This language allows flexible symbolic expressions (e.g., ratios, differences, aggregates) generated through symbolic regression [1, 6], providing interpretable yet expressive formulas.
Example (based on Boston Housing):

$$\frac{\text{RM}}{\text{LSTAT}} > 1.5$$

indicates houses where the ratio of room count to lower-status percentage is high, showing systematic over-prediction.

The goal is to optimize the following objective:

$$d^* = \arg\max_{d \in \mathcal{L}} \lambda \cdot q(d) + (1 - \lambda) \cdot I(d)$$

where $q(d)$ is the subgroup quality (effect size on residuals) [23], $I(d)$ is the interpretability score (inverse of description complexity) [3], and λ is the trade-off parameter

Constraints: The symbolic language $\mathcal{L}_{\text{symb}}$ allows combinations of maximum of 3 (we choose 3, jlt) arithmetic and logical operations built from a restricted operator set $\{+, -, \times, \div, \leq, >, =\}$ and up to two distinct features per description/expression. We permit simple ratio or difference formulas such as $\frac{f_i}{f_j} \leq \tau$ or $(f_i - f_j) > \tau$, which cover many interpretable relationships (e.g., efficiency ratios or deviations). Thresholds τ are quantized to a predefined resolution. This keeps the search space finite and the resulting expressions human-readable.

RQ3: Interpretability–Quality Trade-off. Using RQ2's languages ($\mathcal{L}_{\text{conj}}, \mathcal{L}_{\text{poly}}, \mathcal{L}_{\text{tree}}, \mathcal{L}_{\text{symb}}$), we formalize a trade-off between:

- $q(d)$: **quality** of a description d (e.g., residual exceptionalness / effect size on residuals);
- $I(d)$: **interpretability** of d, modeled as a decreasing function of its complexity.

Interpretability Model. Let complexity(d) aggregate simple, auditable counts:

$$\begin{aligned}\text{complexity}(d) = w_1 &\cdot \#\text{predicates}(d)\\ + w_2 &\cdot \#\text{operators}(d)\\ + w_3 &\cdot \text{depth}(d)\\ + w_4 &\cdot \text{precision}(d)\end{aligned}$$

where precision(d) penalizes overly precise thresholds (e.g., many decimal places). We instantiate $I(d)$ as

$$I(d) = \frac{1}{1 + \text{complexity}(d)} \quad \text{or} \quad I(d) = \exp(-\beta \cdot \text{complexity}(d)),$$

with $\beta > 0$ controlling how sharply interpretability decays with complexity [17].

Complexity Components. We define each component so that interpretability scoring is auditable and intuitive:

- **#predicates(d): number of atomic conditions:** Counts the number of simple boolean tests (e.g., $f > \tau$, $f = v$). Example: $(\text{RM} > 6.5) \wedge (\text{LSTAT} < 10) \Rightarrow$ #predicates = 2, representing homes with many rooms in affluent areas.
- **#operators(d)**: arithmetic/logical operators used: Includes $\{+, -, \times, \div, \wedge, \vee, \leq, >, =\}$. Example: $(\text{RM} > 7) \wedge (\text{TAX} < 300) \Rightarrow$ #operators = 3 ($>, <, \wedge$).
- **depth(d)**: maximum path length from the root of a decision tree to any leaf, measuring the deepest reasoning chain required. Example: LSTAT $< 10 \rightarrow$ RM $> 6 \rightarrow$ Residual $> 2 \Rightarrow$ depth = 2, indicating a simple two-step rule that identifies underpredicted wealthy areas.
- **#splits(d)**: total number of internal decision nodes across the entire tree, reflecting global structural complexity. Example: a decision tree with three internal splits on RM, LSTAT, and TAX has #splits = 3.
- **precision(d): threshold resolution penalty**: Rather than normalizing all numeric attributes, we penalize thresholds that are overly precise relative to the natural measurement scale of each feature. Let each numeric attribute f_j have observed range $R_j = \max(f_j) - \min(f_j)$ and measurement resolution Δ_j (the smallest meaningful increment in the data). For every numeric condition in a description d, we define

$$\text{precision}(d) = \sum_{j \in \text{num}(d)} \log_{10}\left(\frac{R_j}{\Delta_j}\right).$$

