WHO REALLY WINS? STRESS-TESTING SYMBOLIC REGRESSION BENCHMARKS ACROSS METRICS, AGGREGATIONS, AND DATASETS

Anonymous authors
Paper under double-blind review

ABSTRACT

Symbolic regression (SR) benchmarks such as SRBENCH rank methods by averaging a single accuracy metric over repeated runs across evaluated datasets. Yet, the conclusions drawn from such leaderboards depend on implicit design choices, i.e., which error metric is reported, how repeated runs are aggregated, and which datasets are included. These design choices are rarely stress-tested. In this paper, we present a meta-study about the robustness of the SR ranking built on the recently proposed next-generation SRBENCH. Using each performance metric (MSE, MAE, R^2, and training time) and each run-aggregation rule (mean, median, max, min, standard deviation, and four random single-run picks), we generate model rankings with a portfolio of 15 multi-criteria decision-making (MCDM) schemes and introduce two robustness indicators: *ranking-scheme robustness* (sensitivity of rankings to both aggregation methodology and chosen metric) and *dataset-composition robustness* (sensitivity to correlation-aware dataset subsampling). Then, we perform *metric-agnostic* and *aggregation-agnostic* cross-consistency analysis. In addition, we complement our robustness analysis by isolating real-world (black-box) from physical-law-based (first-principles) datasets. Across 23 SR methods and 21 datasets, we find that no single method dominates every regime, but a small core—most consistently EQL and PS-TREE—remains robust across metrics, aggregations, and dataset compositions on real-world data, while QLATTICE and OPERON rise to the top on first-principles problems. We validate five among the overall-top-ranked methods on five held-out external, well-established datasets. Our findings support the *call for action* of the SR community for a living benchmark and show that robustness indicators, rather than single averaged scores, should accompany future SR leaderboards. We publicly release all model scores and derived rankings as reusable benchmark artifacts, enabling new benchmarking studies without rerunning the original experiments.

1 INTRODUCTION

Symbolic regression (SR) is the task of discovering a mathematical formula (or equation) that fits the given data (Koza, 1994; Schmidt & Lipson, 2009; Langdon et al., 2008; Kronberger et al., 2024). More specifically, it searches for a closed-form mathematical expression $\hat{f}(\mathbf{x}, \hat{\theta})$ that fits a set of observations $\{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathcal{X}, y_i \in \mathcal{Y}\}$ such that $y \approx \hat{f}(\mathbf{x}, \hat{\theta})$. Unlike parametric models that start from a fixed functional form, SR searches over both the structure of the expression and its parameters. Furthermore, since these models represent mathematical formulae, in case the discovered solutions are contained in size, they may additionally exhibit interesting interpretability properties (Aldeia & de França, 2022; Rovito et al., 2022; Makke & Chawla, 2024; Nadizar et al., 2024b), which, consequently, cover potential interest in high-stakes applications (Jobin et al., 2019; Molnar, 2020; Rudin, 2019; Nadizar et al., 2024c; Dong & Zhong, 2025; Nadizar et al., 2024a; Crespi et al., 2023), where a responsible and transparent use of machine learning (ML) models is required, as stated in the artificial intelligence (AI) acts by the European Union and the United States (Commission, 2021; 117th US Congress, 2022; Jobin et al., 2019). SR methods have found application in several scenarios (Huynh et al., 2022), such as biology (Schmidt et al., 2011), survival analysis (Rovito & Virgolin, 2025), physics (Schmidt & Lipson, 2009; Llorella et al., 2024; Udrescu &

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

Tegmark, 2020), robotics (Bongard & Lipson, 2005), wind turbines identification (La Cava et al., 2016b), finance (Chen, 2012), fluid dynamics inference (La Cava et al., 2016a), air quality monitoring (Rovito et al., 2026), water alkalinity reconstruction (Tonelli et al., 2026), and climate modeling (Stanislawska et al., 2012). Both the popularity and complexity of SR (Virgolin & Pissis, 2022) have led to the development of several heuristics, most of them based on genetic programming (GP) (Cranmer, 2023; Kommenda et al., 2020; Cava et al., 2019; Bonin et al., 2024; Vanneschi et al., 2014). Each of them was validated under a different evaluation setting, making fair comparison among these methods challenging.

SRBENCH (La Cava et al., 2021) was introduced as a living benchmark to bring order to this landscape, and a recent *call for action* (Imai Aldeia et al., 2025) (hereafter the next-generation SRBENCH) nearly doubled the number of evaluated methods, refined the evaluation metrics, and highlighted trade-offs between accuracy, model complexity, and energy consumption. A central conclusion of that work is that *no single method dominates across all datasets*. Despite this progress, SR leaderboards (like most ML benchmarks) still distill performance into a small number of aggregated statistics. A ranking is typically produced by (i) choosing one error metric (e.g. R^2 or mean squared error), (ii) aggregating the metric over the repeated runs of each method (almost always by taking the mean or median), and (iii) averaging or otherwise pooling across a fixed collection of datasets. Each of these three choices is a *degree of freedom in the evaluation design*, and each can silently change which method looks best. The community has repeatedly flagged that aggregated views can obscure important nuances (De França, 2023; Dick, 2022; Imai Aldeia et al., 2025), but, to the best of our knowledge, the sensitivity of SR rankings to these choices has never been measured systematically. Benchmark rankings may depend not only on model quality but also on evaluation design. Inspired by robustness analysis in text embedding benchmarks (Gjorgjevikj et al., 2026), we examine this issue in the distinct context of SR, focusing on performance metrics, such as mean squared error (MSE), mean absolute error (MAE), R^2 and training time, along with run-aggregation rules (mean, median, max, min, standard deviation, and random individual runs). Our goal is to determine whether the reported winners remain stable across these SR-specific evaluation choices.

Contributions. This paper makes the following contributions. (i) We formulate SR benchmark robustness as a multi-criteria decision-making (MCDM) problem and instantiate two robustness indicators for SR: *ranking-scheme robustness* (sensitivity of rankings to both aggregation methodology and chosen metric) and *dataset-composition robustness* (sensitivity to correlation-aware dataset subsampling). (ii) We apply this framework to the next-generation SRBENCH (Imai Aldeia et al., 2025), analyzing 23 SR methods over 21 datasets, and separately over black-box (real-world) and first-principles (physical-law-based, abbreviated to "law-based" from now on for simplicity) subsets, for four metrics and nine run-aggregation rules. (iii) We introduce *metric-agnostic* and *aggregation-agnostic* cross-consistency analysis, which identifies the methods that remain top-ranked regardless of which metric is reported or how repeated runs are summarized. (iv) We show that a small core of methods (EQL and PS-TREE) is robust on real-world data whereas QLATTICE and OPERON rise to the top on first-principles data, and we confirm the robust core on five held-out external datasets that are well-established in the field. Our results support reporting robustness indicators alongside point estimates on SR leaderboards, helping practitioners distinguish consistently strong methods from those favored by a particular metric, aggregation rule, or dataset mixture. To support reproducibility and future studies without rerunning the original experiments, we release the underlying model scores, derived rankings, code, and analysis artifacts across all evaluated choices at https://anonymous.4open.science/r/RobustSRBench-B094/.

2 RELATED WORK

Benchmarking symbolic regression. Orzechowski et al. (2018) were among the first to point out that SR methods were being compared on simple problems with a lack of standardization, and conducted a large study across more than a hundred datasets. La Cava et al. (2021) formalized these efforts into SRBENCH, a joint effort comparing 14 modern SR methods over datasets retrieved from the Penn machine learning benchmarks (PMLB) (Olson et al., 2017; Romano et al., 2022) and law-based data based on Feynman equations (Feynman et al., 2015; Feynman, 2006). Subsequent works criticized the reliance on aggregated views (De França, 2023; Dick, 2022) and the way ground-truth data for SRBENCH's Feynman datasets was sampled (Matsubara et al., 2024). Later, de Franca et al. (2025) benchmarked a set of modern SR methods on artificially-generated datasets to measure

overall accuracy performance, robustness to noise, capabilities of selecting the relevant features, extrapolation performance, and interpretability. They showed that SR models have the potential to overcome the shortcomings with traditional ML models by using domain knowledge and customized user-experience (Ingelse & Fonseca, 2023). Other worth-mentioning SR-based benchmarking efforts have been applied in cosmology (Thing & Koksbang, 2025), civil engineering (Abdalla et al., 2025), and aerodynamics (Zegklitz & Posik, 2021). Most recently, the *call for action* (Imai Aldeia et al., 2025) for SRBENCH nearly doubled the number of methods to 25, increased the number of independent runs from 10 to 30, curated 24 datasets split into a black-box and a phenomenological & first-principles track, and analyzed energy–accuracy–complexity trade-offs.

Open questions motivating this work. The *call for action* frames benchmarking as an ongoing community responsibility and leaves several questions deliberately open. Its headline conclusion is that *no single method dominates across all datasets*, yet it also cautions that final results are reported with aggregated metrics that "can hide finer details in performance comparisons", and it argues for refining the reporting methodology to compare the nuances among top-performing methods rather than designating a single state-of-the-art (Imai Aldeia et al., 2025). Several concrete questions follow directly from this stance but are not answered there. *(i)* Is the identity of the top ranked methods stable when the reported error metric changes—R^2 versus MSE, MAE, or training time? *(ii)* How much does the choice of run-aggregation alter the ranking, given that the benchmark summarizes each method by the best of its 30 runs while practitioners typically report the mean or median? *(iii)* Are rankings stable when the dataset collection is resampled, i.e., is a method's lead driven by a handful of correlated "easy" datasets? *(iv)* Do the real-world and law-based tracks agree on their winners? Because the next-generation SRBENCH distills performance into aggregated performance-profile plots under a single aggregation rule, it cannot address these questions of *evaluation-design sensitivity*. We close this gap by casting SR ranking as a MCDM problem and quantifying how much the induced rankings vary across metrics, run-aggregations, and dataset compositions. Recent work has highlighted the value of robustness indicators for assessing the sensitivity of benchmark rankings to evaluation-design choices (Gjorgjevikj et al., 2026). Other recent benchmark ranking robustness-related studies employ rank aggregation approaches inspired by social choice theory, where datasets are "voters" and models are "candidates" (Rofin et al., 2023; Colombo et al., 2022; Enevoldsen et al., 2025). These strategies, however, assume equal relevance of datasets and do not account for correlations or application-driven preferences.

3 METHODOLOGY

Let $\mathcal{M} = \{1, \ldots, M\}$ index the set of evaluation *metrics*, and let $\mathcal{A} = \{1, \ldots, A\}$ index the set of run-*aggregation* rules. For a metric $m \in \mathcal{M}$ and aggregation rule $a \in \mathcal{A}$, we evaluate t SR methods on d datasets, producing a performance matrix $P^{(m,a)} \in \mathbb{R}^{t \times d}$, where $P^{(m,a)}_{i,j}$ is the performance of method i on dataset j under metric m, after the u repeated runs for that (method, dataset) pair have been collapsed by aggregation rule a. Rows are methods and columns are datasets. One of the four metrics (R^2) is a maximization criterion and the remaining ones (MSE, MAE, training time) are minimization criteria; MCDM directions are set accordingly. Prior to all downstream analysis, each matrix $P^{(m,a)}$ is normalized column-wise (i.e. per dataset) so that datasets on very different numeric scales contribute comparably. The nine aggregation rules are *mean*, *median*, *max*, *min*, *standard deviation*, and four *random single-run picks* ($\mathrm{random}_0, \ldots, \mathrm{random}_3$), each selecting one of the u runs at random with a fixed seed. The mean and median are the aggregations that practitioners actually report; max and min are optimistic and pessimistic stress tests; standard deviation ranks methods by run-to-run *stability* rather than central performance; and random picks probe how much a single lucky or unlucky run would perturb a ranking. To compare methods we use K ranking schemes, each defining a ranking operator $R^{(s)} : \mathbb{R}^{t \times d} \to \{1, \ldots, t\}^t$ mapping a performance matrix to a vector of ranks (smaller is better).

Dataset subsampling. To probe sensitivity to dataset composition, we build subsamples of datasets with low pairwise correlation across method-performance profiles. For dataset j, its profile is the column vector $p^{(m,a)}_j = P^{(m,a)}_{\cdot,j} \in \mathbb{R}^t$. Datasets whose profiles are strongly correlated (correlation above a threshold τ under a correlation-dissimilarity metric) carry redundant information and are clustered together; from each cluster we randomly draw one representative, generating uncorrelated

subsamples with fixed seeds. Together with the full composition (all datasets), this gives four dataset compositions per (metric, aggregation) pair.

Robustness indicators. For a ranking scheme s and dataset composition c, the ranks are $r^{(s,m,a,c)} = R^{(s)}(P^{(m,a,c)})$. We quantify two sources of variation. First, for a fixed dataset composition c, we measure how much method i's rank varies across the K ranking schemes with $\mathrm{RS}^{(m,a,c)}(i) = \mathrm{Var}_{s=1,\dots,K}\left(r_i^{(s,m,a,c)}\right)$ (*Ranking-scheme robustness (RS)*). Second, for a fixed ranking scheme s, we measure how much method i's rank varies across the dataset compositions with $\mathrm{DS}^{(m,a,s)}(i) = \mathrm{Var}_{c}\left(r_i^{(s,m,a,c)}\right)$ (*Dataset-composition robustness (DS)*). Low variance indicates a stable ranking. When only a single dataset composition is meaningful (e.g. when all datasets fall into one correlation cluster), only RS can be assessed. We call a method *fully robust* for a given (metric, aggregation) pair when it is stable under both indicators, and *RS-robust* when it is stable across schemes but DS could not be evaluated. **Cross-consistency.** The per-(metric, aggregation) analysis answers "which methods are robust when we report metric m aggregated by rule a?". Practitioners, however, often do not know in advance which metric or aggregation they will ultimately report. We therefore also ask which methods remain top-ranked *regardless* of that choice. For each ranking $r^{(s,m,a,c)}$, let the top-η set be $\mathcal{Q}_\eta^{(s,m,a,c)} = \{i \mid r_i^{(s,m,a,c)} \le \eta\}$. For a fixed aggregation rule a, the *metric-agnostic* (task-agnostic) cross-consistency of method i counts in how many metrics it consistently appears among the top-η, as $\mathrm{CT}^{(s,a)}(\eta, i) = \sum_{m=1}^{M} \frac{1}{|S^{(m,a)}|} \sum_{c \in S^{(m,a)}} \mathbf{1}\left(i \in \mathcal{Q}_\eta^{(s,m,a,c)}\right)$, where $S^{(m,a)}$ indexes the dataset compositions. Symmetrically, for a fixed metric m, the *aggregation-agnostic* (language-agnostic) cross-consistency counts in how many aggregation rules the method i stays in the top-η. Higher values indicate a method whose strong position does not depend on the choice of metric or aggregation.

4 Experimental Design

Data and methods. We build directly on the next-generation SRBench (Imai Aldeia et al., 2025), which reports, for each (method, dataset) pair, the metric values of $u = 30$ independent runs. Because our robustness analysis requires every method–dataset combination to be available, we exclude two methods (GP-GOMEA and uDSR) and three datasets (`1089_USCrime`, `first_principles_absorption`, and `first_principles_kepler`) that had missing combinations. This leaves 23 SR methods—AFP, AFP_FE, AFP_EHC, Bingo, Brush, BSR, E2E, EPLEX, EQL, FEAT, FFX, Genetic Engine, GPLearn, GPZGD, ITEA, NeSymRes, Operon, PS-Tree, PySR, QLattice, Rils-Rols, TIR, and TPSR—and 21 datasets (11 real-world black-box datasets and 10 law-based first-principles datasets). **Analysis dimensions.** We evaluate four metrics ($\mathrm{MSE}_{\text{test}}$, $\mathrm{MAE}_{\text{test}}$, R^2_{test}, and training time in seconds) and nine aggregation rules (*mean*, *median*, *max*, *min*, *standard deviation*, and four *random single-run picks*), giving 36 per-(metric, aggregation) analyses per dataset track. Each analysis is repeated on three dataset tracks: *all* datasets, *black-box* (real-world) only, and *first-principles* (law-based) only. This separation lets us detect methods that are strong on real-world data but weak on law-based data and vice versa.

