LEARNING MATERIAL DECISIONS FROM HETEROGENEOUS EVIDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Materials decisions combine incomplete composition ranges, property tables, response curves, processing histories, and text. We propose a query-conditioned framework that retains each observation's value type, role, conditions, and provenance, learns to reconstruct whole withheld evidence views, and applies typed constraints to the completed responses. Existing corpus observations provide supervision without new manual labels for every task. The implemented interface composes trained tree-based, neural, and donor specialists for property completion, joint requirement selection, and response-to-region inference; a fully shared heterogeneous encoder remains an extension. The organizing principle is decision sufficiency: a unique material state is unnecessary when compatible states support the same action. The decision layer preserves joint constraints, nonzero composition or process ranges, and unresolved applicability, distinguishing source-supported conclusions from model-conditional recommendations without autoregressive text generation. Evaluations show that complementary evidence improves completion against strong context-only controls and that learned functional responses support nonzero operating windows. Useful coverage and requirement violations expose tradeoffs that regression error alone misses. Matched controls separate these gains from general parameter transfer and show that joint-decision training does not improve every seed. The result connects fragmented materials knowledge to evaluated decisions over ranges rather than isolated property predictions.

1 INTRODUCTION

Materials knowledge is distributed across documents and data structures with incompatible levels of detail. A composition is often a range; a property can be a measured value, a typical value, a specification minimum, or an entire response curve. Processing histories and surrounding text specify when those observations apply. An engineer combines these fragments to select a related material, alter processing, or determine whether a composition window can meet several requirements. Learning to support these actions requires more than predicting one scalar from a complete fixed-width table.

We propose a query-conditioned framework for learning material decisions from heterogeneous evidence. The input is a variable collection of observations, each carrying a value representation, role, conditions, and provenance. The learning objective hides an entire evidence view and reconstructs it from the remaining collection. Explicit queries specify property, functional, or decision tasks. The implemented interface composes trained specialists under a common evidence/query contract; a fully shared heterogeneous encoder remains an extension. Existing source observations supply supervision without new manual labels for every task.

The central design choice is to couple evidence completion to *decision sufficiency*. Evidence may leave several physical or source-level explanations possible while still determining a useful action. Conversely, an accurate point prediction may not determine whether a whole composition range meets a requirement. We therefore distinguish prediction of missing evidence from a decision layer that combines the predicted response with applicable physical relations, specification constraints, and unresolved conditions. The output can be a conditional admissible region, a rejected candidate

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

under stated assumptions, or a request for evidence that distinguishes remaining actions. It is not necessary to generate an explanatory text sequence in order to produce these structured outputs.

This distinction is particularly relevant to materials manufacturing. Exact composition points suppress the tolerances that an industrial process needs. Independent element intervals can also lose valid compensating relationships. A useful answer should preserve a nonzero region and the conditions under which it is supported. Moreover, asymmetric requirements matter: some properties must remain within a band, while a dangerous shortfall in another property must be avoided. Averaging scalar prediction errors does not capture these differences.

The proposed framework is compatible with both tree-based and neural predictors. Strong tabular models are essential comparators, given their established empirical performance (Grinsztajn et al., 2022; McElfresh et al., 2023). An apparent representation advantage can otherwise be caused by extra information, different missingness patterns, or an unnecessarily weak numeric baseline. Our evaluation accordingly separates the utility of additional evidence from the utility of common learned parameters and from the quality of the resulting decision.

Our contributions are:

1. A typed evidence/query method coupling whole-view learning to joint requirements and admissible regions, with source-supported and predicted answers represented separately in one execution interface.
2. A working prediction-to-window route: a fully hidden functional response is reconstructed, inverted, and eroded by a prescribed input variation before the proposed region is checked against the withheld source function.
3. Controlled evaluations of complementary evidence, simultaneous missing properties, applicability, and useful versus violating window width. They identify both gains and cases where a simpler specialist remains preferable.

The scientific question is whether learning makes partial knowledge more useful for decisions, separating evidence utility, parameter transfer, and decision quality.

2 Related Work

Strong tabular controls are essential: Grinsztajn et al. (2022) examine why trees outperform deep networks on typical medium-sized datasets, while McElfresh et al. (2023) show how rankings depend on data and model selection. We therefore compare with LightGBM (Ke et al., 2017) under matched inputs, missingness information, and groups. TabM (Gorishniy et al., 2025) adds a relevant neural approach through parameter-efficient MLP ensembling; it is not evaluated here.

MaterialsBERT (Shetty et al., 2023) extracts material–property records from polymer literature; its discussion identifies processing and measurement conditions as gaps in the extraction ontology. LLM-Prop (Niyongabo Rubungo et al., 2025) predicts properties from textual descriptions of crystalline structures. These tasks motivate text representations but differ from interpreting handbook conditions and normative intervals. Domain pretraining, supervised adaptation, and frozen-feature utility must therefore be distinguished; success in one does not remove the need for downstream labels.

MODNet (De Breuck et al., 2021) combines physically motivated feature selection and joint learning on small materials datasets, including temperature-dependent properties. MultiMat (Moro et al., 2025) uses self-supervised multimodal training for property prediction and latent-space retrieval. Multi-CrabNet (Maier et al., 2026) learns sparse glass properties with task embeddings and a shared decoder, conditioning on composition, task, and temperature for viscosity. Observed values of other properties are not inputs to that prediction interface, whereas they are central to our completion queries. Learning from related properties, response functions, or multiple modalities is thus established prior work.

Deep Sets (Zaheer et al., 2017) provides permutation-invariant collection processing, and Conditional Neural Processes (Garnelo et al., 2018) predict queried outputs from observed input–output sets. ACE (Strauss & Oliva, 2021) estimates arbitrary conditional distributions; VAEAC (Ivanov et al., 2018) generates missing numerical or categorical features from arbitrary observed subsets.

Our contribution is an application-level learning and evaluation contract: observation roles, units, and operators distinguish bounds, intervals, and curves; masking removes whole target views and their known derivatives; typed requirements turn completed evidence into decisions and nonzero regions. Neither set processing nor masking property subsets is claimed as new.

Elmachtoub & Grigas (2022) evaluate prediction through downstream optimization loss. Our threshold and interval decisions likewise require separate error measures, but do not implement their loss or inherit its guarantees. EDDI (Ma et al., 2018) combines a partial VAE with expected information gain for feature acquisition; our one-observation study instead uses empirical requirement-decision loss. These related methods motivate matched future controls. The present experiments do not establish superiority over their implementations or an advantage for a fully shared heterogeneous encoder.

3 QUERY-CONDITIONED LEARNING FROM MATERIALS EVIDENCE

We propose a framework that learns to complete missing materials evidence and uses the resulting responses in explicitly scoped engineering queries. Its training unit is a linked collection of observations with a hidden view, rather than a row assumed to contain interchangeable measurements. Existing catalogue values and functions supply supervision without requiring new manual labels for every downstream task.

Evidence and queries. An evidence item is $e = (p, o, u, c, r, v, s)$: property p, observation operator o, units u, conditions c, evidential role r, content v, and provenance s. Content includes a recorded scalar, interval, lower bound, sampled curve, or textual condition. Operators distinguish, for example, proof-stress definitions; conditions encode processing, geometry, temperature, and test environment. Roles distinguish normative, measured, derived, or unspecified assertions. A recorded scalar denotes a source assertion, not necessarily an exact physical measurement. Units are converted only when the transformation is supported; gauge lengths, material states, and heat/product analysis remain distinct. Unknown fields remain explicit.

A query $q = (p, o, c, r, \mathcal{T})$ requests a response under a specified definition, role, and conditions, optionally on coordinates $\mathcal{T}$. Let E_i contain permitted evidence linked to record i. The proposed shared representation and decoder are

$$z_i = \mathrm{Enc}_\theta(E_i \setminus M_i), \qquad \widehat{v}_{iq} = \mathrm{Dec}_\theta(z_i, q), \tag{1}$$

where M_i masks the entire queried view and identified derivatives or copies. The encoder respects set membership rather than arbitrary item order; fixed typed slots are one restricted implementation. Source identifiers support grouping and provenance rather than serving as unrestricted predictive shortcuts. Shared grade names do not establish common heats or specimens.

