# Class Symbolic Regression: Gotta Fit 'Em All

Wassim Tenachi, Rodrigo Ibata, Thibaut L. François, Foivos I. Diakogiannis

The Astrophysical Journal Letters, 969:L26 (7pp), 2024 July 10 https://doi.org/10.3847/2041-8213/ad5970


# Class Symbolic Regression: Gotta Fit 'Em All

Wassim Tenachi[1], Rodrigo Ibata[1], Thibaut L. François[1], and Foivos I. Diakogiannis[2]
[1] Université de Strasbourg, CNRS, Observatoire astronomique de Strasbourg, UMR 7550, F-67000 Strasbourg, France; wassim.tenachi@astro.unistra.fr
[2] Data61, CSIRO, Kensington, WA 6155, Australia
*Received 2023 December 4; revised 2024 June 14; accepted 2024 June 16; published 2024 July 2*

## Abstract

We introduce "Class Symbolic Regression" (Class SR), the first framework for automatically finding a single analytical functional form that accurately fits multiple data sets—each realization being governed by its own (possibly) unique set of fitting parameters. This hierarchical framework leverages the common constraint that all the members of a single class of physical phenomena follow a common governing law. Our approach extends the capabilities of our earlier Physical Symbolic Optimization (Φ-SO) framework for symbolic regression, which integrates dimensional analysis constraints and deep reinforcement learning for unsupervised symbolic analytical function discovery from data. Additionally, we introduce the first Class SR benchmark, comprising a series of synthetic physical challenges specifically designed to evaluate such algorithms. We demonstrate the efficacy of our novel approach by applying it to these benchmark challenges and showcase its practical utility for astrophysics by successfully extracting an analytic galaxy potential from a set of simulated orbits approximating stellar streams.

*Unified Astronomy Thesaurus concepts:* Neural networks (1933); Astronomy data analysis (1858); Astronomy software (1855); Open source software (1866); Analytical mathematics (38)

## 1. Introduction

Since the beginning of the scientific revolution, researchers have tried to find repeatable regularities in experiments and observations. Mathematical structures were used in this exploration, and many new ones including functions and differential equations were developed to respond to this need to model nature. Perhaps because of shared symmetries between nature and mathematics, these abstract structures have often been found to work exceedingly well in reproducing and predicting properties of the world, to the point where some have even considered whether the Universe is actually mathematical at heart (Tegmark 2008).

The symbolic regression (SR) that the present contribution is concerned with has a long pedigree. Perhaps its most famous application was by Kepler to planetary ephemerides, thereby finding the fitting law that bears his name (Kepler 1609). This empirical law gave the observational basis upon which Newton was able to build the physical theories developed in his Principia Mathematica (Newton 1687).

Modern SR (Schmidt & Lipson 2009, 2011; Kommenda et al. 2020; Kammerer et al. 2020; Bartlett et al. 2023b; Brence et al. 2021; Jin et al. 2019; Luo et al. 2022; Tohme et al. 2023; Udrescu & Tegmark 2020; Udrescu et al. 2020; Kamienny et al. 2022; Biggio et al. 2020, 2021; Vastl et al. 2024; Kamienny et al. 2023; Martius & Lampert 2017; Brunton et al. 2016; Zheng et al. 2022; Sahoo et al. 2018; Petersen et al. 2021a; Landajuela et al. 2022; Holt et al. 2023; Scholl et al. 2023; Sousa et al. 2024; Fiorini et al. 2024; Shojaee et al. 2024; Zhang & Lei 2024; Cheng & Alkhalifah 2024; He et al. 2024; Makke & Chawla 2022; Angelis et al. 2023; Faris et al. 2024; Tian et al. 2024; Michishita 2024; Melching et al. 2024; Meidani et al. 2024; Li et al. 2024a, 2024b; Chen et al. 2024) aims to use the immense computational resources at our disposal to search through possible analytic descriptions in terms of a set of functions and operators (e.g., $x$, $+$, $-$, $\times$, $/$, sin, cos, exp log, ...) to best fit some numerical data set ($\boldsymbol{x}$, $y$) we wish to model. Concretely, one seeks some analytic function $f\colon \mathbb{R}^n \longrightarrow \mathbb{R}$ that fits $y = f(\boldsymbol{x})$ given those data. It is worth pointing out here that the search space becomes exponentially larger the longer the analytic expression is that we seek to find. Hence the key to SR is to develop efficient schemes to search through the possibilities, and most importantly, to prune out poor choices.

Our modern computational abilities have allowed us to examine nature in unprecedented quantitative detail, with cameras, spectrographs, and other detectors amassing vast quantities of numerical data. It is likely that the clues to next-generation physics and understanding lie therein, and so we are tasked to devise methodologies capable of handling this wealth of information and translating it into coherent, interpretable, and intelligible physical models. The promise of SR is that it may allow us in part to answer this need to find accurate and intelligible empirical laws in giant data sets to best capitalize on the community's observational investments.

While SR has been extensively applied in scientific research, its focus has largely been on single data set analysis, overlooking the rich potential in examining multiple data sets linked to a singular physical phenomenon. The present article extends our Physical Symbolic Optimization framework (Φ-SO; presented in Tenachi et al. 2023a, 2023b) further by allowing the search for a functional form that can simultaneously fit several data sets at once, each realization having (possibly) different fitting parameters. This opens up the new possibility of implementing a functional search on the properties of a class of objects. This approach is relevant across various natural sciences, but it particularly shines in astrophysics, where multiple observations of a single phenomenon are often available, providing a rich multi–data set setup enabling us to devise "universal" laws that apply to a range of celestial objects of interest.

