What Can Text Mining Tell Us About Lithium-Ion Battery Researchers' Habits?

Hassna El-bousiydy, Teo Lombardo, Emiliano Primo, Marc Duquesnoy, Mathieu Morcrette, Patrik Johansson, Patrice Simon, Alexis Grimaud, Alejandro A. Franco

VIP *Very Important Paper*

Special Collection

What Can Text Mining Tell Us About Lithium-Ion Battery Researchers' Habits?

Hassna El-Bousiydy[+],[a, b] Teo Lombardo[+],[a, c] Emiliano N. Primo[+],[a, c] Marc Duquesnoy,[a, c] Mathieu Morcrette,[a, b, c] Patrik Johansson,[b, d] Patrice Simon,[b, c, e, h] Alexis Grimaud,[b, c, f, g] and Alejandro A. Franco*[a, b, c, h]

Artificial Intelligence (AI) has the promise of providing a paradigm shift in battery R&D by significantly accelerating the discovery and optimization of materials, interfaces, phenomena, and processes. However, the efficiency of any AI approach ultimately relies on rapid access to high-quality and interpretable large datasets. Scientific publications contain a tremendous wealth of relevant data and these can possibly, but not certainly, be used to develop reliable AI algorithms useful for battery R&D. To address this, we present here a text mining study wherein we unravel lithium-ion battery researchers' habits when reporting results, reason on how these habits link to issues of lacking reproducibility and discuss the remaining challenges to be tackled in order to develop a more credible and impactful AI for battery R&D.

1. Introduction

The development of rechargeable lithium-ion battery (LIB) technology constitutes one of the most emblematic success stories of deployment of materials science discoveries, leading to a societal change via enabling wide practical application of portable electronics.[1] The recent introduction and plans for numerous giga-factories to manufacture LIBs will further reduce their cost, much driven by the demand for electric vehicles. Yet, more efficient and faster R&D schemes are needed to further improve batteries performance, durability and safety, as well as to lower their manufacturing CO_2 footprint and increase their re-usability and recyclability. Even if LIBs have significantly improved since the first cells successfully commercialized by Sony[2] based on $LiCoO_2$ (LCO), they are still based on LCO-derived layered oxides such as $LiNi_xMn_yCo_zO_2$ (NMC), in spite of enormous materials research efforts. Since recent years, Artificial Intelligence (AI) is raising the battery community expectations to revolutionize the way we search for new materials, optimize interfaces and operation conditions, as illustrated by the roadmaps of several international research initiatives, such as the European Battery 2030+.[3]

Modern AI encompasses numerous types of computer algorithms such as supervised, unsupervised and reinforced machine learning (ML),[4–6] neural networks[7–9] and natural language processing.[10–12] They have tremendous capabilities of automatically mining and finding patterns in very large datasets (Big Data) revealing difficult-to-access information and propose solutions to complex problems. These capabilities have been demonstrated early on in many scientific and engineering fields,[13–18] and start to have significant impact in the battery R&D:[19] AI has proven to be useful for battery materials discovery,[20–23] electrolyte formulation (in combination with robotics),[24] electrode tomography image processing,[25] electrode design,[26] state of charge estimation,[27] aging prediction,[28–31] correlating manufacturing parameters to

[a] H. El-Bousiydy,[+] T. Lombardo,[+] Dr. E. N. Primo,[+] M. Duquesnoy, Dr. M. Morcrette, Prof. Dr. A. A. Franco
Laboratoire de Réactivité et Chimie des Solides (LRCS), UMR CNRS 7314
Université de Picardie Jules Verne, Hub de l'Energie
15, rue Baudelocque, 80039 Amiens Cedex 1, France
E-mail: alejandro.franco@u-picardie.fr
[b] H. El-Bousiydy,[+] Dr. M. Morcrette, Prof. Dr. P. Johansson, Prof. Dr. P. Simon, Dr. A. Grimaud, Prof. Dr. A. A. Franco
ALISTORE-European Research Institute, FR CNRS 3104, Hub de l'Energie
15, rue Baudelocque, 80039 Amiens Cedex 1, France
[c] T. Lombardo,[+] Dr. E. N. Primo,[+] M. Duquesnoy, Dr. M. Morcrette, Prof. Dr. P. Simon, Dr. A. Grimaud, Prof. Dr. A. A. Franco
Réseau sur le Stockage Electrochimique de l'Energie (RS2E), FR CNRS 3459, Hub de l'Energie
15, rue Baudelocque, 80039 Amiens Cedex 1, France
[d] Prof. Dr. P. Johansson
Department of Physics
Chalmers University of Technology
SE-412 96 Göteborg, Sweden
[e] Prof. Dr. P. Simon
CIRIMAT, Université de Toulouse, CNRS, INPT, UPS
Université Toulouse 3 Paul Sabatier, Bât. CIRIMAT
118, route de Narbonne 31062 Toulouse cedex 9, France
[f] Dr. A. Grimaud
UMR CNRS 8260 "Chimie du Solide et Energie"
Collège de France
11 Place Marcelin Berthelot, 75231 Paris Cedex 05, France
[g] Dr. A. Grimaud
Sorbonne Universités - UPMC Univ Paris 06
4 Place Jussieu, 75005 Paris, France
[h] Prof. Dr. P. Simon, Prof. Dr. A. A. Franco
Institut Universitaire de France
103 boulevard Saint Michel, 75005 Paris, France
[+] These authors contributed equally to this work.

Supporting information for this article is available on the WWW under https://doi.org/10.1002/batt.202000288

Special Collection *An invited contribution to a Special Collection on Artificial Intelligence in Electrochemical Energy Storage*

electrode properties,[32] and is being considered for the full cell production chain.[33,34]

AI's "understanding" of a problem and prediction accuracy does not rely on natural science-based models and theories; it rather by-passes them by finding correlations and interdependencies between the variables in a dataset through mathematical operations. Therefore, AI's capabilities depend on the veracity and the completeness of the dataset used with respect to the overall complexity of the system under analysis. The number of variables that should be considered for accelerating LIB R&D in practice arises from an unknown but surely complex numerical expression, involving aspects such as materials synthesis, electrode manufacturing, electrochemical performance, recyclability, environmental impact and cost. The experimental approaches and modeling techniques used nowadays allow to investigate a single or few aspects in multiple scales, *e.g.* electrochemical performance or manufacturing procedure, but none of them enables a holistic view to approach that expression. Instead, we may think at using the massive amount of data already available in scientific publications (more than 27,900 LIB publications already exist and this number is growing rapidly)[35] to generate a holistic view and to unravel correlations between variables by using AI/ML. However, in order to collect all the knowledge scattered in scientific literature, one still needs the capability to accurately recover data and variables from it. Text mining (TM) algorithms can be used to extract these data in the most complete and multi-dimensional possible way. TM algorithms efficiency depends not only on their ability to recover the data, but also on the certainty that the data is explicitly and consistently informed in the literature, *i.e.* the fact that researchers have reported all the information needed to construct the AI models.

