
SYMBOLIC REGRESSION

Veronica Guidetti

*Multi-Objective Optimization and
Symbolic Regression
Lecture IV*

SYMBOLIC MACHINE LEARNING (SML)

SML is a subfield of AI that focuses on processing, manipulating, and producing symbols or concepts rather than numerical data.

PROS:

- **Interpretable results:** can distill complex phenomena into understandable analytical constructs
- **Flexibility:** SML can be adapted to different domains by modifying the rules and knowledge base
- **Knowledge representation:** easy integration of new data-driven knowledge with existing theories

CONS:

- **Need for complete knowledge:** to function correctly. In domains where knowledge is incomplete, SML may not be effective.
- **Scales poorly:** at increasing input dimensions. For example, symbolic regression is NP-hard.
- **Limited ability to learn and adapt:** Symbolic AI models may not allow intensive real-time learning and adaptation.

SYMBOLIC REGRESSION [1,2]

- Symbolic regression (SR) refers to data-driven ML methods for regression analysis that search the space of mathematical formulas to find the model that best fits a given dataset.
- In practice, whatever follows next is not constrained to regression problems but can be extended to
 - Most parametric statistical model
 - Other general problems with multiple parameters and analytic solution/formulation

HOW DOES IT WORK?

- Dataset of observations

$$\{\mathbf{x}_i, y_i\}_{i=1,\dots,N}$$

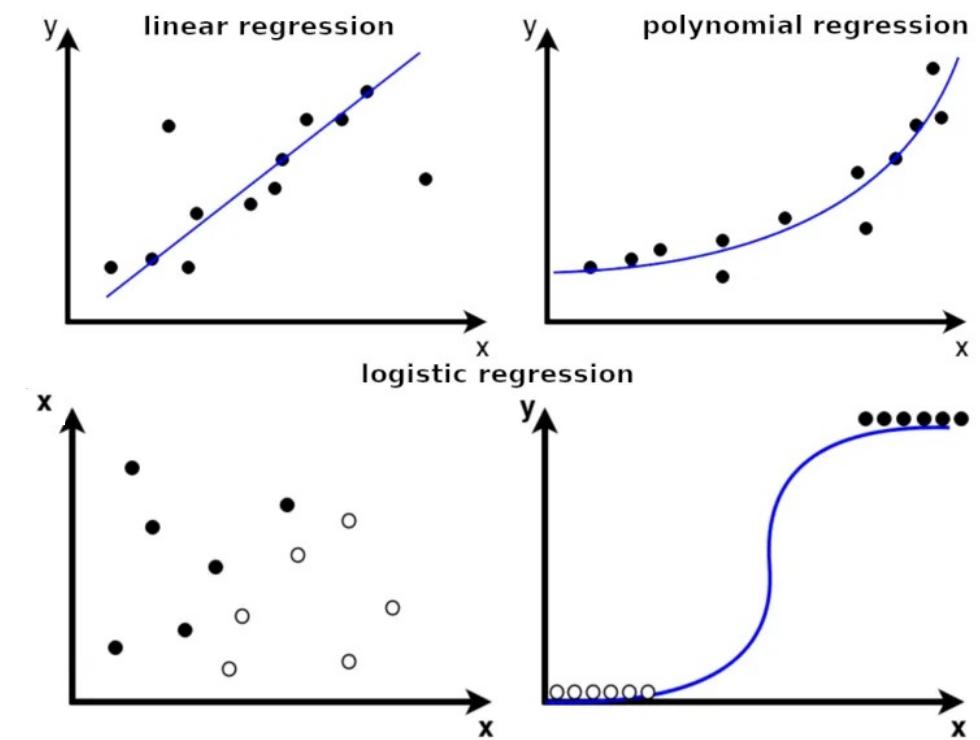
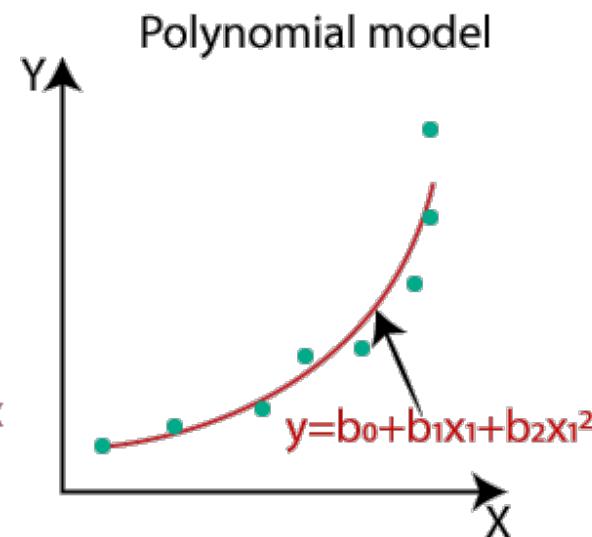
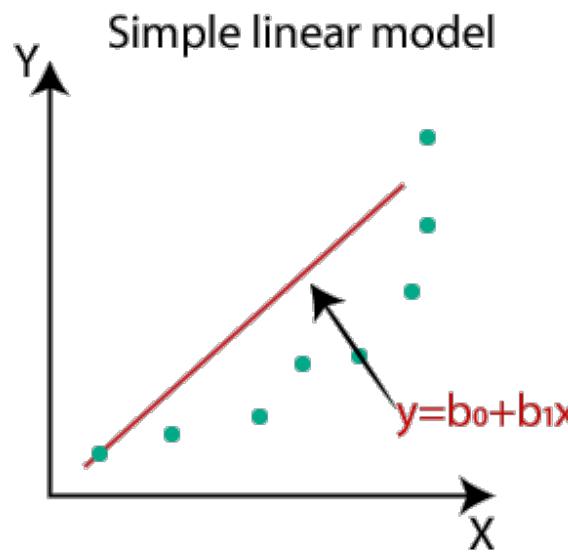
- Representation of candidate solutions

$$x_i \rightarrow f(x_i, \theta)$$

- Adjust model parameters to optimize an objective/loss/fitness function.

This is easy when having a fixed model class (linear/logit/poly),
but what about finding the model as well?

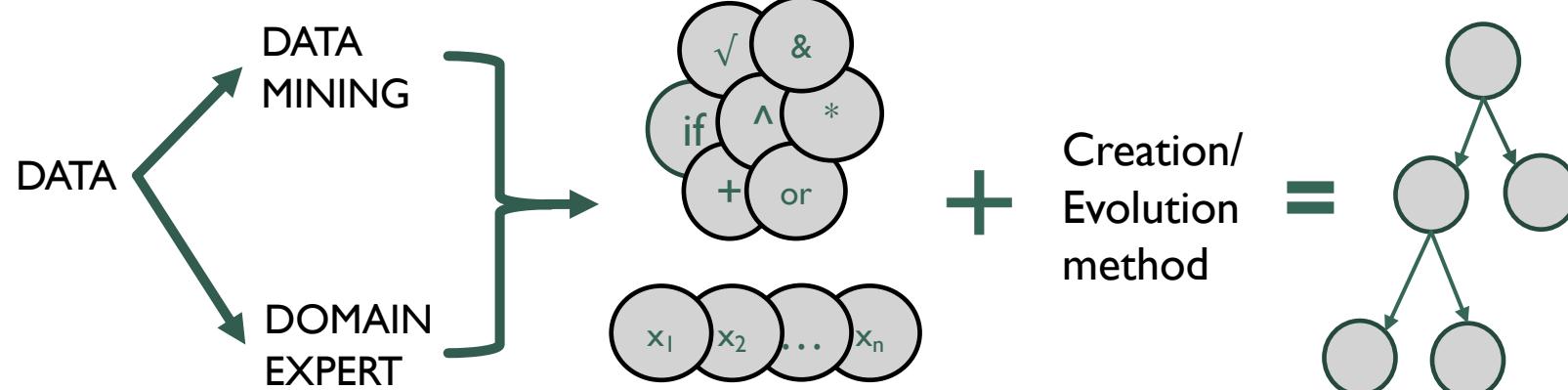
SYMBOLIC REGRESSION [1,2]



