

Benchmark Datasets for Photopolymer Modelling and Machine Learning: Coverage, Leakage, Shift and Fitness for Use

Designing benchmarks that test transfer across materials, machines and states rather than rewarding memorisation of correlated batches or process histories

Juan Seguro

3Dresyns by Resyner Technologies S.L.

© 2026 Juan Seguro. All Rights Reserved.

Correspondence: info@3Dresyns.com · www.3Dresyns.com

ER-398 · Engineering Review · Version 0.1 · 5 October 2026

Abstract

Benchmark datasets can accelerate photopolymer modelling and machine learning, but a benchmark is useful only when its split structure, coverage and metadata reflect the intended engineering deployment. Random row splits can leak batch, formulation, machine or specimen-family information across training and test sets and produce misleadingly optimistic performance. This review defines benchmark fitness in terms of target decision, coverage, independence, leakage control, distribution shift and uncertainty. It proposes split hierarchies and reporting requirements for photopolymer datasets whose dominant dependencies arise from shared material and process histories.

Keywords: benchmark dataset; machine learning; photopolymer; data leakage; dataset shift; validation; transferability

1. A benchmark is a test contract

A dataset becomes a benchmark only after the prediction task, target population, split logic and evaluation metric are defined. The same table can be a useful training set and a poor benchmark. For photopolymers, the intended deployment may be interpolation within one formulation, transfer to a new batch, transfer between printers or prediction for an unseen chemistry family. These are different contracts.

2. Hierarchical dependence creates hidden leakage

Rows from one batch, build plate or specimen can share chemistry and process noise. If related rows are split randomly between training and test sets, a model can exploit family identity rather than learn transferable structure. Leakage control should therefore occur at the highest dependency level relevant to deployment: specimen, build, batch, formulation, printer or laboratory.

Table 1. Split hierarchy for photopolymer data.

Split unit	What it tests	Leakage prevented
measurement row	repeatability only	almost none
specimen	new specimen from same build	within-specimen correlation
build	new build under similar recipe	build-specific effects
batch/formulation	material transfer	chemical identity leakage
machine/lab	hardware or site transfer	instrument and operator leakage

3. Coverage must be described in engineering coordinates

Dataset size alone says little about coverage. A useful benchmark maps material class, wavelength, exposure regime, layer thickness, post-cure, temperature, geometry and endpoint range. Sparse regions and excluded states should be explicit. FAIR principles support reusable metadata [1], but benchmark fitness additionally requires a statement of what operating space is and is not represented.

4. Dataset shift is expected, not exceptional

Dataset shift occurs when training and deployment distributions differ [2]. Photopolymer systems are particularly exposed because commercial formulations change, machines age, optics are recalibrated and environmental conditions vary. Benchmark design should therefore include controlled out-of-distribution tests rather than assume independent and identically distributed deployment.

Table 2. Coverage map.

Axis	Record at minimum	Why
material	family, version/batch	chemical support
process	wavelength, exposure, layer, post-cure	state support
geometry	specimen type, orientation, scale	geometric support
environment	temperature, humidity/conditioning	deployment support
endpoint	definition, unit, method, uncertainty	label comparability

5. Metrics must correspond to engineering loss

Mean error can hide catastrophic failure near a qualification boundary. Evaluation should include calibration, worst-region performance, subgroup errors and decision-threshold consequences. If uncertainty estimates are produced, they should be assessed for calibration on shifted data, not only in the training distribution.

6. Benchmark governance matters after release

A public test set can become indirectly trained on through repeated submissions and model selection. Versioning, hidden holdouts, challenge resets and clear rules for external data reduce this risk. Provenance should record dataset version, corrections and any overlap discovered after release. Provenance models such as PROV-O make dataset derivations, transformations and version relationships explicit [3], while RO-Crate provides a practical packaging pattern for keeping data, metadata and contextual entities together [4].

Table 3. Benchmark fitness checks.

Check	Failure condition
duplicate/near-duplicate	related records cross split
distribution shift	test range narrower than claimed use
metadata completeness	state cannot be reconstructed
metric relevance	score weakly linked to engineering decision
version control	test labels or corrections change silently

7. Fitness-for-use decision

A benchmark supports a deployment claim only when its independence level and shift tests are at least as demanding as the intended use. High scores on an easier split do not transfer automatically. The benchmark should therefore publish a use matrix rather than a single leaderboard number.

