Supporting information for manuscript

Infrared spectroscopy-assisted predicting of impurities in chemicals using machine learning: towards smart self-driving laboratories

by the authors Anastasiia M. Kutskaia and Konstantin S. Rodygin[*]

Saint Petersburg State University, Universitetskaya nab. 7/9, St. Petersburg 199034, Russia.

*Corresponding author. E-mail: k.rodygin@spbu.ru

Contents

S1 Visualization of dataset

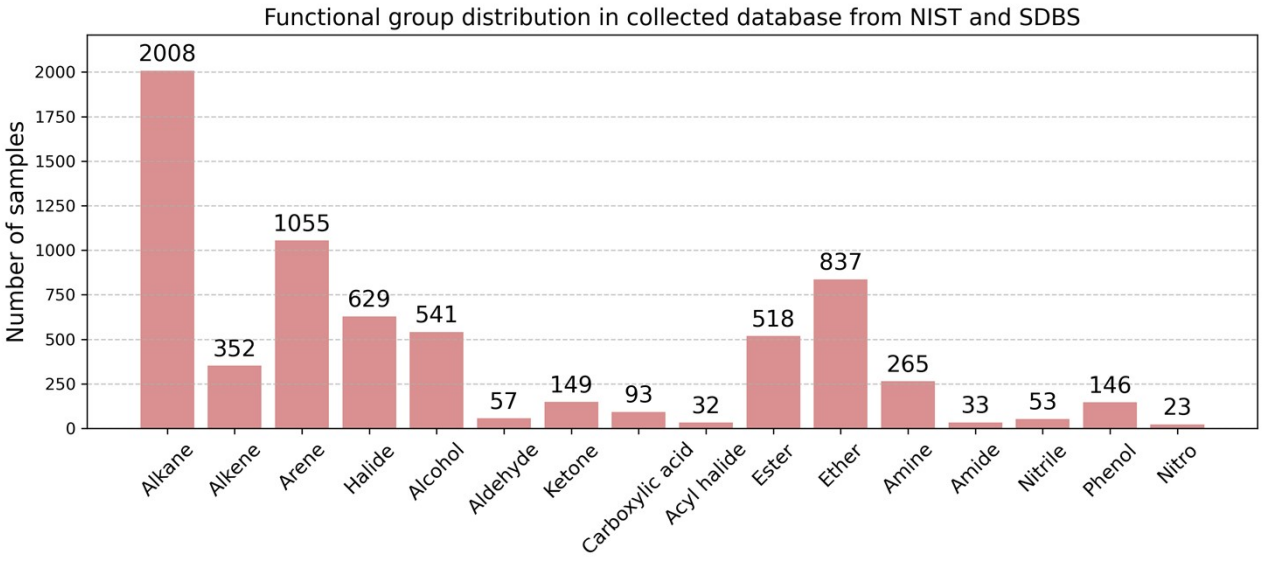

Figure S1. Distribution of functional groups in collected dataset from NIST and SDBS databases.

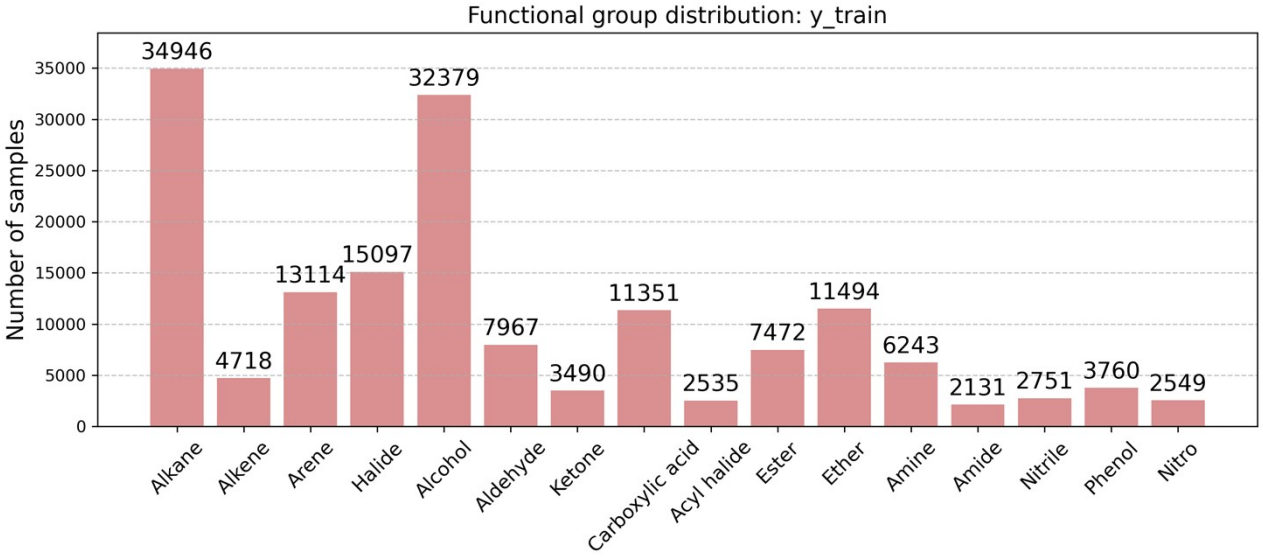

Figure S2. Distribution of functional groups in train set.

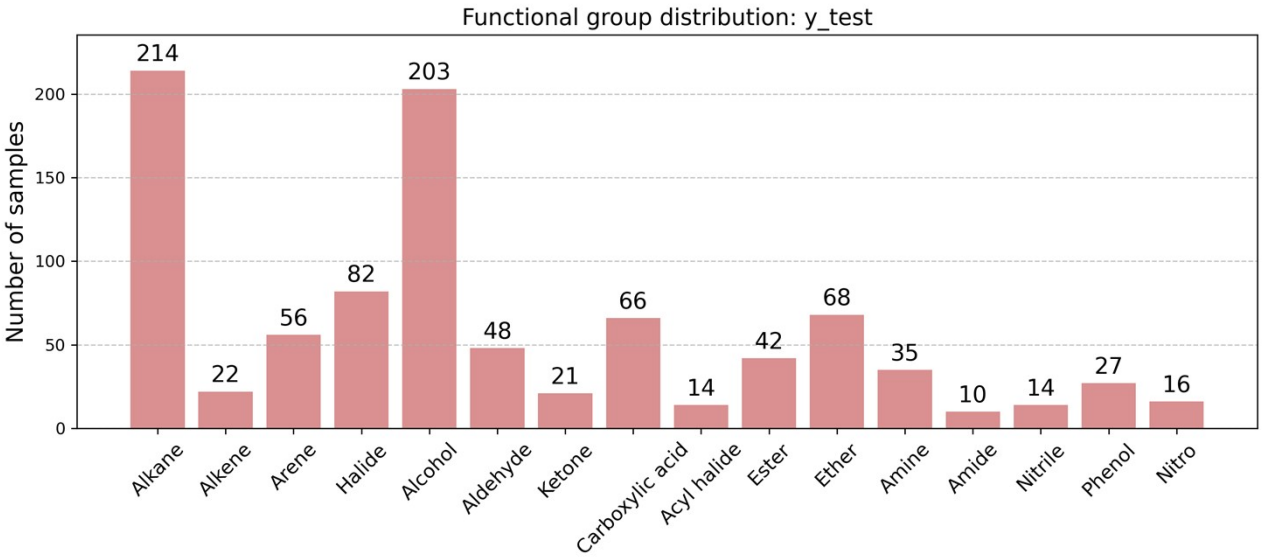

Figure S3. Distribution of functional groups in test set.

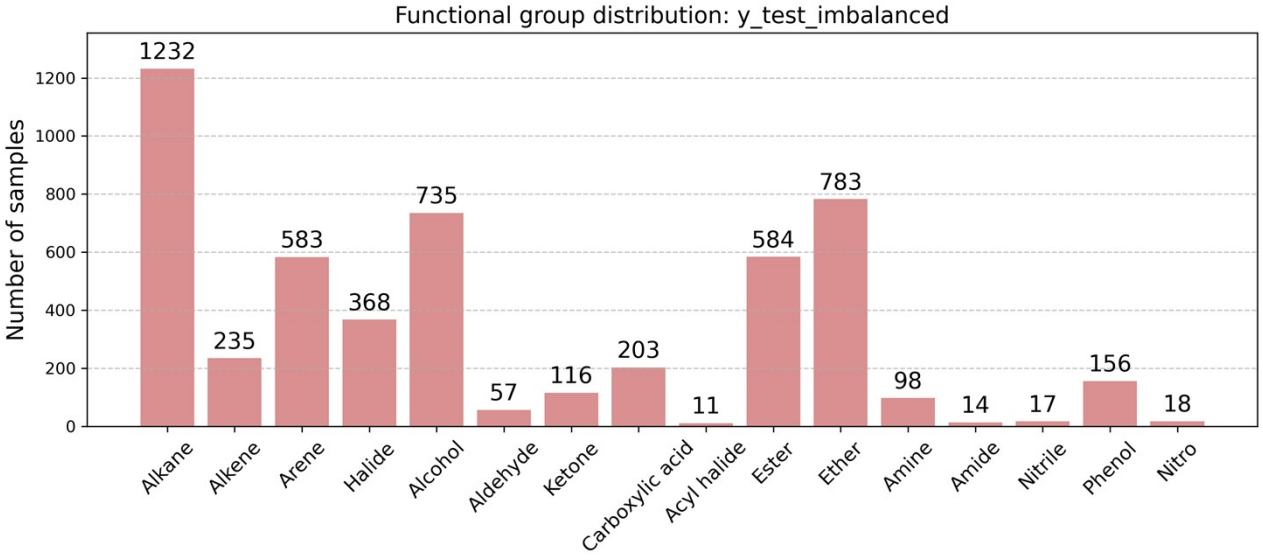

Figure S4. Distribution of functional groups in imbalanced test set.

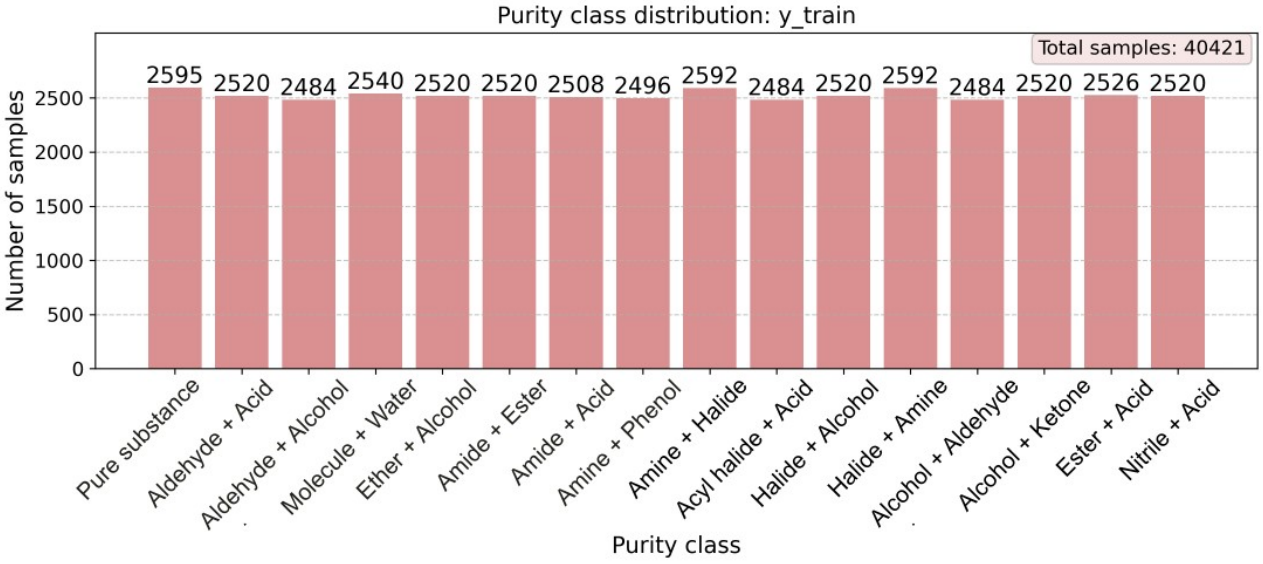

Figure S5. Distribution of purity classes in train set.

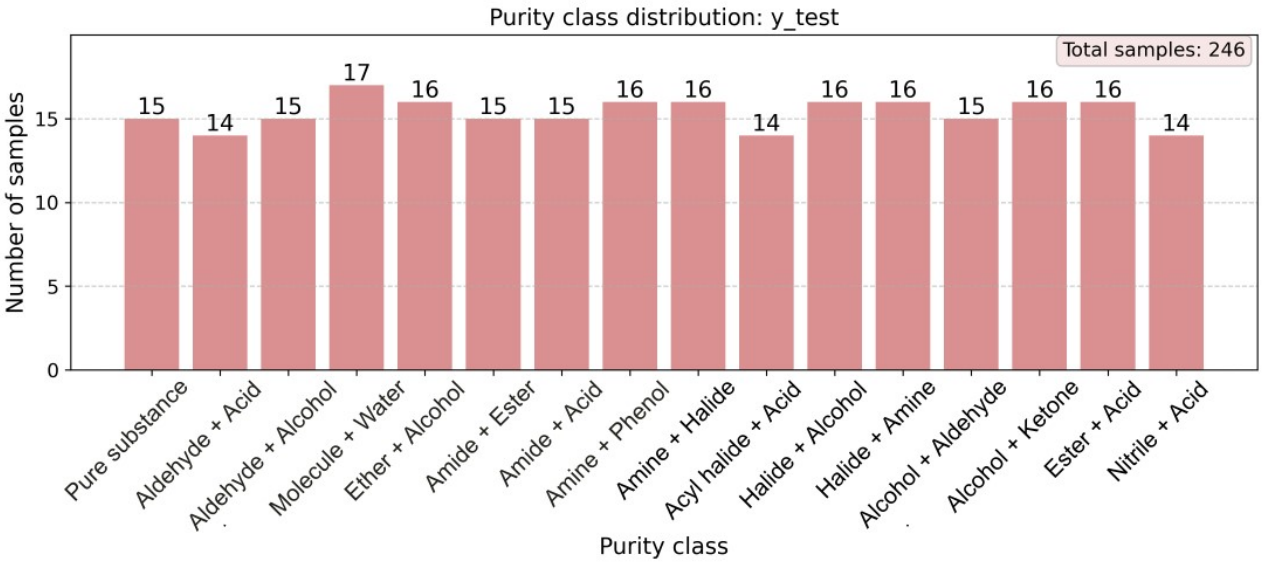

Figure S6. Distribution of purity classes in test set.

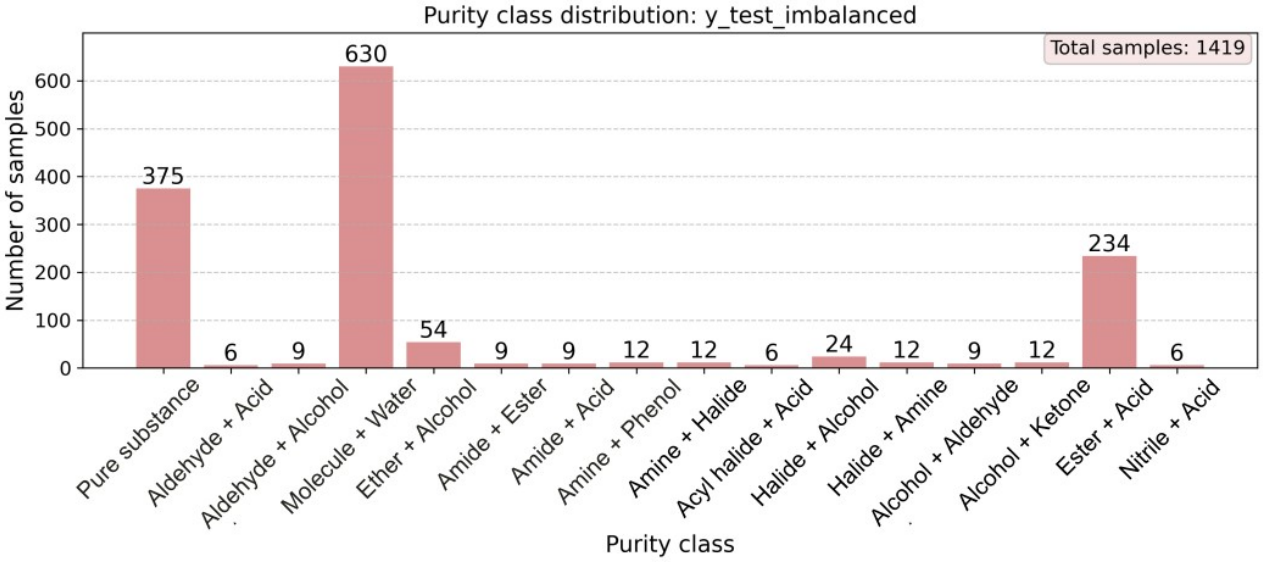

Figure S7. Distribution of purity classes in imbalanced test set.

