Quantification and prediction of uncertainty in coarse-grained models of molecular simulations

Martha Grover Gallivan

with Cihan Oguz and Andres Hernandez Moreno

School of Chemical & Biomolecular Engineering
Georgia Institute of Technology

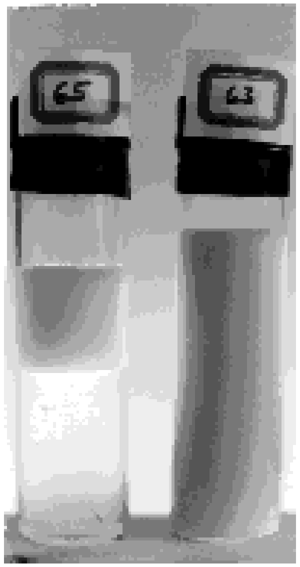
August 29, 2007

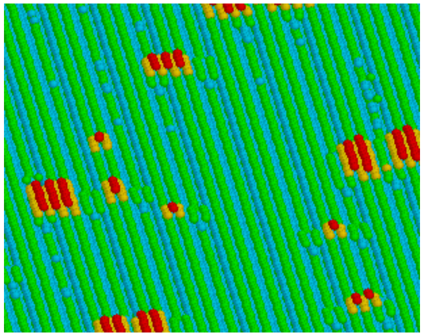
Mathematics of Model Reduction

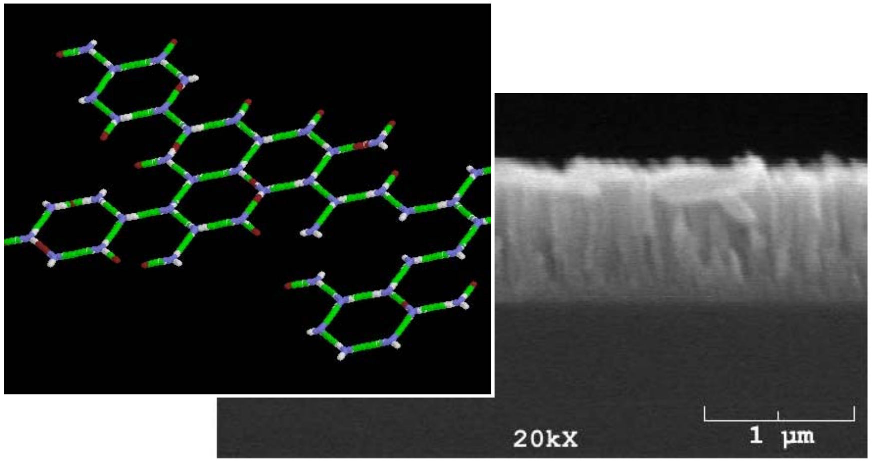
Support from US Air Force Office of Scientific Research: FA9550-04-1-0183, FA9550-07-1-0161

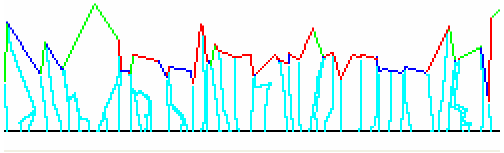
Current practice in materials development

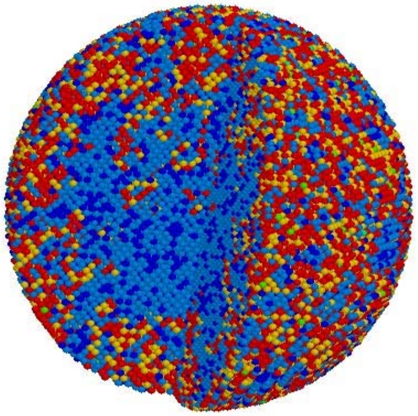
- Design of materials and processes is largely empirical
- Macroscopic models are used in process design, but molecular/microscopic models are not
- Materials properties (advanced materials) require consideration of molecular structure



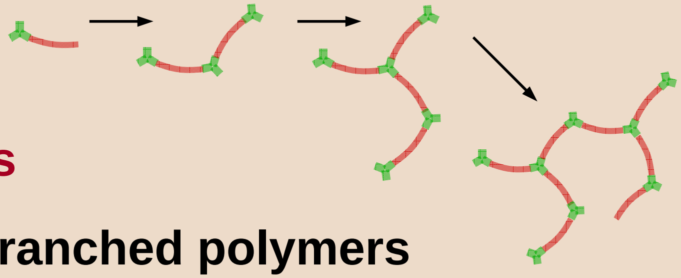






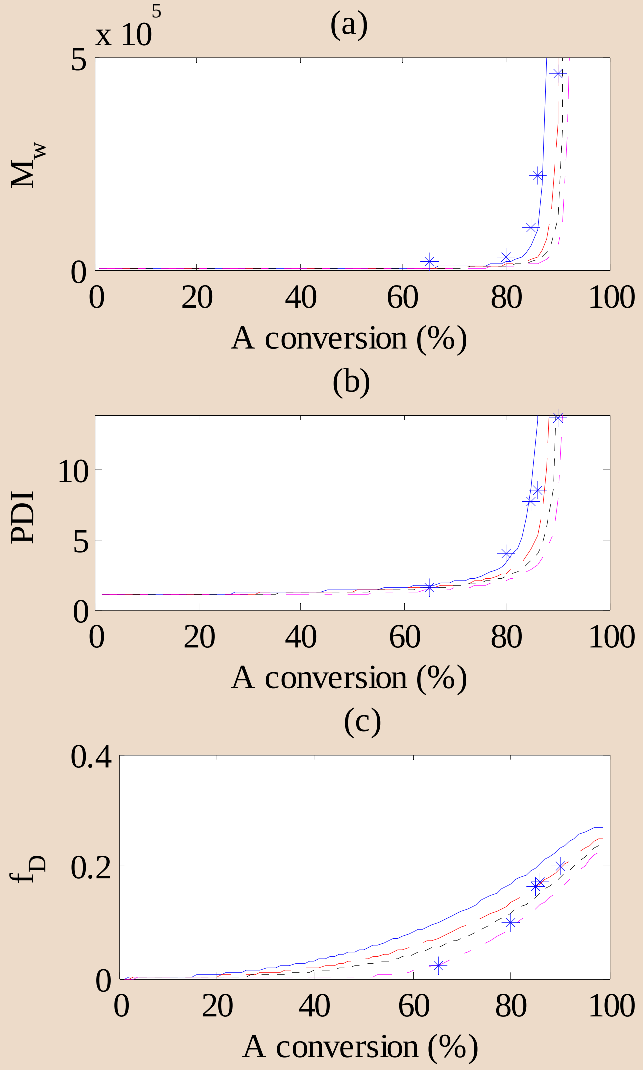


Evolution of polymer networks



Polymer networks

- A_2+B_3 hyperbranched polymers
- No solvent → negligible cycle formation
- NMR measurements provide branching structure
  - NMR data suggests unequal reactivity of free B_3
- Addition of monofunctional A groups (A_2:B_3:A=1:1:1)
  - Non-intuitive effect
  - Not a robust operating point