8. Scope boundary and neighbouring reviews

ER-398 owns benchmark-dataset fitness, including coverage, leakage, distribution shift and evaluation design. ER-397 owns negative results and failed replication that may enter a dataset, while ER-399 owns value-of-information prioritisation of new experiments. ER-398 assesses whether data are fit for a modelling decision; it does not assign the research priority itself.

9. Leakage threat model for photopolymer datasets

Photopolymer data have several nested dependency levels, so a random row split is often invalid. Measurements from the same specimen share fabrication history; specimens from the same build share machine and exposure state; builds from the same batch share chemistry; repeated formulations may differ only slightly; and several datasets may contain reprocessed versions of the same original measurements. Kapoor and Narayanan document how data leakage can produce overoptimistic scientific claims and provide a taxonomy of leakage mechanisms across many ML-based fields [5]. For photopolymer benchmarks, the threat model should therefore be declared before splitting, with the deployment target defining the highest grouping level that must remain independent.

10. Split design should simulate the intended extrapolation

A benchmark for interpolation within one printer should not be used to claim transfer to another printer. Likewise, leaving out random specimens does not test transfer to a new formulation family. ER-398 therefore distinguishes specimen-held-out, build-held-out, batch-held-out, formulation-held-out, printer-held-out and laboratory-held-out splits. These are not interchangeable difficulty settings; they represent different deployment questions. Where data volume permits, nested splits can test several levels. Where it does not, the benchmark should state that the stronger transfer claim is untested rather than estimating it from an easier split.

11. Shift stress tests should be designed, not discovered accidentally

Dataset shift is expected when formulation revisions, optics, environmental conditions or post-cure histories change [2]. A robust benchmark can therefore reserve explicit stress-test partitions: another wavelength, another layer-thickness regime, a later formulation revision, another machine, or an aged specimen state. Performance degradation is then reported against the shift magnitude and engineering consequence. This is more informative than a single pooled score because it reveals where the model ceases to be reliable. Uncertainty estimates should be assessed on the shifted sets as well; a model that becomes confidently wrong outside the training distribution is especially dangerous near qualification boundaries.

12. Metrics need threshold and calibration views

Regression error averaged across a wide range can hide poor performance near a pass/fail threshold. Benchmark reports should therefore pair global metrics with threshold-local error, subgroup performance, calibration and the worst relevant operating region. Classification tasks should report the cost of false pass and false fail separately when those consequences differ. If repeated measurements exist, the irreducible measurement scatter provides context for model error: a model cannot be interpreted sensibly without knowing whether its error is small or large relative to the uncertainty of the reference measurement.

13. Benchmark lifecycle and contamination control

Once a public benchmark is used repeatedly, its test set can become part of model development indirectly through leaderboard feedback, manual tuning or incorporation into later training corpora. The benchmark therefore needs a lifecycle policy: immutable initial release, correction log, hidden or rotating holdout where appropriate, overlap checks against new training data and a rule for issuing a new major version. Provenance relations [3] and packaged contextual metadata [4] support this history. The objective is not to prevent reuse; it is to know when a score still represents an independent test rather than accumulated familiarity with the benchmark.

14. Reference-label quality can dominate model performance

A benchmark is only as meaningful as its target values. If tensile strength, conversion or dimensional error are measured with inconsistent methods, the model is asked to learn a mixture of physics and measurement-system differences. The benchmark should therefore document reference-method uncertainty, preprocessing and quality-control exclusions. Where multiple laboratories contribute labels, site

should be retained as metadata even if the final model does not use it. Label noise should be estimated where repeats exist, and performance should be interpreted relative to that noise floor rather than against a fiction of perfect ground truth.

15. Benchmark construction should precede model tuning

The benchmark protocol should freeze inclusion rules, grouping hierarchy, train/validation/test logic, metrics and primary comparison before extensive model tuning. Otherwise benchmark design can drift toward the model that happens to perform well. The test set should remain inaccessible to routine feature engineering where feasible, and every data correction after release should be versioned. If a corrected benchmark changes leaderboard ordering or qualification conclusions, the previous score remains historically valid for the old bytes but must not be silently compared with the new release.

16. Fitness-for-use matrix

A final benchmark report should include a matrix of intended and unsupported uses. Rows can represent deployment targets such as new specimen from known build, new build from known formulation, new batch, new formulation family, new printer or new laboratory. Columns can report whether the benchmark contains an independent test for that target, how large the shift is, and what performance or calibration criterion was met. This turns a leaderboard into an engineering qualification statement. A model is then 'fit for use' only for the rows actually exercised by the benchmark.

