MUSE: MIXTURE-TO-SUBSTRUCTURE ELUCIDATION VIA DISENTANGLED SET PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Infrared (IR) spectroscopy is a crucial technique for molecular characterization in chemistry and materials science. Existing deep learning-based IR spectral analysis methods are largely confined to the idealized single-compound scenario, with limited exploration of more realistic multi-component mixtures. This challenge stems from two main issues: signal coupling and attribution ambiguity caused by overlapping absorption peaks, and the entanglement between a component's chemical structure and its concentration in the spectral features. To overcome these challenges, we introduce **MUSE**, an end-to-end framework that reformulates **M**ixt**u**re-to-**S**ubstructure **E**lucidation as a one-shot set prediction problem. By formulating a permutation-equivariant decoding scheme via learnable slot queries and cross-attention, the model simultaneously captures the physical additivity and global competition among components in a single forward pass. This joint modeling significantly mitigates the error accumulation that plagues traditional single-target methods. To overcome the intrinsic entanglement between chemical identity and abundance, MUSE partitions each slot embedding into structural and concentration subspaces, penalizing their empirical cross-covariance to encourage linear decorrelation. Building upon this latent disentanglement, it incorporates conditional flow matching during the generative process to provide physically interpretable supervision through generative reconstruction. Extensive experiments on real and simulated datasets demonstrate that MUSE significantly outperforms existing methods, establishing its effectiveness and robustness in complex mixture analysis. Our code can be found at https://anonymous.4open.science.

1 INTRODUCTION

Infrared (IR) spectroscopy is a fundamental technique for molecular characterization in chemistry, drug discovery, and materials science Siddique (2024); Jackson et al. (1997); Rijs & Oomens (2015); Khan et al. (2018). By mapping molecular absorption intensities across a broad range of wavenumbers, IR spectra encode the vibrational signatures of chemical bonds, offering critical insights into molecular structures Czarnecki et al. (2015); Baiz et al. (2020). Driven by the rapid advancements in deep learning, data-driven IR spectral analysis has recently witnessed remarkable progress Lansford & Vlachos (2020). Across tasks ranging from multi-label functional group recognition Jung et al. (2023); Alberts et al. (2024) to *de novo* generation of molecular structures Alberts et al. (2025b); Wu et al. (2025), deep learning models significantly outperform traditional database retrieval and manual analysis by offering superior computational efficiency and broader generalization across diverse scenarios Sarker (2021); Rane et al. (2024). Consequently, by capturing complex, non-linear spectral patterns, these models demonstrate immense potential to transform IR spectroscopy into a general-purpose tool for automated structural elucidation.

However, nearly all existing methods are confined to the idealized single-compound setting, whereas most real-world samples are multi-component mixtures at varying concentrations, including pharmaceutical formulations, reaction intermediates, and environmental specimens Kidd et al. (2024); Cheng et al. (2024); Nassar et al. (2025). Recovering every component's structure from a single mixture spectrum therefore remains a largely open problem, as does the coarser task of identifying each component's substructures. This gap between the single-compound assumption and realistic samples severely limits the practical applicability of current deep spectral models in chemical analysis Venetos et al. (2024); Wang et al. (2025); Chen & Qiu (2026).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

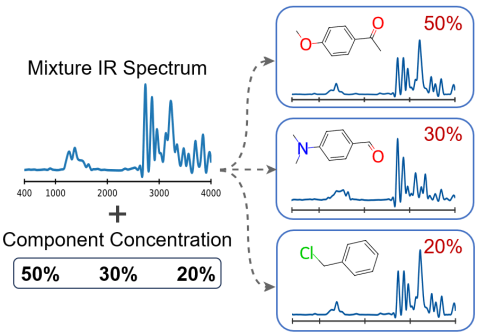

Figure 1: Schematic representation of resolving a mixture IR spectrum for per-component functional group elucidation.

Crucially, this gap stems not merely from the scarcity of structurally annotated mixture spectra, but from two inherent difficulties of the task itself. First, characteristic absorptions of different components may overlap or even cancel within the same fingerprint region, inducing severe signal coupling and attribution ambiguity Baker et al. (2014); Rinnan et al. (2009). Prior attempts Alberts et al. (2025a) cast the task as a single-target conditional prediction that predicts one component at a time, ignoring the physical additivity and global competition across components and leaving per-component predictions isolated and prone to substantial error accumulation. This sequential bottleneck blinds the model to the synergistic interactions that define true multi-component mixtures. Second, within a mixture spectrum, a component's chemical identity is encoded in peak shapes and positions, while its relative abundance is reflected in peak intensities. These two factors share the same spectral features and are further modulated by non-linear effects such as hydrogen bonding. As a result, structural and concentration information are fundamentally difficult to decompose into independent representations, giving rise to a problem of intrinsic disentanglement Tauler (1995).

Here, we propose **MUSE**, an end-to-end framework for **M**ixt**u**re-to-**S**ubstructure **E**lucidation from mixed infrared spectra. Motivated by the unordered nature of chemical mixtures, we recast the elucidation process as a one-shot set prediction problem. Specifically, we employ a permutation-equivariant decoding scheme with learnable slot queries to parallelly extract component-wise representations from a shared spectral encoding. By doing so, MUSE simultaneously captures the physical additivity and global competition among components in a single forward pass. This joint modeling significantly mitigates the isolated prediction and error accumulation that plague traditional single-target paradigms. To disentangle chemical identity from relative abundance, we split each slot embedding into a structural and a concentration subspace and penalize their cross-covariance, encouraging the structural representation to remain stable under concentration shifts. To ground this latent disentanglement in the physical measurement space, we integrate conditional flow matching (CFM) into the generative process. Rather than generating spectra abruptly, CFM models a continuous probability path from mixture spectrum to each pure-component spectrum. By explicitly conditioning this trajectory on the concentration subspace, the generative process physically mirrors the gradual unmixing of overlapping signals, providing rigorous, physically consistent supervision. Consequently, identity-abundance disentanglement is explicitly anchored not only in the geometry of the learned representations but also in the spectral generation process itself. Conditioning this trajectory on the concentration subspace supplies supervision along a target-component enrichment path, anchoring identity–abundance disentanglement in the spectral generation process rather than in the latent geometry alone. As an initial foray into mixture multi-component functional group elucidation, our study primarily focuses on systems with up to 4 components. While seemingly manageable, this setting already exhibits severe signal coupling, allowing us to fundamentally validate the proposed mechanisms before scaling to macroscopic real-world scenarios. Extensive experiments on real and simulated datasets demonstrate that MUSE significantly outperforms existing methods, and its advantage persists under non-ideal conditions such as nonlinear mixing and imperfect priors, and carries over to physically measured mixtures, establishing its effectiveness in complex mixture analysis. Our contributions can be summarized as follows:

- We propose MUSE, the first end-to-end framework that reformulates component-level substructure elucidation from mixed IR spectra as a one-shot set prediction problem.
- We introduce a slot-based decoder that partitions latent embeddings into structural and concentration subspaces, with a cross-covariance penalty that reduces the identity–abundance entanglement.
- We integrate CFM into generative reconstruction, where its continuous probability paths mirror signal unmixing and root latent disentanglement in the spectral generation space.
- Comprehensive evaluations on real and simulated datasets demonstrate that MUSE outperforms existing methods in terms of performance and robustness in multi-component mixture analysis.

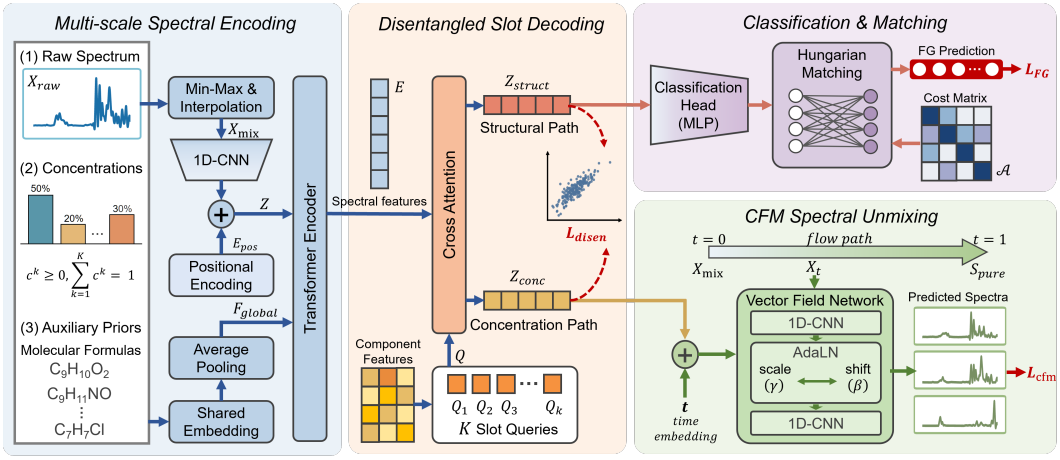

Figure 2: Illustration of the **MUSE** framework. The model takes an infrared spectrum with per-component concentrations and optional formula priors. A CNN encoder extracts spectral features, from which K slot queries aggregate component-wise representations via cross-attention. Each slot embedding is partitioned into a structural subspace for functional-group prediction under Hungarian matching, and a concentration subspace that reconstructs pure-component spectra.

2 Methodology

2.1 Problem Formulation

Let $x_{\text{mix}} \in \mathbb{R}^L$ denote the infrared spectrum of a mixture, where L is the spectral sequence length. We assume that the mixture consists of K latent components with relative physical concentrations $\mathcal{C} = \{c^{(1)}, c^{(2)}, \ldots, c^{(K)}\}$, subject to the constraints

$$c^{(k)} \geq 0, \quad \sum_{k=1}^{K} c^{(k)} = 1. \tag{1}$$

Crucially, since a physical mixture is inherently invariant to the ordering of its components, our objective is to infer an unordered set of per-component functional-groups directly from x_{mix}, optionally conditioned on the concentration prior $\mathcal{C}$. Each component is represented by a multi-label vector $\hat{\mathbf{y}}^{(k)} \in \{0,1\}^G$ over G functional group categories. Thus, the elucidation process is formulated as learning a permutation-equivariant mapping $\mathcal{F}_\Theta$:

$$\mathcal{F}_\Theta : \left(x_{\text{mix}}, \{c^{(k)}\}_{k=1}^K\right) \mapsto \{\hat{\mathbf{y}}^{(k)}\}_{k=1}^K. \tag{2}$$

2.2 Multi-Scale Spectral Encoder

Before feature extraction, the raw spectrum x_{raw} is min–max normalized and interpolated to a uniform length, giving the standardized input $x_{\text{mix}} \in \mathbb{R}^L$. A 1D-CNN with a downsampling branch maps it to deep features $H_{\text{deep}} \in \mathbb{R}^{C_1 \times S}$ (C_1 channels, reduced length S), which are projected to d dimensions and augmented with learned positional encodings E_{pos} to form the spectral sequence $Z \in \mathbb{R}^{S \times d}$:

$$z_i = \text{LayerNorm}(\text{Linear}(H_{\text{deep},i}) + E_{\text{pos},i}), \quad i \in \{1, \ldots, S\}. \tag{3}$$

When formula priors are available, a shared embedding layer followed by sequence-wise average pooling yields $f_k \in \mathbb{R}^d$ for each component. Their mean, F_{global}, is prepended to Z and processed by an L_{enc}-layer Transformer encoder, producing $E \in \mathbb{R}^{(1+S)\times d}$:

$$E = \text{TransformerEncoder}([F_{\text{global}} \parallel Z]), \quad F_{\text{global}} = \frac{1}{K}\sum_{k=1}^{K} f_k \in \mathbb{R}^{1\times d}, \tag{4}$$

where $[\cdot \parallel \cdot]$ denotes concatenation along the sequence dimension.

