



Science Arts & Métiers (SAM)

is an open access repository that collects the work of Arts et Métiers Institute of Technology researchers and makes it freely available over the web where possible.

This is an author-deposited version published in: <https://sam.ensam.eu>
Handle ID: [.http://hdl.handle.net/10985/23142](http://hdl.handle.net/10985/23142)

To cite this version :

Mélanie MÜNCH, Valérie GUILLARD, Sébastien GAUCEL, Sébastien DESTERCHE, Jonathan THÉVENOT, Patrice BUCHE - Composition-based statistical model for predicting CO2 solubility in modified atmosphere packaging application - Journal of Food Engineering - Vol. 340, p.111283 - 2023

Any correspondence concerning this service should be sent to the repository

Administrator : scienceouverte@ensam.eu



22 used as input parameters in the MAP' OPT tool, predicting the evolution of headspace gas
23 composition. Predicted CO₂ solubilities used as input parameters succeeded in representing the
24 CO₂ headspace dynamic as a function of time in the four case studies.

25

26 **Keywords**

27 CO₂ solubility; machine learning models; food composition; Modified Atmosphere Packaging;

28 CO₂ headspace dynamic

29

30 **Abbreviations**

31

32 **1. Introduction**

33 In Modified Atmosphere Packaging (MAP) applications, the packaging atmosphere is generally
34 replaced by a mixture of different gases mainly composed of O₂, CO₂ and N₂ in order to prevent
35 food degradation during storage. CO₂ is often used for its bacteriostatic effect. The
36 concentration of CO₂ injected in the pack is calculated to be close to or above the minimal
37 inhibitory concentration for microorganism's growth (Farber, 1991) and this CO₂ concentration
38 must be maintained as much as possible into the packaging to keep its benefit along the food
39 shelf-life. However, CO₂ concentration varies during storage due to the CO₂ permeation from
40 the internal atmosphere toward the surrounding, and due to the solubilization and diffusion of
41 CO₂ into the food product initially free of dissolved CO₂ (Chaix et al., 2015; Guillard, Couvert,
42 et al., 2016; Simpson et al., 2001). If the loss of CO₂ due to permeation may be well mastered
43 by using high barrier packaging films (Guillard et al., 2017), the CO₂ solubilization into the
44 food is unavoidable and leads to rapid CO₂ partial pressure drop into the packaging, to an extent
45 that depends on the headspace to food volume ratio and nature of the food. The lower the
46 headspace volume is, the higher the CO₂ drop is, due to gas solubilization into the food. This

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

$$50 \quad C_{CO_2} = S_{CO_2}(T) \times p_{CO_2} \quad \text{Eq. 1}$$

51 where $S_{CO_2}(T)$, the inverse function of Henry's law coefficient, is the solubility coefficient, at
52 temperature T, expressed in mol.kg⁻¹. Atm⁻¹. It represents the maximal quantity of CO₂ that could
53 be dissolved in a product for a given partial pressure of CO₂. The value of solubility depends on
54 the nature of the food and reflects the compatibility between CO₂ and the food matrix (Chaix et
55 al., 2014; Rotabakk et al., 2007; Schwartz, 2003). The knowledge of this data is thus of
56 paramount importance to anticipate CO₂ losses by solubilization and its expected effect on food
57 shelf-life. This CO₂ solubility is an input parameter required in MAP modelling tools that permit
58 the prediction of the evolution of internal gas composition as a function of time (Chaix et al.,
59 2015; Guillard, Couvert, et al., 2016; Simpson et al., 2001) and accuracy of this data is crucial
60 for prediction's relevance.

61 CO₂ solubility is generally determined using costly and time-expensive experimental set ups:
62 nowadays there are no low-cost techniques available for a non-invasive determination of gas
63 concentration in solid matrices, which makes automatization of the measurement difficult
64 (Chaix et al., 2014). Lab made experimental set-ups are generally needed and measurement
65 requires equilibrium to be reached (24h-48h). Methodologies used to measure CO₂ solubility
66 have been reviewed by (Chaix et al., 2014). These authors also proposed a first database of
67 values that has been recently updated by (Guillard, Buche, et al., 2016) and (Munch et al.,
68 2022). In the last version, 362 solubility values were available for 81 different food products.
69 If the link between food type and value of CO₂ solubility is not straightforward, it seems
70 nevertheless that food composition (fat, water, proteins or salt content) has a strong impact. For
71 instance, CO₂ solubility was found higher in fat products than in aqueous ones: at 22°C, CO₂

