scientific reports

Check for updates

OPEN

On the use of TabPFN on mass spectrometry analysis of volatile organic compounds

Pablo M. Granitto[1,2]✉, Emanuela Betta[1], Iuliia Khomenko[1], Michele Pedrotti[1], Andrea Romano[1] & Franco Biasioli[1]✉

Volatile organic compounds (VOCs) are key markers in applications ranging from food quality assessment to medical diagnostics that can be profiled, for example, by gas chromatography–mass spectrometry (GC-MS) or by direct injection mass spectrometry (e.g. proton transfer reaction mass spectrometry). The common practice in both cases is to construct a tabular dataset from the raw measurements by performing peak extraction across samples and use statistical or machine learning methods to analyze it. However, modeling VOC profiles is particularly challenging due to high dimensionality, noise, and small sample sizes. In this study, we evaluate the Tabular Prior-data Fitted Network (TabPFN), a foundation model recently introduced for tabular data, across diverse VOC datasets. Without requiring task-specific training, TabPFN achieves state-of-the-art performance in both classification and regression tasks, outperforming classical machine learning methods for most datasets. We further explore new strategies to enhance TabPFN's performance, including ensembling and fine-tuning, finding that a plain ensemble seems to be the best option in this setting. Our results demonstrate that TabPFN is a highly effective modeling tool for VOC profiles obtained with different analytical approaches. It offers robust predictions even in the data-scarce, high-variability scenarios typical of real-world workflows.

Keywords TabPFN, Volatile Organic Compounds, PTR-ToF-MS

Volatile organic compounds (VOCs) are key chemical markers that play a crucial role in a wide range of applications, from monitoring food quality and authenticity in the agro-food industry[1] to serving as non-invasive biomarkers for the detection of diseases and physiological states[2]. Gas chromatography mass spectrometry (GC-MS) is the reference analytical technique for the qualitative and quantitative VOCs analysis. It generally requires sample preparation for extraction and preconcentration steps, and involves long analysis times due to the chromatographic separation. On the other hand, the sensitivity and speed of direct injection mass spectrometry (DIMS) approaches, such as proton transfer reaction time-of-flight mass spectrometry (PTR-ToF-MS), have been proven effective for broad screening and real-time monitoring of VOCs, though this may come at the cost of analytical specificity[3].

In both approaches, the raw data produced by the instrument can be numerically processed to extract quantitative (i.e., relative or absolute amounts) and qualitative (i.e., compound identity) information. In DIMS, qualitative and quantitative information is typically conveyed by measured mass and peak area, respectively[4]. GC-MS also provides retention times as an additional set of coordinates obtained by chromatographic separation. The resulting tabular data are commonly organized into a two-dimensional matrix, with each row representing a sample and each column corresponding to a specific compound, effectively defining a VOC profile for each observation. These profiles can then be used to predict categorical outcomes (e.g., product origin, quality or disease presence) or continuous variables (e.g., concentrations, sensory attributes). Figure 1 summarizes this process. However, the high dimensionality, complexity, and variability of VOC datasets present significant challenges for conventional modeling approaches.

Machine learning (ML) is a powerful tool to address these challenges, offering the ability to model complex, often non-linear relationships with scarce, high-dimensional data[5]. In the last decade, ML has been largely dominated by Deep Learning methods[6]. In spite of its success in almost all types of data (text, images, video, etc.),

[1]Research and Innovation Center, Fondazione Edmund Mach, Via E. Mach 1, San Michele all'Adige, Trento, Italy. [2]CIFASIS, CONICET-UNR, Ocampo y Esmeralda, Rosario, Argentina. ✉email: granitto@cifasis-conicet.gov.ar; franco.biasioli@fmach.it

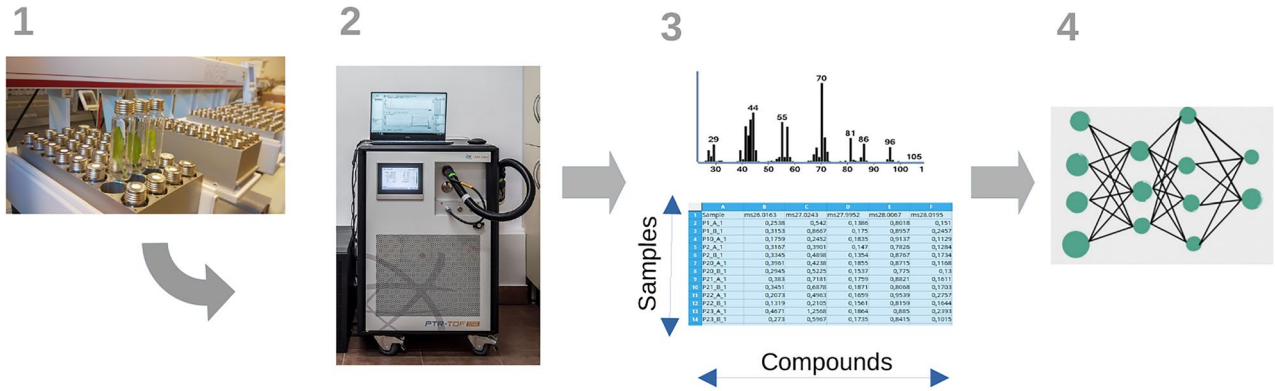

Fig. 1. Typical VOCs analysis process. After preparation (1), samples are analyzed with the instrument (2), using for example Direct Headspace Injection, producing, after data preprocessing and peak extraction, VOCs profiles (3) that are collected into a dataset, which is used (4) to fit a machine learning method.

Deep Learning has not been able to clearly outperform traditional methods in tabular datasets[7], the common format of VOC datasets after pre-processing.

Foundation models[8] are considered a paradigm shift in ML, consisting in large-scale models trained on broad data distributions that can be adapted to a wide variety of related tasks. They were originally developed and popularized in the field of natural language processing, using transformer-based architectures[9]. Models like GPT[10] and BERT[11] can learn very general representations that show strong performance in diverse problems even with minimal or no task-specific fine-tuning. Their strength lies in their re-usability: once trained, a foundation model can be applied across domains, reducing the need for task-specific data and computational resources. This approach is now rapidly extending beyond text, with models learning images[12], astronomical data[13] or tandem mass-spectrometry[14], for example.

An important feature of many foundation models is their ability to generalize beyond the data they were explicitly trained on through in-context learning[15], allowing the models to adapt to a new task or domain using only the information provided in the input prompt, without updating their internal parameters. This allows the model to leverage prior knowledge and adapt its behavior dynamically, making it appropriate to scenarios where annotated data are scarce or diverse, like VOC analysis.

