Supporting Methods

For data evaluation we followed the methods of statistical data analysis as described in detail in ref. 10. The posterior probability $p(\text{model}\,|\,\text{data},\,I)$ that a model is able to fit the experimental data, given some initial information I on the model parameters is given by

$$p(\text{model}\,|\,\text{data},I) = \frac{p(\text{data}\,|\,\text{model},I)\times p(\text{model}\,|\,I)}{p(\text{data}\,|\,I)} \qquad \textbf{[2]}$$

where $p(\text{data}\,|\,\text{model},I)$ gives the likelihood function (e.g., least squares) that the data fit the model, $p(\text{model}\,|\,I)$ describes the prior probability that the model is valid, and $p(\text{data}\,|\,I)$ gives the probability that a certain data set can be measured given I. The prior probability describes our state of knowledge or ignorance on the problem (e.g., positiveness of model parameters).

In general, a model may contain several parameters $a_i,\ i = 1,\ldots n$, of which some ($a_l,\ l = 1,\ldots m$) are of interest and others, so called nuisance parameters ($a_k,\ k = m+1,\ldots n$), are not. By integrating the probability distribution over the nuisance parameters, a process called marginalization, one obtains a new probability distribution that depends only on the relevant parameters:

$$p(\{a_l\}) = \int_{-\infty}^{+\infty}\ldots\int_{-\infty}^{+\infty} p(\{a_i\})da_{m+1}\ldots da_n\,. \qquad \textbf{[3]}$$

The best estimates of these parameters and their uncertainties are then obtained in the usual way: They are given by the parameter values at the maximum of $p(\{a_l\})$ and the inverse of the covariance matrix at that point,

$$[\sigma^2]_{ij} = -[\{\nabla\nabla\log[p(\{a_l\})]\}^{-1}]_{ij}\,. \qquad \textbf{[4]}$$

In our particular case the detector signal $I(\nu,\ d)$ at a certain frequency, ν, and a water layer thickness, d, is described by

$$I(d,\ \nu) = I_{0,\ \nu}\exp(-\alpha(\nu)d) + C \qquad \textbf{[5]}$$

where $I_{0,\ \nu}$, $\alpha(\nu)$, d, and C correspond to the terahertz signal without probe, the absorption coefficient of the probe, the layer thickness of the probe, and the detector offset, respectively. Only α is of interest, so $I_{0,\ \nu}$ and C may be treated as nuisance parameters. Because the model above is linear in both nuisance parameters, the integration can be performed analytically.

A summary of all absorption coefficients with their error bars which were determined following this approach at a frequency of 80 cm^{-1} is shown in Fig. 6. In total, > 100 measurements were carried out over a frequency range from 73 to 91 cm^{-1}.

Equ. **2** may now be used to quantitatively compare how well different models with e.g., a different number of parameters describe a certain data set. It is clear that a larger number of parameters will give a better fit in the least-squares sense. However, are these just fitting parameters or do the additional parameters contain new information?

Consider two models A and B that are used to describe a given data set. The ratio of the posterior probabilities that the models fit the data is given by

$$\frac{p(A|D,I)}{p(B|D,I)} = \frac{p(D|A,I)}{p(D|B,I)} \times \frac{p(A|I)}{p(B|I)}\,. \qquad \textbf{[6]}$$

Under the assumptions that we can estimate a reasonable parameter range for λ

$$\lambda_{\min} \leq \lambda \leq \lambda_{\max}, \qquad \textbf{[7]}$$

that we have no *a priori* knowledge on λ so that the according prior is uniform between $\lambda_{\min}$ and $\lambda_{\max}$ and zero elsewhere and that the probability distribution

$$p(D|\lambda,B,I) = p(D|\lambda_0,B,I) \times \exp(-\frac{(\lambda-\lambda_0)^2}{2\delta\lambda^2}) \qquad \textbf{[8]}$$

is well approximated by a Gaussian probability distribution function, we can use marginalization to finally get

$$\frac{p(A|D,I)}{p(B|D,I)} = \frac{p(A|I)}{p(B|I)} \times \frac{p(D|A,I)}{p(D|\lambda_0,B,I)} \times \frac{\lambda_{\max}-\lambda_{\min}}{\delta\lambda\sqrt{2\pi}}\,. \qquad \textbf{[9]}$$

This equation can be easily generalized to cases of several variables.

In our particular case we want to compare the two-component model A with parameters $\{a_i\}$ and the three-component model B with parameters $\{b_i\}$, including two additional parameters, the effective radius $r_{\rm solv}$ and the absorption coefficient $\alpha_{\rm solv}$ of the solvation shell, with our measurements of the overall absorption coefficient versus frequency. Because the different measurements are statistically independent,

$$p(A|D,I) \propto \prod_{i=1}^{N} \exp(-\frac{(D_i - A(\{a_k\})_i)^2}{\sigma_i^2}), \qquad \textbf{[10]}$$

where D_i is the ith data point, and $A(\{a_k\})_i$ and σ_i are the corresponding model value and measurement uncertainty, respectively. A similar equation holds for model B. The best fit parameters for both models are given by the sets of parameter values $\{a_{k,0}\}$ and $\{b_{k,0}\}$.

For the comparison of the two component and three component models presented in this paper we used an uncertainty range 0 $\mathrm{cm}^{-1} \leq \alpha_{\mathrm{solv}} \leq 1{,}000\ \mathrm{cm}^{-1}$ for the absorption coefficient of the solvation shell and 0 Å $\leq r_{\mathrm{solv}} \leq 20$ Å for its radius. The probability ratio is then given by

$$\frac{p(B|D,I)}{p(A|D,I)} = \exp(-[\sum_{i=1}^{N} \frac{(D_i - B(\{b_{k,0}\})_i)^2}{\sigma_i^2} - \sum_{i=1}^{N} \frac{(D_i - A(\{a_{k,0}\})_i)^2}{\sigma_i^2}]) \times \frac{\alpha_{\mathrm{solv,\,max}} - \alpha_{\mathrm{solv,\,min}}}{\delta\alpha_{\mathrm{solv}}\sqrt{2\pi}} \times \frac{r_{\mathrm{solv,\,max}} - r_{\mathrm{solv,\,min}}}{\delta r_{\mathrm{solv}}\sqrt{2\pi}}\,. \quad \textbf{[11]}$$

The average probability ratio of three-component model versus the two-component model at the investigated frequencies is 7.5×10^{14}.

---

[10] Sivia, D. (1996) *Data Analysis: A Bayesian Tutorial* (Clarendon Press, Oxford).