Composition-based statistical model for predicting CO2 solubility in modified atmosphere packaging application

Mélanie Münch, Valérie Guillard, Sébastien Gaucel, Sébastien Destercke, Jonathan Thévenot, Patrice Buche

Composition-based statistical model for predicting CO_2 solubility in Modified Atmosphere Packaging application

Mélanie Münch[a,b 1], Valérie Guillard[b], Sébastien Gaucel[b], Sébastien Desterckе[c], Jonathan Thévenot[d], Patrice Buche[b]

[a] I2M, Université de Bordeaux, INRAE, 33400 Talence, France

[b] UMR IATE, Université de Montpellier, INRAE, Institut Agro, 34090 Montpellier, France

[c] UMR C.N.R.S. 7253 Heudiasyc, Université de Technologie de Compiègne, 60203 Compiègne, France

[d] ADRIA Food Technology Institute – UMT ACTIA 19.03 ALTER'IX, ZA Creac'h Gwen F29196 Quimper Cedex 1, France

Abstract (100-150 words in length)

Carbon dioxide (CO_2) is an important gas used in modified atmosphere packaging of non-respiring foods where it solubilizes into the aqueous and lipid phases of food and exerts an antimicrobial effect. Prediction of CO2 solubility within food is thus of paramount importance to anticipate its benefit on food preservation. In the present study, machine learning algorithms were applied on a set of 362 values of CO_2 solubilities collected from the scientific literature to tentatively predict the solubility as a function of food composition (water, protein, fat and salt content) and temperature. The best option kept was a random forest algorithm that was used to predict CO_2 solubility in four food case studies (ham, salmon, cheese and pâté) that were further

[1] Corresponding author

used as input parameters in the MAP' OPT tool, predicting the evolution of headspace gas composition. Predicted CO_2 solubilities used as input parameters succeeded in representing the CO_2 headspace dynamic as a function of time in the four case studies.

Keywords

CO_2 solubility; machine learning models; food composition; Modified Atmosphere Packaging; CO_2 headspace dynamic

Abbreviations

1. Introduction

In Modified Atmosphere Packaging (MAP) applications, the packaging atmosphere is generally replaced by a mixture of different gases mainly composed of O_2, CO_2 and N_2 in order to prevent food degradation during storage. CO_2 is often used for its bacteriostatic effect. The concentration of CO_2 injected in the pack is calculated to be close to or above the minimal inhibitory concentration for microorganism's growth (Farber, 1991) and this CO_2 concentration must be maintained as much as possible into the packaging to keep its benefit along the food shelf-life. However, CO_2 concentration varies during storage due to the CO_2 permeation from the internal atmosphere toward the surrounding, and due to the solubilization and diffusion of CO_2 into the food product initially free of dissolved CO_2 (Chaix et al., 2015; Guillard, Couvert, et al., 2016; Simpson et al., 2001). If the loss of CO_2 due to permeation may be well mastered by using high barrier packaging films (Guillard et al., 2017), the CO_2 solubilization into the food is unavoidable and leads to rapid CO_2 partial pressure drop into the packaging, to an extent that depends on the headspace to food volume ratio and nature of the food. The lower the headspace volume is, the higher the CO_2 drop is, due to gas solubilization into the food. This

CO_2 solubilization is governed by Henry's law: at equilibrium and for constant temperature and pressure (Eq. 1) the concentration of dissolved CO_2 (C_{CO2}) in a product is proportional to its partial pressure (p_{CO2}) in the surrounding atmosphere (Chaix et al., 2014; Henry, 1832):

$$C_{CO_2} = S_{CO_2}(T) \times p_{CO_2} \qquad \text{Eq. 1}$$

where $S_{CO2}(T)$, the inverse function of Henry's law coefficient, is the solubility coefficient, at temperature T, expressed in $mol.kg^{-1}.Atm^{-1}$. It represents the maximal quantity of CO_2 that could be dissolved in a product for a given partial pressure of CO_2. The value of solubility depends on the nature of the food and reflects the compatibility between CO_2 and the food matrix (Chaix et al., 2014; Rotabakk et al., 2007; Schwartz, 2003). The knowledge of this data is thus of paramount importance to anticipate CO_2 losses by solubilization and its expected effect on food shelf-life. This CO_2 solubility is an input parameter required in MAP modelling tools that permit the prediction of the evolution of internal gas composition as a function of time (Chaix et al., 2015; Guillard, Couvert, et al., 2016; Simpson et al., 2001) and accuracy of this data is crucial for prediction's relevance.

CO_2 solubility is generally determined using costly and time-expensive experimental set ups: nowadays there are no low-cost techniques available for a non-invasive determination of gas concentration in solid matrices, which makes automatization of the measurement difficult (Chaix et al., 2014). Lab made experimental set-ups are generally needed and measurement requires equilibrium to be reached (24h-48h). Methodologies used to measure CO_2 solubility have been reviewed by (Chaix et al., 2014). These authors also proposed a first database of values that has been recently updated by (Guillard, Buche, et al., 2016) and (Munch et al., 2022). In the last version, 362 solubility values were available for 81 different food products. If the link between food type and value of CO_2 solubility is not straightforward, it seems nevertheless that food composition (fat, water, proteins or salt content) has a strong impact. For instance, CO_2 solubility was found higher in fat products than in aqueous ones: at 22°C, CO_2

solubility was found, respectively, 1.6 and 1.8 times more soluble in olive oil and grape seed oil than in water (Pauchard et al., 1980). (Jakobsen & Bertelsen, 2006) have demonstrated that there is a significant difference between the amounts of CO_2 that can be absorbed in meat with different fat contents. The CO_2 absorption increases along with the increasing content of unsaturated fat. CO_2 solubility was found to significantly decrease in cheese with increased salinity (from 0 to 2.7% NaCl w/w) (Acerbi et al., 2016). The CO_2 solubility of renneted casein matrices was found to decrease linearly with salt-in-moisture content, whereas it increased with increasing pH and non-linearly varied with the moisture-to-protein ratio (Fava & Piergiovanni, 1992; Jakobsen et al., 2009; Lamichhane et al., 2021). In all cases, beyond compositional aspects, temperature was identified as the most impacting parameter on the CO_2 solubility value with, in general, a decrease of solubility with an increase in temperature. Interference between temperature and physical state of lipids into the food formulation was also observed making trends more difficult to interpret and formalize: for instance, in seafood model products with varying lipid profile, liquid fat leads to a similar solubility of CO_2 as water, while CO_2 only being minimally dissolved in solid fats (Abel et al., 2018).

