Completely computational model setup for spectroscopic techniques: the *ab initio* molecular dynamics indirect hard modeling (AIMD-IHM) approach

Justus Wöhl, Wassja A. Kopp, Iryna Yevlakhovych, Leo Bahr, Hans-Jürgen Koß, and Kai Leonhard*

Institute of Technical Thermodynamics, RWTH Aachen University, 52062 Aachen, Germany

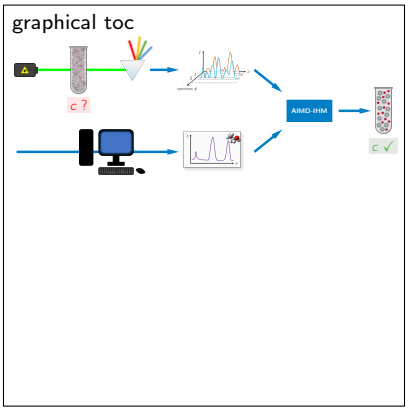
E-mail: kai.leonhard@ltt.rwth-aachen.de

Phone: +49 (0)241 8098174. Fax: +49 (0)241 92255

Abstract

The spectroscopic quantification of mixture compositions usually requires pure compounds and mixtures of known composition for calibration. Since they are not always available, methods to fill such gaps have evolved, which are, however, not generally applicable. Therefore, calibration can be extremely challenging, especially when multiple instable species, *e.g.* intermediates, exist in a system. This study presents a new calibration approach that uses *ab initio* Molecular Dynamics (AIMD)-simulated spectra as to set up and calibrate models for the physics-based spectral analysis method Indirect Hard Modeling (IHM). To demonstrate our approach called AIMD-IHM, we analyze Raman spectra of ternary hydrogen-bonding mixtures of acetone, methanol, and ethanol. The derived AIMD-IHM pure-component models and calibration coefficients are in good agreement with conventionally generated experimental results. The method yields compositions with prediction errors of less than 5 % without any experimental calibration input. Our approach can be extended, in principle, to IR and NMR spectroscopy and allows for the analysis of systems that were hitherto inaccessible to quantitative spectroscopic analysis.

Graphical TOC Entry



Keywords

quantitative mixture analysis, model-based calibration, TRAVIS, spectral simulation, missing-PCS, aixcalibration

Raman, infrared (IR), and nuclear magnetic resonance (NMR) spectroscopy provide fast, non-invasive and *in-situ* quantification of mixture compositions, especially suited for reactive systems. Quantitative spectral analysis of mixtures can be achieved by data-driven methods, like Partial Least Squares Regression (PLSR), or by physics-based methods, such as Classical Least Squares (CLS)[1] and Indirect Hard Modeling (IHM)[2]. In comparison to data-driven methods, physics-based methods require fewer calibration spectra. These include mixture spectra with known compositions as well as pure-component spectra (PCSa)[2]. However, for complex (*e.g.* reactive) systems, specific calibration spectra and some PCSa are often unavailable or experimentally inaccessible. For instance, for acids that dissociate in multiple steps, PCSa cannot be determined experimentally, since the ionic species cannot be extracted as pure components. As a different example, despite their versatile applications, ionic liquids (ILs) suffer from decomposition to often unknown intermediates and products[3,4]. In both systems, yet unknown PCSa and calibration coefficients hamper the use of spectroscopic methods to calculate the composition of mixtures. Other examples include spectroscopic *in operando* determination of intermediates in catalysis[5–7] or quantification of large numbers of metabolites in biological samples[8].

For specific cases, unavailable PCSa can be extracted from mixture spectra applying different algorithms[9–13]. However, these algorithms are only applicable for mixtures either with one unknown component[13] or with multiple components that have distinctive peaks[11,13]. Besides, unavailable PCSa of individual components of the mixture system can sometimes be generated and calibrated externally in alternative systems. For example, for the dissociation of sulfuric acid, unavailable spectra of bisulfate ions were determined in a mixture with (Raman-inactive) lithium ions and calibrated in a mixture with sodium perchlorate[14]. However, these experimentally laborious approaches are not generally applicable.

Instead of acquiring spectra experimentally, we can infer spectra from molecular dynamics (MD) simulations. In *ab initio* molecular dynamics (AIMD) simulations, the interactions of the molecules of the system of interest are computed in a small virtual box at discrete

time steps from quantum chemical methods. Modern density functional theory (DFT) methods provide a particularly feasible compromise of accuracy and cost. The resulting forces are applied according to Newton's equations of motion to update the geometry of the system for each time step. The autocorrelation function of the atom velocities is related to the vibrational spectrum of the system. Amongst others, IR and Raman spectra can be calculated from the autocorrelation function. Dependent on the desired spectrum type, information on polarizability and dipole moment changes must also be obtained from the trajectory[15–18]. In contrast to "static" calculations, *e.g.* simple frequency calculations of ideal-gas equilibrium configurations, such simulations provide information on anharmonic vibrations and liquid phase effects caused by intermolecular interactions such as hydrogen bonds[15]. Such analysis of AIMD simulations has been successfully employed to obtain spectra, *e.g.* of organic solvents[19,20] or ILs[21] like ethyl-methyl-imidazolium (EMIM) acetate[22]. In this work, we present a new approach that uses simulated spectra from AIMD to set up IHM models for the quantification of compositions in mixtures with experimentally inaccessible calibration spectra. Fig. 1 shows a general overview of the AIMD-IHM method that is split into two approaches: missing-PCS and aixcalibration (*ab initio* extended calibration). The two modules derive unavailable pure-component models (PCMs) and calibration coefficients K_i (cf. Eq. 3), respectively, by combining measured mixture spectra with AIMD spectral simulations.

The new method is validated by investigating a well-known ternary mixture system of acetone, methanol, and ethanol. We chose this mixture system mainly for two reasons. First, the mixture system is fully accessible, *i.e.* all PCSa, as well as mixture spectra with known compositions, can be measured. In this work, however, this experimental information is used for reference purposes only. Second, the Raman spectra of acetone, methanol, and ethanol strongly overlap in some spectral regions, causing challenges with common techniques for the extraction of PCSa. By applying the new AIMD-IHM method, we demonstrate that the model to analyze mixtures spectroscopically can be set up completely based on simulations.

