Matrix completion methods for thermodynamic property prediction

F.S. Middleton, J.T. Cripwell*

Department of Chemical Engineering, Stellenbosch University, Banghoek Road, Stellenbosch 7600, South Africa.
Email: *cripwell@sun.ac.za. Tel: +27 21 808 4108

BACKGROUND

- Matrix completion methods (MCMs) are proposed for **pseudo-data generation** towards fundamental model improvement.
- MCMs leverage sparse data sets, offering an advantage over other machine learning methods.
- The MCM was used to predict the excess enthalpy of binary liquid mixtures to determine if the method could be used on **composition dependent data**. It has been used previously on activity coefficients at infinite dilution [1].
- The pseudo-data can be used for parameterising thermodynamic models: potential to decrease reliance on experiments.

ARRAY FORMATION

- **4-way array:** compound 1 and 2 (categorical), mole fraction, and temperature (continuous).
- **Discretised** composition after using **polynomial fits** to experimental data: interpolate random experimental intervals to generate 5% “slices” of compound 1.

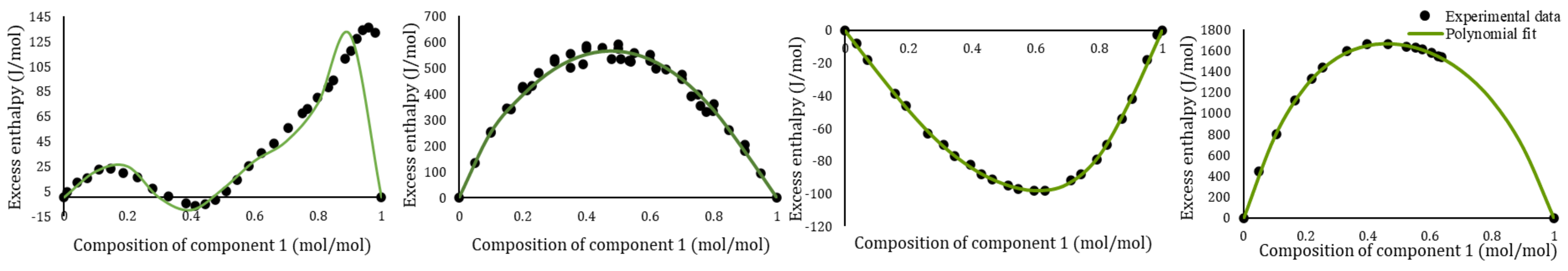


Fig 1: Interpolated data for (left to right) ethane and methanol, 1-hexene and cyclohexane, ethane and propane, and butanone and dodecane at 298.15K.

- Isothermal and constant composition matrix slices can be completed as they have **randomly missing entries** [2].
- Symmetrical matrices across diagonal: halves array size.
- The MCM can find **compound ‘personalities’** using **SVD** (singular value decomposition).