Whole-view learning. Training examples are constructed by hiding available views. Both endpoints of an interval, or a whole sampled response, are removed together with target-revealing notes and availability. Only supported target coordinates contribute to loss; missing answers are not fabricated. For nonempty observed coordinate sets J_{iq} of the hidden view, define

$$\ell_{\mathrm{view}}(i, q) = \frac{1}{|J_{iq}|} \sum_{j \in J_{iq}} \left(\frac{\widehat{v}_{iqj} - v_{iqj}}{s_q} \right)^2, \tag{2}$$

using a train-only scale $s_q > 0$ for each response type. An interval contributes its endpoints; a normative minimum contributes its recorded bound, not an invented specimen value. A curve contributes its supported queries. This operator-conditioned reconstruction allows partial records to participate without homogenizing their physical meaning.

The proposed integrated objective combines reconstruction with functional and query consistency:

$$\mathcal{L} = \operatorname*{Avg}_{g,i,q} \left[\ell_{\mathrm{view}} + \lambda_s \ell_{\mathrm{shape}} + \lambda_d \ell_{\mathrm{decision}} + \lambda_c \ell_{\mathrm{constraint}} \right]. \tag{3}$$

The weights satisfy $\lambda_s, \lambda_d, \lambda_c \geq 0$. All averages omit empty sets. The average balances declared dependence groups, records, and available hidden views. Shape loss compares centered predicted and reference profiles on the same supported coordinates, preventing level accuracy from standing

in for response-shape learning. Decision loss is a weighted binary loss for predefined queries whose answers follow from available reference witnesses; false acceptance and false rejection can receive different fixed costs. Constraint loss penalizes violations of supported relations, such as endpoint ordering, only where their applicability is known. These auxiliary terms are specified in Appendix A.1; their weights must be fixed using training data or a declared validation protocol. They do not turn normative answers into physical failure labels.

Engineering decision layer. The decoder's predicted response is distinct from a source-supported conclusion. For applicable evidence E, let $\mathcal{A}(E)$ denote the retained set of compatible interpretations or admissible states. A requirement R is supported under that contract only when

$$\mathcal{A}(E) \neq \varnothing \quad \text{and} \quad R(a) \text{ holds for every } a \in \mathcal{A}(E). \tag{4}$$

Missing applicability conditions produce an unresolved answer; an empty set signals inconsistent premises rather than vacuous support. A counterexample to universal support does not imply that every real specimen fails. Predicted views may rank candidates or define a model-conditional region, but their use is labeled as predictive evidence. Agreement among predictions does not establish coverage of the true response.

For composition and process selection, the layer preserves joint requirements and dependent ranges rather than optimizing a single nominal point. Under declared linear constraints $Ax \leq b$, a nonempty box $B(c, h) = \{x : |x_j - c_j| \leq h_j\}$, $h_j \geq 0$, satisfies a row throughout exactly when

$$a^\top c + |a|^\top h \leq b_a. \tag{5}$$

This elementary support calculation permits optimization with fixed nonzero widths. Physical or normative constraints retain their source assumptions; a chemically admissible window acquires mechanical properties only through an additional supported relation. Direct catalogue lookup and exact constraint evaluation remain strong controls for the decision layer.

Response-to-region operator. For a predicted scalar response $\widehat{f}(t)$, a lower requirement r, a reserve $\rho \geq 0$, and input variation $\delta \geq 0$, form

$$S_\rho = \{t : \widehat{f}(t) \geq r + \rho\}, \qquad W_{\rho,\delta} = S_\rho \ominus [-\delta, \delta]. \tag{6}$$

Here erosion retains centers whose entire perturbation interval lies in S_ρ. For a piecewise-linear response, threshold crossings and interval erosion are computed exactly on the stored numeric representation. Disconnected components remain separate, and a query may require a minimum component width. If $f(t) \geq \widehat{f}(t) - \rho$ throughout the relevant intervals, every retained center satisfies the requirement under every allowed perturbation. Wolfram and Lean verify this sufficient implication; the empirical reserve is not assumed to satisfy that error bound. Evaluation therefore measures both useful and violating proposed width against the withheld reference response.

Implemented learning and execution. One interface implements property completion, joint requirement selection, composition constraints, source-curve processing windows, and learned-response operating windows. Adapters preserve the observation and query semantics while using boosting, a shared neural operator, or a functional donor as appropriate. For simultaneous missing properties, all queried targets are removed from every conditional input, preventing one query from revealing another's answer. A matched neural campaign implements reconstruction plus joint-decision loss; the full objective in Equation (3) is an extension, not one loss trained across all heterogeneous inputs in these experiments. The functional-window adapter uses stored out-of-group predictions for indexed corpus records. It implements the complete evaluated route without claiming arbitrary-document ingestion or prediction for any unseen composition.

Evaluation contract. Known source, alias, and copying relations are closed before grouped splits; preprocessing is fitted on permitted training data only. Targets retain their recorded meaning: paired catalogue fields need not form a physical uncertainty interval. Missing targets contribute no loss; missing inputs retain their masks. For evaluated groups $\mathcal{G}$, records $\mathcal{I}_g$, and supported queries $\mathcal{Q}_i$, reported error is

$$\mathrm{MAE}_{\mathrm{group}} = \frac{1}{|\mathcal{G}|} \sum_g \frac{1}{|\mathcal{I}_g|} \sum_i \frac{1}{|\mathcal{Q}_i|} \sum_q |\widehat{y}_{iq} - y_{iq}|. \tag{7}$$

Groups encode identifiable dependencies, not common heats or independent physical experiments. Cohorts and previously examined splits remain exploratory; counts are not pooled across studies. We

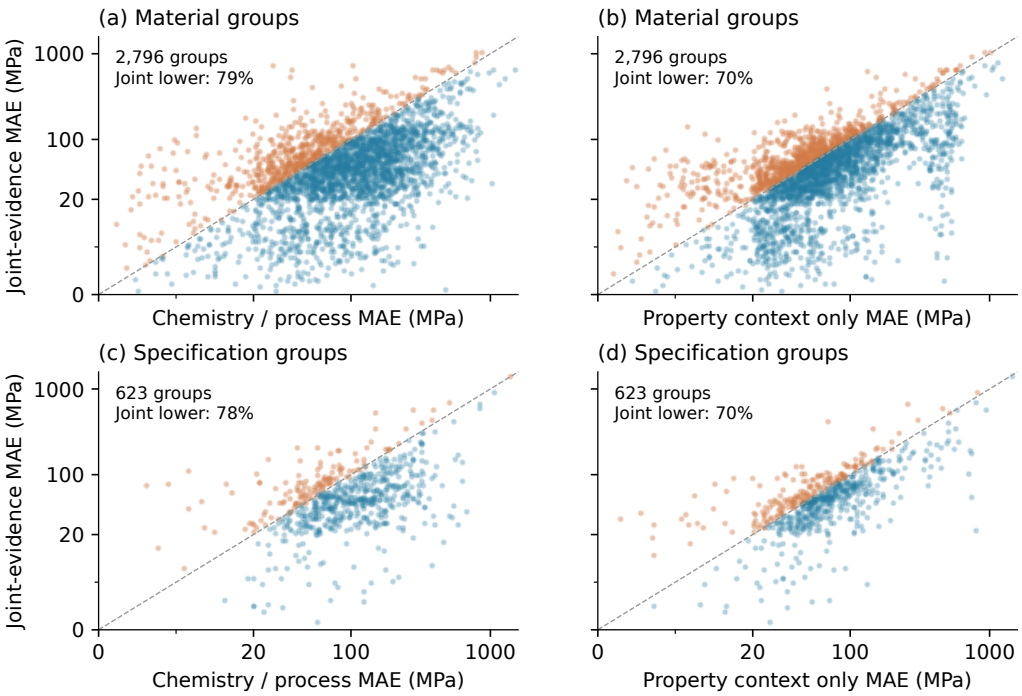

Figure 1: Each point pairs the joint model with chemistry/process alone (left) or property-context boosting (right). Rows use separate material and specification-family splits. Group errors average both UTS endpoints and all records in the group. Blue points below the identity line favor joint evidence; orange points retain failures to improve. All groups are shown. Axes are linear up to 20 MPa and logarithmic above it. The aggregate means reproduce Table 1.

report descriptive level, shape, false-acceptance, false-rejection, and coverage errors without significance claims. Per-experiment configurations and thresholds are fixed, with no retrospective risk calibration. Saved predictions support execution replay; separate perturbation diagnostics examine numerical stability.

4 Evaluation of completion, decisions, and regions

We test complementary evidence, joint decisions, and response-to-region inference under the evaluation contract in Section 3. Task-specific cohorts, splits, and controls are specified below.

