

From Detection to Prioritisation

A Computational Framework for Identifying
Commercially Relevant White Spaces in the
Science–Technology Nexus

Théo Caustur and Grégoire Froment

Théo Caustur (r0878839), Grégoire Froment (r1071535)

Thesis submitted to obtain the degree of
Master Of Business Engineering
Major: Technology and Entrepreneurship

Promoter: Prof. Dr. Bruno Cassiman
Work Leader: Niloufar Farrokhzad
Academic year 2025-2026

Copyright information:

*student paper as part of an academic education
and examination.*

*No correction was made to the paper
after examination.*

From Detection to Prioritisation

A Computational Framework for Identifying Commercially Relevant White Spaces in the Science–Technology Nexus

Science guides technological search, but it does not dictate where invention will take root. Some scientific domains cross rapidly into the patent landscape; others remain structurally disconnected from it, even when they carry knowledge of plausible commercial value. These uncovered regions, referred to as white spaces, matter to firms scouting external science, to investors positioning ahead of technological consolidation, and to policymakers directing innovation funding. Yet existing approaches either detect such gaps without judging their commercial promise, or score individual papers without locating them within the broader topology of the science-technology nexus. This thesis asks whether commercially relevant white spaces can be detected, ranked, and validated within a single framework, and tests the answer on the alternative-fuels sector, a domain in which multiple pathways compete without a dominant design.

Working with roughly 58,600 scientific papers and 34,900 patents, the framework first identifies which scientific clusters sit far enough from the patent frontier to qualify as genuine white spaces, then ranks them by the concentration of commercially promising research they contain. Twenty white spaces emerge from this process, and a strategic matrix positions each along both dimensions. The result reveals an unexpected pattern: structural absence from the patent frontier tells us little about where commercial promise concentrates, and only one cluster scores high on both. A historical test, in which the framework was rerun on data available before 2019, shows that its ranking anticipated where patent-citation attention would subsequently strengthen, an outcome drawn from evidence outside the framework's own construction. The thesis therefore reframes white spaces as a two-dimensional phenomenon, defined by structural absence and characterised by commercial concentration, and shows that distinguishing the two dimensions is what makes the resulting prioritisation strategically informative. The framework does not replace the judgment of firms, investors, or policymakers; it narrows the field within which such judgment becomes worthwhile.

Acknowledgements

We would like to extend our gratitude to everyone who contributed to the completion of this thesis. Our first thanks go to our promoter, Professor Bruno Cassiman, who proposed the original research direction on the science–technology nexus and trusted us to develop it into the framework presented here. Throughout the work, his sharp questioning and methodological rigour shaped its direction at every important turn. His ability to push us toward precision while leaving us the space to develop our own arguments made this thesis stronger than we could have made it on our own.

We are equally grateful to our thesis assistant, Niloufar Farrokhzad, for her thoughtful guidance throughout the methodological development of the pipeline. Her willingness to engage with the technical details, to challenge our reasoning, and to remain available across the many iterations this work required was essential to bringing the framework to its final form.

We also thank the professors of the Faculty of Economics and Business at KU Leuven, whose teaching across the master’s programme gave us the foundation on which this thesis was built. The questions this work asks, and the methods it brings to bear on them, would not have been possible without the intellectual training the faculty provided.

Finally, we thank our families. For their patience through the long stretches of writing, their encouragement during the moments of doubt, and their steady support throughout: this thesis is as much theirs as it is ours.

Table of Contents

1	Introduction	1
2	Literature Review	4
2.1	Introduction	4
2.2	The Science and Technology Nexus	4
2.3	White Spaces: Concept and Prior Detection	7
2.4	Semantic and Predictive Advances	9
2.5	Field Context: Alternative Fuels as a Case	11
2.6	Conclusion and Research Question	12
3	Methodology	14
3.1	Field Eligibility and Research Design	14
3.2	Field Perimeter and Search Strategy	15
3.2.1	Field Perimeter	15
3.2.2	Search Strategy	16
3.2.3	Temporal Scope	17
3.3	Corpus Construction	17
3.3.1	Scientific Corpus	17
3.3.2	Patent Corpus	17
3.3.3	Quality Checks and Preprocessing	18
3.4	Representation Choice and Semantic Embedding	18
3.4.1	Representation Choice	18
3.4.2	Embedding Model	19

3.4.3	Implementation	19
3.5	Clustering and Mapping Design	20
3.5.1	Dimensionality Reduction	20
3.5.2	Clustering Algorithm	20
3.5.3	Cluster Labelling	21
3.6	White Space Detection Logic	21
3.6.1	Distance Computation	22
3.6.2	Distance Gate	23
3.7	Commercial Potential Scoring and Strategic Prioritisation	25
3.7.1	Paper-Level Commercial Potential	26
3.7.2	Matching and Prediction	26
3.7.3	Ranking Valid White Spaces	27
3.7.4	Strategic Positioning With the 2×2 Opportunity Matrix	28
3.7.5	Measurement Constraint in the Compot Layer	30
3.8	Validation Design	31
3.8.1	External Outcome: The Reliance on Science Framework	31
3.8.2	Primary Validation: The Historical Re-Clustering Test	31
3.9	Robustness and Boundary Conditions	33
4	Empirical Findings	35
4.1	Corpus Overview and Topological Landscape	35
4.2	Distance Gate Results and Classification	37
4.3	White-Space Ranking and the 2×2 Opportunity Matrix	38
4.3.1	Ranked List	39

4.3.2	The 2×2 Opportunity Matrix	40
4.3.3	Ranking and Matrix Tell Different Stories	41
4.3.4	What the Four Quadrants Contain	42
4.4	Deep-Dive: PEM Electrolyzer Component Engineering and Transport .	42
4.5	Deep-Dive: Two Deep-Frontier Gap Clusters	45
4.6	Validation Results	46
5	Discussion	49
5.1	What the Pipeline Shows: Distance and Commercial Concentration Are Loosely Coupled	49
5.2	What the Validation Supports: The Pipeline Produces a Predictive Ranking	50
5.3	What the Empirical Findings Cannot Resolve: The Noise Pool and the Static Snapshot	52
5.4	What Kind of Instrument the Pipeline Is	53
6	Conclusion	55
6.1	Research Question and Answer	55
6.2	Contributions	56
6.3	Limitations	57
6.4	Future Work	58
6.5	Closing Remarks	60
	Sources	61
	List of figures	65
	List of tables	67

1 Introduction

The relationship between scientific discovery and technological development is uneven by design: science can guide technological search, but it does not determine where technological activity will actually concentrate (Fleming & Sorenson, 2004; Narin et al., 1997). Some regions of science cross into patented technology and become embedded in industrial use; others remain on the scientific side of the divide for years or decades, despite carrying knowledge that could plausibly be commercialised (Marx & Fuegi, 2020; Shibata et al., 2010). These uncovered regions represent more than an empirical curiosity. They are a strategic opportunity for firms scouting external science, for investors positioning ahead of technological consolidation, and for policymakers directing innovation investment toward areas of high commercial promise. Identifying where such regions sit, and which among them carry the strongest commercial signal, has become a question of practical and methodological consequence, especially because the existing analytical toolkit is only partly adequate.

The phenomenon at the centre of this problem has often been described as a white space. In this thesis, a white space is defined as a region of the science–technology nexus where scientific activity is dense but corresponding technological representation in the observed patent landscape is weak or absent. The concept builds on the work of Shibata et al. (2010) on commercialisation gaps in the solar-cell industry and has since become relevant to innovation studies and business strategy as a way of locating areas where commercially relevant science may exist before it is substantially claimed by technology (Liu et al., 2025; Marx & Fuegi, 2020; Masclans et al., 2025). White spaces matter because they identify where future technological activity may plausibly emerge. Yet the concept has been more clearly described than operationalised. Existing approaches have often treated white spaces as a binary outcome derived from expert interpretation of cluster topics, without providing a continuous measure of which detected regions are more commercially concentrated than others.

The methodological landscape around this problem has changed. Citation-based approaches, exemplified by Shibata et al. (2010), provide a topological logic for comparing scientific and technological landscapes, but they do not rank the gaps they identify. Recent machine-learning approaches to measuring the commercial potential of science, notably Masclans et al. (2025), provide continuous paper-level scores of commercial relevance, but they do not identify whether the papers belong to structurally uncovered regions of the science–technology nexus. These two approaches therefore remain largely separate: one detects structural gaps, while the other scores commercial potential. The resulting limitation is that white-space analysis remains closer to a classification exercise than to a strategic prioritisation framework.

Closing the distance between structural detection and strategic prioritisation requires a framework that does not merely identify white spaces but ranks them by a property external to the detection logic itself, and tests whether that ranking carries meaning beyond the data used to produce it. This is the question addressed in this thesis: *how can commercially relevant white spaces in a multi-paradigm¹ science–technology nexus be detected, ranked, and validated through a computational framework that combines structural mapping with continuous commercial-potential scoring?* The thesis answers this question by separating detection from prioritisation. Structural distance first identifies where science is weakly represented in the observed patent landscape; commercial potential then distinguishes which of those structurally uncovered regions contain a concentration of papers with stronger downstream commercial promise.

To address this problem, the thesis develops a computational pipeline organised as a sequence of stages. First, scientific publications and patent documents are embedded in a shared semantic space. Second, science and patent clusters are identified independently. Third, science clusters are gated according to their structural distance from the patent layer, allowing white-space candidates to be distinguished from already represented or adjacent areas. Fourth, the white spaces identified at the previous step are ranked by how many of their underlying papers carry strong commercial-potential signals. The resulting set is then positioned in a 2x2 strategic matrix that separates structural distance from commercial concentration. Finally, the ranking is validated against an external patent-citation outcome that is independent of the semantic-distance procedure on which the pipeline relies.

The framework is applied to the alternative-fuels sector, a setting well suited to the analysis because of its multi-paradigm structure, sustained scientific and patenting activity, and strategic relevance to energy transition (International Energy Agency, 2024). The empirical corpus contains approximately 58,600 scientific papers and 34,900 patent documents. Applying the framework identifies twenty science clusters as white spaces under the structural criterion. These clusters do not form a single homogeneous opportunity class. Instead, they vary substantially in their distance from the patent landscape and in the commercial-potential concentration of their underlying papers. This supports the thesis’s central claim that structural absence and commercial attractiveness are related but distinct dimensions of the science–technology nexus.

The empirical results also provide external support for the ranking logic. In the strategic matrix, only one cluster, PEM Electrolyzer Component Engineering and Transport,

¹A multi-paradigm sector is a field in which several technological pathways coexist without one having consolidated as the dominant design in the sense of Abernathy and Utterback (1978). In alternative fuels, this refers to the coexistence of pathways such as green hydrogen, advanced biofuels, electrofuels, and sustainable aviation fuels, each with distinct scientific communities, conversion mechanisms, feedstocks, and end-use applications.

combines above-median structural distance with above-median commercial-potential concentration. Applied retrospectively to pre-2019 data, the framework's ranking predicts subsequent patent-citation attention among commercially high-concentrated clusters, with Spearman $\rho = .685$ ($p = .005$). These results suggest that the framework does more than identify semantic gaps: it produces a prioritisation signal that is externally meaningful in relation to later technological attention.

The thesis proceeds as follows. Chapter 2 reviews the literature on the science–technology nexus, white-space detection, the methodological shift from citation-based to semantic approaches, and the alternative-fuels sector as a case. Chapter 3 develops the methodological framework in full, from corpus construction through embedding, clustering, distance gating, commercial-potential scoring, strategic positioning, and validation. Chapter 4 reports the empirical findings produced by applying the framework to the alternative-fuels corpus. Chapter 5 interprets those findings and identifies the conditions under which the framework operates reliably. Chapter 6 states the thesis's contributions, acknowledges its limitations, and proposes directions for extending the framework.

2 Literature Review

2.1 Introduction

This chapter develops the literature against which the thesis's research question is set. It starts from the idea that science and technology are connected, but not mechanically aligned. Scientific knowledge can guide technological search, become visible in patents, and support commercial innovation, yet this movement is uneven. Some scientific domains are rapidly absorbed into technological development, while others remain weakly represented in the patent landscape despite carrying knowledge of possible commercial relevance.

This unevenness is the theoretical starting point of the thesis. If patents capture only part of the science–technology connection, then the absence of patent activity around a scientific domain cannot automatically be interpreted as absence of value. It may instead indicate a structural disconnect between scientific activity and technological appropriation. The white-space concept captures this situation: scientific knowledge exists, but the corresponding patent representation is weak or absent.

The chapter therefore reviews three bodies of work. The first concerns the science–technology nexus and the mechanisms through which science becomes, or fails to become, connected to technological development. The second concerns white spaces and prior attempts to detect them. The third concerns recent semantic and predictive advances that make it possible to move from detecting structural absence to comparing which uncovered scientific clusters contain stronger commercial-potential signals. The chapter closes by explaining why alternative fuels provide an appropriate empirical setting and by deriving the research question that guides the thesis.

2.2 The Science and Technology Nexus

The relationship between science and technology is not a simple linear sequence in which basic research automatically produces applied inventions. Kline and Rosenberg (1986) describe innovation as a chain-linked process in which scientific knowledge, technological development, design, production, and market feedback interact recursively. Scientific research can guide technological search, but technological problems can also redirect scientific inquiry. This view is important for the thesis because it implies that science and technology are connected, but not mechanically aligned.

Empirical work on patents confirms that public science plays an important role in technological development. Narin et al. (1997) showed that US patents increasingly cited scientific publications, especially in high-technology sectors. Fleming and Sorenson (2004) similarly argue that science can act as a map for technological search by helping inventors navigate complex technological landscapes. Marx and Fuegi (2020) extend this tradition by constructing a large-scale database of front-page patent citations to scientific articles, providing evidence that patents frequently rely on prior scientific work. These studies establish that patents can provide a measurable trace of the link between science and technology.

However, the same literature also shows why that trace is incomplete. Science and technology operate as two partly connected but institutionally different systems. Scientific work is organised around publication, cumulative knowledge, and peer recognition, while technological development is organised around appropriability, claims, applications, and market position. Because these systems do not share the same incentives, vocabularies, or time horizons, the movement from science to technology requires translation rather than simple transfer.

The first mechanism behind this uneven movement is therefore a translation problem. Scientific results and patentable inventions are expressed in different languages and organised around different questions. A scientific paper may describe mechanisms, experimental results, materials, or performance characteristics, while a patent describes claims, embodiments, prior art, and possible industrial applications. Bridging these two forms of knowledge requires the ability to recognise which scientific findings can be re-framed as inventions and which complementary assets are needed for commercialisation. This is closely linked to absorptive capacity, meaning the ability of firms to recognise, assimilate, and use external knowledge (Cohen & Levinthal, 1990). Nooteboom et al. (2007) add that cognitive distance is double-edged: some distance between actors can create novelty, but too much distance makes knowledge difficult to absorb. In the context of this thesis, cognitive distance helps explain why scientific clusters may remain disconnected from patent clusters even when their content is potentially useful.

The second mechanism is institutional retreat. Arora et al. (2018) document a shift away from scientific research by large corporations between 1980 and 2006. Their finding does not mean that firms no longer use science, but it suggests that firms may rely more heavily on external scientific knowledge while maintaining less internal capacity to produce it. This matters because firms that do less internal science may also have weaker internal capacity to evaluate, translate, and absorb public scientific knowledge. External science can therefore be commercially relevant without being immediately transformed into patentable technology. When firms have weaker internal scientific capabilities, the movement from public science to private invention may become more selective and uneven (Arora et al., 2018; Cohen & Levinthal, 1990).

The conventional way of observing the science–technology connection is through non-patent literature citations in patents. These citations are useful because they show where patent documents explicitly refer to scientific publications. Yet they have three known limitations.

The first limitation is lag. A scientific article may influence technological development before it is cited by a patent, and additional time may pass before the patent is published and its citations become observable. Citation data therefore records the science–technology connection with delay. This is especially problematic when the objective is to detect emerging scientific regions before they become visibly incorporated into patents.

The second limitation is selection. Patent citations are not a neutral record of intellectual influence. Patent applicants, attorneys, and examiners cite documents partly to establish prior art, define legal boundaries, and support patent examination. The resulting citation record therefore reflects a mixture of knowledge flow, legal strategy, examiner search behaviour, and procedural convention (Alcácer & Gittelman, 2006; Lemley & Sampat, 2012; Marx & Fuegi, 2020). A cited scientific article may not be the only knowledge source behind an invention, and an uncited article may still have influenced technological development. The implication is that patent citations are best read as a hybrid signal: partly evidence of intellectual influence, partly evidence of legal and procedural behaviour. For the purposes of white-space analysis, this hybridity is itself an argument against relying on citation data as the primary detection instrument, because the signal that matters here – whether scientific activity is structurally absent from the patent layer – is precisely the kind of signal that legal-procedural noise is most likely to obscure.

The third limitation is invisibility. Many genuine knowledge flows never appear as formal patent citations. Scientific knowledge can move through researcher mobility, collaborations, technical standards, engineering know-how, conferences, consulting, or informal learning (Cohen & Levinthal, 1990; Marx & Fuegi, 2020). These channels can matter commercially, but they leave little or no trace in patent-to-paper citation data.

Patent citations are therefore informative but partial evidence of the science–technology nexus. They show where science and technology have already become visibly connected, but they cannot fully show where scientific activity exists without corresponding patent representation. This distinction is central to the thesis. Structural disconnects between science and patents should not be treated only as anomalies or data noise. They can emerge from the normal organisation of the science–technology nexus: different incentives, different vocabularies, different institutional locations, imperfect absorptive capacity, and incomplete observability. This is the theoretical basis for studying white spaces.

2.3 White Spaces: Concept and Prior Detection

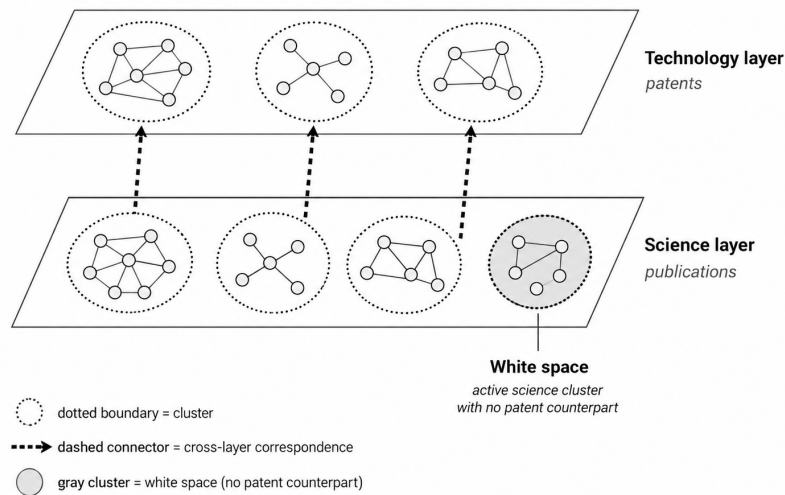
A white space, as used in this thesis, is a region of the science–technology nexus where scientific activity exists but corresponding patent activity is weak or absent. The concept is structural. It does not mean that the science is necessarily better or weak than other science. It means that the science occupies a position that is not strongly represented in the observed patent layer. This distinction matters because a white space is different from a research gap, where scientific knowledge itself is missing, and different from a market gap, where customer demand or commercial adoption is missing. A white space begins from the opposite premise: the scientific knowledge exists, but the patent landscape does not yet reflect it.

The mechanisms responsible for white-space emergence follow from the science–technology nexus reviewed above. Cognitive distance creates a translation problem between scientific knowledge and patentable invention. Institutional retreat weakens the channel through which firms absorb and develop upstream science. A third friction operates at the level of evaluation. Even when commercially relevant science is visible, firms may not recognise its value equally. Industrial search can overweight already validated science, prestigious institutions, familiar domains, or work that fits existing technological trajectories. Masclans et al. (2025) describe this type of evaluation failure as a structural feature of how firms search for commercially relevant science. White spaces can therefore be understood as the result of cognitive, institutional, and evaluative frictions in the science–technology nexus.

Shibata et al. (2010) provide the closest methodological foundation for this thesis. In their study of solar cells, they compare citation networks of scientific publications with citation networks of patents in order to identify commercialisation gaps between science and technology. Their approach clusters publications and patents separately, then compares the two layers to identify scientific regions without corresponding technological representation. Figure 2.1 illustrates this parallel-layer detection logic: scientific publications and patent documents are clustered independently, cross-layer correspondences indicate areas where science is reflected in technology, and the unmatched science cluster represents the white space. This logic is central to the present thesis: science and patent activity should first be mapped independently, then compared across layers.

Figure 2.1

Shibata's parallel-layer detection logic. Dotted boundaries identify clusters, dashed arrows identify cross-layer correspondence between science and technology, and the grey science cluster marks a white space with no patent counterpart.



The strength of Shibata et al. (2010) is that they turn a broad strategic idea into an operational mapping problem. Instead of asking generally whether science is commercialised, they ask where the scientific landscape and the technological landscape fail to overlap. This makes white-space detection strategically actionable because it shifts the analysis from the general claim that science matters to the identification of specific knowledge domains that remain weakly represented in technology. A firm, investor, or policymaker does not need to know only that science matters for technology. They need to know which parts of the scientific landscape are not yet reflected in technological activity.

At the same time, the Shibata et al. (2010) approach has three limitations for the purpose of this thesis. First, it relies on citation networks. This makes the method vulnerable to the citation limitations reviewed above: lag, selection, and invisibility. Second, it primarily identifies gaps as a structural classification. It can show that a scientific cluster has no patent counterpart, but it does not provide a continuous measure of whether that uncovered cluster carries stronger or weaker commercial-potential signals than another uncovered cluster. Third, because the method depends on cluster comparison and expert interpretation, it does not by itself produce a scalable ranking of white spaces across large corpora.

These limitations do not weaken the conceptual importance of Shibata et al. (2010). They define the problem that the present thesis extends. White-space analysis requires structural detection, because a cluster is not a white space unless it is weakly represented in the patent layer. But structural detection alone is not enough for strategic use. Among several structurally uncovered clusters, some may contain science that is closer to future patent activity, while others may remain distant because they are too basic, too systemic, too regulatory, or commercialised through non-patent channels. The literature therefore motivates a second layer: after detecting structural absence, white spaces must be compared by a commercial-relevance signal.

This is why the thesis is structured around two distinct analytical questions. The first question is: does a scientific cluster lie beyond the patent frontier?¹ This is a structural question. The second question is: among the clusters that lie beyond that frontier, which contain the strongest commercial-potential concentration? This is a ranking question. Keeping these questions separate follows directly from the literature on science and technology linkages, where absence from patents and commercial value are related but not identical constructs (Fleming & Sorenson, 2004; Masclans et al., 2025; Shibata et al., 2010). The next section reviews the methodological advances that make it possible to address these two questions together.

2.4 Semantic and Predictive Advances

Recent methodological work makes it possible to address some of the limits of citation-based white-space detection. The first advance is semantic comparison. Instead of linking papers and patents only when a formal citation exists, semantic approaches compare the textual content of documents. Two documents can then be considered related if they discuss similar scientific or technological content, even when no patent-to-paper citation is present. Liu et al. (2025) explicitly compare citation-based and semantic signals in the science–technology nexus and show that natural-language-processing methods can reveal links that are not visible through non-patent-literature citations alone. This is important for the thesis because white spaces are precisely about what is absent or weakly represented in the observable patent layer.

