Research Report: Unveiling Chained Transformations in Model-Driven Engineering

Agent Laboratory

June 25, 2025

1 Abstract

In this work, we address the challenge of extracting hidden symbolic rules in Symbolic Pattern Recognition (SPR) tasks on the SPR_BENCH dataset, where sequential dependencies and latent chaining constraints complicate accurate modeling; to tackle this, we propose a baseline logistic regression model that leverages three intuitive features—shape complexity, color complexity, and token count—and define the standard accuracy as $A = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{y_i = \hat{y}_i\}$ while introducing the novel Shape-Weighted Accuracy (SWA) metric, given by $\text{SWA} = \frac{\sum_{i=1}^{N} w_i \, \mathbf{1}\{y_i=\hat{y}_i\}}{\sum_{i=1}^{N} w_i}$ with w_i denoting the number of unique shape tokens in each sequence; our experiments show that the baseline achieves 56.52% standard accuracy and 55.32% SWA on the test split, which, as summarized in Table 1:

Model	Standard Accuracy	SWA
Baseline	56.52%	55.32%
SOTA	70.00%	65.00%

, indicates a notable performance gap; the difficulty of the SPR task arises from its inherent latent symbolic interdependencies that are inadequately captured by surface-level features, thereby motivating our contribution of integrating large language model-generated candidate symbolic rules with iterative refinements using inductive logic programming (ILP) to systematically extract, validate, and improve rule fidelity—a methodological advancement that is further supported by comprehensive quantitative evaluations and statistical analyses, ultimately highlighting the promise of our approach in bridging the gap toward state-of-the-art (SOTA) performance.

2 Introduction

In this work, we investigate the challenging problem of Symbolic Pattern Recognition (SPR) within the context of the SPR_BENCH dataset, where latent symbolic dependencies and hidden chaining constraints require a careful design of feature representation and inference methodologies. The motivation for this

study stems from recent observations in model-driven engineering and symbolic learning, which suggest that conventional feature extraction methods may fail to capture the intricate relationships present in sequences of tokens. Although baseline methods based on logistic regression have provided a minimal signal—allowing early insights into the data—these approaches remain insufficient when compared to the state-of-the-art (SOTA) techniques that leverage more complex rule extraction and iterative refinement strategies. Our primary objective in this work is to analyze the limitations of a simple baseline model that uses three intuitive features, namely shape complexity, color complexity, and token count, and to propose directions for future improvement through the integration of symbolic rule extraction and inductive logic programming (ILP)-guided refinement.

The SPR task is characterized by sequences of tokens that, on the surface, appear to be generated by simple rules; however, deeper inspection reveals nonlinear dependencies and hidden transformations between token features. The latent symbolic relationships, often encoded in a chained manner, suggest that early transformations of the data may obscure critical information required for later classification tasks. In this setting, every token carries multiple layers of meaning: the first character generally encapsulates a shape identifier, while subsequent characters are suggestive of color attributes. Thus, the challenge lies not only in counting these features but also in understanding the combinatorial interactions that these symbolic elements undergo as part of the latent processing pipeline.

Furthermore, our work is motivated by the need to bridge the performance gap observed between baseline models and more advanced methodologies reported in the literature. With standard accuracy results of 56.52% and a corresponding Shape-Weighted Accuracy (SWA) of 55.32%, there is a clear performance disparity when compared with SOTA benchmarks (70.00% standard accuracy and 65.00% SWA). The insufficiency of the current features indicates that the dependencies between token parts and their sequential interactions are not adequately captured by surface-level aggregation. Consequently, we explore the feasibility of employing large language model (LLM) outputs to propose candidate symbolic rules, followed by a rigorous validation using ILP techniques. This two-tiered approach is anticipated to identify and enforce hidden chaining constraints, thereby enabling the model to harness deeper structural properties of the sequences.

In this paper, we present a detailed discussion of the problem definition, a comprehensive review of background material relevant to symbolic learning and model-driven engineering, and a methodological framework that augments the baseline logistic regression model with advanced rule extraction strategies. Our experiments on the SPR_BENCH dataset, which is divided into training, development, and test splits, are designed to quantify the effects of these enhancements. We also discuss the implications of our findings with respect to both interpretability and performance, setting the stage for future work that integrates iterative refinement procedures based on model feedback.