4.1 Joint evidence improves whole-view completion

The main task retains all 7,797 target-complete low-alloy records in a predefined catalogue region. Both ultimate tensile strength (UTS) endpoints and UTS text are withheld. Inputs comprise chemical bounds and permitted process descriptors, optionally augmented by four yield-strength (YS) and elongation endpoints. Grade and specification identifiers are excluded from predictors. The context is complete for 6,352 records, partial for 841, and entirely missing for 604; none of these input-missingness strata is dropped.

We separately evaluate five-fold splits of 2,796 material components and 623 specification families. Material components connect normalized grade names or identical complete chemical fingerprints. Specification groups normalize the standard designation across revisions. These are separate axes, not simultaneous unseen-material and unseen-source splits; specification families are not independent publishers. MAE averages records within held-out groups, groups, and the two endpoints. All boosted regressors use 160 trees, 15 leaves, learning rate 0.05, and the same fixed regularization and missing-value handling.

Table 1: Whole-UTS-view completion: group-macro MAE (MPa; lower is better). Material and specification-family splits are evaluated separately. The context-only boosting control was added on the same exposed cohort and folds; it is a matched control, not an independent replication.

Predictor	Material	Specification
Training mean	259.86	275.08
Chemistry/process boosting	132.64	130.97
Context-only Ridge	121.79	115.03
Context-only boosting	104.39	100.11
Chemistry/process + disrupted context	118.75	108.86
Chemistry/process + true context	**69.58**	**75.30**

Figure 1 shows the group-level gains and losses. Joint evidence outperforms both chemistry/process and a strong nonlinear context-only control (Table 1). Disruption permutes complete context bundles within specification and coarse chemistry strata, separately in training and test batches. It preserves partial local structure and is a conditional diagnostic, not a deployment procedure. The result supports useful joint information, while physical coupling and shared normative conventions remain competing explanations.

Aggregate improvement is not uniform robustness: among records with no context, specification-split MAE is 263.95 for the joint model versus 254.06 for chemistry/process alone. Independent endpoint regressors also reverse the predicted bounds in 564 material-split and 522 specification-split cases out of 7,797, although no target pair is reversed. These outputs require an explicit consistency policy before interval-based decisions.

4.2 Broad-corpus utility is not specific to neural models

A separate five-fold grade-held evaluation compares full, automatically filtered, and random size-matched auxiliary pools. The full pools contain 90,710–98,216 records; the filter retains 1,514–1,541 using iron content, observed chemistry, context availability, and evidence roles. All pools exclude the entire 1,071-record target cohort, held-out grade names, and matching test chemical fingerprints. Both architectures receive the same local labels from 25% of eligible training grade groups. Evaluation covers 455 YS and 565 UTS records, averaging errors over their two stored fields and then grades.

After adaptation, full-corpus boosting obtains YS/UTS MAE of 97.36/129.13 MPa, versus 131.00/155.15 after filtering and 120.74/140.66 with random size-matched data. The corresponding neural results are 121.65/139.30, 147.72/169.58, and 126.81/148.19 MPa. Thus the broad corpus helps both architectures; a comparison of full-corpus neural training against filtered boosting alone would miss the stronger full-corpus boosting result. The filter changes domain and label availability, not just quality, and is not expert curation. Grade-level gain distributions are retained in Appendix E, Figure 3. Shared preprocessing uses permitted full-pool statistics; adaptation procedures and compute are not identical. This tests one reproducible selection rule, not the value of all data preparation or expert labor.

4.3 Functional completion and decision tradeoffs

For ASME-derived temperature functions, we retain 797 carbon/low-alloy steel records in 166 dependency components. Whole YS functions are hidden. Inputs contain UTS at $100, 200, \ldots, 1000\,^{\circ}\mathrm{F}$, availability masks, and permitted product metadata. Interpolation stays within source support. Five component-held folds merge metadata links and exact curve matches before fitting; this is completion within one handbook, not source transfer. Targets exist for 520 records and 4,742 queries. The 117 evaluable records without any UTS query remain included; predictions for 277 target-missing records cannot be scored.

ExtraTrees (Geurts et al., 2006) uses 64 trees, depth 10, and minimum leaf size 3, without tuning. Its lower absolute MAE does not establish better response shape: Δ evaluates changes relative to the $100\,^{\circ}\mathrm{F}$ anchor, where plain copying is stronger (Table 2). Nor does lower MAE minimize false acceptance. The catalogue query exposes a tradeoff that aggregate regression error conceals.

Table 2: Whole-function completion and a fixed catalogue query. MAE gives equal weight to dependency groups after averaging available temperatures and records. FA/FR count false accepts/rejects for YS ≥ 30 ksi at $700\,^{\circ}\mathrm{F}$ on 469 records, not physical failure rates.

Predictor	MAE (ksi)	Δ MAE (ksi)	FA	FR
Nearest donor, copy	6.740	**2.583**	**40**	46
Nearest donor, additive	6.487	2.873	57	41
Full-context Ridge	7.641	7.451	114	**8**
Full-context ExtraTrees	**5.686**	2.683	71	12

Table 3: A sufficient property value does not establish applicability. This stored example comes from the separate 34,595-query historical cohort; UNKNOWN denotes unsupported applicability, not material failure. Full confusion matrices appear in Appendix E, Figure 4.

Source	Grade 1.4637: elongation A5 $> 35\%$, $20\,^{\circ}\mathrm{C}$, annealed, size < 75 mm.
Query	A5 $\geq 10\%$, same temperature and state, size $= 75$ mm.
Decision	Lookup: **PASS**. Typed pipeline: **UNKNOWN**: the size boundary is excluded.

Concrete substitution query. Full catalogue lookup identifies normalized A537 Class 1 plate as supporting an incumbent A516 Grade 70 query at 10 mm: UTS within 485–620 MPa, YS ≥ 260 MPa, and elongation $\geq 21\%$ at 50 mm gauge length. The alternative specifies 485–620 MPa, 345 MPa, and 22%, respectively. Quenched-and-tempered Classes 2/3 specify UTS up to 690 MPa and therefore do not establish the requested upper-bound guarantee. This is a catalogue candidate, not a certified substitution: composition information mixes heat and product analysis and different thickness scopes. Preserving that unresolved status is part of the decision output, and complete lookup remains the exact baseline.

4.4 Structured decisions and selective evidence acquisition

Table 3 illustrates why a property value must be checked together with its applicability conditions.

On 1,114 fixed condition-applicability queries, zero-shot DiffusionGemma with one denoising step and one noise draw achieves 63.11% accuracy and 247 false applicability assertions. A task-trained extractor followed by exact unit and interval operations achieves 99.46–99.82% with no false applicability assertions across two seeds. This compares ready-to-use structured inference against a supervised specialist, not architectures under equal supervision. The local diffusion implementation reads label logits without generating explanatory text (mmastrac, 2026); the original Jev service is not evaluated.

A separate one-observation acquisition study uses 56 catalogue variants from 14 grades and 1,680 training-derived synthetic requirements. Selecting an anchor through empirical response-scenario decision loss makes 24 errors, versus 47 for a fixed 25 mm anchor. However, a subsequent control conditioned only on the requirement location makes 25 errors. Thus acquisition is useful, while additional threshold-specific benefit remains weak. Appendix B reports protocols, asymmetric errors, and controls; neither study establishes end-to-end framework performance or a manufacturing guarantee.

4.5 Shared execution and simultaneous requirements

A shared execution interface. We integrated frozen property predictors, learned condition extraction, source-relative requirement checks, and composition/process-window operators behind one evidence/query interface. The numerical adapter reproduces all 41,330 historical validation predictions for each of four existing backends. This establishes integration fidelity, rather than independent confirmation of predictive accuracy. Predictive recommendations retain a distinct status from source-supported conclusions.

