WHITE PAPER · METHODS

Graph-Based Material Similarity for Engineering Shortlisting

An explainable method for finding related materials from structured technical evidence.

CAPABILITY BRIEF **Find Similarity** · SURFACES MaterialPoint, MCP

ABSTRACT

Engineering material selection often begins with an incomplete description rather than a perfect specification. Practitioners may know a commercial grade, a family name, a processing route, a small set of target properties, or a test condition, but rarely all of these at once. Conventional text search can retrieve names, and property filters can retrieve narrow numeric subsets, but neither approach captures the structural relationships among composition, processing, properties, and test context. This paper describes a graph-based material similarity tool, deployed within the Ainuric platform as **Find Similarity**, designed to support engineering shortlisting. The method represents each material as a typed evidence graph derived from a materials database and ranks candidate materials according to the similarity of their graph structure and recorded evidence. Queries may begin from a known material or from a partial engineering description such as a family designation, processing hint, or property target. The system returns a ranked shortlist together with matched evidence patterns and data-quality warnings. The approach improves discovery, triage, and comparison, while remaining interpretable enough for engineering review. It is not intended to certify equivalence or replace qualification testing.

AT A GLANCE

partial queries, not exact names · relationships, not isolated fields · explainable shortlists, not black-box scores

1 Introduction

Material selection is rarely a purely numerical exercise. In practical workflows, engineers reason across several dimensions at once: material family, processing route, architecture, constituent system, measured properties, and test conditions. These relationships are often hierarchical and context-dependent. A polymer grade identified as “molded PA66,” for example, carries information about chemistry, processability, and likely property regimes, even before a full datasheet has been assembled.

Materials informatics has increasingly emphasised the importance of structured data, machine learning, and reusable digital representations for decision support in materials engineering [1, 2]. At the same time, work on materials knowledge graphs has shown that materials data are naturally relational and provenance-sensitive rather than flat [3, 4]. The central idea behind Find Similarity is therefore simple: model each material as a graph of engineering evidence and compare those graphs directly.

This yields three practical advantages. First, it supports partial queries rather than requiring exact product names. Second, it captures relationships among facts rather than treating all fields as independent keywords. Third, it produces results that can be explained in human terms, such as shared process route, shared property evidence, or comparable test context.

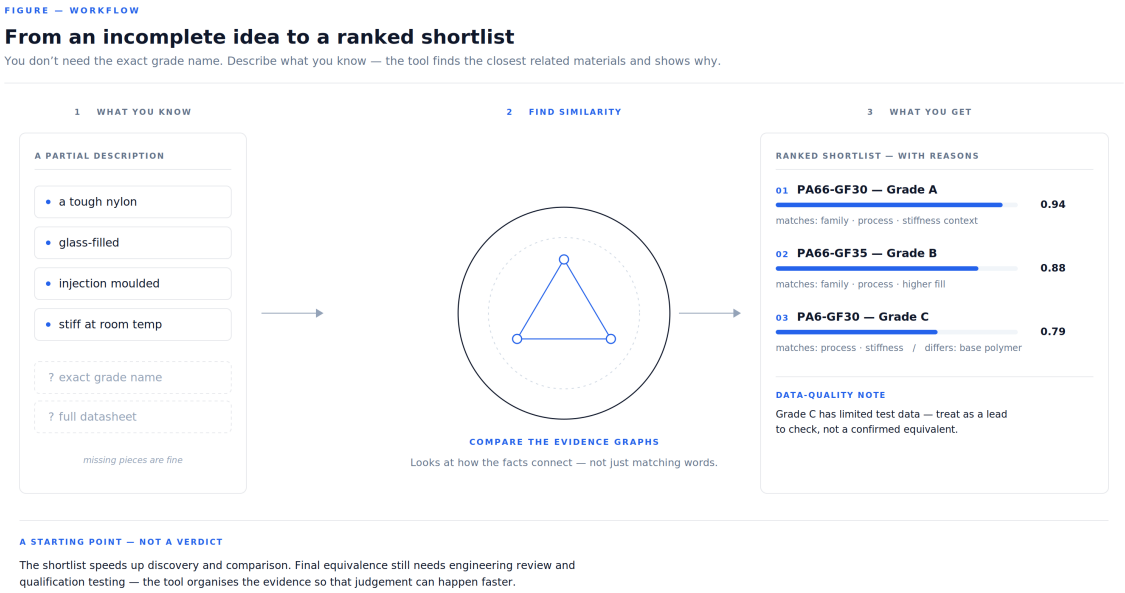

FIGURE 1

From an incomplete description to a ranked, explainable shortlist. The tool accepts partial engineering intent and returns evidence-matched candidates with explicit data-quality warnings.

2 Problem Statement

Many existing materials search tools are optimised for one of two modes:

1. exact or fuzzy retrieval by name, grade, or family;
2. structured filtering by explicit property constraints.

These modes are useful but incomplete. Name-based retrieval can miss functionally adjacent candi-

dates, while property-based retrieval can ignore process route, constituent system, or test context. In engineering practice, the relevant question is often not “Which materials contain this string?” or even “Which materials satisfy this single threshold?” but rather “Which materials look most similar to this material or this need, given the available evidence?”

