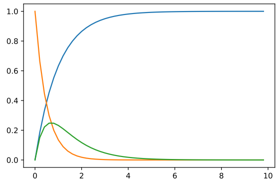
Reactmine: an algorithm for inferring biochemical reactions from time series data

Julien Martinelli, Jeremy Grignard, Sylvain Soliman, Annabelle Ballesta, François Fages

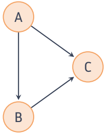
Tuesday, January 3rd 2023

Network Inference from time-series data

Input: time series describing evolution of molecular species



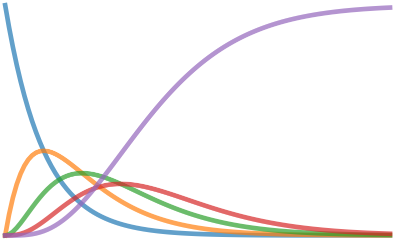
Output: interaction graph



- (un)oriented graph ⟶ Gene Regulatory Network inference
  - Gaussian Processes *(Aalto et al. 2019)*
  - Information theory *(Chan et al. 2017)*
  - Correlation networks *(Krumsiek et al. 2011)*
- oriented / weighted graph ⟶ **Chemical Reaction Network Inference**
  - Evolutionary algorithms *(Choi et al. 2018)*
  - Sparse regression *(Brunton et al. 2016)*

Chemical Reaction Network Inference

Input: single time series data $Y = \left(y_{l,i}\right)_{\substack{1\leqslant l\leqslant n \\ 1\leqslant i\leqslant m}}$



Output:
Chemical Reaction Network

Hidden CRN	Learned CRN
$A \xrightarrow{1} B$	$A \xrightarrow{0.999} B$
$B \xrightarrow{1} C$	$B \xrightarrow{1.001} C$
$C \xrightarrow{1} D$	$C \xrightarrow{1.002} D$
$D \xrightarrow{1} E$	$D \xrightarrow{0.999} E$

For this presentation: $A \xrightarrow{k} B \iff \begin{cases} \dot{A} = -kA \\ \dot{B} = kA \end{cases}$

Framework

Reaction: (R, P, f) with R (resp. P) set of reactants (resp. products) and f rate function.

Chemical Reaction Network (CRN): Finite set of reactions

Framework

Reaction: (R, P, f) with R (resp. P) set of reactants (resp. products) and f rate function.

Chemical Reaction Network (CRN): Finite set of reactions

- 0/1 Stoichiometry
- Elementary reactions: at most two reactants
- At most 1 catalyst (e.g. D in $A + D \xrightarrow{k} B + D$)

Framework

Reaction: (R, P, f) with R (resp. P) set of reactants (resp. products) and f rate function.

Chemical Reaction Network (CRN): Finite set of reactions

- 0/1 Stoichiometry
- Elementary reactions: at most two reactants
- At most 1 catalyst (e.g. D in $A + D \xrightarrow{k} B + D$)

Learning protocol:

- Learn a CRN involving only observed species
- Based on a single trace (no combinatorics of initial states and *knockouts*)

Backbone of most methods: Sparse Identification of Nonlinear Dynamics

$$\Xi = \underset{\Xi \in \mathbb{R}^{p\times m}}{\operatorname{argmin}} \|\dot{Y} - \Theta(\mathrm{Y})\Xi\|_F^2 + \lambda\|\Xi\|_1$$

$\Theta(Y) \in \mathbb{R}^{n\times p}$: library of p functions, e.g. $\begin{bmatrix} | & | & & | & | & & | \\ 1 & \mathrm{Y}_{\bullet,1} & \dots & \mathrm{Y}_{\bullet,m} & \mathrm{Y}_{\bullet,1}\mathrm{Y}_{\bullet,2} & \dots & \mathrm{Y}_{\bullet,m-1}\mathrm{Y}_{\bullet,m} \\ | & | & & | & | & & | \end{bmatrix}$

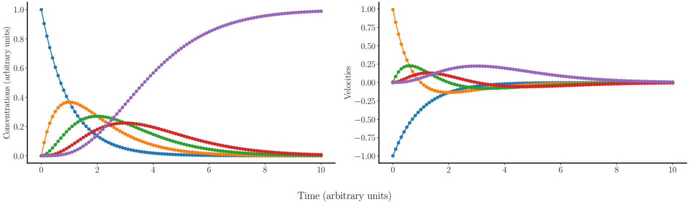
$\mathrm{Y}_{\bullet,i}$: time concentration vector for species i

$\Xi \in \mathbb{R}^{p\times m}$: weight matrix

$\lambda \in \mathbb{R}^{+}$: hyperparameter controlling level of sparsity

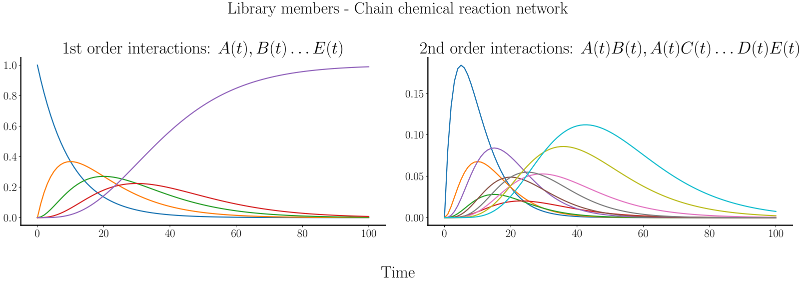
Example - Chain Chemical Reaction Network

$$\begin{cases} \dot{A} = -A \\ \dot{B} = A - B \\ \dot{C} = B - C \\ \dot{D} = C - D \\ \dot{E} = D \end{cases}$$

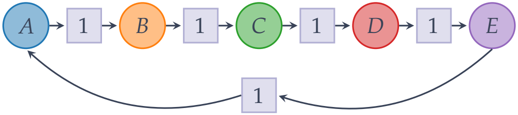


Left/right plots: simulated concentrations, derivatives

SINDy aims to predict the derivatives with the following library functions



SINDy fails at CRN inference

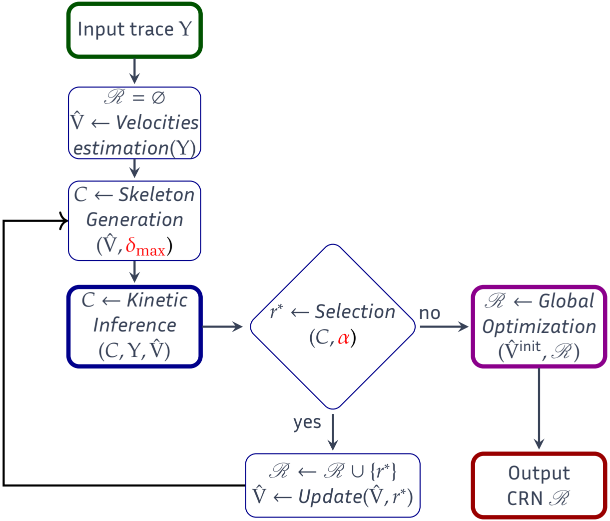


$$\begin{cases} \dot{A} = E - A \\ \dot{B} = A - B \\ \dot{C} = B - C \\ \dot{D} = C - D \\ \dot{E} = D - E \end{cases}$$

