MELD: Multimodal Extraction of Machine-Learning-Ready Materials Data from Scientific Literature

Anonymous ACL submission

Abstract

Machine-learning models for materials design predict properties from composition and processing, and they need large, structured datasets of composition-process-property (CPP) records to train on. Such records are dispersed across the text, tables, and figures of the literature and are reported as N-ary relations in which a property value is tied to both a composition and a processing route. We present MELD, a multimodal pipeline that builds CPP datasets from full-length materials science papers. MELD parses a PDF into a layout-aware representation, reads property values directly from figures with a vision-language model, and aligns them with text-derived compositions and processing routes, normalizing units with a deterministic tool layer. We evaluate MELD on held-out, expert-annotated gold in two alloy systems, copper and aluminum. It reaches a fact-level recall of 0.55 on copper and 0.51 on aluminum; property values are read accurately (property recall 0.81 and 0.77), while the composition-processing binding is the bottleneck, and genuine fabrications account for about 8% of false positives. In a non-circular downstream test, property predictors trained on 231 papers of automatically extracted aluminum data match or exceed those trained on a small expert-curated set on three of four properties, raising the yield-strength test R^2 from 0.58 to 0.87. We release the extracted aluminum CPP dataset.

1 Introduction

Materials development is a high-cost and high-risk undertaking in which synthesis conditions and process variability shape the final properties of a material. To search the space of candidate materials more efficiently, recent work has adopted machine learning models, including graph neural networks, large language models, and vision-language models, for property prediction and inverse design (Jain et al., 2013; Dunn et al., 2020). Composition-based property prediction, crystal-structure-based prediction, and composition-structure multimodal models have each improved the efficiency of materials design and synthesis-recipe search.

The final properties of a synthesized material are not determined by composition or structure alone. For a fixed composition, the processing history, including heat treatment, aging, rolling, quenching, and solution treatment, alters the microstructure and the precipitation behavior, and with them properties such as yield strength, tensile strength, elongation, and electrical conductivity. An experimental materials dataset therefore has to be structured not as composition-property pairs but as composition-process-property (CPP) records. This view is consistent with the processing-structure-property and composition-processing-property relationships that underlie materials design.

Automatic extraction of structured materials data from the literature has been pursued with NLP and LLM methods, from domain-adapted language models such as MatSciBERT (Gupta et al., 2022) to prompt-engineered LLM extractors (Polak and Morgan, 2024; Dagdelen et al., 2024), and more recently to N-ary and multimodal efforts such as POLYIE (Cheung et al., 2024) and MatViX (Khalighinejad et al., 2025), the multimodal benchmark closest to our setting. Section 2 reviews this work in detail.

Existing work does not yet deliver the end-to-end CPP dataset construction that a materials researcher needs. Many systems are restricted to abstracts, selected sentences, tables, or a single modality (Shetty et al., 2023; Khalighinejad et al., 2024), and stop short of taking a full PDF, integrating composition, process, and property information scattered across the article, and converting it into a CPP dataset usable for downstream tabular modeling. Multimodal benchmarks such as MatViX establish the importance of full-length, multimodal extraction, but remain confined to poly-

mer nanocomposites and biodegradation, and their authors name an agentic extraction framework and downstream validation of extracted-data usefulness as open directions (Khalighinejad et al., 2025).

We address these limitations with MELD, a multimodal extraction pipeline that constructs CPP-schema datasets from materials-science literature. MELD takes a full PDF as input, identifies evidence spread across text, tables, figures, and captions, and aligns composition, processing conditions, and property values into sample-level records that form a CPP table for downstream machine learning. Property values reported only in figures are read directly by a vision-language model, and a deterministic tool layer normalizes units so that values reported under different conventions become comparable.

Our study proceeds in four stages. We first develop the extraction pipeline on a small set of copper-alloy papers drawn from an existing curated dataset and measure the contribution of each component. We then evaluate the pipeline on a held-out set of copper-alloy papers, assessing both how well it reproduces curated values and the validity of the records it produces beyond the curated gold, which lets us ask whether automatic extraction can extend the coverage of a human-curated dataset rather than only reproduce it. Next, we adapt the schema to precipitation-hardened aluminum alloys and apply the pipeline to a manually labeled aluminum set to quantify extraction accuracy under domain transfer. Finally, we scale the pipeline to a larger collection of open-access aluminum-alloy papers, build a large CPP dataset, and test whether property prediction is feasible with standard tabular machine-learning models, without a dedicated deep architecture.

These stages outline an end-to-end path from full-document scientific literature to a CPP dataset to downstream property prediction. Our contributions are as follows.

1. A multimodal extraction pipeline that aligns composition, process, and property information from full materials-science PDFs into sample-level CPP records, scored by a frozen, recall-oriented fact-level metric.

2. An evaluation on held-out copper-alloy papers that measures both the reproduction of a curated dataset and the validity of records extracted beyond it.

3. A manually labeled precipitation-hardened aluminum-alloy dataset together with a schema adaptation that quantify extraction accuracy under domain transfer.

4. A demonstration that the automatically extracted CPP data supports downstream tabular property prediction, indicating that literature-scale automatic construction can feed materials-informatics workflows.

