Supporting Information

Structure-preserving joint non-negative tensor factorization to identify reaction pathways using Bayesian networks

Anjana Puliyanda, Kaushik Sivaramakrishnan, Zukui Li, Arno de Klerk, and Vinay Prasad*

Department of Chemical and Materials Engineering, 9211 116 Street NW Edmonton, Alberta T6G 1H9, Canada.

E-mail: vprasad@ualberta.ca

S1 Process Conditions

Table S1: Process conditions for spectral data collection

Spectral sensor	Process conditions	
	Temperature($^\circ C$)	Residence time (min)
FTIR	150	66, 126, 186, 246, 306, 366, 426, 486
	200	66, 126, 186, 246, 306, 486
	250	246
	300	126, 186, 246, 306, 366, 426, 486
	340	6, 66, 126, 246, 486
	360	6, 16.02, 25.98, 36, 66, 246, 583.02
	400	6, 16.02, 25.98, 36, 66, 96, 126
^{1}H-NMR	150	60, 120, 180, 240, 300, 360, 420, 480
	200	60, 120, 180, 240, 300, 360, 420, 480
	250	60, 120, 180, 240, 300, 360, 420, 480
	300	60, 120, 180, 240, 300, 360, 420, 480

S2 Robust formulation of JNTF using subtensors

This section outlines the approach to gradient-based optimization of simultaneously solving for mode matrices. Individual sub-problems in eqn S1-eqn S3 are simple rank R approximations of the mode-n matricized tensors, solved in an ALS-based round robbin scheme.

$$\min_{\mathbf{A}} \sum_{i=1}^{I_1} \sum_{j=1}^{I_2 I_3} \sqrt{(\mathcal{Z}_{(1)} - (\mathbf{A}(\mathbf{C} \odot \mathbf{B})^T)_{ij})^2} \tag{S1}$$

$$\min_{\mathbf{B}} \sum_{i=1}^{I_2} \sum_{j=1}^{I_3 I_1} \sqrt{(\mathcal{Z}_{(2)} - (\mathbf{B}(\mathbf{C} \odot \mathbf{A})^T)_{ij})^2} \tag{S2}$$

$$\min_{\mathbf{C}} \sum_{i=1}^{I_3} \sum_{j=1}^{I_1 I_2} \sqrt{(\mathcal{Z}_{(3)} - (\mathbf{C}(\mathbf{B} \odot \mathbf{A})^T)_{ij})^2} \tag{S3}$$

It is desired to combine these into a single objective function designed to minimize the L_{21} norm of the n^{th} mode matricized tensor. The L_{21} norm of a certain matrix $\mathbf{X}_{m \times n}$ is as given

below:

$$||\mathbf{X}||_{21} = \sum_{i=1}^{n} \sqrt{\sum_{j=1}^{m} x_{ji}^2} \tag{S4}$$

From matrix algebra it is known that $||\mathbf{X}||_F^2 = Tr[\mathbf{X}\mathbf{X}^T]$. Following on these lines for an L_{21} norm has a similar expression in terms of the trace $||\mathbf{X}||_{21} = Tr[\mathbf{X}\ \mathbf{D}\ \mathbf{X}^T]$, with an additional diagonal scaling matrix $\mathbf{D}$ defined as follows:

$$\mathbf{D}(\mathbf{X}) = \frac{I_{n\times n}}{\sqrt{\sum_{i=1}^{m} x_{ij}^2}} \text{for any } \mathbf{X}_{m\times n} \tag{S5}$$

Using eqn S4 and eqn S5 in eqn 15 we have the following formulation of the objective function:

$$\begin{aligned}\min_{\mathbf{A},\mathbf{B},\mathbf{C}\geq 0} F(\mathbf{A},\mathbf{B},\mathbf{C}) = Tr\Big(&\{\mathcal{W}_{(1)} * [\mathcal{Z}_{(1)} - \mathbf{A}(\mathbf{C}\odot\mathbf{B})^T]\}\mathbf{D}_1\{\mathcal{W}_{(1)} * [\mathcal{Z}_{(1)} - \mathbf{A}(\mathbf{C}\odot\mathbf{B})^T]\}^T \\ &+ \{\mathcal{W}_{(2)} * [\mathcal{Z}_{(2)} - \mathbf{B}(\mathbf{C}\odot\mathbf{A})^T]\}\mathbf{D}_2\{\mathcal{W}_{(2)} * [\mathcal{Z}_{(2)} - \mathbf{B}(\mathbf{C}\odot\mathbf{A})^T]\}^T \\ &+ \{\mathcal{W}_{(3)} * [\mathcal{Z}_{(3)} - \mathbf{C}(\mathbf{B}\odot\mathbf{A})^T]\}\mathbf{D}_3\{\mathcal{W}_{(3)} * [\mathcal{Z}_{(3)} - \mathbf{C}(\mathbf{B}\odot\mathbf{A})^T]\}^T\Big)\end{aligned} \tag{S6}$$

where $\mathbf{D}_1 = \mathbf{D}(\mathcal{W}_{(1)} * [\mathcal{Z}_{(1)} - \mathbf{A}(\mathbf{C}\odot\mathbf{B})^T])$, $\mathbf{D}_2 = \mathbf{D}(\mathcal{W}_{(2)} * [\mathcal{Z}_{(2)} - \mathbf{B}(\mathbf{C}\odot\mathbf{A})^T])$, $\mathbf{D}_3 = \mathbf{D}(\mathcal{W}_{(3)} * [\mathcal{Z}_{(3)} - \mathbf{C}(\mathbf{B}\odot\mathbf{A})^T])$ are the diagonal scaling matrices for the n^{th} mode matricized tensor.

The gradients of the objective function in eqn S6 with respect to each of the factor matrices is given below:

$$\begin{aligned}\nabla F_{\mathbf{A}} = &\mathcal{W}_{(1)} * \Big(\mathbf{A}(\mathbf{C}\odot\mathbf{B})^T - \mathcal{Z}_{(1)}\Big)\mathbf{D}_1(\mathbf{C}\odot\mathbf{B}) \\ &+ \frac{\partial(\mathbf{C}\odot\mathbf{A})}{\partial\mathbf{A}}\mathbf{B}^T\mathcal{W}_{(2)} * \Big(\mathbf{B}(\mathbf{C}\odot\mathbf{A})^T - \mathcal{Z}_{(2)}\Big)\mathbf{D}_2(\mathbf{C}\odot\mathbf{A}) \\ &+ \frac{\partial(\mathbf{B}\odot\mathbf{A})}{\partial\mathbf{A}}\mathbf{C}^T\mathcal{W}_{(3)} * \Big(\mathbf{C}(\mathbf{B}\odot\mathbf{A})^T - \mathcal{Z}_{(3)}\Big)\mathbf{D}_3(\mathbf{B}\odot\mathbf{A})\end{aligned} \tag{S7}$$

$$\begin{aligned}\nabla F_{\mathbf{B}} = \mathcal{W}_{(2)} * \Big(\mathbf{B}(\mathbf{C} \odot \mathbf{A})^T - \mathcal{Z}_{(2)}\Big)\mathbf{D}_2(\mathbf{C} \odot \mathbf{A}) \\ + \frac{\partial(\mathbf{C} \odot \mathbf{B})}{\partial \mathbf{B}}\mathbf{A}^T\mathcal{W}_{(1)} * \Big(\mathbf{A}(\mathbf{C} \odot \mathbf{B})^T - \mathcal{Z}_{(1)}\Big)\mathbf{D}_1(\mathbf{C} \odot \mathbf{B}) \\ + \frac{\partial(\mathbf{B} \odot \mathbf{A})}{\partial \mathbf{B}}\mathbf{C}^T\mathcal{W}_{(3)} * \Big(\mathbf{C}(\mathbf{B} \odot \mathbf{A})^T - \mathcal{Z}_{(3)}\Big)\mathbf{D}_3(\mathbf{B} \odot \mathbf{A})\end{aligned} \tag{S8}$$

$$\begin{aligned}\nabla F_{\mathbf{C}} = \mathcal{W}_{(3)} * \Big(\mathbf{C}(\mathbf{B} \odot \mathbf{A})^T - \mathcal{Z}_{(3)}\Big)\mathbf{D}_3(\mathbf{B} \odot \mathbf{A}) \\ + \frac{\partial(\mathbf{C} \odot \mathbf{B})}{\partial \mathbf{C}}\mathbf{A}^T\mathcal{W}_{(1)} * \Big(\mathbf{A}(\mathbf{C} \odot \mathbf{B})^T - \mathcal{Z}_{(1)}\Big)\mathbf{D}_1(\mathbf{C} \odot \mathbf{B}) \\ + \frac{\partial(\mathbf{C} \odot \mathbf{A})}{\partial \mathbf{C}}\mathbf{B}^T\mathcal{W}_{(2)} * \Big(\mathbf{B}(\mathbf{C} \odot \mathbf{A})^T - \mathcal{Z}_{(2)}\Big)\mathbf{D}_2(\mathbf{C} \odot \mathbf{A})\end{aligned} \tag{S9}$$

To tackle the derivatives of the Khatri-rao ($\odot$) aka columnwise Kronecker product ($|\otimes|$) in the above expression for gradients we resort to the use of vectorizing the product expressions using principles of tensor algebra, computing the gradients of the vectors and then re-shaping them to matrices.

