Supporting Information

Raman Spectroscopy Algorithm Based on Convolutional Neural Network and Multilayer Perceptron: Qualitative and Quantitative Analysis of Chemical Warfare Agent Simulants

Jie Wu[a,#], Fei Li[b,#], Jing-Wen Zhou[d], Hongmei Li[d], Zilong Wang[a], Xian-Ming Guo[d], Yue-Jiao Zhang [d], Lin Zhang [c*], Pei Liang [a*] , Shisheng Zheng [d*] and Jian-Feng Li[d, e*]

[a] College of Optical and Electronic Technology, China Jiliang University, Hangzhou 310018 China

[b] School of Optoelectronics, University of Chinese Academy of Sciences, Beijing 101408, China

[c] Institute of Chemical Defense, Academy of Military Sciences, Beijing 102205, China

[d] State Key Laboratory of Physical Chemistry of Solid Surfaces, iChEM, College of Chemistry and Chemical Engineering, College of Energy, College of Materials, College of Electronic Science and Engineering, College of Physical Science and Technology, Fujian Key Laboratory of Ultrafast Laser Technology and Applications, Xiamen University, Xiamen 361005, P. R. China

[e] Innovation Laboratory for Sciences and Technologies of Energy Materials of Fujian Province (IKKEM), Xiamen 361005, P. R. China

[*]Correspondence: zhanglin_zju@aliyun.com; plianghust@gmail.com; zhengss@xmu.edu.cn; li@xmu.edu.cn

# These authors contribute equally

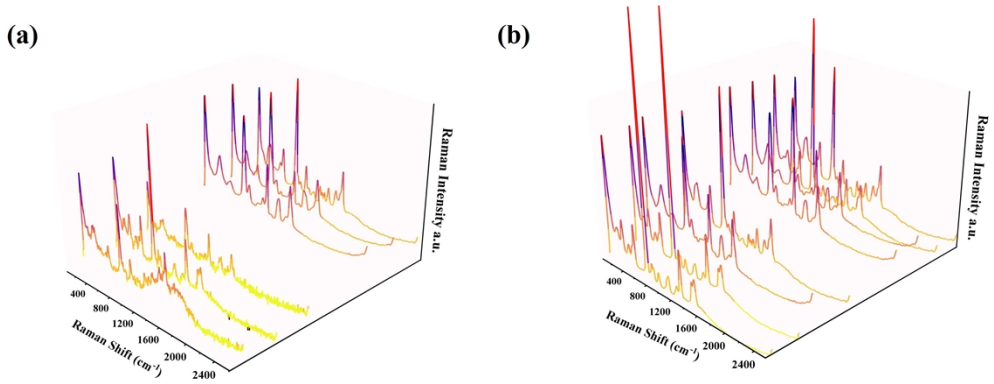

Figure. S1. Spectral comparison of pure substances and mixtures and Dataset. Figure S1 (a) shows the comparison between the pure material data and the simulated ternary mixture data. Figure S1 (b) shows the spectra of some binary and ternary data sets.

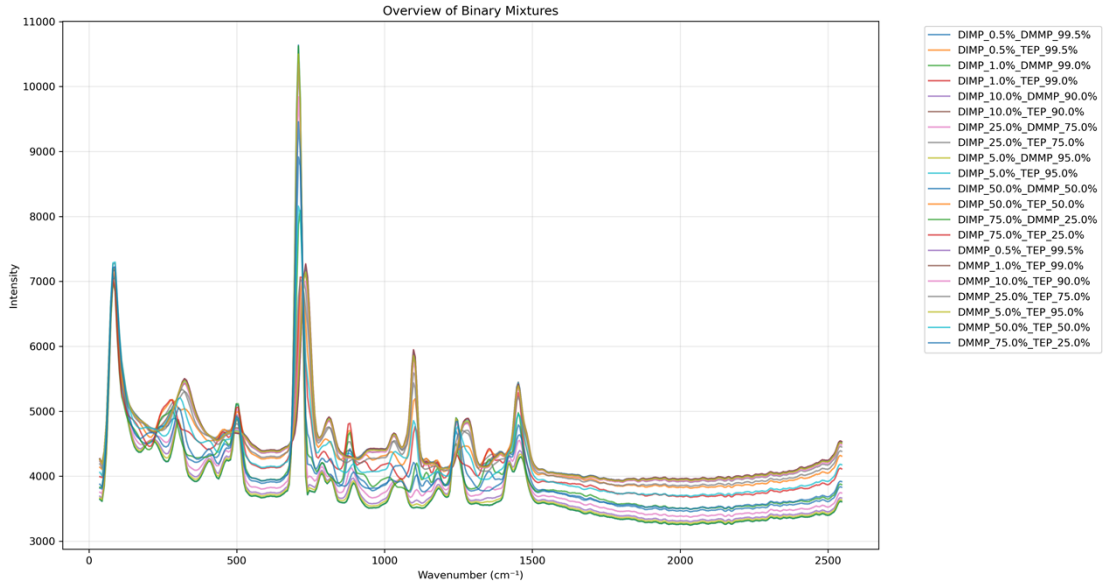

Figure. S2. Demonstration of mixed concentration ratios and spectra for all binary mixtures

