Multiview Symbolic Regression

How to learn laws from examples

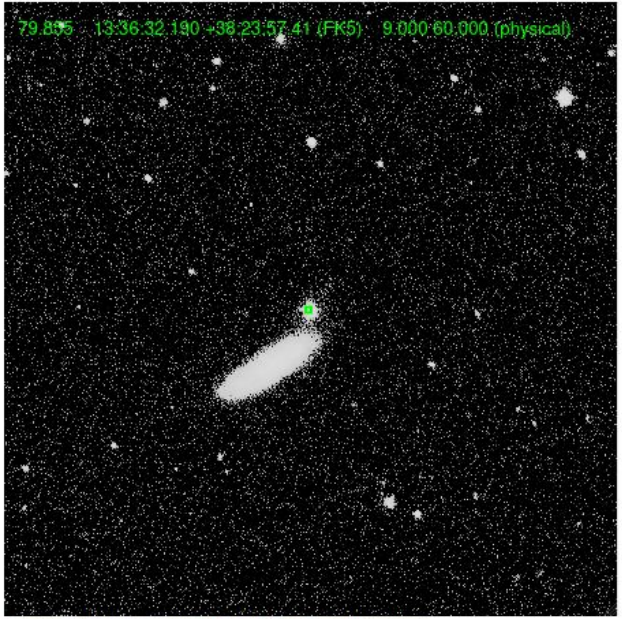
<u>Etienne Russeil</u> - *LPC Université Clermont Auvergne, France*
Fabricio Olivetti - *CMCC Federal University of ABC, Brazil*
Konstantin Malanchev - *University of Illinois Urbana–Champaign, USA*
Emille Ishida - *LPC Université Clermont Auvergne, France*
Emmanuel Gangler - *LPC Université Clermont Auvergne, France*

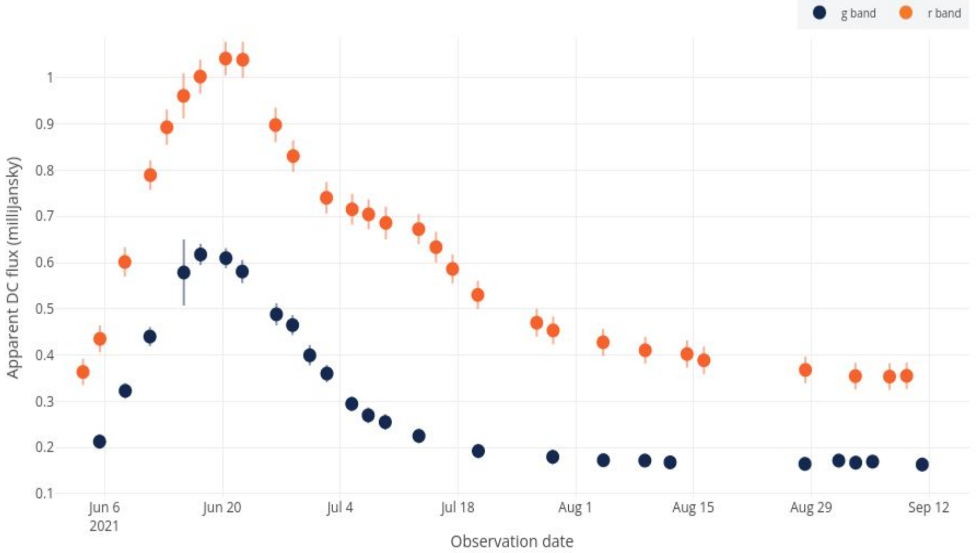
AI for science, science for AI

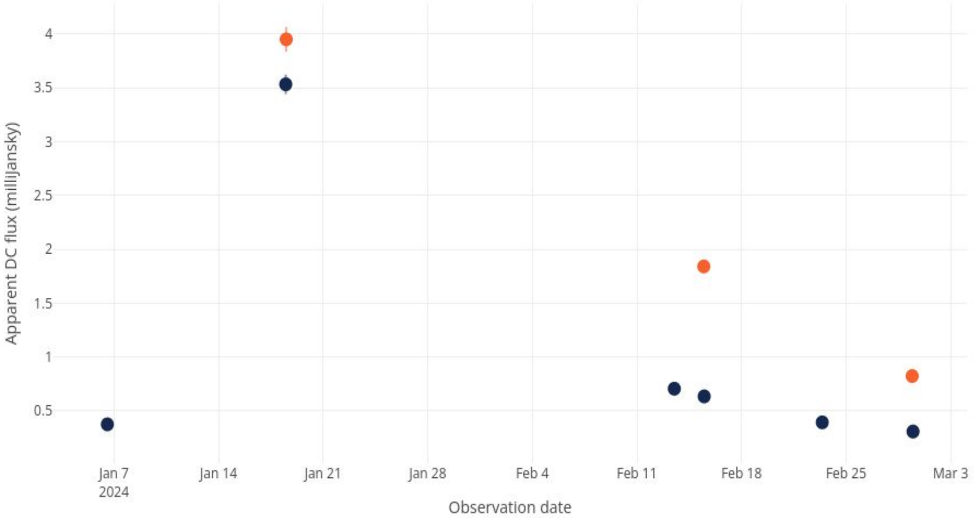
I

Feature extraction

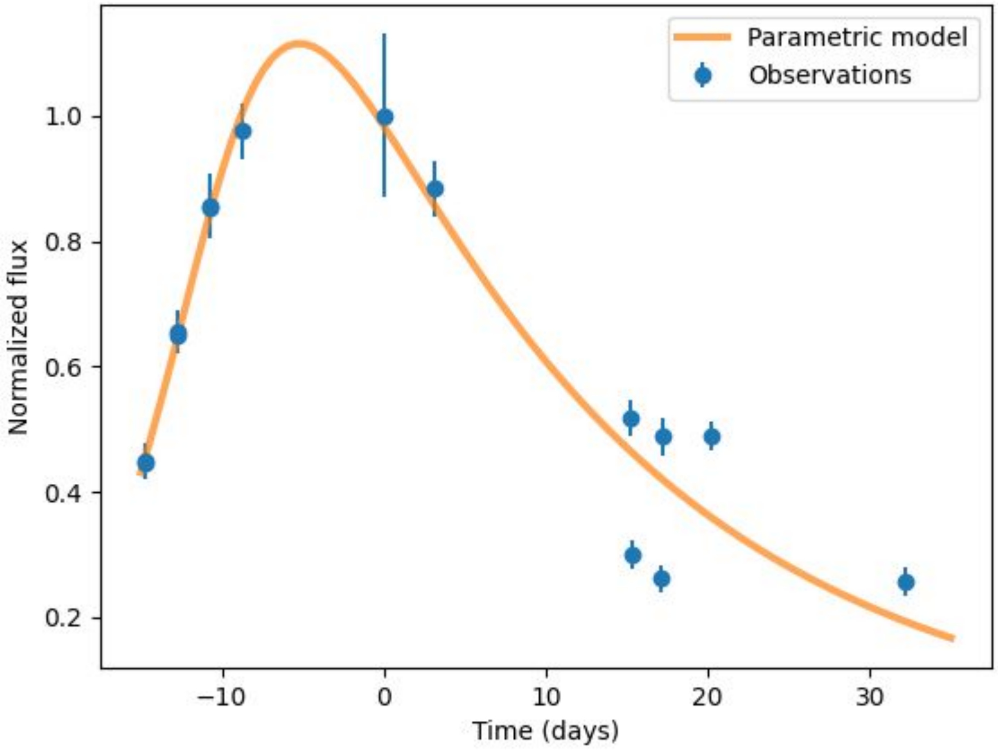
Dealing with irregular sampling (example of astronomy)







One possible solution

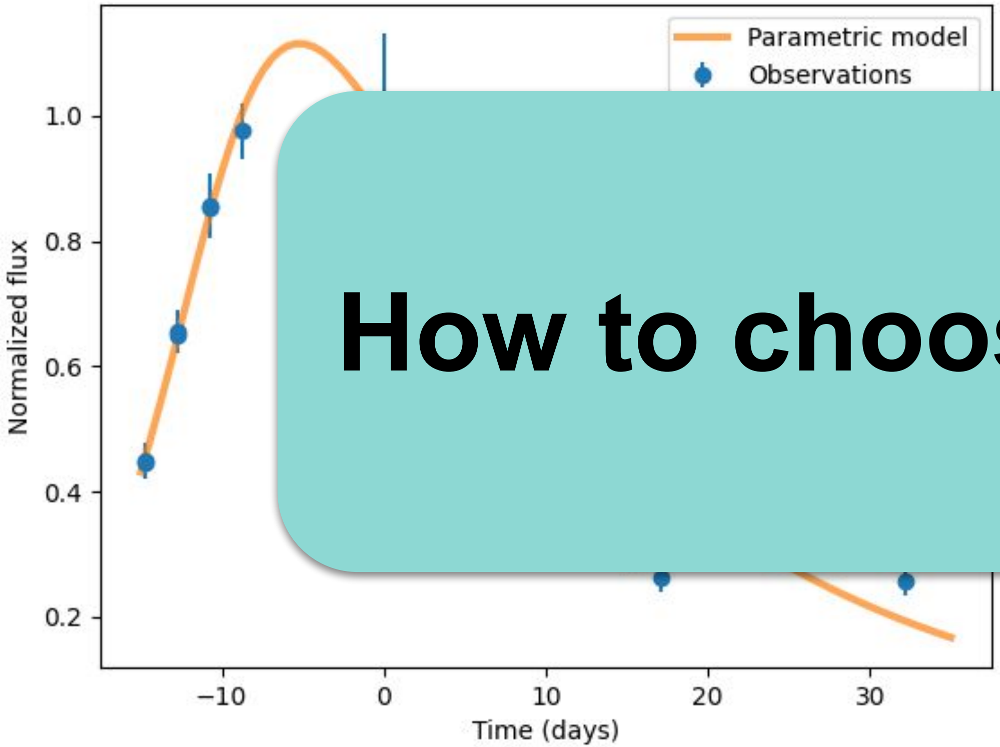


$$f(X; \theta_1, \ldots \theta_n)$$

↓

Extract **n** minimized parameters as features

One possible solution



$$, \ldots \theta_n)$$

How to choose f(X) ?

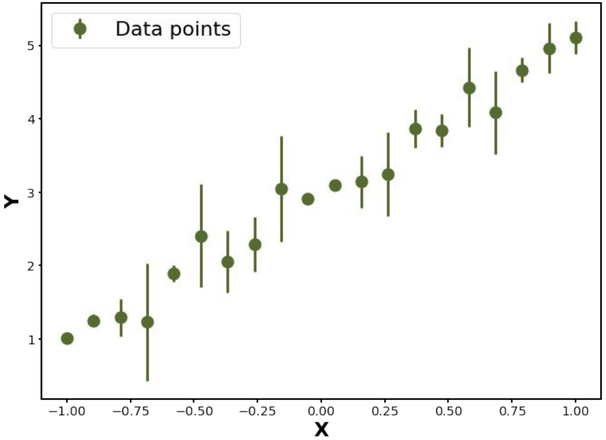
inimized
parameters as features

II

Symbolic Regression

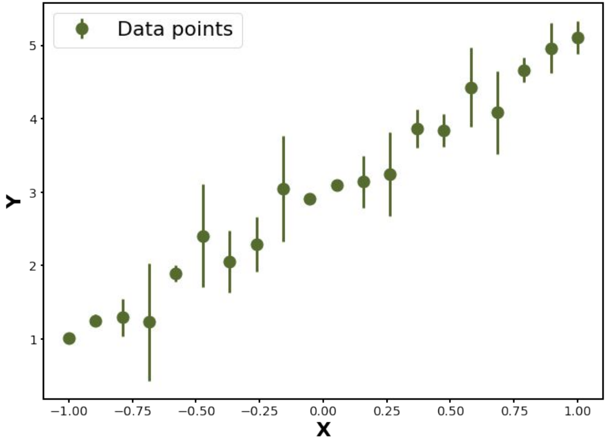
Traditional Symbolic Regression

DATA SET



Traditional Symbolic Regression

DATA SET



RANDOM EQUATIONS

$f(X) = \sin(X) + 2$

$f(X) = X^2 - 1$

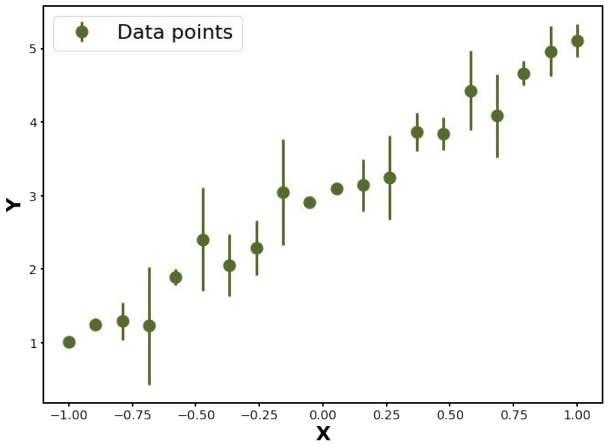
$f(X) = 42$

$f(X) = -4X + 8$

$f(X) = X$

Traditional Symbolic Regression

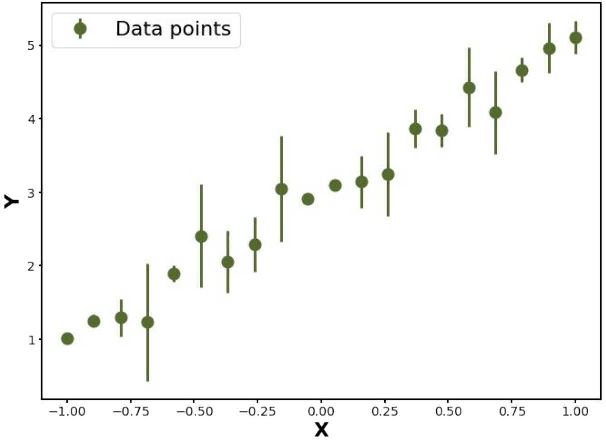
DATA SET



RANDOM EQUATIONS	COST FUNCTION
$f(X) = \sin(X) + 2$	COST = 12
$f(X) = X^2 - 1$	COST = 24
$f(X) = 42$	COST = 43
$f(X) = -4X + 8$	COST = 7
$f(X) = X$	COST = 3

