Meaningful Intermediate Variables for Model-Driven Explainable AI and Qualitative Reasoning with Grammatical Evolution

Dominik Sepioło, **Antoni Ligęza**, **Mikołaj Jarosławski**
AGH University of Krakow
{sepiolo, ligeza, mikjar}@agh.edu.pl

Abstract

Qualitative Reasoning requires representations that expose relevant variables, relations, dependencies, and state distinctions. This paper studies Meaningful Intermediate Variables (MIVs) as candidate relational abstractions between raw numerical variables and symbolic model structure. The proposed framework generates candidate MIVs by a constrained pairwise construction scheme, ranks them using validation-based criteria, and injects selected candidates into a second Grammatical Evolution search. In a controlled synthetic task, the method recovers product, difference, and protected-division components within the top-ranked candidates. The resulting model is more compact and achieves lower prediction error than a baseline search over raw variables, providing a proof of concept for hybrid qualitative–quantitative representation construction.

1 Introduction

Qualitative Reasoning (QR) is concerned with representations that make system structure explicit, including relevant variables, relations, dependencies, mechanisms, and state distinctions [Forbus, 1984; Kuipers, 1994]. In scientific, engineering, medical, and decision-support domains, a model is often required to provide accurate predictions while exposing the structure by which variables are related. This requirement is also central to Model-Driven Explainable Artificial Intelligence (MD-XAI), where explanation is approached as structural model construction rather than as a post-hoc description of a trained black-box model [Sepioło and Ligęza, 2024].

A recurring difficulty is the transition from raw numerical variables to a representation that exposes intermediate structure. Numerical models may fit the data while hiding functional dependencies, ratios, thresholds, or other relation types. This paper addresses this intermediate level by studying MIVs as symbolic components generated and evaluated within a grammar-constrained modeling process.

Post-hoc XAI methods such as LIME and SHAP provide local surrogate explanations or feature attributions [Ribeiro *et al.*, 2016; Lundberg and Lee, 2017]. They are useful for model inspection, but they are not designed to construct reusable symbolic intermediate variables or qualitative model structures. Earlier experiments comparing shallow XAI methods with GE also indicated that symbolic model construction may provide deeper structural explanations than local attribution-based methods [Sepioło and Ligęza, 2023].

From a QR perspective, MIVs are primarily representational constructs. Although the candidates considered in this paper are numerical expressions, they expose relation types that are often central to qualitative modeling: contrast, interaction, proportionality, and state-relevant functional dependence. Such variables can be inspected as symbolic components and may later be mapped to qualitative distinctions by thresholds, intervals, landmarks, or domain-specific regimes. The paper contributes a constrained candidate-generation scheme for MIVs and evaluates whether the generated candidates can recover known intermediate relations and support more compact downstream symbolic models.

2 Background and Method

Symbolic Regression searches for analytical expressions that fit observed data without assuming a fixed parametric form. It is useful when compact and interpretable mathematical relations are sought, rather than only predictive accuracy [Makke and Chawla, 2024]. Grammatical Evolution (GE) is a grammar-based evolutionary method in which a grammar defines the admissible expressions and evolutionary search explores the resulting symbolic model space [Ryan *et al.*, 1998]. The `gramEvol` package implements GE in R by optimizing populations of expressions generated from user-defined grammars [Noorian *et al.*, 2016].

A key element of MD-XAI is the concept of a Meaningful Intermediate Variable [Sepioło and Ligęza, 2024]. An MIV is an auxiliary variable constructed from input variables or other intermediate variables for knowledge aggregation. Examples include the discriminant of a quadratic equation, area-like variables in classification tasks, or the Body-Mass Index in medicine [Sepioło and Ligęza, 2024]. Here, an MIV is distinguished from a *candidate MIV*: the latter is only a proposed intermediate structure generated by symbolic search or by a symbolic construction scheme. Its semantic or causal relevance is not automatically guaranteed.

In QR, useful representations often abstract from raw measurements to variables that encode relations, regimes, or state-relevant quantities [Forbus, 1984; Kuipers, 1994]. MIVs serve a related role: they are computed from numerical data, but they can expose structural information such as product relations, difference relations, ratios, or threshold-dependent quantities. The present paper evaluates the first step: automatic generation and reuse of symbolic intermediate variables before explicit discretization into qualitative states.