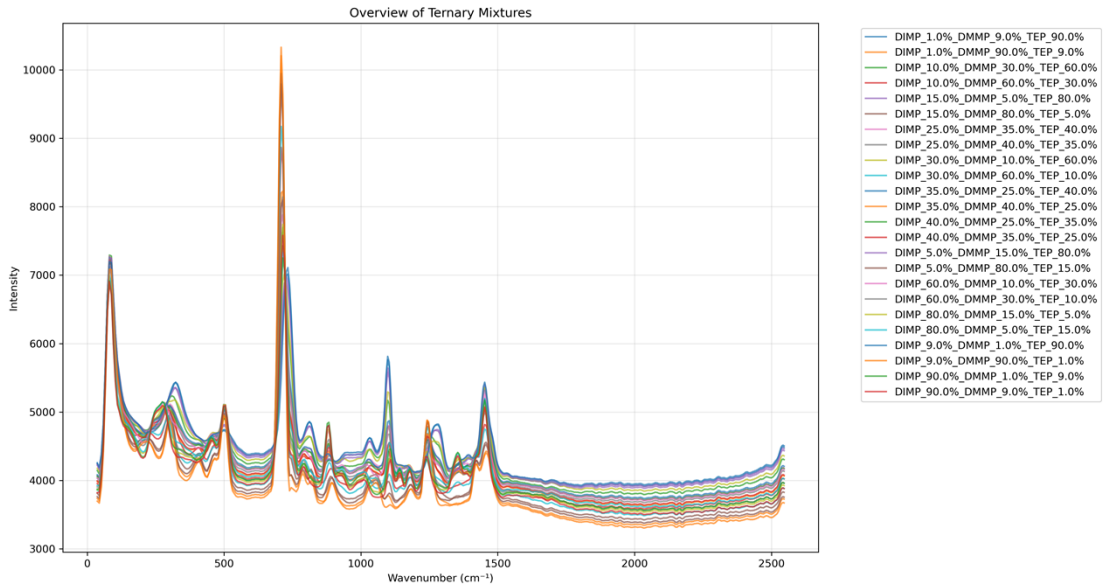

Figure. S3. Demonstration of mixed concentration ratios and spectra for all ternary mixtures

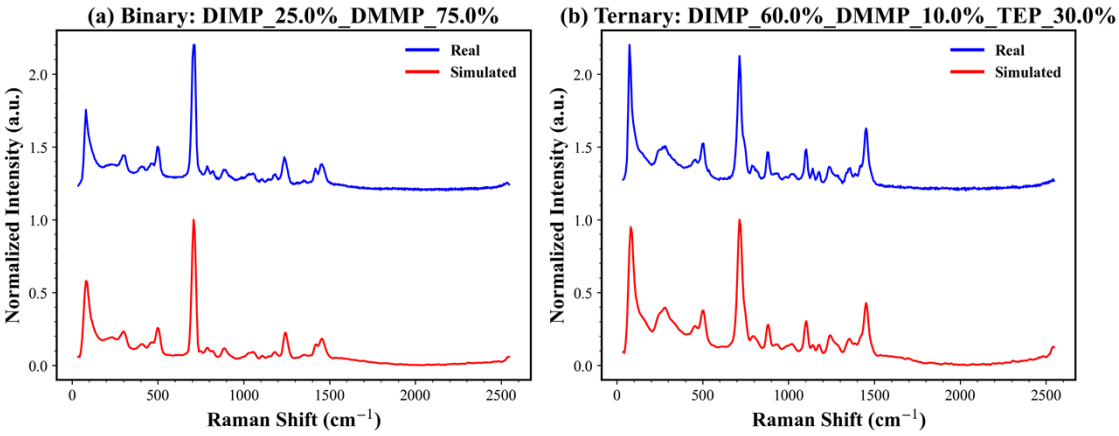

Figure. S4. Mixed probe molecular Raman spectroscopy under different conditions. (a) and (b) show the real spectra of binary mixtures against the simulated hybrid spectra, respectively

In Figure S4, the simulated mixture data is compared with the experimentally collected mixture data. As shown in Figure S4(a), the binary mixture demonstrates strong consistency in both peak positions and intensities. Figure S4(b) presents the comparison for ternary mixtures, where the number and positions of peaks are generally consistent. However, a closer examination reveals minor discrepancies: slight deviations in peak widths in the 700–800 cm^{-1} range and differences in the intensity and relative relationships of smaller peaks in the 1100–1200 cm^{-1} range. Despite these minor imperfections, the overall agreement remains high. These results confirm that the simulated spectra align closely with the experimentally collected spectra, validating that mixed spectra can be approximately described as linear superpositions of pure material spectra. However, considerations such as spectral broadening and noise filtering are necessary for simulations to better approximate real data. This validation provides valuable insights into the application of deep learning in Raman spectroscopy. Simulated data effectively supplement experimental data, addressing the issue of data scarcity and supporting the development of more robust and accurate deep learning models. In addition, by generating diverse and controlled mixed spectral data, the training effect of the neural network can be enhanced to make it more suitable for complex spectral analysis tasks, including the identification and quantitative analysis of chemical warfare agent simulants and their interactions.

