Energy Summer Program Report

Data Driven Understanding of Hard Carbon Development for Sodium Ion Batteries

Literature data mining, validation, and 3D modelling of disordered porous carbon structures

By: Lujain Abdulrahman Alkhammash

Internship Organization: King's College London

Supervised by: Dr. Carla de Tomas

Date: August, 24th 2026

Table of Contents

Abstract 3
Project at a Glance 4
1 Introduction 4
1.1 Background and Motivation 4
1.2 Project Aim and Objectives 5
2 Scientific Background 5
2.1 Disordered and Nanoporous Carbon Structure 5
2.2 Temperature, Ordering, and Graphitization 6
2.3 Why Data Driven Analysis Is Needed 6
3 Methodology 7
3.1 Reference Data and Initial Analysis 7
3.2 Paper Conversion, Retrieval, and First Extraction Architecture 7
3.3 Table Aware Architecture 7
3.4 Validation Strategy 8
3.5 Scale Up and Literature Access 8
4 Results and Discussion 9
4.1 Benchmark and Extracted Data Trends 9
4.2 Final Extraction and Benchmark Performance 11
4.3 Remaining Benchmark Differences 13
4.4 Physical Interpretation from 3D Printing 14
5 Challenges and Lessons Learned 14
6 Future Work 15
7 Conclusion 16
Acknowledgements 17
References and Project Sources 18
Appendix A. Final Validation Figure Files 19

Table of Figures

Figure 1. Effect of density on one million atom disordered carbon structures. Increasing density reduces pore size and increases graphene sheet stacking. Adapted from Putman et al. (2023). 6
Figure 2. Final project pipeline showing the progression from source literature through table reconstruction, structural and semantic recovery, source linked data, benchmark validation, and scientific analysis. 8
Figure 3. Manual benchmark and final extracted elemental compositions as a function of processing temperature. Panels show carbon, hydrogen, nitrogen, and oxygen. The overlap demonstrates reproduction of the main temperature dependent trends. 10
Figure 4. Van Krevelen comparison of the manual benchmark and final extracted dataset. The full view shows the complete atomic H/C and O/C distribution, while the enlarged view provides a clearer comparison of the dense central region. 10
Figure 5. Atomic H/C and O/C ratios as a function of processing temperature for the manual benchmark and final extraction. Both datasets reproduce the general reduction in hydrogen and oxygen relative to carbon as thermal treatment progresses. 10
Figure 6. Element level agreement between the manual benchmark and final extraction for matched records. Panels show C, H, N, and O. The matched values lie almost entirely on the one to one line, showing very high numerical fidelity after record identity is established. 12
Figure 7. Final benchmark validation performance. The panels summarize matched and unmatched benchmark rows, recovery by paper, recovery by feedstock group, and the classified reasons for the remaining benchmark differences. 13
Figure 8. One million atom disordered carbon model proposed for future 3D printing. The larger simulation volume allows structural features to develop with fewer finite size restrictions. Adapted from Putman et al. (2023). 16

Abstract

Hard carbon is one of the most promising anode materials for sodium ion batteries because its disordered carbon layers, enlarged interlayer spacing, defects, and pore network provide several sites for sodium storage. Sodium is abundant and geographically widespread, making sodium ion batteries attractive for reducing dependence on more constrained battery materials. However, the performance of hard carbon depends strongly on precursor type, carbonization conditions, and the resulting pore and carbon structure. The experimental knowledge needed to understand these relationships is distributed across a very large scientific literature using different materials, terminology, measurements, and reporting formats.

This project combines physical understanding of disordered carbon with a data driven literature mining workflow. Atomistic models of disordered and nanoporous carbon were visualized and 3D printed to examine how changes in density and thermal treatment affect pore size, carbon sheet stacking, curvature, and three dimensional connectivity. Several sections of the same models were also printed so that pore continuity and termination could be inspected physically. This provides a simple way to visualize structural pathways that are difficult to understand from two dimensional representations alone.

In parallel, a literature mining pipeline was developed and validated using a manually curated biochar benchmark containing 582 experimental rows from 60 scientific publications. The biochar dataset is not the final scientific target. It provides a controlled and deliberately difficult benchmark containing diverse precursor names, treatment temperatures, elemental compositions, and scientific table structures. The final table aware pipeline produced 521 unique source linked records. The current validated branch represents 33 benchmark papers containing 351 manual benchmark rows. Conservative DOI constrained one to one validation recovered 306 of these rows, corresponding to 87.18% benchmark recovery. Of the 306 matched records, 305 had all comparable carbon, hydrogen, nitrogen, and oxygen values within 0.02 wt% of the manual reference, while the remaining record was within 0.15 wt%.

The importance of the developed workflow is not limited to the current carbon dataset. The same architecture can be adapted by changing the target variables, terminology, scientific constraints, and validation rules. This makes it useful for analysing large scientific literature collections where manual extraction would require substantial time. A validated literature mining system can help researchers identify promising experimental regions, recognise conditions that have already been tested, and prioritise laboratory work before committing materials and experimental resources. The next stage is to adapt the validated extraction schema to hard carbon variables including precursor, carbonization temperature, surface area, pore characteristics, interlayer spacing, Raman measurements, reversible capacity, and initial Coulombic efficiency, followed by application to the much larger hard carbon literature.

Project at a Glance

Item	Current project status
Scientific focus	Data driven understanding of hard carbon development for sodium ion battery anodes
Validation benchmark	582 manually curated biochar rows from 60 scientific publications
Final source linked dataset	521 unique source linked records
Represented validation scope	351 benchmark rows from 33 represented papers
Benchmark eligible extracted rows	420
Conservative one to one matches	306
Benchmark recovery	87.18% of represented benchmark rows
Numerical agreement	305 of 306 matches had all comparable C, H, N, and O values within 0.02 wt%. The remaining match was within 0.15 wt%.
3D modelling	Printed disordered porous carbon models and sectional cuts to inspect pore size, stacking, and connectivity
Next stage	Adapt the validated workflow to hard carbon variables and scale to the larger literature collection

1 Introduction

1.1 Background and Motivation

Sodium ion batteries are being developed as a complementary technology to lithium ion batteries, particularly for applications where material abundance, cost, and supply diversity are important. Sodium resources are abundant and geographically widespread. At the same time, sodium ions are larger than lithium ions, which makes conventional graphite a less effective host under normal sodium ion battery conditions. Hard carbon is therefore a leading anode candidate because its non graphitizable and disordered structure contains enlarged interlayer spacing, defects, and nanopores that can support several sodium storage mechanisms.