For example let us say we have two matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ of dimensions $m \times n$ and $p \times n$ respectively, then the derivative of their column-wise Kronecker product is given by:

$$\frac{\partial(\mathbf{X}_1 \odot \mathbf{X}_2)}{\partial \mathbf{X}_i} = \text{Reshape}\left(\frac{\partial\, vec\{\mathbf{X}_1 \odot \mathbf{X}_2\}}{\partial x_i}\right) = \text{Reshape}\Big(K_i^T K_i\, vec\{\mathbf{X}_i\}\Big) \tag{S10}$$

Expressions for K_i come from the following two equations from tensor algebra:

$$vec\{\mathbf{X}_1 \odot \mathbf{X}_2\} = \quad ([I_N \odot \mathbf{X}_1] \otimes I_P)\, vec\{\mathbf{X}_2\} = \quad K_2 vec\{\mathbf{X}_2\} \tag{S11}$$

$$vec\{\mathbf{X}_1 \odot \mathbf{X}_2\} = \quad [I_{MN} \odot (\mathbf{X}_2\,[I_N \otimes 1_{1\times M}])]\, vec\{\mathbf{X}_1\} = \quad K_1 vec\{\mathbf{X}_1\} \tag{S12}$$

It can be seen that the gradient computation of mode matricized tensors involve the derivatives of the Khatri-Rao products of the matrix modes, the computation of which is memory

intensive for large-scale tensors making it challenge in the implementation of JNTF[1]. Hence a large tensor is typically divided into subtensors, parallelzing the JNTF over the small-sized subtensors using the divide and conquer technique[2].

The concepts discussed in this section are now put together as we extend it to the framework of *joint* weighted robust non-negative tensor factorization with respect to our case of factorizing tensor blocks of FTIR and ^{1}H-NMR data. Since the dimension of the spectral channel modes are much higher than that of the process modes, it is proposed to divide the tensors into subtensors along the spectral channel modes. Hence, the grid tensor factorization (GTF) is also implemented in the high dimensional mode of wavenumbers/chemical shifts. Let N_1, N_2 be the number of FTIR and HNMR subtensors respectively. For FTIR i=1 for HNMR i=2 : $\mathcal{Z}^{[n_i]} \in \Re^{I_1 \times I_2 \times K_{n_i}}$ So from CPD $\mathcal{Z}^{[n_i]} \approx I \times_1 \mathbf{A}^{[n_i]} \times_2 \mathbf{B}^{[n_i]} \times_3 \mathbf{H}_i^{[n_i]}$ where $n_i = 1, 2 \cdots N_i$ and $\mathbf{A}^{[n_i]} \in \Re^{I_1 \times R}, \mathbf{B}^{[n_i]} \in \Re^{I_2 \times R}, \mathbf{H}_i^{[n_i]} \in \Re^{K_{n_i} \times R}$ such that $\sum_{n_i=1}^{N_i} K_{n_i} = I_3$ followed by $\mathbf{H}_i = [\mathbf{H}_i^{[1]T}, \mathbf{H}_i^{[2]T}, \cdots \mathbf{H}_i^{[N_i]T}]^T$

The objective function :

$$\min_{\mathbf{A}^{[n_i]}, \mathbf{B}^{[n_i]}, \mathbf{H}_i^{[n_i]} \geq 0} \sum_{i=1,2} \sum_{n_i=1}^{N_i} ||\mathcal{W}^{[n_i]} * (\mathcal{Z}^{[n_i]} - [[\mathbf{A}^{[n_i]}, \mathbf{B}^{[n_i]}, \mathbf{H}_i^{[n_i]}]])||_{21} \tag{S13}$$

Writing out eqn S13 out explicitly in terms of the matricized n-mode tensor:

$$\begin{aligned} \min_{\mathbf{A}^{[n_i]}, \mathbf{B}^{[n_i]}, \mathbf{H}_i^{[n_i]} \geq 0} F(\mathbf{A}, \mathbf{B}, \mathbf{H}_i) = \sum_{i=1,2} \sum_{n_i=1}^{N_i} & ||\mathcal{W}_{(1)}^{[n_i]} * [\mathcal{Z}_{(1)}^{[n_i]} - \mathbf{A}^{[n_i]} (\mathbf{H}_i^{[n_i]} \odot \mathbf{B}^{[n_i]})^T]||_{21} + \\ & ||\mathcal{W}_{(2)}^{[n_i]} * [\mathcal{Z}_{(2)}^{[n_i]} - \mathbf{B}^{[n_i]} (\mathbf{H}_i^{[n_i]} \odot \mathbf{A}^{[n_i]})^T]||_{21} \\ & + ||\mathcal{W}_{(3)}^{[n_i]} * [\mathcal{Z}_{(3)}^{[n_i]} - \mathbf{H}_i^{[n_i]} (\mathbf{B}^{[n_i]} \odot \mathbf{A}^{[n_i]})^T]||_{21} \end{aligned} \tag{S14}$$

Using eqn S4 and eqn S5 in eqn S14 we have the following formulation of the objective function:

$$\min_{\mathbf{A}^{[n_i]},\mathbf{B}^{[n_i]},\mathbf{H}_i^{[n_i]}\geq 0} F(\mathbf{A},\mathbf{B},\mathbf{H}_i)=\sum_{i=1,2}\sum_{n_i=1}^{N_i} Tr\Big(\{\mathcal{W}_{(1)}^{[n_i]}*[\mathcal{Z}_{(1)}^{[n_i]}-\mathbf{A}^{[n_i]}(\mathbf{H}_i^{[n_i]}\odot\mathbf{B}^{[n_i]})^T]\}\mathbf{D}_1^{[n_i]}\{\mathcal{W}_{(1)}^{[n_i]}*[\mathcal{Z}_{(1)}^{[n_i]}-\mathbf{A}^{[n_i]}(\mathbf{H}_i^{[n_i]}\odot\mathbf{B}^{[n_i]})^T]\}^T$$
$$+\{\mathcal{W}_{(2)}^{[n_i]}*[\mathcal{Z}_{(2)}^{[n_i]}-\mathbf{B}^{[n_i]}(\mathbf{H}_i^{[n_i]}\odot\mathbf{A}^{[n_i]})^T]\}\mathbf{D}_2^{[n_i]}\{\mathcal{W}_{(2)}^{[n_i]}*[\mathcal{Z}_{(2)}^{[n_i]}-\mathcal{B}^{[n_i]}(\mathbf{H}_i^{[n_i]}\odot\mathbf{A}^{[n_i]})^T]\}^T$$
$$+\{\mathcal{W}_{(3)}^{[n_i]}*[\mathcal{Z}_{(3)}^{[n_i]}-\mathbf{H}_i^{[n_i]}(\mathbf{B}^{[n_i]}\odot\mathbf{A}^{[n_i]})^T]\}\mathbf{D}_3^{[n_i]}\{\mathcal{W}_{(3)}^{[n_i]}*[\mathcal{Z}_{(3)}^{[n_i]}-\mathbf{H}_i^{[n_i]}(\mathbf{B}^{[n_i]}\odot\mathbf{A}^{[n_i]})^T]\}^T\Big) \tag{S15}$$

The gradients of the objective function wrt to the factor matrices are given below:

$$\begin{aligned}\nabla F_{\mathbf{A}} = &\sum_{i=1,2}\sum_{n_i=1}^{N_i}\mathcal{W}_{(1)}^{[n_i]}*\Big(\mathbf{A}^{[n_i]}(\mathbf{H}_i^{[n_i]}\odot\mathbf{B}^{[n_i]})^T-\mathcal{Z}_{(1)}^{[n_i]}\Big)\mathbf{D}_1^{[n_i]}(\mathbf{H}_i^{[n_i]}\odot\mathbf{B}^{[n_i]})\\ &+\frac{\partial(\mathbf{H}_i^{[n_i]}\odot\mathbf{A}^{[n_i]})}{\partial\mathbf{A}^{[n_i]}}\mathbf{B}^{[n_i]T}\mathcal{W}_{(2)}^{[n_i]}*\Big(\mathbf{B}^{[n_i]}(\mathbf{H}_i^{[n_i]}\odot\mathbf{A}^{[n_i]})^T-\mathcal{Z}_{(2)}^{[n_i]}\Big)\mathbf{D}_2^{[n_i]}(\mathbf{H}_i^{[n_i]}\odot\mathbf{A}^{[n_i]})\\ &+\frac{\partial(\mathbf{B}^{[n_i]}\odot\mathbf{A}^{[n_i]})}{\partial\mathbf{A}^{[n_i]}}\mathbf{H}_i^{[n_i]T}\mathcal{W}_{(3)}^{[n_i]}*\Big(\mathbf{H}_i^{[n_i]}(\mathbf{B}^{[n_i]}\odot\mathbf{A}^{[n_i]}T)-\mathcal{Z}_{(3)}^{[n_i]}\Big)\mathbf{D}_3^{[n_i]}(\mathbf{B}^{[n_i]}\odot\mathbf{A}^{[n_i]})\end{aligned} \tag{S16}$$

