Chemical Science

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE

View Article Online
View Journal

Cite this: DOI: 10.1039/d6sc01583b

All publication charges for this article have been paid for by the Royal Society of Chemistry

Automatic identification of compounds in molecular mixtures from liquid-phase infrared spectra

Yannah J. U. Melle, [ab] Thanh Nguyen, [c] Jeffrey Lopez *[c] and Daniel Schwalbe-Koda *[a]

Received 24th February 2026
Accepted 26th May 2026

DOI: 10.1039/d6sc01583b

rsc.li/chemical-science

Interpreting spectroscopy data is a critical bottleneck in automating chemical research and industrial characterization. Particularly within infrared (IR) spectroscopy, identifying compounds in complex, liquid-phase chemical mixtures largely relies on expert knowledge, as variable peak assignment, broadening, and shifts hinder data-driven methods. Here, we show that an algorithmic approach can identify components in both simulated and experimental mixture spectra with high accuracy despite nonlinearities in liquid-phase IR data. The method is comprehensively benchmarked with a dataset of over 44 000 simulated liquid-phase IR spectra for mixtures and achieves up to 90% accuracy in identifying molecular components across a dataset of binary and ternary liquid mixtures. Our strategy is robust to perturbation of spectra, and its accuracy is capped by near-identical liquid-phase IR spectra that limit the resolution of chemical identification, imposing theoretical limits on achieving perfect accuracy in structure identification. Finally, we apply the method to automatically interpret IR spectra in experimental settings, correctly identifying the components of nearly all samples within a blind study. This work provides tools and data to advance automated chemical laboratories through algorithmic interpretation of liquid-phase IR spectra of mixtures.

1 Introduction

Identifying the constituents of molecular liquids is essential for the study and design of chemical formulations across applications, from biomedical and pharmaceutical research to energy materials.[1–3] In principle, sufficient and high-resolution characterization data can unambiguously determine all components of a liquid mixture and elucidate their intermolecular interactions. Infrared (IR) spectroscopy is one of the core tools used in chemistry to identify unknown compounds and functional groups in liquid-phase mixtures.[4,5] The technique is fast, nondestructive, and can be paired with complementary measurements such as nuclear magnetic resonance (NMR) and mass spectrometry (MS) to monitor reactions, analyze materials, and identify products in chemical reactions and chemical processes, including under *operando* conditions.

Many experimental IR datasets for pure gas- and liquid-phase spectra are available in standardized digital formats through sources such as the NIST Chemistry Webbook,[6] the NIST Quantitative Infrared Database,[7] and the Japanese AIST Spectral Database for Organic Compounds (SDBS).[8] While these datasets provide reference spectra, their limited coverage of chemical space and restricted accessibility constrain their use for large-scale computational analysis and data-driven modeling. Consequently, compound identification in molecular mixtures using IR spectra relies on expert-driven workflows, including identifying and interpreting spectral signatures using comprehensive functional group tables, performing simplified density functional theory calculations, and conducting pattern-guided searches over a large space of spectra.[5] As chemical laboratory automation and liquid-phase handling continue to advance, the limited ability to automate the interpretation of characterization data becomes the main bottleneck in chemical analysis.

Automation challenges go beyond data availability to the fundamental physics of vibrational spectroscopy. Spectral peak positions and intensities are sensitive to local thermodynamic conditions, intra- and intermolecular interactions, and anharmonic phenomena such as hot bands, overtones, and vibrational coupling.[5] As a result, structure identification from IR spectra is better posed in the gas phase than in the liquid phase. Highly reproducible measurements and sharp and distinct peaks characteristic of the gas phase allow near-unambiguous characterization of compounds. In contrast, liquid phase spectra reflect a broader distribution of molecular

[a]Department of Materials Science and Engineering, University of California, Los Angeles, CA, USA. E-mail: dskoda@ucla.edu

[b]Department of Chemistry and Biochemistry, University of California, Los Angeles, CA, USA

[c]Department of Chemical and Biological Engineering, Northwestern University, Evanston, IL, USA. E-mail: jlopez@northwestern.edu

environments and geometries, giving normal modes a range of frequencies whose overlap broadens and blurs vibrational features. Liquid mixtures further encode the molecular environment directly in the IR spectrum: they exhibit peak shifts and broadening relative to their gas-phase counterparts, and a mixture spectrum cannot always be modeled well as a simple weighted sum of its constituents. Developing data and methods is therefore essential for automating chemical characterization across phases and components.

Historically, chemometric methods based on partial least squares (PLS) have been widely used for predictive modeling of liquid-phase mixtures from IR spectra across pharmaceutical, fuel, and food applications,[9–12] with numerous small-scale studies demonstrating effective property prediction or classification.[13–15] However, their performance depends strongly on spectral preprocessing, reference measurements, and the chemical scope of calibration data, thereby restricting applicability to a very limited, case-by-case basis.[16–18] Furthermore, few large-scale, chemically diverse, liquid-phase spectroscopic datasets have been evaluated with PLS to test these limitations. Partial least squares is used to learn statistical covariances between spectra and properties rather than to enforce a physically constrained mixture model. As a result, while PLS remains useful for liquid-phase spectroscopic analysis, it functions as a pattern recognition-based regression method as opposed to a molecular identification model capable of explicitly representing mixture composition and component identities. Similarly, other least squares methods have been used to deconvolute mixture IR spectra, but only for narrow chemical domains.[19–24]

More recently, gas-phase synthetic spectral datasets and machine learning (ML) have been developed to perform structure elucidation from experimental spectra.[25–28] Such tools have led to establishing spectral-to-structure maps in the gas phase and have supported limited structure prediction from pure-component IR spectra.[25,26,29–31] Additionally, data-driven models have been shown to successfully predict the components and concentrations of gas-phase mixtures from spectra with moderate to high accuracy,[32] with only limited demonstrations extending to liquid-phase synthetic and experimental data.[33] Synthetic spectral generation has also advanced through neural network models capable of predicting infrared spectra from molecular structure in the gas phase to support inverse spectral analysis.[34–36] Despite these advances, accurate molecular identification from liquid-phase spectra remains insufficient for reliable use in practice.

This context highlights two knowledge gaps. On the data side, broader and standardized datasets are needed to understand and evaluate the performance of liquid-phase IR analysis and to enable training newer ML models in the high-data regime. On the modeling side, the limits of traditional algorithms for deconvolving liquid-phase IR data remain unclear beyond narrow chemical spaces, leaving even baseline performance unestablished. In this work, we develop a dataset containing over 44 000 simulated, liquid-phase IR spectra to quantify peak shifting in chemical mixtures, and we develop algorithms to automate chemical identification from mixture spectra. We show that the non-negative least squares (NNLS) algorithm correctly identifies compounds in both gas- and liquid-phase simulated and experimental IR mixture spectra. Interestingly, the NNLS algorithm does so with high accuracy despite nonlinearities in mixing. This result remains robust to spectral perturbations, including noise and artificial peak shifts applied to the pure-component spectra. Component identification reaches 100% in gas-phase mixtures but is limited in liquid-phase IR, with an optimistic performance of up to 90% accurate identification. We show that identification accuracy is bounded not by algorithmic performance, but by near-identical liquid-phase IR spectra that yield degenerate solutions to mixture deconvolution. By characterizing these degenerate solutions, we indicate a potential theoretical limit to spectral deconvolution that may require additional measurements or information to accomplish component identification in the liquid phase. We show how our strategy can be useful in practice to deconvolve mixture IR spectra from multiple experimental samples. In a small-scale blind study where experimental sample identities were withheld during analysis, our algorithm correctly identified the compounds in nearly all experimental samples, demonstrating its applicability in practical settings. In addition to the dataset and quantitative baselines, we discuss potential limitations to the scalability of characterization and molecular identification in automated laboratories.

2 Results

2.1 Rationalizing peak shifts in simulated gas- and liquid-phase IR spectra

To quantify and rationalize the magnitude of peak shifts in simulated IR spectra of gas- and liquid-phase molecules, we created a dataset with 8880 pure gas-phase and 8550 pure liquid-phase spectra using classical simulations (Section 5.1). Examples of simulated gas- and liquid-phase spectra for 3-(dihydroxymethyl)piperidine (molecule A) and 4(5)-vinylimidazole (molecule B) are shown in Fig. 1a. When comparing the gas- and liquid-phase spectra across both pure components and mixtures, several features can be observed. Liquid-phase spectra exhibit spectral broadening relative to the sharp peaks observed in the gas phase in the fingerprint region between 0 and 2000 cm^{-1} for both molecules. For the A + B mixture, the IR spectrum simulated using equal molar amounts of each molecule is not equivalent to the sum of their pure-component liquid-phase spectra, as shown in Fig. 1b. This shows how intermolecular interactions in the mixture can obscure clear peak assignments of individual components in complex mixtures and give rise to nonlinear mixing behavior characteristic of the liquid phase.

