The Journal of Open Source Software

DISCOVER: A Physics-Informed, GPU-Accelerated Symbolic Regression Framework

Udaykumar Gajera[1,2], Mohsen Sotoudeh[3,4], Kanchan Sarkar[2,3], and Axel Groß[2,3]

1 Department of Physics and CSMB, Humboldt-Universität zu Berlin, Berlin, Germany **2** Institute of Theoretical Chemistry, Ulm University, Oberberghof 7, 89081 Ulm, Germany **3** Helmholtz Institute Ulm (HIU), Electrochemical Energy Storage, 89081 Ulm, Germany **4** Karlsruhe Institute of Technology (KIT), P.O. Box 3640, D-76021 Karlsruhe, Germany

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor:

Submitted: 28 January 2026
Published: unpublished

Summary

Symbolic Regression (SR) enables the discovery of interpretable mathematical relationships from experimental and simulation data. These relationships are often coined descriptors which are defined as a fundamental materials property that is directly correlated to a desired or undesired functional property of the material. Although established approaches such as Sure Independence Screening and Sparsifying Operator (SISSO) have successfully identified low-dimensional descriptors within large feature spaces (Ouyang et al., 2018), many existing SR tools integrate poorly with modern Python workflows, offer limited control over the symbolic search space, or struggle with the computational demands of large-scale studies.

This paper introduces DISCOVER (Data-Informed Symbolic Combination of Operators for Variable Equation Regression), an open-source symbolic regression package developed to address these challenges through a modular, physics-motivated design. DISCOVER allows users to guide the symbolic search using domain knowledge, constrain the feature space explicitly, and take advantage of optional GPU acceleration to improve computational efficiency in data-intensive workflows, enabling reproducible and scalable SR workflows. The software is intended for applications in computational physics, computational chemistry, and materials science, where interpretability, physical consistency, and execution time are especially important, and it complements general-purpose SR frameworks by emphasizing the discovery of physically meaningful models (Muthyala et al., 2025).

Statement of Need

Symbolic regression is widely used in scientific domains where interpretability and physical insight are essential, including physics, chemistry, and materials science. This insight can be expressed as a descriptor which corresponds to a correlation between a fundamental materials property and a desired or undesired function of the material (Sotoudeh & Groß, 2022). While many SR methods can recover analytical expressions from data (Udrescu & Tegmark, 2020), practical adoption is often limited by several factors: insufficient integration with Python-based scientific workflows, limited mechanisms for incorporating *a priori* physical knowledge, and high computational cost when exploring large symbolic search spaces. These challenges make it difficult for researchers to apply SR methods efficiently and reproducibly in real-world scientific studies.

DISCOVER addresses this gap by providing a Python-native symbolic regression framework that explicitly supports physics-informed constraints and optional GPU-accelerated computation. By allowing users to define constraints on operators, feature combinations, and physical consistency through a configuration-based interface, DISCOVER lowers the barrier to incorporating domain knowledge into SR workflows. Its design supports reproducible experimentation, efficient exploration of constrained search spaces, and seamless integration into existing scientific Python ecosystems.

State of the Field

Existing tools such as SISSO provide powerful, deterministic strategies for identifying sparse descriptors but are not designed to offer fine-grained, user-defined control over the symbolic search or to leverage modern hardware acceleration as a core feature (Ouyang et al., 2018; Purcell et al., 2023). Conversely, more flexible or physics-informed SR approaches (PiSR) may require complex customization or lack scalable performance (**?**). As a result, researchers often face trade-offs between interpretability, usability, and computational efficiency.

Recent symbolic regression tools have demonstrated impressive capabilities in recovering analytical expressions from data. For example, AI Feynman (Udrescu & Tegmark, 2020) leverages symbolic manipulation and neural-guided search to rediscover known physical laws, while extensions of the SISSO framework, such as SISSO++ (Purcell et al., 2023), continue to advance large-scale descriptor discovery through efficient sparsity-driven screening. These methods represent important progress in the field; however, they often prioritize either fully automated discovery or highly specialized workflows, and may offer limited flexibility for incorporating fine-grained physical constraints, modern Python integration, or hardware acceleration as first-class features.

Software Design

DISCOVER is an open-source symbolic regression package designed for the guided discovery of interpretable mathematical expressions. The software generates candidate symbolic expressions from user-provided features and operator libraries, evaluates them against target data, and identifies parsimonious models that balance accuracy and simplicity. The search process is iterative and incorporates pruning strategies informed by user-defined physical constraints.

