Gaussian Approximation Potential for Multiscale Methods: Investigation of Dynamics of Organic Molecules in Enzymatic Environment

Letif Mones
Mathematics Institute, University of Warwick
Engineering Department, University of Cambridge
l.mones@warwick.ac.uk
Outline

- QM/MM methods for biomolecular simulations
- Gaussian Approximation Potential
- GAP for organic molecules
- Alternative approach: application of GAP(QM)/MM method
QM/MM methods for biomolecular simulations

- Different level of description of all-atom systems:

  **MM level**
  + computationally fast \([O(N \log(N))]\)
  + relatively large systems (~10^6 atoms)
  + relatively long simulation times (~ms)
  + ok with unchanged valence state
  + folding, solvation, binding free energy
  - low chemical accuracy (polarisation effect, reactions, metal ions)

  **QM level**
  + chemical accuracy
  + electronic degrees of freedom are partly or completely taken into account
  + bond formation / cleavage
  - computationally very demanding \([O(n^3)]\)

- Application areas:
  - Investigation of enzymatic reactions (rational enzyme design)
  - Examination of subtle binding phenomena (drug discovery)
Gaussian process regression: theory

- Problems with high level QM:
  - limited number of atoms in QM region
  - performing extensive MD is still challenging

- A possible solution:
  - construct a non-parametric machine learning based potential (GAP)
  - training ~ $O(N^3)$ but prediction ~ $O(N)$

- Gaussian process regression:
  - target function: $f = f(x), f: \mathbb{R}^D \rightarrow \mathbb{R}, x \in \mathbb{R}^D$
  - training data set: $\mathcal{D} = \{X, f, \sigma^2\} = \{x_n, f_n, \sigma_n^2\}_{n=1}^N$
  - Gaussian process: $P(f_1, f_2, \ldots) = \mathcal{N}(0, \mathbf{C})$
  - Bayes’ theorem: $P(f(x^*)|\mathcal{D}) = \frac{P(\mathcal{D}|f(x^*))P(f(x^*))}{P(\mathcal{D})}$
Gaussian process regression: prediction

- Gaussian process regression:
  - prediction: \( \mathbb{E}(f(x^*)) = k^T(K + \sigma^2 I)^{-1}f \)
    \[ \nabla (f(x^*)) = \kappa - (K + \sigma^2 I)^{-1}k \]
  - covariance matrix elements: \( (K)_{ij} = \text{cov}(f(x_i), f(x_j)) \)
  - covariance function: \( \text{cov}_{SE} = \delta^2 \exp \left( -\frac{1}{2} \sum_{k=1}^{D} \frac{(x_i^{(k)} - x_j^{(k)})^2}{\theta_k^2} \right) \)
  - hyperparameters: \( \delta^2 \) typical variance of the function
    \[ \{\theta_k\}_{k=1}^{D} \] smoothness of the function / uncorrelation length scale
    \[ \{\sigma^2_n\}_{n=1}^{N} \] assumed noise / variance of data
  - generalised input / output: function \text{training} from gradient or \text{gradient+function} observables
    - prediction of function gradient
Gaussian Approximation Potential

- Extension for atomistic systems:
  - crystalline materials (diamond, silicon, germanium)
  - water, hydrated ions and gases

QM calculations \( \rightarrow \) GAP fitting \( \rightarrow \) GAP molecular dynamics

- \( N_a \): number of atoms
- \( N \): configurations

\[
\{ E(x_n), F(x_n) \}_{n=1}^{N}
\]

\[
x_n \in \mathbb{R}^{3N_a}
\]

- descriptor:
  \[
  \xi = \xi(x), \xi : \mathbb{R}^{3N} \rightarrow (\mathbb{R}/\mathbb{T})^D
  \]
  - invariant to symmetry operations (permutation, reflection, rotation, translation)

- prediction:
  - energy: \( \mathbb{E}(E(\xi(x^*))) \)
  - forces: \( \mathbb{E}(-\nabla E(\xi(x^*))) \)

- training data:

\[
D = \left\{ \begin{array}{c}
\xi_n = \xi(x_n) \\
E(\xi_n) \\
F(\xi_n) \\
\sigma_n^2(E) \\
\sigma_n^2(F)
\end{array} \right\}_{n=1}^{N}
\]
GAP for organic molecules

- **Test system:**

- **training set:** 500 K MD on MNDO

- **descriptor:** distance matrix

  \[
  \mathbf{R} = \begin{pmatrix}
  0 & r_{12} & r_{13} & \cdots & r_{1N_a} \\
  r_{21} & 0 & r_{23} & \cdots & r_{2N_a} \\
  \vdots & \vdots & \vdots & \ddots & \vdots \\
  r_{N_{a1}} & r_{N_{a2}} & r_{N_{a3}} & \cdots & 0
  \end{pmatrix}
  \]

- **target:** MNDO and MP2/6-311G(2d,p) levels

- "unravel" operator

  \[
  \xi = \hat{U}(n_{\text{bond}}) \mathbf{R} = (r_{12}, r_{13}, \cdots, r_{23}, r_{24}, \cdots)
  \]

  "bond-hops": \(n_{\text{bond}} = 1 - 6\)
Results: GAP(MNDO)

- Prediction:
  - optimal parameters: \( \delta = 1.0 \text{ eV}, \sigma_{\text{energy}} = 10^{-4} \text{ eV}, \sigma_{\text{force}} = 10^{-2} \text{ eV \AA}^{-1}, \theta_{\text{fac}} = 15.0 \text{ \AA} \)
  - \( n_{\text{bond}} = 4 \)
  - \( N = 2000 \)
Results: GAP(MNDO) / MM

GAP(MNDO) / OPLS
Results: MP2(GAP)

- **Problem:** MNDO and MP2/6-311G(2d,p) potentials differ significantly
  - MNDO sampling is poor for MP2 level
  - GAP training is inaccurate

- **Solution:** use a cheap potential that is similar to MP2/6-311G(2d,p)
  - MP2/6-31G

- **Future:**
  - some simple long-range repulsion term?
  - transferability?
  - electrostatic embedding?
Acknowledgement

Alan Nichol  
*University of Cambridge, UK*

Gábor Csányi  
*University of Cambridge, UK*

Daniel Cole  
*University of Cambridge, UK*

Yale University, UK