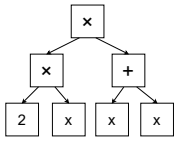
Research Statement - Prof. Fabricio Olivetti de França, Ph.d.

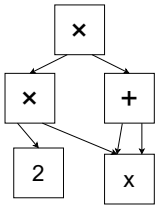
Discovering mathematical equations from observations is an important step for understanding scientific laws. With the growth in the capacity of measuring and acquiring new data, automating this step is crucial for processing the large quantity of available data. Symbolic regression (SR) automates this step by searching the space of mathematical functions and returning a set of alternative hypotheses for the studied law. The current state-of-the-art suffers with many challenges such as: 1) irregular search space, which makes the search inefficient, 2) underspecification, allowing multiple high-quality hypothesis with different extrapolating behavior, 3) lacking a proper integration with domain knowledge, which is essential to find an equation that correctly explains the studied phenomena.

My research involves proposing a symbolic regression algorithm that incorporates multiple sources of domain knowledge and the user's desiderata to guide the search process towards a mathematical expression that accurately describes the data while remaining transparent, allowing the scientist to understand it, verify its validity, and using it to explain the phenomena. Transparency is an important and current topic in AI, especially in sensitive applications that requires a careful consideration of the impact of a model. Understanding the model allows us to debug and correct any critical issue that can lead to an economical, societal, or life threatening impact. Besides avoiding negative impacts, having a transparent model allows a better understanding of the phenomena.

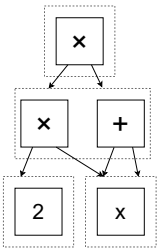
The base of this research is the data structure called *equality graph* (e-graph) that will act as a database of the history of the search that can efficiently retrieve visited expressions and their building blocks. This also allows to perform relational queries to constrain the retrieved expressions to those having the desired properties. So far, a simple proof of concept [6] showed that this significantly improves the efficiency of equation discovery with interesting possibilities to integrate domain knowledge.



(a)



(b)



(c)

Figure 1: The expression $2x(x + x)$ represented as (a) a tree, (b) a directed acyclic graph and, (c) an e-graph.

Equality Graphs: the Future of Symbolic Regression

Equality graphs (e-graphs) extends the traditional graph structure to represent equivalence relationships. It is used in practice for optimization of computer programs to reduce the effects of the phase ordering problem. In symbolic regression, it can be used to represent a mathematical function efficiently as illustrated in Fig.1. While many symbolic regression algorithms conveniently represents expressions as trees (Fig. 1a), a directed acyclic graph (Fig. 1b) can help reducing the computation of repeated components. This is further improved with e-graphs (Fig.1c), in which different components that evaluate to the same value (i.e., they are equivalent) are grouped together (dashed box) indicating that once we evaluate one of them, we do not need to evaluate the other.

This structure is often coupled with the algorithm called **equality saturation** that repeatedly applies equivalence rules to an initial expression extending the e-graph with equivalent components until saturation, at which point every possible ways of writing such expression is compactly stored in the e-graph.

This can be used to simplify any expression by applying equality saturation and extracting the path from the root with the minimal cost. The cost is a heuristic function defining the user preference such as minimum number of nodes, minimum use of numerical parameters, or avoiding repeated variables in the expression. We have found that simplifying expressions generated by different SR algorithms with equality saturation works better than using the mature symbolic mathematics library SymPy [5]. Not only that, but this experiment also revealed a common problem with many SR algorithms, the introduction of redundant parameters which hinders the data fitting process [8].

In another work, we have used this simplification to detected the visitation of duplicated expressions during the SR search [9]. After storing the history of visitations in different runs of a minimalist genetic programming algorithm to different datasets, we found that up to 60% of the visited expressions were already visited at different points of the search in its exact or equivalent form. One reason for this, is that the search favors incremental improvements of the current solutions, but any minimal change to an expression often leads to a significantly negative impact in the model performance, thus, generating an equivalent expression that does not affect the performance, allows the algorithm to navigate the search space. This shows the inefficiency of searching for symbolic models due to the irregularity of the search space.

This navigability problem can be alleviated by constraining the search space to parametric functions that are easier to fit and with a placement of numerical parameters that regulates any change in the expression. This is such the case for Interaction-Transformation [1] and Transformation-Interaction-Rational [3], which constraint the

search to functions that are a linear combination of transformed features. This allows to pose the search problem as an optimization problem, that can be structured as a multi-layer neural network [1], for example.