TM can be defined as the indexing of content or, alternatively, as the extraction of text/number/ideas looking for meaning,[36] and typically includes the following steps: *(i)* retrieving the publication documents, *(ii)* converting from PDF/HTML/XML files into plain text, and *(iii)* mining the desired data. *(i)* is principally hampered by the paywall lock up of peer-reviewed scientific literature behind the publishers' copyright laws, as recently highlighted.[37] In this regard, there has been a long-standing debate on whether or not to use only the typically open-source abstracts rather than full-text articles. Westergaard *et al.* mined 15 million full-text molecular and cell biology articles and showed that mining the full-text article corpus always outperformed the same analysis performed by using abstracts only, *i.e.* it allowed increasing the accuracy of their TM algorithm in terms of information retrieval.[38] Back in 2010, with a considerably smaller dataset, Blake found that authors report $<8\%$ of scientific claims in abstracts.[39] For the field of LIBs, we have estimated that only ~11% of the information found in the full texts could be recovered by using the abstracts only (details in the Methods section). Thus, it is clear that mining full-text articles is strongly preferred. *(ii)* is linked to the widespread digital format today used and it can be a source of errors, as for instance for the conversion from PDF to plain text format for different journals' layout and special characters. Therefore, TM tools should be improved for the sake of building datasets as accurate as possible. Finally, *(iii)* constitutes a major challenge: AI-based algorithms require well-curated inputs for their training, but in the scientific literature most of the data (~80%) is reported as unstructured text.[40] Therefore the TM algorithms need vocabularies and custom dictionaries to assist data identification, extraction and integration.

In general, the task of identifying the structured information of interest (also called entity) in texts is based on Named Entity Recognition (NER). The NER method uses existing databases to identify entities and quantify their occurrence. Substantial progress has been made in NER and information retrieval methods for biomedicine,[41,42,51,43–50] while for chemistry the most known and complete databases available are ChemDataExtractor[52] and PubChem.[53] However, there are no libraries specific to battery R&D able to identify information such as the features of composite electrodes, the electrolyte used, the cycling conditions, etc. Another TM approach widely used today is known as word embedding,[11] which associates to each word in the text a vector and tries to recognize its semantic/syntax as a function of their surroundings.

Despite the lack of dedicated databases, TM (combined or not with AI/ML) has already proved its potential for leading to new discoveries and knowledge in the materials/energy field.[10,54–59] As a benchmark in the battery community, Huang *et al.* constructed a database of battery materials electrochemical properties (such as capacity, conductivity or coulombic efficiency) by mining 229,061 academic papers using the chemistry-aware natural language processing toolkit, ChemDataExtractor.[60] Furthermore, the potential of extracting information and trends from literature of the emerging all-solid-state batteries (ASSBs) field was recently highlighted by Randau *et al.*, which required the use of approximations and hypotheses to manually analyze a small dataset of about 30 publications.[61] Needless to say, such in depth analysis would not be possible for a much more established technology such as LIB, for which thousands of reports must be individually screened: under the hypothesis of reading 200 articles per year, a researcher will need almost 140 years to read all the LIB scientific publications available today![35]

By using an *in-house* TM algorithm and analyzing more than 13,000 LIB/sodium-ion battery (SIB) scientific publications, we aim in this article to provide a general overview on certain valuable electrode's and cell's features extractable out of scientific literature in terms of how often they are reported and about the scattered ways in which they are reported. Considering the lack of specific datasets enabling the use of approaches such as the word-embedding method, this work is based on a keywords search TM algorithm relying on devoted *in-house* libraries complex enough to identify several critical information specific to the battery field. By analyzing this, we would also like to raise the scientific community attention towards the issues that need to be tackled to ensure that in the future accurate and high-quality data coming from scientific literature can feed AI algorithms, raising awareness on the importance of reporting systematically certain basic electrode and cell properties. In addition, we hope that the libraries herein developed

will be further used and improved by the community, with the final goal of setting common battery specific databases to ease information identification and extraction.

2. LIB/SIB Scientific Publication Mining: Researchers' Habits

Our *in-house* text-mining algorithm (provided with this Article) automatically extracts information from the full text of peer-reviewed publications, as schematically presented in Figure 1. This algorithm is based on a combination of keywords linked with logical operators and devoted libraries complex enough to capture as much as possible the different ways in which the screened properties are reported in scientific literature. In stark contrast, classical search engines embedded in online platforms, such as Web of Science™ or Scopus®, search for information in the title, abstract, keywords list and references within an article, and do it in a literal way rather than in context.

Using a semi-automatic download algorithm, we collected a representative dataset of LIB and SIB scientific publications spanning from 1990 to 2019, comprising *ca.* 13,000 articles from different journals and editors (Wiley, Springer, Elsevier, American Chemical Society, IOP Publishing and IEEE Xplore) in PDF or XML format. For the former, the text is extracted from both single and double column configurations and the text contained in tables and figure captions is recovered as well. Subsequently, the text passes through a conversion and pre-processing step, in order to transform the PDF format into a computer-readable text format (TXT) and to remove "noise" and useless text such as HTML tags, links and advertisements. The PDF-to-TXT conversion step can introduce errors due to the different templates used by scientific journals, figures embedded within the text, presence of headers and footers, etc. In our case, different Python libraries were tested, choosing a mixed approach between the pdfminer[80] and tika[81] libraries and an *in-house* code, as discussed more in detail in section S7. The error associated to this step, by manually comparing the PDF to the converted TXT in a randomly chosen sample of 600 papers, was equal to 4.7 %.