2.3 DISENTANGLED SLOT DECODING

To disentangle the highly mixed spectral features into individual component representations, we introduce a permutation-equivariant decoding scheme. Instead of sequential extraction, K learnable slot queries $Q \in \mathbb{R}^{K\times d}$ are deployed as independent structural probes. Through cross-attention with the shared context E, these probes parallelly and competitively distill component-wise features.

To reduce identity–abundance entanglement, we partition the slot embeddings $V \in \mathbb{R}^{K\times d}$ into structural and concentration subspaces. For each component, $z^{(k)}_{\mathrm{struct}}$ captures intrinsic structural features, while $z^{(k)}_{\mathrm{conc}}$ captures concentration-dependent spectral variation for downstream generation:

$$V_k = \mathrm{CrossAttn}(Q_k, E, E), \quad z^{(k)}_{\mathrm{struct}} = V_k\left[1:\frac{d}{2}\right], \quad z^{(k)}_{\mathrm{conc}} = V_k\left[\frac{d}{2}+1:d\right]. \tag{5}$$

To discourage information leakage between these subspaces, a covariance-based decorrelation regularizer is introduced:

$$\mathcal{L}_{\mathrm{disen}} = \|\mathrm{Cov}(z_{\mathrm{struct}}, z_{\mathrm{conc}})\|_F^2 = \left\| \frac{1}{B\cdot K-1}\sum_{i=1}^{BK}(z^{(i)}_{\mathrm{struct}} - \bar{z}_{\mathrm{struct}})(z^{(i)}_{\mathrm{conc}} - \bar{z}_{\mathrm{conc}})^\top \right\|_F^2, \tag{6}$$

Here, B is the batch size and $\|\cdot\|_F$ denotes the Frobenius norm. The penalty encourages linear decorrelation between the two subspaces (Appendix B).

2.4 SPECTRAL UNMIXING VIA CONDITIONAL FLOW MATCHING

To reconstruct high-fidelity pure spectra, we formulate spectral unmixing as a CFM process. For component k, CFM interpolates between the mixture x_{mix} and its pure spectrum $s^{(k)}_{\mathrm{pure}}$. For a randomly sampled progression timestep $t \sim \mathcal{U}(0,1)$:

$$x_t = (1-t)x_{\mathrm{mix}} + t s^{(k)}_{\mathrm{pure}}. \tag{7}$$

Substituting the additive mixture model $x_{\mathrm{mix}} = \sum_j c^{(j)} s^{(j)}_{\mathrm{pure}}$ yields:

$$x_t = \sum_{j=1}^{K} c^{(j)}(t)\, s^{(j)}_{\mathrm{pure}}, \qquad c^{(k)}(t) = (1-t)c^{(k)} + t, \qquad c^{(j)}(t) = (1-t)c^{(j)} \;\; (j\neq k), \tag{8}$$

which stays on the probability simplex while the target coefficient rises monotonically to 1. Each x_t is thus an idealized mixture along a target-enrichment path, motivating conditioning on $z^{(k)}_{\mathrm{conc}}$. A vector field network f_θ, parameterized by a 1D-CNN with Adaptive Layer Normalization (AdaLN) blocks, is trained to predict $s^{(k)}_{\mathrm{pure}}$ directly from x_t. The time embedding of t and $z^{(k)}_{\mathrm{conc}}$ are concatenated into a joint condition $u^{(k)} \in \mathbb{R}^d$, which within each residual block is projected by an MLP to scale γ and shift β that modulate the group-normalized features h:

$$\mathrm{AdaLN}(h, u^{(k)}) = \gamma(u^{(k)}) \odot \mathrm{GroupNorm}(h) + \beta(u^{(k)}). \tag{9}$$

Crucially, the final convolutional layer of f_θ is zero-initialized, acting as an identity map at initialization so that early CFM outputs do not inject noise into the functional-group classification gradients. Training uses the x-prediction objective:

$$\mathcal{L}_{\mathrm{cfm}} = \mathbb{E}_{t\sim\mathcal{U}(0,1)} \left\| f_\theta(x_t, t, z^{(k)}_{\mathrm{conc}}) - s^{(k)}_{\mathrm{pure}} \right\|_2^2. \tag{10}$$

2.5 OPTIMIZATION AND CURRICULUM LEARNING

Since the output is a set, Hungarian matching is used to align predictions with ground truth. We incorporate the concentration prior into the matching cost to disambiguate components with similar functional groups:

$$\mathcal{A}_{i,j} = \mathcal{L}_{\mathrm{BCE}}(\hat{\mathbf{y}}^{(i)}, \mathbf{y}^{(j)}) + \lambda_{\mathrm{match}} \left| c^{(i)} - c^{(j)} \right|, \tag{11}$$

where $c^{(i)}$ denotes the concentration prior associated with the i-th slot and $c^{(j)}$ is the ground-truth concentration of the j-th component. The optimal permutation $\hat{\pi}$ that minimizes the total assignment

cost is determined by $\hat{\pi} = \arg\min_{\pi \in \Pi_K} \sum_{i=1}^{K} \mathcal{A}_{i,\pi(i)}$, where Π_K signifies the space of all candidate permutations. For functional group classification, the structural representation $z_{\text{struct}}^{(k)}$ is routed to a classification head to output the predicted probabilities $\hat{\mathbf{y}}^{(k)}$. A multi-label Focal Loss with class-adaptive weights w_g is then applied to these predictions:

$$\mathcal{L}_{\text{FG}} = -\frac{1}{K} \sum_{k=1}^{K} \sum_{g=1}^{G} w_g \left[y_g^{(k)} \left(1 - \hat{y}_g^{(k)}\right)^{\gamma} \log\left(\hat{y}_g^{(k)}\right) + \left(1 - y_g^{(k)}\right) \left(\hat{y}_g^{(k)}\right)^{\gamma} \log\left(1 - \hat{y}_g^{(k)}\right) \right], \tag{12}$$

where $g \in \{1, \ldots, G\}$ denotes the functional group index, $y_g^{(k)} \in \{0, 1\}$ is the ground-truth label, $\hat{y}_g^{(k)}$ is the sigmoid probability derived from $z_{\text{struct}}^{(k)}$, and γ is the focusing parameter. The total objective is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{FG}} + \lambda_{\text{disen}} \mathcal{L}_{\text{disen}} + \lambda_{\text{cfm}} \mathcal{L}_{\text{cfm}}. \tag{13}$$

To stabilize optimization, we adopt a curriculum that starts with balanced mixture ratios and progressively broadens the concentration distribution to cover increasingly imbalanced mixtures.

3 Experiments and Analysis

In this section, we conduct extensive experiments to demonstrate the effectiveness of our model, addressing the following research questions:

- **RQ1:** Can MUSE accurately distinguish individual components in mixtures?
- **RQ2:** Can MUSE maintain reliable substructure elucidation under signal dilution and corruption?
- **RQ3:** How do individual components contribute to MUSE's overall performance?

3.1 Datasets and Setups

Datasets We construct our datasets using simulated spectra from QM9S Zou et al. (2023) and Alberts et al. Alberts et al. (2024), and ∼3K real spectra from the NIST Chemistry WebBook[1]. Each source is randomly split into train, validation, and test sets at an 8:1:1 ratio to prevent data leakage. Following the Beer–Lambert law and prior work Alberts et al. (2025a); Yue et al. (2025), we synthesize mixtures via online sampling: at each step, K basis spectra are drawn from the corresponding subset, and a concentration vector is sampled within a predefined range. These basis spectra are then linearly superimposed by concentrations, yielding ∼5M simulated and ∼100K real-based mixture spectra. Further details are provided in Appendices C and D.

Baselines To benchmark MUSE, we compare it against several representative architectures: XGBoost, 1D-CNN, OpenNMT-based Transformer Klein et al. (2017), the more advanced patch-attention Transformer proposed by Wu et al. Wu et al. (2025). We also include a conditional prediction Transformer specifically designed for mixture spectra by Alberts et al. Alberts et al. (2025a). All baselines are augmented with concentration and molecular formulas as conditioning signals.

Metrics. For functional group prediction we report F1-score and macro-F1, the latter to emphasize rare groups. For the disentangled spectra we report cosine similarity and Mean Absolute Error (MAE). All experiments are conducted 5 times independently, and the average results with variances are reported. Detailed hyperparameter settings and training cost are provided in Appendix E.

3.2 Model Performance (RQ1)

Obs.1: MUSE achieves superior performance over all baselines. Table 1 reports F1-scores across mixture configurations, with MUSE achieving the best performance in all settings. On the Alberts dataset, MUSE outperforms the strongest baseline by 4.2% and 5.5% in the two- and three-component settings, with consistent gains on QM9S (4.1% and 4.6%). The widening margins suggest that set prediction offers a growing advantage over single-target conditional prediction as mixture complexity increases. We further evaluate reliance on chemical priors by omitting molecular formulas. This reduces performance across all models, reflecting the value of atomic composition constraints for

[1] https://webbook.nist.gov

Table 1: F1-scores for predicting per-component functional groups under varying mixture configurations and auxiliary inputs. Component concentrations are uniformly sampled from $[0.2, 0.8]$. **Bold** indicates the top-performing method within each row, which is statistically significant under a t-test at the 95% confidence level.