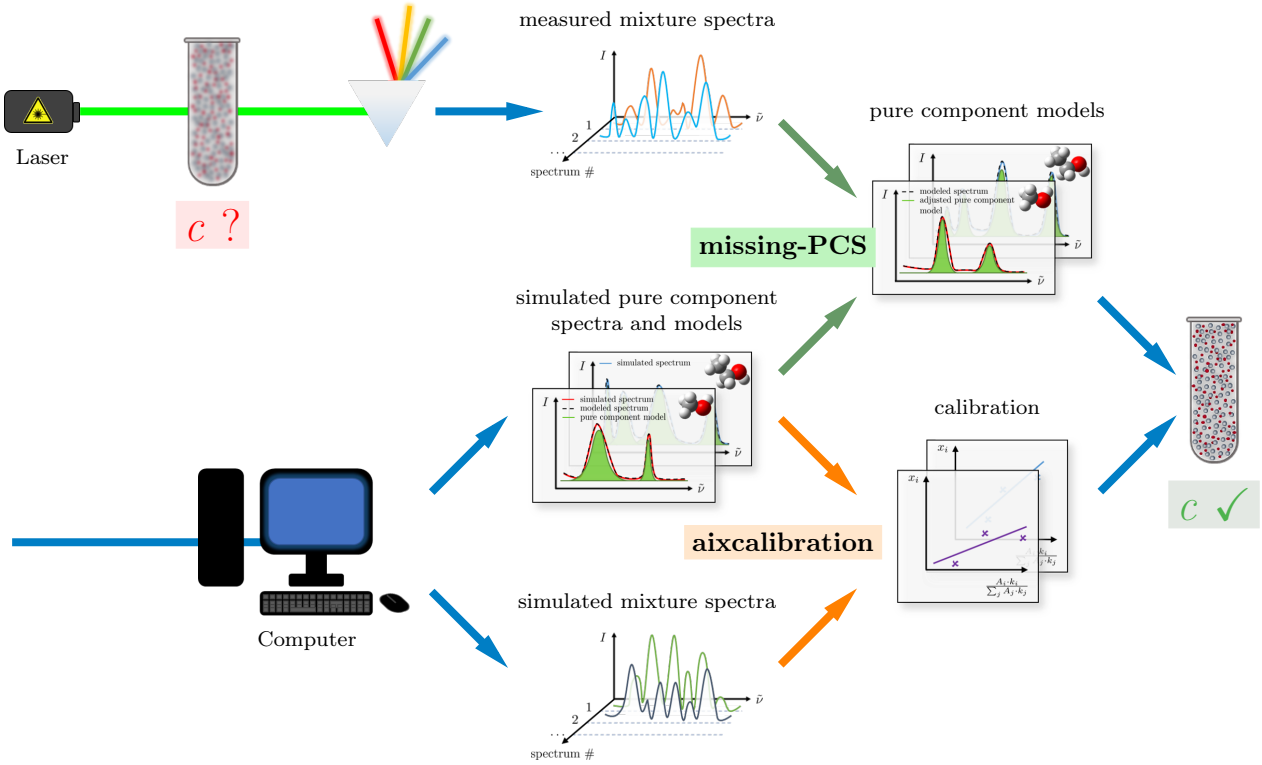


Figure 1: General overview of the two-step AIMD-IHM method using simulated and measured spectra to derive compositions of mixture spectra. The green arrows showcase the "missing-PCS" approach that uses measured mixture spectra and simulated PCSa to set up PCMs for IHM. The orange arrows showcase the "aixcalibration" approach that uses simulated mixture spectra and simulated PCSa to derive calibration coefficients for IHM.

Our novel AIMD-IHM method is based on the spectral analysis method IHM, which is described in detail in the original work of Alsmeyer *et al.*[2]. Physics-based spectral analysis methods like IHM model mixture spectra with a superposition of the corresponding PCSa. However, in contrast to simpler physics-based methods such as CLS[1], IHM uses parametrized models of the PCSa and can thereby account for so-called non-linear effects in mixtures like *e.g.* peak shifts or peak deformations. To quantify compositions from mixture spectra using IHM, a mixture model (containing all PCSa) and a linear calibration model (based on calibration spectra from mixtures with known compositions) have to be generated[2]. Therefore, the PCSa $S_{\mathrm{PC},i}(\tilde{\nu}_w)$ dependent on the Raman shift $\tilde{\nu}_w$ for all n_{comp} components are modeled *via* a sum of p parametrized peak functions P_j, the so-called PCMs $\mathcal{PCM}_i(\tilde{\nu}_w)$.

$$\boldsymbol{\Theta}_i = \underset{\boldsymbol{\Theta}_i, \boldsymbol{\Theta}_\mathrm{B}}{\arg\min} \sum_{w=1}^{n_{\tilde{\nu}}} \left\{ \mathcal{S}_{\mathrm{PC},\,i}(\tilde{\nu}_w) - \underbrace{\sum_{j=1}^{p} \mathcal{P}_j\left(\tilde{\nu}_w, \boldsymbol{\Theta}_i\right)}_{\mathcal{PCM}_i(\tilde{\nu}_w, \Theta_k)} - B(\tilde{\nu}_w, \boldsymbol{\Theta}_\mathrm{B}) \right\}^2 \tag{1}$$

The vector $\boldsymbol{\Theta}_i$ comprises all peak functions (pseudo-Voigt profiles), parametrized by their position, area, width, and shape. The index w counts all $n_{\tilde{\nu}}$ recorded wavenumbers. Potential background signals in the spectra are modeled with a baseline model $B(\tilde{\nu}_w, \boldsymbol{\Theta}_\mathrm{B})$. In this work, we use a linear baseline with the coefficients $\boldsymbol{\Theta}_\mathrm{B}$.

To set up the calibration model, the n_{mix} mixture spectra $S_{\mathbf{mix,m}}(\tilde{\nu}_w)$ with known compositions are modeled *via* a least-squares fitting using a weighted sum (with the weights w_m) of all expected PCMs, cf. Eq. 2. To model non-linear effects in the mixture spectra adequately, *e.g.* peak shifts or peak broadenings, specific peak parameters $\boldsymbol{\Theta}_i^*$ are adjustable during the spectral modeling. The concatenation of all PCMs including the adjustable peak parameters is called the mixture model in the following. The mixture spectra are modeled by computing

the weights of each PCM:

$$w_m = \underset{w_m, \boldsymbol{\Theta}_1^* \ldots \boldsymbol{\Theta}_{n_{\mathrm{comp}}}^*, \boldsymbol{\Theta}_B}{\arg\min} \sum_{w=1}^{n_{\tilde{\nu}}} \left\{ \mathcal{S}_{\mathrm{mix},\, m}(\tilde{\nu}_w) - \sum_{i=1}^{n_{\mathrm{comp}}} \left\{ w_{m,i} \cdot \mathcal{PCM}_i \left(\tilde{\nu}_w, \boldsymbol{\Theta}_i, \boldsymbol{\Theta}_i^* \right) \right\} - B(\tilde{\nu}_w, \boldsymbol{\Theta}_B) \right\}^2 \tag{2}$$

The resulting weights w_m for each mixture m are directly proportional to the concentrations and molar fractions $\boldsymbol{x}$ of the corresponding components in the mixture. From the weights w_m the molar fractions are determined *via* the calibration coefficients K_i, which differ for each component due to the molecule-specific Raman scattering cross-sections. All K_i (concatenated in $\boldsymbol{K}$) are determined *via* linear regression of the molar fractions against the corresponding areas of the components $A_{m,i} = w_{m,i} \cdot A_{0,i}$, where $A_{0,i}$ is the area of the PCM.

$$\boldsymbol{K} = \underset{\boldsymbol{K}}{\arg\min} \left\{ \sum_{m=1}^{n_{\mathrm{mix}}} \sum_{i=1}^{n_{\mathrm{comp}}} \left\{ \boldsymbol{x} - \frac{K_i \cdot \boldsymbol{A}_i}{\sum_{j=1}^{n_{\mathrm{comp}}} K_j \cdot \boldsymbol{A}_j} \right\}^2 \right\} \tag{3}$$

To compensate for experimental influences on the absolute signal intensity, *e.g.* fluctuating laser intensity, we use a ratiometric calibration. In order to conduct the calibration, first the molar fractions of the calibration spectra are calculated using the calibration coefficients K_i from Eq. 3 and the areas resulting from Eq. 2:

$$\boldsymbol{x}_{\mathrm{calc}} = \frac{K_i \cdot \boldsymbol{A}_i}{\sum_{j=1}^{n_{\mathrm{comp}}} K_j \cdot \boldsymbol{A}_j} \tag{4}$$

Secondly, the deviation between the calculated and the known molar fractions is quantified by calculating the root mean squared error (RMSE) of cross-validation of the calculated compositions $RMSECV_{\mathrm{Calib}}$. For analyzing mixtures with unknown compositions, the corresponding mixture spectra are first modeled with the mixture model according to Eq. 2. From the resulting areas, the molar fractions are then calculated from Eq. 4 using $\boldsymbol{K}$.

