

Development of a Data-Driven Scientific Methodology: From Articles to Chemometric Data Products

Ara Carballo-Meilan, Lewis McDonald, Wanawan Pragot, Lukasz Michal Starnawski, Ali Nauman Saleemi², Waheed Afzal

¹School of Engineering, University of Aberdeen, King's College, Aberdeen AB24 3UE

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

Abstract

Information and data science algorithms were combined to predict the outcome of an experiment in chemical engineering. Using the Scientific Method workflow, we started the journey with the formulation of a specific question. At the research stage, the common process of querying and reading articles on scientific databases was substituted by a systematic review with a built-in recursive data mining method. This procedure identifies a specific community of knowledge with the key concepts and experiments that are necessary to address the formulated question. A small subset of relevant articles from a very specific topic among thousands of papers was identified while assuring the loss of the least amount of information through the process. The secondary dataset was bigger than a common individual study. The process revealed the main ideas currently under study and identified optimal synthesis conditions to produce a chemical substance.

Once the research step was finished, the experimental information was compiled and prepared for meta-analysis using a supervised learning algorithm. This is a hypothesis generation stage whereby the secondary dataset was transformed into experimental knowledge about a particular chemical reaction. Finally, the predicted sets of optimal conditions to produce the desired chemical compound were validated in the laboratory.

Keywords

Scientific Method, Data Mining, Meta-Methodology, Chemometrics, Scientometrics, Machine Learning

1 Introduction

2 1.1 Secondary Data and Related Concepts

3 This work relies heavily on the concept of secondary data and secondary data analysis, its
4 procedures, problems, and merits. Secondary data is pre-existing material compiled for a new study,
5 primary data is the information collected by the original investigator and Big Data are just very large
6 secondary datasets [1]. The term secondary data analysis encompasses a methodology of analysis
7 that aims at investigating existing data in new ways than initially thought when it was created [1]. A
8 meta-analysis is a form of secondary data analysis that uses all the relevant literature available on a
9 subject [2]. In this case, the secondary data is taken from published manuscripts through a
10 systematic literature review. A systematic literature review is a method of collecting and
11 synthesizing manuscript data for posterior meta-analysis. Articles are filtered systematically
12 following an inclusion/exclusion criteria and the secondary data analysis is carried out using
13 statistical analysis to the aggregated data [2]. Mixed methodologies are defined as the combination
14 of qualitative and quantitative approaches to resolve a research question [3].

15 Data mining is an interdisciplinary subject using techniques from statistics, machine learning (ML)
16 and pattern recognition for mining knowledge from large amounts of data [4]. Additional domains
17 with strong influence on data mining methods include database systems, information retrieval,
18 visualization, and application domains. The most important topics in data mining research and
19 development are classification, clustering, statistical learning, association analysis and link mining
20 [5]. ML studies how computers could automatically recognize patterns and make decision on data
21 [6]. It can be subdivided in supervised learning (classification), unsupervised learning (clustering),
22 semi-supervised learning (including both supervised and unsupervised approach) and active learning
23 (actively acquiring domain expert knowledge from users). Terms like ML, artificial intelligence (AI)
24 and deep learning are all interrelated and offer sophisticated forms of inference, search and
25 optimization to incorporate in the scientific methodology [7]. Chemometrics is an interdisciplinary
26 field where statistics and data mining are used to solve a wide range of chemistry and chemical
27 engineering problems.

28 There are big secondary dataset compilations in fields like economics, psychology and political and
29 social sciences where the use of secondary data studies is common [1], [8], [9]. Most of this
30 information is published by government bodies and private organizations such as banks, universities,
31 and research institutes but researchers also have the choice of building their own secondary
32 datasets rather than using the ones built by third parties. In chemical engineering, fast simulations
33 produce large data sets that could be used for mining new hypotheses and confirm with an
34 experimentation process [4]. Several factors are likely to increase the popularity of secondary data
35 analysis in future and expand its applications to different fields of study such as a rise in computer
36 capacity, new and larger volumes of data and a more mature data science field [10], [11].

37 Although not free from criticism, secondary data analysis is perceived positively by different
38 institutions that persistently are assembling huge amounts of data free to use. The general
39 consensus is that research studies of secondary data compensate for long hours of trial and error in
40 the laboratory, saves money and avoids the repetition of research [12]. Furthermore, when
41 experiments from multiple studies are gathered to create a secondary dataset, secondary data
42 analysts have at their disposal a more diverse sample with many more variables and wider ranges
43 than a primary dataset could ever have [1]. This potentially could unravel new relationships as the
44 data is likely to be analysed with a different perspective [12].

45 While secondary data analysis saves time of tedious experimental trial and error in the laboratory,
46 they also require long hours of data analysis where sophisticated statistical techniques are applied
47 iteratively using the scientific method to get satisfactory results. Depending on the dataset, other
48 type of disadvantages include the influence of unmeasured variables associated with the response
49 [2]. The way the information is collected influence what you can do with the data later [1]. Ideally, a
50 researcher should design their own data collection method based on the type of research question
51 their study is intended to answer. Taking charge of the data collection stage is time consuming but it
52 could pay off in the long run because it sorts out some of the disadvantages associated with
53 secondary datasets. On one side, the researcher has more control over the collection process and,
54 on the other side, the process helps them to understand their research problem better rather than
55 spending time in understanding complex data structures built by others.

56 Laboratory experimentation is an arduous and time consuming activity where the trial and error
57 approach of the scientific method needs to be balanced with the resources and time available.
58 Typically, researchers conduct literature reviews manually and rely on previous discoveries to
59 circumvent the mentioned shortcomings. However, it is becoming increasingly apparent that the
60 vast number of publications at their disposal confuse rather than inform the scientific community.
61 With close to 2.5 million new scientific papers published every year [13], researchers feel
62 overwhelmed and unable to process this information. Even the most skilful of researchers is at risk
63 of missing key ideas unless he is able to read a significant amount of publications in his or her field of
64 study. Moreover, the navigation problem in the scientific databases is exacerbated by the bias
65 inherit in the metric used for the identification of quality research papers. In order to evaluate the
66 worth of a publication, most of the researchers rely on the number of citations that a piece of work
67 has received. But not many of them are aware of the limitations of this metric. As it turns out,
68 scientists tend to read and cite only the most recent papers in detriment of the older ones [14],
69 which makes them loose systematically old contributions from well-established researchers. It is
70 therefore evident than without an effective bibliometric strategy to manage science database
71 information, scientists could potentially waste energy, time and resources reading poor-quality
72 articles and/or repeating research than someone has already done and publish. A systematic
73 literature review and subsequent meta-analysis is an efficient method to deal with all these
74 undesirable effects that unmanageable amounts of information causes to researchers [2].

75 1.2 Aim and Organization of the work

76 This work reports a statistical mixed methodology called data-driven scientific methodology (DDSM).
77 The procedure empowers individual researchers with the tools and techniques required to build
78 their own secondary datasets from published research studies and mine this data to extract
79 meaningful knowledge and patterns of information in their specific fields of study. In essence, the
80 method is a modular framework likely to change with the nature of the problem. This is why the
81 work that started as a single manuscript ended up being three full articles. This one is solely devoted
82 to explain the idea behind all the steps described below. In this work, we will peel the outer layer of
83 the methodology to reveal a mode of thinking about problems using data and abstract designs.

84 This method was built as a by-product of a real life research struggle in a few organizations over the
85 course of several years. The intention behind the statistical recipe is captured in the following
86 questions: How can an experimentalist with a very limited amount of sample determine the right
87 experimental conditions without doing trial and error in the lab? How can a researcher in a short-
88 term contract, and with limited in-depth knowledge of a research topic, come up with an optimal
89 experiment in the laboratory without reviewing the bulk of the knowledge? Many questions were
90 asked while developing this method: 1) Will the information produced by these meta-models be

91 reliable? 2) How good will be those artificially-created experimental conditions once I apply them in
92 the lab? 3) How many experiments do I need to collect? 4) Will the conclusions drawn from some of
93 the primary datasets be corroborated in the meta-analysis or will a bigger dataset disprove those
94 ideas? This work will answer some of these questions, while others will require further work.

95 The DDSM article is organized as follows: 1) we introduce the basic definitions; 2) state the roots of
96 this research; (3) the generic workflow of the DDSM is shown and its two main mining components
97 are broadly depicted in the context of the scientific method; 4) the case study used to implement
98 this methodology is explained.

99 This research article is complemented with two method articles, each corresponding to a different
100 stage of the article processing methodology: Part 1 is titled "Systematic review using a semi-
101 supervised bibliometric methodology for application in a precipitation process", while Part 2 is
102 "Meta-Analysis of vaterite secondary data reveals the synthesis conditions for polymorphic control";
103 5) an executive summary of main results from these two publications is provided; 6) we discuss the
104 findings against the promises associated with the DDSM, included examples how a DDMS could be
105 customized with different algorithms and designs, some obvious applications and the limitations of
106 this mixed method; finally, 7) we concluded this introductory first article with an interpretation of
107 results and implications of this type of work in the evolution of the scientific method.