Jointly withheld requirements. We next withheld both properties in each of the six pairs drawn from yield strength, tensile strength, elongation, and reduction in area. Both queries receive only the remaining available property observations and permitted context. Nine requirement profiles per

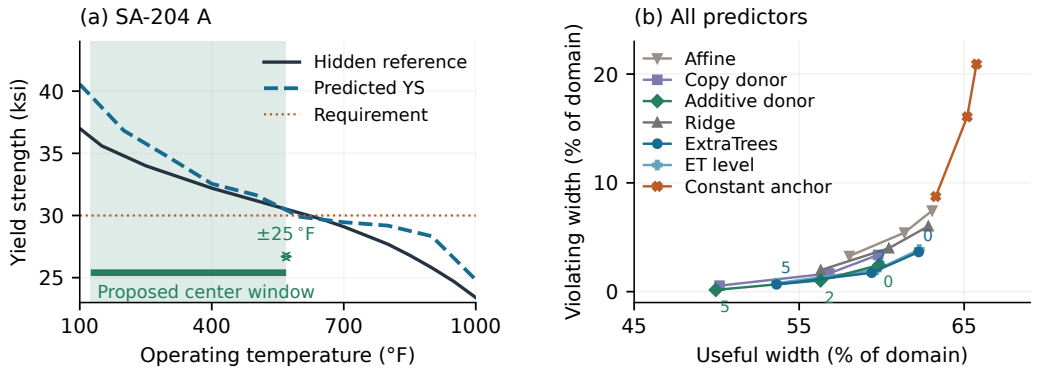

Figure 2: **(a)** The fixed SA-204 A example: predicted YS, withheld reference, and a 30-ksi requirement. Shading marks proposed temperature centers after allowing $\pm 25\,^{\circ}\mathrm{F}$ variation; the reference is used only for evaluation. **(b)** All seven predictors on 508 evaluable records. Each line joins the prespecified reserves 0, 2, and 5 ksi; numbers label the two highlighted methods. Lower-right is preferable. Fractions measure interval width, not failure probability; reserves are not calibrated error bounds.

pair use previously fixed training quantiles. Evaluation uses 10,266 record–pair combinations on the historical validation split; combinations share records. For group-balanced errors, then averaged over pairs and profiles, full-evidence boosting reduces false acceptance from 0.04522 to 0.04431 and false rejection from 0.07162 to 0.07010 relative to the earlier boosting recipe. Thus the component improvement also benefits a joint predictive decision, without establishing full material interchangeability.

Training for the joint decision. A matched campaign trains the degree-two operator from scratch with regression alone or regression plus a differentiable joint-decision loss, for 120 epochs on three seeds. Both targets remain hidden from both conditional queries. With the prespecified cost $5\mathrm{FA} + \mathrm{FR}$, the auxiliary loss changes validation costs from 0.33640 to 0.32512, from 0.33642 to 0.32364, and from 0.30535 to 0.31800. The gain is therefore not consistent across seeds. All models remain weaker on this criterion than the full-evidence boosting control (0.29164). Rejecting every candidate yields 0.29092 but no useful acceptances, illustrating why this scalar cost must be accompanied by useful coverage. These comparisons test a specific learning objective on historical validation, rather than a general impossibility of decision-focused learning. Full trajectories appear in Appendix E, Figure 5.

4.6 Learned responses yield nonzero operating windows

From a hidden response to an operating window. Using the existing grouped whole-function completion experiment, we invert predicted yield-strength curves into operating-temperature windows. All 797 records are retained. Requirements are 20, 30, or 40 ksi over 100–1000°F; we impose a temperature variation of $\pm 25^{\circ}\mathrm{F}$ and a minimum center-window width of 50°F. Predictions and window proposals do not observe the held-out yield-strength curve or its availability domain. Evaluation subsequently uses all available native curve knots, with explicit piecewise-linear reconstruction. A positive-width evaluable center domain exists for 508 records; unsupported predicted width is recorded separately.

At zero reserve (Table 4), ExtraTrees recovers more useful width than the additive donor but also admits more violating width; the constant anchor obtains still greater coverage at a much larger violation cost. Thus the best regression model does not dominate the simpler donor on decision quality. Figure 2 also shows the tradeoffs at reserves of 2 and 5 ksi. These are empirical width measures, not failure probabilities or calibrated error bounds. The conditional guarantee in Equation (6) requires a matching interval-wide one-sided error bound; Wolfram/Lean checks and reproduction details are given in Appendix D.

Table 4: Operating windows with zero additional reserve. Entries are fractions of the evaluable temperature domain, averaged within groups and then over requirements. Larger useful width and smaller violating width are better.

Response predictor	Useful width	Violating width
Constant predicted anchor	0.65750	0.20916
Nearest additive donor	0.59842	0.02443
ExtraTrees, full context	0.62244	0.03647

5 Discussion and limitations

Joint evidence improves completion even against the nonlinear context-only control, while missing-context errors and reversed predicted bounds expose limits of the aggregate gain. Likewise, full-corpus training helps both architectures, but the separate pretraining controls (Appendix C) do not establish useful general parameter transfer or a sample-efficiency advantage. That negative result concerns the tested numerical recipe; broader transfer claims require held-out tasks, explicit information and label budgets, and training-from-scratch controls.

The regression winner does not dominate the donor on operating-window quality, and joint-decision loss improves only two of three seeds. These results support separate evaluation of downstream objectives (Elmachtoub & Grigas, 2022). False acceptance here means disagreement with a source-defined requirement, not industrial failure probability. A feasible composition range inherits mechanical properties only through a supported relation; conditional calculations do not demonstrate a new alloy, manufacturing validation, or cost savings. Direct diffusion label readout is feasible but does not enforce applicability. A supervised extractor plus exact operations performs better under its greater supervision. Acquisition improves over a fixed observation, but the requirement-location control explains nearly all of the additional gain. These controls locate the useful mechanism without attributing it to model scale or a more elaborate policy. Several functional tasks cover few sources; source dependencies and historical encoder exposures are not exhaustively known. Units, permitted inputs, unresolved conditions, and unsupported queries remain explicit. The demonstrated integration reuses a decision procedure with trained specialists. Universal weight transfer, annotation savings relative to external multitask methods, and a learned composition-to-properties design loop remain open; they are not prerequisites for the evaluated prediction-to-window route.

6 Conclusion

Learning material decisions from partial knowledge can be implemented through whole-view completion and typed constraint execution. Complementary observations improve property completion, simultaneous masking supports joint selection, and predicted functions yield nonzero operating windows checked against hidden reference responses. Strong trees and functional donors remain complementary: regression accuracy alone does not choose the best decision specialist. The result is a common method for applying fragmented engineering knowledge, with useful coverage and application errors measured directly.

REPRODUCIBILITY STATEMENT

The appendix specifies observation roles, grouping, target masking, evaluation units, and experiment settings. The supplementary code archive is available at `https://anonymous.4open.science/r/materials-decisions-artifact-3687/`. The accompanying `formal/` directory contains the Lean checks.

AI USE STATEMENT

Generative AI assistants were used in literature search, data inspection, text editing, verification, Lean proof development and checking, and manuscript formatting.

REFERENCES

Pierre-Paul De Breuck, Geoffroy Hautier, and Gian-Marco Rignanese. Materials property prediction for limited datasets enabled by feature selection and joint learning with MODNet. *npj computational materials*, 7(1):83, 2021. doi: 10.1038/s41524-021-00552-2. URL https://www.nature.com/articles/s41524-021-00552-2.

Adam N Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management science*, 68(1): 9–26, 2022. doi: 10.1287/mnsc.2020.3922. URL https://pubsonline.informs.org/doi/10.1287/mnsc.2020.3922.

Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and SM Ali Eslami. Conditional neural processes. In *International conference on machine learning*, pp. 1704–1713. PMLR, 2018. URL https://proceedings.mlr.press/v80/garnelo18a.html.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine learning*, 63(1):3–42, 2006. doi: 10.1007/s10994-006-6226-1. URL https://doi.org/10.1007/s10994-006-6226-1.

Yury Gorishniy, Akim Kotelnikov, and Artem Babenko. TabM: Advancing tabular deep learning with parameter-efficient ensembling. In *International Conference on Learning Representations*, volume 2025, pp. 77899–77935, May 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/c1ba41c694834aeef91ae161711d4939-Abstract-Conference.html.

Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems*, 35:507–520, 2022. doi: 10.52202/068431-0037. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/0378c7692da36807bdec87ab043cdadc-Abstract-Datasets_and_Benchmarks.html.

Oleg Ivanov, Michael Figurnov, and Dmitry Vetrov. Variational autoencoder with arbitrary conditioning. *arXiv preprint arXiv:1806.02382*, 2018. URL https://arxiv.org/abs/1806.02382.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html.

Chao Ma, Sebastian Tschiatschek, Konstantina Palla, José Miguel Hernández-Lobato, Sebastian Nowozin, and Cheng Zhang. EDDI: Efficient dynamic discovery of high-value information with partial VAE. *arXiv preprint arXiv:1809.11142*, 2018. URL https://arxiv.org/abs/1809.11142.