72 solubility was found, respectively, 1.6 and 1.8 times more soluble in olive oil and grape seed
73 oil than in water (Pauchard et al., 1980). (Jakobsen & Bertelsen, 2006) have demonstrated that
74 there is a significant difference between the amounts of CO₂ that can be absorbed in meat with
75 different fat contents. The CO₂ absorption increases along with the increasing content of
76 unsaturated fat. CO₂ solubility was found to significantly decrease in cheese with increased
77 salinity (from 0 to 2.7% NaCl w/w) (Acerbi et al., 2016). The CO₂ solubility of renneted casein
78 matrices was found to decrease linearly with salt-in-moisture content, whereas it increased with
79 increasing pH and non-linearly varied with the moisture-to-protein ratio (Fava & Piergiovanni,
80 1992; Jakobsen et al., 2009; Lamichhane et al., 2021). In all cases, beyond compositional aspects,
81 temperature was identified as the most impacting parameter on the CO₂ solubility value with,
82 in general, a decrease of solubility with an increase in temperature. Interference between
83 temperature and physical state of lipids into the food formulation was also observed making
84 trends more difficult to interpret and formalize: for instance, in seafood model products with
85 varying lipid profile, liquid fat leads to a similar solubility of CO₂ as water, while CO₂ only
86 being minimally dissolved in solid fats (Abel et al., 2018).

87 Faced with the importance of accurate CO₂ solubility predictions and lack of low-cost and rapid
88 methods for its determination, some authors have attempted to develop empirical mathematical
89 models (mostly regressions) between CO₂ solubility and temperature and one or more
90 compositional parameters. One of the first models was proposed by (Fava & Piergiovanni, 1992)
91 and related using multiple linear regressions CO₂ solubility and compositional parameters (fat,
92 protein, moisture, pH, water activity) of different foods at one temperature (7 °C). However, if
93 this model was found suitable for meat products, it failed to predict solubility in dairy products.
94 After this preliminary attempt, a second model was proposed by (Jakobsen et al., 2009) to
95 predict CO₂ solubility in semi-hard cheese based on the weight fraction of water (w_w) and fat
96 (w_f) in the 2-phase cheese system, temperature (T), and the CO₂ solubilities in, respectively,

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

$$102 \quad S_{CO_2} = 37.92 - 0.35 T - 1.21 S/M \quad \text{Eq. 2}$$

103 However, the main drawback of all the modelling attempts mentioned above is that they are no
104 longer valid when they are extrapolated to products that were not initially included in the initial
105 range of data used for their setting up. For instance, (Chaix et al., 2014) have tested linear
106 correlations between fat and water content and CO_2 solubility determined in water, hake
107 sausage, and ham. Although well accurate for those products, these correlations failed to predict
108 solubility into fish products with an error of more than one order of magnitude (about 90%).
109 This limits their usefulness and well illustrates the difficulty of finding a simple and universal
110 linear model or correlations that are valid for large domains. In addition, it is difficult to draw
111 clear and fair conclusions about the impact of food composition on CO_2 solubility because
112 temperature often interferes with other effects, even masking them sometimes, and only one
113 class of food is examined at a time which makes it very difficult to conclude about the real
114 effect of compositional parameters. pH may also interfere by modifying the ratio of dissolved
115 CO_2 species into the food, e.g. carbonic acid, bicarbonate ion, and carbonate ion (Chaix et al.,
116 2014). This lack of generalization of state-of-the-art solubility predictive attempts is a real
117 problem to extend virtual MAP modelling tools (Guillard et al., 2017) to decision-making
118 where multiple simulations with various food products would be required.