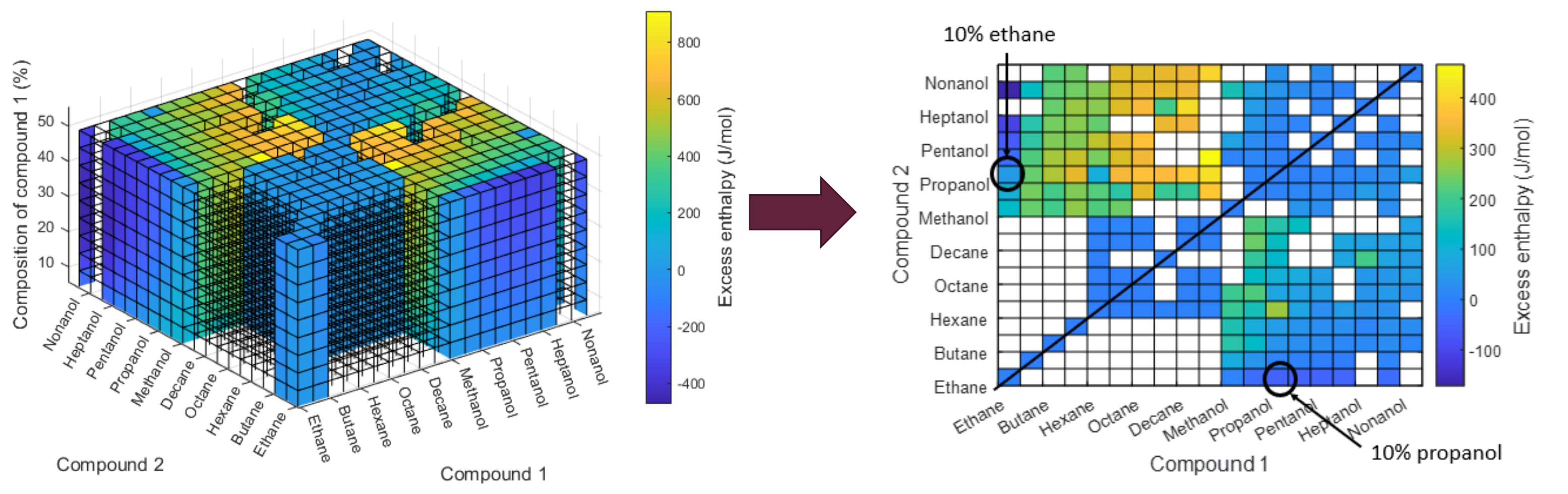


Fig. 2: A small 3-way array at 298.15K and a matrix slice at 10% of compound 1 illustrating diagonal symmetry of composition. The upper triangle contains 10% of ethane and 90% propanol, and the lower triangle contains 10% propanol and 90% ethane.

