Foods 2024 Conference

The 5th International Electronic Conference on Foods

28–30 October 2024 | Online

MDPI

Data analysis of protein-flavor interactions by classification and deep-learning techniques

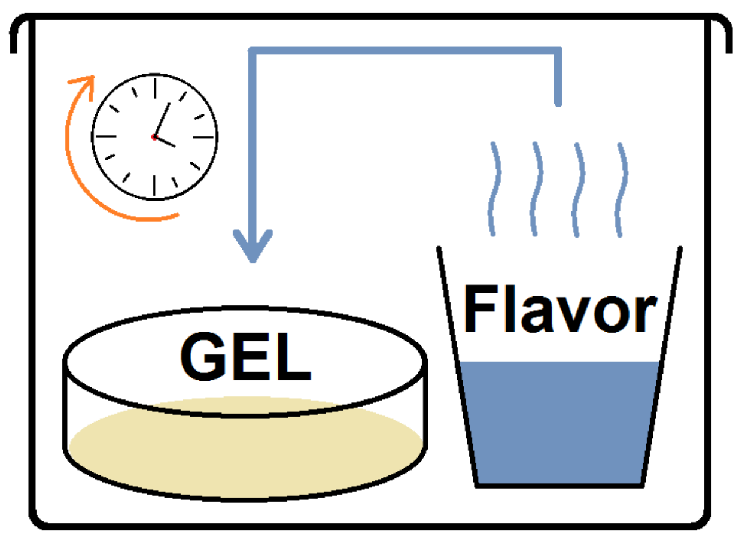
Nicolás Villagrán dos Santos (1)*, Lorena Pepa (2), Yamila Alen (3), Pilar Buera (2), Cristina dos Santos Ferreira (4)

(1) Universidad de Buenos Aires. Facultad de Ciencias Exactas y Naturales. Departamento de Matemática. Buenos Aires, Argentina. (2) CONICET - Universidad de Buenos Aires. Instituto de Tecnología de Alimentos y Procesos Químicos. Buenos Aires, Argentina. (3) CONICET - Universidad de Buenos Aires. Instituto de Cálculo. Buenos Aires, Argentina. (4) Universidad de Buenos Aires. Facultad de Ciencias Exactas y Naturales. Departamento de Química Orgánica. Buenos Aires, Argentina. ***: nicovillagran@gmail.com**

