MMBioS Meeting, Feb 2017

Model Checking Techniques for Systems Biology Modeling: A Case Study of Neurotransmission

Bing Liu

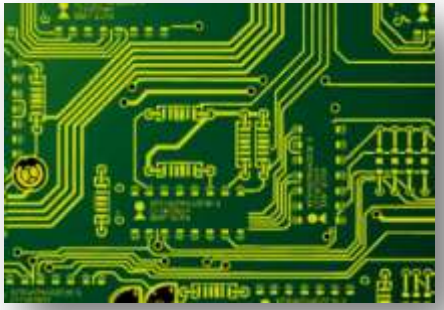
TRD1 & 3, DBP1

Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh

Complex Systems

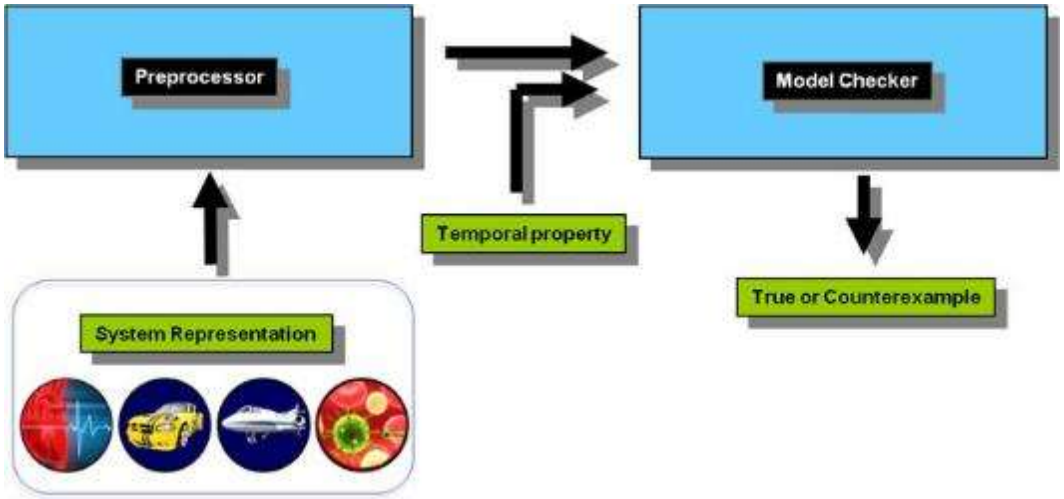
Model Checking

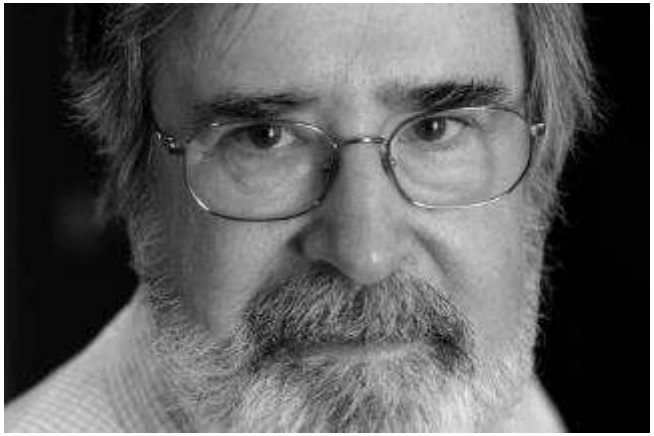
- Detect bugs in a variety of hardware and software applications
  - E.g. microprocessor, railway system, satellite-control software
- Many industrial successes
  - Intel, IBM, Apple, Microsoft, Motorola, Airbus, etc.



Model Checking

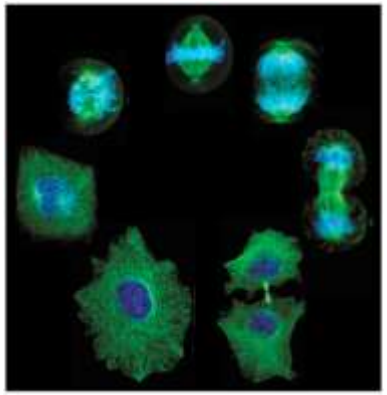
- An automated method to formally verify a system's behavior with respect to a set of properties





Edmund M. Clarke

Biological Machines



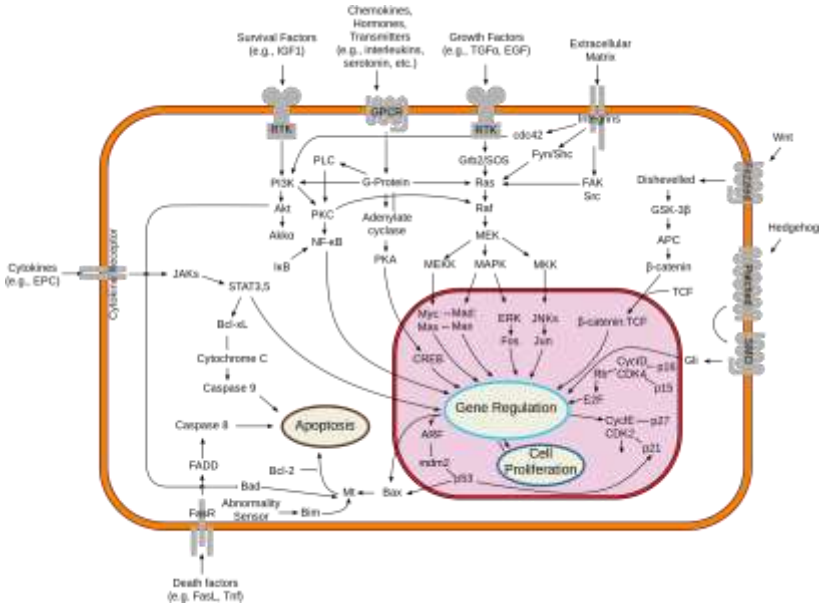
Cell cycle

ATP synthase

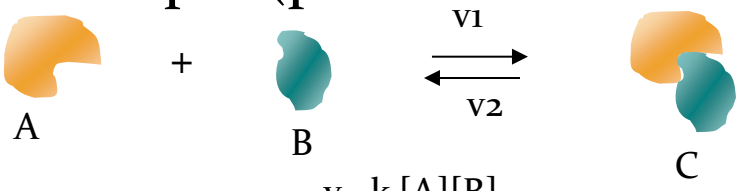
Microtubule assembly, vesicle transport driven by motor proteins, protein synthesis by ribosome, power station mitochondria

Systems Biology Modeling

- Mathematical formalisms
  - Ordinary Differential Equations
  - Petri Nets
  - Hybrid Automata
  - Markov chains (e.g. CTMC)
  - ***BioNetGen language***
  - …



- ODE Example (protein association):



$v_1 = k_1[A][B]$

$v_2 = k_2[C]$

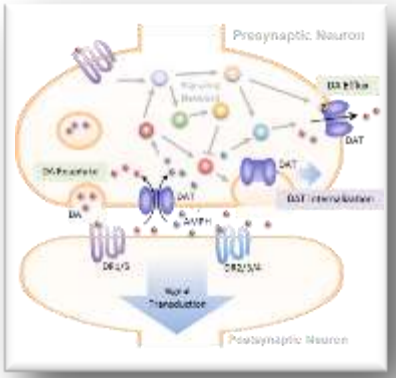
$d[B]/dt = - v_1 + v_2 = k_1[A][B] - k_2[C]$

Mass action law

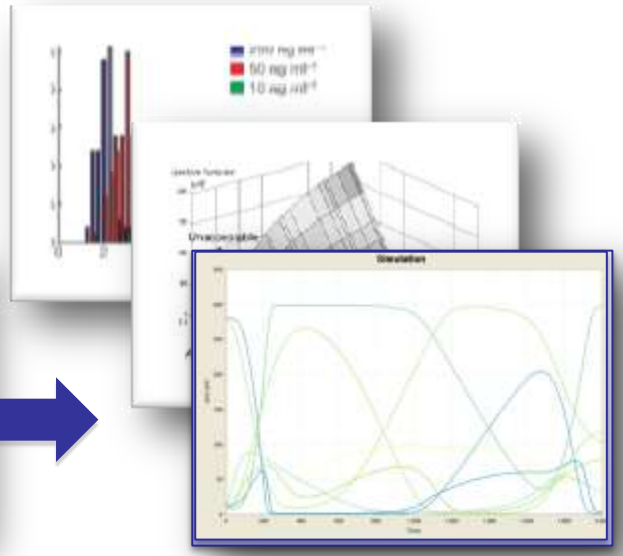
Problems Faced

- Which hypothesis is more plausible?
- **How to estimate unknown model parameters?**
- Which component is critical to the dynamics?
- How to control to system to get a desired behavior?

Model checking can help!

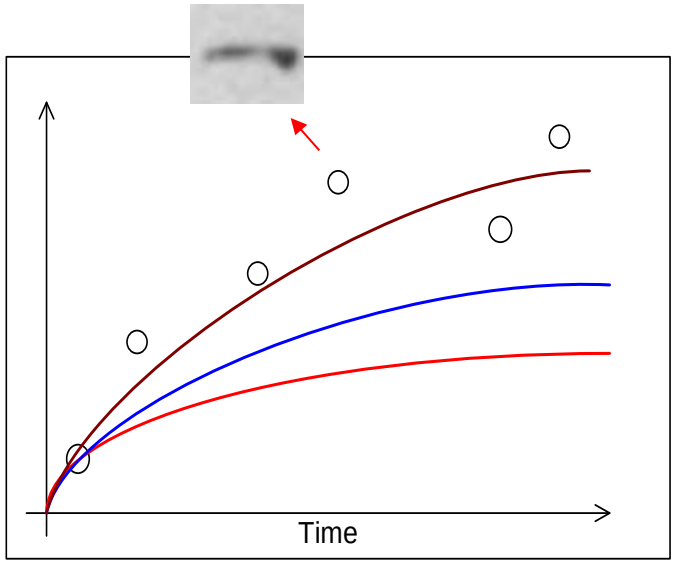



```
begin molecule type
L() R()
end molecule type
begin reaction rules
L(r) + R(l) <-> L(r!1).R(l!1) kp1, km1
end reaction rules
generate_network()
simulate({method=>"ode",t_end=>500,n_s
teps=>500})
```



Parameter Estimation

- Goal:
  - Find values of parameter so that model predictions can match experimental data (e.g. time serials, steady state)



krbNGF = 0.33, KmAkt = 0.16, kpRaf1 = 0.42 … …

target

krbNGF = 0.49, KmAkt = 0.08, kpRaf1 = 0.97 … …

krbNGF = 0.88, KmAkt = 0.21, kpRaf1 = 0.05 … …

Optimization Approach

- Minimize the difference between model prediction and experimental data

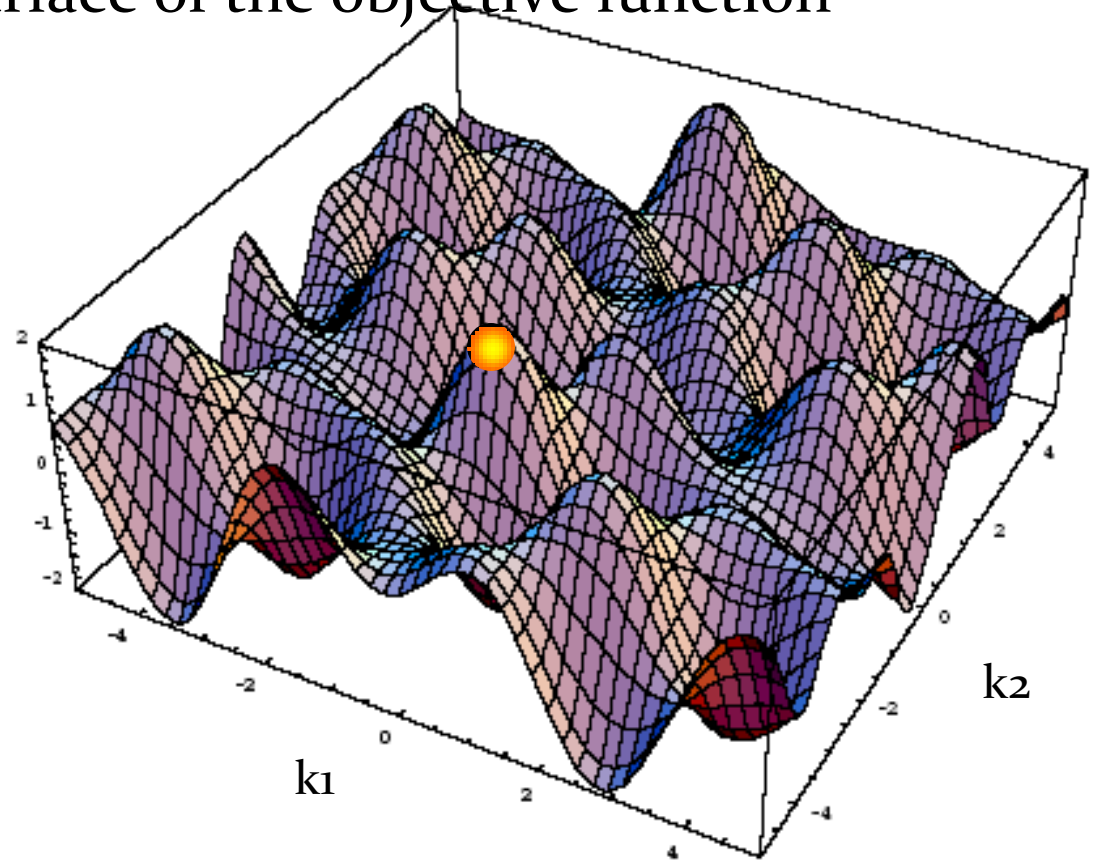
Given data $\tilde{\mathbf{x}}(t_j)$, find $\mathbf{k}$ to

minimize $J(\mathbf{k}) = \sum_j ||\mathbf{x}(t_j; \mathbf{k}) - \tilde{\mathbf{x}}(t_j)||^2$