Semantic methods are especially useful when science and patents use different vocabularies to describe related problems. Scientific documents often describe mechanisms, experiments, materials, or performance results. Patent documents often describe claims, embodiments, applications, and legally relevant boundaries. The same underlying knowledge can therefore appear in different textual forms. Domain-adapted

¹The patent frontier here refers to the full set of patent clusters in the shared semantic space, considered as a coverage surface rather than as a single boundary. A science cluster lies beyond it when no patent cluster sits within a calibrated cosine distance.

language models help reduce this vocabulary problem by representing documents as vectors that capture conceptual similarity rather than only shared keywords (Beltagy et al., 2019; Singh et al., 2022). This supports the methodological move from keyword overlap or citation overlap toward distance in a shared semantic space.

For the present thesis, semantic mapping addresses the detection side of the problem. It can help identify whether a science cluster is close to, adjacent to, or distant from patent clusters. It does not answer which distant clusters are commercially promising. A semantic gap may indicate commercially latent science, but it may also reflect a scientific area that is too early, too general, too systemic, or not well suited to patenting. A second signal is therefore required to compare detected white spaces after the structural gate has been applied.

The second advance is the development of predictive measures of the commercial potential of science. Masclans et al. (2025) introduce a machine-learning measure that estimates the probability that a scientific paper will become linked to commercially valuable patented innovation. Their measure is grounded in downstream patent activity, especially patent citations and renewal behaviour, rather than in scientific attention alone. This makes it distinct from traditional academic impact indicators such as publication counts or citations. A paper can be scientifically influential without being commercially relevant, and a commercially relevant paper may not necessarily be the most cited article within its academic field (Masclans et al., 2025).

The contribution of Masclans et al. (2025) is highly relevant but incomplete for white-space analysis. Their measure operates at the paper level. It estimates the commercial potential of individual scientific publications, but it does not identify whether those papers sit in a region of science that is structurally uncovered by patents. Conversely, the Shibata et al. (2010) tradition identifies structural gaps between science and patents, but it does not rank those gaps by a continuous commercial-potential signal. The two approaches therefore answer different parts of the same problem.

This thesis is positioned at the intersection of these two advances. It takes the structural mapping logic from Shibata et al. (2010), replaces citation-network comparison with semantic cluster comparison, and then adds the commercial-potential logic of Masclans et al. (2025) at the cluster level. The result is not simply a method for finding white spaces. It is a method for detecting which science clusters are structurally distant from patents, then ranking those white spaces by the concentration of high-commercial-potential papers they contain. This integration is the methodological gap that Chapter 3 operationalises. Whether the framework is meaningful in practice depends on whether it is applied to a setting in which white spaces are theoretically expected to emerge; the next section identifies the structural features that make a sector suitable, and explains why alternative fuels meet them.

2.5 Field Context: Alternative Fuels as a Case

The mechanisms reviewed in Section 2.3 imply that white spaces should not be expected uniformly across the science–technology nexus. Their prevalence depends on the structural state of the sector in question. Sectors that have stabilised around a dominant design, in the sense developed by Abernathy and Utterback (1978), tend to direct both scientific and technological activity along a shared architecture. This alignment reduces the likelihood that dense scientific clusters develop without corresponding technological representation. Sectors in which several technological pathways co-exist exhibit the opposite pattern. Scientific work fragments across competing trajectories, while patent activity follows incumbent or emerging architectures unevenly, raising the probability that some scientific clusters remain weakly represented in the patent landscape. Multi-paradigm sectors are therefore well suited to white-space analysis, both because structural gaps should be more prevalent and because identifying them carries greater strategic consequence.

Alternative fuels is one such sector. The decarbonisation pressure associated with the energy transition has generated sustained scientific and technological activity, but no single pathway has consolidated as the dominant design. Green hydrogen, advanced biofuels, electrofuels, renewable diesel, and sustainable aviation fuels remain plausible decarbonisation routes, each involving distinct scientific communities, conversion mechanisms, feedstocks, infrastructures, and end-use applications (Intergovernmental Panel on Climate Change, 2022; International Energy Agency, 2023a, 2023b, 2024). The absence of consolidation is visible in the persistence of parallel R&D trajectories and policy support across these pathways: the most recent IEA outlooks treat them as substitutable rather than ordered options, and public investment continues to flow into multiple routes simultaneously rather than concentrating on a single emerging dominant design (International Energy Agency, 2024). This heterogeneity makes the sector analytically useful because it creates precisely the type of science–technology misalignment that white-space analysis is designed to detect: dense scientific activity in some pathways, but uneven or absent patent representation around them.

Within this broader field, the thesis focuses on alternative fuel pathways in which the link between scientific research and technological implementation can be observed through publication and patent data. This includes fuel molecules, conversion processes, combustion systems, emissions control, and related device or component engineering, while excluding adjacent domains such as upstream electricity generation, primary feedstock production, and broad energy-system modelling. The operational details of the corpus delineation are reserved for Chapter 3, Section 3.2.

2.6 Conclusion and Research Question

The literature reviewed in this chapter leads to four conclusions.

First, the science–technology nexus is connected but uneven. Scientific knowledge can guide technological search and become visible in patents, but this transfer is shaped by translation problems, absorptive capacity, cognitive distance, institutional incentives, and the limits of patent-citation observability (Cohen & Levinthal, 1990; Fleming & Sorenson, 2004; Kline & Rosenberg, 1986; Marx & Fuegi, 2020; Narin et al., 1997; Nooteboom et al., 2007). Structural disconnects between science and patents are therefore theoretically plausible and empirically meaningful.

Second, the white-space concept captures one specific form of this disconnect: scientific activity that is not strongly represented in the patent landscape. Shibata et al. (2010) provide the key prior methodology by comparing publication and patent networks to detect commercialisation gaps. Their work establishes the structural logic of white-space detection, but it does not provide a continuous ranking of which gaps are more commercially promising.

Third, recent methods make such a ranking possible. Semantic approaches can compare papers and patents through textual similarity rather than through citations alone (Liu et al., 2025). Machine-learning measures of commercial potential can estimate the downstream commercial relevance of scientific papers (Masclans et al., 2025). Yet these two advances remain separate in the literature: one improves structural detection, while the other improves commercial relevance measurement.

Fourth, alternative fuels provide an appropriate setting for integrating these approaches. The field is technologically plural, strategically important for decarbonisation, and characterised by multiple coexisting pathways whose scientific and patent trajectories may not align perfectly (Intergovernmental Panel on Climate Change, 2022; International Energy Agency, 2023a, 2024). This makes it a suitable case for studying whether structurally uncovered science can be detected and then ranked by commercial-potential concentration.

The gap in the literature is therefore clear. Existing work provides the conceptual basis for white-space detection and the methodological basis for commercial-potential scoring, but it does not combine them into a framework that detects structural absence, ranks validated white spaces, and tests whether the ranking predicts later technological attention.

This thesis addresses that gap through the following research question:

How can commercially relevant white spaces be detected, ranked, and validated in a multi-paradigm sector by combining structural mapping with cluster-level commercial-potential scoring?

Chapter 3 develops the methodological framework.

3 Methodology

The preceding chapter established why white spaces can emerge in the science–technology nexus and why alternative fuels provides a suitable empirical setting for studying them. This chapter translates that conceptual argument into an operational research design. It explains how the scientific and patent corpora were delimited, prepared, embedded, clustered, compared, ranked, and validated, with the aim of making each methodological choice transparent before the empirical findings are presented in Chapter 4.

The chapter is organised around the sequence followed by the pipeline. It first defines the field perimeter and retrieval strategy, then describes corpus construction and preprocessing. It then explains the representation choice, the use of SPECTER2 embeddings, the clustering design, and the computation of cosine distances between science and patent cluster centroids. These distance measures are used to detect which science clusters sit beyond the patent frontier. Only after this detection step does the chapter introduce cluster-level commercial potential (hereafter compot), which is used to rank validated white spaces rather than to define them. Strategic positioning is added afterwards through a 2×2 opportunity matrix that reintroduces structural distance alongside compot to distinguish clusters that are commercially strong but structurally different. These operations are logically sequential, and conflating them, as earlier iterations of the pipeline demonstrated, produces distorted rankings. The chapter therefore presents each step with its rationale before moving to the next.

3.1 Field Eligibility and Research Design

The study adopts a quantitative exploratory design that combines bibliometrics with computational linguistics. It is quantitative because every stage of the pipeline, from embedding through clustering, distance computation, and external validation, operates on large-scale numerical representations of text rather than on close reading of individual documents. It is exploratory because the set of white spaces is not specified *ex ante*: which science clusters lie beyond the patent frontier, and which of those carry commercial weight, are outcomes of the pipeline rather than inputs to it.

A semantic approach is required because the two corpora, more than 93,000 documents in total, express related ideas in systematically different vocabularies (Kline & Rosenberg, 1986). Manual review is infeasible at that scale, and keyword matching cannot bridge the two vocabularies reliably. Shared semantic embeddings are the only representation in which the science–patent comparison becomes tractable.

The choice of empirical field therefore matters: the pipeline can only operate in a domain where the science and patent layers are both large enough, comparable enough, and temporally deep enough to be meaningfully related. Alternative fuels satisfies these conditions: sufficient scientific density to produce stable clusters, sufficient patent activity to define a meaningful frontier, semantic comparability between scientific and patent abstracts, and more than a decade of temporal depth for downstream outcomes to be observed, with patents that remain informative as a commercialisation proxy (Narin et al., 1997). As established in Section 2.5, the sector's pre-paradigmatic structure means that white spaces are expected to be numerous and varied rather than residual, making it analytically well-suited to a framework designed to detect structural disconnects between science and technology.

3.2 Field Perimeter and Search Strategy

3.2.1 Field Perimeter

The domain was scoped around the alternative fuels S–T nexus, encompassing low-carbon fuel production, fuel conversion, and end-use application in hard-to-abate transport sectors, aviation, heavy-duty freight, and maritime shipping, where direct electrification remains limited and chemical energy carriers are expected to retain a structural role (Intergovernmental Panel on Climate Change, 2022; International Energy Agency, 2023a). Although the analytical focus is on hard-to-abate transport applications, the retrieval perimeter intentionally includes upstream production technologies whose scientific development is not sector-specific but remains constitutive of those downstream pathways.

Within this parameter, three paradigms were included: green hydrogen, advanced bio-fuels, and electrofuels, also referred to as power-to-liquid or power-to-X systems. This scope captures both upstream production science, from electrolysis and catalytic synthesis to fermentation and gasification, and downstream application science, from combustion behaviour and engine adaptation to emissions characterisation and materials compatibility. Two exclusions were made explicit. Battery-electric pathways were excluded because they belong to a fundamentally different innovation system centred on electrochemical storage rather than chemical energy carriers; including them would have contaminated the clustering structure with literature whose technological trajectory and patenting patterns diverge from those of chemical fuels. Broad first-generation biofuel themes, such as corn-ethanol debates and food-versus-fuel arguments, were also not the analytical focus. This filtering logic is reflected in the search strategy described in the following section.

3.2.2 Search Strategy

After defining the substantive perimeter of the field, the next step was to translate that perimeter into reproducible retrieval rules for the scientific and patent corpora. The retrieval logic was designed as a search architecture rather than as a one-shot keyword export, so that the same conceptual perimeter could be applied consistently across document types while still respecting their different vocabularies.

On the science side, the query followed a dual-block design that operationalises the S–T nexus directly:

[Fuel terms] AND ([Production terms] OR [Application terms])

The fuel block included core alternative fuel concepts (green hydrogen, biofuels, biodiesel, electrofuels, sustainable aviation fuel, and related variants). The production block captured upstream conversion processes (electrolysis, pyrolysis, gasification, fermentation, catalysis, Fischer–Tropsch synthesis, CO₂ hydrogenation, and related). The application block captured downstream use-side concepts (combustion, engine and turbine behaviour, emissions, NO_x, soot, aftertreatment, materials compatibility). The complete Web of Science query string is reproduced in Appendix F.

This dual-block structure is not merely a retrieval convenience: it reflects the conceptual architecture of the framework. A white space can arise in two structurally distinct ways, either a fuel is scientifically producible but not yet technologically integrated into downstream applications, or downstream technological development proceeds ahead of the corresponding scientific production base. Requiring at least one production or application term alongside a fuel term keeps retrieved documents within the S–T nexus rather than in adjacent domains such as energy policy or agricultural economics. The query was executed in the Web of Science Core Collection using the Topic Search (TS) field, which simultaneously searches titles, abstracts, author keywords, and Keywords Plus.

The patent corpus was retrieved from Lens.org using a parallel but not identical strategy: the same conceptual perimeter (fuel × production/application) was preserved on the patent side, but operationalised through CPC (Cooperative Patent Classification) classes anchoring the search across the alternative fuels value chain, supplemented by keyword logic to recover field-specific language for emerging pathways such as electrofuels, which may not yet be fully represented in the CPC taxonomy. The full Lens query and CPC list are reported in Appendix B and Appendix F. Both exports were downloaded on 17 February 2026.

The two retrieval strategies were parallel in purpose, both aimed to capture the full alternative fuels value chain from production to application, but adapted in wording and structure to the conventions of each document type. Requiring identical query strings across the two corpora would have biased retrieval toward the subset of documents where scientific and patent vocabularies happen to overlap, systematically missing precisely the conceptual matches that the semantic embedding layer is designed to recover.

3.2.3 Temporal Scope

The analysis covers 2010–2025, capturing the modern advanced-fuels era that emerged following the regulatory shifts of the late 2000s and early 2010s, notably the EU Renewable Energy Directive (RED I) (European Parliament and Council, 2009) and the revised US Renewable Fuel Standard (RFS2) (U.S. Environmental Protection Agency, 2010), which redirected R&D investment toward non-food feedstocks and advanced fuel pathways. The upper bound of 2025 corresponds to the most recent complete calendar year available at the time of data collection. Because patent publication lags filing by approximately 18 months (World Intellectual Property Organization, n.d.), patent absence near the temporal boundary is interpreted cautiously, since some relevant patents may have been filed but not yet published.

3.3 Corpus Construction

3.3.1 Scientific Corpus

The scientific corpus comprises 58,628 records published between 2010 and 2025, retrieved from the Web of Science Core Collection using the dual-block query described in Section 3.2.2. All records contained non-empty abstracts and DOI values at the time of embedding.

3.3.2 Patent Corpus

The patent corpus comprises 34,869 records retrieved from Lens.org using the keyword-and-CPC strategy described in Section 3.2.2. All records carry unique Lens identifiers, cover publication years from 2010 to 2025, and contain non-empty abstracts. The corpus breaks down into patent applications (20,911), limited patents (7,462), granted patents (5,891), and smaller numbers of search reports and patents of addition.

3.3.3 Quality Checks and Preprocessing

Iterative quality checks confirmed topical coverage across the three included paradigms, namely green hydrogen, advanced biofuels, and electrofuels, in both corpora, and verified the absence of systematic contamination from adjacent domains such as conventional petrochemistry or battery storage. These checks informed iterative refinement of the retrieval strategy rather than serving as formal statistical tests.

Before entering the pipeline, raw exports were merged and key metadata fields were cleaned into a consistent format. DOI strings were normalised by removing prefix formatting and enforcing consistent casing, a step that proved critical because the cluster-level commercial-potential matching described in Section 3.7 depends on exact DOI joins between the science corpus and the Zenodo-based compot dataset of Masclans et al. (2025). Publication years, titles, abstracts, and document identifiers were also standardised into the column structure required by the embedding, clustering, and compot-matching scripts. Records with missing or very short abstracts were removed upstream, and both embedding scripts applied an additional runtime filter to retain only records with usable abstract text.

3.4 Representation Choice and Semantic Embedding

3.4.1 Representation Choice

Both corpora were embedded from abstract-level text. This reflects a comparability constraint rather than a preference for abstracts over other document sections. Abstracts are the most consistently available text unit across the two corpora, every record in both datasets contains one, whereas full-text availability varies, and patent claims express legal scope in a register that differs substantially from scientific prose. An abstract-versus-claims design would have introduced an asymmetry in representational depth between the two corpora and made cross-domain cosine distances harder to interpret. Abstracts are also semantically denser than titles, which would have discarded the conceptual content needed for fine-grained clustering. The trade-off is that commercially relevant signals contained in methods sections, results, or patent-claim language remain unobserved by the embedding layer. The same trade-off applies to the commercial-potential scoring stage (Section 3.7) and is revisited as a boundary condition in Section 3.9.

3.4.2 Embedding Model

SPECTER2 (Singh et al., 2022) was selected as the embedding model because the pipeline requires document-level representations of scientific and technical text rather than keyword-level similarity. SPECTER2 extends SciBERT (Beltagy et al., 2019) with citation-guided training: papers that cite each other, or are cited in similar contexts, are encouraged to occupy nearby positions in the embedding space. This is useful for the thesis because white-space detection depends on recovering coherent research areas even when related documents do not share identical vocabulary. In other words, the model is expected to capture topical and conceptual relatedness more effectively than a general-purpose transformer or a purely lexical representation.

The cross-domain requirement is separate. The pipeline measures cosine distances between science-cluster centroids and patent-cluster centroids, so both corpora must be embedded in a single representational space. Applying one encoder to both scientific papers and patent abstracts avoids the incommensurability that would arise from using separately trained models for each corpus. This does not mean that SPECTER2 is a patent-specialised model; rather, it provides a common semantic coordinate system in which science and patent abstracts can be compared consistently.

This choice remains a design assumption. SPECTER2 was trained primarily on scientific literature, not patent documents, and its representational fidelity for patent abstracts is therefore not guaranteed. Patent abstracts are closer in style to scientific abstracts than patent claims are, which makes the transfer plausible, but not automatic. The validation results reported in Section 3.8 provide empirical support that the resulting distances and rankings carry useful signal in the alternative-fuels corpus; they do not establish SPECTER2 as a generally valid science–patent embedding model.

3.4.3 Implementation

The embedding pipeline used the `allenai/specter2_base` model with the `allenai/specter2` adapter from the AdapterHub library (Poth et al., 2023). Each abstract was passed through the model with a maximum input length of 256 tokens, producing a 768-dimensional vector representation per document. The resulting embedding matrices contain 58,628 science vectors and 34,869 patent vectors in a shared 768-dimensional semantic space.

3.5 Clustering and Mapping Design

As developed in this thesis, white spaces are properties of knowledge bodies rather than of individual papers or patents. Whether a research direction sits beyond the patent frontier is a question about whether a group of semantically related scientific work has a corresponding group of patents that cover it. Detection therefore operates on cluster-level objects rather than on raw document vectors: we cluster each corpus independently and represent each cluster by its centroid in the embedding space, so that the science–patent comparison is reduced to an assessment between regions of the shared geometry. This section describes the dimensionality reduction used for clustering, the choice of clustering algorithm, and the procedure used to assign interpretable labels to the resulting clusters.

3.5.1 Dimensionality Reduction

Before clustering, each corpus was reduced from 768 to 50 dimensions using UMAP (McInnes et al., 2018). The two reductions were performed independently: the science corpus was reduced separately from the patent corpus. This independence is methodologically important. A joint reduction would have allowed the distribution of one corpus to influence how the other corpus is compressed, contaminating cluster formation before the science–patent comparison had even begun. Separate reductions ensure that science clusters and patent clusters emerge from the internal structure of their respective corpora alone.

The choice of 50 dimensions over a lower target such as two or three was driven by analytical rather than aesthetic considerations. Two-dimensional projections are useful for visualisation, but they remove too much of the semantic structure for the clustering algorithm to identify coherent groups; a 50-dimensional target preserves far more of the original embedding geometry while still providing the computational benefits of dimensionality reduction. UMAP parameters for both corpora are reported in Appendix C. A joint two-dimensional UMAP projection of the two corpora was also produced, but it serves descriptive visualisation only and plays no role in the analytical pipeline; all subsequent operations, centroid computation, distance measurement, and white-space classification, were carried out in the original 768-dimensional SPECTER2 space.

3.5.2 Clustering Algorithm

The clustering algorithm we used is HDBSCAN (McInnes et al., 2017). It allowed us to identify dense regions in the 50-dimensional reduced space of each corpus. This

choice was preferred over centroid-based alternatives such as K-means for two reasons. First, HDBSCAN does not require the number of clusters to be specified in advance, a critical property in an exploratory design where the cluster count is itself an outcome rather than a prior assumption. Second, HDBSCAN does not force every document into a cluster; documents that do not belong to any dense region are classified as noise rather than being assigned to the nearest centroid. In this pipeline, forcing isolated or topically ambiguous documents into clusters would have distorted cluster centroids and potentially generated spurious white-space classifications downstream.

Baseline parameters were tuned separately for each corpus to reflect their different sizes and density profiles; the full configuration is reported in Appendix C. The science-side clustering produced 39 clusters containing 38,020 documents and 20,608 noise points (35.1%); the patent-side clustering produced 46 clusters containing 21,587 documents and 13,282 noise points (38.1%). Section 3.9 reports the sensitivity analyses that confirm both the rank structure and the noise share are stable across alternative HDBSCAN configurations.

3.5.3 Cluster Labelling

To assign interpretable names to the numerically indexed HDBSCAN clusters, we developed an LLM-assisted labelling procedure. For each cluster, 25 abstracts were randomly sampled and submitted to Qwen3.5-Plus (qwen3.5-plus-2026-02-15), a commercial model available through Alibaba Cloud Model Studio (Alibaba Cloud, 2026). We selected this model because the task required short, structured, domain-specific labels from batches of scientific abstracts rather than open-ended generation, and Qwen3.5-Plus supports structured output and long-context text processing while remaining operationally accessible through a standard API. The model belongs to the broader Qwen family documented by Yang et al. (2025). The prompt instructed the model to identify the specific alternative fuel topic uniting the sample and to return a concise two- to five-word label. The prompt included explicit negative instructions against generic labels and anchored the model’s domain knowledge in the relevant technology families; the full prompt template is reproduced in Appendix B. Labels were reviewed manually and corrected where necessary using a single-cluster patch mode that allowed individual labels to be updated without rerunning the full pipeline.

3.6 White Space Detection Logic

The clustering procedure described in Section 3.5 produced 39 science clusters and 46 patent clusters embedded in the same 768-dimensional SPECTER2 space. These clusters are the building blocks of detection, but on their own they do not yet identify

any white space, which is not a property of a single cluster but a relational property between a science cluster and the surrounding patent landscape. We therefore approach detection as a three-step procedure: clustering first, distance-based detection second (this section), and commercial prioritisation third (Section 3.7). The distance step answers the structural question of whether a science cluster sits beyond the current patent frontier, and is kept separate from the ranking step for the reasons set out at the start of this chapter.

3.6.1 Distance Computation

To answer the structural question framed above, we need a single quantity that summarises, for each science cluster, how far it sits from its nearest patent neighbour in the embedding space. Computing this quantity proceeds in three steps: each cluster is first summarised by its centroid, the cosine distance between centroids is then computed for every science–patent pair, and the result is reduced to the distance to the nearest patent cluster for each science cluster. We treat these in turn.

