Space filling design for calibration sample selection

C. CHARLOTO, M. METZ, M. SERGENT, M. CLAEYS-BRUNO

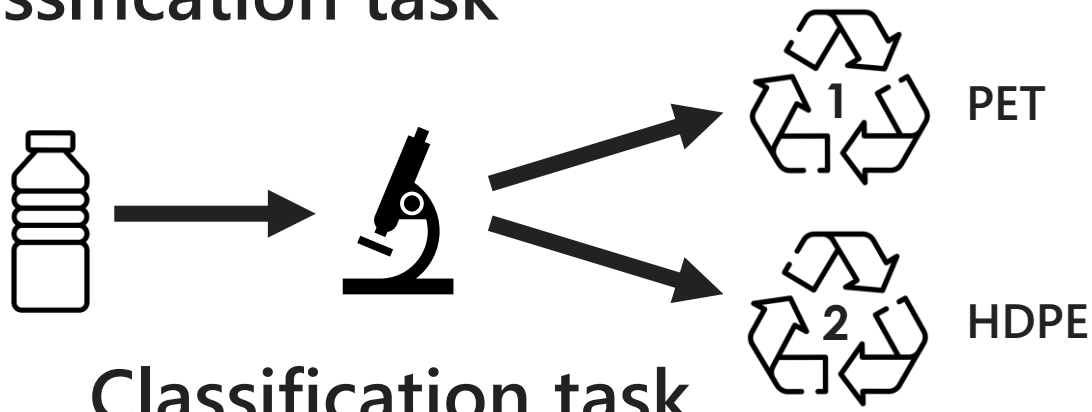
C.CHARLOTO@PELLENCST.COM

Table of contents

IR Spectroscopy – General [1]

- Common analytical technique → many applications including the recycling, agriculture, and oil industries
- Non-destructive, fast and sensitive technique
- Use to achieve regression and classification task

Regression task



Classification task

[1] Stuart, B. H. *Infrared Spectroscopy: Fundamentals and Applications*; John Wiley & Sons, 2004.

IR Spectroscopy – Cost of Training data

Training data [2] is the association of spectral variables and outcomes of interest (Labelling) :

- Spectral measurements (X) can be done in large scale
- Outcomes of interest (Y) are informations linked to the task
  - Need considerable amount of labels
  - High labelling cost
  - High risk of mislabelling
  - More and more frequent

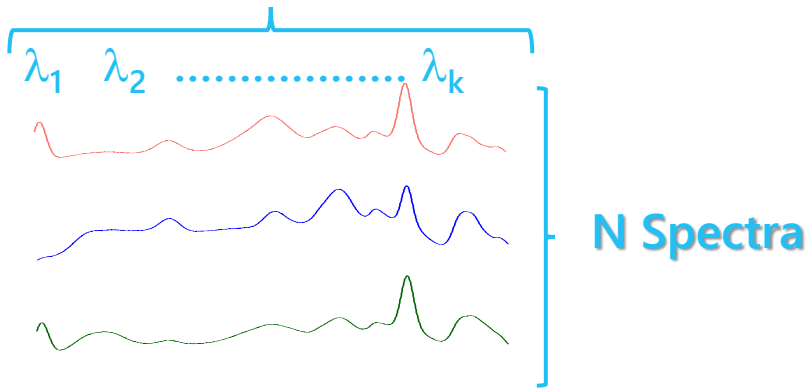
Can we reduce the size of training dataset ?
How to do it ?

[2] Nasteski, V. An Overview of the Supervised Machine Learning Methods. *HORIZONS.B* 2017, *4*, 51–62. https://doi.org/10.20544/HORIZONS.B.04.1.17.P05.

Key idea

Usual training process :

How to select spectra ?