J: objective function

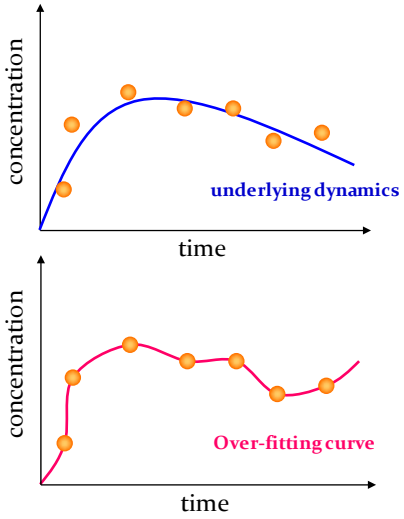
Example: Steepest Decent

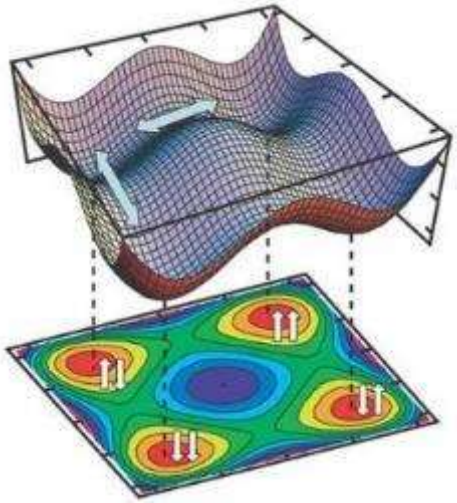
- Update following the direction of steepest descent on the hyper-surface of the objective function



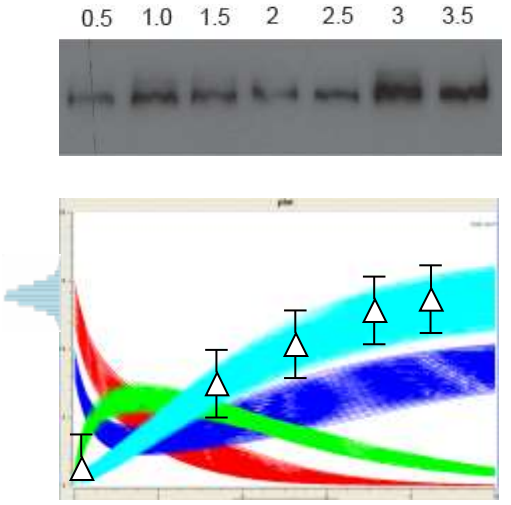
Many Challenges

- The curse of dimensionality
- Over-fitting
- Non-identifiable models
- Inherent uncertainty of data



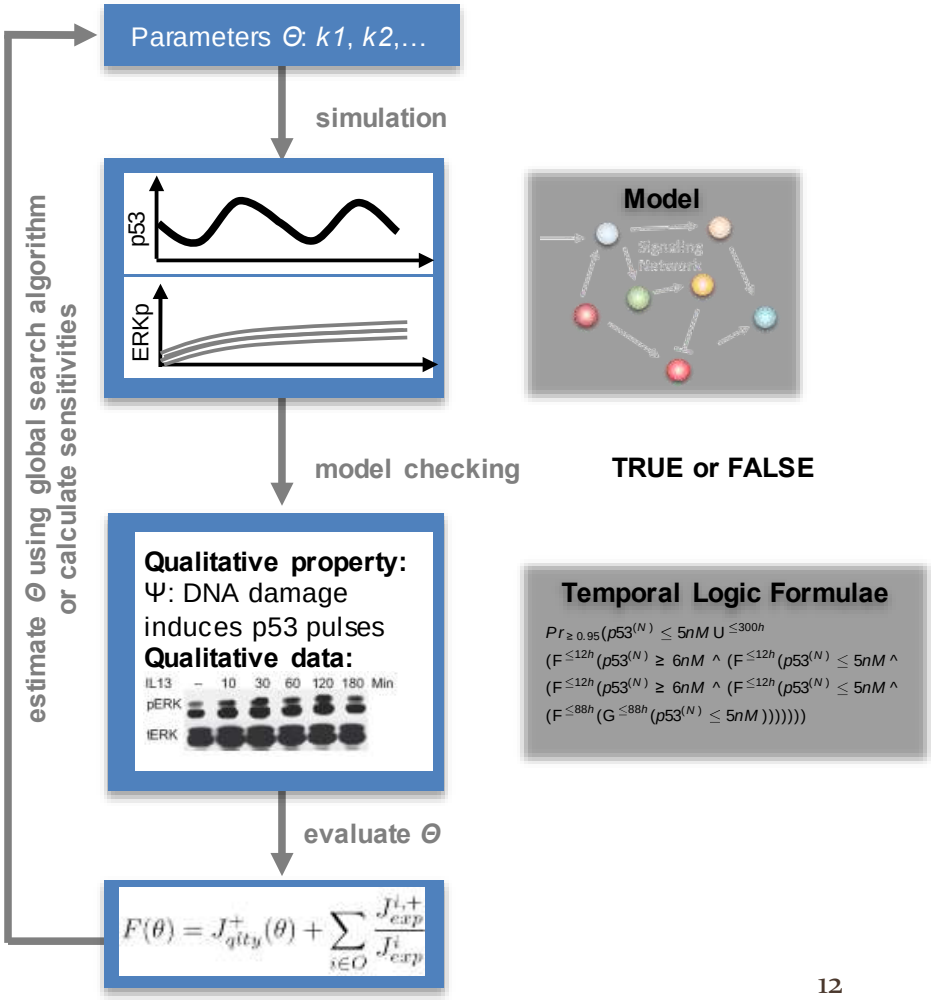


Kim et al. 2007



Our Solution

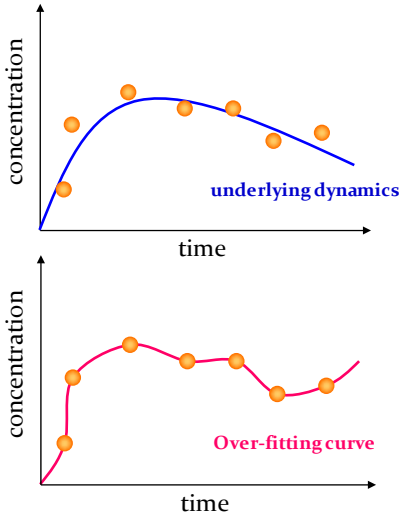
- A statistical model checking (SMC) based approach
  - Encode training data as a ***bounded linear temporal logic (BLTL)*** formula
  - Evaluate candidate parameters using SMC
  - Perform global optimization (e.g. stochastic ranking evolutional strategy, SRES)

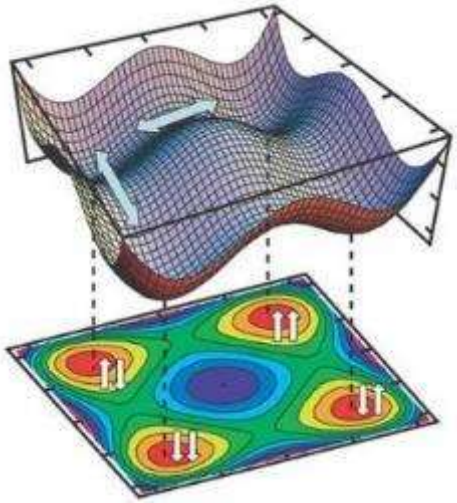


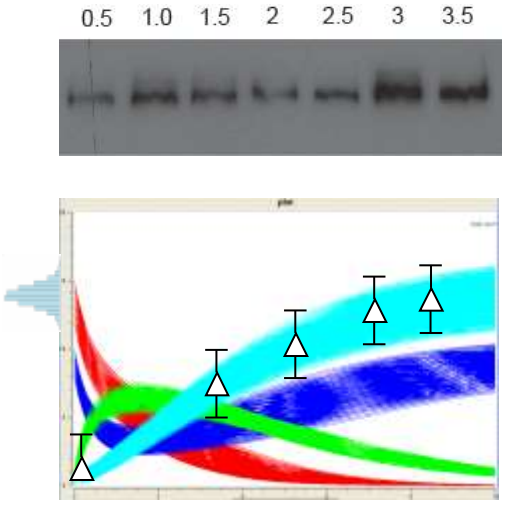
Palaniappan et al, CMSB, 2013, Liu & Faeder, BIBM, 2016

Our Solution

- Advantages
  - Utilize both ***quantitative*** and ***qualitative*** knowledge
  - Deal with uncertainty of the biological system/data
  - Good scalability due to the power of statistical testing







Kim et al. 2007

How to encode knowledge?

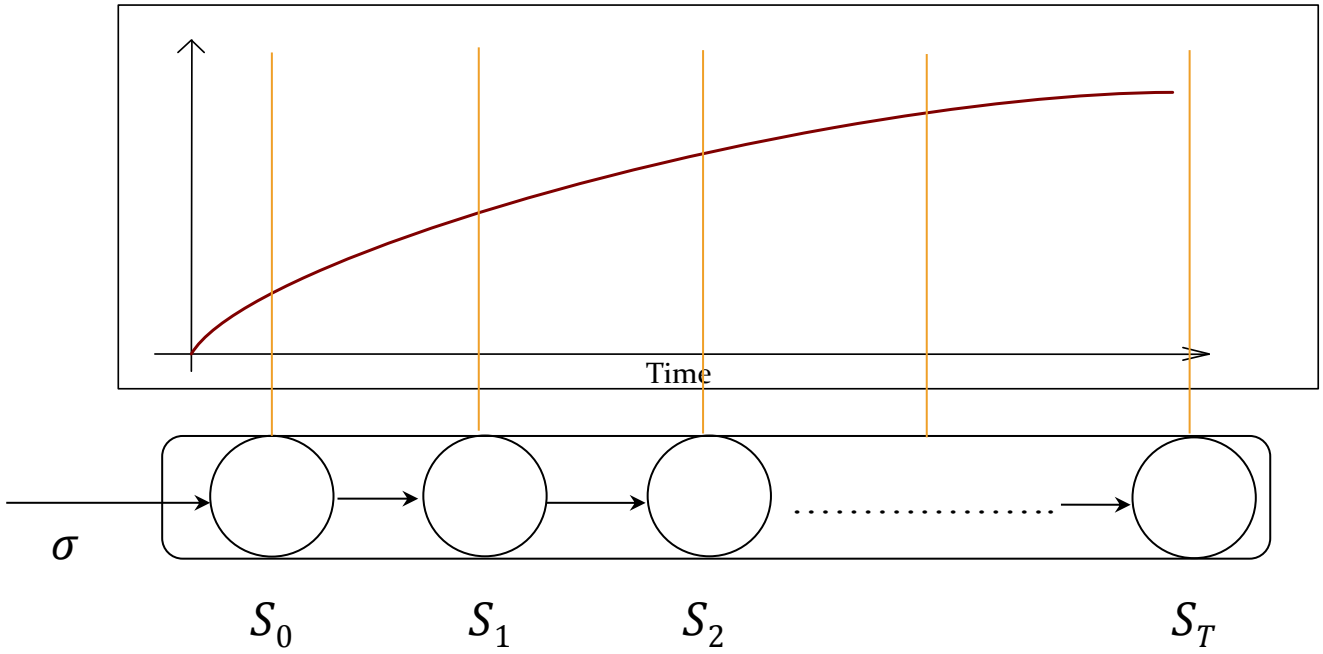
- E.g.
  - "ERKp level is between 10nM and 20nM"
  - "*Caspase-3 level sustains once it reaches threshold 30nM*"
- Temporal logic
  - A bounded linear temporal logic for biological properties (CSMB'13)

$$C3 \le 1nM \mathbf{U}^{10h} (\mathbf{F}^{\le 56h} (C3 \ge 30nM \wedge \mathbf{G}^{\le 44h} (C3 \ge 30nM)))$$

Amir Pnueli

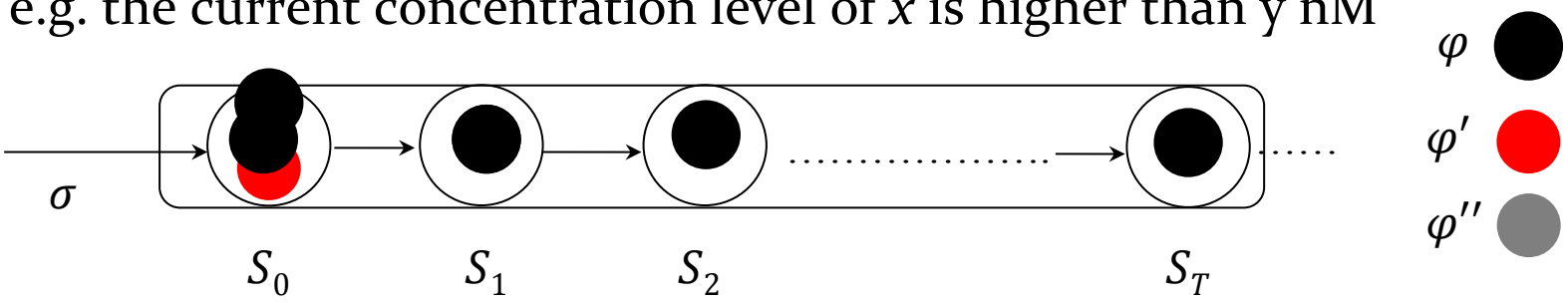
BLTL

- A finite set of time points: $\mathbf{T} = \{0,1,\ldots,T\}$
- A trajectory is represented by
  $\sigma = (s_o, t_o), (s_1, t_1), \ldots (s_T, t_T) \ldots$