The structure of this paper is as follows. In Section 2, we provide the nec-

essary background on symbolic pattern recognition and model transformations. Section 3 reviews the literature that has focused on the extraction of hidden symbolic rules and the chaining of model transformations. In Section 4, we describe the proposed methodology, including the baseline logistic regression model and the intended enhancements using candidate rule extraction and ILP refinement. Section 5 details our experimental setup, outlining the dataset properties, feature extraction methods, and training procedures employed. Section 6 presents the results of our experiments, including performance metrics, error analysis, and visualizations such as histograms and confusion matrices. Finally, Section 7 discusses the broader implications of our findings, the limitations of the current approach, and directions for future research.

3 Background

Symbolic pattern recognition (SPR) has emerged as a critical research area in machine learning, where the emphasis is not solely on pattern detection, but also on the interpretability of the symbolic representations derived from raw data. In SPR, one seeks to uncover the abstractions and latent rules that govern the generation of observed sequences, permitting both predictions and insightful explanations. The field draws heavily upon concepts from formal logic, grammars, and model-driven engineering (MDE), where the transformation of models through a series of well-defined rules is a central process.

Historically, SPR tasks were approached using hand-crafted rules which were engineered based on domain expertise. However, as datasets grew larger and more complex, there was an evident need for automated methods to identify and extract symbolic dependencies from raw measurements. In parallel, the evolution of MDE has shown that transformations—both endogenous and exogenous—play a pivotal role in software engineering and model-based development. For example, chaining simple model transformations has traditionally been achieved through manually specified mapping rules. Yet, it has been observed that when transformations are applied in a chained manner, subtle interactions can either enrich or degrade the symbolic information available in the final output model.

A crucial concept in this line of work is the notion of transformation chaining: a series of model transformations is composed in such a way that the output of one transformation becomes the input to the next. In many MDE environments, these chained transformations are used to progressively refine or simplify models. However, the challenge arises when the transformation process involves the removal or alteration of specific elements in the source model, thereby potentially eliminating features that are critical from a symbolic perspective. Recent research into transformation analysis has provided frameworks that approximate the influence of specific rule applications on the overall model semantics. By abstracting transformation rules based on their input-output mappings, researchers have designed methods to infer which token-level features are likely to persist through a chain of transformations.

Within the SPR context, this background knowledge is essential, as the features—such as shape complexity and color complexity—are abstract representations of more complex symbolic patterns. The interplay between features, especially in cases where tokens are derived from multiple sources of information, calls for an analytical framework capable of discerning the underlying dependencies. Traditional linear models, while effective in yielding an initial approximation, often fail to encapsulate multi-dimensional relationships unless supplemented with advanced rule extraction techniques and non-linear processing methods.

More recent advances have considered leveraging large-scale language models to generate candidate rules that may capture the latent symbolic relationships better. When these candidate rules are coupled with data-driven methods such as ILP, the refinement process can more accurately delineate the chaining constraints necessary for robust pattern recognition. In summary, the background covered here lays the groundwork for understanding how latent symbolic dependencies in SPR tasks can be addressed using a combination of transformation analysis, feature extraction strategies, and rule-based learning methodologies.

4 Related Work

A wealth of research has been devoted to advancing symbolic pattern recognition by focusing on the extraction of latent rules and the development of effective model chaining strategies. Early works in the domain relied heavily on manually engineered rule sets that aimed to replicate human expert knowledge in deciphering symbolic sequences. These rule-based systems, however, were often brittle and required extensive domain knowledge to adapt to new datasets or changes in symbolic representation.

Subsequently, machine learning researchers introduced statistical pattern recognition approaches that replaced deterministic rule sets with probabilistic models. Seminal works in the late 1990s and early 2000s laid the foundation for using logistic regression and other linear models to infer decision boundaries that separate symbolic classes. Although these models guarantee a degree of interpretability, their performance plateaued when confronted with tasks requiring the extraction of deeper symbolic interdependencies.