$Y = f(X)$ from N

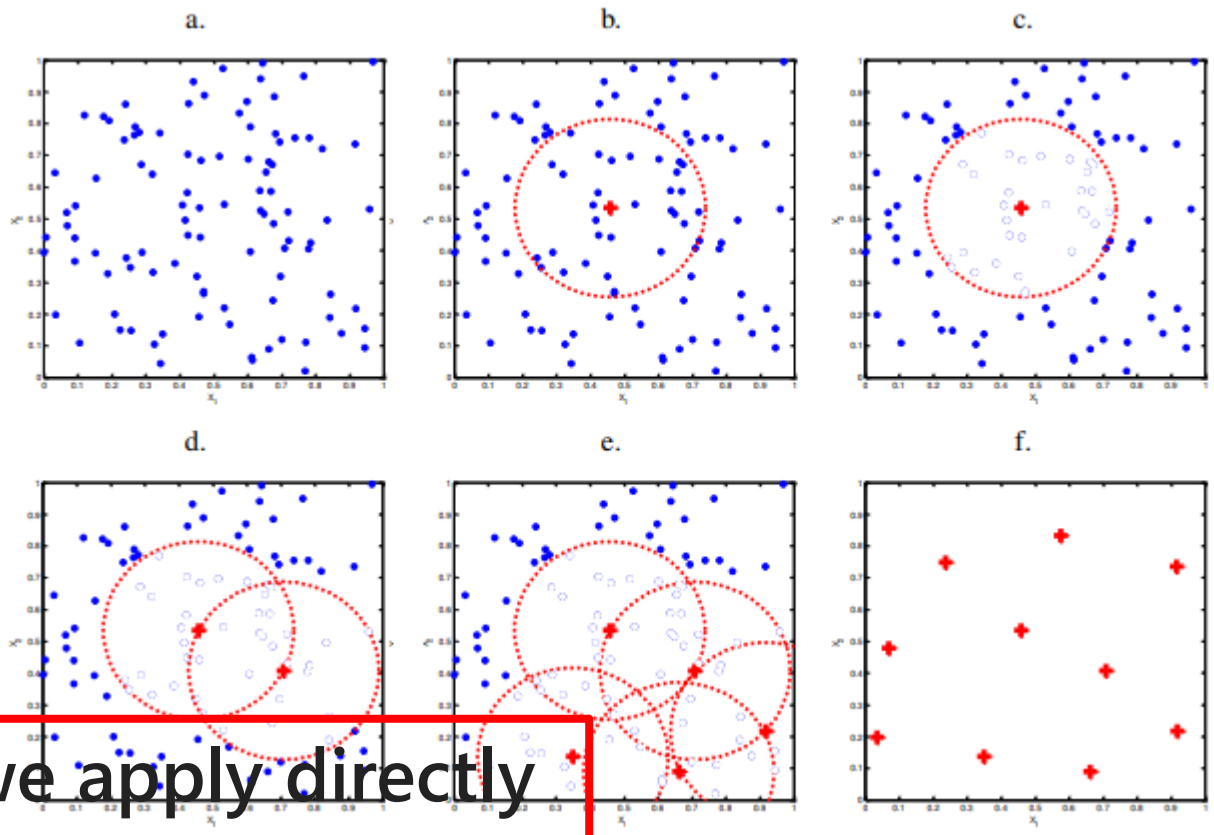
Spectral measurement → Labelling → Training + Validation

Step 1 → Step 2 → Step 3

Addition of an extra step : Samples Selection

Selection algorithms

- Random
- Kennard Stone [3]
- WSP [4]



Should we apply directly selection algorithms on spectra ?

[3] R.W. Kennard, L.A. Stone, Computer Aided Design of Experiments, Technometrics 11 (1969) 137–148. https://doi.org/10.1080/00401706.1969.10490666.
[4] Santiago, J.; Claeys-Bruno, M.; Sergent, M. Construction of Space-Filling Designs Using WSP Algorithm for High Dimensional Spaces. Chemometrics and Intelligent Laboratory Systems 2012, 113, 26–31. https://doi.org/10.1016/j.chemolab.2011.06.003.

Should we select directly on spectra [5] ?

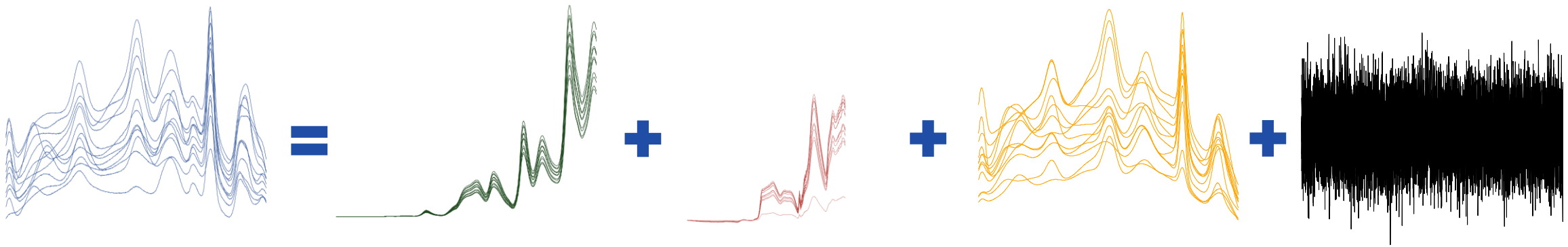
- What's the impact of pretreatment for selection ?
- What's the impact of dimensionality reduction for selection ?
- What's the impact of encoding variables for selection ?
- Which selection algorithm use ?

We investigated the influence of these parameters on multiple datasets

[5] V. Fonseca Diaz, B. De Ketelaere, B. Aernouts, W. Saeys, Cost-efficient unsupervised sample selection for multivariate calibration, Chemometrics and Intelligent Laboratory Systems 215 (2021) 104352. https://doi.org/10.1016/j.chemolab.2021.104352.

Data used I : simulation datasets [6]

- 500 Spectra ranging between 1100-2500 nm with 2 nm resolution.
- 400 as candidates for selection, 100 for validation and test purposes.
- Glucose concentration was chosen as the outcome of interest



Spectra = Alcohol spectra + Glucose spectra + D.S + Noise

[6] Metz, M.; Biancolillo, A.; Lesnoff, M.; Roger, J.-M. A Note on Spectral Data Simulation. *Chemometrics and Intelligent Laboratory Systems* 2020, *200*, 103979. https://doi.org/10.1016/j.chemolab.2020.103979.

