Electronic Supplementary Material (ESI) for Digital Discovery.

Supporting Information

Machine Learning Enhanced Infrared Spectroscopy Analysis for Rapid Mixture Characterization

Andrea Angulo, Lankun Yang, Eray S, Aydil and Miguel A. Modestino*.

Department of Chemical and Biomolecular Engineering, Tandon School of Engineering, New York University

*E-mail: modestino@nyu.edu

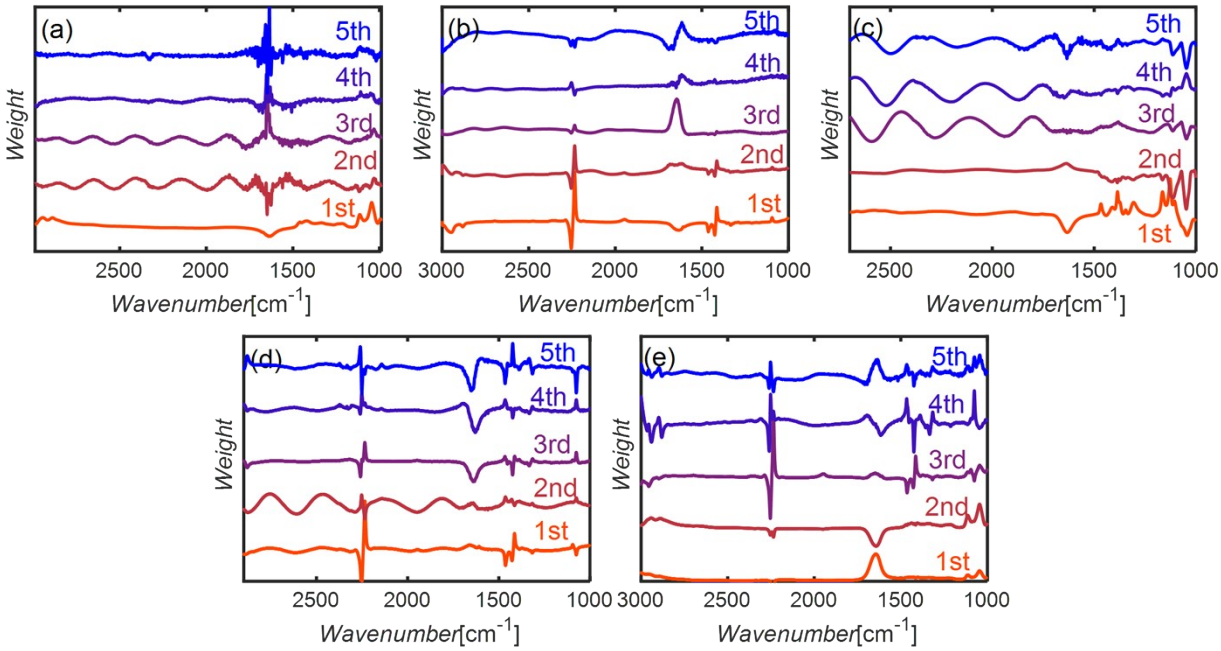

Figure S1. Weight (or loading) per component, for the five components with the highest explained variance vs. wavenumber, resulting from PCA applied to (a) 1-Gly, (b) 2-AN, (c) 3-Gly, (d) 3-AN and (e) 4-AN mixtures