In contrast to the conventional IHM approach that uses experimentally derived PCSa and mixture spectra (with known composition for model setup), our suggested novel AIMD-

IHM approach is set up with simulated pure-component spectra (sim-PCSa) and simulated mixture spectra. Fig. 2 sketches the course of computations and employed software for these simulations. The molecular structures of the pure components (PCs) are provided by the user using the graphical interface GaussView[23]. Aiming for a simulation at an appropriate density and composition, the desired numbers of molecules are packed into a box of specified volume in a non-overlapping way using Packmol[24,25].

The packed box is input to the subsequent MD simulations. These simulations require computations of the interatomic forces, which are predominantly obtained from DFT calculations. Although DFT already represents a good compromise between accuracy and computational cost, for the first preparatory simulation steps an even simpler force field (FF) is used to further reduce computational cost. These preparatory simulation steps include the minimization of the potential energy of the box as well as the subsequent equilibration at the desired temperature. This minimization and a pre-equilibration are done using the fast generalized AMBER FF[26] and the LAMMPS MD code[27]. After the pre-equilibration step, the remaining equilibration as well as the production run is performed using the much more accurate DFT method. However, this increases the computational time compared to the FF calculations by about six orders of magnitude. Our DFT simulations, performed *via* CP2K[28] for 10 ps to 30 ps, take roughly five days wall-time on a 24 core processor. To conduct the remaining equilibration at DFT level, a massive equilibration with one thermostat per degree of freedom is performed. Then a non-massive one with one global thermostat follows, which is also used in the following production run. The production run yields the actual trajectory. To model the change in polarizability needed for the Raman spectrum, additional computations along the previously obtained field-free trajectory are performed with an electric field (details can be found in Thomas *et al.*[15]). The trajectory is subsequently analyzed with TRAVIS which yields five sorts of spectra, cf. Sommers *et al.*[29]: power, spherical, anisotropic, orthogonal, and parallel. As the measured spectra have been recorded on a backscattering Raman spectrometer (see below), we continue with the parallel

spectra.

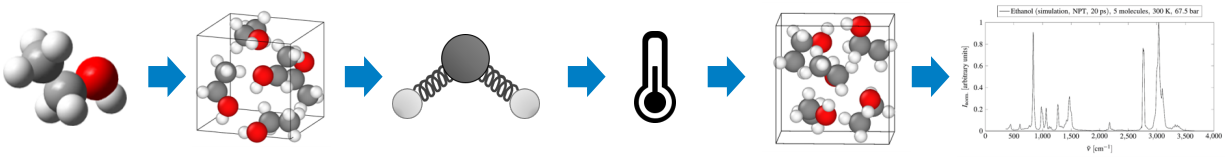


Figure 2: Sequence of required calculations to obtain AIMD results for use in the AIMD-IHM combination.

For the DFT calculations, we chose the functional BLYP with a double-ζ valence polarizable basisset (DZVP-MOLOPT-SR-GTH) as implemented in the CP2K software[28]. The recent benchmark by Taherivardanjani *et al.* revealed that double-ζ basis sets yield overall satisfactory results for liquid-phase frequencies of methanol[30]. For BLYP, Irikura *et al.*[31] report (for basis sets of at least 6-31G(d) size) an accuracy in vibrational frequencies of about 2.5 %. Similarly, Merrick *et al.*[32] report RMSEs for BLYP frequencies of 40 cm^{-1} to 47 cm^{-1}. Therefore, we regard this functional to yield a good compromise between computational effort and accuracy. Further details regarding the AIMD computation settings are given in the Supplementary Information (SI). We also investigated different calculation schemes and durations to determine suitable simulation times and system sizes, which is also reported in the SI. Our simulations have been performed at liquid densities at room temperature. We calculated the densities by interpolating the experimental values of Iglesias *et al.*[33] who report molar excess volumes for the ternary acetone-methanol-ethanol mixture. The simulations yield spectra including anharmonic and solvation effects[17]. We consider these spectra as valuable prior information for the IHM routine and superior to "static" *ab initio* ideal gas phase calculations (as used by *e.g.* Moores *et al.*[34]). Already in the gas phase, such harmonic oscillator (HO) frequencies suffer from the neglect of anharmonicity. In the liquid phase, peaks present in the gas phase may disappear and new peaks may arise due

to intermolecular interactions. The consideration of such effects is especially important for hydrogen-bonding mixtures such as the mixture system investigated in this work.

Depending on the availability of PCSa and calibration spectra, two challenges for spectral evaluation can be targeted by AIMD-IHM. To derive experimentally inaccessible PCSa, we present the "missing-PCS" approach, whereas "aixcalibration" (*ab initio*-extended calibration) enables to derive calibration coefficients without using experimental calibration mixture spectra. In practice, one may face either one of these challenges or both jointly. Hence, missing-PCS and aixcalibration can be applied either individually or together. From the AIMD simulations, IR spectra can also be calculated as shown by Thomas *et al.*, and IHM was repeatedly used for the quantitative analysis of IR spectra as well[13,35]. We therefore presume that our AIMD-IHM approach should in principle also be applicable to IR spectroscopy.

In missing-PCS, sim-PCSa are calculated by the described AIMD routine (cf. 2). However, previous work (own preliminary studies and *e.g.* Thomas *et al.*[15]) has shown that the sim-PCSa usually cannot directly be used for the quantitative analysis of experimental mixture spectra. Peak positions, as well as intensity ratios of the sim-PCSa differ significantly from the experimental spectra. The spectral modeling according to Eq. 2 only results in reliable weights, and thus ultimately compositions, if the peak parameters of the underlying PCMs accurately model the mixture spectra[36]. Nevertheless, the resemblance between simulated and experimental spectra is commonly high enough in order to derive the number and approximate positions, widths, and intensities of peaks from the sim-PCSa. This information from the simulated spectra is used in the missing-PCS approach to generate adjusted pure-component models (adj-PCMs) that can be used to accurately model the experimental spectra.

The missing-PCS approach is depicted in Fig. 3a: initially, the sim-PCSa are modeled with PCMs according to Eq. 1. The resulting PCMs are concatenated in a mixture model defined by the peak parameters $\boldsymbol{\Theta}$. From this starting point, a two-step iterative process for deriving

the adj-PCMs is carried out. In the first step, the mixture model is fitted to at least two experimental mixture spectra (with different compositions) using Eq. 2 (adjusting only the weights w_i). In the second step, the peak parameters $\boldsymbol{\Theta}$ of the PCMs in the mixture model are adjusted in order to optimize the spectral fit of the mixture model to the experimental mixture spectra. Subsequently, steps one and two are repeated until a minimum in the RMSE of the spectral fit is reached. This bilevel minimization can be written as:

$$\boldsymbol{\theta} = \arg\min_{\boldsymbol{\theta}} \overline{RMSE}_{\mathrm{Fit}}(\theta, \boldsymbol{w}, ...) \tag{5}$$

Within the minimization, the peak parameters in $\boldsymbol{\Theta}$ are optimized within specific boundaries around their initial values (derived directly from the AIMD spectra). This is justified by the uncertainty present in the AIMD simulations. In the literature, for gas-phase harmonic frequencies, 1σ-uncertainties between $40\,\mathrm{cm}^{-1}$ to $50\,\mathrm{cm}^{-1}$ (dependent on method and basis set) are reported[31,32]. For liquid-phase frequencies, no systematically obtained uncertainties of *ab initio* methods have been reported. However, we assume additional uncertainties due to more complex molecular interactions compared to the gas phase. Therefore, we decided to constrain the position shift to $\pm 100\,\mathrm{cm}^{-1}$. Regarding the other peak parameters, benchmark data is very scarce. Consequently, we only apply some physically motivated criteria. Peak widths are constrained to positive values with maximal full widths at the half maximum of $500\,\mathrm{cm}^{-1}$ to prevent the modeling of broadband background signals by Raman peaks. Peak intensities are constrained to positive values. In general, the peak parameters are additionally constrained by the fact that the same PCMs must be valid for all mixtures in the above described bilevel minimization (Eq. 5). Hence, all mixture spectra are modeled with the same mixture model, where only the component weights can vary (in this study, to demonstrate the mere effect of combining AIMD and IHM, we have not considered non-linear, *i.e.* concentration-dependent peak shifts).

