AI for Science

Discovering Governing Equations (And Beyond)

Krzysztof Kacprzyk

van_der_Schaar \LAB
vanderschaar-lab.com

UNIVERSITY OF CAMBRIDGE

kk751@cam.ac.uk

linkedin.com/in/krzysztof-kacprzyk/

Outline

1. Introduction
2. Equation discovery (symbolic regression)
    a. Static
    b. Dynamic
3. ODE discovery for treatment effects
4. Beyond equations
    a. Shape Arithmetic Expressions
    b. Transparent time series forecasting

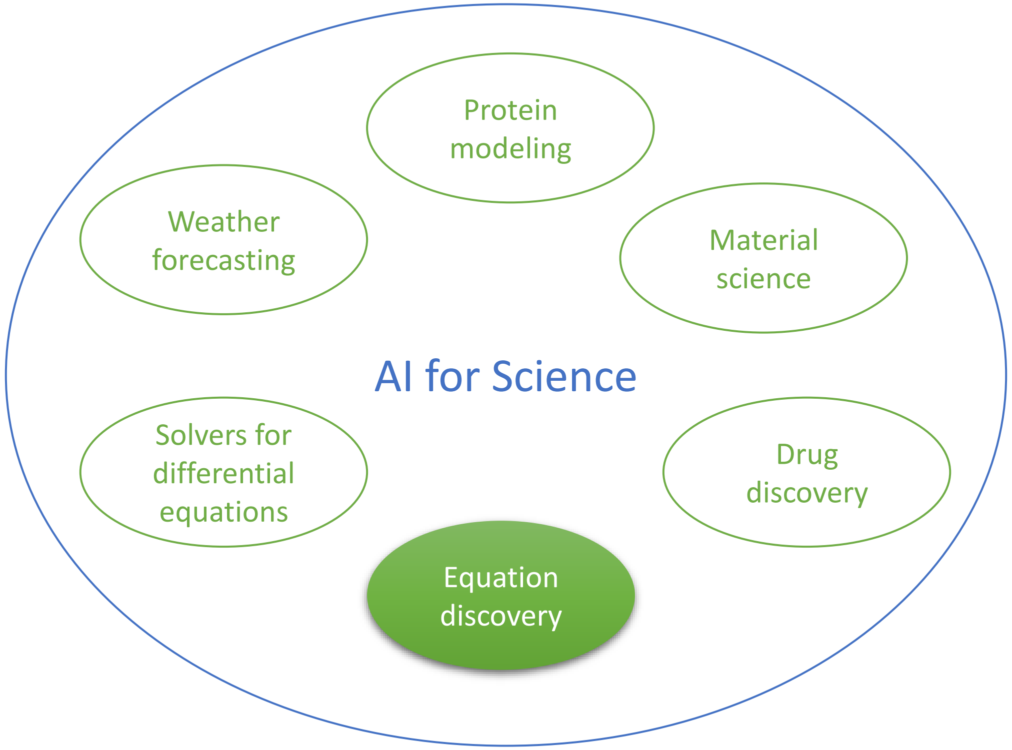
Outline

1. Introduction
2. Equation discovery (symbolic regression)
    a. Static
    b. Dynamic
3. ODE discovery for treatment effects
4. Beyond equations
    a. Shape Arithmetic Expressions
    b. Transparent time series forecasting

AI for Science

Integration of AI into scientific discovery to accelerate research by helping scientists to

- generate hypotheses
- design experiments
- collect and interpret large datasets
- gain insights that might not have been possible using traditional scientific methods alone



AI for Science - community

AI for Science workshops: *ai4sciencecommunity.github.io*
NeurIPS 2021, ICML 2022, NeurIPS 2022, NeurIPS 2023, ICML 2024

Wang, Hanchen, et al. "Scientific discovery in the age of artificial intelligence." *Nature* 620.7972 (2023): 47-60.

github.com/sherrylixuecheng/AI_for_Science_paper_collection

Upcoming workshops at NeurIPS 2024:

- Foundation Models for Science: Progress, Opportunities, and Challenges
- Machine Learning and the Physical Sciences

AI for Science – Equation Discovery

Closed-form expressions

variables

arithmetic operations

$$3\sin(x_1 + x_2) \times e^{2x_3^2}$$

constants

well-known functions

Symbolic Regression (SR)

Applications in physics, biology, medicine, material science

Death After Liver Transplantation: Mining Interpretable Risk Factors for Survival Prediction

Veronica Guidetti, FIM Dept.[1], UNIMORE[2], Modena, Italy, veronica.guidetti@unimore.it
Giovanni Dolci, Infectious Diseases Dept., UNIMORE[2], Modena, Italy, giodolci@hotmail.it
Erica Franceschini, Infectious Diseases Clinic, AOU Policlinico di Modena[3], Modena, Italy, franceschini.erica@aou.mo.it
Erica Bacca, Infectious Diseases Dept., UNIMORE[2], Modena, Italy, erica.bacca@gmail.com

Automated reverse engineering of nonlinear dynamical systems

Josh Bongard*† and Hod Lipson*‡

Symbolic regression in materials science

Yiqun Wang†, Nicholas Wagner†, and James M. Rondinelli, Department of Materials Science and Engineering, Northwestern University, Evanston, IL 60208, USA
Address all correspondence to James M. Rondinelli at jrondinelli@northwestern.edu

ARTICLES
https://doi.org/10.1038/s42256-021-00353-8
nature machine intelligence
Check for updates

Machine learning to guide the use of adjuvant therapies for breast cancer

Ahmed M. Alaa[1], Deepti Gurdasani[2], Adrian L. Harris[3], Jem Rashbass[4] and Mihaela van der Schaar[1,5,6]

SCIENCE ADVANCES | RESEARCH ARTICLE
COMPUTER SCIENCE

AI Feynman: A physics-inspired method for symbolic regression

Silviu-Marian Udrescu[1] and Max Tegmark[1,2]*

Symbolic Regression (SR)

$$E = \frac{mv^2}{2}$$

Dataset

m	v	E
1.0	1.5	1.13
2.3	3.4	13.29
0.5	3.4	2.89
7.8	0.8	2.50
5.6	1.2	4.03
9.8	0.3	0.44

Objective

$$E = f(m, v)$$

$$\frac{1}{D}\sum_{d=1}^{D}\left(E^{(d)} - f(m^{(d)}, v^{(d)})\right)^2$$

How do we search through the space of closed-form expressions?

- Combinatorial in the structural form
- Continuous in the parameters
- NP-hard

Virgolin, M., & Pissis, S. P. (2022). Symbolic Regression is NP-hard. Transactions on Machine Learning Research.

Optimization in SR

Genetic programming Schmidt, M., & Lipson, H. (2009). Distilling Free-Form Natural Laws from Experimental Data. Science, 324(5923), 81–85.

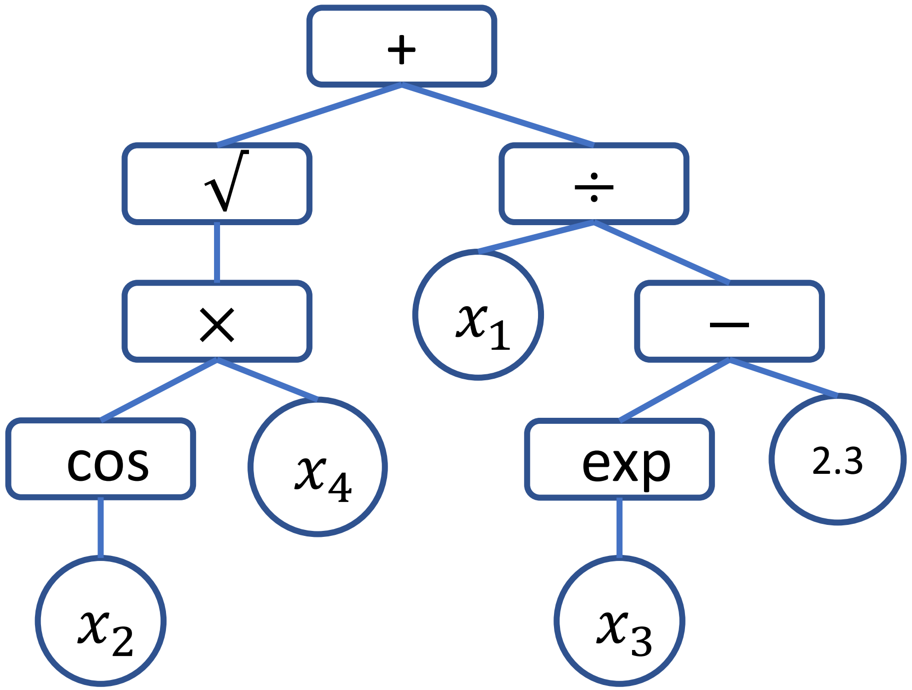
Popular python libraries

- gplearn

✓ pure python implementation
✓ easy to edit and extend
✗ not very efficient

- PySR

✓ fast
✓ complicated constraints
✗ written in Julia



Optimization in SR

Expressions as neural networks

Martius, G. S., & Lampert, C. (2017). Extrapolation and learning equations. *5th International Conference on Learning Representations, ICLR 2017-Workshop Track Proceedings*.

Sahoo, S., Lampert, C., & Martius, G. (2018). Learning Equations for Extrapolation and Control. *Proceedings of the 35th International Conference on Machine Learning*, 4442–4450.

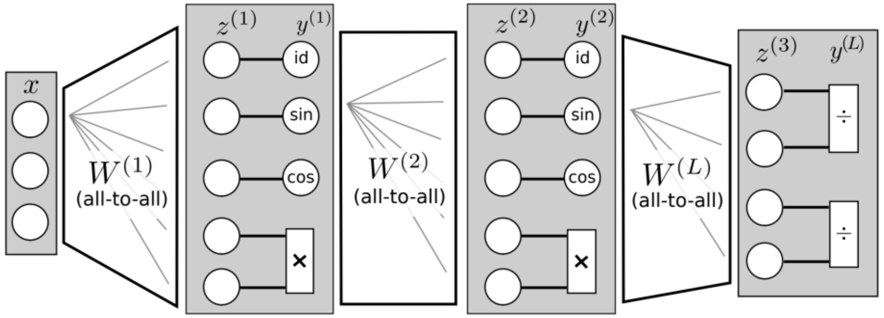


Figure 1. Sahoo et al. (2018)

Optimization in SR

Expressions parametrized through Meijer G-functions

Alaa, A. M., & van der Schaar, M. (2019). Demystifying Black-box Models with Symbolic Metamodels. *Advances in Neural Information Processing Systems*, *32*.

Crabbé, J., Zhang, Y., Zame, W., & van der Schaar, M. (2020). Learning outside the Black-Box: The pursuit of interpretable models.

G-function	Equivalent form
$G^{0,1}_{3,1}\left(\begin{smallmatrix}2,2,2\\1\end{smallmatrix}\middle\vert x\right)$	x
$G^{1,0}_{0,1}\left(\begin{smallmatrix}-\\0\end{smallmatrix}\middle\vert x\right)$	e^{-x}
$G^{1,2}_{2,2}\left(\begin{smallmatrix}1,1\\1,0\end{smallmatrix}\middle\vert x\right)$	$\log(1+x)$
$G^{1,0}_{0,2}\left(\begin{smallmatrix}-\\0,\frac{1}{2}\end{smallmatrix}\middle\vert \frac{x^2}{4}\right)$	$\frac{1}{\sqrt{\pi}}\cos(x)$
$G^{1,2}_{2,2}\left(\begin{smallmatrix}\frac{1}{2},1\\\frac{1}{2},0\end{smallmatrix}\middle\vert x\right)$	$2\arctan(x)$

Table 1. Alaa, A. M., & van der Schaar, M. (2019)

Optimization in SR

Neural Networks to constrain the search space

Udrescu, S.-M., & Tegmark, M. (2020). AI Feynman: A physics-inspired method for symbolic regression. *Science Advances*, *6*(16)

Udrescu, S.-M., Tan, A., Feng, J., Neto, O., Wu, T., & Tegmark, M. (2021). AI Feynman 2.0: Pareto-optimal symbolic regression exploiting graph modularity. *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*.

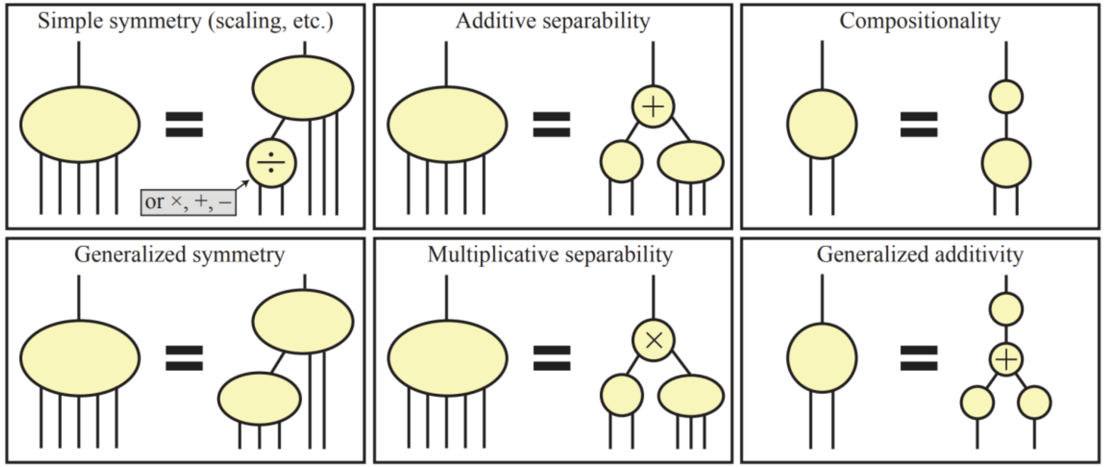


Figure 3. Udrescu et al. (2021).

Optimization in SR

Large pre-trained models

Biggio, L., Bendinelli*, T., Neitz, A., Lucchi, A., & Parascandolo, G. (2021). Neural Symbolic Regression that Scales. *38th International Conference on Machine Learning*.

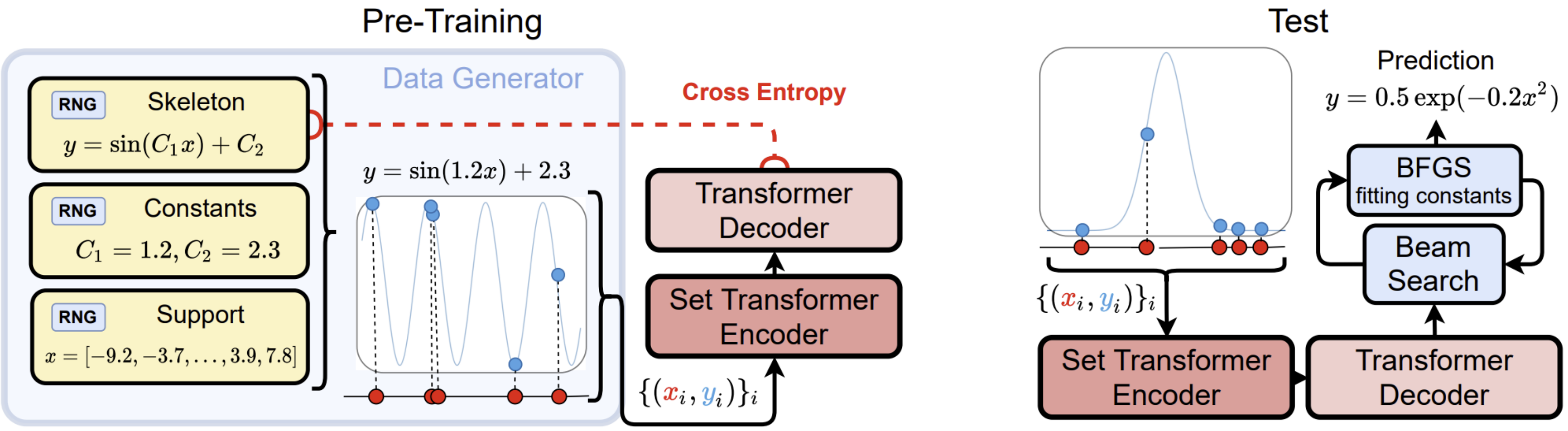


Figure 1. Biggio et al. (2021)

Optimization in SR

Deep Reinforcement Learning

Petersen, B. K., Larma, M. L., Mundhenk, T. N., Santiago, C. P., Kim, S. K., & Kim, J. T. (2021). Deep Symbolic Regression: Recovering Mathematical Expressions From Data via Risk-seeking Policy Gradients. *ICLR 2021*.

Optimization in SR

Recent advances

Holt, S., Qian, Z., & van der Schaar, M. (2023) Deep Generative Symbolic Regression.
In *The Eleventh International Conference on Learning Representations*.

Landajuela, M., Lee, C. S., Yang, J., Glatt, R., Santiago, C. P., Aravena, I., ... & Petersen, B. K. (2022).
A unified framework for deep symbolic regression.
Advances in Neural Information Processing Systems, *35*, 33985-33998.

Kamienny, P. A., d'Ascoli, S., Lample, G., & Charton, F. (2022). End-to-end symbolic regression with transformers.
Advances in Neural Information Processing Systems, *35*, 10269-10281.

