Optimal experimental design for the well and the ill (-posed problems)

André Bardow
Institute of Technical Thermodynamics, RWTH Aachen University

27 July 2011

Acknowledgements

- **P. Domagalski**, V. Göke, X. Liu, H.-J. Koß
- **E. Kriesten, S. Kossack**, W. Marquardt
- M. Voda, F. Casanova, B. Blümich

- M. Ottens

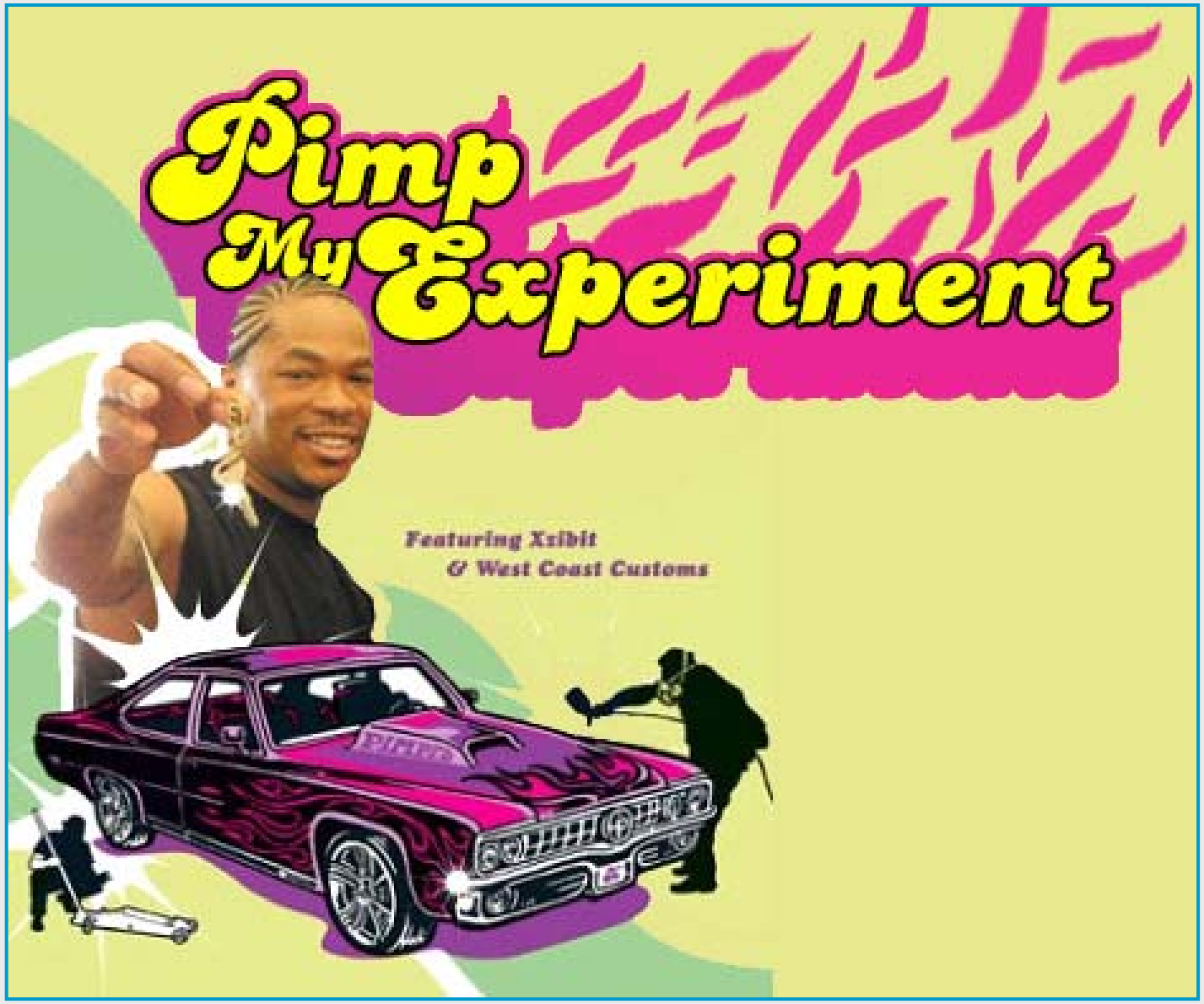
Funding

- German Science Foundation (DFG)
- Delft Research Centers for Life-Sciences and for Sustainable Processes

Design for the well and the ill

- Design of diffusion experiments
- Design for ill-posed problems:
  The METER approach
- OED for physical property models

What is optimal experimental design?



Pimp my diffusion experiment

take an old car…

make it faster, nicer,… – better !

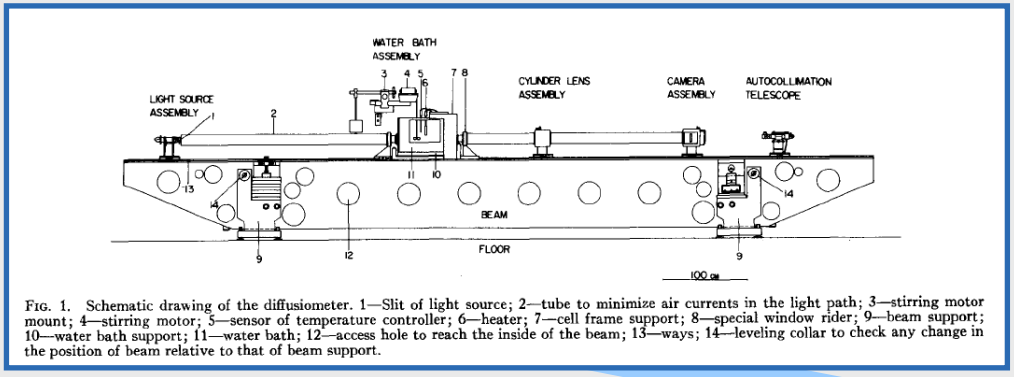
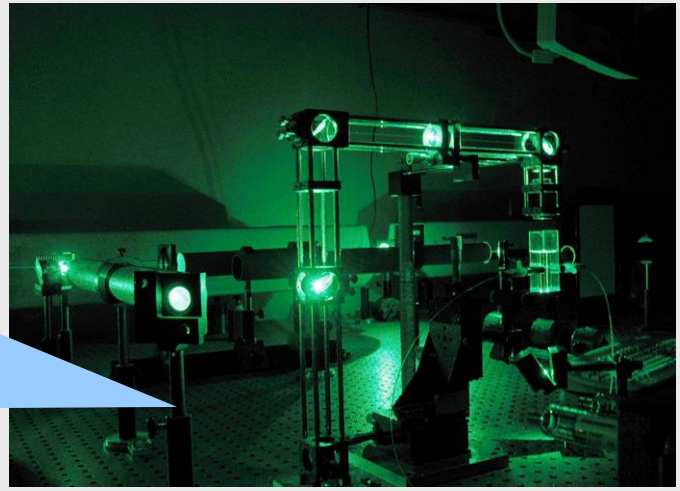


FIG. 1. Schematic drawing of the diffusiometer. 1—Slit of light source; 2—tube to minimize air currents in the light path; 3—stirring motor mount; 4—stirring motor; 5—sensor of temperature controller; 6—heater; 7—cell frame support; 8—special window rider; 9—beam support; 10—water bath support; 11—water bath; 12—access hole to reach the inside of the beam; 13—ways; 14—leveling collar to check any change in the position of beam relative to that of beam support.

take an old experiment…



make it faster, nicer,… – better !

Current diffusion experiments

Method	Accuracy	Temperature range	Pressure range	Experiment time
Diaphragm cell	Good (0.5–1%)	Moderate ($T \leq 400$ K)	Wide†	Days
Capillary				
Open	Moderate	Wide range	Narrow††	Days
Closed	High (0.2%)	Near ambient	Narrow††	
Conductance	High (0.2%)	Near ambient	Narrow††	Days
Taylor dispersion	Moderate (~ 1%)	Wide range ($T \leq 600\ K$)	Moderate Moderate†	< 1 hour
NMR	Moderate (1–2%)	Wide range ($T \leq 700\ K$)	Wide††	< 1 hour
Gouy	High($\leq 0.1\%$)	Near ambient	Narrow	1–2 days
Rayleigh	High ($\leq 0.1\%$)	Near ambient	Narrow	1–2 days