In contrast to alternative methods that identify PCSa directly from experimental mixture

spectra[9–13], the missing-PCS approach can identify and assign hidden peaks correctly in spectral regions where peaks of different components overlap strongly, cf. Fig. 5.

Aixcalibration aims at deriving calibration coefficients directly from AIMD simulations for systems, where calibration coefficients cannot be obtained experimentally. This can be the case when *e.g.* compounds can react with each other and the equilibrium constant is unknown. For instance, certain compounds are only stable in a mixture, like *e.g.* the sulfate and bisuflate ions in æqueous sulfuric acid. The workflow of aixcalibration is depicted in Fig. 3b. First, PCSa, as well as mixture spectra, are simulated *via* our AIMD approach, cf. Fig. 2. The sim-PCSa represent the basis for setting up simulated PCMs (according to Eq. 1), which are then used to model the simulated mixture spectra (according to Eq. 2). As the compositions of the simulated mixture spectra are known, the calibration coefficients are calculated according to Eq. 3.

To evaluate the accuracy of AIMD-IHM and to provide mixture data for missing-PCS, we recorded spectra of the ternary hydrogen-bonding acetone-methanol-ethanol system. We carried out all measurements on a Raman backscattering microscope (Renishaw inVia, $\lambda_{\mathrm{Laser}} = 532\,\mathrm{nm}$, $P(\mathrm{cw}) = 42\,\mathrm{mW}$) detecting the spectral "fingerprint" region from $396\,\mathrm{cm}^{-1}$ to $2107\,\mathrm{cm}^{-1}$. Mixture spectra (Fig. 4b) of 10 different compositions (gravimetrically prepared, see Fig. 4a), as well as experimental pure-component spectra (exp-PCSa) of all components (for comparison purposes only) were recorded. Detailed experimental conditions are given in the SI.

To validate AIMD-IHM, we will first validate the missing-PCS and aixcalibration approaches separately by evaluating the resulting adj-PCMs and calibration coefficients, respectively. Afterwards, we will combine both approaches to predict compositions of the measured exemplary Raman spectra without using any exp-PCSa or calibration spectra.

In the following, we will apply the missing-PCS approach to derive adj-PCMs using sim-PCSa and the recorded mixture spectra without using any of the exp-PCSa. Fig. 5 shows the Raman spectral "fingerprint" region for a 1:1:1 mixture of acetone, methanol, and ethanol

along with the sim-PCSa (5a) and the adj-PCMs (5b). From Fig. 5a it can be seen that all peaks present in the experimental mixture spectra have corresponding, slightly shifted peaks in the simulated spectra. In some regions, *e.g.* around 1700 cm^{-1}, only one of the sim-PCSa (acetone) yields a peak, resulting in a relatively unambigous assignment. In other regions, *e.g.* around 1450 cm^{-1}, peaks of all components overlap. This overlapping causes challenges for a correct peak assignment. To find adequate adj-PCMs, first the sim-PCSa (cf. colored spectra in Fig. 5a) are modeled with pseudo-Voigt peaks to create simulation-based PCMs. Secondly, using these simulation-based PCMs as an initial guess, the model parameters are adjusted in the bilevel optimization (Eq. 5) based on the 10 measured mixture spectra (not considering any information on the concentrations) to create the final adj-PCMs. Fig. 5b demonstrates that the adj-PCMs accurately model both the unambigous spectral regions as well as the complex overlapping region around 1450 cm^{-1}. The two peaks below 1000 cm^{-1} are correctly assigned to acetone and ethanol only. The methanol peak in turn is correctly ascribed to the mixture peak near 1050 cm^{-1}. Furthermore, the strongly overlapping region at 1450 cm^{-1} is correctly decomposed into the contributions of the three components.

Fig. 6 shows the resulting PCSa in comparison to exp-PCSa derived *via* a conventional IHM approach based on experimental PCSa. Intensities, widths, and positions of adj-PCM peaks match the experimental data very well. For example, the average deviation of the peak positions of the adj-PCMs (from missing-PCS) in comparison to the peak positions of the exp-PCSa amounts to 2.7 cm^{-1} for acetone, 1.2 cm^{-1} for methanol, and 2.1 cm^{-1} for ethanol. It is important to note that these exp-PCSa do not at all enter the optimization within the missing-PCS scheme and are just plotted in Fig. 6 for the sake of comparison.

To evaluate the quality of the adj-PCMs, we generate a calibration model based on these adj-PCMs and the experimental mixture spectra (now considering the known compositions) according to Eq. 2, Eq. 3, and 4. The resulting $RMSECV_{\mathrm{Calib}}$ averaged for all three components is 0.67 % and proves the excellent quality of the PCMs resulting from the missing-

PCS scheme.

For comparison, we calibrated the mixture system with our recently published Method for Automatic Generation of Indirect Hard Models using crossvalidation (MAGIC)[36]. Modeling the experimental calibration spectra with these experiment-based PCMs led to a $RMSECV_{\mathrm{Calib}}$, averaged for all three components, of 0.46 %. Hence, for this multi-component mixture of species with partly overlapping peaks, missing-PCS proves to be a valuable tool to derive PCSa from AIMD simulations requiring solely mixture spectra of unknown compositions.

In the following, we will validate the aixcalibration approach. Using the aixcalibration methodology described above, we use sim-PCSa and simulated mixture spectra in order to derive calibration coefficients for the exemplary mixture system of acetone, methanol, and ethanol. Fig. 7 compares calibration coefficients obtained from aixcalibration with calibration coefficients obtained from using exp-PCSa and experimental mixture spectra following the conventional IHM approach outlined above. Both calibrations were carried out using MAGIC[36]. Calibration coefficients for acetone are set to unity since one factor is irrelevant due to the closure constraint for molar fractions. Calibration coefficients for the two alcohols deviate less than 4 % from experimentally obtained calibration coefficients. Aixcalibration thus provides calibration coefficients solely from simulated spectra that can be used directly for the analysis of experimental spectra.

The 95 % confidence interval represents the deviation of calculated compositions from the ideal linear correlation. This confidence interval is significantly larger for the calibration coefficients derived from aixcalibration. This is due to the higher variance in the simulations (cf. Fig. 2–4 in SI). According to our pilot survey for method selection (cf. SI) and the recent benchmark by Taherivardanjani *et al.*[30], the variance decreases with simulation time, system size, and number of mixtures investigated. Hence, there exists a trade-off between calibration coefficient uncertainty and computational effort that has to be found for each application scenario.

