Computational Challenges in Large-Scale Pathway Modeling

Frank Tobin

Scientific Computing and Mathematical Modeling

GlaxoSmithKline

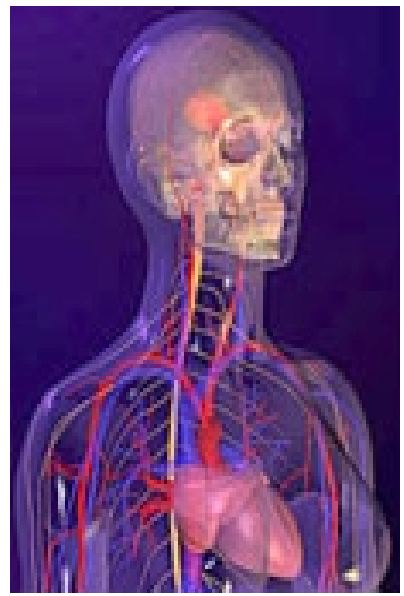
September 22, 2004

Agenda

- **Biological pathways**
  - simple example of a pathway
  - simple example of pharmaceutical interest
- **Building a mathematical model of biological networks**
- **Computational challenges**

Motivation

- **Build as complete a model of as much of a cell or organism as possible**
  - **E. coli is the archetypical prototype**



- **Figure out what to do with it once we get it**

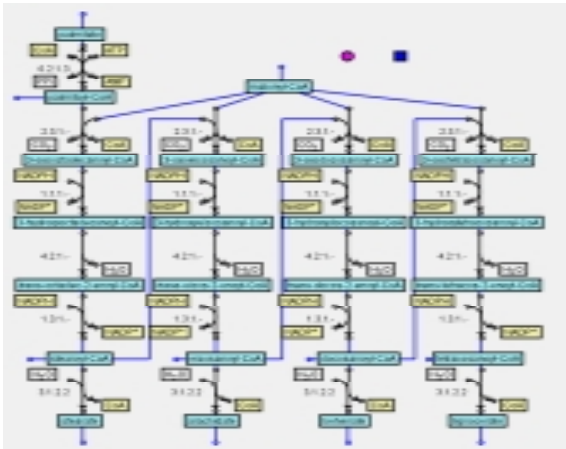
  What if we had a perfect model? Then what?

What is a Pathway?

For the purposes of this talk:

A network of interaction biological entities represented as a directed graph.

So network and pathway are equivalent under this definition.



Saturated Fatty Acid Elongation

Pharmaceutical Interest in Pathways

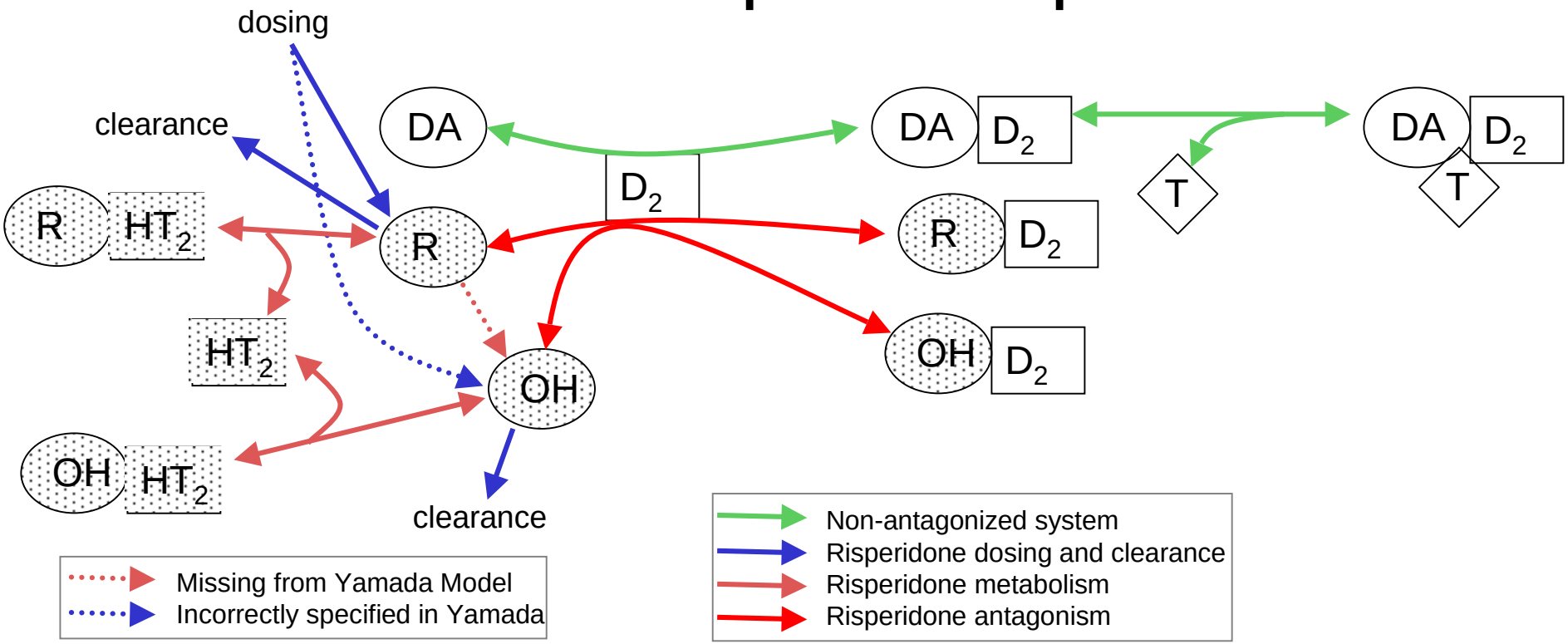
- Predicting culture conditions for overproduction of biopharmaceuticals and drug targets, bioengineering of target assays, enzymes, receptors, etc.
- Understanding compound modes of action
- Identifying novel behaviors and new behaviors of known pathways
  - clues to new intervention approaches
  - selecting and prioritizing of new targets
- Identifying and validating bio-markers
  - animal ⇔ human correlation
- Interpreting and integrating system biology data:
  - transcriptomics, proteomics and metabolomics and other 'omics'

A Simple Pharmaceutical Pathway Example

- Risperidone is a psychotropic agent used for treating schizophrenia or psychosis
- 2.1% of patients develop extrapyramidal symptoms:
  - involuntary movements
  - tremors and rigidity
  - body restlessness
  - muscle contractions
  - changes in breathing and heart rate
- Hypothesis for the extrapyramidal symptoms:

  Dopamine receptor antagonism
  Yamada, et al, Synapse 46, 32-37 (2002)

Mechanism of dopamine receptor inhibition



Receptor Binding:

$DA + D_2 \Leftrightarrow DA \bullet D_2$

Formation of active complex:

$DA \bullet D_2 + T \Leftrightarrow DA \bullet D_2 \bullet T$

Risperidone conversion to 9-hydroxyrisperidone

$R \rightarrow OH$

Binding to D_2 and 5-HT_2 receptors

$R + D_2 \Leftrightarrow R \bullet D_2$

$R + HT_2 \Leftrightarrow R \bullet HT_2$

$OH + D_2 \Leftrightarrow OH \bullet D_2$

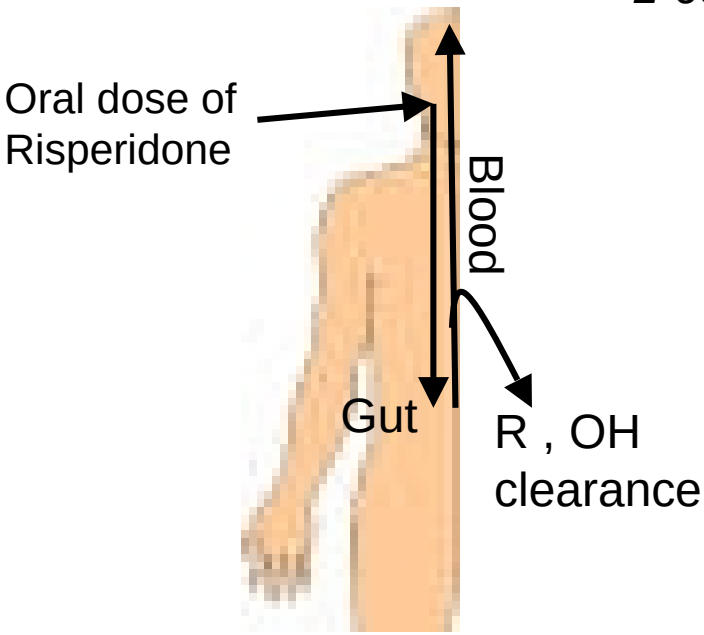
$OH + HT_2 \Leftrightarrow OH \bullet HT_2$