Data used II : Milk dataset [7]

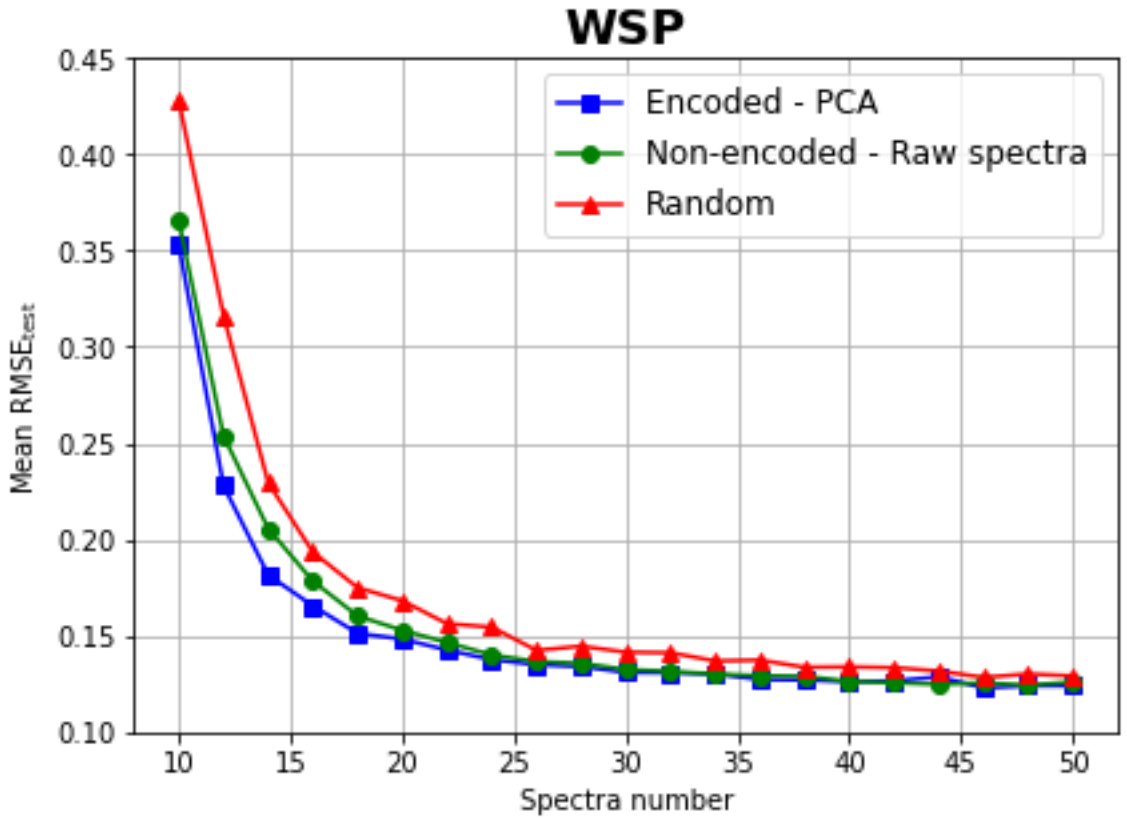
- Total number of spectra: 395
- Spectral range: 900–1700 nm
- Spectral resolution: 3 nm
- Data organization:
  - Candidate set → 316 spectra
  - Validation + Test set→ 79 spectra (39/40)
- Chemical constituent of interest: lactose

[7] J.A. Diaz-Olivares, I. Adriaens, E. Stevens, W. Saeys, B. Aernouts, Online milk composition analysis with an on-farm near-infrared sensor, Computers and Electronics in Agriculture 178 (2020) 105734. https://doi.org/10.1016/j.compag.2020.105734.