Wakeham, Measurement of the transport properties of fluids, IUPAC 1991

Optimal Experimental Design: Idea

set free variables ***x***
such that information
on parameter θ is maximized

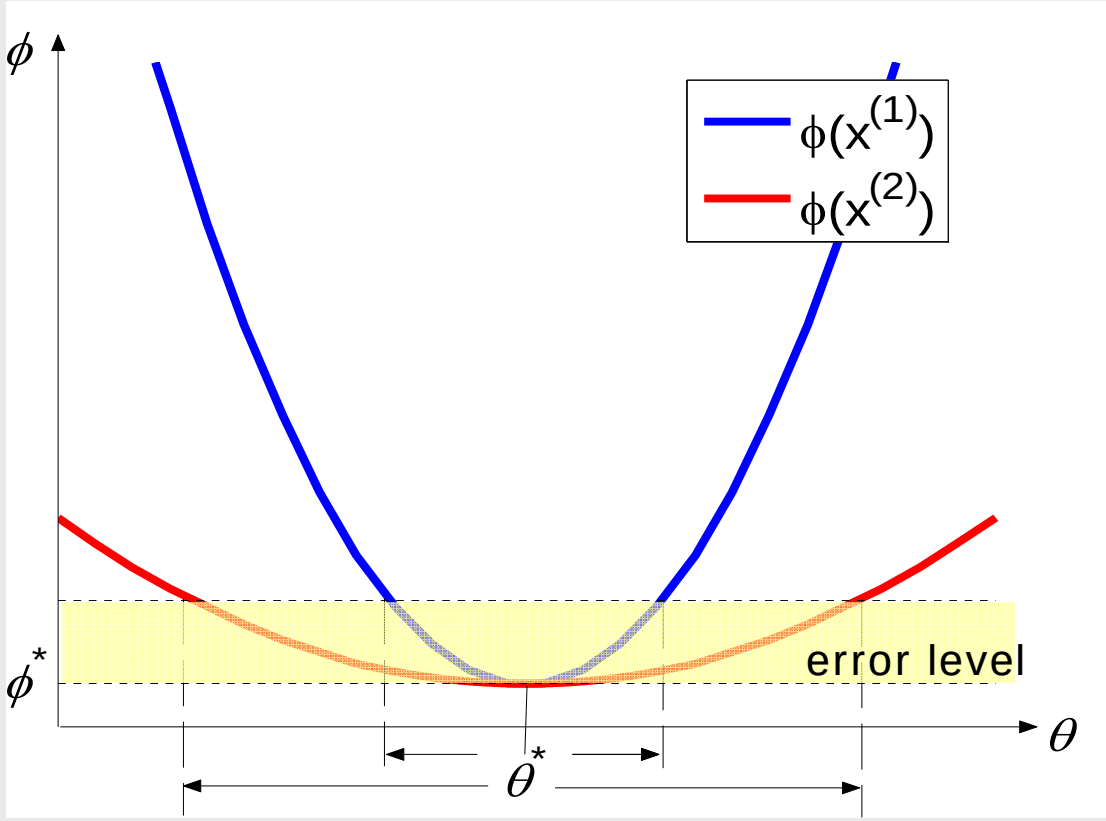
||

maximize curvature
of parameter estimation
objective

OED objective
for parameter estimation

$$\max_{x} \frac{\partial^2 \phi}{\partial \theta^2}$$

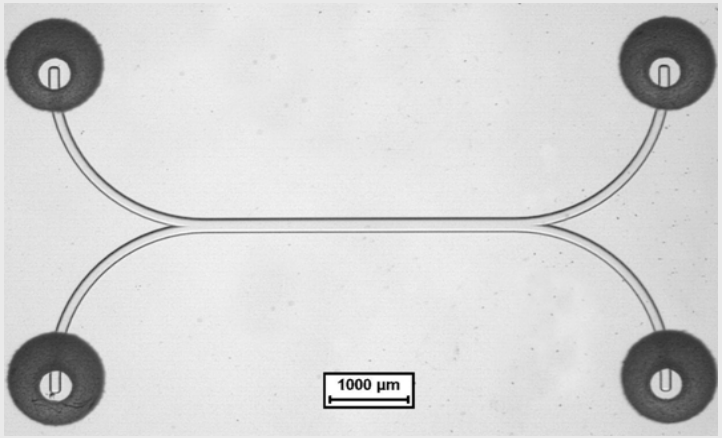
parameter estimation objective

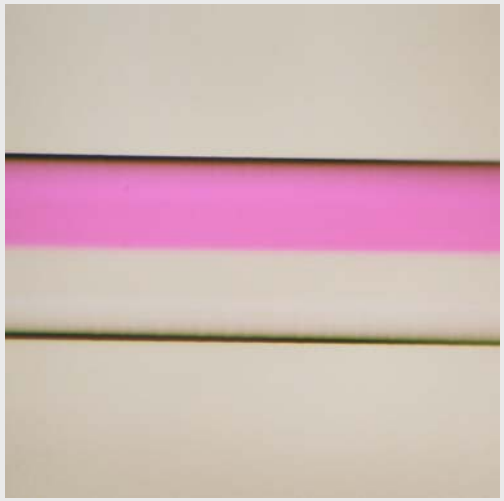


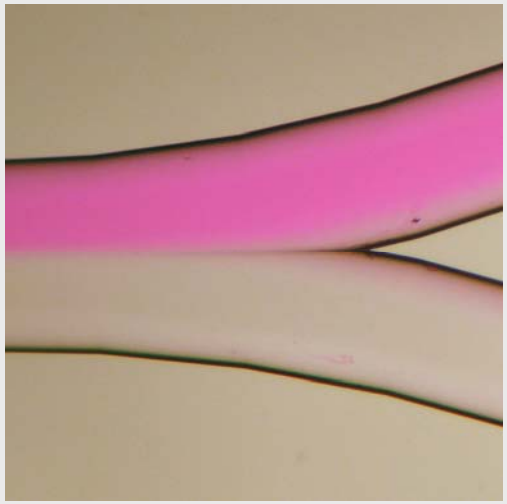
Model-based design of μ-diffusion chip

⇒ Diffusion experiments on the micro-scale

H-cell microchannel







(P. Domagalski, M. Ottens, AB, *submitted*)

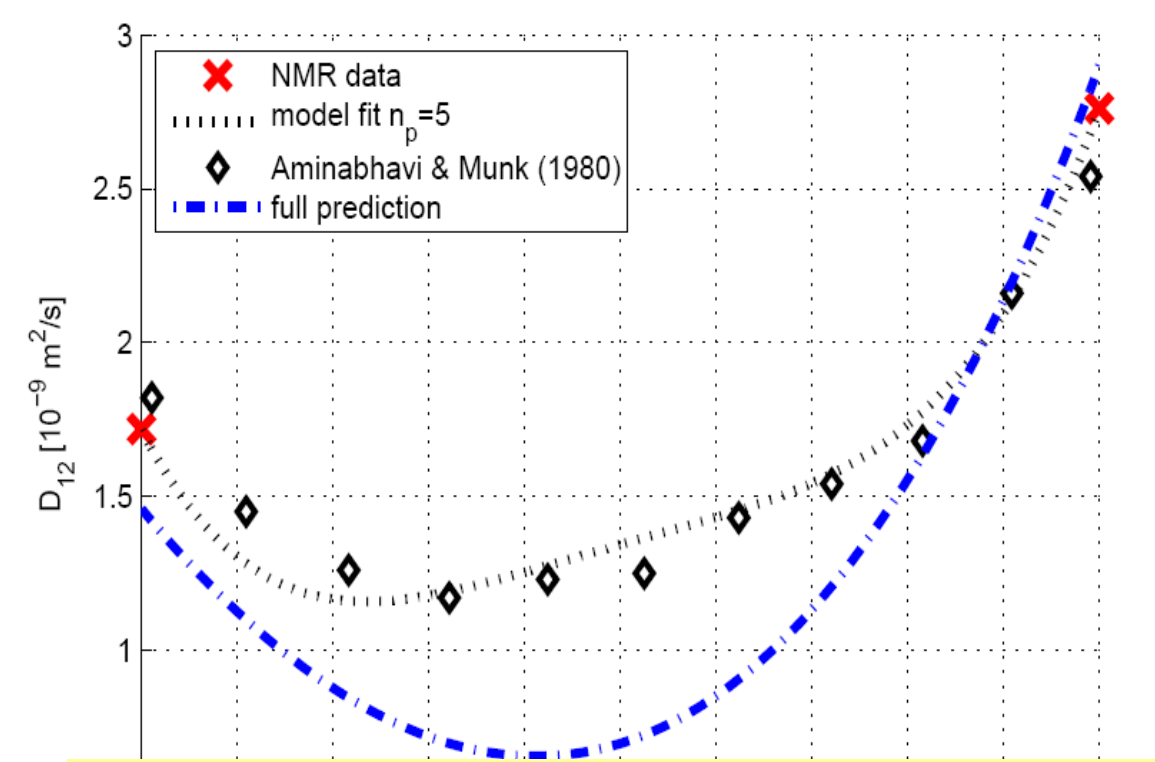
Unpublished slides deleted

Update: diffusion experiments

Method	Accuracy	Temperature range	Pressure range	Experiment time
Diaphragm cell	Good (0.5–1%)	Moderate ($T \leq 400$ K)	Wide†	Days
Capillary				
Open	Moderate	Wide range	Narrow††	Days
Closed	High (0.2%)	Near ambient	Narrow††	
Conductance	High (0.2%)	Near ambient	Narrow††	Days
Taylor dispersion	Moderate (~ 1%)	Wide range ($T \leq 600$ K)	Moderate Moderate‡	< 1 hour
NMR	Moderate (1–2%)	Wide range ($T \leq 700$ K)	Wide‡‡	< 1 hour
Gouy	High($\leq 0.1\%$)	Near ambient	Narrow	1–2 days
Rayleigh	High ($\leq 0.1\%$)	Near ambient	Narrow	1–2 days
Microfluidic H-Cell	Moderate	To be explored		**minutes**

Can we do even better?