BLTL

- Atomic (elementary) proposition: $x\#\, y, \# \in \{>, <, =, \leq, \geq\}$
  - e.g. the current concentration level of *x* is higher than y nM



- The formulas are built over operators $\wedge, \vee, \neg, \mathbf{O}, \mathbf{G}^{\leq t}, \mathbf{G}^{t}, \mathbf{F}^{\leq t}, \mathbf{F}^{t}, \mathbf{U}^{\leq t}, \mathbf{U}^{t}$

$\sigma(o) \vDash \varphi \vee \varphi'$, $\sigma(o) \vDash \varphi \wedge \varphi'$, $\sigma(o) \vDash \varphi$, $\sigma(o) \vDash \sim\varphi''$

$\sigma(o) \vDash O(\varphi)$, φ is true in the next state

$\sigma(o) \vDash \varphi U^{\leq t} \varphi'$, φ will be true until φ' is true

$\sigma(o) \vDash F^{\leq t}(\varphi')$, φ' will be true some time in the future

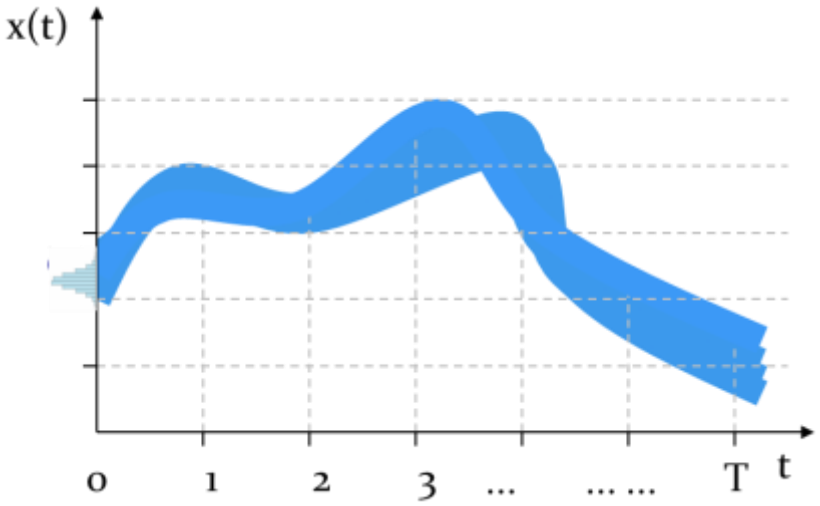
$\sigma(o) \vDash G^{\leq t}(\varphi)$, φ will be globally true in the future

$\sigma(o) \vDash F^{t}(\varphi)$, φ is true at time point t

Probabilistic BLTL

- Example:
  - *Caspase-3 level sustains once it reaches threshold 30nM with a probability at least 0.95*

$$\Pr_{\geq 0.95}(C3 \leq 1nM \mathbf{U}^{10h}(\mathbf{F}^{\leq 56h}(C3 \geq 30nM \wedge \mathbf{G}^{\leq 44h}(C3 \geq 30nM))))$$



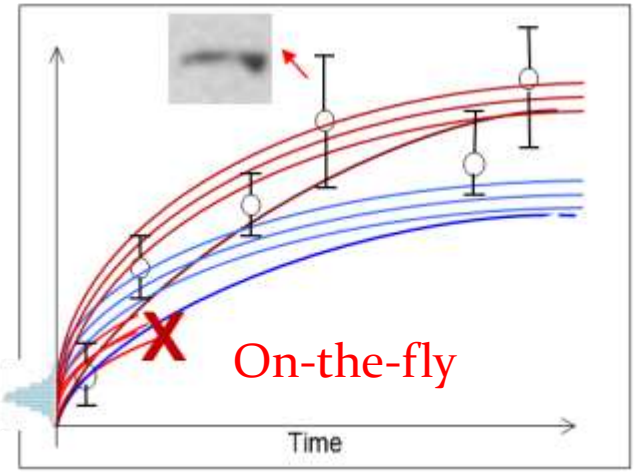
SMC of PBLTL formulas

- Check $M \models \Pr_{\geq r}(\psi)$ using a sequential hypothesis test between

$$\text{H0}: p \geq r + \delta \text{ and } \text{H1}: p \leq r - \delta$$

- Generate a sequence of sample trajectories: $\sigma_1, \sigma_2, \ldots$
- Verify each trajectory and determine whether accept Ho or H1 based on Type I/II error bounds (α, β):

$$q_m = \frac{[r-\delta]^{(\sum_{i=1}^{m} y_i)}[1-[r-\delta]]^{(m-\sum_{i=1}^{m} y_i)}}{[r+\delta]^{(\sum_{i=1}^{m} y_i)}[1-[r+\delta]]^{(m-\sum_{i=1}^{m} y_i)}}$$



Knowledge Encoding

- Quantitative experimental data

$$\psi_i^t = \mathbf{F}^t(l_i^t \le x_i \wedge x_i \le u_i^t)$$

$$\psi_{\exp} = \wedge_{i \in O}(\wedge_{t \in T_i} \psi_i^t)$$

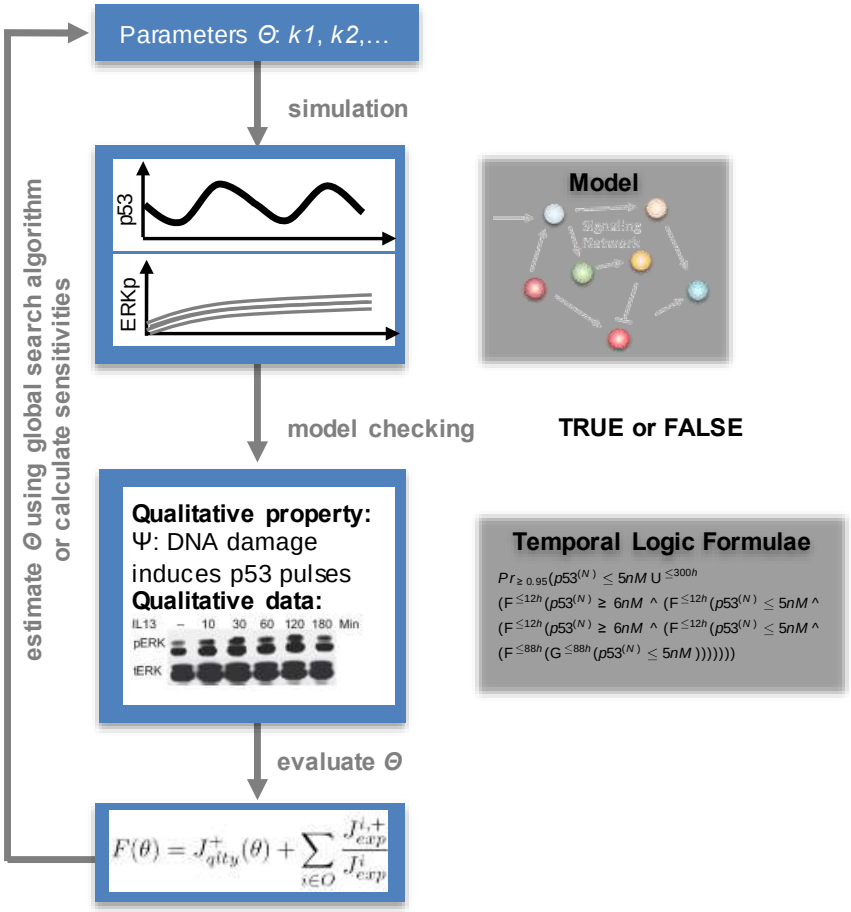
- Qualitative properties of the dynamics
  - E.g. transient/sustained activation, oscillatory behavior, bistable, …
  - *'trend'* formulas: ψ_{qlty}

- PBLTL formula: $\Pr_{\ge r}(\psi_{\exp} \wedge \psi_{qlty})$

SMC based Parameter Estimation

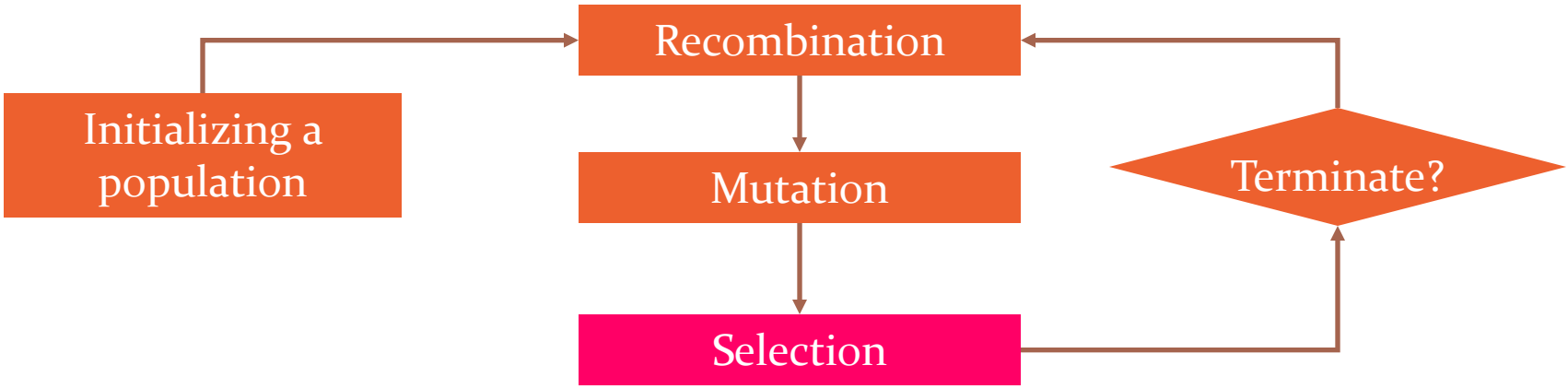
1. Guess θ_l
2. Verify $\psi_{\text{exp}} \wedge \psi_{qlty}$ with the chosen strength
3. Compute $F(\theta_l)$
4. Terminate or make a new guess (based on search strategy e.g. SRES) and repeat step 1

$$F(\theta) = J^{+}_{qlty}(\theta) + \sum_{i \in O} \frac{J^{i,+}_{exp}}{J^{i}_{exp}}$$



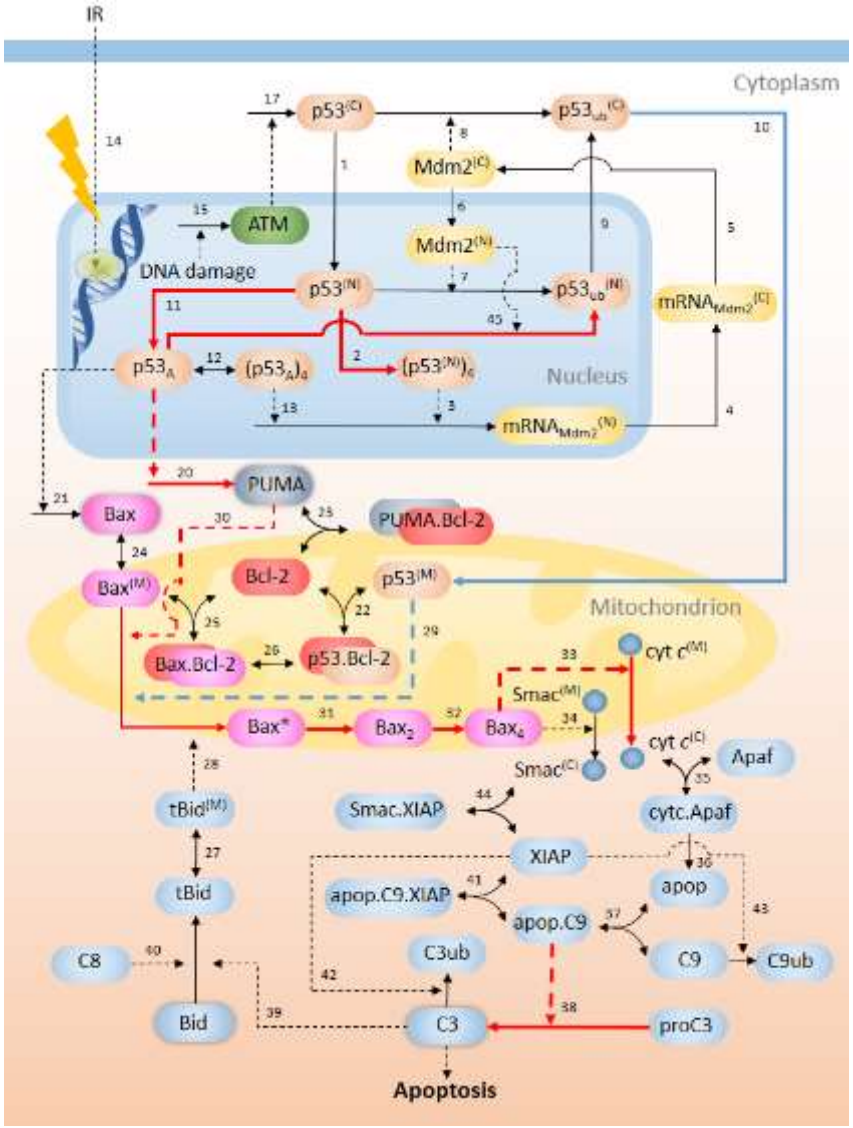
Stochastic Ranking Evolutionary Strategy

- A variation of evolutionary strategy
- Select best λ solutions according to a probabilistic formula
- One of the best performing global method in parameter estimation (*Moles et al, Genome Res 2003*)



