

# Abstract

In a recent paper we presented a new benchmark tool for symbolic regression and tested it with cosmology datasets. This new benchmark tool is easy to use, and we would like to spread the word and encourage other people to try it in their respective field where applicable. Out of the box it has ten different machine learning algorithms, but we also encourage the community to add their own methods to the framework and expand the capabilities. In this talk I will discuss how it works and the paper we published with a test application in cosmology. We find no indication that performance of algorithms on standardized datasets are good indications of performance on cosmological datasets. This suggests that it is not necessarily prudent to choose which symbolic regressions algorithm to use based on their performance on standardized data. Instead, a more prudent approach is to consider a variety of algorithms. Overall, we find that most of the benched algorithms do rather poorly in the benchmark and suggest possible ways to proceed with developing algorithms that will be better at identifying ground truth expressions for cosmological datasets. As part of this publication, we introduce our benchmark algorithm cp3-bench which we make publicly available at <https://github.com/CP3-Origins/cp3-bench>.

# HAMLET-PHYSICS 2024

## Symbolic regression and cp3-bench

University of Southern Denmark

CP3 - Origins



# Motivation

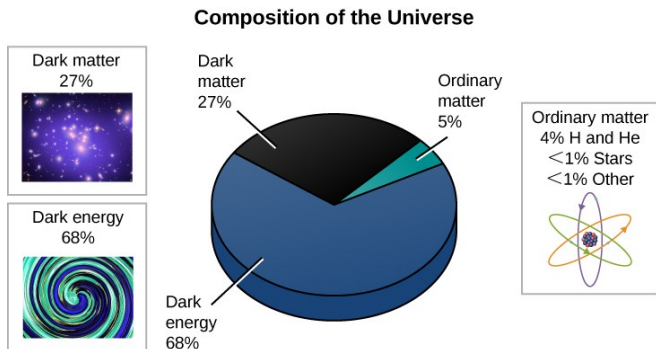
My hope is that today's talk will help inspire us to try new methods to push the bar higher and I will present a new tool to help with symbolic regression. In particle physics and cosmology we have seen great results and progress during the 20th century. However, other than the Higgs discovery in 2012 there has only been limited successes in particle physics. With machine learning and symbolic regression the ultimate goal is to make an AI tool that can discover physics and in this way, help us resolve the remaining mysteries we face.

# Overview

The talk is divided into the following parts:

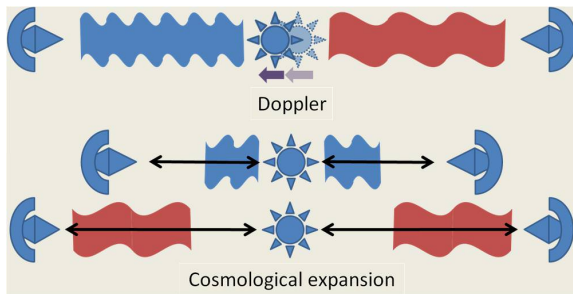
- 1 Problems in cosmology and astroparticle physics
  - Energy composition of the universe
  - Dark energy
  - Dark matter
- 2 cp3-bench
  - Symbolic regression benchmark
  - Results
  - Conclusion
- 3 Questions

# What is the universe made of?



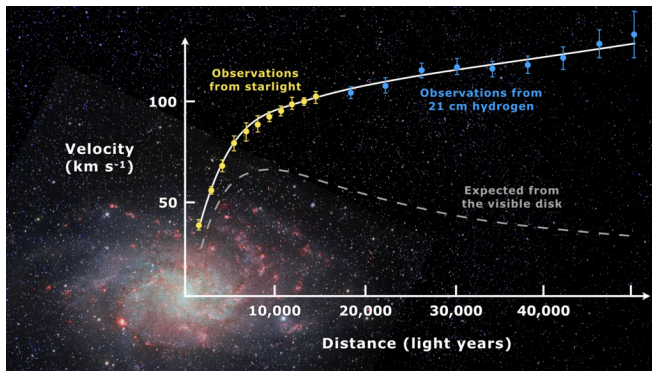
**Figure:** Dark energy is the dominant energy type in the universe, and the second most dominant is dark matter. Credit: OpenStax CNX, CC BY

# Redshift



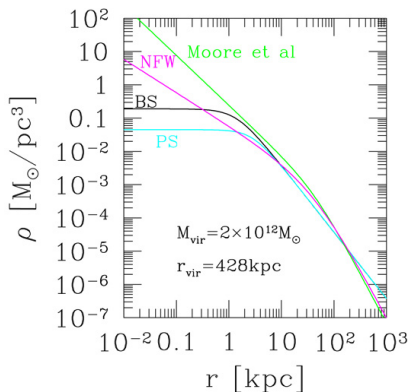
**Figure:** Redshift from Doppler effects and cosmological expansion.  
Credit: Brews Ohare

