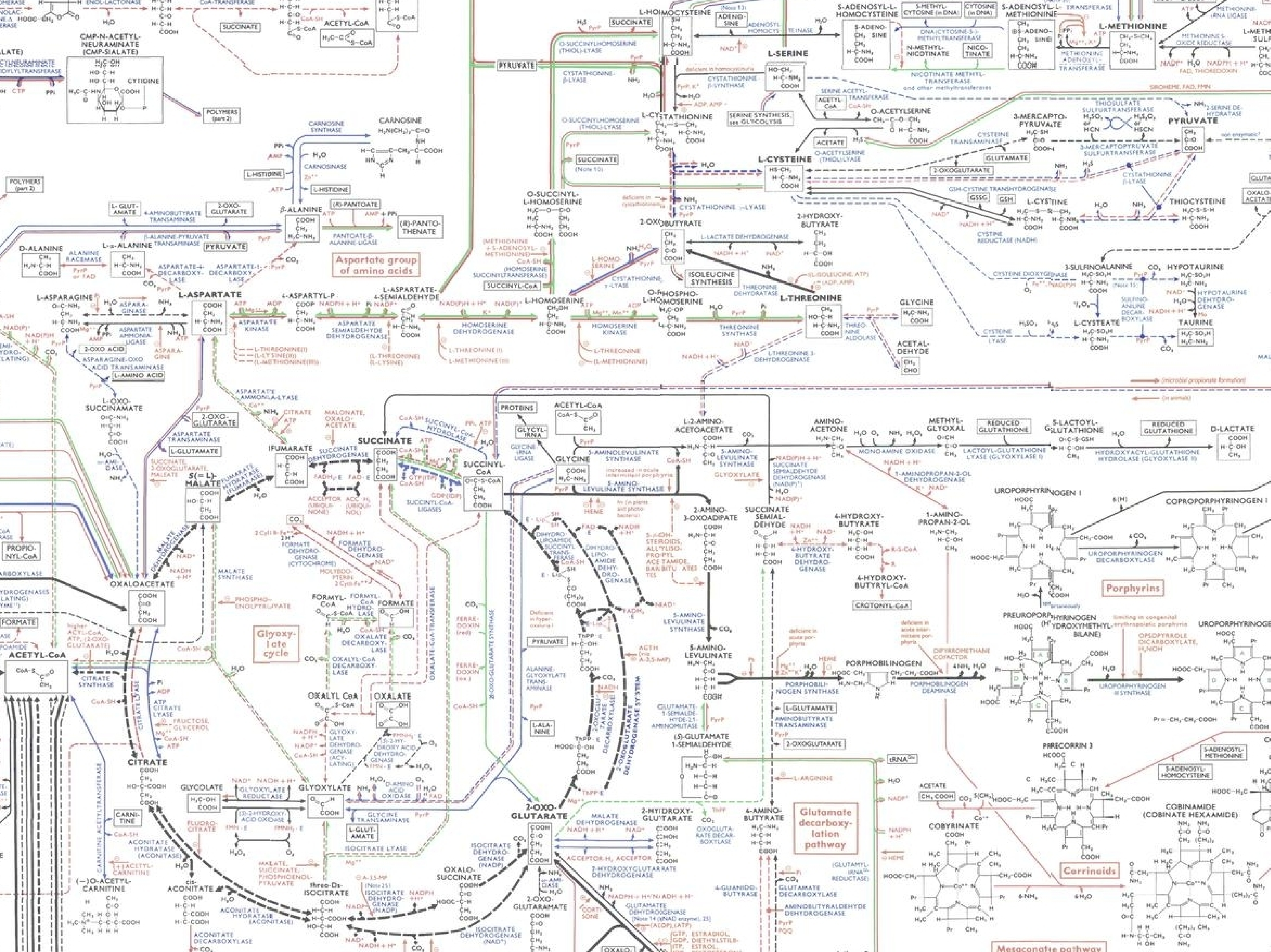
Integration of enzyme kinetic data from various sources

