Scientific Equation Discovery via Evolutionary Program Search with Large Language Models

Parshin Shojaee[1], **Kazem Meidani**[2], **Shashank Gupta**[3], **Amir Barati Farimani**[2], **Chandan K Reddy**[1]

[1]Virginia Tech [2]Carnegie Mellon University [3]Allen Institute for AI

Introduction

Scientific equation discovery is a crucial aspect of computational scientific discovery, traditionally approached through symbolic regression (SR) methods that focus mainly on data-driven equation search. Recent advancements have explored multi-island evolutionary search (Cranmer 2023), language model-guided search at the decoding stage (Shojaee et al. 2024a) and latent space search with multi-modal representation learning bridging mathematical expressions with numeric observations (Meidani et al. 2024). However, these methods often struggle to fully leverage the rich domain-specific knowledge that scientists rely on. Building upon these methods, we propose LLM-SR (Shojaee et al. 2024b), an integrated approach that combines large language models (LLMs) with evolutionary program search to discover scientific equations more effectively and efficiently, while incorporating scientific prior knowledge.

Components and Integration

LLM-SR integrates several key aspects of discovery: scientific knowledge representation and reasoning (via LLMs's prompting and prior knowledge), hypothesis generation (equation skeleton proposals by LLMs), data-driven evaluation and optimization, and evolutionary search for iterative refinement. The system takes as input a natural language description of the scientific problem and relevant variables, a dataset of input-output pairs representing observed data, and structured LLM prompts containing problem specifications and in-context examples. It outputs equation skeletons as Python programs representing mathematical equations with placeholder parameters, optimized equations with parameters fitted to data, and performance metrics such as Normalized Mean Squared Error (NMSE) for both in-domain and out-of-domain data.

The integration of these components occurs in an iterative loop: First, the LLM generates equation skeleton hypotheses based on its scientific prior knowledge, the provided problem description, and in-context examples. These skeletons are represented as Python programs with placeholder parameters. Then, data-driven hypothesis optimization is conducted using off-the-shelf optimizers (e.g., BFGS or Adam) to fine-tune the skeleton parameters (i.e., equation coefficients/constants). The equations with optimized parameters are then evaluated based on validation data, using metrics such as Normalized Mean Squared Error (NMSE). This evaluation signal guides the evolutionary search process, informing the selection and refinement of promising equation candidates. The best-performing equations are stored in an experience buffer, which is used to update the in-context examples for the next iteration, allowing the LLM to learn from and build upon successful discoveries. This integrated approach supports the discovery of interpretable and physically meaningful equations, efficient exploration of the equation search space, generalization to out-of-domain data, and integration of scientific prior knowledge with data-driven optimization.

Findings and Future Directions

We evaluated LLM-SR's effectiveness across three diverse scientific domains: nonlinear oscillators, bacterial growth, and material stress behavior. In each case, LLM-SR discovered accurate and physically meaningful equations. LLM-SR consistently outperforms state-of-the-art symbolic regression baselines, achieving lower NMSE scores on both in-domain (ID) and out-of-domain (OOD) data across all three domains. Our work demonstrates the potential of integrating LLMs with evolutionary search and data-driven optimization for scientific equation discovery. This approach not only improves the accuracy and interpretability of discovered equations but also enhances the efficiency of the search process by leveraging scientific prior knowledge.

Future work could explore integrating LLM-SR with experimental design and data collection modules to create a more comprehensive computational scientific discovery system. By combining the strengths of LLMs, evolutionary algorithms, and data-driven optimization, LLM-SR represents a significant step towards more integrated and effective approaches to computational scientific discovery.

References

Cranmer, M. 2023. Interpretable machine learning for science with PySR and SymbolicRegression. jl. *arXiv preprint arXiv:2305.01582*.

Meidani, K.; Shojaee, P.; Reddy, C. K.; and Farimani, A. B. 2024. SNIP: Bridging Mathematical Symbolic and Numeric Realms with Unified Pre-training. *The Twelfth International Conference on Learning Representations*.

Shojaee, P.; Meidani, K.; Barati Farimani, A.; and Reddy, C. 2024a. Transformer-based planning for symbolic regression. *Advances in Neural Information Processing Systems*, 36.

Shojaee, P.; Meidani, K.; Gupta, S.; Farimani, A. B.; and Reddy, C. K. 2024b. LLM-SR: Scientific equation discovery via programming with large language models. *arXiv preprint arXiv:2404.18400*.