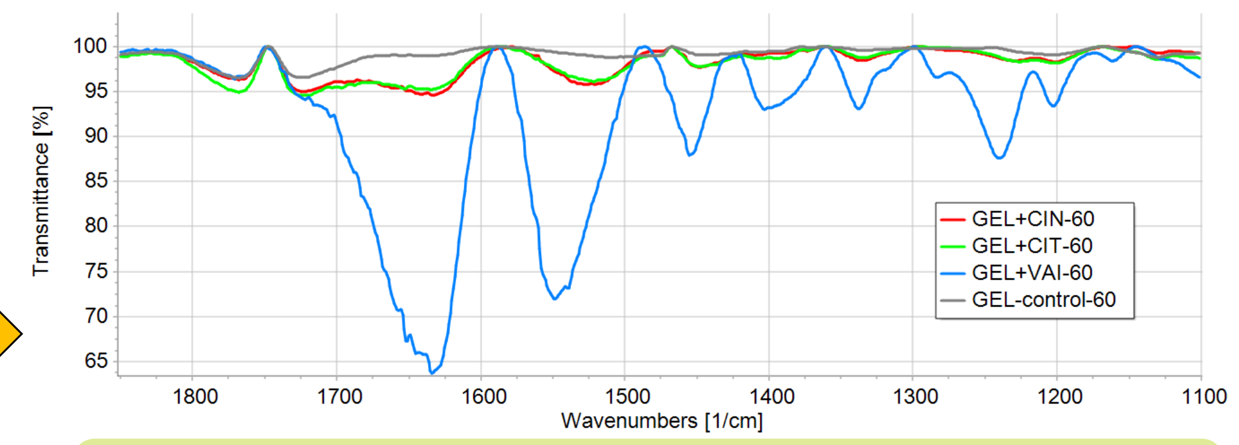
INTRODUCTION & AIM

Analysis of molecular interactions among food components is of key interest for novel food formulations and optimization of their shelf-life, but the complexity of food matrices often poses difficulties. In this work, model systems with gelatin (GEL) placed in indirect contact with different flavors (citral, CIT; cinnamaldehyde, CIN, and vanillin, VAI) and stored at room temperature during 60, 120 and 150 days were studied. The aim of this work was to explore molecular interactions among these food components using FTIR-ATR spectra of the systems, employing data analysis techniques focusing on specific spectral zones around 1850-1100 cm^{-1} (amide I, II, III regions). Data analysis with Python involved principal component analysis (PCA) to differentiate flavors and storage times, followed by classification models using random forest (RFC) and neural network (a multi-class perceptron classification, MPC) techniques. The models achieved high accuracy in flavor (81% RFC, 83% MPC) and storage time (88% RFC, 92% MPC) classification. Key features identified by PCA and RFC coincided, while the MPC showed superior accuracy in system classification. This research showcases innovative data analysis techniques for understanding complex protein-flavor interactions, offering insights valuable for advancing food formulation strategies.

METHOD



Storage of GEL+flavor or GEL alone (control) systems in airtight containers for 60, 120 and 150 days at 25°C.



