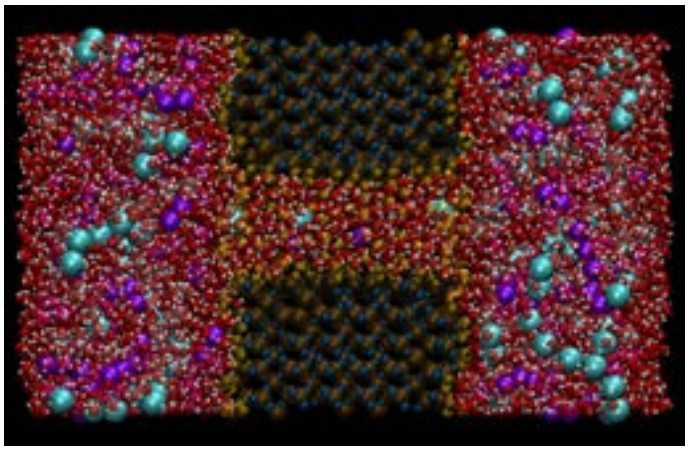
Uncertainty Quantification in Molecular Dynamics

Francesco Rizzi

Department of Mechanical Engineering and Materials Science
Duke University

KAUST

March 2013



JOHNS HOPKINS UNIVERSITY

Duke UNIVERSITY

Background and Motivation

George E. P. Box, 1987

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

Paraphrasing in less drastic terms...

"*All models are inaccurate in some aspects, some models are useful despite that, and for a given purpose some models are more useful than others.*"

- If not enough to convince you on using Uncertainty Quantification (UQ)...
- ...during this talk, I will try to! Specifically focusing on Molecular Dynamics (MD).

Background and Motivation

George E. P. Box, 1987

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

Paraphrasing in less drastic terms...

"*All models are inaccurate in some aspects, some models are useful despite that, and for a given purpose some models are more useful than others.*"

- If not enough to convince you on using Uncertainty Quantification (UQ)...
- ...during this talk, I will try to! Specifically focusing on Molecular Dynamics (MD).

Background and Motivation

George E. P. Box, 1987

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

Paraphrasing in less drastic terms...

"*All models are inaccurate in some aspects, some models are useful despite that, and for a given purpose some models are more useful than others.*"

- If not enough to convince you on using Uncertainty Quantification (UQ)...
- ...during this talk, I will try to! Specifically focusing on Molecular Dynamics (MD).

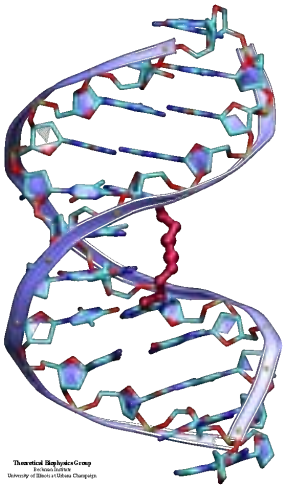
Background and Motivation

- 1957: origin of MD pioneered by Alder and Wainwright.
- 1964: first MD simulation based on a realistic potential (the Lennard-Jones potential), applied to liquid Argon (Rahman).
- 1974: first MD simulation of liquid water (Stillinger and Rahman).

  . . .

- MD is suitable and cheap (vs. experiments) to explore physical properties at the atomic level.
- Now widely employed in both industrial and academic environments:
  - variety of systems: from liquids to solids, to proteins and nucleic acids (DNA, RNA).
- As every simulation technique, MD is an approximation method with a few weaknesses...

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



MD snapshot of DNA (Biophysics group, UIUC)

Background and Motivation

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

$$\mathbf{F}_{i,t} = -\nabla_{\mathbf{r_i}} \Phi(\mathbf{r}_{1,t}, \ldots, \mathbf{r}_{N,t}), \qquad \text{and} \qquad \frac{d^2\mathbf{r}_{i,t}}{dt^2} = \frac{\mathbf{F}_{i,t}}{m_i} \qquad i = 1, \ldots, N.$$

- Φ is the potential (or force-field), must be defined before starting the simulation, and should be tailored to the target application.
- Reliability of an MD simulation mainly depends on the accuracy with which Φ can reproduce the atomic interactions occurring in the real system of interest.
  - e.g.: a potential providing an accurate description for solid fracture simulations might not be suitable to simulate the atomic behavior of a fluid.
- Continuous development of potentials and experience accumulated over the past few decades have made MD reliable for a variety of systems but...
- ...the weakness due to the potential uncertainty still poses a serious problem.
- MD potential represents the main source of uncertainty.
- One example to convince you: water...

Background and Motivation

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

$$\mathbf{F}_{i,t} = -\nabla_{\mathbf{r_i}} \Phi(\mathbf{r}_{1,t}, \ldots, \mathbf{r}_{N,t}), \qquad \text{and} \qquad \frac{d^2\mathbf{r}_{i,t}}{dt^2} = \frac{\mathbf{F}_{i,t}}{m_i} \qquad i = 1, \ldots, N.$$

- Φ is the potential (or force-field), must be defined before starting the simulation, and should be tailored to the target application.
- Reliability of an MD simulation mainly depends on the accuracy with which Φ can reproduce the atomic interactions occurring in the real system of interest.
  - e.g.: a potential providing an accurate description for solid fracture simulations might not be suitable to simulate the atomic behavior of a fluid.
- Continuous development of potentials and experience accumulated over the past few decades have made MD reliable for a variety of systems but...
- ...the weakness due to the potential uncertainty still poses a serious problem.
- MD potential represents the main source of uncertainty.
- One example to convince you: water...

Background and Motivation

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

$$\mathbf{F}_{i,t} = -\nabla_{\mathbf{r_i}} \Phi(\mathbf{r}_{1,t}, \ldots, \mathbf{r}_{N,t}), \qquad \text{and} \qquad \frac{d^2\mathbf{r}_{i,t}}{dt^2} = \frac{\mathbf{F}_{i,t}}{m_i} \qquad i = 1, \ldots, N.$$

- Φ is the potential (or force-field), must be defined before starting the simulation, and should be tailored to the target application.
- Reliability of an MD simulation mainly depends on the accuracy with which Φ can reproduce the atomic interactions occurring in the real system of interest.
  - e.g.: a potential providing an accurate description for solid fracture simulations might not be suitable to simulate the atomic behavior of a fluid.
- Continuous development of potentials and experience accumulated over the past few decades have made MD reliable for a variety of systems but...
- ...the weakness due to the potential uncertainty still poses a serious problem.
- MD potential represents the main source of uncertainty.
- One example to convince you: water...

Background and Motivation: Uncertainty for Water

- More than 50 MD water models available trying to provide suitable descriptions of the water molecule in the form of governing potentials, see (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. Data are obtained from the review by (Guillot,2002) and from listed references.

Background and Motivation: Uncertainty for Water

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

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

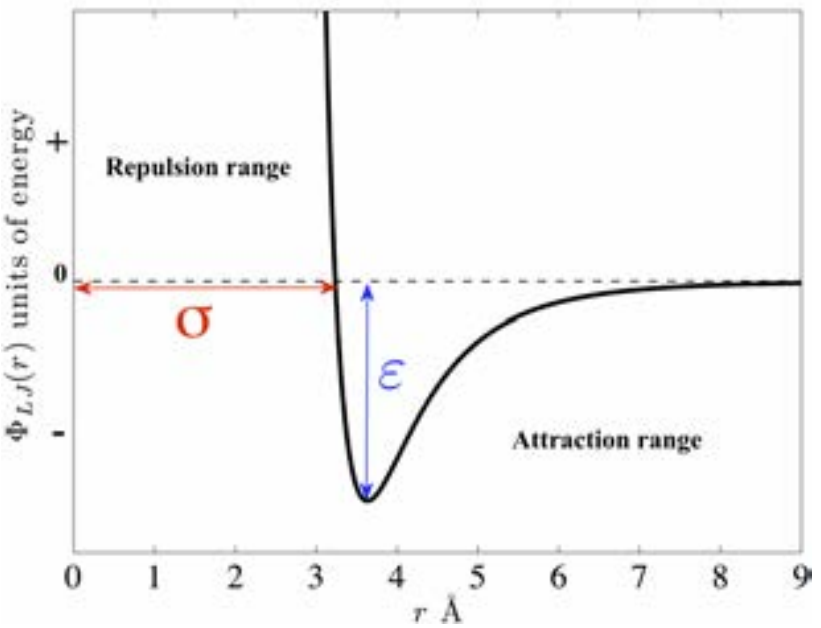
where r is separation between two O atoms of two water molecules.

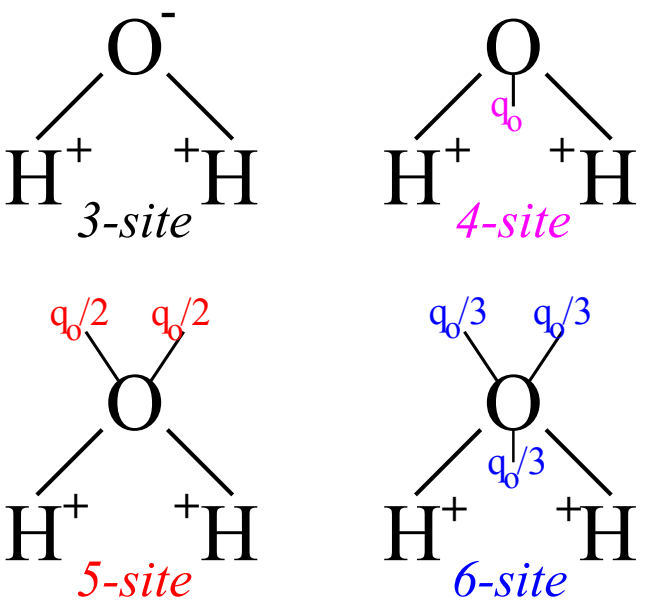
- Different models involve different values of the LJ parameters ε, σ.

b Rigid or flexible molecule.

c Number of sites defining the water structure, ranging from the 3-site H_2O structure, to 6 sites models.
...

- Discussion can be extended to several other systems or materials, indicating that the MD potential is an important source of uncertainty to consider in the MD setting.





Background and Motivation

Q1:

If MD is so “wrong”, why do we use it?

Q2:

Why not resorting to more reliable frameworks?
E.g. quantum mechanical calculations?

Background and Motivation: *Ab Initio* versus Classical MD

- 1984: ab initio MD by Car and Parrinello.

- Newton's law for the atoms trajectories, but the forces are obtained by solving the full quantum mechanical electronic structure problem.
  - √ No need for the potential.
  - X Large computational cost.
  - X Systems of the order of $10^3/10^4$ atoms.
  - X Practical time scales on the order of picoseconds.

Ab initio simulation of protein folding.
Isosurface reflects electrostatic potential -, +
due to the instantaneous electron configurations.
Source: Pietro Faccioli, Univ. of Trento, Italy

- Classical MD:
  - X Need potential.
  - √ Systems of about 10^6 atoms with current supercomputers.
  - √ Practical time-scales on the order of nanoseconds.

- Ideal run time for exploring atomistic systems is nanoseconds (microseconds preferably).

⇒ Feasible time scales is the key factor still making MD the preferred setting.

Background and Motivation: *Ab Initio* versus Classical MD

- 1984: ab initio MD by Car and Parrinello.

  - Newton's law for the atoms trajectories, but the forces are obtained by solving the full quantum mechanical electronic structure problem.
    - √ No need for the potential.
    - X Large computational cost.
    - X Systems of the order of $10^3/10^4$ atoms.
    - X Practical time scales on the order of picoseconds.

Ab initio simulation of protein folding.
Isosurface reflects electrostatic potential -, +
due to the instantaneous electron configurations.
Source: Pietro Faccioli, Univ. of Trento, Italy

- Classical MD:
  - X Need potential.
  - √ Systems of about 10^6 atoms with current supercomputers.
  - √ Practical time-scales on the order of nanoseconds.

- Ideal run time for exploring atomistic systems is nanoseconds (microseconds preferably).

⇒ Feasible time scales is the key factor still making MD the preferred setting.

Background and Motivation: *Ab Initio* versus Classical MD

- 1984: ab initio MD by Car and Parrinello.

  - Newton's law for the atoms trajectories, but the forces are obtained by solving the full quantum mechanical electronic structure problem.
    - $\surd$ No need for the potential.
    - X Large computational cost.
    - X Systems of the order of $10^3/10^4$ atoms.
    - X Practical time scales on the order of picoseconds.

Ab initio simulation of protein folding.
Isosurface reflects electrostatic potential -, +
due to the instantaneous electron configurations.
Source: Pietro Faccioli, Univ. of Trento, Italy

- Classical MD:
  - X Need potential.
  - $\surd$ Systems of about 10^6 atoms with current supercomputers.
  - $\surd$ Practical time-scales on the order of nanoseconds.

- Ideal run time for exploring atomistic systems is nanoseconds (microseconds preferably).

$\Rightarrow$ Feasible time scales is the key factor still making MD the preferred setting.

Background and Motivation: UQ

- Quantitative estimation of uncertainty in a computational study 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 a high-fidelity simulation analysis is of central importance.
- Two main probabilistic methods widely used for UQ: polynomial chaos (PC) expansion (Wiener,1938) and Bayesian inference (Bayes,1763).

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

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

Bayes' theorem: **D** is data - θ 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

The talk presents two tightly connected components of UQ applied to MD.

Part I: Forward Propagation

- Quantify how uncertainty in a set of model parameters affects selected model observables.

- Focus on MD simulations of concentration driven ionic flow in a silica nanopore.
- The heterogeneous nature of the system, due to the several components involved, represents a key complexity of this study.

Part II: Inverse Problem

- Estimation of 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.

Talk Overview

The talk presents two tightly connected components of UQ applied to MD.

Part I: Forward Propagation

- Quantify how uncertainty in a set of model parameters affects selected model observables.

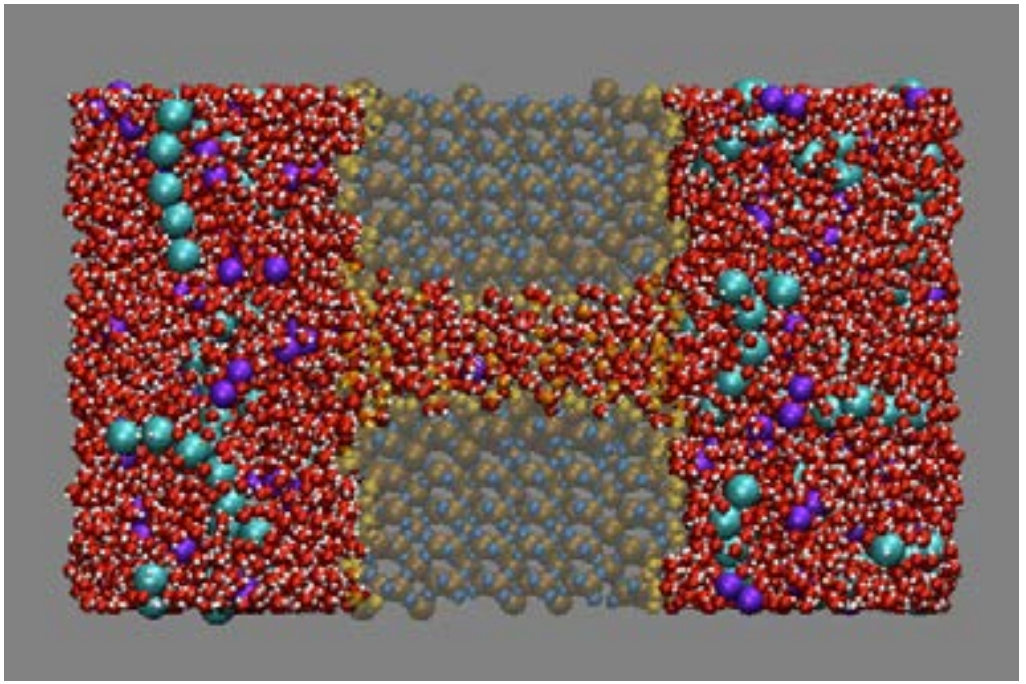
- Focus on MD simulations of concentration driven ionic flow in a silica nanopore.
- The heterogeneous nature of the system, due to the several components involved, represents a key complexity of this study.

Part II: Inverse Problem

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

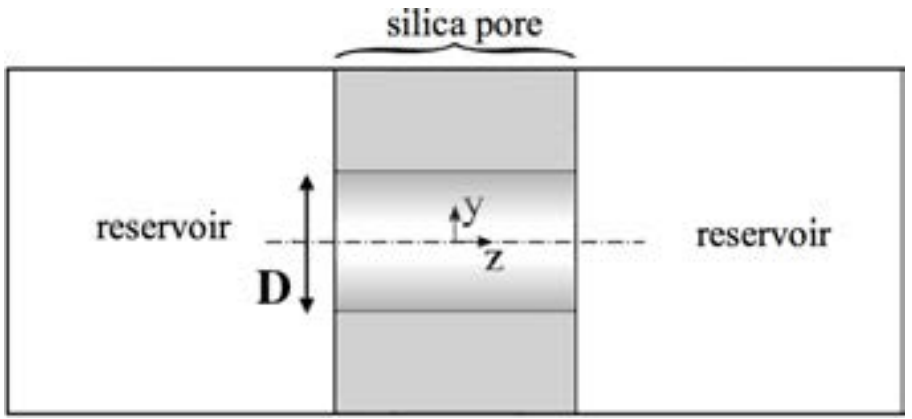
- Focus on MD simulations of bulk water.
- Estimation of potential parameters based on noisy observations of water observables.

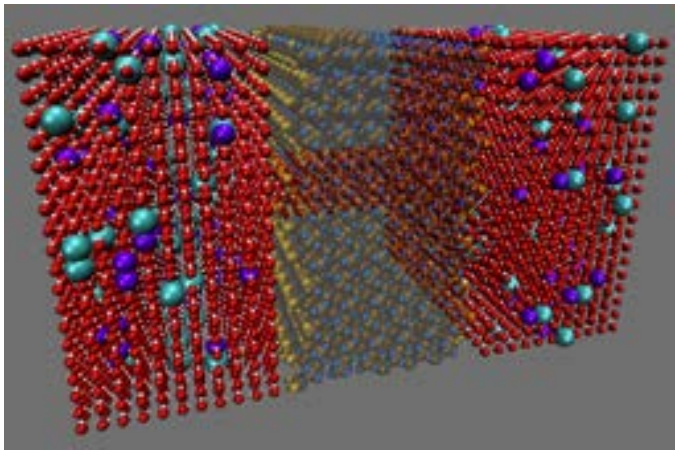
Forward propagation: nanopore flow



Computational System and Geometry

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



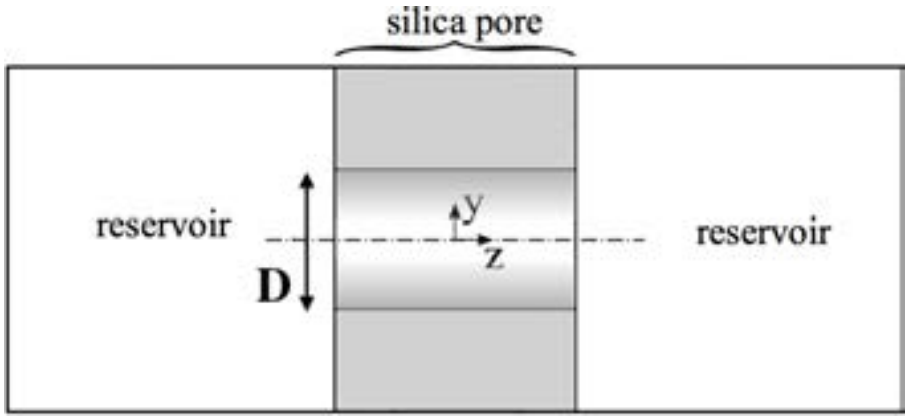


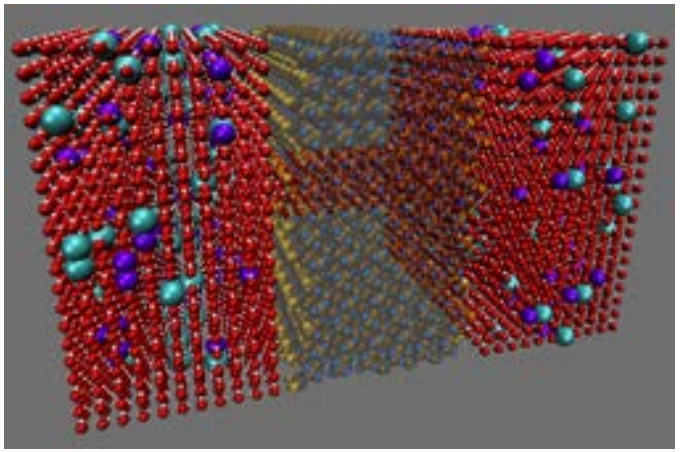
1. α-quartz crystal structure for the silica.
2. Remove the atoms within a cylindrical region of nominal diameter D.
3. Saturate dangling bonds with hydroxide groups (OH^-), to mimic real hydroxylation processes.

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

Computational System and Geometry

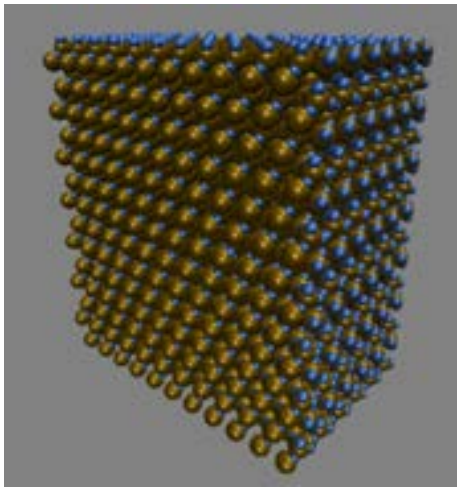
- We consider a silica pore model connecting two reservoirs containing a 1.5 mol/l solution of sodium (Na^+) and chloride (Cl^-) ions in H_2O (white-red).
- Reservoirs communicate only through the pore and PBC are imposed along x and y.





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

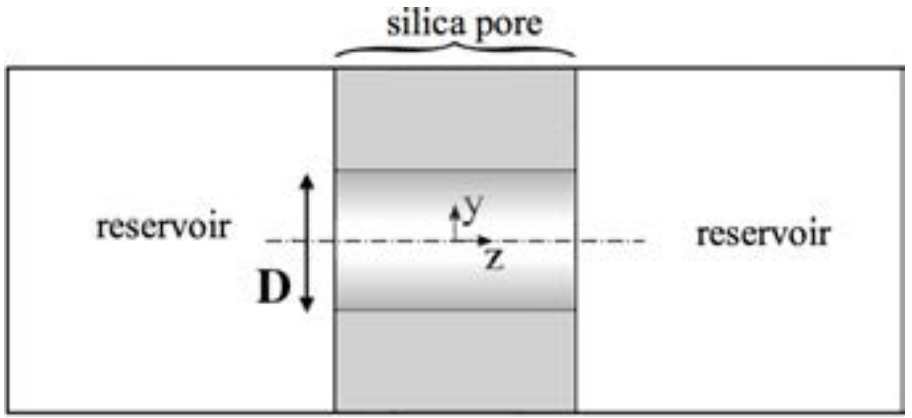
- Domain (xyz) $5.4 \times 6 \times 10.5$ nm^3.
- Total simulation time $\sim$ 8 ns.

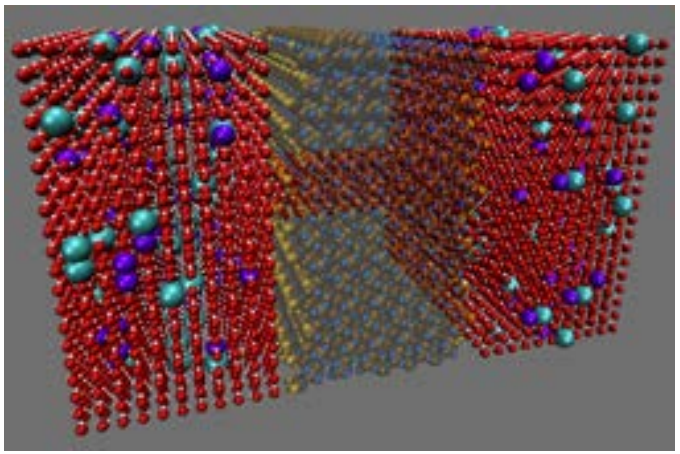


O_{bulk}, Si

Computational System and Geometry

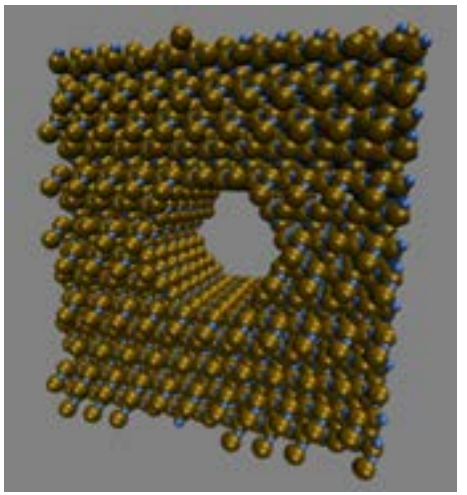
- We consider a silica pore model connecting two reservoirs containing a 1.5 mol/l solution of sodium (Na^+) and chloride (Cl^-) ions in H_2O (white-red).
- Reservoirs communicate only through the pore and PBC are imposed along x and y.





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

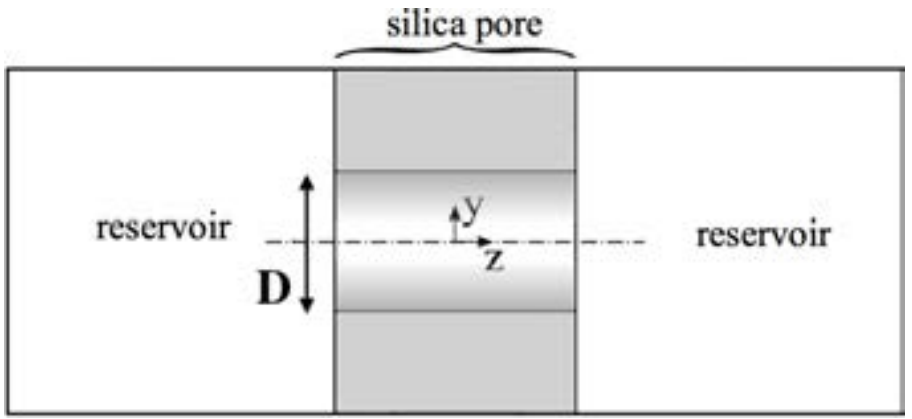
- Domain (xyz) $5.4 \times 6 \times 10.5$ nm^3.
- Total simulation time $\sim$ 8 ns.