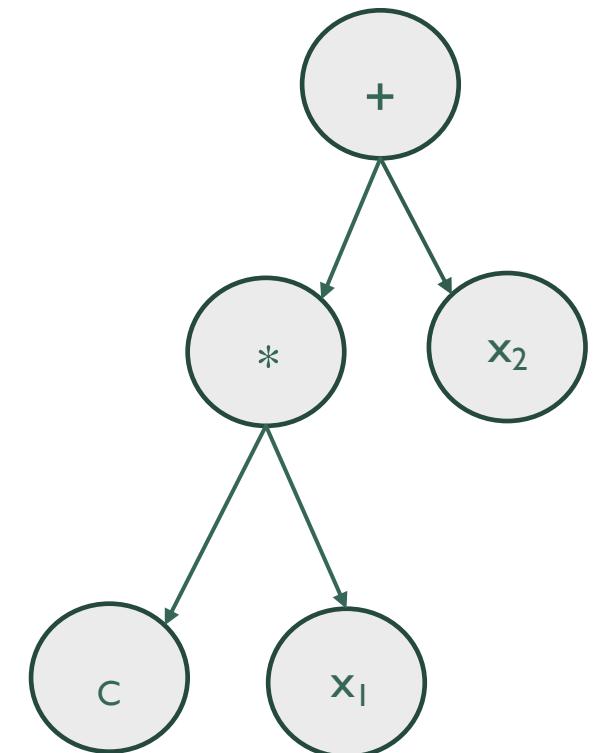
CANDIDATE SOLUTIONS

Knowledge:

symbolic representation of underlying data to be described as the composition of knowledge primitives.
Exploit the Polish notation to create a bijective map between formulas and abstract syntax trees.



$$f(x_1, x_2) = c * x_1 + x_2$$



OBJECTIVE/LOSS/FITNESS FUNCTION

- Maximum likelihood estimation

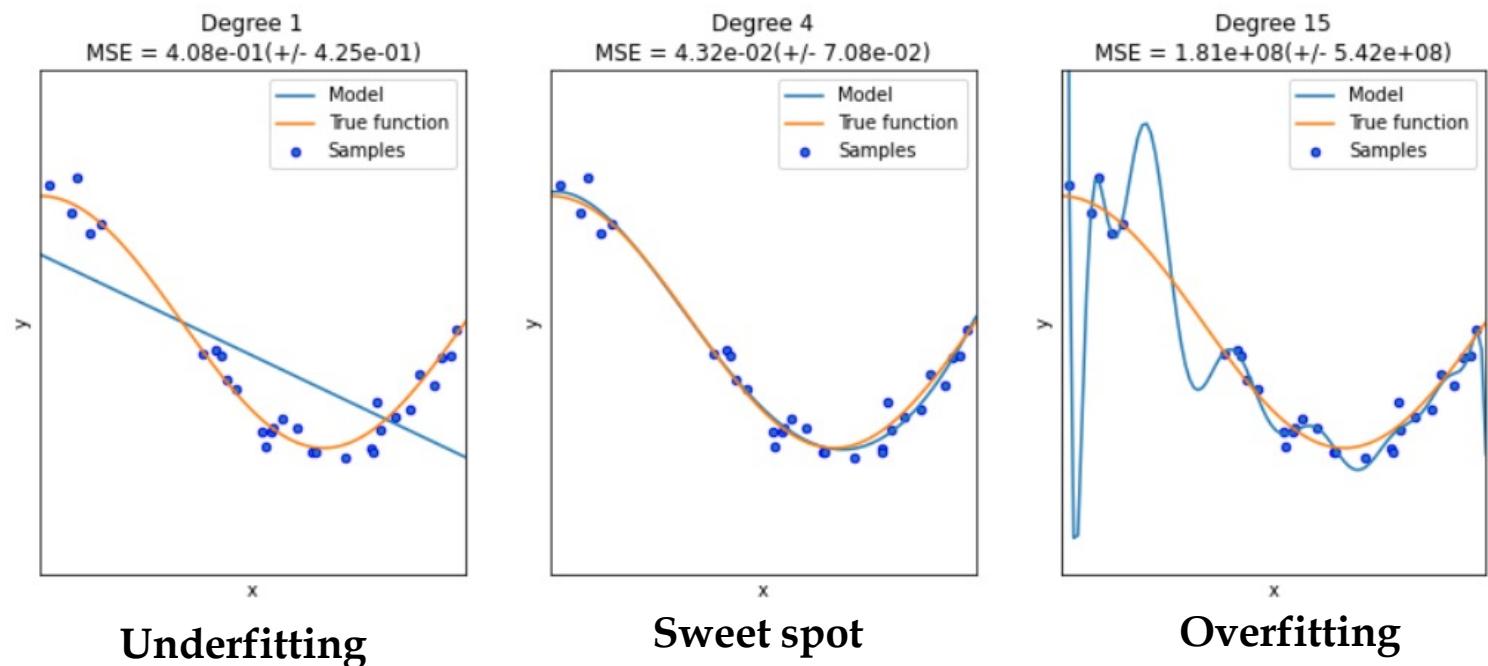
$$\max_{\theta} P(\theta, \mathcal{D}) = \max_{\theta} \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma}} \exp \left\{ -\frac{(y_i - f(x_i, \theta))^2}{2\sigma^2} \right\}$$

OBJECTIVE/LOSS/FITNESS FUNCTION

- Maximum likelihood estimation

$$\max_{\theta} P(\theta, \mathcal{D}) = \max_{\theta} \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma}} \exp \left\{ -\frac{(y_i - f(x_i, \theta))^2}{2\sigma^2} \right\}$$

Prone to
parameter
overfitting!



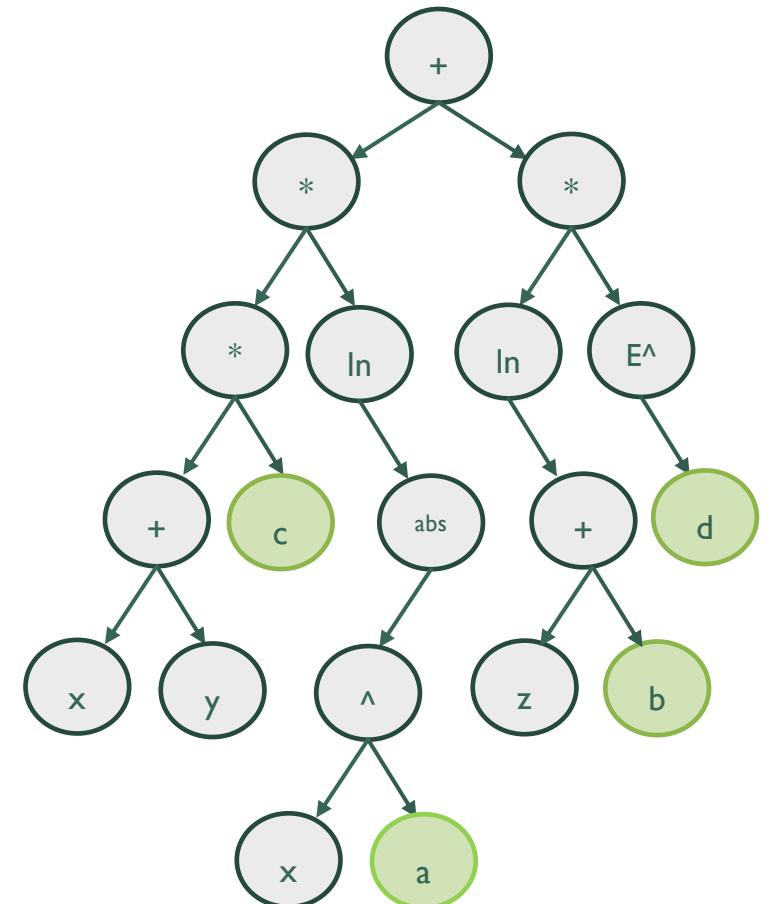
OBJECTIVE/LOSS/FITNESS FUNCTION

- Maximum likelihood estimation
- Akaike/Bayes Information Criterion

$$AIC = 2|\theta| - 2 \log(P(\theta, \mathcal{D}))$$

$$BIC = \log(N)|\theta| - 2 \log(P(\theta, \mathcal{D}))$$

Penalize parameter but not **operator verbosity**



OBJECTIVE/LOSS/FITNESS FUNCTION

- Maximum likelihood estimation
- Akaike/Bayes Information Criterion
- Minimum Description Length

$$L(\mathcal{D}) = L(H) + L(\mathcal{D}|H)$$

$$- \log(P(\theta^*, \mathcal{D})) + \mathcal{N} \log(\mathcal{O}) + |\theta| \log(2) + \sum_{i=1}^{|\theta|} \log(\theta^*/\Delta_i)$$

