DATA ANALYTICS IN NANOMATERIALS DISCOVERY

Michael Fernandez | OCE-Postdoctoral Fellow
September 2016

www.data61.csiro.au
Materials Discovery Process

Materials Genome Project

Integrating computational methods and information with sophisticated computational and analytical tools to shorten the duration of materials development from 10-20 years to 2 or 3 years.
Materials and Molecular Modeling

Collaborators

• Big data and HPC integration
  Piotr Szul
  Yulia Arzhaeva

• Deep learning and GPU computations
  Chris Watkins

Team leader
Amanda Barnard

Sun Baichuan
Outline

• Material Discovery Process
  • Methods for atomistic simulations of materials
  • Experimental confirmation

• Data-driven Computational Nanomaterials Discovery
  • Hypothetical material space sampling (structure generation and statistics)
  • \textit{In silico} high-throughput characterization (atomistic simulations and machine learning)
  • Data storage, analytics, exploitation and integration
Theory can predict properties of materials

- Quantum chemistry methods can cover any chemistry
- Empirical potentials exist for large number of elements
- Computation is scalable and generically deployable

\[ H = E_{\text{nuclei}}(\{R_I\}) - \sum_{i=1}^{N_e} V_i^2 + V_{\text{nuclei}}(r_i) + \frac{1}{2} \sum_{i \neq j}^{N_e} \frac{1}{|r_i - r_j|} \]
Atomistic Simulations of Materials

Computational predictions later confirmed by experiments

- Self-assembly mechanism of nanodiamonds


DFTB simulations of the surface electrostatic potential of dodecahedral diamond nanoparticles of a) 2.2 nm and b) 2.5 nm

The arrow indicates (111)|(111) interface between two 4 nm sized nanodiamonds.
Data Analytics for Nanomaterials | Michael Fernandez

Nano-materials discovery cycle

- Polydispersive systems
- Exponential increase in complexity and diversity
- Nearly infinite combinatorial problem


Modern materials discovery cycle

- Polydispersive systems
- Exponential increase in complexity and diversity
- Nearly infinite combinatorial problem

Nanomaterials Screening

Departing from the Edisonian approach

In silico structure generation

Combinatorial In silico design

VS.

Data Analytics for Nanomaterials | Michael Fernandez
Nanomaterials Screening

- Systematic and extensive materials performance ranking.
- “Big data” discovery of structure-property relationships in unknown materials domains.
- Accelerated identification of high potential candidates and rational design principles.
Data Analytics Challenges

Information representation
• Fingerprints

Information extraction
• Multivariate statistical analysis

Knowledge discovery
• Data mining and machine learning

Knowledge representation
• Visualization
Polydispersity Challenge in Nanomaterials

Polydispersity can be detrimental for high-performing applications.

Purification of polydisperse nanoparticle samples is expensive.
Data Analytics of Nanocarbons

Virtual structures relaxed using TB-DFT

Nanodiamonds

Graphenes
Archetypal Analysis (AA)

Finds a $k \times m$ matrix $Z$ that corresponds to the archetypal or "pure patterns" in the data in such a way that each data point can be represented as a mixture of those archetypes. In other words, the archetypal analysis yields the two $n \times k$ coefficient matrices $\alpha$ and $\beta$, which minimize the residual sum of squares:

$$\text{RSS} = \sum_{i=1}^{n} \left\| X_i - \sum_{j=1}^{k} \alpha_{ij} Z_i \right\|^2 = \sum_{i=1}^{n} \left\| X_i - \sum_{j=1}^{k} \alpha_{ij} \sum_{l=1}^{n} \beta_{ij} X_l \right\|^2$$

$$Z_j = \sum_{l=1}^{n} \beta_{jl} X_l$$


The predictors of $X_i$ are finite mixtures of archetypes $Z_j$, which are convex combinations of the observations.
Archetypal Analysis of Nanocarbons