The useful structure is not simply carbon containing as much empty space as possible. Sodium storage depends on the arrangement of the carbon layers, the accessibility and dimensions of the pores, the balance between open and closed porosity, and the connectivity of the structure. Closed nanopores have particular importance for low voltage sodium storage, while interlayer regions and defects can also contribute. Very high accessible surface area can increase irreversible reactions with the electrolyte. The objective is therefore to identify a useful combination of pore architecture, carbon ordering, interlayer spacing, and surface chemistry.

Hard carbon can be prepared from many carbon rich precursors, including biomass and waste derived materials. This creates opportunities to use abundant and renewable precursor sources, although the environmental impact depends on the complete production and battery life cycle. Precursor diversity also creates a large scientific data

problem. Different studies use different precursor names, carbonization temperatures, atmospheres, measurement methods, units, and reporting formats. Manually comparing a limited number of studies is possible, but doing the same across thousands of publications becomes extremely time consuming.

For this reason, the project first validates the extraction workflow against an independently curated biochar dataset containing 582 rows from 60 publications. Biochar provides a difficult test because the papers contain highly variable precursor names, treatment conditions, elemental compositions, and scientific tables. The purpose of this stage is to determine whether the code can reliably recover the correct experimental record before the extraction variables are changed to those required for hard carbon research.

1.2 Project Aim and Objectives

The aim is to develop a validated and scalable workflow that can extract structured hard carbon data from scientific literature and use the resulting dataset to identify processing, structure, and performance trends relevant to sodium ion battery anodes.

- Understand disordered and nanoporous carbon structures physically through atomistic visualization and 3D printing.
- Reproduce and inspect the manually curated 60 paper benchmark using elemental ratios, feedstock grouping, treatment temperature, and curve fitting.
- Build a traceable literature mining workflow in which every extracted value can be linked back to its paper and evidence.
- Separate deterministic table processing from semantic model interpretation so that the language model is used only where scientific context is genuinely required.
- Validate the workflow on biochar before changing the target variables to hard carbon descriptors and scaling to a much larger literature collection.
- Develop a workflow that can later be adapted to other scientific extraction problems by changing the target variables, keywords, scientific rules, and validation criteria.

2 Scientific Background

2.1 Disordered and Nanoporous Carbon Structure

Disordered carbons are not completely amorphous. They contain short range graphenic order arranged in curved three dimensional networks. The models studied in this project include graphene fragments, micropores, curvature, line defects, and stacked regions. Martin et al. describe these structures as disordered 3D graphene networks and show that negative Gaussian curvature is important for continuous three dimensional connectivity. Their models span densities from 0.5 to 1.5 g cm-3 and contain around 32,000 atoms.

Density has a strong effect on the visible microstructure. In the million atom structures reported by Putman et al., the 0.5 g cm-3 model contains large pores with little significant stacking. As density increases, pore size decreases

and stacking develops. At 1.5 g cm-3 only a small number of pores remain and stacking becomes more regular, while at 2.0 g cm-3 the structure approaches graphite density, with no visible pores and extended stacking. This trend was especially useful in the 3D printed models because the change from open porous pathways to denser stacked regions could be inspected as a physical object rather than only on a screen.

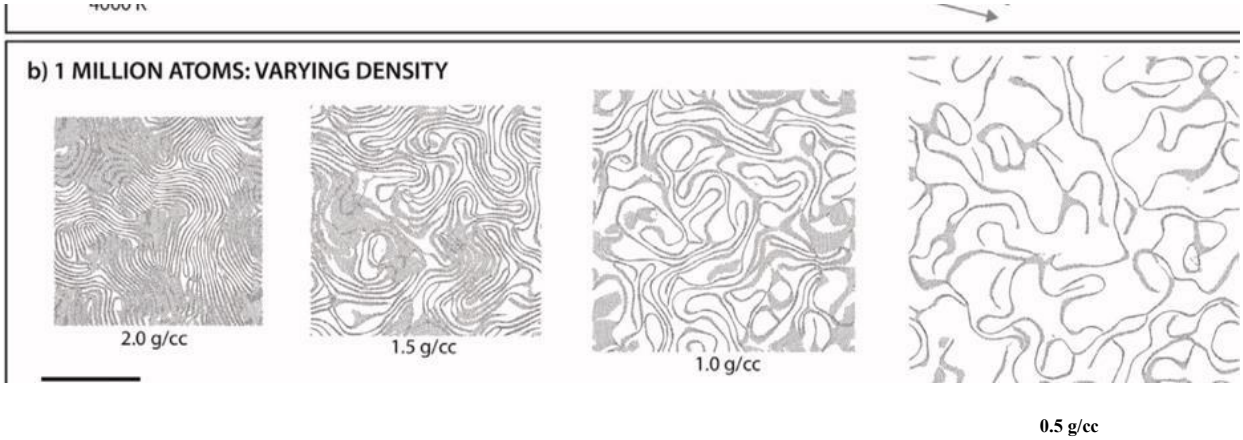


Figure 1. Effect of density on one million atom disordered carbon structures. Increasing density reduces pore size and increases graphene sheet stacking. Adapted from Putman et al. (2023).

2.2 Temperature, Ordering, and Graphitization

Thermal treatment changes the degree of order in nanoporous carbon. The carbide derived carbon simulations of de Tomas et al. reproduce the experimentally observed progression from a disordered microstructure at lower temperature toward graphitic fragments and stacked graphene sheets at higher temperature. This does not mean that all porous carbon becomes hard carbon at one fixed temperature or density. Hard carbon is a non graphitizable carbon class whose final structure depends on precursor chemistry and processing history. Temperature, density, and precursor structure together influence pore architecture, ordering, and stacking.

For sodium ion anodes, the objective is not maximum graphitization. A useful hard carbon structure must retain disorder and appropriate pore architecture while providing sufficient electronic connectivity and sodium storage sites. Excessive ordering can remove pore volume, while excessive open surface area can increase irreversible reactions. The design problem is to find the processing window that gives the required balance of interlayer spacing, closed pores, defects, connectivity, capacity, and initial Coulombic efficiency.