Benchmarking

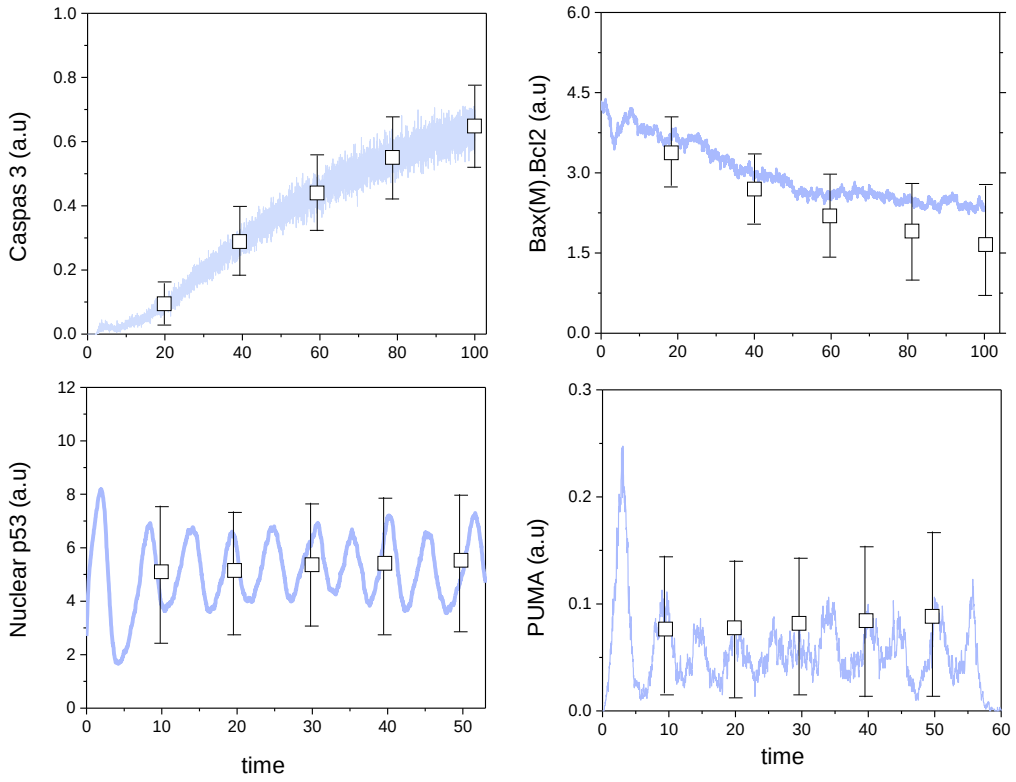
- p53-induced apoptosis
  - 86 rules
  - 160 parameters (10 unknown)
- Synthetic training data
  - 4 species at 5 time points
  - 3 qualitative properties:
    - Mmd2 reaches its peak before p53
    - Sustained caspase-3 once its level reaches certain threshold
    - p53 pulses induce oscillatory behaviors of target genes



Liu et al, Sci Rep, 2014

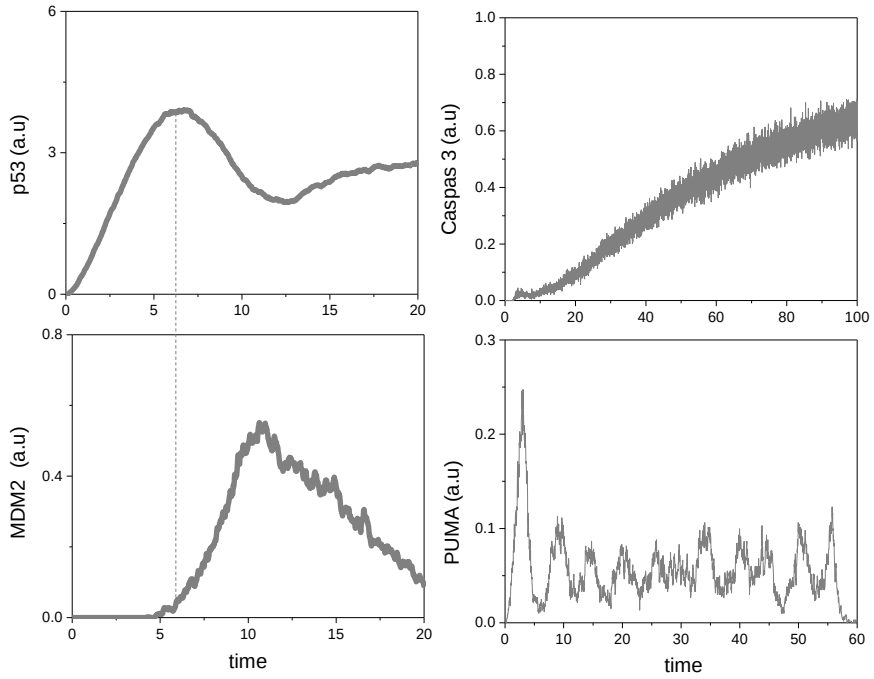
Benchmarking

- Running time: 4.2 hours
- Reproduce quantitative data



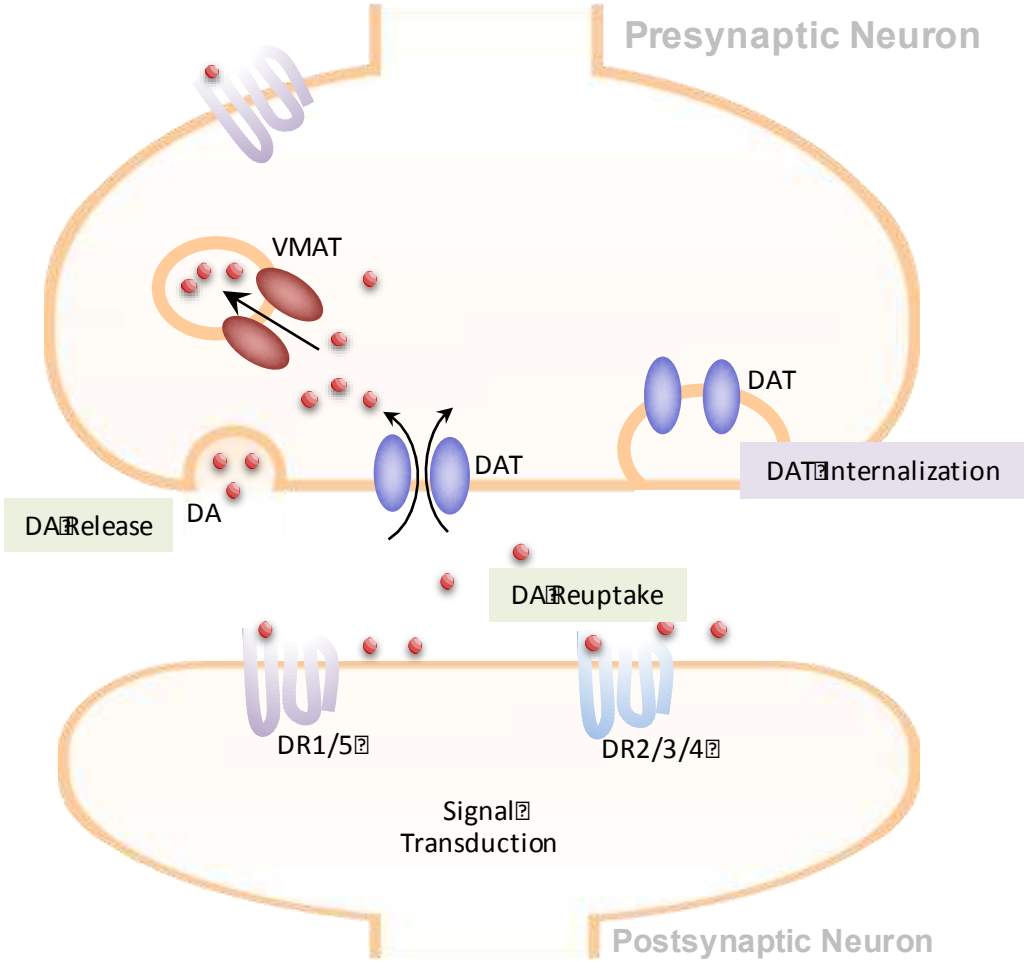
Benchmarking

- Reproduce qualitative behaviors
  - Mmd2 reaches its peak after p53
  - Sustained caspase-3 once its level reaches certain threshold
  - p53 pulses induce oscillatory behaviors of target genes



Amphetamine (AMPH)

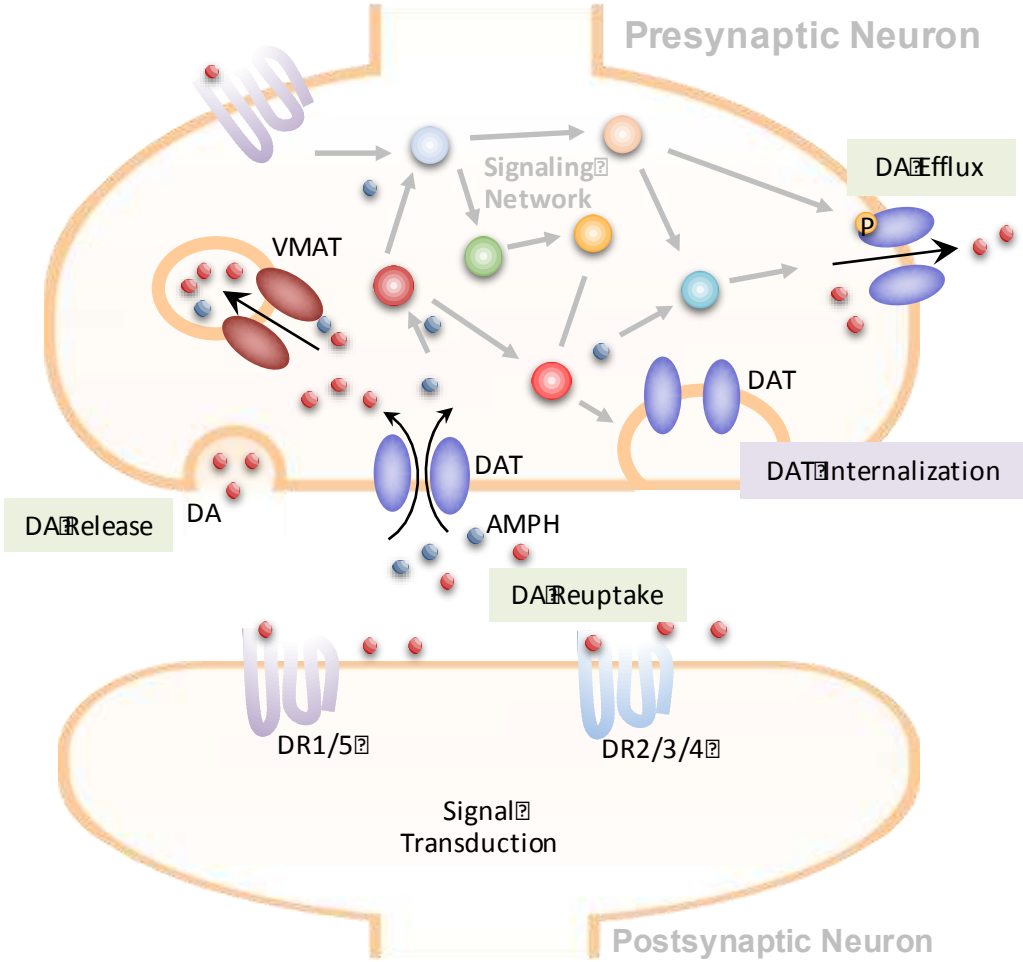
- Induce euphoria and hyperactivity by increasing extracellular dopamine
- AMPH enters DA neurons via DAT
- 'Block' VMAT2
- Enhance DA efflux
  - via PKC, CaMKII, G-protein pathways
- Stimulate DAT internalization
  - via Rho pathway



