A spectral matching algorithm based on the Wasserstein metric

Klaus Neymeyr[a,b], Baoxin Zhang[b], Christoph Kubis[b], Lukas Prestin[a],
Tomass Andersons[a], Mathias Sawall[a], Jan Hellwig[a]

[a] *Universität Rostock, Institut für Mathematik, Ulmenstraße 69, 18057 Rostock, Germany*
[b] *Leibniz-Institut für Katalyse, Albert-Einstein-Straße 29a, 18059 Rostock, Germany*

Abstract

Through visual inspection, scientists can easily judge the similarity of pairs of spectra from different sources, whether they are experimental measurements, spectra libraries, or spectra calculations. Spectral similarity is recognized when the peak patterns appear similar, even if the peak positions are shifted or the peak amplitudes are changed. However, it is desirable to have an objective spectra distance measure that can be calculated in terms of a numerical distance number. This work introduces a move-and-scale spectral matching algorithm based on the Wasserstein metric as a distance measure between pairs of spectra. The focus is on situations where only part of a spectrum (e.g., the spectral signature of a functional group in a molecule) is matched to a specific window of a full-spectrum range with a more complex peak pattern. The spectral matching algorithm is used to assign calculated spectra by density functional theory (DFT) to pure component spectra based on the peak group analysis (PGA) of FTIR spectroscopic data sets from transition metal-catalyzed carbonylation reactions.

Key words: Multivariate curve resolution, Spectral matching, Wasserstein distance, Vibrational mode analysis

1. Spectra similarity

Let two spectra f and g be represented by vectors in the n-dimensional space $\mathbb{R}^n$. Their distance can be measured by evaluating a norm of the difference $f - g$, e.g., the Euclidean norm $\|f - g\|_2 = \sqrt{\sum_{i=1}^n (f_i - g_i)^2}$ or the 1-norm distance $\|f - g\|_1 = \sum_{i=1}^n |f_i - g_i|$. Another distance measure is the acute angle

$$\measuredangle(f, g) = \arccos \frac{|\sum_{i=1}^n f_i g_i|}{\|f\|_2 \|g\|_2}.$$

A statistical measure is the covariance, namely $\mathrm{cov}(f, g) = \mathrm{E}\left[(f - \mathrm{E}[f])(g - \mathrm{E}[g])\right]$ with the mean or expected value $\mathrm{E}[\cdot]$. If the spectra are shifted against each other, then these and other distance measures may not be useful, which can be easily understood by taking the standard basis vectors $f = e_1$ and $g = e_2$. These single-peak, single-channel spectra differ in a peak displacement by one channel index. Moreover, f and g are orthogonal vectors. The norm evaluation for $f - g = e_1 - e_2$ takes maximal values. The norm cannot even distinguish the minor shift from e_1 to e_2 compared to the maximal shift from e_1 to e_n since

$$\|e_1 - e_2\|_p = \|e_1 - e_n\|_p$$

for any p-norm, $p = 1, 2, \ldots, \infty$.

Hence, calculating spectra distances or spectra similarities may not work in this way, even when the peak pattern is very similar. A (partial) solution may consist of considering the cross correlation (or sliding dot product)

$$R_{f,g}(t_1, t_2) = \mathrm{E}[f(t_1) \cdot g(t_2)].$$

so that shifts $\tau = t_2 - t_1$ can be compensated. The shift is also called a lag or displacement. In signal processing, machine learning and convolutional neural networks, the cross correlation is a popular measure of similarity that can be implemented by a convolution [7]. See also other correlation measures as the Pearson correlation [25, 3, 29]. The cross correlation can work well to measure the spectra similarity if the peak displacement is more or less constant for the related pairs of peaks within the two spectra f and g. However, such an assumption cannot generally be justified for experimental spectra. What is needed is a distance measure for pairs of spectra even when the displacement of related peaks is not constant.

In general, the so-called *spectral matching problem* is defined as the challenge of identifying similar pairs of spectra from different origins. Important fields of application are hyperspectral or multispectral images where similar spectra for different pixels can help to identify materials or to detect material similarities. Such methods are used in agriculture, mining, satellite surveillance, medical imaging, pharmaceutical industry and others. One reference

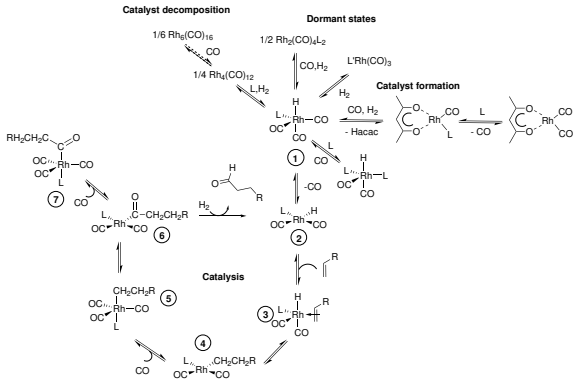

Figure 1: Simplified scheme of rhodium-catalyzed hydroformylation with the catalytic cycle (2-6), the product aldehyde (7), catalyst decomposition and dormant states above from (1) and catalyst formation to the right of (1), cf. [14, 34].

for spectral matching in hyperspectral satellite image processing is [30]. In chemistry, spectral matching can help to identify spectra belonging to the same chemical species. For instance, [26] reports on applications to Raman spectroscopy and automated match with a spectrum library. However, the application environments are often so different that there can hardly be a standardized method for spectral matching.

1.1. An application background in organometallic catalysis

The authors were motivated to consider the spectral matching problem through interdisciplinary collaboration aimed at understanding and optimizing the homogeneous catalyzed hydroformylation process of olefins, see, for example, references [15, 16, 4, 14, 17, 20]. Transition metal carbonyl complexes with changing further ligands operate in catalytic cycles and connected pathways. Specific complexes appear during catalyst formation, in dormant states and during catalyst decomposition, e.g., through the formation of polynuclear metal-carbonyl complexes. While the chemical feedstocks and the reaction products are present at high concentrations, the catalyst species occur at concentrations several orders of magnitude lower. A detailed understanding of the catalytic reactions is of high economic interest, as carbonylation reactions have a high mass throughput in chemical industrial production and related high costs of the catalyst. IR and Raman spectroscopy are ideally suited for analyzing the catalytic processes in real time [5]. In situ IR spectroscopic data for the monitoring and characterization of catalyst complexes and intermediates are obtained using high-pressure transmission IR flow cells. Figure 1 shows a simplified scheme of some of relevant steps of the homogeneously rhodium-catalyzed hydroformylation (oxo) process. This process transforms alkenes with syngas (CO/H_2) into aldehydes. The experimental work is accompanied by an IR spectroscopic data analysis using chemometric multivariate curve resolution methods. We use the Peak Group Analysis (PGA), which is very well-suited for the analysis of IR data from catalytic carbonylation processes. PGA provides a sequence of potential pure component spectra from the time series of mixture spectra as described in [27, 28]. The PGA results contain only a small degree of inherent factor ambiguity [24]. An in-depth understanding of the hydroformylation process requires subsequent structure elucidation steps of the molecular structures of identified catalytic species. Quantum mechanical methods using density functional theory (DFT) calculations are employed to identify thermodynamically favorable configurations of certain catalytically relevant organometallic transition metal complexes and to calculate their vibrational frequencies and band intensities in the carbonyl region. However, DFT often does not precisely reproduce the spectra of pure components as measured experimentally for pure chemical species (if available) or as output by a chemometric method. Depending on the chemical neighborhood, temperature, and pressure conditions in the experimental setup, as well as the restrictions of the quantum mechanical method (e.g., chosen functional, basis set, vibrational scaling factor), a considerable offset must be expected between the peak positions and the relative shifts of the peaks within a spectral window [1, 13]. This makes a spectral matching algorithm a highly valuable tool.