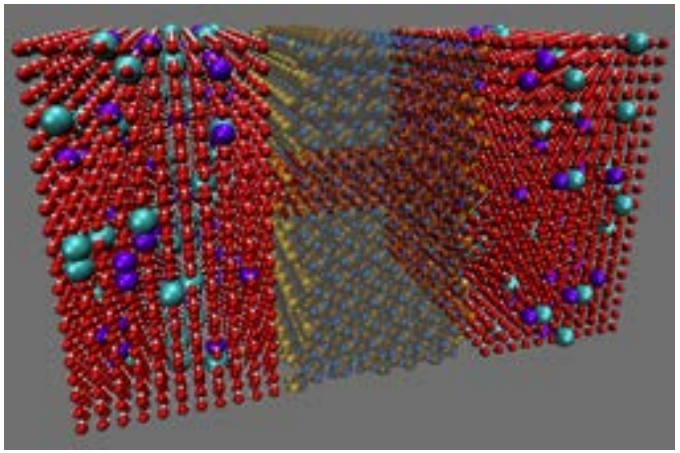


O_{bulk}, Si

Computational System and Geometry

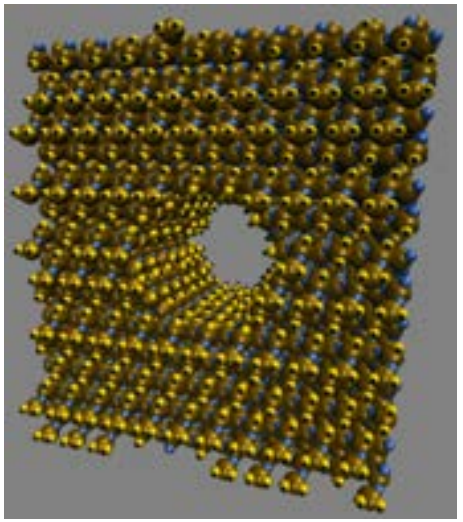
- We consider a silica pore model connecting two reservoirs containing a 1.5 mol/l solution of sodium (Na^+) and chloride (Cl^-) ions in H_2O (white-red).
- Reservoirs communicate only through the pore and PBC are imposed along x and y.





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

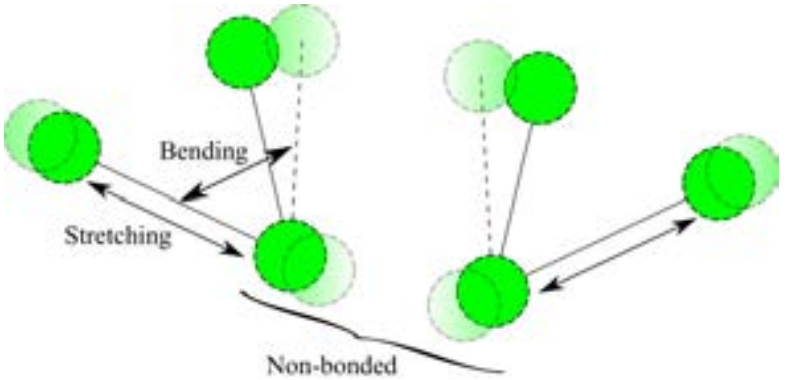
- Domain (xyz) $5.4 \times 6 \times 10.5$ nm^3.
- Total simulation time $\sim$ 8 ns.



O_{bulk}, Si, OH

MD Potential

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



- The non-bonded interactions (Van der Waals, Electrostatic) are modeled as:

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

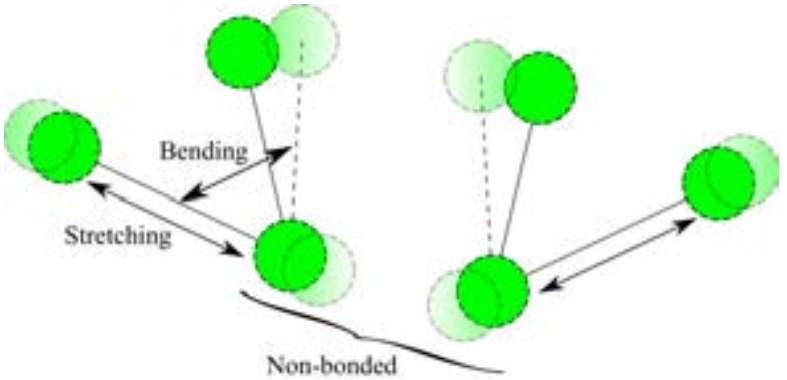
- Define the LJ parameters $\{\varepsilon_{\alpha\beta}, \sigma_{\alpha\beta}\}$ between atoms types α and β, for each homoatomic pair present in the system, i.e. $\alpha = \beta$.
- Calculate the cross-interaction parameters $\{\varepsilon_{\alpha\beta}, \sigma_{\alpha\beta}\}$, $\alpha \neq \beta$, using the Lorentz-Berthelot (LB) mixing rules:

$$\sigma_{\alpha\beta} = \frac{1}{2}(\sigma_\alpha + \sigma_\beta) \qquad \text{and} \qquad \varepsilon_{\alpha\beta} = \sqrt{\varepsilon_\alpha \varepsilon_\beta}$$

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

MD Potential

- We model the potential energy as:
$$\Phi_{total} = \Phi_{bonded} + \underbrace{\Phi_{LJ} + \Phi_{Coulomb}}_{non-bonded}$$



- Φ_{bonded} (bond stretching and bending) is modeled using harmonic potential.

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

- Define the LJ parameters $\{\varepsilon_{\alpha\beta}, \sigma_{\alpha\beta}\}$ between atoms types α and β, for each homoatomic pair present in the system, i.e. $\alpha = \beta$.

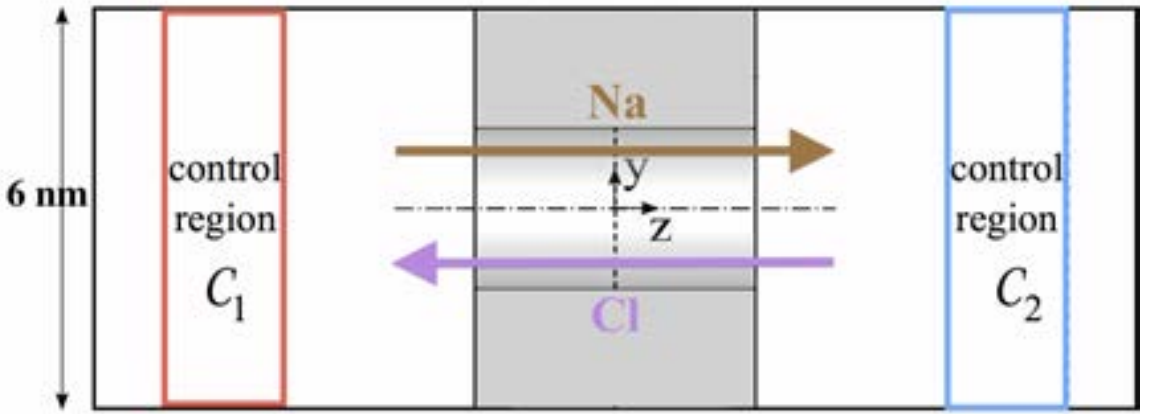
- Calculate the cross-interaction parameters $\{\varepsilon_{\alpha\beta}, \sigma_{\alpha\beta}\}$, $\alpha \neq \beta$, using the Lorentz-Berthelot (LB) mixing rules:
$$\sigma_{\alpha\beta} = \frac{1}{2}(\sigma_\alpha + \sigma_\beta) \qquad \text{and} \qquad \varepsilon_{\alpha\beta} = \sqrt{\varepsilon_\alpha \varepsilon_\beta}$$

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

Concentration Control Algorithm

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

- Inject/remove ions in two control regions $\mathcal{C}_1$ and $\mathcal{C}_2$.
- No ion deletion, only swapping: $\Rightarrow N$ is constant.
- $-\Delta c_{Na^+} = 30/V = \Delta c_{Cl^-}$:
  Flow Na^+: left $\rightarrow$ right
  Flow Cl^-: left $\leftarrow$ right



- Ionic flux magnitude:

$$J = \frac{N_{exchanges}}{\tau A} \quad \text{and condutance} \quad G = \frac{J}{|\overline{\Delta c}|}$$

  - $N_{exchanges}$ is the (net) number of ion exchanges between $\mathcal{C}_1$ and $\mathcal{C}_2$ over a time τ.
  - A is the cross-sectional area of the pore.
  - method was validated against a steady flux measured via integration of the velocity profiles of the ions over the cross-section of the pore.

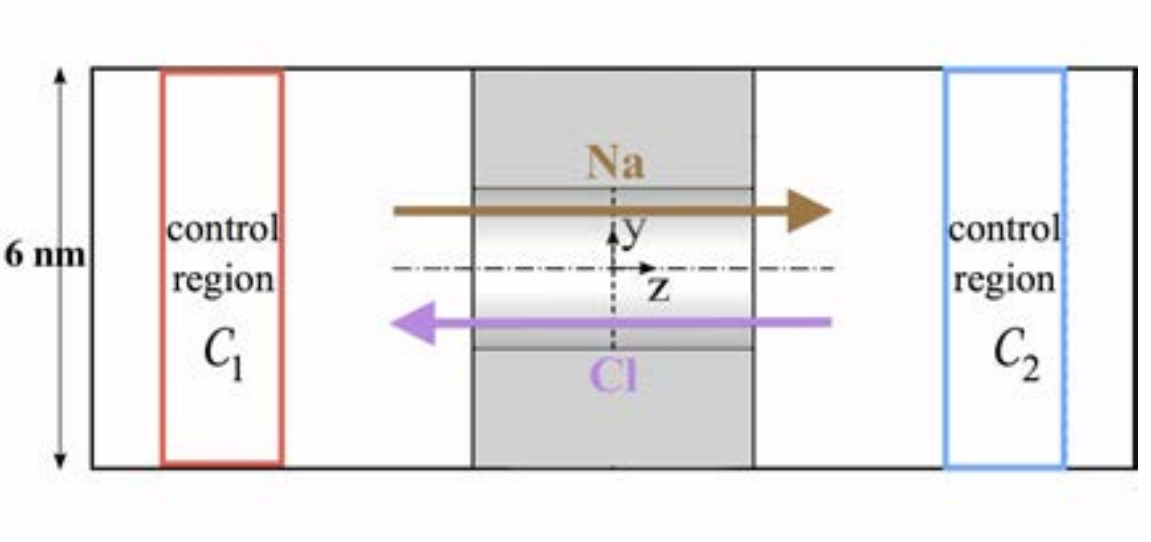
Concentration Control Algorithm

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

- Inject/remove ions in two control regions $\mathcal{C}_1$ and $\mathcal{C}_2$.
- No ion deletion, only swapping: $\Rightarrow N$ is constant.
- $-\Delta c_{Na^+} = 30/V = \Delta c_{Cl^-}$:

  Flow Na^+: left $\rightarrow$ right

  Flow Cl^-: left $\leftarrow$ right



- Ionic flux magnitude:

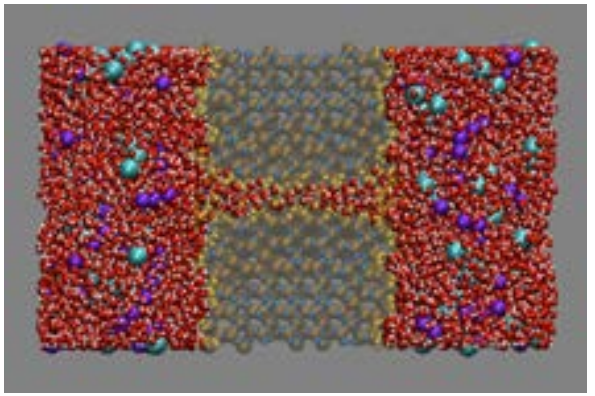
$$J = \frac{N_{exchanges}}{\tau A} \qquad \text{and condutance} \quad G = \frac{J}{|\overline{\Delta c}|}$$

  - $N_{exchanges}$ is the (net) number of ion exchanges between $\mathcal{C}_1$ and $\mathcal{C}_2$ over a time τ.
  - A is the cross-sectional area of the pore.
  - method was validated against a steady flux measured via integration of the velocity profiles of the ions over the cross-section of the pore.

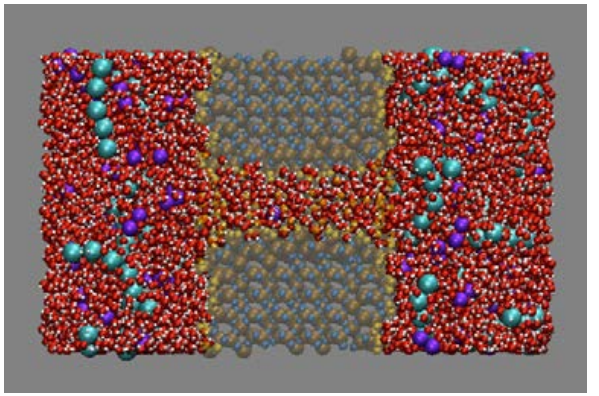
Sensitivity to the pore diameter

Sensitivity to Pore Diameter

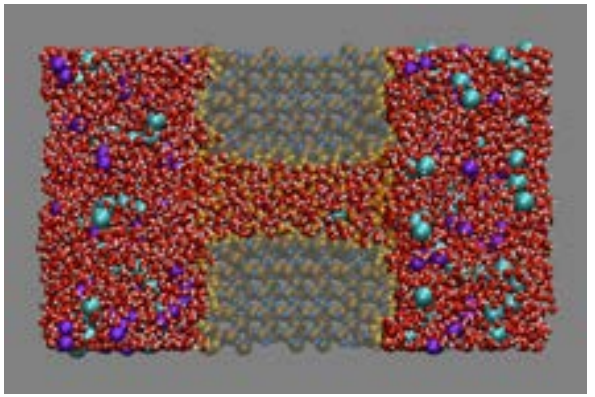
- Values of the pore diameter: $D = 12.5, 17, 21, 27$ Å.
- Total number of atoms:
  - 32124 for $D = 12.5$
  - 31630 for $D = 27$
- $D = 12.5$: smallest practical diameter yielding a non-zero ionic flux through the pore.
- $D = 27$: largest value such that the system does not "feel" effect of images.



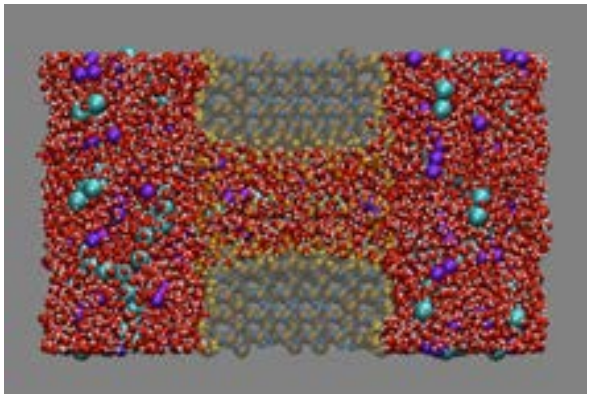
D=12.5 Å



D=17 Å



D=21 Å



D=27 Å

- 5 replica simulations are run for each value of D to account for the effect of the intrinsic (thermal) noise.
- These replicas are obtained using 5 different sets of random numbers to initialize the velocity field of the atoms.

Animation & Velocity Profile

- Animation for one replica of $D = 21$ Å.
- 440 ions and total of 31043 atoms.
- Ions tend to flow along the pore centerline with net mean velocity at steady state.
- At steady state, water is stationary.

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

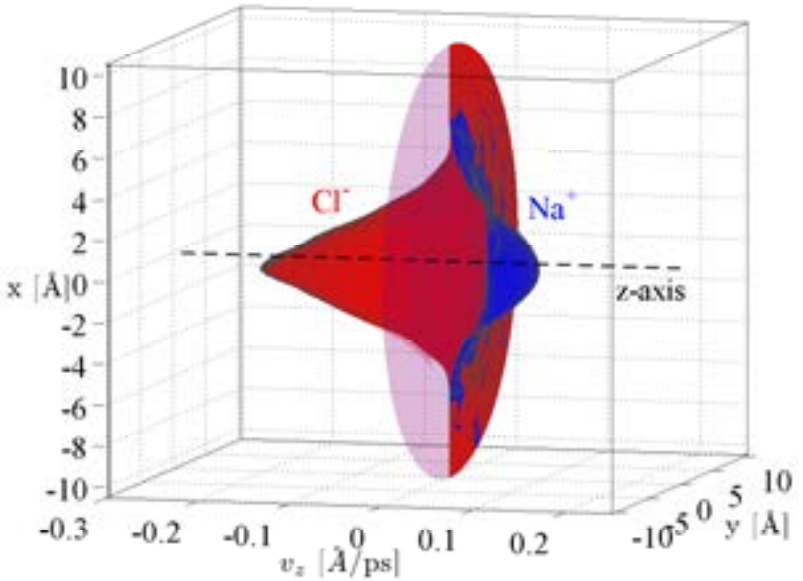
- Perform time/spatial averaging to construct radial profile of the axial velocity v_z.
- Spatial discretization based on "tessellation" binning and time window of 100 time steps.
- Parabolic profile for Na^+; conic for Cl^-.
- $v_z(r) \approx 0$ for $r \geq 7.5$ due to OH groups.
- Area under $v_z(r)$ indicates larger Cl^--flux.

Animation & Velocity Profile

- Animation for one replica of $D = 21$ Å.
- 440 ions and total of 31043 atoms.
- Ions tend to flow along the pore centerline with net mean velocity at steady state.
- At steady state, water is stationary.

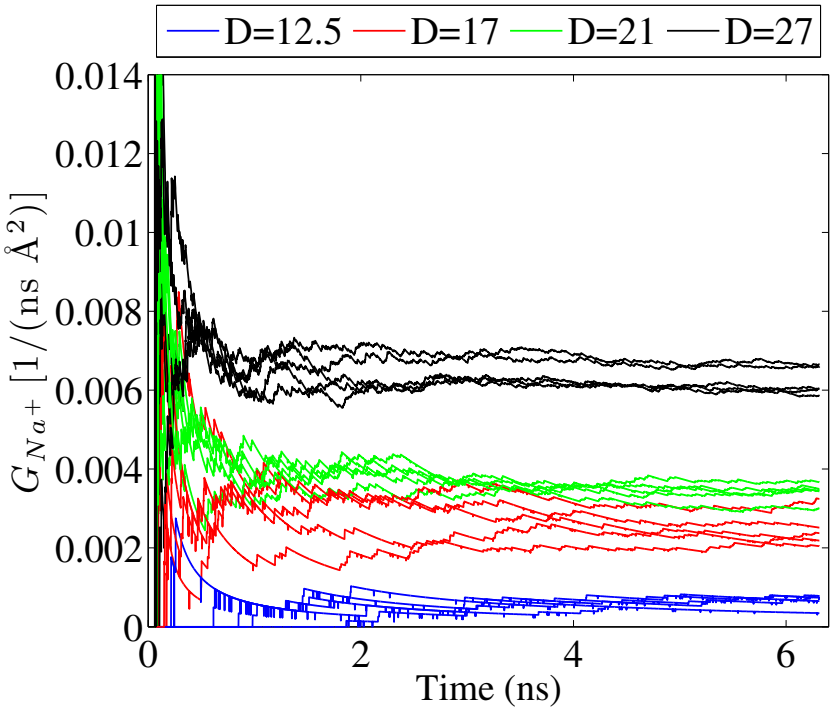
Na^+, Cl^-, H_2O (white-red), O_{bulk}, Si, OH

- Perform time/spatial averaging to construct radial profile of the axial velocity v_z.
- Spatial discretization based on "tessellation" binning and time window of 100 time steps.
- Parabolic profile for Na^+; conic for Cl^-.
- $v_z(r) \approx 0$ for $r \geq 7.5$ due to OH groups.
- Area under $v_z(r)$ indicates larger Cl^--flux.

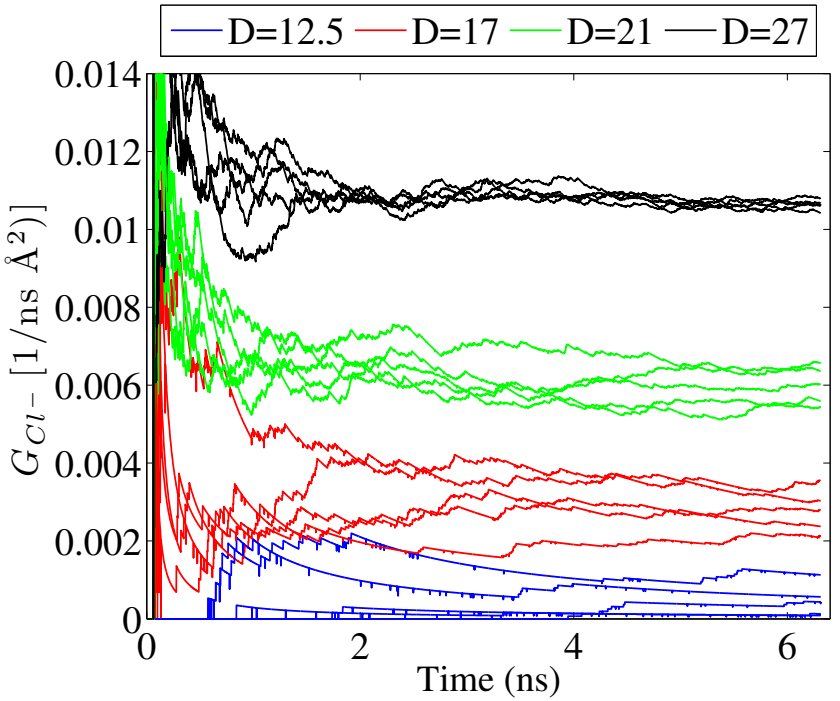


Sensitivity to Pore Diameter: Conductance

- Time evolution of the conductance, $G(t)$, computed for Na^+ and Cl^-, obtained for all 5 replicas at each diameter $D = 12.5$, 17, 21 and 27.
- Initial sharp transient showing strong fluctuations.
- Steady state with stable oscillations around a well-defined mean value.
  - D weakly affects the duration of the transient state.
  - As D increases, the steady state value of G substantially increases.



Na^+ conductance

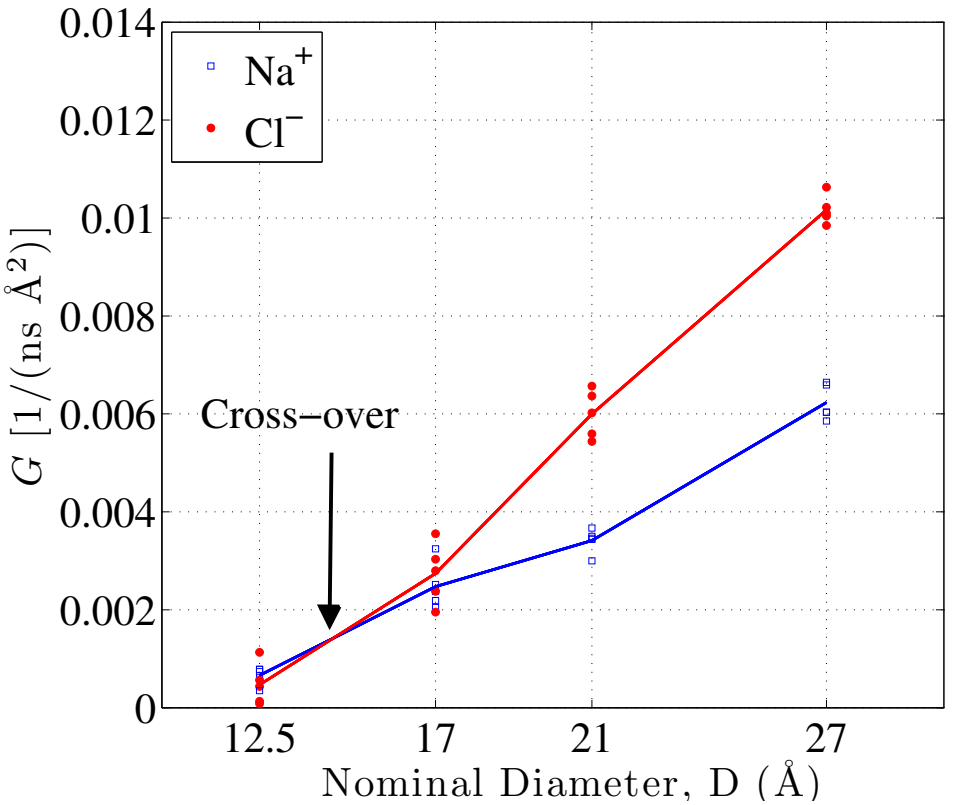


Cl^- conductance

Sensitivity to Pore Diameter: Conductance

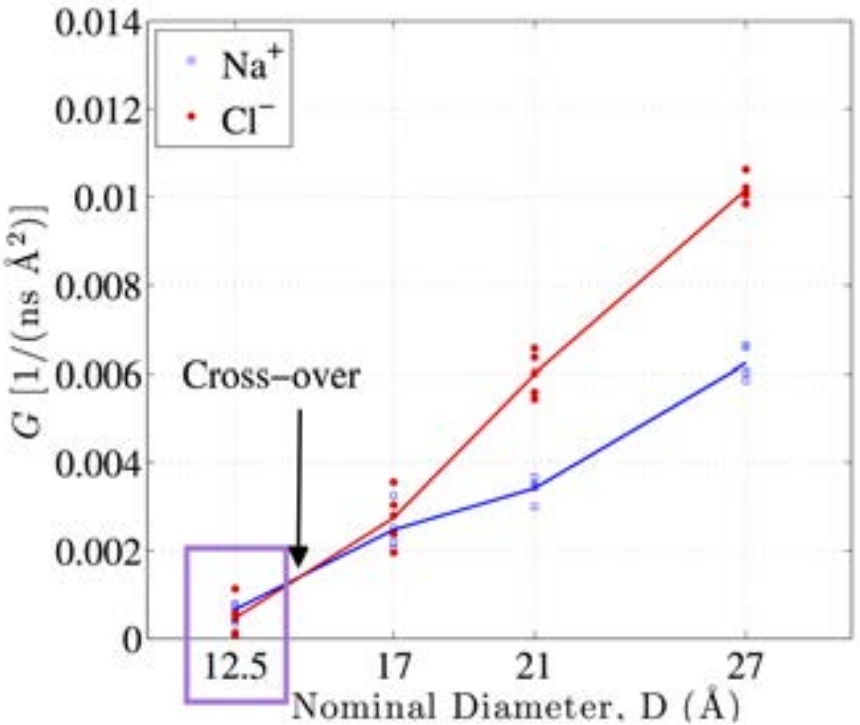
- Steady-state value of G_{Na^+} and G_{Cl^-} as a function of D for all 5 replicas showing the replica values (markers) and the mean trends (solid lines).