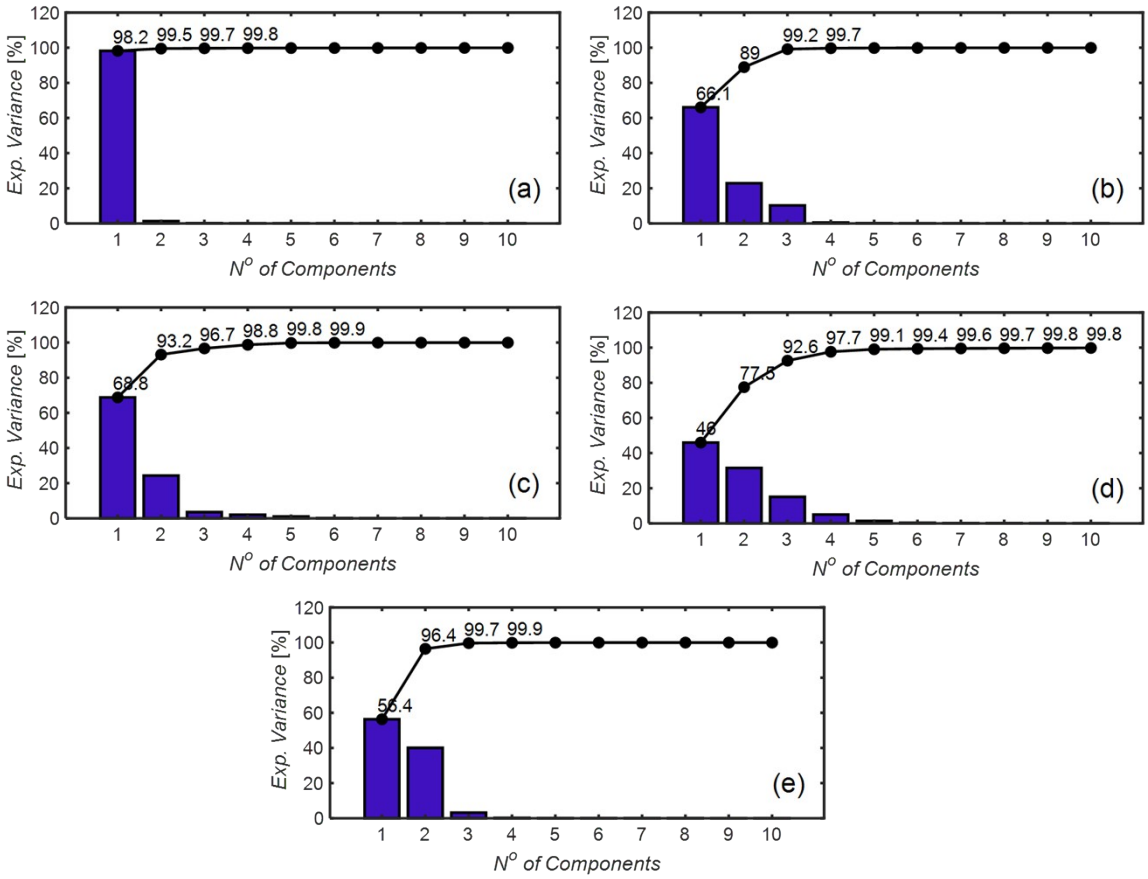

Figure S2. Principal component analysis explained variance, individual (blue bars) and cumulative (black line), for (a) 1-Gly, (b) 2-AN, (c) 3-Gly, (d) 3-AN and (e) 4-AN mixtures

Table S1. Concentrations of the prepared aqueous stock solutions

Prepared Aqueous Solution	% wt
Glycerol	5.0-10.0
IPA	8.8
1-Butanol	7.3
AN	4.1
ADN	4.2
PN	3.8

Table S2. Components considered for systems of different complexity for synthetically generated data

Number of components	Components
1	ADN
2	ADN, AN
3	ADN, AN, PN
4	ADN, AN, PN, EDTA
5	ADN, AN, PN, EDTA, PO_4^{-3}
6	ADN, AN, PN, EDTA, PO_4^{-3}, TMA

Table S3. Components considered for 3 different chemical systems

Mixture Label	Components
ADN-containing	ADN, AN, PN, EDTA, PO_4^{-3}, TMA
Glycerol-containing	Glycerol, acetic acid, dihydroxyacetone, formic acid, glycolic acid, oxalic acid
Random	Ethylene glycol, propionic acid, 1,2-propanediol, phenol, hexane, benzene

Table S4. Hidden layers sizes and activation functions used for ANN models (multilayer perceptron regressor) for each type of mixture considered. The rest of the hyperparameter other than layer sizes and activation function are the same in all cases, which are the following: tol = 1e-5, random_state = 0, solver = 'lbfgs', learning_rate = 'adaptive',batch_size = 80 for simulated data, batch_size = 10 for experimental data

Mixture	Hidden layers size	Activation Function
Simulated data: 1 component AN 2 components AN, ADN 3 components AN, ADN, PN 4 components AN, ADN, PN, EDTA 5 components AN, ADN, PN, EDTA, PO_4^{-3}	(12,)	Rectifier
Simulated data: 6 components AN, ADN, PN, EDTA, PO_4^{-3}, TMA 6 components Acetic acid, Dihydroxyacetone, Formic acid, Glycerol, Glycolic acid, Oxalic acid, Water 6 components Random	(20,)	Rectifier
Experimental data: 1 component Glycerol in water with	(2,)	Rectifier
Experimental data 2 components AN ADN with 3 components Glycerol, IPA, 1-butanol 4 components AN ADN PN Glycerol	(10,10)	Identity
3 components AN ADN PN	(20,)	Rectifier

Table S5. Coefficient of determination R^2 for different activation functions, for ANN regression algorithms for experimental mixtures. Under the coefficient of determination is also noted the number of neurons per layers for each case.

	Activation Function			
Mixture	'identity'	'relu'	'tanh'	'logistic'
1-Gly	0.9800 (2,)	0.9808 (2,)	0.9778 (2,)	0.9808 (2,)
2-AN	0.9810 (10,10)	0.9518 (15,15)	0.9314 (15,15)	0.9262 (10,10)
3-Gly	0.9815 (10,10)	0.959 (15,15)	0.9695 (10,10)	0.9712 (10,)
3-AN	0.9215 (15,15)	0.9686 (20,)	0.9541 (20,)	0.7642 (10,)
4-AN	0.8527 (10,10)	0.7768 (20,)	0.7577 (20,)	0.5832 (20,)