Large scale Bayesian Uncertainty Quantification in Molecular Dynamics simulations

Panagiotis Angelikopoulos

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

With

Dr. Stephen Wu Dr. Panagiotis Hadjidoukas

Prof. Costas Papadimitriou
@University of Thessaly

ETH Zürich

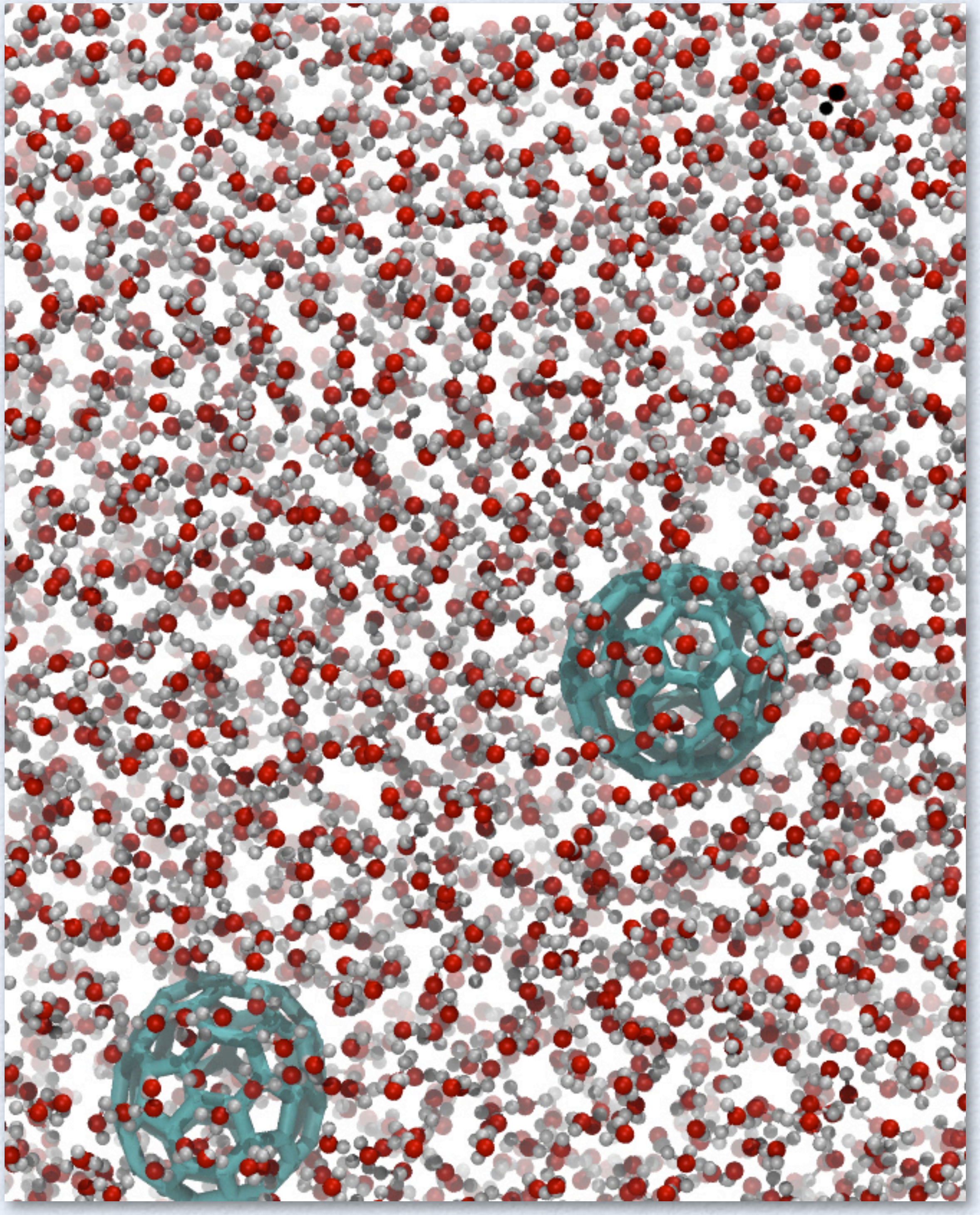
- Prof. Robert Moser (UT Austin)
- Prof. Stacey Finley (USC)

Prof. Petros Koumoutsakos

Overview

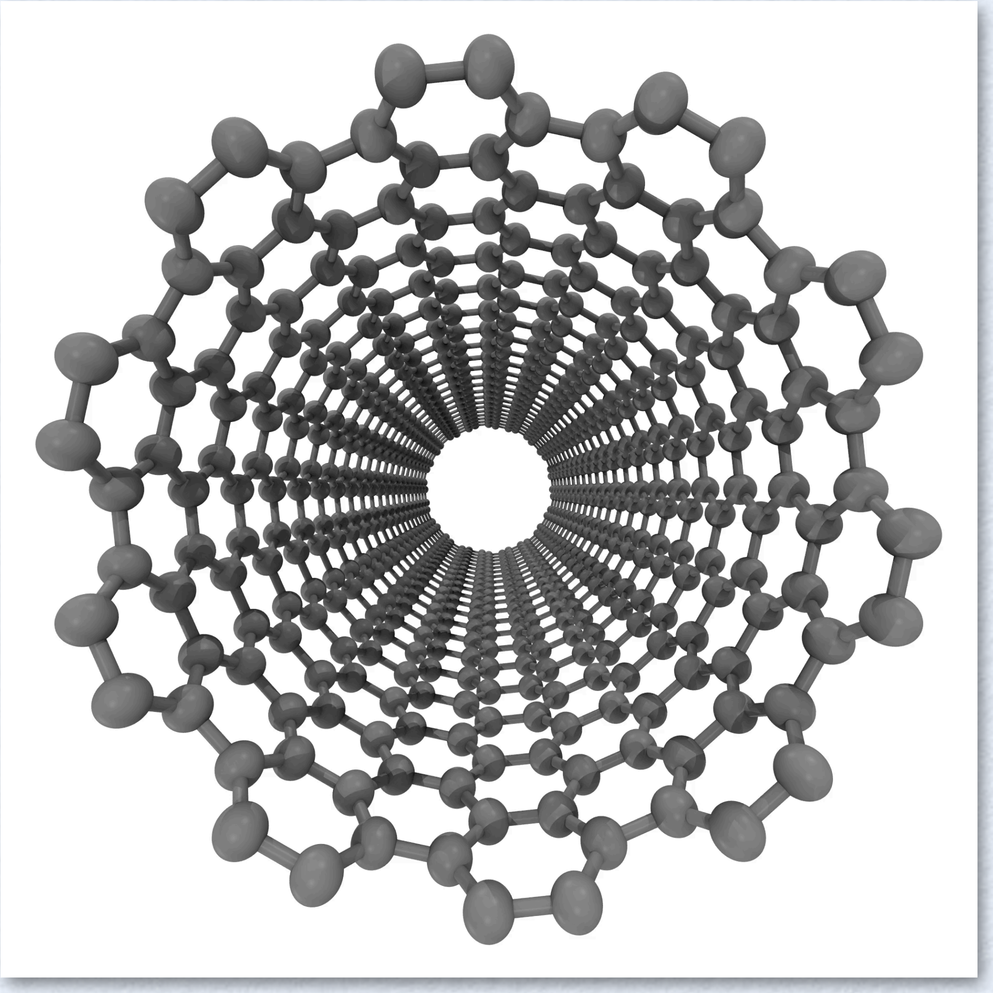
- **Motivation**
  - Molecular Dynamics (MD) - sources of uncertainty
  - MD and heterogeneous data fusion
- **Bayesian UQ in MD**
  - Algorithms
  - Graphical probability models for Bayesian updating
    - *Pooled data* - Water/Graphite interactions
    - *Structured data* - Argon
    - *Heterogeneous data* - Water
  - HPC software - Pi4U
- **Summary and Outlook**

Molecular Dynamics Simulations

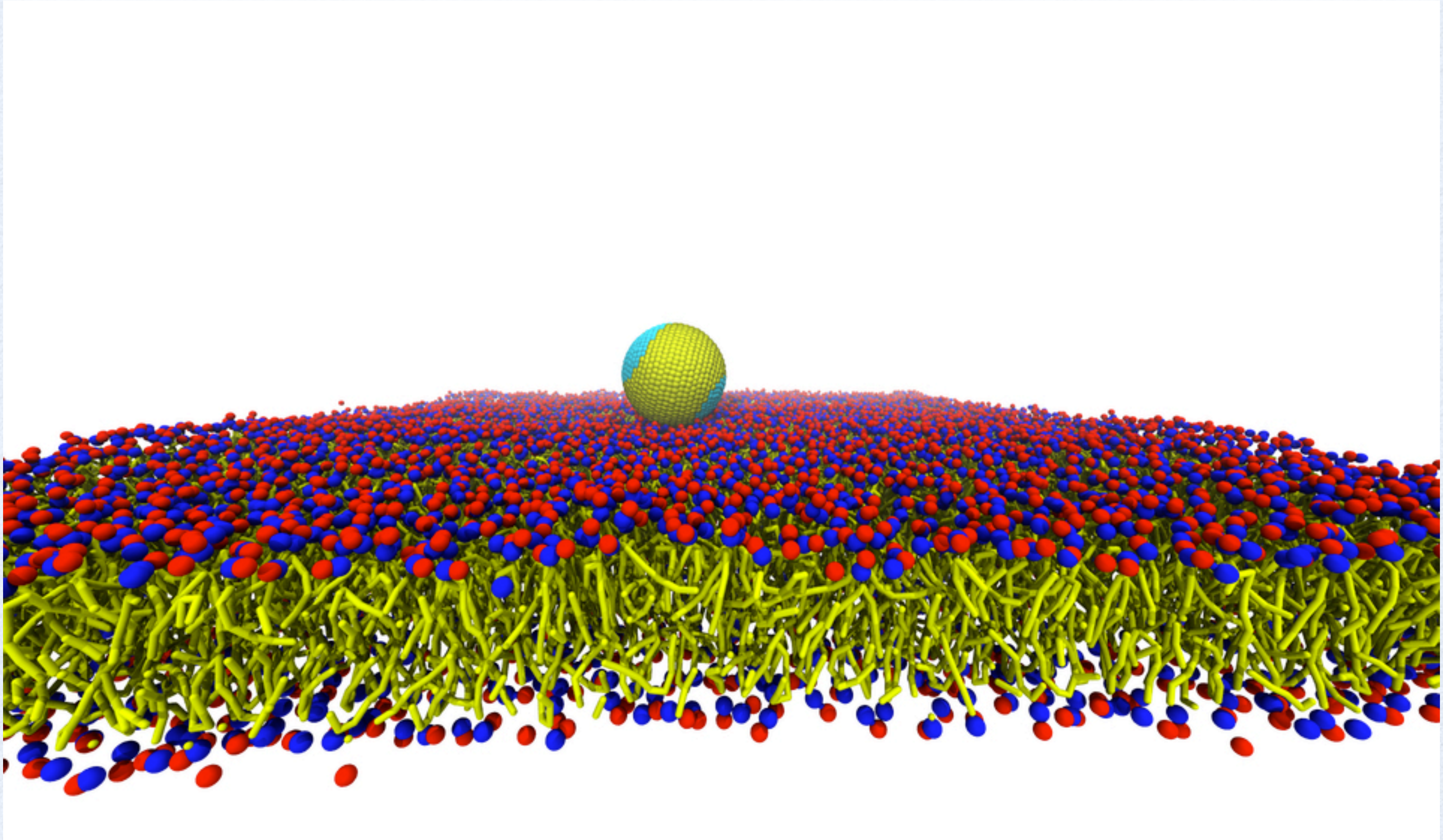


$$\frac{d\mathbf{x_p}}{dt} = \mathbf{u_p}$$

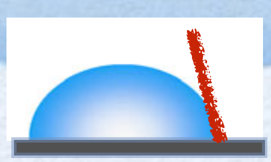
$$m\frac{d\mathbf{u_p}}{dt} = \sum_i^N \mathbf{F_p} \equiv -\frac{\partial U}{\partial \mathbf{x}_p}$$

