Reactmine-2: a Statistical Beam Search Algorithm for Learning Biochemical Reactions from Time Series Data

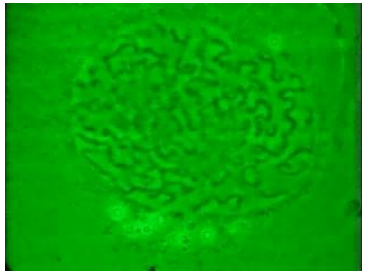
Constant Le Bezvoët, François Fages, Julien Martinelli

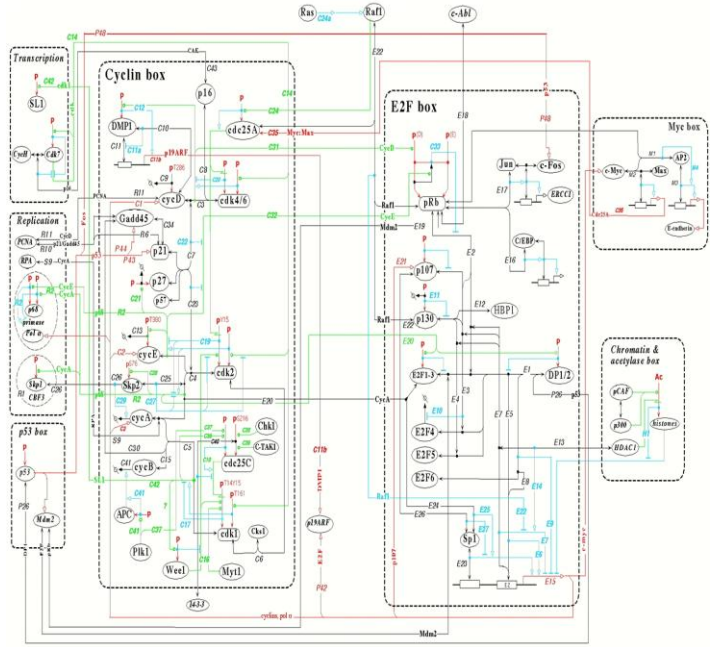
Lifeware group http://lifeware.inria.fr/
on Computational Systems Biology and Optimization

INRIA Saclay, Palaiseau, France

A Computer Scientist View at Cell Biology: Cells Compute

- Cells process signals
- Regulate their metabolism
- Take decisions such as
  - Replication
  - Migration
  - Differentiation
  - Apoptosis (suicide)
- Control those processes





Chemical Reaction Networks (CRN)

➢ What are the programs ?

Continuous time, continuous protein concentrations: analog computation

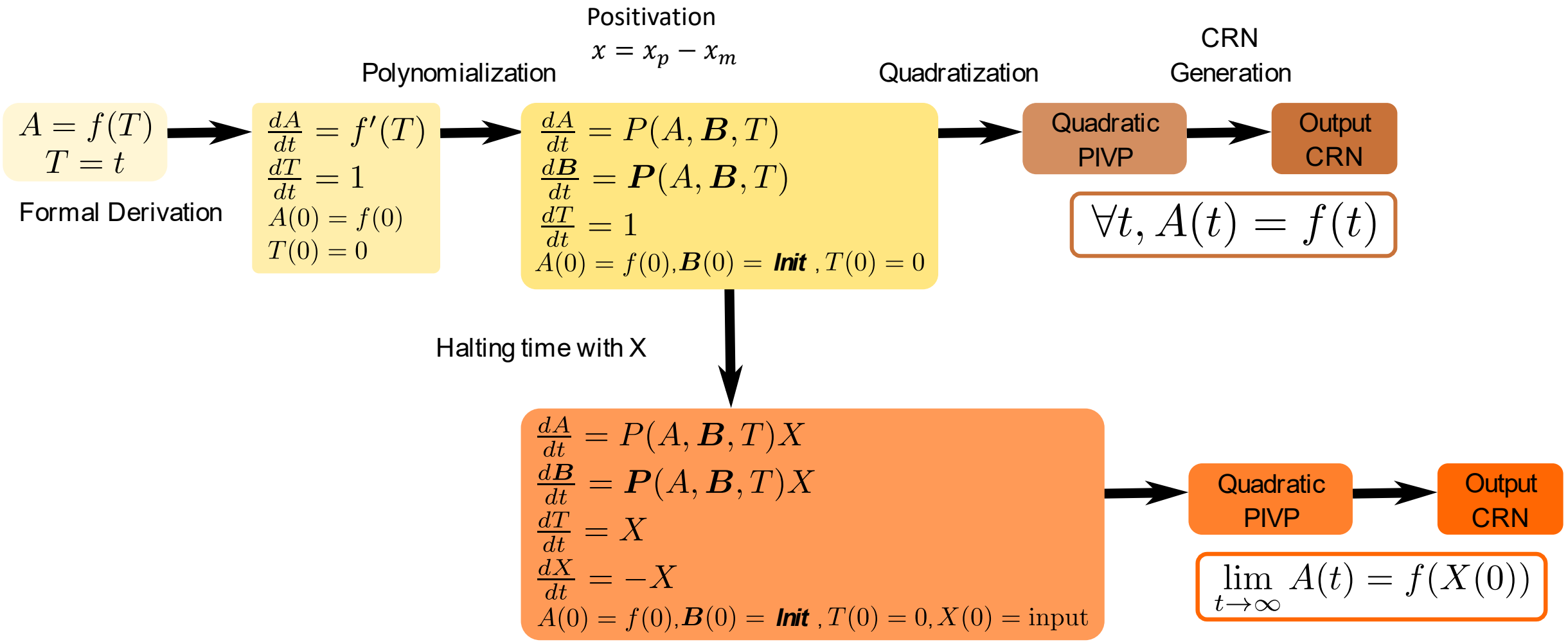
➢ Computational Systems Biology: development of CRN models of cell processes