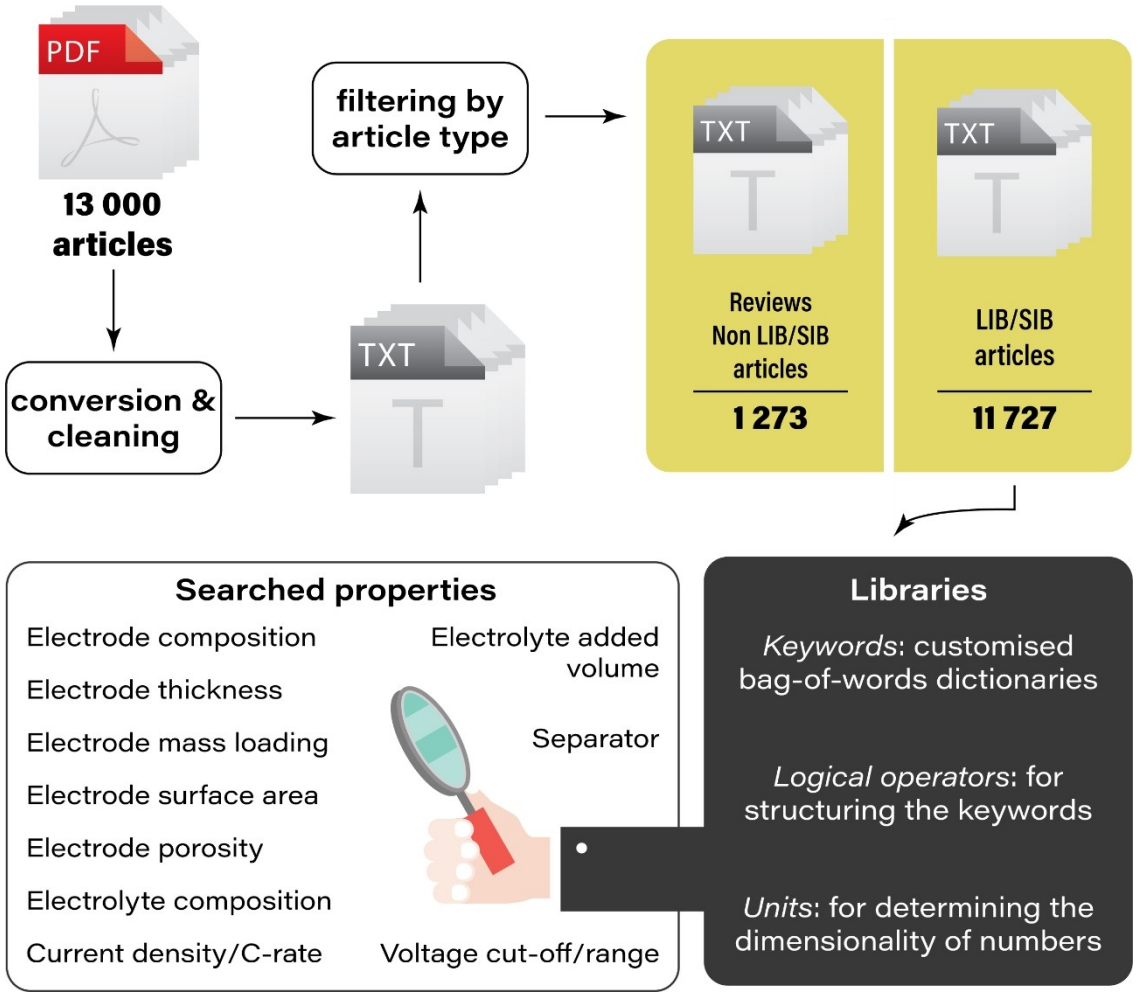


Figure 1. Workflow of our in-house developed TM tool.

Next, a filtering process limits the database to original experimental studies that actually deal with LIBs and SIBs. Two different filters, fully described in section S4, were applied to discard Review articles and to discriminate LIB and SIB studies from other kind of energy storage technologies, such as supercapacitors, K-, Mg-, Al-, Ca- and Zn-Ion batteries, Li- or Na-air/sulfur or redox flow batteries, and also articles dealing with LIB/SIB separators. This filter also classifies articles as LIB or SIB, by frequency analysis. A third filter is applied within the Experimental section to discard articles where no electrode composition is found, to avoid considering articles focusing only on materials physical properties, *e.g.* structural or magnetic characterization of battery materials, or computational studies[82–84] rather than electrochemical performance.

The database was hereby reduced to *ca.* 6,300 (48% remaining), ~5,800 LIB and ~500 SIB, articles. Out of the ~6,700 discarded articles, ~7% were Reviews and ~12% dealt with other energy storage technologies or LIB/SIB separators, but surprisingly ~81% were filtered due to lacking electrode composition. The latter comes from: *(i)* articles without electrochemical testing of composite electrodes *e.g.* modeling studies, *(ii)* complete lack of report of the electrode(s) composition(s) or being placed in the Supplementary Information, *(iii)* data actually being reported, but in a way that did not allow to discriminate it through the search rules defined within the algorithm, as discussed in more detail in sections S5 and S7. While *(ii)* could (partly) be addressed relatively easily by accessing the Supplementary Information, if the electrode composition is therein reported, *(iii)* underlines the complexity of recovering information reported in a non-standardized way and to the associated information losses, even when data is reported. Even if the TM algorithms efficiency is expected to increase, at present and most likely in the next years part of the literature's data will be unreachable, then calling for the need of standardization actions.

From each LIB/SIB property of interest a specific library is defined, fully reported in section S9. These libraries should ideally capture all the different ways a certain information can be reported in the literature *i.e.* being sensitive, while avoiding false hits *i.e.* being selective. The latter is particularly challenging when dealing with common words such as 'thickness' or 'diameter' or similar formats/units being used to report different kinds of information such as electrode and electrolyte composition.

Figure 2 reports the results of our analysis for all the properties investigated (listed in Figure 1) for LIBs in terms of how often they are reported in scientific literature. Similar analysis for SIBs is reported in Figure S1. The electrode, electrolyte and separator properties, and the cycling conditions were screened only in the publications that contained electrode composition. These searched properties were selected for two main reasons: *(i)* they are expected to significantly affect the electrochemical performance of the cell, and *(ii)* they are easy-to-measure and general enough to be relevant for a wide spectrum of electrochemical energy storage technologies, like supercapacitors or ASSBs.[62–64] Out of the screened properties, the electrolyte composition and the cycling conditions (voltage cut-off/range and current density) can be considered as typically reported (>80%). Some other battery properties that are of paramount importance for the performance are often not reported, such as electrode thickness and porosity as well as electrolyte volume (≤ 10%). In addition, the electrode surface area is not found to be systematically disclosed. Since these parameters are critical for *e.g.* high-power applications, observing such trends is highly problematic for the implementation of AI-assisted R&D. Even more revealing, the electrode mass loading was found to be reported in only ~15% and ~27% of the LIB and SIB articles, respectively. Bearing in mind that this property critically determines the battery performance, those percentages were expected to be higher. The extraction of mass loading is however affected by an error arising from the difficulty of the TM algorithm to easily extract a property which can be inferred from two other features, here the electrode's active material mass and its surface area. Yet, even considering this, the percentage of articles reporting the mass loading is significantly lower than expected considering the crucial role of this electrode feature.

In addition to the electrode and cell properties analyzed here, other aspects such as the cycling protocol (formation procedure, waiting time, cycling temperature, use of a constant potential floating step or not, etc.) and the cell format (coin cell, prismatic, pouch, etc.) are known to be important for the battery performance, safety and lifetime. Some aspects, such as the cycling protocol, are more challenging to extract through TM approaches, while others, such as the cell configuration and format, are rather simple to extract. For the latter, we estimate that >75% of the articles reported the cell format and that the vast majority of these (~85%) used coin cells. However, the aim of this article is not to analyze every single aspect that can influence battery performance, but rather to demonstrate the lack of systematic reporting even for basic LIBs electrode and cell properties, calling for action, such as a stronger standardization.

The error associated to each percentage reported in Figure 2 was assessed by analyzing manually 1000 randomly selected articles (100 for each screened property) and comparing the results to the ones obtained through the algorithm. The error analysis, as discussed in more detail in section S7, indicates three main sources of error: *(i)* on the conversion step from PDF to TXT format, *(ii)* incompleteness of the library developed and *(iii)* data not accessible, because it is reported in the Supplementary Information, within a reference or in figures.