2.3 Why Data Driven Analysis Is Needed

Experimental screening remains essential for hard carbon development because the relationship between precursor chemistry, thermal processing, pore structure, and electrochemical performance is too complicated to predict from one simple structural model. However, testing every possible precursor and processing condition experimentally would require substantial time, materials, and laboratory resources.

A large amount of this experimental information already exists in the scientific literature, but it is distributed across papers using different terminology and formats. The purpose of the literature mining pipeline is to convert this information into a structured and traceable dataset that can be analysed systematically. Instead of replacing

experimental work, the objective is to help narrow the experimental search space. Trends identified across previous studies can help researchers decide which precursor and processing regions deserve further investigation and which conditions have already shown poor or inconsistent outcomes.

The value of this pipeline also extends beyond the current hard carbon project. The extraction architecture separates paper processing, table reconstruction, deterministic scientific rules, semantic interpretation, and final validation. By changing the target terminology, expected variables, acceptable ranges, and validation criteria, the same approach can be adapted to other scientific literature mining problems. This is particularly useful in fields where experimental knowledge is spread across large numbers of publications and manually reconstructing that evidence would take considerable time.

3 Methodology

3.1 Reference Data and Initial Analysis

The validation benchmark contains 582 experimental rows from 60 biochar publications. Core fields include feedstock identity, feedstock group, treatment temperature, and carbon, hydrogen, nitrogen, and oxygen composition. Atomic H/C and O/C ratios were calculated from elemental weight percentages and used for Van Krevelen analysis. Temperature dependent elemental trends were also examined, and selected relationships were fitted using a four parameter decreasing logistic function.

3.2 Paper Conversion, Retrieval, and First Extraction Architecture

The 60 source PDFs were converted to machine readable Markdown and linked to DOI identifiers. The first architecture divided papers into evidence chunks, embedded those chunks, retrieved candidate passages for predefined scientific questions, and supplied evidence packages to a locally deployed Qwen2.5 7B model. The model returned structured records rather than narrative summaries. This stage established that semantic retrieval is useful, but it also showed that embedding alone is not sufficient when the source information is stored in scientific tables.

3.3 Table Aware Architecture

The revised pipeline treats tables as scientific structures rather than ordinary text. It first builds a table inventory, identifies elemental analysis tables, repairs multi level headers, recovers transposed elemental tables, and normalizes candidate tables into a consistent representation. Each row is then classified according to whether it is directly extractable, requires semantic interpretation, or should be rejected.

Table semantics are validated deterministically wherever possible. Candidate queues then separate rows that can be accepted by explicit rules from rows that require the language model. For ambiguous rows, semantic recovery packages combine the relevant row with the table header, caption, nearby text, and paper context. The enrichment stage provides additional paper level evidence so that the model can decide what a sample name or temperature means without being asked to reconstruct information that was removed during conversion.

The semantic recovery model is therefore not used as the first validator. It is used after structural repair and deterministic filtering. The final validation checks whether the recovered record is internally consistent and supported by its source evidence. This division of responsibility is important because loops, conditional rules, range checks, schema checks, and one to one constraints are reproducible, while semantic interpretation is reserved for cases where context is genuinely ambiguous.

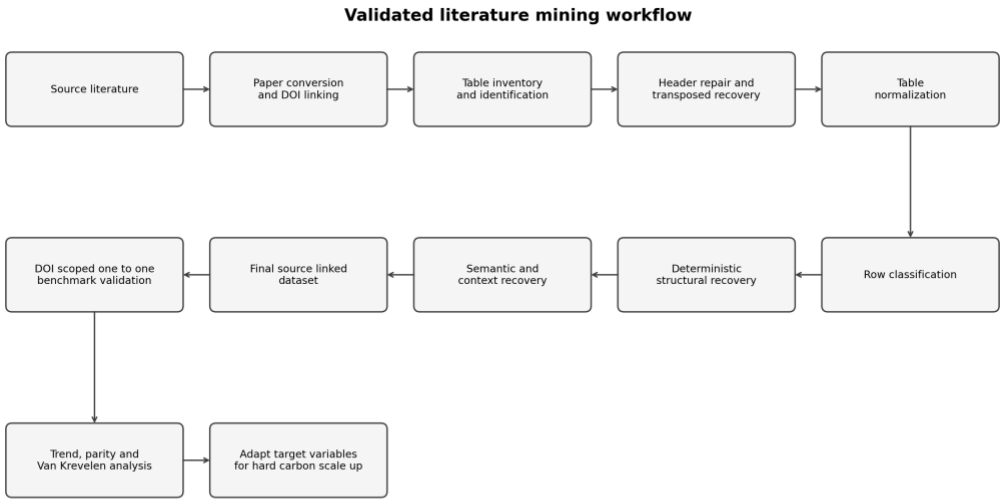


Figure 2. Final project pipeline showing the progression from source literature through table reconstruction, structural and semantic recovery, source linked data, benchmark validation, and scientific analysis.

3.4 Validation Strategy

Validation is performed at several levels rather than using one accuracy number. First, the code checks whether tables and rows satisfy the expected schema and scientific constraints. Second, extracted records are compared against the manual benchmark within the same DOI. Treatment temperature must agree within 0.5 degrees Celsius where both values are resolved, and at least two comparable C, H, N, and O measurements are required for an eligible numerical comparison. Conservative one to one assignment prevents one extracted row from being assigned to several manual rows.

Numerical agreement is then evaluated using the comparable elemental values. Records with all comparable differences within 0.02 wt% are treated as effectively exact, while a broader 0.15 wt% threshold is retained as the reliable comparison limit. The final analysis also separates record recovery from numerical fidelity so that a missing or unresolved record is not confused with an incorrect numerical extraction.

3.5 Scale Up and Literature Access

The validation stage was intentionally completed before attempting the full hard carbon literature extraction. The intended hard carbon corpus is on the scale of several thousand papers, with an initial target of approximately 5,000 candidate publications. A major practical limitation is access to full scientific text. Many relevant

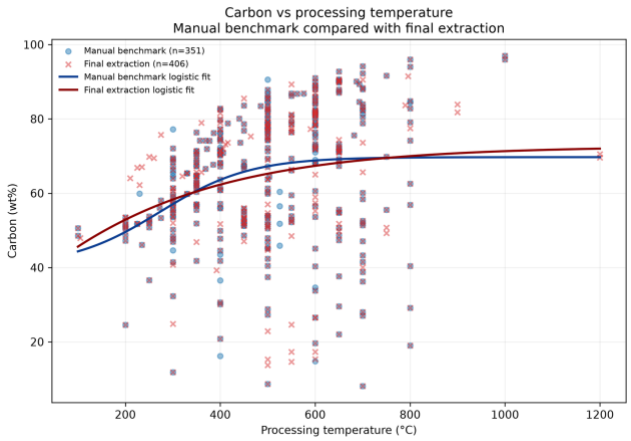
publications are not openly accessible and require institutional access before they can be converted and processed by the pipeline.

