Meta-Analysis of Vaterite Secondary Data Revealed the Synthesis Conditions for Polymorphic Control

Ara Carballo-Meilan, Lukasz Michal Starnawski, Lewis McDonald, Wanawan Pragot, Ali Nauman Saleemi[2], Waheed Afzal

School of Engineering, University of Aberdeen, King's College, Aberdeen AB24 3UE

[2] GlaxoSmithKline, Stevenage, Herts SG1 2NY, United Kingdom

Abstract

The synthesis of vaterite was investigated from a statistical point of view to identify sets of optimal experimental conditions to obtain pure anhydrous calcium carbonate polymorph. Relevant research papers in the field of the precipitation of calcium carbonate were compiled in a secondary dataset using a statistical mixed method described in another of our publications. This statistical mixed method consisted of three distinctive stages: a systematic literature review (Stage 1), followed by a meta-analysis of the acquired secondary data (Stage 2) and the validation in the laboratory (Stage 3).

In this work we present the results of Stages 2 and 3 of the mentioned method. A decision tree was built with the vaterite dataset and obtained good classification performance. A number of if-then decision rules were created covering the occurrence and absence of vaterite. The oven drying temperature, the pH and the concentration of the salt were used to control polymorphism. The best result corresponded to a vaterite polymorphic abundance of 93.6 ± 0.3%. It was possible to carry out a different investigation and arrive at new insights as a result of the unique size and characteristics of the mined data from Web of Science scientific articles.

Keywords

Supervised Learning, Decision tree, Vaterite, Meta-Analysis, Reactive Crystallization

1. Introduction

1.1 Problem Statement

The reactive crystallization of calcium carbonate ($CaCO_3$) polymorphs from the reaction between calcium and carbonate ions has been much studied. Despite the apparent simplicity of this reaction, the simultaneous and rapid occurrence of nucleation, crystal growth and other processes such as agglomeration during precipitation is a challenge for the control of the final properties of the solid. An extensive body of literature on the subject is available; but controlling polymorphism in an industrial process still remains difficult. Vaterite is the most unstable anhydrous form of $CaCO_3$ [1]; its appearance in nature is rare and its synthesis in the lab using the spontaneous precipitation method is difficult [2]. In spite of it, vaterite particles has numerous applications [1], [3], among them, it is the most important form of $CaCO_3$ applied in regenerative medicine, drug delivery and personal care products [3].

Many variables affect the precipitation characteristics of calcium carbonate (i.e. crystal habit and polymorphism). Some of them include the addition of additives like magnesium ions, initial concentration of reactants, initial pH, temperature, CO_3^{2-}/Ca^{2+} molar ratio, Mg^{2+}/Ca^{2+} molar ratio, configuration of the feed, mixing mode and contact time. Typically, a researcher would select subsets of experimental conditions from the variables (also called attributes) that are known to affect more the outcome and then carry out further experimentation in the laboratory to verify the hypothesis. This decision is mainly based on literature searchers conducted by the researcher and his or her previous professional experience.

1.2 The Statistical Mixed Method

This work is the third article of a larger study. The reader is encouraged to start with the main publication titled “Development of a Data-Driven Scientific Methodology: From Articles to Chemometric Data Products” [4]. In that paper, a statistical mixed methodology called data-driven scientific methodology (DDSM) was developed and all the stages described in detail. The first stage corresponds to a second article titled “Systematic Review using a Semi-Supervised Bibliometric Methodology for Application in a Precipitation Process”; there we discussed the process by which scientific articles were collected from Web of Science, transformed into maps and the most influential articles identified using network centrality measures and mapping techniques. Then, numerical data was compiled from these relevant documents to finally obtain the secondary dataset used in the present study.

The main objective of this work is to identify key variables at optimal ranges to control calcium carbonate polymorphism. The task was accomplished creating decision tree models with the secondary dataset. Only the case of vaterite is disclosed. This information was used to develop an adequate experimental design and setup that was tested in a real laboratory.

By comparison, the present work provides the technical details necessary to understand and reproduce the work. A summary of the main findings was included in our previous publication [4]. We have omitted a statement of the main points here to avoid an overlap between both research articles. Nonetheless, all the information is described as part of the analysis in the results section. This article also highlights new findings in the conclusions with a concrete example of how the data-driven approach shaped the experiments and assisted scientific discovery.

2 Methodology

2.1 Research Design: The Meta-Analysis

In this section, a broad perspective of the second stage of the statistical mixed methodology is provided. A sequence of steps were followed to process the secondary dataset and obtain optimal sets of experimental conditions to synthesize single phase vaterite. The steps were: data preparation, exploratory meta-analysis (EMA), subsetting the $CaCO_3$ phases from the overall secondary dataset, building the decision tree models and collating all the results to produce an array of hypothesis from the EMA study and the supervised learning algorithms. Finally, the validity of the meta-model predictions was verified with laboratory experiments. Only the case of vaterite is disclosed. The process is depicted in Figure 1. The word "Meta" indicates that the secondary data was taken from published manuscripts through a systematic literature review, and therefore uses all the relevant literature available on a subject [4]. A bibliometric technique was developed for screening thousands of papers, and identify publications likely to contain optimal experiments of vaterite. The maps obtained after this procedure corresponds to the so called Meta – Mode I in Stage 1.

The nuts and bolts of how the Meta – Model II was built, as well as, a description of the structure of the secondary dataset are provided in the next sections.

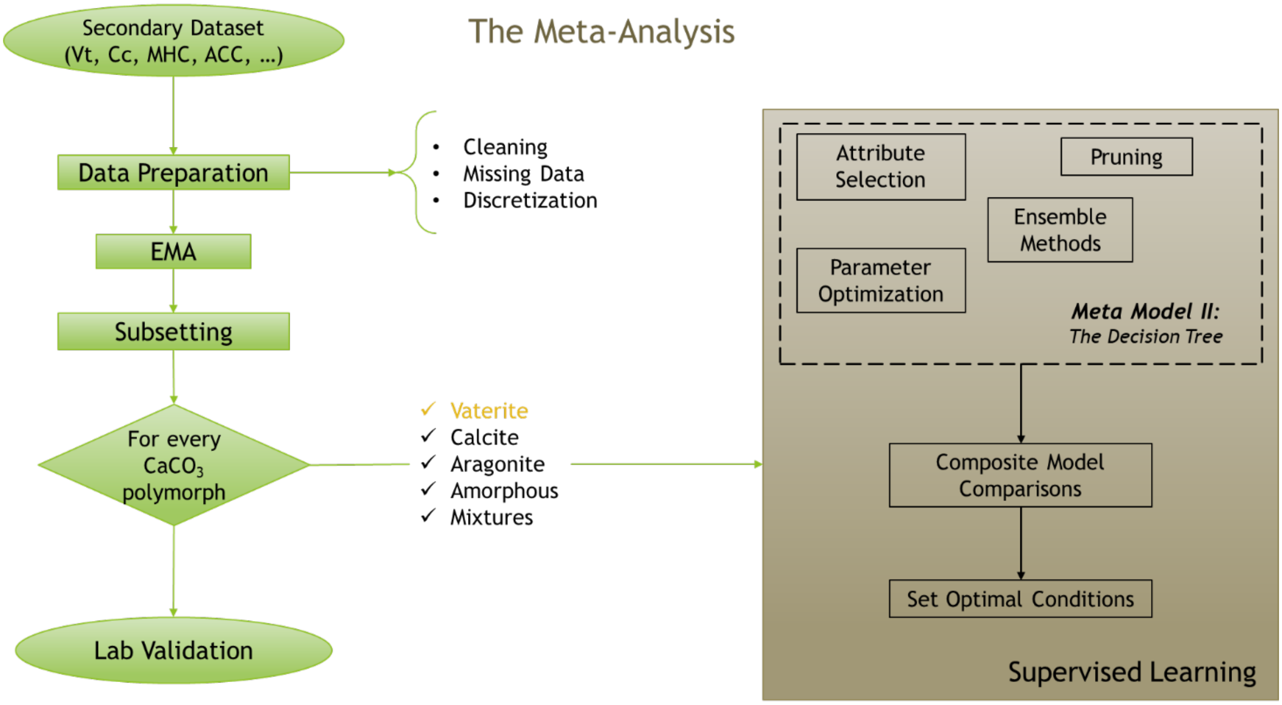

Figure 1 Flow diagram of the second part of the method: the application of secondary data for the development of synthetic routes of all the calcium carbonate polymorphs

2.2 Decision trees

A decision tree (DT) is a supervised learning algorithm used in data mining to classify cases (instances) into categories. Among them, *J4.8* algorithm – a Weka implementation of *C4.5* – is one of the most popular decision tree learners. A tree consists of a root node (the first attribute picked by the algorithm, having the greatest information gain), internal nodes (the attributes), branches (the attribute values) and leaves (the terminal nodes representing the single category or class). The goal of the algorithm is to split the root in two or more branches to produce pure subsets of data belonging to a single class. The splitting is recursive from top to bottom based on the amount of information gained by knowing the value of an attribute [5]. The algorithm computes how many bits of information are gained at each split and pick the attribute with the highest gain of information. The process stops when all nodes are pure, which in many cases occurs when the node contains just one observation. This is a common and undesirable behaviour of decision trees called overfitting. The size of the tree becomes too big and the dataset is fitted too tightly. Pruning the tree is one prerequisite to avoid fitting this noise. Pruning can be achieved in several ways, after building the tree (post-pruning) or during its construction (on-line pruning) [6], [7]. The pruning process removes unnecessary branches using threshold values that controls the size of the tree. There are other implementations such as *Random Forest* that can overcome overfitting issues.

Besides pruning, a study might include attribute selection methods for optimizing the tree. Given all the attributes under study, sometimes it is useful to select a handful of them following different criteria and then build the classifier with this small subset. This is a worthwhile approach to implement since the inclusion of irrelevant attributes is known to affect negatively the performance of data mining algorithms [8]. In this case, wrapper and filter selection methods are available and described elsewhere [8], [9].

In general, DTs offers many advantages: they are easy to read and interpret, can deal effectively with both numeric and categoric variables, as well as, missing and imbalanced data. DTs handle effectively redundant attributes and there are no a priori assumptions about the nature of the data [10], [11]. However, DTs have some disadvantages. As previously mentioned, one of the main disadvantages of decision trees is that they are prone to overfitting. Another one is instability. The output is unstable in the sense that slight changes in the training set usually lead to different attribute selections and attribute splits, producing different trees [10]. A common solution to reduce high variance is to apply ensemble methods such as bagging and boosting.

Ensemble methods for classification such as bagging (bootstrap aggregating), boosting and random forest are used for improving decision tree models. The general idea of these procedures is to produce and then combine multiple trees to yield a single prediction. Bagging reduces the variation of unstable classifiers, while boosting minimize both variance and bias [12]. Bagging is a technique that sample with replacement from the training set to randomly generate data subsets, then grows a decision tree for each bootstrap sample and combines the classifiers' predictions by voting (in the case of classification) [13]. Boosting follows a similar approach but here the subsets are created from the training set sequentially rather than randomly, giving misclassified instances from the previous tree higher preference in the next iteration. Furthermore, weights are used to give more influence to the most successful models, while in bagging all the classifiers receive equal weights [5]. *AdaBoost* is the most commonly used classification boosting algorithm and Weka uses the simpler version *AdaBoost.M1* [12]. Random forest is a meta-learner that constructs random forest by bagging ensembles of random trees using the *Random Tree* algorithm [14]. Similar to bagging, it takes random subsets of data but also random sets of predictors that then uses to grow the trees. Although combined trees have given excellent results in many fields they lack the simplicity of a single tree and are in general more difficult to interpret.