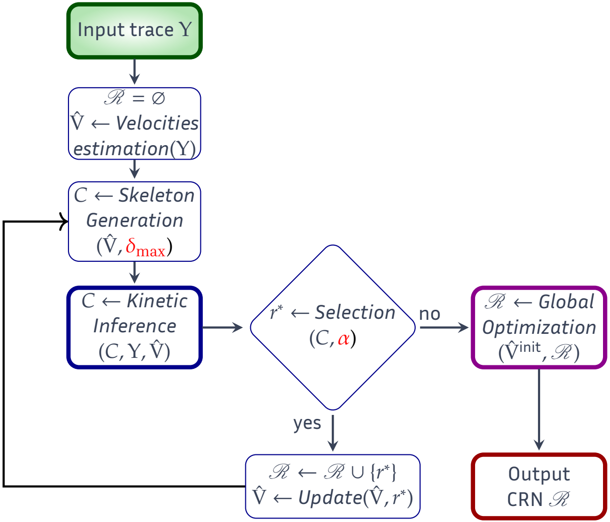
$$\begin{cases} \dot{A} = -1.00A + 1.03E - 0.006D - 0.07AE - 0.06DE \\ \dot{B} = 1.00A - 1.00B + 0.004C + 0.001AB - 0.211AC - 0.092BC \\ \dot{C} = 1.14B - 1.18C - 0.002D - 0.17AB + 0.39CD \\ \dot{D} = 0.35B - 0.35E \\ \dot{E} = 0.39C + 0.457E - 4.21AE \end{cases}$$

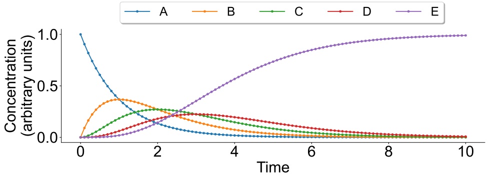
Best ODE system found across all sparsity threshold λ

Core Reactmine sequential algorithm



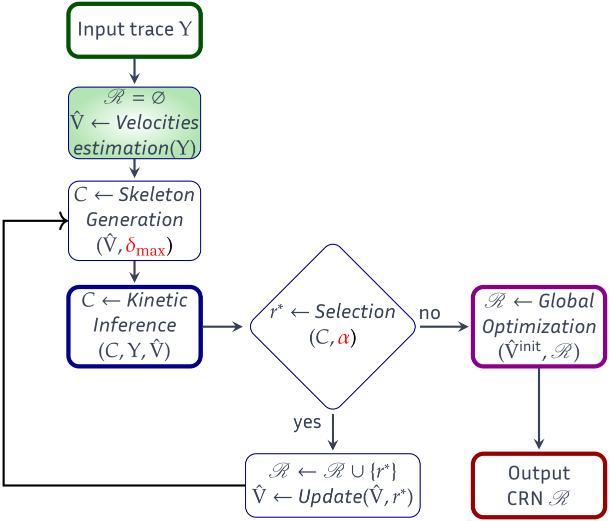
Core Reactmine sequential algorithm





$$Y = \left(y_{l,i}\right)_{\substack{1 \leqslant l \leqslant n \\ 1 \leqslant i \leqslant m}}$$

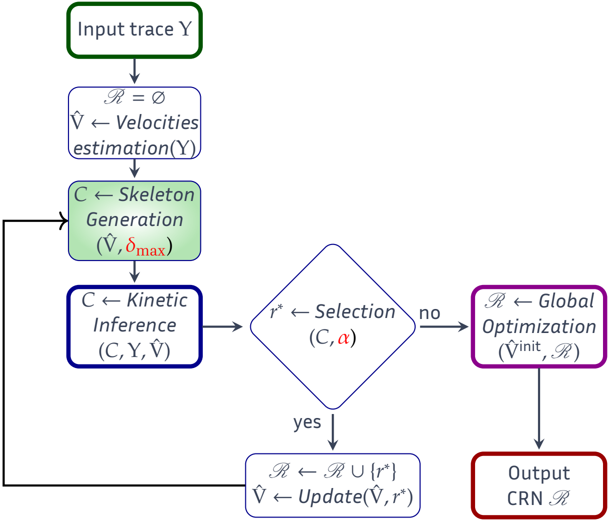
Core Reactmine sequential algorithm



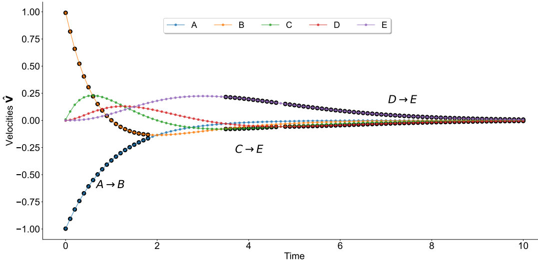
Velocities $\hat{V} = \left(\hat{v}_{l,i}\right)_{\substack{1 \leqslant l \leqslant n \\ 1 \leqslant i \leqslant m}}$

$$\hat{v}_{l,i} = \frac{y_{l+1,i} - y_{l,i}}{t_{l+1} - t_l}$$

Core Reactmine sequential algorithm

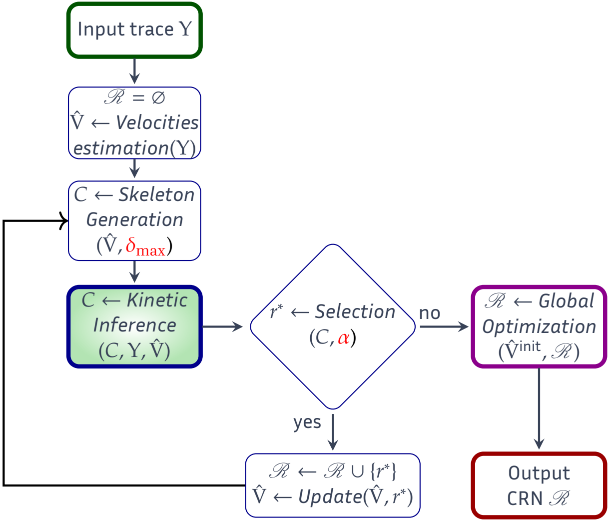


Each reaction skeleton $r = (R, P)$
is inferred based on time points t_l
where it is **preponderant: support set** $\mathscr{T}(r)$



Reactants and products belonging to a skeleton
have similar absolute velocities up to δ_{max}