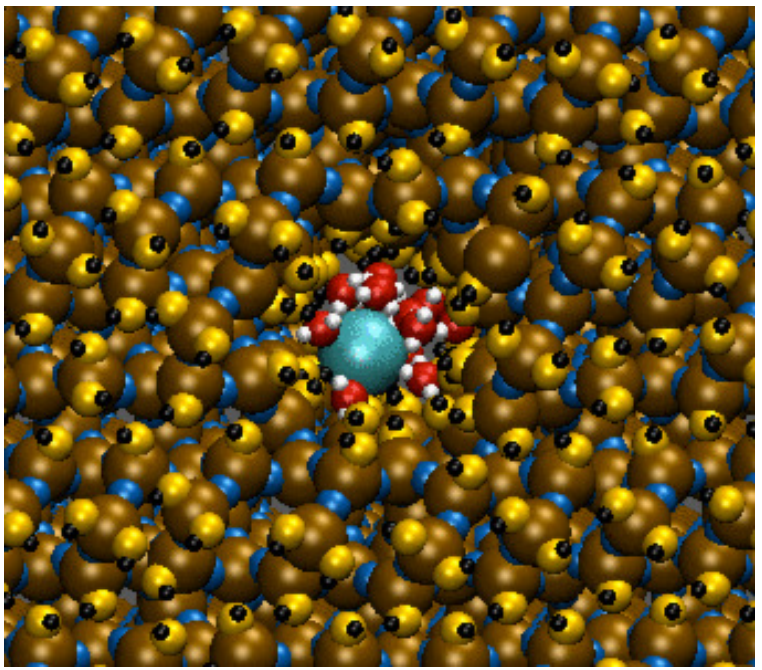
- Clear difference is observed in the steepness of the trend, which is sharper for the Cl^- conductance, G_{Cl^-}.
- Overlapping of replica distributions for small D.
- For $D \geq 17$: $\overline{G}_{Cl^-} > \overline{G}_{Na^+}$.
- The trend reverses for the smallest diameter $D = 12.5$.
- How to explain this physically?



Physical Explanation

- Cross-over is the result of the **interplay** between **size effects** and **ionic mobility**.
- $D = 12.5$: weak solvation shell $\Rightarrow$ strong effect of pore walls and confinement favor ions with smaller size, Na^+ (result seen before in Lyndenbell,1996).
- $D \geq 17$: complete solvation shell around the ions
  $\Rightarrow$ ion's mobility dominates
  $\Rightarrow$ the flux of Cl^- is greater than Na^+, because the diffusivity of Cl^- is larger.

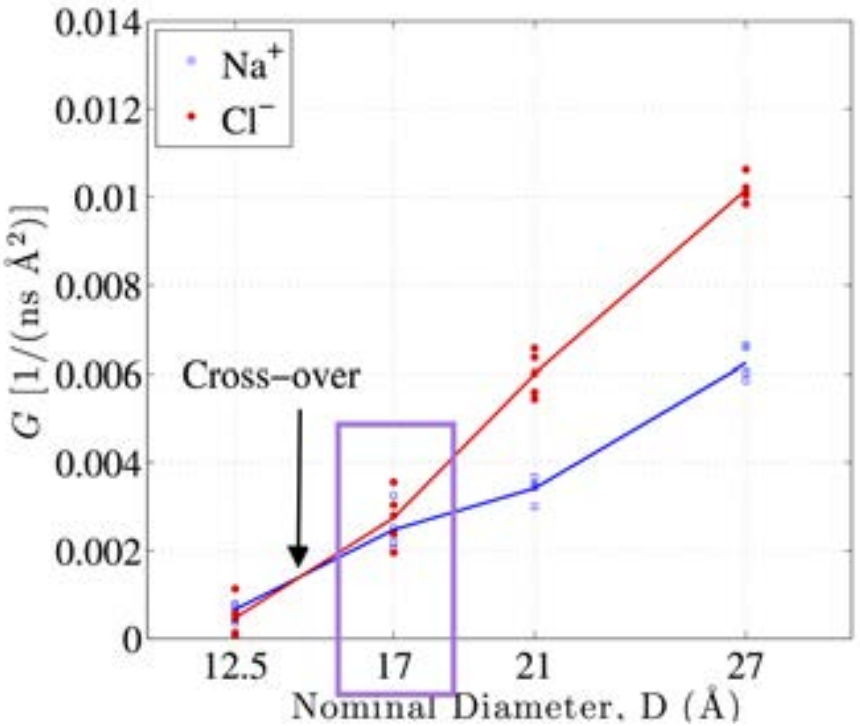


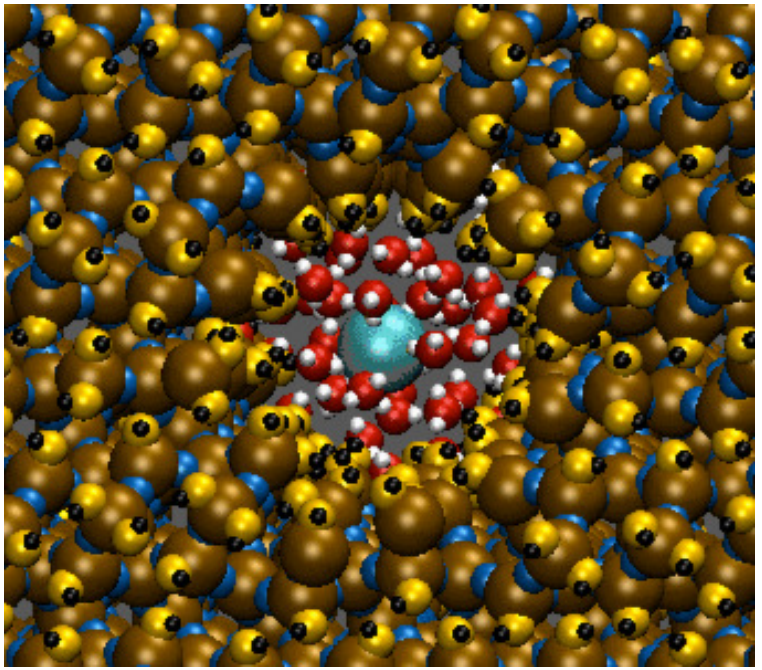


Na^+, Cl^-, H_2O (white-red), OH, Si, O_{bulk}

Physical Explanation

- Cross-over is the result of the **interplay** between **size effects** and **ionic mobility**.
- $D = 12.5$: weak solvation shell $\Rightarrow$ strong effect of pore walls and confinement favor ions with smaller size, Na^+ (result seen before in Lyndenbell,1996).
- $D \geq 17$: complete solvation shell around the ions
  $\Rightarrow$ ion's mobility dominates
  $\Rightarrow$ the flux of Cl^- is greater than Na^+, because the diffusivity of Cl^- is larger.

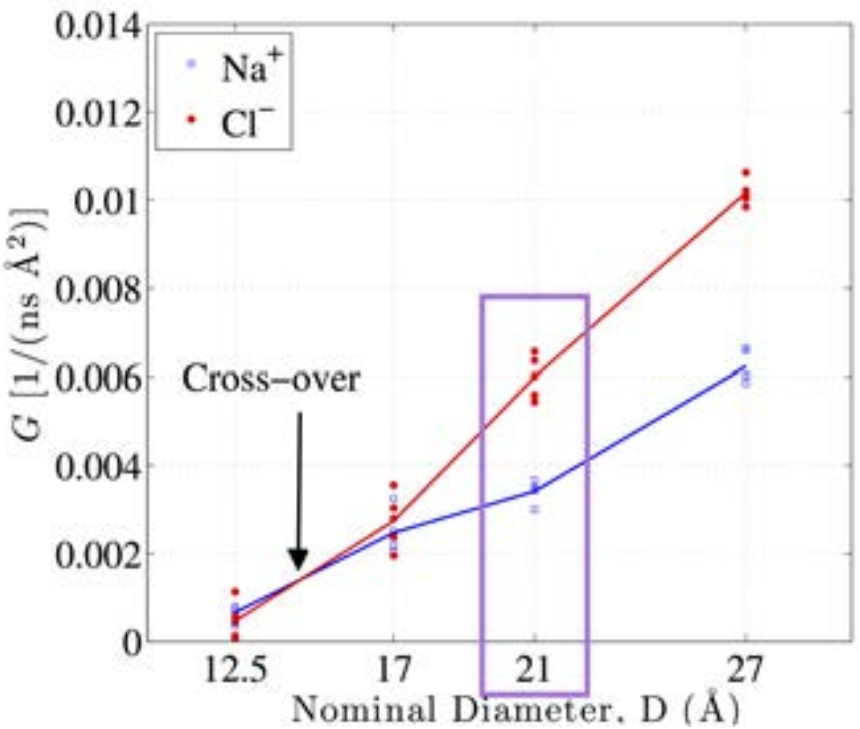


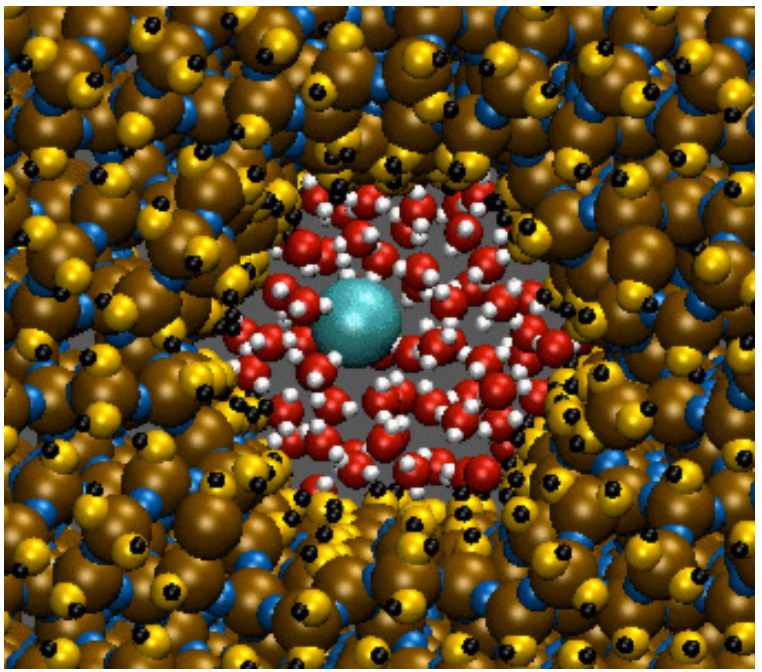


Na^+, Cl^-, H_2O (white-red), OH, Si, O_{bulk}

Physical Explanation

- Cross-over is the result of the **interplay** between **size effects** and **ionic mobility**.
- $D = 12.5$: weak solvation shell $\Rightarrow$ strong effect of pore walls and confinement favor ions with smaller size, Na^+ (result seen before in Lyndenbell,1996).
- $D \geq 17$: complete solvation shell around the ions
  $\Rightarrow$ ion's mobility dominates
  $\Rightarrow$ the flux of Cl^- is greater than Na^+, because the diffusivity of Cl^- is larger.

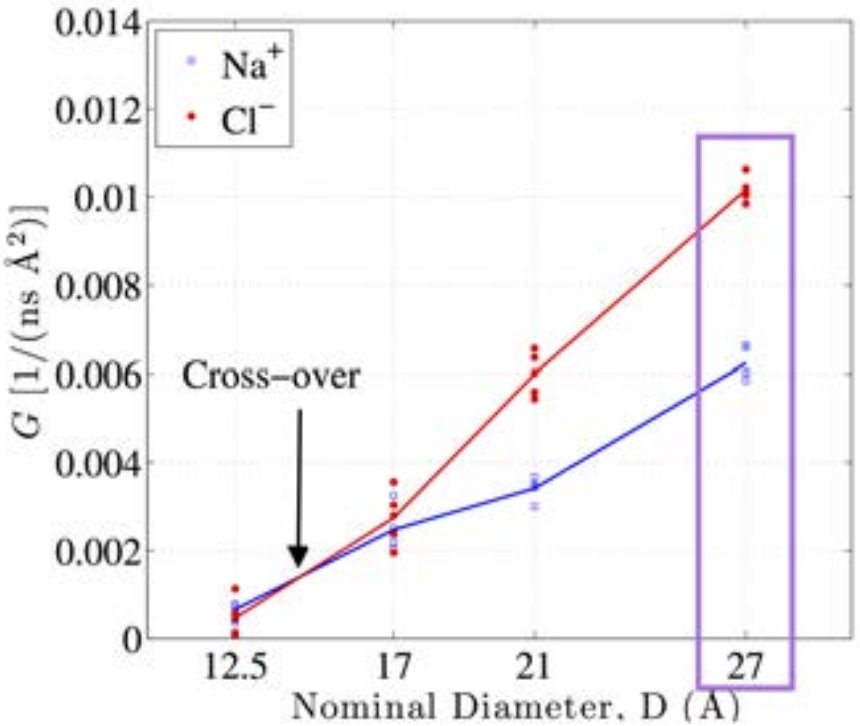


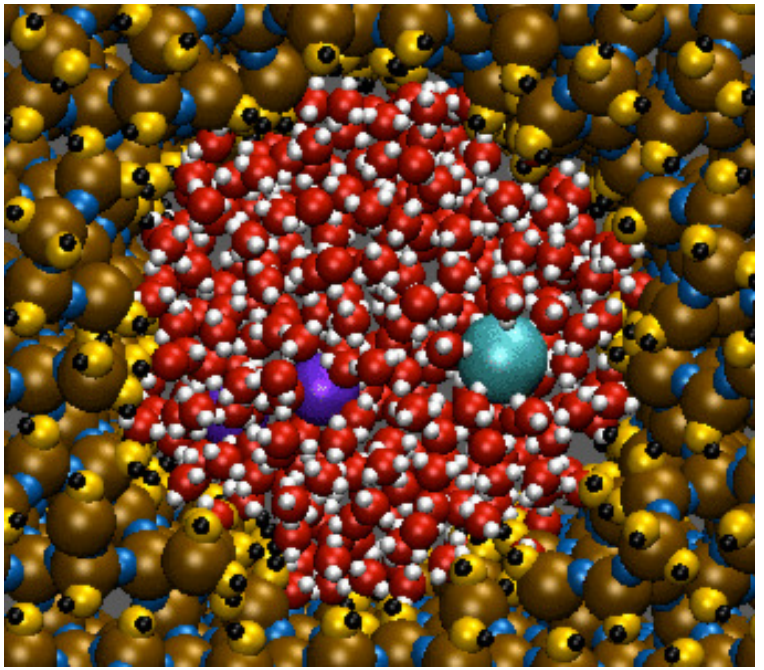


Na^+, Cl^-, H_2O (white-red), OH, Si, O_{bulk}

Physical Explanation

- Cross-over is the result of the **interplay** between **size effects** and **ionic mobility**.
- $D = 12.5$: weak solvation shell $\Rightarrow$ strong effect of pore walls and confinement favor ions with smaller size, Na^+ (result seen before in Lyndenbell,1996).
- $D \geq 17$: complete solvation shell around the ions
  $\Rightarrow$ ion's mobility dominates
  $\Rightarrow$ the flux of Cl^- is greater than Na^+, because the diffusivity of Cl^- is larger.





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

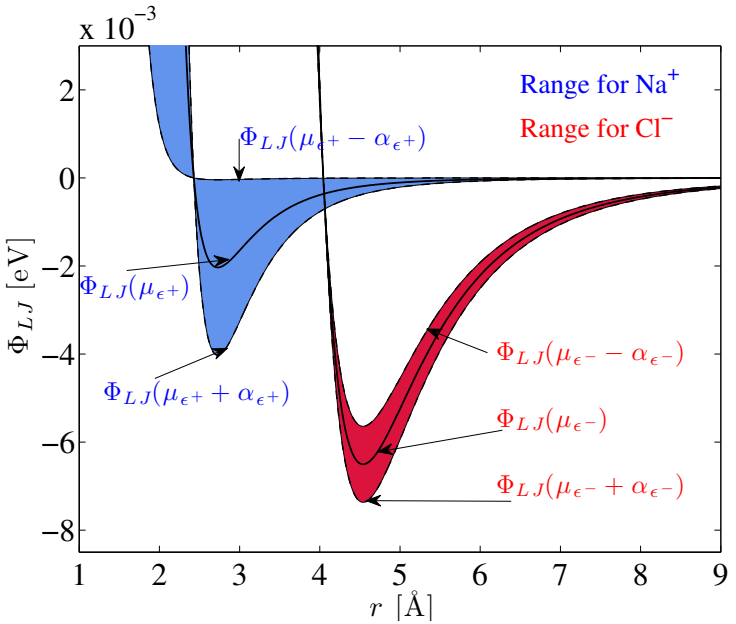
Sensitivity to LJ potential parameters

Problem Definition

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

$$\varepsilon_{Na^+} = 0.002033777 + 0.0019923703\ \xi_1, \qquad [\text{eV}],$$
$$\varepsilon_{Cl^-} = 0.006504600 + 0.0008630547\ \xi_2, \qquad [\text{eV}],$$

  where $\{\xi_1, \xi_2\}$ are *i.i.d.* uniform random variables; values from literature.
- Directly affects the potential for Na^+-Na^+ and Cl^--Cl^- LJ interactions.
- Since $\varepsilon_{\alpha\beta} = \sqrt{\varepsilon_\alpha \varepsilon_\beta}$ between different atom types $\alpha \neq \beta$, it also affects *all* the cross-interactions between each ion and the other atoms.



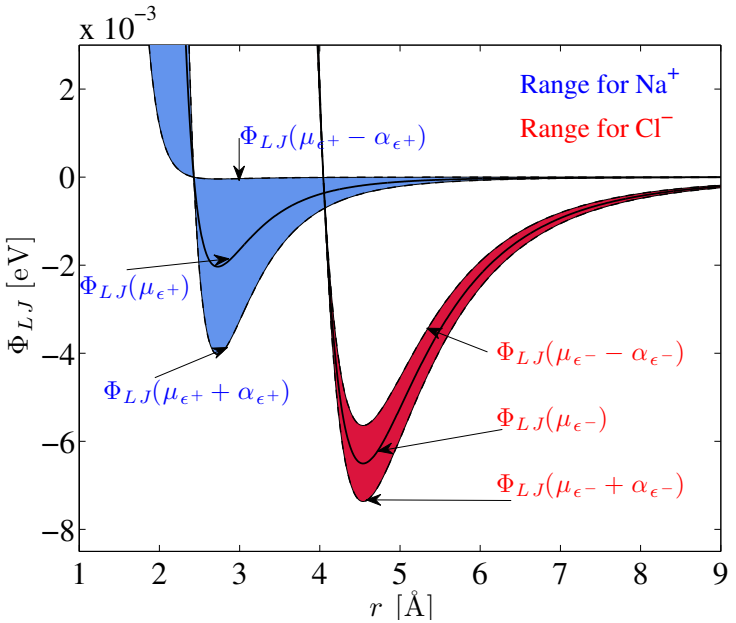
$\Phi_{LJ}(r)$ for Na^+-Na^+ and Cl^--Cl^-

Problem Definition

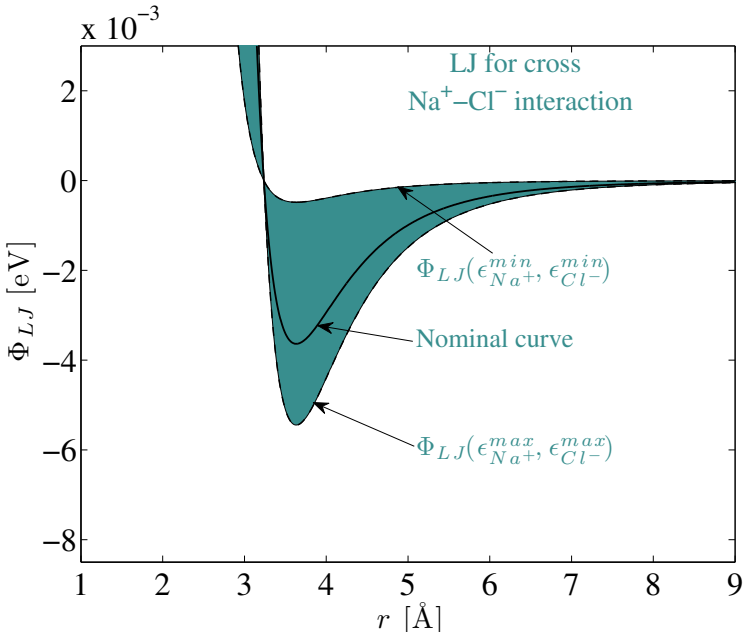
- Fix $D = 21$ Å; choose ε_{Na^+} and ε_{Cl^-}: depths of the LJ potential for Na^+ and Cl^-.

$$\varepsilon_{Na^+} = 0.002033777 + 0.0019923703\ \xi_1, \qquad \text{[eV]},$$
$$\varepsilon_{Cl^-} = 0.006504600 + 0.0008630547\ \xi_2, \qquad \text{[eV]},$$

  where $\{\xi_1, \xi_2\}$ are *i.i.d.* uniform random variables; values from literature.
- Directly affects the potential for Na^+-Na^+ and Cl^--Cl^- LJ interactions.
- Since $\varepsilon_{\alpha\beta} = \sqrt{\varepsilon_\alpha \varepsilon_\beta}$ between different atom types $\alpha \neq \beta$, it also affects *all* the cross-interactions between each ion and the other atoms.



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



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

Objective and Methods

- Stochastic reformulation

$$\begin{cases} \varepsilon_{Na^+} = f_1(\xi_1) \\ \varepsilon_{Cl^-} = f_2(\xi_2) \end{cases}$$

$\Rightarrow$ the MD predictions of the observables extracted from the nanopore simulation are random variables with finite variance.

- Goal: map the uncertainty from ε_{Na^+} and ε_{Cl^-} to the ionic conductance, G.

- Rely on:

$$G \approx \sum_{i=0}^{P} c_i \Psi_i(\xi_1, \xi_2)$$

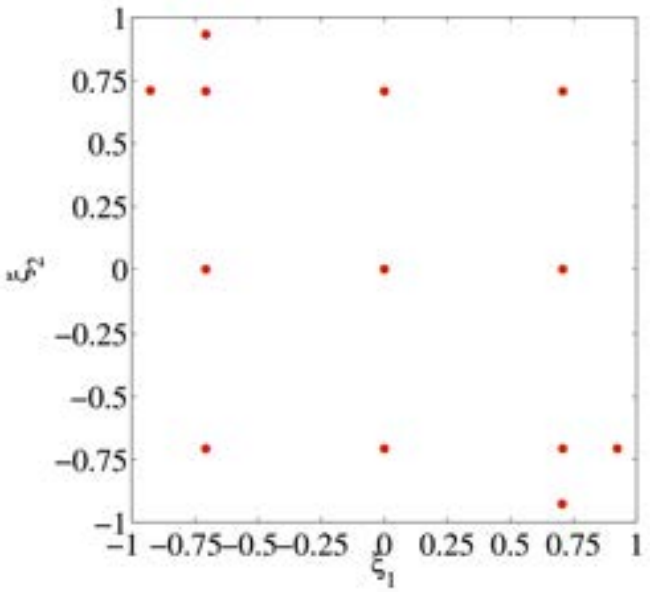
- Employ a Bayesian regression approach to find coefficients.

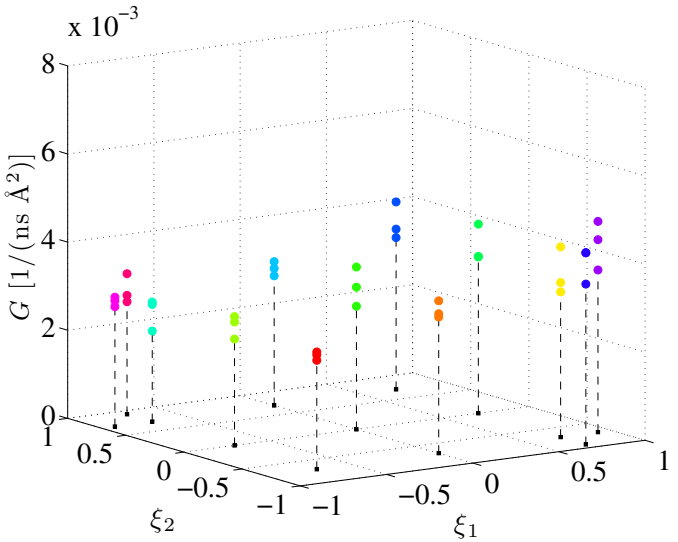
Regression: Collect Data

- Recall $\varepsilon_{Na^+} = f_1(\xi_1)$ and $\varepsilon_{Cl^-} = f_2(\xi_2)$.
- Collect data using 13 nodes and 3 MD replicas.
- $(-1, 1)$: small ε_{Na^+}, large ε_{Cl^-}; opposite for $(1, -1)$.
- Data-set of conductance:

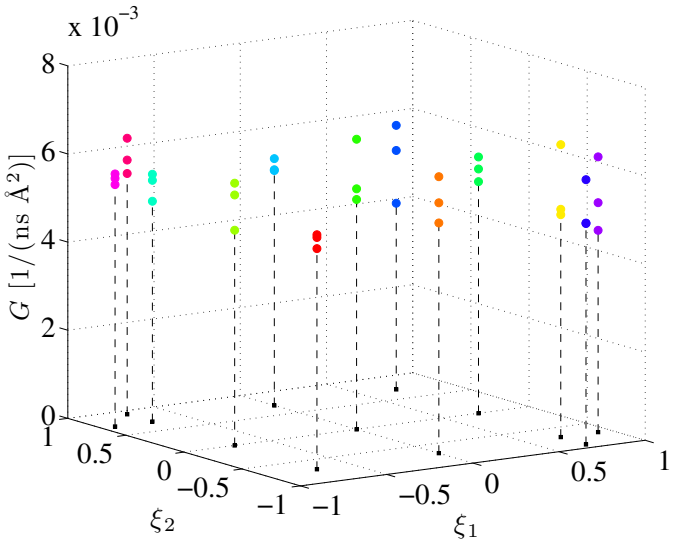
$$\mathbf{G} = \left\{ G_{i,j} \right\}_{i=1,\dots,13}^{j=1,\dots,3},$$

where G denotes the steady-state conductance computed for either Na^+ or Cl^-.





Results for Na^+.

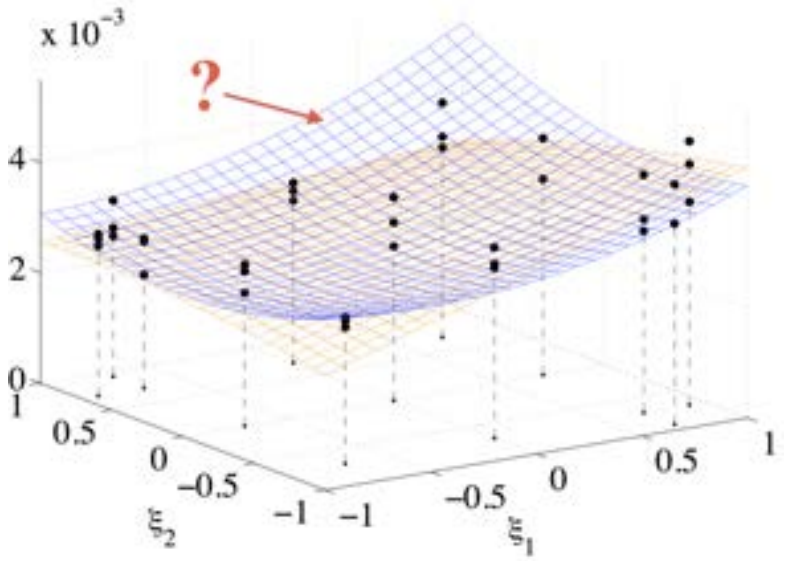


Results for Cl^-.