Each cluster is summarised by its centroid, defined in Equation 3.1 as the arithmetic mean of the SPECTER2 embeddings of its member documents:

$$c_k = \frac{1}{n_k} \sum_{i \in C_k} x_i, \quad \forall k \in \{1, \dots, K\} \quad (3.1)$$

where $x_i \in \mathbb{R}^{768}$ is the SPECTER2 embedding of document i , C_k is the set of documents assigned to cluster k , $n_k = |C_k|$ is the number of documents in cluster k , and K is the total number of clusters in the corpus (39 for the science corpus, 46 for the patent corpus).

Although clustering operates in the UMAP-50D space introduced in Section 3.5.1, all centroids and all cosine distances were computed in the original 768-dimensional SPECTER2 space. The two spaces serve different purposes: UMAP-50D groups documents efficiently but distorts distances between regions of the map, while the full 768D space preserves the cosine geometry SPECTER2 was trained on. Detection depends on measuring distances between clusters rather than identifying them, so the comparison was carried out in 768D.

For each science cluster s and patent cluster p , Equation 3.2 computes the cosine distance between their centroids:

$$d(s, p) = 1 - \frac{c_s \cdot c_p}{\|c_s\| \|c_p\|}, \quad \forall s \in \{1, \dots, S\}, \forall p \in \{1, \dots, P\} \quad (3.2)$$

Cosine distance was selected over Euclidean distance for two reasons. First, SPECTER2 was trained so that related papers point in similar directions in the embedding space and unrelated papers point in different directions, with similarity measured by cosine (Singh et al., 2022). Cosine is therefore the geometry the model was built for. Euclidean distance would measure the same space with the wrong ruler: it factors in the length of each vector as well as its direction, even though SPECTER2 was never trained to make vector length carry any meaning. Second, cosine distance is bounded in a stable, interpretable range and is the standard similarity measure for transformer-based document representations, which makes the resulting thresholds easier to compare across studies using the same model family.

Applying Equation 3.2 to every science–patent pair yields a full 39×46 cross-corpus distance matrix. For each science cluster s , the structural measure of patent-frontier proximity is the distance to its semantically nearest patent cluster:

$$d_{\min}(s) = \min_{p \in \{1, \dots, P\}} d(s, p), \quad \forall s \in \{1, \dots, S\} \quad (3.3)$$

Equation 3.3 identifies the nearest patent cluster, which is the strongest candidate for already covering the science: if any patent cluster were going to overlap topically, it would be the closest one in the shared semantic space, so $d_{\min}(s)$ is the only quantity that enters the detection rule.

3.6.2 Distance Gate

We use $d_{\min}(s)$ to assign each science cluster to one of three structural categories.

Table 3.1

Distance-gate classification rule

Range of $d_{\min}(s)$	Classification
$d_{\min} < 0.020$	FALSE POSITIVE: patents already cover the science
$0.020 \leq d_{\min} < 0.026$	ADJACENT: partial coverage, boundary case
$d_{\min} \geq 0.026$	VALID: genuine white space beyond the patent frontier

Two design choices require justification: why the gate is trinary rather than continuous, and how the two cut-points were chosen.

Why three categories rather than a continuous score. A natural alternative would be to keep $d_{\min}$ as a continuous variable and let it propagate into the ranking stage, so

that clusters further from the patent landscape rank higher. We deliberately avoided this. Detection and ranking answer different questions: is this a white space, and how commercially promising is it? Mixing them creates a perverse outcome: a cluster could end up highly ranked simply because it sits far from any patent, even if its commercial-potential signal is weak. Using distance as a gate rather than a score confines it to its proper role: deciding which clusters enter the ranking stage at all, and leaving the ranking itself to commercial potential.

Why three categories rather than two. A binary covered / not-covered split would be cleaner, but it would force every borderline cluster into one camp or the other and inflate both error types: clusters with genuine partial patent coverage would either be wrongly classified as white spaces or wrongly dismissed as already-patented. The intermediate ADJACENT category absorbs this borderline zone explicitly. Clusters in it are too close to the nearest patent neighbour to be confidently called white spaces, but too far to be dismissed as already-patented science. Setting them aside as a separate class keeps the VALID pool free of borderline cases and the FALSE.POSITIVE pool free of clusters that may still hold opportunity.

How the cut-points were chosen. The gate operates on centroid distance, but the placement of the cut-points required substantive judgement that the distance distribution alone could not provide. Whether a given $d_{\min}$ value marks a meaningful boundary between covered, partially covered, and uncovered science depends on what each pair of science and patent clusters actually contains: whether they address the same topic, draw on the same technological vocabulary, and target the same application space. Without that substantive layer, the choice of any threshold would be arbitrary.

The cut-points were therefore calibrated through an LLM-assisted comparison of each science–patent cluster pair. For each of the 39 science clusters, its nearest patent neighbour was identified, and 25 sampled science abstracts together with the patent cluster’s titles, abstracts, and CPC classifications were submitted to Qwen under a fixed prompt asking for a structured-JSON verdict in {VALID, FALSE.POSITIVE, ADJACENT} based on substantive topical correspondence. The prompt scored pairs on overlap of subject matter, technological function, and CPC scope, and was never given the $d_{\min}$ values themselves, ensuring that the qualitative layer rendered an independent judgement. The full prompt template and an example output are reproduced in Appendix B. The values 0.020 and 0.026 are the $d_{\min}$ levels at which the resulting verdict classes separate most consistently in the alternative fuels corpus; they are calibrated operating points for this domain and should be recalibrated before transfer to any other.

Robustness of the cut-points. Because the cut-points were calibrated rather than read directly off the geometry, their exact placement carries some uncertainty that warrants pressure-testing. We therefore conducted a second LLM-assisted review of the science–patent cluster pairs in the threshold-sensitive region around 0.020 and 0.026. The review used a different model and a richer evidence pack, including cluster labels, top terms from both the science and patent sides, and sampled records from each, and asked a different question: rather than re-classifying each pair, it asked where the cut-points would be placed if calibration were performed afresh on this evidence. The review suggested that the current thresholds are plausible but slightly permissive: stricter operating points around 0.022 and 0.0275 would also be defensible. We retained the original thresholds for consistency with the reported pipeline outputs, and because their substantive impact is small; only a handful of clusters near the boundaries would shift category under stricter cut-points, and the top of the commercial-potential ranking reported in Chapter 4 is essentially unchanged. We therefore treat 0.020 and 0.026 as calibrated operating points rather than as definitive boundaries, with boundary cases interpreted qualitatively in the discussion of results.

Result. Applying the calibrated gate to the 39 science clusters produced 20 VALID, 14 ADJACENT, and 5 FALSE_POSITIVE classifications. The 20 VALID clusters form the structurally confirmed white-space set, and pass forward to the commercial prioritisation step described in Section 3.7.

3.7 Commercial Potential Scoring and Strategic Prioritisation

Section 3.6 established which science clusters sit beyond the current patent frontier: the distance gate produced 20 VALID white spaces out of 39 science clusters. Detection, however, only answers one of the two questions posed in this chapter. It tells us where the patent landscape is silent; it does not tell us whether that silence matters commercially. A white space might mark a genuinely promising area of science, or it might reflect a body of work with weak commercial appropriability. Distinguishing the two requires a second methodological layer, which this section develops.

We import a paper-level commercial-potential signal from Masclans et al. (2025), aggregate it to the cluster level, and use it to rank the 20 VALID white spaces by commercial concentration. A 2×2 opportunity matrix then reintroduces structural distance alongside this ranking to support strategic interpretation in Chapter 4.

3.7.1 Paper-Level Commercial Potential

We operationalise the commercial potential signal using the framework developed by Masclans et al. (2025), whose measure is imported at the paper level and then aggregated at the cluster level, following the cluster-level logic introduced at the start of Section 3.5. In their framework, the commercial potential (hereafter compot) score of a scientific paper is the probability that the paper will later be cited by a patent that is subsequently renewed. This grounds the score in a market-driven outcome rather than in scientific attention: maintaining an issued patent requires the owner to pay periodic fees, and non-payment causes the patent to lapse. Renewal is therefore a revealed-preference signal that the underlying invention is still perceived as commercially valuable. A high compot score indicates that the paper contains textual and conceptual features similar to papers that have previously been linked to commercially valuable patents.

For the present thesis, this signal is repurposed conservatively. We do not modify the compot definition, we import the scores of Masclans et al. (2025) directly wherever we can, and we aggregate the paper-level signal to the cluster level only after the detection stage has run. Compot therefore enters the pipeline as an imported input, not as a newly constructed measure.

3.7.2 Matching and Prediction

To apply the compot signal at the cluster level, we needed a score for every paper in the corpus. Masclans et al. (2025) made the paper-level scores from their study publicly available through a Zenodo release (Masclans et al., 2024), which contains the compot value for each paper in their original sample. However, our thesis corpus only partially overlaps with that sample: the Zenodo release covers the papers selected for their own study, not the alternative fuels papers retrieved through our search strategy. Some of our papers therefore have a published compot score while others do not, and the analysis required a strategy that handled both cases.

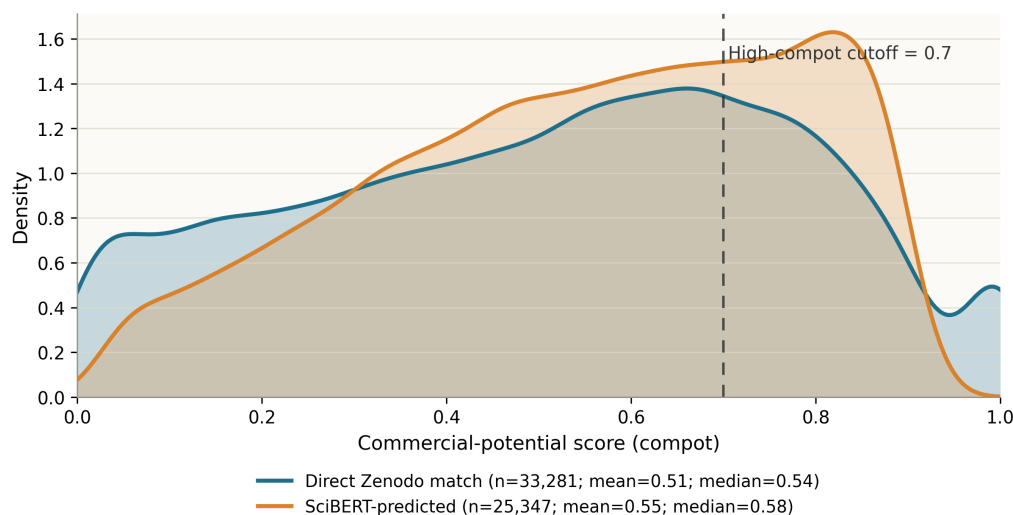
Direct matching against the Zenodo release was the preferred route wherever possible, since it imports the original score without modification. DOI-level matching linked 33,281 papers (56.8% of the corpus) to existing compot scores, leaving 25,347 papers (43.2%) without one. Dropping the unmatched papers would have meant aggregating each cluster’s compot signal over a non-random subset of its members, since the Zenodo release reflects the original corpus selection rather than a random sample of alternative fuels research; the cluster ranking would then have inherited that selection bias. To fill the gap, we fine-tuned a SciBERT regression model (Beltagy et al., 2019) on the matched subset, training it to predict compot from abstract text, and used it to

assign scores to the unmatched papers. The goal was coverage, not improvement on the original measure. Full training configuration and hyperparameters are reported in Appendix C.

Held-out test performance was $R^2 = 0.57$. This value is acceptable for cluster-level aggregation, where paper-level noise averages out across the hundreds or thousands of papers in a cluster, but too noisy to support paper-level claims. We therefore use predicted compot only as an input to cluster-level summaries, never to make statements about individual papers. To keep the two layers traceable, the merged file retains a `compot_source` flag (`zenodo` for directly matched scores, `scibert_predicted` for inferred scores) preserved through every downstream step, making it possible to rerun any cluster-level analysis on the matched-only subset as a robustness check.

Figure 3.7.1

Paper-level commercial-potential score distributions by source. The blue curve shows directly matched Zenodo scores and the orange curve shows SciBERT-predicted scores; the dashed line marks the 0.7 cutoff used to define high-compot papers.



Directly matched scores are imported from the Masclans et al. Zenodo release; unmatched papers receive SciBERT-predicted scores. Density curves use the final paper-level compot score retained for downstream cluster aggregation.

3.7.3 Ranking Valid White Spaces

With a compot score now available for every paper in the corpus, we can turn to the question that motivated this section: which of the 20 VALID white spaces are most commercially promising? Compot is a paper-level measurement, but the white spaces are clusters, so we need a way to summarise the compot signal of each cluster into a single comparable number. The most obvious candidate would be the cluster's mean compot, but a mean tends to compress clusters toward the middle of the score distribution: a cluster with a handful of standout papers and many ordinary ones can end

up looking similar to a cluster of uniformly average papers, even though the two are commercially quite different. The problem is sharpened by the mixed provenance of the paper-level scores in our corpus: some come directly from the Zenodo release, others from the SciBERT predictor of Section 3.7.2, so averaging blurs the signal we most want to surface. We therefore use the concentration-based summary defined in Equation 3.4: the share of papers in a cluster whose compot exceeds 0.7, denoted `pct.above_07`:

$$\text{pct.above_07}(k) = \frac{1}{n_k} \sum_{i \in C_k} \mathbb{I}(\text{compot}_i > 0.7), \quad \forall k \in \{1, \dots, K_{\text{VALID}}\} \quad (3.4)$$

where C_k is the set of papers in VALID cluster k , $n_k = |C_k|$, $K_{\text{VALID}} = 20$, and $\mathbb{I}(\cdot)$ is an indicator function equal to 1 when the condition inside the parentheses is satisfied and 0 otherwise. The sum therefore counts the number of papers in cluster k whose compot score exceeds 0.7.

A high `pct.above_07` indicates that the cluster contains a dense pocket of clearly high-potential papers near the top of the compot distribution, which is the substantive question the ranking is trying to answer. The 0.7 threshold is a corpus-specific operating point chosen to sit well above both the direct Zenodo-matched mean of 0.51 and the full-corpus mean of 0.530, so that `pct.above_07` captures clearly high-potential papers rather than merely above-average ones. To check that the ranking is not driven by this exact cutoff, the metric was recomputed across alternative cutoffs from 0.50 to 0.85; the resulting rank orders remained highly stable, with identical top-5 membership across all tested thresholds. The full sensitivity results are reported in Appendix H.

The 20 VALID white spaces are then ranked in descending `pct.above_07`. Distance does not re-enter the ranking at this stage, having already done its work at the detection gate. The resulting rank order communicates which validated white spaces are most commercially concentrated, but it does not, by itself, reveal everything that matters for strategic interpretation.

3.7.4 Strategic Positioning With the 2×2 Opportunity Matrix

The ranking developed in Section 3.7.3 identifies which validated white spaces are most commercially concentrated, but it tells only half of the strategic story. Two clusters can share a similar `pct.above_07` value while sitting at very different structural positions relative to the patent landscape: one may have just crossed the VALID threshold

and still sit relatively close to existing patent activity, while another may lie deep in semantically uncovered territory. The ranking treats these two cases identically; strategic interpretation should not.

To recover the second dimension, we plot the 20 VALID white spaces on a 2×2 matrix using the two variables the pipeline has already produced. The horizontal axis is $d_{\min}$ (Section 3.6.1) and the vertical axis is `pct_above_07` (Section 3.7.3). Each axis is split at the median of the 20 VALID clusters: $d_{\min} = 0.0319$ for distance and `pct_above_07` = 0.2023 for commercial concentration. The matrix thus partitions the validated white spaces into four quadrants, each carrying a distinct strategic interpretation.

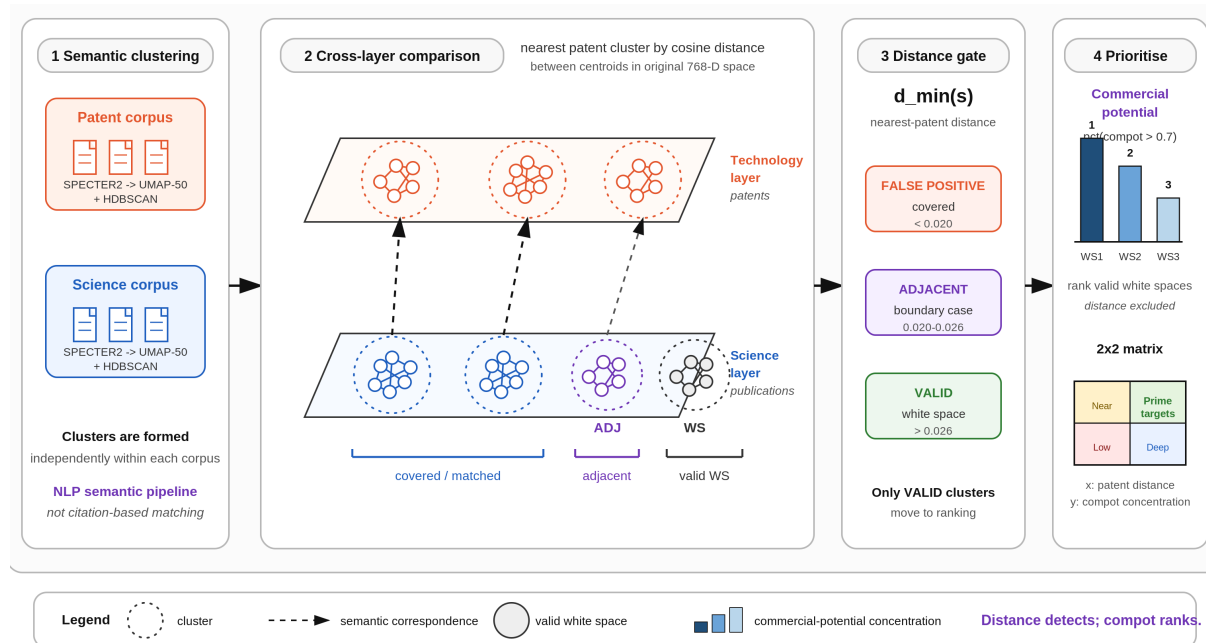
- **Prime Targets** combine high commercial concentration with high structural distance: validated white spaces where a dense pocket of high-potential papers sits in patent territory that remains relatively open. Both axes point in the same direction, making these the most strategically salient clusters in the validated set.
- **Near-Frontier Opportunities** are commercially concentrated but structurally close to existing patent activity. They pass the VALID gate, so they remain white spaces by the detection criterion, but their structural margin is thinner than the rest of the validated set; capturing the opportunity may require navigating around adjacent patent claims.
- **Deep-Frontier Gaps** sit far from the patent landscape but show weaker commercial concentration. The structural opening exists, but the compot signal does not corroborate that the science is heading toward patent-appropriable value. One plausible reading, taken up in Section 3.7.5, is that commercial value in some of these clusters travels through non-patent channels that compot cannot see.
- **Low Priority** clusters score weakly on both dimensions: structurally close to existing patents and without a strong commercial-potential concentration. They remain validated white spaces, but neither axis points to them as strategic priorities.

The matrix does not displace the ranking; it complements it. A cluster's rank reports how concentrated its commercial potential is, and its quadrant position reports how much patent territory remains open around it. These are two different questions that a single rank order cannot answer simultaneously. The labels are deliberately structural-positional rather than temporal: terms such as "emerging," "closing," or "long-term" are avoided, because the matrix is a static cross-sectional positioning device and observes no movement over time. We apply the matrix to the 20 VALID white spaces in Section 4.3.

The full pipeline framework is summarised in Figure 3.7.2.

Figure 3.7.2

Methodological framework: science and patent corpora are clustered semantically, compared by inter-centroid cosine distance, classified as false positives, adjacent, or valid white spaces, and prioritised by commercial-potential concentration.



3.7.5 Measurement Constraint in the Compot Layer

One limitation of the commercial-potential signal is inherent to its definition and must be stated here, at the point where the measurement enters the pipeline, rather than deferred to a later interpretive section. Compot is trained on patent-citation and patent-renewal outcomes. It is therefore structurally blind to commercial value that does not travel through patents. Value realised through technical standards, regulation, consulting, systems integration, or open-source technical infrastructure is not captured by the Masclans et al. (2025) framework, and it is not captured by our extension of it. A cluster in which most commercialisation occurs through non-patent channels will register a low compot signal even when the underlying science is commercially important.

This is a property of the measurement instrument, not a failure of its application. It defines a specific boundary condition under which the ranking should be read: `pct_above_07` tells us which clusters are most likely to generate patent-appropriate commercial value, not which clusters are most commercially valuable in general. The distinction matters most for clusters whose technological content plausibly travels through standards or regulatory pathways, a point revisited interpretively in Chapter 5, but it should

be acknowledged here, at the methodological level, as a structural constraint on what compot can and cannot see.

3.8 Validation Design

The methodology developed in Sections 3.4 through 3.7 is only as credible as the validation that supports it. A pipeline that produces a ranking of 20 white spaces must demonstrate that the ranking corresponds to something real: clusters placed at the top of the ordering should differ in substantively meaningful ways from clusters placed at the bottom, and that this difference can be observed using evidence the pipeline did not itself produce. This section describes how we addressed that requirement, organised around one decisive test: a historical re-clustering of the 2018 data, with patent citations as the outcome.

3.8.1 External Outcome: The Reliance on Science Framework

To test whether the ranking is meaningful, we need an outcome that the pipeline did not itself produce. If validation relied on something computed from the same embedding space as the ranking, the test would risk confirming the ranking by construction rather than by evidence. This concern shaped the choice of outcome here, with earlier alternatives discussed in Appendix G.

Patent citations meet this requirement well. They are recorded directly in patent documents by examiners and applicants, completely outside the embedding pipeline, and they capture exactly what the ranking is implicitly predicting: whether scientific work crosses into patented technological activity.

The citation data was drawn from the Reliance on Science framework (Marx & Fuegi, 2020, 2023), a worldwide mapping of patent-to-paper citations. We matched our 58,628-paper corpus to it through DOI resolution via OpenAlex (Priem et al., 2022), linking 3,429 papers (6.1%) to at least one patent citation. This rate is a technical artefact of cross-database matching, not a sign that science-to-patent transfer is rare in this field. Each match has to pass through three separate identifier systems, and some papers drop out at each step. Even so, the 3,429 matched papers are enough to support cluster-level comparisons.

3.8.2 Primary Validation: The Historical Re-Clustering Test

The central validation question is simple: if we had run this pipeline at the end of 2018, using only information available at that time, would the ranking it produced predict

the later patent-citation attention that actually occurred afterwards? If the ranking is meaningful, clusters the pipeline identified as commercially promising in 2018 should be the ones that drew the most patent-citation attention in the years that followed.

