Machine learning for molecular modeling

Eric CANCES

Ecole des Ponts ParisTech and INRIA, Paris, France

Big data challenges for predictive modeling of complex systems

Hong Kong, November 26-30, 2018
1. Molecular simulation in a nutshell

2. Learning molecular properties, interatomic potentials and force fields
1. **Molecular simulation in a nutshell**

2. **Learning molecular properties, interatomic potentials and force fields**
   
   **Overview of three promising methods**
Outline of the talk

1. Molecular simulation in a nutshell

2. Learning molecular properties, interatomic potentials and force fields

   Overview of three promising methods

   • SOAP: Bartók, Payne, Kondor, Csanyi ’10 (kernel method)
1. Molecular simulation in a nutshell

2. Learning molecular properties, interatomic potentials and force fields

   Overview of three promising methods

   • SOAP: Bartók, Payne, Kondor, Csányi ’10 (kernel method)
   • Solid Harmonic Wavelet Scattering Transform: Mallat et al. ’17
1. Molecular simulation in a nutshell

2. Learning molecular properties, interatomic potentials and force fields

Overview of three promising methods

- SOAP: Bartók, Payne, Kondor, Csanyi ’10  (kernel method)
- Solid Harmonic Wavelet Scattering Transform: Mallat et al. ’17

Stylometry analysis of sculptures?
1. Molecular simulation in a nutshell

2. Learning molecular properties, interatomic potentials and force fields

Overview of three promising methods

- SOAP: Bartók, Payne, Kondor, Csanyi ’10  (kernel method)
- Solid Harmonic Wavelet Scattering Transform: Mallat et al. ’17
- Moment tensor models: Shapeev ’16  (high-dimensional interpolation)
1. Molecular simulation in a nutshell

2. Learning molecular properties, interatomic potentials and force fields
   
   Overview of three promising methods
   
   • SOAP: Bartók, Payne, Kondor, Csanyi ’10  (kernel method)
   • Solid Harmonic Wavelet Scattering Transform: Mallat et al. ’17
   • Moment tensor models: Shapeev ’16  (high-dimensional interpolation)

3. Conclusion and perspectives
1 - Molecular simulation in a nutshell
Molecular simulation

- one of the major application fields of scientific computing
  (1998 and 2013 Nobel prizes in Chemistry)
- based on a broad variety of physical models
- an inexhaustible source of problems for mathematicians and computer scientists

Ubiquitin (Filippo Lipparini, Pisa)
12 among the most cited 100 articles, including 2 in the top 10, are methodological papers that lay the foundations for molecular simulation.
PRACE HPC European projects funded in 2016

37% are molecular simulation projects
(33% of the overall computational time)

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Field</th>
<th>Countries</th>
<th>Awarded hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>CellPhy</td>
<td>Biology</td>
<td>FR, IT, USA</td>
<td>20.8 millions</td>
</tr>
<tr>
<td>MemAlloReg</td>
<td>Biomedicine</td>
<td>UK</td>
<td>17.8 millions</td>
</tr>
<tr>
<td>MAPK</td>
<td>Biomedicine</td>
<td>ES</td>
<td>31 millions</td>
</tr>
<tr>
<td>AmyGoNP</td>
<td>Biomedicine</td>
<td>IT</td>
<td>25 millions</td>
</tr>
<tr>
<td>ESCM</td>
<td>Materials</td>
<td>DE, NO, SE, UK</td>
<td>16 millions</td>
</tr>
<tr>
<td>AIMODIM</td>
<td>Materials</td>
<td>FR, SE</td>
<td>12 millions</td>
</tr>
<tr>
<td>SMoG</td>
<td>Materials</td>
<td>IT</td>
<td>15 millions</td>
</tr>
<tr>
<td>HTESO</td>
<td>Materials</td>
<td>FR, IT</td>
<td>9.9 millions</td>
</tr>
<tr>
<td>QuMOS</td>
<td>Materials</td>
<td>FR</td>
<td>10.6 millions</td>
</tr>
<tr>
<td>HowCP</td>
<td>Chemistry</td>
<td>CH</td>
<td>10 millions</td>
</tr>
<tr>
<td>PHHP</td>
<td>Physics</td>
<td>FR, IT, USA</td>
<td>65 millions</td>
</tr>
</tbody>
</table>
Some applications: materials design

Avere France

CSCV
Some applications: drug design

European Journal of Medicinal Chemistry
Volume 91, 16 February 2015, Pages 4–14
Molecular Dynamics: New Advances in Drug Discovery

Review article
Molecular dynamics in drug design

Hongtao Zhao, Amedeo Caflisch
Department of Biochemistry, University of Zurich, CH-8057 Zurich, Switzerland

Molecular dynamics validation of the binding mode of a ligand of the first bromodomain of BRD4 identified in silico by docking.
In vitro: IC₉₀ = 7.5 µM; ligand efficiency of 0.37 kcal/mol per heavy atom. PDB code: 4PCI
**Journal of Geophysical Research: Planets**

**REVIEW ARTICLE**  
10.1002/2016JE005080

**Special Section:**  
JGR-Planets 25th Anniversary

**Understanding Jupiter’s interior**

Burkhard Militzer\(^1,2\), François Soubiran\(^1\), Sean M. Wahl\(^1\), and William Hubbard\(^3\)

\(^1\)Department of Earth Planetary Science, University of California, Berkeley, California, USA, \(^2\)Department of Astronomy, University of California, Berkeley, California, USA, \(^3\)Lunar Planetary Laboratory, University of Arizona, Tucson, Arizona, USA