Shojaee, P., Meidani, K., Barati Farimani, A., & Reddy, C. (2024). Transformer-based planning for symbolic regression.
Advances in Neural Information Processing Systems, *36*.

Different types of equations

$$F = q(E_f + Bv\sin\theta)$$

$$\begin{cases} \frac{dx}{dt} = \alpha x - \beta xy \\ \frac{dy}{dt} = \delta xy - \gamma y \end{cases}$$

$$E = \frac{mv^2}{2}$$

$$\frac{\partial^2 u}{\partial t^2} = c^2 \nabla^2 u$$

$$\frac{a^3}{T^2} = \frac{GM}{4\pi^2}$$

$$U = mgz$$

$$\frac{\partial u}{\partial t} = \alpha \nabla^2 u$$

$$p = \frac{m_0 v}{\sqrt{1 - v^2/c^2}}$$

$$T = 2\pi\sqrt{\frac{L}{g}}$$

$$m\frac{d^2x}{dt^2} = F(x,t)$$

Differential equations

$$\frac{dN(t)}{dt} = \underbrace{\rho N(t)\log(\frac{K}{N(t)})}_{\text{Growth}} - \underbrace{\beta_c C(t)N(t)}_{\text{Chemotherapy}} - \underbrace{(\alpha d(t) + \beta d(t)^2)N(t)}_{\text{Radiation}}$$

$$\frac{d\rho}{dt} = (\alpha_0 - \mu + \mu \cdot \Lambda) \cdot \rho - \mu \cdot \Lambda \cdot \rho^2$$

$$\frac{dC_1}{dt} = -k_{10}C_1 - k_{12}C_1 + k_{21}C_2$$
$$\frac{dC_2}{dt} = k_{12}C_1 - k_{21}C_2$$

$$\frac{dS}{dt} = -\beta\frac{SI}{N}$$
$$\frac{dI}{dt} = \beta\frac{SI}{N} - \gamma I$$
$$\frac{dR}{dt} = \gamma I$$

Challenges of discovering differential equations

$$E = \frac{mv^2}{2}$$

$$\frac{dx}{dt} = \alpha x - \beta xy$$

Dataset

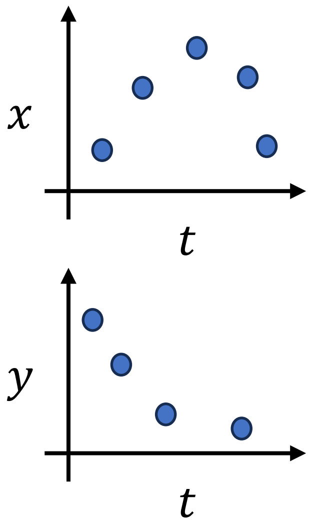
m	v	E
1.0	1.5	1.13
2.3	3.4	13.29
0.5	3.4	2.89
7.8	0.8	2.50
5.6	1.2	4.03
9.8	0.3	0.44

Objective

$$E = f(m, v)$$

$$\frac{1}{D}\sum_{d=1}^{D}\left(E^{(d)} - f\left(m^{(d)}, v^{(d)}\right)\right)^2$$

Dataset



$$\frac{dx}{dt}\ ?$$

Objective

?

SINDy

Brunton, S. L., Proctor, J. L., & Kutz, J. N. (2016). Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, *113*(15), 3932–3937.

$$\dot{x}(t) = \sum_i \alpha_i g_i(x, t)$$

$$\Theta(\mathbf{X}) = \begin{bmatrix} | & | & | & | & & | & | & \\ 1 & \mathbf{X} & \mathbf{X}^{P_2} & \mathbf{X}^{P_3} & \cdots & \sin(\mathbf{X}) & \cos(\mathbf{X}) & \cdots \\ | & | & | & | & & | & | & \end{bmatrix}.$$

$$\dot{\mathbf{X}} = \begin{bmatrix} \dot{\mathbf{x}}^T(t_1) \\ \dot{\mathbf{x}}^T(t_2) \\ \vdots \\ \dot{\mathbf{x}}^T(t_m) \end{bmatrix}$$

$$\dot{\mathbf{X}} = \Theta(\mathbf{X})\Xi.$$

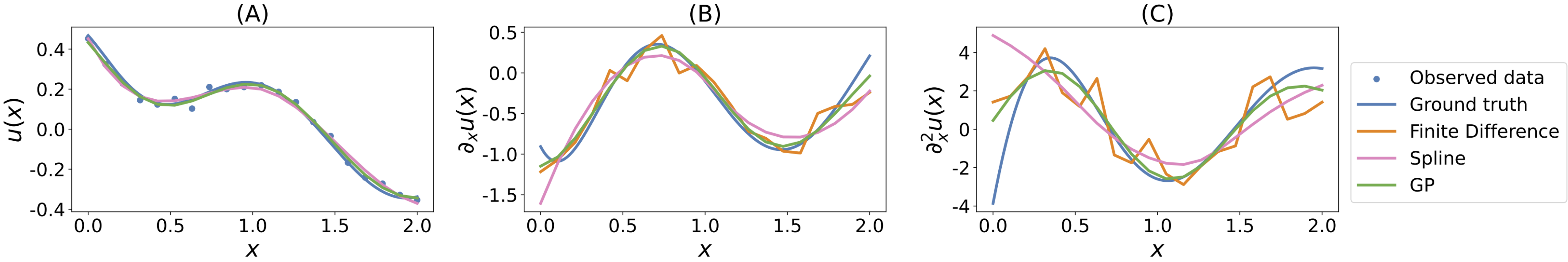
Kaheman, K., Kutz, J. N., & Brunton, S. L. (2020). SINDy-PI: A robust algorithm for parallel implicit sparse identification of nonlinear dynamics. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *476*(2242)

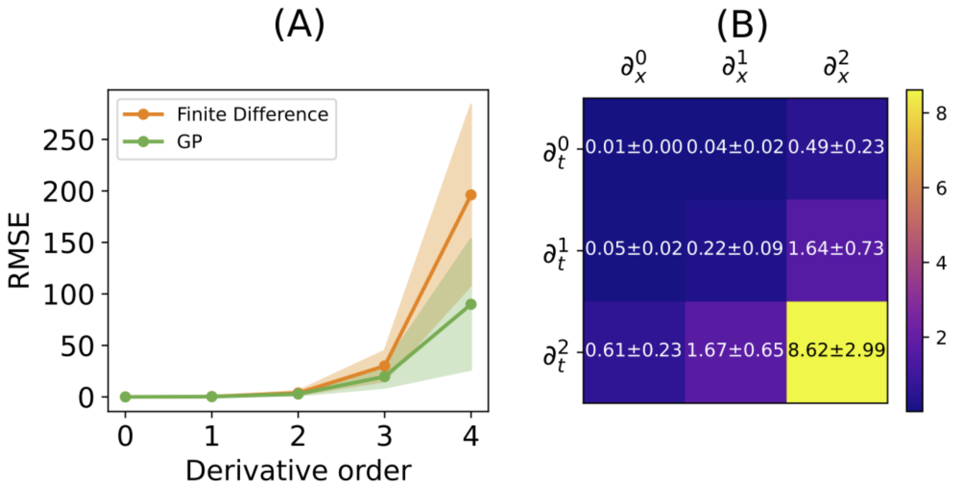
Rudy, S. H., Brunton, S. L., Proctor, J. L., & Kutz, J. N. (2017). Data-driven discovery of partial differential equations. *Science Advances*, *3*(4)

Brunton, S. L., Proctor, J. L., & Kutz, J. N. (2016). Sparse identification of nonlinear dynamics with control (SINDYc).

PySINDy package: https://github.com/dynamicslab/pysindy

Why not just estimate the derivatives?





Our solution: Use variational trick!

$$\frac{dx(t)}{dt} - \alpha x(t) + \beta x(t)y(t) = 0 \Leftrightarrow \int_a^b \left(\frac{dx(t)}{dt} - \alpha x(t) + \beta x(t)y(t)\right)\boldsymbol{\phi(t)dt = 0 \ \ \forall\phi}$$

$$\Leftrightarrow \int_a^b \frac{dx(t)}{dt}\phi(t) - \big(\alpha x(t) - \beta x(t)y(t)\big)\phi(t)dt = 0 \ \ \forall\phi \ \ \Leftrightarrow \int_a^b x(t)\frac{d\phi(t)}{dt} + \big(\alpha x(t) - \beta x(t)y(t)\big)\phi(t)dt = 0 \ \ \forall\phi$$

No $\frac{dx}{dt}$!

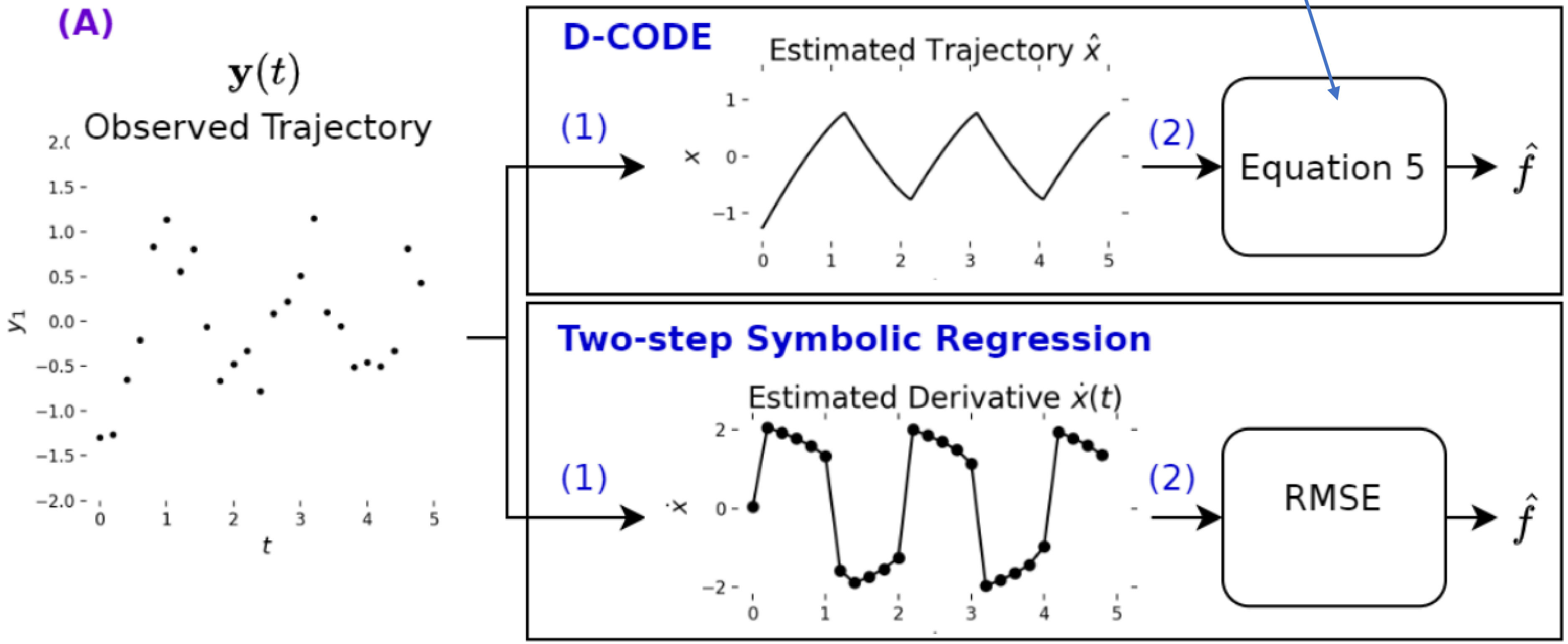
$$\int_a^b \frac{dx(t)}{dt}\phi(t) = -\int_a^b x(t)\frac{d\phi(t)}{dt} + [x(t)\phi(t)]_a^b$$

D-CODE: Algorithm

Qian, Z., Kacprzyk, K. & van der Schaar, M. D-CODE: Discovering Closed-form ODEs from Observed Trajectories. (ICLR 2022)

$$C_j(f, \boldsymbol{x}, g) := \int_0^T f(\boldsymbol{x}(t))g(t)dt + \int_0^T x_j(t)\dot{g}(t)dt$$

$$\hat{f}_j = \arg\min_f \sum_{i=1}^{N}\sum_{s=1}^{S} C_j(f, \hat{\boldsymbol{x}}_i, g_s)^2$$



Kacprzyk, K., Qian, Z. & van der Schaar, M. D-CIPHER: Discovery of Closed-form Partial Differential Equations. (NeurIPS 2023)

What about higher order ODEs and PDEs?

$$\frac{du}{dt}$$

$$\frac{\partial u}{\partial t} \quad \frac{\partial u}{\partial x} \quad \frac{\partial u}{\partial y}$$

$$\frac{\partial^2 u}{\partial t^2} \quad \frac{\partial^2 u}{\partial x^2} \quad \frac{\partial^2 u}{\partial y^2}$$

$$\frac{\partial^2 u}{\partial t \partial x} \quad \frac{\partial^2 u}{\partial t \partial y} \quad \frac{\partial^2 u}{\partial x \partial y}$$

$$u\frac{\partial u}{\partial t} \quad u^2\frac{\partial u}{\partial t} \quad u\frac{\partial u}{\partial x}$$

Difficult to search

Variational trick may not work

Assumptions made by current discovery methods

Evolution assumption: $$\underbrace{\frac{\partial u}{\partial t}} = \underbrace{\alpha \nabla^2 u + \sin(x + t)}$$

✗ $\nabla \cdot \boldsymbol{E} = \frac{\rho}{\epsilon_0 \epsilon_r}$

Linear combinations: $$\frac{dx}{dt} - \alpha x + \beta xy = 0$$

$$\sum_{p=1}^{P} \theta_p f_p(\boldsymbol{x}, \boldsymbol{u}(\boldsymbol{x}), \partial^{[K]} \boldsymbol{u}(\boldsymbol{x})) = 0$$

✗ $\sin(\theta x_i)$

✗ $e^{\theta x_i}$

Weak SINDy

Messenger, D. A., & Bortz, D. M. (2021). Weak SINDy: Galerkin-Based Data-Driven Model Selection. Multiscale Modeling & Simulation, 19(3), 1474–1497. https://doi.org/10.1137/20M1343166

Messenger, D. A., & Bortz, D. M. (2021). Weak SINDy for partial differential equations. *Journal of Computational Physics*, *443*, 110525.

Reinbold, P. A. K., Gurevich, D. R., & Grigoriev, R. O. (2020). Using noisy or incomplete data to discover models of spatiotemporal dynamics. *Physical Review E*, *101*(1), 010203.

Current methods that utilize variational formulation

- make the evolution assumption and
- assume the PDE to be in a linear combination form or
- work only for explicit first order ODEs (D-CODE)

Derivative-bound and derivative-free part

$$\underbrace{f(\boldsymbol{x}, \boldsymbol{u}(\boldsymbol{x}), \partial^{[K]}\boldsymbol{u}(\boldsymbol{x}))}_{\text{derivative-bound}} - \underbrace{g(\boldsymbol{x}, \boldsymbol{u}(\boldsymbol{x}))}_{\text{derivative-free}} = 0$$

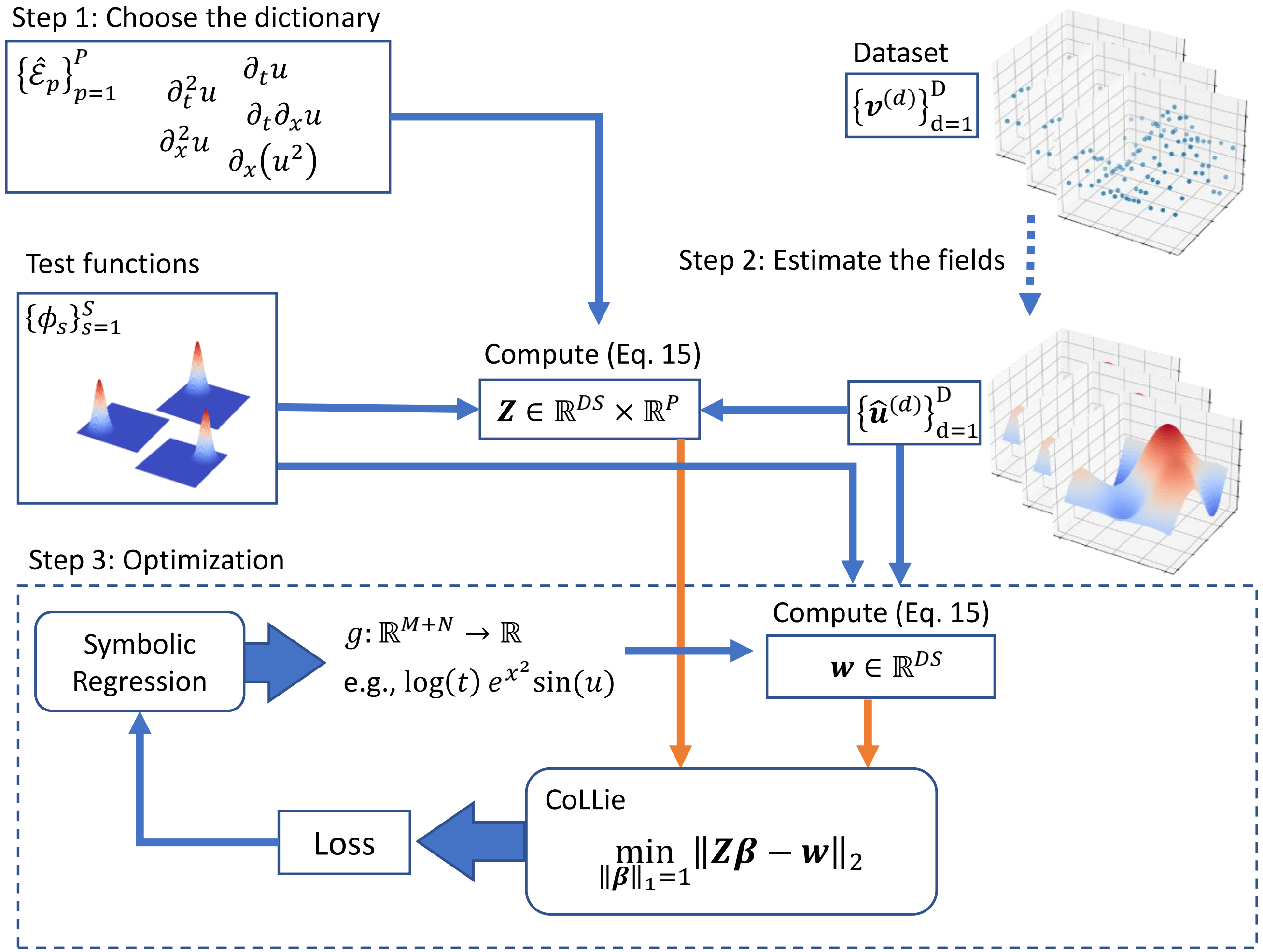
derivative-bound → Requires some technical constraints

derivative-free → No additional constraints!