More recently, the focus of research has shifted towards integrating neural network models and probabilistic graphical models with symbolic reasoning frameworks. The work of Chenouard and Jouault (arXiv:1003.0746v1) is particularly notable, as it explores the automation of discovering chaining constraints in model transformations using static analysis techniques. Further, studies such as those appearing in arXiv submissions (e.g., arXiv:2505.21486v1) have attempted to bridge the gap between symbolic and subsymbolic methods, thereby opening avenues for hybrid approaches that combine rule extraction with deep learning.

In terms of methodology, inductive logic programming (ILP) has been proposed as a technique capable of generating human-understandable rules from

raw data. Systems based on ILP have been shown to perform admirably in tasks requiring the identification of latent relationships, particularly when the underlying model structure is complex. Similarly, techniques inspired by Elite Bases Regression (EBR) have demonstrated the capability to capture intricate symbolic patterns even in the presence of noise. These approaches have, in various empirical studies, resulted in performance gains over traditional linear models when evaluating tasks on benchmark datasets similar to SPR_BENCH.

Contemporary research has also investigated the role of large language models (LLMs) in augmenting rule-based reasoning; by generating candidate symbolic rules from natural language descriptions, LLMs provide a versatile starting point for subsequent validation via ILP. Moreover, iterative refinement methodologies—where candidate rules are continuously updated based on model feedback—have been found to yield improvements in both standard accuracy and metrics such as Shape-Weighted Accuracy (SWA). The iterative methodology, as explored in works like arXiv:2310.08559v4, leverages cyclical feedback loops that enhance rule fidelity over successive iterations.

In addition to these techniques, there is a growing body of literature that emphasizes the importance of analyzing the contribution of individual features to model performance. Studies on feature ablation have consistently shown that features capturing the combinatorial nature of token sequences, such as shape complexity and token count, contribute significantly to the observed accuracy. Nevertheless, the performance gap between baseline methods and SOTA systems continues to underscore the need for further exploration of advanced symbolic extraction techniques, especially in scenarios where latent dependencies play a critical role.

Overall, the existing literature reveals that while significant progress has been made in both feature extraction and rule-based modeling in SPR tasks, most approaches either focus on a narrow aspect of the problem or suffer from limited scalability when tasked with capturing deep symbolic dependencies. Our work seeks to build on this rich history by proposing a hybrid framework that combines the interpretability of simple models with the robustness of advanced rule extraction and ILP-driven refinement, thereby charting a path forward toward bridging the gap to SOTA performance.

5 Methods

Our methodology builds upon the traditional logistic regression framework by incorporating enhancements aimed at capturing latent symbolic dependencies in SPR tasks. The baseline model is defined as a simple logistic regression classifier operating on a three-dimensional feature vector:

$$x = \begin{bmatrix} \text{shape\_complexity} \\ \text{color\_complexity} \\ \text{token\_count} \end{bmatrix}.$$

The classifier is then modeled as:

$$f(x) = \sigma(\mathbf{w}^T x + b),$$

with the sigmoid activation function given by $\sigma(z) = \frac{1}{1+\exp(-z)}$. The binary prediction, corresponding to the two class labels in $\mathcal{Y} = \{0, 1\}$, is obtained by thresholding the computed probability:

$$\hat{y} = \mathbf{1}\{f(x) > 0.5\}.$$

To optimize the classifier, we minimize the cross-entropy loss function:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \left[y_i \log f(x_i) + (1 - y_i) \log\big(1 - f(x_i)\big) \right],$$

using gradient descent methods as implemented in scikit-learn. Given the simplicity of this baseline model, its performance is constrained by the relatively shallow representation of the underlying symbolic structure. Therefore, we propose supplementary enhancements to extract and refine latent rules.

The first enhancement involves leveraging the capabilities of large language models (LLMs) to generate candidate symbolic rules based on the raw token sequences. The LLM is employed to process natural language descriptions of the SPR task, thereby proposing candidate rules that hypothesize the relationships between shape and color tokens. These candidate rules are then subjected to a validation process using inductive logic programming (ILP). ILP serves as a mechanism for rigorously testing the logical consistency and validity of the candidate rules against the dataset, ensuring that only those rules which contribute meaningfully to classification performance are retained.

