December 5, 2018

Alexandre Tkatchenko

Physics and Materials Science (PhyMS), University of Luxembourg

Opportunities and Challenges

Quantum Chemical Space

Machine Learning of Molecular
... optimization
multi-property
design through
Molecular
Molecular chemistry
"new" chemistry,
aromaticity,
trends,
Reactivity
chemical space
Structure of

Insights:
ML

molecular properties
Training data:
Quantum physics/chemistry tomorrow?
Most popular techniques: kernel methods and (deep) neural networks.

BUT: how to do this optimally with good performance on unseen data?

Example: understand chemical compound space, distinguish brain states

\[ y = f(x) \]

Learn/Infer underlying unknown mapping

- given data set \( x \) and labels \( y \) (generated by some joint probability distribution \( p(x, y) \))

Typical scenario: learning from data

[Diagram of data points and a line]

**Machine Learning in a nutshell**
\[ \langle (z)\phi(x)\phi \rangle = (z,x)k \]

\( k \leftarrow \mathcal{X} \times \mathcal{X} : k \)

**Feature space** \( \mathcal{H} \)

**Input space** \( \mathcal{X} \)

- Rewrite linear algorithm to use only inner products
- Implicitly compute inner products there
- Transform samples into higher-dimensional space

**Idea:** Kernel Learning
**Regularized Kernel Ridge Regression**

- Regularized form of ordinary regression
- Regularization prevents over-fitting by penalizing large coefficients
- Use of kernels for non-linearity

Solution has form

\[
f(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x)
\]

Coefficients \( \alpha \) are obtained by solving

\[
\sum_{i=1}^{n} (f(x_i) - y_i)^2 + \lambda \alpha^T K \alpha,
\]

which has solution

\[
\alpha = (K + \lambda I)^{-1} y.
\]
Examples of functions: hyperbolic tangent, rectification. Neuron applies a nonlinear function to its input.
Big Data for Molecules and Materials
Molecular Data in this Talk

(DFT and CCSD(T) levels)

dynamics trajectories from my group

MIDI/ISO17 datasets: Molecular

(Sci. Data 2014) and my group (PRL 2012).
calculations by von Lilienfeld’s group

QM7/QM datasets: Hybrid DFT

htp://gdb.unibe.ch/downloads/

CDB mol Graphs: J. Reymond (U. Bern)
Molecular big data

haystack

Finding needles in a haystack

candidate molecules

At least 10^6 small drug candidates

combinatorial explosion

Graph theory:
Composition vs. Conformation

- Degrees of freedom

Molecule
Properties uniquely defines a set of properties which

Which molecules to use for training?

- Data selection: Which

Between two molecules?

- Metric: how to define distance

Representation of a molecule?

- Descriptor: what is a good

Machine learning for molecular big data
Chemical Compound Space: Freedom of design
Predicting Molecular Properties: From Descriptors to Dressed Atoms to Bag-of-Bonds.

K. R. Muehler, and A. Tkatchenko,
O. A. von Lilienfeld,
R. Ramakrishnan, W. Pronobis,
K. Hansen, F. Bicger,
Predicitng Molecular Properties: QM7 dataset

<table>
<thead>
<tr>
<th>Model</th>
<th>MAE [kcal/mol]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coulomb matrix</td>
<td>1, Laplacian</td>
</tr>
<tr>
<td>Coulomb matrix</td>
<td>1, Gaussian</td>
</tr>
<tr>
<td>Bag of bonds</td>
<td>1, Laplacian</td>
</tr>
<tr>
<td>Bag of bonds</td>
<td>2, Gaussian</td>
</tr>
<tr>
<td>Polynomial pot.</td>
<td>18 = u</td>
</tr>
<tr>
<td>Polynomial pot.</td>
<td>10 = u</td>
</tr>
<tr>
<td>Polynomial pot.</td>
<td>6 = u</td>
</tr>
<tr>
<td>Lennard-Jones potential</td>
<td>sum-overbonds</td>
</tr>
<tr>
<td>dressed atoms</td>
<td></td>
</tr>
</tbody>
</table>
### 2+3 body many-body expansion