Storage and Annotation of Reaction Kinetics Data

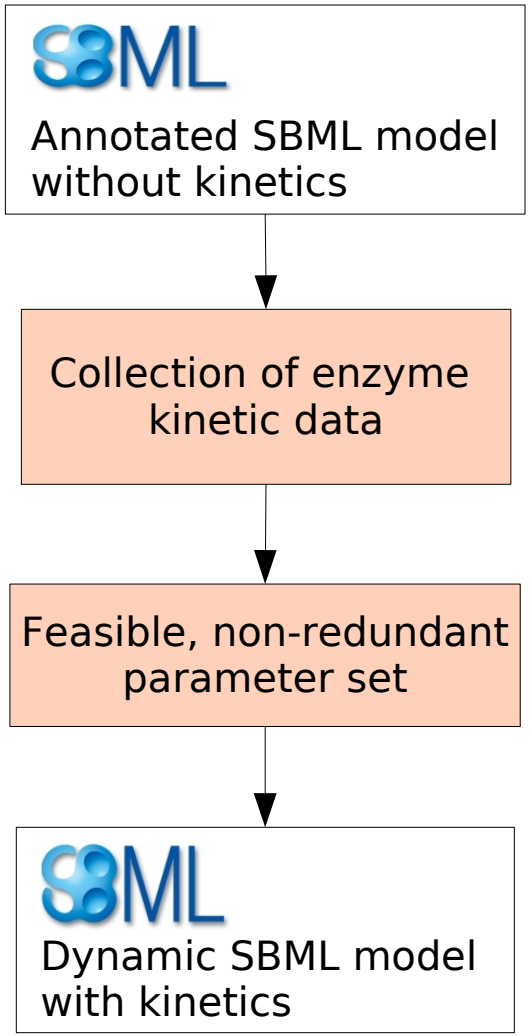
EML, May 21-23, 2007

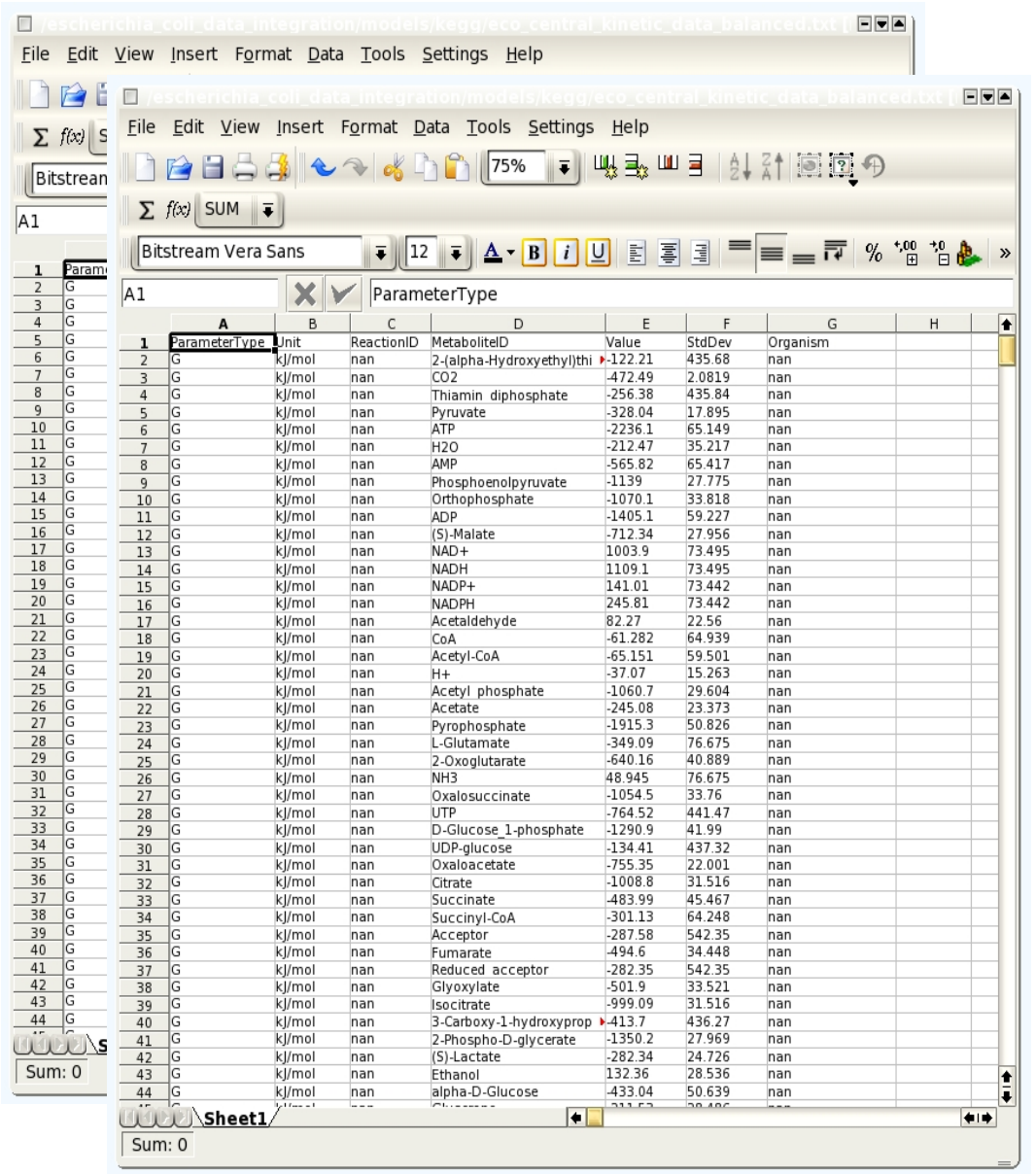
Wolfram Liebermeister, MPI-MG Berlin
Computational systems biology



CMP-N-ACETYL-NEURAMINATE (CMP-SIALATE)
CYTIDINE
POLYMERS (part 2)
CARNOSINE
CARNOSINE SYNTHASE
CARNOSINASE
L-HISTIDINE
β-ALANINE
(R)-PANTOATE
(R)-PANTOTHENATE
PANTOATE-β-ALANINE LIGASE
4-AMINOBUTYRATE TRANSAMINASE
β-ALANINE-PYRUVATE TRANSAMINASE
PYRUVATE
D-ALANINE
ALANINE RACEMASE
L-α-ALANINE
ASPARTATE-4-DECARBOXYLASE
ASPARTATE-1-DECARBOXYLASE
Aspartate group of amino acids
L-ASPARAGINE
ASPARAGINASE
L-ASPARTATE
ASPARTATE AMMONIA-LIGASE
2-OXO ACID
ASPARAGINE-OXO ACID TRANSAMINASE
L-AMINO ACID
L-OXO-SUCCINAMATE
ASPARTATE KINASE
4-ASPARTYL-P
ASPARTATE SEMIALDEHYDE DEHYDROGENASE
L-ASPARTATE-4-SEMIALDEHYDE
HOMOSERINE DEHYDROGENASE
L-HOMOSERINE
HOMOSERINE KINASE
O-PHOSPHO-L-HOMOSERINE
THREONINE SYNTHASE
L-THREONINE
THREONINE ALDOLASE
GLYCINE
ACETALDEHYDE
L-THREONINE 3-DEHYDROGENASE
ISOLEUCINE SYNTHESIS
THREONINE DEHYDRATASE
2-OXOBUTYRATE
L-LACTATE DEHYDROGENASE
2-HYDROXYBUTYRATE
O-SUCCINYL-L-HOMOSERINE
SUCCINYL-CoA
PYRUVATE
L-HOMOCYSTEINE
ADENOSINE
ADENOSYLHOMOCYSTEINASE
SUCCINATE
O-SUCCINYLHOMOSERINE (THIOL)-LYASE
CYSTATHIONINE-β-LYASE
CYSTATHIONINE-β-SYNTHASE
L-CYSTATHIONINE
CYSTATHIONINE γ-LYASE
L-SERINE
SERINE ACETYLTRANSFERASE
O-ACETYLSERINE
O-ACETYLSERINE (THIOL)-LYASE
SERINE SYNTHESIS, see GLYCOLYSIS
S-ADENOSYL-L-HOMOCYSTEINE
S-ADENOSYL-L-METHIONINE
L-METHIONINE
METHIONINE ADENOSYLTRANSFERASE
N-METHYL-NICOTINATE
NICOTINATE
NICOTINATE METHYLTRANSFERASE and other methyltransferases
THIOSULFATE SULFURTRANSFERASE
3-MERCAPTOPYRUVATE
3-MERCAPTOPYRUVATE SULFURTRANSFERASE
CYSTEINE TRANSAMINASE
GLUTAMATE
2-OXOGLUTARATE
L-CYSTEINE
PYRUVATE
GSH-CYSTINE TRANSHYDROGENASE
GSSG
GSH
L-CYSTINE
CYSTINE REDUCTASE (NADH)
THIOCYSTEINE
CYSTATHIONINE β-LYASE
CYSTEINE DIOXYGENASE
3-SULFINOALANINE
HYPOTAURINE
HYPOTAURINE DEHYDROGENASE
L-CYSTEATE
TAURINE
SULFINOALANINE DECARBOXYLASE
CYSTEINE LYASE
(microbial propionate formation)
(in animals)
ASPARTATE AMMONIA-LYASE
2-OXO-GLUTARATE
ASPARTATE TRANSAMINASE
L-GLUTAMATE
FUMARATE
FUMARATE HYDRATASE (FUMARASE)
S(=L)-MALATE
SUCCINATE
SUCCINATE DEHYDROGENASE
SUCCINYL-CoA HYDROLASE
SUCCINYL-CoA
SUCCINYL-CoA LIGASES
PROTEINS
GLYCYL-tRNA
ACETYL-CoA
GLYCINE
5-AMINOLEVULINATE SYNTHASE
L-2-AMINO-ACETOACETATE
5-AMINOLEVULINATE SYNTHASE
AMINOACETONE
MONOAMINE OXIDASE
METHYLGLYOXAL
REDUCED GLUTATHIONE
S-LACTOYL-GLUTATHIONE
LACTOYL-GLUTATHIONE LYASE (GLYOXYLASE I)
HYDROXYACYL-GLUTATHIONE HYDROLASE (GLYOXYLASE II)
D-LACTATE
1-AMINOPROPAN-2-OL DEHYDROGENASE
1-AMINO-PROPAN-2-OL
MALATE DEHYDROGENASE
MALATE SYNTHASE
OXALOACETATE
PHOSPHOENOLPYRUVATE
FORMATE
FORMATE DEHYDROGENASE (CYTOCHROME)
FORMYL-CoA
FORMYL-CoA HYDROLASE
OXALATE DECARBOXYLASE
OXALYL-CoA DECARBOXYLASE
Glyoxylate cycle
OXALATE-CoA-TRANSFERASE
2-OXOGLUTARATE SYNTHASE
DIHYDROLIPOAMIDE SUCCINYLTRANSFERASE
DIHYDROLIPOAMIDE DEHYDROGENASE
2-OXOGLUTARATE DEHYDROGENASE SYSTEM
2-AMINO-3-OXOADIPATE
SUCCINATE SEMIALDEHYDE
4-HYDROXYBUTYRATE
4-HYDROXYBUTYRATE DEHYDROGENASE
4-HYDROXYBUTYRYL-CoA
CROTONYL-CoA
5-AMINOLEVULINATE SYNTHASE
5-AMINOLEVULINATE
PORPHOBILINOGEN SYNTHASE
PORPHOBILINOGEN
PORPHOBILINOGEN DEAMINASE
UROPORPHYRINOGEN I
COPROPORPHYRINOGEN I
UROPORPHYRINOGEN DECARBOXYLASE
Porphyrins
PREUROPORPHYRINOGEN (HYDROXYMETHYLBILANE)
UROPORPHYRINOGEN III SYNTHASE
UROPORPHYRINOGEN III
ACETYL-CoA
CITRATE SYNTHASE
ATP CITRATE LYASE
CITRATE LYASE
CITRATE
ALANINE-GLYOXYLATE TRANSAMINASE
L-ALANINE
OXALYL CoA
OXALATE
GLYOXYLATE DEHYDROGENASE (ACYLATING)
GLYCOLATE
GLYOXYLATE REDUCTASE
GLYOXYLATE
D-AMINO ACID OXIDASE
GLYCINE TRANSAMINASE
L-GLUTAMATE
ISOCITRATE LYASE
(S)-2-HYDROXY-ACID OXIDASE
CARNITINE
CARNITINE ACETYLTRANSFERASE
(−)-O-ACETYL-CARNITINE
ACONITATE HYDRATASE (ACONITASE)
cis-ACONITATE
threo-Ds-ISOCITRATE
ISOCITRATE DEHYDROGENASE (NADP)
OXALOSUCCINATE
ISOCITRATE DEHYDROGENASE (NAD+)
ACONITATE DECARBOXYLASE
2-OXO-GLUTARATE
MALATE DEHYDROGENASE
2-HYDROXY-GLUTARATE
2-HYDROXYGLUTARATE DEHYDROGENASE
2-OXO-GLUTARAMATE
GLUTAMATE DEHYDROGENASE
GLUTAMATE-1-SEMIALDEHYDE-2,1-AMINOMUTASE
(S)-GLUTAMATE 1-SEMIALDEHYDE
L-GLUTAMATE
AMINOBUTYRATE TRANSAMINASE
2-OXOGLUTARATE
OXOGLUTARATE DECARBOXYLASE
4-AMINO-BUTYRATE
4-GUANIDO-BUTYRASE
GLUTAMATE DECARBOXYLASE
AMINOBUTYRALDEHYDE DEHYDROGENASE
Glutamate decarboxylation pathway
(GLUTAMYL-tRNA REDUCTASE)
S-ADENOSYL-METHIONINE
S-ADENOSYL-HOMOCYSTEINE
PRECORRIN 3
COBYRINATE
COBINAMIDE (COBINATE HEXAAMIDE)
Corrinoids
Mesaconate pathway