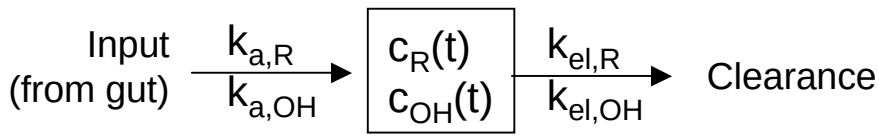
DA: Dopamine
D_2: Receptor
T: Transmitter
R: Risperidone
OH: 9-hydroxyR
HT_2: Receptor

Yamada model for Risperidone PK

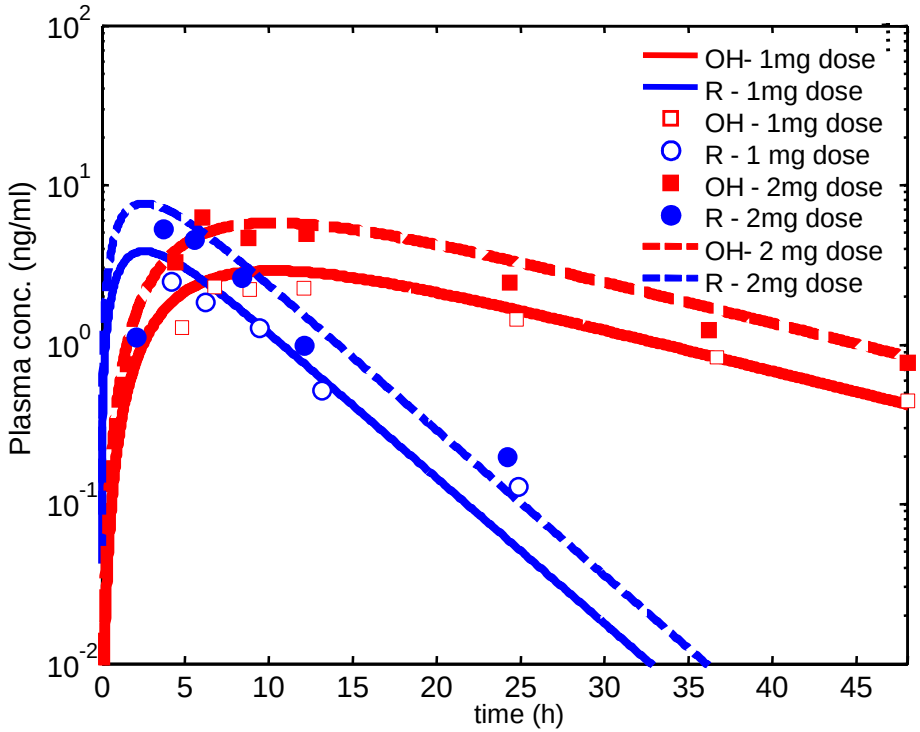
Yamada et al, 2002, Synapse, 46:32-37

1-compartment PK model for Risperidone concentration

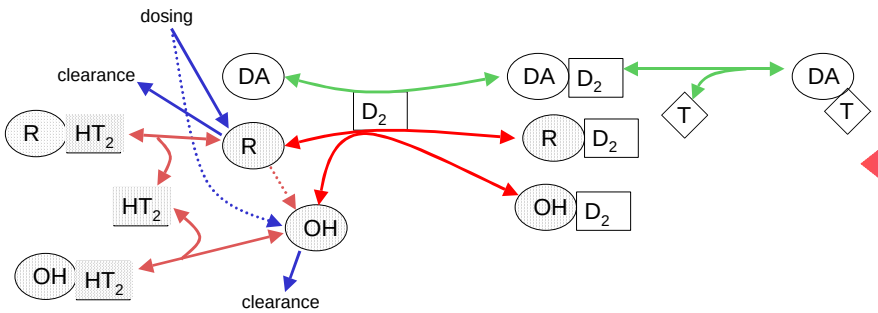




$$c(t) = A(c_0, k_a, k_{el}) \left[\exp(-k_{el}t) - \exp(-k_a t)\right]$$



The ODE Model Approach



Biological model

$$\frac{d[R]^{gut}}{dt} = -k_a^R\,[R]^{gut}$$

$$\frac{d[OH]^{gut}}{dt} = -k_a^{OH}\,[OH]^{gut}$$

$$\frac{d[R]}{dt} = k_a^R\,R^{gut} - k_{el}^R\,[R]$$

$$\frac{d[OH]}{dt} = k_a^{OH}\,[OH]^{gut} - k_{el}^{OH}\,[OH]$$

$$\frac{d[DAgD_2]}{dt} = k_+^{DAgD_2}\,[DA][D_2] - K_A^{DAgD_2} k_+^{DAgD_2}\,[DAgD_2]$$

$$\frac{d[DAgD_2gT]}{dt} = k_+^{DAgD_2gT}\,[DAgD_2][T] - K_A^{DAgD_2gT} k_+^{DAgD_2gT}\,[DAgD_2gT]$$

$$\frac{d[RgD_2]}{dt} = k_+^{RgD_2}\,\beta_R[R][D_2] - K_A^{RgD_2} k_+^{RgD_2}\,[RgD_2]$$

$$\frac{d[OHgD_2]}{dt} = k_+^{OHgD_2}\,\beta_{OH}[OH][D_2] - K_A^{OHgD_2} k_+^{OHgD_2}\,[OHgD_2]$$

$$\frac{d[RgHT_2]}{dt} = k_+^{RgHT_2}\,\beta_R[R][HT_2] - K_A^{RgHT_2} k_+^{RgHT_2}\,[RgHT_2]$$

$$\frac{d[OHgHT_2]}{dt} = k_+^{OHgHT_2}\,\beta_{OH}[OH][HT_2] - K_A^{OHgHT_2} k_+^{OHgHT_2}\,[OHgHT_2]$$

$$[DA]_{total} = [DA] + [DAgD_2] + [DAgD_2gT]$$

$$[T]_{total} = [T] + [DAgD_2gT]$$

$$[D_2]_{total} = [D_2] + [DAgD_2] + [DAgD_2gT] + [RgD_2] + [OHgD_2]$$

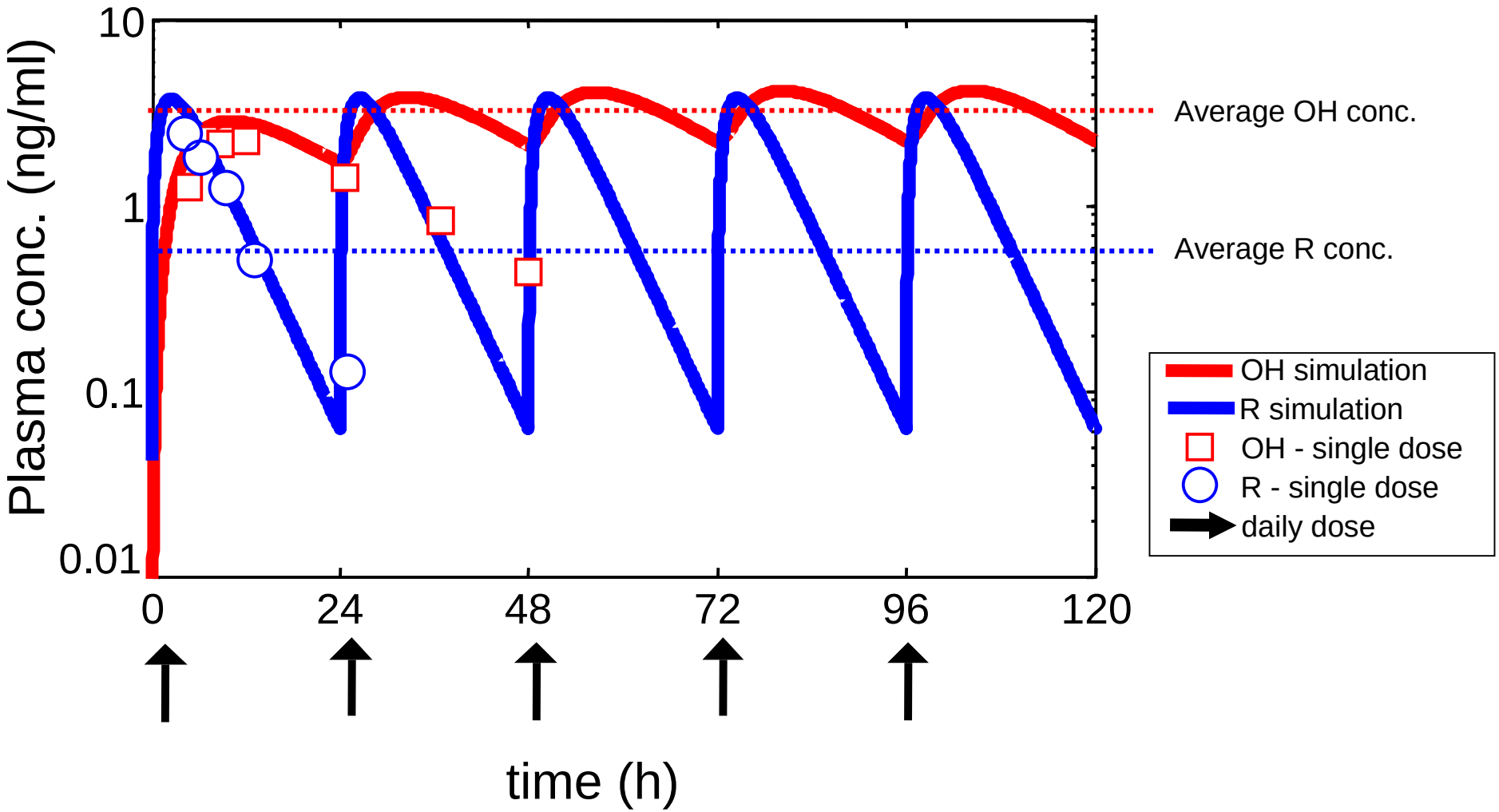
$$[HT_2]_{total} = [HT_2] + [RgHT_2] + [OHgHT_2]$$