Susan Amara, DBP1

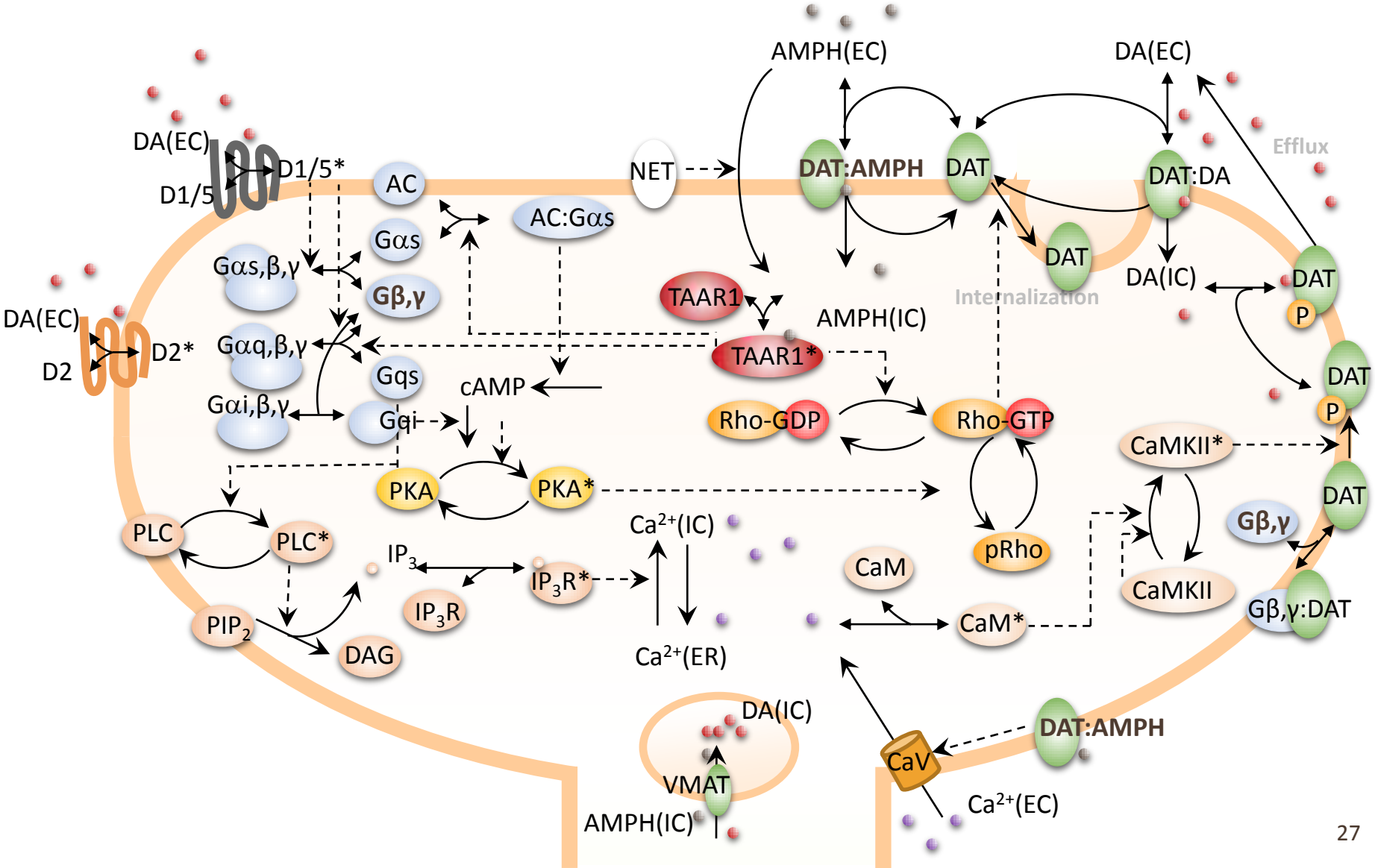
Amphetamine (AMPH)

- Induce euphoria and hyperactivity by increasing extracellular dopamine
- AMPH enters DA neurons via DAT
- 'Block' VMAT2
- Enhance DA efflux
  - via PKC, CaMKII, G-protein pathways
- Stimulate DAT internalization
  - via Rho pathway

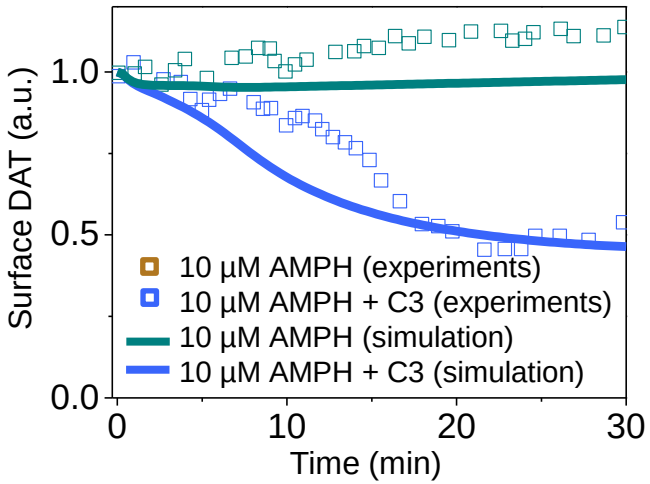


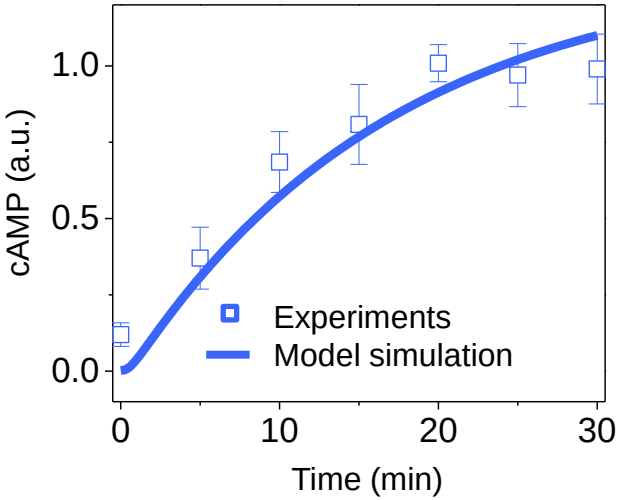
Susan Amara, DBP1

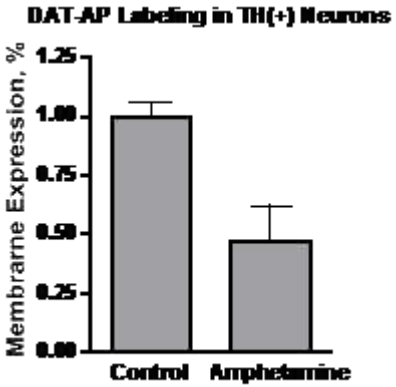
A Kinetic Model

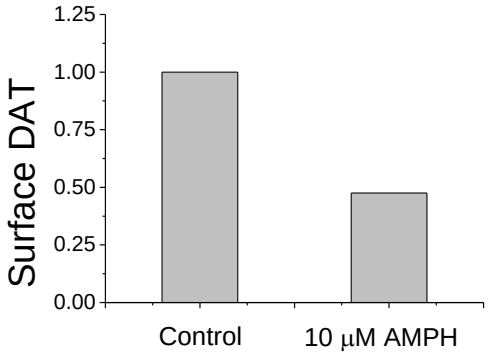


Training data





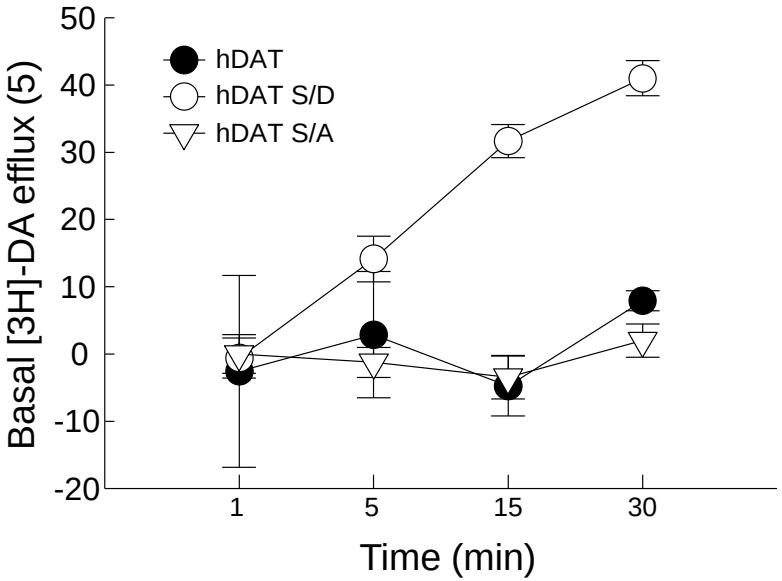


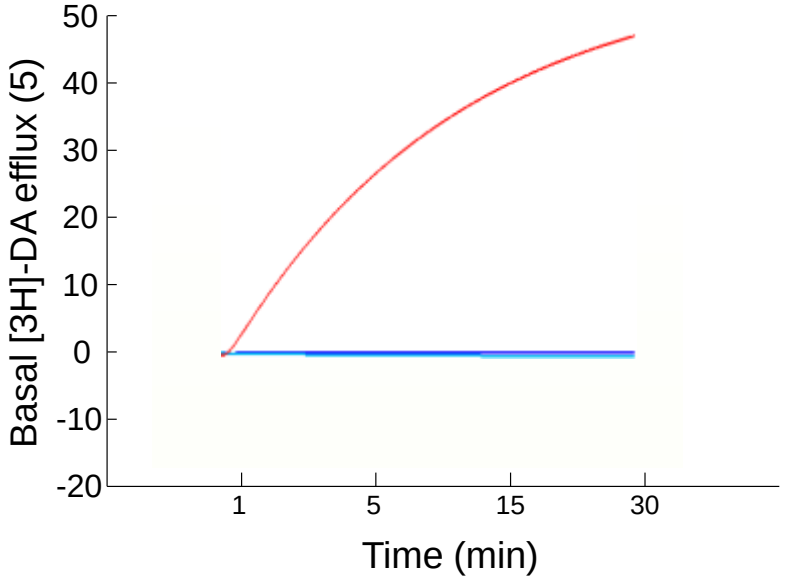


Wheeler et al, PNAS 2015

Training Data

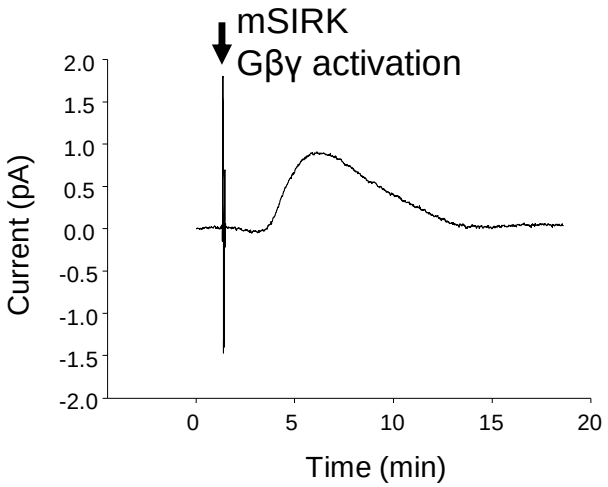
- Effect of N-terminal Serines Phosphorylation on DA efflux

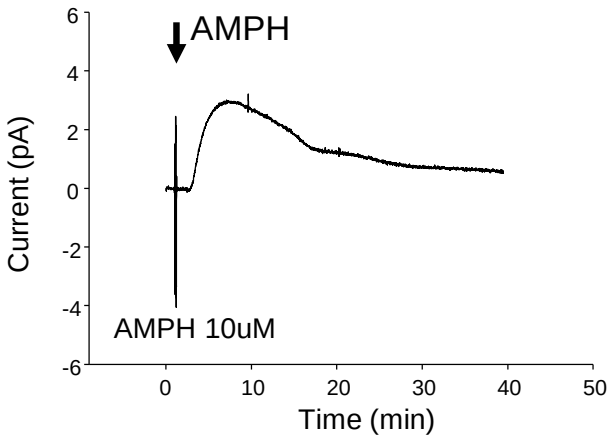


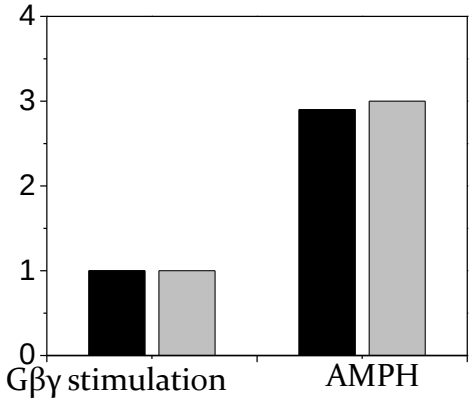


Training Data

- Amperometric recordings for DA efflux after Gβγ stimulation in CHO cells expressing DAT