The test required reconstructing the entire pipeline under historical conditions. Every step, including SPECTER2 embedding, UMAP reduction, HDBSCAN clustering, centroid computation, distance measurement, and cluster-level compot aggregation, was rerun on a restricted corpus containing only papers and patents dated 2018 or earlier, so that no post-2018 information entered the historical ranking at any point. To guard against parameter dependence, we ran the historical pipeline under two HDBSCAN configurations: a baseline historical configuration producing 46 science clusters and a coarser setting producing 26 clusters. The 0.026 distance threshold calibrated on the current corpus does not transfer cleanly to the pre-2019 clustering, so for the historical analysis we ranked clusters directly by `pct_above_07` without applying a distance gate; distance returns as a filter and as a control variable in the second of the two tests described below.

The validation works through two tests, each examining a different population of historical clusters and answering a different version of the predictive question. The first test examines whether the predictive relationship is present across all historical clusters, regardless of their structural position relative to the patent frontier. Two statistics are computed against post-2018 patent-citation incidence. A Mann–Whitney U test compares the upper and lower halves of the pre-2019 `pct_above_07` ranking, asking whether the more-promising group received more citations than the less-promising group on average. A Spearman rank correlation then captures the same relationship as a graded one, across the entire ranking rather than as a split at the median. Together, these two statistics establish whether the ranking has any predictive content across the full set of historical clusters.

The full historical ranking, however, includes clusters that sat close to existing patent activity in 2018. These clusters were not plausibly white spaces at that time, and their citation patterns may reflect ongoing commercial activity rather than new patent attention to uncovered science. Because the thesis claim concerns specifically the subset of clusters that lay beyond the patent frontier, the second test restricts the analysis to the most-plausibly-white-space clusters of 2018. Since the 0.026 distance threshold does not transfer to the pre-2019 clustering, we substitute a top- k filter on nearest-patent distance: the top 15 clusters by distance in the 46-cluster configuration, and the top 10 in the 26-cluster configuration. The choice of 15 and 10 reflects the goal of the test, which is directional rather than threshold-precise: we needed a subset of historical clusters that could plausibly be treated as white spaces under each configuration, large enough to support rank-correlation testing but restricted enough to exclude clusters sitting visibly close to existing patent activity. The two k values were set in proportion to

their respective cluster counts, so that the historical white-space subset is comparable in scale to the 20 VALID clusters identified by the main pipeline. Re-running the LLM-assisted calibration on the two historical configurations would have substantially increased token cost without changing the directional logic of the test.

Within this restricted subset, the test computes a Spearman rank correlation between pre-2019 `pct_above_07` and post-2018 patent citations, asking whether the ranking remains predictive among clusters that genuinely correspond to the thesis’s white-space definition. A complementary partial Spearman correlation then adds the critical control: it removes the part of the citation signal explainable by nearest-patent distance alone, addressing the concern that distant clusters might attract more citations through novelty effects or rapid-growth dynamics rather than through commercial-potential concentration. If the `pct_above_07` ranking were a dressed-up distance measure, the partial correlation should collapse toward zero; if it carries independent predictive content, the partial correlation should remain meaningfully positive. The partial Spearman is therefore the statistic on which the central thesis claim most directly depends.

The full numerical results of both tests are reported in Section 4.6.

3.9 Robustness and Boundary Conditions

The pipeline rests on a number of methodological choices, and this section asks how far its outputs can be trusted given those choices. Some constraints have already been raised at their natural locations, namely the compot measurement constraint (3.7.5), the corpus-specific calibration of the distance gate (3.6.2), and the cutoff sensitivity for `pct_above_07` (3.7.3), and are not fully restated here. What this section adds is what those local caveats cannot capture: robustness checks that span the whole pipeline, and boundary conditions that emerge from interactions between methodological layers.

The first question is whether the rank structure produced by our chosen HDBSCAN configuration survives reasonable alternatives. We reran the full pipeline under four configurations spanning coarse (25 clusters) to fine (62 clusters). Pairwise Spearman rank correlations between the resulting VALID rankings range from $\rho = .976$ to $\rho = 1.000$ (mean $\rho = .988$), confirming that the aggregate rank structure is substantially preserved across resolutions. Individual cluster identities at the top of the ranking are less stable than the rank structure itself. PEM Electrolyzer Component Engineering, for example, is absent from the coarse configuration and splits into two clusters under the fine one. Statements about the overall ranking are therefore robust, but statements about any single named cluster depend on the specific clustering configuration used in the main pipeline.

The science-side noise share itself varies only between 33.3% and 36.3% across these four configurations. The size of the noise pool is therefore a property of the corpus, not of the parameter choice. What that noise pool contains, and what it means for the framework’s coverage of the field, is taken up in Section 5.3.

A second set of boundary conditions arises because the SPECTER2 embedding and the compot regression both operate on abstract text alone. Commercially relevant signals carried in methods sections, results, or patent-claim language are invisible to both layers. SPECTER2 itself was trained primarily on scientific text, so its fidelity for patent abstracts is a design assumption. The validation in Section 3.8 supports the assumption, but does not turn it into a guarantee.

The LLM-assisted layer of the pipeline depends on a single model for two distinct tasks: cluster labelling (3.5.3) and gate calibration (3.6.2). A stronger calibration design would have submitted every abstract in each cluster to an exhaustive pairwise comparison with the corresponding patent cluster, rather than the representative samples used here. This was not feasible within available token budgets. We relied instead on sampled prompts, together with the second-pass review described in 3.6.2.

The central validation in Section 3.8 was bounded by the 6.1% matching rate between our corpus and the Reliance on Science framework. The matched set is large enough to support cluster-level comparisons, but the validation observes only the fraction of citation activity that survived cross-database matching. Stronger linkage between bibliographic databases would have allowed a finer-grained test.

A final boundary condition concerns the detection layer rather than the ranking layer. Scientific work taken up through trade secrets, proprietary development, or other unpublished mechanisms is invisible to the patent corpus. Some apparent white spaces may therefore reflect undisclosed rather than absent technological activity. This parallels the compot blind spot raised in 3.7.5, applied to detection rather than to ranking. Its substantive implications for alternative fuels, where commercial value also flows through standards, regulation, and demonstration projects, are taken up in Section 5.3.

Taken together, these constraints define the scope within which the pipeline gives reliable signals. They do not invalidate it: the rank structure is robust, the boundary conditions are explicit, and the substantive claims of Chapter 4 are bounded by the methodological scope this section makes visible.

4 Empirical Findings

This chapter reports what the pipeline produces when applied to the alternative fuels corpora retrieved in Section 3.2. The findings serve two purposes at once. Substantively, they characterise the structural and commercial topology of the alternative fuels science–technology nexus. Methodologically, and more consequentially for the thesis as a whole, they demonstrate that the pipeline produces interpretable, structured, and externally validated output of the kind it was designed to produce. The alternative fuels findings are the worked instantiation; the pipeline itself is the contribution.

Section 4.1 sets out the empirical landscape. Section 4.2 reports the distance gate's classification of the 39 science clusters. Section 4.3 presents the ranking of the 20 VALID white spaces and the 2×2 opportunity matrix. Sections 4.4 and 4.5 examine three clusters that test the framework's interpretability against substantive content: PEM Electrolyzer Component Engineering and Transport as the only Prime Target, and Marine Alternative Fuel Technologies together with Low Carbon Transport Fuel Pathways as contrasting Deep-Frontier Gap cases. Section 4.6 reports the external validation evidence.

4.1 Corpus Overview and Topological Landscape

The scientific corpus comprises 58,628 peer-reviewed records published between 2010 and 2025. Annual output grew from 1,661 papers in 2010 to 6,428 in 2025, a 3.9-fold increase, with the most pronounced acceleration in the second half of the window, consistent with intensifying policy and commercial investment in low-emissions fuels through the 2010s and early 2020s (International Energy Agency, 2023a, 2023b). Publication is concentrated in a small number of field-defining journals: *Fuel*, *Bioresource Technology*, *Energies*, *International Journal of Hydrogen Energy*, and *Renewable Energy* together cover the main scientific pathways the pipeline was designed to capture.¹

The patent corpus comprises 34,869 publication records retrieved from Lens.org over the same period. Unlike the science corpus, patent publication volume does not show sustained expansion: annual counts plateau in the late 2010s and remain broadly flat through 2025, averaging around 2,300 records per year. This asymmetry is the empirical condition the pipeline was designed for. Scientific activity in the field is accelerating, while patenting activity remains broadly steady. This creates the kind of setting in which

¹*Fuel* and *Energies* anchor combustion and fuel-characterisation research; *Bioresource Technology* leads biofuel production and fermentation science; *International Journal of Hydrogen Energy* covers green hydrogen and electrolysis; *Renewable Energy* captures cross-cutting systems-level analysis. CPC class distribution across the patent corpus is reported in Appendix A.

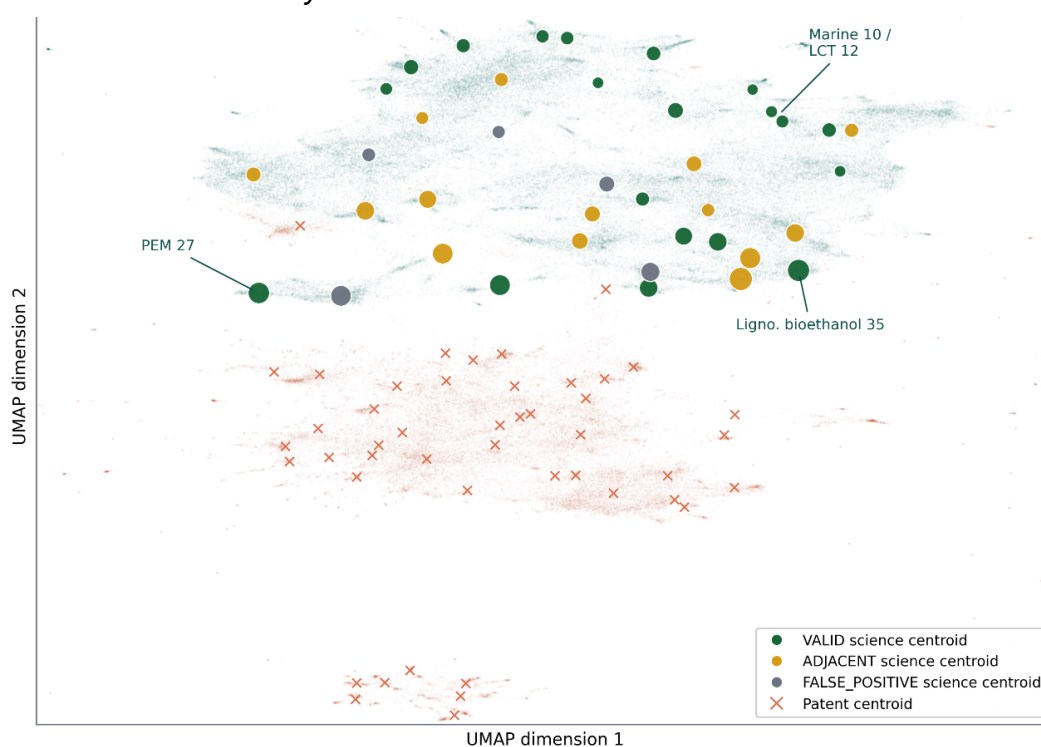
structural white spaces are likely to appear in meaningful numbers, rather than as isolated exceptions.

Compot scores were assigned to each paper through the two-stage procedure described in Section 3.7. Across the 58,628-paper corpus, the mean compot score is 0.530 and roughly 30% of papers (29.7%) sit above the 0.7 cutoff that enters the cluster-level ranking metric. The ranking therefore answers not whether commercially promising science exists in this field, but where it is concentrated. A cluster with `pct_above_07` well above 30% contains substantially more high-commercial-potential papers than the field as a whole.

Figure 4.1 presents the cluster-centroid topology of both corpora on a joint two-dimensional UMAP projection. As established in Section 3.5.1, this projection is a visualisation device only: all distance measurements and gate classifications reported below were carried out in the original 768-dimensional SPECTER2 space, and apparent proximities in the 2D layout should not be read as reliable indicators of true semantic distance.

Figure 4.1

Cluster-centroid topology in shared two-dimensional UMAP space. The plot overlays science-cluster centroids and patent centroids on the shared two-dimensional UMAP projection; colours and symbols distinguish VALID, ADJACENT, FALSE_POSITIVE, and patent centroids. Axes are UMAP dimensions used for visualisation only and should not be read as analytical distance measures.



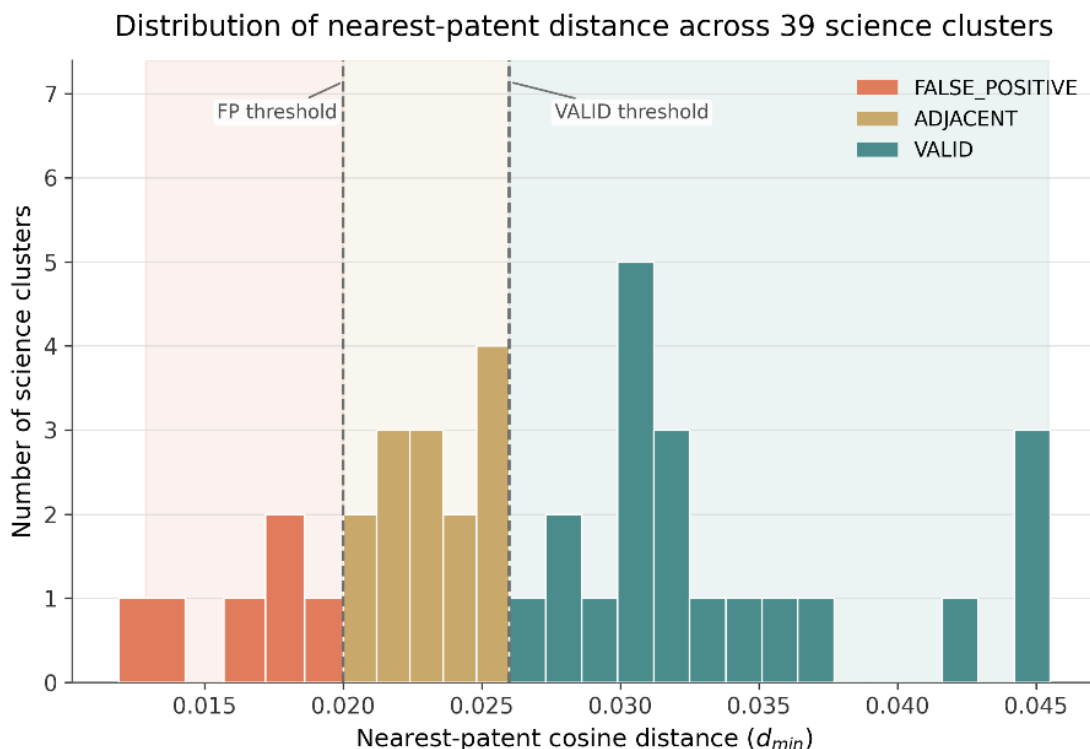
Axes clipped to the central 98% of the combined projection to prevent sparse patent outliers from compressing the dense semantic landscape.

4.2 Distance Gate Results and Classification

For each of the 39 science clusters, we computed the cosine distance to its nearest patent cluster ($d_{\min}$). The distribution ranges from 0.0129 to 0.0455 with a median of 0.0261, forming a graded continuum across the corpus rather than a clean separation into covered and uncovered science. The trinary gate (Section 3.6.2) maps this continuum to three structural categories via calibrated cut-points. Figure 4.2 presents the $d_{\min}$ distribution with the threshold lines (0.020 and 0.026) overlaid.

Figure 4.2

Nearest-patent distance distribution and distance-gate thresholds. The horizontal scale reports $d_{\min}$, the cosine distance from each science cluster to its nearest patent cluster; vertical threshold lines at 0.020 and 0.026 separate FALSE_POSITIVE, ADJACENT, and VALID classifications.



Applying the calibrated gate produces five FALSE_POSITIVE clusters ($d_{\min} < 0.020$), fourteen ADJACENT clusters ($0.020 \leq d_{\min} < 0.026$), and twenty VALID clusters ($d_{\min} \geq 0.026$). The 20 VALID clusters pass forward to the ranking step in Section 4.3; the full $d_{\min}$ values and category assignments for all 39 clusters are reported in Appendix A.

A representative cluster from each category illustrates what the gate’s structural categories contain in this corpus. Biodiesel Production Technology ($d_{\min} = 0.0129$), the prototype FALSE_POSITIVE, covers the core chemistry of biodiesel synthesis, including

transesterification, methyl-ester yields, and catalyst selection, and its nearest patent neighbour is a patent cluster densely covering exactly that chemistry. The science and the patents are doing the same work; the gate correctly identifies the science as already covered.

Heterogeneous Catalytic Biodiesel Production ($d_{\min} = 0.0200$) sits at the lower threshold and illustrates the ADJACENT category. The cluster covers solid-acid and solid-base catalysts for biodiesel production, a methodological refinement of the same broader pathway as the FALSE_POSITIVE example above, but at a specificity level the nearest patent cluster does not exhaust. Part of the science is patented (the general transesterification framework); part is not (the specific heterogeneous-catalyst chemistries). The cluster genuinely sits in the borderline territory the ADJACENT category was designed to capture.

Lignocellulosic Bioethanol Fermentation ($d_{\min} = 0.0276$) crosses into VALID territory but only just, at 0.0016 above the threshold. The cluster covers the fermentation of xylose and other pentose sugars from lignocellulosic biomass into ethanol, where the nearest patent cluster covers broader microalgae and biofuel production but not the specific xylose-fermentation chemistry. This is a structurally confirmed white space with a narrow margin, and as Section 4.3 shows, it tops the commercial-potential ranking. The case previews a pattern the chapter develops in the next section: a cluster sitting near the structural margin of VALID can carry the highest commercial-potential concentration in the validated set, while the most structurally distant cluster (Marine Alternative Fuel Technologies, $d_{\min} = 0.046$) sits in a different position entirely. Structural distance and commercial concentration are two separate dimensions, and a single ranking cannot capture both.

The three-way classification separates the 39 clusters into substantively distinct categories at each end of the distance range and an explicit borderline category in the middle, rather than collapsing into a single covered-or-not split. The categories correspond to three qualitatively different science-patent relationships identified through the calibration described in Section 3.6.2: clusters whose science overlaps with existing patents, clusters where the overlap is partial, and clusters where no substantive overlap exists. The gate operationalises these as distance bands across the full corpus.

4.3 White-Space Ranking and the 2×2 Opportunity Matrix

The distance gate identified 20 VALID white spaces. This section reports what the pipeline produces when those clusters are subjected to the two-layer prioritisation framework developed in Section 3.7: first the commercial-potential ranking, then the 2×2 opportunity matrix that adds structural positioning. The two layers are designed to

capture different dimensions of strategic interest, and a central finding of this section is that they do so in a way the ranking alone could not reveal.

4.3.1 Ranked List

The ranking metric is `pct_above_07`: the share of papers in a cluster with `compot` above 0.7 (Section 3.7.3). Table 4.1 reports the top 10 VALID clusters by `pct_above_07`; the full ranking of all 20 appears in Appendix A.

Table 4.1

Top 10 VALID white spaces by commercial-potential concentration

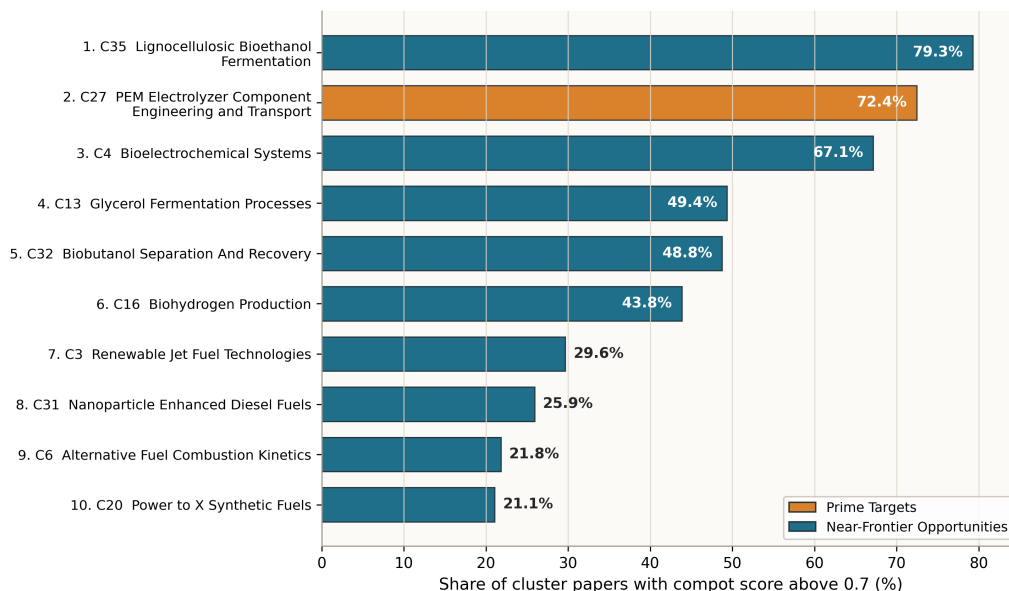
Rank	Cluster	<code>pct</code> > 0.7	$d_{\min}$
1	Lignocellulosic Bioethanol Fermentation	79.3%	0.028
2	PEM Electrolyzer Component Engineering and Transport	72.4%	0.036
3	Bioelectrochemical Systems	67.1%	0.031
4	Glycerol Fermentation Processes	49.4%	0.028
5	Biobutanol Separation and Recovery	48.8%	0.031
6	Biohydrogen Production	43.8%	0.031
7	Renewable Jet Fuel Technologies	29.6%	0.032
8	Nanoparticle Enhanced Diesel Fuels	25.9%	0.026
9	Alternative Fuel Combustion Kinetics	21.8%	0.029
10	Power to X Synthetic Fuels	21.1%	0.031

Note. `pct_above_07` is the share of papers in the cluster with `compot` score above 0.7. Distance is $d_{\min}$, the nearest-patent cosine distance used by the detection gate.

The top of the ranking is dominated by biological production processes, including bioethanol fermentation, bioelectrochemistry, glycerol fermentation, biobutanol recovery, and biohydrogen, together with PEM electrolyser engineering. In the first cluster of the ranking, nearly four in five papers carry a `compot` score above 0.7, against a corpus-wide rate of roughly 30%. The middle of the top 10 transitions into application-side clusters where commercial-potential concentration sits closer to but still above that baseline. Figure 4.3 presents the top 10 as a bar chart coloured by quadrant.

Figure 4.3

Top 10 VALID white spaces by commercial-potential concentration. Bars report pct_above_07 for the highest-ranked VALID clusters; colours indicate whether each cluster falls in the Prime Target or Near-Frontier Opportunity quadrant.



Rows are ordered by pct_above_07. Colours indicate the static 2x2 opportunity-matrix quadrant, showing that most top-ranked clusters are Near-Frontier Opportunities while PEM is the sole Prime Target in the top 10.

The ranking on its own already shows that commercial-potential concentration is unevenly distributed across the validated set; the gap between rank 1 (79.3%) and rank 10 (21.1%) is nearly fourfold, against a corpus baseline of 30%. But the ranking does not yet tell us where these clusters sit relative to the patent frontier, and as the next subsection shows, that second dimension reorganises which clusters look most strategically salient.