In addition to these mixture effects, liquid-phase spectra also show clear shifts in peak positions relative to their gas-phase counterparts. Fig. 1c depicts how the gas-phase peak around 4100 cm^{-1} for molecule A red-shifts when simulated at the liquid phase. To systematically quantify the magnitude and sign of peak shifts, we normalized each molecule's gas- and liquid-phase spectra to unit area and computed the difference

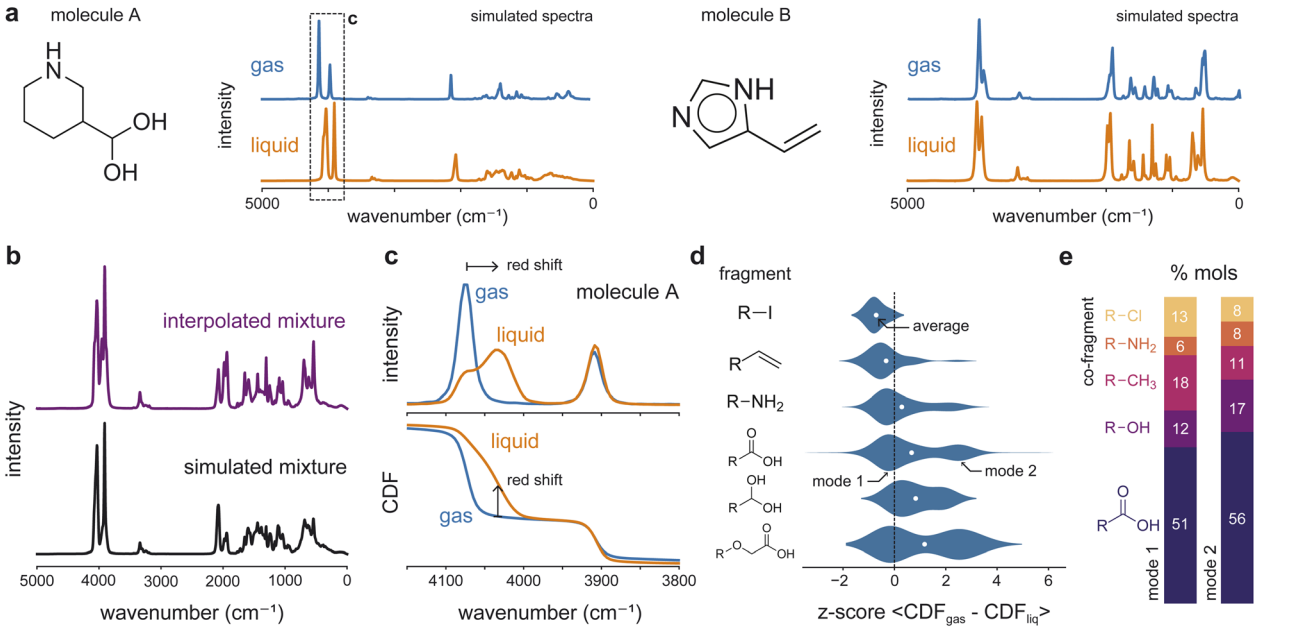

Fig. 1 MD-generated gas- and liquid-phase pure and mixture IR spectra and cumulative intensity difference metric analysis. (a) Simulated gas and liquid spectra for 3-(dihydroxymethyl)piperidine (molecule A) and 4(5)-vinylimidazole (molecule B). (b) Simulated mixture spectrum of a liquid-phase mixture of molecule A and B (bottom) and the equal weight linear combination (average) of the two molecular spectra (top). The linear sum is not equivalent to the true simulated mixture. (c) Raw and cumulative intensities of molecule A's gas- and liquid-phase spectra between 4150–3800 cm^{-1}, illustrating the gas-to-liquid peak shift and broadening. (d) Fragment-driven differences between gas and liquid spectra. Distributions and per-core means of fragment-level z-scores for the cumulative distribution function (CDF) between gas- and liquid-phase spectra. Molecules are decomposed into a Murko-scaffold "core" and their largest remaining fragment. Average CDF values are standardized (z-score) within each core, removing core-specific effects and isolating fragment-dependent contributions. The white dots represent the average of each distribution. (e) Mode-specific relative composition of the most common co-fragments for molecules containing a carboxylic acid fragment (O=CO). Molecules are assigned to one of two modes by fitting a two-component Gaussian to their per-core z-score cumulative intensity differences.

between their cumulative distribution functions (CDFs), where the sign measures the direction of the shift (Section 5.5.1). Using the average distance between CDFs, we then analyzed the dataset for trends in peak shifts and broadenings between both spectra. To probe the structural dependence of these phase-dependent spectral shifts, each molecule was decomposed into its Murcko scaffold "core" and its largest remaining fragment. Fig. 1d highlights how the average CDF difference captures fragment-dependent peak shifts that are consistent with known phase-dependent intermolecular interaction differences. Molecules whose largest fragments are hydrocarbons or halogens exhibit smaller average CDF differences between their gas- and liquid-phase spectra, consistent with their weaker intermolecular interactions and lack of hydrogen bonding. In contrast, fragments such as amines, carboxylic acids, and alcohols show substantial peak shifts and broadening from the gas phase to the liquid phase.

Notably, for some fragments, the distribution of z-scores of the mean CDF differences is bimodal, indicating two subpopulations with different interaction strengths. Decomposing the molecules in these modes into their full fragment compositions reveals distinct compositional differences between the modes, as illustrated in Fig. 1e. Molecules in mode 1 contain more co-fragments that contribute weakly to gas–liquid spectral differences; for instance, methyl and chlorine fragments co-appear more frequently in mode 1 than in mode 2, explaining their lower average shifts. Conversely, molecules in mode 2 contain more carboxyl, alcohol, and amine co-fragments, which account for their higher average peak shifts. This demonstrates that peak shifts in simulated liquid-phase IR spectra can be rationalized according to known intermolecular interaction trends, while the emergence and structure of their distributions are not predicted by chemical intuition alone.

2.2 Robustness of deconvolution of simulated IR spectra of mixtures to small shifts

Given that IR spectra of mixtures often deviate from linear interpolations of pure-component spectra due to liquid-phase peak shifts (Fig. 1), decomposing molecular mixtures into individual components is typically deemed non-trivial. We therefore tested whether linear algorithmic approaches can succeed in predicting molecular components from an unknown IR spectrum even in the presence of non-linearities in the data. Fig. 2a illustrates the main method tested in this work, where components of each mixture spectrum are predicted using linear algorithms and a "basis set" of pure-component spectra. Using the gas-phase pure-component spectra as the basis set, the components of 30 000 gas-phase and 26 996 liquid-phase mixtures were predicted. Using the liquid-phase basis set, molecular constituents of 27 657 two-component and 7985 three-component liquid-phase mixtures were predicted.

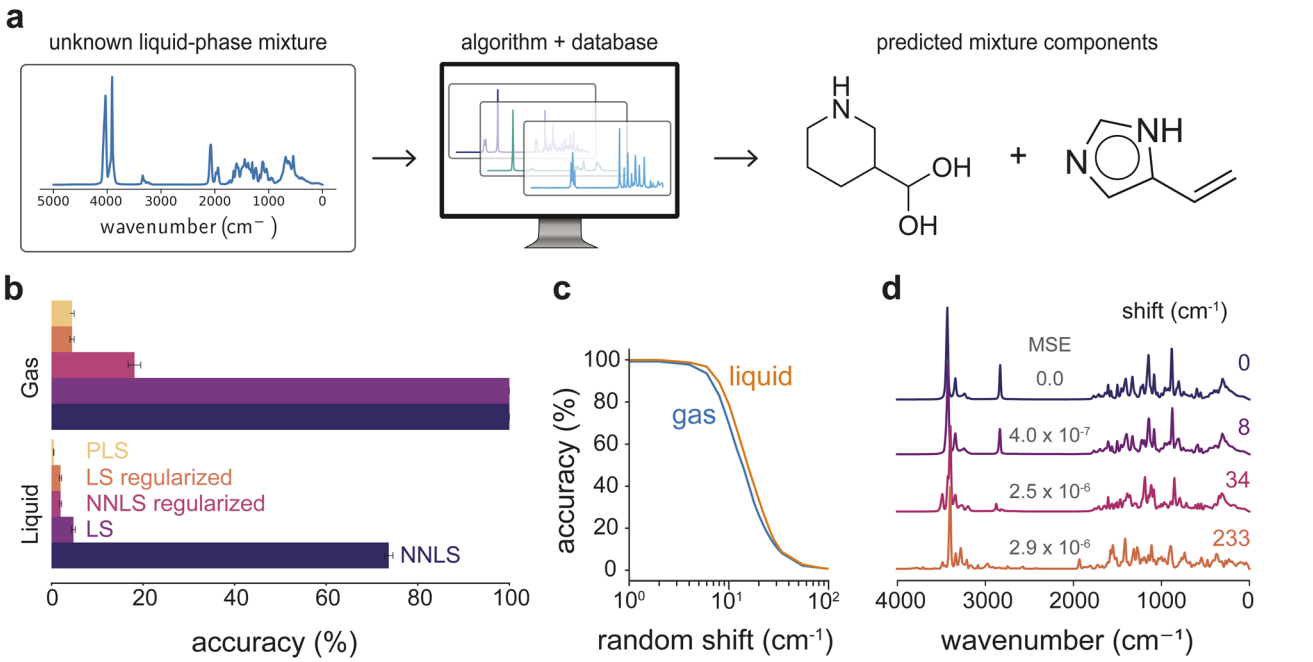

Fig. 2 Identification accuracies of unknown components of two-component liquid-phase mixtures from simulated IR spectra, using MD-generated simulated pure-component IR spectra. (a) Workflow to identify unknown mixture components from a liquid-phase mixture spectrum. Given an unknown spectrum and our database of pure-component spectra, an algorithm is used to predict mixture components. (b) Prediction accuracies using NNLS, LS, and regularized variants to identify two-component mixtures. Gas- and liquid-phase mixtures were predicted using both gas and liquid pure-spectra basis sets, with NNLS achieving the highest liquid-phase accuracy. (c) Prediction accuracies using NNLS for gas and liquid phase mixtures as a function of spectral peak shifts. (d) Examples of spectra with increasing peak shift magnitudes (in cm^{-1}).