Faced with the importance of accurate CO_2 solubility predictions and lack of low-cost and rapid methods for its determination, some authors have attempted to develop empirical mathematical models (mostly regressions) between CO_2 solubility and temperature and one or more compositional parameters. One of the first models was proposed by (Fava & Piergiovanni, 1992) and related using multiple linear regressions CO_2 solubility and compositional parameters (fat, protein, moisture, pH, water activity) of different foods at one temperature (7 °C). However, if this model was found suitable for meat products, it failed to predict solubility in dairy products. After this preliminary attempt, a second model was proposed by (Jakobsen et al., 2009) to predict CO_2 solubility in semi-hard cheese based on the weight fraction of water (w_w) and fat (w_f) in the 2–phase cheese system, temperature (T), and the CO_2 solubilities in, respectively,

pure water ($S_{CO2/W}$) and pure fat (butterfat, $S_{CO2/F}$). This model succeeded in predicting CO_2 solubility in semi hard cheeses and the range of temperature from 0 to 20◦C investigated by the authors. Another more recent model was the one proposed by (Acerbi et al., 2016) that linked the CO_2 solubility (SCO2 in $mmol.kg^{-1}.atm^{-1}$) in Maasdam type cheese to temperature and salt-in-moisture (S/M) content (Eq. 2):

$$S_{CO2} = 37.92 - 0.35\,T - 1.21\,S/M \qquad \text{Eq. 2}$$

However, the main drawback of all the modelling attempts mentioned above is that they are no longer valid when they are extrapolated to products that were not initially included in the initial range of data used for their setting up. For instance, (Chaix et al., 2014) have tested linear correlations between fat and water content and CO2 solubility determined in water, hake sausage, and ham. Although well accurate for those products, these correlations failed to predict solubility into fish products with an error of more than one order of magnitude (about 90%). This limits their usefulness and well illustrates the difficulty of finding a simple and universal linear model or correlations that are valid for large domains. In addition, it is difficult to draw clear and fair conclusions about the impact of food composition on CO_2 solubility because temperature often interferes with other effects, even masking them sometimes, and only one class of food is examined at a time which makes it very difficult to conclude about the real effect of compositional parameters. pH may also interfere by modifying the ratio of dissolved CO_2 species into the food, e.g. carbonic acid, bicarbonate ion, and carbonate ion (Chaix et al., 2014). This lack of generalization of state-of-the-art solubility predictive attempts is a real problem to extend virtual MAP modelling tools (Guillard et al., 2017) to decision-making where multiple simulations with various food products would be required.

Artificial intelligence tools can bring generalization power by inducing global models from data, that are able to deal with such different behaviors, both by (1) learning models for prediction and extrapolation; and (2) structuring available knowledge and extracting new ones

from data. For the first part, different works in machine learning can be noted in the case of solubility prediction for saline solutions (Boobier et al., 2020; Vo Thanh et al., 2022). However, to the best of our knowledge, no attempt has been made yet for predicting CO_2 solubility in food using machine learning algorithms. In this work, we would like to tackle this issue using families of standard machine learning methods, in order to assess their performances. Our main purpose is to evaluate their ability to predict the food product's solubility from the temperature and composition alone. To do so, we compare three families of algorithms: linear, local and ensemble methods (better described in Section 3.2.1.). While the chosen models have different characteristics, they are all dedicated to the prediction of a value (in our case, the solubility) given an entry vector (temperature and composition), and thus represent good candidates for evaluating the ability of machine learning approaches for our problem.

For the second part, knowledge engineering is a sub-domain of artificial intelligence, using methods and tools based on ontologies, that can be helpful to extract knowledge semi-automatically (Lentschat et al., 2022), and to annotate experimental data from scientific papers (Buche et al., 2013) in order to be able to realize meta-analyses. Moreover, a semi-automatic mapping between ontologies dedicated to the food domain description permits to manage the problem of data incompleteness, especially in terms of food product compositional parameters (Buche et al., 2021). In the case of CO_2 solubility prediction, knowledge engineering could be useful for structuring the different relations between the solubility and the different input parameters (e.g. compositional parameters, temperature), as well as retrieving missing information from other databases.

In this context, the aim of this paper is to present an innovative composition-based statistical model of CO_2 solubility as a function of temperature (T) and compositional parameters (fat, moisture, protein and salt contents). To avoid any bias due to the numerical treatment of a specific, focused set of data and to enlarge the analysis to all kinds of foods available in the

147 scientific literature, an exhaustive dataset of all CO_2 solubility was first created. Compositional
148 parameters were retrieved from the original paper or inferred using the MultiDB explorer tool[2]
149 and were capitalized in the dataset too. Multiple machine learning algorithms were then
150 evaluated on the dataset in order to identify the most suitable model for predicting CO_2 solubility
151 as a function of T and composition. Its predictions of CO_2 solubility for 4 different food products
152 (ham, salmon, cheese and pâté) were then used to feed the MAP'OPT modelling tool (Guillard
153 et al., 2017) which predict evolution of CO_2 composition into packaging headspace. The
154 theoretical CO2 headspace composition for these 4 products was confronted to experimental
155 measurements to validate the composition-based statistical model proposed.
156

157 # 2. Material and methods

158 ## 2.1. Food products

159 Ham, salmon, cheese and pâté were purchased in local supermarkets. Nutritional composition
160 information of the food products used for the validation are presented in Table 1.

161 ### 2.2. Shelf-life experiments

162 Exactly 100 g of each food product were packaged in high density polyethylene (HDPE) trays
163 with a volume capacity of 375 cm^3 (530 XX 00, PROMENS, Norway) and a minimal thickness
164 of 200 µm. The gas transmission rates of this tray are 3 and 13 cm^3/day.atm for O_2 and CO_2
165 respectively. Each sample were placed in a cooling cell to reach a core temperature of 4 °C
166 before their sealing with a lidding film in PE (42.0 ± 4.2 µm thick) with less than 5 and 25
167 cm^3/m^2.day.bar of O_2 and CO_2 permeance respectively (Lintop PE HB B 42, LINPAC
168 PACKAGING, France) using an OPE 1000C tray sealer (Guelt, France) configured to modify
169 the headspace atmosphere with a gas mixture of 30% of N_2 and 70% of CO_2. This step took

[2] https://ico.iate.inra.fr/meatylab

place in a laminar flow hood to avoid any microbiological contamination. The samples were stored at 4 °C until analysis for 5 days. Daily monitoring of headspace CO_2 was made using a Check Mate 9900 calibrated annually by the supplier (Dansensor / AMETEK, France). The principle of dosing is based on an infrared sensor for CO_2.

2.3. MAP'OPT: mathematical model for headspace CO_2 dynamic

The mathematical model developed by (Guillard et al., 2017) was used to predict the variation of the O_2 and CO_2 concentration in the headspace of packaged food products in the present shelf-life experiment (i.e., ham, salmon, cheese and pâté). This semi-mechanistic model included (i) $O_2/CO_2/N_2$ transfer between headspace and external atmosphere via permeation through the lid material and the tray in contact with headspace, (ii) O_2/CO_2 sorption or desorption characterized by solubilization and diffusion within the food product, (iii) variations in headspace volume and composition obeying the ideal gas law while maintaining a total pressure equal to the set pressure of the tray sealer and (iv) temperature effect on all the aforementioned mechanisms according to Arrhenius equation. The input parameters needed to run the simulation depend on the characteristics of the packaging (volume capacity, thickness of the tray and lid, exposed area, gas permeation), storage (composition of the gas mixture, temperature, duration preservation) and of the food product (solubilization and diffusion of gases, mass, density, thickness, information on nutritional composition). The O_2 diffusivity and solubility, at 4°C, were fixed respectively to 1.2 x 10^{-9} m²/s and 2 x 10^{-8} mol/(kg.Pa) from (Chaix et al., 2014). The CO_2 diffusivity (in m²/s), at 4°C, was estimated, for each product, according to (Chaix et al., 2014) by:

$$D_{CO_2} = 3 \times 10^{-10}\%\text{fat} + 1 \times 10^{-9} \qquad \text{Eq. 3}$$

Valid in the range of temperature [0, 8°C], where D_{CO_2} is the diffusivity of CO_2 (m²/s) and %fat is the fat content (%w/w in wet basis) of food products.

The CO_2 solubility for the 4 food case studies was predicted using the model developed in this study and were used as input parameters for CO_2 solubility.