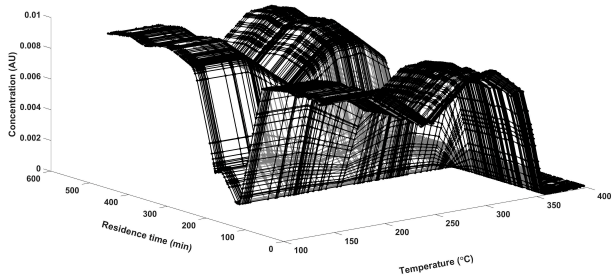
$$\begin{aligned}\nabla F_{\mathbf{B}} = &\sum_{i=1,2}\sum_{n_i=1}^{N_i}\mathcal{W}_{(2)}^{[n_i]}*\Big(\mathbf{B}^{[n_i]}(\mathbf{H}_i^{[n_i]}\odot\mathbf{A}^{[n_i]})^T-\mathcal{Z}_{(2)}^{[n_i]}\Big)\mathbf{D}_2^{[n_i]}(\mathbf{H}_i^{[n_i]}\odot\mathbf{A}^{[n_i]})\\ &+\frac{\partial(\mathbf{H}_i^{[n_i]}\odot\mathbf{B}^{[n_i]})}{\partial\mathbf{B}^{[n_i]}}\mathbf{A}^{[n_i]T}\mathcal{W}_{(1)}^{[n_i]}*\Big(\mathbf{A}^{[n_i]}(\mathbf{H}_i^{[n_i]}\odot\mathbf{B}^{[n_i]})^T-\mathcal{Z}_{(1)}^{[n_i]}\Big)\mathbf{D}_1^{[n_i]}(\mathbf{H}_i^{[n_i]}\odot\mathbf{B}^{[n_i]})\\ &+\frac{\partial(\mathbf{B}^{[n_i]}\odot\mathbf{A}^{[n_i]})}{\partial\mathbf{B}^{[n_i]}}\mathbf{H}_i^{[n_i]T}\mathcal{W}_{(3)}^{[n_i]}*\Big(\mathbf{H}_i^{[n_i]}(\mathbf{B}^{[n_i]}\odot\mathbf{A}^{[n_i]})^T-\mathcal{Z}_{(3)}^{[n_i]}\Big)\mathbf{D}_3^{[n_i]}(\mathbf{B}^{[n_i]}\odot\mathbf{A}^{[n_i]})\end{aligned} \tag{S17}$$

$$\nabla F_{\mathbf{H}_i} = \mathcal{W}_{(3)}^{[n_i]} * \Big(\mathbf{H}_i^{[n_i]}(\mathbf{B}^{[n_i]} \odot \mathbf{A}^{[n_i]})^T - \mathcal{Z}_{(3)}^{[n_i]}\Big)\mathbf{D}_3^{[n_i]}(\mathbf{B}^{[n_i]} \odot \mathbf{A}^{[n_i]})$$
$$+ \frac{\partial(\mathbf{H}_i^{[n_i]} \odot \mathbf{B}^{[n_i]})}{\partial \mathbf{H}_i^{[n_i]}}\mathbf{A}^{[n_i]T}\mathcal{W}_{(1)}^{[n_i]} * \Big(\mathbf{A}^{[n_i]}(\mathbf{H}_i^{[n_i]} \odot \mathbf{B}^{[n_i]})^T - \mathcal{Z}_{(1)}^{[n_i]}\Big)\mathbf{D}_1^{[n_i]}(\mathbf{H}_i^{[n_i]} \odot \mathbf{B}^{[n_i]})$$
$$+ \frac{\partial(\mathbf{H}_i^{[n_i]} \odot \mathbf{A}^{[n_i]})}{\partial \mathbf{H}_i^{[n_i]}}\mathbf{B}^{[n_i]T}\mathcal{W}_{(2)}^{[n_i]} * \Big(\mathbf{B}^{[n_i]}(\mathbf{H}_i^{[n_i]} \odot \mathbf{A}^{[n_i]})^T - \mathcal{Z}_{(2)}^{[n_i]}\Big)\mathbf{D}_2^{[n_i]}(\mathbf{H}_i^{[n_i]} \odot \mathbf{A}^{[n_i]}) \tag{S18}$$

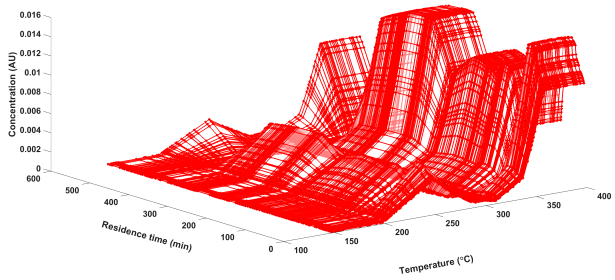
The above problem has been formulated as a gradient-based optimization and is solved using the LBFGSB solver of the Poblano optimization toolbox developed by Sandia Laboratories on Matlab[3].

S3 NTF of synthetically generated FTIR spectra

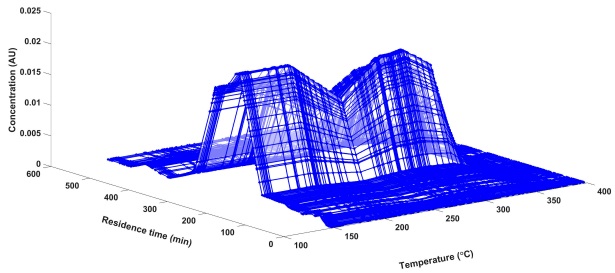
Section 4.1 describes the results of performing robust non-negative tensor factorization on the FTIR spectra for the 41 temperature and residence time conditions given in Table S1, in addition to the baseline spectrum. The absence of spectral data across certain reaction times at each temperature are accorded as missing values, and are imputed in the process of factorization. In this section, we investigate the results of NTF in the event of being able to collect data extensively across all times at each temperature, at several intermediate temperature conditions. The spectral data at the intermediate temperature-time conditions have been generated synthetically by random interpolation of the existing spectral data in Table S1, followed by baseline correction before being fed into the NTF objective.



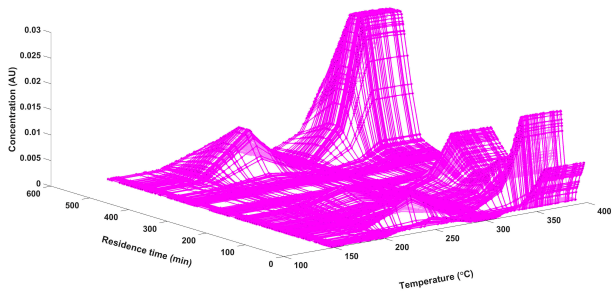
(a) Pseudo-component 1



(b) Pseudo-component 2



(c) Pseudo-component 3



(d) Pseudo-component 4

Figure S1: Concentrations of the pseudo-components across the reaction space of the synthetic FTIR dataset

Figure S1 provides the concentration profiles across the reaction space of temperature and residence times, for the 4 pseudo-components, while Figure S2 gives the extracted spectral profiles obtained by projection onto the FTIR spectral channels for the 4 pseudo-components. It can be seen that the concentration surface of PC_3 is more pronounced at intermediate residence times, whereas PC_1 is seen to have a sharp decreasing trend, while PC_2 and PC_4 have smaller increases in concentration, that later rise at higher temperatures. It can be inferred that PC_1 represents a class of starting reactants that finally give rise to a class of final products, represented by PC_3, while PC_2 and PC_4 could be treated as a class of reaction intermediates obtained by various mechanisms underlying the conversion of $PC_1 \rightarrow PC_3$.

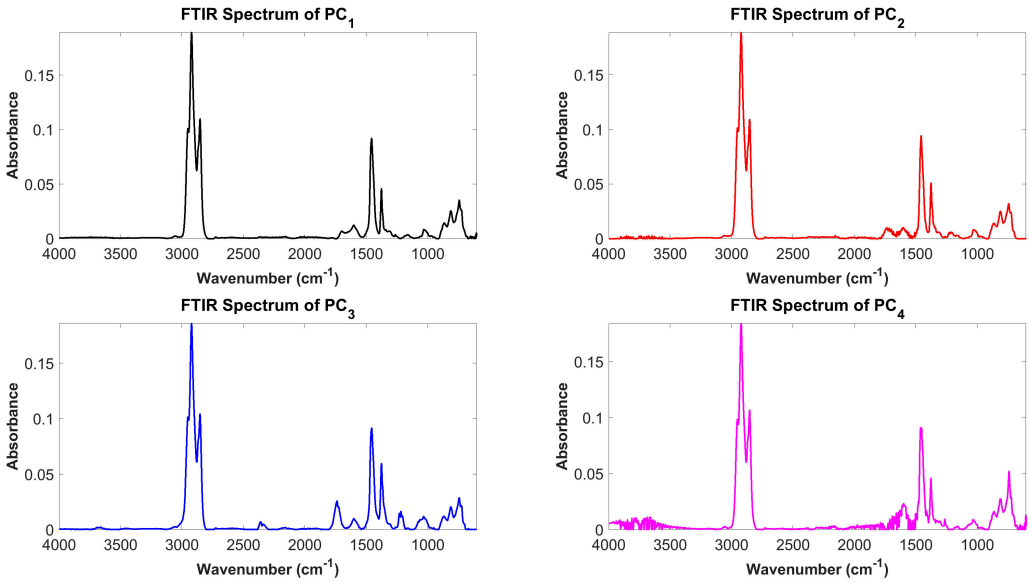


Figure S2: Spectra of pseudo-components from the synthetic FTIR tensor decomposition

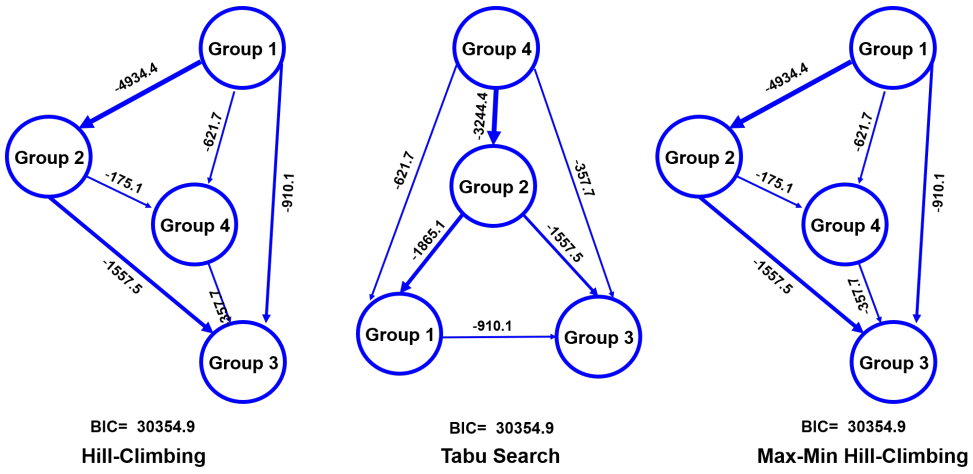


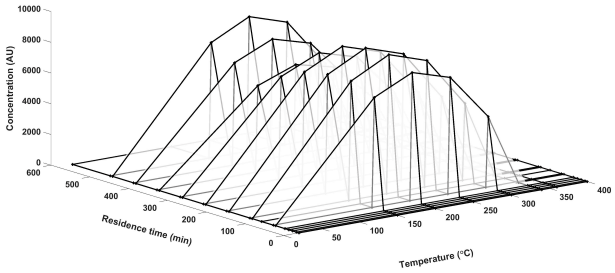
Figure S3: Bayesian networks from the synthetic FTIR pseudo-component spectra

The reaction mechanisms inferred by using Bayesian structure learning among the pseudo-component spectra of Figure S2, as given by Figure S3 is found to corroborate with the qualitative inferences drawn from the concentration profiles. The details of the reaction mechanisms underlying the hypotheses generated from the Bayesian networks can be deciphered by chemically interpreting the functional groups in the spectra of the associated pseudo-components.