Mixture Setting	XGBoost	1D-CNN	Transformer	Wu et al.	Alberts et al.	MUSE
Alberts et al. (Simulated Spectra)						
2-comp. (conc. + formula)	$0.641_{(0.021)}$	$0.710_{(0.024)}$	$0.706_{(0.031)}$	$0.711_{(0.017)}$	$0.816_{(0.020)}$	$\mathbf{0.858}_{(0.009)}$
3-comp. (conc. + formula)	$0.619_{(0.028)}$	$0.694_{(0.017)}$	$0.688_{(0.035)}$	$0.702_{(0.022)}$	$0.764_{(0.018)}$	$\mathbf{0.819}_{(0.016)}$
2-comp. (conc. only)	$0.616_{(0.024)}$	$0.663_{(0.022)}$	$0.678_{(0.028)}$	$0.683_{(0.013)}$	$0.728_{(0.025)}$	$\mathbf{0.767}_{(0.011)}$
3-comp. (conc. only)	$0.598_{(0.031)}$	$0.645_{(0.019)}$	$0.651_{(0.033)}$	$0.652_{(0.015)}$	$0.701_{(0.021)}$	$\mathbf{0.744}_{(0.014)}$
QM9S (Simulated Spectra)						
2-comp. (conc. + formula)	$0.672_{(0.028)}$	$0.728_{(0.019)}$	$0.731_{(0.032)}$	$0.739_{(0.014)}$	$0.832_{(0.012)}$	$\mathbf{0.873}_{(0.013)}$
3-comp. (conc. + formula)	$0.647_{(0.015)}$	$0.713_{(0.024)}$	$0.708_{(0.030)}$	$0.724_{(0.018)}$	$0.801_{(0.019)}$	$\mathbf{0.847}_{(0.014)}$
2-comp. (conc. only)	$0.633_{(0.020)}$	$0.679_{(0.021)}$	$0.694_{(0.029)}$	$0.706_{(0.006)}$	$0.749_{(0.015)}$	$\mathbf{0.799}_{(0.021)}$
3-comp. (conc. only)	$0.618_{(0.029)}$	$0.661_{(0.026)}$	$0.669_{(0.034)}$	$0.677_{(0.021)}$	$0.723_{(0.019)}$	$\mathbf{0.774}_{(0.023)}$
NIST WebBook (Experimental pure spectra, synthetically mixed)						
2-comp. (conc. + formula)	$0.604_{(0.026)}$	$0.673_{(0.025)}$	$0.681_{(0.034)}$	$0.689_{(0.020)}$	$0.761_{(0.026)}$	$\mathbf{0.819}_{(0.011)}$
3-comp. (conc. + formula)	$0.581_{(0.032)}$	$0.658_{(0.021)}$	$0.652_{(0.038)}$	$0.671_{(0.026)}$	$0.749_{(0.021)}$	$\mathbf{0.797}_{(0.015)}$
2-comp. (conc. only)	$0.572_{(0.029)}$	$0.626_{(0.028)}$	$0.643_{(0.035)}$	$0.651_{(0.022)}$	$0.703_{(0.018)}$	$\mathbf{0.746}_{(0.023)}$
3-comp. (conc. only)	$0.554_{(0.025)}$	$0.609_{(0.014)}$	$0.617_{(0.041)}$	$0.622_{(0.028)}$	$0.679_{(0.016)}$	$\mathbf{0.732}_{(0.024)}$

spectral disentanglement. Nevertheless, MUSE retains a robust relative advantage, confirming that its gains stem from deeper spectral modeling rather than prior reliance. On the experimental NIST dataset, despite performance degradation across all models attributable to both the sim-to-real domain gap and the limited size of data, MUSE still achieves an f1-score of 0.819, demonstrating strong domain generalization in real-world scenarios.

Obs.2: MUSE exhibits superior robustness against compositional complexity. As shown in Figure 3(a), increasing the number of mixture components universally degrades baseline performance due to intensified signal coupling and attribution ambiguity. When scaling from two to four components, the f1-score drops by $\sim$11% for 1D-CNN and $\sim$13.2% for Alberts et al. However, MUSE exhibits the most graceful degradation, bounding the decline to merely 11.3% (from 0.858 to 0.761). Furthermore, as the spectral signatures of rare groups become increasingly diluted in the four-component setting, baseline macro-f1 scores experience a sharp decline. MUSE's sustained robustness directly validates our dual-level design: the set prediction paradigm jointly resolves global competition to prevent error accumulation, while CFM-guided reconstruction preserves attenuated signals that would otherwise be overshadowed.

Obs.3: MUSE achieves consistent gains on real mixture spectra. To assess MUSE under realistic measurement conditions, including instrumental imperfections and deviations from ideal linear additivity, we evaluate it on two sets of physically measured mixtures: 30 spectra released by Yue et al. Yue et al. (2025) and 29 two-component mixtures measured in this study. The latter cover 12 small organic molecules (approximately 87–186 Da). On the released dataset, MUSE achieves an F1-score of 0.762, outperforming Alberts et al. (0.709) and Wu et al. (0.663). On our measured mixtures, MUSE again ranks first, reaching 0.769 compared with 0.726 for Alberts et al. and 0.703 for Wu et al. Despite the overall performance decline under limited data and domain shift, MUSE retains its lead and exhibits the smallest degradation. These results support the practical effectiveness of MUSE and its transfer from synthetic training mixtures to real measurements.

3.3 Robustness Against Signal Dilution and Corruption (RQ2)

Obs.4: MUSE exhibits stronger elucidation capability on low-concentration components. Figure 3(b-c) reports f1-scores across concentration intervals for two- and three-component mixtures. As the concentration of an individual component increases, all models' performance improves steadily, since higher abundance yields stronger, less ambiguous absorption cues in the mixture spectrum. However, performance diverges drastically in the low-concentration regime. For trace components (< 0.1 in two-component setting, < 0.05 in three-component setting), all baseline models suffer severe degradation. In contrast, MUSE avoids abrupt collapse, outperforming the strongest baseline by an absolute margin of 11.8% and 12.0% in these respective settings. As concentration increases, the

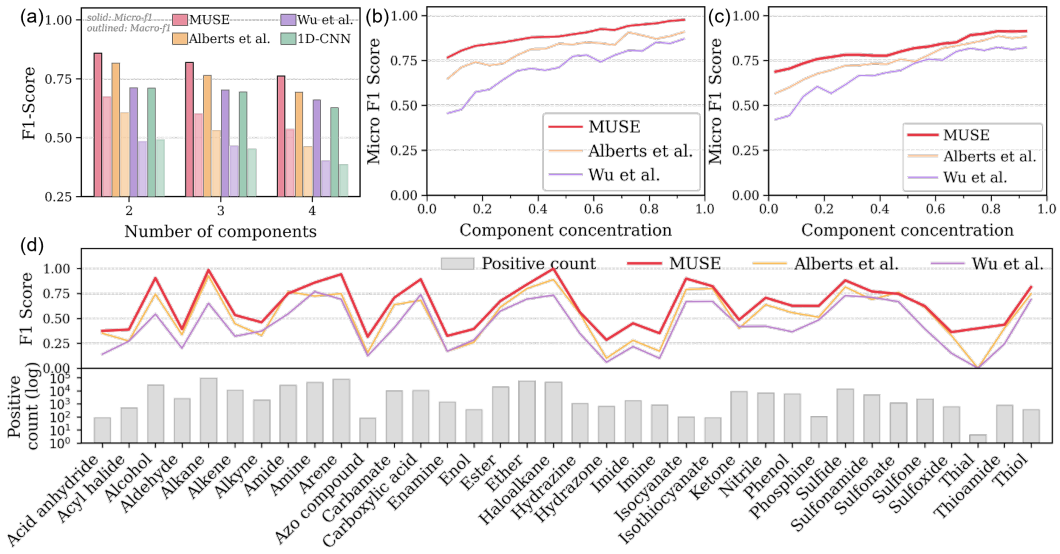

Figure 3: (a) Performance comparison under varying numbers of components, with solid and outlined bars indicating f1-score and macro-f1, respectively. (b)-(c) F1-scores relative to component concentration in two- and three-component settings. (d) Impact of functional group frequency on performance, where bars represent positive sample counts and lines indicate per-class f1-scores.

performance gap gradually narrows, and all models converge to comparable accuracy. This graceful degradation at low abundances confirms that CFM's continuous generative supervision explicitly anchors the concentration subspace, preventing weak, diluted spectral signatures from being drowned out by dominant global signals. We further evaluate MUSE on chemically similar functional groups, strongly overlapping spectra, and constitutional-isomer mixtures (Details in Appendix F).

Obs.5: MUSE remains robust under imperfect input conditions and imperfect priors. In practice, mixture spectroscopy is subject to measurement noise, baseline drift, and uncertainty in the input concentration prior. We evaluate MUSE under each of these perturbations.

Spectral noise. As shown in Figure 4(a), all models degrade as Gaussian noise increases, but MUSE consistently exhibits the slowest decline, retaining an f1-score of 0.762 under severe corruption ($\sigma = 0.10$) while the strongest baseline drops to 0.675. This resilience indicates that the disentangled slot decoding effectively insulates the structural subspace from unstructured intensity perturbations.

Baseline drift. As shown in Figure 4(b), MUSE strongly resists low-frequency distortions, retaining an f1-score of 0.774 at $\alpha = 0.20$, compared to 0.699 for the strongest baseline. We attribute this to the CFM-guided reconstruction, which anchors the structural representation against smooth background modulations that would otherwise be misinterpreted as broadband absorption features.

Concentration prior noise. In practice, concentration priors inevitably contain weighing and pipetting errors, which we model using multiplicative Gaussian noise of relative magnitude σ_c (Table 2). Even at $\sigma_c = 0.10$, which corresponds to coarse manual estimation, MUSE only drops from 0.858 to 0.828 (2-comp) and from 0.819 to 0.787 (3-comp), with smaller drops at $\sigma_c = 0.05$ (0.846 and 0.804, respectively). These results suggest that MUSE relies primarily on spectral evidence, with concentration priors providing auxiliary guidance.

Missing and misspecified priors. We also evaluate missing and misspecified priors (Table 2). For two-component mixtures, removing formulas or concentrations reduces F1 by 9.1% and 7.2%, respectively. With the component number as its only side information, MUSE still achieves 0.742 and 0.704 in the two- and three-component settings. We then overestimate K as well, supplying one slot more than the mixture actually contains and no prior for it, which costs a further 1.4% and 2.2%.

Overall, these results confirm MUSE's reliable performance under realistic input imperfections, supporting real-world chemical applications. Details are provided in Appendices G, H, I and J.

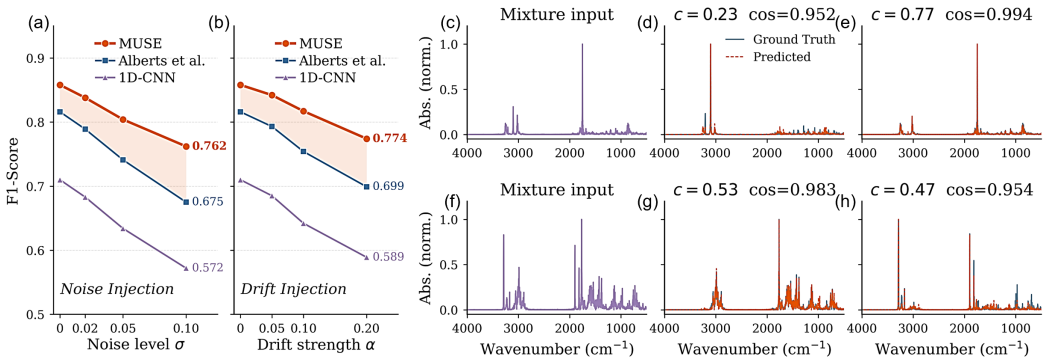

Figure 4: (a) F1-score under varying noise levels. (b) F1-score under varying baseline drift levels. (c–h) Reconstructed pure spectra of individual components, with cosine similarities to ground truth. More prediction results are provided in Appendix K.

