Language Model Enabled Structure Prediction from Infrared Spectra of Mixtures

Marvin Alberts[1,2,3†] **Filippo Ficarra**[1†] **Teodoro Laino**[1,2]
[1]IBM Research [2]University of Zürich [3]NCCR Catalysis
marvin.alberts@ibm.com

Abstract

Transformer language models recently enabled molecular structure prediction directly from infrared (IR) spectra, yet have remained confined to pure compounds. We show that the same architecture learns the correlations embedded in binary mixture spectra and can retrieve the individual molecular components. Trained solely on gas-phase data, our model attains a Top–10 accuracy of 61.4% on balanced synthetic mixtures. When evaluated on 15 mixtures measured with Attenuated Total Reflectance (ATR) IR spectrometer, whose response differs markedly from the training domain, it still achieves 52.0% Top–10 accuracy, evidencing strong cross-instrument transferability. The ability to identify signals of individual molecules within complex spectra extends machine-learning-assisted spectroscopy from idealised samples to realistic laboratory scenarios. All code and pretrained weights are released to accelerate adoption and further development. This advance opens the door to automated structure elucidation using IR data in fields ranging from environmental monitoring to pharmaceutical quality control.

1 Introduction

Infrared (IR) spectroscopy is an essential analytical technique employed across diverse fields, including chemistry, pharmaceuticals, materials science, and forensic science[1–4]. While IR spectroscopy excels at revealing the presence or absence of specific functional groups through characteristic absorption peaks at defined wavelengths, extracting more detailed structural information, such as the molecular scaffold or the complete structure, directly from the spectra without relying on database searches has long been considered an impossible task[5, 6].

The ability of artificial intelligence (AI) models to capture correlations among weak fingerprint signals embedded in complex spectra has opened new avenues for molecular structure elucidation, offering the potential to overcome traditional limitations by directly predicting structures from spectra without the need for database searches. In particular, the work of the IBM team [7, 8] has the merit of showing that Transformer-based language models can learn to translate IR spectra into detailed molecular structures with unprecedented accuracy [7, 8]. More broadly, AI-driven techniques have proven capable of predicting molecular structures directly from various individual spectroscopic modalities, including IR, nuclear magnetic resonance (NMR), and tandem mass spectrometry (MS/MS) [9–14]. Complementing these single-modality advances, multimodal methods [15, 16] that fuse data from multiple spectroscopic sources have begun to unlock deeper structural understanding beyond the reach of traditional approaches.

However, despite the advances in combining different spectroscopical modalities, nearly all existing approaches have been limited to spectra of single, pure compounds. Applying AI to the spectra

[†] Equal Contributions

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Accelerated Materials Design.

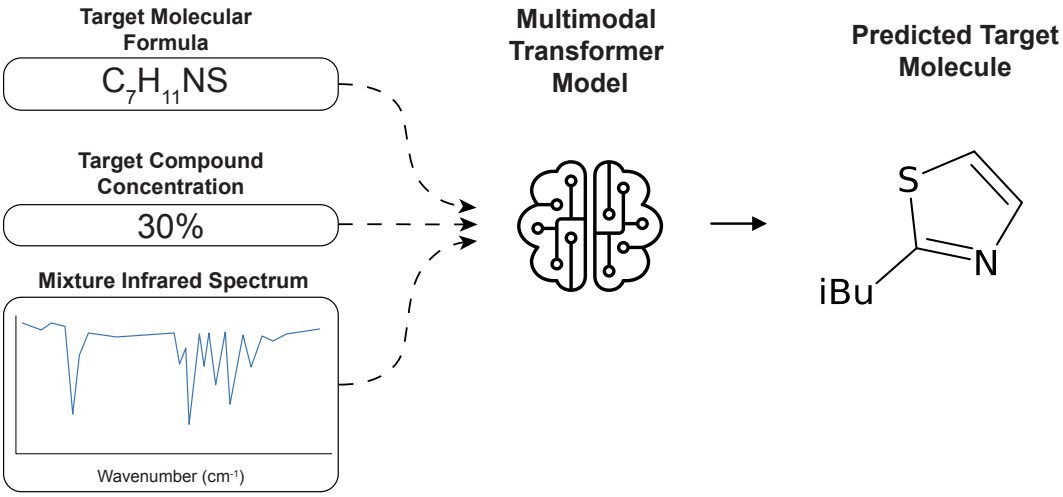


Figure 1: Modelling pipeline overview. Given the IR spectrum of a mixture, the concentration of one of the components present in the mixture and its molecular formula, the model predicts the corresponding molecule.

of mixtures has remained largely unexplored due to several challenges, including the scarcity of annotated spectral data for mixtures, the combinatorial complexity arising from overlapping signals, and the fundamental uncertainty about whether a model can disentangle and correctly assign individual molecular signatures that may interact non-linearly within a mixture. This challenge is further compounded by overlapping bands and spectral interferences, which can mask or distort the peaks of individual molecules, making reliable identification considerably more difficult. Traditional approaches for analysing spectra of mixtures rely heavily on database searches, wherein experimental spectra are compared against reference libraries to identify potential matches [17–20]. While effective in many practical scenarios, this strategy has an inherent limitation: it depends entirely on the coverage of the reference database. If a compound is absent from the database, it cannot be identified.

Here, for the first time, we demonstrate that it is possible to leverage the Transformer architecture to directly identify the individual components of chemical mixtures from IR spectra. This approach marks a substantial departure from traditional database-dependent methods, enabling comprehensive mixture analysis without the limitations imposed by reference libraries. Our model successfully predicts the correct components present in a mixture within the Top–10 candidates in up to 72.3% of cases.

2 Results

2.1 Acquisition of IR Spectra for Mixtures

As with any AI-driven approach, the primary limitation is the availability of sufficient training data. While large datasets of simulated IR spectra and smaller collections of experimental spectra exist, to the best of our knowledge, no publicly available database provides IR spectra of mixtures under acceptable licensing terms. To address this limitation, we designed a well-defined protocol for creating synthetically generated mixture data.

Most database search methods for mixture analysis rely on the principle that the IR spectrum of a mixture can be approximated as a linear combination of the spectra of its individual components [21]. These methods typically identify the reference spectrum with the greatest spectral overlap, subtract it from the mixture spectrum, and iteratively detect additional components [22, 23].

Here, we adopt the same principle to generate synthetic mixture spectra by combining spectra of pure compounds. This data generation strategy is illustrated in Figure 2 A), and a formal definition is provided in SI section 1.

To validate our approach, we measured the IR spectra of four pure compounds and their corresponding binary mixtures: 1) Cyclohexylamine, 2) N,N-Dimethylethylenediamine, 3) 4-Chloronitrobenzene, and 4) 1,4-diazabicyclo[2.2.2]octane.

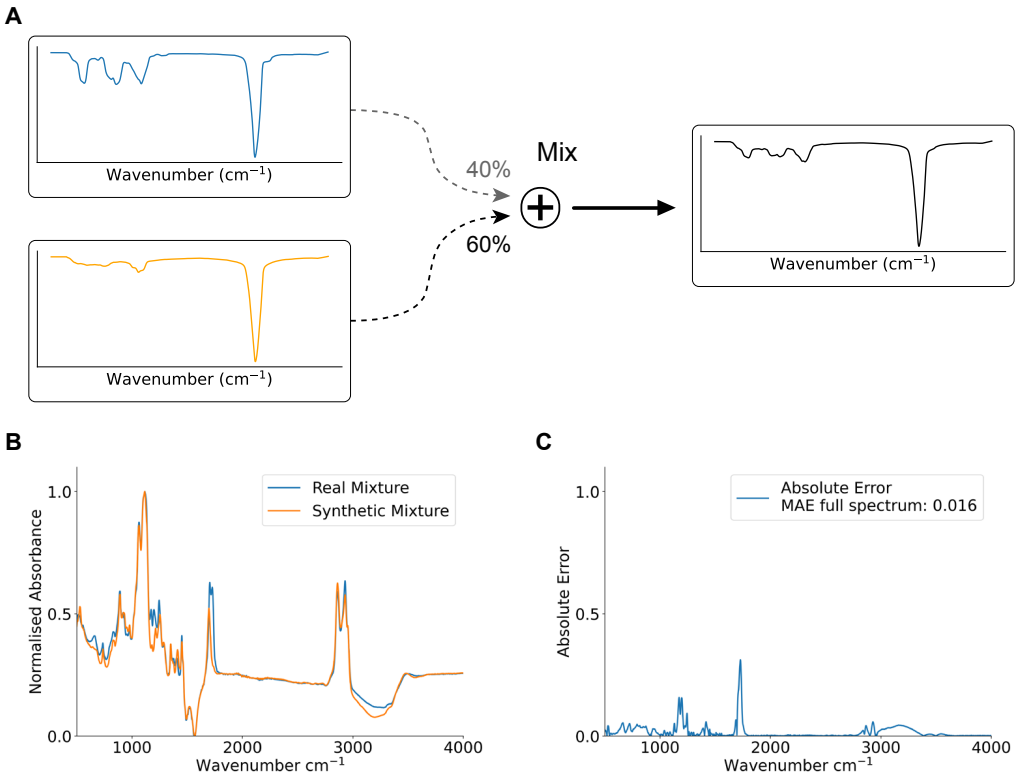


Figure 2: Generating synthetic mixture spectra. A) Spectra are generated as linear combinations of the IR spectra of the individual components. B) Example of the IR spectrum of a real mixture compared to a synthetic linear combination of pure spectra. C) Absolute error between the real mixture spectrum and the synthetic one.

We then evaluated the effectiveness of our synthetic data generation strategy by comparing experimentally measured mixture spectra with spectra obtained by linearly combining the corresponding pure component spectra. We evaluated binary balanced mixtures as well as unbalanced binary mixtures. The results, shown in Figure 2 B) and C) as well as SI Figure 1, demonstrate that linear combinations of individual component spectra provide an excellent approximation of the experimental mixture spectra. The synthetic spectra exhibit a mean absolute error (MAE) of 0.066 relative to the experimental spectra. To contextualise this error, we added Gaussian noise to the experimental mixture spectra, resulting in MAE values of 0.040, 0.056, and 0.080 for noise levels $\sigma_{\text{noise}} = 0.05$, 0.07, and 0.10, respectively. This indicates that the discrepancy between synthetic and experimental mixture spectra is comparable to the magnitude of moderate experimental noise.

Based on these results, we synthetically generated mixture spectra by linearly combining individual component spectra. However, the limited availability of experimental IR spectra remains a constraint. To overcome this, we adopt a two-stage training strategy: we first pretrain our model on simulated spectra and then fine-tune it on experimental data, an approach that has proven effective in numerous previous studies [7, 16].

For pretraining, we utilised the dataset published by Alberts *et al.* [24], which contains simulated IR spectra for approximately 790,000 pure compounds. These spectra were generated using molecular dynamics with the GAFF force field [25]. The scale of this dataset presents unique challenges: although binary combinations of these compounds could theoretically yield over 10^{12} synthetic mixture spectra (assuming equal proportions), generating, storing, and efficiently shuffling data at this scale during training would create substantial computational bottlenecks. We address these challenges in detail in SI section 3.

For fine-tuning, we used the NIST gas-phase infrared database [26], adopting the same subset as Alberts *et al.* [7]. This subset comprises gas-phase IR spectra of 3,453 pure compounds with heavy atom counts ranging from 6 to 13, providing a representative sample of small to medium-sized organic

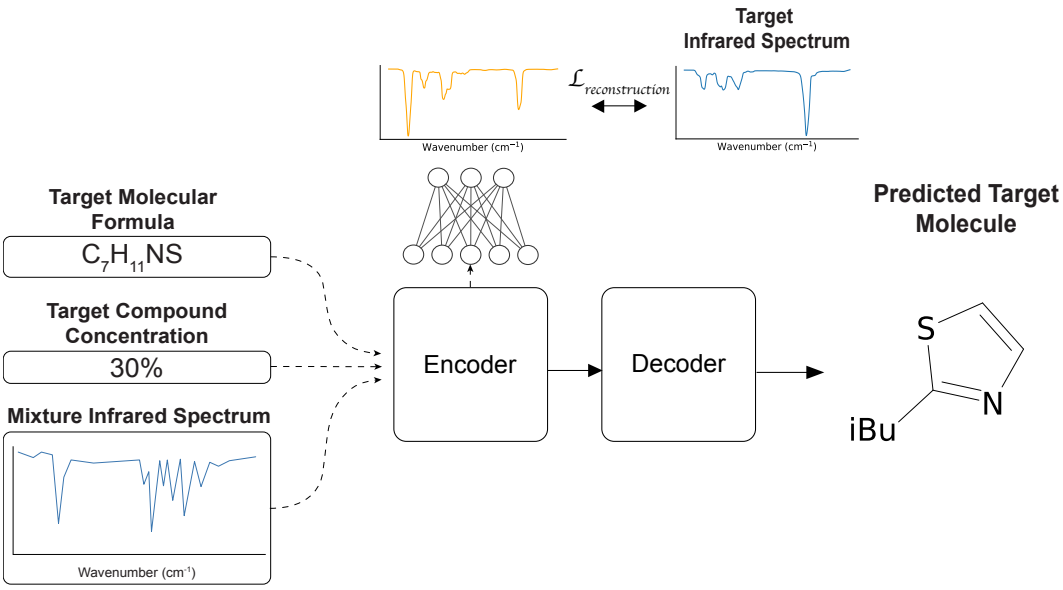


Figure 3: Transformer model with alignment mechanism. A reconstruction network aligns the encoder representations by reconstructing the IR spectrum of the pure target compound. The reconstruction network is trained jointly with the decoder, predicting the molecular structure.

molecules. To prevent data leakage, we partitioned the spectra of pure compounds into training, validation, and test sets before generating mixture spectra for each partition. During fine-tuning on experimental data, we employed five-fold cross-validation to ensure robust and representative performance estimates.