Traditional Symbolic Regression

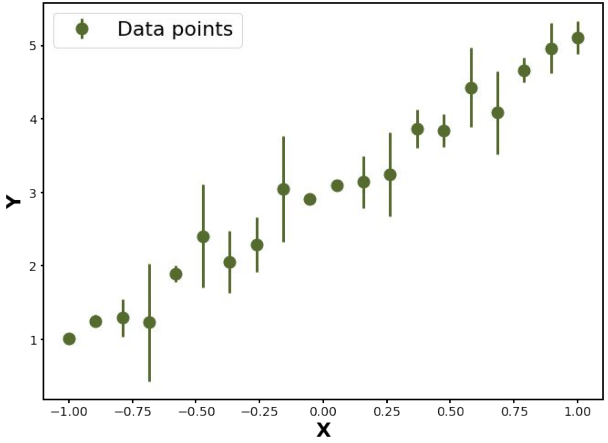
DATA SET



RANDOM EQUATIONS	COST FUNCTION
$f(X) = \sin(X) + 2$	COST = 12
$f(X) = X^2 - 1$	COST = 24
$f(X) = 42$	COST = 43
$f(X) = -4X + 8$	COST = 7
$f(X) = X$	COST = 3

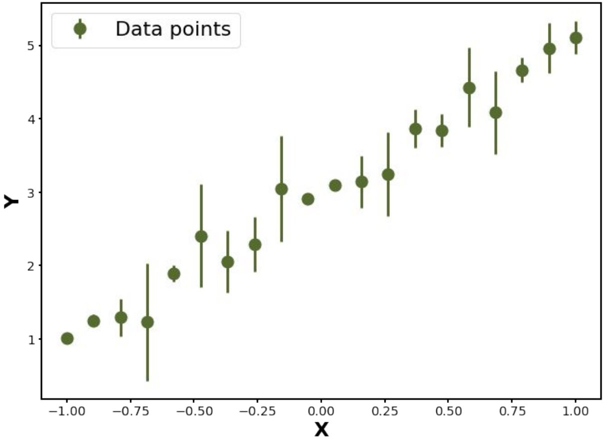
Traditional Symbolic Regression

DATA SET



Traditional Symbolic Regression

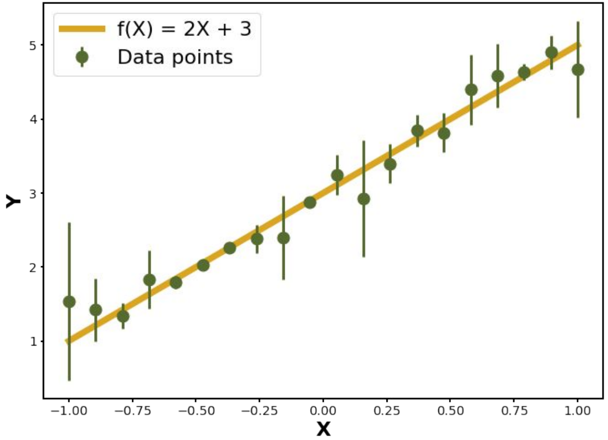
DATA SET



EVOLVED EQUATIONS	COST FUNCTION
f(X) = 2X - 8	COST = 10
f(X) = X	COST = 7
f(X) = 2X + 2	COST = 0.5
f(X) = 2X	COST = 7
f(X) = 1/X	COST = 42

Traditional Symbolic Regression

DATA SET



After many generation

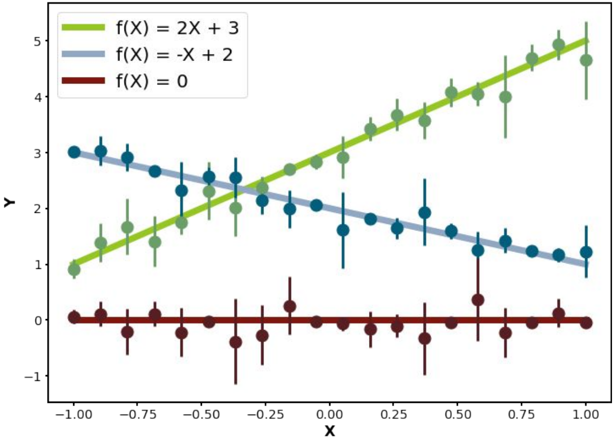
Best answer

f(X) = 2 X + 3

COST ~ 0

Traditional Symbolic Regression

DATA SETS



Best answers

f(X) = 2 X + 3

f(X) = -X + 2

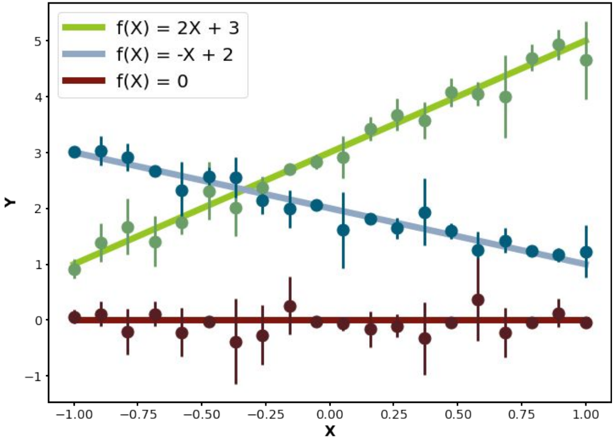
f(X) = 0

III

MultiView Symbolic Regression

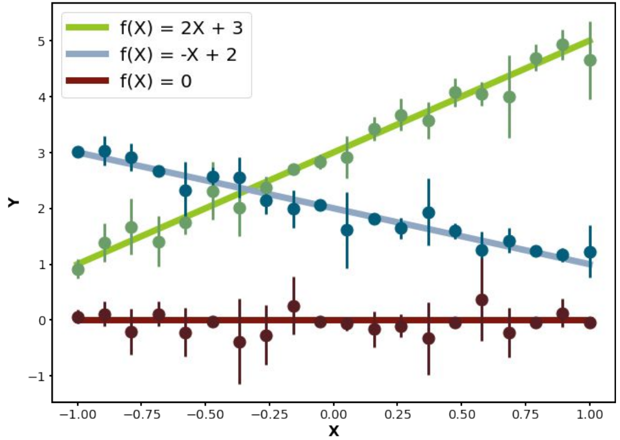
Multiview Symbolic Regression (MvSR)

DATA SETS



Multiview Symbolic Regression (MvSR)

DATA SETS



RANDOM EQUATIONS

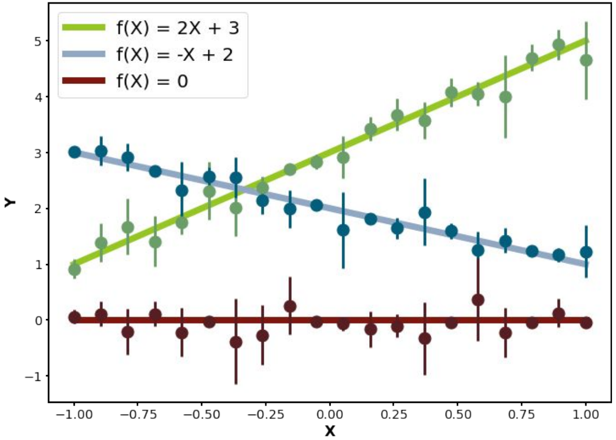
$f(X) = \sin(X) + \mathbf{A}$

$f(X) = \mathbf{A} + \mathbf{B}\, X^2$

$f(X) = \mathbf{A}$

Multiview Symbolic Regression (MvSR)

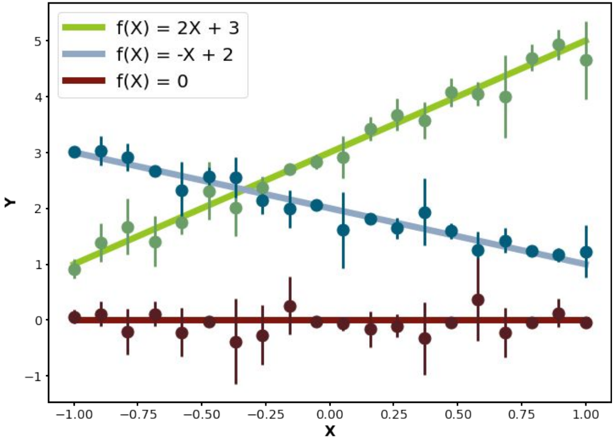
DATA SETS



RANDOM EQUATIONS	COST FUNCTION AFTER MINIMIZATION
f(X) = sin(X) + **A**	COST = 24, 32, 7
f(X) = **A** + **B** X^2	COST = 17, 8, 0
f(X) = **A**	COST = 19, 10, 0

Multiview Symbolic Regression (MvSR)

DATA SETS

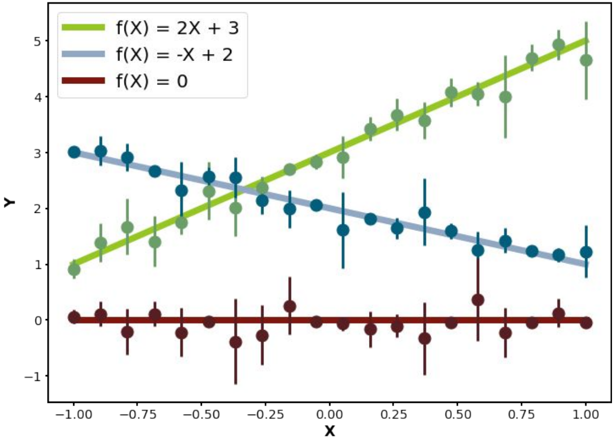


RANDOM EQUATIONS	COST FUNCTION AFTER MINIMIZATION
$f(X) = \sin(X) + \mathbf{A}$	COST = 24, 32, 7
$f(X) = \mathbf{A} + \mathbf{B}\,X^2$	COST = 17, 8, 0
$f(X) = \mathbf{A}$	COST = 19, 10, 0

Average COST

Multiview Symbolic Regression (MvSR)

DATA SETS



After many generation

Best answer

f(X) = A X + B

COST = 0 0 0

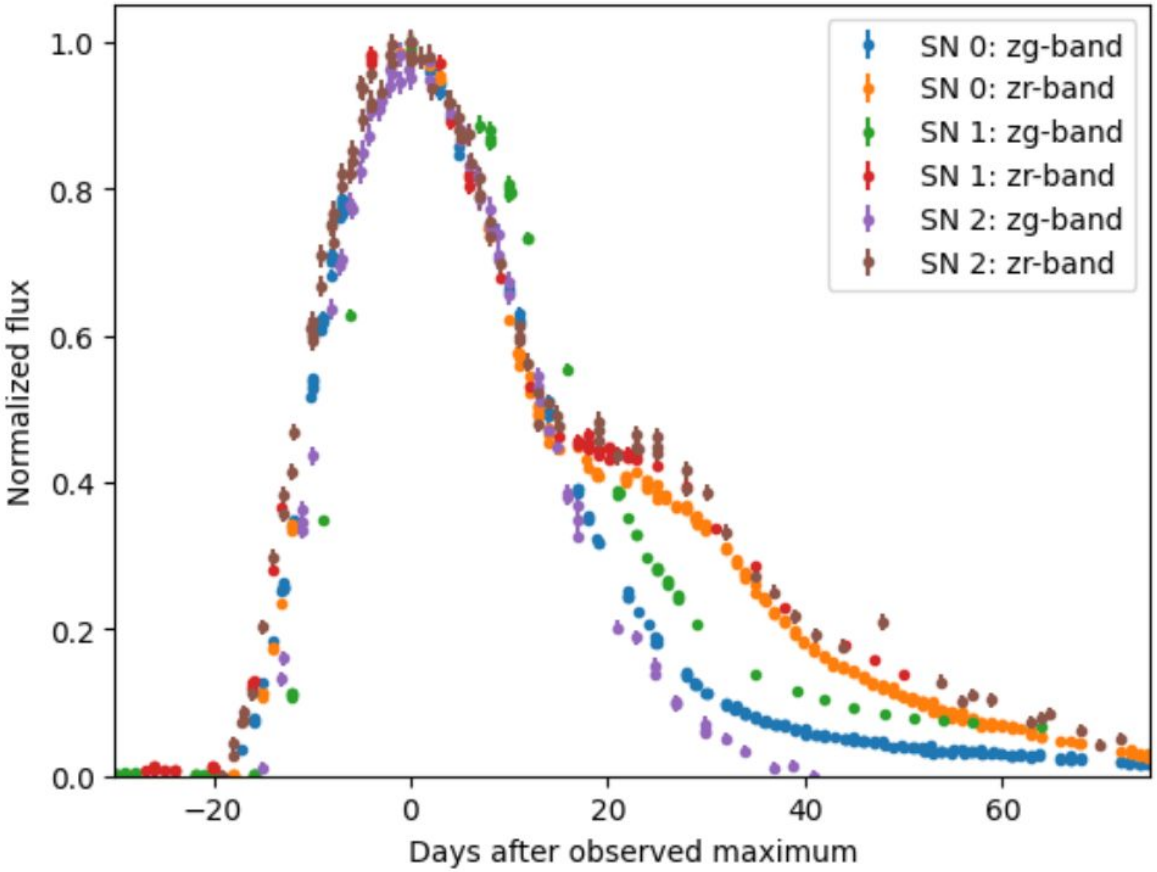
MvSR in a nutshell

- (1) Receive multiple datasets as input.
- (2) Perform a minimization of the parameters independently
- for each dataset.
- (3) Use an aggregation function to compute an overall loss.
- (4) Allow parameters to be repeated.
- (5) Control the maximum number of parameters.
- (6) Penalise solutions based on the number of parameters used

