SAND2015-10607PE

UQ in Molecular Dynamics Simulations: Forward and Inverse Problem

F. Rizzi[†,‡], O. Knio[‡,*], R. Jones[†], H. Adalsteisson[†], H. Najm[†], K. Sargsyan[†], M. Salloum[†], C. Safta[†], B. Debusschere[†]

[†] Sandia National Laboratories, Livermore, CA
[‡] Johns Hopkins University, Baltimore, MD
[*] Duke University, Durham, NC

Sensitivity, Error and Uncertainty Quantification for Atomic, Plasma and Material Data
– Stony Brook Univ., Nov. 2015 –

Supported by the US Department of Energy (DOE)
Advanced Scientific Computing Research (ASCR)

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Background and Motivation

Why Uncertainty Quantification (UQ)?

George E. P. Box, 1919 - 2013

Fundamental work on:

- Time-series analysis
- Box-Cox transformation
- Response surface methodology
- ...

"Remember that all models are essentially wrong; the practical question is how wrong do they have to be to not be useful."

Background and Motivation

Why Uncertainty Quantification (UQ)?

George E. P. Box, 1919 - 2013

Fundamental work on:

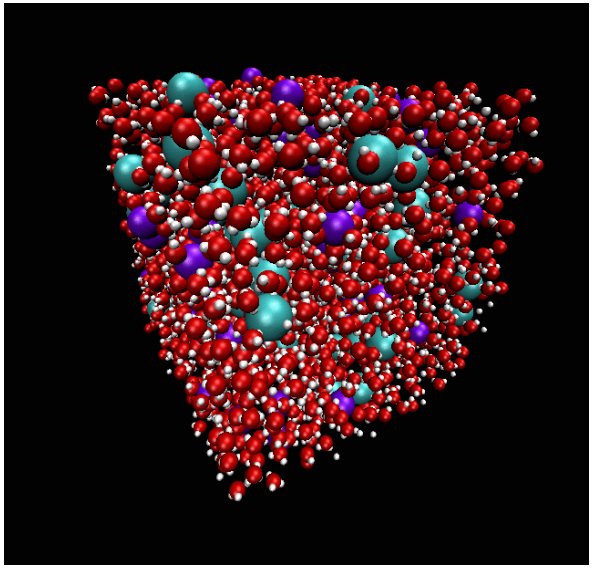
- Time-series analysis
- Box-Cox transformation
- Response surface methodology
- ...

"Remember that all models are essentially wrong; the practical question is how wrong do they have to be to not be useful."

Background and Motivation

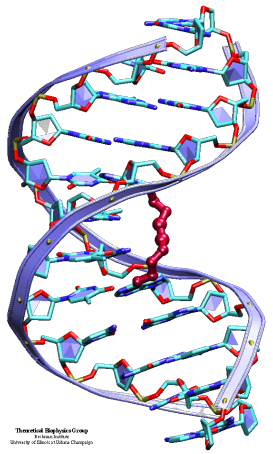
- 1957: seminal work in molecular dynamics (MD), Alder and Wainwright.
- 1964: first MD simulation based on a realistic potential (Lennard-Jones): liquid Ar (Rahman).
- 1974: first MD simulation of liquid water (Stillinger and Rahman).

  . . .



MD simulation of $\mathbf{Na}^{+}$ and $\mathbf{Cl}^{-}$ in water.

- MD is useful and cheap (vs. experiments) to explore physical properties at the atomic level.
- Industrial/academic applications: liquids, solids, proteins and nucleic acids (DNA, RNA).
- As every simulation technique, MD is an approximation method with a few weaknesses...



MD snapshot of DNA (Biophys. group, UIUC)

Background and Motivation: MD overview

- Classical MD simulation (Frenkel,2001; Allen & Tildesley,1987):

$$\frac{d^2\mathbf{r}_{(i,t)}}{dt^2} = \frac{\mathbf{f}_{(i,t)}}{m_i} \qquad \mathbf{f}_{(i,t)} = -\nabla_{\mathbf{r_i}}\Phi\left(\mathbf{r}_{(1,t)}, \ldots, \mathbf{r}_{(N,t)}\right) \qquad i = 1, \ldots, N$$

- Φ is the potential (or force-field), defined *before* starting the simulation.
- Φ should be tailored to the target application.
- Reliability depends on the accuracy of Φ.
- Continuous development of potentials and experience over the years.
- MD potential represents an important source of uncertainty.

Background and Motivation: MD overview

- Classical MD simulation (Frenkel,2001; Allen & Tildesley,1987):

$$\frac{d^2\mathbf{r}_{(i,t)}}{dt^2} = \frac{\mathbf{f}_{(i,t)}}{m_i} \qquad \mathbf{f}_{(i,t)} = -\nabla_{\mathbf{r_i}} \Phi\left(\mathbf{r}_{(1,t)}, \ldots, \mathbf{r}_{(N,t)}\right) \qquad i = 1, \ldots, N$$

- Φ is the potential (or force-field), defined *before* starting the simulation.
- Φ should be tailored to the target application.
- Reliability depends on the accuracy of Φ.
- Continuous development of potentials and experience over the years.
- MD potential represents an important source of uncertainty.

Potential Uncertainty for Water

- More than 50 MD water models available (Guillot,2002; Wallqvist,2007).
- Only some physical properties are reproduced with a good degree of accuracy.

Acronym	Date	Type	Sites	Reference
SPC	1981	rigid	3	(Berendsen,1981)
TIP3P	1981	rigid	3	(Jorgensen,1983)
SPC/F	1985	flexible	3	(Toukan,1985)
SPC/FP	1991	flexible,polarizable	3	(Zhu,1991)
NSPCE	1998	rigid	3	(Errington,1998)
SPC/Fw	2006	flexible	3	(Wu,2006)
BF	**1933**	rigid	4	(Bernal,1933)
RWK	1982	flexible	4	(Reimers,1982)
TIP4P	1983	rigid	4	(Jorgensen,1983)
PTIP4P	1991	polarizable	4	(Sprik,1991)
TIP4P/FQ	1994	polarizable	4	(Rick,1994)
TIP4P-Ew	2004	rigid	4	(Horn,2004)
TIP4P/2005	2005	rigid	4	(Abascal,2005)
ST2	1973	rigid	5	(Stillinger,1974)
TIP5P	2000	rigid	5	(Mahoney,2000)
TIP5P-Ew	2004	rigid	5	(Rick,2004)
NvdE	2003	rigid	6	(Nada,2003)

Table: Reduced list of water models developed since 1933.

Potential Uncertainty for Water

a Most water models use Lennard-Jones (LJ) potential to describe Van der Waals forces.

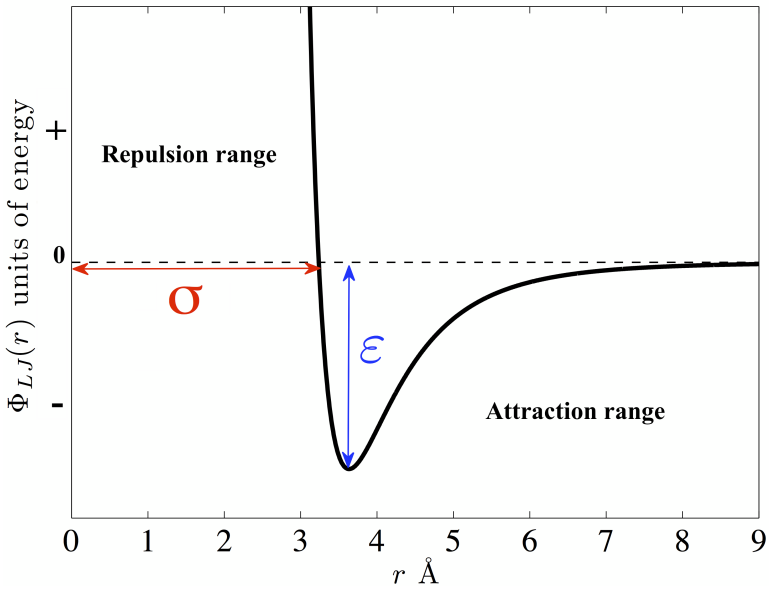
$$\Phi_{LJ}(r) = 4\varepsilon \left\{ \left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6} \right\}$$

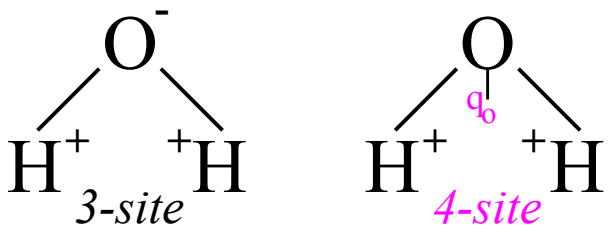
- Different models involve different values of the LJ parameters ε, σ.

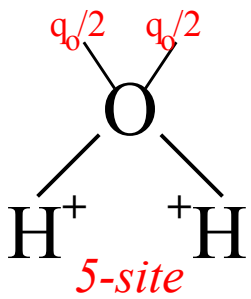
b Rigid or flexible molecule.

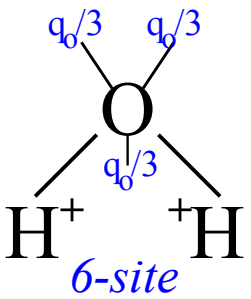
c H_2O structure: from 3-site to 6 sites models.
...

- Discussion holds for several other systems: potential and parameters are important sources of uncertainty to consider.









Uncertainty Quantification (UQ)

- Estimating uncertainty in a model of a physical process of interest.
- Complex non-linear systems: small uncertainties and errors can be largely amplified and strongly affect the model predictions.
- Key role when high-fidelity/risk prediction is of central importance.
- Polynomial chaos (PC) (Wiener,1938) and Bayesian inference (Bayes,1763).

PC expansion: X is a target RV - c_i are coeff. - $\Psi_i(\boldsymbol{\xi})$ polyn. of standard RV ξ

$$X \approx \sum_{i=0}^{\infty} c_i \Psi_i(\boldsymbol{\xi})$$

Bayes' theorem: $\mathbf{D}$ is data - $\boldsymbol{\theta}$ is set of parameters (hypothesis)

$$\overbrace{\mathcal{P}(\boldsymbol{\theta}|\mathbf{D})}^{Posterior} = \frac{\overbrace{\mathcal{P}(\mathbf{D}|\boldsymbol{\theta})}^{Likelihood}\ \overbrace{\mathcal{P}(\boldsymbol{\theta})}^{Prior}}{C}$$

Talk Overview

Part I: Forward Propagation

- How uncertainty in a set of model parameters affects selected predictions.

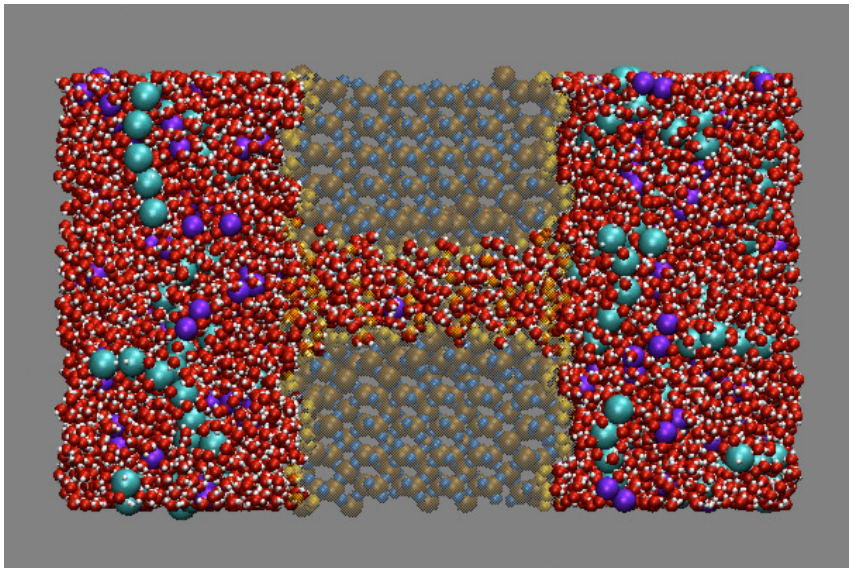
- Focus on MD simulations of concentration driven ionic flow in a silica nanopore.
- System's heterogeneity is a key complexity of this study.

Part II: Inverse Problem

- Estimating target model parameters based on a set of observations.

- Focus on MD simulations of bulk water.
- Estimation of potential parameters based on noisy observations of water observables.

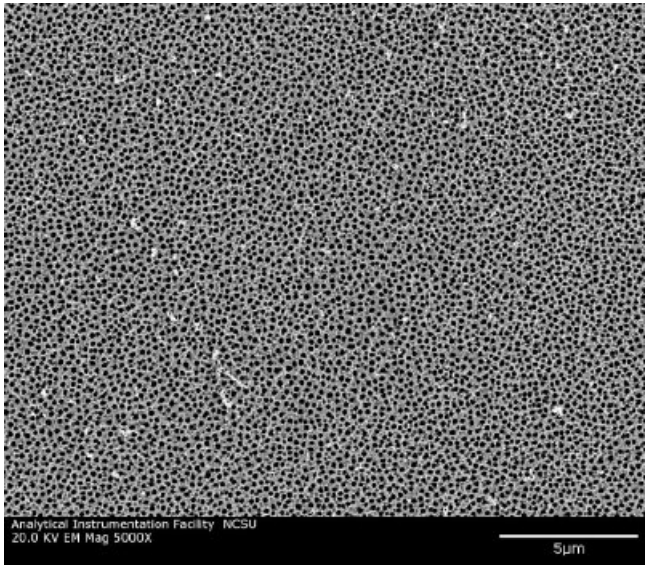
Forward Propagation: Nanopore Flow



- F. Rizzi, R.E. Jones, B.J. Debusschere and O.M. Knio - Part I – *J. Chem. Phys.*, 138:194104, 2013.
- F. Rizzi, R.E. Jones, B.J. Debusschere and O.M. Knio - Part II – *J. Chem. Phys.*, 138:194105, 2013.
- F. Rizzi, Ph.D. thesis, The Johns Hopkins University, Baltimore, MD, 2012.

Why Nanopores?

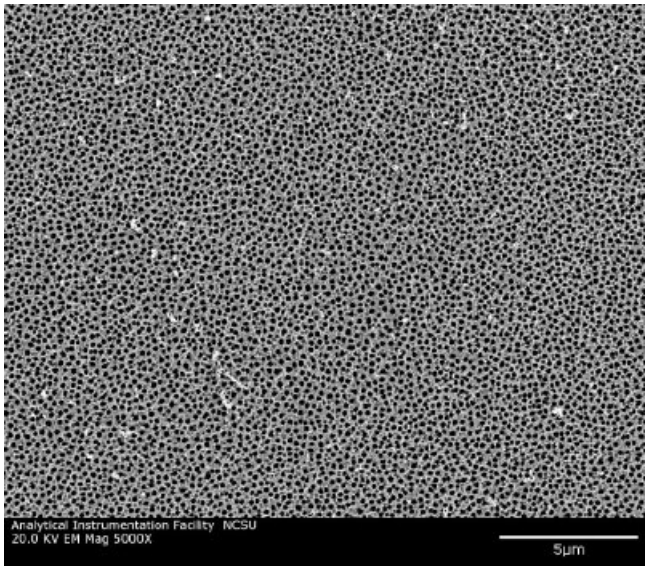
- Synthetic nanopores are widely used in the industry: desalination or other separation tasks.
- Selective control of transport: identify & manipulate physical properties or the interaction between the transported ions (or molecules) and the pore walls.
- Inspired by their biological counterparts: e.g. transmembrane protein channels.
- Complex and highly heterogeneous.
- MD simulations of nanopores are not new.
- UQ applied to nanopore simulations is novel.
  - Characterize uncertanties in the system.
  - Important for improving design capabilities.



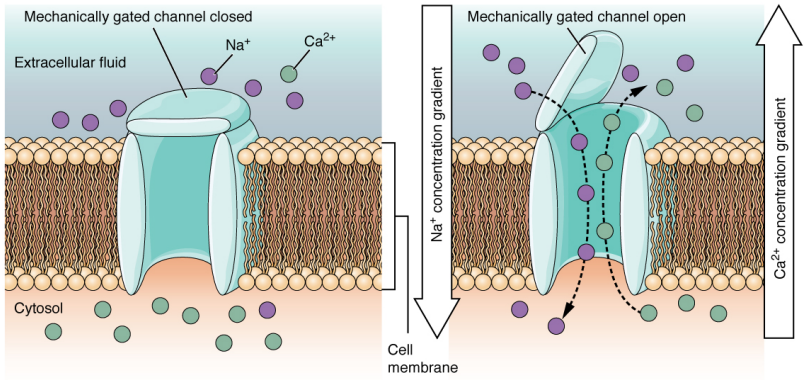
Example of nanoporous membrane: pore size $\sim 20nm$, (R.Narayan, NC State).

Why Nanopores?

- Synthetic nanopores are widely used in the industry: desalination or other separation tasks.
- Selective control of transport: identify & manipulate physical properties or the interaction between the transported ions (or molecules) and the pore walls.
- Inspired by their biological counterparts: e.g. transmembrane protein channels.
- Complex and highly heterogeneous.
- MD simulations of nanopores are not new.
- UQ applied to nanopore simulations is novel.
  - Characterize uncertanties in the system.
  - Important for improving design capabilities.



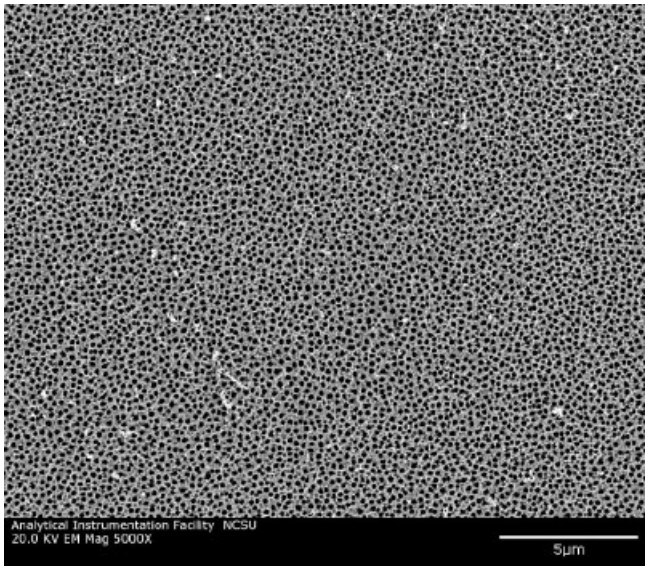
Example of nanoporous membrane: pore size $\sim 20nm$, (R.Narayan, NC State).



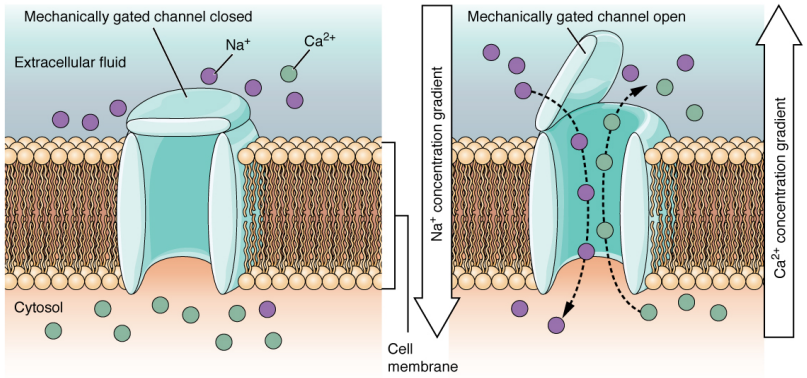
Proten channel schematic (web).

Why Nanopores?

- Synthetic nanopores are widely used in the industry: desalination or other separation tasks.
- Selective control of transport: identify & manipulate physical properties or the interaction between the transported ions (or molecules) and the pore walls.
- Inspired by their biological counterparts: e.g. transmembrane protein channels.
- Complex and highly heterogeneous.
- MD simulations of nanopores are not new.
- UQ applied to nanopore simulations is novel.
  - Characterize uncertanties in the system.
  - Important for improving design capabilities.



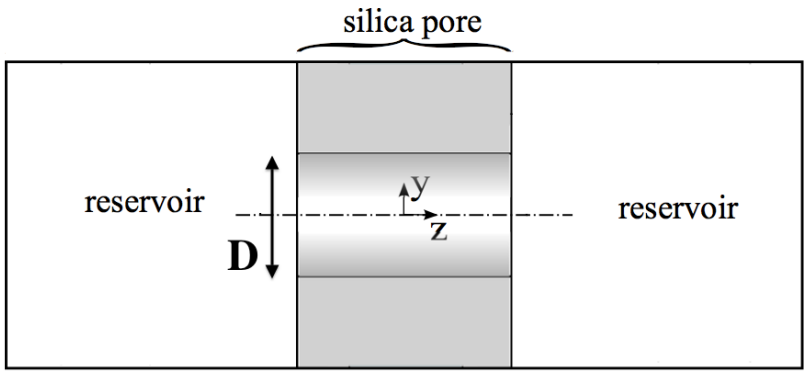
Example of nanoporous membrane: pore size $\sim 20nm$, (R.Narayan, NC State).

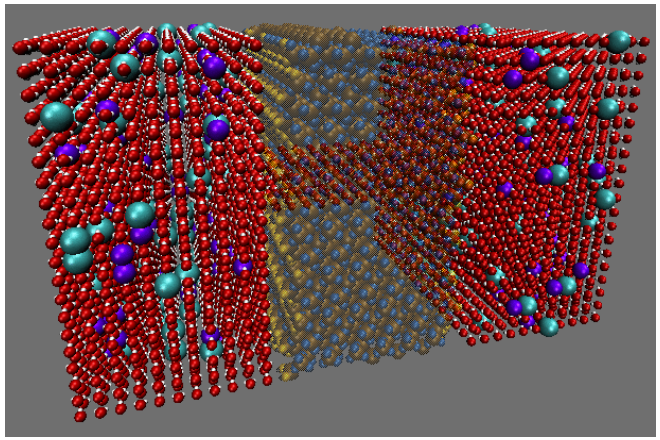


Proten channel schematic (web).

Atomistic System and Geometry

- Silica pore model connecting two reservoirs containing a 1.5 mol/l solution of sodium ($\mathbf{Na^+}$) and chloride ($\mathbf{Cl^-}$) ions in H_2O (white-red).
- Reservoirs communicate only through the pore, PBC are imposed along x and y.



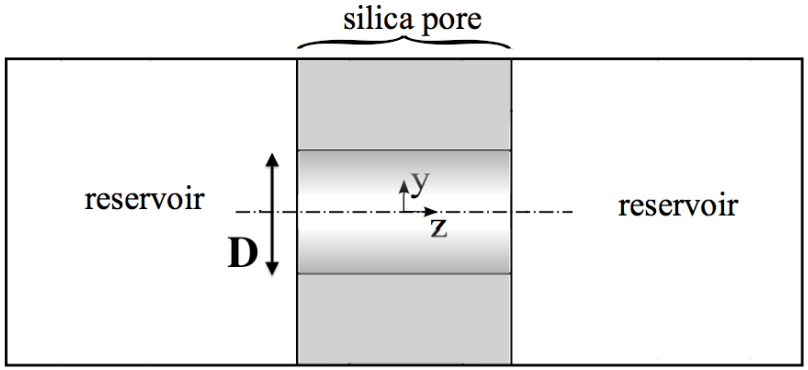


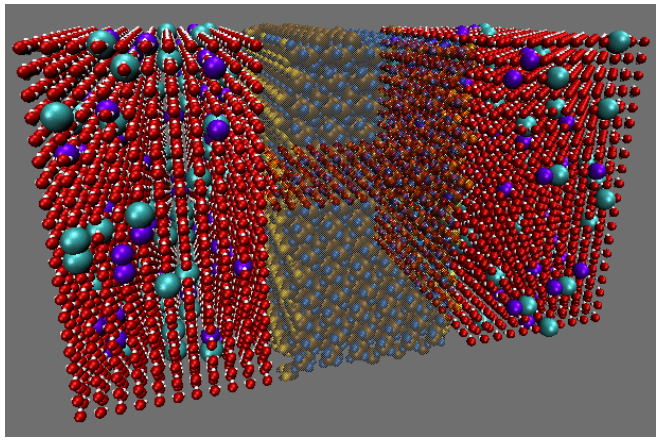
1. α-quartz crystal structure for the silica.
2. Cylindrical region of nominal diameter D.
3. Saturate dangling bonds with hydroxide groups (OH^-), to mimic real hydroxylation processes.

- Domain (xyz) $5.4 \times 6 \times 10.5$ nm^3.
- LAMMPS, simulation time ~ 8 ns.

Atomistic System and Geometry

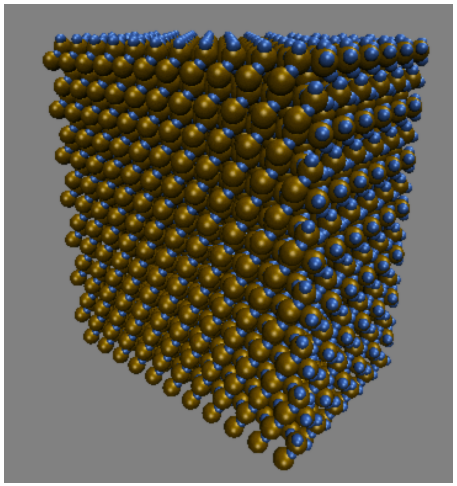
- Silica pore model connecting two reservoirs containing a 1.5 mol/l solution of sodium ($\mathbf{Na^+}$) and chloride ($\mathbf{Cl^-}$) ions in H_2O (white-red).
- Reservoirs communicate only through the pore, PBC are imposed along x and y.





1. α-quartz crystal structure for the silica.
2. Cylindrical region of nominal diameter D.
3. Saturate dangling bonds with hydroxide groups (OH^-), to mimic real hydroxylation processes.

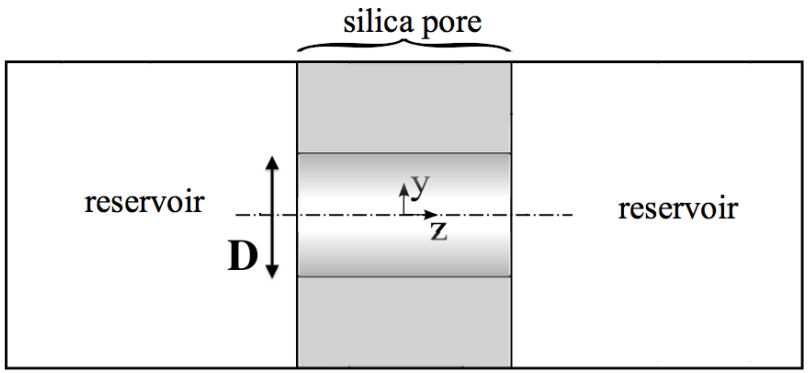
- Domain (xyz) $5.4 \times 6 \times 10.5$ nm^3.
- LAMMPS, simulation time $\sim$ 8 ns.

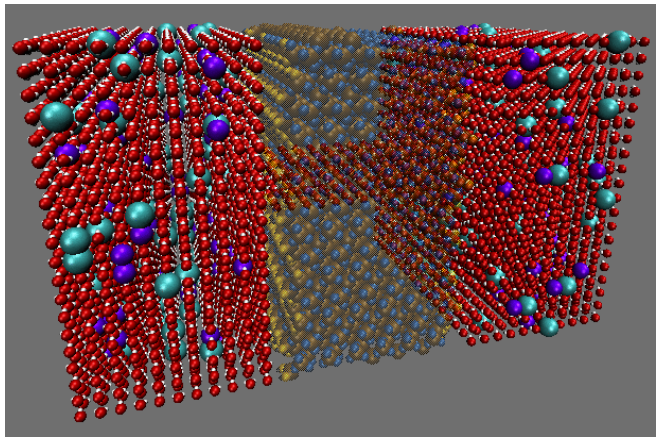


$\mathbf{O}_{bulk}$, **Si**

Atomistic System and Geometry

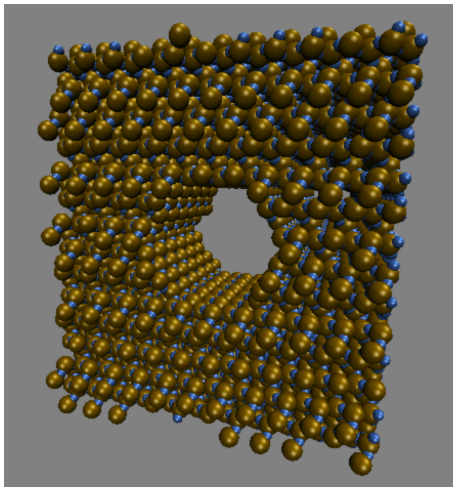
- Silica pore model connecting two reservoirs containing a 1.5 mol/l solution of sodium ($\mathbf{Na^+}$) and chloride ($\mathbf{Cl^-}$) ions in H_2O (white-red).
- Reservoirs communicate only through the pore, PBC are imposed along x and y.





