Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

AI-Driven Data-Centric Computing for Large-Scale Scientific Research: Automated Data Quality Assessment, Feature Discovery, Knowledge Extraction, and Reproducible Machine Learning Pipelines

Syed Muhammad Raza Zaidi
Department of Computer Systems Engineering, NED University of Engineering and Technology, Karachi, Sindh, Pakistan Email: mraza28067@gmail.com

Muhammad Yasir Ahmed Abbasi
Department of Communication Systems and Network, Master of Information Technology, Sindh Agriculture University, Tandojam, Pakistan
Email: yasir110@gmail.com

Shah Wali Ullah
Ph.D. Student, Advanced Functional Materials (AFM) Laboratory, Engineering Physics Department, Institut Teknologi Bandung, Bandung 40132, Indonesia
Email: 33323751@mahasiswa.itb.ac.id

Zeeshan Ali
Mirpur University of Science and Technology (MUST), Azad Jammu and Kashmir, Pakistan Email: alimehduvi313@gmail.com

ABSTRACT

Machine learning is now routine in scientific research, but the reliability of its results depends at least as much on the data as on the model. Label errors, missing values, gross measurement errors and duplicated records are common in scientific datasets, and they are rarely measured before a model is trained. This study presents and evaluates an open, end-to-end data-centric pipeline that combines automated data-quality assessment, feature discovery, knowledge extraction and reproducibility controls, and tests it on six public scientific datasets from astroparticle physics, chemistry, medicine, materials science, civil engineering and geography (569 to 21,263 records). An automated audit of the unmodified data found that 40.1% of the records in the superconductivity dataset share an identical feature vector with at least one other record, with 2,375 groups carrying conflicting critical temperatures, and that a conventional robust z-score rule flags 9.9–57.5% of rows as outliers, which makes it unsuitable as an unsupervised filter. In controlled corruption experiments, cross-validated confident learning ranked injected label errors with an area under the ROC curve (AUROC) of 0.78–0.99, and filtering the flagged samples raised test accuracy at 30% noise from 0.705 to 0.787 (QSAR biodegradation) and from 0.809 to 0.899 (breast cancer). Model-based imputation reduced the normalised reconstruction error on superconductivity from 1.00 (mean imputation) to 0.23, while native missing-value handling in gradient boosting matched imputation in predictive accuracy on the

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

smaller datasets. Local outlier factor detected injected gross errors best (AUROC 0.82–0.91), and robust filtering restored the concrete R^2 from 0.19 to 0.90 at 5% corruption. Near-duplicates inflated reported R^2 by up to 0.036 under random splitting compared with group-aware splitting. Automated feature discovery raised the cross-validated R^2 of a linear model on concrete strength from 0.605 to 0.873, and a log-linear sparse regression recovered the exponents of six physical laws with a maximum error below 0.01 at 10% noise. Knowledge extraction identified the water-to-cement ratio as the first decision rule for concrete strength and thermal-conductivity descriptors as dominant for superconductivity, in agreement with domain knowledge, with importance-rank correlations of 0.95 and 0.69 across bootstrap retraining. Fixed seeds made ten repeated runs bit-for-bit identical, whereas uncontrolled runs varied by up to 0.053 in test score, mostly because of the data split. All quality checks scaled close to linearly with data size. The results show that data-quality assessment, group-aware evaluation and provenance tracking change scientific conclusions measurably and should be treated as standard parts of machine-learning-based research.

Keywords: Data-Centric Artificial Intelligence; Data Quality; Label Noise; Confident Learning; Missing Data; Outlier Detection; Data Leakage; Feature Discovery; Knowledge Extraction; Reproducibility; Scientific Machine Learning

Introduction

Background and motivation

Scientific research has become data-intensive. Instruments, simulations and databases produce records faster than they can be inspected by hand, and machine learning is increasingly used to classify events, predict properties and screen candidates [1, 2]. In this setting the quality of the data sets an upper limit on the quality of the science. A model trained on mislabelled examples learns the errors, a model evaluated on duplicated records reports accuracy that it will not reach on new data, and a result that cannot be rerun cannot be checked [3–5].

For most of the past decade, progress in machine learning was measured by improving models on fixed benchmark datasets. Data-centric artificial intelligence reverses the emphasis: it treats the dataset as the main object of engineering and asks how data can be measured, cleaned, augmented and documented systematically [6–8].

The motivation is practical. Field studies show that data problems propagate through the stages of a project and cause failures long after they are introduced [9], and the technical debt of production machine-learning systems is dominated by data dependencies rather than model code [10]. In science the same problems appear as irreproducible results. A survey of more than 1,500 researchers found that a majority had failed to reproduce another group's experiment [11], and a recent review identified data leakage as a cause of errors in hundreds of machine-learning-based scientific papers across many fields [3].

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

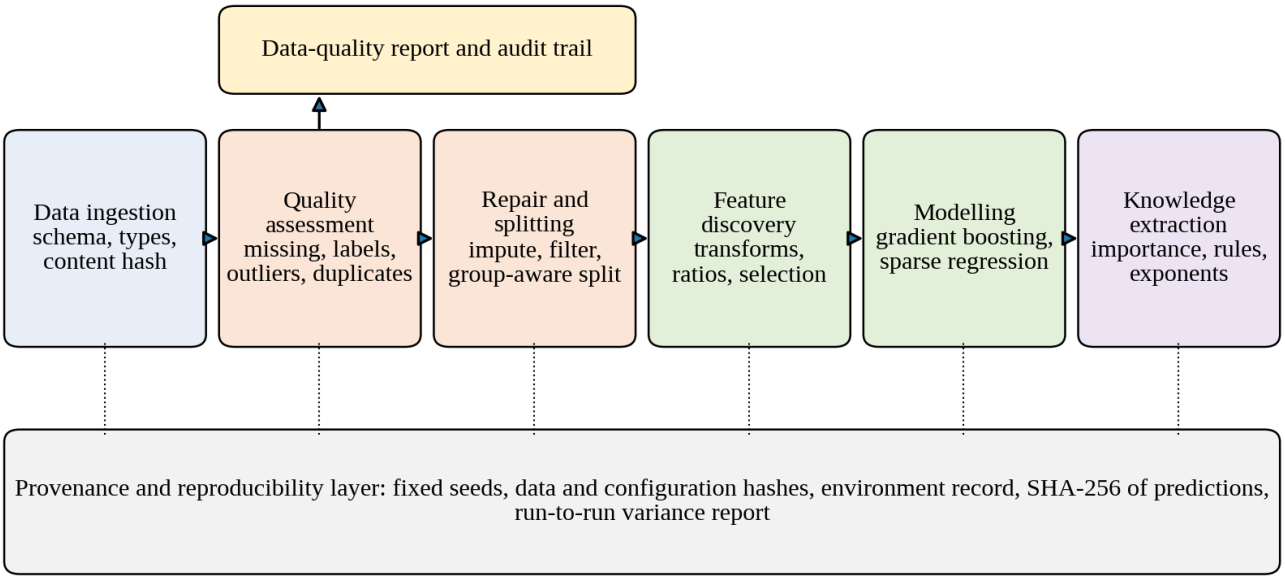


Figure 1. Architecture of the data-centric pipeline evaluated in this study. Six processing stages are connected by a provenance and reproducibility layer that records hashes, seeds and the computing environment at every stage.

Problem statement

Methods exist for each individual data problem: confident learning for label errors [12], multiple imputation for missing values [13], isolation forests and density-based scores for outliers [14, 15], group-aware splitting for leakage [16], automated feature synthesis [17], sparse regression for equation discovery [18] and experiment tracking for reproducibility [19]. In practice, however, scientists rarely apply them together, and published studies seldom report how much the data problems changed the results. Researchers therefore have little quantitative guidance on which checks matter most, how reliable automated detectors are on real scientific data, and what each repair costs or gains.

Research gap

Three gaps motivate this work. First, most evaluations of data-cleaning methods use a single error type on generic benchmarks, whereas scientific datasets combine skewed distributions, correlated descriptors, small samples and domain-specific duplication [7, 20]. Second, studies of feature discovery and equation discovery usually report predictive accuracy or the recovery of a known law, but rarely both in the same framework, and rarely together with a check on whether the extracted knowledge is stable under resampling [21, 22]. Third, reproducibility is discussed mainly as a policy question [2, 23]. The size of run-to-run variation and the relative contributions of the data split and model randomness are seldom measured on scientific tasks [24, 25].

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Objectives and research questions

The objective of this study is to design an integrated data-centric pipeline for scientific machine learning and to quantify, on real scientific datasets, how its components affect the reliability of results. We address four research questions. RQ1: how prevalent are data-quality problems in widely used scientific datasets, and how accurately can they be detected automatically? RQ2: how much do label noise, missing values, gross errors and duplication change model performance and reported performance, and how much of the loss can automated repair recover? RQ3: can automated feature discovery and knowledge extraction produce features, rules and laws that are both predictive and scientifically meaningful? RQ4: how large is run-to-run variability, what causes it, and what does it cost to make a pipeline reproducible and scalable?

Contributions

The study makes five contributions: (i) an integrated pipeline architecture (Figure 1) in which every stage is logged and can be verified by hashing; (ii) an audit of six public scientific datasets that reveals extensive duplication with conflicting targets in a widely used materials dataset; (iii) controlled experiments, with five or three repetitions, on label noise, missing data, gross errors and duplicate leakage; (iv) an evaluation of feature discovery and knowledge extraction against known domain relationships, including six physical laws; and (v) a quantitative analysis of reproducibility and scalability, ending with practical recommendations for scientists who use machine learning.

Literature Review

Data-centric artificial intelligence

Data-centric AI has been defined as the discipline of systematically engineering the data used to build AI systems [6, 8]. Surveys organise it into training-data development (collection, labelling, cleaning, augmentation), evaluation-data development and data maintenance [6, 7]. The approach builds on older work on data quality, which defines quality by its fitness for use and divides it into dimensions such as accuracy, completeness, consistency and timeliness [20, 26]. Documentation frameworks such as datasheets for datasets and model cards add the requirement that the origin, composition and limitations of data and models be reported [27, 28]. In science, the FAIR principles require data to be findable, accessible, interoperable and reusable, which supports reproducibility but does not by itself guarantee that the data are correct [29].

Automated data validation and quality assessment

Industrial systems validate data against schemas and statistical expectations before training. Google's data validation system checks feature types, ranges and distribution drift [30], and Deequ expresses quality constraints as unit tests that run on large datasets in a distributed engine [31]. These tools detect schema violations and drift well but are less suited to errors that only become visible relative to a model, such as mislabelled examples, or to duplication that crosses a train–test split [7, 16]. Robust

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

statistics provide the classical tools for outliers: the median absolute deviation (MAD) and the modified z-score, with a common cut-off of 3.5, are resistant to the outliers they are meant to detect [32, 33]. Model-based detectors such as isolation forest, which isolates anomalies with random partitions [14], and the local outlier factor (LOF), which compares the local density of a point with that of its neighbours [15], extend detection to multivariate structure.

Label noise and confident learning

Label noise reduces accuracy, increases the required sample size and distorts estimates of feature importance [34]. Early work filtered training sets by removing instances misclassified by an ensemble [35]. Confident learning estimates the joint distribution of given and true labels from out-of-sample predicted probabilities and uses per-class thresholds to identify likely errors [12]. Applied to ten widely used benchmarks, it found an average of at least 3.3% label errors in the test sets, enough to change the ranking of models [4]. Methods for deep networks add robust losses and sample selection [36]. Most evaluations use image or text data; tabular scientific data, with small samples and class overlap, have received less attention.