2.4. Evaluating Statistical Models for CO2 solubility

While machine learning algorithms are numerous and can virtually be applied to any cases, their performances often vary greatly between application cases. In order to elect the best model, different algorithms were compared in our study. To do so, we use a 10-folds cross-validation (CV), which allows to separate the dataset into two parts, a learning set (used for learning a model) and a testing test (used for evaluating the learned model). To ensure a good precision in the results, this operation is repeated 10 times, while changing the composition of both the learning and the testing sets. For each fold, a score is computed. The final score represents the mean of these different results, and determines the average predictive performances of the tested algorithm for the given dataset. As shown in Fig. 1, which illustrates a 4-folds validation, testing and training sets do not overlap between folds (i.e., the test sets form a partition of the data). To validate even more, we will also use a LOO (Leave-One-Out) procedure, corresponding to a n-fold cross validation. Note that (Bengio & Grandvalet, 2004) shows that K-fold cross-validation has no unbiased estimator of its variance, meaning that its performances will depend on the internal variation of the considered dataset. This is not a major drawback in our case, as we mainly use these tools to compare different algorithms predictive capabilities.

All experiments were implemented using the Python library Scikit-learn (Pedregosa et al., 2011), which is dedicated to machine learning. Unless otherwise stated, the library's default algorithm's parameters were used. Further explanations of the results were done using the Python SHAP library (Lundberg & Lee, 2017), which allows to compute the relative importance of features in a prediction task using the game-theoretic notion of Shapley value. The choice of this method was motivated by its agnostic aspect: as its results do not depend on

the selected model, it provides insights and explanations that are less dependent on it. Such methods are also applicable to other models, and therefore in future works, one could try to see if using other models with similar capabilities would provide the same explanations.

2.5. Statistical analysis

For shelf-life experiments, significant differences in headspace composition between food matrices and time points were tested using the nonparametric Kruskal test with the "kruskal.test" function in statistical software R 3.6.1 (R. C. Team, 2019). In case of significant food matrix effect, Dunn's test for stochastic dominance among food matrix groups was computed using the function "dunn.test" of the R package "dunn.test" (Dinno, 2019) and $P < 0.1$ was considered as significant.

3. Results and Discussion[3]

3.1. Data collection

362 data from 21 references of the literature were collected and stored in a dedicated database. Solubility unit kept for the following is $mmol.kg^{-1}$. Atm^{-1} for the sake of clarity. Corresponding food compositions were retrieved directly from the original paper or, if not provided in the source paper, retrieved from the Food Composition database (Buche et al., 2021). Four constituents (water, fat, protein and salt) were kept for further analyses (sugar was discarded due to many null or missing values, which would not have brought more information to the model). This choice was motivated by analysis of previous literature on the topic, as fat content was found particularly relevant (Jakobsen & Bertelsen, 2006; Pauchard et al., 1980). However, while the lipid profile and physical state of lipids was also proved to be important, especially its interrelationship with temperature (Abel et al., 2018), it was not possible to consider it in

[3] All data and source codes are available at the following URL: https://doi.org/10.57745/QRBX4Z

this approach because lipids profile was most of the time simply unknown or impossible to retrieve with enough precision. On another hand, protein and moisture contents were also kept because several times quoted as relevant compositional parameters influencing CO_2 solubility (Lamichhane et al., 2021). More specifically, (Fava & Piergiovanni, 1992) considered fat, protein, moisture, pH, water activity in their model of CO_2 solubility. In the present study, pH and water activity were discarded because they are not available in the food composition database. Finally, salt content was also kept as several times quoted for its impact on CO_2 solubility (Acerbi et al., 2016; Duan & Sun, 2003).

In the end, the constituted database presents mainly three categories of food products: dairy products, meat and fish. It was also complemented with measures made on water and oil. While this distinction of "type" was kept for data description purposes, it was not used as a variable during the learning: the composition was considered to be sufficient for predictability purposes. For each food product, temperature was also kept as one of the main factors affecting CO_2 solubility value. Even if the temperature effect was in general well modelled using Arrhenius' law (Chaix et al., 2014), it was decided in the present work to consider it as a parameter in addition to composition in the statistical model and to not model its effect using Arrhenius' law.

Once the data collected, an additional pretreatment was applied after the preliminary descriptive analysis: since some data were repetitions made on a same sample (for instance, there are 12 repetitions for Maasdam cheese at 25°C), the average solubility was considered in those cases in order to reduce the dominance of certain food products. After these pretreatments, 258 data from the original 362 values collected were kept and used in machine learning algorithms.

3.2. Learning models / prediction of CO_2 solubilities

3.2.1. Model used, learning

We considered three types of algorithms: linear methods (which aim to learn linear relationships), local methods (which aim to learn local models for the different parts of the dataset) and ensemble methods (which aim to learn multiple models in order to enhance the predictive performances and reduce variance of the predictions).

Linear methods (and their extensions) are prototypical of statistical parametric methods: they make some strong assumptions about the relationships between the data, meaning that they have a high potential bias but low variance. This means that if their assumption is true, they will require few data to have a very good predictive power and will come with powerful statistical tool to select features, explain results etc. In contrast, if their assumption is false (as will be the case here for linear models), they are likely to produce models with poor predictive power, and will provide potentially misleading conclusions. In contrast, local or regionalized methods typically make very few assumptions, meaning that they have a low bias but a high variance. They are likely to provide good predictive power in all cases, but come with less powerful statistical tools, and can strongly vary if the data are modified, meaning that they can be instable and that one should be careful about their conclusions, especially when having few data points. Due to their localized nature, they are usually interpretable models. Ensemble methods try to achieve a low bias with a low variance, by making very few assumptions and by averaging a (usually large) set of simpler models. Due to their high flexibility and the use of averaging, they usually achieve very high predictive performances, but are by nature poorly interpretable and extendable. They must therefore be complemented by additional tools if one wants to analyze/interpret their results, and should be used in those cases where simpler models failed to deliver satisfactory results. We will see in the next pages that our study falls into this category, at least when one restricts to linear models for the global ones.

For each, we selected a few classical algorithms and performed 10-fold cross validations, whose results are presented in Table 2. We also present the results of a particular type of cross-

validation, the Leave One Out (LOO). Better fitted for small datasets, LOO is learned for each split using all data except one, which is used for the testing part. If we have n data, LOO corresponds to an n-folds cross validation. While it can lead to overfitting (i.e., learning a model that memorize the training set but extrapolate/generalize poorly to unseen data points), it also gives a good overview of the model's performances when trying to predict data close to the original dataset.

As we can see, ensemble methods perform the best. This is not surprising considering our dataset, which presents very different products on variable conditions (temperature, composition). To deal with them, our model needs to be able to (1) describe multiple (possibly linear) regimes of CO_2 evolution depending on the original conditions (which is not fitted for linear methods, that can only describe well one linear regime) and (2) keep a coherent continuity between these different regimes (for which local methods are not fitted, as predictions can change abruptly when modifying slightly conditions). Ensemble methods, on the contrary, are based on the learning of multiple simplified models (decision trees in the case of Random Forests), whose predictions are computed in order to select an average result; this allows both the adaptability and the continuity of the learned model. As a consequence, we adopt for the rest of this article the Random Forest regression, which obtained the best overall score. While its performances are not perfect (which is due, as we will see, to the diversity of our dataset), it presents promising results and seems to be the best suited for our application.

