ACCEPTED MANUSCRIPT

CO2 Reduction Beyond Copper-Based Catalysts: A Natural Language Processing Review From The Scientific Literature

Lucas Bandeira, Henrique Ferreira dos Santos, James Moraes de Almeida, Amauri Jardim de Paula, Gustavo Martini Dalpian

ACS Sustainable Chemistry & Engineering, 12, 11 (2024)

DOI: https://doi.org/10.1021/acssuschemeng.3c06920

This document is the Accepted Manuscript version of a Published Work that appeared in final form in ACS Sustainable Chemistry & Engineering, copyright © 2024 American Chemical Society, after peer review and technical editing by the publisher. To access the final edited and published work see https://doi.org/10.1021/acssuschemeng.3c06920.

CO_2 Reduction Beyond Copper-Based Catalysts: A Natural Language Processing Review From The Scientific Literature

Lucas Bandeira[a], Henrique Ferreira[a,b], James Moraes de Almeida[c], Amauri Jardim de Paula[c], and Gustavo Martini Dalpian[*,d,a]

Abstract: Carbon dioxide (CO_2) is a prominent greenhouse gas contributing significantly to global warming. To combat this issue, one strategy is the conversion of CO_2 into alcohols and hydrocarbons, which can be utilized as fuels and chemical feedstocks. Consequently, a substantial volume of scientific literature has been dedicated to investigating different materials and reaction conditions to facilitate the CO_2 reduction reaction (CO_2RR) into these so-called high-value products. However, the vastness of this literature makes it challenging to stay updated on recent discoveries and review the most promising materials and conditions that have been explored. To address this issue, we applied natural language processing (NLP) tools to extract valuable data from 7292 published articles in the scientific literature. Our analysis revealed the emergence of new materials such as cesium-lead-bromide perovskites and bismuth oxyhalides, that have been recently used to CO_2RR, and identified Bi-based catalysts as the most selective for $HCOO^-$ production. Furthermore, we gleaned insights into the composition of other elements and materials commonly employed in CO_2RR, their relationship to product distribution, and the prevalent electrolytes used in CO_2 electrochemical reduction (CO_2ER). Our findings can serve as a foundation for future investigations in the realm of catalysts for CO_2 reduction reactions, offering insights into the most promising materials and conditions to pursue further research.

Keywords: CO_2 Reduction Reaction, Natural Language Processing, Data Analysis, Photocatalysis, Electrocatalysis

Synopsis: CO_2 is the greenhouse gas that most contributes to global warming. Much research has been published focusing on CO_2 conversion. We use natural language processing to gain insights from this literature.

1 Introduction

Carbon dioxide (CO_2) is one of the greenhouse gases that most contribute to global warming[??]. During the last decades, its concentration in Earth's atmosphere has substantially increased, resulting in a temperature rise of approximately 1ºC[??]. Models presented in the 2021 IPCC report suggest that if the climate policies are kept the same, the average temperature is predicted to increase by 2.5 to 2.9ºC by 2100, causing deleterious effects on living beings[??]. In 2021, fossil fuels (including oil, natural gas, and coal) accounted for 77% of global energy consumption. Moreover, between 2020 and 2021, there was a staggering 17% increase in fossil fuel consumption[?]. One of the directions for tackling global warming and our dependency on fossil fuels is to convert the emitted CO_2 in the so-called high-value products, a process named CO_2 reduction reaction (CO_2RR). Over the past years, the scientific community has devoted considerable attention to this area, leading to a wealth of published articles that explore diverse materials and methodologies aimed at establishing a closed CO_2 cycle[????]. Figure 1 (yellow) depicts the rising trend in works regarding CO_2 and CO reduction reaction as topic.

One of the greatest challenges in CO_2RR is the inherent stability of carbon dioxide molecules. The energy required to break the C=O bond is about 750 kJ/mol, about twice the value for forming C−H, C−C, and C−O bonds[??]. Consequently,

[a] *Federal University of ABC, Av. dos Estados, 5001 - Bangú, Santo André - SP, 09280-560, Brazil.*
[b] *Paulista Faculty of Administration and Informatics, Av. Lins de Vasconcelos, 1222, São Paulo - SP, 01538-001, Brazil.*
[c] *Ilum School of Science (CNPEM), Rua Lauro Vannucci, 1020 - Fazenda Santa Cândida, Campinas - SP, 13087-548, Brazil.*
[d] *Institute of Physics, University of São Paulo, R. do Matão, 1371 - Butantã, São Paulo - SP, 05508-090, Brazil.*
[*] Corresponding author. E-mail: dalpian@if.usp.br

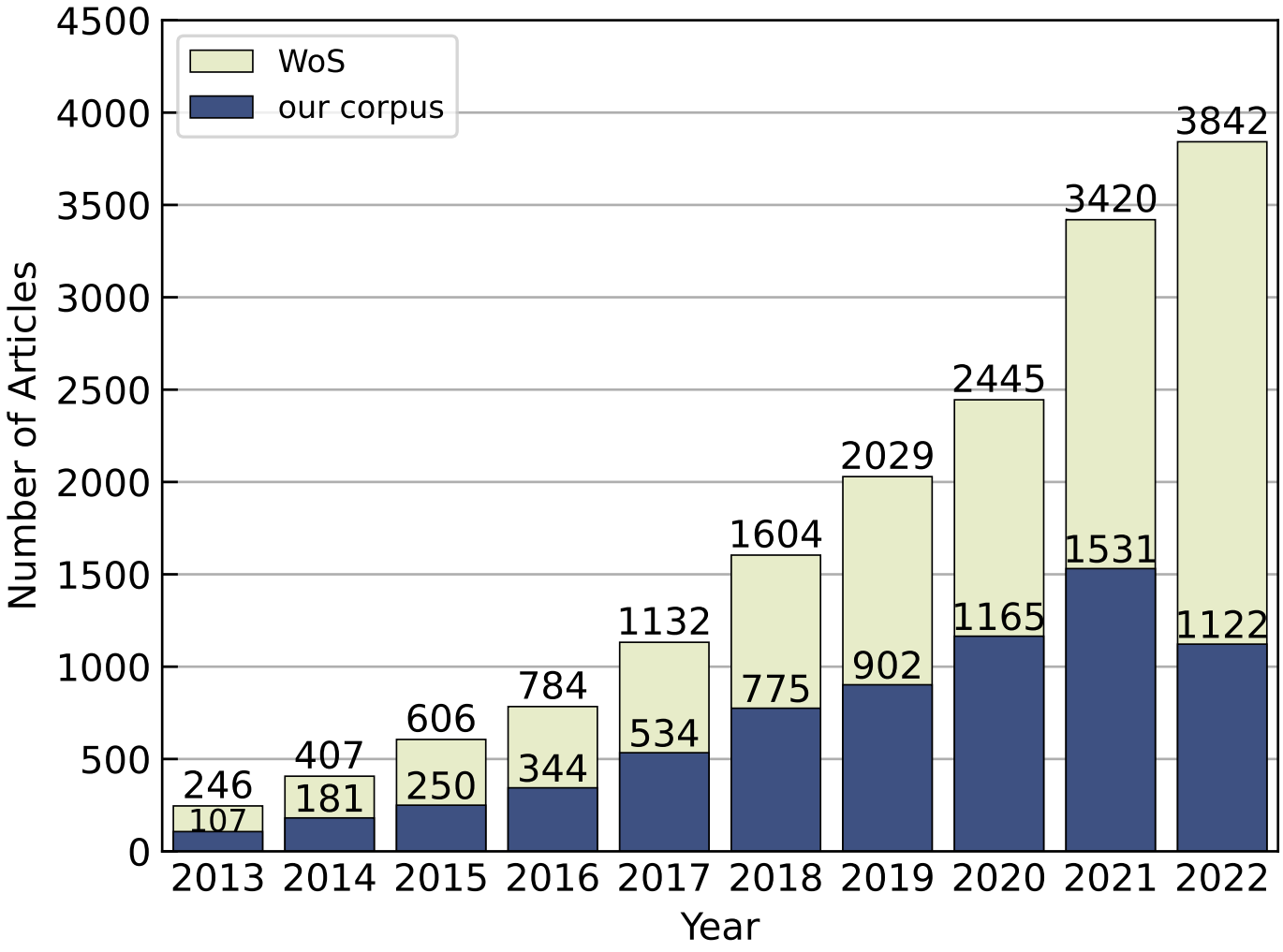

Figure 1 (Yellow) number of articles published and (blue) number of articles in our corpus over the past 10 years. Results were obtained by querying the Web of Science using the terms: *("CO2 *reduction"* OR *"CO *reduction"* OR *"carbon dioxide *reduction"* OR *"carbon monoxide *reduction")* AND *"*cataly**"* as topic.