Find Similarity addresses that problem as a *shortlisting* task. It is designed to identify nearby candidates, not to declare interchangeability. This distinction is important: similarity is a ranking construct grounded in recorded evidence, whereas equivalence remains an engineering judgement.

3 Conceptual Method

At a high level, the method has four stages: evidence representation, query representation, candidate generation, and similarity ranking.

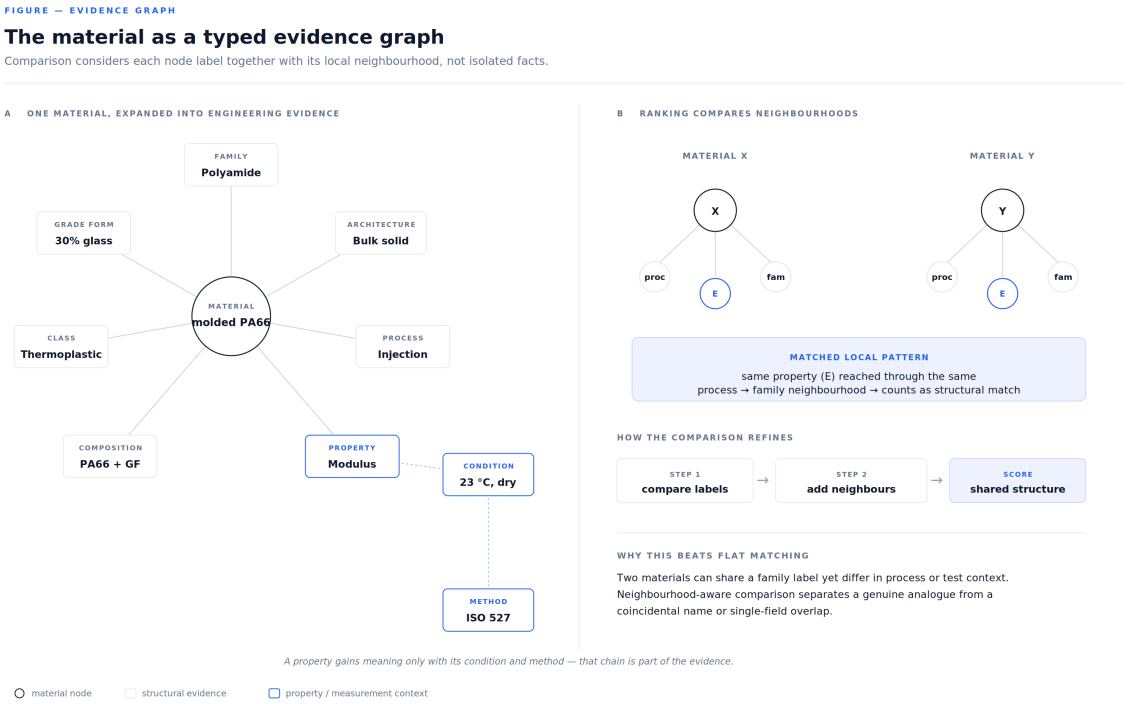

FIGURE 2

Each material is modelled as a typed evidence graph (left); ranking compares local neighbourhoods rather than isolated fields (right), in the spirit of the Weisfeiler–Lehman framework.

3.1 Evidence Representation

Each material in the underlying database is transformed into a typed graph. The graph is centred on the material and expanded through relationships that reflect engineering meaning. These include, at a minimum:

- material family or classification;
- manufacturing or processing route;
- architecture or structural form;
- composition or constituent evidence;
- property evidence;

- test-condition evidence;
- measurement-method context.

This design is motivated by the observation that materials data are inherently relational and hierarchical [3]. A property value without its measurement context is less informative than a property value tied to a condition and acquisition method. Likewise, a material family without process information may be too broad to support engineering shortlisting.

3.2 Query Representation

The tool supports two query modes.

For a **known material**, the reference graph is built directly from the stored database record.

For a **partial engineering description**, the system constructs a smaller "virtual" graph by extracting recognisable signals from the query. These may include a family designation, subtype, process cue, composition clue, target property, or operating condition. The goal is not to reconstruct a full datasheet from text, but to generate enough structured intent to support meaningful retrieval.

This distinction between full material graphs and partial virtual graphs is essential. In practice, engineers often begin with sparse intent and refine from there.

3.3 Candidate Generation

Rather than compare every material against every query with equal effort, the tool first creates a candidate pool using broad relevance signals. This stage acts as a retrieval screen and prioritises materials that are plausibly related by family, process, architecture, or available property evidence.

This design follows a common principle in materials informatics: retrieval and ranking should be separated when data are heterogeneous and sparse [1, 2]. Broad screening helps maintain scale and responsiveness, while the later ranking stage focuses on structural fit.