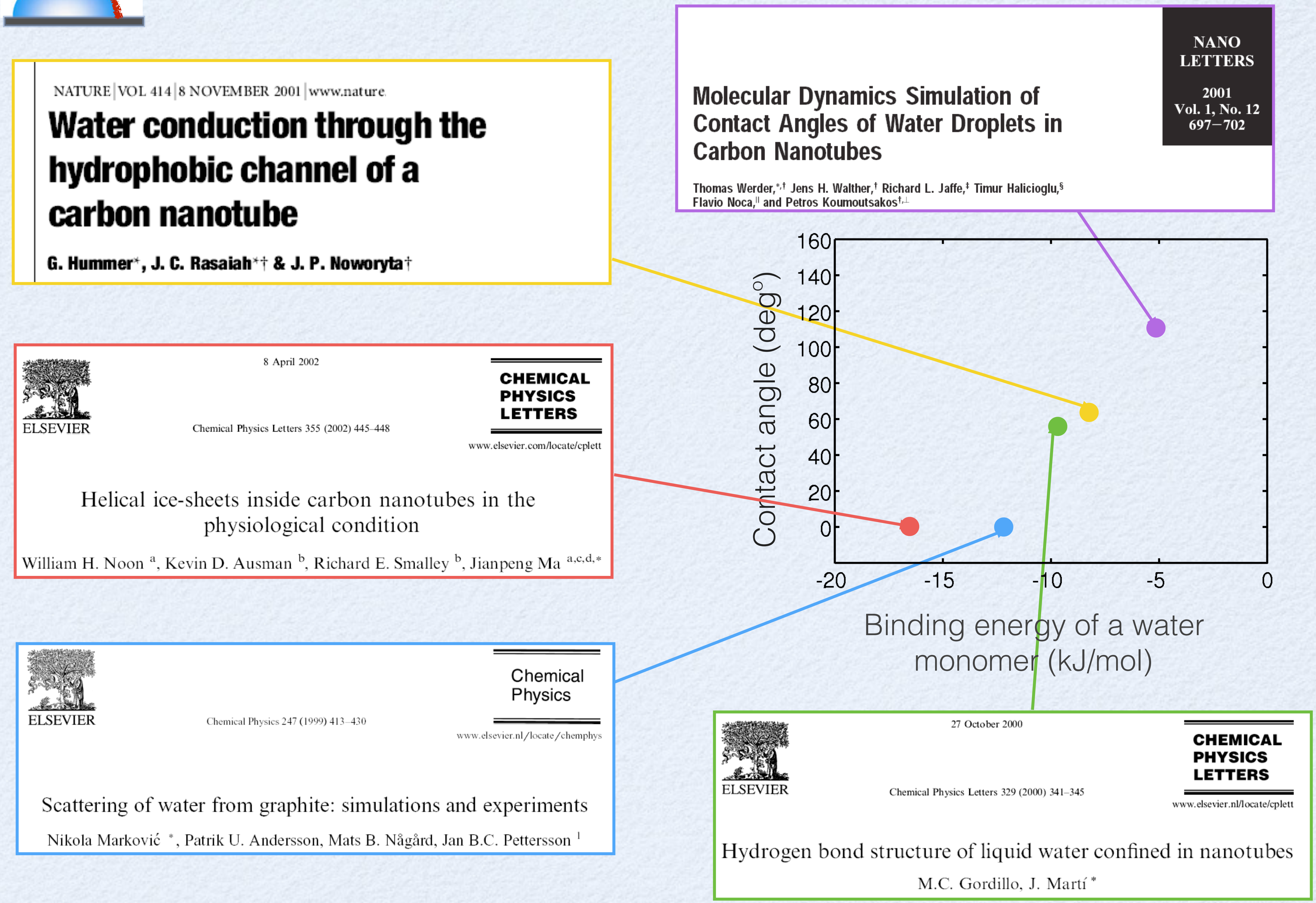


Molecular Dynamics Simulations



Wetting Dependence on MD Potentials





NATURE | VOL 414 | 8 NOVEMBER 2001 | www.nature

Water conduction through the hydrophobic channel of a carbon nanotube

G. Hummer*, J. C. Rasaiah*† & J. P. Noworyta†

NANO LETTERS 2001 Vol. 1, No. 12 697–702

Molecular Dynamics Simulation of Contact Angles of Water Droplets in Carbon Nanotubes

Thomas Werder,*,† Jens H. Walther,† Richard L. Jaffe,‡ Timur Halicioglu,§ Flavio Noca,|| and Petros Koumoutsakos†,⊥

ELSEVIER 8 April 2002 Chemical Physics Letters 355 (2002) 445–448 CHEMICAL PHYSICS LETTERS www.elsevier.com/locate/cplett

Helical ice-sheets inside carbon nanotubes in the physiological condition

William H. Noon [a], Kevin D. Ausman [b], Richard E. Smalley [b], Jianpeng Ma [a,c,d,*]

ELSEVIER Chemical Physics 247 (1999) 413–430 Chemical Physics www.elsevier.nl/locate/chemphys

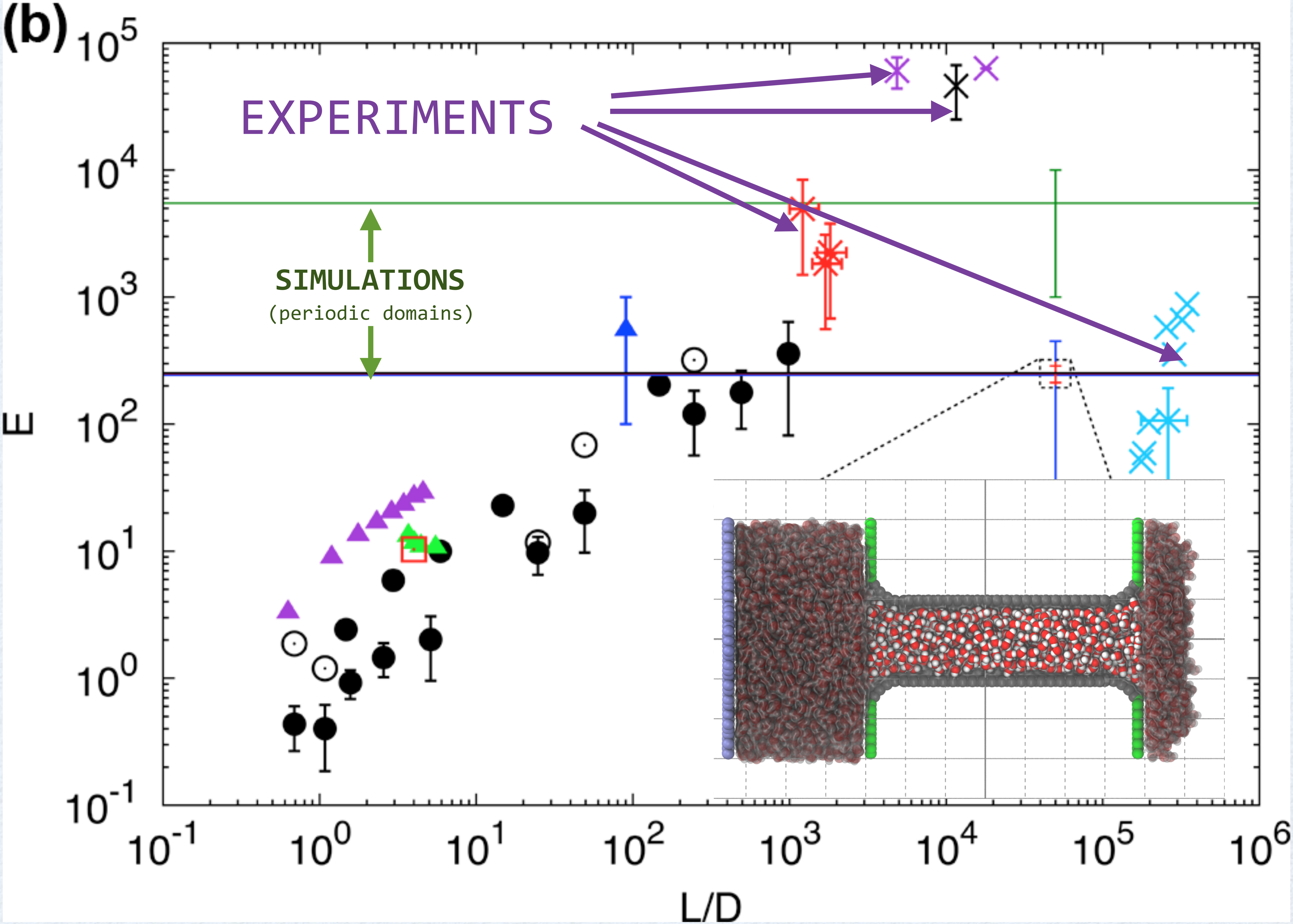
Scattering of water from graphite: simulations and experiments

Nikola Marković *, Patrik U. Andersson, Mats B. Någård, Jan B.C. Pettersson [1]

ELSEVIER 27 October 2000 Chemical Physics Letters 329 (2000) 341–345 CHEMICAL PHYSICS LETTERS www.elsevier.nl/locate/cplett

Hydrogen bond structure of liquid water confined in nanotubes

M.C. Gordillo, J. Martí *

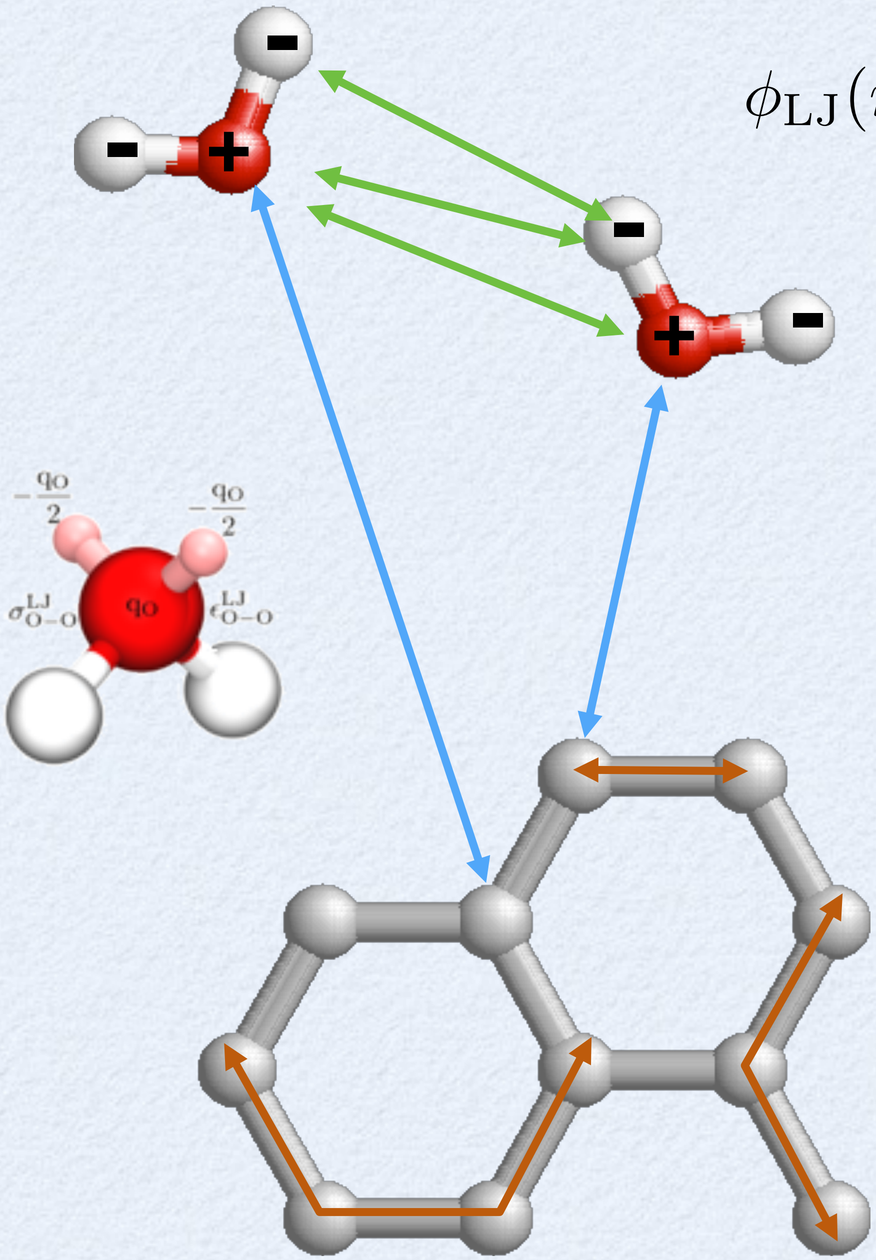


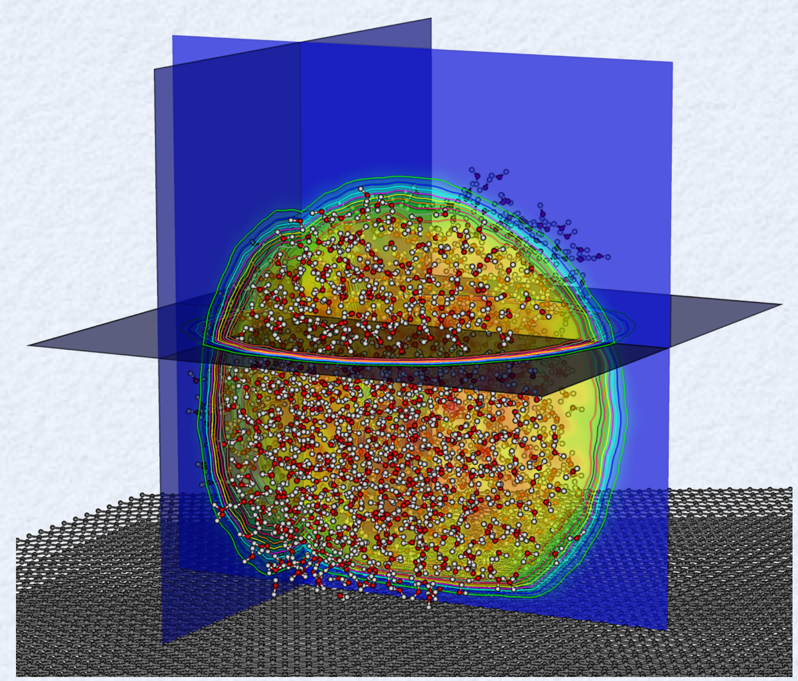
Walther JH, Ritos K, Cruz-Chu E, Megaridis CM, Koumoutsakos P. **Nano Letters**, 2013
Angelikopoulos P. Papadimitriou C., Koumoutsakos P. **JPC B** 2013,

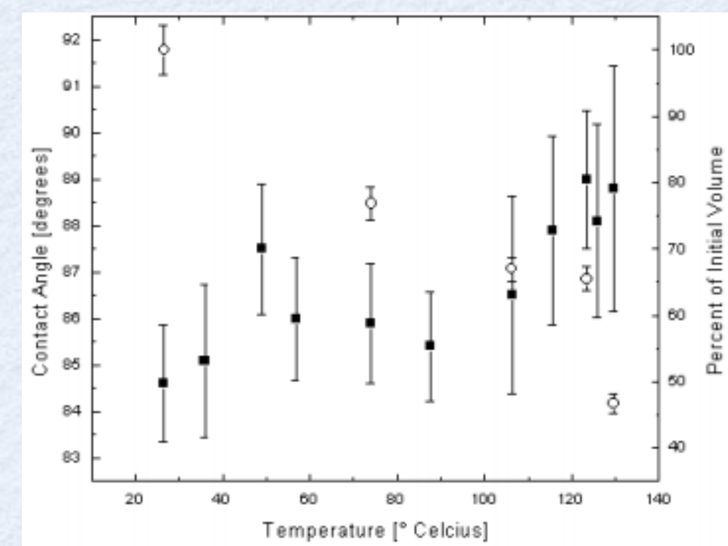
Sources of Uncertainty in Water-Graphite Systems