Scheme for a modelling workflow





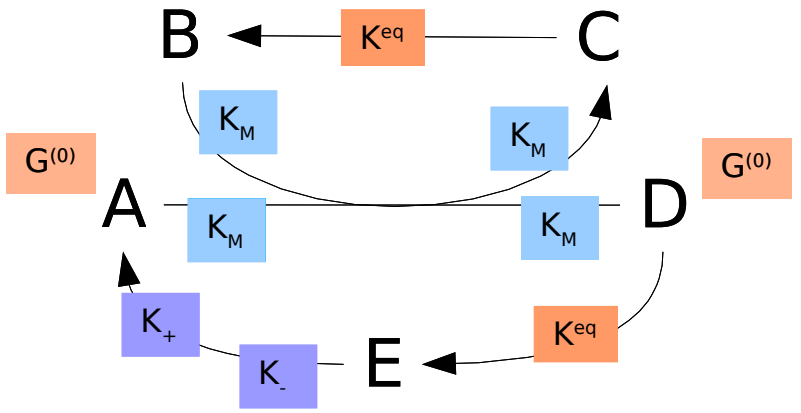
How to combine enzyme kinetic data?

K_+ Turnover rate

K_M K_M value

$G^{(0)}$ Gibbs free energy of formation

K^{eq} Equilibrium constant



Kinetic parameters are dependent!

What is determined by the available data?

How can we account for the error widths?

Can we use prior knowledge to compensate for missing values?

Kinetic parameters are dependent

Reversible Michaelis-Menten kinetics:

$$v = E\frac{k_+^{\mathrm{cat}} a/K_a - k_-^{\mathrm{cat}} b/K_b}{1 + a/K_a + b/K_b}$$

Equilibrium constant K^{eq}

$$k^{\mathrm{eq}} = \left(\frac{b}{a}\right)_{\mathrm{eq}} = \frac{k_+^{\mathrm{cat}}}{k_-^{\mathrm{cat}}}\frac{K_b}{K_a}$$

(1) $$\ln k^{\mathrm{eq}} = \ln k_+^{\mathrm{cat}} - \ln k_-^{\mathrm{cat}} + \ln K_b - \ln K_a$$

that means:
the parameters are dependent given the equilibrium constant

Gibbs free energy of formation $G^{(0)}$

$$k^{\mathrm{eq}} = \mathrm{e}^{-\Delta G^{(0)}/RT}$$

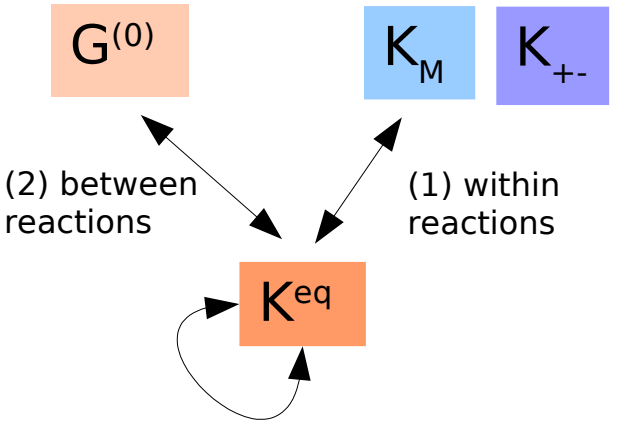
(2) $$\ln \mathbf{k}^{\mathrm{eq}} = -\frac{1}{RT} N^{\mathrm{T}} \mathbf{G}^{(0)}$$

that means:
- the $G^{(0)}$ determine the K^{eq}
- the K^{eq} are dependent

Kinetic parameters are dependent

$$(1) \quad \ln k_l^{\mathrm{eq}} = \ln k_{+l}^{\mathrm{cat}} - \ln k_{-l}^{\mathrm{cat}} + \sum_{\text{reactants } i} N_{il} \ln K_{il}^{\mathrm{M}}$$

$$(2) \quad \ln k_l^{\mathrm{eq}} = -\frac{1}{RT} \sum_i N_{il}^{\mathrm{T}} G_i^{(0)}$$



Problem: complicated relationships between all parameters

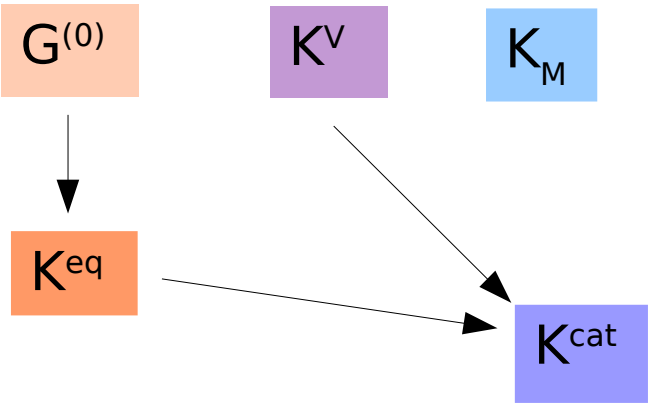
Solution: define a set of INDEPENDENT parameters