1.2. The spectral matching problem

Understanding the formation of a cobalt(II) biphosphine catalyst for the hydroformylation process [20] is a research challenge that leads us to consider the following generalized problem of finding the best spectral matches between three DFT spectra denoted by DFT1, DFT2 and DFT3, see Appendix A for the details on the DFT calculations, and three PGA spectra (PGA1, PGA2 and PGA3). These six spectra are shown in Fig. 2. The DFT spectra without applying a vibrational scaling factor cover the wavenumber range 2000–2300 cm^{-1} whereas the PGA spectra cover the wider range 1505–2150 cm^{-1} and have many more peaks. The dominant peaks of the PGA spectra, together with a few other peaks within an enclosing frequency window, exhibit some similarity to the DFT spectra.

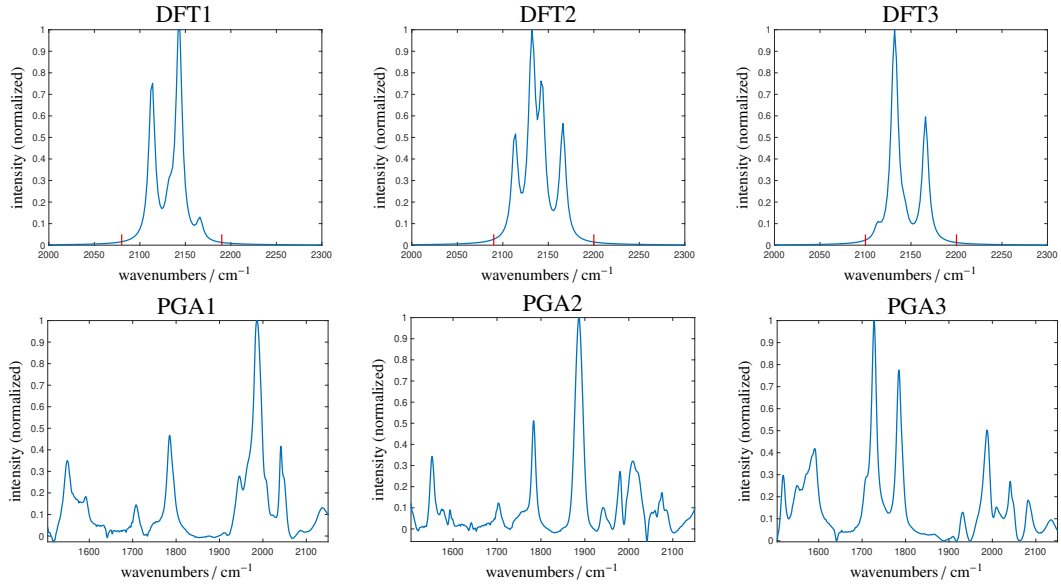

Figure 2: Top row: Three DFT-caculated IR spectra (on C=O vibration in Co(II) complexes). The short red lines indicate the windows used for spectral matching; see Sec. 3.2 for details. Lower row: Three approximate pure component spectra recovered by PGA from experimental FTIR mixture data. The aim is to find a one-to-one spectral matching.

However, there are a considerable wavenumber offset and a different spread of the wavenumber axis, as well as only similar peak height ratios. Solving the spectral matching problem for these spectra should provide objective measures of similarity.

1.3. Overview

Section 2 introduces the Wasserstein metric and explains its relation to optimal transport problems. Sec. 3 introduces an environment in which a move-and-scale algorithm can be used to solve the spectral matching problem with the Wasserstein metric. Demonstrations for sequences of DFT and PGA spectra accompany all algorithmic steps. Finally, Sec. 4 presents a second spectral matching problem from rhodium-catalyzed hydroformylation, along with its solution using the Wasserstein distance-based spectral matching algorithm with move-and-scale.

2. The Wasserstein distance

The Wasserstein distance, also known as Wasserstein metric, is a distance measure defined between probability density functions μ and ν, see Wasserstein [32] and Kantorovich [12] (English translation of the 1939 original in Russian). There are few applications of the Wasserstein metric to shape-matching problems and spectral similarity in chemistry, see [22, 19, 23]. To prepare the ground for the spectral matching problem, we consider the probability density functions to have finite support in the interval $[a, b] \subset \mathbb{R}$. Furthermore, let F_μ and F_ν be their cumulative distribution functions, namely

$$F_\mu(x) = P(\mu \leq x) \tag{1}$$

so that $F_\mu(x)$ is the probability of μ taking a value less than or equal to x. Then the Wasserstein distance of μ and ν is given by (see also [31])

$$d_W(\mu, \nu) = \int_a^b |F_\mu(x) - F_\nu(x)| dx. \tag{2}$$

The Wasserstein metric can be interpreted in terms as a solution to an optimal transport problem. It measures the cost of transforming the density function μ into the density function ν, where the density functions are imagined as piles of sand and the least amount of sand is moved the shortest distance. See Fig. 3 for an illustration and Villani [33] for more on the optimal transport problem. Instead of Eq. (2) one could also use

$$d_p(\mu, \nu) = \left(\int_a^b |F_\mu(x) - F_\nu(x)|^p dx \right)^{1/p}$$

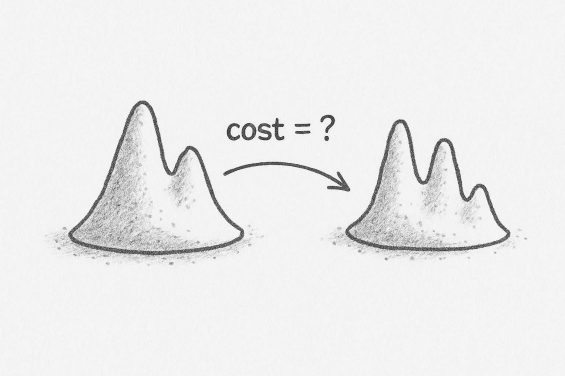

Figure 3: Two sand piles. The cost of transporting the sand to reshape the pile on the left into the pile on the right correlates with the similarity of the two piles. Very different piles correspond to high costs. (The image has been generated by ChatGPT.)

for any natural number p or the Kolmogorov-Smirnov distance ($p = \infty$)

$$d_{\mathrm{KS}}(\mu, \nu) = \sup_{x \in [a,b]} |F_\mu(x) - F_\nu(x)|.$$

However, the Kolmogorov-Smirnov distance is sensitive only to the largest difference, and distances measured by other values of p, except $p = 1$, are not interpretable in the sense of optimal transport theory. Therefore, we use $p = 1$ as given by the Wasserstein distance (2) .

The Wasserstein metric can also be evaluated for any pair of (integrable) nonnegative functions f and g on the same domain after normalization by the 1-norm. With this normalization, $\mu = f/\|f\|_1$ and $\nu = g/\|g\|_1$ are probability density functions. To this end, we evaluate

$$d_W(f/\|f\|_1, g/\|g\|_1),$$

which measures the similarity of the functions. Again, this distance can be interpreted as the transport cost of transforming one normalized function into the other normalized one. The descriptive interpretation of sand piles remains still valid.