IV

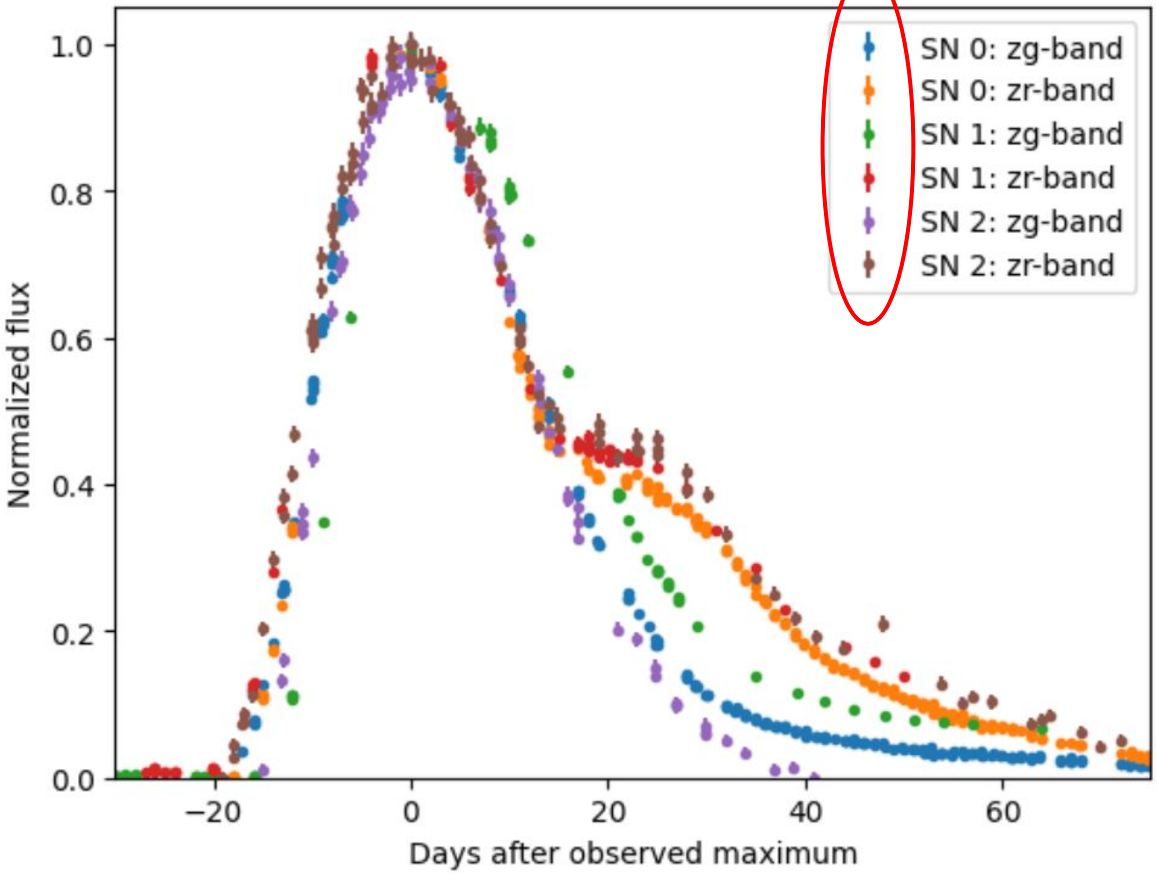
Scientific applications

Astrophysical database

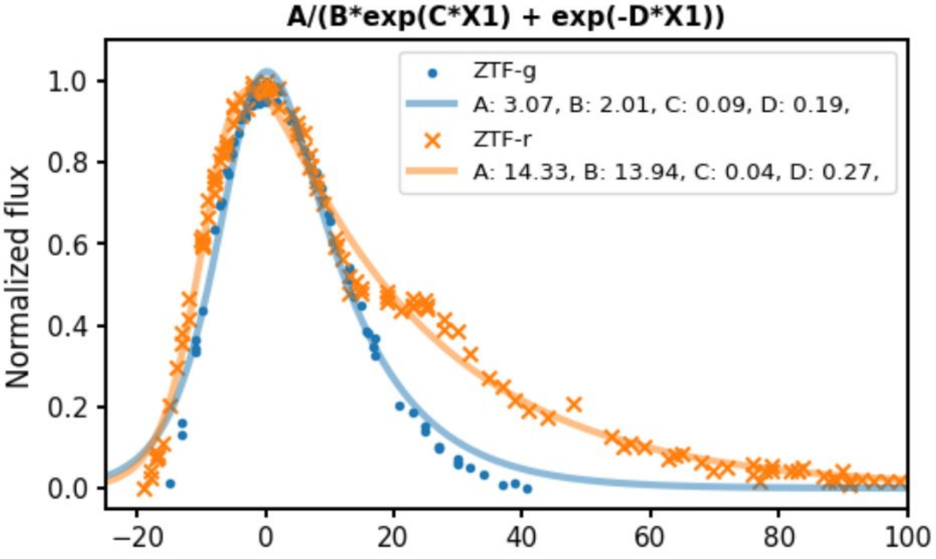


Astrophysical database

Error bars not used :(

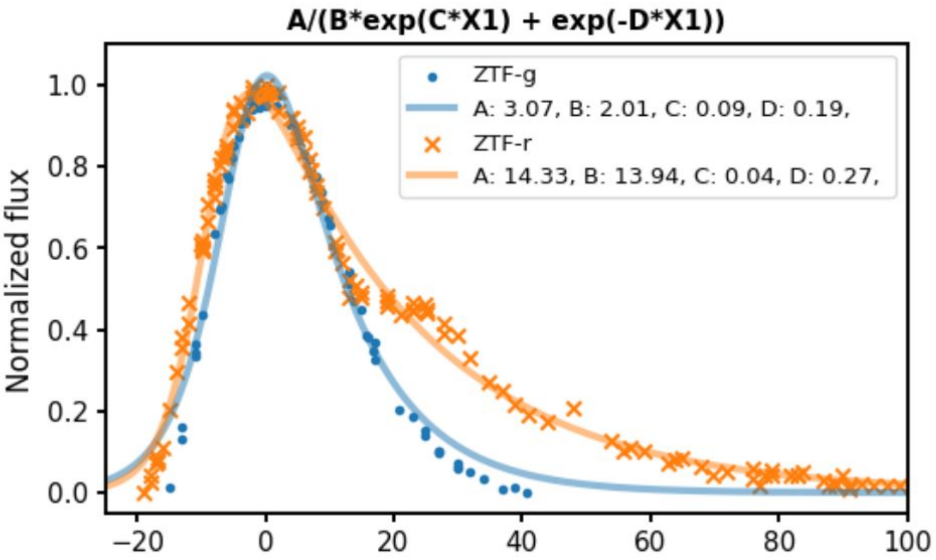


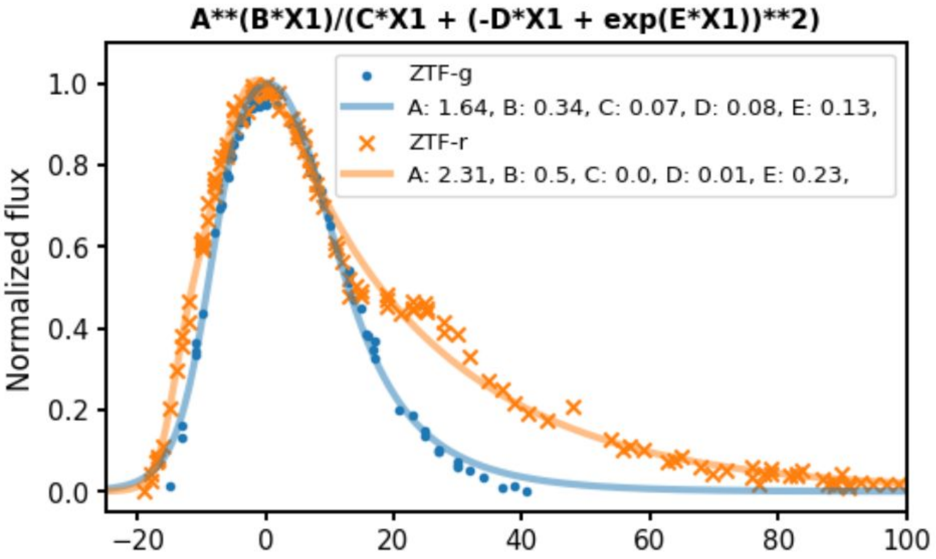
Astrophysical database



MvSR recovers the literature !