the application of a potential is necessary for the reaction to occur. Furthermore, the hydrogenation of CO_2 can yield multiple products, limiting the efficiency of the process[??], and the competition with the hydrogen evolution reaction (HER) further complicates the process. These issues have drawn the attention of a considerable amount of research regarding CO_2RR[???]. The goals are i) to reduce the energy required for the reaction to happen, ii) to improve the selectivity of the process, and iii) to decrease the H_2 formation. These conditions are reflected in the following metrics measured in laboratories and commonly reported in the literature: the applied potential and the light intensity associated with the energy furnished to the system, the faradaic efficiency (FE), and the formation rate of a catalyst toward a specific product, indicating the selectivity of the reaction. Different approaches have been utilized for this task, each with its own advantages and limitations, resulting in distinct performances and outcomes. Keeping track of these metrics is paramount to understanding how research has progressed over the years, and providing some guidance for future works in this field.

As shown in Figure 1 (yellow), the number of publications involving CO_2RR has increased yearly. This extensive literature makes reviewing and staying updated a challenging task. The most comprehensive a manually written review article can be, they just provide a fraction of the information presented in the literature, comprising at most a few hundred references[???]. This limitation is understandable since an exhaustive manual analysis of such large literature would be intractable. To address this challenge, alternative approaches utilizing Natural Language Processing (NLP) tools, such as text mining, have emerged. NLP is an interdisciplinary field of computer science and artificial intelligence that concerns how to use machines to analyze and understand natural (human) language[??] in text and/or audio. Its applications in computer science date back to the mid-50s with Allan Turing proposing what is known today as the Turing test[?], but only recently they have been expanded to materials science due to advancements in machine learning (ML) techniques, allowing for automatic extraction of relevant and high-quality information dug in the scientific literature.

One of the most important tasks in NLP is named entity recognition (NER), i.e., to understand the semantic role of a term within a sentence. It is the pivot of relation extraction, which aims to establish connections between entities in a sentence according to their semantic function in texts. This is the core task in materials science once it enables machines to recognize materials, reaction conditions, reagents, and products and reliably extract relations between them. NER tools can be classified into three categories: Dictionary-lookup, rule-based, and ML-based[??]. The first consists of a dictionary with a set of terms to be exactly matched in texts. This method is accurate; however, it requires a hugely comprehensible dictionary to extract a significant amount of data. Rule-based approaches loosen some restrictions for capturing information by setting rules that need to be satisfied rather than strict terminology. They are less accurate than the former strategy, but they facilitate extractions once they only demand some hand-crafted conditions to be fulfilled. Finally, ML-based methods are the most cutting-edge algorithms employed in NLP pipelines. They can collect information from text without a dictionary of terms or rules but instead from vector semantics[?]. This approach is faster and can extract much more data compared to the previous ones; however, they require large amounts of data to be trained and parameters to be adjusted, being more computationally expensive.

Although relatively new in the scientific domain, these tools have already demonstrated their utility in materials science and other scientific fields by creating datasets[??], providing insights into synthesis routes[??], and developing materials- and chemistry-aware language models[???]. Thus, we employed these NLP tools to gain insights from the scientific literature pertaining to CO_2RR and CORR. By selecting relevant articles through a query in the Web of Science, we were able to uncover valuable details about commonly used catalyst elements, their combinations in materials, faradaic efficiencies reported in the literature for each element-based catalyst, the relationship between catalyst composition and reaction selectivity, frequently used electrolytes in CO_2 electro-reduction reaction (CO_2ER), and how these electrolytes are utilized in combination with the previously identified catalyst elements. Furthermore, by tracking the use of these materials over the years, we could identify trends in the literature and also prominent compositions that have gained the attention of researchers in recent times.

2 Methods

Our method primarily relies on a rule-based strategy utilizing regular expressions and a small dictionary to capture the desired information. These approaches strike a balance between high accuracy obtained through manual extractions and reduced labor facilitated by automated tools. Furthermore, the rules in our pipeline are section-specific, meaning that we extract data from different sections of a paper that we deemed relevant for obtaining the desired information. This helps minimize incorrect extractions and improves the reliability of our approach. Additionally, when collecting numerical data, we utilize the `GPT3.5-turbo API`[?] to verify the obtained relation, rather than relying solely on human review. By selecting specific sentences, we assessed the model's precision, recall, and F1-score before proceeding with further analyses. For some cases where we observed the GPT model was not accurate, we resorted to a manual verification to remove erroneous relations. This rigorous evaluation helps to ensure the accuracy and effectiveness of our strategy. It is worth noting that similar approaches have been attempted recently in the field of CO_2RR, but with a narrower focus on Cu-based catalysts for CO_2ER and relying on only a few articles[??], or using these NLP tools in a general fashion[?]. Unlike these previous analyses, our approach considers a larger corpus and has a broader scope: CO_2 catalysts beyond

Cu-based materials. By expanding the range of materials under investigation, we were able to uncover novel materials that have gained attention in the field of CO_2RR and CORR. Additionally, we employed a different strategy to leverage large language models (LLMs). While these tools offer powerful language processing capabilities, there is a concern regarding the generation of inaccurate or false information, commonly referred to as "hallucinations". However, by limiting their responses and providing appropriate context, we could efficiently mitigate these risks. Figure 2 charts the pipeline adopted in our work.

2.1 Journal Articles Acquisition

The first step in our pipeline was to gather the articles containing the relevant information concerning CO_2/CO reduction reaction into high-value products. On July 23rd, 2022, we queried the Web of Science[?] for publications including the following terms as keywords, either in its title or abstract:

*(CO2 *reduction* OR *CO *reduction* OR *carbon dioxide *reduction* OR *carbon monoxide*reduction)* AND *cataly**.

This query returned 15,559 articles distributed over a myriad of publishers. To refine our dataset, we focused on the top 5 publishers with the highest number of publications: Elsevier, ACS, RSC, Wiley, and Springer. This narrowed down our options to 12,452 articles. Using the DOIs acquired in this search, we downloaded Elsevier publications through its Application Programming Interface (API)[?] and ACS files by direct contacting the organization. Altogether, these two publishers provided a corpus of 7292 papers spanned from 1996 to 2022 in the extensible markup language (XML) format. We chose XML due to its lower susceptibility to conversion errors compared to the portable document format (PDF)[??]. Figure 1 (blue) shows the distribution of the papers in our corpus over the years, starting from 2013. The complete distribution starting from 1996 is depicted in Figure S1.

2.2 Sentence Segmentation and Section Identification

We used four Python libraries, `BeautifulSoup`[?], `re`[?], `regex`[?] and `chemdataextractor` (`CDE`)[??] for splitting the files in our corpus into sentences and identifying sections from which they were extracted in the articles. The latter step is particularly important as some of the information we intend to gather may predominantly reside in specific sections, and relying on data from other sections could potentially yield inaccurate outcomes. For instance, when examining the frequency of elements employed in catalysts, we exclusively considered abstracts as they are more direct, concise summaries of the article's scope and avoid connections with the literature (see subsection 2.3). Including the introduction section could add false counts to a particular element since it might be cited solely for contextualization. Similarly, adding the discussion section could also lead to false positives once authors often compared their results with the existing literature. We employed `BeautifulSoup` to convert the markup language into plain text, `CDE` to segment sentences in a paragraph into individual units, and `re` and `regex` to write regular expressions (*regexes*) to identify the section headings. We applied these expressions to the section titles identified by `BeautifulSoup` through the tag `<ce:section-title>`. We notice here that we did not use regexes for abstracts. This section has a specific tag in XML files, either `<ce:abstract>` or `<abstract>`, depending on the publisher. We assigned a label "abstract", "introduction", "materials-methods", "discussion", or "conclusion" to every sentence according to the section from which it was extracted. Those to which the section headings were not identified were labeled as "None" and overlooked in our analyses. This process yielded a dataset comprising over 1.4 million sentences, distributed over the five categories aforementioned.

