


Multiview Symbolic Regression

How to learn laws from examples

Etienne Russeil - *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*

PLAN

- Motivation
- Symbolic Regression
- Multiview Symbolic regression
- Scientific applications
- Conclusion

Laws in science

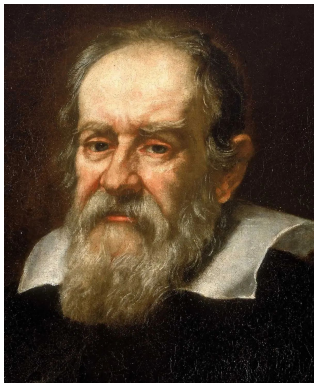
The importance of laws in science

Science's goal is to describes and predicts phenomena.

The ultimate form of that is a mathematical equation

Galileo Galilei

"Mathematics is *the language with which God has written the universe*."



Einstein matter energy equivalence law

$$E = mc^2$$

Newton's gravitation law

$$F = G\frac{m_1 m_2}{d^2}$$

Pythagoras theorem

$$a^2 + b^2 = c^2$$

Maxwell's equations

$$\nabla \cdot E = 0 \quad \nabla \times E = -\frac{1}{c}\frac{\partial H}{\partial t}$$

$$\nabla \cdot H = 0 \quad \nabla \times H = \frac{1}{c}\frac{\partial E}{\partial t}$$

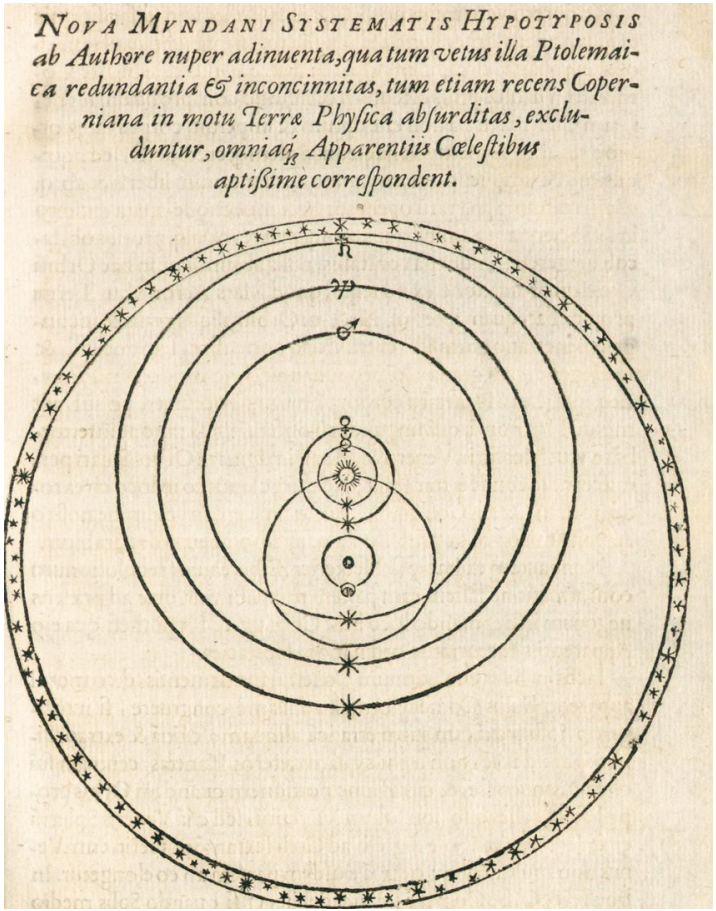
Shannon's information theory

$$H = -\sum p(x) \log p(x)$$

Equations come from observations : Kepler's laws

Tycho Brahe

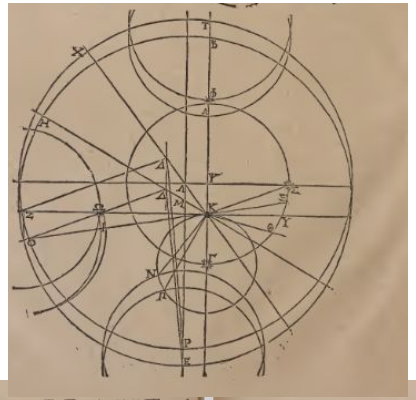
The planets orbit around the sun and I will prove it with measurements !

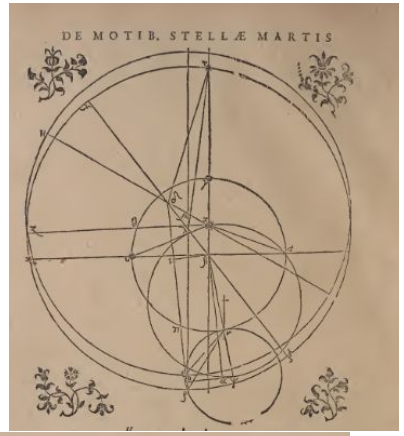


Equations come from observations : Kepler's laws

Johannes Kepler

Thank you for the dataset





Tabula TYCHONIS BRAHE obs vatarum & computatarum oppositionum
MARTIS cum linea me motus Solis, ejusque examen.

Igitur tabula, q qua supra, fuit ista.

Planetæ ♂ motus in suo eccentrico e certis observationibus a nychiis per annos XX (ab LXXX usque M DC) sedulo per nostra
instrumenta habitis respectu variarum dispositionu in subjecta tabula patet, accurata restitutio.

Tempus æquale ♂				Long. obs. respec. circuli ♂				Latitudo vera obs.				Long. obs. respectu ecliptiæ			Differentia			Simpl. Long. ♂			Apog. ♂			Præcess. æquin. nostra			Supput.		
Anni	Mens.	D	H	M	G	M	S		G	M	S		G	M	S	M	S		S	G	M	S	S	G	M	S	G	M	S	G	M	S
1580	Novemb.	17	9	40	6	50	10	♊	1	40	0	B	6	46	10		4	10	A	0	27	29	46	3	25	21	40	27	58	50	6	50	40
1582	Decembr.	28	12	16	16	51	30	♋	4	6	0	B	16	46	10		5	20	A	2	11	34	56	3	25	22	17	28	0	38	16	51	26
1585	Januarii	31	19	35	21	9	50	♌	4	32	10	B	21	10	16		0	36	S	3	22	37	46	3	25	22	55	28	2	25	21	9	41
1587	Martii	7	17	22	25	5	10	♍	3	38	12	B	25	10	20		5	10	S	5	3	27	46	3	25	23	32	28	4	10	25	4	50
1589	Aprilis	15	13	34	3	54	35	♎	1	6	45	B	3	58	10		3	35	S	6	16	53	7	3	25	24	10	28	5	55	3	54	33
1591	Junii	8	16	25	26	+0 +2	33 0	♐ ♑	3	59	0	M	26	32	0		10	20	A	8	7	47	30	3	25	24	48	28	7	47	26	40	23
1593	Augusti	24	2	13	12	35	0	♑	6	3	0	M	12	43	45		8	45	S	10	10	53	50	3	25	25	26	28	9	40	12	34	36
1595	Octobris	29	21	22	17	56	5	♉	0	5	15	B	17	56	15		0	12	A	0	8	26	47	3	25	27	35	28	11	27	17	57	14
1597	Decemb.	13	13	35	2	34	0	♋	3	33	0	B	2	28	0		6	0	A	1	24	55	47	3	25	29	5	28	13	20	2	32	20
1600	Januarii	19	9	40	8	18	45	♌	4	30	50	B	8	18	0		0	45	S	3	6	46	16	3	25	30	6	28	15	5	8	19	57

Equations come from observations : Kepler's laws

Johannes Kepler

1: The orbit of every planet is an ellipse with the Sun at one of the two foci

2: A line joining a planet and the Sun sweeps out equal areas during equal intervals of time

$$\frac{dA}{dt} = \frac{\pi \times a \times b}{T}$$

3: The ratio of the square of an object's orbital period with the cube of the semi-major axis of its orbit is the same for all objects orbiting the same primary.

$$\frac{R^3}{T^2} = constant$$

What if we could automatise this discovery process ?

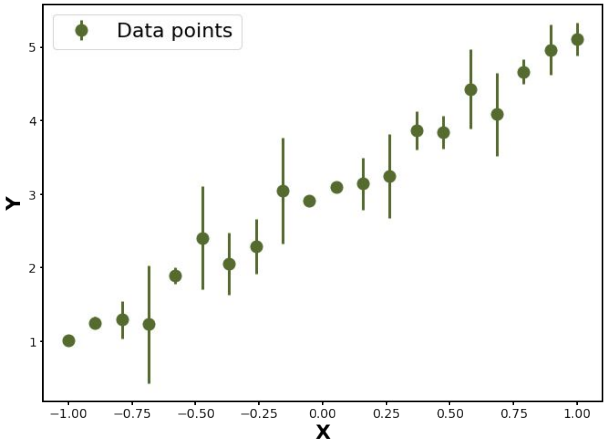
Symbolic Regression

Traditional Symbolic Regression

Caution: I will present a standard method based on Genetic Programming but it is not the only one !
In practice each implementation varies

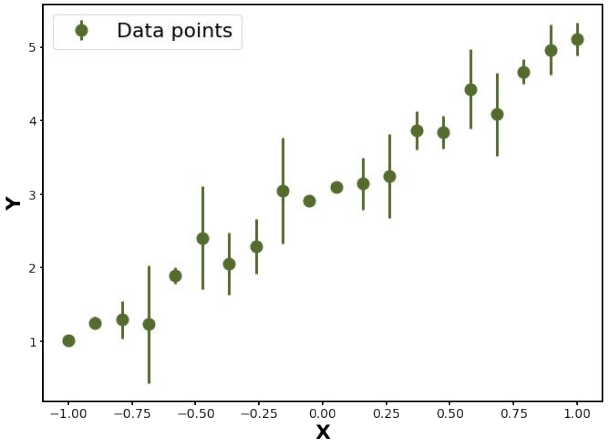
Traditional Symbolic Regression

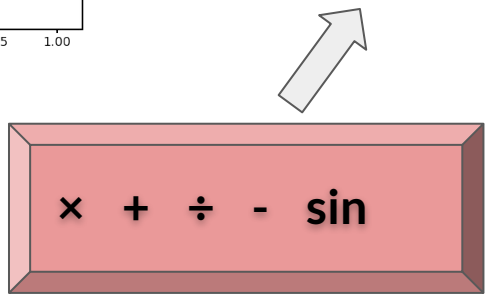
DATA SET

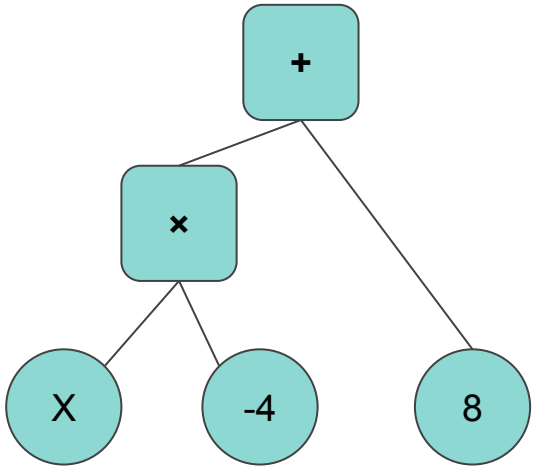


Traditional Symbolic Regression

DATA SET



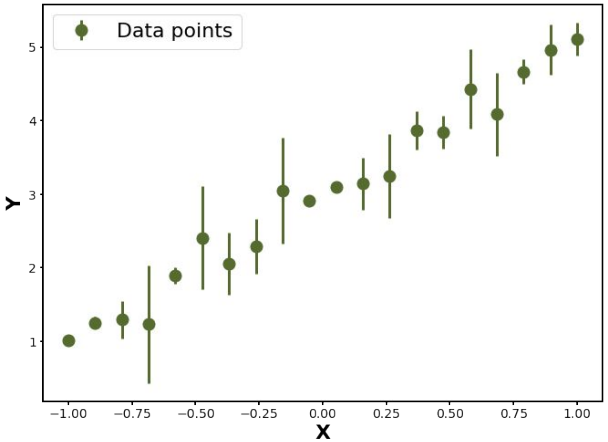




As a first step the algorithm will randomly generate many different equations

Traditional Symbolic Regression

DATA SET



RANDOM EQUATIONS

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

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

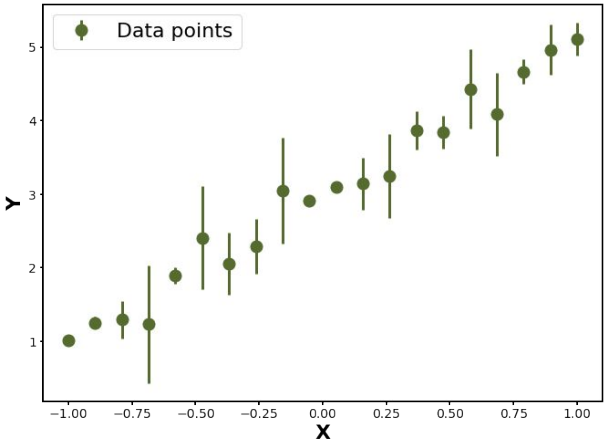
$f(X) = 42$

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

$f(X) = X$

Traditional Symbolic Regression

DATA SET



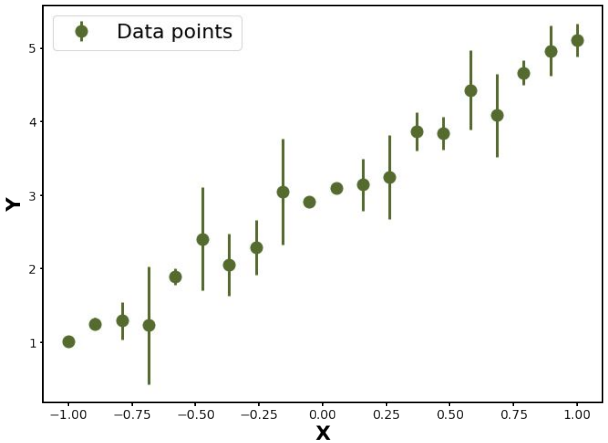
RANDOM EQUATIONS

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



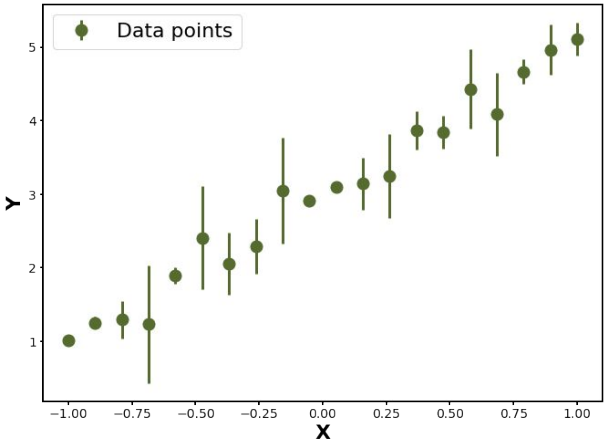
RANDOM EQUATIONS

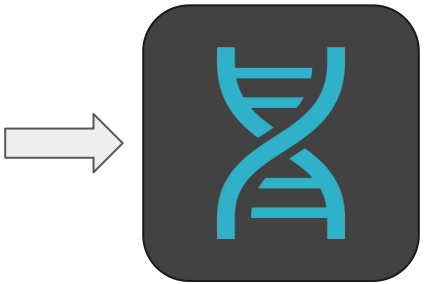
COST FUNCTION

f(X) = sin(X) + 2	COST = 12
f(X) = X² - 1	COST = 24
f(X) = 42	COST = 43
f(X) = -4X + 8	COST = 7
f(X) = X	COST = 3