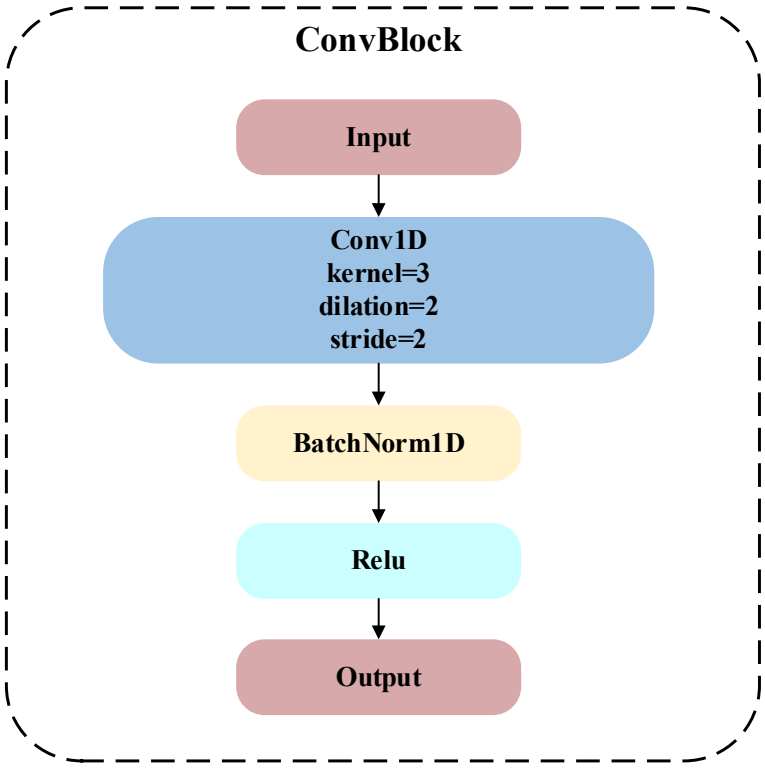

Figure. S5. Block diagram of ConvBlock structure

Figure S5 Structure of the ConvBlock module. The block consists of a 1D convolutional layer (Conv1D) with kernel size 3, dilation rate 2, and stride 2, followed by batch normalization (BatchNorm1D), and ReLU activation. This architecture allows for efficient feature extraction and downsampling of the input signal.

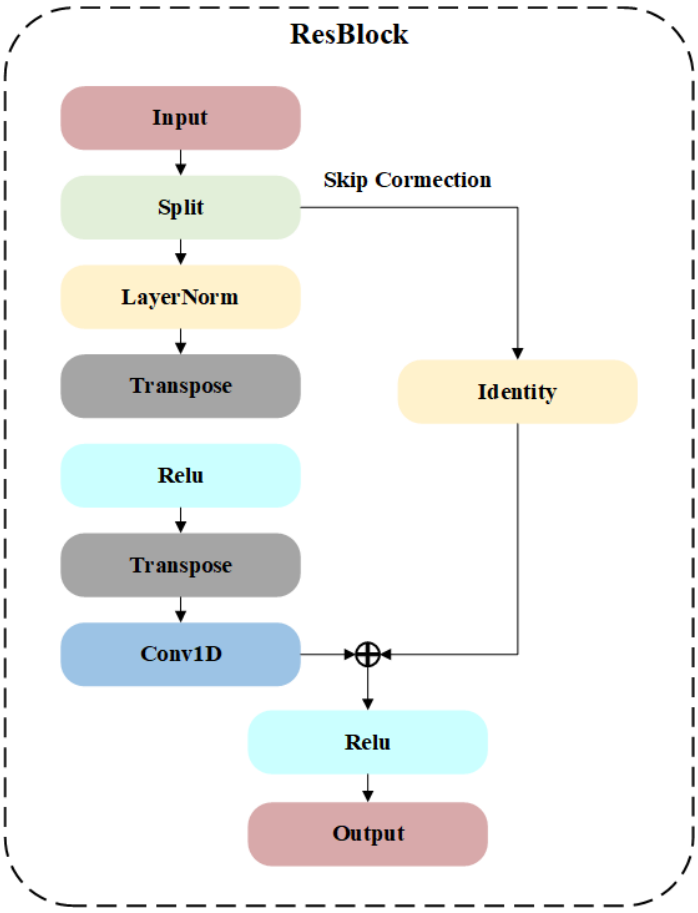

Figure. S6. Schematic representation of the ResBlock architecture

The diagram illustrates the flow of data through multiple layers: starting with an input layer, followed by a split operation that creates a skip connection pathway. The main branch consists of LayerNorm, two Transpose operations interleaved with a ReLU activation, and a Conv1D layer. The skip connection maintains signal identity and merges with the main branch through addition. The final output passes through a ReLU activation layer. This residual block design facilitates gradient flow and enables deeper network architectures while mitigating the vanishing gradient problem.

The architecture demonstrates modern deep learning design principles, combining residual learning with normalization and convolution operations for effective feature transformation while maintaining gradient flow through the network.