TabPFN (Tabular Prior-data Fitted Network)[16,17] was recently introduced as a foundation model for tabular data. It relies on a prior-fitting procedure, in which the model is trained offline once on millions of synthetically generated tabular datasets. This unique training strategy allows the network to approximate Bayesian inference by exposing it to a broad distribution of data-generating mechanisms.

Using in-context learning, i.e., conditioning on small context sets of labeled examples, TabPFN can produce fast and accurate predictions on real-world tabular problems. In fact, Hollmann et al.[17] show that TabPFN performs competitively without requiring fine-tuning on several standard ML benchmarks.

In this study, we investigate the use of TabPFN for modeling VOC profiles obtained from both PTR-MS and GC-MS measurements, mostly from food products but also from other origins. These datasets consist of vectors representing peak coordinates and intensities, often under high-dimensional, low-sample conditions. This context is particularly challenging due to the typical wide structure of the data, where the number of peaks greatly exceeds the number of samples. Moreover, modeling is further complicated by instrumental and biological variability.

Using a variety of real-world VOC datasets, we evaluate TabPFN's performance in both classification and regression tasks, comparing it against traditional machine learning baselines. Foundation models can, in most cases, be improved by adapting them to a particular domain. We therefore explore and evaluate multiple strategies to enhance TabPFN's performance on VOC data, including two ensemble approaches and two fine-tuning methods.

Results

Comparison with other methods

First, we compared TabPFN with a selection of established methods in both classification and regression setups. Section "Comparison methods" describes all baseline methods, including their tuning procedures and experimental setup. The datasets used for the evaluation are introduced in Section "Datasets", with details summarized in Tables 1 through 4.

Classification

Figure 2 presents the results for classification problems. In the top panel, the datasets are ordered first by their origin –PTR-MS on food products (PT), PTR-MS on other data (PTO) and GC-MS on food products (GC)– and then by the sample-to-peak ratio. In first place, the panel shows the inherent difficulty of VOCs analysis, where error rates are typically high. In this context, the performance of TabPFN seems to be as good as reported by Hollmann et al.[17] on other datasets. We did not find clear differences in performance related to the sample-to-

Dataset	Samples	Batches	Peaks	Classes	S/P	Description
Tea	456	21	161	4	2.83	Evaluation of leaves of green and black tea. The classes correspond to the geographical origin of each batch. Experimental details in Yener et al.[26].
Gum2	267	27	167	2	1.60	Evaluation of samples of base material for chewing gum. The classes correspond to the presence or absence of two components in the base material, over a set of several possible combinations of components.
Gum3	267	27	167	2	1.60	Same as Gum2.
Mush 21	593	50	383	6	1.55	Evaluation of samples of diverse species of fungi. The classes correspond to the condition of cultivation. Experimental details in Telagathoti et al.[27].
Mush 13	54	19	125	6	0.43	Evaluation of mushroom samples of diverse species of the genus Armillaria. The classes correspond to the species. Experimental details are similar to Mush 21.
Fish	104	32	259	3	0.40	Evaluation of fish samples. The classes correspond to different cooking processes. Experimental details in Khomenko et al.[28].
Peppers	96	32	253	2	0.38	Evaluation of whole fresh peppers. The classes correspond to two different methods of conservation. Experimental details in Khomenko et al.[29].
Spinach	72	24	333	2	0.22	Same as Peppers for fresh spinach leaves.
Ham	54	18	427	3	0.13	Evaluation of samples of dry cured ham. The classes correspond to the geographical, controlled, and protected origin of each batch. Experimental details in del Pulgar et al.[30].
Lacto	102	20	798	2	0.13	Evaluation of samples of diverse strains of lactic acid bacteria during fermentation, taken at three consecutive times. The classes correspond to the temperature of the fermentation process. Experimental details in Rajendran et al.[31].
Coffee	36	12	563	6	0.06	Evaluation of coffee powder. The classes correspond to the geographical origin of each batch. Experimental details in Yener et al.[32].

Table 1. Details on PTR-ToF-MS agrifood datasets for classification tasks. S/P corresponds to sample-to-peak ratio.

peak ratio. Comparing PTR-ToF-MS with GC-MS datasets, TabPFN's performance on the last group seems to be similar to classical methods, but this finding requires further investigation, as it is based on only five cases.

To aid interpretation, each cell is color-coded according to the method's rank for the corresponding dataset. For example, in the Tea dataset, XGB achieves the lowest error, followed by PDA and TabPFN. In several cases TabPFN achieves the best performance, and in the remaining cases, it performs very closely to the best method. Bottom panels of Fig. 2 summarize this analysis. The bottom right panel shows that TabPFN is the top-performing method on half of the datasets, while the remaining cases are evenly distributed among the other methods. The bottom left panel reports the average rank of each method across all datasets. Again, TabPFN stands out with the best (i.e., lowest) average rank, suggesting that even when it is not the top performer, its results remain highly competitive.

Overall, the results suggest that TabPFN is a highly suitable method for modeling VOCs in classification tasks.

Regression

We conducted the same set of experiments for regression problems. In this case, only datasets corresponding to food VOC profiles measured with PTR-ToF-MS were available. The results are shown in the top panel of Fig. 3, using the same color coding and ranking-based analysis as in the classification setting.

TabPFN once again demonstrates strong performance, consistently ranking among the top methods across datasets. The bottom panels of Fig. 3 quantify the ranking results, showing an even greater advantage of TabPFN over the other three methods compared to the classification scenario. As in the previous case, TabPFN can be considered as a highly suitable method for modeling VOCs in regression settings.

Evaluation of methods to improve TabPFN performance

As mentioned before, a key advantage of TabPFN lies in its capacity to yield accurate results on most tabular problems without requiring explicit training. Our findings in the previous section corroborate this inherent capability. Nevertheless, strategies exist to further enhance the performance of any foundation model in specific tasks. In Section "TabPFN", we detail and discuss four such strategies particularly suited for TabPFN and VOC datasets. These include Post-Hoc Ensembles (Post-Hoc), which is the method proposed by Hollmann et al. (called AutoTabPFN by the authors)[17] involving a weighted (and optimal) combination of previously and independently trained methods; a straightforward plain ensemble of TabPFN models (Plain), where independent models are simply averaged with equal weights, and two distinct fine-tuning approaches. The fine-tuning strategies involve either tuning the model on the same dataset being evaluated (FT Data) or on a group of VOC datasets that excludes the specific target dataset (FT Group).