Mathematical model
Numerical simulations

$$\mathbf{x}' = \mathbf{f}(\mathbf{x}, \lambda) + D(t)$$

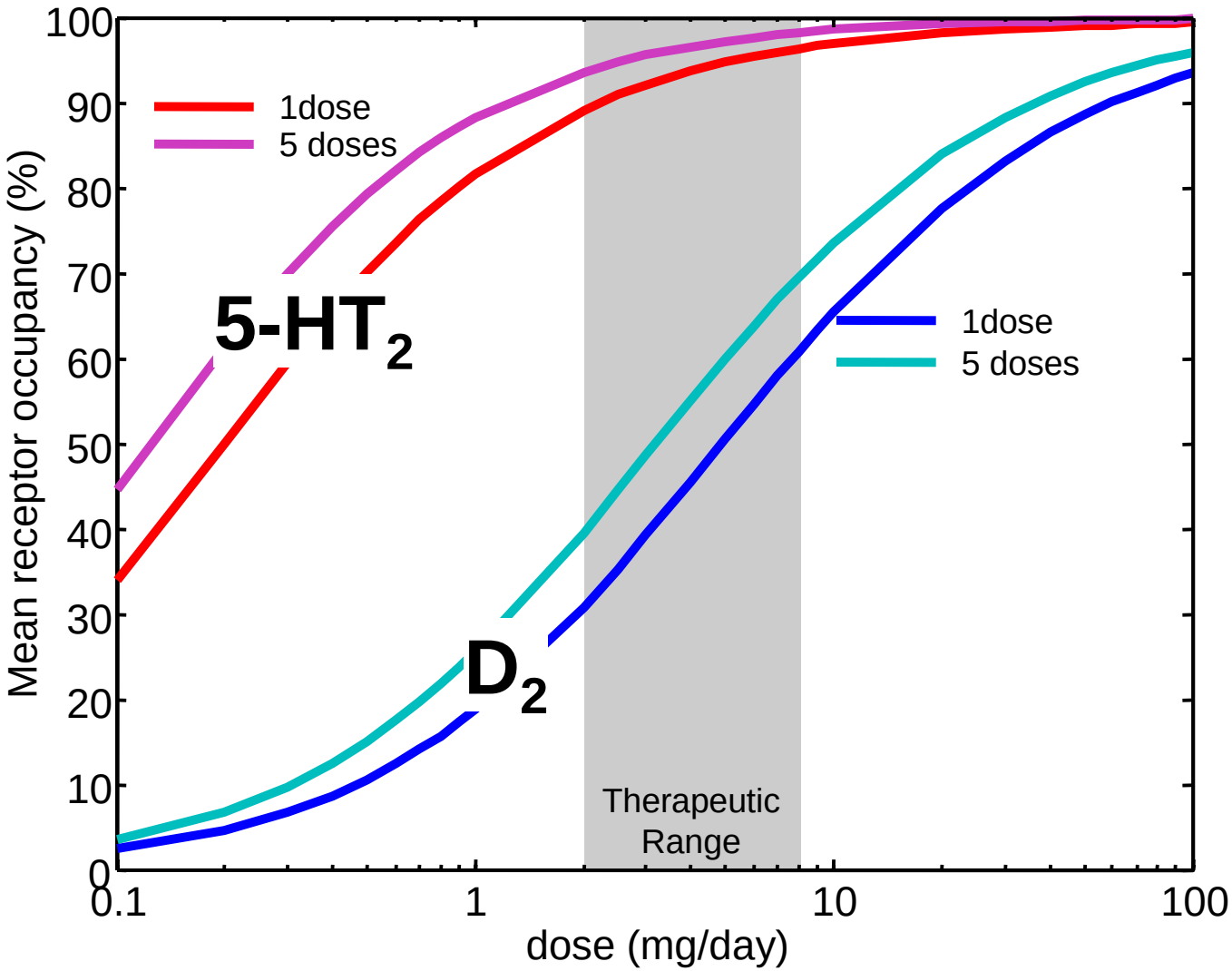
ODEs

Daily Dosing Differs from a Single Dose
Plasma Concentration

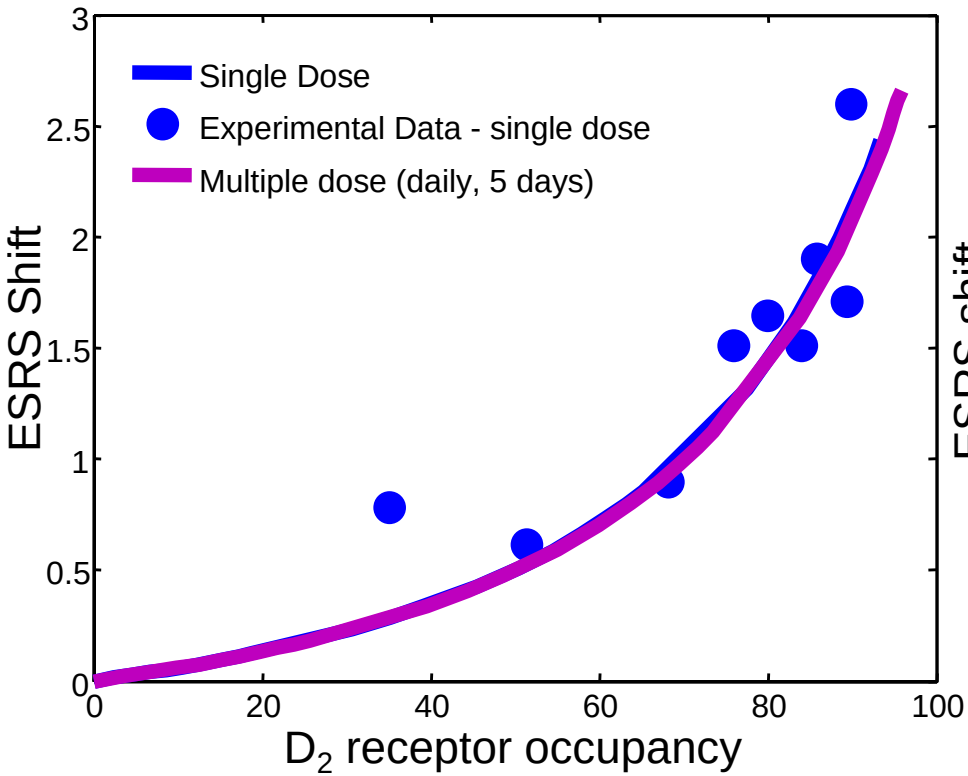