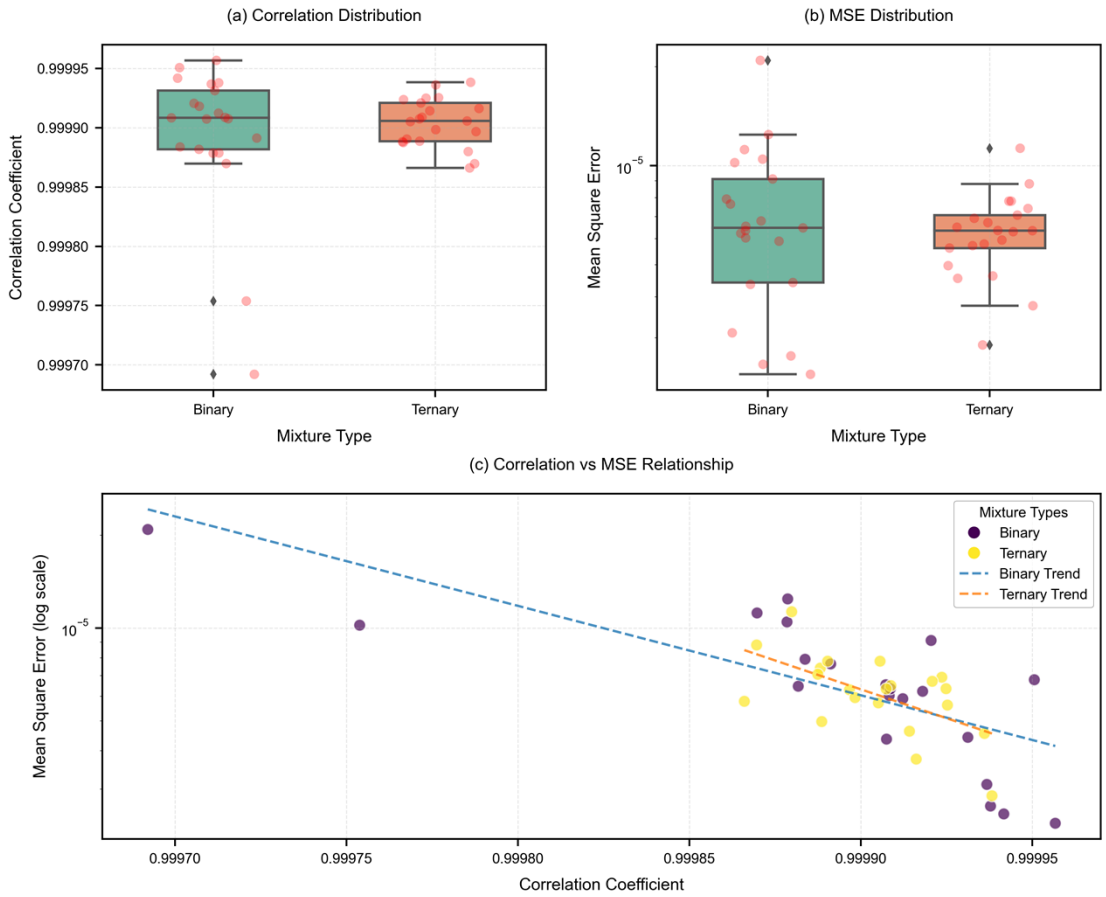

Figure. S7. GAN network evaluation: correlation and mse analysis for binary and ternary mixtures

The box plots in (a) show that both binary and ternary mixtures achieve extremely high correlation coefficients (>0.9997), demonstrating the superior learning ability of GAN networks, with binary mixtures having slightly higher median values. Panel (b) presents MSE distributions on a logarithmic scale, showing similarly low error rates for both mixture types, with ternary mixtures performing slightly better. The scatter plot in (c) reveals an inverse relationship between MSE and correlation coefficient, where higher correlations correspond to lower errors. These results highlight the GAN network's robust and accurate performance across varying system complexities.