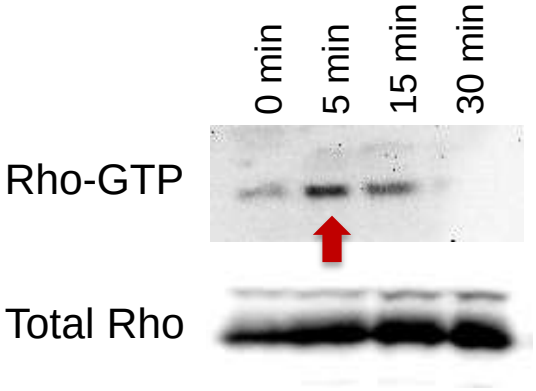


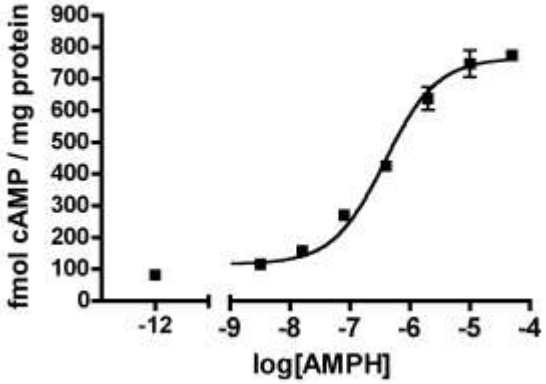


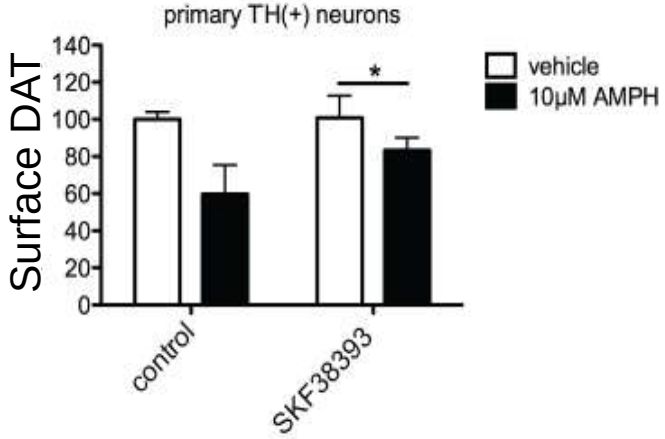


Model reproduces test data

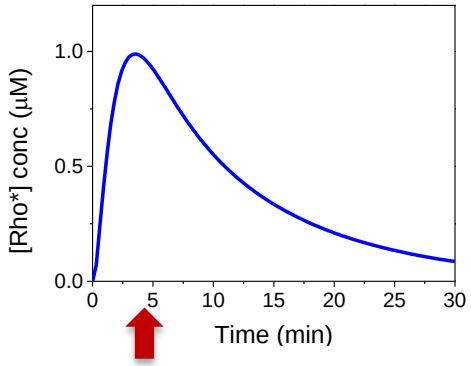
Experimental data

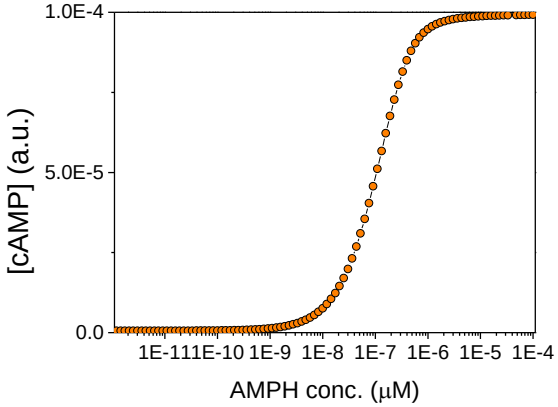


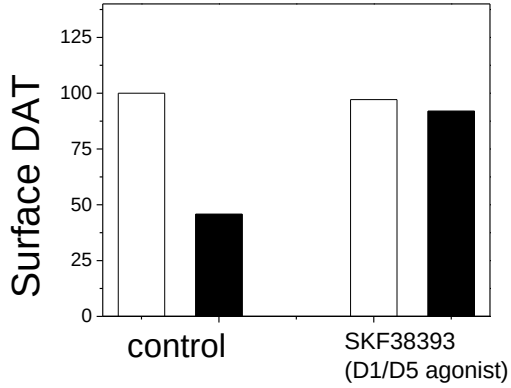




Model prediction



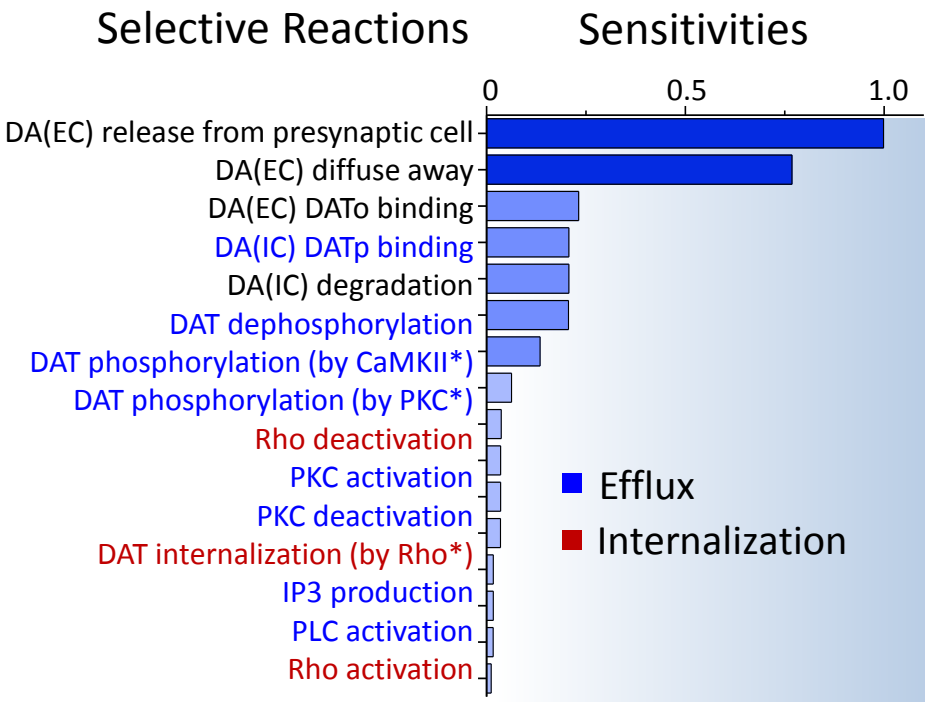


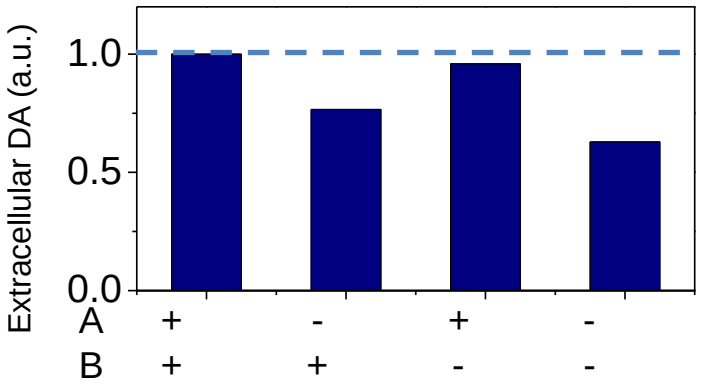


Wheeler et al, PNAS 2015

Model Predictions

- Sensitivity analysis suggests that AMPH modulates DA(EC) level mainly through the DA efflux pathways, than DAT internalization
- Simultaneously block DAT internalization and DA efflux pathways synergistically enhance DA reuptake

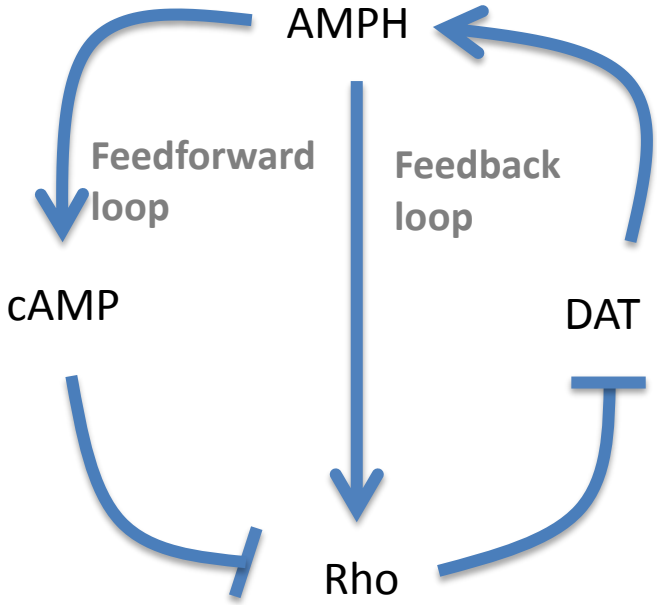


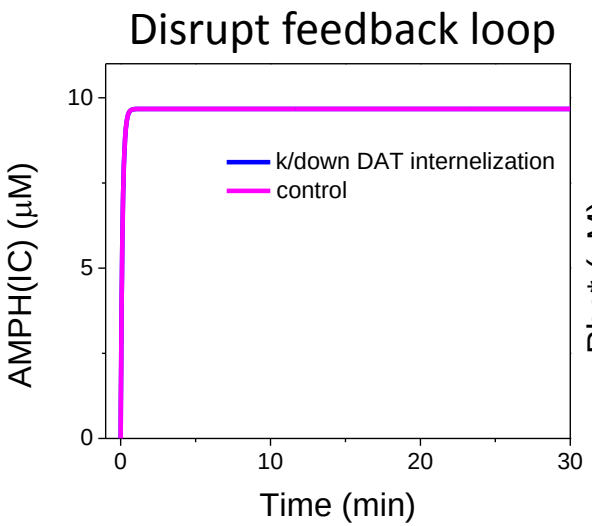


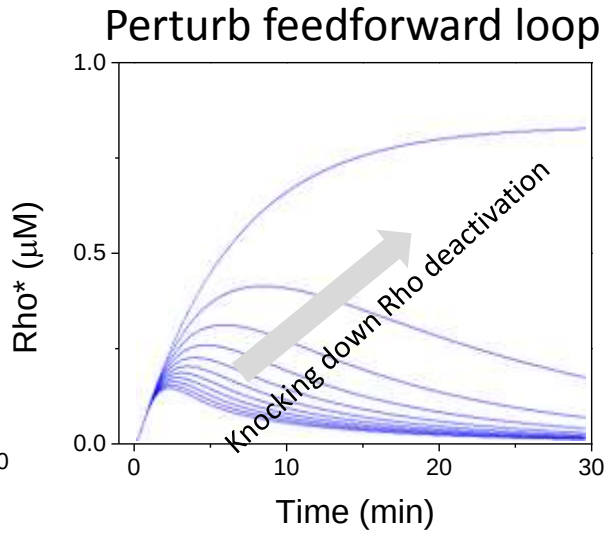
A: DAT phosphorylation by CaMKKII*
B: Rho activation

Model Predictions

- How Rho mediated feedforward/back loops fine-tune AMPH induced DAT internalization
  - The role of the feedback loop is insignificant
  - The feedforward loop governs the time window of Rho activation







Conclusion

- A SMC based approach for the parameter estimation of rule-based models
- Utilize both quantitative and qualitative knowledge
- Deal with uncertainty of biological systems/data
- Good performance due to the power of statistical testing and online model checking

Our MC-based Techniques

Core Technology

System Representation
- DBN (***Bioinformatic, 2012***)
- ODEs (***CMSB'13***)
- Stochastic models (***Sci Rep, 2014***)
- Hybrid Automata (***CMSB'14, HSCC'15, HSB'15***)
- Boolean Network (***CMSB'16***)
- Rule-based models (***BIBM'16***)

Model Checking
- Statistical model checking
- Probabilistic model checking
- δ-decision model checking
- Symbolic model checking

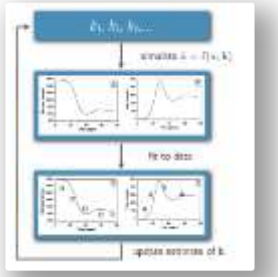
SAT

UNSAT

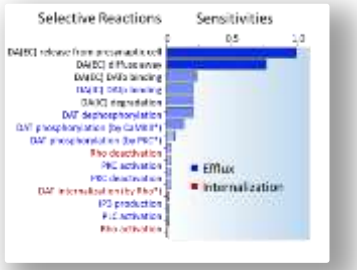
Temporal Property
- Bounded Linear Temporal Logic
- Quantitative property
- Qualitative behaviors

Analysis Methods

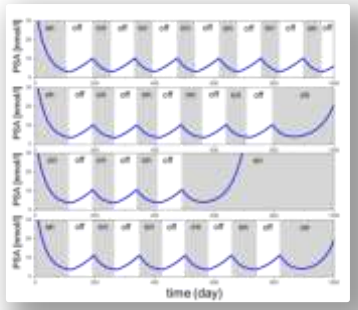
Parameter estimation

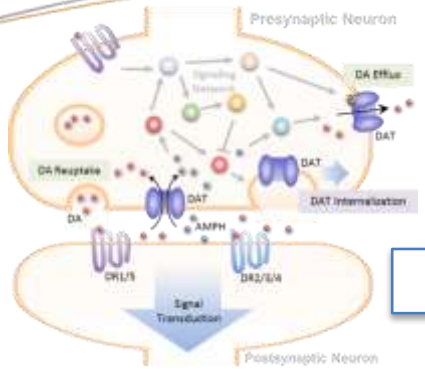


Sensitivity analysis



Predict therapeutic strategies

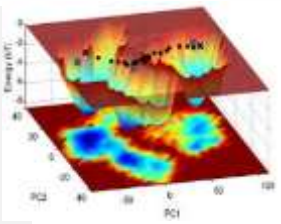




Model Construction

BioNetGen Modeling & Analysis

```
begin reaction rules
L(r) + R(l) <->
L(r!1).R(l!1) kp1, km1
end reaction rules
generate_network()
simulate({method=>"ode",t_e
nd=>500,n_steps=>500})
```

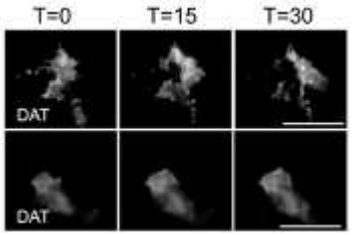


Energy Landscape

Kinetic Modeling

Biological System & Data

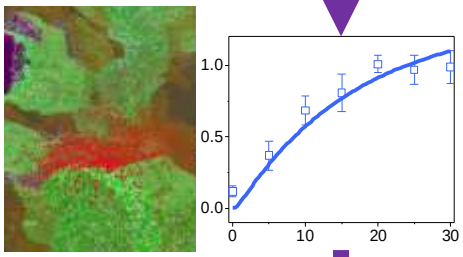
Parameter Estimation & Model Validation

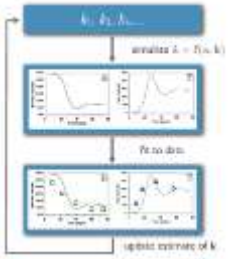


MCell Simulations

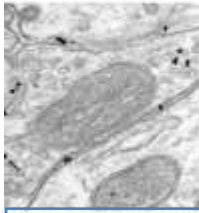
Image analysis

Model Checking Techniques



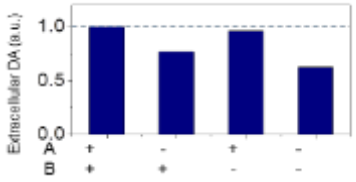


Model Analysis

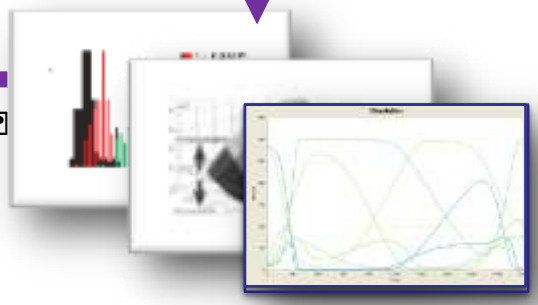


EM Data

Model Refinement



New Insights & Hypotheses



Discovery of new therapeutic strategy

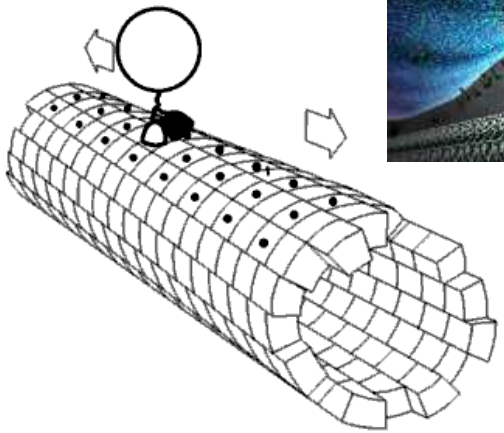
Acknowledgements

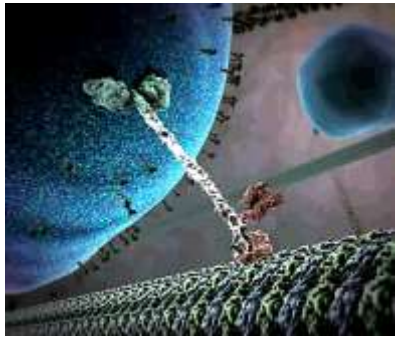
University of Pittsburgh
Ivet Bahar's Lab
James R. Faeder's Lab

Carnegie Mellon University
Edmund M. Clarke's Lab

National University of Singapore
P.S. Thiagarajan's Lab

Questions?





