A Appendix

A.1 Methods: Automated Literature Extraction

A.1.1 Corpus

A comprehensive corpus spanning the SIB-literature was assembled by querying the Lens API [16] using relevant keywords. The acquisition of the papers was made possible through licensing agreements with multiple publishers. Subsequent subdivision into paragraphs was carried out leveraging the group's prior work [18]. This initial corpus was then filtered for cathode materials of interest using several text-based matching rules.

We looked at several cathode active material classes as follows:

Category	Examples
Layered Metal Oxides (LMOs)	NaMnFeO or NFM or NFMO (Sodium Manganese Iron Oxide) NaMnFeNiO or NaNFM (Sodium Nickel Manganese Iron Oxide) NaMnNiO (Sodium Manganese Nickel Oxide) NaFeCo (Sodium Iron Cobalt Oxide) NaFeMnCu (Sodium Iron Manganese Copper Oxide) NaMnNiTi (Sodium Manganese Nickel Titanium Oxide) NaNiMn (Sodium Nickel Manganese Oxide) NaNiMnCo (Sodium Nickel Manganese Cobalt Oxide) NaRuO (Sodium Ruthenium Oxide)
Polyanionics	NVPF or NVPFO (Sodium Vanadium Fluorophosphates) NFPO or NaFeP (Sodium Iron Phosphate) NaVPON
Prussian Blue Analogues (PBAs)	NaFeCN or NaHFC (Sodium Hexaferrocyanates) NaFeFeCN

Table 2: Summary of Cathode Materials

A.1.2 Classifiers

Our pipeline harnessed sentence and phrase-level classifiers to screen and filter papers for pertinent information and underscore key mechanisms as cited by the authors. It embodied two types of classification models:

Sentence Level The initial filtering stage operates at the sentence level where we utilized fine-tuned BERT models to discriminate among three sentence categories: challenge sentences, mitigation sentences, and sentences of neither class. This methodology proved effective in efficiently screening the millions of sentences present in our corpus, subsequently filtering out information of distinct value. Given a sentence s_i, our classifier predicted the probability distribution over the classes c_i, eventually assigning the most suitable label. Our three classes were defined as Challenge Sentences, Mitigation Sentences and Non-Target Sentences.

For the development of our sentence-level screening classifier, we annotated a dataset comprising approximately 2,500 sentences. To expedite the labeling endeavor while securing sufficient quantities of infrequently occurring improvement and challenge sentences, we exploited an in-context learned GPT-esque large-language model for a cycle of weak labeling [31]. Ultimately, selected weak-labeled sentences were supplemented by an equal number of randomly chosen sentences and presented to human annotators, thus assuring expert-level quality. Owing to the presence of highly domain-specific vocabulary, we employed the expertise of battery domain experts for labeling. The main merit of this procedure lies in the cutting of labeling time by augmenting the proportion of scarce challenge and improvement sentences. Moreover, this methodology ensures the inclusion of hard-to-predict sentences that had been incorrectly selected by the weak labeller. Overall sentence diversity is ensured by overcoming potential biases of the weak-labeller by introducing a substantial volume of randomly sampled sentences. We computed the inter-annotator rating to be high with agreements of 80% Cohen's κ among the trio of annotators.

For the development of our classifiers, we commenced by benchmarking a variety of approaches on our dataset, using stratified data splits and hyperparameter optimisation.

Phrase Level The phrases were categorized as mitigation strategies (e.g., "doping with Li"), undesirable material-related outcomes (e.g., "low Mn dissolution"), or performance metrics (e.g., "energy density"). Formally, our task was: given a set of all candidate spans s_i, assign each span to the correct entry in the set of defined entity classes $\mathcal{E}$ using $y_e(s_i) \in \mathcal{E}$. This set of classes encompassed valid and invalid phrases. In the second stage the same spans are then investigated for causal relationships, scrutinizing all potential combinations using $y_r(s_i, s_j) \in \mathcal{R}$.