- ➢ Repositories of thousands of hand-made CRN models, e.g. biomodels.net
- ➢ Model building tools, e.g. BIOCHAM, using graph-theoretic, symbolic computation and constraint solving methods
- ➢ Need for Machine Learning methods to infer CRN reactions from experimental data

➢ Computational Synthetic Biology: design CRNs to implement new functions

Turing Completeness of CRNs and BIOCHAM Compilation Pipeline

[Fages Le Guludec Bournez Pouly CMSB 2017] [Fages Hemery Soliman ACM CCA 2025]



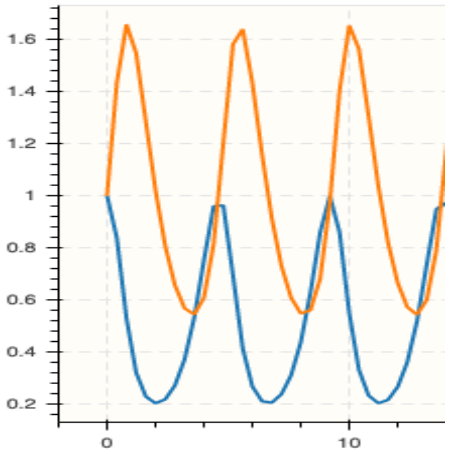
CRN Syntax and *Hierarchy of* Semantics

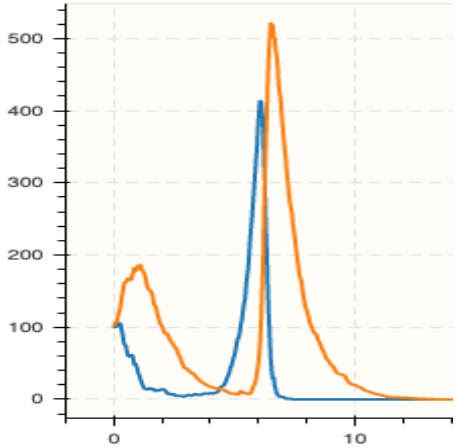
CRN: reaction rules with kinetics

```
k1*A*B for A + B => 2*B
k2*A  for A => 2*A
k3*B for B => _
```

- ODE semantics:

$$\frac{dB}{dt} = k1 * A * B - k3 * B$$
$$\frac{dA}{dt} = k2 * A - k1 * A * B$$





- Stochastic semantics (continuous time Markov chain):
- Discrete (Petri net)
- Boolean semantics:

```
list_stable_states.
```

```
[A-0,B-0]
[A-1,B-0]
```

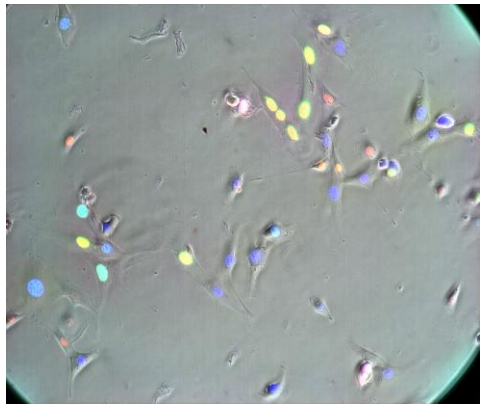
```
generate_ctl_not.
```

```
reachable(stable(A))
reachable(stable(notA))
reachable(stable(notB))
reachable(steady(B))
```

Learning Protocol:
Infer Reactions with Kinetics from Experimental/Simulation Traces

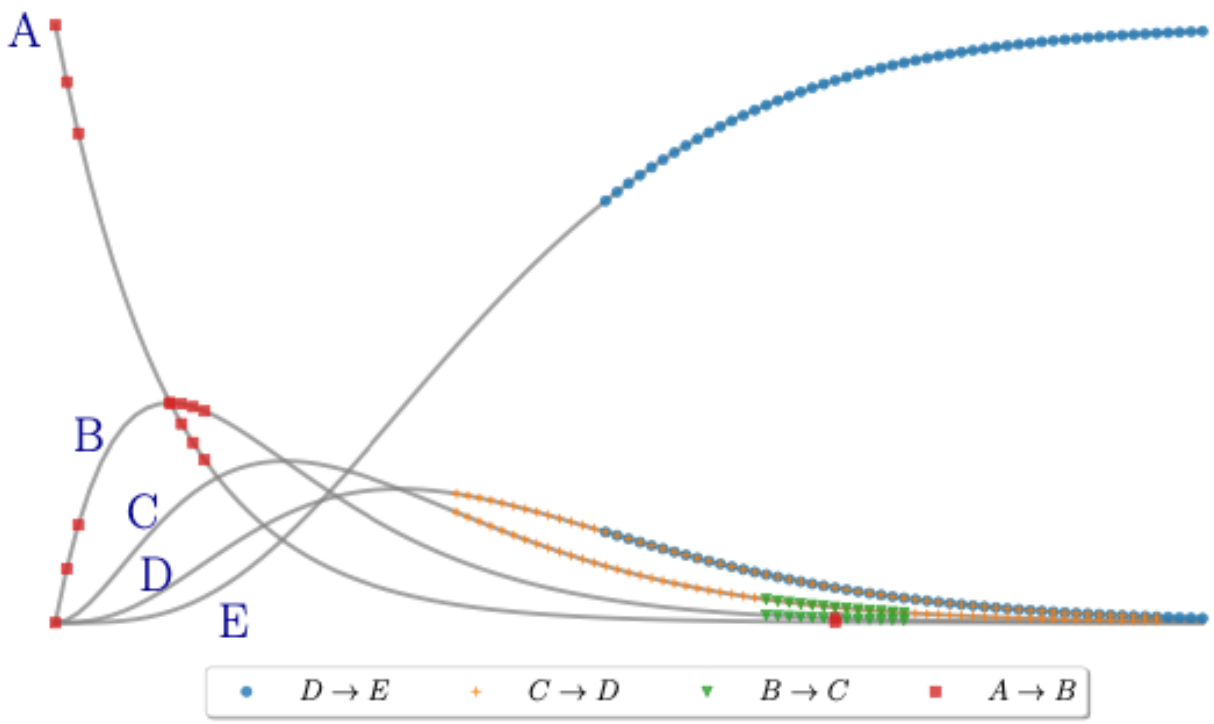
Input: prior knowledge and finite traces,
i.e. set of observed state transitions (velocity matrix)