MODELLING **PARAMETRIC** **COMPUTATIONAL** **MEASUREMENT**

$$\phi_{\mathrm{LJ}}(r_{ij}) = 4\epsilon_{\mathrm{LJ}}\left[\left(\frac{\sigma_{\mathrm{LJ}}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{\mathrm{LJ}}}{r_{ij}}\right)^{6}\right]$$



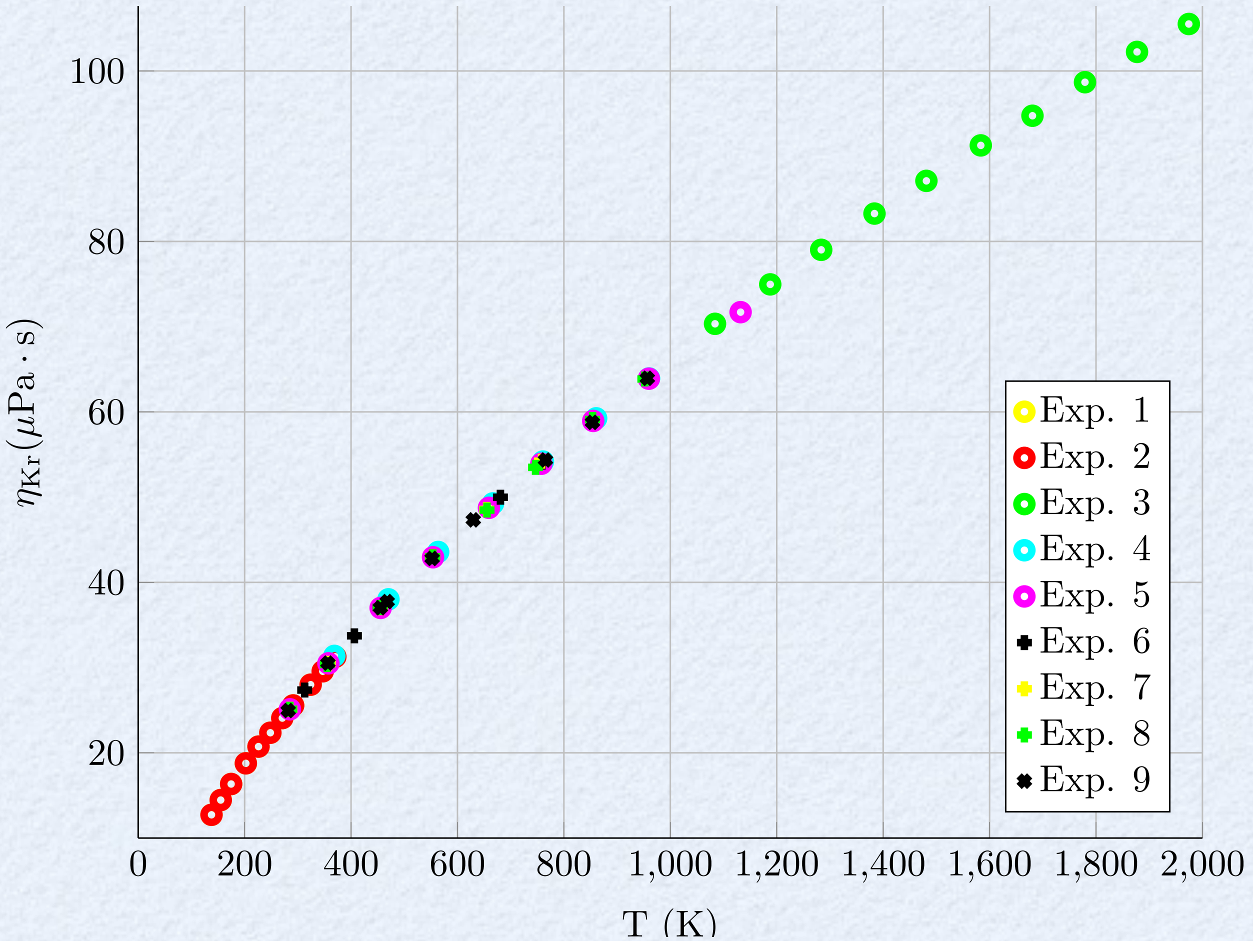




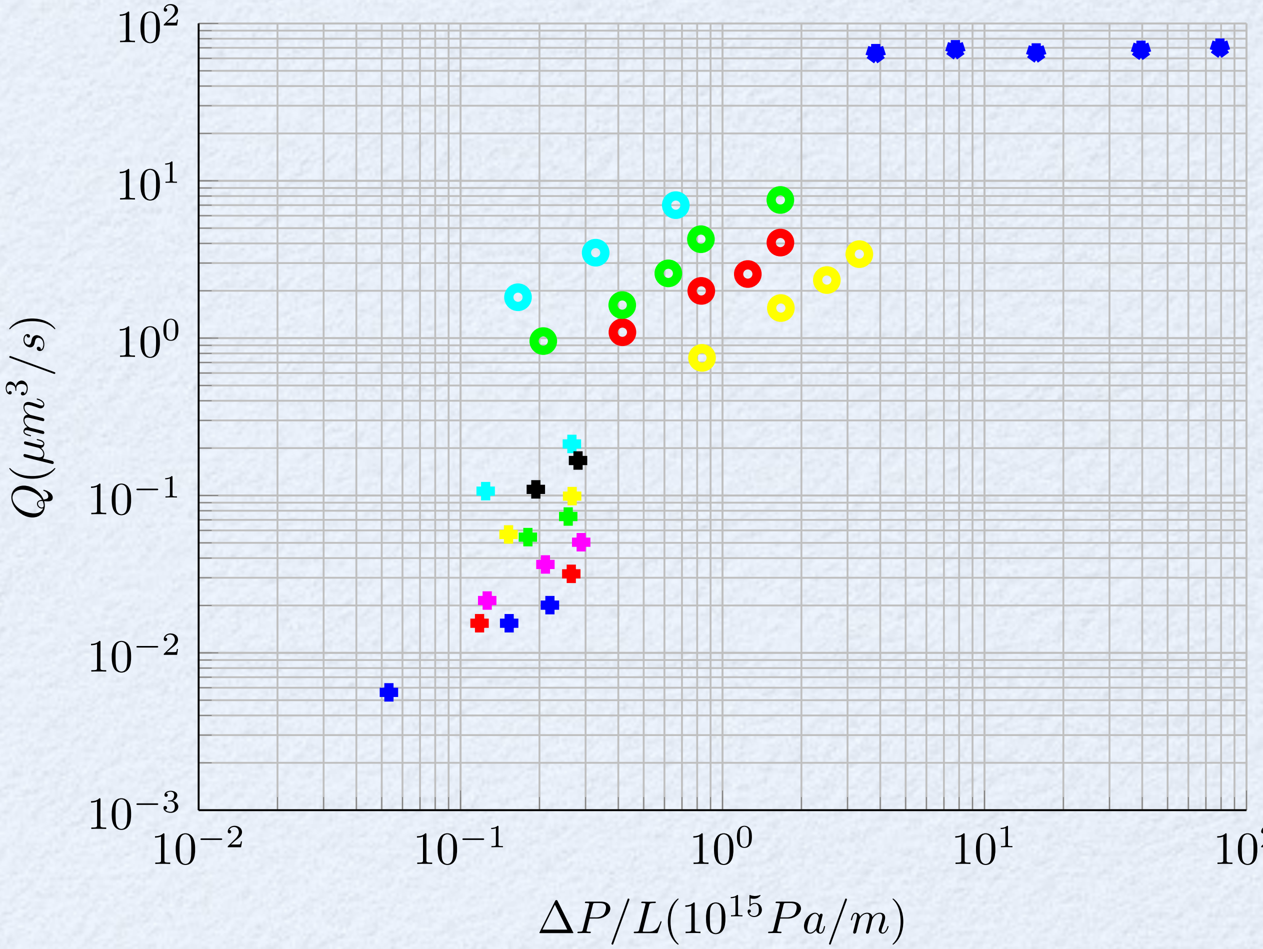
K. Osborne III (2009)

MD Data is Structured

Structure in the data and predictions



Argon viscosity



water flow rate inside carbon nanotubes

MD predictions are Heterogeneous

> As stated above, the goal was to develop the simplest potential function, which reproduces well the density anomaly of liquid water, while simultaneously yielding good thermodynamic and structural properties near 25 °C and 1 atm. Additional studies of three- and four-site models, in-

Mahoney & Jorgensen, J. Phys. Chem. 2000

Multiple Objective Functions

- Pareto front
- Weighted sum of objective functions

How to choose weights?

“Everybody trusts an experiment, but the person that did it.

No-one believes *a simulation* but the person that did it

Why not combine the two and get results everybody can mistrust a little?”

Tony Kordyban

Bayesian Inference and Uncertainty Quantification

Embedding the model in a stochastic model class

From deterministic

$$\theta \in \mathcal{R}^n \longrightarrow \mathcal{MD} \longrightarrow q\left(\theta | \mathcal{MD}\right)$$

Model class

Quantities of Interest (QoI)

To Stochastic

$$\theta \in \mathcal{R}^n \longrightarrow \mathcal{MD} \xrightarrow{q\left(\theta | \mathcal{MD}\right)} f\left(q | \theta, \mathcal{MD}\right)$$

QoI

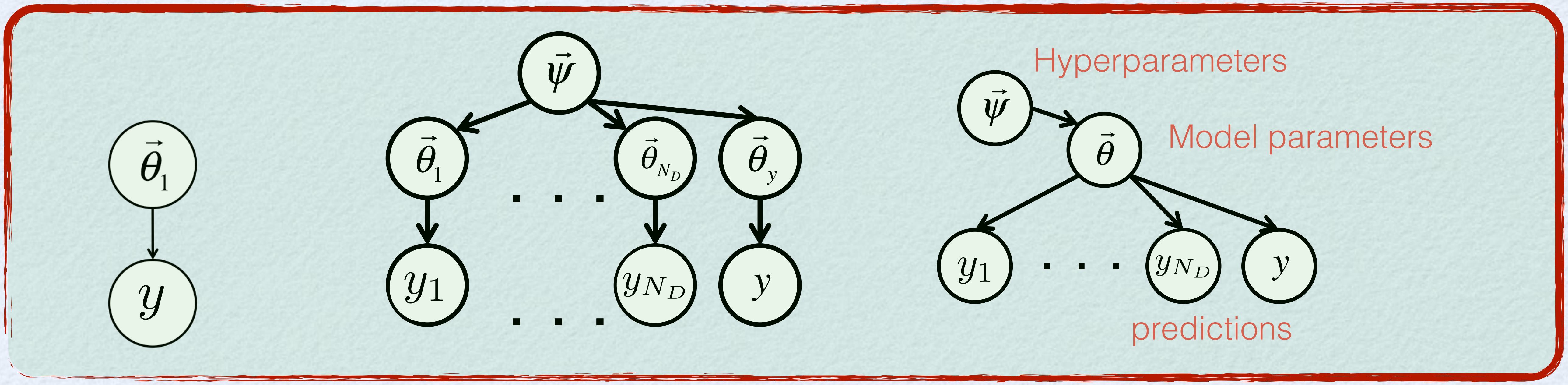
Conditional PDF

Stochastic Model class

Describing the stochastic model class : Graphical probabilistic models

Graph based models of a **joint probability distribution**

A way to describe **conditional independencies**- valid for any type of distributions (not just Gaussians)



e.g $p\left(\vec{\theta_1}, \vec{\theta_2} | \vec{\psi}\right) = p\left(\vec{\theta_1} | \vec{\psi}\right) \cdot p\left(\vec{\theta_2} | \vec{\psi}\right)$

Types : 1) directed acyclic graphs (Bayesian networks)
2) undirected graphs (Markov Random Fields)

Bayesian UQ : Calibration and Model Selection

Experimental Data: D

Use observations to select the model classes and estimate their parameter values such that the model predictions best fit the data

PARAMETER ESTIMATION

$$f\left(\theta_i|D,\mathcal{MD}_i\right) = \frac{f\left(D|\theta_i,\mathcal{MD}_i\right)\pi\left(\theta_i|\mathcal{MD}_i\right)}{f\left(D|\mathcal{MD}_i\right)}$$

Experiments

Physical limitations
Past studies
Expert elicitation

MODEL CLASS SELECTION

$$Pr\left(\mathcal{MD}_i|D\right) = \frac{f\left(D|\mathcal{MD}_i\right)Pr\left(\mathcal{MD}_i\right)}{f\left(D\right)}$$

Evidence of Model

$$f\left(D|\mathcal{MD}_i\right) = \int f\left(D|\theta_i,\mathcal{MD}_i\right)\pi\left(\theta_i|\mathcal{MD}_i\right)d\theta_i$$

Bayesian Uncertainty Propagation

$$\theta \in \mathcal{R}^n \longrightarrow \boxed{\mathcal{MD}} \xrightarrow{q\left(\theta|\mathcal{MD}\right)} f\left(q|\theta, \mathcal{MD}\right)$$

Model class

QoI

Conditional PDF

QUANTITIES OF INTEREST: **Posterior Robust Predictions: PDF**

$$f\left(q|D, \mathcal{MD}\right) = \int \underbrace{f\left(q|\theta, \mathcal{MD}\right)}_{\text{Conditional PDF}} \underbrace{f\left(\theta|D, \mathcal{MD}\right)}_{\text{Posterior PDF}} d\theta$$