2.2 A model optimised for mixtures

Having validated our synthetic data generation strategy, we next addressed the challenge of training a model to identify mixture components. Drawing inspiration from prompt engineering, we formulated the problem as follows: given an IR spectrum of a mixture, the fractional concentration of one component, and its molecular formula, predict the corresponding molecule as an SMILES string [27]. In effect, we prompt the model with the molecular formula and ask it to predict the component present in the mixture that matches this formula. Experimentally, the molecular formula and approximate concentration can be determined using liquid chromatography–mass spectrometry (LC–MS). Our approach can thus be summarised as:

$$\text{MOLECULAR FORMULA}_{\text{TARGET}} \mid \%_{\text{TARGET}} \mid \text{SPECTRUM}_{\text{MIXTURE}} \rightarrow \text{SMILES}_{\text{TARGET}}$$

We extended the encoder–decoder architecture proposed by Alberts *et al.* [8] to predict the components of mixtures. An overview of the modelling approach is shown in Figure 1. Following Wu *et al.* [9] and Alberts *et al.* [8], we embed IR spectra using a patch-based approach with a fixed patch size of 150 cm^{-1}. Molecular formulae and concentrations are tokenised and embedded as text, and we adopt the tokenisation scheme introduced by Schwaller *et al.* [28] to encode SMILES strings.

A key difference between the task of decoding mixture components and previous work on pure compounds is that the encoder representations do not contain spectral information exclusive to the target component. Instead, the mixture spectrum includes overlapping signals from multiple compounds that must be disentangled. To address this challenge, we incorporated an additional inductive bias through a spectral alignment network: a reconstruction module learns to predict the IR spectrum of the pure target compound from the encoder outputs, as illustrated in Figure 3. This architecture encourages the encoder to learn representations that are specific to the target component of interest. With this modification, our total loss function becomes:

$$\mathcal{L} = \mathcal{L}_{S2S} + \lambda \mathcal{L}_{\text{reconstruction}}$$

Table 1: Ablation study comparing reconstruction network architectures, loss functions, and optimal weighting parameters λ on the binary balanced dataset. Best results are highlighted in bold with the models evaluated on the Top–1, 5 and 10 accuracies as well as the Tanimoto similarity of the predicted molecules to the ground truth.

Reconstruction Network	Reconstruction Loss	λ	Top–1 Acc. [%] ↑	Top–5 Acc. [%] ↑	Top–10 Acc. [%] ↑	Tanimoto ↑
None	N/A.	N/A.	16.0	29.0	32.9	0.543
CNN	MSE	5	17.7	31.7	36.3	0.566
CNN	MAE	50	**19.7**	**34.4**	**39.2**	**0.596**
MLP	MSE	1	17.2	30.9	35.0	0.561
MLP	MAE	1	18.9	33.7	38.2	0.579

Table 2: Performance comparison between baseline and aligned models on binary balanced mixtures evaluated based on the Top–1, 5 and 10 accuracy as well as tanimoto similarity. The aligned model uses convolutional reconstruction network, MAE loss, and $\lambda = 50$.

Model Configuration	Top–1 Acc. [%] ↑	Top–5 Acc. [%] ↑	Top–10 Acc. [%] ↑	Tanimoto ↑
Baseline (fine-tuned)	17.0 ± 1.0	46.5 ± 1.0	59.8 ± 2.2	0.553 ± 0.004
Aligned (fine-tuned)	19.9 ± 1.5	48.6 ± 2.0	61.4 ± 1.4	0.576 ± 0.008

where $\mathcal{L}_{S2S}$ is the sequence-to-sequence loss and $\mathcal{L}_{\text{reconstruction}}$ is the spectrum reconstruction loss, while λ is used as a hyperparameter tuning the impact of the reconstruction loss.

We systematically explored different alignment configurations to optimise the spectrum reconstruction. A Multi-Layer Perceptron (MLP) and a Convolutional Neural Network (CNN) were evaluated in combination with either Mean Absolute Error (MAE) or Mean Squared Error (MSE) as the reconstruction loss function. For each configuration, the weighting parameter for the alignment loss, λ, was tuned. All models, including a baseline without an alignment loss, were evaluated on binary balanced mixtures (50% of each component) generated by linearly combining simulated spectra. Performance was assessed using Top-1, Top-5, and Top-10 accuracies, defined as the percentage of correctly predicted molecules within the top n predictions as well as the mean tanimoto similarity between the Top–1 prediction and the ground truth. A tanimoto similarity above 0.5 indicates a high similarity between two molecules. The results are summarised in Table 1.

The results reveal clear preferences for both architecture and loss function. CNN-based reconstruction networks consistently outperform MLP architectures across all evaluation metrics, with the CNN-MAE combination achieving the best performance. Notably, the optimal λ values differ substantially between configurations, with CNN-MAE requiring $\lambda = 50$ compared to $\lambda = 1$ for MLP variants. This suggests that convolutional architectures benefit from stronger weighting of the reconstruction signal.

The superior performance of the CNN architectures likely arises from their ability to capture local spectral patterns and spatial relationships, while the MAE loss better preserves sparse spectral features compared to MSE. Complete ablation results, including additional λ values and architectural variants, are provided in SI Table 2.

We further validated our approach by fine-tuning the models on the NIST database using five-fold cross-validation. Both the baseline model and the optimal alignment configuration (CNN reconstruction network with MAE loss, $\lambda = 50$) were fine-tuned, with the results presented in Table 2. The aligned model shows a clear performance improvement over the baseline, demonstrating the effectiveness of incorporating spectrum reconstruction into the encoder representations.

The alignment mechanism yields substantial improvements across all evaluation metrics. On simulated data, the aligned model achieves a 3.7% increase in Top-1 accuracy (from 16.0% to 19.7%) and consistent gains of 5–6 percentage points for Top-5 and Top-10 accuracies. Although fine-tuning reduces the performance gap between the baseline and aligned models, the aligned model still outperforms the baseline, particularly for Top-1 predictions. These results demonstrate that incorporating

Table 3: Performance of the multi-task model trained simultaneously on multiple imbalance ratios.

	PERCENTAGE	TOP–1 ACC. [%] ↑	TOP–5 ACC. [%] ↑	TOP–10 ACC. [%] ↑	TANIMOTO ↑
Multi-task Fine-tuned	10%	8.7 ± 1.2	26.5 ± 2.2	37.8 ± 2.2	0.472 ± 0.014
	20%	11.6 ± 0.8	31.3 ± 2.0	43.5 ± 2.2	0.509 ± 0.018
	30%	13.6 ± 1.6	34.8 ± 2.8	47.2 ± 2.6	0.534 ± 0.017
	40%	14.9 ± 2.1	37.3 ± 3.1	50.3 ± 2.6	0.549 ± 0.019
	50%	19.6 ± 1.6	45.5 ± 2.4	57.5 ± 2.1	0.573 ± 0.009
	60%	25.3 ± 1.4	53.6 ± 2.3	63.9 ± 2.3	0.606 ± 0.007
	70%	28.4 ± 0.9	57.2 ± 2.3	66.9 ± 2.1	0.628 ± 0.006
	80%	30.5 ± 1.3	60.0 ± 2.3	69.4 ± 2.1	0.644 ± 0.008
	90%	33.2 ± 1.7	62.1 ± 2.4	71.6 ± 1.9	0.658 ± 0.011

an inductive bias through the reconstruction loss enhances the model's ability to predict molecular structures directly from the spectra of mixtures.

2.3 Multitask learning enables the processing of imbalanced mixtures

Our optimised model demonstrates strong performance on binary balanced mixtures; however, real-world applications rarely involve exact 50:50 component ratios. In practice, imbalanced mixtures are far more common. To assess the robustness of our approach under realistic conditions, we conducted experiments on imbalanced mixtures. A fundamental challenge in mixture analysis is that component concentrations derived from LC–MS data are often imprecise, and varying absorption intensities can cause the spectral contribution of each component to deviate from its actual concentration. Consequently, the model must be capable of predicting components across a wide range of mixture compositions.

To address this challenge, we adopted a multitask learning approach and trained a single model to predict the components of imbalanced mixtures. This unified multitask model was trained simultaneously on binary mixtures with the following composition ratios: $\text{BINARY}_{50\%\text{-}50\%}$, $\text{BINARY}_{40\%\text{-}60\%}$, $\text{BINARY}_{30\%\text{-}70\%}$, $\text{BINARY}_{20\%\text{-}80\%}$, and $\text{BINARY}_{10\%\text{-}90\%}$. We employed the best configuration found in section 3.2, incorporating spectral alignment with a convolutional neural network, MAE loss, and $\lambda = 50$, pretrained on simulated spectra and finetuned on synthetic mixture spectra derived from the NIST database[26]. Additionally, we performed an ablation study comparing single-task models trained on individual mixture ratios, with detailed results provided in SI section 5.

The results presented in Table 3 reveal a clear correlation between model performance and the prevalence of the target compound. Performance declines when the target compound constitutes only 10% of the mixture, highlighting the inherent challenge of detecting low-concentration components in spectroscopic data. This trend reflects a fundamental limitation: weaker spectral signals from minority components become increasingly difficult to distinguish from background noise and dominant spectral features of the majority component.

Despite these challenges, the multi-task model maintains reasonable performance across all tested ratios, demonstrating robustness and versatility for diverse analytical scenarios. The gradual performance degradation, rather than abrupt failure, suggests that the model successfully learns to extract relevant spectral features even under challenging conditions.

2.3.1 Extension to Ternary Mixtures

To demonstrate the scalability of our approach, we evaluated the performance of our architecture on ternary balanced mixtures, representing a more complex spectral deconvolution challenge with three components. Adopting the same modelling approach as in section 3.3 and evaluating on spectra from the NIST database[26].

The ternary mixture results shown in Table 4 highlight the increased complexity of multi-component spectral analysis. While the Top-1 accuracy remains modest, the substantial improvements in the

Table 4: Model performance on ternary balanced mixtures. The increased complexity of three-component systems presents additional challenges, with fine-tuning providing substantial improvements in Top-5 and Top-10 accuracy.

	TOP–1 ACC. [%] ↑	TOP–5 ACC. [%] ↑	TOP–10 ACC. [%] ↑	TANIMOTO ↑
Simulated	9.5	19.2	22.7	0.472
Fine-tuned	10.1 ± 0.5	31.9 ± 1.4	44.8 ± 1.6	0.479 ± 0.002

Table 5: Accuracy of our multitask model predicting 30 compounds by using 15 *real mixtures* spectra.

	REJECTION SAMPLING	TOP–1 ACC. [%] ↑	TOP–5 ACC. [%] ↑	TOP–10 ACC. [%] ↑	TANIMOTO ↑
Multi-task Fine-tuned	✗	6.0 ± 5.7	24.7 ± 4.5	40.0 ± 6.3	0.239 ± 0.056
Multi-task Fine-tuned	✓	6.0 ± 5.7	34.0 ± 3.9	52.0 ± 4.0	0.247 ± 0.055

Top-5 and Top-10 metrics indicate that the model effectively narrows the candidate space. This is valuable for practical applications in which human experts can make final identifications from a reduced set of plausible candidates.

2.4 Validation on real spectra

All previous evaluations were conducted on synthetically generated mixture spectra created by linearly combining individual compound spectra. While this approach was validated experimentally, as described in section 3.1, the synthetic spectra still represent high-quality approximations rather than true mixture spectra. To further validate our method, we evaluated the performance of our model on experimentally measured mixture spectra. Although our fine-tuning dataset comprises gas-phase IR spectra, our experimental measurements were limited to ATR-IR spectroscopy. We measured 15 binary mixtures using ATR-IR and tested the multitask model on these spectra. The results are presented in Table 5.

Despite the domain shift between gas-phase and ATR-IR spectra, our model demonstrated robust performance, achieving a Top-10 accuracy of 40.0% on experimentally measured mixture spectra. This result confirms the transferability of models trained on synthetic mixture data to real measurements, with even higher performance expected for models fine-tuned directly on ATR-IR spectra. To further improve performance, we implemented a rejection sampling strategy that discards predictions yielding invalid SMILES strings or molecular structures inconsistent with the provided molecular formula. This approach resulted in an approximate 12% increase in both Top-5 and Top-10 accuracies.

These findings demonstrate that models trained exclusively on synthetic mixture spectra can generalise effectively to real-world spectra. The consistent performance across different instrumental techniques (gas-phase versus ATR-IR spectroscopy) indicates that our approach captures fundamental spectral relationships that are robust to experimental variations.

2.5 Application Domains Demanding Immediate Mixture Analysis

Rapid, database-independent deconvolution of IR spectra has immediate relevance in several applied domains. In environmental monitoring, regulatory bodies increasingly employ field-deployable IR instruments to screen water or air samples for complex contaminant mixtures (e.g., phthalates, per- and polyfluoroalkyl substances, VOCs)[29–31]. Automated identification of minor components at trace levels enables near-real-time decisions during spill response or industrial emission audits, where waiting for laboratory GC–MS confirmation can delay mitigation. In pharmaceutical manufacturing, inline IR probes are used in process chemistry to verify content uniformity, detect residual solvents, and track degradation products in solid-dose and biopharmaceutical formulations[32, 33]. A model that can disentangle overlapping excipient and active-ingredient bands directly on the produc-

tion line would shorten batch-release times and improve quality control robustness. In untargeted metabolomics, high-throughput IR microarrays analyse biofluids whose spectral fingerprints reflect hundreds of metabolites with wide concentration ranges; rapid component prediction could prioritise candidates for follow-up LC–MS/MS, accelerating biomarker discovery in toxicology and clinical studies[34, 35]. Across these scenarios, the framework presented in this paper provides the speed and breadth needed to complement or, in time-critical settings, temporarily replace more labour-intensive chromatographic methods.

3 Conclusion

In this study, we developed a language model-driven approach for predicting individual molecular components directly from IR spectra of mixtures. We demonstrated that synthetic mixture spectra, generated by linearly combining pure compound spectra, can be effectively leveraged to train models capable of analysing real-world data. In addition, we introduced a spectral alignment strategy that reconstructs the pure spectrum of the target compound, resulting in substantial performance gains. This alignment mechanism addresses the fundamental challenge that mixture spectra do not contain isolated information for each component by ensuring that encoder representations retain component specific spectral features. Our optimised model achieves up to 61.4% Top-10 accuracy on synthetic binary balanced mixtures. Using multitask learning, we extended this approach to imbalanced mixtures, maintaining performance even when the target compound accounts for only 10% of the mixture. Most importantly, validation on 15 experimentally measured mixtures confirms the real-world applicability of our method: despite a significant domain shift from gas-phase training data to ATR-IR measurements, our model achieves Top-10 accuracies of up to 52.0%. The ability to train on synthetic mixture spectra and transfer effectively to real experimental conditions opens new possibilities for AI-assisted analysis of complex mixtures. The same principles can be extended to other spectroscopic techniques, such as NMR spectroscopy, or to multimodal approaches that integrate multiple types of spectra. We envision these models as integral parts of a collaborative human–AI workflow that combines the strengths of automated prediction with expert interpretation. AI models can provide rapid initial molecular candidates from spectroscopic data, while chemists apply their expertise to verify predictions, refine structural assignments using additional analytical evidence, and interpret results within the broader chemical context. By providing focused starting points, such models could significantly accelerate the structure elucidation process, transforming a traditional bottleneck into a more efficient and streamlined workflow.