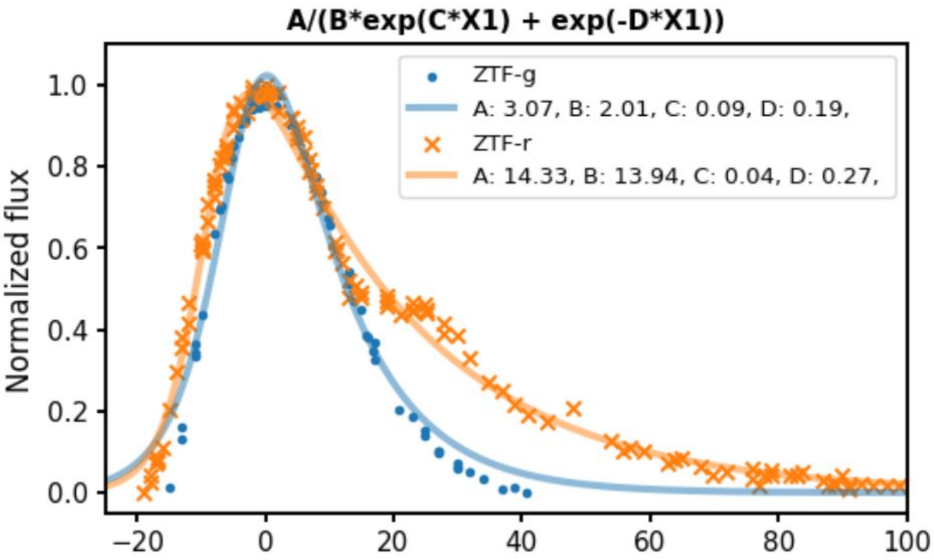
Astrophysical database

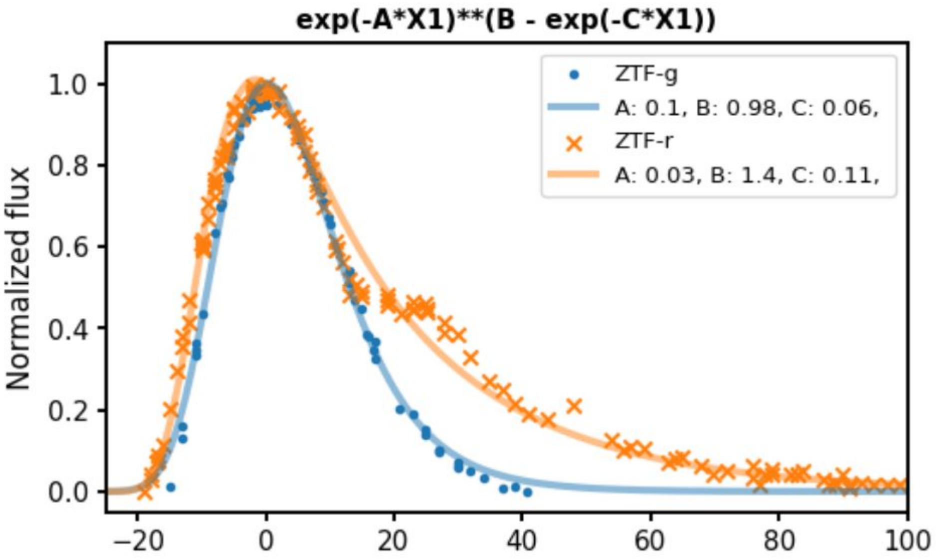




It can also generate more complex and better solutions

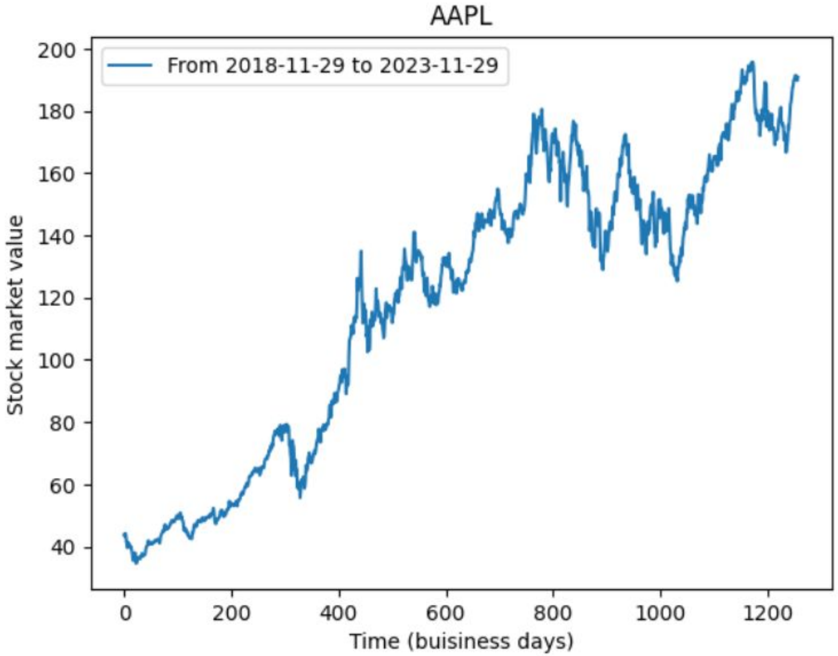
Astrophysical database



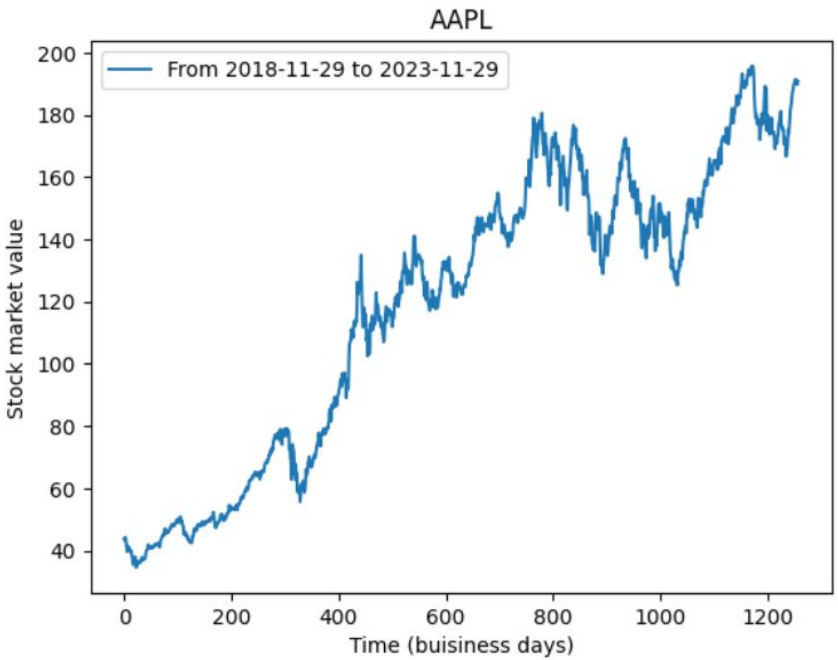


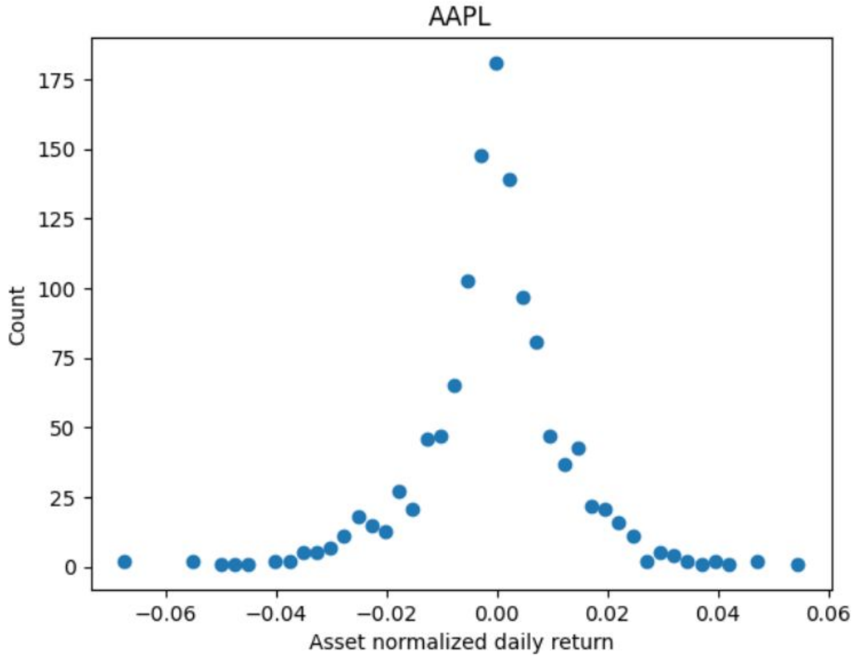
As well as unexpected but very effective forms

Finance database



Finance database





Finance database

S&P500

Stock market of the 500 biggest US companies

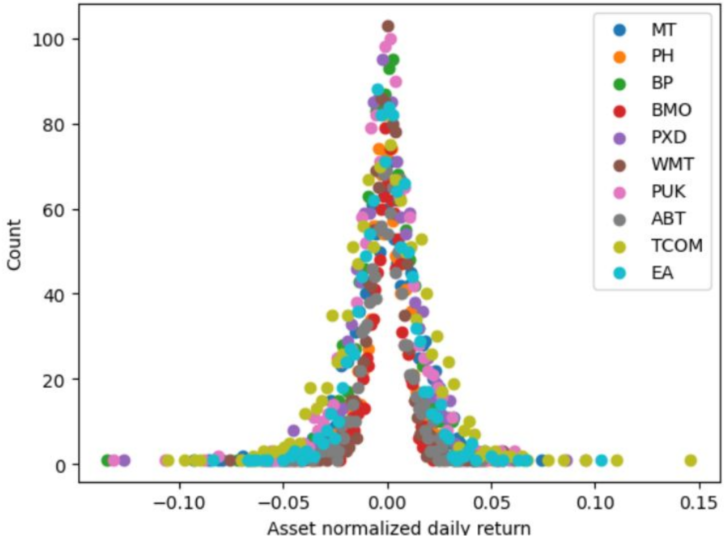
Usually aggregated for analysis

Finance database

S&P500

Stock market of the 500 biggest US companies

10 random companies



Finance database

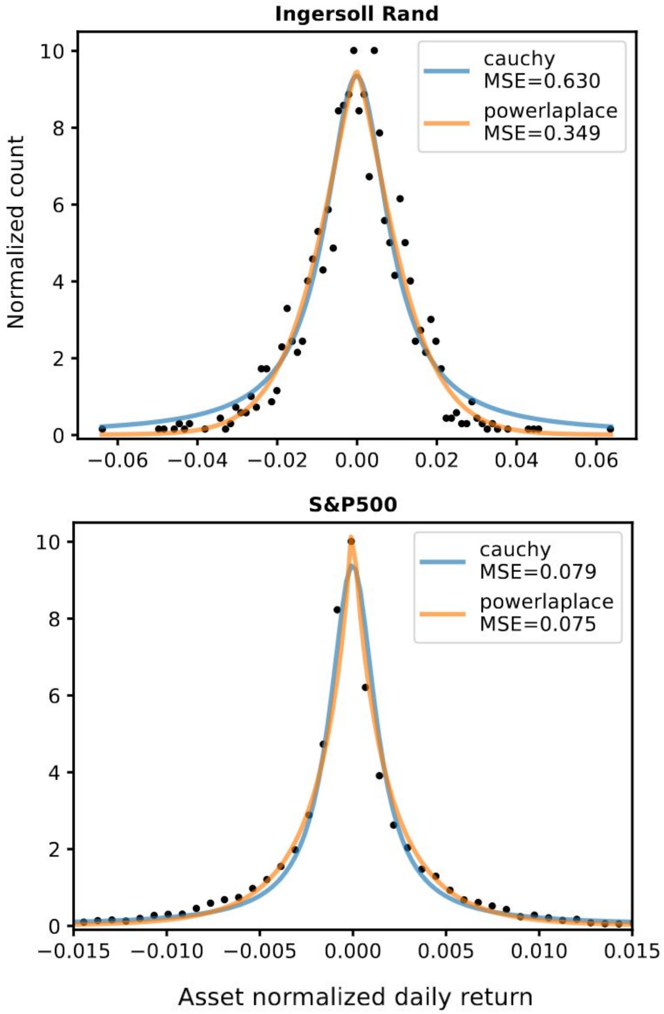
Recover literature

Models	Equation f(x)
Gaussian [2, 5]	$A \cdot e^{-\frac{x^2}{B}}$
Laplace [17]	$A \cdot e^{-B\|x\|}$
Cauchy [20]	$A \cdot B^2/(x^2 + B^2)$
Linear-Laplace	$(A - Bx) \cdot e^{-C\|x\|}$
Exp-Laplace	$A \cdot e^{Bx-C\|x\|}$
Power-Laplace	$A \cdot e^{B\|x\|^C}$

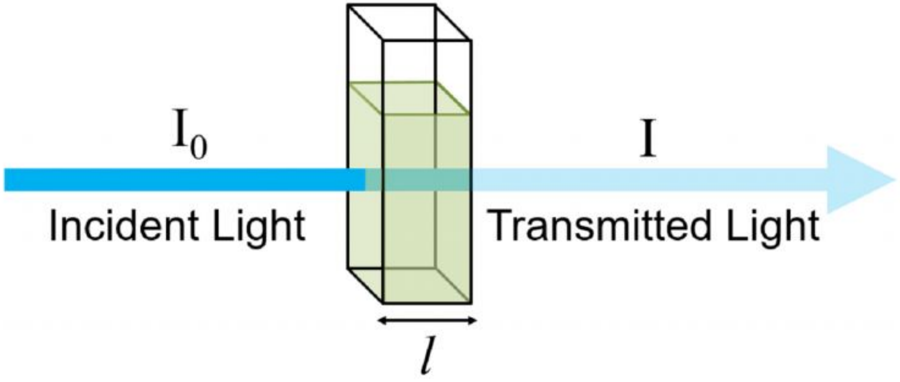
Find new models

Finance database

Models	Equation f(x)
Gaussian [2, 5]	$A \cdot \mathrm{e}^{-\frac{x^2}{B}}$
Laplace [17]	$A \cdot \mathrm{e}^{-B\lvert x\rvert}$
Cauchy [20]	$A \cdot B^2/(x^2 + B^2)$
Linear-Laplace	$(A - Bx) \cdot \mathrm{e}^{-C\lvert x\rvert}$
Exp-Laplace	$A \cdot \mathrm{e}^{Bx - C\lvert x\rvert}$
Power-Laplace	$A \cdot \mathrm{e}^{B\lvert x\rvert^C}$