The use of these complex methods is only justified if their accuracy outperforms other more simple alternatives. In this regard, simple classification algorithms such as *OneR* and *ZeroR* can be a useful reference. *ZeroR* predicts the majority class, ignoring the predictors, and it is included in the analysis to determine the baseline performance of the rest of the classifiers. *OneR* classification algorithm creates one single rule for each variable and then pick up the rule with the smallest error rate [15].

Besides their limitations, DTs are able to solve a wide array of classification problems. For instance, among their applications can be cited citation networks [16], pharmaceutical manufacturing process [17], modelling building energy demand [18], weather forecast [19], diagnosis of diseases [8], [9], detection of forest fires [20], agriculture [21], finance [22], computer vision and many more [23].

2.3 Secondary Dataset Description

The raw vaterite dataset comprised of a total of 256 experiments. The scope of the study was limited to the spontaneous precipitation method [24] and the synthesis of single form vaterite and its mixtures with amorphous calcium carbonate (ACC), calcite and aragonite.

Overall, 56 different attributes described each of the $CaCO_3$ experiments. The complete list of attribute and their definitions are provided in the Supporting Information. The variables corresponding to the vaterite study are shown in Table 1 where each attribute name, type, range, definition and units are described.

Table 1 Dataset description (N = 256 cases, A = 23 attributes)

Attribute	Type	Range	Description
• Operational Categoric Attributes			
SynRoute	Categoric	Single-stage, Multi-stage	Experiments where the experiment was performed in two steps (Multi-stage route) or one step (Single-stage route)
Feeding	Categoric	CarbToSalt, SaltToCarb, Simultaneous	Reactant addition mode
Mixing	Categoric	Dynamic, Static	Agitation mode during precipitation (vigorous stirring versus no stirring)
• Attributes related to reactant concentration			
Volume	Numeric	0.05 –2.0	Total volume of the solution mixture (L)
Ca_M	Numeric	0.001 – 2.0	$CaCl_2$ initial concentration (mol/L)
Mg_M	Numeric	0 – 0.065	$MgCl_2$ initial concentration (mol/L)
CO3_M	Numeric	0 – 2.0	Na_2CO_3 initial concentration (mol/L)
HCO3_M	Numeric	0 – 1.0	$NaHCO_3$ initial concentration (mol/L)
Mg_Ca	Numeric	0 – 6.5	Initial ionic Mg^{2+}/Ca^{2+} molar ratio
CO3_Ca	Numeric	0.025 – 13.3	Initial ionic CO_3^{2-}/Ca^{2+} molar ratio
Mg_Pct	Numeric	0 – 87	Molar percent of Mg in the initial salt solution
• Operational Numeric attributes			
pH	Numeric	7.5 – 12.7	Initial pH
TempRe	Numeric	1 – 96	Reaction Temperature (°C)
TempOv	Numeric	25 – 105	Oven drying Temperature (°C)
time	Numeric	0.15 – 3300	Contact time (min)
• Target Attributes			
VAT, MIX	Categoric	Yes, No	Ocurrence or Non-Ocurrence of a polymorph (Vaterite and Mixtures) in the final precipitate (*Binary targets*)
FstPhase	Categoric	VAT, MIX, ACC, CAL, ARG	Appearance of a polymorph as first phase (Vaterite, Calcite, Amorphous, Aragonite and Mixtures) if polymorphic abundance of at least 85%; (*Multiclass target*)

PolType	Categoric	Hydrate, Anhydrous	Polymorph type. Crystalline nature of the polymorph. Refers to water content (*Binary target*)
PA_Cal, PA_Arg, PA_Vat, PA_ACC	Numeric	0 – 100	Polymorphic abundance (%) of calcite (PA_Cal), aragonite (PA_Arg), vaterite (PA_Vat) and amorphous (PA_ACC) (*Numeric targets*)

2.4 Stage 2: Secondary Data Analysis

The unified data obtained as a result of the repetition of the systematic review for each $CaCO_3$ polymorph was integrated by 732 experiments. The subset of the secondary dataset corresponding to vaterite experiments contained 256 experiments. The structure of this dataset was described in the previous section.

The secondary data analysis consisted of several stages as depicted in Figure 1 and explained below. The 4 major stages in the modelling process of the vaterite decision tree are shown in more detail in Figure 2 and are described in this section: data collection, data preparation, model construction and evaluation.

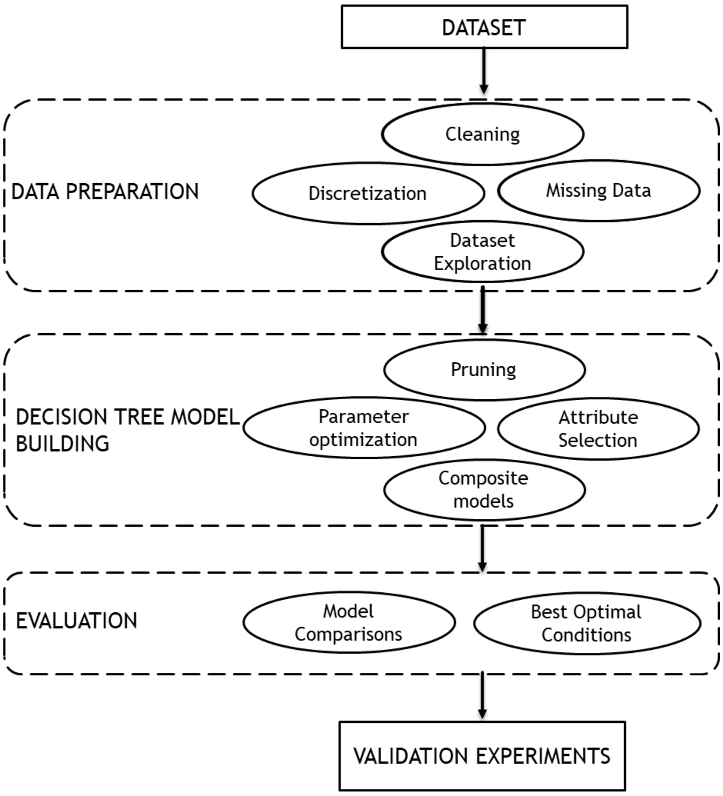

Figure 2 Flow diagram methodology for the construction of Meta Model II: The vaterite decision tree

2.4.1 Data preparation

Dataset preprocessing steps such as cleaning, data transformation, attribute selection and data exploration were used to analyse the initial dataset and prepare it for the subsequent modelling.

The analysis of missing data was performed to describe patterns of missing values, assess if missing values were random and finally decide if a missing value required a multiple imputation method. With regards to cleaning, the numerical attributes were rounded up to the nearest integer or nearest decimal. Once the dataset was collected and cleaned, new features were defined in order to use classification algorithms. Discretization was intended to construct meaningful boundaries that could explain the differences observed in the polymorphism with time. Quantile binning (same number of observations per bin) was performed to transform the numeric time attribute into a 4-class nominal attribute. A comparison between the discretized and original attribute was done.

The 230 instances forming the balanced dataset were split randomly in two groups named training/validation set (90%) and test set (10%). Data exploration was performed over the training set. The training set was also used by the learning scheme to build the classifier, the validation set was used for parameter optimization and to compare and select the best classifier. However, the final true model performance was assessed using only the test set, which was set aside from the beginner of the modelling process. The training set was balanced (same proportion of each class) and the test set also had each class well represented. Once the modelling procedure was finished and a reliable predictive power was obtained using the unbiased test set, the EDA and model were rebuilt with a whole balanced dataset ready for deployment in the Lab Validation stage. Results shown in this work correspond to the complete set of training values at this later stage.

In the data exploration stage, sample distribution analysis using bar charts, box plots and density plots was performed. The worth of each attribute was investigated following feature selection techniques. Two single-attribute evaluators were used, named *GainRatioAttributeEval* and *CorrelationAttributeEval* in Weka. The first evaluates the merit of the attribute based on gain ratio, the measure used by J48 to determine the splits and to select the most important features [9]. The second evaluates the Pearson's correlation between the predictor and the class. Both uses the Ranker method to create an ordered list of attributes, from the most to the least influential with respect to the class.

2.4.2 Modelling a Decision Tree

This section includes the construction, optimization and comparison of the following algorithms: simple classifiers such as ZeroR and OneR, J48 pruned single tree, J48 ensemble trees using bagging and boosting techniques and feature selection modelling. Model construction was done using the training/validation set containing 207 instances and 6 attributes (pH, time, [$CaCl_2$], [$MgCl_2$], TempRe, TempOv). Training dataset was balanced and contained no missing values (except for pH). The binary class target attribute VAT used for classification was formed by 2 categories: *Yes, No*; corresponding to the occurrence and the non-occurrence of vaterite precipitation.

In order to produce a decision tree with good predictive performance, parameter optimization of the J48 algorithm is often required [11]. The pruning confidence factor (-C) and the minimum number of instances in any leaf (*minNumObj* or -M) parameters in J48 were selected for the tuning procedure. The confidence threshold was used to control the complexity or size of the tree [6].-C was modified from 0.1 to 0.9 by an increment of 0.1 and -M from 1 to 10 with 10 steps. Cross-validated parameter selection (*CVParameterEval*) was the performance optimization method used in the Weka Explorer. In the case of VAT, an optimal set of parameter values was found using [C = 0.6, M = 5] for the 2-class training set.

Ensemble methods were configured as follows: The number of iterations (*numIterations*) in the algorithms was optimized in the Experimenter. *AdaBoostM1* used the following J48 weak learner configuration: -U –M2 and 3 iterations. Bagging experiment was carried out with default options. Random forest learning scheme was configured to build 10 boosted trees and the maximum depth (*maxDepth*) parameter was set to 3, corresponding to the number of attributes measured. This setting was selected based on the feature engineering analysis where at least 2 out of the 6 attributes were found relevant for the classification.

The implementation of the feature selection results into an effective classifier was done using a meta-learner called *AttributeSelectedClassifier*, using J48 as the base learner, the wrapper method (*WrapperSubsetEval*) as attribute subset evaluator (wrapping J48 for attribute selection) and *Best First* with forward direction as the search method. This approach builds the classifier selecting a smaller number of attributes based only on training set data and not in the validation set. The process was repeated 10 times in the Weka experimenter to provide a reliable estimate. J48 default options were used (-C0.25 -M2). A scheme-independent attribute subset evaluator, *CFsSubsetEval*, was as well used in conjunction with the mentioned meta-learner. In this case, the selection of the set of attributes is a function of how correlated they are with the class and how little among themselves. The same single-attribute evaluators used in the preprocessing stage were included in the comparison. In this case the number of attributes to retain was fixed to 3.