---

**Some applications:** phenomena out of reach of experimental approaches
Some applications: CAD for nanotechnologies

Xe atoms on Ni surface
(STM image - IBM, 1989)

http://www.thenanoage.com

Nature Chemistry (mars 2015)
Molecular dynamics within Born-Oppenheimer approximation

- atomic nuclei are considered as point-like classical particles
- the (quantum) state of the electrons is assumed to adapt instantly to the positions of the nuclei
- electrostatic interactions between nuclei and electrons

This approximation is used in 99% of molecular simulations

\[
\frac{d\mathbf{R}_k}{dt}(t) = \frac{1}{m_k} \mathbf{P}_k(t)
\]

\[
\frac{d\mathbf{P}_k}{dt}(t) = \mathbf{F}_k(t)
\]

\[
\mathbf{F}_k(t) = -\nabla_{\mathbf{R}_k} V(\mathbf{R}_1(t), \ldots, \mathbf{R}_M(t))
\]

\(V\): interatomic potential
(potential energy surface)
Molecular dynamics within Born-Oppenheimer approximation

- atomic nuclei are considered as point-like classical particles
- the (quantum) state of the electrons is assumed to adapt instantly to the positions of the nuclei
- electrostatic interactions between nuclei and electrons

This approximation is used in 99% of molecular simulations

\[
\frac{d\mathbf{R}_k(t)}{dt} = \frac{1}{m_k} \mathbf{P}_k(t)
\]

\[
d\mathbf{P}_k(t) = \mathbf{F}_k(t)dt - \gamma \mathbf{P}_k(t) dt + \sigma dW_t
\]

\[
\mathbf{F}_k(t) = -\nabla_{\mathbf{R}_k} V(\mathbf{R}_1(t), \cdots, \mathbf{R}_M(t))
\]

\(V\): interatomic potential
(potential energy surface)
Molecular dynamics within Born-Oppenheimer approximation

- atomic nuclei are considered as point-like classical particles
- the (quantum) state of the electrons is assumed to adapt instantly to the positions of the nuclei
- electrostatic interactions between nuclei and electrons

This approximation is used in 99% of molecular simulations
Molecular dynamics within Born-Oppenheimer approximation

- atomic nuclei are considered as point-like classical particles
- the (quantum) state of the electrons is assumed to adapt instantly to the positions of the nuclei
- electrostatic interactions between nuclei and electrons

This approximation is used in 99% of molecular simulations
Simulation of a molecular system

- state of the system at time $t$: positions et velocities of $M$ atomic nuclei
  $$(R_1(t), \cdots, R_M(t), P_1(t), \cdots, P_M(t))$$

- Newton’s laws of classical mechanics
  $$\frac{dR_k(t)}{dt} = \frac{1}{m_k} P_k(t), \quad \frac{dP_k(t)}{dt} = F_k(t)$$

- Schrödinger-Dirac-Born-Oppenheimer model
  $$F_k(t) = - \sum_{j \neq k} z_j z_k e^2 \frac{R_j(t) - R_k(t)}{4\pi\varepsilon_0 |R_j(t) - R_k(t)|^3} + z_k e^2 \int_{\mathbb{R}^3} \rho_{(R_1(t), \cdots, R_M(t))}(x) \frac{x - R_k(t)}{4\pi\varepsilon_0 |x - R_k(t)|^3} dx$$

  ground state electronic density when nuclei are in configuration $(R_1(t), \cdots, R_M(t))$
Simulation of a molecular system

- state of the system at time $t$: positions et velocities of $M$ atomic nuclei
  \[(R_1(t), \cdots, R_M(t), P_1(t), \cdots, P_M(t))\]

- Newton’s laws of classical mechanics
  \[
  \frac{dR_k}{dt}(t) = \frac{1}{m_k}P_k(t), \quad \frac{dP_k}{dt}(t) = F_k(t)
  \]

- Schrödinger-Dirac-Born-Oppenheimer model
  \[
  F_k(t) = -\sum_{j \neq k} z_j z_k e^2 \frac{R_j(t) - R_k(t)}{4\pi\varepsilon_0 |R_j(t) - R_k(t)|^3} + z_k e^2 \int_{\mathbb{R}^3} \rho_{(R_1(t), \cdots, R_M(t))}(x) \frac{x - R_k(t)}{4\pi\varepsilon_0 |x - R_k(t)|^3} dx
  \]

- simulation parameters
  - a few fundamental constants of physics: $\hbar, m_e, e, \varepsilon_0$;
  - the masses $m_1, \cdots, m_M$ and charges $z_1, \cdots, z_M$ of the $M$ nuclei;
  - no need of initial conditions
    \[(R_1(t_0), \cdots, R_M(t_0), V_1(t_0), \cdots, V_M(t_0)):\]

  search for equilibrium configurations
  computation of statistical averages in an ergodic setting
Main issue: curse of dimensionality

\[ \rho_{(R_1, \ldots, R_M)}^{\text{elec}}(x) = N \int_{\mathbb{R}^{3(N-1)}} |\Psi_{(R_1, \ldots, R_M)}(x, x_2, \ldots, x_N)|^2 \, dx_2 \cdots dx_N \]

solution of the electronic Schrödinger equation, a \(3N\)-dimensional PDE, where \(N\) is the number of electrons in the system (\(3N = 306\) for caffeine)

\[
\left( -\frac{\hbar^2}{2m_e} \sum_{i=1}^{N} \Delta_{x_i} - \sum_{i=1}^{N} \sum_{k=1}^{M} \frac{z_k e^2}{4\pi \varepsilon_0 |x_i - R_k|} + \sum_{i<j} \frac{e^2}{4\pi \varepsilon_0 |x_i - x_j|} \right) \Psi(x_1, \ldots, x_N) = E\Psi(x_1, \ldots, x_N)
\]