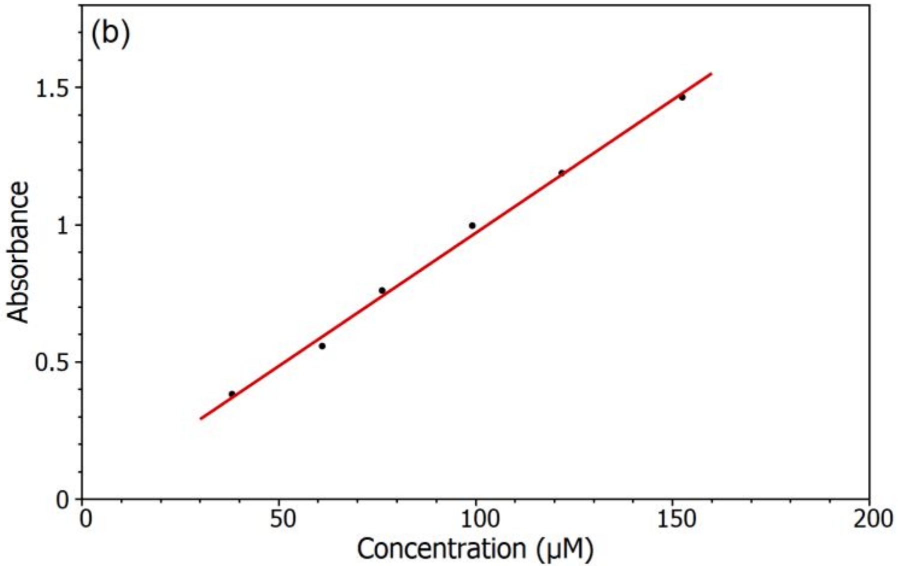


Chemistry database



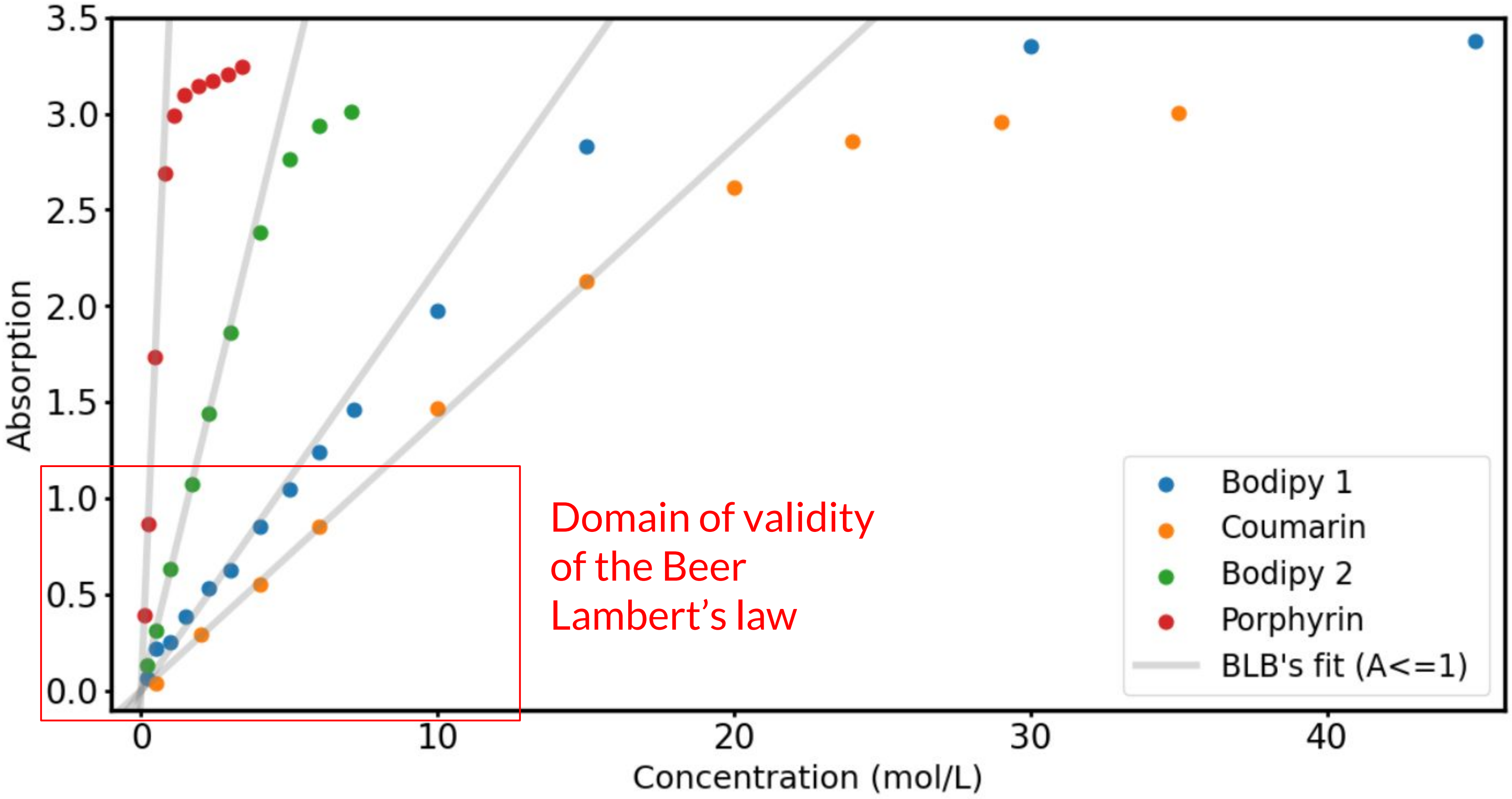
$$log_{10}(\frac{I_0}{I}) = A = k \cdot C$$

Beer Lambert's law

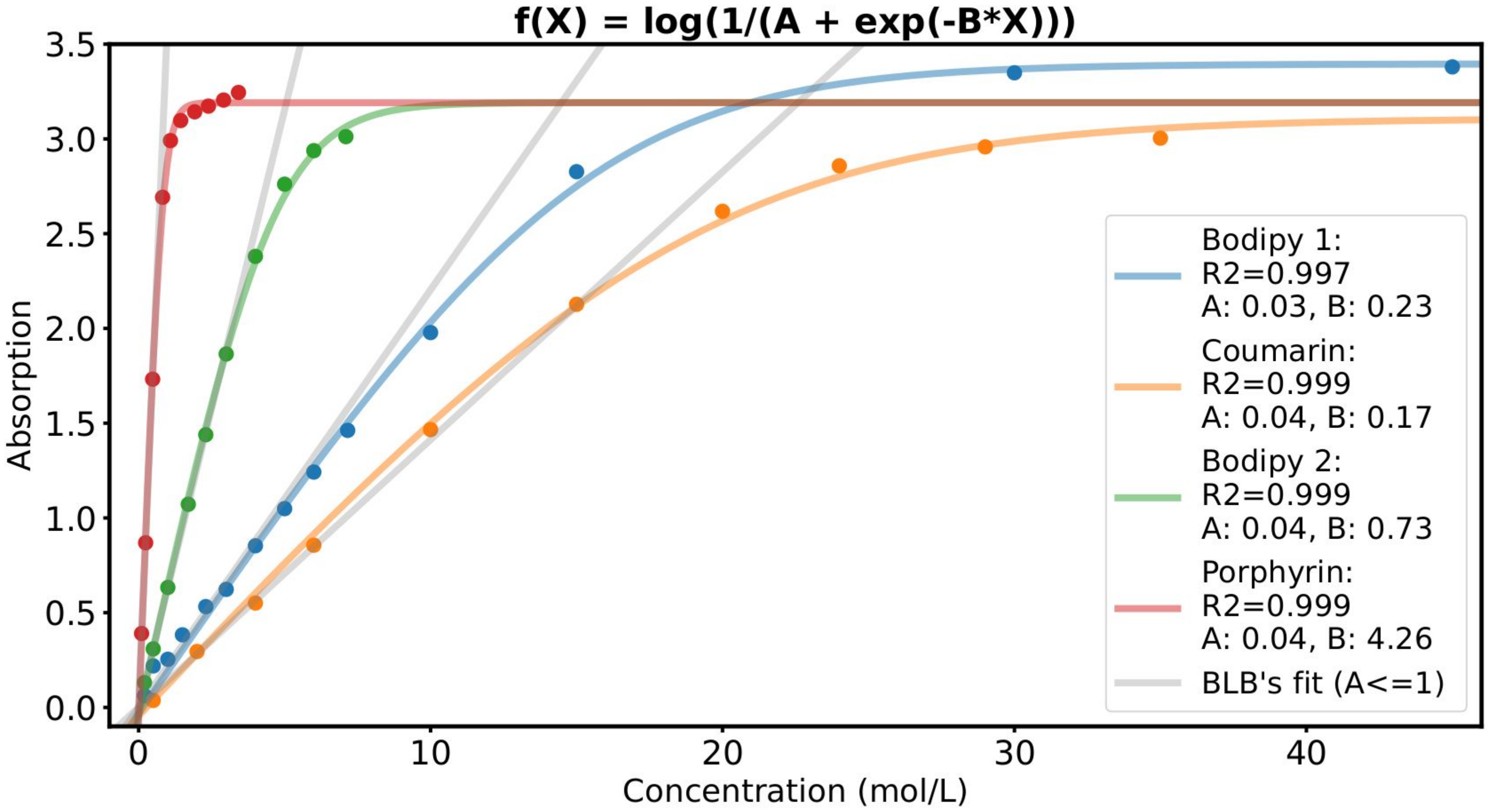


https://www.edinst.com/blog/the-beer-lambert-law/

Chemistry database



Chemistry database



Chemistry database

General Beer Lambert's law:

$$f(x; \mu, \epsilon) = \log\left(\frac{1}{\mu + \mathrm{e}^{-\epsilon x}}\right)$$