Sample Estimate:

$$f\left(q|D, \mathcal{MD}\right) = \frac{1}{N} \sum_{i=1}^{N} \underbrace{f\left(q|\theta^{(i)}, \mathcal{MD}\right)}_{\text{Conditional PDF}}$$

$$\theta^{(i)} \sim f\left(\theta|D, \mathcal{MD}\right)$$

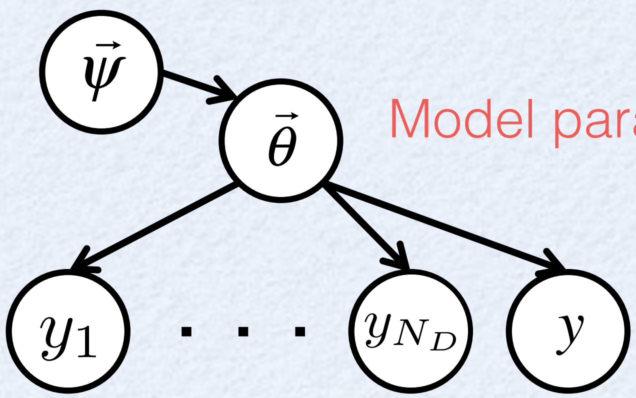
Samples drawn from Posterior PDF

Basic data structure : pooled data

DISCREPANCY BETWEEN MODEL PREDICTIONS AND DATA

$$\boxed{\begin{aligned}\hat{y} &= g(\theta|\mathcal{MD}) + \mathrm{e}\\ e &= e^d + e^f + e^m\end{aligned}}$$

0



Model parameters

predictions

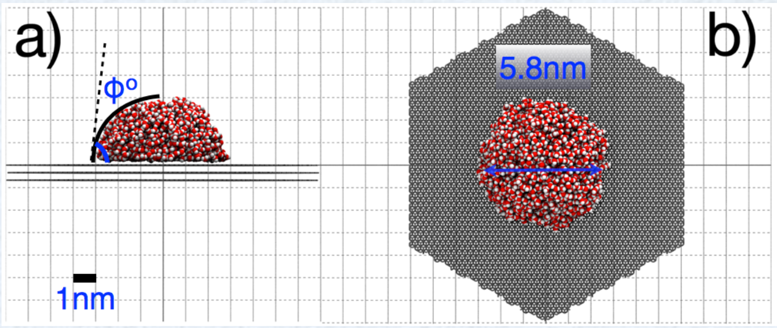
- **Measurement Error** $e^d \sim N\left(\mu^d, \Sigma^d\right)$ $\Sigma^d = diag\left(s_r^2 \hat{y_r}^2\right)$
- **Computational Error** $e^f \sim N\left(\mu^f, \Sigma^f\right)$ $\Sigma^f = diag\left(\sigma_r^2 \hat{y_r}^2\right)$
- **Model Error** $e^m \sim N\left(\mu^m, \Sigma^m\right)$ $\Sigma^m = diag\left(\lambda_r^2 \hat{y_r}^2\right)$

LIKELIHOOD: Assume independent model prediction errors $\Sigma(\theta_e) = \Sigma^d + \Sigma^f + \Sigma^m$

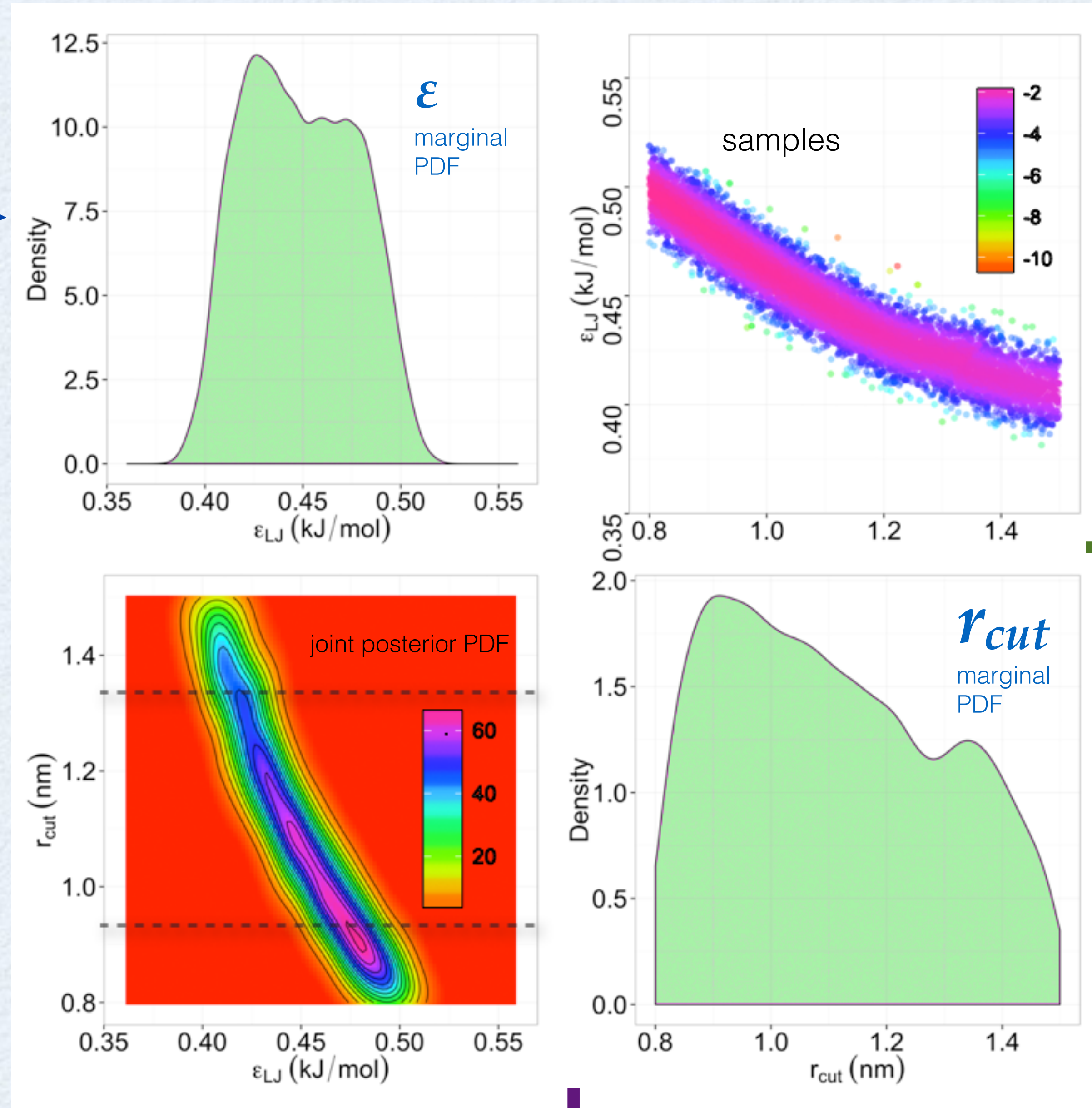
$$p\left(D|\theta, \mathcal{MD}\right) = \frac{|\Sigma|^{-1/2}}{(2\pi)^{m/2}} exp\left[-\frac{1}{2}\left[\hat{y} - f\left(\theta_m|\mathcal{MD}\right) - \mu\left(\theta_e\right)\right]^T \Sigma^{-1}\left(\theta_e\right)\left[\hat{y} - f\left(\theta_m|\mathcal{MD}\right) - \mu\left(\theta_e\right)\right]\right]$$

UQ

water contact angle

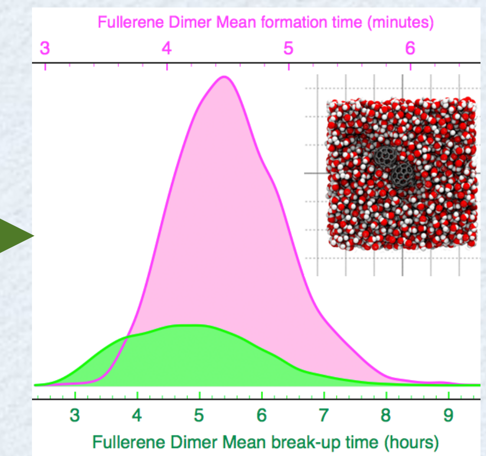


parameters: ε, r_{cut}



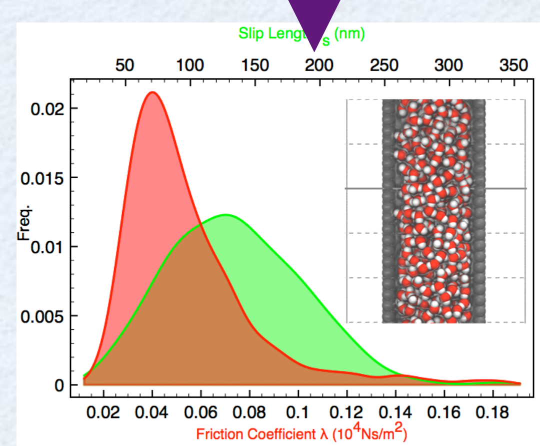
buckyball aggregation

P



200 samples - 7 days

P



Friction coefficient
and slip length of
water inside CNTs

500 samples - 2 days

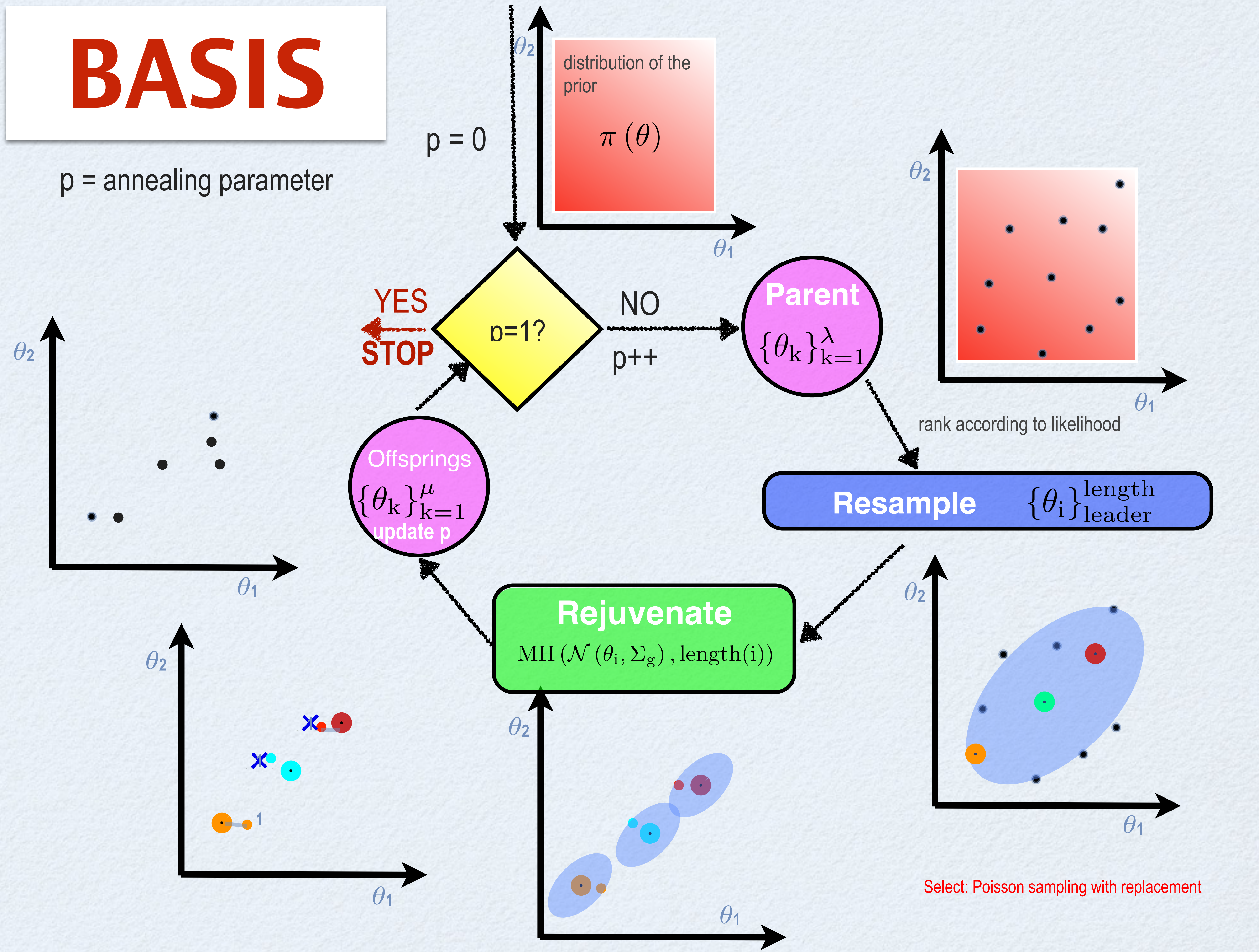
water transport in CNTs

HPC challenges for UQ
in nanoscale flows:

large, variable, computational
cost per calibration or
propagation sample

Resources: 1200 cores, 32GPUs

Angelikopoulos, P., Papadimitriou, C. and Koumoutsakos, P. (2013), J. Phys. Chem. B


BASIS
p = annealing parameter
p = 0
distribution of the prior
$\pi(\theta)$
θ_2
θ_1
p=1?
YES
STOP
NO
p++
Parent
$\{\theta_k\}_{k=1}^{\lambda}$
rank according to likelihood
Resample
$\{\theta_i\}_{leader}^{length}$
Rejuvenate
$MH(\mathcal{N}(\theta_i, \Sigma_g), length(i))$
Offsprings
$\{\theta_k\}_{k=1}^{\mu}$
update p
1
Select: Poisson sampling with replacement