Pauli principle: \(\psi\) antisymmetric

Normalization condition: \(\|\Psi\|_{L^2(\mathbb{R}^{3N})} = 1\)
Main issue: curse of dimensionality

\[
\rho_{(R_1, \ldots, R_M)}^{\text{elec}}(x) = N \int_{\mathbb{R}^{3(N-1)}} |\Psi_{(R_1, \ldots, R_M)}(x, x_2, \ldots, x_N)|^2 \, dx_2 \cdots dx_N
\]

solution of the electronic Schrödinger equation, a \(3N\)-dimensional PDE, where \(N\) is the number of electrons in the system (\(3N = 306\) for caffeine)

Approximate, accurate electronic structure models (quantum chemistry)

- wavefunction methods (e.g. coupled-clusters)
- Density Functional Theory (DFT)
- Quantum Monte Carlo (QMC)

1998 Nobel Prize in Chemistry

Walter Kohn

John Pople
Main issue: curse of dimensionality

\[ \rho_{\text{elec}}^{(R_1,\cdots,R_M)}(x) = N \int_{\mathbb{R}^{3(N-1)}} |\Psi_{(R_1,\cdots,R_M)}(x, x_2, \cdots, x_N)|^2 \, dx_2 \cdots dx_N \]

solution of the electronic Schrödinger equation, a \(3N\)-dimensional PDE, where \(N\) is the number of electrons in the system (\(3N = 306\) for caffeine)

Approximate, accurate electronic structure models (quantum chemistry)

- wavefunction methods (e.g. coupled-clusters)
- Density Functional Theory (DFT)
- Quantum Monte Carlo (QMC)

... too costly for very large molecular systems or large sets of molecular systems
Main issue: curse of dimensionality

\[ \rho_{(R_1, \ldots, R_M)}^\text{elec}(x) = N \int_{\mathbb{R}^{3(N-1)}} |\Psi_{(R_1, \ldots, R_M)}(x, x_2, \ldots, x_N)|^2 \, dx_2 \cdots dx_N \]

solution of the electronic Schrödinger equation, a \(3N\)-dimensional PDE, where \(N\) is the number of electrons in the system (\(3N = 306\) for caffeine)

Approximate, accurate electronic structure models (quantum chemistry)

- wavefunction methods (e.g. coupled-clusters)
- Density Functional Theory (DFT)
- Quantum Monte Carlo (QMC)

... too costly for very large molecular systems or large sets of molecular systems

Empirical interatomic potentials (CHARMM, Amber, EAM...) are much faster

\[ V_{\text{LJ}}(R_1, \ldots, R_N) = \sum_{k < l} 4\varepsilon \left( \left( \frac{|R_k - R_l|}{d} \right)^{12} - \left( \frac{|R_k - R_l|}{d} \right)^6 \right) \]  
LJ pot. for rare gases
Main issue: curse of dimensionality

\[ \rho_{(R_1, \ldots, R_M)}(x) = N \int_{\mathbb{R}^{3(N-1)}} |\Psi_{(R_1, \ldots, R_M)}(x_1, x_2, \ldots, x_N)|^2 \, dx_2 \cdots dx_N \]

solution of the electronic Schrödinger equation, a \(3N\)-dimensional PDE, where \(N\) is the number of electrons in the system (\(3N = 306\) for caffeine)

Approximate, accurate electronic structure models (quantum chemistry)

- wavefunction methods (e.g. coupled-clusters)
- Density Functional Theory (DFT)
- Quantum Monte Carlo (QMC)

... too costly for very large molecular systems or large sets of molecular systems

Empirical interatomic potentials (CHARMM, Amber, EAM...) are much faster

... but lack of accuracy and transferability
2 - Learning molecular properties, interatomic potentials and force fields
Input:

- atomic types and positions \( x = (z_k, R_k)_{1 \leq i \leq M} \in (\mathbb{N}^* \times \mathbb{R}^3)^M \)
- charge of the molecular system (assumed neutral here)
Input:

- atomic types and positions \( x = (z_k, R_k)_{1 \leq i \leq M} \in (\mathbb{N}^* \times \mathbb{R}^3)^M \)
- charge of the molecular system (assumed neutral here)

Learning interatomic potentials and force fields: outputs

- the value of the interatomic potential \( V(x) \in \mathbb{R} \)
- and/or the values of the interatomic forces \( (F_k(x))_{1 \leq k \leq M} \in \mathbb{R}^{3M} \)

\[
x(t) = (z_k, R_k(t))_{1 \leq i \leq M}, \quad m_k \frac{d^2 R_k(t)}{dt^2} = F_k(x(t)), \quad F_k(x) = -\nabla_{R_k} V(x)
\]
Input:

- atomic types and positions $x = (z_k, R_k)_{1 \leq i \leq M} \in (\mathbb{N}^* \times \mathbb{R}^3)^M$
- charge of the molecular system (assumed neutral here)

Learning interatomic potentials and force fields: outputs

- the value of the interatomic potential $V(x) \in \mathbb{R}$
- and/or the values of the interatomic forces $(F_k(x))_{1 \leq k \leq M} \in \mathbb{R}^{3M}$

\[
x(t) = (z_k, R_k(t))_{1 \leq i \leq M}, \quad m_k \frac{d^2 R_k(t)}{dt^2} = F_k(x(t)), \quad F_k(x) = -\nabla_{R_k} V(x)
\]

Learning molecular properties: possible outputs

- atomization energy (molecule) or formation energy (crystal)
- specific chemical, electronic or optical properties

In this case, $x = (z_k, R_k)_{1 \leq i \leq M} \in (\mathbb{N}^* \times \mathbb{R}^3)^M$ is assumed to describe a (meta)stable configuration of the system
Generating data

