Supplementary Information for "Multicomponent Raman spectral regression using complete and incomplete models and convolutional neural networks"

Derrick Boateng,[a] Chuanzhen Hu,[b] Yichuan Dai,[b] KaiqinChu,[c] Jun Du*,[a], Zachary J. Smith*,[b]

[a]National Engineering Laboratory for Speech and Language Information Processing, Department of Electronic Engineering and Information Science, University of Science and Technology of China.

[b]Key Laboratory of Precision Scientific Instrumentation of Anhui Higher Education Institutes, Department of Precision Machinery and Precision Instrumentation, University of Science and Technology of China.

[c]Suzhou Institute of Advanced Research, University of Science and Technology of China.

Table of contents

1. List of ternary mixtures

Table S1 Ternary mixture sets

Ternary Mixture	Component
1	CL CytoC DNA
2	Erg DNA Prot
3	LPC OPC PC
4	PE PI PA
5	PS SPH Prot

2. AsLS performance at different asymmetry parameters

As discussed in Section 3.1 in the main text, the performance of the AsLS is highly dependent on the user choice of asymmetry parameter (the p value) to provide an acceptable result, and in our simulations we found empirically that a p value of 0.001 provides a good fit for the considered mixture datasets. We show in Fig. S1, RMSEP at different asymmetry parameters, and in Fig. S2, examples of spectral fits at different p values. Those results clearly indicate that it's a visual judgement call which p value is considered "best", particularly as different components show different trends versus asymmetry parameter. 0.001 represents the best compromise. However, selecting this parameter required ground truth knowledge of the true concentrations, a condition which is typically not met.

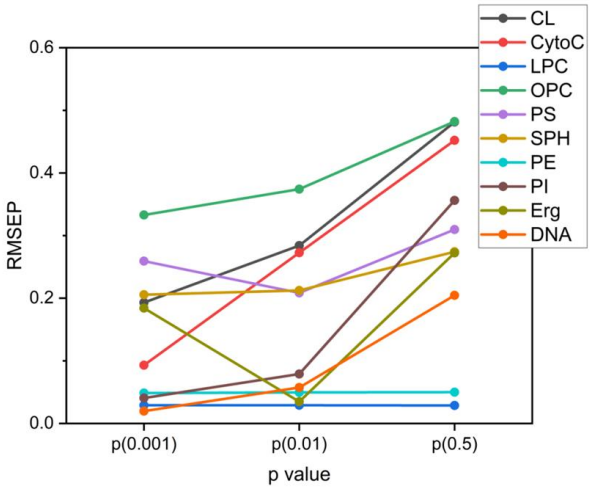

Fig. S1 The RMSEP versus asymmetry parameter for various chemical components in ternary mixtures.

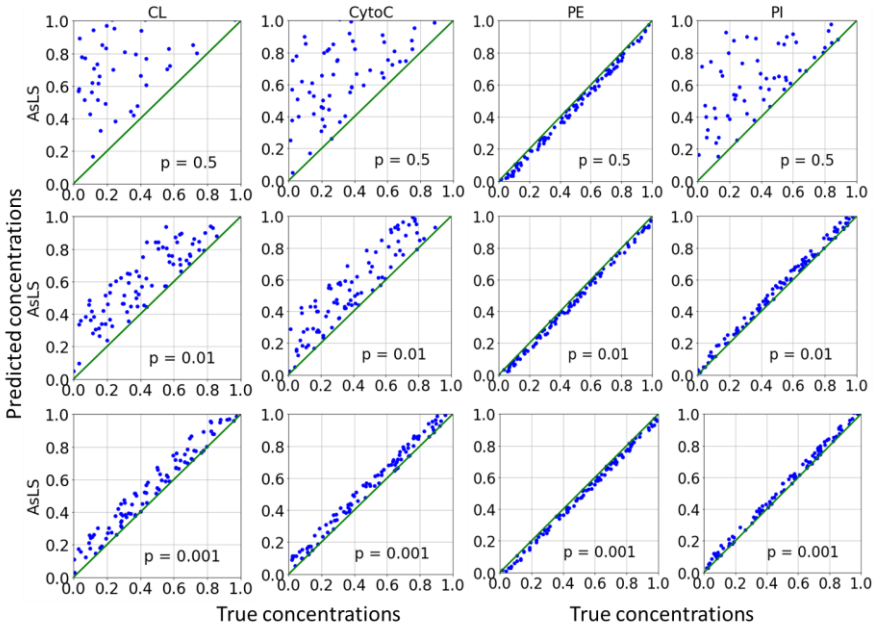

Fig. S2 Examples of spectra fits at the different p values using two representative sets of simulated mixtures.