- diffusion coefficients are concentration-dependent
- usually many experiments to determine this dependency
- here: one Raman-experiment plus 2 NMR data points



⇒ whole dependence from minimal data
⇒ but requires to solve ill-posed problem

(E. Kriesten et al., *Fluid Phase Equilibria*, 2009)

Tikhonov regularization for function estimation

Estimation problem

$$\min_r \|c-\tilde{c}\|$$

$$s.t.\ \ c(t)-c_0=\int_0^t r(t)dt$$

$$g(t)=\int_T K(t,s;d)\,f(s)ds$$

regularization

Regularized problem

$$\min_r \|c-\tilde{c}\|+\lambda\|r\|$$

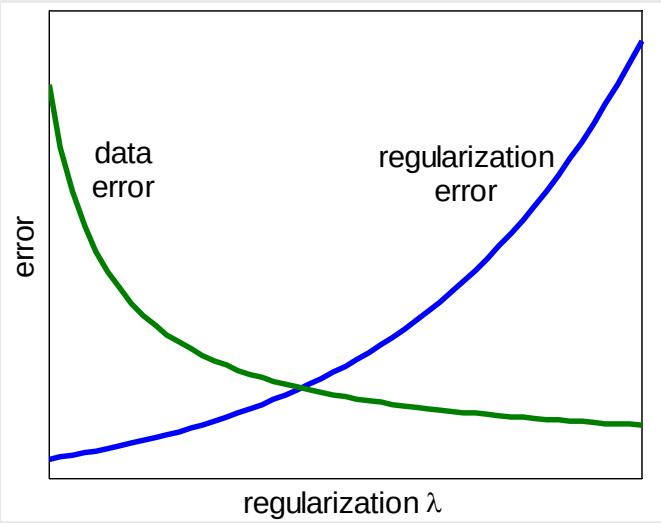
$$s.t.\ \ c(t)-c_0=\int_0^t r(t)dt$$

$$\underbrace{\min_f \sum_{i=1}^{n}\frac{1}{\sigma_i^2}\left(g(t_i)-\int_T K(t_i,s;d)\,f(s)ds\right)^2}_{\text{lack of fit (“least-squares”)}}\ \underbrace{+\lambda\|Lf\|_{L_2}^2}_{\text{smoothness penalty}}$$

lack of fit (“least-squares”) → variance

smoothness penalty → bias

bias-variance trade-off crucial for success



Optimal Experimental Design: Idea

set free variables ***x***
such that information
on parameter θ is maximized

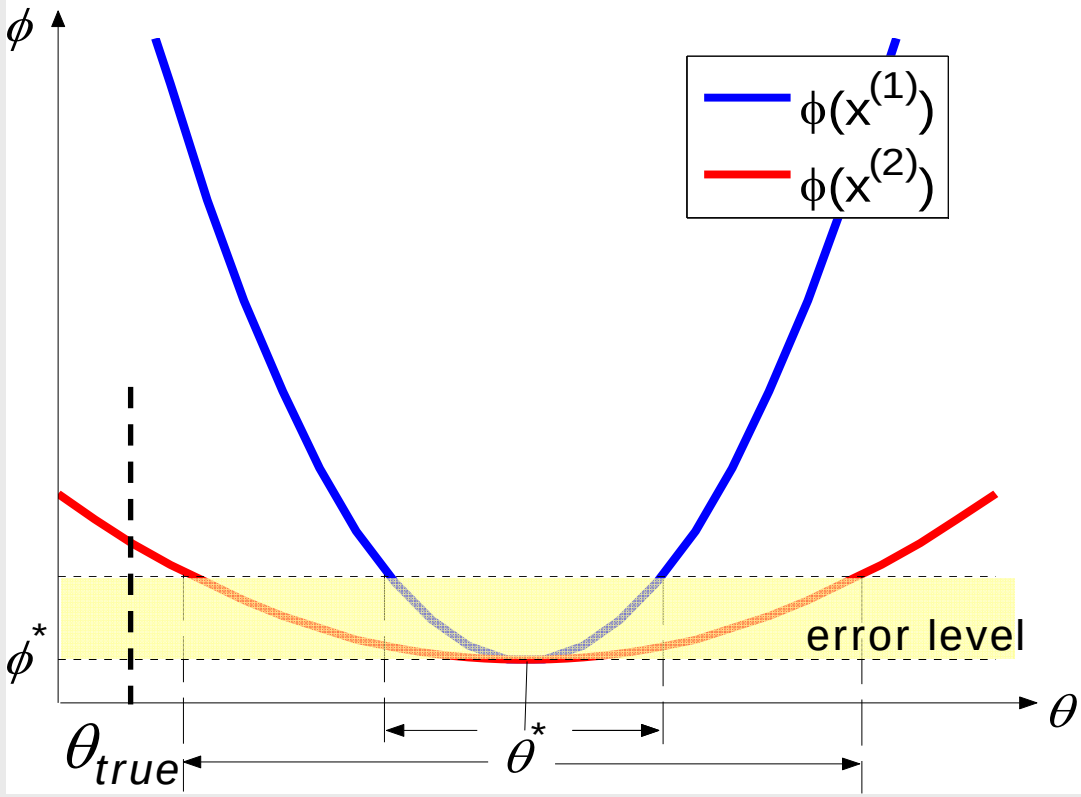
||

maximize curvature
of parameter estimation
objective

OED objective

$$\max_{x} \frac{\partial^2 \phi}{\partial \theta^2}$$

parameter estimation objective



- focus on variance only
- bias contribution is missing
- not applicable

⇒ "estimate the wrong solution with less uncertainty"

Minimum Expected Total Error (METER)

Optimal design:

find settings for experiment that allow best determination of unknown function

classical approach (Box-Lucas school): minimize parameter variance only

approach proposed for ill-posed problems (Box-Draper school):

minimize

$$E\left(\left\| f - \hat{f} \right\|_{L_2}^2\right)$$ expected total error in estimate

$$= \left\| f - K^+(\lambda) K f \right\|_{L_2}^2$$ = *bias* due to regularization

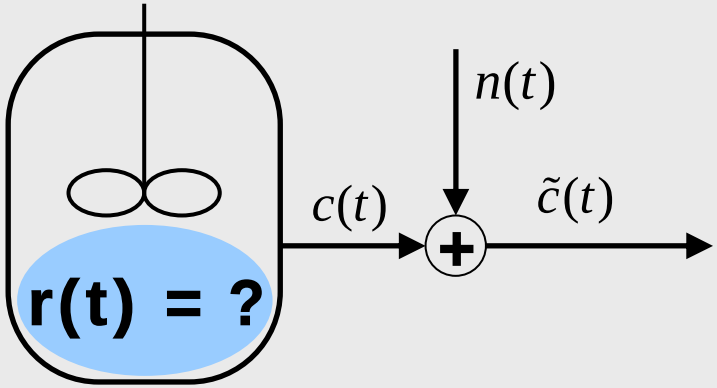
$$+\sigma^2 trace\left(K^+(\lambda)\left(K^+(\lambda)\right)^T\right),$$ + *variance* in estimate

with $K^+(\lambda) = \left(K^T K + \lambda L^T L\right)^{-1} K^T.$ "regularized Fisher information matrix"

(A. Bardow, *Comput. Chem. Eng.* 2008)

Example: Numerical differentiation

Application: Batch reactor



- determine reaction rate *r(t)* from concentration measurements *c(t)*

$$c(t) = \int_0^1 r(t')dt' \quad \Rightarrow \quad r(t) = \frac{dc(t)}{dt}$$

- numerical differentiation of experimental data

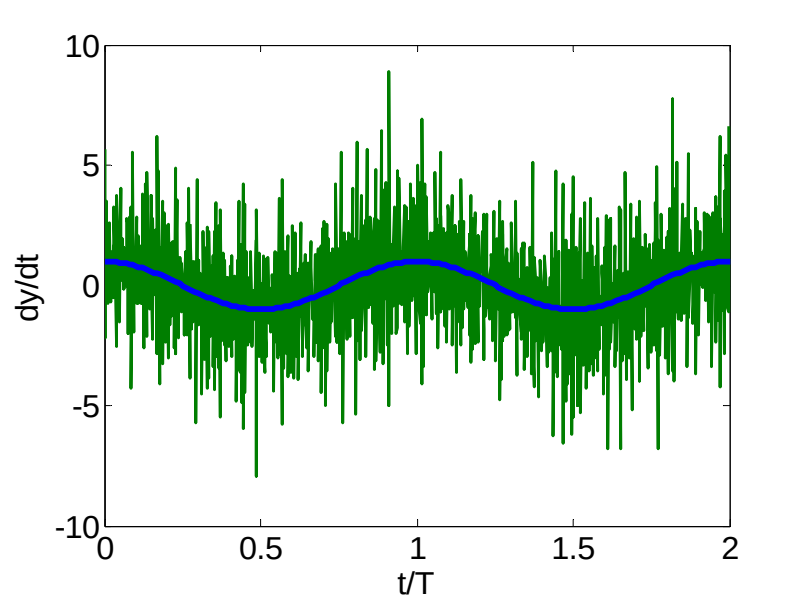
$\Rightarrow$ ill-posed problem

- solution by finite differences

$$\hat{r}(t_i) = \frac{c(t_i) - c(t_{i-1})}{\Delta t}$$

- time step Δt too large $\rightarrow$ approximation error
- time step Δt too small $\rightarrow$ amplification of noise