The metrics used for evaluating the error associated to our TM algorithm outputs were precision, recall and F-score. Precision is the ratio of true positive observations to the total of true and false positive ones, while recall (or sensitivity) is the ratio of true positive observations to the entire observations reporting the screened property, *i.e.* true positives + false negatives (Figure 3). A true/false positive observation is defined as an article correctly/wrongly classified by the TM algorithm as reporting a certain property. The same can be said for the true/false negatives for the case of articles classified as not reporting a certain property.The F-score is a weighted average of

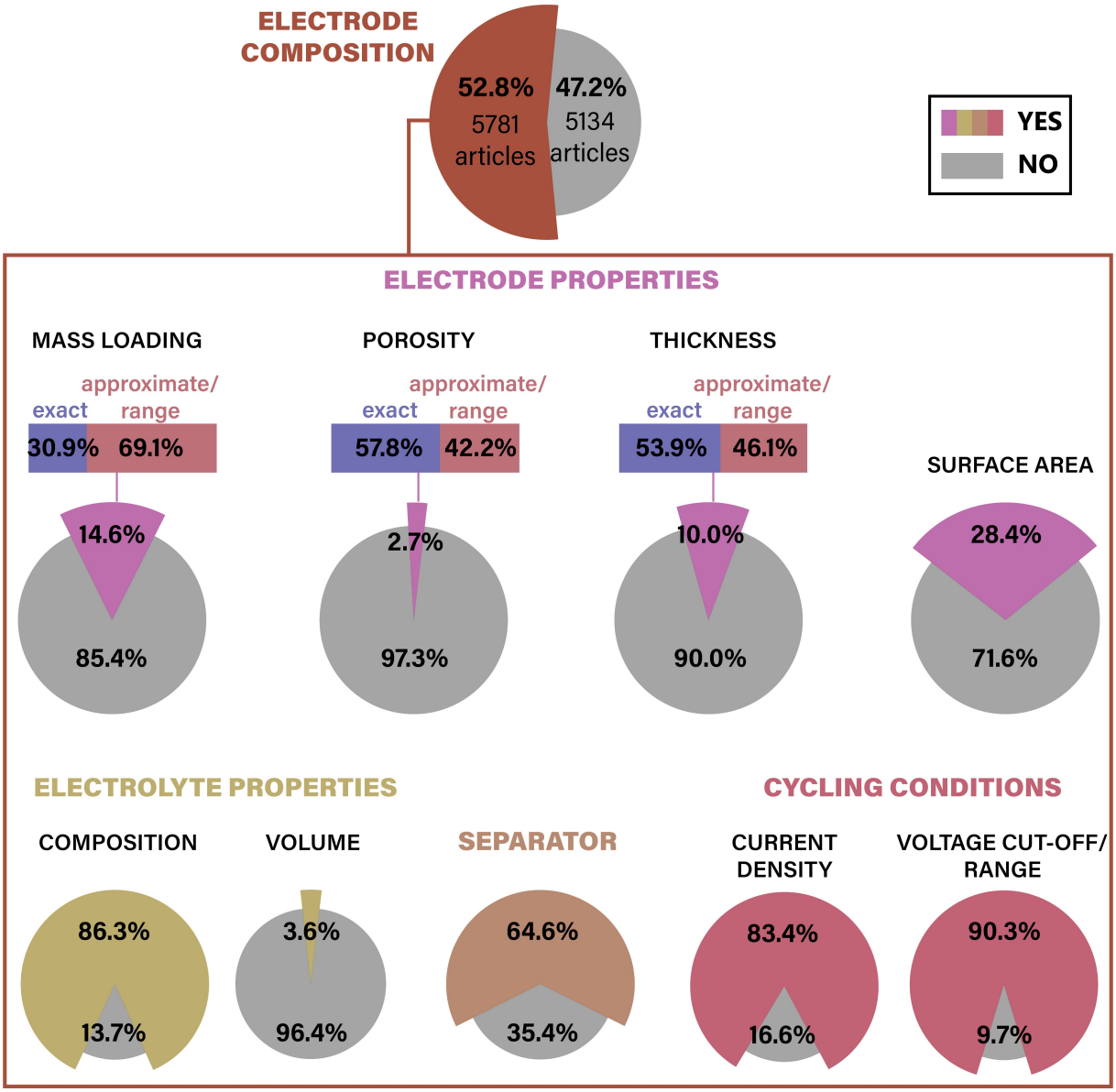


Figure 2. LIB text mining algorithm outputs according to the properties used.

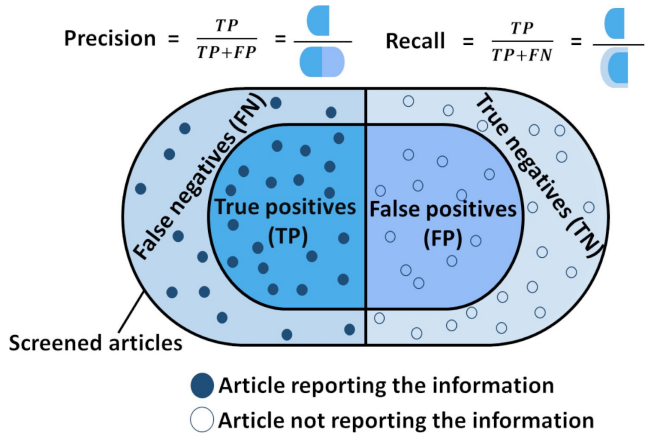


Figure 3. Schematic explanation of the precision and recall concepts.

precision and recall, giving an overall idea about the accuracy of the TM output analyzed. Briefly, F-score ranges between 0 and 1, where 0 indicates that the algorithm is not able to extract any information, while 1 indicates an ideal information extraction process (no error). Figure 4 displays in a spider graph the F-score for the 10 searched properties (section S7 in the Supplementary Information displays all the error evaluation data). Porosity, surface area, mass loading, electrolyte composition and separator have F-scores $\geq$0.85, limit above which the data extraction procedure is considered to be accurate. The lowest F-scores are those of thickness and electrolyte volume. The latter arises from their low precision (Table S1) and the main source of this inaccuracy is that both a volume and a thickness are not exclusively associated to an electrode property, leading to false positives and then to an over-estimation. On the other hand, the source of lower F-scores for