Requires some technical constraints + No additional constraints! → Variational-Ready PDEs (VR-PDEs)

D-CIPHER

- No linear combination assumption
- No evolution assumption
- Searches through all closed-form derivative-free parts
- Uses variational formulation
- Searches through a linear subspace of derivative-bound parts



Kacprzyk, K., Qian, Z. & van der Schaar, M. D-CIPHER: Discovery of Closed-form Partial Differential Equations.
(NeurIPS 2023)

Summary

Symbolic Regression (SR)

Static

Dynamic

Genetic Programming

Neural Networks

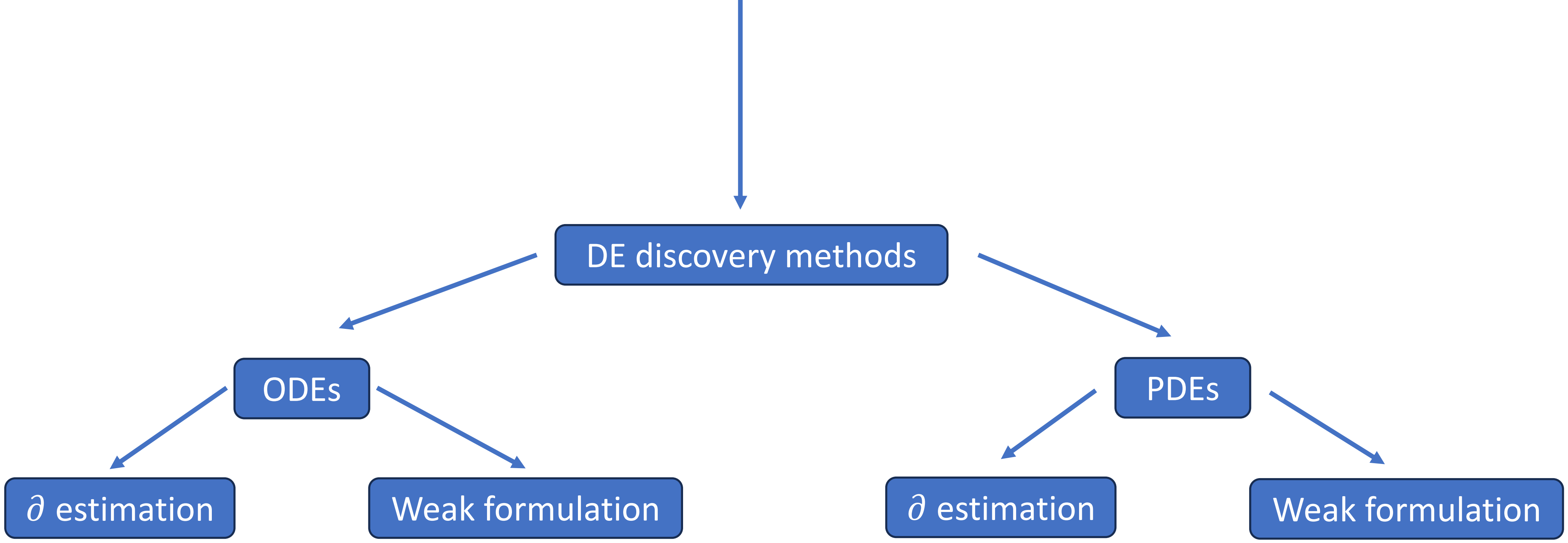
Neurally-guided

Large pre-trained models

Deep Reinforcement Learning

Hybrids

Summary

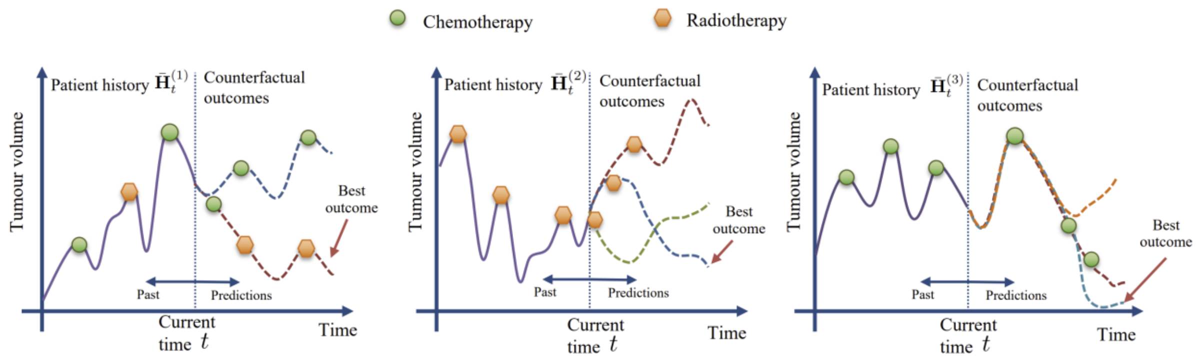


	ODEs: ∂ estimation	ODEs: Weak formulation	PDEs: ∂ estimation	PDEs: Weak formulation
Linear in parameters	SINDy	Weak SINDy	PDE-FIND	Weak SINDy for PDEs
More general	Adapted SR	D-CODE	Adapted SR	D-CIPHER

Outline

1. Introduction
2. Equation discovery (symbolic regression)
    a. Static
    b. Dynamic
3. ODE discovery for treatment effects
4. Beyond equations
    a. Shape Arithmetic Expressions
    b. Transparent time series forecasting

Causality in the real world: treatment effects



(a) Decide treatment plan (b) Decide optimal time of treatment (c) Decide when to stop treatment

Applications:

- **Pharmacology (e.g., drug response)**
- **Physiology (e.g., tumor growth)**
- **Treatment regimes (e.g., treatment plan optimization)**
- **…**
- **Also beyond medicine**

$$\frac{dN(t)}{dt} = \underbrace{\rho N(t)\log\left(\frac{K}{N(t)}\right)}_{\text{Growth}} - \underbrace{\beta_c C(t)N(t)}_{\text{Chemotherapy}} - \underbrace{(\alpha d(t) + \beta d(t)^2)N(t)}_{\text{Radiation}}$$

Let us discover the underlying ODE and use it for treatment effect inference!

*Kacprzyk, K., *Holt, S., *Berrevoets, J., Qian, Z., & van der Schaar, M.
ODE Discovery for Longitudinal Heterogeneous Treatment Effects Inference. (ICLR 2024)

Learning structural (ODE) equation for treatment inference

Advantages over neural networks:

- Interpretable
- Naturally works for irregular sampling and continuous trajectories
- Smaller hypothesis space
- Better performance in certain scenarios

Learning structural (ODE) equation for treatment inference

Treatment Effects Assumptions

Assumption 2.1 (Consistency) *For an observed treatment process $\boldsymbol{A}_{0:T^{(i)}} = \boldsymbol{a}$, the potential outcome is the same as the factual outcome $\boldsymbol{Y}(\boldsymbol{a}) = \boldsymbol{Y}_{0:T^{(i)}}$.*

Assumption 2.2 (Overlap) *The treatment intensity process $\lambda(t|\mathfrak{F}_t)$ is not deterministic given any filtration $\mathfrak{F}_t$*[2] *(Klenke, 2008) and time point $t \in [0, T]$, i.e.,*

$$\gamma < \lambda(t|\mathfrak{F}_t) = \lim_{\delta t \to 0} \frac{p(A_{t+\delta t} - A_t \neq 0|\mathfrak{F}_t)}{\delta t} < 1 - \gamma, \quad \textit{with} \quad \gamma \in (0, 1)$$

Assumption 2.3 (Ignorability) *The intensity process $\lambda(t|\mathfrak{F}_t)$ given the filtration $\mathfrak{F}_t$ is equal to the intensity process that is generated by the filtration $\mathfrak{F} \cup \{\sigma(\boldsymbol{Y}_s) : s > t\}$ that includes the σ-algebras generated by future outcomes $\{\sigma(\boldsymbol{Y}_s) : s > t\}$.*

- Static features not considered in ODE discovery
- ODE discovery methods find only a single equation for a whole dataset
- There are diverse types of treatment: continuous, binary, categorical, or multiple

ODE discovery assumptions

Assumption 3.1 (Existence and Uniqueness) *The underlying process can be modelled by a system of ODEs $\dot{\boldsymbol{x}}(t) = \boldsymbol{F}(\boldsymbol{v}, \boldsymbol{x}(t), \boldsymbol{a}(t))$,*[3] *and for every initial condition $\boldsymbol{x}_0$, $\boldsymbol{v}$ and treatment plan $\boldsymbol{a}$ at t_0, there exists a unique continuous solution $\boldsymbol{x} : [t_0, T] \to \mathbb{R}^d$ satisfying the ODEs for all $t \in (t_0, T)$(Lindelöf, 1894; Ince, 1956).*

Assumption 3.2 (Observability) *All dimensions of all variables in $\boldsymbol{F}$ are observed for all individuals, ensuring sufficient data to identify the system's dynamics and infer the ODE's structure and parameters (Kailath, 1980).*

Assumption 3.3 (Functional Space) *Each ODE in $\boldsymbol{F}$ belongs to some subspace of closed-form ODEs. These are equations that can be represented as mathematical expressions consisting of binary operations $\{+, -, \times, \div\}$, input variables, some well-known functions (e.g., $\{\log, \exp, \sin\}$), and numeric constants (e.g., $\{-0.2, \ldots, 5.2\} \in \mathbb{R}$) (Schmidt & Lipson, 2009).*

Learning structural (ODE) equation for treatment inference

In our paper, we

- provide a general framework which connects ODE discovery with TE

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{F}(\boldsymbol{v}, \boldsymbol{x}(t), \boldsymbol{a}(t)) \quad \mathbb{E}[Y_{t:t+\tau}(\bar{\boldsymbol{a}}_{t:t+\tau})|\boldsymbol{V}, \boldsymbol{X}_{0:t}, \boldsymbol{A}_{0:t}]$$

- reconcile the differences
- propose a 3 step workflow to turn any ODE discovery method into a TE algorithm
- develop INSITE as an instantiation of our framework.

1. New assumptions

	ODE discovery	Treatment effects		Explanation
ref	*assumption*	*assumption*	*ref*	
3.1	*existence & uniqueness*	*consistency*	2.1	2.1 is *implicit* through 3.2.
3.2	*observability*	*overlap*	2.2	2.2 can be relaxed by 3.1 and 3.3
3.3	*functional spaces*	*ignorability*	2.3	2.3 is *similar* as 3.2.

2. Incorporating diverse treatment types

$$\frac{d\boldsymbol{x}(t)}{dt} = \dot{\boldsymbol{x}}(t) = \boldsymbol{F}(\boldsymbol{v}, \boldsymbol{x}(t), \boldsymbol{a}(t)) \quad \text{and} \quad y(t) = g(\boldsymbol{x}(t)),$$

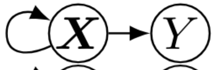
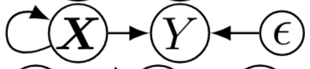
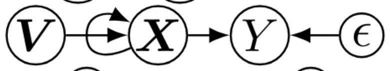
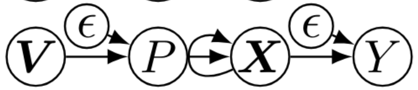
cannot be closed-form if the treatment is categorical

Need to decide how $\boldsymbol{a}$ is incorporated in $\boldsymbol{F}$, so that we can simplify it into simpler closed-form $\boldsymbol{f}$ that we can discover

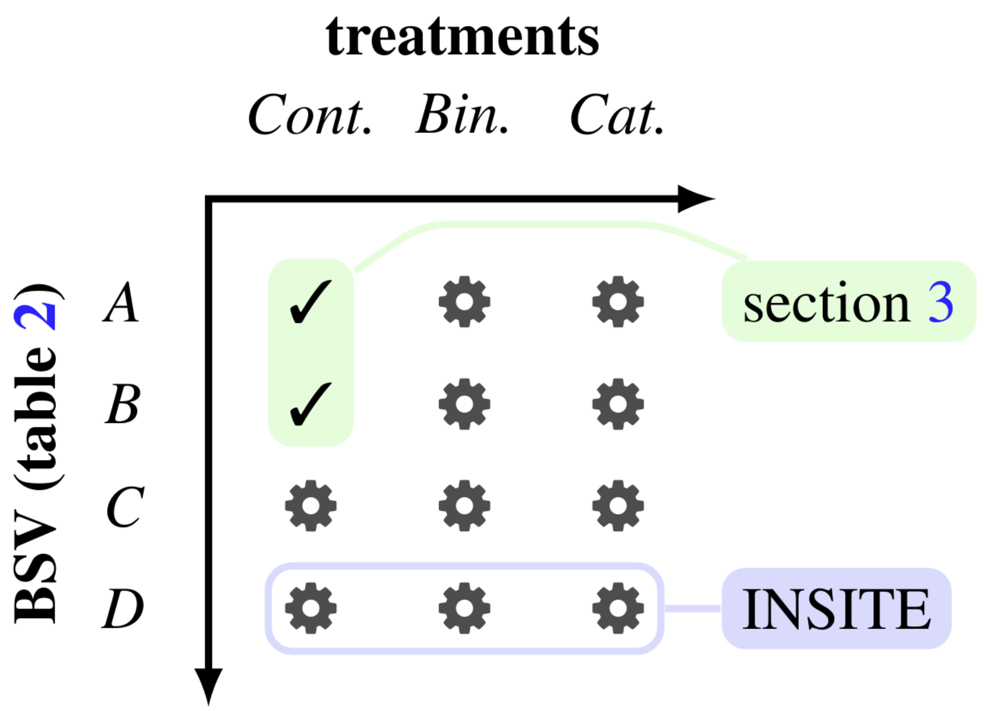
2. Incorporating diverse treatment types

Treatment	S/D	Domain of $\boldsymbol{a}$	Constant	$\boldsymbol{F}(\boldsymbol{x}(t), \boldsymbol{v}, \boldsymbol{a}(t))$
Continuous	S D	$\boldsymbol{a}(t) \in \mathbb{R}^K$	Yes No	$\boldsymbol{f}(\boldsymbol{x}(t), \boldsymbol{v}, \boldsymbol{a}(t))$
Binary	S D	$a(t) \in \{0, 1\}$	Yes Piece-wise	$\boldsymbol{f}_{a(t)}(\boldsymbol{x}(t), \boldsymbol{v})$ or $\boldsymbol{f}_0(\boldsymbol{x}(t), \boldsymbol{v}) + a(t)\boldsymbol{f}_1(\boldsymbol{x}(t), \boldsymbol{v})$
Categorical	S D	$a(t) \in [1, K]$	Yes Piece-wise	$\boldsymbol{f}_{a(t)}(\boldsymbol{x}(t), \boldsymbol{v})$
Multiple	S D	$\boldsymbol{a}(t) \in \{0, 1\}^K$	Yes Piece-wise	$\boldsymbol{f}_{\boldsymbol{a}(t)}(\boldsymbol{x}(t), \boldsymbol{v})$ or $\sum_{i=1}^{K} a_i(t)\boldsymbol{f}_i(\boldsymbol{x}(t), \boldsymbol{v})$

3. Between-subject variability

	(i) ODE.	(ii) +RUV	(iii) +Cov.	(iv) +Dist	Causal graph	Example $y(t)$	Parameters (eq. (5))
A	✓	✗	✗	✗		$x(t)$	$C_0 = c_0, C_1 = c_1$
B	✓	✓	✗	✗		$x(t) + \epsilon$	$C_0 = c_0, C_1 = c_1$
C	✓	✓	✓	✗		$x(t) + \epsilon$	$C_0 = q(c_0), C_1 = q(c_1)$
D	✓	✓	✓	✓		$x(t) + \epsilon$	$C_0 \sim \mathcal{N}(q(c_0), \sigma_0), C_1 \sim \mathcal{N}(q(c_1), \sigma_1)$

Dimensions of our framework



INSITE

From our 3-step plan, we built a method (INSITE)

Based on SINDy – ODE discovery method

Decide how the treatment is modelled and transform the dataset accordingly

Model BSV on level D.
- Covariate model – group level variability
- Parameter distribution – finetune the exact parameters based on the initial trajectory

Outline

1. Beyond equationsIntroduction
2. Equation discovery (symbolic regression)
    a. Static
    b. Dynamic
3. ODE discovery for treatment effects
4. Beyond equations
    a. Shape Arithmetic Expressions
    b. Transparent time series forecasting

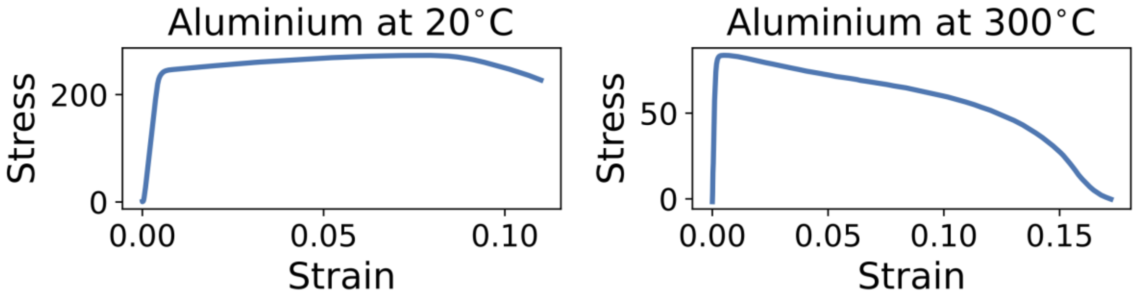
Beyond Closed-Form Equations?