Bayesian Regression Formulation

- Goal: represent a given set of data, **G**, using a target regression function, $M(\xi_1, \xi_2)$, in the form of a PCe:

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



- Regression model: assume additive errors as

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

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

- Assume $\{\gamma_\ell\}_{\ell=1}^{39}$ to be *independent* and *normally* distributed with *mean zero*.
- Independence assumption is justified since the data points result of independent runs of the MD system.
- Consider a space-dependent noise STD $\sigma_\ell = \sigma(\boldsymbol{\xi})$ by parametrizing:

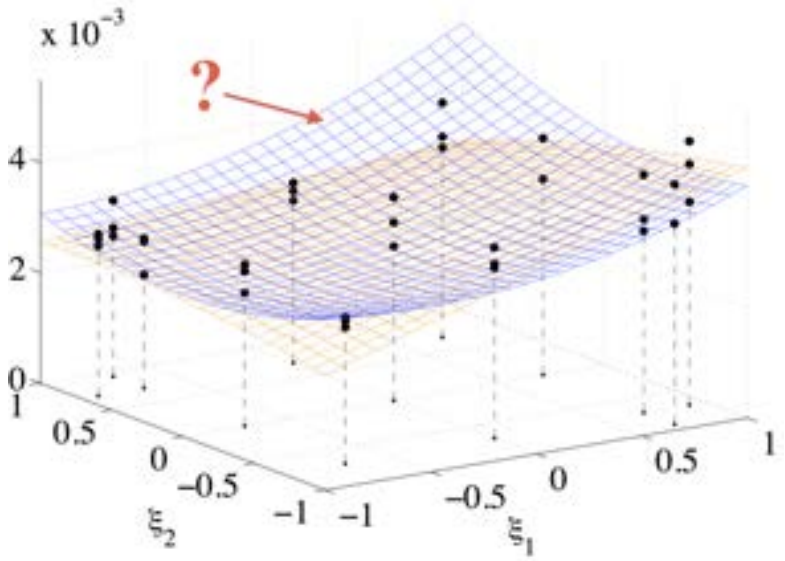
$$\sigma(\boldsymbol{\xi}) = h_0 + h_1\xi_1 + h_2\xi_2.$$

- We thus have: $\gamma_\ell \sim \mathcal{N}(0, \sigma(\boldsymbol{\xi}_\ell))$.

Bayesian Regression Formulation

- Goal: represent a given set of data, **G**, using a target regression function, $M(\xi_1, \xi_2)$, in the form of a PCe:

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



- Regression model: assume additive errors as

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

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

- Assume $\{\gamma_\ell\}_{\ell=1}^{39}$ to be *independent* and *normally* distributed with *mean zero*.
- Independence assumption is justified since the data points result of independent runs of the MD system.
- Consider a space-dependent noise STD $\sigma_\ell = \sigma(\boldsymbol{\xi})$ by parametrizing:

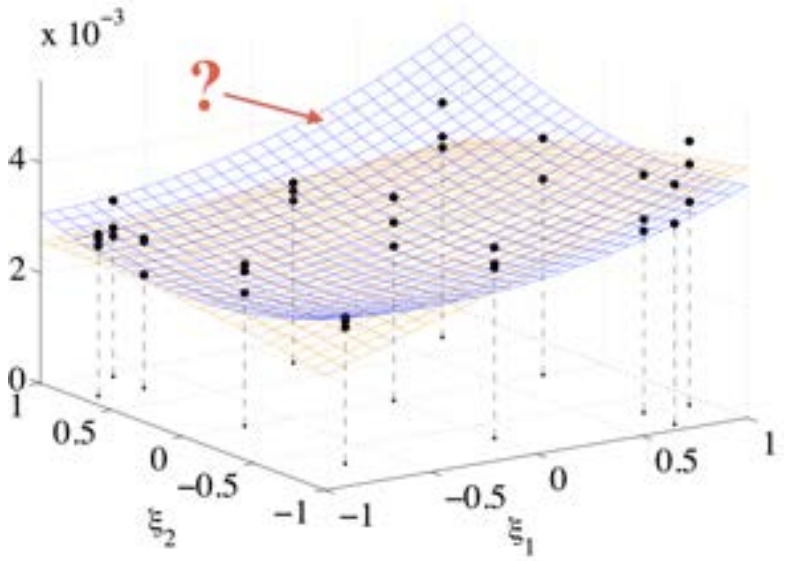
$$\sigma(\boldsymbol{\xi}) = h_0 + h_1\xi_1 + h_2\xi_2.$$

- We thus have: $\gamma_\ell \sim \mathcal{N}(0, \sigma(\boldsymbol{\xi}_\ell))$.

Bayesian Regression Formulation

- Goal: represent a given set of data, **G**, using a target regression function, $M(\xi_1, \xi_2)$, in the form of a PCe:

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



- Regression model: assume additive errors as

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

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

- Assume $\{\gamma_\ell\}_{\ell=1}^{39}$ to be *independent* and *normally* distributed with *mean zero*.
- Independence assumption is justified since the data points result of independent runs of the MD system.
- Consider a space-dependent noise STD $\sigma_\ell = \sigma(\boldsymbol{\xi})$ by parametrizing:

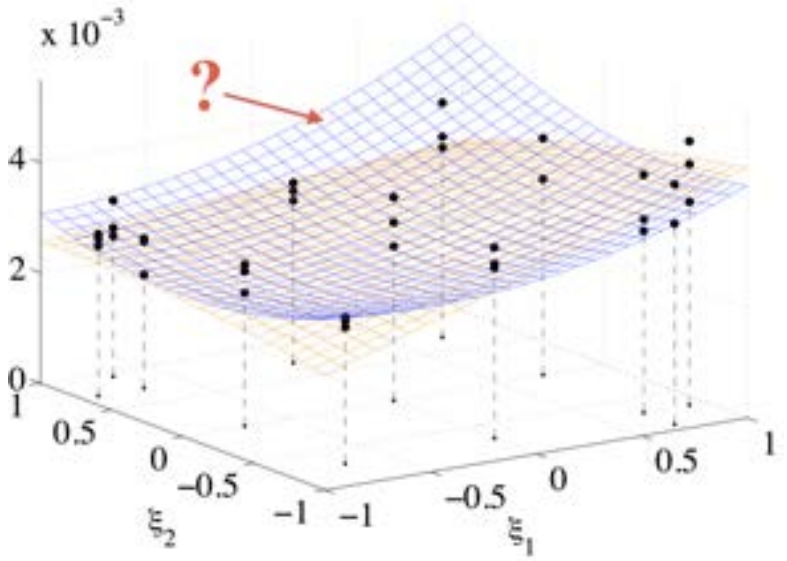
$$\sigma(\boldsymbol{\xi}) = h_0 + h_1\xi_1 + h_2\xi_2.$$

- We thus have: $\gamma_\ell \sim \mathcal{N}(0, \sigma(\boldsymbol{\xi}_\ell))$.

Bayesian Regression Formulation

- Goal: represent a given set of data, **G**, using a target regression function, $M(\xi_1, \xi_2)$, in the form of a PCe:

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



- Regression model: assume additive errors as

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

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

- Assume $\{\gamma_\ell\}_{\ell=1}^{39}$ to be *independent* and *normally* distributed with *mean zero*.
- Independence assumption is justified since the data points result of independent runs of the MD system.
- Consider a space-dependent noise STD $\sigma_\ell = \sigma(\boldsymbol{\xi})$ by parametrizing:

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

- We thus have: $\gamma_\ell \sim \mathcal{N}(0, \sigma(\boldsymbol{\xi}_\ell))$.

Bayesian Regression Formulation

- Treat the coefficients $\{h_k\}_{k=0}^{2}$ as hyperparameters, i.e. they become part of the set of unknowns which becomes: $\{g_0, \ldots, g_P, h_0, h_1, h_2\}$.
- We can construct the following likelihood

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

  where $G_{i,j}$ is the j-th observation obtained at the i-th sampling point, $\boldsymbol{\xi}_i$.
- Bayes' theorem thus yields the following joint posterior

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

  where the priors account for the information avaiable about the unknowns *before* considering the data: use uniform prior with suitably large bounds.
- The posterior π is sampled with a Markov chain Monte Carlo (MCMC) algorithm: random walk in the $\{g_0, \ldots, g_P, h_0, h_1, h_2\}$-space.

Bayesian Regression Formulation

- Treat the coefficients $\{h_k\}_{k=0}^2$ as hyperparameters, i.e. they become part of the set of unknowns which becomes: $\{g_0, \dots, g_P, h_0, h_1, h_2\}$.
- We can construct the following likelihood

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

  where $G_{i,j}$ is the j-th observation obtained at the i-th sampling point, $\boldsymbol{\xi}_i$.
- Bayes' theorem thus yields the following joint posterior

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

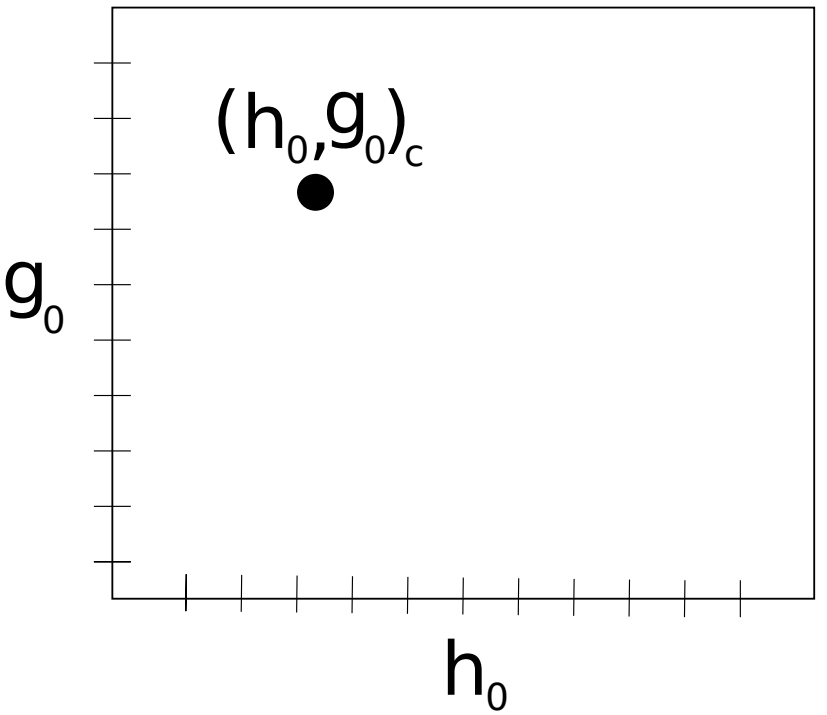
  where the priors account for the information avaiable about the unknowns *before* considering the data: use uniform prior with suitably large bounds.
- The posterior π is sampled with a Markov chain Monte Carlo (MCMC) algorithm: random walk in the $\{g_0, \dots, g_P, h_0, h_1, h_2\}$-space.

Markov chain Monte Carlo (MCMC)

- For simplicity, suppose that we were in 2D to infer $\{g_0, h_0\}$.

$$\pi\left(g_0, h_0 \,\middle|\, data\right) \propto \mathcal{L}\left(data|g_0, h_0\right) U_{g_0} U_{h_0},$$

1. Let $(g_0, h_0)_c$ be an initial guess.
2. Draw a candidate point (g_0', h_0') from a Gaussian centered on the current state: $(g_0', h_0') \sim \mathcal{N}((g_0, h_0)_c, Cov)$.
3. Calculate the ratio:
$$r = \frac{\pi(g_0', h_0'|data)}{\pi((g_0, h_0)_c|data)}$$
4. Draw a sample $\theta \sim U(0, 1)$.
5. The chain then moves according to:
$$(g_0, h_0)_{t=1} = \begin{cases} (g_0', h_0') & \text{if} \quad \theta < r, \\ (g_0, h_0)_c & otherwise. \end{cases}$$
6. Repeat the loop.



Markov chain Monte Carlo (MCMC)

- For simplicity, suppose that we were in 2D to infer $\{g_0, h_0\}$.

$$\pi\left(g_0, h_0 \,\middle|\, data\right) \propto \mathcal{L}\left(data|g_0, h_0\right) U_{g_0} U_{h_0},$$

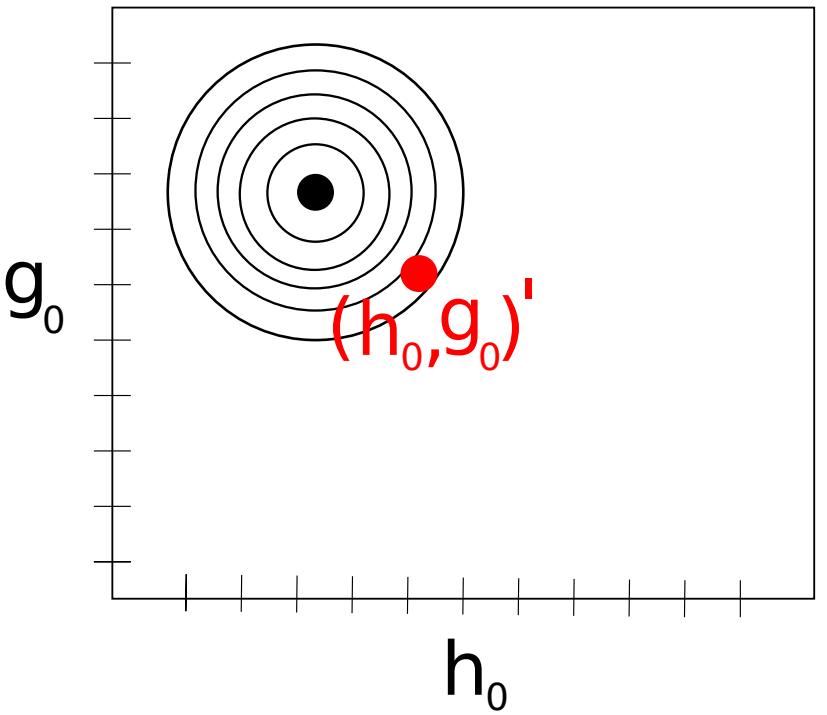
1. Let $(g_0, h_0)_c$ be an initial guess.
2. Draw a candidate point (g_0', h_0') from a Gaussian centered on the current state: $(g_0', h_0') \sim \mathcal{N}((g_0, h_0)_c, Cov)$.
3. Calculate the ratio:

$$r = \frac{\pi(g_0', h_0'|data)}{\pi((g_0, h_0)_c|data)}$$

4. Draw a sample $\theta \sim U(0, 1)$.
5. The chain then moves according to:

$$(g_0, h_0)_{t=1} = \begin{cases} (g_0', h_0') & \text{if} \quad \theta < r, \\ (g_0, h_0)_c & otherwise. \end{cases}$$

6. Repeat the loop.



Markov chain Monte Carlo (MCMC)

- For simplicity, suppose that we were in 2D to infer $\{g_0, h_0\}$.

$$\pi\left(g_0, h_0 \,\middle|\, data\right) \propto \mathcal{L}\left(data|g_0, h_0\right) U_{g_0} U_{h_0},$$

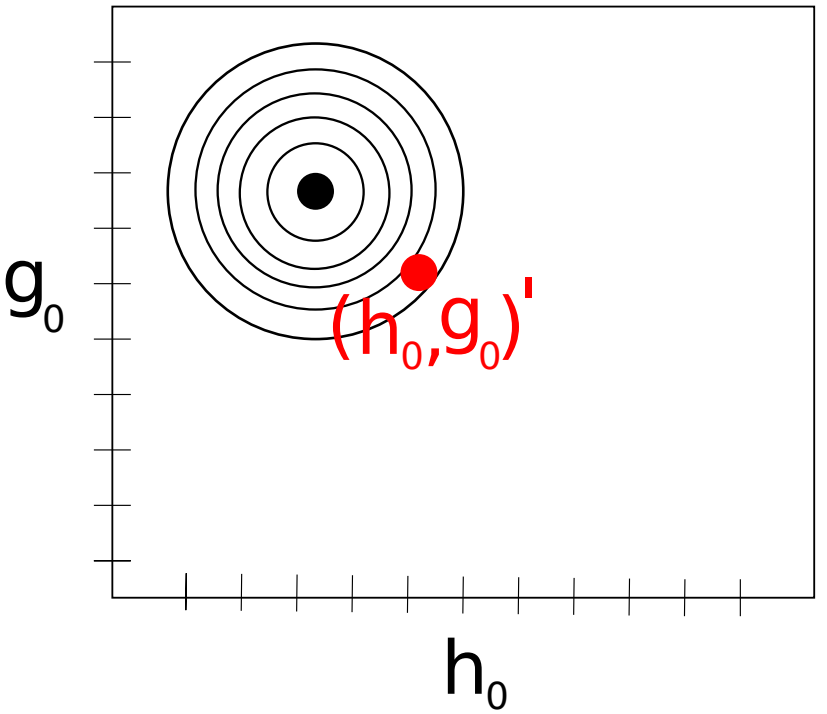
1. Let $(g_0, h_0)_c$ be an initial guess.
2. Draw a candidate point (g_0', h_0') from a Gaussian centered on the current state: $(g_0', h_0') \sim \mathcal{N}((g_0, h_0)_c, Cov)$.
3. Calculate the ratio:

$$r = \frac{\pi(g_0', h_0'|data)}{\pi((g_0, h_0)_c|data)}$$

4. Draw a sample $\theta \sim U(0, 1)$.
5. The chain then moves according to:

$$(g_0, h_0)_{t=1} = \begin{cases} (g_0', h_0') & \text{if} \quad \theta < r, \\ (g_0, h_0)_c & otherwise. \end{cases}$$

6. Repeat the loop.



Markov chain Monte Carlo (MCMC)

- For simplicity, suppose that we were in 2D to infer $\{g_0, h_0\}$.

$$\pi\left(g_0, h_0 \,\middle|\, data\right) \propto \mathcal{L}\left(data|g_0, h_0\right) U_{g_0} U_{h_0},$$

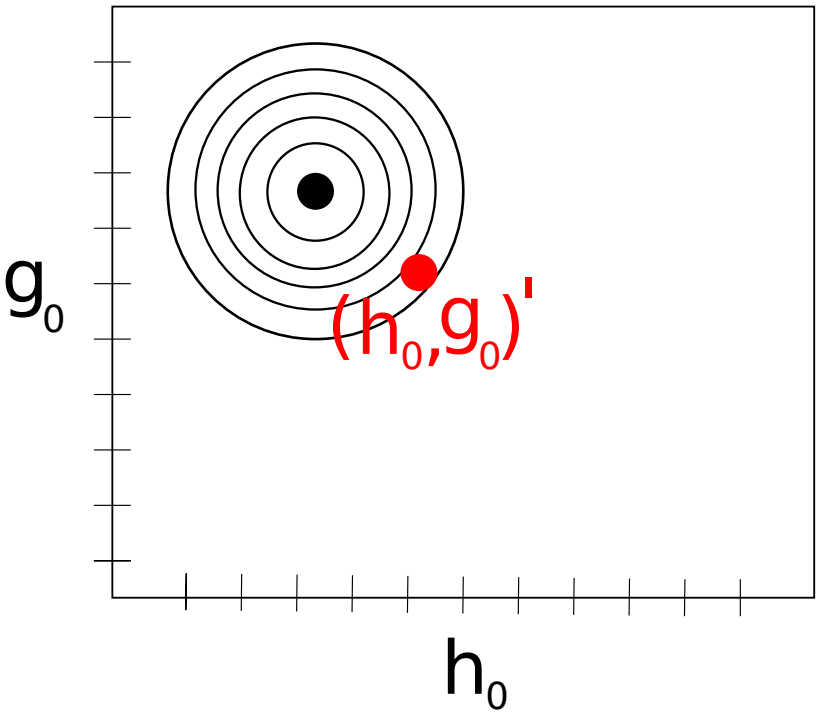
1. Let $(g_0, h_0)_c$ be an initial guess.
2. Draw a candidate point (g_0', h_0') from a Gaussian centered on the current state: $(g_0', h_0') \sim \mathcal{N}((g_0, h_0)_c, Cov)$.
3. Calculate the ratio:

$$r = \frac{\pi(g_0', h_0'|data)}{\pi((g_0, h_0)_c|data)}$$

4. Draw a sample $\theta \sim U(0, 1)$.
5. The chain then moves according to:

$$(g_0, h_0)_{t=1} = \begin{cases} (g_0', h_0') & \text{if} \quad \theta < r, \\ (g_0, h_0)_c & otherwise. \end{cases}$$

6. Repeat the loop.

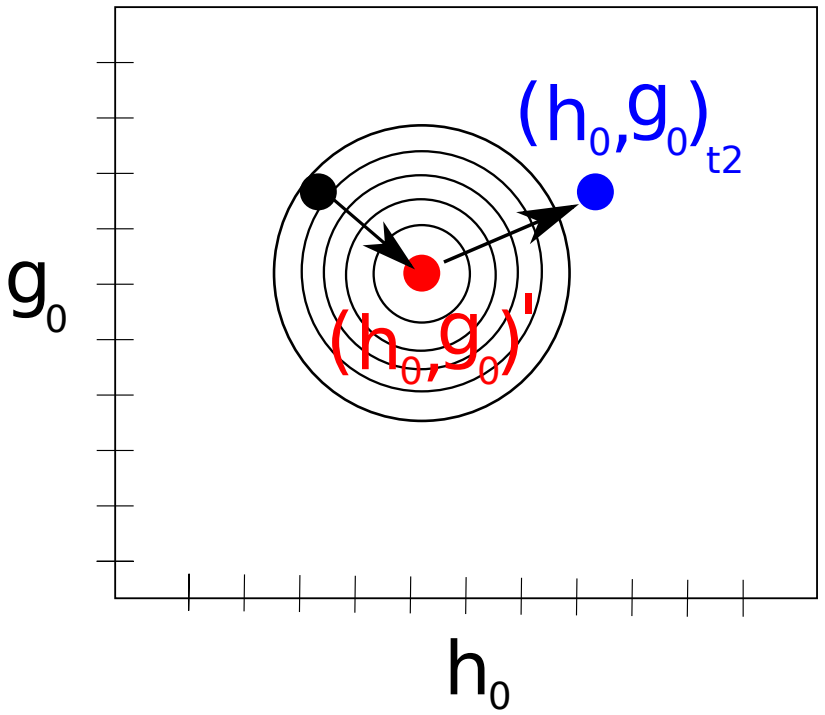


Markov chain Monte Carlo (MCMC)

- For simplicity, suppose that we were in 2D to infer $\{g_0, h_0\}$.

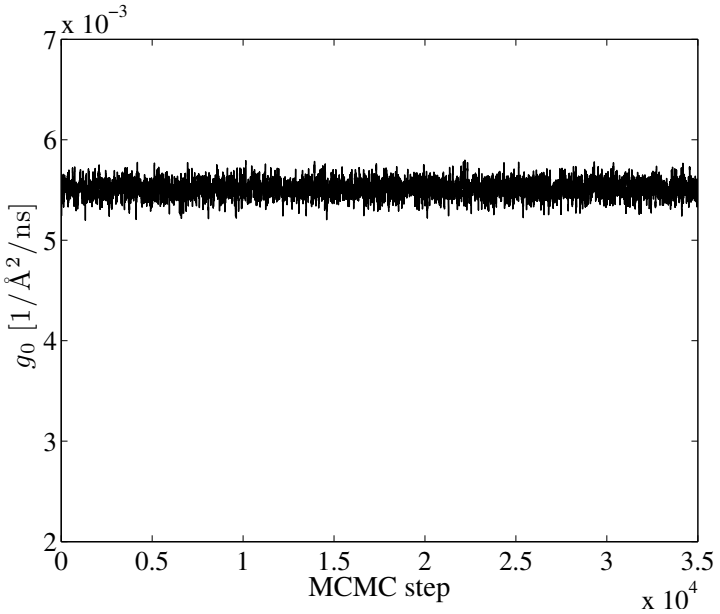
$$\pi\left(g_0, h_0 \,\middle|\, data\right) \propto \mathcal{L}\left(data|g_0, h_0\right) U_{g_0} U_{h_0},$$

1. Let $(g_0, h_0)_c$ be an initial guess.
2. Draw a candidate point (g_0', h_0') from a Gaussian centered on the current state: $(g_0', h_0') \sim \mathcal{N}((g_0, h_0)_c, Cov)$.
3. Calculate the ratio:
$$r = \frac{\pi(g_0', h_0'|data)}{\pi((g_0, h_0)_c|data)}$$
4. Draw a sample $\theta \sim U(0, 1)$.
5. The chain then moves according to:
$$(g_0, h_0)_{t=1} = \begin{cases} (g_0', h_0') & \text{if} \quad \theta < r, \\ (g_0, h_0)_c & \textit{otherwise}. \end{cases}$$
6. Repeat the loop.

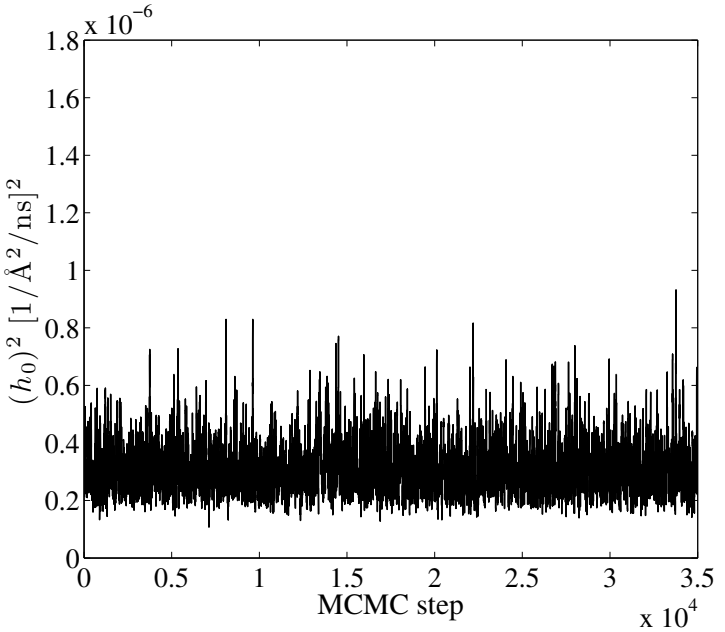


Regression: Results

- Results of MCMC is a “chain” of samples for each unknown:



. . .



. . .

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

Which order of the regression function?

- Run the inference using constant, linear, quadratic and cubic PCe $M(\xi_1, \xi_2)$.
- Which order is the most appropriate?
- Bayes factor: non-dimensional number that allows one to compare and discriminate between two "models" describing a given set of data.
- $\theta_p = \{g_0, \ldots, g_P, h_0, h_1, h_2\}$: parameter vector using a p-th order PCe $M(\xi)$.
- Given θ_{p_1} and θ_{p_2}, associated with regression functions of order p_1 and p_2, respectively, the (ln) of Bayes factor, $B(\theta_{p_1}, \theta_{p_2})$, is given by:

$$\ln(B(\theta_{p_1}, \theta_{p_2})) = \ln \frac{\int_{\Omega_1} \pi\left(\theta_{p_1} | \mathbf{G}\right) \, d\theta_{p_1}}{\int_{\Omega_2} \pi\left(\theta_{p_2} | \mathbf{G}\right) \, d\theta_{p_2}}$$

Which order of the regression function?