The framework has two layers: symbolic search and candidate MIV construction. GE is used both as a baseline symbolic regression method and as a downstream search procedure after candidate MIVs have been added to the variable set. Let

$$\mathcal{X} = \{X_1, \ldots, X_p\}$$

be the set of raw input variables. A candidate-generation scheme defines a finite set

$$\mathcal{C} = \{c_1, \ldots, c_q\},$$

where each c_i is a symbolic expression over a subset of $\mathcal{X}$.

In the pairwise setting used in the experiment, for each pair (X_i, X_j), $i < j$, the scheme generates:

$$X_i + X_j, \quad X_i - X_j, \quad X_j - X_i, \quad X_i X_j,$$
$$\mathrm{pdiv}(X_i, X_j), \quad \mathrm{pdiv}(X_j, X_i),$$

Here, pdiv denotes protected division. The construction scheme contains additive, subtractive, multiplicative, and ratio-type candidate MIVs, but avoids arbitrary expression growth.

Each candidate expression $c \in \mathcal{C}$ is evaluated on a validation set to identify candidates that are simple, predictive, and useful beyond the raw variables. The ranking score combines validation correlation, one-candidate predictive usefulness, incremental usefulness over the raw variables, a small operator-type bonus, and a complexity penalty:

$$\mathrm{score}(c) = \alpha_1 |\operatorname{corr}(c, Y)| + \alpha_2 U(c) + \alpha_3 I(c) + B(c) - C(c).$$

Here, $U(c)$ denotes the validation RMSE reduction obtained by a linear model using only c, relative to a constant baseline. The term $I(c)$ denotes the additional validation RMSE reduction obtained when c is added to a linear model over the raw variables. In the experiment, the score was instantiated as

$$\mathrm{score}(c) = 12 |\operatorname{corr}(c, Y)| + 10 U(c) + 25 I(c) + B(c) - C(c).$$

The selected candidates are added to the original raw inputs as additional variables:

$$\mathcal{X}' = \mathcal{X} \cup \{m_1, \ldots, m_k\}.$$

A second GE run is then performed over $\mathcal{X}'$.

3 A Proof-of-Concept Experiment

The experiment evaluates whether a constrained pairwise MIV candidate-generation scheme can recover known intermediate variables and whether injecting the selected candidates into a second GE phase improves the resulting symbolic model. The experiment is intentionally synthetic because the ground-truth MIVs are known and can therefore be evaluated directly.

3.1 Synthetic Task

The data-generating process contains three latent intermediate variables:

$$z_1 = X_1 X_2,$$
$$z_2 = \mathrm{pdiv}(X_3, X_2),$$
$$z_3 = X_1 - X_3.$$

The target variable is defined as

$$Y = z_1 + z_3 - z_2 + \varepsilon,$$

where ε is Gaussian noise. The variable X_2 is sampled from a positive interval to keep the denominator in z_2 away from zero. The remaining variables are sampled independently from bounded uniform distributions. The variable X_4 is included as an irrelevant input to expose spurious pairwise candidates involving unused variables.

The synthetic task tests the recovery of heterogeneous intermediate structures: a product, a difference, and a protected-division ratio. All three ground-truth intermediate variables are expressible by the pairwise candidate-generation scheme. The ratio component makes it possible to evaluate whether protected division can be recovered as a meaningful candidate MIV rather than appearing only as an artifact of symbolic search.

Two variants are compared. Baseline GE searches over the original variables X_1, X_2, X_3, X_4. Pairwise-MIV-GE first ranks candidate intermediate variables generated by the scheme described above and then runs a second GE phase over the original variables and the injected candidates.

3.2 Implementation and Metrics

For four raw variables, the pairwise scheme generated 36 candidate intermediate variables: six candidates for each unordered pair of variables. Candidate generation was deterministic and used only the raw input variables. It did not use test-set information or access to the ground-truth intermediate variables.