2.3 Information Extraction and Data Analysis

We conducted different sorts of information extraction according to the sections from which sentences were identified. Starting with the abstract, we gleaned data about the metallic elements most used in catalysts for CO_2RR. We selected this section as its sentences tend to be shorter, direct (without comparison with the literature), and positive (mentioning the materials used rather than those not used). We determined the top 10 most common metal elements employed by the number of articles, i.e., how many articles mentioned a catalyst based on that element. It is important to note that, for each element, we only considered whether it was present or not in the abstract but not the frequency it appeared; thereby, if an element was mentioned ten times within a single abstract, it would only be counted once. Additionally, aiming to capture only elements in catalysts, we discarded any composition after the symbols "@", "/", and "-", or mentioned after the prepositions "on" and "at", so these results do not account for elements used in supports.

In a similar analysis, we identified the top ten electrolytes commonly utilized in carbon dioxide electro-reduction (CO_2ER). Nevertheless, unlike the previous analysis that relied on abstracts, we exclusively focused on the experimental section. This decision was based on heuristics since details about reaction conditions are typically described in this section and oftentimes dismissed in the abstract unless the work emphasizes the impact of electrolytes on CO_2ER. To validate our approach, we utilized the manually extracted data from[?]. By comparing our results with their analysis, we were able to assess the accuracy and reliability of our methodology.

To gain a more comprehensive understanding of the literature, we integrated the aforesaid analyses to understand how those electrolytes are combined with the selected elements. Since all the information required is seldom available in a single sentence, our strategy involved combining relevant data based on file names. In other words, we considered a pair *element-electrolyte* only if both element and electrolyte were extracted from the same file. Furthermore, we also analyzed the co-occurrence of elements and electrolytes with CO_2RR and HER products. We investigated how frequently they are positively mentioned with the products CH_4, CH_3OH, C_2H_4, C_2H_5OH, C_3H_7OH, HCOOH, $HCOO^-$, H_3CCOOH, H_3CCOO^-, C_3H_8, C_4H_{10}, C_4H_9OH, CO and H_2. By "positive" we mean that sentences that do not contain negative terms such as "no", "don't" or "didn't". We intentionally disregarded sentences such as "we observed **no** methane production during the reaction" to avoid false positives in pairs *element-product* or *electrolyte-product*. Although it is not guaranteed that a particular element can favor a specific product simply by analyzing the frequency they co-occur, this result can provide an overview of the relationship between different elements or electrolytes and their impact on the selectivity of the reaction. We also notice here that this is the only step in our pipeline in which we used a dictionary lookup, to match product formulae.

We continued extending our extractions now incorporating the faradaic efficiency towards each product, but focused solely on elements in catalysts. This type of information is often encountered in the results and discussion sections. Therefore, we carried out these extractions uniquely considering sentences labeled as "discussion" by our previous approach. To avoid capturing misleading relations, we devised a set of rules to be fulfilled in every single sentence: (i) The same number of materials, products, and FE values should be reported, (ii) the same number of materials and FEs should be reported for a single CO_2RR product, or (iii) the same number of FEs and CO_2RR products should be reported with a single material composition. In addition, the sentence should explicitly contain the term "faradaic efficiency" or "FE" and a value reported with "%" as a unit. Finally, to ensure the quality of the data we employed the GPT API[?] on the selected sentences to verify the veracity of the relation extracted. We utilized the model `GPT3.5-turbo` providing the following prompt:

> *Is the information that a FE of* `[value]` toward `[product]` *was obtained using a* `[element]`*-based catalyst present in this sentence* `[sentence]`*? 1 (for "Yes") or 0 (for "No")*

where `[value]`, `[product]`, `[element]` and `[sentence]` are variables containing the value of the faradaic efficiency extracted, the target product, one of the ten elements obtained in our first analysis, and the sentence from which the relation was extracted. We set all parameters to zero except the maximum token length, which we set to 1, thereby, drastically reducing the probability of the model "hallucinating".

If one considers our text corpus with 1.4 million sentences, it is virtually impossible to identify the number of sentences that contain certain information we want to extract (e.g, the catalyst definition or the FE value). Although it is necessary to determine this number to directly calculate recall and precision values for our NLP approach, its most accurate calculation would require human reading of the entire corpus, which is unfeasible. In this context, we used a benchmark model (`GPT3.5-turbo`) for indirectly evaluating precision and recall. To ensure the reliability of this strategy, we selected 20 sentences per element to manually confirm the data extracted. By contrasting our labels with those proposed by `GPT3.5-turbo`, we computed the precision, recall, and F1-score of the model, defined as

$$Precision = \frac{TP}{TP+FP}; \tag{1}$$

$$Recall = \frac{TP}{TP+FP}; \tag{2}$$

$$F1_{score} = \frac{2 \times Recall \times Precision}{Precision + Recall}. \tag{3}$$

where TP (true positives) and TN (true negatives) indicate labels predicted by the GPT model that matched our analysis, and FP (false positives) represent labels assigned as positive by the model, but as negative in our manual analysis. For the particular case where we observed `GPT3.5-turbo` did not perform as accurately, we resorted to a manual analysis to remove erroneous relations obtained. Guaranteed the effectiveness of our approach and the quality of our dataset, we utilized sentences recognized as relevant in the downstream task of linking elements, products, and FEs from the papers in our corpus.

Additionally, we explored the composition of materials employed in CO_2RR. Using the same regular expressions to match elements and specific rules to avoid wrong extractions, we determined how elements co-occur in binary and ternary compounds. We selected the most prominent results from each category and search for these materials in reviews in the literature regarding this topic. In cases we did not encounter any discussion about the compositions identified in the checked review articles, we analyzed the sentences extracted in order to gain some insights on the application of these

materials in carbon dioxide reduction. Finally, we investigated how the number of publications about these materials has changed over the years. Applying this strategy we were able to identify promising structures that only recently have gained attention in the literature regarding CO_2RR.

3 Results and Discussion

3.1 Elements in CO_2RR Catalysts and Electrolytes in CO_2ER

Initially, we focused on extracting information regarding metallic elements most employed in catalysts for CO_2 reduction reaction from abstracts of the papers in our corpus. By using a set of regular expressions to identify elements and clean the sentences, we selected the top 10 most frequent metals, depicted in Figure 3(a). We remember here these results account for materials solely employed in catalysts but not in supports. It can be seen that Cu and Ti are the most common elements encountered. Copper is widely recognized as a highly promising material for electrochemically converting carbon dioxide into hydrocarbons and alcohols[???], while titanium serves as a reference material for CO_2 photo-reduction[???]. Additionally, we observed significant counts for Ni, Ag, and Bi. These elements are also frequently employed in carbon dioxide reduction both in the electrochemical and photochemical approaches. Nevertheless, it is interesting to note the growing trends in Bi-based catalysts in this field, and that our approach was able to identify it.