S2 Data preprocessing and augmentation

The spectral data and molecular information infrared spectrum-compound pairs were collected from the National Institute of Standards and Technology (NIST) database and the Spectra Database of Organic Compounds (SDBS) database of the National Institute of Advanced Industrial Science and Technology (AIST). The spectrum in png or jcamp format and the corresponding InChI string in text format were extracted using the unique compound identifier.

The spectral data were obtained using CCl_4 as a solvent. The collected set contained 2440 spectra. Each spectrum corresponded to a molecule encoded as a one-hot vector of 16 organic functional groups and a purity class.

Data preprocessing was introduced according to the methodology described by Cole and co-autors.[1] Under data preprocessing step, all the spectra were converted to absorption spectra in order to obtain a IR spectrum of mixtures using Beer mixing law. For the same reason, all the spectra were baseline-corrected and normalized: spectra from different databases may significantly vary. The data preparation provided prediction the presence of contaminants in concentrations more than 5%. Each IR spectrum consisted of a series of 600 points over a range of 400 cm^{-1} to 4000 cm^{-1} with a resolution of ~6 cm^{-1}.

A classification of mixtures into 16 classes, including pure substances, was carried out, based on possible impurities. Additionally, we focused on the number of substance-impurity pairs in the collected database: cases with at least 7 examples per group were selected. The algorithm for searching for a substance-impurity pairs was implemented using Reaction SMARTS templates: the function accepted all the examples that corresponded to the condition of the presence of a characteristic label in the main substance (*e.g.*, all the alcohols) as input, and the corresponding contaminant was determined based on the specified transformations. If the search of the library resulted in a complete match of the InChI string of the impurity and the substance in the database, the pair was added to the database. The corresponding vector of functional group presence for mixtures was the element-wise OR of the functional group vectors of the initial compounds.

Before the spectra were generated, the data in the form of substance-impurity pairs were divided in a ratio of ~ 72:13:15 for training, validation and test sets. The indexes of the pairs main compound/impurity were divided as follows: augmented data were not presented in the validation and test sets for a fairer evaluation of the model. It was an indicator, the model was not overtrained on the features of the training set, but rather generalized on data on characteristic vibrational frequencies.

According to Beer's law, spectra of a substance with a corresponding impurity were synthetically generated. For a binary mixture, it can be written as follows:

$$\begin{bmatrix} A_1(\tilde{\upsilon}_1) & A_1(\tilde{\upsilon}_2) & \dots & A_1(\tilde{\upsilon}_l) \\ A_2(\tilde{\upsilon}_1) & A_2(\tilde{\upsilon}_2) & \dots & A_2(\tilde{\upsilon}_l) \\ \vdots & \vdots & \ddots & \vdots \\ A_m(\tilde{\upsilon}_1) & A_m(\tilde{\upsilon}_2) & \dots & A_m(\tilde{\upsilon}_l) \end{bmatrix} = \begin{bmatrix} \varphi_{11} & \varphi_{12} \\ \varphi_{21} & \varphi_{22} \\ \vdots & \vdots \\ \varphi_{m1} & \varphi_{m2} \end{bmatrix} \begin{bmatrix} K_1(\tilde{\upsilon}_1) & K_1(\tilde{\upsilon}_2) & \dots & K_1(\tilde{\upsilon}_l) \\ K_2(\tilde{\upsilon}_1) & K_2(\tilde{\upsilon}_2) & \dots & K_2(\tilde{\upsilon}_l) \end{bmatrix}$$

where m is number of samples (m is individual for each class), l is wavenumber points (600 in this work); $K_i(\tilde{\upsilon})$– intensities of a pure compound. The coefficients for each linear combination were assigned using uniform distribution.

The concentration of a main substance (φ_{m1}) is in the range of 75-95% due to Beer's law limitations. The simple and efficient approach resulted in good correlation of generated and real spectra. Quantitative analysis or strict modeling of mixture spectra was not the aim of the study. Qualitative addition of an impurity was enough to observe it in IR-spectrum. Recording IR-spectra in CCl_4 met the requirements of handling with dilute solutions, that was appropriate for Beer's approximation. In a case of very dilute solutions, a medium insignificantly impact on the shift in vibration frequencies due to the environment of a molecule is constant and there is no impact of the anisotropic polarizability of a solvent.

Horizontal shifting and noise incorporation were applied on training data to make a sample balanced for the purity prediction task. It was carried out due to limitation of examples for some classes of purity ("Nitrile is contaminated with carboxylic acid" – 7 examples, "Aldehyde is contaminated with carboxylic acid" – 11 examples, "Aldehyde is contaminated with alcohol" – 14 examples). Noise level, number of linear combinations for a single pair, and the composition of horizontally shifted spectra were selected individually in order to obtain ≈2500 examples per class. The set of augmentations provided prediction of the class even with a very limited set of examples.

Artificial noise was incorporated according to the methodology described in Modestino and co-autors work[2]: each class was assigned individual noise coefficient ranging from 0 (no noise) to 1 (maximum noise level). The coefficient was multiplied by the maximum intensity deviation value of ± 0.02 a.u., and the resulting noise value was added to the signal intensity.

The test set is highly imbalanced due to the test data set was not augmented (Section S1 Fig. S7). According to fundamental principle of ML model evaluation, the distribution in test set and in data for production must be the same. Considering the true distribution of reagent contamination cases is unknown and the existing class distribution was not representative in this work, it was decided to balance the dataset. Balancing the test set corrected the impact of arbitrary class frequencies incorporated after data collection and increased the impact of false predictions for classes with limited examples (the prediction results for the imbalanced test set were given in Section S8).

S3 Model architecture

The 1D-CNN encoder processes the 600-dimensional one-chanel input through successive convolutional and fully-connected layers. Three 1D convolutional blocks (with channel depths 10, 20, 40) are applied sequentially. Each block consists of a 1D convolution, batch normalization, ReLU activation, and 1D max-pooling. These layers extract local patterns from a signal at progressively higher levels of abstraction. Output of final convolutional block is flattened and passed through three dense layers with sizes of 1015, 733, and 529. Dropout is applied between these layers to regularize the model, where p = 0.388079443185074. Each of the first two fully-connected layers uses ReLU activation. This 1D-CNN encoder serves as one expert in the MMoE framework, where 529-dimensional representation from final layer is used as the expert output.

The core MMoE module contains four shared experts, each of which learns different feature extractors for the input. These networks are not weights-shared but share the overall architecture, so that each can specialize in different aspects of the data. For each task, the model includes a gating network implemented as a linear layer, with a softmax activation. The gating network uses the same 529-dimensional input and computes a 4-dimensional output of weights (sum of weights is 1) – one weight per expert. These gate values are multiplied with the corresponding expert outputs and summed to produce a single 529-dimensional fused representation for that task.

Thus, for task *t*, the model computes

$$g^{(t)} = softmax\left(W^{(t)}x + b^{(t)}\right);\ h^{(t)} = \sum_{i=1}^{4} g_i^{(t)} e_i\,,$$

where x is the input and e_i are the expert outputs. The gating networks thus dynamically re-weight the shared expert outputs per task. This multi-gate structure explicitly learns to model task relationships from data by allowing each task to attend to different experts.

Each task has its own decoder that project the 529-dimensional fused representation h^t to a 16-dimensional output. Then, the first task (prediction of set of functional groups, multi-label classification) produces 16 independent scores passed through a sigmoid activation. Each output dimension represents a separate binary label. The task is trained with a multi-label binary cross-entropy (BCE) loss (summing the per-dimension BCE losses). The second task (prediction of purity class, multiclass classification) produces 16 class logits. These 16 outputs are interpreted as a 16-way multiclass task and trained with a margin-based multiclass loss (PyTorch's implementation of MultiMarginLoss). A margin-based loss is independent of the exact probability scores, in contrast with cross entropy loss; instead, it focuses on ensuring that the score of the correct class is higher than the scores of incorrect classes by a specified margin. For the main task, using MultiMarginLoss as loss function results in to a smoother convergence and a more stable training process because the loss function does not overreact to every minor prediction error.

While training, the losses from the two tasks were combined *via* summing, effectively weighting the tasks equally in the objective. In our implementation, each task's loss contributes 50% to the total loss.

Although we do not explicitly encode cross-task constraints, the shared experts and gating allow the model to capture implicit task relationships. For example, if the data has patterns such as "whenever class "Amine is contaminated with halide" is predicted, label indices "amine" and "halide" are active", the network can learn this correlation through the shared expert features and task gates. In summary, the model uses joint representation learning (shared experts) combined with task-specific gates and decoders to perform both tasks simultaneously, based on the data to inform any structured dependencies.

Table S1. Parameters of learning

Name	Value
Weighting	Equal weighting
Optimizer	Adam
Learning rate	0.0001
Batch size	60
Scheduler	ReduceOnPlateau
Scheduler mode	Max
Scheduler factor	0.9
Scheduler patience	5
Scheduler cooldown	2
Scheduler threshold	0.001
Epoch for training	60
Best model epoch	53

S4 External test set preparation

Cyclohexanone (99.8%), propionic acid (99%), propionyl chloride (98%), 1-bromohexane (>99%) were purchased from Acros Organic. Di-*n*-butyl ether (99%) was purchased from abcr. Cyclohexanol, hexan-1-ol, valeric acid, butan-1-ol, phthalic acid, ethanol, ethyl acetate, CCl_4 were purchased from local chemical company Vecton. Methyl valerate was synthesized using triglyceride hydrolysis reaction.

All the substances were used without further purification, except CCl_4, which was distilled to remove water according to the standard procedure. The concentrations of the substances and mixtures in CCl_4 were selected individually taking into account the signal intensities: concentrations of about 5% in CCl_4 were used, and lower concentrations were used if a substance contained a strong oscillator (for example, C=O functionality).

An infrared spectrum of a sample was recorded using a Shimadzu IRAffinity-1 Fourier transform infrared spectrophotometer (Spectra 1-14, 16-18). Spectrum recording parameters were as follows: wavenumbers range of 4000–400 cm^{-1}, number of scans was 32, resolution of 2 cm^{-1}, Happ-Genzel apodization function. An analyzed solution was placed in a 0.025 mm thick cuvette. The atmospheric spectrum and the CCl_4 spectrum were recorded before each substance were subtracted from the recorded spectrum. All the data were converted to a wavenumbers range of 4000 to 400 cm^{-1}, with a resolution of ~6 cm^{-1}. The intensities in the ranges of 1500-1650 cm^{-1} and 650-860 cm^{-1} were set to 0.001 to remove peaks corresponding to solvent absorption.

An additional external infrared spectra of a samples were recorded using a Bruker Tensor 27 Fourier transform infrared spectrophotometer to evaluate reproducibility of the model results over instruments (Spectra 15, S19-S22). Spectrum recording parameters were as follows: wavenumbers range of 4000–400 cm^{-1}, number of scans was 40, resolution of 2 cm^{-1}, Blackman-Harris 3-Term apodization function. *Spectral data S19-S22 were not included in the data for F_1-score calculation on the external set.*

While generating the training set: all the substances and pairs from the external test set were strictly transferred to the generated test set using the swap_elements function only in this case of if those substances and pairs were in the generated set.

The data separation procedure is as follows:

- For each substance in the external test, a check was performed to see if the spectrum of this pure substance was present in processed_dataset. This was done manually.

- If it is, the row index of this substance in processed_dataset was written to a separate list (example, test, test_13 in the generate_db.py script).

- The function swap_elements operates by identifying elements in the train set that also contain in the external test set, treating these overlaps as data leaks that must be removed from train and transferred into the test set. The process begins by identifying common elements, termed leaked_from_train elements, present in both train and external_test.

- If no leakage is detected, the original sets are returned. If leakage is detected, the function computes the size of the necessary swap and partitions the data: the new train and test set is formed a pure_train set and pure_candidates_from_test, excluding the elements from external_test.

- A subset of these pure candidates (to_move_from_test), equal in size to num_to_swap, is extracted from the test set. These elements are then inserted into the new_train set, replacing the removed leaked elements.

- The new_test set is updated. The leaked_from_train elements are moved from the original train set into the new_test set, replacing the first element of the to_move_from_test subset that was just used to fill up the train set.

It was applied to the following compounds - 1-bromohexane, methyl valerate, di-*n*-butyl ether, cyclohexanone (spectral data for all these compounds are presented in the SDBS database), and pairs of compounds – 1-bromohexane and hexan-1-ol; di-*n*-butyl ether and butan-1-ol; cyclohexanol and cyclohexanone.

The presented interpretation of the infrared spectra is based on manual decoding and presumptive identification of absorption peaks.

List of abbreviations in pictures:

str – stretching vibration;

bend – bending vibration;

sciss – scissoring vibration;

wag – waggiing vibration.

S5 External test set visualization and prediction

Spectrum 1: Cyclohexanol (0.0931 g) and cyclohexanone (0.0290 g) were dissolved in CCl_4 (2.4465 g) to obtain the mixture of an alcohol and a ketone in a 76.25 : 23.75 mass ratio (4.75 wt% solution in CCl_4) and to simulate class «alcohol is contaminated with ketone» (class 13). Then, 0.1390 g of this solution and 1.3287 g of CCl_4 were taken to reduce the concentration of the mixture by ~ 10 times and obtain 0.45 wt% solution in CCl_4.

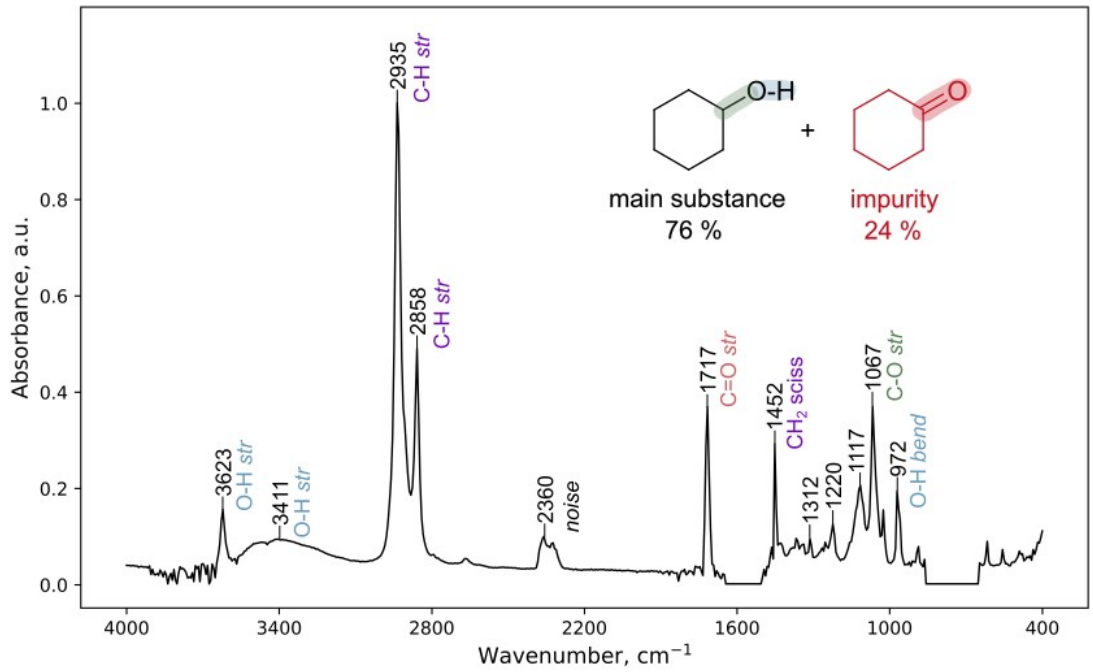

Figure S8. The IR-spectrum of the mixture of cyclohexanone and cyclohexanol in 76:24 ratio (0.45 wt% solution in CCl_4).

Table S2. Predictions of the model

Labels	Alkane	Alcohol	Ketone	Purity class
True	+	+	+	alcohol is contaminated with ketone
Predicted	+	+	+	alcohol is contaminated with ketone

Spectrum 2: Cyclohexanone (0.0788 g) were dissolved in CCl_4 (1.6367 g) (4.59 wt% solution in CCl_4).

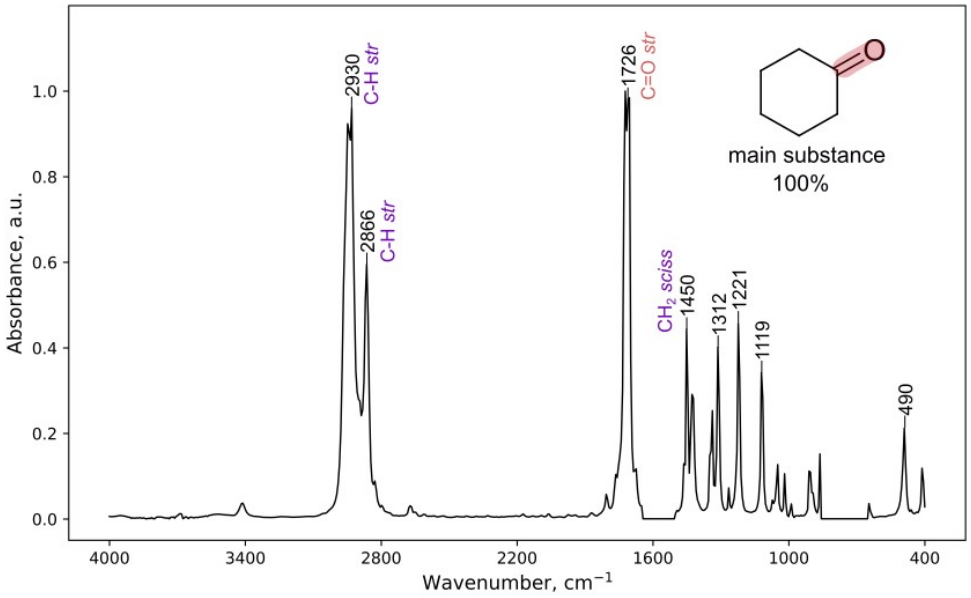

Figure S9. The IR-spectrum of pure cyclohexanone (4.6 wt% solution in CCl_4).