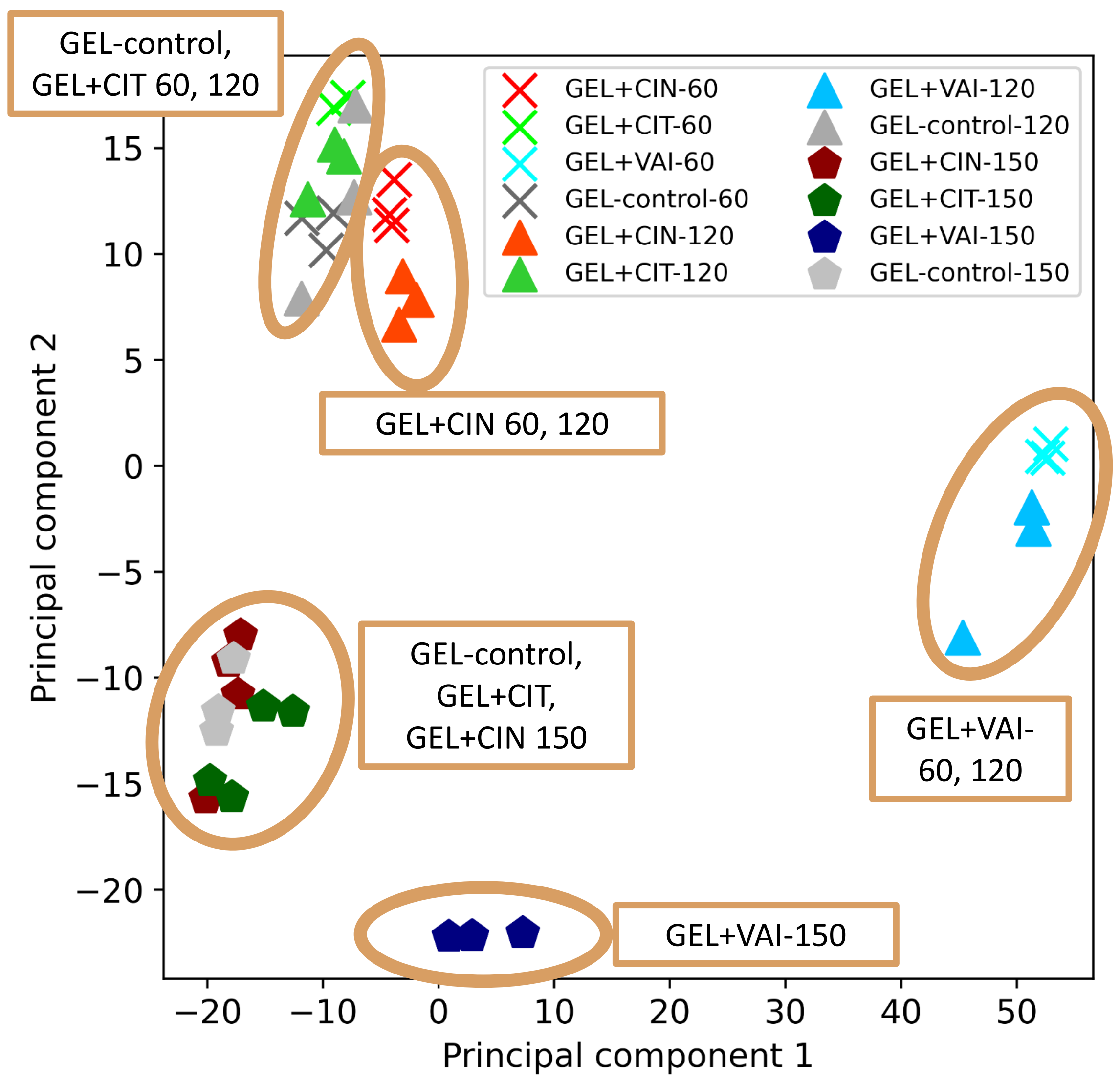
FTIR-ATR spectra

Principal component analysis + classification models (Python)

RESULTS & DISCUSSION

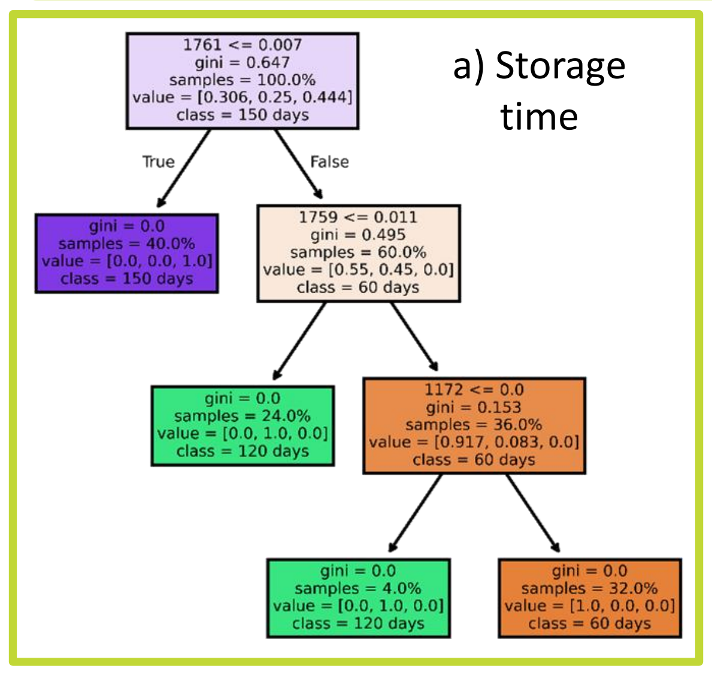
Data (37 FTIR-ATR spectra with 750 features) classified according to composition (Control, CIT, CIN, VAI) or storage time (60, 120, 150 days).