  This term grows when thresholds use unrealistically fine granularity compared to how the variable is measured. For example, if RM ranges from 3 to 9 (so $R = 6$) and is recorded to the nearest 0.1, a threshold such as RM > 6.2 incurs a small penalty, while RM > 6.23 incurs a larger one because it uses unnecessarily fine precision.
- **#terms(d): number of monomials in a polynomial:** Example: $a_0 + a_1 f_1 + a_2 f_2 + a_{12} f_1 f_2 \Rightarrow$ #terms = 4. This measures how many interaction components are combined in the polynomial rule.
- **degree(d): polynomial degree:** Example: $0.6 \cdot \text{RM}^2 - 0.4 \cdot \text{LSTAT} \Rightarrow$ degree = 2, capturing non-linear housing relationships.
- **#tokens(d): symbolic token count:** Total of operands (variables, constants) and operators. This metric quantifies the symbolic compactness of an expression. Example: $\frac{\text{RM}}{\text{LSTAT}} > 1.5 \Rightarrow$ #tokens = 5 (RM, LSTAT, 1.5, /, >). [27]

We then compute:

$$\begin{aligned}\text{complexity}(d) = w_1 &\cdot \#\text{predicates}(d)\\ + w_2 &\cdot \#\text{operators}(d)\\ + w_3 &\cdot \text{depth}(d)\\ + w_4 &\cdot \text{precision}(d)\end{aligned}$$

and derive interpretability:

$$I(d) = \frac{1}{1 + \text{complexity}(d)}.$$

Language-specific Operationalization.

$$\begin{aligned}\mathcal{L}_{\text{conj}} &: \text{complexity}(d) = \#\text{predicates}(d).\\ \mathcal{L}_{\text{tree}} &: \text{complexity}(d) = \text{depth}(d) + \#\text{splits}(d).\\ \mathcal{L}_{\text{poly}} &: \text{complexity}(d) = \#\text{terms}(d) + \text{degree}(d).\\ \mathcal{L}_{\text{symb}} &: \text{complexity}(d) = \#\text{tokens}(d) \text{ (operands + operators)}.\end{aligned}$$

DATASET METADATA TABLE
(FOR 151 CONTEXT)

Trade-off Objective. We search for descriptions that balance quality and interpretability:

$$d^* = \arg\max_{d \in \mathcal{L}} \lambda \cdot q(d) + (1 - \lambda) \cdot I(d), \quad \lambda \in [0, 1].$$

The trade-off parameter $\lambda \in [0, 1]$ is user-controlled: $\lambda \rightarrow 1$ emphasizes subgroup quality, while $\lambda \rightarrow 0$ prioritizes interpretability.

Constraints (i.e., Readability Guards).

- #predicates$(d) \leq 5$ (concision);
- depth$(d) \leq 3$ for trees; polynomial degree$(d) \leq 2$;
- thresholds quantized to sensible units (e.g., \$1000 increments for MEDV, 0.1 rooms for RM), to ensure interpretability and prevent overfitting to measurement noise.

At each iteration, candidate subgroups are scored by $q(d)$ and $I(d)$, and Pareto-optimal solutions are retained to visualize the interpretability–quality frontier. FIGURE(S)?

5.2 Exceptional Subgroups (Under-performance: $\mathcal{L}_{\text{conj}}$ Baseline)

The performance baseline (see Table 1) is established by analyzing the top Pareto-optimal rules defining systematic under-performance (i.e., high positive residuals) using the Conjunctive language.

The $\mathcal{L}_{\text{conj}}$ baseline effectively detects highly exceptional failure regions (RQ1), with interpretable rules found in low-dimensional datasets like `Boston Housing` and complex ones like `Forest Fires` (i.e., late weekend fires during dry season).

5.3 Polynomial Subgroups ($\mathcal{L}_{\text{poly}}$ Analysis)

The Polynomial language, constrained to a maximum degree of two, aimed to introduce non-linear interaction terms (see Table 2).

The $\mathcal{L}_{\text{poly}}$ rule for `Forest Fires` shows a simple conjunctive structure being preferred by the algorithm, while the `Year Prediction MSD` rule demonstrates the language's core advantage by capturing a non-linear interaction term: timbre_avg_3 × timbre_cov_3_3.