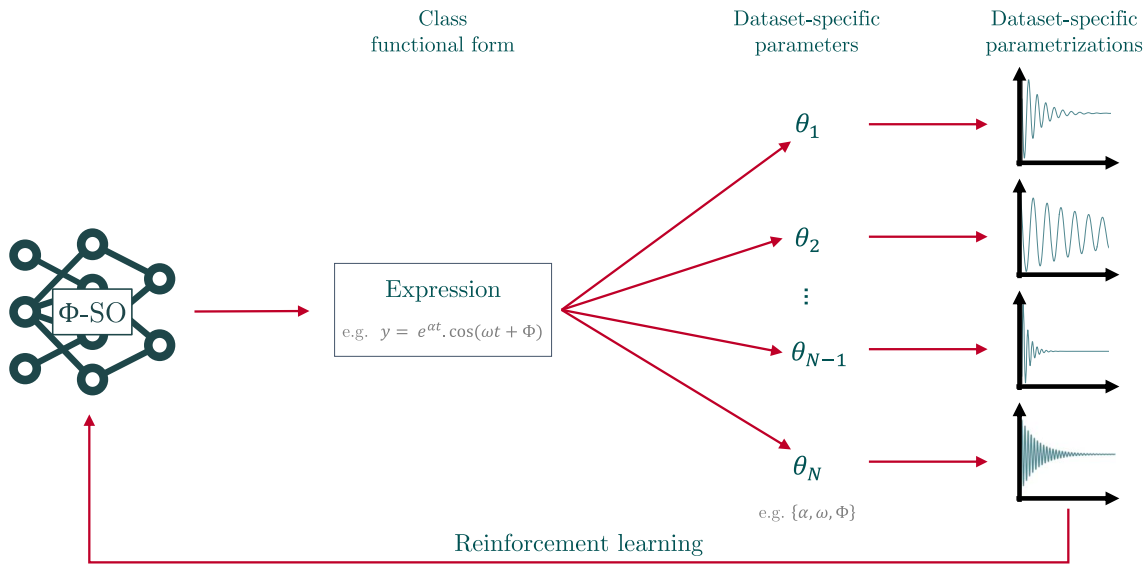


**Figure 1.** Class symbolic regression framework sketch: searching for a unique functional form simultaneously fitting multiple data sets. The process starts at the left-hand side; a batch of trial class analytical expressions are generated using our Φ-SO framework (Tenachi et al. 2023a). The free parameters appearing in those expressions are then optimized in a data set–specific manner i.e., allowing each data set to have its own unique associated values for each parameter. The neural network used to generate the trial expressions is then reinforced based on the fit quality of the trial symbolic functions. This process is repeated until convergence.

In particular, we apply this new framework to the recovery of a milky Way–like analytic Galactic potential from simulated orbits that can be inferred from stellar streams. Specifically, our approach recovers a single analytical form for the energy of stellar stream members, incorporating a "universal" term that encapsulates the dark matter distribution alongside a nuisance term that accounts for the specifics of individual streams—containing parameters allowed to have object-specific values. Unlike traditional black-box deep learning methods, such as autoencoders, our method generates a physically meaningful, low-dimensional model in the form of an analytical model.

The layout of the paper is as follows: in Section 2, we present the methodology of our approach. In Section 3 details the first benchmark for Class SR, consisting of a series of physics problems designed to assess the performance of Class SR systems; here, we also evaluate our method against these benchmarks. In Section 4, we illustrate the practical application of our method in the more complex scenario of a Milky Way–like potential recovery from orbits. Finally, Section 5, offers a discussion and a conclusion.

## 2. Method

We build our "Class Symbolic Regression" (Class SR) framework on the Φ-SO framework for SR. This framework combines deep reinforcement learning with in situ dimensional analysis constraints to construct solutions that avoid physically nonsensical combinations of units. This algorithm currently achieves state-of-the-art performance on physics data sets, and significantly outperforms competitors on the standard Feynman SR benchmark (La Cava et al. 2021) in exact symbolic recovery in the presence of even slight levels of noise (exceeding 0.1%).

Figure 1 gives an overview of our Class SR framework. Using Φ-SO we generate a batch of analytical expressions via a recurrent neural network (RNN). In these expressions, class parameters ($\boldsymbol{c}$)—which are shared across the entire class and have consistent values across all data sets—can appear alongside realization-specific parameters ($\boldsymbol{k}$). Subsequently, we optimize the free parameters appearing in each expression ($\boldsymbol{c}$, $\boldsymbol{k}$), assigning unique values to realization-specific parameters $\{k_i\}_{i<N_r}$ for each of the $N_r$ data sets.

This optimization is conducted using the L-BFGS nonlinear optimization routine (Zhu et al. 1997). Encoding our mathematical symbols with `PyTorch` (Paszke et al. 2019) enables us to use `PyTorch`'s implementation of the L-BFGS routine, which benefits from `PyTorch`'s autodifferentiation capabilities to efficiently and simultaneously optimize both class and realization-specific parameters employing a mean squared error (MSE) cost function: $\mathrm{MSE} = \frac{1}{N_r \cdot \sum_{i=1}^{N_r} N(i)} \sum_{i=1}^{N_r} \sum_{j=1}^{N(i)} (y_{ij} - f(\boldsymbol{c}, \boldsymbol{k}_i, \boldsymbol{x}_{ij}))^2$, where $\boldsymbol{x}_{ij}$ are the input variables, $y_{ij}$ are the target values, and $N(i)$ is the number of samples, which depends on the data set.

We then use reinforcement learning to update the RNN's parameters following a risk-seeking gradient policy (Petersen et al. 2021a), as detailed in Tenachi et al. (2023a). This update is based on a reward $R = (1 + \mathrm{NRMSE})^{-1}$ that is representative of the fit quality of the trial functional form $f$ across all data sets—evaluated using a normalized root mean squared error (NRMSE): $\mathrm{NRMSE} = \frac{1}{\sigma_y} \sqrt{\frac{1}{N_r \cdot \sum_{i=1}^{N_r} N(i)} \sum_{i=1}^{N_r} \sum_{j=1}^{N(i)} (y_{ij} - f(\boldsymbol{c}, \boldsymbol{k}_i, \boldsymbol{x}_{ij}))^2}$, where $\sigma_y$ is the standard deviation of target values evaluated across all data sets. We repeat this process until the RNN converges to a unique high-quality expression and its associated parameter values simultaneously fitting all data sets.