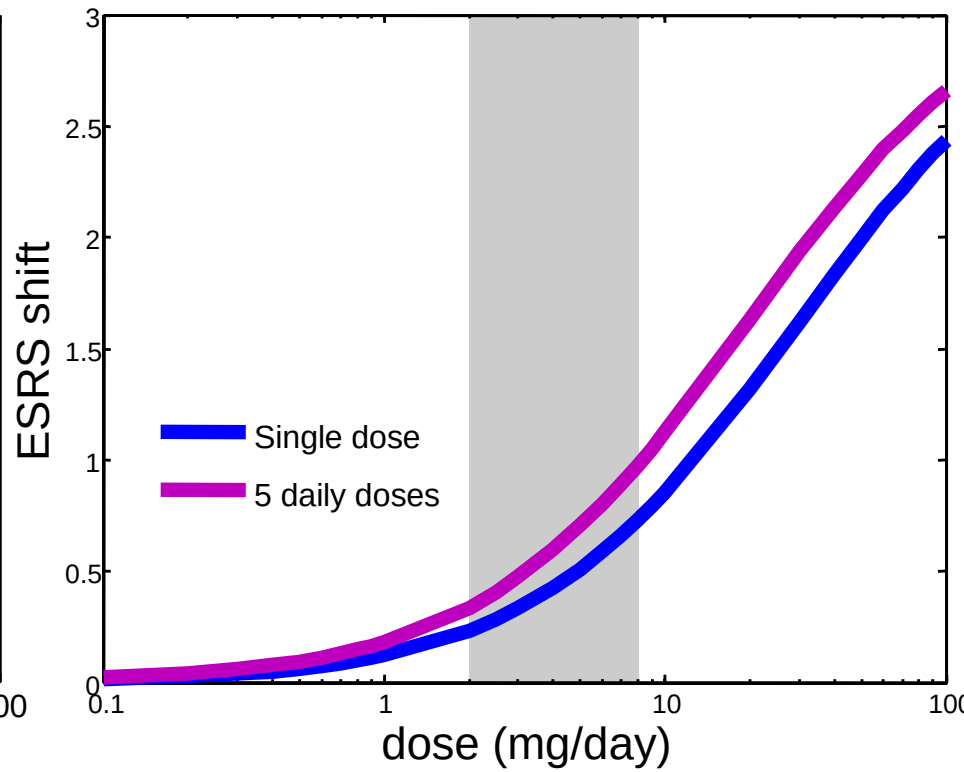
R, OH exptl data from Ishigooka et al., Clin Eval 19, 93-163 (1991)

Effect of multiple dosing on receptor occupancy



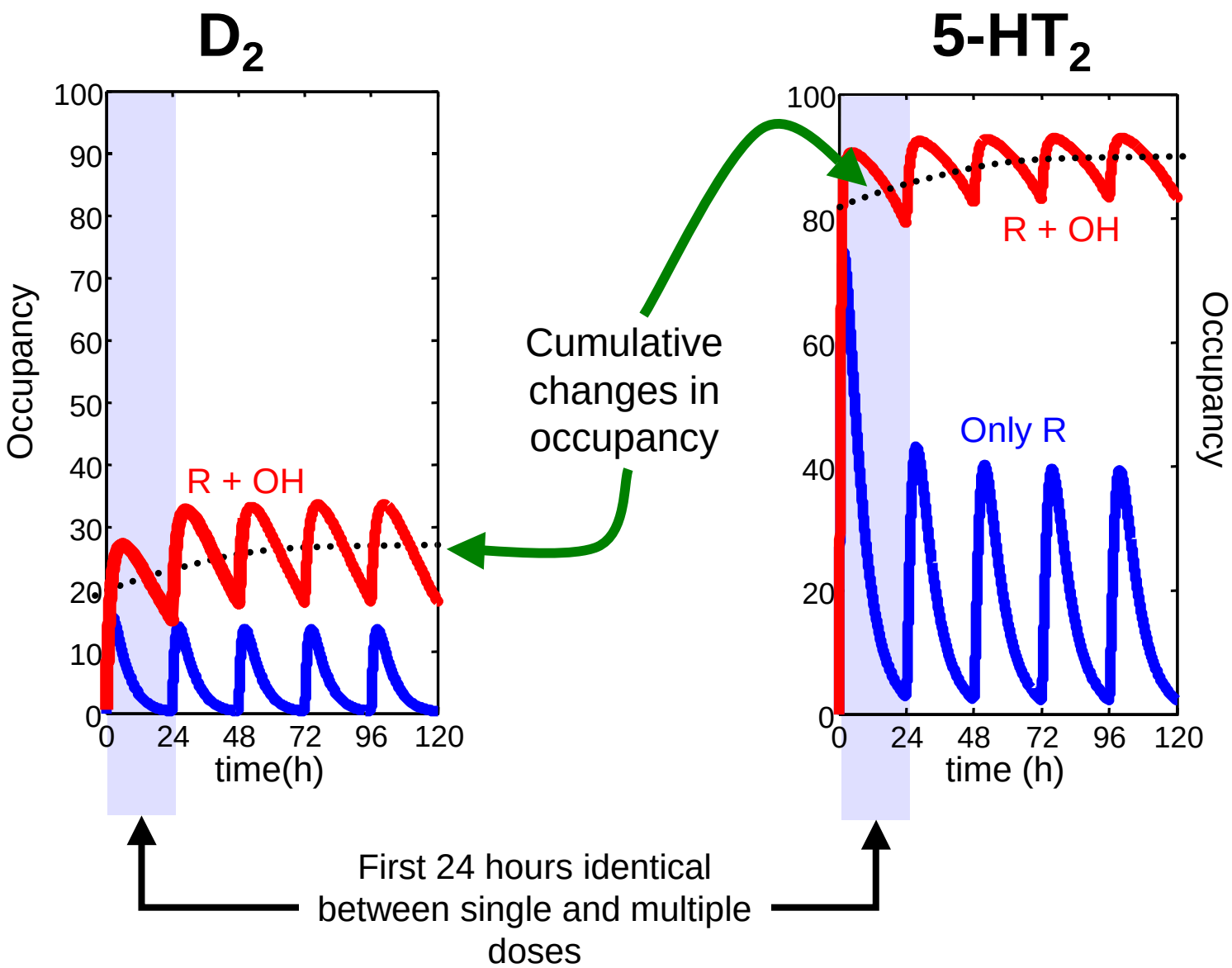
Daily dosing causes differences in predicted side-effects