2 Related Work

Materials information extraction from text and tables. Materials information extraction began with rule-based and statistical pipelines over text. ChemDataExtractor parses chemical entities, properties, and their values with a grammar-based toolkit (Swain and Cole, 2016), and surveys document the broader shift toward data-driven materials research through natural language processing (Olivetti et al., 2020). Domain adaptation then improved entity and relation extraction: unsupervised word embeddings trained on materials text capture latent knowledge (Tshitoyan et al., 2019), large-scale named entity recognition normalizes materials and properties across millions of abstracts (Weston et al., 2019), recipe mining structures inorganic synthesis at corpus scale (Kononova et al., 2019), and MatSciBERT (Gupta et al., 2022) adapts SciBERT (Beltagy et al., 2019) to the domain. Large language models raised accuracy and lowered annotation cost: general schemes structure scientific text into records (Dagdelen et al., 2024), ChatExtract reaches high precision through prompt engineering (Polak and Morgan, 2024), and a pipeline built on a domain language model extracts property records from polymer abstracts at scale (Shetty et al., 2023). These systems operate mainly on text and tables and stop at entities or binary relations, so they do not produce the figure-grounded, N-ary CPP records that downstream materials modeling needs.

N-ary and multimodal materials IE. A closer line of work targets N-ary and multimodal extraction. POLYIE annotates variable-length N-ary relations among materials, properties, values, and conditions in polymer literature (Cheung et al., 2024), and a unified framework couples entity recognition with an LLM-guided assembly step to structure N-ary property relations (Phi and Matsumoto, 2025). On the multimodal side, full-length polymer nanocomposite documents have

been mined for sample-level records (Khalighinejad et al., 2024), and MatViX contributes a multimodal benchmark over full-length articles, evaluating zero-shot vision-language models with a curve-similarity metric (Khalighinejad et al., 2025). MatViX is the closest prior work, and MELD departs from it in three ways. First, MELD is a deployable extraction pipeline rather than a benchmark of off-the-shelf models. Second, it treats the processing route as a first-class N-ary axis on which the property value depends; our error analysis shows that the composition-processing binding, not value reading, is the dominant bottleneck. Third, it is scored with a fact-level, recall-oriented metric and is validated by downstream property prediction, the two open directions that MatViX names.

Document understanding and chart parsing. Our figure reading builds on document understanding and chart parsing. Layout-aware models such as LayoutLM learn text and layout jointly (Xu et al., 2020), Pix2Struct pretrains on screenshot parsing for visually situated language (Lee et al., 2023), and DePlot translates a plot into a linearized table for downstream reasoning (Liu et al., 2023). We use layout parsing to recover document structure, but read figures with a vision-language model directly instead of converting plots to tables, because plot-to-table conversion discards the binding between a property value and the composition and processing route it belongs to, which is exactly the relation a CPP record must preserve.

3 Task and Data

3.1 Problem Definition

We frame composition-process-property extraction as filling a structured record for every material sample that a paper reports. Let a paper D describe a set of samples $\mathcal{S} = \{s_1, \ldots, s_n\}$. Each sample is a triple

$$s_j = (c_j,\, p_j,\, y_j), \tag{1}$$

where c_j is a composition, a map from chemical element to weight percent; p_j is a processing route, an ordered set of process steps with their conditions; and y_j is the set of property values measured for that sample. Given the full PDF of D, the task is to recover $\mathcal{S}$.

A property value is interpretable only together with the composition and the processing route that produced it. We therefore take a *fact*, a single (composition, processing, property value) triple, as the unit of evaluation, so a record with k property values contributes k facts. A fact is correct only when its composition, its processing route, and its value are all correct and bound to one another, which makes the target an N-ary relation rather than a set of independent fields.

3.2 Schema

A frozen schema fixes the vocabulary so that records from different papers are comparable and can be used directly as tabular features. The schema shares one backbone across domains and is specialized per alloy system.

The backbone has three axes. Compositions range over 27 alloying elements, expressed as weight percent and balanced by the base metal. A processing route is assembled from step types that each carry a temperature and a duration, with cold rolling also carrying a thickness reduction. Properties are numeric and are matched within property-specific tolerances that reflect reporting precision and figure read-off error rather than exact equality; the tolerance is 10 percent relative for composition and 3 to 5 percent for the property values, and the full schedule is given in Appendix A.

The two alloy systems differ in which properties carry the signal and in how processing is expressed. For copper alloys we score Vickers hardness and electrical conductivity in percent IACS, the two properties present in at least 99 percent of the curated database, and the processing route covers solution treatment, cold rolling, aging, and a secondary thermomechanical step. For precipitation-hardened aluminum alloys conductivity is seldom the reported target, so we score yield strength, ultimate tensile strength, elongation, and hardness, and we extend the processing vocabulary to the steps that govern precipitation, adding pre-aging and pre-deformation alongside solution treatment and aging. The matching algorithm is identical across the two systems; only the scored property set and the processing vocabulary change.

3.3 Gold Datasets

Copper gold is derived from a curated copper-alloy database. We hold out 38 papers and score them curation-free, supplying no per-paper composition or processing priors, which tests the pipeline on papers it was not tuned on. The held-out set contains 2,476 gold facts over the two primary properties. A separate development set of six copper papers, dis-

Set	Papers	Facts	Role
Cu held-out	38	2,476	extraction eval
Al held-out	30	413	domain transfer
Al corpus	231	n/a	downstream train

Table 1: Corpus statistics. The Cu held-out set is scored on hardness and conductivity; the Al held-out set on strength, elongation, and hardness. The 231-paper Al corpus has no gold and supplies the downstream training data, whose labels come from the disjoint Al held-out annotations.

joint from the 38, is used for pipeline development and for the component ablation in Section 6.

Aluminum gold consists of 30 papers annotated by materials-science experts over all scored properties, giving 413 gold facts. The aluminum schema and corpus are independent of the copper database, so this set measures extraction under domain transfer rather than reproduction of a known source.

For the downstream study we run the pipeline on 231 open-access aluminum-alloy papers and use the resulting records as a training set. The downstream evaluation labels are the expert aluminum annotations, which are disjoint from the 231 training papers at the paper level, so no test paper is seen during training. Dataset sizes and roles are summarized in Table 1.