Furthermore, the sequential nature of expression generation in our Φ-SO framework enables the incorporation of various priors regarding the resulting expressions as demonstrated in Tenachi et al. 2023a, Bartlett et al. 2023a, Petersen et al. 2021b, and Kim et al. 2021. This allows for customized constraints on the generated expressions such as adherence to the rules of dimensional analysis (which was one of the focal points of our previous study, Tenachi et al. 2023a) but also simpler priors such as constraints on the number of occurrences of given parameters, the length of the expression, and more.

## 3. Multi–Data Set SR Challenges

Despite existing research efforts to establish benchmarks for SR (La Cava et al. 2021; Matsubara et al. 2022; Marinescu et al. 2023;

**Table 1**
Class Symbolic Regression Challenges

| # | Challenge | Formula | Variables | Realization-specific Free Parameters |
|---|---|---|---|---|
| 1 | Harmonic Oscillator | $A\cos(\Phi + \omega t)$ | $t \in [0.0, 9.4]$ … | $A \in [0.6, 1.2]$<br>$\omega \in [0.2, 1.5]$<br>$\Phi \in [0.9, 1.1]$ |
| 2 | Radioactive Decay | $n_0 e^{\frac{-t}{T}}$ | $t \in [0.5, 6.0]$ … | $n_0 \in [0.4, 2.0]$<br>$T \in [0.9, 1.4]$ |
| 3 | Free Fall | $\frac{1}{2}9.81t^2 + tv_0 + z_0$ | $t \in [0.0, 1.0]$ … | $v_0 \in [-2.0, 8.0]$<br>$z_0 \in [-3.0, 3.0]$ |
| 4 | Damped Harmonic Oscillator A | $e^{-kt}\cos(\Phi + 1.389t)$ | $t \in [0.0, 9.4]$ … | $k \in [0.2, 1.0]$<br>$\Phi \in [-0.2, 0.3]$ |
| 5 | Damped Harmonic Oscillator B | $e^{-0.345t}\cos(\Phi + \omega t)$ | $t \in [0.0, 9.4]$ … | $\omega \in [0.6, 1.4]$<br>$\Phi \in [-0.2, 0.3]$ |
| 6 | Black Body Photon Count | $\frac{1}{e^{5.9\nu/T} - 1}$ | $\nu \in [1.0, 5.0]$ … | $T \in [1.0, 5.0]$ … |
| 7 | Ideal Gas Law | $\frac{n8.314T}{V}$ | $T \in [1.0, 5.0]$<br>$V \in [1.0, 5.0]$ | $n \in [1.0, 5.0]$ … |
| 8 | Free Fall Terminal Velocity | $\sqrt{\frac{2m9.807}{0.47A\rho}}$ | $m \in [1.0, 10.0]$<br>$A \in [1.0, 5.0]$ | $\rho \in [1.0, 6.0]$ … |

**Note.** Each row details a distinct challenge, with the objective being the exact symbolic recovery of the designated target expression using multiple synthetic data sets. Each data set being generated using unique realization-specific parameter sets sampled from the given parameter ranges by sampling from the target expression within the given variable ranges.

Graham et al. 2013), a benchmark tailored specifically for Class SR has yet to be developed, reflecting the innovative nature of this approach. To address this, we introduce our own Class SR challenges, designed to evaluate a system's ability to analyze multiple data sets. These data sets represent varied observations of a similar phenomenon occurring at different scales but governed by a consistent functional form. Table 1 outlines these challenges, each focusing on accurately recovering the symbolic expression from synthetic data sets having varied scale parameter values. To heighten the challenge, we include multiple scenarios incorporating class parameters that are common to all realizations in addition to other realization-specific parameters.

We evaluate our algorithm by randomly sampling 10 data sets of $10^2$ samples for each of the eight challenges described in Table 1 and allowing a maximum of 200,000 expressions to be explored during each run. In order to ensure robustness, for each challenges, this procedure was repeated five times, each with a unique random seed, and the recovery rates were subsequently averaged. The whole benchmark tests were conducted across four noise levels: 0%, 0.1%, 1%, and 10% with noise being added individually to each data set as per the `SRBench` (La Cava et al. 2021) standardized SR benchmarking protocol : $y_{\text{noise}} = y + \epsilon, \quad \epsilon \sim \mathcal{N}\left(0, \gamma\sqrt{\frac{1}{N}\sum_i y_i^2}\right)$, where $\gamma$ is the level of noise. We conduct runs having access to a single data set (SR) and having access to all 10 data sets (Class SR), leading to the total evaluation of 64,000,000 expressions through 320 runs.

We run our algorithm using the hyperparameters detailed in Tenachi et al. (2023a), but with dimensional analysis disabled to ensure a fair comparison with other algorithms (as a consequence the batch size is lowered to 2000). This adjustment allows future comparisons with our system to be focused solely on the machine-learning technique used (here reinforcement learning), rather than the problem simplification achieved through dimensional analysis. We allow the use of the following operations: $\{+, -, \times, /, 1/\,\square, \sqrt{\square}, \square^2, -\,\square, \exp, \log, \cos, \sin\}$, a constant equal to one $\{1\}$, two adjustable realization-specific free constants $\boldsymbol{k} = \{k_1, k_2\}$ allowed to have data set–specific values and one adjustable class free constant $\boldsymbol{c} = \{c_1\}$. The recovery rate is evaluated by examining each expression in the Pareto front, which contains optimum expressions found in conciseness/accuracy, i.e., best-fitting expressions at each level of complexity generated by our algorithm. Successful recovery is noted if an expression on the Pareto front is symbolically equivalent to the target expression. Exact symbolic recovery is assessed by formally comparing these expressions with the target expression using the `SymPy` library for symbolic mathematics (Meurer et al. 2017), following a methodology similar to the one in the `SRBench` (La Cava et al. 2021). Specifically, expressions are deemed equivalent if their fraction is symbolically equivalent to 1 or a constant or if their difference is symbolically equivalent to 0 or a constant.