Figure 4, top panel, shows the corresponding results for classification problems. Two datasets (Fish and Lacto) were not included in this analysis as they give error zero for TabPFN and all candidate methods. Also, the Bees dataset was excluded because it is an order of magnitude larger in samples than the others, and the time needed to create an ensemble was out of the allocated time budget. As in previous cases, we use a color code to show relative rankings and present the qualitative results in the bottom panels. We also include Random Forest (RF) and (original) TabPFN results as references. If we analyze the ensemble strategies, it is clear that our plain ensemble outperforms the Post-Hoc ensemble proposed by the authors of TabPFN. This result is related to the high-variability/low-sample situation typical of VOCs analysis. It is well-known that a plain ensemble is more adapted to high-variability situations than an ensemble that learns weighted combinations from the

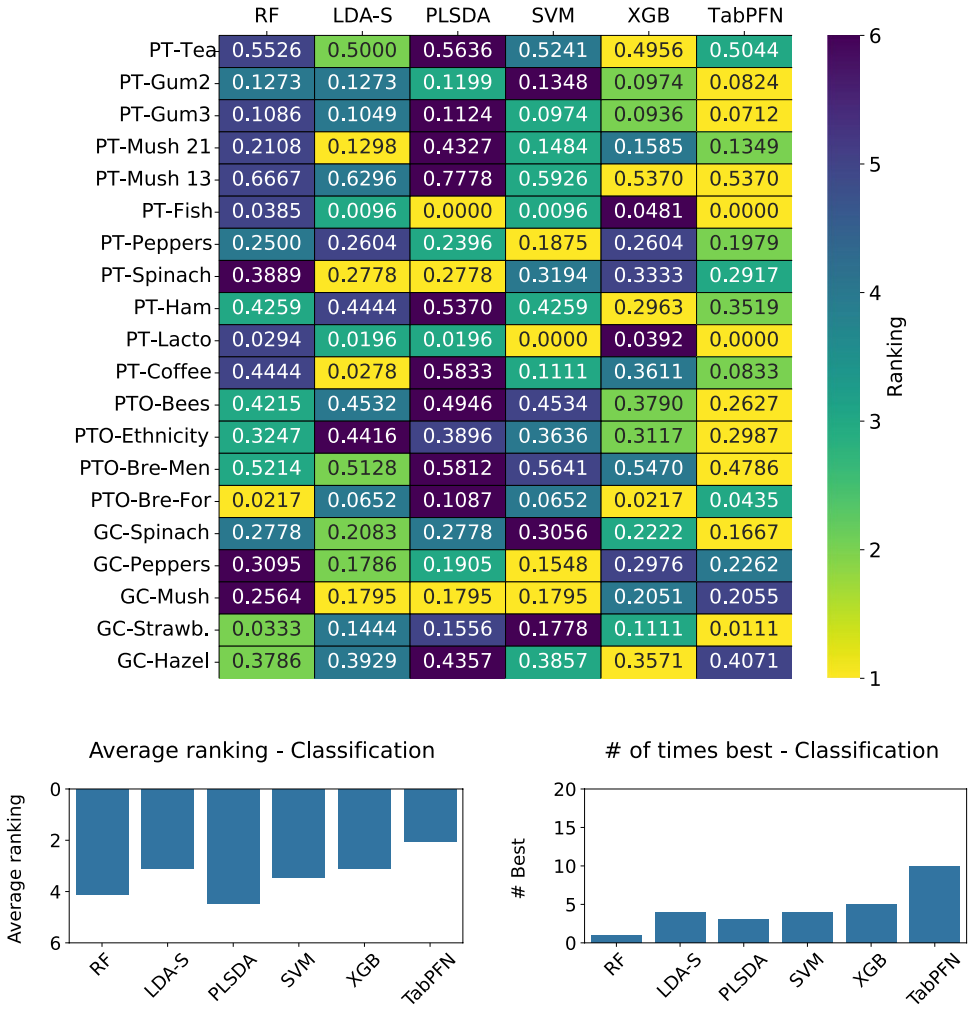

Fig. 2. Comparison of TabPFN with other methods in classification: Random Forest (RF), Linear Discriminant Analysis with shrinkage (LDA-S), Partial Least Squares coupled with LDA (PLSDA), Support Vector Machines (SVM), and XGBoost (XGB). Top panel: The table shows the mean classification error obtained using a leave-group-out evaluation for each method and dataset. The color of each cell represents the rank of the corresponding method for that dataset (lower is better). Bottom panels: Summary comparison of methods. Left: Average rank across all datasets (lower is better). Right: Number of datasets for which each method achieved the best performance (higher is better).

same training data, because the latter can easily produce overfitting when adjusting the weights to the training data. Comparing now the two strategies selected for fine-tuning, our results suggest that tuning directly on the same dataset that we want to predict is better than training on several other similar problems. We believe that tuning on other similar datasets leads TabPFN to bias the response to those problems, reducing the method's generalization capability, instead of adding relevant information, but this hypothesis requires further and dedicated investigation. Overall, the Plain ensemble strategy seems to be the best idea, but the average ranking results (bottom left panel of Fig. 4) are similar to the FT-data fine-tuning method and to directly use the original TabPFN.

We repeated the experiments on the regression problems. Figure 5 shows the corresponding results. Comparisons among strategies are similar to classification problems, plain ensemble seems to be clearly better than Post-Hoc, and FT-Data looks better than FT-Group. In this case, the Plain ensemble seems to be preferable over FT-data and the original TabPFN (which are similar).

Discussion

In this work we evaluated the use of TabPFN, a recently introduced foundation model for tabular data, on VOCs datasets obtained with PTR-ToF-MS and GC-MS, mostly on food related samples, but also from other sources.