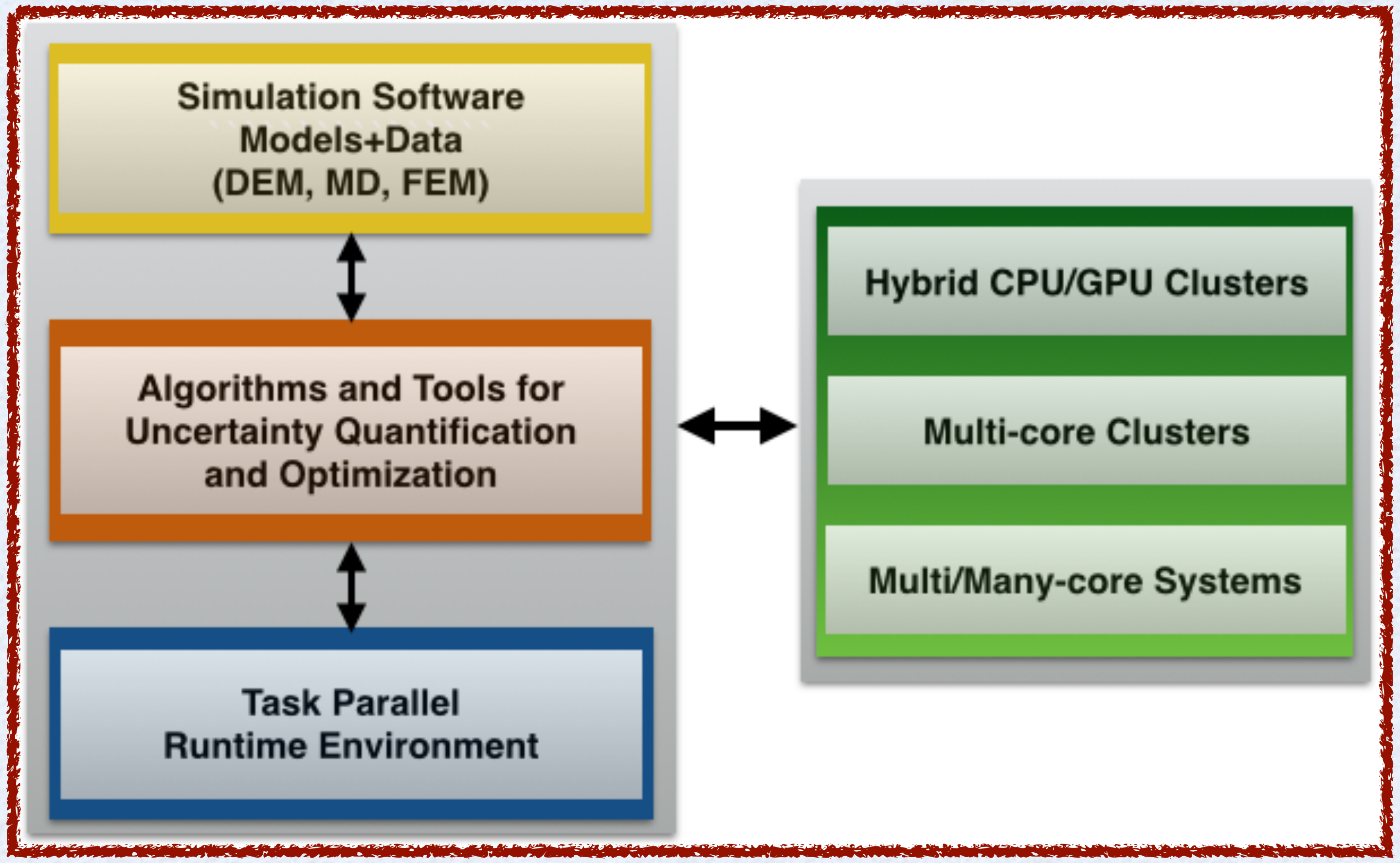
Π4U: HPC Framework for Bayesian UQ

Cluster Setup:

Piz Daint Cray XC30, 512 nodes x 8-core Intel Xeon E5-2670 + 1 Tesla GPU

=> **4,096 cores + 512 GPUs in total**

96% efficiency



- ✓ Automatic Load balancing
- ✓ Multi-level Nested parallelism
- ✓ Extensible/ built upon tasking Library TORC
- ✓ Slurm/ LSF queuing systems compatibility
- ✓ Intel Phi support

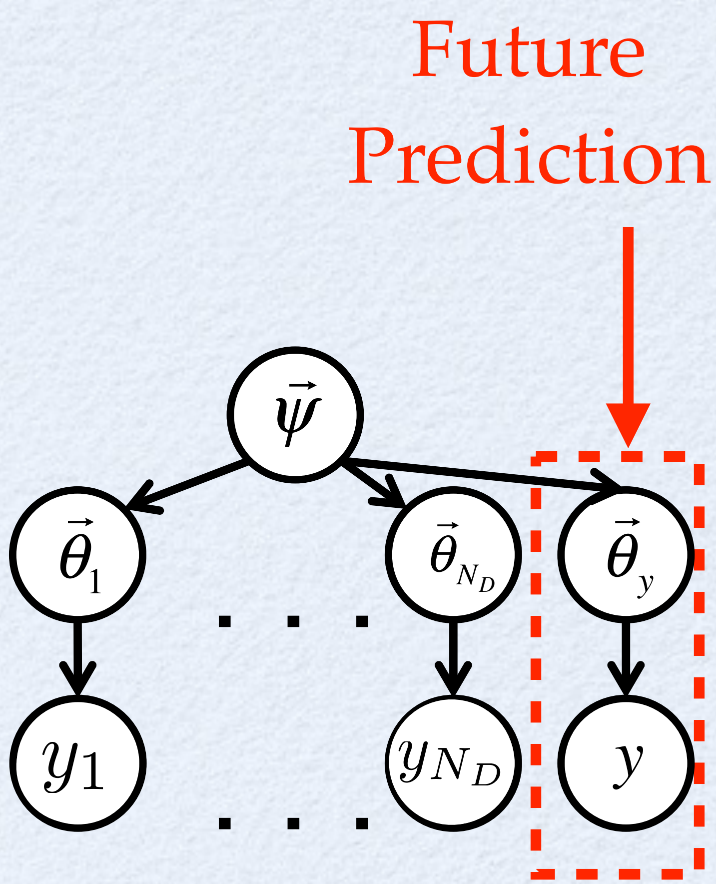
Algorithms implemented:

- ✓ BASIS
- ✓ SubSet simulation
- ✓ ABC-Subsim
- ✓ CMA-ES/AMalgam

"Hierarchical" Bayesian Framework

- Given N sets of heterogeneous data:
  1. Calibrate θ_i for each data set D_i
  2. Link all θ_i with hyperparameters ψ
  3. Combine heterogeneous data sets based on Evidence of D_i in Bayesian inference

- Evidence combines data-fitting & Ockham's razor



Pitfalls of more complicated probability graphs

$$\cancel{p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}}$$

$$p(\theta|D) = \int p(\theta|\psi)\boxed{p(\psi|D)}\,\mathrm{d}\psi$$

Double Integral

↓

Nested MCS

↓

Very High Computational cost

$$\left[\; p(\psi|D) = \frac{\boxed{p(D|\psi)}p(\psi)}{p(D)} \;\right]$$

$$\left[\; p(D|\psi) = \prod_{i=1}^{N} \boxed{p(D_i|\psi)} \quad \textit{Evidence of } D_i \right.$$

$$\left. = \prod_{i=1}^{N} \int p(D_i|\theta_i,\psi)p(\theta_i|\psi)\,\mathrm{d}\theta_i \;\right]$$

Importance Sampling for marginal likelihood

$$p(D_i|\psi, \sigma_y) = \int \frac{p(D_i|\theta_i, \sigma_y)p(\theta_i|\psi)}{\pi(\theta_i|D_i)} \pi(\theta_i|D_i)\, \mathrm{d}\theta_i$$

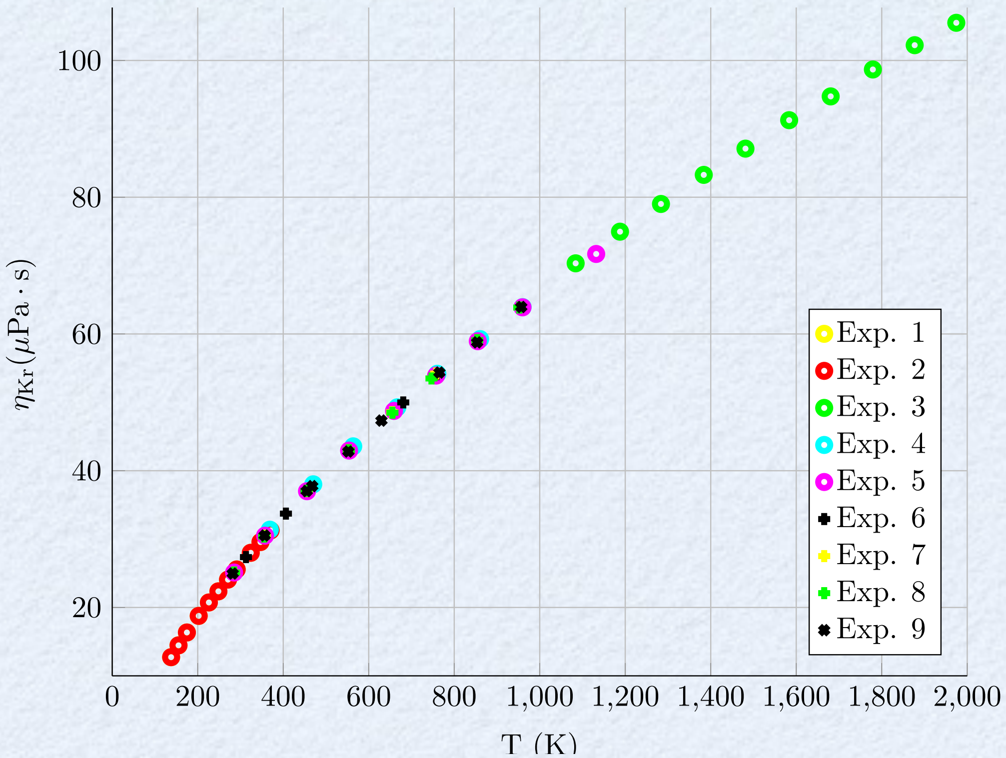
$$\left[\theta_i^{(j)} \sim \pi(\theta_i|D_i)\right] \qquad \approx \frac{1}{N_{s,i}} \sum_{j=1}^{N_{s,i}} \frac{p(D_i|\theta_i^{(j)}, \sigma_y)p(\theta_i^{(j)}|\psi)}{\pi(\theta_i^{(j)}|D_i)}$$

Basic idea: use posterior samples for each data Di to estimate the integral with varying (ψ,σ_y)

Pros: Usefull to do UQ on individual data set anyway, HB becomes a post-processing step with no extra likelihood evaluations