Figure 2 presents a comparison of exact symbolic recovery rates between our Class SR framework and the traditional SR approach under both noiseless and noisy conditions using an `SRBench`-style benchmarking pipeline, with detailed challenge-by-challenge results published online (see Section 5). Our results demonstrate the superiority of Class SR over traditional SR in exact symbolic recovery, particularly in noisy scenarios where noise overfitting is generally an important concern (La Cava et al. 2021).

While one might consider employing traditional SR individually on each data set and subsequently aggregating the results, this approach would not only be substantially more computationally demanding, but it would also fail to differentiate class constants from realization-specific scale parameters, thus yielding a less interpretable model. Furthermore, our analysis uncovers several instances where traditional SR did not successfully identify the correct expression in any of the five attempts but in which Class SR effectively

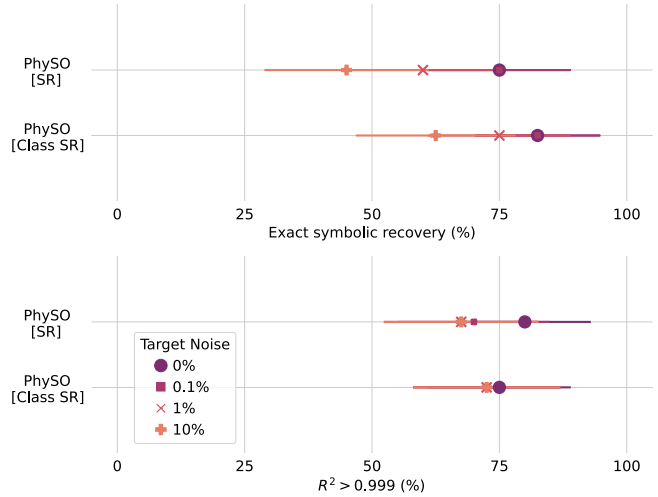


**Figure 2.** Comparison of exact symbolic recovery rates and rate of accurate expressions (having $R^2 > 0.999$) between Class SR and standard SR on our Class SR challenges using an `SRBench`-style benchmarking pipeline (La Cava et al. 2021). This figure demonstrates the enhanced effectiveness of Class SR in identifying common underlying functions across multiple data sets with varying scale parameters, resulting in a higher success rate compared to the traditional SR method exploiting only one data set at a time—especially in the presence of noise.

discovered the correct expressions. This concerns Problem #3 and #6 at 10% noise level scenarios, as well as Problem #5 across all noise levels. These findings highlight the superior robustness and efficiency of Class SR over traditional methods.

Following the `SRBench` protocol, we also include, on Figure 2, the rate of accurate expressions (having $R^2 > 0.999$) with the $R^2$ metric defined as $R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - f(\mathbf{x}_i))^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$. We evaluate fit quality by refitting all constants of candidate expressions on newly generated previously unseen test data sets. This approach ensures a fair comparison between Class SR expressions, whose numerical parameters must accommodate multiple observations, and expressions derived from traditional SR, which only fit a single observation. Our results demonstrate that Class SR is not only more efficient at recovering the exact expressions but also more effective at deriving accurate approximations than traditional SR, in scenarios with noise levels exceeding 0.1%.

## 4. Recovering an Analytic Potential from Stellar Streams

We now turn to an astrophysical application of the algorithm: to try to find the underlying potential of a gravitational system from a set of orbit segments within it. This could be practically applicable for finding an analytic potential model of a galaxy from a set of stellar streams. These linear structures form from the tidal dissolution of globular clusters and dwarf satellite galaxies. When their progenitors are of low mass, the escaping stars have similar energy to the progenitor, and therefore follow a similar orbit. Hence stellar streams approximate orbits in the host galaxy. As has recently been shown by Ibata et al. (2024), for many real streams one can calculate a "correction function" to convert an orbit model into a stream track, and these functions are relatively insensitive to the adopted potential. This procedure can be inverted to give the orbit from the stream.

For this test we imagine having access to full six-dimensional phase-space measurements of a sample of streams. For each structure $i$, the kinetic energy per unit mass $E_{i,\mathrm{kin}}(\boldsymbol{x})$ is simply

$$\frac{1}{2}\boldsymbol{v}^2 = E_{\mathrm{t}}^{i} - \Phi(\boldsymbol{x}). \tag{1}$$

The total energy per unit mass $E_{\mathrm{t}}^{i}$, which is constant, but different, for each stream, can be considered to be nuisance terms in our search for the underlying potential $\Phi$.

We run our algorithm with the objective of recovering the analytic form for $E_{i,\mathrm{kin}}(\boldsymbol{x})$. We use the the hyperparameters detailed in Tenachi et al. (2023a), allowing the use of the following operations: $\{+, -, \times, /, 1/\,\square, \sqrt{\square}, \square^2, -\square, \exp, \log,\}$, a constant equal to one $\{1\}$, one adjustable realization-specific free constant (having units of energy), and three adjustable class free constants (one having units of energy, one having length units, and the other having dimensionless units).