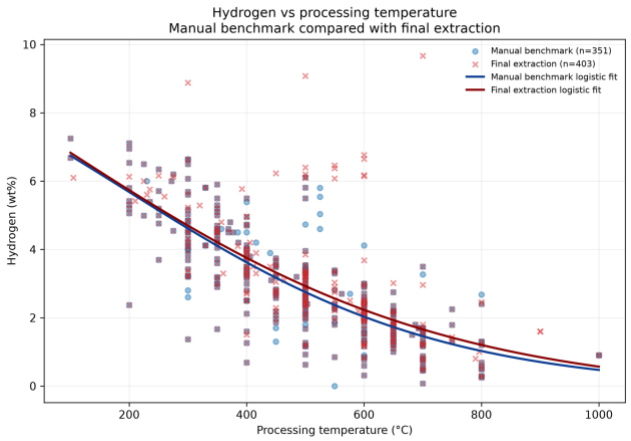
Large scale processing also introduces a significant computational limitation. GPU acceleration is already required for the current 60 paper benchmark because the semantic recovery stage relies on a locally deployed language model, and processing the same workload on the laptop without GPU support would not be practical. Most deterministic table reconstruction and validation steps are relatively inexpensive, but ambiguous records require semantic interpretation and therefore considerably more computation. Extending the same process from 60 papers to several thousand publications will require substantially greater GPU resources and processing time. For this reason, the present work focuses on validating the extraction architecture before committing the additional access and computational resources required for full hard carbon scale up.

4 Results and Discussion

4.1 Benchmark and Extracted Data Trends

The manually curated benchmark provides the reference distribution against which the final extraction can be compared. Across the benchmark, higher temperature records generally move toward lower H/C and O/C ratios, consistent with progressive loss of hydrogen and oxygen containing species during thermal treatment. The final extracted records reproduce the same broad elemental and atomic ratio trends while also retaining source linked rows that are not necessarily represented in the manual benchmark.





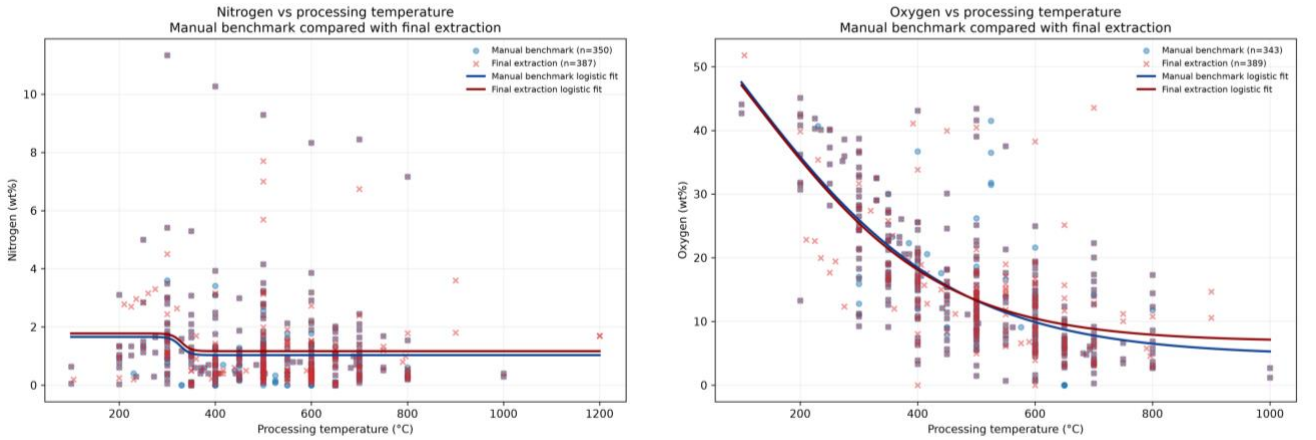


Figure 3. Manual benchmark and final extracted elemental compositions as a function of processing temperature. Panels show carbon, hydrogen, nitrogen, and oxygen. The overlap demonstrates reproduction of the main temperature dependent trends.

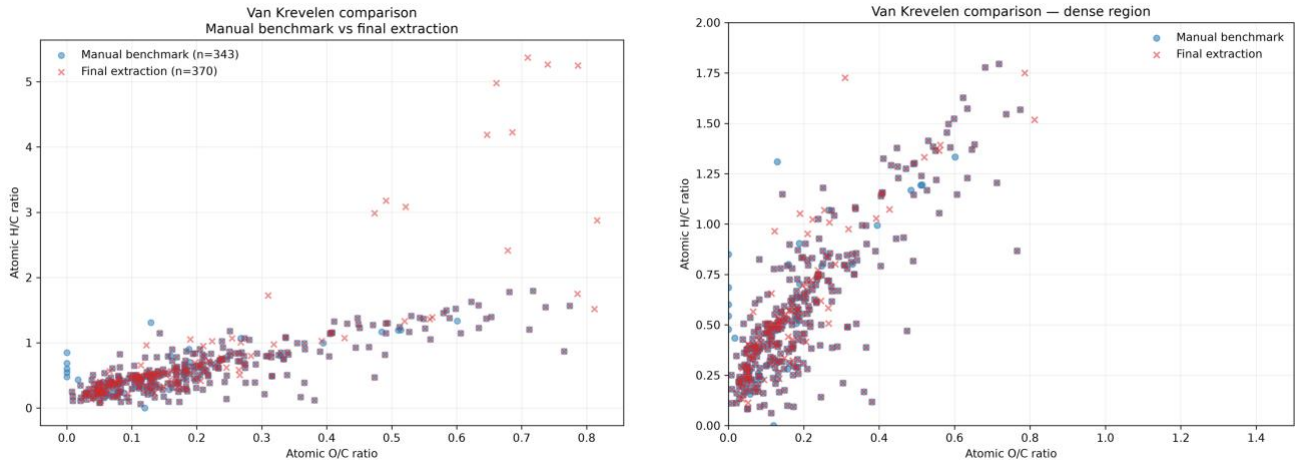


Figure 4. Van Krevelen comparison of the manual benchmark and final extracted dataset. The full view shows the complete atomic H/C and O/C distribution, while the enlarged view provides a clearer comparison of the dense central region.