Cons: potentially high variance for the estimation

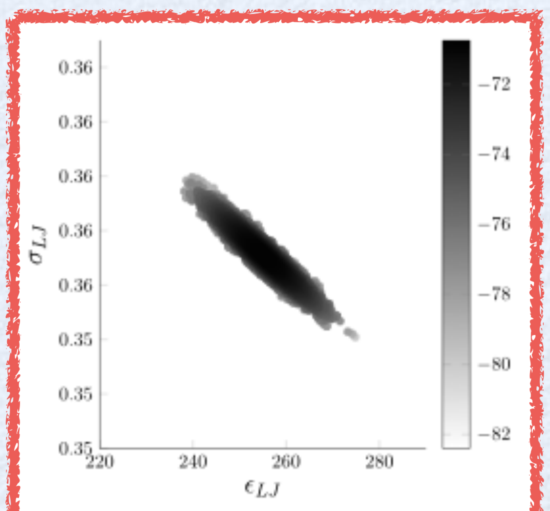
Data structure matters – Argon

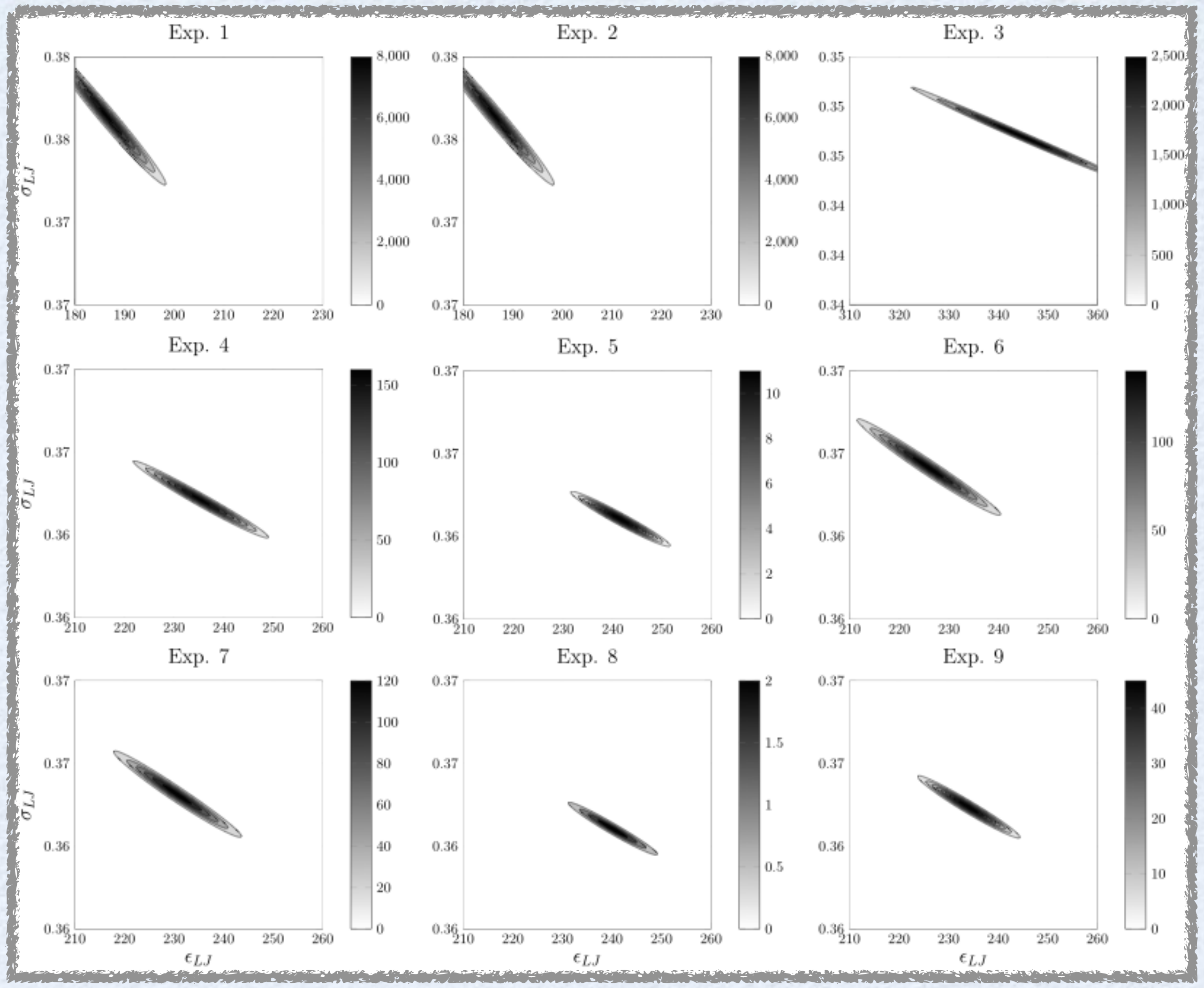


Argon viscosity

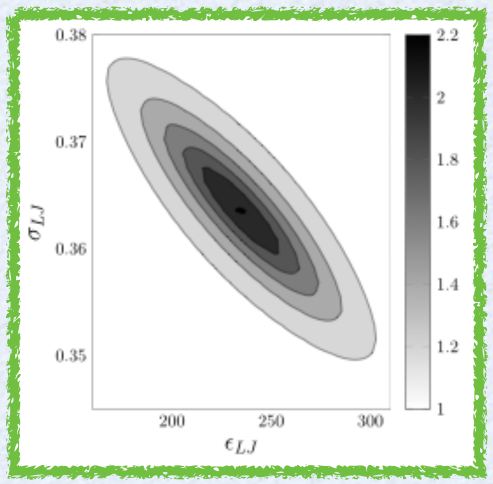
Kestin et al. 1984

Pooled Data





with structured Data

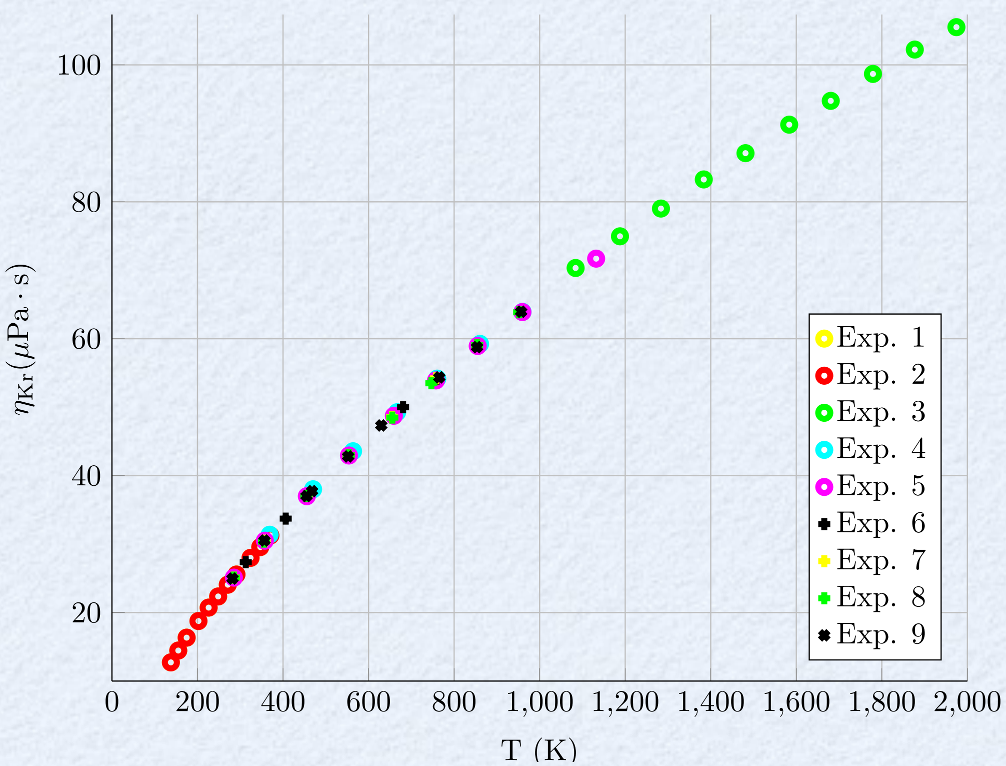


Wu, S., Angelikopoulos, P., Moser, R., Papadimitriou, C. and Koumoutsakos, P. (2015), Phil. Trans. Royal. Soc. A.