Predicted mixture components were obtained by ranking the coefficients by magnitude and selecting the top two (binary) or three (ternary) values, matching the known number of components in each mixture.

Each pure-component basis set included spectra for all molecules present in the mixtures, along with an equal number of spectra from molecules not appearing in any mixture, selected at random. Each dataset of mixture-basis set combinations was evaluated eight times, using a different random selection of additional molecular spectra in each iteration, shown in Fig. 2b. Gas-phase mixtures are predicted with the highest accuracy, reaching up to 100% accuracy with zero standard deviation across prediction runs. This result reflects the linear additivity of gas-phase spectra: sharp and distinct peaks arising from non-interacting molecules make each spectrum perfectly distinguishable, allowing linear unmixing algorithms to recover the exact components under a linear mixing assumption. For liquid-phase mixtures, using pure gas-phase spectra as the basis set yields a prediction accuracy of only 15.4% (Fig. S1), indicating that gas-phase spectral signatures differ too substantially from their liquid-phase counterparts to serve as effective reference data for deconvolving condensed-phase mixtures. This limitation persists even when the gas-phase data are artificially broadened (Fig. S1). When using a basis set of pure liquid-phase spectra that captures liquid-phase spectral effects, NNLS identifies liquid-phase mixture components with 73.6% accuracy, outperforming all other algorithms.

Given the reasonable success of NNLS in identifying mixture components despite peak shifts arising from gas-to-liquid phase mixing behavior, we quantified the robustness of this algorithm by deliberately introducing random shifts into liquid-phase pure-component spectra used to construct mixtures (Section S2.2). Fig. 2c shows that liquid-phase mixtures tolerate slightly larger frequency shifts than gas-phase mixtures. While any peak shift is expected to reduce identification accuracy, the greater tolerance observed for liquid-phase deconvolution under moderate wavenumber shifts is consistent with broader liquid-phase spectral features, which make them less sensitive to small positional shifts. In contrast, sharper gas-phase peaks would be expected to be more vulnerable to small positional perturbations that disrupt features essential for accurate deconvolution. However, gas- and liquid-phase identification accuracy degrades at similar rates, with accuracy remaining above 80% for random peak shifts up to 8 cm^{-1}. While both phases show a rapid decline in accuracy beyond approximately 15–20 cm^{-1}, gas-phase identification accuracy remains consistently lower than that of the liquid phase across all shifts, reflecting greater sensitivity to deviations in peak positions. Fig. 2d illustrates how small shifts still preserve many of the original spectral features, while larger shifts produce pronounced spectral changes that coincide with the observed loss in identification accuracy.

2.3 Accuracy limits for deconvolution of IR spectra of mixtures with linear methods

Considering that NNLS can deconvolve spectra of liquid-phase mixtures with meaningful accuracy (Fig. 2b), we assessed the

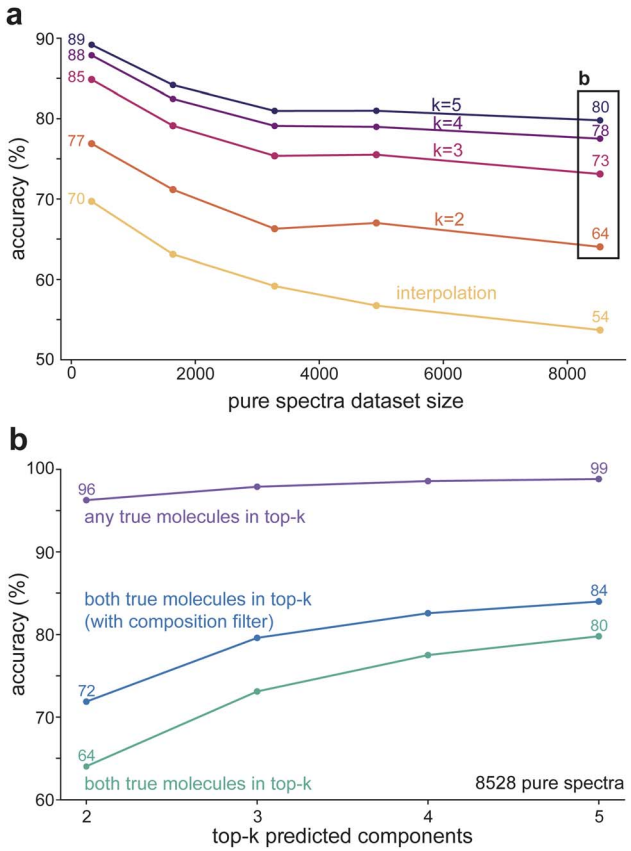

Fig. 3 Two-component liquid-phase mixture identification accuracies obtained with NNLS as a function of pure liquid-phase spectra dataset size and prediction criteria. (a) Identification accuracies as a function of pure liquid-phase basis set size. Accuracy is reported for identifying all true components from the largest $k = 2$–5 NNLS coefficients compared with the interpolation baseline, which selects the top two coefficients from a brute-force convex interpolation over all spectrum pairs. For $k = 2$, NNLS achieves higher identification accuracy than the interpolation baseline across all dataset sizes. (b) Identification accuracies as the prediction criterion increases from $k = 2$–10 are evaluated by (i) requiring all true components to appear within the top k, (ii) requiring at least one (any) true component appears within the top k coefficients, and (iii) applying an atom-count filter that restricts candidate components whose combined atomic compositions match the mixture's atom count (as would be available from mass spectrometry (MS)).

limits of structure identification with this method. First, correctly identifying all components of a mixture is the strictest measure of accuracy for an algorithm. In practice, however, identifying a set of candidate molecules with similar spectral features (and thus structural motifs) can already support indirect peak assignment and interpretation of spectral data. Second, linear algorithms rely on a spectral database (Fig. 2a), whose size can vary and constrain identification accuracy. To quantify the limits of accuracy under relaxed criteria and dataset sizes, we evaluated two accuracy metrics: (1) both true molecules are within the top-k molecules predicted by the algorithm; or (2) any of the true mixture components is within the top-k molecules predicted by the algorithm. Fig. 3a compares how the accuracy of NNLS varies as a function of k and the dataset size under criterion (1). At the smallest dataset size (*i.e.*, the minimum number of pure components needed to fully identify all mixtures), NNLS correctly predicts 77% of mixtures when $k = 2$, corresponding to exact recovery of the mixture. Accuracy increases to 89% when both true molecules are within the top-5 candidates. When the dataset of pure spectra is increased to the maximum size in this study of 8528 spectra (26 times larger than the minimum dataset size needed), NNLS identifies both true molecules with 64% accuracy when $k = 2$ and with 80% accuracy when $k = 5$. The algorithm's performance first plateaus before declining to lower accuracies, indicating the increasing difficulty of identifying the correct components as the candidate space grows. On the other hand, the "brute-force" interpolation approach, which evaluates all molecular pairs independently (Section 5.3), resulted in a monotonic decrease in accuracy from 70% to 54%.

At the largest dataset size, we examined the NNLS accuracy under a range of evaluation strategies to characterize achievable performance. Incorporating an atom-type filter during evaluation substantially improved performance at the largest pure dataset size, increasing the top-2 accuracy from 64% to 72% (Fig. 3b). Computationally, this filter prioritizes pairs whose elements are known to be present in the mixture, narrowing the pool of candidate molecules (Supplemental methods S2.3). Experimentally, the presence or absence of atoms that frequently confound NNLS (such as I *vs.* Cl) is often known. In cases of true unknown identification, elemental analysis can readily provide the required information to distinguish between two candidate molecules. To test whether the observed accuracy limit reflects NNLS failing to recover the true components entirely, we assessed whether either of the correct molecules is recovered among top-k candidates predicted by the algorithm. In this case, Fig. 3b shows that the accuracy of the algorithm is nearly perfect, at 99.1% for $k = 5$, demonstrating that NNLS almost always identifies at least one of the true components from the mixture IR spectrum. Practically, this accuracy is sufficient for dramatic acceleration of interpretation of IR spectra compared to traditional manual functional group analysis. The small remaining accuracy gap suggests that further gains may be limited not only by the algorithm, but also by the underlying spectral data.