Table 2: F1-scores under reduced and misspecified input priors on the Alberts et al. dataset.

Setting	K+conc.+formula	K+formula	K+conc.	K only	$\hat{K} = K + 1$ (K only)
2-comp.	0.858	0.786	0.767	0.742	0.728
3-comp.	0.819	0.737	0.744	0.704	0.682

Obs.6: MUSE achieves stronger performance on low-frequency functional groups. As shown in Figure 3(d), the label distribution is severely imbalanced, ranging from 96k positive samples for Alkane to only 4 for Thial. Although all methods perform comparably on majority classes, baseline performance deteriorates markedly in the low-frequency regime. On rare groups such as Imine and Azo compounds, baseline F1-scores fall as low as 0.06. MUSE maintains a clear advantage on these categories, achieving up to four times the performance of the strongest baseline. We attribute this advantage to disentangled slot decoding: separating structural identity from concentration helps prevent weak component-specific spectral features from being overshadowed by dominant signals.

4 Ablation Study and Sensitivity Analysis (RQ3)

Obs.7: CFM reconstructs pure spectra and improves component recognition. Figure 4(c–h) shows that CFM recovers characteristic peaks of both dominant and trace components, with MAE of 0.011/0.014 and cosine similarities of 0.833/0.673 for two-/three-component mixtures. Table 3 examines how this supervision interacts with the prediction architecture. Slot-only prediction without disentanglement or reconstruction reaches 0.812/0.758, while independent branches retaining CFM achieve 0.820/0.762; neither configuration recovers the performance of full MUSE. Replacing CFM with concentration prediction yields 0.832/0.778, nearly matching the variant without reconstruction (0.831/0.776). A direct CNN decoder provides a larger recovery (0.838/0.798), but remains below full MUSE (0.858/0.819). These comparisons support the combined architecture and show that direct concentration supervision provides little of the benefit of spectral reconstruction. CFM's further advantage over the direct decoder is consistent with the benefit of supervision along a continuous target-enrichment path. Baseline architectures are described in Appendix L.

Table 3: Attribution of MUSE's performance (f1-score).

Model Variant	**2-Comp**	**3-Comp**
Independent branches	0.820	0.762
Slot-only	0.812	0.758
w/o Reconstruct	0.831	0.776
w/ Direct Decoder	0.838	0.798
w/ Conc. Prediction	0.832	0.778
w/o Subspace Split	0.815	0.772
w/o $\mathcal{L}_{\text{disen}}$	0.817	0.776
Full MUSE (Ours)	**0.858**	**0.819**

Obs.8: Disentanglement decouples concentration from structure. Figure 5(a) visualizes test-set concentration embeddings using t-SNE, colored by ground-truth concentration. The smooth color gradient suggests that z_{conc} encodes concentration-dependent variation. In Table 3, removing $\mathcal{L}_{\text{disen}}$ reduces F1 by 4.1% and 4.3% for two- and three-component mixtures; additionally removing the subspace split increases these losses to 4.3% and 4.7%. These results support combining dedicated structural and concentration representations with cross-covariance regularization. To examine what

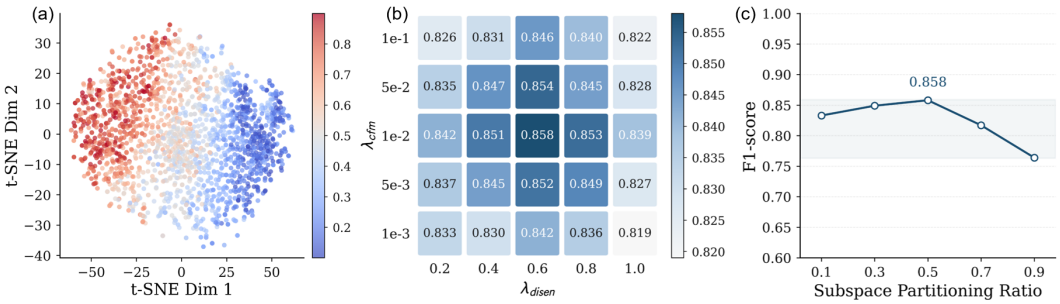

Figure 5: (a) t-SNE of $z_{\text{conc}}^{(k)}$ colored by ground-truth concentration. (b) F1-scores across varying disentanglement (λ_{disen}) and flow matching (λ_{cfm}) weights. (c) Impact of the subspace partitioning ratio on component recognition accuracy.

each subspace encodes, we freeze the trained MUSE model and train small prediction networks to recover functional-group labels or concentrations from each subspace separately. Functional-group prediction achieves an F1-score of 0.84 from z_{struct}, compared with only 0.15 from z_{conc}. Conversely, concentration prediction is more accurate from z_{conc} (Pearson's $R = 0.88$) than from z_{struct} ($R = 0.06$). These results suggest that the two subspaces preferentially encode their intended factors, with limited information leakage. We also place the same molecule in mixtures at different concentrations and compare its structural embeddings. Their mean cosine similarity of 0.95 indicates that the structural representation remains largely stable as concentration changes (Details in Appendix B).

Obs.9: Sensitivity analysis. Figure 5(b–c) examines the two loss weights jointly and the subspace partitioning ratio. F1-scores are stable across a broad central region and fall off at both extremes, suggesting that weak penalties provide insufficient disentanglement, while excessive regularization may interfere with classification. For the partitioning ratio, performance peaks at an equal split, supporting a balanced allocation of representation capacity between structural classification and spectral reconstruction.

5 Broader Impacts and Limitations

Broader Impacts. MUSE elucidates per-component functional groups directly from mixture spectra, supporting analytical chemistry while reducing reliance on costly instrumentation like mass spectrometry. Its disentangled representation framework also offers a transferable paradigm for inverse problems such as blind source separation and multi-modal decomposition in medical imaging.

Limitations. Despite its strengths, MUSE still has limitations. First, it uses IR as its only spectral modality, as multimodal mixture datasets remain scarce and incorporating additional modalities may improve identification accuracy. Second, it predicts per-component functional groups rather than complete molecular structures. Since the CFM branch already reconstructs pure-component spectra, adding a molecular decoder is a natural next step. Third, MUSE requires a predefined component count, without which decomposition of a single mixture spectrum is generally underdetermined unless additional constraints are imposed. Finally, scaling to mixtures of five or more components remains challenging due to severe signal overlap and increasing assignment ambiguity.

6 Conclusion

In this work, we introduced MUSE, an end-to-end framework that reformulates substructure elucidation from mixed infrared spectra as one-shot set prediction. Its permutation-equivariant slot decoder models global component competition and mitigates the error accumulation associated with sequential extraction. To separate chemical identity from relative abundance, MUSE combines latent subspace disentanglement with CFM, providing a generative constraint on the concentration subspace that is interpretable within the linear absorbance model. Extensive evaluations demonstrate that MUSE outperforms existing baselines and remains robust to compositional complexity, measurement noise, and signal dilution in trace components. These findings support physically grounded representation learning as a promising direction for AI-driven complex mixture analysis.

AI USE STATEMENT

Generative AI tools were used to assist with language editing and to improve the clarity, organization, and readability of parts of the manuscript. The authors carefully reviewed and revised all AI-assisted content, including its text, claims, results, and other materials.

REFERENCES

Michael S Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. *arXiv preprint arXiv:2209.15571*, 2022.

Marvin Alberts, Oliver Schilter, Federico Zipoli, Nina Hartrampf, and Teodoro Laino. Unraveling molecular structure: A multimodal spectroscopic dataset for chemistry. *Advances in Neural Information Processing Systems*, 37:125780–125808, 2024.

Marvin Alberts, Filippo Ficarra, and Teodoro Laino. Language model enabled structure prediction from infrared spectra of mixtures. In *AI for Accelerated Materials Design-NeurIPS 2025*, 2025a.

Marvin Alberts, Federico Zipoli, and Teodoro Laino. Setting new benchmarks in ai-driven infrared structure elucidation. *Digital Discovery*, 4(7):1936–1943, 2025b.

Carlos R Baiz, Bartosz Błasiak, Jens Bredenbeck, Minhaeng Cho, Jun-Ho Choi, Steven A Corcelli, Arend G Dijkstra, Chi-Jui Feng, Sean Garrett-Roe, Nien-Hui Ge, et al. Vibrational spectroscopic map, vibrational spectroscopy, and intermolecular interaction. *Chemical reviews*, 120(15):7152–7218, 2020.

Matthew J Baker, Júlio Trevisan, Paul Bassan, Rohit Bhargava, Holly J Butler, Konrad M Dorling, Peter R Fielden, Simon W Fogarty, Nigel J Fullwood, Kelly A Heys, et al. Using fourier transform ir spectroscopy to analyze biological materials. *Nature protocols*, 9(8):1771–1791, 2014.

Montgomery Bohde, Mrunali Manjrekar, Runzhong Wang, Shuiwang Ji, and Connor W Coley. Diffms: Diffusion generation of molecules conditioned on mass spectra. *arXiv preprint arXiv:2502.09571*, 2025.

Biqing Chen and Xiaohong Qiu. Deep learning-empowered sers: deciphering the multidimensional information code of complex biological samples. *Analytical Methods*, 2026.

Ricky TQ Chen and Yaron Lipman. Flow matching on general geometries. *arXiv preprint arXiv:2302.03660*, 2023.

Yu Cheng, Jue Ding, Catherine Estefany Davila Arenas, Markus Brinkmann, and Xiaowen Ji. A brief review on the assessment of potential joint effects of complex mixtures of contaminants in the environment. *Environmental Science: Advances*, 3(5):661–675, 2024.

Mirosław Antoni Czarnecki, Yusuke Morisawa, Yoshisuke Futami, and Yukihiro Ozaki. Advances in molecular structure and interaction studies using near-infrared spectroscopy. *Chemical reviews*, 115(18):9707–9744, 2015.

Sriram Devata, Bhuvanesh Sridharan, Sarvesh Mehta, Yashaswi Pathak, Siddhartha Laghuvarapu, Girish Varma, and U. Deva Priyakumar. Deepspinn – deep reinforcement learning for molecular structure prediction from infrared and 13c nmr spectra. *Digital Discovery*, 3:818–829, 2024. doi: 10.1039/D4DD00008K. URL `http://dx.doi.org/10.1039/D4DD00008K`.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky TQ Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. Discrete flow matching. *Advances in Neural Information Processing Systems*, 37: 133345–133385, 2024.