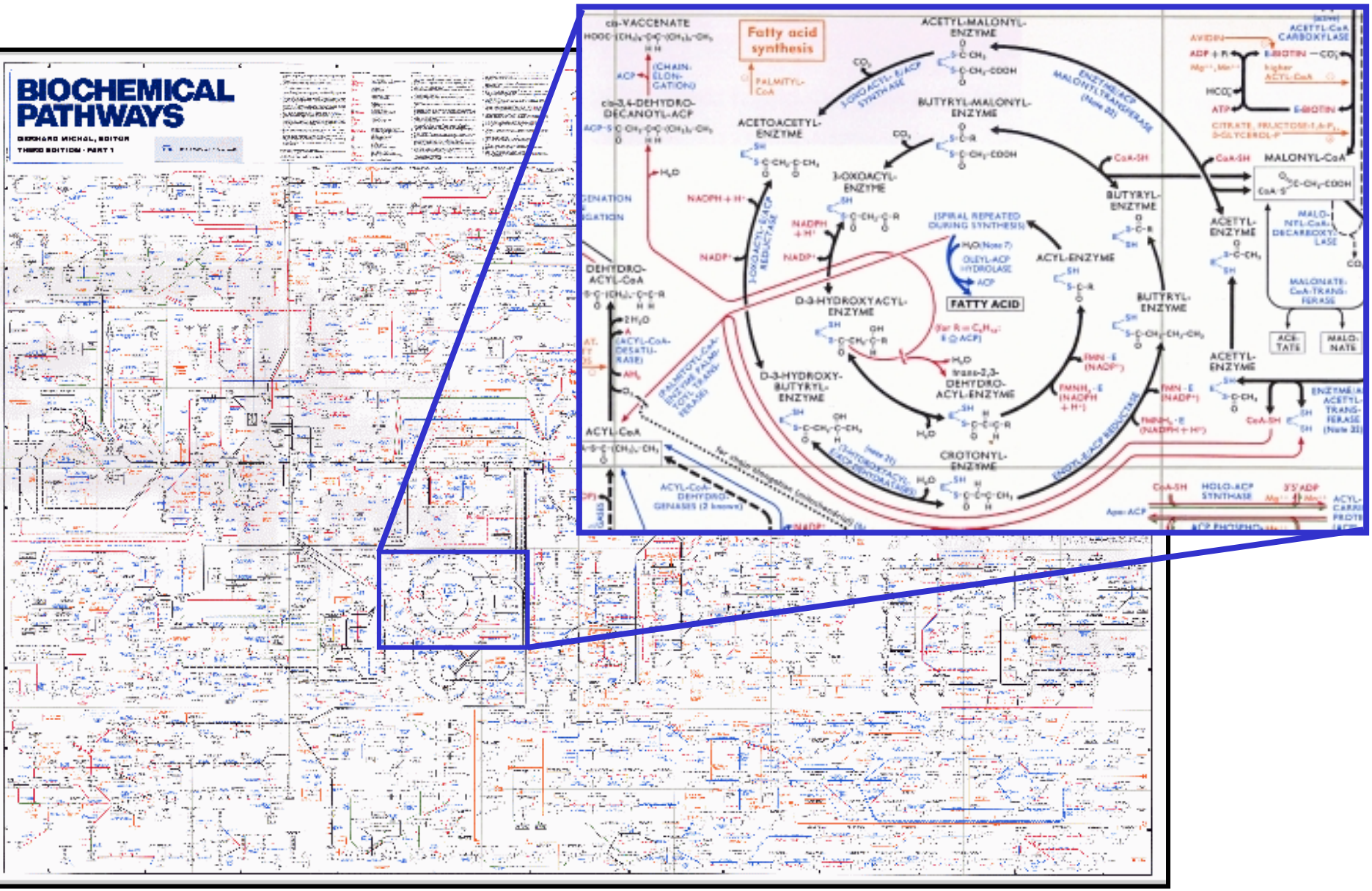


Multiple dosing results in increased ESRS shift, increasing with daily dose administered

Receptor Occupancy as a function of cumulative dosing

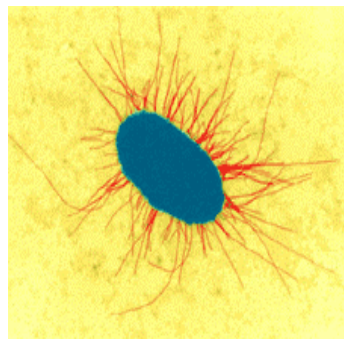


Real Pathways are More Complex



Mathematical Complexity

- **Consider a small, relatively unsophisticated bacterium: Escherichia coli**
  - **≈ 2000 genes**
  - **2500 proteins**
  - **at least several hundred small molecules**
  - **3 interactions per entity X 5000 entities**
  - **3 parameters per equation**
  - **≈ 15 000 equations with 45 000 parameters!**



$$
\begin{aligned}
X' &= F(X;\lambda) && \text{continuous, discrete, stochastic} \\
0 &= G(X;\lambda) && \text{analytic constraints} \\
0 &= H(X;\lambda) && \textit{non}-\text{analytic constraints}
\end{aligned}
$$

- **Now add on spatial change - 15 000 PDEs!**

The Modeling Process

1A Building the model -- forward problem

- Static
- Kinetic
  - Rate law determination
  - Parameter determination

1B Reconstructing the model -- inverse problem

Building the model

2 Validating the model

- Experimental data comparison
- Plausible biology from analytic analysis/simulation
- Examining and assertions testing results

Getting it right

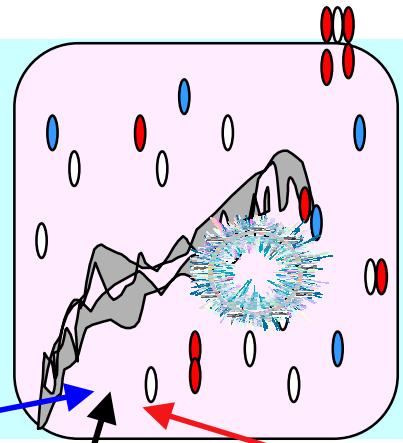
3 Simulation

- Hypothesis testing
- Hypothesis generation

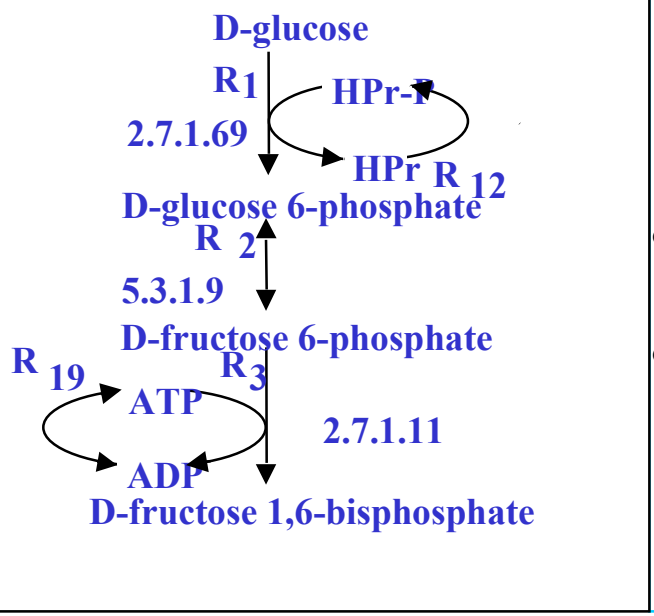
Getting value out of it

Static Model

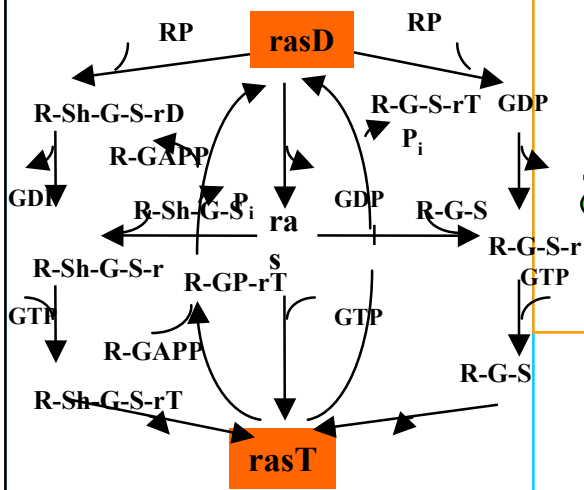
- Only connectivity (topology) of the interactions
- Visualised as connection or interaction graph
- Used for initial model verification and testing
- Types
  - Metabolic
  - Gene Regulation
  - Gene-Product, and Protein-Protein Interactions



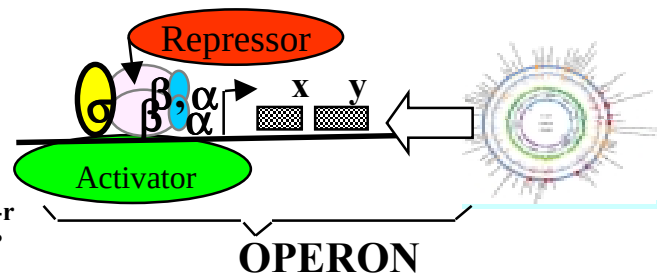
Metabolic network



Gene-Product interactions network



Genetic network

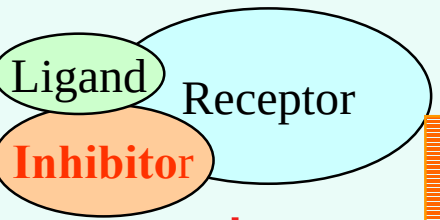


Kinetic Model

- **First phase:** Kinetic models - time dependency incorporated
  - Kinetic behaviour (rate laws) added to static model
  - May or may not obey mass action kinetics
- **Second phase:** Kinetic constants determined from experimental data
- **Third phase** Mathematical model - equations generated
  - Time variation of all concentrations and fluxes can be simulated
  - Model analyses possible: sensitivity, linear stability theory, asymptotic analysis, etc.