5.4 Decision Tree Subgroups ($\mathcal{L}_{\text{tree}}$ Analysis)

The Decision Tree language captured sequential and hierarchical relationships (see Table 3).

The $\mathcal{L}_{\text{tree}}$ rule for `Forest Fires` is highly specific to a narrow temperature band, showing the tree's ability to partition data based on thresholds, achieving a q_{residual} competitive with $\mathcal{L}_{\text{conj}}$.

5.5 Symbolic Expressions Subgroups ($\mathcal{L}_{\text{symb}}$ Analysis)

$\mathcal{L}_{\text{symb}}$ provided the most compact and highly effective rules, particularly on the complex Year Prediction MSD dataset (see Table 4).

The $\mathcal{L}_{\text{symb}}$ rule for `Forest Fires` is the simplest possible, using a single feature predicate (temp > 24.6), demonstrating its preference for maximum parsimony.

5.6 Exceptional Subgroups (Over-performance: RQ1 Extension)

Analysis of over-performing rules (i.e., exceptionally small residuals; see Table 5) completes the diagnosis (RQ1).

The consistent discovery of low-complexity rules with significant negative residual deviation demonstrates that the methodology successfully identifies regions where the model is highly accurate, often corresponding to the most common or easily characterized segments of the input space. This completes the RQ1 objective.

5.7 Trade-off Analysis (RQ2 and RQ3) with Comparative Summary and Discussion

The comparison across all four languages provides a comprehensive answer to RQ2 and RQ3 (see Table 6).

Discussion on RQ2 (Balancing Performance and Interpretability). Hypothesis H2, which suggested that shallow trees or low-degree polynomials offer the best balance, is refuted by the dominance of Symbolic Expressions ($\mathcal{L}_{\text{symb}}$) in terms of both peak performance and complexity. $\mathcal{L}_{\text{symb}}$'s ability to achieve extremely high $q_{residual}$ scores with minimal complexity (e.g., $q_{residual} = 330.52$ with Complexity 3.0 on Year Prediction MSD) shows that a highly constrained search over fundamental arithmetic expressions yields superior results for diagnostic insights.

Discussion on RQ3 (Influence of Expressiveness). The efficacy of each language is domain-dependent. $\mathcal{L}_{\text{conj}}$ remains the most robust generalist, succeeding where model errors align with orthogonal features, achieving the highest peak quality on three out of five datasets. $\mathcal{L}_{\text{symb}}$ provides the greatest overall gain in insight and performance, achieving the absolute highest q_{residual} score and capturing succinct non-linear functional relationships (e.g., ratios, sums, differences). $\mathcal{L}_{\text{poly}}$ and $\mathcal{L}_{\text{tree}}$ introduce structural complexity that often limits the search's ability to find the absolute exceptionality peak, making them less competitive for this specific residual-based EMM task under tight complexity constraints.

6 Conclusions

NEED TO REITERATE, ZOOM OUT, SKETCH CONTEXT AGAIN

This work ~~successfully~~ framed the diagnosis of black-box regression models as an Exceptional Model Mining task targeting model residuals (RQ1). By systematically exploring four distinct description languages (RQ2), quantitative evidence for the complex trade-off between expressiveness and interpretability was demonstrated (RQ3). Furthermore, a complete diagnostic tool for model generalization was provided by analyzing both systematic under-performance (i.e., failures) and over-performance (i.e., overfitting/easy cases).

The current findings indicate that the most effective language for balancing exceptional subgroup quality (q_{residual}) and human interpretability ($I(d)$) is the Symbolic Expressions language ($\mathcal{L}_{\text{symb}}$). This rejects the general idea that moderately rich languages like decision trees or polynomials offer the best sweet spot (i.e., hypothesis H2). $\mathcal{L}_{\text{symb}}$ demonstrated an unprecedented ability to capture peak exceptionality with minimal complexity, leveraging simple arithmetic relationships that are often more insightful than complex conjunctive rules. While the Conjunctive language ($\mathcal{L}_{\text{conj}}$) proved highly competitive on lower-dimensional data, $\mathcal{L}_{\text{symb}}$ proved superior in finding both simple and complex, yet concise, patterns. LINK BACK TO TABLES/ SECTIONS