Traditional Symbolic Regression

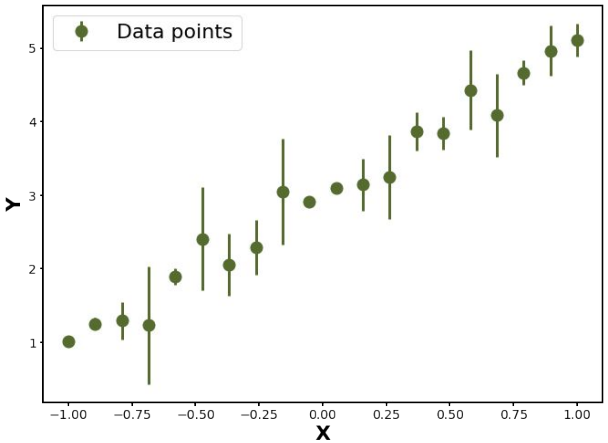
DATA SET





Traditional Symbolic Regression

DATA SET



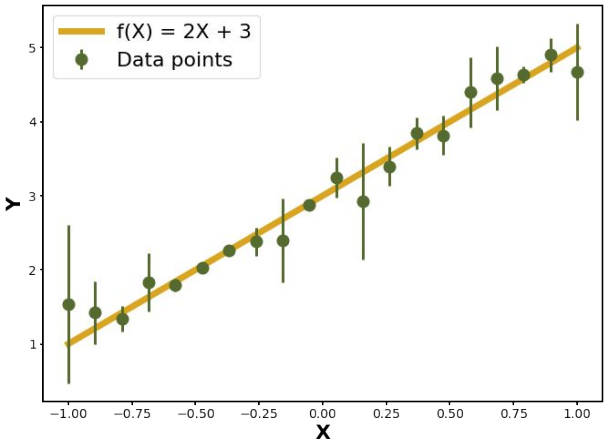
EVOLVED EQUATIONS

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

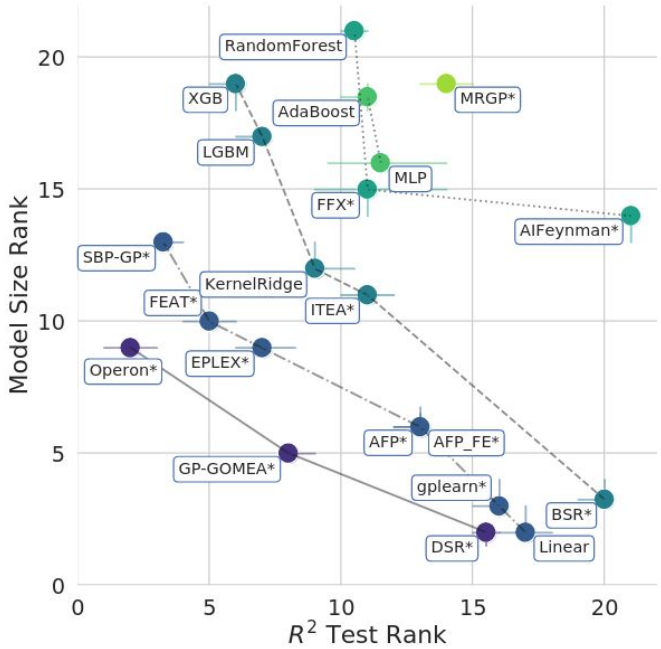
Best answer

f(X) = 2 X + 3

COST ~ 0

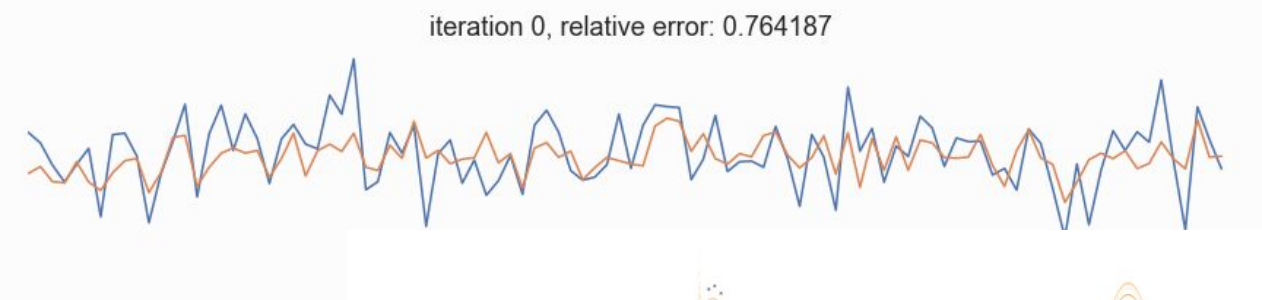
Traditional Symbolic Regression

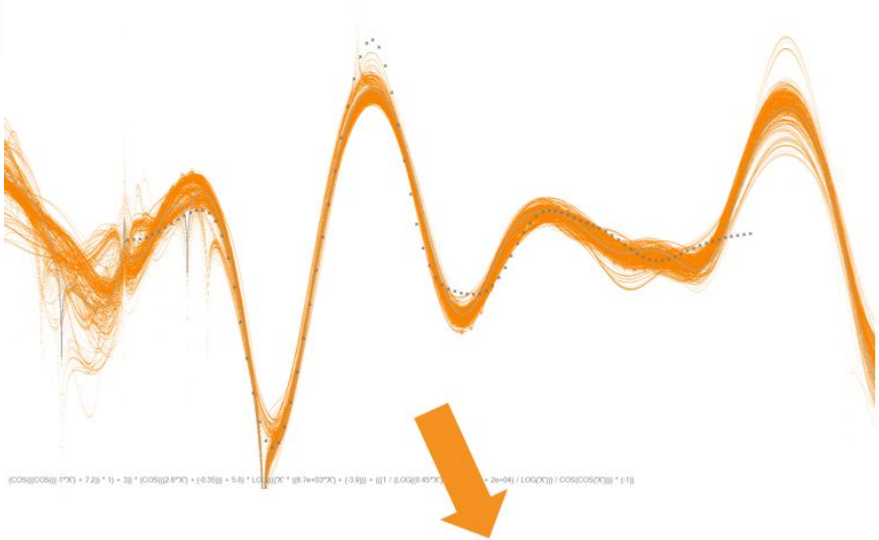
Many different implementations of SR available !



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

Traditional Symbolic Regression





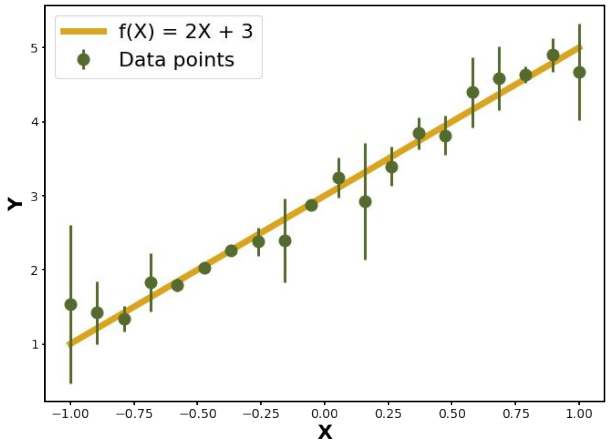
$$e^{-x}x^3\cos(x)\sin(x)(cos(x)sin(x)^2-1)$$

...

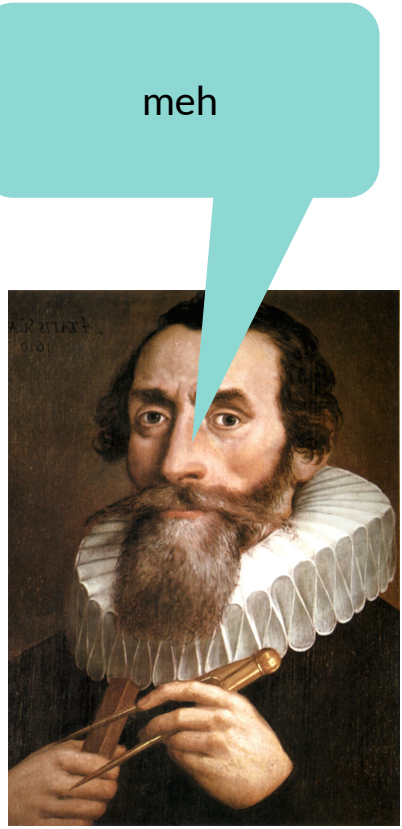
https://github.com/heal-research/operon

Traditional Symbolic Regression

DATA SET

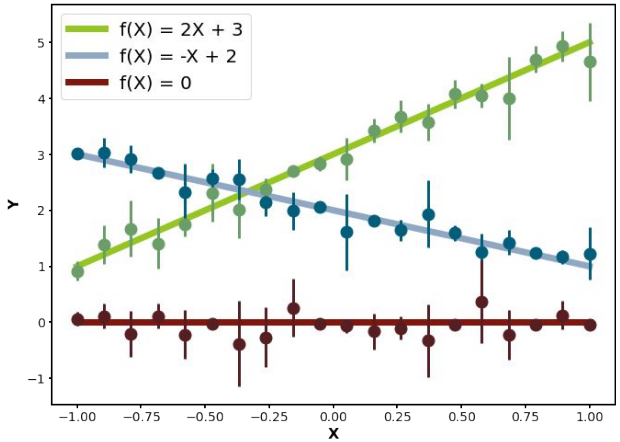


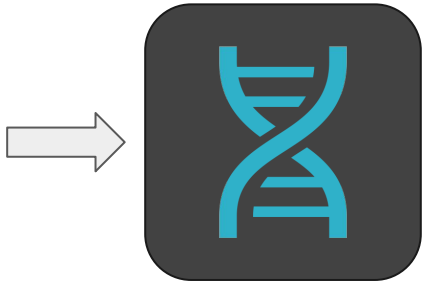
does not output a general law



Traditional Symbolic Regression

DATA SETS

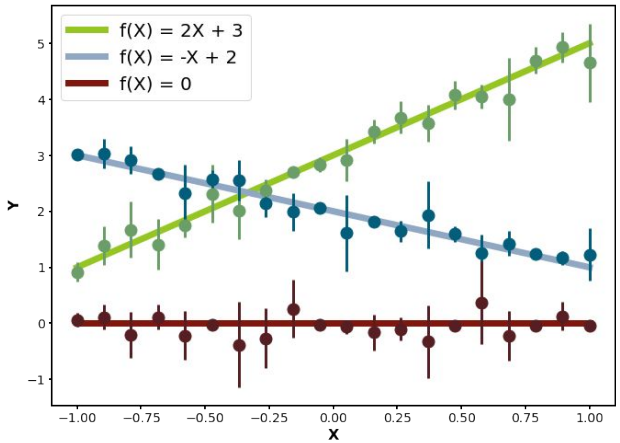




Here look

Traditional Symbolic Regression

DATA SETS



Best answers

f(X) = 2 X + 3

f(X) = -X + 2

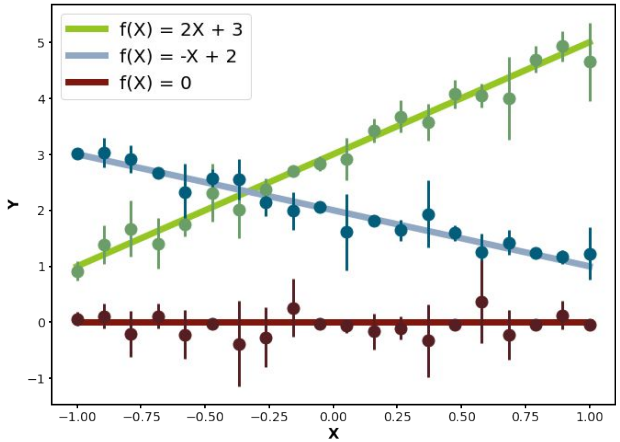
f(X) = 0

I would have found
f(X) = A X + B

MultiView Symbolic Regression

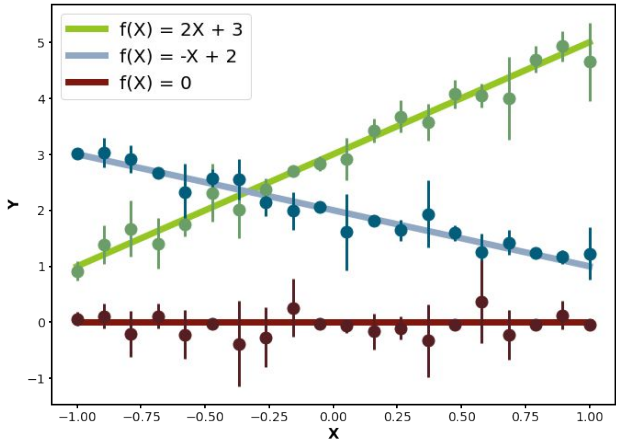
Multiview Symbolic Regression (MvSR)