3. Sensitivity of CNN to hyperparameter selection

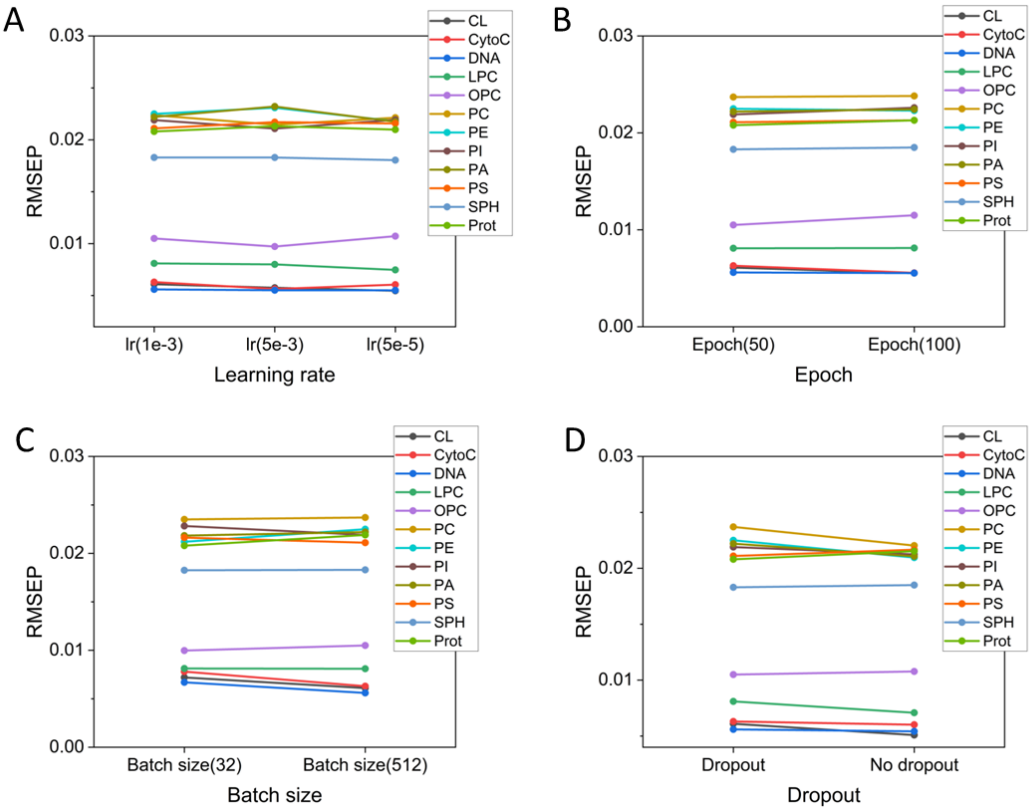

Fig. S3 Sensitivity of the model performance to hyperparameter selection. (A) The effect of different learning rates. (B) The effect of epoch number. (C) The effect of batch size. (D) The effect of dropout regularization.

CNN models typically have a large number of hyperparameters that are often set at default values or adjusted to optimize the learning process. As an advantage of CNN observed in the main text is less reliance on expert users, we conducted a sensitivity analysis of the CNN model performance to changes in the configuration of the hyperparameters settings using four representative spectral mixture sets in the complete spectral model scenario, namely the learning rate, epoch, dropout and batch size.

To investigate each hyperparamter's effect we hold all other hyperparameters fixed as baseline model setting and vary the hyperparameter of interest. For the setting conditions of each hyperparameter we evaluated the model's performance in terms of the RMSEP. We show in Fig. S3 the results of the sensivity analysis. As shown in Fig. S3, the CNN has minimal dependence on hyperparameter selection, highlighting that a "vanilla" architecture trained using default hyperparameters produces state-of-the-art results without requiring user tuning.