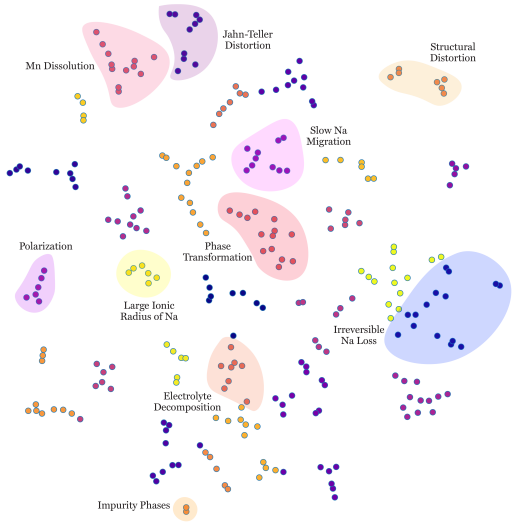

Figure 6: Challenge sentences space extracted using our NLP framework from the sodium-ion batteries literature on NFMO (sodium iron manganese oxide) cathodes with labelled approximate regions of prominent micro-challenges.

To assemble a dataset for our task, we curated annotation guidelines and utilize our sentence classifiers to evaluate the sodium-ion battery corpus. We randomly sampled 600 sentences and around $2,000$ phrases, with equal shares of challenge and improvement types for expert annotation. Comparing the inter-annotator agreement scores amongst a subset of 60 sentences, we calculated the average inter-annotator agreement by F1 scores for pairwise comparison. Our examination revealed that the discrepancy is primarily caused by the high complexity leading to divergences in annotators' span delimitation. We found the annotations to remain accurate and preserved the key messages conveyed by the sentences in our dataset. This underscored the high degree of flexibility inherent in our task. The comparison of inter-annotator agreement and classifier performance demonstrated the outstanding level of capability attained by our model. It effectively exploited sentence context to discern lengthy phrases within the sentences, providing insights into mitigation strategies as well as challenges. Our model successfully discerns even non-trivial relationships.

	F1	P	R
MATSCIBERT	83.1 (1.2)	83.4 (1.1)	83.8 (1.3)
SCIBERT	**84.1 (1.7)**	**84.4 (1.5)**	**84.2 (1.4)**
SENTENCE BERT	79.2 (1.8)	79.3 (2.0)	79.2 (1.8)
GPT3 @10 SHOTS	73.2	75.9	72.5
TF-IDF	70.2	70.1	70.5
RANDOM	51.4	51.2	51.8

	Sentences	Entities	Relations
CHALLENGE	84.1	67.5	39.4
IMPROVEMENT	83.1	67.9	50.8

Table 3: Model Comparison on the Sentence Dataset (Left) and Results for the Open Information Extraction Task (Right).

A.1.3 Database Creation and Accessibility

The core of our methodology was formed by a sequential application of the developed methods. The extraction and presentation of information were facilitated through four main steps. Initially, the acquired publications were scrutinized for the studied cathode active material discussed, which includes layered metal oxides, polyanionic compounds, and Prussian blue analogues. We utilized text-guided rules to match elemental formulas and delineate materials by composition. Subsequently, the papers were introduced to our sentence classifier to discern sentences related to challenges and mitigation strategies. Following this, in the phrase and relation identification phase, we highlighted sections in the literature of particular interest. Lastly, we visualized the

database of identified challenges and mitigation strategies by encoding each sentence using sentence BERT models [29, 4].

By applying this methodology to our corpus of approximately 2200 papers on selected cathode chemistries, we obtained a database of 31,000 challenge and mitigation sentences. Out of these, our classifiers identified 9,000 relations. Analyzing the diversity of papers in the final relational database, we noted that 91% of the papers are represented in the final improvement, and 82% in the challenge database, underscoring the comprehensive coverage of our source material. To evaluate the accuracy of our database, we randomly selected 200 entries, equally distributed between challenge and mitigation strategy entries. Two domain experts then assessed these entries for their accuracy and completeness. Our findings indicated very good overall correctness exceeding 90% in our database. Enabling usage of our data for further analysis.