Evaluation strategy

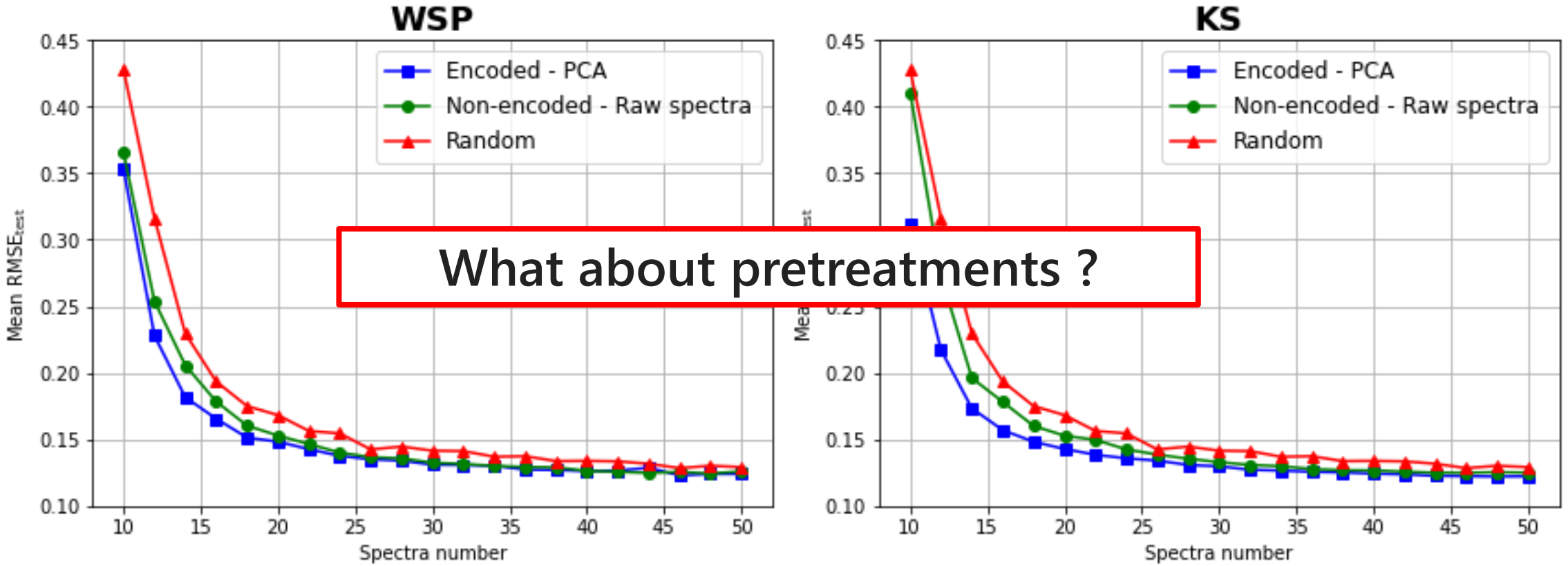
Datasets	**Simulated**			**Milk**		
Selection algorithms	**WSP**	**RAN**	**KS**	**WSP**	**RAN**	**KS**
Parameters tested	**Dimensionality reduction + Encoding**			**Dimensionality reduction + Encoding+ Pretreatment**		
Type of model	**PLS-R**					
Metric	**RMSE on test sets**					

Simulated datsets



Random selection performs worse
Best configuration : Encoding + PCA

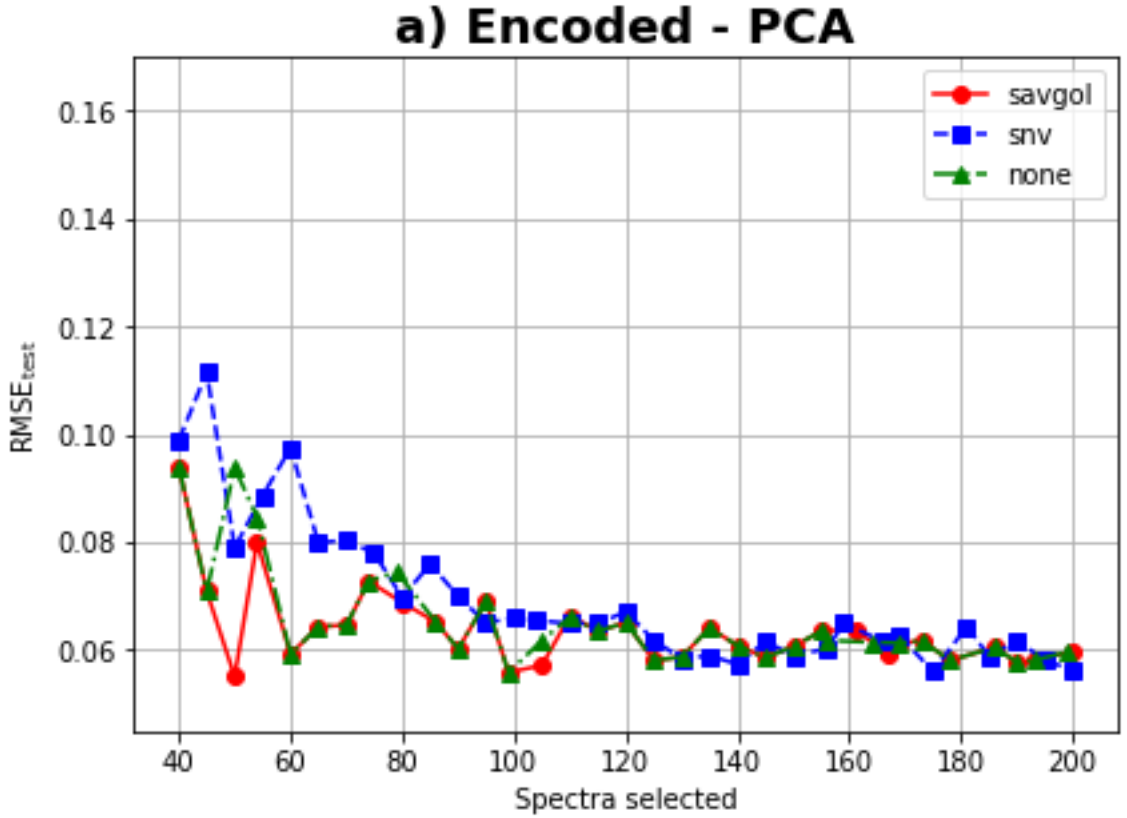
Simulated datsets



What about pretreatments ?