Missing data

Rubin's framework distinguishes data missing completely at random (MCAR), at random (MAR) and not at random (MNAR) [37]. Simple mean imputation preserves the sample size but shrinks variance and weakens correlations. Nearest-neighbour imputation was popularised for gene-expression data [38], and multiple imputation by chained equations models each variable conditionally on the others [13]. Random-forest imputation handles mixed data and nonlinear relations [39]. Modern gradient-boosting libraries can learn a default direction for missing values at each split, which avoids imputation altogether [40, 41]. Which approach is best for a scientific dataset depends on the correlation structure of the variables and on whether the goal is reconstruction or prediction.

Duplicates and data leakage

Leakage occurs when information that would not be available at prediction time enters training or model selection [16]. Duplicated or near-duplicated records across the training and test sets are a common form: deduplication changes the measured performance of language models [42] and of models of source code [43]. Kapoor and Narayanan catalogued eight types of leakage and showed that they affect scientific papers in at least seventeen fields [3]. Group-aware splitting, in which all records that share an identity are placed on the same side of the split, is the standard remedy, but it requires the groups to be detected first.

Feature discovery and equation discovery

Feature engineering remains one of the most time-consuming steps in applied machine learning. Deep feature synthesis generates features automatically by stacking aggregation and transformation operators over relational data [17], and automated machine-learning systems search over preprocessing and model choices [44, 45]. Selection is needed to control the resulting growth in dimensionality; filter, wrapper

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

and embedded methods, including mutual information and the lasso, are standard [46, 47]. In the physical sciences the aim is often an equation rather than a predictor. Symbolic regression recovered conservation laws from motion-tracking data [48], sparse identification of nonlinear dynamics (SINDy) selects terms from a library of candidate functions [18], and AI Feynman combines neural fitting with physics-inspired simplifications to recover all 100 equations of a Feynman benchmark [21].

Knowledge extraction and interpretability

Knowledge extraction aims to turn a trained model into statements that a scientist can check. Global surrogates approximate a complex model with an interpretable one, such as a shallow decision tree [49]. Permutation importance measures the loss of performance when a feature is shuffled [22, 50], and additive attribution methods such as SHAP explain individual predictions [51]. Local surrogates such as LIME fit simple models around single predictions [52]. Rudin argued that for high-stakes decisions inherently interpretable models should be preferred to post-hoc explanations [53]. A recurring concern is stability: explanations that change under small perturbations of the data cannot support scientific claims [22].

Reproducibility in machine-learning research

A study of 400 AI papers found that few documented enough of their method, data and code to be reproduced [23]. The NeurIPS 2019 reproducibility programme introduced a checklist and code submission policy [2]. Even with code available, results vary with random seeds: Bouthillier et al. showed that conclusions drawn from single runs can be overturned by variance [24], and a systematic scan of seeds for a vision model found accuracy differences large enough to be reported as improvements [25]. Experiment-tracking platforms such as MLflow record parameters, code versions and artefacts [19], and notebooks support literate computing [54], but neither guarantees that a run can be repeated bit for bit.

Data quality in scientific domains

Scientific data have features that make generic cleaning rules risky. Measurements often follow skewed or heavy-tailed distributions, so values that are rare are not necessarily wrong [33]. Descriptors computed from structures or formulae, as in cheminformatics and materials informatics, are strongly correlated and can map distinct entities to identical inputs [55, 56]. Labels may come from simulations, expert annotation or threshold rules applied to noisy measurements, each with its own error pattern [34, 57]. Sample sizes are often small relative to the number of descriptors, which makes cross-validated detectors noisier [12]. Finally, scientific conclusions are often drawn from feature importances or fitted coefficients, not from predictions, so data problems that barely change accuracy can still change the conclusion [22, 53]. Evaluations of data-centric methods should therefore include scientific datasets and measure effects on interpretation as well as on accuracy.

Synthesis and research gap

Table 1 compares representative systems. Each addresses one or two stages well, but none combines quality assessment, repair, feature and equation discovery, knowledge

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

extraction and verifiable reproducibility in one evaluated pipeline, and few report how much each stage changes the scientific result. This study addresses that gap with the pipeline in Figure 1 and the experiments described in Section 3.

Table 1. Comparison of representative data-centric tools and methods with the pipeline evaluated in this study.

System or method	Quality assessment	Repair	Feature / equation discovery	Knowledge extraction	Reproducibility
Data validation (TFX) [30]	Schema, drift	No	No	No	Partial
Deequ [31]	Constraint tests	No	No	No	Partial
Confident learning [12]	Label errors	Filtering	No	No	No
MICE / MissForest [13, 39]	Missingness	Imputation	No	No	No
Deep feature synthesis [17]	No	No	Features	No	No
Auto-sklearn [44]	No	Imputation	Preprocessing search	No	Partial
SINDy / AI Feynman [18, 21]	No	No	Equations	Equations	No
MLflow [19]	No	No	No	No	Tracking
This study	Labels, missingness, outliers, duplicates	Filtering, imputation, group split	Features and power laws	Importance, rules, exponents	Hashes, seeds, variance report

3. Methodology

3.1 Research design

The study is an empirical, computational evaluation. We implemented the pipeline in Figure 1 and applied it to public datasets in three modes: an audit of the unmodified data (RQ1), controlled corruption experiments in which known errors were injected into clean training data so that detection and repair could be scored against ground truth (RQ1, RQ2), and analyses of feature discovery, knowledge extraction, reproducibility and scalability (RQ3, RQ4). All test sets were left uncorrupted, so that differences in test performance can be attributed to the training data.

3.2 Datasets

Six datasets were obtained from OpenML [58] and represent different scientific domains, sizes and data types (Table 2). MAGIC contains simulated Cherenkov telescope image parameters for separating gamma-ray signals from hadronic background [57]. QSAR-biodeg contains 41 molecular descriptors for ready biodegradability [56]. WDBC contains 30 cell-nucleus features from breast-mass images [59]. Superconductivity contains 81 descriptors derived from the chemical

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

formulae of 21,263 superconductors and their critical temperatures [55]. Concrete contains mixture proportions and age for compressive strength [60]. California housing contains census block-group data [61], with one-hot encoding of the categorical ocean-proximity variable.

Table 2. Datasets used in this study.

Dataset	Domain	Records	Features	Task and target	Source
MAGIC	Astroparticle physics	19,020	10	Classification, gamma vs hadron	[57]
QSAR-biodeg	Chemistry	1,055	41	Classification, ready biodegradability	[56]
WDBC	Medicine	569	30	Classification, malignant vs benign	[59]
Superconductivity	Materials science	21,263	81	Regression, critical temperature (K)	[55]
Concrete	Civil engineering	1,030	8	Regression, compressive strength (MPa)	[60]
California housing	Geography, economics	20,640	13	Regression, median house value	[61]

All datasets were retrieved from OpenML [58]. California housing has 13 features after one-hot encoding.

Automated data-quality assessment

The audit computed, for each dataset, the proportion of missing cells; exact duplicate rows; rows whose feature vectors (rounded to ten decimals and hashed) occur more than once; duplicate groups whose targets disagree; constant features; feature pairs with absolute Pearson correlation above 0.95; and the proportion of rows with at least one modified z-score above 3.5, where $z = 0.6745|x - \text{median}|/\text{MAD}$ [32]. For classification data, suspected label errors were identified by confident learning [12]: out-of-sample class probabilities p were obtained by five-fold stratified cross-validation with a histogram gradient-boosting classifier, the self-confidence of each sample was p(given label), and a sample was flagged when its self-confidence fell below the mean self-confidence of its class and the predicted class differed from the given one.

Controlled corruption experiments

Label noise. Each classification dataset was split 70/30 with stratification. A fraction $r \in \{0.05, 0.10, 0.20, 0.30\}$ of training labels was flipped at random, the confident-learning detector was applied, and three models were trained: on noisy labels, on noisy labels after removing flagged samples, and on the original labels. Detection was scored by precision, recall, F1 and AUROC of the self-confidence ranking. Five random seeds were used.

Missing values. Values were removed completely at random at rates of 10–40% from both training and test features of the three regression datasets. Four strategies were

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

compared: native missing-value handling in gradient boosting, mean imputation, k-nearest-neighbour imputation with k = 5 [38] and iterative imputation with Bayesian ridge regression and five iterations [13]. Reconstruction was scored by the root-mean-square error of imputed values divided by the feature standard deviation (nRMSE), and prediction by test R^2. Superconductivity and California housing were subsampled to 6,000 records to keep the KNN and iterative imputers tractable; three seeds were used.

Gross errors. In 1–10% of training rows one of three errors was injected with equal probability: a feature multiplied by 10, 100 or 0.01 (a unit or scale error), a feature replaced by its negative minus three standard deviations (a sign or offset error), or the target multiplied by 10 or 0.1. Three detectors were applied to the joint feature–target matrix: the maximum modified z-score, isolation forest with 200 trees [14] and LOF with 20 neighbours [15]. Detection was scored by AUROC and average precision, and repair by removing the 1.5r highest-scoring rows and retraining.

Duplicates and leakage. Natural duplication in superconductivity was assessed by comparing a random 70/30 split with a group-aware split in which all records with an identical feature vector were kept together [16]. In concrete and California housing, near-duplicates were injected at 10–40% of the original size by copying random rows and adding Gaussian noise of 0.001 standard deviations; reported R^2 under random and group-aware splits was compared, and a hashing detector based on standardised values rounded to two decimals was scored.

Feature discovery

For each regression dataset the eight features with the highest mutual information with the target were selected; candidate features were then generated as pairwise products and ratios of these eight and as $\log(1 + x)$ of every non-negative feature, and the 60 candidates with the highest mutual information were retained [17, 46]. Raw and discovered feature sets were compared with ridge regression (regularisation chosen by internal cross-validation) and gradient boosting, using five-fold cross-validated R^2.

Equation discovery was tested on six laws with known exponents: Newton's law of gravitation, Coulomb's law, Kepler's third law, the ideal-gas law, the period of a simple pendulum and kinetic energy. For each law, 500 samples were drawn with variables log-uniform on [0.5, 50], multiplicative log-normal noise of 1%, 5% or 10% was applied to the output, and three irrelevant variables were added. Taking logarithms turns a power law into a linear model, so the exponents were estimated by lasso regression on the logged variables, with the penalty chosen by five-fold cross-validation [18, 47]. We report the maximum absolute exponent error and the largest coefficient assigned to an irrelevant variable over five seeds.

Knowledge extraction

For superconductivity (8,000-record subsample) and concrete, a gradient-boosting model was trained on eight bootstrap resamples of the training set. Permutation importance on the test set (three repeats) was computed for each resample [22, 50]. Stability was measured by the mean pairwise Spearman correlation of the importance vectors and by the mean Jaccard index of the top-10 feature sets. Global surrogates

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

were obtained by fitting decision trees of depth 1–8 to the predictions of the full model [49]; fidelity is the R^2 between surrogate and model predictions on the test set, and accuracy is the R^2 of the surrogate against the true target. For concrete, the water-to-cement ratio was included as a candidate feature because of its central role in concrete science [60].