Log-likelihood

Tree-extension

Parameter sign and precision

CREATION/EVOLUTION METHODS

I. Genetic Algorithms [1-7]

- Initialize a population of random formulas
- Evolve it via
 - genetic mutations
 - objective evaluation

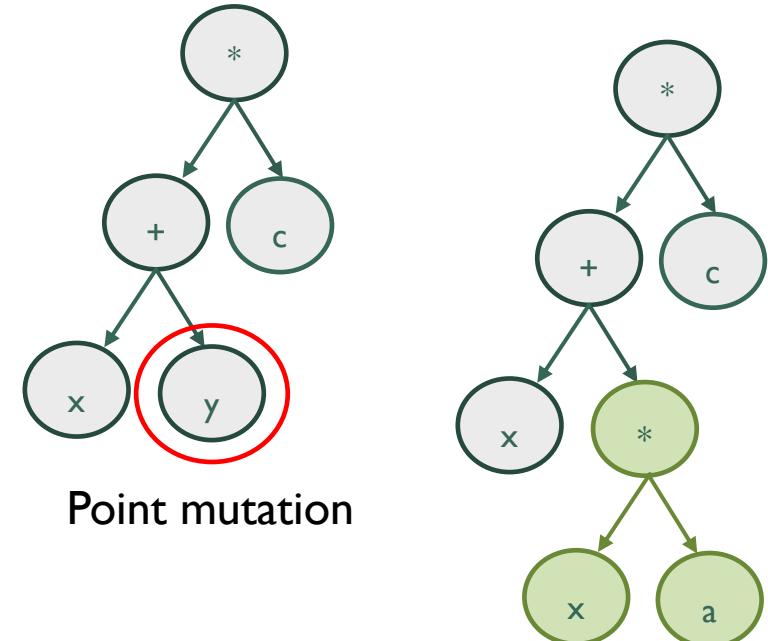
Pros:

No intrinsic limits

Can accommodate Multi-Objective optimization

Cons:

Poor convergence (NP-hard) for high input dimensions



CREATION/EVOLUTION METHODS

1. **Genetic Algorithms**
2. **Exhaustive searches [8]**

- Compute all valid graph structures up to a max complexity
- Find global minimum on that set

Pros:

Good for small formulas

Cons:

Constrained domain search

CREATION/EVOLUTION METHODS

1. **Genetic Algorithms**
2. **Exhaustive searches**
3. **Recursive approaches [9,10]**

- combine neural network fitting and physics-inspired rules and transformations to find optimal models (AI-Feynman)
- Detect symmetries in data by studying modularity in the formula graph

Pros:

Smart management of units of measure

Cons:

Symmetries need to be hard-coded
operation order is static

CREATION/EVOLUTION METHODS

1. **Genetic Algorithms**
2. **Exhaustive searches**
3. **Recursive approach**
4. **Graph neural network approach [11]**

- Map (parts) of the dynamical system into a graph (NN)
- Augment data
- Train GNN

Pros:

Split the problem into simpler ones by inferring part of GNN

Cons:

dynamical graph-like settings

data augmentation need to partially exploit previous knowledge

CREATION/EVOLUTION METHODS

1. **Genetic Algorithms**
2. **Exhaustive searches**
3. **Recursive approach**
4. **Graph neural network approach**
5. **Transformer-based architecture [12-14]**

- Treat mathematics as a language
- Create a training set for a specific objective
- Train

Pros:

does not need training each problem from scratch

Cons:

still poor scaling, no guarantees when predicting

CREATION/EVOLUTION METHODS

- 1. **Genetic Algorithms**
- 2. **Exhaustive searches**
- 3. **Recursive approach**
- 4. **Graph neural network approach**
- 5. **Transformer-based architecture**
- 6. **...**
-
-
- *many others*

POSSIBLE CASE STUDIES

SR may act on

- Experimental Data
- Numerical simulation to distill results
- Datasets to create standard predictive modelling (classification/regression/survival)
- Datasets to optimize multiple non-derivative objectives

SR AND EPISTEMIC UNCERTAINTY [26,27]

This is a nice result but what about quantifying model uncertainty?

We can use parametric bayesian inference on model parameters or directly go for model selection.

BAYESIAN FORMULATION OF SYMBOLIC REGRESSION

To do so we need to define:

- **Prior over models:** compile a corpus of closed-form mathematical models and formalize a prior distribution.
- Rules to explore the posterior over model expressions via **MCMC sampling** (proposal distribution).

BAYESIAN FORMULATION FOR MODEL SELECTION

Model posterior distribution is given by:

$$p(f_i | D) = \frac{1}{Z} \int_{\Theta_i} d\theta_i p(D | f_i, \theta_i) p(\theta_i | f_i) p(f_i) = \frac{\exp [-\mathcal{L}(f_i)]}{Z}$$

Evidence P(D) likelihood parameter posterior model prior

where

$$\mathcal{L}(f_i) \equiv -\log [p(D, f_i)]$$

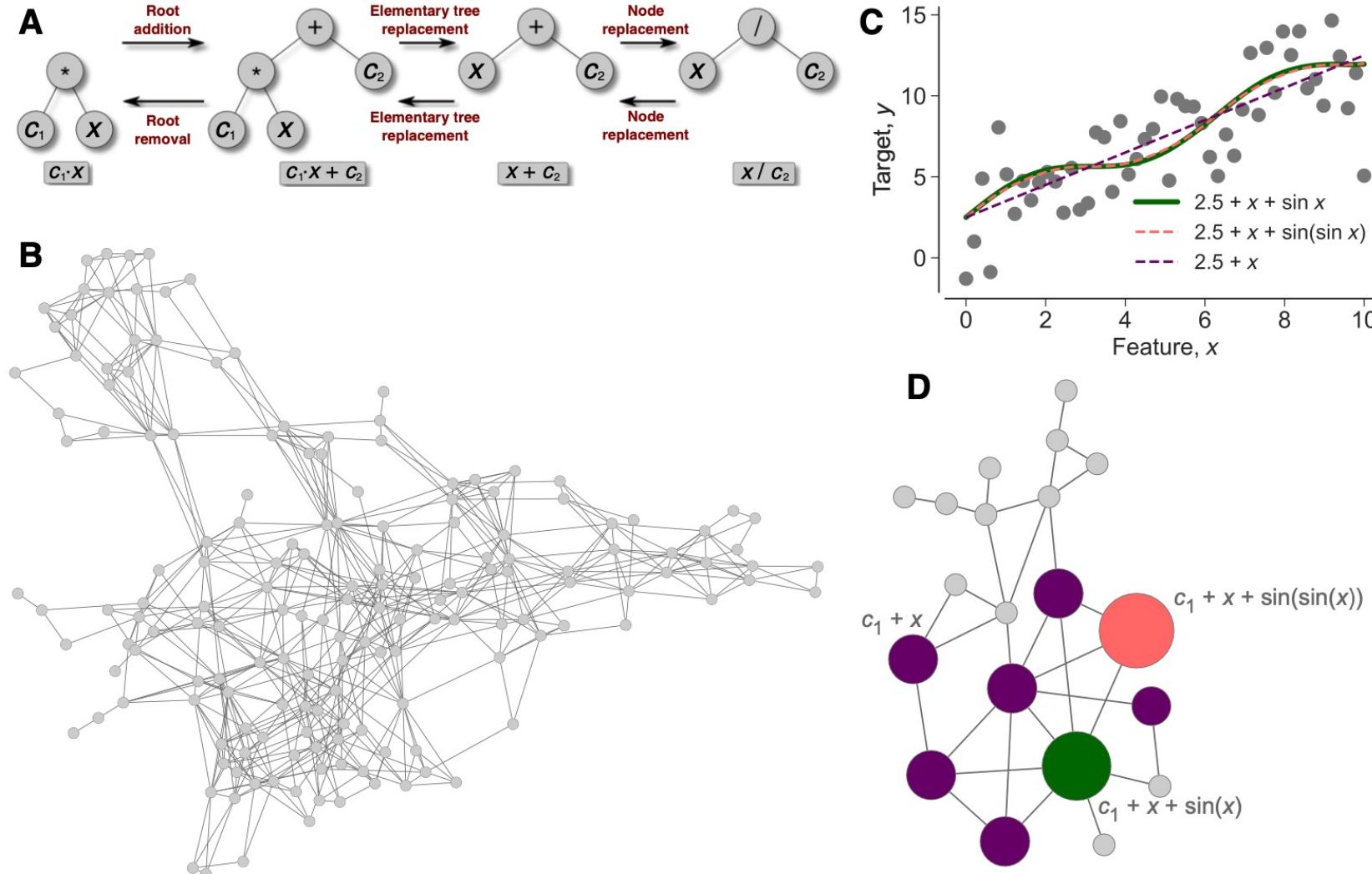
$$= -\log \left[\int_{\Theta_i} d\theta_i p(D | f_i, \theta_i) p(\theta_i | f_i) p(f_i) \right] \approx \frac{B(f_i)}{2} - \log p(f_i)$$

using Laplace's method to integrate the (likelihood \times parameter posterior) over the parameters.