4 Methods

4.1 Data

Simulated Data: The simulated data used in this study is sourced from our earlier paper introducing a multimodal spectroscopic dataset[24]. We use all molecules available in the dataset, filtering out only those that also appear in any of the experimental datasets. The IR spectra in this dataset were simulated using molecular dynamics with the GAFF force field. We further filter the dataset by excluding molecules with a heavy atom count (all atoms except hydrogen) outside the range of 5 to 35, as well as molecules containing elements other than carbon, hydrogen, oxygen, nitrogen, sulfur, phosphorus, or the halogens.

Experimental IR spectra: IR spectra were sourced from the NIST EPA Gas-Phase database[26]. All entries in the datasets were filtered to remove molecules either consisting of more than 13 or fewer than 6 heavy atoms or with elements not contained in the simulated dataset matching earlier work[7, 8].

4.2 Tokenisation and Preprocessing

The chemical formula and molecules represented as SMILES were provided to the model as text with the tokenisation procedures outlined below. IR spectra were embedded via patches.

Molecular Formula: All chemical formulae were tokenised using the following regular expression: `([A-Z]{1}[a-z]?[0-9]*)`

Concentration: Similarly, the concentration of our compound in the mixture is treated as text and tokenised using the following regular expression: `\d*\.\d+`

IR spectra: IR spectra were segmented into patches, with each patch projected into the embedding space via a multilayer perceptron (MLP). For all experiments, we used a patch size of 75.

Molecules: All molecules were canonicalised using RDKit[36] and tokenised using the same regular expression as employed by Schwaller *et al.* [28].

4.3 Model

The model employed in this work follows the encoder-decoder transformer architecture. Building upon the original implementation by Vaswani *et al.* [37] we leverage post layer normalisation[38], learned positional embeddings[39] and gated linear units[40]. The following hyperparameters were used to construct the model:

```
Layers: 6
Heads: 8
Embedding Dimension: 512
Feedforward Dimension: 2048
```

Alignment: To align the encoder representations, a neural network was used to convert the representations to the target molecule signal. The configuration for the CNN architecture used in the experiments is the following:

```
hidden_dimension: 256
conv_channels: 512
kernel_size: 5
output_dimension: 1800
loss_lambda: 50
loss_function: mae
```

Train-test splitting: For pretraining the data, a 70/20/10 train, test and validation split was used to further combine the spectra online as introduced in SI section 3. With the mixture combination, our training data consists of up to $3.2 * 10^8$ spectra, while test and validation consist of 10,000 spectra. All fine-tuning experiments were carried out with five-fold cross-validation using the same seed to ensure reproducibility, also with a 70/20/10 train, test and validation split.

Training settings: Training of the models was carried out on eight Nvidia A100 GPUs with an average pretraining time of $\sim$ 20h. When evaluating models, the best validation checkpoint was used. For each training run, the following training parameters were used. No distinction was made between pretraining and finetuning experiments:

```
Optimiser: AdamW
Learning Rate: 0.001
Dropout: 0.1
Warmup steps: 8000
Batch size: 128
```

4.4 Experimental Data Collection

Chemicals: Cyclohexylamine, N,N-Dimethylethylenediamine, 4-Chloronitrobenzene, 1,4-diazabicyclo-[2.2.2]-octane, Trans-1-Phenyl-1,3-butadiene, Diethylene glycol butyl ether, Cyclohexanecarboxylic acid and 2,6-Lutidine were purchased from SigmaAldrich; Cyclobutylamine was purchased from Alfa Aesar; Methyl-3-Bromopyruvate was purchased from Honeywell Fluka; N-Methylmorpholine was purchased from Apollo Scientific; Methyl-Glycolate was purchased from Tokyo Chemical Industry, and Propylene carbonate was purchased from Carl Roth GmbH.

IR spectra: IR spectra were acquired using a PerkinElmer Spectrum Two FTIR spectrometer with a diamond anvil ATR attachment (450–4000 cm^{-1}, 16 scans, 2 cm^{-1} resolution).

References

1. Stuart, B. H. *Infrared Spectroscopy: Fundamentals and Applications* (John Wiley & Sons, Ltd, 2004).
2. Visser, T. in *Encyclopedia of Analytical Chemistry* (John Wiley & Sons, Ltd, 2006).
3. Avram, M. & Mateescu, G. D. *Infrared Spectroscopy: Applications in Organic Chemistry* (R. E. Krieger Publishing Company, 1978).
4. Chalmers, J. E., Edwards, H. G. M. & Hargreaves, M. D. *Infrared and Raman Spectroscopy in Forensic Science* (John Wiley & Sons, Ltd, 2012).
5. *KnowItAll IR Spectral Database Collection* `https://sciencesolutions.wiley.com/solutions/technique/ir/knowitall-ir-collection/`.
6. Lin-Vien, D., Colthup, N. B., Fateley, W. G. & Grasselli, J. G. *The Handbook of Infrared and Raman Characteristic Frequencies of Organic Molecules* (Academic Press, 1991).
7. Alberts, M., Laino, T. & Vaucher, A. C. Leveraging infrared spectroscopy for automated structure elucidation. *Communications Chemistry* **7,** 1–11 (2024).
8. Alberts, M., Zipoli, F. & Laino, T. *Setting New Benchmarks in AI-driven Infrared Structure Elucidation* ChemRxiv: 10.26434/chemrxiv-2025-9p2dw. 2025.
9. Wu, W., Leonardis, A., Jiao, J., Jiang, J. & Chen, L. Transformer-Based Models for Predicting Molecular Structures from Infrared Spectra Using Patch-Based Self-Attention. *The Journal of Physical Chemistry* **129,** 2077–2085 (2025).
10. Jonas, E. *Deep imitation learning for molecular inverse problems* in *Advances in Neural Information Processing Systems* **32** (2019).
11. Sridharan, B., Mehta, S., Pathak, Y. & Priyakumar, U. D. Deep Reinforcement Learning for Molecular Inverse Problem of Nuclear Magnetic Resonance Spectra to Molecular Structure. *The Journal of Physical Chemistry Letters* **13,** 4924–4933 (2022).
12. Schilter, O., Alberts, M., Zipoli, F., Vaucher, A. C., Schwaller, P. & Laino, T. *Unveiling the Secrets of* ^{1}*H-NMR Spectroscopy: A Novel Approach Utilizing Attention Mechanisms* in *NeurIPS 2023, AI4Science Workshop* (2023).
13. Devata, S., Sridharan, B., Mehta, S., Pathak, Y., Laghuvarapu, S., Varma, G. & Priyakumar, U. D. DeepSPInN – deep reinforcement learning for molecular structure prediction from infrared and 13C NMR spectra. *Digital Discovery* **3,** 818–829 (2024).
14. Stravs, M. A., Dührkop, K., Böcker, S. & Zamboni, N. MSNovelist: de novo structure generation from mass spectra. *Nature Methods* **19,** 865–870 (2022).
15. Priessner, M., Lewis, R., Janet, J. P., Lemurell, I., Johansson, M., Goodman, J. & Tomberg, A. *Enhancing Molecular Structure Elucidation: MultiModalTransformer for both simulated and experimental spectra* ChemRxiv: 10.26434/chemrxiv-2024-zmmnw. 2024.
16. Alberts, M., Hartrampf, N. & Laino, T. *Automated Structure Elucidation at Human-Level Accuracy via a Multimodal Multitask Language Model* ChemRxiv: 10.26434/chemrxiv-2025-q80r9. 2025.
17. Rasmussen, G. T., Isenhour, T. L., Lowry, S. R. & Ritter, G. L. Principal component analysis of the infrared spectra of mixtures. *Analytica Chimica Acta* **103,** 213–221 (1978).
18. Smith, B. C. *Infrared Spectral Interpretation: A Systematic Approach* (CRC Press, Boca Raton, 1999).
19. Smith, B. More Theory and Practice: The Thorny Problem of Mixtures and More on Straight Chain Alkanes. *Spectroscopy* **30,** 26–31, 48 (2015).
20. *FTIR Analysis Software* `https://www.thermofisher.com/ch/en/home/industrial/spectroscopy-elemental-isotope-analysis/molecular-spectroscopy/fourier-transform-infrared-spectroscopy/software.html`.
21. Myers, T. L., Bernacki, B. E., Wilhelm, M. J., Jensen, K. L., Johnson, T. J., Primera-Pedrozo, O. M., Tonkyn, R. G., Smith, S. C., Burton, S. D. & Bradley, A. M. Influence of intermolecular interactions on the infrared complex indices of refraction for binary liquid mixtures. *Physical Chemistry Chemical Physics* **24,** 22206–22221 (2022).
22. Chen, C.-S., Li, Y. & Brown, C. W. Searching a mid-infrared spectral library of solids and liquids with spectra of mixtures. *Vibrational Spectroscopy* **14,** 9–17 (1997).
23. Li, J., Hibbert, D. B., Fuller, S. & Vaughn, G. A comparative study of point-to-point algorithms for matching spectra. *Chemometrics and Intelligent Laboratory Systems* **82,** 50–58 (2006).

24. Alberts, M., Schilter, O., Zipoli, F., Hartrampf, N. & Laino, T. *Unraveling molecular structure: A multimodal spectroscopic dataset for chemistry* in *Advances in Neural Information Processing Systems* **37** (2024), 125780–125808.
25. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general amber force field. *Journal of computational chemistry* **25,** 1157–1174 (2004).
26. Stein, S. E. *NIST Standard Reference Database 35: NIST/EPA Gas-Phase Infrared Database - JCAMP Format* Accessed: 2025-03-28. 2008. `https://www.nist.gov/srd/nist-standard-reference-database-35`.
27. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences* **28,** 31–36 (1988).
28. Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Hunter, C. A., Bekas, C. & Lee, A. A. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Central Science* **5,** 1572–1583 (2019).
29. D'Arco, A., Mancini, T., Paolozzi, M. C., Macis, S., Mosesso, L., Marcelli, A., Petrarca, M., Radica, F., Tranfo, G., Lupi, S. & Della Ventura, G. High Sensitivity Monitoring of VOCs in Air through FTIR Spectroscopy Using a Multipass Gas Cell Setup. *Sensors* **22,** 5624 (2022).
30. Zhou, J., Al Husseini, D., Li, J., Lin, Z., Sukhishvili, S., Coté, G. L., Gutierrez-Osuna, R. & Lin, P. T. Detection of volatile organic compounds using mid-infrared silicon nitride waveguide sensors. *Scientific Reports* **12,** 5572 (2022).
31. Baker, T. J., Tonkyn, R. G., Thompson, C. J., Dunlap, M. K., Koster van Groos, P. G., Thakur, N. A., Wilhelm, M. J., Myers, T. L. & Johnson, T. J. An infrared spectral database for gas-phase quantitation of volatile per- and polyfluoroalkyl substances (PFAS). *Journal of Quantitative Spectroscopy and Radiative Transfer* **295,** 108420 (2023).
32. Roggo, Y., Chalus, P., Maurer, L., Lema-Martinez, C., Edmond, A. & Jent, N. A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies. *Journal of Pharmaceutical and Biomedical Analysis. Drug Analysis 2006* **44,** 683–700 (2007).
33. Flores, Y. V., Polak, A., Jambet, J., Stothard, D. & Haertelt, M. Point of Interest Mid-Infrared Spectroscopy for Inline Pharmaceutical Packaging Quality Control. *IEEE Sensors Journal* **23,** 16115–16122 (2023).
34. Martens, J., Berden, G., van Outersterp, R. E., Kluijtmans, L. A. J., Engelke, U. F., van Karnebeek, C. D. M., Wevers, R. A. & Oomens, J. Molecular identification in metabolomics using infrared ion spectroscopy. *Scientific Reports* **7,** 3363. (2025) (2017).
35. Jeppesen, M. J. & Powers, R. Multiplatform untargeted metabolomics. *Magnetic Resonance in Chemistry* **61,** 628–653 (2023).
36. *RDKit* (Accessed April 14, 2025). `https://www.rdkit.org/`.
37. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. Attention is all you need. *Advances in neural information processing systems* **30** (2017).
38. Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., Zhang, H., Lan, Y., Wang, L. & Liu, T. *On Layer Normalization in the Transformer Architecture* in *Proceedings of the 37th International Conference on Machine Learning* **119** (2020), 10524–10533. `https://proceedings.mlr.press/v119/xiong20b.html`.
39. Gehring, J., Auli, M., Grangier, D., Yarats, D. & Dauphin, Y. N. *Convolutional Sequence to Sequence Learning* arXiv:1705.03122. 2017.
40. Shazeer, N. *GLU Variants Improve Transformer* arXiv:2002.05202. 2020.

A Formal definition for the generation of spectra of mixtures

Given a set of component spectra $\{s_i\}_{i=0}^{n}$, where each spectrum $s_i \in \mathbb{R}^m$ consists of m datapoints, and a vector of mixture proportions $\boldsymbol{\alpha} \in \mathbb{R}^n$ where $\boldsymbol{\alpha}_i$ represents the concentration of the i-th compound in the mixture, we compute the mixture spectrum $s_{\text{mixture}} \in \mathbb{R}^m$ as:

$$s_{\text{mixture}} = \sum_{i=0}^{n} \boldsymbol{\alpha}_i s_i \tag{1}$$

B Measured vs Synthetic Mixture Infrared Spectra

To justify the abstraction of **mixture IR spectra** as explained above, we measured four different compound's IR spectra and their combination. The compound we measured in lab are the following:

- **Sample1:** Cyclohexylamine.
- **Sample2:** N,N-Dimethylethylenediamine.
- **Sample3:** 4-Chloronitrobenzene.
- **Sample4:** 1,4-diazabicyclo[2.2.2]octane.

Figure 4 demonstrates that by linearly combining the spectra of individual components, we can obtain an approximation of the spectrum of a mixture.