2.4 Characterizing misidentification profiles and spectral ambiguity

To explain the accuracy limits in Fig. 3b, we analyzed which molecules are often misidentified by NNLS. We categorized false positive misidentifications according to classes depicted in Fig. 4a. False positives tend to occur due to (1) structural isomers, *i.e.*, molecules with the same chemical formula but different atomic arrangements; (2) single atom substitutions, where one element is replaced by another; and (3) molecules that differ only in total carbon count where the false positive has one more or one fewer carbon than the true positive. Together, these cases explain about 35% of all false positives predicted by NNLS (Fig. 4b). The remaining 65% of false positives fall into the "other" category and involve molecules that share similar

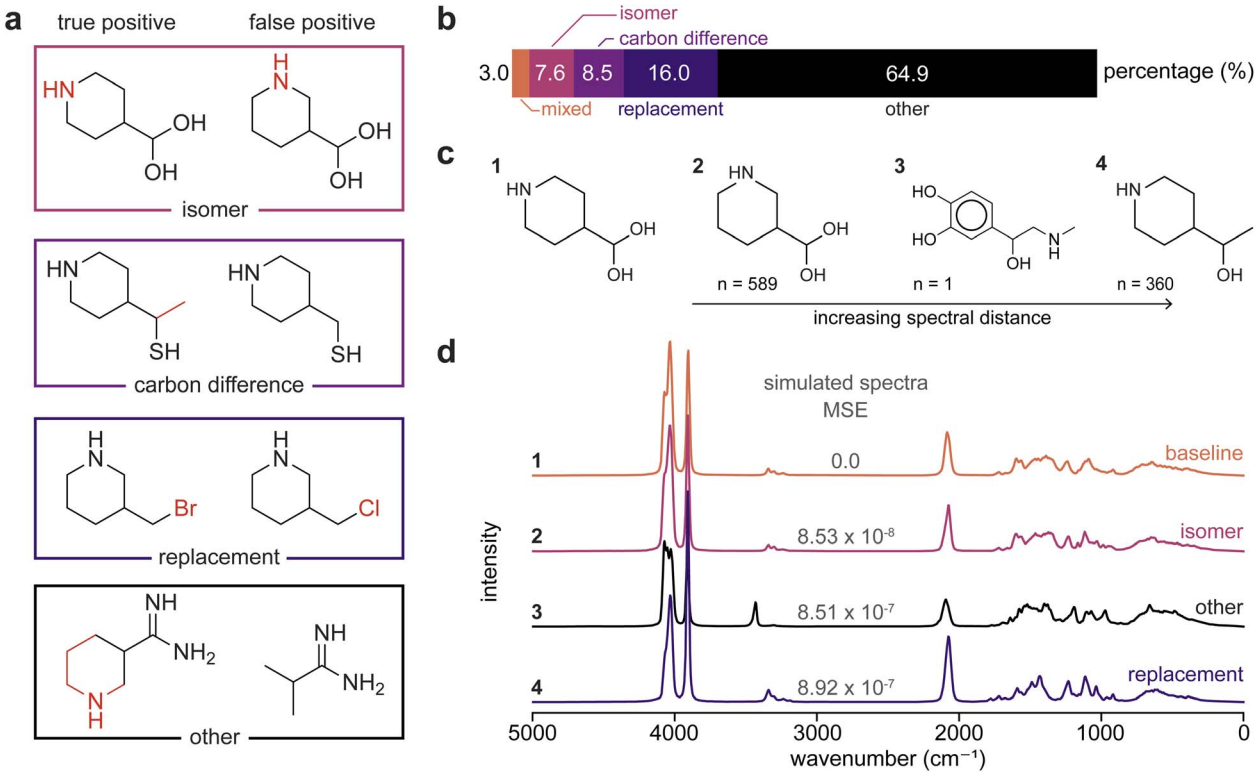

Fig. 4 Misidentification profiles for predicting all components in two-component liquid-phase mixtures using the top $k = 2$ NNLS coefficients. (a) True *vs.* falsely predicted components for characteristic misidentification examples: (i) a predicted component differs by the addition or removal of a carbon relative to the true component; (ii) a predicted component is an isomer of the true component; (iii) a predicted component differs by one-atom substitution; and (iv) misidentification not covered by (i)–(iii). (b) Percentage of two-component mixtures that were misidentified when using the top $k = 2$ NNLS coefficients to identify both components, aggregated across pure-component dataset sizes. "Mixed" indicates mixtures where multiple misidentification categories (carbon difference, isomer, substitution) apply to the true-false component pair. (c) Molecules closest to the true component, 4-(dihydroxymethyl)piperidine, by spectral distance (MSE). The spectral similarity among these molecules makes them ambiguous to the NNLS algorithm that minimizes squared error (MSE-equivalent). The variable "*n*" indicates the number of times the nearest neighbor molecule was incorrectly predicted instead of the true component, across all evaluated mixtures and pure-spectra dataset sizes. (d) Spectra of the molecules in (c) and their spectral distance (MSE) from the true component spectrum (baseline), with corresponding misclassification profile labels.

molecular characteristics despite being structurally dissimilar. One such example is 2-methylpropanimidamide and piperidine-3-carboximidamide (Fig. 4a), two carboximidamides that are structurally distinct yet have very similar simulated IR spectra. The common C=N bond and amine functional groups result in similar major features of the spectra, while the characteristic peaks of the piperidine ring overlap with the broad amine peak (3200–3300 cm^{-1}) and C–H/C–C peaks. These overlaps result in a mean squared error (MSE) difference of 7.3 $\times 10^{-7}$ cm^{-1} between the true and false positive spectra.

Fig. 4c and d further illustrate this spectral ambiguity by comparing the first three nearest spectral neighbors of a representative molecule (4-(dihydroxymethyl)piperidine), identified using the MSE between liquid-phase IR spectra. The resulting MSE values, on the order of 10^{-8} to 10^{-7}, are much smaller compared to the average MSE between two arbitrary molecules, 7.16 $\times 10^{-6}$ cm^{-1}. These small errors illustrate how distinct molecular structures can produce nearly indistinguishable liquid-phase IR spectra, leading NNLS to assign false-positive identifications. Fig. 4d emphasizes the degree of spectral similarity among these nearest neighbors, particularly the "isomer" and "replacement" misidentification spectra, where differences are primarily limited to changes in relative peak intensities. When small structural variations make molecules effectively indistinguishable in liquid-phase IR spectroscopy, misidentification reflects limitations in the discriminative information available from spectra rather than limitations of the prediction algorithm itself. This suggests that the accuracy limits reported in Fig. 3 arise from intrinsic constraints of linear mixture deconvolution applied to liquid-phase IR spectra. Thus, misidentification of liquid-phase IR data must be interpreted carefully, even though accuracies approaching 90% already enable reasonable automation of molecular structure deconvolution.

2.5 Interpreting NNLS coefficients to evaluate component contributions and identify *n*-component mixtures

Although the analysis thus far has been performed on binary mixtures, the methods are, in principle, applicable to mixtures containing an arbitrary number of components. Using 7,985 simulated spectra of three-component liquid-phase mixtures, the NNLS algorithm correctly identified all three components

within the top $k = 3$ candidates with an accuracy of 73% (Fig. S2). Despite the presence of similar misidentification profiles discussed in Section 2.4, the identification accuracy of NNLS for these mixtures remains comparable to the results shown in Fig. 3 for two-component mixtures. This suggests that the algorithm can be applied for automated identification pipelines without further modification.

When the number of components in a mixture is unknown, the coefficients obtained from NNLS deconvolution can be used to infer both the number of components present and their relative contributions to the resulting mixture IR spectrum. Starting from a spectrum with zero intensity, pure component spectra are added sequentially in decreasing NNLS coefficient order, each weighted by its predicted coefficient. Then, the cumulative variance of the total spectrum that is explained after each added component hints at the likely number of true components (Fig. 5a). Once all real components have been included, the explained variance fraction plateaus and additional components yield minimal improvement. Fig. 5b exemplifies this behavior. Although at least six molecules have non-zero coefficients, the cumulative explained variance saturates after the third component. Additional metrics described in the SI (Section S1.4) suggest that, within reasonable signal-to-noise thresholds, n-component mixtures may be analyzed even when the number n is unknown *a priori*.

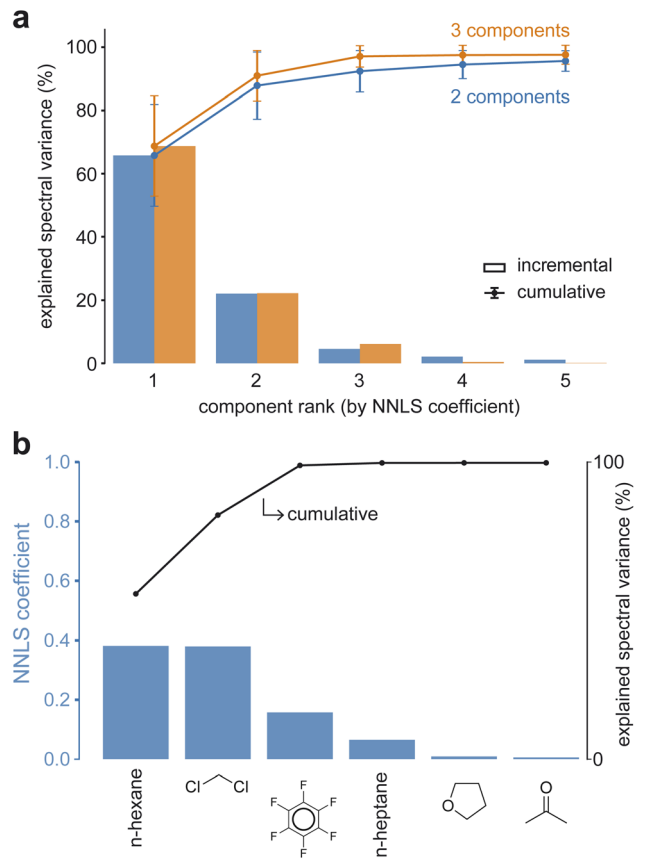

Fig. 5 Fraction of mixture spectra explained by components ranked by decreasing NNLS coefficient for two- and three-component liquid mixtures. (a) Average cumulative and incremental percentage of the explained spectrum across all two- and three-component liquid-phase mixtures, ranked in decreasing NNLS coefficient order, as a function of basis set size with associated errors. (b) NNLS coefficients for the top six components and the percentage of the spectrum explained for a three-component liquid mixture. The plateau in the explained-spectra curve indicates the likely number of components in the mixture, as additional component spectra weighted by their coefficients do not further contribute to explaining the mixture.