Example: Inhibition of a Ligand-Receptor Complex Formation

Static model



+

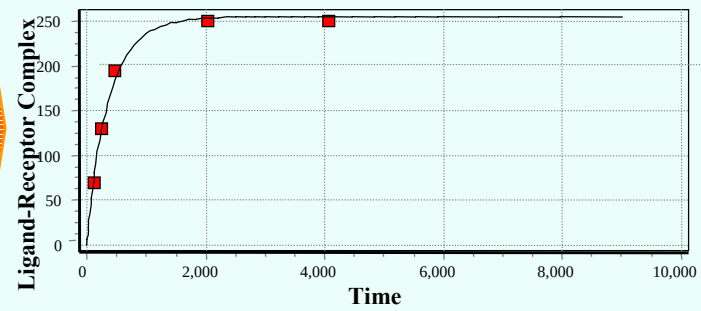
Kinetic Model

$$R + L \Leftrightarrow R \cdot L$$
$$R + I \Leftrightarrow R \cdot I$$

Mathematical Model

$$\begin{aligned}
[R]' &= -k_1[R][L] + k_2[RL] - k_3[R][I] + k_4[RI] \\
[RL]' &= k_1[R][L] - k_2[RL] \\
[RI]' &= k_3[R][I] - k_4[RI] \\
[L]' &= -k_1[R][L] + k_2[RL] \\
[I]' &= -k_3[R][I] + k_4[RI] \\
L_0 &= [L] + [RL] \\
I_0 &= [I] + [RI] \\
R_0 &= [R] + [RL] + [RI]
\end{aligned}$$

Numerical Simulation

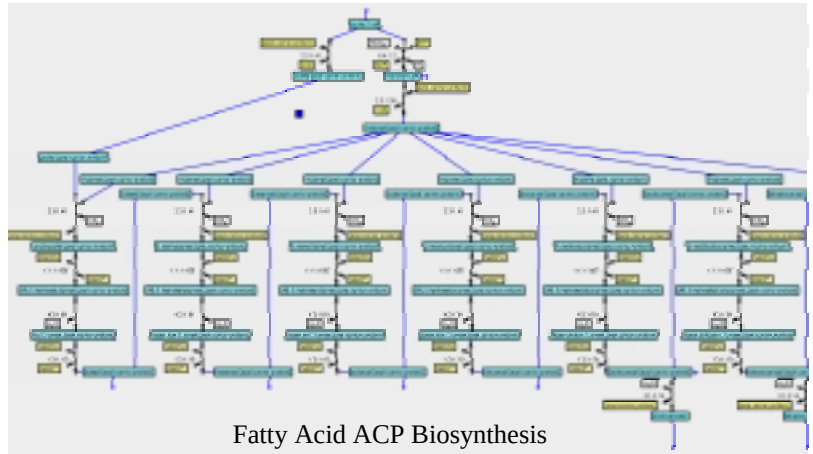


The Resulting System
Very Large, Flawed, and Damned Useful!

- The resulting system of equations:

$$x' = F(x,l)$$
$$0 = G(x,l) \quad \text{algebraic relationships}$$
$$0 = H(x,l) \quad \text{analytic constraints}$$
$$0 = \mathrm{I}(x,l) \quad \text{non-analytic constraints}$$

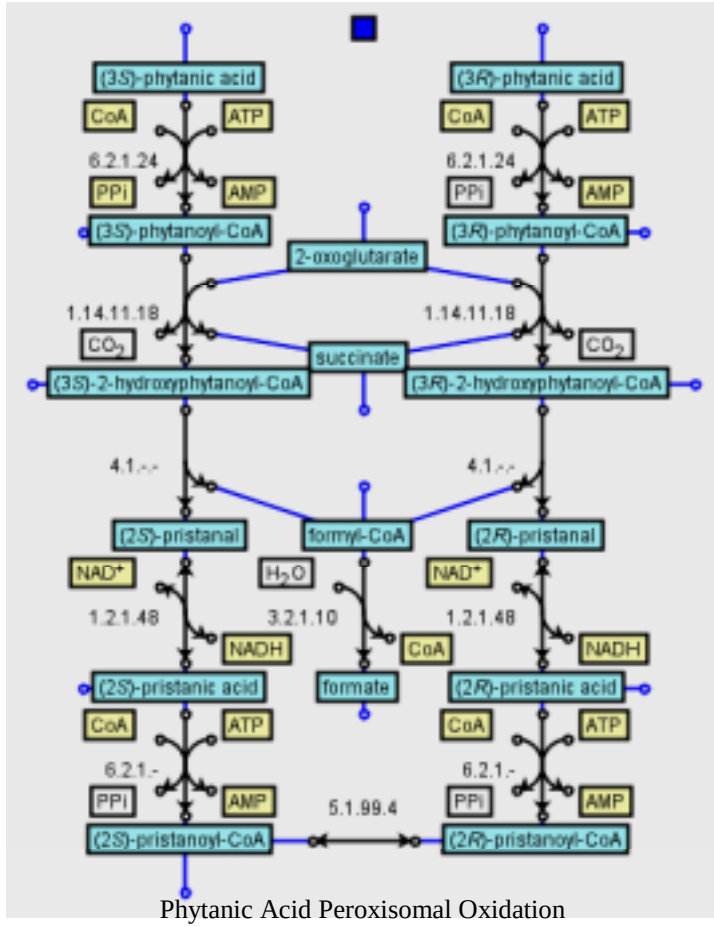
- Very large dimensionalities in:
  - the number of species, X
  - the number of interactions
  - the number of parameters, λ
  - the number of constraint equations


Fatty Acid ACP Biosynthesis

- Uncertainty, error, ambiguity, approximations, etc

As the pathways grow large, the nature of the problems change.

- **Building the model**
  - knowledge management
  - knowledge updating
  - incomplete knowledge
  - Automation
  - Updating the model - versioning
- **Analysis of the model**
  - Too much for a human to peruse
  - Theory gaps
  - Automation
- **Analysis of the simulation results**
  - Too much for a human to peruse
  - New techniques
  - Automation

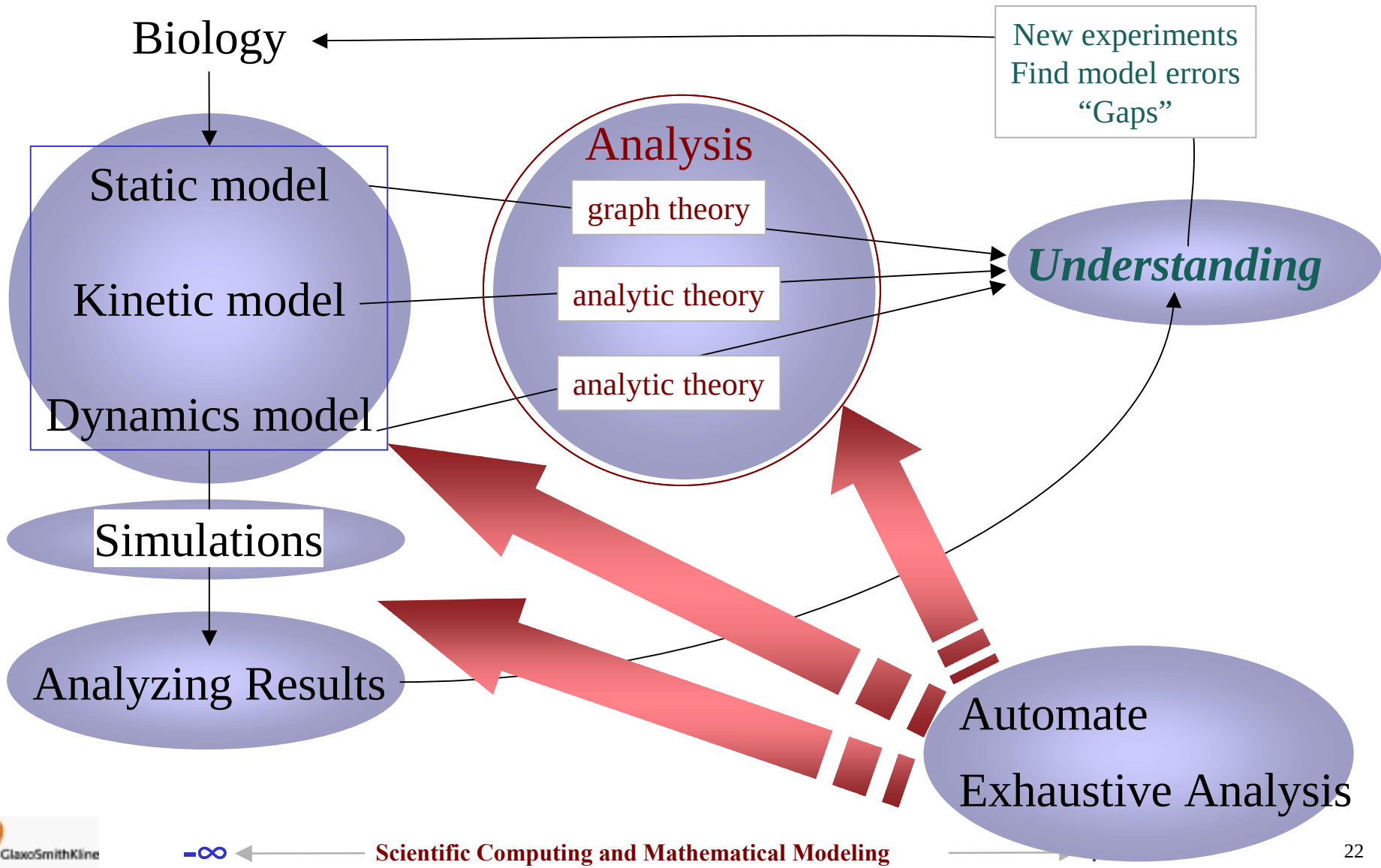


Phytanic Acid Peroxisomal Oxidation

Automation

- **No human intervention whatsoever**
  - **None, nada, zip!**
  - **If it takes a human to setup, run or analyze - its not automated**
- **Robust algorithms**
  - **Graceful failure**
  - **Knowledge of domain of applicability**
  - **Pathological data happens very often - Murphy is omnipresent**
- **Not as easy as it main seem at first**
- **Many existing algorithms are not automatable in current usage**