The Wasserstein metric can also be applied to discrete functions represented by nonnegative vectors $f, g \in \mathbb{R}^n$. Let $f, g \geq 0$ satisfy the normalization constraints

$$1 = \|f\|_1 = \sum_{i=1}^{n} f_i, \qquad \text{and} \qquad 1 = \|g\|_1 = \sum_{i=1}^{n} g_i.$$

Furthermore, let $x \in \mathbb{R}^n$ be the associated vector of equidistant arguments of the discrete functions f and g and $\delta_x = x_2 - x_1$. Then the counterpart of the cumulative distribution function (1) is the cumulative sum of $f = (f_1, \ldots, f_n)^T \in \mathbb{R}^n$

$$\mathrm{cumsum}(f) = (f_1,\ f_1 + f_2,\ f_1 + f_2 + f_3, \ldots,\ f_1 + f_2 + \ldots + f_n).$$

and the discrete Wasserstein distance of f and g with the joint vector of arguments $x \in \mathbb{R}^n$ reads

$$d_W(f, g) = \delta_x \|\mathrm{cumsum}(f) - \mathrm{cumsum}(g)\|_1 = \delta_x \sum_{i=1}^{n} |(\mathrm{cumsum}(f))_i - (\mathrm{cumsum}(g))_i| \tag{3}$$

The Wasserstein metric could be implemented in Matlab as follows (the argument vector x is assumed to have equidistant values) :

```
function wst=wasserstein(x,f,g);
 f=f/norm(f,1);
 g=g/norm(g,1);
 wst=(x(2)-x(1))*norm(cumsum(f)-cumsum(g),1);
```

2.1. Distance to a shifted signal

To illustrate this distance measure, consider two identical asymmetrical triplet signal groups within the frequency interval $[0, 30]$. One group is centered at 10, and the other is centered at 12, see Fig. 4. Each peak is modeled by a Gaussian. The discrete representations of the two signal groups are given by vectors $f, g \in \mathbb{R}^{400}$.

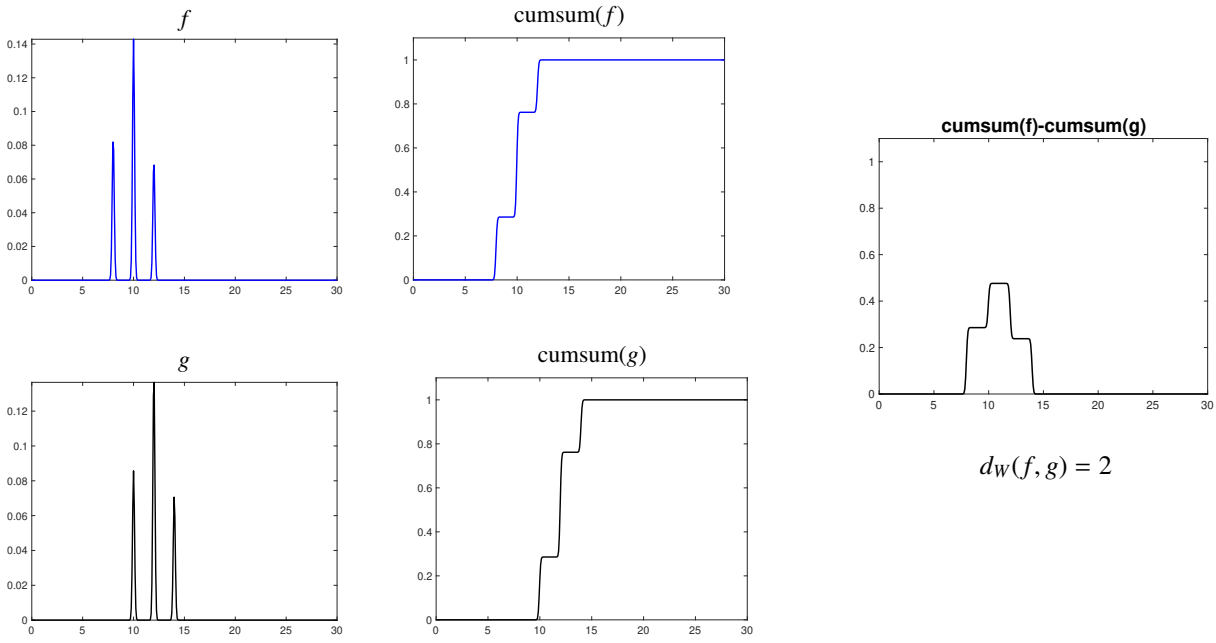

Figure 4: Left: Two identical asymmetrical triplet signals, but the second is shifted to higher frequencies by 2. Center: The cumulative sums of f and g. Right: Difference of the two cumulative sums. Its 1-norm equals the Wasserstein distance.

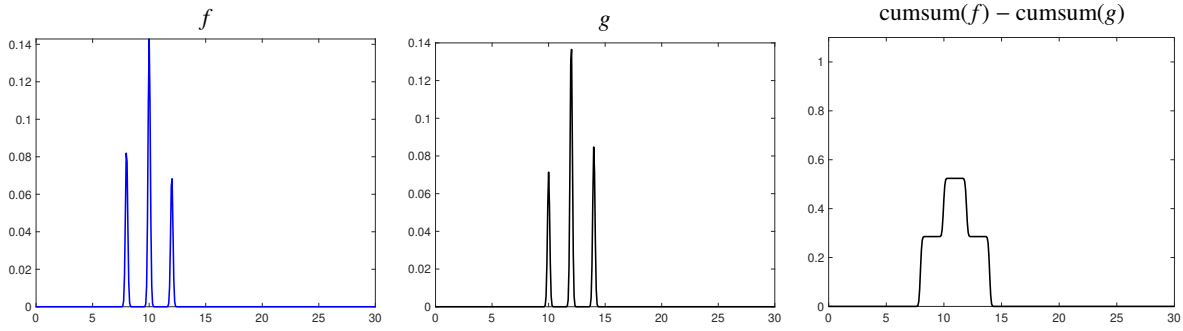

Figure 5: The two triplet signals, where g is shifted by 2 to higher frequencies and mirrored. Right: Difference of the cumulative sums of f and g whose 1-norm equals $d_W(f, g) = 2.1905$.

Calculating the 1-norm of the difference of the two cumulative sums yields the Wasserstein distance

$$d_W(f, g) = \delta_x \|\text{cumsum}(f) - \text{cumsum}(g)\|_1 = 2.$$

This is an easily interpretable result because transporting the normalized triplet f to the normalized triplet f (shifted by 2) has exactly the cost 2.

2.2. Distance to a shifted and flipped signal

Next, we study the Wasserstein distance for the example from Sec. 2.1, but with one of the asymmetrical triplet signals mirrored. The profiles f and g are shown in Fig. 5. The amplitude pattern of g has been flipped. The numerical evaluation of the Wasserstein distance results in $d_W(f, g) = 2.1905$. The distance measure works well and is sensitive to changes in the peak pattern, since the transport cost of moving f to the center argument 12 is 2. Furthermore, the small surplus in the left peak of f can be transported to the right peak of g at a cost of 0.1905. However, this transport cost is dominated by the cost of transporting the triplet to its shifted position. Changes in the peak amplitude pattern result in a much smaller transport cost.

Since the effect of shifting on d_W is much larger than correcting the mirrored peak pattern, shifts can make correct solution to the peak assignment problem impossible. Therefore, the next goal is to split off the impact of shifts in Wasserstein distance calculations.