ODE Dynamics

We assume that :

$x_i(t) \in [L_i, U_i]$, where $0 < L_i < U_i$

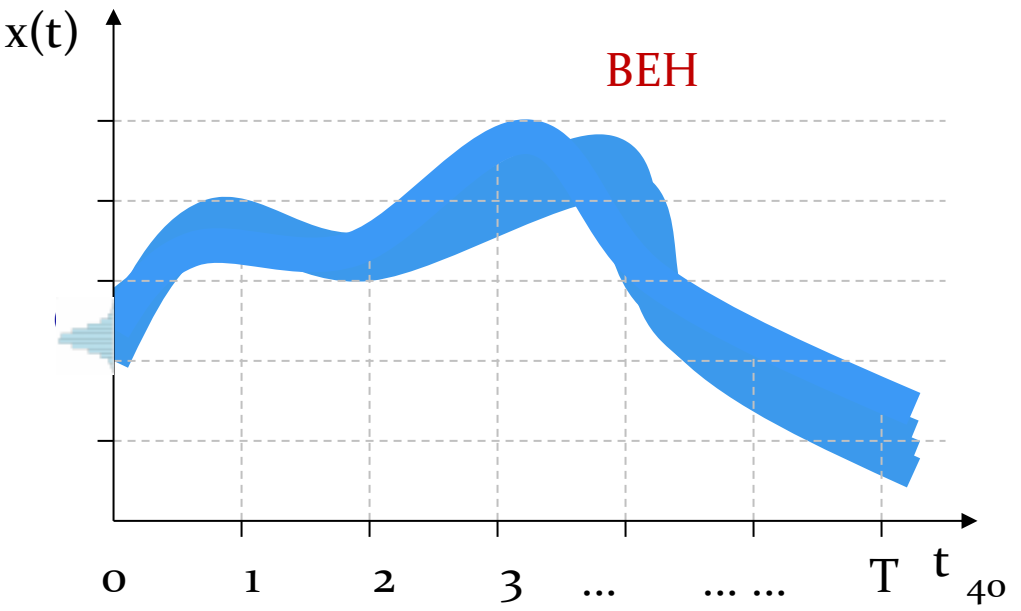
State space : $\mathbf{V} = [L_1, U_1] \times [L_2, U_2] \times ... [L_n, U_n] \subseteq \mathbf{R}_+^n$

$INIT = [L_1^{init}, U_1^{init}] \times [L_2^{init}, L_2^{init}] \times ... [L_n^{init}, L_n^{init}]$

$f_i \in C^1$ for each i, hence $F : \mathbf{V} \to \mathbf{V} \in C^1$

As a result, for each $\mathbf{v} \in \mathbf{V}$, the ODE system will have a unique solution $\mathbf{X}_\mathbf{v}(t)$

$$\frac{d\mathbf{x}}{dt} = F(\mathbf{x}), \mathbf{x} = \{x_1, x_2, ..., x_n\}, \text{and } F(\mathbf{x}(i)) = f_i$$



We define :

flow $\Phi : \mathbf{R}_+ \times \mathbf{V} \to \mathbf{V}$ for arbitrary initial vector $\mathbf{v}$

$\Phi(t, \mathbf{v}) = \mathbf{X}_\mathbf{v}(t)$

Then $\Phi(t, \cdot) \in C^0$

Fixing a maximal time point T, we define a *trajectory* strating from $\mathbf{v} \in \mathbf{V}$ denoted $\sigma_\mathbf{v}$

$\sigma_\mathbf{v} : [0, T] \to \mathbf{V}, \sigma_\mathbf{v}(t) = \Phi(t, \mathbf{v})$

Behavior of our dynamical system is the set of trajectories :

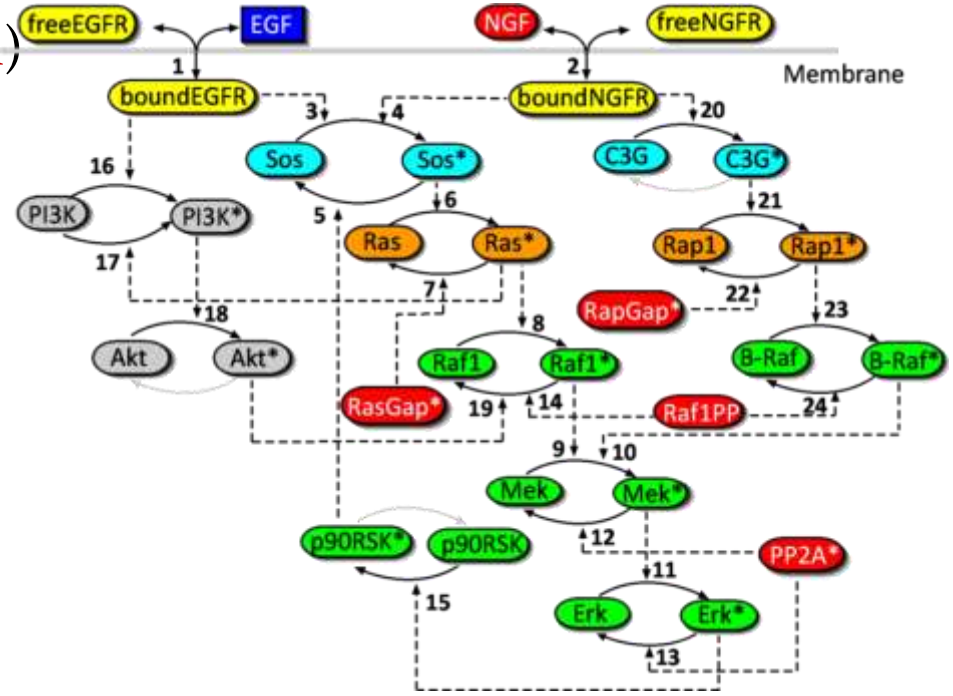
$BEH = \{\sigma_\mathbf{v} \mid \mathbf{v} \in INIT\}$

Case Studies

- Pathway models taken from BioModels database
- Nominal parameters
- Synthetic experimental data
- Qualitative trend

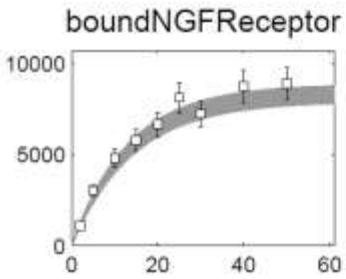
EGF-NGF Pathway

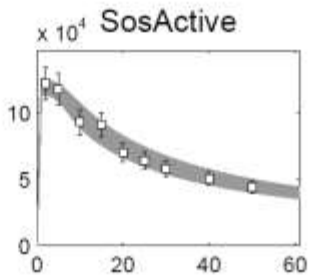
- ODE model (*Brown et al. 2004*)
  - 32 species
  - 48 parameters (20 unknown)
- Training data
  - 7 species, 9 time points
- Test data
  - 2 species, 9 time points

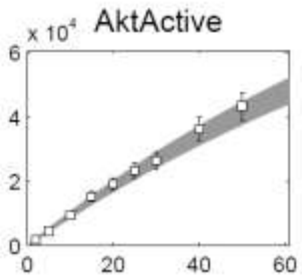


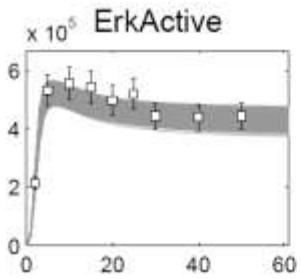
EGF-NGF Pathway

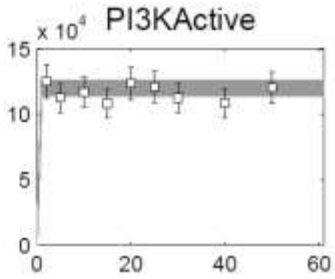
- Running time: 2.23 hours









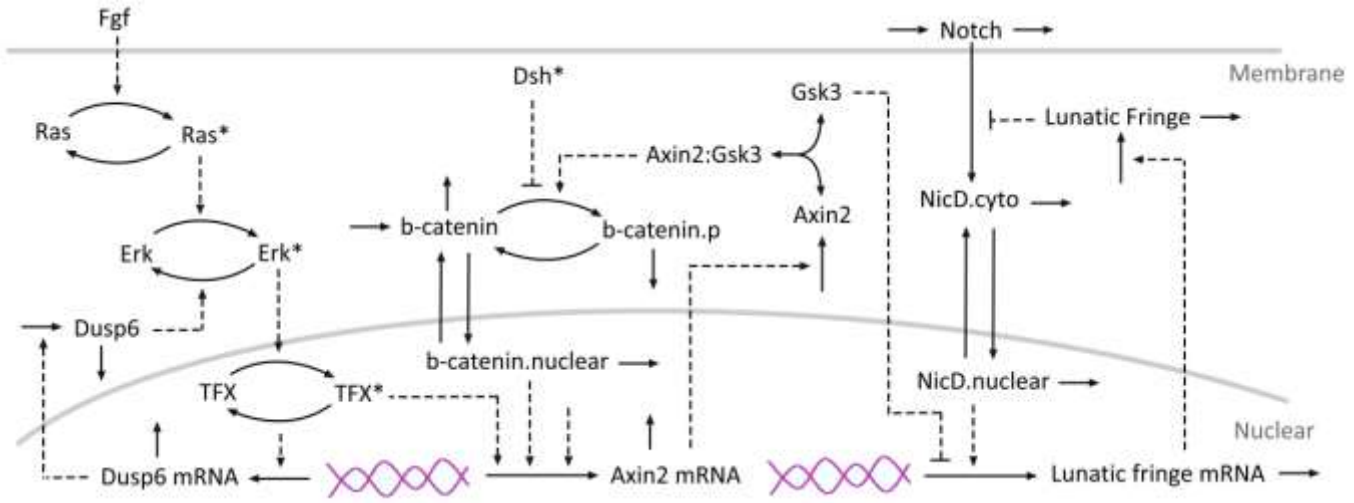


Training data

Test data

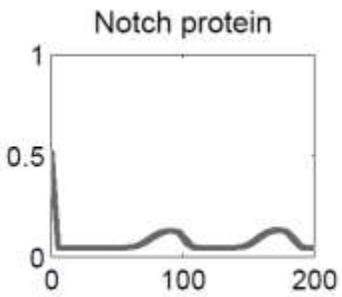
Segmentation Clock Network

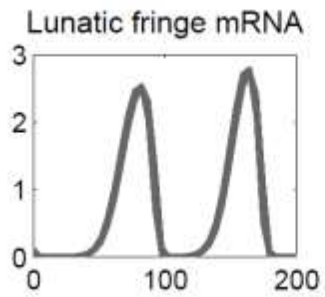
- ODE model (*Goldbeter et al. 2008*)
  - 22 species, 75 parameters (40 unknown)
- Training data
  - Time serials: Axin2 mRNA, 14 time points
  - Qualitative trend: 5 species, oscillatory behavior
    - E.g. $(([LmRNA \leq 0.4] \wedge (\mathbf{F}([LmRNA \geq 2.2] \wedge \mathbf{F}([LmRNA \leq 0.4] \wedge (\mathbf{F}([LmRNA \geq 2.2] \wedge \mathbf{F}([LmRNA \leq 0.4]))))))$
- Test data: Dusp6 protein, qualitative trend