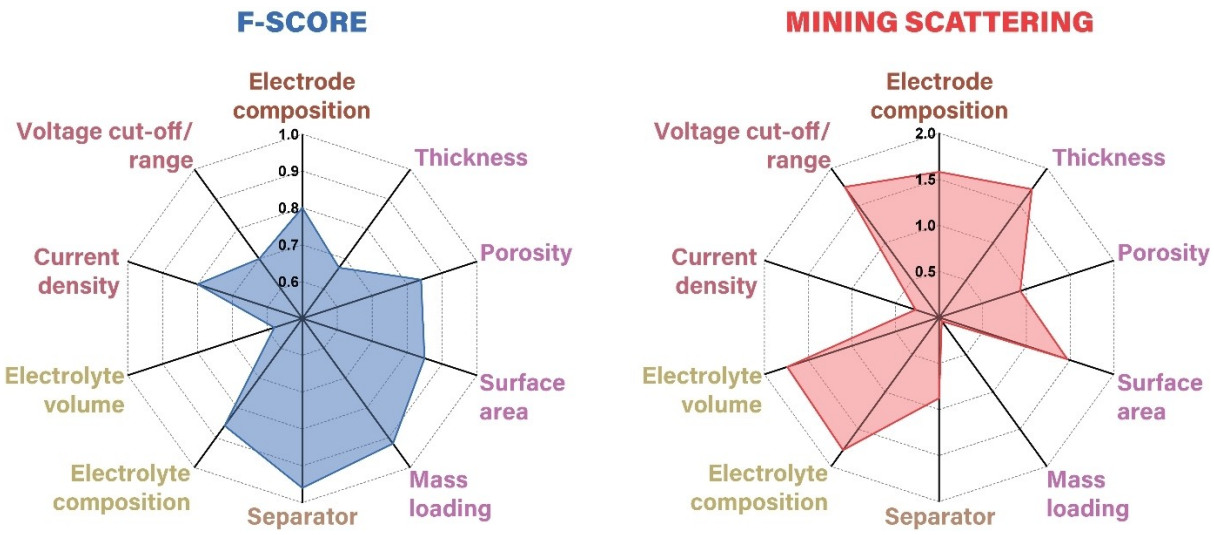


Figure 4. F-score and mining scattering from the analysis of the information within the articles according to the 10 properties investigated.

electrode composition, current density and voltage cut-off/range is their lower sensitivity (Table S1). For the first property, the high number of false negatives are due to huge variability in corpora and lexicons or due to the information reported in a reference within the article, outside the experimental section or in the Supplementary Material. The error on the current density is due to conversion issues on its units or to information reported in the figures. Lastly, the false negatives found for the voltage cut-off/range are systematically due to information that can be inferred from figures, which are inaccessible for the TM algorithm.

For the case of properties associated to a number, our TM tool is also able to detect if that value is exact or if it is reported as a range or approximate value. To discriminate between these two cases, specific libraries and algorithmic functions were developed for the cases of electrode porosity, thickness and mass loading. While the electrode thickness and porosity are expressed $>50\,\%$ of times as exact values, the mass loading is often reported as an approximate value or a range of values ($\sim 70\,\%$) (bars in Figure 2). Although rarely acknowledged, disclosing an electrode property as a range of values, sometimes with differences between the extremes exceeding 100%, makes quite difficult to both reproduce experiments and get reliable information for AI algorithms. The F-scores for these results (Table S2 in the Supplementary Information) show that for the case of porosity there is a low accuracy (yet F-score >0.8), coming mainly from the false negatives (false ranges), and a F-score >0.85 for both thickness and mass loading.

One last aspect that should be considered is how electrode and cell properties are reported in scientific literature, *i.e.* if they are reported similarly or in significantly different ways. In order to quantify the complexity of our mining procedure, the concept of Shannon Entropy[65] from the information theory inspired the calculation of the mining scattering (MS). Briefly, MS [Eq. (S2)] depends on how many different ways researchers refer to a certain information in the literature and to the probability (p_i) to find each of them in the analyzed articles. Unlike F-score, which computes the validity of the TM outputs, the MS measures the information scattering in scientific literature associated with the algorithm's libraries. For instance, if a certain information is expressed in only two possible ways with the same frequency ($p_1 = p_2 = {}^1/_2$), the associated MS would be 0.3, while for three equally probable ways ($p_1 = p_2 = p_3 = {}^1/_3$), the MS would increase to ~ 0.48. Therefore, if researchers report a certain information similarly, its associated MS value would be low, while high MS values indicate that the habits of reporting a certain information in scientific literature is highly scattered, hampering its recovery. For more details about the implementation and use of the MS in our study, readers are referred to section S8.

The MS results for the 10 screened properties are reported in the right spider graph of Figure 4. The current density can be easily recovered by searching selectively its units ($A\,g^{-1}$ or $A\,cm^{-2}$ and its variations) or the C-rate. As these two possibilities are not equally probable, ~ 0.31 and ~ 0.69 when searching by units and C-rate, respectively, the MS for current density for LIBs is <0.3. Similarly, the MS associated to the mass loading is extremely low, as it is always reported in a similar way and can thus be easily captured. However, if considering the case when electrode mass and surface are reported separately, the MS value would then be expected to increase. Aside from these two parameters, all other properties are highly complex to recover, as highlighted by their high MS. Note that the properties which had low F-score values are amongst the ones that have higher values of MS, implying that their lower accuracy is not related to an un-developed library, but rather to the unstructured way in which they are presented within the articles, which hinders their recovery. Furthermore, SIB spider graph (Figure S1) displays very similar results, suggesting that the information retrieval complexity is independent on the technology, and rather depend on the researchers.

Although the percentages reported in Figure 2 might change by developing further improved TM tools, the trends presented in it are not expected to change. Elementary electrode properties as the ones mined in this work are far from being routinely disclosed. Yet, they are basic cell/electrode features for assessing the applicability of a material. For instance, the electrode porosity is such an important property that not only defines the total accessible volume of the Li-ions, but also its tuning allows to improve the electrode electronic conductivity.[66,67] Despite that, having 156 (out of 5,781) articles disclosing its value is, at the least, a severe lack of the "full picture". Mass loading, which is intimately related with the electrode thickness, has a negative correlation with the areal capacity and C-rate-dependence performance.[68] Having less than 15% of the articles reporting it (and often as an approximate value) hampers the validity of reporting a rate capability test. Something similar could be said about the added electrolyte volume.[69]

3. About Battery Reproducibility Crisis and its Connection to Data Reporting

Much has been said about the "reproducibility crisis" in current science[70,71] and several recommendations have been proposed to tackle this issue. The highly competitive academic system forces us to publish, more often than not, focusing on final positive ground-breaking results rather than a consistent and well-elaborated experimental procedure. In 2016, Nature carried out a survey amongst its readers trying to shed light on this so-called crisis.[72] 90% of the researchers acknowledged this crisis, to a certain extent, and the top reason for this irreproducible research was (with over 70% of consensus) the selective reporting of data. Ironically, when possible solutions for tackling this problem were asked, compulsory disclosure of experimental conditions and setting standards for reporting them was not among the 11 proposed improvements. It is surprising that compulsory data reporting was not considered as the easiest and cheapest way to overcome this crisis.

Well-curated data is the way to improve the quality of the battery (and all disciplines) scientific publications. Furthermore, AI and ML need these data for their development and set up. As we saw with the handful electrode properties we tried to mine throughout this article, there is shortage of systematically collected, standardized and accessible experimental battery data.