2.4.3 Evaluation of the Classifiers

The performance of the studied classifiers (ZeroR, OneR, J48 pruned, Bagging, AdaBoost, Random Forest, cost-sensitive and attribute selection schemes) was calculated using both *Acc* (accuracy or percent of correctly classified instances) and *AUC* (Area under the ROC curve) as a combined measure of the overall quality [25], [26]. Differences in *AUC* and *Acc* among classifiers were determined using stratified 10x10-fold cross validation in the Weka Experimenter and the corrected paired t-test statistic with 95% confidence level (two tailed). This corresponded to a total of 100 experimental runs per dataset and classifier. Finally, a decision list was extracted from the decision trees and interpreted in the context of a precipitation experiment.

2.5 Stage 3 - Laboratory Validation

2.5.1 Design of experiments

Full factorial design was adopted to study the simultaneous effect of pH, salt content (M) and the oven drying temperature (°C). The treatment objective was to achieve vaterite single phase. A total of 11 experiments (also called runs) were performed by designing a full factorial with 3 centre points, 3 factors and no replicates. All terms were free from aliasing, including main effects and 2-way interactions. By default, all experiments were randomized to reduce the effect of experimental bias. The independent variables (also called factors) were the pH, oven temperature (°C) and ratio CO3/Ca (M). Their levels low (-1), middle (0) and high (1) are the following: pH (8.7 – 9.3 – 10.0), oven temperature (30 – 40 – 50 °C), and CO3/Ca (3 – 6 – 9). The polymorphic abundance of vaterite R_{VAT} (0 – 100 %) was set as the main response. The following formula was used to compare the results from the different runs

$$R_{VAT} = \frac{VAT}{(Total - Halite)} \cdot 100 \qquad (1)$$

where VAT, $Halite$ and $Total$ represent the vaterite ($CaCO_3$), sodium chloride (NaCl) and total content of a fully dried precipitate using XRD quantitative phase analysis and the Rietveld multiphase refinement method.

2.5.2 Run description

The experiment took place in a 2 L borosilicate glass reactor at a fixed temperature of 19 °C. The concentration of the $CaCl_2$ salt solution was prepared using $CaCl_2$ flakes (purity 77%). Both the $CaCl_2$ salt concentration and the concentration of the carbonate/bicarbonate solution (Na_2CO_3/$NaHCO_3$)

were modified based on the run conditions. The $CaCl_2$ concentration (0.9 L) ranged from 0.11 to 0.33 M depending on the CO3/Ca values. The carbonate solution (0.9 L) was prepared using a sodium carbonate/bicarbonate mixture with different molar ratios as described by the pH value (5, 25 and 55 % Na_2CO_3). The total carbonate concentration (Na_2CO_3 + $NaHCO_3$) was kept constant and equal to 1 mol/L. The experimental methodology included the adjustment of the pH of the salt solution using NaOH (6M). The amount of NaOH necessary to raise the pH depended on both the required initial pH and salt concentration. Thus, the following reactant concentrations were used in the experimental design: at low level CO3/Ca = 3, [$CaCl_2$] = 0.33 M, pH = 8.7, 5 %molar Na_2CO_3, [Na_2CO_3] = 0.05 M, [$NaHCO_3$] = 0.95 M; at middle level CO3/Ca = 6, [$CaCl_2$] = 0.17 M, pH = 9.3, 25 %molar Na_2CO_3, [Na_2CO_3] = 0.25 M, [$NaHCO_3$] = 0.75 M; and at high level CO3/Ca = 9, [$CaCl_2$] = 0.11 M, pH = 10.0, 55 %molar Na_2CO_3, [Na_2CO_3] = 0.55 M, [$NaHCO_3$] = 0.45 M.

Both, carbonate and salt solutions were simultaneously added to the reactor at a constant rate of 400 rpm using two 323Du Watson Marlow pumps. Vigorous stirring was provided during the duration of the run. After a contact time of 4 min, the solids from the reactive suspension were quenched by vacuum filtration and washed with water several times. Then, they were immediately dried overnight in an oven at different temperatures. A Memmert's universal oven UF110 was used. Air circulation inside the oven was constant by fixing the fan setting to 10% and air flap to 100%. The rate of water evaporation from the sample was a function of the fan settings and was seen to have an effect on the distribution of polymorphs.

The procedure is depicted in Figure 4. Out of the four steps involved (preparation of solutions, precipitation, physical separation and drying), the separation step was the one that introduced more uncertainty in the measurements. Unlike the other three steps, where all the variables involved were well controlled, the separation was not so meticulously supervised. Potential sources of error coming from the filtration and washing step included the unequal overall filtration times and the unequal thickness of the cake relative to the volume of water added during washing. These parameters varied in an undetermined and uncontrollable manner. The variability created by this stage affected the amount of NaCl extracted from the solid. This inequality was reflected in the halite content of the centre points of the experimental design. The elimination of the NaCl contribution determined by XRD decreased the error variance and, hence, the power of the experimental design increased. Qualitative and quantitative phase analysis was done using X-ray diffraction (XRD) in a Panalytical X'Pert Powder diffractometer and the Rietveld multiphase refinement method to determine phase abundance.

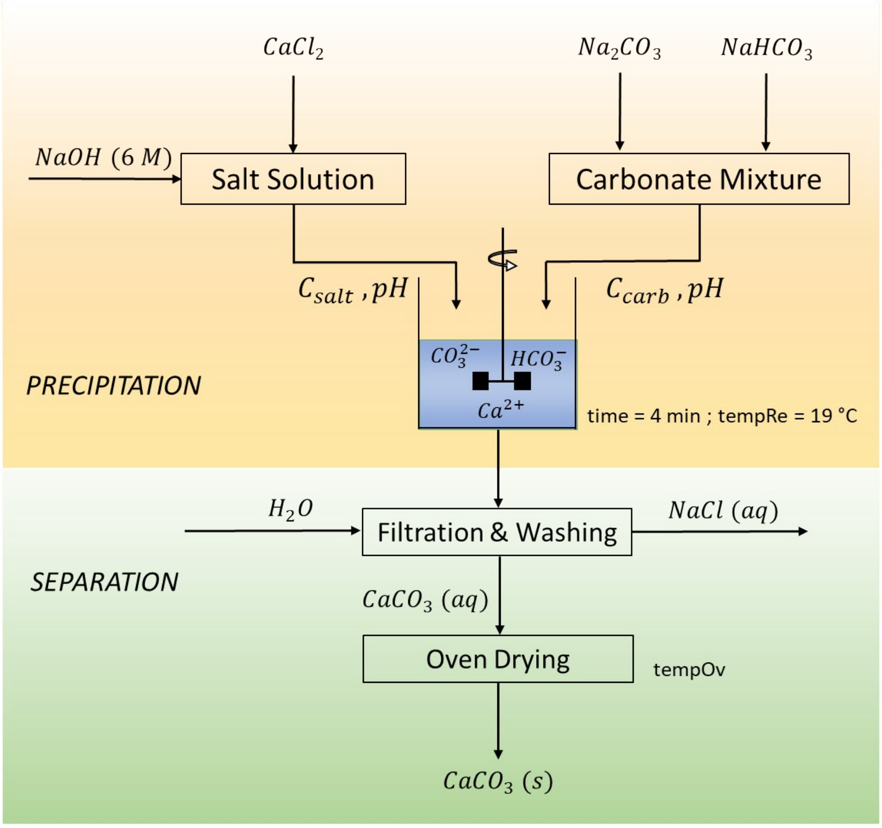

Figure 3 Lab experimental set up in the synthesis of vaterite $CaCO_3$ polymorph

2.5.3 X-ray Diffraction Analysis

X-ray diffraction was carried out using a Malvern Panalytical XPert Powder Diffractometer. The samples were placed into sample holders prepared such that a smooth powder surface was produced. The samples were then placed into the diffractometer where they were subject x-rays. The x-rays were produced from a copper radiation source with Kα wavelength of 1.54 ˚A. The angle between radiation source and detector continually increased with time from 5◦2ϴ to 60◦2ϴ. Analysis of the resulting patterns was conducted using the HighScore Plus which allowed for phase identification and Rietveld refinement of the XRD diffractograms. To prepare the diffractograms for Rietveld refinement, the background noise was removed in HighScore Plus, all phases were identified by matching each peak to a dataset from the Open Crystalography Database that had a high match score in the software. The Rietveld refinement was then run, which uses a least squares method to quantify the contribution from each dataset to the provided diffractogram and rank the contribution of each to the peaks.

2.6 Software

Data preprocessing was performed in IBM SPSS Statistics version 24 (missing data analysis), Minitab 17.1.0 and Rattle version 5.1.0, a free graphical interface for data science with R (data exploration, discretization and design of experiments). Waikato analysis for knowledge environment (Weka version 3.8.1) [27] was used as data mining software to assist the decision tree model construction and evaluation process.

3 Results

3.1 Secondary Data Analysis

The main idea behind the meta-analysis was to describe under which experimental conditions a researcher is most likely to find a particular polymorph such as vaterite after the reactive crystallization process. Furthermore, the meta-analysis was as well used to indicate which of the studied parameters were more relevant for the classification, and therefore able to play a greater role during precipitation.

3.1.1 Preprocessing & EMA

This section describe the application of the previous steps to the modelling process, including discretization, missing values treatment and data exploration. An attribute selection was included here as exploratory tool but is also part of the modelling stage.

- Discretization

A categorical attribute with 4 groups was created using the numeric time attribute, representing the precipitation contact time. The following 4 classes were built during the time discretization using quantile binning (Bins [min]): *low* (0.07 – 30), *medium* (30 – 120) *high* (120 – 720) and *very high* (720 – 3300). In general, attribute transformations can be accomplished in different ways: normalization (standardize), discretization, principal components, among others [5]. With regards to *discretization*, the transformation of a numerical attribute into categories can be done in two main ways: using equal-width bins and using equal-frequency binning. Overall, the best approach isn't obvious since discretization is data dependent, so the most suitable discretization technique was determined experimentally. Although some information might be lost during the discretization process, binning is useful in that it helps to simplify the models [28]. To test the effect of discretization the classification was performed with the original and discretized time attribute. Better classification results were obtained with the original attribute (data not shown).

- Missing data

We followed the missing data methodology described in the help manual of IBM SPSS Statistics software but the analysis was not included in the article. The following numeric attributes were discarded prior to the modelling as more than 50% of their values were missing: Size (nm), Rate (ml/min), Yield (%), amount of Mg in the polymorph (%). Regarding the variables under study in the vaterite dataset, there were 18 variables with no missing values (FstPhase, PolType, SynRoute, Feeding, Ca_M, Mg_M, CO3_M, HCO3_M, volume, Mg_Ca, Mg_Pct, CO3_Ca, tempRe, tempOv, time, mixing, VAT, MIX) and only pH was included in the analysis and had missing values. The analysis of missing data was performed to describe patterns of missing values using a tabulated patterns table, assess if the values were missing at random (Little's MCAR test) and finally decide if a missing values multiple imputation method was required. The pattern of incomplete pH data was analysed to determine if there was randomness in the way the data was missing. There was no systematic difference between the instances with missing and nonmissing observations. No multiple imputation was applied. The pH of the initial solution is by far the most important operational variable in the control of $CaCO_3$ polymorphism. This is a statement that is not demonstrated in this paper as it focuses only on the vaterite dataset rather than in the bulk of the compiled cases. For this reason, the pH was included in all the studies despite the fact that it contained a substantial amount of missing values (55 %).