DATA SETS



Multiview Symbolic Regression (MvSR)

DATA SETS



RANDOM EQUATIONS

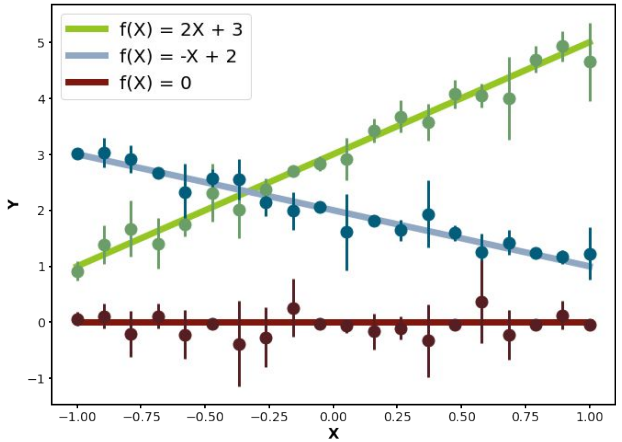
$f(X) = \sin(X) + \mathbf{A}$

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

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

Multiview Symbolic Regression (MvSR)

DATA SETS



RANDOM EQUATIONS

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

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

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

COST FUNCTION AFTER MINIMIZATION

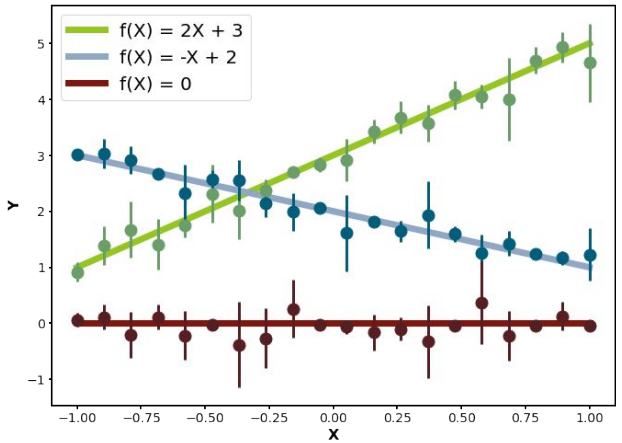
COST = 24
32
7

COST = 17
8
0

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

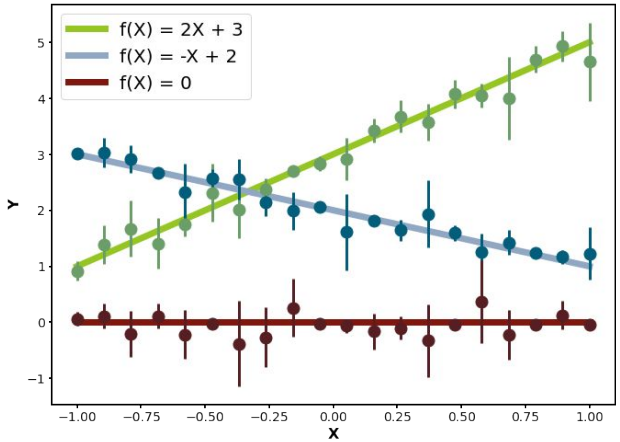
$f(X) = \mathbf{A}$ COST = 19 10 0

Average COST

Multiview Symbolic Regression (MvSR)

Sehr gut !

DATA SETS



After many generation

Best answer

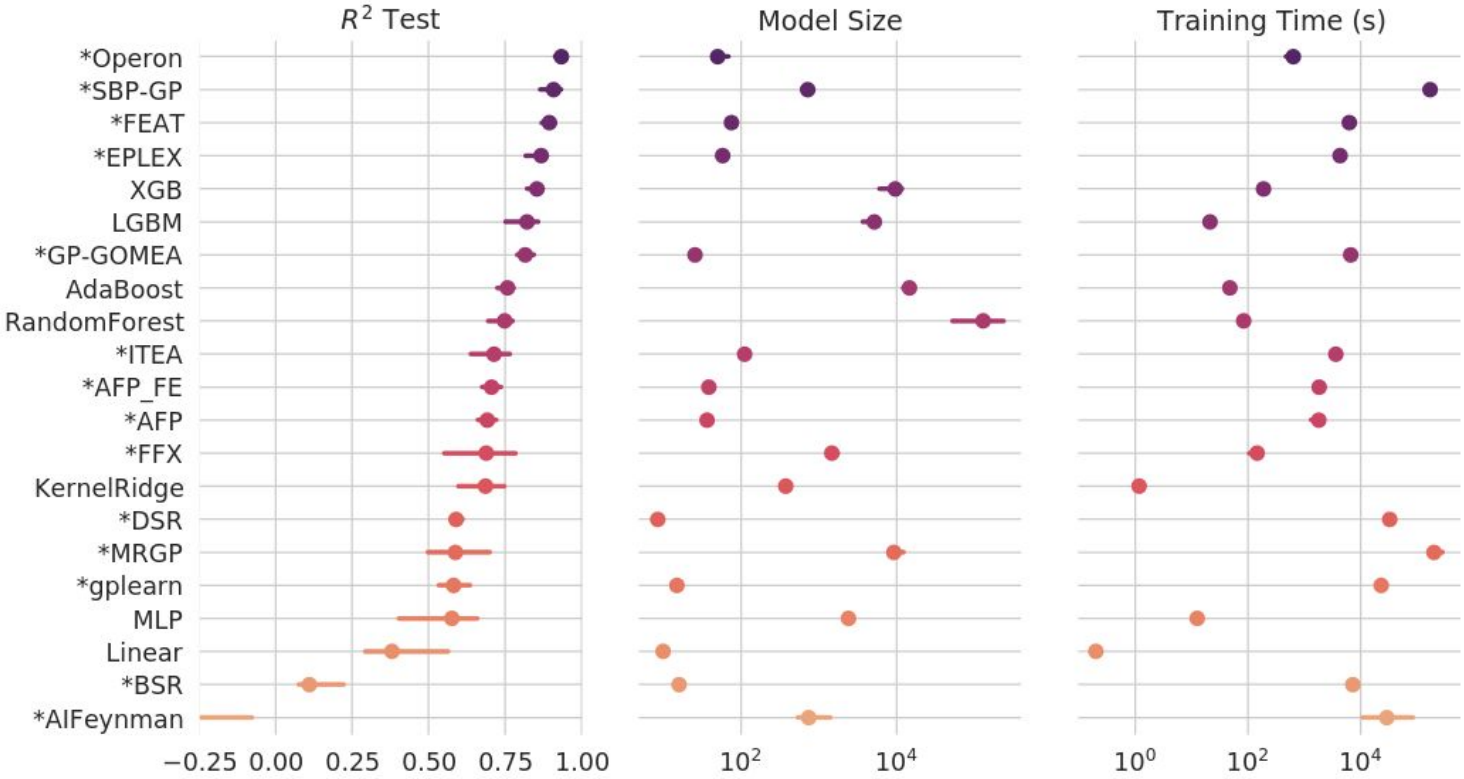
f(X) = A X + B

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

Multiview Symbolic Regression (MvSR)

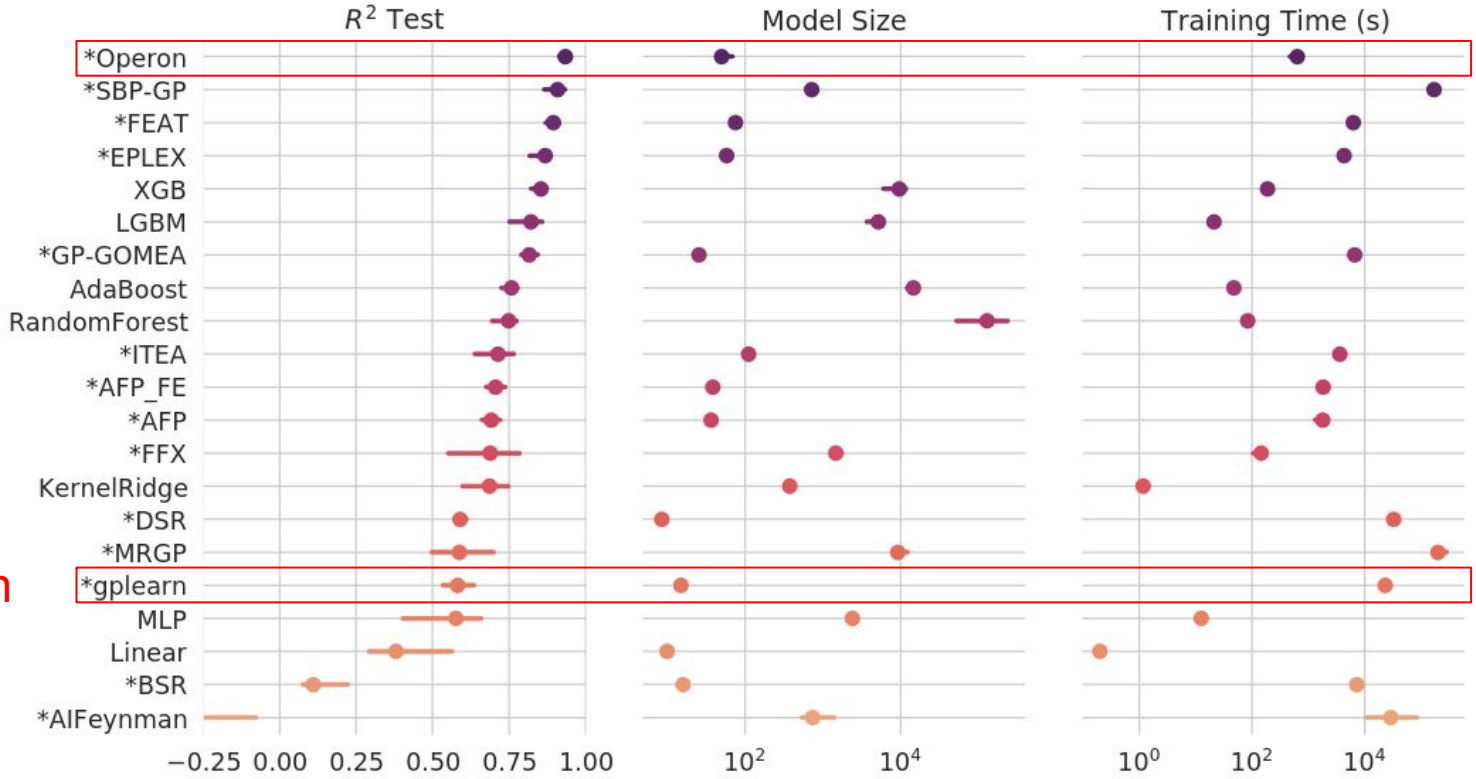


LaCava et al., 2021. Contemporary Symbolic Regression Methods and their Relative Performance, arXiv:cs https://arxiv.org/abs/2107.14351

Multiview Symbolic Regression (MvSR)

C++

Python



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

IV

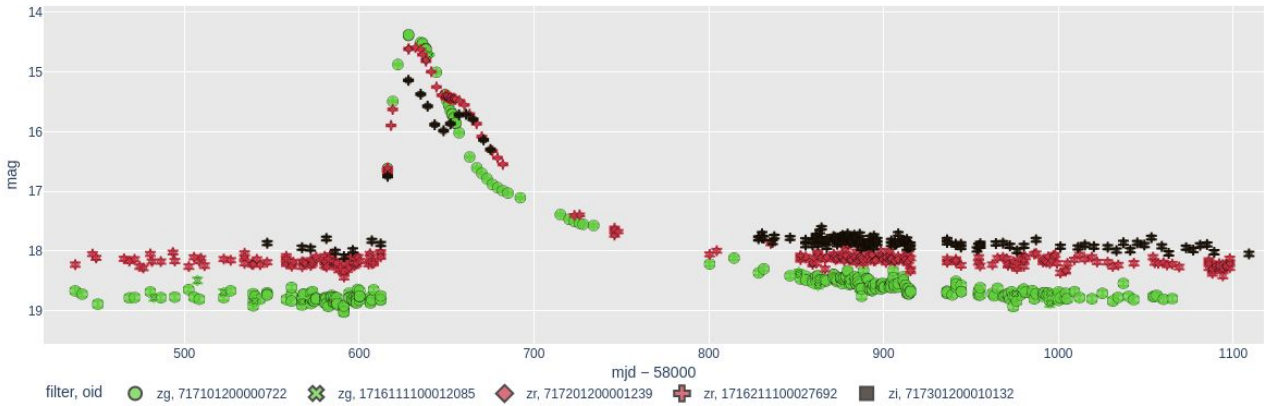
Scientific applications

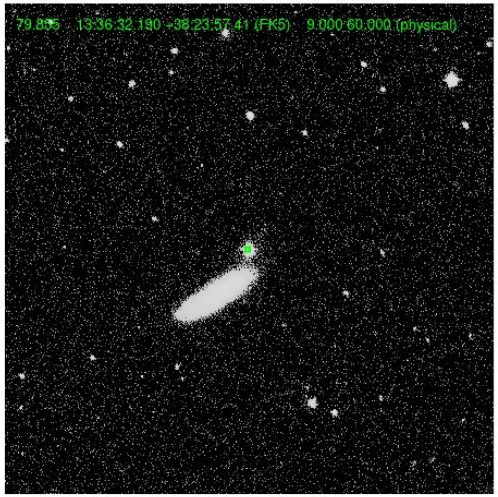
SNAD ZTF DR17 object viewer

ZTF DR OID or SNAD name 633207400004730 or 202 GO

Coordinates or name 00h00m00s +00d00m00s or M57 radius, arcsec 1 GO

717101200000722





Download PNG, PDF, CSV

MJD range: 50000 — 70000

50000.0 ≤ MJD ≤ 70000.0

ZTF Private Survey: 58194.0 ≤ MJD ≤ 59524.0

Full light curve Folded light curve

Closest Antares object, diff-photometry Closest Pan-STARRS object, apparent Closest Gaia object, apparent