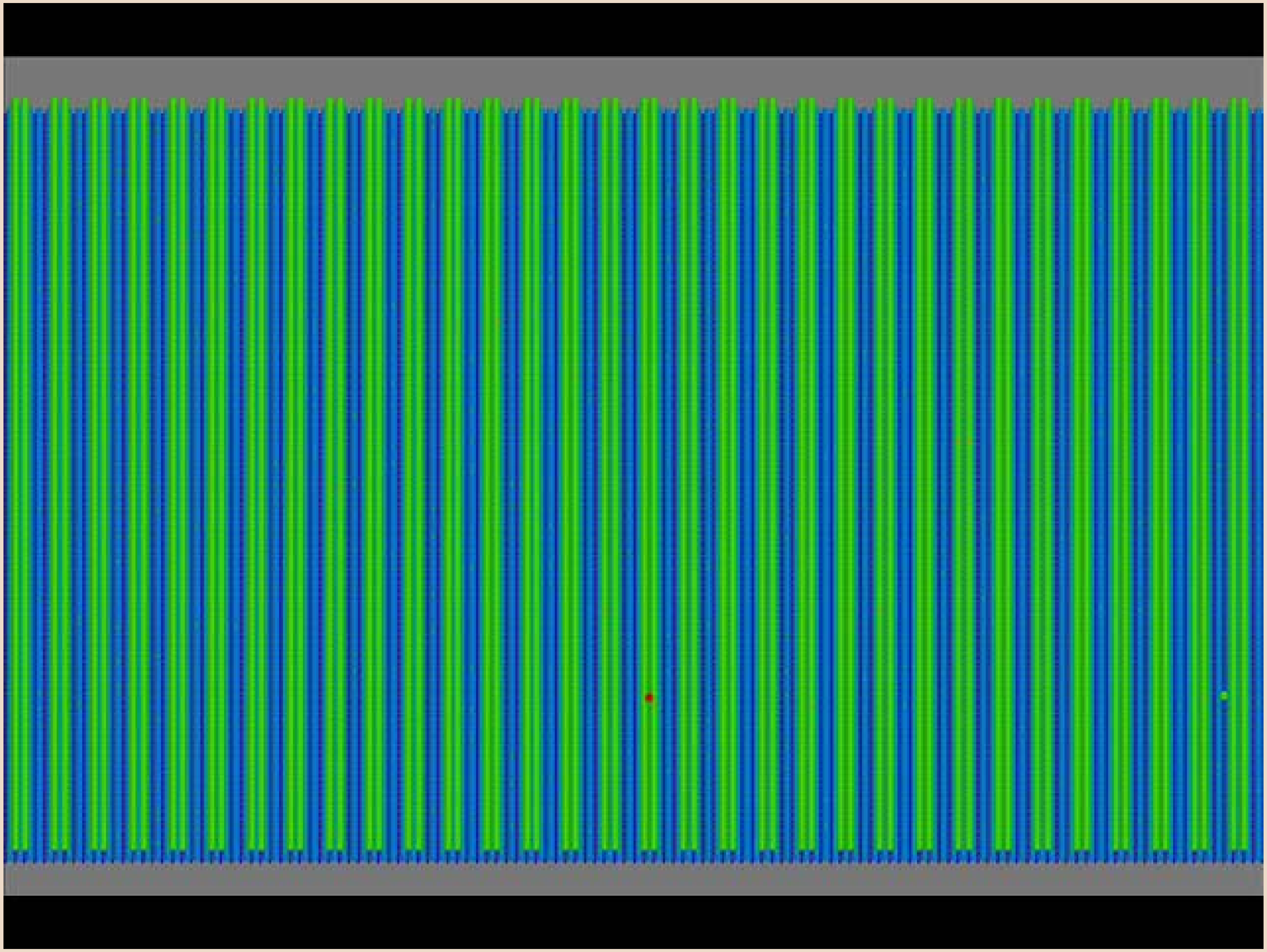
What is the state of the polymer network?

Oguz, Unal, Long, and Gallivan, *Macromolecules*, in press. (ARO DAAD 19-02-1-0275)

Background

- Objectives
  - Use complex simulations to control and engineer nanoscale material structure
  - Understand and predict the uncertainty
- Technical approach
  - Build reduced order (reduced computation) models based on discrete configurations using the full simulations
    - Aggregation
    - Discrete number of states
  - Use spatial statistics to model the error
    - Errors in a reduced order model are correlated
  - Current state: multiple modeling approaches, error analysis is ad hoc or non-existent
    - Adaptive tabulation (Pope 1995)

Plant model



Itoh 2000

Key question

What is the mathematical structure of a molecular system?