Again we conduct runs at four noise levels (0%, 0.1%, 1%, and 10%), having access to a single orbit (SR), 25% of the orbits, 50% of the orbits, and 100% of the orbits (Class SR), repeating experiments 16 times with different random seeds and allowing a maximum of 250,000 expressions to be explored during each run, leading to the total evaluation of 64,000,000 expressions through 256 runs.

For the present analysis we generated a sample of artificial orbit data (shown in Figure 3) that approximates the sample of 29 thin and long streams studied by Ibata et al. (2024). To this end we used the present day progenitor positions estimated by Ibata et al. (2024), and integrated orbits within a universal Navarro–Frenk–White (NFW) dark matter halo model (Navarro et al. 1997) that very roughly approximates the large-scale mass distribution in the Milky Way. The adopted potential (Łokas & Mamon 2001) is

$$\Phi_{\mathrm{NFW}} = -M_{200}\, g \log[1 + r/R](R/r), \tag{2}$$

where $M_{200}$ is the virial mass of the halo, $g \equiv (\ln(1+c) - c/(1+c))^{-1}$ is a function of the halo concentration $c$ and $R$ is the scale radius. We chose $M_{200} = 10^{12}\, M_{\odot}$, $c = 10$, and $R = 20.0$ kpc. The orbits consist of 100 phase-space points at locations between $\pm 1$ Gyr from the current progenitor location.

Figure 4 presents the results of our analysis in terms of exact symbolic recovery and fit quality, evaluated using the $R^2$ metric. This metric was determined by refitting candidate expressions on noiseless test data and computing the median across various random seeds.

As anticipated, our results underscore that utilizing more realizations during the SR process significantly enhances model accuracy and the likelihood of exact symbolic recovery. This trend is particularly evident as noise levels rise, reinforcing our findings of Section 3. Notably, at a 1% noise level, none of the 16 runs that analyzed stellar stream individually succeeded in recovering the correct functional form. In contrast, when all 29 stellar streams were utilized, the correct functional form was identified nearly half of the time, showcasing the advantages of Class SR under noisy conditions.

We observe that the inability of our algorithm to recover the exact symbolic expression in the presence of 10% noise can be attributed to the fact that, under such high noise conditions, the difference in fit quality between the expressions typically identified by our algorithm and the true solution yields only a minimal improvement in terms of reward, $\Delta R \sim 10^{-5}$. This minute improvement, which corresponds to a difference in $R^2$

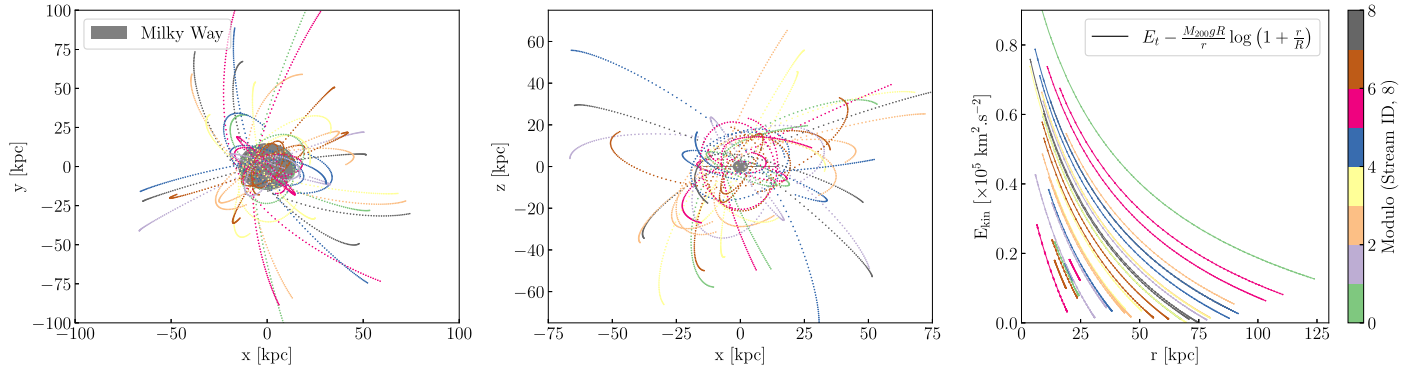


**Figure 3.** Synthetic stellar stream data utilized by our algorithm to recover the Galactic potential. The left and middle panels display the spatial positions of stream members relative to the Milky Way, while the right panel illustrates the kinetic energy of these members as a function of their radial distance from the Galactic center.

of approximately ($10^{-6}$), is the sole metric available to guide the algorithm, as it operates on a trial-and-error basis. Unfortunately, such a small difference often remains undetected due to it falling below the tolerance threshold of the free constants optimization procedure. This scenario highlights a known intrinsic limitation of purely empirical SR, where degeneracies in the space of functional forms can go undetected.

Excluding scenarios where noise levels render the numerically found expression indistinguishable from the true solution, our Class SR algorithm typically converges toward the correct functional form by exploring under 250,000 expressions, despite the presence of multiple alternative functional forms that provide a near-perfect fit to individual streams. Φ-SO identifies an offset parameter specific to each stream (corresponding to $E_t^i$) and a functional form parameterized by class parameters common to all streams corresponding to $\Phi_{\rm NFW}$. These results show that our algorithm can effectively recover a concise interpretable model for a Milky Way–like potential in the form of an analytic expression based solely on stellar positions and velocities without any prior information about the system.

## 5. Discussion and Conclusions

We presented the first framework for discovering symbolic analytical functions that simultaneously fit multiple data sets by allowing for (possibly) unique data set–specific parameter values. This new framework, which we dub "Class Symbolic Regression" is built upon our earlier Φ-SO framework, which already delivers state-of-the-art performances in symbolic recovery in the presence of noise.

We demonstrated the efficacy of Class SR through simple textbook physics examples that we compiled into a first Class SR benchmark, finding better performance in exact symbolic recovery over traditional SR, especially in noisy situations. Additionally, we applied our method to a more complex astrophysical scenario, successfully rediscovering an NFW galaxy potential model from orbits approximating stellar streams.