Kehan Guo, Bozhao Nan, Yujun Zhou, Taicheng Guo, Zhichun Guo, Mihir Surve, Zhenwen Liang, Nitesh Chawla, Olaf Wiest, and Xiangliang Zhang. Can llms solve molecule puzzles? a multimodal benchmark for molecular structure elucidation. *Advances in Neural Information Processing Systems*, 37:134721–134746, 2024.

Michael Jackson, Michael G Sowa, and Henry H Mantsch. Infrared spectroscopy: a new frontier in medicine. *Biophysical Chemistry*, 68(1-3):109–125, 1997.

Guwon Jung, Son Gyo Jung, and Jacqueline M Cole. Automatic materials characterization from infrared spectra using convolutional neural networks. *Chemical Science*, 14(13):3600–3609, 2023.

Shahid Ali Khan, Sher Bahadar Khan, Latif Ullah Khan, Aliya Farooq, Kalsoom Akhtar, and Abdullah M Asiri. Fourier transform infrared spectroscopy: fundamentals and application in functional groups and nanomaterials characterization. In *Handbook of materials characterization*, pp. 317–344. Springer, 2018.

Karen A Kidd, Thomas Backhaus, Tomas Brodin, Pedro A Inostroza, and Erin S McCallum. Environmental risks of pharmaceutical mixtures in aquatic ecosystems: Reflections on a decade of research. *Environmental Toxicology and Chemistry*, 43(3):549–558, 2024.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pp. 67–72, Vancouver, Canada, July 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P17-4012.

Joshua L Lansford and Dionisios G Vlachos. Infrared spectroscopy data-and physics-driven machine learning for characterizing surface microstructure of complex materials. *Nature communications*, 11(1):1513, 2020.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Mohamed W Nassar, Ahmed Serag, Mohamed A Hasan, and Mohamed Kamel. Development and validation of a rp-hplc method for simultaneous determination of five covid-19 antiviral drugs in pharmaceutical formulations. *Scientific reports*, 15(1):25470, 2025.

Nitin Liladhar Rane, Mallikarjuna Paramesha, Saurabh P Choudhary, and Jayesh Rane. Machine learning and deep learning for big data analytics: A review of methods and applications. *Partners Universal International Innovation Journal*, 2(3):172–197, 2024.

Anouk M Rijs and Jos Oomens. Ir spectroscopic techniques to study isolated biomolecules. *Gas-Phase IR Spectroscopy and Structure of Biological Molecules*, pp. 1–42, 2015.

Åsmund Rinnan, Frans Van Den Berg, and Søren Balling Engelsen. Review of the most common pre-processing techniques for near-infrared spectra. *TrAC Trends in Analytical Chemistry*, 28(10): 1201–1222, 2009.

Iqbal H Sarker. Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. *SN computer science*, 2(6):1–20, 2021.

Iqtiar Siddique. Exploring functional groups and molecular structures: a comprehensive analysis using ftir spectroscopy. *Chemistry Research Journal*, 9(2):70–76, 2024.

Roma Tauler. Multivariate curve resolution applied to second order data. *Chemometrics and intelligent laboratory systems*, 30(1):133–146, 1995.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *arXiv preprint arXiv:2302.00482*, 2023.

Maxwell C Venetos, Masha Elkin, Connor Delaney, John F Hartwig, and Kristin A Persson. Deconvolution and analysis of the 1h nmr spectra of crude reaction mixtures. *Journal of chemical information and modeling*, 64(8):3008–3020, 2024.

Bo Wang, Pu Zhang, Xiangping Zhu, Hua Wang, Wenzhen Ren, Chuan Jin, and Wei Zhao. Multiplet network for one-shot mixture raman spectrum identification. In *Photonics*, volume 12, pp. 295. MDPI, 2025.

Wenjin Wu, Ales Leonardis, Jianbo Jiao, Jun Jiang, and Linjiang Chen. Transformer-based models for predicting molecular structures from infrared spectra using patch-based self-attention. *The Journal of Physical Chemistry A*, 129(8):2077–2085, 2025.

Zhuo Yang, Jiaqing Xie, Shuaike Shen, Daolang Wang, Yeyun Chen, Ben Gao, Shuzhou Sun, Biqing Qi, Dongzhan Zhou, Lei Bai, et al. Spectrumworld: Artificial intelligence foundation for spectroscopy. *arXiv preprint arXiv:2508.01188*, 2025.

Xiaoyu Yue, Tao Song, Jiaqi Cao, Xiaolong Zhang, Shuo Feng, Weiwei Shang, Jun Jiang, and Linjiang Chen. Ir-bot: An autonomous robotic system for real-time chemical mixture analysis via infrared spectroscopy and machine learning. *CCS Chemistry*, pp. 1–14, 2025.

Zihan Zou, Yujin Zhang, Lijun Liang, Mingzhi Wei, Jiancai Leng, Jun Jiang, Yi Luo, and Wei Hu. A deep learning model for predicting selected organic molecular spectra. *Nature Computational Science*, 3(11):957–964, 2023.

A Related Work

A.1 Molecular Structure Elucidation via Spectral Analysis

Deep learning for spectral analysis spans a spectrum of tasks and methodological paradigms. At the level of partial structure recognition, Jung et al. Jung et al. (2023) employed convolutional neural networks to identify functional groups from infrared spectra, treating elucidation as a multi-label classification problem. Moving toward complete structure prediction, Wu et al. Wu et al. (2025) proposed a patch-based attention Transformer that partitions infrared spectra into local segments and models their long-range dependencies to predict full molecular structures. Devata et al. Devata et al. (2024) took a different methodological route with DeepSPInN, formulating structure elucidation as a sequential decision-making problem solved by reinforcement learning over infrared and ^{13}C-NMR spectra, without querying any spectral reference database. Beyond methods targeting individual modalities, Alberts et al. Alberts et al. (2024) contributed a multimodal spectroscopic dataset together with baseline models, providing infrastructure for approaches that integrate heterogeneous spectral sources.

A more recent line of work reframes spectral analysis through generative modeling and general-purpose foundation models. DiffMS Bohde et al. (2025) adopts a diffusion process over molecular graphs conditioned on mass spectral peaks, with combinatorial optimization enforcing chemical validity during generation. In parallel, large language models have been examined as general spectral reasoners: Guo et al. Guo et al. (2024) constructed a multimodal benchmark to assess whether LLMs can parse spectroscopic signals and reconstruct molecular structures, while SpectrumWorld Yang et al. (2025) introduces a multi-agent framework and benchmark suite aimed at standardizing deep learning research across spectroscopic modalities.

A.2 Flow Matching and Conditional Generative Modeling

Flow matching has emerged as a powerful framework for training continuous-time generative models. Lipman et al. Lipman et al. (2022) introduced flow matching as a simulation-free objective for training continuous normalizing flows by regressing a vector field along a prescribed probability path, and demonstrated that conditional flow matching yields tractable gradients identical to those of the intractable marginal objective, providing a natural foundation for conditional generation. Concurrently, Liu et al. Liu et al. (2022) proposed rectified flow, which learns linear interpolations between source and target samples and applies a reflow procedure to straighten trajectories for few-step generation, while Albergo et al. Albergo & Vanden-Eijnden (2022) developed stochastic interpolants, unifying flow-based and diffusion-based generative modeling under a common transport-equation viewpoint. Building on this foundation, Tong et al. Tong et al. (2023) introduced minibatch optimal-transport couplings to further straighten trajectories and to formulate conditional generation as a coupling problem between paired distributions, Chen et al. Chen & Lipman (2023) generalized the framework to non-Euclidean domains through Riemannian flow matching, and Gat et al. Gat et al. (2024) extended it to categorical data with discrete flow matching. The paradigm has also demonstrated strong scalability in conditional settings, with Esser et al. Esser et al. (2024) applying rectified-flow Transformers to high-resolution text-conditioned image synthesis.

B Disentanglement Diagnostics

In this section, we provide two quantitative diagnostics that test whether each subspace encodes the intended factor and only that factor, so that the claimed disentanglement is verified rather than assumed.

Cross-prediction probes. To verify that chemical identity and relative abundance are carried by different parts of the slot embedding, we probe each subspace directly. We freeze the trained model, extract $z^{(k)}_{\text{struct}}$ and $z^{(k)}_{\text{conc}}$ on the test split, and train lightweight two-layer MLP probes to predict either the functional-group labels or the true concentration from each subspace. A well-separated representation should give a strong diagonal in Table 4, where z_{struct} predicts structure and z_{conc} predicts concentration, and a weak off-diagonal. As reported in Table 4, MUSE exhibits exactly this pattern: z_{struct} predicts functional groups at 0.84 f1 while z_{conc} reaches only 0.15, and z_{conc} predicts the true concentration at $R = 0.88$ while z_{struct} attains merely 0.06. Removing $\mathcal{L}_{\text{disen}}$ raises the two

Table 4: Cross-prediction probes on frozen representations. Diagonal entries should be high and off-diagonal entries low. FG is reported as f1, concentration as Pearson R.

Model	$z_{\text{struct}} \to$FG $\uparrow$	$z_{\text{conc}} \to$FG $\downarrow$	$z_{\text{struct}} \to c \downarrow$	$z_{\text{conc}} \to c \uparrow$
MUSE	**0.84**	**0.15**	**0.06**	**0.88**
w/o $\mathcal{L}_{\text{disen}}$	0.77	0.41	0.30	0.72
w/o Subspace Split	0.73	0.45	0.40	0.66

Table 5: Invariance of the structural subspace and monotonicity of the concentration subspace, measured for the same molecule placed at different abundances across mixtures.

Model	z_{struct} similarity $\uparrow$	z_{conc} vs. c Pearson $\uparrow$
MUSE	**0.95**	**0.91**
w/o $\mathcal{L}_{\text{disen}}$	0.79	0.74
w/o Subspace Split	0.69	0.62

off-diagonal entries to 0.41 and 0.30, and removing the subspace split as well raises them to 0.45 and 0.40, while the diagonal entries themselves fall from 0.84 to 0.73 and from 0.88 to 0.66, indicating that information leaks in both directions once either mechanism is absent.

Invariance and monotonicity. The second diagnostic tests the property the design is meant to deliver, namely that the structural representation of a molecule does not move when its abundance changes. We place the same molecule in different mixtures at concentrations spanning the sampled range, and report the mean pairwise cosine similarity among its structural embeddings, which should be near-invariant, together with the Pearson correlation between z_{conc} and the true concentration, which should track it monotonically. As shown in Table 5, the structural embeddings of MUSE remain nearly unchanged across abundances, with a mean cosine similarity of 0.95, while its concentration subspace tracks the true concentration at a Pearson correlation of 0.91. Both quantities fall steeply for the two ablations, to 0.79 and 0.74 without $\mathcal{L}_{\text{disen}}$ and to 0.69 and 0.62 without the subspace split, showing that the invariance is produced by these two mechanisms rather than emerging on its own.