Nanodiamonds

Graphene nanoflakes

Nanocarbons Prototypes

Nanodiamonds prototypes

Graphene prototypes

Estimation of Nanodiamonds Properties

Structural Diversity Challenge

Graphene nanoflakes

- Trigonal
- Rectangular
- Hexagonal

Defects, oxidation and edge passivation yield large structural diversity
Structural Diversity Challenge

Silicon qbits

Single Si substitution by P yields a large structural diversity
Structural Diversity Challenge

Metal-Organic Framework (MOF)

benzimidazole

nitroimidazole

Zn$^{2+}$

CO$_2$ capture and sequestration
(Science, 2008)

ZIF-68
Structural Diversity Challenge

Metal-Organic Framework (MOF)

In-silico Combinatorial design
Modification with 35 functional groups gives a total of ~1.5 million (35^4) unique combinations.
Machine Learning Approach

Machine learning prediction of functional properties

Feature fingerprints

Machine learning
Binary decision tree of the Band Gap of graphene

Accuracy 80%

Features
- Surface area
- Number of atoms
- Shape aspect ratio

Estimation of the graphene **Band Gap** from **topological features**

\[
\text{ATS}^L = \sum_{ij} \delta_{ij}^L \times P_i \times P_j
\]

where

\[
\delta_{ij}^L = \begin{cases} 
1 & \text{if } (L = d_{ij}) \\
0 & \text{otherwise}
\end{cases}
\]

\( P_i \) and \( P_j \), are the values of a bond order of the carbon atoms in graphene, while \( L \) is the topological distance, whilst \( \delta_{ij}^L \) is a delta function.

Machine Learning of Graphene

Radial Distribution Function (RDF) scores for graphene

\[
RDF(R) = \sum_{i,j}^N \exp^{-B(r_{ij} - R)^2}
\]

the summation is over the \( N \) atom pairs in the graphene structure, and \( r_{ij} \) is the distance of these pairs and \( B \) is a smoothing parameter set to 10.

Machine Learning vs. Atomistic Simulations

Energy of the Fermi level

Ionization Potential

From RDF scores

Machine Learning vs. Atomistic Simulations

Machine learning prediction of gas adsorption in MOF

Accuracy and Coverage Challenges

Accuracy of electronic calculations methods vs. system size

- **5000 atoms**: TBDF/Semiempirical
- **500 atoms**: Density Functional
- **100 atoms**: Quantum Monte Carlo Coupled Cluster
- **20 atoms**: Quantum Monte Carlo Coupled Cluster

System size vs. Accuracy

- 5000 atoms
- 500 atoms
- 100 atoms
- 20 atoms
Data-driven Challenge

Machine learning for large material spaces

Machine learning predictions:
- Functional property value or threshold
- Accuracy of different quantum-chemistry methods

\[ \Delta = f(\text{structure}) \]

Challenges and Limitations

Accuracy Gap Between Different Levels of Theory

6,095 isomers of $\text{C}_7\text{H}_{10}\text{O}_2$

$\Delta E$

B3LYP  QMC

big gap

Density

$\Delta H$ (kcal/mol)
Accuracy Gap Predictions

Machine learning calibration
Data-driven High-throughput Screening

- Input jobs
- Queue management
- Computational resources
- Resubmit or kill failed runs
- Finished runs
- Data storage
- Outputs
- Accuracy refinement

Data analytics
Self-organization Map (SOM) of NPs

Ag-NP → Electrostatic Potential → SOM
Deep Learning for Nanomaterials

Feature extraction

- Input 32x32
- Image 32x32
- 5x5 convolution
- 2x2 subsampling
- C₁ Feature maps 28x28
- S₁ Feature maps 14x14
- C₂ Feature maps 10x10
- S₂ Feature maps 5x5
- B₁
- Fully connected
- B₂
- Output

Classification

Fully connected
Thank you

Data61
Michael Fernandez
e  michael.fernandezllamosa@csiro.au
w  https://research.csiro.au/mmm/

www.data61.csiro.au