3.4 Similarity Ranking

For structurally rich comparisons, the tool uses a graph-kernel style approach inspired by the Weisfeiler–Lehman framework [5]. In conceptual terms, the ranking stage compares not only node labels, but also the local neighbourhoods around those labels. A property connected to a specific condition and method is therefore treated differently from the same property appearing in a different relational context.

This matters because engineering similarity is rarely defined by isolated facts. Two materials may share a family label but differ significantly in process route or property context. Conversely, two materials may differ in commercial name while still appearing close in structure and evidence.

For sparse queries, the tool uses a more conservative relevance view so that small but exact matches are not undervalued simply because the query graph contains less information than a fully documented material record. This helps keep the ranking interpretable for early-stage exploratory use.

4 Why a Graph-Based Approach

The value of the graph approach is not merely technical elegance. It reflects the actual structure of engineering reasoning.

A materials engineer does not typically evaluate density, process, composition, and test condition as unrelated spreadsheet columns. These factors are linked. A property is measured under some

condition. A condition pertains to a material in some processed state. A state arises from a process. A process acts on a formulation or constituent system. The graph formalism preserves these relationships in a way that flat similarity metrics often do not.

The use of a Weisfeiler–Lehman style graph comparison is particularly attractive for this setting because it offers a balance between expressive structural comparison and computational practicality [5]. It is also easier to explain than many opaque latent-space approaches — a property that matters in engineering environments where traceability and user trust are paramount.

5 Outputs and User Value

The tool is intended to return more than a ranked list. Its outputs are designed to support engineering review and shortlist formation. Typical outputs include:

- a ranked set of similar candidate materials;
- a similarity score for prioritisation;
- matched evidence patterns, such as shared process or property context;
- an estimate of evidence coverage;
- warnings where data are missing or sparse.

This makes the tool useful across several workflows:

- broadening a shortlist beyond exact grade-name matches;
- identifying adjacent candidates for substitution studies;
- finding comparison targets for benchmarking and testing;
- starting from incomplete user intent rather than a finished specification.

The broader materials-informatics literature has repeatedly emphasised that the practical value of digital tools depends not only on predictive capability but also on data quality, interpretability, and workflow fit [1, 2]. The present approach is aligned with that view.

6 Scope and Limitations

This method should be understood as an engineering discovery tool, not an automated qualification engine.

A high similarity score indicates that two materials are close with respect to the structured evidence available in the database and the query representation used at search time. It does not guarantee equivalence in performance, safety, regulatory status, process robustness, or application suitability.

Its quality depends on three things:

- the completeness of the underlying database;
- the quality of the query representation;
- the appropriateness of graph features for the decision context.

These are not flaws unique to this method; they are normal constraints in applied materials informatics [1, 2]. For that reason, results should be interpreted as shortlist guidance that supports further comparison, testing, and expert review.

7 Positioning within the Ainuric Platform

Find Similarity sits between simple search and full predictive modelling.

It is more expressive than keyword retrieval because it uses structured relationships, not just names. It is more interpretable than many black-box embedding systems because its outputs can be tied back to matched evidence. And it is lighter-weight than building a dedicated supervised model for every engineering decision problem.

Within the Ainuric architecture, the tool is delivered through two complementary surfaces. **MaterialPoint** provides the focused commercial environment for production engineering use, while the **Material Community Platform (MCP)** makes the same evidence-graph capability available across its broader, open-access toolset spanning structural metals, polymers, composites, ceramics, and other tracks. In both surfaces, the tool serves the practical front end of materials engineering: finding relevant nearby options, surfacing evidence-based analogues, and accelerating material selection and comparison.

8 Conclusion

Graph-based material similarity provides a practical way to formalise how engineers already reason about related materials. By representing each material as a typed evidence graph and comparing graph structure rather than isolated fields, the method supports better shortlisting from both known materials and incomplete engineering descriptions.

The approach is especially valuable where data are heterogeneous, queries are partial, and interpretability matters. Its purpose is not to replace expert judgement, but to organise evidence, improve retrieval quality, and help engineers reach the next defensible decision faster.

REFERENCES

01 R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, and C. Kim. Machine learning in materials informatics: recent applications and prospects. *npj Computational Materials* **3**, 54 (2017).
nature.com/articles/s41524-017-0056-5

02 R. Batra, L. Song, and R. Ramprasad. Emerging materials intelligence ecosystems propelled by machine learning. *Nature Reviews Materials* **6**, 655–678 (2021).
nature.com/articles/s41578-020-00255-y

03 Y. Zheng et al. A materials terminology knowledge graph automatically constructed from text corpus. *Scientific Data* **11**, 585 (2024).
nature.com/articles/s41597-024-03448-0

04 V. Venugopal and E. Olivetti. MatKG: An autonomously generated knowledge graph in Material Science. *Scientific Data* **11**, 217 (2024).
nature.com/articles/s41597-024-03039-z

05 N. Shervashidze, P. Schweitzer, E. J. van Leeuwen, K. Mehlhorn, and K. M. Borgwardt. Weisfeiler–Lehman Graph Kernels. *Journal of Machine Learning Research* **12**, 2539–2561 (2011).
jmlr.org/papers/v12/shervashidze11a.html