Exhaustive Symbolic Regression: Learning astrophysics directly from data

Harry Desmond[1], Deaglan Bartlett[2], Pedro Ferreira[3]

[1] Institute of Cosmology and Gravitation, University of Portsmouth
[2] Institut Astrophysique de Paris; [3] Department of Physics, University of Oxford

harry.desmond@port.ac.uk

arxiv: 2211.11461, 2301.04368, 2310.16786 **github.com/DeaglanBartlett/ESR**

(Exhaustive) Symbolic Regression and the MDL principle

- **Symbolic Regression** (SR) algorithms learn analytic expressions which fit data accurately and simply in a highly interpretable way.
- We have developed a new SR method – **Exhaustive Symbolic Regression** (ESR) – which efficiently considers **all possible equations** up to some complexity. Unlike other methods, ESR is guaranteed to find the true optimum and complete function ranking.
- ESR represents functions as trees and finds all that form unique functions.

The *minimum description length (MDL) principle* posits that the **best functional representation of a dataset is the one that compresses it most**, so that the fewest units of information are needed to communicate the data with the help of the function. The total amount of information to send is

$$L(D) = L(H) + L(D|H),$$

where L denotes codelength, $L(D|H)$ is an accuracy term and $L(H)$ penalises more complex hypotheses (functions). Lower-$L(D)$ models are those which *i)* fit the data more accurately, *ii)* contain fewer operators and parameters and *iii)* do not need as finely tuned parameter values to fit the data well.

Application 1 – Is cosmic expansion Friedmann?

- Can we derive the law of the Universe's expansion without assuming GR?
- Apply ESR to cosmic chronometers, stellar "standard clocks".
- We find that **ΛCDM is not the best equation for the data!** It does, however, lie in the top 40 of the 5.2 million functions up to complexity 10.