4.3.2 The 2×2 Opportunity Matrix

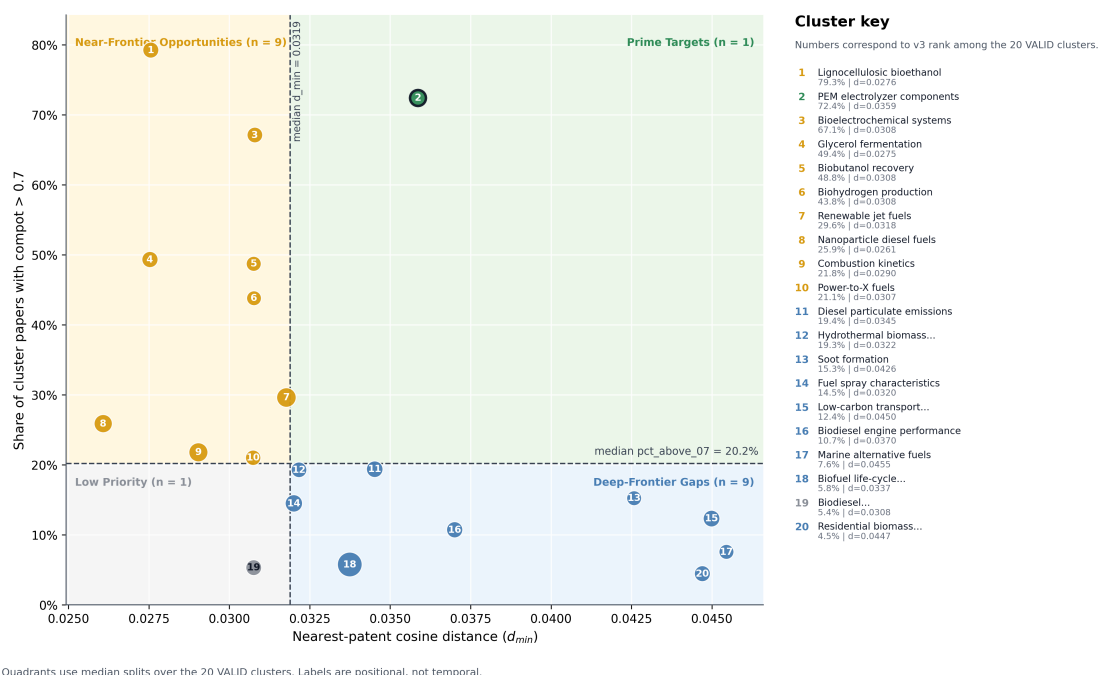
To recover the structural dimension the ranking sets aside, we plot the 20 VALID clusters on a 2×2 matrix using the two variables the pipeline has already produced: pct_above_07 from the ranking and $d_{\min}$ from the gate (Section 3.7.4). Each axis is split at the median of the 20 VALID clusters (pct_above_07 median = 0.2023; $d_{\min}$ median = 0.0319), partitioning the validated set into four quadrants: Prime Targets, Near-Frontier Opportunities, Deep-Frontier Gaps, and Low Priority.

Applied to the 20 VALID clusters, the matrix produces a notably asymmetric distribution: 1 Prime Target, 9 Near-Frontier Opportunities, 9 Deep-Frontier Gaps, and 1 Low Priority. Almost all validated white spaces sit on the off-diagonal axis: either commercially concentrated and close to the patent frontier, or commercially weaker and far

from it. Only two clusters occupy the diagonal corners where both dimensions point in the same direction. Across the 20 clusters, commercial-potential concentration and nearest-patent distance are negatively correlated (Spearman $\rho = -.58$), confirming the asymmetric pattern numerically. Figure 4.4 presents the matrix as a scatter plot with the median split lines shown explicitly.

Figure 4.4

Static opportunity matrix for the 20 VALID white-space clusters. Points are positioned by nearest-patent distance and pct_above_07; dashed lines show the VALID-set medians used to define the four quadrants.



4.3.3 Ranking and Matrix Tell Different Stories

The matrix's asymmetric distribution has a direct consequence for the ranking: the highest-ranked cluster is not the Prime Target. Lignocellulosic Bioethanol Fermentation leads the ranking at 79.3% but sits at $d_{\min} = 0.028$, below the distance median, placing it in Near-Frontier Opportunities. PEM, ranked second at 72.4%, sits at $d_{\min} = 0.036$, above the median, and is the only Prime Target. This is not a contradiction; it is the reason both layers of the framework are needed. The ranking captures commercial-potential concentration; the matrix captures structural openness. Collapsing them into a single rank order would erase the distinction. The fact that the framework's two layers reorganise the same validated set in different ways is direct evidence that the layers are doing analytically independent work, exactly as Section 3.7 was designed to ensure.

4.3.4 What the Four Quadrants Contain

The asymmetric distribution also tells us something substantive about the validated set: the 20 VALID clusters are not homogeneous. Near-Frontier Opportunity clusters are commercially concentrated but retain less structural margin. They remain meaningfully beyond the gate, but closer to existing patent activity than the rest of the VALID set. Deep-Frontier Gap clusters sit far from the frontier but their compot signal is comparatively low, possibly because their commercialisation travels through the non-patent channels discussed in Section 3.7.5, a pattern Chapter 5 develops further. The Prime Target quadrant, where high distance and high compot align, is the rare configuration in which both conditions hold simultaneously. PEM is the only cluster occupying it, and Section 4.4 examines what makes that combination possible.

4.4 Deep-Dive: PEM Electrolyzer Component Engineering and Transport

The previous section identified PEM Electrolyzer Component Engineering and Transport as the sole occupant of the Prime Target quadrant. This section reports the empirical evidence underlying that placement: what the cluster contains, why its structural distance from the patent frontier is genuine rather than artefactual, and what makes its combination of commercial concentration and structural distance unique among the 20 VALID clusters. Strategic interpretation is reserved for Chapter 5.

PEM is the only cluster satisfying both quadrant criteria simultaneously: `pct_above_07` of 72.4% (second in the ranking) and $d_{\min}$ of 0.036 (well above the distance median of 0.032). The cluster comprises 410 papers with a median publication year of 2024. PEM electrolyzers use renewable electricity to split water into hydrogen and oxygen; the hydrogen feeds many of the alternative-fuel pathways in the corpus, including green hydrogen, ammonia, and Power-to-X synthetic fuels. The papers focus not on the hydrogen itself but on the device: specifically the components surrounding the catalyst, including porous transport layers, flow-field plate designs, membrane-electrode assemblies, and degradation diagnostics. These are the engineering layers that determine whether a PEM electrolyser operates efficiently and durably at scale.

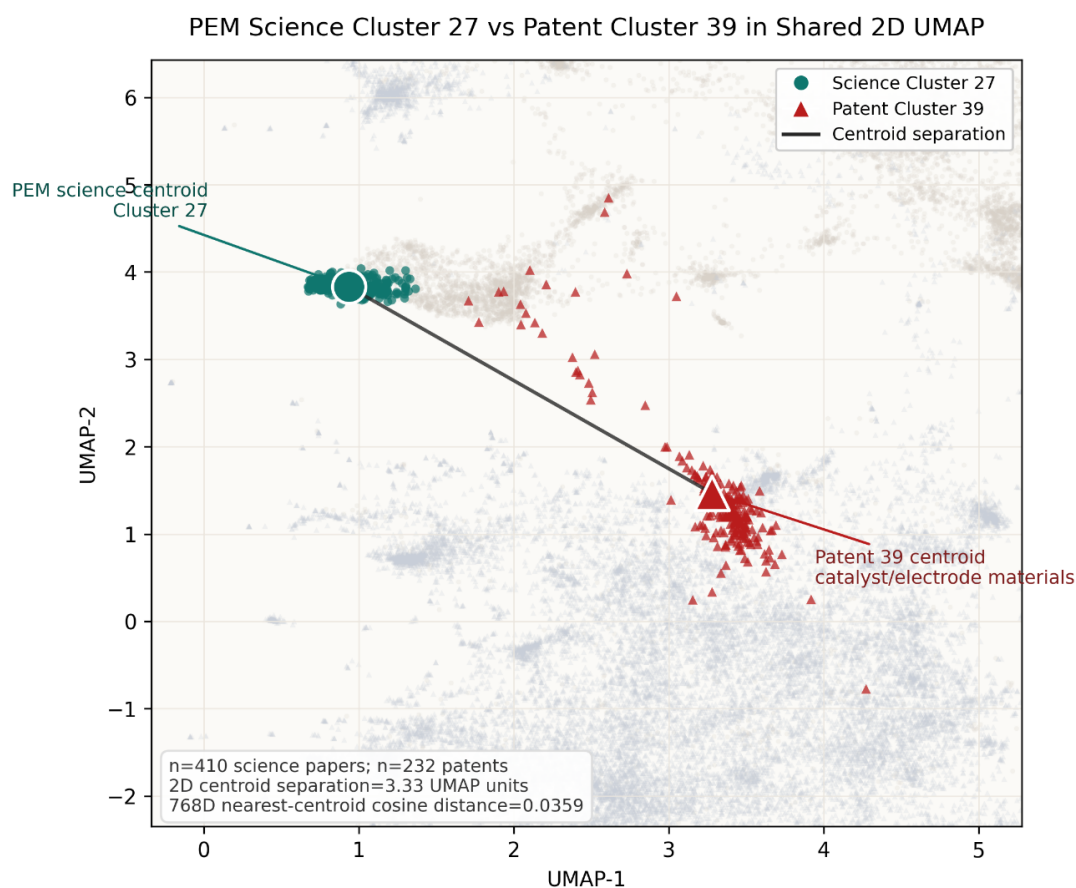
PEM is a device, not a fuel. A reader expecting the flagship Prime Target to be a fuel chemistry may notice that the pipeline instead identifies an upstream production technology. This is a substantive feature of the result: the field perimeter defined in Section 3.2 explicitly includes upstream production technologies constitutive of downstream fuel pathways, on the grounds that bottlenecks in production propagate downstream into fuel availability. The pipeline's identification of electrolyser engineering as

the only Prime Target reflects where commercial-potential concentration in this field is structurally located today.

The structural distance rests on a specific empirical finding. The nearest patent cluster (Patent Cluster 39, $d_{\min} = 0.036$) is dominated by catalyst-and-electrode-materials patents covering chemistries across multiple transition metals and electrochemical applications. PEM-specific examples are present but are a minority within a broader catalyst-materials cluster. Within these 232 patents, targeted scans for stack architecture, bipolar plates, flow fields, and porous-transport-layer terms return zero hits. Membrane chemistry appears only as application context for catalyst-use claims, not as the substance of what is being protected. The patents protect *what catalysts are made of*; they do not protect *the apparatus that holds them*. Since this is the geometrically closest patent territory in the entire corpus, the more general patent landscape addresses the engineering layer the science is developing even less. Figure 4.5 shows the separation directly.

Figure 4.5

PEM electrolyzer component engineering and nearest patent-cluster separation. The figure compares the PEM science cluster with its nearest patent cluster and highlights that the closest patent territory is centred on catalyst and electrode materials rather than PEM stack architecture, flow fields, bipolar plates, or porous transport layers.



Two further observations strengthen the Prime Target designation. First, the temporal trajectory of compot within the cluster moves from 58.3% in 2013–2017 to 81.7% in 2025–2026, a sustained signal whose intensity has accelerated, confirming that the commercial concentration is not driven by recent SciBERT-predicted scores alone.² Second, PEM’s nearest-patent distance shows the strongest negative trajectory in the validated set: patent-corpus snapshots taken at successive points between 2018 and 2025 show patent activity moving closer to the PEM science cluster more clearly than for the other VALID clusters. PEM papers were also cited by post-2018 patents at six times the rate of pre-2018 patents, the highest post-2018 citation momentum in the validated set, indicating that patents began drawing on the cluster’s science substantially

²Per-period pct.above.07 values for all five windows (2013–2017, 2018–2020, 2021–2022, 2023–2024, 2025–2026) are reported in Appendix D. The trajectory diagnostic methodology and full cluster-level results are in Appendix E.

more intensively after 2018. PEM is therefore the clearest case where the static matrix and the dynamic diagnostics converge.

Taken together, the evidence supports PEM's Prime Target classification on three grounds: it satisfies both quadrant criteria more clearly than any other cluster; the structural distance corresponds to a measurable absence of patent coverage on the engineering layer the science is actively developing; and both the static matrix and the temporal diagnostic identify the same case uniquely. Whether this translates into actionable strategic opportunity is the subject of Chapter 5.

4.5 Deep-Dive: Two Deep-Frontier Gap Clusters

Nine clusters occupy the Deep-Frontier Gap quadrant, meaning they sit far from the patent frontier but low on commercial-potential concentration. Marine Alternative Fuel Technologies and Low Carbon Transport Fuel Pathways illustrate this configuration most clearly and show that low compot does not invalidate a white-space classification; it reveals a different commercialisation pathway than the one PEM exemplifies.

Marine Alternative Fuel Technologies (cluster 10, $d_{\min} = 0.046$, pct.above.07 = 7.6%) is the most structurally distant cluster in the corpus. Its 224 papers focus on the practical problem of moving the maritime sector onto alternative fuels: how ammonia and methanol burn in marine engines, how hydrogen fuel-cell systems integrate with ship propulsion, how bunkering infrastructure should be configured for new fuels at scale, and how fleets can be retrofitted within the IMO's decarbonisation timeline (DNV, 2023; International Maritime Organization, 2023). The science is operationally oriented: it asks how the maritime sector adopts alternative fuels, not how those fuels are produced. The nearest patent territory (cluster 11, 146 patents) is a synthetic-fuel production cluster covering syngas, methane reforming, and Fischer-Tropsch synthesis. A targeted scan returns zero hits on bunkering, retrofit feasibility, ammonia engine specificity, or marine vessel terms beyond six incidental mentions. The science is about adoption on ships; the nearest patents are about production upstream.

The cluster's low compot reflects how commercialisation actually proceeds in this sector. Adoption of alternative fuels in shipping advances primarily through standards and regulatory compliance (IMO frameworks, MARPOL annexes, fuel specifications, classification society approval), through operational and infrastructure engineering (bunkering systems, retrofit feasibility, fuel-handling protocols), and through demonstration projects on pilot vessels. These pathways generate substantial commercial activity but relatively few novel patentable inventions. The science engages with the levers that actually move the sector, but those levers are structurally less patent-generating than

the catalyst chemistries and device-engineering pathways that dominate the upper half of the ranking.

Low Carbon Transport Fuel Pathways (cluster 12, $d_{\min} = 0.045$, $\text{pct_above_07} = 12.4\%$) follows the same structural logic but with different channels. The 574 papers address which fuel and powertrain combination each transport mode should adopt, comparing alternatives across road, rail, and heavy-duty transport using life-cycle assessment, well-to-wheel analysis, and fleet-level policy modelling (International Energy Agency, 2023c). The work is comparative and evaluative rather than synthetic: it does not produce new fuels or new technologies but assesses how existing options should be deployed. Here the non-patent channel is more direct than for Marine: comparative modelling, scenario analysis, and policy work do not generate patentable inventions at all, so the gap between the science and the nearest patent cluster (cluster 22, 1,098 fuel-formulation patents) is a clean separation between what the fuel recipe is and which fuel and powertrain mix a transport mode should adopt.

Both clusters score low on compot because their commercial value travels primarily through non-patent channels, the structural blind spot acknowledged in Section 3.7.5, but the specific channels differ. Marine’s commercialisation runs through standards, operational engineering, and demonstration; Low Carbon Transport’s runs through modelling, assessment, and policy. Read alongside PEM, they point to the broader pattern the matrix captures: production-side science concentrates at the high end of the ranking; deployment-side and policy-side science concentrates at the low end. Chapter 5 takes up the substantive interpretation.

4.6 Validation Results

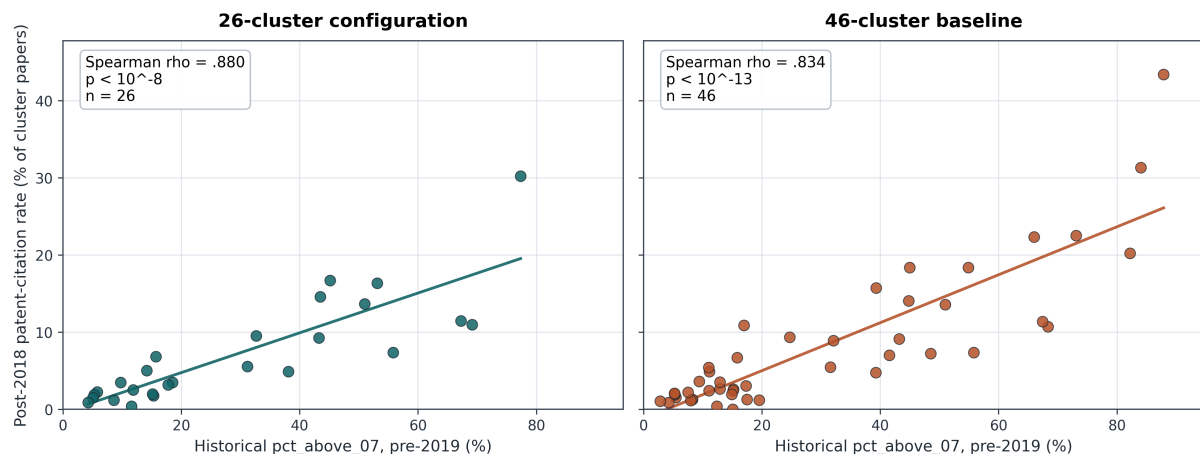
Sections 4.1 through 4.5 demonstrated that the pipeline produces interpretable output. This section turns to direct evidence: the results of the historical re-clustering test described in Section 3.8, which asks whether the rankings the pipeline would have produced at the end of 2018 predict the patent citation activity that actually followed.

Across the full historical ranking, clusters the 2018 pipeline judged more commercially promising went on to receive substantially more post-2018 patent citations. In the 26-cluster configuration, a Mann-Whitney U test on the upper-half-versus-lower-half split returned $p < 10^{-4}$, and the pattern held when the outcome was restricted to new citations, meaning papers first cited by post-2018 patents only, ruling out the trivial case in which already-cited papers continue to attract citations ($p < 10^{-3}$). The graded version of the same relationship, captured by Spearman rank correlation across the full ranking, returned $\rho = .880$ ($p < 10^{-8}$) in the 26-cluster configuration and $\rho = .834$ ($p < 10^{-13}$) in the 46-cluster configuration. Figure 4.6.1 presents the cluster-level scatter for

both configurations. A ranking constructed entirely from pre-2019 information therefore explains a substantial share of the variance in post-2018 patent activity.

Figure 4.6.1

Historical commercial-potential ranking and later patent-citation attention. Each point is a historical cluster; the x-axis is pre-2019 pct_above_07 and the y-axis is the share of papers cited by post-2018 patents.



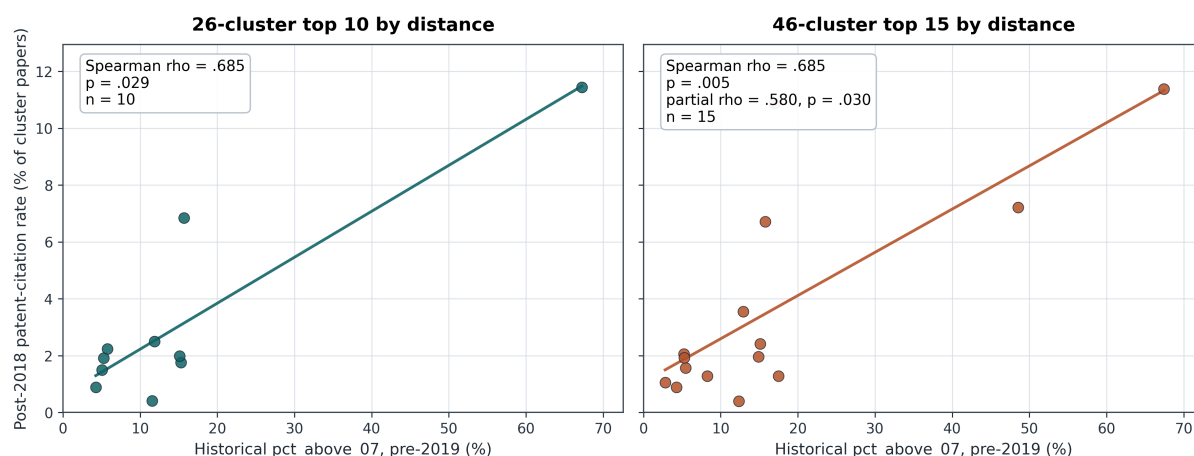
This is a strong starting point but not yet a test of the thesis claim, which concerns specifically the subset of clusters that lay beyond the patent frontier in 2018. Restricted to that historical white-space subset, the predictive relationship holds. In the 46-cluster configuration, the top 15 clusters by nearest-patent distance produced Spearman $\rho = .685$ ($p = .005$). The 26-cluster configuration produced an identical correlation strength on its top 10 clusters by distance ($\rho = .685$, $p = .029$); the smaller subset reduces the precision of p-value estimation but the effect size is unchanged. Among the clusters that genuinely correspond to the thesis's white-space definition, those the pipeline ranked higher by pct_above_07 attracted more post-2018 patent citations.

The remaining concern is whether the predictive relationship in the restricted subset is really tracking commercial-potential concentration or simply tracking structural distance, which is itself correlated with citation activity. The partial Spearman addresses this directly. In the 46-cluster restricted subset, the correlation between pre-2019 pct_above_07 and post-2018 citations, after removing the part of the citation signal explainable by nearest-patent distance alone, is $\rho = .580$ ($p = .030$). The effect size drops from the unadjusted $\rho = .685$, indicating that some of the predictive content does overlap with distance, but the majority survives. Even among clusters held at the same structural distance from the patent frontier, those with higher pre-2019 pct_above_07 received more post-2018 patent citations. The ranking is not a dressed-up distance

measure; it carries predictive content that distance alone does not explain. Figure 4.6.2 reports this restricted validation.

Figure 4.6.2

Validation inside historical distance-filtered white-space candidates. The right panel also reports the partial correlation controlling for nearest-patent distance.



The validation therefore delivers what the design in Section 3.8 set out to test. The pipeline's rankings, constructed entirely from pre-2019 information, predict patent activity that occurred in the years that followed, and the prediction holds in the subset of clusters that most plausibly constituted genuine white spaces, and is not reducible to either prior commercial activity or to the structural distance the gate already used to define the white-space set.

Taken together, the findings of this chapter establish that the pipeline produces output of the kind it was designed to produce. The distance gate separates the corpus into substantively distinct structural categories; the ranking and the opportunity matrix surface different dimensions of strategic interest in a way the ranking alone could not reveal; the cluster deep-dives show that the framework's classifications correspond to verifiable patterns in the patent landscape; and the historical re-clustering test shows that the rankings carry predictive content beyond their internal interpretability. Chapter 5 takes up the substantive implications and boundary conditions.

5 Discussion

This chapter interprets the findings reported in Chapter 4. It asks what the pipeline teaches us about the conditions under which commercially promising science becomes visible to patents, and where the framework’s reach ends. The argument proceeds in four steps. We first show that structural openness and commercial-potential concentration behave as distinct dimensions of the science–technology nexus (Section 5.1). We then explain what the historical validation supports, and what it does not (Section 5.2). We turn to the limits the empirical findings reveal directly: the conservative blind spot introduced by HDBSCAN noise and the static-snapshot character of the framework’s outputs (Section 5.3). We close with the methodological and practical implications of treating white-space detection and ranking as separable questions (Section 5.4).