The data were split into training, validation, and test sets. Baseline GE was trained on the training set. Candidate intermediate variables were ranked on the validation set using the scoring function from the previous section. The final predictive performance of Baseline GE and Pairwise-MIV-GE was reported on the held-out test set. Each split contained 300 training, 300 validation, and 300 test samples. The experiment was repeated over 10 random seeds. The noise standard deviation was set to 0.03. The top five ranked candidates were injected into the second GE phase. GE was run for 140 generations in both the baseline and MIV-augmented phases, with a population size of 180.

Candidate recovery is evaluated using the known ground-truth MIVs:

$$z_1 = X_1 X_2, \quad z_2 = \mathrm{pdiv}(X_3, X_2), \quad z_3 = X_1 - X_3.$$

For each ground-truth MIV, its rank in the candidate list is reported. A candidate is considered to match a ground-truth MIV if it is numerically equivalent or affine-equivalent to the target component in the training data. Affine equivalence accounts for expressions such as $X_3 - X_1$, which is equivalent to $X_1 - X_3$ up to sign.

Table 1: Candidate recovery and operator composition in top-k candidates.

Top-k	Rec.	pdiv	Prod.	Diff.
1	1.0	0.0	0.0	1.0
3	2.0	0.0	1.0	2.0
5	3.0	2.0	1.0	2.0
10	3.0	2.0	2.0	4.9
20	3.0	3.5	2.5	10.0

Table 2: Predictive and structural results.

Method	Train	Test	Gap	Nodes
Baseline	0.111	0.119	0.007	15.0
Pairwise-MIV	0.029	0.030	0.000	5.0

Downstream symbolic modeling is evaluated using train RMSE, test RMSE, the train–test generalization gap, and expression node count. The generalization gap is defined as

$$\mathrm{gap} = \mathrm{RMSE}_{test} - \mathrm{RMSE}_{train} .$$

In the MIV-augmented setting, injected candidate MIVs are counted as single terminal nodes in the final expression. Therefore, a lower node count indicates a more compact final symbolic expression, but it does not account for the internal symbolic cost of computing the injected intermediate variables themselves.

3.3 Results

The candidate-ranking results show stable recovery of all three ground-truth MIVs. The difference component $z_3 = X_1 - X_3$ was ranked first in every run. The product component $z_1 = X_1 X_2$ had a mean rank 3.0, and the protected-division component $z_2 = \mathrm{pdiv}(X_3, X_2)$ had a mean rank 4.8. No ground-truth component was missing from the ranked candidate lists.

Table 1 confirms that all three ground-truth MIVs were recovered within the top-5 candidates. In the top-1 position, one component was recovered on average. In the top-3 candidates, two components were recovered. In the top-5, top-10, and top-20 candidates, all three components were recovered.

Table 2 shows that, with five injected candidates, the downstream modeling results align with the candidate-recovery results. Pairwise-MIV-GE achieved substantially lower test error than Baseline GE while also producing smaller final expressions. Baseline GE achieved mean test RMSE 0.119, whereas Pairwise-MIV-GE achieved mean test RMSE 0.030. The mean node count decreased from 15.0 to 5.0. The mean generalization gap also decreased from 0.007 to 0.0004.

The selected candidates explain this change. The top-5 injected set always contained the difference component, the product component, and the true protected-division component. Thus, the second GE phase had access to all three ground-truth MIVs and could construct compact expressions algebraically equivalent to $z_1 + z_3 - z_2$. Redundant selected candidates, including the sign-reversed difference, were generally not needed in the final compact model.

4 Discussion

From a QR perspective, the recovered MIVs can be interpreted as candidate structural variables. They correspond to distinct relation types: multiplicative interaction, contrast, and ratio-like dependence. They are computed numerically, but their role is representational: they expose intermediate relations that are not explicit in the raw input space. Such variables could later be mapped to qualitative levels by thresholding, interval abstraction, landmarks, or domain-specific discretization.

The experiment separates two effects. First, the pairwise candidate-generation scheme recovered the relevant intermediate variables, including the ratio-type component $\mathrm{pdiv}(X_3, X_2)$, within the top-5 candidates in all runs. Second, once these candidates were injected, the downstream GE phase could construct compact expressions equivalent to the ground-truth structure. For MD-XAI and QR-oriented modeling, this supports evaluating candidate recovery and final symbolic modeling as related but distinct targets.