Reproducibility, provenance and scalability

Every dataset and prediction vector was fingerprinted with SHA-256. For four datasets, the complete train–evaluate pipeline was run ten times with fixed seeds (split and model seed 42) and ten times with seeds drawn at random. Variance was decomposed by varying only the split seed or only the model seed. The model used early stopping with an internal validation fraction of 0.1, which makes it sensitive to its seed. Scalability was measured on synthetic data with 20 features, 5% missing cells and 10^3 to 10^6 rows; the slope of log time against log size estimates the empirical complexity of each check.

Evaluation metrics

Detection performance was measured by precision $P = TP/(TP + FP)$, recall $R = TP/(TP + FN)$ and their harmonic mean $F1 = 2PR/(P + R)$, where TP, FP and FN count correctly flagged, wrongly flagged and missed corrupted records. Because detectors produce scores rather than decisions, threshold-free ranking quality was measured by the area under the receiver operating characteristic curve (AUROC), the probability that a randomly chosen corrupted record receives a higher score than a randomly chosen clean one, and by average precision, which weights the ranking towards its top. Regression performance was measured by the coefficient of determination $R^2 = 1 - \Sigma(y - \hat{y})^2/\Sigma(y - \bar{y})^2$ and the mean absolute error, and classification by accuracy. Imputation quality was measured by nRMSE = [mean over masked cells of $((\hat{x} - x)/s_j)^2]^{1/2}$, where s_j is the training standard deviation of feature j, so that nRMSE = 1 corresponds to the error of a constant prediction. Stability of importance rankings was measured by the Spearman rank correlation between importance vectors and by the Jaccard index $|A \cap B|/|A \cup B|$ of top-10 feature sets A and B [22].

Implementation and statistical reporting

All experiments were implemented in Python 3.12 with scikit-learn 1.8 [62], NumPy 2.4, pandas 3.0 and SciPy 1.17, and executed on a single x86-64 processor core with 3 GB of memory. Gradient boosting used scikit-learn's histogram-based implementation, which follows the LightGBM algorithm [40]. Results are reported as means over seeds, with standard deviations where shown. No hyperparameters were tuned on test data.

Results

Data-quality audit of unmodified scientific datasets

Figure 2 and Table 3 summarise the audit. Superconductivity stands out: 8,519 of its 21,263 records (40.1%) share a feature vector with at least one other record, and 2,375 such groups have different critical temperatures. The cause is the construction of the

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

dataset: its 81 descriptors are statistics of elemental properties computed from the chemical formula [55], so materials with the same elements in proportions that give the same statistics, or measured several times, receive identical inputs. A model cannot separate such records, which places a floor under the achievable error and creates a channel for leakage. Concrete has 25 exact duplicate rows and nine conflicting groups, MAGIC has 115 exact duplicates, and California housing contains the dataset's only missing values (207 cells, 0.08%).

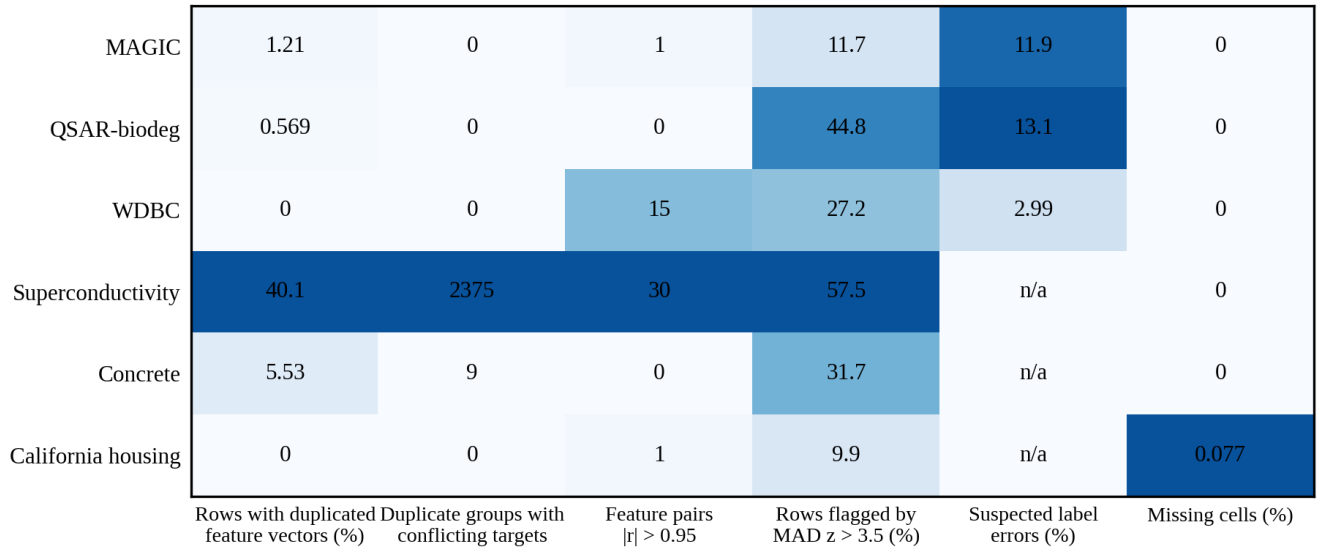


Figure 2. Automated data-quality audit of the six unmodified datasets. Cell shading is normalised within each column; values are counts or percentages as labelled. n/a, not applicable to regression targets.

The modified z-score rule flagged 9.9–57.5% of rows as containing at least one outlier. Scientific descriptors are often skewed, discrete or heavy-tailed, and on such variables the rule labels legitimate values as extreme; on QSAR-biodeg, many of whose descriptors are integer counts, it flagged 44.8% of rows. Used as an automatic filter, it would discard a large share of valid data. Confident learning flagged 11.9% of MAGIC and 13.1% of QSAR-biodeg samples as possible label errors, but only 3.0% of WDBC. These figures should be read as rates of ambiguous samples, not confirmed errors: MAGIC is known to have overlapping classes [57], and the controlled experiments below show that the detector's precision is limited when true noise is low.

Table 3. Results of the automated audit of the unmodified datasets.

Dataset	Missing cells (%)	Exact duplicate rows	Rows in duplicated feature groups	Conflicting groups	Pairs |r| > 0.95	MAD-flagged rows (%)	Suspected label errors (%)
MAGIC	0	115	230	0	1	11.7	11.9
QSAR-biodeg	0	3	6	0	0	44.8	13.1
WDBC	0	0	0	0	15	27.2	3.0

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Dataset	Missing cells (%)	Exact duplicate rows	Rows in duplicated feature groups	Conflicting groups	Pairs |r| > 0.95	MAD-flagged rows (%)	Suspected label errors (%)
Superconductivity	0	69	8,519	2,375	30	57.5	n/a
Concrete	0	25	57	9	0	31.7	n/a
California housing	0.08	0	0	0	1	9.9	n/a

Detection and repair of label noise

Confident learning ranked injected errors well on all three datasets (Figure 3a, Table 4). AUROC was 0.91–0.94 on MAGIC, 0.78–0.91 on QSAR-biodeg and 0.89–0.99 on WDBC, decreasing as the noise rate rose. Recall stayed between 0.70 and 0.94. Precision, by contrast, rose with the noise rate, from 0.22–0.52 at 5% noise to 0.51–0.71 at 30%. At low noise rates the flagged set is dominated by naturally ambiguous samples, which is consistent with the audit. For binary tasks the per-class threshold gave almost the same F1 as the simpler rule of flagging every misclassified sample (difference below 0.005).

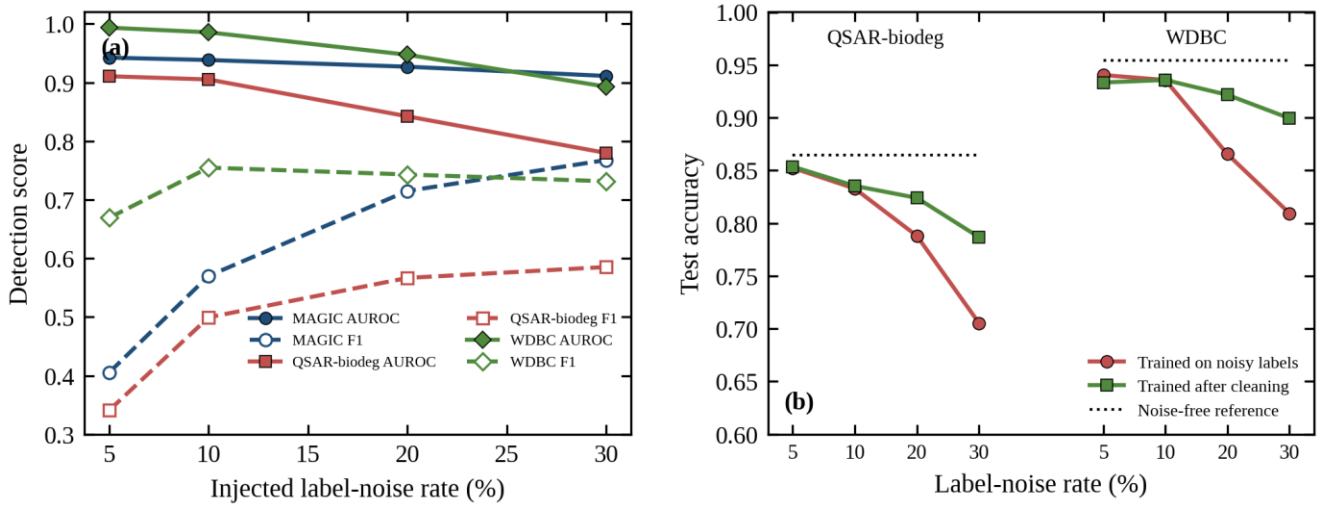


Figure 3. Label-noise experiments. (a) AUROC of the self-confidence ranking and F1 of confident-learning flags as a function of the injected noise rate. (b) Test accuracy of gradient boosting trained on noisy labels, after removal of flagged samples, and on noise-free labels. Means of five seeds.

The benefit of cleaning depended on the noise level (Figure 3b). At 30% noise, filtering raised test accuracy from 0.705 to 0.787 on QSAR-biodeg and from 0.809 to 0.899 on WDBC, recovering 52% and 62% of the accuracy lost to noise. At 20% the gains were 0.036 and 0.057. At 5–10% noise, however, filtering gave no gain or a small loss (for example 0.876 to 0.871 on MAGIC), because removing 16–20% of the training data, most of it correctly labelled but difficult, cost more than the errors it removed. Gradient boosting is fairly robust to low levels of symmetric noise, so automated filtering should be applied when the estimated noise is high or when the

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

detector's flags are reviewed by a domain expert.

Table 4. Detection performance of confident learning for injected label noise (means of five seeds).

Dataset	Noise rate	Precision	Recall	F1	AUROC	Removed (%)
MAGIC	0.05	0.264	0.873	0.405	0.942	16.2
MAGIC	0.10	0.424	0.872	0.570	0.938	20.3
MAGIC	0.20	0.609	0.865	0.715	0.927	28.2
MAGIC	0.30	0.712	0.833	0.768	0.911	34.9
QSAR-biodeg	0.05	0.216	0.814	0.341	0.911	18.9
QSAR-biodeg	0.10	0.356	0.839	0.499	0.905	24.4
QSAR-biodeg	0.20	0.456	0.754	0.567	0.842	34.1
QSAR-biodeg	0.30	0.505	0.696	0.585	0.780	39.7
WDBC	0.05	0.520	0.941	0.669	0.993	8.4
WDBC	0.10	0.635	0.939	0.755	0.986	14.3
WDBC	0.20	0.644	0.878	0.743	0.947	27.6
WDBC	0.30	0.656	0.830	0.731	0.893	38.1

Standard deviations of F1 were at most 0.008 for MAGIC, 0.047 for QSAR-biodeg and 0.046 for WDBC.