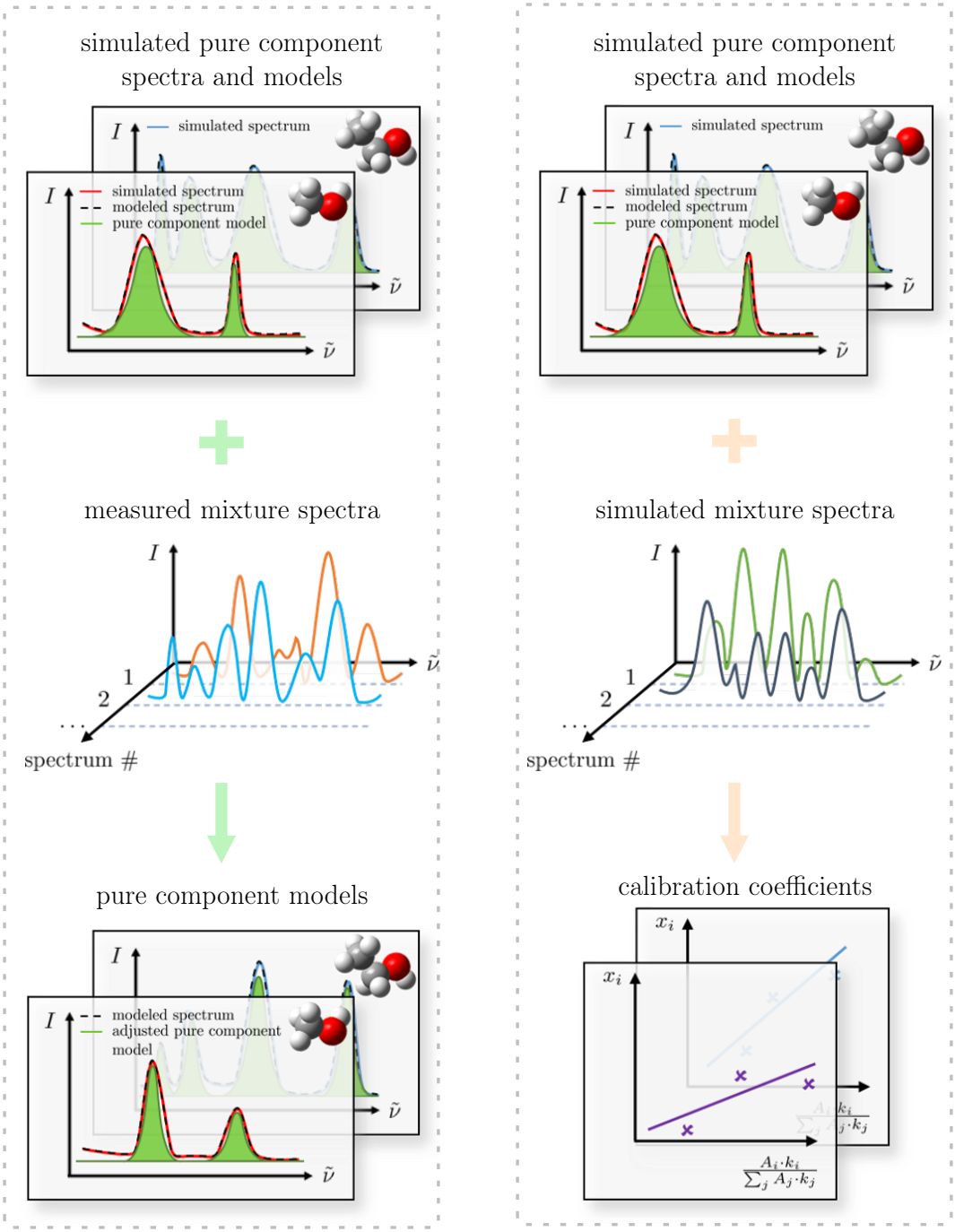
Finally, we applied the combination of missing-PCS and aixcalibration (blue arrows in the flowchart from Fig. 2) to the above described experimental mixture spectra of unknown composition. This is necessary when neither the composition of the measured mixtures nor the PCSa are experimentally available. The resulting PCMs from the missing-PCS approach are used to model measured mixture spectra, and the calibration coefficients from aixcalibration are used to finally obtain the mixture compositions. The resulting calculated molar fractions $\boldsymbol{x}_{\mathrm{calc}}$ compared to the known experimental molar fractions $\boldsymbol{x}_{\mathrm{real}}$ are shown in Fig. 8. The molar fractions for methanol show the lowest deviation with a $RMSECV_{\mathrm{Calib}}$ of 1.13 %. The calculated molar fractions for ethanol and acetone show higher deviations with $RMSECVs_{\mathrm{Calib}}$ of 4.30 % and 4.81 %, respectively. As can be seen in Fig. 8, there is a systematic bias in all mixtures in overestimating the molar fraction of ethanol while underestimating the molar fraction of acetone. This leads to an averaged $RMSECV_{\mathrm{Calib}}$ for all three components of 3.41 %. This is significantly higher than the values from conventional IHM, missing-PCS, or aixcalibration applied independently. However, considering that in the "full" AIMD-IHM approach with combined missing-PCS and aixcalibration no experimental calibration data has been used, the result of prediction errors less than 5 % is still very promising.

In conclusion, our AIMD-IHM combination allows for the prediction of PCSa and calibration coefficients for mixture systems that are (partly) unavailable for quantitative vibrational spectroscopy methods. In this work, we showed the application to Raman spectroscopy. Applications to other techniques such as IR and NMR spectroscopy are possible and will be evaluated in the future. We validated our new approach on Raman measurements for the acetone-methanol-ethanol system. The missing-PCS approach allows obtaining PCSa of excellent quality despite strongly overlapping peaks based on a few mixture spectra alone. In the acetone-methanol-ethanol system, predicted and experimental peak positions agree within $2\,\mathrm{cm}^{-1}$. Aixcalibration computes calibration coefficients also for mixtures of unknown composition (*e.g.* reactive systems, decomposing ILs). Calibration coefficients obtained this

way deviated less than 4 % from experimentally obtained calibration coefficients.

The combined use of missing-PCS and aixcalibration (where no experimental calibration data at all was used) led to predicted molar fractions that deviate less than 5 % from known compositions. After having validated the new AIMD-IHM method in this work, we will in the next step analyze reactive mixtures, *e.g.* æqueous sulfuric acid, which are so far truly not accessible for the analysis with Raman spectroscopy.

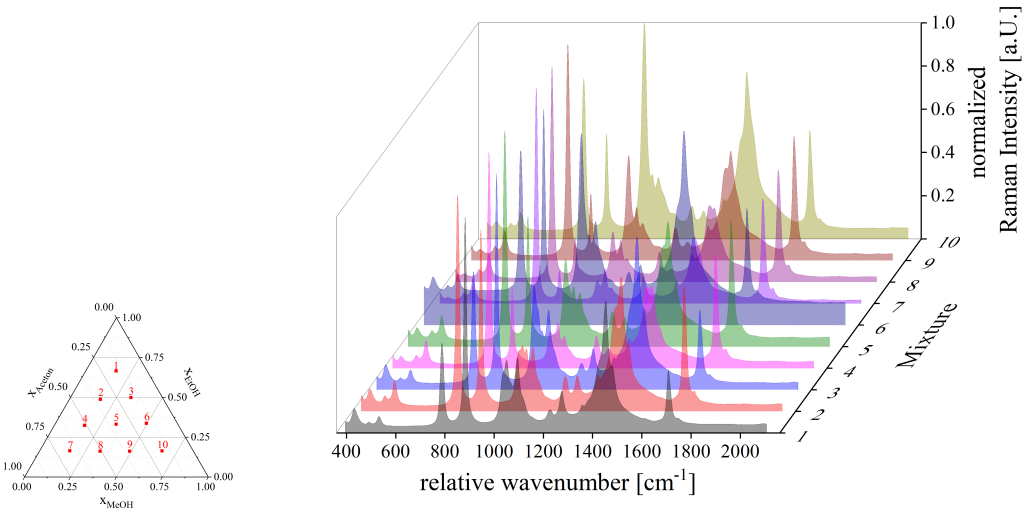
Combined with a reactive trajectory analyzer as ChemTraYzer[37,38], the AIMD-IHM tool can be extended even further to analyze systems where the number and type of intermediate and product species may be unknown, *e.g.* thermally decomposing ILs[22], intermediates of *in operando* catalysis investigations[5–7], and metabolites[8]. Larger, more complex systems can be treated by DFT- and MD-acceleration techniques or by using machine-learning potentials[39]. Hence, AIMD-IHM shows great potential for the quantitative spectral analysis of mixture systems that have so far been experimentally inaccessible.