1. α-quartz crystal structure for the silica.
2. Cylindrical region of nominal diameter D.
3. Saturate dangling bonds with hydroxide groups (OH^-), to mimic real hydroxylation processes.

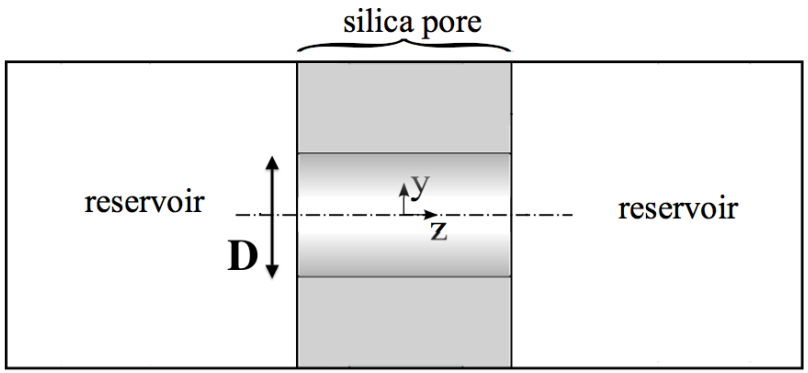
- Domain (xyz) $5.4 \times 6 \times 10.5$ nm^3.
- LAMMPS, simulation time $\sim$ 8 ns.

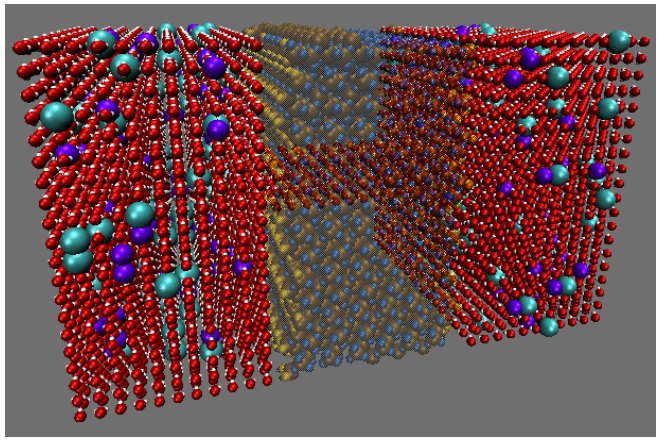


$\mathbf{O}_{bulk}$, **Si**

Atomistic System and Geometry

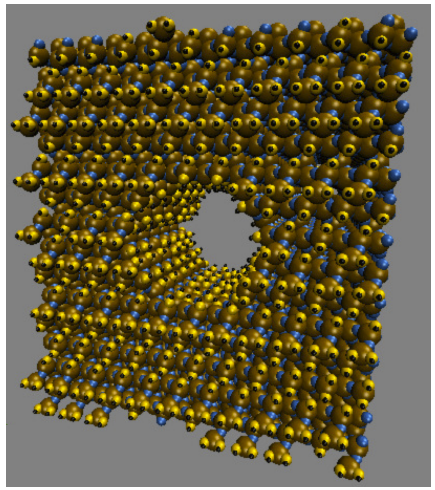
- Silica pore model connecting two reservoirs containing a 1.5 mol/l solution of sodium ($\mathbf{Na^+}$) and chloride ($\mathbf{Cl^-}$) ions in H_2O (white-red).
- Reservoirs communicate only through the pore, PBC are imposed along x and y.





1. α-quartz crystal structure for the silica.
2. Cylindrical region of nominal diameter D.
3. Saturate dangling bonds with hydroxide groups (OH^-), to mimic real hydroxylation processes.

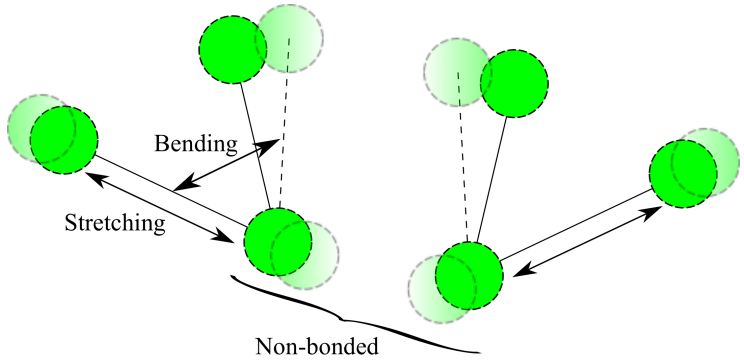
- Domain (xyz) $5.4 \times 6 \times 10.5$ nm^3.
- LAMMPS, simulation time $\sim$ 8 ns.



$\mathbf{O}_{bulk}$, **Si**, **OH**

MD Potential

- $\Phi_{total} = \Phi_{bonded} + \underbrace{\Phi_{LJ} + \Phi_{Coulomb}}_{non-bonded}$
- Φ_{bonded} (bond stretching and bending) is modeled using harmonic potential.



- Non-bonded interactions (Van der Waals + Electrostatic) are modeled as:

$$\Phi_{non-bonded} = \sum_{i=1, j>i}^{n_{atoms}} \left[\underbrace{4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right]}_{\Phi_{LJ}(r_{ij})} + \underbrace{\frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}}_{\Phi_{Coulomb}(r_{ij})} \right]$$

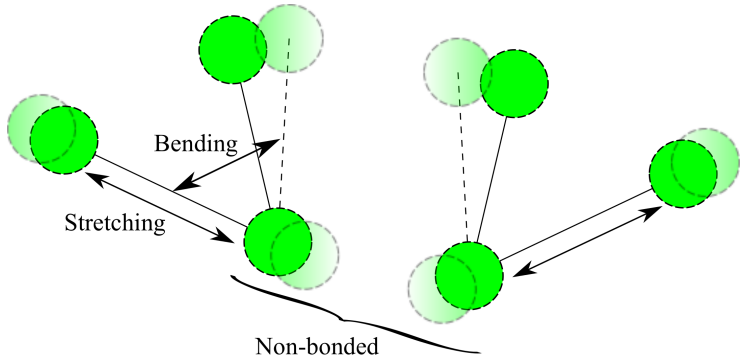
- LJ parameters $\{\varepsilon_{\alpha\beta}, \sigma_{\alpha\beta}\}$ between atoms types α and β is defined for each homoatomic pair present in the system, i.e. $\alpha = \beta$. E.g. $O \sim O$, $H \sim H$, etc.
- Cross-interactions $\{\varepsilon_{\alpha\beta}, \sigma_{\alpha\beta}\}$, $\alpha \neq \beta$, based on Lorentz-Berthelot (LB) rules:

$$\sigma_{\alpha\beta} = (\sigma_\alpha + \sigma_\beta)/2, \qquad \text{and} \qquad \varepsilon_{\alpha\beta} = \sqrt{\varepsilon_\alpha \varepsilon_\beta}.$$

- Parameters: silica (Lopes,2006), water (Jorgensen,1984), ions (Patra,2002).

MD Potential

- $\Phi_{total} = \Phi_{bonded} + \underbrace{\Phi_{LJ} + \Phi_{Coulomb}}_{non-bonded}$
- Φ_{bonded} (bond stretching and bending) is modeled using harmonic potential.



- Non-bonded interactions (Van der Waals + Electrostatic) are modeled as:

$$\Phi_{non-bonded} = \sum_{i=1,j>i}^{n_{atoms}} \left[\underbrace{4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right]}_{\Phi_{LJ}(r_{ij})} + \underbrace{\frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}}_{\Phi_{Coulomb}(r_{ij})} \right]$$

- LJ parameters $\{\varepsilon_{\alpha\beta}, \sigma_{\alpha\beta}\}$ between atoms types α and β is defined for each homoatomic pair present in the system, i.e. $\alpha = \beta$. E.g. $O \sim O$, $H \sim H$, etc.
- Cross-interactions $\{\varepsilon_{\alpha\beta}, \sigma_{\alpha\beta}\}$, $\alpha \neq \beta$, based on Lorentz-Berthelot (LB) rules:

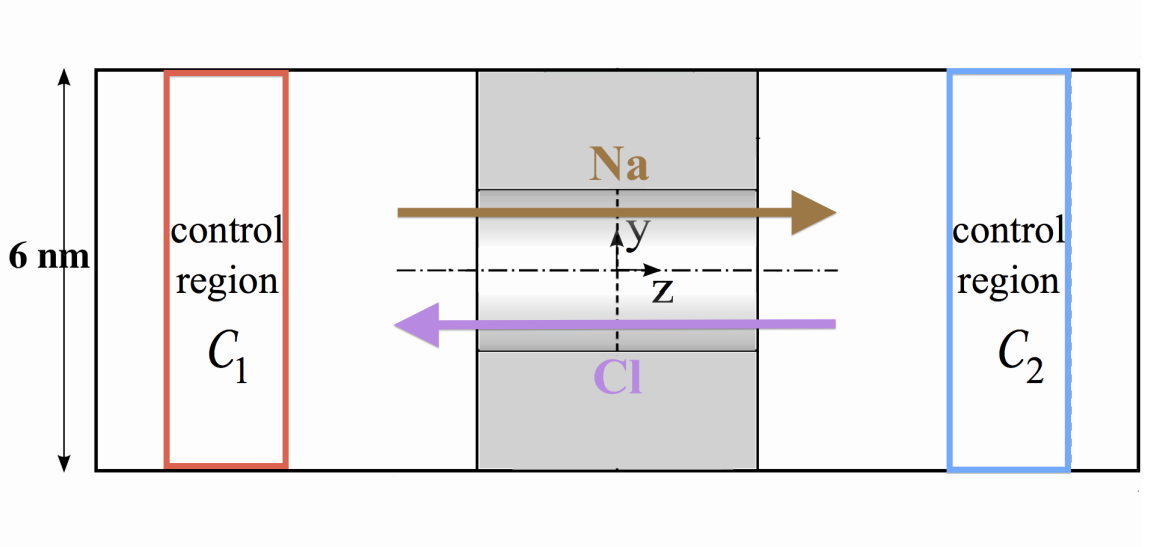
$$\sigma_{\alpha\beta} = (\sigma_\alpha + \sigma_\beta)/2, \qquad \text{and} \qquad \varepsilon_{\alpha\beta} = \sqrt{\varepsilon_\alpha \varepsilon_\beta}.$$

- Parameters: silica (Lopes,2006), water (Jorgensen,1984), ions (Patra,2002).

Concentration Control Algorithm

- Monitor concentration difference $\Delta c(t) = c_2(t) - c_1(t)$:
  $c_i(t) =$ (molar) concentration at t of a target ionic species in i-th reservoir.

- Inject/remove ions in $\mathcal{C}_1$ and $\mathcal{C}_2$.
- No ion deletion, only swapping:
  $\Rightarrow N_{atoms}$ is constant.
- $-\overline{\Delta c}_{Na^+} = 60/V_{fluid} = \overline{\Delta c}_{Cl^-}$
  Flow Na^+: left $\rightarrow$ right
  Flow Cl^- : left $\leftarrow$ right

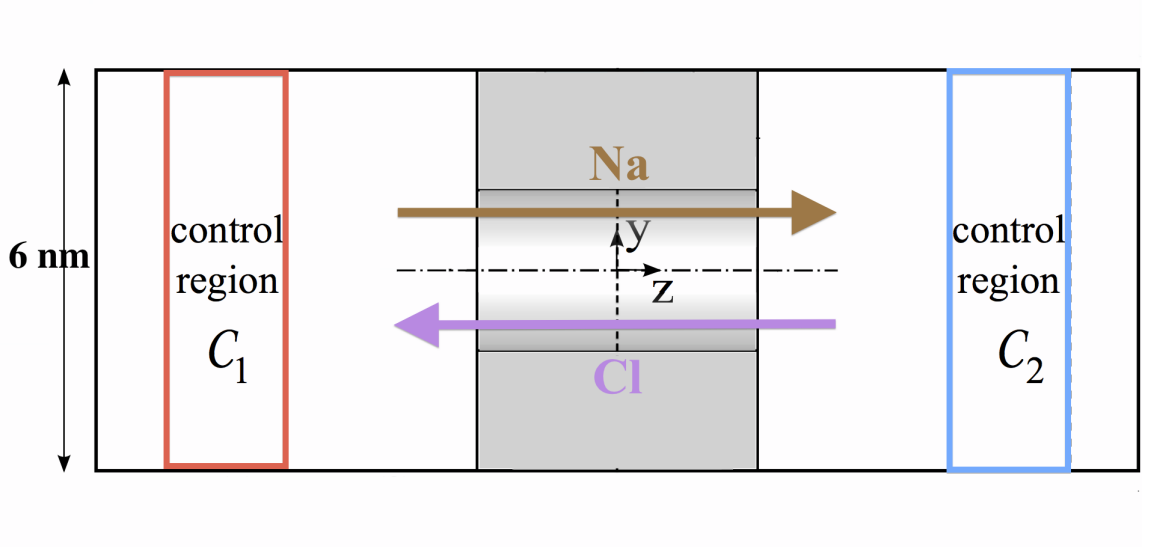


- Ionic flux (magnitude): $J(t) = \frac{N_{exchanges}(t)}{tA}$ and conductance: $G(t) = \frac{J(t)}{|\overline{\Delta c}|}$
  - $N_{exchanges}$ is the number of ion exchanges between $\mathcal{C}_1$ and $\mathcal{C}_2$ over the time t.
  - A is the nominal cross-sectional area of the pore.
  - Validated against a steady flux measured via integration of the velocity profiles of the ions over the cross-section of the pore.

Concentration Control Algorithm

- Monitor concentration difference $\Delta c(t) = c_2(t) - c_1(t)$:
  $c_i(t) =$ (molar) concentration at t of a target ionic species in i-th reservoir.

- Inject/remove ions in $\mathcal{C}_1$ and $\mathcal{C}_2$.
- No ion deletion, only swapping:
  $\Rightarrow N_{atoms}$ is constant.
- $-\overline{\Delta c}_{Na^+} = 60/V_{fluid} = \overline{\Delta c}_{Cl^-}$
  Flow Na^+: left $\rightarrow$ right
  Flow Cl^- : left $\leftarrow$ right



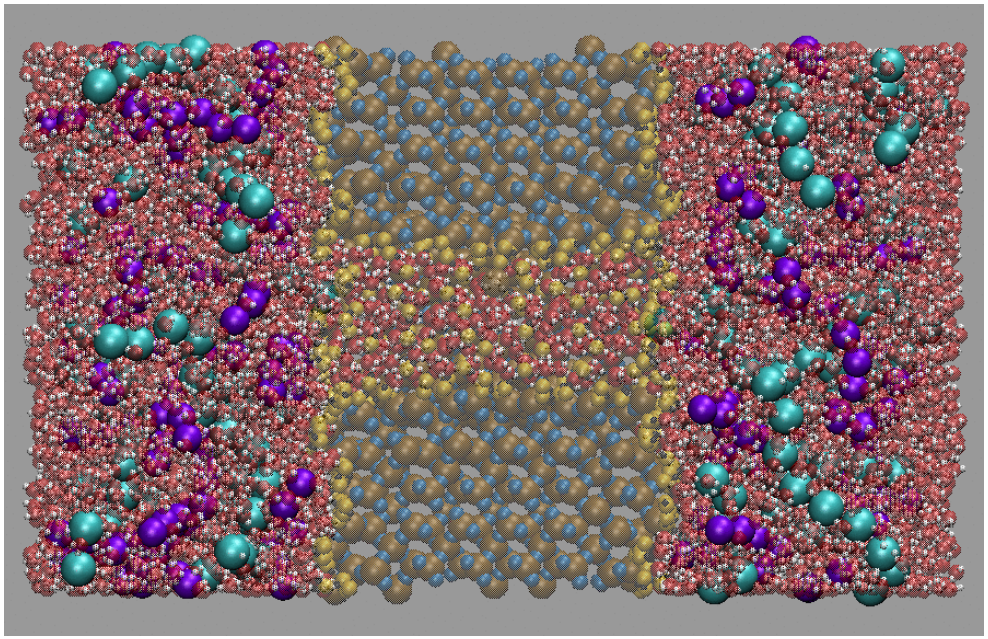
- Ionic flux (magnitude): $J(t) = \frac{N_{exchanges}(t)}{tA}$ and conductance: $G(t) = \frac{J(t)}{|\overline{\Delta c}|}$
  - $N_{exchanges}$ is the number of ion exchanges between $\mathcal{C}_1$ and $\mathcal{C}_2$ over the time t.
  - A is the nominal cross-sectional area of the pore.
  - Validated against a steady flux measured via integration of the velocity profiles of the ions over the cross-section of the pore.

Dependence on the pore diameter

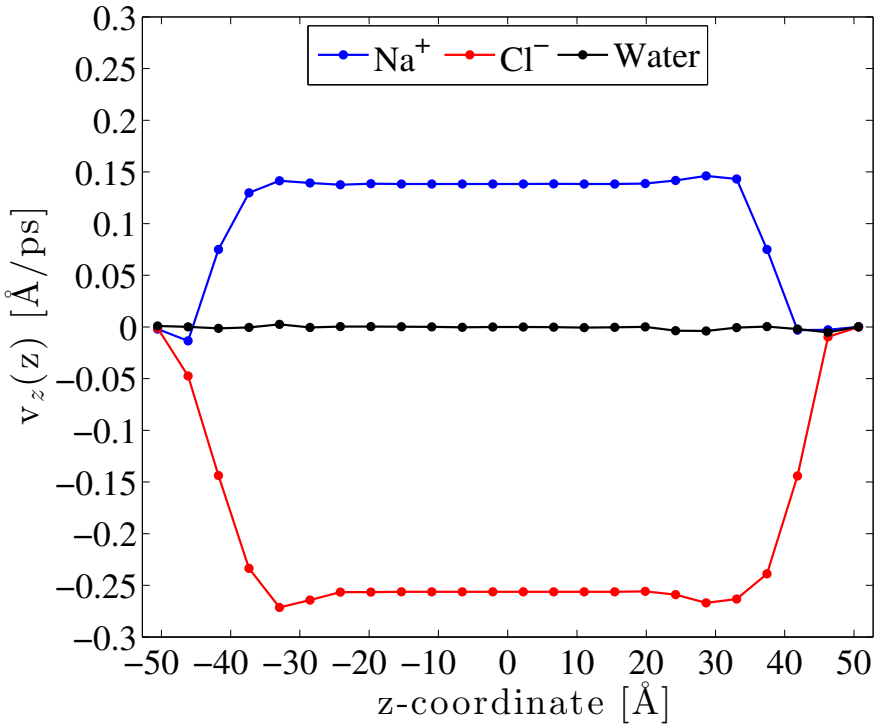
Animation & Velocity Profile

- $D = 12.5$, 17, 21 and 27 Å.
- 5 replica simulations at each D to account for intrinsic (thermal) noise.
- Different initial velocities and random seed for CC algorithm.

- Time/spatial averaging of axial velocity over 24 slabs orthogonal to pore axis.
- Water is stationary.
- Ions tend to flow along the pore centerline with net mean velocity.



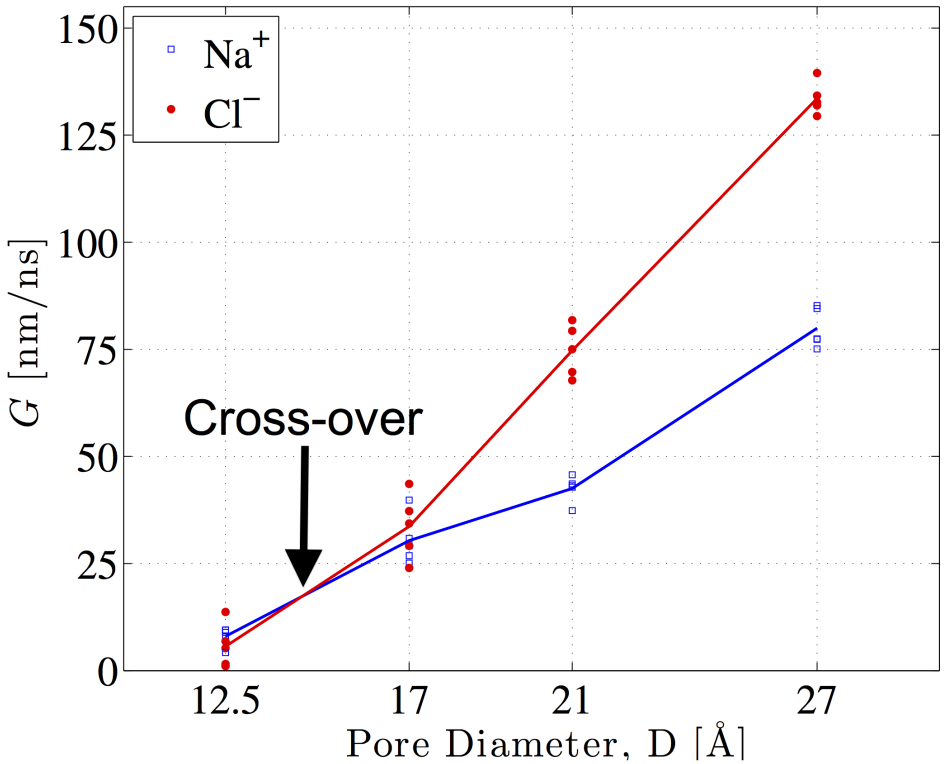
Na^+, **Cl^-**, **H_2O** (white-red), **O_{bulk}**, **Si**, **OH**



Effect of the Pore Diameter: Conductance

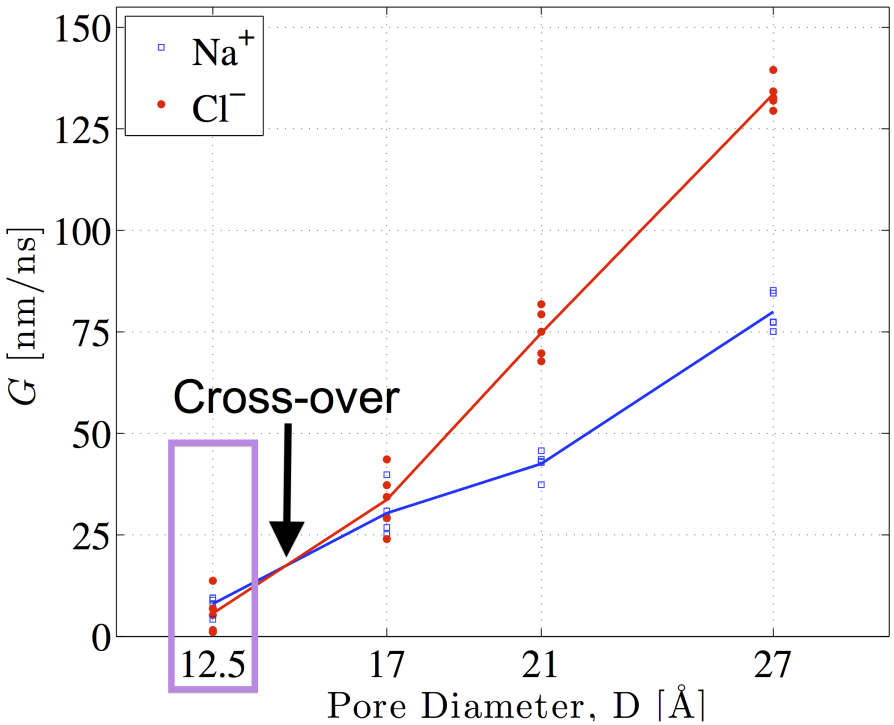
- Extract steady state value of conductance G for Na^+ and Cl^-.
- Steady state when the coefficient of variation based on 500 values is below 1%.
- Steady-state value of G_{Na+} and G_{Cl-} as a function of D for all 5 replicas showing the replica values (markers) and the mean trends (solid lines).

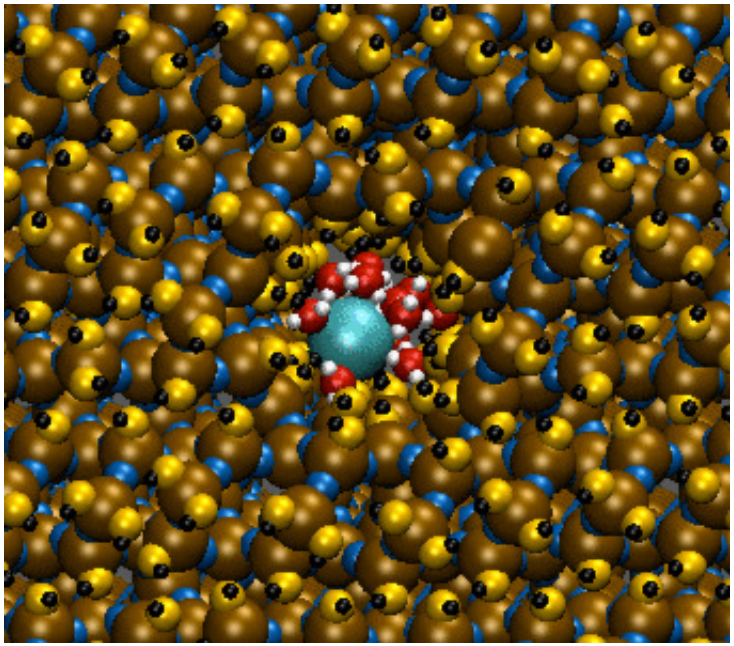
- Slope of G_{Cl-} is sharper than G_{Na+}.
- Overlapping of distributions for small D.
- For $D \geq 17$: $\overline{G}_{Cl-} > \overline{G}_{Na+}$.
- The trend reverses for $D = 12.5$.
- Physical explanation?



Physical Explanation

- Cross-over is the result of the **interplay** between **size effects** and **ionic mobility**.
- $D = 12.5$ Å: weak solvation shell $\Rightarrow$ strong effects of pore walls and confinement favor ions with smaller ionic radius, i.e. Na^+ (as seen by Lyndenbell,1996).
- $D \geq 17$ Å: complete solvation shell around the ions
  $\Rightarrow$ ion's mobility dominates
  $\Rightarrow Flux_{Cl^-} > Flux_{Na^+}$ because the diffusivity of Cl^- is larger.

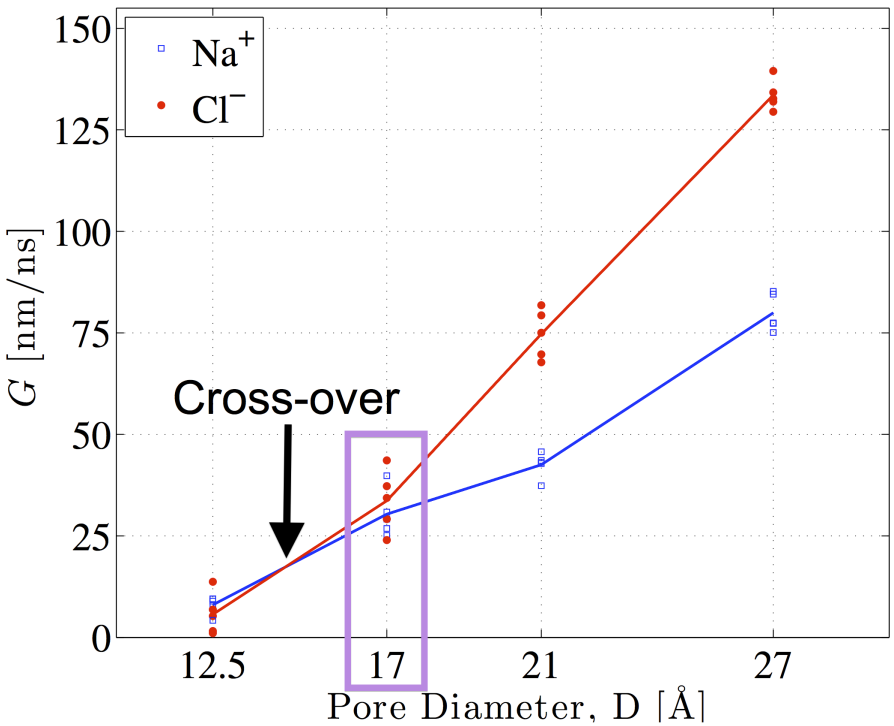


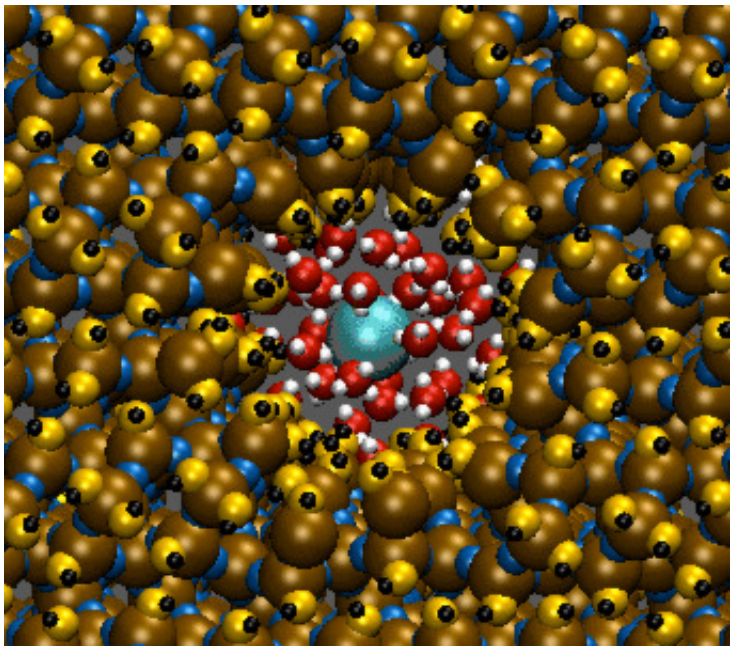


Na^+, Cl^-, H_2O (white-red), OH, Si, O_{bulk}

Physical Explanation

- Cross-over is the result of the **interplay** between **size effects** and **ionic mobility**.
- $D = 12.5$ Å: weak solvation shell $\Rightarrow$ strong effects of pore walls and confinement favor ions with smaller ionic radius, i.e. Na^+ (as seen by Lyndenbell,1996).
- $D \geq 17$ Å: complete solvation shell around the ions
  $\Rightarrow$ ion's mobility dominates
  $\Rightarrow Flux_{Cl-} > Flux_{Na+}$ because the diffusivity of Cl^- is larger.