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

122 from data. For the first part, different works in machine learning can be noted in the case of
123 solubility prediction for saline solutions (Boobier et al., 2020; Vo Thanh et al., 2022). However,
124 to the best of our knowledge, no attempt has been made yet for predicting CO₂ solubility in food
125 using machine learning algorithms. In this work, we would like to tackle this issue using
126 families of standard machine learning methods, in order to assess their performances. Our main
127 purpose is to evaluate their ability to predict the food product's solubility from the temperature
128 and composition alone. To do so, we compare three families of algorithms: linear, local and
129 ensemble methods (better described in Section 3.2.1.). While the chosen models have different
130 characteristics, they are all dedicated to the prediction of a value (in our case, the solubility)
131 given an entry vector (temperature and composition), and thus represent good candidates for
132 evaluating the ability of machine learning approaches for our problem.

133 For the second part, knowledge engineering is a sub-domain of artificial intelligence, using
134 methods and tools based on ontologies, that can be helpful to extract knowledge semi-
135 automatically (Lentschat et al., 2022), and to annotate experimental data from scientific papers
136 (Buche et al., 2013) in order to be able to realize meta-analyses. Moreover, a semi-automatic
137 mapping between ontologies dedicated to the food domain description permits to manage the
138 problem of data incompleteness, especially in terms of food product compositional parameters
139 (Buche et al., 2021). In the case of CO₂ solubility prediction, knowledge engineering could be
140 useful for structuring the different relations between the solubility and the different input
141 parameters (e.g. compositional parameters, temperature), as well as retrieving missing
142 information from other databases.

143 In this context, the aim of this paper is to present an innovative composition-based statistical
144 model of CO₂ solubility as a function of temperature (T) and compositional parameters (fat,
145 moisture, protein and salt contents). To avoid any bias due to the numerical treatment of a
146 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 CO₂ solubility was first created. Compositional
148 parameters were retrieved from the original paper or inferred using the MultiDB explorer tool²
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 CO₂ solubility
151 as a function of T and composition. Its predictions of CO₂ 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 CO₂ 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 cm³ (530 XX 00, PROMENS, Norway) and a minimal thickness
164 of 200 µm. The gas transmission rates of this tray are 3 and 13 cm³/day.atm for O₂ and CO₂
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 cm³/m².day.bar of O₂ and CO₂ 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 N₂ and 70% of CO₂. This step took

² <https://ico.iate.inra.fr/meatylab>

170 place in a laminar flow hood to avoid any microbiological contamination. The samples were
171 stored at 4 °C until analysis for 5 days. Daily monitoring of headspace CO₂ was made using a
172 Check Mate 9900 calibrated annually by the supplier (Dansensor / AMETEK, France). The
173 principle of dosing is based on an infrared sensor for CO₂.

174 **2.3. MAP'OPT: mathematical model for headspace CO₂ dynamic**

175 The mathematical model developed by (Guillard et al., 2017) was used to predict the variation
176 of the O₂ and CO₂ concentration in the headspace of packaged food products in the present shelf-
177 life experiment (i.e., ham, salmon, cheese and pâté). This semi-mechanistic model included (i)
178 O₂/CO₂/N₂ transfer between headspace and external atmosphere via permeation through the lid
179 material and the tray in contact with headspace, (ii) O₂/CO₂ sorption or desorption characterized
180 by solubilization and diffusion within the food product, (iii) variations in headspace volume
181 and composition obeying the ideal gas law while maintaining a total pressure equal to the set
182 pressure of the tray sealer and (iv) temperature effect on all the aforementioned mechanisms
183 according to Arrhenius equation. The input parameters needed to run the simulation depend on
184 the characteristics of the packaging (volume capacity, thickness of the tray and lid, exposed
185 area, gas permeation), storage (composition of the gas mixture, temperature, duration
186 preservation) and of the food product (solubilization and diffusion of gases, mass, density,
187 thickness, information on nutritional composition). The O₂ diffusivity and solubility, at 4°C,
188 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
189 CO₂ diffusivity (in m²/s), at 4°C, was estimated, for each product, according to (Chaix et al.,
190 2014) by:

$$191 \quad D_{CO_2} = 3 \times 10^{-10} \%fat + 1 \times 10^{-9} \quad \text{Eq. 3}$$

192 Valid in the range of temperature [0, 8°C], where D_{CO_2} is the diffusivity of CO₂ (m²/s) and %fat
193 is the fat content (% w/w in wet basis) of food products.

194 The CO₂ solubility for the 4 food case studies was predicted using the model developed in this
195 study and were used as input parameters for CO₂ solubility.