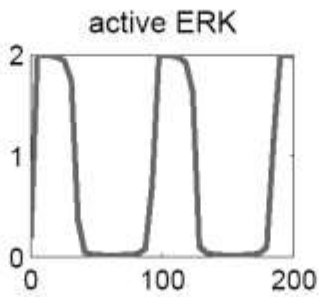


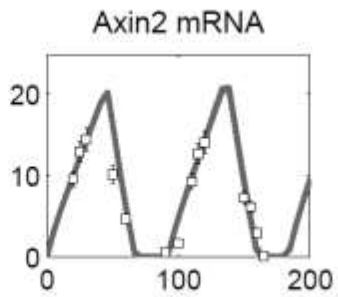
Segmentation Clock Network

- Running time: 2.2 hours

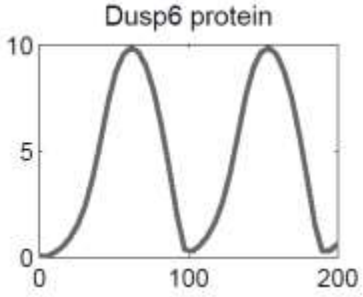








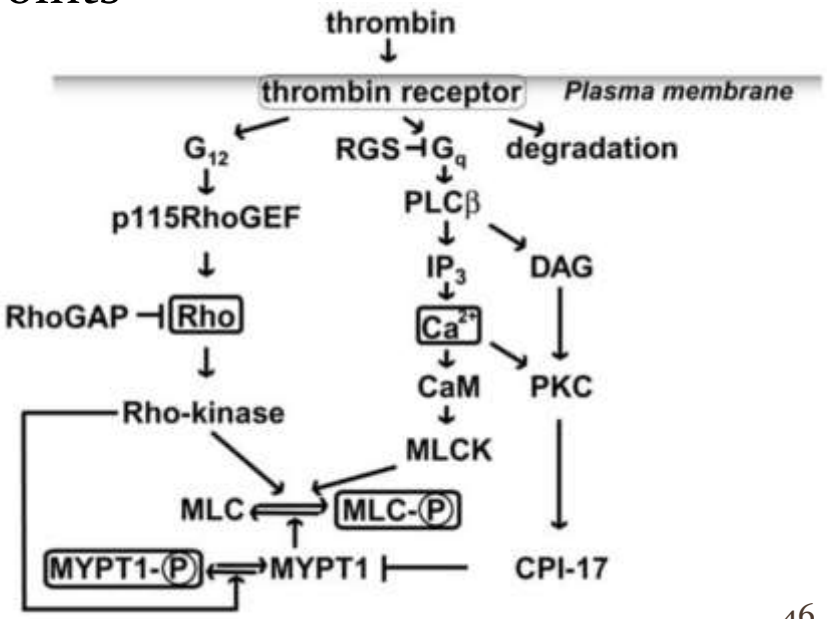
Training data



Test data

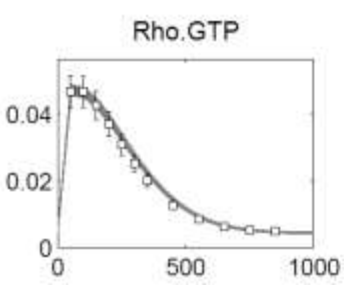
MLC Phosphorylation Pathway

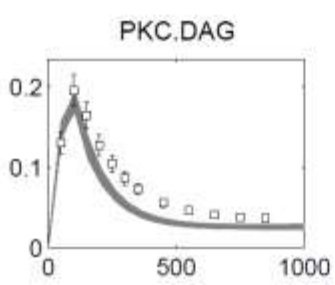
- Regulates the contraction of endothelia cells
- ODE model (*Maeda et al 2006*)
  - **105** species, 197 parameters (**100** unknown parameters)
- Training data
  - Time serials: 8 species, 12 time points
  - Qualitative trend: 2 species
- Test data
  - 2 species, 12 time points

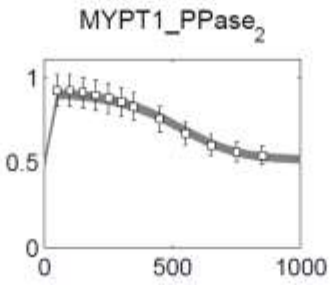


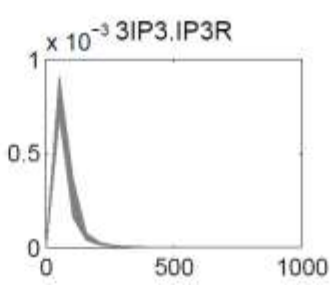
MLC Phosphorylation Pathway

- Running time: 50.67 hours

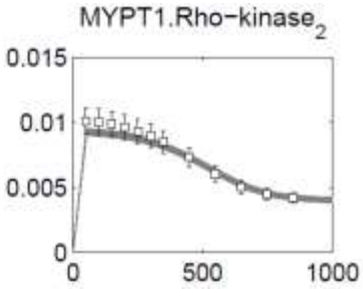








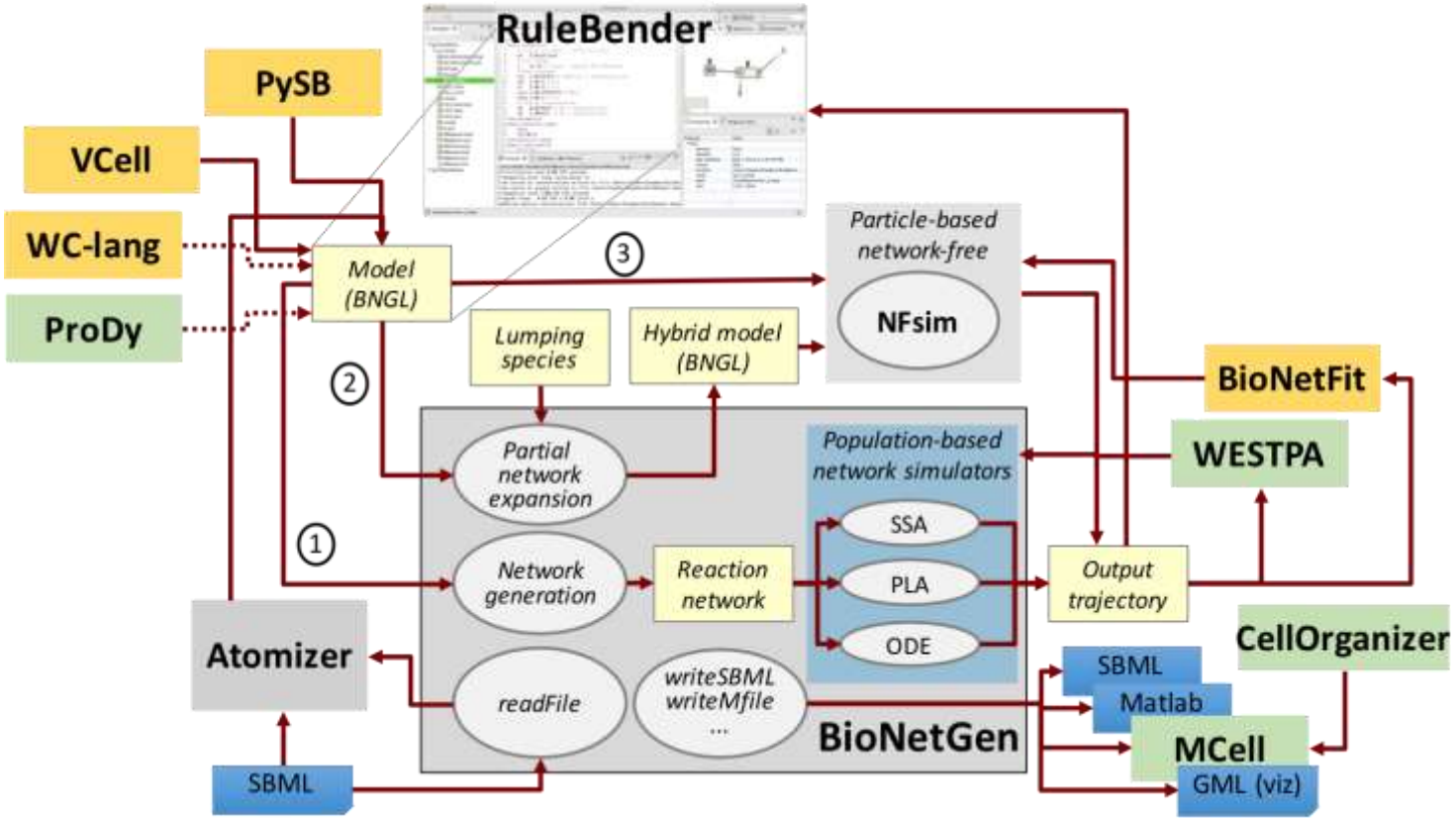
Training data



Test data

Parameter Estimation for BioNetGen

- Current solutions: ptempest, BioNetFit, SBML tools



JR Faeder, unpublished

Our MC-based Techniques

Core Technology

System Representation
- DBN (***Bioinformatic, 2012***)
- ODEs (***CMSB'13***)
- Stochastic models (***Sci Rep, 2014***)
- Hybrid automata (***CMSB'14, HSCC'15, HSB'15***)
- Boolean network (***CMSB'16***)
- Rule-based models (***BIBM'16***)

Model Checking
- Statistical model checking
- Probabilistic model checking
- δ-decision model checking
- Symbolic model checking

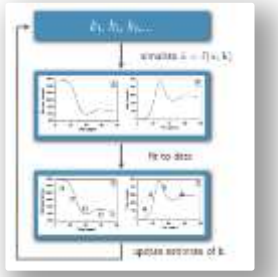
SAT

UNSAT

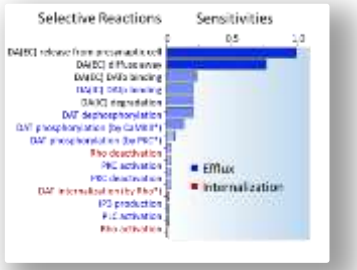
Temporal Property
- Bounded Linear Temporal Logic

Analysis Methods

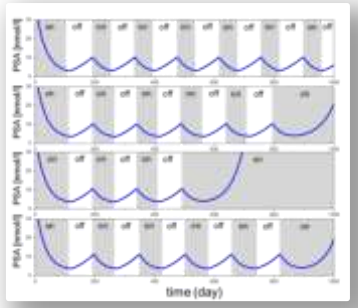
Parameter estimation



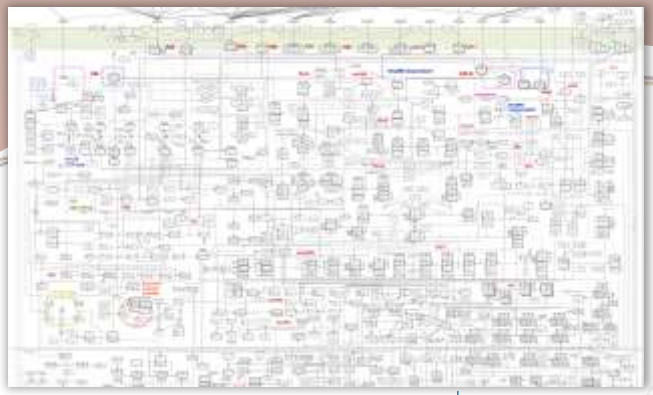
Sensitivity analysis



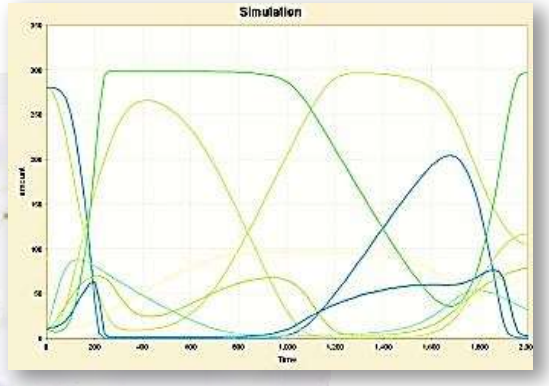
Predict therapeutic strategies



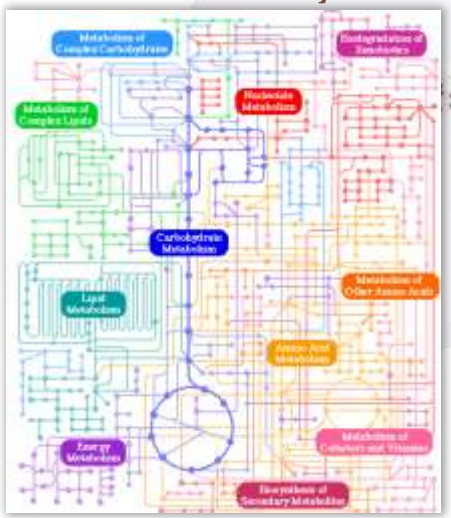
Biological Networks



Signaling Pathways
- Cell death
- Cell differentiation
- Cell proliferation
- Cell migration
- … …

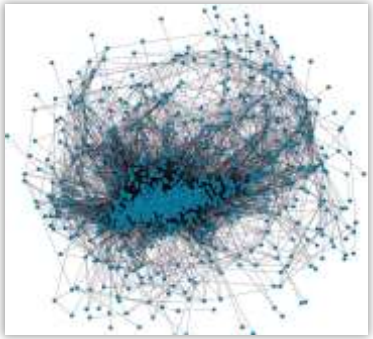


Metabolic Pathways



Macrophage

Gene Regulatory Network



Model Parameters

- Two types of model parameters
  - Initial conditions
  - Rate constants
- Experimental measurements
  - Expensive
  - Not possible to measure all parameters
  - *In vitro* measurements may not reflect the actual physiological conditions in the cell (*Minton, J Biol Chem, 2001*)
  - Cell population-based measurements are not very accurate (*Kim & Price, Phys Rev Lett, 2010*)

On-the-fly Model Checking

- Model checking and generation of trace are coupled i.e. simulate as much as you need.
- Algorithm
  - At each time point we maintain the minimum subset of formulas that need to be true at the state.
  - Based on the simulation, we check the validity of the elements in this set to verify the property
  - Simulation is stopped once the formula has been asserted true/false by the model checking algorithm.
  - We repeat the process of generating simulations and verification until we run enough simulations to satisfy the Wald's statistical test.