.

3.4.3. Expanding Dataset with cVAE

The algorithm for dataset expansion using a conditioned Variational Autoencoder (cVAE) consists of the following steps:

Step 1. Training the cVAE on the original training set of experimental optical absorption spectra. During training, the corresponding known values of target parameters (ion concentrations) associated with these spectra are used.

Step 2. Generating synthetic patterns using the cVAE decoder from random vectors in the latent space, sampled according to the parameters of the multivariate normal distribution in the latent space learned by the cVAE encoder. For generating additional patterns, concentration sets already present in the experimental training dataset were selected. If the normalized optical density in any spectral channel was negative, it was set to zero; if greater than 1, it was clamped to 1.

Step 3. Training regression neural networks autonomously to predict ion concentrations on the expanded training set. This doubled-size dataset consists of half of experimental spectra and an equal number of spectra generated by the cVAE following the described procedure.

The computational workflow for dataset expansion using the conditioned VAE is illustrated in Fig. 5.

3.5. Model Parameters

3.5.1. Architecture of Regression Neural Networks

Multilayer perceptrons (MLPs) were used as regression neural networks. Each MLP architecture consisted of two hidden layers with 64 and 16 neurons, respectively, followed by a single-neuron output layer. The hidden layers employed

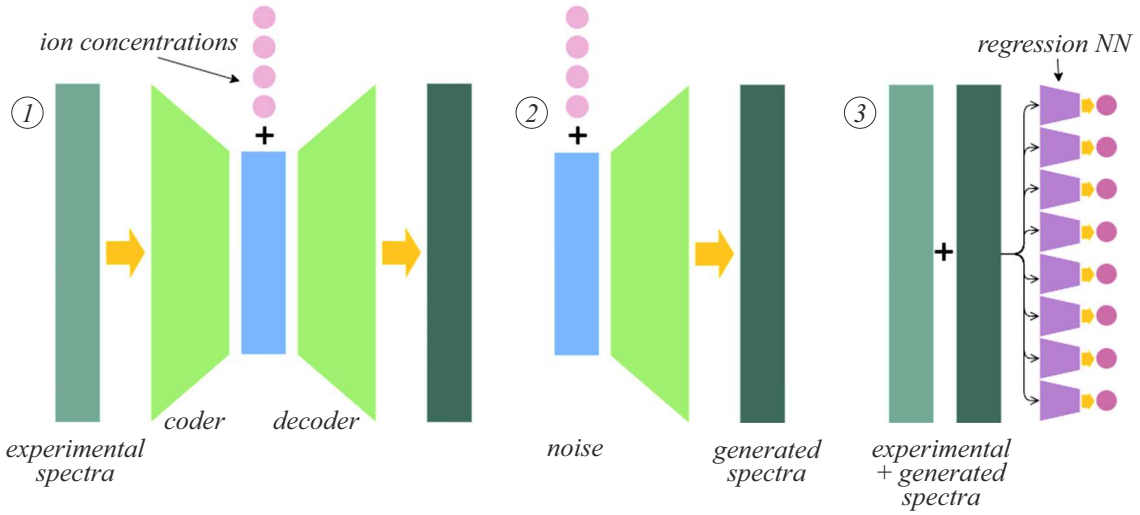

Figure 5. Computational workflow for dataset expansion using a conditioned VAE (cVAE): *1*. Training the cVAE on the training set of experimental spectra; *2*. Using the cVAE decoder to generate synthetic spectra from latent vectors with the target distribution, corresponding to specified concentration sets; *3*. Training regression NNs on the expanded dataset (doubled in size by adding synthetic spectra generated in Step 2 to the experimental spectra).

sigmoid activation functions, while the output layer used a linear activation function.

3.5.2. VAE Architecture

Encoder: The encoder is implemented as a multilayer perceptron (MLP) with a single hidden layer. The hidden layer consists of 256 neurons with sigmoid activation, while the output layer contains $2 \times LS = 2 \times 91$ neurons with linear activation. Here, LS denotes the dimensionality of the latent space. The output of the encoder is a two-dimensional vector of length $LS = 91$, where one component represents the mean μ and the other the variance σ^2 of a normal distribution along each coordinate of the latent space.

Decoder: The decoder is also an MLP with a single hidden layer. The input layer contains $LS = 91$ neurons with sigmoid activation. The hidden layer has 256 neurons with sigmoid activation, and the output layer consists of 911 neurons with linear activation. The input to the decoder is a one-dimensional latent vector $h = v^g_{noise} \times \sigma + \mu$, of length $LS = 91$, where v^g_{noise} is a random vector of dimensionality $LS = 91$, sampled from a standard normal distribution.

3.5.3. cVAE Architecture

Encoder: The encoder is implemented as a multilayer perceptron (MLP) with a single hidden layer. The hidden layer consists of 256 neurons with sigmoid activation, while the output layer contains $2 \times LS = 2 \times 91$ neurons with linear activation. Here, LS denotes the dimensionality of the latent space. The output of the encoder is a two-dimensional vector of length $LS = 91$, where one component represents the mean μ and the other the variance σ^2 of a normal distribution along each coordinate of the latent space.