<table>
<thead>
<tr>
<th>Energy [kcal/mol]</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.8</td>
<td>MAE</td>
</tr>
<tr>
<td>4.3</td>
<td>Coulomb matrix $(I', \text{Laplacian})^{16}$</td>
</tr>
<tr>
<td>1.0</td>
<td>Coulomb matrix $(\mathcal{Z}, \text{Gaussian})^{17}$</td>
</tr>
<tr>
<td>1.5</td>
<td>Bag of bonds $(I', \text{Laplacian})$</td>
</tr>
<tr>
<td>4.5</td>
<td>Bag of bonds $(\mathcal{Z}, \text{Gaussian})$</td>
</tr>
<tr>
<td>3.0</td>
<td>Polynomial potential $(18 = u)$</td>
</tr>
<tr>
<td>6.9</td>
<td>Polynomial potential $(10 = u)$</td>
</tr>
<tr>
<td>8.7</td>
<td>Polynomial potential $(6 = u)$</td>
</tr>
<tr>
<td>6.9</td>
<td>Lennard-Jones potential sum-overbonds</td>
</tr>
<tr>
<td>15.1</td>
<td>dressed atoms model</td>
</tr>
</tbody>
</table>

### Predicting Molecular Properties: GMP dataset
Bag-of-Bonds (Bob): Non-Locality in Chemical Space
QM9 dataset: Evolution from Coulomb Matrix to Many-Body Representation

The QM9 dataset consists of 131k molecules.
QM9 dataset: Extensive and Intensive Properties

Zoo of Descriptors for Molecules and Solids
Deep Tensor Neural Networks (DTNN)

Learning the Representation
Deep Tensor Neural Networks (DTNN)
Mean absolute error on QM9: 0.2 kcal/mol.

\[
(\sum_{\text{structures}} \frac{|\Delta E|}{N})^{1/2} = \epsilon
\]

Prediction via atom-wise contributions:

\[
(\sum_{\text{structures}} |p| \cdots |p|) \sum_{a} (\sum_{a} \ldots \sum_{a} (x_{a} + \cdots + x_{a})) \phi_{a} + (x + \cdots + x) = (\sigma + \cdots + \sigma) x
\]

Sequential refinement steps \(\sigma = 1\)...

Add interaction with environment using \(\sigma \in \mathbb{R}\)

\[
p \in \mathbb{R} \Rightarrow Z x = x
\]

Embedding of based on atom types:

Input: Atomic numbers and interatomic distances.

Deep Tensor Neural Networks (DTNN)
Molecular DTNN: What Did it Learn?
Quantum Chemical Insights: Aromaticity

Energies in kcal mol⁻¹

# 281 - 290

-841.7

-845.1

-843.8

-842.1

-841.4

-841.9

-841.2

-841.9

-841.1

-841.9

-857.3

-856.9

-856.8

-857.4

-857.4

-859.9

-858.3

-857.8

-857.4
Learning Full Chemical Space with DTNN?

Accurately representing BOTH compositional and conformational degrees of freedom is difficult.

For C7O2H10 isomer and MD data, the error grows to > 1.0 kcal/mol

Beating the Hell out of Data:

Gradient-Domain Machine Learning (GDML)
Gradient-Domain Machine Learning (GDML)

Beating the Hell out of Data!
Gradient-Quasimatrix Learning (GQML)

Beating the Hell out of Data:
Gradient-Domain Machine Learning (GDM)

Beating the Hell out of Data.
Globally accurate force field from only 100s of conformations.

Towards Exact Molecular Force Field
Symmetrized Gradient Domain Machine Learning:

Quantized Electrons (CCSD(T)) and Nuclear MD for Molecules: Embarrassingly Quantum MD
Quantized Electrons (CCSD(T)) and Nuclei (PIMD)

Embarassingly Quantum MD for Molecules
Exact Free Energy Surfaces vs. Empirical Force Fields

Methyl angle (°)

Hydroxy angle (°)

Ethanol probability distribution of dihedral angles

Ab initio

SCG@CSD (T)
Towards rational design of molecules and materials in chemical space

More and better (big) data

Can ML suggest better approximations for Hamiltonians?

Can we learn (approximate) Hamiltonians and interaction models?

How to combine ML with physical laws (symmetries) excited states, spectra?

How to learn intensive properties: energy levels, materials, metrics?

What is chemical space: descriptors of molecules and Grand Challenges for Machine Learning