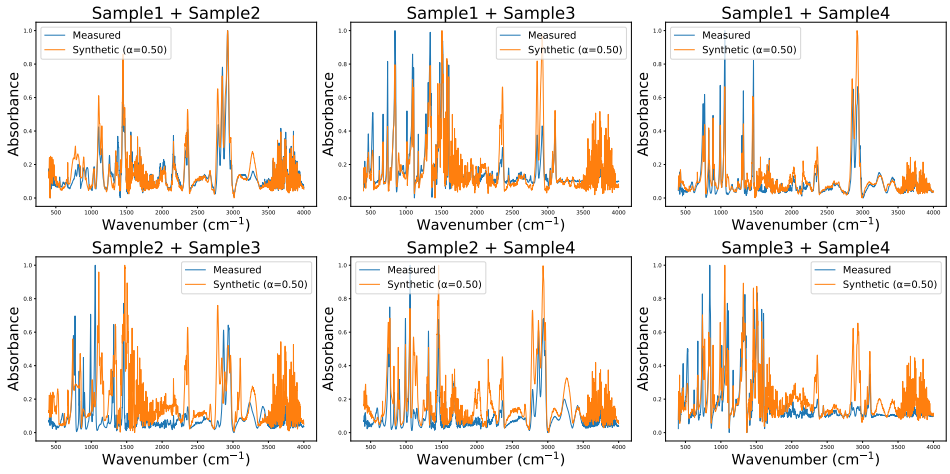


Figure 4: Measured mixture spectra vs linear combinations of mixture's components spectra.

We then extended this analysis to mixtures of alcohols and acids, a system with likely interactions influencing the IR spectrum. For this, we measured spectra both for binary balanced and binary unbalanced mixtures. Cyclohexanecarboxylic acid and diethylene glycol monobutyl ether were used, referred to as "Acid" and "Alcohol". Results are shown in Figure 5.

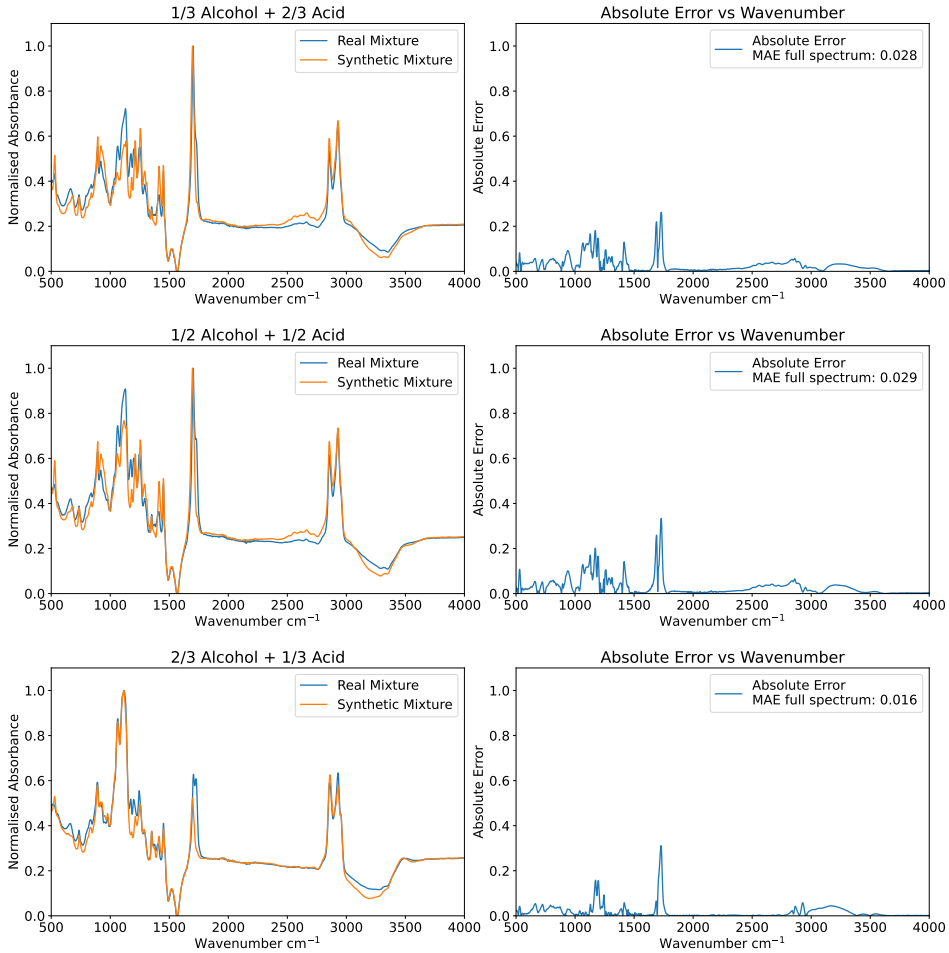


Figure 5: Synthetic mixture spectra compared to real measured mixtures.

C Mixture Dataset Generation

To generate the dataset for the mixtures we combine (or permute, when we consider imbalanced mixture) spectra. Let us consider the case of binary mixtures. The train split consists of $625,433$ spectra and when generating the mixtures we theoretically get:

- #combinations $\rightarrow \binom{625,433}{2} = 195,582,906,028$
- #permutations $\rightarrow \frac{625,433!}{(625,433-2)!} = 391,165,812,056$

Since the number of combinations (or permutations) are infeasible to be generated in a shuffled way, to avoid bias on over-represented training samples, we sampled using a random generator.

By sampling at random we do not guarantee that any combination (or permutation) is repeated, hence we have performed the following theoretical analysis.

Let s be the number of samples we draw from a set of N elements. The expected value of resamples R can be defined as

$$\mathbb{E}[R] = s - \mathbb{E}[U]$$

where U is the number of unique values sampled. Let $X_i = \begin{cases} 1 & \text{if } i \text{ is sampled at least once} \\ 0 & \text{otherwise} \end{cases}$

$$
\begin{aligned}
\implies \mathbb{E}[R] &= s - \mathbb{E}[U] \\
&= s - E\left[\sum_{i}^{N} X_i\right] \\
&= s - N\mathbb{E}[X_i] \\
&= s - N\left(1 - \left(1 - \frac{1}{N}\right)^{s}\right) \\
&\approx s - N\left(1 - e^{-\frac{s}{N}}\right)
\end{aligned}
$$

In the case of combinations (or permutations) N is #combinations (#permutations). We acknowledge that for increasing s we have an exponential number of conflicts, but in our case when the samples are #nsamples, the effect is almost negligible. Note that, when training in epochs usually samples are repeated *#epochs* times, so our concern is not merely on the number of repeated samples but on the bias towards over-represented population. Table 6 shows the empirical results for $\mathbb{E}[C]$ for permutations as presented in the proof above. The same result holds for combinations.

SAMPLES	EXPECTED RESAMPLES	RESAMPLES	NON-UNIQUE RESAMPLES
20M	511.28	487	0
40M	2,045.09	2061	0
80M	8,180.12	8125	1
160M	32,718.24	32979	6
320M	130,855.10	130559	40

Table 6: Permutations expected number of resamples, the actual number of resamples in our experiments and the number of elements resamples more the one time. This shows that even though for increasingly amount of data there is an exponential resampling, the effect is negligible since most of the resamples are present at most once.

D Ablation Study on the Encoder Alignment

Table 7 shows the result of or ablation study on the encoder alignment. We systematically explored various alignment configurations to optimise the spectrum reconstruction approach. The combination of two architectures for the reconstruction network, Convolutional Neural Network (CNN) and Multilayer Perceptron (MLP), two reconstruction losses, Mean Absolute Error (MAE) and Mean Squared Error (MSE), and in addition $\lambda \in [1, 5, 50]$.

E Single Task Models for Imbalanced Mixture

Real-world spectroscopic applications often involve imbalanced mixtures where the target compound represents only a small fraction of the total samples. To evaluate our approach's robustness under such conditions, we conducted comprehensive experiments on imbalanced mixtures and developed a unified multi-task model.

E.1 Performance on Imbalanced Binary Mixtures

We trained both vanilla and alignment-enhanced models on three imbalanced binary datasets: $\text{BINARY}_{40\%\text{-}60\%}$, $\text{BINARY}_{30\%\text{-}70\%}$, and $\text{BINARY}_{10\%\text{-}90\%}$. These datasets simulate realistic scenarios where target compounds appear with varying concentrations in spectroscopic measurements.

Results are shown in Table 8. Performance degrades when the target compound represents only 10% of the mixture, highlighting the challenge of detecting minority components in spectroscopic data. Fine-tuning consistently improves performance across all imbalance ratios, with the most substantial gains observed in the Top-5 and Top-10 metrics.

RECONSTRUCTION NETWORK	RECONSTRUCTION LOSS	λ	TOP–1 ↑	TOP–5 ↑	TOP–10 ↑
CNN	MSE	50	15.8	29.2	33.5
CNN	MSE	5	17.7	31.7	36.3
CNN	MSE	1	17.6	31.1	35.7
CNN	MAE	50	**19.7**	**34.4**	**39.2**
CNN	MAE	5	15.7	29.2	33.8
CNN	MAE	1	19.1	33.8	38.4
MLP	MSE	50	15.6	29.1	33.6
MLP	MSE	5	16.6	30.6	35.0
MLP	MSE	1	17.2	30.9	35.0
MLP	MAE	50	18.0	32.1	36.5
MLP	MAE	5	13.4	25.4	29.9
MLP	MAE	1	18.9	33.7	38.2

Table 7: Ablation study on RECONSTRUCTION NETWORK, RECONSTRUCTION LOSS and λ. The model were trained on the $\text{BINARY}_{50\%\text{-}50\%}$ datasets.

	DATASET	PERCENTAGE	TOP–1 ↑	TOP–5 ↑	TOP–10 ↑
Simulated	$\text{BINARY}_{40\%\text{-}60\%}$	40%	14.5	27.0	31.1
		60%	18.8	32.5	37.7
	$\text{BINARY}_{30\%\text{-}70\%}$	30%	13.2	24.9	28.8
		70%	20.9	37.3	41.0
	$\text{BINARY}_{10\%\text{-}90\%}$	10%	6.3	14.3	17.7
		90%	**25.6**	**43.6**	**48.6**
Fine-tuned	$\text{BINARY}_{40\%\text{-}60\%}$	40%	14.8 ± 1.3	42.7 ± 2.4	55.8 ± 2.2
		60%	21.1 ± 0.4	52.5 ± 1.3	64.0 ± 1.9
	$\text{BINARY}_{30\%\text{-}70\%}$	30%	13.1 ± 1.2	39.5 ± 3.0	53.2 ± 3.0
		70%	24.0 ± 1.2	56.2 ± 2.7	67.9 ± 2.9
	$\text{BINARY}_{10\%\text{-}90\%}$	10%	8.2 ± 1.1	27.4 ± 2.3	40.2 ± 2.8
		90%	$\mathbf{29.1 \pm 1.4}$	$\mathbf{62.1 \pm 2.6}$	$\mathbf{72.3 \pm 2.3}$

Table 8: Performance comparison of vanilla and fine-tuned models on imbalanced binary datasets. Each test set contains 5000 samples for each percentage class. Results demonstrate the models' sensitivity to class imbalance, with performance declining as target prevalence decreases.

E.2 Impact of Alignment on Imbalanced Data

We next evaluated whether our alignment strategy could mitigate the performance degradation observed with imbalanced data. Using a convolutional neural network as the reconstruction network with MAE loss and $\lambda = 50$, we applied the alignment technique to all imbalanced datasets.

Table 9 demonstrates that alignment consistently improves performance across most imbalance scenarios. The enhancement is particularly pronounced for minority class detection, where the alignment mechanism helps the model focus on relevant spectral features despite class imbalance. While improvements are modest for the 90% case (majority class), the alignment strategy maintains performance without degradation.

	Dataset	Percentage	Top–1 ↑	Top–5 ↑	Top–10 ↑
Alignment	$\text{BINARY}_{40\%\text{-}60\%}$	40%	17.7	32.6	37.8
		60%	20.8	37.2	43.3
	$\text{BINARY}_{30\%\text{-}70\%}$	30%	16.1	29.8	34.7
		70%	24.0	40.2	45.6
	$\text{BINARY}_{10\%\text{-}90\%}$	10%	8.3	18.6	22.9
		90%	**26.7**	**44.1**	**49.5**
Fine-Tuned + Alignemnt	$\text{BINARY}_{40\%\text{-}60\%}$	40%	16.5 ± 0.1	42.4 ± 1.3	56.2 ± 1.4
		60%	23.0 ± 0.5	51.8 ± 1.1	64.7 ± 1.3
	$\text{BINARY}_{30\%\text{-}70\%}$	30%	15.1 ± 1.3	40.6 ± 2.5	53.8 ± 2.3
		70%	25.2 ± 0.6	54.5 ± 0.4	67.3 ± 0.7
	$\text{BINARY}_{10\%\text{-}90\%}$	10%	9.7 ± 0.7	29.5 ± 2.8	43.1 ± 2.8
		90%	$\mathbf{29.5 \pm 1.5}$	$\mathbf{61.1 \pm 1.7}$	$\mathbf{73.0 \pm 1.5}$

Table 9: Performance of alignment-enhanced models on imbalanced binary datasets. The alignment strategy provides consistent improvements across most imbalance ratios, with the most significant gains for lower concentrations.

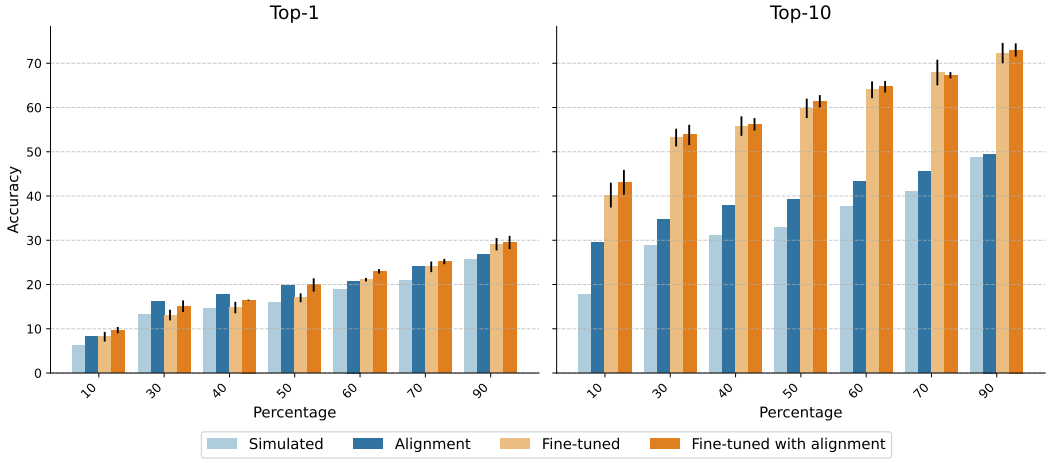


Figure 6: Bar plot comparing the performance of the Baseline and Aligned models both pretrained on Simulated spectra and fine-tuned on real spectra. The plot summarizes Table 8 and 9. Aligning the encoder representations consistently improves performance for both pretraining and fine-tuning. Furthermore, the model performance—unsurprisingly—positively correlate with the concentration of the target compound present in the mixture.