C Further Details on Synthetic Mixture Spectra

According to the Beer-Lambert Law, under ideal conditions, the total absorbance of a mixture is equal to the linear sum of the absorbances of its pure components. Based on this physical principle, we employ the following method to generate synthetic mixture spectra: first, the standard infrared spectra of the pure components are obtained and strictly converted to the absorbance mode to ensure linear additivity. Subsequently, based on randomly sampled or predefined relative physical concentration distributions $\mathcal{C} = \{c^{(1)}, c^{(2)}, \ldots, c^{(K)}\}$, these pure component spectra are linearly weighted and combined as $x_{\text{mix}} = \sum_{k=1}^{K} c^{(k)} s_{\text{pure}}^{(k)}$. In this way, we can efficiently and physically consistently generate large-scale, annotated mixture spectral datasets for subsequent deep learning model training.

To verify the feasibility and rationality of this linear superposition-based method, we utilized a recently open-sourced dataset of mixture spectra obtained from real experiments by Yue et al. Yue et al. (2025). Specifically, these real mixture spectra were acquired by drop-casting real reaction solutions onto calcium fluoride (CaF_2) substrates and drying them into films, using an automated robotic platform (IR-Bot) integrated with a Fourier transform infrared spectrometer (Nicolet IS50 FT-IR). We selected three compounds from this dataset: 4-(4,4,5,5-tetramethyl-1,3,2-dioxaborolan-2-yl)-benzonitrile, benzoyl benzonitrile, and benzoic acid.

In the validation experiment, we converted these real open-source spectra into absorbance mode and used our method to linearly weight the corresponding pure component spectra based on the real concentration ratios provided in the experiment, generating synthetic mixture spectra. The comparison results are shown in Figure 6. It can be clearly observed that, although limited by factors such as instrumental baseline drift in the real experimental environment—which leads to minor shifts or absolute intensity differences at certain peak positions between the real and synthetic spectra—the overall spectral trends, the positional distribution of characteristic absorption peaks, and the relative

profiles of the two maintain a high degree of similarity. This high level of consistency demonstrates that by linearly combining the spectra of individual components, we can obtain an approximation of the spectrum of a mixture. This provides solid experimental evidence to support our use of synthetic mixture spectra for driving model training. For the pure component spectra and the real mixture spectra, please refer to Figure 7 and Figure 8.

D Definition of Functional Groups

To systematically analyze molecular structures, we employ a set of predefined functional groups to identify key structures within a given molecular dataset.

The functional groups investigated in this study, together with their SMARTS representations, are summarized in Table 6. The selection criteria prioritized chemical relevance to organic and medicinal chemistry, covering widely encountered moieties including hydroxyl (-OH), carbonyl (C=O), and amine (-NH_2) groups. Structural identification was carried out via subgraph matching algorithms applied to the molecular graphs.

Table 6: Selected Functional Groups and Their SMARTS Patterns

Functional Group	**SMARTS Pattern**
Acid anhydride	`[CX3](=[OX1])[OX2][CX3](=[OX1])`
Acyl halide	`[CX3](=[OX1])[F,Cl,Br,I]`
Alcohol	`[#6][OX2H]`
Aldehyde	`[CX3H1](=O)[#6,H]`
Alkane	`[CX4;H3,H2]`
Alkene	`[CX3]=[CX3]`
Alkyne	`[CX2]#[CX2]`
Amide	`[NX3][CX3](=[OX1])[#6]`
Amine	`[NX3;H2,H1,H0;!$(NC=O)]`
Arene	`[cX3]1[cX3][cX3][cX3][cX3][cX3]1`
Azo compound	`[#6][NX2]=[NX2][#6]`
Carbamate	`[NX3][CX3](=[OX1])[OX2H0]`
Carboxylic acid	`[CX3](=O)[OX2H]`
Enamine	`[NX3][CX3]=[CX3]`
Enol	`[OX2H][#6X3]=[#6]`
Ester	`[#6][CX3](=O)[OX2H0][#6]`
Ether	`[OD2]([#6])[#6]`
Haloalkane	`[#6][F,Cl,Br,I]`
Hydrazine	`[NX3][NX3]`
Hydrazone	`[NX3][NX2]=[#6]`
Imide	`[CX3](=[OX1])[NX3][CX3](=[OX1])`
Imine	`[$([CX3]([#6])[#6]),$([CX3H][#6])]=[$([NX2][#6]),$([NX2H])]`
Isocyanate	`[NX2]=[C]=[O]`
Isothiocyanate	`[NX2]=[C]=[S]`
Ketone	`[#6][CX3](=O)[#6]`
Nitrile	`[NX1]#[CX2]`
Phenol	`[OX2H][cX3]:[c]`
Phosphine	`[PX3]`
Sulfide	`[#16X2H0]`
Sulfonamide	`[#16X4]([NX3])(=[OX1])(=[OX1])[#6]`
Sulfonate	`[#16X4](=[OX1])(=[OX1])([#6])[OX2H0]`
Sulfone	`[#16X4](=[OX1])(=[OX1])([#6])[#6]`
Sulfonic acid	`[#16X4](=[OX1])(=[OX1])([#6])[OX2H]`
Sulfoxide	`[#16X3]=[OX1]`
Thial	`[CX3H1](=S)[#6,H]`
Thioamide	`[NX3][CX3]=[SX1]`
Thiol	`[#16X2H]`

The functional group identification is performed using cheminformatics libraries such as RDKit, which allows for efficient substructure searches within molecular datasets. This approach enables us to extract chemically meaningful information and facilitate downstream tasks such as molecular property prediction, reactivity analysis, and structure-based clustering.

Table 7: Discrimination within the carbonyl family (aldehyde, ketone, and ester).

Method	Subset F1 ↑	Cross-class confusion ↓
Wu et al.	0.58	0.39
Alberts et al.	0.69	0.27
MUSE	**0.75**	**0.18**

Table 8: F1-scores grouped by the cosine similarity between the two pure-component spectra. Higher similarity indicates greater spectral overlap.

Method	$[0.0, 0.4)$	$[0.4, 0.7)$	$[0.7, 1.0)$
Wu et al.	0.74	0.63	0.56
Alberts et al.	0.79	0.72	0.64
MUSE	**0.86**	**0.79**	**0.72**

E Training Settings

MUSE takes infrared spectra of length $L = 1800$ as input. The spectra are first encoded by a 1D-CNN with channel sizes of $[32, 128]$, and then processed by a Transformer backbone with a hidden dimension of $d = 256$, 4 attention heads. The CFM vector field network is built from 4 AdaLN residual blocks with 64 hidden channels, and we use 10 steps for integration. Optimization is performed with AdamW at a learning rate of 3×10^{-4}, weight decay of 0.01, and batch size of 512. The learning rate follows a schedule with a 10% linear warmup and cosine annealing thereafter. To better handle complex mixtures, we adopt a 3-stage curriculum learning strategy, in which the concentration sampling bounds are gradually expanded from $[0.45, 0.55]$ to $[0.20, 0.80]$ over 13 epochs. The overall training objective combines a multi-label Focal Loss with $\gamma = 2.0$ for functional group classification, together with several weighted auxiliary terms. MUSE is trained on two NVIDIA A100 40GB GPUs, with a total training time of approximately 10 hours.

F Chemically Similar Groups, Spectral Overlap, and Isomers

We conduct three complementary evaluations in the two-component setting of the Alberts et al. dataset to examine component-level prediction under chemical similarity, spectral overlap, an constitutional isomerism.

Chemically similar functional groups. We first evaluate discrimination within the carbonyl family, whose members share the dominant C=O stretching band. For the aldehyde, ketone, and ester subset, we report the F1-score and cross-class confusion rate, defined as the fraction of positive predictions incorrectly assigned to another member of this family. As shown in Table 7, MUSE achieves the highest subset F1-score (0.75) and the lowest confusion rate (0.18), compared with 0.69/0.27 for Alberts et al. and 0.58/0.39 for Wu et al. These results indicate improved discrimination among functional groups with similar characteristic absorptions.

Strongly overlapping spectra. For 100 two-component test mixtures, we compute the cosine similarity between the two pure-component spectra and group the mixtures into three similarity bins. Higher similarity indicates greater overlap in overall spectral shape. Table 8 reports the F1-score within each bin. MUSE leads across all three bins, with its F1-score decreasing from 0.86 in the lowest-similarity bin to 0.72 in the highest. The corresponding decreases are 0.79 to 0.64 for Alberts et al. and 0.74 to 0.56 for Wu et al. MUSE therefore maintains its advantage under strong overlap, exceeding the strongest baseline by 8 percentage points in the highest-similarity bin.

Constitutional-isomer mixtures. We additionally curate 100 two-component mixtures whose constituents are constitutional isomers. Both slots receive the same molecular formula, so the formula prior cannot distinguish the two components within a mixture. We compare full MUSE with a variant without formula input and a formula-only control without spectral input.

Table 9: Performance on constitutional-isomer mixtures. The final two rows report full MUSE on subsets defined by whether the two components have distinct or identical functional-group label sets.

Setting / subset	N	**F1**
Formula-only (no spectrum)	100	0.25
MUSE (w/o formula)	100	0.78
MUSE (full)	100	0.77
MUSE (full): distinct label sets	94	0.77
MUSE (full): identical label sets	6	0.80

As shown in Table 9, the formula-only control achieves an F1-score of 0.25, whereas full MUSE reaches 0.77. Removing the formula prior yields a comparable score of 0.78, supporting the role of spectral evidence in component discrimination.

To clarify what these results establish, we further separate the mixtures according to whether their components have distinct functional-group label sets. Of the 100 isomer pairs, 94 have distinct label sets. Full MUSE achieves an F1-score of 0.77 on these pairs, matching its score on the complete isomer subset. Its performance therefore remains consistent when evaluation is restricted to pairs that require component-specific functional-group predictions.

The remaining six pairs share identical functional-group label sets, on which MUSE achieves an F1-score of 0.80. For these pairs, recovering the shared label set evaluates functional-group prediction but does not establish discrimination between the underlying molecular identities. This limitation follows from the resolution of the functional-group target; distinguishing such isomers requires a finer output representation, such as molecular topology.