We first compared TabPFN with several classical ML methods across both regression and classification tasks. Overall, TabPFN represents a promising approach for modeling VOC datasets. These datasets are tabular in nature, characterized by high levels of noise and low samples/features ratio that present significant challenges for most ML algorithms. While no single method can be expected to consistently outperform others across all

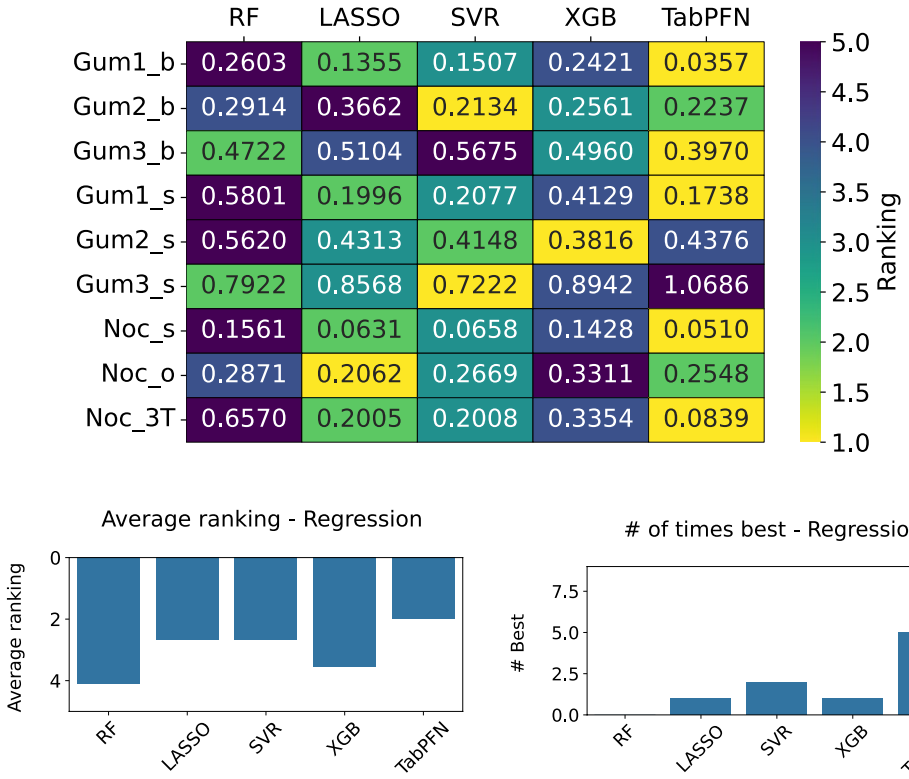

Fig. 3. Comparison of TabPFN with other methods in regression: Random Forest (RF), Least Angle Regression (LASSO), Support Vector Regression (SVR), and XGBoost (XGB). Top panel: The table shows the NMSE obtained using a leave-group-out evaluation for each method and dataset. The color of each cell represents the rank of the corresponding method for that dataset (lower is better). Bottom panels: Summary comparison of methods. Left: Average rank across all datasets (lower is better). Right: Number of datasets for which each method achieved the best performance (higher is better).

scenarios, TabPFN demonstrated robust performance, consistently achieving the lowest average rank across all evaluations.

We also discussed four strategies to improve the performance of TabPFN, including three original proposals (plain ensembles and two fine-tuning methods). Our experiments show that, in our particular setting, the use of plain ensembles of several TabPFN models seems to produce the best results, but using fine-tuning over the same data, or even using the original TabPFN gives almost similar results. As one of the advantages of TabPFN is that it can be used as an "off the shelf" method, the choice to use or not a more elaborate strategy with only a small increase in performance will depend on the needs of each individual problem.

The performance of ML methods, in particular in tabular data, cannot be directly extrapolated among domains. In this case, after our detailed evaluation, TabPFN shows the same good performance in this domain as in several others analyzed in the original work.

Our work exploited PTR-ToF-MS data as prototypical example of DIMS (Direct Injection Mass Spectrometry). Future research includes adding new data origins to the evaluation and considering other types of DIMS data, as for example the more complex data obtained by adding ion mobility spectrometry to the time of flight data. In addition, a foundation model capable of handling raw PTR-ToF-MS data (before all preprocessing) is under development.

Methods

TabPFN

As we established in the Introduction, foundation models are large-scale models trained on broad data distributions that can be adapted to a wide variety of related tasks. They were originally developed in the field of natural language processing, using transformer-based architectures[9]. They learn very general representations that show strong performance in diverse problems, even with minimal or no task-specific fine-tuning. Most foundation models have the ability to generalize beyond the data they were explicitly trained on through in-context learning.

TabPFN was recently introduced as the first foundation model for tabular data. TabPFN uses a prior-fitting procedure, where the model is trained offline once on millions of synthetically generated tabular datasets. After training, TabPFN uses in-context learning, conditioning on small context sets of labeled examples, in order to produce accurate predictions on test sets.

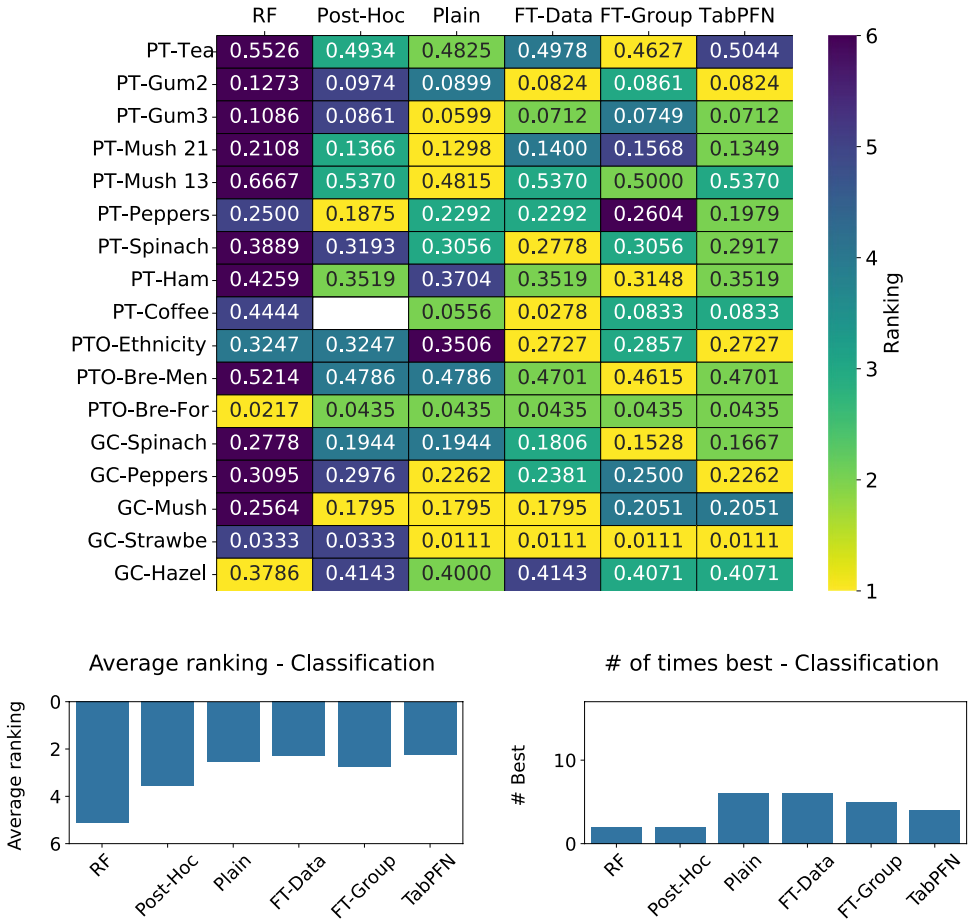

Fig. 4. Comparison of diverse strategies to improve TabPFN in classification. Details are similar to Fig. 2. The Post-Hoc ensemble cannot model the PT-Coffee dataset due to the low number of samples per class.