(a) The missing-PCS approach determines unknown PCMs from AIMD initial information on PCs and (usually measured) mixture data.

(b) Aixcalibration determines calibration coefficients needed for composition analysis solely from simulation data: AIMD initial information on pure components and simulations of mixtures

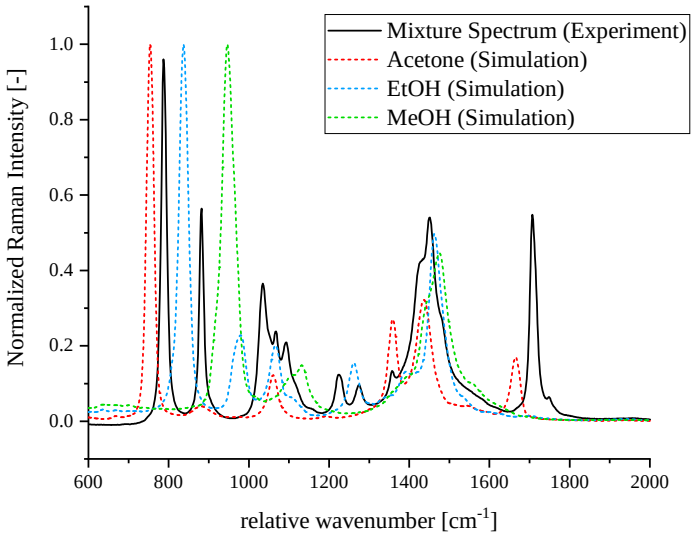
Figure 3: Overview of missing-PCS (3a) and aixcalibration (3b), two steps within the AIMD-IHM approach.



(a) Compositions of measured mixtures

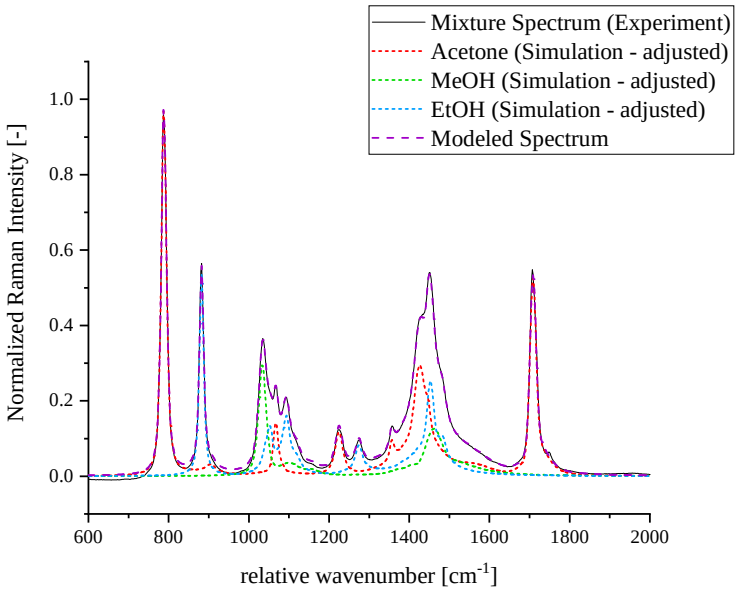
(b) Spectra, numbers on the z-axis correspond to the numbers in Fig. 4a.

Figure 4: Overview of the 10 compositions of the measured acetone-ethanol-methanol mixtures and their respective Raman spectra.



;

(a) before model adjustment



;

(b) after model adjustment

Figure 5: missing-PCS for 1:1:1 acetone-ethanol-methanol mixture: Comparison of the fingerprint region of an equimolar experimental mixture spectrum with sim-PCSa that are used for generating starting values for Eq. 5.

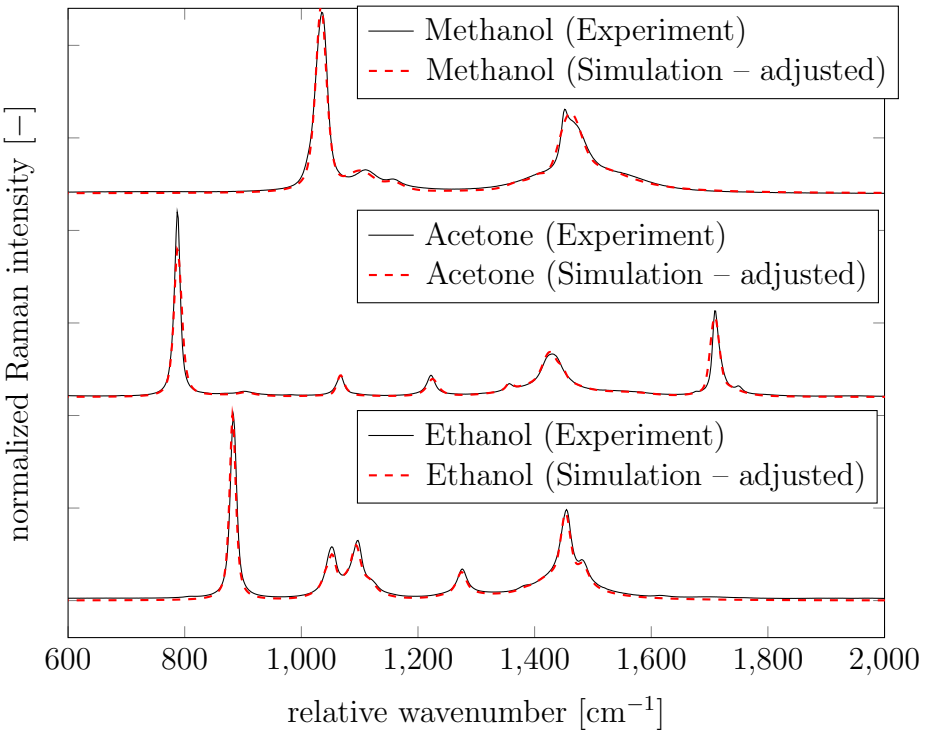


Figure 6: missing-PCS for acetone-methanol-ethanol mixtures: resulting PCMs from Eq. 5 using the sim-PCSa from Fig. 5a in order to generate initial values and the 10 mixture spectra of acetone-methanol-ethanol in comparison to exp-PCSa.