Using another set of regexes, we collected information regarding electrolytes used in CO_2ER. As electrolytes are essential for any electrochemical reaction, this information is typically presented in the methodology section of the articles, rather than the abstract, except when the work focuses on the impact of salts in CO_2ER selectivity. To prove this statement, we attempted this extraction solely on sentences from the abstract. This result is shown in Figure S2. It can be seen that just a small amount of sentences comprised the desired information. Figure 3(b) exhibits the results obtained with our extractions. As it can be seen, $KHCO_3$ emerged as the most commonly employed electrolyte in CO_2ER, followed by Na_2SO_4, and KOH. These salts represent different reaction conditions utilized in CO_2 electroreduction, specifically neutral, acidic, and alkaline media. In fact, they are the most commonly used in each of these conditions[?], with $KHCO_3$ being the most frequent electrolyte overall. Additionally, we conducted a comparison of Figure 3(b) with the manually extracted results published in[?] which also focused on the same objective, electrolytes employed in CO_2ER. Notably, the top five electrolytes identified in their analysis are also present in our top five, albeit in a different order. This correspondence indicates the accuracy of our automated approach in extracting relevant information. Nevertheless, it is important to highlight that our approach encompasses a significantly larger dataset, comprising six times more papers than the study mentioned, and considers not only copper-based catalysts but a broader range of catalysts. Thereby, our analysis provides a more comprehensive view of the literature and better indicates trends and insights in this field.

Afterward, we performed a third analysis by combining the data from Figure 3(a) and 3(b) to gain insights into the utilization of electrolytes based on the catalyst composition. Since the likelihood of finding all this information in a single sentence is low, we linked the data using the files' IDs. The underlying concept behind this strategy is as follows: If Article A mentions a Cu-based catalyst and CH_4 as a product in the abstract, and describes $KHCO_3$ as the electrolyte in the methodology, it is highly probable that $KHCO_3$ was employed as the electrolyte for methane production. Figure 3(c) illustrates the frequency of each electrolyte in Figure 3(b) co-occurring with the elements in Figure 3(a), normalized by the total frequency of each element. Notably, $KHCO_3$ is the electrolyte most used with all the elements considered, except for Ti and Pt. These metals are more frequently associated with the use of Na_2SO_4 as the electrolyte, which might indicate a preference for acid media conditions.

Our approach also provided valuable insights into product selectivity based on the presence of specific elements in a catalyst. We conducted two types of analyses using different sections of the articles: the abstract and the discussion. In the first, we simply examined positive co-occurrence between catalyst elements and products of CO_2RR and HER. The result of this analysis is portrayed in Figure 4(a); the numbers in the heatmap indicate the frequency of a pair *element-product* co-occurs in the abstract. We initially searched for CH_4, CH_3OH, C_2H_4, C_2H_5OH, C_3H_9, C_3H_7OH, HCOOH, $HCOO^-$, H_3CCOOH, H_3CCOO^-, C_4H_{10}, C_4H_9OH, and H_2. Nevertheless, we had only one occurrence for C_3H_7OH and none for C_3H_9, C_4H_{10}, and C_4H_9OH. It is important to note that these compounds are indeed generated in CO_2RR but usually in small quantities. Consequently, they are not commonly mentioned in abstracts, where only major products are typically cited. Furthermore, we excluded CO as a target product in this analysis. This choice was based on the nature of our corpus, which consists of articles focusing on carbon dioxide and carbon monoxide reduction. Therefore, including CO as a product would introduce a bias towards higher frequencies for *element-CO* pairs. We conducted a similar analysis for *electrolyte-product* pairs. These results are shown in Figure S3.

From Figure 4(a) we can see that either H_2 or CH_4 are the products that most co-occur with all the elements, with one notable exception: bismuth, primarily associated with $HCOO^-$. Moreover, it is interesting to note that bismuth exhibits the lowest co-occurrence with H_2, followed by copper. Copper is widely recognized as an excellent catalyst for carbon dioxide due to its lower selectivity towards the HER compared to other metals. Our findings suggest the possibility that bismuth might be even less favorable to HER, and potentially more efficient to CO_2 reduction. These observations indicate that our analysis may have captured valuable insights into the product selectivity associated with different catalyst elements.

We expanded our investigations to consider the FE towards products of CO_2RR on different element-based materials. To gather this information, we relied solely on the discussion section, where most of the numerical data is presented. We also employed `GPT3.5-turbo` API to check whether the relation extract was correct or not. However, before we continued with this approach, we selected about 20 sentences for each element to manually evaluate the accuracy of the data collected. Based on our labels, we evaluate the model performance in identifying correct extractions. The precision, recall, and F1-score obtained were 0.84, 0.96, and 0.90, which are very high scores (the maximum is 1.0 for each one). Therefore, we decided to keep with our approach and neglected sentences labeled as 0 (i.e., wrong) by `GPT3.5-turbo`. Nevertheless, for the particular case of Ag, we notice our regexes capture also relations where "Ag" were extracted from "Ag/AgCl", a reference electrode. Even though we devised our regexes to ignore such cases, few of them were matched. Nevertheless, the GPT model was not able to identify some of these errors. Therefore, specifically for Ag, we manually remove these inaccurate extractions. After checking the quality of the dataset, we utilized sentences recognized as relevant in the downstream task of linking elements, products, and FEs from the papers in our corpus. The results of this analysis are depicted in Figure 4(b), where the numbers in the heatmap indicate the median of the Faradaic efficiency extracted for an *element-product* pair, and "N/A" means no such relation was found. Boxplots illustrating the range of FE collected and the number of papers from which this information was extracted are presented in Figure S4. Additionally, we point out that, different than our previous analysis in Figure 4(a), we included CO as a product in both Figure 4(b) and Figure S4 since the set of rules devised are satisfactory for not capturing carbon monoxide as a reactant.

The analysis of Figure 4(b) reveals that CO and H_2 exhibit the highest median faradaic efficiencies (FE), which is true for most of the elements. This observation can be attributed to the fact that CO is the primary reduction product of CO_2 and reducing water requires less energy compared to CO_2. This trend is further supported by the data presented in Figure S4, which demonstrates that the majority of the collected data corresponds to these two products. However, contrasting this trend in Figure 4(b), we can notice the FE values towards HCOOH, and H_3CCOOH for Cu and Ti, and $HCOO^-$ for Zn and Bi that surpass those obtained for CO and H_2. It is important to consider that, on closer examination of Figure S4, the high FE values for HCOOH and H_3CCOOH are based on a single data point, rendering these results statistically unreliable. On the other hand, for Zn, the FE in Figure 4(b) is the median of eight data points, which, although still a relatively small sample size, can be considered statistically reliable. Nevertheless, for Bi, the relatively high median FE of 86% towards formate ($HCOO^-$), is supported by 67 data points. The abundance of data and the elevated FE suggest a potential selectivity of Bi and Bi-based materials in promoting formate production ($HCOO^-$), which aligns with similar reasoning applied in the analysis of Figure 4(a). These findings related to bismuth may offer an explanation for the increased application of Bi in CO_2RR over the years, as evidenced in Figure S2. It is important to mention that reproducibility is indirectly assessed in these results by the dispersions of FE values reported in the literature for each catalyst/product pair. The median is a statistical metric that determines the 50% separation point of the data. For example, the FE value of 86% for Bi in the reaction towards formate ($HCOO^-$) indicates that 50% of the articles report an FE greater than 86% for this reaction, while the other 50% report an FE lower than 86%. In addition, when there is a low standard deviation in the distribution of FE values for a specific catalyst/product pair it means that there is convergence in the values reported in the literature. This also occurs for the Bi/$HCOO^-$ pair. Finally, the skewness of a distribution of FE values for a specific catalyst/product pair can also be analyzed. As an example, the dispersion for FE values of the Bi/$HCOO^-$ pair is negatively skewed, thus indicating that a measured FE value most likely will be higher than the value expressed by the median for this pair (86%).

3.2 Materials Composition Commonly Employed in CO_2RR

Using similar regexes as those described in subsection 3.1 to match elements and clean the sentences, we conducted an analysis on how elements co-occur in binary and ternary compounds for carbon dioxide electro-reduction, once more solely focusing on abstracts. The chord diagrams in Figure 5 illustrate the top ten binary compositions (upper panels) and the top 5 ternary systems (lower panels) that most frequently appeared in our search. The other 5 most common compositions in ternary materials are portrayed as chord diagrams in Figure S5. From Figure 5(a), we observe a significant presence of oxides in binary systems, particularly for Ti and Cu. TiO_2 is a well-known reference material used in carbon dioxide photo-reduction. Its wide bandgap facilitates carbon dioxide reduction in addition to water oxidation, despite being cheap, stable, and non-toxic[?] . CuO or Cu_2O, on the other hand, can be used in either CO_2 electro- or photo-catalysis, however, being most famous for its outstanding performance in CO_2ER[?] .