STEPS

- Prior over expressions acts as a model regularizer
- Explore functional space: elementary jumps and a proposal movement distribution $g(f_i, f_f)$
- Use MCMC (metropolis rule) algorithms to sample from approximate posterior distributions

$$p_{\text{accept}}(f_i \rightarrow f_f) = \min \left\{ 1, \frac{p(f_f | D)g(f_i | f_f)}{p(f_i | D)g(f_f | f_i)} \right\}$$



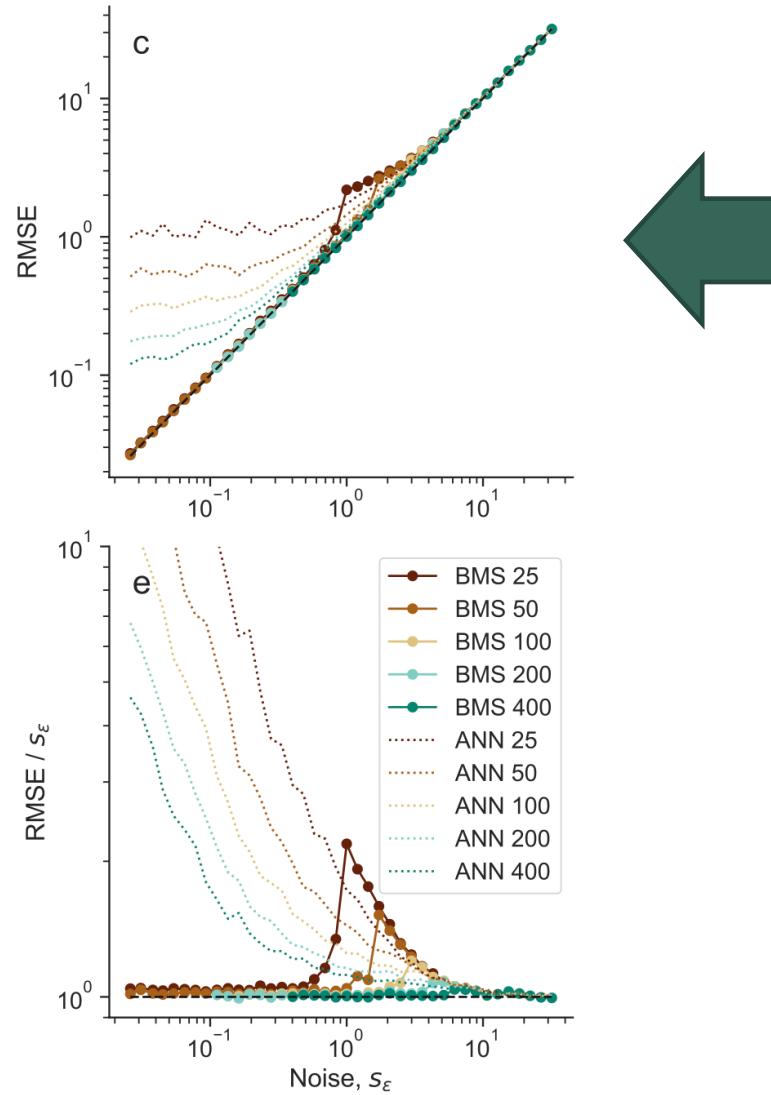
Taken from [26]: Guimerà, Roger, et al. "A Bayesian machine scientist to aid in the solution of challenging scientific problems." (2020).

STUDY MODEL LEARNABILITY

Every experimental data comes with noise s_ϵ

low s_ϵ	high s_ϵ
true model	trivial model priors

Increasing s_ϵ we see a transition to the unlearnable regime: phase transition driven by changes in the model plausibility/description length landscape



BAYESIAN SR vs. NNs

s_ϵ is the irreducible error

predictions on the diagonal
 $RMSE = s_\epsilon$ are **optimal**.

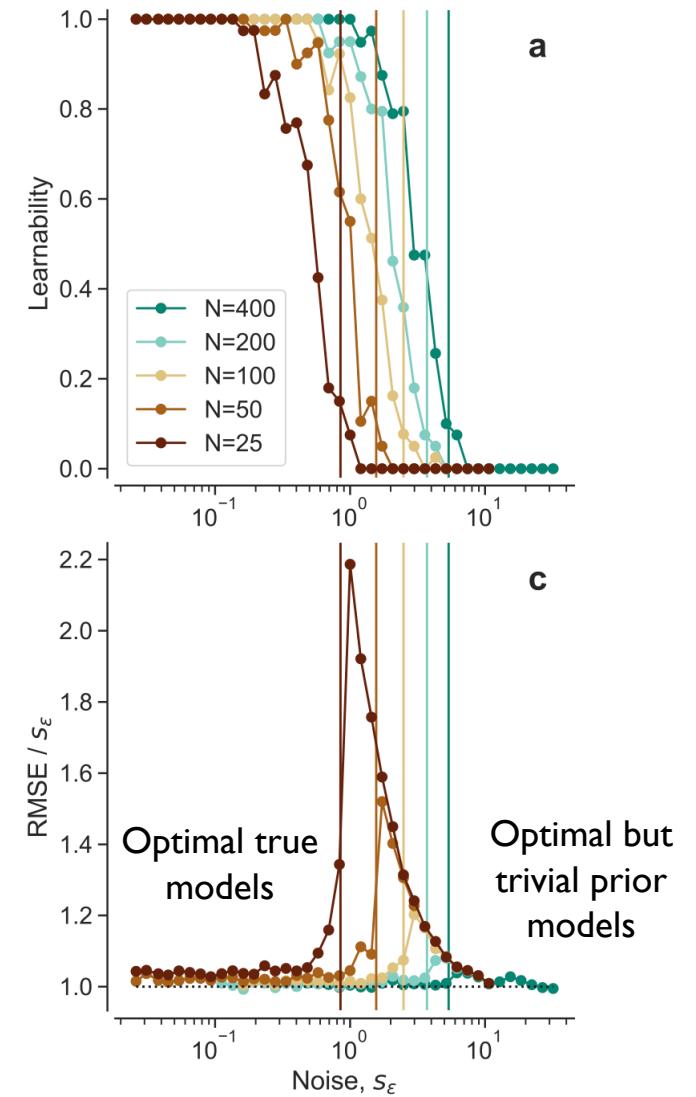
Bayesian SR is quasi-optimal
 NN are optimal only for large s_ϵ

Taken from [27]: Fajardo-Fontiveros, Oscar, et al. "Fundamental limits to learning closed-form mathematical models from data." (2023)