108 2 The Article Processing Methodology

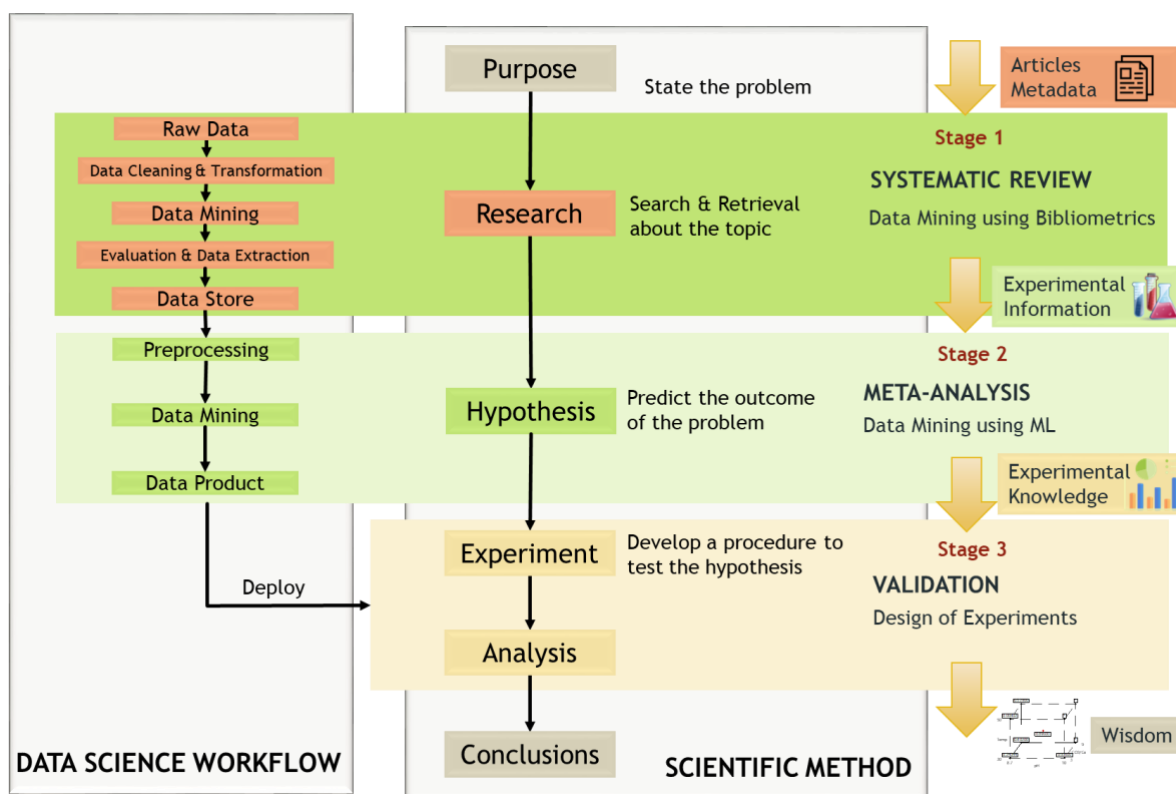
109 This work is an example of integration of data science, information science and domain knowledge
110 for compiling experiments based on a stated problem and then running an analysis for hypothesis
111 generation. The incorporation of data science in the scientific life cycle has been broadly depicted by
112 other authors [15] to indicate how data science could be used to generate hypotheses, as well as
113 design and analyse experiments. Hypothesis-driven research is deductive and corresponds to the
114 traditional scientific method. In contrast, data-driven research are inductive methods used to find
115 patterns, trends or principles [16]. The history of science suggest that a mixture of these methods is
116 usually involved [17].

117 Our DDSM follows the general data science workflow in the background (Figure 1), from the
118 collection of raw data (articles) to the production of a chemical engineering data product ready for
119 deployment in the laboratory. In this case the data product is a predictive model containing optimal
120 sets of experimental conditions to synthesize a chemical product. It is a statistical mixed
121 methodology constituted by three distinctive stages with two mining steps: a systematic literature
122 review for mining document co-occurrence networks (Figure 1 – Stage 1), followed by a meta-
123 analysis for mining chemical conditions from the acquired secondary data (Figure 1 – Stage 2) and
124 the validation stage in the laboratory (Figure 1 – Stage 3). At the higher level, there are two recursive
125 executions of the data science workflow embedded in an iterative scientific process. This is better
126 understood after reading the details of each stage.

127 Using a parallelism with the traditional scientific method: the journey starts with the formulation of
128 a specific question. At the research stage, the common process of searching and reading key text on
129 scientific databases is substituted by a systematic review with a built-in recursive data mining
130 method. The data mining methodology consists of several bibliometric techniques in combination
131 with a semi-supervised learning procedure that transforms scientific articles from Web of Science
132 (WoS) into comprehensive maps. They are used to identify a specific community of knowledge with
133 the key concepts and experiments that are necessary to address the formulated question.

134 Once the research step is finished, the experimental information is compiled in a single document
 135 and prepared for a statistical meta-analysis corresponding to the second stage of the DDSM. This is a
 136 hypothesis generation stage whereby the secondary dataset is transformed into patterns and
 137 experimental knowledge. In our case, the transformation produces a number of if-then decision
 138 rules covering the occurrence and absence of a particular chemical reaction. Finally, the predicted
 139 sets of optimal experimental conditions to produce the desired product are validated in the
 140 laboratory.

141 The combination of these different algorithms to collect, filter, rank, model, analyse and interpret
 142 scientific data taken from the literature is unknown to the average engineer and provides a novel
 143 approach not seen before. At its core, the first procedure is able to identify a small subset of
 144 relevant articles from a very specific topic among thousands of papers while assuring the loss of the
 145 least amount of information through the process. This is also a unique result of our method.
 146 Although this methodology was applied to the precipitation of calcium carbonate, it can be used and
 147 adapted to a broad range of engineering processes.



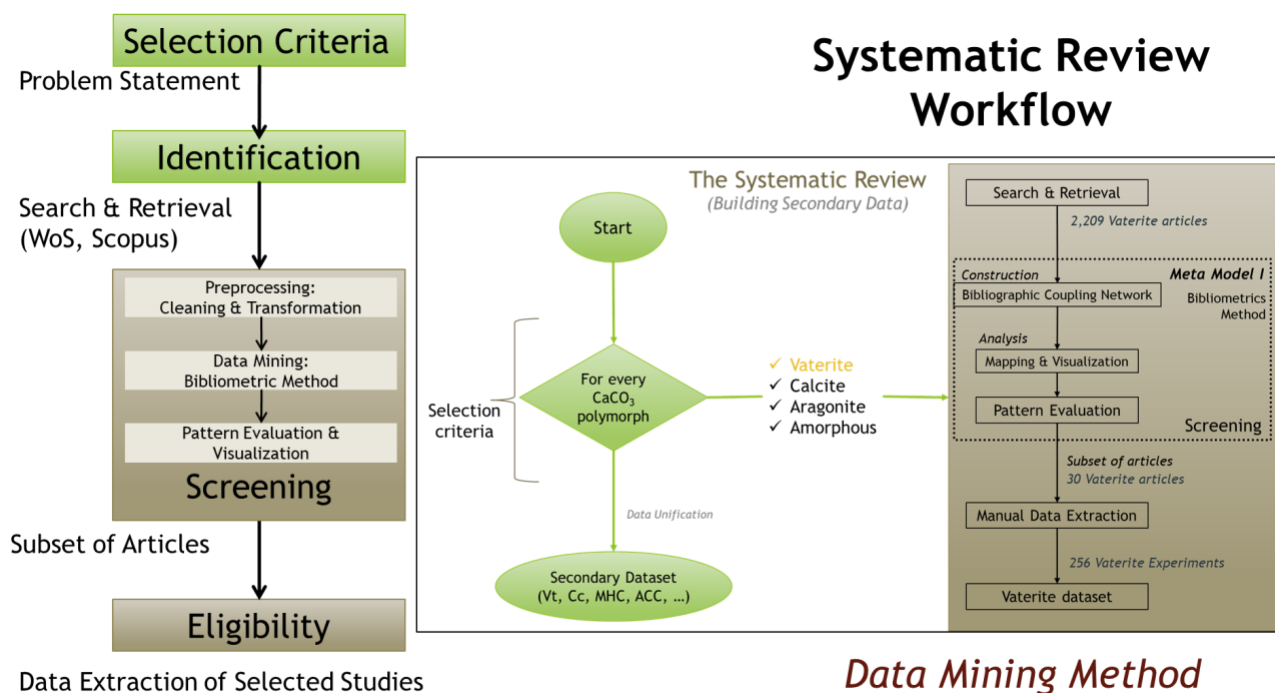
148
 149 *Figure 1 Article processing methodology (APM) towards the automation of the scientific method. This is a statistical mixed*
 150 *methodology constituted by three distinctive stages executed sequentially*

151 The two stages of the DDSM will be explained using a case study and Figure 2 and Figure 3 for stages
 152 1 and 2, respectively. Only an overview is shown, the full technical description can be found in their
 153 associated articles.

154 2.1 Stage 1: Systematic Review

155 In general, a systematic literature review method is executed following four basic steps (Figure 2 –
 156 Left) [18]: selection criteria (scope and definition of problem), identification (search and retrieval in
 157 scientific databases), manual screening to reduce the bulk of the information to a suitable subset,
 158 and eligibility (full text reading and data extraction of the selected studies).

159 An atypical study on the synthesis of calcium carbonate (CaCO_3) was conducted using the approach
 160 outlined in the previous section. CaCO_3 can precipitate in six solid forms: three anhydrous
 161 polymorphs (calcite, vaterite and aragonite) and three hydrated forms (amorphous CaCO_3 , CaCO_3
 162 monohydrate or monohydrocalcite, and CaCO_3 hexahydrate or ikaite) [19]. The precipitation involves
 163 the reaction between calcium and carbonate ions: $\text{CaCl}_2(\text{aq}) + \text{Na}_2\text{CO}_3(\text{aq}) \rightarrow \text{CaCO}_3(\text{s}) +$
 164 $2\text{NaCl}(\text{aq})$. Several methods were described to prepare calcium carbonate products synthetically
 165 [20]. The scope of the study was limited to the spontaneous precipitation method and the synthesis
 166 of single form vaterite and its mixtures with amorphous calcium carbonate (ACC), calcite and
 167 aragonite. The influence of the additive MgCl_2 was considered. Further details about the problem
 168 statement were relevant during the meta-analysis (Stage 2) and therefore this information was
 169 included in the corresponding method article: “Meta-Analysis of vaterite secondary data revealed
 170 the synthesis conditions for polymorphic control”.