Independent system parameters

Velocity constant $k^{\mathrm{V}} = \sqrt{k_+^{\mathrm{cat}} k_-^{\mathrm{cat}}}$

$$\ln k_l^{\mathrm{eq}} = -\frac{1}{RT} \sum_i N_{il}^{\mathrm{T}} G_i^{(0)}$$

$$\ln k_{\pm}^{\mathrm{cat}} = \ln k^{\mathrm{V}} \pm \frac{1}{2} \ln k^{\mathrm{eq}}$$

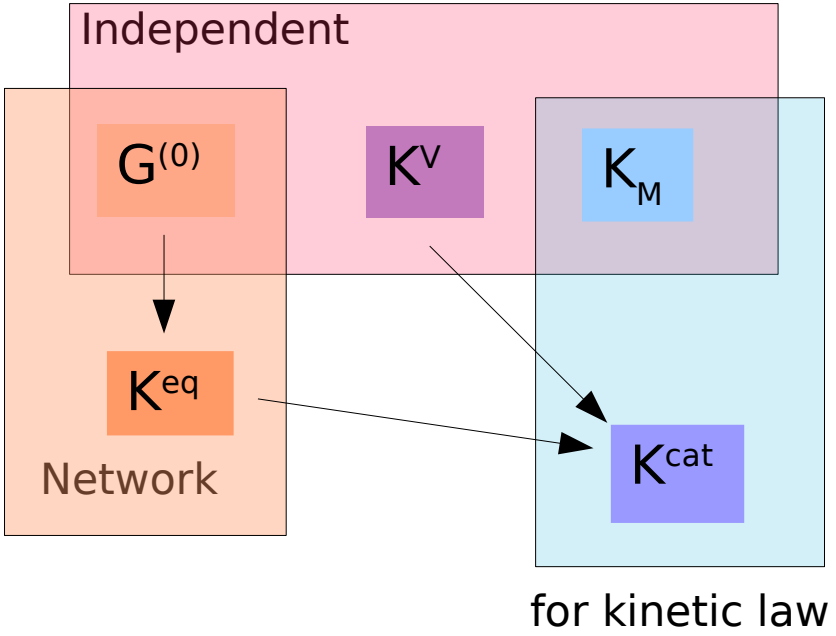


Independent system parameters

Velocity constant $k^{\mathrm{V}} = \sqrt{k_+^{\mathrm{cat}} k_-^{\mathrm{cat}}}$

$$\ln k_l^{\mathrm{eq}} = -\frac{1}{RT} \sum_i N_{il}^{\mathrm{T}} G_i^{(0)}$$

$$\ln k_{\pm}^{\mathrm{cat}} = \ln k^{\mathrm{V}} \pm \frac{1}{2} \ln k^{\mathrm{eq}}$$

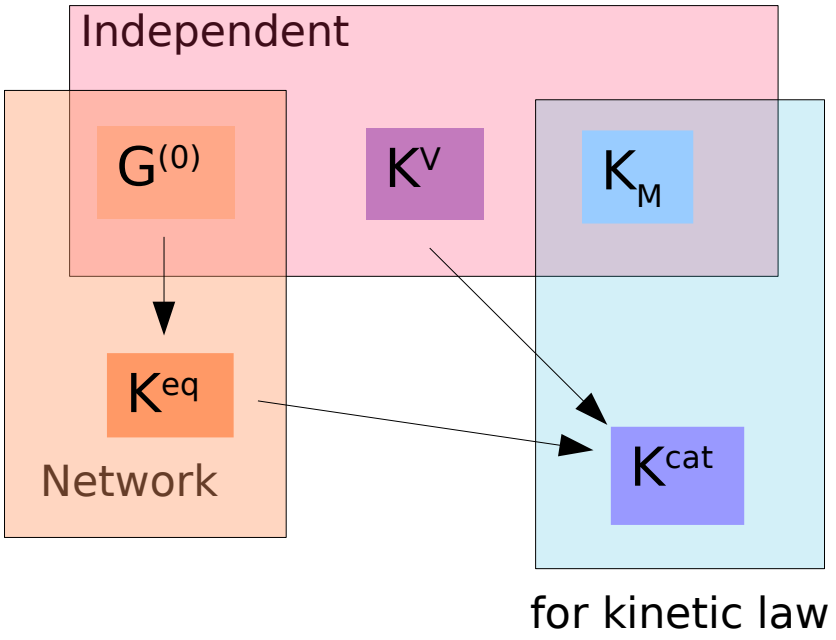


Independent system parameters

Velocity constant $k^{\mathrm{V}} = \sqrt{k_+^{\mathrm{cat}} k_-^{\mathrm{cat}}}$

$$\ln k_l^{\mathrm{eq}} = -\frac{1}{RT} \sum_i N_{il}^{\mathrm{T}} G_i^{(0)}$$

$$\ln k_\pm^{\mathrm{cat}} = \ln k^{\mathrm{V}} \pm \frac{1}{2} \ln k^{\mathrm{eq}}$$



for kinetic law

Linear dependencies for logarithms:

$$x = R\ p$$

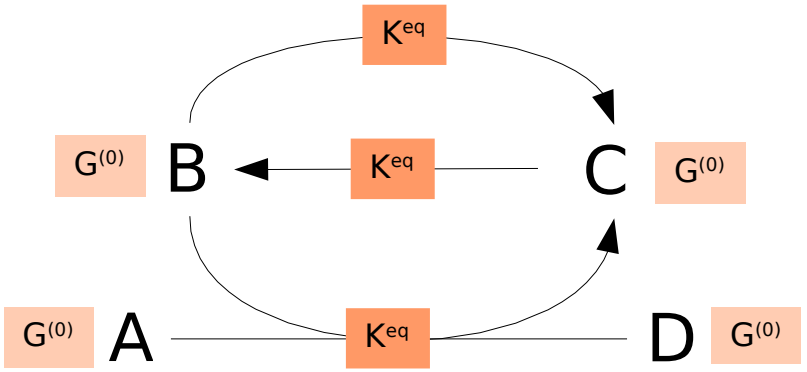
log. observable parameters

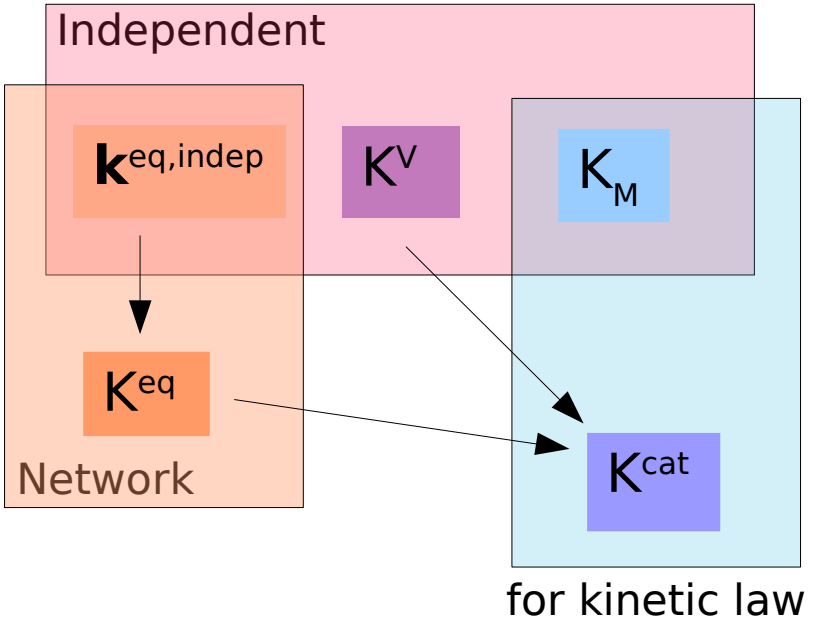
log independent parameters

Dependence matrix

$$x = \begin{pmatrix} G^{(0)} \\ \ln k^{\mathrm{V}} \\ \ln K^{\mathrm{M}} \\ \ln k^{\mathrm{eq}} \\ \ln k_+^{\mathrm{cat}} \\ \ln k_-^{\mathrm{eq}} \end{pmatrix} \quad p = \begin{pmatrix} G^{(0)} \\ \ln k^{\mathrm{V}} \\ \ln K^{\mathrm{M}} \end{pmatrix} \quad R = \begin{pmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \\ -\frac{1}{RT} N^{\mathrm{T}} & 0 & 0 \\ -\frac{1}{2}\frac{1}{RT} N^{\mathrm{T}} & I & 0 \\ +\frac{1}{2}\frac{1}{RT} N^{\mathrm{T}} & I & 0 \end{pmatrix}$$

An alternative: independent equilibrium constants





Define independent equilibrium constants:

Set $N = N_0 L$ (N_0 has independent columns)

$$\ln k^{\mathrm{eq}} = -\frac{1}{RT} N^{\mathrm{T}} G^{(0)}$$

$$= -\frac{1}{RT} L^{\mathrm{T}} N_0^{\mathrm{T}} G^{(0)} = L^{\mathrm{T}} \left(-\frac{1}{RT} N_0^{\mathrm{T}} G^{(0)}\right) = L^{\mathrm{T}} k^{\mathrm{eq,indep}}$$

Exploiting the independent parameters

Are the parameter data feasible?
given data x^*, check whether

$$\exists\, p\colon x^* = R\, p$$

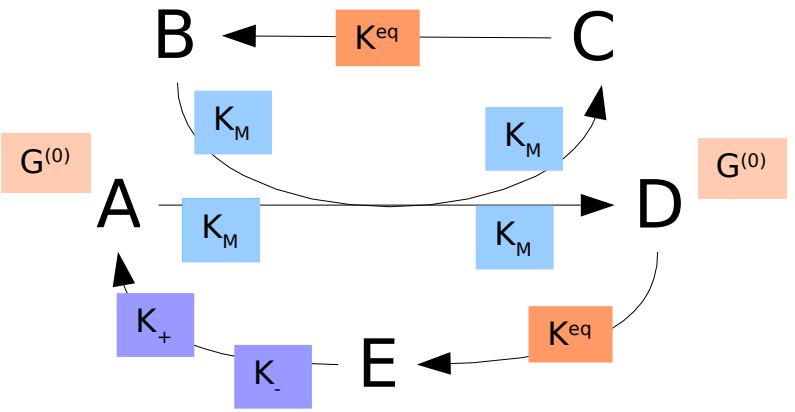
Balance contradicting data
from redundant, contradictory x^*,
obtain complete feasible x:

- solve $x^* = R^*\, p$ by least squares,
- set $x = R\, p$

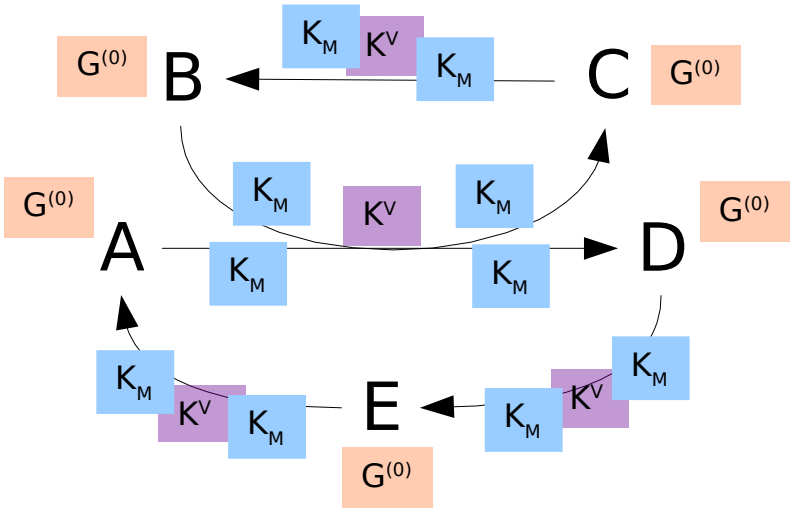
But if there are too few data?
-> Values remain undetermined

Use plausible parameter ranges
as priors in Bayesian estimation

log data x^*

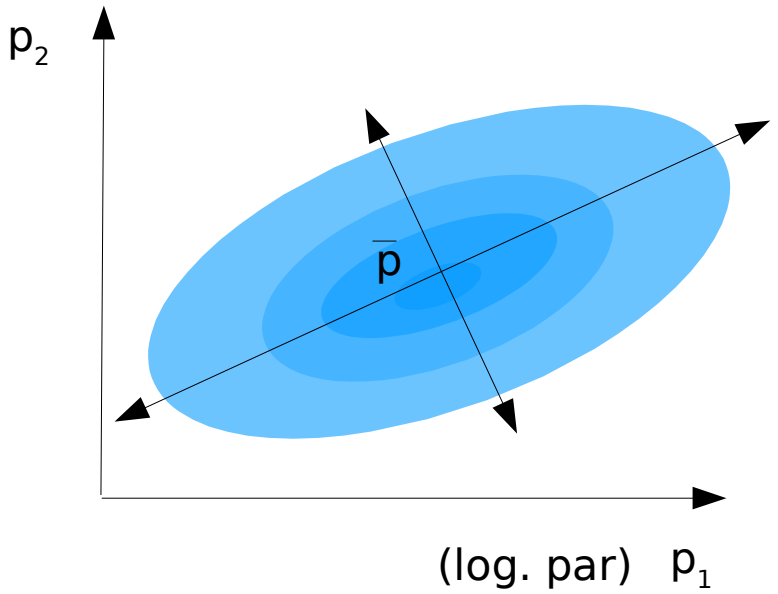


log. independent parameters p

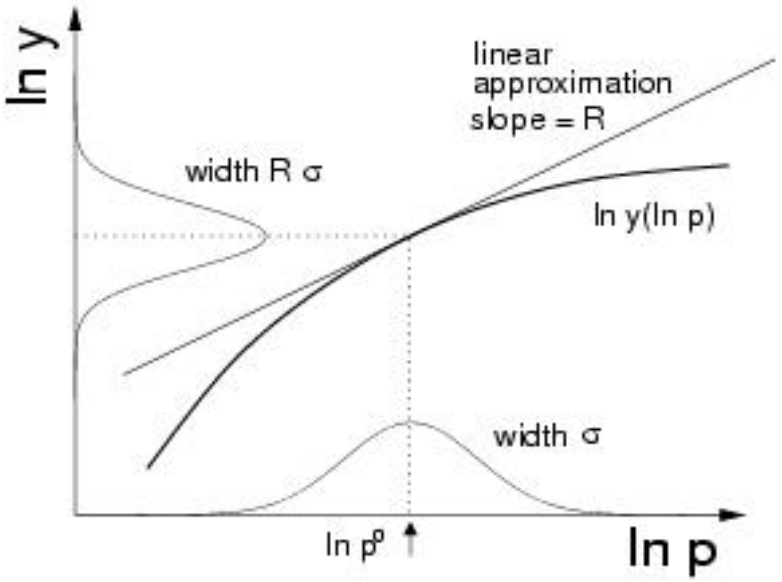


Uncertain parameters: distributions

Linear relationships between **log** parameters
-> **Gaussian** distributions of **log** parameters