Options

1. Probabilistic representation

- **Master equation or Liouville equation**
- **State-affine control system**
- **Graph structure**

2. Stochastic simulations of time-dependent behavior

- **Molecular dynamics (many body Hamiltonian)**
- **Kinetic Monte Carlo (Poisson statistics)**
- **State is not meaningful as a dynamic state**

3. Moment equations

- **Not closed for many properties of interest**

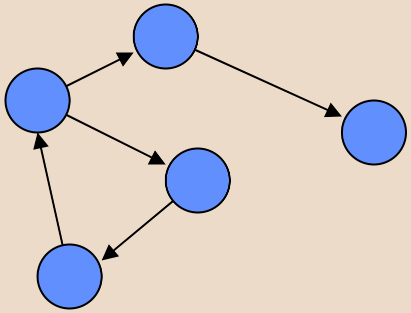
$$\frac{dx}{dt} = A(u)x$$

$$y = Cx$$

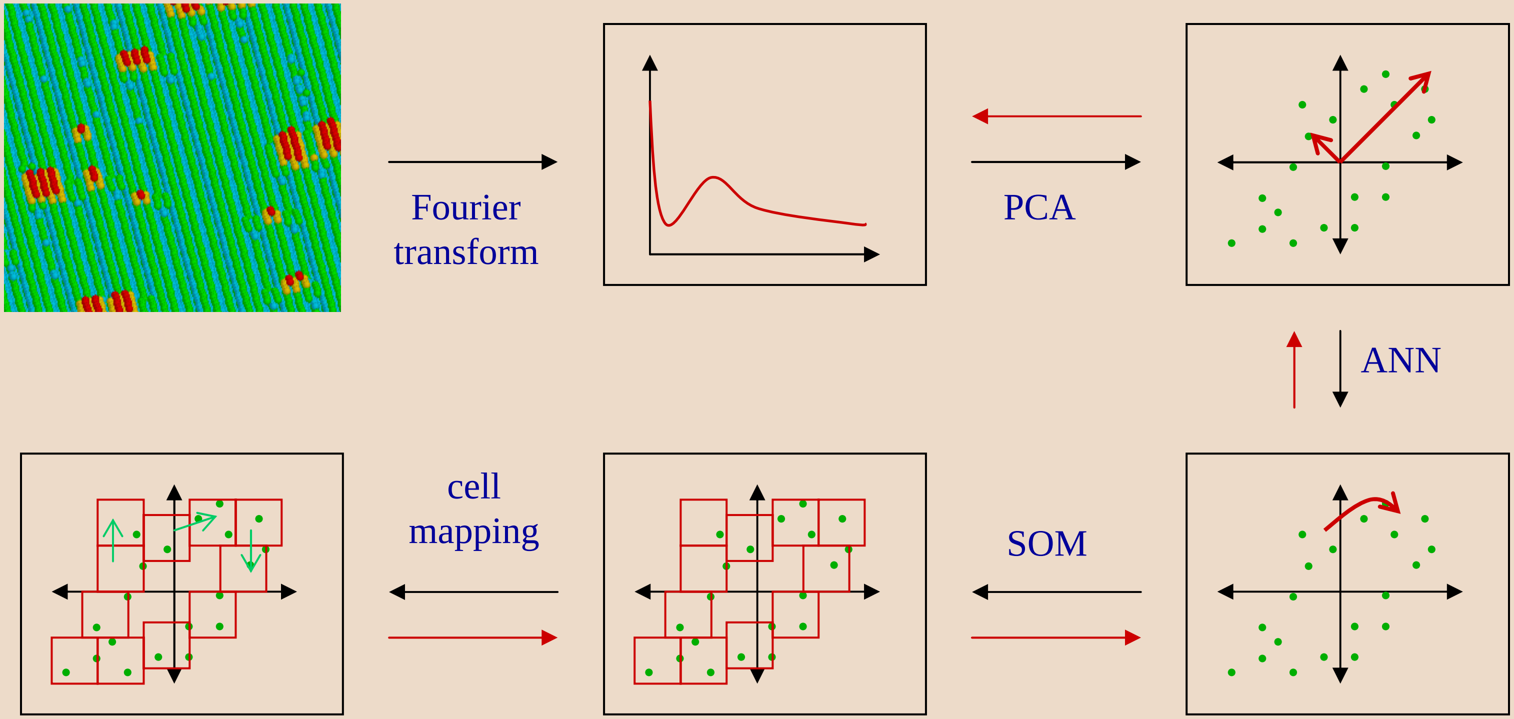
$$x_i \geq 0$$

$$\sum_{i=1}^{n} x_i$$

$$x \in R^n$$



Reduction Approach



molecular simulation

Fourier transform

spatial correlation function

PCA

linear subspace

ANN

nonlinear manifold

SOM

discrete space

cell mapping

process model

Characterizing the state space

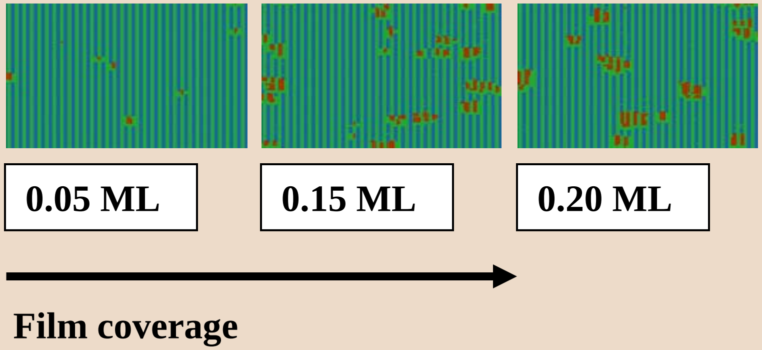
Simulations with constant and varying Ga flux profiles

Run a **set of simulations** under different conditions

Record **surface snapshots**

Quantify the **microstructure** of the surface snapshots

Performed **76 KMC simulations**
Growth Temperature: **580 °C**
Incident As_2 flux: **0.4 ML/s**
Incident Ga flux: **0.06-0.20 ML/s**
Lattice size: **300x300**



1521 surface snapshots are recorded

Use a **step-step correlation** (SSC) function.

Only interested in **relative positions** of the steps.