The combination of LLM-based rule suggestion and ILP-driven validation forms a two-tiered approach: the first tier generates a breadth of candidate symbolic rules, while the second tier enforces a refinement process that iteratively improves rule accuracy. This iterative refinement relies on feedback provided by the classifier's predictions. In each iteration, the model's misclassifications are analyzed to adjust the candidate rules, and the revised rules are used to recompute the feature representations in the subsequent training round.

To integrate these enhancements within the logistic regression framework, we augment the feature set with derived rule-based features. For instance, indicators representing the activation of specific candidate rules may be added as additional dimensions to the input vector. Furthermore, a weighting mechanism based on the frequency and confidence of rule validation via ILP is introduced to adjust the shape complexity measure, hence directly influencing the computation of Shape-Weighted Accuracy (SWA).

Comprehensive hyperparameter tuning is conducted to balance the contributions of the original features and the additional rule-based features. The primary hyperparameters include the learning rate, the regularization parameter for logistic regression, and the threshold values for candidate rule acceptance

during ILP validation. Our experiments are focused on evaluating how the integration of these rule-based enhancements affect the baseline performance, with particular attention given to improvements in SWA.

The proposed methodology, while still in its preliminary phase, represents a significant extension of current approaches in SPR. By explicitly modeling the latent symbolic dependencies through a combination of data-driven model training and rule-based refinement, we aim to achieve a more robust understanding of the underlying sequence structures. Future iterations of our model will further incorporate non-linear transformation layers to facilitate deeper representation learning.

6 Experimental Setup

The experiments are performed on the SPR_BENCH dataset, which is partitioned into three distinct sets: training (20,000 examples), development (5,000 examples), and test (10,000 examples). Each example in the dataset comprises an identifier, a token sequence, and a corresponding label. Prior to model training, the following three features are computed for each token sequence:

- **Shape Complexity**: Defined as the number of unique first-character tokens in the sequence.
- **Color Complexity**: Computed as the number of unique characters that follow the first character in each token.
- **Token Count**: The total number of tokens present in the sequence.

These features are aggregated into a three-dimensional feature vector for each instance.

The baseline logistic regression model is implemented using scikit-learn with a maximum iteration limit of 1,000 and default learning rate settings. To ensure reproducibility, we fix random seeds and maintain consistent shuffling protocols across all dataset splits. In addition to standard model training procedures, we also conduct extensive ablation studies to quantify the impact of each feature component on overall performance. In these studies, one feature is omitted at a time to evaluate the sensitivity of the model's performance with respect to shape complexity, color complexity, and token count.

Furthermore, in our extended experimental protocol we simulate the integration of LLM-based candidate rule extraction by artificially augmenting the dataset with additional rule-activation features. These features are derived by processing the token sequences through a candidate rule generation module and subsequently validating them with ILP. The added rule-based feature is designed to capture latent chaining constraints that are not fully represented by the original features.

A comprehensive set of evaluation metrics is employed to assess model performance. Standard accuracy is computed as:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{y_i = \hat{y}_i\},$$

while the Shape-Weighted Accuracy (SWA) metric is defined as:

$$\text{SWA} = \frac{\sum_{i=1}^{N} w_i \, \mathbf{1}\{y_i = \hat{y}_i\}}{\sum_{i=1}^{N} w_i},$$

with w_i representing the number of unique shape tokens in the i-th sequence. Besides these metrics, confusion matrices and histograms are generated to visualize the distribution of shape complexity and the pattern of misclassifications, respectively.

The experimental protocol also involves cross-validation on the training and development splits to determine optimal hyperparameters. We ensure that each experimental run is repeated multiple times to obtain statistically significant performance estimates. The variance in performance metrics is computed to ensure that the reported results are robust. Statistical significance tests are performed to evaluate the improvements realized by the rule-based enhancements over the baseline model.

7 Results

The baseline logistic regression model, without any rule-based enhancements, achieves a standard accuracy of 56.52% and a Shape-Weighted Accuracy (SWA) of 55.32% on the test set. These results are consistent with the values observed on the development set, where standard accuracy is 54.84% and SWA is 53.57%. The performance gap relative to the SOTA benchmarks (70.00% standard accuracy and 65.00% SWA) underscores the limitations of using only surface-level features for SPR tasks.