17. Hierarchical validation should mirror deployment

Photopolymer datasets naturally contain nested structure: measurements within specimens, specimens within builds, builds within batches, batches within formulations and formulations within material families. Validation should split at the highest level relevant to the intended deployment. A random row split may estimate interpolation among near-duplicates; leave-one-build-out tests process transfer; leave-one-batch-out tests material consistency; leave-one-formulation-family-out tests chemical extrapolation. Reporting several levels can reveal where apparent model performance collapses as the task becomes more realistic. The split hierarchy is therefore part of the engineering claim, not a secondary implementation detail. [2,5]

18. Baselines, ablations and uncertainty

A benchmark should include simple baselines appropriate to the task, such as mean prediction, linear model or physically motivated rule, so that model complexity is justified by measurable gain. Ablation tests should show whether performance depends on variables that will actually be available in deployment. Predictive uncertainty or calibration should be reported when decisions use thresholds, because two models with equal average error can produce very different risk near an acceptance boundary. If uncertainty grows under dataset shift, that behaviour is itself useful evidence and should not be hidden by a single aggregate score.

19. Benchmark release and anti-contamination controls

Once a benchmark becomes public, repeated tuning against its test set can turn the test into training information. Governance should therefore record benchmark versions, challenge submissions where relevant, and any public labels or examples that could leak into future model development. New benchmark releases should add genuinely independent data or rotate held-out groups rather than merely reshuffling old records. Where external foundation models may already have seen public data, the report should disclose that contamination cannot be ruled out. These controls preserve the distinction between a reusable demonstration set and an independent qualification test.

20. External validation and domain shift

Internal cross-validation cannot establish deployment performance in a new laboratory, printer family or material chemistry. At least one external validation set should be reserved when the intended use includes such transfer. The external set should differ along the same axis that matters operationally, not merely come from a random holdout. Performance loss should be reported alongside the magnitude of the shift in process and material coordinates. If no external dataset exists, the benchmark should say explicitly that extrapolation beyond the represented domain remains unqualified. [2,5]

21. Decision-centred error analysis

Aggregate metrics can hide the errors that matter most. Error analysis should stratify cases near engineering acceptance thresholds, by material family, process state and other high-consequence subgroups. False passes and false rejects can have different costs, so confusion matrices, calibration curves or threshold-specific losses may be more informative than a single R-squared or mean error. Representative failure cases should be retained to show whether the model breaks under known shifts or simply exhibits random residual error.

22. Reproducibility package for benchmark claims

A benchmark release should preserve the exact dataset version, split assignments, preprocessing code, metric implementation, baseline models and random seeds where relevant. If licensing prevents redistribution of raw data, stable identifiers and a deterministic reconstruction recipe should be provided. Every leaderboard or model comparison should name the benchmark version. This package allows later users to distinguish a genuine modelling advance from a change in data cleaning or split logic and makes benchmark conclusions auditable in the same way as experimental qualification results. [1,3,4]

23. Limitations and research priorities

There is currently no universal benchmark that represents the full chemical, machine and process diversity of photopolymer engineering. A major priority is therefore multi-site datasets designed around deployment shifts rather than convenience sampling. These should preserve hierarchical grouping, measurement uncertainty and exact material/process state, and reserve genuinely external test domains. Benchmark studies should compare random splits with build-, batch-, formulation- and laboratory-held-out evaluations to quantify the optimism created by leakage. Public benchmarks also need governance against repeated test-set tuning and foundation-model contamination. ER-398 consequently treats benchmark scores as conditional evidence tied to one release and one deployment target, not as transferable model qualifications in themselves. [1-5]

References

1. Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*. 2016;3:160018. doi:10.1038/sdata.2016.18.
2. Quinonero-Candela J, Sugiyama M, Schwaighofer A, Lawrence ND, editors. *Dataset Shift in Machine Learning*. MIT Press; 2009. doi:10.7551/mitpress/9780262170055.001.0001.
3. Lebo T, Sahoo S, McGuinness D, et al. PROV-O: The PROV Ontology. W3C Recommendation; 30 April 2013. <https://www.w3.org/TR/prov-o/>.
4. Sefton P, O Carragain E, Soiland-Reyes S, editors. *RO-Crate Metadata Specification 1.1*. Research Object Crate Community; 2023. doi:10.5281/zenodo.7867028.
5. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*. 2023;4(9):100804. doi:10.1016/j.patter.2023.100804.

Competing interests

The author is Founder, CEO and CTO of 3Dresyns by Resyner Technologies S.L.; ER-398 is a methodological engineering review and does not provide a comparative commercial recommendation.

Data availability

ER-398 is a literature-based engineering review; no new experimental dataset was generated.