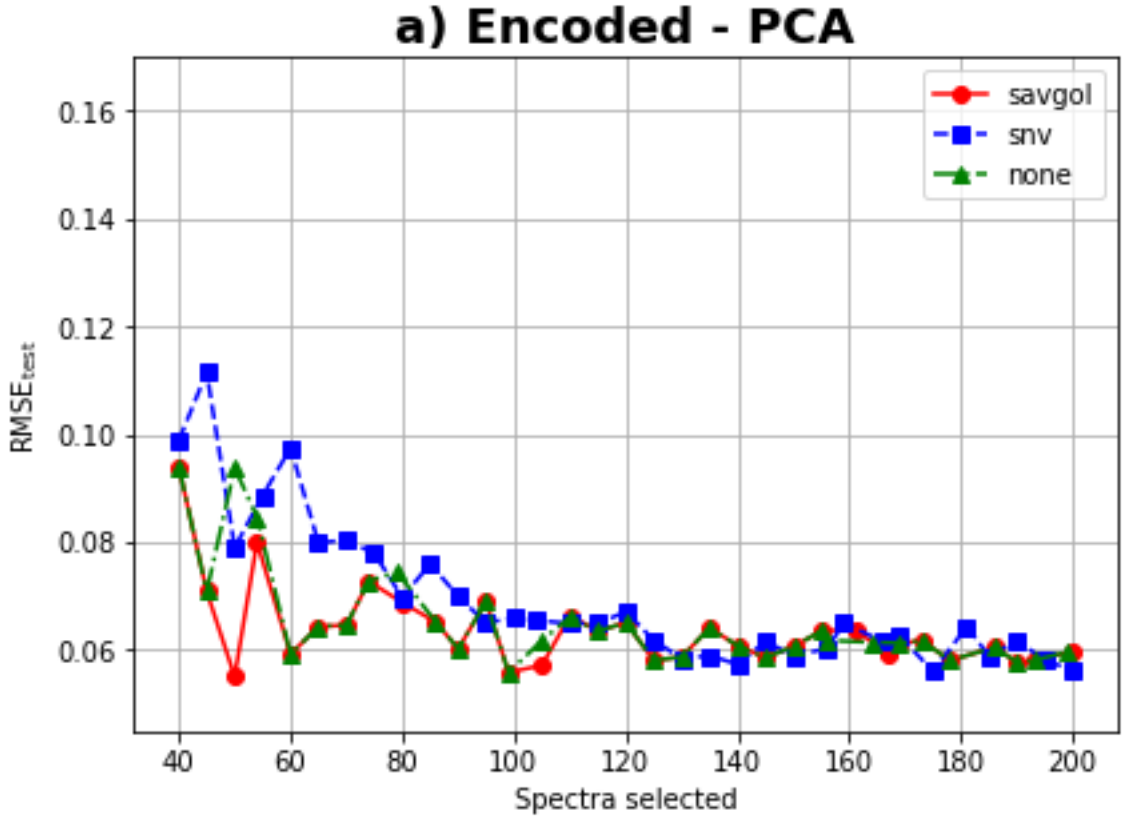
Both selection algorithms perform better than random
Encoding and PCA improve performances for both selection