4. Prediction results in simulated underdetermined models

In the main text Section 3.2, we showed in Fig. 5 the prediction performance using a representative simulated ternary mixture of LPC, OPC and PC, where PC was deleted. In Figs. S4 and S5, the performance on the complete group of simulated mixtures (summarized in Table 2) is presented.

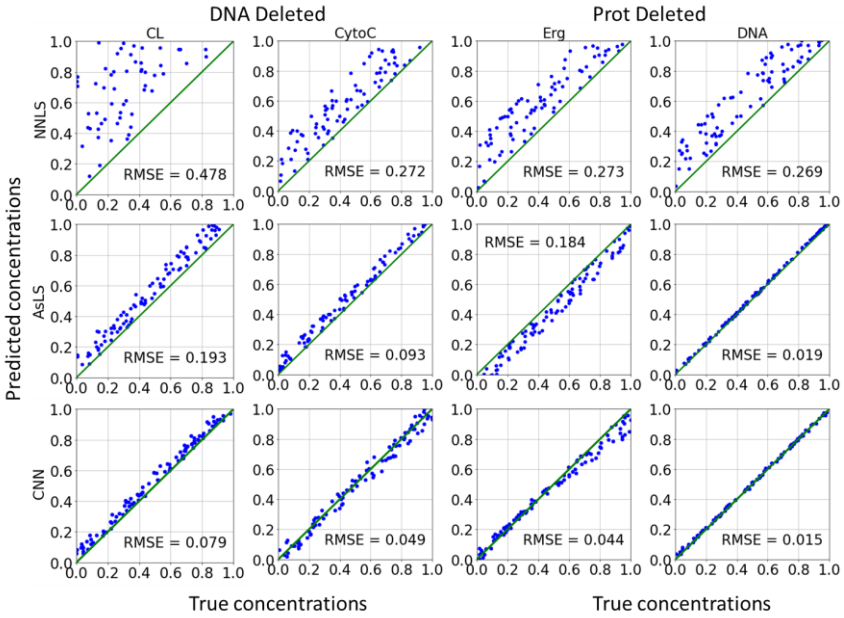

Fig. S4 Comparison of the prediction performance by the different methods in incomplete spectral models

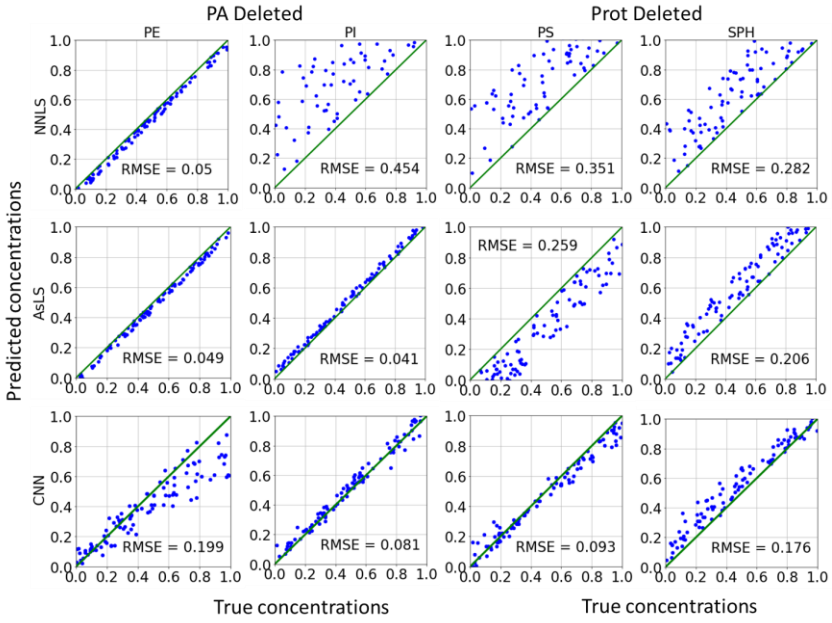

Fig. S5 Comparison of the prediction performance by the different methods in incomplete spectral models.