5.1 What the Pipeline Shows: Distance and Commercial Concentration Are Loosely Coupled

The central empirical pattern is not the list of white spaces. It is the geometry of how the validated white spaces distribute across the two strategic dimensions used to characterise them. When the clusters are positioned by median splits on both dimensions, only one falls in the high-distance / high-compot quadrant (Prime Target); one falls in the below-median-distance / below-median-compot quadrant (Low Priority); and the remaining 18 split evenly between Near-Frontier Opportunities (below-median distance, high compot) and Deep-Frontier Gaps (high distance, below-median compot). This 1/9/9/1 distribution is the finding.

Within the VALID set, the two dimensions are moderately anti-correlated (Spearman $\rho = -0.58$): clusters higher on commercial-potential concentration tend to sit closer to the patent frontier, and clusters further from the frontier tend to score lower on compot. The 1/9/9/1 count and this correlation together indicate that, within the VALID set, the clusters furthest from the patent frontier are rarely the same clusters that concentrate the highest commercial potential. The two dimensions are loosely coupled rather than coincident. This loose coupling is the empirical form of the conceptual distinction drawn in Section 2.3: structural absence and commercial relevance were argued there to be related but non-identical, and the 1/9/9/1 distribution is the observable signature of that non-identity.

Several mechanisms could plausibly contribute to this pattern, though discriminating between them is outside the scope of this thesis. One candidate is grounded in firm

behaviour: patent-active firms tend to draw on science whose outputs they can already see how to use (Fleming, 2001; Fleming & Sorenson, 2004), and science whose outputs are immediately usable in this sense tends to sit close to the existing patent frontier rather than far from it. The further a region of science sits from the frontier, the less likely it is that firms have already worked out how to translate it into patentable inventions, and the lower its commercial-potential concentration on the patent channel would tend to be. A second candidate is methodological rather than behavioural: compot is trained to predict whether a paper will be cited by patents, and papers that historically attract patent citations tend to sit close to existing patent activity by construction. The scoring model could therefore learn to assign higher compot to science that textually resembles work already absorbed into the patent corpus, structurally biasing high-distance regions toward lower scores. Both mechanisms are consistent with the observed pattern; neither is tested here, and disentangling them would require longitudinal cluster-tracking beyond the scope of this thesis.

PEM Electrolyzer Component Engineering and Transport is exceptional precisely because it does not follow this pattern. It combines a nearest-patent distance of 0.036 (above the VALID-set median of 0.032) with a `pct.above_07` of 72.4%, more than three times the median. Although its distance margin above the median is modest, it is the only cluster in the VALID set that occupies above-median values on both dimensions simultaneously. Its strategic interest follows from this combination. PEM does not rank as the single Prime Target because it is the most distant cluster, nor because it has the highest compot concentration in absolute terms; it ranks there because the alternative-fuels frontier rarely produces clusters that combine distance from the patent frontier with concentrated commercial potential.

Two implications follow. First, the off-diagonal pattern is consistent both with established accounts of how firms select science for patenting and with how a patent-citation-trained scoring model would be expected to assign commercial potential; the 1/9/9/1 distribution observed here is therefore not easily dismissed as a quirk of this corpus, though the present analysis cannot establish which mechanism dominates. Second, the Prime Target quadrant matters because of how rarely it can be occupied: what the pipeline contributes is the ability to identify the joint condition PEM satisfies, which neither dimension alone could reveal.

5.2 What the Validation Supports: The Pipeline Produces a Predictive Ranking

The historical re-clustering test reported in Section 4.6 asked whether the pipeline, run on data available before 2019, would have produced a cluster-level ordering that

corresponded to where patent-citation attention subsequently concentrated. It returned a clear result. The pre-2019 ranking predicted post-2018 patent activity across the full set of historical clusters; the signal survived a coarser HDBSCAN configuration; and within the restricted subset of clusters that genuinely corresponded to historical white spaces, the predictive relationship held even after the contribution of nearest-patent distance was statistically removed.

This is a claim about the pipeline as a construction, not about its individual inputs. The predictive power of the compot score is not what this validation establishes: that result was already demonstrated extensively by Masclans et al. (2025), across multi-million-paper holdout tests, university technology-transfer outcomes, and stock-market-based patent-value measures. What this thesis tests is whether aggregating compot at the cluster level, positioning clusters relative to the patent frontier through SPECTER2-embedded distance, and ranking the resulting white-space candidates produces an ordering with external predictive content. The historical test shows that it does. A version of the pipeline run at the end of 2018, given only information available at that time, would have ranked the structurally open clusters of the alternative-fuels corpus in a way that anticipated which of them subsequently drew more patent-citation attention.

The partial Spearman result is the part of this finding that most directly supports the two-dimension framework developed in this thesis. If `pct_above_07` were simply re-expressing distance, meaning that the ranking was a dressed-up distance measure with a different label, then holding distance statistically constant should collapse the correlation. The partial Spearman remained meaningfully positive. Both dimensions therefore carry predictive weight that the other does not capture, which is the empirical counterpart to the conceptual argument made in Section 5.1: distance and commercial-potential concentration are loosely coupled, capture different parts of the science–technology translation process, and need to be read together. The opportunity matrix is not a presentational device imposed on a one-dimensional ordering; it reflects a structure the validation has now corroborated.

Why this ordering holds up against later patent-citation evidence is a separate question, and one the validation cannot directly answer. The result is consistent with the absorptive-capacity logic of Cohen and Levinthal (1990): science with higher commercial-potential concentration may be more recognisable, more codified, or more directly usable by patent-active actors, raising the probability that it enters the patent literature as cited prior art. We do not observe firm search behaviour directly, so this remains an interpretation rather than a tested mechanism. The validation establishes that the ranking has predictive content; absorptive capacity offers a plausible account of why it might.

Two boundary conditions constrain the claim. The Reliance on Science citation match reaches 6.1% of the thesis corpus, a rate that reflects cross-database identifier attrition rather than rarity of science-to-patent transfer, but that nonetheless narrows the test population. Patent citations also capture only one channel through which science enters technology; commercialisation routed through publication, collaboration, or non-patented use sits outside what this validation can observe. The supported claim is therefore specific: within historical white-space candidates, the pipeline's ranking identified where subsequent patent-citation attention strengthened. It does not amount to a general patent-forecasting model, and it does not characterise the share of science-to-technology transfer that bypasses the patent system.

5.3 What the Empirical Findings Cannot Resolve: The Noise Pool and the Static Snapshot

The methodological boundaries identified in Section 3.9 and the thesis-level limitations developed in Section 6.3 concern what the pipeline's design constrains and what the framework as a whole does not establish. This section addresses something different: two specific patterns in the empirical findings of Chapter 4 that have interpretive consequences for how the pipeline's outputs should be read, and that the static design of the present study cannot fully resolve.

The first concerns the noise pool. As stated in Section 3.5.2, HDBSCAN does not force every paper into a cluster; it leaves weakly-consolidated papers in a noise pool rather than fabricating artificial centroids around them. In the alternative-fuels corpus, this design choice assigned 20,608 papers, 35.2% of the science corpus, to noise. Of those, 6,083 carry a commercial-potential score above 0.7. The final ranking is therefore a ranking of coherent clustered white spaces, not a complete inventory of every commercially promising paper. The pipeline prioritises the reliability of cluster structure over exhaustive recall, and a substantial share of the corpus sits outside the structures it is willing to evaluate.

Whether the exclusion of these papers from the ranking represents an important loss depends on what they contain. The high-compot noise papers concentrate on the same terminological territory as the clustered material, including production, biodiesel, hydrogen, biomass, biofuels, fermentation, and combustion, which means their exclusion from the ranking is a question of clustering geometry rather than topical irrelevance. The year distribution sharpens this. High-compot noise papers grew from 397 in 2021 to 929 in 2025, more than doubling over four years, while the total noise pool itself grew at a much slower rate. The growth concentrates specifically in the high-compot fraction. This raises the possibility that some part of the noise pool represents

emerging or weakly-consolidated commercial pockets that have not yet acquired the density required to form their own clusters. HDBSCAN's refusal to impose structure where structure is weak protects the pipeline against false centroids, but it also means that emerging material may remain invisible to the ranking until it stabilises in its own right. Whether the recent high-compot growth reflects emerging topics on a path to consolidation or transient clutter is a question the present cross-sectional design cannot answer.

The second pattern is the static-snapshot character of the empirical results themselves. Chapter 4 reports a matrix configuration, an exceptional cluster in PEM, a 1/9/9/1 distribution across the quadrants, and the noise-pool growth just described. Each of these findings raises a temporal question. Do Near-Frontier Opportunities close over time as patent-active firms catch up to the science, or do they remain open? Do Deep-Frontier Gaps consolidate into testable opportunities as scientific density accumulates, or do they persist as structurally distant regions? Does the recent expansion of the high-compot noise pool produce observable cluster-level trajectories, with new white spaces emerging into the ranking as topics gain density, or does it stay diffuse? The pipeline as currently constructed cannot answer any of these. It produces a single snapshot, and its outputs are positions in a matrix at one point in time rather than trajectories through it.

Both limitations point in the same direction. The noise pool's composition and growth, and the temporal questions the empirical findings invite, are limits the present design accepts in exchange for cluster reliability and tractable validation. They are not flaws to be hidden, but specific aspects of what the pipeline does not yet do and what extensions of it might. A longitudinal version of the pipeline, running on rolling temporal windows and tracking how clusters move between matrix quadrants, would address both: the noise pool could be examined for emerging-versus-residual structure across time, and the static snapshot could be replaced by trajectory data capable of testing the temporal questions Chapter 4 raises. This is taken up in Section 6.4.

5.4 What Kind of Instrument the Pipeline Is

Taken together, the findings of the previous three sections characterise the framework not as a forecasting model, nor as a substitute for domain expertise, but as a specific kind of upstream research instrument. Section 5.1 showed that the framework distinguishes strategically heterogeneous clusters within a structurally homogeneous set, with distance and commercial-potential concentration loosely coupled rather than coincident. Section 5.2 showed that the resulting ranking carries external predictive content

beyond what either dimension provides on its own. Section 5.3 showed that the framework is not exhaustive: a substantial share of the corpus sits in a noise pool that the pipeline declines to evaluate, and the pipeline's single-snapshot design leaves temporal questions outside its reach. Each of these findings constrains what the framework can claim, and together they specify what it is.

The pipeline operates as a structured prioritisation device. Given a corpus of scientific publications and a corresponding corpus of patents, it produces a ranked, calibrated map of where strategically interesting white spaces sit: which clusters are structurally absent from the patent frontier, which of those carry commercially concentrated science, and how the two dimensions distribute across the set. The map is not a decision; it is a starting point for one. The pipeline narrows a fifty-eight-thousand-paper corpus into a small set of clusters worth qualitative attention, and the validation establishes that this narrowing is not arbitrary, since clusters the framework places high are clusters that historically did attract subsequent patent-citation attention. But the framework stops there. It identifies where to look first; it does not tell anyone what they will find when they look.

This bounded character is what makes the framework useful across audience contexts. The pipeline produces the same map for everyone; what differs is the downstream work done with it. A firm scouting for partnership candidates and freedom-to-operate analysis will read Prime Targets and Near-Frontier Opportunities differently than a research-strategy team deciding where to invest exploratory capacity, and both will read them differently than a policy actor examining whether structural absence in a region of science reflects institutional rather than scientific gaps. None of these uses follows automatically from the map; each requires the kind of domain expertise the pipeline does not encode and was not designed to encode. What the framework changes is the starting point from which that expertise operates, not the need for it.

The pipeline's current form is therefore one realisation of the framework rather than the limit of what the framework could become. Extensions that the present design cannot accommodate, including longitudinal tracking of cluster trajectories, integration of non-patent commercialisation signals, and hybrid configurations in which expert judgement enters as a calibration input rather than as a downstream consumer of outputs, would broaden the range of strategic questions the framework can address without altering its central logic. These are taken up as future-work directions in Section 6.4. What this chapter has established is narrower than a contribution claim and prior to it: that the framework's outputs behave as the conceptual and methodological argument predicted, and that the chapter has been explicit about where those outputs should not be over-read. Translating this into a stated contribution, and setting it against the limitations that bound it, is the work of the final chapter.

6 Conclusion

The preceding chapter established what kind of instrument the pipeline is: a structured prioritisation device that narrows a large corpus into a strategically differentiated set of white spaces, produces a ranking with externally validated predictive content, and identifies the boundaries beyond which its outputs should be supplemented rather than trusted alone. This chapter draws those threads to a close. Section 6.1 restates how the research question was answered. Section 6.2 formalises the contributions that follow from those findings. Section 6.3 acknowledges the limitations that bound the thesis's claims. Section 6.4 outlines directions for future work that respond to what Chapters 4 and 5 left open. Section 6.5 closes with a statement of broader significance.

6.1 Research Question and Answer

This thesis asked whether the structural detection and commercial ranking of white spaces could be integrated into a single validated framework. The answer, within explicit boundary conditions, is yes.

The pipeline joins detection and ranking into a single sequence and reads their outputs together in a strategic matrix. A cosine-distance gate, calibrated through LLM-assisted comparison of science–patent cluster pairs, qualifies clusters as structurally absent from the patent frontier. A cluster-level concentration measure then ranks the qualifying clusters by the share of papers carrying high commercial-potential scores. The matrix positions the validated white spaces along both dimensions, allowing strategic differentiation within a set that the detection step alone would have treated as homogeneous.

Applied to the alternative-fuels sector, the framework yielded the outcomes the research question was designed to test. The ranking carried predictive content corroborated against an outcome external to the pipeline's own semantic geometry. The matrix revealed a pattern of strategic heterogeneity within structural homogeneity, a pattern neither dimension alone would have made visible, and one the two-dimension construction was specifically designed to surface.

The remainder of this chapter states what these results contribute to the literature, where the framework's claims are bounded, and what work the results invite.

6.2 Contributions

The thesis makes three contributions to the literature on science–technology mapping.

The first is conceptual. Chapter 2 identified a gap between two bodies of work: Shibata et al. (2010) established that white spaces can be detected through topological comparison of science and patent clusters, but treated all detected gaps as structurally equivalent. Masclans et al. (2025) introduced a continuous measure of commercial potential at the paper level, but did not embed it in a topological structure that could distinguish covered science from uncovered science. The thesis bridges this gap by proposing that white spaces are defined by a single property, structural absence from the patent frontier, but characterised along a second dimension the literature had not previously operationalised: the concentration of commercially promising papers within each structurally absent cluster. The contribution lies in the framing rather than the empirical pattern: by defining white spaces along two operationalised dimensions instead of one, the framework gives the literature a vocabulary for distinguishing strategically heterogeneous opportunities that prior work, stopping at detection, treated as equivalent.

The second contribution is methodological. The pipeline operationalises the conceptual separation through a gate-then-rank architecture in which the two analytical questions cannot compensate for one another: a cluster’s commercial concentration cannot rescue it from failing the structural gate, and structural distance cannot substitute for commercial concentration in the ranking. Three design choices distinguish the pipeline from its antecedents. Citation networks are replaced by SPECTER2 semantic embeddings, overcoming the lag, selection bias, and vocabulary mismatch that constrain citation-based detection (Section 2.2). The commercial-potential score is lifted from the paper level to the cluster level through a concentration metric (`pct.above_07`), providing the strategic dimension along which validated white spaces are compared. And the validation strategy relies on patent citations from the Reliance on Science dataset (Marx & Fuegi, 2020), an outcome external to the embedding pipeline, which avoids the problem, documented in Appendix G, of validating a distance-based framework with outcomes derived from the same geometry.

The third contribution is the empirical demonstration that the framework operates as the conceptual and methodological claims predict. The alternative-fuels application is not the thesis’s contribution in itself; it is the evidence that the framework produces the kind of output it was designed to produce. The findings that matter for this purpose are reported in Chapter 4 and interpreted in Chapter 5: the two dimensions diverge within the validated set, the ranking carries predictive content that survives restriction to the white-space subset and statistical control for distance, and the framework surfaces

strategic heterogeneity that neither dimension alone could reveal. These are properties of the framework, demonstrated through a particular corpus. Whether they generalise beyond alternative fuels is an empirical question the present thesis cannot answer, and it is the first limitation to which the next section turns.

6.3 Limitations

The claims made in this thesis are bounded by four constraints, each of which has been developed at its natural location in the preceding chapters. This section does not re-argue them; it states what they mean, taken together, for the strength of the thesis's overall contribution.

The first constraint is internal to the commercial-potential layer. Approximately 43% of papers in the corpus carry SciBERT-predicted compot scores rather than direct Masclans et al. (2025) scores, and the share of predicted scores varies across the 20 validated clusters from 28% to 66%. Diagnostic tests reported in Appendix I indicate that this variation is recency-driven rather than rank-distorting. The predicted-score share rises with publication year but is uncorrelated with cluster rank, and the rank structure is robust to the prediction layer. Full resolution would nonetheless require ground-truth scores for the currently predicted subset.

The second constraint is sectoral scope. The framework was operationalised and tested on a single sector. Alternative fuels satisfies the field-eligibility conditions set out in Section 3.1, but the thesis does not test whether the gate-then-rank logic, the distance thresholds, or the median-split matrix construction reproduce in sectors with different patenting densities, different degrees of semantic overlap between scientific and patent texts, or different balances between patent and non-patent commercialisation routes. The framework's structural generality is a hypothesis supported by design logic, not an empirical result.

The third constraint is the most consequential because it limits what the ranking can see rather than how precisely it measures what it does see. The commercial-potential score is anchored in patent-citation and patent-renewal outcomes (Masclans et al., 2025). It is therefore structurally less sensitive to commercial value that travels through non-patent channels: standards, regulation, consulting, systems integration, demonstration projects, and open technical infrastructure. Section 3.7.5 stated this as a property of the measurement instrument; Section 4.5 showed its substantive consequences through the Marine and Low Carbon Transport clusters, where low compot reflected a redirection of commercial value rather than its absence. The ranking should be read as identifying where patent-appropriable commercial potential concentrates, not where commercial relevance concentrates in general.

The fourth constraint follows from the pipeline's cross-sectional design. Section 5.3 showed that the noise pool contains over 6,000 high-compot papers whose number has been growing rapidly, and that the opportunity matrix captures positions at a single point in time rather than trajectories through it. Whether Near-Frontier Opportunities close, whether Deep-Frontier Gaps consolidate, and whether new white spaces emerge from the noise pool as topics gain density are questions the present snapshot cannot answer. The framework identifies where strategically interesting configurations sit today; it does not yet track how they move.

These four constraints converge on a single characterisation. The pipeline produces a reliable, externally validated prioritisation of patent-channel white spaces within a single sector at a single point in time. Each qualifier in that sentence marks a boundary the thesis has chosen not to cross. Extending beyond those boundaries is the subject of Section 6.4.

6.4 Future Work

The four boundaries identified in Section 6.3 define a connected set of research directions. This section organises them into three clusters: extending the framework's inputs, deepening its internal logic, and testing its practical uses.

The first cluster extends what the pipeline can see. Cross-domain replication is the natural starting point. The interesting question is not whether the absolute thresholds reproduce across sectors, since they are corpus-specific by construction, but whether the gate-then-rank separation continues to produce strategically informative matrices under different structural conditions. Would sectors with denser patenting regimes produce a more populated Prime Target quadrant, or would the off-diagonal asymmetry reappear? Would the negative correlation between distance and commercial concentration hold in fields where the patent frontier sits closer to the scientific frontier than it does in alternative fuels? Each answer would clarify whether the 1/9/9/1 pattern observed here is a general property of how firms select science for patenting or a feature of this particular landscape. A complementary extension addresses the patent-channel constraint directly. The Marine and Low Carbon Transport clusters already showed what a multi-channel commercial-potential measure would need to capture: standards adoption, regulatory uptake, demonstration-project activity, and systems-integration consulting. Constructing such a measure and running it alongside the existing compot signal would test whether the Deep-Frontier Gap quadrant shrinks, stays the same, or reorganises entirely when non-patent commercial value becomes visible to the ranking. A narrower but immediately actionable extension concerns the validation layer. The historical re-clustering test relied on a 6.1% citation-match rate between

the thesis corpus and the Reliance on Science dataset, a rate driven by cross-database identifier attrition rather than by the rarity of science-to-patent transfer. Improved identifier linkage across bibliographic databases, or access to a patent-citation dataset with broader DOI coverage, would raise the matched fraction substantially and allow the validation to operate at finer cluster-level resolution than the present test permits.

The second cluster deepens the framework's internal logic by treating its outputs as objects of study rather than as endpoints. The 2×2 matrix in its current form is a sorting mechanism: clusters are positioned by their coordinates, but the substantive question of what makes each quadrant the kind of thing it is remains unanswered. Asking what Near-Frontier Opportunities have in common beyond their matrix coordinates, whether in the type of science they contain or in the commercialisation channels they depend on, would turn the matrix from a sorting device into a typology generator. The same logic applies to Deep-Frontier Gaps, which the thesis has tentatively associated with non-patent commercialisation routes but has not characterised in their own right. A second direction within this cluster responds to the static-snapshot constraint. A longitudinal version of the pipeline, running on rolling temporal windows and tracking how clusters move between matrix quadrants, would convert the static map into a dynamic one. The noise-pool growth pattern identified in Section 5.3, where high-compot papers more than doubled between 2021 and 2025 while the total noise pool grew much more slowly, already suggests the first temporal question such an extension should ask: are the high-compot noise papers of 2025 the white-space clusters of 2028?

The third cluster moves from extending the framework to validating its claimed practical uses. Section 5.4 argued that Prime Targets, Near-Frontier Opportunities, and Deep-Frontier Gaps suggest distinct strategic implications for different audiences. Each of these implications is currently a hypothesis about what happens when the framework is used, and testing them requires moving from output description to outcome measurement. For firms, this means tracking whether clusters identified as Prime Targets actually result in patentable inventions, R&D partnership formation, or successful technology transfers when integrated into industrial scouting workflows. For investors, it means embedding the matrix into screening processes and measuring whether it improves selection outcomes relative to existing methods. A more methodologically ambitious version of this direction would bring expert judgement inside the pipeline as a calibration input rather than leaving it as a downstream consumer of outputs. Domain-expert ratings on a sample of clusters could tune the distance threshold, the commercial-potential aggregation, or the matrix median splits, producing a hybrid framework that combines computational scaling with the substantive judgement no fully automated method can replace. This is not a robustness check on the existing framework; it is a methodological contribution in its own right, and it responds directly to

the calibrated claim throughout this thesis that the pipeline narrows expert judgement without displacing it.

Together, these directions trace the natural arc of the framework's development. The first cluster extends what the pipeline can observe; the second deepens what it can interpret; the third tests what it can support. None of them is preliminary to the others; they would draw on different data, address different audiences, and answer different versions of the question this thesis has begun to ask. What they share is a single underlying claim: that the integration of detection, ranking, and validation demonstrated here is a starting point for a research program rather than a finished instrument.