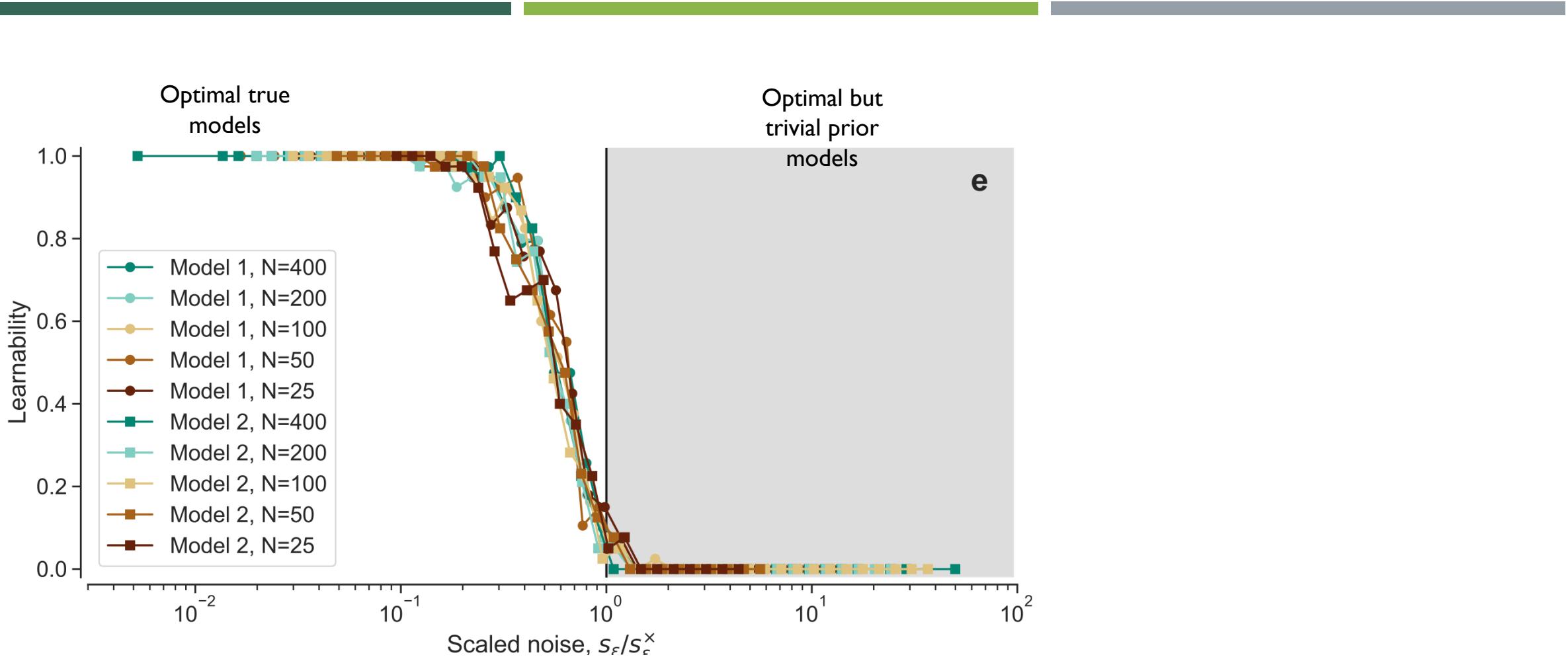
Model learnability: the fraction of training datasets D at a given noise s_ϵ for which the true generating model is learnable.

learnability phase transition shifts towards higher values of s_ϵ with increasing size of D .

BAYESIAN SR
Learnability vs. Noise



Taken from [27]: Fajardo-Fontiveros, Oscar, et al. "Fundamental limits to learning closed-form mathematical models from data." (2023)



Taken from [27]: Fajardo-Fontiveros, Oscar, et al. "Fundamental limits to learning closed-form mathematical models from data." (2023)

EFFICIENT TRADEOFF

Empirical studies show that sampling functions via MCMC may not lead to optimal solutions [3].

For this reason, efficient tradeoff approaches may be used. For example:

- Genetic programming to sample solutions from the space of closed-form mathematical functions
- Bayesian inference on a set of theoretically grounded models from the final population

Moreover, GP allows to explore Multi-Objective optimization setups!

WHY MOVING TO MULTIOBJECTIVE OPTIMIZATION?

You may need to optimize multiple derivative and non-derivative objectives.

Prominent example: Symbolic ML must generate interpretable models

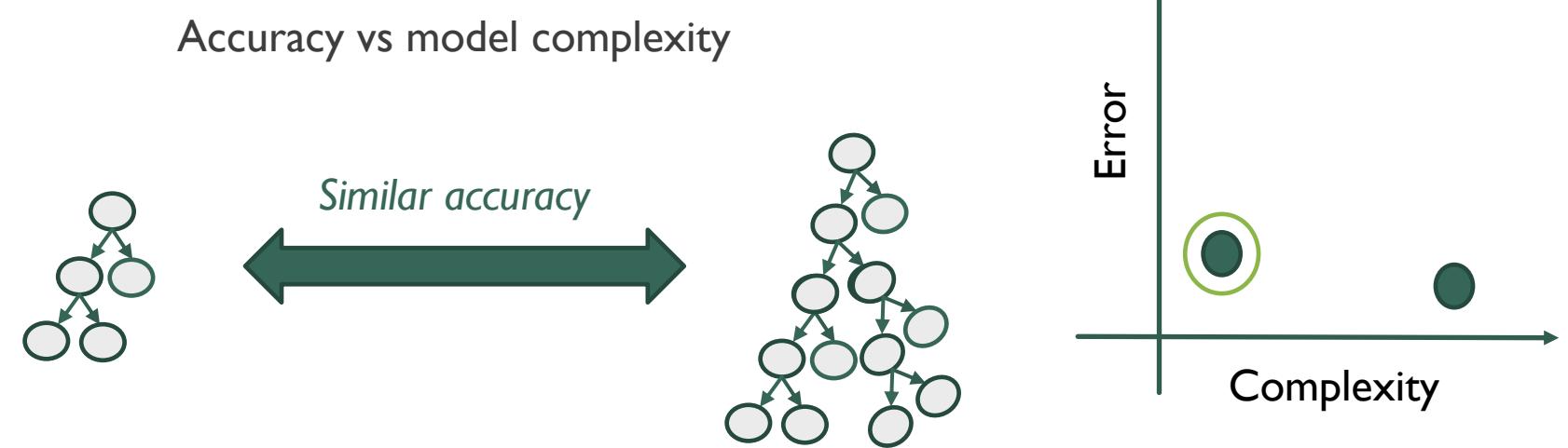
Accuracy vs model complexity



WHY MOVING TO MULTIOBJECTIVE OPTIMIZATION?

You may need to optimize multiple derivative and non-derivative objectives.

Prominent example: Symbolic ML must generate interpretable models



MOSR WITH GENETIC PROGRAMMING

I. Genetic Algorithms [1-7]

- Initialize a population of random formulas
- Compute fitness function(s)
- Apply genetic mutations
- Evaluate fitness and select offspring
- Repeat until convergence / for N generations

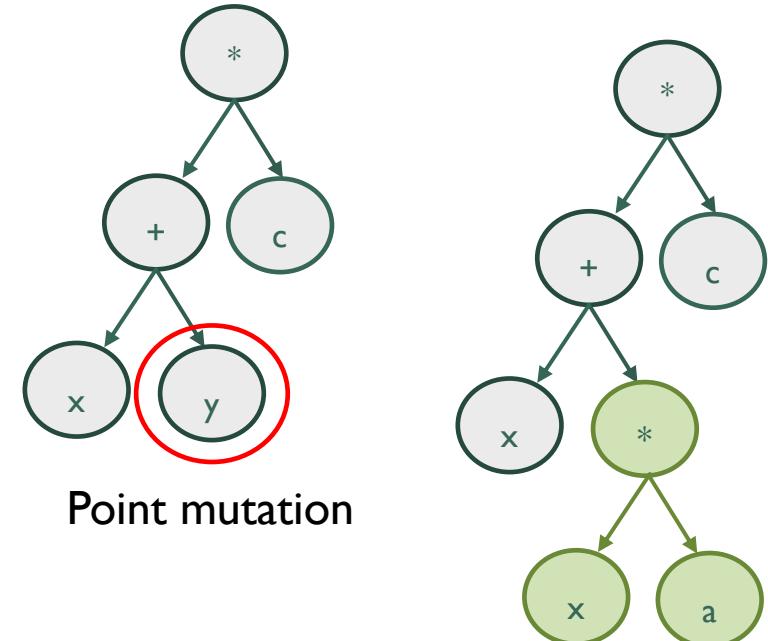
Pros:

No intrinsic limits

Can accommodate Multi-Objective optimization

Cons:

Poor convergence (NP-hard) for high input dimensions



CREATION/EVOLUTION METHODS

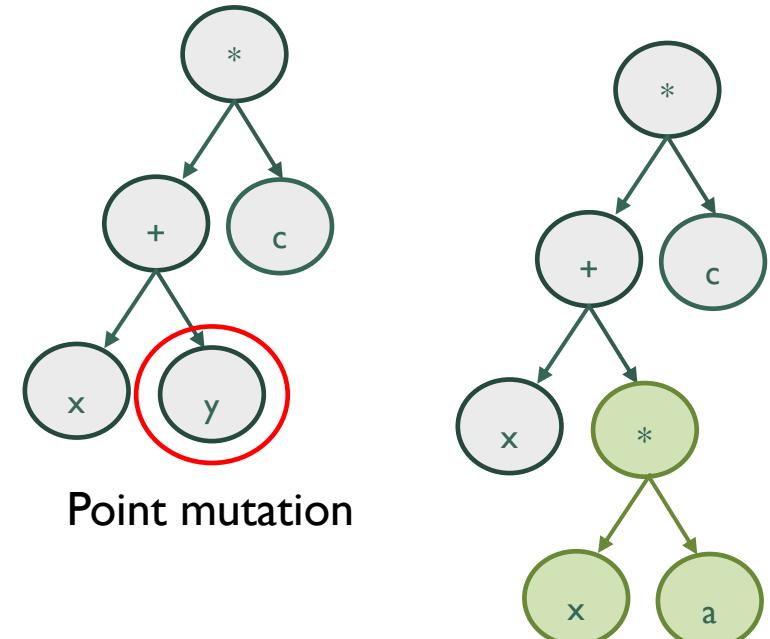
I. Genetic Algorithms [1-7]

- Initialize a population of random formulas

Generate random trees using 2 hyperparameters
and random number generation

Hyperparameters: Parsimony: p, Parsimony decay: d

```
if r = random(0,1) < p:  
    pick a random operator,  
    p  $\leftarrow$  p * d  
else  
    pick a variable or constant
```



CREATION/EVOLUTION METHODS

I. Genetic Algorithms [1-7]

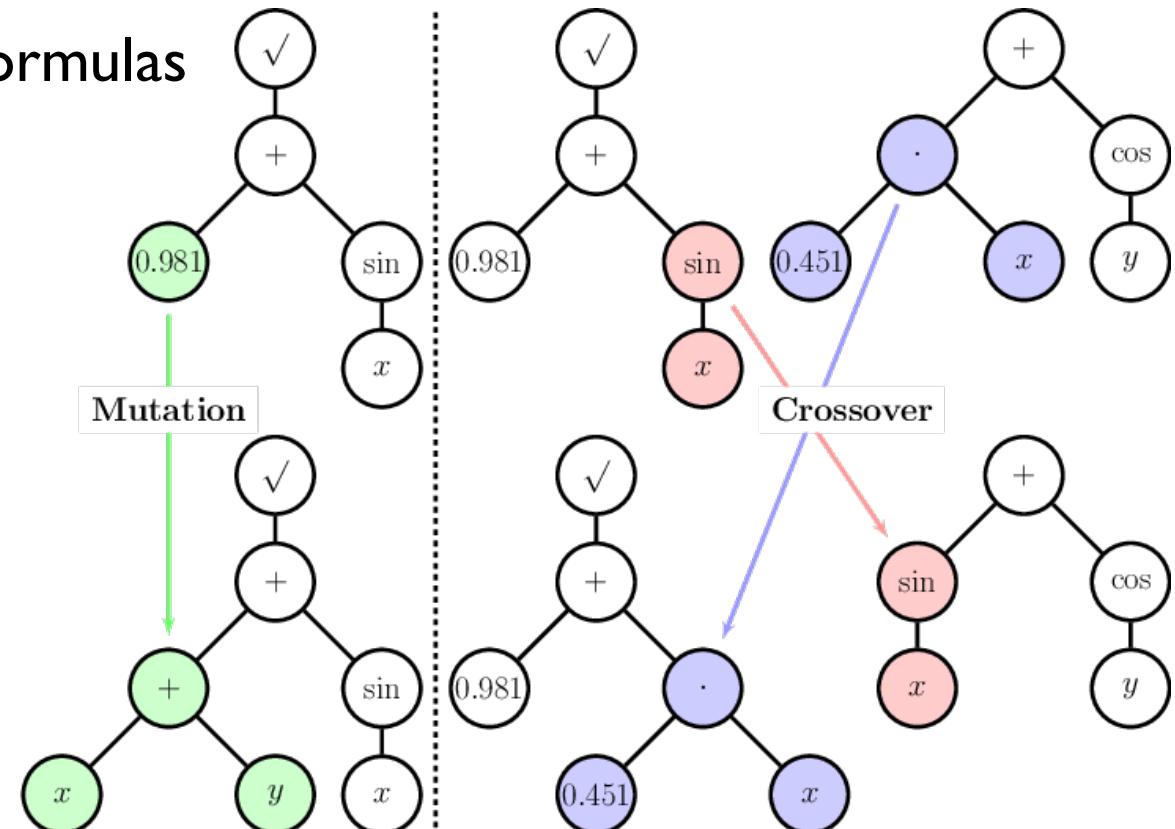
- Initialize a population of random formulas
- Compute fitness function(s)
- Apply genetic mutations

What are genetic mutations?

Classical: Mutation and Crossover

Others:

- Node insertion
- Node deletion
- Recalibration
- ...

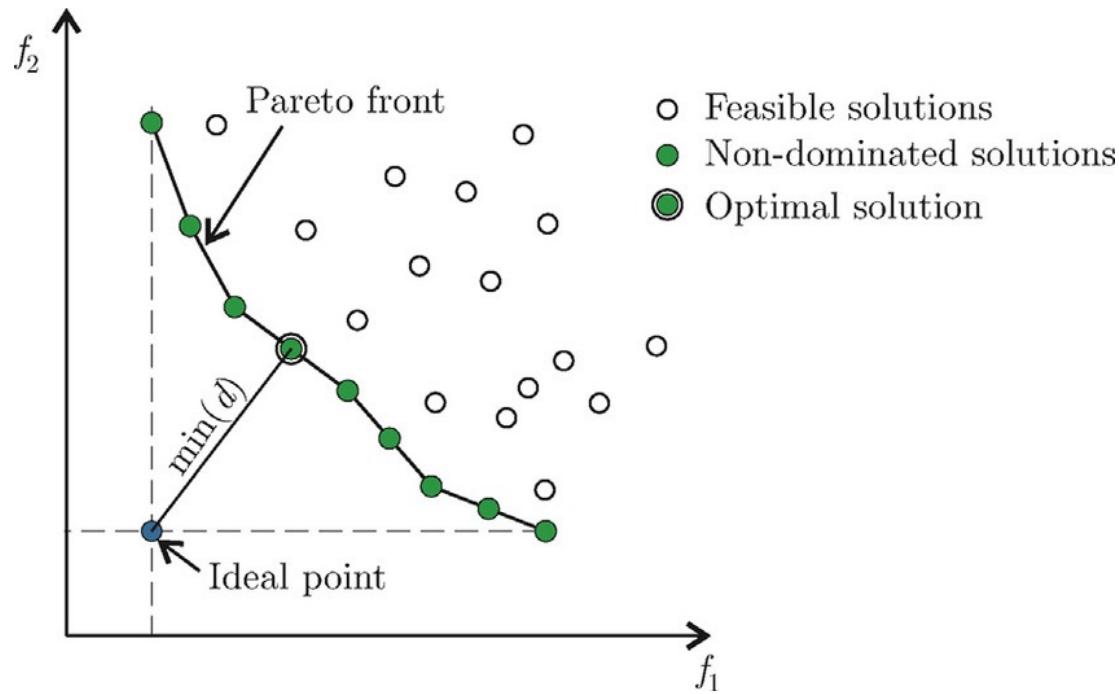


CREATION/EVOLUTION METHODS

I. Genetic Algorithms [1-7]

Multi-objective SR:

- No global ordering
- Constrained problem
- Desiderata



Training Pipeline:

- Initialize random population
- Compute fitnesses
- Compute dominance and crowding distance
- Apply mutations
- Select population