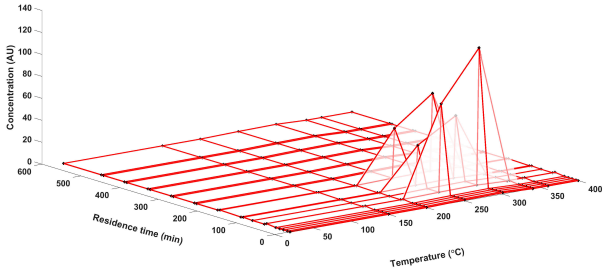
S4 Individual analysis of ^{1}H-NMR data

The extracted concentration profiles of the pseudo-components from tensor decomposition in the reaction space of the temperature and residence time modes are given in Figure S4. The extracted ^{1}H-NMR profiles for the 4 pseudo-components are given in Figure S5. The Bayesian networks depicting causal relationships among the 4 groups are given in Figure S6. Hill climbing and the maximum minimum hill climbing score search methods result in similar network structures that indicate PC_1 as the reactant species. The concentration of PC_1 is seen to be much higher than the other pseudo-components at all temperatures and residence times, corroborate with PC_1 being the starting reactant species. The concentrations of PC_2 are prominent at higher temperatures and lower residence times, while PC_3 and PC_4 appear

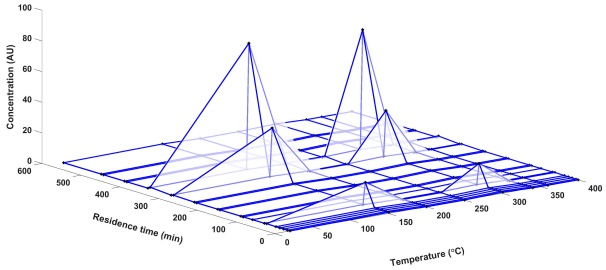
at lower temperatures reacted over longer durations, towards the later part of the reaction residence time, indicating that they represent a class of the product species, as indicated by the Bayesian networks as well.



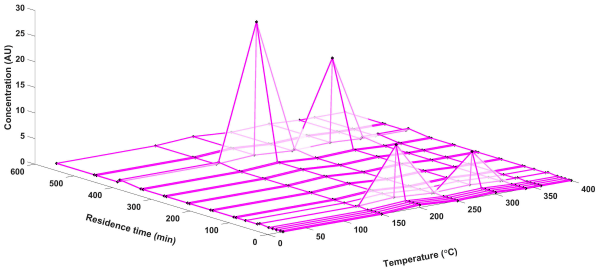
(a) Pseudo-component 1



(b) Pseudo-component 2



(c) Pseudo-component 3



(d) Pseudo-component 4

Figure S4: Concentrations of the pseudo-components across the reaction space of the ^{1}H-NMR spectra

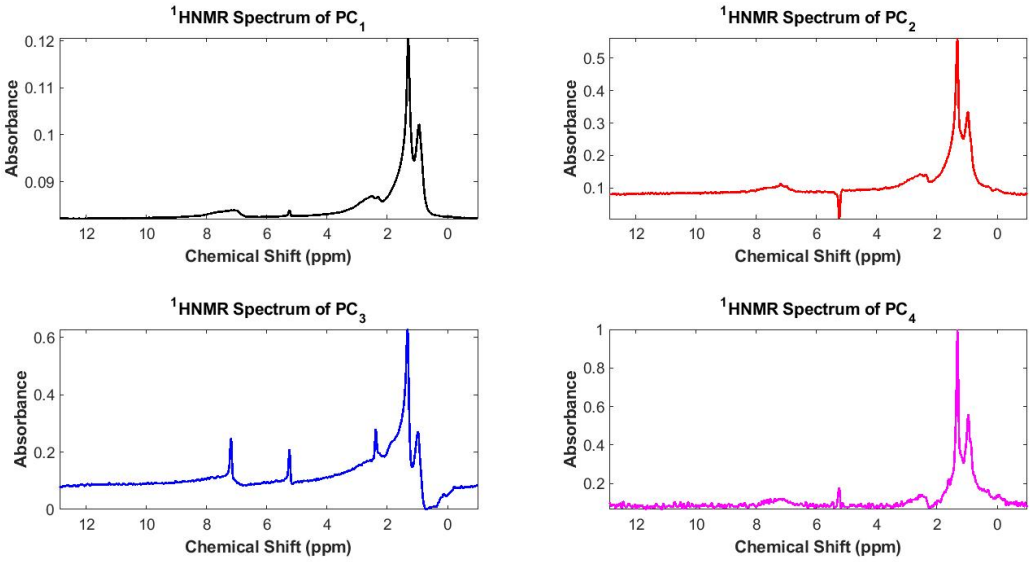


Figure S5: Spectra of pseudo-components from ^{1}HNMR tensor decomposition

^{1}H-NMR spectra alone does not provide as much information as the FTIR spectra, especially in the aromatic region since it just shows a single overlapped lump from 7 – 9 ppm (Figure S5), except for PC_3 that has a distinct aromatic hydrogen peak at ∼7.2 ppm. Peaks for aliphatic methylene and methyl protons are distinct and common to all pseudo-components with CH_2 showing higher intensity. All spectra also show the peak for benzylic proton at ∼2.5 ppm confirming the presence of aromatics, but this does not indicate the number of substitutions. Another distinct characteristic of ^{1}H-NMR profiles is the peak at ∼5.2 ppm that depicts hydrogen from methylene chloride which points to the solvent that remains in the converted samples. This is present in all pseudo-components, although an inverted peak in PC_2, and also falls in the olefinic range. Overall, not much conversion chemistry can be proposed from ^{1}H-NMR profiles alone so it is worthwhile to look at the joint decomposition in section 4.2.

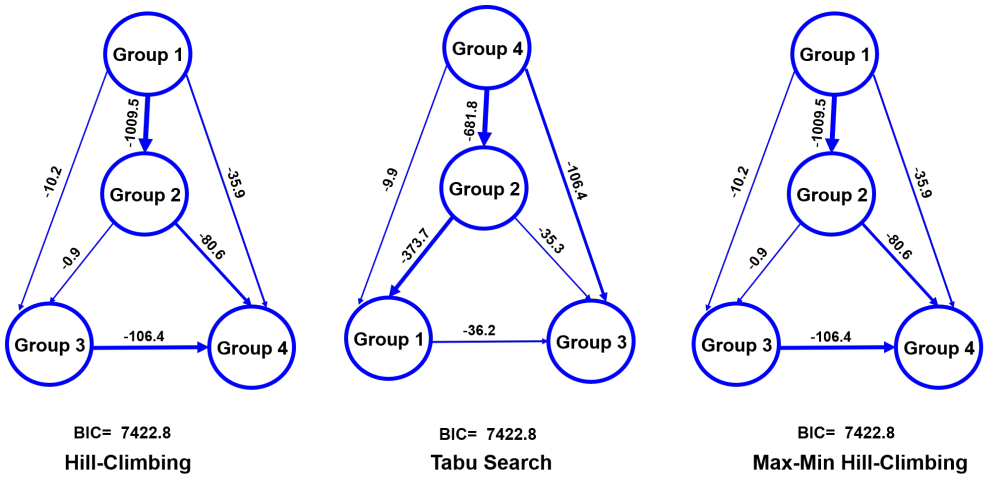
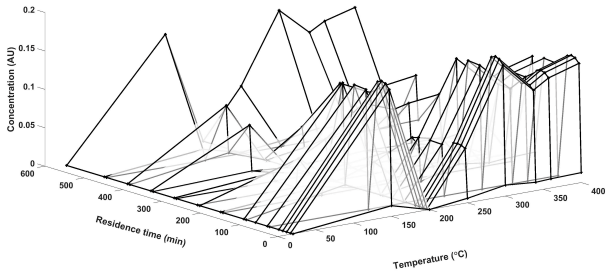


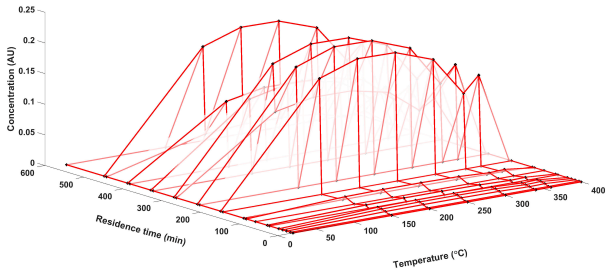
Figure S6: Bayesian networks from the unique ^{1}H-NMR pseudo-component spectra