G Noise Injection Protocol

To evaluate the robustness of MUSE against realistic measurement noise, we introduce a controlled noise injection protocol. For a given clean mixture spectrum $x_{\text{mix}} \in \mathbb{R}^L$, we generate its noisy observation $\tilde{x}_{\text{mix}}$ by adding zero-mean Gaussian noise scaled by the peak spectral intensity:

$$\tilde{x}_{\text{mix}}(\lambda) = x_{\text{mix}}(\lambda) + \epsilon(\lambda), \quad \epsilon(\lambda) \sim \mathcal{N}(0, (\sigma I_{\max})^2),$$

where $I_{\max} = \max_\lambda x_{\text{mix}}(\lambda)$ is the maximum absorption intensity and $\sigma \in \{0.00, 0.02, 0.05, 0.10\}$ controls the relative noise level. This intensity-dependent formulation ensures consistent perturbation scaling across spectra with varying dynamic ranges, faithfully simulating the heteroscedastic noise profiles typical of real-world instrument acquisitions.

H Baseline Drift Injection Protocol

Beyond high-frequency Gaussian noise, real infrared measurements are commonly affected by baseline drift—a smooth, low-frequency background distortion arising from detector non-uniformity, film thickness variation, and environmental factors. Unlike random noise, drift is highly correlated across adjacent wavenumbers and probes a complementary axis of measurement non-ideality.

For a clean mixture spectrum $x_{\text{mix}} \in \mathbb{R}^L$, we generate its drift-perturbed observation as

$$\tilde{x}_{\text{mix}}(\lambda) = x_{\text{mix}}(\lambda) + \alpha \cdot I_{\max} \cdot b(\lambda), \quad b(\lambda) = \tfrac{1}{2}\left[\sin\left(2\pi f \tilde{\lambda} + \phi\right) + 1\right], \tag{14}$$

where $I_{\max} = \max_\lambda x_{\text{mix}}(\lambda)$, $\tilde{\lambda} \in [0, 1]$ is the normalized wavenumber, and $\alpha \in \{0.00, 0.05, 0.10, 0.20\}$ controls the relative drift magnitude. The frequency f is randomly sampled from $\{0.5, 1.0, 1.5, 2.0\}$ and the phase ϕ from $[0, 2\pi]$ for each spectrum, yielding diverse drift shapes ranging from monotonic offsets to S-shaped curves. The α values are chosen larger than the corresponding noise levels, reflecting that low-frequency drift is typically less disruptive than high-frequency noise of equal amplitude.

I Concentration Noise Injection Protocol

In real-world chemical analysis, the input concentration prior is rarely exact and is typically subject to measurement uncertainty arising from weighing imprecision, pipetting variability, and dilution errors. Such deviations are inherently *relative*: a 5% pipetting error perturbs each component by a fraction of its own concentration rather than by an absolute offset. To faithfully simulate this regime, we inject multiplicative Gaussian noise into the concentration prior.

For a clean concentration vector $c = (c^{(1)}, c^{(2)}, \ldots, c^{(K)})$ satisfying $\sum_k c^{(k)} = 1$, we generate its perturbed observation as:

$$\tilde{c}^{(k)} = c^{(k)} \cdot \big(1 + \epsilon^{(k)}\big), \quad \epsilon^{(k)} \sim \mathcal{N}(0, \sigma_c^2), \tag{15}$$

followed by clipping and re-normalization to preserve the simplex constraint:

$$\tilde{c}^{(k)} \leftarrow \frac{\max\big(\tilde{c}^{(k)}, 0\big)}{\sum_j \max\big(\tilde{c}^{(j)}, 0\big)}. \tag{16}$$

The relative noise level $\sigma_c \in \{0.00, 0.02, 0.05, 0.10\}$ spans the practical range of concentration measurement errors, from high-precision analytical balances ($\sigma_c \approx 0.02$) to coarse manual estimation ($\sigma_c \approx 0.10$). Larger noise levels are not considered, as deviations beyond 10% are uncommon in standard laboratory practice and would no longer correspond to realistic measurement scenarios.

Unlike noise and drift injection, which perturb the spectral input itself, concentration noise targets the auxiliary prior fed into the Hungarian matching cost (Eq. 11). This setup directly probes the model's reliance on the precision of concentration priors, complementing the spectral perturbation studies by characterizing robustness along the prior-input axis.

J The Sensitivity of Input Priors

Problem setting. Recovering individual components from a single mixture spectrum is an underdetermined problem, particularly when the component count is unknown. MUSE takes the component count K as part of the task specification, while per-component concentrations $\mathcal{C}$ and molecular formulas serve as optional auxiliary priors. Below, we evaluate its sensitivity to missing priors, misspecified K, and noisy concentrations.

Reduced priors. Table 2 in the main text reports f1-scores when the optional priors are withheld. Relative to the full setting, removing the molecular formulas costs 9.1 points and removing the concentrations 7.2 points for two-component mixtures, while for three-component mixtures the same removals cost 7.5 and 8.2 points. With K as the sole prior, MUSE still attains 0.742 and 0.704; the latter already matches what the strongest baseline achieves with the concentration prior supplied (0.701 in Table 1). Interestingly, formulas help more than concentrations in the two-component case while the ordering reverses for three components, consistent with concentrations becoming more valuable for disambiguation as the number of components, and hence the number of possible assignments, grows.

Misspecified K. We additionally overestimate the component number, instantiating $\hat{K} = K + 1$ slots while supplying neither a concentration nor a formula for the extra slot. Hungarian matching is then performed between $\hat{K}$ slots and K ground-truth components; unmatched slots receive no classification target and are ignored when computing f1. Performance degrades from 0.742 to 0.728 for $\hat{K} = 3$ with $K = 2$, and from 0.704 to 0.682 for $\hat{K} = 4$ with $K = 3$, i.e. the redundant slot is largely absorbed without disrupting the assignment of the components that are actually present. Underestimating K is not evaluated, since with fewer slots than components at least one component has no slot to be assigned to and the task is ill-posed by construction.

Noisy priors. Perturbation of the concentration prior with multiplicative Gaussian noise is reported in Obs.5 with the protocol in Appendix I. Together, these three axes, absent priors, misspecified K, and noisy concentrations, indicate that MUSE extracts component identity primarily from the spectrum itself and treats the priors as auxiliary rather than load-bearing inputs.

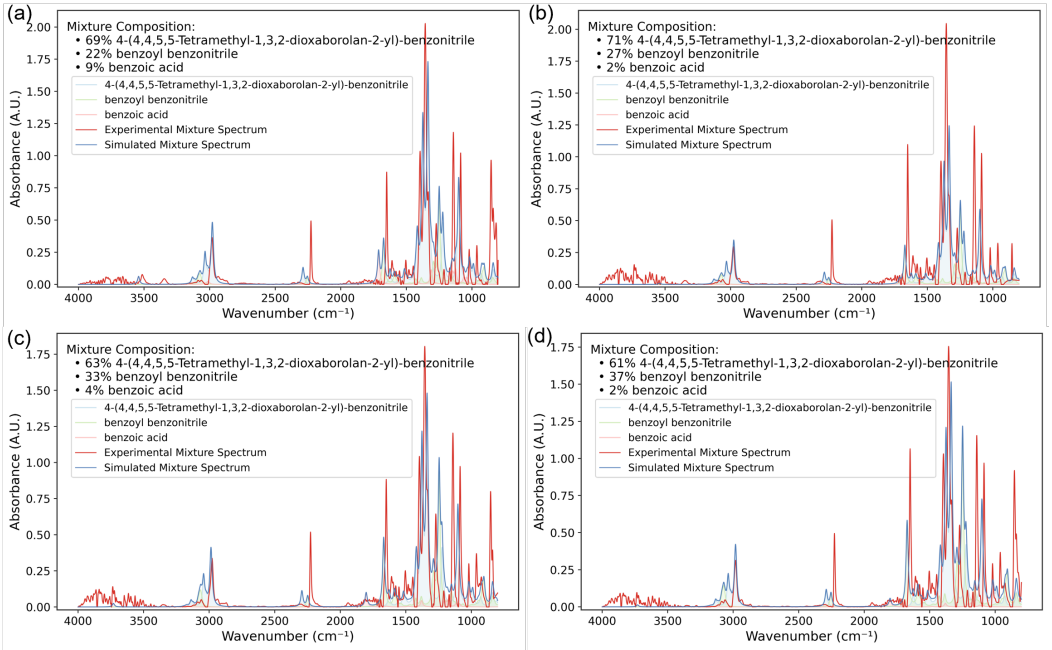

Figure 6: Comparison of experimental and synthetic mixture infrared spectra. The red lines represent the real mixture spectra obtained from experiments, while the blue lines denote the synthetic spectra generated by the linear combination.

K More Spectra Prediction Results

We provide additional predicted pure-component spectra together with their cosine similarities to the ground truth. Results on 2-component and 3-component mixtures are shown in Fig. 9 and Fig. 10, respectively.

L Matched Baseline Architectures

To separate the contribution of set prediction from those of the subspace split and the generative objective, Table 3 includes three controls that share the encoder, curriculum, optimizer, and training budget of MUSE.

Slot-only. This variant keeps the K learnable slot queries, the cross-attention over the shared spectral encoding, and Hungarian matching, but removes the subspace split, the cross-covariance penalty, and the CFM branch. Its classification head consumes the full d-dimensional slot embedding rather than the $d/2$-dimensional structural block, so it recovers most of the parameters freed by deleting the CFM vector field, which is itself a lightweight four-block 1D-CNN with 64 hidden channels. The two models are therefore comparable in scale, and the comparison isolates permutation-equivariant set prediction as such.

Independent branches with reconstruction. This variant removes slot competition while retaining set prediction and reconstruction. The global spectral representation is replicated K times, each copy is concatenated with the corresponding concentration and formula embedding, and K independent MLP branches each perform functional-group prediction and CFM reconstruction. Hungarian matching is retained, so the output is still an unordered set, but the components no longer compete for a shared spectral context through cross-attention.

Non-generative auxiliary head. This variant replaces the CFM branch with a two-layer MLP that regresses the concentration $c^{(k)}$ from $z_{\text{conc}}^{(k)}$ under a mean-squared-error loss, weighted so that its magnitude is comparable to $\mathcal{L}_{\text{cfm}}$. It therefore supplies auxiliary supervision of similar scale and cost that is attached to the same subspace, but without generative reconstruction of the pure spectrum.