- quantum chemistry calculations
  - massive, easy-to-generate consistent data (but possibly wrong)
Generating data

- quantum chemistry calculations
  - massive, easy-to-generate consistent data (but possibly wrong)
- experiments
  - less abundant, more noisy
Generating data

- quantum chemistry calculations
  - massive, easy-to-generate consistent data (but possibly wrong)
- experiments
  - less abundant, more noisy

Collecting data

- quantum chemistry databases
  e.g. QM9 database: equilibrium geometries of 133,855 molecules containing of up to five different elements (H, C, N, O, and F)
Generating data

- quantum chemistry calculations
  - massive, easy-to-generate consistent data (but possibly wrong)
- experiments
  - less abundant, more noisy

Collecting data

- quantum chemistry databases
  e.g. QM9 database: equilibrium geometries of 133,855 molecules containing of up to five different elements (H, C, N, O, and F)
- repositories
  - about 25,000 scientific papers containing first-principle calculations published in 2018
  - initiatives to collect the outputs in public repositories
    e.g. NOMAD https://nomad-coe.eu (launched in 2015) on Nov. 18, 2018: 44,249,030 open access total energy calculations
Machine learning in molecular simulation (some refs. among many others)

- Neural networks for fitting empirical interatomic potentials
  Sumper and Noid ’92...
  Behler and Parrinello ’07...
  Smith, Isayev and Roitberg ’17

- Kernel methods for approximating interatomic potentials
  Ho and Rabitz ’96...
  Bartók, Payne, Kondor, Csányi ’10 (SOAP)...

- Neural networks for computing properties
  Hu et al. ’03...

- Kernels for computing properties
  Rupp, Tkatchenko, Müller, von Lilienfeld ’12 (Coulomb matrix)...

- Moment tensor models for fitting interatomic potentials and forces
  Shapeev ’16... (Moment Tensor Potentials)

- Image-processing based methods for computing properties
  Hirn, Mallat, Poilvert ’17... (Solid Harmonic Wavelet Scattering Transform)
Machine learning in molecular simulation (some refs. among many others)

- Neural networks for fitting empirical interatomic potentials
  Sumper and Noid ’92...
  Behler and Parrinello ’07...
  Smith, Isayev and Roitberg ’17

- Kernel methods for approximating interatomic potentials
  Ho and Rabitz ’96...
  Bartók, Payne, Kondor, Csányi ’10 (SOAP)...

- Neural networks for computing properties
  Hu et al. ’03...

- Kernels for computing properties
  Rupp, Tkatchenko, Müller, von Lilienfeld ’12 (Coulomb matrix)...

- Moment tensor models for fitting interatomic potentials and forces
  Shapeev ’16... (Moment Tensor Potentials)

- Image-processing based methods for computing properties
  Hirn, Mallat, Poilvert ’17... (Solid Harmonic Wavelet Scattering)
**Invariance property** (for scalar outputs $f(x)$)

- **Euclidean group**

\[ \forall g \in E(3), \quad f(gx) = f(x), \quad x = (z_k, R_k)_{1 \leq i \leq M}, \quad gx = (z_k, gR_k)_{1 \leq i \leq M} \]

- **relabeling of identical atoms**

  if $z_i = z_j$ for some $i < j$, then $f(\tau_{ij}x) = f(x)$,

where $\tau_{ij}$ swaps the positions of atoms $i$ and $j$
Invariance property (for scalar outputs $f(x)$)

- Euclidean group

$$\forall g \in E(3), \quad f(gx) = f(x), \quad x = (z_k, R_k)_{1 \leq i \leq M}, \quad gx = (z_k, gR_k)_{1 \leq i \leq M}$$

- relabeling of identical atoms

if $z_i = z_j$ for some $i < j$, then $f(\tau_{ij}x) = f(x)$,

where $\tau_{ij}$ swaps the positions of atoms $i$ and $j$

Physically relevant approximations of $f$ must satisfy these invariance properties
The locality assumption

\[ f(x) \approx \sum_{k=1}^{M} f_{z_k}(\mathcal{N}_k(x)), \quad \mathcal{N}_k(x) = (z_j, R_j - R_k)_{j\neq k}, |R_j - R_k| \leq R_{\text{cut}} \]

\( \mathcal{N}_k(x) \): local environment around atom \( k \), \( R_{\text{cut}} \): cut-off radius

The approximation of \( f \) then is translation invariant by construction

If \( f = V \), \( f_{z_k}(\mathcal{N}_k(x)) \) can be interpreted as the site energy of atom \( k \)
The locality assumption

\[ f(x) \simeq \sum_{k=1}^{M} f_{z_k}(\mathcal{N}_k(x)), \quad \mathcal{N}_k(x) = (z_j, R_j - R_k)_{j \neq k}, \quad |R_j - R_k| \leq R_{\text{cut}} \]

\( \mathcal{N}_k(x) \): local environment around atom \( k \), \( R_{\text{cut}} \): cut-off radius

The approximation of \( f \) then is translation invariant by construction

If \( f = V \), \( f_{z_k}(\mathcal{N}_k(x)) \) can be interpreted as the site energy of atom \( k \)

The locality assumption (for "reasonable" values of \( R_{\text{cut}} \))

- can be proved for some crude models (e.g. tight-binding), see Ortner et al. ’16
- is only valid for some classes of real molecular systems
Smooth overlap of atomic positions (SOAP) (Csányi et al.)

- based on the locality assumption

\[
f(x) \simeq \sum_{k=1}^{M} f_{z_k}(N_k(x)), \quad N_k(x) = (z_j, \mathbf{R}_j - \mathbf{R}_k)_{j \neq k}, |R_j - R_k| \leq R_{\text{cut}}
\]

\[\rightarrow \quad \text{translational invariance property}\]
Smooth overlap of atomic positions (SOAP) (Csányi et al.)