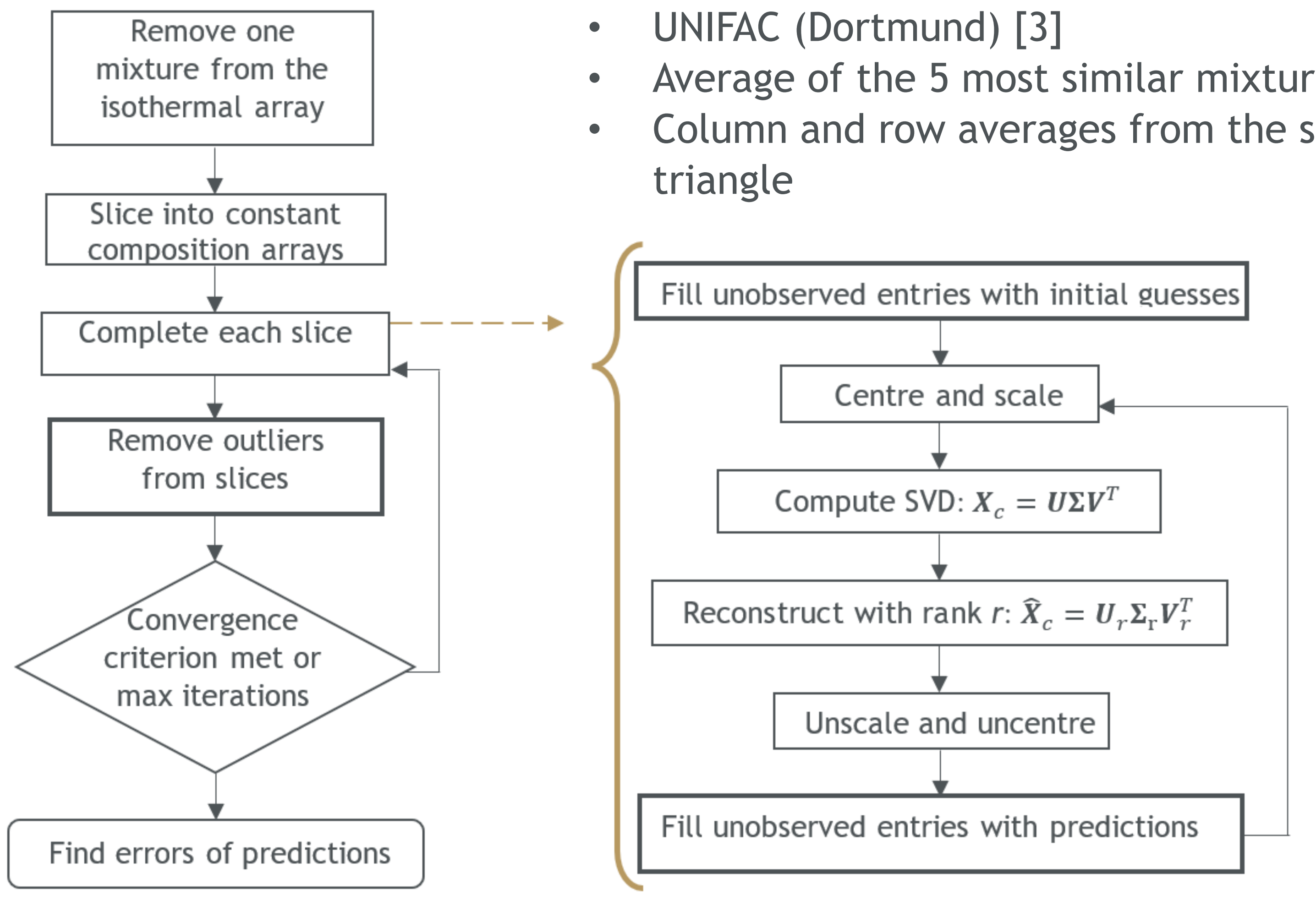
METHODOLOGY

- MCM algorithm repeated for every array mixture (LOOCV).
- **Initial guesses** must be used for SVD to be applied.
- The **coherence of predictions** was maintained by removing outliers.

Approach to initial guesses

Types of guesses attempted for missing data:

- UNIFAC (Dortmund) [3]
- Average of the 5 most similar mixtures
- Column and row averages from the same triangle


Remove one mixture from the isothermal array

Slice into constant composition arrays

Complete each slice

Remove outliers from slices

Convergence criterion met or max iterations

Find errors of predictions

Fill unobserved entries with initial guesses

Centre and scale

Compute SVD: $X_c = U\Sigma V^T$

Reconstruct with rank r: $\hat{X}_c = U_r \Sigma_r V_r^T$

Unscale and uncentre

Fill unobserved entries with predictions

Figure 3: MCM algorithm for an isothermal array.

RESULTS

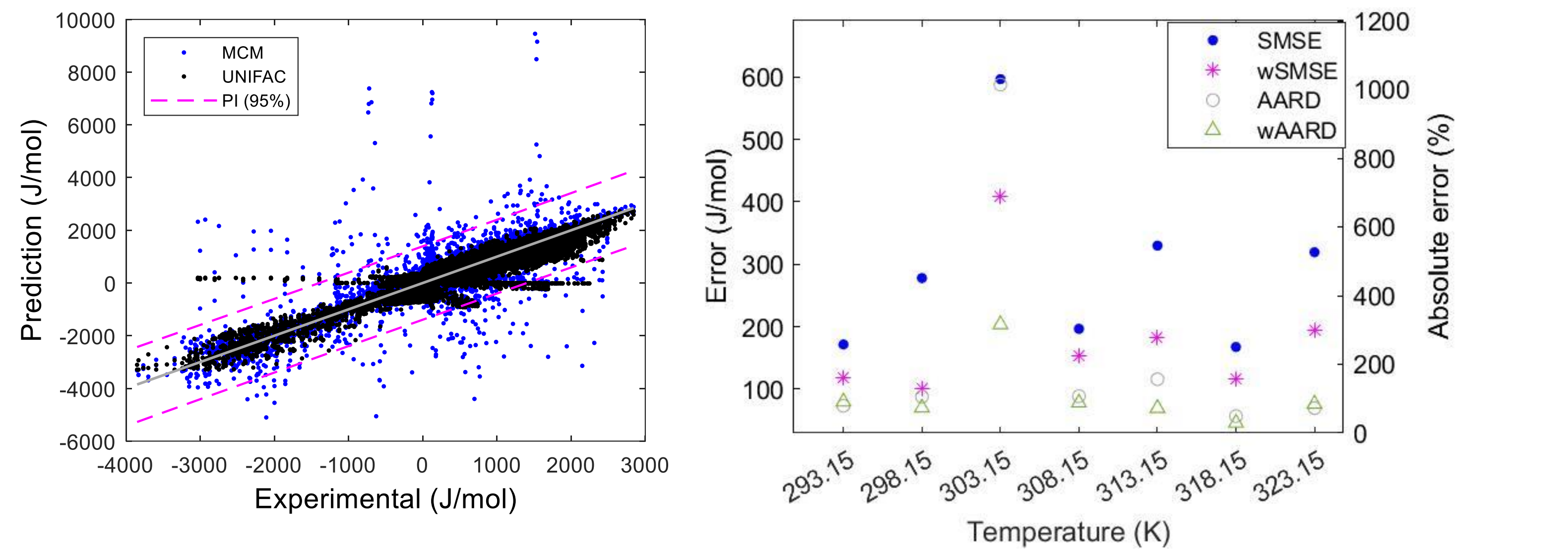


Fig 4: Parity plot of the MCM predictions at 298.15K (left), square root of the mean squared error (SMSE) and average absolute relative deviation (AARD) of predictions and winsorized counterparts (5% best and worst predictions removed) for the predictions for the temperatures attempted (right), using UNIFAC (Dortmund) initial guesses.