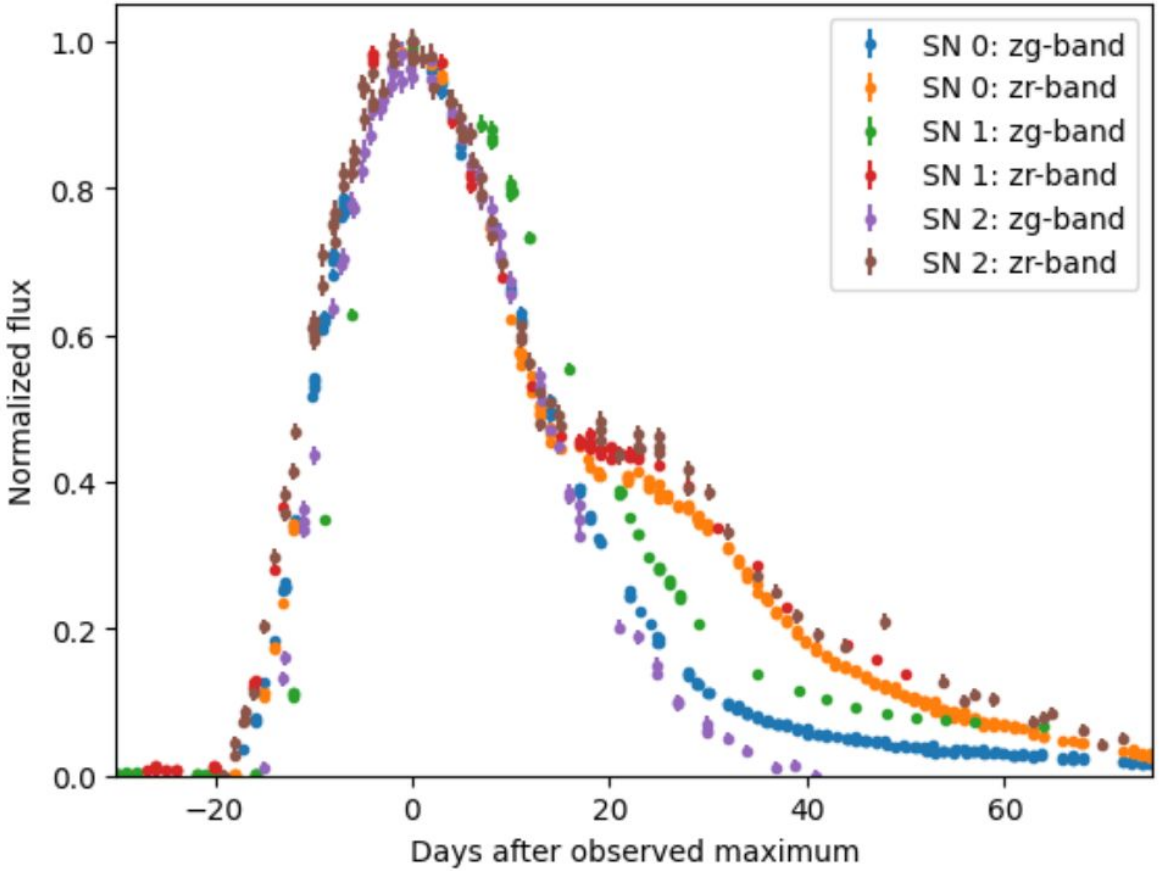
Magnitude Flux diff Magnitude diff Flux

Open in JS9 Download FITS Product directory

No minor planets found in 15″

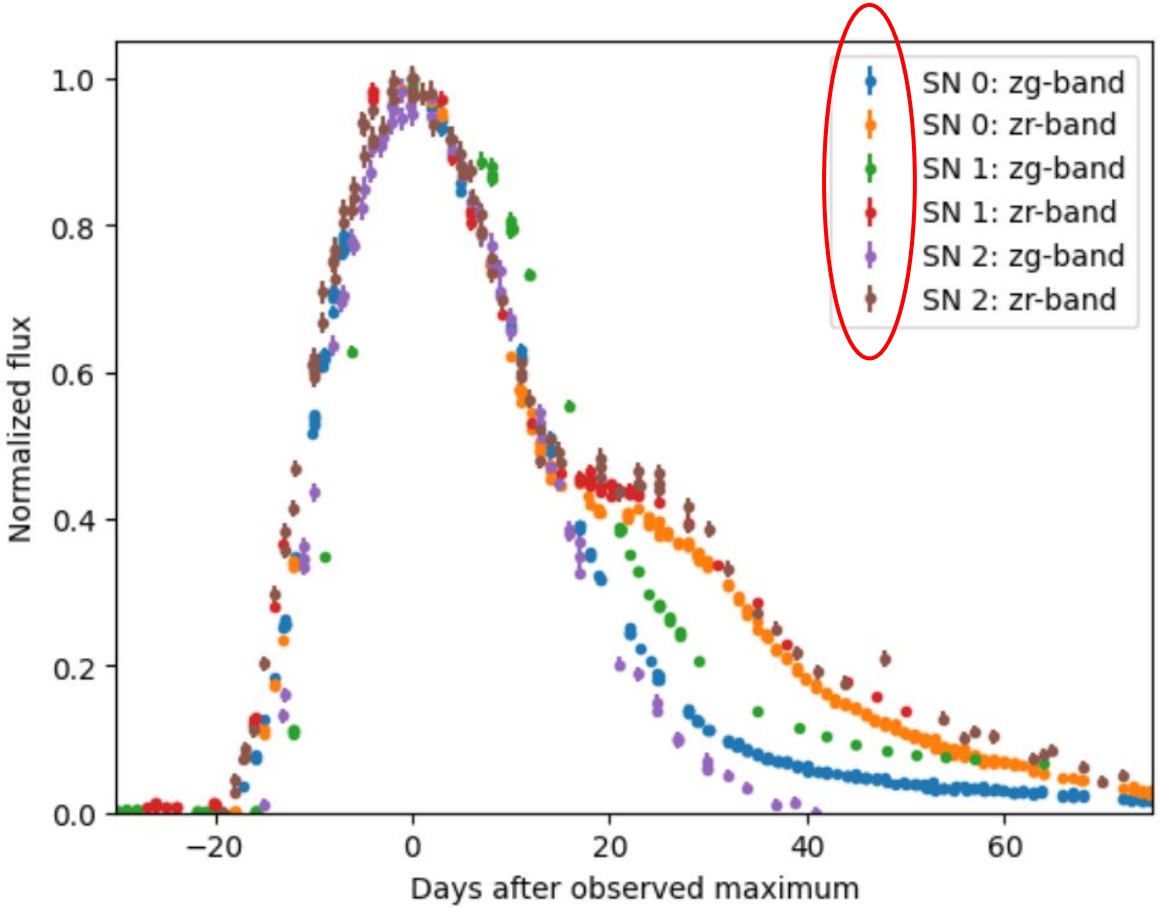
https://ztf.snad.space/dr17/view/717101200000722

Astrophysical database

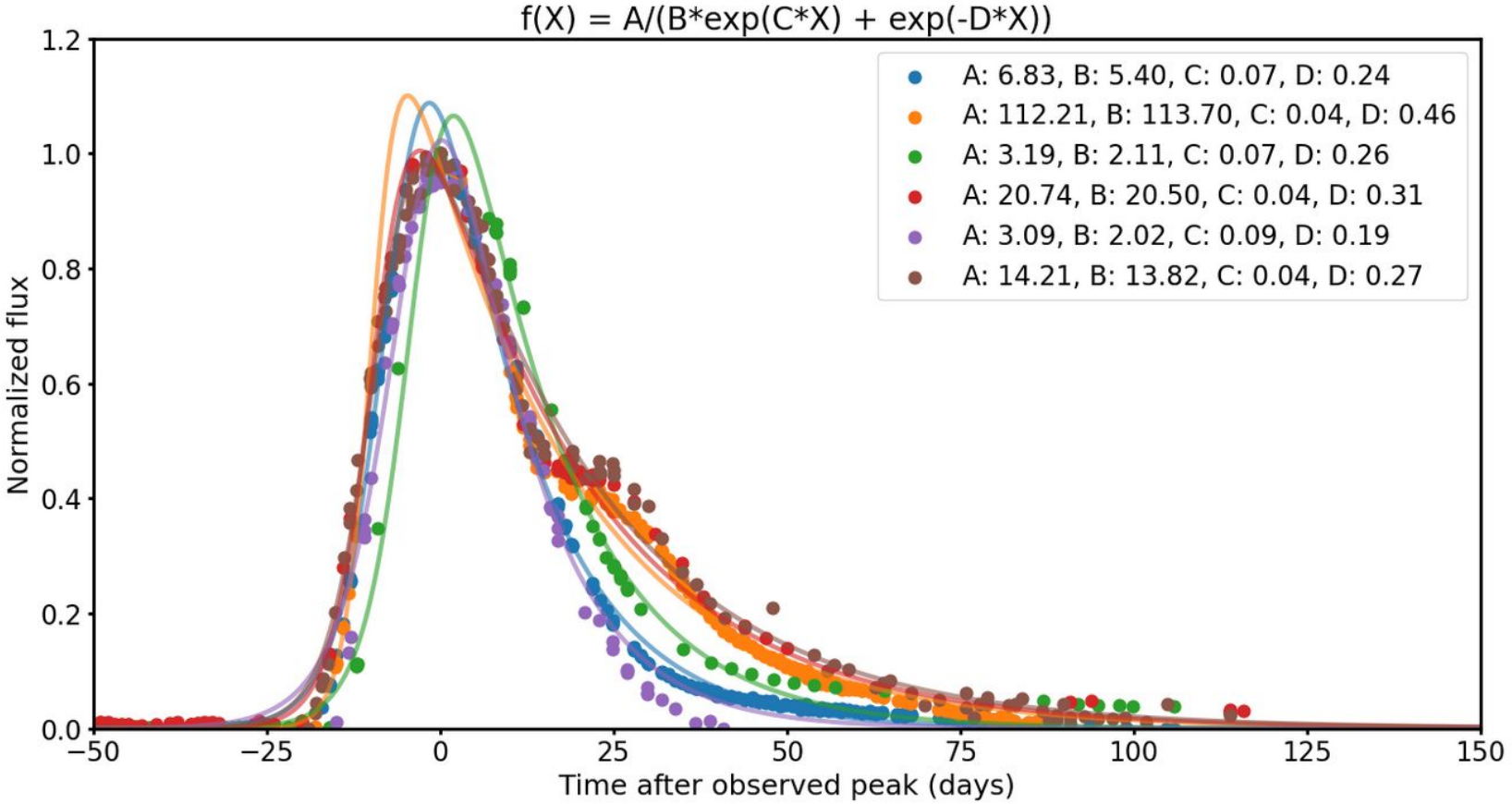


Astrophysical database

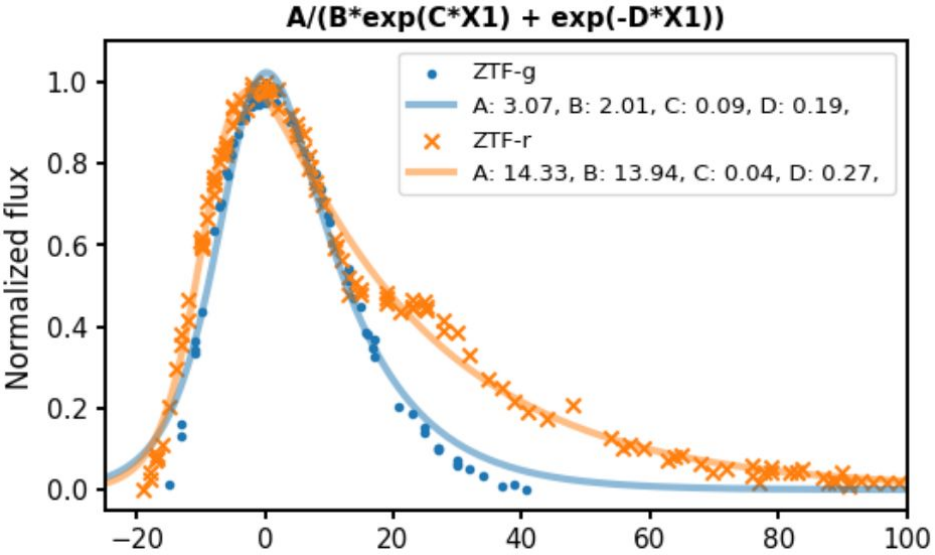
Error bars not used :(



Astrophysical database

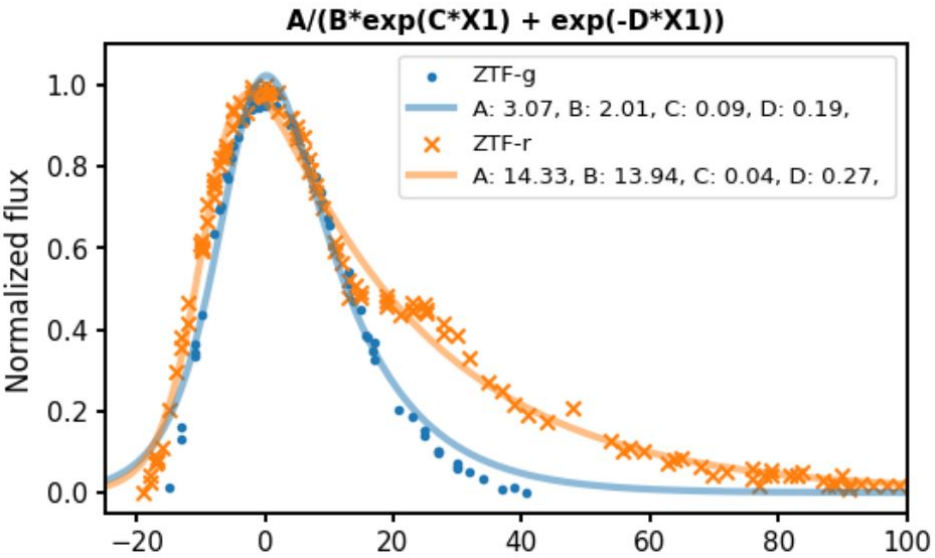


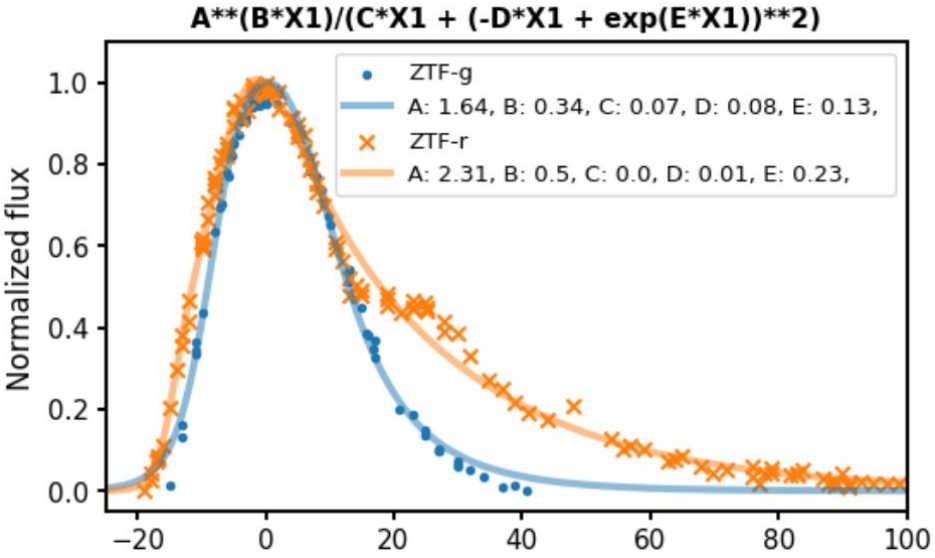
Astrophysical database



MvSR recovers the literature !