Gregor Maier, Jan Hamaekers, and Benedikt Ziebarth. Multitask learning via task embeddings for glass property prediction with improved sample efficiency. *Computational Materials Science*, 263:114407, 2026. doi: 10.1016/j.commatsci.2025.114407. URL https://doi.org/10.1016/j.commatsci.2025.114407.

Duncan McElfresh, Sujay Khandagale, Jonathan Valverde, Vishak Prasad C, Ganesh Ramakrishnan, Micah Goldblum, and Colin White. When do neural nets outperform boosted trees on tabular data? *Advances in neural information processing systems*, 36:76336–76369, 2023. doi: 10.52202/075280-3337. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f06d5ebd4ff40b40dd97e30cee632123-Abstract.html.

mmastrac. diffgemma. GitHub software repository, 2026. URL https://github.com/mmastrac/diffgemma/tree/6f6c825dc9a7bad15451f8dc35174209c14480a3. Evaluated revision 6f6c825; accessed 26 September 2026.

Viggo Moro, Charlotte Loh, Rumen Dangovski, Ali Ghorashi, Andrew Ma, Zhuo Chen, Samuel Kim, Peter Y Lu, Thomas Christensen, and Marin Soljačić. Multimodal foundation models for material property prediction and discovery. *Newton*, 1(1):100016, 2025. doi: 10.1016/j.newton.2025.100016. URL https://doi.org/10.1016/j.newton.2025.100016.

Andre Niyongabo Rubungo, Craig Arnold, Barry P. Rand, and Adji Bousso Dieng. LLM-Prop: predicting the properties of crystalline materials using large language models. *npj Computational Materials*, 11(1):186, 2025. doi: 10.1038/s41524-025-01536-2. URL https://www.nature.com/articles/s41524-025-01536-2.

Pranav Shetty, Arunkumar Chitteth Rajan, Chris Kuenneth, Sonakshi Gupta, Lakshmi Prerana Panchumarti, Lauren Holm, Chao Zhang, and Rampi Ramprasad. A general-purpose material property data extraction pipeline from large polymer corpora using natural language processing. *npj Computational Materials*, 9(1):52, 2023. doi: 10.1038/s41524-023-01003-w. URL https://www.nature.com/articles/s41524-023-01003-w.

Ryan Strauss and Junier B Oliva. Arbitrary conditional distributions with energy. *Advances in Neural Information Processing Systems*, 34:752–763, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/06c284d3f757b15c02f47f3ff06dc275-Paper.pdf.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/f22e4747da1aa27e363d86d40ff442fe-Abstract.html.

A REPRODUCTION CONTRACTS AND CONDITIONAL ALGEBRA

A.1 PROPOSED JOINT OBJECTIVE AND ITS IMPLEMENTATION BOUNDARY

Equation (3) specifies a proposed integration of evidence reconstruction and decision-oriented learning. For a hidden functional view with at least two supported coordinates, let $\overline{v}$ and $\overline{\widehat{v}}$ be means on the same set J. One concrete shape term is

$$\ell_{\text{shape}} = \frac{1}{|J| s_q^2} \sum_{j \in J} \left[(\widehat{v}_j - \overline{\widehat{v}}) - (v_j - \overline{v}) \right]^2 . \tag{8}$$

It is omitted for a view without the required support. For a predefined query with a defined catalogue-reference answer $d \in \{0, 1\}$, a concrete decision term is

$$\ell_{\text{decision}} = -c_{\text{FR}} d \log p_\theta - c_{\text{FA}} (1 - d) \log(1 - p_\theta), \qquad p_\theta = \sigma(m_R(\widehat{v})/\tau), \tag{9}$$

where $c_{\text{FA}}, c_{\text{FR}}, \tau > 0$ are declared training constants. For a lower-threshold query, $m_R(\widehat{v}) = \widehat{v} - r$; for interval containment, one possible margin is $\min\{\widehat{v}_{\text{lo}} - L, U - \widehat{v}_{\text{hi}}, \widehat{v}_{\text{hi}} - \widehat{v}_{\text{lo}}\}$. The score p_θ is not automatically a calibrated physical probability. Queries unresolved by their reference evidence are not assigned negative labels. For applicable relations $g_k(v) \le 0$, a concrete constraint term is the average of $\max\{g_k(\widehat{v})/s_k, 0\}^2$ over the supported relations, with declared positive scales s_k; the term is omitted if none apply. No cross-record physical relation is introduced solely from a shared grade name.

These formulas make the proposed learning contract explicit. Their joint optimization and choice of weights have not been validated by the completed experiments below. The implemented numerical shared model used observed- endpoint squared reconstruction; centered shape and decision errors in the functional and text studies were evaluation quantities, not a retrospectively claimed joint training objective. Exact decision and box logic were checked separately. An eventual integrated experiment must fix its task distribution, weights, exposure boundary, and strong same-information controls before reporting performance.

A.2 EVIDENCE SUPPORT, MASKING, AND DEPENDENCE

The numerical catalogue experiments use a Granta GMS extract. Fatigue and thermal-response experiments use JAHM functional views and their property, unit, and parameter metadata. A separate functional-completion component uses ASME-derived strength functions. Database and standard versions are not uniformly established across these recovered extracts. These source names identify provenance; they do not imply independence among entries or permission to redistribute licensed raw records.

The empirical tasks concern agreement with the stated source operator. Normative endpoints and published approximations are retained as useful targets with their roles; they are not relabeled as independent measurements. Unknown metadata are represented explicitly rather than supplied by a plausible but unsupported interpretation. Reusing a source-derived formula to reproduce a curve calculated by that formula is an operator-recovery test, not an independent physical validation.

Whole-view masking removes the queried endpoints together. In the shared numerical experiment, all ultimate-tensile-strength values, associated notes, target roles, and target availability are excluded from auxiliary training. Each queried auxiliary property's endpoints and availability are masked together. In the fatigue experiments, the target curve and its support are not inputs. Permitted text retains specified processing information, while target-bearing clauses, bibliography, and identified aliases are removed according to the declared preprocessing rule. This does not ensure that every physically useful condition survives parsing.

The property-transfer experiment predicts yield strength from available ultimate tensile strength at the same catalogue record and temperature. The fixed temperature grid is 293.15, 473.15, and 673.15 K. Each property must have one unambiguously decoded series, supported stress units, and a temperature argument. Internal queries use linear interpolation. Queries outside support remain unsupported; a singleton supports only its matching temperature. An endpoint tolerance of 0.001 K accommodates stored floating-point coordinates and is not a process tolerance. Inputs use the

ultimate-strength note; the yield-strength note is audit metadata, not a feature. Same-record pairing does not resolve contradictory treatment notes or establish a common specimen.

Grouping is task-specific. The shared numerical experiment closes normalized grade identifiers and sufficiently observed exact composition patterns over the available corpus. Functional and textual tasks additionally use declared source, graph, coordinate, reference, and copying links. These are conservative exclusion rules, not a reconstruction of every source genealogy. Connected components are kept intact across the relevant training and test boundary. Known previously examined components in the property-transfer experiment are training-only. Its three new evaluated source components support a descriptive comparison, not a precise population-level uncertainty estimate.

A.3 Shared numerical training and controls

The numerical pretraining corpus contains $99{,}328$ recovered records. The fixed downstream scope retains $11{,}552$ records, including $3{,}755$ without ultimate-strength targets. The auxiliary queries are ten composition intervals—C, Mn, Si, Cr, Ni, Mo, V, S, P, and Fe—plus yield-strength and elongation endpoints. There is no new manual annotation. Existing target answers nevertheless remain labels, so this experiment is not label-free.

For each of three component-held-out folds, exactly 64 target-bearing training components supply downstream answers. All their available target endpoints are shared across learners. The conditional network has two hidden layers of widths 128 and 64, ReLU activations, a two-coordinate output, and a 13-slot query indicator. Inputs contain standardized permitted values, availability indicators, and predefined condition and standard-family flags. Train-only moments are used; missing standardized values become zero with their missingness retained separately. Target normalization is property-wise, and downstream target moments use only the selected local answers.