Missing values

Prediction accuracy fell steadily with the missing rate for every strategy (Figure 4), but the best strategy depended on the dataset. On superconductivity, whose descriptors are strongly correlated (30 pairs with $|r| > 0.95$), iterative imputation reconstructed the removed values with a mean nRMSE of 0.23, compared with 0.53 for KNN and 1.00 for mean imputation, and, together with KNN imputation, gave the highest test R^2 at every rate (0.884 at 10% and 0.846 at 40%, against 0.868 and 0.808 for native handling). On concrete and California housing, whose features are weakly correlated, the strategies differed by less than 0.04 in R^2, apart from KNN imputation on California housing, and at 30–40% missingness native handling was equal to or better than imputation (0.579 against 0.537 for iterative imputation on California housing at 40%).

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

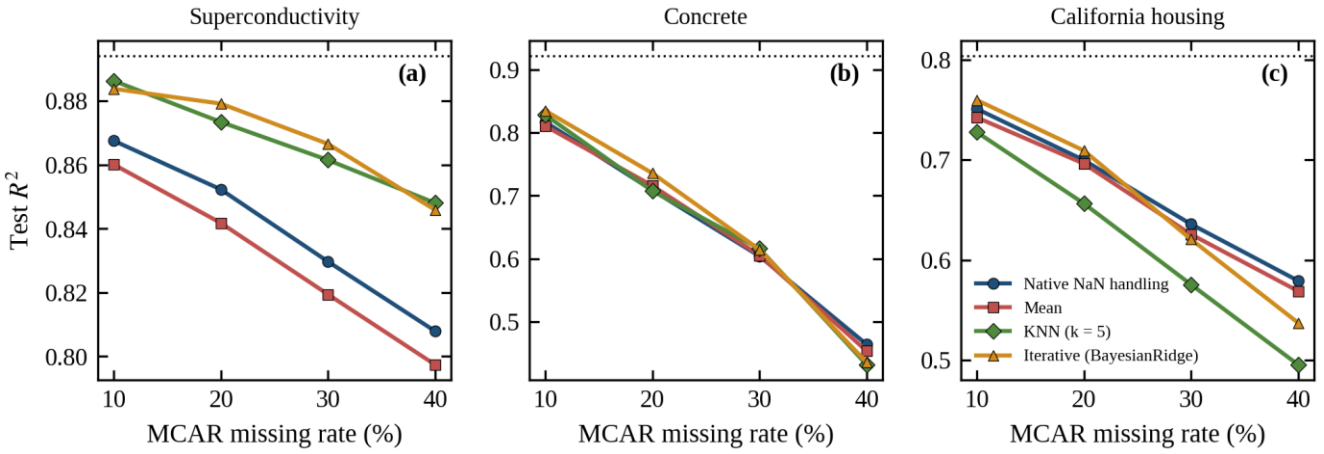


Figure 4. Test R^2 as a function of the MCAR missing rate for four strategies on (a) superconductivity, (b) concrete and (c) California housing. Dotted lines show the R^2 with complete data. Means of three seeds.

KNN imputation performed worst on California housing (nRMSE 0.97), because its distance metric was dominated by large-valued features such as population and number of rooms. The result is a reminder that the default settings of imputation methods assume comparable scales. Iterative imputation was also the most expensive method on the high-dimensional dataset, at 7.7 s per run against 0.8 s for mean imputation (Table 5). Good reconstruction does not guarantee good prediction: on California housing, iterative imputation had the lowest nRMSE (0.57) but was not the best predictor at high missing rates.

Table 5. Reconstruction error and run time of imputation strategies (means over missing rates and seeds).

Dataset	Mean nRMSE	KNN nRMSE	Iterative nRMSE	Mean time (s)	KNN time (s)	Iterative time (s)	Native time (s)
Superconductivity	1.001	0.530	0.229	0.82	7.04	7.67	1.23
Concrete	0.988	0.791	0.744	0.13	0.17	0.19	0.13
California housing	0.961	0.968	0.572	0.22	1.84	0.42	0.26

nRMSE: root-mean-square error of imputed values divided by the feature standard deviation; native handling does not impute. Times include model training.

Gross errors and outliers

Injected gross errors damaged models severely (Figure 5b). On concrete, 5% corrupted rows lowered test R^2 from 0.922 to 0.189, and 10% to 0.076, mainly because of target-scale errors, which gradient boosting with squared loss fits faithfully. LOF ranked corrupted rows best on all three datasets, with mean AUROC of 0.91 (concrete), 0.88 (California housing) and 0.82 (superconductivity) (Figure 5a). The MAD-based score reached 0.67–0.84, and isolation forest was close to chance on

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

superconductivity (0.53), where 81 dimensions dilute the effect of a single corrupted feature on random partitions.

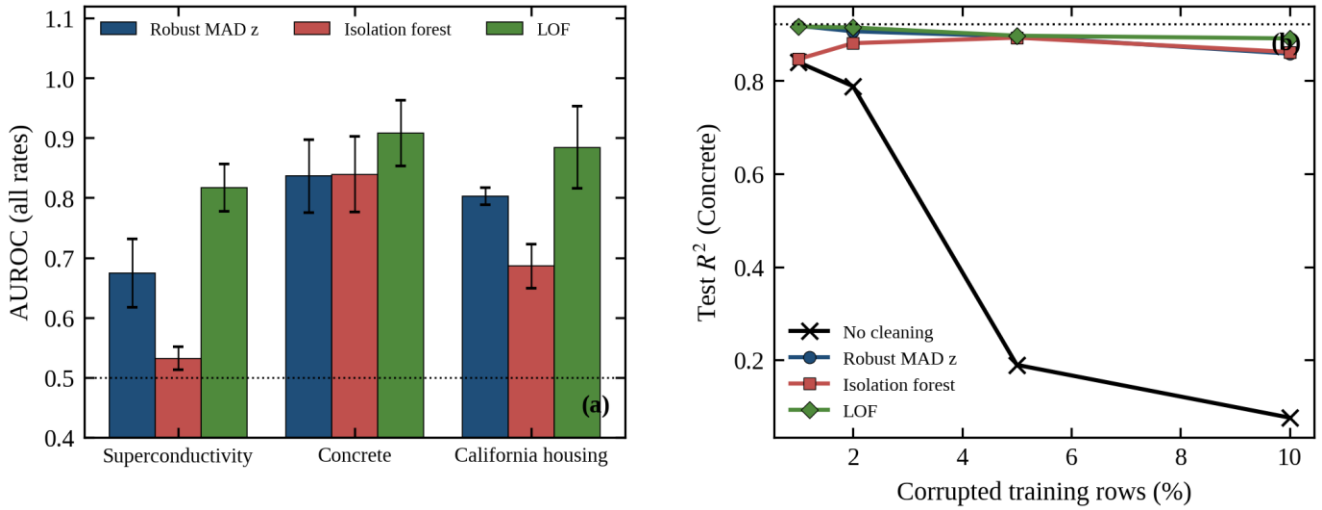


Figure 5. Gross-error experiments. (a) AUROC of three detectors averaged over corruption rates of 1–10% (error bars, standard deviation). (b) Test R² on concrete without cleaning and after removing the 1.5r highest-scoring rows. The dotted line shows the clean-data R².

Detection quality and repair quality did not rank the same way. Table 6 gives the full results. When the 1.5r top-scoring rows were removed, the MAD score restored concrete to R² = 0.895 at 5% corruption and California housing to 0.818–0.826 at all rates, close to the clean values of 0.922 and 0.827. LOF gave the best repair at low corruption but failed at 10% on California housing (0.459) and superconductivity (0.227), where it ranked target errors low and spent its removal budget on unusual but valid rows. Because single-feature and target-scale errors are extreme in one coordinate, the coordinate-wise robust score finds them reliably even though it has lower overall AUROC. A combination of a coordinate-wise robust score for gross errors and a multivariate density score for subtler anomalies is therefore advisable. Without a defined removal budget, the unsupervised MAD rule should not be used, as Section 4.1 showed.

Table 6. Test R² after removing the 1.5r highest-scoring training rows for each detector (means of three seeds).

Dataset	Corrupted rows (%)	Clean data	No cleaning	Robust MAD z	Isolation forest	LOF
Superconductivity	1	0.892	0.847	0.884	0.838	0.885
Superconductivity	5	0.892	0.524	0.726	0.515	0.625
Superconductivity	10	0.892	0.386	0.670	0.365	0.227
Concrete	1	0.922	0.840	0.918	0.846	0.917
Concrete	5	0.922	0.189	0.895	0.893	0.897
Concrete	10	0.922	0.076	0.858	0.862	0.891
California housing	1	0.827	0.776	0.826	0.780	0.814

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Dataset	Corrupted rows (%)	Clean data	No cleaning	Robust MAD z	Isolation forest	LOF
California housing	5	0.827	0.682	0.825	0.744	0.717
California housing	10	0.827	0.542	0.818	0.769	0.459

Mean AUROC over all rates: MAD 0.674, 0.836, 0.803; isolation forest 0.532, 0.839, 0.686; LOF 0.817, 0.908, 0.884 for superconductivity, concrete and California housing, respectively.

Duplicates and evaluation leakage

Duplication inflated reported performance in a dose-dependent way (Figure 6). With 40% injected near-duplicates, a random split reported R^2 = 0.947 on concrete and 0.846 on California housing. Group-aware splitting gave 0.910 and 0.828, values close to those for the original data (0.922 and 0.827). The inflation reached 0.036 and 0.018, larger than many improvements claimed for new models. On superconductivity, 35.6% of test records under a random split shared a feature vector with a training record. Group-aware splitting lowered R^2 from 0.908 to 0.902 and raised the mean absolute error from 6.56 K to 6.74 K (2.7%). The effect is small here because conflicting targets within groups limit what can be memorised, but it is consistent in the mean over five seeds.