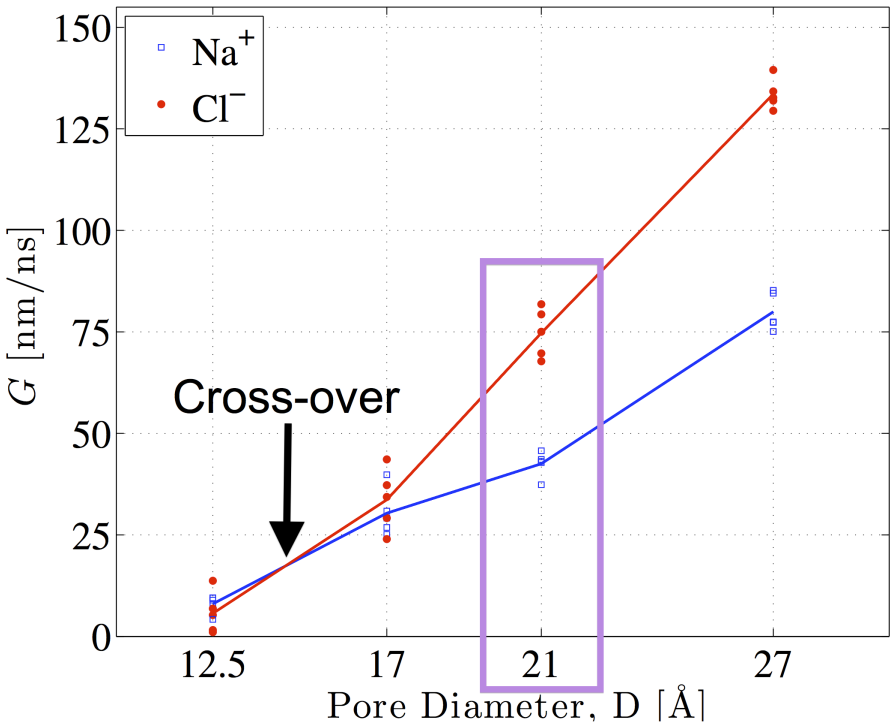


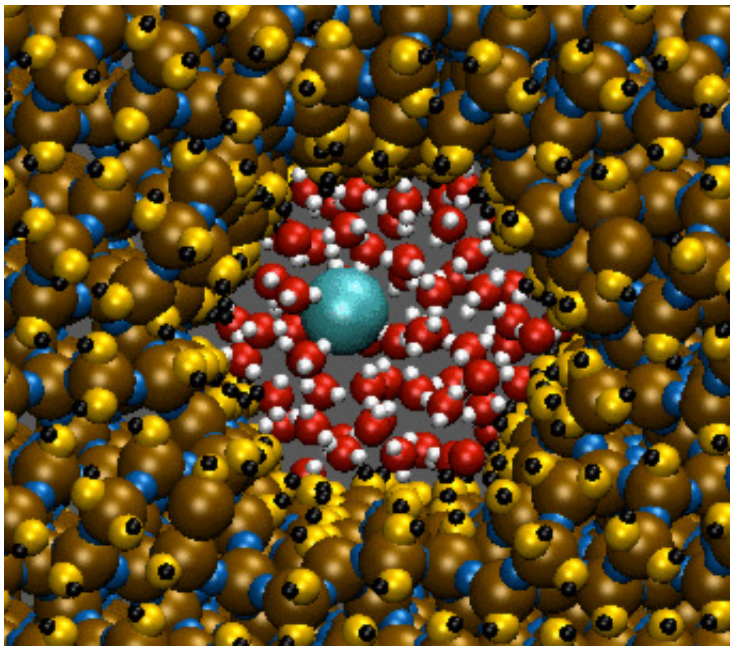


Na^+, Cl^-, H_2O (white-red), **OH**, **Si**, O_{bulk}

Physical Explanation

- Cross-over is the result of the **interplay** between **size effects** and **ionic mobility**.
- $D = 12.5$ Å: weak solvation shell $\Rightarrow$ strong effects of pore walls and confinement favor ions with smaller ionic radius, i.e. Na^+ (as seen by Lyndenbell,1996).
- $D \geq 17$ Å: complete solvation shell around the ions
  $\Rightarrow$ ion's mobility dominates
  $\Rightarrow Flux_{Cl-} > Flux_{Na+}$ because the diffusivity of Cl^- is larger.

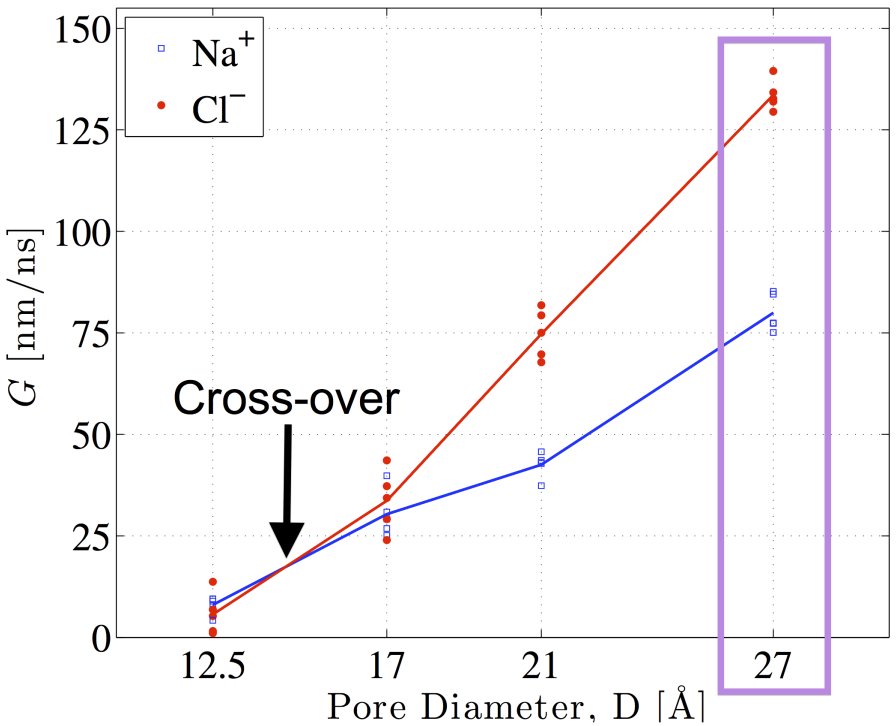


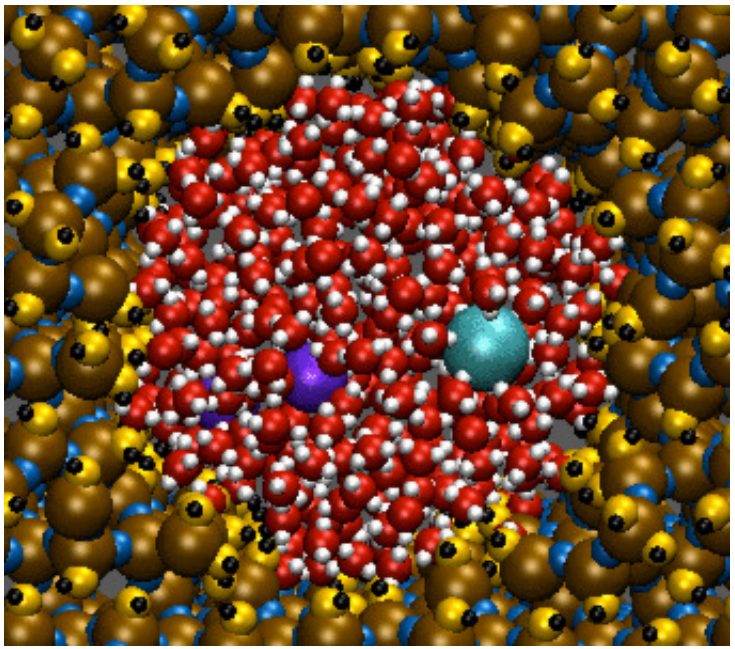


Na^+, Cl^-, H_2O (white-red), **O**H, Si, O_{bulk}

Physical Explanation

- Cross-over is the result of the **interplay** between **size effects** and **ionic mobility**.
- $D = 12.5$ Å: weak solvation shell $\Rightarrow$ strong effects of pore walls and confinement favor ions with smaller ionic radius, i.e. Na^+ (as seen by Lyndenbell,1996).
- $D \geq 17$ Å: complete solvation shell around the ions
  $\Rightarrow$ ion's mobility dominates
  $\Rightarrow Flux_{Cl-} > Flux_{Na+}$ because the diffusivity of Cl^- is larger.





Na^+, **Cl^-**, H_2O (white-red), **OH**, **Si**, **O_{bulk}**

Sensitivity to LJ potential parameters

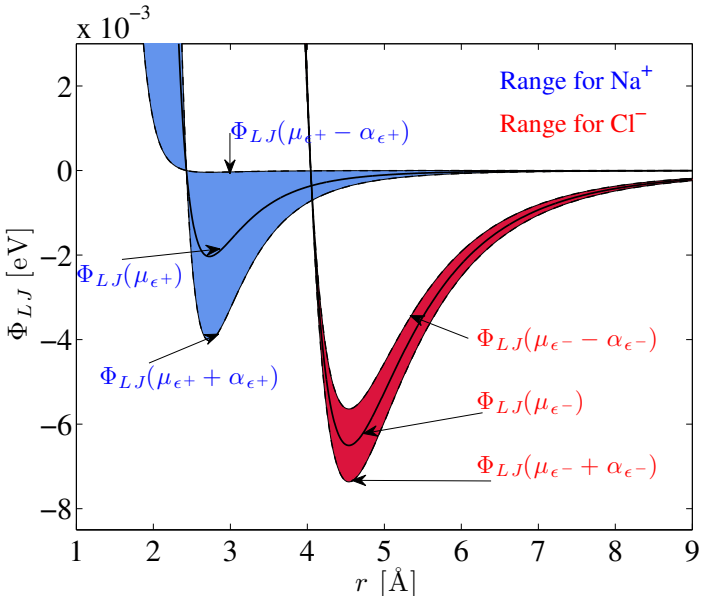
Problem Definition

- Fix $D = 21$ Å; choose ε_{Na^+} and ε_{Cl^-}, depths of the LJ potential for Na^+ and Cl^-.

$$\begin{aligned} \varepsilon_{Na^+} &= 0.002033777 + 0.001992370\ \xi_1, \qquad [\text{eV}], \\ \varepsilon_{Cl^-} &= 0.006504600 + 0.000863055\ \xi_2, \qquad [\text{eV}], \end{aligned}$$

  where $\{\xi_1, \xi_2\}$ are *i.i.d.* uniform random variables $\mathcal{U}(-1, 1)$; values from literature.

- Directly affects the LJ potential for Na^+-Na^+ and Cl^--Cl^- interactions.
- Since $\varepsilon_{\alpha\beta} = \sqrt{\varepsilon_\alpha \varepsilon_\beta}$ for atom types $\alpha \neq \beta$, it affects *all* the cross-interactions.



$\Phi_{LJ}(r)$ for Na^+-Na^+ and Cl^--Cl^-

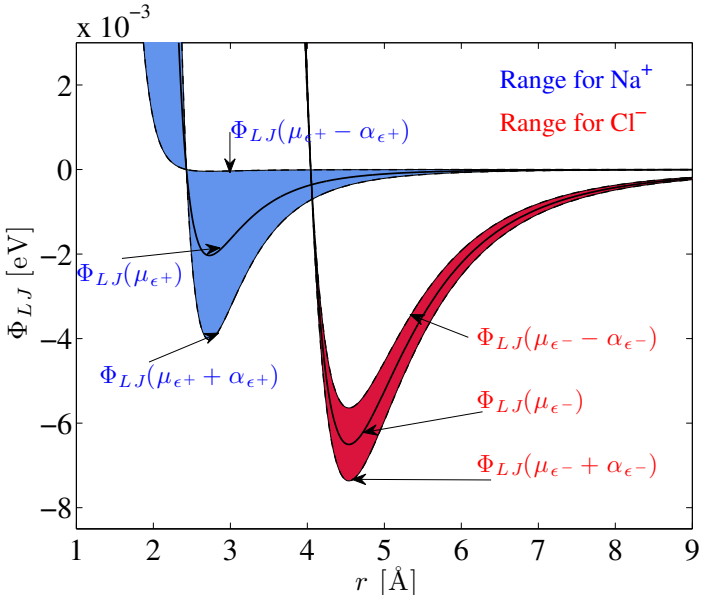
Problem Definition

- Fix $D = 21$ Å; choose ε_{Na^+} and ε_{Cl^-}, depths of the LJ potential for Na^+ and Cl^-.

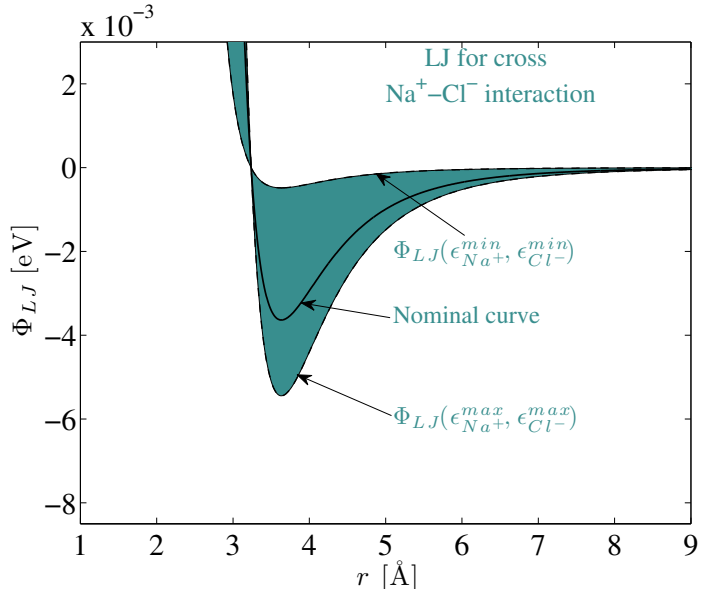
$$\varepsilon_{Na^+} = 0.002033777 + 0.001992370\,\xi_1, \qquad \text{[eV]},$$
$$\varepsilon_{Cl^-} = 0.006504600 + 0.000863055\,\xi_2, \qquad \text{[eV]},$$

  where $\{\xi_1, \xi_2\}$ are *i.i.d.* uniform random variables $\mathcal{U}(-1, 1)$; values from literature.

- Directly affects the LJ potential for Na^+-Na^+ and Cl^--Cl^- interactions.
- Since $\varepsilon_{\alpha\beta} = \sqrt{\varepsilon_\alpha \varepsilon_\beta}$ for atom types $\alpha \neq \beta$, it affects *all* the cross-interactions.



$\Phi_{LJ}(r)$ for Na^+-Na^+ and Cl^--Cl^-



$\Phi_{LJ}(r)$ for cross Na^+-Cl^-

Objective and Methods

- $\varepsilon_{Na^+} = f_1(\xi_1), \quad \varepsilon_{Cl^-} = f_2(\xi_2), \quad$ with $\quad \xi_{1,2} \sim \mathcal{U}(-1,1)$

$\Rightarrow$ nanopore observables (flux, conductance) can be considered as random variables.

- How to "map" the uncertainty from $\varepsilon_{Na^+}, \varepsilon_{Cl^-}$ to the observables?
- E.g. Monte Carlo: Sample ξ_1, ξ_2 to generate samples of ε_{Na^+} and ε_{Cl^-}; run one full MD simulation for each sample; collect data and generate statistics.
- We rely on Polynomial Chaos expansions (PCe):

$$(conductance)\ G \approx \sum_{i=0}^{P} g_i \Psi_i(\xi_1, \xi_2) \qquad [\text{e.g. linear PCe: } \ G \approx g_0 + g_1\xi_1 + g_2\xi_2]$$

- $P+1 = (order+2)!/(order!\,2!)$; $\Psi() =$ Legendre Polyn., and $\mathbf{g} =$ PC coefficients.
- PCe is an orthogonal expansion: scalar product, then pseudo-spectral approach.
  - ⋆ Regularity of data; noisy might be problematic; constraints on sampling.
- For noisy systems, as MD due to thermal noise, Bayesian regression works best.

Objective and Methods

- $\varepsilon_{Na^+} = f_1(\xi_1), \quad \varepsilon_{Cl^-} = f_2(\xi_2), \quad$ with $\quad \xi_{1,2} \sim \mathcal{U}(-1,1)$

$\Rightarrow$ nanopore observables (flux, conductance) can be considered as random variables.

- How to "map" the uncertainty from $\varepsilon_{Na^+}, \varepsilon_{Cl^-}$ to the observables?
- E.g. Monte Carlo: Sample ξ_1, ξ_2 to generate samples of ε_{Na^+} and ε_{Cl^-}; run one full MD simulation for each sample; collect data and generate statistics.
- We rely on Polynomial Chaos expansions (PCe):

$$(conductance)\ G \approx \sum_{i=0}^{P} g_i \Psi_i(\xi_1, \xi_2) \qquad [\text{e.g. linear PCe: } \ G \approx g_0 + g_1\xi_1 + g_2\xi_2]$$

- $P+1 = (order+2)!/(order!\,2!)$; $\Psi() =$ Legendre Polyn., and $\mathbf{g} =$ PC coefficients.
- PCe is an orthogonal expansion: scalar product, then pseudo-spectral approach.
  - ⋆ Regularity of data; noisy might be problematic; constraints on sampling.
- For noisy systems, as MD due to thermal noise, Bayesian regression works best.

Objective and Methods

- $\varepsilon_{Na^+} = f_1(\xi_1), \quad \varepsilon_{Cl^-} = f_2(\xi_2), \quad$ with $\quad \xi_{1,2} \sim \mathcal{U}(-1,1)$

$\Rightarrow$ nanopore observables (flux, conductance) can be considered as random variables.

- How to "map" the uncertainty from $\varepsilon_{Na^+}, \varepsilon_{Cl^-}$ to the observables?
- E.g. Monte Carlo: Sample ξ_1, ξ_2 to generate samples of ε_{Na^+} and ε_{Cl^-}; run one full MD simulation for each sample; collect data and generate statistics.
- We rely on Polynomial Chaos expansions (PCe):

$$(conductance)\ G \approx \sum_{i=0}^{P} g_i \Psi_i(\xi_1, \xi_2) \qquad [\text{e.g. linear PCe:} \quad G \approx g_0 + g_1\xi_1 + g_2\xi_2]$$

- $P+1 = (order+2)!/(order!\,2!)$; $\Psi() =$ Legendre Polyn., and $\mathbf{g} =$ PC coefficients.
- PCe is an orthogonal expansion: scalar product, then pseudo-spectral approach.
  - ⋆ Regularity of data; noisy might be problematic; constraints on sampling.
- For noisy systems, as MD due to thermal noise, Bayesian regression works best.

Objective and Methods

- $\varepsilon_{Na^+} = f_1(\xi_1), \quad \varepsilon_{Cl^-} = f_2(\xi_2), \quad$ with $\quad \xi_{1,2} \sim \mathcal{U}(-1,1)$

$\Rightarrow$ nanopore observables (flux, conductance) can be considered as random variables.

- How to "map" the uncertainty from $\varepsilon_{Na^+}, \varepsilon_{Cl^-}$ to the observables?
- E.g. Monte Carlo: Sample ξ_1, ξ_2 to generate samples of ε_{Na^+} and ε_{Cl^-}; run one full MD simulation for each sample; collect data and generate statistics.
- We rely on Polynomial Chaos expansions (PCe):

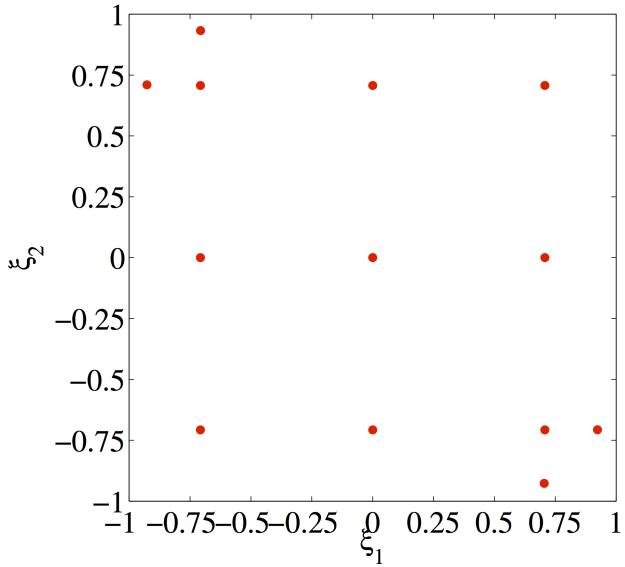
$$(conductance)\ G \approx \sum_{i=0}^{P} g_i \Psi_i(\xi_1, \xi_2) \qquad \text{[e.g. linear PCe: } G \approx g_0 + g_1\xi_1 + g_2\xi_2\text{]}$$

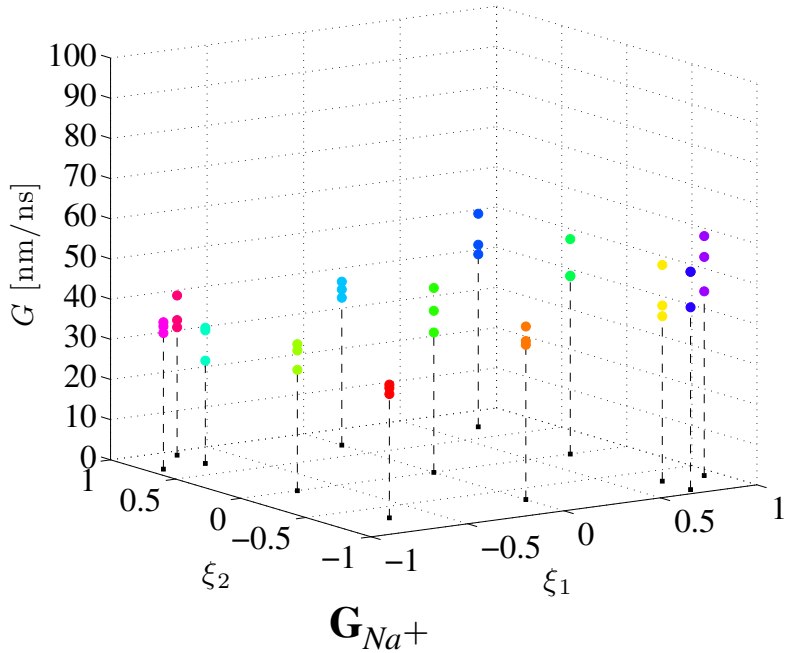
- $P+1 = (order+2)!/(order!\,2!)$; $\Psi() =$ Legendre Polyn., and $\mathbf{g} =$ PC coefficients.
- PCe is an orthogonal expansion: scalar product, then pseudo-spectral approach.
  - ⋆ Regularity of data; noisy might be problematic; constraints on sampling.
- For noisy systems, as MD due to thermal noise, Bayesian regression works best.

Bayesian Regression: Collecting Data

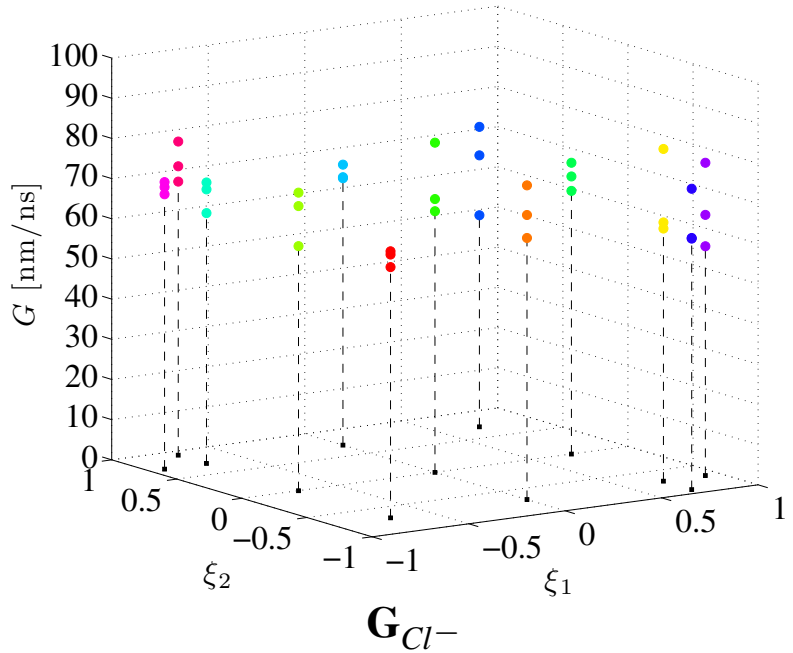
- $\varepsilon_{Na^+} = f_1(\xi_1)$ and $\varepsilon_{Cl^-} = f_2(\xi_2)$.
- Sampling grid of 13 nodes (3 replicas each).
- Data-set of steady-state conductance values:

$$\mathbf{G}_{Na^+,Cl^-} = \left\{ G_{i,j}^{Na^+,Cl^-} \right\}_{i=1,\ldots,13}^{j=1,\ldots,3},$$





$\mathbf{G}_{Na^+}$

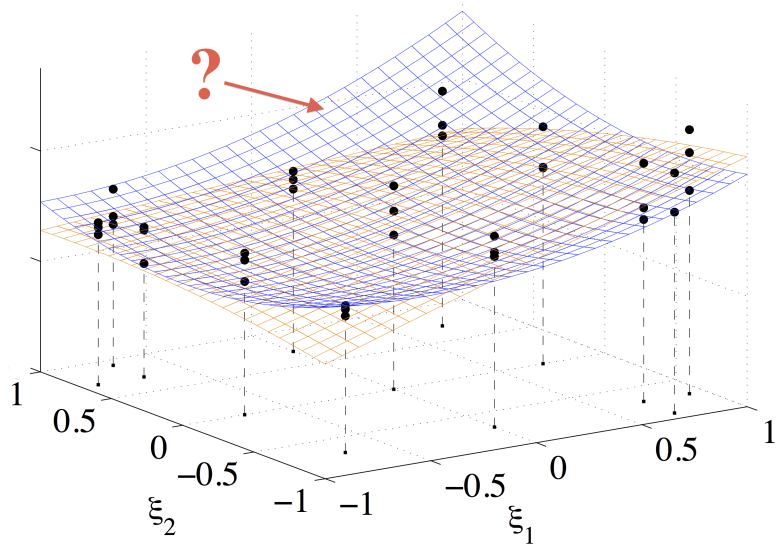


$\mathbf{G}_{Cl^-}$

Bayesian Regression: Formulation

- Regression function is a PCe:

$$M(\xi_1, \xi_2) = \sum_{k=0}^{P} g_k \Psi_k(\xi_1, \xi_2),$$



- Regression model as

$$G_\ell = M(\boldsymbol{\xi}_\ell) + \gamma_\ell, \qquad \ell = 1, \ldots, 39.$$

  - $\boldsymbol{\xi}_\ell$ is the coordinate of the ℓ-th data point G_ℓ
  - γ_ℓ is a RV capturing the discrepancy between data and model prediction.