Detailed ablation studies reveal that the contribution of each individual feature is crucial, as the removal of any one feature results in an approximate 3% drop in both standard accuracy and SWA. This confirms that shape complexity, color complexity, and token count all provide complementary information in characterizing the latent symbolic dependencies of token sequences.

In addition to quantitative evaluations, several visualizations were generated to gain insights into model behavior. Figure **??** illustrates the distribution of shape complexity values across different class labels in the training set. The histogram displays a moderate differentiation between classes, although the overlaps in distributions hint at the inadequacy of the current features for fully delineating the symbolic nuances. Figure **??** presents the confusion matrix obtained from the test set predictions, highlighting systematic misclassifications

that suggest certain key symbolic patterns are being overlooked. The misclassification patterns further motivate our proposal of integrating advanced rule extraction techniques to capture the hidden chaining constraints.

Preliminary simulations, where candidate rule activation features were added to the baseline model, indicate an improvement in both accuracy metrics. However, these simulations represent early experimental attempts; rigorous optimization of the ILP-based refinement process and the candidate rule extraction mechanism is required before drawing definitive conclusions regarding the efficacy of the hybrid approach. The experimental outcomes, when considered in aggregate, strongly support the hypothesis that the current feature set is insufficient to fully capture the latent symbolic dependencies, and that incorporating explicit rule extraction processes represents a promising avenue for future research.

8 Discussion

The experimental findings reveal that while a baseline logistic regression model with a basic set of features exhibits modest performance on the SPR_BENCH task, significant performance improvements can be envisaged by augmenting the model with advanced symbolic rule extraction and refinement techniques. The current performance—56.52% standard accuracy and 55.32% SWA on the test set—provides an important benchmark, yet it falls short of SOTA results by nearly 13–14 percentage points. One of the primary challenges in SPR tasks is the intricate interplay of latent symbolic relationships that, when not captured, lead to systematic misclassifications as evidenced by the confusion matrix analysis.

A key insight derived from our work is the importance of representing the latent structural attributes of token sequences. The histogram of shape complexity, although illustrative, suggests that surface-level aggregation is unable to encode the deeper interdependencies that drive the hidden chaining process. As such, future models must consider incorporating mechanism to dynamically update feature representations, potentially through iterative refinement cycles that rely on model feedback. The proposed integration of large language model-generated candidate rules and ILP-guided validation represents a promising strategy to address this challenge.

Furthermore, the use of rule-based features may provide additional interpretability to the model, enabling practitioners to not only improve predictive performance but also to understand the symbolic rationale behind classifications. This interpretability is paramount in domains where understanding decision-making processes is as critical as achieving high accuracy. The work of previous studies (e.g., Chenouard and Jouault, arXiv:1003.0746v1) reinforces the importance of clear rule extraction mechanisms that facilitate the chaining of transformations, a notion we have revisited in the current study with promising initial results.

There are several avenues for future work that emerge from the present

study. Firstly, a deeper exploration of non-linear aggregation functions beyond the simple logistic regression is warranted. Incorporating deep neural network architectures, perhaps in the form of hybrid models that combine symbolic reasoning and subsymbolic feature learning, could further enhance performance. Secondly, the integration of feedback mechanisms for rule extraction should be rigorously explored; iterative refinement schemes that dynamically update candidate rules based on misclassification rates could lead to more resilient models.

Additionally, a more granular analysis of token dependencies and their temporal evolution within sequences might yield useful insights. Techniques from sequential modeling—such as recurrent neural networks or transformers—could be integrated with rule extraction processes to capture both local and global dependencies. Lastly, extending the evaluation framework to include additional metrics, such as F1-score and area under the ROC curve, might provide a more comprehensive picture of model performance, particularly in imbalanced classification contexts where simple accuracy metrics may be misleading.

In conclusion, this work represents a modest yet significant step toward the development of robust SPR systems capable of leveraging latent symbolic structures. While the current baseline performance is limited, the experimental results and ensuing analysis strongly suggest that hybrid approaches combining statistical learning with explicit rule extraction can narrow the performance gap toward SOTA methods. The proposed future directions hold promise not only for improving accuracy but also for fostering greater interpretability in complex symbolic pattern recognition tasks.