6.5 Closing Remarks

The literature on white spaces has approached detection and ranking as separate operations: as a topological exercise that stops at gap identification, or as a paper-level scoring of commercial relevance untethered from structural position. This thesis brought the two operations into a single sequence and tested whether the integration would survive external validation. The integration held, within the boundary conditions the thesis has made explicit. The resulting prioritisation carries predictive content beyond what either dimension delivers alone, and the matrix construction reveals strategic differences that neither operation, taken separately, would have made visible.

The conceptual shift is small but consequential. White spaces are not a binary phenomenon. They are a two-dimensional one, defined by structural absence and characterised by commercial concentration. The first dimension tells us where technology has not yet followed science. The second tells us which of those open regions may warrant strategic attention. In practical terms, this gives an organisation scanning a large scientific corpus a defensible way to decide where to look first: rather than triaging thousands of papers by hand or by citation count, it can direct scarce analytical and scouting capacity toward the small set of clusters the framework places high, with the historical validation providing evidence that this narrowing is not arbitrary. The framework does not replace the judgement of firms, policymakers, or investors. It narrows the field within which such judgement becomes worthwhile.

Sources

- Abernathy, W. J., & Utterback, J. M. (1978). Patterns of industrial innovation. *Technology Review*, 80(7), 40–47.
- Alcácer, J., & Gittelman, M. (2006). Patent citations as a measure of knowledge flows: The influence of examiner citations. *The Review of Economics and Statistics*, 88(4), 774–779. <https://doi.org/10.1162/rest.88.4.774>
- Alibaba Cloud. (2026). *Supported models and capabilities overview: Model studio* [Model documentation for Qwen3.5-Plus, including qwen3.5-plus-2026-02-15]. Retrieved May 18, 2026, from <https://www.alibabacloud.com/help/en/model-studio/models>
- Arora, A., Belenzon, S., & Patacconi, A. (2018). The decline of science in corporate r&d. *Strategic Management Journal*, 39(1), 3–32. <https://doi.org/10.1002/smj.2693>
- Beltagy, I., Lo, K., & Cohan, A. (2019). Scibert: A pretrained language model for scientific text. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 3615–3620. <https://doi.org/10.18653/v1/D19-1371>
- Cohen, W. M., & Levinthal, D. A. (1990). Absorptive capacity: A new perspective on learning and innovation. *Administrative Science Quarterly*, 35(1), 128–152. <https://doi.org/10.2307/2393553>
- DNV. (2023). *Maritime forecast to 2050* (Energy Transition Outlook). DNV. Høvik, Norway. Retrieved May 12, 2026, from <https://www.dnv.com/maritime/maritime-forecast/>
- European Parliament and Council. (2009). Directive 2009/28/ec of the european parliament and of the council on the promotion of the use of energy from renewable sources. <https://eur-lex.europa.eu/eli/dir/2009/28/oj/eng>
- Fleming, L. (2001). Recombinant uncertainty in technological search. *Management Science*, 47(1), 117–132. <https://doi.org/10.1287/mnsc.47.1.117.10671>
- Fleming, L., & Sorenson, O. (2004). Science as a map in technological search. *Strategic Management Journal*, 25(8–9), 909–928. <https://doi.org/10.1002/smj.384>

- Intergovernmental Panel on Climate Change. (2022). *Climate change 2022: Mitigation of climate change. contribution of working group iii to the sixth assessment report of the intergovernmental panel on climate change*. Cambridge University Press. <https://doi.org/10.1017/9781009157926>
- International Energy Agency. (2023a). *Aviation and shipping*. <https://www.iea.org/reports/aviation-and-shipping>
- International Energy Agency. (2023b). *Global hydrogen review 2023*. <https://www.iea.org/reports/global-hydrogen-review-2023>
- International Energy Agency. (2023c). *Transport: Tracking clean energy progress* (tech. rep.). International Energy Agency. Paris. Retrieved May 12, 2026, from <https://www.iea.org/energy-system/transport>
- International Energy Agency. (2024). *Renewable fuels*. <https://www.iea.org/reports/renewables-2024/renewable-fuels>
- International Maritime Organization. (2023). 2023 IMO strategy on reduction of GHG emissions from ships. *International Maritime Organization*. Retrieved May 12, 2026, from <https://www.imo.org/en/OurWork/Environment/Pages/2023-IMO-Strategy-on-Reduction-of-GHG-Emissions-from-Ships.aspx>
- Kline, S. J., & Rosenberg, N. (1986). An overview of innovation. In R. Landau & N. Rosenberg (Eds.), *The positive sum strategy: Harnessing technology for economic growth* (pp. 275–305). National Academy Press. <https://doi.org/10.17226/612>
- Lemley, M. A., & Sampat, B. (2012). Examiner characteristics and patent office outcomes. *The Review of Economics and Statistics*, 94(3), 817–827. https://doi.org/10.1162/REST_a.00194
- Liu, J., Amdaoud, M., Mescheba, W., Quemener, J., Sapinho, D., Deltorn, J.-M., & Guellec, D. (2025). Npl v. nlp: Analysing the links between science and technology with citations and semantics. *Scientometrics*, 130(11), 6063–6090. <https://doi.org/10.1007/s11192-025-05460-6>
- Marx, M., & Fuegi, A. (2020). Reliance on science: Worldwide front-page patent citations to scientific articles. *Strategic Management Journal*, 41(9), 1572–1594. <https://doi.org/10.1002/smj.3145>

- Marx, M., & Fuegi, A. (2023). Reliance on science. <https://doi.org/10.5281/zenodo.7903131>
- Masclans, R., Hasan, S., & Cohen, W. M. (2024). The commercial potential of science. <https://doi.org/10.5281/zenodo.10815144>
- Masclans, R., Hasan, S., & Cohen, W. M. (2025). Measuring the commercial potential of science. *Strategic Management Journal*, 46(9), 2199–2236. <https://doi.org/10.1002/smj.3720>
- McInnes, L., Healy, J., & Astels, S. (2017). Hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11), 205. <https://doi.org/10.21105/joss.00205>
- McInnes, L., Healy, J., Saul, N., & Grossberger, L. (2018). Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29), 861. <https://doi.org/10.21105/joss.00861>
- Narin, F., Hamilton, K. S., & Olivastro, D. (1997). The increasing linkage between u.s. technology and public science. *Research Policy*, 26(3), 317–330. [https://doi.org/10.1016/S0048-7333\(97\)00013-9](https://doi.org/10.1016/S0048-7333(97)00013-9)
- Nooteboom, B., Van Haverbeke, W., Duysters, G., Gilsing, V., & van den Oord, A. (2007). Optimal cognitive distance and absorptive capacity. *Research Policy*, 36(7), 1016–1034. <https://doi.org/10.1016/j.respol.2007.04.003>
- Poth, C., Sterz, H., Paul, I., Purkayastha, S., Englander, L., Imhof, T., Vulic, I., Ruder, S., Gurevych, I., & Pfeiffer, J. (2023). Adapters: A unified library for parameter-efficient and modular transfer learning. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 149–160. <https://aclanthology.org/2023.emnlp-demo.13>
- Priem, J., Piwowar, H., & Orr, R. (2022). Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. <https://doi.org/10.48550/arXiv.2205.01833>
- Shibata, N., Kajikawa, Y., & Sakata, I. (2010). Extracting the commercialization gap between science and technology: Case study of a solar cell. *Technological Forecasting and Social Change*, 77(7), 1147–1155. <https://doi.org/10.1016/j.techfore.2010.03.008>

Singh, A., D'Arcy, M., Cohan, A., Downey, D., & Feldman, S. (2022). Scirepeval: A multi-format benchmark for scientific document representations. <https://arxiv.org/abs/2211.13308>

U.S. Environmental Protection Agency. (2010). Renewable fuel standard (rfs2): Final rule. <https://www.epa.gov/renewable-fuel-standard-program>

World Intellectual Property Organization. (n.d.). *Patent cooperation treaty faqs* [n.d.]. Retrieved February 17, 2025, from <https://www.wipo.int/pct/en/faqs/faqs.html>

Yang, A., et al. (2025). Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

List of Figures

2.1	Shibata’s parallel-layer detection logic	8
3.7.1	Paper-level commercial-potential score distributions by source	27
3.7.2	Methodological framework	30
4.1	Cluster-centroid topology in shared two-dimensional UMAP space	36
4.2	Nearest-patent distance distribution and distance-gate thresholds	37
4.3	Top 10 VALID white spaces by commercial-potential concentration	40
4.4	Static opportunity matrix for the 20 VALID white-space clusters	41
4.5	PEM engineering and patent-cluster separation	44
4.6.1	Historical commercial-potential ranking and later patent-citation attention	47
4.6.2	Validation inside historical distance-filtered white-space candidates . . .	48

List of Tables

3.1	Distance-gate classification rule	23
4.1	Top 10 VALID white spaces by commercial-potential concentration	39
A1	CPC class distribution across the patent corpus	69
A2	Full distance-gate table for all science clusters	70
A3	Full VALID white-space ranking by pct_above_07	71
B1	Documented CPC classes used for patent search scoping	72
B2	Qwen cluster-labelling prompt template and generation settings	72
B3	Full Qwen validation prompt template	73
B4	Example structured Qwen output: PEM science cluster 27 versus patent cluster 39	73
C1	UMAP parameters used for clustering and visualisation	75
C2	Baseline HDBSCAN configuration and cluster outputs	75
C3	SciBERT regression configuration and performance	75
D1	Per-period pct_above_07 values for all VALID clusters	77
E1	Nearest-patent distance trajectories for VALID clusters	78
E2	Patent-citation momentum for VALID clusters	78
E3	Static quadrant and trajectory category for VALID clusters	79
E4	Trajectory category by current opportunity-matrix quadrant	80
F1	Web of Science query components for the scientific corpus	81
F2	Lens.org query components for the patent corpus	81

F3	Corpus source provenance and export dates	82
G1	Control diagnostics for abandoned geometry-derived validation attempts	83
H1	Spearman correlations and top-5 stability versus the 0.70 cutoff	84
H2	Full Spearman rank-correlation matrix across pct_above cutoffs	84
H3	Cluster ranks across alternative pct_above cutoffs	84
I1	Correlation between predicted-score share and ranking variables across the 20 VALID clusters	86
I2	Top 10 VALID clusters: full ranking vs. Zenodo-only ranking	86

Appendices

A Patent Corpus, Distance Gate, and Final VALID Ranking Tables

This appendix reports the tabular evidence behind the patent corpus composition, the distance gate, and the final ranking of VALID white-space clusters. CPC counts are non-exclusive because a patent record can carry multiple CPC classes. The full derived CPC distribution across all observed CPC class prefixes is saved as a CSV in the appendices folder; Table A1 keeps the PDF concise by reporting the top observed classes plus all query-relevant classes.

Table A1

CPC class distribution across the patent corpus

CPC class	Description	Patents	Share	Query class
Y02E	Climate-mitigation technologies in energy	8,314	23.8%	
Y02P	Climate-mitigation technologies in production	3,808	10.9%	
C10L	Fuels not otherwise provided for	2,712	7.8%	Yes
Y02T	Climate-mitigation technologies in transport	2,229	6.4%	
C10G	Cracking hydrocarbon oils; fuel refining	1,669	4.8%	Yes
F01K	Steam-engine plants and power-generation systems	1,586	4.5%	
B01D	Separation and purification processes	1,526	4.4%	
C25B	Electrolytic processes	1,465	4.2%	Yes
C01B	Hydrogen and inorganic gases	1,414	4.1%	Yes
B01J	Catalysis and chemical/physical processes	1,351	3.9%	
C07C	Acyclic or carbocyclic compounds	1,171	3.4%	
C12P	Fermentation and bioprocess production	1,007	2.9%	Yes
F02C	Gas-turbine plants	979	2.8%	
C02F	Treatment of water or wastewater	887	2.5%	
Y02A	Climate adaptation technologies	886	2.5%	
C10J	Gasification and synthesis-gas production	856	2.5%	
F02M	Fuel supply and fuel-injection apparatus	853	2.4%	Yes
F01D	Turbines	836	2.4%	
C12G	Wine and alcoholic beverages	798	2.3%	
Y02W	Climate technologies in waste management	778	2.2%	
H01M	Electrochemical cells and batteries	776	2.2%	
C11C	Fatty acids, oils, and waxes	754	2.2%	
F02B	Internal combustion engines	732	2.1%	Yes
F02D	Engine control	715	2.1%	
C12N	Microorganisms or enzymes	543	1.6%	
F05D		517	1.5%	
F03B		515	1.5%	
C10K		508	1.5%	
Y02C		462	1.3%	
C12M		458	1.3%	

Table A2 reports all 39 science clusters and their nearest patent-cluster distance. The operational classes are FALSE_POSITIVE for $d_{\min} < 0.020$, ADJACENT for $0.020 \leq d_{\min} < 0.026$, and VALID for $d_{\min} \geq 0.026$.

Table A2

Full distance-gate table for all science clusters

Cluster	Label	Papers	Nearest patent	d_min	Class
0	Residential Biomass Combustion Emissions	425	25	0.0447	VALID
1	Biodiesel Glycerol Valorization Processes	849	30	0.0244	ADJACENT
2	Biodiesel Corrosion and Materials	392	27	0.0308	VALID
3	Renewable Jet Fuel Technologies	1228	22	0.0318	VALID
4	Bioelectrochemical Systems	362	39	0.0308	VALID
5	Biofuel and Biolubricant Tribology	342	23	0.0182	FALSE_POSITIVE
6	Alternative Fuel Combustion Kinetics	1114	22	0.0290	VALID
7	Lipase Catalyzed Biodiesel Production	1171	30	0.0255	ADJACENT
8	Thermochemical Conversion of Algal Biomass	296	31	0.0254	ADJACENT
9	Microalgae Biofuel Production	2131	31	0.0203	ADJACENT
10	Marine Alternative Fuel Technologies	224	11	0.0455	VALID
11	Fischer Tropsch Synthetic Fuel Catalysis	341	39	0.0221	ADJACENT
12	Low Carbon Transport Fuel Pathways	574	22	0.0450	VALID
13	Glycerol Fermentation Processes	462	31	0.0275	VALID
14	Hydrothermal Biomass Conversion	347	26	0.0322	VALID
15	Biodiesel Engine Performance Optimization	475	22	0.0370	VALID
16	Biohydrogen Production	203	31	0.0308	VALID
17	Biomass Thermochemical Conversion Processes	3098	26	0.0166	FALSE_POSITIVE
18	Biofuel Life Cycle Assessment	2698	26	0.0337	VALID
19	Biomass Gasification For Fuel Synthesis	249	25	0.0218	ADJACENT
20	Power to X Synthetic Fuels	304	34	0.0307	VALID
21	Biodiesel Physicochemical Properties	505	30	0.0232	ADJACENT
22	Electrocatalyst and Photocatalyst Material Design	1367	39	0.0175	FALSE_POSITIVE
23	Diesel Biodiesel Particulate Emissions	546	22	0.0345	VALID
24	Soot Formation and Oxidation	216	22	0.0426	VALID
25	Alternative Fuel Spray Characteristics	717	22	0.0320	VALID
26	Green Hydrogen Production	1676	34	0.0234	ADJACENT
27	PEM Electrolyzer Component Engineering and Transport	410	39	0.0359	VALID
28	Heterogeneous Catalytic Biodiesel Production	3687	30	0.0200	ADJACENT
29	Biodiesel Production Technology	3150	30	0.0129	FALSE_POSITIVE
30	Biodiesel Blend Engine Performance	4107	22	0.0217	ADJACENT
31	Nanoparticle Enhanced Diesel Fuels	915	22	0.0261	VALID
32	Biobutanol Separation And Recovery	240	31	0.0308	VALID
33	Waste To Biofuel Conversion	268	36	0.0253	ADJACENT
34	Lignocellulosic Ethanol Production	1401	31	0.0231	ADJACENT
35	Lignocellulosic Bioethanol Fermentation	429	31	0.0276	VALID
36	Biobutanol Fermentation Processes	263	31	0.0251	ADJACENT
37	Microbial Lipid Biodiesel Production	587	31	0.0186	FALSE_POSITIVE
38	Engineered Microbial Biofuel Production	251	31	0.0238	ADJACENT

Table A3*Full VALID white-space ranking by pct.above_07*

Rank	Cluster	Label	pct.above_07	d_min	Quadrant	Papers
1.0	35	Lignocellulosic Bioethanol Fermentation	79.3%	0.0276	Near-Frontier Opps.	429
2.0	27	PEM Electrolyzer Component Engineering and Transport	72.4%	0.0359	Prime Targets	410
3.0	4	Bioelectrochemical Systems	67.1%	0.0308	Near-Frontier Opps.	362
4.0	13	Glycerol Fermentation Processes	49.4%	0.0275	Near-Frontier Opps.	462
5.0	32	Biobutanol Separation And Recovery	48.8%	0.0308	Near-Frontier Opps.	240
6.0	16	Biohydrogen Production	43.8%	0.0308	Near-Frontier Opps.	203
7.0	3	Renewable Jet Fuel Technologies	29.6%	0.0318	Near-Frontier Opps.	1228
8.0	31	Nanoparticle Enhanced Diesel Fuels	25.9%	0.0261	Near-Frontier Opps.	915
9.0	6	Alternative Fuel Combustion Kinetics	21.8%	0.0290	Near-Frontier Opps.	1114
10.0	20	Power to X Synthetic Fuels	21.1%	0.0307	Near-Frontier Opps.	304
11.0	23	Diesel Biodiesel Particulate Emissions	19.4%	0.0345	Deep-Frontier Gaps	546
12.0	14	Hydrothermal Biomass Conversion	19.3%	0.0322	Deep-Frontier Gaps	347
13.0	24	Soot Formation and Oxidation	15.3%	0.0426	Deep-Frontier Gaps	216
14.0	25	Alternative Fuel Spray Characteristics	14.5%	0.0320	Deep-Frontier Gaps	717
15.0	12	Low Carbon Transport Fuel Pathways	12.4%	0.0450	Deep-Frontier Gaps	574
16.0	15	Biodiesel Engine Performance Optimization	10.7%	0.0370	Deep-Frontier Gaps	475
17.0	10	Marine Alternative Fuel Technologies	7.6%	0.0455	Deep-Frontier Gaps	224
18.0	18	Biofuel Life Cycle Assessment	5.8%	0.0337	Deep-Frontier Gaps	2698
19.0	2	Biodiesel Corrosion and Materials	5.4%	0.0308	Low Priority	392
20.0	0	Residential Biomass Combustion Emissions	4.5%	0.0447	Deep-Frontier Gaps	425

B Patent Search Classes and LLM Validation Materials

This appendix reports the CPC classes used to scope the Lens.org patent search and the LLM/Qwen materials used for cluster labelling and distance-gate validation. The complete Lens.org query and corpus download date are reported alongside the Web of Science query in Appendix F.

Table B1

Documented CPC classes used for patent search scoping

CPC class	Description
C10L	Fuels not otherwise provided for; fuel compositions and alternative fuels
C10G	Cracking hydrocarbon oils; fuel refining
C01B3	Hydrogen
C25B	Electrolytic processes
C12P7	Fermentation to produce organic compounds
F02B	Internal combustion engines
F02M	Fuel-injection apparatus

Qwen Cluster-Labelling Prompt Template

Table B2

Qwen cluster-labelling prompt template and generation settings

Prompt block	Template text
Model and settings	qwen3.5-plus-2026-02-15; temperature = 0.2; maximum output = 20 tokens.
System message	You are a concise, accurate scientific assistant.
Role and domain	You are an expert in alternative fuels, clean energy technology, and combustion science. You specialise in green hydrogen, biofuels, electrofuels and power-to-liquid fuels, ammonia as fuel, Fischer–Tropsch synthesis, and combustion systems.
Input framing	Below are abstracts from scientific papers that have been mathematically clustered into the same group. All papers come from the alternative-fuels research domain.
Abstract sample	For each cluster, 25 abstracts were sampled with random seed 42; each abstract was truncated to 500 characters in the prompt and listed as [1], [2], . . .
Task instruction	Identify the specific alternative fuel topic or technology that unites these papers. The label must reflect the precise fuel type, process, or application, not generic terms such as fuel research or energy production.
Specificity constraint	The label must distinguish this cluster from neighbouring research areas. Avoid broad sector names such as Alternative Fuels, Sustainable Aviation Fuels, or Biofuel Research that could describe multiple clusters; focus on the specific process, molecule, application, or mechanism that makes the cluster unique.
Output format	Provide a concise two- to five-word label. Do not use quotes or punctuation. Return only the label itself.

Qwen Distance-Gate Validation Prompt Template

Table B3

Full Qwen validation prompt template

Prompt block	Template text
Role	You are an expert in alternative fuels, clean energy technology, and patent analysis.
Case framing	Below are patents from patent cluster {nearest_patent_cluster_id} and papers from science cluster {science_cluster_id} ('{science_cluster_label}').
Task	Judge whether this science cluster is a real white space relative to its nearest patent cluster.
Classes	VALID: patents and science cover genuinely different topics and the white space is real. FALSE_POSITIVE: patents already cover what the science cluster covers. ADJACENT: patents and science are related, but operate at a different specificity or application boundary.
Instruction	Focus on semantic topic overlap, not on publication counts or hype. Be precise about what the patents do and do not cover.
Patent block	PATENT_CLUSTER {nearest_patent_cluster_id} ({patent_cluster_size} patents total / {patent_cluster_pct_of_corpus:.1f}% of corpus, showing {shown_patents}): {patent_block}
Science block	SCIENCE_CLUSTER {science_cluster_id} - {science_cluster_label} ({science_cluster_size} papers total, showing {n_sampled_papers} sampled): {paper_block}
Required JSON keys	patent_cluster_topic; science_cluster_topic; verdict; verdict_confidence; claimed_overlap_boundary; false_positive_check; adjacent_check; valid_check; explanation.
Formatting	verdict must be one of VALID, FALSE_POSITIVE, ADJACENT. verdict_confidence must be HIGH, MEDIUM, or LOW. Topic fields should be 2-3 sentences; boundary/check fields one sentence; explanation 3-4 sentences. Return JSON only, with no markdown fences and no extra commentary.

Table B4

Example structured Qwen output: PEM science cluster 27 versus patent cluster 39

Field	Value
science_cluster_id	27
science_cluster_label	PEM Electrolyzer Component Engineering and Transport
nearest_patent_cluster_id	39
nearest_patent_distance	0.035864
expected_classification	VALID
qwen_model	qwen3.5-plus-2026-02-15
qwen_verdict	VALID
verdict_confidence	HIGH
patent_cluster_topic	Focuses on catalyst composition and synthesis methods for CO2 hydrogenation into fuels (SAF, methanol, formic acid) and electrolysis electrode material fabrication (Ni/Co-based alloys).
science_cluster_topic	Concentrates on PEM electrolyzer device engineering, specifically porous transport layer (PTL) design, gas/liquid transport phenomena, water management, and operando diagnostics.
claimed_overlap_boundary	Both domains involve hydrogen production technologies, but patents target chemical activity and material synthesis while science targets physical transport and component architecture.
false_positive_check	FALSE_POSITIVE does not fit because the patents lack claims on PTL structure, wettability, or transport modeling central to the science cluster.
adjacent_check	ADJACENT does not fit because the operational boundary is distinct (material synthesis vs. device transport engineering), not merely a difference in specificity.