For future work, an interesting direction would be to validate these findings on real-world industrial streaming data (e.g., wind turbine telemetry), and investigate adaptive complexity constraints

EXTRA PARAGRAPH AT THE START

MAKE DISTINCTIVE CAPTIONS

YOU WANT THE LANGUAGE IN THE TABLE CAPTION

Table 1: Example Subgroup Descriptions and Statistics

Dataset	Subgroup Description (d)	$q_{\text{residual}}(S)$	Complexity	Size (|S|)	Avg. Resid.
Boston Housing	PTRATIO > 19.7 ∧ LSTAT ≤ 11.45 ∧ DIS ≤ 2.1	14.82	5.0	8	24.47
auto-mpg	weight > 2155 ∧ model > 79 ∧ cylinders > 4	7.70	4.0	14	6.99
CMC	Contraceptive_Method ≤ 1 ∧ Wife_Education > 3 ∧ Wife_religion > 0	10.56	5.0	132	37.72
Forest Fires	day > 6 ∧ month > 8 ∧ ISI > 8.56	8.81	4.0	6	24,018.55
Year Prediction MSD	timbre_avg_6 > −8.1 ∧ timbre_avg_3 > 16.18 ∧ timbre_cov_3_3 > 19.42	119.57	5.0	65,956	24.63

Table 2: Example Subgroup Descriptions and Statistics

Dataset	Subgroup Description (d)	$q_{\text{residual}}(S)$	Complexity	Size (|S|)	Avg. Resid.
Boston Housing	CHAS > 0 ∧ TAX > 403	8.30	3.0	8	14.34
auto-mpg	model > 80 ∧ cylinders > 4	5.62	3.0	11	5.95
CMC	Contraceptive_Method ≤ 1 ∧ Wife_Education > 3	8.75	4.0	175	30.73
Forest Fires	day > 6 ∧ temp > 24.1	5.25	3.0	16	9,131.29
Year Prediction MSD	timbre_avg_6 > 0.04687 ∧ timbre_avg_3 × timbre_cov_3_3 > 988.5	112.99	4.0	41,227	27.18

Table 3: Example Subgroup Descriptions and Statistics

Dataset	Subgroup Description (d)	$q_{\text{residual}}(S)$	Complexity	Size (|S|)	Avg. Resid.
Boston Housing	DIS ≤ 1.34	8.05	3.0	10	12.63
auto-mpg	model > 79.5 ∧ displacement > 212.5	7.41	3.0	6	9.77
CMC	Husband_Occupation > 1.5 ∧ Contraceptive_Method > 1.5 ∧ ...	3.19	8.0	181	6.33
Forest Fires	temp > 25.05 ∧ temp ≤ 25.45	8.55	4.0	6	23,331.97
Year Prediction MSD	timbre_avg_6 > −8.76 ∧ timbre_avg_3 > 26.79 ∧ ...	79.95	7.0	31,870	24.13

Table 4: Example Subgroup Descriptions and Statistics

Dataset	Subgroup Description (d)	$q_{\text{residual}}(S)$	Complexity	Size (|S|)	Avg. Resid.
Boston Housing	CHAS + RAD > 24	8.30	3.0	8	14.34
auto-mpg	model > 79	5.42	2.0	89	2.72
CMC	target − Children > 38	11.66	5.0	197	35.32
Forest Fires	temp > 24.6	3.84	2.0	71	3,537.61
Year Prediction MSD	target ≤ 1988	330.52	3.0	75,562	45.31

ARE THESE ALWAYS THE TOP-1 SUBGROUPS FOUND, OR ARE THESE HANDPICKED?

FULL TOP-Q SUBGROUP LISTS ON GITHUB?