196 **2.4. Evaluating Statistical Models for CO₂ solubility**

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

213 All experiments were implemented using the Python library Scikit-learn (Pedregosa et al.,
214 2011), which is dedicated to machine learning. Unless otherwise stated, the library's default
215 algorithm's parameters were used. Further explanations of the results were done using the
216 Python SHAP library (Lundberg & Lee, 2017), which allows to compute the relative
217 importance of features in a prediction task using the game-theoretic notion of Shapley value.
218 The choice of this method was motivated by its agnostic aspect: as its results do not depend on

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

222 **2.5. Statistical analysis**

223 For shelf-life experiments, significant differences in headspace composition between food
224 matrices and time points were tested using the nonparametric Kruskal test with the
225 “kruskal.test” function in statistical software R 3.6.1 (R. C. Team, 2019). In case of significant
226 food matrix effect, Dunn’s test for stochastic dominance among food matrix groups was
227 computed using the function “dunn.test” of the R package “dunn.test” (Dinno, 2019) and $P <$
228 0.1 was considered as significant.

229

230 **3. Results and Discussion³**

231 **3.1. Data collection**

232 362 data from 21 references of the literature were collected and stored in a dedicated database.
233 Solubility unit kept for the following is $\text{mmol.kg}^{-1} \cdot \text{Atm}^{-1}$ for the sake of clarity. Corresponding
234 food compositions were retrieved directly from the original paper or, if not provided in the
235 source paper, retrieved from the Food Composition database (Buche et al., 2021). Four
236 constituents (water, fat, protein and salt) were kept for further analyses (sugar was discarded
237 due to many null or missing values, which would not have brought more information to the
238 model). This choice was motivated by analysis of previous literature on the topic, as fat content
239 was found particularly relevant (Jakobsen & Bertelsen, 2006; Pauchard et al., 1980). However,
240 while the lipid profile and physical state of lipids was also proved to be important, especially
241 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>

242 this approach because lipids profile was most of the time simply unknown or impossible to
243 retrieve with enough precision. On another hand, protein and moisture contents were also kept
244 because several times quoted as relevant compositional parameters influencing CO₂ solubility
245 (Lamichhane et al., 2021). More specifically, (Fava & Piergiovanni, 1992) considered fat, protein,
246 moisture, pH, water activity in their model of CO₂ solubility. In the present study, pH and water
247 activity were discarded because they are not available in the food composition database. Finally,
248 salt content was also kept as several times quoted for its impact on CO₂ solubility (Acerbi et al.,
249 2016; Duan & Sun, 2003).

250 In the end, the constituted database presents mainly three categories of food products: dairy
251 products, meat and fish. It was also complemented with measures made on water and oil. While
252 this distinction of “type” was kept for data description purposes, it was not used as a variable
253 during the learning: the composition was considered to be sufficient for predictability purposes.
254 For each food product, temperature was also kept as one of the main factors affecting CO₂
255 solubility value. Even if the temperature effect was in general well modelled using Arrhenius’
256 law (Chaix et al., 2014), it was decided in the present work to consider it as a parameter in
257 addition to composition in the statistical model and to not model its effect using Arrhenius’
258 law.

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

264

265 **3.2. Learning models / prediction of CO₂ solubilities**

266 **3.2.1. Model used, learning**

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

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

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

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

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

311 Random Forests are an ensemble method based on the learning of multiple decision trees from
312 a random sample of the whole dataset. This approach avoids the over-fitting tendencies of
313 decision trees through averaging, and proposes a better adaptability to the data's inner
314 variations. Since the number of trees has an impact over the final result, we have used another
315 10-folds CV to fine tune the parameter and find the best possible combination. We have tested
316 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 CO₂ 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 CO₂ 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 CO₂ 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 CO₂ solubility
340 observed.

341 Similarly to many black-boxes models and in contrast to using, e.g., on decision tree, random
342 forests can be hard to interpret. As predictions are based on the combination of multiple decision
343 trees, explanations are not direct as we have no clear dependency between the parameters and
344 the final result. In order to understand the role of the compositional parameters in the prediction,
345 the Shapley's value of each nutrient was computed using the SHAP Python library. Shapley's
346 values are used in game theory to express a property's contribution to a final result, considering
347 both its individual contribution as well as its marginal contribution when combined with other
348 properties (accounting for interactions): the higher it is in absolute value, the more this property
349 has influenced the final decision. In the following, we distinguish positive and negative
350 influences: in our case, the first tends to increase the CO₂'s solubility value, while the second
351 tends to decrease it.