4 The MELD Framework

MELD maps a full PDF to a set of CPP records through a fixed sequence of stages (Figure 1). Composition and processing are recovered from the text, property values are read from figures, and one model call binds them into records that a deterministic tool layer then normalizes. Binding is the only stage where the model assembles a record, which keeps the pipeline simple to run and to ablate.

4.1 Layout Parsing

A layout-aware OCR model converts the PDF into a typed document representation in which each block is labeled as text, table, figure, caption, or section header. This representation exposes the reading order and the figures as croppable regions to the later stages, and it is computed once per paper and cached.

4.2 Composition and Processing from Text

Compositions and processing routes come from the text and tables. We resolve each alloy name or label to a composition by spelling out its elements and balancing the remainder with the base metal, so that a nominal name and an explicit element list map to the same record. Processing steps and their conditions, such as a solution-treatment temperature or an aging time, are read from the surrounding text and attached to the sample they describe.

4.3 Reading Property Values from Figures

A large share of property measurements appears only in figures. For each figure, a vision-language model reads the axis variables and emits the plotted points as value pairs, separating the discrete condition on the horizontal axis from the property value on the vertical axis. We read values directly rather than converting plots to linearized tables, as plot-to-table methods do (Liu et al., 2023), because a value must stay attached to the condition and the sample it belongs to; a table of numbers detached from that binding is not enough to form a fact.

4.4 Single-Shot Binding

A single model call binds the text-derived compositions and processing routes to the figure-derived property values, producing nested records of the form (composition, processing, properties). Binding the whole paper in one pass lets the model use the full context, the text carriers together with the figure series, when it decides which composition and which processing route a value belongs to.

4.5 Deterministic Unit Normalization

Reported values use heterogeneous units: composition in atomic or molar percent, conductivity as resistivity, temperature in Celsius, and time in seconds or days. A deterministic tool layer converts these to the schema units (weight percent, percent IACS, kelvin, and hours) with explicit functions rather than the model. This step is what makes values comparable to the gold. Removing it collapses fact recall to near zero (Section 6), because correctly read values fail to match the gold when their units differ.

5 Experimental Setup

We evaluate on the two held-out sets of Section 3, summarized in Table 1.

5.1 Metric

A fact is a (composition, processing, single property) triple (Section 3). A predicted fact matches a gold fact under the strict metric when three conditions hold together: the composition element sets

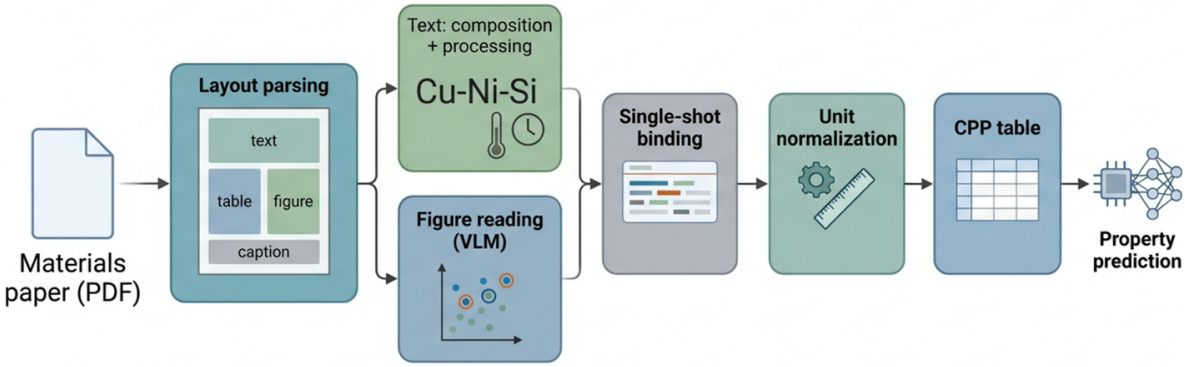

Figure 1: Overall MELD framework: a single-shot multimodal pipeline that parses a full PDF into typed blocks, reads property values directly from figures with a vision-language model, binds them to text-derived compositions and processing routes in one model call, and normalizes units deterministically to produce a machine-learning-ready CPP table.

are identical and every shared weight-percent value is within tolerance, the processing field sets are identical and every value is within tolerance, and the property value is within tolerance; the per-field tolerances are listed in Appendix A. Matching is greedy and one-to-one within a paper. We report staged, micro-averaged precision, recall, and F1, together with the underlying true-positive, false-positive, and false-negative counts at the composition, processing, and fact levels. Component recall is nested, so composition recall is at least processing recall, which is at least fact recall, while property recall is binding-free and isolates value reading. We treat recall as the primary quantity and pair it with the fraction of false positives that are genuine fabrications, because an exact-set metric over a non-exhaustive curated gold penalizes substantively correct output: a near miss on a single field yields both a false positive (the predicted variant) and a false negative (the unmatched gold fact). As a secondary view we report a property-relaxed tolerance of about 10% relative error. The metric is frozen, and we never modify it to move a number.

5.2 Baselines

Our comparison is internal and follows a single-variable principle: we remove exactly one component at a time from the full pipeline, reusing the same cached layout parse and figure reading so that any change is attributable to the removed stage alone. This leave-one-out ablation is reported in Section 6 and Appendix B. We also compare the deployed single-shot design against two alternative pipeline designs that we built and then set aside, and report those comparisons as negative results in Section 8. We do not report an external system as a baseline, because prior multimodal extractors target different schemas and corpora and are not directly runnable on the CPP task we define.