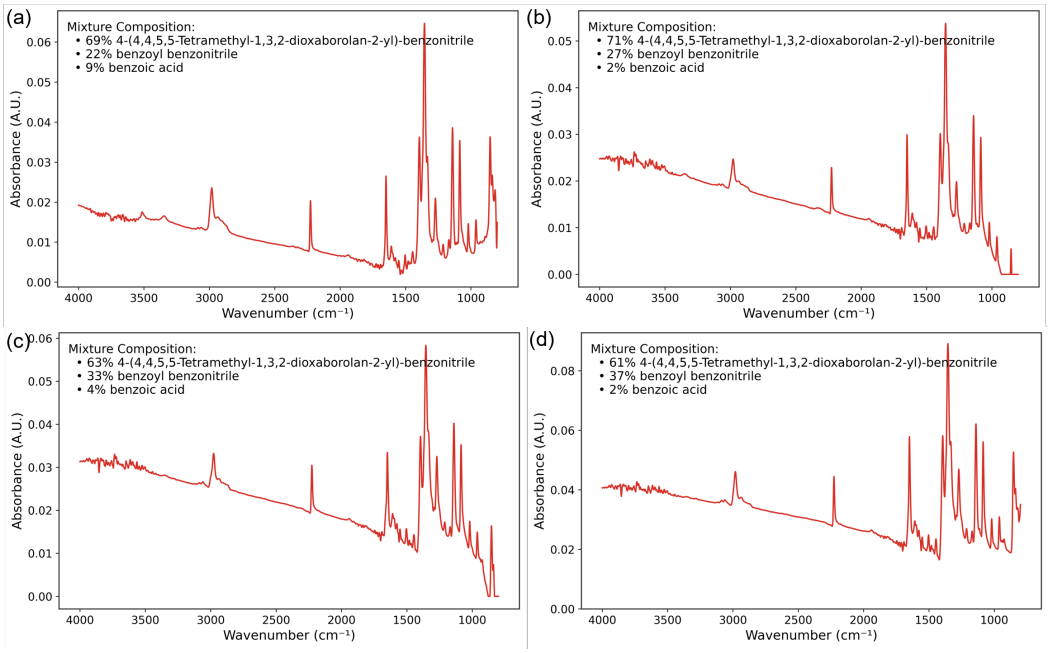

Figure 7: Experimental spectra of the real mixtures. The panel labels (a)–(d) correspond to those in Figure 6

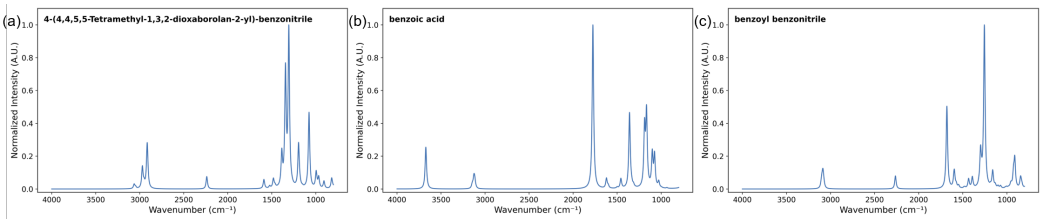

Figure 8: Simulated infrared spectra of the three individual pure components.

It serves as the control for the hypothesis that the gain of CFM is merely the effect of adding any auxiliary objective.

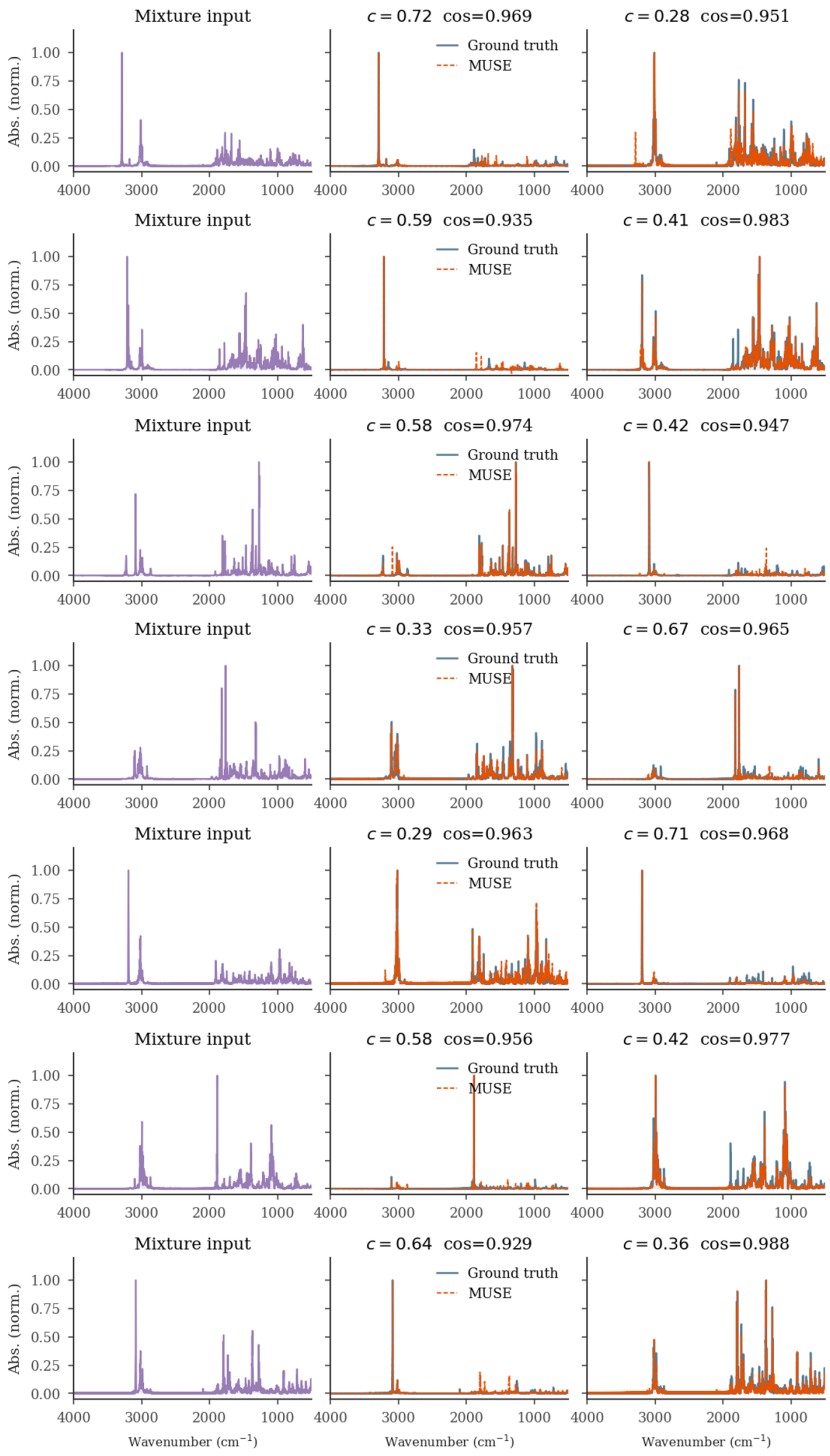

Figure 9: Additional predicted pure-component spectra on 2-component mixtures. For each mixture input (left), MUSE reconstructs the constituent components (orange) and we compare them with the ground truth (blue). The concentration and cosine similarity to the ground truth are reported above each panel.

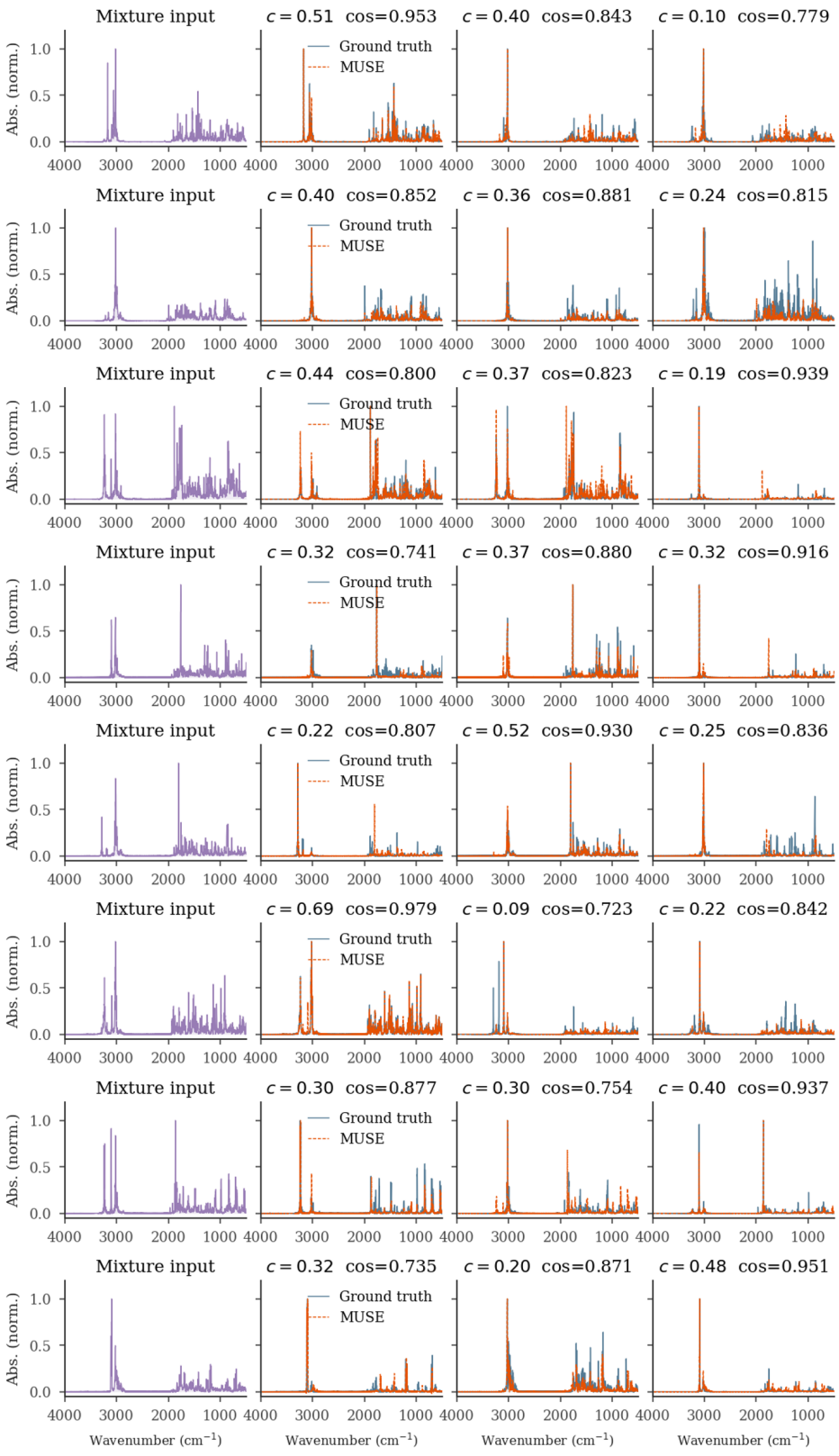

Figure 10: Additional predicted pure-component spectra on 3-component mixtures. For each mixture input (left), MUSE reconstructs the three constituent components (orange) and we compare them with the ground truth (blue). The concentration and cosine similarity to the ground truth are reported above each panel.