Standardization of the battery data reported is therefore critically needed. Standardization within science, *i.e.* the action to establish norms most people agree with,[73] can only be successful when implemented in a way not seen as a burden by the scientific community, but as a tool to further support and ensure the creativity process, maximizing simultaneously researchers' freedom and efficiency. Some steps have been taken in that direction by the Journal of Power Sources,[64,74] as they have published a series of guidelines and good practices for publishing batteries and supercapacitors research articles. The field of photovoltaics have long been discussing around a standard report sheet/minimal data to report, and some journals have already started to demand the filling of a datasheet condensing experimental conditions, cell properties and main experimental outputs.[75,76] Furthermore, the standardized approach starts to be adopted in several initiatives at the European level as the European Materials Modelling Council (EMMC),[77] European Materials Characterization Council (EMCC)[78] or the Photon and Neutron Open Science Cloud (PaNOSC).[79]

The use of a standardized template to report experimental data would bring significant advantages to:

- The whole battery community, which would benefit from an extremely valuable database generated and updated continuously by the community itself. These datasets (used in agreement with the specific Journal policies) would have the potential to boost the research of both academia and industry. As a result, data comparison and evaluation of reproducibility will be strongly simplified.
- Journals, editors, reviewers and readers, which would benefit of a faster and easier reviewing process in terms of consistency of the reported results and completeness of the experimental information.
- The AI algorithms, for which it will be easier to exploit the data.

We strongly believe that the adoption of data templates, such as the example reported as supplementary material of this article, will assist in overcoming the broad standardization challenge. In order to reach a widespread use, journals could implement such templates to be used at the submission stage. These templates would allow the creation of databases from which data could be extracted easily and accurately, and as a consequence, the wide use of AI approaches will be unlocked. Finally, specific templates should be developed and applied for distinctly different energy storage technologies, such as ASSBs and supercapacitors, and for materials synthesis and manufacturing – which is intrinsically more complicated due to the large variety of techniques and instruments used.

A new era is awaiting us and we are dreamful on the idea of being more prolific by spending less time deciphering about the veracity and completeness of reported data. With the emergence of AI, scientists will not become robots filling out templates and clicking checkboxes, but the access to reliable and trustful data will enhance scientific inspiration, cleaver reasoning, and innovative theoretical approaches, all of which are key enablers for technological breakthroughs.

Methods

Aiming to increase the TM algorithm accuracy as much as possible, several strategies to discriminate between different sections of the articles (as title, abstract, experimental section and keywords) were developed. The latter allows searching information in specific regions of the articles, reducing the frequency of false hits. For the interested readers, a complete discussion on the strategies developed to identify the different articles' sections is reported in section S3, in the Supplementary Information.

The most critical aspects in our analysis were the development of accurate filters to discriminate between LIBs, SIBs and other battery related articles and the development of keywords + logical operators based rules allowing a sensitive and selective identification of the searched properties, as briefly discussed in the main text of this article. Similar considerations can be said for the libraries devoted to identify if a certain property is reported as exact or approximate/range of values. These filters and rules can be applied either to certain section(s) of the articles (as the filter for identifying the review articles) or to the whole article (as the search of the current density used). For further information on the strategies developed along this work to define these filters, libraries and rules, the interest readers are referred to sections S4, S5 and S6 in the Supplementary Information.

The concept of precision, recall and F1-score (briefly explained in the main text) are widely used in the TM field and were used here as a metric to evaluate the error associated to each TM output. For more details on the error analysis performed along this work, the interested readers are referred to section S7.

The concept of MS [Eq. (S2)] was defined and calculated to quantify the complexity of information retrieval for the searched properties and by using the libraries developed here (section S9). Briefly, the concept of MS was identified as suitable to quantify the complexity of information retrieval because *(i)* MS value raises when the different ways of reporting a certain property in scientific articles increase and *(ii)* for a constant number of ways in which a certain property is referred to, MS increase if all (or several) of them are commonly used in scientific literature, which makes their correct recovery challenging. However, if only one (or few) of these ways are commonly used, MS value will be lower, indicating an easier information retrieval process. For more insight on the MS mathematical definition the interested readers are referred to section S8.

Finally, the comparison between information retrieval while using full text or abstracts only was performed as follows. The abstracts of the 5,781 LIBs articles classified as reporting the electrode composition were extracted. Then, the libraries developed for the searched properties (bottom of Figure 1) were applied to those abstracts. The number of articles in which one or more properties were detected in the abstracts were counted and compared to the ones found by using the full texts (Figure 2). This allows calculating the percentage of information that could be extracted by using abstracts compared to the one extractable if using the full text, for each property analyzed here: mass loading (~5%), porosity (~13%), thickness (~10%), surface area (~4%), electrolyte composition (~7%), electrolyte volume (0%), separator (~5%), current density (~43%) and voltage cut-off (~14%). The value reported in the main text (~11%) is the arithmetic average of the percentages detailed above.

Acknowledgements

H.E.-B., P.J. and A.A.F. acknowledge the ALISTORE European Research Institute for the funding support of H.E.-B. Ph.D. thesis. A.A.F., T. L., E. P. and M.D. acknowledge the European Union's Horizon 2020 research and innovation programme for the funding support through the European Research Council (grant agreement 772873, "ARTISTIC" project). A.A.F. and P.S. acknowledge Institut Universitaire de France for the support. P.J. acknowledges the financial support from the Swedish Energy Agency "Batterifondsprogrammet" as well as the continuous support from several of Chalmers Areas of Advance: Materials Science, Energy, and Transport. H. E.-B., T. L., E. N. P., M. D., M. M., P. S., A. G. and A. A. F. thank the French National Research Agency for its support through the Labex STORE-EX project (ANR-10LABX-76-01). All the authors gratefully acknowledge the RS2E network for the inspirational discussions that originated this Article and for its contribution at building the publications database used in our text mining analysis. The authors acknowledge Prof. Jean-Marie Tarascon for very useful discussions and support.

Conflict of Interest

The authors declare no conflict of interest.