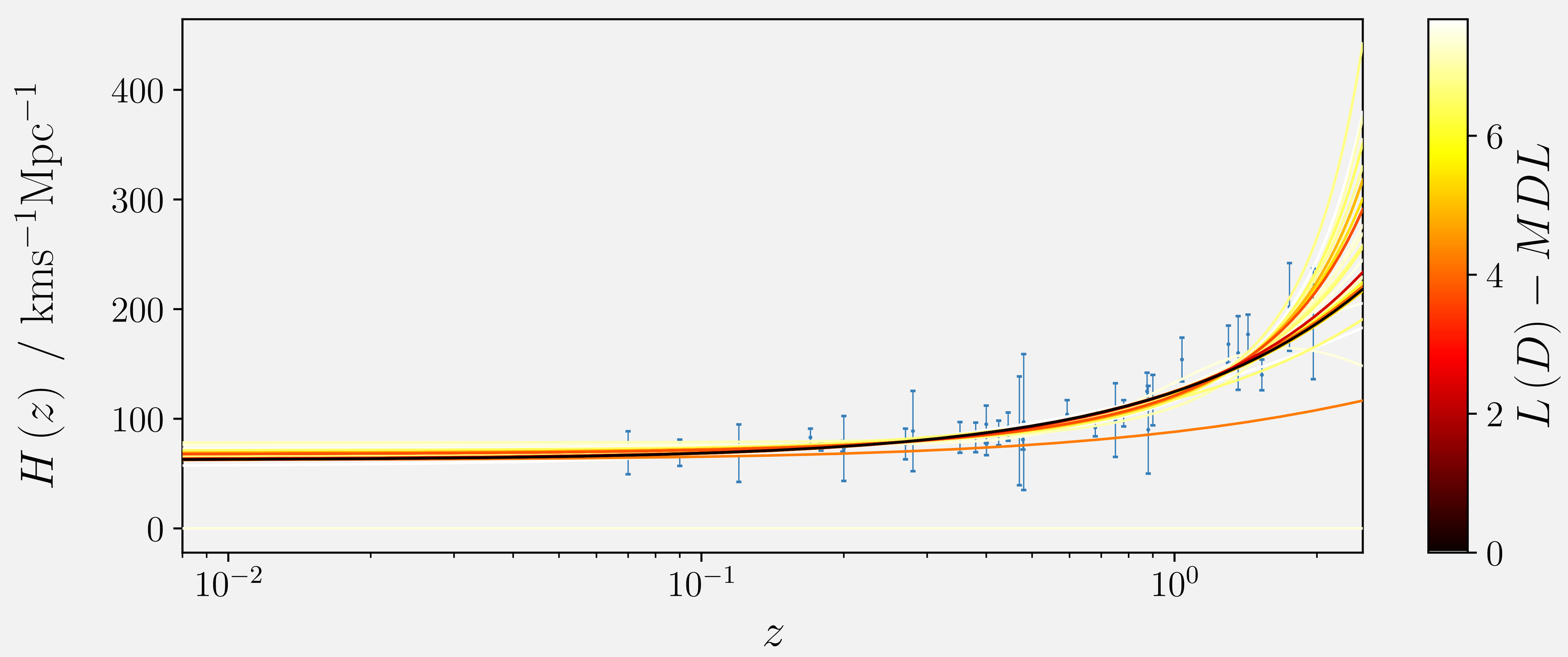

Figure 1: Expansion rate, H, as a function of redshift, z, learned by applying ESR to the cosmic chronometer data (blue points). Functions are colour-coded by their description length.