- Data points result from independent but statistically equivalent MD runs.
- Assume $\{\gamma_\ell\}_{\ell=1}^{39}$ to be independent and $\gamma_\ell \sim \mathcal{N}(0, \sigma_\ell)$, $\ell = 1, \ldots, 39$.
  - ⋆ Gaussian model (verified): data extracted from MD using running averages.
- Consider a space-dependent noise $\sigma_\ell = \sigma(\boldsymbol{\xi})$ by parametrizing:

$$\sigma(\boldsymbol{\xi}) = h_0 + h_1\xi_1 + h_2\xi_2.$$

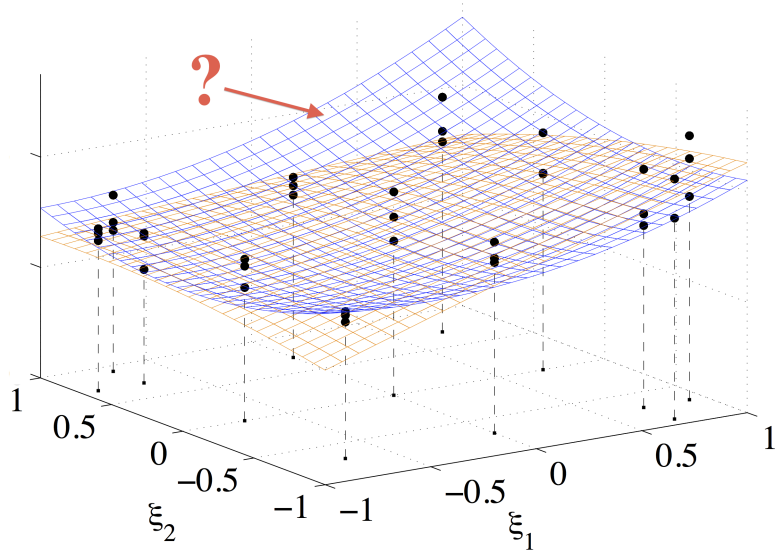
Bayesian Regression: Formulation

- Regression function is a PCe:

$$M(\xi_1, \xi_2) = \sum_{k=0}^{P} g_k \Psi_k(\xi_1, \xi_2),$$

- Regression model as

$$G_\ell = M(\boldsymbol{\xi}_\ell) + \gamma_\ell, \qquad \ell = 1, \ldots, 39.$$



  - $\boldsymbol{\xi}_\ell$ is the coordinate of the ℓ-th data point G_ℓ
  - γ_ℓ is a RV capturing the discrepancy between data and model prediction.

- Data points result from independent but statistically equivalent MD runs.
- Assume $\{\gamma_\ell\}_{\ell=1}^{39}$ to be independent and $\gamma_\ell \sim \mathcal{N}(0, \sigma_\ell)$, $\ell = 1, \ldots, 39$.
  - ⋆ Gaussian model (verified): data extracted from MD using running averages.
- Consider a space-dependent noise $\sigma_\ell = \sigma(\xi)$ by parametrizing:

$$\sigma(\boldsymbol{\xi}) = h_0 + h_1 \xi_1 + h_2 \xi_2.$$

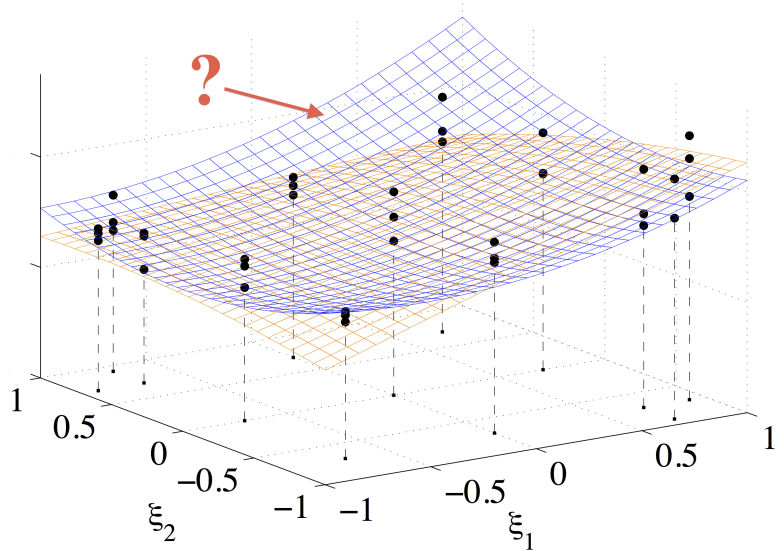
Bayesian Regression: Formulation

- Regression function is a PCe:

$$M(\xi_1, \xi_2) = \sum_{k=0}^{P} g_k \Psi_k(\xi_1, \xi_2),$$

- Regression model as

$$G_\ell = M(\boldsymbol{\xi}_\ell) + \gamma_\ell, \qquad \ell = 1, \ldots, 39.$$



  - $\boldsymbol{\xi}_\ell$ is the coordinate of the ℓ-th data point G_ℓ
  - γ_ℓ is a RV capturing the discrepancy between data and model prediction.

- Data points result from independent but statistically equivalent MD runs.
- Assume $\{\gamma_\ell\}_{\ell=1}^{39}$ to be independent and $\gamma_\ell \sim \mathcal{N}(0, \sigma_\ell)$, $\ell = 1, \ldots, 39$.
  - ⋆ Gaussian model (verified): data extracted from MD using running averages.
- Consider a space-dependent noise $\sigma_\ell = \sigma(\xi)$ by parametrizing:

$$\sigma(\boldsymbol{\xi}) = h_0 + h_1\xi_1 + h_2\xi_2.$$

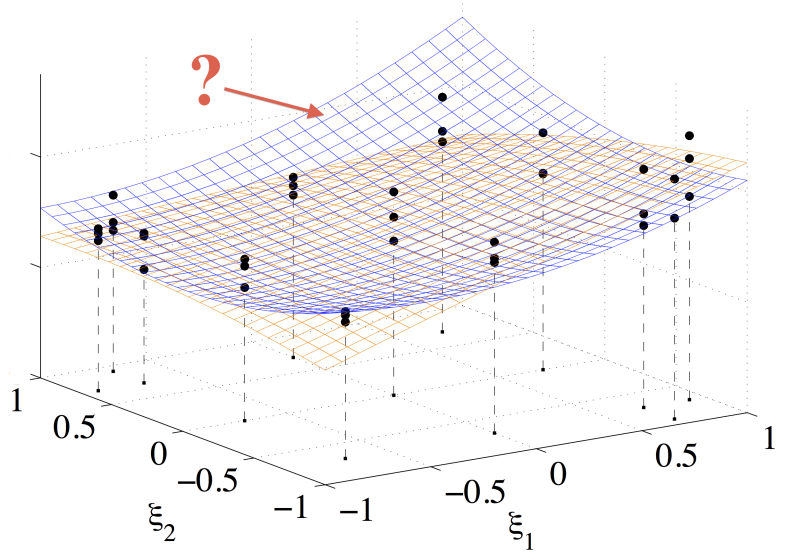
Bayesian Regression: Formulation

- Regression function is a PCe:

$$M(\xi_1, \xi_2) = \sum_{k=0}^{P} g_k \Psi_k(\xi_1, \xi_2),$$

- Regression model as

$$G_\ell = M(\boldsymbol{\xi}_\ell) + \gamma_\ell, \qquad \ell = 1, \ldots, 39.$$



  - $\boldsymbol{\xi}_\ell$ is the coordinate of the ℓ-th data point G_ℓ
  - γ_ℓ is a RV capturing the discrepancy between data and model prediction.

- Data points result from independent but statistically equivalent MD runs.
- Assume $\{\gamma_\ell\}_{\ell=1}^{39}$ to be independent and $\gamma_\ell \sim \mathcal{N}(0, \sigma_\ell)$, $\ell = 1, \ldots, 39$.
  - ⋆ Gaussian model (verified): data extracted from MD using running averages.
- Consider a space-dependent noise $\sigma_\ell = \sigma(\boldsymbol{\xi})$ by parametrizing:

$$\sigma(\boldsymbol{\xi}) = h_0 + h_1\xi_1 + h_2\xi_2.$$

Bayesian Regression: Formulation

- $\{h_k\}_{k=0}^{2}$ are hyperparameters, i.e. part of the unknowns: $\{g_0, \ldots, g_P, h_0, h_1, h_2\}$.
- Likelihood becomes:

$$\mathcal{L} = \prod_{i=1}^{13} \prod_{j=1}^{3} \frac{1}{\sqrt{2\pi[h_0 + h_1\xi_{1,i} + h_2\xi_{2,i}]^2}} \exp\left(- \frac{[G_{i,j} - \sum_{k=0}^{P} g_k \Psi_k(\xi_{1,i}, \xi_{2,i})]^2}{2[h_0 + h_1\xi_{1,i} + h_2\xi_{2,i}]^2} \right),$$

  where $G_{i,j}$ is the j-th observation obtained at the i-th sampling point, $\boldsymbol{\xi}_i$.

- Bayes' theorem yields the joint posterior

$$\pi\left(\{g_k\}_{k=0}^{P}, \{h_l\}_{l=0}^{2} \,\middle|\, \mathbf{G} \right) \propto \mathcal{L}\left(\mathbf{G} \middle| \{g_k\}_{k=0}^{P}, \{h_l\}_{l=0}^{2} \right) \text{Prior}(\{g_k\}_{k=0}^{P}, \{h_l\}_{l=0}^{2}),$$

  For the priors we use uniform distributions.

- Sample $\pi()$ with a Markov chain Monte Carlo (MCMC) method based on Adaptive Metropolis (AM): "walk" in the $\{g_0, \ldots, g_P, h_0, h_1, h_2\}$-space.

Bayesian Regression: Formulation

- $\{h_k\}_{k=0}^{2}$ are hyperparameters, i.e. part of the unknowns: $\{g_0, \ldots, g_P, h_0, h_1, h_2\}$.
- Likelihood becomes:

$$\mathcal{L} = \prod_{i=1}^{13} \prod_{j=1}^{3} \frac{1}{\sqrt{2\pi[h_0 + h_1\xi_{1,i} + h_2\xi_{2,i}]^2}} \exp\left(- \frac{[G_{i,j} - \sum_{k=0}^{P} g_k \Psi_k(\xi_{1,i}, \xi_{2,i})]^2}{2[h_0 + h_1\xi_{1,i} + h_2\xi_{2,i}]^2} \right),$$

  where $G_{i,j}$ is the j-th observation obtained at the i-th sampling point, $\boldsymbol{\xi}_i$.

- Bayes' theorem yields the joint posterior

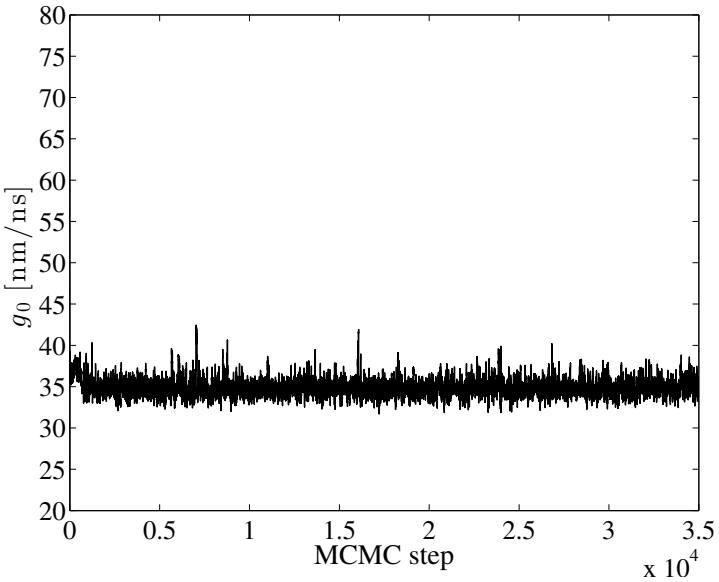
$$\pi\left(\{g_k\}_{k=0}^{P}, \{h_l\}_{l=0}^{2} \,\middle|\, \mathbf{G} \right) \propto \mathcal{L}\left(\mathbf{G} \middle| \{g_k\}_{k=0}^{P}, \{h_l\}_{l=0}^{2} \right) \text{Prior}(\{g_k\}_{k=0}^{P}, \{h_l\}_{l=0}^{2}),$$

  For the priors we use uniform distributions.

- Sample $\pi()$ with a Markov chain Monte Carlo (MCMC) method based on Adaptive Metropolis (AM): "walk" in the $\{g_0, \ldots, g_P, h_0, h_1, h_2\}$-space.

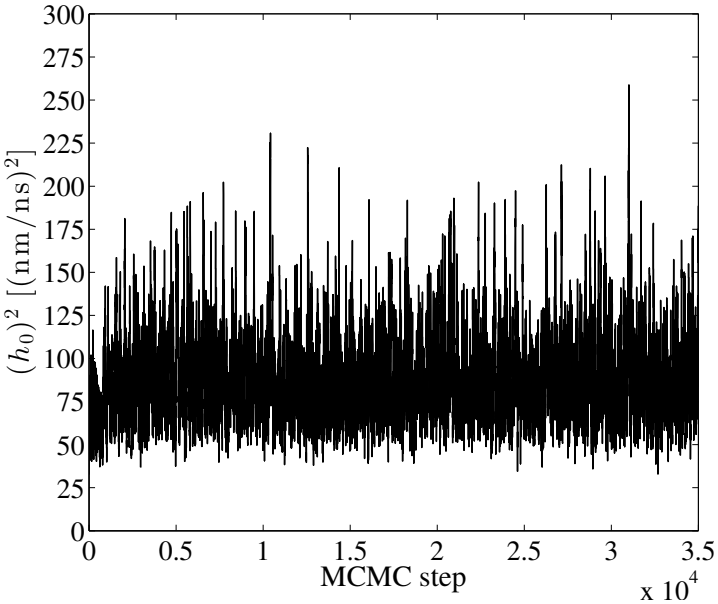
Baysian Regression: Results

- Regression function order: e.g. $ord = 1$ (linear), $ord = 2$ (quadratic), etc.
- $P + 1 = (ord + 2)!/(ord!\ 2!)$.
- MCMC yields a "chain" in the $\{g_0, \ldots, g_P, h_0, h_1, h_2\}$-space:



MCMC for g_0 of the regression function $M(\boldsymbol{\xi}) = \sum_{k=0}^{P} g_k \Psi_k(\boldsymbol{\xi})$ for Na^+.

$\ldots$



MCMC for $(h_0)^2$ of the noise variance $\sigma(\boldsymbol{\xi}) = h_0 + h_1\xi_1 + h_2\xi_2$ for Na^+.

$\ldots$

- Samples are used to derive statistics of the posterior $\pi(g_0, \ldots, g_P, h_0, h_1, h_2)$: mean, variance, joint distributions ...

Regression Function: Linear? Quadratic? ...

- Regression function $M(\xi_1, \xi_2)$: constant? linear? higher-order?
- Bayes factor: discriminate between two "models" describing the same set of data.
- $\theta_p = \{g_0, \ldots, g_P, h_0, h_1, h_2\}$: set of model parameters for a p-th order $M(\xi)$.
- Integrates over the full parameter space.
- The ($\log_n$) of Bayes factor, $B(\theta_{p_1}, \theta_{p_2})$, is given by:

$$\log_n(B(\theta_{p_1}, \theta_{p_2})) = \log_n \frac{\int \mathcal{L}(G|\theta_{p_1}) Pr(\theta_{p_1})\, d\theta_{p_1}}{\int \mathcal{L}(G|\theta_{p_2}) Pr(\theta_{p_2})\, d\theta_{p_2}}$$

- The more positive $\log_n(B(\theta_{p_1}, \theta_{p_2}))$, the stronger the support for θ_{p_1} (Kass,1995).

Regression Function: Linear? Quadratic? ...

- Regression function $M(\xi_1, \xi_2)$: constant? linear? higher-order?
- Bayes factor: discriminate between two "models" describing the same set of data.
- $\boldsymbol{\theta}_p = \{g_0, \ldots, g_P, h_0, h_1, h_2\}$: set of model parameters for a p-th order $M(\boldsymbol{\xi})$.
- Integrates over the full parameter space.
- The ($\log_n$) of Bayes factor, $B(\boldsymbol{\theta}_{p_1}, \boldsymbol{\theta}_{p_2})$, is given by:

$$\log_n(B(\boldsymbol{\theta}_{p_1}, \boldsymbol{\theta}_{p_2})) = \log_n \frac{\int \mathcal{L}(G|\boldsymbol{\theta}_{p_1}) Pr(\boldsymbol{\theta}_{p_1})\, d\boldsymbol{\theta}_{p_1}}{\int \mathcal{L}(G|\boldsymbol{\theta}_{p_2}) Pr(\boldsymbol{\theta}_{p_2})\, d\boldsymbol{\theta}_{p_2}}$$

- The more positive $\log_n(B(\boldsymbol{\theta}_{p_1}, \boldsymbol{\theta}_{p_2}))$, the stronger the support for $\boldsymbol{\theta}_{p_1}$ (Kass,1995).

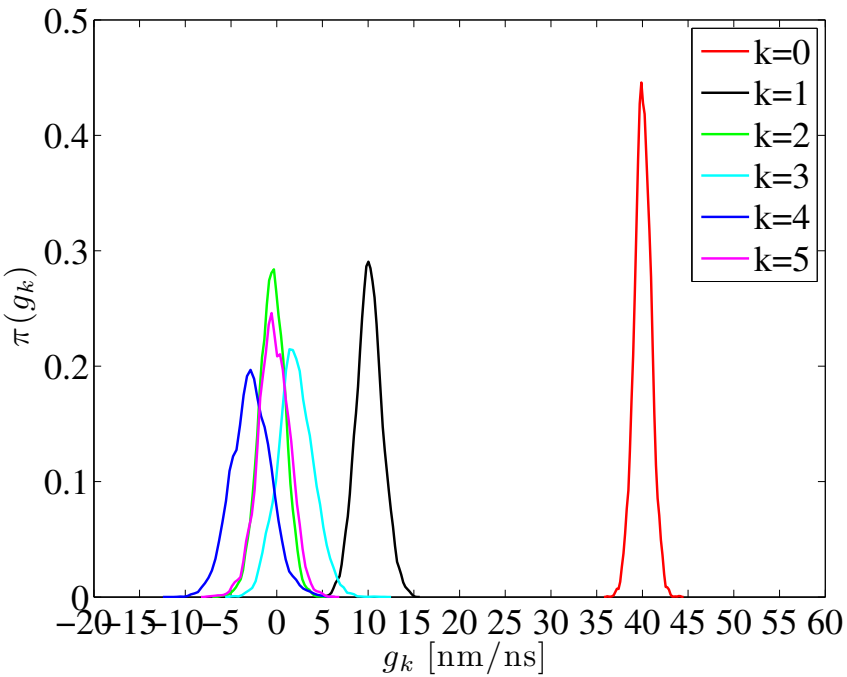
	Na^+				Cl^-			
	$p_2 = 0$	$p_2 = 1$	$p_2 = 2$	$p_2 = 3$	$p_2 = 0$	$p_2 = 1$	$p_2 = 2$	$p_2 = 3$
$p_1 = 0$	–	-19.341	-19.933	-16.285	–	**2.484**	**2.737**	**6.499**
$p_1 = 1$	19.341	–	-0.593	3.055	-2.284	–	0.254	4.015
$p_1 = 2$	**19.933**	**0.593**	–	**3.648**	-2.737	-0.254	–	3.761
$p_1 = 3$	16.285	-3.055	-3.648	–	-6.499	-4.015	-3.761	–

Posterior Uncertainty & Response Surface for Na^+

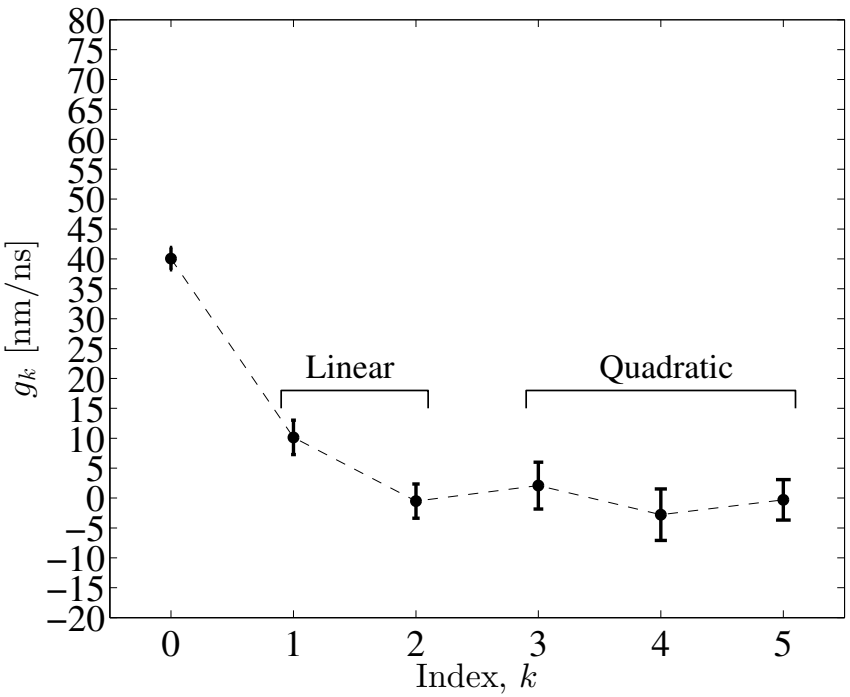
- Quadratic regression model for G_{Na^+}:

$$M_{G_{Na^+}}(\boldsymbol{\xi}) = g_0 + g_1\xi_1 + g_2\xi_2 + g_3(3\xi_1^2 - 1)/2 + g_4\xi_1\xi_2 + g_5(3\xi_2^2 - 1)/2$$

- From Bayesian regression: $\Rightarrow$ *uncertain* PCe.
- Generate sample spectra $\{\mathbf{g}_i\}_{i=1}^{500}$ by sampling $\pi(\mathbf{g}) = \pi(g_0, \ldots, g_5)$
- g_0, g_1 dominant in magnitude, higher-order modes play a minor role.
- Plot sample response surfaces: clear trend present.



Marginalized posteriors of PC coefficients.



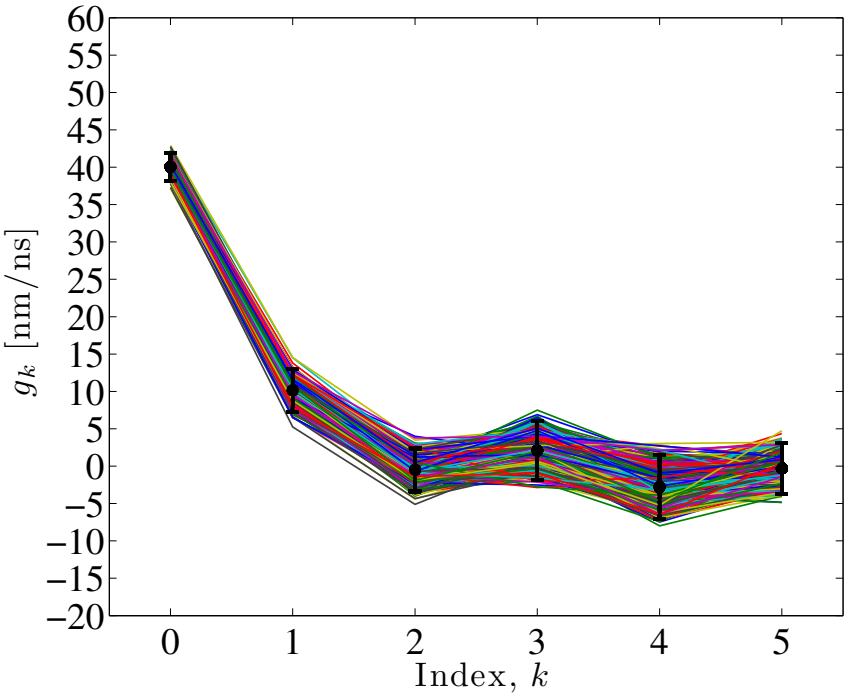
Basic statistics of PC spectrum.

Posterior Uncertainty & Response Surface for Na^+

- Quadratic regression model for G_{Na^+}:

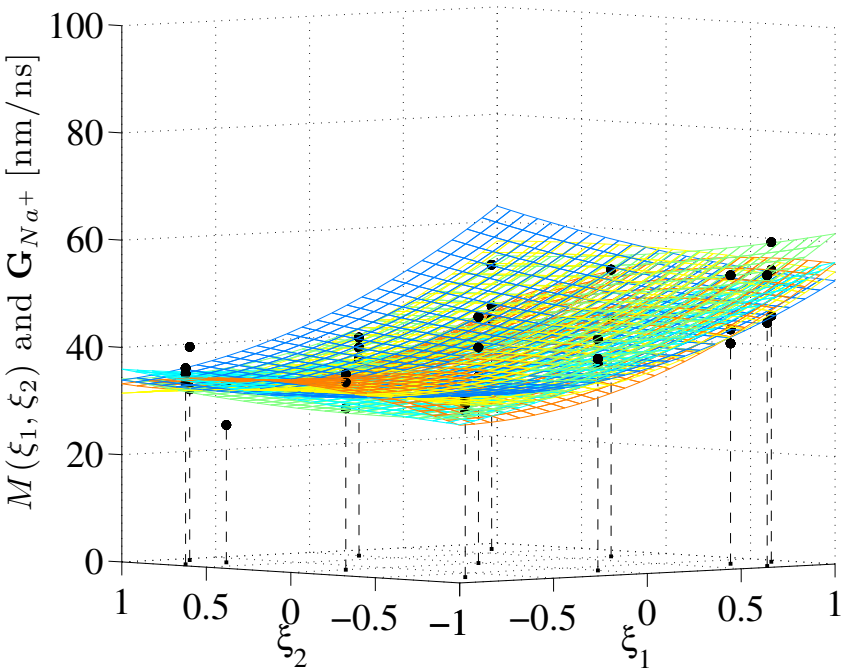
$$M_{G_{Na^+}}(\boldsymbol{\xi}) = g_0 + g_1\xi_1 + g_2\xi_2 + g_3(3\xi_1^2 - 1)/2 + g_4\xi_1\xi_2 + g_5(3\xi_2^2 - 1)/2$$

- From Bayesian regression: $\Rightarrow$ *uncertain* PCe.
- Generate sample spectra $\{\mathbf{g}_i\}_{i=1}^{500}$ by sampling $\pi(\mathbf{g}) = \pi(g_0, \ldots, g_5)$



Sample spectra.

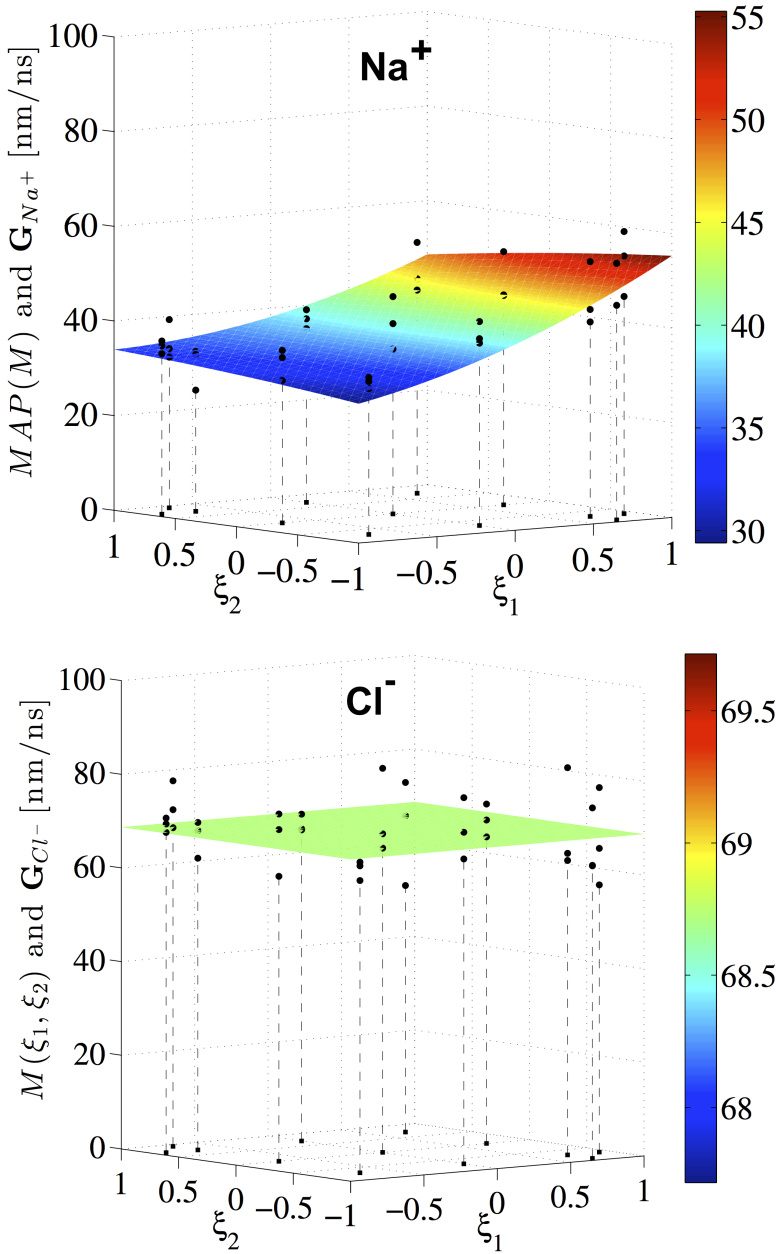
- g_0, g_1 dominant in magnitude, higher-order modes play a minor role.
- Plot sample response surfaces: clear trend present.