- time evolution of molecular species concentrations
- coming from biological experiments



Fluorescence videomicroscopy
C. Feillet F. Delaunay 2013

- or from simulations of a hidden CRN



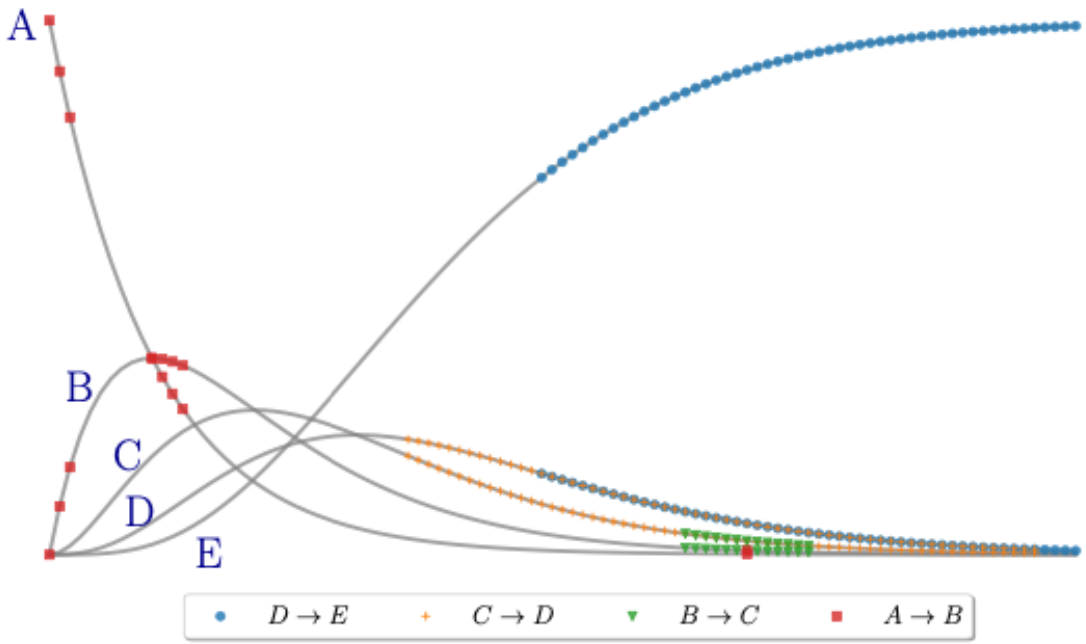
Output: CRN, i.e. finite set of kinetic reactions

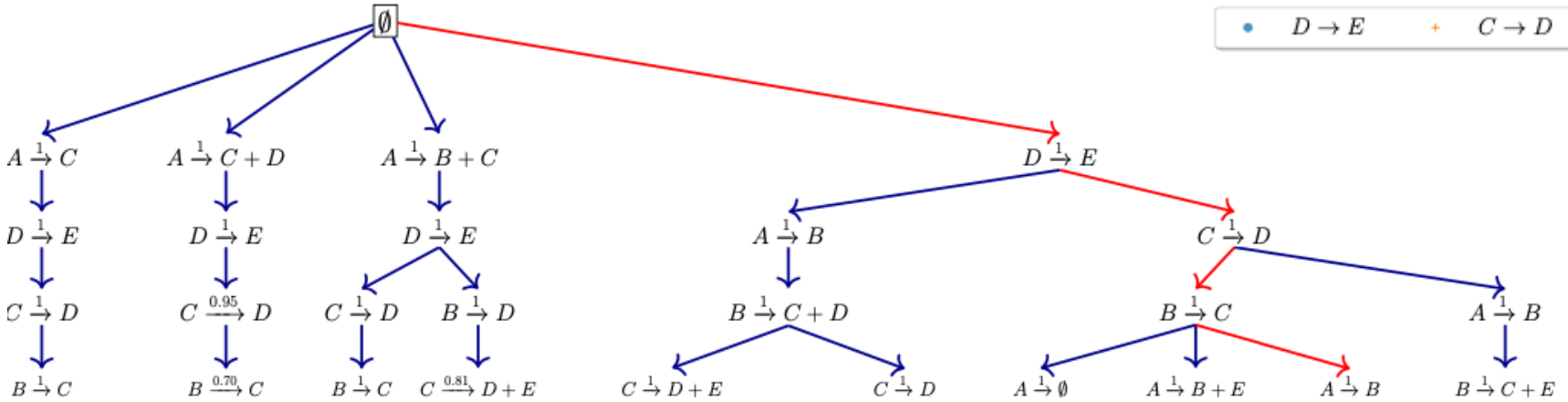
Reactmine Statistical Tree Search Algorithm