We use five MCDM methods: the Weighted Sum Model (WSM), the Technique for Order Preference by Similarity to an Ideal Solution (TOPSIS) (Chen & Hwang, 1992), the VlseKriterijumska Optimizacija I Kompromisno Resenje (VIKOR) (Opricovic, 1998), and the Ranking Organization Method for Enrichment Evaluations (PROMETHEE II) (Jean-Pierre, 1982), alias PRM2, with the usual (PRM2U) and Gaussian (PRM2G) preference functions. Each MCDM method is combined with three objective weighting strategies: the equal weighting, the Method based on the Removal Effects of Criteria (MEREC) (Keshavarz-Ghorabaee et al., 2021), and the CRiteria Importance Through Inter-criteria Correlation (CRITIC) (Diakoulaki et al., 1995). This yields $K = 15$ ranking schemes drawn from distinct families (value-based, distance-based, compromise-based, and outranking), ensuring scheme diversity rather than arbitrary selection. For details, see Appendix A. For each (metric, aggregation, track) we produce a summary dendrogram (ranks across all 15 schemes $\times$ 4 dataset compositions, i.e., all datasets and 3 subsets of performance uncorrelated datasets), an RS sensitivity heatmap, and a DS sensitivity heatmap. We additionally produce the metric-agnostic (one per aggregation rule) and aggregation-agnostic (one per metric) cross-consistency heatmaps

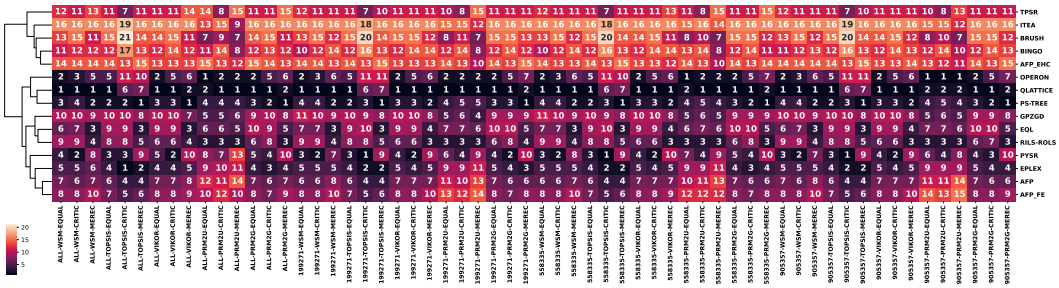


Figure 1: Summary dendrogram for R^2_{test} (mean, all datasets). Rows are methods (union of top-10); columns are the 15 ranking schemes $\times$ four dataset compositions.

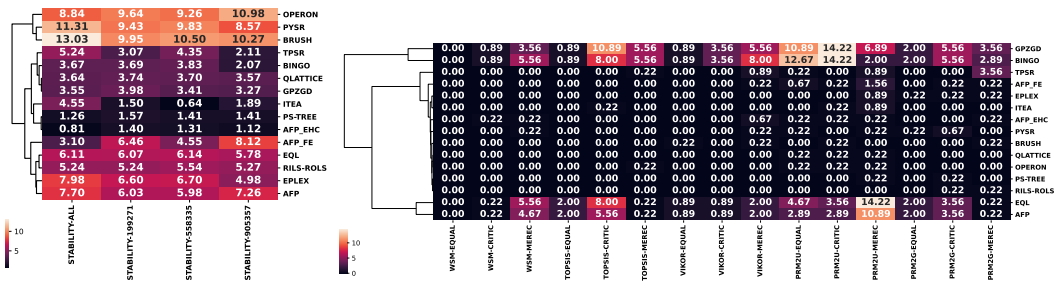
for every track. For the cross-consistency calculation, we set $\eta = 10$, taking the union of top-10 methods across ranking schemes in the visualizations.

Validation. As external datasets for validation that are not in the *call for action* version of SRBENCH (Imai Aldeia et al., 2025), we adopt five datasets, consolidated in the SR community (White et al., 2013; Orzechowski et al., 2018; Vanneschi et al., 2014; Bonin et al., 2024; Rovito et al., 2025a;b). In particular, we use three real-world and two law-based datasets. The real-world datasets are: `airfoil` with 1502 records and 5 variables (Brooks et al., 1989), `concrete` with 1029 records and 8 variables (Castelli et al., 2013), and `slump` with 102 records and 9 variables (Yeh, 2009). The law-based are: `keijzer-6` with 170 records and 1 variable (Keijzer, 2003; Streeter & Becker, 2003), and `pagie-1` with 676 records and 2 variables (Pagie & Hogeweg, 1997). To check that the robust core is not a benchmark datasets' artifact, we re-run the five overall-best methods that are the outcome of our analysis (QLATTICE, OPERON, RILS-ROLS, PS-TREE, and EQL) on the aforementioned datasets and repeat the robustness-driven pipeline for this held-out *validation* track.

5 RESULTS

Metric-specific, aggregation-specific robustness. Figure 1 presents the summary dendrogram for R^2_{test} aggregated by the mean over all repetitions. Rows are the methods in the union of top-10 lists that are results from all applied ranking schemes; columns are the 15 ranking schemes crossed with the four dataset compositions (i.e., *all* datasets and three subsets of uncorrelated datasets, with correlation threshold 0.9, identified by the random seeds 199271, 558335, and 905357). A sensitivity analysis using correlation thresholds of 0.7 and 0.8 is presented in Appendix G. We selected these thresholds based on the guidance for strong correlations in Schober et al. (2018). Both thresholds yield similar patterns, suggesting that the findings are not sensitive to this choice. The appendix reports the two robustness indicators for each threshold; the corresponding ranking heatmaps are available in our repository. A complementary analysis of ranking concordance via Kendall's W is reported in Appendix H, confirming that rankings are highly stable across different stochastic dataset compositions regardless of the random seed employed, with only isolated sensitivity observed when varying weighting criteria and aggregators.

First, a compact top cluster of methods occupies the best ranks under nearly every scheme and composition. For mean R^2 this cluster is led by: OPERON, QLATTICE, and PS-TREE, followed by EQL, RILS-ROLS, GPZGD, and PYSR, with the latter two occasionally exhibiting bad ranks (outside the top-10) for some schemes and compositions. PYSR has rank 13 for `ALL-PRM2U-MEREC`, while GPZGD has rank 11 for `199271-WSM-EQUAL` and `558335-WSM-EQUAL`. Notice that also OPERON itself sometimes exhibits a rank of 11. However, since in all the other schemes and compositions this method is consistently ranked good (contrary to GPZGD), it is, anyway, clustered with QLATTICE and PS-TREE, which are the ones exhibiting the best ranks overall in the dendrogram. Second, the bottom of the dendrogram is dominated by methods like: TPSR, AFP_EHC, and ITEA. In general, the bottom of the dendrograms for the other metrics, aggregation rules, schemes, and compositions is mostly dominated by methods like: TPSR, AFP, AFP_EHC, AFP_FE and ITEA, as shown in Appendix B. The separation between these groups is preserved across columns, indicating that the coarse good/bad split is robust to the choice of scheme and dataset composition.



(a) RS (b) DS

Figure 2: Robustness indicators for R^2_{test} (mean, all datasets). Darker (lower) values indicate more robust rankings (less variance).

Figure 2 isolates the two robustness indicators for the same setting (mean R^2). The RS map (left) shows that the top cluster in Figure 1, composed by QLATTICE and PS-TREE, attains small rank variance across the 15 schemes, i.e. these two methods would be ranked the same (top rank) regardless of the MCDM method used. Also the worst methods in Figure 1 (i.e., ITEA and AFP_EHC) exhibit small rank variance, since they are consistently ranked bad regardless of the MCDM, and this trend also holds for GPZGD, which exhibits rank between 8 and 11 in most cases. Other methods exhibit higher rank variance because, depending upon the MCDM, they are assigned substantially different ranks (e.g., OPERON is in the top-3 in most cases, but for some specific schemes it is ranked 10 or 11). The DS map (right) shows that for most methods and schemes the variance is basically zero, meaning that most methods are robust w.r.t. dataset compositions. Notable exceptions are: GPZGD and BINGO for PRM2U and `TOPSIS-CRITIC`, EQL and AFP for `PRM2U-MEREC`.

Comparison with baseline rankings. Figure 55 (see Appendix I) shows a heatmap containing the Spearman rank correlation coefficient (Spearman, 2010) values computed between each ranking scheme and a simple baseline where SR methods are ranked according to their median performance across all the datasets (a baseline adopted by the SRBENCH). Given a metric and an aggregation rule, the Spearman coefficient appears to be stable across all schemes, except for TOPSIS, particularly in combination with certain metrics and weighting methods (e.g., R^2 mean and CRITIC), where the correlation is slightly lower.

Effect of the aggregation rule. Sweeping the aggregation rule (Appendix B) shows that mean and median produce almost identical top clusters, confirming that the standard practice of reporting either is stable. A notable case is represented by PS-TREE, which shows substantially worse performance when represented by median w.r.t. the mean aggregation, suggesting that even though median performance of this method is modest, it occasionally finds extremely good solutions. The optimistic *max* aggregation—closest to the "best-of-30-runs" view adopted by the next-generation SRBENCH (Imai Aldeia et al., 2025)—slightly promotes stochastic search methods that occasionally find excellent expressions, while the *standard deviation* aggregation, which ranks by run-to-run stability rather than central accuracy, reorders the middle of the field substantially: some moderately accurate methods are comparatively *unstable* across runs (e.g., EQL, GPZGD, PYSR), a distinction invisible to a mean-based leaderboard. The random single-run aggregations lie between mean and max and almost never overturn the identity of the top cluster, except for PS-TREE, RILS-ROLS, and EQL that occasionally exhibit bad ranks due to single, unlucky runs. QLATTICE, on the other hand, is consistently both stable and top-performer even in the single-run random picks.

Metric-agnostic robustness. Figure 3 shows the metric-agnostic cross-consistency for the mean aggregation over all datasets: for each method it counts, across the four metrics, how often it appears among the top-10 under each ranking scheme. A small set of methods—again GPZGD, RILS-ROLS, and EQL, together with PS-TREE, PYSR, OPERON and QLATTICE—remain top performers across all four metrics, indicating that their strength is not tied to a particular error definition. Training time is the metric that most reshuffles the ranking: methods with heavy inner parameter-optimization loops (which are highly accurate, such as OPERON and QLATTICE) fall in the time ranking, while lightweight methods climb (e.g., GPZGD, E2E, FFX). Consequently, methods that are simultaneously accurate *and* fast (e.g., EQL and RILS-ROLS) form an even smaller robust core,

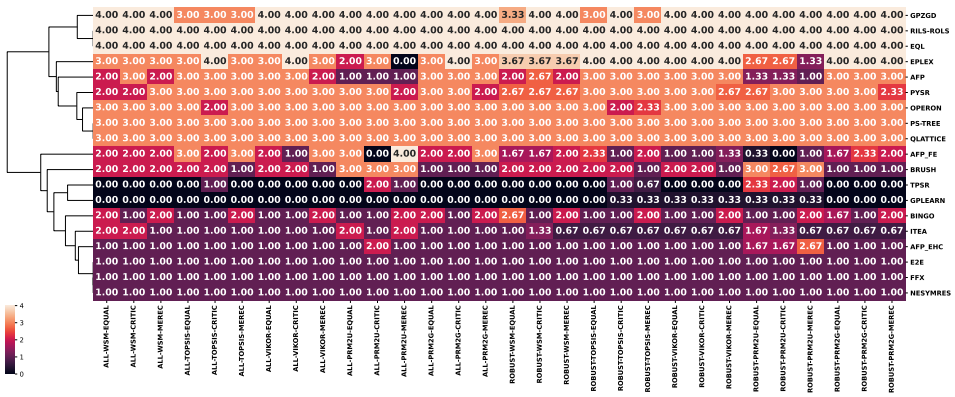


Figure 3: Metric-agnostic cross-consistency (mean, all datasets). Cell values count how often each method (row) is in the top-10 across the four metrics under each ranking scheme (column).

a trade-off that a single accuracy leaderboard cannot express. Broken down by metric, the three accuracy metrics agree closely—MSE, MAE, and R^2 provide essentially the same robust core (OPERON, EQL, QLATTICE, with RILS-ROLS and PS-TREE)—whereas training time is the lone outlier that demotes the optimization-heavy methods and promotes lightweight ones. This metric-agnostic core is moreover stable across the other run-aggregation rules with some exceptions (e.g., in the median aggregation, BINGO and BRUSH exhibit high cross-consistency while PS-TREE drops). Other metric-agnostic heatmaps are reported in Appendix C.

Aggregation-agnostic robustness. Figure 4 shows the complementary aggregation-agnostic analysis for R^2_{test}: for each method it counts, across the nine aggregation rules, how often it stays in the top-10 under each ranking scheme. OPERON, QLATTICE, and RILS-ROLS are the most aggregation-agnostic methods, remaining top-ranked whether one reports the mean, median, max, min, standard deviation, or a random single run; PS-TREE, EQL, PYSR, BINGO, BRUSH, and GPZGD follow. This means the apparent superiority of these methods is not an artifact of *how* the repeated runs are summarized. In contrast, several mid-field methods (e.g., EPLEX, FEAT, GPLEARN, E2E) drift in and out of the top-10 as the aggregation changes, and appear stronger or weaker purely depending on a reporting convention. Broken down by aggregation rule, the mean, median, and the four random single-run picks yield almost identical aggregation-invariant cores, the optimistic max mildly favors stochastic search methods, and the standard-deviation view rewards run-to-run stability rather than peak accuracy. The same aggregation-agnostic pattern recurs for every error metric except for the training time, where only EQL, GPZGD, and RILS-ROLS remain in the top-cluster. A set of per-metric aggregation-agnostic heatmaps are provided in Appendix D.

Real-world versus law-based datasets. A recurring concern in the SR community is that a method may excel on law-based, ground-truth problems yet fail on messy real-world data (or the reverse) (Matsubara et al., 2024; Imai Aldeia et al., 2025). Splitting the analysis into the black-box (real-world) and first-principles (law-based) tracks reveals exactly such a divergence. On the real-world track (Figure 5), the robust core is led by EQL, PS-TREE, GPZGD, and PYSR, with OPERON and QLATTICE exhibiting bad ranks. On the law-based track (Figure 6), QLATTICE and OPERON rise to the very top instead, alongside BINGO and PS-TREE, while, on the other hand, EQL and GPZGD drop. QLATTICE's probabilistic, structure-searching approach appears especially suited to recovering the compact closed forms underlying first-principles data, but this advantage does not transfer as strongly to noisy real-world regression. This is a concrete example of a method whose rank is *not* robust to the real-versus-law-based composition of the benchmark, and it cautions against reporting a single pooled leaderboard.

Validation on external datasets. When repeating the full pipeline for the five overall top methods on five held-out external datasets, we observe that OPERON and PS-TREE are consistently in the top-3 for the regression metrics and all aggregation rules, under both robustness indicators, while EQL is the fastest approach. By also analyzing Figure 7 and accounting for the separation between real-world and law-based datasets, we observe that the robust core identified on the benchmark is not an artifact of the specific benchmark datasets: EQL and PS-TREE lead on the three real-world

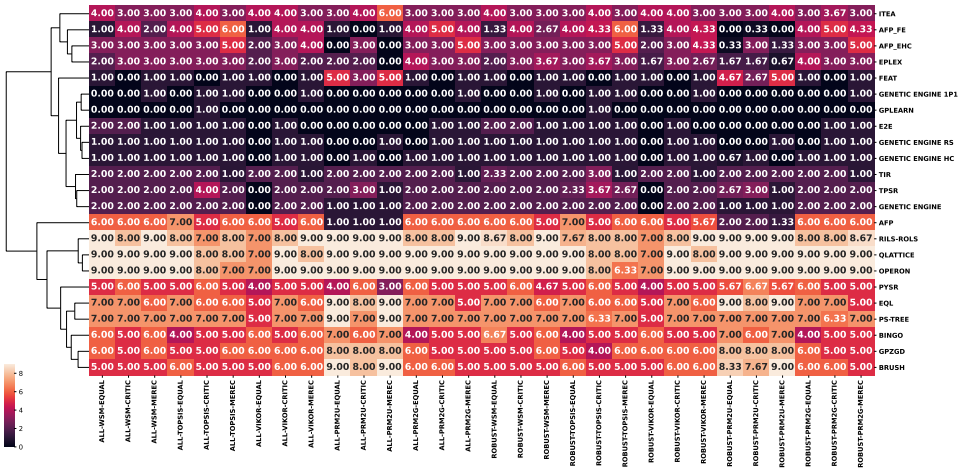


Figure 4: Aggregation-agnostic cross-consistency for R^2_{test} (all datasets). Cell values count how often each method stays in the top-10 across the nine aggregation rules under each ranking scheme.