Table S3. Predictions of the model

Labels	Alkane	Ketone	Purity class
True	+	+	pure substance
Predicted	+	+	pure substance

Spectrum 3: Propionyl chloride (0.1285 g) was dissolved in CCl_4 (2.4586 g) (4.97 wt% solution in CCl_4).

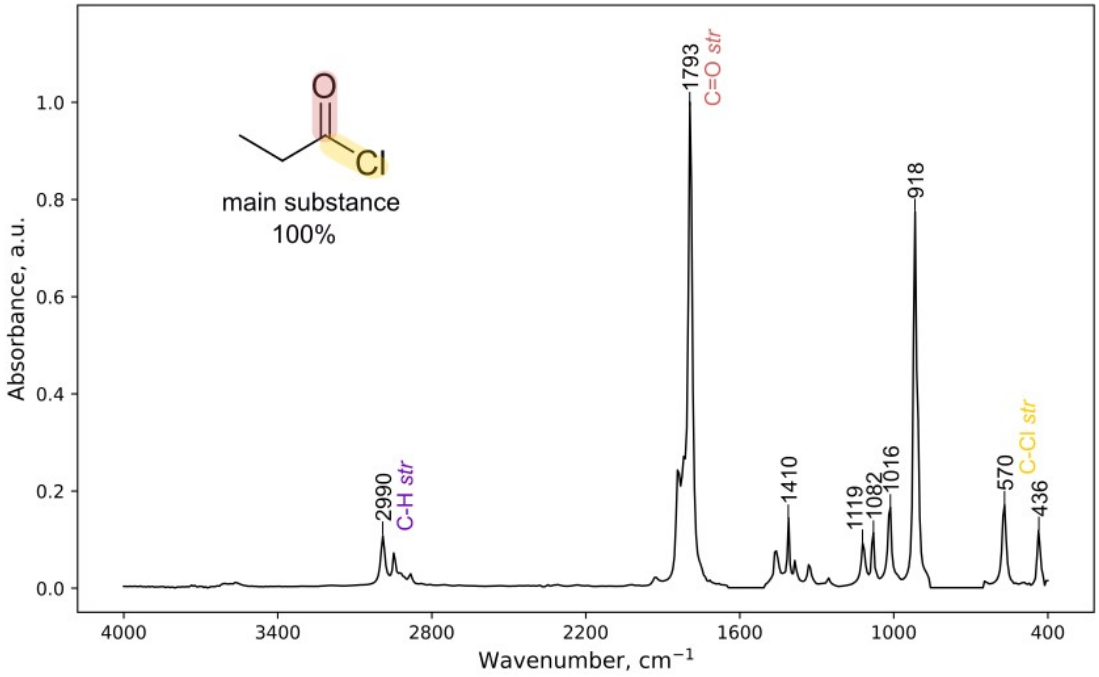

Figure S10. The IR-spectrum of pure propionyl chloride (5 wt% solution in CCl_4).

Table S4. Predictions of the model

Labels	Alkane	Haloalkane	Acyl halide	Purity class
True	+	+	+	pure substance
Predicted	+	+	+	pure substance

Spectrum 4: Propionyl chloride (0.1134 g) and propionic acid (0.0285 g) were dissolved in CCl_4 (2.7000 g) to obtain the mixture of an acyl halide and an acid in 79.9 : 20.1 mass ratio (4.99 wt% solution in CCl_4) and to simulate class «acyl halide is contaminated with acid» (class 9).

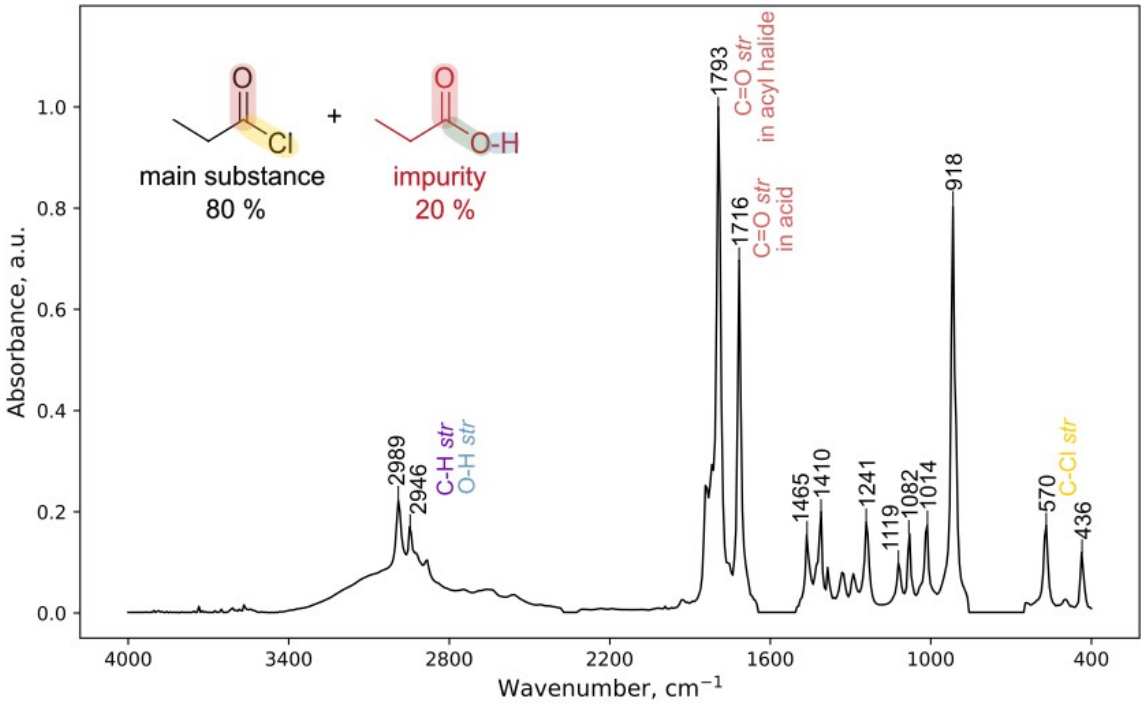

Figure S11. The IR-spectrum of the mixture of propionyl chloride and propionic acid in 80:20 ratio (5 wt% solution in CCl_4).

Table S5. Predictions of the model

Labels	Alkane	Haloalkane	Alcohol	Carboxylic acid	Acyl halide	Purity class
True	+	+	+	+	+	Acyl halide is contaminated with acid
Predicted	+	+	+	+	+	Acyl halide is contaminated with acid

Spectrum 5: Propionyl chloride (0.1078 g) and propionic acid (0.0118 g) were dissolved in CCl_4 (2.2574 g) to obtain the mixture of an acyl halide and an acid in 90.1 : 9.9 mass ratio (5.03 wt% solution in CCl_4) and to simulate class «acyl halide is contaminated with acid» (class 9).

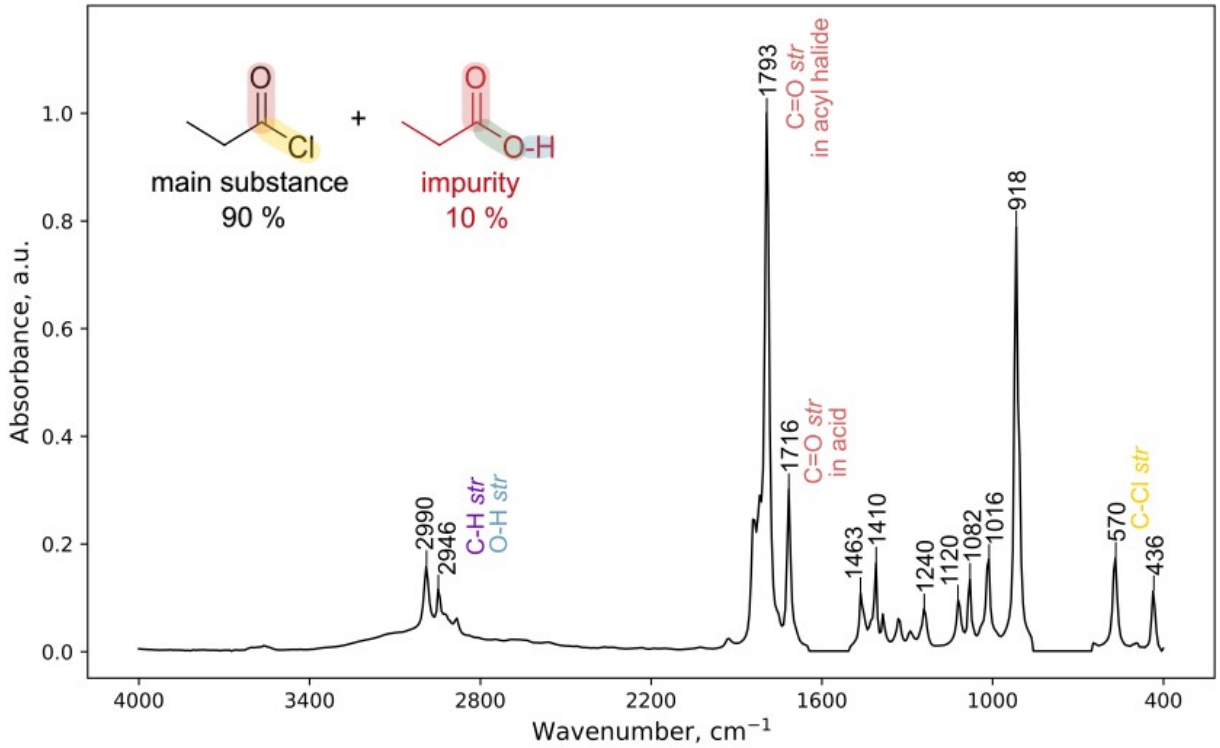

Figure S12. The IR-spectrum of the mixture of propionyl chloride and propionic acid in 90:10 ratio (5 wt% solution in CCl_4).

Table S6. Predictions of the model

Labels	Alkane	Halide	Alcohol	Carboxylic acid	Acyl halide	Purity class
True	+	+	+	+	+	Acyl halide is contaminated with acid
Predicted	+	+	+	+	+	Acyl halide is contaminated with acid

Spectrum 6: Propionic acid (0.0293 g) was added to a solution 2 to obtain the mixture of an acyl halide and an acid in ~ 72.4 : 27.6 mass ratio (~ 6 wt% solution in CCl_4) and to simulate class «acyl halide is contaminated with by acid» (class 9).

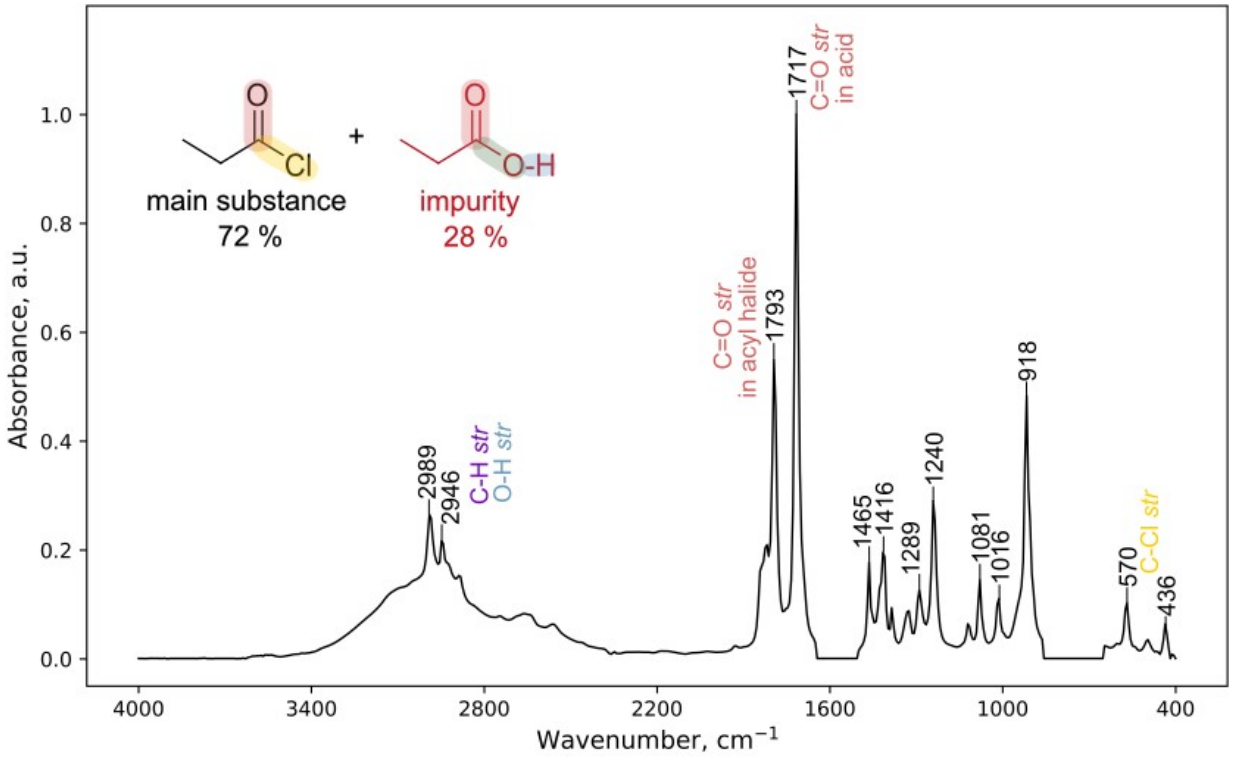

Figure S13. The IR-spectrum of the mixture of propionyl chloride and propionic acid in 72:28 ratio (6 wt% solution in CCl_4).

Table S7. Predictions of the model

Labels	Alkane	Halide	Alcohol	Carboxylic acid	Acyl halide	Purity class
True	+	+	+	+	+	Acyl halide is contaminated with acid
Predicted	+	+	+	+	-	Pure

Spectrum 7: Methyl valerate (0.0215 g) was dissolved in CCl_4 (0.0428 g) (4.78 wt% solution in CCl_4).

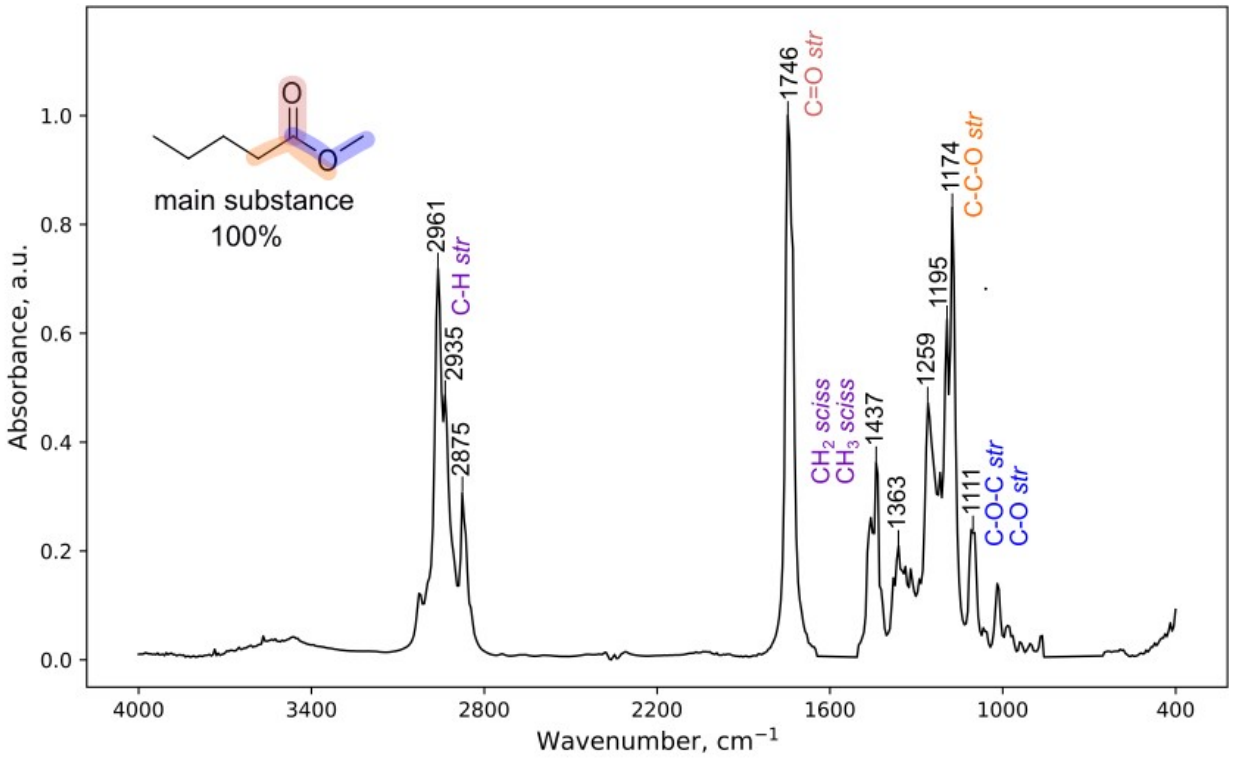

Figure S14. The IR-spectrum of pure methyl valerate (4.8 wt% solution in CCl_4).