5.3 Implementation

The layout-aware parser of Section 4 is dots.mocr, an open OCR model that produces the typed-block document representation. The vision-language model that reads figures and performs the single-shot binding is a Gemini model (`gemini-3.1-pro-preview`); because its parameter count is not public, we report the exact model version (Gemini Team, Google, 2023). To test robustness to the backbone we also run an open-weights model, Gemma-4-31B-it (about 31B parameters), in its place. Decoding is fixed at temperature 0 with JSON-constrained output and a fixed seed (seed 0). Single-shot binding is non-deterministic even at temperature 0, with run-to-run variation of about ± 0.03 F1, so we report a single seed and leave multi-seed error bars to ongoing work. The Gemini model is accessed through a hosted API, so we report budget as API calls and wall-clock time rather than GPU hours, while the open-weights model is self-hosted. Unit normalization is deterministic (Section 4), and all extraction outputs are cached per paper and seed, so re-scoring does not call the model.

6 Results

6.1 Extraction Performance

Table 2 reports fact-level performance on the two held-out sets under the metric of Section 5. We read recall as the primary quantity, because the

	Component R			Fact		
Set	comp	proc	prop	P	**R**	F1
Cu held-out	0.746	0.626	0.814	0.344	**0.546**	0.422
Al held-out	0.639	0.518	0.765	0.538	**0.508**	0.523

Table 2: Main extraction performance on the held-out sets (strict, micro-averaged). Recall is the primary metric. Property values are read well; the loss is in the composition-processing binding.

goal is to recover as many correct CPP facts as the literature reports. MELD reaches a fact recall of 0.546 on copper and 0.508 on aluminum. The component recalls show a consistent pattern: property values are read accurately (property recall 0.814 for copper and 0.765 for aluminum), while fact recall is lower because a value must also be bound to the correct composition and processing route. The gap between property recall and fact recall, about 0.27 in both domains, places the bottleneck in binding rather than in reading. Processing is the weakest carrier (processing recall 0.626 for copper and 0.518 for aluminum), so the composition-processing join is where recall is lost.

Precision is lower than recall, 0.344 for copper and 0.538 for aluminum, and we analyze it in Section 6.2. The aluminum annotations omit the aluminum balance element on some carriers, which makes the predicted and gold element sets differ; neutralizing this single convention raises the aluminum fact F1 from 0.523 to 0.593.

6.2 Gold-Discordant Outputs

The strict exact-set metric counts every predicted fact that does not match the curated gold as a false positive, yet most of these outputs are not fabrications. Table 3 divides the false positives into three kinds. In both domains only about 8 percent are genuine fabrications, where the predicted composition has no counterpart in the paper. The majority are records whose composition does appear in the paper but is absent from the non-exhaustive curated gold, or that fail the exact-set match for an annotation reason such as an omitted balance element. A smaller share are mis-bindings, where a correct value is attached to a wrong or collapsed composition. Low precision therefore reflects exact-set strictness over a non-exhaustive gold rather than hallucination, and it shows that MELD recovers valid records beyond the curated set. The full taxonomy is given in Appendix D.

Set	**FP**	**valid/gap**	**mis-bind**	**fabricated**
Cu held-out	2,577	75%	17%	8%
Al held-out	180	62%	30%	8%

Table 3: Gold-discordant output analysis. valid/gap: the predicted alloy appears in the paper but the fact is rejected by exact-set matching against a non-exhaustive curated gold, including annotation artifacts such as an omitted balance element; mis-bind: a correct value attached to a wrong or collapsed composition; fabricated: no corresponding alloy in the paper. Genuine fabrication is about 8 percent in both domains.

6.3 Component Ablation

We measure each component by removing it from the full pipeline on the development set and reading the change in fact recall (full-pipeline recall 0.80; the per-component table is in Appendix B). Two components are indispensable. Removing figure reading lowers recall by 0.77, because most property values originate in figures, and removing the deterministic unit normalization lowers recall by 0.80, because correctly read values no longer match the gold once their units differ. Processing normalization is a smaller but real lever, worth 0.08 in recall, and the composition resolver is recall-neutral at 0.002, since it improves composition precision rather than recall and so does not register in a recall-based ablation.

7 Downstream Utility

Extraction metrics measure whether records match a gold annotation, not whether the records are useful. We test usefulness directly with a property-prediction task that is non-circular by construction: the training data and the evaluation data come from disjoint papers and differ in provenance.

7.1 Setup

We train on the records MELD extracts from the 231-paper aluminum corpus (1,581 rows) and evaluate on the 31 expert-annotated papers (223 rows), which share no paper with the training set. Each property is a separate regression problem with composition (weight percent, absent elements set to zero, base computed from the balance) and numeric processing variables (median-imputed) as features. We compare seven standard tabular regressors: random forests (Breiman, 2001), XGBoost (Chen and Guestrin, 2016), LightGBM (Ke et al., 2017), CatBoost (Prokhorenkova et al., 2018), and elastic net, support vector regression, and a multilayer percep-

tron from scikit-learn (Pedregosa et al., 2011). We report R^2, MAE, RMSE, and MAPE.

Because the extracted training values contain reading errors, we compare three filters on the training target: none (raw), Tukey 1.5 IQR removal (data-driven), and clipping to a priori aluminum physical ranges taken from standard-temper values (Davis, 1993). The physical ranges are set without reference to the test labels and are wide enough to keep every real datum, removing only unit-confusion artifacts such as Brinell or Rockwell numbers misread as Vickers hardness. As a small-data reference we also train on the expert gold set alone with 5-fold cross-validation.

7.2 Results

Table 4 reports, for each property, the test R^2 of the human-curated baseline (R^2_{human}) and of the best regressor trained on the extracted data (R^2_{MELD}). On yield strength, tensile strength, and elongation the MELD-trained model clearly exceeds the small human-curated baseline; yield-strength R^2, for example, rises from 0.58 to 0.87. The gain holds under a like-for-like comparison with the same model, so it does not come from model selection. Scale compensates for extraction noise: a larger, noisier literature-mined dataset carries more usable signal than a small clean one.