Keywords: artificial intelligence · battery · reproducibility crisis · standards · text mining

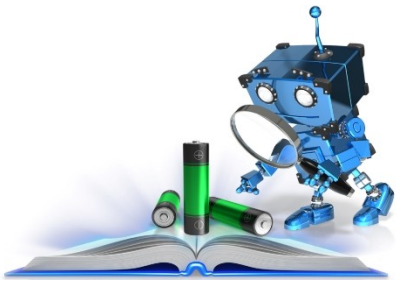
[1] G. Armstrong, *Nat. Chem.* **2019**, *11*, 1076.
[2] T. Placke, R. Kloepsch, S. Dühnen, M. Winter, *J. Solid State Electrochem.* **2017**, *21*, 1939.
[3] Battery 2030. Available at: https://battery2030.eu/. (Accessed: June 2020).
[4] C. M. Bishop, *Pattern Recogniton and Machine Learning*; Springer-Verlag: New York, 2006.
[5] Z. Zhou, X. Li, R. N. Zare, *ACS Cent. Sci.* **2017**, *3*, 1337.
[6] B. Sanchez-Lengeling, A. Aspuru-Guzik, *Science* **2018**, *361*, 360.
[7] S. J. Russell, P. Norvig, Artificial Intelligence: A Modern ApproachRussell, S. J., & Norvig, P. (2010). Artificial Intelligence: A Modern Approach. Artificial Intelligence. https://doi.org/10.1017/S0269888900007724; 2010.
[8] G. H. Yann LeCun, Yoshua Bengio, *Nature* **2015**.
[9] A. C. Mater, M. L. Coote, *J. Chem. Inf. Model.* **2019**, *59*, 2545.
[10] A. Torayev, P. C. M. M. Magusin, C. P. Grey, C. Merlet, A. A. Franco, *J. Phys. Mater.* **2019**.
[11] V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K. A. Persson, G. Ceder, A. Jain, *Nature* **2019**, *571*, 95.
[12] B. Writer, *Lithium-Ion Batteries*: A Machine-Generated Summary of Current Research; Springer, Ed.; **2019**.
[13] N. Jones, *Nature* **2017**.
[14] R. Rada, *ACM SIGBIO Newsl.* **1983**.
[15] Artificial intelligence methods in the environmental sciences; Haupt, S. E.; Pasini, A.; Marzban, C., Eds.; Springer, 2009.
[16] M. Haghighatlari, J. Hachmann, *Curr. Opin. Chem. Eng.* **2019**, *23*, 51.
[17] T. Dimitrov, C. Kreisbeck, J. S. Becker, A. Aspuru-Guzik, S. K. Saikin, *ACS Appl. Mater. Interfaces* **2019**, *11*, 24825.
[18] J. G. Freeze, H. R. Kelly, V. S. Batista, *Chem. Rev.* **2019**, *119*, 6595.
[19] S. Li, J. Li, H. He, H. Wang, *Energy Procedia* **2019**, *159*, 168.
[20] E. D. Cubuk, A. D. Sendek, E. J. Reed, *J. Chem. Phys.* **2019**, *150*, 214701.
[21] R. P. Joshi, J. Eickholt, L. Li, M. Fornari, V. Barone, J. E. Peralta, *ACS Appl. Mater. Interfaces* **2019**, *11*, 18494.
[22] H. Wang, Y. Ji, Y. Li, *Rev. Comput. Mol. Sci.* **2020**, *10*, 1.
[23] C. Chen, Y. Zuo, W. Ye, X. Li, Z. Deng, S. P. Ong, *Adv. Energy Mater.* **2020**, *1903242*, 1.
[24] A. Dave, J. Mitchell, K. Kandasamy, S. Burke, B. Paria, B. Poczos, J. Whitacre, V. Viswanathan, Autonomous discovery of battery electrolytes with robotic experimentation and machine-learning. *arXiv* **2019**.
[25] L. Petrich, D. Westhoff, J. Feinauer, D. P. Finegan, S. R. Daemi, P. R. Shearing, V. Schmidt, *Comput. Mater. Sci.* **2017**, *136*, 297.
[26] Y. Takagishi, T. Yamanaka, T. Yamaue, *Batteries* **2019**, *5*.
[27] E. Chemali, P. J. Kollmeyer, M. Preindl, A. Emadi, *J. Power Sources* **2018**, *400*, 242.
[28] M. Berecibar, *Nature* **2019**, *568*, 325–326.
[29] D. A. Howey, *Electrochem. Soc. Interface* **2019**.
[30] K. A. Severson, P. M. Attia, N. Jin, N. Perkins, B. Jiang, Z. Yang, M. H. Chen, M. Aykol, P. K. Herring, D. Fraggedakis, M. Z. Bazant, S. J. Harris, W. C. Chueh, R. D. Braatz, *Nat. Energy* **2019**, *4*, 383.