F Multitask Model

To develop a model capable of predicting components from spectra of mixtures with widely varying concentrations, we trained a multitask model using data spanning all concentration levels, as described in subsection E.2. The pretraining performance of this model, which was not included in the main paper, is reported in Table 10.

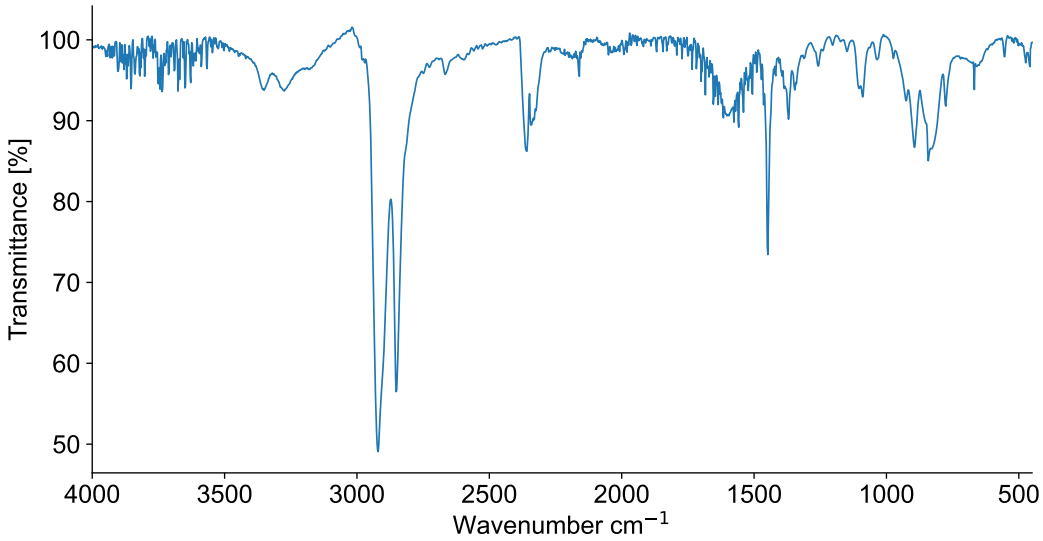
	Percentage	Top–1 ↑	Top–5 ↑	Top–10 ↑
Multi-task	10%	3.4	9.4	11.7
	20%	5.3	11.6	14.9
	30%	7.2	14.6	17.4
	40%	7.3	16.6	20.2
	50%	8.0	17.7	20.9
	60%	9.9	18.9	23.3
	70%	12.3	23.3	28.6
	80%	14.1	26.6	31.9
	90%	16.4	29.9	33.9

Table 10: Performance of the multi-task model trained simultaneously on multiple imbalance ratios. The test set contains 2500 samples at 50% and 1250 samples for each other concentrations. Multi-task learning enables the model to generalize across different imbalance scenarios while maintaining robust performance.