Auxiliary training uses two complete passes, batch size 512, and a fixed property-query sampler. Downstream adaptation uses 400 updates of batch size 128. Adam uses learning rate 10^{-3} and weight decay 10^{-4}. The loss is squared error over observed standardized endpoints, with inverse component-frequency weights normalized within the batch. Two fixed initialization seeds are paired with scratch controls; they are optimization sensitivity checks, not independent datasets. Scratch uses the same initial parameters and downstream updates. Pooled boosting receives the same saved auxiliary examples, masked designs, queries, and downstream answers; local boosting and Ridge provide additional controls. Paired auxiliary endpoints are also shuffled within query, coarse standard family, and a declared note marker to test correspondence beyond those preserved distributions. This shuffle is not a physical intervention or an exchangeability-based test.

A.4 Frozen representations, readouts, and text controls

The fatigue block retains 85 views, including 68 target-bearing curves and 217 supported queries at 10^4, 10^5, 10^6, and 10^7 cycles. Stress amplitude and maximum stress retain their source definitions. Three outer folds hold source components intact. Each fold has eight eligible labeled training sources; budgets of four and eight use fixed source orders and all supported answers from selected sources. The full budget is not refitted merely to duplicate an order. Two low-budget orders happen to select the same set in one fold; those selections are not independent replications.

Frozen MaterialsBERT, a historical continued-pretraining checkpoint, and TextJEPA are evaluated as available representations. Historical training exposure and exact base revision are not fully controlled, so the comparison does not identify a causal effect of domain pretraining. All arms receive the same permitted numerical context, masks, condition flags, and target-safe text. An embedding z is centered on selected training records and divided by one scalar root-mean-square coordinate scale; no principal-component selection is fitted. TF–IDF uses word/numeric unigrams and bigrams, sublinear term frequency, ℓ_2 normalization, and at most 2,000 features; vocabulary and document frequencies are fitted on selected training records.

Write $t = (\log_{10} N - 5.5)/1.5$. Ridge uses $[x, z, t, tx, tz]$, allowing context-dependent slopes rather than a common straight line. It uses an intercept, LSQR tolerance 10^{-8}, at most 10,000 iterations, and $\alpha_B = 10 n_B / n_{\text{full}}$, where n_B counts selected training answers. For n_i answers in record i and n_g labeled records in source g, raw answer weights are $1/(n_g n_i)$ and are rescaled to mean one. With the library's weighted sum-of-squares objective, the scaling keeps the relative penalty constant

across budgets within an outer fold. The numerical and numerical-plus-text LightGBM controls use 160 trees, 15 leaves, learning rate 0.05, minimum child count five, ℓ_2 leaf regularization one, seed 42, and two threads, without a hyperparameter sweep.

Residual partners fit two inner source-disjoint numerical boosters. Every residual target is generated by a model that excluded that residual's source. A Ridge correction is then added with fixed coefficient one to the full numerical booster. Residuals from smaller inner models need not match the full model's error distribution. A direct embedding-plus-boosting comparison was added after the original results as a disclosed exploratory completion; it is not a retrospectively preregistered confirmation.

The subsequent text diagnostic compares full permitted text, canonical serialization of existing numerical features, numerical/unit/style tokens, and source-constrained reassignment while retaining each record's own composition. Canonical text is a deterministic function of existing inputs and adds no information. Reassignment preserves only its specified grouping variables, can disrupt valid physical combinations, and may leave singletons unchanged. The changed support is reported. These controls identify alternative predictive explanations, not causal semantics. The transfer experiment refits all five predefined numerical/text procedures for the new property relation; it does not transfer the fatigue model's weights.

A.5 Error hierarchy and decision contracts

Equation (7) averages queries, then records, then evaluated groups. Unscored predictions do not increase the evaluation denominator. Record- and source-level errors are retained rather than treating all curve coordinates as independent observations. For functional shape, the reference and predicted profiles are each centered using their means over the same supported query set before taking absolute error. Anchor-change error instead subtracts the respective value at the declared anchor. These are distinct objectives; neither is implied by a smaller absolute-response error.

For a fixed requirement r, a false accept is $\widehat{d} = 1, d = 0$, and a false reject is $\widehat{d} = 0, d = 1$, where $d = \mathbf{1}\{y \geq r\}$. The fatigue comparison uses 400 MPa at 10^6 cycles, with amplitude and maximum-stress results reported separately. The numerical endpoint experiment asks whether the whole stored ultimate- strength interval lies inside $[600, 1000]$ MPa. Predicted endpoints must be ordered; they are not silently sorted for that query. These thresholds are catalogue evaluation scenarios, not universal engineering requirements. The yield-from-ultimate-strength transfer defines no acceptance threshold. Error counts, hierarchical rates where computed, coverage, and average prediction error are reported separately. No fitted risk calibration follows from these summaries.

For the proportional transfer control, nonnegative training weights w_j and $\sum_j w_j u_j^2 > 0$ give the unique least-squares coefficient

$$a^* = \frac{\sum_j w_j u_j y_j}{\sum_j w_j u_j^2}, \qquad \widehat{y} = a^* u. \tag{10}$$

It has no intercept or post-fit clipping. These weights balance sources; they are not asserted to be inverse physical-noise variances. This control is fitted by squared error and evaluated by the declared absolute error.

A.6 Agreement requires containment

Let L, U, y, r be real, with $L \leq U$, and let a response x accept exactly when $r \leq x$. All responses in $[L, U]$ accept if and only if $r \leq L$; all reject if and only if $U < r$. If $L \leq y \leq U$, either unanimous decision agrees with the designated reference y. For a finite nonempty set of alternatives, its minimum and maximum give the same unanimous threshold decisions without asserting attainability of intermediate values.

Wrong unanimous acceptance requires $y < r \leq L$, and wrong unanimous rejection requires $U < r \leq y$. With $a_+ = \max(a, 0)$,

$$(L - y)_+ + (y - U)_+ = \max\{L - y, y - U, 0\} = \operatorname{dist}(y, [L, U]). \tag{11}$$

The gap vanishes exactly when the reference is contained. These threshold equivalences, containment soundness, and the algebraic identity were checked in Lean/Mathlib and independently sym-

bolically checked in Wolfram. The interpretation as the Lebesgue length of erroneous threshold bands follows elementarily from the displayed intervals, but was not formalized in Lean. On a finite threshold grid one counts actual grid points, not gap divided by step size. Neither quantity is a failure probability without an appropriate distribution and physical reference. In the checked diagnostic, the reference was a formula from the same source, so disagreement concerns source representations rather than independent physical truth.

A.7 Dependent ranges and exact linear support

For real a, c, h with $h_j \geq 0$, $\max_{x \in B(c,h)} a^\top x = a^\top c + |a|^\top h$: choose each endpoint according to the sign of a_j. This proves Equation (5). With fixed positive proposed halfwidths $\bar{h}$, one may maximize $0 \leq \lambda \leq 1$ subject to $Ac + |A|\lambda\bar{h} \leq b$ and declared coordinate bounds. Any optimum concerns this specified box family, not every feasible shape. Floating-point solver outputs are checked separately; no directed-rounding certificate is claimed for arbitrary numerical optimization.

The catalogue composition example uses the declared continuous interpretation

$$\mathrm{Mn} \leq 1.60, \qquad 6\mathrm{C} + \mathrm{Mn} \leq 6C_{\max} + 1.30. \tag{12}$$

For a nonempty C–Mn rectangle, the second constraint holds throughout exactly when $6C_{\mathrm{upper}} + \mathrm{Mn}_{\mathrm{upper}} \leq 6C_{\max} + 1.30$. At $\mathrm{Mn}_{\mathrm{upper}} = 1.60$, this is equivalent to $C_{\mathrm{upper}} \leq C_{\max} - 0.05$. The old rectangle $C \leq C_{\max}, \mathrm{Mn} \leq 1.30$ is contained in the compensated halfspace. These concrete implications and a saved five-coordinate box were checked with exact rational constants in Lean/Mathlib and Wolfram. The result establishes conditional algebra, not the source's continuous interpretation, heat/product-analysis equivalence, or mechanical response of a newly chosen composition. Its proof is elementary and is not proposed as a novel mathematical contribution.

