Towards Data-Driven Scientific Discovery

Arush Tagade[a], **Leo Mckee-Reid**[a], **Robbie McCorkell**[a], **Andrew Cusick**[a], **Sebastian Sosa**[a], **Matthieu Platre**[b], **Jessica Rumbelow**[a], **Zohreh Shams**[a]

[a] *Leap Labs, UK* {arush, leo, robbie, andrew, sebastian, jessica, zohreh}@leap-labs.com

[b] *Montpellier Institute of Plant Sciences (IPSIM), France* matthieu.platre@inrae.fr

1. Introduction

We propose a novel system for automated scientific knowledge discovery at scale, referred to as Discovery Engine (DE). Traditional data analysis methods rely on human assumptions and linear patterns, while performant machine learning models find complex patterns but remain opaque. Our system overcomes these limitations by combining machine learning's pattern recognition ability with the SOTA interpretability methods [1, 2] that broadly aim at shedding light on decision making of machine learning models. This enables human-understandable insights from complex data that would otherwise remain hidden.

While existing work shows AI's potential for scientific discovery, automation across diverse datasets remains unexplored. The exception to this are Large Language Model (LLM) driven discovery pipelines that aim to go from data to insight directly [3, 4]. However, in practice due to data confidentiality and/or context length issues, it is not trivial to expose the LLM to the entire dataset. This makes it very difficult to judge what part of the discovery is in fact specifically related to the dataset, rather than the LLM's knowledge based on its training data. This is on top of LLM shortcomings in mathematical tasks in general [5, 6]. DE addresses these limitations, making AI-driven discovery data-driven, automated, systematic and reproducible.

Below we provide an overview of DE components, followed by two case studies. For the case study that has ground truth patterns, we make a comparison with the ground truth, and results from a pure LLM pipeline. In the second case study we validate the patterns by exposing them to the academic expert who has collected the dataset for the study.

2. Discovery Engine

Figure 1 shows an overview of DE. In what follows we provide details about each component.

Data Ingestion: In this step the data gets preprocessed automatically. Operations done include imputation of missing values, duplication removal, elimination of correlated columns as well as handling of categorical and continuos variables.

AutoML The pre-processed data is modelled using an AutoML component. Whilst AutoML tools exist, they are typically unsuitable for interpretability-driven scientific discovery as they rely on transfer learning from general pre-trained models, and do not optimise for interpretability. Our AutoML is tailored for scientific discovery and includes several models in various categories: linear (e.g., linear regression), tree-based (e.g., XGBoost [7]), kernel-based (e.g., SVM [8]) and deep learning models (e.g., autoencoders [9]). Whilst discovery often requires sophisticated models capable of capturing complex patterns, we cater for simpler models as well to ensure that a highly parameterised model overfitting a simple dataset will not form the basis of discovery.

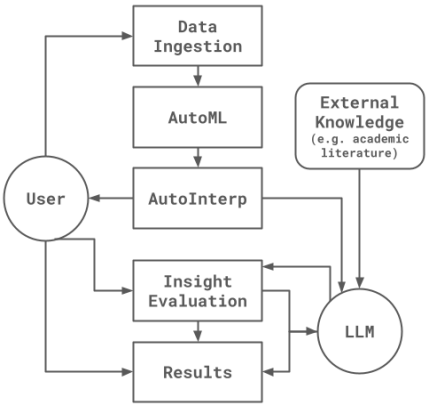

Fig. 1: The Discovery Engine Schematic

AutoInterp: In this component a suite of interpretability methods is applied to the trained models to elicit patterns in the data. These methods include, but are not limited to (i) feature importance methods that highlight the features a model focuses on the most; (ii) top examples, which are training points that are most representative of a class or a high regression value; (iii) prototypical examples, which are synthetic data points generated via an optimisation process, such that they maximally activate a certain class or low/high regression value. These examples are at the heart of the discovery process as they exaggerate the patterns learnt by a model; (iv) global counterfactuals, which are also synthetic data points, and with minimal changes swap from being a member of one class to another or from very high to low regression values.