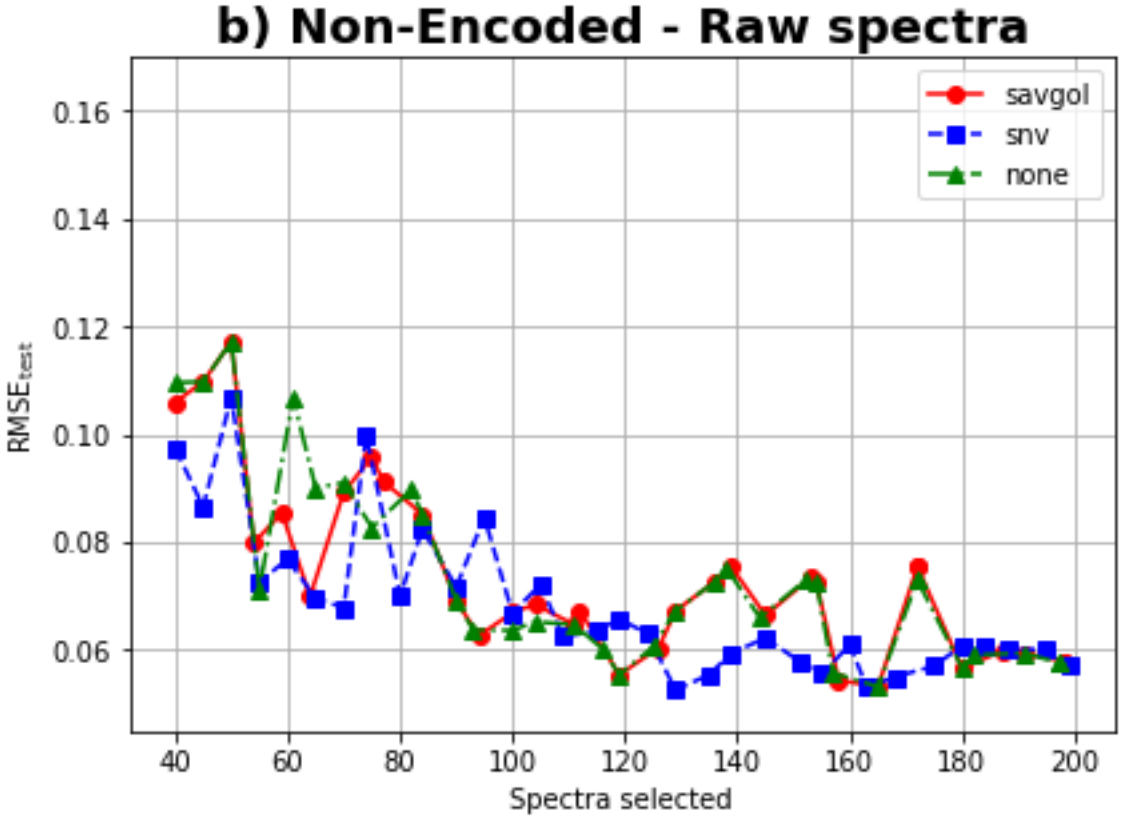
Milk dataset results : WSP



No pretreatments stand out
Performances stabilize after 110–115 spectra

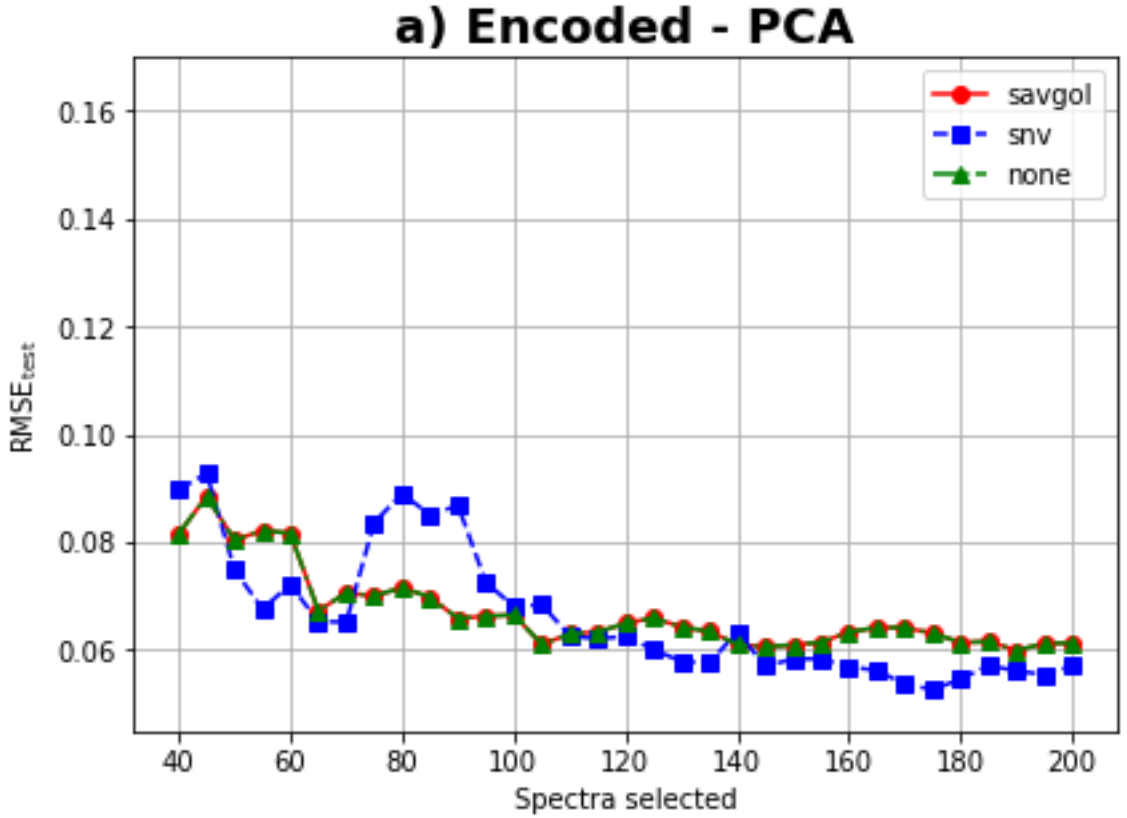