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

359 However, most of the compositional parameters' influence cannot be characterized as easily:
360 the fat, for instance, seems to have low SHAP values, lower than those obtained for water. Thus,
361 it appears that fat might have lower effect than water on CO₂ solubility and would positively or
362 negatively impact this solubility (both positive and negative Shapley's values were observed
363 for fat), depending on other factors that are not shown in this figure and may be absent from the
364 data set. Since Table 3 has highlighted a strong interaction between compositional parameters
365 and the temperature, we display Shapley values on two axes (temperature + constituent) to

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

377 However, it is important to note that in both cases, we have represented in red the combinations
378 represented in the learning dataset. This is important, as we can see in the case of the
379 temperature/fat graph that nearly all predictions between 30 and 100% of fat are inferred, as
380 there was no product with that quantity of fat in our learning dataset (which is credible, since
381 apart from certain particular food products such as oil or butter, products with fat content above
382 30% are rather scarce). As a consequence, the model has extrapolated the result (and the
383 importance of the parameters in its prediction) from similar results, and not from concrete and
384 observed data. This could lead to false interpretations, and highlight the limit of our model in
385 its current state: while predicting solubility of items similar to the ones used for the learning
386 can be reasonably trusted, the more a food product will be remote from the original learning
387 set, the more difficult and not trust-worthy its prediction will be. Put another way, while
388 provided inferences in unexplored areas appear plausible, they should be further checked by
389 concrete experiments.

390 **3.2.3. Comparison with mechanistic models from the literature**

391 The literature well highlighted the impact of temperature on CO₂ 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 CO₂
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. CO₂
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 CO₂ 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 CO₂ 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 CO₂ 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, CO₂ 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 CO₂ solubility is governed by the fat phase

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

420 Fig.4 (b) shows interaction of the protein and fat contents and the corresponding Shapley's
421 value. It shows that for products with fat content below 30%, protein content tends to negatively
422 impact CO₂ solubility. On the contrary, above 30% of fat content, protein content positively
423 impacts solubility. In other words, below 20% of protein, increasing fat content has a slight
424 positive impact on CO₂ solubility until a threshold value of 30%. Above this threshold value of
425 30% of fat, this positive effect turns into a negative one. However, in both cases, the effect is
426 low with absolute SHAP-value below 1. In addition, for fat content higher than 30%, there are
427 only few data (red open symbols on Fig. 4 (a)) and data are thus mostly extrapolated by the
428 model and should be considered cautiously. This interaction between protein and fat contents
429 was never related in the literature. If the impact of protein contents was clearly identified on
430 CO₂ solubility, it was never clearly stated to what extent it would affect these solubility values.
431 For instance, (Jakobsen & Bertelsen, 2006) observed that CO₂ absorption increases along with the
432 increasing fat content (from 2 to 65%) into mixtures of muscle and fat (from pig meat) but they
433 did not mention the protein contents of their samples making difficult to align their study on
434 the results shown in Fig. 4 (b). Nevertheless, supposing that pig meat contains a maximum of
435 20% of proteins (from the French food composition table (Anses, 2020)), we can estimate that
436 protein content varies from 19.6% for 2% of fat content to 7% at the lowest for the fattiest
437 mixture. We are thus below the threshold value of 20% of proteins where increasing fat content
438 tends to increase CO₂ solubility into such samples (Fig. 4 (a)). Findings of (Jakobsen & Bertelsen,
439 2006) tend to confirm the prediction of our model.

440

441 The impact of proteins on CO₂ 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 CO₂
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 CO₂ solubility and moisture-to-protein ratio which is
449 ascribed to interactive effects of moisture and protein content on CO₂ 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 CO₂ 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 CO₂ headspace dynamic for the 4 food case studies**

460 Headspace CO₂ composition was followed during the shelf-life experiment (Figure 5).
461 Following the sealing, the CO₂ content decreases in the headspace over time for each of the
462 food matrices. After 5 days, CO₂ 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

465 ($P < 0.1$). CO₂ content was not different over the first 5 days for cheese and salmon packs (P
466 > 0.18).