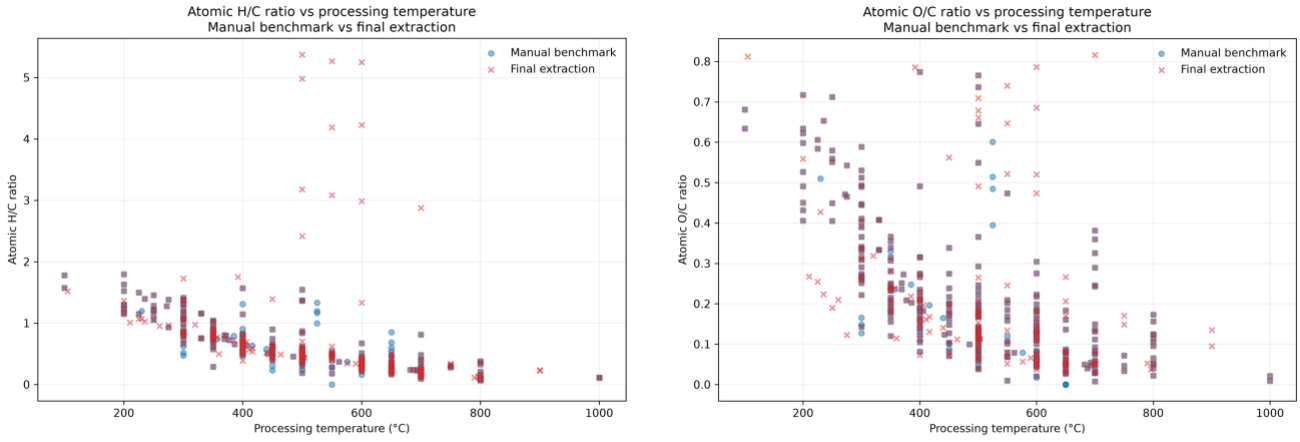


Figure 5. Atomic H/C and O/C ratios as a function of processing temperature for the manual benchmark and final extraction. Both datasets reproduce the general reduction in hydrogen and oxygen relative to carbon as thermal treatment progresses.

4.2 Final Extraction and Benchmark Performance

The final table aware pipeline produced 521 unique source linked records. Of these, 466 were recovered through deterministic structural processing, 52 required semantic recovery, and three retained a source supported temperature range. The final dataset contained 428 records with an exact treatment temperature, 71 for which temperature was not applicable, three range values, and 19 unresolved temperatures.

The final extraction branch represented 33 of the 60 benchmark papers. These papers contained 351 manual benchmark rows, while the extraction produced 420 rows eligible for benchmark comparison. Conservative DOI constrained one to one matching recovered 306 of the 351 represented benchmark rows, corresponding to a benchmark recovery of 87.18%. Forty five represented benchmark rows remained unmatched, and 114 eligible extracted rows were not paired under the conservative benchmark procedure.

The numerical agreement after successful record identification was substantially stronger than the recovery percentage alone suggests. Of the 306 matched records, 305 had every comparable carbon, hydrogen, nitrogen, and oxygen measurement within 0.02 wt% of the manually curated reference. The final remaining match was within 0.15 wt%. Carbon was compared across 298 records, hydrogen across 305, nitrogen across 283, and oxygen across 296. The mean absolute errors were approximately 0.00003 wt% for carbon, 0.00008 wt% for hydrogen, zero for nitrogen, and 0.00034 wt% for oxygen.

This distinction is important when interpreting the final result. The 87.18% value measures conservative record recovery within the represented benchmark subset. It does not mean that the numerical extraction is wrong in approximately 13% of cases. Once the pipeline identifies the corresponding benchmark record, the extracted elemental values are almost identical to the manually recorded values.