Sequential inference of reactions with kinetics;

1. on some transition support where the preponderant changes can be paired in one preponderant reaction;
2. ranked by kinetics of lowest variance;
3. trace update by subtraction of the inferred reaction;

Final global kinetic parameter re-optimization





Inference of both Reaction's Structure and Kinetics

At each node of the search tree with **bounded depth $\leq \gamma$** (i.e. maximum number of reactions to infer)

1) For each observed transition consider all reaction skeletons on concentration changes with highest velocities

$$i_l^{\max} = \underset{i}{\operatorname{argmax}} \, |\tilde{v}_{l,i}| \qquad \tilde{v}_{l,i} = \frac{\hat{v}_{l,i}}{\max\limits_{1 \leq k \leq n} |\hat{v}_{k,i}|}$$

$$R_\delta(t_l) = \left\{ i \in \{1, \ldots, m\},\ \hat{v}_{l,i} < 0,\ \left| \frac{\hat{v}_{l,i^{\max}}}{\hat{v}_{l,i}} \right| \leqslant \delta \right\}$$

with **velocity ratio $\leq \delta$**

$$P_\delta(t_l) = \left\{ i \in \{1, \ldots, m\},\ \hat{v}_{l,i} > 0,\ \left| \frac{\hat{v}_{l,i^{\max}}}{\hat{v}_{l,i}} \right| \leqslant \delta \right\}$$

2) Rank reaction candidates (k, R, P) according to kinetic variance $\hat{\sigma_j} = \frac{1}{|T(r)|} \sum\limits_{l \in T(r)} \left(\frac{|v_{l,j}|}{\prod_{i \in R} y_{l,u}} - \hat{k_j} \right)^2$

on their supporting transitions $T(r) = \{ l \in \{1, \ldots, n\} \mid \exists \delta \in [1, \delta_{\max}], r^\delta(t_l) = (R, P) \}$

3) Create successor nodes for least kinetic variance candidates with **beam width $\leq \beta$**
and subtract reaction effect on the trace

Finally, re-optimize the inferred kinetic parameters on the whole trace $\mathbf{k} = \underset{\mathbf{k} \in \mathbb{R}_+^p}{\operatorname{argmin}} \, \|\mathbf{V} - \mathbf{F}(\mathbf{Y}, \mathbf{k})\mathbf{S}\|_F^2$

SINDy Non-Linear Sparse Regression Algorithm

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

- Reconstructs an ODE system from transition traces, as weighted sums of library functions

$$\mathbf{V} = \boldsymbol{\Theta}(\mathbf{Y})\boldsymbol{\Xi} \tag{9}$$

$\boldsymbol{\Theta}(\mathbf{Y}) \in \mathbb{R}^{n \times p}$ is a library of p functions constructed from the input variables $\mathbf{Y}$ including, for instance, first to m-order polynomial interactions, e.g. $\mathbf{Y}_{\bullet,j} \odot \mathbf{Y}_{\bullet,j'}$ the sin and cos functions, e.g. $\sin(\mathbf{Y}_{\bullet,j})$, or even more sophisticated user-defined functions. The dynamics of each variable is then captured by a weighted combination of library members, the weights being encompassed in $\boldsymbol{\Xi}$. Because it is thought that the expression of the dynamics should be sparse within the library $\boldsymbol{\Theta}(\mathbf{Y})$, SINDy proposes to obtain $\boldsymbol{\Xi}$ using sparse regression.

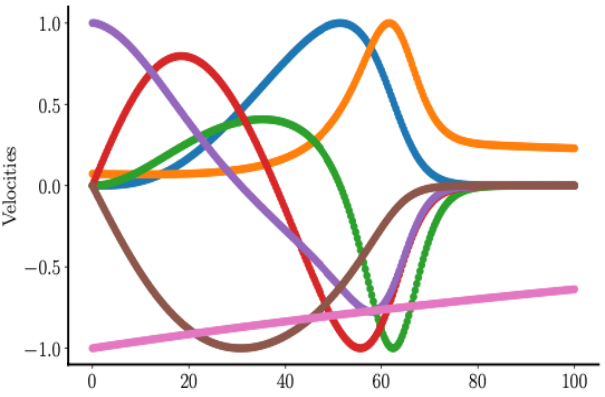
$$\boldsymbol{\Xi} = \underset{\boldsymbol{\Xi} \in \mathbb{R}^{p \times m}}{\operatorname{argmin}} \|\mathbf{V} - \boldsymbol{\Theta}(\mathbf{Y})\boldsymbol{\Xi}\|_F^2 + \lambda \|\boldsymbol{\Xi}\|_1 \tag{10}$$

- Ensures sparsity by $||\quad||_1$ regularization (Lasso regression)
- One single hyperparameter λ.