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

481 The relatively good fitting is also confirmed by the RMSE values equal to 2.78% for ham,
482 2.09% for salmon, 2.50% for cheese and 3.26% for pâté. We obtained a low value of RMSE
483 which indicated that we can reasonably consider a validation of the gas concentration
484 prediction. Considering the multiples sources of uncertainty in the MAP OPT simulation, taken
485 together, the simulation results validate the composition-based statistical model predicted CO₂
486 solubilities developed in this study and its generic use for a wide range of products
487 conventionally packaged in MAP. The composition-based statistical model could be included
488 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 CO₂ 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 CO₂ 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 CO₂) 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 CO₂ 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 CO₂ solubility. This allowed us to build a learning base with 362 values of CO₂ solubility from
533 which different machine learning algorithms were tested in order to select a model able to
534 predict CO₂ 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 CO₂ concentrations measured on 4 different foodstuffs

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

542 The purpose of this work is to propose a novel approach to the CO² solubility prediction, using
543 classical machine learning algorithms. The interest was both its simplicity (in order to learn a
544 model, we only needed a dataset with the raw values), and the possibilities of explanation
545 provided by tools such as the SHAP values. We wanted to assess whether rather generic
546 machine learning methods were enough to tackle our problem. While we have demonstrated it,
547 it should be interesting to compare their results to more statistical approaches, such as
548 extensions of linear models. Moreover, as mentioned in Section 4, the model's prediction could
549 benefit from the addition of new products, not only from the types considered here, but also
550 from others: this should strengthen the precision of our predictions.

551

552 **Funding**

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

556

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

559

560 **Declaration of Interests**

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

563

564 **Ethics statements**

565 This work neither involves human subject nor animal experiments.

566

567 **CRedit Authors Statements**

568 Patrice Buche: Conceptualization

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

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

572 Sébastien Gaucel: Conceptualization, Methodology, Review and Editing

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

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

576

577 **References**

578

579 Abel, N., Rotabakk, B. T., Rustad, T., & Lerfall, J. (2018). The influence of lipid composition, storage
580 temperature, and modified atmospheric gas combinations on the solubility of CO₂ in a seafood
581 model product. *Journal of Food Engineering*, *216*, 151–158.
582 <https://doi.org/10.1016/j.jfoodeng.2017.08.020>

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

586 Anses. (2020). *Ciqual French food composition table*. <https://Ciqual.Anses.Fr>.

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

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

592 Buche, P., Cufi, J., Dervaux, S., Dibie, J., Ibanescu, L., Oudot, A., & Weber, M. (2021). How to Manage
593 Incompleteness of Nutritional Food Sources? *International Journal of Agricultural and*
594 *Environmental Information Systems*, *12*(4), 1–26. <https://doi.org/10.4018/IJAEIS.20211001.0a4>

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](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., Lorient, 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., Lorient, 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>

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

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

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

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

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

649 Munch, M., Buche, P., Guillard, V., & Gaucel, S. (2022). CO₂ solubility and composition data of food
650 products annotated from the scientific literature. <https://doi.org/10.15454/4SFE64>.

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

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

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

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

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

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

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

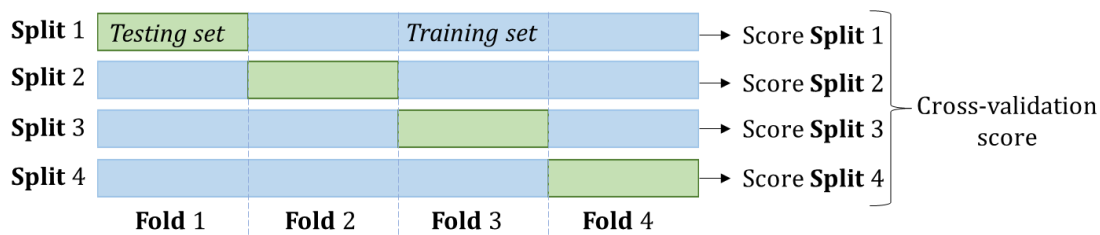
672

673

674 **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%

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



676

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

680 **Table 2.** Performances of different models on our dataset. *Average R^2 [variance computed over*
 681 *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]