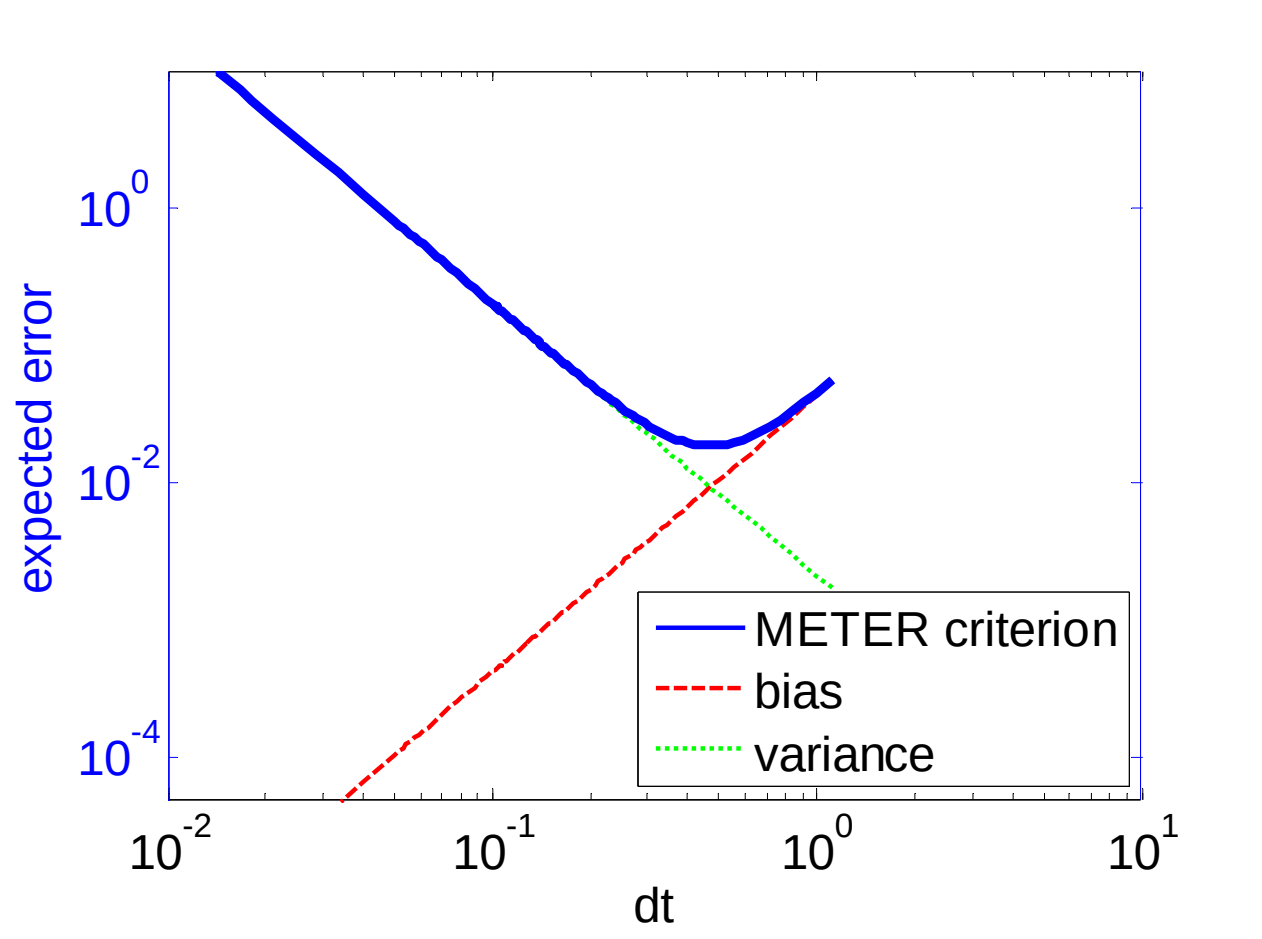
determine best time step in optimal design



Optimal sampling time by METER Criterion

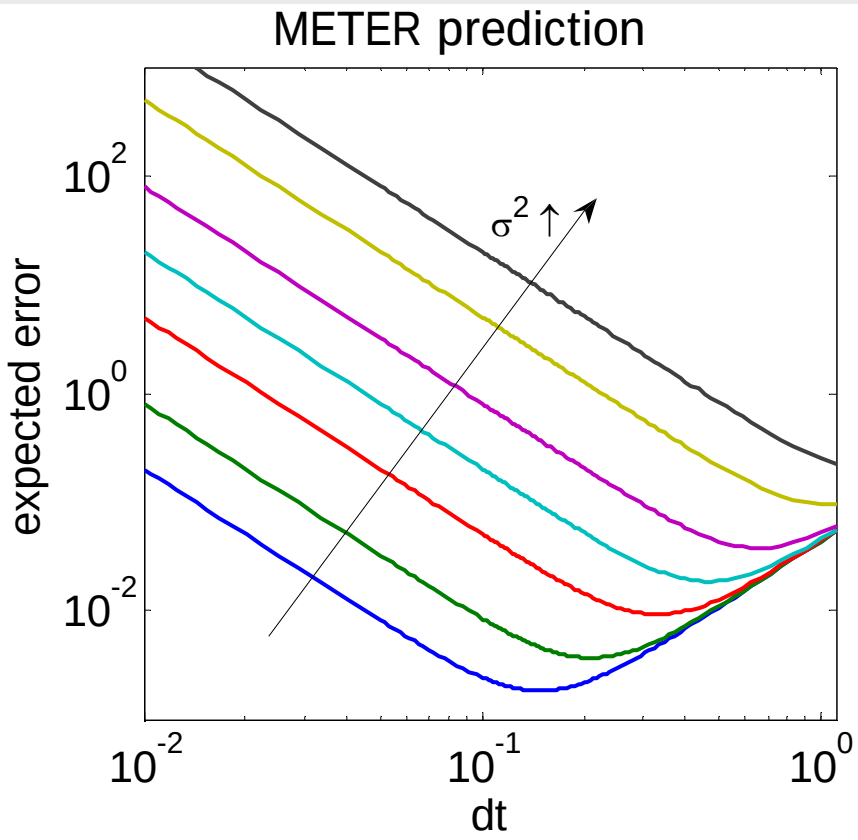
$r(t) = \exp(-t)$

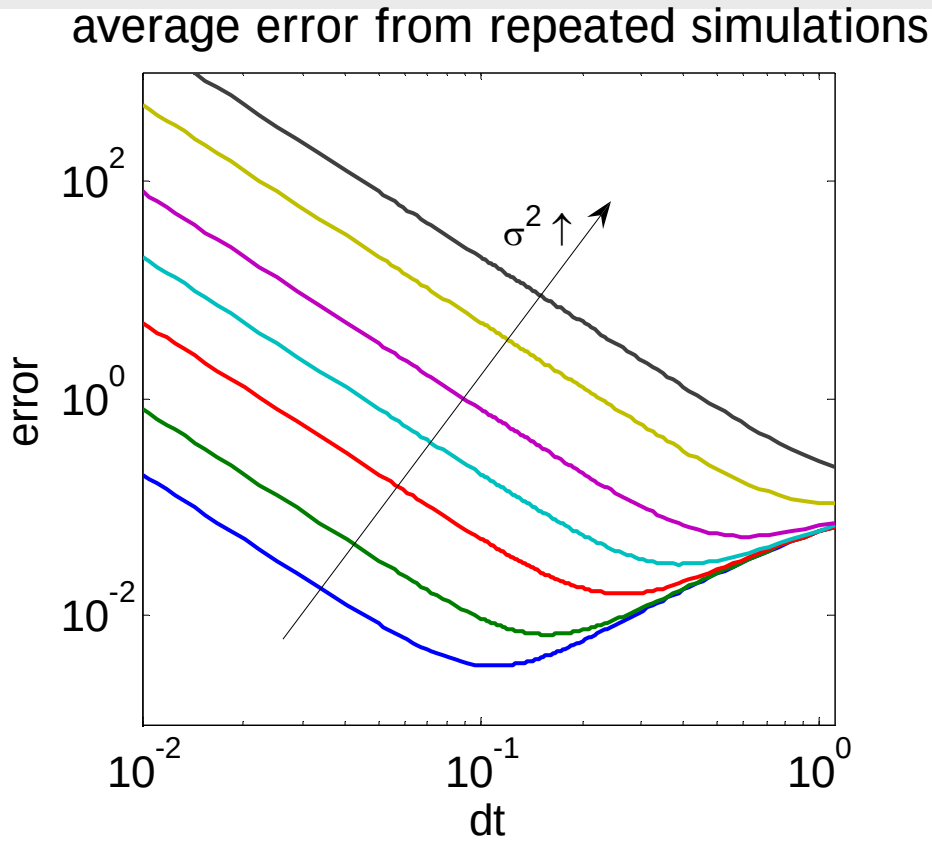
$t = [0,10]$



- bias and variance show the expected trends
- optimal sampling time as trade-off balance by noise level

Comparison with Numerical Computations





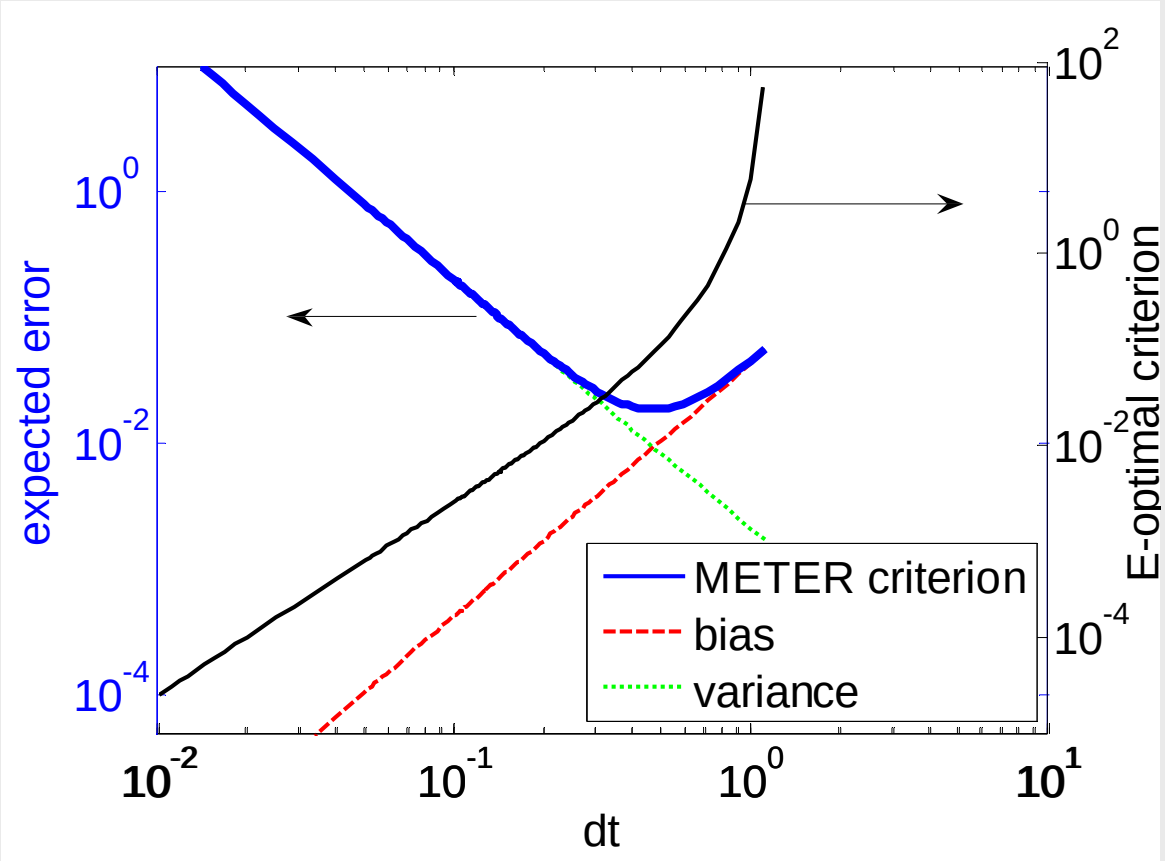
- predicted behaviour is found in practice
- small difference from approximate bias term