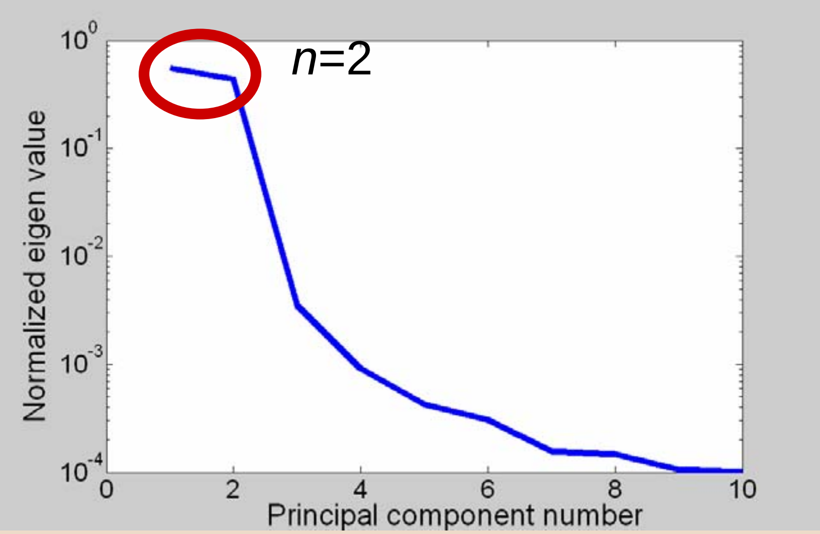
Each snapshot is described by a **(300x16)** SSC matrix.

Principal component analysis

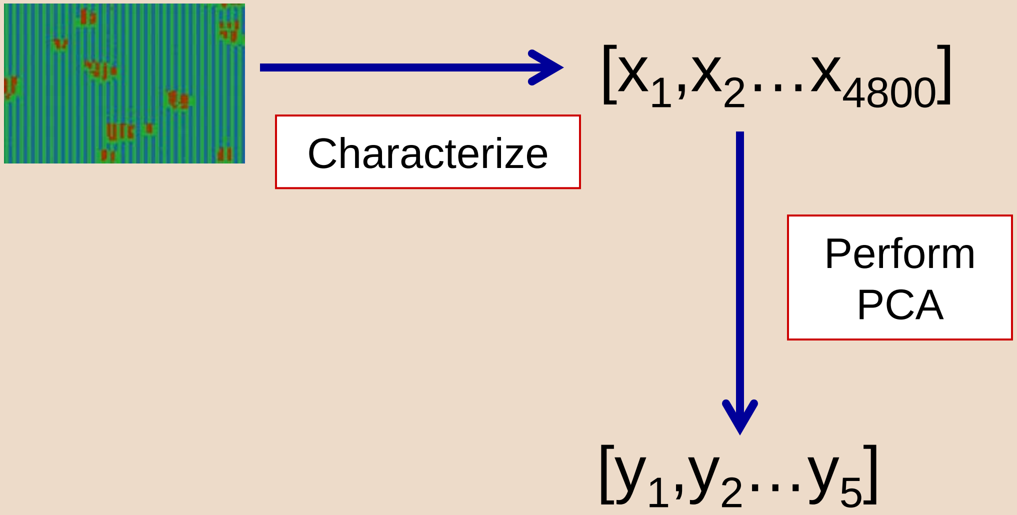
Reducing the dimensions of the simulation data

PCA retains most of the information:

- **Find the principal components**
- **Plot eigenvalues versus PCs**
- **Pick the first '*n*' PCs that can capture most of the variance**



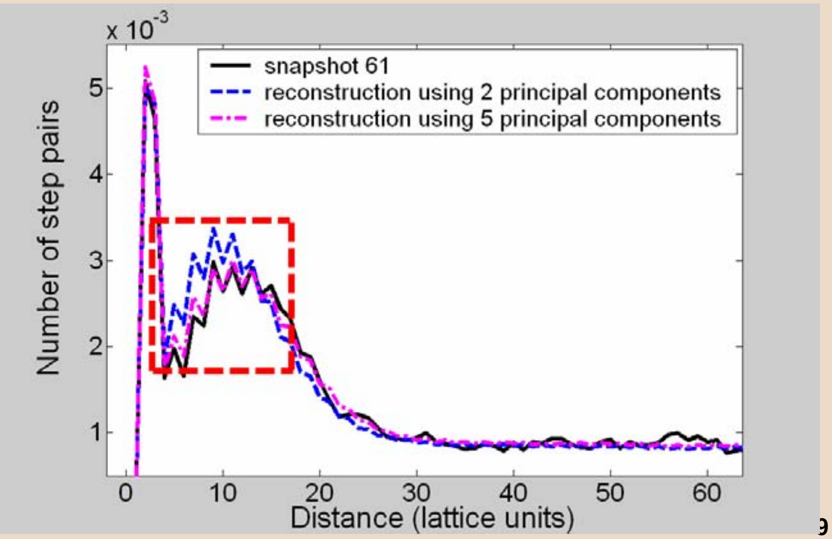
Data reconstruction showed that we need 5 PCs


$[x_1,x_2\ldots x_{4800}]$

Characterize

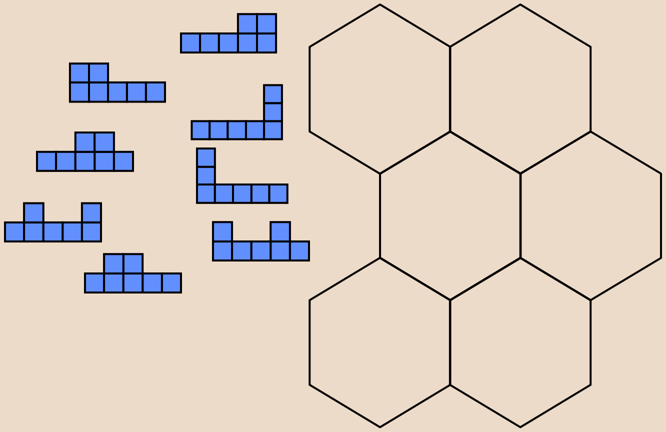
Perform PCA

$[y_1,y_2\ldots y_5]$



The self organizing map

An algorithm used for grouping similar surface snapshots

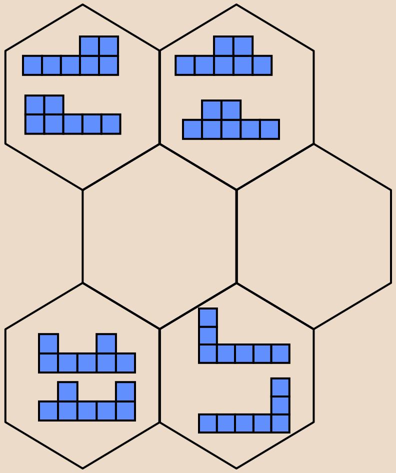


Before SOM training:

- **Each** surface snapshot **is described by a 5-D data vector.**
- **Each map node is described by a 5-D** prototype vector.