Random Forests are an ensemble method based on the learning of multiple decision trees from a random sample of the whole dataset. This approach avoids the over-fitting tendencies of decision trees through averaging, and proposes a better adaptability to the data's inner variations. Since the number of trees has an impact over the final result, we have used another 10-folds CV to fine tune the parameter and find the best possible combination. We have tested with 50, 100, 150, 250 and 500 trees, without denoting a drastic change in the performances; as

317 a consequence, our final model was learned with the dataset previously presented and 100 trees,
318 which correspond to the default value in Scikit-learn. To be noted, due to the multiplicity of
319 methods and features to optimize, we only focus in this article on fine-tuning the method elected
320 after the cross-validation made on the library's default parameters.
321
322 **3.2.2. Impact of food composition on predicted CO_2 solubility**
323 The impact of both food composition and temperature on the predicted solubility can be now
324 analyzed from the learned model. First of all, we analyze the sensitivity of each parameter by
325 learning multiple models with truncated information (only one parameter, then two, etc.). The
326 objective is to compute the scores' difference (and thus the quantity of knowledge) brought by
327 the addition of information. Part of the results are presented in Table 3, where we can see that
328 adding the temperature's value to a nutrient composition drastically enhances the quality of the
329 model, confirming the key role of temperature on the reliability of CO_2 solubility prediction.
330 Indeed, while temperature or compositional parameters alone are not enough to predict the
331 solubility, the combination of temperature and at least one of the compositional parameters can
332 give a rather good prediction, which can be further improved by adding the other compositional
333 parameters. On the contrary, the combination of multiple compositional parameters alone is not
334 enough: for instance, a model learned solely with the fat and water parameters has a score of
335 0.35; which is very close to the score of a model learned with all four compositional parameters
336 without knowledge of the temperature (0.40). This result well highlights the importance of
337 considering both criteria, temperature and compositional parameters, for an accurate prediction
338 of the CO_2 solubility. To be noted, a model learned with temperature alone has a very bad score
339 (0.04). Temperature alone is thus not enough to explain the variability of CO_2 solubility
340 observed.

Similarly to many black-boxes models and in contrast to using, e.g., on decision tree, random forests can be hard to interpret. As predictions are based on the combination of multiple decision trees, explanations are not direct as we have no clear dependency between the parameters and the final result. In order to understand the role of the compositional parameters in the prediction, the Shapley's value of each nutrient was computed using the SHAP Python library. Shapley's values are used in game theory to express a property's contribution to a final result, considering both its individual contribution as well as its marginal contribution when combined with other properties (accounting for interactions): the higher it is in absolute value, the more this property has influenced the final decision. In the following, we distinguish positive and negative influences: in our case, the first tends to increase the CO_2's solubility value, while the second tends to decrease it.

Fig. 2 shows the evolution of Shapley's values depending on the parameters for every measure of our dataset. We can see that the repartition of the Shapley's values for the temperature are strongly correlated to its value, as expected from the state of the art: the higher the temperature is (to the left of the figure, as indicated in the Feature value's legend), the lower the Shapley's value is, indicating a negative impact over the final solubility. On the contrary, a low temperature (this time on the right side) is correlated to positive Shapley's values, and thus will have a positive effect on the solubility value.

However, most of the compositional parameters' influence cannot be characterized as easily: the fat, for instance, seems to have low SHAP values, lower than those obtained for water. Thus, it appears that fat might have lower effect than water on CO_2 solubility and would positively or negatively impact this solubility (both positive and negative Shapley's values were observed for fat), depending on other factors that are not shown in this figure and may be absent from the data set. Since Table 3 has highlighted a strong interaction between compositional parameters and the temperature, we display Shapley values on two axes (temperature + constituent) to

observe impact of this interaction on the final prediction in order to describe precisely these results. Fig. 3 shows an example of this interaction in the case of the water and temperature (a) and the fat and temperature (b). In contrast with Fig. 2, this now clearly shows how the addition of temperature increases the precision of the model. If we consider again the example of the fat (right-most figure), we can see that the Shapley's value varies between -1 and 1.5 depending of the fat value and the temperature: for instance, given a fat composition of 10, lower temperatures (under 10°C) have a rather positive impact; while higher temperatures (over 15°C) have a rather negative impact. This tendency shifts for pure-fat products: here, a high temperature will have a positive impact, while a low temperature has a negative impact. On the other hand, the left-most figure shows an inverse tendency for the interaction between water and solubility on the temperature.

However, it is important to note that in both cases, we have represented in red the combinations represented in the learning dataset. This is important, as we can see in the case of the temperature/fat graph that nearly all predictions between 30 and 100% of fat are inferred, as there was no product with that quantity of fat in our learning dataset (which is credible, since apart from certain particular food products such as oil or butter, products with fat content above 30% are rather scarce). As a consequence, the model has extrapolated the result (and the importance of the parameters in its prediction) from similar results, and not from concrete and observed data. This could lead to false interpretations, and highlight the limit of our model in its current state: while predicting solubility of items similar to the ones used for the learning can be reasonably trusted, the more a food product will be remote from the original learning set, the more difficult and not trust-worthy its prediction will be. Put another way, while provided inferences in unexplored areas appear plausible, they should be further checked by concrete experiments.

3.2.3. Comparison with mechanistic models from the literature

391 The literature well highlighted the impact of temperature on CO_2 solubility, which generally
392 decreased with temperature following a Van't Hoff type equation with a negative enthalpy of
393 sorption (Chaix et al., 2014). For instance, (Acerbi et al., 2016) found a decrease of CO_2
394 solubility with increasing temperature in the range 2-25°C for Maasdam cheeses, in agreement
395 with previous observations made by (Jakobsen et al., 2009) in similar semi-hard cheeses. CO_2
396 solubility of water is decreasing with temperature too (Carroll et al., 1991; Dean, 1999).
397 However, this effect of temperature seems to interact with compositional parameters. Thus,
398 solubility of CO_2 was found to slightly increase in pure dairy fat (99% fat) with increasing
399 temperature from 3 to 19 °C (Jakobsen et al., 2009). Therefore, a compensating effect may
400 occur for products rich in fat, resulting in smaller temperature variation than expected for
401 example in cheese with high fat content as observed by (Jakobsen et al., 2009) or even an
402 increase of CO_2 solubility with temperature as observed in fatty meat samples (Jakobsen &
403 Bertelsen, 2006). This effect of temperature and its interaction with fat content effect is well
404 captured by our model. Indeed, as shown on Fig.3 (a), for water content above 60-70%, the
405 temperature has a strong negative effect on CO_2 solubility as generally experimentally observed
406 in aqueous-based phases with low fat content. In agreement with those findings, at low fat
407 contents (below 30%) and, thus, corresponding assumed high moisture content, CO_2 solubility
408 is negatively correlated to temperature increase (Fig. 3 (b)). On the opposite, above the
409 threshold fat content of 30% (and corresponding supposed lower moisture content) solubility
410 becomes positively correlated with temperature, confirming findings of literature studies
411 (Jakobsen et al., 2009; Jakobsen & Bertelsen, 2006).
412
413 This antagonistic effect between fat and moisture contents is also obvious on Fig. 4 (a)
414 presenting the interaction of the water and fat contents and the corresponding Shapley's value.
415 It is clearly visible that above 30% of fat content, the CO_2 solubility is governed by the fat phase

that tends to negatively impact the solubility, while for fat content below this threshold value, moisture phase's impact predominates with a slight trend to positively increase solubility until nevertheless a certain extend; above a threshold value of 60-70% of water content, its influence tends to become slightly negative.