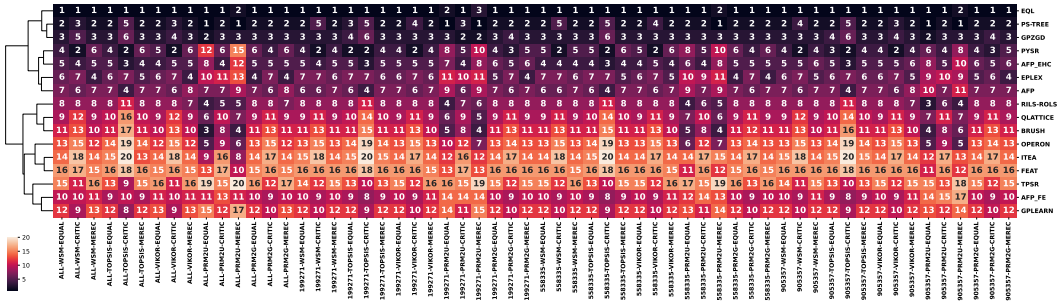


Figure 5: Summary dendrograms for R^2_{test} (mean) on the real-world track.

datasets, with especially PS-TREE winning against all other methods in this validation setting, while they perform worse on the law-based. At the same time, QLATTICE and OPERON lead on the two law-based datasets, while QLATTICE performs the worst among the five methods on the three real-world datasets. The different distributions depicted in Figure 7 are exactly what a single averaged leaderboard number hides, and it motivates asking how stable the induced *rankings* are once we vary the metric, the aggregation of these runs, and the datasets.

6 DISCUSSION

Robustness profiles. Aggregating our findings, three profiles emerge: (i) *Fully robust methods* occupy the best ranks across ranking schemes, dataset compositions, metrics, and aggregations. EQL is the clearest example, with PS-TREE close behind on real-world data. (ii) *Regime-dependent methods* are strong only under particular choices—QLATTICE and OPERON on law-based data, stochastic search methods under the optimistic max aggregation, or lightweight methods under the training-time metric. Their apparent quality is genuine but conditional, and a single leaderboard would misrepresent them in one direction or the other. (iii) *Consistently weak methods*, chiefly the transformer/neural approaches (E2E, NESYMRES, TPSR, AFP) in this tabular-regression regime, stay low-ranked everywhere. **Implications for SR benchmarking.** Our analysis operationalizes part of the community's *call for action* (Imai Aldeia et al., 2025). First, because mean and median rankings are nearly identical, except for the outlined exception represented by PS-TREE, but the standard-deviation and training-time rankings differ substantially, leaderboards should report at least a central-tendency ranking *and* a stability/efficiency ranking, rather than a single accuracy number. Second, because the real-world and law-based tracks disagree on their winners, benchmarks should keep the two tracks separate and report per-track robustness. Third, and most directly,

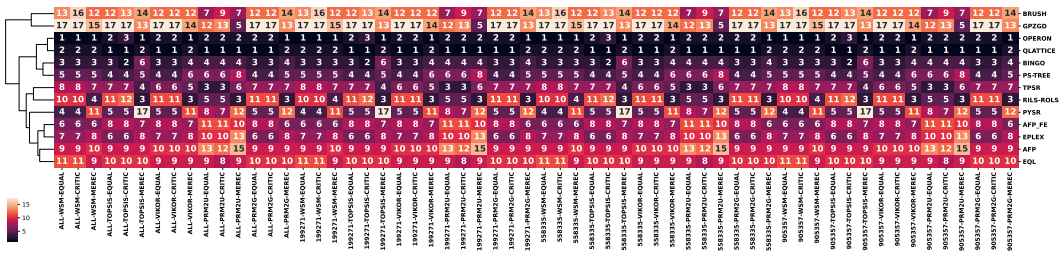


Figure 6: Summary dendrograms for R^2_{test} (mean) on the law-based track.

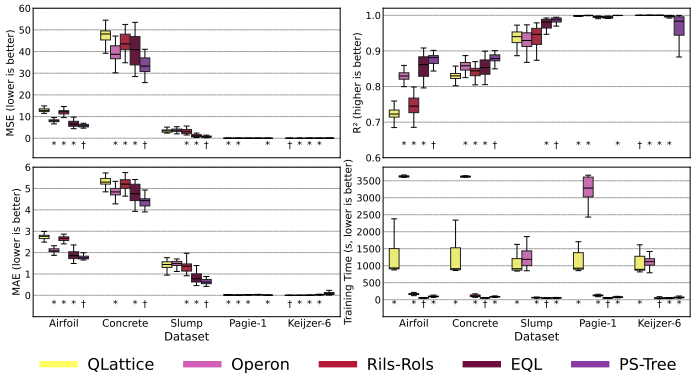


Figure 7: Distribution of four evaluation metrics across five top benchmark methods on external datasets. Within each group, an asterisk marks methods that outperform at least one other (Mann-Whitney U test (Mann & Whitney, 1947)) and a cross marks the method that outperforms all others after Holm-Bonferroni correction (Holm, 1979). All tests are preceded by a Kruskal-Wallis test (Kruskal & Wallis, 1952) at $\alpha = 0.05$.

we recommend that future SRBENCH iterations attach the two robustness indicators (RS and DS) to every reported ranking, so that a claimed state-of-the-art is accompanied by evidence that it survives reasonable variation in the evaluation design. Crucially, none of these choices introduces opaque engineering—the MCDM portfolio and the correlation threshold are fully documented and their sensitivity is itself quantified, which is precisely the transparency that a bare average across datasets lacks. **Limitations.** Our study operates in the space of *reported performance* and is a benchmarking meta-study, not a model-selection method; it does not model per-dataset uncertainty beyond the aggregation rules, and the rankings are ordinal and can obscure the size of performance gaps. We also do not analyze model complexity as a fifth criterion, since for most methods in SRBENCH it was not computed. Assessing model complexity, however, is a natural extension given its central role in interpretable SR. Finally, our results reflect a single benchmark snapshot, but the code can be rerun as the benchmark evolves. Although we report rankings, the repository also provides the underlying continuous values, allowing readers to assess the magnitude of performance differences.

7 CONCLUSION

We presented a robustness meta-study of symbolic regression (SR) benchmark rankings, adapting a multi-criteria robustness framework to the metric and run-aggregation axes that are specific to SR evaluation. Across 23 methods and 21 datasets, and separately for real-world and law-based data, we found that only a small core of methods—most consistently PS-TREE and EQL on real-world data and QLATTICE and OPERON on law-based data—remains top-ranked regardless of the ranking scheme, dataset composition, error metric, or run-aggregation rule. The remaining methods are strong only under specific evaluation designs. These results support the community's call for a living SR benchmark and argue concretely that robustness indicators, rather than single averaged scores, should accompany SR leaderboards. All analysis artifacts are released to enable the analysis to be re-run as the benchmark grows.

AI USE STATEMENT

In this work, we used generative AI tools for performing minor text manipulation to improve overall flow and readability. We have not used generative AI tools for interpreting the results or generating synthetic data. Additionally, we used generative AI to summarize specific paragraphs of the paper with the sole aim to better fit the page limit. We have reviewed all AI-assisted work, and we take responsibility for the final content of this work.

ETHICS STATEMENT

After reading, we fully adhere to the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

The analysis is built entirely on the publicly released per-run scores of the next-generation SRBENCH (Imai Aldeia et al., 2025). Section 3 and Section 4 specify the exact set of methods, datasets, metrics, aggregation rules, aggregators and weighting strategies (ranking schemes), dataset-subsampling procedure, and robustness indicators. All summary, sensitivity, and cross-consistency figures for every (metric, aggregation, track) combination are provided in the Appendix, and the full set of generated figures is released with the paper also via our dedicated repository at https://anonymous.4open.science/r/RobustSRBench-B094/.

AUTHOR CONTRIBUTIONS

Anonymous statement for double-blindness.

ACKNOWLEDGMENTS

Anonymous statement for double-blindness.

REFERENCES

117th US Congress. Algorithmic accountability act act, 2022. URL https://www.congress.gov/bill/117th-congress/house-bill/6580/.

Jamal Abdalla, M Naser, Saleh Alogla, Alireza Ghasemi, and Ahmad Naser. Evaluation and benchmarking of performance of machine learning and symbolic regression: Datasets, software tools and prediction models. *ES General*, 7:1352, 2025. ISSN 2576-9898. doi: 10.30919/esg1352. URL http://dx.doi.org/10.30919/esg1352.

Guilherme Seidyo Imai Aldeia and Fabrício Olivetti de França. Interpretability in symbolic regression: a benchmark of explanatory methods using the Feynman data set. *Genetic Programming and Evolvable Machines*, pp. 1–41, 2022. doi: 10.1007/s10710-022-09435-x.

Josh C Bongard and Hod Lipson. Nonlinear system identification using coevolution of models and tests. *IEEE Transactions on Evolutionary Computation*, 9(4):361–384, 2005.

Lorenzo Bonin, Luigi Rovito, Andrea De Lorenzo, and Luca Manzoni. Cellular geometric semantic genetic programming. *Genetic Programming and Evolvable Machines*, 25(1):1–32, 2024. doi: 10.1007/s10710-024-09480-8.

Thomas F Brooks, D Stuart Pope, and Michael A Marcolini. Airfoil self-noise and prediction. Technical report, NASA, 1989.

Mauro Castelli, Leonardo Vanneschi, and Sara Silva. Prediction of high performance concrete strength using genetic programming with geometric semantic genetic operators. *Expert Systems with Applications*, 40(17):6856–6862, 2013. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2013.06.037. URL https://www.sciencedirect.com/science/article/pii/S0957417413004326.

William La Cava, Tilak Raj Singh, James Taggart, Srinivas Suri, and Jason Moore. Learning concise representations for regression by evolving networks of trees. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Hke-JhA9Y7.

Shu-Heng Chen. *Genetic algorithms and genetic programming in computational finance*. Springer Science & Business Media, 2012.

Shu-Jen Chen and Ching-Lai Hwang. *Fuzzy Multiple Attribute Decision Making Methods*, pp. 289–486. Springer Berlin Heidelberg, Berlin, Heidelberg, 1992. ISBN 978-3-642-46768-4. doi: 10.1007/978-3-642-46768-4_5. URL https://doi.org/10.1007/978-3-642-46768-4_5.

Pierre Colombo, Nathan Noiry, Ekhine Irurozki, and Stephan Clémençon. What are the best systems? new perspectives on nlp benchmarking. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 26915–26932. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/ac4920f4085b5662133dd751493946a6-Paper-Conference.pdf.

European Commission. Artificial intelligence act, 2021. URL https://artificialintelligenceact.eu/.

Miles Cranmer. Interpretable machine learning for science with pysr and symbolicregression.jl, 2023. URL https://arxiv.org/abs/2305.01582.

Marco Crespi, Andrea Ferigo, Leonardo Lucio Custode, and Giovanni Iacca. A population-based approach for multi-agent interpretable reinforcement learning. *Applied Soft Computing*, 147: 110758, 2023. ISSN 1568-4946. doi: https://doi.org/10.1016/j.asoc.2023.110758. URL https://www.sciencedirect.com/science/article/pii/S1568494623007767.

F. O. de Franca, M. Virgolin, M. Kommenda, M. S. Majumder, M. Cranmer, G. Espada, L. Ingelse, A. Fonseca, M. Landajuela, B. Petersen, R. Glatt, N. Mundhenk, C. S. Lee, J. D. Hochhalter, D. L. Randall, P. Kamienny, H. Zhang, G. Dick, A. Simon, B. Burlacu, Jaan Kasak, Meera Machado, Casper Wilstrup, and W. G. La Cava. Srbench++: Principled benchmarking of symbolic regression with domain-expert interpretation. *IEEE Transactions on Evolutionary Computation*, 29(4): 1127–1137, 2025. doi: 10.1109/TEVC.2024.3423681.

Fabrício Olivetti De França. Transformation-interaction-rational representation for symbolic regression: A detailed analysis of srbench results. *ACM Trans. Evol. Learn. Optim.*, 3(2), June 2023. ISSN 2688-299X. doi: 10.1145/3597312. URL https://doi.org/10.1145/3597312.

D. Diakoulaki, G. Mavrotas, and L. Papayannakis. Determining objective weights in multiple criteria problems: The critic method. *Computers & Operations Research*, 22(7):763–770, 1995. ISSN 0305-0548. doi: https://doi.org/10.1016/0305-0548(94)00059-H. URL https://www.sciencedirect.com/science/article/pii/030505489400059H.

Grant Dick. Genetic programming, standardisation, and stochastic gradient descent revisited: initial findings on srbench. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, GECCO '22, pp. 2265–2273, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392686. doi: 10.1145/3520304.3534040. URL https://doi.org/10.1145/3520304.3534040.

Junlan Dong and Jinghui Zhong. Recent advances in symbolic regression. *ACM Comput. Surv.*, 57 (11), June 2025. ISSN 0360-0300. doi: 10.1145/3735634. URL https://doi.org/10.1145/3735634.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Gabriel Sequeira, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, Ömer Çağatan, Akash Kundu, Martin Bernstorff, Shitao Xiao, Akshita Sukhlecha, Bhavish Pahwa, Rafał Poświata, Kranthi Kiran GV, Shawon Ashraf,

Daniel Auras, Björn Plüster, Jan Philipp Harries, Loïc Magne, Isabelle Mohr, Mariya Hendriksen, Dawei Zhu, Hippolyte Gisserot-Boukhlef, Tom Aarsen, Jan Kostkan, Konrad Wojtasik, Taemin Lee, Marek Šuppa, Crystina Zhang, Roberta Rocca, Mohammed Hamdy, Andrianos Michail, John Yang, Manuel Faysse, Aleksei Vatolin, Nandan Thakur, Manan Dey, Dipam Vasani, Pranjal Chitale, Simone Tedeschi, Nguyen Tai, Artem Snegirev, Michael Günther, Mengzhou Xia, Weijia Shi, Xing Han Lù, Jordan Clive, Gayatri Krishnakumar, Anna Maksimova, Silvan Wehrli, Maria Tikhonova, Henil Panchal, Aleksandr Abramov, Malte Ostendorff, Zheng Liu, Simon Clematide, Lester James Miranda, Alena Fenogenova, Guangyu Song, Ruqiya Bin Safi, Wen-Ding Li, Alessia Borghini, Federico Cassano, Hongjin Su, Jimmy Lin, Howard Yen, Lasse Hansen, Sara Hooker, Chenghao Xiao, Vaibhav Adlakha, Orion Weller, Siva Reddy, and Niklas Muennighoff. Mmteb: Massive multilingual text embedding benchmark, 2025. URL https://arxiv.org/abs/2502.13595.

R.P. Feynman. *Lectures on Physics: The Definitive Edition*. Pearson/ Addison Wesley, 2006. ISBN 9780805390476. URL https://books.google.com.br/books?id=AbruAAAAMAAJ.

R.P. Feynman, R.B. Leighton, and M. Sands. *The Feynman Lectures on Physics, Vol. I: The New Millennium Edition: Mainly Mechanics, Radiation, and Heat*. Basic Books, 2015. ISBN 9780465040858. URL https://books.google.com.br/books?id=d76DBQAAQBAJ.

Ana Gjorgjevikj, Barbara Koroušić Seljak, and Tome Eftimov. On the robustness of multilingual text embedding rankings across learning tasks, languages, and benchmark datasets, 2026. URL https://arxiv.org/abs/2605.31142.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pp. 65–70, 1979. URL https://www.jstor.org/stable/4615733.

Quang Huynh, Hemant Singh, Tapabrata Ray, and Akira Oyama. Improved genetic programming for symbolic regression: Case studies on practical applications. In *2022 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1135–1142, 2022. doi: 10.1109/SSCI51031.2022.10022279.