Hardness is the exception. It is the noisiest target to extract, because tables and figures report Brinell and Rockwell numbers that are easily mistaken for Vickers values, and even after physical filtering the extracted model (R^2 0.61) stays just below the human-curated baseline (0.66). This is consistent with the extraction error analysis, where hardness reading is the least reliable component.

Two patterns hold across properties. Gradient-boosted trees (CatBoost, XGBoost, LightGBM) outperform the linear and kernel models by a wide margin, and CatBoost is most often best. The physical filter is more effective than statistical outlier removal, and the difference is largest where unit confusion is worst: raw hardness training yields a negative R^2, since the contaminated values dominate the trees, while clipping to physical ranges restores it above 0.5. Domain knowledge about plausible ranges is a more reliable cleaning signal than a distribution-based rule. The full model and filter panel is in Appendix C.

Property	R^2_{human}	R^2_{MELD}
Yield strength	0.58	**0.87**
Tensile strength	0.45	**0.72**
Elongation	0.31	**0.57**
Hardness (HV)	0.66	0.61

Table 4: Downstream property prediction (test R^2). R^2_{human} is a regressor trained on the expert human-curated gold under 5-fold cross-validation; R^2_{MELD} is the best regressor trained on the 231-paper MELD-extracted set and evaluated on the held-out expert gold, which shares no paper with the training set. The MELD-trained model exceeds the human-curated baseline on the first three properties; hardness is the exception. Per-property error metrics (MAE, RMSE, MAPE) are in Appendix C.

8 Discussion

The accuracy of MELD comes from a deliberate division of labor: the model reads and binds in a single pass, and deterministic code normalizes the result. We tested this choice by building the two extensions that prior work names as open directions (Khalighinejad et al., 2025), a multi-call agentic extractor and an independent verifier that grounds each record against the paper. Neither improves on the single-shot pipeline. Both extensions act on the same object, the binding of a composition, a processing route, and a property value into one fact, and both show why that binding resists further automation in this setting.

8.1 A Multi-Call Agent Does Not Beat a Single Pass

We built a variant that replaces the single binding call with a sequence of model calls. Separate calls build and verify the composition list, the processing route, and the bindings in turn, and each step may correct the previous output but may not delete it. The variant does not overtake the single-shot pipeline. Its fact recall is 0.50 on the development set and 0.39 on the held-out copper papers, against 0.80 and 0.55 for the single-shot pipeline; fact F1 shows the same gap (0.42 and 0.36 against 0.67 and 0.42). Two effects explain the loss. Splitting the binding across calls makes the output non-deterministic even at temperature zero, so the score varies between runs on the papers that contribute the most facts. The single binding call also standardizes processing-step names against the schema vocabulary while it has the full context in view, whereas steps produced in a separate call are named

in the words of the paper and are then dropped by the exact-set matcher. Added calls introduce variance and forfeit a standardization that the single call performs at no cost, without improving the binding that the single call already makes.

8.2 Evidence Grounding Works Against Figure-Derived Records

We also built an independent verifier that runs after extraction, grounds each predicted record in the paper, and drops the records it cannot support, with a separate check for physical plausibility. On a copper paper whose properties are reported only in figures, the verifier removed almost all correct records and cut fact recall from 0.85 to 0.02, while precision did not rise. The cause is structural and not a matter of tuning. The verifier asks whether the paper states the (composition, processing, value) tuple as a single unit. A record read from one table row passes this test, but a record assembled from a composition and a fixed process in the text and a value read from a scatter plot does not: the paper never writes that join in one place, so a correct inference is treated as unsupported and rejected. Supplying the figure series to the verifier did not recover the records, because what the paper omits is the binding, not the value. Evidence grounding fits table-based extraction and works against a figure-derived N-ary target, so MELD leaves it out.

The two results bound the same lesson. In a figure-heavy alloy CPP setting, accuracy follows from binding once with full context in view and then normalizing deterministically, rather than from adding model-driven structure or model-driven judgment on top of that step. The downstream validation that the same prior work names as its other open direction is the experiment we report in Section 7.

9 Conclusion

We presented MELD, a single-shot multimodal pipeline that builds composition-process-property datasets from full-length materials-science papers. MELD parses a PDF into layout-aware blocks, reads property values from figures with a vision-language model, binds them to text-derived compositions and processing routes in one model call, and normalizes units with a deterministic tool layer. On held-out, expert-annotated gold in copper and aluminum it reads property values accurately, with the composition-processing binding as the bottleneck and genuine fabrication at about 8 percent of false positives. In a non-circular downstream test, property predictors trained on 231 papers of automatically extracted aluminum data match or exceed those trained on a small expert-curated set on three of four properties, which shows that literature-scale automatic extraction can feed materials-informatics workflows in spite of its reading noise. We release the extracted aluminum CPP dataset so that this larger, noisier training source can be reused.

Limitations

We evaluate on two alloy systems, copper and aluminum. The schema backbone is shared, but the scored properties and the processing vocabulary are set per system, so generalization to other materials classes is untested.

Extraction is reported from a single run at temperature zero. Because the binding is one model call, its output is not fully deterministic, and per-paper scores carry run-to-run variance that we do not average over; the multi-seed behavior we observe for the agent variant in Section 8 applies to the single-shot binding as well.

The primary metric is a strict exact-set match against a curated gold that is not exhaustive, which undercounts precision. As Section 6.2 shows, most false positives are records that are valid in the paper or rejected for an annotation reason rather than fabrications, so the reported F1 is a lower bound on usable accuracy rather than a direct estimate of it.

The downstream study is a feasibility test of whether the extracted data supports tabular property prediction. We do not isolate the contribution of the processing axis to prediction accuracy, so the paper does not claim that processing features improve prediction; the case for extracting processing rests on its role in the extraction task and on prior materials-science evidence, not on a prediction ablation.