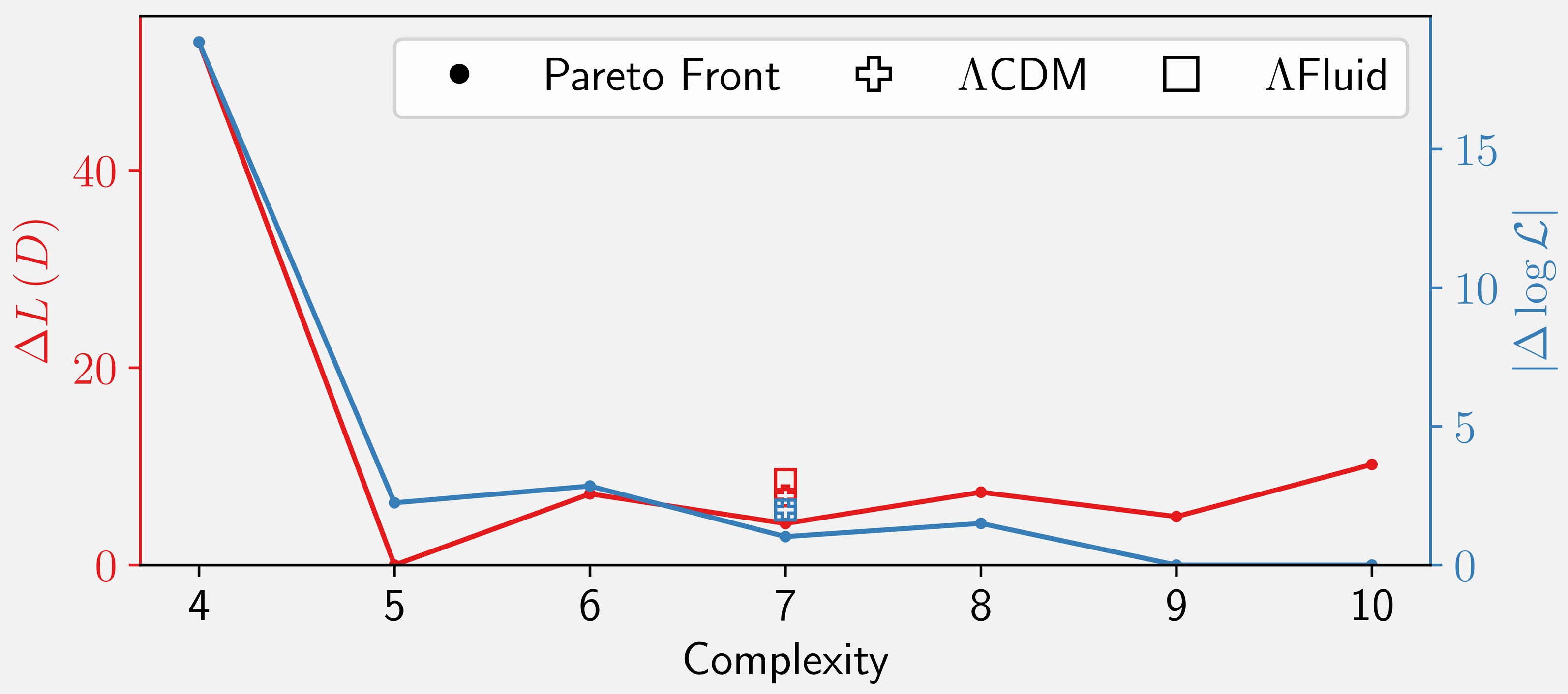

Figure 2: The "Pareto front" shows the best $L(D)$ and likelihood $\mathcal{L}$ achievable at any complexity. ΛCDM lies above this line: it is "Pareto-dominated" by the ESR results.

Rank	$H(z)^2$ / km^{2}s^{-2}Mpc^{-2}	Complexity	$-\log\mathcal{L}$	$L(D)$	
1	$\theta_0(1+z)^2$	5	8.36	16.39	←Best
2	$\|\theta_0\|^{(1+z)^{\theta_1}}$	5	7.97	18.70	
3	$\theta_0\,\|\theta_1\|^{-(1+z)}$	5	7.57	20.08	
⋮	⋮	⋮	⋮	⋮	
39	$\theta_0+\theta_1(1+z)^3$	9	7.28	23.51	←ΛCDM
⋮	⋮	⋮	⋮	⋮	

Table 1: Highest ranked functions for $H(z)^2$ inferred by ESR. 38 surpass ΛCDM.

Application 2 – Is galaxy dynamics MOND?

- The Radial Acceleration Relation (RAR) describes the coupling between galaxies' visible and dynamical mass. It is claimed to support MOND by having asymptotic slopes of 1/2 & 1 ("deep-MOND" & "Newtonian" regimes).
- We apply ESR to the SPARC RAR to ask whether the best functions have these limits, and how good they are compared to the classic MOND functions.
- We find scant support for the deep-MOND limit, and **many functions better than those of MOND**.