To support sparse model discovery, DISCOVER provides access to multiple sparsifying search strategies, including heuristic, optimization-based, and stochastic approaches such as Orthogonal Matching Pursuit (OMP) (Tropp & Gilbert, 2007), Mixed-Integer Quadratic Programming (MIQP) (Lazimy, 1982), and Simulated Annealing (Eglese, 1990). The software architecture is modular and Python-native, enabling straightforward integration with common scientific libraries. Computationally intensive operations such as feature generation and model evaluation are parallelized and executed on hardware accelerators when available. For large-scale studies, DISCOVER supports optional GPU acceleration via CUDA on NVIDIA GPUs and Metal Performance Shaders (MPS) on Apple Silicon devices, while maintaining efficient CPU-based execution for standard workloads. This hardware-aware design enables scalable symbolic regression workflows on both high-performance computing systems and local development environments.

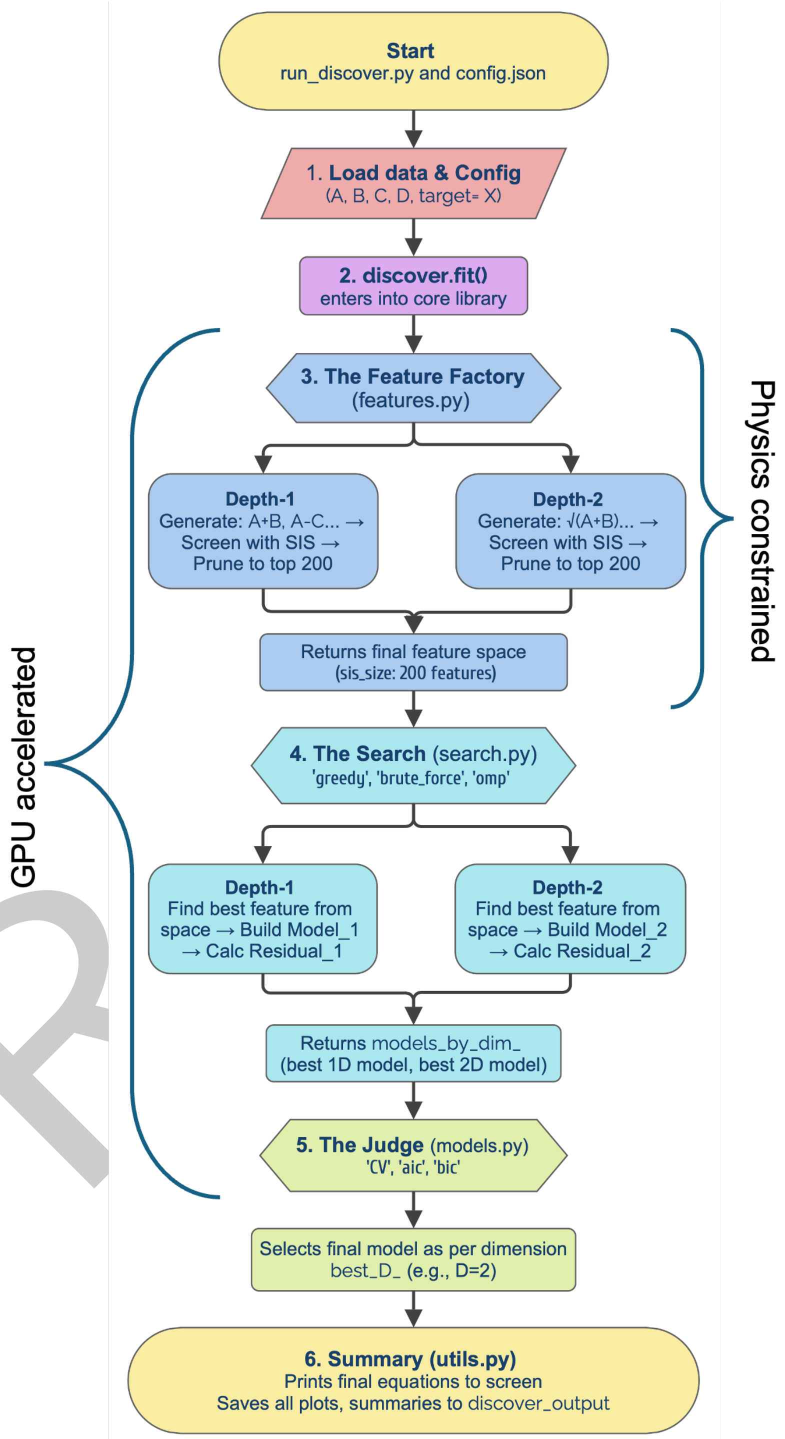

Figure 1: Overview of the DISCOVER workflow, illustrating iterative feature generation, physics-informed screening, and sparse model selection.