5. Full comparison of liposome data between CNN and AsLS

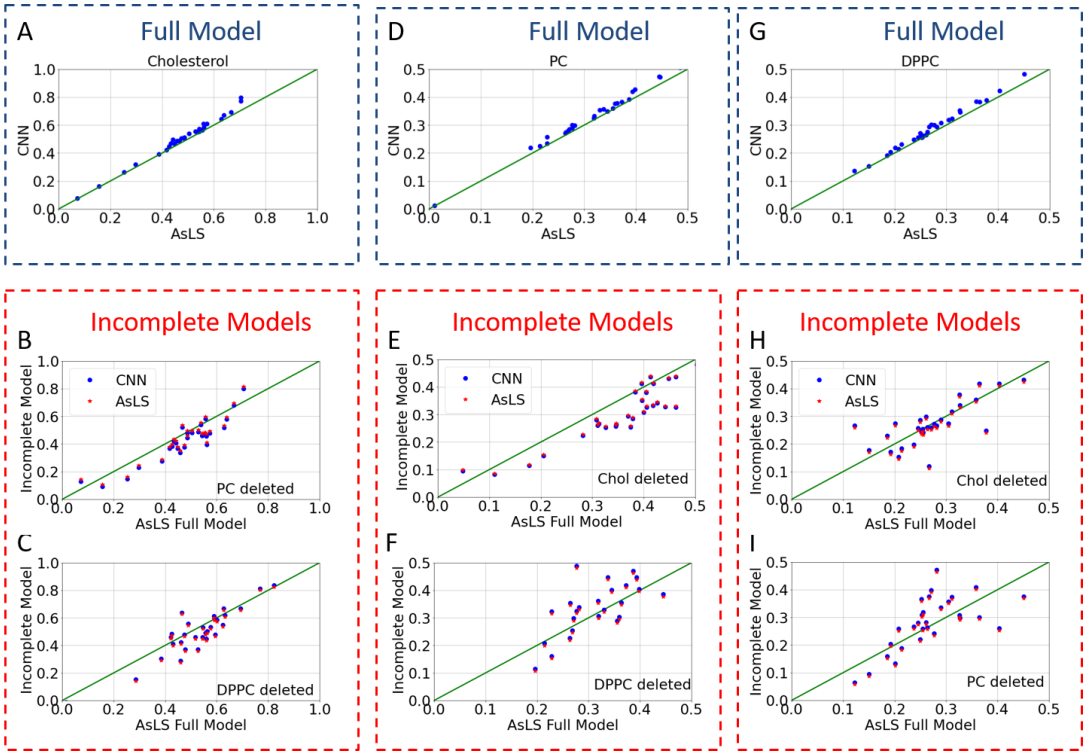

Fig. S6 Comparison of prediction results in experimental liposome data. (A) CNN and AsLS agreement in full spectral model for Cholesterol. (B) Comparison of prediction results by CNN and AsLS using an incomplete model with deletion of PC. (C) Comparison of CNN and AsLS prediction results using an incomplete model with deletion of DPPC. (D) CNN and AsLS agreement in full spectral model for PC. (E) Comparison of prediction results by CNN and AsLS using an incomplete model with deletion of Cholesterol. (F) Comparison of prediction results by CNN and AsLS using an incomplete model with deletion of DPPC. (G) CNN and AsLS agreement in full spectral model for DPPC. (H) Comparison of prediction results by CNN and AsLS using an incomplete model with deletion of Cholesterol. (I) Comparison of prediction results by CNN and AsLS using an incomplete model with deletion of PC.

6. Prediction results using data after preprocessing

We note in the Conclusion Section in the main text that we explored using CNN to predict concentrations using data after pre-processing. In Figs. S7 and S8, the results of the CNN evaluation using raw and pre-processed spectra in complete spectral models and incomplete spectra models for a ternary mixture is presented.