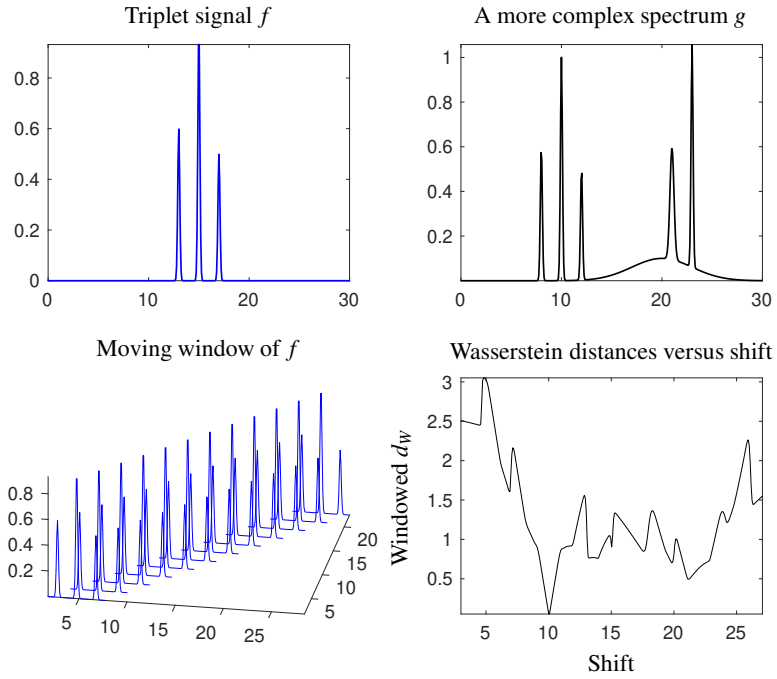

Figure 6: Upper row: A triplet signal f (left) and a more complex spectrum g (right). Lower row: The triplet signal window $[12, 18]$ of f is moved through the frequency range of g (left plot). The Wasserstein distance within the shifted window of f to g is plotted versus the shift parameter (lower right plot). The smallest distance is $\min d_{W,\text{window}} = 0.0524$. For $\sigma = 10$ the two triplets match. The non-smooth course of the Wasserstein distances curve is influenced by the fact that in the active window the restriction of g is always normalized with respect to the 1-norm as part of the Wasserstein distance calculation.

3. A move-and-scale Wasserstein distance measure

3.1. Moving window approach

As explained in Sec. 2.2, Wasserstein distances between pairs of similar spectra can be dominated by the displacement of signals or groups of signals along the frequency axis. In other words, it can be much more costly to transport a signal to a displaced position in another spectrum than it is to reshape one spectrum into a similar one locally. For the given spectral matching problem, the second type of cost is important. To split off the first type of cost, we define a window that encompasses the signals for which the spectral matching problem is to be solved. Then, we move the window through the entire spectral range of the second spectrum. For each shift, the Wasserstein distance is calculated only within the window.

This procedure is illustrated in Fig. 6. We start with a triplet signal f and a more complex spectrum g, which not only contains the displaced triplet, but also other peaks. The triplet f is contained within the spectral window $[12, 18]$. This window is then moved through the frequency range $[0, 30]$ of g. The process is started with the initial window $[12, 18]$ after moving its center to the far left of the full interval $[0, 30]$. We calculate the Wasserstein distances for each shift parameter by evaluating only within the shifted window. The curve of Wasserstein distances versus the shift parameter is shown in the lower-right subplot of Fig. 6. The minimum of this curve is $\min d_{W,\text{window}} = 0.0524$. This minimum is attained when the shift parameter is 10 and the center frequency of the triplet in f is initially fixed at the origin. All of this confirms a successful spectral matching of the two triplet signals.

As shown in the top row of Fig. 6, we note that the Wasserstein distance of the pure triplets f and g, equals $d_W(f, g) = 5$ with respect to the full frequency interval $[0, 30]$. This much larger distance confirms the usefulness of the moving window approach, which uses a much smaller windowed Wasserstein distance of $\min d_{W,\text{window}} = 0.0524$.

To verify that the moving window approach can split off the impact of signal displacement, while remaining sensitive to small changes in the signal shape, we modify the scenario depicted in Fig. 6 in a way that the triplet is mirrored as discussed in Sec. 2.2. Fig. 7 shows the results. The minimum of the windowed Wasserstein distance curve is now $\min d_{W,\text{window}} = 0.1941$. This distance is approximately 0.14 greater than the minimum in the first experiment. The higher transport costs can be traced back to the small changes in the amplitude pattern. The difference value is also consistent with the changes from Figure 4 to 5 (where the sensitivity of the Wasserstein distance under the same amplitude pattern mirroring is tested for the two triplet signals without displacement).

3.2. Application of the moving window approach to DFT-PGA data from Sec. 1.1

To apply the moving window approach to the spectral data from the hydroformylation process, see Sec. 1.1, we must first establish a situation to which the moving window approach is applicable. We observe that the DFT spec-

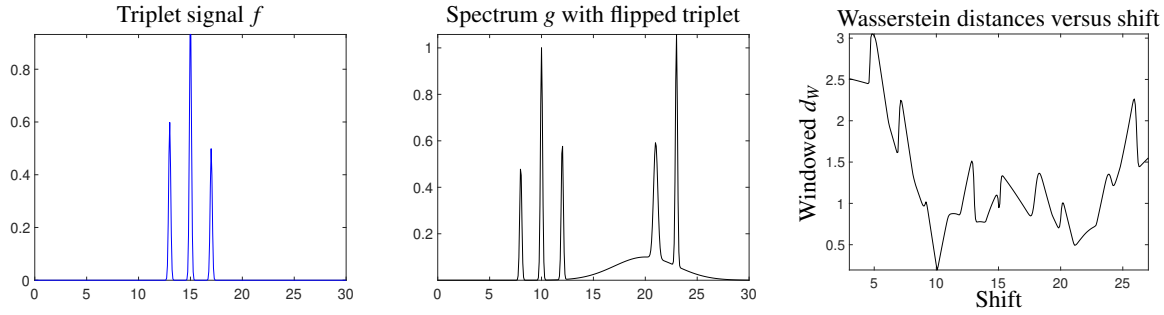

Figure 7: The model problem shown in Fig. 6 is modified in a way that the triplet in the spectrum g is mirrored. The smallest windowed Wasserstein distance is then $\min d_{W,\text{window}} = 0.1941$.

tra cover the wavenumber range $[2000 - 2300]$ cm^{-1} and the PGA spectra are given on the wider range $[1505 - 2150]$ cm^{-1}. Furthermore, while the DFT spectra each show only a single compact group of peaks, the PGA spectra show many more peaks. Therefore, the first step is to define frequency windows that encompass the majority of these peaks. The endpoints of these windows are marked by short red lines in the upper row of Fig. 2. Within these windows, we determine the maximal distances between the peak centers of the dominant peaks within the peak clusters. The peak center frequencies and the associated distances $\Delta_1, \Delta_2, \Delta_3$ are as follows:

DFT1:	2114 cm^{-1}	2144 cm^{-1}	$\Delta_1 = 30$ cm^{-1}
DFT2:	2114 cm^{-1}	2166 cm^{-1}	$\Delta_2 = 52$ cm^{-1}
DFT3:	2121 cm^{-1}	2166 cm^{-1}	$\Delta_3 = 34$ cm^{-1}

Guided by the concept that the dominant peaks in the DFT spectra are somehow correlated with the dominant peaks in the PGA spectra (see the lower row of Fig. 2, we determine the distances $\tilde{\Delta}_1, \tilde{\Delta}_2, \tilde{\Delta}_3$ between the neighboring dominant peaks in the dominant signal clusters of the PGA spectra. The results are as follows:

PGA1:	1943 cm^{-1}	2042 cm^{-1}	$\tilde{\Delta}_1 = 99$ cm^{-1}
PGA2:	1784 cm^{-1}	1888 cm^{-1}	$\tilde{\Delta}_2 = 104$ cm^{-1}
PGA3:	1727 cm^{-1}	1784 cm^{-1}	$\tilde{\Delta}_3 = 57$ cm^{-1}

When we compare the peak center distances Δ_i in the DFT spectra with the peak center distances $\tilde{\Delta}_i$ in the PGA spectra, we see that the quantum mechanical DFT calculations tend to underestimate the experimentally observed peak distances. Consequently, the moving window approach in Sec. 3.1 cannot successfully match the three DFT spectra windows to the dominant peak groups of the PGA spectra due to its construction . Instead, the moving window approach must allow the DFT spectra to be spread (scaled with respect to the abscissa). By comparing the distances Δ_i with the distances $\tilde{\Delta}_i$, we conclude that the unknown scaling factors must cover a range of at least 1.7 to 3.5.