Core Reactmine sequential algorithm



For each reaction skeleton $r = (R, P)$ associate kinetic rate

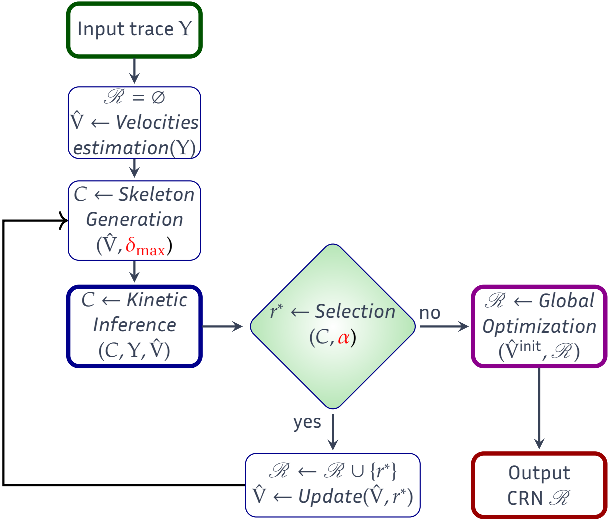
$$\forall j \in R \cup P,\ \forall l \in \{1, \dots, n\},\ |v_{l,j}| = k \prod_{u \in R} y_{l,u}$$

Estimate k reliably on the support set $\mathcal{T}(r)$

$$\hat{k} = \frac{1}{\#\mathcal{T}(r)} \sum_{l \in \mathcal{T}(r)} \frac{|\hat{v}_{l,j}|}{\prod_{u \in R} y_{l,u}}$$

Coefficient of variation (CV) $\rho = \frac{\sigma}{|\hat{k}|}$

Core Reactmine sequential algorithm

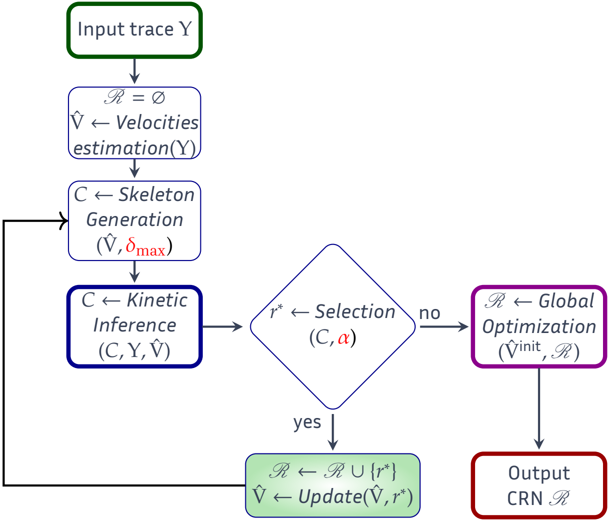


Select reaction minimizing CV

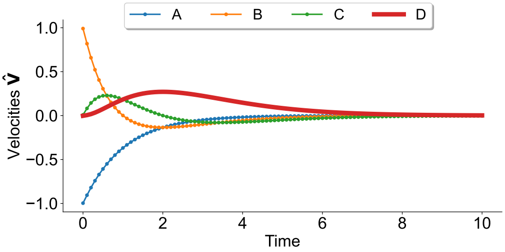
$$r^* = \underset{r}{\operatorname{argmin}}\, \rho_r$$

Accept r^* if $\rho_{r^*} < \alpha$

Core Reactmine sequential algorithm



Remove the effect of accepted reaction on the velocities



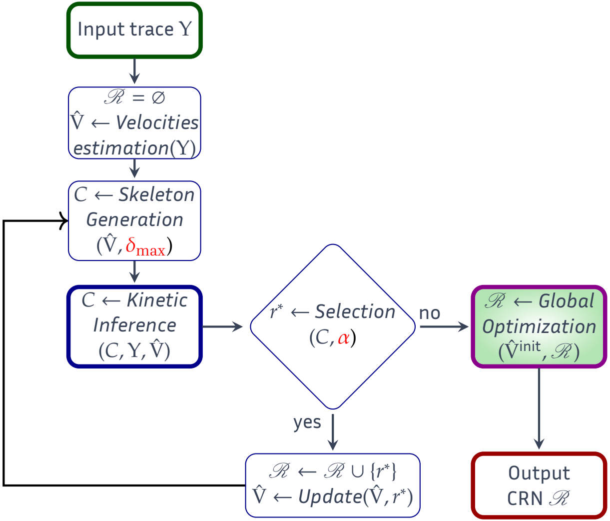
$$\hat{V} \leftarrow \hat{V} - \begin{pmatrix} f(Y_{1,\bullet}) \\ \vdots \\ f(Y_{n,\bullet}) \end{pmatrix} s^T$$

effect of the reaction

stoichiometry vector

$Y_{l,\bullet}$: species concentration vector at time t_l

Core Reactmine sequential algorithm

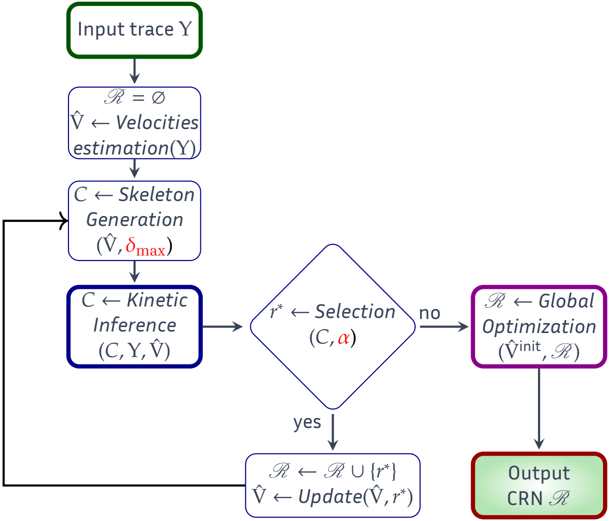


Joint optimization
of kinetic parameters
over whole trace

$$k = \underset{k \in \mathbb{R}_+^p}{\operatorname{argmin}} \|\hat{V}^{init} - F(Y, k)S\|_F^2$$