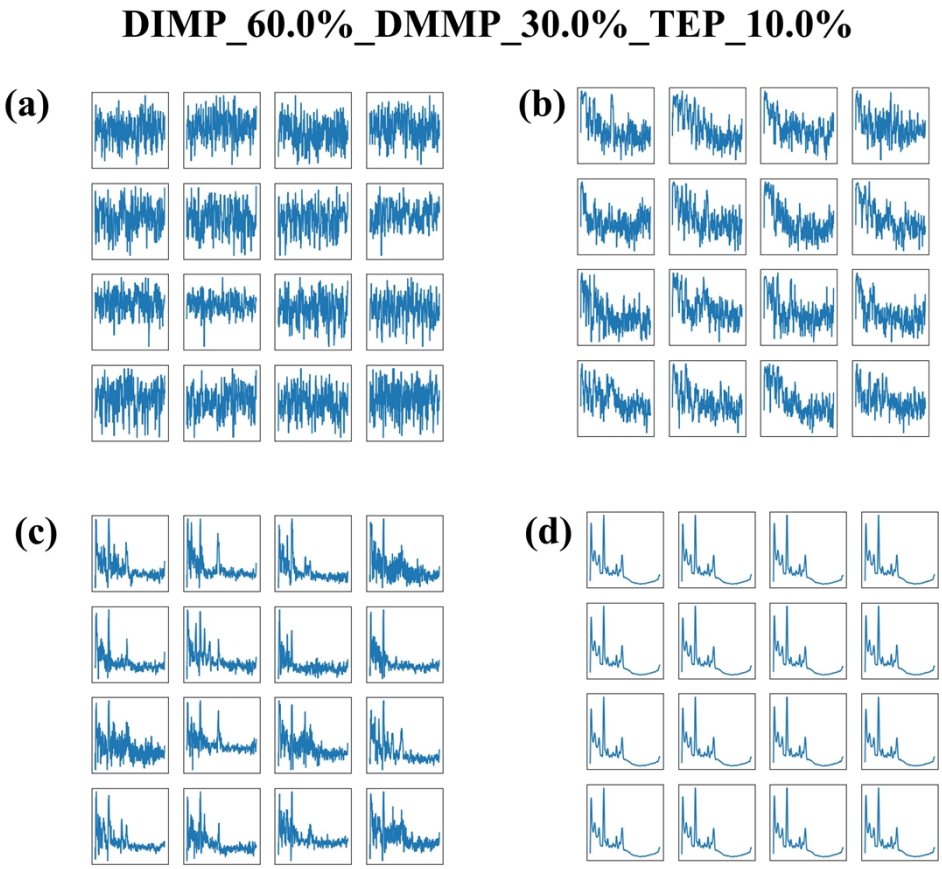

Figure. S8. GAN training process for binary mixtures. (a) Initial training stage showing noisy, unstructured spectral outputs from the generator; (b) Intermediate training stage demonstrating emerging spectral features and pattern formation; (c) Advanced training stage showing more refined spectral characteristics; (d) Final training stage outputs exhibiting well-defined peak features and stable spectral patterns.

The training process and final performance of the GAN framework for binary mixtures are comprehensively illustrated in Figure S8. Subplot (a) shows the generator's output during the initial training stages. At this phase, the generated spectra exhibit significant noise and lack well-defined peak structures, reflecting the generator's early learning phase in capturing the complex characteristics of multi-component spectra. As training progresses, subplot (b) demonstrates the intermediate training stage, where the generated spectra start to show emerging spectral features. The spectral patterns begin to take shape, indicating that the GAN is gradually learning the underlying spectral characteristics of the ternary mixture.

In the advanced training stage, as depicted in subplot (c), the spectral features become more refined and stable, showing improved consistency across different generations. Finally, in subplot (d), the generated spectra display well-defined peak features and highly stable spectral patterns, demonstrating the GAN's successful learning of complex three-component spectral relationships and its ability to generate high-quality spectral data for ternary mixtures.

DIMP_60.0%_DMMP_30.0%_TEP_10.0%

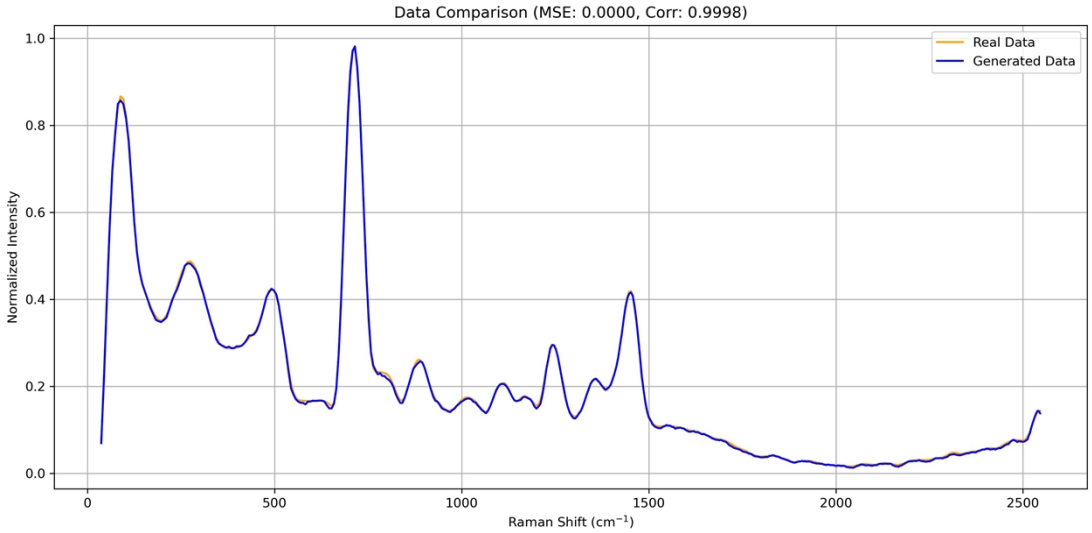

Figure. S9. GAN comparison of generated and real data for binary mixtures

Figure S9 provides a direct comparison between the normalized Raman spectra of real and GAN-generated data. The nearly overlapping curves, with a correlation coefficient of 0.9998 and an MSE of 0.0000, emphasize the model's remarkable accuracy in replicating real spectral features.

These results confirm the robustness and effectiveness of the GAN framework in generating high-quality spectral data for binary mixtures, addressing the challenges of limited experimental datasets while ensuring predictive accuracy and reliability.