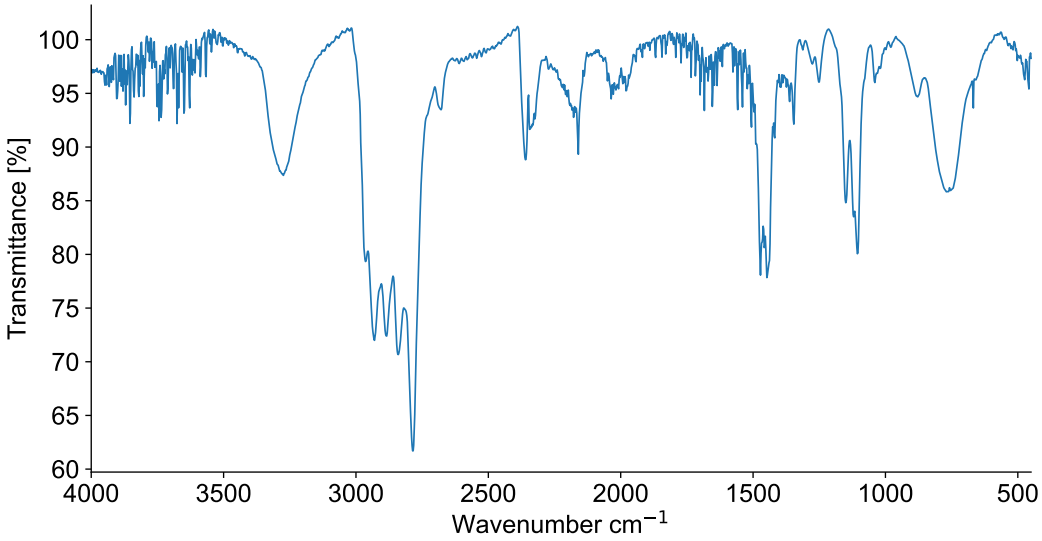
G Experimental IR spectra

G.1 Spectra of pure compounds

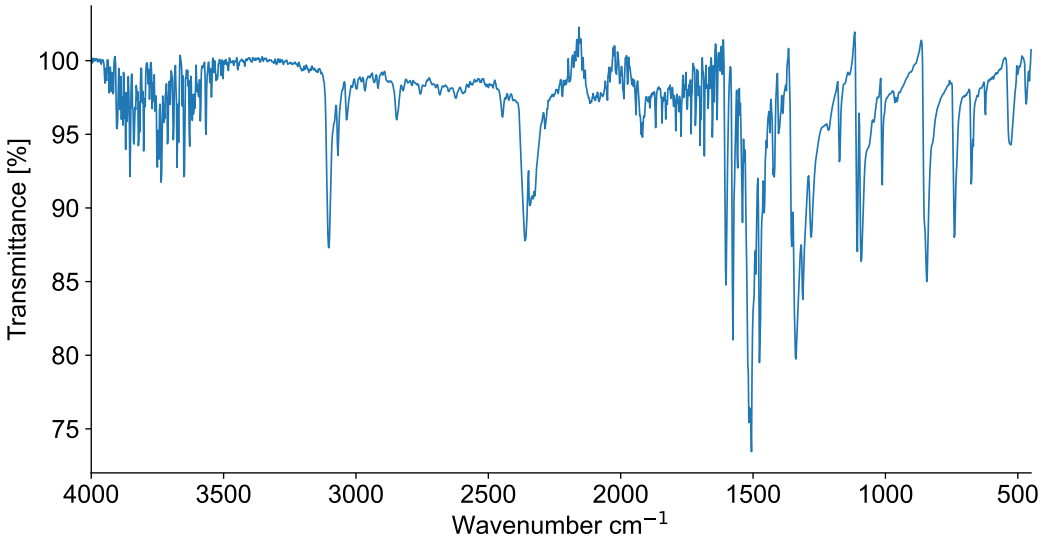
Cyclohexylamine



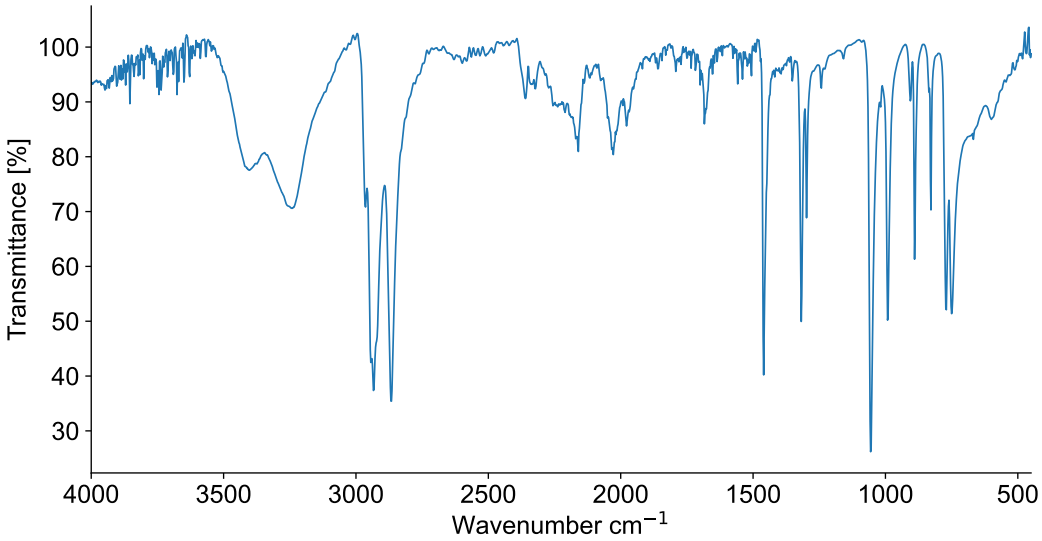
N,N-Dimethylethylenediamine



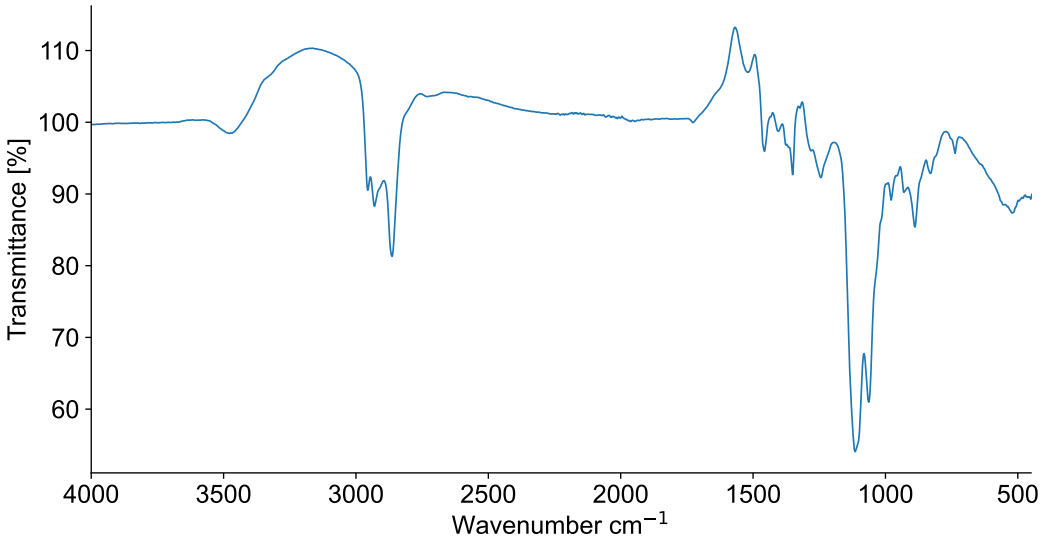
4-Chloronitrobenzene



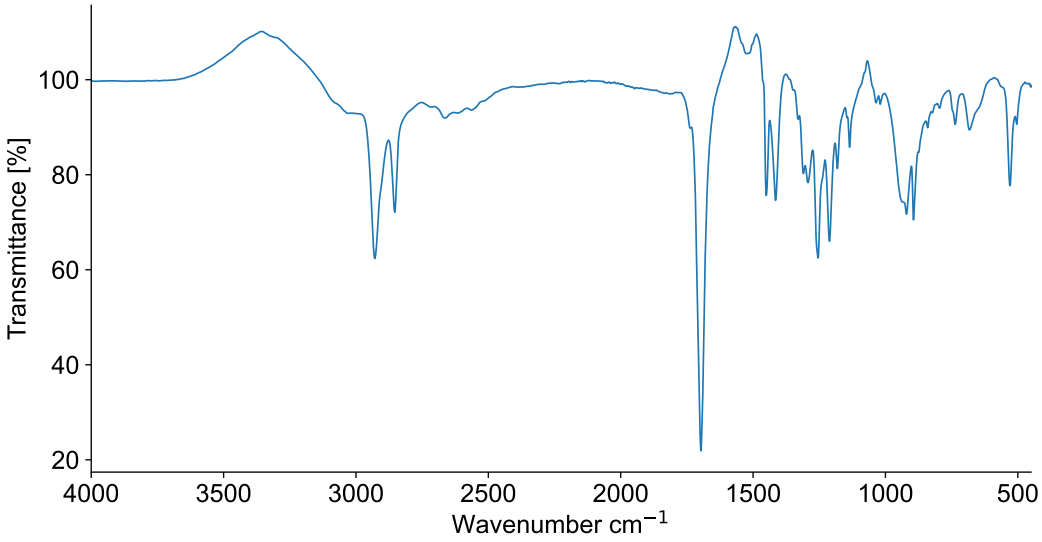
1,4-diazabicyclo[2.2.2]octane



Cyclohexancarboxylic acid

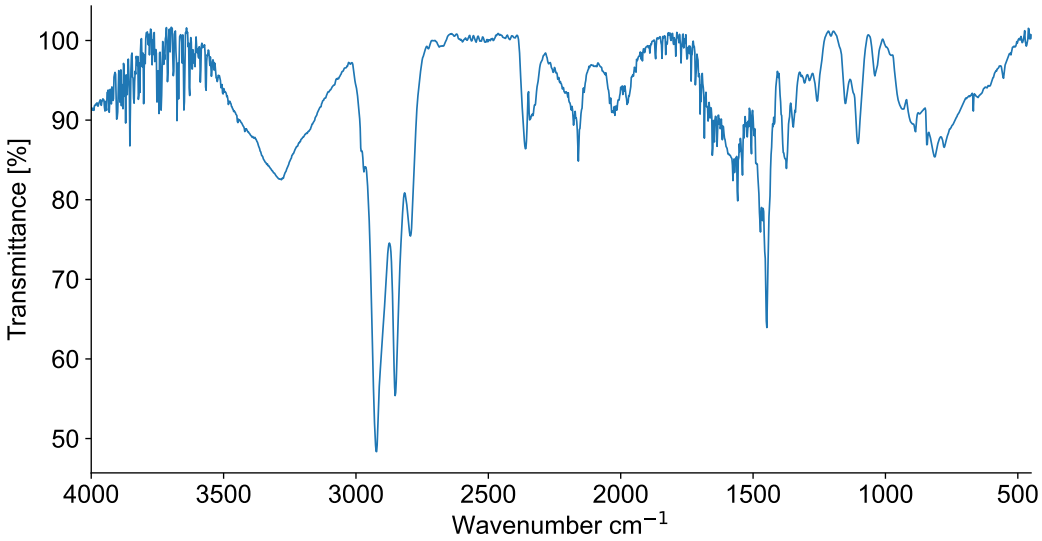


Diethylene glycol monobutyl ether

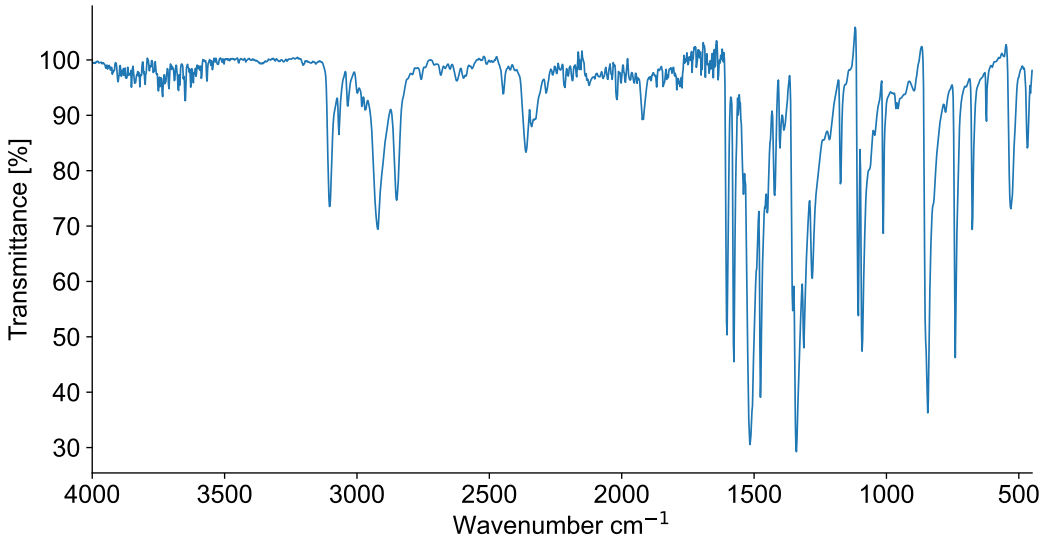


G.2 Spectra of mixtures used for validation

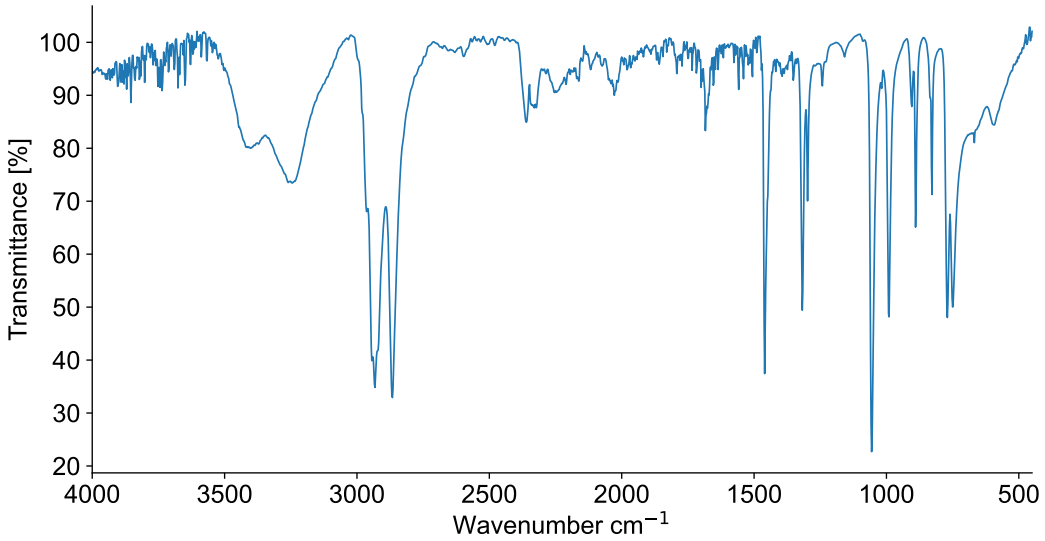
Cyclohexylamine + N,N-Dimethylethylenediamine



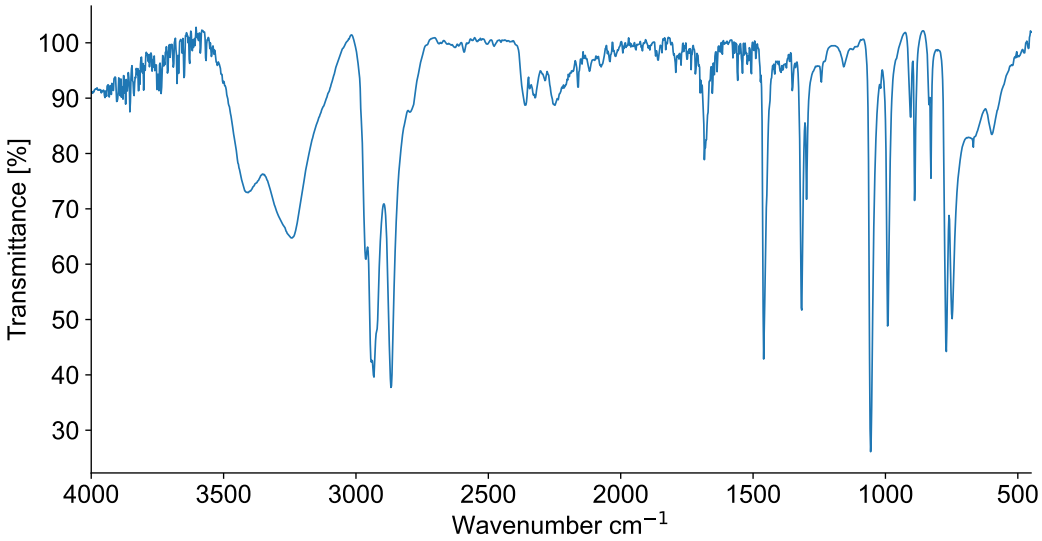
Cyclohexylamine + 4-Chloronitrobenzene



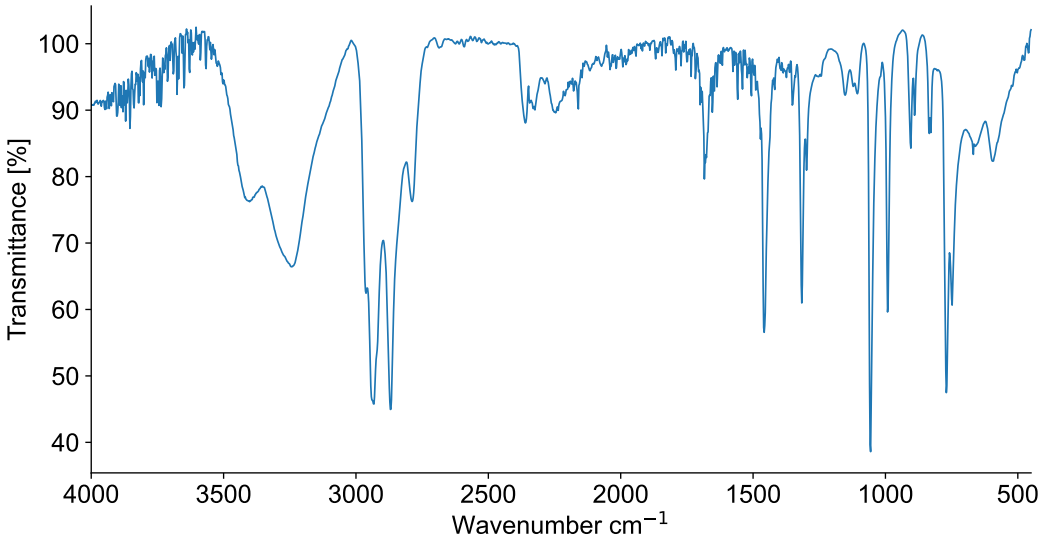
Cyclohexylamine + 4-Chloronitrobenzene



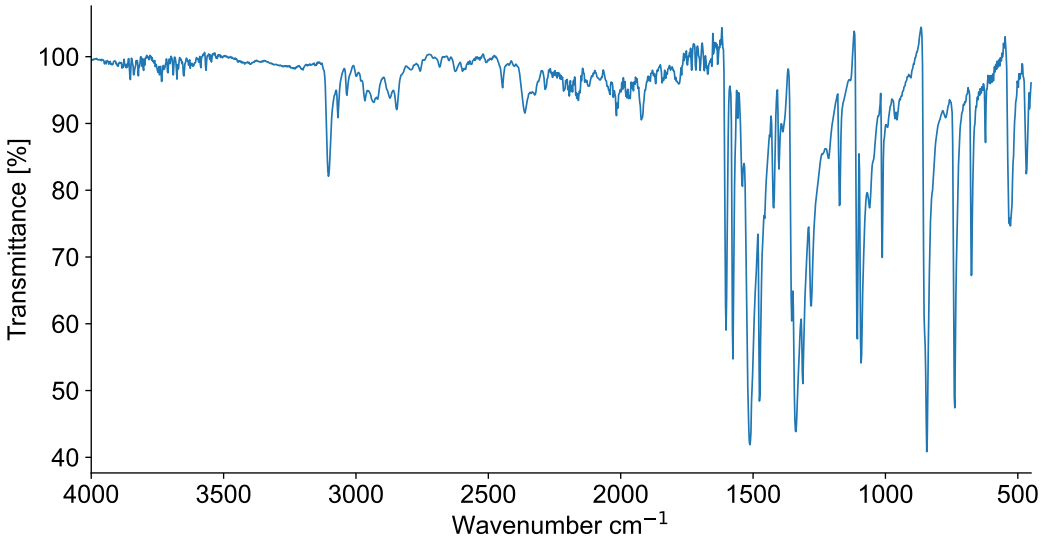
N,N-Dimethylethylenediamine + 4-Chloronitrobenzene



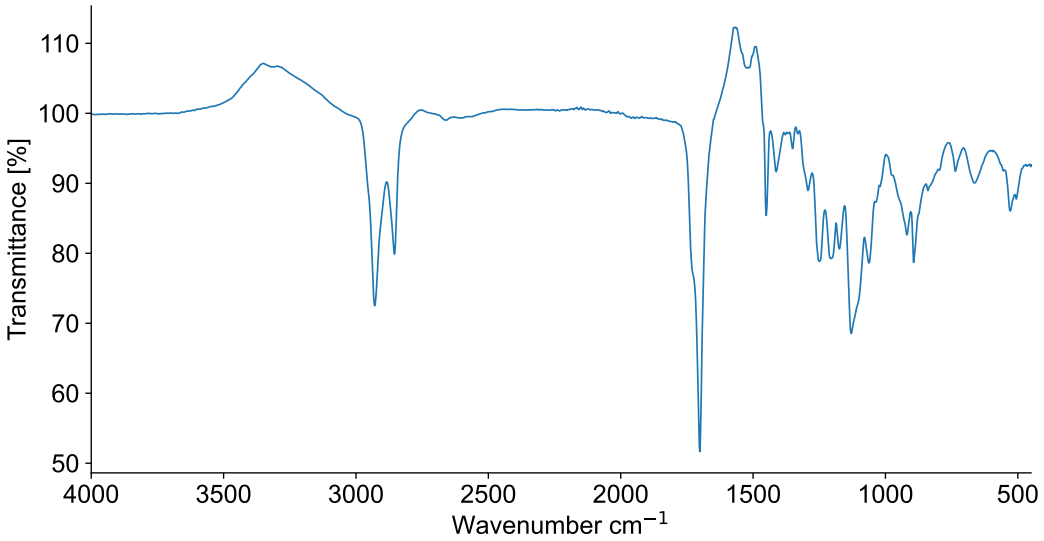
N,N-Dimethylethylenediamine + 4-1,4-diazabicyclo[2.2.2]octane



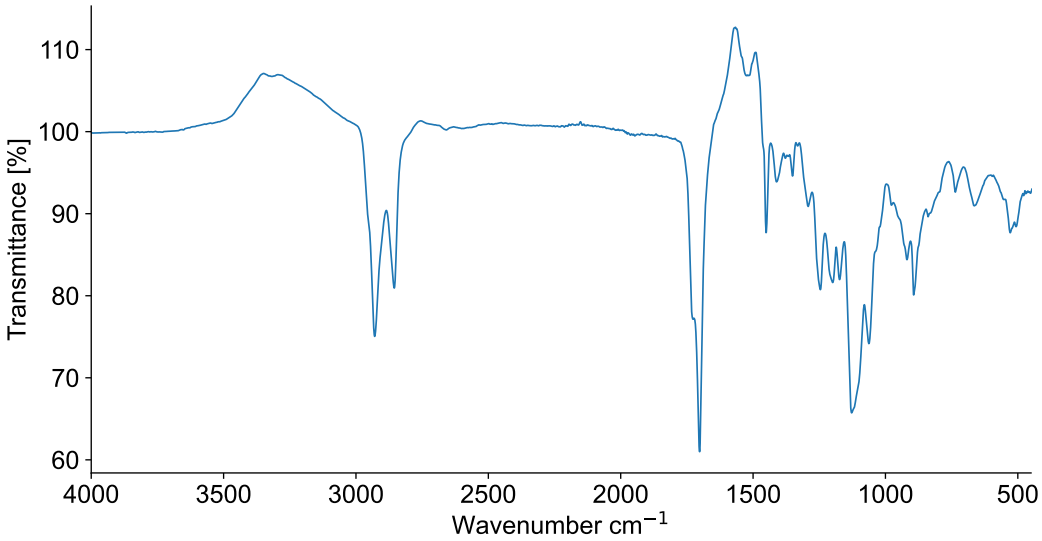
4-Chloronitrobenzene + 1,4-diazabicyclo[2.2.2]octane