171
 172 *Figure 2 Stage 1 or Systematic Review. On the left, the generic flow chart. On the right, the specific workflow of our example*
 173 *on the synthesis of vaterite*

174 Search and retrieval in WoS core collection was split by CaCO_3 polymorph using a divide and conquer
 175 strategy (Figure 2 – Right). For every CaCO_3 polymorph, a set of papers were retrieved from the
 176 scientific databases representing the core of their research. In the case of vaterite, all the articles
 177 and references containing the keyword Vaterite in the title-abstract-keyword were downloaded
 178 (April, 2020). This search retrieved 2209 papers with publications years spanned from 1925 to 2020.

179 The manual screening step of the systematic literature review was substituted by a bibliometric
 180 method (Meta – Model I). At the higher-level, the modified screening method repeats the same data
 181 science scheme that was used to create the backbone of the DDSM methodology. This recursive
 182 design was depicted in Table 1 where DDSM was compared with other applications of similar
 183 information technologies. The third step of the systematic review workflow, pre-processes, models
 184 and analyses all the information available on a studied topic to facilitate an unbiased selection
 185 criteria of subsets of articles and the assurance of missing the least amount of information through
 186 the process. The specific goal was to identify the publications more likely to contain “successful”

187 synthesis outcomes (i.e. the experiments giving pure phases of CaCO₃). The process also revealed
188 the main ideas and topics currently under study.

189 2.1.1 The Bibliometric Method

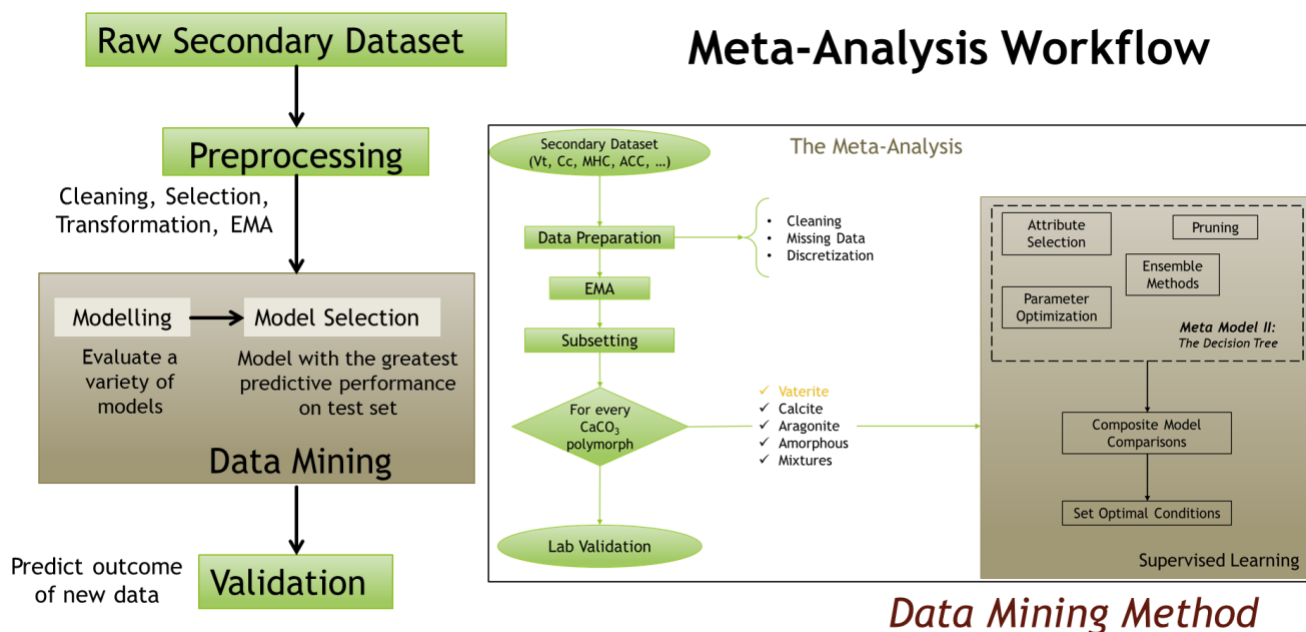
190 Several visualizations were produced adapting the methodological steps described by [21], [22]:
191 document retrieval, preprocessing, network extraction, normalization, mapping, analysis and
192 visualization. Articles were downloaded from WoS and transformed into maps using document
193 citation-based analysis and information visualization techniques to produce accurate
194 representations of the structure of CaCO₃ research. Two maps based on bibliographic data were
195 produced using two different types of studies: (1) A keyword co-occurrence analysis using all
196 keywords as the unit of analysis and (2) A bibliographic coupling analysis using articles. The co-
197 occurrence map of keywords was built in VOSviewer using the visualization of similarities technique
198 (VOS) [23], an alternative mapping technique to multidimensional scaling (MDS), and a variant of a
199 modularity-based clustering [24]. The bibliographic coupling network was created in Sci² Tool [25].
200 The clusterization of the bibliographically coupled (BC) document citation network was achieved
201 applying both the Leiden and the Louvain stochastic methods (modularity-based clustering) in Gephi.
202 The computation of the backbone of the network was performed using link reduction algorithms
203 (minimum spanning trees).

204 2.2 Stage 2: Meta-Analysis

205 The main idea behind the meta-analysis was to describe under which experimental conditions a
206 researcher is most likely to find a particular polymorph after the reactive crystallization process.
207 Specifically, the meta-analysis was applied to (1) indicate which of the studied parameters were
208 more relevant for the classification, and therefore able to play a greater role during precipitation, (2)
209 identify optimal sets of experimental conditions at optimal ranges to synthesize single phases and
210 control CaCO₃ polymorphism, and (3) develop an adequate experimental design and setup that was
211 tested in a real laboratory.

212 A sequence of steps were followed to process the secondary dataset (Figure 3 – Left): (1) data
213 preprocessing steps such as cleaning, data transformation, attribute selection, and exploratory
214 meta-analysis (EMA) were used to analyse the initial dataset and prepare it for the subsequent
215 modelling; (2) subsetting the CaCO₃ phases from the overall secondary dataset; (3) building the
216 decision tree models and collating all the results to produce an array of hypothesis from the EMA
217 study and the supervised learning algorithms. At the higher-level, the meta-analysis follows as well
218 the data science scheme used in the previous stage and in the overall structure of the DDSM
219 methodology. Finally, (4) the validity of the meta-model predictions was verified with laboratory
220 experiments. Full factorial experimental design was adopted to study the simultaneous effect of pH,
221 salt content (M) and the oven drying temperature (°C). The treatment objective was to achieve
222 vaterite single phase.

223 Data preprocessing was performed in IBM SPSS Statistics version 24 (missing data analysis), Minitab
224 17.1.0 and Rattle version 5.1.0, a free graphical interface for data science with R (data exploration,
225 discretization and design of experiments). Waikato analysis for knowledge environment (Weka
226 version 3.8.1) [26] was used as data mining software to assist the decision tree model construction
227 and evaluation process.



228

229 *Figure 3 Stage 2 or Meta-Analysis. On the left, the generic flow chart. On the right, the specific workflow of our example on*
 230 *the synthesis of vaterite*

231 The subset of the secondary dataset corresponding to vaterite experiments had 256 experiments
 232 with 36 different attributes describing each experiments. The variables represented general
 233 characteristics of the final precipitate such as the identity of the polymorph First phase (FstPhase),
 234 its molecular water content (PolType), its polymorphic abundance (%), the CaCO₃ precipitated yield
 235 (%), the amount of Mg (molar %) contained in the first phase and the mean particle size (nm).
 236 System attributes included the type of reactants (carbonate source and calcium and/or magnesium
 237 salts), their initial molar concentrations, solution volumes and mole ratios, the synthetic route
 238 (SynRoute), the reaction temperature, the oven drying temperature, the initial and final pH, the
 239 sampling location, the contact time (min), the stirring speed (rpm), the feeding order, the mixing
 240 mode and the reactant rate of addition (ml/min).

241 The analysis of missing data was performed to describe patterns of missing values, assess if missing
 242 values were random and finally decide if a missing value required a multiple imputation method.
 243 With regards to cleaning, the numerical attributes were rounded up to the nearest integer or
 244 nearest decimal. Once the dataset was collected and cleaned, new features were defined in order to
 245 use classification algorithms. Discretization was intended to construct meaningful boundaries that
 246 could explain the differences observed in the polymorphism with time. The 230 instances forming
 247 the balanced dataset were split randomly in two groups named training/validation set (90%) and
 248 test set (10%). Data exploration was performed over the training set. The training set was also used
 249 by the learning scheme to build the classifier, the validation set was used for parameter optimization
 250 and to compare and select the best classifier. However, the final true model performance was
 251 assessed using only the test set, which was set aside from the beginner of the modelling process.
 252 The training set was balanced (same proportion of each class) and the test set also had each class
 253 well represented. Once the modelling procedure was finished and a reliable predictive power was
 254 obtained using the unbiased test set, the EMA and model were rebuilt with a whole balanced
 255 dataset ready for deployment in the Lab Validation stage.

256 In the data exploration stage, sample distribution analysis using bar charts, box plots and density
 257 plots was performed. The worth of each attribute was investigated following feature selection

258 techniques. Two single-attribute evaluators were used, named *GainRatioAttributeEval* and
259 *CorrelationAttributeEval* in Weka.

260 A classification predictive model was created at the data mining step (Meta – Model II in Figure 3 –
261 Right) using supervised learning algorithms such as J48 decision tree and the meta-learners
262 associated with it to classify cases into categories. Pruning, feature selection and ensemble trees
263 using bagging and boosting techniques were used to optimize the tree and avoid overfitting. Model
264 construction was done using the training/validation set containing 207 instances and several
265 attributes such as reaction temperature (TempRe), oven temperature (TempOv), reactants
266 concentration ($[CaCl_2]$, $[MgCl_2]$), the initial pH and the contact time. They were used to split the
267 vaterite secondary data into pure nodes or subsets belonging to a single class. Training dataset was
268 balanced and contained no missing values (except for pH). The binary class target attribute VAT used
269 for classification was formed by 2 categories: *Yes*, *No*; corresponding to the occurrence and the non-
270 occurrence of vaterite precipitation. The performance of the studied classifiers (ZeroR, OneR, J48
271 pruned, Bagging, AdaBoost, Random Forest, cost-sensitive and attribute selection schemes) was
272 calculated using both *Acc* (accuracy or percent of correctly classified instances) and *AUC* (Area under
273 the ROC curve) as a combined measure of the overall quality [29], [30]. Differences in *AUC* and *Acc*
274 among classifiers were determined using stratified 10x10-fold cross validation in the Weka
275 Experimenter and the corrected paired t-test statistic with 95% confidence level (two tailed). This
276 corresponded to a total of 100 experimental runs per dataset and classifier. Finally, a decision list
277 was extracted from the decision trees and interpreted in the context of a precipitation experiment.