[31] P. M. Attia, A. Grover, N. Jin, K. A. Severson, T. M. Markov, Y.-H. Liao, M. H. Chen, B. Cheong, N. Perkins, Z. Yang, P. K. Herring, M. Aykol, S. J. Harris, R. D. Braatz, S. Ermon, W. C. Chueh, *Nature* **2020**, *578*, 397.
[32] R. P. Cunha, T. Lombardo, E. N. Primo, A. A. Franco, *Batteries & Supercaps* **2020**, *3*, 60; Supercaps **2020**, 3, 60.
[33] A. Turetskyy, S. Thiede, M. Thomitzek, N. von Drachenfels, T. Pape, C. Herrmann, *Energy Technol.* **2020**, *8*, 1.
[34] S. Thiede, A. Turetskyy, A. Kwade, S. Kara, C. Herrmann, *CIRP Ann.* **2019**, *68*, 463.
[35] Boolean search performed in Web of Science on November 2020 searching for the keywords "Lithium" AND "Ion" AND "Batteries.".
[36] In *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*; Miner, G.; Delen, D.; Elder, J.; Fast, A.; Hill, T.; Nisbet, R. A. B. T.-P. T. M. and S. A. for N. T. D. A., Eds.; Academic Press: Boston, 2012; pp. 3–27.
[37] P. Pulla, *Nature* **2019**, *571*, 316–318.
[38] D. Westergaard, H.-H. Stærfeldt, C. Tønsberg, L. J. Jensen, S. Brunak, *PLoS Comput. Biol.* **2018**, *14*, e1005962.
[39] C. Blake, *J. Biomed. Inf.* **2010**, *43*, 173.
[40] Structured vs. Unstructured Data. Available at: https://www.datamation.com/big-data/structured-vs-unstructured-data.html. (Accessed: June 2020).
[41] E. Kim, Z. Jensen, A. Van Grootel, K. Huang, M. Staib, S. Mysore, H. S. Chang, E. Strubell, A. McCallum, S. Jegelka, E. Olivetti, *J. Chem. Inf. Model.* **2020**, *60*, 1194.
[42] Y. Garten, A. Coulet, R. B. Altman, *Pharmacogenomics* **2010**, *11*, 1467.
[43] K. Fukuda, A. Tamura, T. Tsunoda, T. Takagi, *Pac. Symp. Biocomput.* **1998**, 707.
[44] D. L. Rubin, C. F. Thorn, T. E. Klein, R. B. Altman, *J. Am. Med. Informatics Assoc.* **2005**, *12*, 121.
[45] M. V Plikus, Z. Zhang, C.-M. Chuong, *BMC Bioinf.* **2006**, *7*, 424.
[46] R. Hoffmann, A. Valencia, *Bioinformatics* **2005**, *21*, 252.
[47] Y. Garten, R. B. Altman, *BMC Bioinf.* **2009**, *10*, S6.
[48] H. M. Müller, E. E. Kenny, P. W. Sternberg, *PLoS Biol.* **2004**, *2*.
[49] R. Winnenburg, T. Wächter, C. Plake, A. Doms, M. Schroeder, *Briefings Bioinf.* **2008**, *9*, 466.
[50] L. J. Jensen, J. Saric, P. Bork, *Nat. Rev. Genet.* **2006**, *7*, 119.
[51] Y. Tsuruoka, J. Tsujii, *J. Biomed. Inf.* **2004**, *37*, 461.
[52] http://chemdataextractor.org/.
[53] https://pubchem.ncbi.nlm.nih.gov/.
[54] A. Moro, G. Joanny, C. Moretti, *Futures* **2020**, *117*, 102511.
[55] Z. Jensen, E. Kim, S. Kwon, T. Z. H. Gani, Y. Román-Leshkov, M. Moliner, A. Corma, E. Olivetti, *ACS Cent. Sci.* **2019**, *5*, 892.
[56] E. Kim, K. Huang, A. Saunders, A. McCallum, G. Ceder, E. Olivetti, *Chem. Mater.* **2017**, *29*, 9436.
[57] M. C. Swain, J. M. Cole, *J. Chem. Inf. Model.* **2016**, *56*, 1894.
[58] C. J. Court, J. M. Cole, *npj Comput. Mater.* **2020**, *6*, 18.
[59] M. Krallinger, O. Rabal, A. Lourenço, J. Oyarzabal, A. Valencia, *Chem. Rev.* **2017**, *117*, 7673.
[60] S. Huang, J. M. Cole, *Sci. Data* **2020**, *7*, 1.
[61] S. Randau, D. A. Weber, O. Kötz, R. Koerver, P. Braun, A. Weber, E. Ivers-Tiffée, T. Adermann, J. Kulisch, W. G. Zeier, F. H. Richter, J. Janek, *Nat. Energy* **2020**, *5*, 259.
[62] Z. Lin, T. Liu, X. Ai, C. Liang, *Nat. Commun.* **2018**, *9*, 5262.
[63] R. Nölle, K. Beltrop, F. Holtstiege, J. Kasnatscheew, T. Placke, M. Winter, *Mater. Today* **2020**, *32*, 131.
[64] J. Li, C. Arbizzani, S. Kjelstrup, J. Xiao, Y. yao Xia, Y. Yu, Y. Yang, I. Belharouak, T. Zawodzinski, S. T. Myung, R. Raccichini, S. Passerini, *J. Power Sources* **2020**, *452*, 227824.
[65] C. E. Shannon, *Bell Syst. Tech. J.* **1948**, *27*, 379.
[66] V. Laue, F. Röder, U. Krewer, *Electrochim. Acta* **2019**, *314*, 20.
[67] D. Parikh, T. Christensen, J. Li, *J. Power Sources* **2020**, *474*, 228601.
[68] S.-H. Park, R. Tian, J. Coelho, V. Nicolosi, J. N. Coleman, *Adv. Energy Mater.* **2019**, *9*, 1901359.
[69] S. Chen, C. Niu, H. Lee, Q. Li, L. Yu, W. Xu, J.-G. Zhang, E. J. Dufek, M. S. Whittingham, S. Meng, J. Xiao, J. Liu, *Joule* **2019**, *3*, 1094.
[70] P. B. Stark, *Nature* **2018**, *557*, 613.
[71] D. Fanelli, *Proc. Nat. Acad. Sci.* U. S. A. **2018**, 115, 2628.
[72] M. Baker, *Nature* **2016**, *533*, 452–454.
[73] NORM | meaning in the Cambridge English Dictionary. Available at: https://dictionary.cambridge.org/dictionary/english/norm. (Accessed: September 2020).
[74] C. Arbizzani, Y. Yu, J. Li, J. Xiao, Y. yao Xia, Y. Yang, C. Santato, R. Raccichini, S. Passerini, *J. Power Sources* **2020**, *450*, 227636.
[75] Nature Solar Cells Reporting Summary. Available at: https://www.nature.com/documents/nr-photovoltaic-reporting.pdf. (Accessed: September 2020).
[76] M. V. Khenkin, E. A. Katz, A. Abate, G. Bardizza, J. J. Berry, C. Brabec, F. Brunetti, V. Bulović, Q. Burlingame, A. Di Carlo, R. Cheacharoen, Y. B. Cheng, A. Colsmann, S. Cros, K. Domanski, M. Dusza, C. J. Fell, S. R. Forrest, Y. Galagan, D. Di Girolamo, M. Grätzel, A. Hagfeldt, E. von Hauff, H. Hoppe, J. Kettle, H. Köbler, M. S. Leite, S. Frank, Liu, Y. L. Loo, J. M. Luther, C. Q. Ma, M. Madsen, M. Manceau, M. Matheron, M. McGehee, R. Meitzner, M. K. Nazeeruddin, A. F. Nogueira, Ç. Odabaşı, A. Osherov, N. G. Park, M. O. Reese, F. De Rossi, M. Saliba, U. S. Schubert, H. J. Snaith, S. D. Stranks, W. Tress, P. A. Troshin, V. Turkovic, S. Veenstra, I. Visoly-Fisher, A. Walsh, T. Watson, H. Xie, R. Yıldırım, S. M. Zakeeruddin, K. Zhu, M. Lira-Cantu, *Nat. Energy* **2020**, *5*, 35.
[77] https://emmc.info/.
[78] European Materials Characterisation Council (EMCC). Available at: http://www.characterisation.eu/. (Accessed: September 2020).
[79] The Photon and Neutron Open Science Cloud (PaNOSC). https://www.panosc.eu/. (Accessed: September 2020).
[80] https://pypi.org/project/pdfminer/.
[81] https://pypi.org/project/tika/.
[82] M. Reynaud, J. Rodríguez-Carvajal, J. N. Chotard, J. M. Tarascon, G. Rousse, *Phys. Rev. B: Condens. Matter Mater. Phys.* **2014**, *89*, 1.
[83] T. Bamine, E. Boivin, F. Boucher, R. J. Messinger, E. Salager, M. Deschamps, C. Masquelier, L. Croguennec, M. Ménétrier, D. Carlier, *J. Phys. Chem. C* **2017**, *121*, 3219.
[84] Y. Deng, C. Eames, L. H. B. Nguyen, O. Pecher, K. J. Griffith, M. Courty, B. Fleutot, J. N. Chotard, C. P. Grey, M. S. Islam, C. Masquelier, *Chem. Mater.* **2018**, *30*, 2618.

Manuscript received: January 4, 2021
Revised manuscript received: January 15, 2021
Accepted manuscript online: January 29, 2021
Version of record online: ■■, ■■■

CONCEPTS

Breaking the status quo: An *in-house* text mining algorithm is used here to study the Na- and Li-ion battery researchers' habits in terms of how often certain key electrode and cell features are reported and about the scattered ways in which they are reported in scientific literature. Our results clearly show a systematic lack of certain key data, calling for standardization actions.



*H. El-Bousiydy, T. Lombardo, Dr. E. N. Primo, M. Duquesnoy, Dr. M. Morcrette, Prof. Dr. P. Johansson, Prof. Dr. P. Simon, Dr. A. Grimaud, Prof. Dr. A. A. Franco**