Guilherme Seidyo Imai Aldeia, Hengzhe Zhang, Geoffrey Bomarito, Miles Cranmer, Alcides Fonseca, Bogdan Burlacu, William G. La Cava, and Fabrício Olivetti de França. Call for action: towards the next generation of symbolic regression benchmark. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, GECCO '25 Companion, pp. 2529–2538, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400714641. doi: 10.1145/3712255.3734309. URL https://doi.org/10.1145/3712255.3734309.

Leon Ingelse and Alcides Fonseca. Domain-aware feature learning with grammar-guided genetic programming. In Gisele Pappa, Mario Giacobini, and Zdenek Vasicek (eds.), *Genetic Programming*, pp. 227–243, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-29573-7.

B Jean-Pierre. L'ingénièrie de la décision; elaboration d'instruments d'aide à la decision. la méthode promethee. *L'aidea la Décision: Nature, Instruments et Perspectives d'Avenir*, 1982.

Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of ai ethics guidelines. *Nature machine intelligence*, 1(9):389–399, 2019. doi: 10.1038/s42256-019-0088-2.

Maarten Keijzer. Improving symbolic regression with interval arithmetic and linear scaling. In Conor Ryan, Terence Soule, Maarten Keijzer, Edward Tsang, Riccardo Poli, and Ernesto Costa (eds.), *Genetic Programming*, pp. 70–82, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg. ISBN 978-3-540-36599-0.

Mehdi Keshavarz-Ghorabaee, Maghsoud Amiri, Edmundas Kazimieras Zavadskas, Zenonas Turskis, and Jurgita Antucheviciene. Determination of objective weights using a new method based on the removal effects of criteria (merec). *Symmetry*, 13(4), 2021. ISSN 2073-8994. doi: 10.3390/sym13040525. URL https://www.mdpi.com/2073-8994/13/4/525.

Michael Kommenda, Bogdan Burlacu, Gabriel Kronberger, and Michael Affenzeller. Parameter identification for symbolic regression using nonlinear least squares. *Genetic Programming and Evolvable Machines*, 21(3):471–501, 2020.

John R Koza. Genetic programming as a means for programming computers by natural selection. *Statistics and computing*, 4(2):87–112, 1994.

Gabriel Kronberger, Bogdan Burlacu, Michael Kommenda, Stephan M Winkler, and Michael Affenzeller. *Symbolic regression*. Chapman and Hall/CRC, 2024.

William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952. doi: 10.1080/01621459.1952.10483441.

William La Cava, Kourosh Danai, and Lee Spector. Inference of compact nonlinear dynamic models by epigenetic local search. *Engineering Applications of Artificial Intelligence*, 55:292–306, 2016a.

William La Cava, Kourosh Danai, Lee Spector, Paul Fleming, Alan Wright, and Matthew Lackner. Automatic identification of wind turbine models using evolutionary multiobjective optimization. *Renewable Energy*, 87:892–902, 2016b.

William La Cava, Bogdan Burlacu, Marco Virgolin, Michael Kommenda, Patryk Orzechowski, Fabrício Olivetti de França, Ying Jin, and Jason H Moore. Contemporary symbolic regression methods and their relative performance. *Advances in neural information processing systems*, 2021(DB1):1, 2021.

William B Langdon, Riccardo Poli, Nicholas F McPhee, and John R Koza. Genetic programming: An introduction and tutorial, with a survey of techniques and applications. *Computational intelligence: A compendium*, pp. 927–1028, 2008. doi: 10.1007/978-3-540-78293-3_22.

Fabio Llorella, José Antonio Cebrián, Alberto Corbi, and Antonio María Pérez. Fostering scientific methods in simulations through symbolic regressions. *Physics Education*, 59(4):045010, may 2024. doi: 10.1088/1361-6552/ad3cad. URL https://doi.org/10.1088/1361-6552/ad3cad.

Nour Makke and Sanjay Chawla. Interpretable scientific discovery with symbolic regression: a review. *Artificial Intelligence Review*, 57(1):2, 2024. doi: 10.1007/s10462-023-10622-0.

H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60, 1947. ISSN 00034851. URL https://www.jstor.org/stable/2236101.

Yoshitomo Matsubara, Naoya Chiba, Ryo Igarashi, and Yoshitaka Ushiku. Rethinking symbolic regression datasets and benchmarks for scientific discovery, 2024. URL https://arxiv.org/abs/2206.10540.

C Molnar. Interpretable machine learning. *Lulu. com*, 2020.

Giorgia Nadizar, Eric Medvet, and Dennis G. Wilson. Naturally interpretable control policies via graph-based genetic programming. In Mario Giacobini, Bing Xue, and Luca Manzoni (eds.), *Genetic Programming*, pp. 73–89, Cham, 2024a. Springer Nature Switzerland. ISBN 978-3-031-56957-9.

Giorgia Nadizar, Luigi Rovito, Andrea De Lorenzo, Eric Medvet, and Marco Virgolin. An analysis of the ingredients for learning interpretable symbolic regression models with human-in-the-loop and genetic programming. *ACM Trans. Evol. Learn. Optim.*, 4(1), feb 2024b. doi: 10.1145/3643688.

Giorgia Nadizar, Luigi Rovito, Dennis G. Wilson, and Eric Medvet. Interpretable control competition. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, GECCO '24 Companion, pp. 11–12, New York, NY, USA, 2024c. Association for Computing Machinery. ISBN 9798400704956. doi: 10.1145/3638530.3664051. URL https://doi.org/10.1145/3638530.3664051.

Randal S Olson, William La Cava, Patryk Orzechowski, Ryan J Urbanowicz, and Jason H Moore. Pmlb: a large benchmark suite for machine learning evaluation and comparison. *BioData mining*, 10(1):36, 2017.

Serafim Opricovic. Multicriteria optimization of civil engineering systems. *Faculty of civil engineering, Belgrade*, 2(1):5–21, 1998.

Patryk Orzechowski, William La Cava, and Jason H. Moore. Where are we now? a large benchmark study of recent symbolic regression methods. In *Proceedings of the Genetic and Evolutionary Computation Conference*, GECCO '18, pp. 1183–1190, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356183. doi: 10.1145/3205455.3205539. URL https://doi.org/10.1145/3205455.3205539.

Ludo Pagie and Paulien Hogeweg. Evolutionary consequences of coevolving targets. *Evolutionary Computation*, 5(4):401–418, 1997. doi: 10.1162/evco.1997.5.4.401.

Mark Rofin, Vladislav Mikhailov, Mikhail Florinsky, Andrey Kravchenko, Tatiana Shavrina, Elena Tutubalina, Daniel Karabekyan, and Ekaterina Artemova. Vote'n'rank: Revision of benchmarking with social choice theory. In Andreas Vlachos and Isabelle Augenstein (eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 670–686, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.48. URL https://aclanthology.org/2023.eacl-main.48/.

Joseph D Romano, Trang T Le, William La Cava, John T Gregg, Daniel J Goldberg, Praneel Chakraborty, Natasha L Ray, Daniel Himmelstein, Weixuan Fu, and Jason H Moore. Pmlb v1.0: an open-source dataset collection for benchmarking machine learning methods. *Bioinformatics*, 38(3):878–880, 01 2022. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab727. URL https://doi.org/10.1093/bioinformatics/btab727.

Luigi Rovito and Marco Virgolin. *Interpretable Non-linear Survival Analysis with Evolutionary Symbolic Regression*, pp. 453–462. Association for Computing Machinery, New York, NY, USA, 2025. ISBN 9798400714658. URL https://doi.org/10.1145/3712256.3726446.

Luigi Rovito, Lorenzo Bonin, Luca Manzoni, and Andrea De Lorenzo. An evolutionary computation approach for twitter bot detection. *Applied Sciences*, 12(12):5915, 2022. doi: 10.3390/app12125915.

Luigi Rovito, Lorenzo Bonin, Davide Farinati, Leonardo Vanneschi, Luca Manzoni, Andrea De Lorenzo, and Gloria Pietropolli. Exploring the integration of cellular structures in genetic programming-based methods. In Bing Xue, Luca Manzoni, and Illya Bakurov (eds.), *Genetic Programming*, pp. 120–138, Cham, 2025a. Springer Nature Switzerland. ISBN 978-3-031-89991-1. doi: 10.1007/978-3-031-89991-1_8.

Luigi Rovito, Lorenzo Bonin, Davide Farinati, Leonardo Vanneschi, Luca Manzoni, Andrea De Lorenzo, and Gloria Pietropolli. Semantic-based recombination and mutation in cellular-inspired genetic programming. *Genetic Programming and Evolvable Machines*, 26(2):27, 11 2025b. ISSN 1573-7632. doi: 10.1007/s10710-025-09524-7. URL https://doi.org/10.1007/s10710-025-09524-7.

Luigi Rovito, Paolo Diviacco, Rodrigo José Carbajales, Lorenzo Grio, Alessandro Busato, Mihai Burca, Alberto Viola, Andrea De Lorenzo, Luca Manzoni, Tommaso Padoan, and Alex Rodriguez. Machine learning and symbolic regression for interpretable pm10 estimation in urban air quality monitoring. *Discover Artificial Intelligence*, May 2026. ISSN 2731-0809. doi: 10.1007/s44163-026-01271-7. URL https://doi.org/10.1007/s44163-026-01271-7.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019. doi: 10.1038/s42256-019-0048-x.

Michael Schmidt and Hod Lipson. Distilling free-form natural laws from experimental data. *science*, 324(5923):81–85, 2009.

Michael D Schmidt, Ravishankar R Vallabhajosyula, Jerry W Jenkins, Jonathan E Hood, Abhishek S Soni, John P Wikswo, and Hod Lipson. Automated refinement and inference of analytical models for metabolic networks. *Physical biology*, 8(5):055011, 2011.

Patrick Schober, Christa Boer, and Lothar A Schwarte. Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia*, 126(5):1763–1768, 2018.

Charles Spearman. The proof and measurement of association between two things. *International journal of epidemiology*, 39(5):1137–1150, 2010.

Karolina Stanislawska, Krzysztof Krawiec, and Zbigniew W Kundzewicz. Modeling global temperature changes with genetic programming. *Computers & Mathematics with Applications*, 64(12): 3717–3728, 2012.

Matthew Streeter and Lee A Becker. Automated discovery of numerical approximation formulae via genetic programming. *Genetic Programming and Evolvable Machines*, 4(3):255–286, 2003.

M.E. Thing and S.M. Koksbang. cp3-bench: a tool for benchmarking symbolic regression algorithms demonstrated with cosmology. *Journal of Cosmology and Astroparticle Physics*, 2025(01):040, January 2025. ISSN 1475-7516. doi: 10.1088/1475-7516/2025/01/040. URL http://dx.doi.org/10.1088/1475-7516/2025/01/040.

Teresa Tonelli, Gloria Pietropolli, Luigi Rovito, Luca Manzoni, and Gianpiero Cossarini. An interpretable machine learning approach for alkalinity reconstruction in the mediterranean sea. *Applied Computing and Geosciences*, 30:100345, 2026. ISSN 2590-1974. doi: https://doi.org/10.1016/j.acags.2026.100345. URL https://www.sciencedirect.com/science/article/pii/S2590197426000297.

Silviu-Marian Udrescu and Max Tegmark. Ai feynman: A physics-inspired method for symbolic regression. *Science Advances*, 6(16):eaay2631, 2020. doi: 10.1126/sciadv.aay2631. URL https://www.science.org/doi/abs/10.1126/sciadv.aay2631.

Leonardo Vanneschi, Sara Silva, Mauro Castelli, and Luca Manzoni. Geometric semantic genetic programming for real life applications. *Genetic Programming Theory and Practice XI*, pp. 191–209, 2014. doi: 10.1007/978-1-4939-0375-7_11. URL https://doi.org/10.1007/978-1-4939-0375-7_11.

Marco Virgolin and Solon P. Pissis. Symbolic regression is np-hard, 2022. URL https://arxiv.org/abs/2207.01018.

David R White, James McDermott, Mauro Castelli, Luca Manzoni, Brian W Goldman, Gabriel Kronberger, Wojciech Jaśkowski, Una-May O'Reilly, and Sean Luke. Better gp benchmarks: community survey results and proposals. *Genetic Programming and Evolvable Machines*, 14: 3–29, 2013.

I.-C. Yeh. Simulation of concrete slump using neural networks. *Proceedings of the Institution of Civil Engineers - Construction Materials*, 162(1):11–18, 02 2009. ISSN 1747-650X. doi: 10.1680/coma.2009.162.1.11. URL https://doi.org/10.1680/coma.2009.162.1.11.

Jan Zegklitz and Petr Posik. Benchmarking state-of-the-art symbolic regression algorithms. *Genetic programming and evolvable machines*, 22(1):5–33, 2021.

A MCDM Portfolio Diversity

To capture the sensitivity of the rank to the performance score aggregation mechanism, a heterogeneous portfolio of MCDM methods was created, comprising *WSM*, *TOPSIS* (Chen & Hwang, 1992), *VIKOR* (Opricovic, 1998) and *PROMETHEE II* (Jean-Pierre, 1982). These methods encode substantially different views on what constitutes a preferable alternative (SR method) from the full portfolio. *WSM* rewards alternatives according to their cumulative weighted performance on the criteria (datasets in this case). *TOPSIS* adopts a reference points perspective, favoring alternatives that are simultaneously close to a positive ideal solution and distant from the least desirable solution (negative

ideal solution). *VIKOR* approaches the problem through compromise optimization, balancing overall performance across criteria against the largest criterion-specific disadvantage. *PROMETHEE II*, in contrast, constructs a complete ranking from pairwise preference comparisons, where the difference between two alternatives on each criterion is transformed into a degree of preference through a criterion-specific preference function. This mechanism makes *PROMETHEE II* particularly flexible, as different preference functions can encode different use case specific requirements on how performance differences should influence the ranking. For example, the *usual* preference function maps any positive difference to a value of 1.0 and negative to 0.0, whereas the *Gaussian* preference function smoothly translates differences into increasing preference intensities between 0.0 and 1.0. Considering alternative preference functions additionally allows us to examine whether the resulting rankings depend on how pairwise performance differences are interpreted within the outranking procedure. Consequently, the agreement among these diverse MCDM methods provides strong evidence that a ranking is not an artifact of a particular aggregation logic and can reliably be considered in subsequent studies or real use cases.

The criteria weighting methods is also selected to capture different assumptions about the relative importance of the evaluation criteria. We consider *Equal*, *CRITIC* (Diakoulaki et al., 1995), and *MEREC* (Keshavarz-Ghorabaee et al., 2021) weighting methods, i.e., three objective MCDM weighting strategies. *Equal weighting* assigns equal importance to all criteria and therefore serves as a reference in which no criterion is favored a priori. In contrast, *CRITIC* derives criterion weights from the information contained in the decision matrix, considering both the variability of each criterion and its correlation with the remaining criteria. Criteria that exhibit greater contrast among alternatives while providing information that is less redundant with other criteria consequently receive greater importance. *MEREC* follows a different data-driven principle, i.e., criteria importance is determined by evaluating how strongly the overall assessment of the alternatives changes when a criterion is removed. A criterion whose exclusion produces a larger change in the resulting assessment is assigned a greater weight. Together, these strategies provide different views on the importance of the criteria, allowing us to assess whether the ranking of SR methods remains stable not only across different aggregation mechanisms, but also across different mechanisms to define the criteria importance.

B Aggregation Sensitivity on All Datasets

This appendix reports summary dendrograms for every evaluation metric aggregated by the mean, median, and max over all 21 datasets. Rows are methods (union of top-10); columns are the 15 ranking schemes crossed with the four dataset compositions. Mean and median yield near-identical top clusters, whereas the optimistic max aggregation mildly promotes stochastic search methods.

In this section, we plot:

- The summary dendrograms for R^2_{test} aggregated by mean (Figure 8), median (Figure 9), and max (Figure 10) over all datasets;
- The summary dendrograms for MSE_{test} aggregated by mean (Figure 11), median (Figure 12), and max (Figure 13) over all datasets;
- The summary dendrograms for MAE_{test} aggregated by mean (Figure 14), median (Figure 15), and max (Figure 16) over all datasets;
- The summary dendrograms for training time aggregated by mean (Figure 17), median (Figure 18), and max (Figure 19) over all datasets.