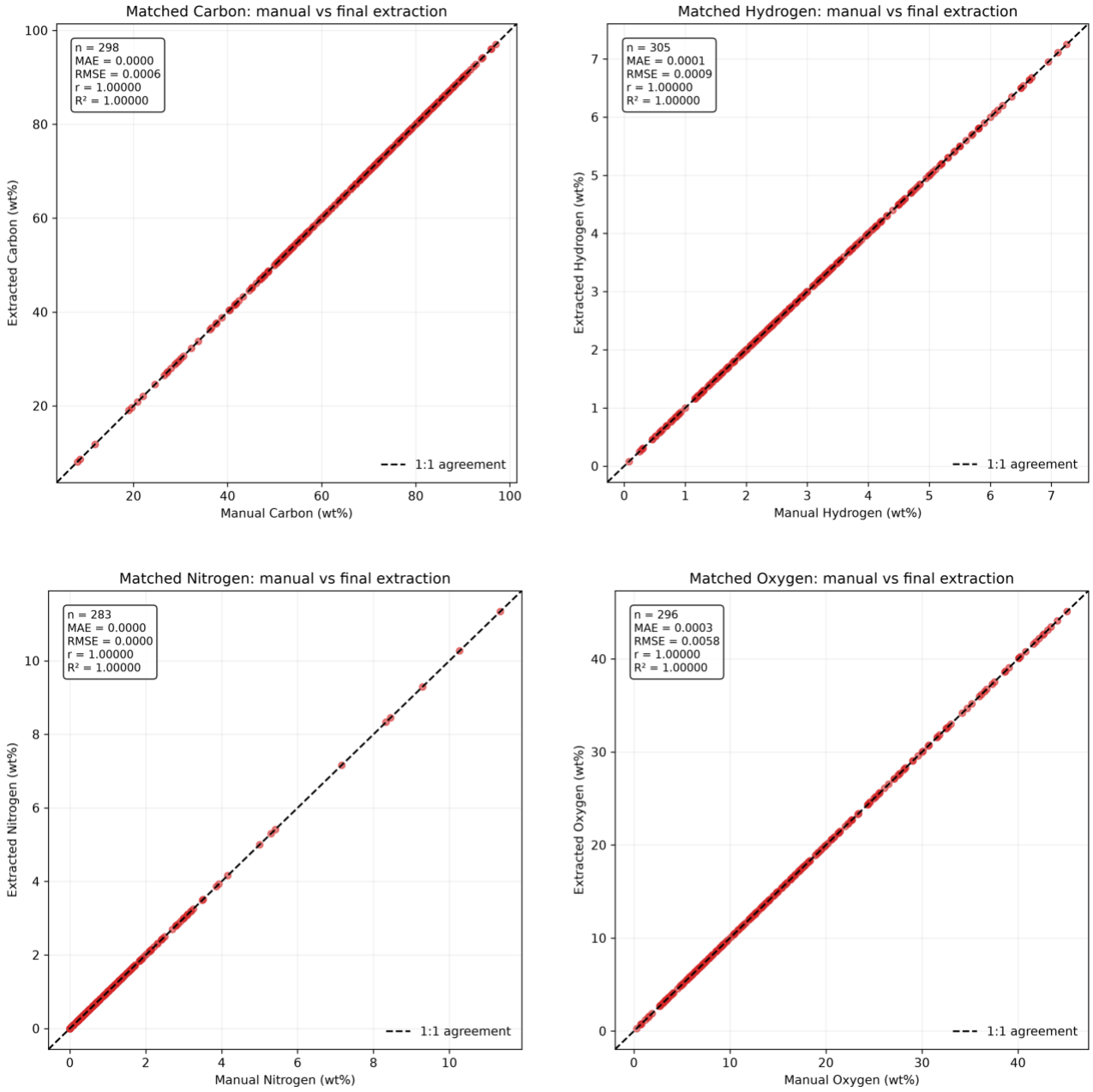


Figure 6. Element level agreement between the manual benchmark and final extraction for matched records. Panels show C, H, N, and O. The matched values lie almost entirely on the one to one line, showing very high numerical fidelity after record identity is established.

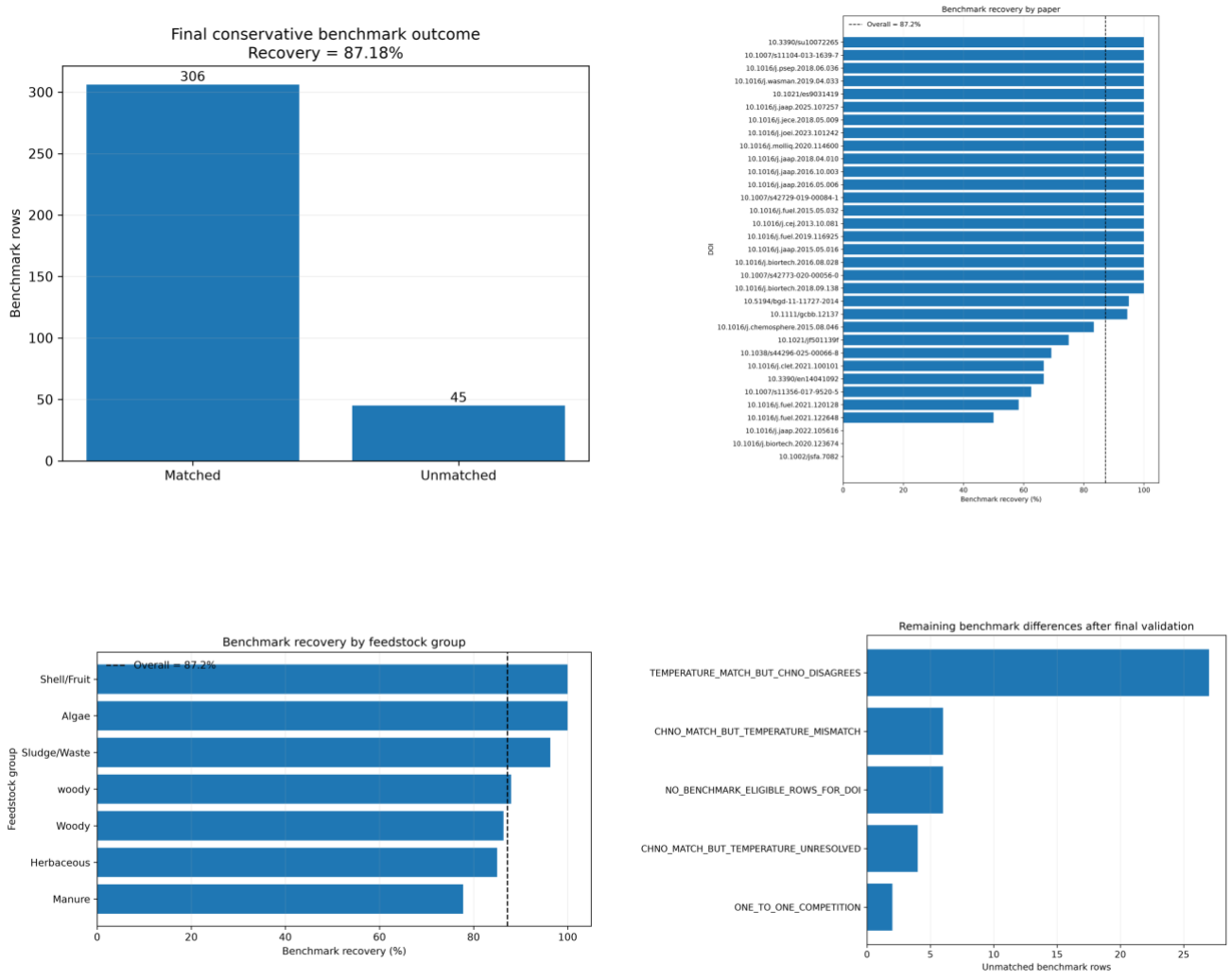


Figure 7. Final benchmark validation performance. The panels summarize matched and unmatched benchmark rows, recovery by paper, recovery by feedstock group, and the classified reasons for the remaining benchmark differences.

4.3 Remaining Benchmark Differences

Forty five represented benchmark rows remained unmatched after final validation. These remaining cases were classified according to the stage at which agreement failed. Twenty seven rows had a compatible treatment temperature but disagreed in one or more elemental values. Six rows had compatible elemental composition but a different treatment temperature. Six rows belonged to DOI cases for which no eligible extracted benchmark candidate was available. Four rows had matching elemental composition but an unresolved treatment temperature, and two rows remained because of competition within the one to one matching constraint.