2.6 Structure identification for experimental mixture IR spectra

To demonstrate the applicability of the present analysis beyond simulated spectra, we performed a blind study to predict both the identities and number of components in experimentally prepared two- and three-component liquid-phase mixtures. Our experimental team prepared nine different mixtures and collected experimental IR spectra for the pure compounds and their mixtures, composed of common laboratory solvents, as shown in Fig. 6a. Then, the computational team ranked and predicted the components of the mixtures using the methods discussed in previous sections. After unblinding the results, we noticed that the NNLS approach accurately identified all true components within the top $k = 2$ or $k = 3$, effectively deconvolving the real mixture spectra when reference pure-component spectra are available. Fig. 6b illustrates predictions for three experimental mixtures along with the NNLS coefficients used to infer how many components are present in each mixture. Fig. 6c showcases how the reconstructed IR spectra change as components are added sequentially, weighted by their NNLS coefficient. For example, in the three-component mixtures (left and middle panels in Fig. 6b and c), a peak present in the mixture spectrum appears only after the third component spectrum is added to the cumulative reconstructed spectrum (green box). On the other hand, in the two-component mixture case (right), characteristic mixture peaks are already reproduced after the first two component spectra are added. The second mixture in Fig. 6b and c also exhibits a peak shift in the C=O stretch near 1600 cm^{-1} between the cumulatively weighted spectrum including all three components and the experimental mixture (blue box), indicating that the third component introduces a shift consistent with the observed experimental mixture spectrum and is accurately captured by the linear algorithm. Thus, beyond correct identification, analysis of component contributions and coefficient rankings provides a framework for interpreting complex experimental mixtures that enables automated identification of liquid-phase mixtures in laboratory settings.

3 Discussion

Interpreting liquid-phase IR spectra often requires identifying nonlinearities that can be explained by intermolecular interactions or chemical arguments, hindering automation of spectroscopic analysis. Despite the existence of these nonlinearities in spectral data, this work shows that linear decomposition algorithms are sufficient to automatically identify liquid-phase mixtures from IR spectra provided knowledge of the pure-

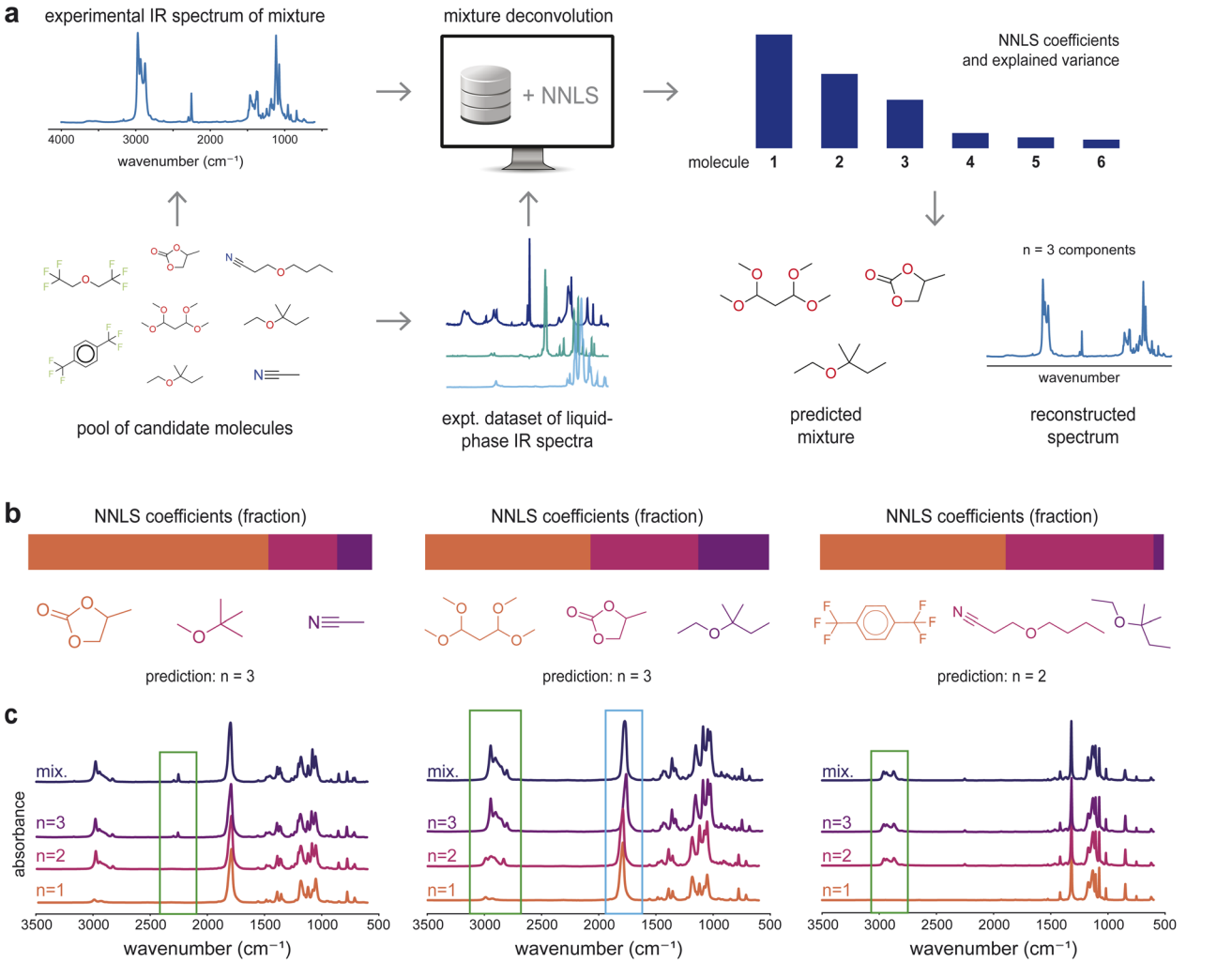

Fig. 6 Automatic identification of experimental mixtures in a blind study. (a) Components of liquid-phase mixtures were identified from experimentally observed mixture spectra using a basis set of experimentally measured pure liquid-phase spectra and the approach developed in this work. (b) Examples of top-3 mixture components ranked by NNLS coefficients for three different mixtures. (c) Cumulative weighted spectrum reconstructions obtained by sequentially adding the top $n = 1, 2, 3$ NNLS-ranked component spectra, each weighted by its NNLS coefficient. The spectrum labeled as "mix" is the measured IR spectrum for the mixture.

component liquid-phase spectra. The accuracy of the approach seems to be limited by the degeneracy of IR spectra in the liquid phase rather than nonlinearities in the data. We established a benchmark for chemical identification in liquid-phase mixtures and highlighted the necessity of liquid-phase spectral data for reliable chemical automation, using a non-negative least squares (NNLS) algorithm and an extensive dataset of simulated spectra.

The presented results demonstrate that the methods can be transferred directly to experimental data, as shown in the identification of candidate components in real experimental mixtures. NNLS coefficients provide chemically informative signals, indicating partial spectral similarity to true components and revealing plausible structural motifs that guide interpretation of spectroscopic data (Section S1.4). Importantly, NNLS accurately identifies two- and three-component mixtures and remains robust to peak shifts and to increases in pure-component dataset sizes while preserving chemically interpretable coefficients. The algorithm outperforms other linear methods tested because it restricts the space of solutions to physically meaningful, sparse, and well-separated coefficient distributions that better isolate dominant mixture components (Section S1.3).

Poor identification accuracy when using gas-phase pure-component spectra as the "basis set" for liquid-phase mixture deconvolution highlights the need to create larger liquid-phase IR spectral datasets. Existing spectroscopic datasets for liquid-phase IR are either limited in size or are protected by licensing requirements that prevent their use in large-scale identification efforts, while the simulated liquid-phase IR spectra here exhibit systematic deviations from experimental data due to the approximations of the force fields. Future efforts to curate and share experimental liquid-phase IR spectral data can be invaluable for automation of chemical identification in laboratory or industrial settings.

The MD simulations used to generate our spectral dataset capture liquid-phase spectral behavior, even though the spectra do not exactly reproduce experimental measurements. The MD-generated IR spectra exhibit self-consistent peak positions, shifts, and broadenings that mirror experimentally observed gas–liquid spectral differences. These intermolecular interaction-driven differences were quantified using cumulative distribution functions of spectral intensities, confirming that the simulations capture the intermolecular effects that distinguish gas- and liquid-phase spectra. Importantly, these gas–liquid shifts correlate with known chemical functionality, with molecules that promote stronger intermolecular interactions exhibiting larger spectral shifts. Interaction-driven features also persist in simulated mixture spectra, indicating that the simulations capture the anharmonic and nonlinear mixing behavior characteristic of the liquid phase.

Interpreting NNLS coefficients enables deeper analysis into the deconvolution of mixture IR spectra. While accuracy is high in gas-phase identification, misidentification profiles of liquid-phase IR spectra show that prediction failures arise primarily when the IR spectra of true and false-positive candidates are nearly indistinguishable. In these cases, misidentifications reflect fundamental limits in discriminative IR information rather than just shortcomings of the linear inversions (Section S1.5). The close correspondence between spectral similarity and shared structural motifs of pure components further supports this interpretation. Furthermore, linear unmixing becomes slower and harder for mixtures composed of highly degenerate spectra, which may reflect challenges in coefficient convergence when weights must be distributed across many nearly indistinguishable candidate spectra. Even though non-linear methods, including machine learning strategies, could exhibit better performance in component prediction, there will be numerical limits to the accuracy due to degenerate spectra.