Furthermore, it is worth noticing that almost every metal in Figure 5(a) occurs as oxide. This is not surprising considering the abundance of oxygen on Earth, which contributes to the stability and cost-effectiveness of these materials. However, it is interesting to observe that Mo and Cd do not co-occur with oxygen, but exclusively with sulfur. This result is highlighted in Figure 5(b) where we overlook O-containing materials. In this modified chord diagram, we can clearly see that sulfur co-occurs with most of the elements, not only Mo and Cd but also Zn and Cu. We notice that the utilization of metal sulfides for CO_2RR is a common strategy described in the literature, being MoS_2 and CdS the main materials in this subgroup[??] . Nevertheless, as mentioned by them, ZnS is emerging as a promising material for carbon dioxide photo-reduction. Despite, S–containing materials, we can also identify TiC in Figure 5(b), with a significant amount of

occurrences. We searched for a material containing these elements in different review papers about CO_2RR[??], but we could not find any specific mention of this composition. However, upon analyzing our extracted sentences, we discovered that this composition refers to Ti_3C_2, an MXene material that has recently garnered attention in the field of carbon dioxide photo-reduction.

On the other hand, Figure 5(d) provides insights into the elements that frequently co-occur in ternary materials employed in CO_2RR. Similarly to Figure 5, we can observe significant occurrences of oxides, primarily BiVO and BiOBr. BiVO from bismuth vanadates ($BiVO_4$) are commonly cited as alternative materials to TiO_2 in CO_2 photo-reduction, once it allows for the same reactions to occur, but in contrast, are activity at visible light[??]. Similarly, BiOBr appears to be another promising material for CO_2RR. Although we extensively searched for information about this composition in several review papers[????], we could not find any specific mention of it. Therefore, we manually examined the sentences selected in our analysis to gain some insight into this composition. BiOBr, named bismuth oxybromide, is a semiconductor that "has obtained wide attention because of its unique narrowed band-gaps, low cost, non-toxicity as well as higher structure stability"[?]. Additionally, during our examination of the review papers, we came across a reference to a similar material, BiOCl. Bismuth oxychloride, in addition to bismuth oxyiodide (BIOI), has also a high-frequency score in our analysis, as portrayed in Figure S5, however, not as high as BiOBr. Also interesting is the fact we did not find any mention of BiOF.

Nonetheless, the compositions with the highest scores in Figure 5(d) are CsPbBr and NaBH. We reproduced this diagram in Figure 5(e), but excluding O-containing materials for a better view. The NaBH composition originates from $NaBH_4$, which is a commonly used reducing agent in materials synthesis. We found it quite interesting that this compound appeared in our results, considering that we exclusively analyzed the abstracts of the papers in our corpus. We expected this sort of compound to be present in the methodology section. Once again, we investigated review articles for the applications of this salt, but we did not find any specific mentions of it. By analyzing in detail the extracted sentences, we confirmed that most occurrences referred to the use of $NaBH_4$ as a reducing agent in materials synthesis. It was interesting to find, however, that this compound has already been directly applied to reduce carbon dioxide, a process called CO_2 hydroboration[??????]. It is worth noting that only a few articles mentioned the application of sodium borohydride for this purpose, with the majority of them published between 2014-2016.

CsPbBr, on the other hand, belongs to the significant family of halide perovskites, which has been extensively studied for its applications in solar cells. However, it seems that these materials have only recently gained attention in CO_2RR. When searching for these structures in reviews about CO_2 reduction, we just found brief mentions of it[??]. Nevertheless, we did come across reviews specifically discussing the application of these structures in CO_2 photo-reduction, indicating their potential as promising catalysts in this field. Interestingly, we noted that these reviews were all published after 2020[???]. Upon analyzing the sentences extracted from our corpus, we discovered that the first mention of this structure was in a paper published in 2017 by Xu et al.[?], which coincidentally was also cited in the review by Wu et al.[?]. Actually, according to the own authors of this paper "is the first report on artificial photosynthesis based on halide perovskite QDs". Therefore, it is likely that our corpus includes one of the early papers that explored the application of halide perovskites in carbon dioxide reduction.

Additionally, we investigated binary and ternary compounds in CO_2RR but exclusively considering cations. This allowed us to gain insights into the metal alloys employed, as shown in Figure 5(c) and 5(f). In Figure 5(c), we can observe the prevalence of Cu as the metallic element that most appear in alloys, followed by Ni and Pd. Numerous articles in the literature have explored the alloying of Cu with different metals to modulate the faradaic efficiency (FE) and applied potential for CO_2RR[????]. Nevertheless, the highest frequency scores occur for NiFe and PdAu alloys. NiFe catalysts are commonly utilized in the methanation reaction, to hydrogenate carbon monoxide into methane[?]. Papers in our corpus, however, mention the application of this material for both CORR and CO_2RR. AuPd alloys, on the other hand, are also known compositions for CO_2RR[?]. One of the articles in our corpus mentions that "Pd_1Au_{24} can convert CO_2 to CO with $\tilde{1}$00% faradaic efficiency" at applied potentials ranging from -0.6 to -1.2V vs RHE[?]. However, we are surprised these alloys have a higher citation score than any Cu-based materials. Another interesting result in Figure 5(f) is that copper only occurs in a single composition, NiFeCu, despite having a lower score than CoAlLa and CoNiPt. Among the identified sentences, CoAlLa is predominantly mentioned as a layered double hydroxide (LDH). These structures are cited in one of the reviewed articles[?] as a strategy for designing catalysts for CO_2RR. However, it is worth noticing that the papers in our corpus that mention CoAlLa were all published after 2019. Conversely, publications on PtCoNi alloys are slightly older, with the most recent one published in 2020, and all of them primarily employed PtCoNi alloys for the hydrogen evolution reaction (HER).

Finally, we performed an annual analysis of the articles referring to the four compositions mentioned earlier (TiC, CsPbBr, BiOBr, and NaBH), which were not found in the reviewed articles. The results are presented in Figure 7; the values over the bars indicate the percentage of papers in our corpus published in the corresponding years that cited the respective material, (a) for CsPbBr, (b) for BiOBr, (c) for TiC, and (d) for NaBH. The trends observed in these plots are very interesting. For CsPbBr, the first article on this structure was published in 2017, and since then, the number of papers involving this material has steadily increased. The publication count almost doubled from 2021 to 2022 and more

than doubled compared to 2020. This indicates a significant upward trend for the utilization of CsPbBr in the literature. A similar pattern can be observed for BiOBr, with the number of publications more than doubling from 2021 to 2022. However, unlike the previous result, the publication count from 2016 onwards has not shown a consistent upward trend. We complemented this analysis including all the bismuth oxyhalides semiconductors. This result is shown in Figure S6(a) and as can be seen, the number of publications using these materials has also grown over the years. When a material shows a rise in the number of publications, it indicates there is a growing interest around it. We understand that the pure number of publications does not mean that the material is efficient per se. However, it is very unlikely to observe a rising trend in the number of published articles with negative results, such as those reporting non-reproducible synthesis methods or materials with poor properties.

In Figure 7(c), we can observe a significant increase in the number of papers regarding TiC starting from 2019. The publication count almost tripled from 2019 to 2020 and doubled from 2020 to 2021. However, in 2022, there was a decrease in the percentage of publications. We notice that in 2009, publications about TiC accounted for 3% of the total literature on CO_2RR (Figure S6(b)), but from 2009 onwards to 2014, we did not find any further mentions of TiC in the articles included in our corpus. However, our corpus comprises only 33 papers published in 2009, as can be seen in Figure S1. This value of 3.03% accounts for a single publication. Therefore, we can say that TiC has indeed recently gained attention in CO_2RR. Regarding NaBH, from Figure 7(d) we can observe that there is a decreasing trend in the number of publications about this compound, starting from 2014, when it accounted for 2.76% of the papers. The percentage steadily declined to 0.52% in 2018, and since then, has remained relatively constant, representing around 0.50% of the papers published. It is interesting to note that most of the articles employing NaBH for direct CO_2 reduction were published between 2014 and 2016, with a higher concentration in 2014, indicating a specific trend during that period, which may explain the highest score in that particular year.