S5 Gaussian tensor factorization

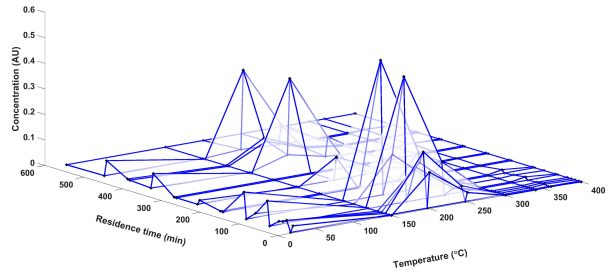
S5.1 Individual tensor factorization of FTIR spectra



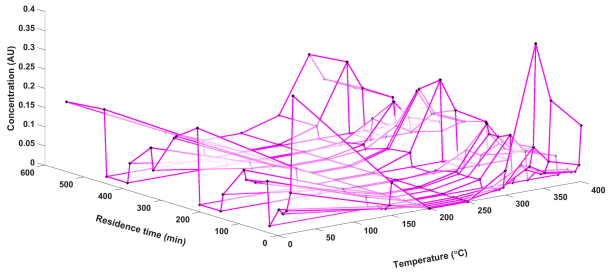
(a) Pseudo-component 1



(b) Pseudo-component 2



(c) Pseudo-component 3



(d) Pseudo-component 4

Figure S7: Concentrations of the pseudo-components across the reaction space of the FTIR spectra

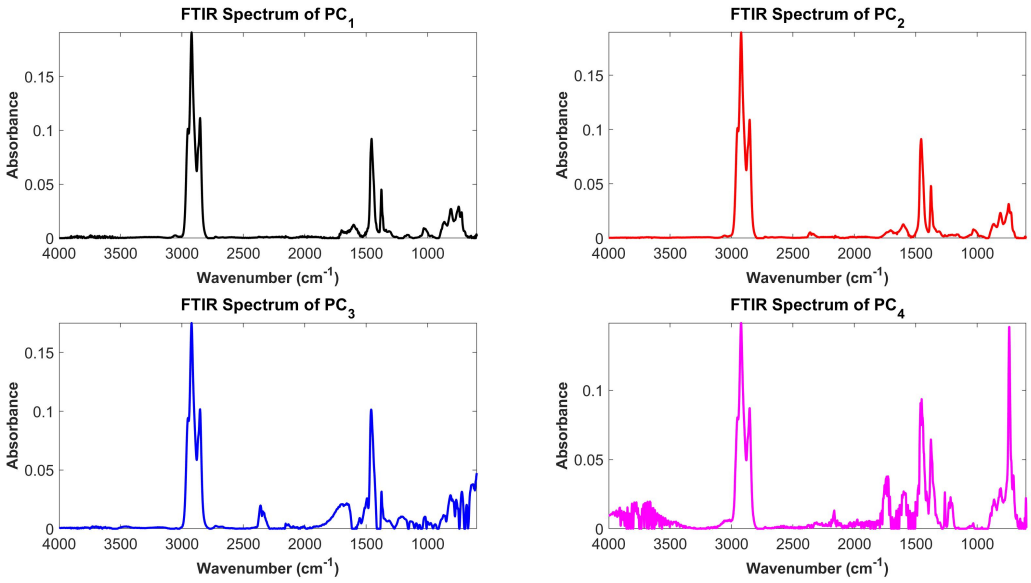


Figure S8: Spectra of pseudo-components from FTIR tensor decomposition

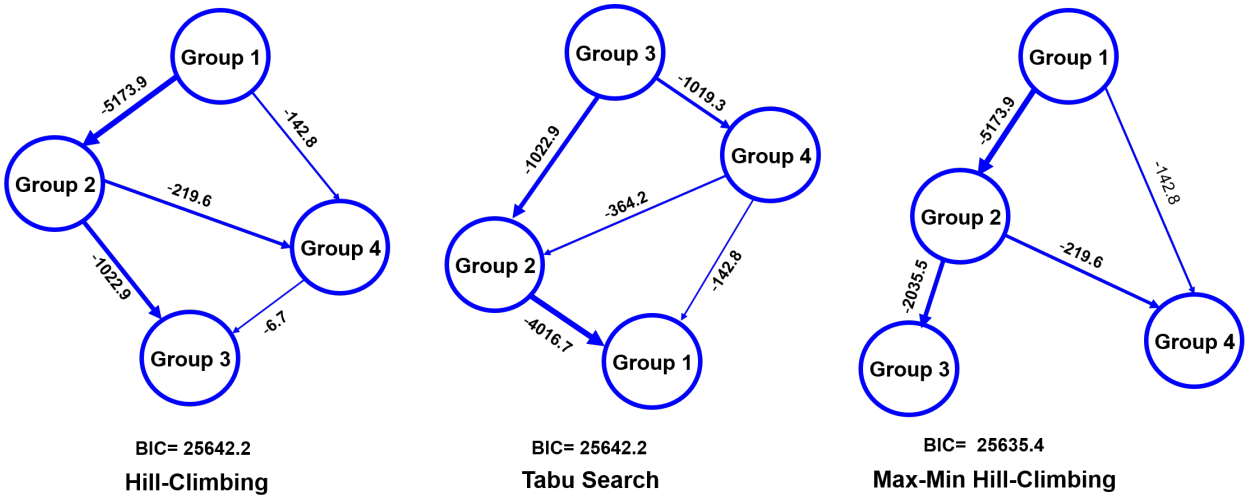


Figure S9: Bayesian networks from the unique FTIR pseudo-component spectra

The major peaks in the FTIR spectra of the pseudo-components have been tabulated in Table S2

Table S2: Absorption regions for all groups in robust FTIR formulation.

Wavenumber (cm^{-1})	Functional group	Vibration type	PCs/groups present
1597	C=C aromatic	Stretch	PC1, PC2, PC4
1701	C=O of carboxylic acid	Stretch	All 4 PCs
1172 – 1203	C-O of acyl group	Stretch	PC1, PC2, PC3
1018	C-O of aliphatics	Stretch	PC1, PC2, PC3
862	C-H in p-substituted aromatics	Bend	Least intensity but present in all 4 PCs
810	C-H in m-substituted aromatics	Bend	Clearly present in all 4 PCs
740	C-H in o-substituted aromatics	Bend	All 4 PCs but highest for PC4
723	C-H in mono-substituted aromatics	Bend	All 4 PCs – as a shoulder with 740 cm-1
1730	C=O in esters/anhydrides	Stretch	PC4
2360	S-H in thiols	Stretch	PC2, PC3
2150	Alkyne triple bond	Stretch	PC3, PC4

For PC_1, absorption at 1700 cm^{-1} indicated the presence of carboxylic acid and its co-existence with C-O acylic group at 1175 cm^{-1} confirmed this observation. Presence of aliphatic alcohol was also marked by absorption at 1018 cm^{-1}. All sp^2 C-H bends for aromatics in the 700 – 900 cm^{-1} region were of almost equal intensity (0.035 units) except the p-compounds as already mentioned. The representative compounds for each group are shown in Figure S10, Figure S11, Figure S12 and Figure S13 that depict the proposed reaction pathways based on the results of Bayesian networks from Gaussian tensor decomposition of FTIR data. Compound (1) is a representative molecule for G1 since it has a carboxylic acid, aliphatic alcohol in the naphthene ring, a side chain and an aromatic ring that is substituted in o-, m- and p- positions. The chemical composition of G2 species is not much different than G1 but it was speculated to be a condensed version of the tri-cyclic compound (1), where the

middle or the third ring becomes aromatic in addition to the already existing aromatic first ring. When the middle ring turns aromatic, it leads to a phenolic entity (compound (2)) while if the end ring turns aromatic, it remains an aliphatic alcoholic species. Probability of the end ring turning into aromatic is lower than that of the middle ring due to the requirement of the loss of a lower number of hydrogens but since G2 has a higher intensity for alkoxy C-O absorption (Figure S8 and Table 1), compound (3) could represent G2 species better. Nevertheless, both compound (2) and compound (3) in Figure S10 are good representatives of G2/PC2.

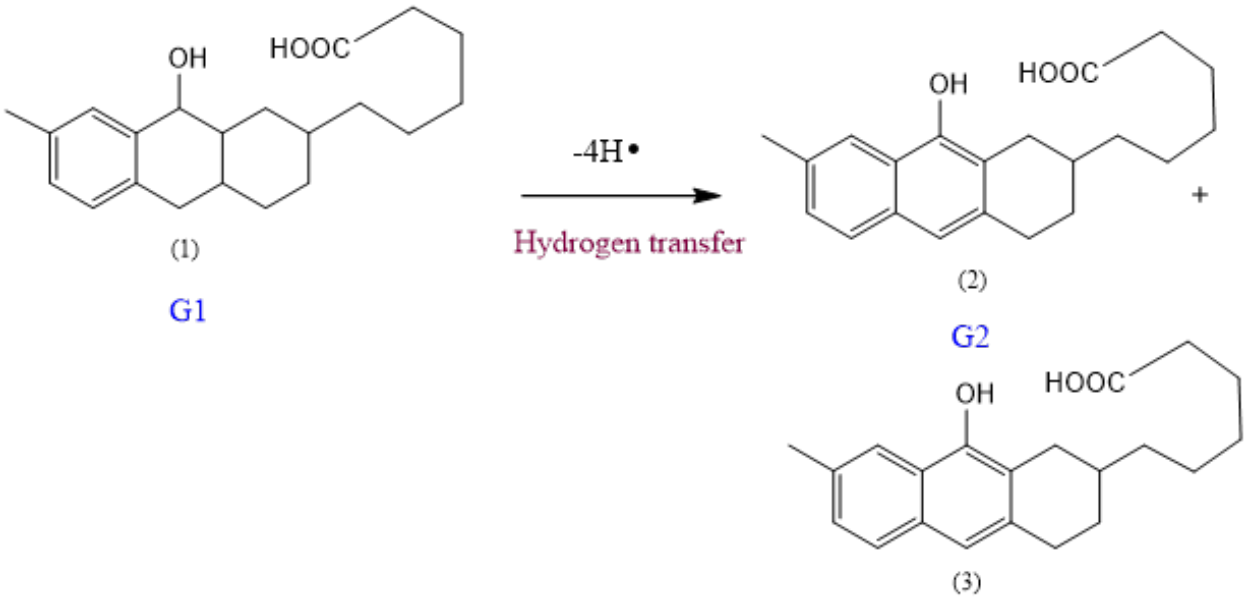


Figure S10: Proposed reaction pathway of group 1 to group 2 conversion.