278 The lab validation experimental setup and the level of the variables were built and selected using all
279 the results from the secondary data analysis. Full factorial design was adopted to study the
280 simultaneous effect of pH, salt content (M) and the oven drying temperature ($^{\circ}C$). The treatment
281 objective was to achieve vaterite single phase. A total of 11 experiments (also called runs) were
282 performed by designing a full factorial with 3 centre points, 3 factors and no replicates. All terms
283 were free from aliasing, including main effects and 2-way interactions. By default, all experiments
284 were randomized to reduce the effect of experimental bias. The independent variables (also called
285 factors) were the pH, oven temperature ($^{\circ}C$) and mole ratio CO_3/Ca (M). Their levels low (-1), middle
286 (0) and high (1) are the following: pH (8.7 – 9.3 – 10.0), oven temperature (30 – 40 – 50 $^{\circ}C$), and
287 CO_3/Ca (3 – 6 – 9). The polymorphic abundance of vaterite R_{VAT} (0 – 100 %) was set as the main
288 response. Qualitative and quantitative phase analysis was done using X-ray diffraction (XRD) in a
289 Panalytical X'Pert Powder diffractometer and the Rietveld multiphase refinement method to
290 determine phase abundance.

291 3 Results and Discussion

292 3.1 Stage 1: Systematic Review Findings

293 The aim of the clustering procedure was to identify a single cluster containing data on the
294 spontaneous precipitation method among all the $CaCO_3$ communities. Each cluster in a bibliographic
295 coupling network represents a common topic and display similar language. This led to the
296 assumption that most if not all of the vaterite spontaneous precipitation experiments would belong
297 to their own cluster because they share the same topic. However, the identification of those subjects
298 in a BC network or how many research divisions are under the overall structure is not known a priori
299 using unsupervised learning.

300 As a first approximation to the unknown segmentation, a semi-supervised procedure was developed
301 at the clustering step. It was semi-supervised because clusterization occurred with constraints on

302 clusters based on the identity and size of the clusters specified by the classification technique, thus,
303 controlling the non-deterministic nature of modularity. To accomplish this task, the keyword co-
304 occurrence network and the top review papers (the ones with the highest weighted degree) were
305 used to identify the research topics and to label the identified clusters, respectively. A total of 9
306 research topics (the classes) emerged from the analysis: biomimetic synthesis, drug delivery
307 applications, regenerative medicine applications, manufacturing applications, natural geological
308 systems, water treatment applications, natural biological systems, photoluminescence and
309 microrheology. These are important vaterite research divisions investigated at present.

310 Following this identification, a document tracing method with two components (the classes and a
311 document tracer) was implemented on top of the community detection procedure. It evaluated the
312 movement of nodes with spontaneous precipitation experiments between communities during the
313 recursive modularity optimization of the bibliographically coupled network. The document tracer
314 was used as a metric to evaluate who and where was the cluster with the higher amount of relevant
315 vaterite experiments during community detection. It consisted of a set of 30 randomly selected
316 papers containing suitable spontaneous precipitation experiments of vaterite. The accuracy of the
317 partition (percent of correctly classified nodes) was defined as the fraction of tracing documents
318 found within a cluster divided by the total number of documents in the tracer. The number of
319 classes (categories) and their identity were supplied by the co-word analysis.

320 The bibliographic coupling clusterization procedure accomplished one the main objectives of the
321 systematic review: the identification of a research community with many publications on the
322 inorganic synthesis of vaterite (CaCO_3) using the spontaneous precipitation method: $\text{CaCl}_2 +$
323 $\text{Na}_2\text{CO}_3 \rightarrow \text{CaCO}_3 + 2\text{NaCl}$. This relevant community included 170 documents associated mainly
324 with two research topics: manufacturing (use of CaCO_3 as additive in cements) and natural geological
325 systems (CaCO_3 formation in geological environments). These research studies contained work about
326 the spontaneous precipitation method ($\text{Na}_2\text{CO}_3/\text{NaHCO}_3$ (l) – CaCl_2 (l) system) and the CO_2 diffusion
327 method (NaOH/CO_2 (g) – CaCl_2 (l) system), both in batch and semi-batch crystallization experiments.
328 The network was densely connected and with overlapping communities which hindered the
329 identification process. Nonetheless, the inspection of the 75 nodes with the highest weighted degree
330 within the target cluster indicated that 44% of these documents corresponded to the specific
331 process of interest. This statistic was underestimated because experiments adding unwanted
332 additives were not computed in this counting but it still shows the validity of the methodology to
333 identify relevant information.

334 Once this cluster was found and isolated from the rest of the network, the set of papers closely
335 related with our research problem were ranked as a function of their influence and role within the
336 network using statistical network analysis. The experimental works with the highest weighted
337 degree were included in the secondary dataset. Once the number of articles was reduced to a
338 manageable number doing multiple-level data sorting, experimental information was collected
339 manually from the document subset. In the case of vaterite, the automated screening reduced the
340 total number of articles from 2,209 to 30 and a total of 256 experiments were collected. Each case
341 was defined by 36 numeric and categoric attributes.

342 3.2 Stage 2: Meta-Analysis and Validation Results

343 3.2.1 Data Preparation

344 EMA was used to identify the variables that have the greatest impact on a specific polymorph,
345 describe the effect of those key variables on the class (the absence or presence of a phase) and
346 determine the most optimal values of those variables to produce specific polymorphs with the

347 particular emphasis on vaterite. The layering of the polymorph datasets revealed some interesting
348 patterns and gave an additional perspective to the way in which the problem was previously
349 thought. This allowed the meta-analysis to become a hypothesis generation stage as well as a
350 modelling tool. Adjustments to the experimental setup were made to test the hypothesis that the
351 selective control of the pH would lead to the synthesis of all the phases in a short period of time.
352 Statistical significance was found to this phenomena currently under research in the literature [27],
353 [28]. Contrary to what might be thought, there was a greater number of successful experiments in
354 the literature obtaining amorphous CaCO_3 and its formation and persistence was more sensitive to
355 aqueous pH than the crystalline phases. The analysis also indicated that the inorganic synthesis of
356 vaterite was mainly performed in the absence of magnesium and the importance of temperature as
357 a means to obtain purity was highlighted in many of the analysed documents. This was a distinctive
358 feature of vaterite and some of the other CaCO_3 polymorphs such as aragonite. Vaterite was the
359 CaCO_3 phase less likely to occur, being found only in 18% of all the collected experiments. The
360 presence of mixtures in the final product was a widespread issue. Experiments where the vaterite
361 occurrence was positive had in common: a contact time lower than 60 min, a CaCl_2 salt solution with
362 no Mg content and a concentration of 0.1 M, reaction performed at ambient temperature and an
363 initial pH of at least 10.0. Additionally, setting the oven drying temperature higher than 25 °C (the
364 median was 50 °C).

365 3.2.2 Model Construction and Evaluation

366 The modelling section included the construction, optimization and comparison of several algorithms:
367 *J48 pruned*, *AdaBoost*, bagging, *Random Forest*, *OneR* and *ZeroR*. Pruning optimization, ensemble
368 learning, cost sensitive evaluation of a binary classifier and attribute selection meta-learners were
369 some of the strategies used in this work to improve the decision tree classification performance. The
370 attributes more relevant for classification were the temperature, the salt concentration and the
371 time. A number of if-then decision rules were created covering the occurrence and absence of
372 vaterite.

373 The classifiers with the best performance were those having simultaneously high accuracy and high
374 *AUC*. The larger is this area, the better is the model [31]. In general, an ideal prediction has *AUC*
375 values around 1, while a random decision will show an *AUC* of 0.5. Results from the Weka
376 Experimenter indicated that the J48 pruned tree had an average accuracy rate of $73.8 \pm 8.7\%$ (10
377 iterations) at the 95% confidence interval. The metalearners (boosting, bagging and random forest)
378 outperformed J48 and all the other classifiers. They showed the greatest accuracy and largest *AUC* in
379 all sets: the validation, test and lab sets. Their prediction on the lab test set was good (*AUC* greater
380 than 0.8 and Acc around 90%). In particular, the AdaBoost classifier was able to successfully predict
381 the presence or absence of vaterite in the final precipitate. Its model performance evaluation
382 showed the greatest accuracy and largest *AUC* in all sets (validation, test and lab sets). It consisted of
383 just 3 decision trees and the excellent performance of this metalearner was attributed to the fact
384 that the classification algorithm primarily reduces the bias but it is also able to reduce the variance
385 [5].

386 3.2.3 Validation

387 The lab validation experiments were uncompleted due to the Coronavirus pandemic but the current
388 findings already meet the target: the synthesis of pure phase vaterite (equivalent to a polymorphic
389 abundance of at least 85%) using the spontaneous precipitation method. The best result
390 corresponded to an experiment with the following conditions: pH = 10.0, time = 4 min, tempRe = 25
391 °C, tempOv = 30 °C, $[\text{CaCl}_2] = 0.33 \text{ M}$, $[\text{Na}_2\text{CO}_3] = 0.55 \text{ M}$, $[\text{NaHCO}_3] = 0.45 \text{ M}$; the obtained vaterite
392 polymorphic abundance was $93.6 \pm 0.3\%$. At low level of pH, an interaction effect between oven

393 drying temperature and calcium concentration was identified. The effect of the drying temperature
394 on the response was more pronounced at 50 °C than at 30 °C. The effect of temperature and calcium
395 was determined only at low level of pH (pH = 8.7) because at high level (pH = 10) most of the
396 experiments were missing. Adding too much calcium with respect to the amount of carbonate
397 produced less vaterite when the experiment was performed at high temperature. The opposite
398 effect of temperature was observed on the response at lower calcium concentrations (same amount
399 of carbonate).