Propagation of uncertainties

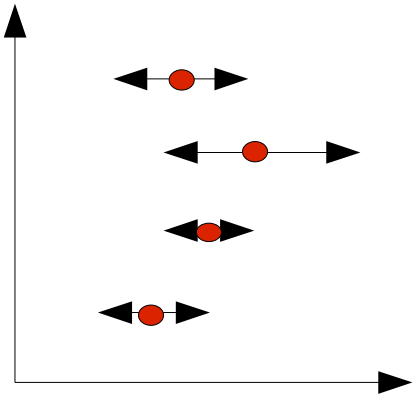


Linear functions $a = R\ b$

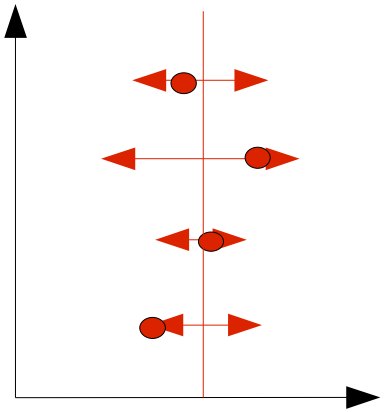
a multivariate Gaussian -> *b* multivariate Gaussian

$$\begin{aligned} \bar{a} &= R\bar{b} \\ \mathrm{cov}(a) &= R\ \mathrm{cov}(b)\ R^{\mathrm{T}} \end{aligned}$$

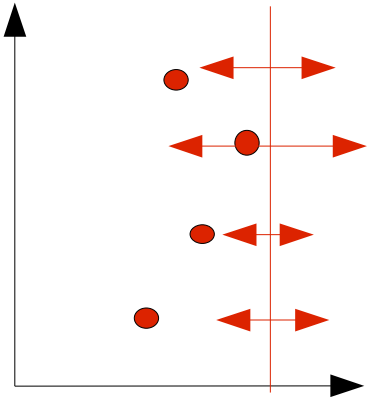
What was maximum likelihood again?



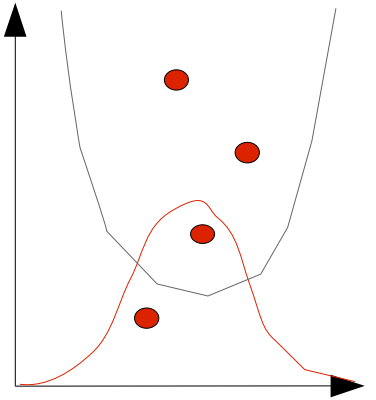
Observed values x^*



Value x



Another value x



Likelihood: *Prob(x^*|x)*
Sum of square errors

Maximum likelihood estimator: $x_{\text{est}} = \text{argmax}_x \, \text{Prob}(\text{data } x^* | \text{value } x)$

Observations with Gaussian errors: $x_{\text{est}} = \text{argmin}_x \sum_i \frac{(x_i^* - x)^2}{\sigma_i^2}$

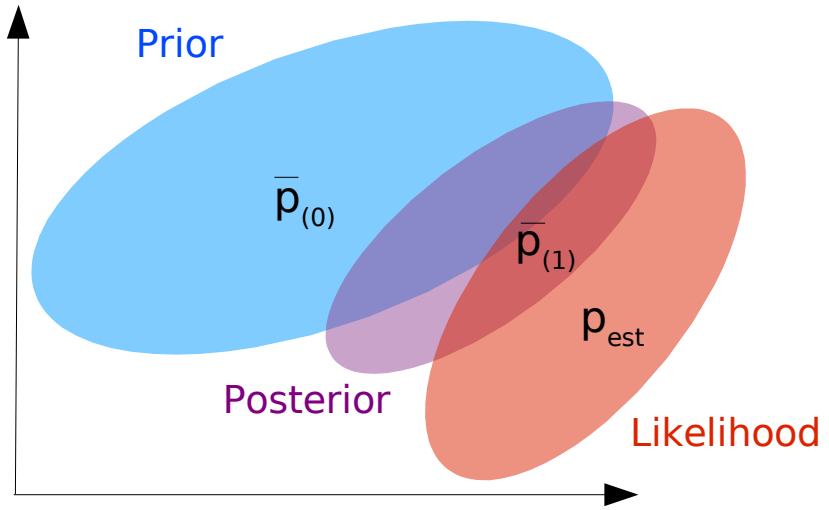
Estimate vector p from observations of x =R p: $x^* = \mathcal{N}(x = R\ p, C_x)$

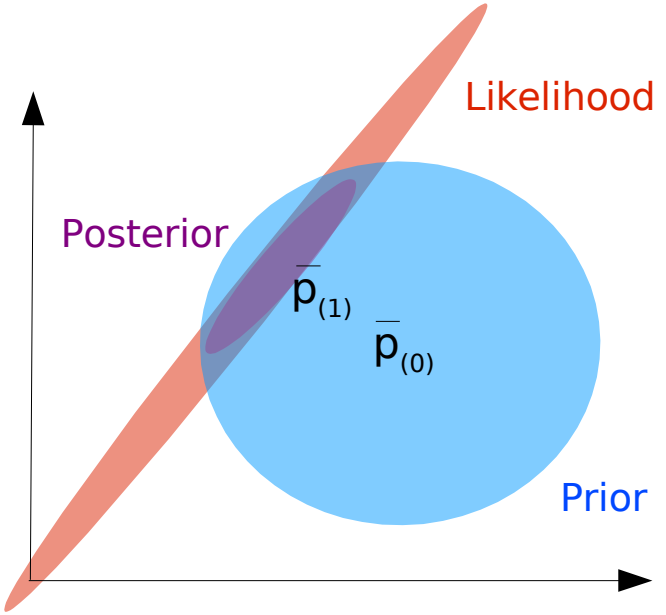
$$p_{\text{est}} = \text{argmax}_p \, \text{Prob}(\text{data } x^* | \text{value } p)$$

$$p_{\text{est}} = \text{argmin}_p (x^* - Rp)^{\mathrm{T}} C_x^{-1} (x^* - Rp)$$

$$= (R^{\mathrm{T}} C_x R)^{-1} R^{\mathrm{T}} C_x^{-1} x^*$$

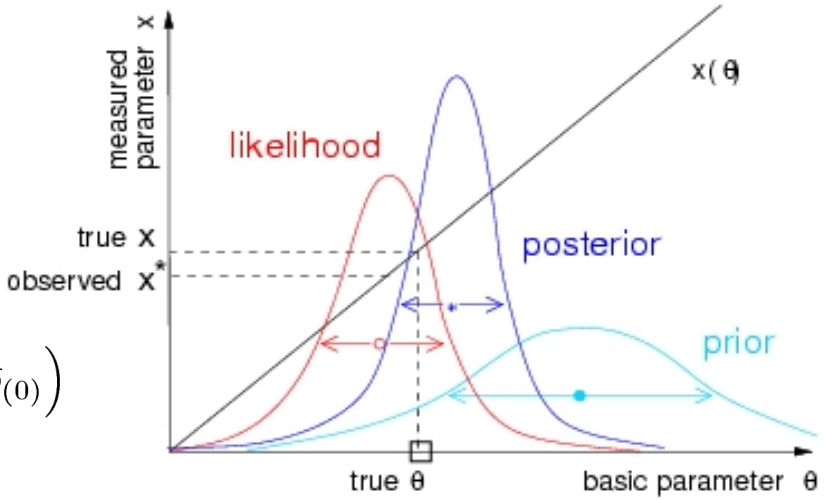
Bayesian parameter estimation





Additional prior knowledge in form of a distribution *Prob(p)*:

Prob(p| x) ~ Prob(x*|p) Prob(p)*

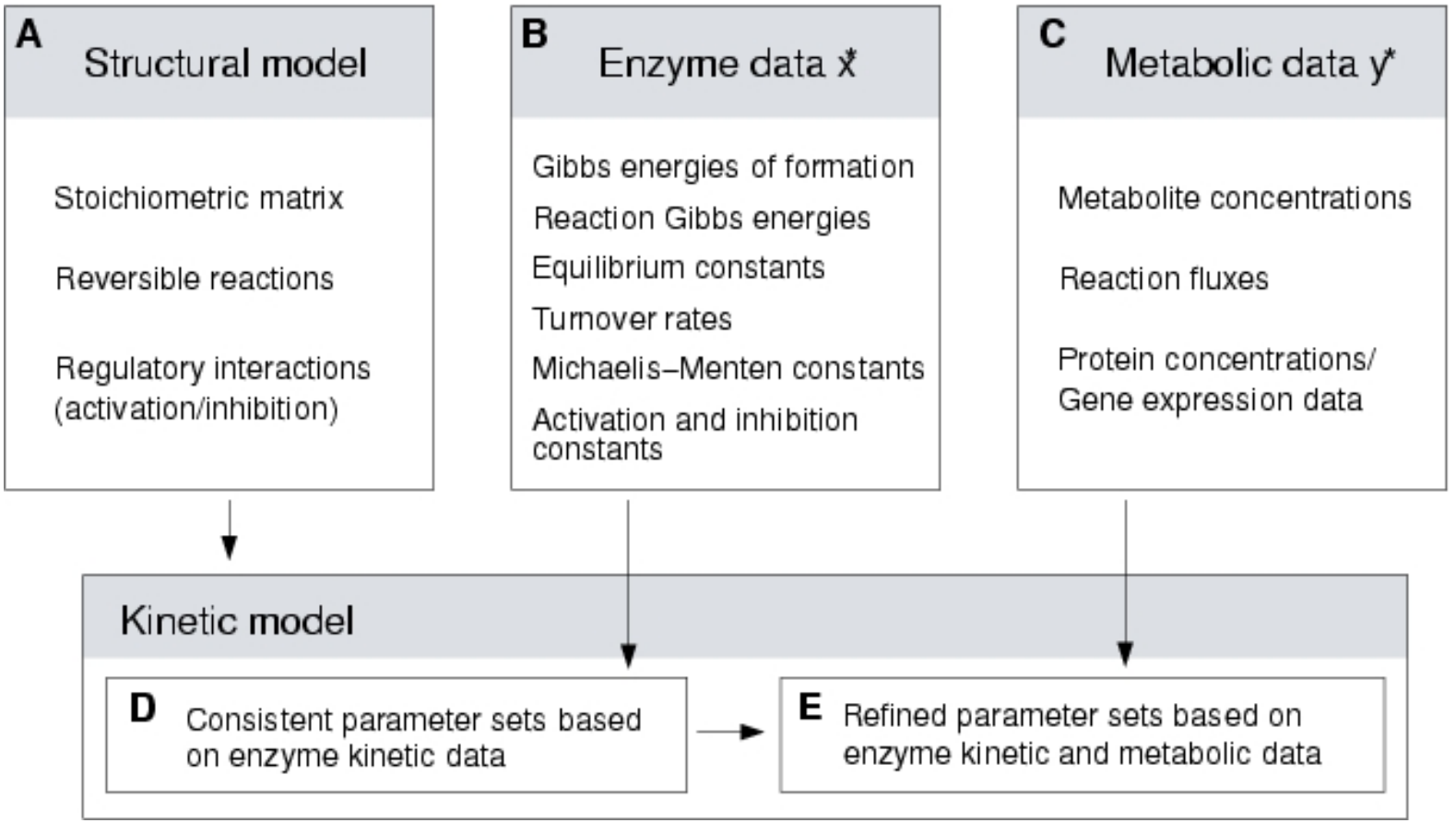


Given data $x^*=R\ p$:
mean and covariance matrix

$$\bar{p}_{(1)} = \left(C_{(0)}^{-1} + (R)^{\mathrm{T}} C_{\mathrm{x}}^{-1} R\right)^{-1} \left((R)^{\mathrm{T}}\ C_{\mathrm{x}}^{-1}\ x^* + C_{(0)}^{-1}\ \bar{p}_{(0)}\right)$$

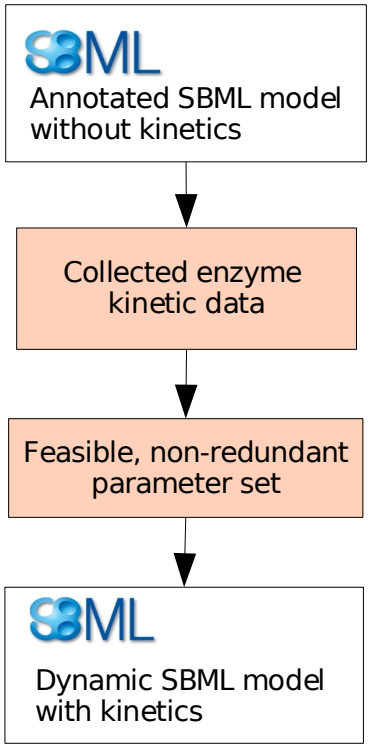
$$C_{(1)} = \left(C_{(0)}^{-1} + (R)^{\mathrm{T}} C_{\mathrm{x}}^{-1} R\right)^{-1}$$

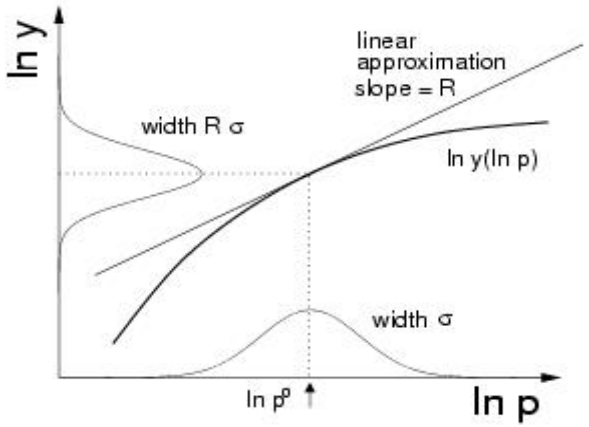
Improving the parameter estimates: integration of metabolic data

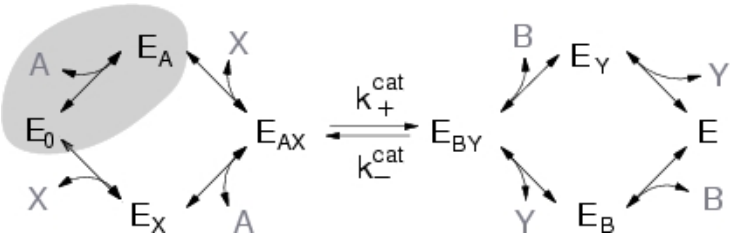


The workflow

1. Define model structure
2. Use generalised Michaelis-Menten kinetics ("convenience kinetics")
3. Retrieve and map kinetic data relevant for the model
4. Determine priors
5. Balance parameters by Bayes estimation
6. Insert kinetics into SBML
7. Predict probabilistic model behaviour