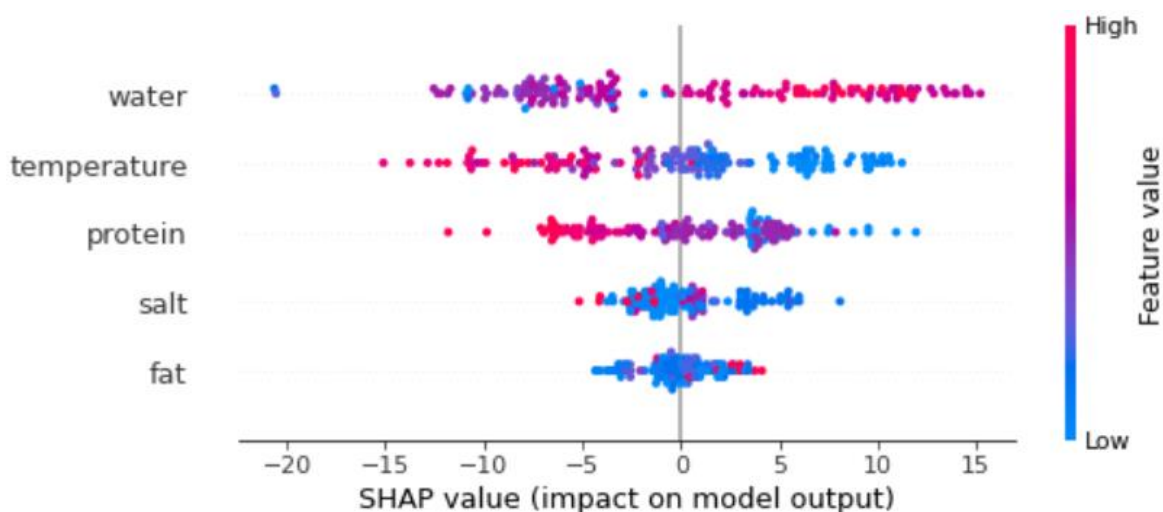
LOO	0.42 [0]	0.42 [0]	0.55 [0.02]	0.58 [0]	0.69 [0]	0.70 [0.0]
-----	----------	----------	-------------	----------	----------	------------

682

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

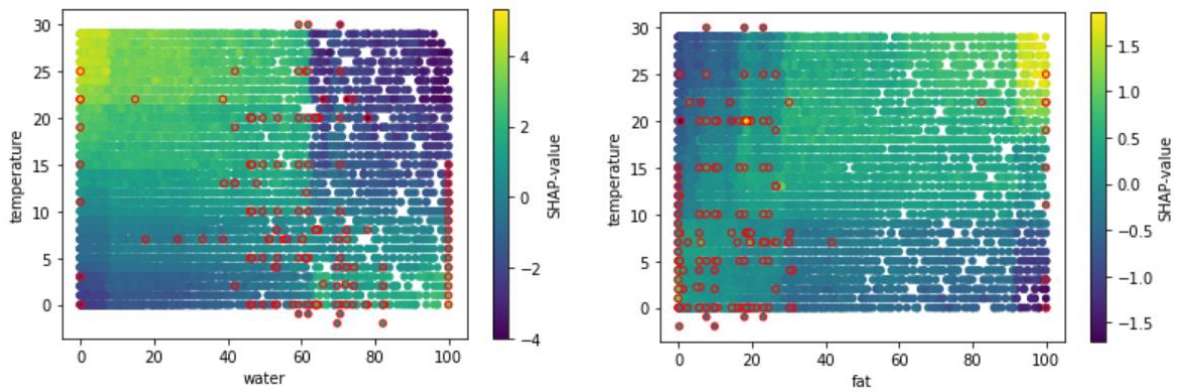
	Without Temperature	With Temperature
Fat	0.32 [0.008]	0.44 [0.008]
Proteins	0.40 [0.006]	0.53 [0.01]
Water	0.35 [0.009]	0.60 [0.003]