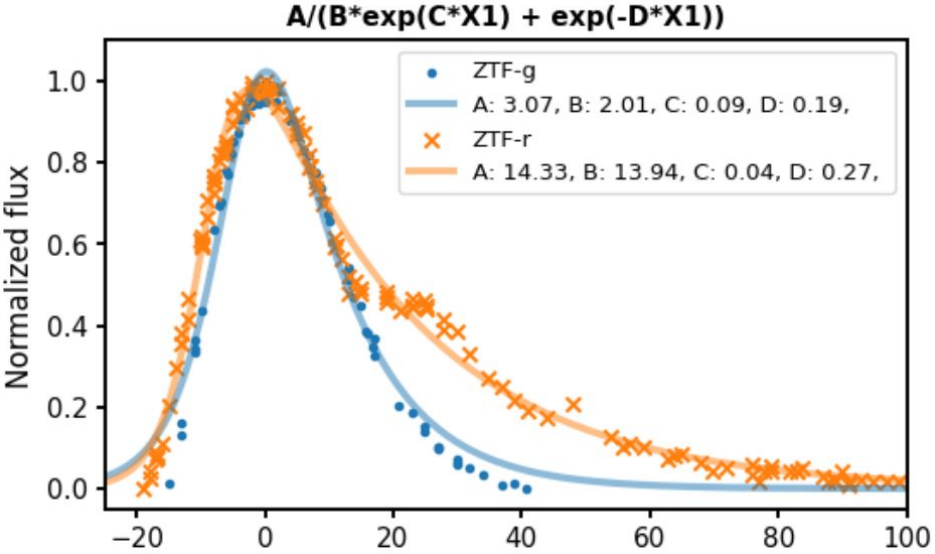
Astrophysical database

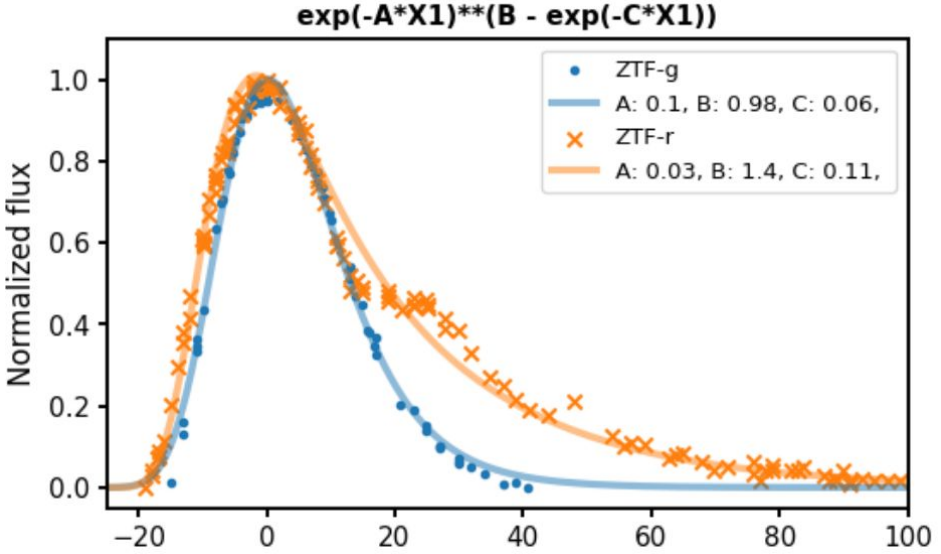




It can also generate more complexe and better solution

Astrophysical database

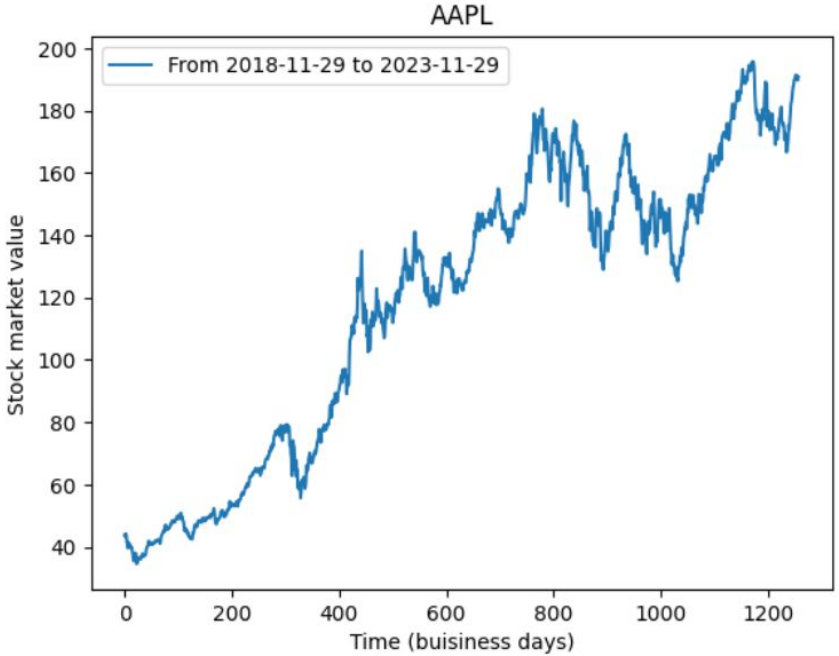




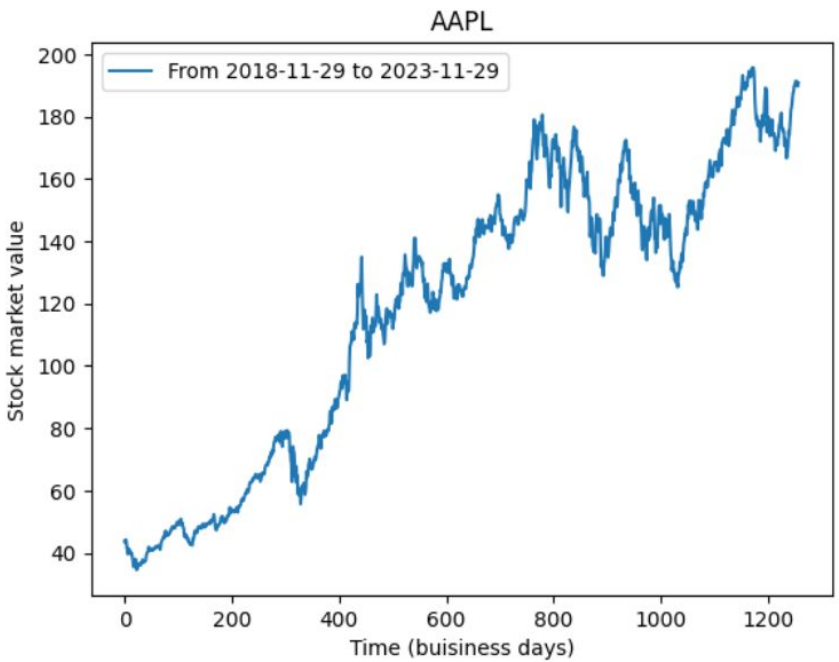
As well as unexpected but very effective forms

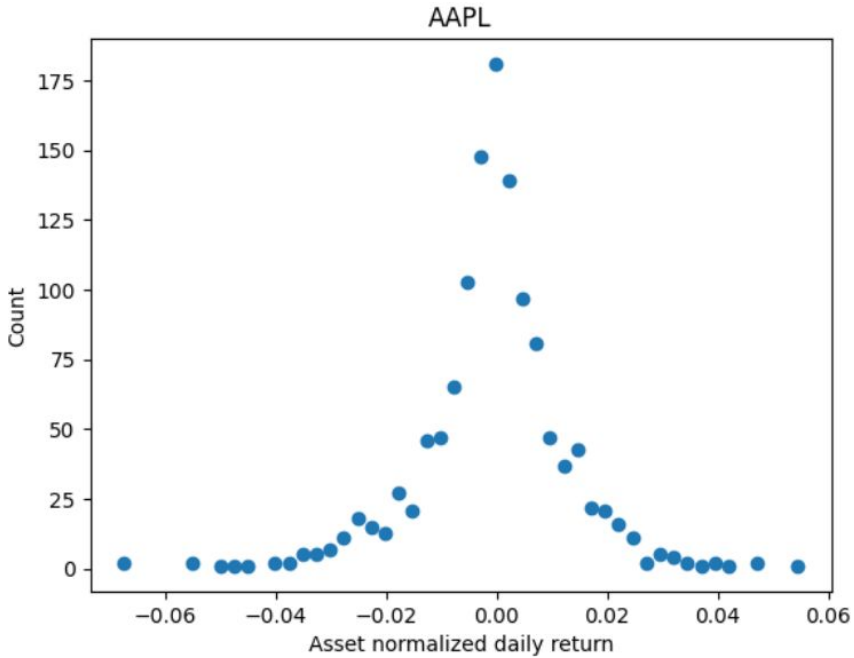
Let’s try other use cases !