- Run the inference using constant, linear, quadratic and cubic PCe $M(\xi_1, \xi_2)$.
- Which order is the most appropriate?
- Bayes factor: non-dimensional number that allows one to compare and discriminate between two "models" describing a given set of data.
- $\boldsymbol{\theta}_p = \{g_0, \ldots, g_P, h_0, h_1, h_2\}$: parameter vector using a p-th order PCe $M(\boldsymbol{\xi})$.
- Given $\boldsymbol{\theta}_{p_1}$ and $\boldsymbol{\theta}_{p_2}$, associated with regression functions of order p_1 and p_2, respectively, the (ln) of Bayes factor, $B(\boldsymbol{\theta}_{p_1}, \boldsymbol{\theta}_{p_2})$, is given by:

$$\ln(B(\boldsymbol{\theta}_{p_1}, \boldsymbol{\theta}_{p_2})) = \ln \frac{\int_{\Omega_1} \pi\left(\boldsymbol{\theta}_{p_1} \middle| \mathbf{G}\right) \, d\boldsymbol{\theta}_{p_1}}{\int_{\Omega_2} \pi\left(\boldsymbol{\theta}_{p_2} \middle| \mathbf{G}\right) \, d\boldsymbol{\theta}_{p_2}}$$

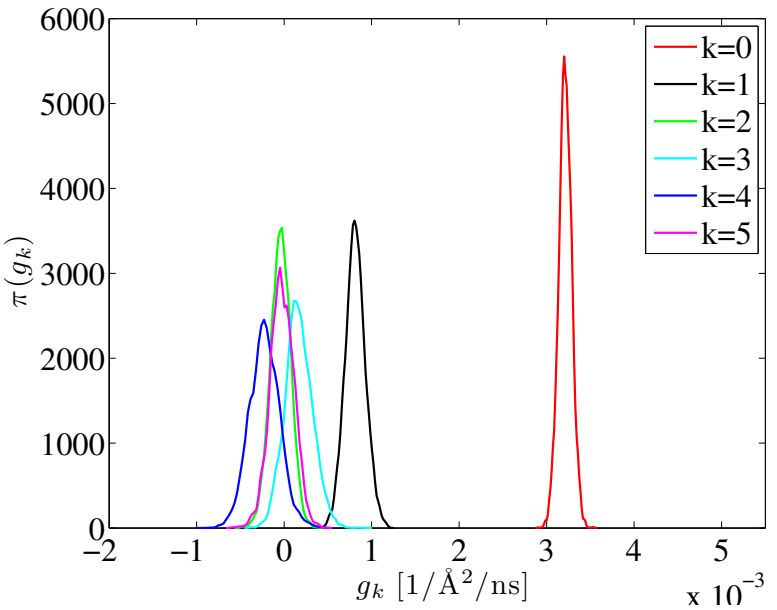
	Na^+				Cl^-			
	$p_2 = 0$	$p_2 = 1$	$p_2 = 2$	$p_2 = 3$	$p_2 = 0$	$p_2 = 1$	$p_2 = 2$	$p_2 = 3$
$p_1 = 0$	–	-19.341	-19.933	-16.285	–	**2.484**	**2.737**	**6.499**
$p_1 = 1$	19.341	–	-0.593	3.055	-2.284	–	0.254	4.015
$p_1 = 2$	**19.933**	**0.593**	–	**3.648**	-2.737	-0.254	–	3.761
$p_1 = 3$	16.285	-3.055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

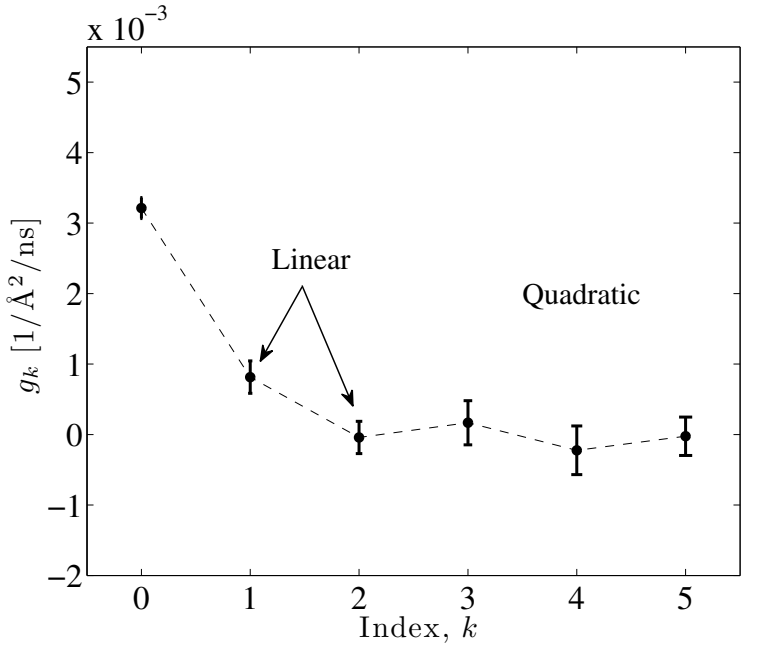
Posterior Uncertainty & Response Surface for Na^+

- Quadratic ($p = 2$) PCe, $M(\xi_1, \xi_2)$, for G_{Na^+}:

$$M = g_0 + g_1\Psi_1 + \ldots + g_5\Psi_5$$

- "Information" is mainly contained in g_0, g_1.
- Bayesian regression $\Rightarrow$ *uncertain* PCe.
- Generate $\{\mathbf{g}_i\}_{i=1}^5$ by sampling $\pi(g_0, ..., g_5)$ and plot response surfaces: clear trend present.



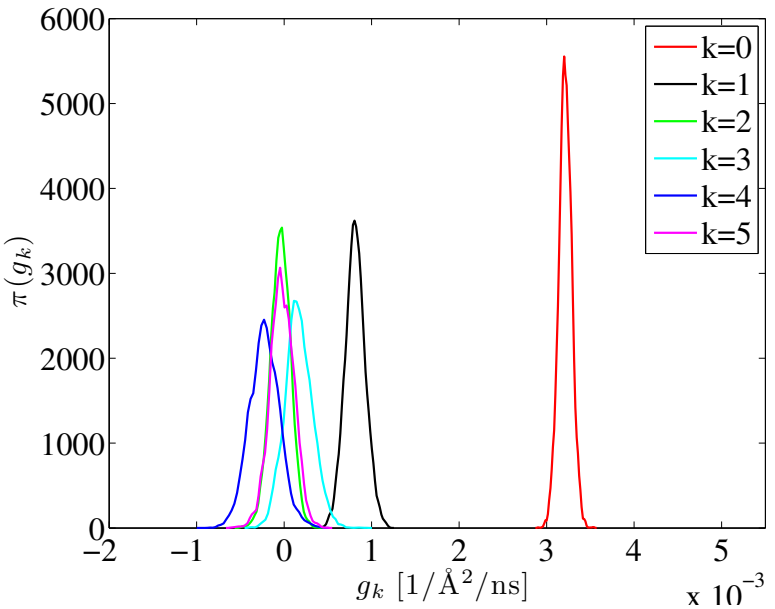


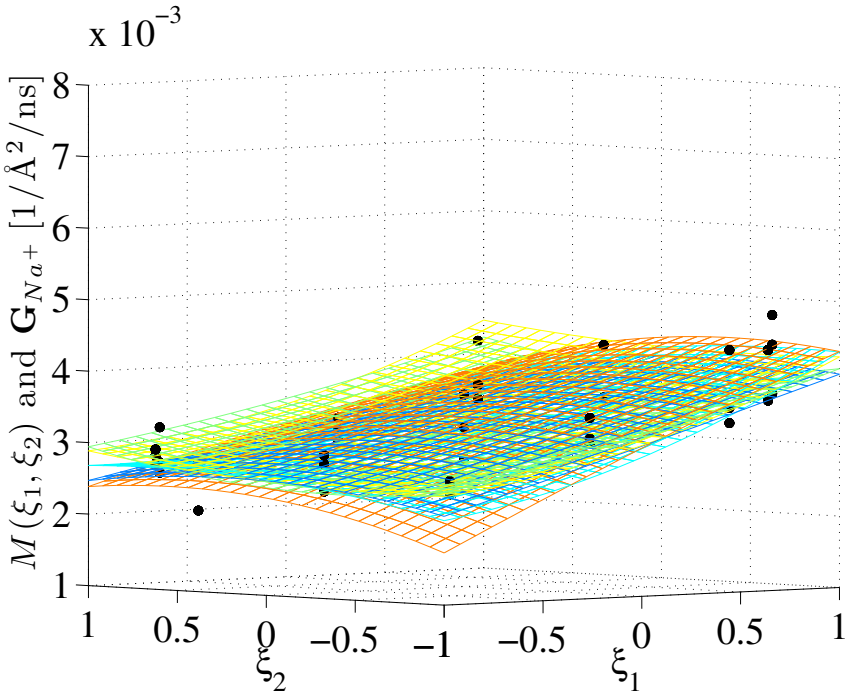
Posterior Uncertainty & Response Surface for Na^+

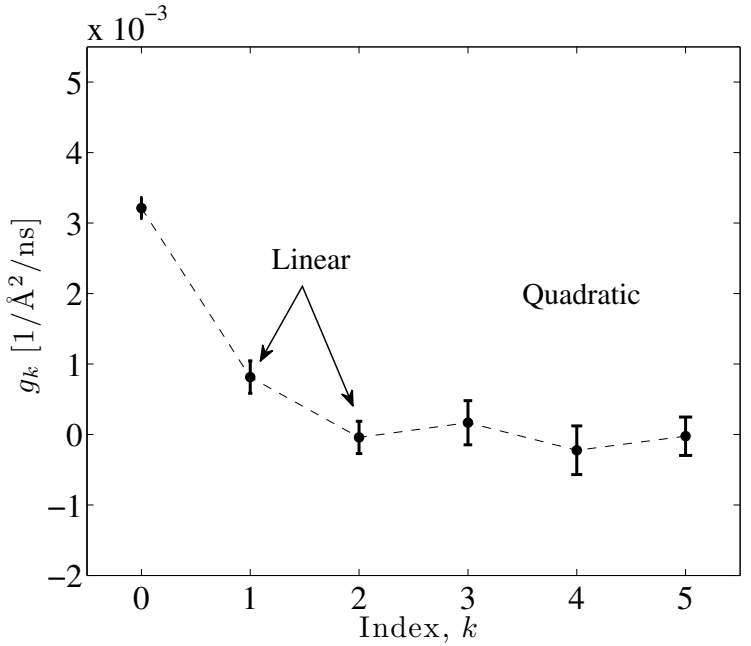
- Quadratic ($p = 2$) PCe, $M(\xi_1, \xi_2)$, for G_{Na^+}:

$$M = g_0 + g_1\Psi_1 + \ldots + g_5\Psi_5$$

- "Information" is mainly contained in g_0, g_1.
- Bayesian regression $\Rightarrow$ *uncertain* PCe.
- Generate $\{\mathbf{g}_i\}_{i=1}^5$ by sampling $\pi(g_0, ..., g_5)$ and plot response surfaces: clear trend present.

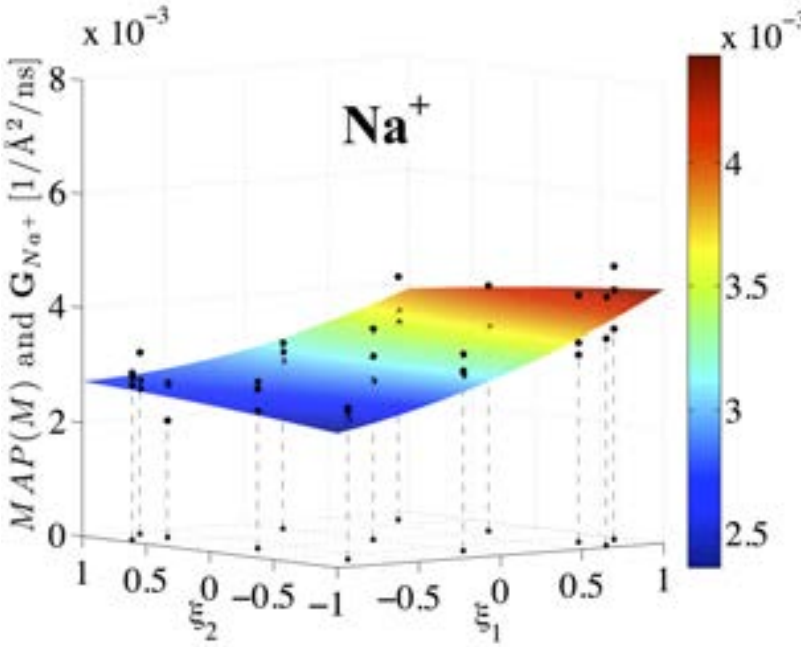


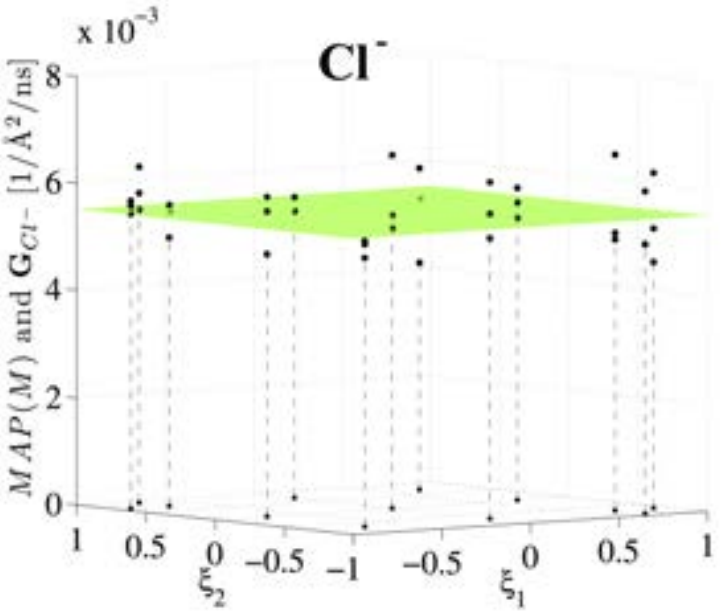




Posterior Uncertainty & Response Surface

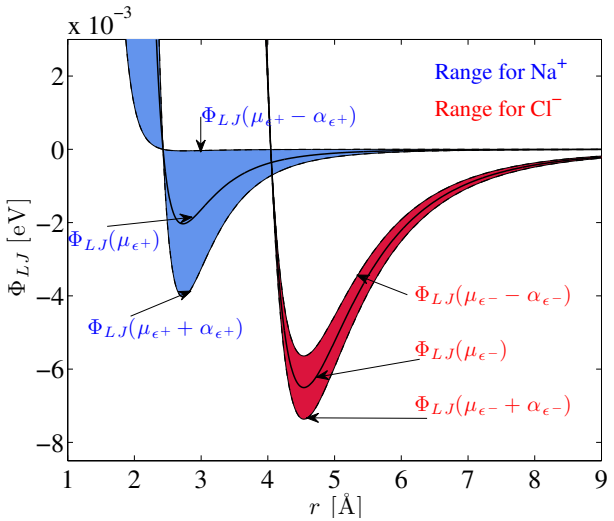
- MAP estimate of response surface for Na^+: G_{Na^+} increases as ε_{Na^+} (i.e. ξ_1) increases.
- For Cl^- Bayes factor suggested to use a *constant M* to represent G_{Cl^-}.
- Can we explain this difference?
- $G_{Cl^-} \sim \varepsilon_{Cl^-}$ similarly to $G_{Na^+} \sim \varepsilon_{Na^+}$ but:
  - smaller range of uncertainty chosen for ε_{Cl^-} gives a smaller absolute variation.
  - trend of G_{Cl^-} with respect to ε_{Cl^-} is obscured by the substantial noise level.

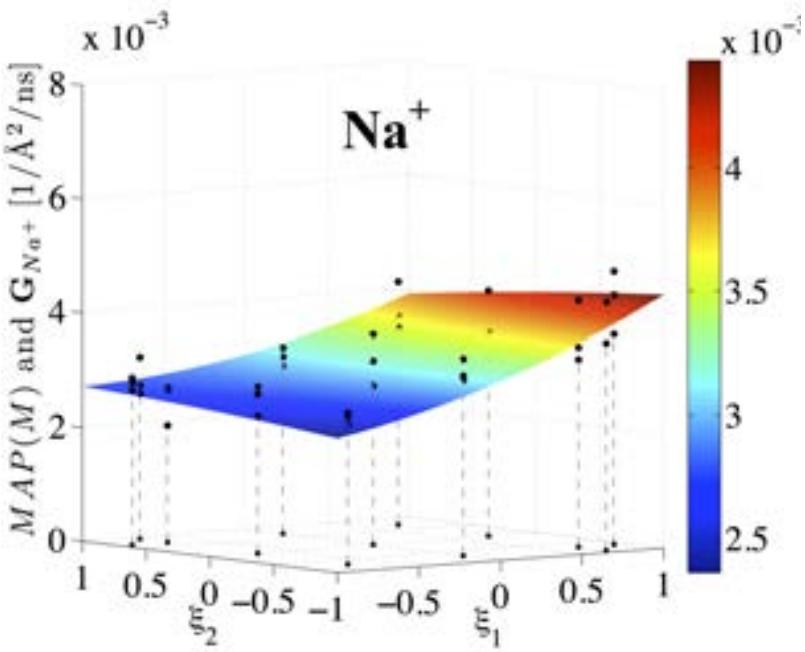


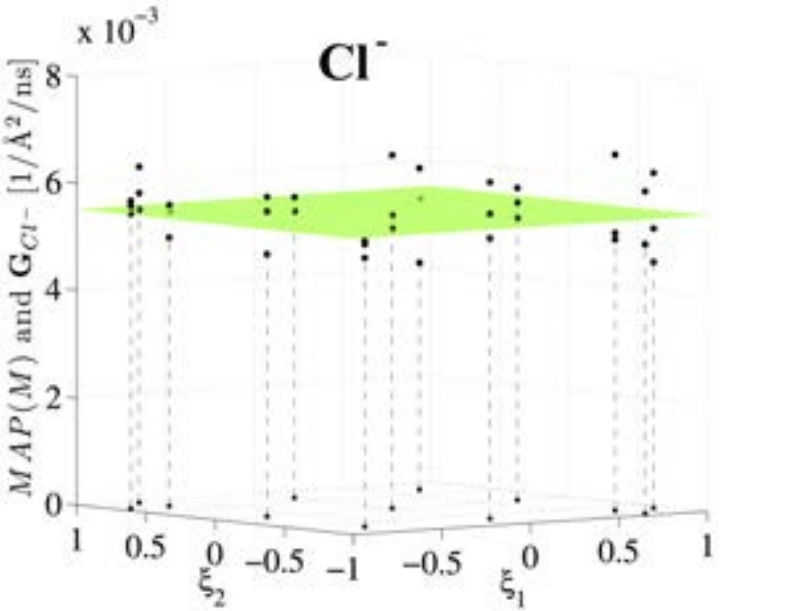


Posterior Uncertainty & Response Surface

- MAP estimate of response surface for Na^+: G_{Na^+} increases as ε_{Na^+} (i.e. ξ_1) increases.
- For Cl^- Bayes factor suggested to use a *constant M* to represent G_{Cl^-}.
- Can we explain this difference?
- $G_{Cl^-} \sim \varepsilon_{Cl^-}$ similarly to $G_{Na^+} \sim \varepsilon_{Na^+}$ but:
  - smaller range of uncertainty chosen for ε_{Cl^-} gives a smaller absolute variation.
  - trend of G_{Cl^-} with respect to ε_{Cl^-} is obscured by the substantial noise level.







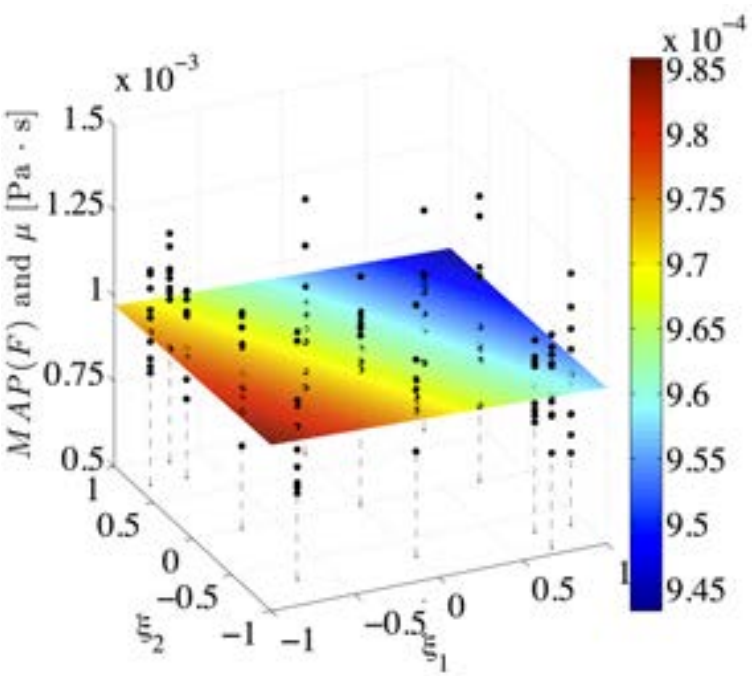
Transport Coefficients

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

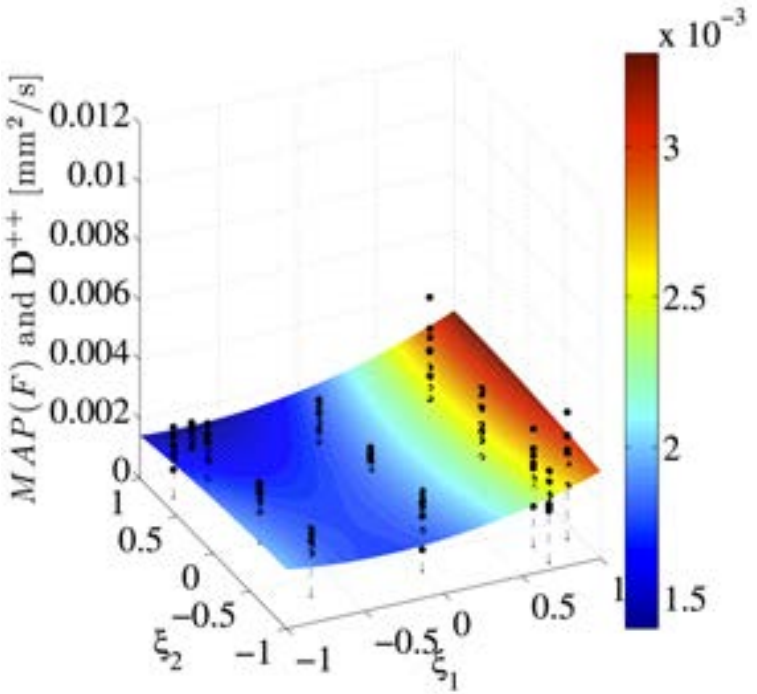
$$\mu = \frac{V}{3k_B T} \int_0^\infty \left\langle \varsigma(0) \cdot \varsigma(t) \right\rangle \mathrm{d}t \tag{2}$$

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 Ionic Conductance and Transport

- Let $F(\xi_1, \xi_2) = \mathbf{f}\,\boldsymbol{\Psi}$ be the PCe of one transport coefficient, μ or D^{++}.
- Let $M(\xi_1, \xi_2) = \mathbf{g}\,\boldsymbol{\Psi}$ be the PCe of the Na^+ conductance, G_{Na^+}.

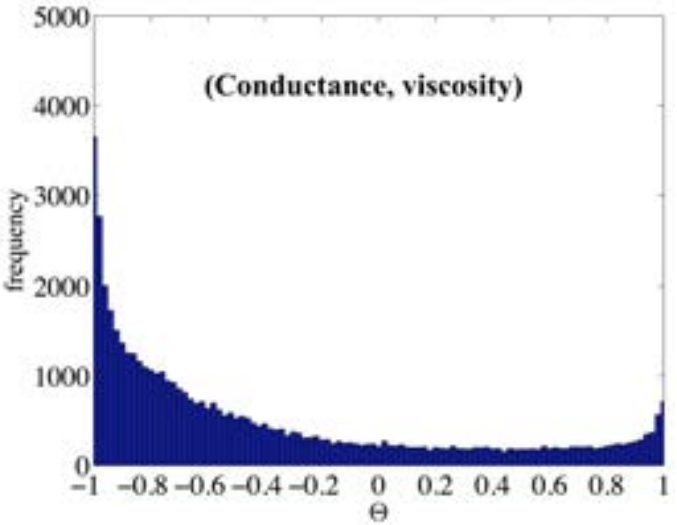
$$\begin{aligned} \mathrm{Cov}(M, F) &= \mathbb{E}\big[(M - \mathbb{E}[M])(F - \mathbb{E}[F])\big] \\ &= \sum_{k=1}^{min(P, P_F)} f_k g_k \, \mathbb{E}\big[\Psi_k^2(\boldsymbol{\xi})\big], \end{aligned}$$

- 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: ion flow decreases when the viscosity increases.
- Strong correlation between G_{Na^+} and D^{++}: convincing result, since we expect the flux of Na^+ to be mostly affected by the diffusivity of Na^+.

Correlations between Ionic Conductance and Transport

- Let $F(\xi_1, \xi_2) = \mathbf{f}\,\boldsymbol{\Psi}$ be the PCe of one transport coefficient, μ or D^{++}.
- Let $M(\xi_1, \xi_2) = \mathbf{g}\,\boldsymbol{\Psi}$ be the PCe of the Na^+ conductance, G_{Na^+}.