Identifying the plausible structure of the Data

Parameters: Lennard-Jones

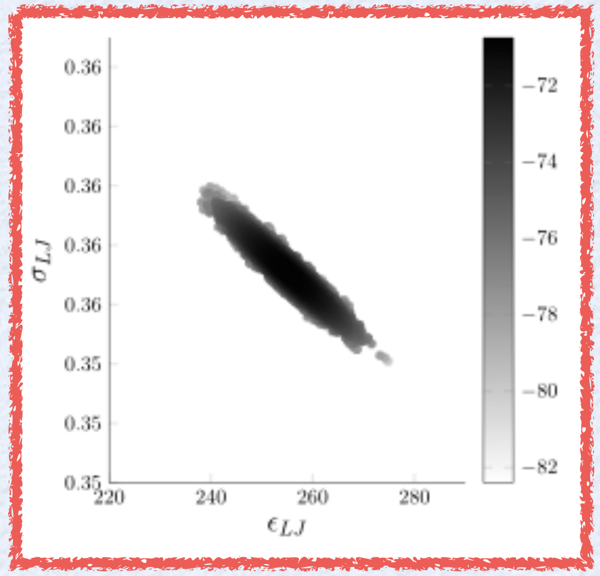
ε_{LJ}, σ_{LJ}

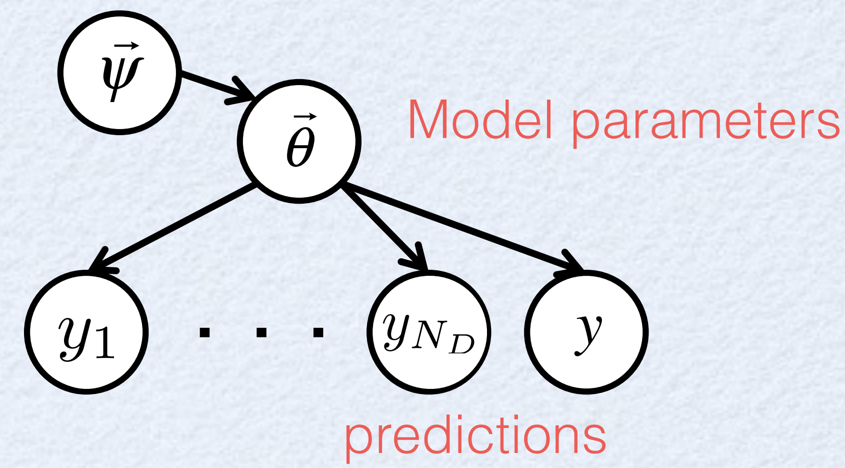


Argon viscosity
Kestin et al. 1984

Pooled Data

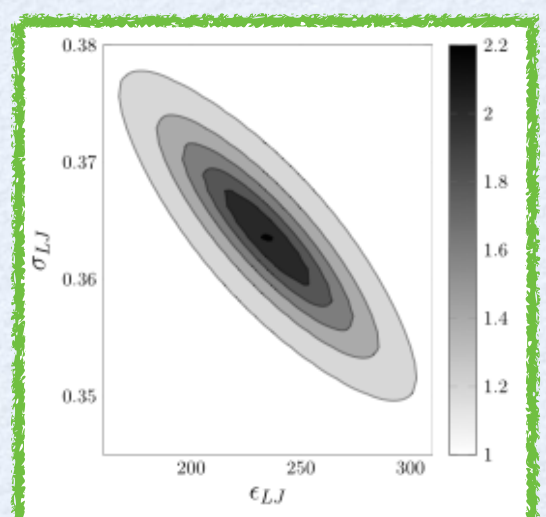
M_1 Evidence : -77.02

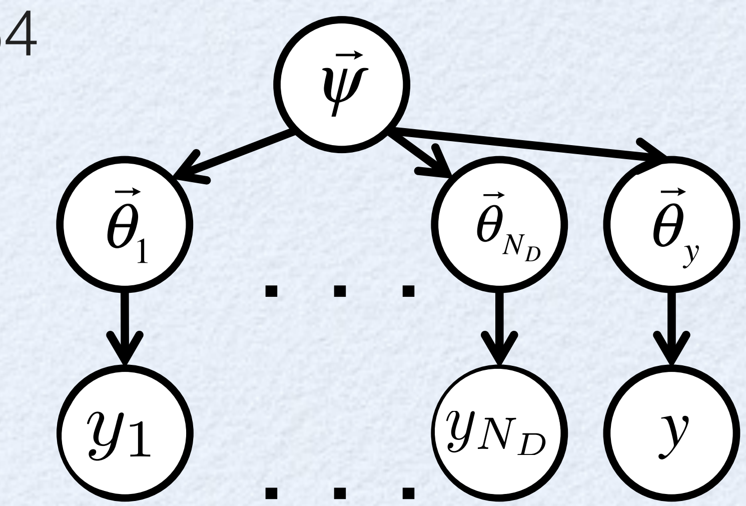




Tree Data Structure

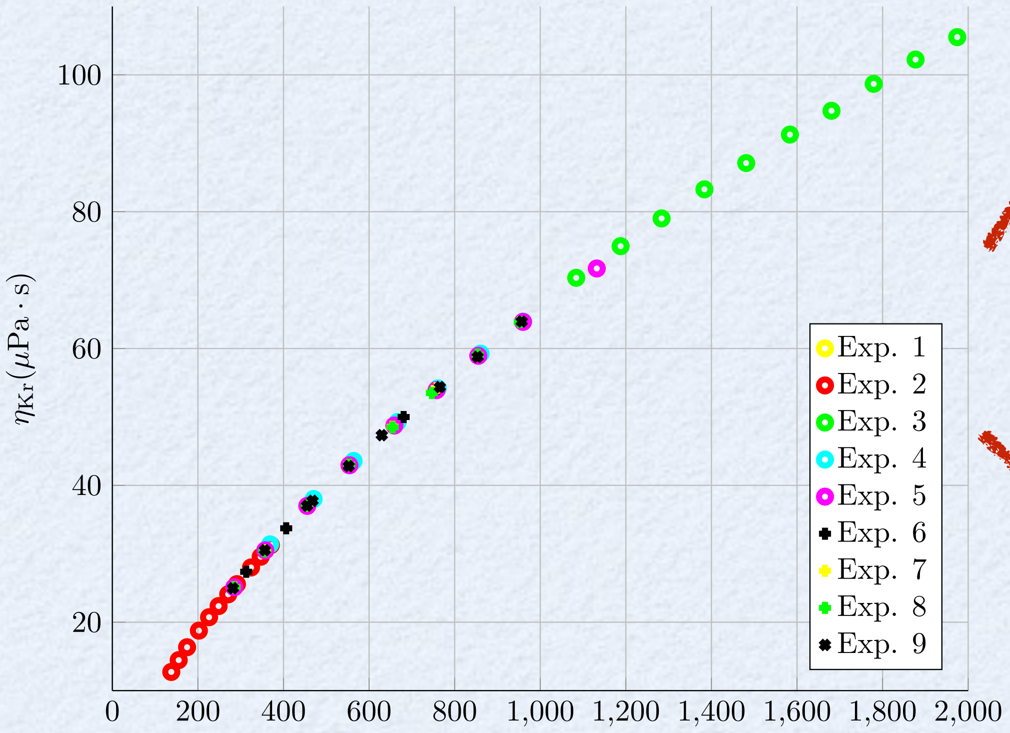
M_2 Evidence : -20.64



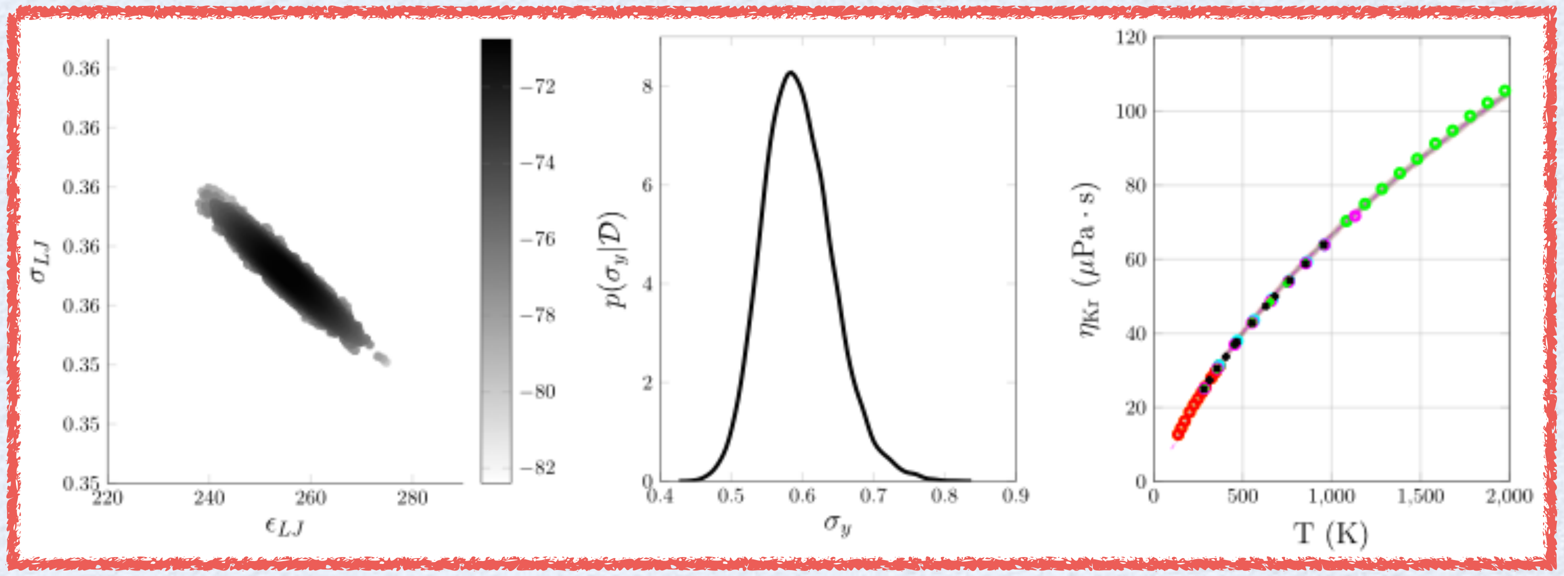


Effect of graphical model on prediction

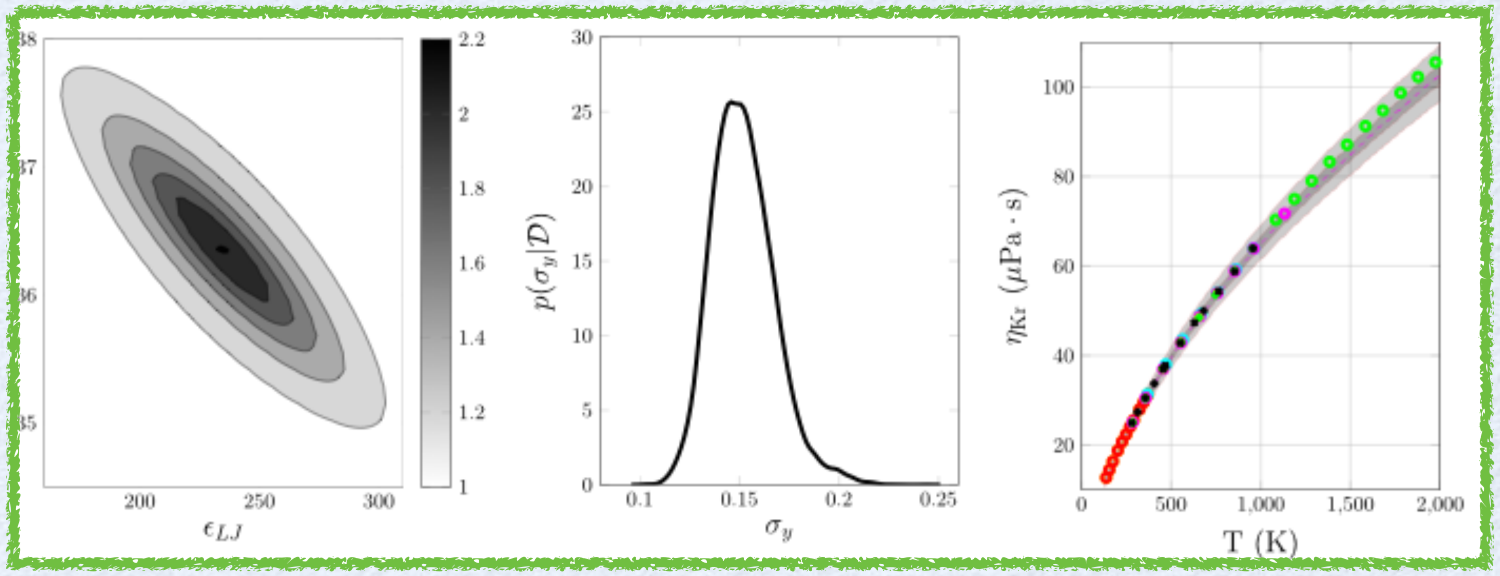
Effect on robust prediction



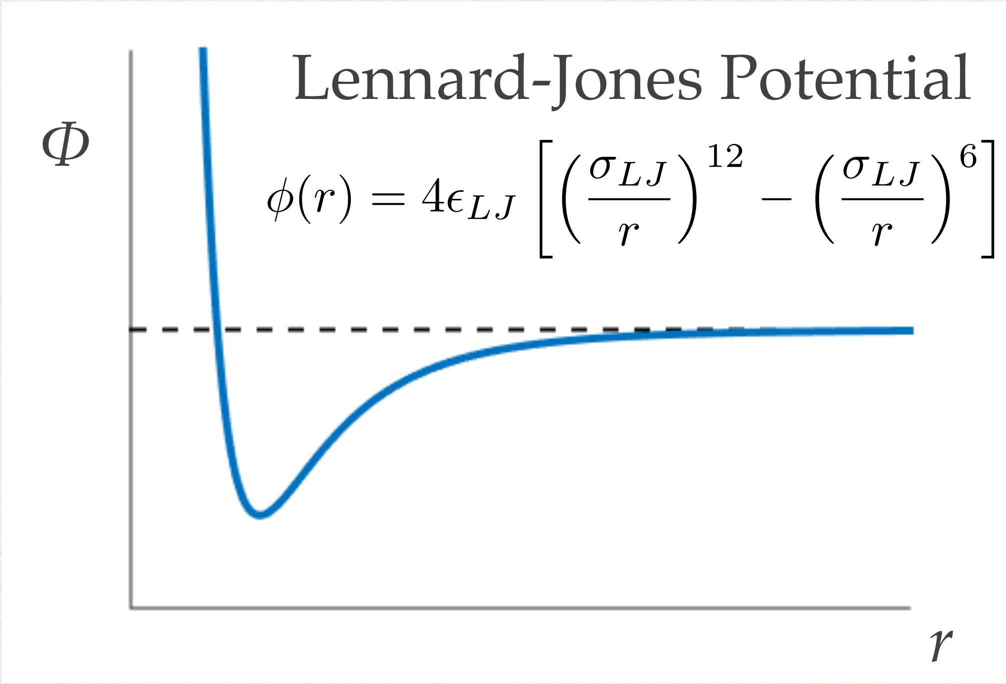
without Hierarchical Data Structure

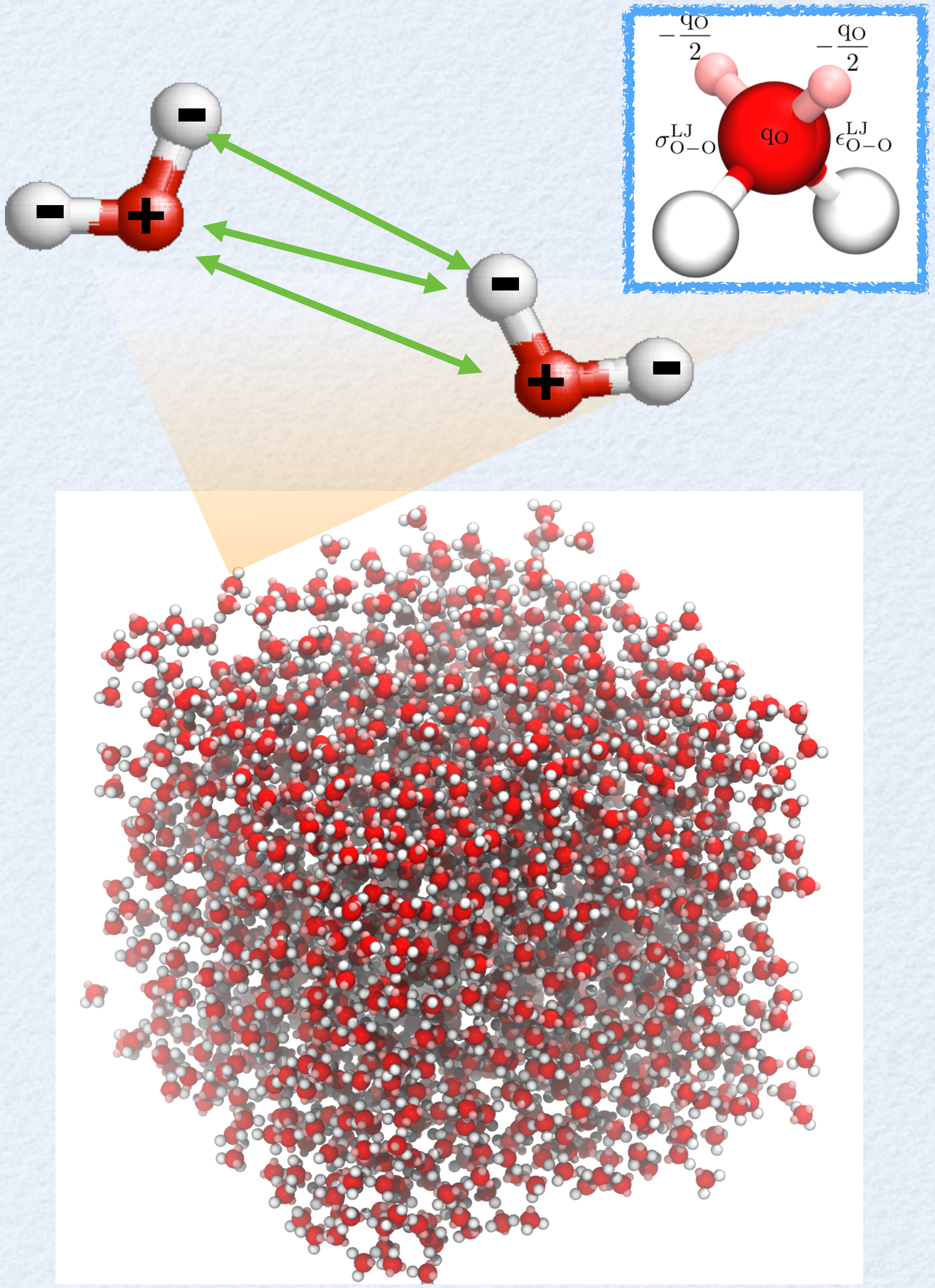


with Hierarchical Data Structure



EXAMPLE: MD Simulations for Water

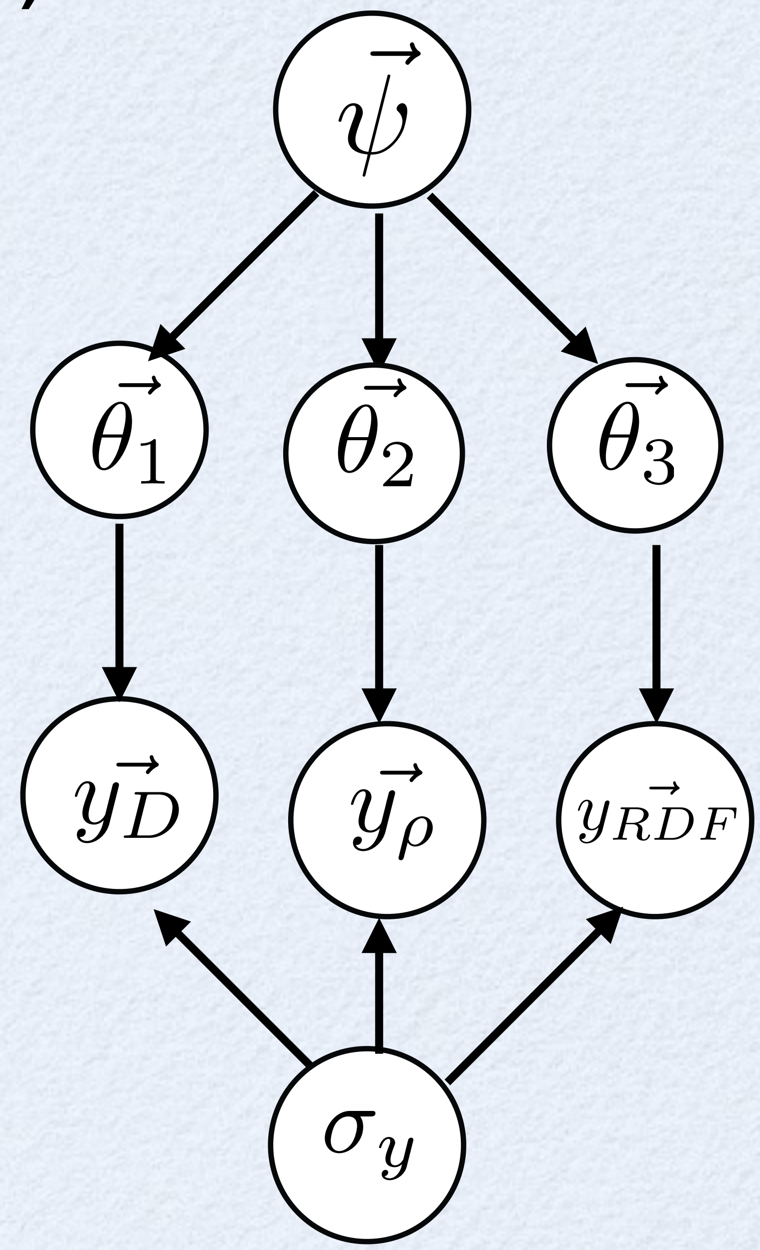




- MD parameter calibration - θ
  - O-O Lennard-Jones (*ε_{LJ}*, *σ_{LJ}*)
  - O-O, O-H Coulomb charges (*q*)

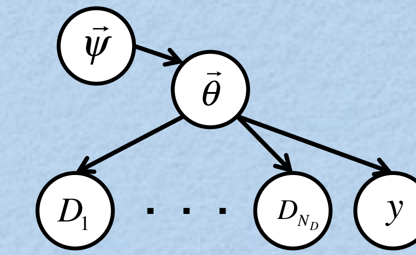
DATA for Water

- Heterogeneous Data Sets (as a function of temperature)
  - Diffusion coefficient
  - Density
  - Radial Distribution Function (RDF)