Regular SR, when applied to a single data set, often risks overfitting to specific characteristics of an observation, such as observational biases or transient events, and noise. In contrast, our Class SR framework should facilitate the finding of universal analytical laws that apply to a range of observations within a single class of physical phenomena. This enables our framework to model the underlying physics rather than the specifics of individual observations, with data set–specific free parameters modeling the unique aspects of each observation. For instance, an application within galactic dynamics that we intend to explore in a future contribution is the analysis of galactic rotation curves. Here, a universal law derived through Class SR could provide insights into the general behavior of dark matter, whereas traditional SR, if applied to a single galaxy, might merely find the specific attributes of that galaxy.

It should be noted that while Class SR might superficially resemble regular SR applied to unbalanced data sets with data set–specific parameters being akin to additional input variables, this comparison is not entirely accurate. In Class SR, these additional degrees of freedom represent unknown values that must be determined, differentiating it as a distinct problem with its own unique challenges.

A persistent issue in SR is model selection as the correct expression can often be overlooked in favor of those that fit better or are less complex (these concerns led to, e.g., the development of single objective criterion; Bartlett et al. 2023b). Our framework, by searching for expressions that fit multiple data sets, effectively utilizes information about the physical phenomena's class structure. This approach significantly mitigates model selection challenges, helping avoid incorrect model choices influenced by data set–specific peculiarities. In addition, exploiting multiple data sets with regular SR techniques would require fitting the individual data sets independently, and then identifying the solutions in common between the objects, which may not be possible if the measurements are uncertain, would be computationally inefficient and would result in lower performances in exact symbolic recovery and fit quality alike in the presence of noise.

Finally, we note that after the first submission of our paper, another Class SR approach built on `Operon` (Kommenda et al. 2020)—a genetic algorithm approach to SR—was applied to supernovae photometry in Russeil et al. (2024).

In future work, we intend to improve on the machine-learning aspects of our method to more effectively leverage multiple data sets. As each data set might distinctly highlight certain symbolic terms or subexpressions more prominently than others, a promising strategy could be to periodically shift the neural network's training emphasis between data sets. This technique could potentially refine the performance of Class SR by sequentially learning different segments of the expression, rather than attempting to learn the entire expression simultaneously, thereby facilitating the learning process.

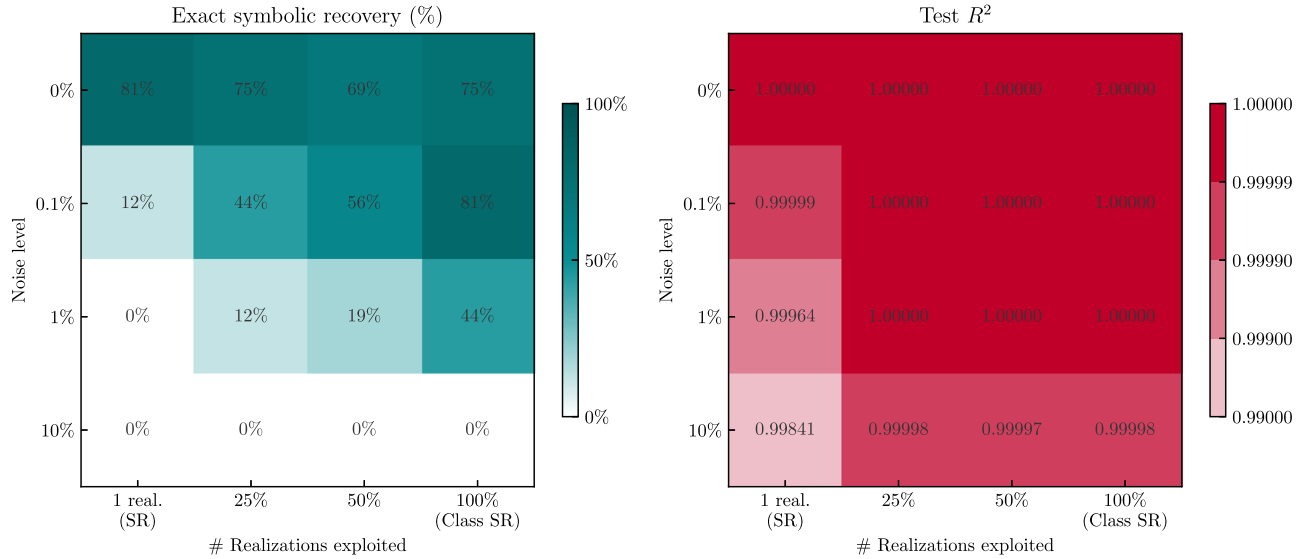


**Figure 4.** This figure presents the exact symbolic recovery rate and median $R^2$ achieved by our Class SR algorithm in the task of recovering an NFW dark matter halo model (Navarro et al. 1997) from synthetic data sets of stellar stream positions and velocities. The performance metrics are displayed as functions of noise levels and the number of realizations exploited. The edge case, in which a single realization is used, corresponds to the conditions of traditional SR. The results distinctly demonstrate that Class SR substantially outperforms traditional SR, particularly in noisy environments.


## Acknowledgments

R.I. acknowledges funding from the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation program (grant agreement No. 834148). The authors would like to acknowledge the High Performance Computing Center of the University of Strasbourg for supporting this work by providing scientific support and access to computing resources. Part of the computing resources were funded by the Equipex Equip@Meso project (Programme Investissements d'Avenir) and the CPER Alsacalcul/Big Data.