INTEGER? IF SO, DROP .0

ALIGN RIGHT IF BETTER

THE REVIEWER WILL WANT TO KNOW WHAT IS IN THESE DOTS (OR WHERE TO FIND THAT INFORMATION)

Table 5: Example Subgroup Descriptions, Metrics, and Interpretations

Dataset	Language	Subgroup Description (d)	$q_{\text{residual}}(S)$	Avg. Resid.	Interpretation
Boston Housing	$\mathcal{L}_{\text{tree}}$	DIS > 1.34 ∧ RM ≤ 8.28 ∧ ...	2.94	0.79	Residential area with moderate rooms, far from employment.
auto-mpg	$\mathcal{L}_{\text{tree}}$	model ≤ 79.5 ∧ displacement > 90.5 ∧ ...	2.43	0.18	Older, mid-to-heavy cars. Model confidently predicts low mileage.
CMC	$\mathcal{L}_{\text{tree}}$	Husband_Occupation > 1.5 ∧ Contraceptive_Method > 1.5 ∧ ...	3.34	3.66	Married women with high-mid occupation/method use.
Forest Fires	$\mathcal{L}_{\text{tree}}$	temp ≤ 25.05 ∧ DMC ≤ 103.55 ∧ ...	1.16	20.08	Fires occurring under specific low temperature/dry fuel conditions.
Year Prediction MSD	$\mathcal{L}_{\text{tree}}$	timbre_avg_6 ≤ −8.76 ∧ timbre_avg_1 > 42.61 ∧ ...	62.44	4.10	Highly constrained, specific area of the timbre space.

Table 6: Comparison of Best q_{residual}, Interpretability, and Structural Insight Across Languages

Dataset	Best q_{residual} (Language)	Highest Interpretability (Language)	Best Structural Insight
Boston Housing	14.82 ($\mathcal{L}_{\text{conj}}$)	0.33 ($\mathcal{L}_{\text{symb}}$ / $\mathcal{L}_{\text{tree}}$)	$\mathcal{L}_{\text{conj}}$ (most specific hyperbox)
auto-mpg	7.70 ($\mathcal{L}_{\text{conj}}$)	0.33 ($\mathcal{L}_{\text{symb}}$)	$\mathcal{L}_{\text{conj}}$ (multivariate linear boundary)
CMC	11.66 ($\mathcal{L}_{\text{symb}}$)	0.20 ($\mathcal{L}_{\text{symb}}$)	$\mathcal{L}_{\text{symb}}$ (difference arithmetic)
Forest Fires	8.81 ($\mathcal{L}_{\text{conj}}$)	0.33 ($\mathcal{L}_{\text{symb}}$ / $\mathcal{L}_{\text{poly}}$)	$\mathcal{L}_{\text{conj}}$ (time-dependent conjunction)
Year Prediction MSD	330.52 ($\mathcal{L}_{\text{symb}}$)	0.25 ($\mathcal{L}_{\text{symb}}$)	$\mathcal{L}_{\text{symb}}$ (simple target threshold)

"BEST" OFTEN COMES WITH ASSUMPTION OF STATISTICAL SIGNIFICANCE.

BE CAUTIOUS WHILE WRITING ABOUT THIS.

that dynamically adjust L based on the inherent sparsity of the underlying residual error surface.

References

[1] Guilherme Seidyo Imai Aldeia and Fabricio Olivetti de Franca. 2024. Interpretability in Symbolic Regression: a benchmark of Explanatory Methods using the Feynman data set. arXiv:2404.05908 [cs.LG]

[2] Jakob Bach. 2024. Using Constraints to Discover Sparse and Alternative Subgroup Descriptions. *arXiv* (2024). doi:10.48550/arxiv.2406.01411

[3] Pablo Barceló, Mikaël Monet, Jorge Pérez, and Bernardo Subercaseaux. 2020. Model Interpretability through the Lens of Computational Complexity. arXiv:2010.12265 [cs.AI]

[4] Nikki M. Branderhorst. 2021. *Goodness-of-Fit in Exceptional Regression Model Mining*. Master's thesis. Eindhoven University of Technology. Available at https://research.tue.nl/en/studentTheses/goodness-of-fit-in-exceptional-regression-model-mining.

[5] Aleksey Buzmakov. 2020. Towards Polynomial Subgroup Discovery by means of FCA. In *CEUR Workshop Proceedings*, Vol. 2729. 1–12.