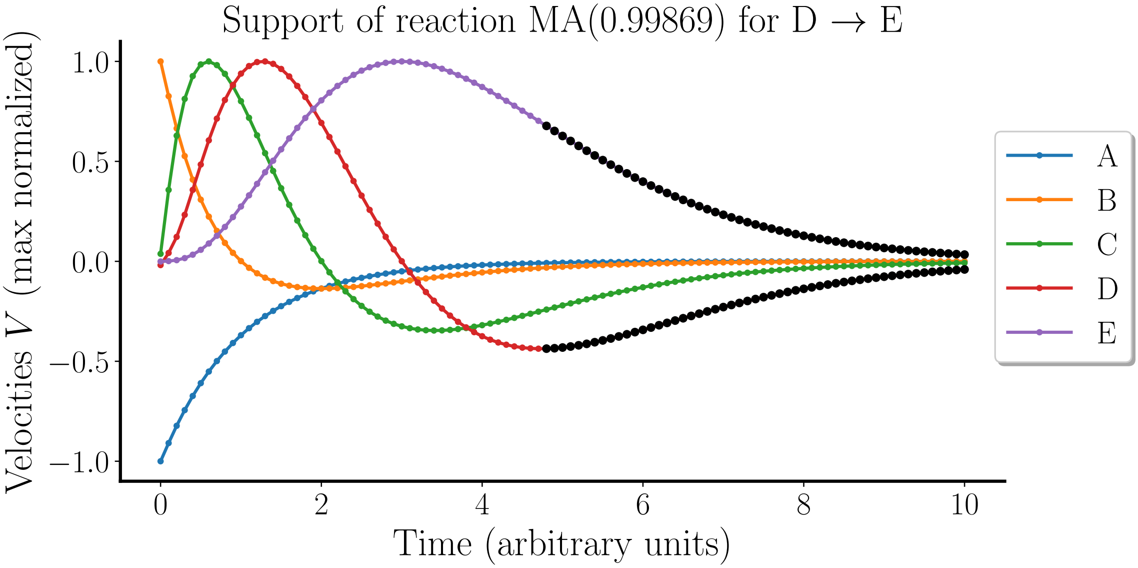
$= \Delta$ whole trace CRN transition discrepancy

Core Reactmine sequential algorithm

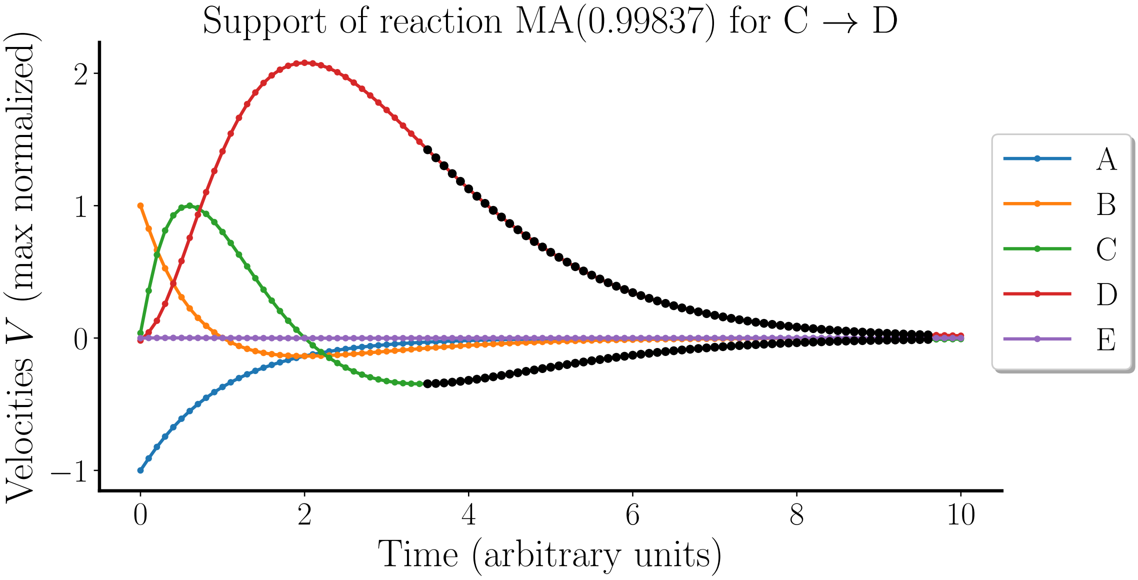


Learned CRN
$A \xrightarrow{0.999} B$
$B \xrightarrow{1.001} C$
$C \xrightarrow{1.002} D$
$D \xrightarrow{0.999} E$

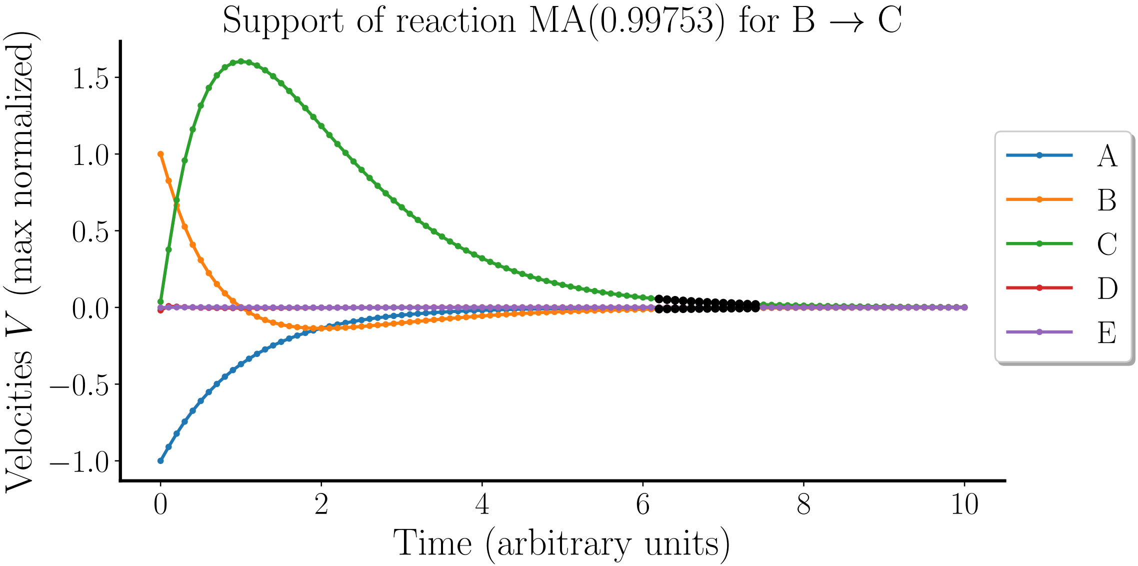
Example - iterative inference of reactions for Chain CRN