Results compared to SINDy on Simple CRN with One Single Trace

Name	Hidden CRN	CRN inferred by Reactmine	ODE system inferred by SINDy
Chain	$\mathbf{A} \xrightarrow{1} B$ $B \xrightarrow{1} C$ $C \xrightarrow{1} D$ $D \xrightarrow{1} E$	$D \xrightarrow[1.00041]{0.99869} E$ $C \xrightarrow[1.00071]{0.99836} D$ $B \xrightarrow[0.99753]{1.00102} C$ $A \xrightarrow[1.00107]{1.00069} B$	$\begin{cases} \dot{A} = -1.00A \\ \dot{B} = 1.00A - 1.00B \\ \dot{C} = 1.03B - 1.03C + 0.01D - 0.06AB \\ \dot{D} = 0.33B - 0.64DE \\ \dot{E} = 1.00D \end{cases}$
Loop	$\mathbf{A} \xrightarrow{1} B$ $B \xrightarrow{1} C$ $C \xrightarrow{1} D$ $D \xrightarrow{1} E$ $E \xrightarrow{1} \mathbf{A}$	$A \xrightarrow[1.00005]{0.99958} B$ $B \xrightarrow[0.99243]{0.99748} C$ $C \xrightarrow[0.99156]{0.99772} D$ $D \xrightarrow[0.92045]{0.99771} E$ $E \xrightarrow[1.00007]{0.99758} A$	$\begin{cases} \dot{A} = -1.00A + 1.03E - 0.006D - 0.07AE - 0.06DE \\ \dot{B} = 1.00A - 1.00B + 0.004C + 0.001AB - 0.211AC - 0.092BC \\ \dot{C} = 1.14B - 1.18C - 0.002D - 0.17AB + 0.39CD \\ \dot{D} = 0.35B - 0.35E \\ \dot{E} = 0.39C + 0.457E - 4.21AE \end{cases}$
Reactant Parallel	$\mathbf{A} + C \xrightarrow{3} B + C$ $\mathbf{D} \xrightarrow{2} C$ $\mathbf{E} \xrightarrow{1} C$	$D \xrightarrow[1.97698]{2.00109} C$ $E \xrightarrow[0.00526]{1.00101} C$ $A + C \xrightarrow[2.95790]{2.94380} B + C$	$\begin{cases} \dot{A} = -1.12A - 510000276790.32C - 0.87AB + 1.03AD \\ \quad + 510000276788.57AC + 510000276790.32BC - 0.04DC \\ \dot{B} = 1.12A + 510000276790.24C + 0.87AB - 1.04AD \\ \quad - 510000276788.49AC - 510000276790.24BC + 0.04DC \\ \dot{C} = -2.20D + 0.02E - 0.19AD + 0.40DE \\ \dot{D} = 0.07A + 1.96D + 1.02E - 0.06DE \\ \dot{E} = -1.00E \end{cases}$
Product Parallel	$\mathbf{A} + \mathbf{C} \xrightarrow{1} B + \mathbf{C}$ $\mathbf{C} \xrightarrow{2} D$ $\mathbf{C} \xrightarrow{3} E$	$C \xrightarrow[3.12657]{2.93602} E$ $C \xrightarrow[2.08438]{1.95735} D$ $A + C \xrightarrow[1.04002]{0.96372} B + C$	$\begin{cases} \dot{A} = -0.93C \\ \dot{B} = 0.93C \\ \dot{C} = 68364.34 - 84189.93A + 32887.73C + 14190.64D \\ \quad + 21285.97E - 17066.849AC - 40273.85AE + 990.87CD \\ \quad + 990.83CD + 1486.24CE + 7919.89DE \\ \dot{D} = 3.73C - 1.84AC \\ \dot{E} = 65012.55A - 60751.24C - 77885.60E - 4258.48AC \\ \quad - 10214.65CE - 27068.38DE \end{cases}$