During SOM training:

- **Prototype vectors are** initialized randomly **and modified during training.**
- **Each snapshot is** mapped **onto a particular node.**
- **Similar snapshots are** mapped **onto the** same **map node.**



After SOM training:

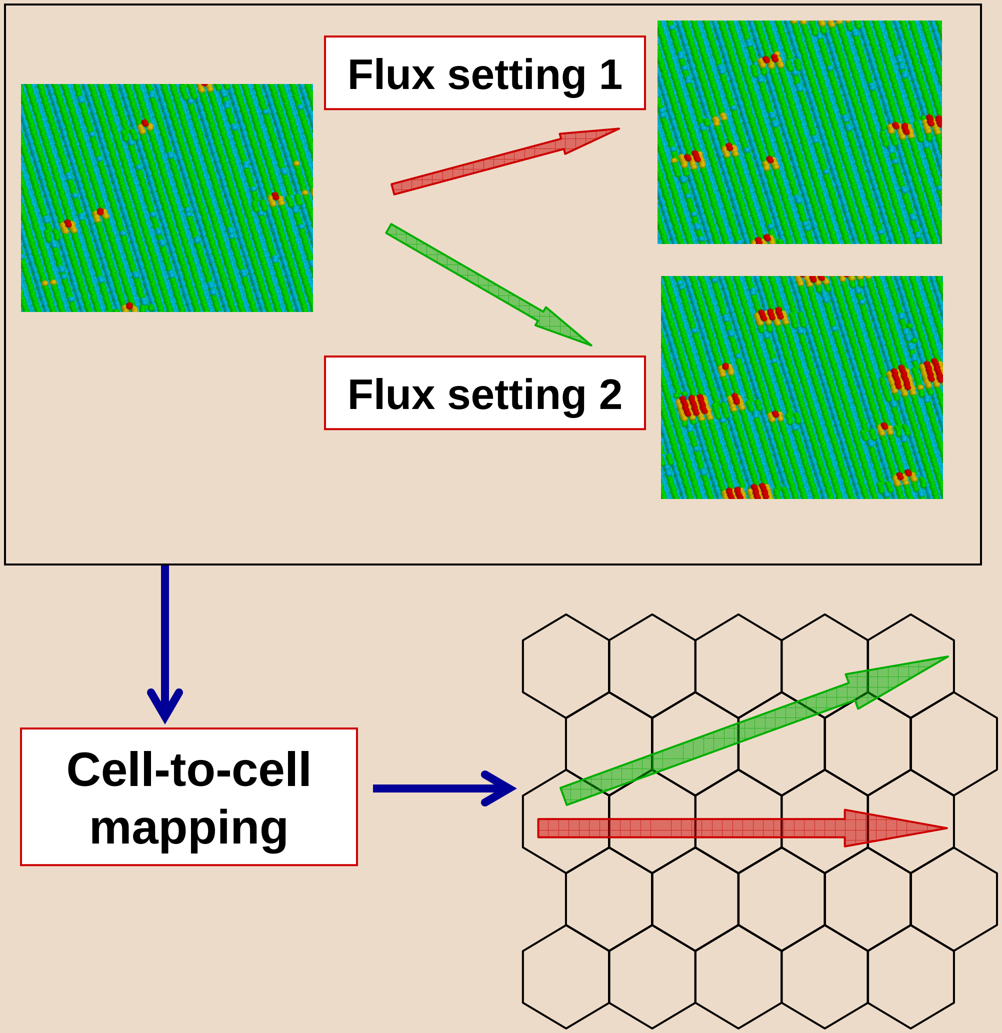
- **1521 snapshots are grouped in** 175 map nodes**.**

Cell mapping

Transitions between different snapshot groups

- Performing system identification:
  - Pick one snapshot from each map node.
  - Run additional simulations starting from selected snapshots under each different flux setting.
  - Identify and record the map node that the system reaches in each case.

- Cell mapping provides a dynamic model:
  - Relationship between the system state and the surface coverage under different flux profiles.



Flux setting 1

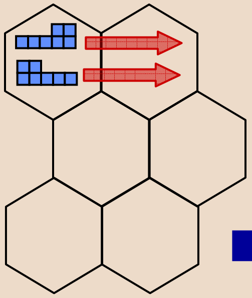
Flux setting 2

Cell-to-cell mapping

Local (one-step) error associated with cell mapping

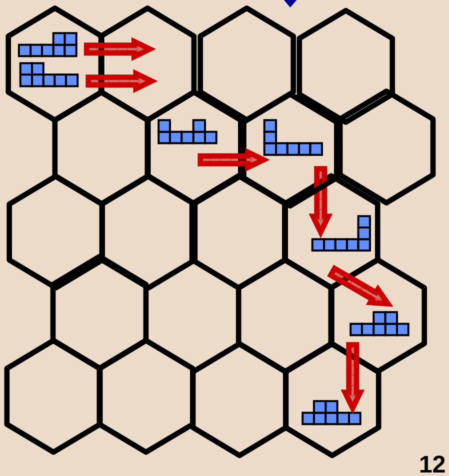
Assumption: Structures in the same node should show identical dynamic behavior under same input.

If the assumption is correct for one step

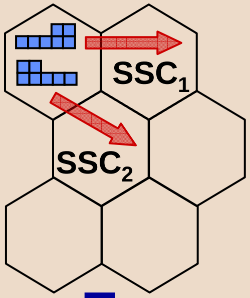


Cell mapping error=0

If the assumption is correct for multiple steps



If the assumption is incorrect

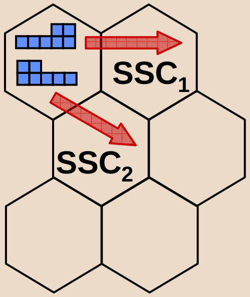


Cell Mapping error:

$||SSC_1-SSC_2|| / [(||SSC_1|| + ||SSC_2||) / 2]$

SSC functions are constructed from prototype vectors.