Milk dataset results : WSP





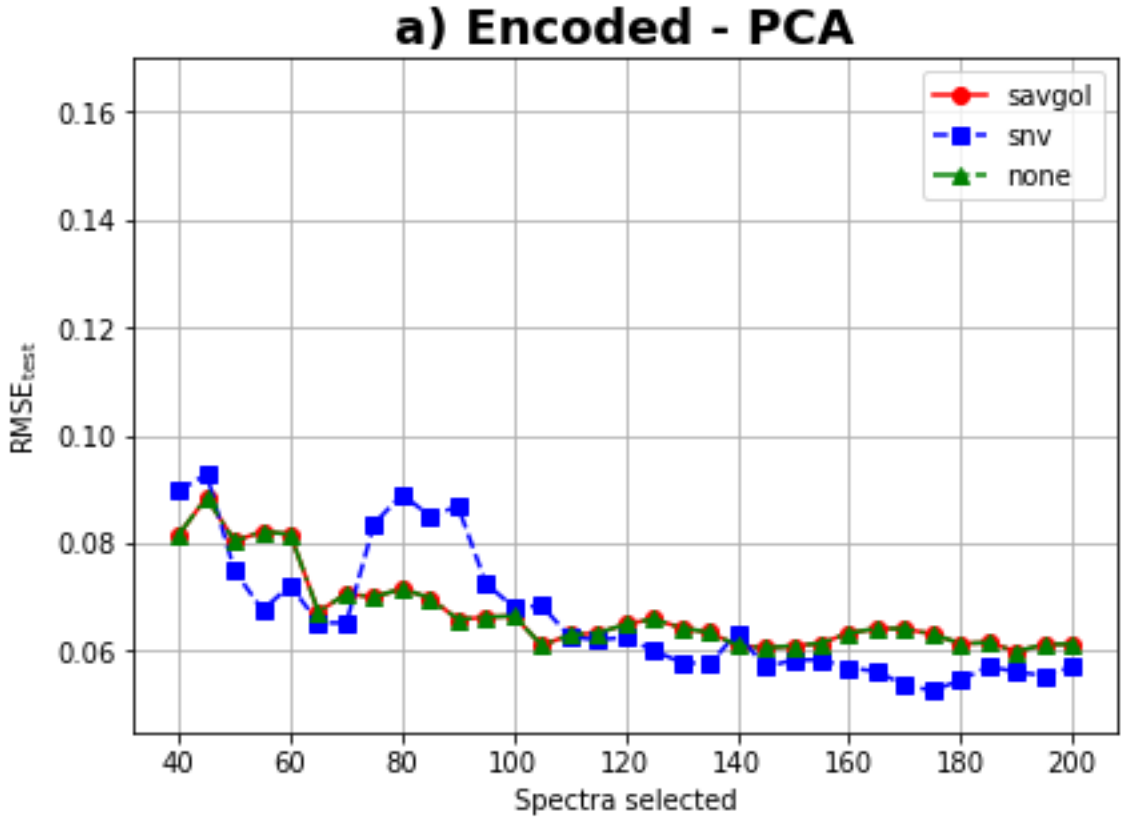
Both pretreatments impact the selection differently for each configuration
Performance stabilizes more quickly with Encoding and PCA