The unresolved cases therefore do not represent one single type of extraction failure. They include source and benchmark differences, incomplete source information, temperature interpretation problems, coverage limitations, and conservative matching constraints. One example is a group of four benchmark records reported at 525 degrees Celsius while the source evidence available to the extraction pipeline only supports a processing interval of 500 to 550 degrees Celsius. These records were intentionally left unresolved rather than inserting the benchmark value

into the extraction. This preserves the principle that the extracted dataset must be based on the source paper rather than information taken from the validation reference.

A final contextual correction also demonstrated why sample identifiers need to be interpreted scientifically rather than numerically. Labels containing values such as oxygen percentages or mixture ratios had initially been vulnerable to temperature interpretation. Resolving these cases from their surrounding experimental context recovered 12 additional benchmark rows and increased recovery from 83.76% to 87.18%, an improvement of 3.42 percentage points.

4.4 Physical Interpretation from 3D Printing

The 3D printed models provide a complementary physical interpretation of the carbon structure examined computationally. Lower density structures contain larger void regions and more open pathways. As density increases, the carbon sheets move closer together, pore dimensions decrease, and more regular stacking develops. This makes it possible to physically inspect how a disordered carbon network changes as the structure becomes denser.

Multiple sectional cuts were also produced from the same carbon model. Looking at the structure at different cut positions makes it possible to see whether individual void regions continue through the structure, terminate internally, or connect with neighbouring pores. This does not directly simulate sodium diffusion, but it provides a physical representation of pore connectivity that can influence ion transport and storage. The printed models therefore help connect the atomistic description of disordered carbon with the pore and structural parameters that will later be extracted from the hard carbon literature.

5 Challenges and Lessons Learned

- Scientific tables are structural. A number can only be interpreted correctly when its row label, column header, units, caption, footnotes, and surrounding context are preserved.
- Validation code is part of the scientific method. Loops, conditional rules, range checks, one to one matching, and provenance checks prevent a model output from being treated as correct simply because it looks plausible.
- Model capability and evidence quality must be separated. A stronger language model cannot recover a table relationship that was destroyed during PDF conversion.
- The 60 paper biochar benchmark is useful because it is difficult and independently curated. It allows extraction logic to be improved before the target is changed to hard carbon.
- Intermediate outputs and failures should be retained so that an error can be traced to conversion, table repair, retrieval, semantic interpretation, normalization, or matching.
- Full scale hard carbon analysis depends on obtaining sufficient full text access and on computational resources for the semantic recovery stage.

6 Future Work

The benchmark development stage has now reached a final validated extraction result. The next stage is therefore to transfer the workflow from the biochar benchmark to the scientific variables required for hard carbon anode development. These include precursor identity, carbonization temperature and atmosphere, BET surface area, pore volume and pore size information, interlayer spacing, Raman ID/IG, structural ordering, reversible capacity, plateau capacity, initial Coulombic efficiency, current density, and cycling performance.

The intended hard carbon literature collection is substantially larger than the current 60 paper validation benchmark. The literature search is being developed toward a corpus on the order of several thousand candidate publications. Full scale processing cannot begin immediately for every candidate because a significant proportion of the relevant literature requires institutional access rather than being openly available. The next practical stage will therefore include obtaining the accessible full text corpus, maintaining DOI and provenance links, and processing the papers through the validated extraction architecture.

Scaling the workflow will also require additional computational resources because semantic recovery becomes expensive when applied across thousands of papers. Deterministic table reconstruction and scientific validation will continue to be used wherever possible so that the language model is reserved for cases where contextual interpretation is genuinely necessary.

Once the hard carbon dataset has been extracted, the objective will be to identify relationships between precursor choice, thermal processing, pore architecture, interlayer spacing, and electrochemical performance. These trends can then be used to identify promising experimental regions and provide laboratory researchers with clearer starting conditions, reducing unnecessary repetition of previously tested parameter combinations.

The physical modelling work can also be extended. Putman et al. generated disordered carbon structures containing approximately one million atoms, which reduces many of the finite size restrictions of smaller atomistic models and allows larger crystallites, extended stacking, and a wider pore network to develop. Producing a printable representation and sectional models of this structure would make it possible to compare pore connectivity, graphenic stacking, and structural changes over a larger volume.

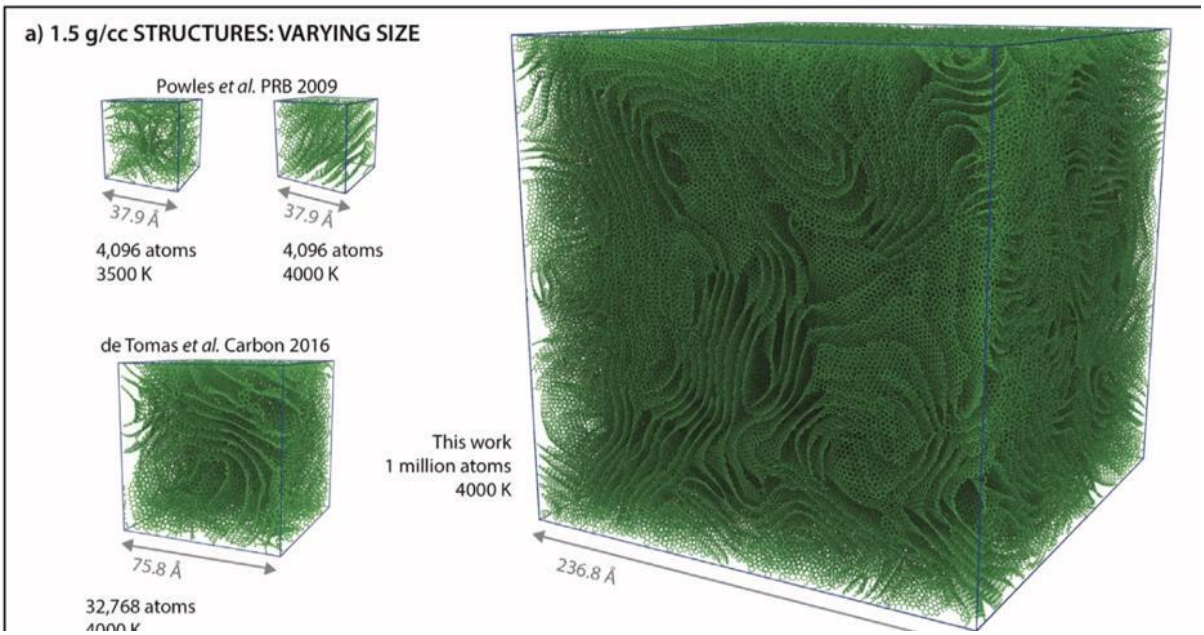


Figure 8. One million atom disordered carbon model proposed for future 3D printing. The larger simulation volume allows structural features to develop with fewer finite size restrictions. Adapted from Putman et al. (2023).