- Exploratory Meta-Analysis

This section describes some of the information contained in the dataset using descriptive statistics. The box plot distribution of several attributes by the VAT class values is shown in Figure 6 and the density plot distribution of each predictor by the target in Figure 7. This last figure corresponds to a histogram that used kernel smoothing to flatten the noise. The distribution of the categorical attribute – the feeding (i.e. the order of addition of the reactants) – was performed using bar charts (Figure 5). Some cases were identified as outliers at this early stage and deleted from the dataset (e.g. time > 1440 min).

Once the dataset was built something became apparent; the number of experiments where vaterite was synthesized in single form was much lower than the number of experiments where vaterite was present in the final product as part of a mixture. This observation was true for all the $CaCO_3$ polymorphs. The identification of sets of conditions where mixtures occurred was considered relevant because in order to synthesize pure phases, regions where mixtures occur more frequently should be avoided. A typical mixture in the vaterite dataset contained a combination of the following phases in different proportions: vaterite, calcite, aragonite and ACC. Its relative quantity depended on the initial conditions of the independent variables. The inorganic synthesis of vaterite was mainly performed in the absence of magnesium and the importance of temperature as a means to obtain purity was always highlighted in most of the documents analysed. However, vaterite was also present in the composition of mixtures whenever a salt solution contained magnesium.

The most common feeding configuration was the simultaneous addition of the salt and carbonate solutions under vigorous stirring. The occurrence of vaterite, aragonite, ACC and their mixtures was seen in these three feeding modes. Based on Figure 5, the addition of the carbonate on the salt solution (CarbToSalt) could favour the appearance of mixtures as compared with the other two feeding modes and therefore be detrimental to the synthesis of single phases. Vaterite was the $CaCO_3$ phase less likely to occur, being found only in 18% of all the collected experiments. Conversely, the phase more common in the final precipitate was calcite and mixtures occurring in 54% and 43% of the cases, respectively. The presence of mixtures in the final product is a widespread issue.

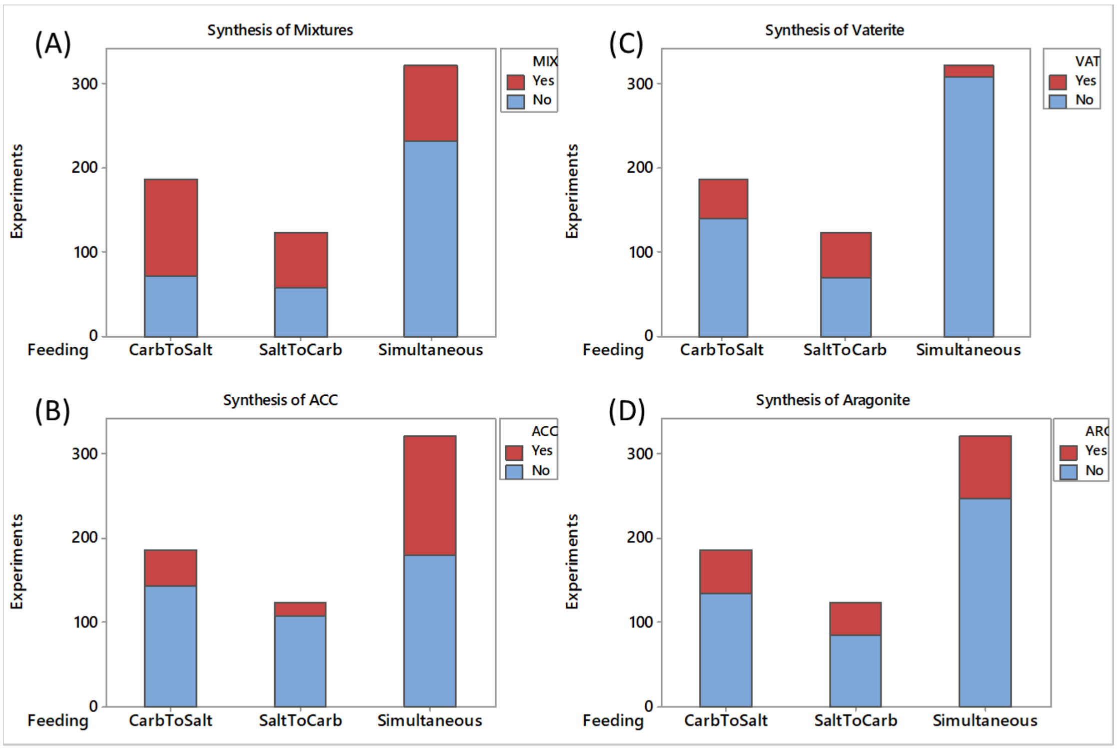

Figure 4 Bar chart of the occurrence and absence of vaterite, aragonite, ACC and their mixtures in the final precipitate as a function of the different feeding combinations.

The distribution of the binary target attribute: VAT describing the occurrence (Yes) and non-occurrence (No) of vaterite was analysed. The median of the *Yes* class in the $CaCl_2$ distribution was 0.1 M, a value not statistically significant from the *No* class (Figure 6 – A). Experimentalists obtained a higher number of positives cases when the $CaCl_2$ salt concentration increased. Most of the experiments were carried out at $CaCl_2 < 0.5$ M and CO3/Ca = 1.0 (Figure 7 – A). However, the Yes class happened more often at values of CO3/Ca lower than 1.0 (62% of the cases found below 1.0 corresponded to the class Yes, as opposed to, less than half of the cases were positives when the value was set to 1.0). Regarding the oven drying temperature in Figure 6 – B, the appearance of vaterite was seen at both high and low oven drying temperatures, and the median for the occurrence of vaterite was 50 °C. The reaction temperature most commonly used for experimentation within the compiled cases was 25 °C (Figure 7 – B). Both, the occurrence and non-occurrence of VAT happened at this setting. The median Mg (% molar) for the occurrence of VAT was significantly different from the median of its non-occurrence. The direction of this difference, within the compiled cases, indicates that researchers are more likely to find VAT as precipitate in the absence of magnesium. In the case of the pH, the median of the Yes class was 10.5 and most of the No values were seen at a pH value around 9.0 (Figure 7 – C). Value of pH above 10 look good because the Yes class was found more often and the No class did not happened.

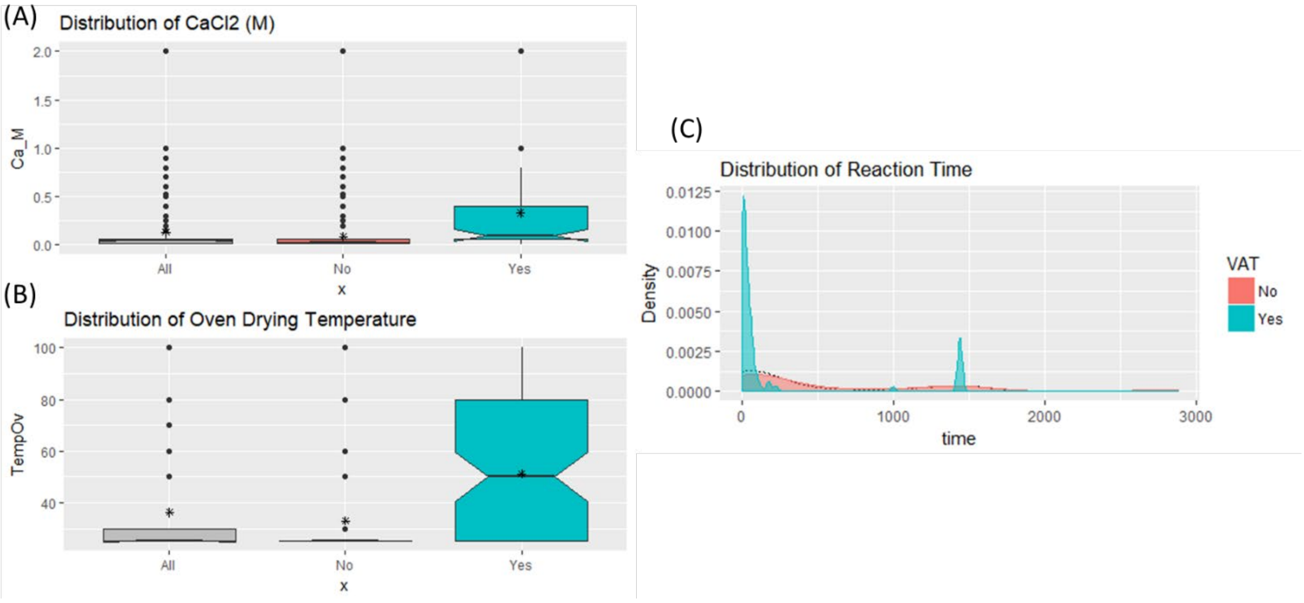

Figure 5 Box plots of the distribution of (A) $CaCl_2$ (M), and (B) tempOv; and density plot of the distribution of (C) time by the occurrence (Yes) and the non-occurrence (No) of the VAT polymorph

Most experiments were carried out at contact times lower than 60 min (71% of the dataset). The appearance of vaterite was seen more often in the precipitation experiments performed under an hour as compared with longer runs (Figure 6 – C).

The distribution of the numerical attributes by the multiclass target attribute: FstPhase was also considered in the exploratory analysis, although the plots are not displayed. We found that, on average, VAT was found in single form (considering purities higher than 85% or the only phase identified by the researchers) using low molar ratio CO3/Ca (median CO3_Ca = 1.0), low Mg molar content (median Mg_Pct = 0 %), high pH (median pH = 11) and high tempOv (median tempOv = 50 °C). Both aragonite and vaterite showed the lowest CO3/Ca and Mg (%) median values, and highest pH and tempOv median values when compared with the other phases (Calcite, ACC and Mixtures). Comparatively single phase vaterite experiments were carried out using more concentrated $CaCl_2$ salt solutions (the median was 0.75 M) than for the synthesis of the other single phases. These results are in agreement with the median value obtained with the constructed binary target attribute VAT where the centre of CO3/Ca was 1.0, Mg content was 0 % and pH median was 10.5. Another distinctive characteristic of the vaterite synthesis (also seen in the case of the aragonite single phase) was higher temperature conditions during reaction and/or higher oven drying temperatures. The oven drying temperature required for aragonite synthesis was higher (median tempOv = 80 °C) than for vaterite precipitation (median tempOv = 50 °C).