Our results show that NNLS deconvolution for liquid-phase mixture identification is ready for use in automated laboratories. The approach offers interpretable coefficients, robustness to peak shifts, and applicability across both synthetic and experimental data, while providing transparent analysis of spectral reconstruction quality and chemical interpretability (Sections S1.3 and S1.4). The practical value of this interpretable framework is reinforced by further improvements in identification accuracy when incorporating mixture atomic composition information, which is easily available in experimental settings. Future work can address the limitations of linear algorithms and dataset sizes to increase the reliability of predictions and further automate the interpretation of nonlinearities in mixture IR spectra. Explicitly quantifying spectral degeneracy provides a pathway toward defining the fundamental limits of IR-based mixture identification under any deconvolution algorithm, thereby enabling uncertainty estimates that are particularly valuable in experimental applications.

4 Conclusion

This work presented a framework for the automatic identification of components in liquid-phase mixtures using IR spectra and algorithmic approaches that are useful for chemical research and industrial processes. Automatic identification was performed on a dataset of simulated binary and ternary mixture spectra using a database of simulated pure-component gas- and liquid-phase spectra. In particular, we showed that linear deconvolution using the non-negative least squares (NNLS) algorithm enabled component identification from IR spectra of mixtures with up to 90% accuracy. The NNLS algorithm yields component coefficients that are interpretable as relative contributions to the mixture spectrum and remain informative even in the cases of false-positive component identifications. Misidentification profiles were shown to stem from spectral and chemical similarity between candidate compounds, originating from the degeneracy between IR spectra rather than shortcomings of the linear decomposition method. Finally, this method was shown to be applicable to experimental settings in a blind study on experimentally prepared mixtures, in which the framework successfully identified the mixture components. Consequently, this work defines a benchmark for liquid-phase mixture identification from IR spectra and establishes a scalable workflow for automated mixture identification. Future advances in algorithmic methods, spectral modeling methods, and the availability and expansion of spectral databases can continue to improve automated interpretation of complex chemical mixtures.

5 Methods

5.1 Simulation methods

5.1.1 Molecular simulations. Gas- and liquid-phase pure component spectra along with liquid-phase two- and three-component mixture spectra were generated by molecular dynamics (MD) simulations using OpenMM[37] and a custom-made code. The potential energy surface was represented using OpenFF's Sage (v. 2.0.0) force field, as implemented in the OpenForceField codebase.[38–40] Simulations were initialized by instantiating an OpenFF Molecule object using the SMILES string, and creating a conformer in the gas phase using OpenFF's in-built functions. All simulations were performed at a constant temperature of 300 K, enforced by a Langevin thermostat with a coupling constant of 1 ps^{-1}, and a time step of 2 fs. Although the high time step leads to peak shifts corresponding to C–H, O–H, and N–H bonds, the shifts are systematic across the dataset and do not influence the identification methods. For liquid simulations, simulation boxes were initialized with a constant density of 0.5 g cm^{-3} containing 100 molecules using PACKMOL.[41] All mixture simulations were performed in the liquid phase with equimolar ratios, thus with 50 molecules of each species inside the simulation box. A Monte Carlo algorithm implemented in OpenMM was used as a barostat in the equilibration stage of the liquid phase simulations to simulate the effect of constant pressure, with a frequency of 200 fs^{-1}. Throughout all simulations, equilibration was performed for 0.5 ns for every system. Afterwards, production simulations were performed for 0.5 ns in the NVT ensemble. In liquid-phase simulations, the equilibrium density of the box was taken as the average density of the last 100 000 steps of the equilibration trajectory. These simulations showed good convergence with

respect to density and IR spectra, and thus were adopted as default parameters for this study. However, approximations related to the force fields prevent direct comparison with experimental data across this study. Additionally, while not all systems are necessarily miscible nor liquid under the simulation conditions, they span enough conformational diversity to enable a systematic study on chemical identification. Thus, the thermodynamics of the liquids and mixtures were not considered in this study.

5.1.2 Calculation of the IR spectra. To compute the IR spectra consistently for both gas- and liquid-phase molecules, the net molecular dipole of the simulation box was computed at each timestep. Then, the total system dipole was recorded in a file for each time step. The IR spectrum was obtained by following the implementation from Braun,[25,42] taking the Fourier transform of the autocorrelation function of this dipole over the trajectory to compute the frequency-dependent absorbance:

$$\tilde{C}(\nu) = \int_0^\infty C(t)\cos(2\pi\nu t)\mathrm{d}t, \tag{1}$$

$$C(t) = \frac{\left\langle \vec{M}(0)\cdot\vec{M}(t)\right\rangle}{\left\langle \vec{M}(0)\cdot\vec{M}(0)\right\rangle}. \tag{2}$$

where $\vec{M}(t)$ is the total dipole moment of the system at time t, ν is the frequency, and $\langle\cdot\rangle$ denotes an ensemble average over the trajectory. To account for coupling with an electromagnetic field and thermal weighting of vibrational modes, the spectrum is modified with correction factors:

$$F(\nu) = \nu\left(1 - \mathrm{e}^{-\frac{\hbar\nu}{k_\mathrm{B}T}}\right) \tag{3}$$

$$Q(\nu) = \frac{\nu}{1 - \mathrm{e}^{-\frac{\hbar\nu}{k_\mathrm{B}T}}} \tag{4}$$

where $\hbar$ is the reduced Planck constant, k_B is the Boltzmann constant, T is the temperature, and $F(\nu)$ and $Q(\nu)$ are the field description and quantum correction factors, respectively, used to adjust the classical spectrum to a quantum corrected spectrum:

$$S(\nu) = \tilde{C}(\nu)F(\nu) \tag{5}$$

$$S_\mathrm{qm}(\nu) = S(\nu)Q(\nu) \tag{6}$$

The final expression for the IR spectrum is thus given by:

$$S_\mathrm{qm}(\nu) = \tilde{C}(\nu)\nu^2 \tag{7}$$

Gas-phase spectra were generated using a single molecule per simulation box with molecules considered sufficiently far apart so that intermolecular dipoles are effectively uncorrelated. Under these conditions, mixtures of gas-phase spectra are assumed to have vanishing correlations between the dipoles of different molecules. Thus, the IR spectrum of a gas-phase mixture is simply the mole-fraction weighted linear combination of the pure-component spectra.

5.1.3 Automated calculation workflow. Automation of the IR spectra calculations was implemented using the mkite software.[43] The software suite allows for the generation of jobs with an arbitrary number of inputs in a matrix-like fashion, thus allowing the calculation of two- and three-component mixtures within a single database. Data management was performed using a new plugin, developed for this work at https://github.com/mkite-group/mkite-infrared.

Every spectrum is represented by a one-dimensional vector of 1250 wavenumber indices, corresponding to a uniform 4 cm^{-1} grid spanning 0 to 5000 cm^{-1}, with every value being the intensity at each wavenumber. The raw MD-derived intensities were interpolated onto this common grid, Gaussian smoothed to reduce noise, and negative intensities where clipped to zero. Each spectrum was normalized by its total integrated intensity to produce a probability-density-like spectrum vector. This unit-area normalization is used for all simulated spectra analyzed in this work.

5.1.4 Dataset of molecules. The dataset of about 8880 molecules used for calculations of gas- and liquid-phase spectra was created by combining the AqSolDB[44] (originally with 9982 compounds) with two other custom-made datasets. One custom dataset was about 50 molecules that were readily accessible to the experimental team. The other dataset was around 380 molecules generated by combinatorially matching one molecular "core" scaffold with one fragment. To generate the combinatorial core-fragment library, a curated set of structurally diverse fragments, defined as monovalent R-group substituents, was combined with a collection of reactive core scaffolds. The core scaffolds comprised related structural backbones that differed in the position of the fragment attachment site, enabling single-point R-group modification at distinct locations. All possible combinations of the defined scaffold core and R-group fragments were enumerated to produce a dataset of closely related yet structurally distinct molecules. The SMILES of the resulting molecules were canolicolized and subjected to chemical sanitization to ensure data consistency, valence correctness, and structural validity.

After removing structures for which MD simulations failed to converge, we obtained 8880 pure gas-phase and 8550 pure liquid-phase simulations. Mixture molecules were selected using two approaches. From the AqSolDB dataset, binary and ternary mixtures were constructed by randomly selecting pairs or triplets of compounds. From the core-fragment dataset, mixtures were generated by enumerating all possible pairwise combinations.

5.2 Experimental IR spectra

Three two-component and six three-component experimentally prepared liquid-phase mixtures were evaluated with all algorithms using 143 liquid-phase spectra as the basis set. The basis set contained twenty pure components that were true components of the mixtures, measured using Fourier transform infrared spectroscopy with attenuated total reflectance (FTIR-

ATR). The ATR apparatus (Pike instrument), using diamond as the ATR prism, was connected to the IR spectrometer (Nicolet iS50, Thermo Scientific) with a deuterated triglycine sulfate (DTGS) detector. The incident IR light was fixed at 45°. The IR background and sample spectra (4000 to 400 cm^{-1} range) were collected with a spectral resolution of 4 cm^{-1} and averaged over 128 scans. ATR correction was performed in OMNIC software (Thermo Scientific). Ethylene carbonate (EC, Gotion), dimethyl carbonate (DMC, Sigma-Aldrich), diethyl carbonate (DEC, Sigma-Aldrich), ethyl methyl carbonate (EMC, Gotion), fluoroethylene carbonate (FEC, Gotion), propylene carbonate (PC, Sigma-Aldrich), acetonitrile (Thermo Scientific), *tert*-butyl methyl ether (Sigma-Aldrich), *tert*-amyl ethyl ether (Sigma-Aldrich), 1,4-bis (trifluoromethyl) benzene (TCI America), 1,1,3,3-tetramethoxypropane (Sigma-Aldrich), 3-butoxypropionitrile (Sigma-Aldrich), bis(2,2,2-trifluoroethyl)ether (Synquest Labs, Inc), 2,2,2-trifluoroethyl acetate (Synquest Labs, Inc), hexafluorobenzene (Sigma-Aldrich), 1,3-dioxolane (DOL, Sigma-Aldrich), ethylenediamine (TCI America), ethyl methyl sulfone (EMS, TCI America), 1,2-dimethoxyethane (DME, Sigma-Aldrich), and 1,1,2,2-tetrafluoroethyl 2,2,3,3-tetrafluoropropylether (TTE, TCI America) were used as received. An additional 123 pure-component spectra were obtained from the Open Specy IR spectral library,[45] to augment the basis set and increase prediction difficulty.