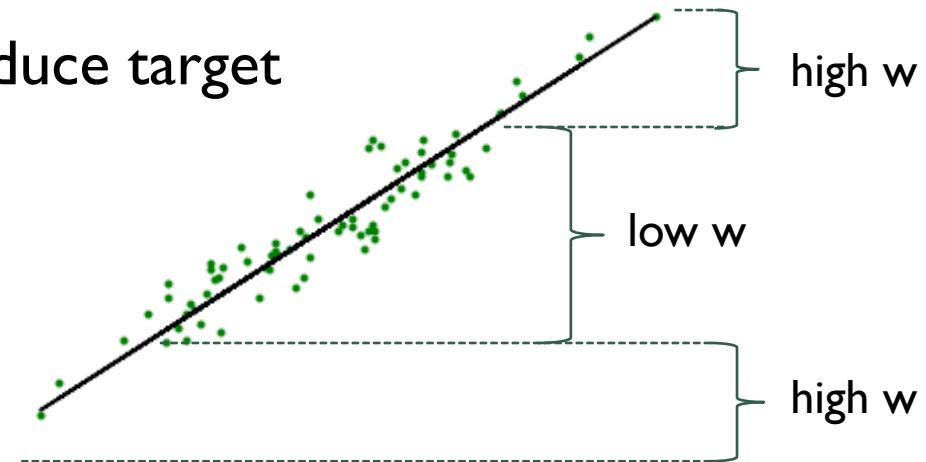
WHICH OBJECTIVES?

Regression

- **wMSE/wAIC/wMDL**
- **Kendall correlation**
- **Wasserstein distance**



Reproduce target



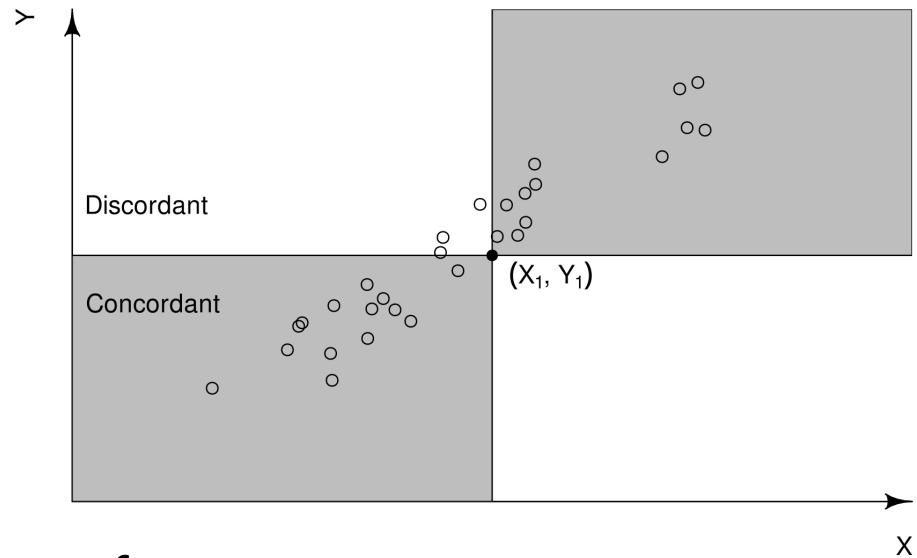
WHICH OBJECTIVES?

Regression

- wMSE/wAIC/wMDL
- Kendall correlation
- Wasserstein distance



Reproduce stratification



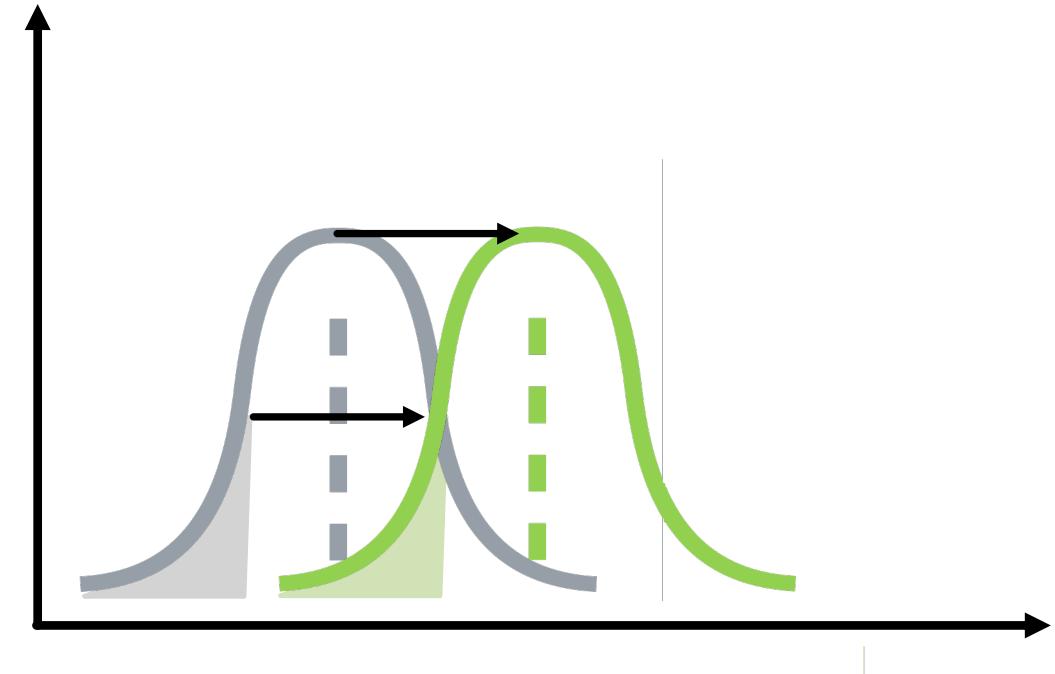
WHICH OBJECTIVES?

Regression

- wMSE/wAIC/wMDL
- Kendall correlation
- Wasserstein distance



Reproduce (im)balancing



WHICH OBJECTIVES?

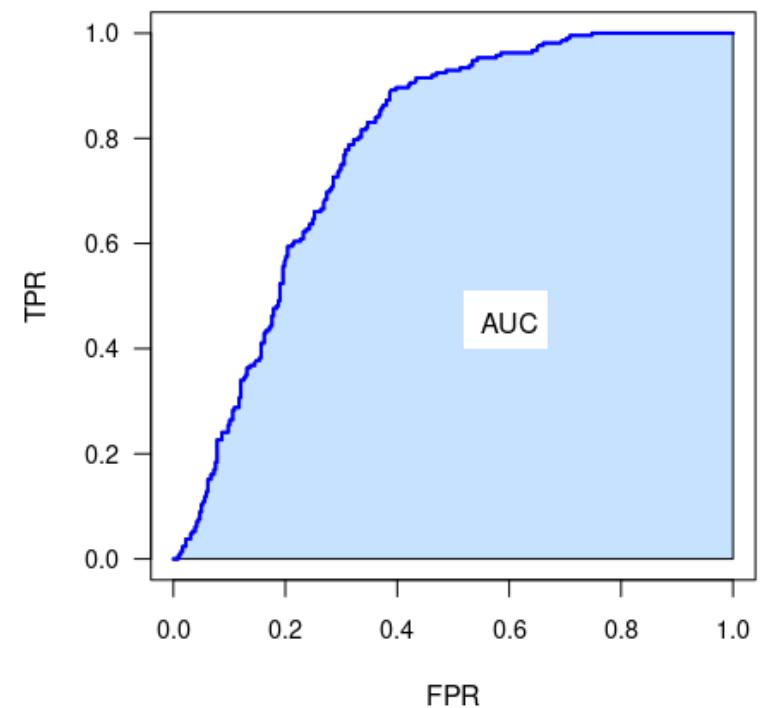
Classification (binary)