Can we have machine learning models that are not closed-form equations but are equally transparent and interpretable?

Kacprzyk, K., & van der Schaar, M. (2024). Shape Arithmetic Expressions: Advancing Scientific Discovery Beyond Closed-form Equations. AISTATS 2024

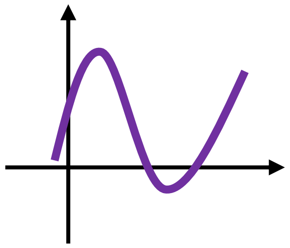
Kacprzyk, K., Liu, T., & van der Schaar, M. (2024). Towards Transparent Time Series Forecasting. ICLR 2024

SR Struggles With Expressions That Are Not Closed-Form



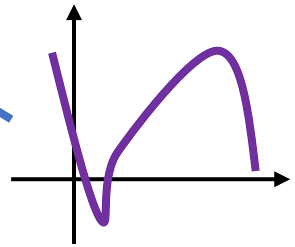
Equation	Size	R^2 score
$y = 63.3e^{-x}$	4	0.163
$y = 78.8 - 285x$	5	0.529
$y = 74.9\cos(7.78x)$	6	0.733
$y = 71.2\cos\left(\frac{x}{x-0.277}\right)$	8	0.750
$y = 147\cos(8.58x - 0.429) - 71.5$	10	0.770
$y = -428x + 428\cos(0.0711\log(x)) - 324$	11	0.836
$y = 428\cos(3.31x - 0.0751\log(1.16x)) - 320$	15	0.933
$y = 168\cos(((7.23 - \cos(e^{-421x}))(x - 2.03)))$	18	0.970

Towards Flexibility: Generalized Additive Models (GAMs)



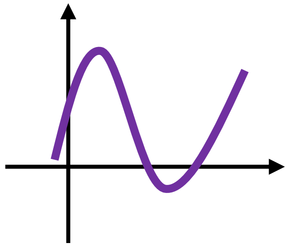
shape functions

$$g(y) = f_1(x_1) + \cdots + f_n(x_n)$$



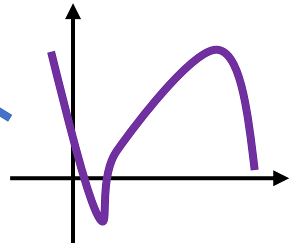
e.g., $g(y) = y$ or $g(y) = (1 + e^{y})^{-1}$

Towards Flexibility: Generalized Additive Models (GAMs)



shape functions

$$g(y) = f_1(x_1) + \cdots + f_n(x_n)$$



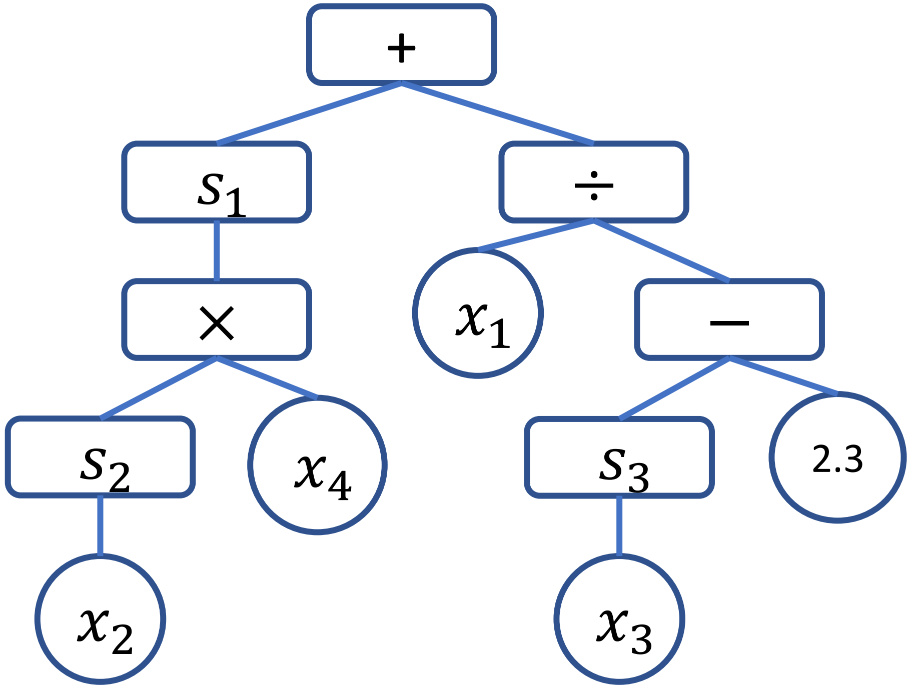
e.g., $g(y) = y$ or $g(y) = (1 + e^y)^{-1}$

Shape Arithmetic Expressions (SHAREs)

shape functions — arithmetic operations

$$s_1(x_4 \times s_2(x_2)) + \frac{x_1}{s_3(x_3) - 2.3}$$

variables — constants



Rule-based Transparency

Not every closed-form expression is transparent (compact enough to understand it).

$$\Downarrow$$

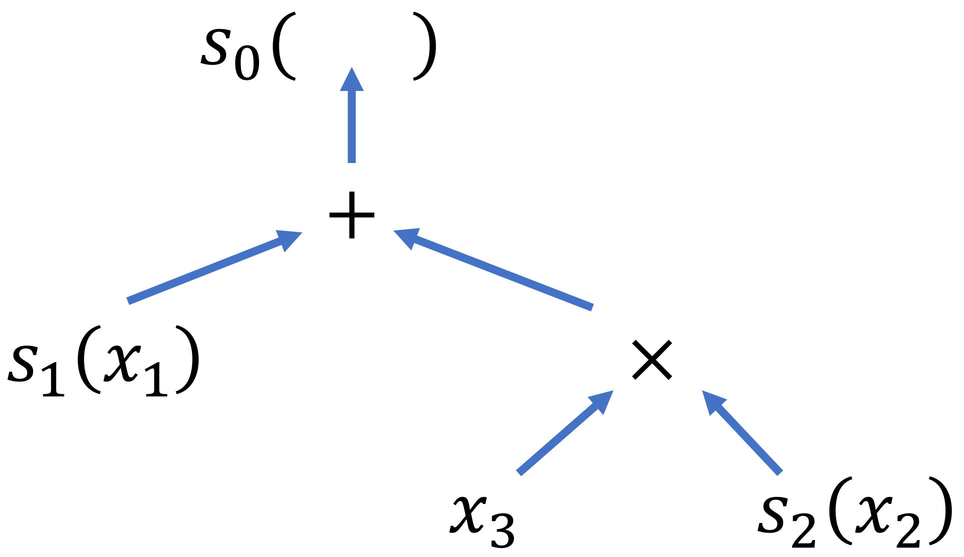
Not every SHARE is transparent. $\Rightarrow$ We need rules for building transparent models.

Number of terms in the expression

Depth of the expression tree

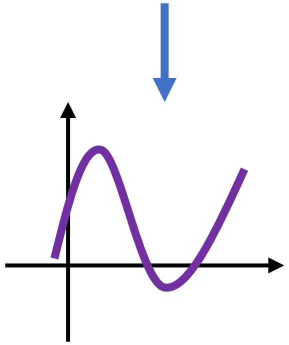
Understanding by Decomposing

$$s_0(s_1(x_1) + x_3 \times s_2(x_2))$$



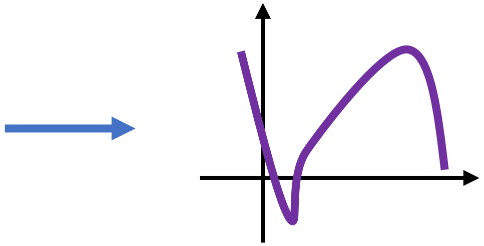
Understanding by Decomposing

$s_1(x_1)$

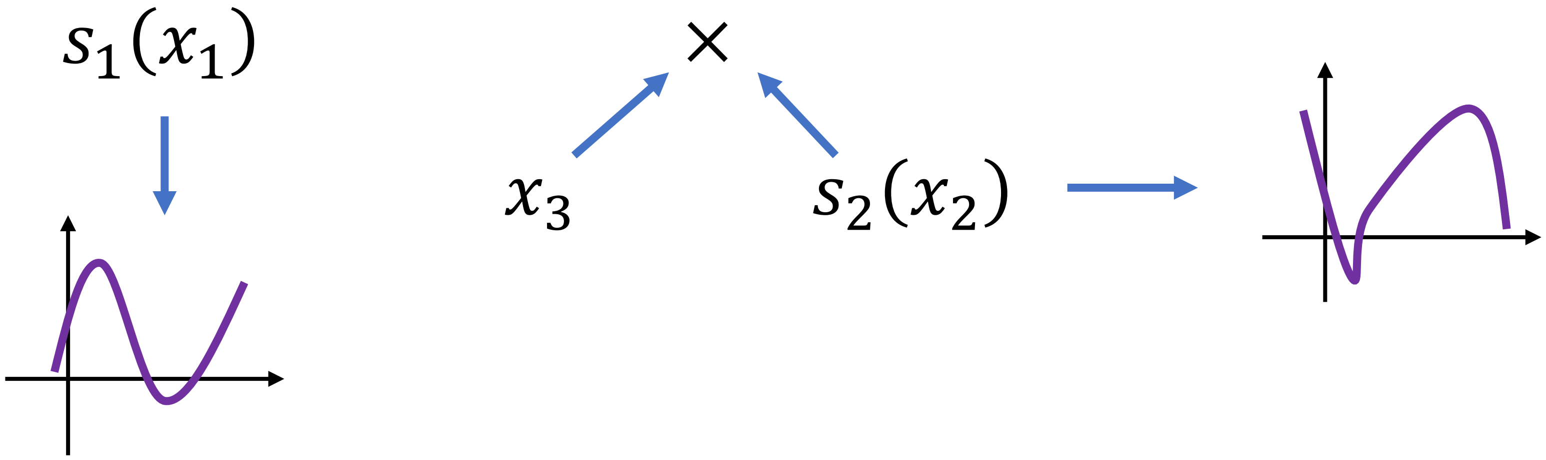


x_3

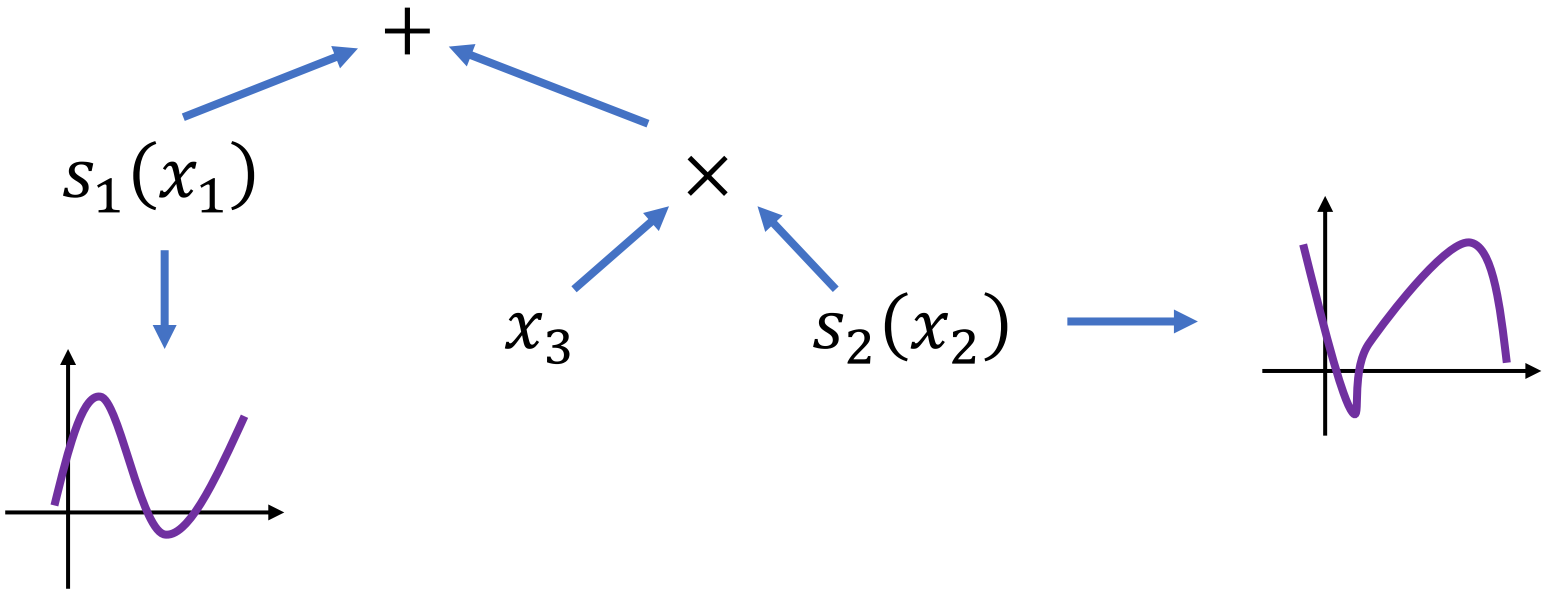
$s_2(x_2)$



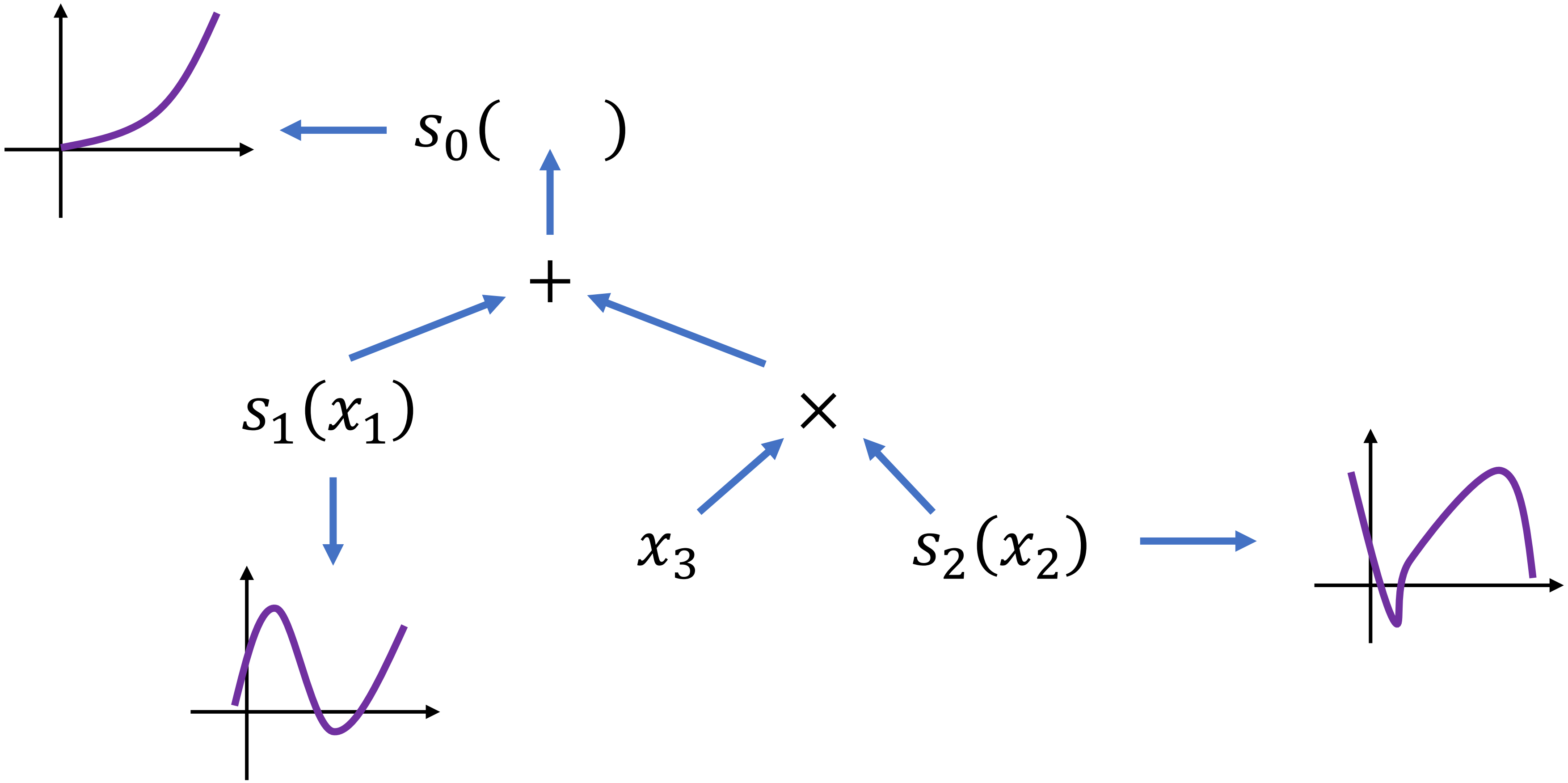
Understanding by Decomposing



Understanding by Decomposing



Understanding by Decomposing



Rule-based Transparency

Not every closed-form expression is transparent (compact enough to understand it).