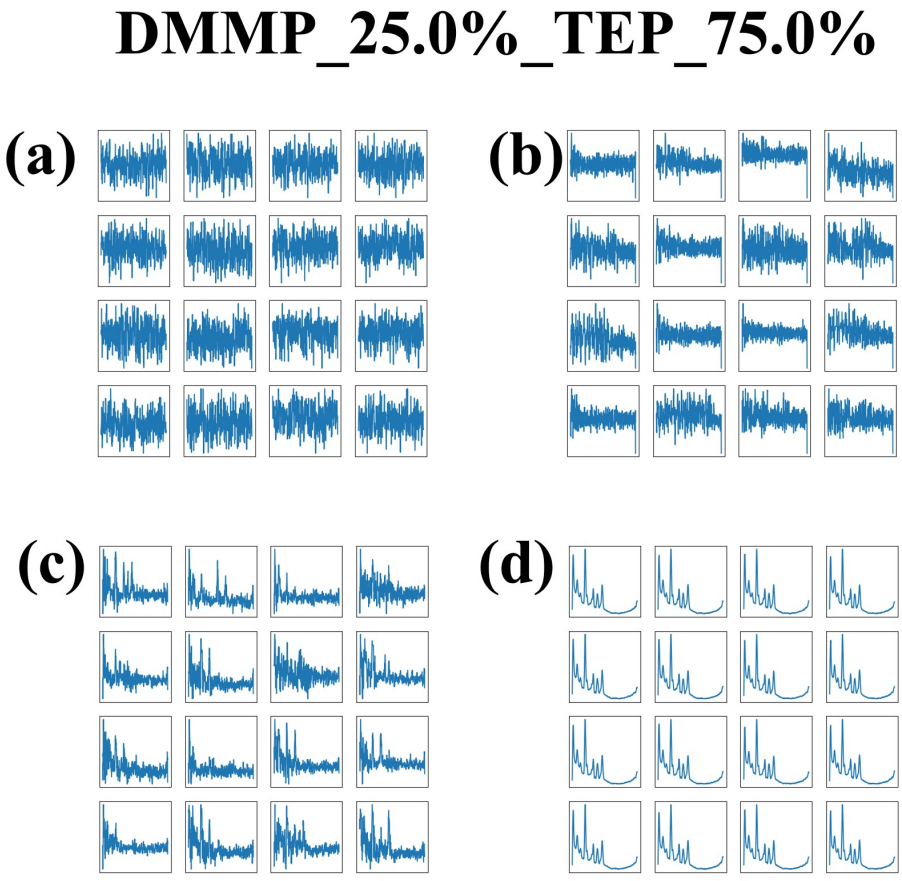

Figure. S10. GAN training process for ternary mixtures

DMMP_25.0%_TEP_75.0%

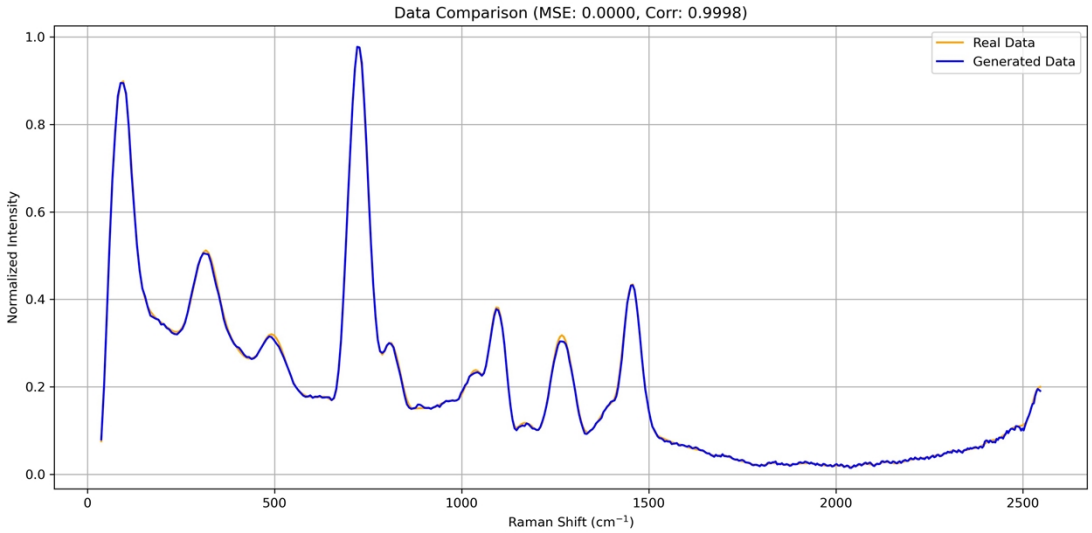

Figure. S11. GAN comparison of generated and real data for ternary mixtures

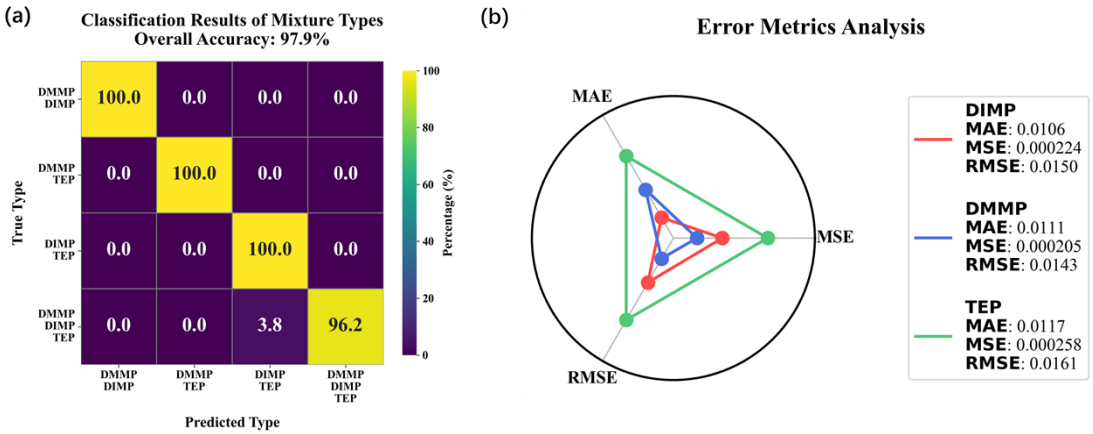

Figure. S12. Evaluation chart of the effectiveness of models enhanced with traditional data

We re-amplified the data using translation and compression transformations, and followed the same modeling approach to train the model, obtaining model evaluation plots for the traditional data augmentation approach.