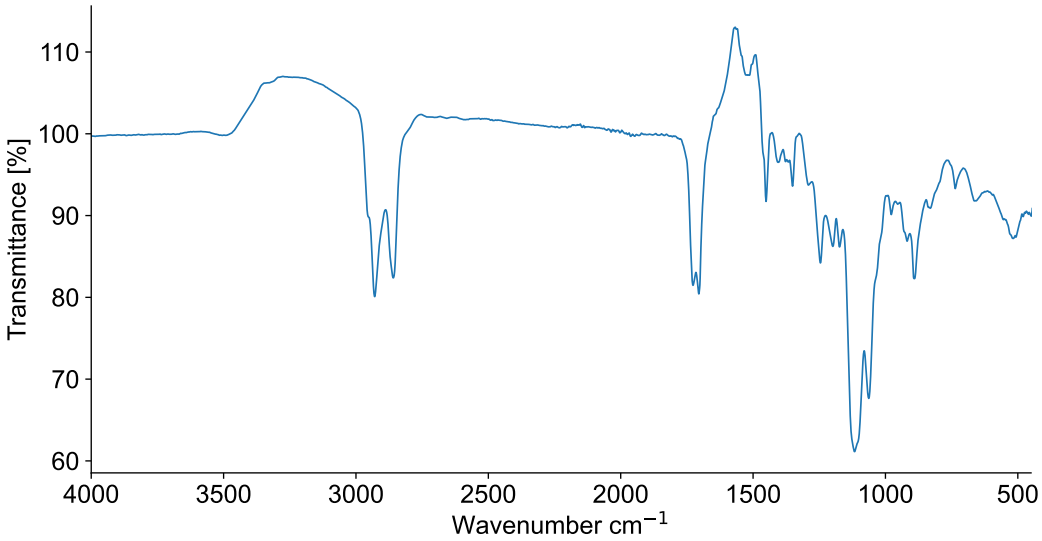
1/3 Cyclohexancarboxylic acid + 2/3 Diethylene glycol monobutyl ether



1/2 Cyclohexancarboxylic acid + 1/2 Diethylene glycol monobutyl ether

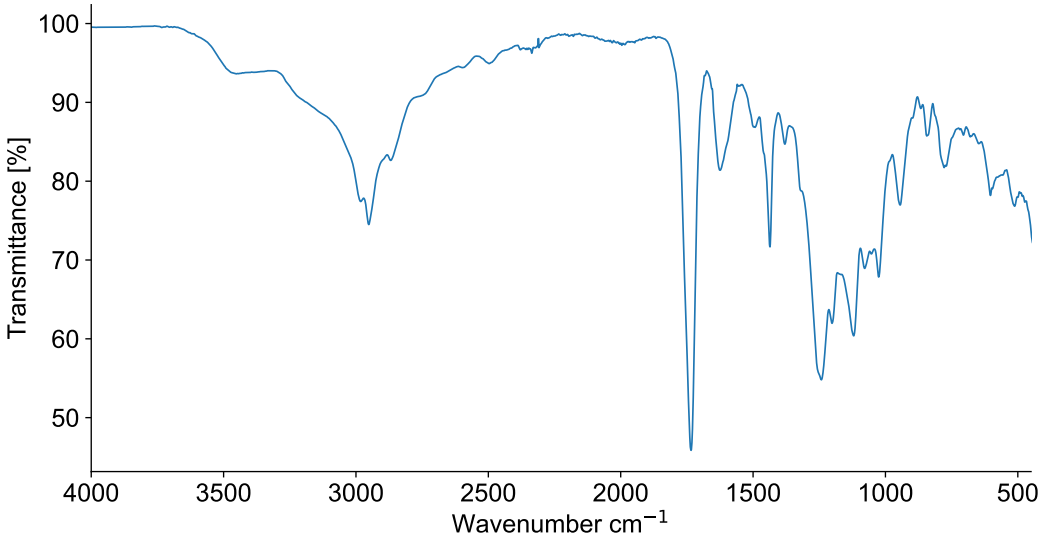


2/3 Cyclohexancarboxylic acid + 1/3 Diethylene glycol monobutyl ether

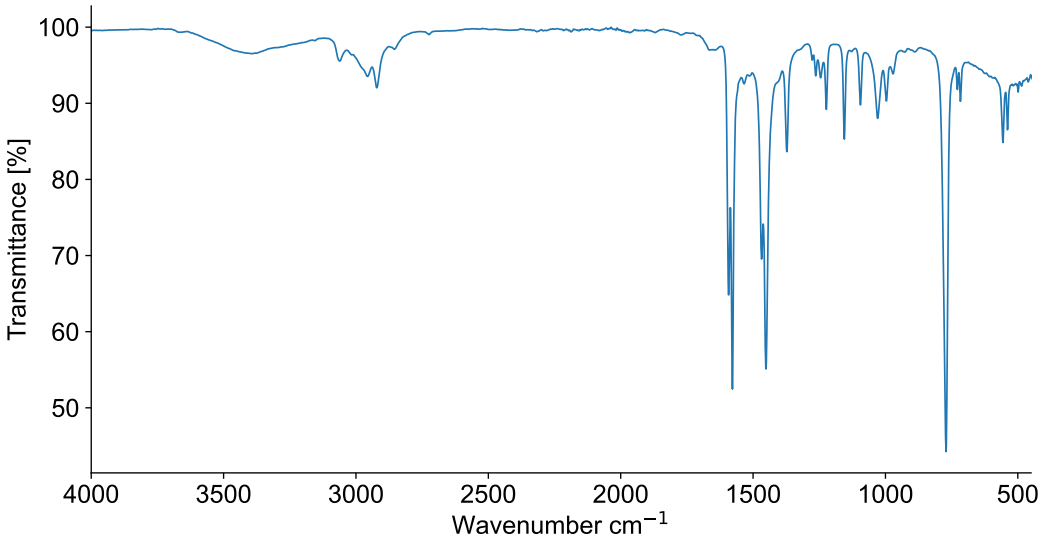


G.3 Spectra of mixtures used for evaluation

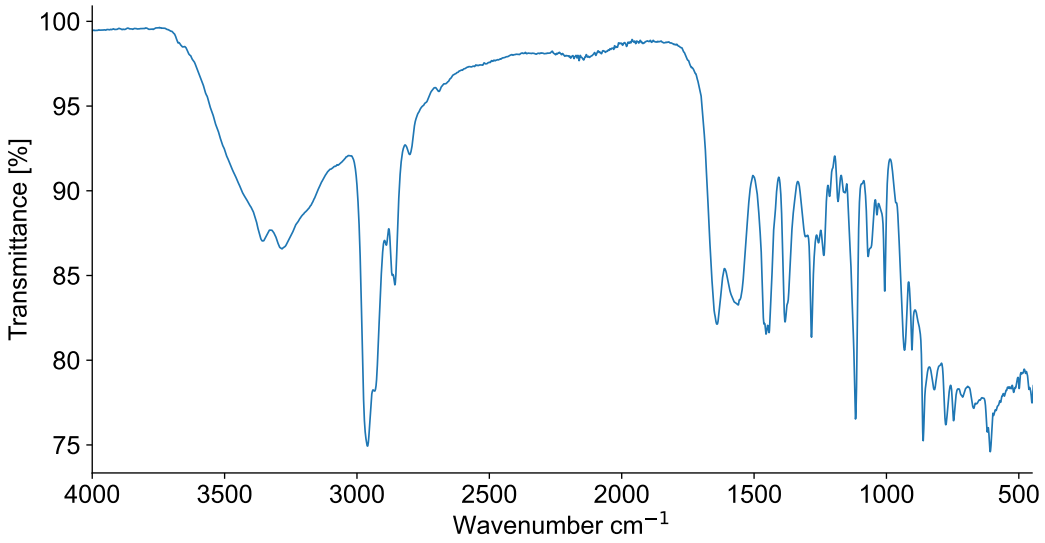
Cyclobutylamine + Methyl-3-Bromopyruvate