Good MCM predictions in evidence for:

- UNIFAC initial guesses → encodes explicit features.
- High % observed data for similar mixtures.
- >12% observed entries in the array → array at 303.15K was 11% observed.
- Binary association code (BAC) groups [4] were used to assess performance for different types of intermolecular forces. BAC_5 (mixtures in which self-association takes place) performed best.

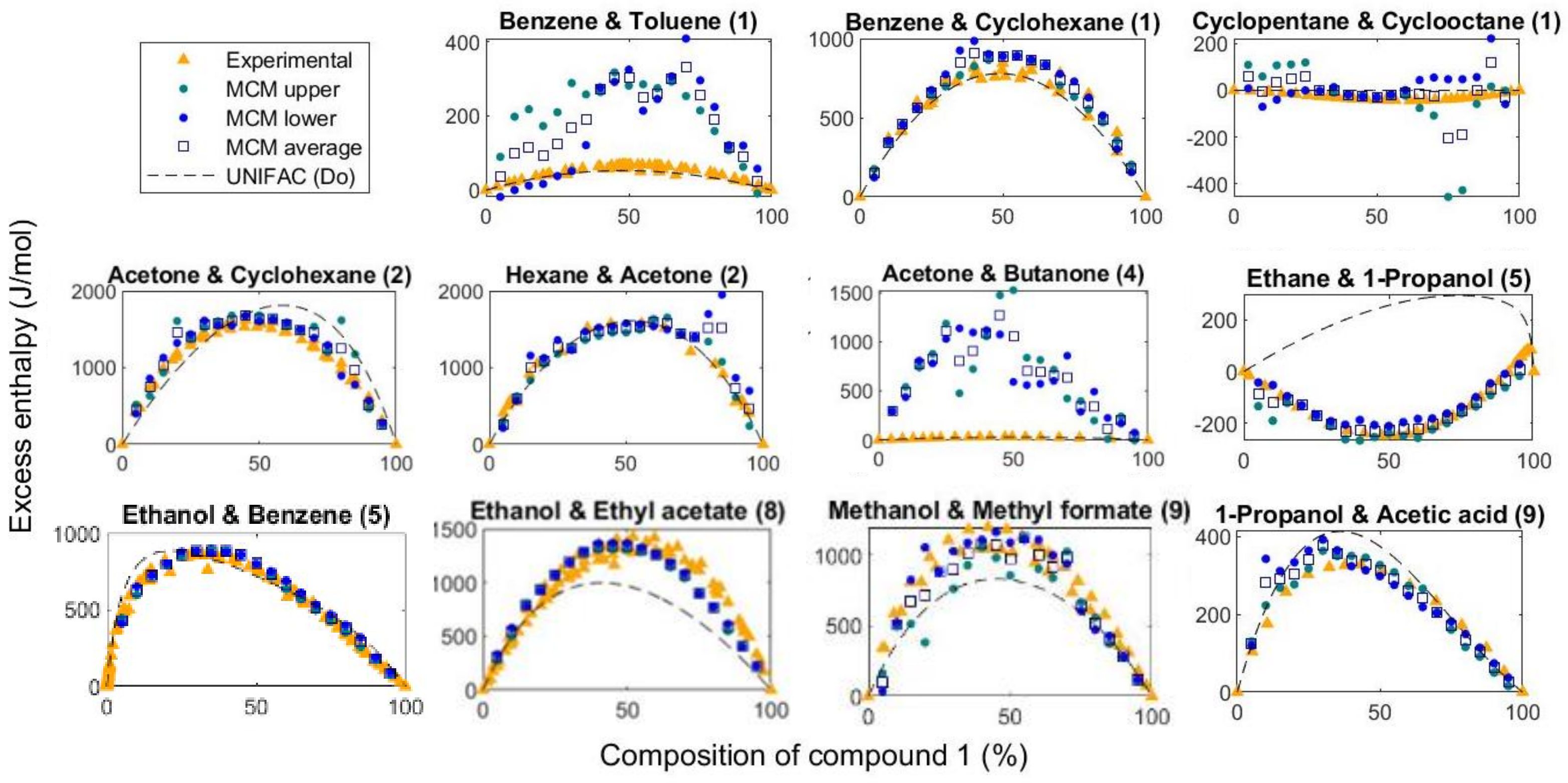


Fig 5: Some results of the MCM on the 298.15K using UNIFAC (Dortmund) initial guesses, compared to UNIFAC (Do) and experimental data. BAC groups given in brackets.