Core Optimization Objective

All search strategies implemented in DISCOVER are designed to approximate or solve a common underlying optimization problem. Given a set of M candidate symbolic features $\mathbf{\Phi} = \{\Phi_1, \Phi_2, \dots, \Phi_M\}$ and a target property vector $\mathbf{y}$, the objective is to identify a sparse linear combination of features that accurately models the data. This problem can be expressed as an L_0-regularized least-squares regression:

$$\min_{\beta} \|\mathbf{y} - \mathbf{\Phi}\beta\|_2^2 \quad \text{subject to} \quad \|\beta\|_0 \leq D,$$

where β is the coefficient vector and $\|\beta\|_0$ denotes the number of nonzero entries, enforcing a maximum descriptor dimensionality D. This formulation is common in sparse symbolic regression and descriptor discovery and is known to be NP-hard. To put it simply, there is no known mathematical shortcut to efficiently find the exact optimal solution for this type of problem. Therefore, instead of trying to calculate the impossible 'perfect' answer, we must rely on smart approximation strategies to find a high-quality model within a reasonable timeframe. As a result, DISCOVER offers multiple heuristic, approximate, and specialized search strategies to explore this objective efficiently under user-defined physical and computational constraints.

Physics-Informed Constraints

A central design goal of DISCOVER is to facilitate the explicit incorporation of domain knowledge into the symbolic regression process. Physical constraints are specified through a configuration-based interface and applied during expression generation and evaluation. Dimensional consistency is enforced through integration with the `pint` unit-handling library, enabling unit-aware symbolic operations and validation of candidate expressions. By tracking physical units throughout the search process, DISCOVER can exclude dimensionally invalid expressions early, reducing the effective search space and promoting the discovery of physically meaningful and interpretable models.

Design Philosophy and Constraints

A core design goal of DISCOVER is to enable direct incorporation of domain expertise into the symbolic regression process. Rather than relying solely on automated sparsity or heuristic search (Talapatra et al., 2022), DISCOVER allows users to specify constraints via a configuration file without modifying source code. Supported constraints include enforcement of dimensional consistency (Tenachi et al., 2023), restrictions on allowed operators or expression complexity, and user-defined rules governing variable combinations and functional forms (Kronberger et al., 2022). These constraints reduce the effective search space, improve interpretability, and help ensure that discovered expressions are physically meaningful. This approach is particularly useful in scientific domains where prior knowledge is well established and model plausibility is as important as predictive accuracy (Keren et al., 2023).

Research Impact Statement

DISCOVER is intended for scientific applications where symbolic regression is used as a tool for model discovery rather than purely predictive performance. Typical use cases include identifying low-dimensional descriptors for physical or chemical properties, such as crystal structure stability (Gajera et al., 2022) or ion mobility in energy storage materials (Sotoudeh & Groß, 2022). The software is especially suited to computational physics, computational chemistry, and materials science workflows that benefit from Python integration and hardware-accelerated computation, spanning from battery cathode discovery (Lu et al., 2021) to accurate discrimination of magnetic structure (Suzuki et al., 2023).

Limitations

The effectiveness of DISCOVER depends on the quality of the input features and the appropriateness of user-defined constraints. Overly restrictive constraints may exclude valid expressions, while insufficient constraints can lead to large search spaces with increased computational cost. Although GPU acceleration improves performance for many workloads, DISCOVER is not optimized for fully unconstrained searches over extremely large feature spaces compared to specialized low-level implementations such as SISSO (Ouyang et al., 2018). Ongoing development focuses on expanded operator libraries, improved benchmarking, and scalability enhancements.

AI Usage Disclosure

During the preparation of this work, the authors used large language models to assist in refactoring the source code. Specifically, AI tools were utilized to remove redundant functions, generate explanatory comments for complex logic, and standardize function naming conventions (e.g., renaming legacy short-form functions like `r_dis()` to the more descriptive `run_discover()`).