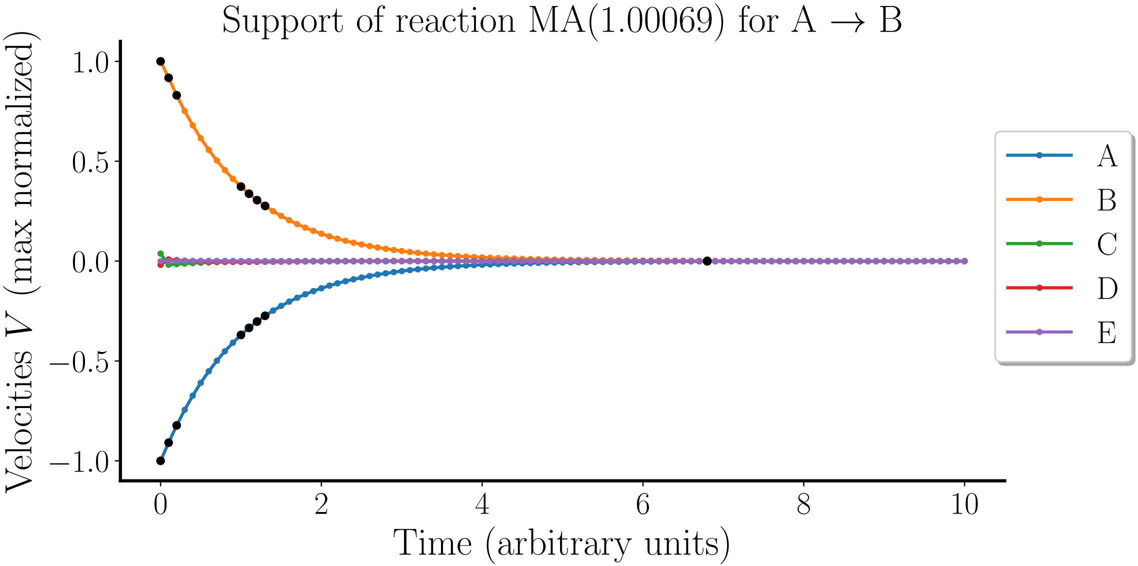
Example - iterative inference of reactions for Chain CRN



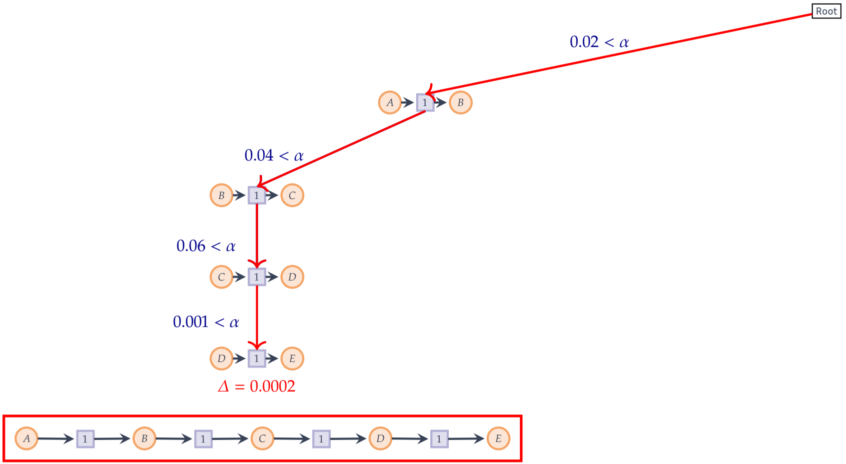
Example - iterative inference of reactions for Chain CRN



Example - iterative inference of reactions for Chain CRN



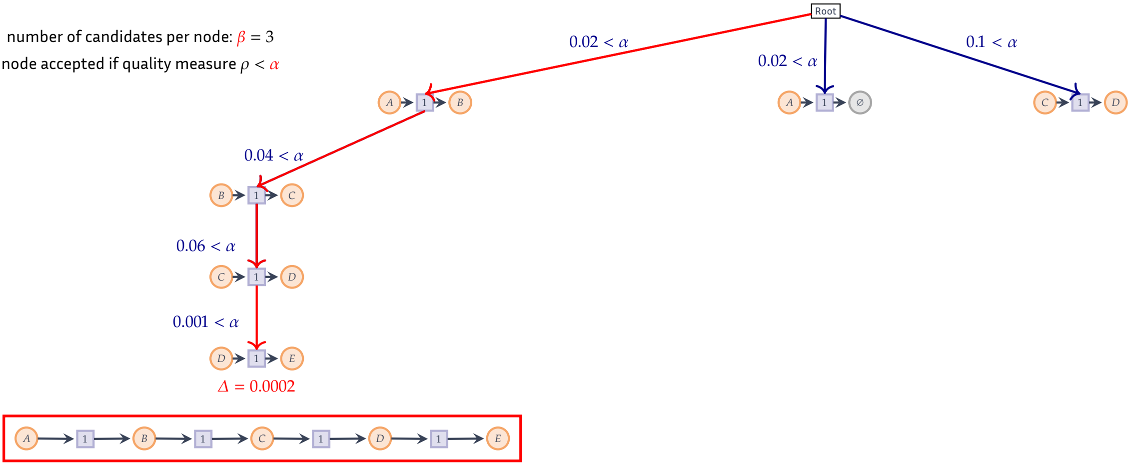
Reactmine search algorithm



Reactmine search algorithm

number of candidates per node: $\beta = 3$

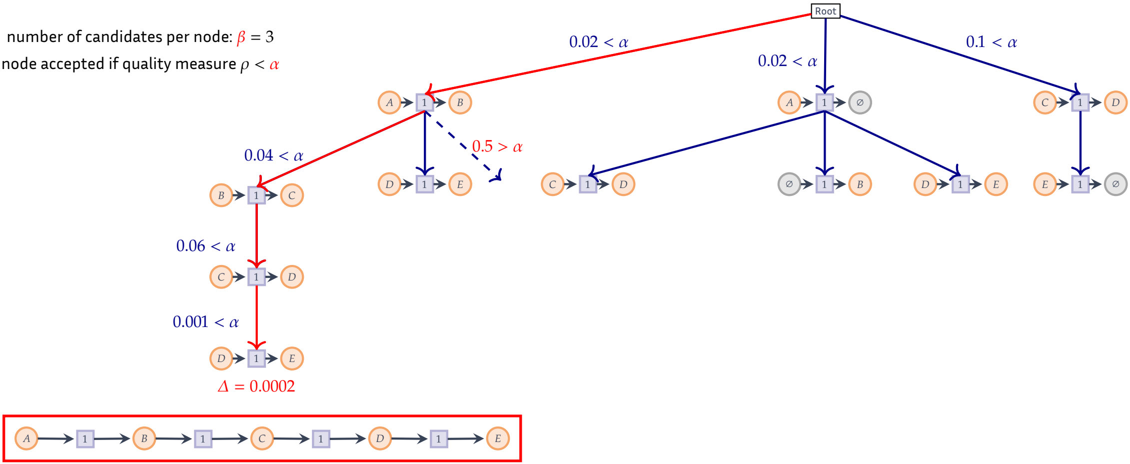
node accepted if quality measure $\rho < \alpha$