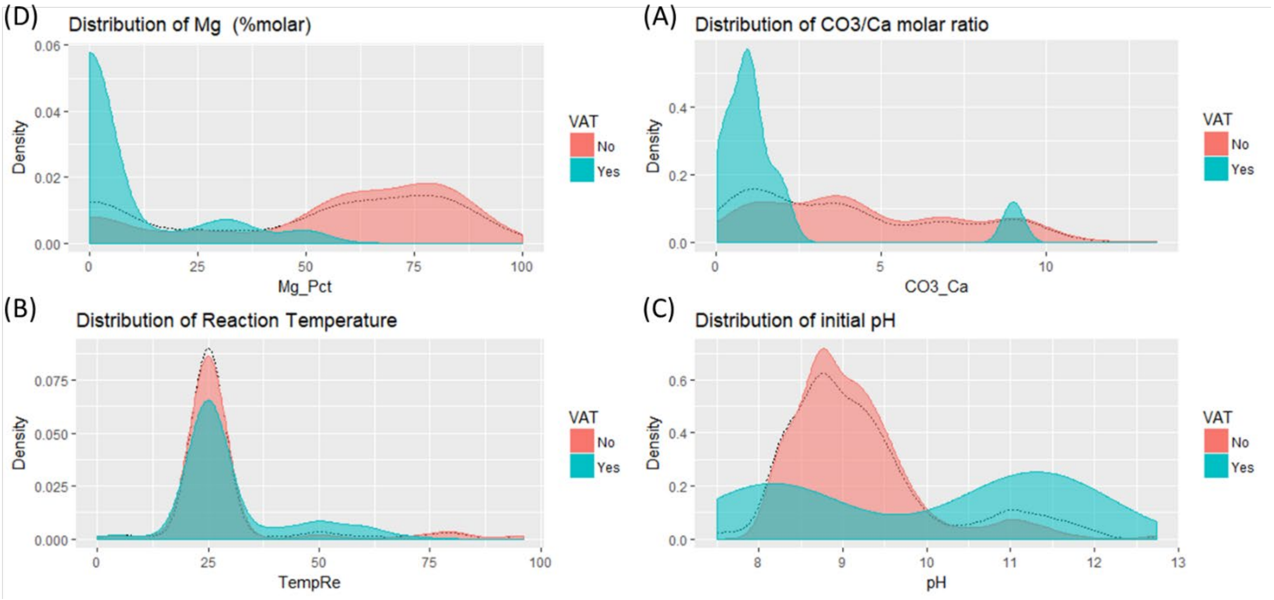

Figure 6 Density plots of the distribution of (A) $Na_2CO_3/CaCl_2$ (molar ratio), (B) TempRe, (C) pH and (D) Mg (% molar) for the two target levels

It could be concluded from this section that an optimal set of experimental conditions for the synthesis of VAT would be selecting a contact time lower than 60 min, preparing a salt solution with no Mg content and a $CaCl_2$ solution of 0.1 M, performing the reaction at ambient temperature and an initial pH of at least 10.0. Additionally, setting the oven drying temperature higher than 25 °C (median was 50 °C) would also aid the production of this anhydrous polymorph. This conclusion was included in the summary of results we provided in a previously published work [4].

- Attribute Selection

A sorted list of the best attributes for the class VAT was created using attribute evaluators (Table 2). At the top of the list, the concentration of $CaCl_2$ and the reaction temperature were the two single attributes more correlated with the class in this dataset. The metric gain ratio is the measure used by J48 to determine the splits and to select the most important features in the classification. Its value indicates the amount of information gained by selecting the attribute for the classification. In this case, values equal to 0 mean no information was gained and values close to 1 indicate that the attribute contained a high amount of information relevant for the classification. Overall, it seems that $CaCl_2$ (M), tempRe and TempOv are the 3 most relevant attributes affecting the occurrence of VAT. Probably time could be considered as well since it appeared in 9 out of 10-fold using the scheme-dependent attribute subset evaluator. The least relevant attributes were $MgCl_2$ and pH, this last attribute with a high amount of missing values.

Table 2 Ranked list of attributes based on correlations (top left) and gain ratio (top right) calculations. Number of times the attribute appears in the subsets using attribute subset evaluators (bottom row). Select attribute panel in Weka and the 2-class dataset (Test: 10-fold cross-validation) was used for these experiments.

Evaluator: CorrelationAttributeEval Scheme: Ranker			Evaluator:GainRatioAttributeEval Scheme: Ranker		
Pearson's Corr.	Avg rank	Attribute	Gain Ratio	Avg rank	Attribute
0.259 ± 0.016	1 ± 0	tempRe	0.152 ± 0.021	1 ± 0	tempRe
0.185 ± 0.02	2 ± 0	$CaCl_2$ (M)	0.104 ± 0.014	2 ± 0	$CaCl_2$ (M)
0.123 ± 0.022	3.1 ± 0.3	tempOv	0.004 ± 0.004	3.1 ± 0.3	tempOv
0.101 ± 0.018	3.9 ± 0.3	time	0	4.1 ± 0.3	pH

0.053 ± 0.021	5.1 ± 0.3	$MgCl_2$ (M)	0	5.1 ± 0.3	$MgCl_2$ (M)
0.017 ± 0.015	5.9 ± 0.3	pH	0.006 ± 0.018	5.7 ± 0.9	time
WrapperSubsetEval (J48 -C0.25 -M2) Search: Best First (Forward direction)			CfsSubsetEval Search: Best First (Forward direction)		
Number of folds		Attribute	Number of folds		Attribute
10 (100%)		$CaCl_2$ (M)	10 (100%)		$CaCl_2$ (M)
10 (100%)		tempRe	10 (100%)		tempRe
9 (90%)		time	1 (10%)		time
4 (40%)		pH	0 (0%)		$MgCl_2$ (M)
3 (30%)		tempOv	0 (0%)		tempOv
2 (20%)		$MgCl_2$ (M)	0 (0%)		pH

3.1.2 Modelling & Evaluation

- Simple classifiers

ZeroR predicted the class value *Yes* with a success rate of 48.3 ± 1.1 % in the binary dataset VAT (Figure 9). This performance value can be considered as the model baseline. Any classifier built with this dataset should perform significantly better than the baseline in order to be considered useful [5]. OneR can be considered a 1-level decision tree [29]. The 1-rule classifier has one parameter called minimum bucket size (*minBucketSize*) that controls the discretization of the numeric attributes and thus the complexity of the rule to avoid overfitting. It indicates the minimum number of cases in a bucket. This means that when this parameter increases, the splits of the attribute are reduced and the rules are simplified. On the contrary, the lower the *minBucketSize* is, the higher becomes the accuracy and complexity of the rule. In Weka, the *minBucketSize* is also referred as -B in the corresponding configuration window. Its value was optimized using the cross-validated parameter selection (*CVParameterEval*). The attribute with the highest success rate was $CaCl_2$, thus the one chosen by OneR learning scheme to produce the single rule: IF ($CaCl_2 \geq 0.029$ M) THEN VAT = Yes ELSE VAT = No.

- J48 Decision Tree & Ensemble Methods

Figure 8 shows the J48 decision tree drawn by Weka after training the model. The run description and confusion matrix of this single experiment were added on Supporting Information. The best single predictor to start classification was the reaction temperature. Early nodes were also formed by the calcium salt concentration and the time. Decision trees can be read as If-Then decision rules by following the path from the root node to the leaves in every branch [18]. The confidence of each node is indicated by the number of correctly classified instances. The number of correctly classified instances at that leaf is indicated in parenthesis. As it can be read below in *Rule 1*, there were 73 false positive observations among the 244 cases (repeated instances are included in the boostrapped tree) covered by the rule: "IF (*Ca* > 0.026 M) THEN VAT = Yes". This rule was able to classify correctly a total of 171 instances. The higher the number of cases correctly classified in the node, the more confident we can be of the given decision. This means that for a rule to be considered better than others has to cover as many cases as possible of the class it is defining. To extract the decision list from the trees (J48 pruned and ensemble classifiers), all the rules for a single class were compiled and then each subset was ranked by its success (from higher to lower accuracy). Some of the rules with a support greater than 30 correctly classified instances are shown below in descending order, from the most to the least successful rule. Rules with higher error rate than the ones shown below were omitted although they also contributed to the whole classification rate (e.g. IF (*tempRe* > 70 °C) THEN VAT = No (11)). Some of these trees also contained duplicate rules.

IF (*Ca* > 0.026 M) THEN VAT = Yes (171/73) *Rule 1*

IF (*tempRe* ≤ 60 °C) AND (*time* < 100 min) AND (Mg/Ca ≤ 0.125) THEN VAT = Yes (95/25) *Rule 2*

IF (20 < *tempRe* ≤ 40 °C) AND (0.026 < *Ca* ≤ 0.6 M) THEN VAT = Yes (86/18) *Rule 3*

IF (19 < *tempRe* ≤ 40 °C) AND (Mg/*Ca* ≤ 0) THEN VAT = Yes (86/24) *Rule 4*

IF (*time* > 120 min) THEN VAT = No (85/11) *Rule 5*

IF (20 < *tempRe* ≤ 60 °C) AND (*time* ≤ 100 min) AND (pH > 10.2) THEN VAT = Yes (64/18) *Rule 6*

IF (*tempRe* ≤ 60 °C) AND (*time* > 100 min) AND (CO3/Ca > 0.845) THEN VAT = No (36/7) *Rule 7*

Some of them are very generic (e.g. IF (*Ca* > 0.026 M) THEN VAT = Yes (171/73)). More specific rules from the AdaBoost classifier are listed below. They include more attributes and cover less number of cases.

IF (19 < *tempRe* ≤ 70 °C) AND (2 < *time* ≤ 120 min) AND (Ca > 0.016 M) AND (pH > 10.7) *Rule 8*
THEN VAT = Yes (33/2)

IF (19 < tempRe ≤ 25 °C) AND (time > 4 min) AND (Ca > 0.067 M) AND (Mg ≤ 0.065 M) *Rule 9*
THEN VAT = Yes (36/3)

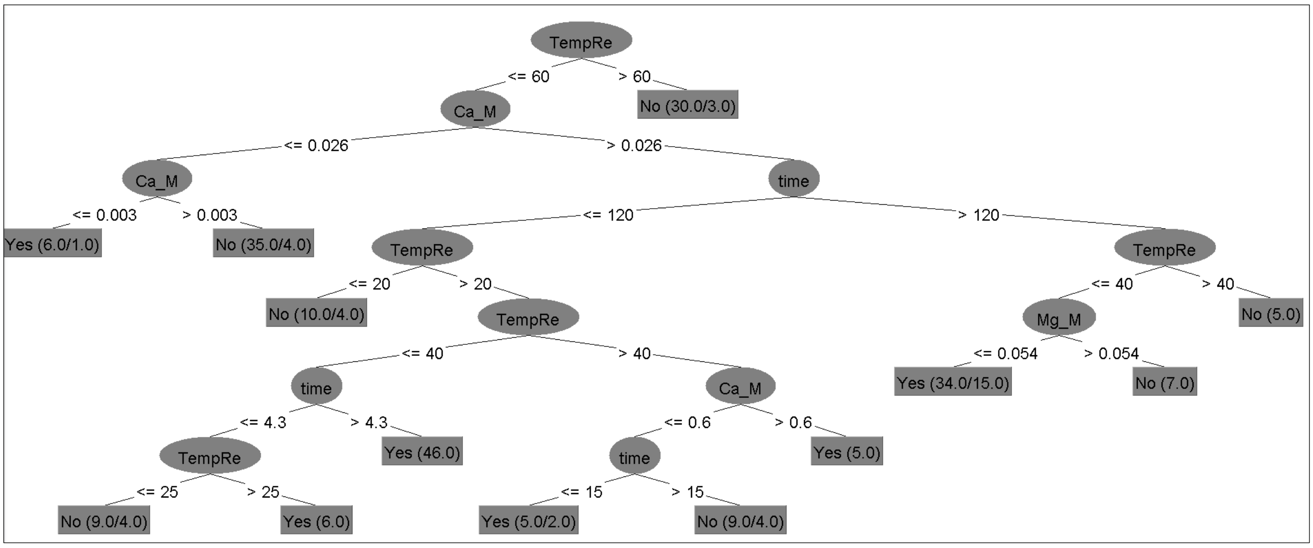

Figure 7 Decision tree of Vaterite (J48 pruned C0.6 M5, size of the tree: 25, number of leaves: 13) with 163 correctly classified instances (78.7% accuracy) as shown by the Weka Explorer (single experiment)