To categorize the identified mitigation and challenge mentions, we utilized BERT-based clustering methods to visualize the extracted phenomena. We collated a dataset of 1100 micro-challenges related to materials, which led to key performance degradation. These root cause or "mirco-challenges" were carried and of several types: elemental phenomena (e.g., electronic conductivity), structural phenomena (e.g., secondary phase formation), morphological phenomena (e.g., surface impurities), and key performance metrics (e.g., energy density). Subsequently, we assessed various clustering methods based on their cluster purity, which evaluated the model's ability to group text mentions of related phenomena closely together, modeling topics of interest. The most successful results were attained when phrases were embedded using Sentence-BERT models [29], which yielded the highest cluster purity values. We thus incorporated this approach into our methodology and introduced an additional visualization step. Utilizing the UMAP [32] algorithm, we condensed the dimensionality into 2D space, generating an interactive map of the curated database. Here, relevant strategies were grouped together, providing a conducive platform for comprehensive exploration of pairs of challenges and mitigation strategies. It further facilitated their selection for evaluation of large-scale manufacturing feasibility. Beyond its application in the realm of SIBs, our methodology was successfully transposed to the field of LIBs. Assessing database accuracy demonstrates a tolerable reduction in performance, with the correctness of database entries dropping to 85%. This successful domain shift paves the way for us to extend the array of investigated mitigation strategies to those strategies reported for LIBs, which have already seen industrial-scale adoption.

A.2 Case Study Identification and Evaluation

We can use our developed methodology for downstream tasks to identify and evaluate case studies and combine it with a process-based cost modeling for scalability inspection as shown in Figure 7.

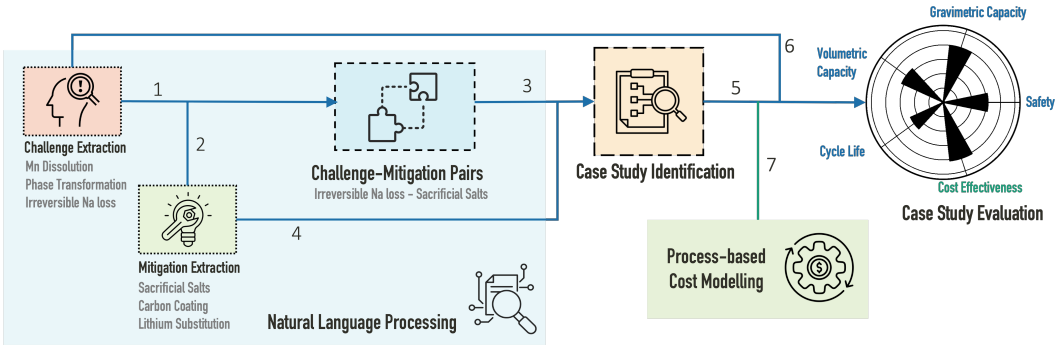

Figure 7: **Combining NLP techniques with Process-based Cost Modeling**: Extracted challenges and mitigation strategies were mapped to constitute challenge-mitigation pairs (Arrows 1 and 2). These challenge-mitigation pairs represent individual cases that can be further evaluated (Arrow 3). To get a broader overview of all represented strategies, we utilized the visualized space of mitigation strategies (Figure 3 (a)) to downselect challenge-mitigation case studies (Arrow 4) and create a database of sentences of interest and corresponding DOIs. After the identification of the case studies (Arrow 5), we used the extracted challenges database of sentences and corresponding DOIs (Arrow 6) along with a cost model built in-house for sodium-ion battery cathode materials (Arrow 7) to quantitatively assess the scalability barriers. The blue parameters were obtained via NLP and the green through cost modeling.