Results on MAPK Signalling CRN with One Single Trace



Reactmine

$$A \xrightarrow[0.0045]{0.00449} Ap$$

$$Bp + Ap \xrightarrow[499.96702]{499.96702} Bpp + Ap$$

$$B + Ap \xrightarrow[500.01476]{500.01476} Ap$$

$$ApB \xrightarrow[150.03923]{150.04220} Ap + Bp$$

$$Ap + B \xrightarrow[501.19119]{501.19113} ApB$$

$$ApB \xrightarrow[150.36767]{150.37064} Ap + B$$

$$Ap + Bp \xrightarrow[517.78581]{517.78303} ApBp$$

$$ApBp \xrightarrow[155.32578]{155.33508} Ap + Bp$$

$$Ap + B \xrightarrow[500.18851]{500.18846} ApB + B$$

SINDy

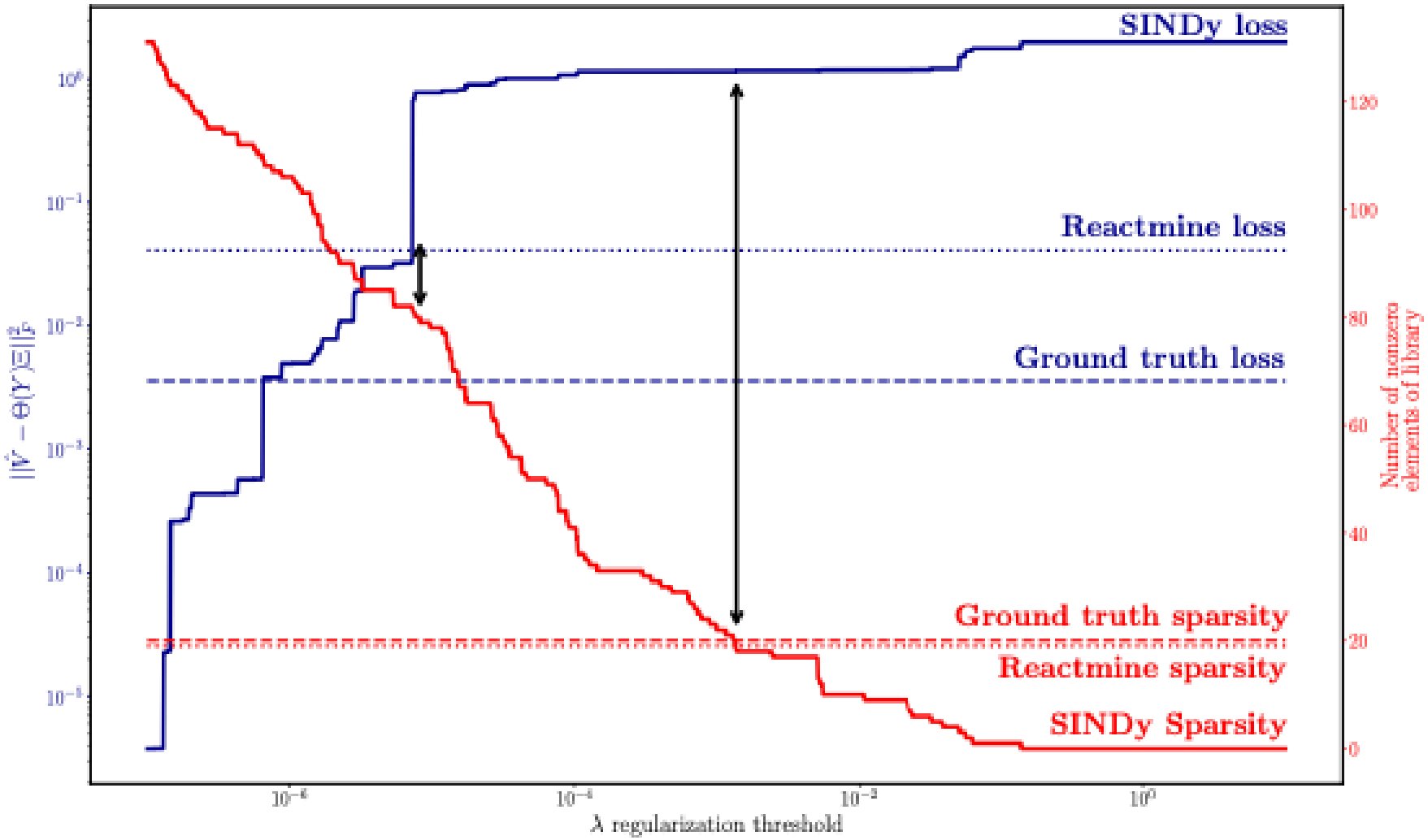
$$\begin{cases}
\dot{Bpp} = 11764.89 - 9818.81Bpp - 21809.21Ap - 64774.82ApBp - 9881.63Bp \\
\quad +109653.06ApB - 10102.57B - 23028.83A + 23087.90Bpp \times Ap \\
\quad +47383.24Bpp \times ApBp + 0.01Bpp \times Bp - 94598.54Bpp \times ApB \\
\quad +0.05Bpp \times B + 24104.25Bpp \times A + 58119.14Ap \times Bp \\
\quad +117674.08Ap \times B + 68314.42ApBp \times Bp + 239788.36ApBp \times B \\
\quad -171491.97Bp \times ApB + 0.03Bp \times B + 45027.82Bp \times A + 118690.34B \times A \\
\dot{Ap} = 0.003Bp - 0.001Bpp \times Bp - 0.004Bpp \times B - 0.002Bp \times B \\
\dot{ApBp} = -0.002Bp + 0.001Bpp \times Bp + 0.004Bpp \times B + 0.002Bp \times B \\
\dot{Bp} = -11345.814 + 9469.53Bpp + 20253.98Ap + 62939.32ApBp \\
\quad +9525.97Bp - 105529.77ApB + 9401.39B + 22299.11A \\
\quad -21772.25Bpp \times Ap - 45730.13Bpp \times ApBp - 0.01Bpp \times Bp \\
\quad +91003.53Bpp \times ApB - 0.05Bpp \times B - 23476.51Bpp \times A \\
\quad -56249.34Ap \times Bp + 996.55Ap \times B - 64537.43ApBp \times Bp \\
\quad -113908.83ApBp \times B + 163092.38Bp \times ApB - 0.03Bp \times B \\
\quad -42276.50Bp \times A + 113704.11ApB \times B - 763.549B \times A \\
\dot{ApB} = 0 \\
\dot{B} = -668.78 + 557.83Bpp + 1724.07Ap + 2552.29ApBp \\
\quad +564.79Bp - 5063.51ApB + 551.20B + 784.94A \\
\quad -1609.99Bpp \times Ap + -1960.37Bpp \times ApBp - 0.001Bpp \times Bp \\
\quad +4399.19Bpp \times ApB - 0.01Bpp \times B - 827.38Bpp \times A \\
\quad -3441.84Ap \times Bp + 543.68Ap \times B - 4279.75ApBp \times Bp \\
\quad -8537.02ApBp \times B + 10869.45Bp \times ApB - 0.003Bp \times B \\
\quad -3146.138Bp \times A + 6610.523ApB \times B + 1384.463B \times A \\
\dot{A} = 0
\end{cases}$$