3.3 Conclusions

Our comprehensive pipeline for extracting various types of information from scientific publications focused on the reduction of CO_2/CO into high-value products. The pipeline encompasses the extraction of categorical and numerical data, specifically the composition of catalysts, the utilization of electrolytes in CO_2ER, and the faradaic efficiency (FE) towards different products of CO_2RR. Furthermore, we combined these extracted data to gain valuable insights from the literature and conducted an annual analysis to identify trends in this research field. Our findings suggest that bismuth (Bi) exhibits promise as a material for hydrogenating CO_2 into formate ($HCOO^-$). Bi is frequently mentioned alongside formate and demonstrates a significantly high median FE while displaying a distinct pattern of least co-occurrence with hydrogen (H_2). Additionally, we discovered that $CsPbBr_3$ and bismuth oxyhalides are novel structures that have recently gained attention in the field of CO_2 reduction. The number of publications involving these materials has shown a notable increase over the years, indicating their growing significance in the research community. Our analyses demonstrate the utilization of natural language processing (NLP) tools for investigating CO_2RR and other catalyzed reactions. By employing rule-based techniques, such as regular expressions and dictionary lookups, we strike a balance between automated analysis to reduce workload and the precision achieved through manual extractions. The high-quality data obtained using these methods can be leveraged for various NLP tasks. It is worth noting that while we adopted these particular strategies, alternative NLP methodologies also exist. To the best of our knowledge, our work extends previous research by integrating multiple types of information to offer a comprehensive and conclusive review of the advancements in CO_2RR. Notably, our investigation incorporates a larger corpus and encompasses a broader range of target materials, enhancing the breadth and depth of our findings.

Notes

There are no conflicts to declare.

Supporting Information

Number of papers published in our corpus; Annual distribution of papers for each of the elements from Fig. 3a; Co-occurrence analysis between electrolytes and products; Range of faradaic efficiencies for each product for all elements on Fig. 3a; Chord diagram for supporting Fig 5(d,e,f); Annual analysis of occurrences for different materials; Statistical analysis of FE.

Acknowledgements

This work has been performed in the scope of the Materials Informatics INCT (CNPq-Brazil). The authors also thank Fapesp, FINEP and CAPES for financial support. LB was funded by the PRH-ANP-UFABC project. The CCM-UFABC is also acknowledged for the computational time. Finally, we thank the ACS, Elsevier, and Wiley for collaborating on our project, providing mass access to the papers.

409 *ForTableofContentsUseOnly*

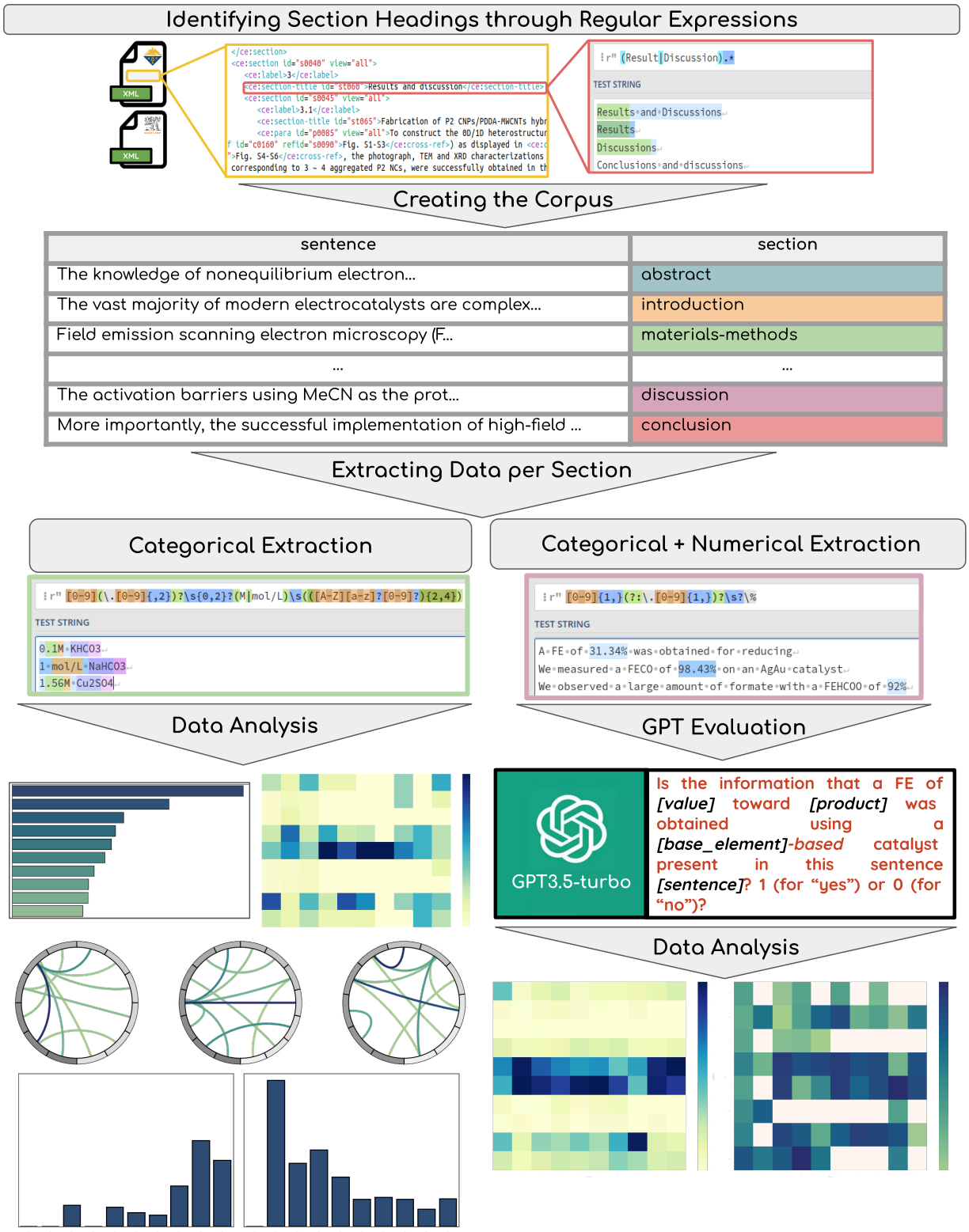

Figure 2 An illustration of the pipeline used in this work. First, we identified section headings in XML files using regexes and created a corpus with sentences separated according to the labels in the table (explained in subsection 2.2). Afterward, we applied regexes to extract information depending on the section sentences were removed. When gathering numerical and categorical data in long spans of sentences, we used `GPT3.5-turbo` API to evaluate whether the information was correctly extracted. Additionally, we combined the content spread over different sections of the same article by the files' ID to gain a broader understanding of the literature in this field.