$\Downarrow$

Not every SHARE is transparent. $\Longrightarrow$ We need rules for building transparent models.

Rule 1 (Univariate composition): Let s be any univariate function. $s(x)$ is transparent. If f is transparent then $s \circ f$ is also transparent

*Rule 2 (Disjoint binary operation): Let $b \in \{+, -, \times, \div\}$ be a binary operation. If f and g are transparent and have **disjoint** sets of arguments then $b \circ (f, g)$ is also transparent.*

Transparent SHAREs

A transparent SHARE is a SHARE that satisfies the following criteria:

- Any binary operator is applied to two functions with disjoint sets of variables.
- The argument of a shape function cannot be an output of another shape function, i.e., $s_1(s_2(x))$ is not allowed.
- It does not contain numeric constants.

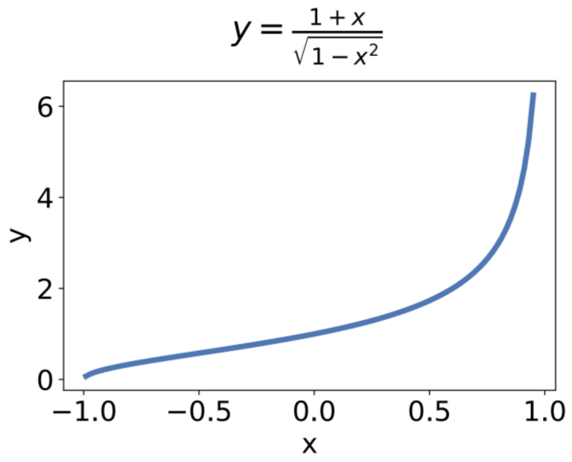
Let $f: \mathbb{R}^n \to \mathbb{R}$ be a transparent SHARE. Then

- The depth of the expression tree of f is at most $2n$
- The number of nodes in the expression tree of f is at most $4n - 2$

Closed-Form Equations Considered as Transparent SHAREs

$$\omega = \frac{1 + v/c}{\sqrt{1 - v^2/c^2}}\omega_0$$

$$\omega = s_1\left(\frac{v}{c}\right)\omega_0$$



SHAREs in Action

Implementation: genetic programming + univariate neural networks

Problem: Given m grams of water (in a liquid or solid form) of temperature t_0 (in °C), what would be the temperature of this water (in a solid, liquid, or gaseous form) after heating it with energy E

SHAREs in Action

Equations found by **SR** when fitted to the temperature dataset.

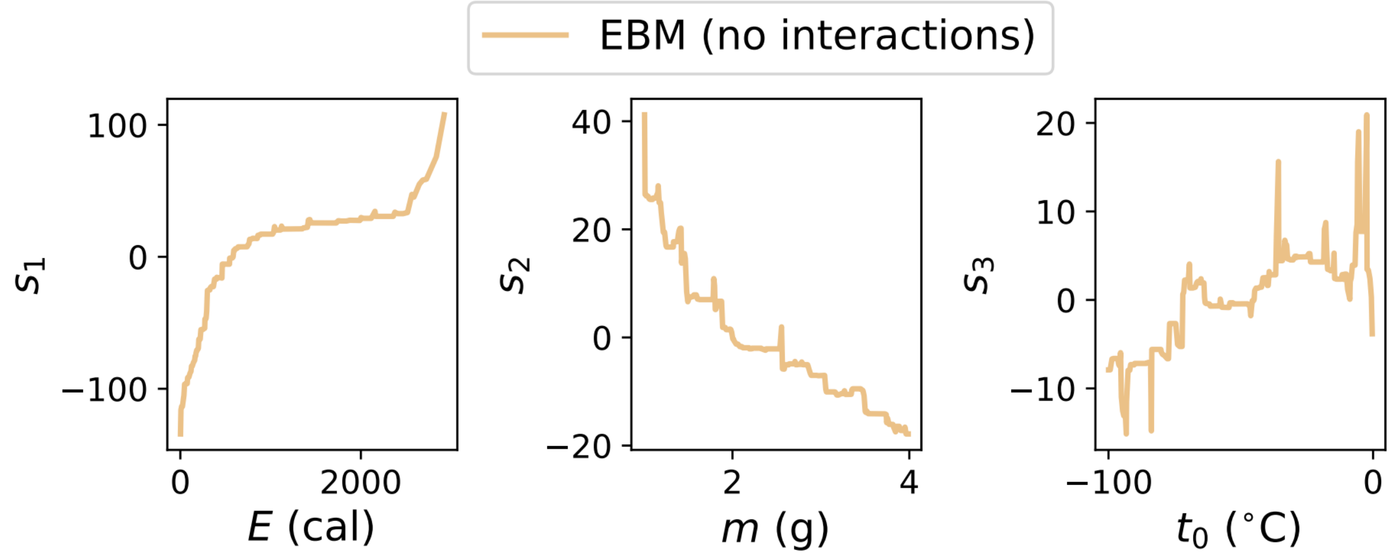
Equation	Size	R^2 score
$y = 13.5\log(E)$	4	0.384
$y = \frac{0.193E}{m}$	5	0.485
$y = 39.4\log\left(\frac{E}{m}\right) - 141$	8	0.733
Appendix C.3 *Equation 7*	17	0.768
Appendix C.3 *Equation 8*	23	0.817
Appendix C.3 *Equation 9*	33	0.840
Appendix C.3 *Equation 10*	40	0.867

Equation 10

$$y = \frac{t_0}{\log(E)} - 1.72e^{e^{\cos\left(\frac{0.0103E}{m}\right)}} + 80.1 + 56.3 \times \cos\left(\log\left(\frac{0.58E}{m} + 31.7\cos\left(\frac{0.021E}{m} + 0.93\right)\right)\right)$$

SHAREs in Action

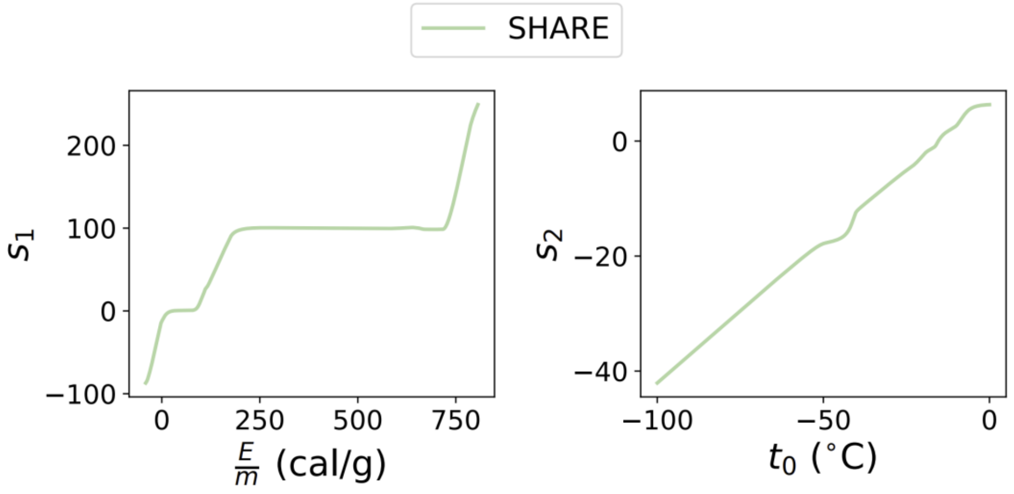
Shape functions from the **GAM** fitted to temperature dataset. R^2 *score: 0.758.*



SHAREs in Action

Found **SHAREs** when fitted to the temperature dataset.

#s	Equation	R^2 score
0	$t = m$	-3.513
1	$t = s_1\left(\frac{E}{m} + t_0\right)$	0.970
2	$t = s_1\left(\left(\frac{E}{m} + s_2(t_0)\right)\right)$	0.999
3	$t = s_1\left(\frac{E+s_2(t_0)}{s_0(m)}\right)$	0.988
4	$t = s_1\left(\frac{E}{s_0(s_3(m)s_2(t_0))}\right)$	0.942



Property	From s_1	Ground truth
Spec. heat cap. of ice ($\frac{cal}{g^\circ C}$)	0.53	0.50
Spec. heat cap. of water ($\frac{cal}{g^\circ C}$)	1.01	1.00
Spec. heat cap. of steam ($\frac{cal}{g^\circ C}$)	0.50	0.48
Heat of fusion ($\frac{cal}{g}$)	78.85	79.72
Heat of vaporization ($\frac{cal}{g}$)	540.91	540.00

What about dynamical systems?

Static Symbolic Regression:

↓

Shape Arithmetic Expressions

Transparent models:

- Generalized additive models
- Linear regression
- Decision trees
- Decision rules/sets

static predictions

$y = 0.87$

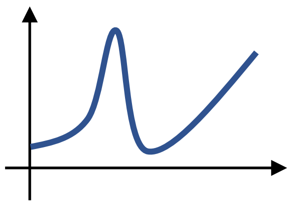
$y =$ „cat”

Dynamic Symbolic Regression:

↓

?

time series forecasting

$y =$ 

?

What does it mean for the model to be transparent when the target is a whole trajectory?

What about dynamical systems?

Published as a conference paper at ICLR 2024

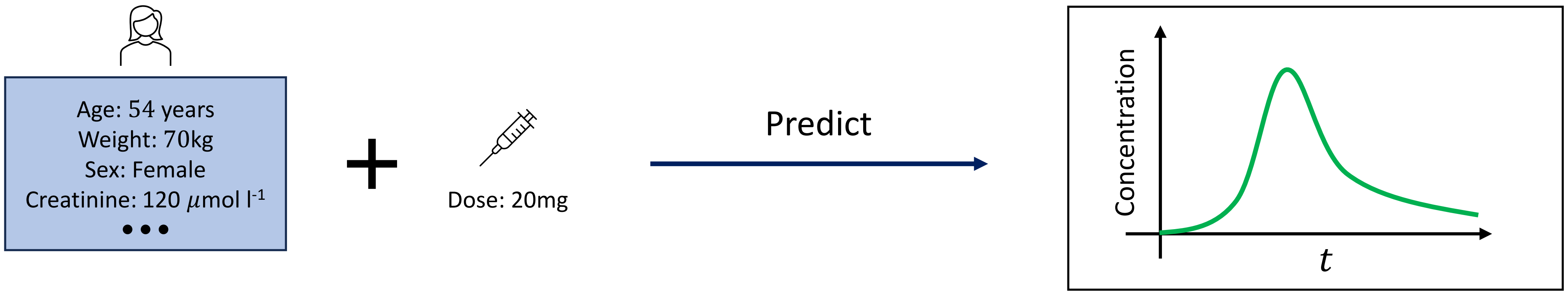
TOWARDS TRANSPARENT TIME SERIES FORECASTING

Krzysztof Kacprzyk
University of Cambridge
kk751@cam.ac.uk

Tennison Liu
University of Cambridge
tl522@cam.ac.uk

Mihaela van der Schaar
University of Cambridge
The Alan Turing Institute
mv472@cam.ac.uk

Population Pharmacokinetic (PopPK) Models



Crucial part of drug development

- How different factors impact drug exposure and if therapeutic individualisation is needed
- Efficacy and safety endpoints
- Dose adjustment

Important for clinical practice

- Dosing guidelines recommend adjusting the dose to achieve a target AUC from PopPK model

Population Pharmacokinetic (PopPK) Models

Age: 54 years
Weight: 70kg
Sex: Female
Creatinine: 120 μmol l^{-1}
• • •

+

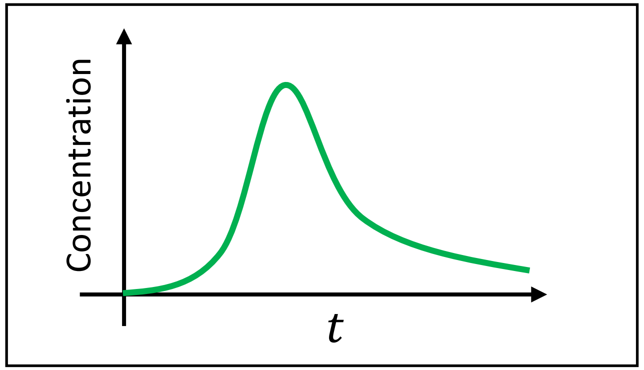
Dose: 20mg

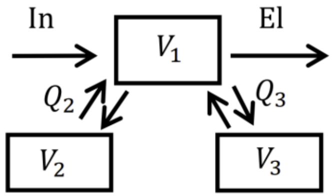
Neural Networks?

Transformers?

Diffusion models?