Results of the CME (local error) analysis

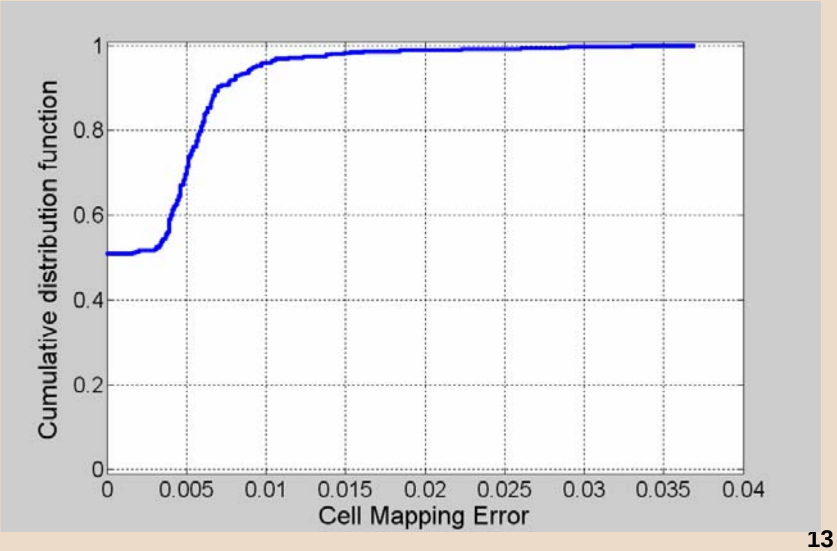


Compute the error for each node under each flux setting

Discretize the error domain into bins

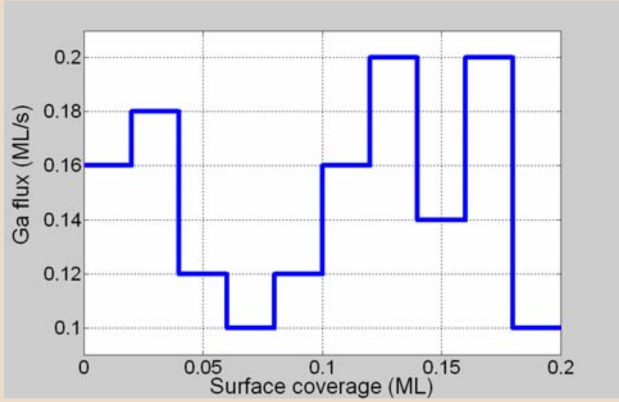
Compute the probability of having certain error values

- 52% of the mappings turned out to be identical
  - With a 0.52 probability, the mapping error is ‘0’
- With a 0.9 probability:
  - Mapping error < 0.75%
- Surface structures in the same groups show similar dynamic behavior.
- A larger SOM can decrease the CME.
  - Larger SOM= Larger cell map
  - Computational load for cell mapping would increase.



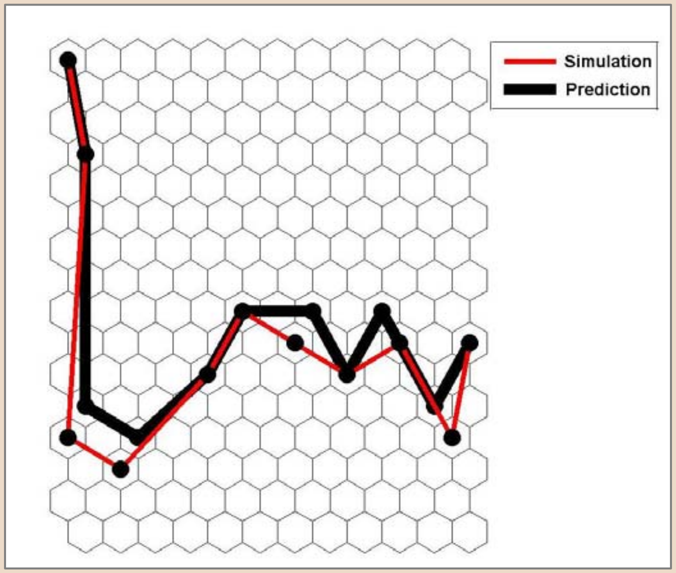
Testing the dynamic model

Run test simulations and evaluate model performance

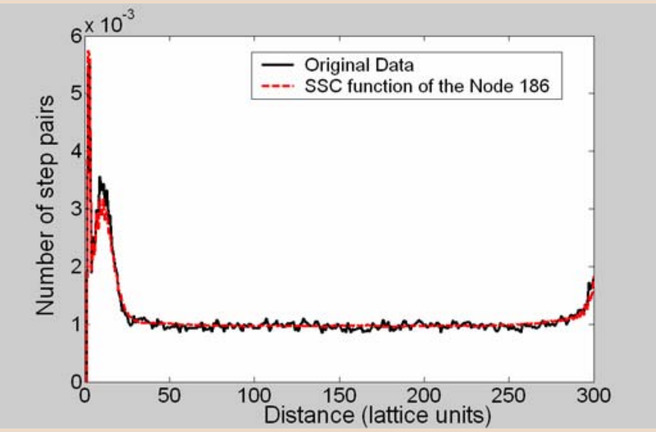


Test KMC simulation with random input profile

Estimate the trajectory and compare with the real trajectory of the KMC simulation