686



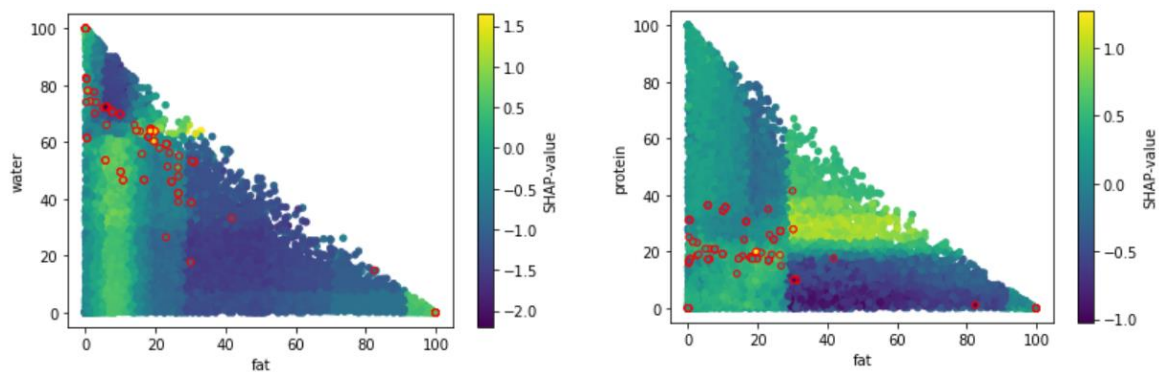
687

688 **Figure 2.** Variation of the SHAP value (no unit) for each feature of the model. For a given line,
 689 each dot represents a measure of our learning dataset. The SHAP value axis shows the
 690 importance of the given feature on the solubility's value's prediction. A positive SHAP value
 691 represents a positive impact (for instance, the more water there is, the higher the predicted
 692 solubility will be); on the contrary, a negative SHAP value has a negative impact (for instance,
 693 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

694 **Figure 3.** Interaction of the water (a) and fat (b) (expressed in %) with the temperature
 695 (expressed in °C), and their Shapley's value. Red points show data represented in the learning
 696 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

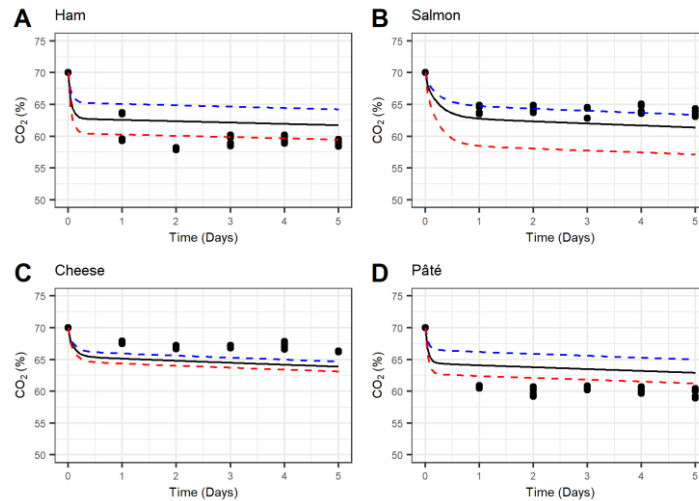
697 **Figure 4.** Interaction of water (a) and protein (b) with fat content (expressed in %)
 698 Shapley's value (no unit). Red dots show data present in the learning dataset; other points are
 699 simulated and represent how the model would infer their impact on the solubility. As the sum
 700 of constituents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

702 **Table 4.** Solubility values predicted with the machine learning model for the food case studies
 703 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]
--	-------------------	---------------------	---------------------	-------------------

704



705

706 **Figure 5:** Impact of food composition on CO₂ concentration in the headspace. A: Ham; B:
 707 Salmon; C: Cheese; D: Pâté; dot: experimental measurement; black solid line: run with the CO₂
 708 solubilities predicted by the machine learning model as inputs; red dashed line: model output
 709 with the upper predicted CO₂ solubilities as inputs; blue dashed line: model output with the
 710 lower predicted CO₂ solubilities as inputs.

711 **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 x 10 ⁻⁹	5.5 x 10 ⁻⁹	9.25 x 10 ⁻⁹	7.6 x 10 ⁻⁹
--	-------------------	-------------------------	------------------------	-------------------------	------------------------

712 * The CO₂ diffusion coefficient of each food matrix was calculated using the linear regression $DCO_2 = 3 \times 10^{-10} \% \text{ fat} + 1 \times 10^{-9}$

713 (Chaix et al., 2014).