## Code Availability

The documented code for the Φ-SO algorithm, along with demonstration notebooks, benchmark and results analysis pipelines is accessible on GitHub at https://github.com/WassimTenachi/PhySO, complete with comprehensive documentation. A frozen version related to this work on Class Symbolic Regression is released under tag v1.1.0 (https://github.com/WassimTenachi/PhySO/releases/tag/v1.1.0) and deposited on Zenodo (doi:10.5281/zenodo.11663147; Tenachi et al. 2024).

We offer to the community a convenient interface for using our Class SR benchmark, running: `pb = ClassProblem(i)` will instantiate challenge `i` ∈{0, 1,...,7} of the Class benchmark presented in Table 1. This interface offers simple ways to generate data points (via `pb.generate_data_points`) and compare a canddiate expression to the target (via `pb.get_sympy`).

In addition, we include challenge-by-challenge and run-by-run performances results tables: see https://github.com/WassimTenachi/PhySO/tree/v1.1.0/benchmarking/ClassBenchmark/results for results pertaining to the Class SR benchmark and https://github.com/WassimTenachi/PhySO/tree/v1.1.0/demos/class_sr/demo_milky_way_streams/results for results pertaining to the stellar stream problem.

Finally, for the sake of result reproducibility, we offer a straightforward method to replicate the outcomes presented in Figure 2 by simply executing the following command: `python classbench_run.py --equation i --noise n --n_reals Nr`. This command will run `PhySO` on challenge number `i` ∈{0, 1,...,7} of the Class benchmark presented in Table 1, employing a noise level of `n` ∈[0, 1] and exploiting `Nr` ∈ℕ realizations. We also include the script we used to estimate performances post-run : `classbench_results_analysis.py`

Similarly, we offer a straightforward method to replicate the outcomes presented in Figure 3 by simply executing the following command: `python MW_streams_run.py --noise n --frac_real fr`. This command will run `PhySO` on the stellar stream problem described in Section 4, employing a noise level of `n` ∈[0, 1] and exploiting a fraction of `fr` ∈[0, 1] realizations. Again, we include our results analysis script: `MW_streams_results_analysis.py`

## ORCID iDs


Wassim Tenachi https://orcid.org/0000-0001-8392-3836
Rodrigo Ibata https://orcid.org/0000-0002-3292-9709
Thibaut L. François https://orcid.org/0009-0001-0314-7038
Foivos I. Diakogiannis https://orcid.org/0000-0002-8788-8174


## References


Angelis, D., Sofos, F., & Karakasidis, T. E. 2023, Arch. Comput. Methods Eng., 30, 3845
Bartlett, D., Desmond, H., & Ferreira, P. 2023a, Proc. of the Companion Conf. on Genetic and Evolutionary Computation, GECCO '23 Companion (New York: Association for Computing Machinery), 2402
Bartlett, D. J., Desmond, H., & Ferreira, P. G. 2023b, IEEE Transactions on Evolutionary Computation, 109, 083524
Biggio, L., Bendinelli, T., Lucchi, A., & Parascandolo, G. 2020, in Learning Meets Combinatorial Algorithms at NeurIPS2020, https://openreview.net/pdf?id=W7jCKuyPn1
Biggio, L., Bendinelli, T., Neitz, A., Lucchi, A., & Parascandolo, G. 2021, in Proc. of the 38th Int. Conf. on Machine Learning, 139, ed. M. Meila & T. Zhang (New York: PMLR), 936, https://proceedings.mlr.press/v139/biggio21a.html
Brence, J., Todorovski, L., & Džeroski, S. 2021, KBS, 224, 107077
Brunton, S. L., Proctor, J. L., & Kutz, J. N. 2016, PNAS, 113, 3932
Chen, T., Xu, P., & Zheng, H. 2024, arXiv:2401.09748