Reactmine search algorithm

number of candidates per node: $\beta = 3$

node accepted if quality measure $\rho < \alpha$

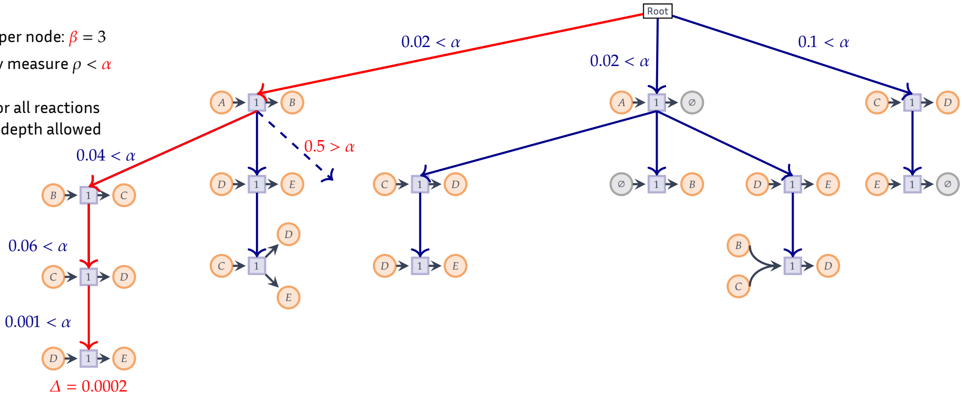


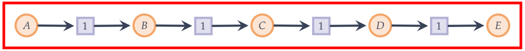
Reactmine search algorithm

number of candidates per node: $\beta = 3$

node accepted if quality measure $\rho < \alpha$

Termination: $\rho > \alpha$ for all reactions
or reached γ maximal depth allowed



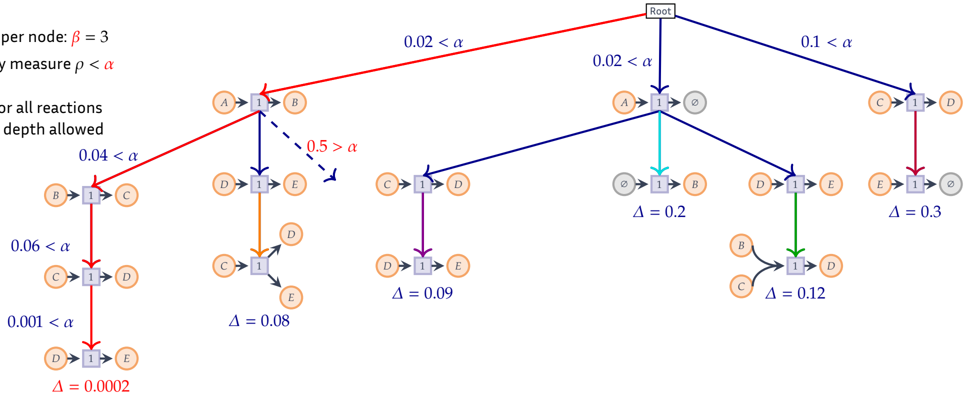


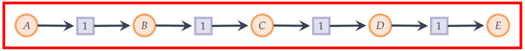
Reactmine search algorithm

number of candidates per node: $\beta = 3$

node accepted if quality measure $\rho < \alpha$

Termination: $\rho > \alpha$ for all reactions
or reached γ maximal depth allowed



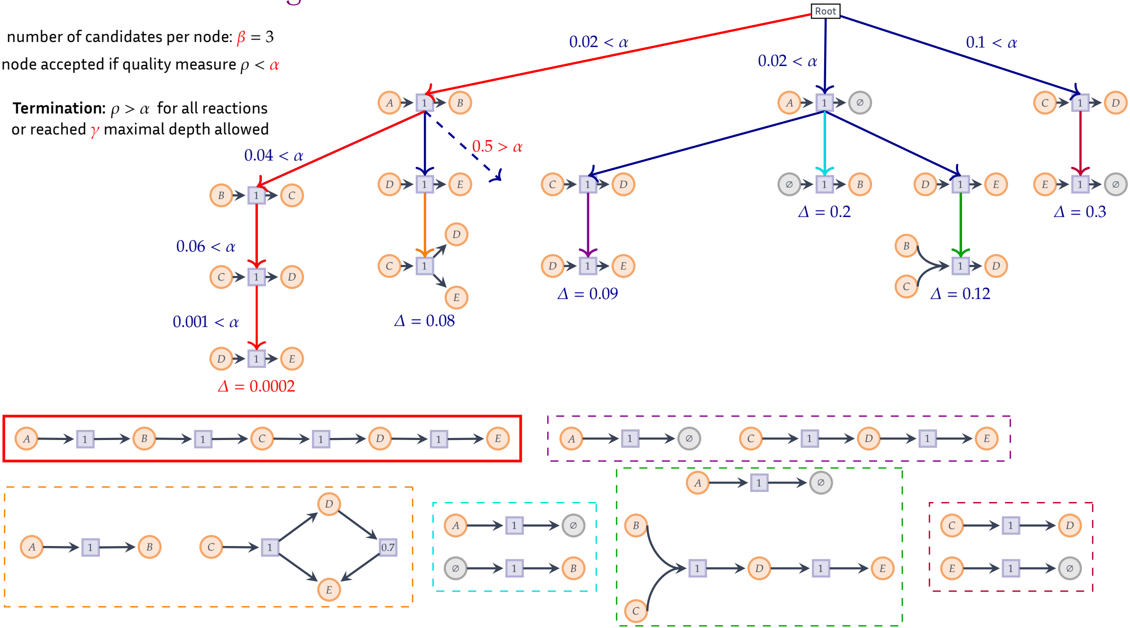


Reactmine search algorithm

number of candidates per node: $\beta = 3$

node accepted if quality measure $\rho < \alpha$

Termination: $\rho > \alpha$ for all reactions
or reached γ maximal depth allowed

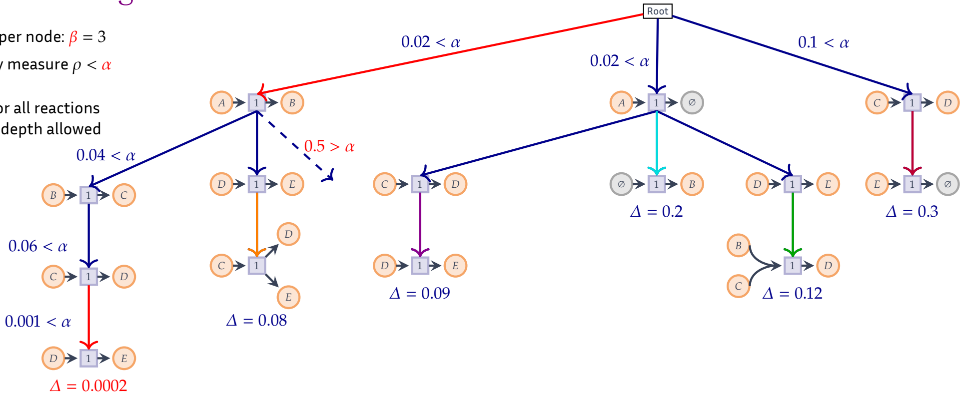


Reactmine search algorithm

number of candidates per node: $\beta = 3$

node accepted if quality measure $\rho < \alpha$

Termination: $\rho > \alpha$ for all reactions
or reached γ maximal depth allowed