Table S8. Predictions of the model

Labels	Alkane	Ester	Ether	Purity class
True	+	+	+	pure substance
Predicted	+	+	+	pure substance

Spectrum 8: Methyl valerate (0.1801 g) and valeric acid (0.0316 g) were dissolved in CCl_4 (1.8953 g) to obtain the mixture of an ester and an acid in 85.07 : 14.93 mass ratio (10.05 wt% solution in CCl_4) and to simulate class «ester is contaminated with acid» (class 14).

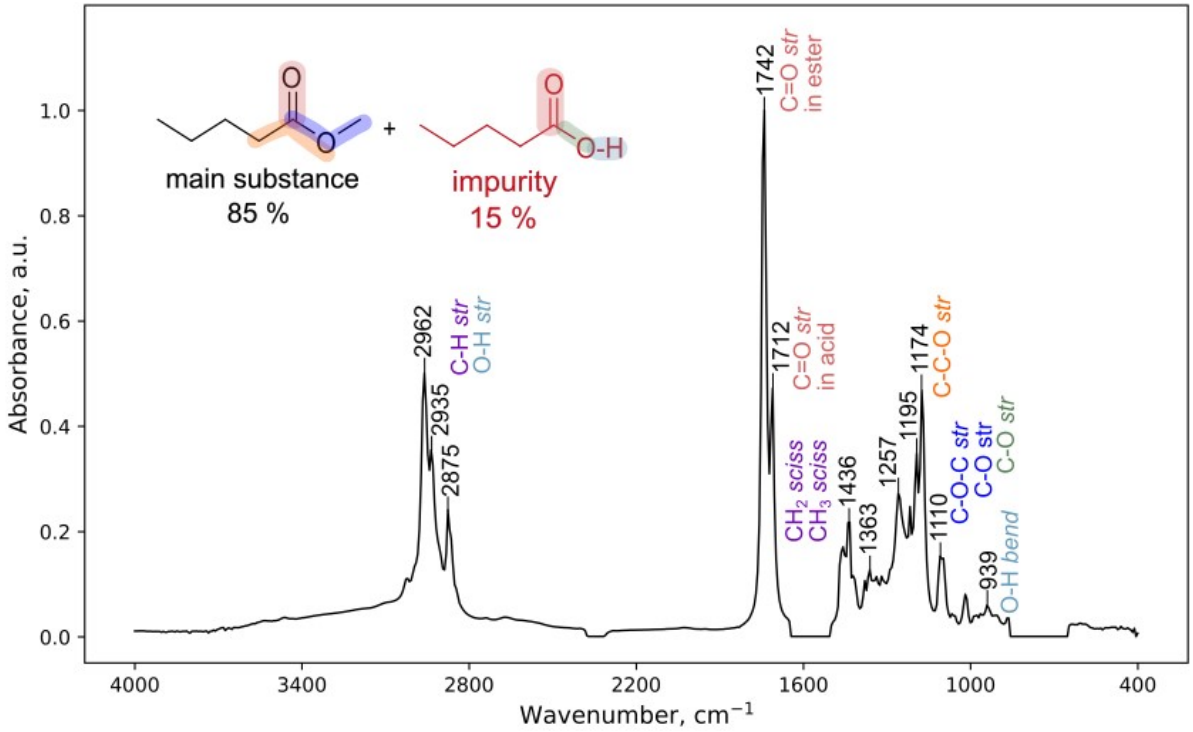

Figure S15. The IR-spectrum of the mixture of methyl valerate and valeric acid in 85:15 ratio (10.1 wt% solution in CCl_4).

Table S9. Predictions of the model

Labels	Alkane	Alcohol	Carboxylic acid	Ester	Ether	Purity class
True	+	+	+	+	+	Ester is contaminated with acid
Predicted	+	+	+	+	+	Ester is contaminated with acid

Spectrum 9: Valeric acid (0.0164 g) and CCl_4 (0.1714 g) was added to the solution 2 to obtain the mixture of an ester and an acid in ~ 78.96 : 21.04 mass ratio (~ 9.94 wt% solution in CCl_4) and to simulate class «ester is contaminated with acid» (class 14).

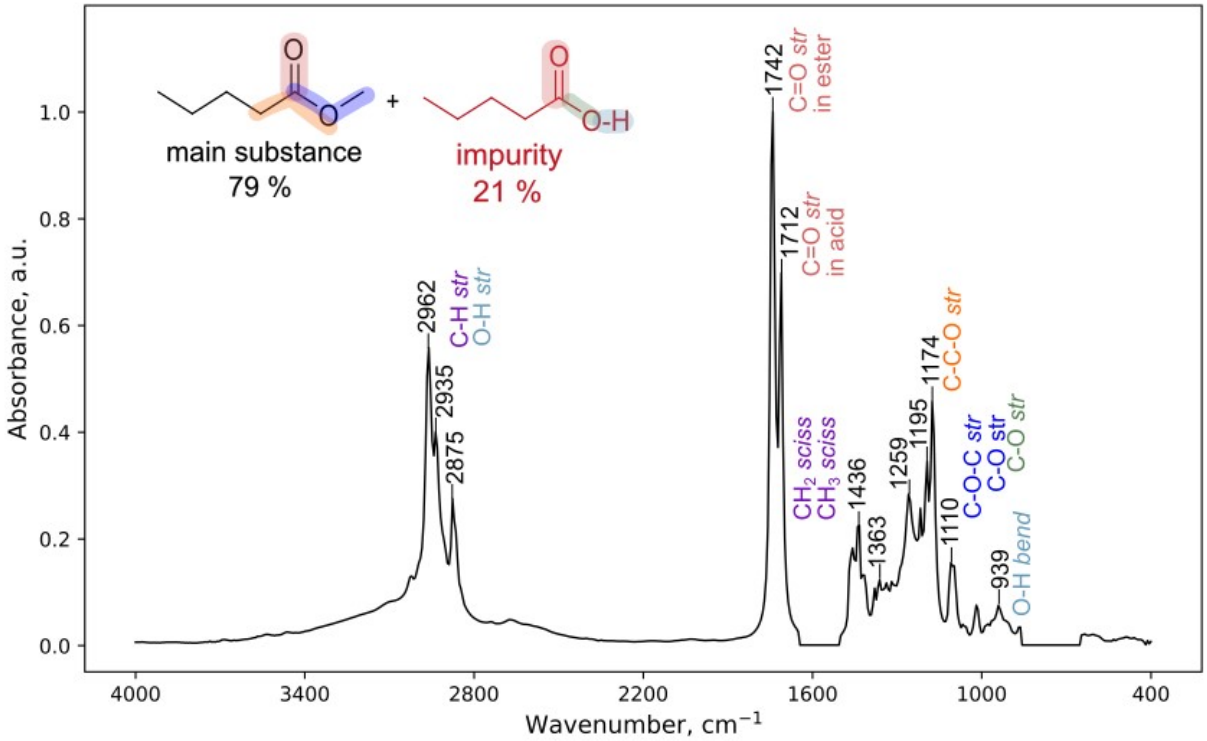

Figure S16. The IR-spectrum of the mixture of methyl valerate and valeric acid in 79:21 ratio (~ 9.9 wt% solution in CCl_4).

Table S10. Predictions of the model

Labels	Alkane	Alcohol	Carboxylic acid	Ester	Ether	Purity class
True	+	+	+	+	+	Ester is contaminated with acid
Predicted	+	+	+	+	+	Ester is contaminated with acid

Spectrum 10: Valeric acid (0.0540 g) was dissolved in CCl_4 (0.4910 g) (9.91 wt% solution in CCl_4).

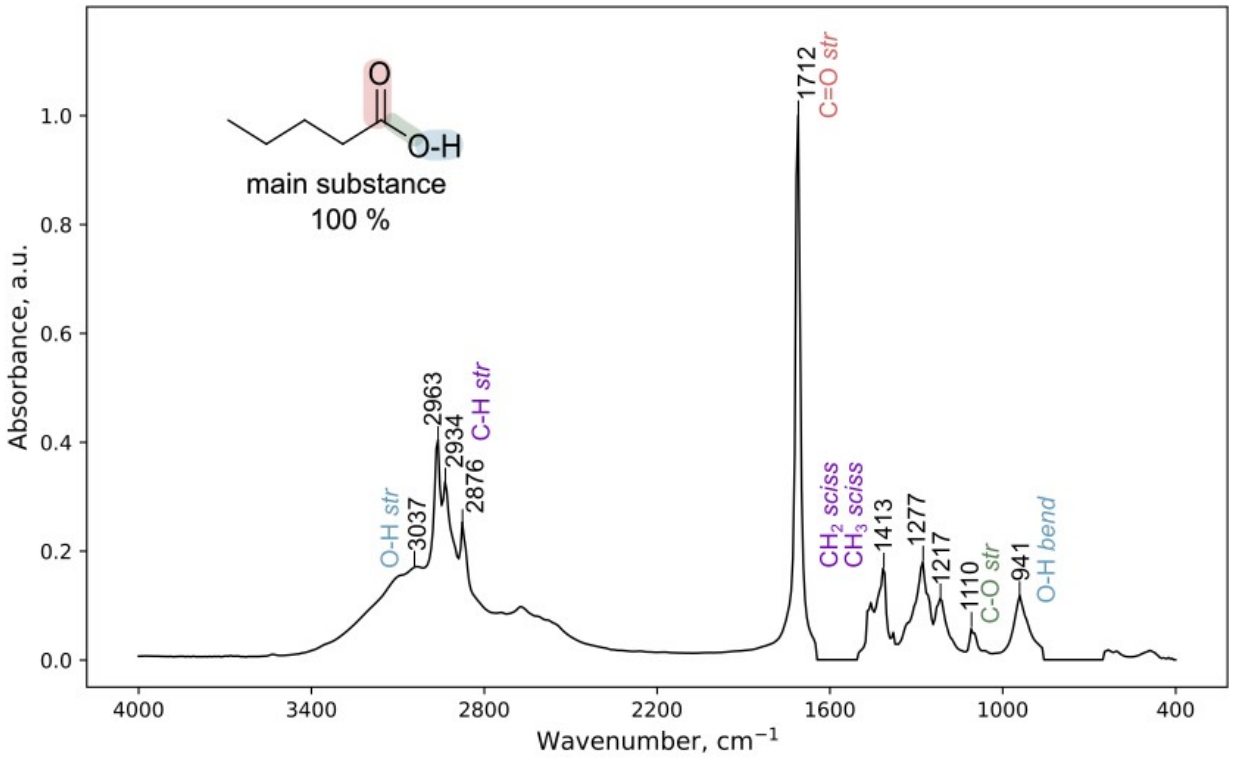

Figure S17. The IR-spectrum of pure valeric acid (9.9 wt% solution in CCl_4).

Table S11. Predictions of the model

Labels	Alkane	Alcohol	Carboxylic acid	Purity class
True	+	+	+	pure substance
Predicted	+	+	+	pure substance

Spectrum 11: Di-*n*-butyl ether (0.0267 g) was dissolved in CCl_4 (0.4982 g) (5.09 wt% solution in CCl_4).

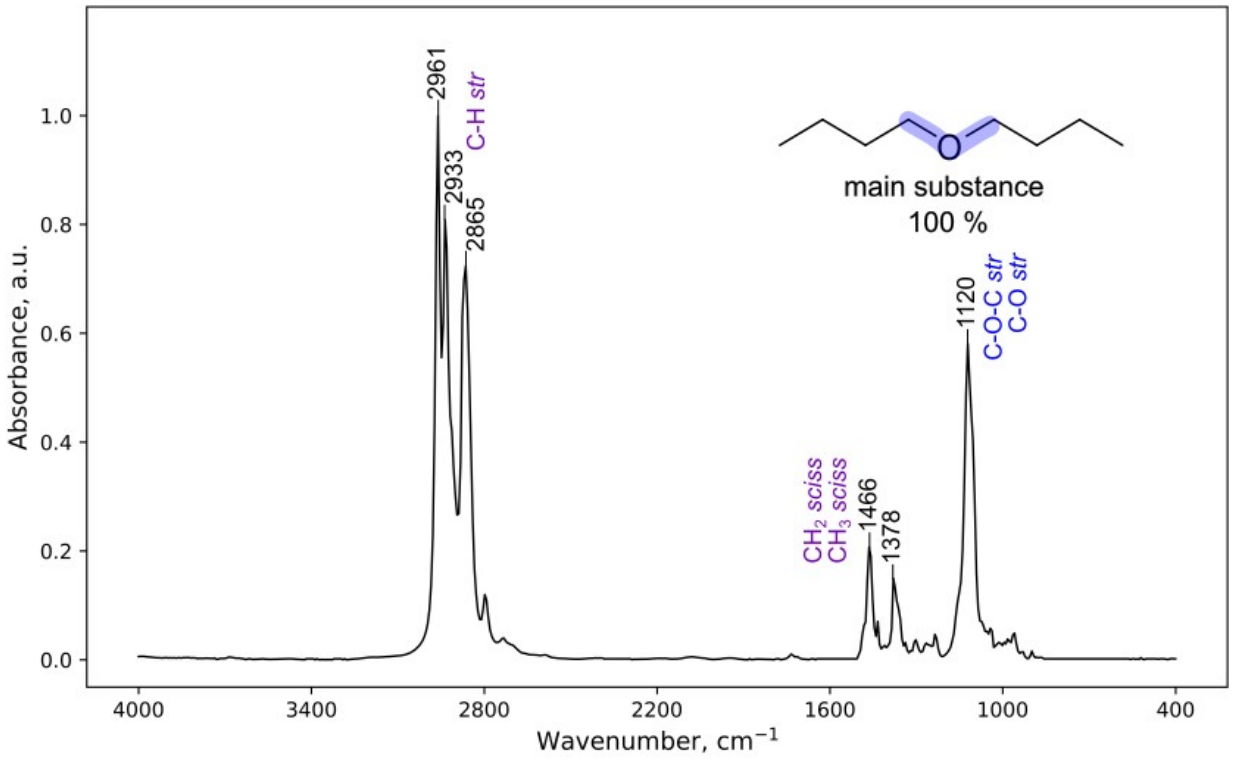

Figure S18. The IR-spectrum of pure di-*n*-butyl ether (5.1 wt% solution in CCl_4).

Table S12. Predictions of the model

Labels	Alkane	Ether	Purity class
True	+	+	pure substance
Predicted	+	+	pure substance

Spectrum 12: Di-*n*-butyl ether (0.1348 g) and butan-1-ol (0.0337 g) were dissolved in CCl_4 (1.5275 g) to obtain the mixture of an ether and an alcohol in 80 : 20 mass ratio (9.94 wt% solution in CCl_4) and to simulate class «ether is contaminated with alcohol» (class 4).

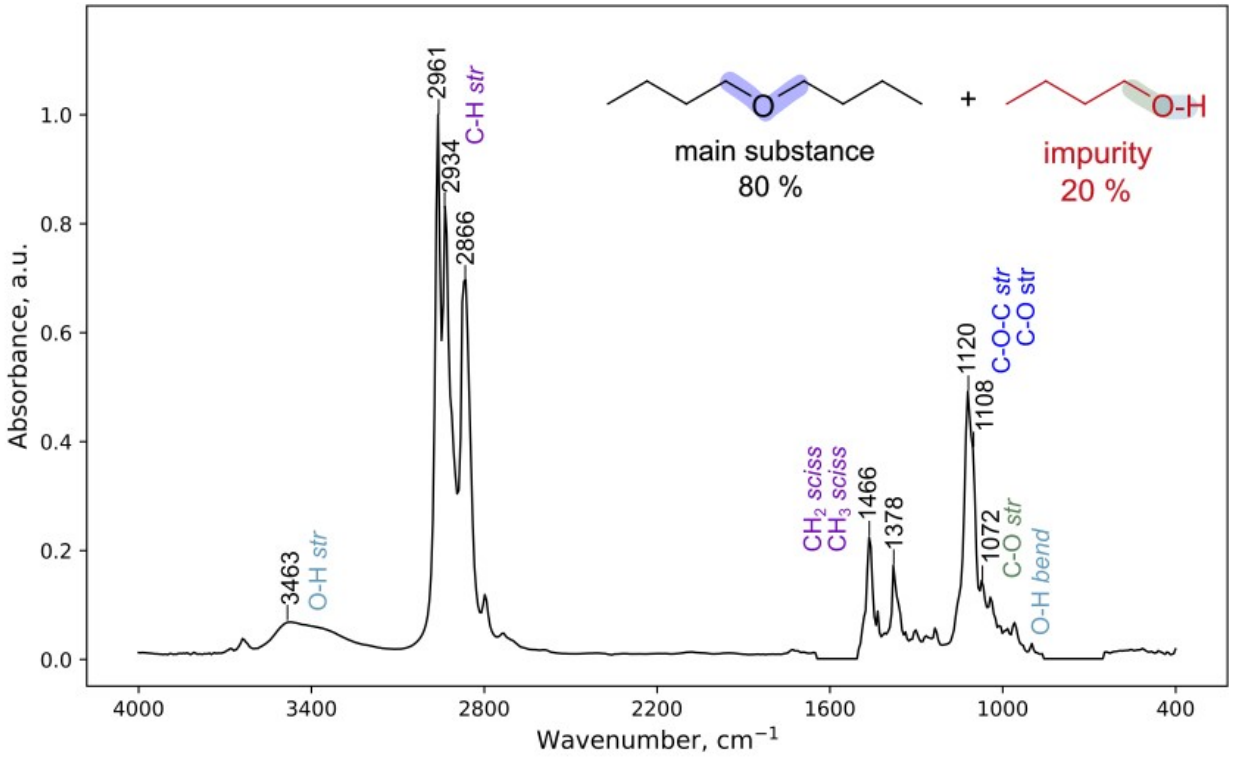

Figure S19. The IR-spectrum of the mixture of di-*n*-butyl ether and butan-1-ol in 80:20 ratio (9.9 wt% solution in CCl_4).