Fig.4 (b) shows interaction of the protein and fat contents and the corresponding Shapley's value. It shows that for products with fat content below 30%, protein content tends to negatively impact CO_2 solubility. On the contrary, above 30% of fat content, protein content positively impacts solubility. In other words, below 20% of protein, increasing fat content has a slight positive impact on CO_2 solubility until a threshold value of 30%. Above this threshold value of 30% of fat, this positive effect turns into a negative one. However, in both cases, the effect is low with absolute SHAP-value below 1. In addition, for fat content higher than 30%, there are only few data (red open symbols on Fig. 4 (a)) and data are thus mostly extrapolated by the model and should be considered cautiously. This interaction between protein and fat contents was never related in the literature. If the impact of protein contents was clearly identified on CO_2 solubility, it was never clearly stated to what extent it would affect these solubility values. For instance, (Jakobsen & Bertelsen, 2006) observed that CO_2 absorption increases along with the increasing fat content (from 2 to 65%) into mixtures of muscle and fat (from pig meat) but they did not mention the protein contents of their samples making difficult to align their study on the results shown in Fig. 4 (b). Nevertheless, supposing that pig meat contains a maximum of 20% of proteins (from the French food composition table (Anses, 2020)), we can estimate that protein content varies from 19.6% for 2% of fat content to 7% at the lowest for the fattiest mixture. We are thus below the threshold value of 20% of proteins where increasing fat content tends to increase CO_2 solubility into such samples (Fig. 4 (a)). Findings of (Jakobsen & Bertelsen, 2006) tend to confirm the prediction of our model.

441 The impact of proteins on CO_2 solubility is quite complex and singular behavior has been
442 observed in the literature that is not completely well captured by our model. For instance,
443 (Lamichhane et al., 2021) noted that the relationship between moisture-to-protein ratio and CO_2
444 solubility was non-linear in casein matrices (~0% fat content). An increase of solubility was
445 first observed for moisture-to-protein ratio ~0.03 to ~0.5 (e.g. protein content ~90 to ~70), then
446 a slight decrease from ~0.5 to ~1.7 moisture-to-protein ratio followed by a small and significant
447 increase (from ~1.7 to ~2.7 moisture-to-protein ratio, e.g. ~35 to ~23% of proteins). Such
448 complex relationships observed between CO_2 solubility and moisture-to-protein ratio which is
449 ascribed to interactive effects of moisture and protein content on CO_2 solubility, is not
450 represented by our model (Fig. 4 (b), points obtained for fat contents close to 0) probably
451 because those data with various moisture-to-protein ratios were not considered in the model
452 learning.
453

454 ## 3.3. Validation experiment

455 ### 3.3.1. Prediction of CO_2 solubility in the 4 food case studies

456 The composition-based learned model previously presented was used to predict the solubility
457 values for the 4 food case studies used in the validation approach.
458 Results are presented in Table 4.

459 ### 3.3.2. Experimental and predicted CO_2 headspace dynamic for the 4 food case studies

460 Headspace CO_2 composition was followed during the shelf-life experiment (Figure 5).
461 Following the sealing, the CO_2 content decreases in the headspace over time for each of the
462 food matrices. After 5 days, CO_2 contents in the ham packs and pâté packs were the lowest
463 (respectively 59.0 +/- 0.5% and 59.9 +/- 0.7%, n = 4, P = 0.45) compared to the others (66.2
464 +/- 0.1% (n = 2) for the cheese packs (P < 0.03) and 63.7 +/- 0.6% (n = 4) for the salmon packs

($P < 0.1$)). CO_2 content was not different over the first 5 days for cheese and salmon packs ($P > 0.18$).

Simulations were carried out with the MAP OPT tool with the predicted CO_2 solubilities as inputs (§ 3.3.1). Values of each parameter used in the MAP OPT tool were presented in Table 5. Simulated data, with any adjustment of any input parameters, are shown in Figure 5. As evidenced in this figure, the composition-based statistical model predicted CO_2 solubilities used as input parameters in the MAP OPT tool rather succeeded in representing the CO_2 headspace dynamic as a function of time in the four case studies. Some variations of CO2 concentration into headspace are nevertheless noted. For cheese and paté, the prediction falls outside the upper/lower predicted curves corresponding to min and max of solubility predicted, respectively. It could be ascribed to uncertainty on the solubility model that tends to deviate when applied to food products that are not well represented in the database. Other sources of uncertainty may occur such as uncertainty on film CO2 permeability or on MAP OPT model hypothesis such as the fact that volume variations are neglected. We can nevertheless consider that the CO2 solubility model is quite satisfactory, in the sense that the error remains of limited value.

The relatively good fitting is also confirmed by the RMSE values equal to 2.78% for ham, 2.09% for salmon, 2.50% for cheese and 3.26% for pâté. We obtained a low value of RMSE which indicated that we can reasonably consider a validation of the gas concentration prediction. Considering the multiples sources of uncertainty in the MAP OPT simulation, taken together, the simulation results validate the composition-based statistical model predicted CO_2 solubilities developed in this study and its generic use for a wide range of products conventionally packaged in MAP. The composition-based statistical model could be included in the MAP OPT tool as a first estimation before further experimental refinement of CO2

489 solubility. It should also be noted that those results are obtained with very few features and a
490 relatively small-size data set, meaning that there is still room for improvements.

491 ## 4. Discussion

492 It is important to keep in mind that our model best shines when presented with predictions
493 similar to data represented in the learning dataset. Indeed, while it may be easy to consider a
494 model learned using machine learning algorithms as objective, it is important to gauge the
495 multiple hidden assumptions that guide its construction. Firstly, as we have seen, the dataset
496 used for learning can easily be biased toward specific food's compositions. Indeed, some food
497 products are over-represented: for instance, the cheese product studied in Sect. 3.3.2. has a
498 compositional profile very close to other food products in our dataset. On another hand, the pâté
499 product, which is the least well predicted, has fewer products with the same profile in the
500 dataset. This is verified by the fact that in Sect. 3.3.2., the second best result has been made on
501 the cheese product (RMSE of 2.50% against 3.26% for pâté), which represents about half of
502 our dataset. Yet, in this article, we propose a proof of concept of the feasibility to predict, using
503 machine learning approaches, CO2 solubility based on food composition and temperature data.
504 Even if extrapolation may be carried out to other food categories not yet quoted in the database
505 used for machine learning, the composition-based statistical model proposed here would be
506 more precise for products whose compositional profile closely matches the ones already
507 represented in the database. Knowing that, it is clear that, for MAP applications where
508 composition fall outside these limits, predictions will be less accurate in a extend that still need
509 to be quantified. However, the database can always be enriched with other data to refine the
510 overall precision, as predictions tend to be better when close to already represented products.
511 Furthermore, it would be possible to send warnings to the user in case a product for which a
512 prediction is given is poorly represented in the data base.
513

514 Moreover, the interpretations (and especially causal interpretations) proposed in this article, in
515 particular using Shapley's values, are made under the assumption that food composition has an
516 impact over the CO_2 solubility; which, as presented in our introduction, has been demonstrated
517 by multiple previous works. In this article, we have verified these assumptions and given a
518 general model able to quantify the impact of these parameters on CO_2 solubility. Indeed,
519 machine learning approaches can be used both to explore hypothesis and make predictions, with
520 the former goal being at least as important as the later in experimental sciences.
521 In the end, machine learning algorithms best shine to represent main tendencies and correlations
522 within a given dataset. They allow to confirm hypotheses (in our case, the influence of a food
523 product's composition and the temperature's measure over its solubility to CO_2) and highlight
524 the importance of a parameter in the final decision; however, one must keep in mind their
525 dependency to the initial assumptions made during their learning and the selected features, in
526 order to avoid abusive extrapolations.
527

528 ## 5. Conclusion

529 In this article, we have presented a novel approach for predicting CO_2 solubility for food
530 products, given their compositional characteristics and their temperature. To do so, we have
531 first compiled an original dataset from 21 references over the past 40 years on the subject of
532 CO_2 solubility. This allowed us to build a learning base with 362 values of CO_2 solubility from
533 which different machine learning algorithms were tested in order to select a model able to
534 predict CO_2 solubility based on temperature on compositional parameters, with a reasonable
535 precision margin.
536 The model presented in this work is a Random Forest, which has been validated by two
537 approaches: (1) theoretically by comparing to state-of-the-art results; and (2) experimentally by
538 confronting experimental headspace CO_2 concentrations measured on 4 different foodstuffs

packed in modified atmosphere packaging (MAP) with predicted ones using a virtual MAP modelling tool integrated the solubility values predicted by our best Random Forest model. In both cases, we have demonstrated the accuracy and genericity of our model.