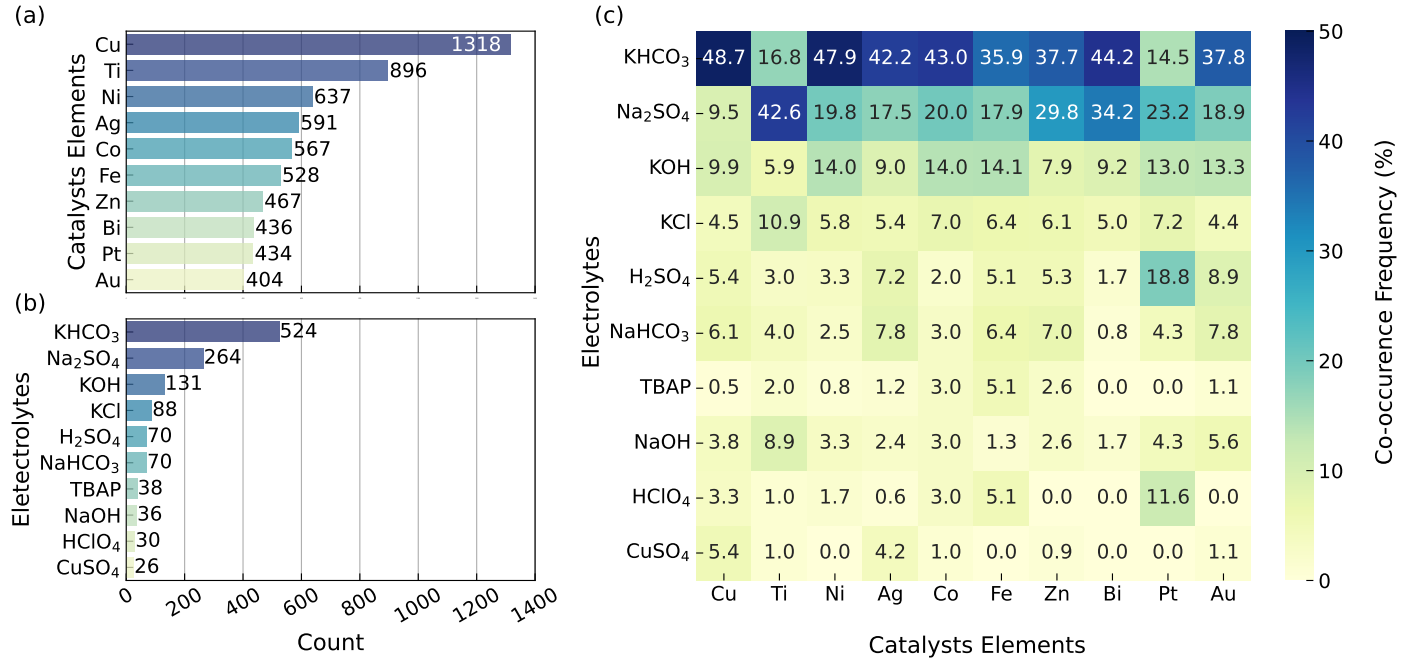

Figure 3 (a) Top ten elements most used in catalysts for CO_2RR. The numbers at the top of the bars indicate the number of papers mentioning a material containing that element. (b) Top 10 electrolytes most commonly employed in CO_2ER. The values above each bar indicate the number of articles mentioning the corresponding electrolyte. (c) Co-occurrence matrix of electrolytes and elements most used in catalysts. The frequency was normalized by the total elements count.

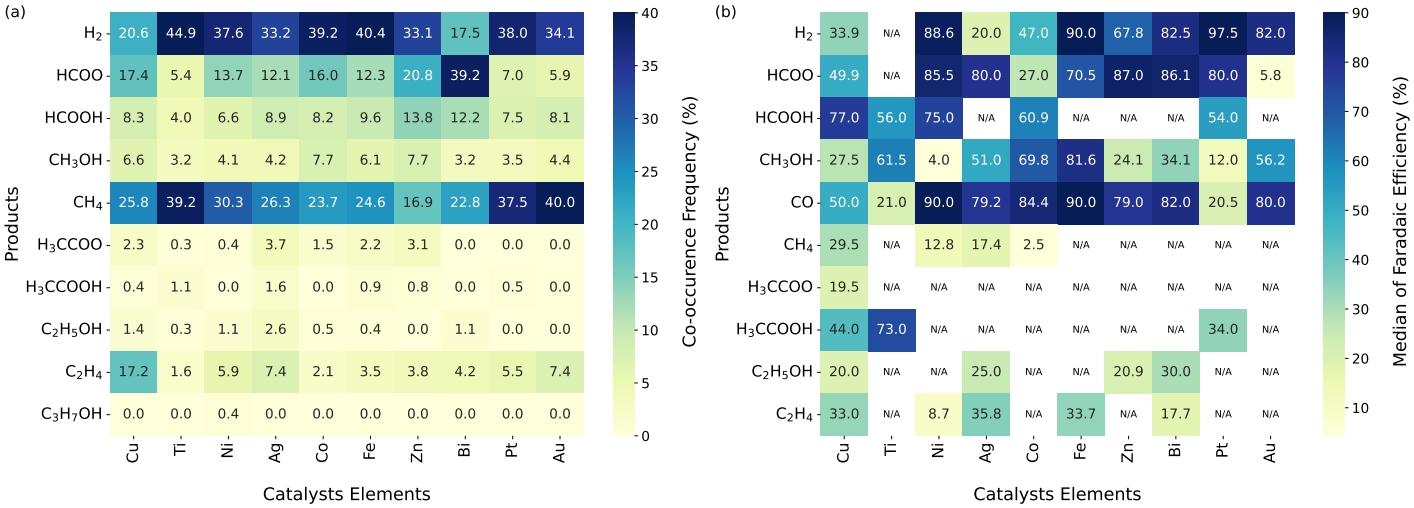

Figure 4 (a) Co-occurrence frequency of elements and CO_2RR products. (b) The median of FE. Cells without co-occurrence were marked with N/A.