Insight Evaluation: The interpretability artefacts generated are analysed using this component in order to identify and prioritise novel patterns. The analysis ranges from discarding noisy artefacts to ranking and grouping them in order to infer robust patterns derived from them collectively. Whilst the heavy lifting of modelling and mathematical reasoning is done in the AutoML and AutoInterp compo-

nents, here we utilise LLMs to explain and contextualise interpretability results (with reference to external sources of knowledge, such as arXiv). The extracted patterns, whether LLM-driven or not, are subjected to validation prior to being revealed.

3. Experiments

3.1 ITR Dataset

We present a case study on an Interfacial Thermal Resistance (ITR) dataset collected by Wu et al [10] to understand how material properties affect the ITR between two interfacing films. We provide more detail about the dataset and the hypotheses that have been inferred from studying it in Appendix A.

We treat these hypotheses as ground truth and evaluate the extent to which they can be generated by DE and a pure LLM-driven discovery system (DiscoveryBench (DB) [3]). DB is an agentic pipeline that provides LLM agents with access to a dataset and tools to generate Python code that can then be applied to the data to analyse it. This pipeline relies on the existence of a ground truth hypothesis framed as a query that the pipeline aims to investigate. The result of the investigation forms hypotheses that may or may not be aligned with the ground truth ones. To score the similarity of the generated hypotheses with the ground truth ones, the pipeline includes an evaluation procedure that matches context across generated and ground truth hypotheses as well as the variables they use and the relationship between such variables. The context match score, variable overlap score and variable relationship score are obtained through LLM-based evaluation as a number between 0 and 1 and multiplied to give a final score.

Unlike DB, DE is not necessarily hypothesis driven, however for fairness, we expose both pipelines to queries extracted based on ITR ground truth hypotheses listed in Appendix B. We then use OpenAI's o1 as the LLM of choice in both pipelines and report on results in Table 1. Since o1 doesn't allow temperature to be set to 0, we run the evaluation a total of 5 times to control for LLM randomness. DE scores better than DB for all queries. In fact in 3 out of 4 cases, where DB scores 0, the agent returns code containing errors leading to hypotheses about the errors rather than addressing the queries.

Table 1: Comparison of evaluation scores between our Discovery Engine (DE) and DiscoveryBench [3] (DB) automated pipelines

	Q1	Q2	Q3	Q4	Avg
DE	0.25 ± 0	0.35 ± 0.09	0.20 ± 0.08	0.195 ± 0.05	**0.25**
DB	0 ± 0	0 ± 0	0.32 ± 0.02	0 ± 0	0.08

3.2 Plant Bio Dataset

In collaboration with Mattieu Platre from the Montpellier Institute of Plant Sciences, we analysed a dataset on the early root architecture of *Arabidopsis Thaliana* within the first 16 days of growth. Details of dataset can be found in Appendix C.

We use DE for pattern discovery on this dataset and divide our findings into: reproducing known benchmark patterns, novel patterns validated by the data, and hypotheses requiring experimental validation. We mainly focus on genotype and nutrient features due to their biological significance and less explored nature compared to other features in the data.

Reproducing benchmark patterns: DE found patterns that have been confirmed by our collaborator and existing literature. One example is the combination of the WT (the "Wild Type" genotype found in nature) and the N110_275 nutrient (a mid-range level of Nitrate) for maximizing the total root length. At this level of nitrate, the root system triggers the plant's foraging mode, which increases the total root length much more than when provided at sufficient levels [11]. Additionally, the results reaffirmed that total root length increases with time and certain known temperature and CO_2 levels.

Validated novel patterns: Our model unexpectedly favoured the BRL3-2 genotype when optimizing total root length under nitrate limiting conditions. Since BRL3 is involved in the brassinosteroid signalling pathway and no clear phenotypes have been reported for this mutant under nitrate limiting conditions, this was surprising. Further analysis revealed that BRL3-2 alone had a lower mean root length than the dataset average, but when paired with N11400_110, it produced significantly higher total root length. This suggests a synergistic effect, where certain genotype-nutrient combinations outperform their individual contributions.