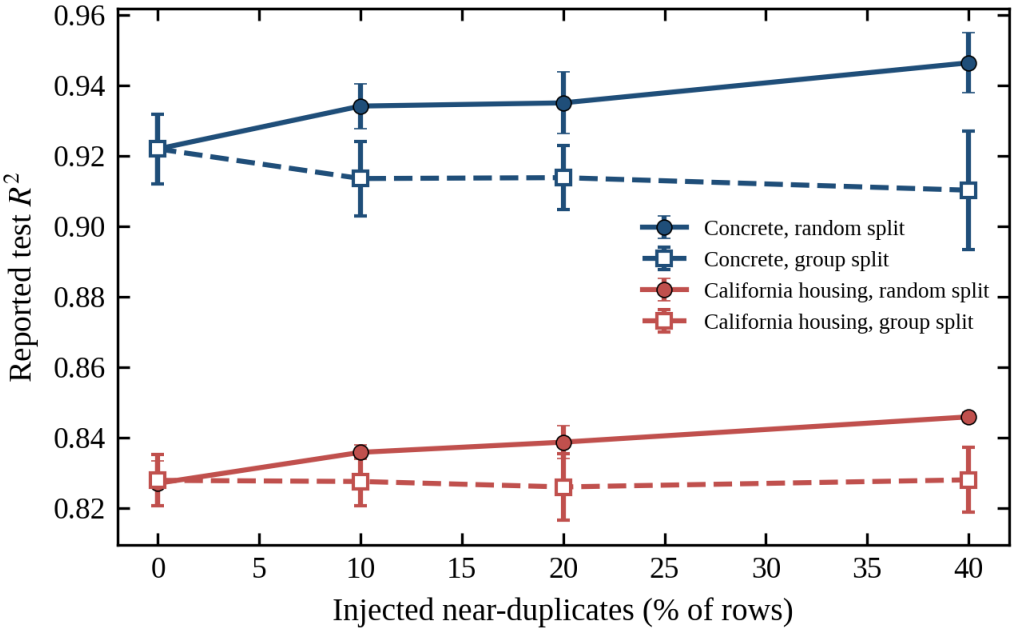


Figure 6. Reported test R^2 under random and group-aware splitting as a function of the fraction of injected near-duplicates. Error bars show standard deviations over three seeds.

The rounding-based hash detector found near-duplicates with perfect precision on California housing but a recall of only 0.28–0.34, and with precision of 0.59–0.86 and recall of 0.54–0.61 on concrete. Small perturbations push values across rounding boundaries, so exact hashing of rounded values is a fast first pass that must be followed by nearest-neighbour or locality-sensitive methods when near-duplicates are

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

expected. The precision below 1 on concrete reflects its natural duplicates, which the detector correctly found but which were not part of the injected ground truth.

Feature discovery

Automated feature discovery raised the cross-validated R^2 of ridge regression on concrete from 0.605 ± 0.024 to 0.873 ± 0.015 and on California housing from 0.645 to 0.657 (Figure 7a, Table 7). Products and ratios with age and the binders capture the nonlinear hardening of concrete, which a linear model on raw proportions cannot represent. For gradient boosting the discovered features made almost no difference on concrete (0.927 to 0.927) and California housing (0.829 to 0.827), because trees already model interactions. On superconductivity, discovery slightly reduced R^2 for both models (ridge 0.736 to 0.722; boosting 0.899 to 0.887). Selecting the top 60 candidates by mutual information discarded some of the 81 informative raw descriptors, a reminder that filter-based selection should be validated against the raw feature set.

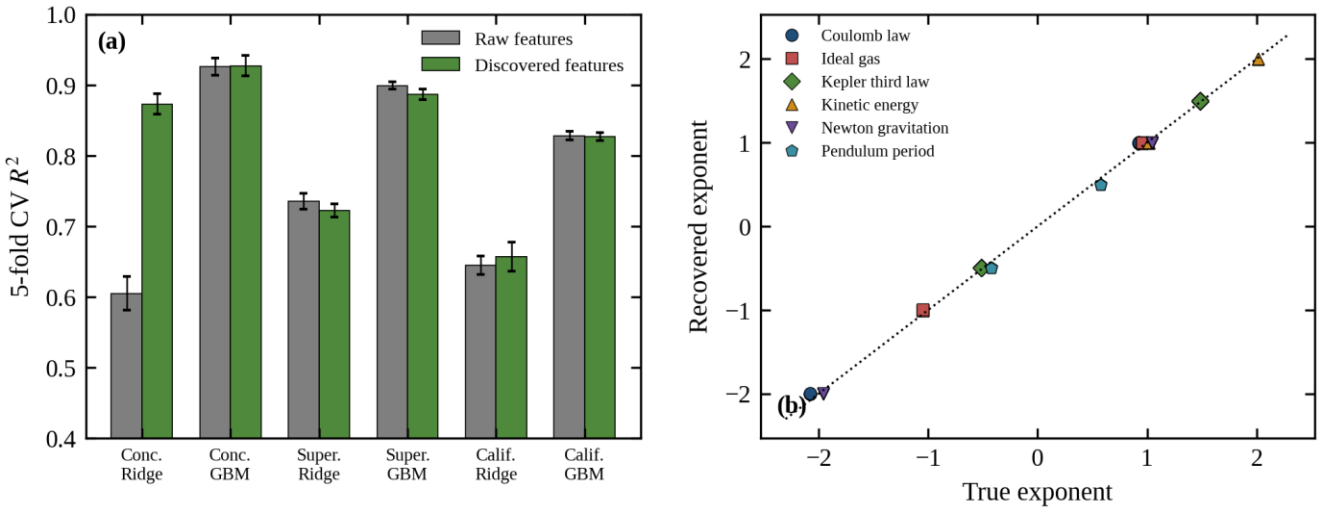


Figure 7. (a) Five-fold cross-validated R^2 with raw and automatically discovered features for ridge regression and gradient boosting (GBM). (b) Exponents recovered by log-linear lasso regression for six physical laws at 10% multiplicative noise compared with their true values (five seeds per law).

Table 7. Cross-validated R^2 (mean ± standard deviation over five folds) with raw and discovered features.

Dataset	Model	Raw features	Discovered features	Number of features (raw → discovered)
Concrete	Ridge	0.605 ± 0.024	0.873 ± 0.015	8 → 60
Concrete	GBM	0.927 ± 0.012	0.927 ± 0.014	8 → 60
Superconductivity	Ridge	0.736 ± 0.011	0.722 ± 0.009	81 → 60
Superconductivity	GBM	0.899 ± 0.005	0.887 ± 0.007	81 → 60
California housing	Ridge	0.645 ± 0.013	0.657 ± 0.021	13 → 60
California housing	GBM	0.829 ± 0.006	0.827 ± 0.006	13 → 60

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

The equation-discovery test was successful for all six laws (Figure 7b, Table 8). At 10% noise the mean maximum exponent error ranged from 0.003 (kinetic energy) to 0.006 (Newton and Coulomb), the largest coefficient assigned to an irrelevant variable was 0.007, and the fit explained at least 98.9% of the variance of the logged output. Fractional exponents such as the 3/2 of Kepler's law and the ±1/2 of the pendulum period were recovered as accurately as integer exponents. Newton's and Coulomb's laws have identical mathematical form, so they gave identical estimates for the same seeds; this is a useful check that the procedure depends only on the structure of the data. Lasso shrinkage biases coefficients slightly towards zero, which explains why the error at 1% noise (up to 0.009) was not smaller than at 10%; refitting the selected terms by ordinary least squares would remove this bias.

The candidates ranked highest by mutual information were also informative in their own right. For concrete, the top candidate was the product of age and cement content, followed by ratios of age to water and to the aggregates, which together express the joint effect of binder quantity and curing time. For superconductivity, ratios of the entropy of atomic radius and the range of first ionisation energy to the weighted mean valence ranked above every raw descriptor. For California housing, the product of longitude and latitude, a simple location encoding, ranked second only to longitude itself. Figure 8 shows that the equation-discovery procedure degraded gracefully with noise: errors and spurious coefficients stayed below 0.02 for every law and noise level.

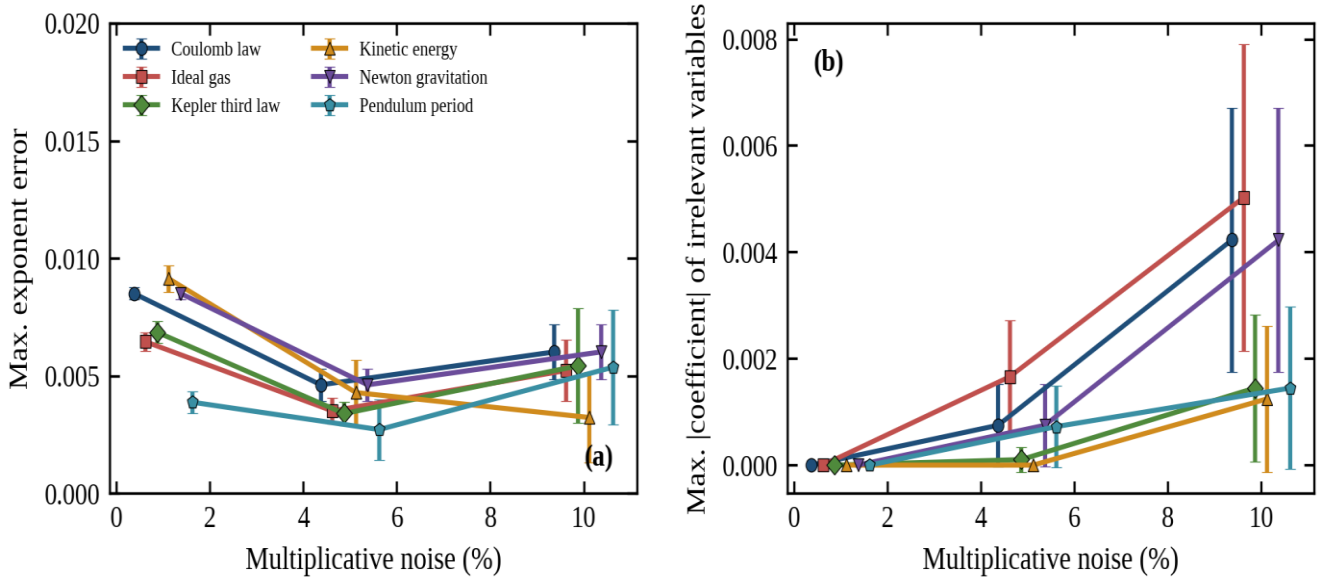


Figure 8. Robustness of log-linear sparse regression for equation discovery. (a) Maximum absolute exponent error and (b) largest coefficient assigned to an irrelevant variable as functions of multiplicative noise (means and standard deviations of five seeds).

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Table 8. Exponents recovered for six physical laws at 10% multiplicative noise (means of five seeds).

Law	True exponents	Recovered exponents	Max. error	Max. spurious	R^2 (log scale)
Newton gravitation	m_1: 1, m_2: 1, r: −2	0.996, 0.998, −1.999	0.006	0.004	0.999
Coulomb law	q_1: 1, q_2: 1, r: −2	0.996, 0.998, −1.999	0.006	0.004	0.999
Kepler third law	a: 1.5, M: −0.5	1.497, −0.495	0.005	0.001	0.998
Ideal gas	n: 1, T: 1, V: −1	0.997, 0.999, −1.000	0.005	0.005	0.998
Pendulum period	L: 0.5, g: −0.5	0.497, −0.495	0.005	0.001	0.989
Kinetic energy	m: 1, u: 2	0.997, 2.000	0.003	0.001	0.999

Max. spurious: largest absolute coefficient assigned to any of the three irrelevant variables.

Knowledge extraction

For concrete, the extracted knowledge agreed with established materials science. Age and the water-to-cement ratio were the two most important features, with permutation importances of 0.74 ± 0.03 and 0.59 ± 0.06 (as R^2 lost when shuffled). The depth-3 surrogate tree split first on water/cement ≤ 0.467 and then on age (Figure 9, Table 9). Its rules state that strength increases with curing age and decreases with the water-to-cement ratio, which is the content of Abrams' law and the basis of mix design [60]. Rankings were highly stable: the mean Spearman correlation across eight bootstrap models was 0.95 (minimum 0.88), and the top-10 sets were identical in every pair (Jaccard index 1.0).