Predict





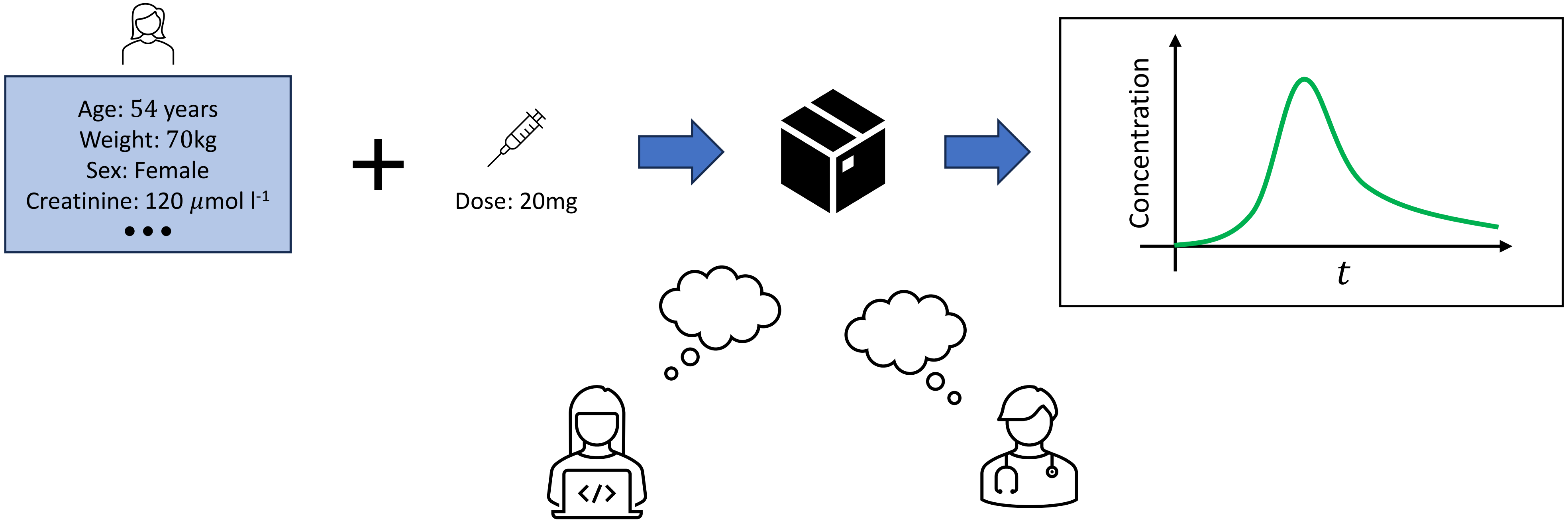
$$\frac{dA_c}{dt} = In - El + k_{21}A_2 - k_{12}A_C + k_{31}A_3 - k_{13}A_C$$

$$\frac{dA_2}{dt} = -k_{21}A_2 + k_{12}A_C$$

$$\frac{dA_3}{dt} = -k_{31}A_3 + k_{13}A_C$$

$$k_{n1} = \frac{Q_n}{V_n} \qquad k_{1n} = \frac{Q_n}{V_1} \qquad Cc = \frac{A_C}{V_1}$$

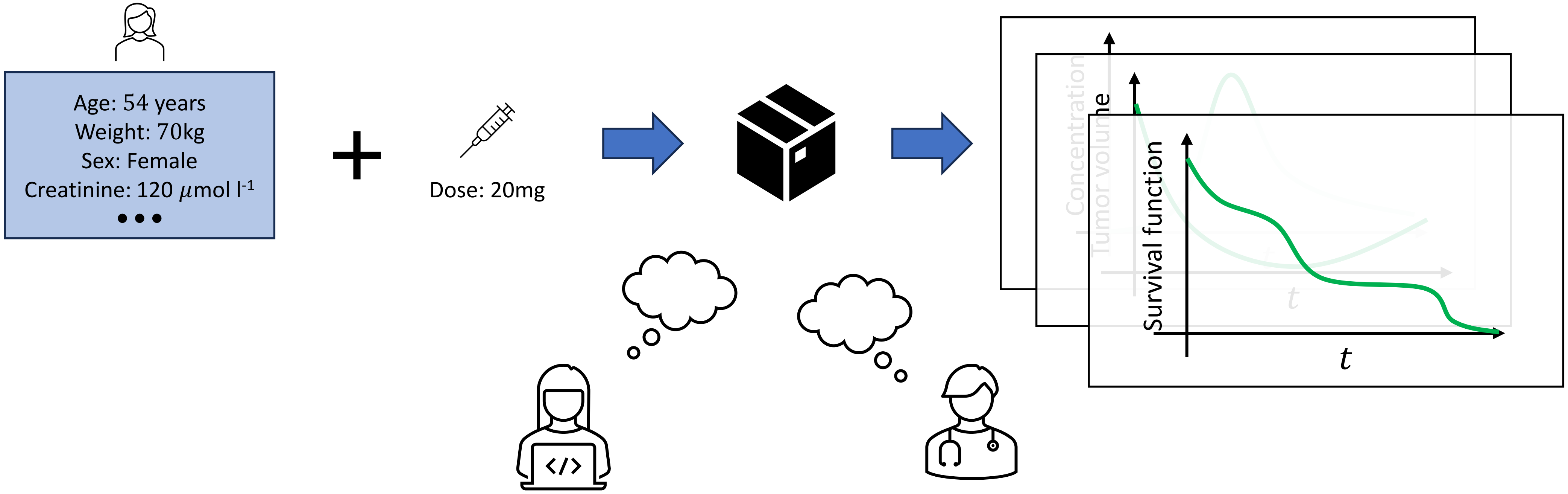
Population Pharmacokinetic Models



- Validate the model
- Debug the model
- Certify

- Understand how various factors influence the prediction

Population Pharmacokinetic Models



- Validate the model
- Debug the model
- Certify

- Understand how various factors influence the prediction

Time Series Forecasting from Static Features

$$\boldsymbol{x} \in \mathbb{R}^N \xrightarrow{\text{Predict}} y: [0, T] \to \mathbb{R}$$

- **How important** is a specific covariate for the prediction?
- **How similar** is this instance to other instances in the dataset?

- **What if**: "What would happen to the model's prediction if a specific covariate changes?"
- **How to be that**: "How should the covariates be modified to get a different prediction?"
- **How to still be this**: "What range of drug dose values keeps the prediction the same?"

Why is it Challenging?

- **What if**: "What would happen to the model's prediction if a specific covariate changes?"
- **How to be that**: "How should the covariates be modified to get a different prediction?"
- **How to still be this**: "What range of drug dose values keeps the prediction the same?"

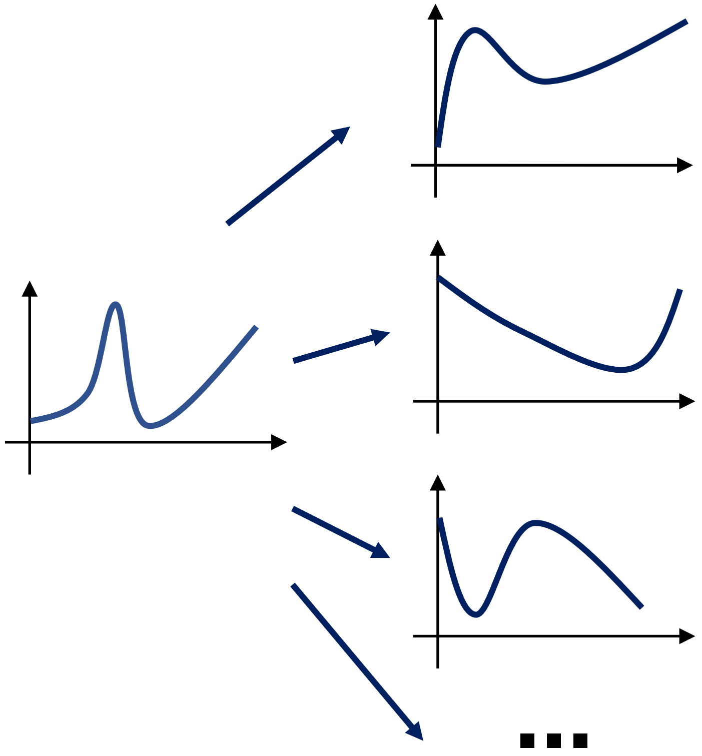
Understanding/Measuring **change** in the trajectory.

More challenging than understanding change in single-label output.

$y = 0.87$ ↓↑→

$y =$ „cat" ⟲ „dog"

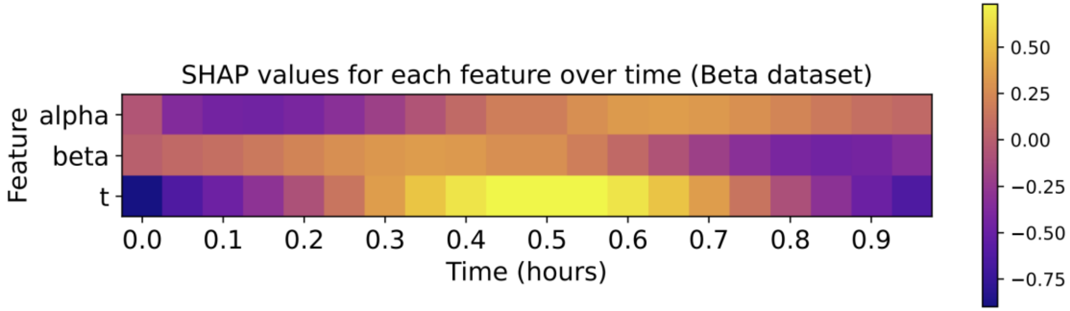
$y =$



Bottom-Up Approach

Feature importance methods have been extended to time series **inputs** but not time series **outputs**.

Bottom-Up: trajectory is understood by looking at its values at individual time points



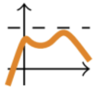
Example: How important is the dose for the drug concentration at $t = 1.5$ *hours?*

Often we want to comprehend the whole trajectory at once.

Example: When administering a drug, we may be less interested in the concentration of the drug every few hours but rather in understanding the entire curve, including properties like the peak plasma concentration and the time when it is achieved

Top-Down Approach

Humans tend to describe trajectories by referring to the trends and properties it exhibits rather than just the values it attains

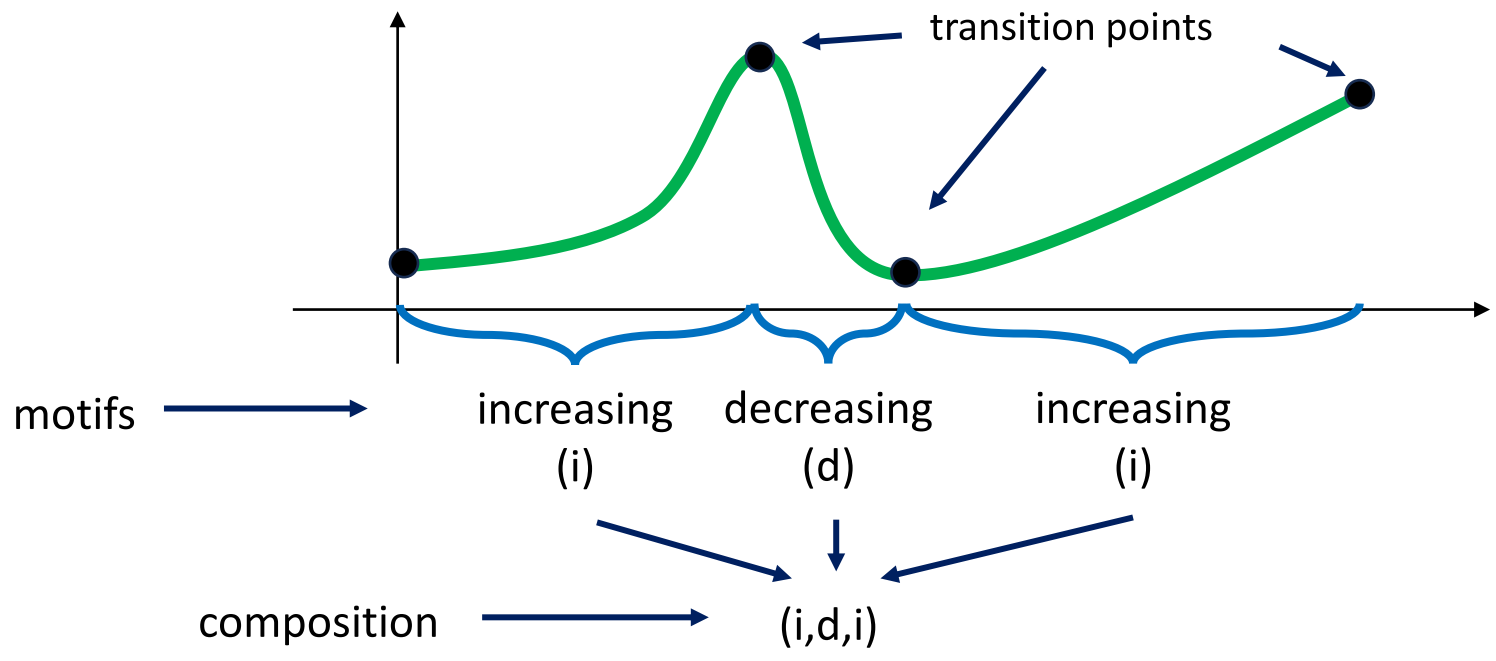
Description	Trend	Properties	Visualization
"The GDP has been steadily increasing for the last 10 years"	increasing	for the last 10 years	
"The blood sugar level in non-diabetic patients should stay below 100mg/dl while fasting"	stay below	below 100mg/dl	
"Tumor volume decreases, obtains a minimum after 6 months, and then increases"	decreases then increases	minimum at 6 months	

Bi-level transparency for time series forecasting.

- Level 1: understanding how the trend (the general shape of the trajectory) changes as we modify the input
- Level 2: understanding how the properties of the current trend (e.g., minimum value) change as we modify the input.

Motifs and Compositions

Motif describes the shape of the trajectory at a particular interval.
E.g., a set of motifs may be: *increasing, decreasing, constant.*



Composition is the shortest sequence of motifs that describes the trajectory.

Dynamical Motifs

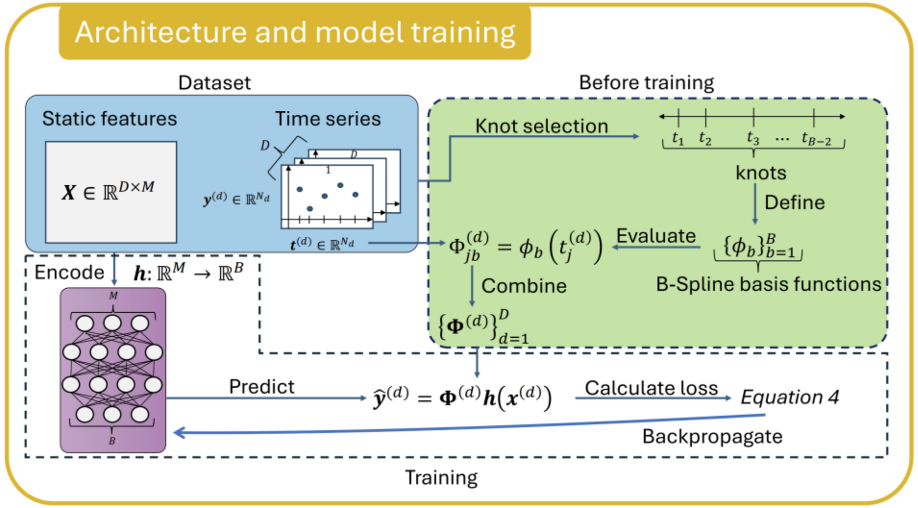
Symbol	Name	Definition	Visualization
s_{+0}	Straight line with positive slope	$f(x) = ax + b,\ a > 0, b \in \mathbb{R}$	
s_{-0}	Straight line with negative slope	$f(x) = ax + b,\ a < 0, b \in \mathbb{R}$	
s_{00}	Straight line with zero slope	$f(x) = b,\ b \in \mathbb{R}$	
s_{++}	Increasing and strictly convex	$f'(x) > 0,\ f''(x) > 0$	
s_{+-}	Increasing and strictly concave	$f'(x) > 0,\ f''(x) < 0$	
s_{-+}	Decreasing and strictly convex	$f'(x) < 0,\ f''(x) > 0$	
s_{--}	Decreasing and strictly concave	$f'(x) < 0,\ f''(x) < 0$	

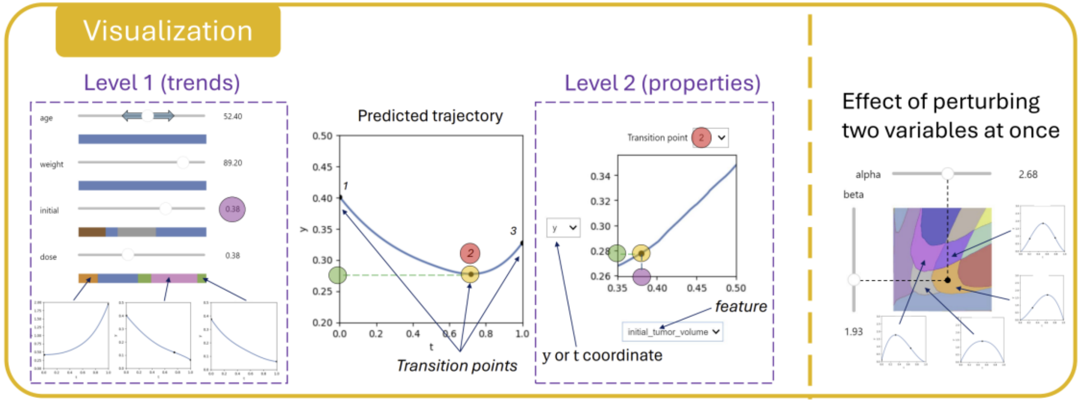
✓ Encode information about the trajectory's first and second derivatives.

✓ Transition points correspond to local minima, maxima, and inflection points.