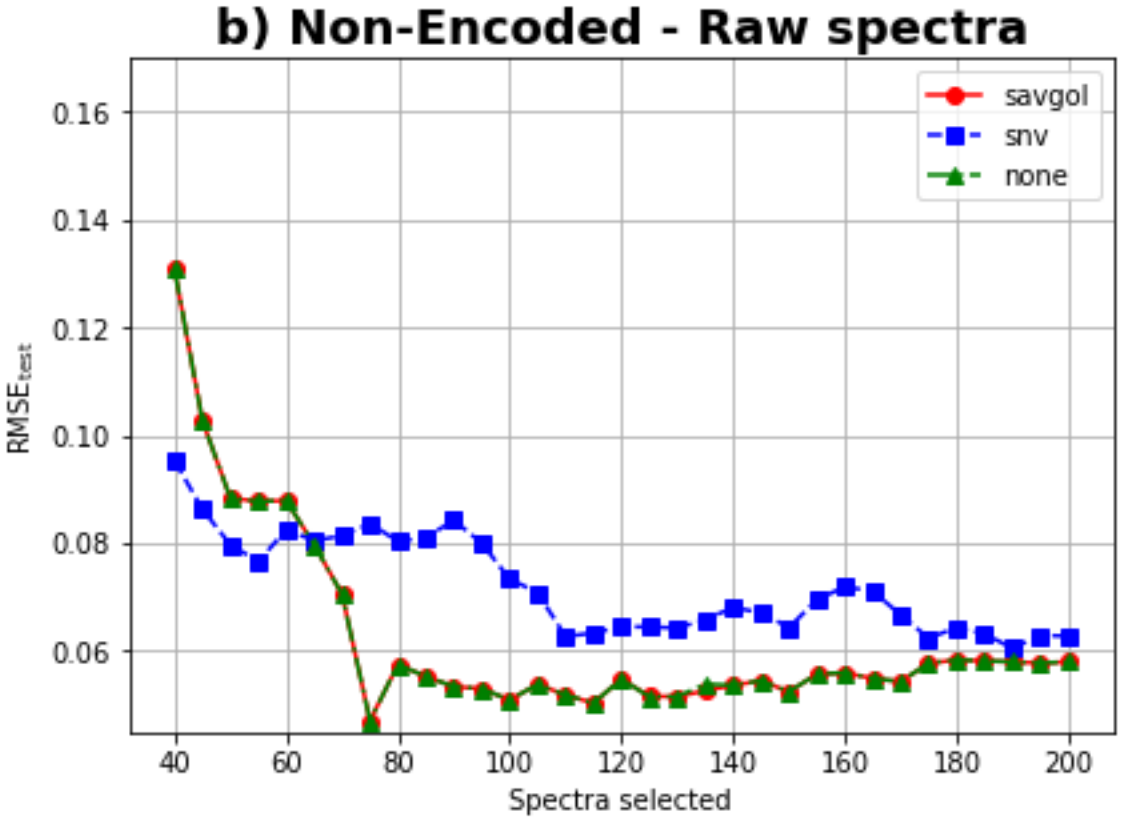
Milk dataset results : KS



No pretreatment stands out
Performance stabilizes after 180 spectra

Milk dataset results : KS





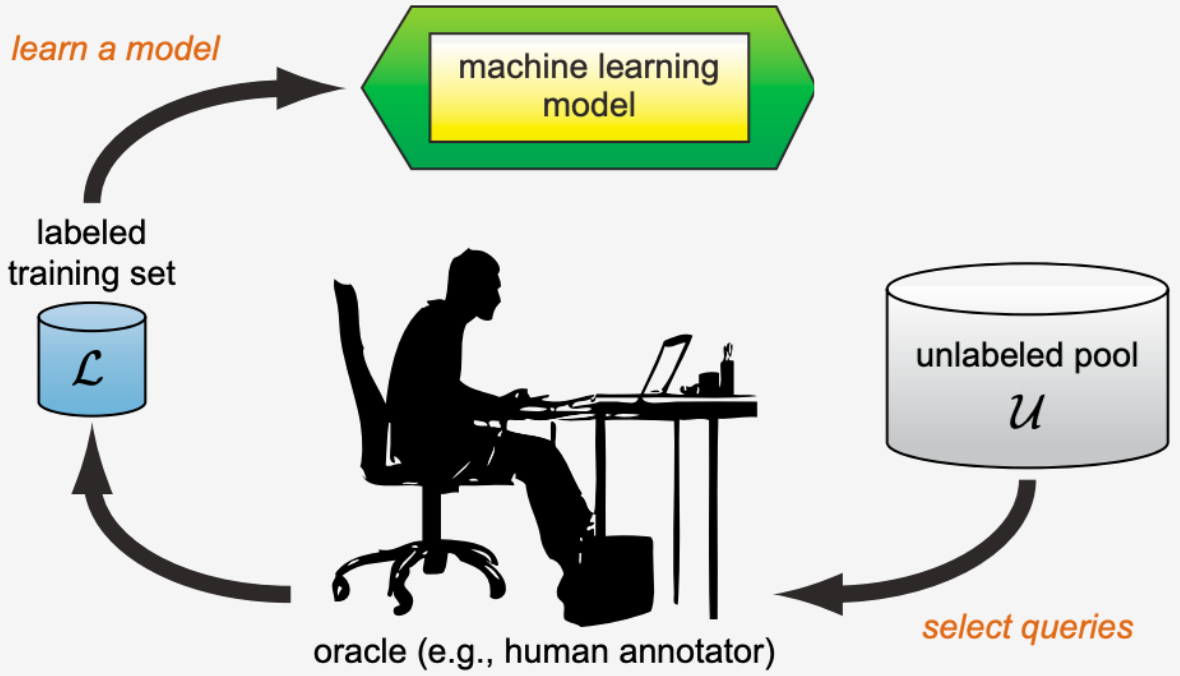
Both pretreatments impact the selection differently for each configuration
Performance peak for the non-encoded and raw spectra for this selection algorithm

Key messages

- Selections achieved better performance than not selecting
- Parameters such as encoding and dimensionality reduction achieved better performance for WSP on both studied cases and for KS on simulated data
- Pretreatments had varying effect depending on configuration, highlighting the need for expertise to choose it, particulary for KS

What's the next step ?

Perspectives : Active learning [8]



Spectral measurement — Samples Selection — Labelling + Training + Validation

Further works will be focused to evolve WSP approach to complete the loop and add complementary selection criteria to WSP

Step 1 → Step 1' → Step 2 and 3

[8] Settles, B. *Active Learning Literature Survey*; Technical Report; University of Wisconsin-Madison Department of Computer Sciences, 2009. https://minds.wisconsin.edu/handle/1793/60660 (accessed 2024-06-26).

Thanks for your attention !

C. CHARLOTO, M. METZ, M. SERGENT, M. CLAEYS-BRUNO

C.CHARLOTO@PELLENCST.COM