After training the model, the classifiers were evaluated using a holdout method and 10-fold cross-validation. The overall model classification performace was measured in terms of its *Acc* (accuracy or percent of correctly classified instances) and *AUC* (Area under the ROC curve). The larger is this area, the better is the model [5]. In general, an ideal prediction has *AUC* values around 1, while a random decision will show an *AUC* of 0.5. The classifiers with the best performance were those having simultaneously high accuracy and high *AUC*. The paired t-test showed that the differences in *Acc* and *AUC* between the simple classifiers (OneR, ZeroR) and J48 were significant. ZeroR was significantly worse than J48 and the rest of the classifiers in this dataset at the 95% confidence level. For instance, results from the Weka Experimenter indicated that the J48 pruned tree had an average accuracy rate of 73.8 ± 8.7% (10 iterations), value significantly better than ZeroR (48.3 ± 1.1%) and OneR (64.9 ± 8.6%) at the 95% confidence interval. In terms of accuracy, the model showed a significant improvement in the *AUC* values using stratified 10x10-fold cross validation (Figure 9).

Based on Figure 9, the metalearners (boosting, bagging and random forest) outperformed J48 and all the other classifiers. They showed the greatest accuracy and largest *AUC* in all sets: the validation, test and lab sets. The prediction on the lab test set was good. Some models such as J48 and CfsAttributeEval performed well in the validation and test sets but failed to predict the outcome of our laboratory experiments. Having a single well-performing tree is a more advantageous result as it is easier to interpret than an ensemble of them. Random forest and bagging are less interpretable but the results from the AdaBoost classifier can be understood to some extent because the classifier consisted of just 3 decision trees (some of the rules were shown above). The excellent performance of the AdaBoost metalearner in this and other datasets could be attributed to the fact that the classification algorithm primarily reduces the bias but it is also able to reduce the variance [30].

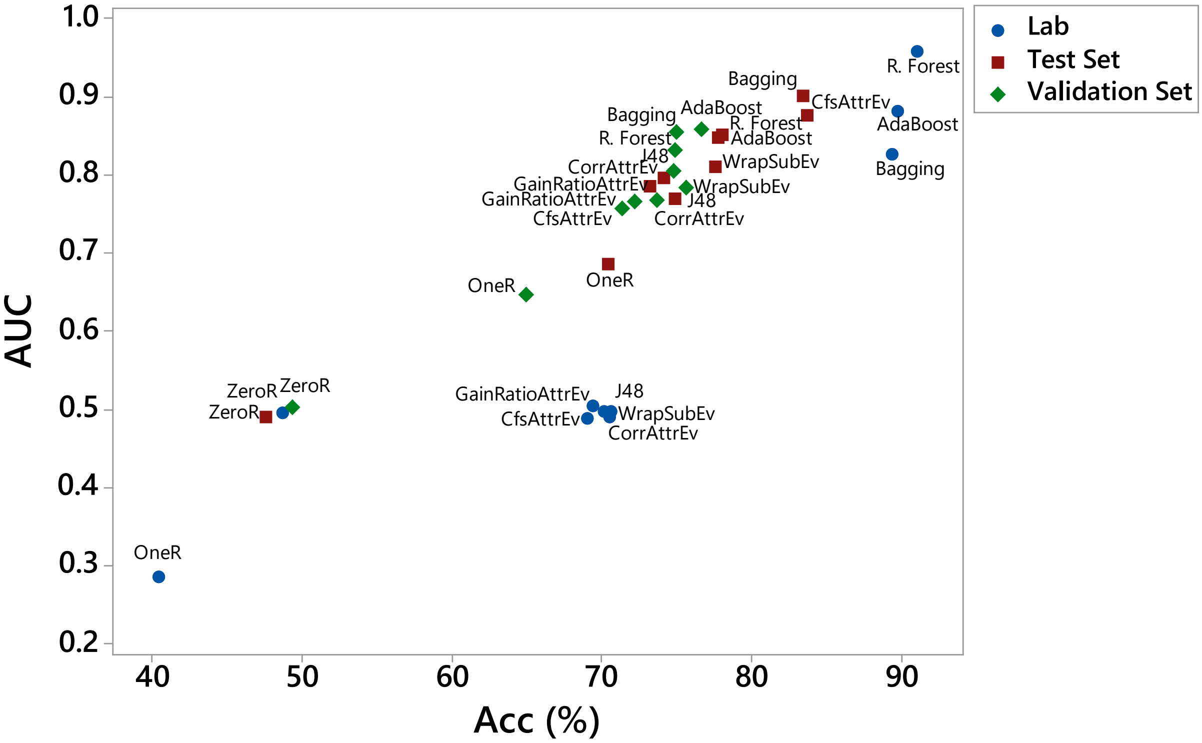

Figure 8 Model performance evaluation (Area under ROC curve versus percent of correctly classified instances) for the vaterite dataset. Colour groups indicate results for the cross-validation set (green), test set (red) and lab set (blue)

In conclusion, we created a classification predictive model using three metalearners that – given some initial conditions of pH, time, reaction temperature, oven drying temperature and reactant concentrations – successfully predict the presence or absence of vaterite in the final precipitate. However, this outcome tells nothing about how abundant vaterite will be in the crystalline product (will the phase be found pure or as part of a mixture?). Once a set of optimal conditions to predict the occurrence of vaterite was found, the next step in Figure 1 involves the repetition of this meta modelling procedure using another polymorphic data subset. In the case of mixtures, the decision tree is not shown but it provided with additional and complementary information to determine suitable and unsuitable experimental regions. There is a certain range of attribute values that made more likely the appearance of mixtures. The avoidance of these zones contributed to the success of the laboratory validation experiments.

Once the secondary data is available, the possibilities for analysis are broad, if enough time and effort is invested. For instance, we could built a model that calculate the polymorphic abundance of the precipitate using the mentioned attributes or different ones. In this case with a multilinear regression model to infer the effect of some variables on the numeric target attribute *polymorphic abundance* described in table 2 (PA_VAT), instead of the binary categoric attribute *VAT* used in the classification problem. However, this would require a completely different data processing to be able to meet all the assumptions of this type of analysis. Schmack et al. [31] provided a good example on how multivariate analysis can be coupled with a supervised learning strategy to examine relationships in a secondary dataset. The authors used a classification method with nested classes to refine the multiple regression model iteratively. The classes were built using reference tables from textbooks and expert knowledge hypotheses. The combination of multiple streams of data for meta-analysis was suggested by them as a way to improve the results.

3.2 Laboratory Validation

The main strategy in the synthesis of $CaCO_3$ was to promote the lifespan of ACC to minimize the production of mixtures and have a better control of polymorphism. Given the measurable influence of pH in the discrimination of single phases and the undeniable effect that time had on the precipitation of $CaCO_3$, effort was placed in controlling these two factors as a means of obtaining better persistence of ACC without full isolation of the material. A number of single phases (NEQ, CAL, ACC and MHC) were obtained with high purity in this way which confirms the success of the concept. Accordingly, a common experiment was designed and depicted in Figure 4 where the reaction time and the reaction temperature were fixed based on previous polymorph models (data not published). The selection of attributes in the synthesis of vaterite was influenced as well by our previous studies on polymorphism.

The response was sorted in descending order (Table 3) to identify what was the variable with the greatest effect on the response. Experiments where vaterite was synthesized in greater amount had in common a higher pH than the runs where the pH was at its lowest level (pH = 8.7). The effect of temperature is unclear because the design of experiments was left unfinished due to the COVID-19 lockdown. Results from runs 8, 9 and 11 are missing (run 10 is a centre point). However, the AdaBoost classifier predicts the presence of vaterite in the three missing experiments. Given the current results, a combination of low temperature (30 °C) and high pH (10.0) seems to be the best setting to maximize the response. The effect of rising pH can be observed by comparing run 3 and 2. At a fixed low level of tempOv (30 °C) and molar ratio (CO3/Ca = 3 corresponding to a high level concentration of calcium), changing the pH from 8.7 (more bicarbonate than carbonate) to 10.0 (more carbonate than bicarbonate) had a profound effect on the synthesis of vaterite (response changed from R_{VAT} = 0.240 at pH = 8.7 to R_{VAT} = 0.936 at pH = 10.0).

Table 3 Full factorial results sorted by the response in descending order. The actual and predicted occurrence of vaterite calculated with the AdaBoost classifier of VAT dataset is shown

Run Order	CP	pH	tempOv (°C)	CO3/Ca	R_{VAT}	Actual Occurrence	Predicted VAT (Yes, No)
8	1	10.00	30	9	-	?	Yes

9	1	10.00	50	3	-	?	Yes
10	0	9.35	40	6	-	Yes	Yes
11	1	10.00	50	9	-	?	Yes
2	1	10.00	30	3	0.936	Yes	Yes
4	0	9.35	40	6	0.899	Yes	Yes
5	0	9.35	40	6	0.893	Yes	Yes
6	1	8.70	50	9	0.529	Yes	Yes
1	1	8.70	30	9	0.402	Yes	Yes
3	1	8.70	30	3	0.240	Yes	Yes
7	1	8.70	50	3	0.004	No	Yes

The interaction effect between temperature and calcium at pH = 8.7 is plotted in Figure 10 – B. There was a strong interaction between these two factors at low levels of pH. Adding too much calcium when the amount of carbonate is low (at pH = 8.7 there is more bicarbonate than carbonate in the system) produced less vaterite when the experiments were performed at high temperature (R_{VAT} = 0 at 50 °C versus R_{VAT} = 0.240 at 30 °C). However, when the amount of calcium was reduced then an increase in temperature produced the opposite results and the synthesis of vaterite was favoured (R_{VAT} = 0.402 at 30 °C versus R_{VAT} = 0.529 at 50 °C). Comparatively, having a high level of molar ratio (CO3/Ca = 9) is preferred independently of the value of the temperature (Figure 10 – B). Main effect plot are not analysed when an interaction between the variables exist. The effect of the drying temperature on the response was more pronounced at 50 °C than at 30 °C. The effect of temperature and calcium was determined only at low level of pH (pH = 8.7) because at high level (pH = 10) most of the experiments were missing (Figure 10 – A).