The current public version of TabPFN works on medium-size numerical datasets (up to 10000 datapoints and 500 features), both in multiclass (up to 10 classes) or regression problems. These capabilities cover most VOCs datasets, the only real limit being the maximum number of features, which can be exceeded by the number of peaks in some particular PTR-ToF-MS datasets.

In order to improve the performance of TabPFN, Hollmann et al.[17] evaluated the use of Post-Hoc ensembles over a portfolio of diverse versions of their original model. They use greedy ensemble selection to learn the optimal weights for combining the predictions of the selected models. In this procedure a validation subset is first extracted from the training data, and an iterative procedure of adding a model to the ensemble, finding the optimal weights on part of the training data and measuring the performance on the validation set is repeated for a fixed amount of time, after which the best ensemble is returned. As they report improved results with this ensemble, we also evaluate its performance on VOCs datasets. Figure 6 explains this method in its top panel. The learning set is used first to adjust the relative weight of each member of the ensemble. After that, the same learning set, together with the inputs of the test set, are used as input to the full ensemble model, which returns predictions for each data point in the test set.

A greedy strategy can lead to overfitting in high-variability/low-sample situations, as is our case. When ensembles use small training and validation sets, the optimal solution for the (small) validation set does not always lead to good performance on unseen data. A plain ensemble, where several models are simply averaged, is usually more appropriate for this setting, as averaging helps canceling noise. We made a few simple modifications to the code provided by Hollmann et al., creating a simple voting/averaging ensemble of TabPFN models. As shown in the second panel of Fig. 6, we use n models, each one with equal weight. In both ensemble methods the final number of models n is variable, based on a time limit for the construction of the ensemble. This typically produces a larger ensemble for smaller datasets.

The other well-known strategy to improve foundation models is fine-tuning. In this case, for each dataset the base model is slightly adjusted to the specific data under modeling.

We explored two different fine-tuning strategies. In the first approach, referred to as FT-Data, we fine tune the model using the training set with a very low learning rate and a limited number of epochs. This method is illustrated in the third panel of Fig. 6. After fine-tuning, as in all our experiments, both the training and test sets are used together as in-context input for TabPFN predictions. This setup involves using the training set twice, first during fine-tuning and again during inference, which may increase the risk of overfitting.

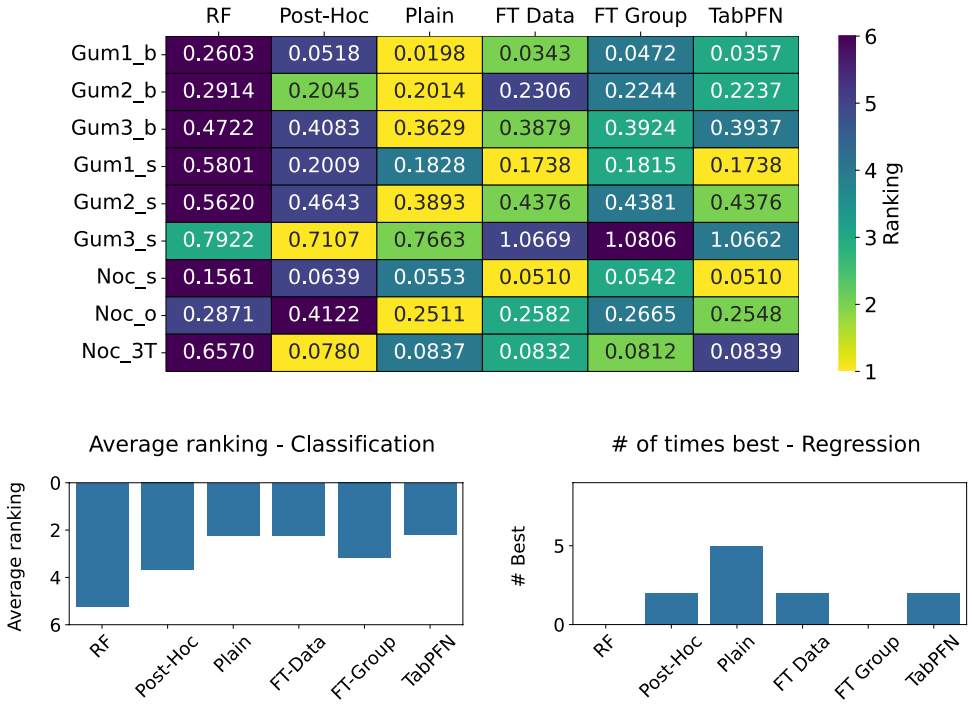

Fig. 5. Comparison of diverse strategies to improve TabPFN in regression. Details are similar to Fig. 3.

In order to avoid this potential hazard, we developed a last model, FT-Group, shown in the bottom panel of Fig. 6. In FT-Group we repeatedly adjusted the model over a random sequence of related datasets, not including the one we want to predict. Again, we use a very low learning rate and a few epochs each time. We train each model 30 times, selecting at each step one dataset at random (meaning that each dataset can be seen several times or not at all). After that, the tuned model uses the (at this point unseen) training data and test inputs to predict the corresponding outputs.

Comparison methods

We selected some classical tabular ML methods to produce a comprehensive and unbiased evaluation of TabPFN.

For classification datasets we chose five very diverse methods. RF[18] is a popular ML base method for any classification or regression problem, which can be used as an "off-the-shelf" method. It is particularly efficient for noisy problems and we consider it as the base comparison method for our VOCs datasets. We also use Linear Discriminant Analysis with shrinkage (LDA-S)[19], a regularized variant of the well-known statistical method. Partial Least Squares coupled with LDA (PLSDA)[20] is one of the classical methods for mass spectrometry. We also use Support Vector Machines (SVM)[21], an ML method with a strong mathematical basis that shows excellent results when properly tuned. Last, we use XGBoost, a gradient boosting algorithm[22], which builds an ensemble sequentially, focusing on hard instances, but also using random sampling of both samples and features. It is considered as one of the most efficient methods for tabular data[7].