A.8 Leaf geometry and numerical reproducibility

The functional experiment retains 80 related-steel records, including 18 without supported targets; 62 provide 412 conductivity queries on $300, 400, \ldots, 900$ K. Separate three-fold material- and source-component splits use the same block, with material-component purging in the source split. Each fold fits one LightGBM only to training $k(300\,\mathrm{K})$: 100 trees, seven leaves, learning rate 0.05, minimum child count three, 63 histogram bins, ℓ_2 regularization one, and seed 42. Inputs comprise composition endpoints, available heat-capacity/density/thermal-expansion functions, masks, and eight state flags. Conductivity, its support, and thermal diffusivity are excluded. At each query, the prediction averages donors from up to three distinct training material components supporting that query. Ties use a fixed record hash. Thus different temperatures can use different donors; the resulting profile is not one observed specimen. Chemistry, full-context, predicted-anchor, mask, and mean-profile controls use the declared common information and donor rules. Neither split is an independent replication of the other.

For one fixed ensemble of $T \geq 1$ trees, concatenate the one-hot terminal- leaf codes into $\phi(x)$. Then $\|\phi(x)\|^2 = T$, and the fraction of matching leaves is the positive-semidefinite kernel

$$K(x, x') = \frac{\langle \phi(x), \phi(x') \rangle}{T}, \qquad d_H(x, x') = 1 - K(x, x') = \frac{\|\phi(x) - \phi(x')\|^2}{2T}. \tag{13}$$

Indeed, $\sum_{ij} a_i a_j K(x_i, x_j) = \|\sum_i a_i \phi(x_i)\|^2 / T \geq 0$. Leaf-constant ensemble output is a linear functional of this code, so different routes may have identical scalar output. On original inputs, d_H is a pseudometric; threshold crossings and hard donor selection can cause discontinuities. Codes from separately fitted folds are not a shared global coordinate system. This algebra establishes neither latent physical states nor stability of a manufacturing window.

Reproduction uses saved splits, permitted features, masks, transformations, models, and predictions. Independent implementations recheck raw support, group closure, model inference, and aggregate arithmetic without retraining. Equivalent floating-point reduction orders can produce slightly different standardized inputs; tree outputs can change at thresholds despite exact replay of the persisted matrix. Arithmetic sensitivity is therefore reported separately from execution reproducibility and from physical input uncertainty. Formal checks use Lean 4.26.0 with Mathlib and Wolfram 15.0.1; they certify only explicitly stated real-valued premises, not statistical coverage or physical completeness.

B STRUCTURED INFERENCE AND EVIDENCE-ACQUISITION CONTROLS

Frozen condition-applicability evaluation. The task asks whether a supplied catalogue condition applies to a query's processing state, temperature, and size. The 1,114 single-clause validation queries include 198 queries from clauses unseen during specialist training. Queries and targets are generated under an explicit source-derived contract: unit conversion is exact, explicit inequalities retain their direction, unspecified endpoints of bare ranges yield `UNKNOWN`, and a definite contradiction takes precedence over other missing fields. This is a constrained condition task, not arbitrary-document understanding. Both model families receive identical source/query inputs; answer keys are kept outside requests.

The frozen `diffgemma` Rust/Metal implementation uses revision `6f6c825`, with the author's q4 DiffusionGemma weights at revision `be312db`. We fix one denoising step, one noise draw, seed 42, and no generated reasoning. There is no task-specific adaptation or validation-driven prompt revision. The specialist is trained to extract condition spans and uses an exact rational unit/interval operator; its direct-classification control is separately trained on the same task. Table 5 therefore assesses zero-shot readiness, not equal-supervision sample efficiency.

Table 5: Condition-applicability validation. FA counts predictions of `APPLIES` when the target is either rejection or unknown. There are 728 such negative/unknown targets overall. Unseen-clause queries are a subset, not an independent test. Ranges cover two specialist seeds.

Method	Accuracy (%)	FA	Unseen accuracy (%)
Zero-shot DiffusionGemma, q4	63.11	247	62.63
Task-trained direct classifier	91.20–91.56	18–40	87.37–90.40
Task-trained extraction + exact execution	99.46–99.82	0	96.97–99.49

All 1,114 requests completed locally on an Apple M5 Max with 128 GB unified memory. Median request latency was 0.70 s; the first request took 7.79 s. Transport, option-to-token mapping, source inputs, answer keys, step/sample counts, and aggregate accuracy were replay-checked. These checks are not an independent model replication. The result does not rule out improvements from adaptation, additional denoising, or another quantization/runtime; model probabilities are not calibrated engineering risks.

Selecting one additional curve observation. We retain 56 whole-variant held-outs from 14 catalogue steel grades. The existing chemistry-selected donor is from the same grade, which does not establish common heat or processing identity. Target values at 1.5, 3, and 5 mm are initially available. Exactly one additional value may be opened at 6.5, 13, 25, or 50 mm. All four candidate anchor positions are excluded from the common evaluation grid. For each evaluation position, three thresholds are training-only donor order statistics, yielding 30 synthetic requirements per variant. These are diagnostic queries, not actual part specifications.

For target T and chosen donor D, the fixed update is

$$\widehat{T}_h(d) = D(d) + (1-w)\delta_5 + w\delta_h, \quad \delta_j = T(j) - D(j), \quad w = \min\{1, \max\{0, (d-5)/(h-5)\}\}. \tag{14}$$

Only $T(h)$ is newly revealed. Training scenarios retain entire within-grade pair differences, centered at 5 mm. Before opening an anchor, the policy simulates each scenario's requirement decision and the decision after this update. It minimizes their mismatch rate, averaging first within training grade and then across grades. The held-out variant is excluded from both ends of every training pair. This is empirical finite-scenario risk minimization, not a calibrated posterior or a new parametric architecture.

The location-only control averages estimated losses over the three training-derived thresholds at that location, without conditioning on the particular requested threshold. It explains almost all of the apparent query-conditioning gain (Table 6). Reused variants, repeated thresholds, and overlapping training folds preclude treating the 1,680 queries as independent confirmations. An earlier construction with independent coordinate uncertainty boxes abstained on every query; its conservative radius prevented useful decision refinement. The scenario policy instead issues fallible binary decisions. This exploratory sequence motivates structured response models and strong location-only controls, without establishing threshold-specific superiority or physical coverage guarantees.

Table 6: One-observation acquisition on 1,680 synthetic queries. FA/FR are false accept/reject counts. Macro error averages within each grade and then over 14 grades. Rows marked $*$ are post-result diagnostic controls.

Acquisition rule	Errors	FA	FR	Macro error (%)
Fixed 13 mm	77	54	23	3.69
Fixed 25 mm	47	19	28	2.24
Requirement location and threshold	24	9	15	1.27
Requirement location only*	25	10	15	1.28
Cyclic threshold shift, same location*	28	10	18	1.45
Nearest observation coordinate*	43	21	22	2.56

Table 7: General conditional-transfer screen: mean endpoint group-macro MAE (MPa). Neural entries average two fixed seeds. This three-fold, limited-label protocol differs from Table 1; errors are not comparable across the two tables.

Predictor	MAE
Pretrained shared MLP	164.218
Identical MLP from scratch	164.261
Pretrained MLP, shuffled auxiliary answers	164.224
Local boosting	136.086
Pooled conditional boosting + adaptation	135.981
Ridge	**134.639**

C Additional transfer and text controls

We additionally test a shared query-conditioned MLP on chemical, YS, and elongation views, excluding UTS entirely from auxiliary inputs and targets. Three outer component-held folds retain 11,552 downstream records, including 7,797 with UTS targets. Each fold supplies the same 64 locally labelled UTS groups to every method. Pretraining receives two auxiliary passes; adaptation uses 400 steps. Controls include identical initialization and adaptation from scratch, shuffled auxiliary answers, and pooled conditional boosting using the same auxiliary query examples.

Neither the pretrained-versus-scratch difference nor the disruption control establishes useful transfer (Table 7). This limits the tested training recipe; it does not refute textual encoders, JEPA, or heterogeneous learning generally.

Text provides a separate, qualified signal. On 68 fatigue curves with 217 targets from 12 labelled source components, numerical-plus-TF–IDF boosting achieves 136.903 MPa versus 185.526 for numerical boosting under the same three source-held folds. At a fixed source-relative query, false accepts/rejects change from 17/3 to 1/8 over 66 answers. However, numeric/style-only text features reproduce all 66 decisions; semantic causation is unestablished. An earlier Ridge configuration on these reused targets achieved 131.636 MPa and 14/1, so the newer boosting result is not a task-wide MAE record. On a subsequent UTS-to-YS task with 12 evaluable records, 35 queries, and three held-out source components, adding full text worsens MAE from 77.211 to 125.251 MPa. This tests a retrained procedure, not transfer of fatigue-model weights.