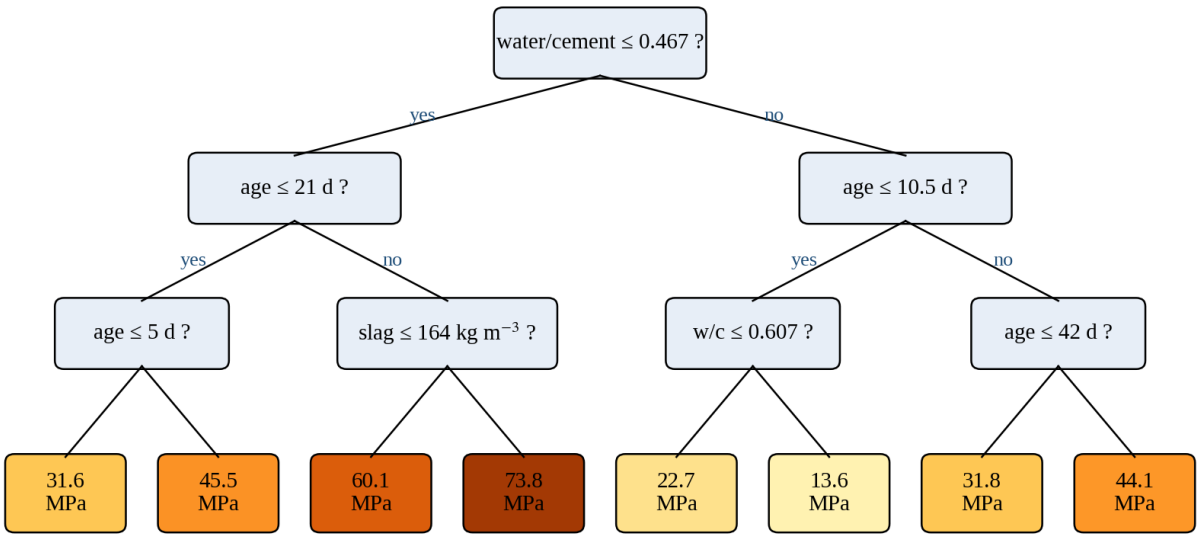


Figure 9. Depth-3 surrogate decision tree extracted from the gradient-boosting model

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

for concrete compressive strength. Leaves give the mean predicted strength; colour intensity increases with strength. w/c, water-to-cement ratio.

Read from left to right, the leaves of Figure 9 form a compact design rule. Mixes with a water-to-cement ratio of 0.467 or less reach a predicted 60–74 MPa after more than 21 days, the higher value when more than 164 kg m^{-3} of slag is present. Mixes above this ratio remain at 14–23 MPa in the first ten days and at 32–44 MPa thereafter. The second split on water-to-cement ratio within the early-age branch (0.607) reproduces the monotonic decrease of strength with water content. Each of these statements can be checked against experimental knowledge of concrete. The rule set is a scientific summary of what the model has learned, not merely a visualisation of it [49, 53].

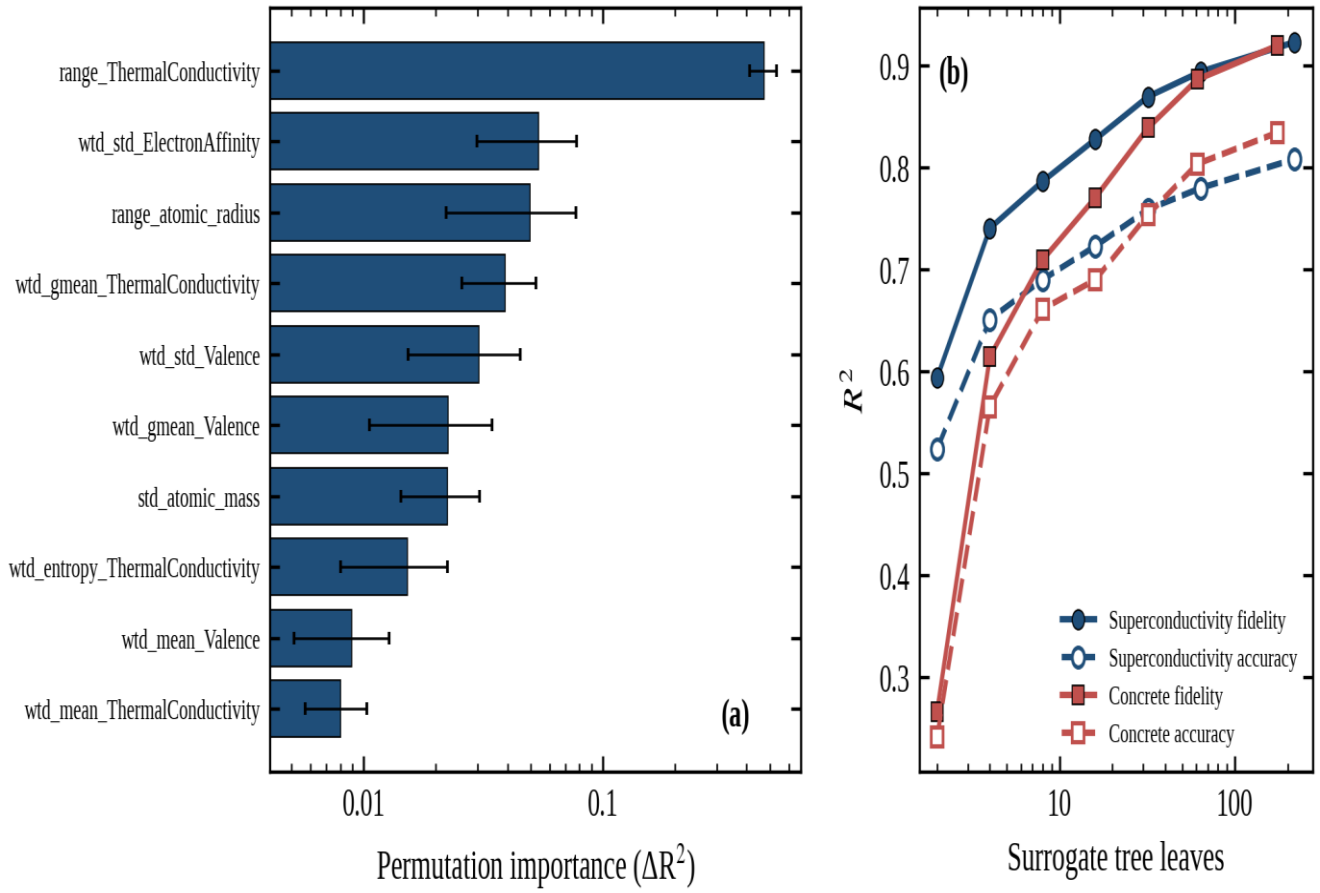


Figure 10. Knowledge extraction. (a) Permutation importance of the ten leading descriptors for superconducting critical temperature (mean and standard deviation over eight bootstrap models; logarithmic scale). (b) Fidelity to the gradient-boosting model and accuracy against the true target of decision-tree surrogates of increasing size.

For superconductivity, a single descriptor dominated: the range of elemental thermal conductivity had an importance of 0.47 ± 0.06, nine times that of the next feature (Figure 10a). Thermal-conductivity descriptors made up three of the top eight and supplied the first split and a second-level split of the surrogate tree, in agreement with the original analysis of this dataset, which also found thermal-conductivity features most important [55]. Stability was lower than for concrete. The mean Spearman

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

correlation was 0.69 and the top-10 Jaccard index 0.56, and only the first four descriptors appeared in the top 10 of every bootstrap model. Many of the 81 descriptors are strongly correlated, so models can exchange one for another without loss of accuracy, which is the multiplicity described by Fisher et al. [22]. Only the stable core should be interpreted scientifically.

Table 9. Knowledge-extraction results: model accuracy, stability of importance rankings and the depth-3 surrogate rule set.

Quantity	Concrete	Superconductivity
Gradient-boosting test R^2	0.924	0.895
Mean / minimum Spearman ρ of importances	0.953 / 0.883	0.686 / 0.583
Mean Jaccard index of top-10 sets	1.000	0.556
Most important feature	Age (0.74)	Range of thermal conductivity (0.47)
First surrogate split	water/cement ≤ 0.467	range_ThermalConductivity ≤ 399.9
Surrogate fidelity, depth 3 (8 leaves)	0.710	0.787
Surrogate fidelity, depth 6	0.887	0.894
Depth-3 rule with highest strength / T_C	w/c ≤ 0.467, age > 21 d, slag > 164 → 73.8 MPa	range_TC > 399.9, wtd_gmean_TC > 0.81, std_atomic_mass > 66.8 → 85.7 K

TC, thermal conductivity; w/c, water-to-cement ratio.

The surrogate analysis quantifies the trade-off between interpretability and fidelity (Figure 10b). Eight-leaf trees reproduced 71% (concrete) and 79% (superconductivity) of the variance of the full model's predictions, and 61–64 leaves raised fidelity to 0.89. Surrogate accuracy against the true target was 0.02–0.11 lower than fidelity at every size, so a compact rule set is a faithful summary of the model rather than a replacement for it.

Reproducibility and provenance

With fixed seeds, ten repeated runs on each dataset produced identical predictions, confirmed by a single SHA-256 hash per dataset, and identical scores (Figure 11a, Table 10). With uncontrolled seeds, all ten runs produced different predictions, and test scores spread by 0.053 on concrete, 0.050 on QSAR-biodeg, 0.024 on California housing and 0.013 on MAGIC. These ranges are of the same order as the effects of the data repairs in Sections 4.2–4.5, so a single unseeded run could make a repair look beneficial or harmful by chance alone.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

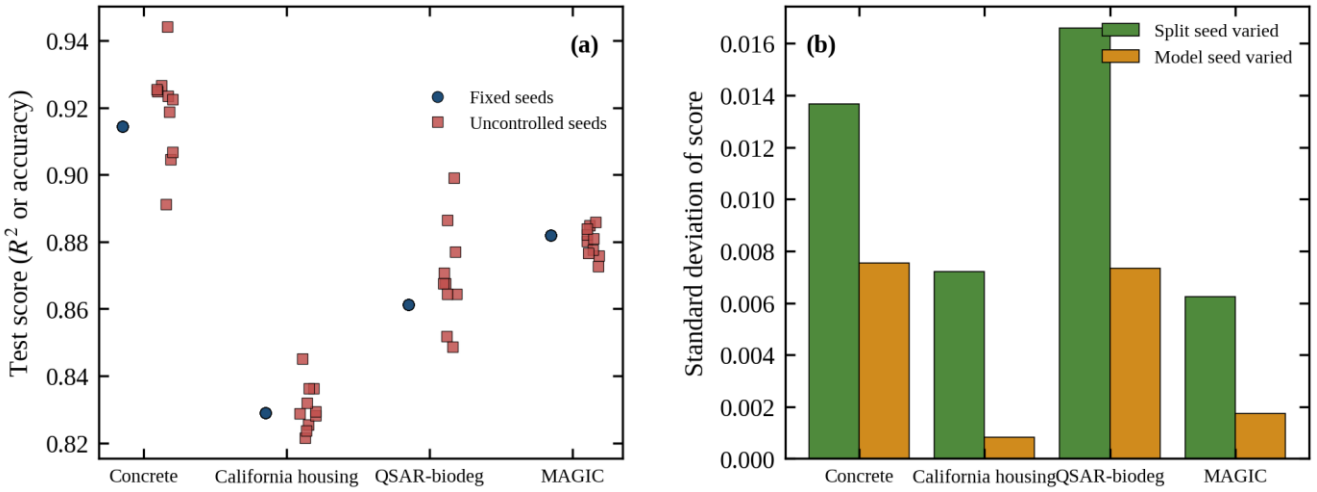


Figure 11. Reproducibility analysis. (a) Test scores of ten runs with fixed seeds (circles, coincident) and with uncontrolled seeds (squares). (b) Standard deviation of the score when only the split seed or only the model seed is varied.