C Metric-Agnostic Robustness

For each aggregation rule, the metric-agnostic heatmap counts how often each method stays in the top-10 across the four metrics under each ranking scheme. A small core (Operon, EQL, GPZGD, RILS-ROLS) is robust regardless of which metric is reported.

In this section, we plot:

- The metric-agnostic cross-consistency for the mean aggregation (all datasets) in Figure 20;

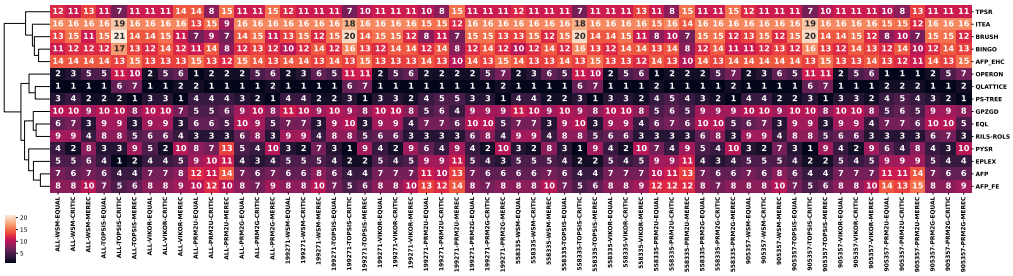


Figure 8: Summary dendrogram for R^2_{test} aggregated by the mean over all datasets.

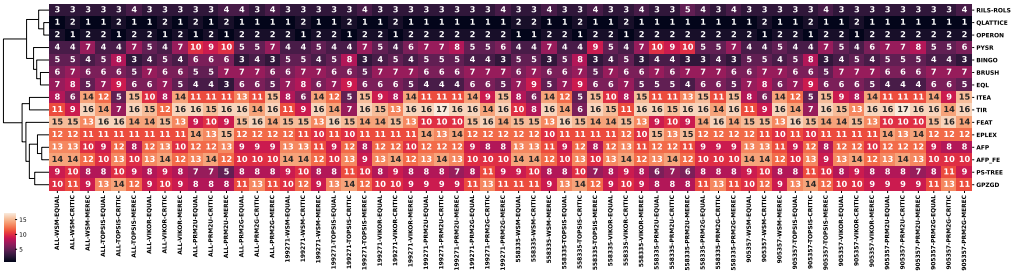


Figure 9: Summary dendrogram for R^2_{test} aggregated by the median over all datasets.

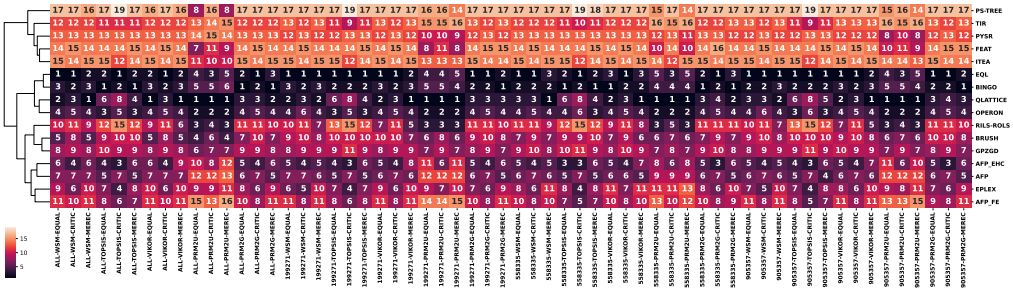


Figure 10: Summary dendrogram for R^2_{test} aggregated by the max over all datasets.

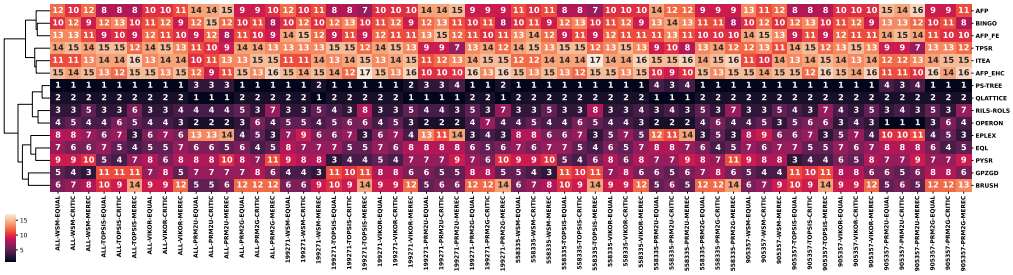


Figure 11: Summary dendrogram for MSE_{test} aggregated by the mean over all datasets.

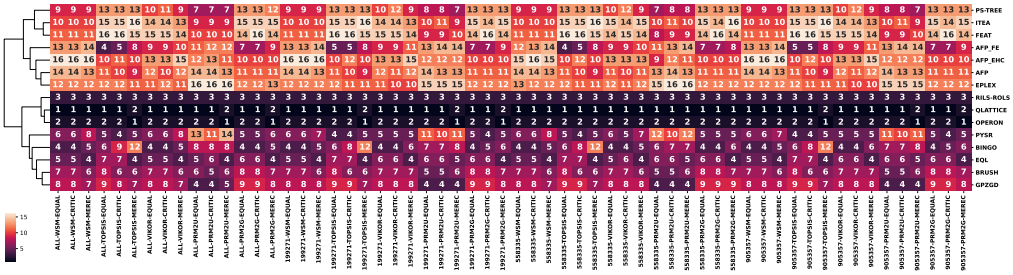


Figure 12: Summary dendrogram for MSE_{test} aggregated by the median over all datasets.

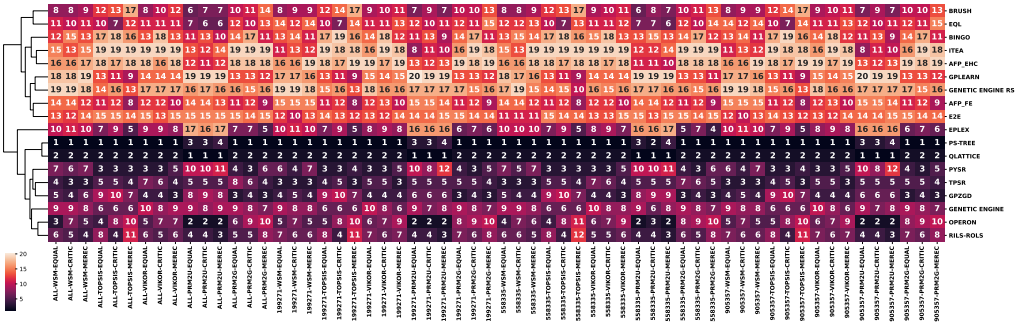


Figure 13: Summary dendrogram for MSE_{test} aggregated by the max over all datasets.

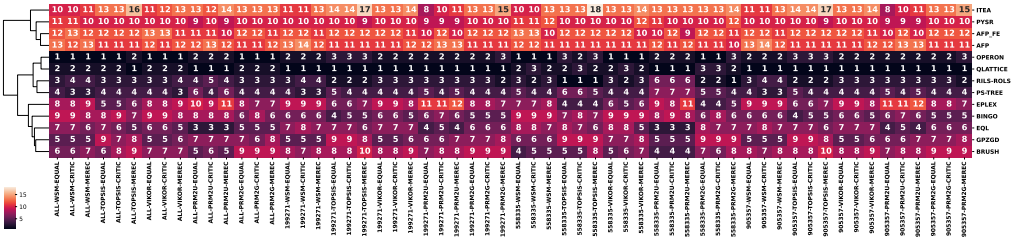


Figure 14: Summary dendrogram for MAE_{test} aggregated by the mean over all datasets.

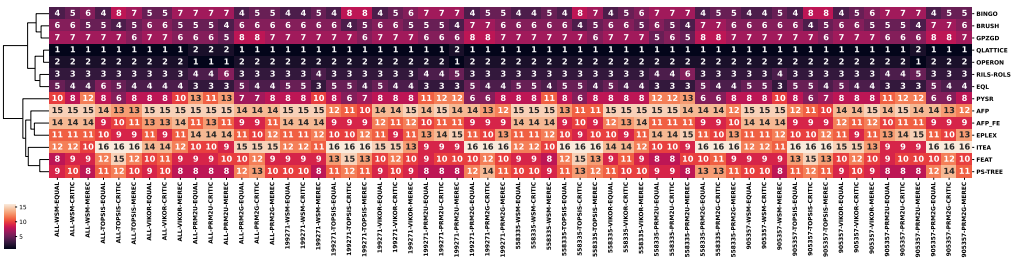


Figure 15: Summary dendrogram for MAE_{test} aggregated by the median over all datasets.

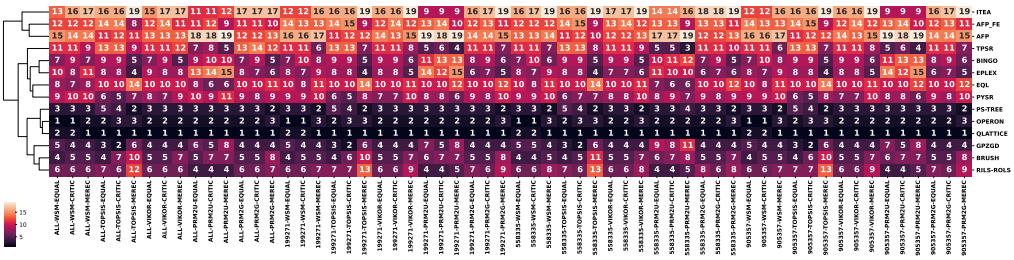


Figure 16: Summary dendrogram for MAE_{test} aggregated by the max over all datasets.

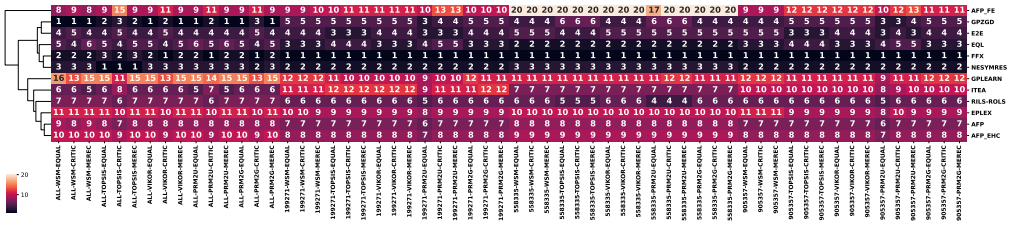


Figure 17: Summary dendrogram for training time aggregated by the mean over all datasets.

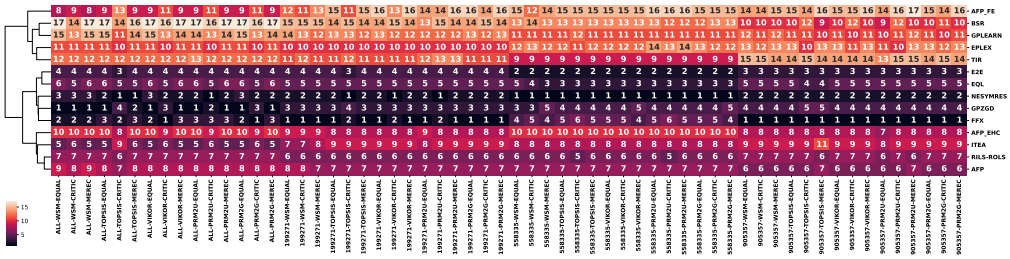


Figure 18: Summary dendrogram for training time aggregated by the median over all datasets.

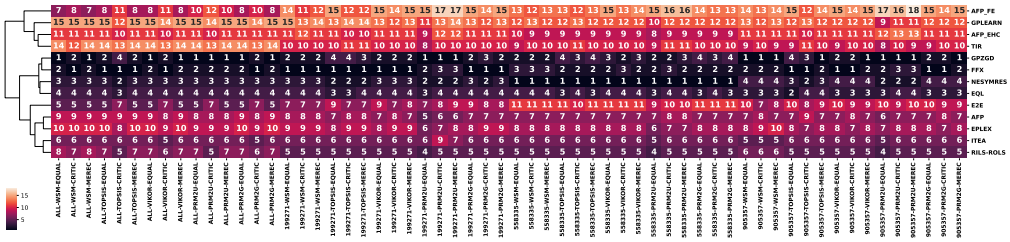


Figure 19: Summary dendrogram for training time aggregated by the max over all datasets.

- The metric-agnostic cross-consistency for the median aggregation (all datasets) in Figure 21;
- The metric-agnostic cross-consistency for the max aggregation (all datasets) in Figure 22;
- The metric-agnostic cross-consistency for the min aggregation (all datasets) in Figure 23;
- The metric-agnostic cross-consistency for the standard deviation aggregation (all datasets) in Figure 24;
- The metric-agnostic cross-consistency for the mean aggregation (black-box track) in Figure 25;
- The metric-agnostic cross-consistency for the mean aggregation (first-principles track) in Figure 26;

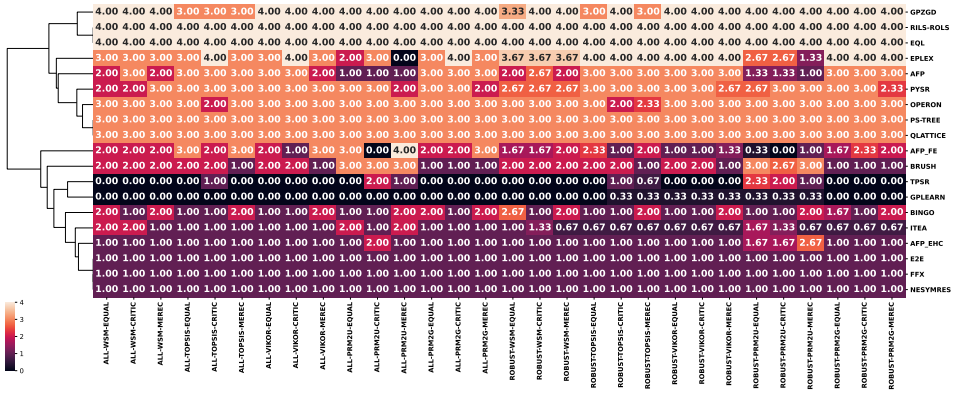


Figure 20: Metric-agnostic cross-consistency for the mean aggregation (all datasets).

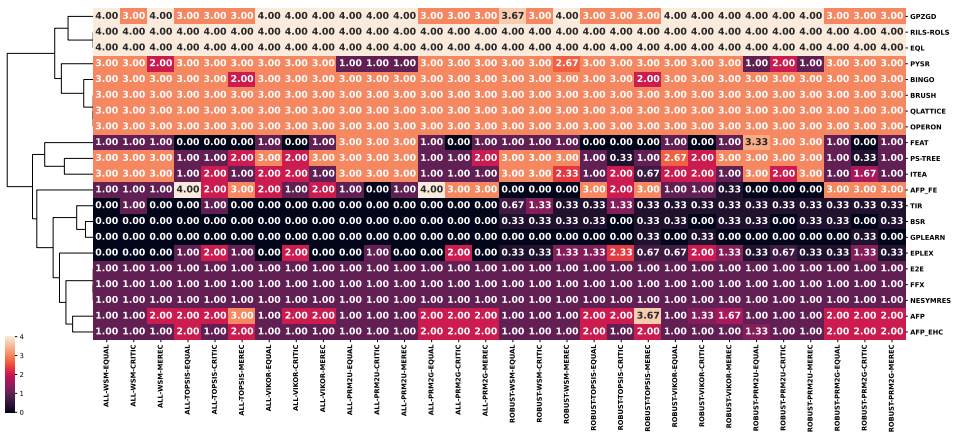


Figure 21: Metric-agnostic cross-consistency for the median aggregation (all datasets).

D AGGREGATION-AGNOSTIC ROBUSTNESS

For each metric, the aggregation-agnostic heatmap counts how often each method stays in the top-10 across the nine aggregation rules under each ranking scheme, reported for the *all*, real-world, and law-based tracks. A small core (OPERON, EQL, GPZGD, RILS-ROLS) is robust regardless of which aggregation rule is reported, except for the training time.