Field	Value
valid_check	VALID fits because the science cluster addresses device-level transport bottlenecks distinct from the catalyst-level innovations in the patent cluster.
explanation	The patent cluster is dominated by CO ₂ conversion catalysts and electrolysis electrode material synthesis (e.g., Ni-alloys, Co-phosphides), whereas the science cluster is highly specialized on PEM electrolyzer transport layers (PTL), gas management, and water distribution imaging. While patents mention porous electrodes, they do not cover the specific PTL engineering, wettability control, or transport modeling prevalent in the science papers. This indicates a genuine white space where scientific research on device engineering has not yet been fully captured by patents in this specific catalyst-focused cluster.

C Pipeline Parameters and SciBERT Commercial-Potential Prediction Layer

This appendix records the main implementation parameters behind the dimensionality-reduction, clustering, and prediction-layer steps referenced in Sections 3.5.1, 3.5.2, and 3.7.2. The prediction layer estimates Masclans/Zenodo compot scores for thesis-corpus papers that could not be matched directly to Zenodo. It predicts the existing compot score, not direct commercial outcomes, and preserves score provenance through the `compot_source` field.

UMAP and HDBSCAN Configuration

Table C1

UMAP parameters used for clustering and visualisation

Use	n_components	n_neighbors	min_dist	metric	random state	Role
Science clustering	50	15	0.0	cosine	42	HDBSCAN input for science corpus
Patent clustering	50	5	0.0	cosine	42	HDBSCAN input for patent corpus
Joint 2D map	2	15	0.1	cosine	42	Descriptive visualisation only

Table C2

Baseline HDBSCAN configuration and cluster outputs

Corpus	Input space	min_cluster_size	min_samples	Clusters	Clustered documents	Noise points
Science	UMAP-50D	200	100	39	38,020	20,608 (35.1%)
Patents	UMAP-50D	100	50	46	21,587	13,282 (38.1%)

SciBERT Prediction-Layer Configuration

Table C3

SciBERT regression configuration and performance

Item	Value
Backbone	allenai/scibert_scivocab_uncased
Matched direct-score papers	33,281
Predicted-score papers	25,347
Train/validation/test split	26,624 / 3,328 / 3,329
Split method	80/10/10, stratified by compot decile

Item	Value
Max sequence length	256 tokens
Batch size	8
Gradient accumulation	2 steps; effective batch size 16
Optimiser	AdamW
Learning rate	2×10^{-5}
Weight decay	0.01
Dropout	0.3
Warmup	6 percent linear warmup, then linear decay
Epochs	5
Mixed precision	fp16
Gradient clipping	max norm 1.0
GPU	NVIDIA GeForce GTX 1650, 4 GB VRAM
CUDA / PyTorch / Python	CUDA 12.4 / PyTorch 2.6.0+cu124 / Python 3.12.3
Training time	Approximately 3 hours per epoch; approximately 15 hours total
Best validation result	Val MSE 0.0278; Val MAE 0.1253; Val R^2 0.6038
Held-out test result	MAE 0.1283; RMSE 0.1748; R^2 0.5665

Training was interrupted after Epoch 3 by an OS-level sleep event and resumed from the checkpoint. The final model weights and held-out test metrics are the values reported in Table C1.

D Period-Based pct_above_07 Trajectories

Table D1 reports the share of papers above the 0.70 compot cutoff for each VALID cluster across the five windows referenced in Chapter 4. Papers before 2013 are excluded from this display because the appendix follows the chapter's five-window trajectory framing. The derived long-format audit file, including counts and predicted-score share, is saved as `appendices/appendix_d_pct_above_07_by_period.csv`.

Table D1

Per-period pct_above_07 values for all VALID clusters

Cluster	Label	2013-2017	2018-2020	2021-2022	2023-2024	2025-2026
35	Lignocellulosic Bioethanol Fermentation	86.2%	77.6%	66.2%	71.4%	80.8%
27	PEM Electrolyzer Component Engineering and Transport	58.3%	58.3%	61.5%	70.8%	81.7%
4	Bioelectrochemical Systems	74.1%	65.5%	60.7%	56.6%	85.4%
13	Glycerol Fermentation Processes	56.5%	42.7%	33.3%	39.4%	79.2%
32	Biobutanol Separation And Recovery	56.6%	35.6%	34.5%	58.8%	61.5%
16	Biohydrogen Production	43.8%	29.6%	44.4%	46.6%	59.1%
3	Renewable Jet Fuel Technologies	28.1%	25.7%	28.1%	38.5%	44.3%
31	Nanoparticle Enhanced Diesel Fuels	14.6%	7.1%	14.8%	36.9%	55.9%
6	Alternative Fuel Combustion Kinetics	21.9%	15.5%	22.9%	26.9%	38.6%
20	Power to X Synthetic Fuels	36.4%	20.0%	13.8%	18.7%	30.0%
23	Diesel Biodiesel Particulate Emissions	24.5%	15.0%	19.7%	33.9%	23.1%
14	Hydrothermal Biomass Conversion	29.4%	14.8%	11.7%	13.7%	32.8%
24	Soot Formation and Oxidation	19.6%	12.5%	17.5%	17.6%	18.2%
25	Alternative Fuel Spray Characteristics	18.5%	8.4%	12.5%	19.2%	37.8%
12	Low Carbon Transport Fuel Pathways	9.1%	13.2%	18.1%	17.4%	13.4%
15	Biodiesel Engine Performance Optimization	3.0%	0.0%	5.4%	9.9%	39.7%
10	Marine Alternative Fuel Technologies	0.0%	10.5%	8.2%	9.1%	7.5%
18	Biofuel Life Cycle Assessment	10.0%	3.6%	2.6%	1.6%	4.4%
2	Biodiesel Corrosion and Materials	7.7%	5.4%	5.6%	4.7%	4.8%
0	Residential Biomass Combustion Emissions	7.4%	3.7%	4.5%	0.0%	15.8%

E Trajectory Diagnostic of VALID White Spaces

This exploratory diagnostic computes nearest patent-cluster distance trajectories across 2018, 2020, 2022, and 2025 patent snapshots; patent-attention momentum from post-2018 versus pre-2018 patent citations; science publication growth; and median cluster age. Patent snapshots were reclustered independently with UMAP reduced to 50 dimensions using random state 42 and HDBSCAN with minimum cluster size 100 and minimum samples 50. Science centroids were fixed from the current 768-dimensional science-cluster embeddings, while patent centroids were recomputed for each snapshot.

Table E1

Nearest-patent distance trajectories for VALID clusters

Cluster	Label	Current d	2018	2020	2022	2025	Slope
35	Lignocellulosic Bioethanol Fermentation	0.0276	0.0635	0.0640	0.0643	0.0645	0.000142
27	PEM Electrolyzer Component Engineering and Transport	0.0359	0.0549	0.0550	0.0544	0.0530	-0.000285
4	Bioelectrochemical Systems	0.0308	0.0537	0.0540	0.0537	0.0528	-0.000127
13	Glycerol Fermentation Processes	0.0275	0.0568	0.0578	0.0585	0.0590	0.000308
32	Biobutanol Separation And Recovery	0.0308	0.0430	0.0437	0.0441	0.0443	0.000180
16	Biohydrogen Production	0.0308	0.0537	0.0544	0.0545	0.0536	-0.000029
3	Renewable Jet Fuel Technologies	0.0318	0.0317	0.0326	0.0330	0.0336	0.000254
31	Nanoparticle Enhanced Diesel Fuels	0.0261	0.0384	0.0393	0.0399	0.0404	0.000279
6	Alternative Fuel Combustion Kinetics	0.0290	0.0320	0.0328	0.0333	0.0335	0.000219
20	Power to X Synthetic Fuels	0.0307	0.0341	0.0346	0.0345	0.0339	-0.000032
23	Diesel Biodiesel Particulate Emissions	0.0345	0.0456	0.0464	0.0470	0.0479	0.000325
14	Hydrothermal Biomass Conversion	0.0322	0.0464	0.0471	0.0476	0.0475	0.000156
24	Soot Formation and Oxidation	0.0426	0.0466	0.0474	0.0478	0.0482	0.000230
25	Alternative Fuel Spray Characteristics	0.0320	0.0349	0.0356	0.0360	0.0368	0.000267
12	Low Carbon Transport Fuel Pathways	0.0450	0.0432	0.0436	0.0437	0.0440	0.000113
15	Biodiesel Engine Performance Optimization	0.0370	0.0407	0.0414	0.0420	0.0425	0.000261
10	Marine Alternative Fuel Technologies	0.0455	0.0384	0.0391	0.0393	0.0393	0.000118
18	Biofuel Life Cycle Assessment	0.0337	0.0489	0.0494	0.0496	0.0500	0.000151
2	Biodiesel Corrosion and Materials	0.0308	0.0354	0.0357	0.0360	0.0360	0.000079
0	Residential Biomass Combustion Emissions	0.0447	0.0480	0.0483	0.0484	0.0488	0.000108

Table E2

Patent-citation momentum for VALID clusters

Cluster	Label	pct.above.07	Cites pre	Cites post	Momentum
35	Lignocellulosic Bioethanol Fermentation	79.3%	241	240	1.00
27	PEM Electrolyzer Component Engineering and Transport	72.4%	1	6	6.00
4	Bioelectrochemical Systems	67.1%	35	26	0.74
13	Glycerol Fermentation Processes	49.4%	79	56	0.71

Cluster	Label	pct.above.07	Cites pre	Cites post	Momentum
32	Biobutanol Separation And Recovery	48.8%	14	13	0.93
16	Biohydrogen Production	43.8%	4	8	2.00
3	Renewable Jet Fuel Technologies	29.6%	16	26	1.62
31	Nanoparticle Enhanced Diesel Fuels	25.9%	4	10	2.50
6	Alternative Fuel Combustion Kinetics	21.8%	5	9	1.80
20	Power to X Synthetic Fuels	21.1%	25	40	1.60
23	Diesel Biodiesel Particulate Emissions	19.4%	1	2	2.00
14	Hydrothermal Biomass Conversion	19.3%	28	12	0.43
24	Soot Formation and Oxidation	15.3%	2	1	0.50
25	Alternative Fuel Spray Characteristics	14.5%	7	2	0.29
12	Low Carbon Transport Fuel Pathways	12.4%	8	9	1.12
15	Biodiesel Engine Performance Optimization	10.7%	0	0	0.00
10	Marine Alternative Fuel Technologies	7.6%	0	1	1.00
18	Biofuel Life Cycle Assessment	5.8%	129	40	0.31
2	Biodiesel Corrosion and Materials	5.4%	7	6	0.86
0	Residential Biomass Combustion Emissions	4.5%	1	2	2.00

Table E3

Static quadrant and trajectory category for VALID clusters

Cluster	Label	Quadrant	Trajectory
35	Lignocellulosic Bioethanol Fermentation	Near-Frontier Opportunities	Other Dynamic
27	PEM Electrolyzer Component Engineering and Transport	Prime Targets	Emerging Frontier
4	Bioelectrochemical Systems	Near-Frontier Opportunities	Other Dynamic
13	Glycerol Fermentation Processes	Near-Frontier Opportunities	Other Dynamic
32	Biobutanol Separation And Recovery	Near-Frontier Opportunities	Other Dynamic
16	Biohydrogen Production	Near-Frontier Opportunities	Other Dynamic
3	Renewable Jet Fuel Technologies	Near-Frontier Opportunities	Other Dynamic
31	Nanoparticle Enhanced Diesel Fuels	Near-Frontier Opportunities	Other Dynamic
6	Alternative Fuel Combustion Kinetics	Near-Frontier Opportunities	Other Dynamic
20	Power to X Synthetic Fuels	Near-Frontier Opportunities	Other Dynamic
23	Diesel Biodiesel Particulate Emissions	Deep-Frontier Gaps	Other Dynamic
14	Hydrothermal Biomass Conversion	Deep-Frontier Gaps	Other Dynamic
24	Soot Formation and Oxidation	Deep-Frontier Gaps	Dormant/Non-Patent Gap
25	Alternative Fuel Spray Characteristics	Deep-Frontier Gaps	Persistent Gap
12	Low Carbon Transport Fuel Pathways	Deep-Frontier Gaps	Other Dynamic
15	Biodiesel Engine Performance Optimization	Deep-Frontier Gaps	Other Dynamic
10	Marine Alternative Fuel Technologies	Deep-Frontier Gaps	Other Dynamic
18	Biofuel Life Cycle Assessment	Deep-Frontier Gaps	Dormant/Non-Patent Gap
2	Biodiesel Corrosion and Materials	Low Priority	Other Dynamic
0	Residential Biomass Combustion Emissions	Deep-Frontier Gaps	Other Dynamic

PEM Electrolyzer Component Engineering and Transport is the highest patent-attention momentum case in this diagnostic, with a post-2018/pre-2018 citation ratio of 6.0×. The diagnostic remains exploratory because it covers only 20 VALID clusters, uses snapshot reclustering that is HDBSCAN-parameter sensitive, and depends on patent-citation matching with limited DOI/OpenAlex/Reliance-on-Science coverage.

Table E4*Trajectory category by current opportunity-matrix quadrant*

Trajectory category	Prime Targets	Near-Frontier Opportunities	Deep-Frontier Gaps	Low Priority
Emerging Frontier	1	0	0	0
Persistent Gap	0	0	1	0
Dormant/Non-Patent Gap	0	0	2	0
Other Dynamic	0	9	6	1

F Search Query Provenance

This appendix reports the query provenance for both corpora. The Web of Science and Lens.org exports were both downloaded on 17 February 2026. Tables F1 and F2 use the same structure so that the scientific and patent retrieval designs can be compared directly; Table F3 summarises the corpus-level provenance.

Web of Science Scientific Corpus Query

Table F1

Web of Science query components for the scientific corpus

Query block	Terms
Field	Web of Science Topic Search, TS, covering titles, abstracts, author keywords, and Keywords Plus.
Boolean structure	[Fuel terms] AND ([Production terms] OR [Application terms])
Fuel terms	green hydrogen; biofuel*; biodiesel; bio-jet; electrofuel*; e-fuel*; synthetic fuel*; power-to-liquid; PtL; PtX; power-to-x; SAF; sustainable aviation fuel; methanol-to-gasoline; renewable diesel; HEFA; alcohol-to-jet
Production terms	electrolysis; pyrolysis; gasification; fermentation; catalysis; Fischer-Tropsch; metabolic engineering; CO ₂ hydrogenation; transesterification; hydroprocessing; reverse water gas shift; RWGS
Application terms	combustion; engine; turbine; ignition; flame speed; injection strategy; materials compatibility; corrosion; swelling; emissions; NO _x ; soot; aftertreatment; performance characteristics
Download date	17 February 2026

Lens.org Patent Corpus Query

Table F2

Lens.org query components for the patent corpus

Query block	Terms
Field	Lens.org Abstract field for keyword terms; CPC Classification Code field for classification filters.
Boolean structure	[Abstract fuel terms] AND [Abstract production/application terms] AND [CPC classification filter]
Normalisation note	The Lens interface query contained repeated biodiesel and power-to-liquid clauses; they are listed once here because duplicate OR terms do not change retrieval.
Abstract fuel terms	green AND hydrogen; biofuel; biodiesel; bio-jet; electrofuel; e-fuel; synthetic AND fuel; power-to-liquid; sustainable AND aviation AND fuel; SAF; HEFA; alcohol-to-jet; renewable AND diesel; methanol-to-gasoline
Abstract production/application terms	electrolysis; pyrolysis; gasification; fermentation; catalysis; Fischer-Tropsch; CO ₂ AND hydrogenation; hydroprocessing; transesterification; RWGS; combustion; turbine; ignition; aftertreatment; emissions; NO _x ; soot; fuel AND injection; corrosion
CPC classification filter	C10L*; C10G*; C01B3*; C25B*; C12P7*; F02B*; F02M*
Download date	17 February 2026

Table F3*Corpus source provenance and export dates*

Corpus	Source database	Export date	Final corpus size
Scientific publications	Web of Science Core Collection	17 February 2026	58,628 records
Patent publications	Lens.org patent database	17 February 2026	34,869 records

G Rejected Geometry-Derived Validation Attempts

Earlier validation attempts used outcomes derived from the same science-patent embedding geometry as the detection pipeline. They were rejected as validation evidence because such outcomes did not provide an external test of the ranking claim. After controls, baseline distance structure dominated the apparent effects, so the thesis moved to post-2018 patent-to-paper citations as the validation outcome reported in Chapter 4.

Table G1

Control diagnostics for abandoned geometry-derived validation attempts

Control test	Result	Interpretation
Partial correlation	Raw $\rho = -0.225101$, $p < .001$; partial $\rho = 0.017899$, $p = .322515$	The compot-compression signal disappears after controlling for baseline distance.
Distance-stratified compression	CLOSE $p = .045285$; MIDDLE $p = .349229$; FAR $p = .591374$	The effect is not stable across comparable distance strata.
Within Group A	Compression vs compot $\rho = 0.024125$, $p = .511447$; compression vs pre-distance $\rho = 0.442421$, $p < .001$	Compression is much more related to pre-existing distance than to compot.
Matched-distance comparison	655 matched pairs, $p = .180005$	Matching on baseline distance removes the apparent effect.
Regression	classification dummy $p = .197923$; compot final $p = .131233$; baseline distance $p < .001$	Baseline distance dominates the model.

These checks are methodological triage rather than negative evidence against the thesis claim: they show why a clean validation outcome had to be external to the embedding geometry used for detection and ranking.

H pct_above_07 Cutoff Sensitivity Analysis

This appendix reports the robustness checks for the high-compot cutoff used to construct pct_above_07. The full Spearman matrix is available for the cutoff-by-cutoff rank order; p-values are available for each cutoff against the 0.70 baseline.

Table H1

Spearman correlations and top-5 stability versus the 0.70 cutoff

Cutoff	rho vs 0.70	p-value	Top-5 cluster IDs	Jaccard
0.50	0.988	4.80e-16	35,27,4,32,13	1.0
0.55	0.995	7.23e-20	35,27,4,13,32	1.0
0.60	0.994	9.58e-19	35,27,4,13,32	1.0
0.65	0.997	1.89e-21	35,27,4,13,32	1.0
0.70	1.000	0.0	35,27,4,13,32	1.0
0.75	0.994	9.58e-19	35,27,4,13,32	1.0
0.80	0.980	3.69e-14	35,27,4,13,32	1.0
0.85	0.950	1.45e-10	35,27,4,32,13	1.0

Table H2

Full Spearman rank-correlation matrix across pct_above cutoffs

cutoff	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85
0.50	1.000	0.989	0.991	0.988	0.988	0.979	0.959	0.929
0.55	0.989	1.000	0.994	0.991	0.995	0.988	0.974	0.952
0.60	0.991	0.994	1.000	0.997	0.994	0.985	0.970	0.937
0.65	0.988	0.991	0.997	1.000	0.997	0.992	0.979	0.944
0.70	0.988	0.995	0.994	0.997	1.000	0.994	0.980	0.950
0.75	0.979	0.988	0.985	0.992	0.994	1.000	0.994	0.970
0.80	0.959	0.974	0.970	0.979	0.980	0.994	1.000	0.979
0.85	0.929	0.952	0.937	0.944	0.950	0.970	0.979	1.000

Table H3

Cluster ranks across alternative pct_above cutoffs

Cluster	Label	Base	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85
35	Lignocellulosic Bioethanol Fermentation	1	1	1	1	1	1	1	1	1
27	PEM Electrolyzer Component Engineering and Transport	2	2	2	2	2	2	2	2	2
4	Bioelectrochemical Systems	3	3	3	3	3	3	3	3	3
13	Glycerol Fermentation Processes	4	5	4	4	4	4	4	4	5
32	Biobutanol Separation And Recovery	5	4	5	5	5	5	5	5	4
16	Biohydrogen Production	6	6	6	6	6	6	6	6	6
3	Renewable Jet Fuel Technologies	7	8	7	7	7	7	7	7	7
31	Nanoparticle Enhanced Diesel Fuels	8	7	8	8	8	8	8	8	11
6	Alternative Fuel Combustion Kinetics	9	9	9	10	10	9	9	9	9
20	Power to X Synthetic Fuels	10	10	11	9	9	10	11	12	14

Cluster	Label	Base	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85
23	Diesel Biodiesel Particulate Emissions	11	11	10	11	11	11	10	10	8
14	Hydrothermal Biomass Conversion	12	12	12	12	12	12	12	11	10
24	Soot Formation and Oxidation	13	14	13	14	14	13	14	15	13
25	Alternative Fuel Spray Characteristics	14	13	14	13	13	14	13	13	12
12	Low Carbon Transport Fuel Pathways	15	17	15	15	15	15	15	14	15
15	Biodiesel Engine Performance Optimization	16	16	17	17	16	16	16	17	18
10	Marine Alternative Fuel Technologies	17	15	16	16	17	17	18	19	19
18	Biofuel Life Cycle Assessment	18	19	19	19	18	18	17	16	16
2	Biodiesel Corrosion and Materials	19	18	18	18	19	19	20	20	20
0	Residential Biomass Combustion Emissions	20	20	20	20	20	20	19	18	17

I Predicted-Score Robustness Diagnostics

This appendix reports diagnostic tests assessing whether the share of predicted, rather than directly matched Zenodo, compot scores distorts the ranking of VALID clusters. The concern is that clusters with higher proportions of SciBERT-predicted scores might move mechanically in the ranking, rather than because they contain a stronger concentration of commercially promising papers.

Table I1

Correlation between predicted-score share and ranking variables across the 20 VALID clusters

Relationship	Spearman ρ	p -value
Predicted-score share vs. cluster median publication year	.742	< .001
Predicted-score share vs. pct.above.07	.041	.865
Predicted-score share vs. current rank	-.041	.865

As a second diagnostic, the full ranking was recomputed using only the 33,281 directly matched Zenodo scores, dropping the 25,347 SciBERT-predicted scores. The Spearman correlation between the Zenodo-only ranking and the full ranking is .965, with maximum rank shifts of approximately three positions. The prediction layer is therefore recency-driven, because more recent clusters have higher predicted-score shares as their papers are less likely to appear in the Zenodo release, rather than rank-distorting.

Table I2

Top 10 VALID clusters: full ranking vs. Zenodo-only ranking

Cluster	Full-ranking position	Zenodo-only position
Lignocellulosic Bioethanol Fermentation	1	1
PEM Electrolyzer Component Engineering and Transport	2	3
Bioelectrochemical Systems	3	2
Glycerol Fermentation Processes	4	4
Biobutanol Separation And Recovery	5	5
Biohydrogen Production	6	6
Renewable Jet Fuel Technologies	7	7
Nanoparticle Enhanced Diesel Fuels	8	11
Alternative Fuel Combustion Kinetics	9	9
Power to X Synthetic Fuels	10	12

FACULTY OF BUSINESS AND ECONOMICS

Naamsestraat 69 bus 3500

3000 LEUVEN, België

feb.leuven@kuleuven.be

www.feb.kuleuven.be