Table S13. Predictions of the model

Labels	Alkane	Alcohol	Ether	Purity class
True	+	+	+	Ether is contaminated with alcohol
Predicted	+	+	+	Ether is contaminated with alcohol

Spectrum 13: Di-*n*-butyl ether (0.5145 g) and butan-1-ol (0.0560 g) were dissolved in CCl_4 (3.2383 g) to obtain the mixture of an ether and an alcohol in 90.2 : 9.8 mass ratio (14.98 wt% solution in CCl_4) and to simulate class «ether is contaminated with alcohol» (class 4).

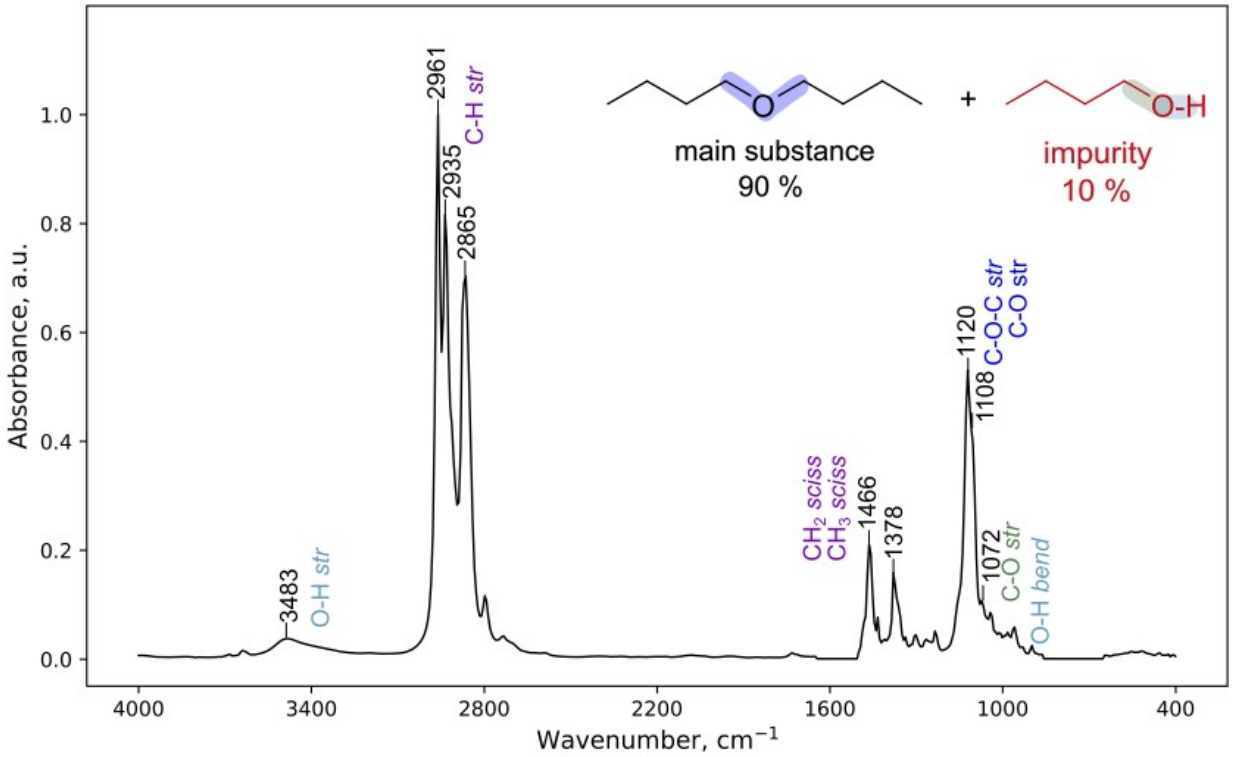

Figure S20. The IR-spectrum of mixture of di-n-butyl ether and butan-1-ol in 90:10 ratio (15 wt% solution in CCl_4).

Table S14. Predictions of the model

Labels	Alkane	Alcohol	Ether	Purity class
True	+	+	+	Ether is contaminated with alcohol
Predicted	+	+	+	Ether is contaminated with alcohol

Spectrum 14: Di-*n*-butyl ether (0.1779 g) and butan-1-ol (0.0307 g) were dissolved in CCl_4 (1.1833 g) to obtain the mixture of an ether and an alcohol in 85.3 : 14.7 mass ratio (14.99 wt% solution in CCl_4) and to simulate class «ether is contaminated with alcohol» (class 4).

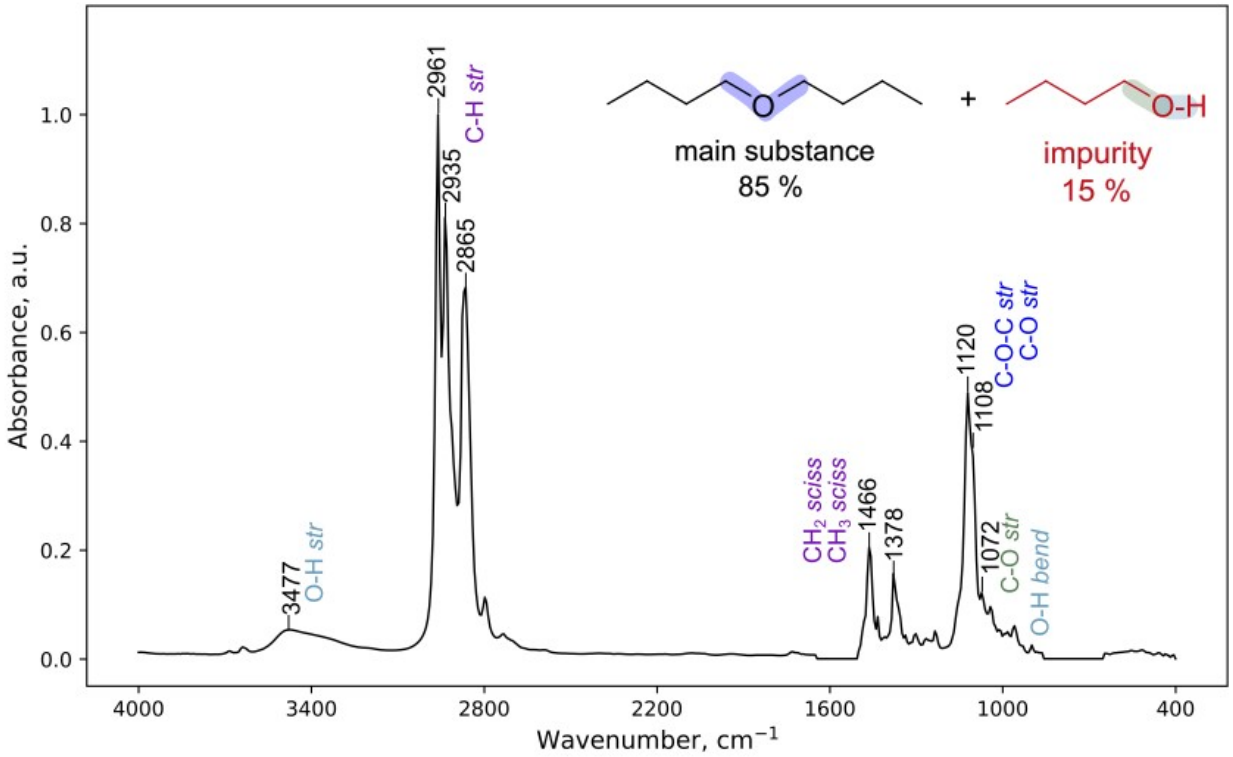

Figure S21. The IR-spectrum of mixture of di-*n*-butyl ether and butan-1-ol in 85:15 ratio (15 wt% solution in CCl_4).

Table S15. Predictions of the model

Labels	Alkane	Alcohol	Ether	Purity class
True	+	+	+	Ether is contaminated with alcohol
Predicted	+	+	+	Ether is contaminated with alcohol

Spectrum 15: Di-*n*-butyl ether (0.8057 g) and butan-1-ol (0.0426 g) were dissolved in CCl_4 (4.7465 g) to obtain the mixture of an ether and an alcohol in 95 : 5 mass ratio (15.16 wt% solution in CCl_4) and to simulate class «ether is contaminated with alcohol» (class 4).

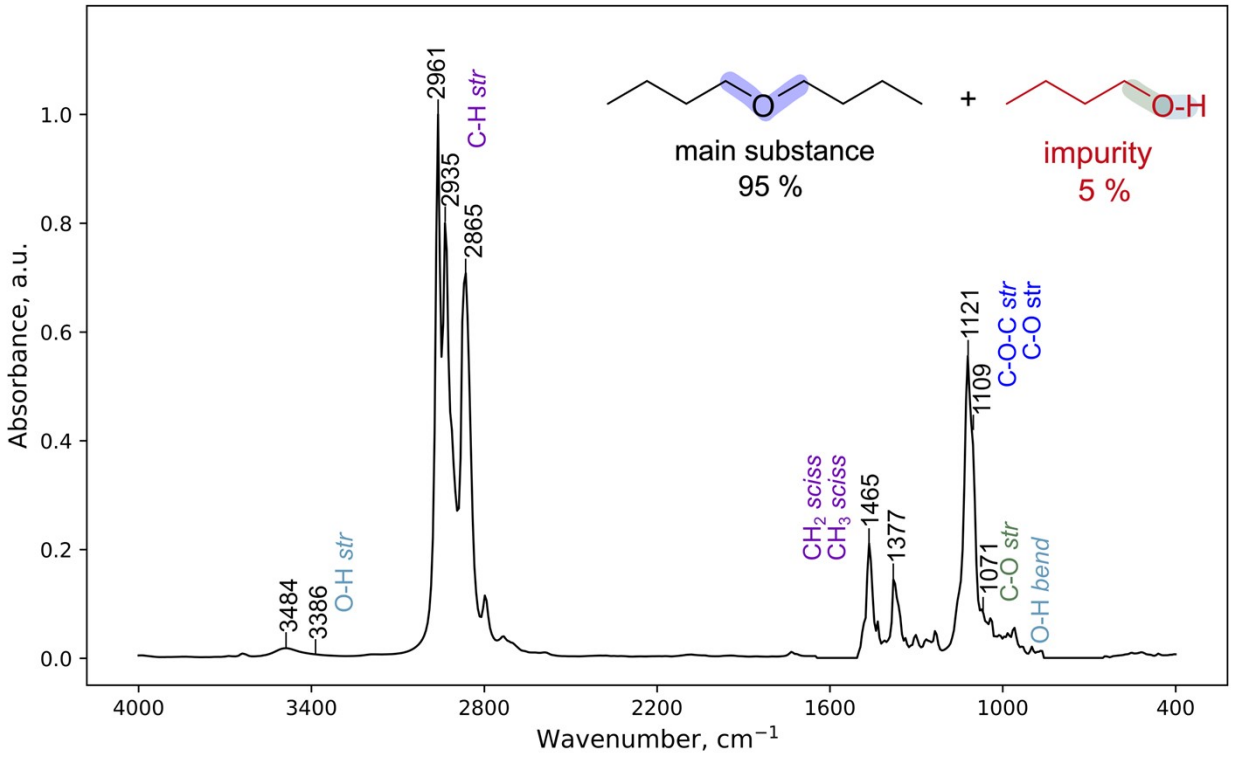

Figure S22. The IR-spectrum of mixture of di-n-butyl ether and butan-1-ol in 95:5 ratio (15.2 wt% solution in CCl_4).

Table S16. Predictions of the model

Labels	Alkane	Alcohol	Ether	Purity class
True	+	+	+	Ether is contaminated with alcohol
Predicted	+	+	+	Ether is contaminated with alcohol

Spectrum 16: 1-bromohexane (0.0632 g) was dissolved in CCl_4 (0.3532 g) (15.18 wt% solution in CCl_4).

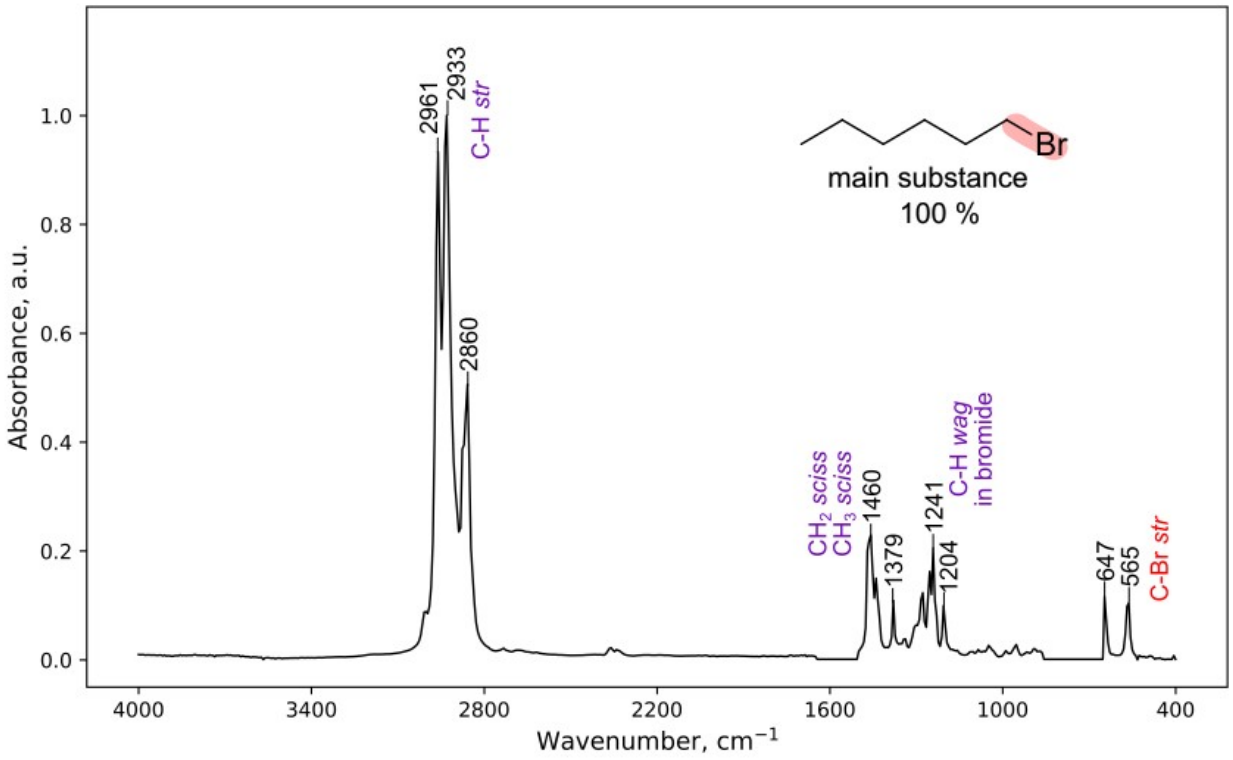

Figure S23. The IR-spectrum of pure 1-bromohexane (15.2 wt% solution in CCl_4).

Table S17. Predictions of the model

Labels	Alkane	Halide	Purity class
True	+	+	Pure substance
Predicted	+	+	Pure substance

Spectrum 17: 1-bromohexane (0.1268 g) and hexan-1-ol (0.0229 g) were dissolved in CCl_4 (0.8460 g) to obtain the mixture of a halide and an alcohol in 84.7 : 15.3 mass ratio (15 wt% solution in CCl_4) and to simulate class «halide is contaminated with alcohol» (class 10).