In this section, we plot:

- The aggregation-agnostic cross-consistency for R^2_{test} (all datasets), in Figure 27;

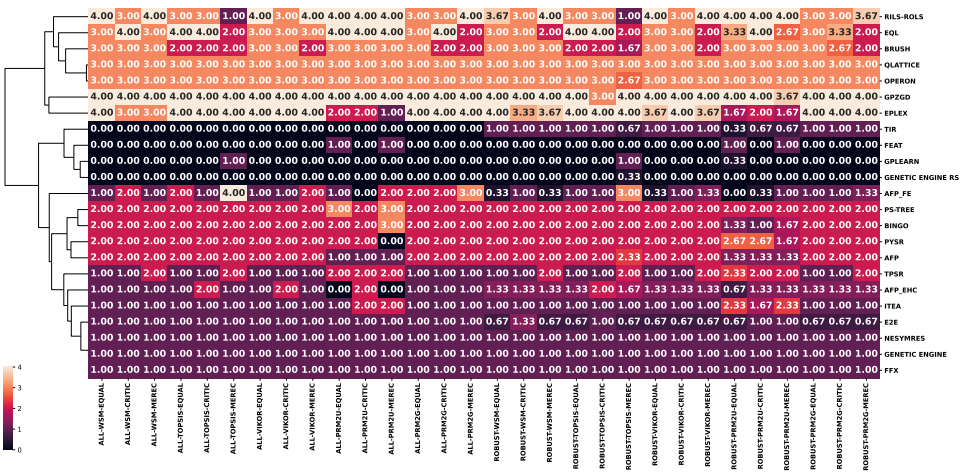


Figure 22: Metric-agnostic cross-consistency for the max aggregation (all datasets).

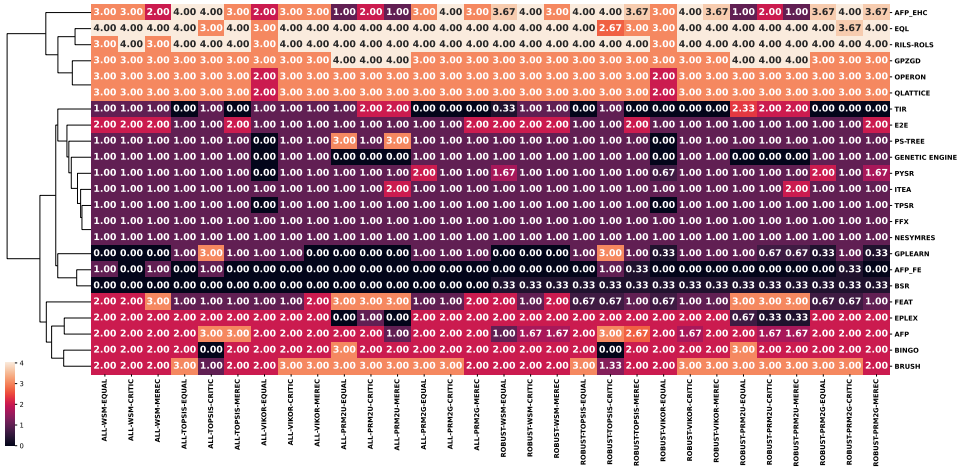


Figure 23: Metric-agnostic cross-consistency for the min aggregation (all datasets).

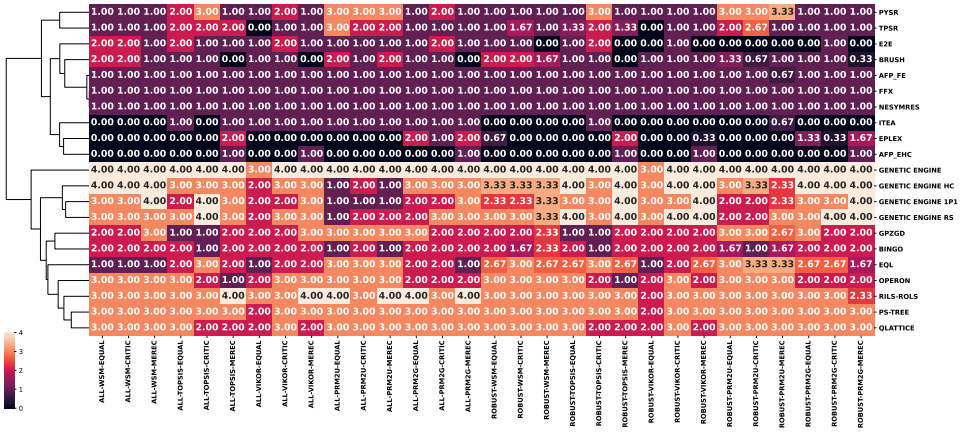


Figure 24: Metric-agnostic cross-consistency for the standard deviation aggregation (all datasets).

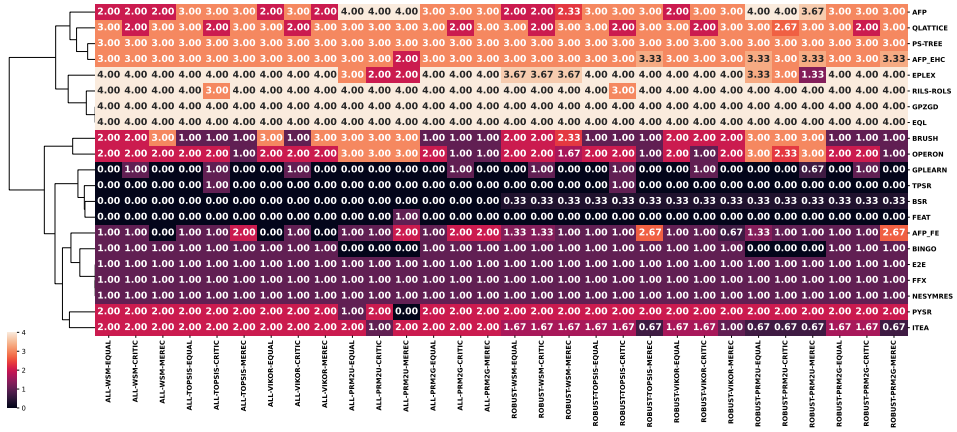


Figure 25: Metric-agnostic cross-consistency for the mean aggregation on the real-world (black-box) track.

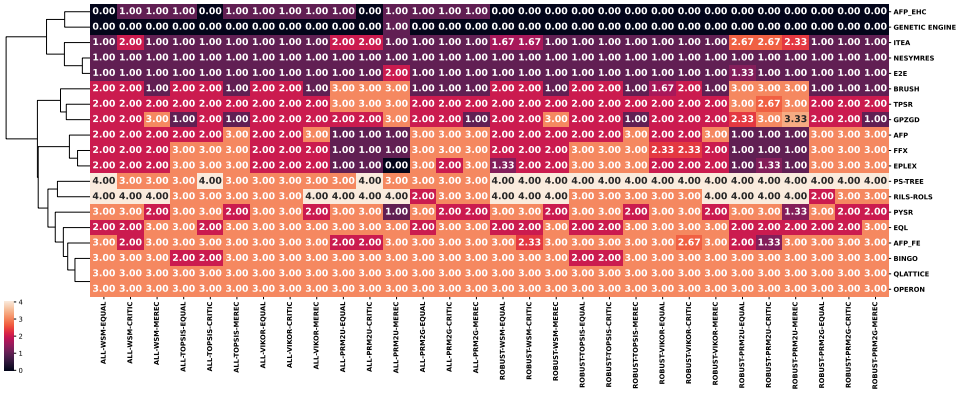


Figure 26: Metric-agnostic cross-consistency for the mean aggregation on the law-based (first-principles) track.

- The aggregation-agnostic cross-consistency for R^2_{test} (black-box track), in Figure 28;
- The aggregation-agnostic cross-consistency for R^2_{test} (first-principles track), in Figure 29;
- The aggregation-agnostic cross-consistency for MSE_{test} (all datasets), in Figure 30;
- The aggregation-agnostic cross-consistency for MSE_{test} (black-box track), in Figure 31;
- The aggregation-agnostic cross-consistency for MSE_{test} (first-principles track), in Figure 32;
- The aggregation-agnostic cross-consistency for MAE_{test} (all datasets), in Figure 33;
- The aggregation-agnostic cross-consistency for MAE_{test} (black-box track), in Figure 34;
- The aggregation-agnostic cross-consistency for MAE_{test} (first-principles track), in Figure 35;
- The aggregation-agnostic cross-consistency for training time (all datasets), in Figure 36;
- The aggregation-agnostic cross-consistency for training time (black-box track), in Figure 37;
- The aggregation-agnostic cross-consistency for training time (first-principles track), in Figure 38.

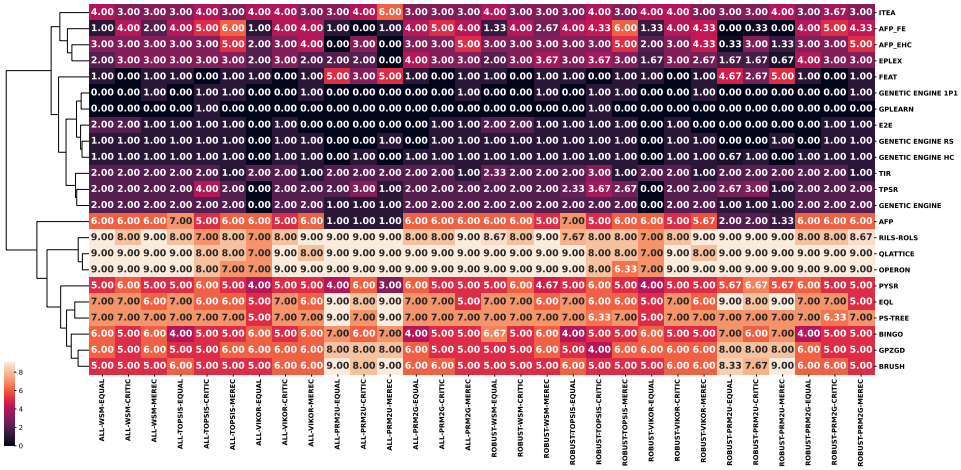


Figure 27: Aggregation-agnostic cross-consistency for R^2_{test} (all datasets).

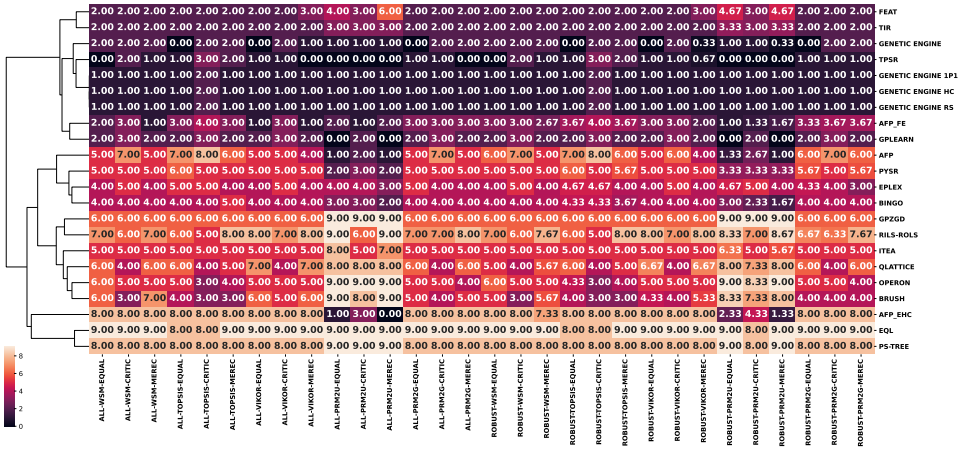


Figure 28: Aggregation-agnostic cross-consistency for R^2_{test} (real-world).

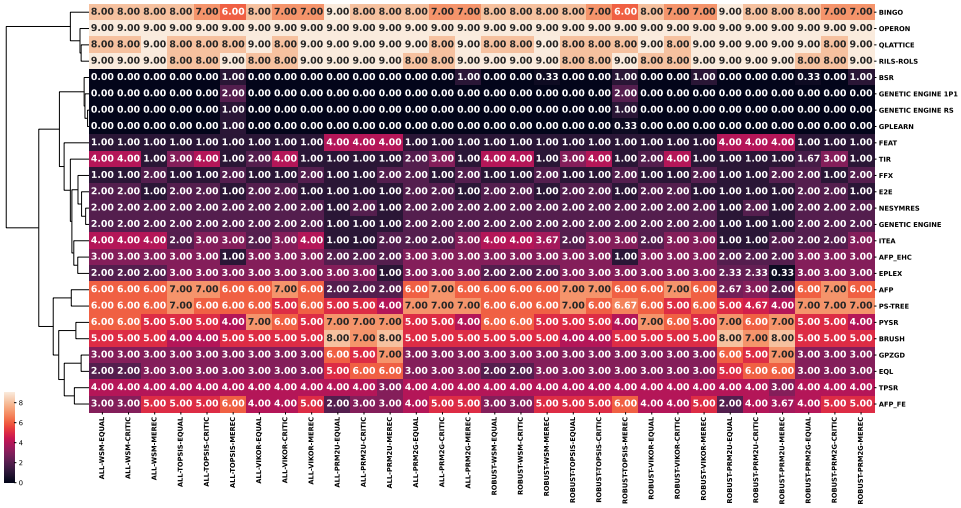


Figure 29: Aggregation-agnostic cross-consistency for R^2_{test} (law-based).

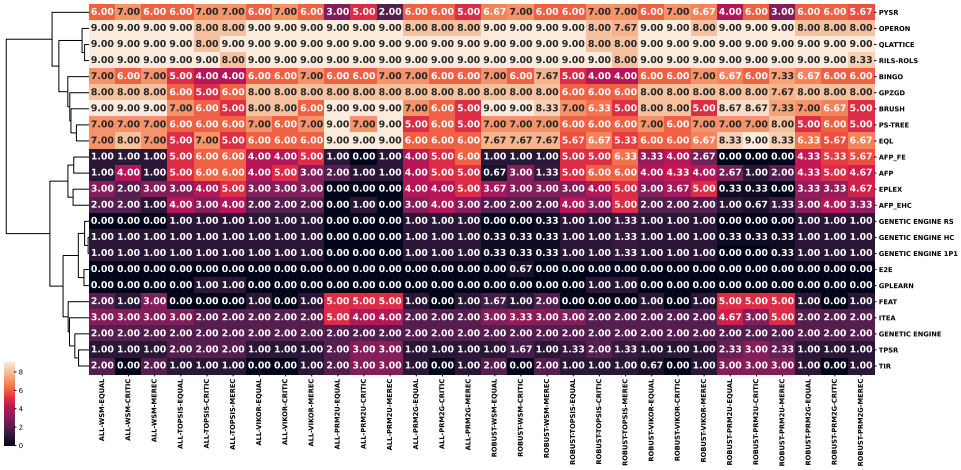


Figure 30: Aggregation-agnostic cross-consistency for MSE_{test} (all datasets).

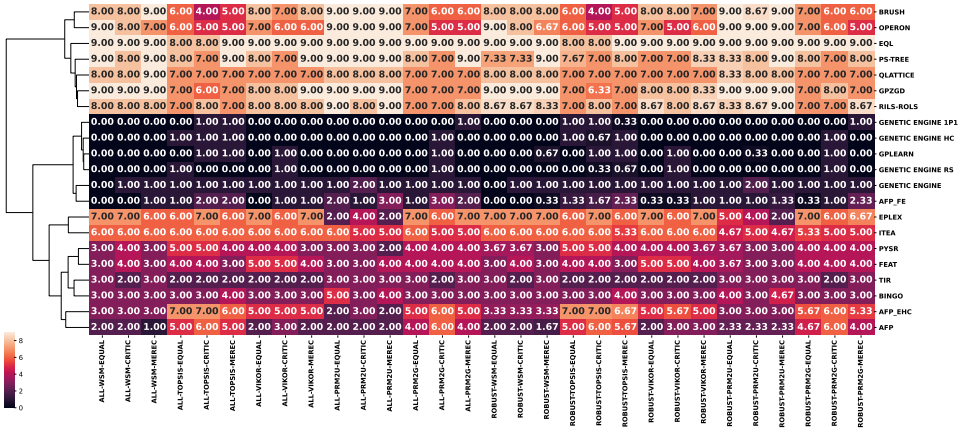


Figure 31: Aggregation-agnostic cross-consistency for MSE_{test} (real-world).