Important points in data collection

Each data point needs an error range

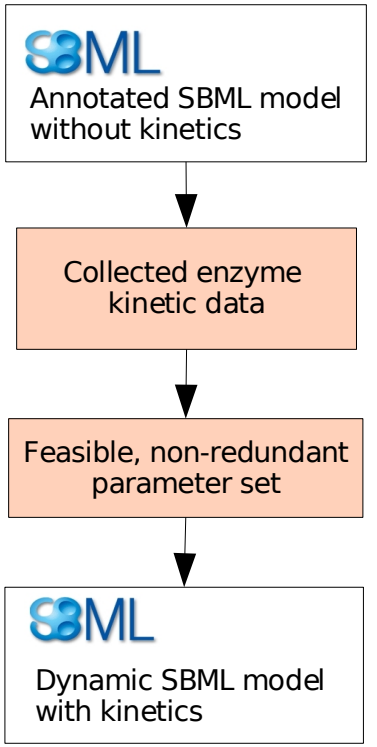
- Must be realistic
- How to compare error ranges for different kinds of data?
- Different error sources

Data must stem from independent sources

- never use the same database entry twice
- never reuse data that were already used in the computations

It is important to keep track of correlations

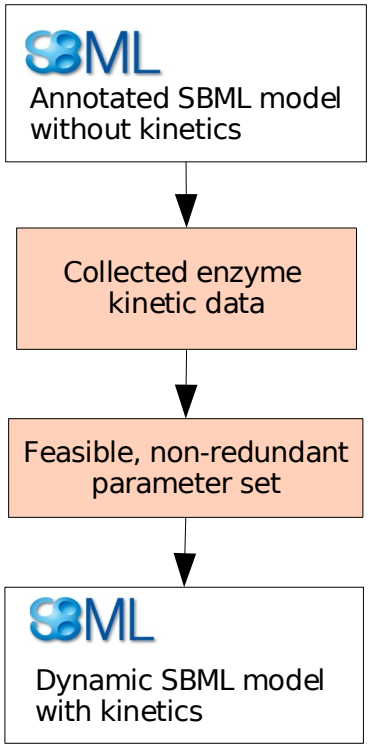
- “posterior is the new prior”
- some knowledge about equilibrium constants is stored in $G^{(0)}$ correlations



Aims and weaknesses of the workflow:

How can I use it?

- model-driven data collection and mapping of relevant data
- assess the possible ranges and correlations of parameters;
- provide parameter priors for model fitting



What's the use of low-fi models??

- initial model as starting point for detailed models
- complete high quality models, fill gaps (better than nothing)
- determine which additional data are most needed

What are their weaknesses?

- convenience kinetics differs from true laws
- definition of parameters may differ
- computed, in-vitro, or inferred values differ from the ones needed in the model
- automatic data mapping requires detailed and reliable annotations
- general weaknesses of bottom-up-approaches

What do we need to make the workflow useful?

Data collection and mapping

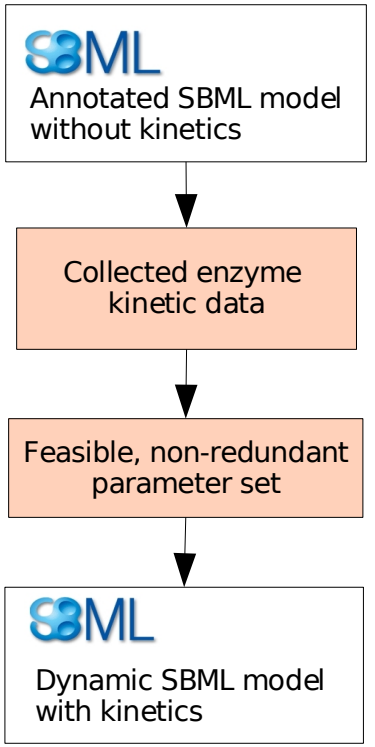
- easy access to databases
- precise annotations in models and kinetic data
- standard exchange formats for enzymatic data

Additional data sources

- databases
- statistical learning methods
- ab-initio calculations

Treatment of uncertainties & error propagation

- all data should come with error bars
- standards to describe uncertainties in SBML and enzyme data
- standards to describe relationships between parameters
- standards to quantify parameter correlations
- users: model fitting tools that use parameter priors



Thanks to ...

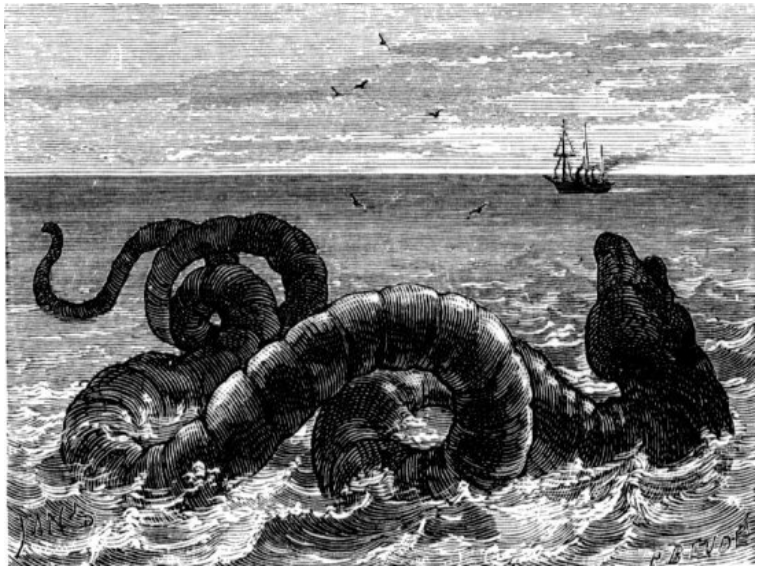
MPI for Molecular Genetics
Simon Borger
Jannis Uhlendorf
Anselm Helbig
Edda Klipp

Dietmar Schomburg lab
Sebastian Buchinger
Eva Vaylann

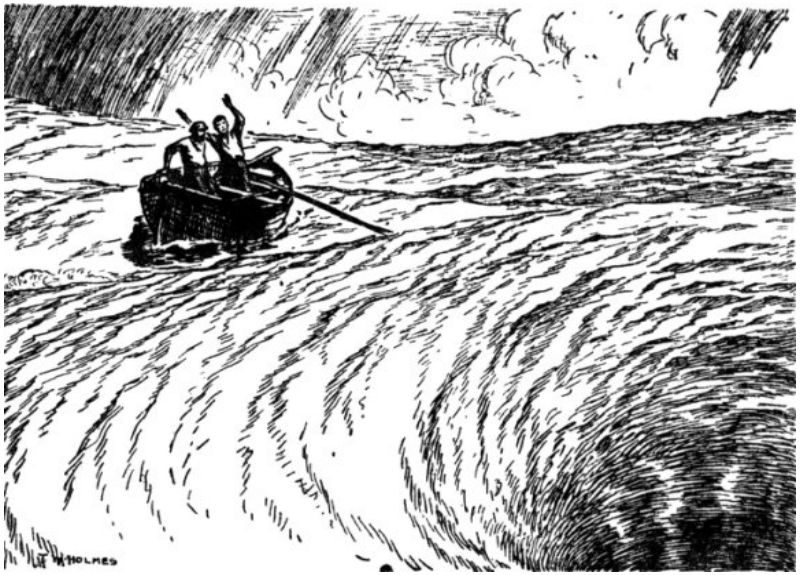
Genoscope, Evry
Vincent Schachter
Maxime Durot

... and to you !!!

Bottom-up and top-down model building



Scylla (“bottom-up”)
in-vitro data may be wrong ...
data may not be transferable ...
models SHOULD not fit all data ...



Charybdis (“top-down”)
the fitted model may work,
but the parameters have an
unclear interpretation ...