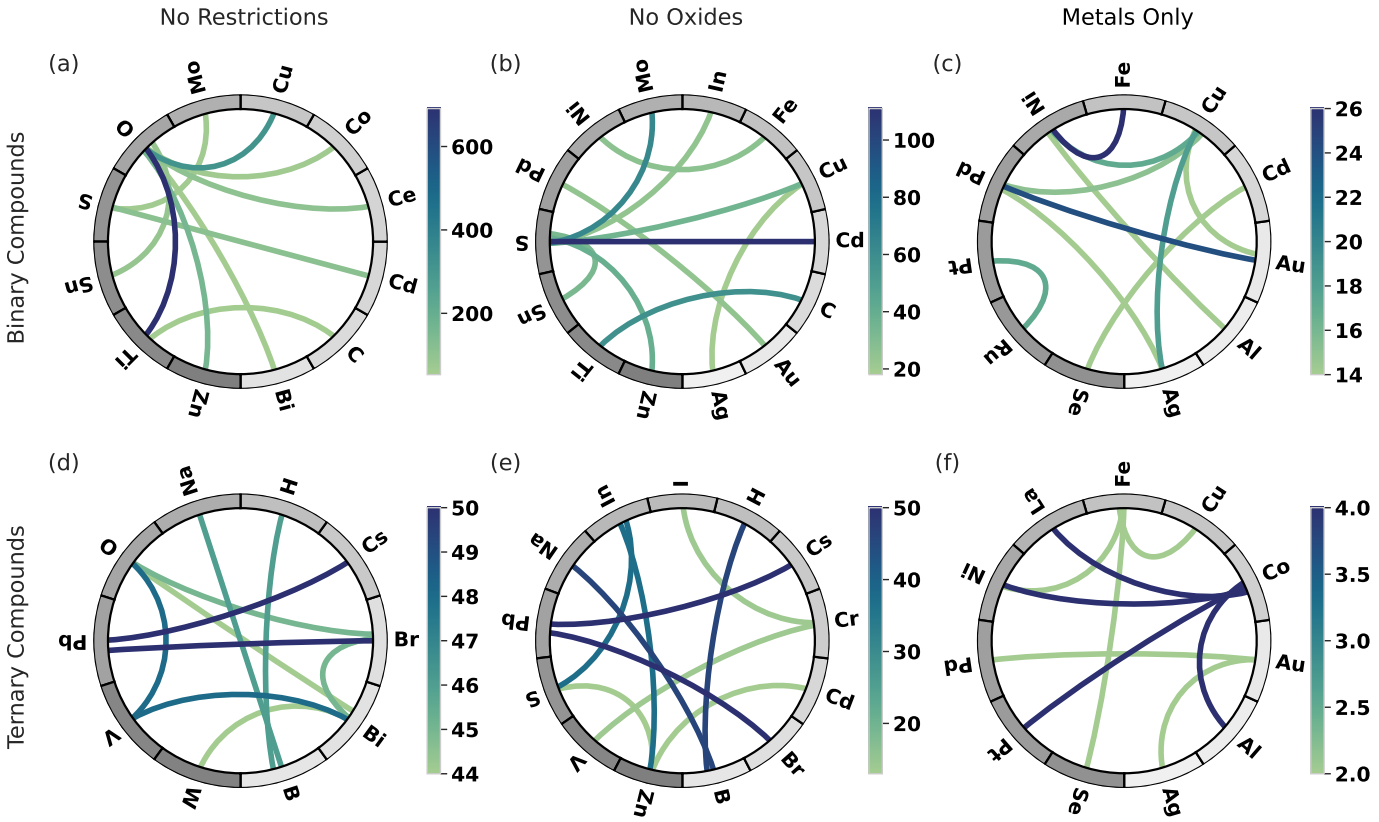

Figure 5 Co-occurrence of elements in binary (upper panels) and ternary (lower panels) compounds. Panels (a) and (d) show unrestricted element combinations, (b) and (e) consider only materials without oxygen, whereas (c) and (f) focus exclusively on cations. The colors indicate the frequency, measured by the number of articles, in which each set of elements appears together.

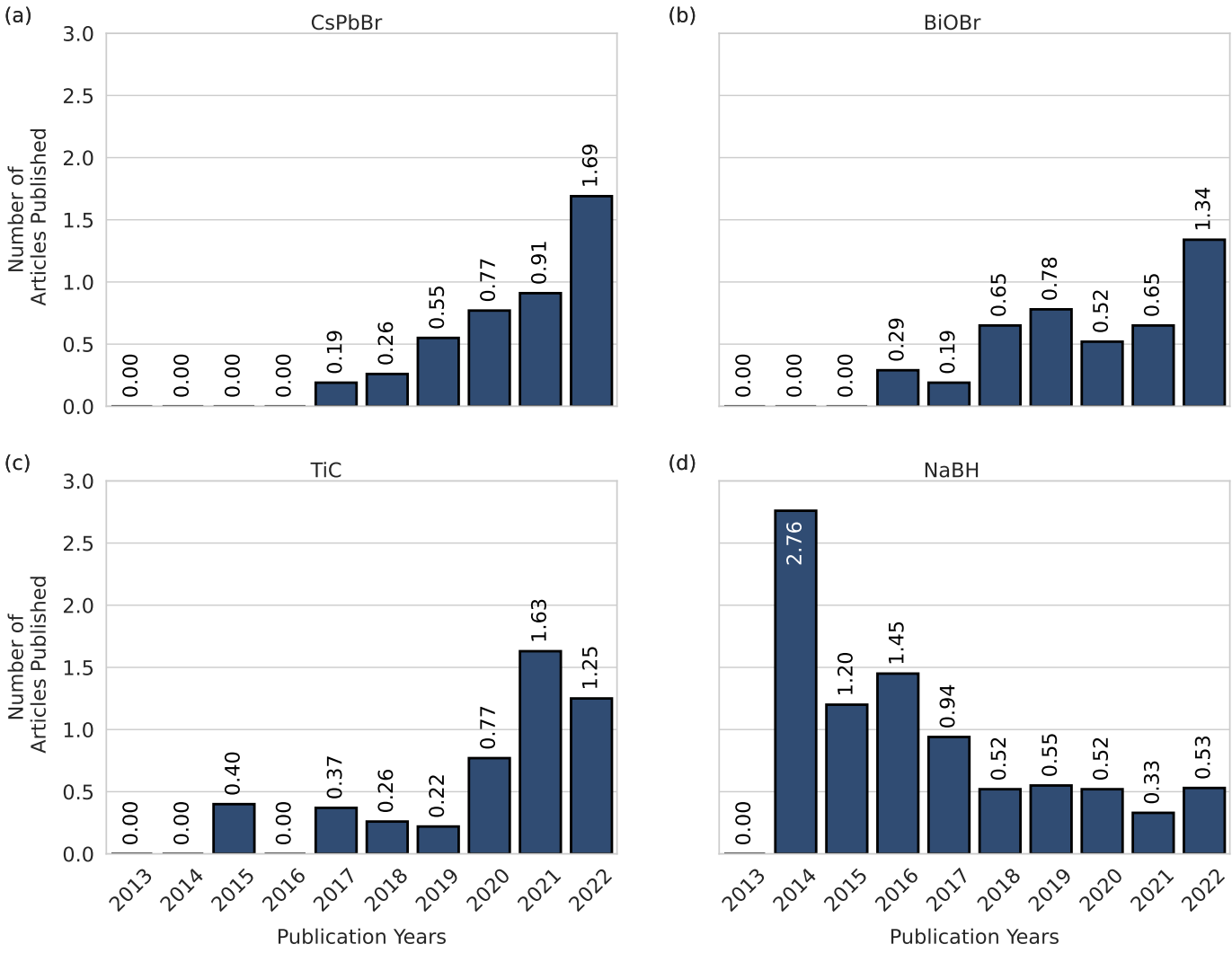

Figure 6 Percentange of paper in our corpus per year mentioning the compositions (a) CsPbBr, (b) BiOBr, (c) TiC, and (d) NaBH. We selected these compounds since we found no reference to them in the consulted review articles about CO_2RR.

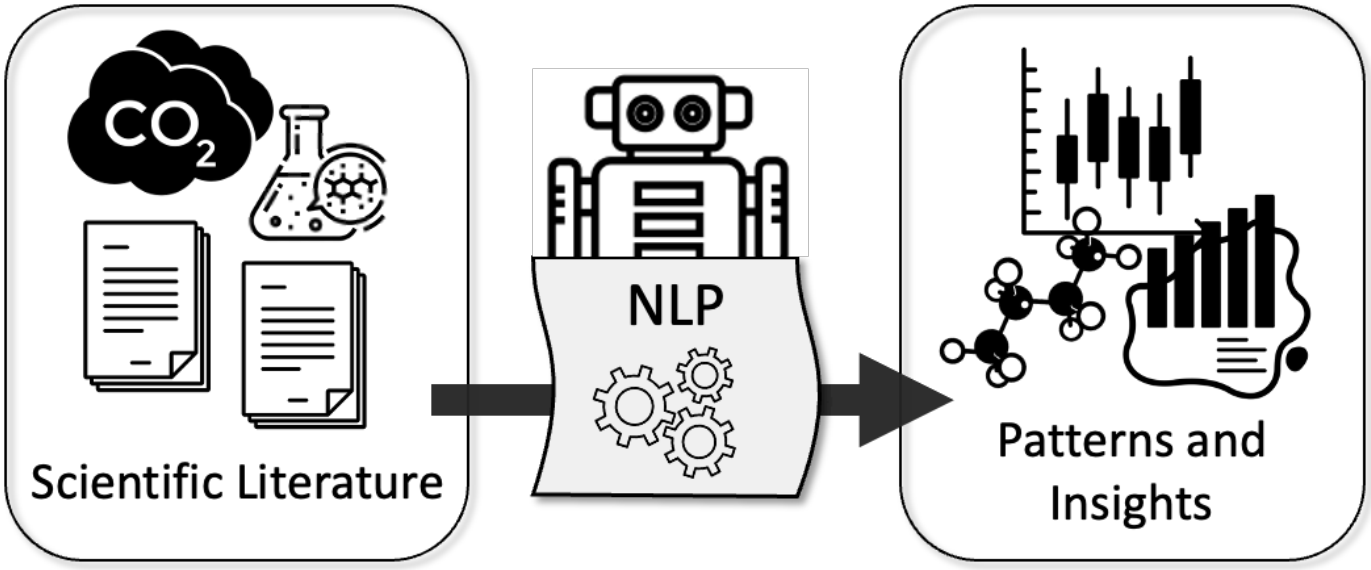

Figure 7 Table of Contents (TOC) graphic page.