Finance database



Finance database





Finance database

S&P500

Stock market of the 500 biggest american companies

Finance database

S&P500

Stock market of the 500 biggest american companies

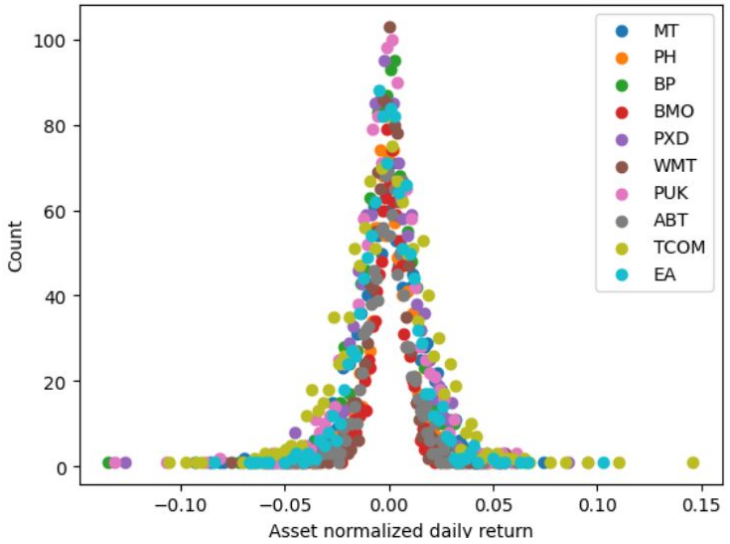
Usually aggregated for analysis

Finance database

S&P500

Stock market of the 500 biggest american companies

10 random companies



490 to check

Finance database

Recover literature

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

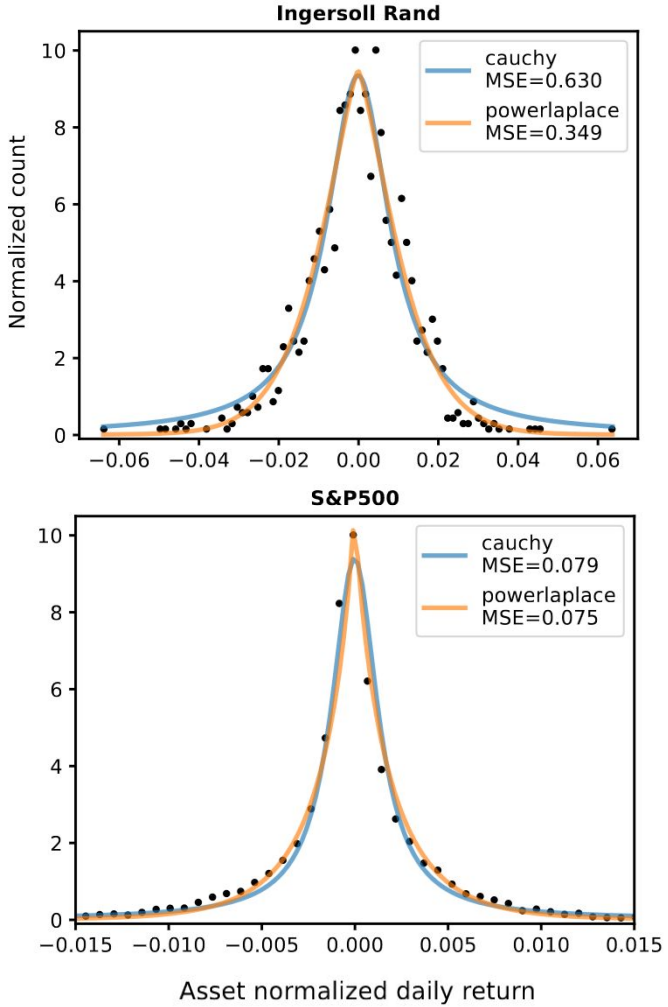
Find new models

Finance database

Recover literature

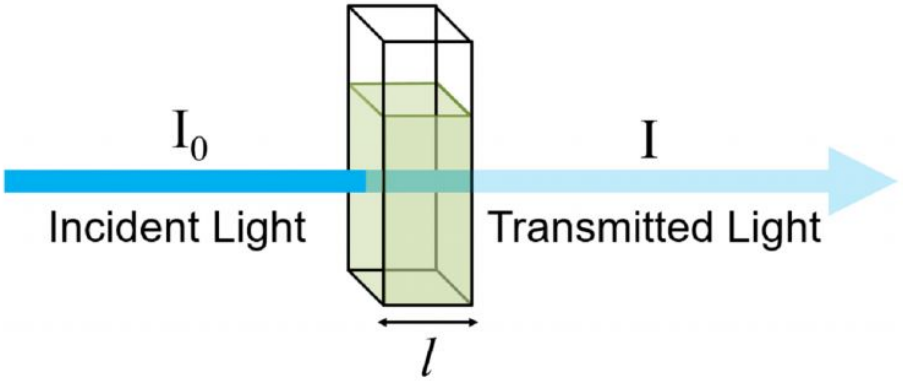
Models	Equation f(x)	*med(MSE)*
Gaussian [2, 5]	$A \cdot e^{-\frac{x^2}{B}}$	0.363
Laplace [17]	$A \cdot e^{-B\|x\|}$	0.342
Cauchy [20]	$A \cdot B^2/(x^2 + B^2)$	0.305
Linear-Laplace	$(A - Bx) \cdot e^{-C\|x\|}$	0.327
Exp-Laplace	$A \cdot e^{Bx - C\|x\|}$	0.328
Power-Laplace	$A \cdot e^{B\|x\|^C}$	***0.246***

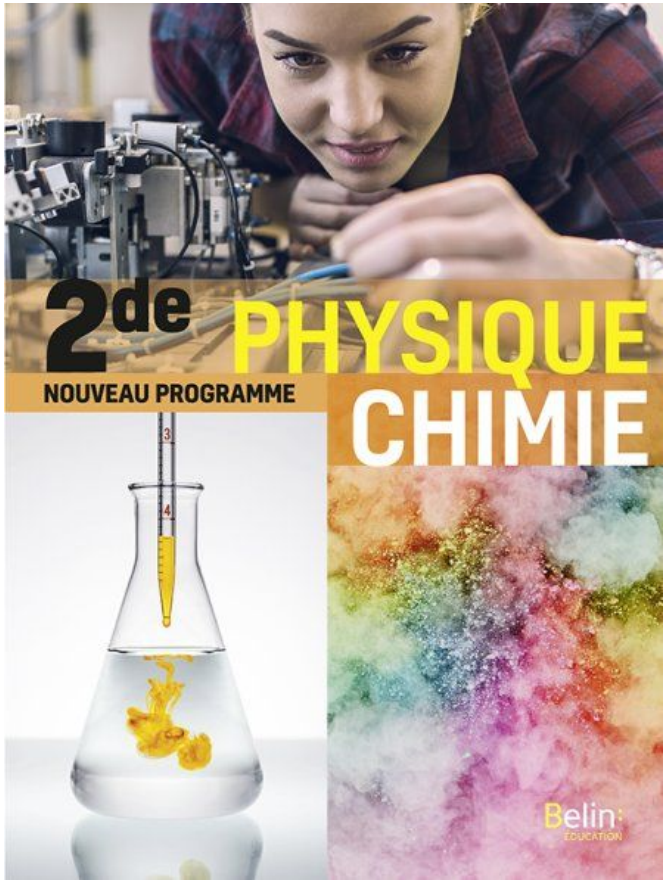
Find new models



And another one ?

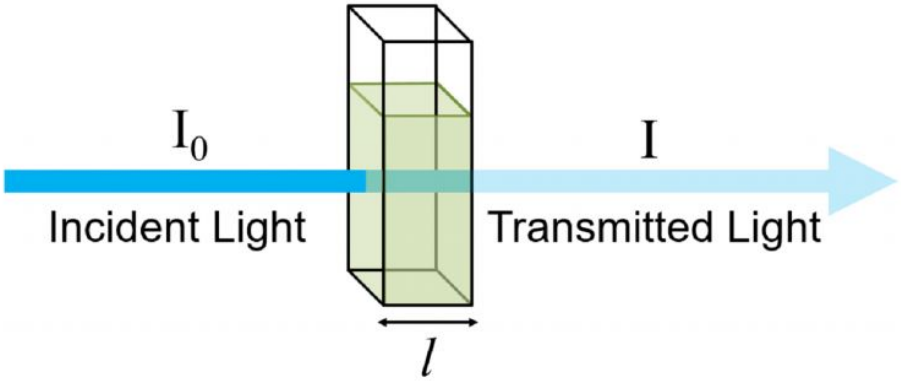
Chemistry database





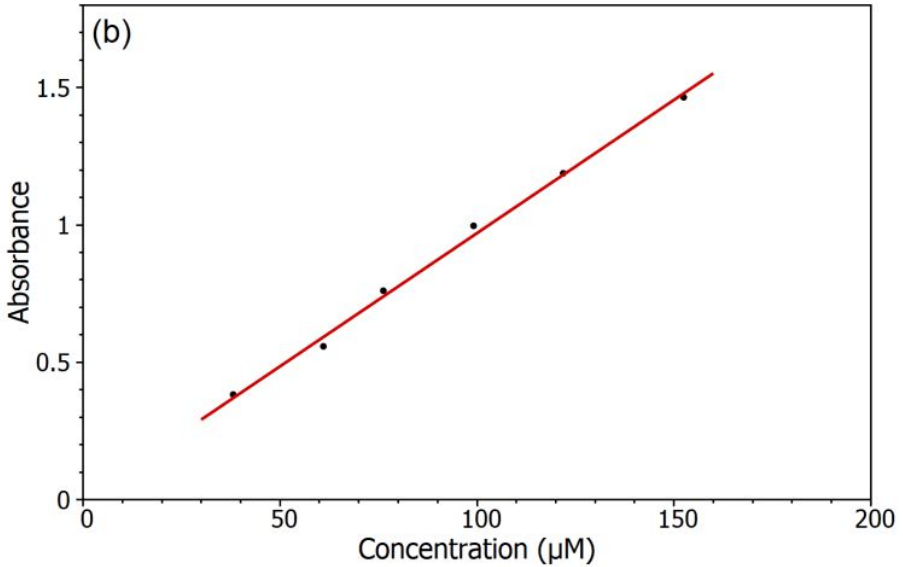
https://www.edinst.com/blog/the-beer-lambert-law/

Chemistry database



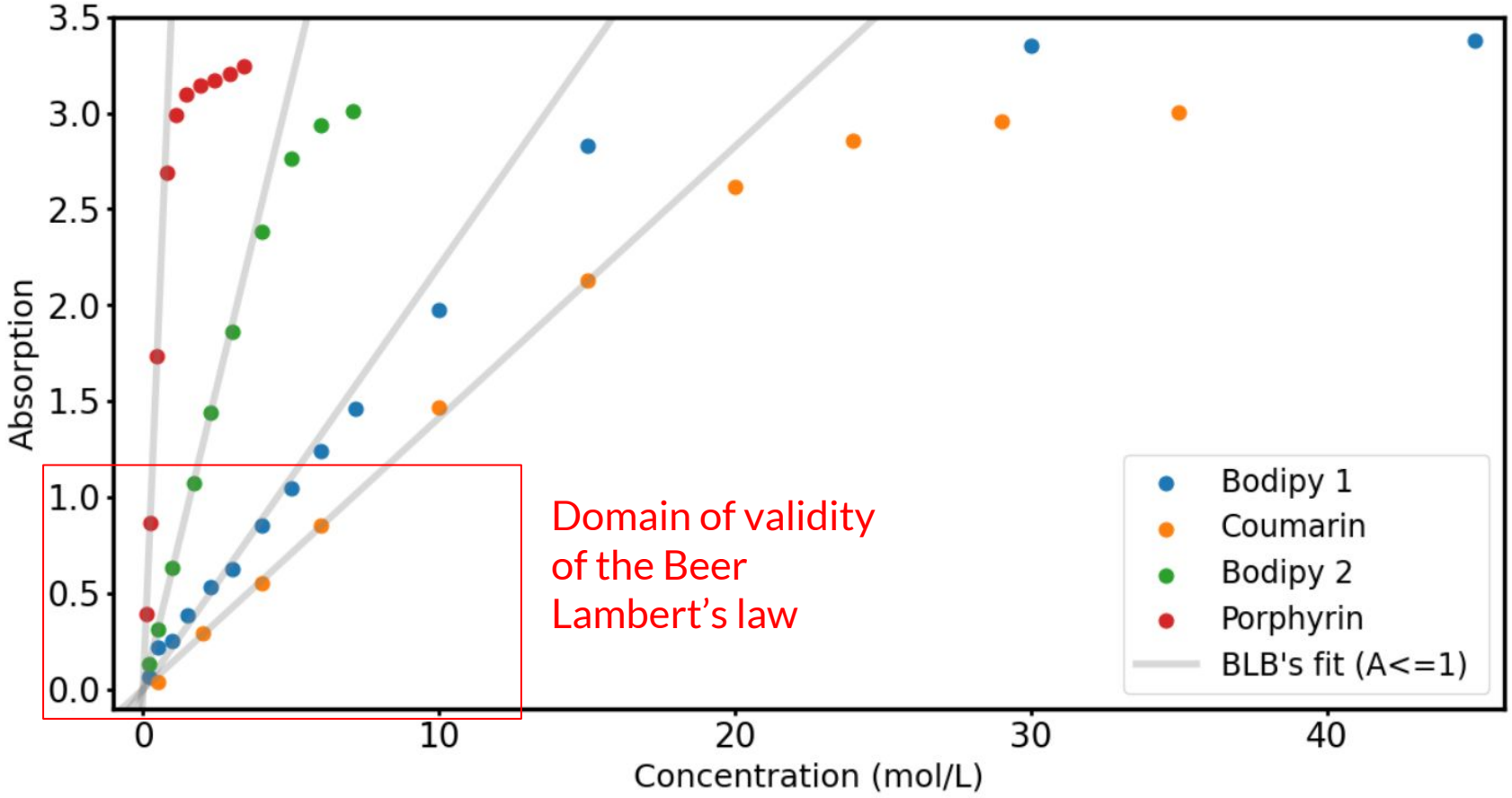
$$log_{10}(\frac{I_0}{I}) = A = k \cdot C$$

Beer Lambert's law

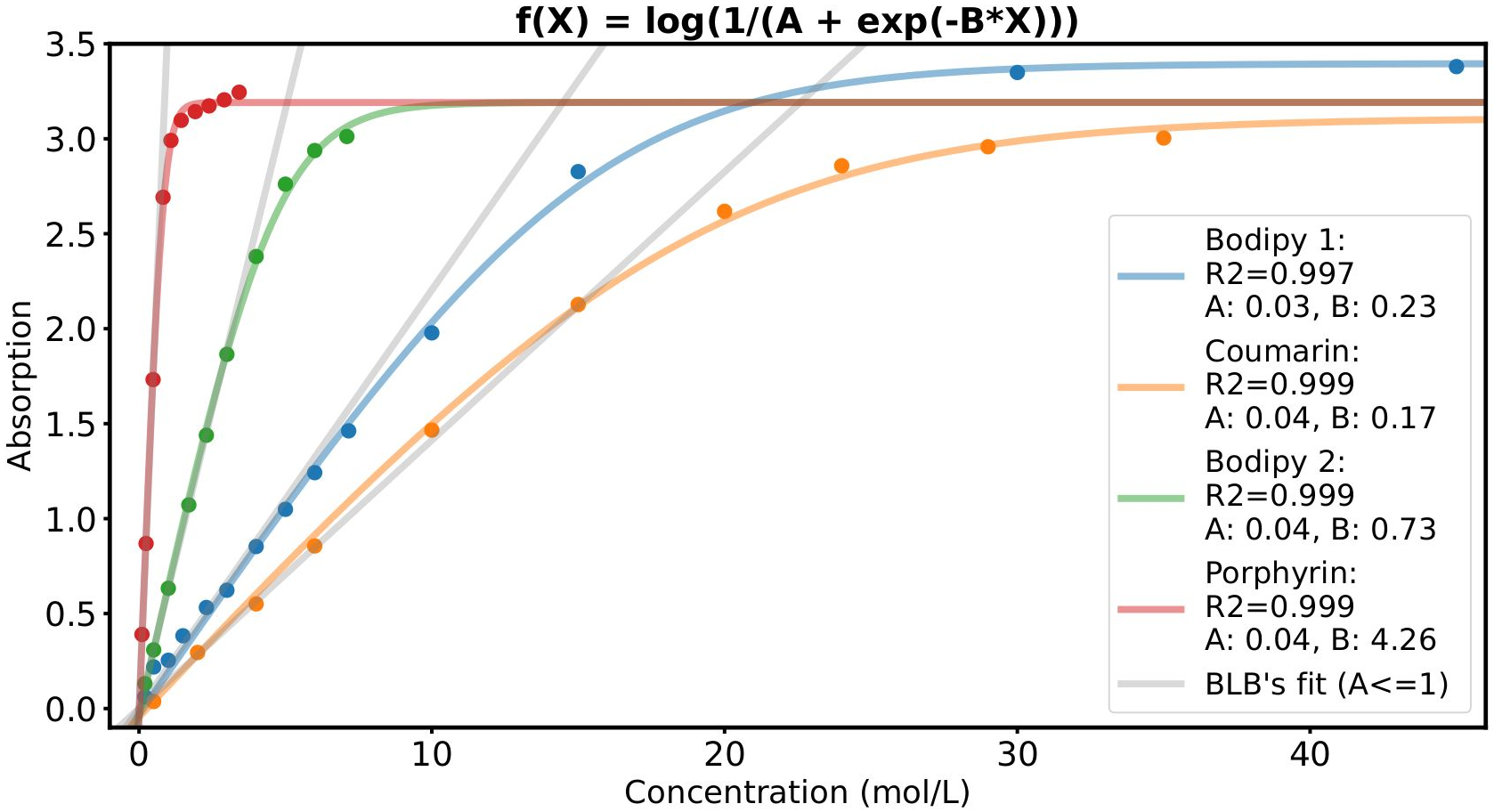


https://www.edinst.com/blog/the-beer-lambert-law/

Chemistry database



Chemistry database



Chemistry database

General Beer Lambert's law:

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

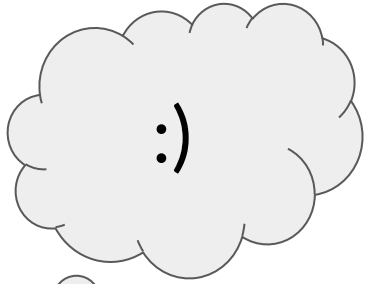
Saturation parameter

Linear slope

VI

Conclusion

Conclusion

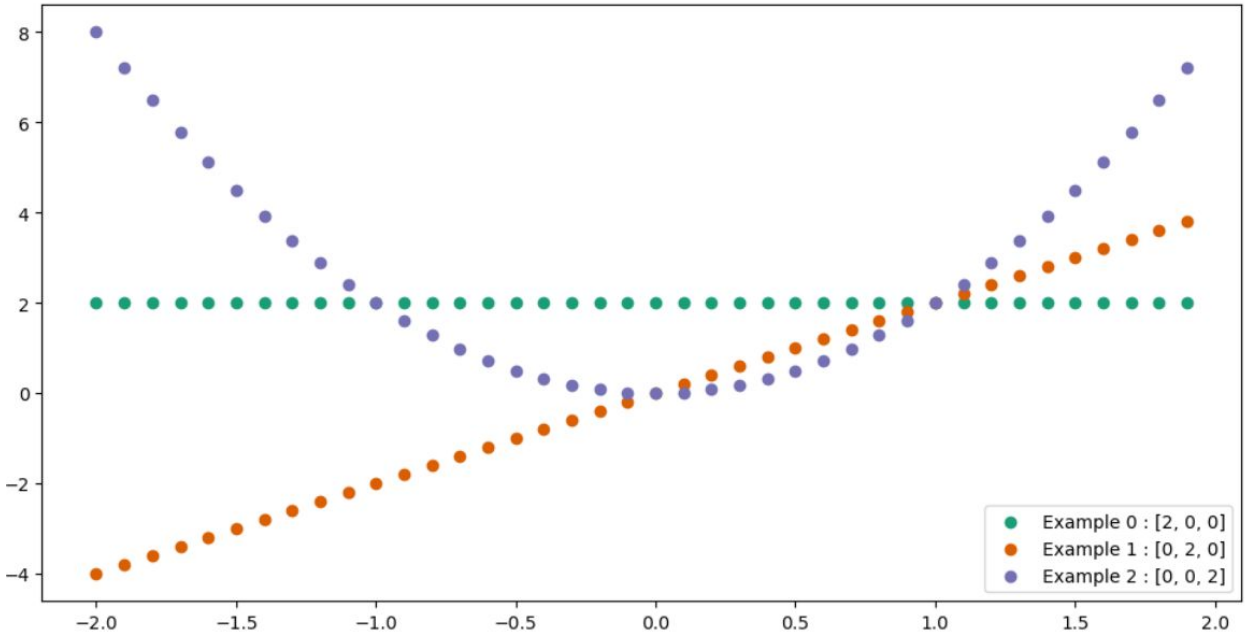


- MvSR is working, have a look at the arXiv
- It has potential to be used in every science
- If you see a potential use case for you science come and talk to me
- Still need some work on our side for a proper full implementation

https://arxiv.org/abs/2402.04298

Multiview Symbolic Regression (MvSR)

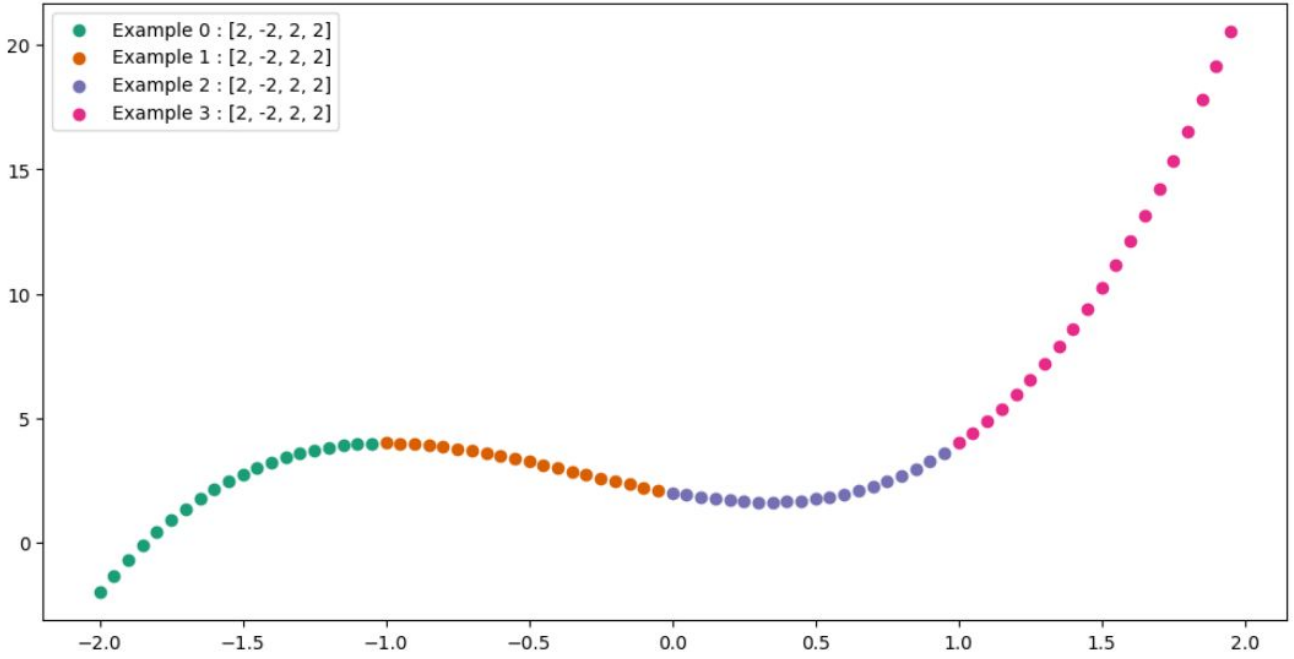
Toy data illustration



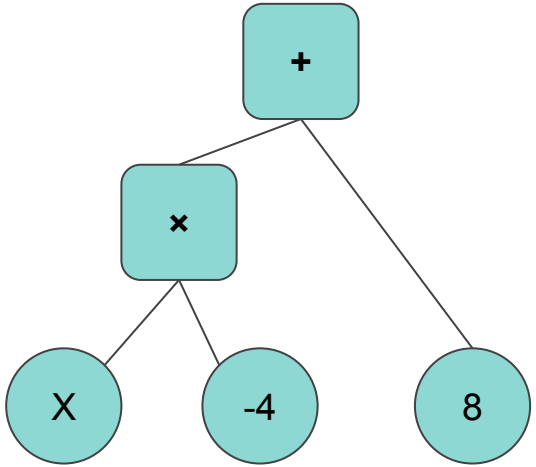
$$f(X) = A + BX + CX^2$$

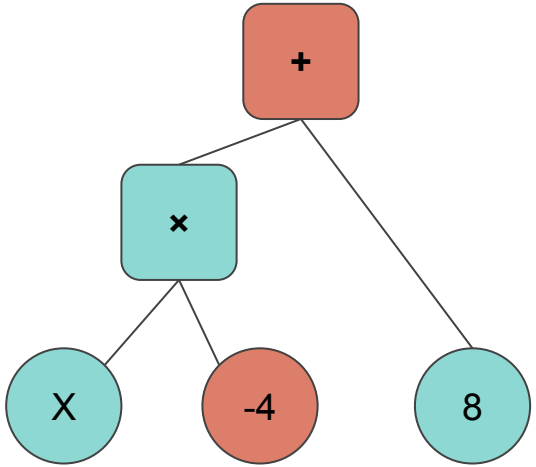
Multiview Symbolic Regression (MvSR)