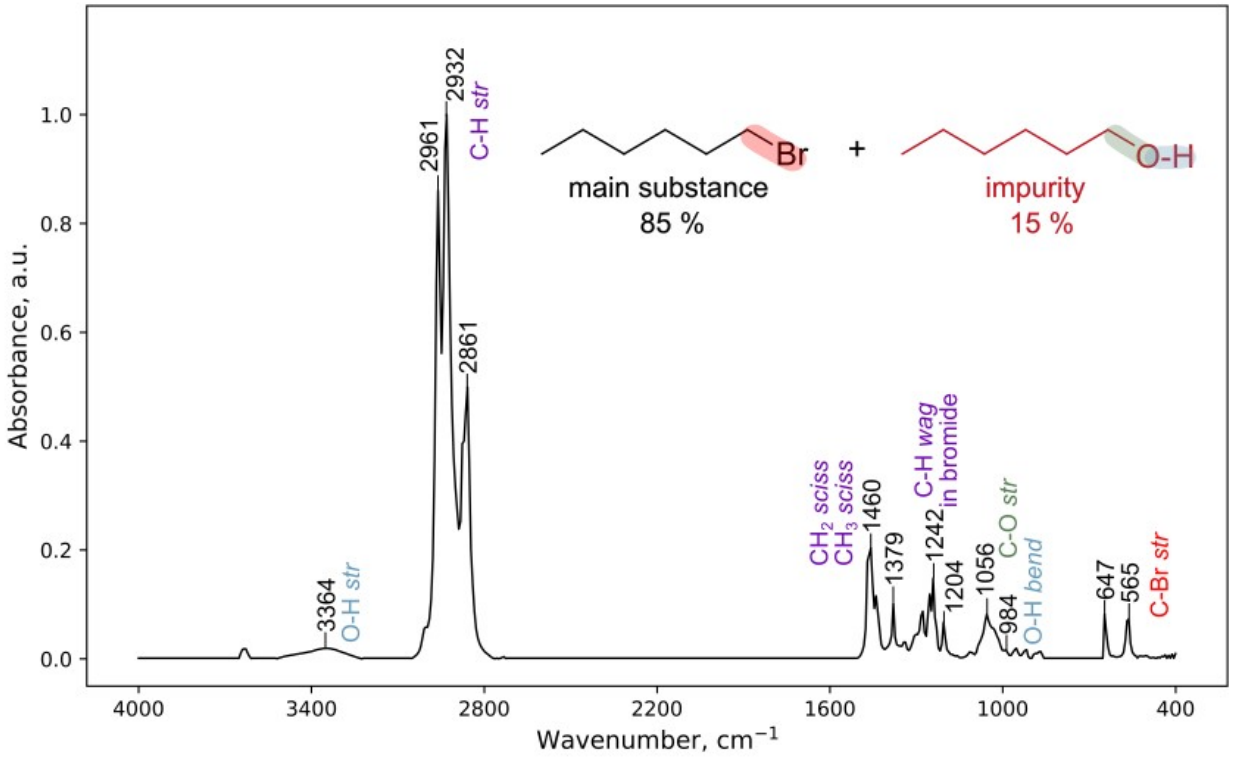

Figure S24. The IR-spectrum of the mixture of 1-bromohexane and hexan-1-ol in 85:15 ratio (15 wt% solution in CCl_4).

Table S18. Predictions of the model

Labels	Alkane	Halide	Alcohol	Purity class
True	+	+	+	Halide is contaminated with alcohol
Predicted	+	+	+	Halide is contaminated with alcohol

Spectrum 18: 1-bromohexane (0.1723 g) and hexan-1-ol (0.0427 g) were dissolved in CCl_4 (1.2314 g) to obtain the mixture of a halide and an alcohol in 80.1 : 19.9 mass ratio (14.9 wt% solution in CCl_4) and to simulate class «halide is contaminated with alcohol» (class 10).

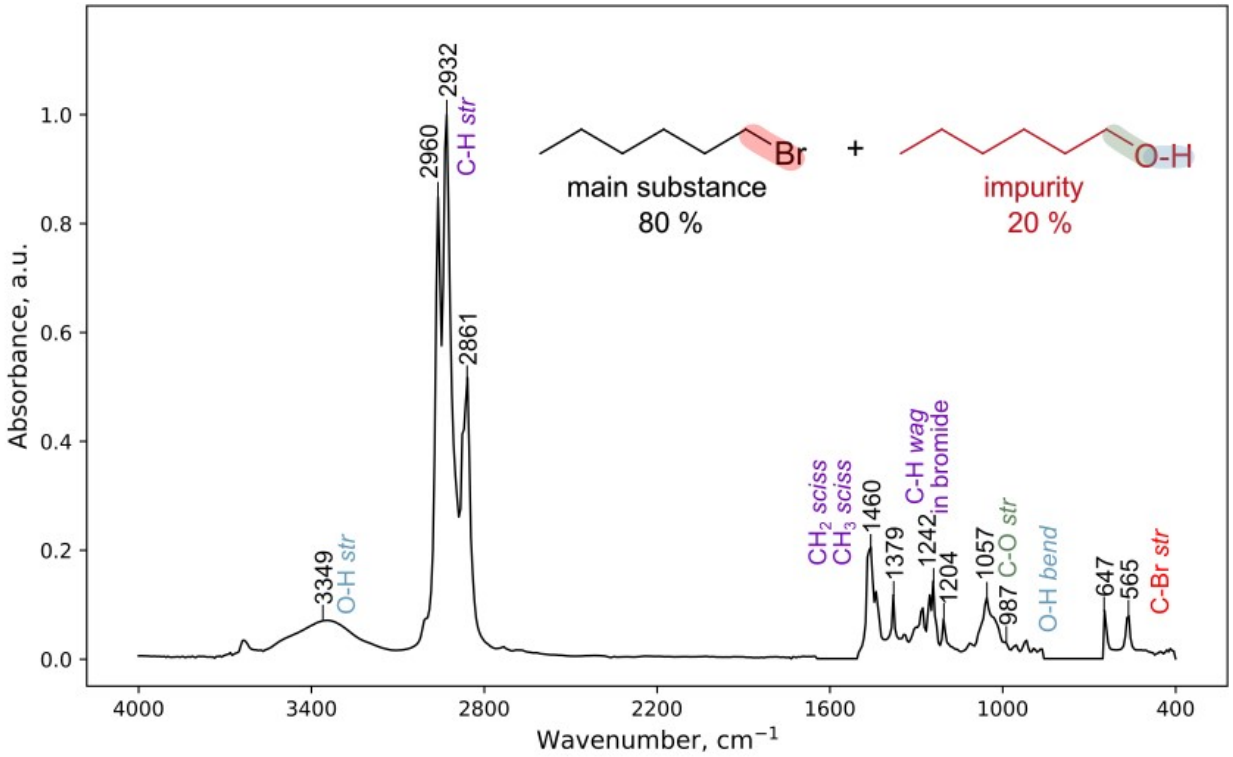

Figure S25. The IR-spectrum spectrum of the mixture of 1-bromohexane and hexan-1-ol in 80:20 ratio (14.9 wt% solution in CCl_4).

Table S19. Predictions of the model

Labels	Alkane	Halide	Alcohol	Purity class
True	+	+	+	Halide is contaminated with alcohol
Predicted	+	+	+	Halide is contaminated with alcohol

S6. An additional set of spectral data that was not included for model evaluation, showing the performance of the model

Spectrum S19: Di-*n*-butyl ether (0.0272 g) was dissolved in CCl_4 (0.5015 g) (5.14 wt% solution in CCl_4).

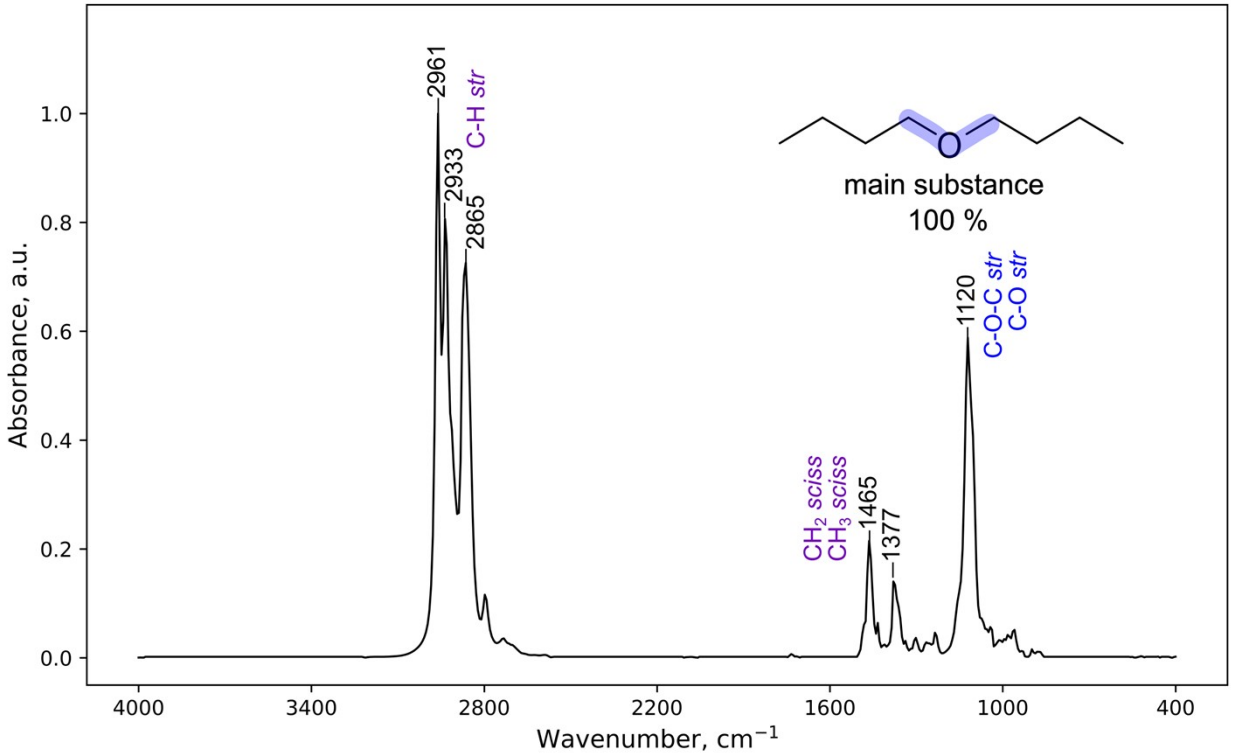

Figure S26. The IR-spectrum of pure di-*n*-butyl ether (5 wt% solution in CCl_4).

Table S20. Predictions of the model

Labels	Alkane	Ether	Purity class
True	+	+	pure substance
Predicted	+	+	pure substance

Spectrum S20: Di-*n*-butyl ether (0.5447 g) and butan-1-ol (0.0593 g) were dissolved in CCl_4 (3.4237 g) to obtain the mixture of an ether and an alcohol in 90.2 : 9.8 mass ratio (15 wt% solution in CCl_4) and to simulate class «ether is contaminated with alcohol» (class 4).

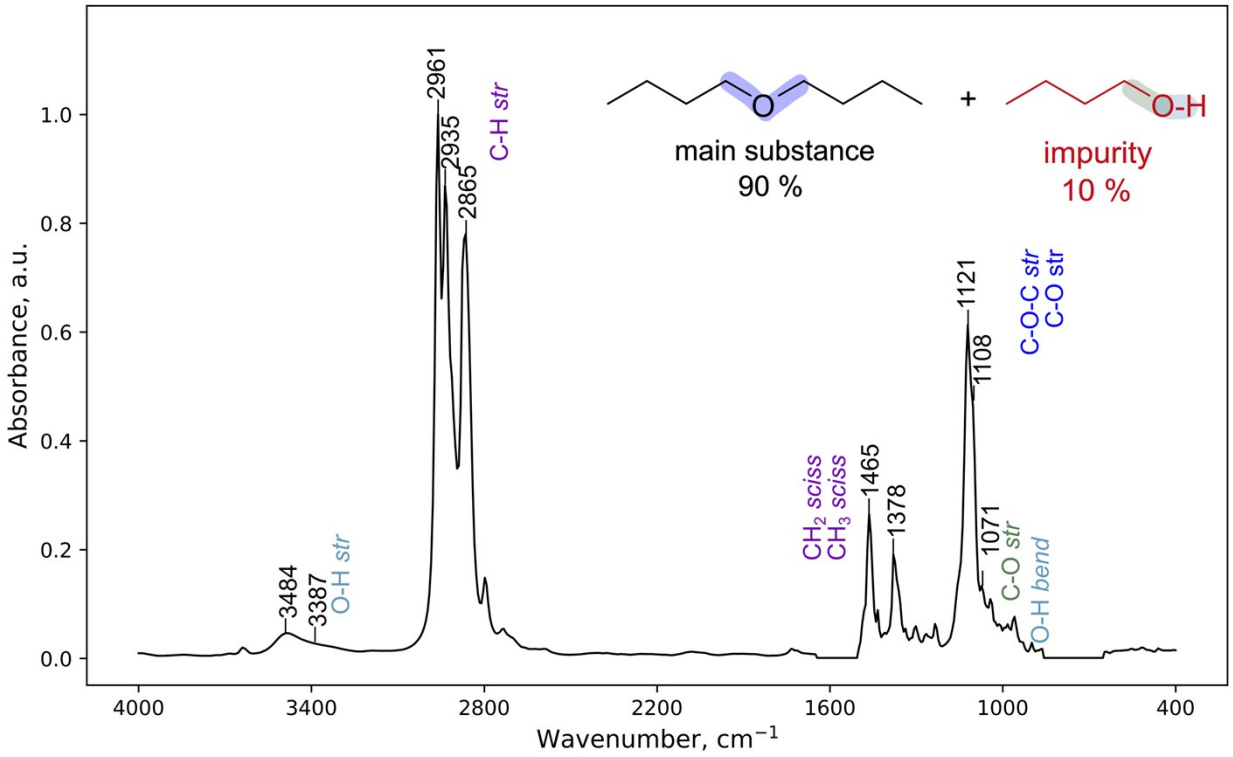

Figure S27. The IR-spectrum of the mixture of di-*n*-butyl ether and butan-1-ol in 90:10 ratio (15 wt% solution in CCl_4).

Table S21. Predictions of the model

Labels	Alkane	Alcohol	Ether	Purity class
True	+	+	+	Ether is contaminated with alcohol
Predicted	+	+	+	Ether is contaminated with alcohol

Spectrum S21: Propionyl chloride (0.1165 g) and propionic acid (0.0300 g) were dissolved in CCl_4 (2.7094 g) to obtain the mixture of an acyl halide and an acid in 79.5 : 20.5 mass ratio (5.13 wt% solution in CCl_4) and to simulate class «acyl halide is contaminated with acid» (class 9).

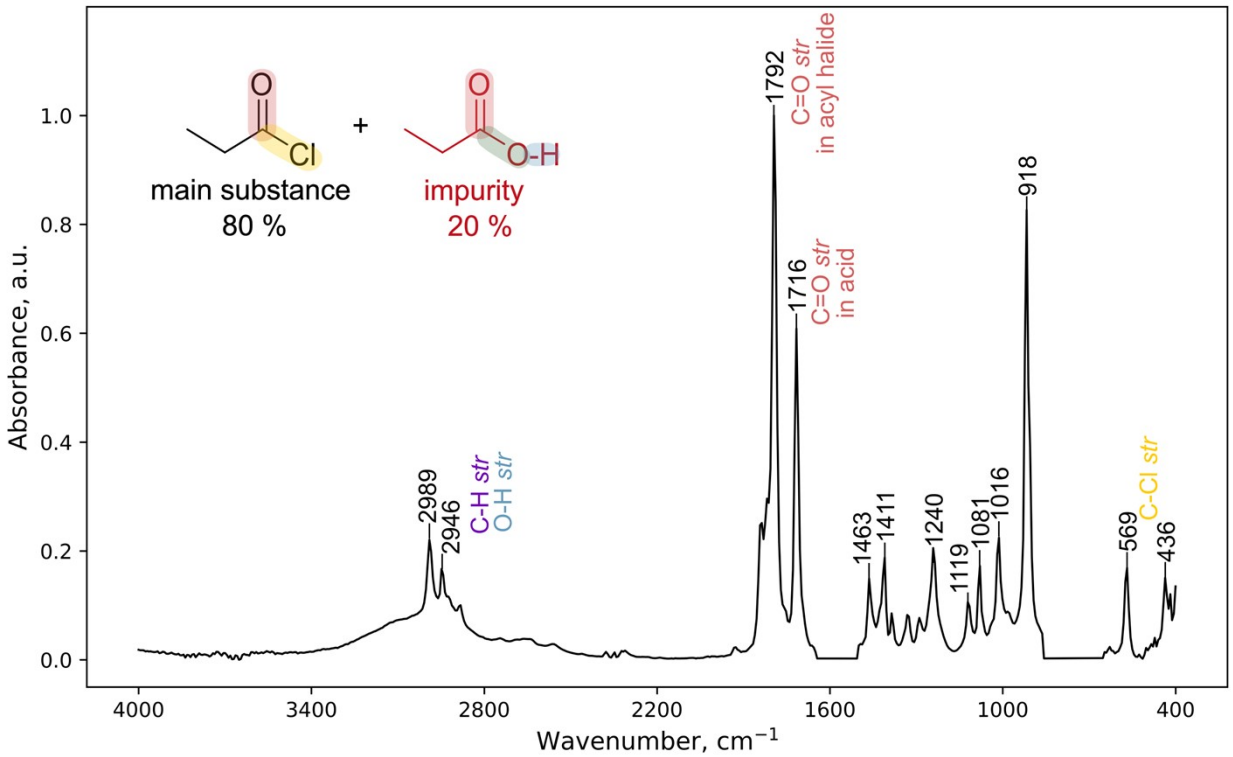

Figure S28. The IR-spectrum of the mixture of propionyl chloride and propionic acid in 80:20 ratio (5.1 wt% solution in CCl_4).