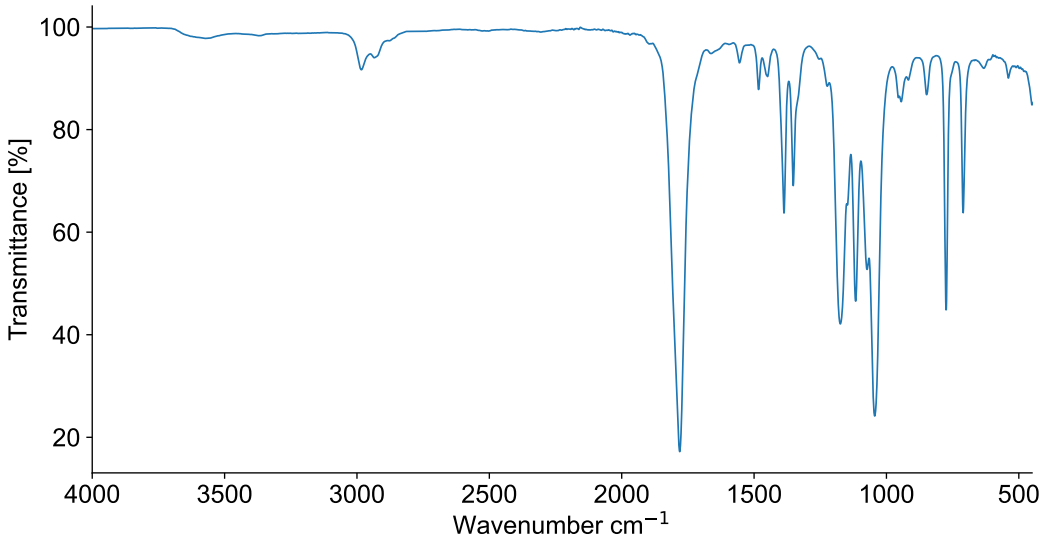
Cyclobutylamine + 2,6-Lutidine



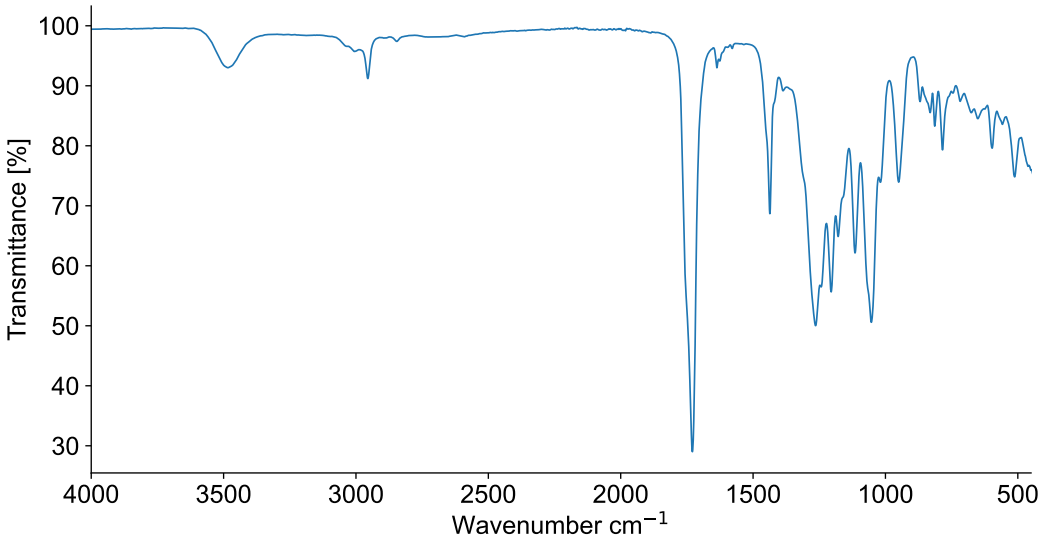
Cyclobutylamine + N-Methylmorpholine



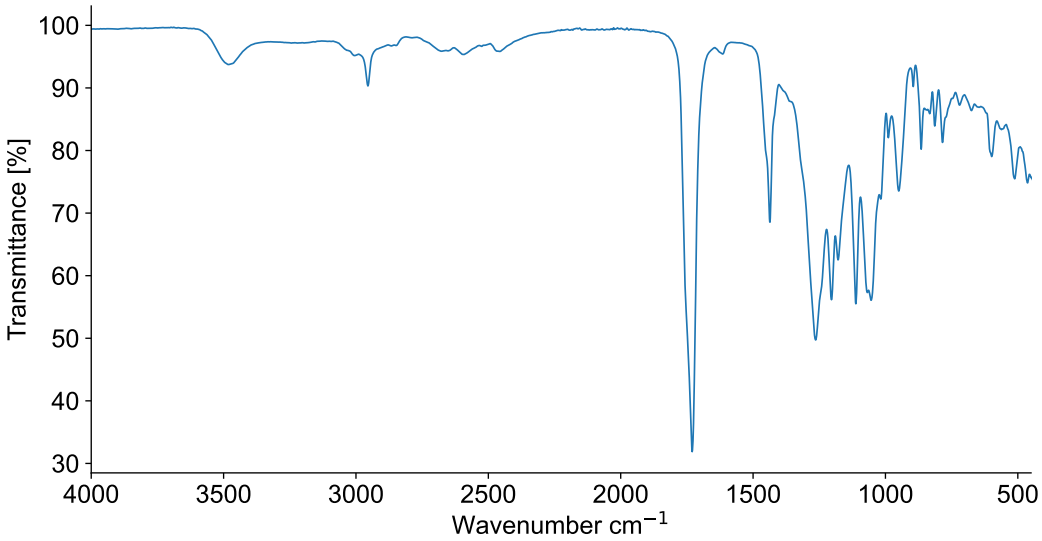
Cyclobutylamine + Propylene-carbonate



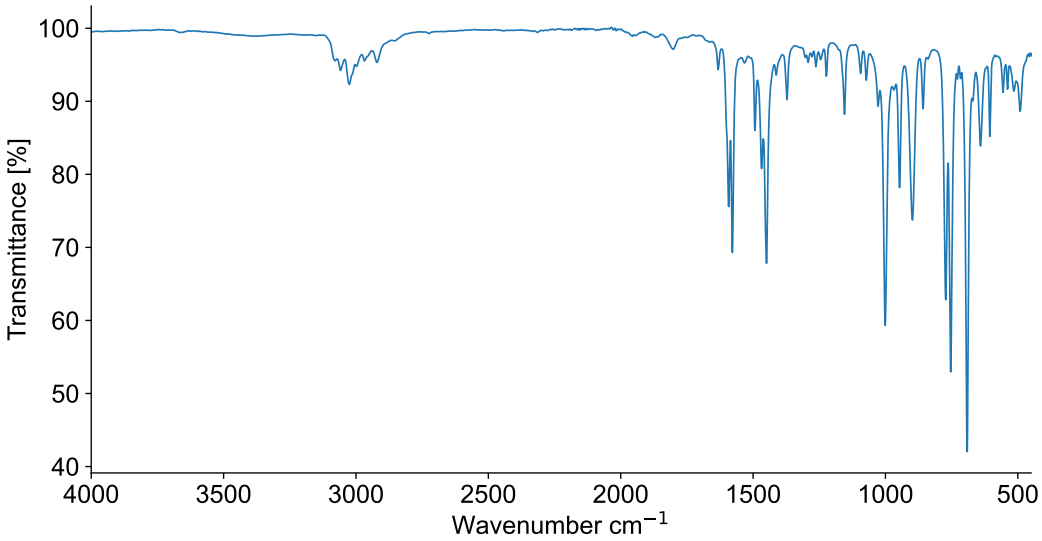
Methyl-3-Bromopyruvate + 2,6-Lutidine



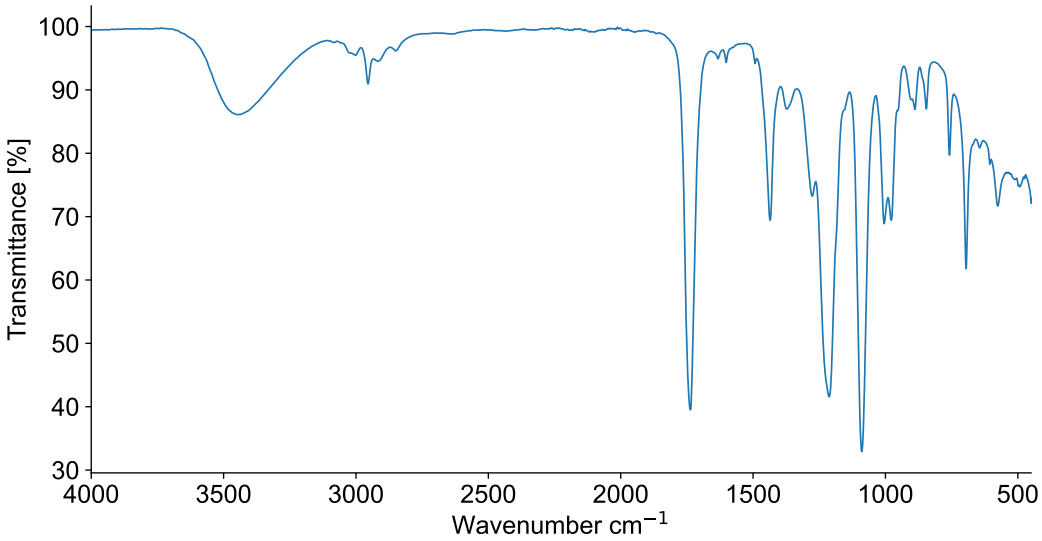
Methyl-3-Bromopyruvate + N-Methylmorpholine



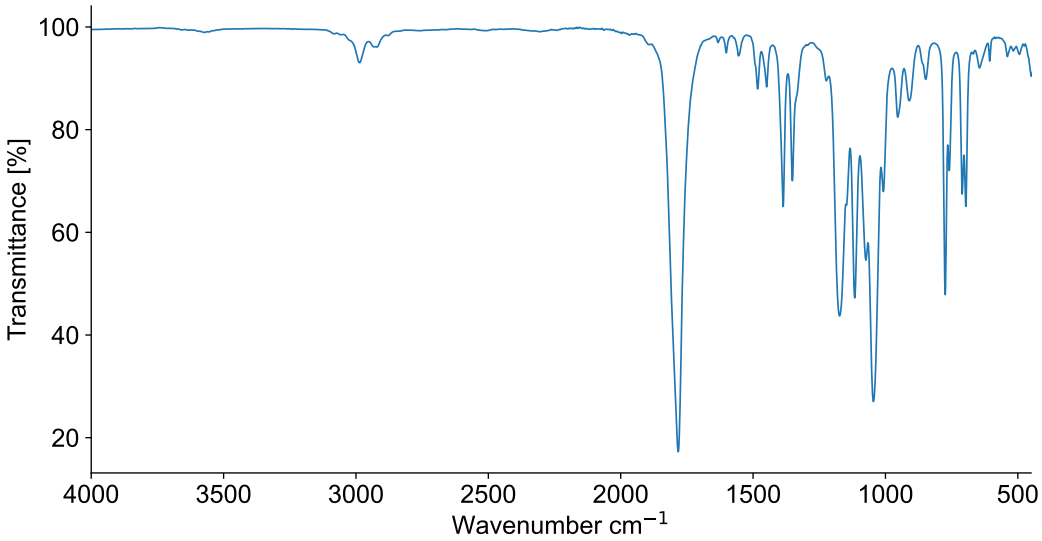
Trans-1-Phenyl-1,3-butadiene + 2,6-Lutidine



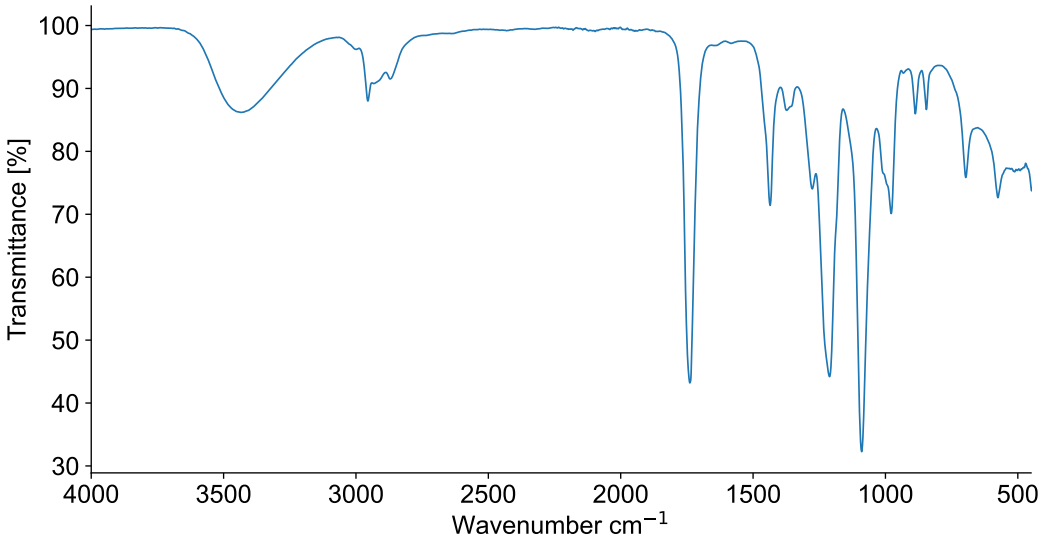
Trans-1-Phenyl-1,3-butadiene + Methyl-glycolate



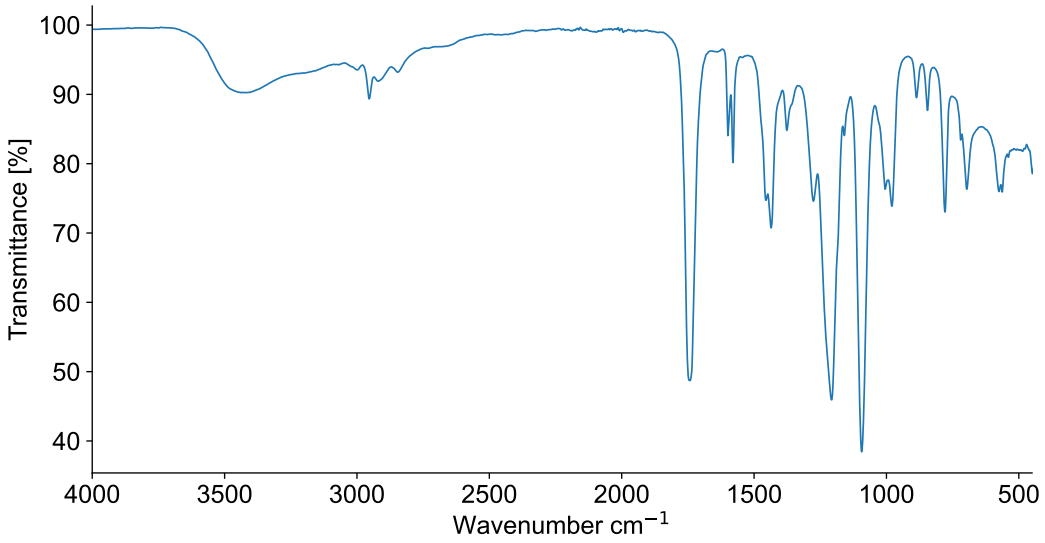
Trans-1-Phenyl-1,3-butadiene + Propylene-carbonate



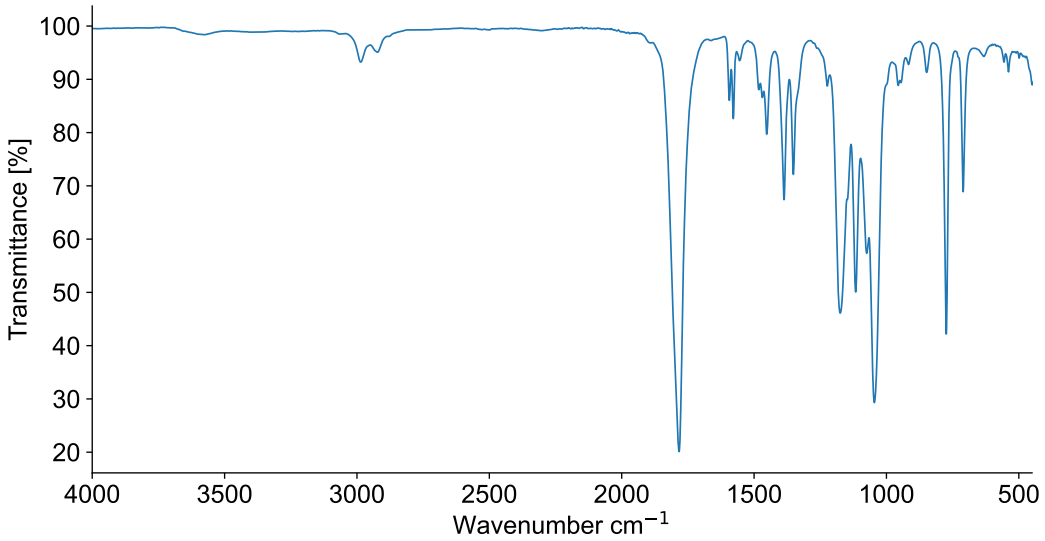
Diethylene-glycol-butyl-ether + Methyl-glycolate



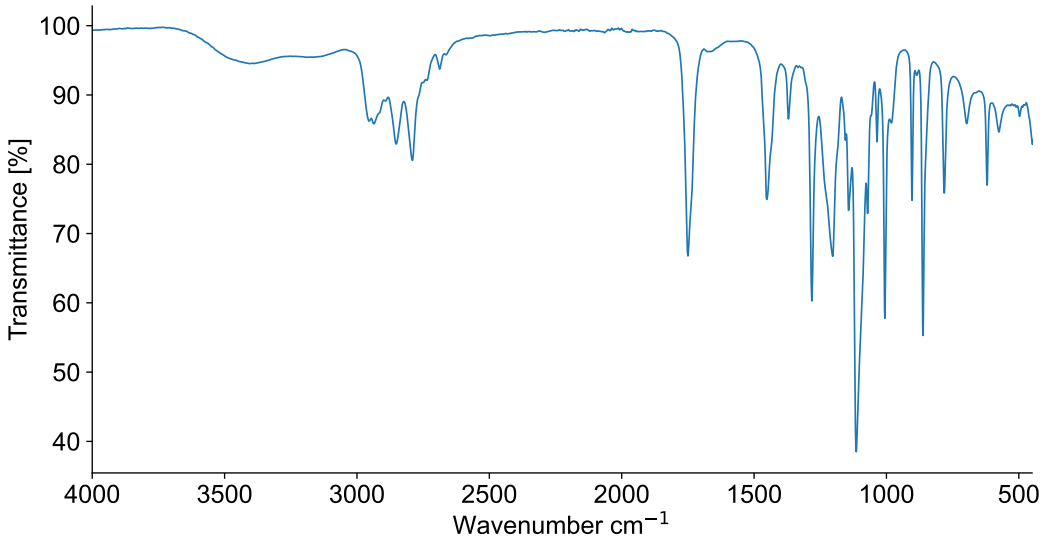
2,6-Lutidine + Methyl-glycolate



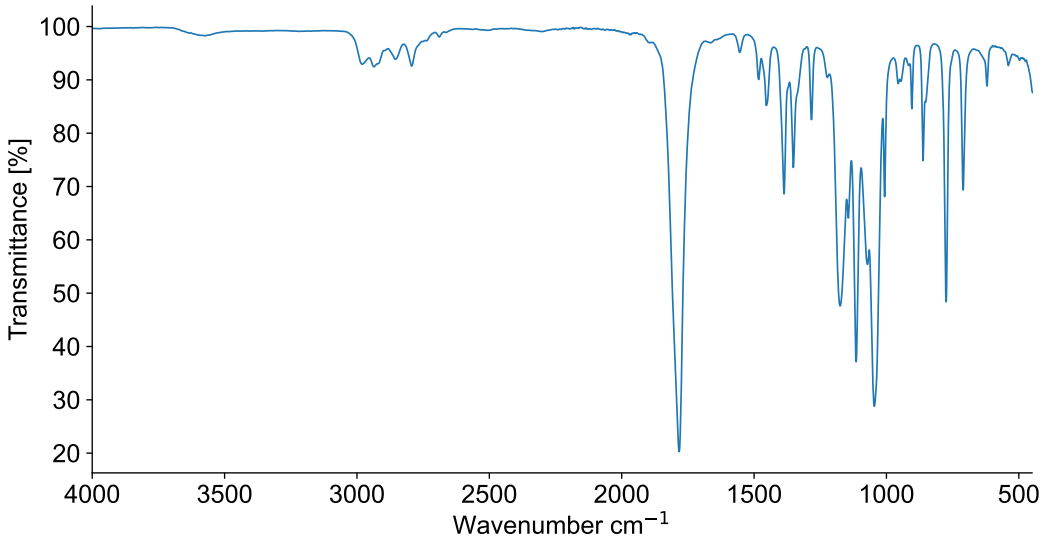
2,6-Lutidine + Propylene-carbonate



N-Methylmorpholine + Methyl-glycolate



N-Methylmorpholine + Propylene-carbonate



Methyl-glycolate + Propylene-carbonate