Toy data illustration

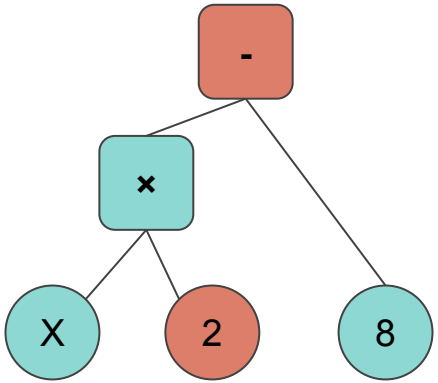


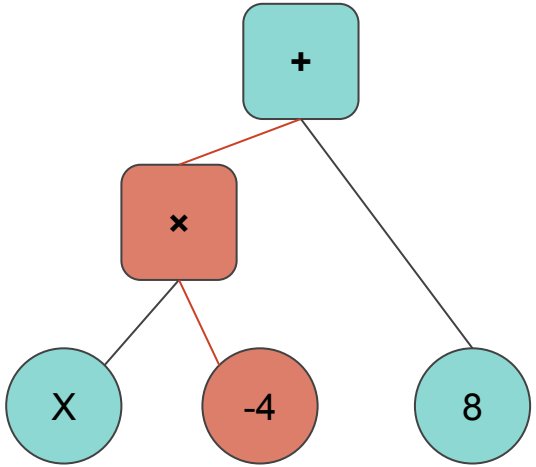
$$f(X) = A + BX + CX^2 + DX^3$$


+
×
X
-4
8

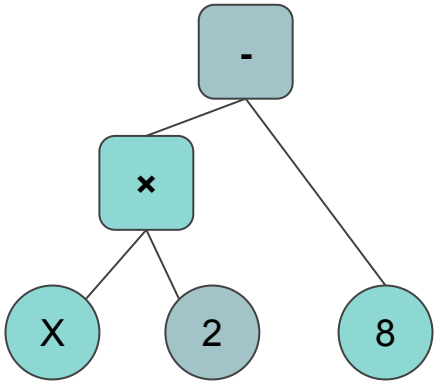


Point mutations

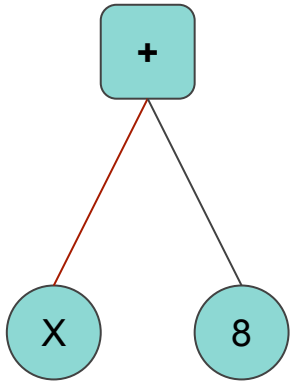


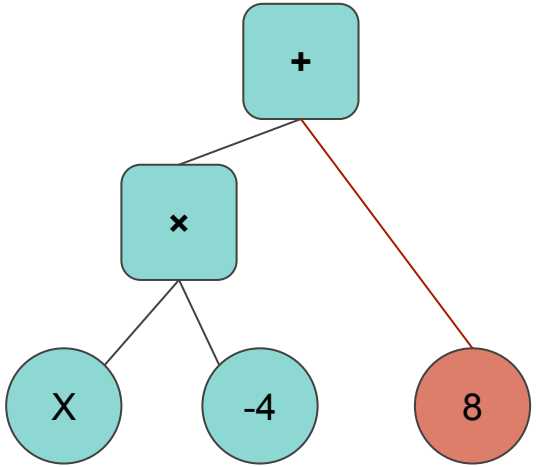


Point mutations

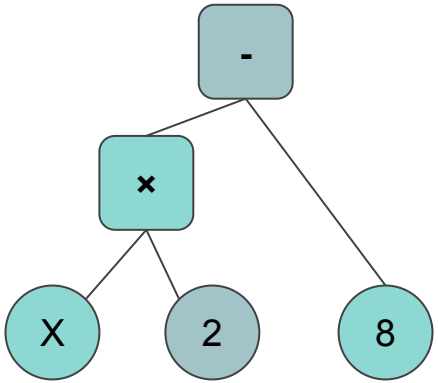


Hoist mutations

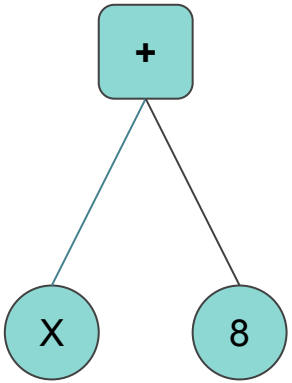




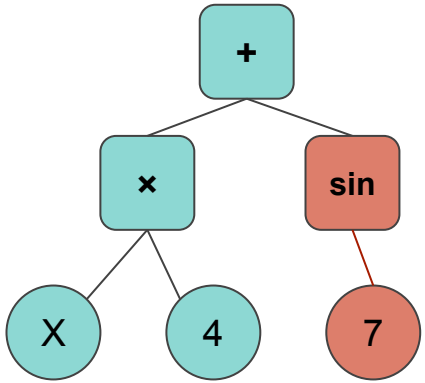
Point mutations

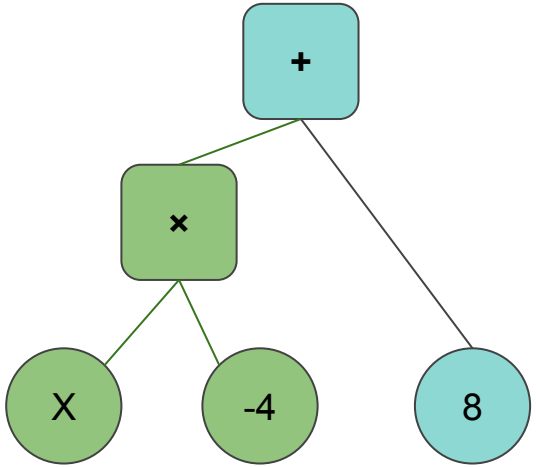


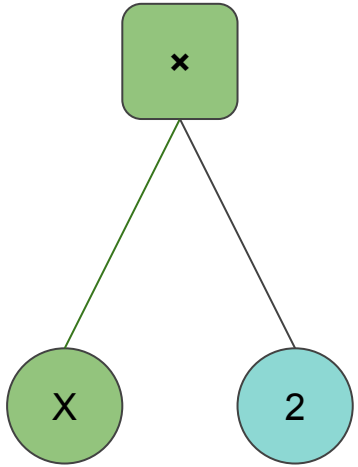
Hoist mutations



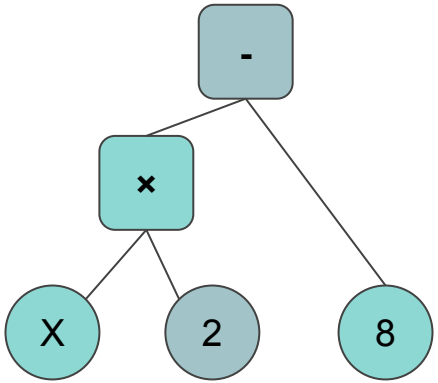
Subtree mutations



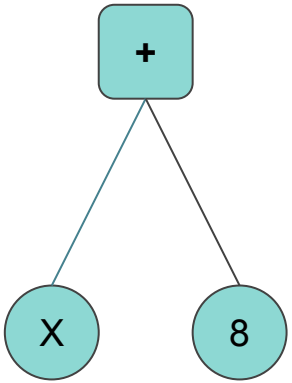




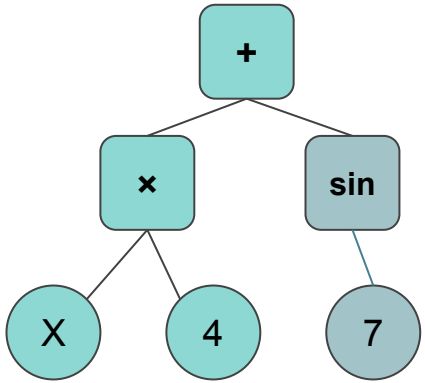
Point mutations



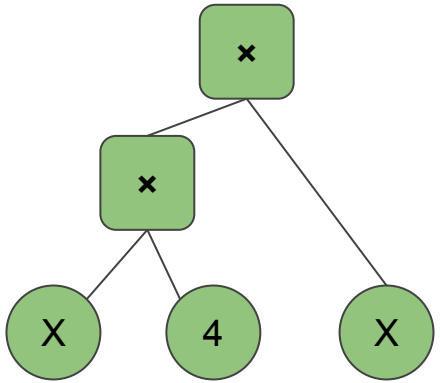
Hoist mutations



Subtree mutations

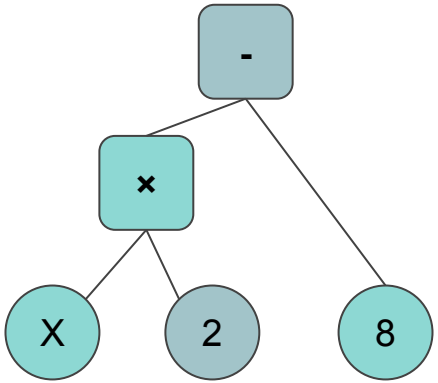


Crossover

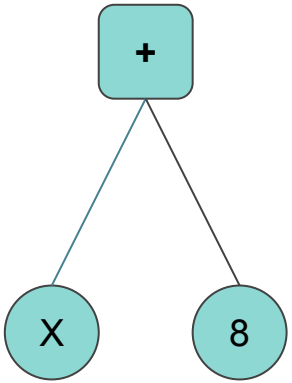


Create a new population from the previous best candidates

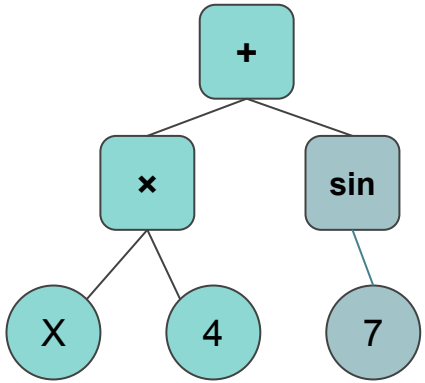
Point mutations



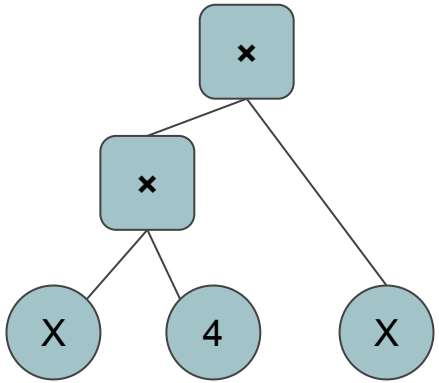
Hoist mutations



Subtree mutations



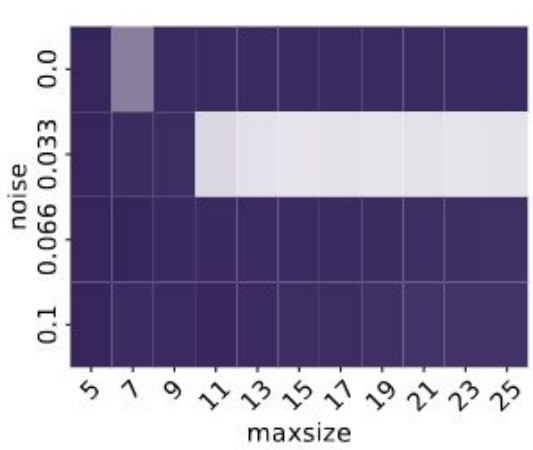
Crossover



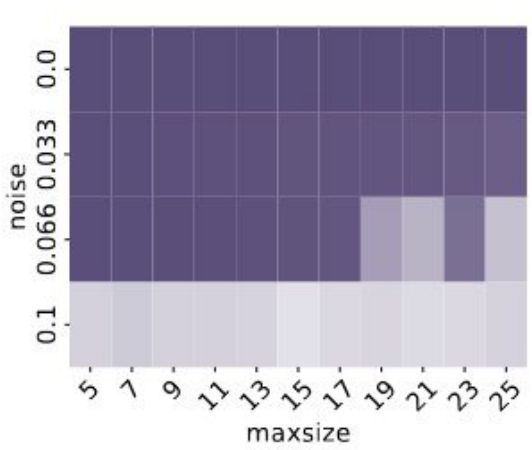
$$f_1(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3$$

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

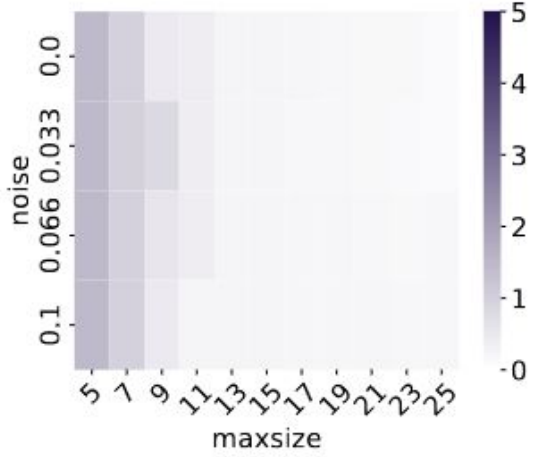
$$f_3(x) = (\theta_0 x_0^2 + (\theta_1 x_1 x_2 - \frac{\theta_2}{(\theta_3 x_1 x_3 + 1)})^2)^{0.5}$$


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

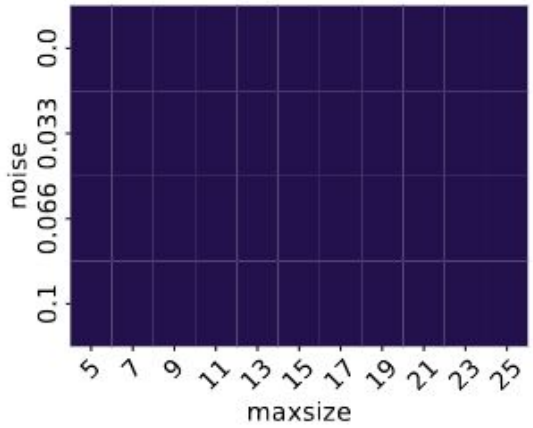
(g)


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

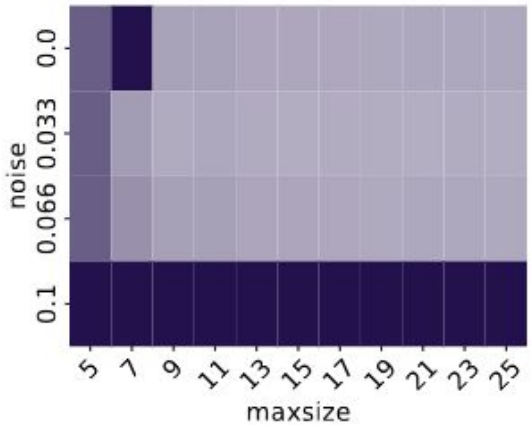
(h)


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

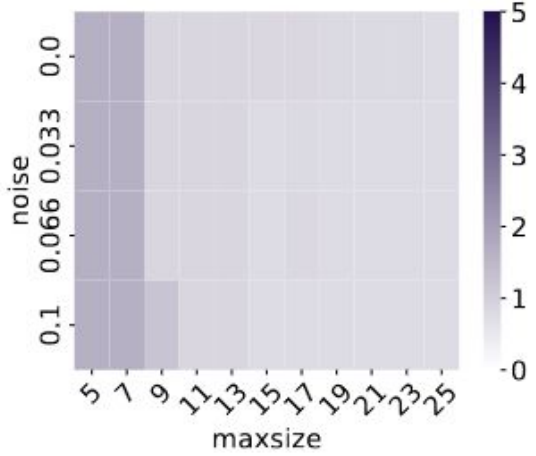
(i)


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

(j)


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

(k)

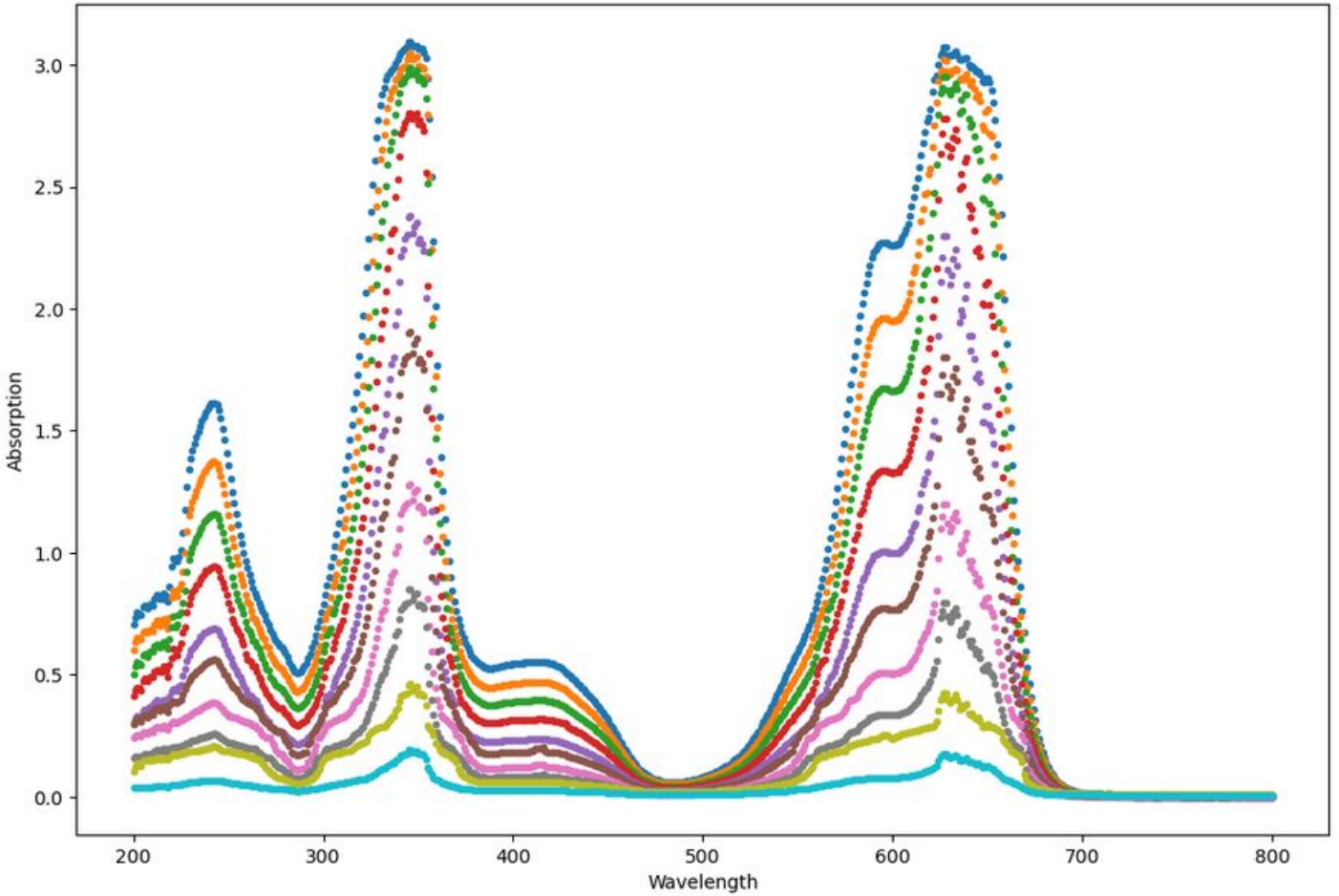

noise
0.0
0.033
0.066
0.1
5 7 9 11 13 15 17 19 21 23 25
maxsize
5
4
3
2
1
0

(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


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