400 Although this work focuses only on vaterite and the analysis was limited to 256 cases, a total of 732
401 experiments with information about many of the anhydrous and hydrous forms of CaCO₃ were used
402 in a bigger study. Some conclusions on CaCO₃ polymorphism – not seen before using primary data –
403 include: (A) The effect some attributes have on only certain phases; (B) The identification of 3
404 variables that can be fixed at a certain range to obtain all the single phases of CaCO₃, just by fine
405 tuning 1 extra parameter that changes depending on the polymorph (we named it as “the inclusive
406 experimental region”); (C) The identification of the experimental conditions where mixtures are less
407 likely to occur; (D) The identification of the experimental conditions where only specific single
408 phases will be obtained (we called this “the exclusive experimental regions”). Overall, it was possible
409 to carry out a different investigation as a result of the unique size and characteristics of the collected
410 data. The secondary dataset included numerous cases and variables from primary data but also new
411 features were created as a result of this combination. In the absence of the secondary dataset, a
412 significant amount of experiments would be required to arrive at the same conclusions and it seems
413 unlikely that we would have generate this new insight without this approach.

414 3.3 Summary of Evidence

415 Data science is stimulated not just by the development of mathematical algorithms but by the
416 creativity of researchers to apply current algorithms to new grounds and their crafty skills to
417 harmoniously blend different statistical techniques and emerge knowledge from both primary and
418 secondary datasets. However, for this new mixed methodologies to flourish, the structure of the
419 dataset have to capture the complexity of the problem, its depth, be organized rather than chaotic
420 and have the right scale. With these requirements as a premise, the pair bibliometrics – ML was a
421 convenient union for the application of statistical mixed methodologies to chemical engineering
422 problems. Stage 1 was key to build the fingerprint of the problem. The emphasis was not in
423 designing a method to collect lots of information but to determine a valid procedure to capture
424 optimal knowledge, that is, articles that most likely contained the solution we were pursuing.

425 The screening step of the bibliometric method could reduce the bias associated with manual
426 screening [18] but do not necessarily eliminate publication bias in the individual studies. Publication
427 bias is the tendency in the literature to publish only positive results and arguments that lead to the
428 particular idea that a researcher wants to prove or promote [32]. In this regard, the second DDSM
429 stage might tackle this source of bias better. The selective citation and repetition of experiments
430 with positive findings was observed during network analysis. For instance, it is widely reported in the
431 literature that an optimal molar ratio of magnesium to calcium ions of 20% Mg content produces
432 single phase monohydrocalcite [33]. The meta-analysis and validation of this polymorph indicated
433 that other optimum regions also exist, thus not only corroborates the previous suspicion but also
434 indicates that this type of analysis can highlight different synthetic pathways.

435 If we believe that the peer review process feeds the scientific databases with quality primary data
436 and if the statistical methods under development are adequate, then it follows that the combination
437 of this data with technology has the potential to bypass conventional scientific research. Some fields

438 of study are well researched and this also ensures quantity of information. Given the staggering
439 amount of data at our disposal [13], it does not seem wise to keep loading Science with more
440 experiments when we are unable first to digest the ones we collectively own. We can find already
441 articles in the literature pointing out in this direction [15] and additional examples of successful
442 meta-analysis using different sources of information [34].

443 One of the main criticisms to the use of secondary data analysis is that the data is full of errors to
444 arrive at valid conclusions [12]. It could be argued extensively about all the technicalities that makes
445 this approach wrong, many of them related with how perfect we would like the data to be. Indeed,
446 there is not such a thing as data without errors (including the one described in this work). However,
447 all it is needed is the selection of the right technique, an understanding of the data limitations and a
448 careful approach to capitalize on the secondary data. With enough information common trends will
449 be revealed even in the presence of error in individual studies [35]. When a validation procedure
450 confirms the model prediction of the meta-study, then it follows that even if the data is not perfect,
451 the method could be useful. What we achieved after model development was the confidence of
452 being able to enter the lab with a couple of experiments to test in a short period of time. This mixed
453 methodology helped to successfully complete two engineering projects belonging to different fields
454 of study (adsorption and crystallization processes). In its infancy, the idea was sketched and applied
455 with relative success to model sorption processes. Then given its apparent good performance, the
456 concept was fully developed into what constitute this paper. The overall outcome of the study was
457 an experimental set up able to produce single phases of CaCO_3 just by modifying two attributes,
458 magnesium content and oven temperature, with the premise of meticulously controlling the pH of
459 both carbonate and salt solutions.

460 3.4 Examples of Systematic Reviews and Meta-Analysis

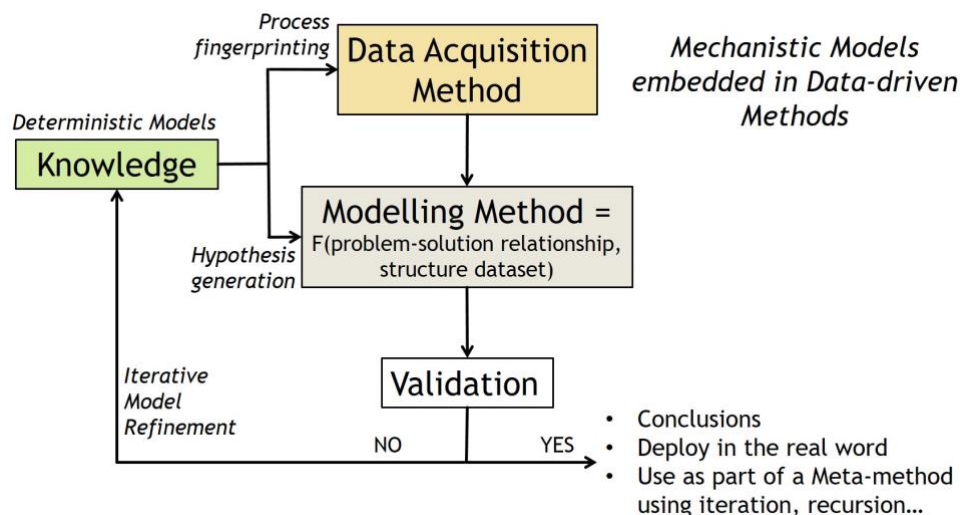
461 Table 1 explores some of the latest advancements towards the automatization of the traditional
462 scientific method in different applied science fields. The list is not exhaustive and only shows a few
463 examples on how systematic reviews and meta-analysis are applied to solve science and engineering
464 problems.

465 A recent systematic review included a bibliometric method with a supervised learning technique
466 [35]. In this work the authors used a keyword co-occurrence analysis to group articles in categories
467 using K-means clustering. The size of clustered data was kept arbitrarily lower than 250 data points
468 because it was considered manageable for manual checking. Then a supervised learning strategy was
469 coupled with the previous step to identify which cluster was the most relevant. In this case
470 maximum entropy algorithm was used as the classification algorithm using a training data subset
471 with 2 classes “relevant” and “not relevant”. The cluster with the greater similarity to the “relevant”
472 class was used to extract the secondary data from the primary publications. Meta-analysis was done
473 using univariate statistics (e.g. polynomial regression analysis with Pearson’s correlation, 2-tailed T-
474 test). This example corroborates that semi-supervised bibliometric methods are feasible and can be
475 tuned with different algorithms to satisfy the problem requirements.

476 Systematic literature reviews with manual screening are tedious and require long times for
477 completion. For instance, a study that started with 2650 papers down to 525 final included studies
478 required almost three years, with two authors independently screening and extracting data [36]. In
479 our case manual screening of publications was reduced a 99% (from 2,209 to 30 articles) and the
480 time of completion of the automated systematic review per polymorph from the initial design,
481 search, automated screening, eligibility and data extraction was less than three months (full time). In
482 the ML aided systematic review described in the previous paragraph [35], the time in manual

483 identification was as well decreased and manual screening of publications was reduced an 87%
484 (from 4,177 to 555 articles).

485 The way determinism was integrated in the data-driven methodology can be further analysed using
486 the following examples. In general, the combination of deterministic and data-driven approaches
487 can take place from the point of view of the mechanistic modelling or using mechanistic hypothesis
488 within a machine-learning-based method [37]. The generic framework depicted in Figure 4 follows
489 this latest case and represents research methodologies that integrate expert knowledge within a
490 data analytics process with a supervised learning approach. Supervised learning is the most common
491 form of ML [38].