Finally, figure reading depends on a proprietary vision-language model, which affects the cost and the exact reproducibility of the figure-derived values.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th*

International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3615–3620.

Leo Breiman. 2001. Random forests. *Machine Learning*, 45(1):5–32.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 785–794.

Jerry Cheung, Yuchen Zhuang, Yinghao Li, Pranav Shetty, Wantian Zhao, Sanjeev Grampurohit, Rampi Ramprasad, and Chao Zhang. 2024. POLYIE: A dataset of information extraction from polymer material scientific literature. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2370–2385.

John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S. Rosen, Gerbrand Ceder, Kristin A. Persson, and Anubhav Jain. 2024. Structured information extraction from scientific text with large language models. *Nature Communications*, 15(1):1418.

Joseph R. Davis, editor. 1993. *ASM Specialty Handbook: Aluminum and Aluminum Alloys*. ASM International, Materials Park, OH.

Alexander Dunn, Qi Wang, Alex Ganose, Daniel Dopp, and Anubhav Jain. 2020. Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm. *npj Computational Materials*, 6(1):138.

Gemini Team, Google. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Tanishq Gupta, Mohd Zaki, N. M. Anoop Krishnan, and Mausam. 2022. MatSciBERT: A materials domain language model for text mining and information extraction. *npj Computational Materials*, 8(1):102.

Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. 2013. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*.

Ghazal Khalighinejad, Defne Circi, L. Brinson, and Bhuwan Dhingra. 2024. Extracting polymer nanocomposite samples from full-length documents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13163–13175.

Ghazal Khalighinejad, Sharon Scott, Ollie Liu, Kelly L. Anderson, Rickard Stureborg, Aman Tyagi, and Bhuwan Dhingra. 2025. MatViX: Multimodal information extraction from visually rich articles. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3636–3655.

Olga Kononova, Haoyan Huo, Tanjin He, Ziqin Rong, Tiago Botari, Wenhao Sun, Vahe Tshitoyan, and Gerbrand Ceder. 2019. Text-mined dataset of inorganic materials synthesis recipes. *Scientific Data*, 6(1):203.

Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. 2023. Pix2Struct: Screenshot parsing as pretraining for visual language understanding. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pages 18893–18912.

Fangyu Liu, Julian Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhu Chen, Nigel Collier, and Yasemin Altun. 2023. DePlot: One-shot visual language reasoning by plot-to-table translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10381–10399.

Elsa A. Olivetti, Jacqueline M. Cole, Edward Kim, Olga Kononova, Gerbrand Ceder, Thomas Yong-Jin Han, and Anna M. Hiszpanski. 2020. Data-driven materials research enabled by natural language processing and information extraction. *Applied Physics Reviews*, 7(4):041317.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Van-Thuy Phi and Yuji Matsumoto. 2025. A unified framework for N-ary property information extraction in materials science. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 2369–2388.

Maciej P. Polak and Dane Morgan. 2024. Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nature Communications*, 15(1):1569.

Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2018. CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*.

Axis	Field	Tolerance
Composition	element wt%, base-balanced	10% (floor 0.05)
Processing	step temperature (K)	5 K
Processing	step duration (h)	10%
Processing	cold-rolling reduction (%)	10%
Property	hardness (HV)	5%
Property	conductivity (% IACS)	3%
Property	yield / tensile strength (MPa)	5%
Property	elongation (%, Al)	5% (floor 0.5)

Table 5: Schema axes and the per-field matching tolerance.

Pranav Shetty, Arunkumar Chitteth Rajan, Chris Kuenneth, Sonakshi Gupta, Lakshmi Prerana Panchumarti, Lauren Holm, Chao Zhang, and Rampi Ramprasad. 2023. A general-purpose material property data extraction pipeline from large polymer corpora using natural language processing. *npj Computational Materials*, 9(1):52.

Matthew C. Swain and Jacqueline M. Cole. 2016. ChemDataExtractor: A toolkit for automated extraction of chemical information from the scientific literature. *Journal of Chemical Information and Modeling*, 56(10):1894–1904.

Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A. Persson, Gerbrand Ceder, and Anubhav Jain. 2019. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571(7763):95–98.

Leigh Weston, Vahe Tshitoyan, John Dagdelen, Olga Kononova, Amalie Trewartha, Kristin A. Persson, Gerbrand Ceder, and Anubhav Jain. 2019. Named entity recognition and normalization applied to large-scale information extraction from the materials science literature. *Journal of Chemical Information and Modeling*, 59(9):3692–3702.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. LayoutLM: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1192–1200.

A Schema Tolerances

Table 5 lists the per-field matching tolerance for every schema axis. Relative tolerances apply to the gold value and a floor is an absolute minimum. Copper is scored on hardness and conductivity; aluminum is scored on the strength, elongation, and hardness rows.

B Component Ablation

Table 6 gives the leave-one-out component ablation on the six-paper development set at seed 0. Each row removes one component from the full pipeline and reports processing recall, fact recall, and the change in fact recall relative to the full pipeline.

Configuration	proc R	fact R	ΔR
Full pipeline	0.851	0.802	(ref.)
− resolver	0.849	0.800	−0.002
− degenerate-CR	0.827	0.779	−0.023
− process norm.	0.757	0.721	−0.081
− figure reading	0.046	0.032	−0.770
− unit normalization	0.000	0.000	−0.802

Table 6: Leave-one-out component ablation on the six-paper development set (seed 0). Removing figure reading or unit normalization collapses fact recall.