Principal component analysis

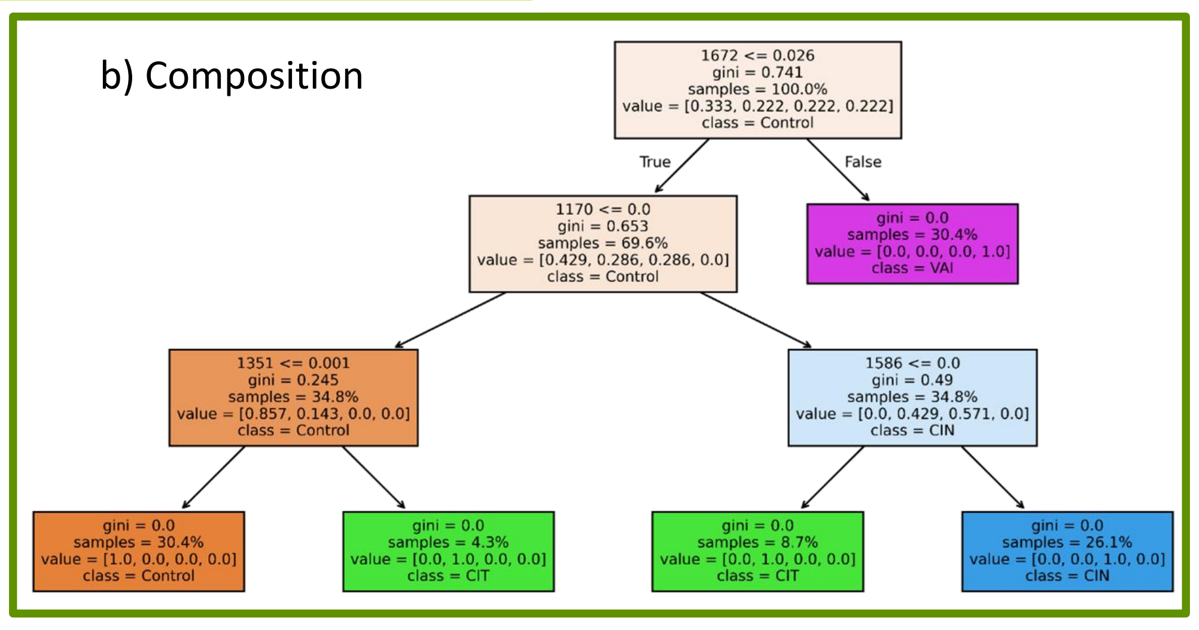


- First two principal components (PCs) explained >94% of variance.
- Data separated into five classes.
- Along PC1, GEL+VAI systems at all times were separated from all others.
- PC2 separated systems at 60 or 120 days from 150 days.
- Feature study in loading plots:
  - PC1 is correlated with changes in amide I and formation of Maillard reaction products, corresponding to protein-flavor interactions.
  - PC2 is associated to C=O stretching in amide I probably linked to structural changes during storage.
  - Key features are similar to those later reported by random forest classification.

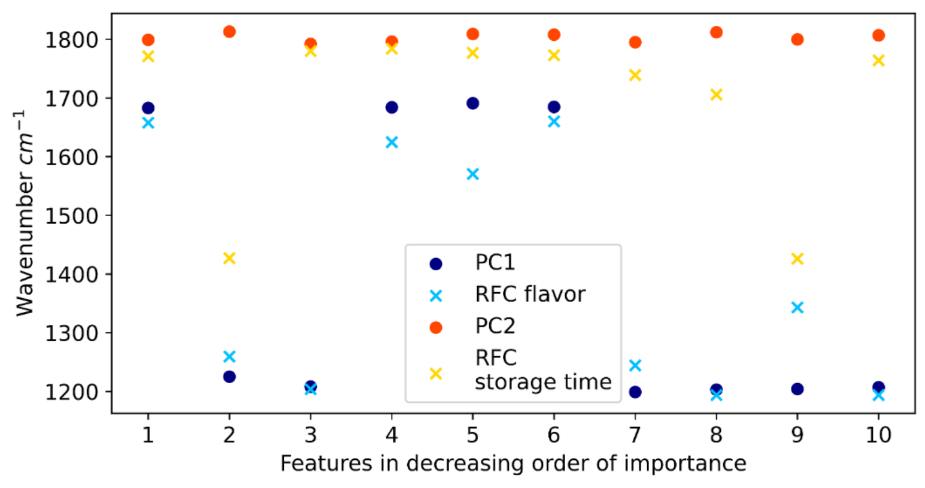
Random forest classification (RFC)



Classification of samples according to:
a) storage time (60, 120 or 150 days)
b) system composition (control, CIT, CIN, VAI)
+ leave-one-out cross-validation