Cheng, S., & Alkhalifah, T. 2024, arXiv:2404.17971
Faris, J. G., Hayes, C. F., Goncalves, A. R., et al. 2024, in The Sixteenth Workshop on Adaptive and Learning Agents, Pareto Front Training For Multi-Objective Symbolic Optimization, https://openreview.net/forum?id=e0gswuNjcb
Fiorini, C., Flint, C., Fostier, L., et al. 2024, arXiv:2404.15742
Graham, M. J., Djorgovski, S., Mahabal, A. A., Donalek, C., & Drake, A. J. 2013, MNRAS, 431, 2371
He, Y., Sheng, B., & Li, Z. 2024, Appl. Sci., 14, 2929
Holt, S., Qian, Z., & van der Schaar, M. 2023, in The Eleventh Int. Conf. on Learning Representations, Deep Generative Symbolic Regression, https://openreview.net/forum?id=o7koEEMA1bR
Ibata, R., Malhan, K., Tenachi, W., et al. 2024, ApJ, 967, 89
Jin, Y., Fu, W., Kang, J., Guo, J., & Guo, J. 2019, arXiv:1910.08892
Kamienny, P., Lample, G., Lamprier, S., & Virgolin, M. 2023, in Proc. of Machine Learning Research, Deep Generative Symbolic Regression with Monte-Carlo-Tree-Search, 202, ed. A. Krause et al. (New York: PMLR), 15655, https://proceedings.mlr.press/v202/kamienny23a.html
Kamienny, P.-A., d'Ascoli, S., Lample, G., & Charton, F. 2022, in Advances in Neural Information Processing Systems, End-to-end Symbolic Regression with Transformers, ed. A. Oh, https://openreview.net/forum?id=GoOuIrDHG_Y
Kammerer, L., Kronberger, G., Burlacu, B., et al. 2020, Genetic Programming Theory and Practice XVII (Berlin: Springer), 79
Kepler, J. 1609, Astronomia Nova (Pragae)
Kim, J. T., Landajuela, M., & Petersen, B. K. 2021, in 1st Mathematical Reasoning in General Artificial Intelligence, Int. Conf. on Learning Representations (ICLR), https://www.osti.gov/servlets/purl/1782518
Kommenda, M., Burlacu, B., Kronberger, G., & Affenzeller, M. 2020, Genet. Program. Evolvable Mach., 21, 471
La Cava, W., Orzechowski, P., Burlacu, B., et al. 2021, in Proc. of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, Contemporary Symbolic Regression Methods and their Relative Performance, ed. J. Vanschoren & S. Yeung, https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/c0c7c76d30bd3dcaefc96f40275bdc0a-Paper-round1.pdf
Landajuela, M., Lee, C. S., Yang, J., et al. 2022, in Advances in Neural Information Processing Systems, A Unified Framework for Deep Symbolic Regression, 35, ed. S. Koyejo (New York: Curran Associates, Inc.), 33985, https://proceedings.neurips.cc/paper_files/paper/2022/file/dbca58f35bddc6e4003b2dd80e42f838-Paper-Conference.pdf
Li, Y., Li, W., Yu, L., et al. 2024a, arXiv:2401.14424
Li, Y., Liu, J., Li, W., et al. 2024b, arXiv:2402.18603
Łokas, E. L., & Mamon, G. A. 2001, MNRAS, 321, 155
Luo, C., Chen, C., & Jiang, Z. 2022, Int. J. Comput. Methods, 19, 2142002
Makke, N., & Chawla, S. 2022, arXiv:2211.10873
Marinescu, I., Strittmatter, Y., Williams, C., & Musslick, S. 2023, in NeurIPS 2023 AI for Science Workshop, https://openreview.net/forum?id=i3PecpoiPG
Martius, G., & Lampert, C. H. 2017, in Int. Conf. on Learning Representations, Extrapolation and learning equations, https://openreview.net/forum?id=BkgRp0FYe
Matsubara, Y., Chiba, N., Igarashi, R., & Ushiku, Y. 2022, in NeurIPS 2022 AI for Science: Progress and Promises, SRSD: Rethinking Datasets of Symbolic Regression for Scientific Discovery, https://openreview.net/forum?id=oKwyEqClqkb
Meidani, K., Shojaee, P., Reddy, C. K., & Farimani, A. B. 2024, in The Twelfth Int. Conf. on Learning Representations, SNIP: Bridging Mathematical Symbolic and Numeric Realms with Unified Pre-training, https://openreview.net/forum?id=KZSEgJGPxu
Melching, D., Paysan, F., Strohmann, T., & Breitbarth, E. 2024, arXiv:2403.10320
Meurer, A., Smith, C. P., Paprocki, M., et al. 2017, PeerJ Computer Science, 3, e103
Michishita, Y. 2024, arXiv:2311.12713
Navarro, J. F., Frenk, C. S., & White, S. D. M. 1997, ApJ, 490, 493
Newton, I. 1687, Philosophiae Naturalis Principia Mathematica (London: Royal Society)
Paszke, A., Gross, S., Massa, F., et al. 2019, in Advances in Neural Information Processing Systems, 32, ed. H. Wallach et al. (New York: Curran Associates, Inc.), https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf
Petersen, B. K., Larma, M. L., Mundhenk, T. N., et al. 2021a, in Int. Conf. on Learning Representations, Deep symbolic regression: Recovering mathematical expressions from data via risk-seeking policy gradients, https://openreview.net/forum?id=m5Qsh0kBQG
Petersen, B. K., Santiago, C., & Landajuela, M. 2021b, in 8th ICML Workshop on Automated Machine Learning (AutoML), Incorporating domain knowledge into neural-guided search via in situ priors and constraints, https://openreview.net/forum?id=yAis5yB9MQ
Russeil, E., de França, F. O., Malanchev, K., et al. 2024, arXiv:2402.04298
Sahoo, S., Lampert, C., & Martius, G. 2018, in Proc. of Machine Learning Research, Proc. of the Int. Conf. on Machine Learning, 35, ed. J. Dy & A. Kraus (New York: PMLR), 4442
Schmidt, M., & Lipson, H. 2009, Sci, 324, 81
Schmidt, M., & Lipson, H. 2011, Age-Fitness Pareto Optimization (New York: Springer), 129
Scholl, P., Bieker, K., Hauger, H., & Kutyniok, G. 2023, arXiv:2310.05537
Shojaee, P., Meidani, K., Gupta, S., Farimani, A. B., & Reddy, C. K. 2024, arXiv:2404.18400
Sousa, T., Bartlett, D. J., Desmond, H., & Ferreira, P. G. 2024, PhRvD, 109, 083524
Tegmark, M. 2008, FoPh, 38, 101
Tenachi, W., Ibata, R., & Diakogiannis, F. I. 2023a, ApJ, 959, 99
Tenachi, W., Ibata, R., & Diakogiannis, F. I. 2023b, arXiv:2312.03612
Tenachi, W., Ibata, R., François, T. L., & Diakogiannis, F. 2024, PhySO-v1.1.0, Zenodo, doi:10.5281/ZENODO.11663147
Tian, Y., Zhou, W., Dong, H., Kammer, D. S., & Fink, O. 2024, arXiv:2402.05306
Tohme, T., Liu, D., & Youcef-Toumi, K. 2023, TMLR, https://openreview.net/forum?id=lheUXtDNvP
Udrescu, S.-M., Tan, A., Feng, J., et al. 2020, Adv Neural Inf Process Syst, 33, 4860
Udrescu, S.-M., & Tegmark, M. 2020, SciA, 6, 2631
Vastl, M., Kulhánek, J., Kubalík, J., Derner, E., & Babuška, R. 2024, IEEE Access, 12, 37840
Zhang, B., & Lei, J. 2024, arXiv:2405.08656
Zheng, W., Sharan, S., Fan, Z., et al. 2022, arXiv:2212.14849
Zhu, C., Byrd, R. H., Lu, P., & Nocedal, J. 1997, ACM Transactions on Mathematical Software (TOMS), 23, 550