Model	Filter	YS	UTS	EL	HV
RF	raw	0.77	0.67	0.48	−2.88
	iqr	0.73	0.67	0.26	0.54
	phys	0.75	0.67	0.55	0.55
XGBoost	raw	0.80	0.72	0.52	0.04
	iqr	0.80	0.71	0.27	0.55
	phys	0.83	0.72	0.51	0.56
LightGBM	raw	0.82	0.68	0.33	−0.28
	iqr	0.82	0.68	0.25	0.49
	phys	0.80	0.67	0.57	0.52
CatBoost	raw	0.85	0.71	0.49	0.50
	iqr	0.87	0.71	0.27	0.56
	phys	0.86	0.70	0.56	0.61

Table 7: Test R^2 on the 31-paper gold set for tree models trained on the extracted data under three training-value filters (raw; Tukey 1.5 IQR; physical bounds). Columns are yield strength, tensile strength, elongation, and hardness. Raw hardness collapses to a negative R^2 from unit-confusion outliers, and physical filtering restores it.

C Downstream Model and Filter Panel

Table 7 reports test R^2 for the four tree-based regressors under the three training-value filters. Elastic net, support vector regression, and the multilayer perceptron were also evaluated and underperformed the tree models throughout (R^2 below 0.5 on most targets, and negative for the perceptron on tensile strength and hardness); their full metrics are in the released logs.

Table 8 lists the per-property error metrics of the best extracted (MELD-trained) model, the model reported as R^2_{MELD} in Table 4.

Property	R^2	MAE	RMSE	MAPE
Yield strength	0.87	32.8	49.0	20.4
Tensile strength	0.72	47.6	80.6	32.9
Elongation	0.57	4.8	10.3	32.8
Hardness (HV)	0.61	19.1	27.7	29.8

Table 8: Error metrics of the best extracted (MELD-trained) model per property. Units: strength MPa, hardness HV, elongation percent; MAPE percent.

D Error Analysis

We adjudicate every error the system makes on the two held-out sets in order to locate where schema-conditioned CPP extraction breaks down. All examples are verbatim facts taken from the system output and the gold annotation, and none is synthetic.

D.1 Protocol

We evaluate on the two held-out sets of Section 3, namely 38 copper-alloy papers scored curation-free on the primary properties (HV and IACS) and 30 aluminum-alloy papers with hand annotations over all properties, using the frozen fact-level strict metric of Section 5 with greedy one-to-one matching within a paper. We assign every false negative and every false positive to a single error type with a deterministic rule set. The main thresholds are: a property value within 10% relative error is a tolerance artifact and one off by at least 50% is a gross misread; a composition difference confined to elements below 0.1 wt% is a trace mismatch; a near-null or room-temperature processing step is trivial; and a composition whose declared fractions sum far from 100 wt% is an arithmetic artifact. The rules are heuristics, released with the code, and the four artifact claims that anchor the aluminum analysis were verified against the source gold. We group both domains under one scheme of four families: reading errors (the carrier is correct but the value is wrong), carrier and binding errors (the value is present but attached to the wrong carrier), recall failures (the fact is not produced at all), and metric or annotation artifacts (the output is substantively correct but fails the exact-set match).

D.2 Headline

On Cu-38 the system reaches fact recall 0.546, precision 0.344, and F1 0.422 over 2,476 gold and 3,929 predicted facts (TP 1,352; FP 2,577; FN 1,124), with component recall 0.746 for composition, 0.626 for processing, and 0.814 for property.

False-negative type (Cu-38)	n	%
Reading errors	**328**	**29.2**
property gross misread ($\geq$50%)	55	4.9
property point-selection (10 to 50%)	73	6.5
processing value or unit	200	17.8
Carrier / binding errors	**320**	**28.5**
composition misread or mis-bound	226	20.1
processing incomplete	90	8.0
processing mismatch	4	0.4
Recall failures	**271**	**24.1**
not recovered	216	19.2
property missing	55	4.9
Metric / annotation artifacts	**205**	**18.2**
property tolerance ($\leq$10%)	93	8.3
trace-element set mismatch	70	6.2
greedy contention	42	3.7
Total	1,124	100.0

Table 9: Copper-alloy (Cu-38) false-negative taxonomy, adjudicated under the four shared families.

On Al-30 it reaches fact recall 0.508, precision 0.538, and F1 0.523 over 413 gold facts (TP 210; FP 180; FN 203), with component recall 0.639 for composition, 0.518 for processing, and 0.765 for property. In both domains property recall is the highest component and processing recall the lowest, so values are read well while the composition-processing binding is the bottleneck: the binding gap (property recall minus fact recall) is 0.268 for copper and 0.257 for aluminum.

D.3 False-negative taxonomy

Tables 9 and 10 adjudicate the recall losses under the four shared families. Genuine system errors (the first three families) account for 81.8% of copper false negatives and 65.5% of aluminum false negatives, and the remainder are metric or annotation artifacts. The largest copper family is reading errors (29.2%), dominated by processing values written in the wrong unit (200 of 1,124), followed by carrier and binding errors (28.5%). The largest aluminum family is metric or annotation artifacts (34.5%), dominated by the aluminum-balance omission described below, followed by recall failures (33.0%).

D.4 False-positive taxonomy

Table 11 adjudicates the precision losses, the detailed form of the main-text summary in Table 3. In both domains genuine fabrication, a predicted fact with no corresponding gold composition, is only about 8% of false positives (206 of 2,577 for copper, 15 of 180 for aluminum), roughly 4 to 5%

False-negative type (Al-30)	n	%
Reading errors	**5**	**2.5**
Carrier / binding errors	**61**	**30.0**
composition mis-bound	22	10.8
processing incomplete	32	15.8
processing mismatch	7	3.4
Recall failures	**67**	**33.0**
not recovered	63	31.0
property missing	4	2.0
Metric / annotation artifacts	**70**	**34.5**
aluminum-balance omission	38	18.7
composition sum invalid	9	4.4
trivial processing step	8	3.9
greedy contention	8	3.9
trace-element set mismatch	7	3.4
Total	203	100.0

Table 10: Aluminum-alloy (Al-30) false-negative taxonomy under the same four families. The single reading error is a processing value off tolerance.