Comparison with Classic Criteria

$r(t) = \exp(-t)$

$t = [0,10]$

E-optimal design: max min eigenvalue of Fisher matrix

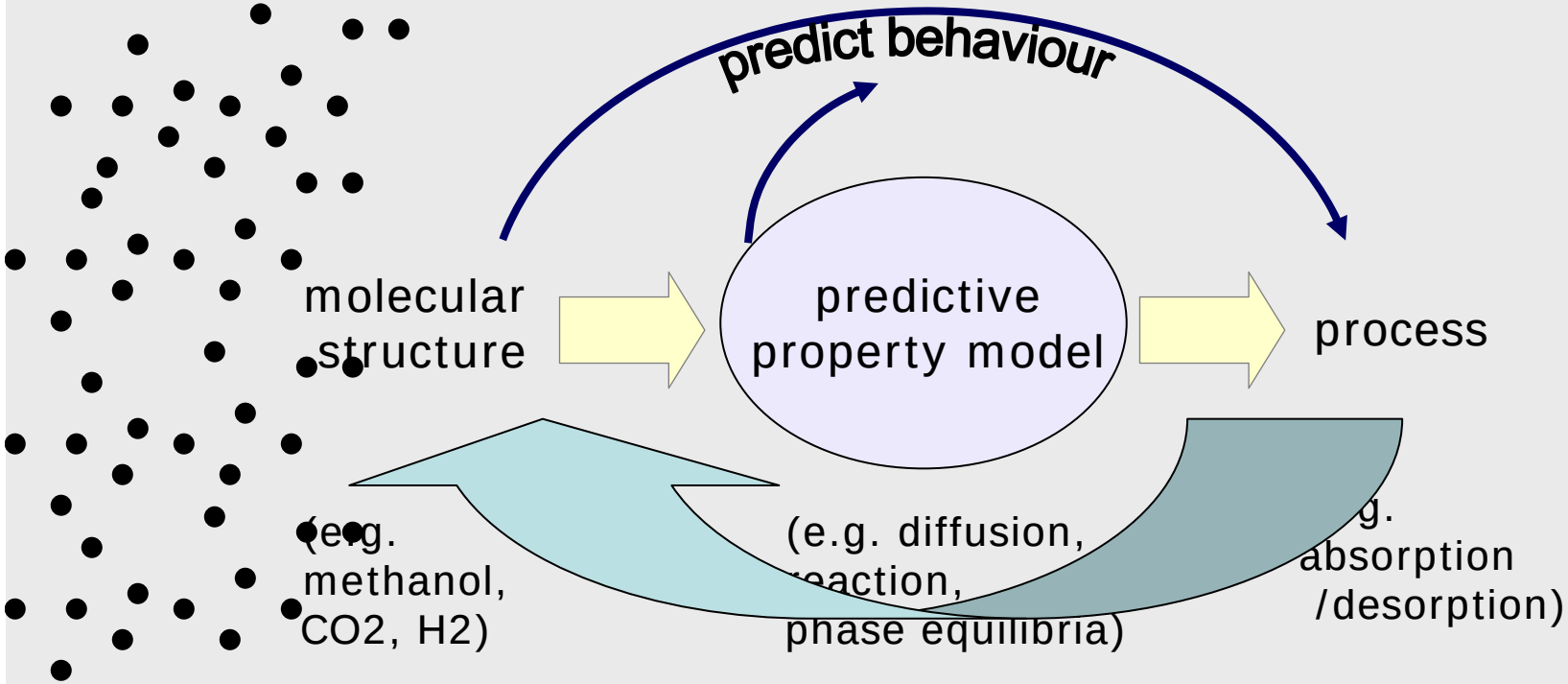


- standard criteria favor maximal time step
- predictions are even qualitatively in error

Discussion

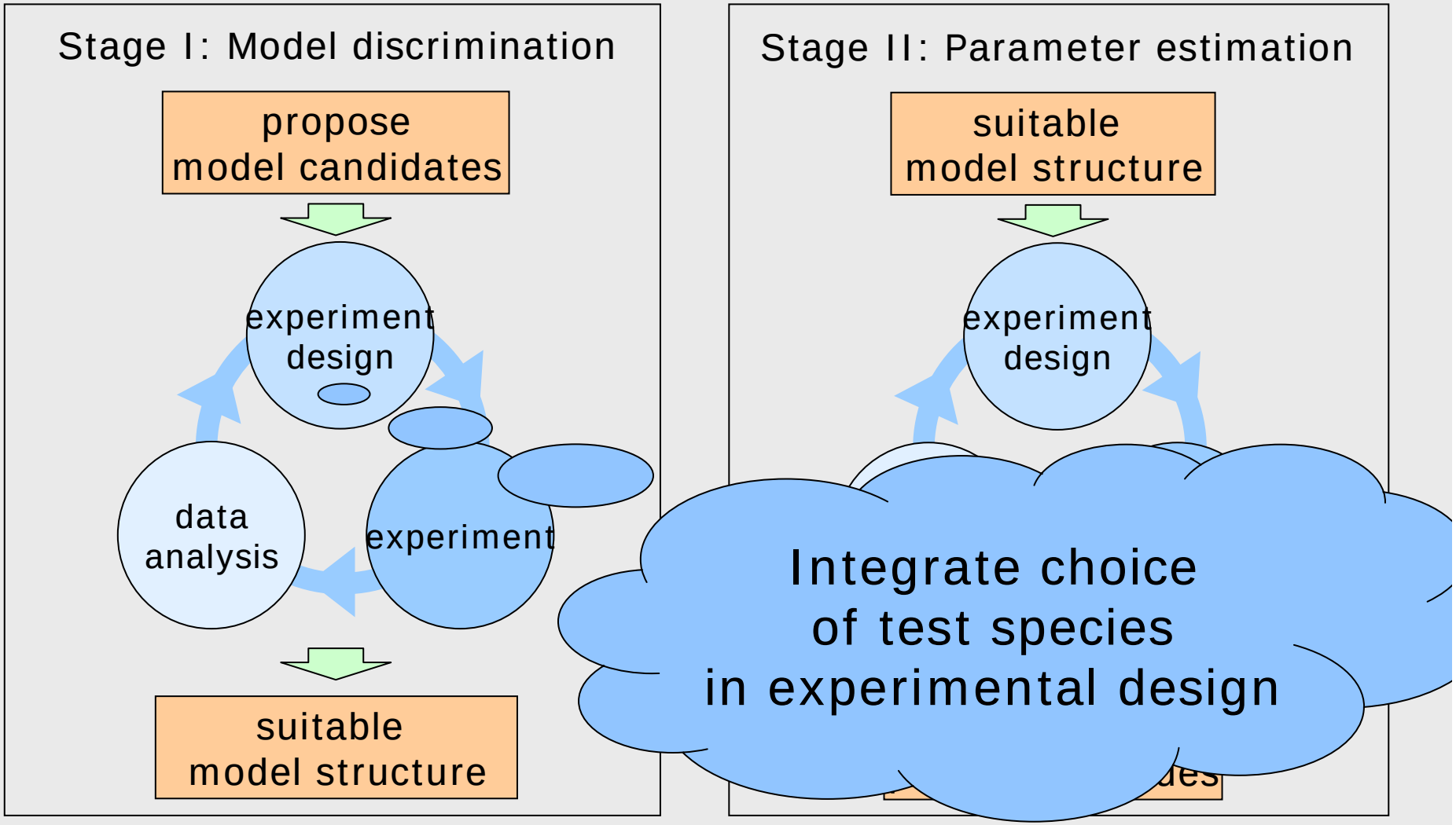
- *METER* criterion captures bias-variance trade-off crucial in the solution of ill-posed problems
- classical methods are even qualitatively wrong
- extension to nonlinear problems is straightforward by linearization and local analysis as in standard theory
- regularization parameter λ incorporates naturally
- main drawback: initial guess of unknown function f required $\Rightarrow$ robust METER design formulations

Predictive physical property models



Which species should be tested
to develop predictive property models?