Recently, we adapted a minimalist genetic programming algorithm to store all visited expressions on a single e-graph [6]. This allowed us to modify the recombination and perturbation operators to enforce the creation of novel expressions, which led to a boost in the efficiency of the search. The results show that by doing this simple modification to the search operators, the algorithm obtained results comparable to the state-of-the-art. It should be noted that this change can benefit any SR algorithm as it simply avoids visiting and evaluating redundant expressions.

During this stage of my research, I have implemented an open-source library supporting different functionalities needed for SR algorithms such as: evaluation, automatic differentiation, support to different distributions (Gaussian, Poisson, Bernoulli), statistical analysis and confidence intervals, and an interface with e-graph and equality saturation algorithm. This library, called **srtree**, is currently available at `https://github.com/folivetti/srtree`, together with different example applications that makes use of such library.

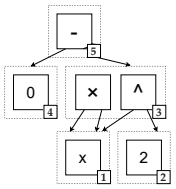


Figure 2: Example of e-graph with each group containing an unique id.

One of the highlights of this implementation, is the possibility to store additional information about the expressions in a relational database fashion. In the e-graph, each one of the dashed boxes in Fig. 1c is assigned to an unique id number. This id can be referenced in different data structures to assert any property of the expression rooted at that group. For example, if we have one of such group that represents the expressions $-x^2, -x \times x$ (Fig. 2), we can assert that the expressions with id = 3 will always be positive and convex and those with id = 5 will be negative and concave. If we assert that id = 1 is always positive, we can say that id = 3 will be monotonically increasing and id = 5 monotonically decreasing. These assertions can be supported by the use of interval arithmetic, providing we know the domain of the input variables. This can be useful to guide the search towards expressions that adhere to some shape-constraints (i.e., monotonicity, convexity) known *a priori* [2].

Additionally, by exploiting a locality sensitive hashing that approximates the Euclidean distance, it is possible to store a relationship between each id and a signature representing the predictions for input points of interest. With this structure, we can group together expressions that behave similarly inside this input domain enabling us to select expressions within a certain distance during recombination, promoting the combination of functions that are specialized in different regions of the data. This can also help to alleviate the underspecification problem [4], when we have multiple alternative functions that behave almost the same in the training data but can be divergent outside that boundaries. By exploiting the e-graph structure, we can list all the visited alternative functions and select a subset of those based on some desiderata associated with known properties.

A glimpse of these possibilities is already available through the rEGGression tool [7] also part of the *srtree* library, which allows to load a list of expressions or an e-graph binary file generated by many SR algorithms and explore the set with a SQL-like query. Among its features, it is possible to search for a functional pattern and retrieve statistics about the building blocks, such as the top building blocks with the highest average fitness. This can help users to explore the set of visited expression, not being limited to the single best expression returned by the algorithm.

References

[1] Interaction–transformation evolutionary algorithm for symbolic regression. *Evolutionary computation*, 29(3):367–390, 2021.

[2] Shape-constrained symbolic regression—improving extrapolation with prior knowledge. *Evolutionary computation*, 30(1):75–98, 2022.

[3] Fabrício Olivetti de França. Transformation-interaction-rational representation for symbolic regression. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 920–928, 2022.

[4] Fabricio Olivetti De Franca. Fighting underspecification in symbolic regression with fitness sharing. In *Proceedings of the Companion Conference on Genetic and Evolutionary Computation*, pages 551–554, 2023.

[5] Fabricio Olivetti de Franca and Gabriel Kronberger. Reducing overparameterization of symbolic regression models with equality saturation. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 1064–1072, 2023.

[6] Fabrício Olivetti de França and Gabriel Kronberger. Improving genetic programming for symbolic regression with equality graphs. *arXiv preprint arXiv:2501.17848*, 2025.

[7] Fabrício Olivetti de França and Gabriel Kronberger. reggression: an interactive and agnostic tool for the exploration of symbolic regression models. *arXiv preprint arXiv:2501.17859*, 2025.

[8] Gabriel Kronberger and Fabrício Olivetti de França. Effects of reducing redundant parameters in parameter optimization for symbolic regression using genetic programming. *Journal of Symbolic Computation*, 129:102413, 2025.

[9] Gabriel Kronberger, Fabricio Olivetti de Franca, Harry Desmond, Deaglan J Bartlett, and Lukas Kammerer. The inefficiency of genetic programming for symbolic regression. In *International Conference on Parallel Problem Solving from Nature*, pages 273–289. Springer, 2024.