D Joint decisions and response-window reproduction

The joint-requirement experiment uses the six pairs of YS, UTS, elongation, and reduction in area. Both target values are absent from both conditional queries. The neural campaign uses 43,441 training record–pairs and 10,266 historical validation record–pairs, three seeds (42, 1729, 2026), and two matched arms trained for 120 epochs. Arms share initialization and sampling. The regression-plus-decision arm uses a product of two sigmoid feasibility scores with temperature 0.25 in standardized target units, weighted binary cross-entropy with false-acceptance weight 5, normalization by 3, and objective weight 0.1. Adam uses learning rate 0.001 and weight decay 0.0001, with batches of 512. Training thresholds are sampled from the three fixed training quartiles per target. A prespecified

extension rule based on improvements in both training MSE and validation decision cost at epochs 48, 72, and 120 was not triggered. The reported 120-epoch results are not early-stopped pilot results.

The operating-window study reuses five grouped folds of the functional completion experiment; it adds no fits. It retains all 797 records, including 277 without the target function and 12 further records without a positive-width evaluable center domain after support restriction and perturbation erosion. Seven response predictors, three requirements, and three reserves (0, 2, and 5 ksi) yield 50,211 scenarios, not independent samples. Window proposals are saved before evaluation reads the hidden native curve coordinates. The reference is piecewise-linear through all available native knots, rather than through only the prediction grid. Proposed width outside known reference support is recorded as unscorable, not as a supported or violated requirement. Widths are normalized by each evaluable center domain, averaged within groups, then across groups and requirements. Missed width is a pointwise domain measure; it may include feasible segments shorter than the minimum issuable window. The public query route reproduces all 797 primary 30-ksi window outputs from stored out-of-group predictions. Replay is an integration check on the same cohort. The conditional window-transfer proof uses Lean 4.26 and Mathlib, with a separate Wolfram 15.0.1 check; neither supplies an empirical error bound.

E ADDITIONAL DIAGNOSTIC FIGURES

The following diagnostics retain the full grade-level distributions, condition-applicability matrices, and three-seed training trajectories underlying the main-text summaries. They introduce no additional experiments.

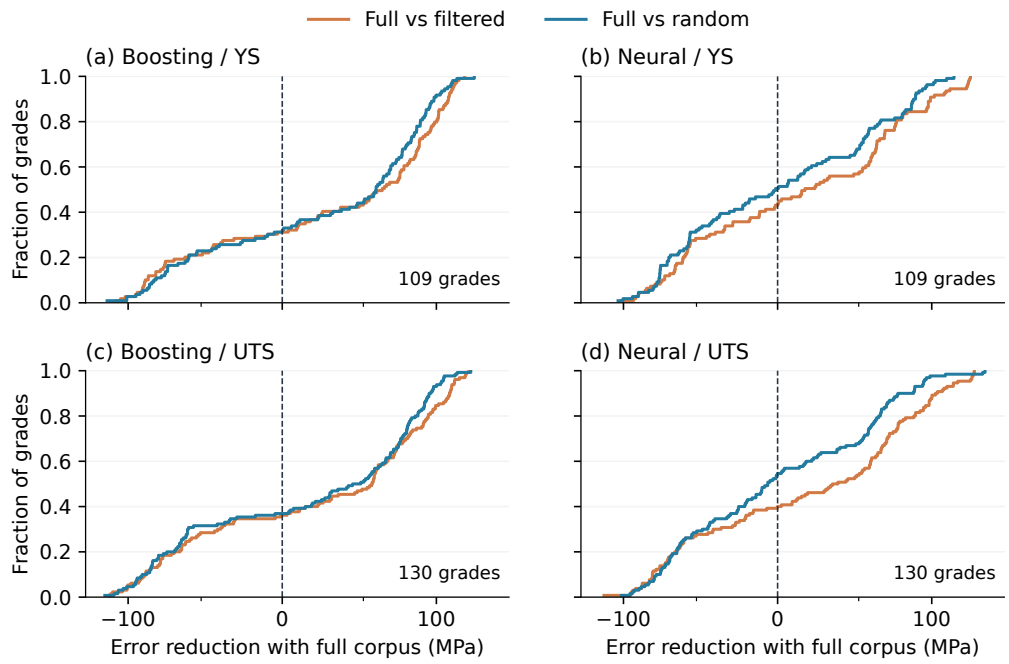

Figure 3: **Broad-corpus gains are heterogeneous across grades.** Empirical cumulative distributions of paired grade-level MAE differences: filtered minus full (orange), and size-matched random minus full (blue). Positive differences favor full-corpus training. Each grade has equal weight; all 109 YS grades and 130 UTS grades are retained. Curves use saved adapted predictions, without new fits or smoothing. The horizontal scale is linear within ± 10 MPa and logarithmic in the tails. At zero, the cumulative fraction records grades with no improvement. Filtering changes domain and label availability as well as size.

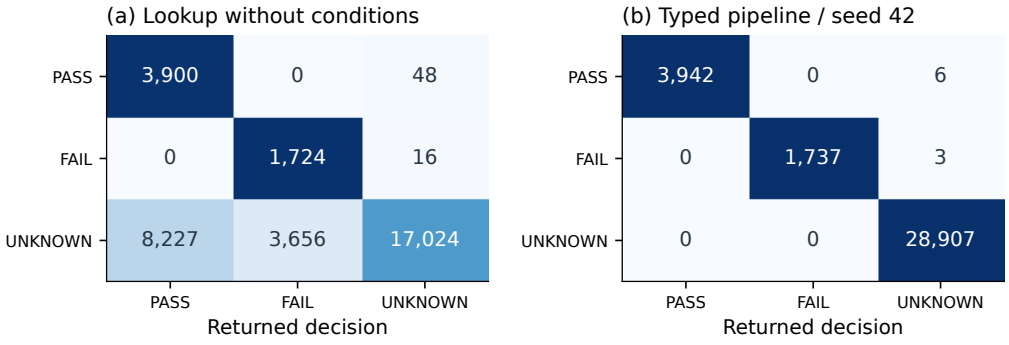

Figure 4: **Applicability errors.** Full reference-to-output matrices for 34,595 historical queries, comparing lookup with the typed pipeline (seed 42). Labels are counts; shared color intensity is the within-reference-row fraction. The parser defines the source reference. UNKNOWN means unsupported applicability, not material failure. This cohort is separate from the 1,114-query diffusion comparison.

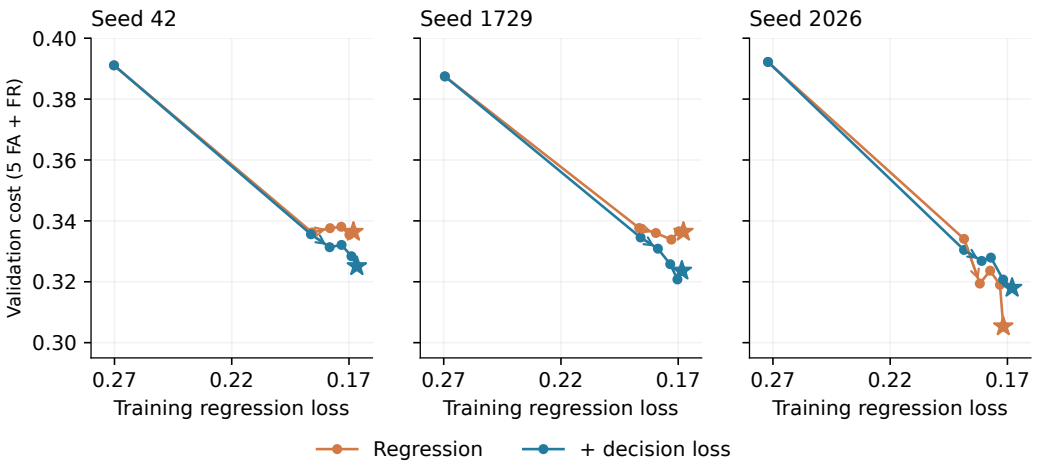

Figure 5: **Better training fit need not improve the engineering decision.** Each panel shows one matched seed, with saved epochs 0, 24, 48, 72, 96, and 120 connected in temporal order. Horizontal position is training regression loss; vertical position is validation decision cost, not the same loss on a second split. The reversed horizontal axis follows decreasing training loss to the right; stars mark epoch 120. Regression is orange and regression plus decision loss is blue. All three seeds and the complete trajectories are retained. This diagnostic reuses the full campaign and adds no training or smoothing.