The Computational Modeling Roadmap



Theory Gap for Large Systems

- Large but not infinite dimensionality is the problem
- Analytical and numerical determination:
  - Finding 'true' null states - there may be a great number
  - Finding linear null states- there may be a great number
  - Asymptotic behaviors
  - Controllability, predictability, integrability, ...
  - Steady state, non-linear behaviors
  - Bifurcation analyses
  - Perturbed behaviors - drug dosing, environment, mutants, etc.
  - ...
- How to calculate in a computationally efficient manner
- Can't afford to calculate everything
- Need to a priori determine which are to be done

Continuousness / Stochasticity/ Discreteness / Ambiguity

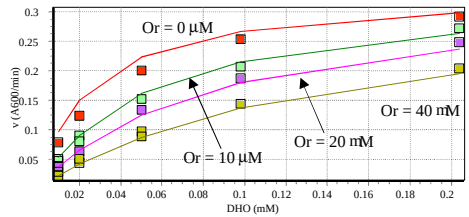
- Continuous approximation breaks down
  - Need to use master equations or some other form of involving stochasticity
  - May need to dynamically switch as system evolves
- Some processes are truly discrete
  - Consider cellular automatons, Petri Nets, discrete events, etc.
- Some parts of the model are only known qualitatively
  - Qualitative simulation techniques.
- Uncertainty and variation in the system
  - Initial conditions
  - Rate constants and rate laws
  - Population variations
  - Interval or fuzzy integration
- Multiscale - time, length, concentration, etc.
- Constraints - DAEs

The challenge: one hybrid integrator

Parameter challenges

- **The larger the model:**
  - **the more parameters compared to the experiments**
- **Static guessing - filling in the gaps**
  - **guessing gene function by analogy**
  - **looking for missing reactions - i.e. enzyme**

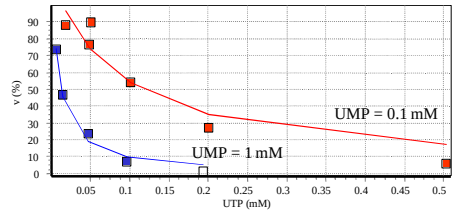
PyrD: DHO + Q = Or + QH_2



- **Kinetic guessing - integrating kinetic islands - guessing plausible rate laws and parameters**
  - **Analogy approaches, similarity across species('multiple alignment')**
  - **From flux analysis?**
- **Do we need to know all parameters? Accuracy?**

Parameter challenges

- **Determine parameters of rate laws from an optimization to fit experimental kinetics data**
  - **noisy and incomplete data**
  - **ill-posed, possibly severely**



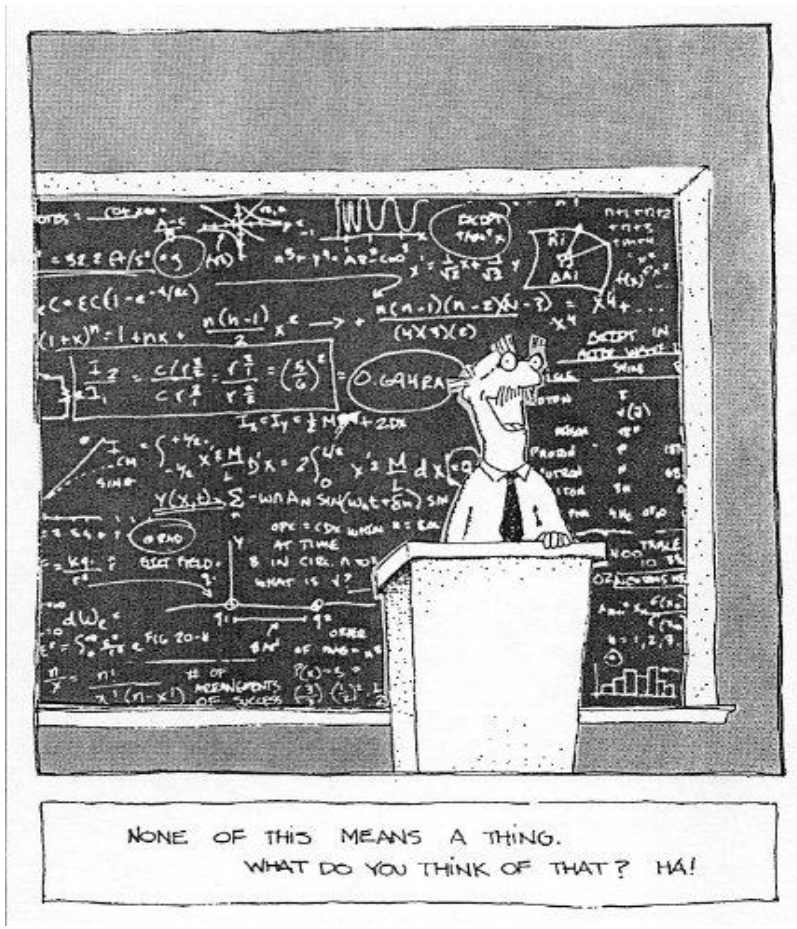
- **How do we scale this up as the model gets bigger?**
  - **One huge model fitting? - Can we even afford this approach?**
  - **One sub-systems at a time fitting?**
  - **Hierarchical fitting? - Stitching together pieces individually calibrated does not a priori mean the model is calibrated**

- **What's the best way to optimize?**
  - **Is L_2 the best objective function?**
  - **Constraints - incorporating and coming up with better ones**

- **How do we know how well we've done?**

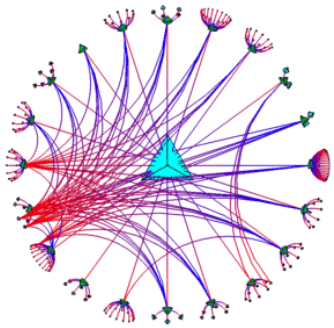
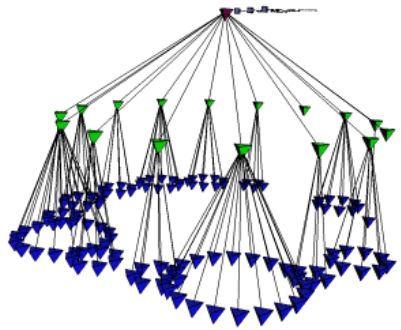
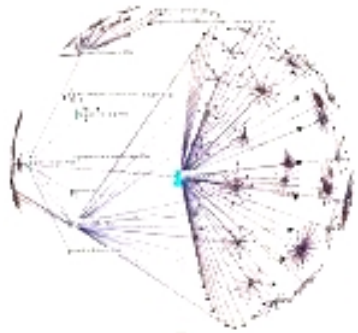
Inverse Problems and Biological Plausibility

- **What makes a model more biological than another?**
  - thermodynamic constraints
  - numerical integrity - semi-definite solutions
  - asymptotic behaviors
  - stability properties
  - information theory constraints
  - physico-chemical constraints
  - environmental constraints
  - evolution constraints
  - flux distributions
  - mass and energy balance
- **Parameter determination needs also**



Visualization Challenges

- **Visualization in a large graph with *too much detail***
  - **Analysis of results - what's interesting?**
  - **Drill down, hyperbolic viewers, database driven for large models**
  - **Visualizing fluxes in a meaningful way**
- **How do you visualize huge networks?**