Saturation parameter

Linear slope

V

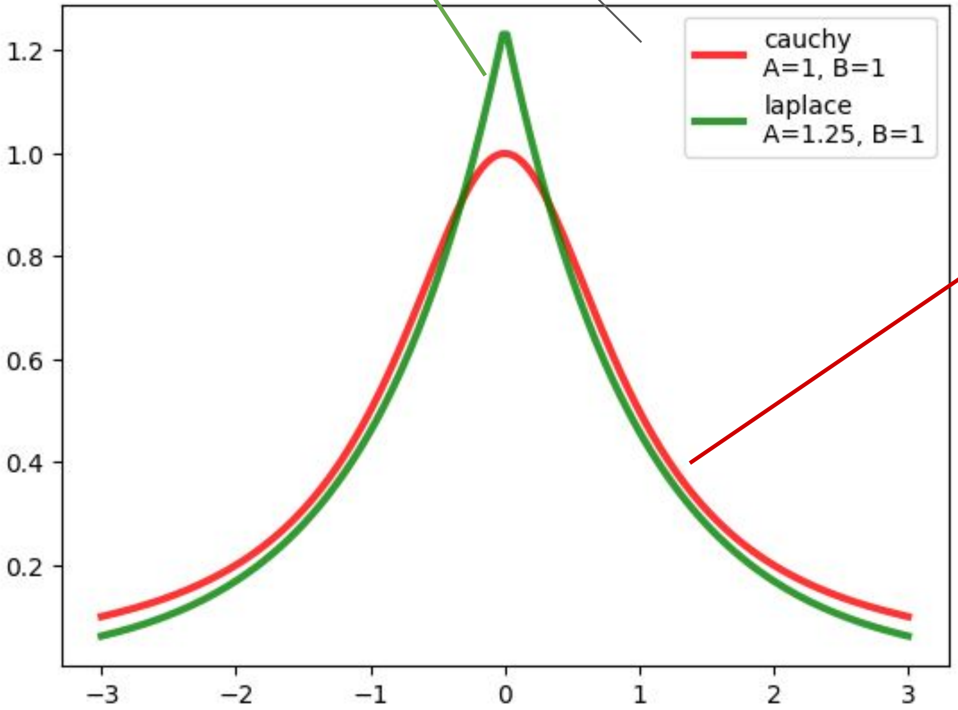
Simple anomaly detection

Generate toy data

100

Hidden anomalies

$$f(X) = A \cdot e^{-B \cdot |X|}$$

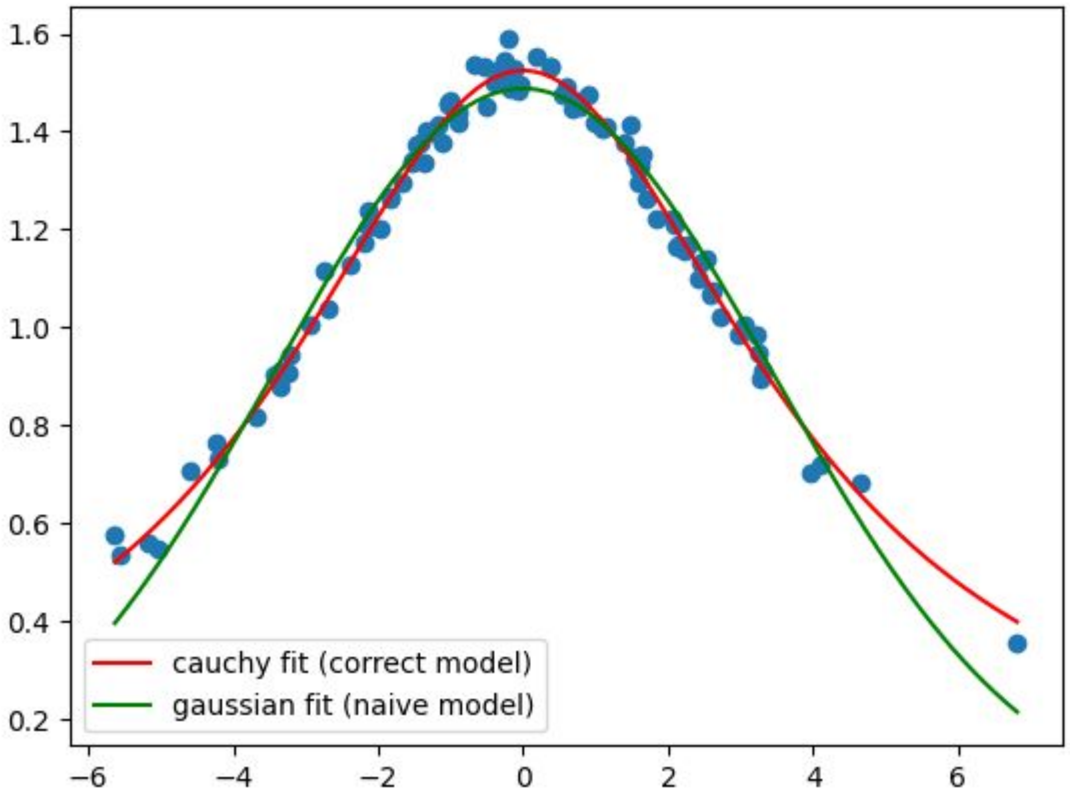


5000

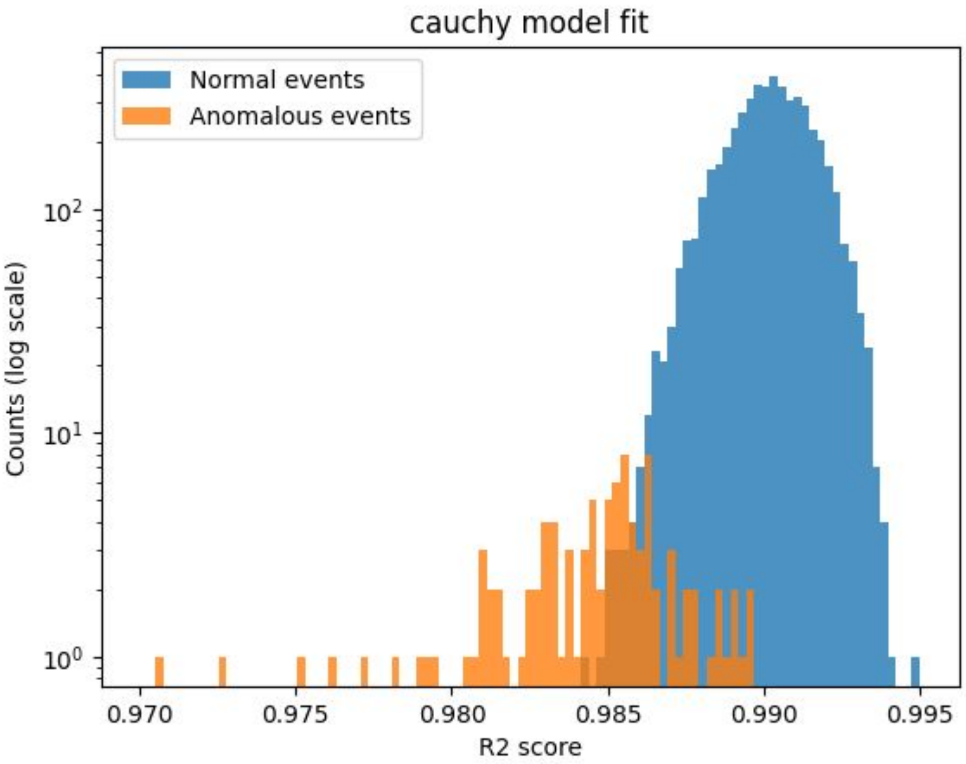
Normal behavior

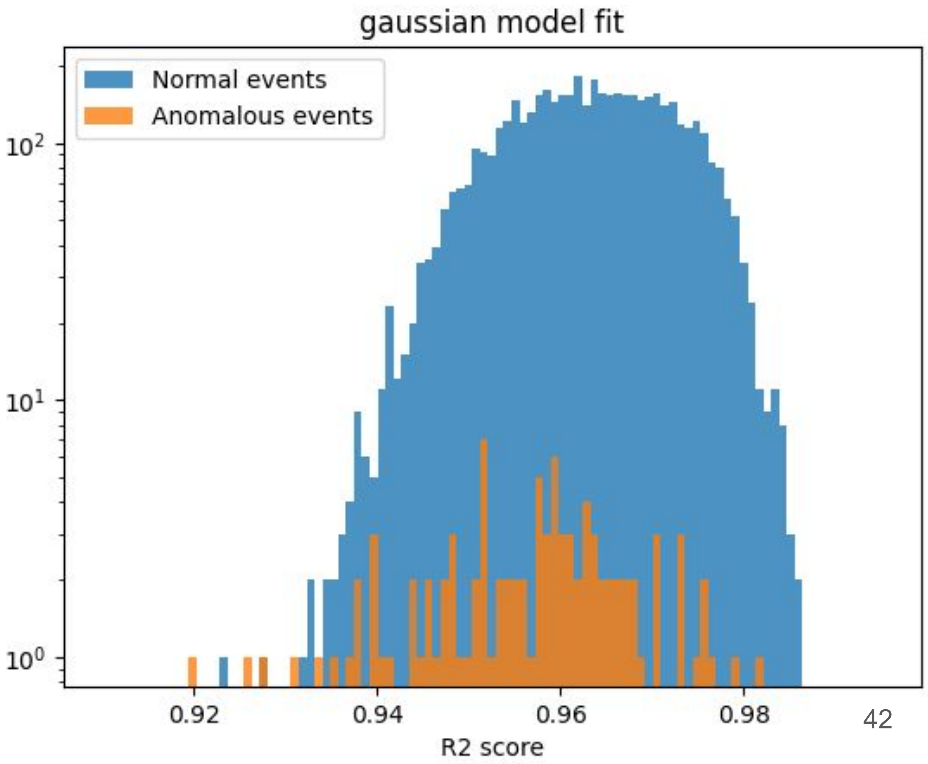
$$f(X) = A \cdot \frac{B^2}{X^2 + B^2}$$

Data example



Simple isolation forest results





VI

Conclusion

Conclusion

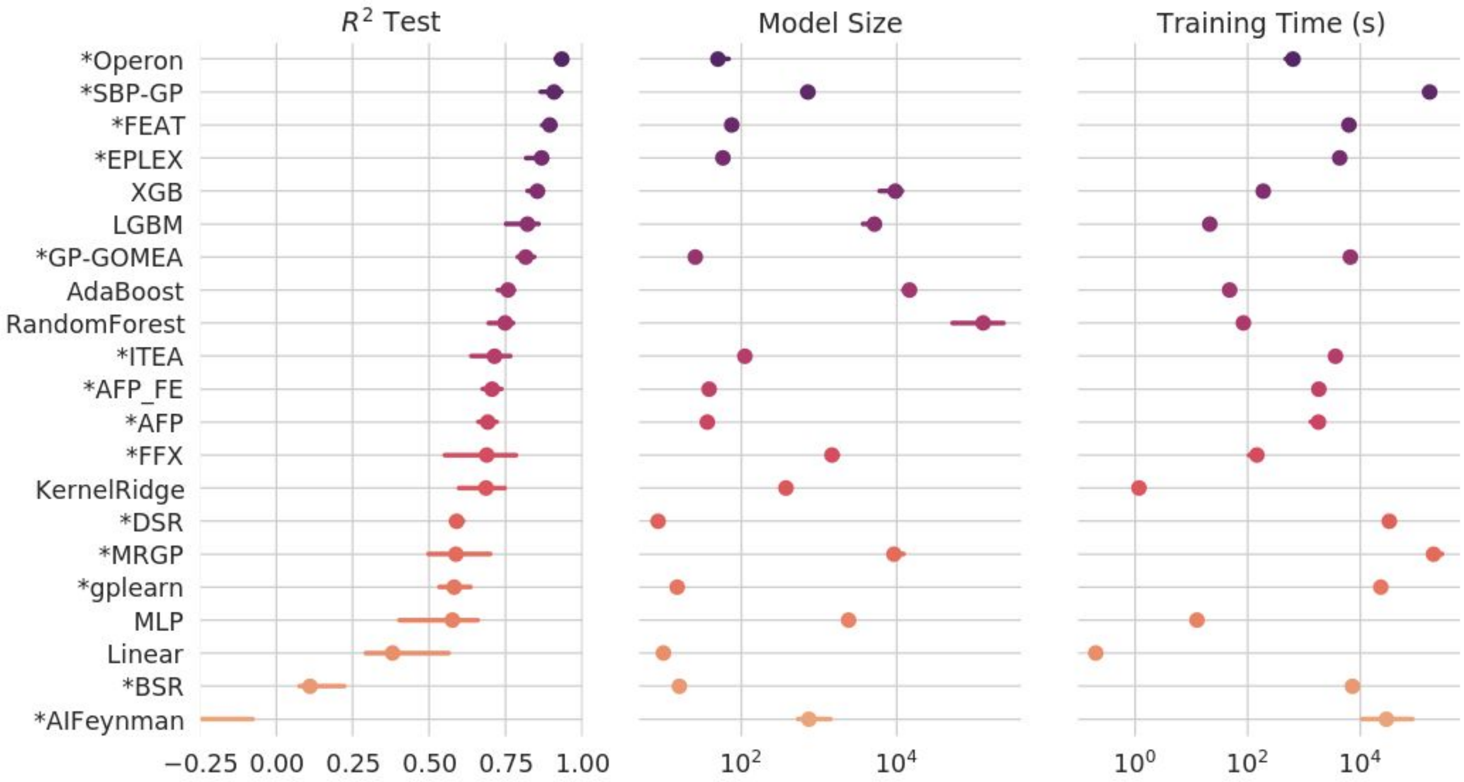
- **MvSR is working**, have a look at the arXiv
- It has potential to be used in **every science**
- It represent the first step of **future anomaly detection** studies
- Still need some work on our side for a proper full implementation

https://arxiv.org/abs/2402.04298

AI for science, science for AI

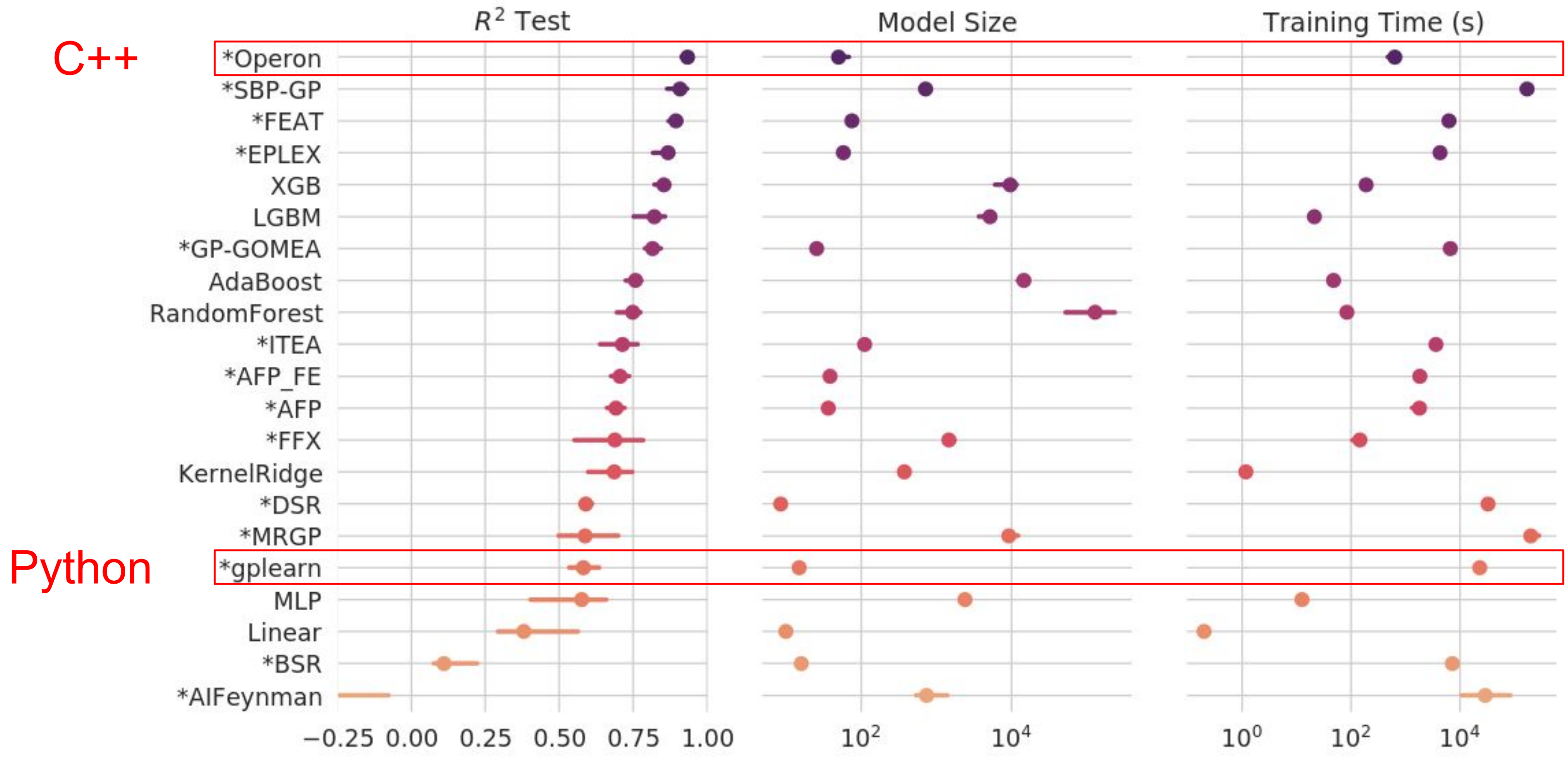
BACKUP SLIDES

Multiview Symbolic Regression (MvSR)



LaCava et al., 2021. Contemporary Symbolic Regression Methods and their Relative Performance, arXiv:cs https://arxiv.org/abs/2107.14351

Multiview Symbolic Regression (MvSR)



LaCava et al., 2021. *Contemporary Symbolic Regression Methods and their Relative Performance*, arXiv:cs https://arxiv.org/abs/2107.14351

[Submitted on 6 Feb 2024]

Multi-View Symbolic Regression

Etienne Russeil, Fabrício Olivetti de França, Konstantin Malanchev, Bogdan Burlacu, Emille E. O. Ishida, Marion Leroux, Clément Michelin, Guillaume Moinard, Emmanuel Gangler

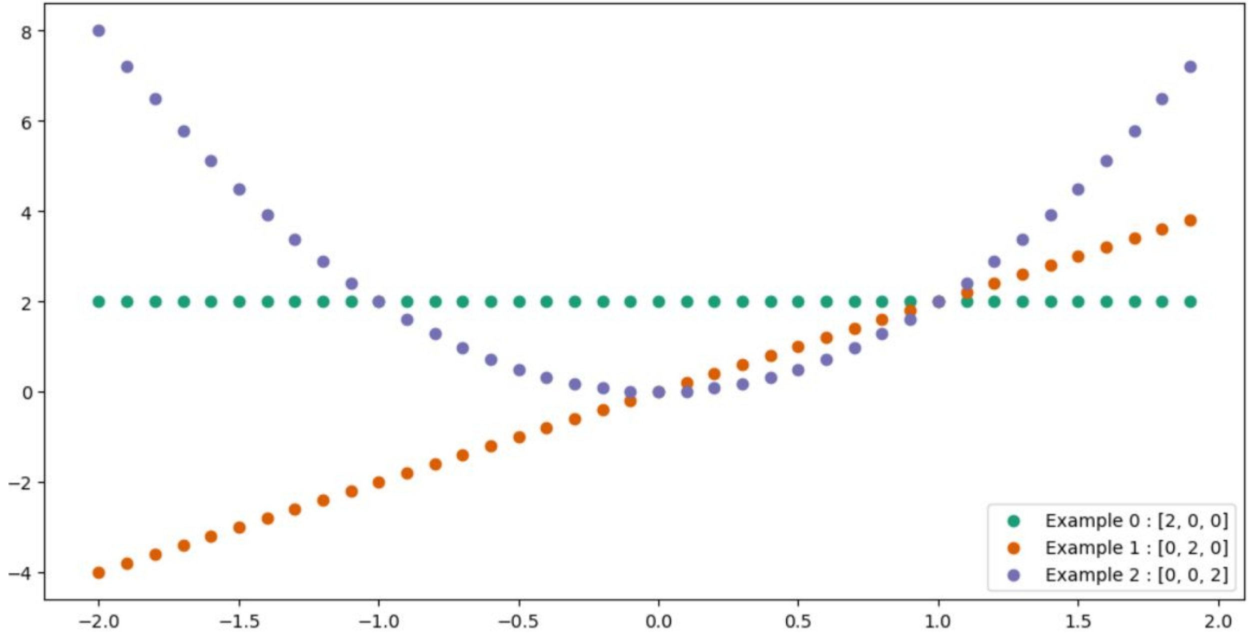
Symbolic regression (SR) searches for analytical expressions representing the relationship between a set of explanatory and response variables. Current SR methods assume a single dataset extracted from a single experiment. Nevertheless, frequently, the researcher is confronted with multiple sets of results obtained from experiments conducted with different setups. Traditional SR methods may fail to find the underlying expression since the parameters of each experiment can be different. In this work we present Multi-View Symbolic Regression (MvSR), which takes into account multiple datasets simultaneously, mimicking experimental environments, and outputs a general parametric solution. This approach fits the evaluated expression to each independent dataset and returns a parametric family of functions f(x; \theta) simultaneously capable of accurately fitting all datasets. We demonstrate the effectiveness of MvSR using data generated from known expressions, as well as real-world data from astronomy, chemistry and economy, for which an a priori analytical expression is not available. Results show that MvSR obtains the correct expression more frequently and is robust to hyperparameters change. In real-world data, it is able to grasp the group behaviour, recovering known expressions from the literature as well as promising alternatives, thus enabling the use SR to a large range of experimental scenarios.