The study is a proof of concept, not a complete QR system. Since the ground-truth components are expressible by the pairwise candidate-generation scheme, the experiment does not test open-ended qualitative model discovery. The predictive comparison is also not a definitive benchmark against all symbolic regression baselines, because Pairwise-MIV-GE benefits from an additional representational step before the second GE phase.

The present experiment is limited by the simplicity and redundancy of the pairwise construction scheme. Pairwise candidates can recover only two-variable intermediate structures, and the selected top-5 set may contain redundant candidates such as $X_1 - X_3$ and $X_3 - X_1$. Richer QR-oriented representations would require diversity-aware candidate selection, compositional candidate generation, monotonicity constraints, threshold or relay-like operators, and explicit discretization of selected MIVs into qualitative states. Grammar refinement, including feedback-driven adaptation of production rules, is one possible route toward such richer representations [Sepioło *et al.*, 2025].

5 Conclusion

This paper examined Meaningful Intermediate Variable generation as a mechanism for constructing intermediate representations in hybrid qualitative–quantitative reasoning. The proof-of-concept experiment showed that a constrained pairwise candidate-generation scheme recovered all ground-truth intermediate variables within the top-5 ranked candidates and enabled Pairwise-MIV-GE to obtain lower prediction error and a more compact final expression than Baseline GE. The main contribution is the construction of an explicit intermediate relational layer between numerical inputs and symbolic model structure. This layer can serve as a basis for later qualitative interpretation through thresholds, regimes, or domain-specific abstraction. Future work should examine diversity-aware candidate selection, compositional MIV construction, qualitative discretization, and QR-oriented grammar adaptation.

References

[Forbus, 1984] Kenneth D. Forbus. Qualitative process theory. *Artificial Intelligence*, 24(1-3):85–168, 1984.

[Kuipers, 1994] Benjamin Kuipers. *Qualitative Reasoning: Modeling and Simulation with Incomplete Knowledge*. MIT Press, 1994.

[Lundberg and Lee, 2017] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, pages 4765–4774, 2017.

[Makke and Chawla, 2024] Nour Makke and Sanjay Chawla. Interpretable scientific discovery with symbolic regression: A review. *Artificial Intelligence Review*, 57, 2024.

[Noorian *et al.*, 2016] Farzad Noorian, Anthony M. de Silva, and Philip H. W. Leong. gramevol: Grammatical evolution in r. *Journal of Statistical Software*, 71(1):1–26, 2016.

[Ribeiro *et al.*, 2016] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.

[Ryan *et al.*, 1998] Conor Ryan, JJ Collins, and Michael O. Neill. Grammatical evolution: Evolving programs for an arbitrary language. In Wolfgang Banzhaf, Riccardo Poli, Marc Schoenauer, and Terence C. Fogarty, editors, *Genetic Programming*, pages 83–96, Berlin, Heidelberg, 1998. Springer Berlin Heidelberg.

[Sepioło and Ligęza, 2023] Dominik Sepioło and Antoni Ligęza. A comparison of shallow explainable artificial intelligence methods against grammatical evolution approach. In Adam Wojciechowski and Piotr Lipiński, editors, *Progress in Polish Artificial Intelligence Research 4*, number 2437 in Monografie Politechniki Łódzkiej, pages 89–94. Wydawnictwo Politechniki Łódzkiej, Łódź, 2023.

[Sepioło and Ligęza, 2024] Dominik Sepioło and Antoni Ligęza. Towards Model-Driven Explainable Artificial Intelligence: Function identification with Grammatical Evolution. *Applied Sciences*, 14(13):5950, 2024.

[Sepioło *et al.*, 2025] Dominik Sepioło, Mikołaj Jarosławski, and Antoni Ligęza. Grammar refinement in grammatical evolution using Large Language Models. In *PP-RAI'2025: Progress in Polish Artificial Intelligence Research 6*, pages 22–28. Silesian University Publishers, 2025.