492

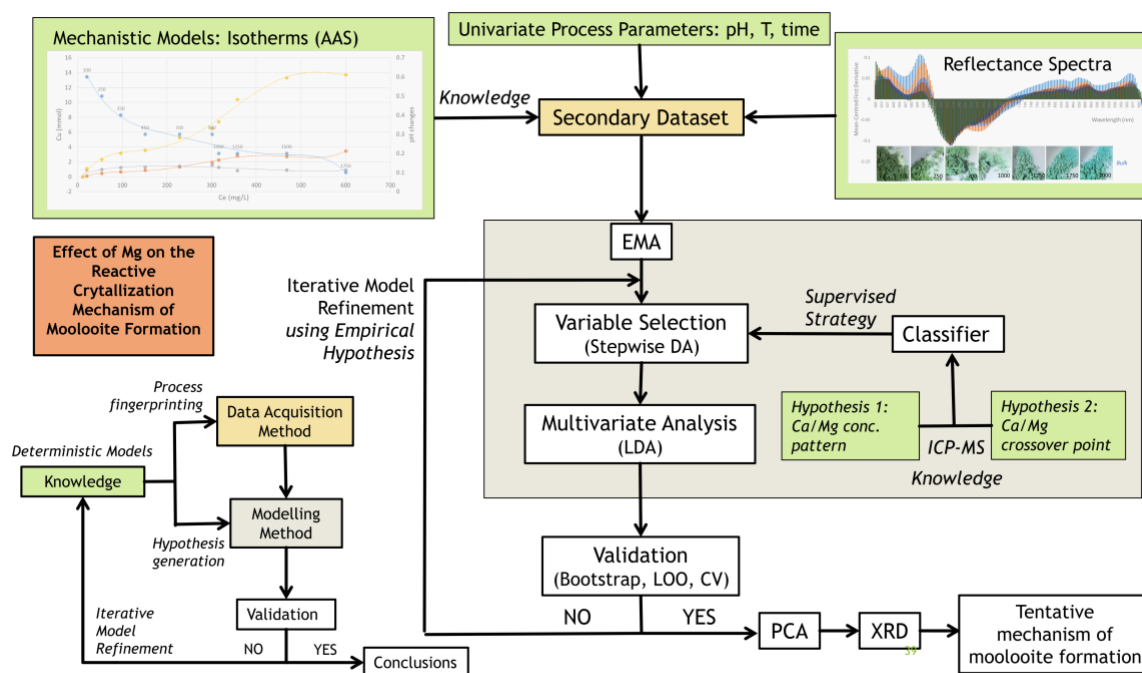
493 *Figure 4 Method sketch of a research framework for the integration of mechanistic modelling in data-driven methodologies*

494 Two specific stages determine the inclusion of facts: the data acquisition stage and the modelling
495 stage. The formulation of the problem and its analysis through deterministic models can be coupled
496 with relevant data from simple sensors (e.g. pH meter, thermometer) and/or complex analysers (e.g.
497 spectroscopic instruments) for process fingerprinting in the corresponding data acquisition method.
498 Literature data also helps to understand how a process works and to identify what are the most
499 important attributes. The modelling stage represents the solution methodology. Specific supervised
500 techniques are designed to test a hypothesis. Iteration occurs as part of this method and can involve
501 empirical observations, categorical attributes or numerical cut-off values depending on the structure
502 of the problem and the intended solution (problem-solution relationship).

503 The examples depicted in Figure 5 to Figure 7 show different configurations of a data acquisition
504 method. The simplest case is described in Figure 7. The data was captured with a single analytical
505 instrument (FTIR) using a rigorous sampling method based on the nature of the problem. The
506 chemical composition of wood extracted from the FTIR data was used to discriminate wood samples
507 between division, class, subclass, order and family, taken groups from the current plant APG II
508 classification system. More complex examples of a data acquisition method involve the application
509 of sophisticated information science methods like the one discussed in this paper (Figure 6) or data
510 fusion from multiple sources (Figure 5). In Figure 5, the role of magnesium in the transformation of
511 malachite into moolooite during the biosorption of copper on nopal fibres was discovered using a
512 combination of biosorption data at equilibrium (isotherm models) and spectroscopic analysers (ICP-
513 MS, XRD...).

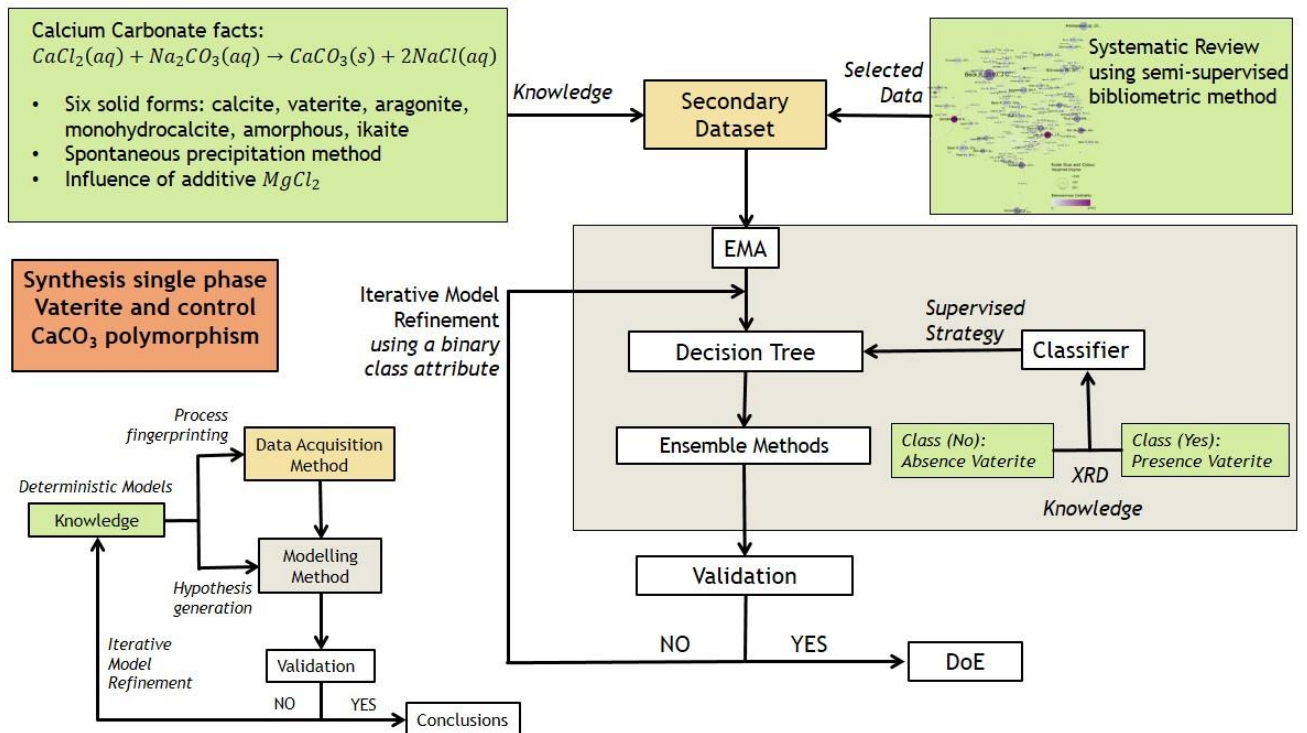
514 Two meta-analysis with supervised techniques were described at the modelling stage [34], [39]. The
 515 generic workflow depicted in Figure 5 correspond to one of these two examples. In this case, the
 516 correlation between the colour (micro-spectrophotometer), XRD analysis, pH shifts and hard cations
 517 released from the biosorbent into the solution (ICP-MS) suggested the existence of two crystal
 518 formations, malachite and moolooite. The shift of the molar ratio Mg/Ca was located between the
 519 two inflexion points of the crystal growth transition. This specific location at the sorption isotherm
 520 was correlated with the colour evolution by a linear discriminant model confirming its association
 521 with the polymorphs. A similar meta-analysis method to formulate hypotheses with a high level of
 522 abstraction have been described [34]. In this case the authors created the secondary dataset using
 523 experiments from the literature and reference tables from textbooks (melting points, formation
 524 enthalpies, carbonate decomposition temperatures...). In Figure 7, unknown samples of trees were
 525 classified correctly using a combination of chemometric techniques (stepwise discriminant analysis,
 526 partial least-squares analysis for classification PLS, linear discriminant analysis LDA, principal
 527 components analysis PCA) and the cross-sectional variations in wood. The methodology developed
 528 relied on multiple independently constructed sub-models (one model per taxonomic level).

529 After analysing these examples, it seems that the development of more thoughtful methodologies
 530 rooted on the structure of the problem and the expert knowledge produces less linear pathways to
 531 the solution in contrast to applications that are less targeted and rely solely on the data analytics
 532 scheme and an emphasis on the collection of big amounts of data coupled with rigorous
 533 mathematical approaches.



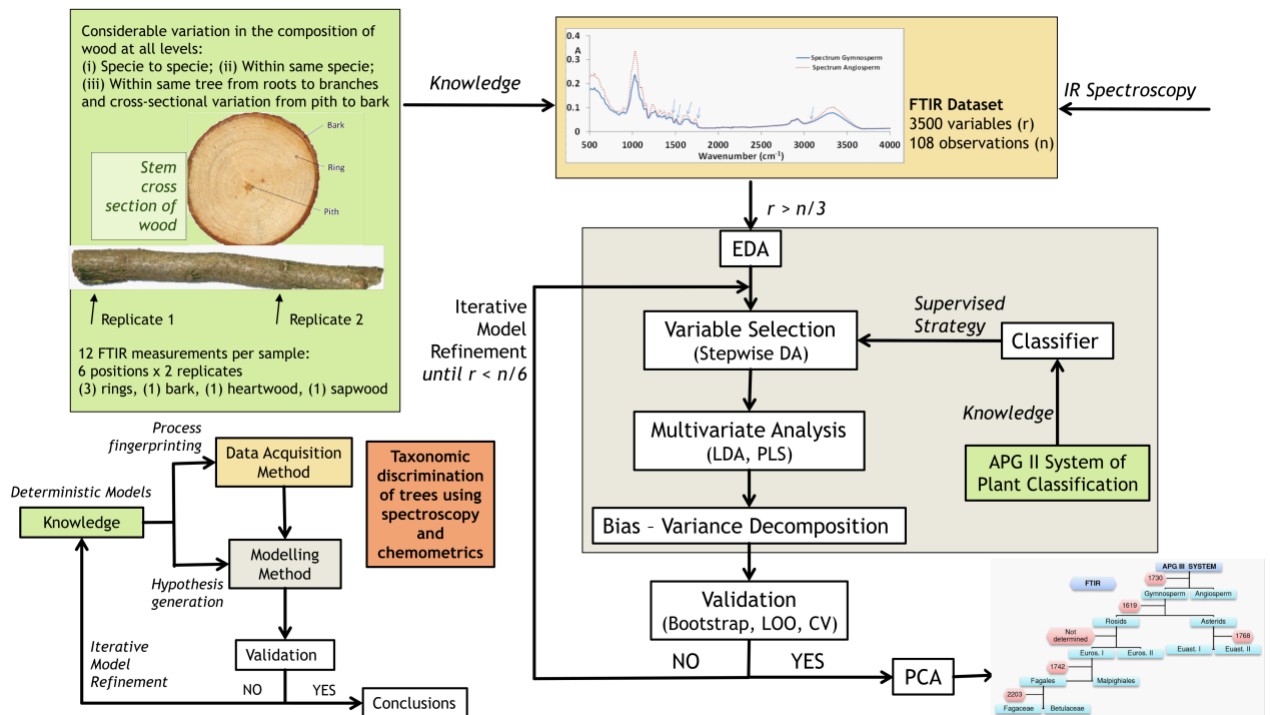
534

535 *Figure 5 Meta-Analysis implementing multivariate analysis with supervised learning [39]*