The purpose of this work is to propose a novel approach to the CO^2 solubility prediction, using classical machine learning algorithms. The interest was both its simplicity (in order to learn a model, we only needed a dataset with the raw values), and the possibilities of explanation provided by tools such as the SHAP values. We wanted to assess whether rather generic machine learning methods were enough to tackle our problem. While we have demonstrated it, it should be interesting to compare their results to more statistical approaches, such as extensions of linear models. Moreover, as mentioned in Section 4, the model's prediction could benefit from the addition of new products, not only from the types considered here, but also from others: this should strengthen the precision of our predictions.

Funding

The data acquired within the framework of the OPTIMAP project was supported by grants from the Regional Council of Brittany, the Departmental Council of Finistère and Quimper Bretagne Occidentale to ADRIA.

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 773375 (GLOPACK project).

Declaration of Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethics statements

This work neither involves human subject nor animal experiments.

CRediT Authors Statements

Patrice Buche: Conceptualization

Sébastien Destercke: Conceptualization, Methodology, Formal Analysis, Review and Editing

Mélanie Münch: Conceptualization, Software, Formal Analysis, Writing – Original, Review and Editing, Visualization

Sébastien Gaucel: Conceptualization, Methodology, Review and Editing

Valérie Guillard: Conceptualization, Validation, Formal Analysis, Writing – Original, Review and Editing

Jonathan Thévenot: Resources, Software, Validation, Writing – Original, Review and Editing

References

Abel, N., Rotabakk, B. T., Rustad, T., & Lerfall, J. (2018). The influence of lipid composition, storage temperature, and modified atmospheric gas combinations on the solubility of CO 2 in a seafood model product. *Journal of Food Engineering*, *216*, 151–158. https://doi.org/10.1016/j.jfoodeng.2017.08.020

Acerbi, F., Guillard, V., Guillaume, C., & Gontard, N. (2016). Impact of selected composition and ripening conditions on CO2 solubility in semi-hard cheese. *Food Chemistry*, *192*, 805–812. https://doi.org/10.1016/j.foodchem.2015.07.049

Anses. (2020). *Ciqual French food composition table*. Https://Ciqual.Anses.Fr.

Bengio, Y., & Grandvalet, Y. (2004). No Unbiased Estimator of the Variance of K-Fold Cross-Validation. *The Journal of Machine Learning Research*, *5*, 1089–1105.

Boobier, S., Hose, D. R. J., Blacker, A. J., & Nguyen, B. N. (2020). Machine learning with physicochemical relationships: solubility prediction in organic solvents and water. *Nature Communications*, *11*(1), 5753. https://doi.org/10.1038/s41467-020-19594-z

Buche, P., Cufi, J., Dervaux, S., Dibie, J., Ibanescu, L., Oudot, A., & Weber, M. (2021). How to Manage Incompleteness of Nutritional Food Sources? *International Journal of Agricultural and Environmental Information Systems*, *12*(4), 1–26. https://doi.org/10.4018/IJAEIS.20211001.oa4

595 Buche, P., Dibie-Barthelemy, J., Ibanescu, L., & Soler, L. (2013). Fuzzy Web Data Tables Integration
596 Guided by an Ontological and Terminological Resource. *IEEE Transactions on Knowledge and*
597 *Data Engineering*, *25*(4), 805–819. https://doi.org/10.1109/TKDE.2011.245

598 Carroll, J. J., Slupsky, J. D., & Mather, A. E. (1991). The Solubility of Carbon Dioxide in Water at Low
599 Pressure. *Journal of Physical and Chemical Reference Data*, *20*(6), 1201–1209.
600 https://doi.org/10.1063/1.555900

601 Chaix, E., Broyart, B., Couvert, O., Guillaume, C., Gontard, N., & Guillard, V. (2015). Mechanistic
602 model coupling gas exchange dynamics and Listeria monocytogenes growth in modified
603 atmosphere packaging of non respiring food. *Food Microbiology*, *51*, 192–205.
604 https://doi.org/10.1016/j.fm.2015.05.017

605 Chaix, E., Guillaume, C., & Guillard, V. (2014). Oxygen and Carbon Dioxide Solubility and Diffusivity in
606 Solid Food Matrices: A Review of Past and Current Knowledge. *Comprehensive Reviews in Food*
607 *Science and Food Safety*, *13*(3), 261–286. https://doi.org/10.1111/1541-4337.12058

608 Dean, J. (1999). Physical properties. Solubilities of gases in water. In *Lange's Handbook of Chemistry*
609 *(15e Ed)* (McGraw-Hill Inc., pp. 375–380).

610 Dinno, A. (2019). *dunn.test: Dunn's test of multiple comparisons using rank sums*. R Package Version
611 1.3.5.

612 Duan, Z., & Sun, R. (2003). An improved model calculating CO2 solubility in pure water and aqueous
613 NaCl solutions from 273 to 533 K and from 0 to 2000 bar. *Chemical Geology*, *193*(3–4), 257–271.
614 https://doi.org/10.1016/S0009-2541(02)00263-2

615 Farber, J. M. (1991). Microbiological Aspects of Modified-Atmosphere Packaging Technology - A
616 Review1. *Journal of Food Protection*, *54*(1), 58–70. https://doi.org/10.4315/0362-028X-54.1.58

617 Fava, P., & Piergiovanni, L. (1992). Carbon dioxide solubility in foods packaged with modified
618 atmosphere. 2: Correlation with some chemical-physical characteristics and composition. *Ind.*
619 *Aliment*, 297–302.