3.3. A move-and-scale spectral matching algorithm

Next, we enhance the moving window approach of Sec. 3.1 by additional scaling of the window and determining the best match based on the smallest Wasserstein distance between the scaled, moved window of the spectrum f and the same window of g. This yields the following move-and-scale algorithm:

Move-and-scale algorithm:

1. Move the DFT signal group window $\mathcal{I}$ through the frequency range of the PGA spectra. If the parameter α denotes the shift, then $\alpha + \mathcal{I}$ is the shifted window.
2. Calculate the minimum of d_W in the respective windows for scaling factors σ in the range 1.0 to 4.0.
3. Determine the minimal Wasserstein distance d_W with respect to the moved and scaled window.

Thus, the spectral matching problem for the sequences of spectra f_i and g_i, $i = 1, \ldots, s$, can be solved by executing

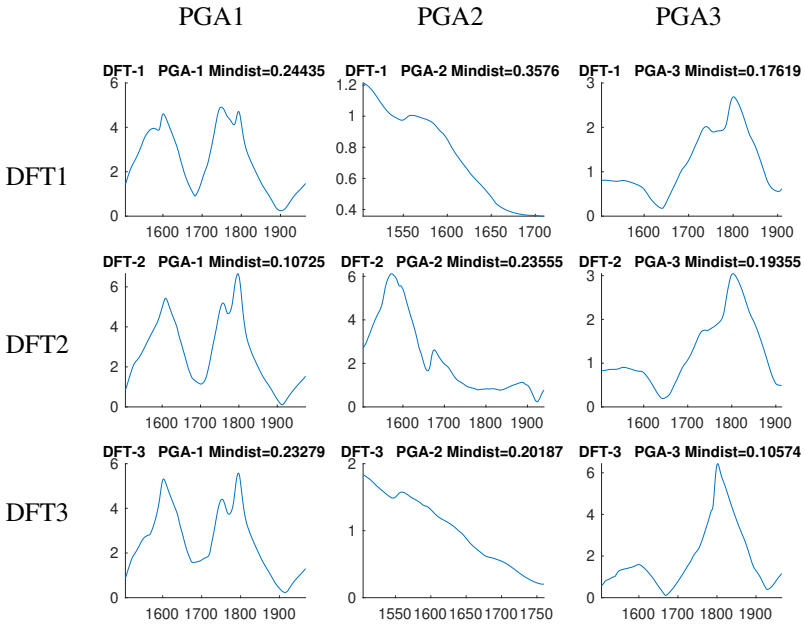

Figure 8: The Wasserstein distance curves are plotted versus the shift parameter α for the nine combinations of three DFT spectra, DFTi, for i=1,2,3, and the three PGA spectra, PGAj, j=1,2,3. For each shift parameter the minimal Wasserstein distance is plotted for optimal scaling parameters in the interval $[1, 4]$.

the two-level minimization:

$$\min_{\alpha} \min_{\sigma} \; d_W(f_i(\alpha + \sigma\mathcal{I}), P_{\alpha+\sigma\mathcal{I}}(g_j)), \quad i, j = 1, \ldots, s, \tag{4}$$

$\mathcal{I}$	window of f_i,
α	shift parameter to move the window to $\alpha + \mathcal{I}$,
σ	scaling parameter $\in [1, 4]$ to move-and-scale the window to $\alpha + \sigma\mathcal{I}$,
$f_1, \ldots, f_s$	first set of spectra, e.g., the DFT spectra,
$g_1, \ldots, g_s$	second set of spectra, e.g., the PGA spectra,
$P_{\alpha+\sigma\mathcal{I}}(g_j)$	Restriction operator to evaluate g_j on the moved and scaled window.

3.4. Application of the move-and-scale algorithm to the DFT-PGA data from Sec. 1.1

The move-and-scale algorithm is applied to the three DFT spectra, DFTi, i=1,2,3, and the three PGA spectra, PGAj for j=1,2,3. These six spectra are plotted in Fig. 2. The minimal Wasserstein distances for the nine possible combinations (i, j) can be represented in terms of a 3×3 matrix D_W, which is given by

$$D_W = \begin{pmatrix} 0.2443 & 0.3576 & 0.1762 \\ 0.1073 & 0.2355 & 0.1935 \\ 0.2328 & 0.2019 & 0.1057 \end{pmatrix}. \tag{5}$$

The associated Wasserstein distance curves versus the shift parameter α are plotted in Fig. 8. Only the inner optimization, see Eq. (4), with respect to the scaling parameter σ is executed for these curves. The minimum of the curve corresponds to the outer minimization in Eq. (4) with respect to the shift parameter σ.

The move-and-scale algorithm provides the following solution to the spectral matching problem.

1. The minimal entry of the distance matrix is $(D_W)_{33} = 0.1057$, indicating that the spectrum DFT3 belongs to the PGA3 spectrum. For a chemical interpretation of this result see [20].
2. As the third DFT spectrum DFT3 has been paired with a certain PGA spectrum, consider only DFTi, for $i = 1, 2$, and search for the smallest matrix element in the first and second rows of D_W. This is the element $(D_W)_{21} = 0.1073$. This means that DFT2 matches to PGA1.
3. Only one combination remains, namely that DFT1 matches PGA2. However, the minimal distance 0.3576 is relatively large. This matching is artificial and is not within the scope of interpretation in [20].

The best spectral matches for the optimal parameters α and σ are shown in Fig. 9. The figure shows all combinations (DFTi, PGAj) for $i, j = 1, 2, 3$. The PGA spectra are plotted in blue, but in red in the active window $\alpha + \sigma\mathcal{I}$ for

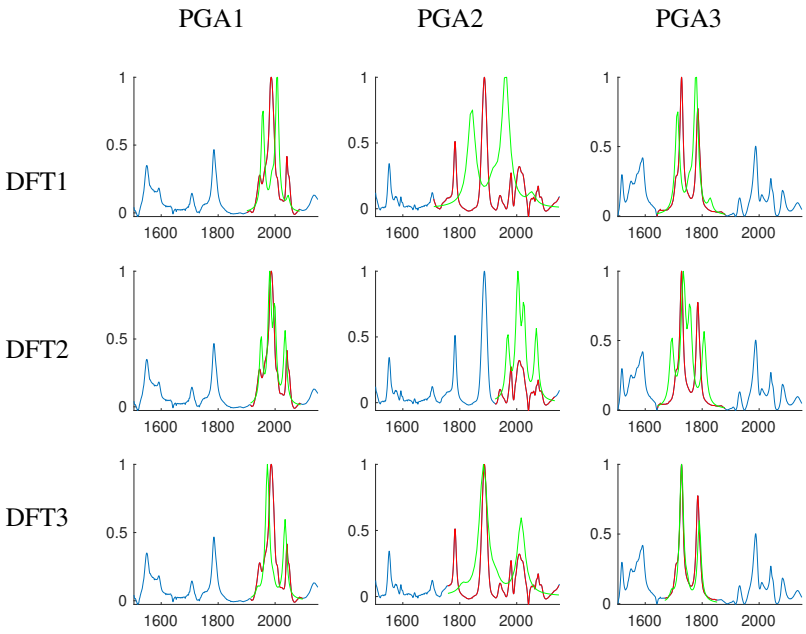

Figure 9: Plot of the best fits with respect to the optimal shift parameter α and the optimal scaling parameter σ. All combinations (DFTi, PGAj) for $i, j = 1, 2, 3$ are shown. The PGA spectra are plotted in blue and red in the active window $\alpha + \sigma\mathcal{I}$, for which the best match is attained. The moved and scaled DFT function for the optimal parameters is plotted in green in the same window where the PGA spectrum is shown in red. Once again, the important, chemically relevant spectral match (DFT3, PGA3) exhibits high similarity between the green and red profiles.

which the best match is attained. The moved and scaled DFT function for the optimal parameters is plotted in green in the same window in which the PGA spectrum is shown in red. The important, chemically relevant spectral match (DFT3, PGA3) shows a high similarity between the green and red profiles, cf. [20].