CONCLUSION

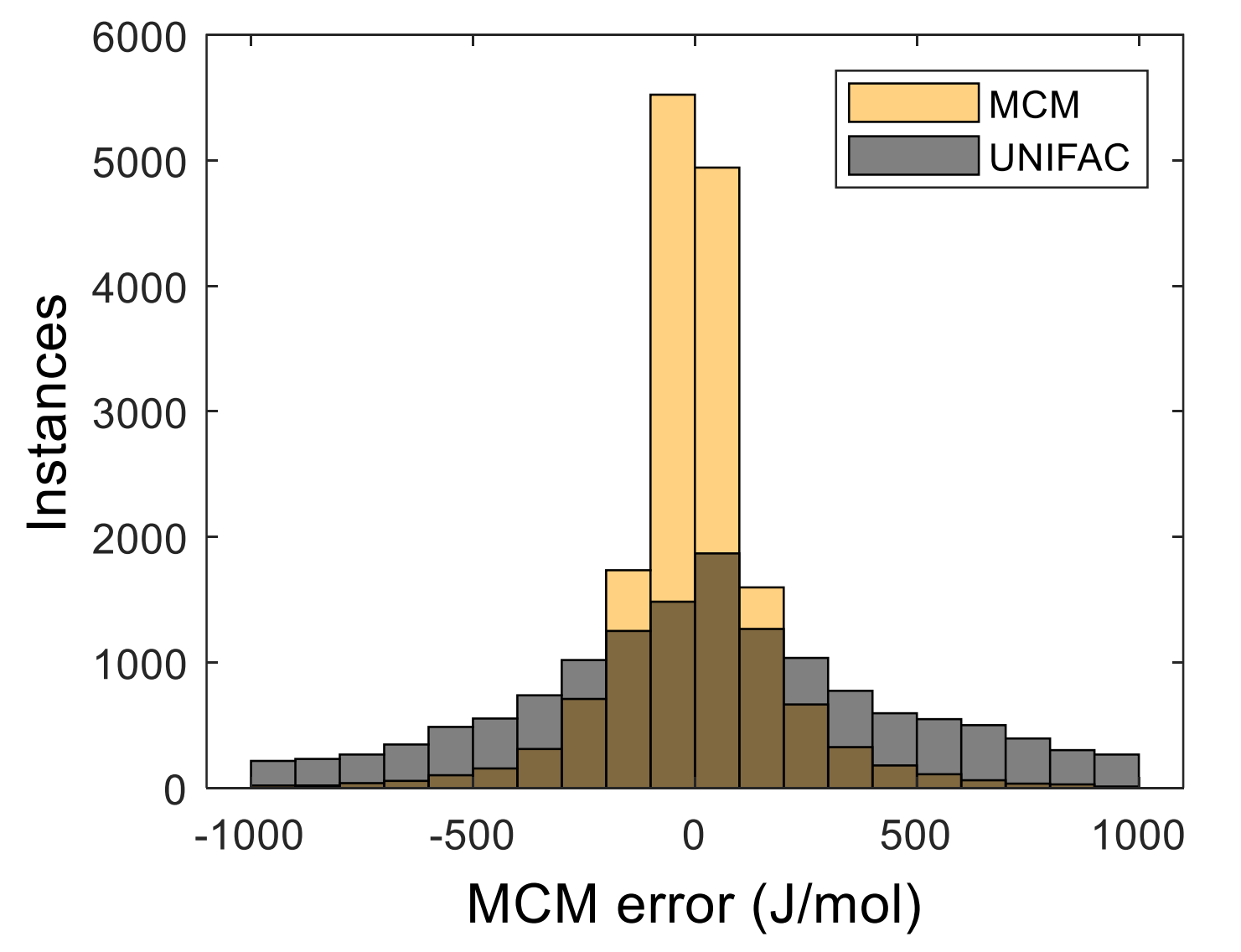


Fig 6: Histogram of the MCM and UNIFAC predictions at 298.15K.

- Smooth predictions for compositional variation, therefore the coherence constraint was successful.
- The MCM outperformed UNIFAC (Dortmund) for 85% of mixtures.
- The MCM can be used for pseudo-data generation.

References: [1] *J. Phys. Chem. Lett.* **11** (2020): 981-985; [2] *Chemometr. Intell. Lab. Syst.* **75 (2)** (2005) 163-180; [3] *Ind. Eng. Chem. Res.* **26** (1987) 1372-1381; [4] *Ind. Eng. Chem. Res.* **59** (2020) 14981-15027

Scan me to visit my GitHub repository