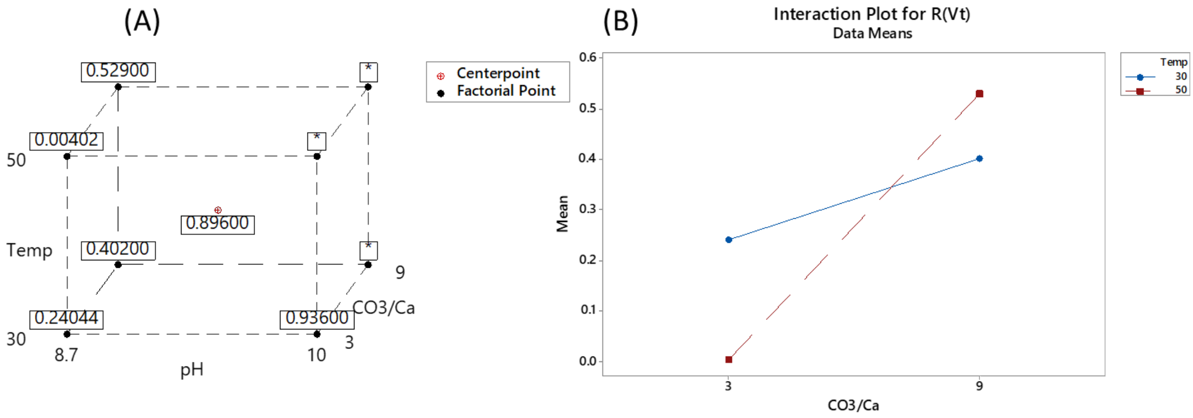

Figure 9 (A) Cube plot of pH, temperature and reactants molar ratio. The response label is shown above the vertexes of the cube. (B) Interaction plot between temperature and calcium amount for the synthesis of vaterite; constant pH = 8.7

4 Conclusions

Our first attempts to synthesize single phases of calcium carbonate started in a conventional manner, replicating One Paper at a Time (OPAT), but the outcome was difficult to control and polymorph mixtures were found often. Instead, the DDSM provided the focus that OPAT was lacking. This section summarizes the experimental insight gained from applying a data-driven approach. This mode of working could be classified under the second dimension of the scientific method: data-mining-inspired induction [32].

An inflexion point in the experimental strategy occurred when graph theory concepts were applied to understand amorphous calcium carbonate (ACC) research literature. Document and keyword co-occurrence networks provided an accurate representation of the structure of this topic. The co-occurrence map of keywords resembled the brain of a human with the right side representing biologically produced ACC and the left side the synthetically obtained amorphous material. A paper [33] was identified at the centre of this dichotomy using the document co-citation network. The uniqueness and understanding of this knowledge structure shifted our perspective and experimental efforts. The importance of ACC was also highlighted during secondary data analysis.

Results indicated that ACC was found more often in the final product than the other phases. Attributes such as time, pH and composition of reactants were statistically more significant in discriminating between the occurrence and absence of the amorphous phase. The synthesis of single phase ACC was optimal at short contact times and when the reactants were added simultaneously in the precipitation vessel. Based on the attribute selection procedure, ACC formation and persistence was more sensitive to aqueous pH than the crystalline phases. Information on the most relevant variables to discriminate between the appearance and the absence of each phase was compiled from the meta-analysis. The study included as well the identification of their optimal values (one decision tree per phase). Comparisons were drawn to identify experimental differences and similarities between the phases, and to determine the phases more sensitive to the variables with the greatest effect on ACC.

From here, a hybrid operation between the single-stage route and the multi-stage route described in Section 2.2 was created. Thus, the transformation from a precursor, metastable form, to a more stable polymorph was not done in the solution where the precursor was formed like in a traditional spontaneous precipitation experiment (single-stage route). The metastable precipitate of interest was ACC and the conditions to promote its lifespan were considered as a strategy to minimize the production of mixtures and control polymorphism. Moreover, it was not isolated at an early stage of the process like in a multi-stage route. Instead the reaction was delayed until it reached the oven. ACC was persistent after the separation stage for at least two hours. Phase transformation from ACC to vaterite occurred primarily in the oven and not in the solution. XRD characterization confirmed that samples reached the oven in an amorphous state and the polymorphic transformation occurred during drying operations and not in the solution during precipitation (Figure 4). This means that the optimization of operating variables such as the rate of water evaporation from the sample as a function of the fan settings and the drying time became relevant and had a direct effect in the distribution of polymorphs. The precipitation of $CaCO_3$ was easier to control in this way.

This was a concrete example on how the knowledge from different statistical approaches was applied dynamically to shape the experimental setup and arrive to an optimum result for all the phases.

5 Acknowledgements

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

6 References

[1] D. Konopacka-Łyskawa, “Synthesis methods and favorable conditions for spherical vaterite precipitation: A review,” *Crystals*, vol. 9, no. 4, 2019.

[2] L. Brecevic and D. Kralj, “On Calcium Carbonates: From Fundamental Research to Application,” *Croat. Chem. Acta*, vol. 80, no. 3–4, pp. 467–484, 2007.

[3] D. B. Trushina, T. V. Bukreeva, M. V. Kovalchuk, and M. N. Antipina, “CaCO3 vaterite microparticles for biomedical and personal care applications,” *Mater. Sci. Eng. C*, vol. 45, pp. 644–658, 2014.

[4] A. Carballo-Meilan, L. McDonald, W. Pragot, L. M. Starnawski, A. N. Saleemi, and W. Afzal, “Development of a data-driven scientific methodology: From articles to chemometric data products,” *Chemom. Intell. Lab. Syst.*, vol. 225, no. March, p. 104555, 2022.

[5] I. H. Witten, E. Frank, and M. A. Hall, *Data mining*. 2011.

[6] S. Drazin and M. Montag, “Decision Tree Analysis using Weka,” *Machine Learning-Project II, University of Miami*. pp. 1–3, 2012.

[7] N. Patel and S. Upadhyay, “Study of Various Decision Tree Pruning Methods with their Empirical Comparison in WEKA,” *Int. J. Comput. Appl.*, vol. 60, no. 12, pp. 20–25, 2012.

[8] Y. Wang *et al.*, “Gene selection from microarray data for cancer classification - A machine learning approach,” *Comput. Biol. Chem.*, vol. 29, no. 1, pp. 37–46, 2005.

[9] A. G. Karegowda, A. S. Manjunath, and M. A. Jayaram, “Comparative Study of Attribute Selection Using Gain Ratio and Correlation Based Feature Selection,” *Int. J. Inf. Technol. Knowl. Manag.*, vol. 2, no. 2, pp. 271–277, 2010.

[10] Y. Zhao and Y. Zhang, “Comparison of decision tree methods for finding active objects,” *Adv. Sp. Res.*, vol. 41, no. 12, pp. 1955–1959, 2008.

[11] R. G. Mantovani, T. Horvath, R. Cerri, J. Vanschoren, and A. C. De Carvalho, “Hyper-Parameter Tuning of a Decision Tree Induction Algorithm,” in *5th Brazilian Conference on Intelligent Systems*, 2016, pp. 37–42.

[12] Y. Freund and R. E. Schapire, “Experiments with a New Boosting Algorithm,” *13th Int. Conf. Mach. Learn.*, pp. 148–156, 1996.

[13] L. Breiman, “Bagging predictors,” *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.

[14] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[15] J.-H. Song and H.-Y. Woo, “A study on AQ (adversity quotient), job satisfaction and turnover intention according to work units of clinical nursing staffs in Korea,” *Indian J. Sci. Technol.*, vol. 8, pp. 74–78, 2015.

[16] N. Shibata, Y. Kajikawa, and I. Sakata, “Link prediction in citation networks,” *J. Am. Soc. Inf. Sci. Technol.*, vol. 63, no. 1, pp. 78–85, 2012.

[17] M. Gams, M. Horvat, M. Ožek, M. Luštrek, and A. Gradišek, “Integrating Artificial and Human Intelligence into Tablet Production Process,” *AAPS PharmSciTech*, vol. 15, no. 6, pp. 1447–

1453, 2014.

[18] Z. Yu, F. Haghighat, B. C. M. Fung, and H. Yoshino, “A decision tree method for building energy demand modeling,” *Energy Build.*, vol. 42, no. 10, pp. 1637–1646, 2010.

[19] J. A. S. Sá, A. C. Almeida, B. R. P. Rocha, M. A. S. Mota, J. R. S. Souza, and L. M. Dentel, “Lightning Forecast Using Data Mining Techniques On Hourly Evolution Of The Convective Available Potential Energy,” in *10th Brazilian Congress on Computational Intelligence*, 2011, no. November, pp. 8–11.

[20] D. Stojanova, A. Kobler, S. Džeroski, and K. Taškova, “Learning to Predict Forest Fires with Different Data Mining Techniques,” in *9th International multiconference Information Society*, 2006.

[21] S. J. Cunningham and G. Holmes, “Developing innovative applications in agriculture using data mining,” in *SEARCC’99 conference proceedings*, 1999.

[22] D. L. Olson, D. Delen, and Y. Meng, “Comparative analysis of data mining methods for bankruptcy prediction,” *Decis. Support Syst.*, vol. 52, no. 2, pp. 464–473, 2012.

[23] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, “Random forests and decision trees,” *IJCSI Int. J. Comput. Sci. Issues*, vol. 9, no. 5, pp. 272–278, 2012.

[24] Y. Boyjoo, V. K. Pareek, and J. Liu, “Synthesis of micro and nano-sized calcium carbonate particles and their applications,” *Journal of Materials Chemistry A*, vol. 2, no. 35. pp. 14270–14288, 2014.

[25] L. I. Kuncheva, V. J. del Rio Vilas, and J. J. Rodríguez, “Diagnosing scrapie in sheep: A classification experiment,” *Comput. Biol. Med.*, vol. 37, no. 8, pp. 1194–1202, 2007.

[26] C. Nadeau and Y. Bengio, “Inference for the generalization error,” in *Advances in Neural Information Processing Systems*, 2000, vol. 12, pp. 239–281.

[27] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The WEKA data mining software: An Update,” *ACM SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, 2009.

[28] G. Williams, *Data Mining with Rattle and R The Art of Excavating Data for Knowledge Discovery*. 2011.

[29] R. C. Holte, “Very Simple Classification Rules Perform Well on Most Commonly Used Datasets,” *Mach. Learn.*, vol. 11, no. 1, pp. 63–90, 1993.

[30] X. Wu and V. Kumar, *The Top Ten Algorithms in Data Mining*. Chapman & Hall/CRC, 2009.

[31] R. Schmack, A. Friedrich, E. V. Kondratenko, J. Polte, A. Werwatz, and R. Kraehnert, “A meta-analysis of catalytic literature data reveals property-performance correlations for the OCM reaction,” *Nat. Commun.*, vol. 10, no. 1, 2019.

[32] E. O. Voit, “Perspective: Dimensions of the scientific method,” *PLoS Comput. Biol.*, vol. 15, no. 9, 2019.

[33] L. Addadi, S. Raz, and S. Weiner, “Taking advantage of disorder: Amorphous calcium carbonate and its roles in biomineralization,” *Advanced Materials*, vol. 15, no. 12. pp. 959–970, 2003.

Secondary Dataset Description

Overall, 56 different attributes were compiled in Table 1 where each attribute name, type, range, definition and units are described.