536

537 Figure 6 Meta-Analysis implementing ML algorithm on secondary data obtained from bibliometric method (this work)



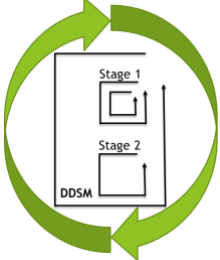
538

539 Figure 7 Multivariate analysis for the classification of wood [40], [41]

540

541

Table 1 Comparison between different statistical mixed methodologies for article processing and application in applied science domains (EMA = Exploratory meta-analysis, MA = Multivariate analysis, ML = Machine Learning)

	RESEARCH	HYPHOTHESIS	
	Systematic Review ⁽¹⁾ method to identify manuscript data suitable for inclusion in the meta-analysis	Meta-Analysis ⁽²⁾ method with the secondary data obtained from the systematic review	Reference
Methodology	✔ Bibliometric method using a supervised learning technique	✔ EMA, ML, and validation	[Ours] Field: Chemical engineering Application: Determine the synthesis conditions of calcium carbonate to obtain pure polymorphs in the precipitation process Case of Study: Vaterite Synthesis
Algorithms	 <ul style="list-style-type: none"> Bibliographic coupling analysis ensembled with modularity clustering algorithm and network analysis Supervised technique: recursive modularity optimization by a classification method (keyword co-occurrence analysis coupled with a document tracer) 	<ul style="list-style-type: none"> EMA: density/box plots, several correlations Supervised ML: J48 decision tree and ensemble trees using bagging and boosting techniques Validation: DOE analysis 	
Merits	Manual screening of publications was reduced a 99% (from 2,209 to 30 articles) Time of completion of the automated systematic review per polymorph from the initial design, search, automated screening, eligibility and data extraction was less than three months (full time)		
Methodology	✔ Bibliometric method using a supervised learning technique	✔ EMA	[35] Field: Physiology, medicine Application: Determine the association between diabetes mellitus and new-onset atrial fibrillation
Algorithms	<ul style="list-style-type: none"> Keyword co-occurrence analysis ensembled with K-means clustering algorithm Supervised technique: maximum entropy classification algorithm with training set 	Polynomial regression analysis with Pearson's correlation, 2-tailed T-test	
Merits	Decreased time in manual identification of relevant studies Provided an objective criteria for the screening of scientific investigations Obtained a more robust effect estimate with the aggregated data Identification and selection of relevant articles using an intelligent automated approach Manual screening of publications was reduced an 87% (from 4,177 to 555 articles)		

Methodology	✔ Manual Screening	✔ EMA, multivariate analysis with supervised strategy, and validation	[34] Field: Heterogeneous catalysis Application: Identify the structure-activity relationships of a catalyst Case of Study: Oxidative coupling of methane (OCM reaction)
Algorithms	None (manual screening from different sources)	<ul style="list-style-type: none"> • <i>EMA</i>: density distribution • <i>MA</i>: Multivariate regression • Supervised technique: iterative model refinement by a classification method with nested classes • <i>Validation</i>: the influence of CO₂ predicted by the model was experimentally verified 	
Merits	Robust and statistically significant property-performance model despite the data heterogeneity and potential publication bias Obtained a generalized set of physico-chemical properties that discriminate between high- and low-performing OCM catalysts		
Technique	✔ Bibliometric method using a ML technique	✘ No Meta-Analysis	[42] Field: Scientometrics Application: Predict quantitatively the evolution of physics research Case of Study: Physics research
Algorithms	<ul style="list-style-type: none"> • Bibliographic coupling and co-citation analysis paired with modularity clustering algorithm and network analysis • Supervised technique: group evolution discovery (GED) method to track topic cluster evolution and learn a ML classifier 	None	
Merits	Used ML and network science together to predict the future of physics research at the community level		
Technique	✔ Manual screening	✔ EMA, univariate analysis	[43] Field: Chemical Engineering Application: Evaluate performance of GAC to remove micropollutants
Algorithms	None (manual screening from technical reports)	two-way ANOVA, breakthrough of adsorbers	
Merits	Used 44 studies with data from the adsorption of three micropollutants to evaluate the adsorption performance of granular activated carbon (GAC) in pilot- and large-scale plants		
Technique	✔ Manual screening	✔ EMA, bias assessment	[44] Field: Chemical Engineering Application: Effect of process variables on
Algorithms	None (manual screening from articles)	<ul style="list-style-type: none"> • <i>EMA</i>: box plot, histogram, scatter plot, forest plot • <i>Statistical heterogeneity</i>: I², Tau², Q-statistic 	

		<ul style="list-style-type: none"> • <i>Quality assessment</i>: Egger's test, Begg's test 	adsorption of dyes onto carbon materials
Merits	Used 87 studies to evaluate the adsorption capacity of dyes on different carbon-based materials		
Technique	✔ Manual screening	✔ EMA, multivariate analysis	[45]
Algorithms	None (manual screening from articles)	<ul style="list-style-type: none"> • <i>EMA</i>: Pearson's correlations • <i>MA</i>: Cluster analysis, PCA, non-metric multidimensional scaling 	Field: Environmental Science Application: Evaluate the heavy metal content of surface water bodies throughout the world
Merits	A multivariate analysis was applied to determine the sources of heavy metals in the water bodies Pollution indices were calculated to evaluate the overall surface water quality		
Technique	✔ Manual screening	✔ EMA, univariate analysis	[46]
Algorithms	None (manual screening from articles)	two-tailed Spearman correlation analysis, nonparametric tests (Mann-Whitney U, Kruskal-Wallis H)	Field: Environmental Science Application: Evaluate the organic contaminants in Chinese sludge
Merits	Used 159 relevant papers for the meta-analysis of 35 classes of chemicals in sewage sludge		

3.5 Applications

In industrial manufacturing and at the basic level, process analytical technology (PAT) corresponds to the application of real-time instrument analysis and chemometrics for process understanding and control but it can also involve relationships among multiple data streams, typically process instrument data, engineering variables and analytical laboratory reference data [47]. In the latest case, the combination of multiple streams of data for process control and understanding is also called multivariate statistical process control or monitoring (MSPC or MSPM). Statistical process control (SPC) is one of the major parts of statistical quality control (SQC) devoted to the collection, analysis, and interpretation of data for use in quality control and deal with the technical aspects of quality in manufacturing [48]. Some emerging trends on quality engineering aims to incorporate information technology and improve quality engineering tools and techniques. Automated SPC systems based on pattern recognition (e.g. neural networks and fuzzy sets) and AI have also been developed as part of technological developments that allows for rapid data acquisition, analysis, and a more significant process improvement [48] but new MSPC methods are needed to digest industrial process data with hundreds of variables [49]. Data-driven method for batch data analysis for process monitoring and quality prediction are also part of the new data-driven paradigm [50].

DDSM could be useful for gaining understanding in process optimization at the early phase of a process development to find the most suitable operating conditions. It gathers valuable information about a greater number of attributes, wider ranges and creation of new features for the analysis of multiple effects and interactions. This application would be limited to processes for which information is already available or the processes share similarities and not applicable for entirely new ones.

3.6 Limitations

At the outcome level, the risk of bias was not calculated. However, we mentioned in the previous section some of the benefits of this methodology to deal with this particular issue. At the review level, the incomplete retrieval of research could affect the results slightly or moderately depending on how much relevant articles were omitted. Although special care was taken during the data collection process, human error cannot be entirely ruled out. Data collection was done manually from pdf articles. In this regard, it would have been easier if the numbers were embedded in tables rather than in the body of text. Current results might fluctuate slightly or moderately depending on how many errors were present (e.g. misplaced classification of some values, wrong numerical recording of some other values). Well-performing algorithms were used in this study, however the inclusion of more data in the training set is likely to produce better results than good algorithms [31]. Extrapolation of the present work should not be done if the studies include other type of experimental conditions outside the ranges mentioned in this work. Generalizations to other types of conditions would require first their addition to the model and a reassessment of the operating conditions. Finally, the method relies more on statistical approaches than on engineering knowledge. Nonetheless, engineering expertise was essential all along the journey.

3.7 Future work

We are providing an at-line implementation whereby the article's data is sampled manually with an offline algorithm methodology. A more advance approach would be an in-line/on-line method that allows the automatic coupling of primary data from scientific publications to create secondary datasets using several data processing tools. Given the huge volume of publications already in circulation and the slow development of DDSM, a combination of different implementation approaches is likely to coexist depending on the complexity of the project, the research question, field of study, volume of data, type of application, and so on.

The DDSM was an iterative procedure with recursive components where recursion was used at the stage level and at the data mining level within the first stage. It remains an open question if multiple scientific methods could be thought exploiting this feature systematically. Recursion is a property of the human thought that we apply creatively in a broad range of domains such as language, theory of mind, counting, computation, use and manufacture of tools, and modern technology [51].

It is expected that in future, the article elements such as the numerical data within the materials and method section will be stored in a standard formats to allow automatic retrieval and processing [52]. As it was beautifully stated in a recent industry-led review focusing on advanced information processing to grow the AI industry in the UK [10]: "...for published research the right to read is also the right to mine data". There is a trend of opening access to research data but most of the research data in circulation is not accessible in machine readable formats and, therefore, cannot be mined easily because it is restricted by contractor or copyright or because the articles are not accessible in standard formats, or both. Advanced search and retrieval technology based on AI will be develop to facilitate data manipulation [53].