As shown in Figure S12, the classification results and error metrics of the traditional data augmentation method were compared with the results obtained using the spectra generated by WGAN-GP, as presented in Figure 4 of the manuscript. The classification accuracy of the traditional data augmentation method was 97.9%, with a misclassification rate of approximately 3.8% for the DMMP-DIMP-TEP mixture. The confusion matrix in Figure S12(a) shows that the misclassifications mainly occurred between the DMMP-DIMP and DMMP-TEP categories. In contrast, the spectra generated by WGAN-GP achieved perfect classification results, with an overall accuracy of 100%. This was due to our adjustment of the GAN parameters, allowing it to closely approximate the original spectra. After multiple attempts, the best training epoch interval was selected, which ensured that only the intensity was altered while the Raman shift remained nearly unchanged, and each spectrum was different. This effectively simulated machine errors during data acquisition.

The error metrics analysis further clarifies the differences between the two methods. As shown in the radar plot in Figure S12(b), the traditional data augmentation method shows higher Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) compared to the radar plot of the WGAN-generated spectra in Figure 4(b), indicating superior data fidelity and model performance.

These results highlight the advantages of using spectra generated by WGAN over traditional data augmentation techniques, as the former not only improves classification accuracy but also minimizes error metrics, providing a more reliable and robust model for predicting mixture types. This comparison emphasizes the potential of WGAN to generate high-quality synthetic data that better reflects the complex spectral patterns inherent in the original mixtures, without introducing artifacts or distorting key features such as peak intensity.

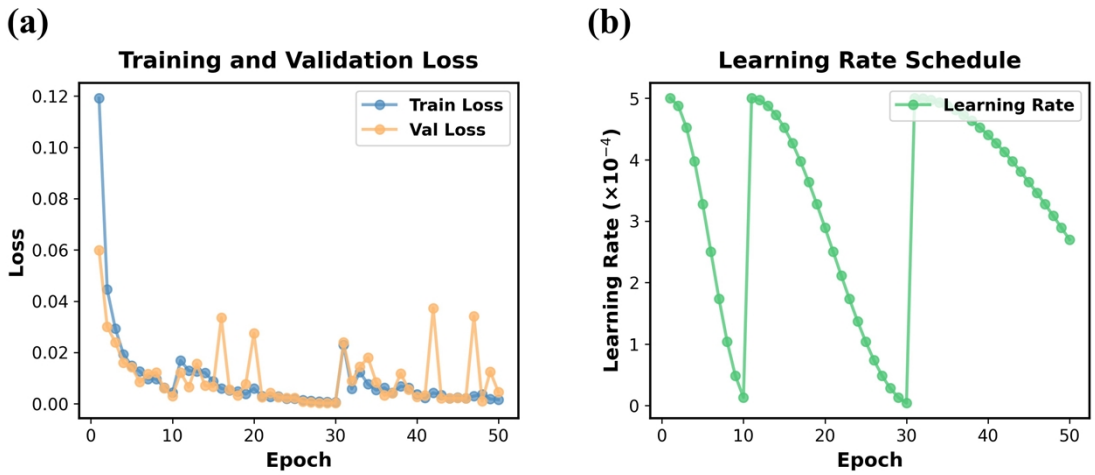

Figure. S13. Training and validation loss curves and learning rate schedule

The loss plot illustrates the training and validation loss curves across 50 epochs. Both the training loss (blue) and validation loss (orange) decrease significantly during the initial epochs, indicating that the model is learning effectively. By approximately the 10th epoch, the losses stabilize, suggesting convergence. Despite occasional fluctuations in validation loss, the gap between training and validation losses remains minimal, demonstrating the model's strong generalization capability and absence of overfitting.

The learning rate plot depicts the learning rate schedule applied during training. The learning rate (green) follows a cyclic pattern, decreasing gradually within each cycle and resetting to a higher value at the start of a new cycle. This approach helps the model escape local minima and ensures stable convergence by allowing periodic exploration of the loss surface. The schedule is particularly effective in optimizing the model's performance over the entire training process.

These results highlight the robustness of the training process, with an efficient balance between loss minimization and learning rate adjustment.