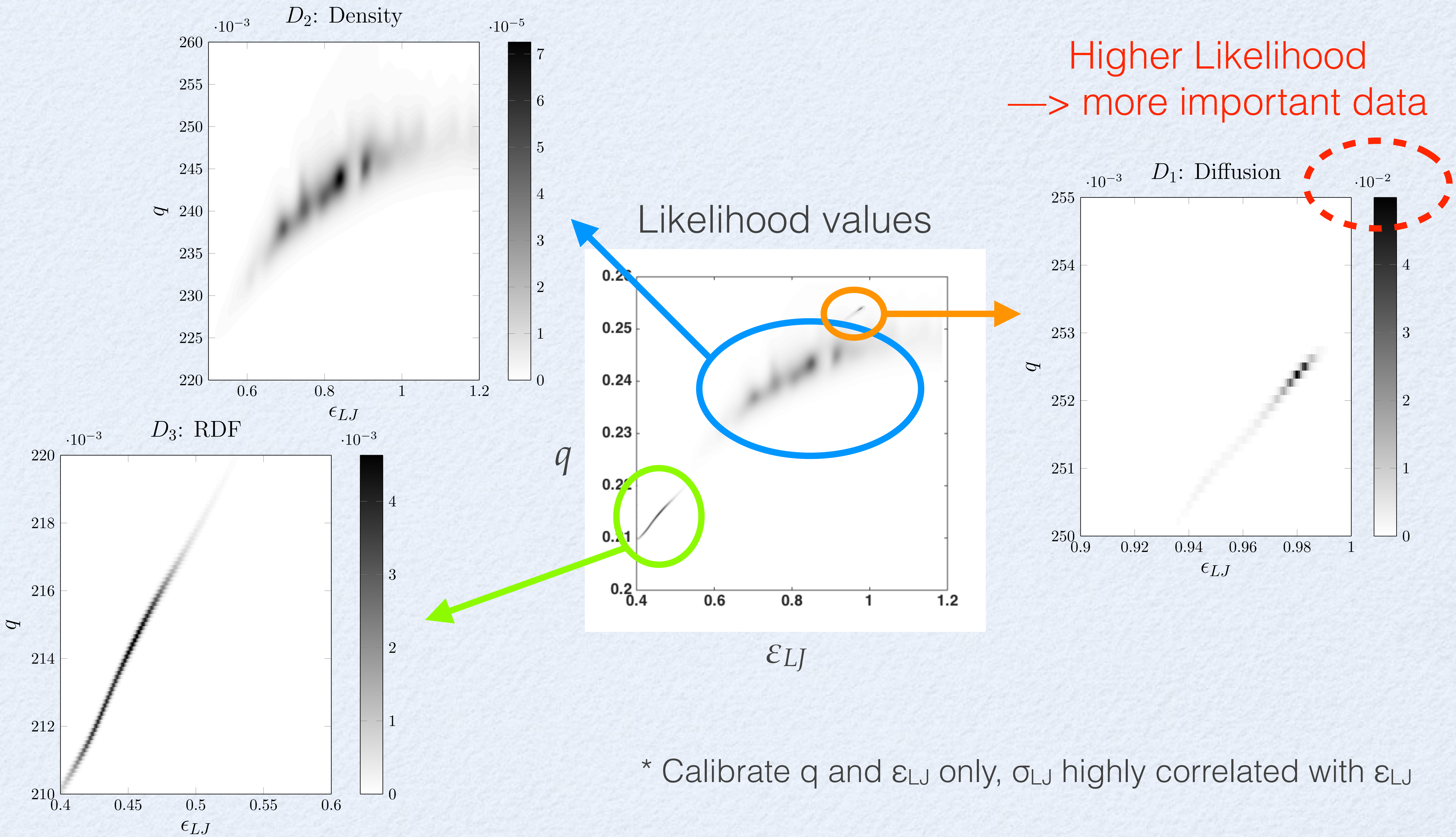


Data sources:
Holz et al. 2000; Jones & Harris 1992; Soper 2013

Bayesian Inference for pooled data



Calibrate for each data set individually…



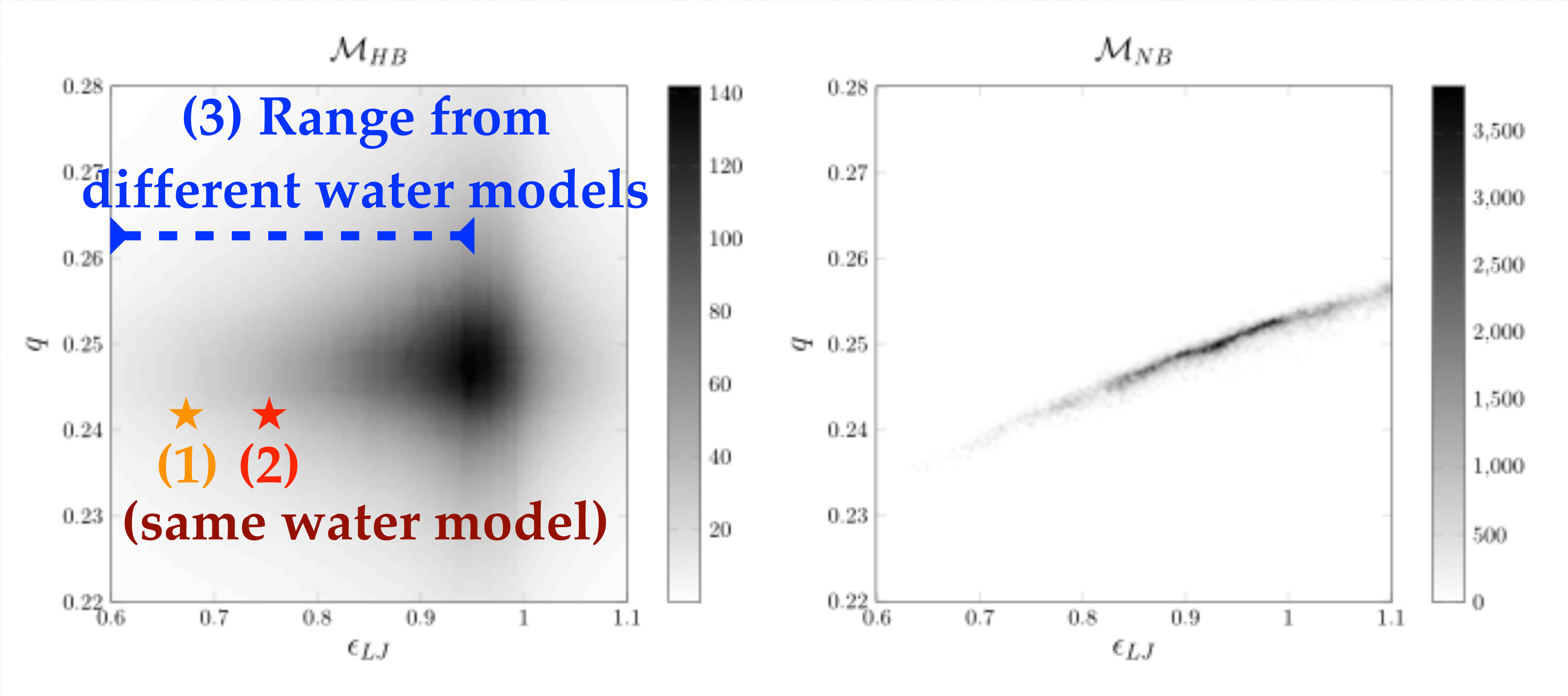
* Calibrate q and ε_{LJ} only, σ_{LJ} highly correlated with ε_{LJ}

Posterior of Model Parameters

Compare Hierarchical Bayesian model with Independent Likelihoods model

Hierarchical (HB)

Independent Likelihoods (NB)



Model Selection

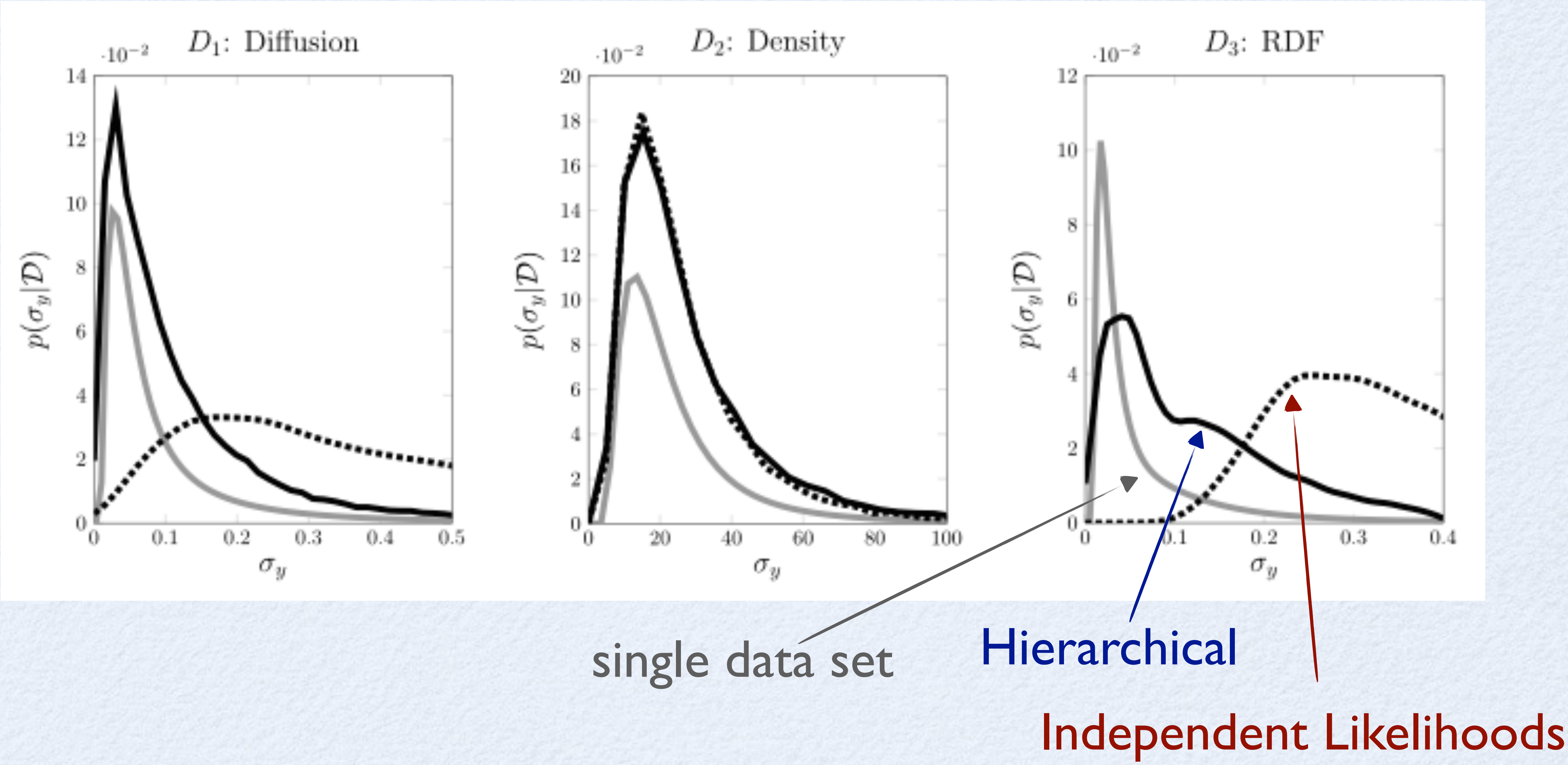
Model	P(Model|D)
HB	0.98
NB	0.02

(1) Mahoney & Jorgensen 2000 (2) Rick et al. 2004

(3) http://www1.lsbu.ac.uk/water/water_models.html

Posterior of Measurement Noise

Results from Hierarchical Bayesian model overlaps with results from individual data set



SUMMARY

- Bayesian UQ in MD: Data Structure and Computational Intensity

  - Data Structuring
    - Calibration data for MD are heterogeneous →
    - Assuming independence for all data is NOT ENOUGH to reach good calibration and prediction!
      - Combine heterogeneous data based on Evidence
      - balance data-fitting and information gain
  - Computational intensity
    - “Evolutionary” algorithms (BASIS/SMC)
    - HPC framework for Bayesian UQ (Π4U)
    - Surrogate models

www.cse-lab.ethz.ch/software/Pi4U

Thank you!

Some References:

- Wu,S., Angelikopoulos, P. Moser, R. Papadimitriou, C. and Koumoutsakos, P. (2015) “A Hierarchical Bayesian Framework for Force Field Selection in Molecular Dynamics Simulations”, *Philosophical Transactions of the Royal Society B.*
- Angelikopoulos, P., Papadimitriou, C. and Koumoutsakos, P. (2015) "X-TMCMC: Adaptive Kriging for Bayesian inverse modeling", *Computer Methods in Applied Mechanics and Engineering*
- Hadjidoukas, P.E., Angelikopoulos, P., Papadimitriou, C. and Koumoutsakos, P. (2015) "Pi4U: A high performance computing framework for Bayesian uncertainty quantification of complex models", *Journal of Computational Physics*
- Hadjidoukas, P.E., Angelikopoulos, P., Rossinelli, D., Alexeev, D., Papadimitriou, C. and Koumoutsakos, P. (2014) "Bayesian Uncertainty Quantification and propagation for Discrete Element Simulations of granular materials", *Computer Methods in Applied Mechanics and Engineering*
- Angelikopoulos, P., Papadimitriou, C. and Koumoutsakos, P. (2013) "Data Driven, Predictive Molecular Dynamics for Nanoscale Flow Simulations Under Uncertainty", *Journal of Physical Chemistry B*
- Angelikopoulos, P., Papadimitriou, C. and Koumoutsakos, P. (2012) "Bayesian uncertainty quantification and propagation in molecular dynamics simulations: A high performance computing framework", *Journal of Chemical Physics*