The variables represented general *characteristics of the final precipitate* such as the identity of the polymorph (FstPhase), its molecular water content (PolType), its polymorphic abundance (%), the $CaCO_3$ precipitated yield (%), the amount of Mg (molar %) contained in the first phase and the mean particle size (nm).

The presence and absence of single phases (CAL, ARG, ACC, MHC, VAT…), and the presence and absence of mixtures were recorded as categorical variables. These *binary target attributes* have class Yes and No, corresponding to the occurrence and non-occurrence of a particular polymorph. If the polymorph was identified in the solid phase, then the case was labelled as Yes, otherwise a No was written. The additional binary target MIX indicated if the solid was pure or more than one phase was formed.

FstPhase represents a multiclass target attribute where the authors of that particular case identified first and second phases using XRD. The first phase is the most prominent phase when more than one phase were present. If abundance of the first phase is equal or greater than 85%, then the final precipitate was considered pure, and therefore named as ACC, CAL, ARG, MHC... Otherwise, the case was labelled as mixture (MIX). In this context, mixtures means that the characterized solid contained more than one polymorph.

System attributes included the type of reactants (carbonate source and calcium and/or magnesium salts), their initial molar concentrations, solution volumes and molar ratios, the synthetic route (SynRoute), the reaction temperature, the oven drying temperature, the initial and final pH, the sampling location (Sam_Loc), the contact time (min), the stirring speed (rpm), the feeding order (Feeding), the mixing mode and the reactant rate of addition (ml/min).

The exact definition of these *mole ratios* and percentages reads as follows:

$$Mg(\%) = \frac{[MgCl_2](M)}{[MgCl_2](M) + [CaCl_2](M)} \cdot 100 \tag{1}$$

$$R\left({}^{Mg}/_{Ca}\right) = \frac{[MgCl_2](M)}{[CaCl_2](M)} \tag{2}$$

$$R\left({}^{CO_3}/_{Ca}\right) = \frac{[Na_2CO_3](M) + [NaHCO_3](M)}{[CaCl_2](M)} \tag{3}$$

where $Mg(\%)$ is the molar percent of magnesium in the initial salt solution (corresponds to Mg_Pct in Table 1), Mg is the initial magnesium salt concentration (mol/L), Ca is the initial calcium salt concentration (mol/L) and CO_3 is the initial carbonate concentration (mol/L). These equations describe bulk compositions before mixing. $R(Mg/Ca)$ and $R(CO_3/Ca)$ were designated as Mg_Ca and CO3_Ca in Table 1. Regarding the type of salt, $CaCl_2$ and $MgCl_2$ were the source of Ca^{2+} and Mg^{2+} ions in these experiments. In the case of the carbonate ions researchers varied more their approach using sometimes only carbonates (K_2CO_3, Na_2CO_3), only bicarbonates ($NaHCO_3$) or a combination of both as initial source of carbonate ions.

Feeding described different ways to combine the salt and carbonate solutions at the initial stage of the precipitation process (simultaneous addition of both reactants, pour the salt solution on the carbonate solution – SaltToCarb, and the opposite arrangement – CarbToSalt). Once those reactants are combined, mixing and precipitation takes place. The different ways of mixing the suspension define the second categorical attribute called *Mixing*. In this case researchers have the option of vigorous stirring (dynamic setting), unstirred system (static) or a combination of both (first stirring then aging without agitation).

The attribute *SynRoute* represented two different approaches followed by the experimentalists to carry out the $CaCO_3$ synthesis. They were named as single-stage route and multi-stage route. The differences in the methodology of these two synthetic routes is described as follows: In the single-stage route, the transformation from a precursor, metastable form, to a more stable polymorph was done in the same solution where the precursor was formed, just by letting the system age. In the case of the multi-stage route, a 2-step synthesis method was done by the experimentalist. The metastable precipitate is isolated at an early stage of the process, filtered, dried and stored until the solid is resuspended in deionized water, in its mother liquor or in another freshly-made salt solution. It is in this second stage where the stable form is produced. These two scenarios were considered independently, so the dataset was split based on these two routes.

Table 1 Dataset description (A = 56 attributes)

Attribute	Type	Range	Description
Categoric Attributes related to reactant concentration			
Ca_Salt	Categoric	None, $CaCl_2$	Calcium salt
Mg_Sal	Categoric	None, $MgCl_2$	Magnesium salt
Carbonate	Categoric	None, K_2CO_3, Na_2CO_3	Carbonate source
Bicarbonate	Categoric	None, $NaHCO_3$	Bicarbonate source
Numeric Attributes related to reactant concentration			
V_CaSalt	Numeric	0 – 1.0	Volume of the calcium salt solution (L)
V_MgSalt	Numeric	0 – 0.5	Volume of the magnesium salt solution (L)
V_Carb	Numeric	0 – 1.2	Volume of the carbonate solution (L)
V_Bicarb	Numeric	0 – 0.5	Volume of the bicarbonate solution (L)
Volume	Numeric	0.05 – 2.0	Total volume of the solution mixture (L)
Ca_M	Numeric	0.001 – 2.0	$CaCl_2$ initial concentration (mol/L)
Mg_M	Numeric	0 – 0.9	$MgCl_2$ initial concentration (mol/L)
CO3_M	Numeric	0 – 2.0	Na_2CO_3 initial concentration (mol/L)
HCO3_M	Numeric	0 – 2.0	$NaHCO_3$ initial concentration (mol/L)
Mg_Ca	Numeric	0 – 10.0	Initial ionic Mg^{2+}/Ca^{2+} molar ratio
Mg_Pct	Numeric	0 – 91	Molar percent of Mg in the initial salt solution
CO3_Ca	Numeric	0 – 13.3	Initial ionic CO_3^{2-}/Ca^{2+} molar ratio
CO3_Mg	Numeric	0 – 18.0	Initial ionic CO_3^{2-}/Mg^{2+} molar ratio
Operational Categoric Attributes			
SynRoute	Categoric	Single-stage, Multi-stage	Experiments where the experiment was performed in two steps (Multi-stage route) or one step (Single-stage route)
Pathway	Categoric	None, ACC, VAT, MHC	Metastable precursor leading to stable form in multi-stage route
Feeding	Categoric	CarbToSalt, SaltToCarb, Simultaneous	Reactant addition mode

Mixing	Categoric	Static, Dynamic, Dyn_Stat	Mixing modes: with agitation, without stirring and a combination of both
Sam_Loc	Categoric	Bulk, Top, Bottom	Sampling location in the crystallizer
Operational Numeric Attributes			
Precursor	Numeric	0 – 48	Isolated metastable form in multi-stage route (g)
Rate	Numeric	3 – 200	Feeding addition rate (ml/min)
pH	Numeric	5.2 – 12.7	Initial pH (rich case solution)
F_pH	Numeric	5.2 – 12.7	Final pH
Var_pH	Numeric	-10.0 – 4.2	Variations in pH between final and initial conditions
TempRe	Numeric	5 – 100	Reaction Temperature (°C)
TempOv	Numeric	25 – 105	Oven drying temperature (°C)
t_min	Numeric	1 – 70,080	Contact time (min)
Mixing	Numeric	0 – 1000	Stirring of reactants (rpm)
Rate	Numeric	1 – 200	Rate of addition of reactants (mL/min)
Numeric Target Attributes			
CAL_Pt, ARG_Pt, MHC_Pt, ACC_Pt, VAT_Pt, NQ_Pt	Numeric	0 – 100	Polymorphic abundance (%) of calcite, aragonite, monohydrocalcite, vaterite, amorphous, nesquehonite…
Yield	Numeric	0 – 100	Total $CaCO_3$ precipitate yield
Mg_sld	Numeric	0 – 38	Amount of Mg in the polymorph (molar %)
Size	Numeric	90 – 40,000	Mean particle size (nm)
Categoric Target Attributes			
FstPhase	Categoric	VAT, CAL, ARG, ACC, MHC, NEQ, IKA, MIX	Appearance of a polymorph as first phase (Vaterite, Calcite, Aragonite, Amorphous, Monohydrocalcite, Nesquehonite, Ikaite, and Mixtures) if polymorphic abundance at least 85%; (*Multiclass target*)
PolType	Categoric	Hydrate, Anhydrous	Polymorph type. Crystalline nature of the polymorph. Refers to water content (*Binary target*)
ACC, CAL, ARG, MHC, MIX, MHC, VAT, IKA, NEQ, MG, HMG, DOL, LAN	Categoric	Yes, No	Ocurrence or Non-Ocurrence of a polymorph in the final precipitate (Amorphous, Calcite, Aragonite, Monohydrocalcite, Mixtures, Nesquehonite, Ikaite, Magnesite, Hydromagnesite, Lansfordite, Dolomite, Northupite,…) (*Binary targets*)

Decision Tree Model

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.6 -M 5

Relation:     Training-weka.filters.unsupervised.attribute.Remove-R1-2,5-6,14-16,18-
weka.filters.unsupervised.attribute.Remove-R3-5

Instances:    207

Attributes:   7

              Ca_M

              Mg_M

              TempRe

              TempOv

              pH

              time

              VAT

Test mode:    10-fold cross-validation


=== Classifier model (full training set) ===

J48 pruned tree

------------------

TempRe <= 60

|   Ca_M <= 0.026

|   |   Ca_M <= 0.003: Yes (6.0/1.0)

|   |   Ca_M > 0.003: No (35.0/4.0)

|   Ca_M > 0.026

|   |   time <= 120

|   |   |   TempRe <= 20: No (10.0/4.0)

|   |   |   TempRe > 20

|   |   |   |   TempRe <= 40

|   |   |   |   |   time <= 4.3

|   |   |   |   |   |   TempRe <= 25: No (9.0/4.0)
|   |   |   |   |   |   TempRe > 25: Yes (6.0)
|   |   |   |   |   time > 4.3: Yes (46.0)
|   |   |   |   TempRe > 40
|   |   |   |   |   Ca_M <= 0.6
|   |   |   |   |   |   time <= 15: Yes (5.0/2.0)
|   |   |   |   |   |   time > 15: No (9.0/4.0)
|   |   |   |   |   Ca_M > 0.6: Yes (5.0)
|   |   time > 120
|   |   |   TempRe <= 40
|   |   |   |   Mg_M <= 0.054: Yes (34.0/15.0)
|   |   |   |   Mg_M > 0.054: No (7.0)
|   |   |   TempRe > 40: No (5.0)
TempRe > 60: No (30.0/3.0)

Number of Leaves  :      13

Size of the tree :       25

Time taken to build model: 1.99 seconds

=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances        163               78.744  %
Incorrectly Classified Instances       44               21.256  %
Kappa statistic                         0.5748
Mean absolute error                     0.2914
Root mean squared error                 0.4078
Relative absolute error                58.2677 %
Root relative squared error            81.5465 %
Total Number of Instances             207

=== Detailed Accuracy By Class ===
              TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
              0.777    0.202    0.792      0.777    0.784      0.575    0.832     0.835     Yes
              0.798    0.223    0.783      0.798    0.790      0.575    0.832     0.774     No
Weighted Avg.    0.787    0.213    0.788      0.787    0.787      0.575    0.832     0.804

=== Confusion Matrix ===
  a  b   <-- classified as
 80 23 |  a = Yes
 21 83 |  b = No
```