Reaction $A_p + B \rightarrow^{1000} A_pB$ inferred with 3 algebraically equivalent reactions $A_p + B \rightarrow^{500} A_p \mid A_pB \mid A_pB + B$

Reaction $A_pB_p \rightarrow^{150} A_p + B_{pp}$ replaced by $A_P + B_p \rightarrow^{500} A_p + B_{pp}$ which Is a good approximaiton:

direct production without intermediary complex leading to similar simulation traces

No sparse ODE system with a good loss function is inferred for any value of λ



Correctness of Lasso Regression [Yu Zhao 2006]

The Lasso estimates $\hat{\beta}^n(\lambda)$ are defined by $\hat{\beta}^n(\lambda) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \left\{ \|Y_n - X_n\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \right\}$

For $\lambda = 0$, $p \geq n$ and X_n of full rank we have a closed form solution $\hat{\beta} = (X_n^T X_n)^{-1} X_n^T Y_n$

Let $X_n(1), X_n(2)$ be the columns of X_n multiplied by non-zero, zero βcoefficients and $C^n = \frac{1}{n} X_n X_n^T$

The (low correlation) Strong irrepresentable Condition is $\left| C_{21}^n \left(C_{11}^n\right)^{-1} \operatorname{sign}\left(\beta_{(1)}^n\right) \right| < \mathbf{1}$ $C^n = \begin{pmatrix} C_{11}^n & C_{12}^n \\ C_{21}^n & C_{22}^n \end{pmatrix}$

Implied by low correlations between covariates in the model and not in the model

Theorem (Sufficient condition for correctness of Lasso model selection with exponential rate in n)

For fixed q and p , and under regularity conditions (1) and (2), Lasso is strongly sign consistent if the Strong Irrepresentable Condition holds. That is, when Strong Irrepresentable Condition holds ,$\forall \lambda_n$ that satisfies $\frac{\lambda_n}{n} \xrightarrow{n} 0$ and $\frac{\lambda_n}{n^{(1+c)/2}} \xrightarrow{n} +\infty$ with $0 < c < 1$, we have :

$$\mathbb{P}(\hat{\beta}^n(\lambda_n) =_s \beta) = 1 - \mathcal{O}(\exp(-n^c))$$

Where the conditions (1) and (2) are : $C^n \xrightarrow{n} C$ $\frac{1}{n} \max_{1\leq i\leq n} \left((x_i^n)^\top x_i^n\right) \xrightarrow{n} 0$

Multi-traces on Different Initial Conditions with Zeros

To make SinDy work, we concatenate m traces stemming from different initial conditions. The resulting input matrix when the library function is the products of concentrations is of the form:

$$\Theta(X_n) = \begin{bmatrix} x_1^1(0) & x_2^1(0) & \cdots & x_1^1(0)x_2^1(0) & \cdots \\ x_1^1(\Delta t) & x_2^1(\Delta t) & \cdots & \cdots & \cdots \\ \vdots & \vdots & \ddots & \cdots & \cdots \\ x_1^1((n-1)\Delta t) & x_2^1((n-1)\Delta t) & \cdots & \cdots & \cdots \\ x_1^2(0) & x_2^2(0) & \cdots & x_1^2(0)x_2^2(0) & \cdots \\ \vdots & \vdots & \ddots & \cdots & \cdots \\ x_1^m((n-1)\Delta t) & x_2^m((n-1)\Delta t) & \cdots & \cdots & \cdots \end{bmatrix}$$

where: $x_k^i(l\Delta t)$ is the concentration of species k at time $l\Delta t$ on the i-th trace.
n is the number of time points
n_s is the number of species (used later)
Δt is the time step

Type 1 CRN: Atmost One Reactant per Reaction on Multitraces with 1 Non-zero Initial Concentrations

Proposition

For a CRN of type 1, suppose we concatenate the traces as above and choose the optimal library function. Suppose also that the time step is such that $\Delta T =_{n\to\infty} o(\frac{1}{n})$. Then Zhao's irrepresentable condition holds for a big enough n where n is the number of time points.

Proof

$$C^n_{1,1}(i,j) = \frac{1}{nm}\sum_{k=1}^{m}\sum_{l=0}^{n-1} X_i^k(l\Delta T)X_j^k(l\Delta T)$$

$$= \frac{1}{m}\sum_{k=1}^{m}\frac{1}{n}\langle X_i^k, X_j^k\rangle$$

$$= \mathbb{E}_m\left[\frac{1}{n}\langle X_i^k, X_j^k\rangle\right]$$

$$\frac{1}{n}\langle X_i^k, X_j^k\rangle = \frac{1}{n}\sum_{l=0}^{n-1} X_i^k(\frac{l}{n^2}), X_j^k(\frac{l}{n^2}) \text{by Cesaro lemma}$$

$$\xrightarrow[n\to\infty]{} X_i^k(0), X_j^k(0)$$

$$C^n_{1,1}(i,j) \xrightarrow[n\to\infty]{} \mathbb{E}_m\left[X_i^k(0)X_j^k(0)\right] \text{by dominated convergence theorem}$$

$$C^n_{1,1} \xrightarrow[n\to\infty]{} I_p \qquad (C^n_{1,1})^{-1} \xrightarrow[n\to\infty]{} I_p$$

$$C^n_{2,1}(i,j) \xrightarrow[n\to\infty]{} \mathbb{E}_m\left[X_{a_i}^k(0)X_{bj}^k(0)\right] \qquad C^n_{2,1}(i,j) \xrightarrow[n\to\infty]{} 0 \qquad |C^n_{2,1}(C^n_{1,1})^{-1}\beta^1| < 1 \text{ for large enough } n$$