False-positive category	n	%
Cu-38		
over-generation (real alloy)	1,932	75.0
mis-bound (misread composition)	439	17.0
fabrication or gold-incomplete	206	8.0
Al-30		
over-generation (real alloy)	59	32.8
mis-bound (wrong alloy)	20	11.1
degenerate composition (pure base)	34	18.9
balance-omission artifact	43	23.9
composition-sum artifact	9	5.0
fabrication (no gold)	15	8.3

Table 11: False-positive taxonomy for both held-out sets. Cu-38 totals 2,577 and Al-30 totals 180.

of all predictions. The bulk of precision loss is over-generation on alloys that genuinely appear in the paper but whose carrier set fails the exact match against a non-exhaustive curated gold (75.0% for copper), together with composition errors that mirror the false-negative side: a value carried on a misread composition, or a collapse to the pure base element. The low precision therefore reflects metric strictness and gold curation far more than fabrication.

D.5 Representative examples

We write a fact as {composition wt%} | {processing} | property and show gold → prediction, with temperatures in K, times in h, and the discrepancy in bold.

- *Reading, processing unit* (`jinstmet1952.62.4_301`): `{Cu 97.94, Ni 1.39, Si 0.67}`, gold `sol 1223·`**1**, `age 773·`**0.083** → pred `sol 1223·`**3.6**, `age 773·`**0.00028**. Times are reported in kiloseconds (1 h is 3.6 ks); the model writes the raw kilosecond value into the hour field, a unit outside the range handled by the deterministic converter.
- *Reading, figure scale* (`10.1016/j.matlet.2017.04.039`): `{Cu 99.84, Cr 0.10, Zr 0.06}` | `sol 1273·1` | HV = **71** → **0.70**. Hardness is read off the y axis at roughly one-hundredth scale, systematically across the paper.
- *Carrier and binding, composition* (`10.1016/j.jallcom.2018.09.296`; HV 300 read correctly): gold `{Cu 92.57, Be 1.66,` **Mg 2, Ti 3**`, Ni 0.33, Si 0.31, Al 0.13}` → pred `{Cu` **97.57**`, Be 1.66, Ni 0.33, Si 0.31, Al 0.13}`. Dropping Mg and Ti inflates the copper balance and binds a correct hardness to a wrong composition.
- *Carrier and binding, collapse to pure base* (Al `ref0201`): gold `{Al 91.75, Mg 1.85, Zn 6.4}` carries hardness 66 and 140, yet the model emits `{Al` **100**`}` for the same values. The gold composition is correct here, so this is a true composition error rather than an annotation artifact.
- *Recall failure, series omitted* (Al `ref0201`): the gold carrier `{Al 91.75, Mg 1.85, Zn 6.4}` with a natural-aging hardness series `{80, 95, 115}` has **no prediction**.
- *Artifact, balance omission* (Al `ref3010`, AA7075): the hand annotation lists `{Cr, Cu, Fe, Mg, Mn, Ti, Zn}` summing to 10.13 wt% with **no Al**, while the model supplies `Al 89.87`. The aluminum balance is blank for every row of this paper, so the prediction is correct and the gold is incomplete; this single paper produces 36 false negatives with no true positive.
- *Artifact, greedy contention* (`10.1016/j.jallcom.2014.07.010`): two gold rows with IACS **78** and **79** both fall within tolerance of one prediction **79.1**; one-to-one matching consumes one and leaves the other a false negative.
- *Fabrication* (`10.1016/j.matdes.2021.109929`): pred `{`**Co 0.5**`, Cu 96.3, Ni 2.5, Si`

	Cu-38	Al-30
fact F1 (strict)	0.422	0.523
binding gap (prop R − fact R)	0.268	0.257
processing recall	0.626	0.518
FN artifact rate	18.2%	34.5%
gold provenance	curated DB	hand annotation

Table 12: Cross-domain contrast. Both sets share the binding and processing bottleneck; the aluminum set carries a higher annotation-artifact rate.

`0.7} | age 723·2, CR 70%, sol 1223·3 |` IACS 46.7 has no gold counterpart, either a fabrication or a real composition absent from the curated gold.

D.6 Cross-domain contrast

Table 12 compares the two domains under the same taxonomy. They share the same real failure signature, namely values read well while binding and processing limit recall, but they differ sharply in artifact rate. The aluminum false-negative artifact rate (34.5%) is roughly twice the copper rate (18.2%), and the difference is an annotation-provenance effect: the hand-annotated aluminum gold often omits the aluminum balance and sometimes encodes nominal spec ranges or arithmetically invalid compositions, whereas the curated copper database gold does not. Neutralizing the aluminum-balance convention alone, by dropping the aluminum field on both sides in a diagnostic rescoring that leaves the reported metric untouched, lifts aluminum F1 from 0.523 to 0.593, a relative gain of 13%.

D.7 Takeaways

The bottleneck is composition-process-property binding and processing normalization, not value reading: property recall is 0.814 (copper) and 0.765 (aluminum) while joint fact recall is 0.546 and 0.508. A large share of recall and precision loss is metric or annotation artifact rather than model failure (18.2% of copper false negatives and 34.5% of aluminum false negatives), and genuine fabrication is about 8% of false positives in both domains. Errors are concentrated: the eight worst copper papers account for 72% of copper false negatives, and the top six aluminum papers for about 60%. Reporting the exact-set F1 alongside a relaxed value-level score and a fabrication rate therefore gives a truer picture of extraction quality than the strict F1 alone.