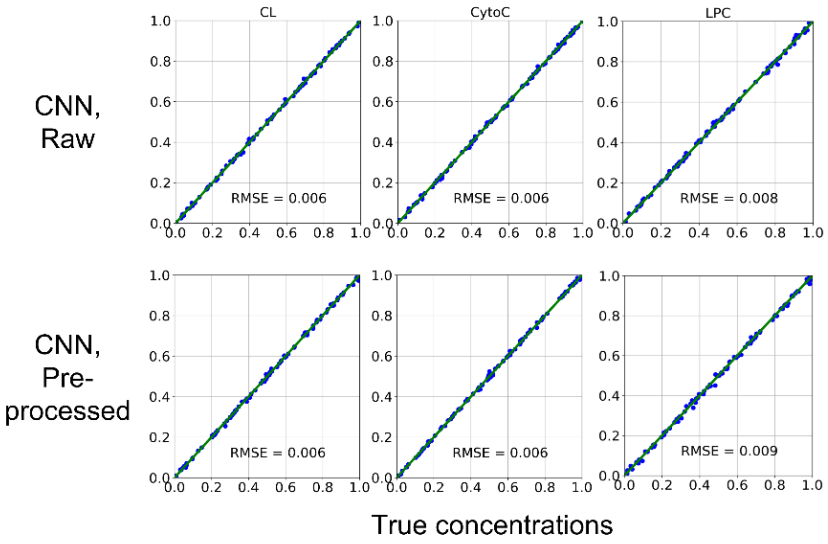

Fig. S7 Performance of CNN regression on raw spectra (top) and pre-processed spectra (bottom) for a ternary mixture and complete spectral model.

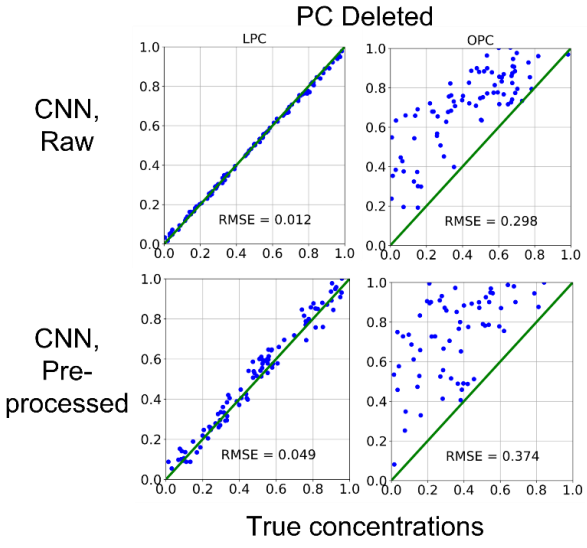

Fig. S8 Performance of CNN regression on raw spectra (top) and pre-processed spectra (bottom) for a ternary PC, LPC, OPC mixture where PC was not included in the spectral model.

7. CNN evaluation in complete spectral models using raw and pre-processed spectra

Table S2 CNN evaluation in complete spectral models using raw and pre-processed spectra

Pure component	CNN (raw)	CNN (pre-processed)
CL	0.0061	0.0064
CytoC	0.0063	0.0060
DNA	0.0056	0.0059
Prot	0.0132	0.0148
LPC	0.0081	0.0093
OPC	0.0105	0.0093
PC	0.0237	0.0225
PE	0.0225	0.0219
PI	0.0219	0.0189
PA	0.0222	0.0252
PS	0.0211	0.0201
SPH	0.0183	0.0184
Average	0.0131	0.0149

8. CNN evaluation in incomplete spectral models using raw and pre-processed spectra

Table S3 CNN evaluation in incomplete spectral models using raw and pre-processed spectra[a]

Mixture	Pure component	CNN (raw spectra)	CNN (preprocessed spectra)
	CL	0.0792	0.0635
1	CytoC	0.0493	0.0706
	DNA	×	×
	LPC	0.0116	0.0494
2	OPC	0.2982	0.3738
	PC	×	×
	PE	0.1985	0.1835
3	PI	0.0811	0.0970
	PA	×	×
	PS	0.0925	0.0783
4	SPH	0.1765	0.2423
	Prot	×	×
	Average	0.1233	0.1448

[a]Italicized text indicates missing components