Moving on to PC_3, it was interesting to note that although it had aromatic C-H bends in the 700 – 900 cm^{-1} region, it had more olefinic characteristics due to the C=C stretch at 1650 cm^{-1} (Table 1).In order to realize PC_3, we need to look at Figure S11 that gives the conversion pathway of G2 to G3 species as proposed from the developed Bayesian networks.

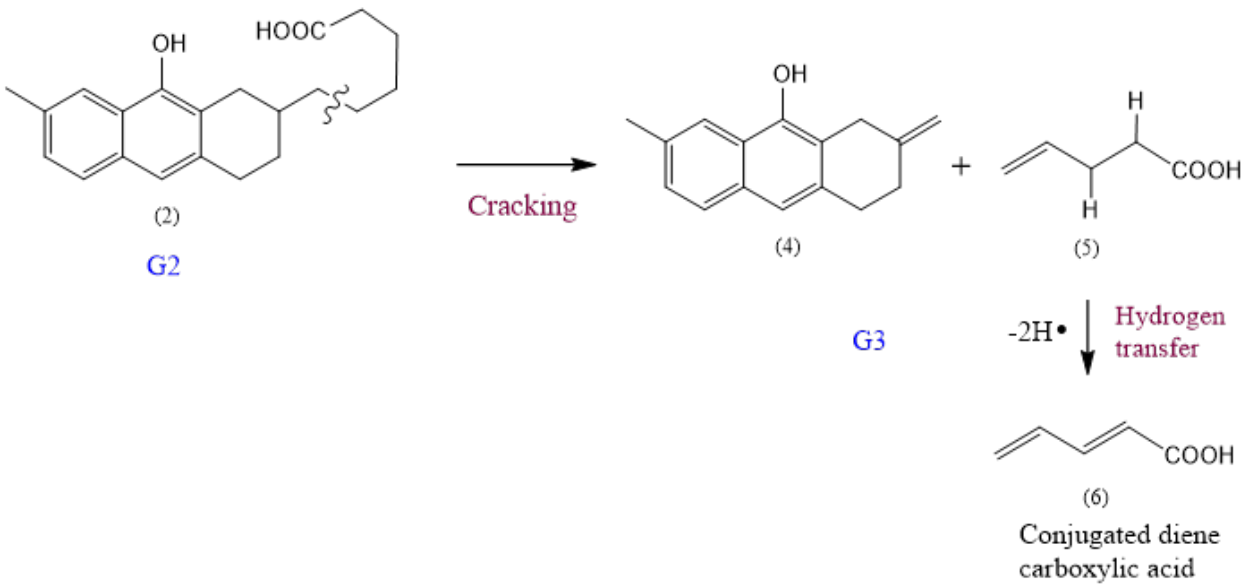


Figure S11: Proposed reaction pathway of group 2 to group 3 conversion.

Compounds (4), (5) and (6) are all representatives of G3, where (4) has aromatic and olefinic C=C bonds, (5) is an olefinic carboxylic acid while (6) has conjugated C=C double bonds with the C=O of the carboxylic acid group. In a similar way, to realize the composition of PC4, we look at Figure S12 and Figure S13 that depict the conversion of G2 to G4 and G1 to G4 respectively. Stretching of C=O at 1730 cm^{-1} and the absorption of acylic C-O at 1202 cm^{-1} for PC_4 indicated the presence of ester/anhydride-type species. Furthermore, among the aromatic C-H bends, the intensity for the ortho-substituted aromatics was the highest.

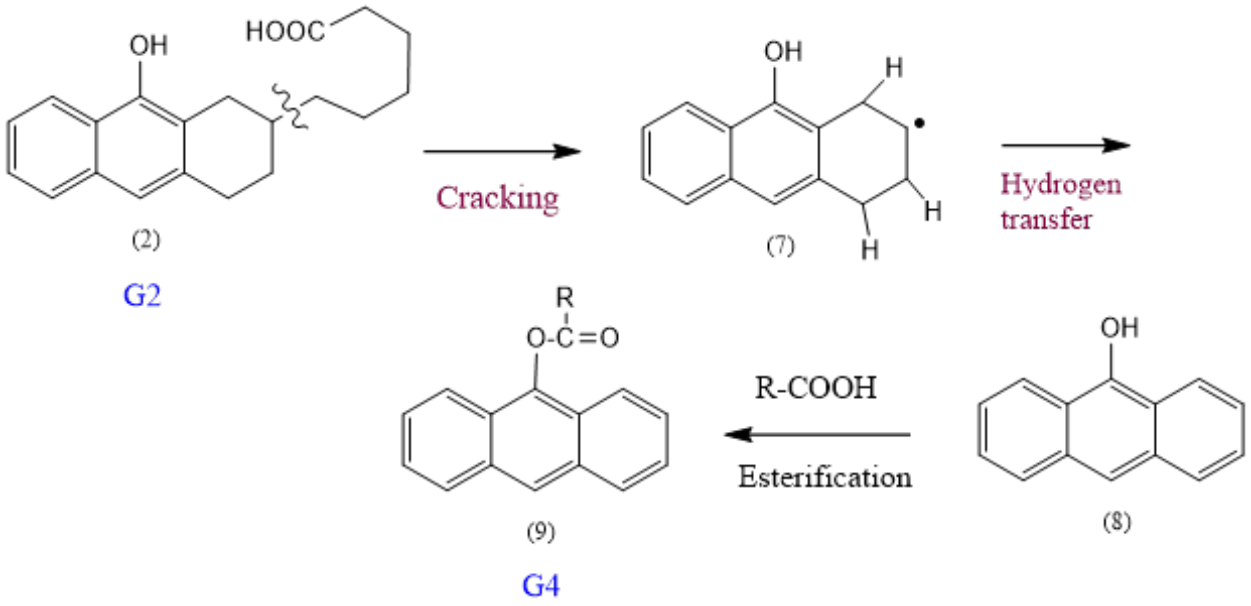


Figure S12: Proposed reaction pathway of group 2 to group 4 conversion.

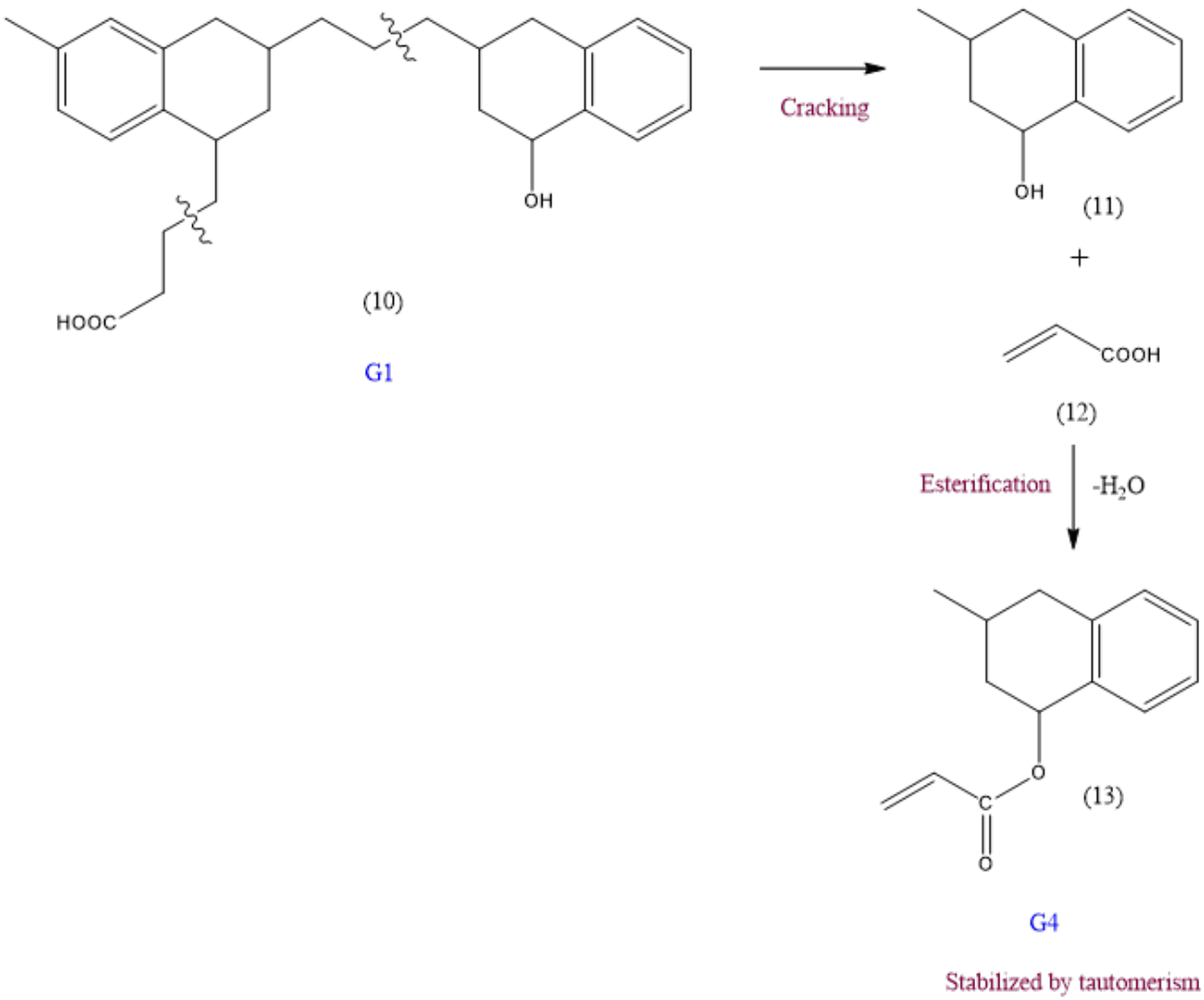


Figure S13: Proposed reaction pathway for group 1 to group 4 conversion.