Comments: Submitted to GECCO-2024. 10 pages, 6 figures
Subjects: **Machine Learning (cs.LG)**; Instrumentation and Methods for Astrophysics (astro-ph.IM); Applications (stat.AP)
Cite as: arXiv:2402.04298 **[cs.LG]**
(or arXiv:2402.04298v1 **[cs.LG]** for this version)

On arXiv since yesterday: https://arxiv.org/abs/2402.04298

Multiview Symbolic Regression (MvSR)

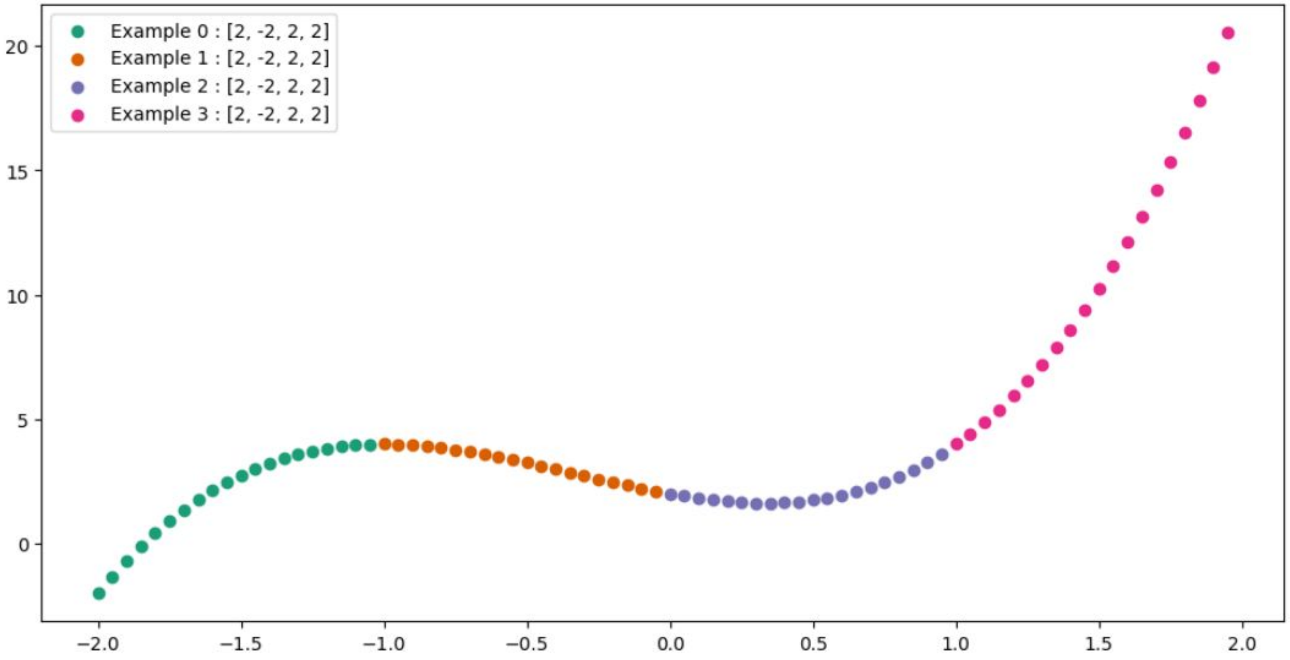
Toy data illustration



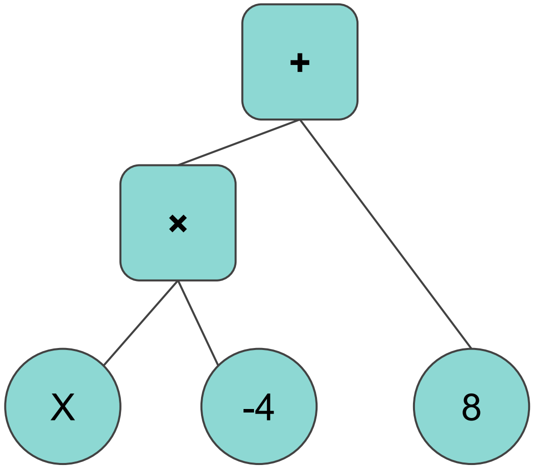
$$f(X) = A + BX + CX^2$$

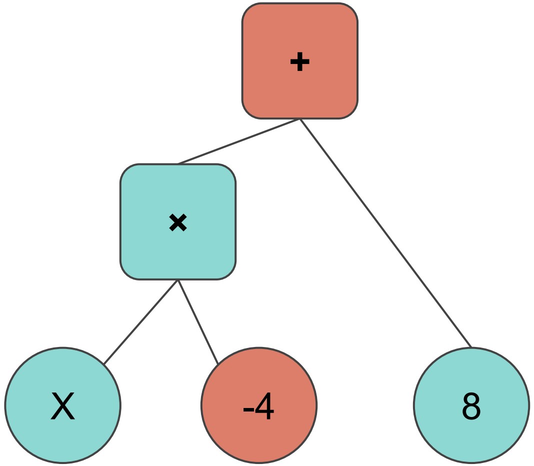
Multiview Symbolic Regression (MvSR)