4 Hyperparameters:

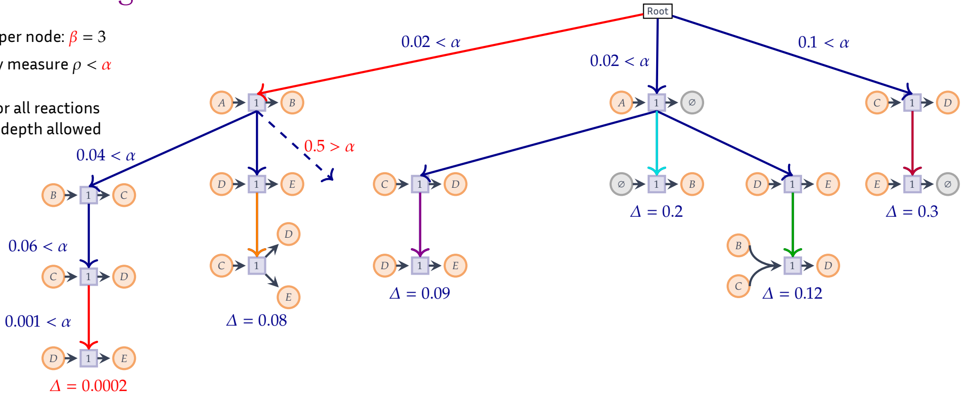
- δ_{max} Species variations similarity threshold
- α CV threshold
- γ CRN size limit
- β Number of reaction candidates per node

Reactmine search algorithm

number of candidates per node: $\beta = 3$

node accepted if quality measure $\rho < \alpha$

Termination: $\rho > \alpha$ for all reactions
or reached γ maximal depth allowed

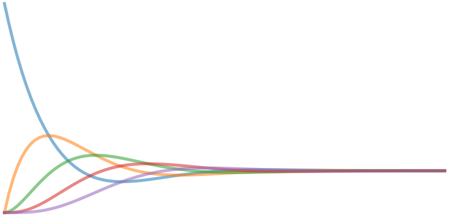


4 Hyperparameters:

- δ_{max} Species variations similarity threshold
- α CV threshold
- γ CRN size limit
- β Number of reaction candidates per node

Hyperparameter selection by minimization of Δ

Evaluation on Loop CRN



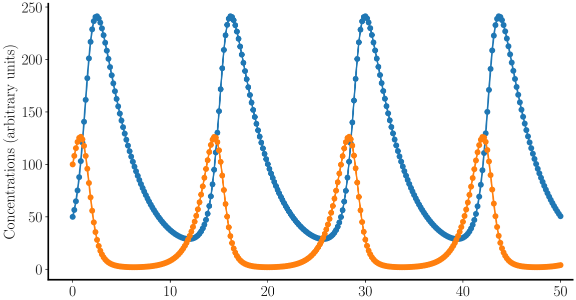
Hidden CRN	Learned CRN
$A \xrightarrow{1} B$	$A \xrightarrow{1} B$
$B \xrightarrow{1} C$	$B \xrightarrow{1} C$
$C \xrightarrow{1} D$	$C \xrightarrow{1} D$
$D \xrightarrow{1} E$	$D \xrightarrow{1} E$
$E \xrightarrow{1} A$	$E \xrightarrow{1} A$

Inference difficult → each species takes part in two reactions

$$\begin{cases} \frac{dA}{dt} = k_5 E - k_1 A \\ \frac{dB}{dt} = k_1 A - k_2 B \\ \frac{dC}{dt} = k_2 B - k_3 C \\ \frac{dD}{dt} = k_3 C - k_4 D \\ \frac{dE}{dt} = k_4 D - k_5 E \end{cases}$$

Lokta-Volterra

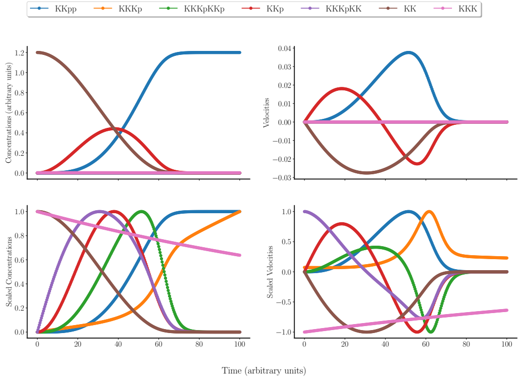
Ground-truth	Learned CRN	SINDy ODE
$\mathrm{A} \xrightarrow{0.3} \varnothing$	$\mathrm{A} \xrightarrow{0.298} \varnothing$	$\begin{cases} \dot{A} = -0.299A + 0.010AB \\ \dot{B} = 0.995B - 0.010AB \end{cases}$
$B \xrightarrow{1} 2B$	$B \xrightarrow{0.994} 2B$	
$A + B \xrightarrow{0.01} 2A$	$A + B \xrightarrow{0.010} 2A$	



Simplified MAPK Cascade

$$KKK \xrightarrow{0.0045} KKKp$$
$$KKKp + KK \xrightarrow{1000} KKKpKK$$
$$KKKpKK \xrightarrow{150} KKKp + KK$$
$$KKKpKK \xrightarrow{150} KKKp + KKp$$
$$KKKp + KKp \xrightarrow{1000} KKKpKKp$$
$$KKKpKKp \xrightarrow{150} KKKp + KKp$$
$$KKKpKKp \xrightarrow{150} KKKp + KKpp$$

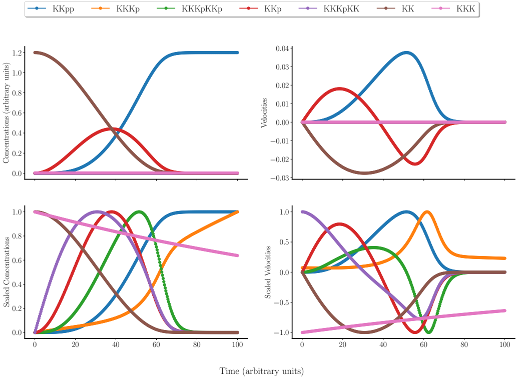
$$KKK \xrightarrow{0.0045} KKKp$$
$$KKp + KKKp \xrightarrow{499.97} KKpp + KKKp$$
$$KK + KKKp \xrightarrow{501.01} KKKp$$
$$KKKpKK \xrightarrow{150.04} KKKp + KKp$$
$$KKKp + KK \xrightarrow{501.19} KKKpKK$$
$$KKKpKK \xrightarrow{150.37} KKKp + KK$$
$$KKKp + KKp \xrightarrow{517.78} KKKpKKp$$
$$KKKpKKp \xrightarrow{155.34} KKKp + KKp$$
$$KKKp + KK \xrightarrow{500.19} KKKpKK + KK$$