Table S22. Predictions of the model

Labels	Alkane	Haloalkane	Alcohol	Carboxylic acid	Acyl halide	Purity class
True	+	+	+	+	+	Acyl halide is contaminated with acid
Predicted	+	+	+	+	+	Acyl halide is contaminated with acid

Spectrum S22: Propionyl chloride (0.1111 g) and propionic acid (0.0118 g) were dissolved in CCl_4 (2.2606 g) to obtain the mixture of an acyl halide and an acid in 90.4 : 9.6 mass ratio (5.16 wt% solution in CCl_4) and to simulate class «acyl halide is contaminated with acid» (class 9).

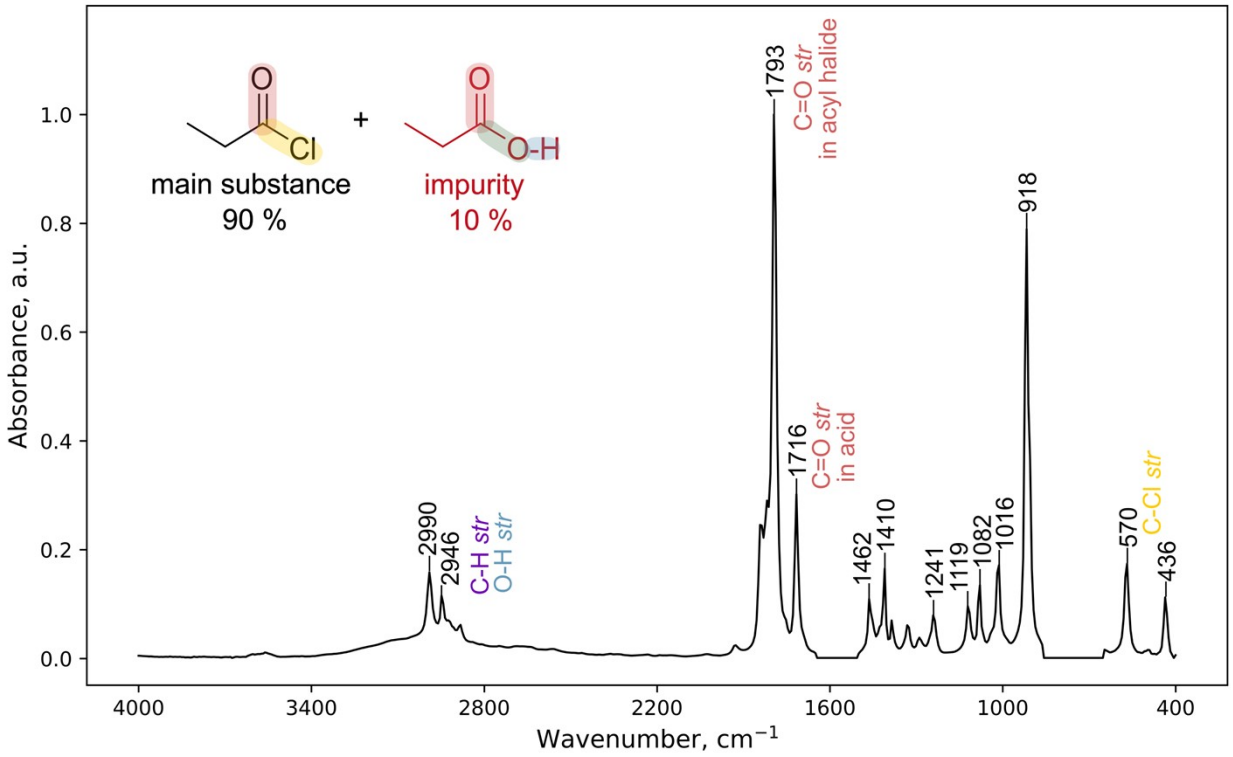

Figure S29. The IR-spectrum of the mixture of propionyl chloride and propionic acid in 90:10 ratio (5.2 wt% solution in CCl_4).

Table S23. Predictions of the model

Labels	Alkane	Halide	Alcohol	Carboxylic acid	Acyl halide	Purity class
True	+	+	+	+	+	Acyl halide is contaminated with acid
Predicted	+	+	+	+	+	Acyl halide is contaminated with acid

Spectra S23: Ethanol (0.2043 g) was added to ethyl acetate (1.0098 g) to obtain a mixture of ester and alcohol in a ~ 83.2 : 16.8 mass ratio (in KBr) and to simulate class «Ester is contaminated with alcohol» (class 14).

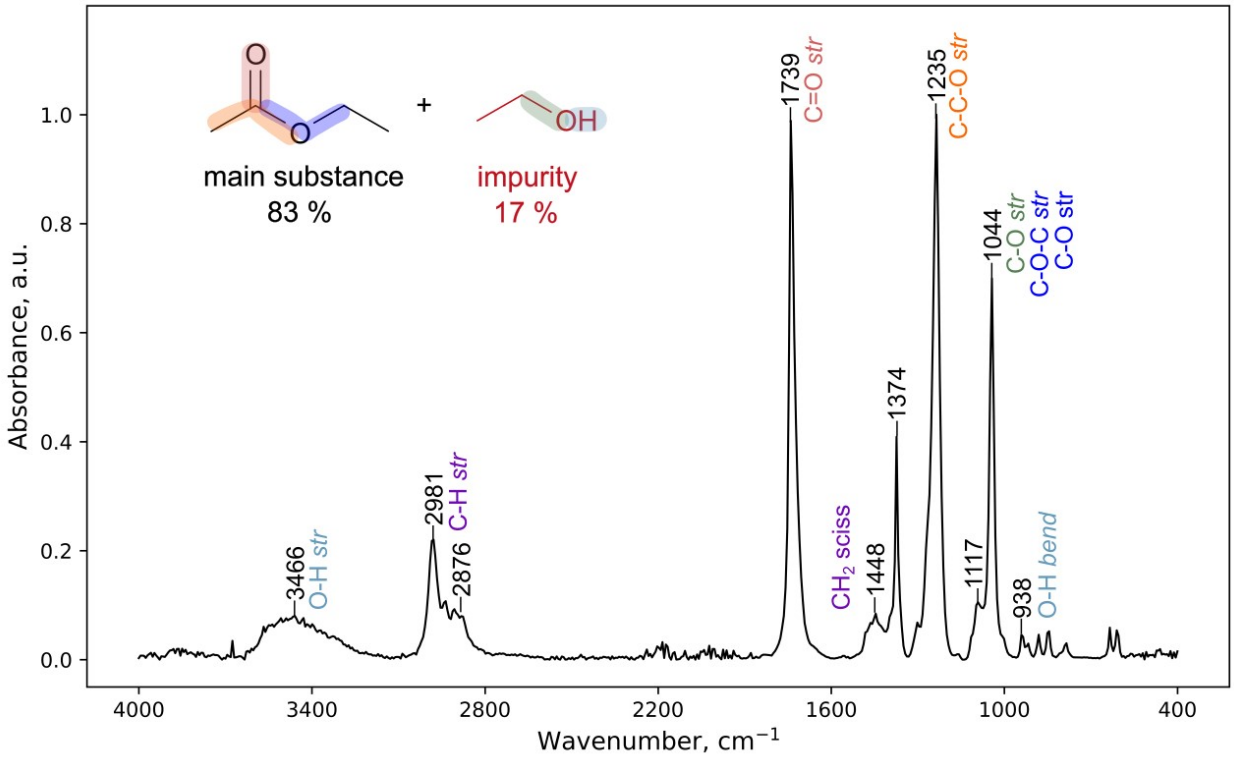

Figure S30. The IR-spectrum of the mixture of ethyl acetate and ethanol in 83:17 ratio (in KBr).

Table S24. Predictions of the model

Labels	Alkane	Alcohol	Ester	Ether	Purity class
True	+	+	+	+	Ester is contaminated with alcohol
Predicted	+	+	+	+	Ester is contaminated with alcohol

Spectra S25: Water (0.1381 g) was added to propionic acid (0.8010 g) to obtain a mixture of acid and water in a ~ 85.3 : 14.7 mass ratio and to simulate class «molecule is contaminated with water » (class 3).

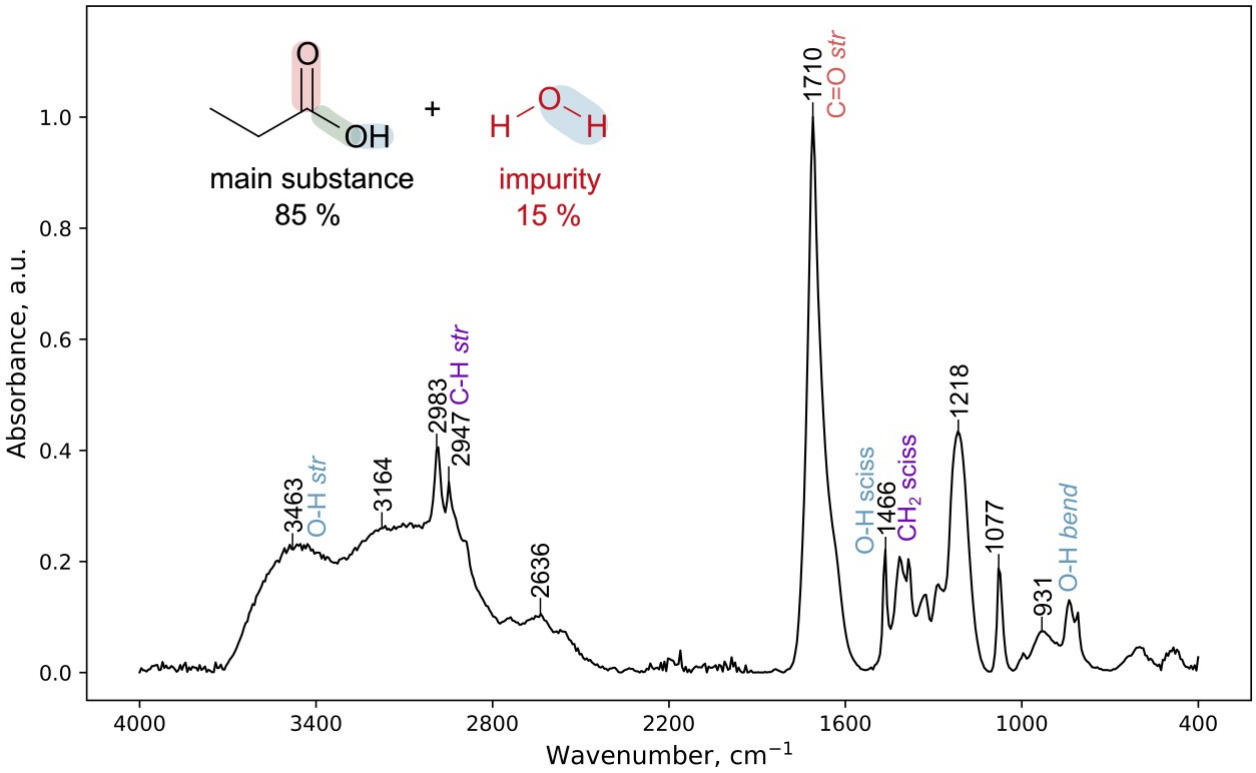

Figure S31. The IR-spectrum of the mixture of propionic acid and water in 85:15 ratio (in KBr).

Table S25. Predictions of the model

Labels	Alkane	Alcohol	Carboxylic acid	Purity class
True	+	+	+	Molecule is contaminated with water
Predicted	+	+	+	Molecule is contaminated with water

Spectra S26: Pure phthalic acid (in KBr)

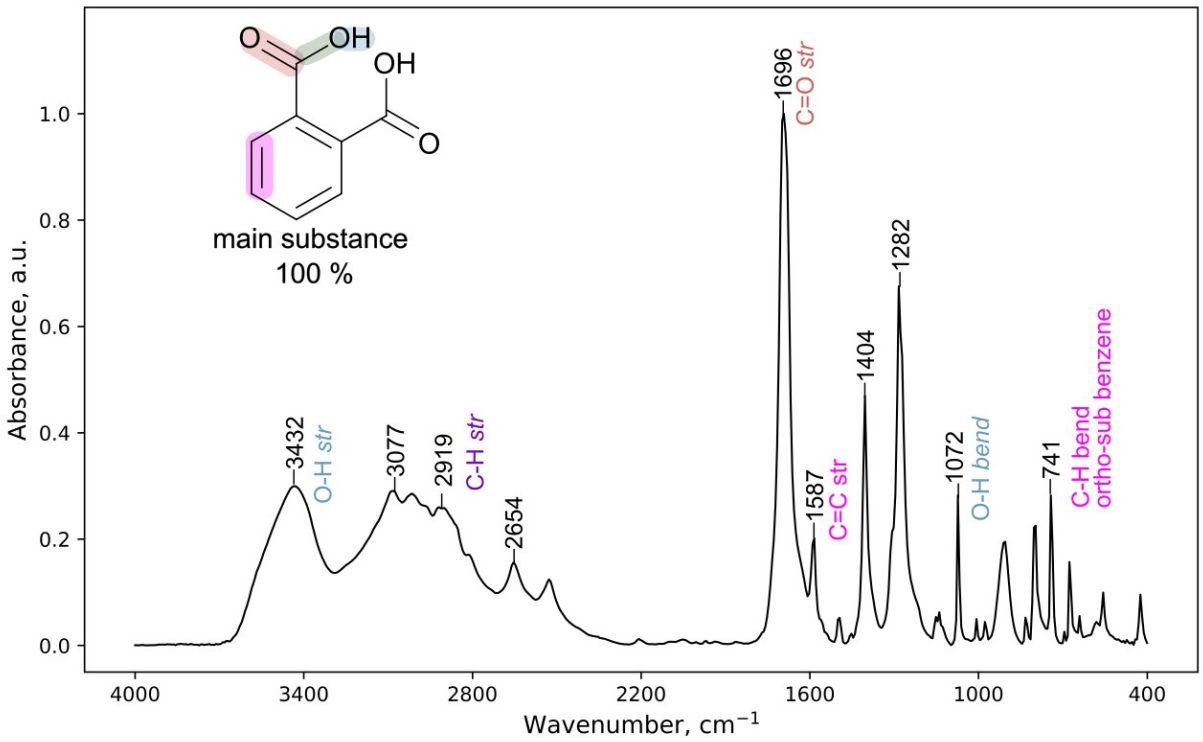

Figure S32. The IR-spectrum of pure phthalic acid (in KBr).

Table S26. Predictions of the model

Labels	Arene	Haloalkane	Alcohol	Carboxylic acid	Purity class
True	+	-	+	+	Pure substance
Predicted	+	+	+	+	Molecule is contaminated with water

Spectra S26: Pure propionic acid (in KBr)

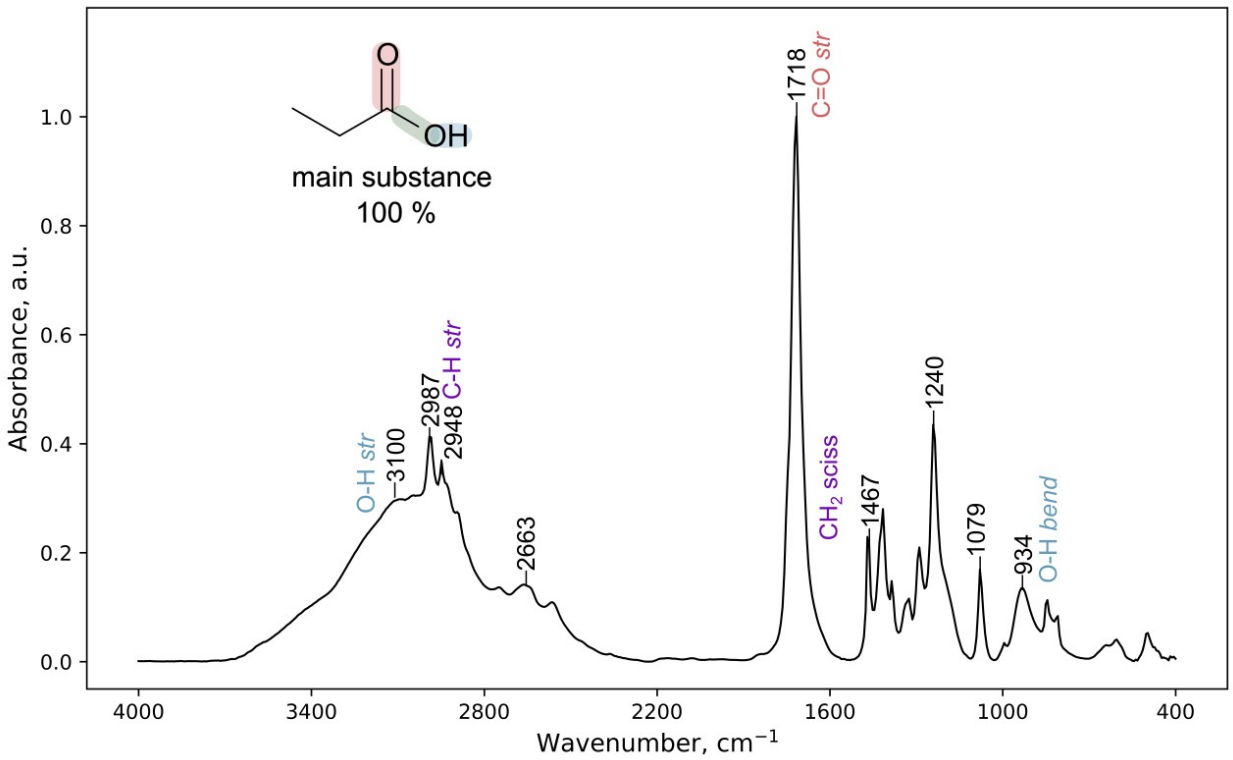

Figure S33. The IR-spectrum of pure propionic acid (in KBr).