During evaluation, the true number of mixture components was hidden. The cumulative fractional spectral variance, mean squared error, and algorithm coefficients (described in Sections 5.4 and 5.5) were used to predict mixture components and infer the number of components.

5.3 Deconvolution algorithms for IR spectra data

In this work, we treated single-component IR spectral data as the "basis set" of a multi-component mixture IR spectrum. Least squares (LS), non-negative least squares (NNLS), and their regularized variants were tested to estimate component contributions in simulated mixture spectra.

The LS method minimizes the cost function

$$\min_{C} \|\mathbf{Y} - \mathbf{XC}\|_2^2 + \lambda \|\mathbf{C}\|^2 \tag{8}$$

whereas the NNLS method minimizes the same cost function while imposing a non-negative restriction on the coefficients,

$$\min_{C \geq 0} \|\mathbf{Y} - \mathbf{XC}\|_2^2 + \lambda \|\mathbf{C}\|^2 \tag{9}$$

In eqn (8) and (9), **Y** denotes the set of simulated mixture spectra and **X** denotes the set of simulated pure component spectra. The coefficient matrix **C** is defined such that every row corresponds to a mixture and each column corresponds to a component in the basis set. From this, every value in a mixture row of **C** specifies the coefficient of a particular component in that mixture.

In the non-regularized formulations, the regularization coefficient λ is set to zero.

In addition to the algorithms above, a "brute-force" interpolation method was also implemented as a baseline for two-component mixtures. It exhaustively evaluates all pairs of single-component spectra and solves for the least-squares optimal convex mixing coefficients. The interpolation method serves as a reference for evaluating component coefficient accuracy under the same linear mixing assumption as in the gas phase.

5.4 Evaluating identification accuracy

The coefficients corresponding to the single-component spectra in the basis set, as given by the linear algorithms, were used to identify the components of a mixture from its spectrum. Identification accuracy was evaluated using two approaches. In both cases, the coefficients obtained were ranked in descending order of their absolute values, and identification accuracy was evaluated by determining whether the top k coefficients corresponded to the spectra of the true components in the mixture.

5.4.1 Exact top-*k* component identification. The top-2 and top-3 coefficients were used for binary and ternary mixtures, respectively. In this method, the pure-component basis set included spectra for all molecules present in the mixtures, as well as an equal number of spectra from molecules not appearing in any of the mixtures. These spectra were selected at random. Identification accuracy was evaluated eight times, each using a different random selection of additional molecular spectra.

5.4.2 Criterion-based top-*k* identification. Identification accuracy was also evaluated for finding the correct mixture components from the top-k coefficients evaluated by (1) requiring all true components to appear within the top k, (2) requiring at least one (any) true component to appear within the top-k, and (3) applying an atom-count filter that restricts candidate components whose combined atomic compositions match the mixtures atom composition (described in SI Section S2.3). Using these criteria, identification accuracy was evaluated as a function of pure-component basis set size. Only the components contained in all the mixtures were part of the smallest basis set size. For larger pure-component dataset sizes, additional pure component spectra were added to the basis set, ensuring that each subsequently large basis set size contained all the pure-component spectra in the smaller basis sets.

When computing coefficients using varying basis set sizes, NNLS failed to converge for the two largest basis set sizes. Using a basis set size of 4920 pure components, NNLS failed for 12 mixtures, corresponding to 0.043% of the mixture dataset. Using a basis set size of 8528, 22 mixtures failed to converge, corresponding to 0.080% of the mixture dataset. Mixtures for which NNLS did not converge were considered unidentifiable and counted as incorrect predictions when accuracy metrics were computed. These failures were attributable to the NNLS solver reaching its iteration limit under increased basis collinearity at larger basis sizes.

5.5 Spectral analysis metrics

Three metrics were used to quantify a predicted component's contribution to a mixture spectrum: the average cumulative

distribution function difference, fractional spectral variance, and mean squared error (MSE).

5.5.1 Cumulative distribution function difference. The difference between the cumulative distribution functions (CDFs) of two spectra was used to quantify relative spectral differences. At a given wavenumber, a larger cumulative intensity difference indicates red shifting of one spectrum relative to the other. Cumulative intensity curves can also be used to identify raw intensity differences, reflected by vertical differences between the curves, and relative spectral broadening, indicated by their horizontal differences. The average CDF difference, defined as the mean difference between two cumulative spectra across all wavenumbers, was used as a quantitative metric for comparing two spectra.

5.5.2 Fractional spectral variance. The fractional spectral variance metric measures how well a mixture spectrum is reconstructed as components are added in decreasing order of their NNLS coefficient values. Starting from a zero-intensity spectrum, pure component spectra were sequentially added in descending order of their NNLS coefficient rankings. At each addition step, the selected component spectrum was scaled by its predicted coefficient and cumulatively summed to form a partial reconstruction of the mixture spectrum. After each addition, the partial reconstruction $\hat{y}$ was compared to the true mixture y, and a cumulative explained variance metric was computed:

$$R^2 = 1 - \frac{\|y - \hat{y}\|^2}{\|y\|^2} \quad (10)$$

This value measures the fraction of variance in the true spectrum explained by the partial reconstruction.

The same procedure was applied using MSE and the average CDF difference to assess each component's contribution to reconstruction quality. The incremental contribution of each component was evaluated using these three metrics: the stepwise increase in R^2, the decrease in MSE, and the decrease in average CDF difference as the mixture spectrum was assembled from zero intensity.

A complementary analysis was performed by removing each component from the basis set, recomputing the NNLS coefficient vector for the same mixture spectrum, and calculating the resulting R^2, MSE, and average CDF difference for the predicted mixture spectrum relative to the true spectrum. This method quantifies the degradation in fit to the true spectrum when a component is unavailable, indicating the importance of that component's spectral features in explaining the mixture spectrum. Component contributions were quantified by the drop in R^2 and the corresponding increases in MSE and average CDF difference when that component was removed.

Author contributions

Y. J. U. M.: methodology; software; validation; formal analysis; investigation; data curation; writing – original draft; writing – review & editing; visualization. T. N.: methodology; validation; investigation; writing – review & editing. J. L.: conceptualization; methodology; formal analysis; investigation; resources; writing – review & editing; supervision; project administration; funding acquisition. D. S.-K.: conceptualization; methodology; software; formal analysis; investigation; resources; data curation; writing – original draft; writing – review & editing; visualization; supervision; project administration; funding acquisition.

Conflicts of interest

The authors have no conflicts to disclose.

Data availability

The simulated and experimental IR spectra used for this work can be found at https://zenodo.org/records/19637299. The additional dataset referenced in 5.2 can be accessed at https://pubs.acs.org/doi/suppl/10.1021/acs.analchem.1c00123/suppl_file/ac1c00123_si_002.zip.

Code availability: the code used to obtain infrared spectra from molecular dynamics simulations is available at https://github.com/digital-synthesis-lab/autoir and https://github.com/digital-synthesis-lab/mkite_infrared. The notebooks/analysis code used to reproduce the results/plots from this work is accessible at https://github.com/digital-synthesis-lab/automatic-ir-id.

Supplementary information (SI): supplementary discussion on methods, additional results, and supporting figures. See DOI: https://doi.org/10.1039/d6sc01583b.

Acknowledgements

Financial support for this publication results from Scialog grant #SA-AUT-2024-024 from Research Corporation for Science Advancement and Frederick Gardner Cottrell Foundation. Y. J. U. M. acknowledges supplementary funding from the Amazon AI PhD Fellowship at UCLA Samueli. This work used computational and storage services associated with the Hoffman2 Shared Cluster provided by UCLA Office of Advanced Research Computing's Research Technology Group, as well as Delta CPU and Delta GPU at NCSA through allocation MAT240040 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296. The work performed at the Reactor Engineering and Catalyst Testing (REACT) Core Facility of the Northwestern University Center for Catalysis and Surface Science (CCSS) was supported by a grant from the DOE (DE-SC0001329). Partial support for instrumentation in REACT is also provided by Northwestern's MRSEC program (NSF DMR-2308691). This work made use of the Keck-II facility (RRID: SCR_026360) of Northwestern University's NUANCE Center, which has received support from the IIN and Northwestern's MRSEC program (NSF DMR-2308691). The authors thank Gabe Gomes, Jose Regio, and Jiawei Guo for useful discussions.