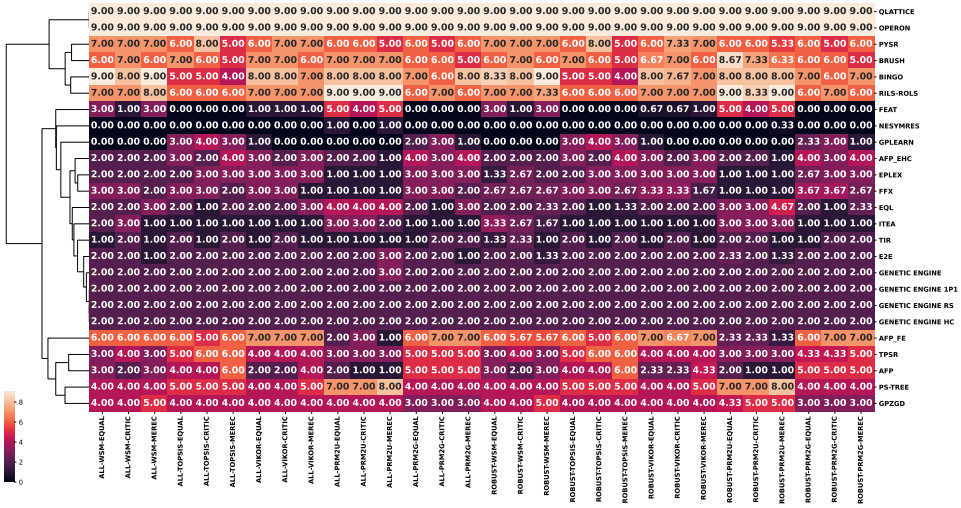


Figure 32: Aggregation-agnostic cross-consistency for MSE_{test} (law-based).

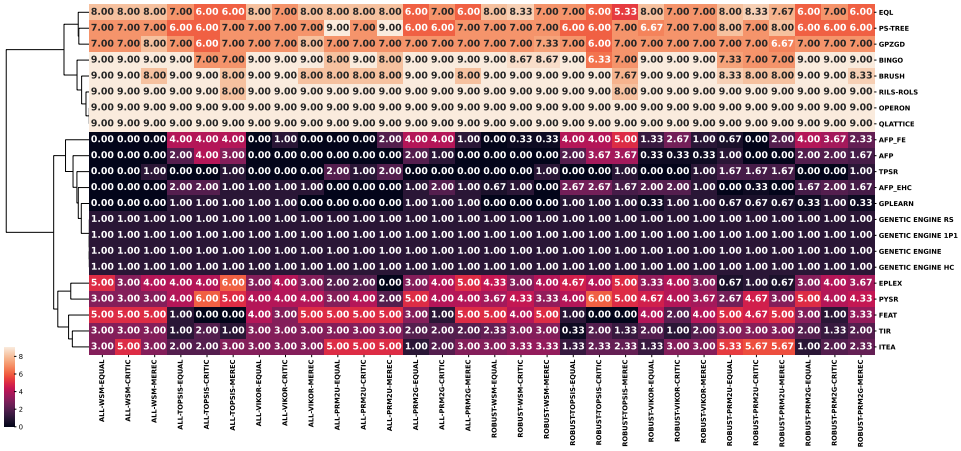


Figure 33: Aggregation-agnostic cross-consistency for MAE_{test} (all datasets).

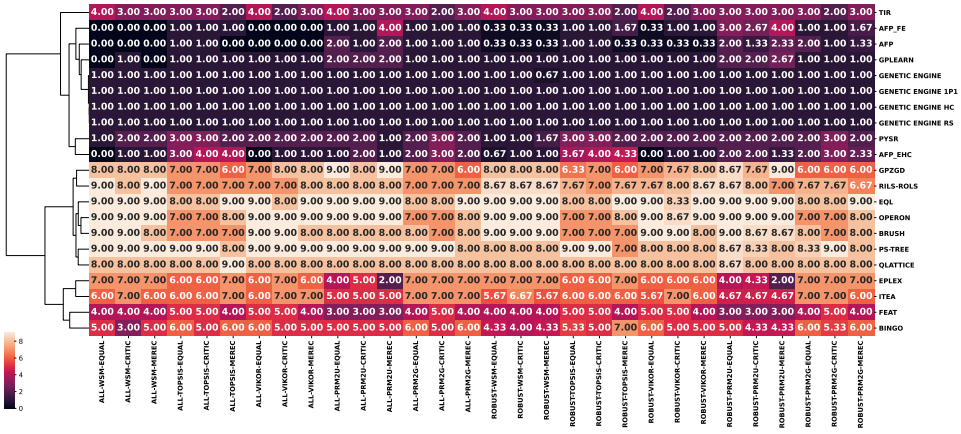


Figure 34: Aggregation-agnostic cross-consistency for MAE_{test} (real-world).

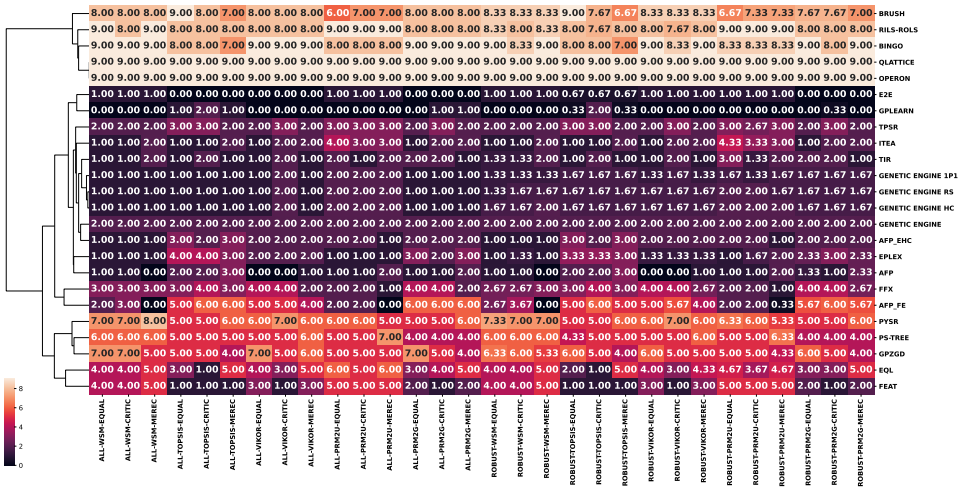


Figure 35: Aggregation-agnostic cross-consistency for MAE_{test} (law-based).

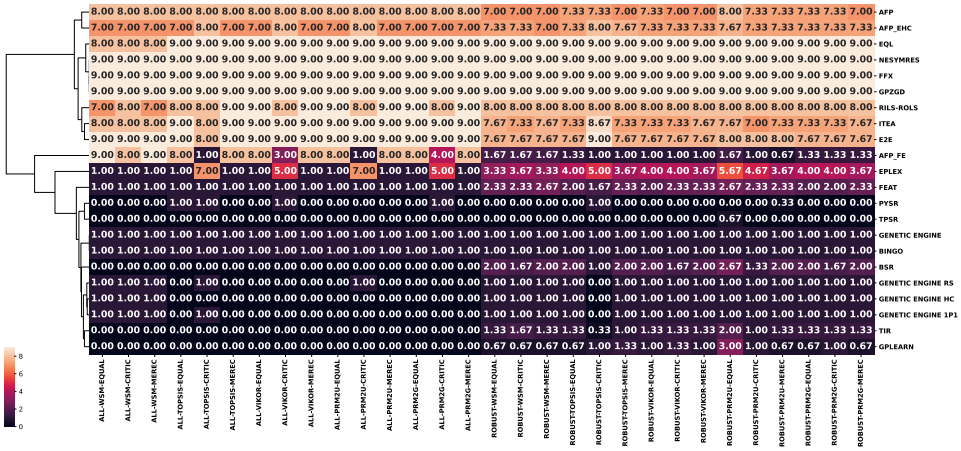


Figure 36: Aggregation-agnostic cross-consistency for training time (all datasets).

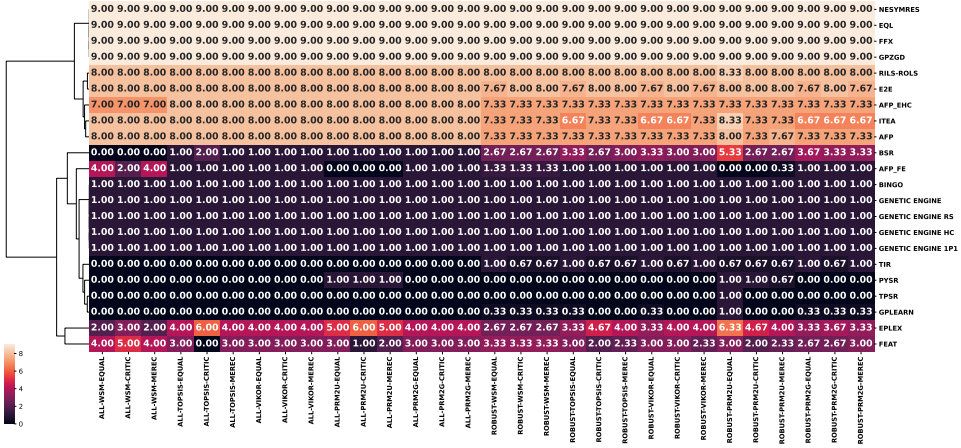


Figure 37: Aggregation-agnostic cross-consistency for training time (real-world).



Figure 38: Aggregation-agnostic cross-consistency for training time (law-based).

E Real-World versus Law-based Summaries (Additional Metrics)

Summary dendrograms (mean aggregation) for the error and training-time metrics on the real-world and law-based tracks, complementing the R^2 comparison in the main text.

In this section, we plot:

- The summary dendrograms for MSE_{test} (mean), over the datasets in the black-box track (Figure 39) and first-principles track (Figure 40);
- The summary dendrograms for MAE_{test} (mean), over the datasets in the black-box track (Figure 41) and first-principles track (Figure 42);
- The summary dendrograms for training time (mean), over the datasets in the black-box track (Figure 43) and first-principles track (Figure 44).

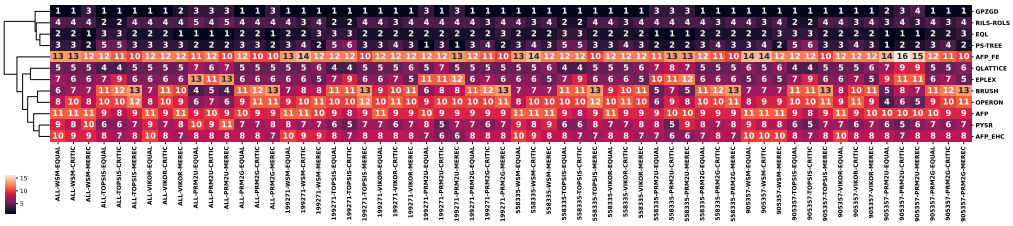


Figure 39: Summary dendrogram for MSE_{test} (mean) on the real-world (black-box) track.

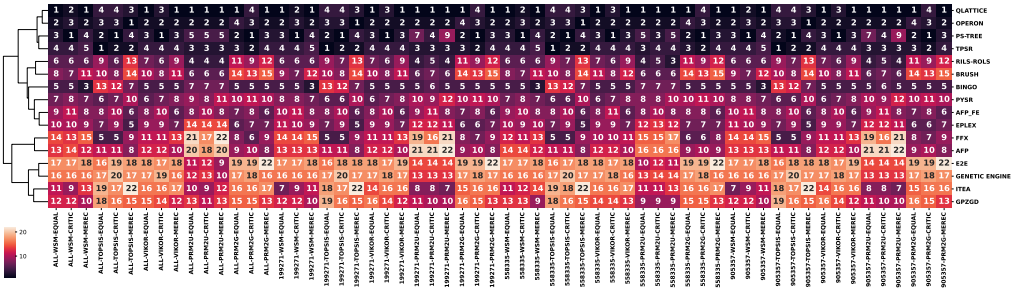


Figure 40: Summary dendrogram for MSE_{test} (mean) on the law-based (first-principles) track.

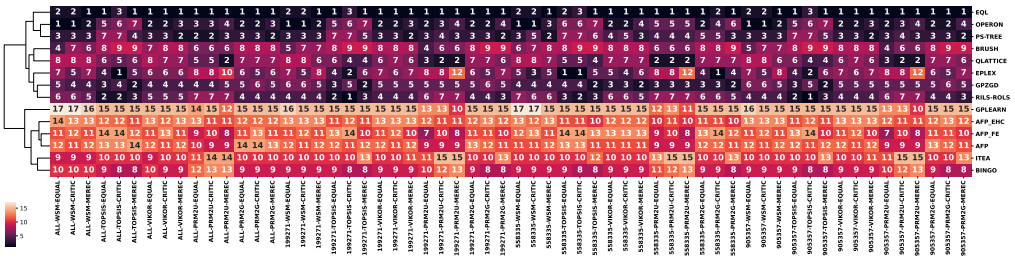


Figure 41: Summary dendrogram for MAE_{test} (mean) on the real-world (black-box) track.

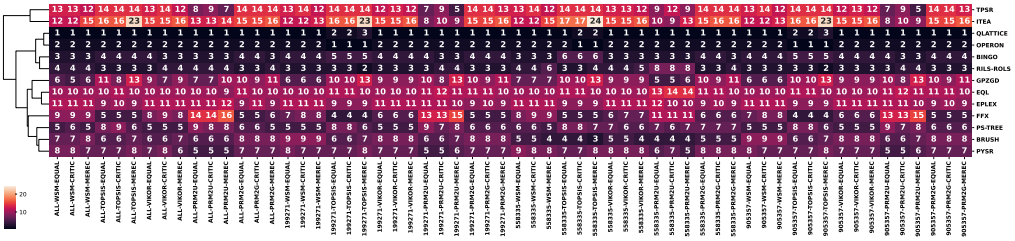


Figure 42: Summary dendrogram for MAE_{test} (mean) on the law-based (first-principles) track.

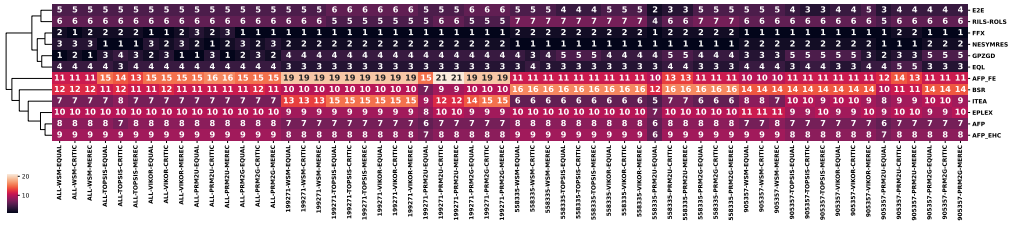


Figure 43: Summary dendrogram for training time (mean) on the real-world (black-box) track.

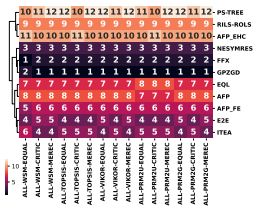


Figure 44: Summary dendrogram for training time (mean) on the law-based (first-principles) track.

F Validation on External Datasets

Robustness analysis restricted to the five overall top methods (QLattice, Operon, Rils-Rols, PS-Tree, EQL) on five external held-out datasets (`airfoil`, `concrete`, `slump`, `keijzer-6`, `pagie-1`).

In this section, we plot:

- The validation summary dendrogram for R^2_{test} (mean), over the external datasets, in Figure 45;
- The validation summary dendrogram for MSE_{test} (mean), over the external datasets, in Figure 46;
- The validation summary dendrogram for MAE_{test} (mean), over the external datasets, in Figure 47;
- The validation summary dendrogram for training time (mean), over the external datasets, in Figure 48;
- The metric-agnostic cross-consistency (mean) on the external validation datasets, in Figure 49.