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

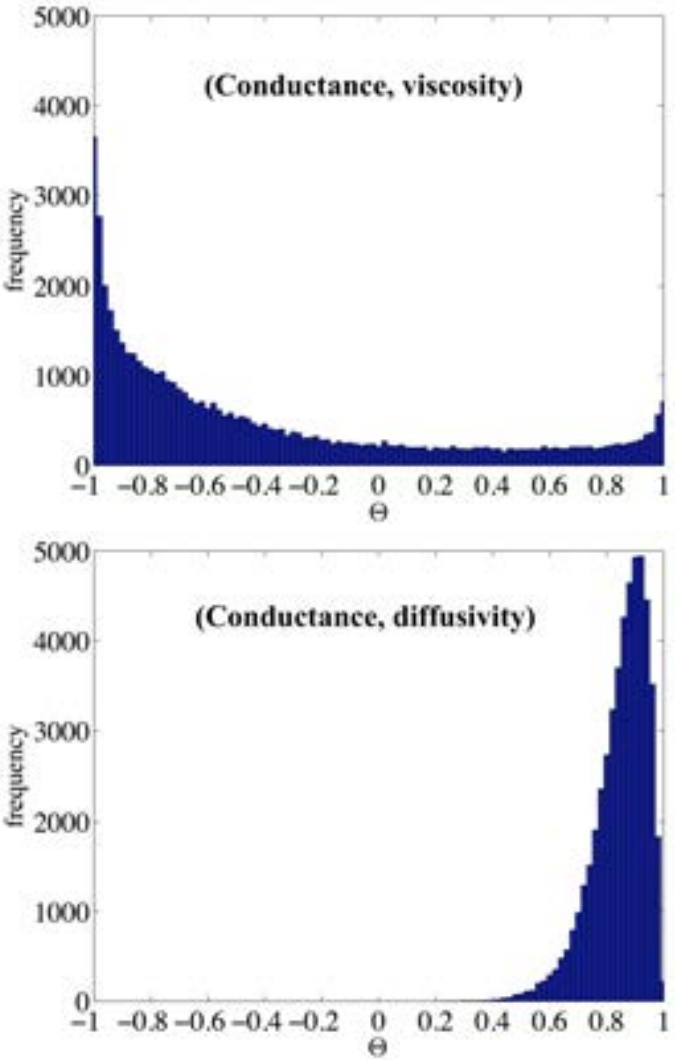


- Sample $\pi(\mathbf{g})$ and $\pi(\mathbf{f}) \Rightarrow \{\mathbf{g}_i\}_{i=1}^{50000}$ and $\{\mathbf{f}_i\}_{i=1}^{50000}$.
- Each $(\mathbf{g}_i, \mathbf{f}_i)$ gives one value of covariance.
- Plot histogram of correlation coefficient Θ.
- (G_{Na^+}, μ) correlation is mainly negative: ion flow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

Correlations between Ionic Conductance and Transport

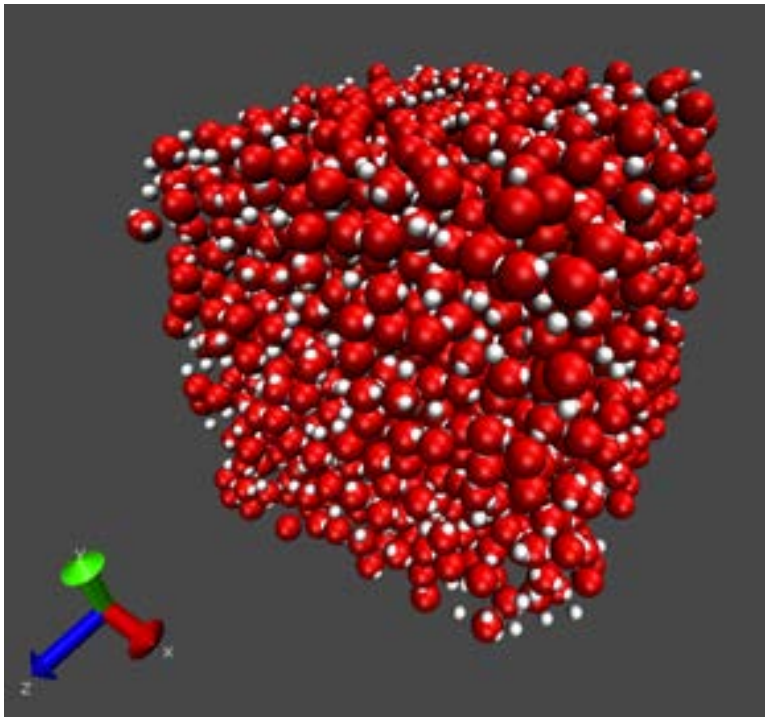
- Let $F(\xi_1, \xi_2) = \mathbf{f}\,\boldsymbol{\Psi}$ be the PCe of one transport coefficient, μ or D^{++}.
- Let $M(\xi_1, \xi_2) = \mathbf{g}\,\boldsymbol{\Psi}$ be the PCe of the Na^+ conductance, G_{Na^+}.

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



- 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: ion flow decreases when the viscosity increases.
- Strong correlation between G_{Na^+} and D^{++}: convincing result, since we expect the flux of Na^+ to be mostly affected by the diffusivity of Na^+.

Inverse problem for MD simulations of water

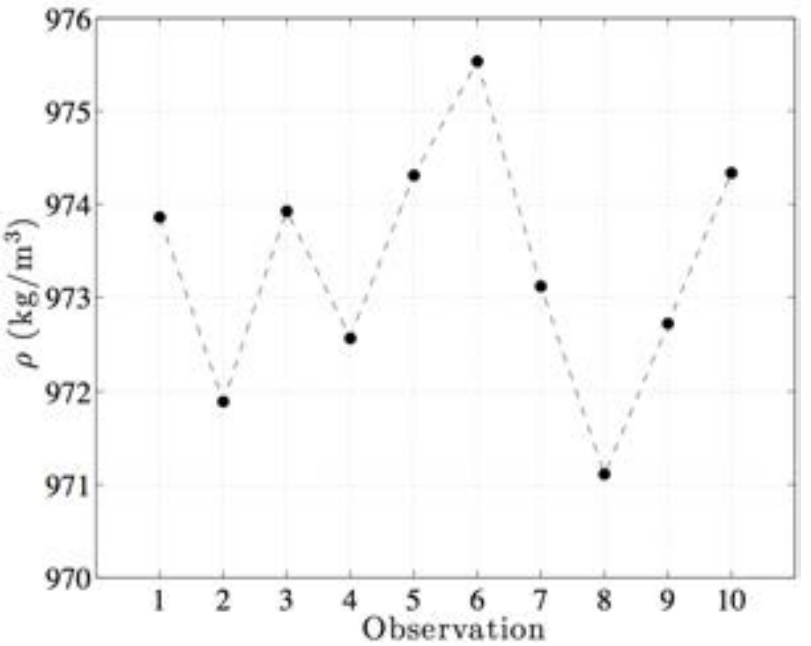


Problem Statement

- Focus on MD simulations of bulk water at ambient conditions: $T = 298$ K, $P = 1$ atm.
- Let $\{\alpha_1, \alpha_2, \alpha_3\}$ be three potential parameters of interest.
- Objective: given data of one or more macroscale observables, infer $\{\alpha_1, \alpha_2, \alpha_3\}$.
- Test case: synthetic problem based on presumed "true" values of the parameters:

$$\begin{aligned} \hat{\alpha}_1 &= 0.17, \\ \hat{\alpha}_2 &= 3.15, \\ \hat{\alpha}_3 &= 0.14, \end{aligned}$$

- Run multiple MD replicas to generate a set of noisy density observations $\boldsymbol{\rho} = \{\rho_i\}_{i=1}^{N=10}$.



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

Using the data $\{\rho_i\}_{i=1}^{N=10}$ how well can we recover the "true" parameters?

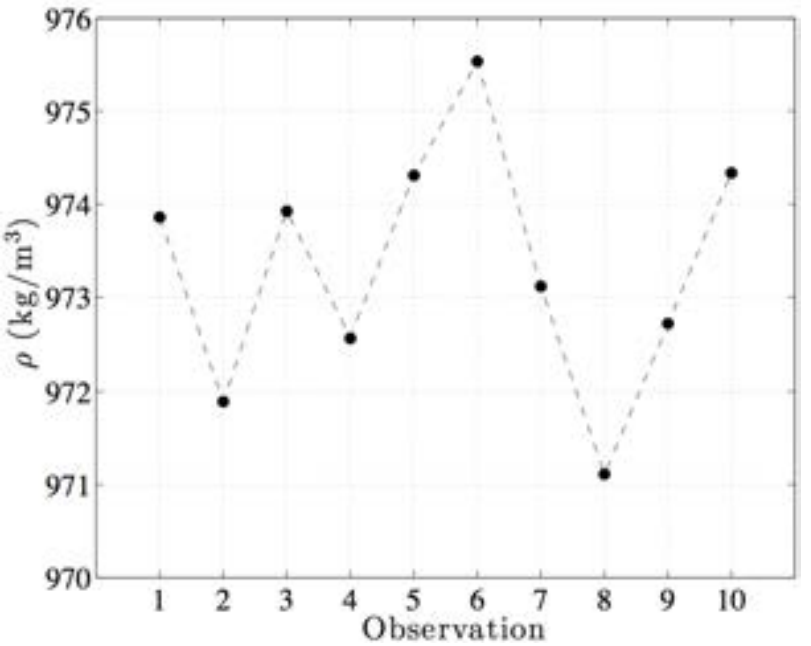
Problem Statement

- Focus on MD simulations of bulk water at ambient conditions: $T = 298$ K, $P = 1$ atm.
- Let $\{\alpha_1, \alpha_2, \alpha_3\}$ be three potential parameters of interest.
- Objective: given data of one or more macroscale observables, infer $\{\alpha_1, \alpha_2, \alpha_3\}$.

- Test case: synthetic problem based on presumed "true" values of the parameters:

$$\begin{aligned} \hat{\alpha}_1 &= 0.17, \\ \hat{\alpha}_2 &= 3.15, \\ \hat{\alpha}_3 &= 0.14, \end{aligned}$$

- Run multiple MD replicas to generate a set of noisy density observations $\boldsymbol{\rho} = \{\rho_i\}_{i=1}^{N=10}$.



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

Using the data $\{\rho_i\}_{i=1}^{N=10}$ how well can we recover the "true" parameters?

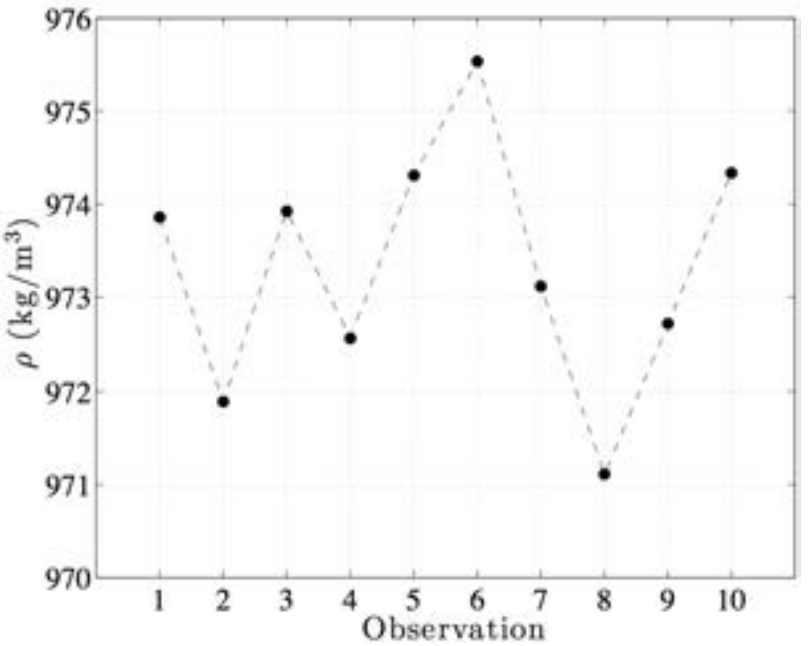
Problem Statement

- Focus on MD simulations of bulk water at ambient conditions: $T = 298$ K, $P = 1$ atm.
- Let $\{\alpha_1, \alpha_2, \alpha_3\}$ be three potential parameters of interest.
- Objective: given data of one or more macroscale observables, infer $\{\alpha_1, \alpha_2, \alpha_3\}$.

- Test case: synthetic problem based on presumed "true" values of the parameters:

$$\begin{aligned} \hat{\alpha}_1 &= 0.17, \\ \hat{\alpha}_2 &= 3.15, \\ \hat{\alpha}_3 &= 0.14, \end{aligned}$$

- Run multiple MD replicas to generate a set of noisy density observations $\boldsymbol{\rho} = \{\rho_i\}_{i=1}^{N=10}$.



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

Using the data $\{\rho_i\}_{i=1}^{N=10}$ how well can we recover the "true" parameters?

"Expensive" Inference

- Bayes' theorem: $\underbrace{\pi\left(\alpha_1, \alpha_2, \alpha_3 \mid \boldsymbol{\rho}\right)}_{\textit{Posterior}} \propto \underbrace{\mathcal{L}\left(\boldsymbol{\rho} \mid \alpha_1, \alpha_2, \alpha_3\right)}_{\textit{Likelihood}} \underbrace{\mathcal{P}(\alpha_1, \alpha_2, \alpha_3)}_{\textit{Prior}}$
- A **direct** (expensive) likelihood:

$$\rho_i = \mathcal{F}(\alpha_1, \alpha_2, \alpha_3) + \gamma_i \quad \Rightarrow \quad \mathcal{L}_{\mathcal{F}}\left(\boldsymbol{\rho} \,|\alpha_1, \alpha_2, \alpha_3\right) = \prod_{i=1}^{N} p_{\gamma}\left(\rho_i - \mathcal{F}(\alpha_1, \alpha_2, \alpha_3)\right)$$

  - $\mathcal{F}$ represents a full MD run
  - γ_i: RV capturing the discrepancy between data, ρ_i, and the MD prediction, $\mathcal{F}$.
- MCMC exploration of π requires $\sim 10^4$ evaluations of $\mathcal{F}$: prohibitive due to the large computational cost associated with a single MD computation.
- Replace the *full* MD prediction with a suitable *surrogate* model.
- A surrogate representation is a model relating the observables to the parameters such that:
  - the accuracy of the representation is comparable to the high fidelity system
  - the evaluation cost is considerably reduced

"Expensive" Inference

- Bayes' theorem: $\underbrace{\pi\left(\alpha_1, \alpha_2, \alpha_3 \mid \boldsymbol{\rho}\right)}_{\textit{Posterior}} \propto \underbrace{\mathcal{L}\left(\boldsymbol{\rho} \mid \alpha_1, \alpha_2, \alpha_3\right)}_{\textit{Likelihood}} \underbrace{\mathcal{P}(\alpha_1, \alpha_2, \alpha_3)}_{\textit{Prior}}$
- A **direct** (expensive) likelihood:

$$\rho_i = \mathcal{F}(\alpha_1, \alpha_2, \alpha_3) + \gamma_i \quad \Rightarrow \quad \mathcal{L}_{\mathcal{F}}\left(\boldsymbol{\rho} \,|\alpha_1, \alpha_2, \alpha_3\right) = \prod_{i=1}^{N} p_{\gamma}\left(\rho_i - \mathcal{F}(\alpha_1, \alpha_2, \alpha_3)\right)$$

  - $\mathcal{F}$ represents a full MD run
  - γ_i: RV capturing the discrepancy between data, ρ_i, and the MD prediction, $\mathcal{F}$.
- MCMC exploration of π requires $\sim 10^4$ evaluations of $\mathcal{F}$: prohibitive due to the large computational cost associated with a single MD computation.
- Replace the *full* MD prediction with a suitable *surrogate* model.
- A surrogate representation is a model relating the observables to the parameters such that:
  - the accuracy of the representation is comparable to the high fidelity system
  - the evaluation cost is considerably reduced

"Cheap" Inference

- Following Marzouk *et al.* (2007), we can use a PC expansion of density as a surrogate model:

$$M_\rho(\alpha_1, \alpha_2, \alpha_3) = \sum_{k=0}^{P} c_k \tilde{\Psi}_k(\alpha_1, \alpha_2, \alpha_3)$$

- Given $\alpha_{1,i}, \alpha_{2,i}, \alpha_{3,i}$, evaluating $M_\rho(\alpha_{1,i}, \alpha_{2,i}, \alpha_{3,i})$ yields a corresponding prediction for the target observable, density.
- $M_\rho(\alpha_1, \alpha_2, \alpha_3)$ is a surrogate representation of the "expensive" MD run $\mathcal{F}$.
- The "cheap" likelihood becomes

$$\rho_i = M_\rho(\alpha_1, \alpha_2, \alpha_3) + \gamma_i \quad \Rightarrow \quad \mathcal{L}_M\left(\boldsymbol{\rho} \,\middle|\, \alpha_1, \alpha_2, \alpha_3\right) = \prod_{i=1}^{N} p_\gamma\left(\rho_i - M_\rho(\alpha_1, \alpha_2, \alpha_3)\right)$$

- M_ρ is a *polynomial*: substantial improvement in the computational efficiency.
- Problem reduces to sampling the posterior:

$$\pi(\alpha_1, \alpha_2, \alpha_3 \mid \boldsymbol{\rho}) \propto \mathcal{L}_M\left(\boldsymbol{\rho} \,\middle|\, \{\alpha_1, \alpha_2, \alpha_3\}\right) \mathcal{P}(\alpha_1, \alpha_2, \alpha_3)$$

"Cheap" Inference

- Following Marzouk *et al.* (2007), we can use a PC expansion of density as a surrogate model:

$$M_\rho(\alpha_1, \alpha_2, \alpha_3) = \sum_{k=0}^{P} c_k \tilde{\Psi}_k(\alpha_1, \alpha_2, \alpha_3)$$

- Given $\alpha_{1,i}, \alpha_{2,i}, \alpha_{3,i}$, evaluating $M_\rho(\alpha_{1,i}, \alpha_{2,i}, \alpha_{3,i})$ yields a corresponding prediction for the target observable, density.
- $M_\rho(\alpha_1, \alpha_2, \alpha_3)$ is a surrogate representation of the "expensive" MD run $\mathcal{F}$.
- The "cheap" likelihood becomes

$$\rho_i = M_\rho(\alpha_1, \alpha_2, \alpha_3) + \gamma_i \quad \Rightarrow \quad \mathcal{L}_M\left(\boldsymbol{\rho} \,\middle|\, \alpha_1, \alpha_2, \alpha_3\right) = \prod_{i=1}^{N} p_\gamma\left(\rho_i - M_\rho(\alpha_1, \alpha_2, \alpha_3)\right)$$

- M_ρ is a *polynomial*: substantial improvement in the computational efficiency.
- Problem reduces to sampling the posterior:

$$\pi(\alpha_1, \alpha_2, \alpha_3 \mid \boldsymbol{\rho}) \propto \mathcal{L}_M\left(\boldsymbol{\rho} \,\middle|\, \{\alpha_1, \alpha_2, \alpha_3\}\right) \mathcal{P}(\alpha_1, \alpha_2, \alpha_3)$$

Noise Model

- A set of observations, $\boldsymbol{\rho} = \{\rho_i\}_{i=1}^{N}$, for density to use for the inference.
- A surrogate (i.e. cheap) model directly relating the observable to the parameters:

$$M_\rho(\alpha_1, \alpha_2, \alpha_3) = \sum_{k=0}^{P} c_k \tilde{\Psi}_k(\alpha_1, \alpha_2, \alpha_3) = c_0 + c_1\alpha_1 + c_2\alpha_2 + c_3\alpha_3 + \dots$$

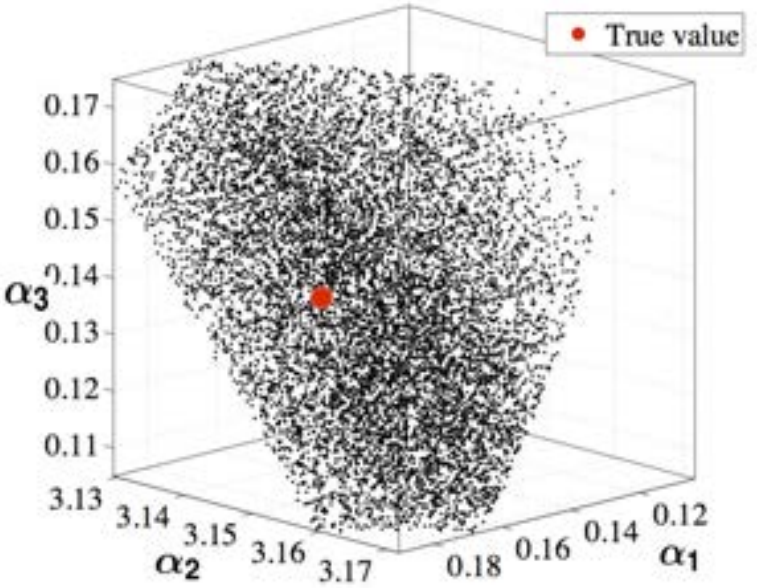
  where $(c_0, c_1, \dots)$ are *deterministic* coefficients.
- Gaussian noise model with noise variance σ^2 as hyperparameter yields:

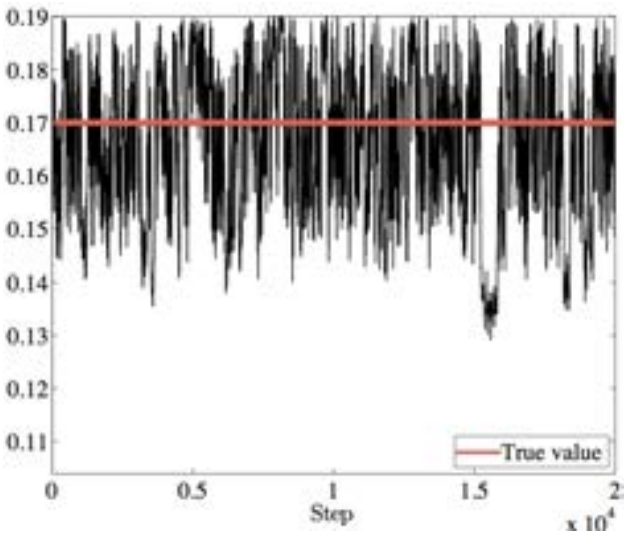
$$\pi(\alpha_1, \alpha_2, \alpha_3, \sigma^2 \mid \boldsymbol{\rho}) \propto \underbrace{\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{[\rho_i - M_\rho(\alpha_1, \alpha_2, \alpha_3)]^2}{2\sigma^2}\right)}_{\textit{Likelihood}} \underbrace{\mathcal{P}(\alpha_1, \alpha_2, \alpha_3, \sigma^2)}_{\textit{Prior}}$$

  - where σ^2 is the hyperparameter
  - priors for $\alpha_1, \alpha_2, \alpha_3$ are uniform distributions, prior for σ^2 is: $\mathcal{P}(\sigma^2) = 1/\sigma^2$.
- Sample the posterior using a MCMC method: it involves a random walk in the $(\alpha_1, \alpha_2, \alpha_3, \sigma^2)$-space.

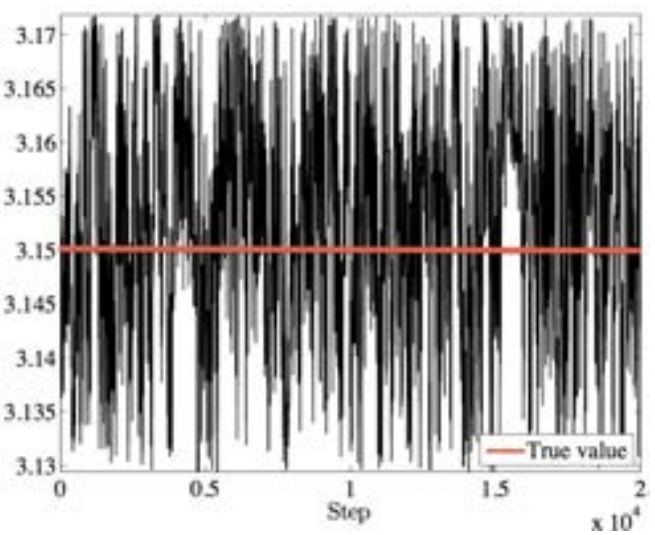
Single Observable: Density

- Inference is performed using a cubic PC expansion as surrogate model.
- Plots of 20000 MCMC samples.
- Underdetermined problem, yielding large posterior uncertainties.
- Posterior densities are nearly uniform.

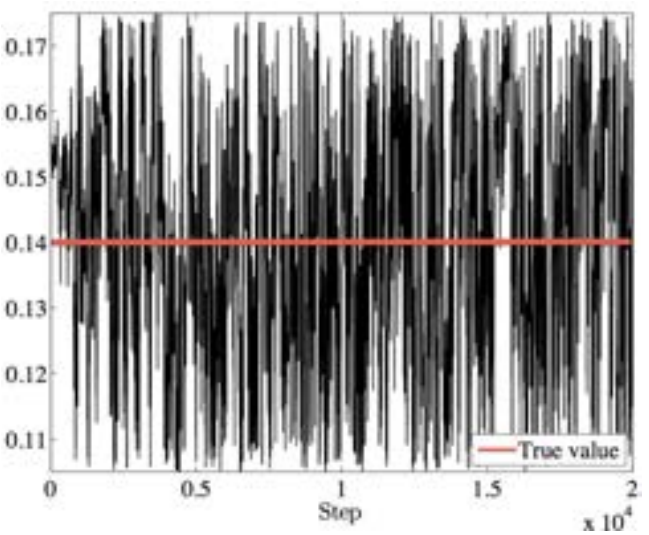




α_1



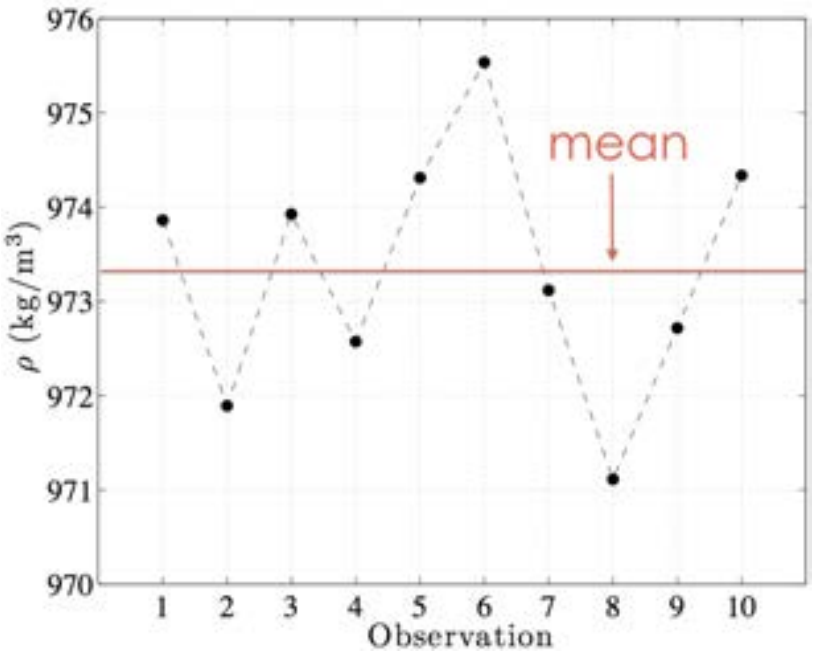
α_2

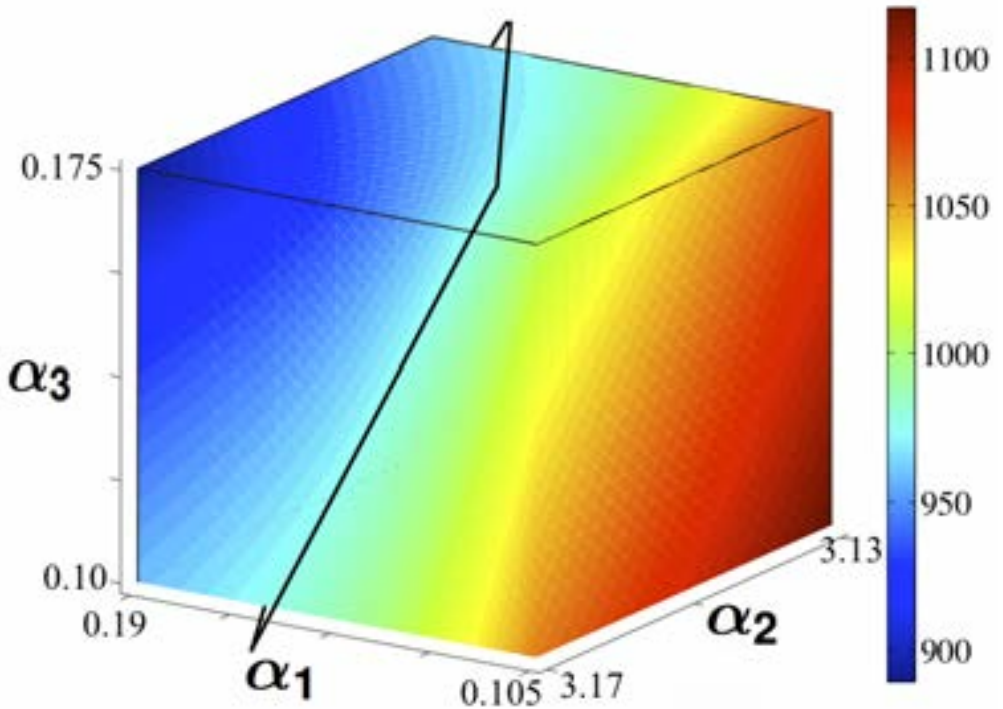


α_3

Single Observable: Density

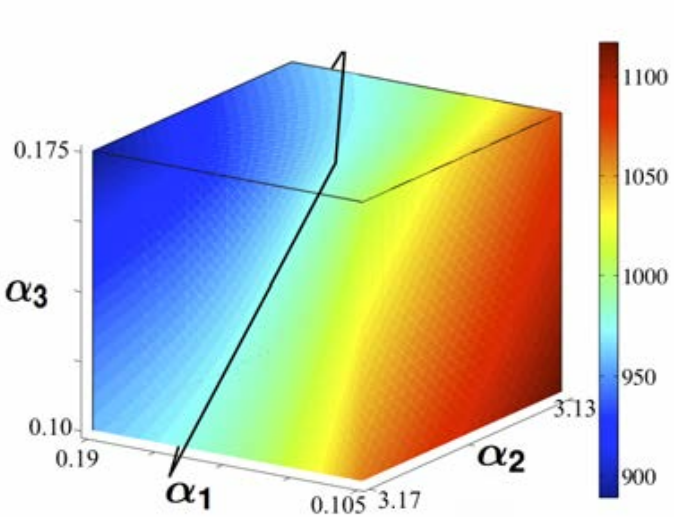
- Result can be explained by analyzing the structure of the surrogate model.
- Compute the mean of the observations of density $\overline{\rho}$.
- Extract from the surrogate model, $M(\alpha_1, \alpha_2, \alpha_3)$, the isosurface connecting points such that: $M(\alpha_1, \alpha_2, \alpha_3) = \overline{\rho}$.