Sample response surfaces.

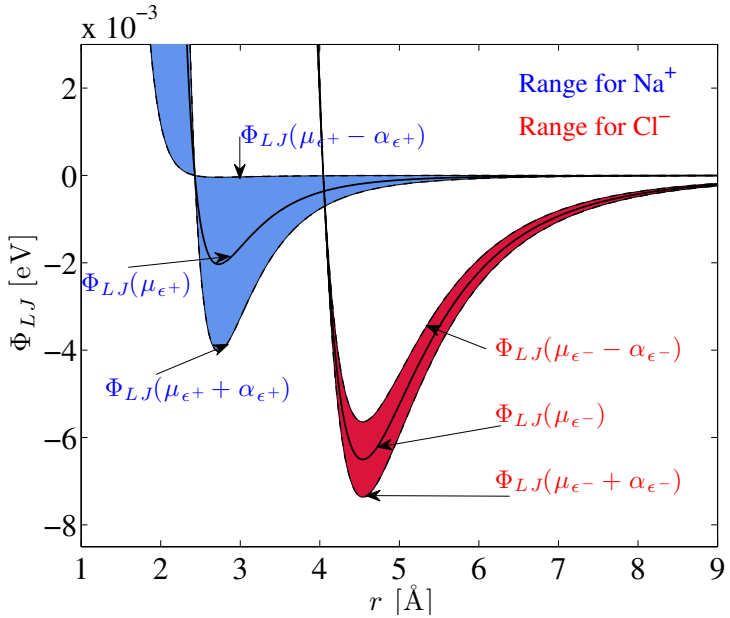
Posterior Uncertainty & Response Surface: Differences

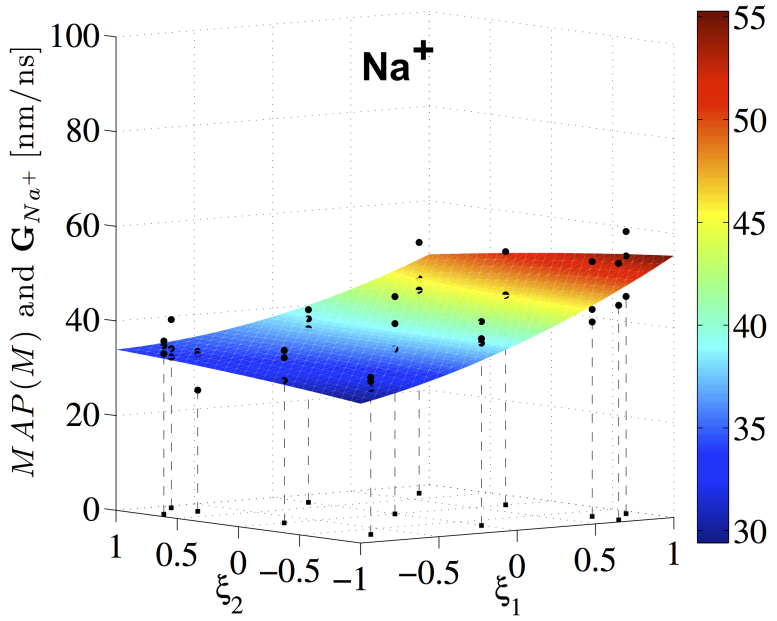
- MAP estimate response surface for Na^+: G_{Na+} increases as ε_{Na+} (i.e. ξ_1) increases.
- For Cl^- Bayes factor supported the use of a *constant M* to represent G_{Cl-}.
- Insight: $G_{Cl-} \sim \varepsilon_{Cl-}$ and $G_{Na+} \sim \varepsilon_{Na+}$ but...
  - smaller uncertainty range for ε_{Cl-} yields a smaller absolute variation for G_{Cl-};
  - trend of G_{Cl-} hidden by the noise level.

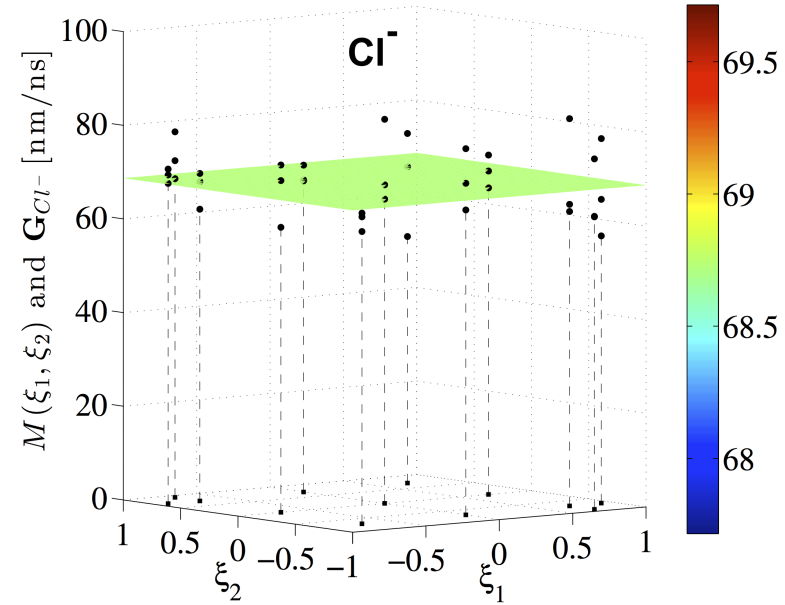


Posterior Uncertainty & Response Surface: Differences

- MAP estimate response surface for Na^+: G_{Na+} increases as ε_{Na+} (i.e. ξ_1) increases.
- For Cl^- Bayes factor supported the use of a *constant M* to represent G_{Cl-}.
- Insight: $G_{Cl-} \sim \varepsilon_{Cl-}$ and $G_{Na+} \sim \varepsilon_{Na+}$ but...
  - smaller uncertainty range for ε_{Cl-} yields a smaller absolute variation for G_{Cl-}.
  - trend of G_{Cl-} hidden by the noise level.







Analysis of the Non-Deterministic PC Model

- The Bayesian regression yields a non-deterministic PC representation:

$$M(\xi_1, \xi_2) = g_0 \Psi_0(\xi_1, \xi_2) + \ldots + g_P \Psi_P(\xi_1, \xi_2)$$

  where $\{g_l\}_{l=0}^{P}$ is a random vector defined by a $(P+1)$-dim density.

- The PC regression model, $M(\xi_1, \xi_2)$, depends on:
  1. the parametric uncertainty in the potential through the RVs ξ_1, ξ_2.
  2. thermal noise through the uncertainty in the PC coefficients.

- Interpretation:
  - Draw m samples of the parameters $\left\{\xi_1^{(j)}, \xi_2^{(j)}\right\}_{j=1}^{m}$
  - For any given $\left\{\xi_1^{(j)}, \xi_2^{(j)}\right\}$, we can draw n different sample-spectra of PC coefficients $\left\{\mathbf{g}_i\right\}_{i=1}^{n}$, from their joint distribution $\pi(\mathbf{g})$.
  - We thus obtain $n \times m$ predictions for the target observable $\left\{\left(M\right)_{i,j}\right\}_{i,j=1}^{n,m}$.

- Each realization of the parameters, due to the random coefficients can be associated with an arbitrary number of predictions of the observable M.

Analysis of the Non-Deterministic PC Model

- The Bayesian regression yields a non-deterministic PC representation:

$$M(\xi_1, \xi_2) = g_0 \Psi_0(\xi_1, \xi_2) + \ldots + g_P \Psi_P(\xi_1, \xi_2)$$

  where $\{g_l\}_{l=0}^{P}$ is a random vector defined by a $(P+1)$-dim density.

- The PC regression model, $M(\xi_1, \xi_2)$, depends on:
  1. the parametric uncertainty in the potential through the RVs ξ_1, ξ_2.
  2. thermal noise through the uncertainty in the PC coefficients.

- Interpretation:
  - Draw m samples of the parameters $\left\{\xi_1^{(j)}, \xi_2^{(j)}\right\}_{j=1}^{m}$
  - For any given $\left\{\xi_1^{(j)}, \xi_2^{(j)}\right\}$, we can draw n different sample-spectra of PC coefficients $\left\{\mathbf{g}_i\right\}_{i=1}^{n}$, from their joint distribution $\pi(\mathbf{g})$.
  - We thus obtain $n \times m$ predictions for the target observable $\left\{\left(M\right)_{i,j}\right\}_{i,j=1}^{n,m}$.

- Each realization of the parameters, due to the random coefficients can be associated with an arbitrary number of predictions of the observable M.

Analysis of the Non-Deterministic PC Model

- The Bayesian regression yields a non-deterministic PC representation:

$$M(\xi_1, \xi_2) = g_0 \Psi_0(\xi_1, \xi_2) + \ldots + g_P \Psi_P(\xi_1, \xi_2)$$

  where $\{g_l\}_{l=0}^{P}$ is a random vector defined by a $(P+1)$-dim density.

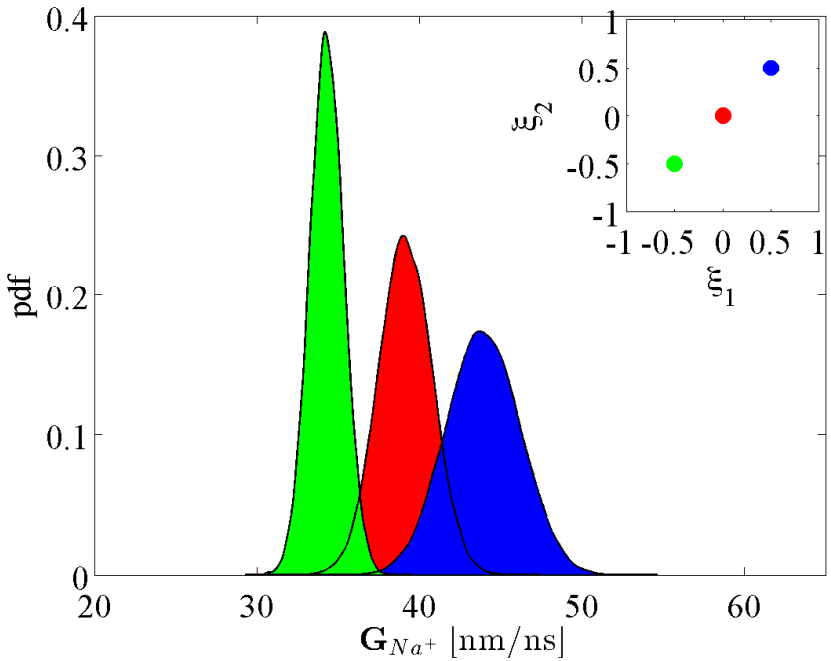
- The PC regression model, $M(\xi_1, \xi_2)$, depends on:
  1. the parametric uncertainty in the potential through the RVs ξ_1, ξ_2.
  2. thermal noise through the uncertainty in the PC coefficients.

- Interpretation:
  - Draw m samples of the parameters $\left\{\xi_1^{(j)}, \xi_2^{(j)}\right\}_{j=1}^{m}$
  - For any given $\left\{\xi_1^{(j)}, \xi_2^{(j)}\right\}$, we can draw n different sample-spectra of PC coefficients $\left\{\mathbf{g}_i\right\}_{i=1}^{n}$, from their joint distribution $\pi(\mathbf{g})$.
  - We thus obtain $n \times m$ predictions for the target observable $\left\{\left(M\right)_{i,j}\right\}_{i,j=1}^{n,m}$.

- Each realization of the parameters, due to the random coefficients can be associated with an arbitrary number of predictions of the observable M.

PDF of Ionic Conductance

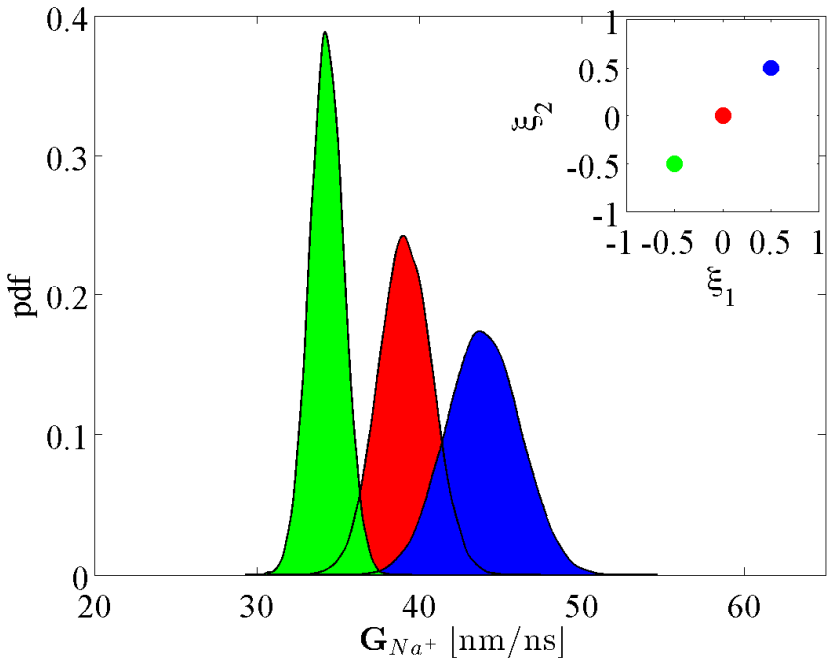
- The PC representation $M(\xi_1, \xi_2)$ is useful to derive statistics.
- Given a value of ξ_1, ξ_2 (ε_{Na+}, ε_{Cl-}), we can sample the PC spectrum and obtain the corresponding uncerainty.
- Estimate the full uncertainty in the observable by sampling both the germ ξ and the PC coefficients.



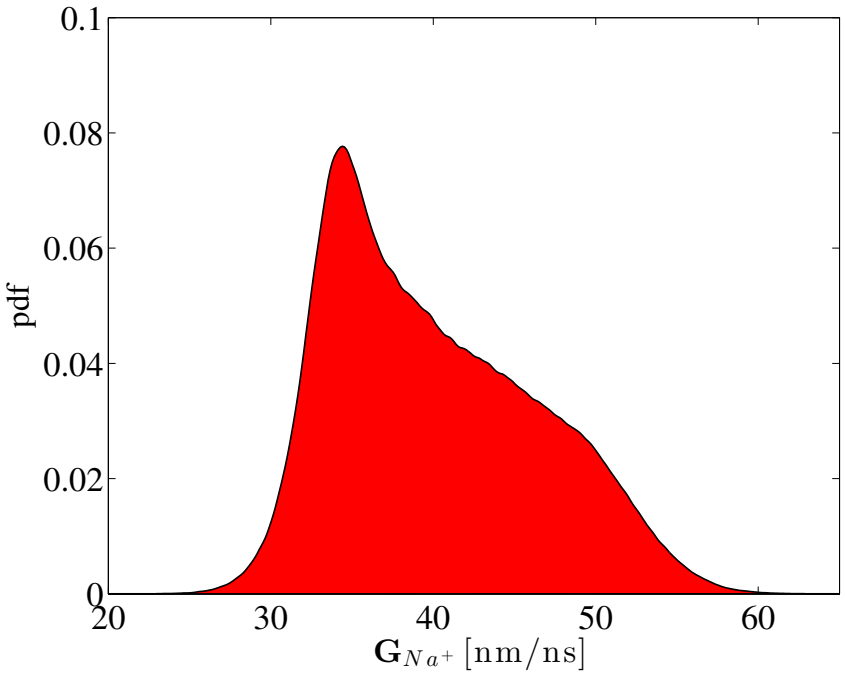
PDF of G_{Na+} for three values of the potential parameters

PDF of Ionic Conductance

- The PC representation $M(\xi_1, \xi_2)$ is useful to derive statistics.
- Given a value of ξ_1, ξ_2 (ε_{Na+}, ε_{Cl-}), we can sample the PC spectrum and obtain the corresponding uncerainty.
- Estimate the full uncertainty in the observable by sampling both the germ $\boldsymbol{\xi}$ and the PC coefficients.



PDF of G_{Na+} for three values of the potential parameters



Full PDF of G_{Na+}

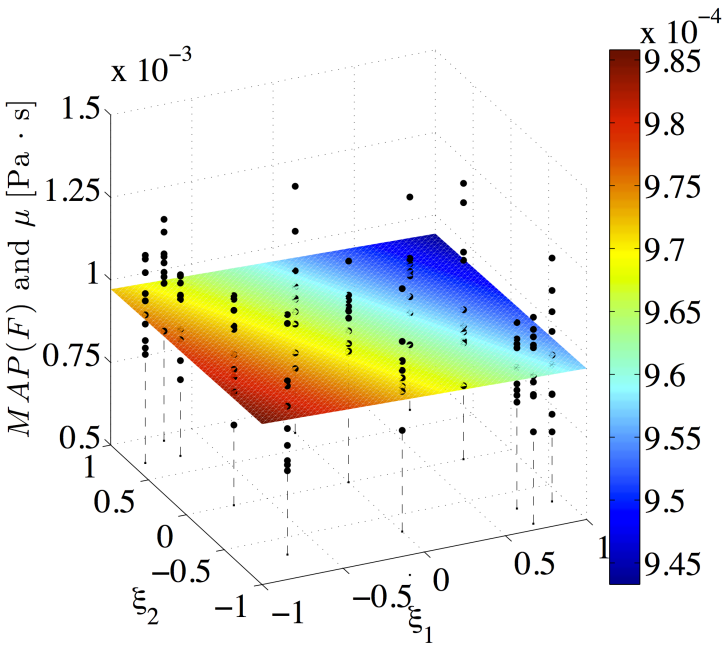
Transport Coefficients

- Separate MD study to compute fluid transport coefficients using Green-Kubo.
- For example, the Green-Kubo formula for dynamics viscosity, μ, is:

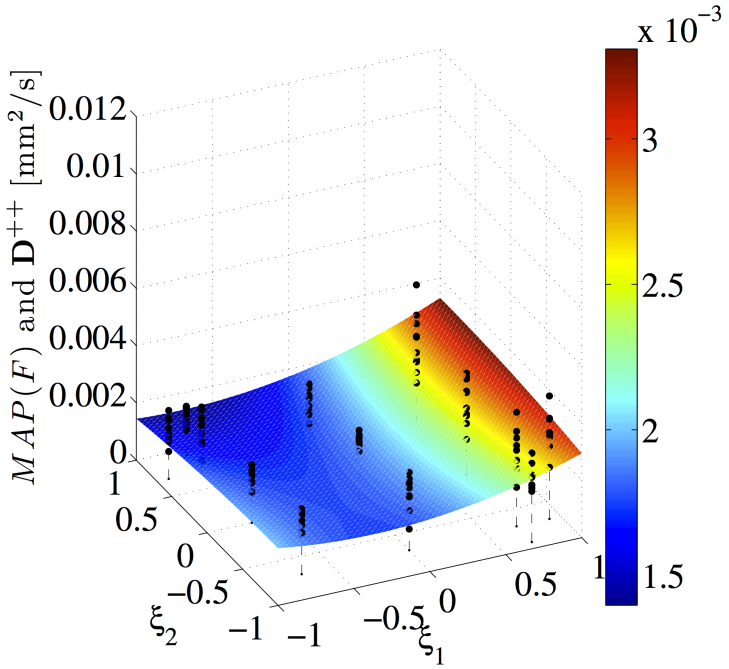
$$\mu = \frac{V}{3k_B T} \int_0^\infty \left\langle \varsigma(0) \cdot \varsigma(t) \right\rangle \mathrm{d}t,$$

where $\varsigma(t)$ is the deviatoric stress, and k_B is the Boltzmann's constant.

- Construct PC expansion, $F(\varepsilon_{Na+}(\xi_1), \varepsilon_{Cl-}(\xi_2))$, for μ and Na^+ diffusivity D^{++}



MAP of PCe response for μ



MAP of PCe response for D^{++}

Correlations between the Ionic Conductance and Transport

- $F(\boldsymbol{\xi}) = \mathbf{f}\,\boldsymbol{\Psi}(\boldsymbol{\xi})$: PCe of one transport coefficient, μ or D^{++}.
- $M(\boldsymbol{\xi}) = \mathbf{g}\,\boldsymbol{\Psi}(\boldsymbol{\xi})$: PCe of the Na^+ conductance, G_{Na^+}.

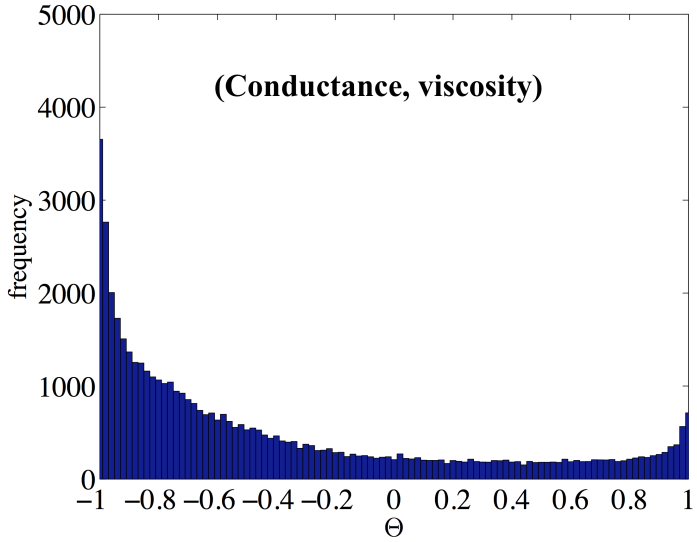
$$\begin{aligned} \mathrm{Cov}(M,F) &= \mathbb{E}\big[(M-\mathbb{E}[M])(F-\mathbb{E}[F])\big] \\ &= \sum_{k=1}^{min(P,P_F)} f_k g_k\, \mathbb{E}\big[\Psi_k^2(\boldsymbol{\xi})\big] \end{aligned}$$

- Sample $\pi(\mathbf{g})$ and $\pi(\mathbf{f}) \Rightarrow \{\mathbf{g}_i\}_{i=1}^{50000}$ and $\{\mathbf{f}_i\}_{i=1}^{50000}$.
- Each $(\mathbf{g}_i, \mathbf{f}_i)$ gives one value of covariance.
- Plot histogram of correlation coefficient Θ.
- (G_{Na^+}, μ) correlation is mainly negative: ionic flux decreases when viscosity increases.
- Strong correlation between G_{Na^+} and D^{++}: we expect the flux of Na^+ to be mostly affected by the diffusivity of Na^+.

Correlations between the Ionic Conductance and Transport

- $F(\boldsymbol{\xi}) = \mathbf{f}\,\Psi(\boldsymbol{\xi})$: PCe of one transport coefficient, μ or D^{++}.
- $M(\boldsymbol{\xi}) = \mathbf{g}\,\Psi(\boldsymbol{\xi})$: PCe of the Na^+ conductance, G_{Na^+}.

$$\begin{aligned} \mathrm{Cov}(M,F) &= \mathbb{E}\big[(M-\mathbb{E}[M])(F-\mathbb{E}[F])\big] \\ &= \sum_{k=1}^{min(P,P_F)} f_k g_k\, \mathbb{E}\big[\Psi_k^2(\boldsymbol{\xi})\big] \end{aligned}$$



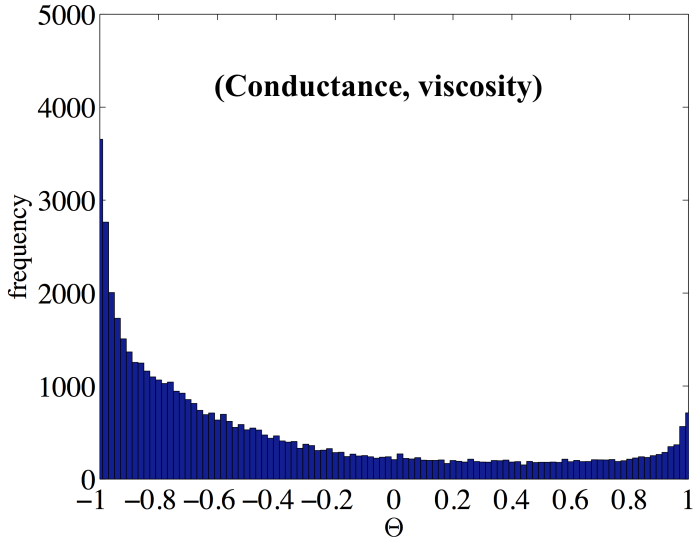
- Sample $\pi(\mathbf{g})$ and $\pi(\mathbf{f}) \Rightarrow \{\mathbf{g}_i\}_{i=1}^{50000}$ and $\{\mathbf{f}_i\}_{i=1}^{50000}$.
- Each $(\mathbf{g}_i, \mathbf{f}_i)$ gives one value of covariance.
- Plot histogram of correlation coefficient Θ.
- (G_{Na^+}, μ) correlation is mainly negative: ionic flux decreases when viscosity increases.
- Strong correlation between G_{Na^+} and D^{++}: we expect the flux of Na^+ to be mostly affected by the diffusivity of Na^+.

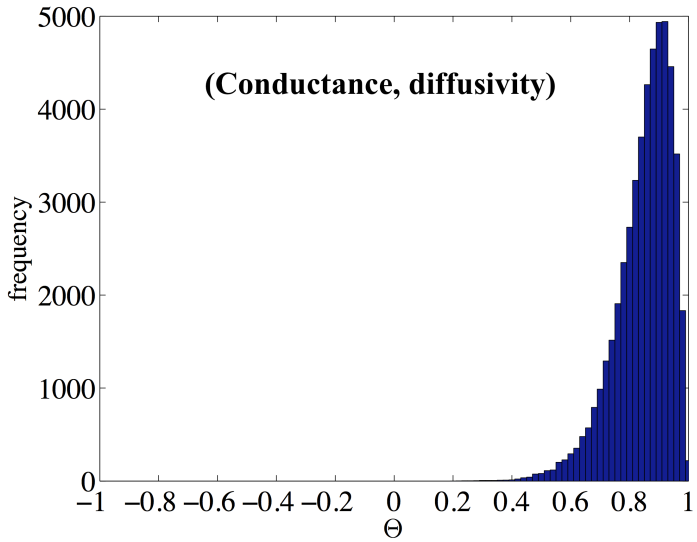
Correlations between the Ionic Conductance and Transport

- $F(\boldsymbol{\xi}) = \mathbf{f}\,\Psi(\boldsymbol{\xi})$: PCe of one transport coefficient, μ or D^{++}.
- $M(\boldsymbol{\xi}) = \mathbf{g}\,\Psi(\boldsymbol{\xi})$: PCe of the Na^+ conductance, G_{Na^+}.

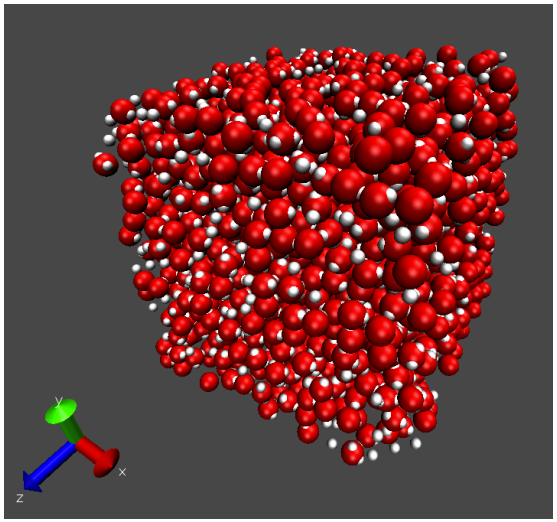
$$\begin{aligned} \mathrm{Cov}(M,F) &= \mathbb{E}\big[(M-\mathbb{E}[M])(F-\mathbb{E}[F])\big] \\ &= \sum_{k=1}^{min(P,P_F)} f_k g_k\, \mathbb{E}\big[\Psi_k^2(\boldsymbol{\xi})\big] \end{aligned}$$

- Sample $\pi(\mathbf{g})$ and $\pi(\mathbf{f}) \Rightarrow \{\mathbf{g}_i\}_{i=1}^{50000}$ and $\{\mathbf{f}_i\}_{i=1}^{50000}$.
- Each $(\mathbf{g}_i, \mathbf{f}_i)$ gives one value of covariance.
- Plot histogram of correlation coefficient Θ.
- (G_{Na^+}, μ) correlation is mainly negative: ionic flux decreases when viscosity increases.
- Strong correlation between G_{Na^+} and D^{++}: we expect the flux of Na^+ to be mostly affected by the diffusivity of Na^+.