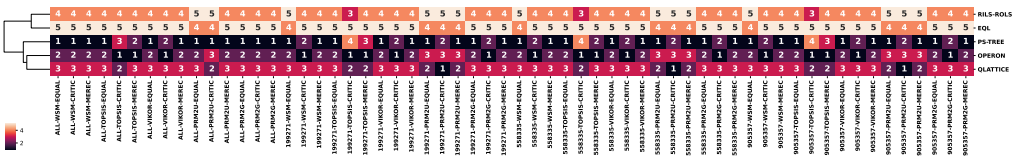


Figure 45: Validation summary dendrogram for R^2_{test} (mean) on the external datasets.

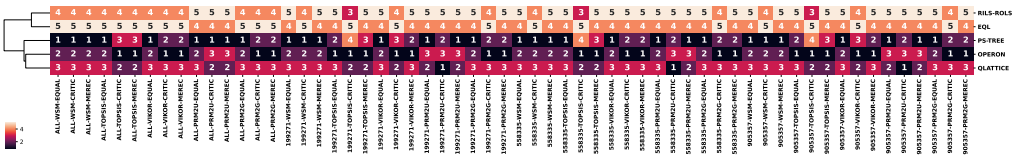


Figure 46: Validation summary dendrogram for MSE_{test} (mean) on the external datasets.

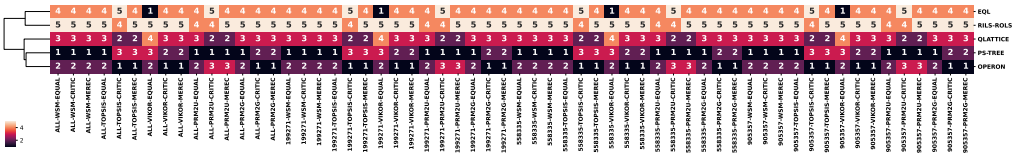


Figure 47: Validation summary dendrogram for MAE_{test} (mean) on the external datasets.

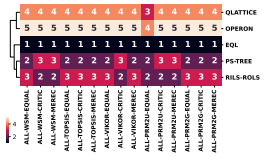


Figure 48: Validation summary dendrogram for training time (mean) on the external datasets.

G Sensitivity of Correlation Thresholds

In this appendix section, we report the robustness indicators RS and DS for the main track under correlation thresholds of 0.8 (Figure 50) and 0.7 (Figure 51), to assess the sensitivity of the rankings to this hyperparameter relative to the value of 0.9 used in the main experimental pipeline.

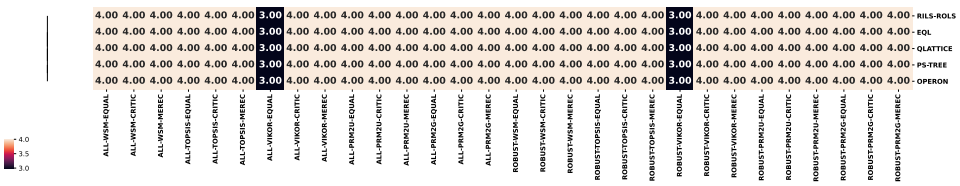
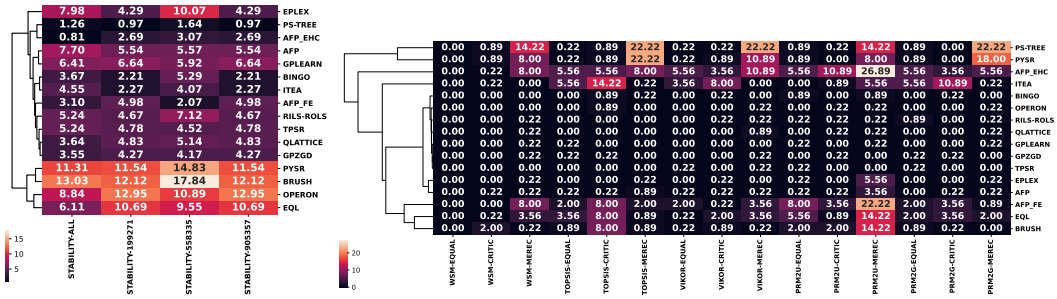


Figure 49: Metric-agnostic cross-consistency (mean) on the external validation datasets.

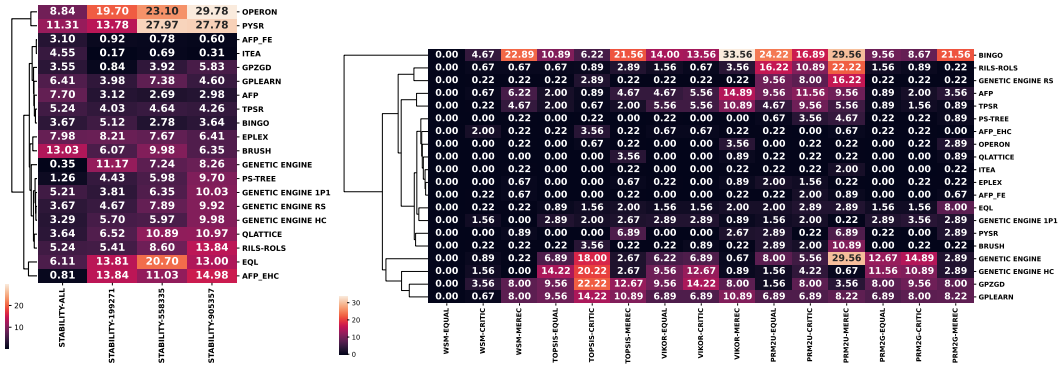
Inspecting the RS plots, we observe that OPERON, EQL, and PYSR consistently exhibit the highest variance across all three threshold values, while PS-TREE, GPZGD, QLATTICE, and RILS-ROLS show low variance at 0.8, in line with the 0.9 case, but higher variance at 0.7 for the dataset compositions associated with random seeds 558335 and 905357. This suggests that lower correlation thresholds can yield less stable rankings depending on the specific dataset composition, an expected outcome since a low threshold may group datasets that are insufficiently correlated.

Inspecting the DS plots, the picture remains largely stable with respect to the 0.9 case for most methods, with isolated exceptions involving high variance for specific ranking schemes. Notably, PS-TREE and PYSR are affected under most *MEREC* configurations at threshold 0.8, and RILS-ROLS under *PRM2U* at threshold 0.7, indicating that certain aggregators and weighting criteria may be sensitive to the choice of correlation threshold.



(a) RS (b) DS

Figure 50: Robustness indicators for R^2_{test} (mean, all datasets) when correlation threshold is 0.8. Darker (lower) values indicate more robust rankings (less variance).



(a) RS (b) DS

Figure 51: Robustness indicators for R^2_{test} (mean, all datasets) when correlation threshold is 0.7. Darker (lower) values indicate more robust rankings (less variance).

H SENSITIVITY OF RANKING SCHEMES AND DATASET COMPOSITIONS

In this appendix section, we analyze how performance rankings are sensitive to the choice of aggregator, weighting criterion, and dataset composition. Specifically, given a metric and an aggregation rule, we compute Kendall's concordance coefficient W across different stochastic dataset compositions for a fixed aggregator and weighting criterion, focusing on stochastic correlation-aware dataset subsampling to assess whether the specific random subsample influences the resulting ranking (Figure 52). All experiments use the main track comprising both black-box and first-principles datasets, with correlation threshold 0.9.

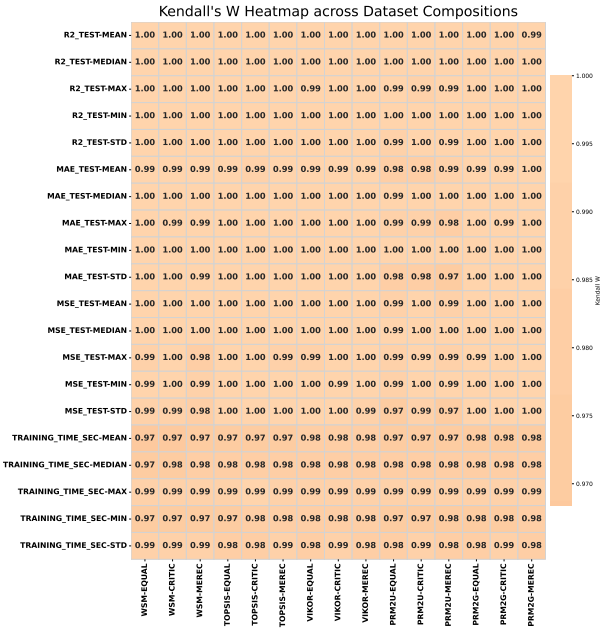


Figure 52: Kendall's concordance W heatmap assessing robustness with respect to stochastic correlation-aware dataset composition.

The heatmap reveals that rankings remain highly consistent regardless of the random seed used for dataset subsampling. We repeat the analysis by fixing the aggregator and dataset composition, and computing Kendall's W across weighting criteria (Figure 53).

Rankings remain stable in most configurations, with the notable exception of *VIKOR* under certain R^2 aggregation rules. Finally, we fix the weighting criterion and dataset composition, and compute Kendall's W across aggregators (Figure 54).

Rankings are again stable across most configurations, with lower concordance observed when the *EQUAL* weighting criterion is paired with certain R^2 aggregation rules, and marginally reduced concordance in several *MEREC* configurations relative to the average case.

I COMPARISON WITH THE BASELINE

In this appendix section, we report heatmaps of the Spearman rank correlation coefficient (Spearman, 2010) between each ranking scheme and a simple baseline in which SR methods are ranked by their median performance across all datasets, as adopted by SRBENCH. Figure 55 shows a concise version for readability, focusing on regression evaluation metrics and the main track without stochastic correlation-aware dataset subsampling, while Figure 56 provides the full extended version for completeness.

Overall, the Spearman correlation is high and stable across schemes for any given metric and aggregation rule, indicating broad agreement with the median baseline. The main exception is TOPSIS,

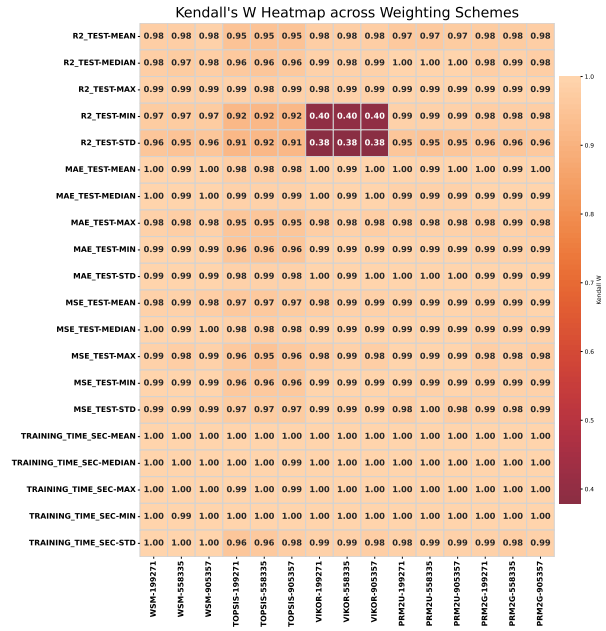


Figure 53: Kendall's concordance W heatmap assessing robustness with respect to weighting criteria.

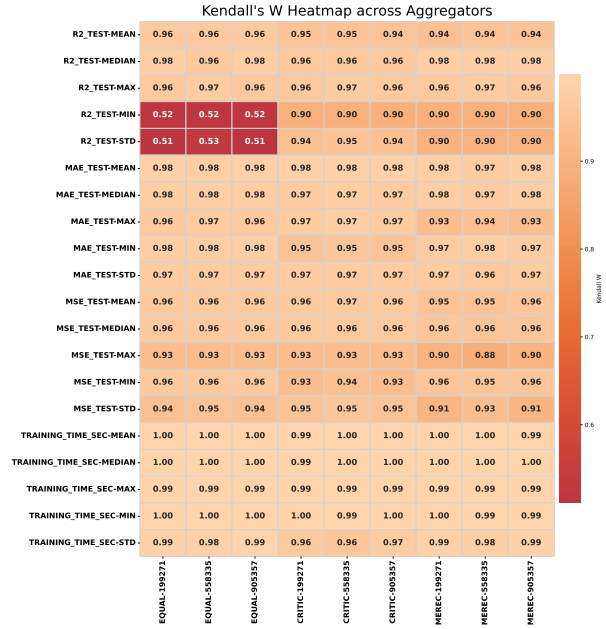


Figure 54: Kendall's concordance W heatmap assessing robustness with respect to aggregators.

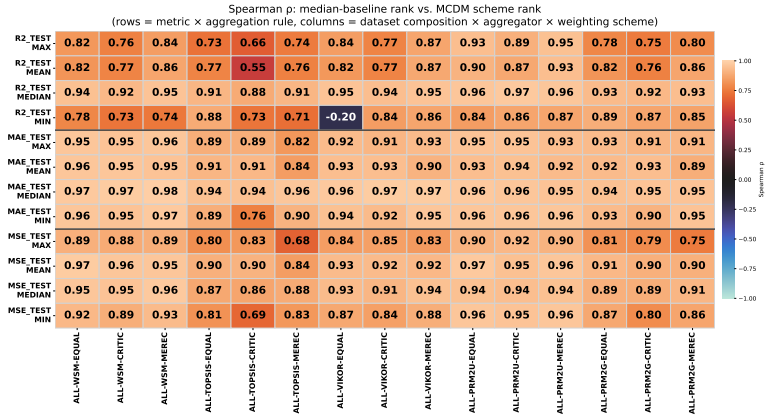


Figure 55: Heatmap showing, for each metric and aggregation rule (with focus on regression evaluation metrics only and the main track containing all datasets without stochastic correlation-aware dataset compositions), the Spearman rank correlation between the ranking provided by the <dataset composition, aggregator, weighting strategy> triple and a ranking provided by simply sorting the models' median performance across all datasets.

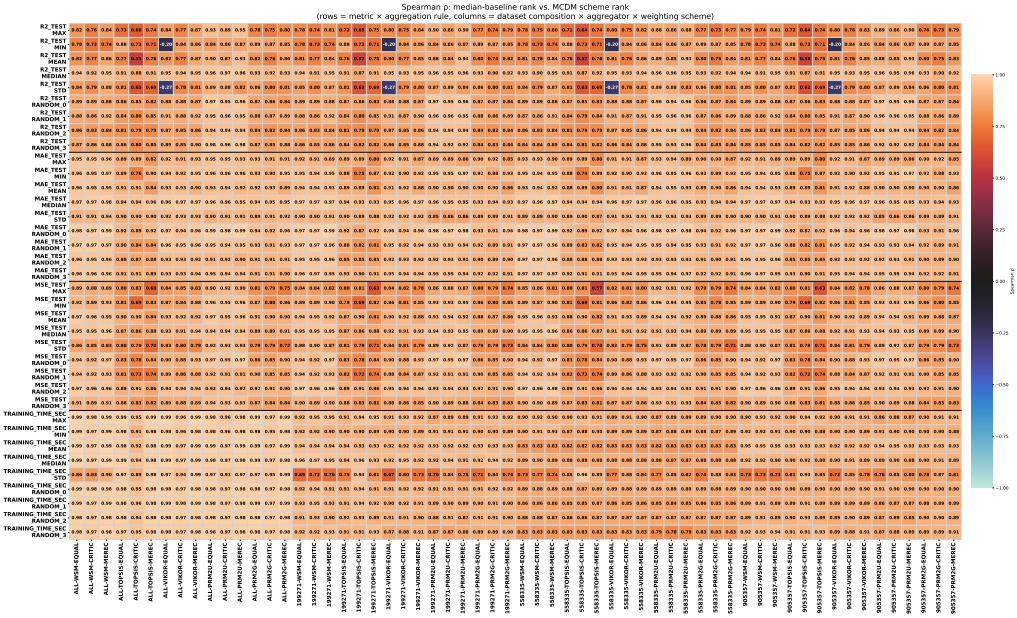


Figure 56: Heatmap showing, for each metric and aggregation rule, the Spearman rank correlation coefficient between the rankings provided by the <dataset composition, aggregator, weighting strategy> triple and a ranking provided by simply sorting the models' median performance across the datasets.

which exhibits slightly lower correlation in specific combinations of metric and weighting criterion, most notably under R^2 with mean aggregation and CRITIC weighting.