# Rotation curves



**Figure:** Total rotation from spiral galaxy Messier 33 with expectation from visible matter. The data and the model predictions are from Corbelli and Salucci 2000. Credit: Mario De Leo, CC BY-SA 4.0

# Dark matter density profiles



**Figure:** Comparison of the cusp models, Moore et al and NFW, and the core models, BS and PS. Credit: Suk Sien Tie



# Introducing cp3-bench

## cp3-bench: A tool for benchmarking symbolic regression algorithms tested with cosmology

Mattias E. Thing, Sofie M. Koksang


We present a benchmark study of ten symbolic regression algorithms applied to cosmological datasets. We find that the dimension of the feature space as well as prec no indication that inter-dependence of features in datasets are particularly important, meaning that it is not an issue if datasets e.g. contain both  $z$  and  $H(z)$  as feature indications of performance on cosmological datasets. This suggests that it is not necessarily prudent to choose which symbolic regressions algorithm to use based on tl variety of algorithms. Overall, we find that most of the benched algorithms do rather poorly in the benchmark and suggest possible ways to proceed with developing alg As part of this publication we introduce our benchmark algorithm cp3-bench which we make publicly available at [this https URL](https://github.com/mthing/cp3-bench). The philosophy behind cp3-bench is th easy additions of new algorithms and datasets.

Comments: 45 pages, 26 tables, 2 figures

Subjects: **Instrumentation and Methods for Astrophysics (astro-ph.IM)**; Cosmology and Nongalactic Astrophysics (astro-ph.CO)

Cite as: arXiv:2406.15531 **astro-ph.IM**

(or arXiv:2406.15531v1 **astro-ph.IM** for this version)

<https://doi.org/10.48550/arXiv.2406.15531> 

### Submission history

From: Mattias Ermakov Thing [[view email](#)]

**fv11** Fri, 21 Jun 2024 13:31:47 UTC (149 KB)

# What is it?

cp3-bench is a tool for running multiple algorithms in parallel or individually for a given dataset. This allows users to easily test one dataset against different types of symbolic regression algorithms. It is also easy to add your own custom algorithm. As an output you get a list with the best equation from each algorithm and the MSE value associated with the algorithm.

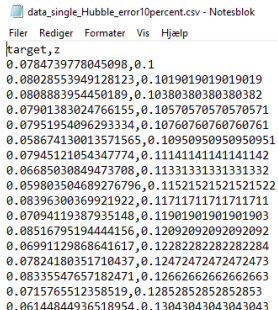
## Supported algorithms

Algorithm Name	Method	Year
AI-Feynman	PI,NN	2020
Deep Symbolic Optimization (DSO)	RNN,RL	2021
Deep Symbolic Regression (DSR)	RNN,RL,GP	2021
Fast Function Extraction (FFX)	GP	2011
Genetic Engine	GP	2022
GPZGD	GP	2020
ITEA	GP	2019
PySR	GP,NN	2023
QLattice Clinical Omics (QLattice)	GP	2022
Unified Deep Symbolic Regression (uDSR)	RNN,RL	2022

**Table:** Physics Inspired = PI, Neural Network = NN, Recurrent Neural Network = RNN, Reinforcement Learning = RL, Genetic Programming = GP.

# Input

Select the path to some csv file with a column called target being the expected output value and any other parameter will be considered the input.



```

data_single_Hubble_error10percent.csv - Notesblok
Filer Rediger Formater Vis Hjælp
target,z
0.0784739778045098,0.1
0.08028553949128123,0.1019019019019019
0.0808883954450189,0.10380380380380382
0.07901383024766155,0.10570570570570571
0.07951954096293334,0.10760760760760761
0.058674130013571565,0.10950950950950951
0.07945121054347774,0.11141141141141142
0.06685030849473708,0.11331331331331332
0.059803504689276796,0.11521521521521522
0.08396300369921922,0.11711711711711711
0.07094119387935148,0.11901901901901903
0.08516795194444156,0.12092092092092092
0.06991129868641617,0.12282282282282284
0.07824180351710437,0.12472472472472473
0.08335547657182471,0.12662662662662663
0.0715765512358519,0.12852852852852853
0.06144844936518954,0.13043043043043043
  