Compound (9) in Figure S12 and compound (13) in Figure S13 are good representatives for G4 species. Compound (9) has 3 fused aromatic rings out of which the first and the 3rd ring excluding the middle one is ortho-substituted while compound (13) is entirely ortho-substituted. Although compound (13) is stabilized by tautomerism due to the olefinic conjugation with the C=O of the ester group, compound (9) is a better representation of G4 since the middle ring has para- and meta- substitutions as well.

Once the representative molecules for each pseudo-component were identified, a reaction pathway was developed according the algorithms in Bayesian structure learning. Here, the hill-climbing and MMHC networks are chosen and the reason for this has been highlighted

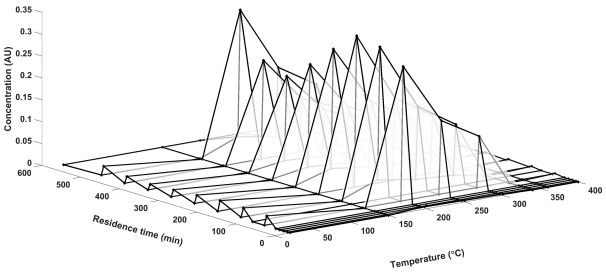
at the start of this section. From Figure S9, it can be seen that G1 →G2 has maximum arc-strength indicating the most probable reaction, followed by G2 → G3, G2 → G4 and G1 → G4 in decreasing order. The proposed conversion chemistries are given in Figure S10–S13. Conversion of G1 → G2 is the easiest since it involves only hydrogen transfer from the middle or end rings to terminate other free radicals in the bitumen matrix or alternatively get transferred to other aromatics. Bond dissociation energy of benzylic C-H is 301 kJ/mol, which is 30 kJ/mol lesser than C-H in aliphatics.[4] Compound (2) can undergo cracking in the aliphatic side chain to yield olefins (4) and (5), which can further lose 2 hydrogens to give a conjugated diene pentanoic acid. The conjugated dienoic acid is stabilized by double bond resonance. This chemistry provides a path from G2 → G3, that requires an additional step as compared to G1 →G2 and is depicted in Figure S11. The same sequence of reactions is possible with compound (3) as the starting material for G2 but in that case, only the olefin will be present only in the carboxylic acid product and the benzylic free radical would be stabilized by a hydrogen or an alkyl free radical. In this case, the alkoxy group in the middle ring would also exist, supporting the absorption at 1018 cm^{-1} for G2 species.

Next, in order to account for the formation of esters from G2-type species, cracking at the carbon attached to the naphthene ring in compound (2) needs to be considered (Figure S12). This would not be possible in (3) since it is much more difficult to break an sp^2 C-sp^3 C bond rather than an sp^3 C-C bond at these milder reaction conditions of ¡ 400 °C. Once the sp^3 C-C bond breaks, the ring can lose 3 more H free radicals to produce a tri-cyclic condensed aromatic phenol (8) (Figure S12). This can add to a carboxylic acid from the reaction medium to give an ester (9), that has all the characteristics of a G4 entity. The path from G2 → G4 involved an additional esterification step apart from cracking and hydrogen transfer through hydrogen disproportionation and hence is concomitant with the Bayesian networks produced from HC and MMHC where this path is the third most probable. G4 → G3 would involve hydrolysis of an ester but that requires the presence of water which is unlikely at these temperatures of bitumen conversion. This could be an explanation for the

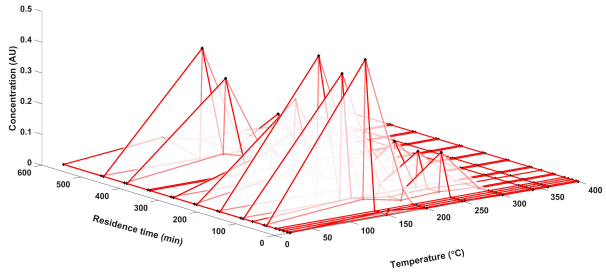
absence of this path in the MMHC network and being the least probable pathway in the HC network.

Lastly, to explain the conversion of G1 $\rightarrow$ G4 even if that was the least probable pathway in the MMHC-produced network, we consider a separate compound that satisfied the absorptions of G1 (compound (10) in Figure S13). This has characteristics to the archipelago structure[5] of asphaltenes where 2 aromatic cores are bridged by aliphatic chains. Compound (10) can crack in the aliphatic bridge and yield an o-substituted alcohol (11) while the other part is m- and p-substituted as well and is not shown. The side chain possessing a COOH group in (10) can crack and add to (11) and compound (13), which is an ester and also stabilized by tautomerism between the C=C and C=O groups. Compound (13) is another representative of G4.

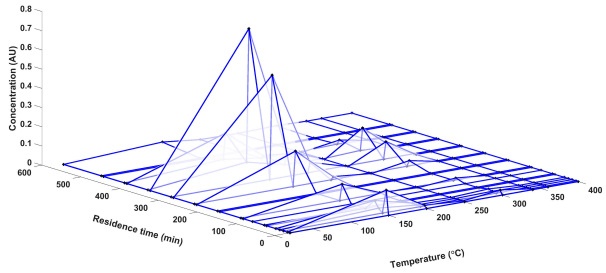
S5.2 Individual tensor factorization of ^{1}H-NMR spectra



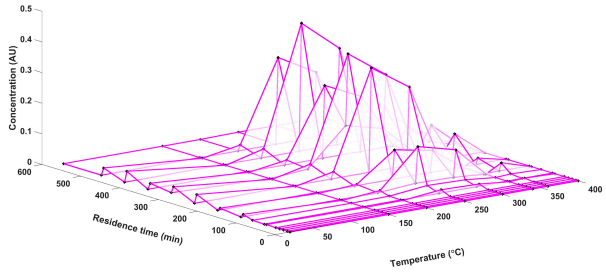
(a) Pseudo-component 1



(b) Pseudo-component 2



(c) Pseudo-component 3



(d) Pseudo-component 4

Figure S14: Concentrations of the pseudo-components across the reaction space of the ^{1}H-NMR spectra

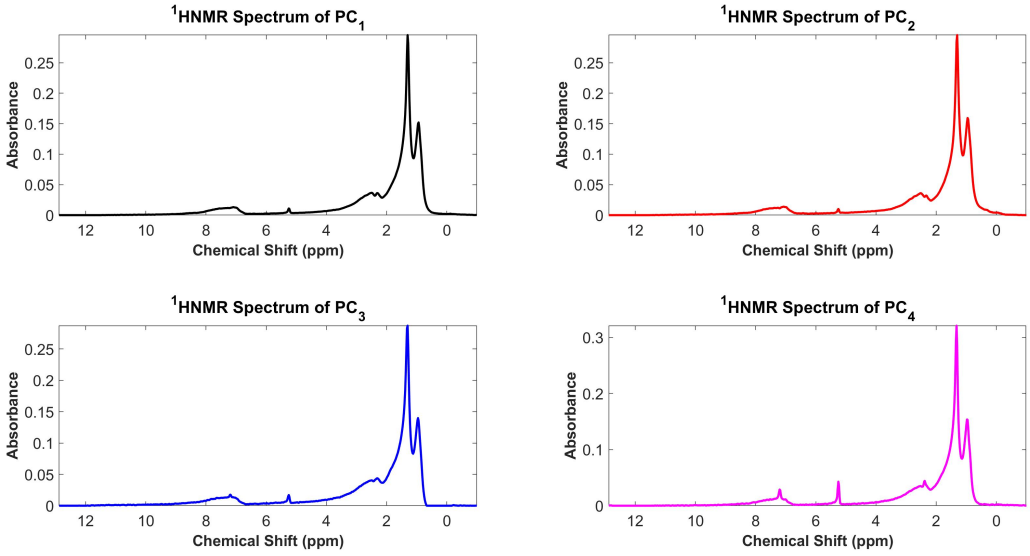


Figure S15: Spectra of pseudo-components from ^{1}H-NMR tensor decomposition

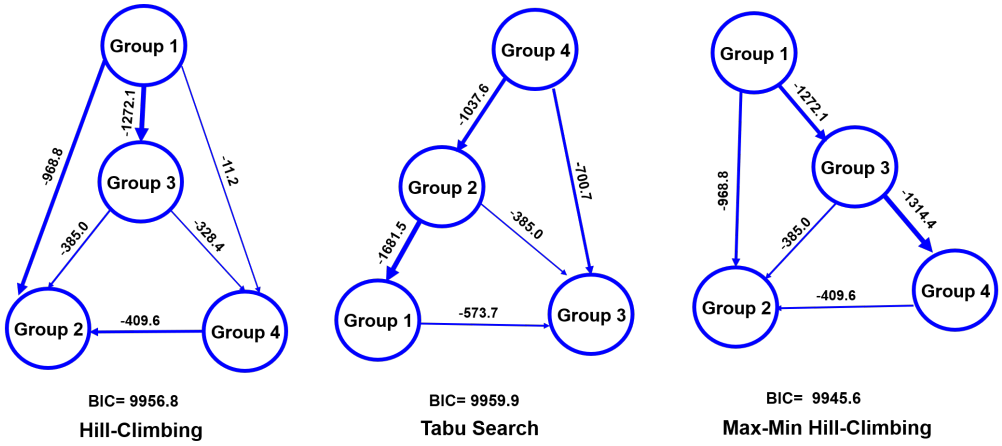
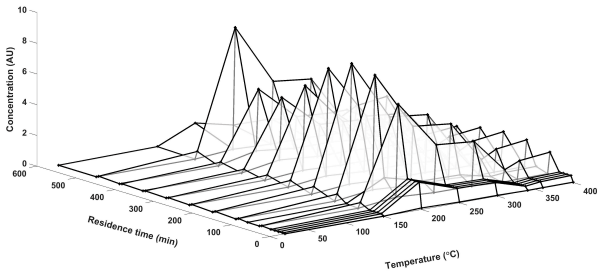