Varying the seeds separately showed that the train–test split caused most of the variation (Figure 11b). The standard deviation due to the split was 0.006–0.017, and that due to the model seed 0.001–0.008, two to nine times smaller. Reporting the mean and spread over several splits, or using repeated cross-validation, therefore matters more than fixing the model seed alone. Provenance was cheap: hashing a full dataset took 0.07–1.8 ms, negligible beside training.

Table 10. Run-to-run variability of test scores (R^2 for regression, accuracy for classification) over ten runs.

Dataset	Fixed seeds: score (unique prediction hashes)	Uncontrolled: mean ± SD	Uncontrolled: range	SD, split only	SD, model only	Dataset hash time (ms)
Concrete	0.9143 (1)	0.919 ± 0.014	0.053	0.014	0.008	0.07
California housing	0.8290 (1)	0.831 ± 0.007	0.024	0.007	0.001	1.83
QSAR-biodeg	0.8612 (1)	0.870 ± 0.014	0.050	0.017	0.007	0.35
MAGIC	0.8819 (1)	0.880 ± 0.004	0.013	0.006	0.002	1.61

Environment: Python 3.12.3, scikit-learn 1.8.0, NumPy 2.4.4, x86-64.

Scalability

All quality checks scaled close to linearly with the number of rows (Figure 12). Fitted log–log slopes between 10^4 and 10^6 rows were 1.05 for the missingness profile, 1.07 for hash-based duplicate detection and 1.01 for the robust z-score. Isolation forest (slope 0.70) and the cross-validated label audit (0.69) were sublinear in this range because fixed costs still dominated. On one processor core, the first three checks took 0.06–1.4 s and isolation forest 5.5 s for one million rows; the five-fold label audit took 15 s for 300,000 rows. These costs are small compared with training and imply that a

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

full audit of datasets with tens of millions of rows is feasible on a single node, while distributed engines such as those used by Deequ [31] can extend the same checks further.

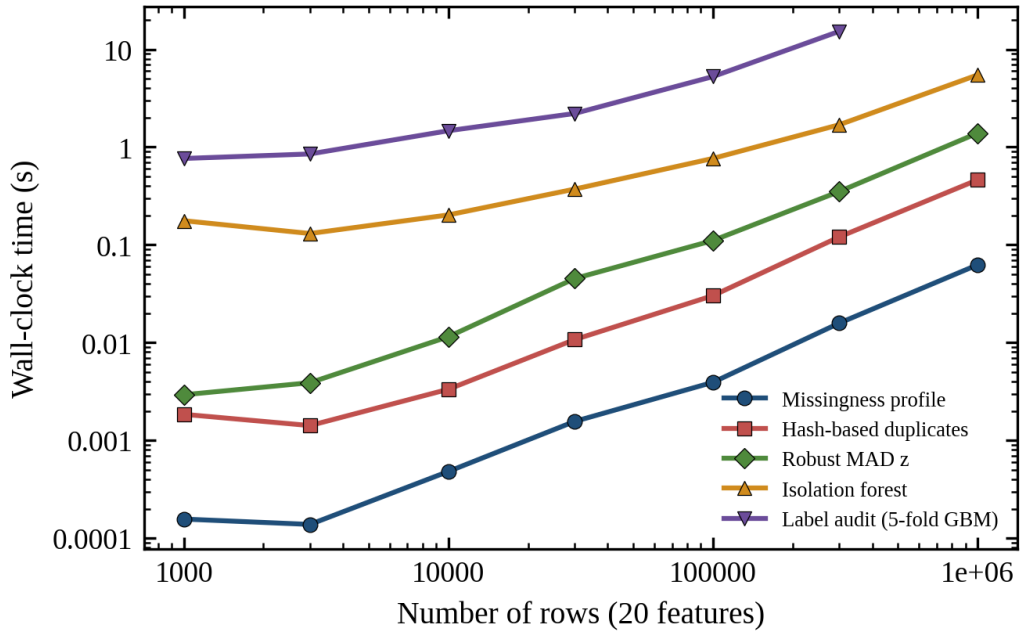


Figure 12. Wall-clock time of the automated quality checks as a function of dataset size (synthetic data, 20 features, single core). The label audit was not run at 10^6 rows.

Discussion

Prevalence and detectability of data problems (RQ1)

The audit shows that data problems are present in widely used scientific datasets and that some are structural rather than accidental. The duplication in superconductivity follows from representing materials by statistics of their elemental composition [55]. Similar many-to-one mappings occur wherever inputs are computed descriptors, as in cheminformatics and materials informatics. Detectability varied with the error type. Label errors and gross errors were ranked well by automated detectors (AUROC up to 0.99 and 0.91), but at low noise rates confident learning flagged mostly difficult examples, and the robust z-score labelled large fractions of skewed data as outliers. Automated detectors are therefore best used to rank records for review, or with explicit budgets, not as unsupervised filters [12, 33].

5.2 Consequences of data problems and the value of repair (RQ2)

Data problems affected results in two ways. They degraded the models: 30% label noise cost QSAR-biodeg 0.16 in accuracy, and 5% gross errors cost concrete 0.73 in R^2. They also inflated the reported performance: near-duplicates added up to 0.036 to R^2 under random splitting. The second effect is more dangerous for science because it produces optimistic conclusions that look valid [3, 16]. Repair recovered much of the degradation when the problem was large (52–62% of the label-noise loss at 30%, and nearly all of the gross-error loss with the MAD-based filter), but it brought little gain or a small loss when the problem was mild. The results support a principle of proportionate repair: measure first, then repair only where the measured problem is

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

large enough to justify discarding or altering data.

Feature discovery and scientific knowledge (RQ3)

Feature discovery helped most where the model class could not represent the relevant interactions, as in the linear model for concrete, and gave no benefit for flexible tree ensembles. Its main value for science is therefore interpretability, not accuracy: a linear model on discovered features is easier to examine than a boosted ensemble with the same R^2. Equation discovery recovered exact exponents for all six physical laws when the correct variables and a log-linear form were available, consistent with the success of sparse-regression approaches such as SINDy [18]. In real data the candidate library and the measurement noise are rarely so favourable, and laws involving sums of terms need richer libraries or symbolic regression [21, 48]. Knowledge extraction rediscovered the water-to-cement rule for concrete and the importance of thermal conductivity for superconductors. The stability analysis showed that only part of an importance ranking may deserve trust when descriptors are correlated. We recommend reporting rank stability together with any importance-based claim [22, 53].

Reproducibility and scale (RQ4)

Bit-for-bit reproducibility was achieved simply by fixing seeds and recording the environment, and it can be verified cheaply by hashing predictions. The variance analysis supports earlier warnings that single runs are unreliable [24, 25], and adds that, for tabular scientific tasks, the data split contributes more variance than the model seed. A reproducible result should therefore be distinguished from a robust one: fixing the seed makes a single number repeatable but does not show that it is representative. Reporting the distribution over several splits addresses both. The near-linear scaling of the checks removes computational cost as an argument against routine data auditing.

Implications for scientific practice and publishing

The results have implications beyond individual projects. Benchmark datasets are reused by many groups, so a structural problem such as the duplicate groups in superconductivity affects every study that uses a random split, and published comparisons of models on such data may partly reflect leakage rather than skill [3, 4]. Dataset providers could prevent this by publishing recommended group-aware splits and quality reports alongside the data, in the spirit of datasheets for datasets [27]. Journals and reviewers could ask for three inexpensive items that this study shows to matter: a data-quality audit, the splitting strategy and its justification, and the spread of results over several splits or seeds [2, 23]. For authors, the cost of these items is small: in our experiments the audit and provenance steps took seconds, while the effects they revealed were comparable to, or larger than, the differences that typically separate competing models [24]. Data-centric checks are therefore part of the scientific method when machine learning is used, not an optional engineering refinement.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Practical recommendations

Table 11 condenses the results into recommendations for researchers who apply machine learning to scientific data. Each is supported by a measured effect in Section 4 and can be implemented with standard open-source tools [19, 62].

Table 11. Recommendations for data-centric scientific machine learning derived from the results.

Stage	Recommendation	Evidence in this study
Audit	Hash feature vectors and check for duplicate groups with conflicting targets before modelling	40.1% duplicated vectors and 2,375 conflicting groups in superconductivity
Outliers	Do not use a fixed robust z cut-off as an unsupervised filter; combine coordinate-wise and density scores with a removal budget	9.9–57.5% of rows flagged; MAD filtering restored R^2 0.19 → 0.90
Labels	Use confident learning to rank samples for review; filter automatically only when estimated noise is high	Accuracy +0.08 to +0.09 at 30% noise; no gain at 5–10%
Missing data	Prefer model-based imputation for correlated descriptors; compare with native missing-value handling	nRMSE 0.23 vs 1.00; native handling equal or better on weakly correlated data
Evaluation	Use group-aware splits whenever records can share an identity	Up to +0.036 R^2 inflation with random splits
Features	Use automated feature discovery for interpretable models; validate selection against raw features	Ridge R^2 0.605 → 0.873 on concrete; small losses on superconductivity
Knowledge	Report the stability of importance rankings and the fidelity of surrogate rules	Spearman ρ 0.95 vs 0.69; fidelity 0.71–0.79 at 8 leaves
Reproducibility	Fix and record seeds, hash data and predictions, and report the spread over several splits	Identical hashes with fixed seeds; score range up to 0.053 otherwise

Limitations

The study has limitations. The controlled experiments used synthetic corruptions (symmetric label flips, MCAR missingness and three types of gross error) that are cleaner than real errors, which are often class-dependent, missing not at random or correlated across features [34, 37]. The datasets are tabular and moderate in size; image, spectral, sequence and graph data raise further problems. Two datasets were subsampled for the imputation, outlier, feature and knowledge experiments to keep computation on a single core tractable. Equation discovery was tested on power laws with the correct variables present, which is a favourable case. Finally, suspected label errors in the unmodified data were not verified by domain experts, so the audit rates are upper bounds on ambiguity, not counts of errors.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Future work

Four directions follow. First, extending the corruption models to class-conditional label noise, MNAR missingness and correlated sensor faults would bring the experiments closer to real scientific data. Second, near-duplicate detection based on nearest neighbours or locality-sensitive hashing should replace rounding-based hashing for large datasets. Third, equation discovery could be combined with the feature-discovery stage, so that candidate libraries are generated automatically and checked for physical units [21]. Fourth, the provenance layer could be connected to FAIR data repositories and experiment trackers, so that every published result carries its data hash, seeds and quality report [19, 29].