```

Figure: Example of input file

# Architecture

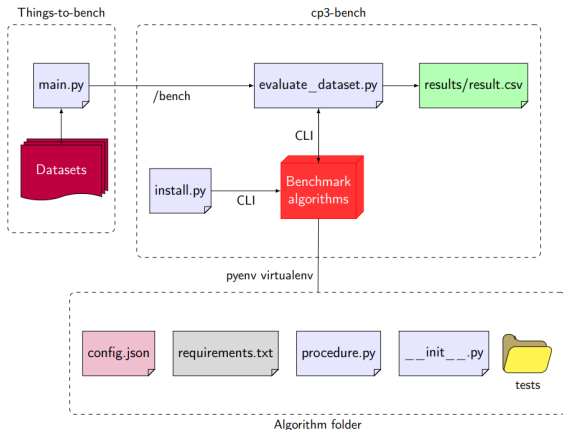


Figure: Architecture for using cp3-bench

# Output

The output contains the name of the algorithm, MSE, best fit equation and runtime. Note the format of the output of the equation is currently not fully standardized.

```
C204 - Notebook
File  Rediger  Formater  Vis  Hjelp
method, mse, equation, run_time
AI-Feynman, 0.00020737999189700002, 0.530059849497+log(cos(cos((pi-sqrt((x0+1))))**(-1))))), 08:53:25
DSO, 0.000251670020405, "

$$\sqrt{x_1 - 3}$$

DSR, 0.00021320278896000002, "

$$\frac{4}{\sqrt{x_1 + 5x_1 + 2\sin(2x_1)}}$$

FFX, 0.00020061633638100002, "0.0030 + 0.0397*x0 + 0.0149*max(0, X0-0.480) + 0.0143*max(0, X0-1.24) + 0.
GeneticEngine, 0.00020888951829300002, log(1 + Abs(0.1**z/(0.001*z - log(Abs(z) + 1)) + 0.0040622776
gozgd, 0.000511403250052, "1.34007e-01+4.30021e-02*((((-1.53168e-01*x[; ; 0]))*(4.78503e+00*x[; ; 0]))**np.
ITFA, 0.00019883940997600002, 0.16115 + 2.10702e-3*(x0**(3)) + -7.82409e-4*Log(x0**(-3)) + 1.36482e-3*lc
PySt, 0.000311869553428400006, 0.07763134846788875*z + 0.07763134846788875*cos(sin(z)), 00:34:39
QLattice, 0.00019738003992200002, 0.16512*tanh(0.56534*z - 1.04296)+0.19878, 01:11:31
OSR, 0.00019544167430700002, "

$$0.576118x_1^5 - 2.07853x_1^4 + 2.60802x_1^3 - 1.30941x_1^2 + 0.278604x_1 + \sin(x_1)$$


$$+ \cos(x_1)$$

| - 0.931898", 02:03:41
```

Figure: Example of output file

# Raw table

Cosmological dataset results										
Dataset	Algorithm									
	AI-Feynman	DSO	DSR	FFX	Genetic Engine	GPZGD	ITEA	PySR	QLattice	uDSR
C1	$10^{-8}$	$10^{-5}$	$10^{-5}$	$10^{-6}$	$10^{-6}$	$10^{-13}$	$10^{-14}$	$10^{-9}$	$10^{-10}$	✓
C2	$10^{-14}$	$10^{-14}$	$10^{-14}$	$10^{-14}$	$10^{-14}$	$10^{-14}$	$10^{-14}$	$10^{-14}$	$10^{-14}$	$10^{-14}$
C3	$10^{-6}$	$10^{-4}$	$10^{-4}$	$10^{-6}$	$10^{-5}$	$10^{-6}$	$10^{-9}$	$10^{-6}$	$10^{-9}$	✓
C4	$10^{-5}$	$10^{-6}$	$10^{-5}$	$10^{-5}$	$10^{-6}$	$10^{-5}$	$10^{-8}$	$10^{-6}$	$10^{-7}$	✓
C5	✓	$10^{-4}$	$10^{-4}$	✓	$10^{-5}$	$\sqrt{(10^{-31})}$	✓	✓	✓	✓
C6	$10^{-3}$	$10^{-4}$	$10^{-4}$	$10^{-6}$	$10^{-5}$	$10^{-7}$	$10^{-7}$	$10^{-5}$	✓	$10^{-8}$
C7	$10^{-6}$	$10^{-4}$	$10^{-5}$	$10^{-4}$	$10^{-6}$	$10^{-10}$	$10^{-13}$	$10^{-7}$	$10^{-11}$	$10^{-13}$
C8	$10^{-4}$	$10^{-3}$	$10^{-4}$	$10^{-4}$	$10^{-4}$	$10^{-7}$	$10^{-10}$	$10^{-6}$	$10^{-9}$	$10^{-11}$
C9	$10^{-3}$	$10^{-3}$	$10^{-3}$	$10^{-4}$	$10^{-4}$	$10^{-5}$	$10^{-6}$	$10^{-3}$	$10^{-6}$	$10^{-7}$
C10	$10^{-3}$	$10^{-2}$	$10^{-2}$	$10^{-3}$	$10^{-3}$	$10^{-3}$	$10^{-5}$	$10^{-2}$	$10^{-5}$	$10^{-6}$
C11	$10^{-3}$	$10^{-3}$	$10^{-3}$	$10^{-4}$	$10^{-3}$	$10^{-4}$	$10^{-7}$	$10^{-3}$	$10^{-6}$	$10^{-3}$
C12	$10^{-4}$	$10^{-2}$	$10^{-3}$	$10^{-3}$	$10^{-2}$	$10^{-1}$	$10^{-6}$	$10^{-3}$	$10^{-6}$	$10^{-2}$
C13	$10^{-3}$	$10^{-2}$	$10^{-3}$	$10^{-4}$	$10^{-4}$	$10^{-3}$	$10^{-5}$	$10^{-4}$	$10^{-6}$	$10^{-2}$
C14	$10^{-4}$	$10^{-3}$	$10^{-3}$	$10^{-4}$	$10^{-4}$	$10^{-5}$	$10^{-2}$	$10^{-3}$	$10^{-6}$	$10^{-3}$
C15	10	10	10	10	10	10	10	10	10	10
C16	1	$10^{-1}$	$10^{-1}$	$10^{-1}$	$10^{-2}$	$10^{-2}$	$10^{-2}$	$10^{-2}$	$10^{-2}$	$10^{-2}$
C17	$10^{-1}$	1	$10^{-1}$	$10^{-1}$	$10^{-2}$	$10^{-2}$	$10^{-2}$	$10^{-2}$	$10^{-2}$	1
C18	1	$10^{-1}$	$10^{-1}$	$10^{-2}$	$10^{-2}$	$10^{-3}$	$10^{-3}$	$10^{-2}$	$10^{-3}$	1
C19	$10^{-3}$	$10^{-4}$	$10^{-3}$	$10^{-3}$	$10^{-4}$	$10^{-4}$	$10^{-4}$	$10^{-4}$	$10^{-4}$	$10^{-2}$