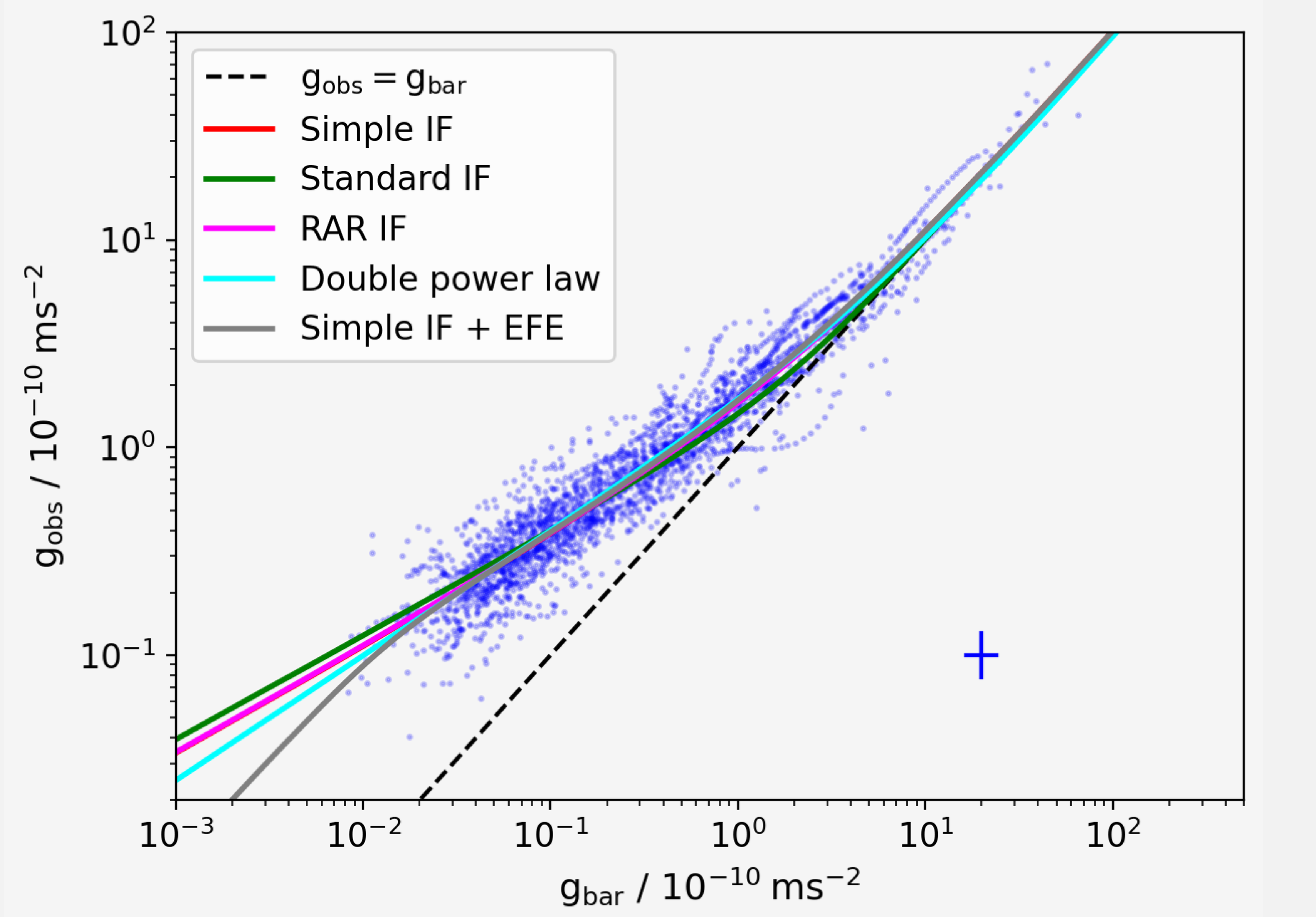

Figure 3: The SPARC RAR (blue points) and its classic MOND fits ("IFs").