Decoder: The decoder is also an MLP with a single hidden layer. The input layer contains $LS + 8 = 91 + 8$ neurons with sigmoid activation. The hidden layer has 256 neurons with sigmoid activation, and the output layer consists of 911 neurons with linear activation. Here, LS denotes the dimensionality of the latent space, 8 — the number of target parameters (i. e., the number of ions to be predicted). The input to the decoder is a one-dimensional latent vector $h = v^g_{noise} \times \sigma + \mu$, of length $LS = 91$, where v^g_{noise} is a random vector of dimensionality $LS = 91$, sampled from a standard normal distribution and an 8-dimensional vector containing the specified set of concentrations.

3.5.4. Model Training Parameters

In all neural network training experiments, the stopping criterion was defined as reaching 250 epochs after the last decrease in the validation loss. The Adam optimization algorithm [24], a variant of stochastic gradient descent, was used with a learning rate of 0.001 and a batch size of 64 patterns.

The loss function for the regression neural networks was the mean squared error (MSE), defined as:

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

where y_i^{true} is the ground truth value for the i-th pattern; y_i^{pred} is the output of the network for the i-th pattern; N — dataset size.

The loss function for variational autoencoders is the sum of the mean squared error (MSE) and the Kullback-

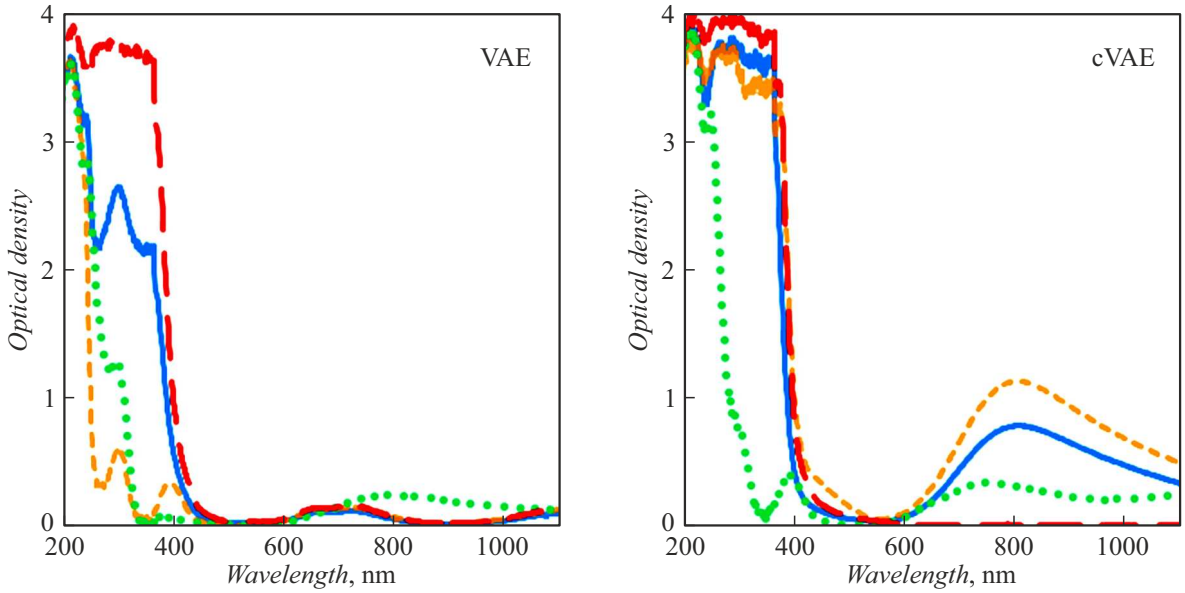

Figure 6. Examples of spectra generated by VAE and cVAE.