Conclusion

This study designed and evaluated an AI-driven, data-centric pipeline for scientific machine learning that integrates automated data-quality assessment, repair, feature discovery, knowledge extraction and reproducibility controls. Applied to six public scientific datasets, the pipeline found substantial structural duplication with conflicting targets in a widely used materials dataset and showed that a common outlier rule is unsuitable as an automatic filter for skewed scientific data. Controlled experiments showed that label noise, missing values, gross errors and duplicates both degrade models and inflate reported performance, and that automated repair recovers much of the loss when problems are large but should be applied in proportion to the measured problem. Feature discovery transformed the accuracy of an interpretable linear model for concrete. Sparse log-linear regression recovered the exponents of six physical laws to within 0.01, and knowledge extraction rediscovered the water-to-cement rule for concrete and the dominance of thermal-conductivity descriptors for superconductors, with stability measures that show which conclusions can be trusted. Fixed seeds and hashing made results bit-for-bit reproducible at negligible cost, while uncontrolled runs varied by up to 0.053 in test score, mainly because of the data split. All checks scaled close to linearly with data size. Data-quality assessment, group-aware evaluation, stability reporting and provenance tracking are inexpensive and change scientific conclusions measurably, and they should become standard practice in machine-learning-based research.

Abbreviations

Table 12. Abbreviations used in this paper.

Abbreviation	Meaning	Abbreviation	Meaning
AI	Artificial intelligence	MCAR / MAR / MNAR	Missing completely at random / at random / not at random
AUROC	Area under the receiver operating characteristic curve	MICE	Multiple imputation by chained equations
CL	Confident learning	nRMSE	Normalised root-mean-square error
FAIR	Findable, accessible, interoperable, reusable	QSAR	Quantitative structure–activity relationship

Online ISSN 3007-3197
Print ISSN 3007-3189

Abbreviation	Meaning	Abbreviation	Meaning
GBM	Gradient-boosting machine	R^2	Coefficient of determination
KNN	k-nearest neighbours	SHA-256	Secure hash algorithm, 256-bit
LOF	Local outlier factor	SINDy	Sparse identification of nonlinear dynamics
MAD	Median absolute deviation	WDBC	Wisconsin diagnostic breast cancer

Declarations

Funding. This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Conflict of interest. The authors declare no conflict of interest.

Data and code availability. All datasets are publicly available from OpenML. The code used for the experiments, the per-seed results and the figure scripts are available from the corresponding author on reasonable request.

References

[1] Hey T, Tansley S, Tolle K, editors. The Fourth Paradigm: Data-Intensive Scientific Discovery. Redmond (WA): Microsoft Research; 2009.

[2] Pineau J, Vincent-Lamarre P, Sinha K, Larivière V, Beygelzimer A, d'Alché-Buc F, et al. Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program). J Mach Learn Res. 2021;22(164):1–20.

[3] Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns. 2023;4(9):100804.

[4] Northcutt CG, Athalye A, Mueller J. Pervasive label errors in test sets destabilize machine learning benchmarks. In: Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks; 2021.

[5] Peng RD. Reproducible research in computational science. Science. 2011;334(6060):1226–1227.

[6] Zha D, Bhat ZP, Lai KH, Yang F, Jiang Z, Zhong S, et al. Data-centric artificial intelligence: a survey. arXiv:2303.10158 [Preprint]. 2023.

[7] Whang SE, Roh Y, Song H, Lee JG. Data collection and quality challenges in deep learning: a data-centric AI perspective. VLDB J. 2023;32:791–813.

[8] Jarrahi MH, Memariani A, Guha S. The principles of data-centric AI. Commun ACM. 2023;66(8):84–92.

[9] Sambasivan N, Kapania S, Highfill H, Akrong D, Paritosh P, Aroyo LM. "Everyone wants to do the model work, not the data work": data cascades in high-stakes AI. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems; 2021. p. 1–15.

[10] Sculley D, Holt G, Golovin D, Davydov E, Phillips T, Ebner D, et al. Hidden technical debt in machine learning systems. Adv Neural Inf Process Syst. 2015;28:2503–2511.

[11] Baker M. 1,500 scientists lift the lid on reproducibility. Nature. 2016;533(7604):452–454.

Online ISSN 3007-3197
Print ISSN 3007-3189

[12]Northcutt CG, Jiang L, Chuang IL. Confident learning: estimating uncertainty in dataset labels. J Artif Intell Res. 2021;70:1373–1411.
[13]van Buuren S, Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations in R. J Stat Softw. 2011;45(3):1–67.
[14]Liu FT, Ting KM, Zhou ZH. Isolation forest. In: Proceedings of the 8th IEEE International Conference on Data Mining; 2008. p. 413–422.
[15]Breunig MM, Kriegel HP, Ng RT, Sander J. LOF: identifying density-based local outliers. In: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data; 2000. p. 93–104.
[16]Kaufman S, Rosset S, Perlich C, Stitelman O. Leakage in data mining: formulation, detection, and avoidance. ACM Trans Knowl Discov Data. 2012;6(4):15.
[17]Kanter JM, Veeramachaneni K. Deep feature synthesis: towards automating data science endeavors. In: Proceedings of the IEEE International Conference on Data Science and Advanced Analytics (DSAA); 2015. p. 1–10.
[18]Brunton SL, Proctor JL, Kutz JN. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. Proc Natl Acad Sci USA. 2016;113(15):3932–3937.
[19]Zaharia M, Chen A, Davidson A, Ghodsi A, Hong SA, Konwinski A, et al. Accelerating the machine learning lifecycle with MLflow. IEEE Data Eng Bull. 2018;41(4):39–45.
[20]Batini C, Cappiello C, Francalanci C, Maurino A. Methodologies for data quality assessment and improvement. ACM Comput Surv. 2009;41(3):16.
[21]Udrescu SM, Tegmark M. AI Feynman: a physics-inspired method for symbolic regression. Sci Adv. 2020;6(16):eaay2631.
[22]Fisher A, Rudin C, Dominici F. All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously. J Mach Learn Res. 2019;20(177):1–81.
[23]Gundersen OE, Kjensmo S. State of the art: reproducibility in artificial intelligence. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2018;32(1):1644–1651.
[24]Bouthillier X, Laurent C, Vincent P. Unreproducible research is reproducible. In: Proceedings of the 36th International Conference on Machine Learning. PMLR. 2019;97:725–734.
[25]Picard D. Torch.manual_seed(3407) is all you need: on the influence of random seeds in deep learning architectures for computer vision. arXiv:2109.08203 [Preprint]. 2021.
[26]Wang RY, Strong DM. Beyond accuracy: what data quality means to data consumers. J Manag Inf Syst. 1996;12(4):5–33.
[27]Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Daumé III H, et al. Datasheets for datasets. Commun ACM. 2021;64(12):86–92.
[28]Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, et al. Model cards for model reporting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency; 2019. p. 220–229.
[29]Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and

Online ISSN 3007-3197
Print ISSN 3007-3189

stewardship. Sci Data. 2016;3:160018.
[30]Breck E, Polyzotis N, Roy S, Whang SE, Zinkevich M. Data validation for machine learning. In: Proceedings of Machine Learning and Systems (MLSys); 2019.
[31]Schelter S, Lange D, Schmidt P, Celikel M, Biessmann F, Grafberger A. Automating large-scale data quality verification. Proc VLDB Endow. 2018;11(12):1781–1794.
[32]Iglewicz B, Hoaglin DC. How to Detect and Handle Outliers. Milwaukee (WI): ASQC Quality Press; 1993.
[33]Rousseeuw PJ, Hubert M. Robust statistics for outlier detection. WIREs Data Min Knowl Discov. 2011;1(1):73–79.
[34]Frénay B, Verleysen M. Classification in the presence of label noise: a survey. IEEE Trans Neural Netw Learn Syst. 2014;25(5):845–869.
[35]Brodley CE, Friedl MA. Identifying mislabeled training data. J Artif Intell Res. 1999;11:131–167.
[36]Song H, Kim M, Park D, Shin Y, Lee JG. Learning from noisy labels with deep neural networks: a survey. IEEE Trans Neural Netw Learn Syst. 2023;34(11):8135–8153.
[37]Rubin DB. Inference and missing data. Biometrika. 1976;63(3):581–592.
[38]Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001;17(6):520–525.
[39]Stekhoven DJ, Bühlmann P. MissForest: non-parametric missing value imputation for mixed-type data. Bioinformatics. 2012;28(1):112–118.
[40]Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: a highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst. 2017;30:3146–3154.
[41]Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016. p. 785–794.
[42]Lee K, Ippolito D, Nystrom A, Zhang C, Eck D, Callison-Burch C, et al. Deduplicating training data makes language models better. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics; 2022. p. 8424–8445.
[43]Allamanis M. The adverse effects of code duplication in machine learning models of code. In: Proceedings of the ACM SIGPLAN International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software (Onward!); 2019. p. 143–153.
[44]Feurer M, Klein A, Eggensperger K, Springenberg J, Blum M, Hutter F. Efficient and robust automated machine learning. Adv Neural Inf Process Syst. 2015;28:2962–2970.
[45]Hutter F, Kotthoff L, Vanschoren J, editors. Automated Machine Learning: Methods, Systems, Challenges. Cham: Springer; 2019.
[46]Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res. 2003;3:1157–1182.
[47]Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Series

Online ISSN 3007-3197
Print ISSN 3007-3189

B Stat Methodol. 1996;58(1):267–288.
[48]Schmidt M, Lipson H. Distilling free-form natural laws from experimental data. Science. 2009;324(5923):81–85.
[49]Craven MW, Shavlik JW. Extracting tree-structured representations of trained networks. Adv Neural Inf Process Syst. 1996;8:24–30.
[50]Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
[51]Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30:4765–4774.
[52]Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?": explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016. p. 1135–1144.
[53]Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206–215.
[54]Kluyver T, Ragan-Kelley B, Pérez F, Granger B, Bussonnier M, Frederic J, et al. Jupyter Notebooks: a publishing format for reproducible computational workflows. In: Loizides F, Schmidt B, editors. Positioning and Power in Academic Publishing: Players, Agents and Agendas. Amsterdam: IOS Press; 2016. p. 87–90.
[55]Hamidieh K. A data-driven statistical model for predicting the critical temperature of a superconductor. Comput Mater Sci. 2018;154:346–354.
[56]Mansouri K, Ringsted T, Ballabio D, Todeschini R, Consonni V. Quantitative structure–activity relationship models for ready biodegradability of chemicals. J Chem Inf Model. 2013;53(4):867–878.
[57]Bock RK, Chilingarian A, Gaug M, Hakl F, Hengstebeck T, Jiřina M, et al. Methods for multidimensional event classification: a case study using images from a Cherenkov gamma-ray telescope. Nucl Instrum Methods Phys Res A. 2004;516(2–3):511–528.
[58]Vanschoren J, van Rijn JN, Bischl B, Torgo L. OpenML: networked science in machine learning. SIGKDD Explor. 2014;15(2):49–60.
[59]Street WN, Wolberg WH, Mangasarian OL. Nuclear feature extraction for breast tumor diagnosis. In: Proceedings of SPIE 1905, Biomedical Image Processing and Biomedical Visualization; 1993. p. 861–870.
[60]Yeh IC. Modeling of strength of high-performance concrete using artificial neural networks. Cem Concr Res. 1998;28(12):1797–1808.
[61]Pace RK, Barry R. Sparse spatial autoregressions. Stat Probab Lett. 1997;33(3):291–297.
[62] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–2830.