Experimental design stages



Example: Multicomponent Diffusion

Fick | Maxwell-Stefan | Thermo

$$\underbrace{\begin{bmatrix} D_{11}(\boldsymbol{x}) & D_{12}(\boldsymbol{x}) \\ D_{21}(\boldsymbol{x}) & D_{21}(\boldsymbol{x}) \end{bmatrix}}_{\text{measured}} = f\left(\underbrace{\begin{bmatrix} - & Đ_{12}(\boldsymbol{x}) & Đ_{13}(\boldsymbol{x}) \\ Đ_{21}(\boldsymbol{x}) & - & Đ_{23}(\boldsymbol{x}) \\ Đ_{31}(\boldsymbol{x}) & Đ_{32}(\boldsymbol{x}) & - \end{bmatrix}}_{\text{theory and predictive models}} , \underbrace{\Gamma(\boldsymbol{x})}_{\text{poorly known}} \right)$$

Example: Multicomponent Diffusion

Fick — Maxwell-Stefan — Thermo

$$\begin{bmatrix} D_{11}(x) & D_{12}(x) \\ D_{21}(x) & D_{21}(x) \end{bmatrix} = f\left(\begin{bmatrix} - & Đ_{12}(x) & Đ_{13}(x) \\ Đ_{21}(x) & - & Đ_{23}(x) \\ Đ_{31}(x) & Đ_{32}(x) & - \end{bmatrix}, \Gamma(x) \right)$$

$$D_{11}^{x_1 \to 0} = \left(\frac{x_2}{(Đ_{12}^{0})^{x_2} (Đ_{12}^{x_3 \to 1})^{x_3}} + \frac{x_3}{(Đ_{13}^{0})^{x_3} (Đ_{13}^{x_2 \to 1})^{x_2}} \right)^{-1} = \hat{D}_{1}^{x_1 \to 0}$$

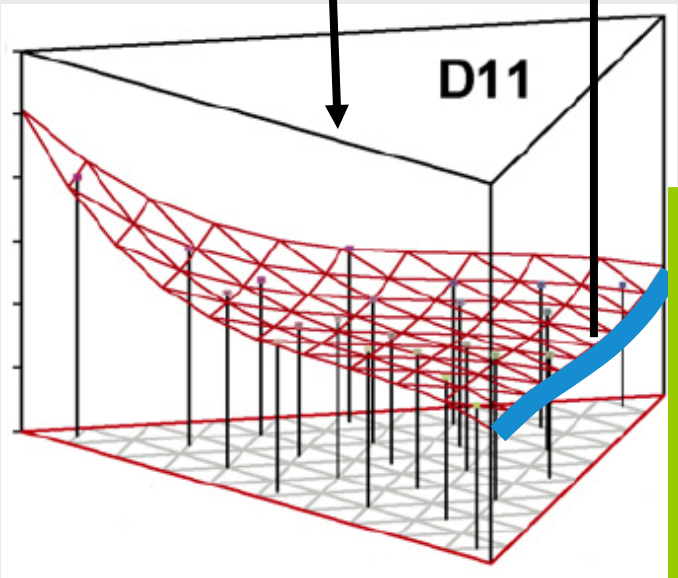
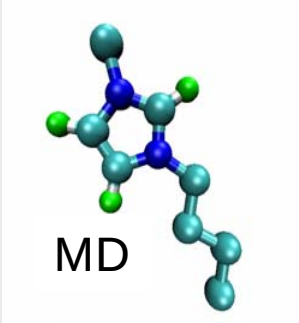


Diagram from Rehfeldt et al. (2007)

Predictive models for $Đ_{ij}^{x_k \to 1}$

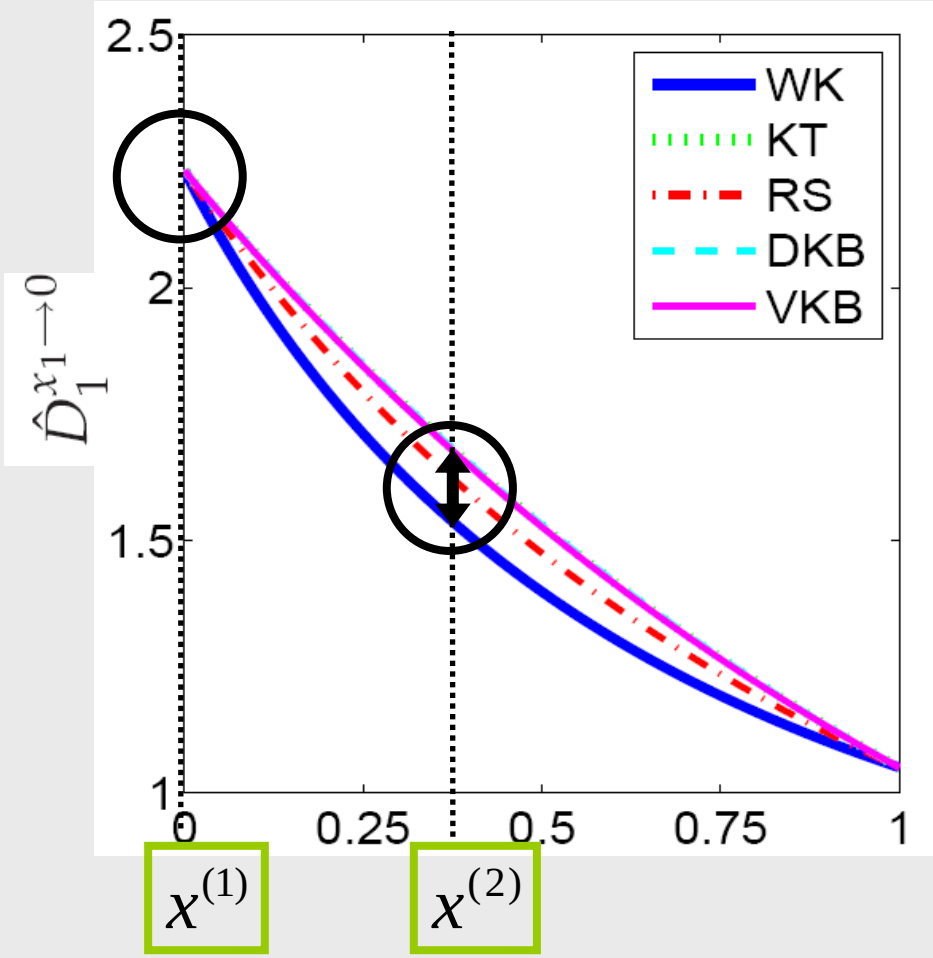
- Wesselingh-Krishna (**WK**)
- Kooijman-Taylor (**KT**)
- Rehfeldt-Stichlmair (**RS**)
- Vignes-Krishna-van Baten (**DKB**)
- Darken-Krishna-van Baten (**VKB**)



(Bardow et al., *Fluid Phase Equlibria.* 2009, Liu et al, *Fluid Phase Equilibria*, 2011)

Classical OED for Model Discrimination

Question: At which **concentration** should we measure?



bad choice $x^{(1)}$

since

$$y^M(x^{(1)}, \theta^M) = y^H(x^{(1)}, \theta^H)$$

Hunter-Reiner criterion (1965):

maximize difference between model responses

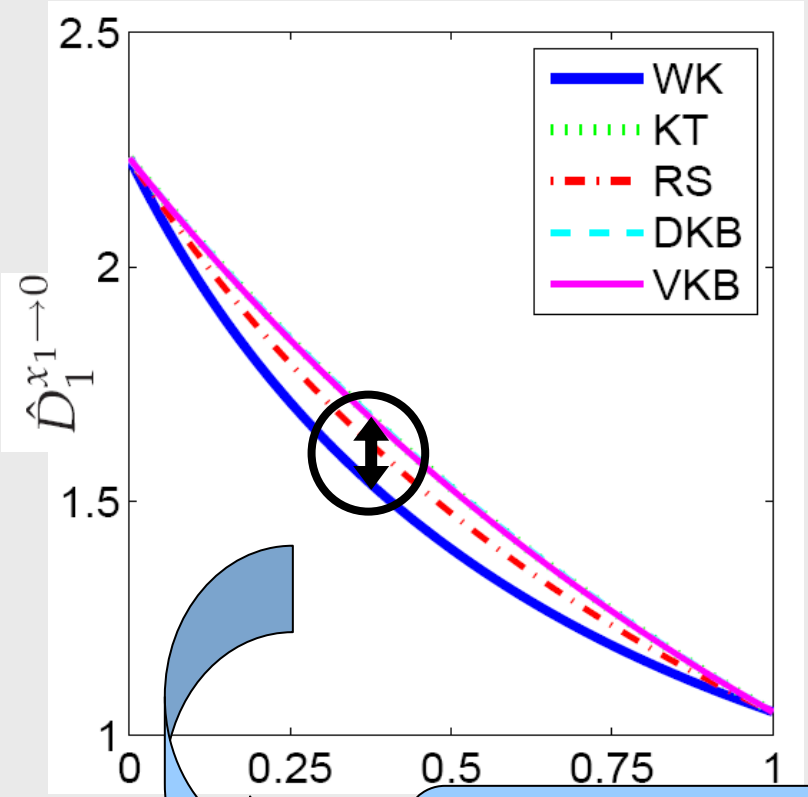
$$\max_x \left[y^M(x, \theta^M) - y^H(x, \theta^H) \right]^2$$

here: Buzzi-Ferraris & Forzatti (1983)

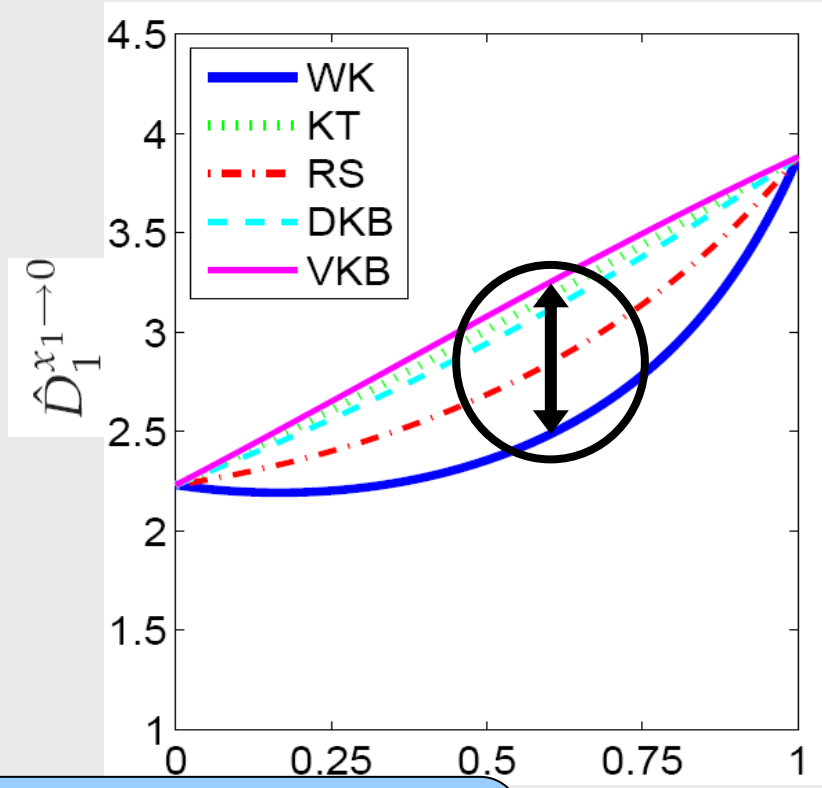
Molecular design for Model Discrimination

Question: Which **chemical system** should be selected?