- based on the locality assumption

\[ f(x) \simeq \sum_{k=1}^{M} f_{z_k}(\mathcal{N}_k(x)), \quad \mathcal{N}_k(x) = (z_j, R_j - R_k)_{j \neq k}, |R_j - R_k| \leq R_{\text{cut}} \]

- SOAP density around atom \( k \)

\[ \rho_{x,k}(r) = \sum_{(z_j, r_j) \in \mathcal{N}_k(x)} c(z_j) \exp \left( -\frac{|r - r_j|^2}{2\sigma^2} \right), \quad c(z_j) > 0, \quad \sigma > 0 \]

\[ \longrightarrow \quad \text{atom relabeling invariance property} \]
Smooth overlap of atomic positions (SOAP) (Csányi et al.)

- based on the locality assumption

\[ f(x) \simeq \sum_{k=1}^{M} f_{z_k}(\mathcal{N}_k(x)) , \quad \mathcal{N}_k(x) = (z_j, R_j - R_k)_{j \neq k}, |R_j - R_k| \leq R_{\text{cut}} \]

- SOAP density around atom \( k \)

\[ \rho_{x,k}(r) = \sum_{(z_j, r_j) \in \mathcal{N}_k(x)} c(z_j) \exp \left( -\frac{|r - r_j|^2}{2\sigma^2} \right), \quad c(z_j) > 0, \quad \sigma > 0 \]

- kernel method for approximating \( f_{z_k}(\mathcal{N}_k(x)) \)

\[ f_{z_k}(\mathcal{N}_k(x)) = \sum_{n=1}^{N} \alpha_{z_k,i} K \left( \rho_{x(i),k}, \rho_{x,k} \right), \quad (x_i)_{1 \leq i \leq N} \text{ training set} \]

(\( \alpha_{z_k,i} \) obtained by Gaussian process regression)
Smooth overlap of atomic positions (SOAP) (Csányi et al.)

- based on the locality assumption

\[ f(x) \simeq \sum_{k=1}^{M} f_{z_k}(N_k(x)), \quad N_k(x) = (z_j, R_j - R_k)_{j\neq k}, |R_j-R_k| \leq R_{\text{cut}} \]

- SOAP density around atom \( k \)

\[ \rho_{x,k}(r) = \sum_{(z_j,r_j) \in N_k(x)} c(z_j) \exp \left( -\frac{|r-r_j|^2}{2\sigma^2} \right), \quad c(z_j) > 0, \quad \sigma > 0 \]

- kernel method for approximating \( f_{z_k}(N_k(x)) \)

\[ f_{z_k}(N_k(x)) = \sum_{n=1}^{N} \alpha_{z_k,i} K \left( \rho_{x(i),k}, \rho_{x,k} \right), \quad (x_i)_{1 \leq i \leq N} \text{ training set} \]

- rotational invariant covariance kernels between two local environments respectively described by the densities \( \rho \) and \( \rho' \) can be defined as

\[ K_{\nu,\zeta}(\rho, \rho') = \left( \frac{k_\nu(\rho, \rho')}{\sqrt{k_\nu(\rho, \rho') k_\nu(\rho', \rho')}} \right)^\zeta, \quad k_\nu(\rho, \rho') = \int_{\text{SO}(3)} \left( \int_{\mathbb{R}^3} \rho(r) \rho'(Rr) \, dr \right)^\nu d\mu_H(R), \]

where \( \nu \in \mathbb{N}^* \) (in practice \( \nu = 2 \)), \( \zeta \in \mathbb{N}^* \), \( \mu_H \) is the Haar measure on \( \text{SO}(3) \).
SOAP: practical calculation of $k_2(\rho, \rho')$

- expansion of $\rho$ and $\rho'$ in an orthonormal basis of $L^2(\mathbb{R}^3)$

\[
\rho(r) = \sum_{nlm} c_{nlm}[\rho] \ g_n(r) \ Y_{lm} \left( \frac{r}{|r|} \right), \quad \rho'(r) = \sum_{nlm} c_{nlm}[\rho'] \ g_n(r) \ Y_{lm} \left( \frac{r}{|r|} \right)
\]

$(g_n)_{n \in \mathbb{N}}$ orthonormal basis of $L^2(\mathbb{R}_+; r^2 \, dr)$, $Y_{lm}$ spherical harmonics
SOAP: practical calculation of $k_2(\rho, \rho')$

- expansion of $\rho$ and $\rho'$ in an orthonormal basis of $L^2(\mathbb{R}^3)$

$$
\rho(r) = \sum_{nlm} c_{nlm}[\rho] \ g_n(r) \ Y_{lm} \left( \frac{r}{|r|} \right), \quad \rho'(r) = \sum_{nlm} c_{nlm}[\rho'] \ g_n(r) \ Y_{lm} \left( \frac{r}{|r|} \right)
$$

$(g_n)_{n \in \mathbb{N}}$ orthonormal basis of $L^2(\mathbb{R}_+; r^2 \, dr)$, $Y_{lm}$ spherical harmonics

- it holds

$$
k_2(\rho, \rho') = \sum_{n,n',l} p_{n,n',l}[\rho] \ p_{n,n',l}[\rho'], \quad p_{n,n',l}[\rho] = \sum_{m=-l}^{l} c_{n,l,m}[\rho] \ c_{n',l,m}[\rho]
$$
SOAP: practical calculation of $k_2(\rho, \rho')$