[6] F. O. de Franca, M. Virgolin, M. Kommenda, et al. 2023. Interpretable Symbolic Regression for Data Science: Analysis of the 2022 Competition. arXiv:2304.01117 [cs.LG]

[7] Xin Du, Sikun Yang, Wouter Duivesteijn, and Mykola Pechenizkiy. 2025. *Conformalized Exceptional Model Mining: Telling Where Your Model Performs (Not) Well.* Springer Nature Switzerland, Cham, 528–544. doi:10.1007/978-3-032-06066-2_31

[8] Wouter Duivesteijn, Ad Feelders, and Arno Knobbe. 2012. Different slopes for different folks: mining for exceptional regression models with cook's distance. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '12)*. Association for Computing Machinery, 868–876. doi:10.1145/2339530.2339668

[9] Wouter Duivesteijn, Ad J. Feelders, and Arno Knobee. 2016. Exceptional Model Mining. *Data Mining and Knowledge Discovery* 30, 1 (2016), 47–98. doi:10.1007/s10618-015-0403-4

[10] Wouter Duivesteijn and Julia Thaele. 2014. Understanding Where Your Classifier Does (Not) Work - The SCaPE Model Class for EMM. *Proceedings - IEEE International Conference on Data Mining, ICDM* (01 2014), 809–814. doi:10.1109/ICDM.2014.10

[11] Jack H. Good, Torin Kovach, Kyle Miller, and Artur Dubrawski. 2023. Feature Learning for Interpretable, Performant Decision Trees. In *Advances in Neural Information Processing Systems*, Vol. 36. 1–11.

[12] Andreas Henelius, Kai Puolamäki, Henrik Boström, Lars Asker, and Panagiotis Papapetrou. 2014. A peek into the black box: exploring classifiers by randomization. *Data Mining and Knowledge Discovery* 28, 5–6 (Sept. 2014), 1503–1529. doi:10.1007/s10618-014-0368-8

[13] Andreas Henelius, Kai Puolamäki, Isak Karlsson, Jing Zhao, Lars Asker, Henrik Boström, and Panagiotis Papapetrou. 2015. GoldenEye++: a Closer Look into the Black Box. In *Statistical Learning and Data Sciences. SLDS 2015. Lecture Notes in Computer Science*, Vol. 9047. 96–105. doi:10.1007/978-3-319-17091-6_5

[14] Francisco Herrera, Cristóbal José Carmona, Pedro González, and María José Del Jesus. 2011. An overview on subgroup discovery: foundations and applications. *Knowledge and Information Systems* 29, 3 (2011), 495—-525. doi:10.1007/s10115-010-0356-2

[15] Kimia Kamal and Bilal Farooq. 2022. Ordinal-ResLogit: Interpretable Deep Residual Neural Networks for Ordered Choices. arXiv:2204.09187 [cs.LG]

[16] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. An Empirical Study of Rich Subgroup Fairness for Machine Learning. arXiv:1808.08166 [cs.LG] https://arxiv.org/abs/1808.08166

[17] John P. Lalor and Hong Guo. 2022. Measuring algorithmic interpretability: A human-learning-based framework and the corresponding cognitive complexity score. arXiv:2205.10207 [cs.AI]

[18] Dennis Leman, Ad Feelders, and Arno Knobbe. 2008. Exceptional Model Mining. In *Machine Learning and Knowledge Discovery in Databases*, Vol. 24. Springer Berlin Heidelberg, 1–16. doi:10.1007/978-3-540-87481-2_1

[19] Lu Liu. [n. d.]. Causal Learning for Heterogeneous Subgroups Based on Nonlinear Causal Kernel Clustering. https://arxiv.org/html/2501.11622v1#abstract. 2025.

[20] Michael Mampaey, Siegfried Nijssen, Ad Feelders, and Arno Knobbe. 2012. Efficient Algorithms for Finding Richer Subgroup Descriptions in Numeric and Nominal Data. In *Proceedings of the 2012 IEEE 12th International Conference on Data Mining (ICDM '12)*. IEEE Computer Society, USA, 499–508. doi:10.1109/ICDM.2012.117

[21] Marvin Meeng and Arno Knobbe. 2021. For real: a thorough look at numeric attributes in subgroup discovery. *Data Mining and Knowledge Discovery* 35 (01