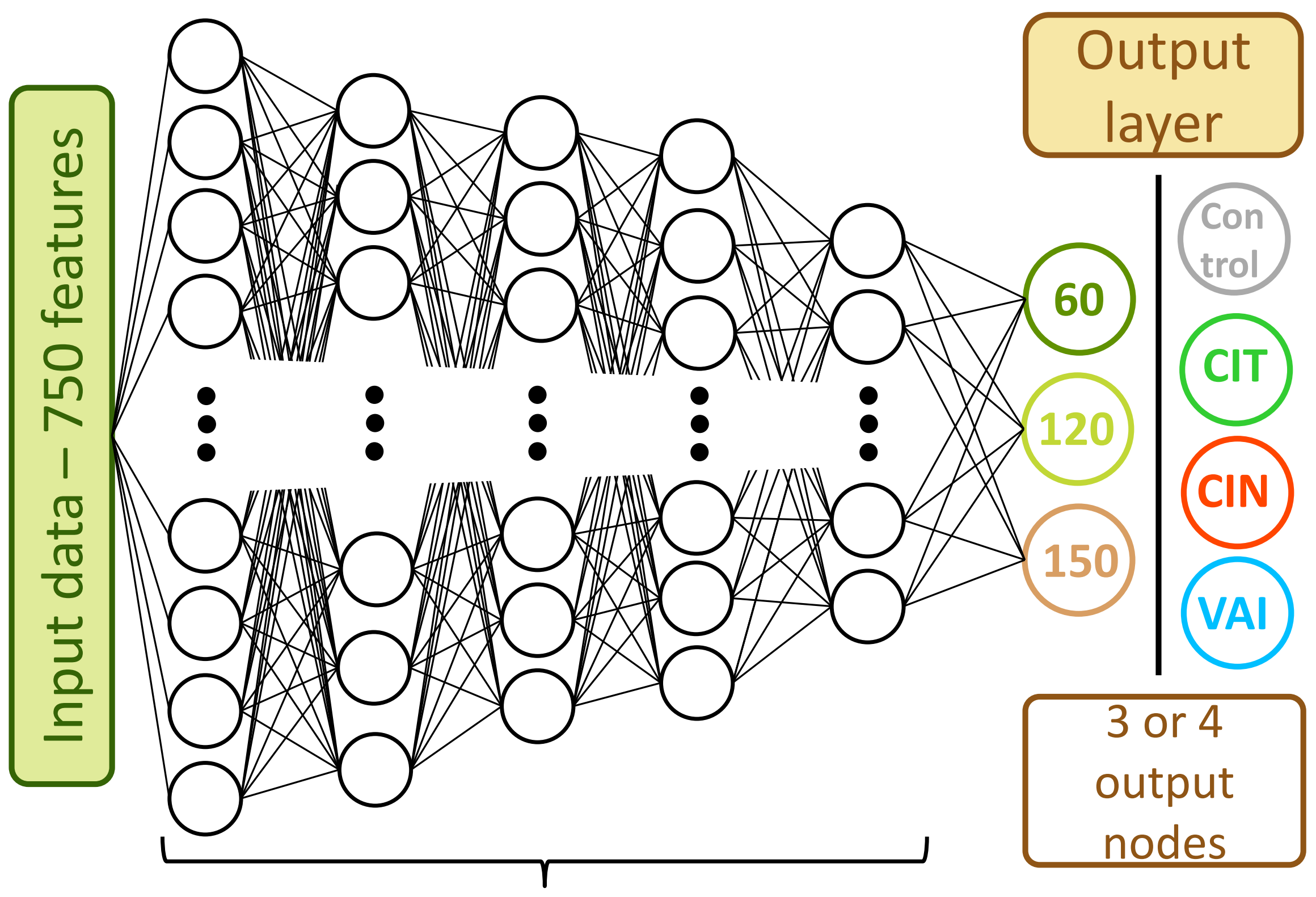
From RFC models => key features (wavenumbers).
Comparison to PCA key features (associated to PC1 and PC2):



- Storage time classification: 150 days samples were more easily separated than those at 120 or 60 days.
- System composition classification: GEL+VAI are the first to be classified, followed by GEL+CIN. GEL+CIT and control systems are more difficult to classify.
- This classification and feature analysis corroborated PCA.
- RFC was quick, efficient and accurate.

Multi-class perceptron classification (MPC)

A multi-layer feedforward neural network of fully connected neuron layers capable of classifying non-linearly separated data into output nodes corresponding to either storage time (60, 120 or 150) or composition (control, CIT, CIN, VAI) classes.



Hidden layers with 1024, 256, 32, 16 neurons

Model employed	Accuracy (%)	
	Classification according to…	
	Composition	Storage time
RFC	81 %	88 %
MPC	83 %	92 %

- MPC models were correctly fitted to the data.
- MPC models showed greater accuracy than RFC models for both storage time and composition classification.
- These neural networks are however more difficult and time-consuming to train than RFC models.
- A preselection of key features (given by PCA/RFC) shows promise to enhance MPC efficiency without greater accuracy loss.

CONCLUSIONS

Present results show that PCA allowed a detailed study of the key features associated to molecular interactions between proteins and flavors. A behavioral separation of the data was then validated by RFC and MPC, both showing a high accuracy. RFC models were quicker to train, while MPC were more accurate. This multidisciplinary approach employing data analysis techniques for the study of complex systems offers useful insights for novel food formulation strategies and a deeper understanding of food ingredients interactions.

We would like to thank the Universidad de Buenos Aires (grants UBACYT 20020220400331BA y 20020220300231BA) and the ANPCyT (PICT-2021 SERIEA-01909) for supporting this work.

https://sciforum.net/event/Foods2024