Cyclohexane + 1,4-Dioxane + Toluene



Cyclohexane + n-Hexane + Toluene



look for molecule
that allows for maximum
model discrimination

Chemical Product Design

Systematic method to develop new or novel products.

Define needs

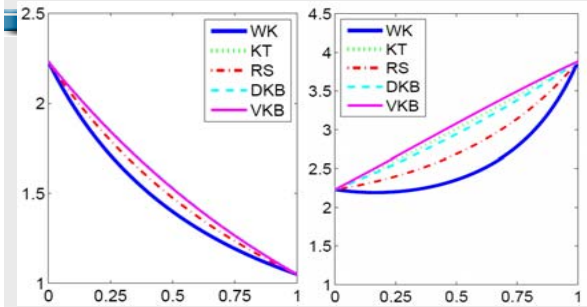
Generate ideas to meet these needs

Screen and select the best ideas

Decide how the product is manufactured

Cussler and Moggridge, 2001

Optimal Experiment Design as Product Design Problem



Define needs

Computer-aided molecular design

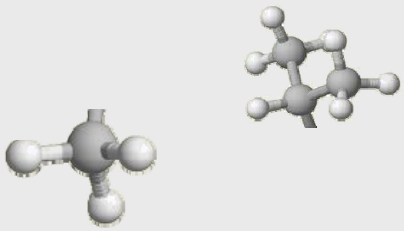
Experiment and model analysis

Generate ideas to meet these needs

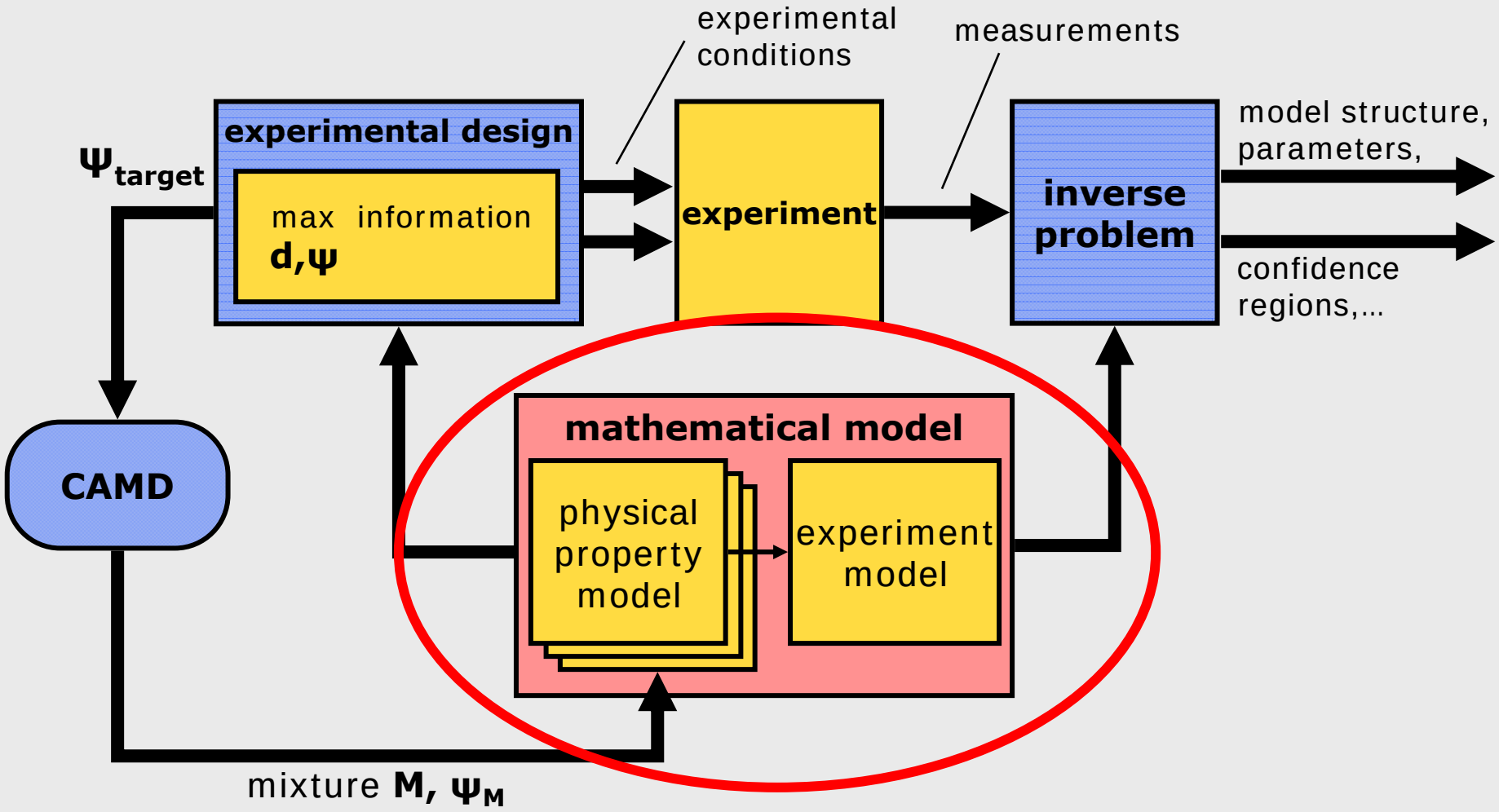
Screen and select the best ideas

Combine Molecular Building Blocks

Group-Contribution Methods & Objective function



Combined OED-CAMD work process



(Bardow, Kossack, Kriesten, Marquardt, *Computer-Aided Chemical Engineering*, 2008)

MEXA-CAMD for Diffusion Case Study

Binary Input data

$$Đ^0_{ij}$$

physical property model

$$\sqrt{Đ^0_{ij}Đ^0_{ji}} = Đ^{x_k\to 1}_{ij}$$

$$\sqrt{Đ^0_{ik}Đ^0_{ki}} = Đ^{x_k\to 1}_{ij}$$

$$\left(Đ^0_{ik}Đ^0_{jk}Đ^0_{ij}Đ^0_{ji}\right)^{1/4} = Đ^{x_k\to 1}_{ij}$$

$$\left(Đ^0_{ik}\right)^{\frac{x_i}{x_i+x_j}}\left(Đ^0_{jk}\right)^{\frac{x_j}{x_i+x_j}} = Đ^{x_k\to 1}_{ij}$$

$$\frac{x_i}{x_i+x_j}Đ^0_{jk} + \frac{x_j}{x_i+x_j}Đ^0_{ik} = Đ^{x_k\to 1}_{ij}$$

experiment model

$$\sum_{j=1;j\neq i}^{n_c}\frac{x_j}{\left(D^0_{ij}\right)^{x_j}\left(D^{x_k\to 1}_{ij}\right)^{x_k}} = \hat{D}^{x_i\to 0}_i$$

$$\hat{D}^{x_i\to 0}_i$$

$$f^m(\boldsymbol{\psi}^m, \boldsymbol{d}^m) = \hat{D}^{x_i\to 0}_i$$

$$\boldsymbol{\psi}^m = \left[Đ^{0,m}_{ij}, Đ^{0,m}_{ji}, Đ^{0,m}_{ik}, Đ^{0,m}_{ki}, Đ^{0,m}_{jk}, Đ^{0,m}_{kj}\right]$$

$$\boldsymbol{d}^m = \left[\boldsymbol{x}^m_i, \boldsymbol{x}^m_j, \boldsymbol{x}^m_k\right]$$

Continuous property targeting in OED

$$\hat{D}_i^{x_i \to 0} = f^m(\boldsymbol{\psi}^m, \boldsymbol{d}^m)$$

$$\begin{aligned} \boldsymbol{\psi}^m &= \left[Đ_{ij}^{0,m}, Đ_{ji}^{0,m}, Đ_{ik}^{0,m}, Đ_{ki}^{0,m}, Đ_{jk}^{0,m}, Đ_{kj}^{0,m}\right] \\ \boldsymbol{d}^m &= \left[\boldsymbol{x}_i^m, \boldsymbol{x}_j^m, \boldsymbol{x}_k^m\right] \end{aligned}$$

Step 1: Relaxed MEXA-OED problem

$$\max_{\boldsymbol{\psi},\boldsymbol{d}} \sum_{m=1}^{M} \sum_{m'=m+1}^{M} \left(\boldsymbol{f}^m - \boldsymbol{f}^{m'}\right)^T \boldsymbol{V}^{-1} \left(\boldsymbol{f}^m - \boldsymbol{f}^{m'}\right) \longrightarrow \boldsymbol{\psi}^{OPT}$$

continuous search in property space → standard NLP problem

Step 2: CAMD

$\boldsymbol{\psi}^{OPT}$ → **CAMD** → real mixtures **M, $\boldsymbol{\psi}_M$**

CAMD targets from OED

Goal: Find **third component** in **Cyclohexane+Toluene+?** allowing for optimum model discrimination