The scaling factors σ associated with the best match between the ith DFT spectrum and the jth PGA spectrum are stored in the matrix

$$\Sigma = \begin{pmatrix} 1.69 & 4.00 & 2.19 \\ 1.60 & 1.91 & 2.15 \\ 1.83 & 3.90 & 1.84 \end{pmatrix}. \tag{6}$$

For example, the best spectral match (DFT3, PGA3) corresponds to $\sigma = 1.84$, as illustrated in Fig. 9.

The following modifications to the objective function, based on the Wasserstein metric, have been implemented to stabilize the numerical calculations:

1. Window-width scaling: The Wasserstein distances are multiplied by the factor 1000/(width_of_scaled_window). The window width is expressed in wavenumbers. This factor compensates for the impact of the window width σ for which the optimization (4) is performed. (The simple example of the standard basis vectors $e_1, e_n \in \mathbb{R}^n$ shows that the Wasserstein distance $d_W(e_1, e_n) = n - 1$ scales linearly with the window width n.)

2. Norm-g scaling: The Wasserstein distances are multiplied by the factor $1/\|P_{\alpha+\sigma\mathcal{I}}(g)\|_1$, which is the 1-norm of the restriction of g to the spectral window $\alpha + \sigma\mathcal{I}$. This scaling favors regions of g with a higher signal intensity for successful spectral matching. It also prevents regions with a low signal intensity from being competitive for successful spectral matching, since the Wasserstein distance measure uses the normalized function $g/\|g\|_1$.

4. A second experimental data set from rhodium catalyzed hydroformylation

A second experimental data set from a study on P-ligand free rhodium catalyzed hydroformylation is considered [18]. The precursor complex [Rh(acac)$(CO)_2$] ($1 \cdot 10^{-3}$ mol L^{-1}) dissolved in dodecane was treated with synthesis gas ($CO/H_2 = 1 : 1$, P = 20 bar) at $\vartheta = 100\,^{\circ}$C which induced the consecutive formation of $Rh_4(CO)_{12}$ and $Rh_6(CO)_{16}$. Three DFT spectra were calculated (Gaussian, PBE, DGDZVP) for the three species Rh(acac)$(CO)_2$ (short DFT1), $Rh_4(CO)_{12}$ (short DFT2) and $Rh_6(CO)_{16}$ (short DFT3). No vibrational scaling factor was applied. PGA analysis extracted three spectra PGAj, $j = 1, \ldots, 3$, from experimental FTIR data. For this dataset, visual inspection immediately suggests the spectral matching of DFTi with PGAi for $i = 1, \ldots, 3$ even though most of the peaks show a nonnegligble frequency displacement.

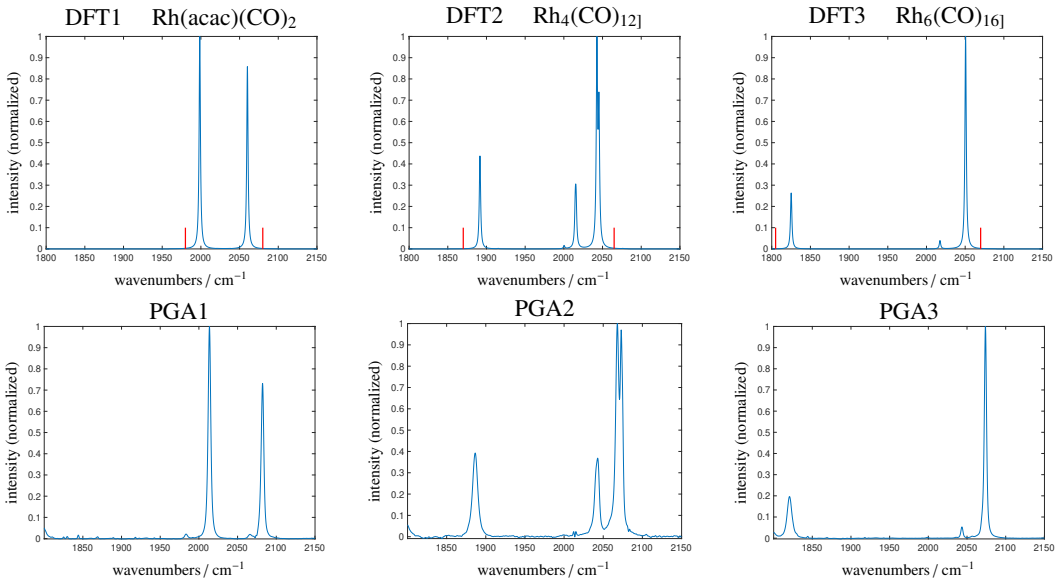

Figure 10: Top row: Three DFT-computed IR spectra (Gaussian, PBE, DGDZVP) for different Rh complexes. The short red lines indicate the selected window for the application of the move-and-scale algorithm. Lower row: Three approximate pure component spectra recovered by PGA from experimental FTIR mixture data. Spectral matches within each of the three columns are immediately suggested by visual inspection.

Fig. 11 shows the Wasserstein distance curves versus the shift parameter α in the same way as in Fig. 8. Additional scaling of the Wasserstein distances, as introduced at the end of Sec. 3.4, is unnecessary for this dataset. This explains why the distance values are larger compared to Sec. 3.4.

The associated 3×3 matrix of minimal Wasserstein distances reads

$$D_W = \begin{pmatrix} 2.44 & 11.21 & 6.43 \\ 15.76 & 5.20 & 15.36 \\ 51.19 & 15.83 & 4.35 \end{pmatrix}.$$

Thus, the solution of the spectral matching problem is as follows: The minimal distance is attained at $(i, j) = (1, 1)$ with the pair (DFT1,PGA1). The next smallest distance is found for the pair (DFT3,PGA3) with $(i, j) = (3, 3)$. Finally, $(i, j) = (2, 2)$ leads to (DFT2,PGA2). Any non-diagonal pair (DFTi,PGAj) with $i \neq j$ has a much larger Wasserstein distance, clearly confirming the diagonal pairs. Finally, the matrix of optimized scaling factors, cf. Eq. (6), is as follows

$$\Sigma = \begin{pmatrix} 1.115 & 1.000 & 1.010 \\ 1.000 & 1.0750 & 1.1950 \\ 1.000 & 1.000 & 1.0350 \end{pmatrix}.$$

The algorithm was permitted to use window scaling factors in the range $\sigma \in [1, 1.3]$. The results show that the optimal scaling factors fall within the range 3.5% to 11.5%. Fig. 12 shows the 3×3 plot of the optimal fits for (DFTi, PGAj). for $i, j = 1, 2, 3$. The PGA spectra are plotted in blue and the DFT spectra within their specific spectral windows are drawn in red. As originally anticipated, the diagonal pairs clearly show the best spectral matches for the given sets of spectra.