Check if the prediction for the final film structure accurate

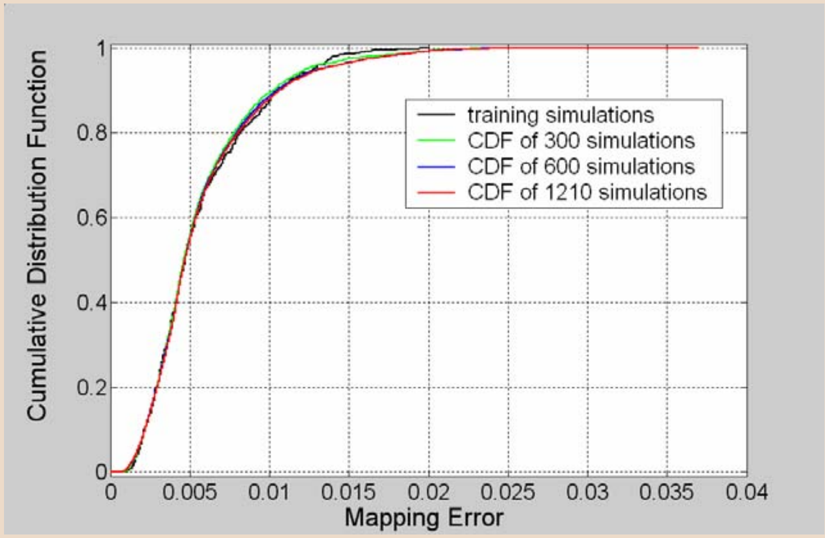


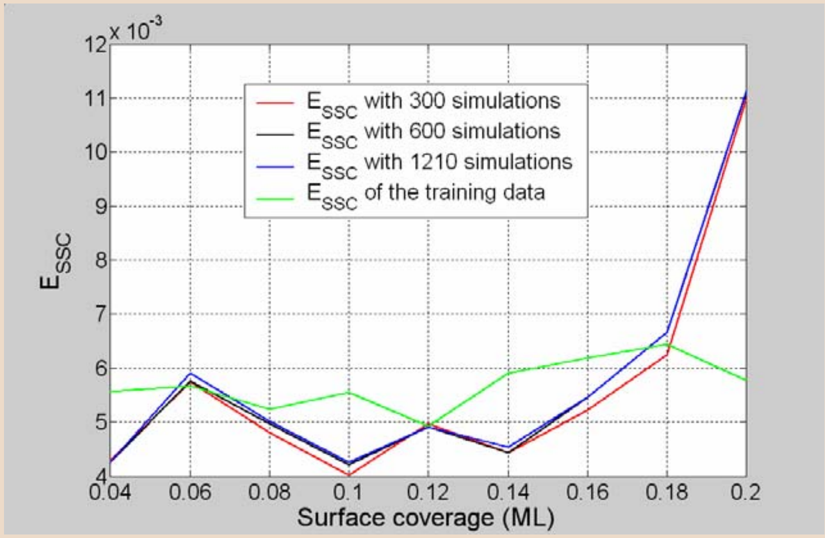
How can we quantify prediction error?

$||SSC_s - SSC_p|| \,/\, ||SSC_s||$

Global (multi-step) prediction error

- Cumulative distribution function (CDF) of the error
  - With a probability of 0.99, error is less than 2.5 %.
- The mean value of error at different film coverage levels (E_{SSC}<1.2%)
  - Mean E_{SSC} increases steadily at high film coverage (prediction gets worse)
- Error at 0.2 ML is lower for simulations in the training data
  - Dynamic model is more familiar with the film structures in the training data
- No need to run more test simulations

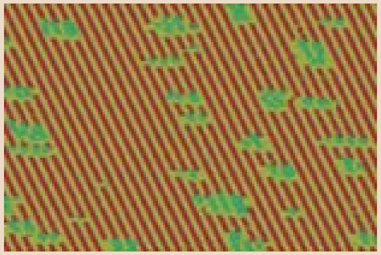




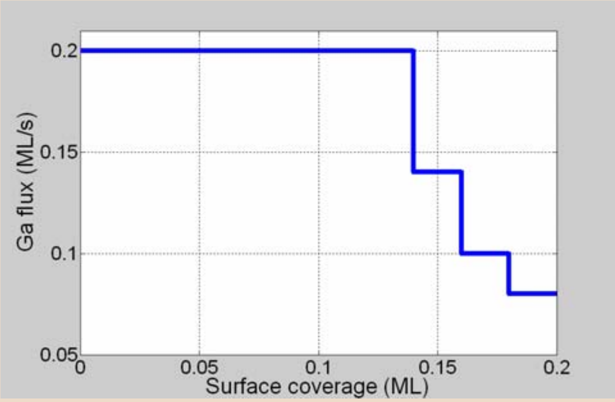
Optimizing film structure

Minimizing the deposition time

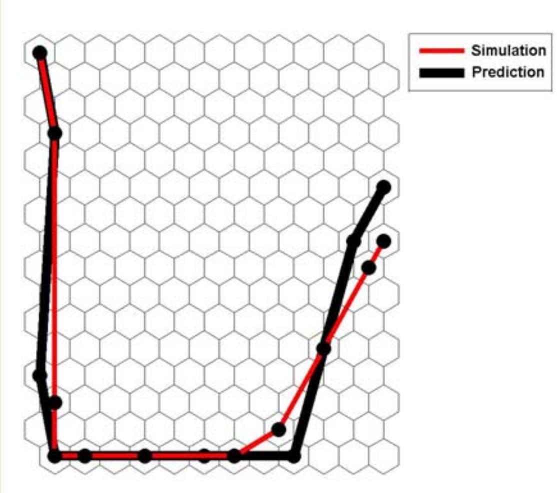
Find the most regular film structure



Find the optimal flux profile to reach that structure

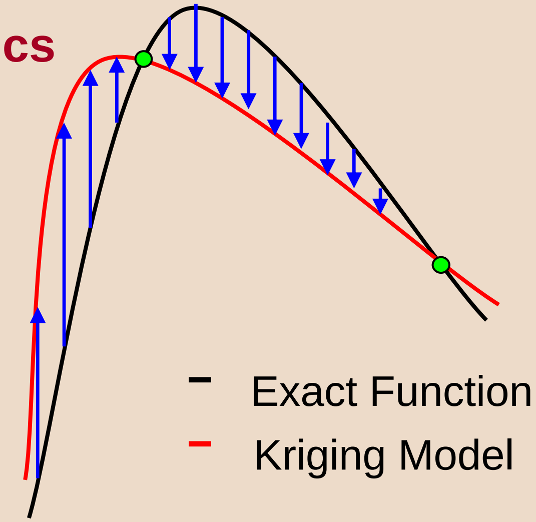


Simulation vs prediction



- **Used eight flux settings (0.06, 0.08 … 0.20 ML/s).**
  - **10 surface coverage intervals.**
  - **8^{10} possible flux profiles.**
- **48% reduction in the deposition time.**
- **Optimal profile is found without running 8^{10} KMC simulations**
  - **It would have taken 2.9 million years using an Intel Xeon processor (2.66 GHz speed).**
  - **Took 5 minutes using the dynamic model**