Inverse problem for MD of bulk water



- F. Rizzi, H. Najm, B. Debusschere, K. Sargsyan, M. Salloum, H. Adalsteinsson and O. Knio - Part I – *SIAM Multiscale Modeling & Simulation*, 10(4), 1428-1459, 2012.
- F. Rizzi, H. Najm, B. Debusschere, K. Sargsyan, M. Salloum, H. Adalsteinsson and O. Knio - Part II – *SIAM Multiscale Modeling & Simulation*, 10(4), 1460-1492, 2012.
- F. Rizzi, Ph.D. thesis, The Johns Hopkins University, Baltimore, MD, 2012.

Inverse Problem

- Consider a generic forward model: $\mathbf{G} = \phi(\mathbf{H})$.
- The associated inverse problem becomes:

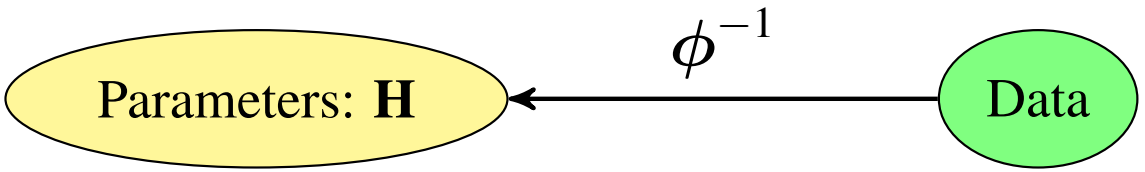
Given data, what can we say about $\mathbf{H}$?

Which $\mathbf{H}$ yields the best match between $\mathbf{G}$ and data?

- If formulated as an optimization, it yields single value for $\mathbf{H}$.
- Bayesian approach yields a *joint* probability density function (PDF) on $\mathbf{H}$.
  - Joint PDFs contain correlations.
  - Ideal for risk assesssment.

Inverse Problem

- Consider a generic forward model: $\mathbf{G} = \phi(\mathbf{H})$.
- The associated inverse problem becomes:



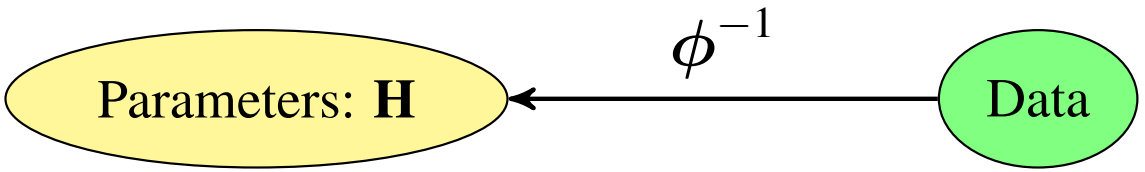
Given data, what can we say about **H**?

Which **H** yields the best match between **G** and data?

- If formulated as an optimization, it yields single value for **H**.
- Bayesian approach yields a *joint* probability density function (PDF) on **H**.
  - Joint PDFs contain correlations.
  - Ideal for risk assesssment.

Inverse Problem

- Consider a generic forward model: $\mathbf{G} = \phi(\mathbf{H})$.
- The associated inverse problem becomes:



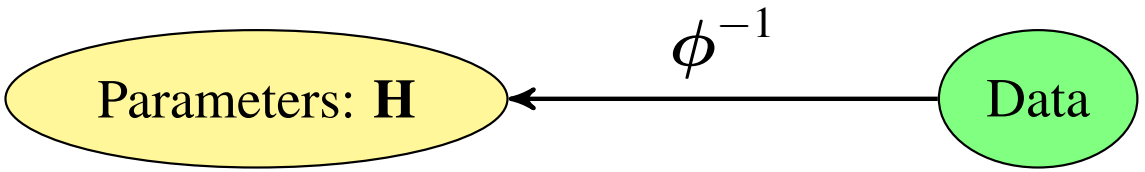
  Given data, what can we say about **H**?

  Which **H** yields the best match between **G** and data?

- If formulated as an optimization, it yields single value for **H**.
- Bayesian approach yields a *joint* probability density function (PDF) on **H**.
  - Joint PDFs contain correlations.
  - Ideal for risk assesssment.

Inverse Problem

- Consider a generic forward model: $\mathbf{G} = \phi(\mathbf{H})$.
- The associated inverse problem becomes:



  Given data, what can we say about **H**?

  Which **H** yields the best match between **G** and data?

- If formulated as an optimization, it yields single value for **H**.
- Bayesian approach yields a *joint* probability density function (PDF) on **H**.
  - Joint PDFs contain correlations.
  - Ideal for risk assesssment.

Example

- Consider $(x, y) = \phi(x, y, t; a, b)$:

$$\begin{aligned} \dot{x}(t) &= a^2 - b^2 \\ \dot{y}(t) &= ab + 0.01\sin(x) \end{aligned}$$

- a, b are **model parameters**:

$$(a, b) \sim \mathcal{N}([2\ 1], Cov)$$

- Two cases:

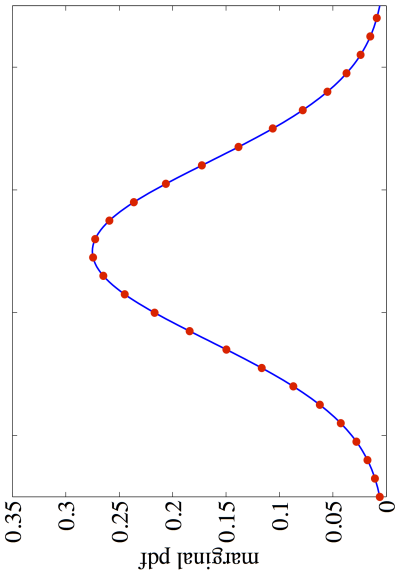
  Uncorrelated parameters:

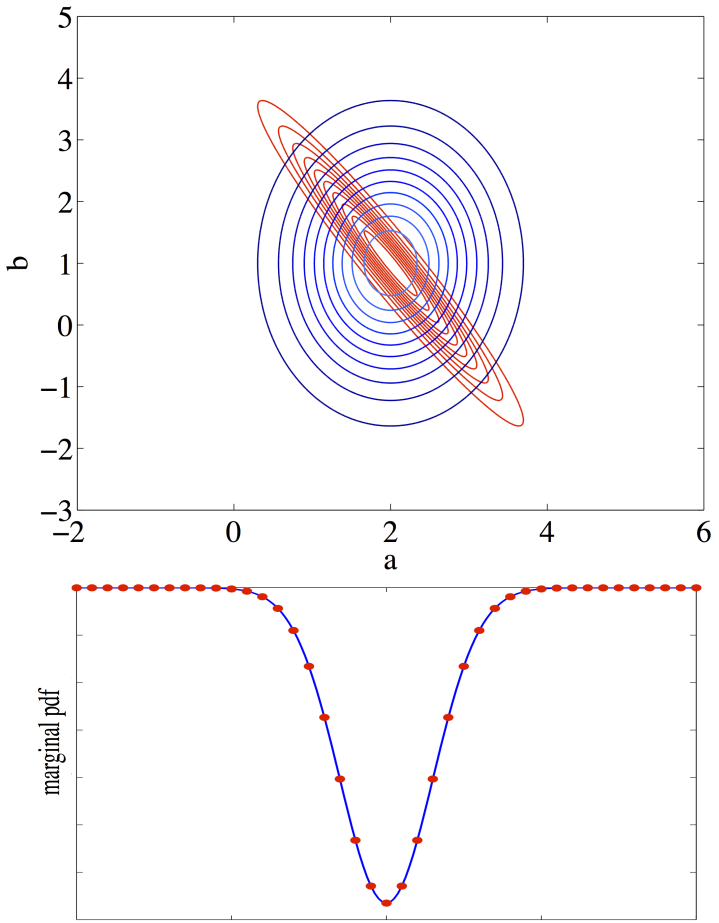
$$Cov = \begin{bmatrix} 0.6 & 0.0 \\ 0.0 & 1.45 \end{bmatrix}$$

  Correlated parameters:

$$Cov = \begin{bmatrix} 0.6 & -0.9 \\ -0.9 & 1.45 \end{bmatrix}$$

- Same marginal densities.
- What is the impact of the correlation?





Example

- Consider $(x, y) = \phi(x, y, t; a, b)$:

$$\begin{aligned} \dot{x}(t) &= a^2 - b^2 \\ \dot{y}(t) &= ab + 0.01\sin(x) \end{aligned}$$

- a, b are **model parameters**:

$$(a, b) \sim \mathcal{N}([2\ 1], Cov)$$

- Two cases:

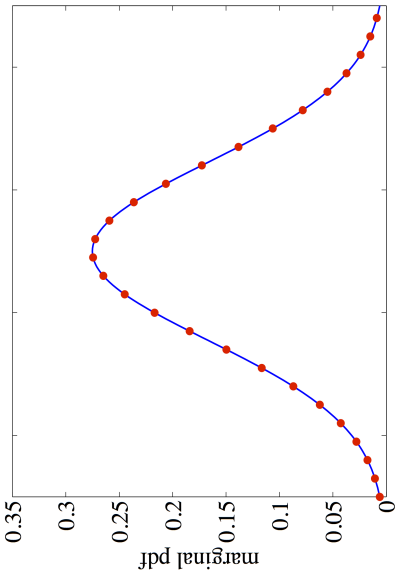
  Uncorrelated parameters:

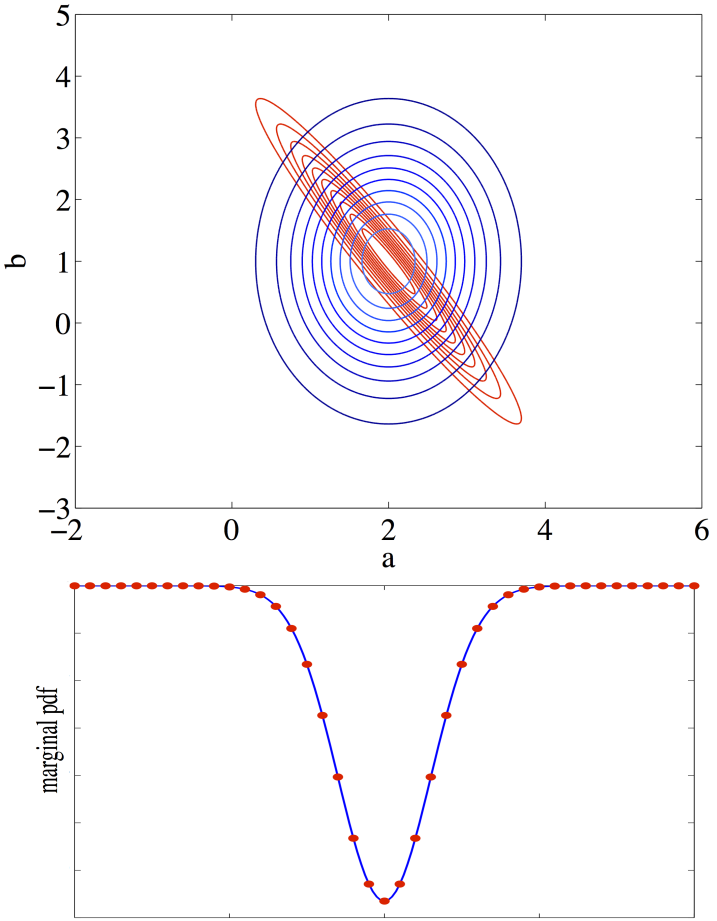
$$Cov = \begin{bmatrix} 0.6 & 0.0 \\ 0.0 & 1.45 \end{bmatrix}$$

  Correlated parameters:

$$Cov = \begin{bmatrix} 0.6 & -0.9 \\ -0.9 & 1.45 \end{bmatrix}$$

- Same marginal densities.
- What is the impact of the correlation?





Example

- Consider $(x, y) = \phi(x, y, t; a, b)$:

$$\begin{aligned} \dot{x}(t) &= a^2 - b^2 \\ \dot{y}(t) &= ab + 0.01\sin(x) \end{aligned}$$

- a, b are **model parameters**:

$$(a, b) \sim \mathcal{N}([2\ 1], Cov)$$

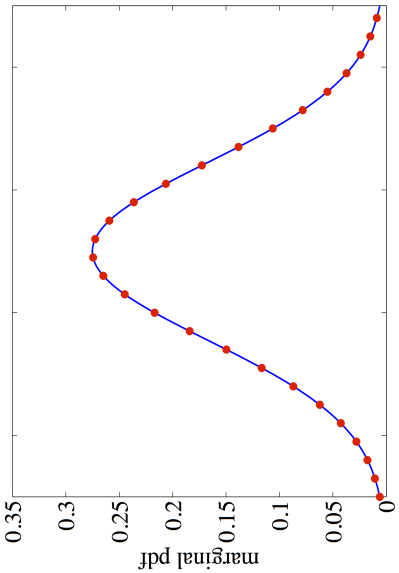
- Two cases:

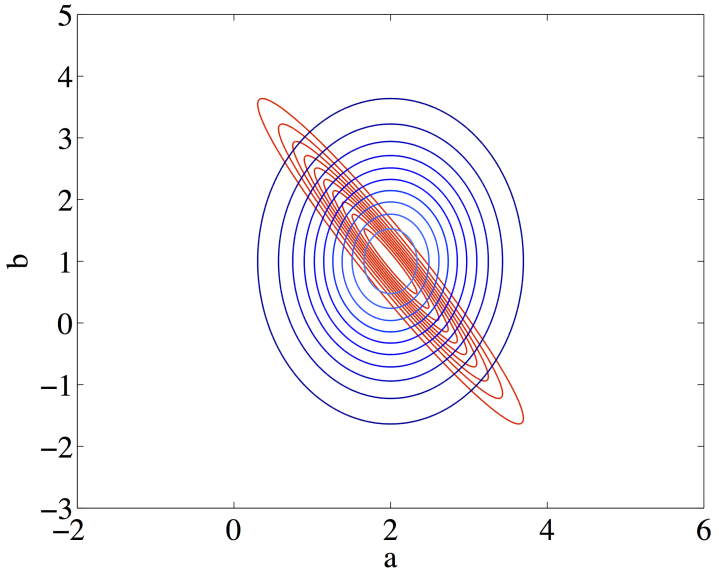
  Uncorrelated parameters:

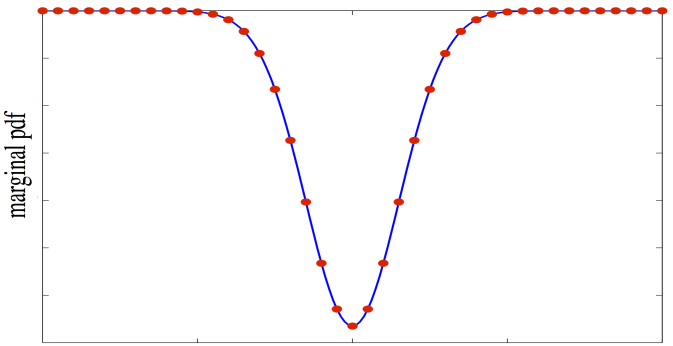
$$Cov = \begin{bmatrix} 0.6 & 0.0 \\ 0.0 & 1.45 \end{bmatrix}$$

  Correlated parameters:

$$Cov = \begin{bmatrix} 0.6 & -0.9 \\ -0.9 & 1.45 \end{bmatrix}$$



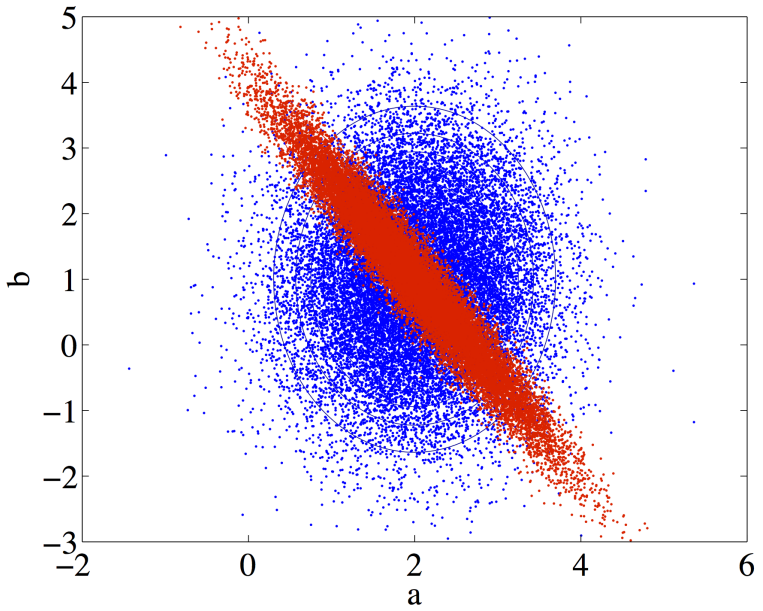




- Same marginal densities.
- What is the impact of the correlation?

Example: results

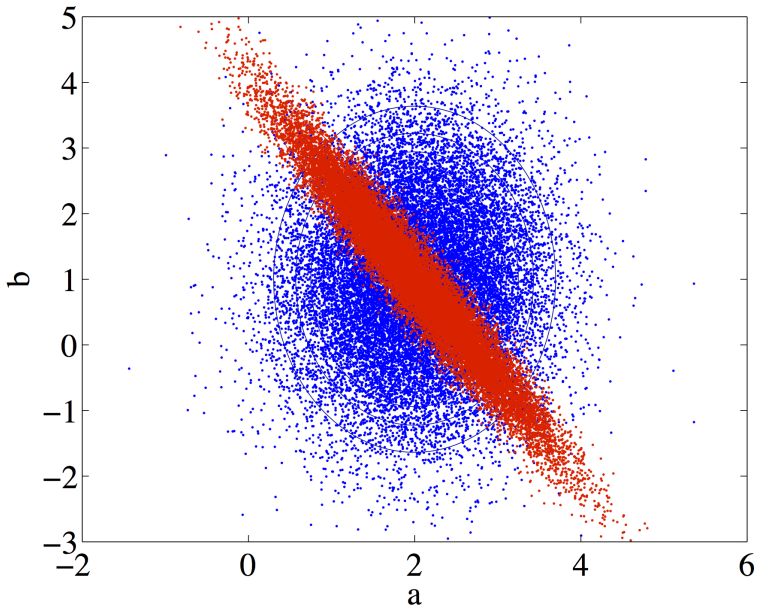
- Sample the joint PDFs: $\{(a_i, b_i)^{U,C}\}_{i=1}^{n}$
- Compute trajectories from $(x_0 = 1, y_0 = 0.5)$.
- Two sets of predictions: $\{(x_j, y_j)^{U,C}\big|_T\}_{j=1}^{n}$
- Estimate the joint PDFs.



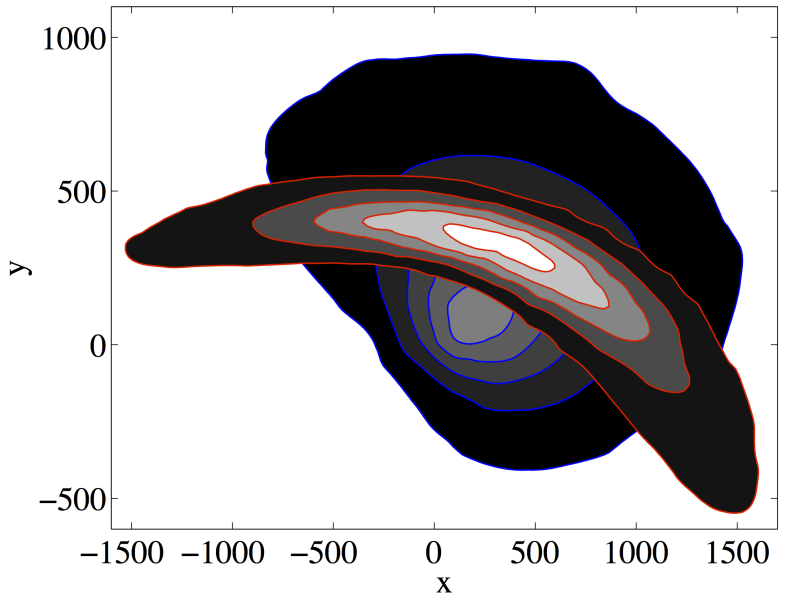
- Model predictions are substantially different.
- Correlation has large impact.
- Especially important for more complicated and non-linear systems.

Example: results

- Sample the joint PDFs: $\{(a_i, b_i)^{U,C}\}_{i=1}^{n}$
- Compute trajectories from $(x_0 = 1, y_0 = 0.5)$.
- Two sets of predictions: $\{(x_j, y_j)^{U,C}\big|_T\}_{j=1}^{n}$
- Estimate the joint PDFs.



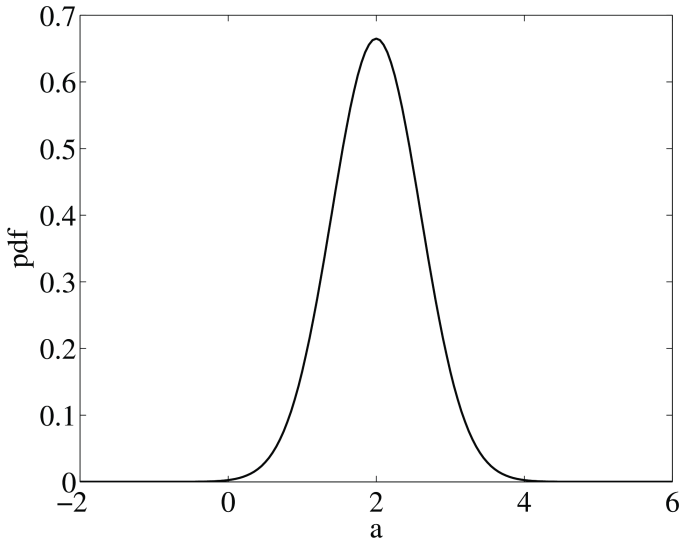
- Model predictions are substantially different.
- Correlation has large impact.
- Especially important for more complicated and non-linear systems.

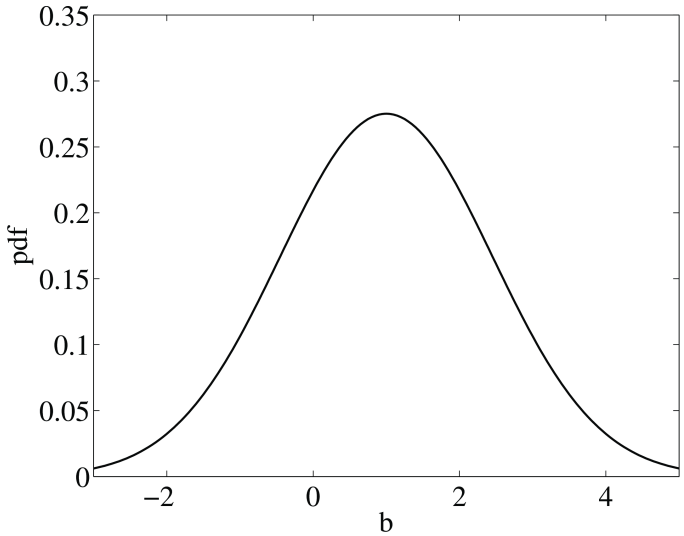


Example: results

- **Joint** distribution of the model inputs is a key information.

- In practice, nominal values with associated confidence intervals and no details on the joint density (or correlation).
- The most **common approach** is to presume **independence** of model parameters, and to use a convenient distribution for each based on the known nominal values and bounds.
- Obtaining **joint distribution**? Probabilistic parameter estimation.
- Bayesian inference yields the joint distribution of the model parameters.





Problem Statement

- MD simulations of bulk water at $T = 298$ K, $P = 1$ atm; run with LAMMPS.

- Suppose person A selects three potential parameters $\{\alpha_1, \alpha_2, \alpha_3\}$, and runs the forward UQ to infer PCEs for some water observables: density, viscosity, etc.

$$M^{(\rho)}(\alpha_1, \alpha_2, \alpha_3) = c_0 \Psi_0(\alpha_1, \alpha_2, \alpha_3) + \ldots + c_P \Psi_P(\alpha_1, \alpha_2, \alpha_3)$$

with $\{c_\ell\}_{\ell=0}^{P}$ described by a $(P+1)$-dim joint density.

- Person B secretly chooses three values of the parameters: $\hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3$.

- Runs MD replicas and collects density observations $\boldsymbol{\rho} = \{\rho_i\}_{i=1}^{N=10}$.

- Given: PCe as a surrogate model and $\{\rho_i\}_{i=1}^{N=10}$.

- Challenge: to recover the "true" parameters chosen by person B.

Problem Statement

- MD simulations of bulk water at $T = 298$ K, $P = 1$ atm; run with LAMMPS.

Suppose person A selects three potential parameters $\{\alpha_1, \alpha_2, \alpha_3\}$, and runs the forward UQ to infer PCEs for some water observables: density, viscosity, etc.

$$M^{(\rho)}(\alpha_1, \alpha_2, \alpha_3) = c_0 \Psi_0(\alpha_1, \alpha_2, \alpha_3) + \ldots + c_P \Psi_P(\alpha_1, \alpha_2, \alpha_3)$$

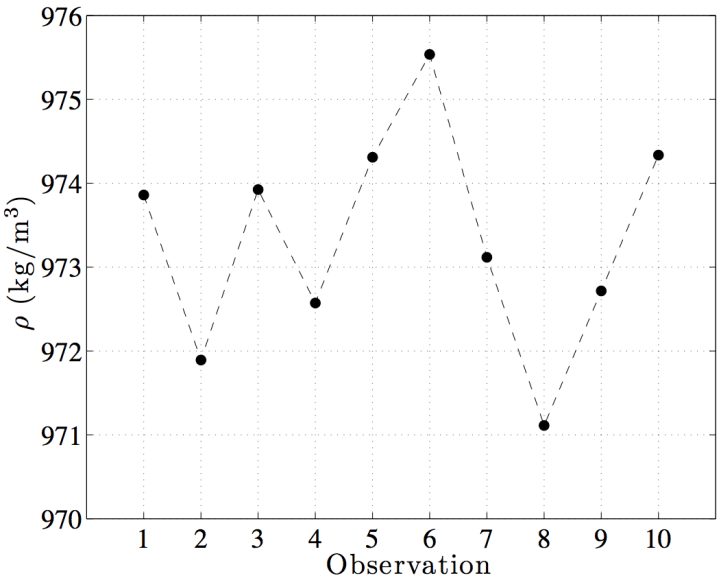
with $\{c_\ell\}_{\ell=0}^{P}$ described by a $(P+1)$-dim joint density.

Person B secretly chooses three values of the parameters: $\hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3$.

Runs MD replicas and collects density observations $\boldsymbol{\rho} = \{\rho_i\}_{i=1}^{N=10}$.

Given: PCe as a surrogate model and $\{\rho_i\}_{i=1}^{N=10}$.

Challenge: to recover the "true" parameters chosen by person B.



Data: $\{\rho_i\}_{i=1}^{10}$

Problem Statement

- MD simulations of bulk water at $T = 298$ K, $P = 1$ atm; run with LAMMPS.

- Suppose person A selects three potential parameters $\{\alpha_1, \alpha_2, \alpha_3\}$, and runs the forward UQ to infer PCEs for some water observables: density, viscosity, etc.

$$M^{(\rho)}(\alpha_1, \alpha_2, \alpha_3) = c_0 \Psi_0(\alpha_1, \alpha_2, \alpha_3) + \ldots + c_P \Psi_P(\alpha_1, \alpha_2, \alpha_3)$$

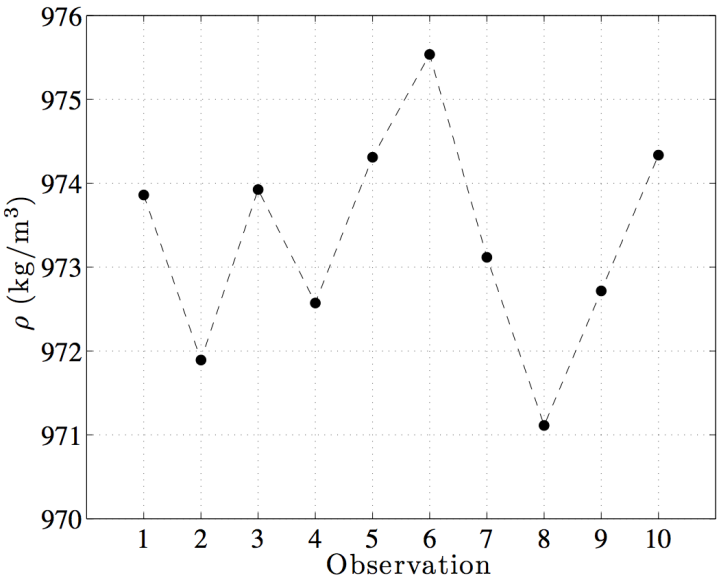
with $\{c_\ell\}_{\ell=0}^{P}$ described by a $(P+1)$-dim joint density.