4 Conclusions

A hybrid method that combines several information and data science techniques was developed and applied with success to solve a conventional experimental problem in engineering. Information and data science algorithms were combined to predict successfully the outcome of an experiment in chemical engineering. The success of this approach opens the door for its application and adaptation to different engineering optimization problems. We believe in the potential of statistical mixed methodologies to solve engineering problems. Although, we also realize it comes associated with many challenges. Studies of secondary data allows the formulation of different research questions and creates new connexions between the information than can lead to the discovery of new insights. True innovation will not come from tweaking slightly previous experiments and researchers should explore other ways of interacting with the available scientific resources. We presented here one idea of many that are attainable and encourage researchers to get inspired and explore more possibilities.

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] R. K. Schutt and D. F. Chambliss, "Investigating the Social World: The Process and Practice of Research," in *Making Sense of the Social World: Methods of Investigation*, 2011.
- [2] D. F. Penson, J. T. Wei, and L. J. Greenfield, *Clinical research methods for surgeons*. 2007.
- [3] R. Yousefi Nooraie, J. E. M. Sale, A. Marin, and L. E. Ross, "Social Network Analysis: An Example of Fusion Between Quantitative and Qualitative Methods," *J. Mix. Methods Res.*, vol. 14, no. 1, pp. 110–124, 2020.
- [4] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2012.
- [5] X. Wu and V. Kumar, *The Top Ten Algorithms in Data Mining*. Chapman & Hall/CRC, 2009.
- [6] L. Mutihac and R. Mutihac, "Mining in chemometrics," *Analytica Chimica Acta*. 2008.
- [7] S. Succi and P. V. Coveney, "Big data: The end of the scientific method?," *Philosophical*

- Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. 2019.
- [8] K. H. Trzesniewski, M. B. Donnellan, and R. E. Lucas, *Secondary Data Analysis: An Introduction for Psychologists*. American Psychological Association, 2011.
- [9] T. P. Vartanian, *Secondary Data Analysis*. 2011.
- [10] W. Hall and J. Pesenti, "Growing the artificial intelligence industry in the UK," 2017. [Online]. Available: <https://www.gov.uk/government/publications/growing-the-artificial-intelligence-industry-in-the-uk>. [Accessed: 11-Jul-2021].
- [11] D. Donoho, "50 Years of Data Science," *Journal of Computational and Graphical Statistics*. 2017.
- [12] E. Smith, *Using Secondary Data in Educational and Social Research*. Open University Press, 2008.
- [13] A. Jinha, "Article 50 million: An estimate of the number of scholarly articles in existence," *Learn. Publ.*, vol. 23, no. 3, pp. 258–263, 2010.
- [14] J. A. Evans, "Electronic publication and the narrowing of science and scholarship," *Science (80-)*, 2008.
- [15] D. Ezer and K. Whitaker, "Data science for the scientific life cycle," *Elife*, vol. 8, pp. 1–10, 2019.
- [16] E. O. Voit, "Perspective: Dimensions of the scientific method," *PLoS Computational Biology*. 2019.
- [17] P. M. K. Illari, F. Russo, and J. Williamson, *Causality in the Sciences*. 2011.
- [18] D. Moher *et al.*, "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement," *PLoS Medicine*. 2009.
- [19] L. Brecevic and D. Kralj, "On Calcium Carbonates : from Fundamental Research to Application," *Croat. Chem. Acta*, vol. 80, pp. 467–484, 2007.
- [20] 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.
- [21] K. Börner, C. Chen, and K. W. Boyack, "Visualizing knowledge domains," *Annu. Rev. Inf. Sci. Technol.*, 2003.
- [22] M. J. Cobo, A. G. López-Herrera, E. Herrera-Viedma, and F. Herrera, "Science mapping software tools: Review, analysis, and cooperative study among tools," *J. Am. Soc. Inf. Sci. Technol.*, 2011.
- [23] N. J. van Eck, L. Waltman, R. Dekker, and J. van den Berg, "A Comparison of Two Techniques for Bibliometric Mapping: Multidimensional Scaling and VOS," *J. Am. Soc. Inf. Sci. Technol.*, vol. 61, no. 12, pp. 2405–2416, 2010.
- [24] J. J. Ng and K. H. Chai, "A bibliometric analysis of Project Management research," *IEEE Int. Conf. Ind. Eng. Eng. Manag.*, vol. 2016-Janua, no. 2002, pp. 976–980, 2016.
- [25] Sci2 Team, "Science of Science (Sci2) Tool." Indiana University and SciTech Strategies, 2009.
- [26] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update," *SIGKDD Explor.*, vol. 11, no. 1, 2009.

- [27] J. H. E. Cartwright, A. G. Checa, J. D. Gale, D. Gebauer, and C. I. Sainz-Díaz, "Calcium carbonate polyamorphism and its role in biomineralization: How many amorphous calcium carbonates are there?," *Angewandte Chemie - International Edition*, vol. 51, no. 48, pp. 11960–11970, 2012.
- [28] M. Farhadi-Khouzani, D. M. Chevrier, P. Zhang, N. Hedin, and D. Gebauer, "Water as the Key to Proto-Aragonite Amorphous CaCO₃," *Angew. Chemie - Int. Ed.*, vol. 55, no. 28, pp. 8117–8120, 2016.
- [29] 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.
- [30] H. Canada and C. Nadeau, "Inference for the Generalization Error," *Heal. (San Fr.)*, no. 514, pp. 1–49, 2001.
- [31] I. H. Witten, E. Frank, and M. A. Hall, *Data mining*. 2011.
- [32] K. Dickersin, "Publication Bias: Recognizing the Problem, Understanding Its Origins and Scope, and Preventing Harm," in *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*, 2006.
- [33] T. Kimura and N. Koga, "Monohydrocalcite in comparison with hydrated amorphous calcium carbonate: Precipitation condition and thermal behavior," *Cryst. Growth Des.*, vol. 11, no. 9, pp. 3877–3884, 2011.
- [34] 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.*, 2019.
- [35] Z. Xiong *et al.*, "A machine learning aided systematic review and meta-analysis of the relative risk of atrial fibrillation in patients with diabetes mellitus," *Front. Physiol.*, 2018.
- [36] S. Hafeez *et al.*, "CO₂ capture using membrane contactors: a systematic literature review," *Frontiers of Chemical Science and Engineering*. 2021.
- [37] R. E. Baker, J. M. Peña, J. Jayamohan, and A. Jérusalem, "Mechanistic models versus machine learning, a fight worth fighting for the biological community?," *Biol. Lett.*, vol. 14, no. 5, pp. 1–4, 2018.
- [38] G. H. Yann LeCun, Yoshua Bengio, "Deep learning," *Nature*, 2015.
- [39] A. Carballo-Meilan *et al.*, "Biosorption of copper using nopal fibres: moolooite formation and magnesium role in the reactive crystallization mechanism," *Cellulose*, 2020.
- [40] A. Carballo-Meilan, A. M. Goodman, M. G. Baron, and J. Gonzalez-Rodriguez, "A specific case in the classification of woods by FTIR and chemometric: Discrimination of Fagales from Malpighiales," *Cellulose*, 2014.
- [41] A. Carballo-Meilán, A. M. Goodman, M. G. Baron, and J. Gonzalez-Rodriguez, "Application of chemometric analysis to infrared spectroscopy for the identification of wood origin," *Cellulose*, 2016.
- [42] W. Liu, S. Saganowski, P. Kazienko, and S. A. Cheong, "Predicting the evolution of physics research from a complex network perspective," *Entropy*, vol. 21, no. 12, pp. 1–22, 2019.
- [43] F. Benstoem *et al.*, "Performance of granular activated carbon to remove micropollutants from municipal wastewater—A meta-analysis of pilot- and large-scale studies," *Chemosphere*. 2017.

- [44] A. Azari, R. Nabizadeh, S. Nasser, A. H. Mahvi, and A. R. Mesdaghinia, "Comprehensive systematic review and meta-analysis of dyes adsorption by carbon-based adsorbent materials: Classification and analysis of last decade studies," *Chemosphere*. 2020.
- [45] V. Kumar *et al.*, "Global evaluation of heavy metal content in surface water bodies: A meta-analysis using heavy metal pollution indices and multivariate statistical analyses," *Chemosphere*. 2019.
- [46] X. Z. Meng *et al.*, "Organic Contaminants in Chinese Sewage Sludge: A Meta-Analysis of the Literature of the Past 30 Years," *Environmental Science and Technology*. 2016.
- [47] K. A. Bakeev, *Process Analytical Technology: Spectroscopic Tools and Implementation Strategies for the Chemical and Pharmaceutical Industries: Second Edition*. 2010.
- [48] A. Hassan, M. S. N. Baksh, and A. M. Shaharoun, "Issues in quality engineering research," *International Journal of Quality and Reliability Management*. 2000.
- [49] F. A. P. Peres and F. S. Fogliatto, "Variable selection methods in multivariate statistical process control: A systematic literature review," *Comput. Ind. Eng.*, 2018.
- [50] R. Rendall, L. H. Chiang, and M. S. Reis, "Data-driven methods for batch data analysis – A critical overview and mapping on the complexity scale," *Computers and Chemical Engineering*. 2019.
- [51] M. C. Corballis, "The Uniqueness of Human Recursive Thinking," *American Scientist*. [Online]. Available: <https://www.americanscientist.org/article/the-uniqueness-of-human-recursive-thinking>.
- [52] H. Juergens *et al.*, "Evaluation of a novel cloud-based software platform for structured experiment design and linked data analytics," *Sci. Data*, 2018.
- [53] S. Viswanath, J. W. Fennell, K. Balar, and P. Krishna, "An Industrial Approach to Using Artificial Intelligence and Natural Language Processing for Accelerated Document Preparation in Drug Development," *J. Pharm. Innov.*, 2021.