Modeling of the Error



Error quantification and prediction via spatial statistics

- **Develop procedures for spatial statistics**
  1. **the sample points**
  2. **the form of the spatial correlation function**
  3. **a set of regression functions**
  4. **the method for parameter identification**

- **Apply and generalize kriging for static systems to the dynamic models**

$$\boxed{x_{k+1} = F(x_k, u_k)}$$

$$x_{k+1} = \sum_{i=1}^{p} \beta_i f_i(x_k, u_k) + Z(x_k, u_k)$$

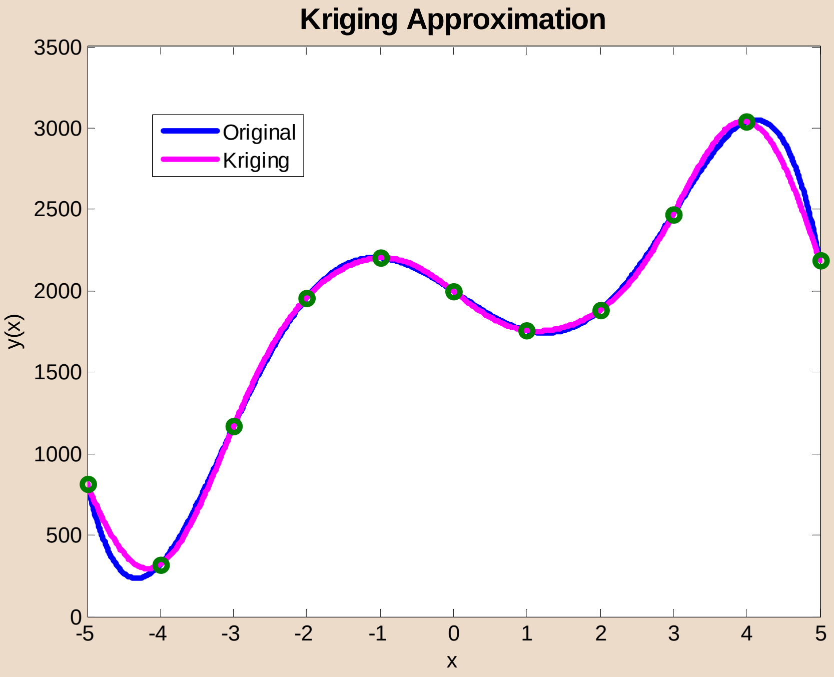
$$Z \sim N(0, \sigma^2)$$

$$Cov[Z(x_i), Z(x_j)] = \sigma^2 \cdot R(x_i, x_j), i, j = 1, \ldots n$$

- **Discrete time models**
- **Kriging is a method, initially developed by geologists, which uses the sample points as a “true” reference points to infer the value of the unknown points.**

A Simple Example

- Parameter Identification
  - MLE is a good method for its simplicity, easy to program, fast response and accurate solution.
- DACE
  - A standard experimental design approach causes high error near the boundaries due to the local approximations performed in kriging.
- Regression function
  - A constant (not necessarily the mean)
- Model of error correlation
  - Gaussian



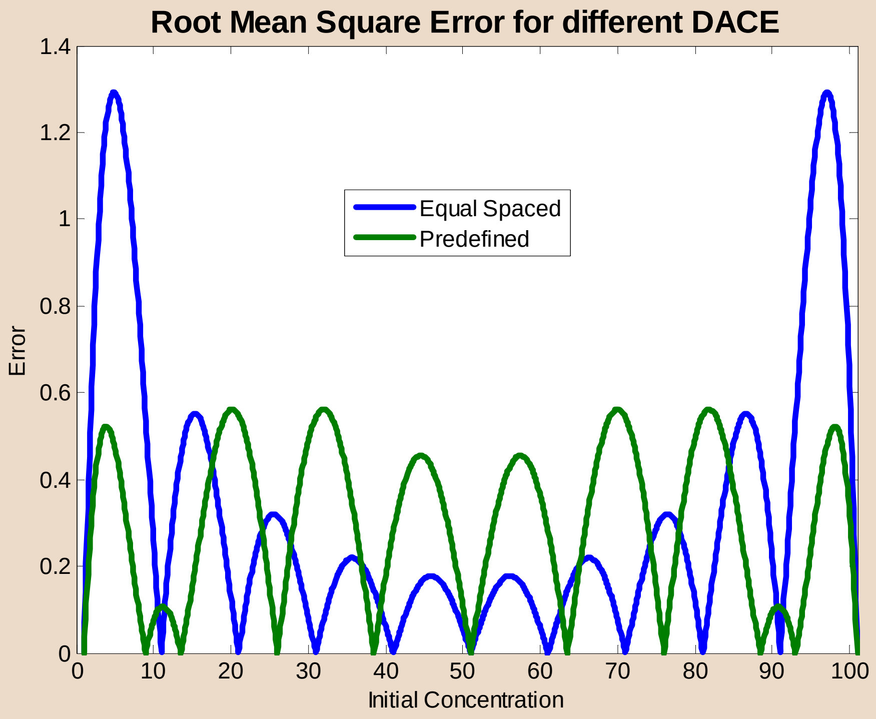
Prediction of the Error Variance

- Use variance as an estimate for uncertainty in the model

$$\boxed{x_{k+1} = F(x_k, u_k)}$$

- Observations
  - No uncertainty at the sampled points
  - Uniform sampling leads to high uncertainty near the boundaries

- Questions
  - Where to sample?
  - How to use the snapshots?
  - How to resample?
  - What regression functions to use?



Impact

- Empirical models based on large simulations are used in many applications
  - Tabulation models in combustion and reacting flow
  - Equation-free computing, tabulation, and Markov modeling in molecular simulations
  - Potential applications in multi-vehicle systems

- Methods must be developed to predict and control the uncertainty in the reduced models (variance v. bounds)
  - Suggest when to resample
  - Steer away from uncertain regions

- Spatial statistics provide a flexible method for modeling error across this spectrum of empirical models

Support from US Air Force Office of Scientific Research: FA9550-04-1-0183, FA9550-07-1-0161