Figure S16: Bayesian networks from the unique ^{1}H-NMR pseudo-component spectra

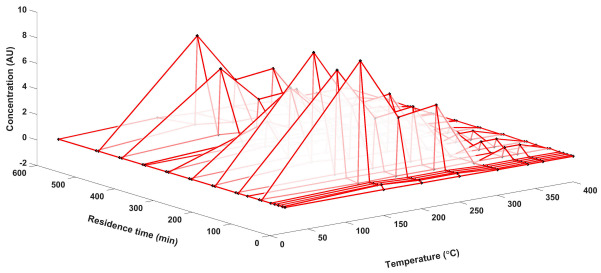
The drawback of this section is that NMR spectra alone does not provide as much information as the FTIR spectra, especially in the aromatic region since it just shows a single overlapped lump from 7 – 9 ppm (Figure S15). Peaks for aliphatic methylene and methyl protons are distinct and common to all pseudo-components with CH_2 showing higher intensity. All spectra also show the peak for benzylic proton at ∼2.5 ppm confirming the presence of aromatics but does not indicate the number of substitutions. One interesting observation was that PC_3 and PC_4 showed a peak for hydrogen attached to an alkyne group at 3.1 ppm

and this was also reflected in the FTIR spectra for the same pseudo-components (Figure S8). Triple bonds are quite stable and their possible participation in the reaction could be such that hydrogens from disproportionation could add across the triple bond. Another distinct characteristic of NMR profiles is the peak at ∼5.2 ppm that depicts hydrogen from methylene chloride that might be remaining in the converted samples. This is present in all pseudo-components but of higher intensity in PC_3, PC_4 and also falls in the olefin range. Overall, not much conversion chemistry can be proposed from NMR profiles alone so it is worthwhile to look at the joint decomposition in Section S5.3.

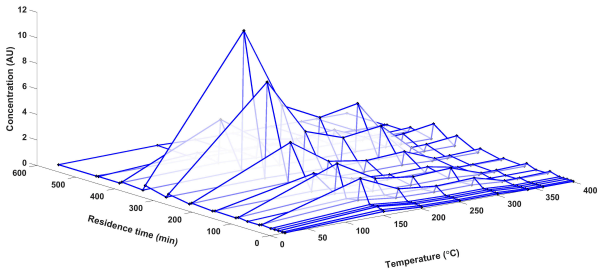
S5.3 Joint Gaussian tensor factorization



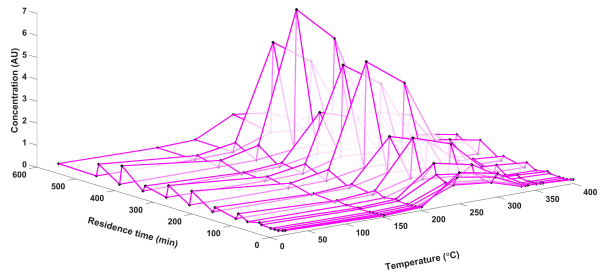
(a) Pseudo-component 1



(b) Pseudo-component 2



(c) Pseudo-component 3



(d) Pseudo-component 4

Figure S17: Concentrations of the pseudo-components across the reaction space from the joint decomposition of FTIR and ^{1}H-NMR spectra

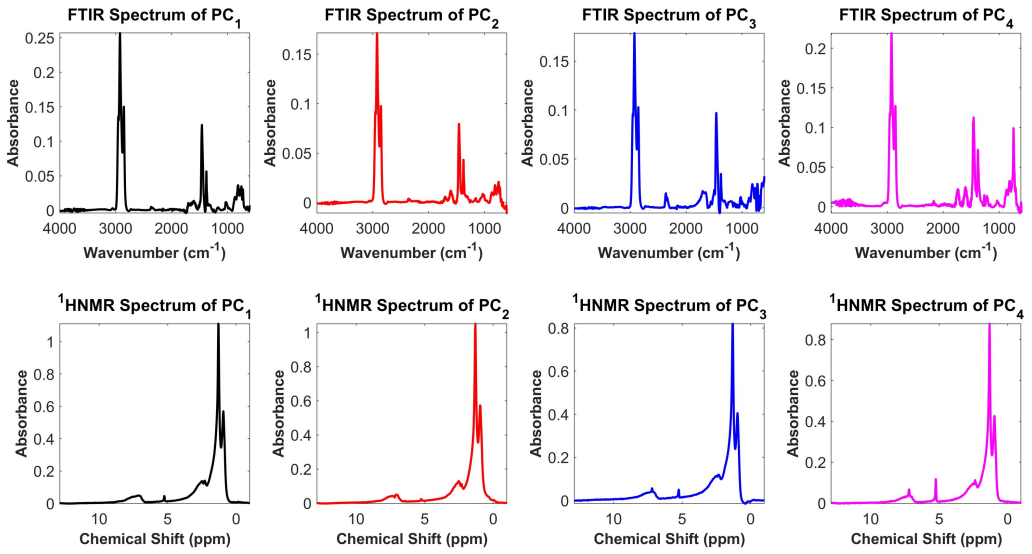


Figure S18: Spectra of pseudo-components from joint tensor decomposition

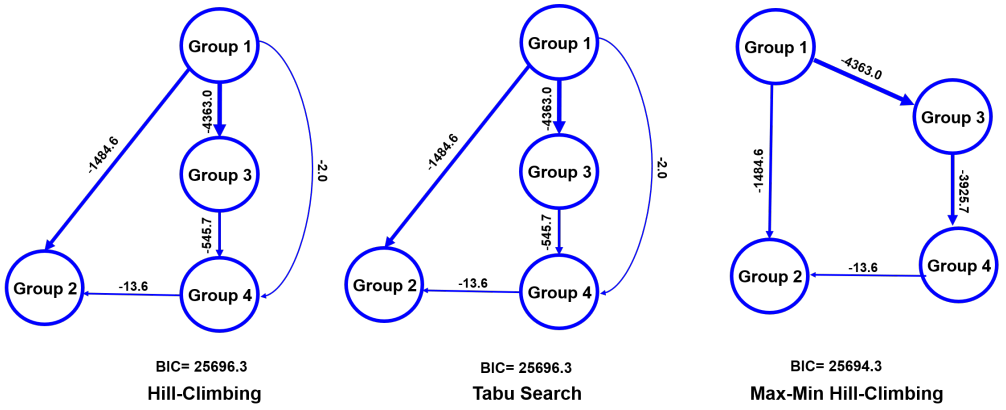


Figure S19: Bayesian networks from the unique joint pseudo-component spectra

Figure S17, Figure S18 and Figure S19 give the concentration profiles across time and temperature modes, spectral profiles for all 4 pseudo-components and the Bayesian networks obtained from tensor decomposition of FTIR and ^{1}H-NMR fused data, respectively, for the non-robust formulation. The absorption peaks for all pseudo-components were similar to those reported in Table 1. The Bayesian network structures are as reported in Figure S19. Here, G1 $\rightarrow$G3 was the most probable pathway which meant cracking leading to olefin formation had a higher chance of occurring than hydrogen transfer. Alcohol groups in these

olefins have more probability of finding carboxylic acids from the matrix to yield an ester (G4) and this pathway is the second most and third most probable in the MMHC and HC network, respectively. This is because carboxylic acids are more prominent and available to react in bitumen than alcohols.[6]

G4→ G2 had the lowest arc strength in the MMHC network and this meant that hydrolysis of esters was least probable which corroborated with the observations from the robust method. In conclusion, the robust method indicates a better flow in the reaction chemistry as hydrogen transfer occurs more easily than cracking. Also, a conjugated double bonded carboxylic acid like (6) that belongs to G3 would react slower than an unconjugated carboxylic acid (G2) to yield G4 esters, which is captured in robust formulation. Hence, overall, it is suggested that the robust formulation gives a better representation of bitumen conversion chemistry at these process conditions.

References

(1) Phan, A. H.; Cichocki, A. Fast Nonnegative Tensor Factorization for Very Large-Scale Problems Using Two-Stage Procedure. *CAMSAP 2009 - 2009 3rd IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing* **2009**, 297–300.

(2) Nguyen, V.-D. Contributions to Fast Matrix and Tensor Decompositions. Ph.D. thesis, Université d'Orléans, 2016.

(3) Dunlavy, D. M.; Kolda, T. G.; Acar, E. *Poblano V1.0: A Matlab Toolbox for Gradient-Based Optimization*; 2010.

(4) Blanksby, S. J.; Ellison, G. B. Bond Dissociation Energies of Organic Molecules. *Accounts of Chemical Research* **2003**, *36*, 255–263.

(5) McKenna, A. M.; Marshall, A. G.; Rodgers, R. P. Heavy Petroleum Composition. 4. Asphaltene Compositional Space. *Energy & Fuels* **2013**, *27*, 1257–1267.

(6) Naghizada, N.; Prado, G. H. C.; de Klerk, A. Uncatalyzed Hydrogen Transfer During 100–250 °C Conversion of Asphaltenes. *Energy & Fuels* **2017**, *31*, 6800–6811.