References

1 P. R. Griffiths and J. A. de Haseth, Infrared spectroscopy: Fundamentals and applications, in *Encyclopedia of Analytical Chemistry*, John Wiley & Sons, Ltd, Chichester, UK, 2006.
2 V. H. Paschoal, L. F. O. Faria and M. C. C. Ribeiro, Vibrational spectroscopy of ionic liquids, *Chem. Rev.*, 2017, **117**, 7053–7112.
3 *Biological and Biomedical Infrared Spectroscopy. Advances in Biomedical Spectroscopy*, ed. Barth A. and Haris P. I., IOS Press, Amsterdam, Netherlands, 2009.
4 B. H. Stuart, *Infrared Spectroscopy: Fundamentals and Applications*, John Wiley & Sons, 2004.
5 B. C. Smith, *Infrared Spectral Interpretation: A Systematic Approach*, CRC Press, Boca Raton, FL, 2009.
6 P. J. Linstrom and W. G. Mallard, NIST Chemistry WebBook, NIST Standard Reference Database Number 69, 2025, https://webbook.nist.gov/chemistry/.
7 P. M. Chu, F. R. Guenther, G. C. Rhoderick and W. J. Lafferty, The NIST quantitative infrared database, *J. Res. Natl. Inst. Stand. Technol.*, 1999, **104**, 59–69, NIST Standard Reference Database 79 (SRD 79), primary citation for the dataset accessed on July 21, 2015.
8 National Institute of Advanced Industrial Science and Technology (AIST), SDBS spectral database for organic compounds, https://sdbs.db.aist.go.jp/, accessed on July 21, 2015.
9 D. Granato, *et al.*, Trends in chemometrics: Food authentication, microbiology, and effects of processing, *Compr. Rev. Food Sci. Food Saf.*, 2018, **17**, 663–677.
10 K. J. Johnson, R. E. Morris and S. L. Rose-Pehrsson, Evaluating the predictive powers of spectroscopy and chromatography for fuel quality assessment, *Energy Fuels*, 2006, **20**, 727–733.
11 Y. Roggo, *et al.*, A review of near infrared spectroscopy and chemometrics in pharmaceutical technologies, *J. Pharm. Biomed. Anal.*, 2007, **44**, 683–700.
12 J. C. Prata and P. Martins da Costa, Fourier transform infrared spectroscopy use in honey characterization and authentication: A systematic review, *ACS Food Sci. Technol.*, 2024, **4**, 1817–1828.
13 C. L. Cunha, *et al.*, Predicting the properties of biodiesel and its blends using mid-FT-IR spectroscopy and first-order multivariate calibration, *Fuel*, 2017, **204**, 185–194.
14 E. Sufriadi, *et al.*, Partial least squares-discriminant analysis classification for patchouli oil adulteration detection by Fourier transform infrared spectroscopy in combination with chemometrics, *ACS Omega*, 2023, **8**, 12348–12361.
15 Y. Wang, *et al.*, On estimating physical and chemical properties of hydrocarbon fuels using mid-infrared FTIR spectra and regularized linear models, *Fuel*, 2019, **255**, 115715.
16 H.-J. van Manen, J. Gerretzen, M. Smout, G. Postma and J. J. Jansen, Quantitative vibrational spectroscopy on liquid mixtures: concentration units matter, *Analyst*, 2021, **146**, 3150–3156.
17 E. Al Ibrahim and A. Farooq, Prediction of the derived cetane number and carbon/hydrogen ratio from infrared spectroscopic data, *Energy Fuels*, 2021, **35**, 8141–8152.
18 Á. M. Ní Fhuaráin, C. P. O'Donnell, J. Luo and A. A. Gowen, A Review on MIR, NIR, Fluorescence and Raman Spectroscopy Combined with Chemometric Modeling to Predict the Functional Properties of Raw Bovine Milk, *ACS Food Sci. Technol.*, 2024, **4**, 2258–2271.
19 D. M. Haaland, R. G. Easterling and D. A. Vopicka, Multivariate least-squares methods applied to the quantitative spectral analysis of multicomponent samples, *Appl. Spectrosc.*, 1985, **39**, 73–84.
20 D. M. Haaland and R. G. Easterling, Application of new least-squares methods for the quantitative infrared analysis of multicomponent samples, *Appl. Spectrosc.*, 1982, **36**, 665–673.
21 F. Cahn and S. Compton, Multivariate calibration of infrared spectra for quantitative analysis using designed experiments, *Appl. Spectrosc.*, 1988, **42**, 865–872.
22 H.-k. Xiao, S. P. Levine and J. B. D'Arcy, Iterative least-squares fit procedures for the identification of organic vapor mixtures by Fourier transform infrared spectrophotometry, *Anal. Chem.*, 1989, **61**, 2708–2714.
23 D. J. Griffin, M. A. Grover, Y. Kawajiri and R. W. Rousseau, Robust multicomponent ir-to-concentration model regression, *Chem. Eng. Sci.*, 2014, **116**, 77–90.
24 G. Livanos, M. Zervakis, N. Pasadakis, M. Karelioti and G. Giakos, Deconvolution of petroleum mixtures using mid-FTIR analysis and non-negative matrix factorization, *Meas. Sci. Technol.*, 2016, **27**, 114005.
25 M. Alberts, T. Laino and A. C. Vaucher, Leveraging infrared spectroscopy for automated structure elucidation, *Commun. Chem.*, 2024, **7**, 268.
26 M. Alberts, F. Zipoli and T. Laino, Setting New Benchmarks in AI-Driven Infrared Structure Elucidation, *Digital Discovery*, 2025, **4**, 1936–1943.
27 X.-Y. Lu, *et al.*, Deep learning-assisted spectrum-structure correlation: State-of-the-art and perspectives, *Anal. Chem.*, 2024, **96**, 7959–7975.
28 A. A. Enders, N. M. North, C. M. Fensore, J. Velez-Alvarez and H. C. Allen, Functional Group Identification for FTIR Spectra Using Image-Based Machine Learning Models, *Anal. Chem.*, 2021, **93**, 9711–9718.
29 F. Zipoli, M. Alberts and T. Laino, IR-NMR multimodal computational spectra dataset for 177 k patent-extracted organic molecules, *Sci. Data*, 2025, **12**, 1375, DOI: 10.1038/s41597-025-05729-8.
30 G. C. Kanakala, B. Sridharan and U. D. Priyakumar, Spectra to structure: Contrastive learning framework for library ranking and generating molecular structures for infrared spectra, *Digital Discovery*, 2024, **3**, 2417–2423.
31 E. J. French *et al.*, Revolutionizing spectroscopic analysis using sequence-to-sequence models i: From infrared spectra to molecular structures, *ChemRxiv*, preprint, 2025, DOI: 10.26434/chemrxiv-2025-n4q84.

32 T. Jin, Q. Zhao, A. B. Schofield and B. M. Savoie, Deductive machine learning models for product identification, *Chem. Sci.*, 2024, **15**, 11995–12005, DOI: 10.1039/D3SC04909D.
33 F. Ficarra, M. Alberts, and T. Laino, Language model enabled structure prediction from infrared spectra of mixtures, *ChemRxiv*, preprint, 2025, https://chemrxiv.org/engage/chemrxiv/article-details/686249a91a8f9bdab5bfefee.
34 S. N. Abdul Al and A.-R. Allouche, Neural network approach for predicting infrared spectra from 3d molecular structure, *Chem. Phys. Lett.*, 2024, **856**, 141603.
35 A. R. Kartha, D. P. Ajayakumar, M. Idris and G. Ragupathy, Unlocking the potential of machine learning in enhancing quantum chemical calculations for infrared spectral prediction, *ACS Omega*, 2025, **10**, 19224–19234.
36 C. McGill, M. Forsuelo, Y. Guan and W. H. Green, Predicting infrared spectra with message passing neural networks, *J. Chem. Inf. Model.*, 2021, **61**, 2594–2609.
37 P. Eastman, *et al.*, Openmm 8: Molecular dynamics simulation with machine learning potentials, *J. Phys. Chem. B*, 2023, **128**, 109–116.
38 S. Boothroyd, *et al.*, Development and benchmarking of open force field 2.0.0: the sage small molecule force field, *J. Chem. Theory Comput.*, 2023, **19**, 3251–3275.
39 J. Wagner, M. Thompson, D. Dotson, and J. Rodríguez-Guerra *et al.*, *OpenForceField/openff-forcefields: Version 2.0.0 "Sage"*, Zenodo, 2021.
40 D. L. Mobley, *et al.*, Escaping atom types in force fields using direct chemical perception, *J. Chem. Theory Comput.*, 2018, **14**, 6076–6092.
41 L. Martínez, R. Andrade, E. G. Birgin and J. M. Martínez, Packmol: A package for building initial configurations for molecular dynamics simulations, *J. Comput. Chem.*, 2009, **30**, 2157–2164.
42 E. Braun Open Source Code: Calculating an IR Spectra from a LAMMPS Simulation, Zenodo, 2016, DOI: 10.5281/zenodo.154672.
43 D. Schwalbe-Koda, MKite: A distributed computing platform for high-throughput materials simulations, *Comput. Mater. Sci.*, 2023, **230**, 112439.
44 M. C. Sorkun, A. Khetan and S. Er, Aqsoldb, a curated reference set of aqueous solubility and 2d descriptors for a diverse set of compounds, *Sci. Data*, 2019, **6**, 143, DOI: 10.1038/s41597-019-0151-1.
45 W. Cowger, *et al.*, Microplastic spectral classification needs an open source community: Open specy to the rescue!, *Anal. Chem.*, 2021, **93**, 7543–7548, DOI: 10.1021/acs.analchem.1c00123.