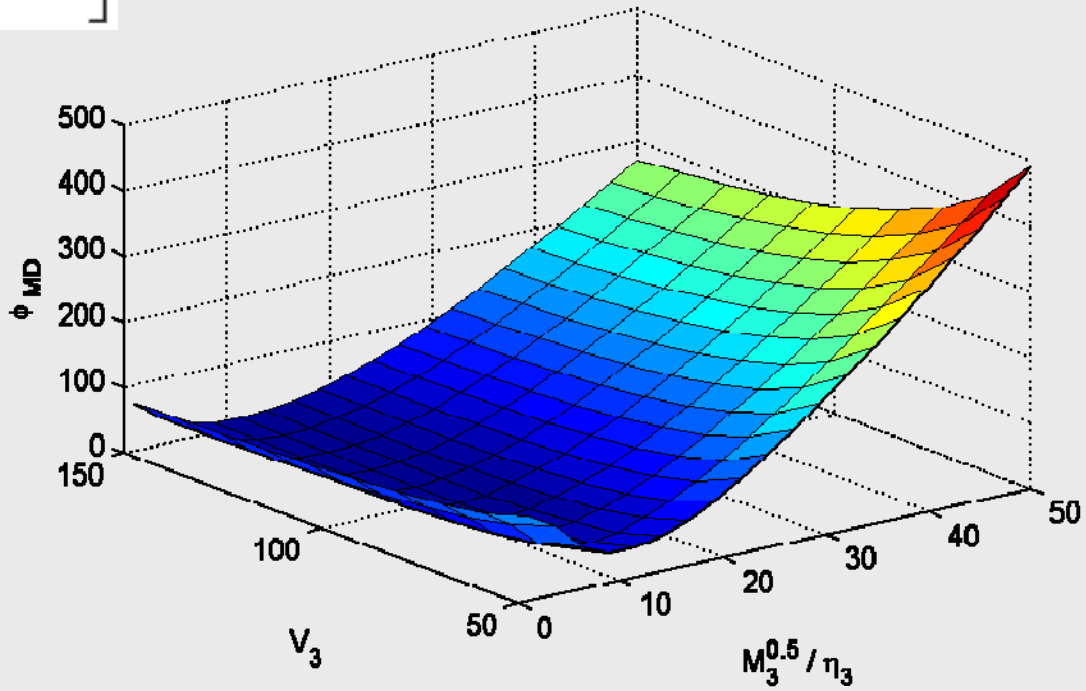
⇨ find $\psi^{OPT} = \left[Đ^0_{13}, Đ^0_{31}, Đ^0_{23}, Đ^0_{32}\right]$

Problem can be further reduced using Wilke-Chang

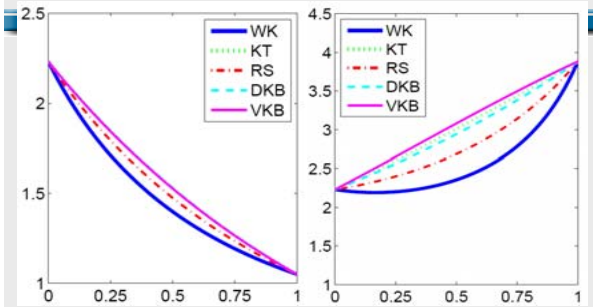
$$Đ^0_{ij} = \alpha \frac{\sqrt{M_j}T}{\eta_j (V_i)^{0.6}}$$

⇨ find $\psi^{OPT} = \left[\frac{\sqrt{M_3}}{\eta_3}, V_3\right]$

targets for CAMD algorithm



Mixture OED as Product Design Problem



Define needs

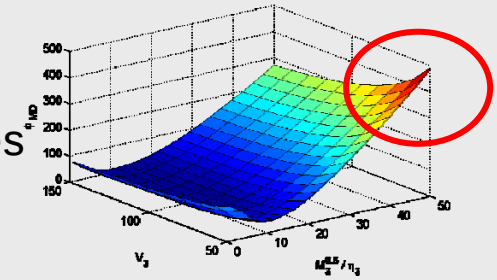
Computer-aided molecular design

Generate ideas to meet these needs

Screen and select the best ideas

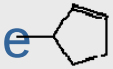
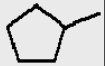
Experiment and model analysis

Software package used: ICAS (Gani et al.)
Includes Group-contribution methods and rules to generate only meaningful molecules



CAMD results

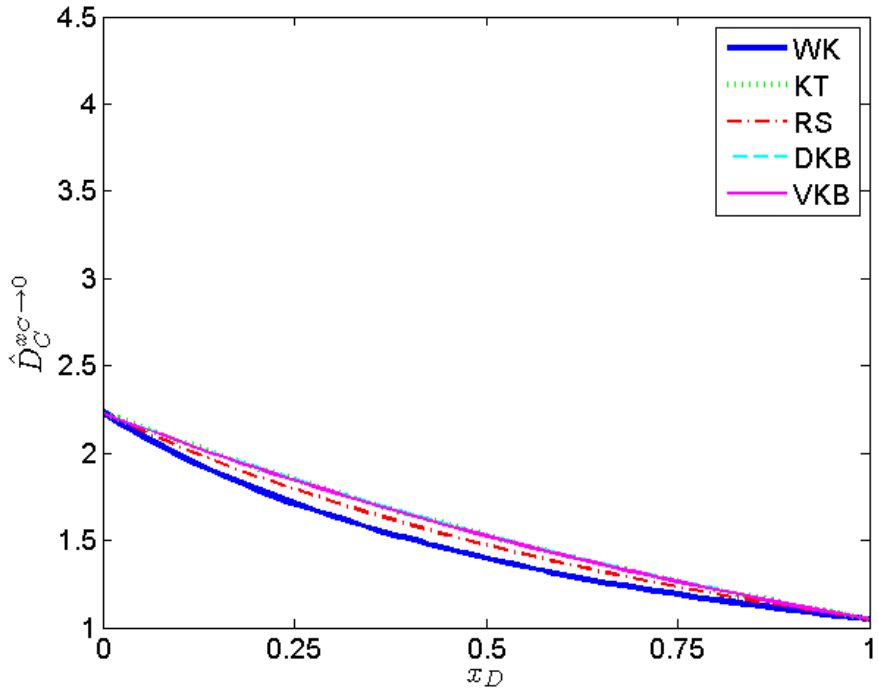
Goal: Find **third component** in **Cyclohexane**+**Toluene+?**
maximal discrimination of property model candidates

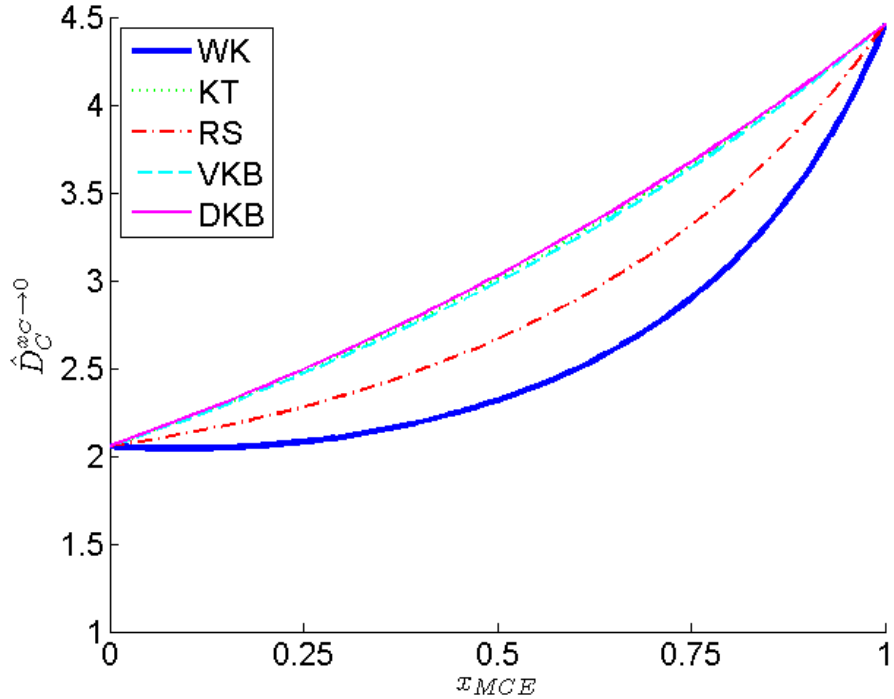
Component name	**M (g/mol)**	**η (cP)**	**V_m (cm^3/mol)**	**MD objective**
3-methyl-cyclopentene 	82.14	0.27	119.57	130
methyl-cyclopentane 	84.16	0.41	120.69	46
cyclohexene 	82.14	0.54	108.04	22

factor six in experimental effort
between top 3 molecules

CAMD results

Goal: Find **third component** in **Cyclohexane+Toluene+?**
maximal discrimination of property model candidates



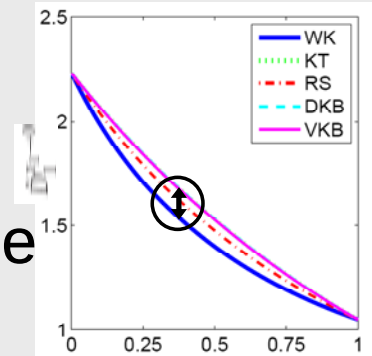


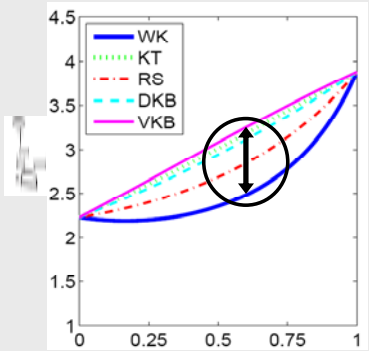
application to NMR measurements and molecular simulations

- none of the existing models works in a robust fashion
- theoretically-based model derived and validated

(Bardow et al., *Fluid Phase Equlibria.* 2009, Liu et al, *Fluid Phase Equilibria*, 2011, Liu et al., *Ind. Eng. Chem. Res.*, 2011)

CAMD for OED





- add test molecules and mixture to experimental design
- integration CAMD into OED work process
- extends to all OED problems

$\Rightarrow$ reduction in experimental effort

$\Rightarrow$ quantitative assessment of model validity

Summary

- OED leads non-intuitive designs

  reducing experimental effort by orders
- software appearing for 1D PDAE systems
- work on ill-posed problems has started
- seamless integration of mixture selection possible

$\Rightarrow$ OED mandatory in model-based experimentation
$\Rightarrow$ comprehensive methodological toolbox available