- expansion of $\rho$ and $\rho'$ in an orthonormal basis of $L^2(\mathbb{R}^3)$

$$\rho(r) = \sum_{nlm} c_{nlm}[\rho] g_n(r) Y_{lm} \left( \frac{r}{|r|} \right), \quad \rho'(r) = \sum_{nlm} c_{nlm}[\rho'] g_n(r) Y_{lm} \left( \frac{r}{|r|} \right)$$

$(g_n)_{n \in \mathbb{N}}$ orthonormal basis of $L^2(\mathbb{R}_+; r^2 \, dr)$, $Y_{lm}$ spherical harmonics

- it holds

$$k_2(\rho, \rho') = \sum_{n,n',l} p_{n,n',l}[\rho] p_{n,n',l}[\rho'], \quad p_{n,n',l}[\rho] = \sum_{m=-l}^{l} c_{n,l,m}[\rho] c_{n',l,m}[\rho]$$

- if $\rho(r) = \sum_{j} c(z_j) \exp \left( -\frac{|r - r_j|^2}{2\sigma^2} \right)$ is a sum of Gaussian functions, then

$$c_{n,l,m}[\rho] = \int_{0}^{+\infty} \rho_{l,m}(r) g_n(r) \, dr,$$

where, denoting by $i_l$ the $l^{th}$ modified spherical Bessel function of the 1$^{st}$ kind,

$$\rho_{l,m}(r) = \sum_{j} 4\pi Z_j \exp \left( -\frac{r^2 + |r_j|^2}{2\sigma^2} \right) i_l \left( \frac{r|r_j|}{\sigma^2} \right) Y_{lm} \left( \frac{r_j}{|r_j|} \right)$$
Solid Harmonic Wavelet Scattering Transform (Mallat et al.)

- associate to any configuration $x$ a family of surrogate densities $\rho_x : \mathbb{R}^3 \rightarrow \mathbb{R}_+^n$

→ identical atom relabeling invariance property
Solid Harmonic Wavelet Scattering Transform (Mallat et al.)

• associate to any configuration $x$ a family of surrogate densities $\rho_x : \mathbb{R}^3 \rightarrow \mathbb{R}_+^n$

• consider the solid harmonic wavelets

$$\psi_{j,l}^m(\mathbf{r}) = 2^{-3j} \psi_l^m(2^{-j} \mathbf{r}), \quad \psi_l^m(\mathbf{r}) = \frac{1}{(2\pi)^{3/2}} e^{-|\mathbf{r}|^2/2} |\mathbf{r}|^l Y_l^m \left( \frac{\mathbf{r}}{|\mathbf{r}|} \right)$$
Solid Harmonic Wavelet Scattering Transform (Mallat et al.)

- associate to any configuration $x$ a family of surrogate densities $\rho_x : \mathbb{R}^3 \rightarrow \mathbb{R}^n_+$
- consider the solid harmonic wavelets
  $$\psi_{j,l}^m(r) = 2^{-3j} \psi_l^m(2^{-j} r), \quad \psi_l^m(r) = \frac{1}{(2\pi)^{3/2}} e^{-|r|^2/2} |r|^l Y_l^m \left( \frac{r}{|r|} \right)$$
- for $\rho : \mathbb{R}^3 \rightarrow \mathbb{R}_+$, consider the first and second-order wavelet transforms
  $$U[j,l] \rho(r) := \left( \sum_{m=-l}^l |(\rho \ast \psi_{j,l}^m)(r)|^2 \right)^{1/2}, \quad U[j,j',l] \rho(r) := \left( \sum_{m=-l}^l |(U[j,l] \rho \ast \psi_{j',l}^m)(r)|^2 \right)^{1/2}$$
  for all $g \in E(3)$, $\pi_g \circ U[j,l] = U[j,l] \circ \pi_g$ and $\pi_g \circ U[j,j',l] = U[j,j',l] \circ \pi_g$
**Solid Harmonic Wavelet Scattering Transform** (Mallat et al.)

- associate to any configuration $x$ a family of surrogate densities $\rho_x : \mathbb{R}^3 \to \mathbb{R}_+$
- consider the solid harmonic wavelets
  \[ \psi_{j,l}^m(r) = 2^{-3j} \psi_l^m(2^{-j} r), \quad \psi_l^m(r) = \frac{1}{(2\pi)^{3/2}} e^{-|r|^2/2} |r|^l Y_l^m \left( \frac{r}{|r|} \right) \]
- for $\rho : \mathbb{R}^3 \to \mathbb{R}_+$, consider the first and second-order wavelet transforms
  \[ U[j, l] \rho(r) := \left( \sum_{m=-l}^l |(\rho * \psi_{j,l}^m)(r)|^2 \right)^{1/2}, \quad U[j, j', l] \rho(r) := \left( \sum_{m=-l}^l |(U[j, l] \rho * \psi_{j',l}^m)(r)|^2 \right)^{1/2} \]
  for all $g \in E(3)$, $\pi_g \circ U[j, l] = U[j, l] \circ \pi_g$ and $\pi_g \circ U[j, j', l] = U[j, j', l] \circ \pi_g$
- compute the first and second-order $\mathbb{R}_+$-valued coefficients
  \[ S_{\rho, \mu}[j, l, q] = \int_{\mathbb{R}^3} |U[j, l] \rho_{\mu}(r)|^q \, dr, \quad S_{\rho, \mu}[j, j', l, q] = \int_{\mathbb{R}^3} |U[j, j', l] \rho_{\mu}(r)|^q \, dr \]
  $1 \leq \mu \leq n, \quad 0 \leq l \leq L - 1, \quad 0 \leq j, j' \leq J - 1, \quad j < j'$, $q \in Q = \{1/2, 1, 2, 3, 4\}$ (e.g.)