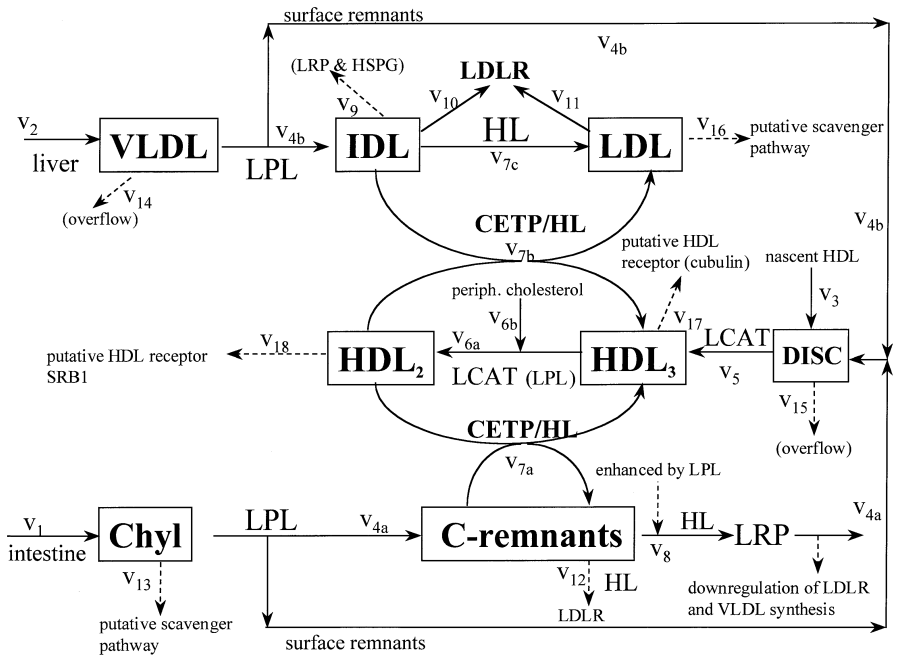




Experiments By T. Munzer, UBC, for visualizing Web connections

- **Tools needed for panning, zooming, drill-down, scalable, incrementally updatable from a database, etc.**
- **Pathway editors for input**
- **Animation - visualizing temporal fluxes**

Discovering "*New*" Biology

Assumption: **if we didn't know anything any biology per se, could we rediscover it from the model?**



Caveat: **if we can find "*old*" biology, then presumably we could find "*new*" biology**

Discovering "*New*" Biology

- **Finding new *cooperative* or *emergent* phenomena:**
  - **pathways and "*distinguishable*" sub-systems**
  - **cycles and "*clocks*"**
  - **oscillatory systems**
  - **regulatory systems**
  - **"states" or "modes" of the system**
- **The resulting biology acts as plausible checks on the model**
- **Some ideas:**
  - **Persistent - pathway behavior *is* or *is not* independent of initial conditions**
  - **Conditional - pathway is active only for certain initial conditions - the nub of course is how to identify this**
  - **Model $\Rightarrow$ graph $\Rightarrow$ matrix $\Rightarrow$ permutation matrix reordering $\Rightarrow$ structure $\Rightarrow$ biology?**
  - **Pattern recognition approaches. Model comparison? Different organisms/species?**
  - **Some type of flux or domain decomposition?**

How do you know they're right? Assertions checking

- Provide a means to formally represent biology that went into the model
  - aspects of computer language parsing, AI-knowledge representation, inference
- Purposes
  - as a formal computer language for incorporation into software
  - for automation of the biology knowledge comparisons against data
  - allow checking model accuracy
  - used as criteria for optimisation - e.g. parameter determination of rate laws
- Consequences of the assertions
  - require certain behaviours to be present in the model
  - expect, but not require some behaviours
  - search for speculative behaviours
  - provide diagnostic tools for examining the quality of the data

Assertions - Bacterial Aerobicity Example

Different genes are expressed under different environments conditions - temperature, media composition, pH, and oxygen. Regulatory systems control expression, but assertions can be used to ensure the basic regulatory processes of the model are accurate.

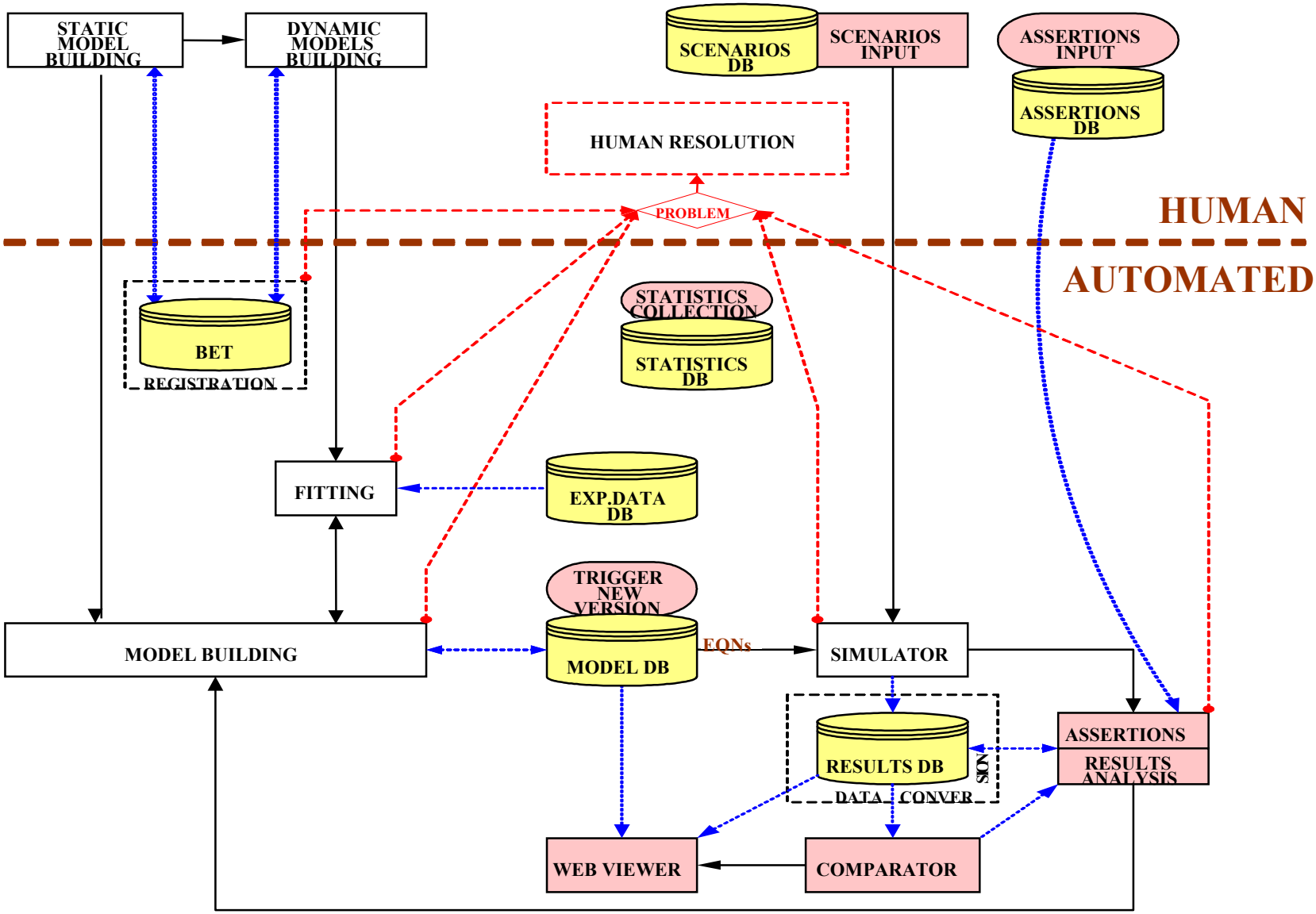
```
# Find the time when the system changes from anaerobic to aerobic behaviour and then
# make sure that the key regulations appear to be happening

Regulation_time =       time > change.time('ANAEROBIC', 'AEROBIC')
                        AND    'ArcA-P' >> 'ArcA'          #positive regulation (activation) of ArcA by ArcA-P
                        OR     'FNR-ox' >> 'FNR-red'       #FNR repressed aerobically

#Then, if regulation appears to be happening, for each protein behaving aerobically:

ForEach aerobic_protein=aerobic(*)                         #look at each aerobic protein, one at a time
   {
    b = flux.value(aerobic_protein);                       #get the flux of each aerobic protein concentration
    c = gene.name(aerobic_protein);                        #time course of expression of the parent gene
    if (regulation_time AND (b > 0))
     {
     Success Action:                                       #if the assertion for this protein is true
          Message ("'AerobicityState' confirmed by the expression profile of gene %s",c)
     Failure Action:                                       #if the assertion for this protein is false
         Message ("Gene %s does not have the expected 'AerobicityState' expression pattern",c)
         Status = WARNING                                  #indicate a non-fatal problem
     }
    }
```

The Pathway Modeling Factory Concept



What Else Is There?
Much, Much More !

Only limited by our imaginations

"Come on, people. We need a creative epiphany right now. Who has one?"

Acknowledgements