Toy data illustration

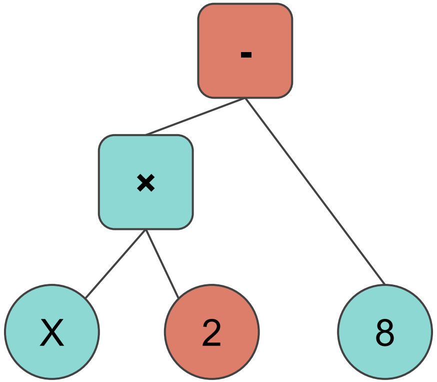


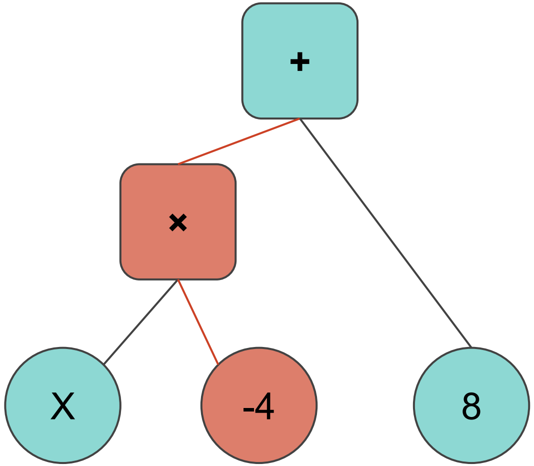
$$f(X) = A + BX + CX^2 + DX^3$$


+
×
X
-4
8

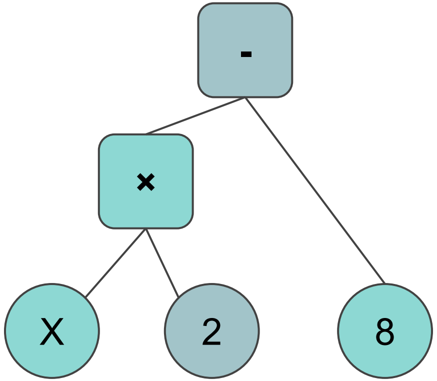


Point mutations

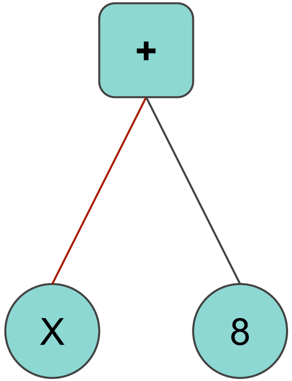


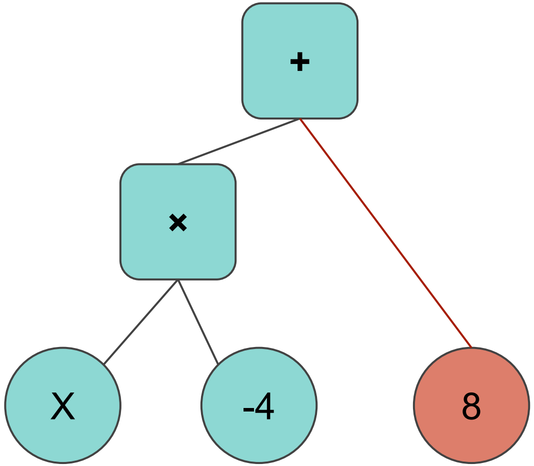


Point mutations

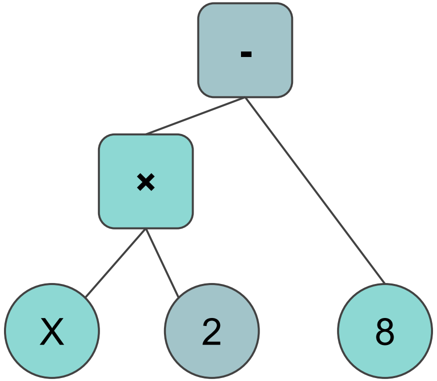


Hoist mutations

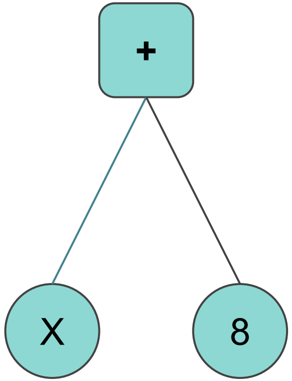




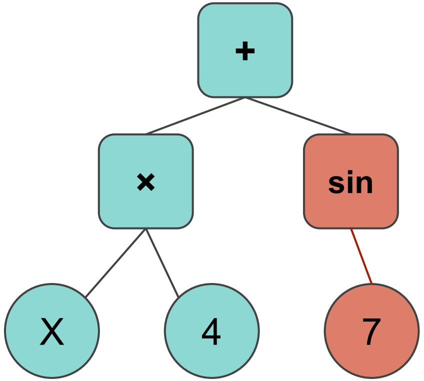
Point mutations

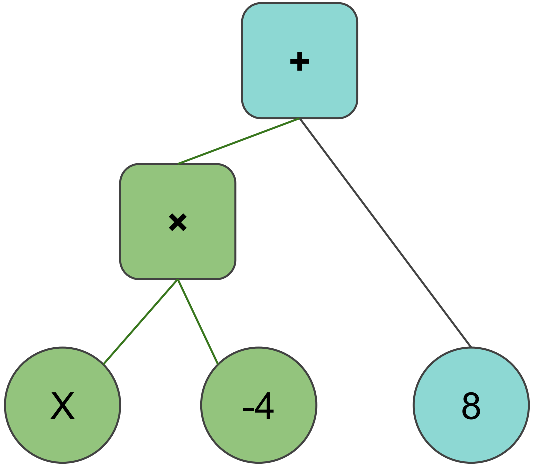


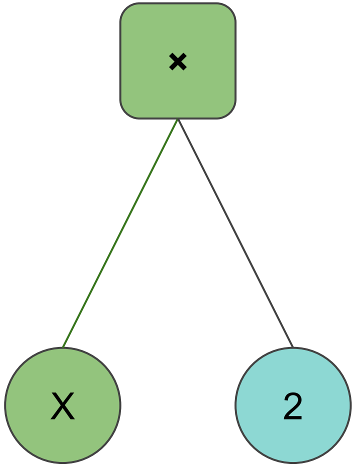
Hoist mutations



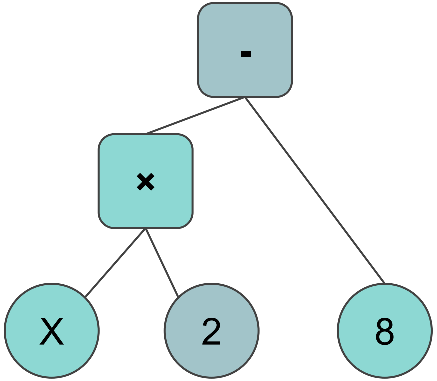
Subtree mutations



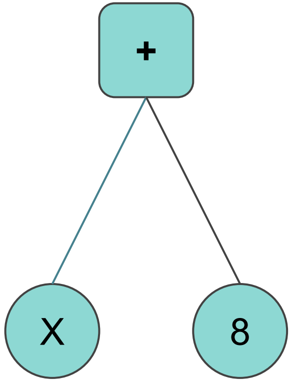




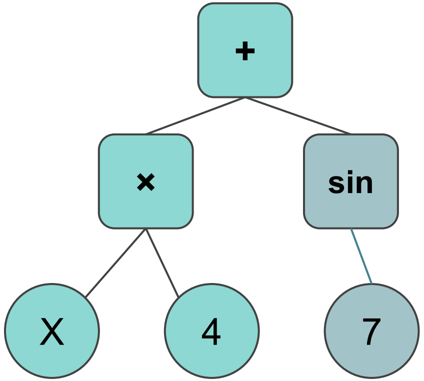
Point mutations



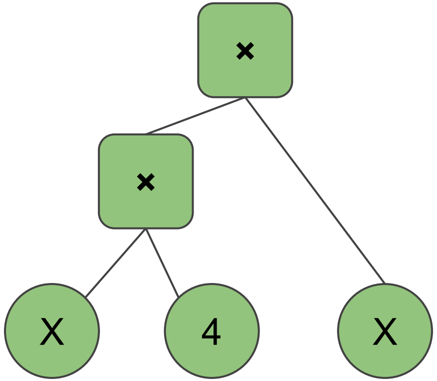
Hoist mutations



Subtree mutations

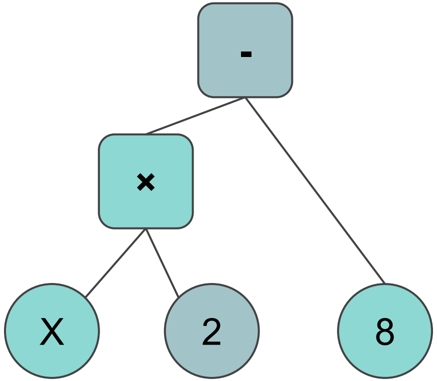


Crossover

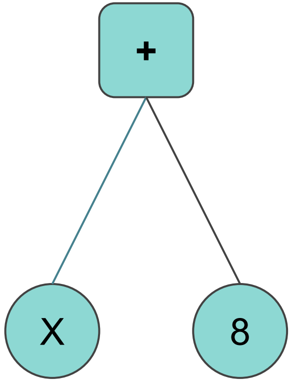


Create a new population from the previous best candidates

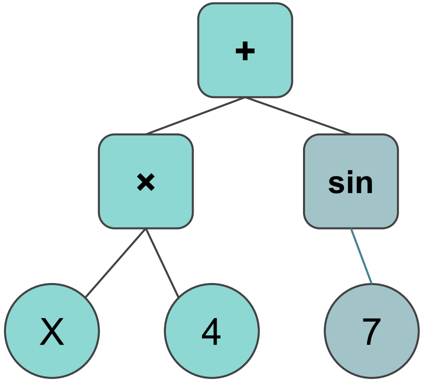
Point mutations



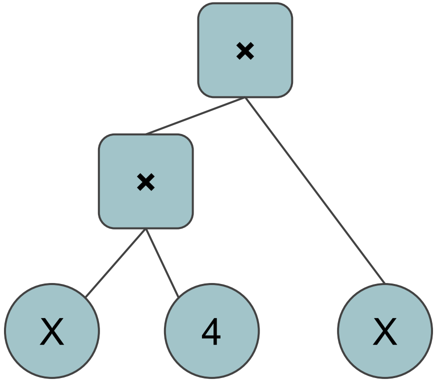
Hoist mutations



Subtree mutations



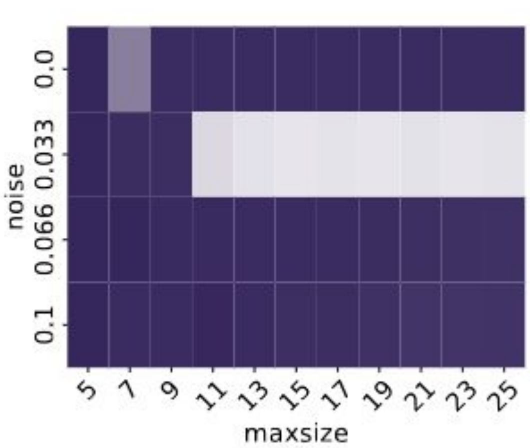
Crossover



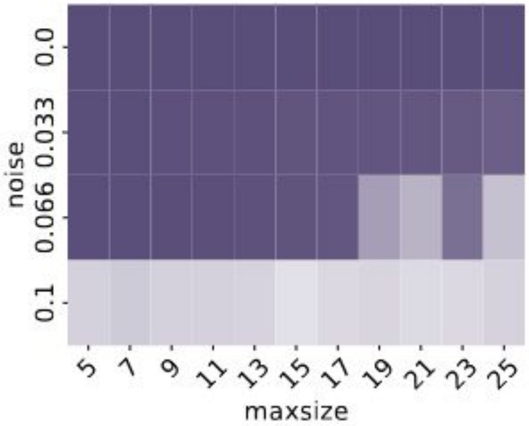
$$f_1(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$$

$$f_2(x) = \sin(\theta_0 x_0 x_1) + \theta_1 (x_2 - \theta_2)^2 + \theta_3 x_3 + x_4$$

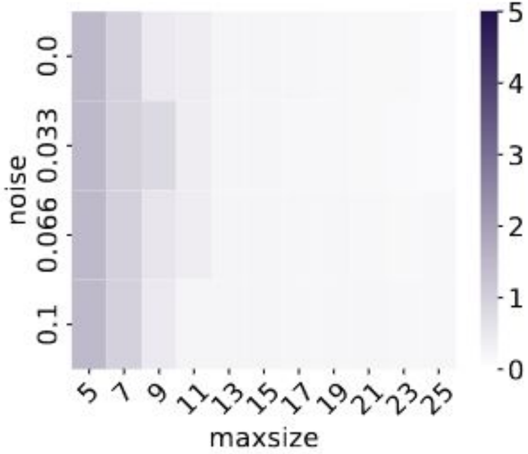
$$f_3(x) = (\theta_0 x_0^2 + (\theta_1 x_1 x_2 - \frac{\theta_2}{(\theta_3 x_1 x_3 + 1)})^2)^{0.5}$$


noise
0.0
0.033
0.066
0.1
maxsize
5 7 9 11 13 15 17 19 21 23 25

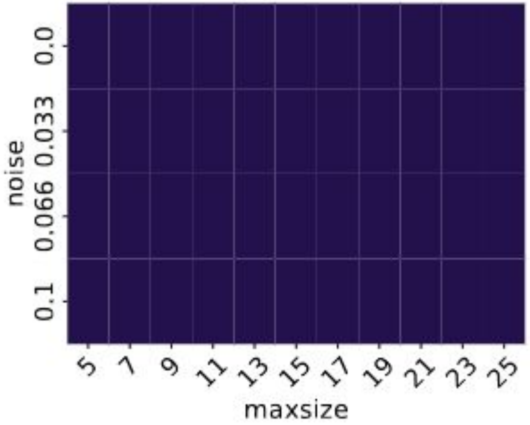
(g)


noise
0.0
0.033
0.066
0.1
maxsize
5 7 9 11 13 15 17 19 21 23 25

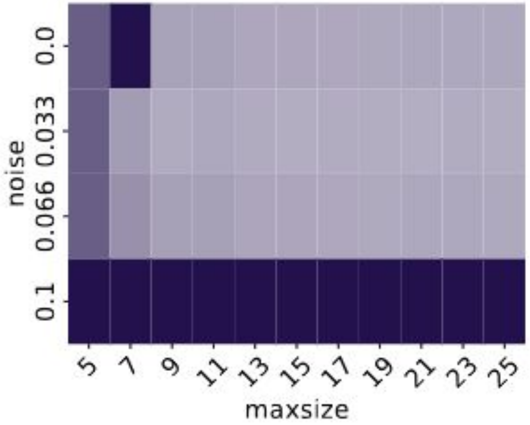
(h)


noise
0.0
0.033
0.066
0.1
maxsize
5 7 9 11 13 15 17 19 21 23 25
5
4
3
2
1
0

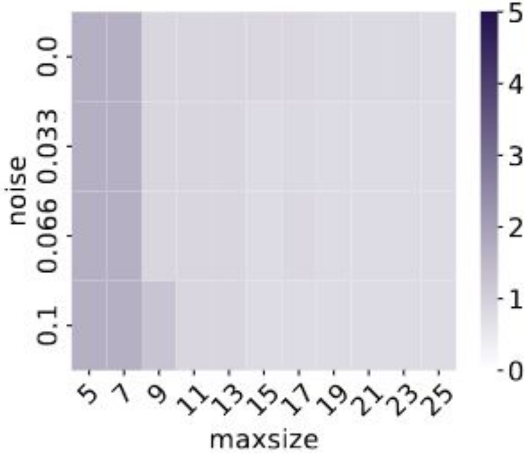
(i)


noise
0.0
0.033
0.066
0.1
maxsize
5 7 9 11 13 15 17 19 21 23 25

(j)


noise
0.0
0.033
0.066
0.1
maxsize
5 7 9 11 13 15 17 19 21 23 25

(k)

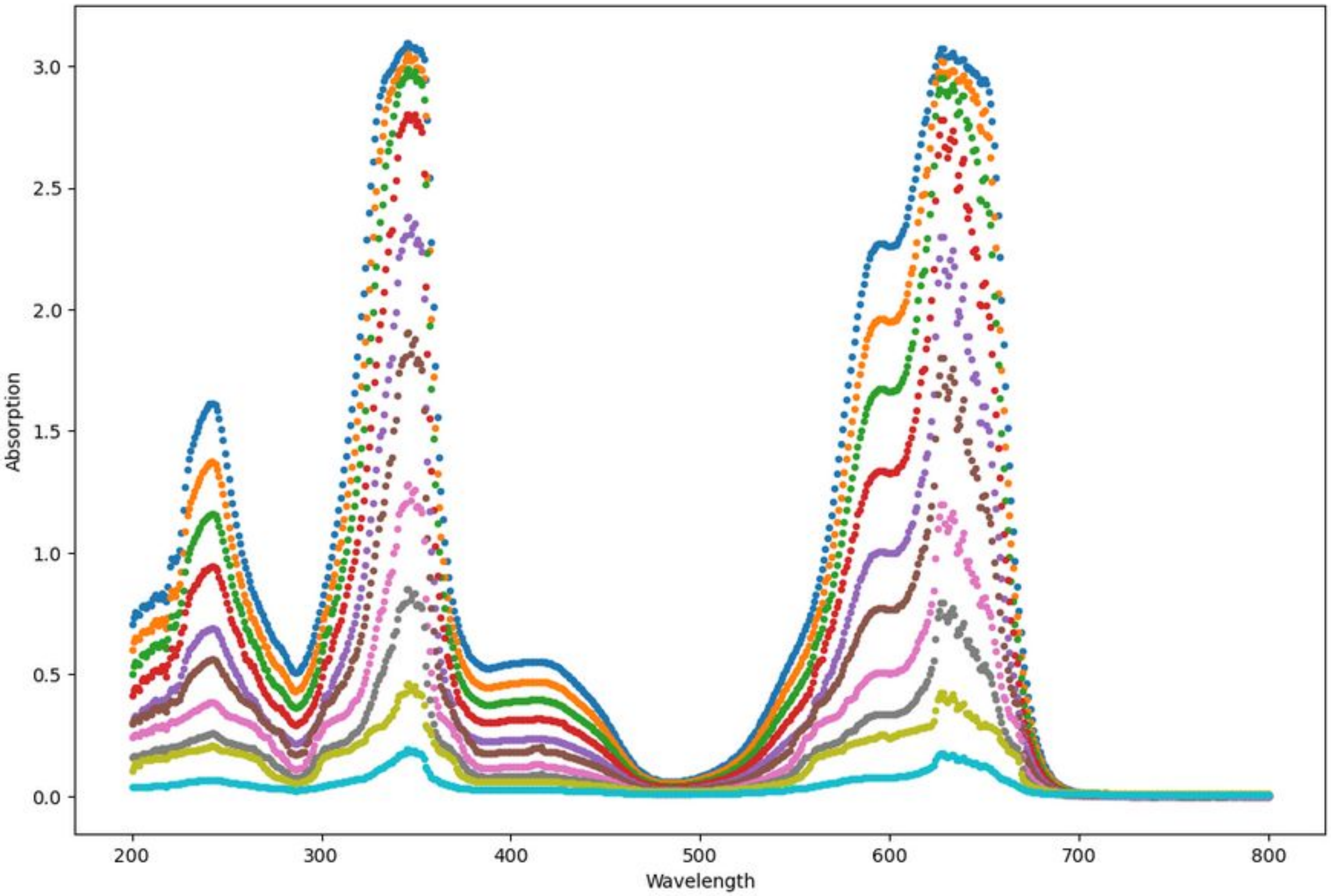

noise
0.0
0.033
0.066
0.1
maxsize
5 7 9 11 13 15 17 19 21 23 25
5
4
3
2
1
0

(l)

	View 1	View 2	View 3	View 4	Partial view
θ_0	2	0	0	2	2
θ_1	2	2	0	0	-2
θ_2	0	2	2	0	2
θ_3	0	0	2	2	2

Parameter	Value
population size	1000
number of evaluations	100000000
pool size	5
error metric	*MSE*
prob. cx	1.0
prob. mut.	0.25
max depth.	10
optim. iterations	100
aggregation function	max
operators	add, sub, mul, div, square, exp, sqrt, sin (f_2 only)

Models	Equation f(x)	$med(MSE)$	$MSE_{S\&P}$
Gaussian [2, 5]	$A \cdot e^{-\frac{x^2}{B}}$	0.363	0.260
Laplace [17]	$A \cdot e^{-B\|x\|}$	0.342	0.084
Cauchy [20]	$A \cdot B^2/(x^2 + B^2)$	0.305	0.079
Linear-Laplace	$(A - Bx) \cdot e^{-C\|x\|}$	0.327	0.065
Exp-Laplace	$A \cdot e^{Bx-C\|x\|}$	0.328	***0.063***
Power-Laplace	$A \cdot e^{B\|x\|^C}$	***0.246***	0.075


3.0
2.5
2.0
1.5
1.0
0.5
0.0
Absorption
200
300
400
500
600
700
800
Wavelength