For regression datasets we use the corresponding variants of RF, SVM and XGBoost, and compare also with the LASSO variant of the least angle regression method[23], which is specifically designed for high-dimensional problems.

In all cases, we used the same optimization strategy, applying CV with a grid search, leaving the rest of the setup with the default values of the Scikit-Learn implementations[24]. For RF we used a fixed setup with 1000 trees for both regression and classification. For LDA-S we used the automatic selection of the shrinkage parameters and for PLSDA we tuned the number of PLS components. In classification with SVM we optimized the constant C with a linear kernel, while in regression we also used a linear kernel but optimized both C and the size of the tube, ϵ. For this last parameter we made the selection over a set of fractions of the standard deviation of the target value of each dataset. For LASSO we optimized α, the regularization parameter. Last, for XGBoost we tuned the maximum depth of the trees (*max_depth*), the proportion of data randomly selected to train each tree (*subsample*) and the proportion of features chosen for that task (*colsample_bytree*). We keep the number of trees that form the ensemble fixed at a large number.

For classification problems, we measured the number of samples with an incorrect label, i.e., the classification error. For regression problems, we measured the Normalized Mean Square Error (NMSE), defined as the MSE divided by the variance of the target variable. Statistical significance tests were not reported, as post-hoc corrections for multiple comparisons resulted in non-significant outcomes in nearly all cases. This outcome is expected given the relatively small number of datasets (20 for classification and 9 for regression) combined with several methods being compared, which limits the statistical power of pairwise tests.

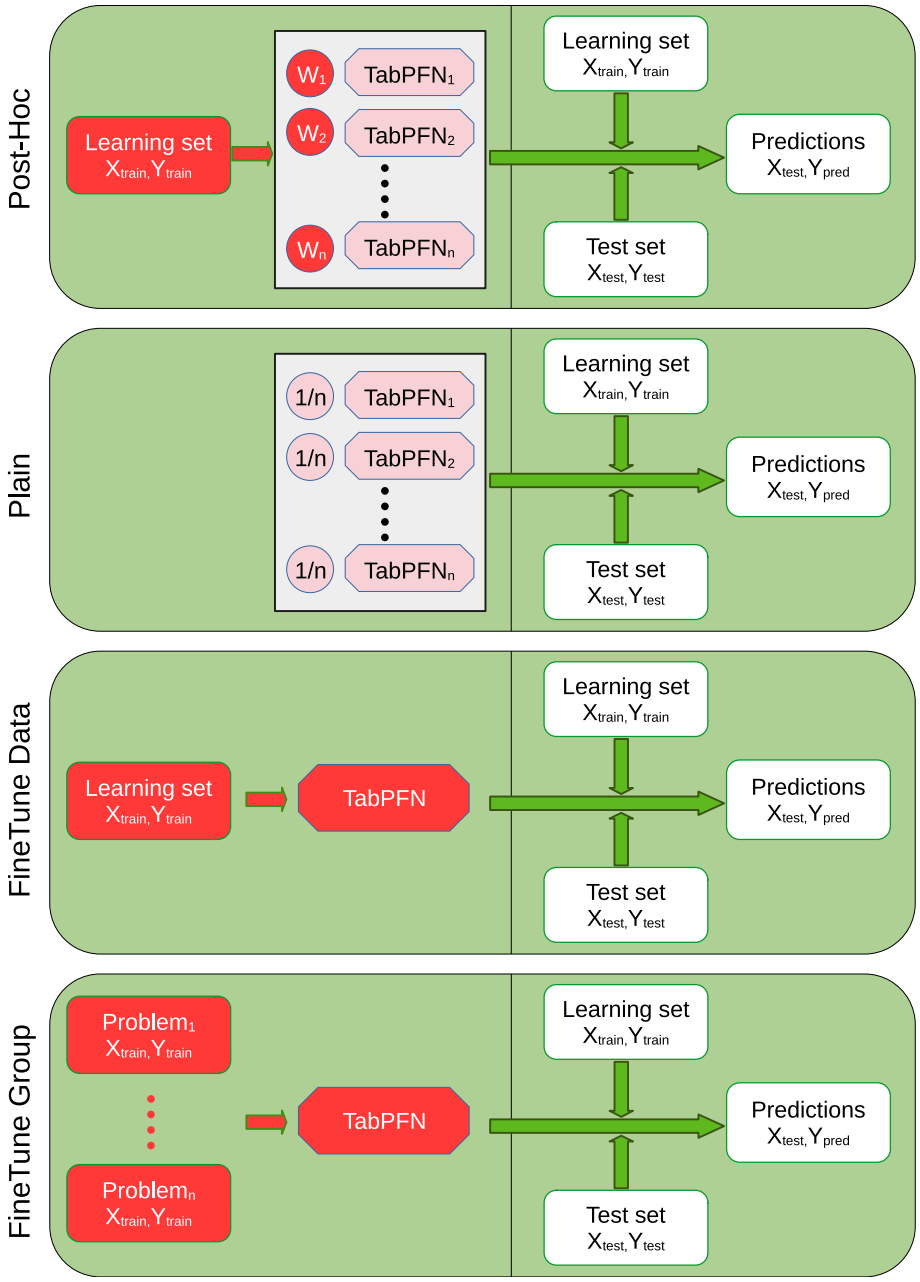

Fig. 6. Explanation of the set of methods considered in this work to improve the performance of plain TabPFN. Red color highlights the elements (data or parts of the models) used at the learning stage.

Datasets

We compare all the methods using diverse datasets that mostly comprise food-related product evaluation with PTR-ToF-MS and GC-MS, but also some data from other origins (only with PTR-ToF-MS). We discuss here the characteristics of our datasets that are more relevant to the modeling process.

In general, tabular data were extracted from PTR-ToF-MS raw data following the procedure described by Cappellin et al.[4], which includes m/z calibration and various pre-processing steps for noise reduction and baseline removal before peak extraction to obtain concentrations for every sample.

A similar procedure was applied to GC-MS data[25]. Data were acquired and processed with Agilent MassHunter. All target peaks were identified on the basis of the corresponding MS spectra and linear retention index. Peak areas were calculated on the response of a target ion. Up to two qualifier ions were monitored in order to increase the reliability of the peak area extraction.