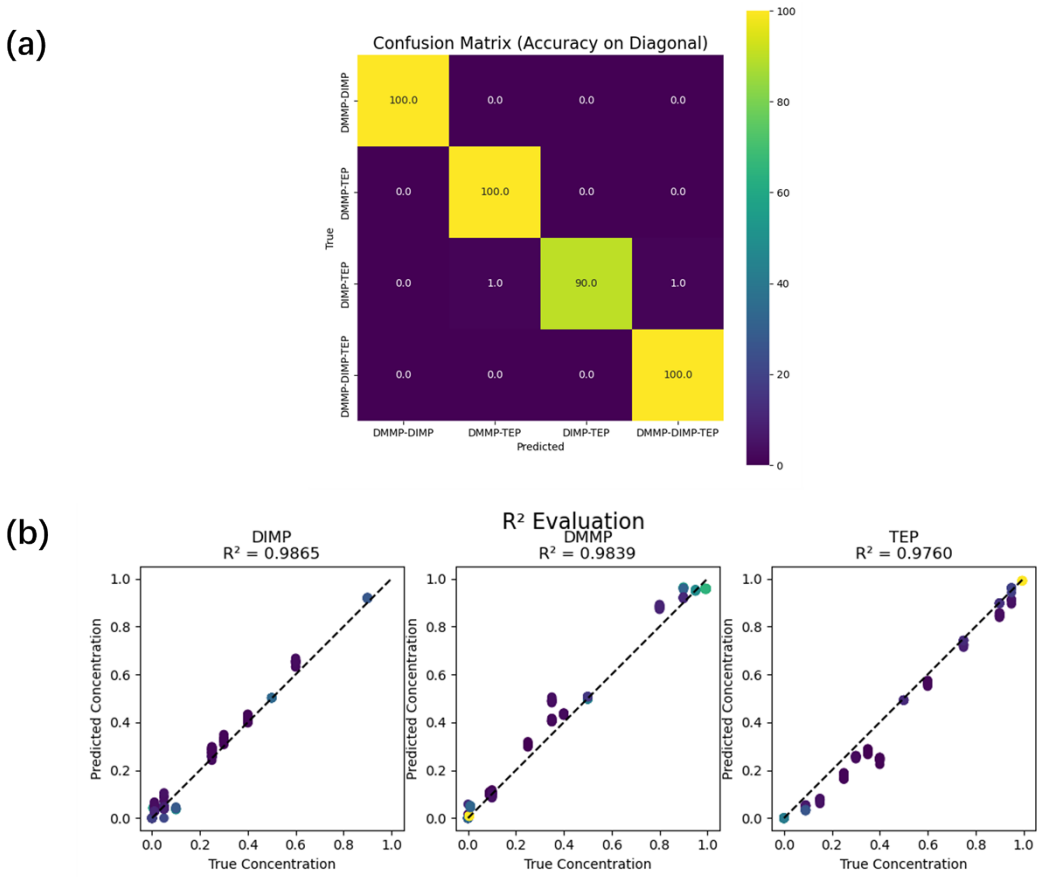

Figure. S14. Test evaluation diagram of real mixture data. (a) The confusion matrix for training a model on real mixture data testing simulated data is shown. (b) The R^2 plot of the training model is shown for testing simulated data with real mixture data

The evaluation results of the model trained on simulated data and the model tested on real mixture data are shown in Figure S14. Figure (a) shows the confusion matrix, which shows that the model effectively predicts the correct concentration for most of the real mixture data. The prediction accuracy for each category is very high, meaning that the model successfully distinguishes between different mixtures. At the same time, it also reflects the shortcomings of the simulation hybrid algorithm, and the interaction simulation of DIMP+TEP combination is not accurate enough. This also points the way for our next research.

Figure (b) shows the R^2 curves for DMMP, DMMP-TEP, and TEP mixtures, showing a strong linear relationship between predicted and true concentrations. The R^2 values for DMMP, DMMP-TEP, and TEP are 0.9865, 0.9839, and 0.9760, respectively, confirming that the model has good predictive power even when applied to actual mixed data. These results show that the model trained on the simulated data can be effectively generalized to real-world applications, showing reliable performance in predicting concentrations in real-world mixture scenarios.

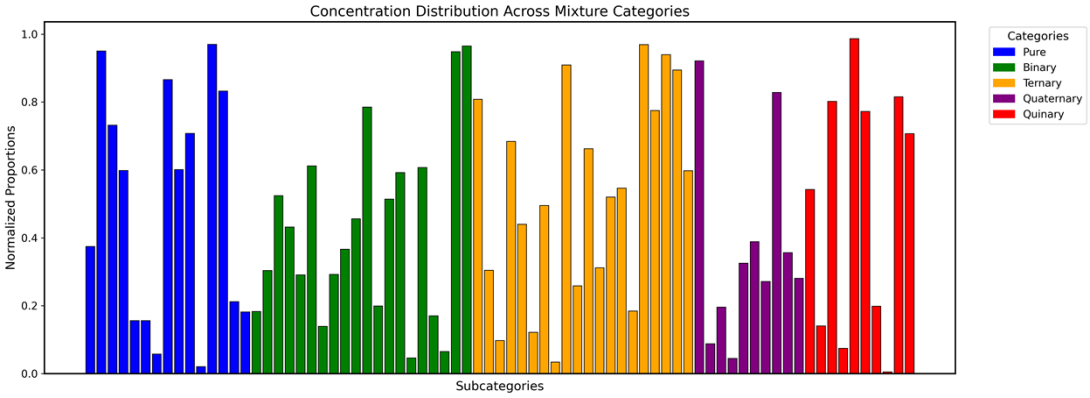

Figure. S15. Concentration distribution across mixture categories. 75 The mixtures consist of five major groups of substances in varying proportions, containing both extreme and homogeneous proportions.