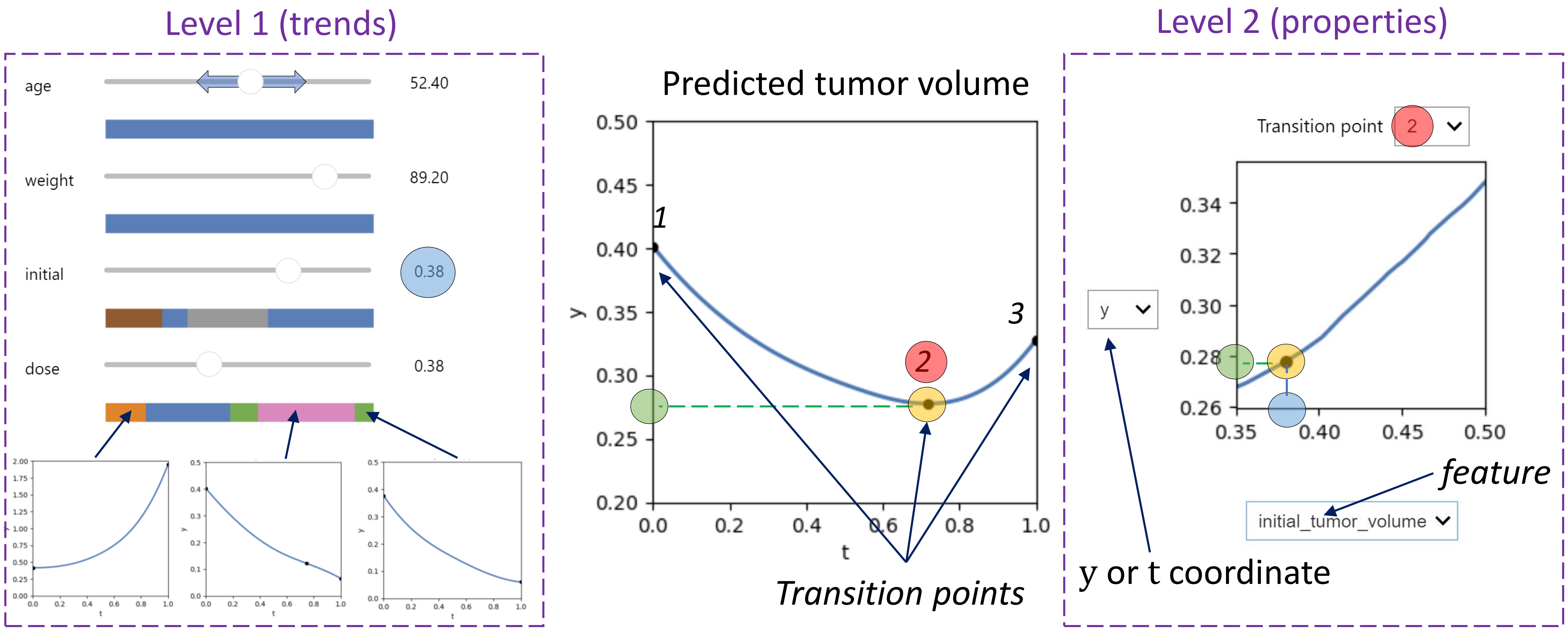
TIMEVIEW

Predictive Model ← TIMEVIEW → Visualization



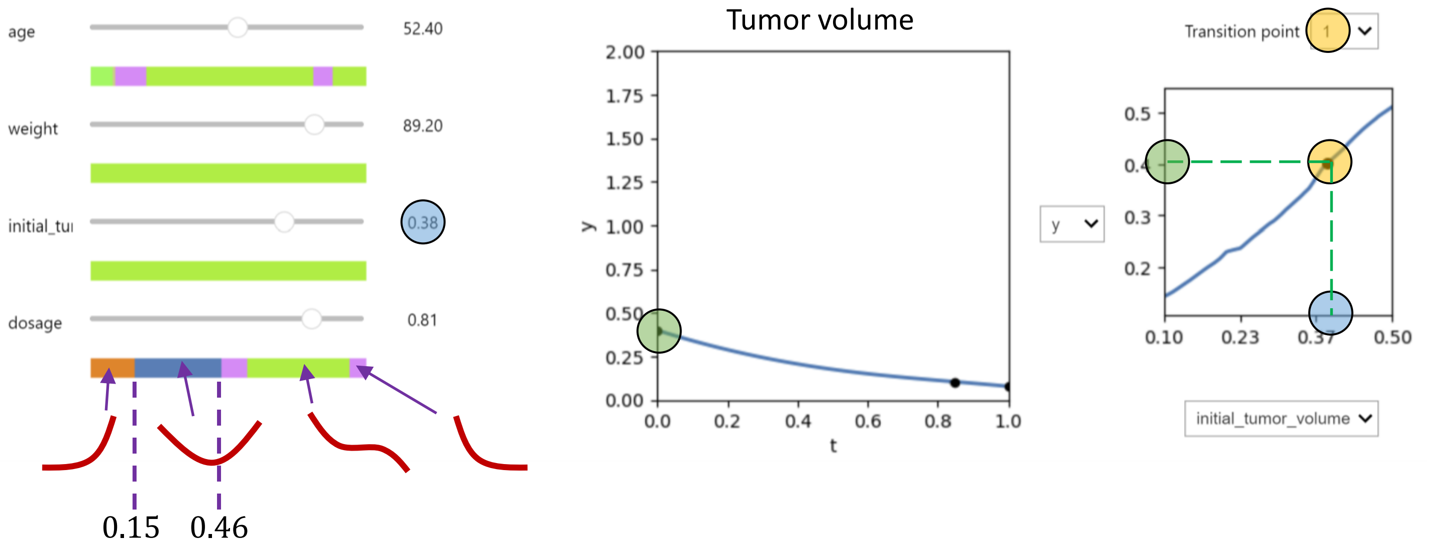


Visualization



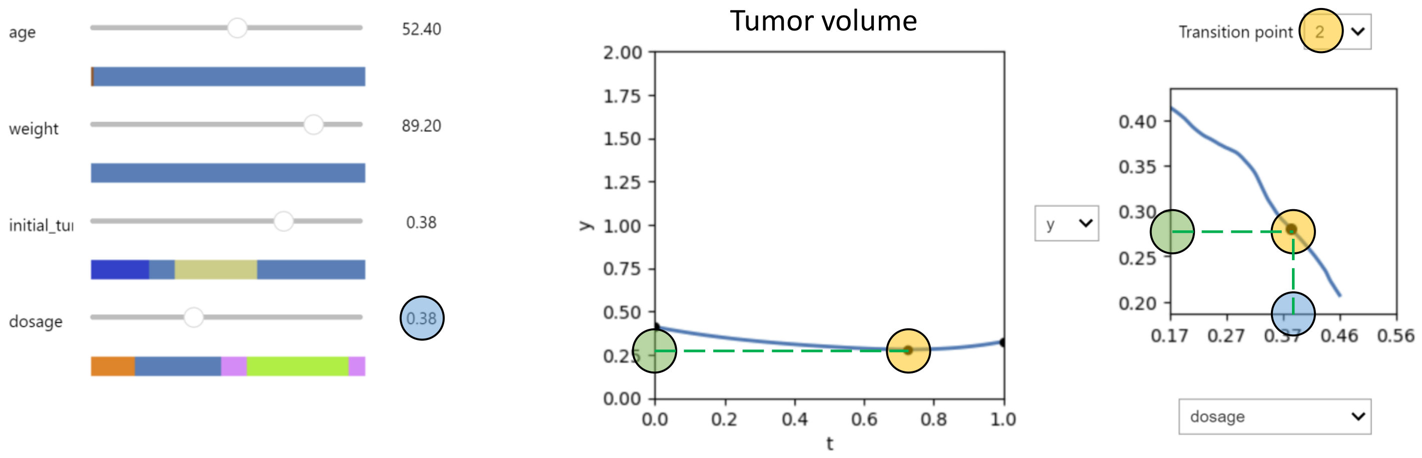
Tumor Example

Would the predicted tumor volume keep decreasing if we adjusted the treatment?



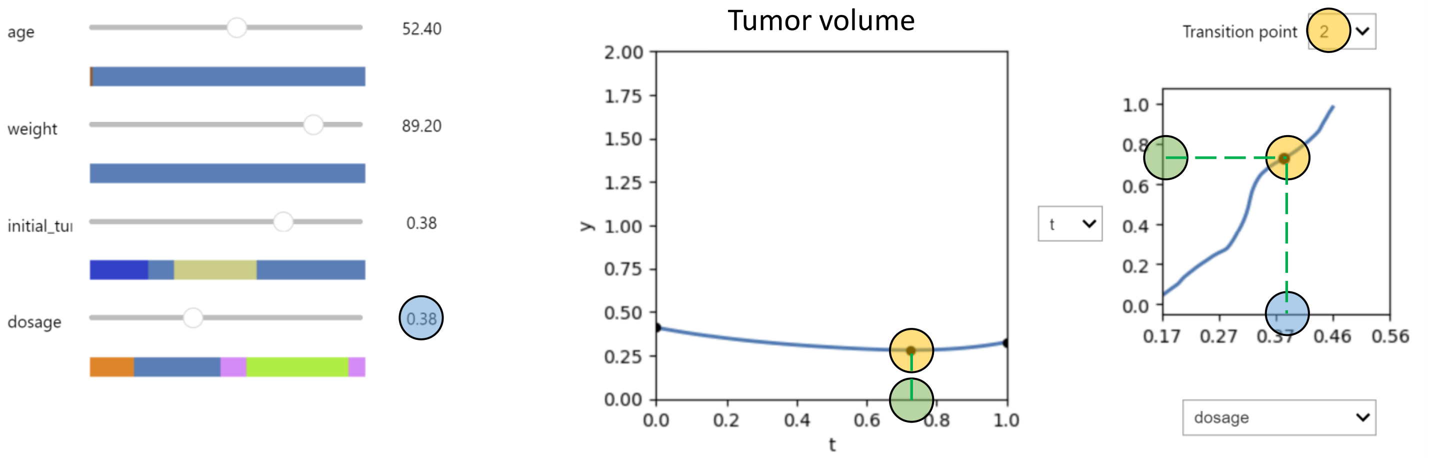
Tumor Example

What feature changes would lower the minimum tumor volume?

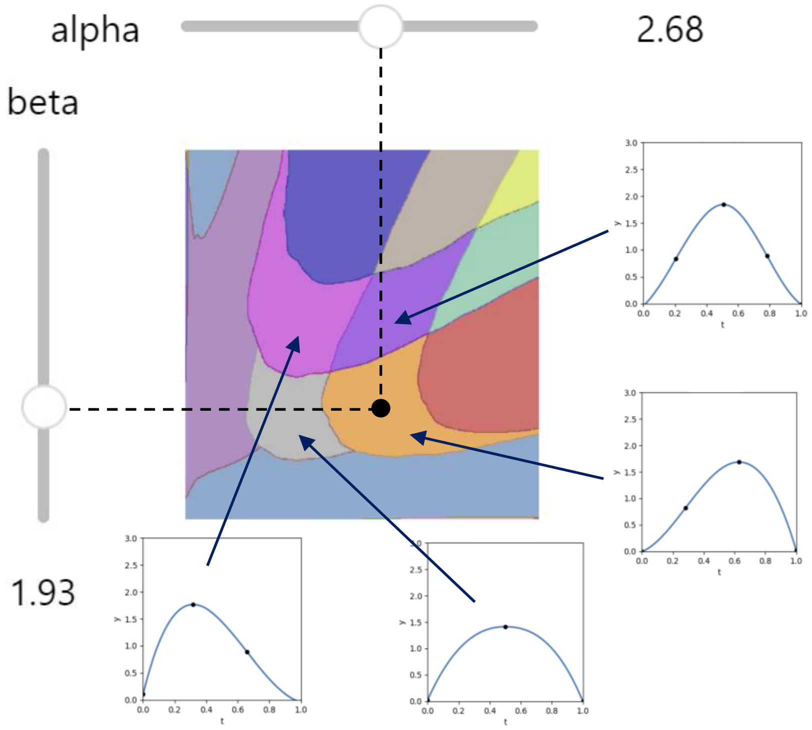


Tumor Example

How feature changes would impact the time this minimum is achieved?



Visualizing interactions



TIMEVIEW

To realize bi-level transparency through dynamical motifs, we need to

1. understand the relation between the feature vectors $\boldsymbol{x}$ and the *compositions* of the predicted trajectories
2. understand the relation between the feature vectors $\boldsymbol{x}$ and the *transition points* of a given composition.

Representing Time Series Using Cubic Splines

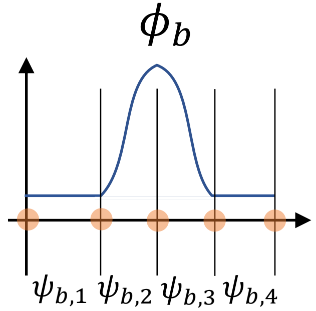
We need to choose a set of predicted trajectories such that:

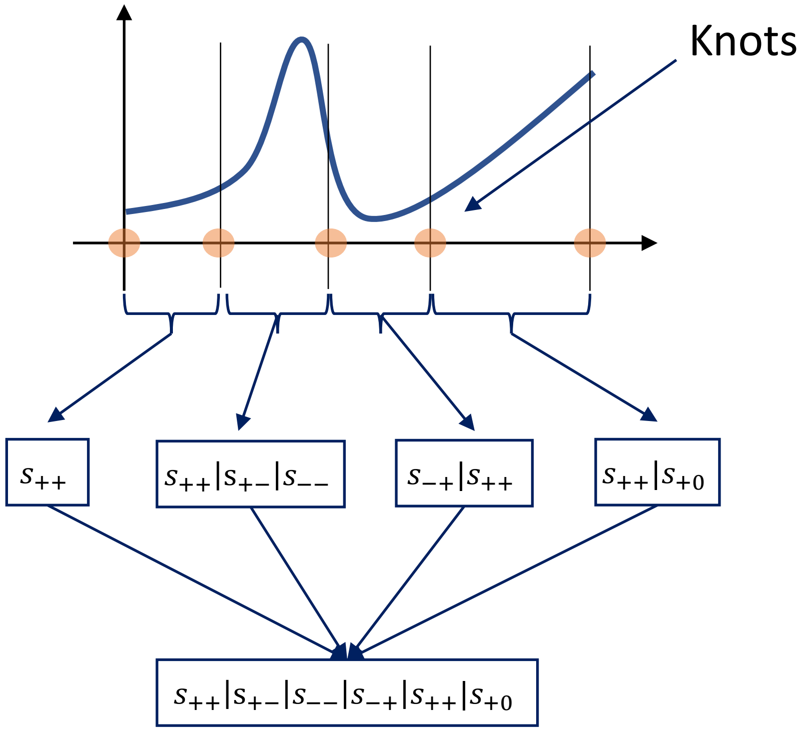
1. To every predicted trajectory we can uniquely assign a composition constructed from dynamical motifs
2. For every such trajectory we can calculate its composition efficiently

Natural choice: cubic splines!

Can be described using B-Spline basis functions.