Peak concentrations for PTR-ToF-MS data were normalized to area 1. No other preprocessing steps were taken. All extracted peaks were considered. There were no missing values on the datasets used in this work.

Dataset	Samples	Batches	Peaks	Classes	S/P	Description
Bees	4420	4	229	2	19.3	Measurements on a small controlled space where bees were freely allowed to enter or leave. The classes correspond to the number of bees present at the time. Work in preparation.
Ethnicity	77	29	44	2	1.75	Evaluation of flavour release of chewing gum by real-time, in-nose monitoring of VOCs. The classes correspond to the ethnicity (Chinese or European) of the subject. Experimental details in Pedrotti et al.[33]
Mentine	117	29	356	4	0.33	Evaluation of flavour release of chewing gum by real-time, in-nose monitoring of VOCs. The classes correspond to different compositions of the samples. Experimental details as in Pedrotti et al.[33], work in preparation.
Forest	46	23	384	2	0.12	Breath sample analysis of VOCs before and after a 20-minute walk in a forest. The classes correspond to before/after. Experimental details as in Pedrotti et al.[33], work in preparation.

Table 2. Details on PTR-ToF-MS datasets for classification tasks, from other sources.

Dataset	Samples	Batches	Peaks	Classes	S/P	Description
Spinach	72	12	54	2	1.33	Evaluation of the same fresh spinach leaves as in PTR-ToF-MS datasets. The classes correspond to two different methods of conservation. Experimental details in Khomenko et al.[29]
Peppers	84	14	67	2	1.25	Same as the Spinach dataset for fresh peppers.
Mush	39	5	46	2	0.85	Same as the Spinach dataset for mushrooms.
Strawb.	90	15	111	2	0.81	Evaluation of fresh strawberries under two different methods of conservation. More experimental details in Farneti et al.[34]
Hazel	140	4	183	3	0.77	Evaluation of hazelnut paste samples, obtained by processing raw kernels. The classes correspond to the place of origin of each sample. The batches correspond to production batches. More experimental details in Mazzucotelli et al.[35]

Table 3. Details on GC-MS datasets for classification tasks, all of agrifood origin.

Dataset	Samples	Batches	Peaks	S/P	Description
Gum1_s	267	27	167	1.60	Evaluation of samples of base material for chewing gum production. The predicted value corresponds to the percentage concentration of a component in the base material, over a set of several possible combinations of components. Each batch corresponds to technical replicates of a single sample
Gum2_s	267	27	167	1.60	Same as Gum1_s for a different component
Gum3_s	267	27	167	1.60	Same as Gum1_s for a different component
Gum1_b	267	3	167	1.60	Same as Gum1_s, but the batches correspond to production batches
Gum2_b	267	3	167	1.60	Same as Gum1_b for a different component
Gum3_b	267	3	167	1.60	Same as Gum1_b for a different component
Noc_S	72	24	380	0.19	Evaluation of hazelnut paste samples, obtained by processing raw kernels (Corylus avellana L.) from different geographical origins and different years. Samples were roasted using different times and temperatures. The predicted value corresponds to the roasting time. The batches correspond to technical replicates of a single sample. More experimental details in Mazzucotelli et al.[35]
Noc_O	72	3	380	0.19	Same as Noc_S but the batches correspond to the origin of the hazelnuts.
Noc_3T	60	20	383	0.16	Same as Noc_S, but the predicted value corresponds to the temperature of the oven during the roasting.

Table 4. Details on datasets for regression tasks.

The final data are represented as a table containing the estimated concentration for each peak and each sample, with sample identification in the first column and the accurate m/z values of each mass peak (or compound identification for GC-MS) as column headings.

Tables 1, 2, 3 and 4 summarize the key characteristics of each dataset, including references to the original work. In all tables, datasets are ordered by the sample-to-peak ratio, presented in the previous to last column. Our selection encompasses a wide range of real-world scenarios in terms of the number of classes and sample-to-peak ratios.

All four tables also include the number of batches considered for each dataset. The concept of "batch" is essential to modeling real-world samples, as it represents a group of samples that are independent of other groups but not necessarily within themselves. Even when a batch contains diverse samples rather than simple technical replications, internal dependencies typically exist. A batch may correspond, for instance, to samples from a specific factory, a particular place of origin, or a specific production year or season.

The choice of batch composition can significantly influence the nature of the modeling problem. The same dataset may contain samples from different geographical origins and multiple years of production, leading to diverse predictive challenges. For example, predicting a product's property for a new geographical origin versus predicting it for a new production year can pose fundamentally different problems. In regression tasks, we accounted for two of these scenarios in most datasets.

Based on these considerations, we employed a leave-group-out approach for all evaluations, using a CV strategy in which each group corresponds to a batch. In this setup, there is no possibility of repeats, so a single pass is used. In addition, randomization seeds and prevention of batch leakage are not required. Error measures were averaged over samples, not over groups, giving each sample the same weight in the result.

For breath analysis datasets involving human subjects, approvals from ethical committees are detailed in the original publication (see Table 2) where the authors state that: "The ethical committee of Wageningen University provided an official opinion that the EC approval is not needed for this study". All experiments were performed in accordance with relevant regulations and written informed consent was obtained from all participants.

Data availibility

We used open-source Python implementations of each method, including an extended use of the Scikit-Learn library[24]. All our code is available at https://github.com/CIFASIS/TabPFN-VOCS. The datasets analyzed during the current study are available in the same repository, in their final version, to help with the reproducibility of our results.

Received: 6 August 2025; Accepted: 14 November 2025
Published online: 02 December 2025