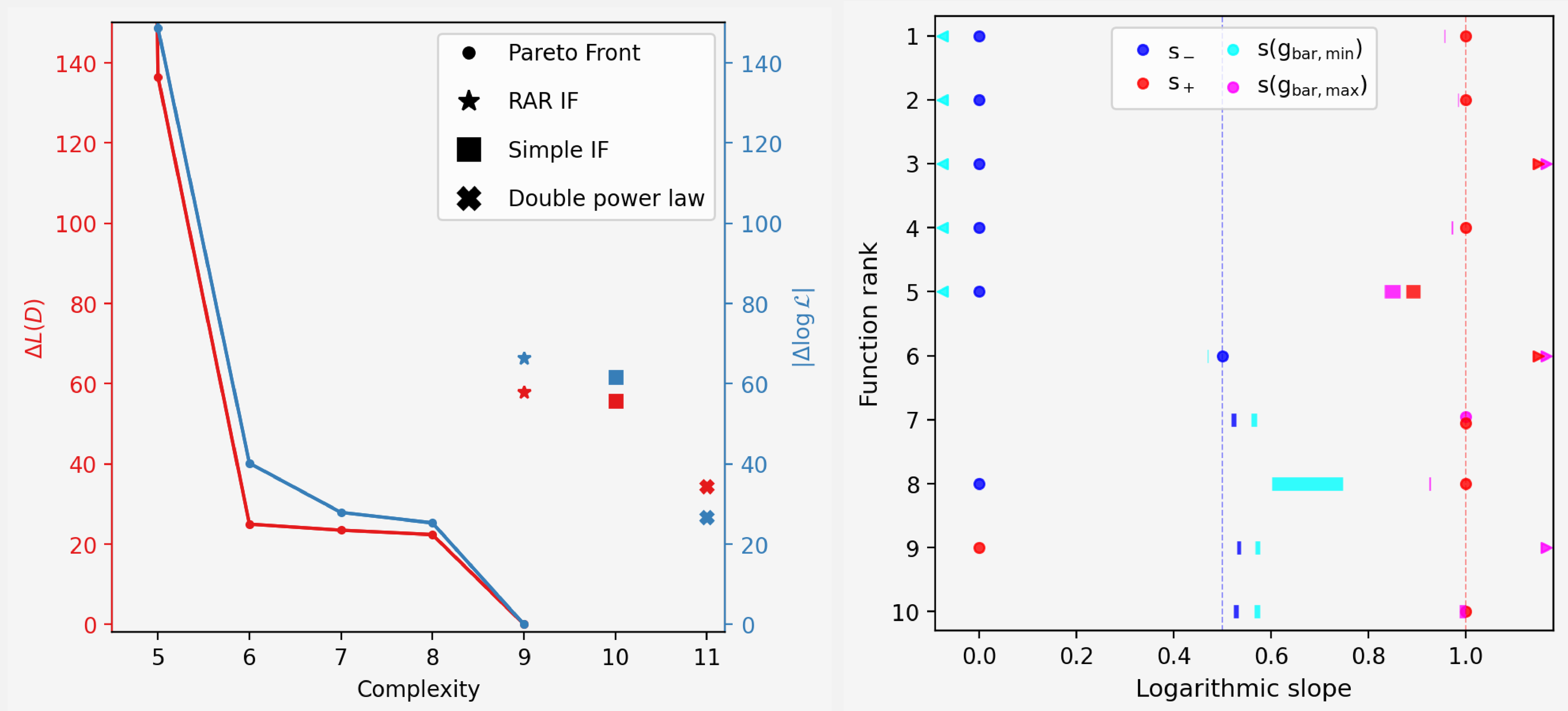

Figure 4: *Left*: Pareto front of $L(D)$ found by ESR vs those of classic MOND IFs and double power law. *Right*: The limiting slopes of the 10 best functions. MOND functions would have blue and red points on the corresponding dotted lines, which is typically not the case.

Application 3 – Is inflation quadratic, quartic or Starobinsky?

- Use ESR to score all possible functional forms for the single-field inflaton potential using A_s, n_s and r constraints from the CMB.
- Consider also a **language-model (Katz) prior** which upweights functions similar to those in a training set (the *Encyclopaedia Inflationaris*).
- Compare to literature solutions such as quadratic, quartic and Starobinsky. **We find thousands of potentials better than these!**

Rank	$V(\phi)$	Comp.	$L(D)$	Rank	$V(\phi)$	Comp.	$L(D)$
1	$e^{-e^{e^{e^{\phi}}}}$	6	-6.06	1	$\theta_0\phi^{\theta_1/\phi}$	7	-2.59
2	$\theta_0 e^{-e^{\phi}}$	5	-5.16	2	$\theta_0(\theta_1+\phi^{\phi})$	7	-1.39
3	$\|\theta_0\|^{e^{e^{\phi}}}$	5	-5.09	3	$\theta_0\phi^{\phi^{\theta_1}}$	7	-0.63
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1272	$\theta_0(1-e^{-\sqrt{2/3}\phi})^2$	9	5.57	12	$\theta_0(1-e^{-\sqrt{2/3}\phi})^2$	9	0.70
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
8697	$\theta_0\phi^2$	4	38.01	5401	$\theta_0\phi^2$	4	38.03

Table 2: Best inflaton potentials found by ESR, plus quadratic and Starobinsky inflation. *Left*: without Katz prior; *Right*: with Katz prior.

Where next?

- In all cases investigated so far we find **clear gains over literature standards**.
- **Improvements to the algorithm**: autodiff, integer snap, increased efficiency of tree operations, improved likelihoods. Plus hybrid exhaustive/stochastic approaches, using ESR to inform the operation of genetic algorithms.
- **New applications**: Halo profiles (from data and simulations), galaxy and halo mass functions, bias relations, ...
- **Your ideas are wanted!!**