Table: Result of evaluating the cosmological datasets

# Redshift data

We see that the simpler symbolic parameterization performs better even though there are inter-dependence of parameters.

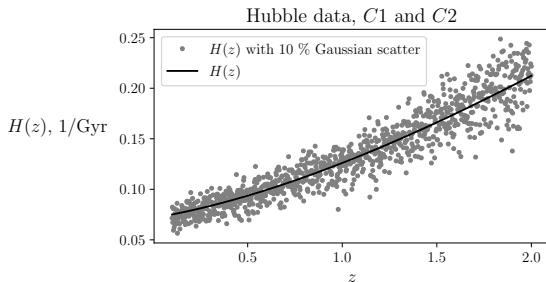
$$\delta z / \delta t_0(z, H(z, \omega)) = 0.0716 / \text{Gyr} \cdot (1 + z) - H(z, \omega) \quad (1)$$

$$\delta z / \delta t_0(z, \omega) = 0.0716 / \text{Gyr} \cdot [(1 + z) - (1 + z)^{3(1+\omega)/2}] \quad (2)$$



# Hubble data

$$H(z) = 0.0716/\text{Gyr} \cdot \sqrt{0.3 \cdot (1+z)^3 + 0.7} \quad (3)$$



# Unknown solution data

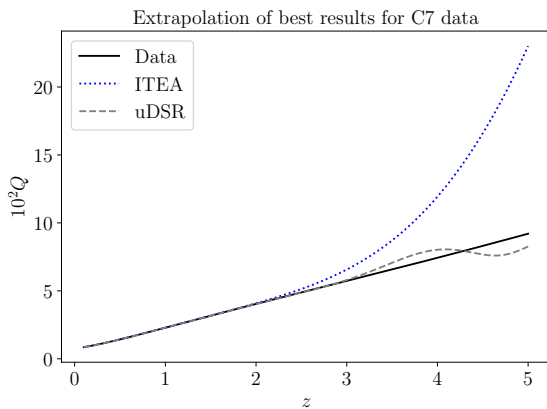


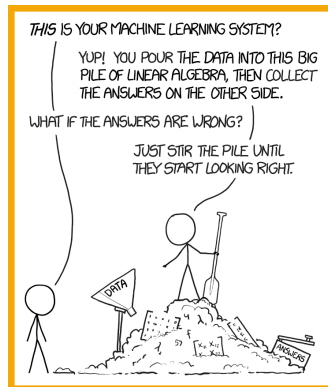
Figure: Attempt at finding unknown solution

# Key points

- A model trained on "standard" datasets doesn't necessarily perform well in a physical scenario.
- Inter-dependence of parameters is less important compared to having a simple symbolic form.
- Noise is a problem.
- Using multiple algorithms hopefully one will find a good result.
- We suggest work should be done on constraining models to not try unrealistic equations.

# TL;DR

- Physics is hard, and we ought not to ignore any possible tool.
- Machine learning is great but it is not magic.
- Mathematics is still the foundation of physics.



# Questions

Any questions from the audience?

Thank you

Thank you