Novel hypotheses for experimental validation: DE proposed some genotype-nutrient pairings not observed in the data, such as WT with N11400_550. These predictions represent testable hypotheses that our collaborator has expressed interest in validating through laboratory experiments.

4. Summary

Despite their limitations, when successful, pure LLM pipelines allow a level of scalability that far surpasses the common task-specific AI discovery (where a certain modelling paradigm is chosen and trained for a dataset and then analysed for insights). We propose a pipeline that offers the best of both worlds: it allows an end-to-end scientific discovery process, where interpretability artefacts capture the essence of the data in a succinct and precise manner that can then be investigated using LLMs, domain experts or validated by contrasting to the data.

We verify DE performance quantitatively (ITR case study) and qualitatively (Plant Bio case study). In the former, whilst DE outperforms DB, the gap could have been even wider if the evaluation procedure was more open ended rather than as restrictive as comparing variable names. In the latter we show the biological relevance of the findings that has motivated the experimental validation of our results.

References

[1] Christoph Molnar. *Interpretable Machine Learning*. 2 edition, 2022.

[2] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable AI: A review of machine learning interpretability methods. *Entropy*, 23(1):18, 2020.

[3] Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. Discoverybench: Towards data-driven discovery with large language models. *arXiv preprint arxiv.org/abs/2407.01725*, 2024.

[4] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arxiv.org/abs/2408.06292*, 2024.

[5] Seyed Iman Mirzadeh et al. GSM-Symbolic: Understanding the limitations of mathematical reasoning in large language models. To appear in the proceeding of ICLR 2025.

[6] Johan Boye and Birger Moell. Large language models and mathematical reasoning failures. *arXiv preprint arXiv:2502.11574*, 2025.

[7] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[8] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[9] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[10] Yen-Ju Wu, Tianzhuo Zhan, Zhufeng Hou, Lei Fang, and Yibin Xu. Physical and chemical descriptors for predicting interfacial thermal resistance. *Scientific Data*, 7(1):36, 2020.

[11] Benjamin D Gruber, Ricardo FH Giehl, Swetlana Friedel, and Nicolaus von Wirén. Plasticity of the arabidopsis root system under nutrient deficiencies. *Plant physiology*, 163(1):161–179, 2013.

Appendix A. ITR Dataset

The ITR dataset comprises 1318 data points denoting 457 interface combinations across 54 materials, including metals, insulators, and semiconductors. The 457 interfaces are defined by their films, interlayers, substrate materials, and experimental conditions. Besides experimental conditions like the temperature at which ITR was measured, the dataset also contains information about the thickness of films in the interface and material properties like melting point, binding energy etc.

The ground truth hypotheses inferred by studying the ITR dataset are as follows:

1. Film melting point, and film/substrate mass have a linear correlation with ITR
2. There is a strong linear correlation between atomic coordinates, binding energy and ionic potential
3. Material systems Bi/graphite, Bi/diamond, and Bi/B are predicted to have high ITR and are not present in the original dataset
4. f_melt (Film melting point) and s_melt (Substrate melting point) show opposite linear trends

Appendix B. Crafted Queries

The queries crafted to enable evaluating our Discovery Engine and DiscoveryBench [3] to align with the ground truth hypotheses presented in Section 3.1 are as follows:

- Q1: How is film melting point and film/substrate mass related to ITR?
- Q2: How are atomic coordinates, binding energy and ionic potential related to each other?
- Q3: What material systems are predicted to have high ITR and are not present in the original dataset?
- Q4: What is the relation between f_melt and s_melt?

Appendix C. Plantbio Dataset

This dataset contains over 10 thousand plant samples, each measuring 18 root system architecture (RSA) measurements and 6 experimental features (days since planting, CO_2 levels, room temperature, genotype, soil nutrients, and sorbitol levels). While we explored multiple target variables, this paper focuses on understanding how these 6 features impact only one of the RSA measurements: total root length.