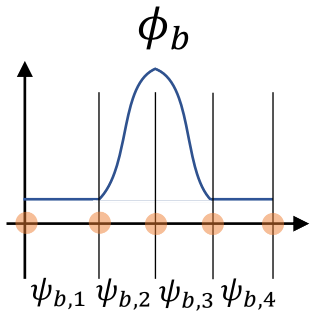
$$\sum_{b=1}^{B} c_b \psi_b$$



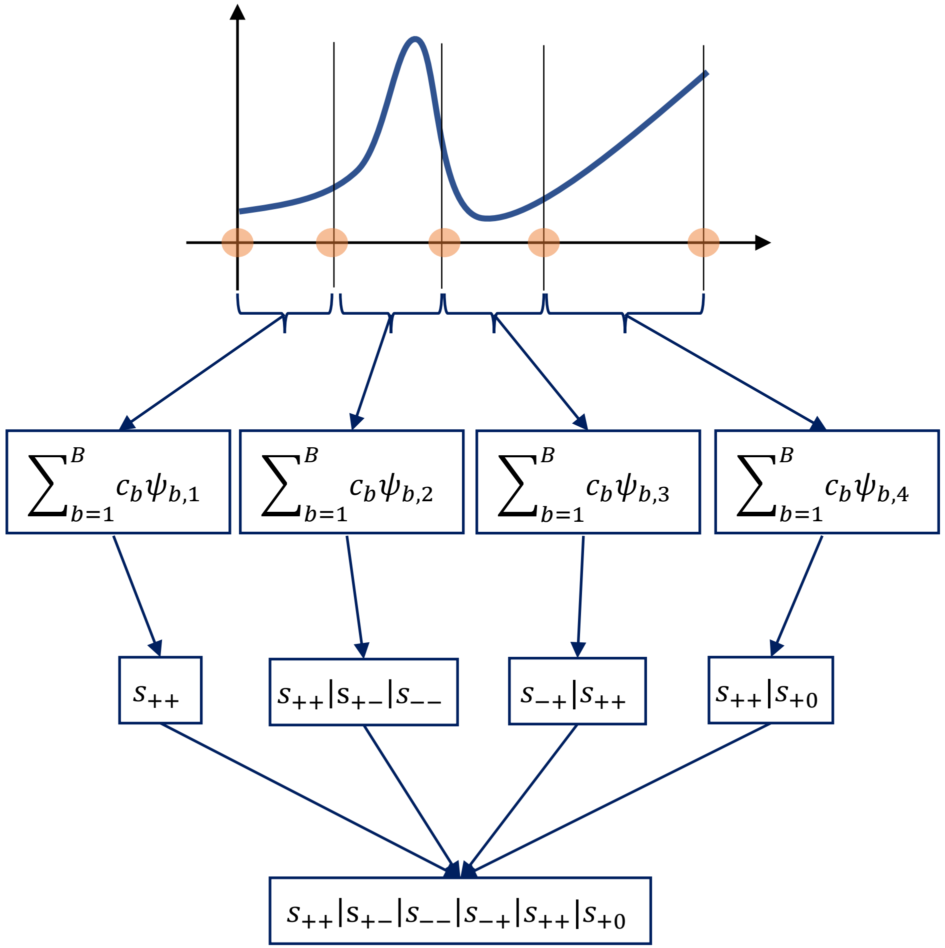


Representing Time Series Using Cubic Splines

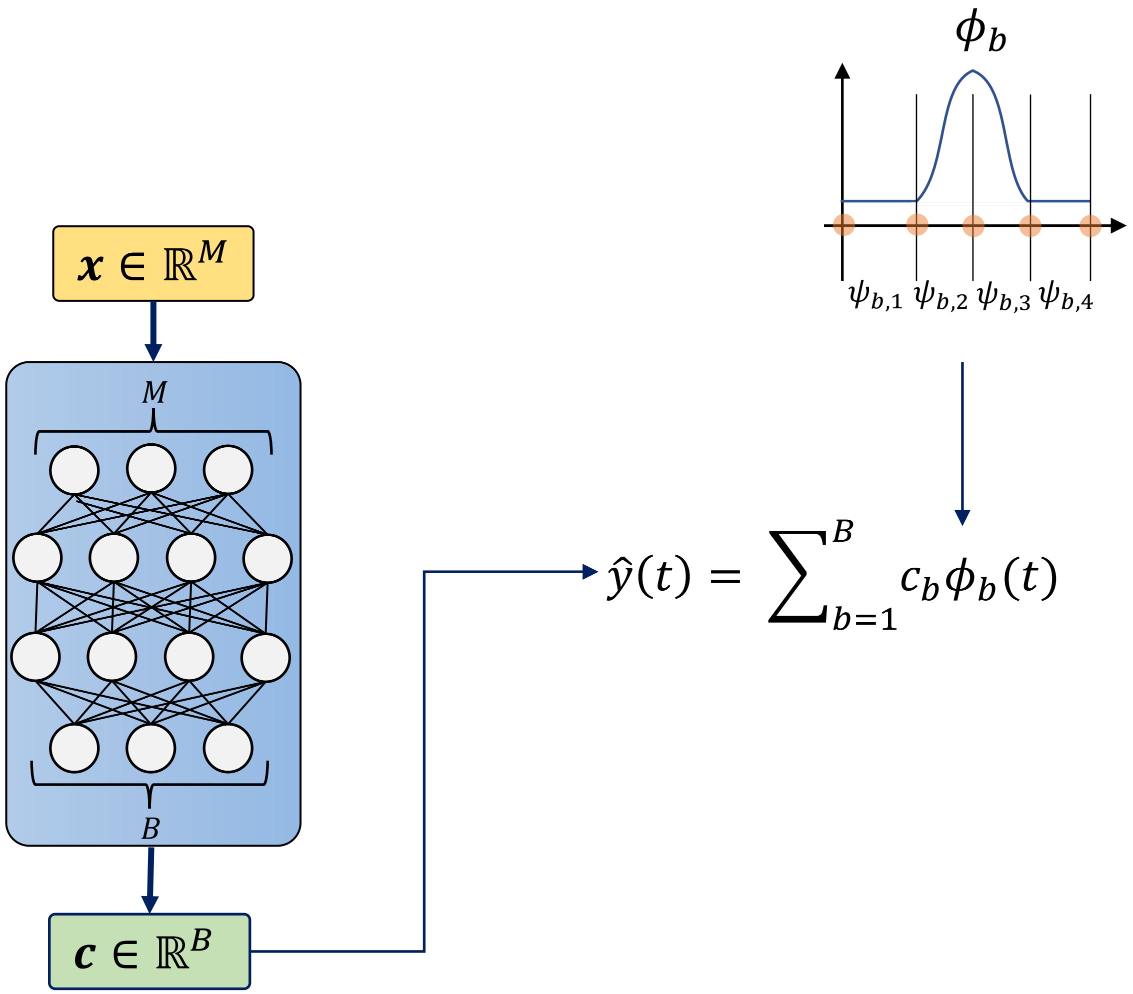
$$\sum_{b=1}^{B} c_b \psi_b$$



Calculate the cubic for each of those before training



Architecture

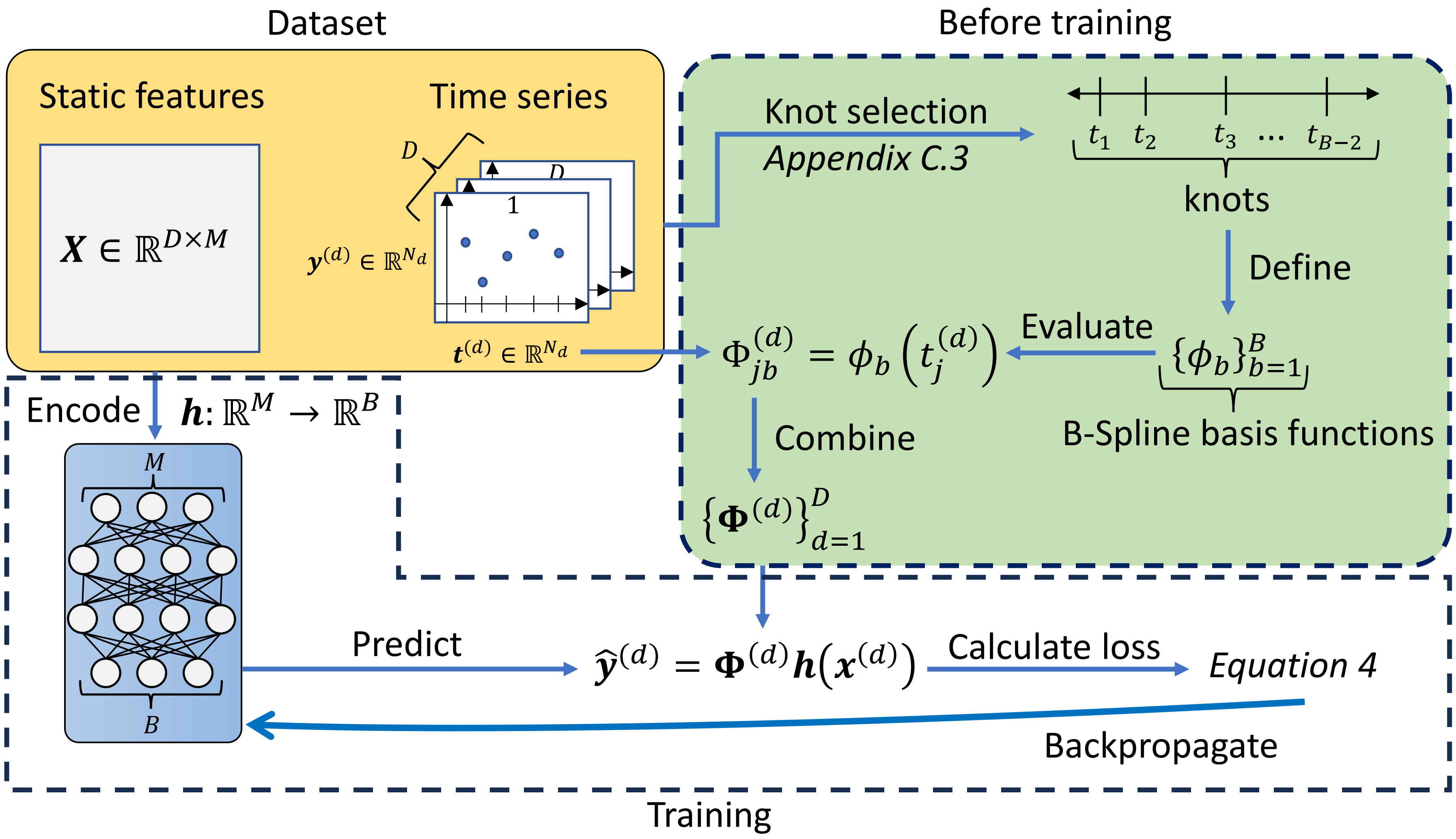


van_der_Schaar
\ LAB

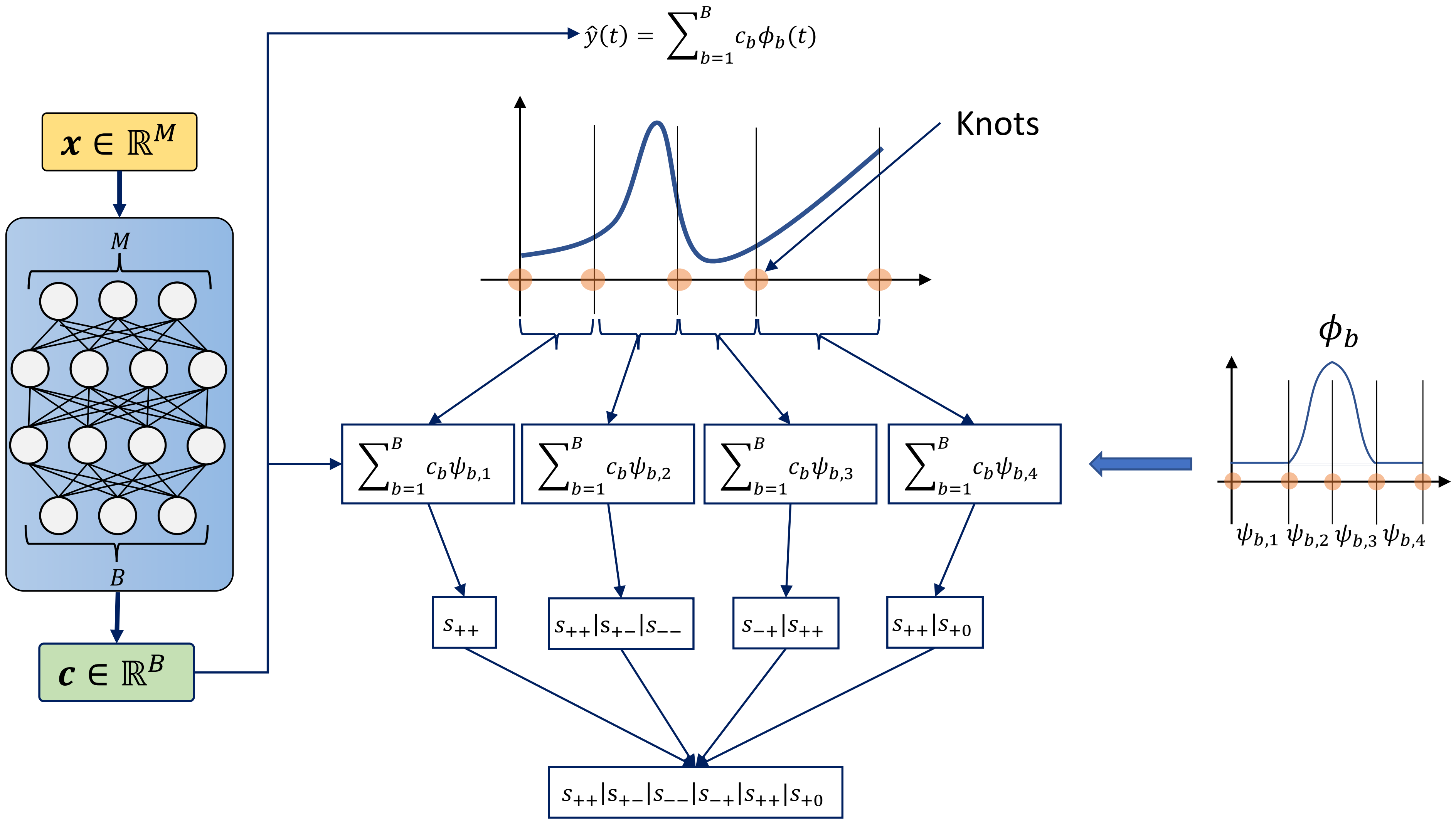
vanderschaar-lab.com

UNIVERSITY OF CAMBRIDGE

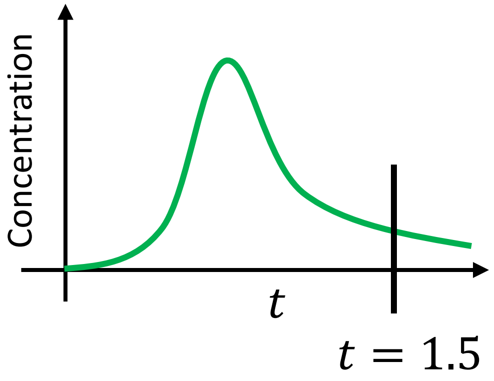
Model Training

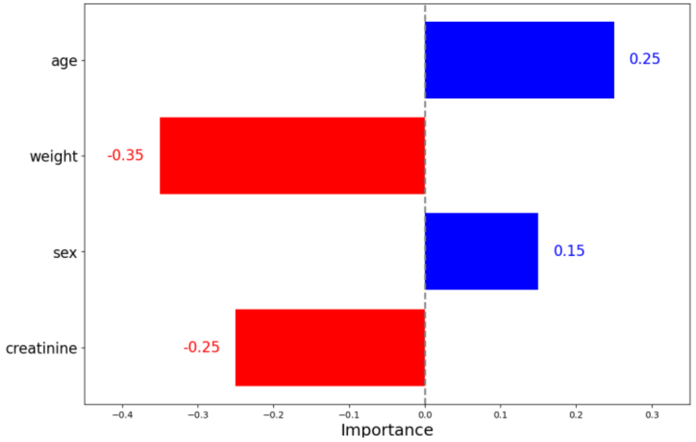


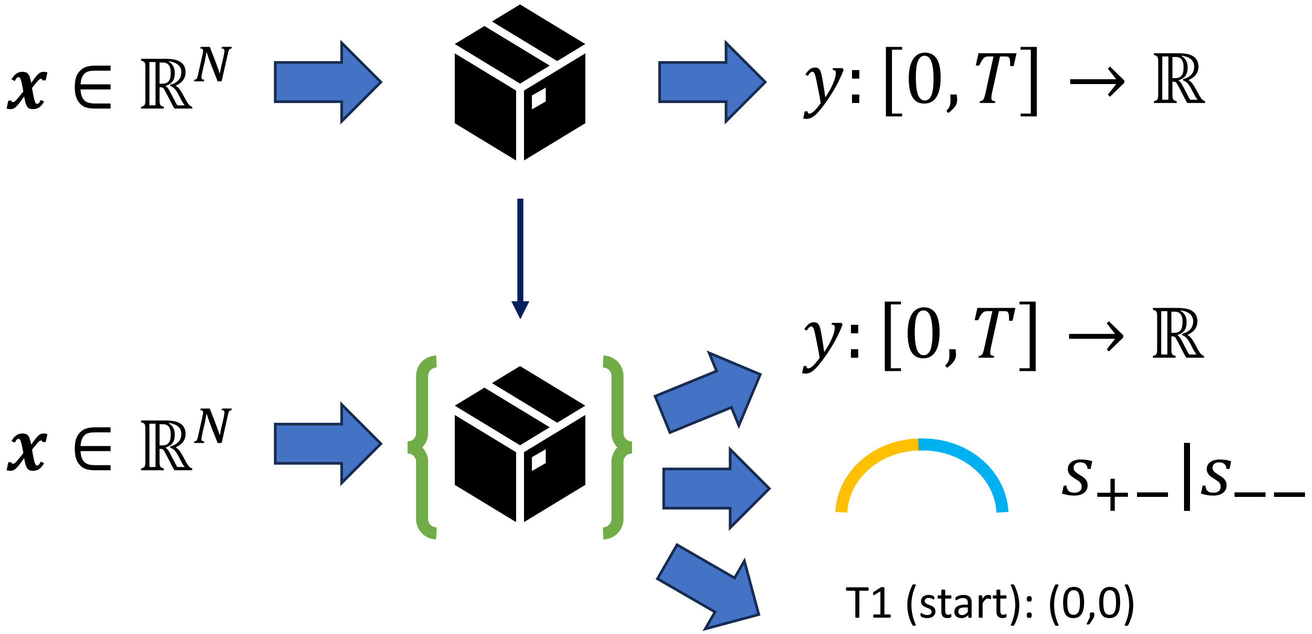
Composition Extraction



Meaningful Explanations







T1 (start): (0,0)
T1 (inflection): (0.5,0.8)
T2 (max): (1.4,1.5)
T3 (inflection): (2.7,0.6)
T4 (end): (3.5,0.2)

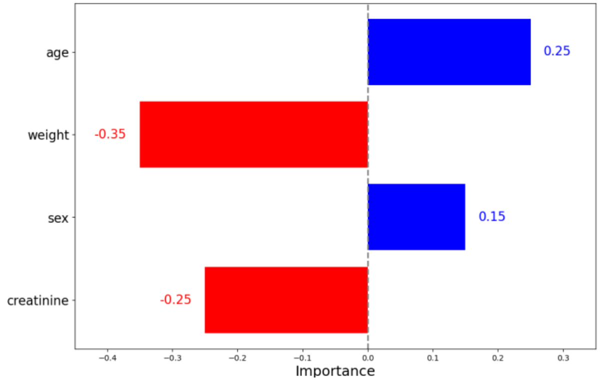
Meaningful Explanations

$x \in \mathbb{R}^N$ → [black box] → $y: [0,T] \to \mathbb{R}$

$x \in \mathbb{R}^N$ → {[black box]} → $y: [0,T] \to \mathbb{R}$

$s_{+-} | s_{--}$

T1 (start): (0,0)
T1 (inflection): (0.5,0.8)
T2 (max): (1.4,1.5)
T3 (inflection): (2.7,0.6)
T4 (end): (3.5,0.2)



$$z = 1.5x_{age} - 3.4x_{weight} + 0.4x_{sex}$$

Future directions

New optimization algorithms for SHAREs

Univariate functions and plotting as first-class citizens in symbolic regression

Personalized ODEs

Meaningful explanations for time series forecasting models