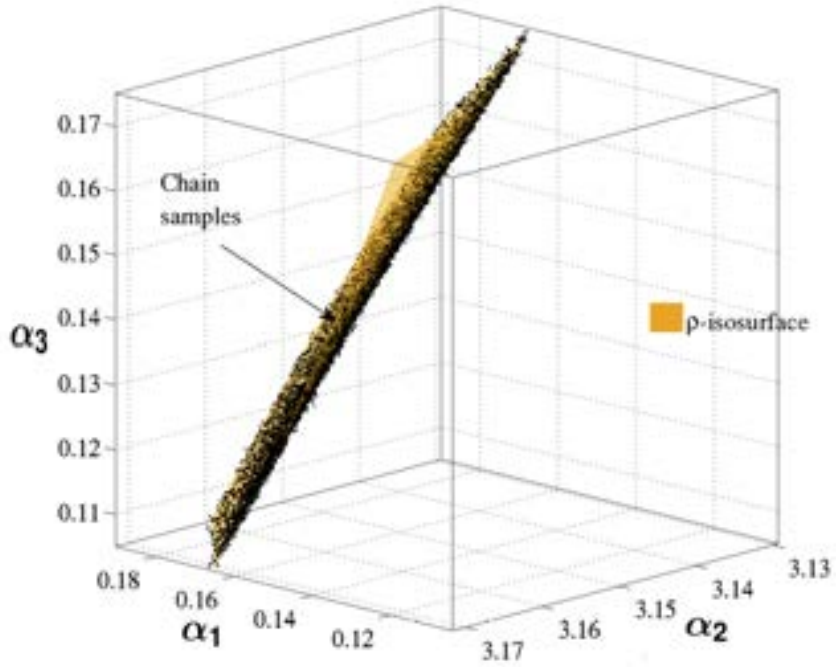


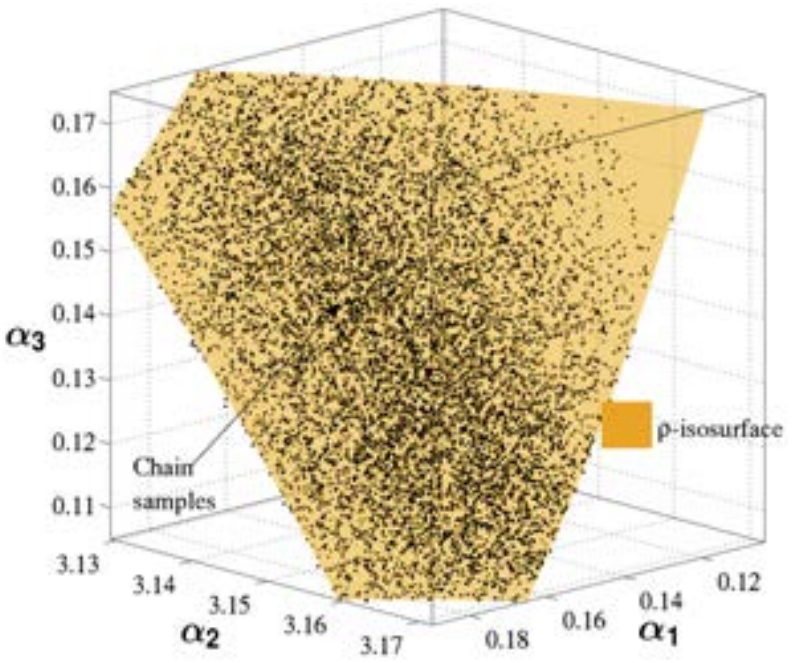


Single Observable: Density

- Superimposing the chain to the isosurface reveals the overlapping.
- The chain is constrained by the structure of the surrogate model.
- Suggests that using *one observable* to infer *three parameters* leaves two additional degrees of freedom missing.

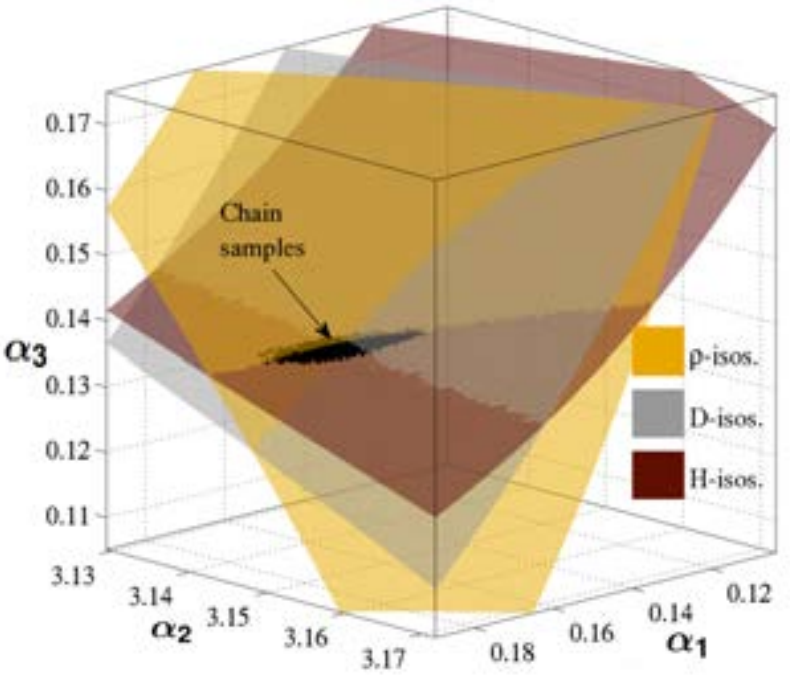


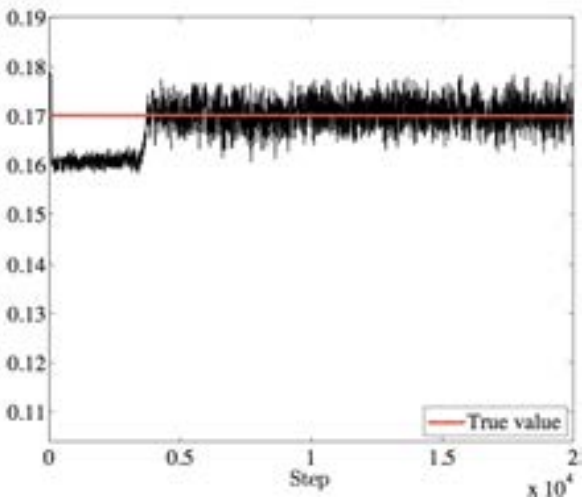




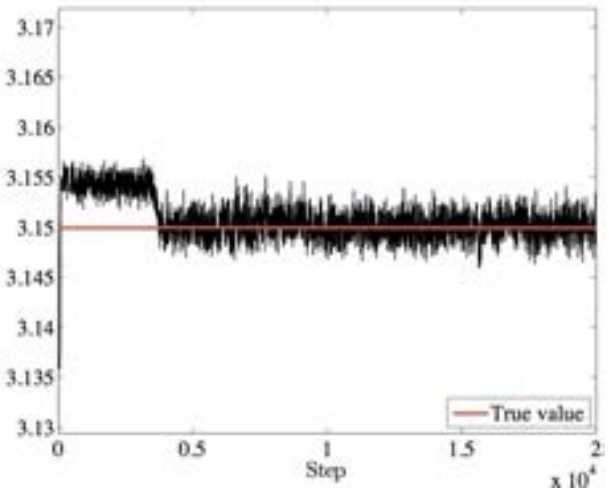
Three Observables: Density, Self-diffusion and Enthalpy

- Run inference using data of two more observables, i.e. $\{\rho_i, D_i, H_i\}_{i=1}^{10}$.
- Surfaces of constant density, self-diffusivity, and enthalpy intersect in a point, leading to a well-defined problem.
- Chain localizes at the intersection of the isosurfaces extracted from the PC surrogate of each observable.
- True parameters are recovered with excellent accuracy.

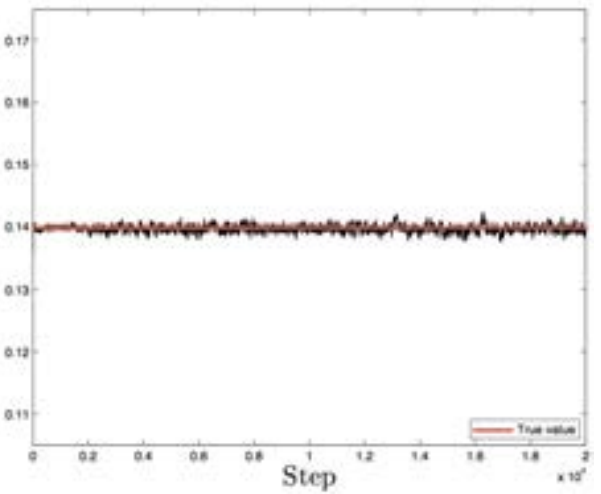




α_1



α_2



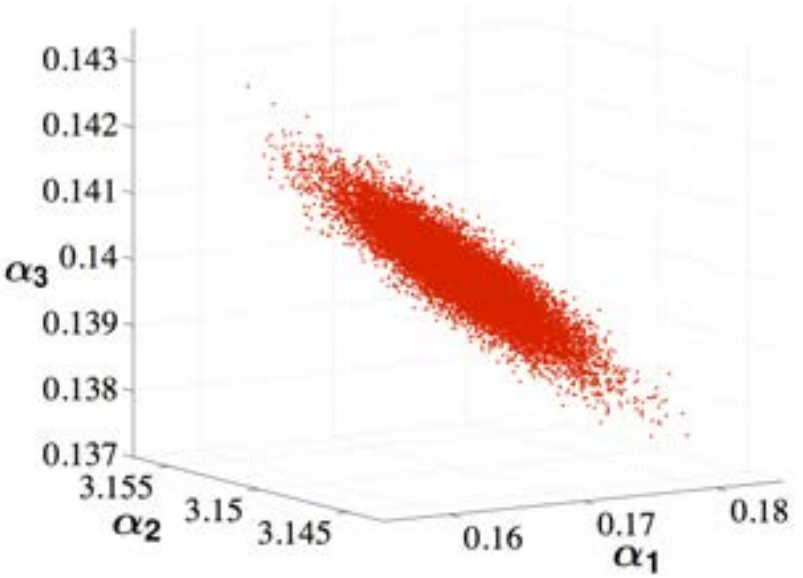
α_3

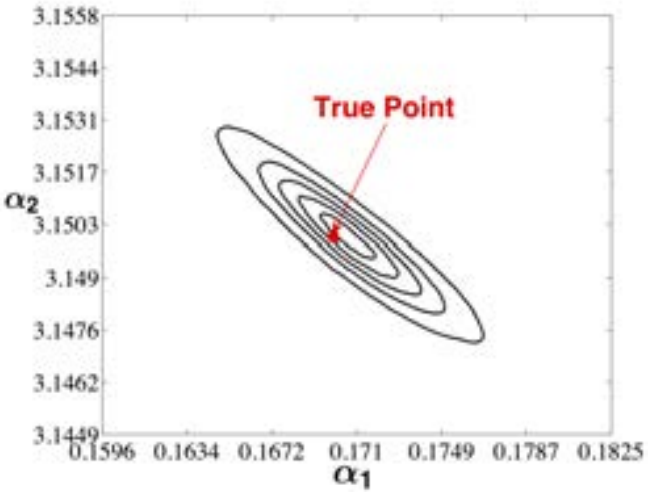
Posterior Correlations

- 3D-joint posterior based on the MCMC samples.
- Advantage of Bayesian approach:

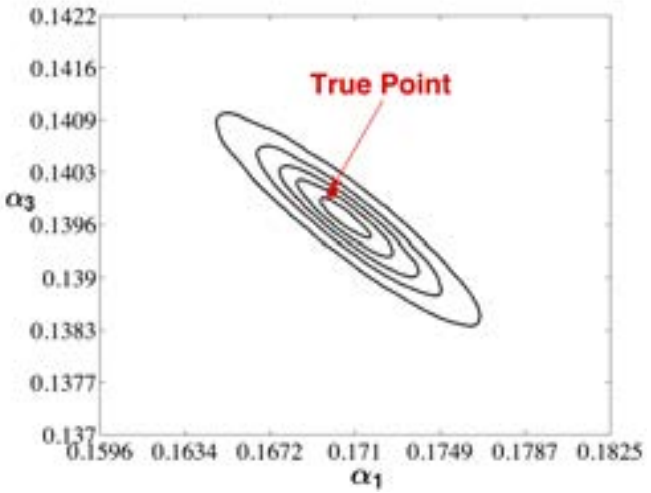
  (a) joint PDFs for the potential parameters, whose spread stem from the noise in the data.

  (b) correlation structure, a priori unknown because originally assumed independent parameters.

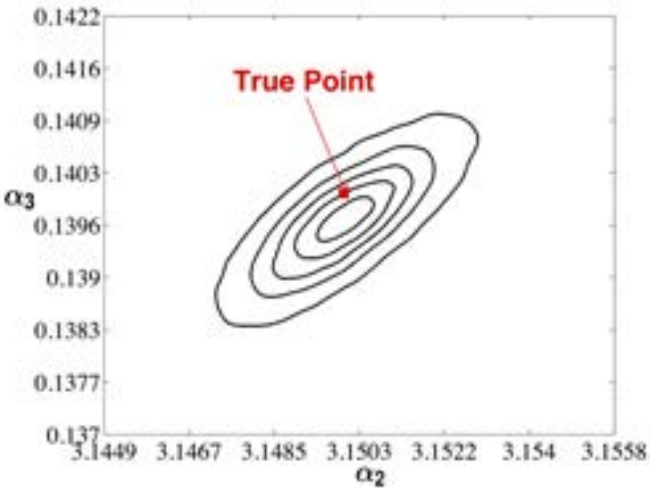




$\pi(\alpha_1, \alpha_2)$



$\pi(\alpha_1, \alpha_3)$



$\pi(\alpha_2, \alpha_3)$

Inverse Problem based on Non-Deterministic Surrogate

- Non-deterministic surrogate:

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

  where the set of PC coefficients $\{c_l\}_{l=0}^{P}$, is a random vector defined by a $(P+1)$-dimensional joint probability density.

- The surrogate model prediction, $M(\alpha_1, \alpha_2, \alpha_3)$, depends on:
  1. the parametric uncertainty through the uncertain parameters $\alpha_1, \alpha_2, \alpha_3$.
  2. the uncertainty in the PC coefficients.

- Interpretation:
  - Draw m samples of the parameters $\left\{\{\alpha_1, \alpha_2, \alpha_3\}^{(j)}\right\}_{j=1}^{m}$
  - For any given $\{\alpha_1, \alpha_2, \alpha_3\}^{(j)}$, we can draw n different sample-spectra of PC coefficients $\left\{\mathbf{c}_i\right\}_{i=1}^{n}$, from their joint distribution.
  - We thus obtain $n \times m$ predictions for the target observable $\left\{(M)_{i,j}\right\}_{i,j=1}^{n,m}$.

- In other words, each realization of the parameters $(\{\alpha_1, \alpha_2, \alpha_3\}^{(j)}$, due to the uncertainty in the coefficients, can be associated with an arbitrary number of predictions of the observable M.

Inverse Problem based on Non-Deterministic Surrogate

- Non-deterministic surrogate:

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

  where the set of PC coefficients $\{c_l\}_{l=0}^{P}$, is a random vector defined by a $(P+1)$-dimensional joint probability density.

- The surrogate model prediction, $M(\alpha_1, \alpha_2, \alpha_3)$, depends on:
  1. the parametric uncertainty through the uncertain parameters $\alpha_1, \alpha_2, \alpha_3$.
  2. the uncertainty in the PC coefficients.

- Interpretation:
  - Draw m samples of the parameters $\left\{\{\alpha_1, \alpha_2, \alpha_3\}^{(j)}\right\}_{j=1}^{m}$
  - For any given $\{\alpha_1, \alpha_2, \alpha_3\}^{(j)}$, we can draw n different sample-spectra of PC coefficients $\left\{\mathbf{c}_i\right\}_{i=1}^{n}$, from their joint distribution.
  - We thus obtain $n \times m$ predictions for the target observable $\left\{(M)_{i,j}\right\}_{i,j=1}^{n,m}$.

- In other words, each realization of the parameters $(\{\alpha_1, \alpha_2, \alpha_3\}^{(j)}$, due to the uncertainty in the coefficients, can be associated with an arbitrary number of predictions of the observable M.

Inverse Problem based on Non-Deterministic Surrogate

- Non-deterministic surrogate:

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

  where the set of PC coefficients $\{c_l\}_{l=0}^{P}$, is a random vector defined by a $(P+1)$-dimensional joint probability density.

- The surrogate model prediction, $M(\alpha_1, \alpha_2, \alpha_3)$, depends on:
  1. the parametric uncertainty through the uncertain parameters $\alpha_1, \alpha_2, \alpha_3$.
  2. the uncertainty in the PC coefficients.

- Interpretation:
  - Draw m samples of the parameters $\left\{\{\alpha_1, \alpha_2, \alpha_3\}^{(j)}\right\}_{j=1}^{m}$
  - For any given $\{\alpha_1, \alpha_2, \alpha_3\}^{(j)}$, we can draw n different sample-spectra of PC coefficients $\left\{\mathbf{c}_i\right\}_{i=1}^{n}$, from their joint distribution.
  - We thus obtain $n \times m$ predictions for the target observable $\left\{(M)_{i,j}\right\}_{i,j=1}^{n,m}$.

- In other words, each realization of the parameters $(\{\alpha_1, \alpha_2, \alpha_3\}^{(j)}$, due to the uncertainty in the coefficients, can be associated with an arbitrary number of predictions of the observable M.

Inverse Problem based on Non-Deterministic Surrogate

- The uncertainty in the coefficients may be an important information and should be taken into account in the inverse problem.
- Given a sample $\boldsymbol{\alpha} = \{\alpha_1, \alpha_2, \alpha_3\}^{(j)}$, we can construct the constant row vector

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

- We can interpret each **non-deterministic** PC representation, $M(\alpha^{(j)})$, as a **linear combination** of the random vector $\boldsymbol{c} = \{c_0, \ldots, c_P\}^T$, according to

$$M(\alpha^{(j)}) = c_0\Psi_0(\alpha^{(j)}) + c_1\Psi_1(\alpha^{(j)}) + \ldots + c_P\Psi_P(\alpha^{(j)}) = \boldsymbol{yc}.$$

- If the $(P+1)$-distribution of the random vector $\mathbf{c}$ can be approximated by a $\mathcal{MVN}(\mu, \mathbf{Z})$, then

$$\boldsymbol{yc} = \Psi_0(\alpha^{(j)})c_0 + \ldots + \Psi_P(\alpha^{(j)})c_P,$$

is distributed as a *univariate* gaussian with mean $(\boldsymbol{y}\mu)$ and variance $(\boldsymbol{yZy}^T)$, in short notation $\mathcal{N}((\boldsymbol{y}\mu), (\boldsymbol{yZy}^T))$.

Inverse Problem based on Non-Deterministic Surrogate

- The uncertainty in the coefficients may be an important information and should be taken into account in the inverse problem.
- Given a sample $\boldsymbol{\alpha} = \{\alpha_1, \alpha_2, \alpha_3\}^{(j)}$, we can construct the constant row vector

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

- We can interpret each **non-deterministic** PC representation, $M(\boldsymbol{\alpha}^{(j)})$, as a **linear combination** of the random vector $\boldsymbol{c} = \{c_0, \ldots, c_P\}^T$, according to

$$M(\boldsymbol{\alpha}^{(j)}) = c_0\Psi_0(\boldsymbol{\alpha}^{(j)}) + c_1\Psi_1(\boldsymbol{\alpha}^{(j)}) + \ldots + c_P\Psi_P(\boldsymbol{\alpha}^{(j)}) = \boldsymbol{yc}.$$

- If the $(P+1)$-distribution of the random vector $\mathbf{c}$ can be approximated by a $\mathcal{MVN}(\mu, \mathbf{Z})$, then

$$\boldsymbol{yc} = \Psi_0(\boldsymbol{\alpha}^{(j)})c_0 + \ldots + \Psi_P(\boldsymbol{\alpha}^{(j)})c_P,$$

is distributed as a *univariate* gaussian with mean $(\boldsymbol{y}\mu)$ and variance $(\boldsymbol{yZy}^T)$, in short notation $\mathcal{N}((\boldsymbol{y}\mu), (\boldsymbol{yZy}^T))$.

Inverse Problem based on Non-Deterministic Surrogate

- The uncertainty in the coefficients may be an important information and should be taken into account in the inverse problem.
- Given a sample $\boldsymbol{\alpha} = \{\alpha_1, \alpha_2, \alpha_3\}^{(j)}$, we can construct the constant row vector

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

- We can interpret each **non-deterministic** PC representation, $M(\boldsymbol{\alpha}^{(j)})$, as a **linear combination** of the random vector $\boldsymbol{c} = \{c_0, \ldots, c_P\}^T$, according to

$$M(\boldsymbol{\alpha}^{(j)}) = c_0\Psi_0(\boldsymbol{\alpha}^{(j)}) + c_1\Psi_1(\boldsymbol{\alpha}^{(j)}) + \ldots + c_P\Psi_P(\boldsymbol{\alpha}^{(j)}) = \boldsymbol{yc}.$$

- If the $(P+1)$-distribution of the random vector $\mathbf{c}$ can be approximated by a $\mathcal{MVN}(\boldsymbol{\mu}, \mathbf{Z})$, then

$$\boldsymbol{yc} = \Psi_0(\boldsymbol{\alpha}^{(j)})c_0 + \ldots + \Psi_P(\boldsymbol{\alpha}^{(j)})c_P,$$

is distributed as a *univariate* gaussian with mean $(\boldsymbol{y\mu})$ and variance $(\boldsymbol{yZy}^T)$, in short notation $\mathcal{N}((\boldsymbol{y\mu}), (\boldsymbol{yZy}^T))$.

Inverse Problem based on Non-Deterministic Surrogate

- With an independent additive error model, the discrepancy between each observation and the non-deterministic surrogate model prediction is

$$G^i = \boldsymbol{y}\boldsymbol{c} + \gamma^i, \qquad i = 1, \ldots, N,$$

where $\{\gamma^i\}_{i=1}^N$ are *i.i.d.* RV's with density p_γ. With $p_\gamma = \mathcal{N}(0, \tilde{\sigma}^2)$, we have:

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

- Combines both surrogate uncertainty and data noise in a self-consistent manner. For each data point, the likelihood reaches its maximum if the data and the surrogate mean coincide. Deviations from this mean are weighted by the sum of variances of the noise *and* the 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 based on Non-Deterministic Surrogate

- With an independent additive error model, the discrepancy between each observation and the non-deterministic surrogate model prediction is

$$G^i = \boldsymbol{y}\boldsymbol{c} + \gamma^i, \qquad i = 1, \ldots, N,$$

where $\{\gamma^i\}_{i=1}^N$ are *i.i.d.* RV's with density p_γ. With $p_\gamma = \mathcal{N}(0, \tilde{\sigma}^2)$, we have:

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

- Combines both surrogate uncertainty and data noise in a self-consistent manner. For each data point, the likelihood reaches its maximum if the data and the surrogate mean coincide. Deviations from this mean are weighted by the sum of variances of the noise *and* the 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 based on Non-Deterministic Surrogate

- With an independent additive error model, the discrepancy between each observation and the non-deterministic surrogate model prediction is

$$G^i = \boldsymbol{y}\boldsymbol{c} + \gamma^i, \qquad i = 1, \ldots, N,$$

where $\{\gamma^i\}_{i=1}^N$ are *i.i.d.* RV's with density p_γ. With $p_\gamma = \mathcal{N}(0, \tilde{\sigma}^2)$, we have:

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

- Combines both surrogate uncertainty and data noise in a self-consistent manner. For each data point, the likelihood reaches its maximum if the data and the surrogate mean coincide. Deviations from this mean are weighted by the sum of variances of the noise *and* the 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 based on Non-Deterministic Surrogate

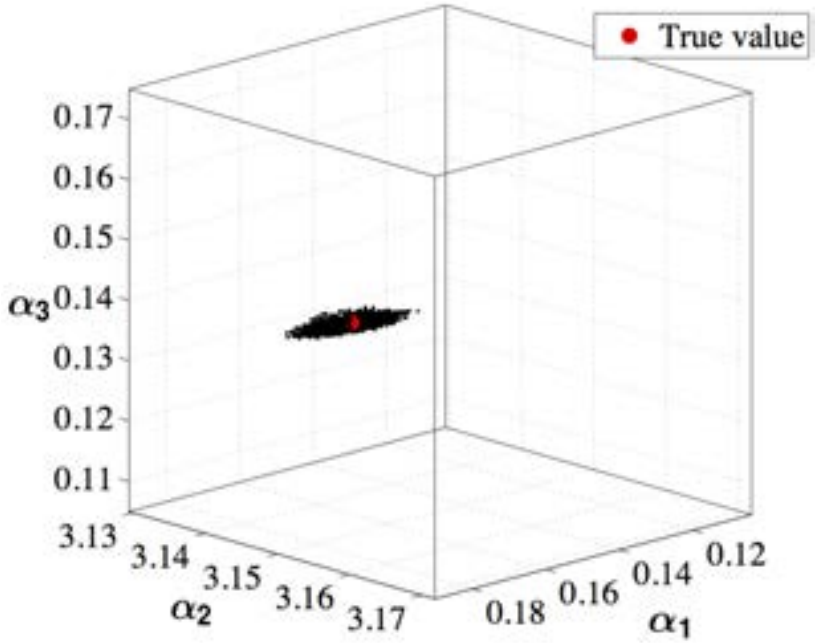
- From Bayes' theorem, the joint posterior distribution is given by

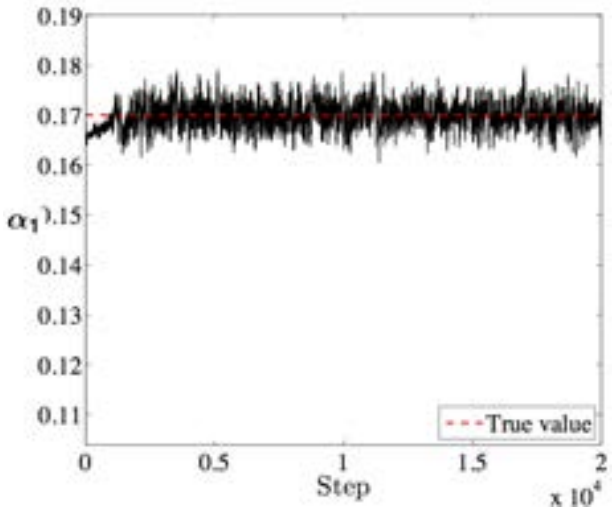
$$\pi\left(\{\alpha_1,\alpha_2,\alpha_3\}, \textit{hyperp}\middle|\ \{G\}_{i=1}^{N}\right) \propto \mathcal{L}\left(\{G\}_{i=1}^{N}\middle|\{\alpha_1,\alpha_2,\alpha_3\}, \textit{hyperp}\right) \textit{Priors}$$

- The problem then reduces to sampling the posterior using a suitable algorithm, e.g. Adaptive Metropolis.