5. Conclusion

The Wasserstein metric is an approach to solving the spectral matching problem that interprets spectra as piles of transportable media. The method calculates the transport cost to transform the profiles into each other. In the suggested move-and-scale spectral matching algorithm, we combine the Wasserstein metric with a two-level optimization in order to determine optimal shifts to correct frequency displacements of the spectra, as well as optimal scaling parameters to deal with pairs of spectra with different scaling of the frequency axis.

We have also tested other approaches to solving the spectral matching problem. One is the dynamic time warping method, which measures the similarity of time-dependent functions with differences in peak speed. Time and speed can be substituted with other parameter dependencies. Alternatively, one can calculate cross-correlations for the scaled and moved windows. A related approach is Fourier transform correlation, which determines the cross-correlation in the frequency domain and that is therefore not sensitive to shifts in the spectral range. However, we found that the move-and-scale approach based on the Wasserstein metric is the most successful one.

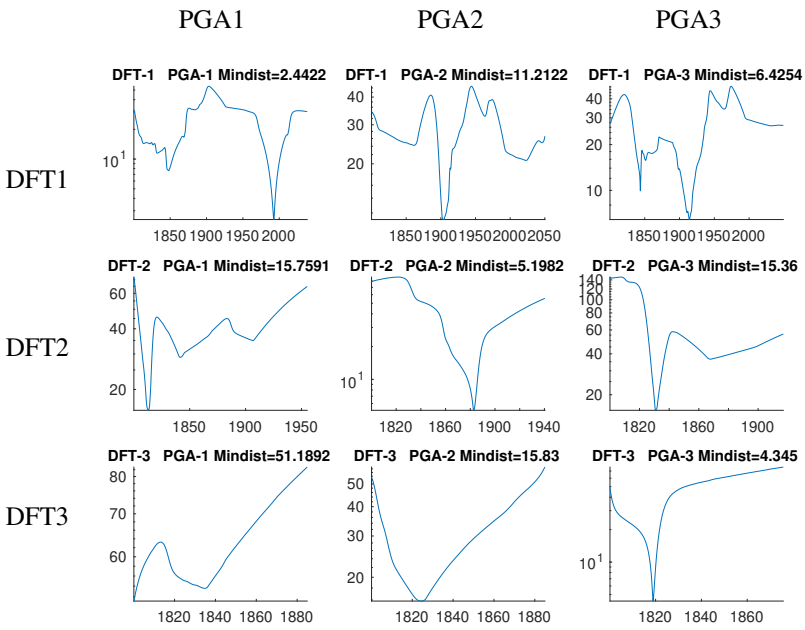

Figure 11: Wasserstein distance curves versus the shift parameter α for the nine combinations of three DFT spectra, DFTi for i=1,2,3 and the three PGA spectra, PGAj, j=1,2,3. The minimal Wasserstein distance is plotted for each shift parameter and all scaling parameters in the interval $[1, 4]$.

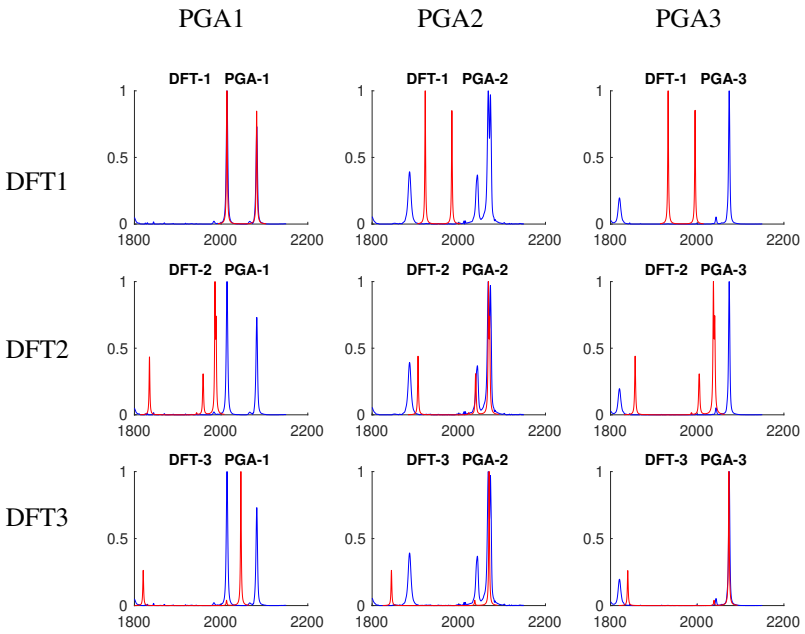

Figure 12: Plot of the best spectral matches for (DFTi, PGAj) for $i, j = 1, 2, 3$. The PGA spectra are shown in blue and the moved and scaled DFT spectra within their respective spectral windows are shown in red. Since the scaling factors are restricted to $\sigma \in [1, 1.3]$, some of the non-diagonal pairs may show better matches for $\sigma > 1.3$. However, the algorithm is not allowed to work with such large deformations since smaller shifts and scaling factors already allow for convincing spectral matching of the diagonal pairs.

Acknowledgement

Funding by the DFG Research Training Group 2943 "SPECTRE" (project number 507189291) is gratefully acknowledged.

A. Details on the DFT computations for the cobalt catalyst

All DFT calculations were performed using Gaussian 16 [6]. The M06-L [37] functional was employed for its proven accuracy in reproducing experimental IR spectra of related cobalt carbonyl complexes [2]. Geometry optimizations were carried out at two levels of theory. The first level of theory (BS1_1) used the LANL2DZ [10] basis set for Co and the 6-31G(d,p) basis set for all other atoms. The second level (BS1_2) employed the LANL2DZ basis set for Co and the 6-31+G(d,p) basis set with diffuse functions for other atoms.

Single-point energy corrections were performed on the optimized structures at the M06-L/BS2 level, where BS2 denotes the def2-TZVP [35, 36] basis set for Co and the 6-311++G(d,p) basis set for other atoms. The Gibbs free energies are expressed as M06-L/BS2//BS1_1 or M06-L/BS2//BS1_2, depending on the optimization level. Thermodynamic corrections were applied under two sets of conditions: 413.15 K and 49.3 atm (50 bar), and 433.15 K and 49.3 atm (50 bar).

The Shermo program [21] was used, implementing Grimme's quasi-rigid rotor harmonic oscillator (quasi-RRHO) method [8] to account for vibrational entropy corrections. The initial geometries of the $19e^-$ cobalt complexes were derived from published geometry structures [11]. The doublet spin state was confirmed by electron paramagnetic resonance (EPR) for $[Co(acac)(dppBz)]^+$, and DFT calculations also indicated that the quartet state is thermodynamically less stable than the doublet state [9]. Therefore, only the low-spin doublet state was considered and discussed in this study.

References

[1] A.D. Allian, Y. Wang, M. Saeys, G.M. Kuramshina, and M. Garland. The combination of deconvolution and density functional theory for the mid-infrared vibrational spectra of stable and unstable rhodium carbonyl clusters. *Vib. Spectrosc.*, 41(1):101–111, 2006.

[2] M.K. Assefa, J.L. Devera, A.D. Brathwaite, J.D. Mosley, and M.A. Duncan. Vibrational scaling factors for transition metal carbonyls. *Chem. Phys. Lett.*, 640:175–179, 2015.

[3] D. Edelmann, T.F. Móri, and G.J. Székely. On relationships between the Pearson and the distance correlation coefficients. *Statistics & Probability Letters*, 169:108960, 2021.