- **wBCE/wAIC/wMDL**
- **Precision/Recall**
- **Area under the ROC curve**

Predicted			
Actual	0	1	
	0	TN	FP
1	FN		TP

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$



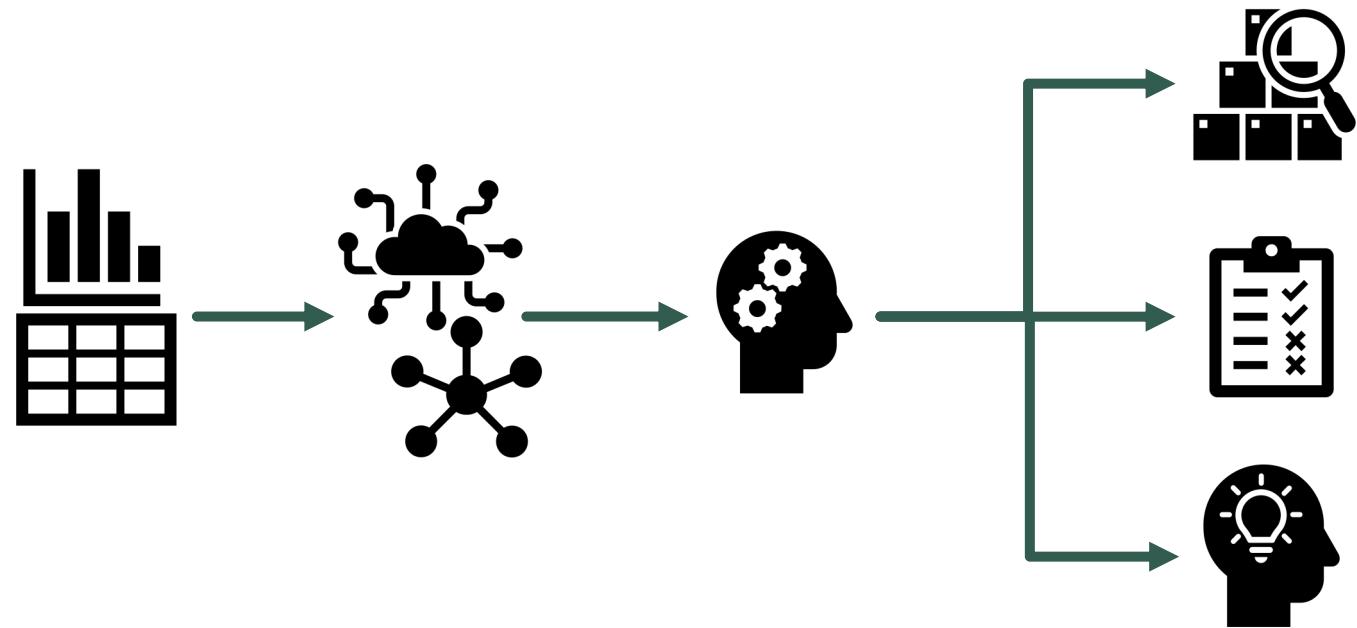
WHICH OBJECTIVES?

Others:

- **Constraints**
- **Fairness criteria**
- **Correlation with some quantity**
-

WHY SYMBOLIC REGRESSION?

- Parsimonious formulas
- Limited number of parameters
- Tailored non-linear functions



EXPLAINABILITY IS NOT INTERPRETABILITY!

SYMBOLIC REGRESSION: HOW?

Some available libraries for GP-SR

- PySR: <https://github.com/MilesCranmer/PySR>
- GP-learn: <https://gplearn.readthedocs.io/en/stable/>
-
- **Ours!** https://github.com/davideferrari92/multiobjective_symbolic_regression
https://github.com/mattiabilla/mosr_survival/

... We need to find an acronym...

Our library was specifically developed to perform MO-SR and implements NSGA-II and SMS-EMOEA algorithms (for now, more is coming).

OPEN PROBLEMS

Symbolic Regression has not yet reached its final formulation! Some non-exhaustive list of open problems:

1. Define model (tree) complexity in a rigorous manner
2. Define a model probability prior which can be largely accepted for many problems
3. Efficiently manage functional redundancies
4. Mutations are only defined based on graph structure, but ignore the overall impact on the functional behavior
5. Efficiently simplify formulas not only based on functional equivalencies (e.g. trigonometric identities)

REFERENCES

- [1] Koza, J. R. (1994). Genetic programming as a means for programming computers by natural selection. *Statistics and computing*, 4, 87-112.
- [2] Michael Schmidt and Hod Lipson. Distilling free-form natural laws from experimental data science, 324(5923):81–85, 2009.
- [3] La Cava, W., Orzechowski, et al. (2021). Contemporary symbolic regression methods and their relative performance. arXiv preprint arXiv:2107.14351.
- [4] Bogdan Burlacu, et al. Parsimony measures in multi-objective genetic programming for symbolic regression. In Proceedings of GECCO, 2019.
- [5] Qi Chen, et al. Rademacher complexity for enhancing the generalization of genetic programming for symbolic regression. *IEEE Transactions on Cybernetics*, 2020.
- [6] Guidetti, Veronica, et al. "Death After Liver Transplantation: Mining Interpretable Risk Factors for Survival Prediction." 2023 IEEE 10th International Conference DSAA. 2023.
- [7] Ferrari, Davide, Veronica Guidetti, and Federica Mandreoli. "Multi-Objective Symbolic Regression for Data-Driven Scoring System Management." 2022 IEEE ICDM, 2022.
- [8] Bartlett, D. J., Desmond, H., & Ferreira, P. G. (2023). Exhaustive symbolic regression. *IEEE Transactions on Evolutionary Computation*.
- [9] Udrescu, S. M., & Tegmark, M. (2020). AI Feynman: A physics-inspired method for symbolic regression. *Science Advances*, 6(16), eaay2631.
- [10] Liu, Z., & Tegmark, M. (2021). Machine learning conservation laws from trajectories. *Physical Review Letters*, 126(18), 180604
- [11] Miles Cranmer, et al., Discovering symbolic models from deep learning with inductive biases 2020
- [12] Vastl, M., Kulhánek, J., Kubalík, J., Derner, E., & Babuška, R. (2022). Symformer: End-to-end symbolic regression using transformer-based architecture. arXiv:2205.15764.
- [13] Kamienny, P. A., d'Ascoli, S., Lample, G., & Charton, F. (2022). End-to-end symbolic regression with transformers. *Advances in Neural Information Processing Systems*
- [14] Aurelien Dersy, Matthew D. Schwartz, and Xiaoyuan Zhang, Simplifying polylogarithms with machine learning 2022

REFERENCES

- [15] Ana Maria Delgado, et al, Modeling the galaxy-halo connection with machine learning, 2022
- [16] Pablo Lemos, Niall Jeffrey, Miles Cranmer, Shirley Ho, and Peter Battaglia, Rediscovering orbital mechanics with machine learning 2022
- [17] Konstantin T. Matchev, et al., Analytical modeling of exoplanet transit spectroscopy with dimensional analysis and symbolic regression 2022
- [18] Desmond, H., Bartlett, D. J., & Ferreira, P. G. (2023). On the functional form of the radial acceleration relation. *Monthly Notices of the Royal Astronomical Society*
- [19] Suyong Choi, Construction of a kinematic variable sensitive to the mass of the Standard Model Higgs boson in $H \rightarrow WW \rightarrow l\bar{b}vl-v-$ using symbolic regression, 2011
- [20] Anja Butter, Tilman Plehn, Nathalie Soyberman, and Johann Brehmer, Back to the formula—LHC edition 2021
- [21] Abdulhakim Alnuqaydan, Sergei Gleyzer, and Harrison Prosper, SYMBA: Symbolic computation of squared amplitudes in high energy physics with machine learning, 2023
- [22] Dong, Z., et al. (2023). Is the machine smarter than the theorist: Deriving formulas for particle kinematics with symbolic regression. *Physical Review D*, 107(5), 055018.
- [23] Sousa, T., Bartlett, D. J., Desmond, H., & Ferreira, P. G. (2023). The Simplest Inflationary Potentials. *arXiv:2310.16786*.
- [24] D. J. Bartlett, H. Desmond, and P. G. Ferreira, in *The Genetic and Evolutionary Computation Conference 2023* *arXiv:2304.06333 [cs.LG]*.
- [25] J. Martin, C. Ringeval, and V. Vennin, *Encyclopaedia Inflationaris Phys. Dark Univ.* 5-6, 75 (2014), *arXiv:1303.3787 [astro-ph.CO]*.
- [26] Guimerà, Roger, et al. "A Bayesian machine scientist to aid in the solution of challenging scientific problems." *Science advances* 6.5 (2020).
- [27] Fajardo-Fontiveros, Oscar, et al. "Fundamental limits to learning closed-form mathematical models from data." *Nature Communications* 14.1 (2023): 1043.