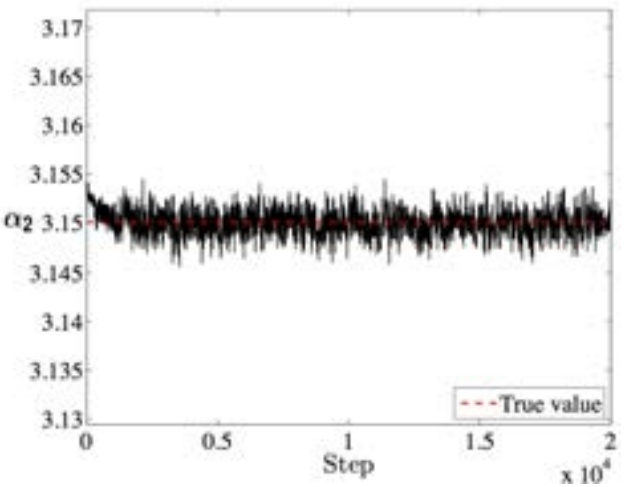
NON-DETERMINISTIC PC surrogate

- Chain localizes at the intersection of the isosurfaces extracted from the PC surrogate of each observable according to: $\mathrm{MAP}(M_k)(\xi)$, $k = 1, 2, 3$.
- True value is recovered with good accuracy.
- Results look similar to the deterministic setting.

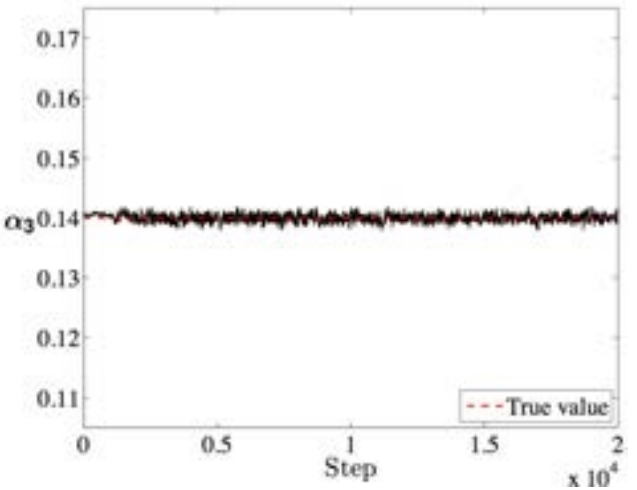




α_1



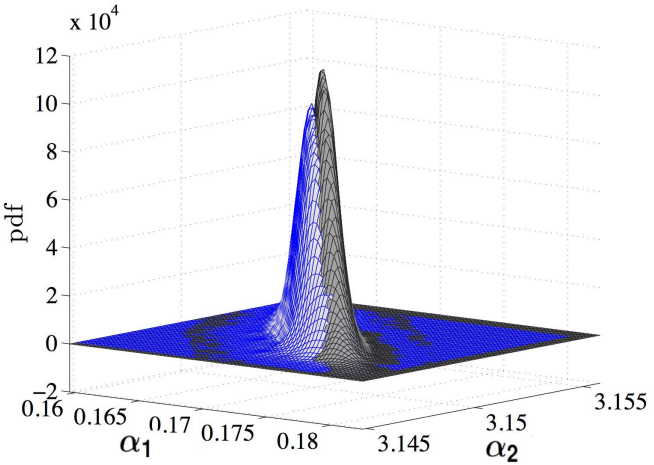
α_2

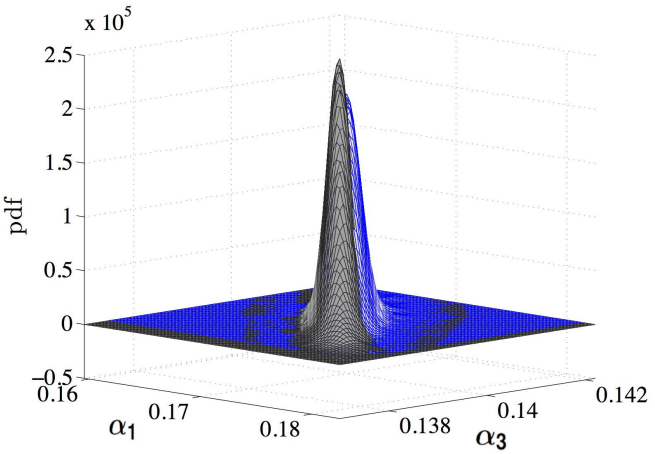


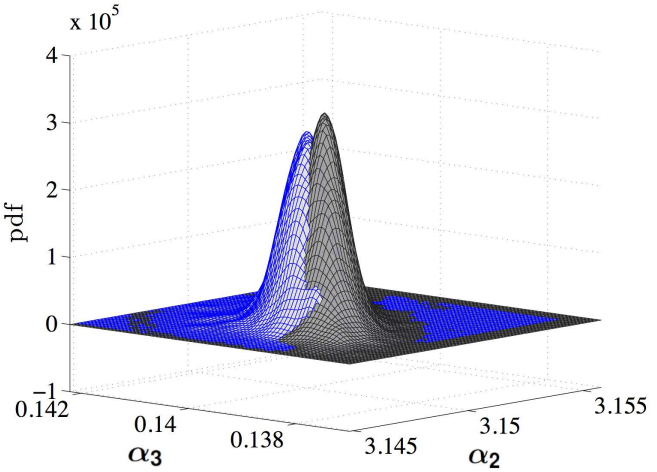
α_3

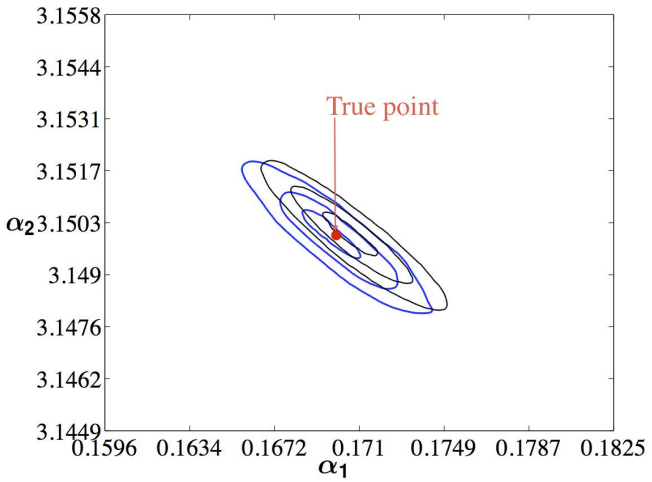
DETERMINISTIC vs. NON-DETERMINISTIC surrogates.

- Joint posteriors based on **Deterministic** and **Non-Deterministic** surrogates.
- Substantial correlations stemming from the forward model solution, and manifested during the inference through the PC surrogate.

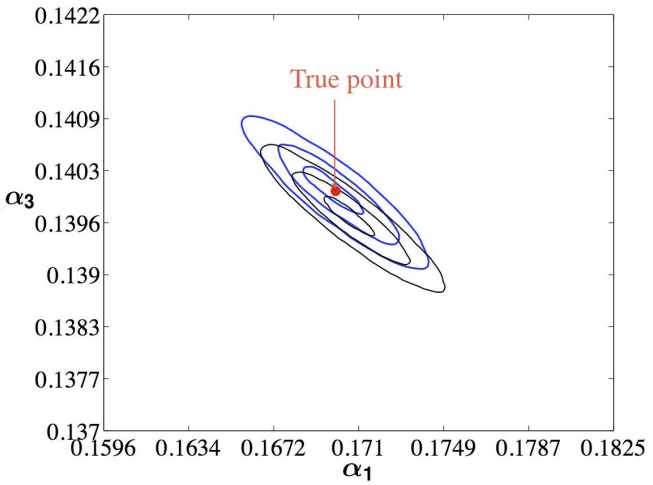




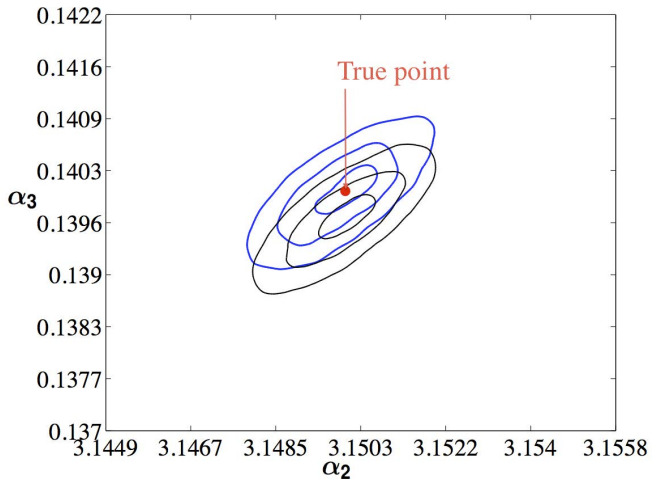




$\pi(\alpha_1, \alpha_2)$

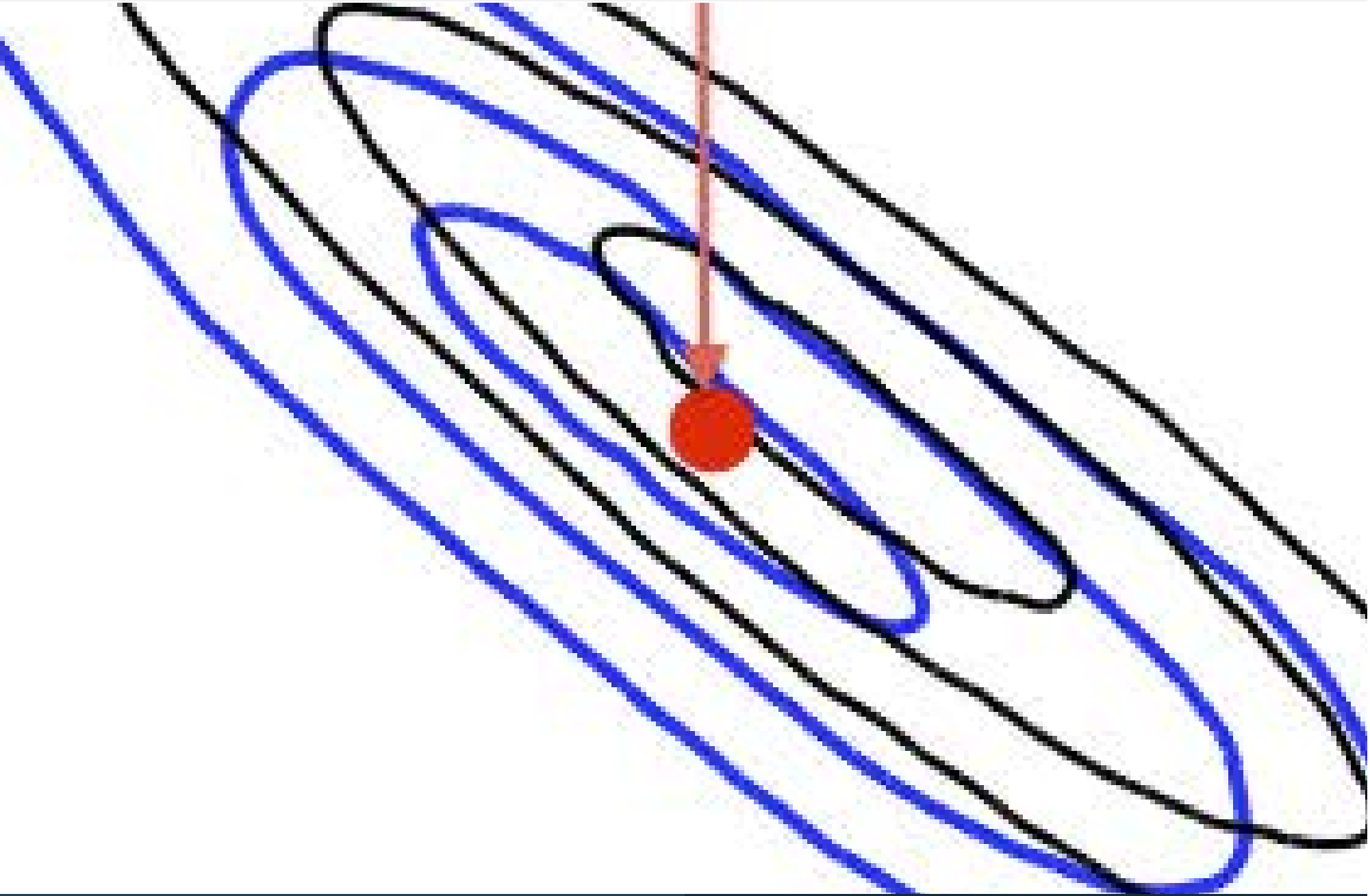


$\pi(\alpha_1, \alpha_3)$

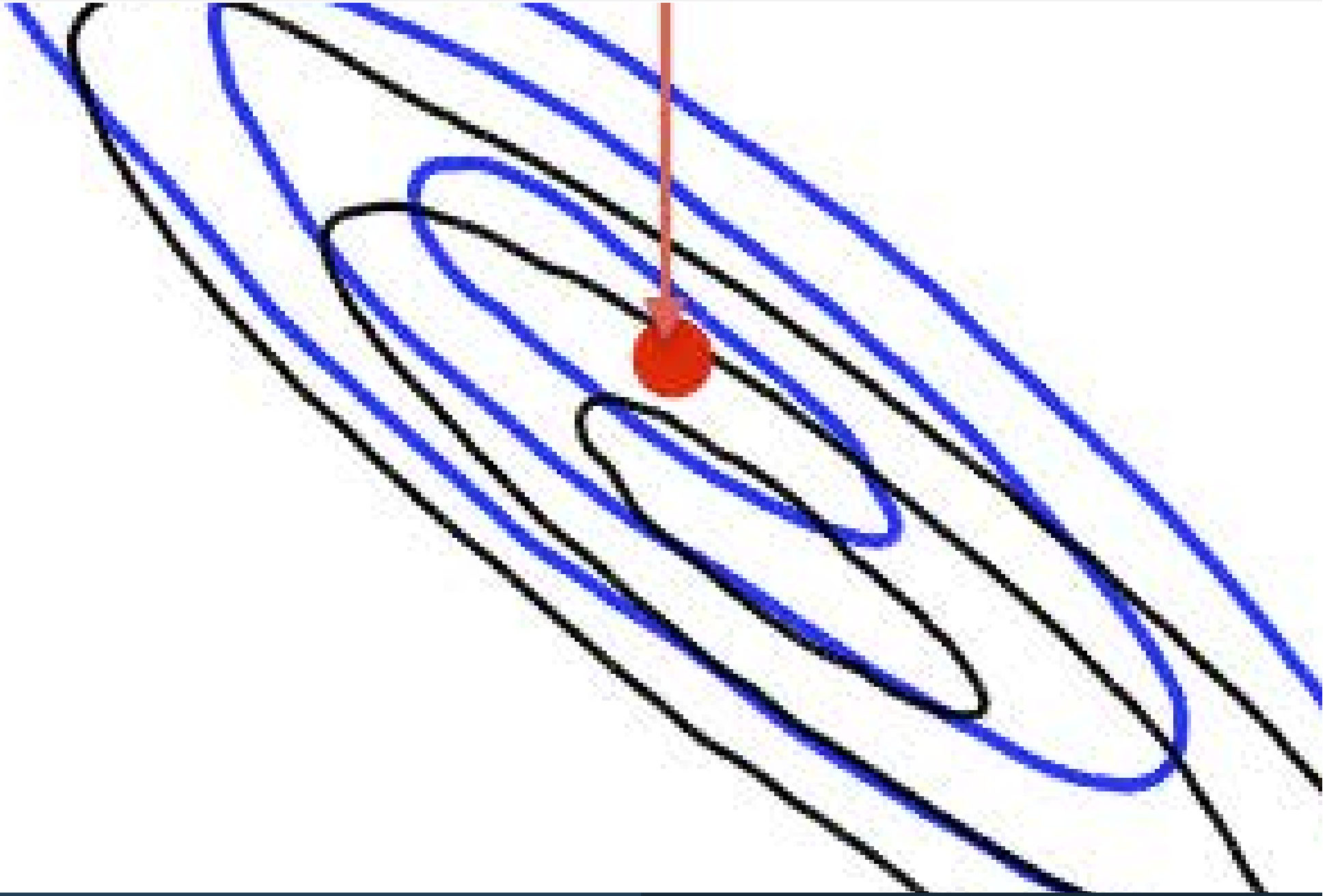


$\pi(\alpha_2, \alpha_3)$

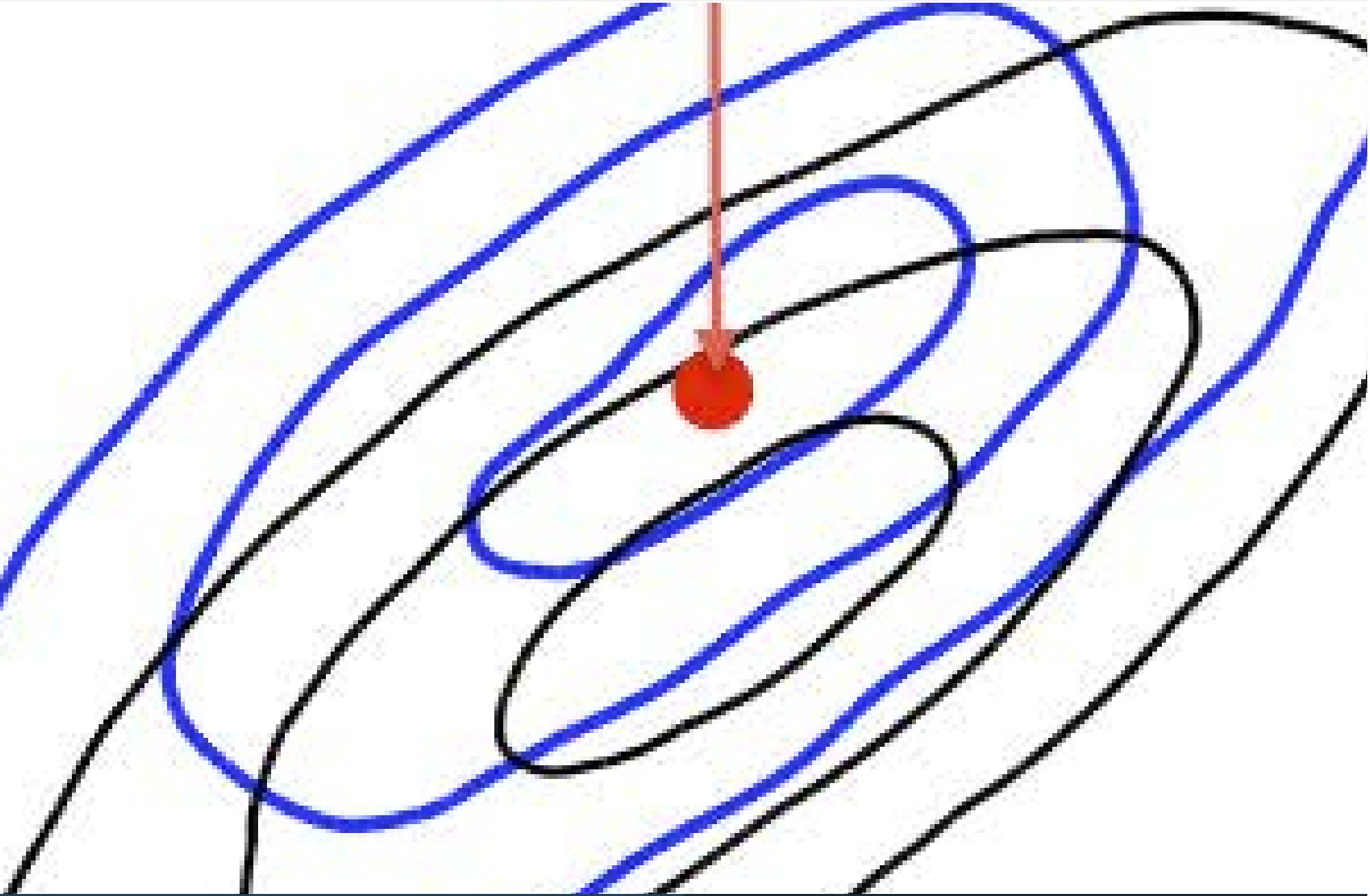
DETERMINISTIC vs. NON-DETERMINISTIC surrogates.



DETERMINISTIC vs. NON-DETERMINISTIC surrogates.



DETERMINISTIC vs. NON-DETERMINISTIC surrogates.



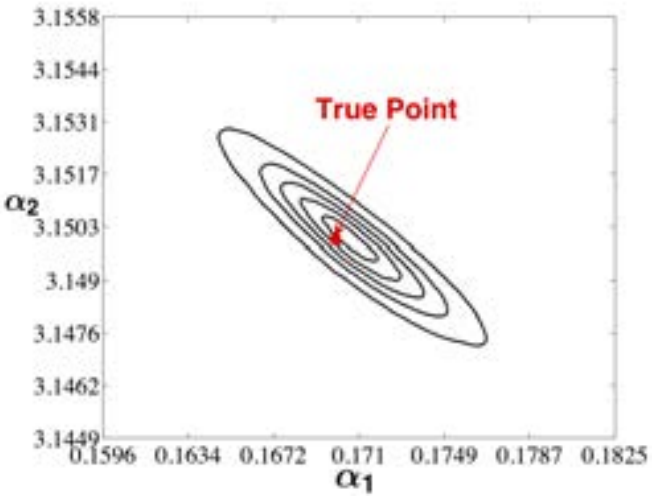
Summary & Conclusions

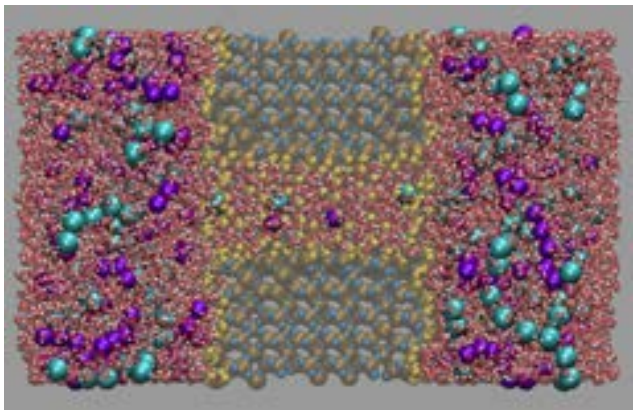
- UQ can be successfully applied to MD simulations.
- Two distinct sources of uncertainty were investigated:
  parametric uncertainty in the potential and intrinsic (thermal) noise.
- PC expansions and Bayesian inference were exploited to develop a framework to isolate the impact of parametric uncertainty on the MD predictions and, at the same time, properly quantify the effect of the intrinsic noise.
- Uncertain PC surrogates provide a suitable tool, especially in the presence of noisy data.
- We addressed the UQ problem in both its main components, the forward propagation and the inverse problem, focusing on two different MD systems.
- Specifically, we showed to suitability of using PCe in the MD context
  for both the forward propagation and inverse problem.
- In part I: we described few important physical mechanisms occurring in a nanopore flow, due to physical parameter effects (diameter, gating charge) as well as effects stemming from potential uncertainty.
- In part II: we successfully showed how to use PCe to infer *atomistic* quantities using *macroscale* observables, obtaining a PDF on the potential parameters.

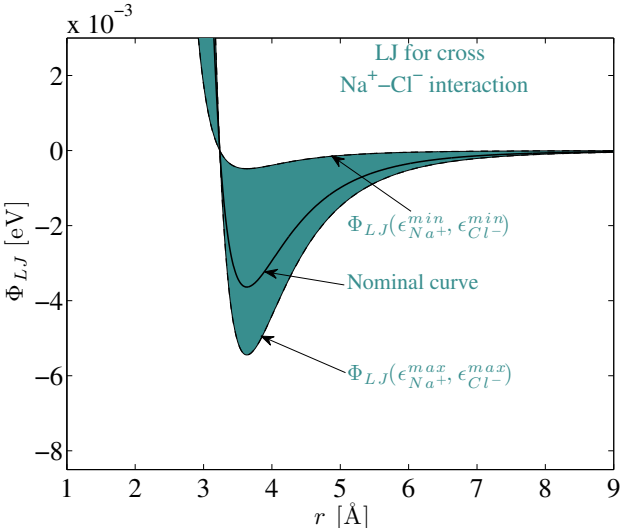
Acknowledgments & Collaborators

- Prof. Omar Knio (Duke, JHU, KAUST)
- Dr. Reese Jones (Sandia)
- Dr. Habib Najm, Dr. Bert Debusschere, Dr. Kachick Sargsyan, Dr. Maher Salloum, Dr. Helgi Adalsteinsson (Sandia).

Thank you for your attention







Research supported by:
US, DOE - Office of Advanced Scientific Computing Research