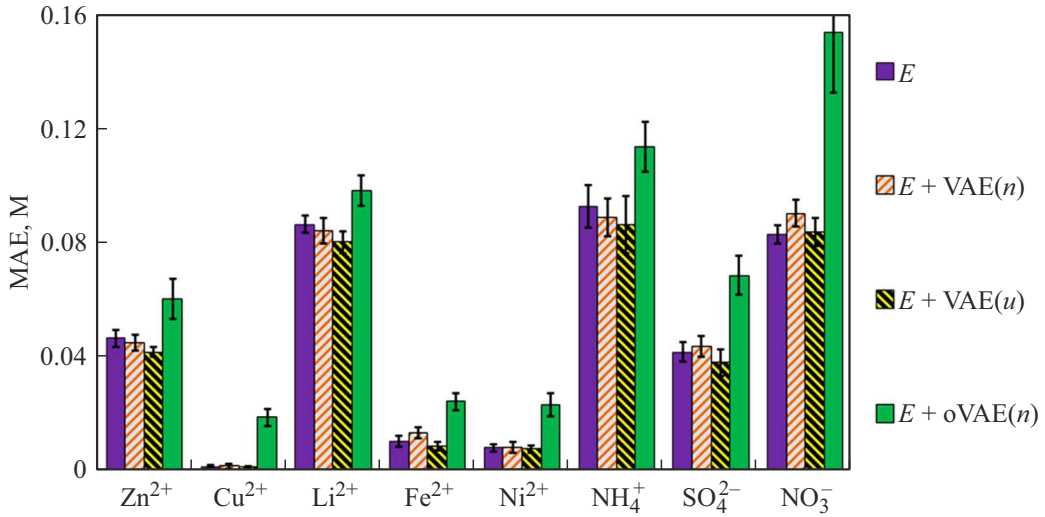

Figure 7. Error metrics obtained solving the inverse spectroscopy problem using the reference method (E) and with dataset expanded using: – VAE with sampling from a normal distribution in the latent space (E+VAE(n)), – VAE with sampling from a uniform distribution in the latent space (E+VAE(u)), and – cVAE with sampling from a normal distribution (E+cVAE(n)).

Leibler (KL) divergence D_{KL}:

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

where $\{x_i\}$ — input vector, $\{y_i\}$ — output vector, D_{KL} — Kullback-Leibler (KL) divergence, ω_N — normal distribution, ω_D — data distribution in the latent space, N — dataset size.

4. Results

To evaluate the quality of the IP solution, the Mean Absolute Error (MAE) was used. MAE calculates the average absolute difference between the predicted values of the target variable (network outputs) and the true values. MAE is defined by the equation:

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

where y_i^{true} is the ground truth value for the i-th pattern; y_i^{pred} is the output of the network for the i-th pattern; N — dataset size.

4.1. Generated Spectra

The spectra generated using both VAE and conditioned VAE (cVAE) exhibit visual similarity to the experimental

spectra (Fig. 6). Furthermore, the generated spectra demonstrate reduced noise levels in the short-wavelength region compared to their experimental counterparts.

4.2. Regression Task accuracy

The error metrics obtained for solving the inverse spectroscopy problem using the reference method and with dataset expansion using synthetic spectra generated by VAE and cVAE are presented in Fig. 7.

These results indicate that the performance of regression neural networks using VAE-generated synthetic spectra sampled from a normal distribution in the latent space remains essentially unchanged compared to the reference method, within the limits of experimental uncertainty. The inclusion of synthetic patterns in the training set did not lead to any statistically significant improvement or degradation in the model's accuracy.

On the other hand, generating synthetic spectra from a uniform distribution in the latent space yields, on average, better results than generation from a normal distribution. For Zn^{2+} and Li^{+} ions, a statistically significant improvement over the reference method was observed.

In contrast, the use of a conditioned VAE leads to an increase in test error, in some cases substantially. This issue may potentially be mitigated by refining the strategy for selecting concentration sets used during generation.

Conclusion

In this study, we address the inverse problem of spectroscopy for multicomponent aqueous solutions of salts containing eight ions: Zn^{2+}, Cu^{2+}, Li^{+}, Fe^{3+}, Ni^{2+}, NH_4^{+}, SO_4^{2-}, and NO_3^{-}. The problem is approached using artificial neural networks. We investigate the possibility of expanding the experimental spectroscopic training dataset using variational autoencoders (VAEs). The objective is to improve the accuracy of the inverse problem solution relative to a reference model trained solely on the original experimental data.

We propose and describe algorithms for dataset expansion based on both standard and conditioned variational autoencoders (cVAEs). Using trained VAEs, we produce synthetic spectra that are subsequently combined with experimental spectra during the training of regression neural networks designed to solve the inverse problem.

To test the hypothesis that the proposed methods can improve dataset representativity by altering data distribution, a series of experiments were conducted involving the training of regression neural networks on different training sets: (1) the original experimental dataset; (2) a dataset expanded using a conditioned variational autoencoder (cVAE); and (3) datasets expanded using a standard (non- conditioned) variational autoencoder (VAE) with pattern generation from either a normal or a uniform distribution in the latent space. Accuracy metrics for the inverse problem solution were obtained for each case.

The cVAE-based dataset expansion approach resulted in decreased accuracy compared to the reference method. The method using a VAE with latent-space sampling from a normal distribution produced results comparable to the reference model. In contrast, VAE-based dataset expansion approach with uniform sampling in the latent space led to a slight but statistically significant reduction in prediction error for the Zn^{2+} and Li^{+} ions on the test set. However, for most target components, the accuracy metrics of networks trained on expanded dataset remained within the error bounds of the reference metrics.

These results suggest that the proposed approach holds promise, but further investigation is needed to optimize the conditions and parameters of the computational experiment. In particular, future work should focus on identifying optimal latent-space sampling strategies for generating synthetic spectra and effectively expanding the training dataset.

Funding

This study has been conducted at the expense of the Russian Science Foundation, grant no. 24-11-00266, https://rscf.ru/en/project/24-11-00266/.

Conflict of interest

The authors declare that they have no conflict of interest.

References

[1] Yu.N. Vodyanitsky, D.V. Ladonin, A.T. Savichev. *Zagryaznenie pochv tyazhelymy metallami* (Tipographia Rosselkhozakademii, M., 2012 (in Russain))
[2] G.A. Teplaya. Astrakhan bulletin of ecological education, **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] 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