Simplified MAPK Cascade

$$KKK \xrightarrow{0.0045} KKKp$$
$$KKKp + KK \xrightarrow{1000} KKKpKK$$
$$KKKpKK \xrightarrow{150} KKKp + KK$$
$$KKKpKK \xrightarrow{150} KKKp + KKp$$
$$KKKp + KKp \xrightarrow{1000} KKKpKKp$$
$$KKKpKKp \xrightarrow{150} KKKp + KKp$$
$$KKKpKKp \xrightarrow{150} KKKp + KKpp$$

$$KKK \xrightarrow{0.0045} KKKp$$
$$KKp + KKKp \xrightarrow{499.97} KKpp + KKKp$$
$$KK + KKKp \xrightarrow{501.01} KKKp$$
$$KKKpKK \xrightarrow{150.04} KKKp + KKp$$
$$KKKp + KK \xrightarrow{501.19} KKKpKK$$
$$KKKpKK \xrightarrow{150.37} KKKp + KK$$
$$KKKp + KKp \xrightarrow{517.78} KKKpKKp$$
$$KKKpKKp \xrightarrow{155.34} KKKp + KKp$$
$$KKKp + KK \xrightarrow{500.19} KKKpKK + KK$$



- Adding up Reaction 3 and 9 is ODE-equivalent to $KKKp + KK \rightarrow KKKpKK$
- This reaction has already been inferred (5^{th}) $\rightarrow$ simplification
- The “simplified” inferred CRN has 7 reactions 6 of which are accurate.

SINDy inferred ODE system for MAPK

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

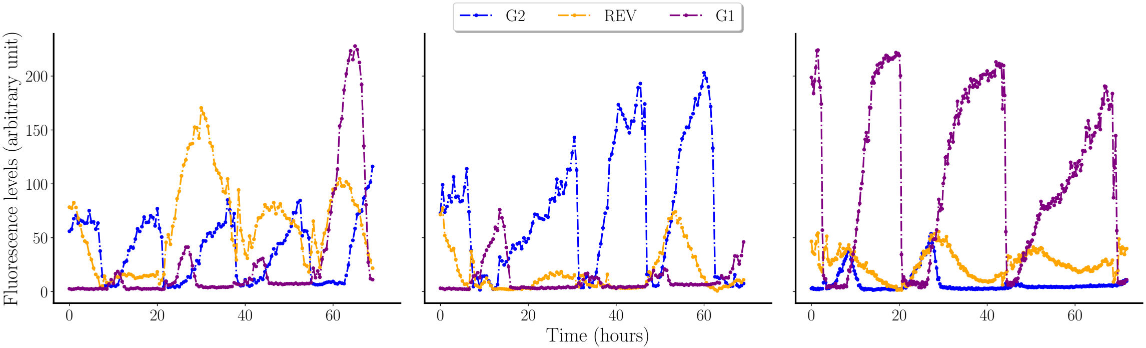
Application on real data: videomicroscopy

- NIH3T3 embryonic mouse fibroblasts left to proliferate in regular medium supplemented with 20% FBS concentration

- Time lapse videomicroscopy, one image taken every 15 minutes during 72 hours

- Cell tracking using three different fluorescent markers of the circadian clock and the cell cycle:
  - Reverbα circadian clock protein reporter
  - Fluorescence Ubiquitination Cell Cycle Indicators, Cdt1 and Geminin, two cell cycle proteins which accumulate during the G1 and S/G2/M phases, respectively.

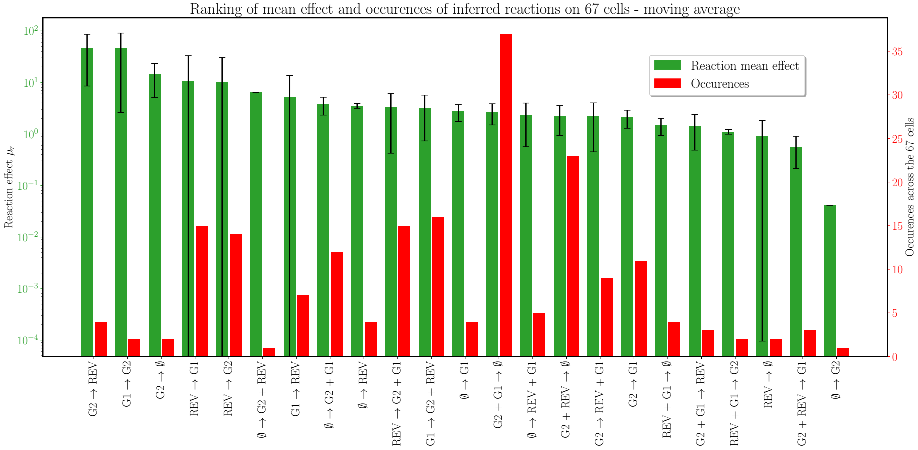
Highly heterogeneous cell behavior



- 67 cells after curation
- Data smoothing using a moving average
- High heterogeneity $\longrightarrow$ infer one CRN per cell
- We search for Michaelis-Menten reactions: $f(y) := v_{\max} \dfrac{y}{K_m + y}$

Distribution of inferred reactions

1. For each cell, select the best CRN inferred
2. Compute $n_{occurrences}$ of reaction $r = (R, P, f)$ and mean effect $\mu_r = \frac{1}{nC} \sum_{c=1}^{C} \sum_{l=1}^{n} f(y_l^{(c)})$



$G2 \rightarrow REV$ and $G1 \rightarrow G2$ recovered and present in literature

Conclusion

- A method to **sequentially** infer biochemical reactions.
  - **Parsimony** of the inferred network integrated by construction.
- Philosophy: **“mining” reactions** at specific time points where they are preponderant.
  - More reliable estimation of reaction kinetics based on support
  - **Explainability** of the method through the support set of inferred reaction
- Successfully tackled multi-scaled / cyclic CRNs

Short and long term perspectives

Noise and real data:

- Proper evaluation/treatment against noisy data (e.g. bootstrap)
- Variables non observed at the same time points
- Non uniform grid of observation time points

Short and long term perspectives

Noise and real data:

- Proper evaluation/treatment against noisy data (e.g. bootstrap)
- Variables non observed at the same time points
- Non uniform grid of observation time points

Hidden species:

- Assume two species A, B for which $A(0)$ and $B(0)$ are available, but we only have the time series $X(t) = A(t) + B(t)$
  $\longrightarrow$ Can we still infer a network involving A and B?
- Infer completely unobserved hidden species?
- **→Evolutionary algorithm**

Scaling:

- Consider larger networks (10 species)