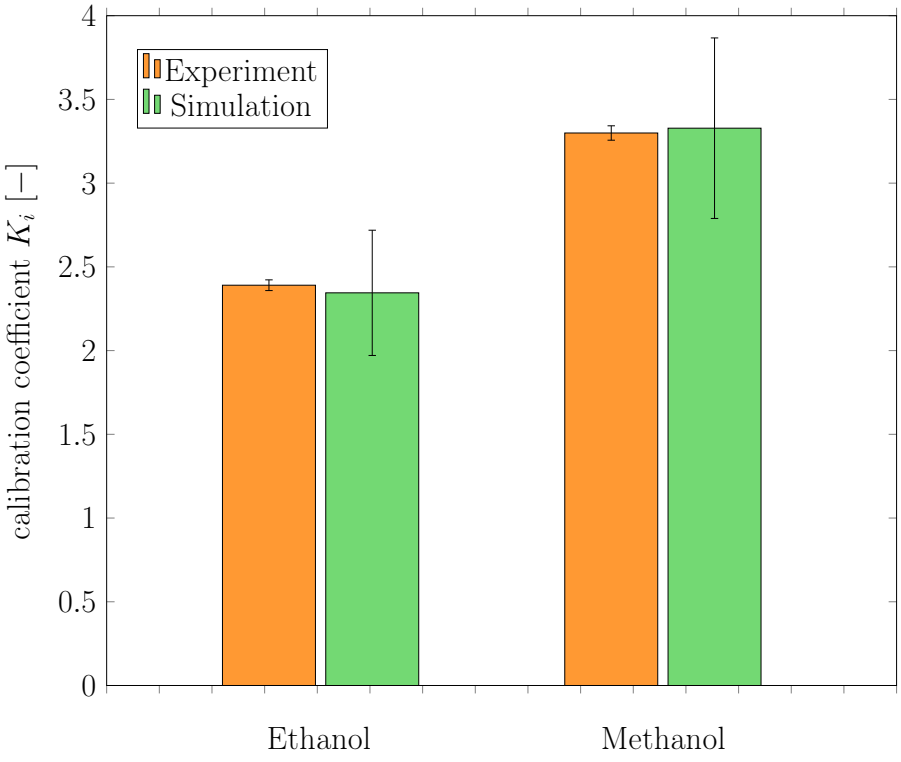


Figure 7: Aixcalibration for acetone-methanol-ethanol mixtures: resulting calibration coefficients according to Eq. 3 using only simulated spectra (Simulation) and using only experimental spectra (Experiment) including the 95 % confidence interval indicated by the error bars. Acetone is taken as reference compound and therefore per definition has $K = 1$ both for simulation and experiment.

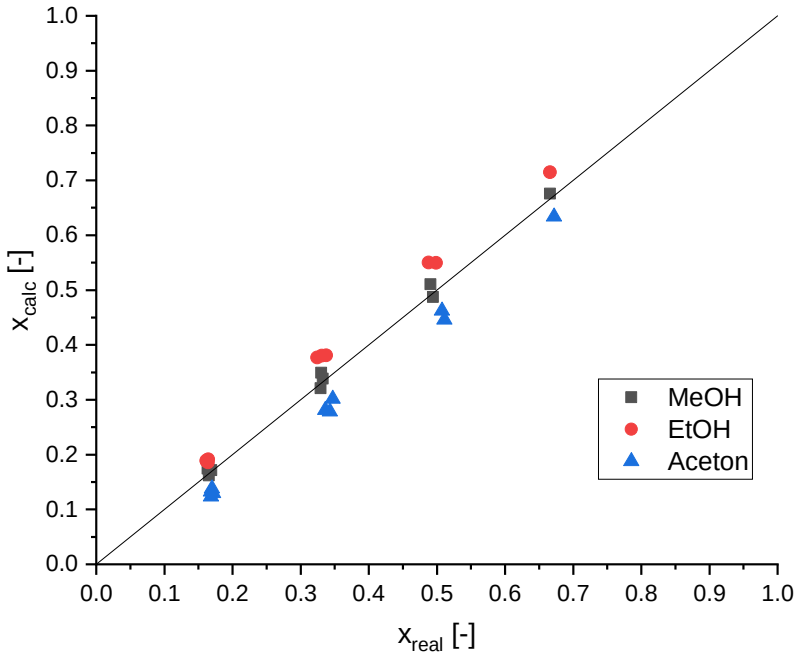


Figure 8: Subsequent application of missing-PCS and aixcalibration to acetone-methanol-ethanol mixtures: calculated mole fractions of the 10 mixture spectra of acetone-methanol-ethanol according to Eqs. 2 and 4 in comparison to the known, gravimetrically measured mole fractions using the PCMs from Fig. 6 and the calibration coefficients from Fig. 7.

Acknowledgement

The authors thank students Nico Henn, Felix Melzer, Jonas Bühner and Robert Göllinger. Simulations were performed with computing resources granted by RWTH Aachen University under project rwth0070. The present work has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – "Cluster of Excellence 2186 The Fuel Science Center" – ID: 390919832. Wassja A. Kopp acknowledges funding by DFG for project 407078203.

Supporting Information Available

The following files are available free of charge:

- SI-AIMDIHM.pdf: detailed description of experimental and computational methods and of results of AIMD simulations

Bibliography

(1) Martens, H. *Multivariate calibration*; John Wiley & Sons, 1992.

(2) Alsmeyer, F.; Koß, H.-J.; Marquardt, W. Indirect Spectral Hard Modeling for the Analysis of Reactive and Interacting Mixtures. *Appl. Spectrosc.* **2004**, *58*, 975–985.

(3) Clough, M. T.; Geyer, K.; Hunt, P. A.; Mertes, J.; Welton, T. Thermal decomposition of carboxylate ionic liquids: trends and mechanisms. *Phys. Chem. Chem. Phys.* **2013**, *15*, 20480–20495.

(4) Zaitsau, D. H.; Abdelaziz, A. The study of decomposition of 1-ethyl-3-methylimidazolium bis(trifluoromethylsulfonyl)imide by using Termogravimetry: Dissecting vaporization and decomposition of ILs. *J. Mol. Liq.* **2020**, *313*, 113507.

(5) Li, G.; Hu, D.; Xia, G.; Conrad Zhang, Z. Methanol Partial Oxidation on MoO3/SiO2 Catalysts: Application of Vibrational Spectroscopic Imaging Techniques in a High Throughput Operando Reactor. *Top. Catal.* **2009**, *52*, 1381–1387.

(6) Rabeah, J.; Bentrup, U.; Stößer, R.; Brückner, A. Selective Alcohol Oxidation by a Copper TEMPO Catalyst: Mechanistic Insights by Simultaneously Coupled Operando EPR/UV-Vis/ATR-IR Spectroscopy. *Angew. Chem. Int. Ed.* **2015**, *54*, 11791–11794.

(7) Yang, Y.; Xiong, Y.; Zeng, R.; Lu, X.; Krumov, M.; Huang, X.; Xu, W.; Wang, H.; DiSalvo, F. J.; Brock, J. D. et al. Operando Methods in Electrocatalysis. *ACS Catal.* **2021**, *11*, 1136–1178.

(8) Häckl, M.; Tauber, P.; Schweda, F.; Zacharias, H. U.; Altenbuchinger, M.; Oefner, P. J.; Gronwald, W. An R-Package for the Deconvolution and Integration of 1D NMR Data: MetaboDecon1D. *Metabolites* **2021**, *11*, 452.

(9) Windig, W.; Markel, S. Simple-to-use interactive self-modeling mixture analysis of FTIR microscopy data. *J. Mol. Struct.* **1993**, *292*, 161–170.

(10) Chen, J.; Wang, X. Z. A New Approach to Near-Infrared Spectral Data Analysis Using Independent Component Analysis. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 992–1001, PMID: 11500115.