- Person B secretly chooses three values of the parameters: $\hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3$.

- Runs MD replicas and collects density observations $\rho = \{\rho_i\}_{i=1}^{N=10}$.

- Given: PCe as a surrogate model and $\{\rho_i\}_{i=1}^{N=10}$.

- Challenge: to recover the "true" parameters chosen by person B.



Data: $\{\rho_i\}_{i=1}^{10}$

Bayesian Inference Framework

- Bayesian theory suites inverse problems involving uncertainties and noisy data.
- Bayesian inference uses a set of observations $\boldsymbol{\rho} = \{\rho_i\}_{i=1}^{N=10}$ (evidence) to calculate the probability that the hypothesis $\mathcal{H} = \{\alpha_1, \alpha_2, \alpha_3\}$ is true.

Bayes' theorem

$$\underbrace{\pi\left(\alpha_1, \alpha_2, \alpha_3 \mid \boldsymbol{\rho}\right)}_{\textit{Posterior}} \propto \underbrace{\mathcal{L}\left(\boldsymbol{\rho} \mid \alpha_1, \alpha_2, \alpha_3\right)}_{\textit{Likelihood}} \underbrace{\mathcal{P}(\alpha_1, \alpha_2, \alpha_3)}_{\textit{Prior}}$$

- ◇ PRIOR: knowledge/information about $\mathcal{H}$ before considering the data.
- ◇ LIKELIHOOD: probability of "seeing" the data given a realization of $\mathcal{H}$.
- ◇ POSTERIOR: probability of the hypothesis given the data.

- An "update" of the current state of knowledge in view of new observations.
- Likelihood formulation?

Inverse Problem Formulation

- Additive Gaussian error model accounting for the deviation between each observation, ρ_i, and the PCe surrogate prediction $M^{(\rho)}(\alpha_1, \alpha_2, \alpha_3)$

$$\rho_i = M^{(\rho)}(\alpha_1, \alpha_2, \alpha_3) + \gamma_i, \qquad i = 1, \ldots, N,$$

  where $\{\gamma_i\}_{i=1}^N$ are *i.i.d.* Gaussian RVs with density $p_\gamma = \mathcal{N}(0, \tilde{\sigma}^2)$.

- Recall that the surrogate PC model is *uncertain*

$$M^{(\rho)}(\alpha_1, \alpha_2, \alpha_3) = c_0 \Psi_0(\alpha_1, \alpha_2, \alpha_3) + \ldots + c_P \Psi_P(\alpha_1, \alpha_2, \alpha_3)$$

  because $\{c_\ell\}_{\ell=0}^P$ are described by a $(P+1)$-dim joint density.

- The PC coefficients thus have an associated variance.

- How do we account for the *uncertainty* in the PC model?

Inverse Problem Formulation

- Additive Gaussian error model accounting for the deviation between each observation, ρ_i, and the PCe surrogate prediction $M^{(\rho)}(\alpha_1, \alpha_2, \alpha_3)$

$$\rho_i = M^{(\rho)}(\alpha_1, \alpha_2, \alpha_3) + \gamma_i, \qquad i = 1, \ldots, N,$$

  where $\{\gamma_i\}_{i=1}^N$ are *i.i.d.* Gaussian RVs with density $p_\gamma = \mathcal{N}(0, \tilde{\sigma}^2)$.

- Recall that the surrogate PC model is *uncertain*

$$M^{(\rho)}(\alpha_1, \alpha_2, \alpha_3) = c_0 \Psi_0(\alpha_1, \alpha_2, \alpha_3) + \ldots + c_P \Psi_P(\alpha_1, \alpha_2, \alpha_3)$$

  because $\{c_\ell\}_{\ell=0}^P$ are described by a $(P+1)$-dim joint density.

- The PC coefficients thus have an associated variance.

- How do we account for the *uncertainty* in the PC model?

Inverse Problem Formulation

- For a given sample $\boldsymbol{\alpha}^{(j)} = \{\alpha_1^{(j)}, \alpha_2^{(j)}, \alpha_3^{(j)}\}$, construct the constant vector

$$\boldsymbol{y} = \{\Psi_0(\boldsymbol{\alpha}^{(j)}), \ldots, \Psi_P(\boldsymbol{\alpha}^{(j)})\}^T$$

- This implies that the **non-deterministic** PC model

$$\begin{aligned} M(\boldsymbol{\alpha}^{(j)}) &= c_0\Psi_0(\boldsymbol{\alpha}^{(j)}) + c_1\Psi_1(\boldsymbol{\alpha}^{(j)}) + \ldots + c_P\Psi_P(\boldsymbol{\alpha}^{(j)}) \\ &= \boldsymbol{y}^T\boldsymbol{c} \end{aligned}$$

represents a **linear combination** of the random vector $\boldsymbol{c} = \{c_0, \ldots, c_P\}^T$.

- If the random vector $\mathbf{c} \sim \mathcal{MVN}(\boldsymbol{\mu}, \mathbf{Z})$ (verified) then, by definition, we have

$$\boldsymbol{y}^T\boldsymbol{c} \sim \mathcal{N}(\boldsymbol{y}^T\boldsymbol{\mu},\ \boldsymbol{y}^T\boldsymbol{Z}\boldsymbol{y})$$

i.e. a *univariate* Gaussian with mean $(\boldsymbol{y}^T\boldsymbol{\mu})$ and variance $(\boldsymbol{y}^T\boldsymbol{Z}\boldsymbol{y})$.

Inverse Problem Formulation

- For a given sample $\boldsymbol{\alpha}^{(j)} = \{\alpha_1^{(j)}, \alpha_2^{(j)}, \alpha_3^{(j)}\}$, construct the constant vector

$$\boldsymbol{y} = \{\Psi_0(\boldsymbol{\alpha}^{(j)}), \ldots, \Psi_P(\boldsymbol{\alpha}^{(j)})\}^T$$

- This implies that the **non-deterministic** PC model

$$\begin{aligned} M(\boldsymbol{\alpha}^{(j)}) &= c_0\Psi_0(\boldsymbol{\alpha}^{(j)}) + c_1\Psi_1(\boldsymbol{\alpha}^{(j)}) + \ldots + c_P\Psi_P(\boldsymbol{\alpha}^{(j)}) \\ &= \boldsymbol{y}^T\boldsymbol{c} \end{aligned}$$

  represents a **linear combination** of the random vector $\boldsymbol{c} = \{c_0, \ldots, c_P\}^T$.

- If the random vector $\mathbf{c} \sim \mathcal{MVN}(\boldsymbol{\mu}, \mathbf{Z})$ (verified) then, by definition, we have

$$\boldsymbol{y}^T\boldsymbol{c} \sim \mathcal{N}(\boldsymbol{y}^T\boldsymbol{\mu}, \boldsymbol{y}^T\boldsymbol{Z}\boldsymbol{y})$$

  i.e. a *univariate* Gaussian with mean $(\boldsymbol{y}^T\boldsymbol{\mu})$ and variance $(\boldsymbol{y}^T\boldsymbol{Z}\boldsymbol{y})$.

Inverse Problem Formulation

- For a given sample $\boldsymbol{\alpha}^{(j)} = \{\alpha_1^{(j)}, \alpha_2^{(j)}, \alpha_3^{(j)}\}$, construct the constant vector

$$\boldsymbol{y} = \{\Psi_0(\boldsymbol{\alpha}^{(j)}), \ldots, \Psi_P(\boldsymbol{\alpha}^{(j)})\}^T$$

- This implies that the **non-deterministic** PC model

$$\begin{aligned} M(\boldsymbol{\alpha}^{(j)}) &= c_0\Psi_0(\boldsymbol{\alpha}^{(j)}) + c_1\Psi_1(\boldsymbol{\alpha}^{(j)}) + \ldots + c_P\Psi_P(\boldsymbol{\alpha}^{(j)}) \\ &= \boldsymbol{y}^T\boldsymbol{c} \end{aligned}$$

  represents a **linear combination** of the random vector $\boldsymbol{c} = \{c_0, \ldots, c_P\}^T$.

- If the random vector $\mathbf{c} \sim \mathcal{MVN}(\boldsymbol{\mu}, \mathbf{Z})$ (verified) then, by definition, we have

$$\boldsymbol{y}^T\boldsymbol{c} \sim \mathcal{N}(\boldsymbol{y}^T\boldsymbol{\mu},\ \boldsymbol{y}^T\boldsymbol{Z}\boldsymbol{y})$$

  i.e. a *univariate* Gaussian with mean $(\boldsymbol{y}^T\boldsymbol{\mu})$ and variance $(\boldsymbol{y}^T\boldsymbol{Z}\boldsymbol{y})$.

Inverse Problem Formulation

- Consequently, the error model becomes

$$\underbrace{\rho_i}_{\text{data}} = \underbrace{\boldsymbol{y}^T\boldsymbol{c}}_{\text{model prediction}} + \underbrace{\gamma_i}_{\text{additive noise}}, \qquad i = 1, \ldots, N,$$

$$\gamma_i \sim \mathcal{N}(0, \tilde{\sigma}^2), \text{ for } i = 1, \ldots, N, \quad \text{and} \quad \boldsymbol{y}^T\boldsymbol{c} \sim \mathcal{N}\left(\boldsymbol{y}^T\boldsymbol{\mu}, \boldsymbol{y}^T\boldsymbol{Z}\boldsymbol{y}\right).$$

- Leading to the following likelihood:

$$\mathcal{L}\left(\{\rho_i\}_{i=1}^N \middle| \alpha_1, \alpha_2, \alpha_3\right) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\left(\boldsymbol{y}^T\boldsymbol{Z}\boldsymbol{y} + \tilde{\sigma}^2\right)}} \exp\left(-\frac{\left[\rho_i - \boldsymbol{y}^T\boldsymbol{\mu}\right]^2}{2\left(\boldsymbol{y}^T\boldsymbol{Z}\boldsymbol{y} + \tilde{\sigma}^2\right)}\right)$$

- Combines both surrogate uncertainty and data noise in a self-consistent manner.
- For each data, $\mathcal{L}$ is maximum if the data and surrogate mean coincide. Deviations are weighted by the variances of the noise *and* uncertain surrogate.
- Regions of high data-noise or large surrogate-uncertainty are both penalized with lower weighting on discrepancies between the data and the mean-surrogate model.

Inverse Problem Formulation

- Consequently, the error model becomes

$$\underbrace{\rho_i}_{\text{data}} = \underbrace{\boldsymbol{y}^T\boldsymbol{c}}_{\text{model prediction}} + \underbrace{\gamma_i}_{\text{additive noise}}, \qquad i = 1, \ldots, N,$$

$$\gamma_i \sim \mathcal{N}(0, \tilde{\sigma}^2), \text{ for } i = 1, \ldots, N, \quad \text{and} \quad \boldsymbol{y}^T\boldsymbol{c} \sim \mathcal{N}\left(\boldsymbol{y}^T\boldsymbol{\mu}, \boldsymbol{y}^T\boldsymbol{Z}\boldsymbol{y}\right).$$

- Leading to the following likelihood:

$$\mathcal{L}\left(\{\rho_i\}_{i=1}^N \middle| \alpha_1, \alpha_2, \alpha_3\right) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\left(\boldsymbol{y}^T\boldsymbol{Z}\boldsymbol{y} + \tilde{\sigma}^2\right)}} \exp\left(-\frac{\left[\rho_i - \boldsymbol{y}^T\boldsymbol{\mu}\right)\right]^2}{2\left(\boldsymbol{y}^T\boldsymbol{Z}\boldsymbol{y} + \tilde{\sigma}^2\right)}\right)$$

- Combines both surrogate uncertainty and data noise in a self-consistent manner.
- For each data, $\mathcal{L}$ is maximum if the data and surrogate mean coincide. Deviations are weighted by the variances of the noise *and* uncertain surrogate.
- Regions of high data-noise or large surrogate-uncertainty are both penalized with lower weighting on discrepancies between the data and the mean-surrogate model.

Inverse Problem Formulation

- Consequently, the error model becomes

$$\underbrace{\rho_i}_{\text{data}} = \underbrace{\boldsymbol{y}^T\boldsymbol{c}}_{\text{model prediction}} + \underbrace{\gamma_i}_{\text{additive noise}}, \qquad i = 1, \ldots, N,$$

$$\gamma_i \sim \mathcal{N}(0, \tilde{\sigma}^2), \text{ for } i = 1, \ldots, N, \quad \text{and} \quad \boldsymbol{y}^T\boldsymbol{c} \sim \mathcal{N}\left(\boldsymbol{y}^T\boldsymbol{\mu}, \boldsymbol{y}^T\boldsymbol{Z}\boldsymbol{y}\right).$$

- Leading to the following likelihood:

$$\mathcal{L}\left(\{\rho_i\}_{i=1}^N \middle| \alpha_1, \alpha_2, \alpha_3\right) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\left(\boldsymbol{y}^T\boldsymbol{Z}\boldsymbol{y} + \tilde{\sigma}^2\right)}} \exp\left(-\frac{\left[\rho_i - \boldsymbol{y}^T\boldsymbol{\mu}\right)\right]^2}{2\left(\boldsymbol{y}^T\boldsymbol{Z}\boldsymbol{y} + \tilde{\sigma}^2\right)}\right)$$

- Combines both surrogate uncertainty and data noise in a self-consistent manner.
- For each data, $\mathcal{L}$ is maximum if the data and surrogate mean coincide. Deviations are weighted by the variances of the noise *and* uncertain surrogate.
- Regions of high data-noise or large surrogate-uncertainty are both penalized with lower weighting on discrepancies between the data and the mean-surrogate model.

Inverse Problem Formulation

- Consequently, the error model becomes

$$\underbrace{\rho_i}_{\text{data}} = \underbrace{\boldsymbol{y}^T\boldsymbol{c}}_{\text{model prediction}} + \underbrace{\gamma_i}_{\text{additive noise}}, \qquad i = 1, \ldots, N,$$

$$\gamma_i \sim \mathcal{N}(0, \tilde{\sigma}^2), \text{ for } i = 1, \ldots, N, \quad \text{and} \quad \boldsymbol{y}^T\boldsymbol{c} \sim \mathcal{N}\left(\boldsymbol{y}^T\boldsymbol{\mu}, \boldsymbol{y}^T\boldsymbol{Z}\boldsymbol{y}\right).$$

- Leading to the following likelihood:

$$\mathcal{L}\left(\{\rho_i\}_{i=1}^N \middle| \alpha_1, \alpha_2, \alpha_3\right) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\left(\boldsymbol{y}^T\boldsymbol{Z}\boldsymbol{y} + \tilde{\sigma}^2\right)}} \exp\left(-\frac{\left[\rho_i - \boldsymbol{y}^T\boldsymbol{\mu}\right)\right]^2}{2\left(\boldsymbol{y}^T\boldsymbol{Z}\boldsymbol{y} + \tilde{\sigma}^2\right)}\right)$$

- Combines both surrogate uncertainty and data noise in a self-consistent manner.
- For each data, $\mathcal{L}$ is maximum if the data and surrogate mean coincide. Deviations are weighted by the variances of the noise *and* uncertain surrogate.
- Regions of high data-noise or large surrogate-uncertainty are both penalized with lower weighting on discrepancies between the data and the mean-surrogate model.

Inverse Problem: Two Possible Likelihoods

Uncertain PC model

$$\mathcal{L}\left(\{\rho_i\}_{i=1}^N \middle| \alpha_1, \alpha_2, \alpha_3\right) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\left(\boldsymbol{y}^T \boldsymbol{Z} \boldsymbol{y} + \tilde{\sigma}^2\right)}} \exp\left(-\frac{\left[\rho_i - \boldsymbol{y}^T \boldsymbol{\mu}\right)\right]^2}{2\left(\boldsymbol{y}^T \boldsymbol{Z} \boldsymbol{y} + \tilde{\sigma}^2\right)}\right)$$

Deterministic PC model

- What if the uncertainty in the PC surrogate coefficients is zero or negligible?
- The random vector $\boldsymbol{c} = \{c_0, \ldots, c_P\}^T$ has covariance zero $\mathbf{Z} \approx \mathbf{0}$
  $\Rightarrow$ the PC surrogate model is now *deterministic*

$$\mathcal{L}\left(\{\rho_i\}_{i=1}^N \middle| \alpha_1, \alpha_2, \alpha_3\right) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\left(\cancelto{0}{\boldsymbol{y}^T \boldsymbol{Z} \boldsymbol{y}} + \tilde{\sigma}^2\right)}} \exp\left(-\frac{\left[\rho_i - \boldsymbol{y}^T \boldsymbol{\mu}\right)\right]^2}{2\left(\cancelto{0}{\boldsymbol{y}^T \boldsymbol{Z} \boldsymbol{y}} + \tilde{\sigma}^2\right)}\right)$$

- Surrogate uncertainty drops out.
- The likelihood now involves only the data noise.

Inverse Problem: Two Possible Likelihoods

Uncertain PC model

$$\mathcal{L}\left(\{\rho_i\}_{i=1}^N \middle| \alpha_1, \alpha_2, \alpha_3\right) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\left(\boldsymbol{y}^T \boldsymbol{Z} \boldsymbol{y} + \tilde{\sigma}^2\right)}} \exp\left(-\frac{\left[\rho_i - \boldsymbol{y}^T \boldsymbol{\mu}\right)\right]^2}{2\left(\boldsymbol{y}^T \boldsymbol{Z} \boldsymbol{y} + \tilde{\sigma}^2\right)}\right)$$

Deterministic PC model

- What if the uncertainty in the PC surrogate coefficients is zero or negligible?
- The random vector $\boldsymbol{c} = \{c_0, \ldots, c_P\}^T$ has covariance zero $\mathbf{Z} \approx \mathbf{0}$
  $\Rightarrow$ the PC surrogate model is now *deterministic*

$$\mathcal{L}\left(\{\rho_i\}_{i=1}^N \middle| \alpha_1, \alpha_2, \alpha_3\right) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\left(\cancelto{0}{\boldsymbol{y}^T \boldsymbol{Z} \boldsymbol{y}} + \tilde{\sigma}^2\right)}} \exp\left(-\frac{\left[\rho_i - \boldsymbol{y}^T \boldsymbol{\mu}\right)\right]^2}{2\left(\cancelto{0}{\boldsymbol{y}^T \boldsymbol{Z} \boldsymbol{y}} + \tilde{\sigma}^2\right)}\right)$$

- Surrogate uncertainty drops out.
- The likelihood now involves only the data noise.

Inverse Problem: Posterior Sampling

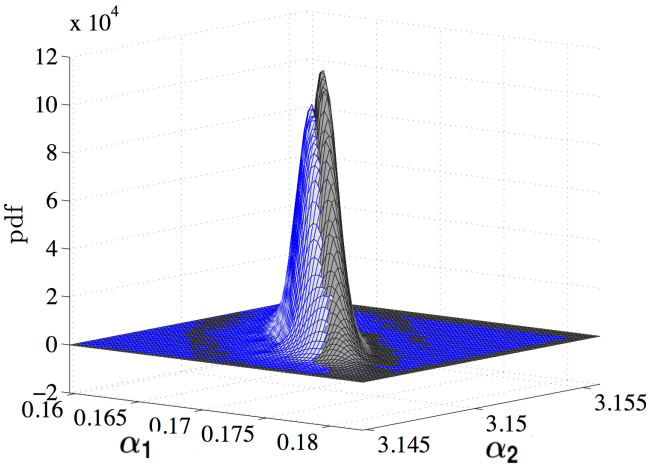
- Previous formulation was limited to density data $\{\rho_i\}_{i=1}^N$.
- In the real study the formulation is based on two additional sets of data, namely water enthalpy $\{h_i\}_{i=1}^N$ and self-diffusion $\{D_i\}_{i=1}^N$.
- The data set is thus: $data = \{\rho_i, h_i, D_i\}_{i=1}^N$.
- From Bayes' theorem, the joint posterior distribution is given by

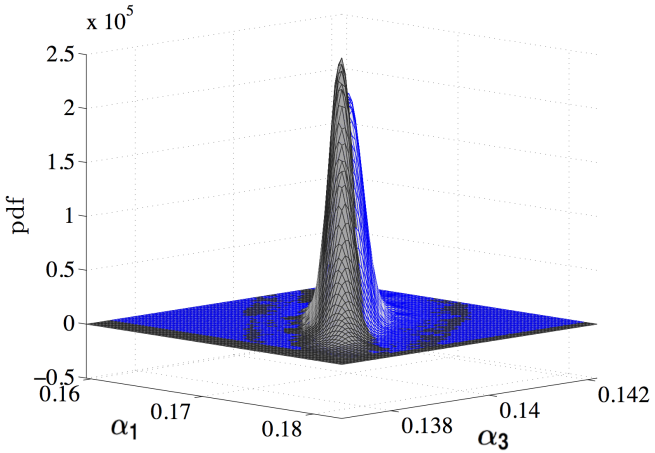
$$\pi\left(\{\alpha_1,\alpha_2,\alpha_3\}, hyperp \middle| \, data\right) \propto \mathcal{L}\left(data \middle| \{\alpha_1,\alpha_2,\alpha_3\}, hyperp\right) Priors$$

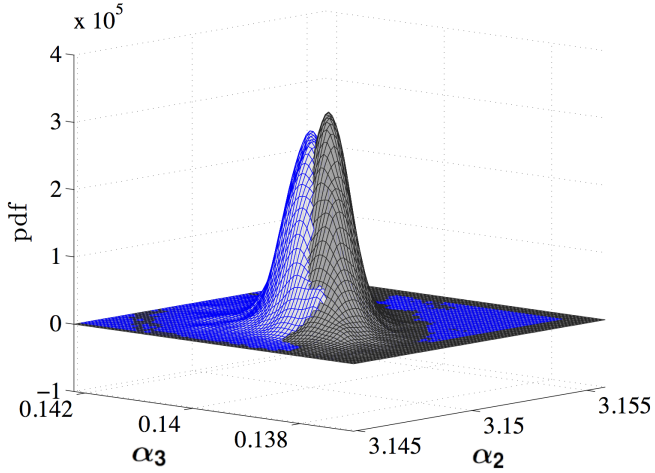
- Sample the posterior using MCMC based on adaptive Metropolis.
- MCMC samples are used to construct posterior statistics.

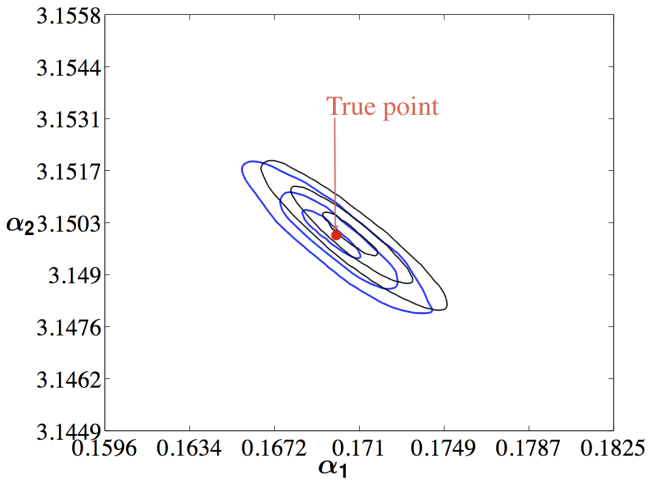
Inverse Problem: Results

- Joint posteriors based on **Deterministic** and **Non-Deterministic** surrogates.
- Substantial correlations are captured by the inference.