$\rightarrow$ Euclidean invariance property
Solid Harmonic Wavelet Scattering Transform (Mallat et al.)

- associate to any configuration $x$ a family of surrogate densities $\rho_x : \mathbb{R}^3 \to \mathbb{R}_+$
- consider the solid harmonic wavelets
  $\psi_{j,l}^m(r) = 2^{-3j} \psi_l^m(2^{-j} r), \quad \psi_l^m(r) = \frac{1}{(2\pi)^{3/2}} e^{-|r|^2/2} |r|^l Y_l^m \left( \frac{r}{|r|} \right)$
- for $\rho : \mathbb{R}^3 \to \mathbb{R}_+$, consider the first and second-order wavelet transforms
  $U[j,l] \rho(r) := \left( \sum_{m=-l}^l |(\rho \ast \psi_{j,l}^m)(r)|^2 \right)^{1/2}$,  $U[j,j',l] \rho(r) := \left( \sum_{m=-l}^l |(U[j,l] \rho \ast \psi_{j',l}^m)(r)|^2 \right)^{1/2}$
  for all $g \in E(3)$, $\pi_g \circ U[j,l] = U[j,l] \circ \pi_g$ and $\pi_g \circ U[j,j',l] = U[j,j',l] \circ \pi_g$
- compute the first and second-order $\mathbb{R}_+$-valued coefficients
  $S_{\rho_x,\mu}[j,l,q] = \int_{\mathbb{R}^3} |U[j,l] \rho_{x,\mu}(r)|^q \, dr$,  $S_{\rho_x,\mu}[j,j',l,q] = \int_{\mathbb{R}^3} |U[j,j',l] \rho_{x,\mu}(r)|^q \, dr$
  $1 \leq \mu \leq n$,  $0 \leq l \leq L - 1$,  $0 \leq j, j' \leq J - 1$,  $j < j'$,  $q \in Q = \{1/2, 1, 2, 3, 4\}$ (e.g.)
- collect all these coefficients into a vector $S \rho_x \in \mathbb{R}_+^N$,  $N = n \times L \times \frac{J(J+1)}{2} \times |Q|$  $\longrightarrow$ invariance w.r.t. Euclidean transforms and identical atom relabeling
Solid Harmonic Wavelet Scattering Transform (Mallat et al.)

- associate to any configuration $x$ a family of surrogate densities $\rho_x : \mathbb{R}^3 \rightarrow \mathbb{R}_+$
- consider the solid harmonic wavelets
  \[
  \psi^m_{j,l}(r) = 2^{-3j} \psi^m_l(2^{-j} r), \quad \psi^m_l(r) = \frac{1}{(2\pi)^{3/2}} e^{-|r|^2/2} |r|^l Y^m_l \left( \frac{r}{|r|} \right)
  \]
- for $\rho : \mathbb{R}^3 \rightarrow \mathbb{R}_+$, consider the first and second-order wavelet transforms

\[
U[j, l] \rho(r) := \left( \sum_{m=-l}^{l} |(\rho \ast \psi^m_{j,l})(r)|^2 \right)^{1/2}, \quad U[j, j', l] \rho(r) := \left( \sum_{m=-l}^{l} |(U[j, l] \rho \ast \psi^m_{j',l})(r)|^2 \right)^{1/2}
\]

for all $g \in E(3)$, $\pi_g \circ U[j, l] = U[j, l] \circ \pi_g$ and $\pi_g \circ U[j, j', l] = U[j, j', l] \circ \pi_g$
- compute the first and second-order $\mathbb{R}_+$-valued coefficients

\[
S_{\rho_x, \mu}[j, l, q] = \int_{\mathbb{R}^3} |U[j, l] \rho_{x, \mu}(r)|^q \, dr, \quad S_{\rho_x, \mu}[j, j', l, q] = \int_{\mathbb{R}^3} |U[j, j', l] \rho_{x, \mu}(r)|^q \, dr
\]

$1 \leq \mu \leq n$, $0 \leq l \leq L - 1$, $0 \leq j, j' \leq J - 1$, $j < j'$, $q \in Q = \{1/2, 1, 2, 3, 4\}$ (e.g.)
- collect all these coefficients into a vector $S\rho_x \in \mathbb{R}_+^N$, $N = n \times L \times \frac{J(J+1)}{2} \times |Q|$
- find the best approximation of $f(x)$ in the space of $r$-multilinear functions of the entries of $S\rho_x$
Solid Harmonic Wavelet Scattering Transform (continued)

- example of densities $\rho_{x,\mu}$
  
  - sum of atomic-like densities
    
    $$\rho_{x,\mu}(r) = \sum_{k=1}^{M} c_{\mu}(z_k) g(r - R_k) \in \mathbb{R}^n, \quad g(r) = \frac{1}{(2\pi\sigma^2)^{3/2}} e^{-\frac{|r|^2}{2\sigma^2}}, \quad c(z) \in \mathbb{R}^n$$
    
    where $c_{\mu}(z_k)$ can be equal to $z_k$, or to the number of core (resp. valence) electrons in the chemical element of atomic number $z_k$;

  - sum of bond-like densities
    
    $$\rho_{x}^{\text{bonds}}(r) = C \sum_{(i,j) \in B} \frac{\gamma_{ij}}{|R_i - R_j|} \exp \left(-\frac{d_{ij}(r)^2}{2d_0^2}\right)$$
Solid Harmonic Wavelet Scattering Transform (continued)