620 Guillard, V., Buche, P., Dibie, J., Dervaux, S., Acerbi, F., Chaix, E., Gontard, N., & Guillaume, C. (2016).
621 CO2 and O2 solubility and diffusivity data in food products stored in data warehouse structured
622 by ontology. *Data in Brief*, *7*, 1556–1559. https://doi.org/10.1016/j.dib.2016.04.044

623 Guillard, V., Couvert, O., Stahl, V., Buche, P., Hanin, A., Denis, C., Dibie, J., Dervaux, S., Loriot, C.,
624 Vincelot, T., Huchet, V., Perret, B., & Thuault, D. (2017). MAP-OPT: A software for supporting
625 decision-making in the field of modified atmosphere packaging of fresh non respiring foods.
626 *Packaging Research*, *2*(1), 28–47. https://doi.org/10.1515/pacres-2017-0004

627 Guillard, V., Couvert, O., Stahl, V., Hanin, A., Denis, C., Huchet, V., Chaix, E., Loriot, C., Vincelot, T., &
628 Thuault, D. (2016). Validation of a predictive model coupling gas transfer and microbial growth
629 in fresh food packed under modified atmosphere. *Food Microbiology*, *58*, 43–55.
630 https://doi.org/10.1016/j.fm.2016.03.011

631 Henry, W. (1832). Experiments on the quantity of gases absorbed by water, at different
632 temperatures, and under different pressures. *Abstracts of the Papers Printed in the*
633 *Philosophical Transactions of the Royal Society of London*, *1*, 103–104.
634 https://doi.org/10.1098/rspl.1800.0063

Jakobsen, M., & Bertelsen, G. (2006). Solubility of carbon dioxide in fat and muscle tissue. *Journal of Muscle Foods*, *17*(1), 9–19. https://doi.org/10.1111/j.1745-4573.2006.00029.x

Jakobsen, M., Jensen, P. N., & Risbo, J. (2009). Assessment of carbon dioxide solubility coefficients for semihard cheeses: the effect of temperature and fat content. *European Food Research and Technology*, *229*(2), 287–294. https://doi.org/10.1007/s00217-009-1059-3

Lamichhane, P., Sharma, P., Kelly, A. L., Risbo, J., Rattray, F. P., & Sheehan, J. J. (2021). Solubility of carbon dioxide in renneted casein matrices: Effect of pH, salt, temperature, partial pressure, and moisture to protein ratio. *Food Chemistry*, *336*, 127625. https://doi.org/10.1016/j.foodchem.2020.127625

Lentschat, M., Buche, P., Menut, L., Guari, R., & Roche, M. (2022). Partial n-Ary relation instances on food packaging composition and permeability extracted from scientific publication tables. *Data in Brief*, *41*, 108000. https://doi.org/10.1016/j.dib.2022.108000

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, *30*, 4765–4774.

Munch, M., Buche, P., Guillard, V., & Gaucel, S. (2022). *CO2 solubility and composition data of food products annotated from the scientific litterature*. Https://Doi.Org/10.15454/4SFE64.

Pauchard, J., Flückiger, E., Bosset, J., & Blanc, B. (1980). CO2 Löslichkeit, Konzentration bei Entstehung der Löcher und Verteilung in Emmentalerkäse. *Schweizerische Milchwirtschaftliche Forschung*, 69–73.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830.

R. C. Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Rotabakk, B. T., Lekang, O. I., & Sivertsvik, M. (2007). Volumetric method to determine carbon dioxide solubility and absorption rate in foods packaged in flexible or semi rigid package. *Journal of Food Engineering*, *82*(1), 43–50. https://doi.org/10.1016/j.jfoodeng.2007.01.013

Schwartz, S. (2003). Presentation of Solubility Data: Units and Applications. In P. G. T. Fogg & J. Sangster (Eds.), *Chemicals in the Atmosphere - Solubility, Sources and Reactivity*. Brookhaven National Laboratory.

Simpson, R., Almonacid, S., & Acevedo, C. (2001). Mass transfer in Pacific Hake (Merluccius australis) packed in refrigerated modified atmosphere. *Journal of Food Process Engineering*, *24*(6), 405–421. https://doi.org/10.1111/j.1745-4530.2001.tb00551.x

Vo Thanh, H., Yasin, Q., Al-Mudhafar, W. J., & Lee, K.-K. (2022). Knowledge-based machine learning techniques for accurate prediction of CO2 storage performance in underground saline aquifers. *Applied Energy*, *314*, 118985. https://doi.org/10.1016/j.apenergy.2022.118985

Table 1: Nutritional composition information of food products used for the validation

Food product	Moisture content [1]	Proteins [2]	Salt [2]	Carbohydrates [2]	Fibers [2]	Lipids [2]
Ham	70.7%	22%	1.9%	0.6%	0%	4.8%
Salmon	66.5%	20%	0.09%	0.5%	0%	15%
Cheese	40%	27%	1.5%	0.1%	0%	27.5%
Pâté	51.2%	15%	2.2%	0.5%	1.1%	22%

[1] From the ANSES-CIQUAL French food composition table (Anses, 2019); [2] From nutrition facts label of food product.

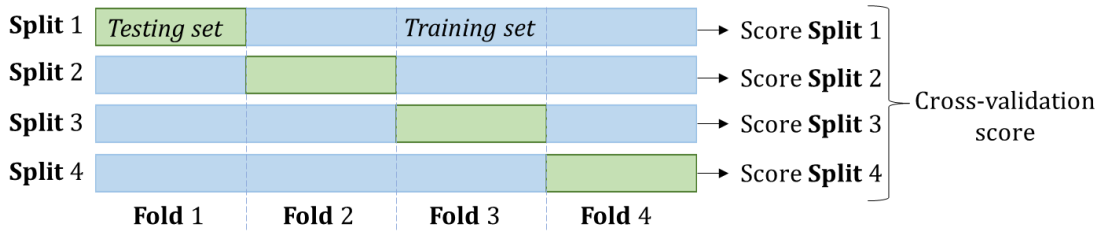


Figure 1. 4-folds cross validation. In order to evaluate the performance of a model, the dataset is separated into four folds with two sets each: the training set (used to learn a model), and the testing set (used to test the learned model).

Table 2. Performances of different models on our dataset. *Average R^2 [variance computed over 10 repetitions] (Higher=better)*

	Linear methods		Local methods		Ensemble methods	
	Linear Regression	Ridge Regression	Decision Tree	K-nearest neighbors	Gradient Boosting	Random Forest
10-folds CV	0.38 [0.03]	0.35 [0.03]	0.44 [0.04]	0.51 [0.03]	0.56 [0.17]	0.68 [0.03]

LOO	0.42 [0]	0.42 [0]	0.55 [0.02]	0.58 [0]	0.69 [0]	0.70 [0.0]

Table 3. R2 scores calculated from a 10-folds CV with a model learned from a single compositional parameter with and without the temperature. *Average R^2 [variance computed over 10 repetitions]*

	Without Temperature	With Temperature
Fat	0.32 [0.008]	0.44 [0.008]
Proteins	0.40 [0.006]	0.53 [0.01]
Water	0.35 [0.009]	0.60 [0.003]