[4] R. Franke, D. Selent, and A. Börner. Applied hydroformylation. *Chem. Rev.*, 112:5675–5732, 2012.

[5] M. Garland. Combining operando spectroscopy with experimental design, signal processing and advanced chemometrics. State of the art and a glimpse of the future. *Catal. Today*, 155:266–270, 2010.

[6] Gaussian, Inc. *Gaussian 16, Revision C.01*, 2016. `https://gaussian.com/citation/`.

[7] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio. *Deep learning*. MIT press, Cambridge, 2016.

[8] S. Grimme. Supramolecular binding thermodynamics by dispersion-corrected density functional theory. *Chem. Eur. J.*, 18(32):9955–9964, 2012.

[9] J. Guo, D. Zhang, and X. Wang. Mechanistic insights into hydroformylation catalyzed by cationic cobalt (ii) complexes: in silico modification of the catalyst system. *ACS Catal.*, 10(22):13551–13559, 2020.

[10] P.J. Hay and W.R. Wadt. Ab initio effective core potentials for molecular calculations. Potentials for K to Au including the outermost core orbitals. *J. Chem. Phys.*, 82(1):299–310, 1985.

[11] D.M. Hood, R. A. Johnson, A.E. Carpenter, J.M. Younker, D.J. Vinyard, and G.G. Stanley. Highly active cationic cobalt (ii) hydroformylation catalysts. *Science*, 367(6477):542–548, 2020.

[12] L. V. Kantorovich. Mathematical methods of organizing and planning production. *Management Science*, 6(4):366–422, 1960.

[13] E. Kohls and M. Stein. Vibrational scaling factors for Rh (I) carbonyl compounds in homogeneous cataylsis. *Contributions: MASA, Section of Natural, Mathematical & Biotechnical Sciences*, 38(1):43–56, 2017.

[14] C. Kubis, M. König, B.N. Leidecker, D. Selent, H. Schröder, M. Sawall, W. Baumann, A. Spannenberg, A. Brächer, K. Neymeyr, R. Franke, and A Börner. Interplay between catalyst complexes and dormant states: In situ spectroscopic investigations on a catalyst system for alkene hydroformylation. *ACS Catal.*, 13(8):5245–5263, 2023.

[15] C. Kubis, R. Ludwig, M. Sawall, K. Neymeyr, A. Börner, K.-D. Wiese, D. Hess, R. Franke, and D. Selent. A comparative in situ HP-FTIR spectroscopic study of bi- and monodentate phosphite-modified hydroformylation. *ChemCatChem*, 2:287–295, 2010.

[16] C. Kubis, M. Sawall, A. Block, K. Neymeyr, R. Ludwig, A. Börner, and D. Selent. An operando FTIR spectroscopic and kinetic study of carbon monoxide pressure influence on rhodium-catalyzed olefin hydroformylation. *Chem.-Eur. J.*, 20(37):11921–11931, 2014.

[17] C. Kubis, D. Selent, M. Sawall, R. Ludwig, K. Neymeyr, W. Baumann, R. Franke, and A. Börner. Exploring between the extremes: Conversion dependent kinetics of phosphite-modified hydroformylation catalysis. *Chem. Eur. J.*, 18(28):8780–8794, 2012.

[18] B.N. Leidecker, D. Peña-Fuentes, C. Wei, M. Sawall, K. Neymeyr, R. Franke, A. Börner, and C. Kubis. In situ FTIR spectroscopic investigations on rhodium carbonyl complexes in the absence of phosphorus ligands under hydroformylation conditions. *New J. Chem.*, 48(43):18365–18375, 2024.

[19] A. Lipp and P. Vermeesch. The Wasserstein distance as a dissimilarity metric for comparing detrital age spectra and other geological distributions. *Geochronology*, 5(1):263–270, 2023.

[20] J. Liu, K. Neymeyr, R. Franke, and B. Zhang. Understand the catalyst formation of cationic cobalt(II) biphosphine hydroformylation catalysis. Technical Report, Leibniz Institute for Catalysis, Rostock, Germany.

[21] T. Lu and Q. Chen. Shermo: A general code for calculating molecular thermochemistry properties. *Comp. Theor. Chem.*, 1200:113249, 2021.

[22] S. Majewski, M.A. Ciach, M. Startek, W. Niemyska, B. Miasojedow, and A. Gambin. The Wasserstein distance as a dissimilarity measure for mass spectra with application to spectral deconvolution. In Laxmi Parida and Esko Ukkonen, editors, *18th International Workshop on Algorithms in Bioinformatics (WABI 2018)*, volume 113 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 25:1–25:21, Dagstuhl, Germany, 2018. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

[23] F. Mémoli. Spectral Gromov-Wasserstein distances for shape matching. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*, pages 256–263, 2009.

[24] K. Neymeyr, C. Kubis, L. Prestin, R. Franke, and M. Sawall. Why is the Peak Group Analysis so effective for IR spectra analysis? Technical report, University of Rostock, 2025.

[25] K. Pearson. Mathematical contributions to the theory of evolution: regression, heredity, and panmixia. *Philos. Trans. R. Soc. Lon.*, 187:253 – 318, 1896.

[26] A.Z. Samuel, R. Mukojima, S. Horii, M. Ando, S. Egashira, T. Nakashima, M. Iwatsuki, and H. Takeyama. On selecting a suitable spectral matching method for automated analytical applications of Raman spectroscopy. *ACS Omega*, 6(3):2060–2065, 2021. PMID: 33521445.

[27] M. Sawall, C. Kubis, E. Barsch, D. Selent, A. Börner, and K. Neymeyr. Peak group analysis for the extraction of pure component spectra. *J. Iran. Chem. Soc.*, 13(2):191–205, 2016.

[28] M. Sawall, C. Kubis, B. N. Leidecker, L. Prestin, T. Andersons, M. Beese, J. Hellwig, R. Franke, A. Börner, and K. Neymeyr. An automated Peak Group Analysis for vibrational spectra analysis. *Chemom. Intell. Lab. Syst.*, 254:105234, 2024.

[29] P. Schober, C. Boer, and L.A. Schwarte. Correlation coefficients: Appropriate use and interpretation. *Anesth. Analg.*, 126(5):1763–1768, 2018.

[30] S. Shanmugam and P. Srinivasa-Perumal. Spectral matching approaches in hyperspectral image processing. *International Journal of Remote Sensing*, 35(24):8217–8251, 2014.

[31] S.S. Vallender. Calculation of the Wasserstein distance between probability distributions on the line. *Theory Probab. Its Appl. (Engl. Transl.)*, 18(4):784–786, 1974.

[32] L.N. Vaserstein. Markov processes over denumerable products of spaces, describing large systems of automata. *Problemy Peredachi Informatsii*, 5(3):64–72, 1969.

[33] C. Villani. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2021.

[34] C. Wei, B.N. Leidecker, D. Peña-Fuentes, H. Schröder, M. Sawall, K. Neymeyr, E.V. Kondratenko, A. Börner, R. Franke, and C. Kubis. Impact of the P-ligand concentration on the formation of hydroformylation catalysts: An in situ FTIR spectroscopic study. *Chem. Ing. Tech.*, 96(12):1657–1667, 2024.

[35] F. Weigend. Accurate Coulomb-fitting basis sets for H to Rn. *Phys. Chem. Chem. Phys.*, 8(9):1057–1065, 2006.

[36] F. Weigend and R. Ahlrichs. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.*, 7(18):3297–3305, 2005.

[37] Y. Zhao and D.G. Truhlar. A new local density functional for main-group thermochemistry, transition metal bonding, thermochemical kinetics, and noncovalent interactions. *J. Chem. Phys.*, 125(19):194101, 2006.