Acknowledgements

U.G. acknowledges primary support from the NFDI consortium FAIRmat, funded by the Deutsche Forschungsgemeinschaft (DFG) under project 460197019. Furthermore, this work contributes to the research performed at CELEST (Center for Electrochemical Energy Storage Ulm-Karlsruhe). Support by the German Research Foundation (DFG) through the POLiS Cluster of Excellence, Project ID 390874152, and by the Dr. Barbara Mez-Starck-Foundation is gratefully acknowledged.

References

Eglese, R. W. (1990). Simulated annealing: A tool for operational research. *European Journal of Operational Research*, *46*(3), 271–281. https://doi.org/10.1016/0377-2217(90)90001-R

Gajera, U., Storchi, L., Amoroso, D., Delodovici, F., & Picozzi, S. (2022). Toward machine learning for microscopic mechanisms: A formula search for crystal structure stability based on atomic properties. *Journal of Applied Physics*, *131*(21), 215703. https://doi.org/10.1063/5.0088177

Keren, L. S., Liberzon, A., & Lazebnik, T. (2023). A computational framework for physics-informed symbolic regression with straightforward integration of domain knowledge. *Scientific Reports*, *13*(1), 1249. https://doi.org/10.1038/s41598-023-28328-2

Kronberger, G., Franca, F. O. de, Burlacu, B., Haider, C., & Kommenda, M. (2022). Shape-constrained symbolic regression—improving extrapolation with prior knowledge. *Evolutionary Computation*, *30*(1), 75–98. https://doi.org/10.1162/evco_a_00294

Lazimy, R. (1982). Mixed-integer quadratic programming. *Mathematical Programming*, *22*(1), 332–349. https://doi.org/10.1007/BF01581047

Lu, Z., Zhu, B., Shires, B. W. B., Scanlon, D. O., & Pickard, C. J. (2021). Ab initio random structure searching for battery cathode materials. *The Journal of Chemical Physics*, *154*(17), 174111. https://doi.org/10.1063/5.0049309

Muthyala, M. R., Sorourifar, F., Peng, Y., & Paulson, J. A. (2025). *SyMANTIC: An efficient symbolic regression method for interpretable and parsimonious model discovery in science and beyond*. https://doi.org/10.1021/acs.iecr.4c03503

Ouyang, R., Curtarolo, S., Ahmetcik, E., Scheffler, M., & Ghiringhelli, L. M. (2018). SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. *Phys. Rev. Mater.*, *2*, 083802. https://doi.org/10.1103/PhysRevMaterials.2.083802

Purcell, T. A. R., Scheffler, M., & Ghiringhelli, L. M. (2023). *Recent advances in the SISSO method and their implementation in the SISSO++ code*. https://doi.org/10.1063/5.0156620

Sotoudeh, M., & Groß, A. (2022). Descriptor and scaling relations for ion mobility in crystalline solids. *JACS Au*, *2*(2), 463–471. https://doi.org/10.1021/jacsau.1c00505

Suzuki, M.-T., Nomoto, T., Morooka, E. V., Yanagi, Y., & Kusunose, H. (2023). High-performance descriptor for magnetic materials: Accurate discrimination of magnetic structure. *Phys. Rev. B*, *108*, 014403. https://doi.org/10.1103/PhysRevB.108.014403

Talapatra, A., Gajera, U., P, S. P., Arout Chelvane, J., & Mohanty, J. R. (2022). Understanding the magnetic microstructure through experiments and machine learning algorithms. *ACS Applied Materials & Interfaces*, *14*(44), 50318–50330. https://doi.org/10.1021/acsami.2c12848

Tenachi, W., Ibata, R., & Diakogiannis, F. I. (2023). Deep symbolic regression for physics guided by units constraints: Toward the automated discovery of physical laws. *The Astrophysical Journal*, *959*(2), 99. https://doi.org/10.3847/1538-4357/ad014c

Tropp, J. A., & Gilbert, A. C. (2007). Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, *53*(12), 4655–4666. https://doi.org/10.1109/TIT.2007.909108

Udrescu, S.-M., & Tegmark, M. (2020). AI feynman: A physics-inspired method for symbolic regression. *Science Advances*, *6*(16), eaay2631. https://doi.org/10.1126/sciadv.aay2631