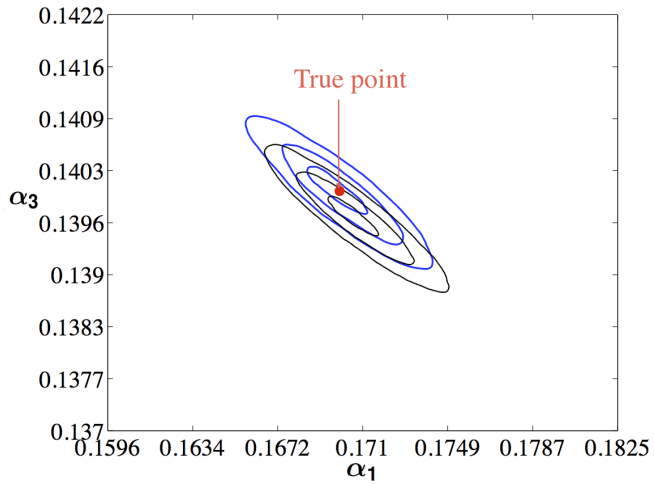




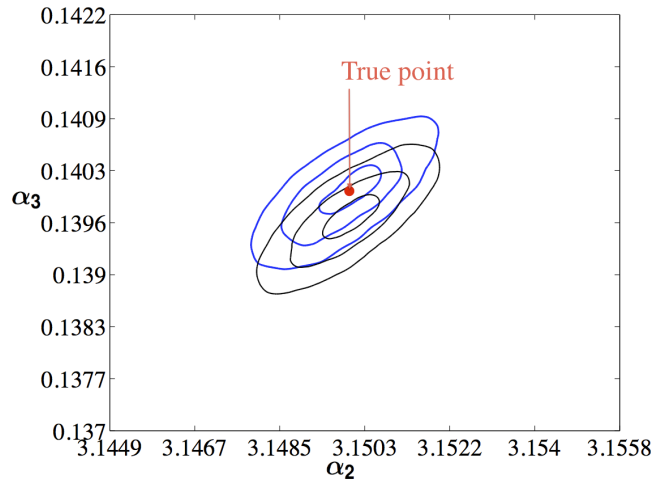




$\pi(\alpha_1, \alpha_2)$

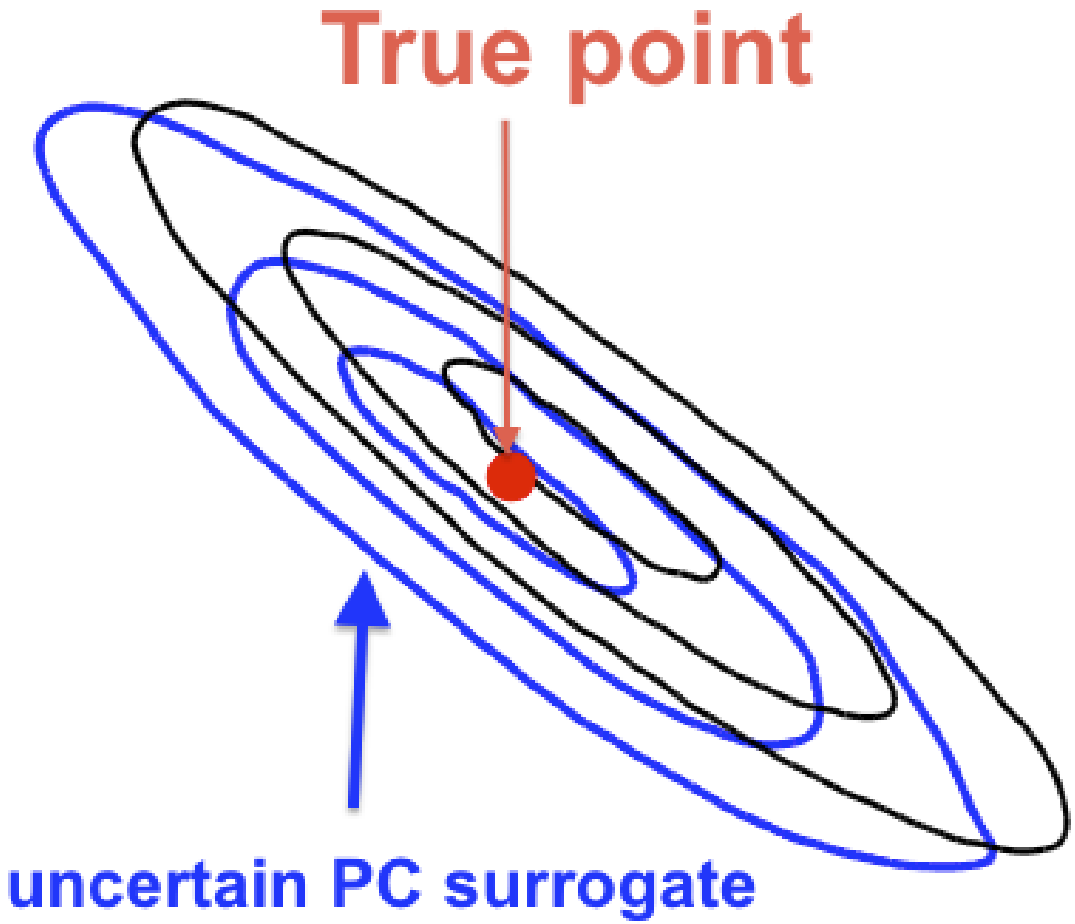


$\pi(\alpha_1, \alpha_3)$



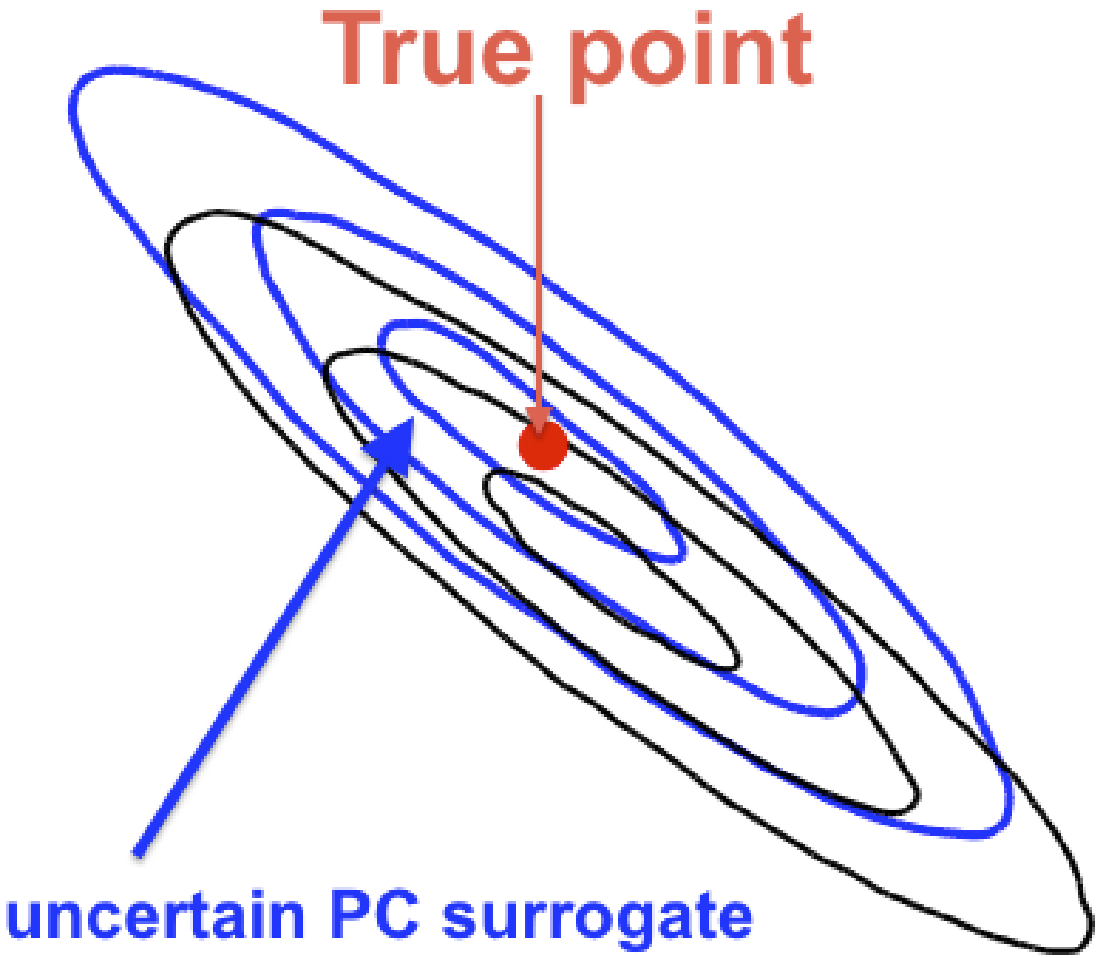
$\pi(\alpha_2, \alpha_3)$

Detailed View of Posterior Performance



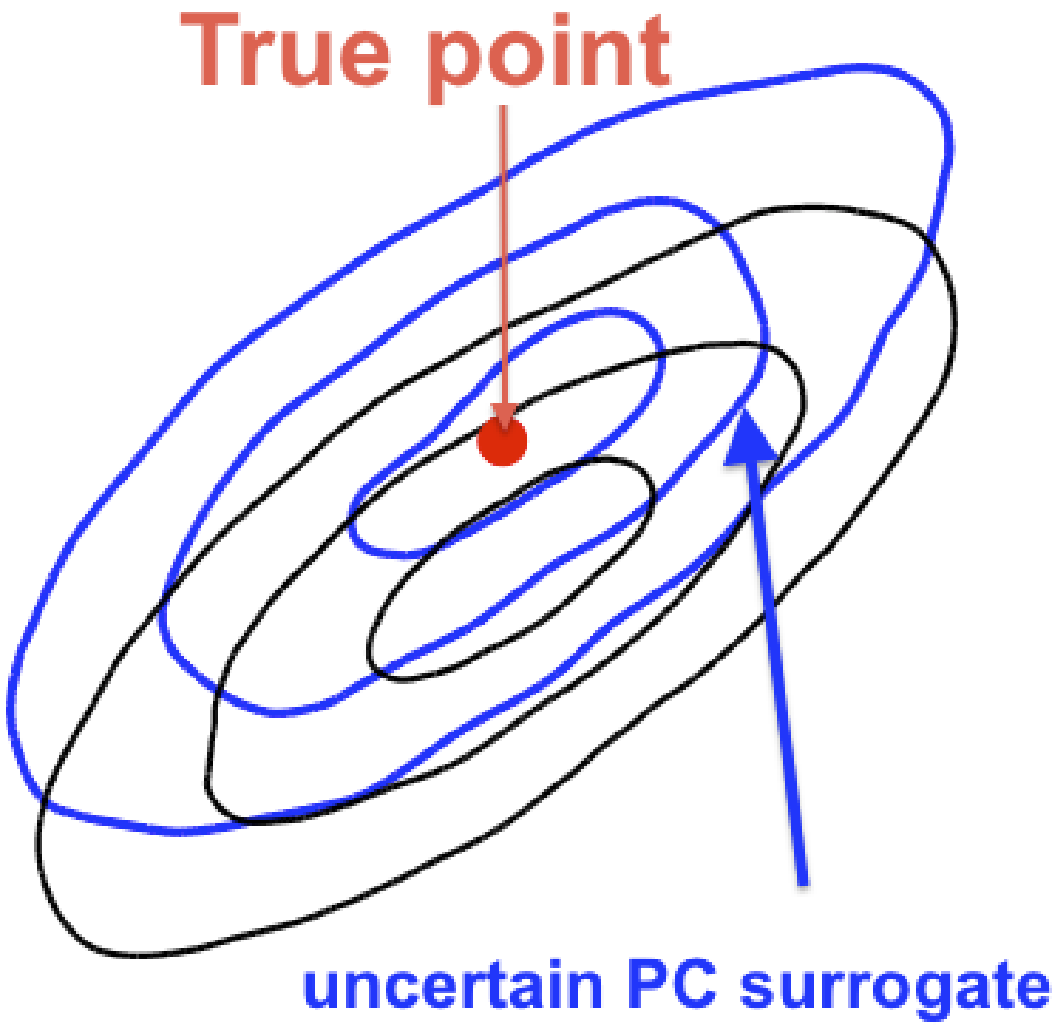
$$\pi(\alpha_1, \alpha_2)$$

Detailed View of Posterior Performance



$$\pi(\alpha_1, \alpha_3)$$

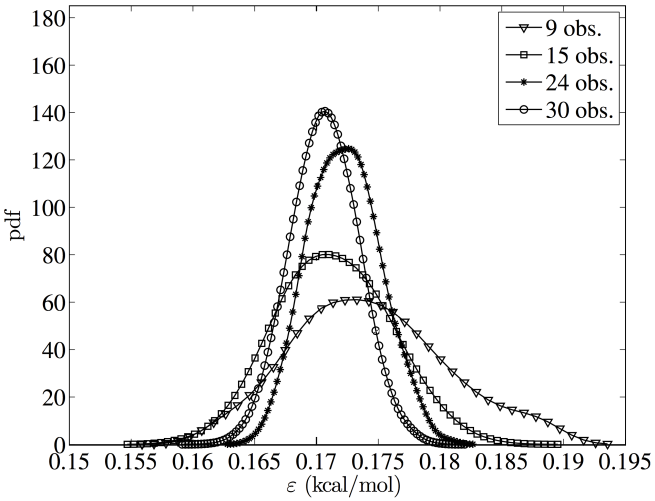
Detailed View of Posterior Performance



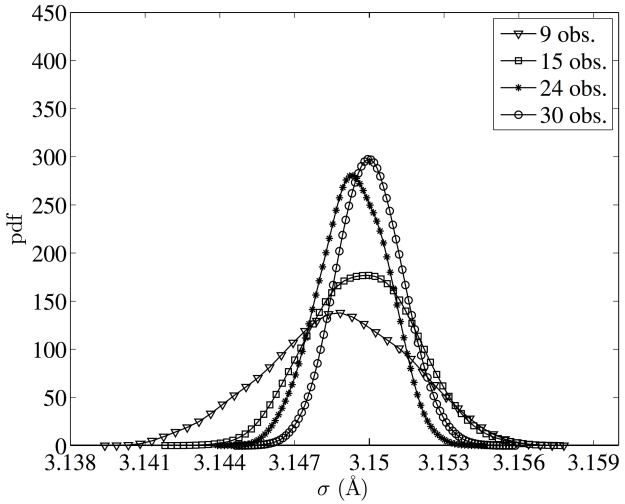
$$\pi(\alpha_2, \alpha_3)$$

Three observables: density, self-diffusion and enthalpy

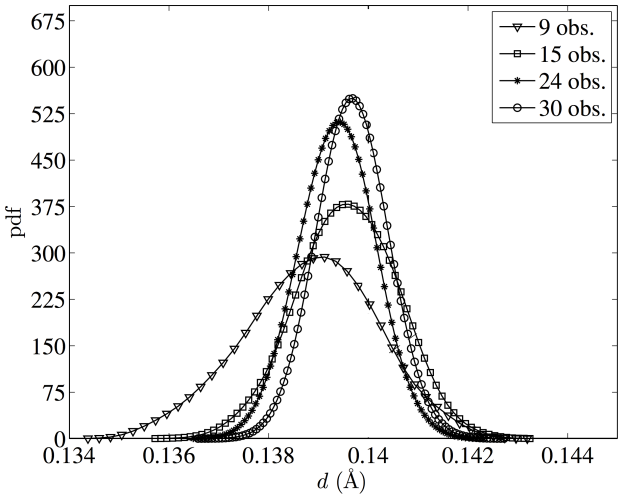
- Explore impact of number of data points.
- N data points for each observable, so $3N$ = total number of data points.
- More information available (larger N), less variance (uncertainty) in the posterior.
- Peak of PDF however varies slightly.



$\pi(\varepsilon)$



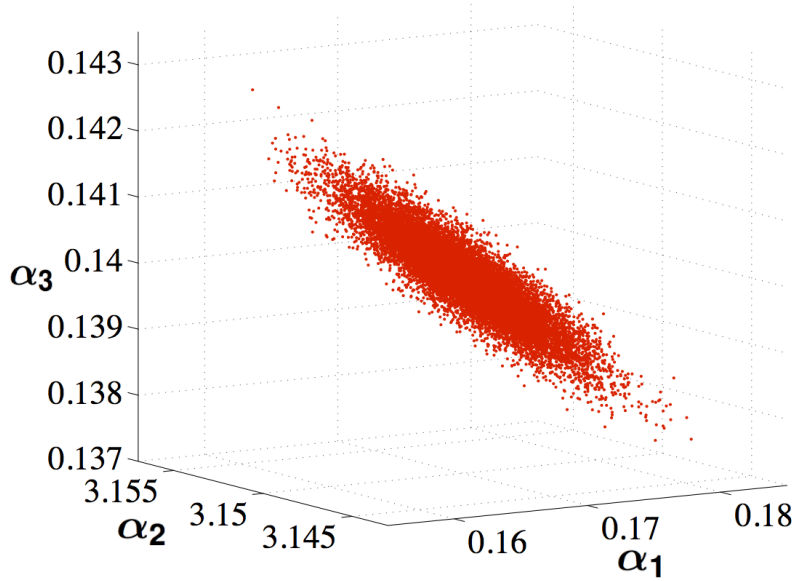
$\pi(\sigma)$

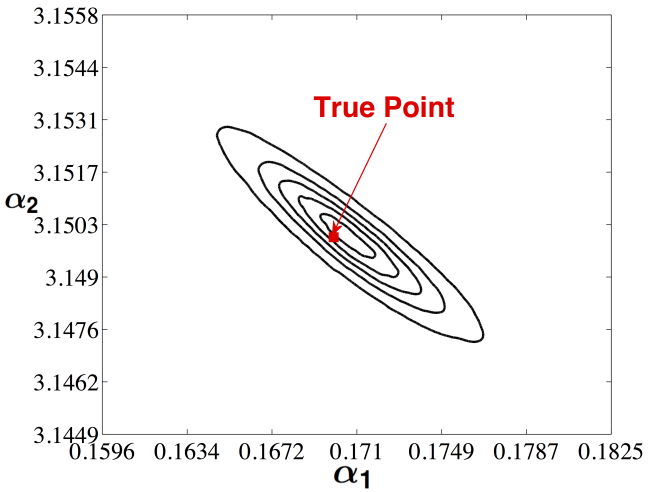


$\pi(d)$

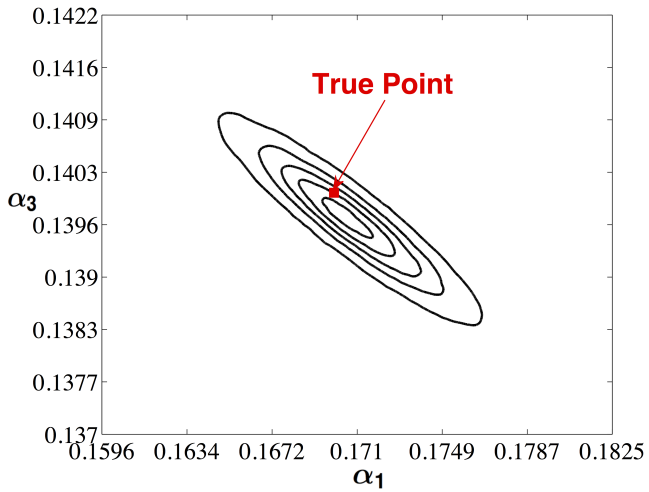
Posterior Correlations

- 3D-joint posterior based on the MCMC samples.
- New information: substantial correlation. (Parameters initially assumed independent.)
- Correlation stems from the "physics"/data and manifests during the inference.

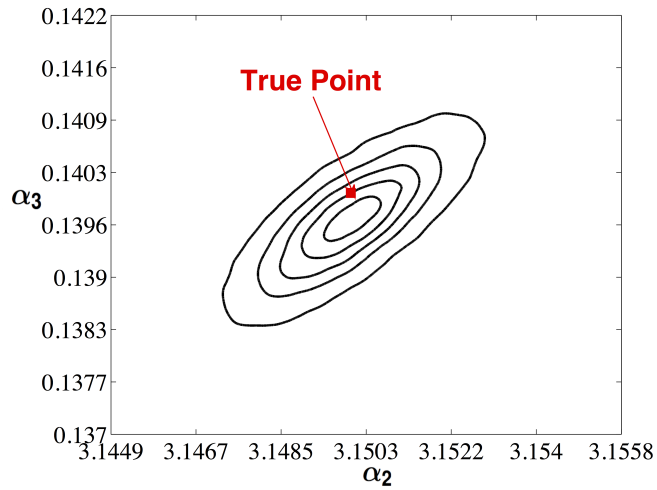




$\pi(\varepsilon, \sigma)$



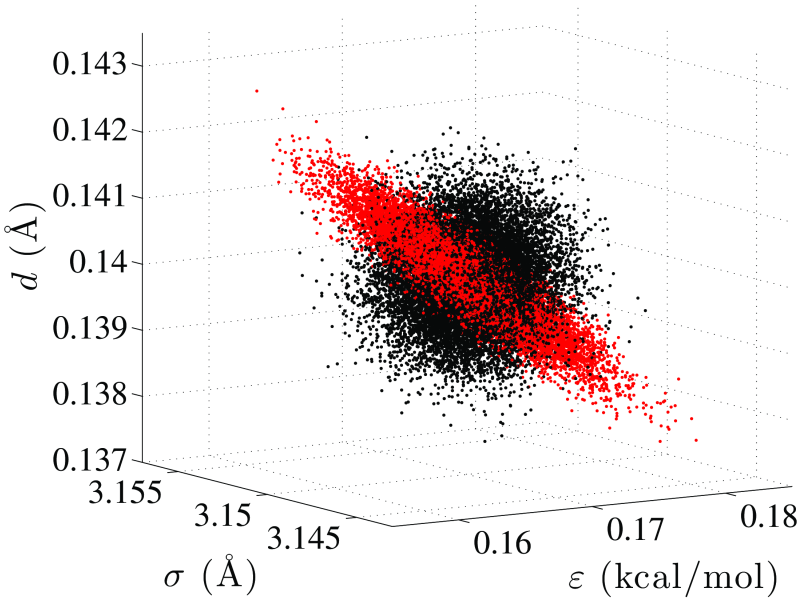
$\pi(\varepsilon, d)$

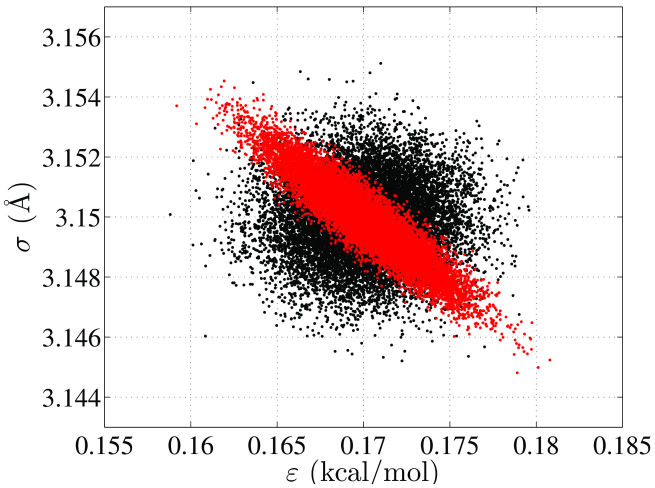


$\pi(\sigma, d)$

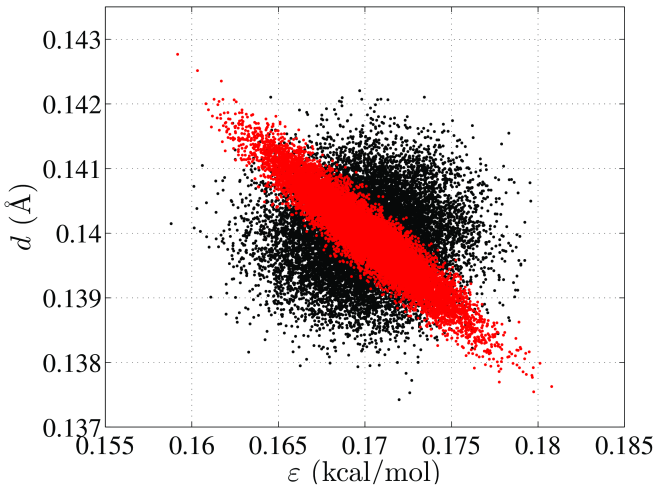
What is the role of correlation?

- Is the correlation important?
- Build PDF with same mean but zero off-diagonal elements = non-correlated.
- Same marginal densities.
- Goal: to analyze the effect of the **correlated** samples vs. **non-correlated** samples.

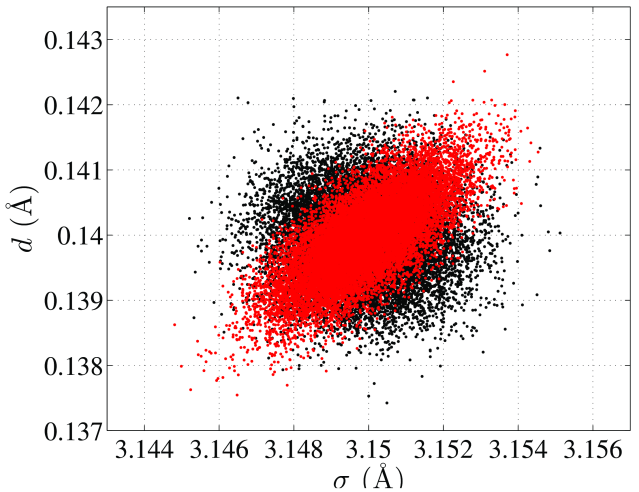




(ε, σ)



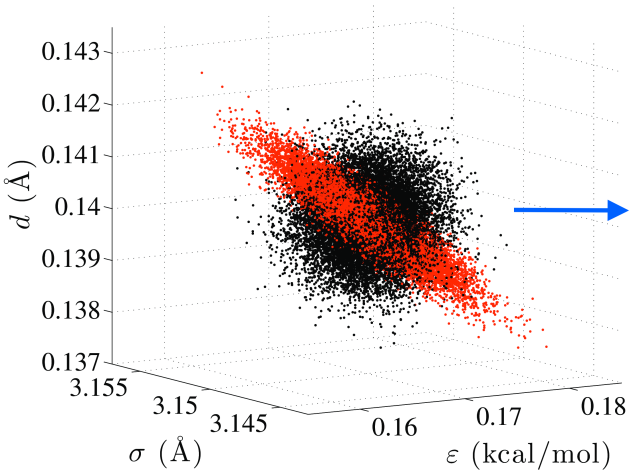
(ε, d)

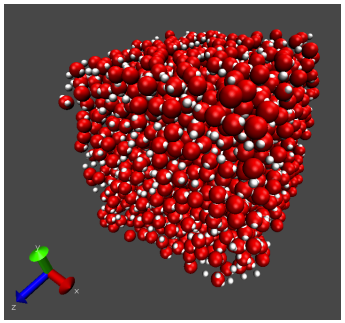


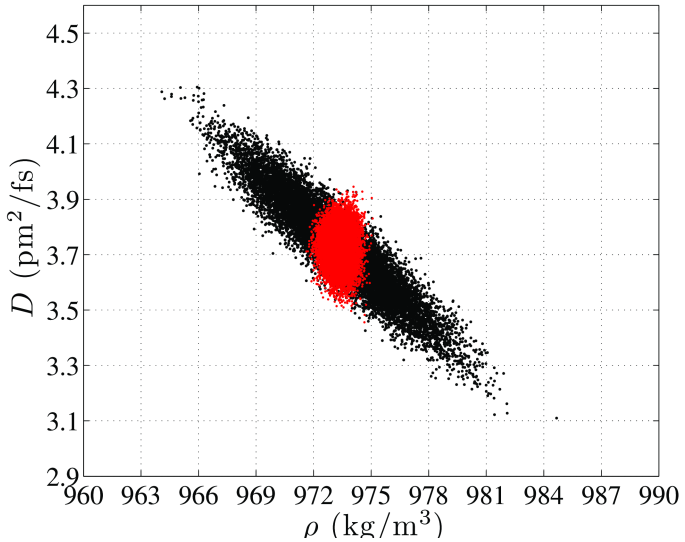
(σ, d)

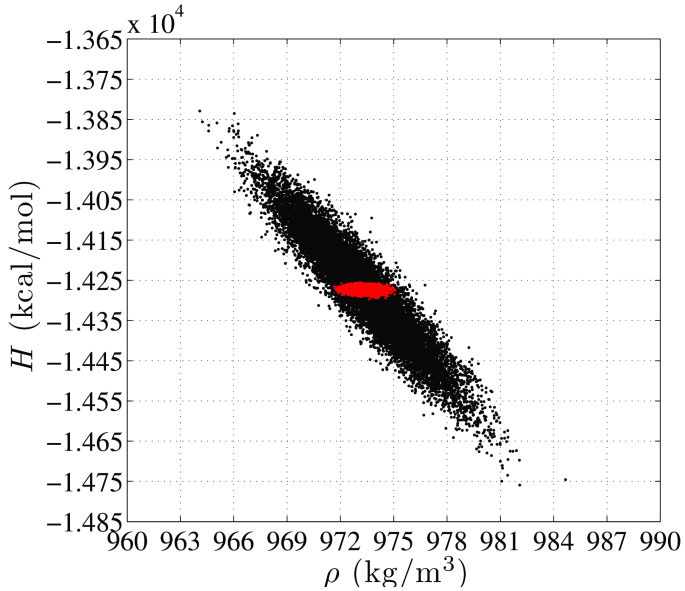
What is the role of correlation?

- Push forward **correlated** and **uncorrelated** samples to compute predictions.
- Plot the predictions:
  1. Data used for the inference.
  2. Predictions from correlated PDF.
  3. Predictions from uncorrelated PDF.









Summary & Conclusions

✓ UQ successfully applied to MD simulations.

- Two distinct sources of uncertainty:
  1. parametric uncertainty in the potential
  2. intrinsic (thermal) noise

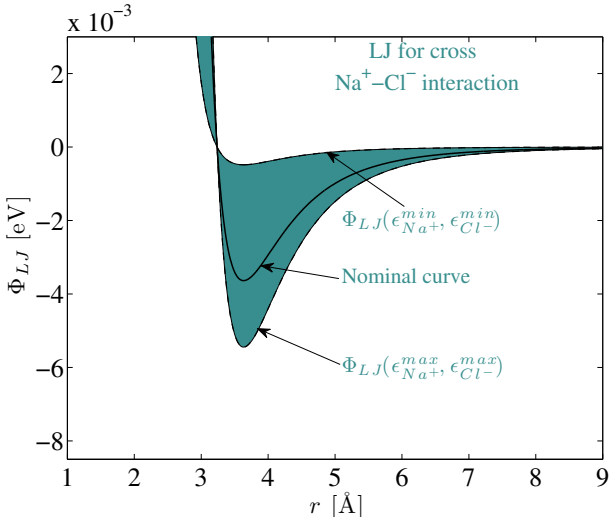
- Part I focused on the impact of potential uncertainties on key observables of the nanopore, revealing how thermal noise can play a key role.
- Part II showed the importance of taking account the uncertainty in the PC coefficients when running the inverse problem in noisy systems.
- PC expansions and Bayesian inference allowed us to isolate the impact of parametric uncertainty and properly capture the effect of the intrinsic noise.
- Showed the suitability of using PCe in the MD context for both the forward propagation and inverse problem.
- Potential for application to experimental data.

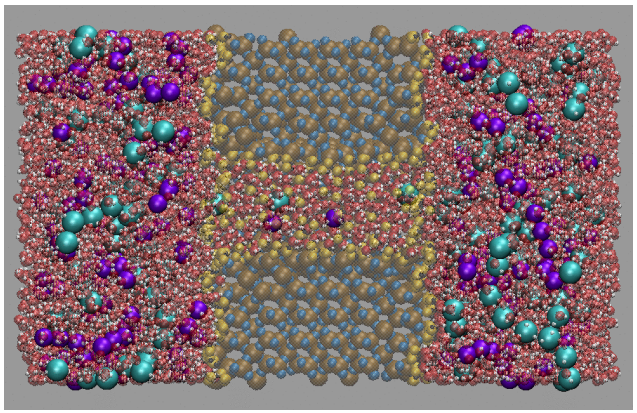
Summary & Conclusions

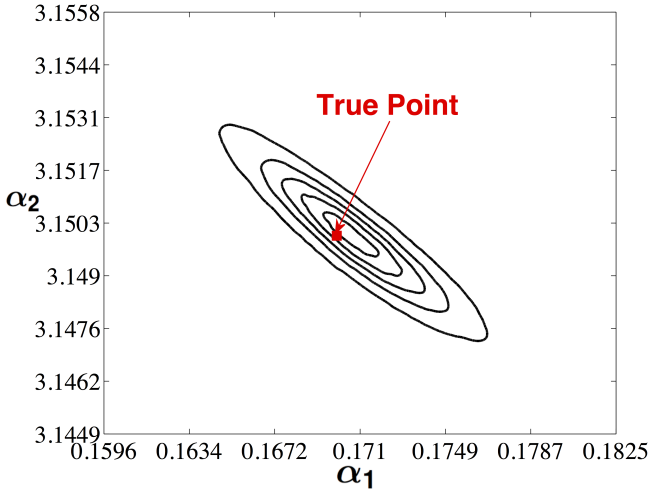
- Our group currently working on resilience computing for PDEs.
- Approach/implementation targeting resilience to:
  - Silent / Soft errors such as bit-flips.
  - Missing data due to communication issues or node failures.
- Approach involves casting PDE into sampling problem, followed by resilient data manipulation to get solution update.
- How about resilience for MD?
  - Data loss.
  - Reconstruct missing parts and full physical structure.
  - etc...

U.S. DoE, Office of Science, ASCR, under Award Number 13-016717.

Thank you for your attention







Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.