Type 2 CRN: At most 2 Reactants per Reaction on Multitraces with 2 Non-zero Initial Conditions

Multitraces with at most two non-zero initial conditions

$$C = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 1 & 0 & \cdots & 0 & 0 \\ 1 & 0 & 1 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \ddots & \cdots & \\ 0 & 0 & 0 & \cdots & 1 & 1 \end{bmatrix}$$

if $C^n(i,j)$ has 3 or 4 species :

$$C^n_{i,j} \xrightarrow[n\to\infty]{} \mathbb{E}_m\left[X^k_{a_i}(0)X^k_{b_i}(0)X^k_{a_j}(0)\right] = 0$$

if $C^n(i,j)$ has 2 species :

$$C^n_{i,j} \xrightarrow[n\to\infty]{} \mathbb{E}_m\left[X^k_{a_i}(0)^{\epsilon_1}X^k_{a_j}(0)^{\epsilon_2}\right] = \frac{1}{m} = \frac{2}{n_s(n_s+1)} \quad \epsilon_1, \epsilon 2 \in (1,2)$$

if $C^n(i,j)$ has 1 specie :

$$C^n_{1,1} = \begin{bmatrix} A & D^\top \\ D & \frac{1}{m} I_{\frac{n_s(n_s-1)}{2}} \end{bmatrix}$$

Non-sparse covariance matrix

Yu & Zhao's correctness condition is not satisfied

SINDy Evaluation with Multitraces with 2 Non-zero Initial Conditions

Ground Truth		Simple trace		Multiple trajectories	
Equation	Rate	Equation	Rate	Equation	Rate
$A \to B$	1	$A \to B$	1.00	$A \to B$	1.00
$B \to C$	1	$B \to C$	1	$B \to C$	1.00
$C \to D$	1	$C \to D$	0.48	$C \to D$	1.00
$D \to E$	1	$D + E \to$	1.02	$D \to E$	1.06

Ground Truth		Simple trace		Multiple trajectories	
Equation	Rate	Equation	Rate	Equation	Rate
$A \to B$	1	$A \to B$	1.00	$A \to B$	1.00
$B \to C$	1	$B \to C$	1.00	$B \to C$	1.00
$C \to D$	1	$C \to D$	1.00	$C \to D$	1.00
$D \to E$	1	$D \to E$	1.00	$D \to E$	1.00
$E \to A$	1	$E \to A$	1.00	$E \to A$	1.00

Ground Truth		Simple trace		Multiple trajectories	
Equation	Rate	Equation	Rate	Equation	Rate
$A \to B$	1	$A \to B$	1.01	$A \to B$	1.01
$D \to C$	2			$D \to C$	1.99
$E \to C$	3			$E \to C$	3.01

Ground Truth		Simple trace		Multiple trajectories	
Equation	Rate	Equation	Rate	Equation	Rate
$A \to B$	1			$A \to B$	1.01
$C \to D$	2			$C \to D$	2.04
$C \to E$	3	$C + D \to E$	0.75	$C \to E$	3.06

Ground Truth		Single trace		Multiple trajectories	
Equation	Rate	Equation	Rate	Equation	Rate
$B + D \to E$	0.3			$B + D \to E$	0.3
$2A \to B$	0.1	$2A \to B + C$	0.20	$2A \to B$	0.1
$A \to C$	0.2	$A + E \to$	0.16	$A \to C$	0.2
$C \to D$	0.13	$C + D \to A$	0.09	$C \to D$	0.13

Conclusion and Perspectives

- State-of-the-art sparse regression algorithms such as Lasso and SINDy are provably correct under strong non-correlation hypotheses between dynamic library terms
  - Generally not satisfied on data time series obtained from one single initial state (wild type trace)
  - Proven satisfied here for monomolecular CRNs on multitraces from initial conditions with only one non zero
  - SINDy's ODE discovery algorithm uses
    - 1 single hyperparameter
    - May fail on single trace data from very simple CRN
    - Needs (unrealistic) multitraces obtained from initial states with many 0, shown to entail correctness of Lasso sparsity
- Reactmine variance-based tree search algorithm
  - Reduction to 1 main hyperparameter by recourse to beam search
  - Good results obtained on single simulation traces of a hidden CRN, e.g. cell signaling and cell cycle models
  - Applied to inference of main systemic regulators of circadian clocks in accordance to model-based analysis [Matinelli Dulong Li Teboul Soliman Levi F- Ballesta *Bioinformatics* 2021]
  - Applied to videomicroscopy data with inference of main conclusions of model-based analysis (upregulation of rev-erb during G2 phase) [Traynard Feillet Soliman Delaunay F- *Biosystem* 2016]
  - Challenge of learning unobserved variables: hopes by analysis of non-supported transitions.
- Not aware of good results obtained yet with LLM or GNN on such complex dynamical systems