(11) Chew, W.; Widjaja, E.; Garland, M. Band-Target Entropy Minimization (BTEM):? An Advanced Method for Recovering Unknown Pure Component Spectra. Application to the FTIR Spectra of Unstable Organometallic Mixtures. *Organometallics* **2002**, *21*, 1982–1990.

(12) de Juan, A.; Tauler, R. Chemometrics applied to unravel multicomponent processes and mixtures: Revisiting latest trends in multivariate resolution. *Analytica Chimica Acta* **2003**, *500*, 195–210, ANALYTICAL HORIZONS - An International Symposium celebrating the publication of Volume 500 of Analytica Chimica Acta.

(13) Kriesten, E.; Mayer, D.; Alsmeyer, F.; Minnich, C.; Greiner, L.; Marquardt, W. Identification of unknown pure component spectra by indirect hard modeling. *Chemom. Intell. Lab. Syst.* **2008**, *93*, 108–119.

(14) Lund Myhre, C. E.; Christensen, D. H.; Nicolaisen, F. M.; Nielsen, C. J. Spectroscopic Study of Aqueous H2SO4 at Different Temperatures and Compositions:? Variations in Dissociation and Optical Properties. *J. Phys. Chem. A* **2003**, *107*, 1979–1991.

(15) Thomas, M.; Brehm, M.; Fligg, R.; Vöhringer, P.; Kirchner, B. Computing Vibrational Spectra From Ab Initio Molecular Dynamics. *Phys. Chem. Chem. Phys.* **2013**, *15*, 6608–6622.

(16) Thomas, M.; Brehm, M.; Kirchner, B. Voronoi dipole moments for the simulation of bulk phase vibrational spectra. *Phys. Chem. Chem. Phys.* **2015**, *17*, 3207–3213.

(17) Brehm, M.; Thomas, M.; Gehrke, S.; Kirchner, B. TRAVIS—A free analyzer for trajectories from molecular simulation. *J. Chem. Phys.* **2020**, *152*, 164105.

(18) Brehm, M.; Kirchner, B. TRAVIS - A Free Analyzer and Visualizer for Monte Carlo and Molecular Dynamics Trajectories. *J. Chem. Inf. Model.* **2011**, *51*, 2007–2023, PMID: 21761915.

(19) Luber, S.; Iannuzzi, M.; Hutter, J. Raman spectra from ab initio molecular dynamics and its application to liquid S-methyloxirane. *J. Chem. Phys.* **2014**, *141*, 094503.

(20) Brehm, M.; Thomas, M. Computing Bulk Phase Raman Optical Activity Spectra from ab initio Molecular Dynamics Simulations. *J. Phys. Chem. Lett.* **2017**, *8*, 3409–3414, PMID: 28685571.

(21) Paschoal, V. H.; Faria, L. F. O.; Ribeiro, M. C. C. Vibrational Spectroscopy of Ionic Liquids. *Chem. Rev.* **2017**, *117*, 7053–7112, PMID: 28051847.

(22) Thomas, M.; Brehm, M.; Hollóczki, O.; Kelemen, Z.; Nyulászi, L.; Pasinszki, T.; Kirchner, B. Simulating the vibrational spectra of ionic liquid systems: 1-Ethyl-3-methylimidazolium acetate and its mixtures. *J. Chem. Phys.* **2014**, *141*, 024510.

(23) Dennington, R.; Keith, T. A.; Millam, J. M. GaussView Version 6. 2019; Semichem Inc. Shawnee Mission KS.

(24) Martiínez, L.; Andrade, R.; Birgin, E. G.; Martínez, J. M. Packmol: A package for building initial configurations for molecular dynamics simulations. *J. Comput. Chem.* **2009**, *30*, 2157–2164.

(25) Martínez, J. M.; Martínez, L. Packing optimization for automated generation of complex system's initial configurations for molecular dynamics and docking. *J. Comput. Chem.* **2003**, *24*, 819–825.

(26) Case, D.; Cerutti, D.; Cheatham, T.; III,; Darden, T.; Duke, R.; Giese, T.; Gohlke, H.; Goetz, A.; Greene, D. et al. Amber. 2017; University of California.

(27) Plimpton, S. J. Fast Parallel Algorithms for Short-Range Molecular Dynamics. *J. Comput. Phys.* **1995**, *117*, 1–19.

(28) Kühne, T. D.; Iannuzzi, M.; Del Ben, M.; Rybkin, V. V.; Seewald, P.; Stein, F.; Laino, T.; Khaliullin, R. Z.; Schütt, O.; Schiffmann, F. et al. CP2K: An electronic structure and molecular dynamics software package - Quickstep: Efficient and accurate electronic structure calculations. *J. Chem. Phys.* **2020**, *152*, 194103.

(29) Sommers, G. M.; Calegari Andrade, M. F.; Zhang, L.; Wang, H.; Car, R. Raman spectrum and polarizability of liquid water from deep neural networks. *Phys. Chem. Chem. Phys.* **2020**, *22*, 10592–10602.

(30) Taherivardanjani, S.; Elfgen, R.; Reckien, W.; Suarez, E.; Perlt, E.; Kirchner, B.

Benchmarking the Computational Costs and Quality of Vibrational Spectra from Ab Initio Simulations. *Adv. Theory Simul.* **2021**, *n/a*, 2100293.

(31) Irikura, K. K.; Johnson III, R. D.; Kacker, R. N. Uncertainties in Scaling Factors for ab initio Vibrational Frequencies. *J. Phys. Chem. A* **2005**, *109*, 8430–8437.

(32) Merrick, J. P.; Moran, D.; Radom, L. An evaluation of harmonic vibrational frequency scale factors. *J. Phys. Chem. A* **2007**, *111*, 11683–11700.

(33) Iglesias, M.; Orge, B.; Piñeiro, M. M.; de Cominges, B. E.; Marino, G.; Tojo, J. Thermodynamic Properties of the Ternary Mixture Acetone + Methanol + Ethanol at 298.15 K. *J. Chem. Eng. Data* **1998**, *43*, 776–780.

(34) Moores, M.; Gracie, K.; Carson, J.; Faulds, K.; Graham, D.; Girolami, M. Bayesian modelling and quantification of Raman spectroscopy. 2018.

(35) Viell, J.; Marquardt, W. Concentration Measurements in Ionic Liquid-Water Mixtures by Mid-Infrared Spectroscopy and Indirect Hard Modeling. *Appl. Spectrosc.* **2012**, *66*, 208–217.

(36) Woehl, J.; Meltzow, F.; Koß, H.-J. Method for Automatic Generation of Indirect Hard Models using crossvalidation (MAGIC) for the spectral analysis of mixture spectra. *Chemometr. Intell. Lab.* **2021**, *217*, 104419.

(37) Döntgen, M.; Przybylski-Freund, M.-D.; Kröger, L. C.; Kopp, W. A.; Ismail, A. E.; Leonhard, K. Automated discovery of reaction pathways, rate constants and transition states using reactive molecular dynamics simulations. *J. Chem. Theory Comput.* **2015**, *11*, 2517–2524.

(38) Döntgen, M.; Schmalz, F.; Kopp, W. A.; Kröger, L. C.; Leonhard, K. Automated Chemical Kinetic Modeling via Hybrid Reactive Molecular Dynamics and Quantum Chemistry Simulations. *J. Chem. Inf. Model.* **2018**, *58*, 1343–1355, PMID: 29898359.

(39) Raimbault, N.; Grisafi, A.; Ceriotti, M.; Rossi, M. Using Gaussian process regression to simulate the vibrational Raman spectra of molecular crystals. *New J. Phys.* **2019**, *21*, 105001.