A further development will be to extend the pipeline beyond text and tables to scientific figures and plots. Many important experimental relationships, especially electrochemical performance data such as reversible capacity, plateau capacity, rate capability, and cycling performance, are often reported graphically rather than in numerical tables. A text only extraction system can therefore miss useful information. Future work will investigate methods for identifying relevant figures, interpreting axes and variables, extracting approximate plotted values where appropriate, and connecting those results with the corresponding text and table evidence from the same publication.

This would allow capacity trends and other electrochemical relationships reported in individual papers to be compared across the wider literature. For example, the effect of carbonization temperature, precursor type, pore structure, or current density on capacity could be compared across multiple studies. These cross paper comparisons could provide an additional consistency check for the extracted dataset and help identify trends that are repeatedly supported, trends that depend strongly on material or experimental conditions, and cases where published results disagree. The aim would be to combine numerical tables, text evidence, and graphical capacity trends into a more complete evidence base for evaluating hard carbon performance.

7 Conclusion

This project approaches hard carbon development through physical structural understanding and large scale scientific data analysis. The 3D modelling work provides a direct way to examine how changes in density and thermal treatment influence pore size, carbon sheet stacking, curvature, and three dimensional connectivity.

Printing multiple sections of the same structure also provides a physical way to inspect how pores connect or terminate within the carbon network.

The literature mining work addresses a different problem, which is the scale and heterogeneity of published experimental evidence. A manually curated biochar dataset containing 582 rows from 60 publications was used as an independent benchmark for developing and validating the extraction architecture before applying it to hard carbon. The final pipeline produced 521 unique source linked records. Within the 33 represented benchmark papers, 306 of 351 manual rows were recovered using conservative DOI constrained one to one validation, giving a benchmark recovery of 87.18%. More importantly, 305 of the 306 matched records reproduced all comparable elemental values within 0.02 wt%, showing that numerical extraction is extremely accurate once the correct experimental record has been identified.

The remaining differences were mainly associated with record coverage, temperature interpretation, source and benchmark disagreement, and conservative one to one matching rather than a general inability to extract numerical values. These cases were retained and classified instead of being hidden or manually forced to agree with the benchmark.

The broader value of the work is that the developed architecture is not restricted to this benchmark or to hard carbon. The target terminology, variables, scientific constraints, and validation criteria can be modified for other literature mining problems. A system of this type can help researchers process evidence at a scale that would be impractical manually and can provide experimental teams with a clearer understanding of which conditions have already been investigated and which regions deserve further study.

The immediate next step is to adapt the validated workflow to hard carbon specific structural and electrochemical variables and apply it to the much larger literature collection. The resulting dataset can then be used to connect processing conditions with carbon structure and sodium storage performance, providing more informed targets for future hard carbon experiments.

Acknowledgements

I would like to thank KAUST Academy and KACARE for sponsoring this research experience and providing the opportunity to learn and contribute to energy storage research. I am especially grateful to my supervisor, Dr. Carla de Tomas, for her guidance throughout the project, for sharing her knowledge of carbon materials, and for helping me connect the computational work with the underlying materials science. I also thank her research group for their support and discussions. I would like to thank Francisco Martin Martinez and his PhD student Xiaoqi for sharing their insights and assisting with the data driven part of the project. Their manually curated dataset provided the controlled benchmark that made it possible to test, diagnose, and improve the extraction workflow.

References and Project Sources

[1] C. de Tomas, I. Suarez-Martinez, F. Vallejos-Burgos, M. J. Lopez, K. Kaneko, and N. A. Marks, "Structural prediction of graphitization and porosity in carbide derived carbons," Carbon, vol. 119, pp. 1-9, 2017. https://doi.org/10.1016/j.carbon.2017.04.004

[2] J. W. Martin, C. de Tomas, I. Suarez-Martinez, M. Kraft, and N. A. Marks, "Topology of Disordered 3D Graphene Networks," Physical Review Letters, vol. 123, 116105, 2019. https://doi.org/10.1103/PhysRevLett.123.116105

[3] K. J. Putman, M. R. Rowles, N. A. Marks, C. de Tomas, J. W. Martin, and I. Suarez-Martinez, "Defining graphenic crystallites in disordered carbon: Moving beyond the platelet model," Carbon, vol. 209, 117965, 2023. https://doi.org/10.1016/j.carbon.2023.03.040

[4] Project code repository, Hard Carbon Literature Mining Pipeline: https://github.com/mythoflogic/hard-carbon-literature-mining

Appendix A. Final Validation Figure Files

The report figures above were assembled from the final validation image set. The individual files are retained separately for use in the final report layout and for detailed comparison.

- temperature_overlay_C.png: Manual and extracted carbon composition versus temperature.
- temperature_overlay_H.png: Manual and extracted hydrogen composition versus temperature.
- temperature_overlay_N.png: Manual and extracted nitrogen composition versus temperature.
- temperature_overlay_O.png: Manual and extracted oxygen composition versus temperature.
- parity_C.png: Carbon parity plot.
- parity_H.png: Hydrogen parity plot.
- parity_N.png: Nitrogen parity plot.
- parity_O.png: Oxygen parity plot.
- van_krevelen_overlay_full.png: Full Van Krevelen comparison.
- van_krevelen_overlay_zoom.png: Zoomed Van Krevelen comparison.
- H_C_atomic_vs_temperature_overlay.png: Atomic H/C ratio versus temperature.
- O_C_atomic_vs_temperature_overlay.png: Atomic O/C ratio versus temperature.
- final_benchmark_outcomes.png: Final matched and unmatched benchmark outcome.
- benchmark_recovery_by_paper.png: Benchmark recovery by paper.
- benchmark_recovery_by_group.png: Benchmark recovery by feedstock group.
- remaining_failure_breakdown.png: Breakdown of remaining benchmark differences.