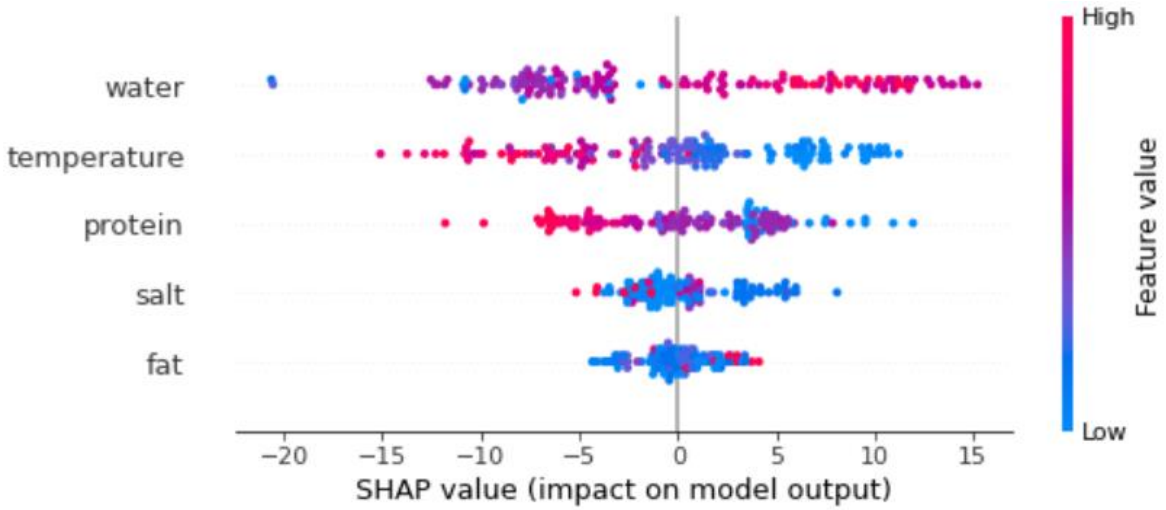
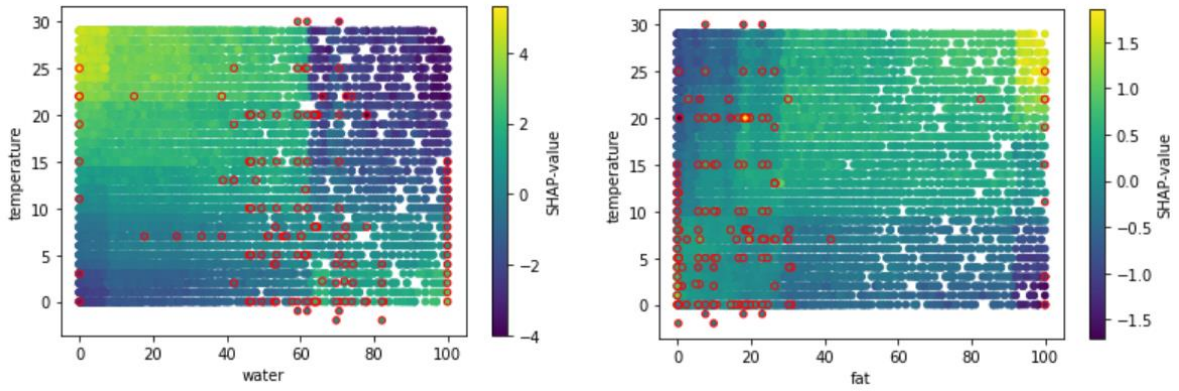
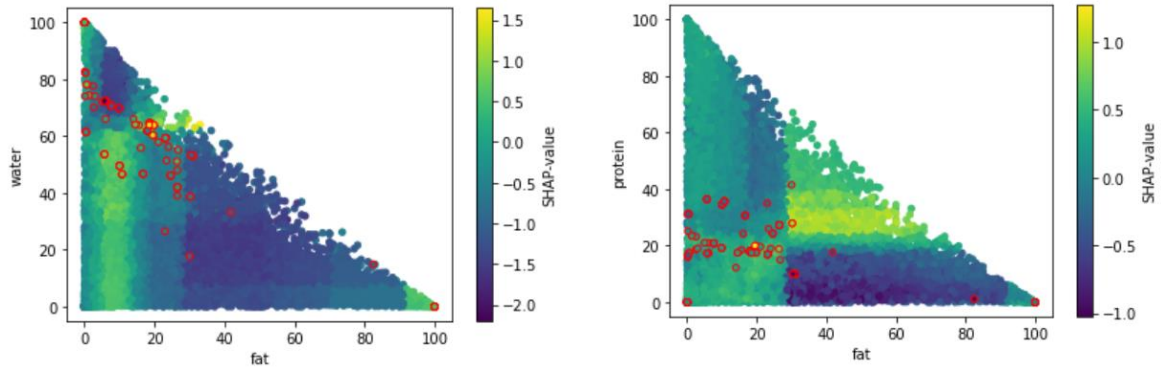


Figure 2. Variation of the SHAP value (no unit) for each feature of the model. For a given line, each dot represents a measure of our learning dataset. The SHAP value axis shows the importance of the given feature on the solubility's value's prediction. A positive SHAP value represents a positive impact (for instance, the more water there is, the higher the predicted solubility will be); on the contrary, a negative SHAP value has a negative impact (for instance, the higher the temperature is, the more it will have a negative impact on the solubility).



(a) Interaction of the temperature and water (b) Interaction of the temperature and fat

Figure 3. Interaction of the water (a) and fat (b) (expressed in %) with the temperature (expressed in °C), and their Shapley's value. Red points show data represented in the learning dataset; other points are simulated and represent how the model would infer their solubility.



(a) Interaction of the water and fat content (b) Interaction of the protein and fat content

Figure 4. Interaction of water (a) and protein (b) with fat content (expressed in %) and their Shapley's value (no unit). Red dots show data present in the learning dataset; other points are simulated and represent how the model would infer their impact on the solubility. As the sum of constituents cannot be greater than 100, we only showed physically feasible points on the graph (i.e., below the line x+y=100).

Table 4. Solubility values predicted with the machine learning model for the food case studies used in the validation approach. Intervals represent the prediction with a confidence of 90%.

Food Product	Ham	Pâté	Cheese	Salmon

CO_2 solubility (mmol.kg-1.atm-1)	55.4 [35.5;74]	42.9 [26.3;56.3]	34.7 [28.4;40.9]	54 [38.1;89.1]

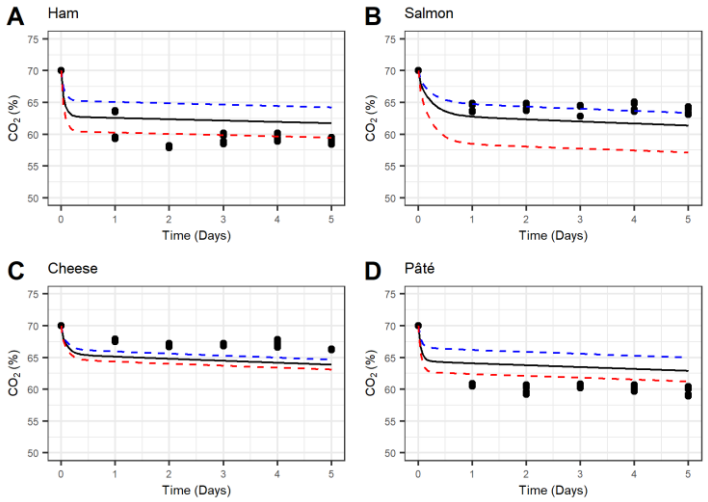


Figure 5: Impact of food composition on CO_2 concentration in the headspace. A: Ham; B: Salmon; C: Cheese; D: Pâté; dot: experimental measurement; black solid line: run with the CO_2 solubilities predicted by the machine learning model as inputs; red dashed line: model output with the upper predicted CO_2 solubilities as inputs; blue dashed line: model output with the lower predicted CO_2 solubilities as inputs.

Table 5: Fixed parameters used in simulations

Argument	Unit	Ham	Salmon	Cheese	Pâté
Tray exposed area	cm^2	260			
Lid exposed area	cm^2	167			
Food thickness	cm	0.6	1.8	1.5	1
Food surface	cm^2	165	60	80	100
Density	-	1.00	1.06	1.20	1.00

Diffusion coefficient of CO_2*	m^2/s	2.44×10^{-9}	5.5×10^{-9}	9.25×10^{-9}	7.6×10^{-9}

712 * The CO_2 diffusion coefficient of each food matrix was calculated using the linear regression $DCO_2 = 3 \times 10^{-10}$ %fat + 1×10^{-9}

713 (Chaix et al., 2014).