- **example of densities** $\rho_{x,\mu}$
  - sum of atomic-like densities

  $$\rho_{x,\mu}(r) = \sum_{k=1}^{M} c_\mu(z_k) g(r - R_k) \in \mathbb{R}^n, \quad g(r) = \frac{1}{(2\pi\sigma^2)^{3/2}} e^{-\frac{|r|^2}{2\sigma^2}}, \quad c(z) \in \mathbb{R}^n$$

  where $c_\mu(z_k)$ can be equal to $z_k$, or to the number of core (resp. valence) electrons in the chemical element of atomic number $z_k$;

  - sum of bond-like densities

  $$\rho_{x}^{\text{bonds}}(r) = C \sum_{(i,j) \in B} \frac{\gamma_{ij}}{|R_i - R_j|} \exp \left( -\frac{d_{ij}(r)^2}{2d_0^2} \right)$$

- **trilinear regression models** give excellent results on QM9
Solid Harmonic Wavelet Scattering Transform (continued)

- **example of densities** $\rho_{x,\mu}$
  
  - sum of atomic-like densities
    \[
    \rho_{x,\mu}(r) = \sum_{k=1}^{M} c_{\mu}(z_k) g(r - R_k) \in \mathbb{R}^{n}, \quad g(r) = \frac{1}{(2\pi\sigma^2)^{3/2}} e^{-\frac{|r|^2}{2\sigma^2}}, \quad c(z) \in \mathbb{R}^{n}
    \]
    where $c_{\mu}(z_k)$ can be equal to $z_k$, or to the number of core (resp. valence) electrons in the chemical element of atomic number $z_k$;
  
  - sum of bond-like densities
    \[
    \rho_{x,\mu}^{\text{bonds}}(r) = C \sum_{(i,j)\in B} \frac{\gamma_{ij}}{|R_i - R_j|} \exp \left( -\frac{d_{ij}(r)^2}{2d_0^2} \right)
    \]

- **trilinear regression models** give excellent results on QM9

- **parameter optimization**: quadratic loss function minimization by Adam’s stochastic gradient
Local Moment Tensor Models (Shapeev ’16)

- based on the locality assumption

\[ f(x) \simeq \sum_{k=1}^{M} f_{z_k}(N_k(x)), \quad N_k(x) = (z_j, R_j - R_k)_{j \neq k}, |R_j - R_k| \leq R_{cut} \]
Local Moment Tensor Models (Shapeev ’16)

- based on the locality assumption

\[
f(x) \simeq \sum_{k=1}^{M} f_{z_k}(\mathcal{N}_k(x)), \quad \mathcal{N}_k(x) = (z_j, R_j - R_k)_{j \neq k}, |R_j - R_k| \leq R_{\text{cut}}
\]

- approximation of the functions \( f_z \) (systematically improvable in principle)

\[
f_z(n) = \sum_{\alpha \leq N_1} \xi_{z,\alpha} B_{z,\alpha}(n), \quad n = (z_j, r_j)_j
\]

\[
B_{z,0}(n) = M_{z,0,0}(n), \quad B_{z,1}(n) = M_{z,0,1}(n) \cdot M_{z,0,1}(n), \quad B_{z,2}(n) = M_{z,0,0}(n) (M_{z,0,2}(n) : M_{z,0,2}(n)),
\]

\[
\ldots
\]

\[
M_{z,\mu,\nu}(n) = \sum_j f_{z,\mu}(z_j, |r_j|) r_j \otimes \cdots \otimes r_j, \quad f_{z,\mu}(z', r') = \sum_{k \leq N_2} c_{\mu,z,z'}^{(k)} T_k(r') (R_{\text{cut}} - r')^2,
\]

\( T_k \) Chebyshev polynomials on the interval \([R_{\text{min}}, R_{\text{cut}}]\)

\( R_{\text{min}} > 0 \): minimal distance between 2 atoms

- model parameters: \( \theta = (\xi_{z,\alpha}, c_{\mu,z,z'}^{(k)}) \)
Nonlocal Moment Tensor Models (Shapeev et al.)

- polynomial interpolation of several local models

\[
f_\gamma(x) = \sum_{k=1}^{M} f_{\gamma,k}(N_k(x)), \quad 1 \leq \gamma \leq \Gamma
\]

\[
f(x) = \sum_{\beta \in \mathbb{N}^\Gamma, ||\beta|| \leq m} p_\beta f_1^{\beta_1}(x) \cdots f_\Gamma^{\beta_\Gamma}(x)
\]

- model parameters: \( \theta = (\xi_{\gamma,z,\alpha, c_{\gamma,\mu,z,z',p_\beta}}) \)
Nonlocal Moment Tensor Models (Shapeev et al.)

- polynomial interpolation of several local models

\[ f_\gamma(x) = \sum_{k=1}^{M} f_{\gamma,k}(N_k(x)), \quad 1 \leq \gamma \leq \Gamma \]

\[ f(x) = \sum_{\beta \in \mathbb{N}^\Gamma, ||\beta|| \leq m} p_\beta f_{1}^{\beta_1}(x) \cdots f_{\Gamma}^{\beta_\Gamma}(x) \]

- model parameters: \( \theta = (\xi_{\gamma,z,\alpha}, c^{(k)}_{\gamma,\mu,z,z'}, p_\beta) \)

Moment Tensor Models

- seem to need less training data than SOAP to achieve chemical accuracy

  e.g.: QM9 database of 130k molecules: tens of thousands of samples for SOAP, a few thousands of samples for MTM

  rule of thumb: \( n \geq 2m \), \( n \): # of training data, \( m \): # of model parameters

- do not require regularization
Conclusions and perspectives
Machine learning for molecular simulation

- excellent results on small, simple molecular systems (e.g. QM9 database)
- some promising results on slightly more complex problems
- still a long way to go to address complex systems,

but things are moving fast

Another interesting application of ML: learning functionals for DFT

- Burke et al.
- Wu, Zhou, Xu
- ...