References

1. Biasioli, F., Gasperi, F., Yeretzian, C. & Märk, T. D. Ptr-ms monitoring of vocs and bvocs in food science and technology. *TrAC Trends Anal. Chem.* **30**, 968–977. https://doi.org/10.1016/j.trac.2011.03.009 (2011) (**Biogenic Volatile Organic Compounds S.I**).
2. Moser, B. et al. Mass spectrometric profile of exhaled breath-field study by ptr-ms. *Respir. Physiol. & Neurobiol.* **145**, 295–300. https://doi.org/10.1016/j.resp.2004.02.002 (2005).
3. Mazzucotelli, M. et al. Proton transfer reaction mass spectrometry: A green alternative for food volatilome profiling. *Green Anal. Chem.* **3**, 100041. https://doi.org/10.1016/j.greeac.2022.100041 (2022).
4. Cappellin, L. et al. On data analysis in ptr-tof-ms: From raw spectra to data mining. *Sensors Actuators B: Chem.* **155**, 183–190. https://doi.org/10.1016/j.snb.2010.11.044 (2011).
5. Bishop, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)* (Springer-Verlag, 2006).
6. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* (MIT Press, 2016). http://www.deeplearningbook.org.
7. Grinsztajn, L., Oyallon, E. & Varoquaux, G. Why do tree-based models still outperform deep learning on typical tabular data?. *Adv. neural information processing systems* **35**, 507–520 (2022).
8. Bommasani, R. et al. On the opportunities and risks of foundation models. *arXiv e-prints* arXiv:2108.07258 (2021).
9. Vaswani, A. et al. Attention is all you need. In Guyon, I. et al. (eds.) *Advances in Neural Information Processing Systems*, vol. 30 (Curran Associates, Inc., 2017).
10. Radford, A. et al. Language models are unsupervised multitask learners. *OpenAI blog* **1**(8), 9 (2019).
11. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186 (2019).
12. Radford, A. et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763 PmLR, 2021).
13. Leung, H. W. & Bovy, J. Towards an astronomical foundation model for stars with a transformer-based model. *Mon. Notices Royal Astron. Soc.* **527**, 1494–1520 (2024).
14. Bushuiev, R. et al. Self-supervised learning of molecular representations from millions of tandem mass spectra using dreams. *Nat. Biotechnol.* 1–11 (2025).
15. Brown, T. et al. Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. & Lin, H. (eds.) *Advances in Neural Information Processing Systems*, vol. 33, 1877–1901 (Curran Associates, Inc., 2020).
16. Hollmann, N., Müller, S., Eggensperger, K. & Hutter, F. Tabpfn: A transformer that solves small tabular classification problems in a second. In *International Conference on Learning Representations (ICLR)* (2023).
17. Hollmann, N. et al. Accurate predictions on small data with a tabular foundation model. *Nature* https://doi.org/10.1038/s41586-024-08328-6 (2025).
18. Breiman, L. Random forests. *Mach. learning* **45**, 5–32 (2001).
19. Guo, Y., Hastie, T. & Tibshirani, R. Regularized linear discriminant analysis and its application in microarrays. *Biostatistics* **8**, 86–100 (2007).
20. Barker, M. & Rayens, W. Partial least squares for discrimination. *J. Chemom. A J. Chemom. Soc.* **17**, 166–173 (2003).
21. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297. https://doi.org/10.1023/A:1022627411411 (1995).
22. Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics* 1189–1232 (2001).
23. Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. Least angle regression. *The Annals of statistics* **32**, 407–451 (2004).
24. Pedregosa, F. et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
25. Corvino, A. et al. Rapid profiling of volatile organic compounds associated with plant-based milks versus bovine milk using an integrated ptr-tof-ms and gc-ms approach. *Molecules* **30**(4), 761. https://doi.org/10.3390/molecules30040761 (2025).
26. Yener, S. et al. Rapid and direct volatile compound profiling of black and green teas (camellia sinensis) from different countries with ptr-tof-ms. *Talanta* **152**, 45–53. https://doi.org/10.1016/j.talanta.2016.01.050 (2016).
27. Telagathoti, A., Probst, M., Khomenko, I., Biasioli, F. & Peintner, U. High-throughput volatilome fingerprint using ptr-tof-ms shows species-specific patterns in mortierella and closely related genera. *J. Fungi* **7**(1), 66. https://doi.org/10.3390/jof7010066 (2021).
28. Khomenko, I. et al. Ptr-tof-ms voc profiling of raw and cooked gilthead sea bream fillet (sparus aurata): Effect of rearing system, season, and geographical origin. *Molecules* **30**(2), 402. https://doi.org/10.3390/molecules30020402 (2025).
29. Khomenko, I. et al. Integrated approach for the evaluation of food loss and waste of fresh spinach during its storage. In *8th MS Food Day, Torre Canne (BR), October 16-18, 2024*, 207–208 (IT, 2024).
30. del Pulgar, J. S. et al. Rapid characterization of dry cured ham produced following different pdos by proton transfer reaction time of flight mass spectrometry (ptr-tof-ms). *Talanta* **85**, 386–393. https://doi.org/10.1016/j.talanta.2011.03.077 (2011).
31. Rajendran, S. et al. The effect of different medium compositions and lab strains on fermentation volatile organic compounds (vocs) analysed by proton transfer reaction-time of flight-mass spectrometry (ptr-tof-ms). *Fermentation* **10**(6), 317. https://doi.org/10.3390/fermentation10060317 (2024).
32. Yener, S. et al. Tracing coffee origin by direct injection headspace analysis with ptr/sri-ms. *Food Res. Int.* **69**, 235–243. https://doi.org/10.1016/j.foodres.2014.12.046 (2015).
33. Pedrotti, M., Spaccasassi, A., Biasioli, F. & Fogliano, V. Ethnicity, gender and physiological parameters: Their effect on in vivo flavour release and perception during chewing gum consumption. *Food Res. Int.* **116**, 57–70. https://doi.org/10.1016/j.foodres.2018.12.019 (2019).
34. Farneti, B. et al. Direct injection and chromatography, mass spectrometry and ion mobility: a synergic approach for strawberry volatilome analysis. In *7 MS Food Day, Florence, Italy, October 5-7, 2022*, 289–291 (IT, 2022).

35. Mazzucotelli, M. et al. Characterization of hazelnut volatilome evolution during roasting by ptr-tof-ms, gc-ims, gc-ms and advanced data mining methods. In *Contributions 9th International Conference on Proton Transfer Reaction Mass Spectrometry and its Applications*, 174–176 (Innsbruck University Press, 2024).

Acknowledgements
Part of the results presented in this work were obtained using the facilities of the CCT-Rosario Computational Center, member of the High Performance Computing National System (SNCAD, Argentina).

Author contributions
P.M.G. conceived the study, developed all the software, performed the numerical experiments, analyzed the results and wrote the original draft. M.P., E.B., and I.K. produced and provided the datasets. M.P. and A.R. contributed to the analysis and revised the original draft. F.B. co-conceived the study, revised the manuscript, and supervised the project. All authors contributed to the writing and approved the final version of the manuscript.

Funding
PMG thanks funding from Provincia Autonoma di Trento (PAT) through the "Visiting in Trentino" 2024 call, and CONICET. This work has been partially supported by the SISTERS project, which has received funding from the European Union's Horizon 2020 research and innovation program under Grant Agreement No. 101037796.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to P.M.G. or F.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.