Table S27. Predictions of the model

Labels	Alkane	Haloalkane	Alcohol	Carboxylic acid	Purity class
True	+	-	+	+	Pure substance
Predicted	+	+	+	+	Molecule is contaminated with water

S7 Balanced test set prediction

Table S28. Performance summary of trained model in balanced test data for purity prediction task.

Class ID	Class name	Precision	Recall	F_1
0	Pure substance	0.64	0.93	0.76
1	Aldehyde is contaminated with carboxylic acid	1	0.93	0.96
2	Aldehyde is contaminated with alcohol	0.94	1	0.97
3	Alcohol, or carboxylic acid, or aldehyde, or ester, or ether, or ketone, or phenol is contaminated with water	0.85	1	0.92
4	Ether is contaminated with alcohol	0.93	0.81	0.87
5	Amide is contaminated with ester	1	1	1
6	Amide is contaminated with carboxylic acid	1	1	1
7	Aromatic amine with nitro group is contaminated with phenol	1	1	1
8	Amine is contaminated with halide	1	1	1
9	Acyl halide is contaminated with carboxylic acid	1	1	1
10	Halide is contaminated with alcohol	1	0.88	0.93
11	Halide is contaminated with amine	1	0.75	0.86
12	Alcohol is contaminated with aldehyde	1	1	1
13	Alcohol is contaminated with ketone	1	0.94	0.97
14	Ester is contaminated with carboxylic acid or alcohol	0.88	0.94	0.91
15	Nitrile is contaminated with carboxylic acid	1	0.86	0.92
Average		0.95	0.94	0.94

Table S29. Performance summary of trained model in balanced test data for functional group prediction task.

Class ID	Class name	Precision	Recall	F_1	Frequency
0	Alkane	1.00	0.98	0.99	214
1	Alkene	0.94	0.73	0.82	22
2	Arene	0.98	1.00	0.99	56
3	Halide	0.99	0.98	0.98	82
4	Alcohol	0.99	0.98	0.99	203
5	Aldehyde	1.00	1.00	1.00	48
6	Ketone	1.00	0.81	0.89	21
7	Carboxylic acid	1.00	0.95	0.98	66
8	Acyl halide	1.00	1.00	1.00	14
9	Ester	1.00	1.00	1.00	42
10	Ether	1.00	1.00	1.00	68
11	Amine	1.00	0.74	0.85	35
12	Amide	1.00	1.00	1.00	10
13	Nitrile	1.00	1.00	1.00	14
14	Phenol	1.00	0.93	0.96	27
15	Nitro	1.00	1.00	1.00	16
Average		0.99	0.94	0.97	

S8 Imbalanced test set prediction

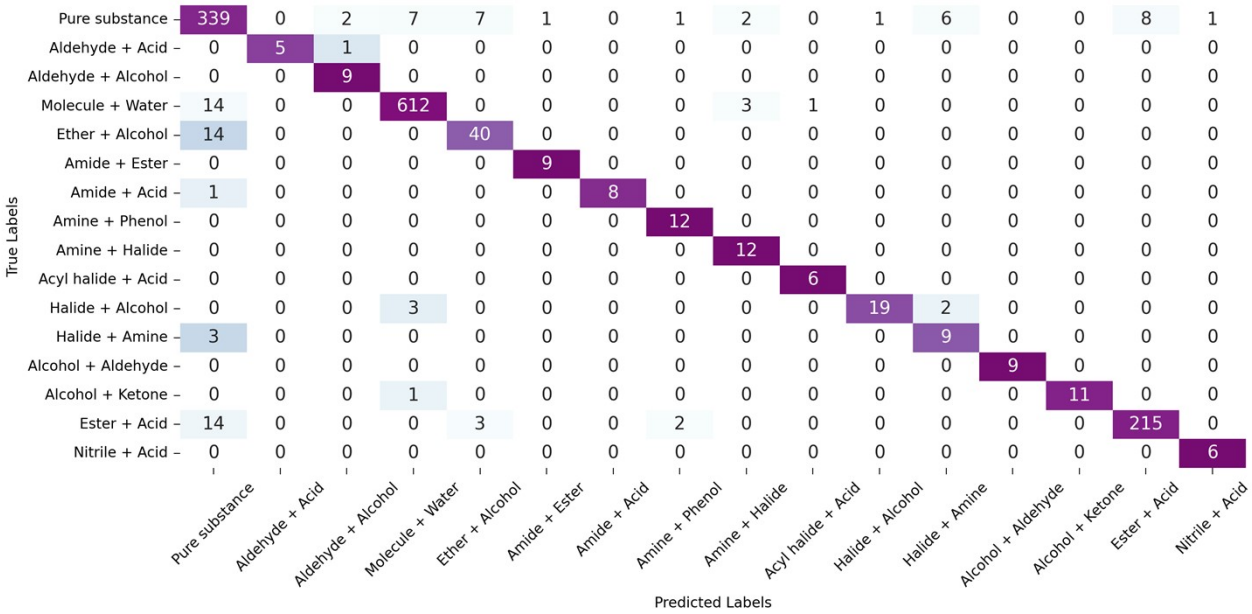

Figure S34. Confusion matrix for purity-predictions task for imbalanced test set.

Table S30. Performance summary of trained model in imbalanced test data for purity prediction task.

Class ID	Class name	Precision	Recall	F_1
0	Pure substance	0.88	0.90	0.89
1	Aldehyde is contaminated with carboxylic acid	1	0.83	0.91
2	Aldehyde is contaminated with alcohol	0.75	1	0.86
3	Alcohol, or carboxylic acid, or aldehyde, or ester, or ether, or ketone, or phenol is contaminated with water	0.98	0.97	0.98
4	Ether is contaminated with alcohol	0.80	0.74	0.77
5	Amide is contaminated with ester	0.90	1	0.95
6	Amide is contaminated with carboxylic acid	1	0.88	0.94
7	Aromatic amine with nitro group is contaminated with phenol	0.8	1	0.89
8	Amine is contaminated with halide	0.71	1	0.83
9	Acyl halide is contaminated with carboxylic acid	0.86	1	0.92
10	Halide is contaminated with alcohol	0.95	0.79	0.86
11	Halide is contaminated with amine	0.53	0.75	0.62
12	Alcohol is contaminated with aldehyde	1	1	1
13	Alcohol is contaminated with ketone	1	0.92	0.96
14	Ester is contaminated with carboxylic acid or alcohol	0.96	0.92	0.94
15	Nitrile is contaminated with carboxylic acid	0.86	1	0.92
Average		0.87	0.92	0.89

Table S31. Performance summary of trained model in imbalanced test data for functional group prediction task.

Class ID	Class name	Precision	Recall	F_1	Frequency
0	Alkane	0.98	1.00	0.99	1232
1	Alkene	0.94	0.91	0.92	235
2	Arene	0.99	0.98	0.98	583
3	Halide	0.92	0.97	0.94	368
4	Alcohol	0.97	0.95	0.96	735
5	Aldehyde	0.95	1.00	0.97	57
6	Ketone	0.98	0.93	0.96	116
7	Carboxylic acid	0.94	0.91	0.93	203
8	Acyl halide	0.79	1.00	0.88	11
9	Ester	0.99	0.99	0.99	584
10	Ether	0.98	0.99	0.99	783
11	Amine	0.83	0.86	0.84	98
12	Amide	0.70	1.00	0.82	14
13	Nitrile	0.92	0.71	0.80	17
14	Phenol	0.99	0.91	0.95	156
15	Nitro	0.90	1.00	0.95	18
Average		0.92	0.94	0.93	

S9 Comparison of real and generated spectra

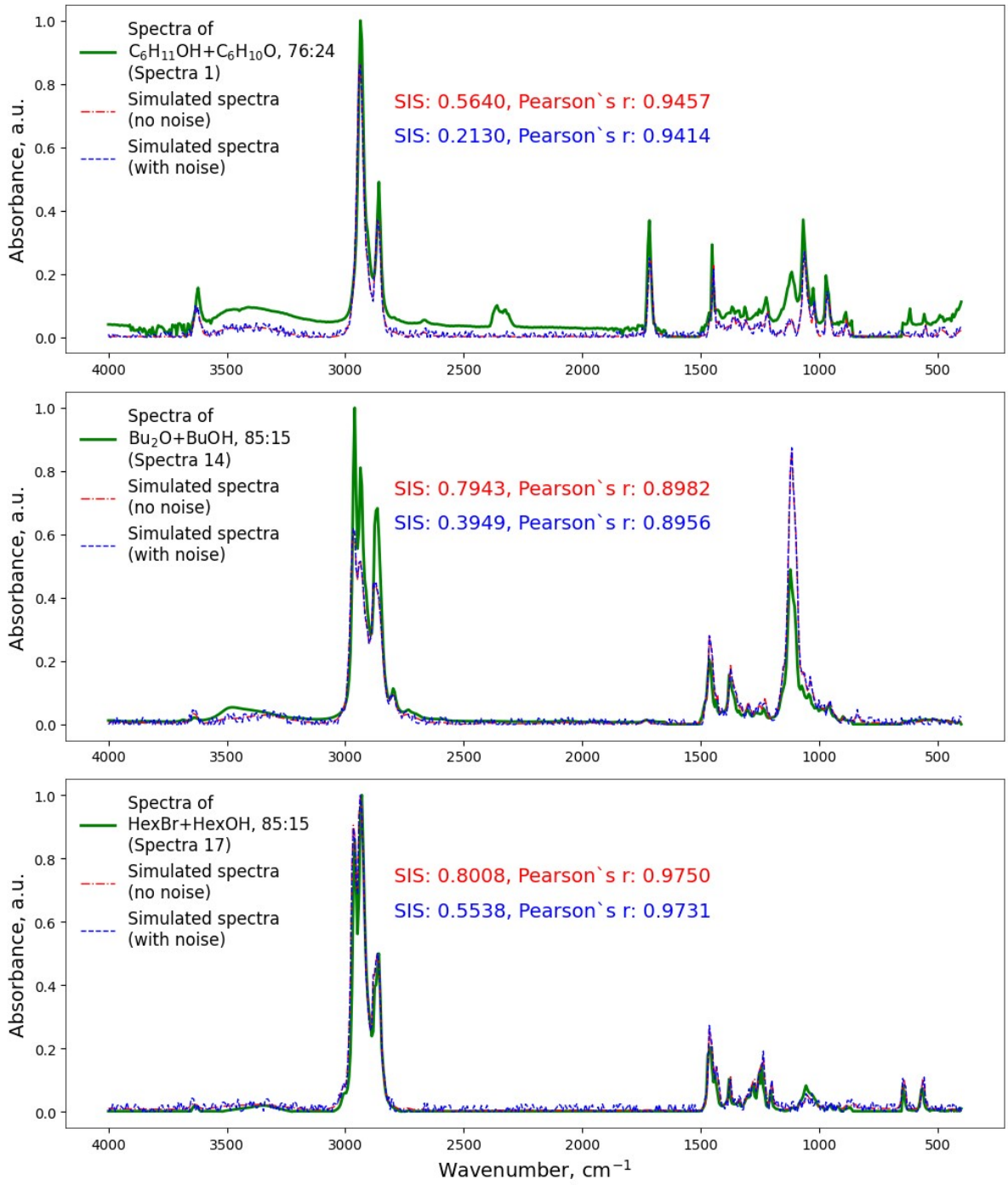

Figure S35. Part 1. Evaluation of the effectiveness of the approach to spectrum generation using linear combination: generated infrared spectra of mixture with added noise (blue line) and without noise added (red line) and actual IR spectra of the corresponding mixtures (green line).

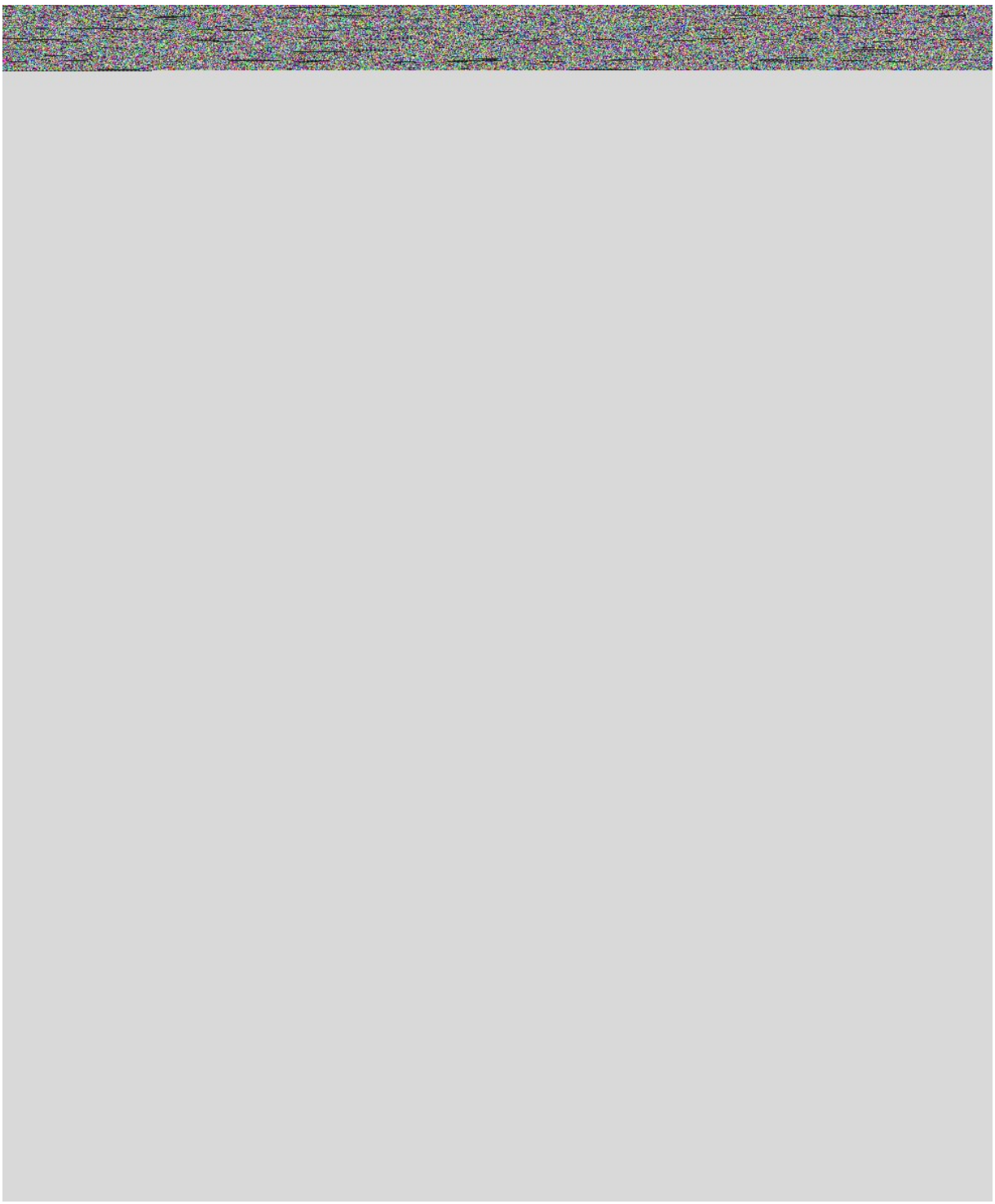

Figure S35. Part 2. Evaluation of the effectiveness of the approach to spectrum generation using linear combination: generated infrared spectra of mixture with added noise (blue line) and without noise added (red line) and actual IR spectra of the corresponding mixtures (green line).

S10 References

1. G. Jung, S. G. Jung and J. M. Cole, Automatic materials characterization from infrared spectra using convolutional neural networks, *Chem. Sci.*, 2023, **14**, 3600-3609.

2. A. Angulo, L. Yang, E. S. Aydil and M. A. Modestino, Machine learning enhanced spectroscopic analysis: towards autonomous chemical mixture characterization for rapid process optimization, *Digital Discovery*, 2022, **1**, 35-44.