Learning atoms to discover new materials

Quan Zhou
The Shoucheng Zhang memorial workshop
05/04/2019
Over the years with Shoucheng

• Always keep your curiosity and open mind

• Ask big questions

• Believe in simplicity, fundamentality and beauty

• Be bold on the road less traveled

  *Why do you fear the abysses and peaks in the long journey?*
Learning atoms to discover new materials
Success of ML and AI
The Big Question

Turing Test

“Can machines do what we (thinking human) can do?”

Conversational Behavior

Can a human evaluator (C) judge natural language conversation between a human (B) and a machine (A) designed to generate human-like responses?

To test highly intelligent behaviors with original insight

“Can machines do science??”
Sanskrit names for ‘missing’ elements

**Al** [“aluminium”] $\rightarrow$ **Ga** [“gallium” (modern name), “ekaaluminium” (Mendeleev’s name)]

**Si** [“silicon”] $\rightarrow$ **Ge** [“germanium” (modern name), “ekasilicon” (Mendeleev’s name)]

“eka” = “one” in Sanskrit
AI for Language

Machines understand words via corpus

Word Embedding Vector: word2vec, GloVe, ...

Machines learn language via composition models

CNN, RNN, attention, ...

Semantic Similarity
Meaning of vector difference
The First Principle

**Distributional Hypothesis**

similar words tend to appear in similar contexts

Let us *learn* materials from data.
Let us *study* materials from data.

**Compositionality Principle**

sentence meaning is determined by words and composition rules

"let us learn materials from data"

word ~ atom
dependency in sentence ~ bond in crystal/molecule
AI for Materials Discovery?

Supervised Machine Learning

Materials Data

Material

Property

Material Feature

Learning Atoms?

Atom Descriptor

Formation Energy
Energy Band Gap
Elasticity
Conductivity

......

Atomic Number
Atomic Mass
Atomic Radius
Electronegativity

......
Materials Data

The Materials Project
Harnessing the power of supercomputing and state of the art electronic structure methods, the Materials Project provides open web-based access to computed information on known and predicted materials as well as powerful analysis tools to inspire and design novel materials.

Database Statistics

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inorganic Compounds</td>
<td>133,691</td>
</tr>
<tr>
<td>Structures</td>
<td>58,329</td>
</tr>
<tr>
<td>Molecules</td>
<td>21,954</td>
</tr>
<tr>
<td>Nanoporous Materials</td>
<td>530,243</td>
</tr>
<tr>
<td>Elastic Tensors</td>
<td>13,942</td>
</tr>
<tr>
<td>Piezoelectric Tensors</td>
<td>3,028</td>
</tr>
<tr>
<td>Intercalation Electrodes</td>
<td>3,628</td>
</tr>
<tr>
<td>Conversion Electrodes</td>
<td>16,128</td>
</tr>
</tbody>
</table>

Distribution Plot of Chemical Symbols

- Chart showing the frequency of chemical symbols with a 1% threshold.
Atom2Vec – A model-free version

Distribution over environments represents atom’s property

associations between atom and environment

- Bi$_3$Sb$_y$, Bi$_3$Se$_y$, Bi$_3$Te$_y$
- Sb$_2$Se$_y$, Sb$_2$Te$_y$, Bi$_3$O$_y$
- Sb$_2$S$_y$, ......
Atom2Vec – A model-free version

Learn atom and environment collaboratively

From atom-environment Matrix $X$

Normalization $\chi_{ij} = X_{ij}/(\sum_j X_{ij}^p)^{1/p}$

Singular value decomposition $X = UDV^T$

Atom vectors $F = [f_1, f_2, \cdots, f_N]^T = \tilde{U} \tilde{D}$

$p = 2$ almost conserves cosine distance metric

$\text{dist}(f_1, f_2) = 1 - f_1 \cdot f_2$
Further analyses

1\textsuperscript{st} principal component: alkali metals
2\textsuperscript{nd} principal component: alkali earth metals
3\textsuperscript{rd} principal component: valence trend among metals
Atom2Vec – Model-based Attempts

Composition function $C$

$$f_{env} = C(f_{i_1}, f_{i_2}, \ldots, f_{i_k})$$

Score function $S$

$$S(f_i, f_{env})$$

Learn atom vector by minimizing the loss

$$loss = \mathbb{E}_{dataset}[-\ln s(f_{i_e}, C(f_{i_1}, f_{i_2}, \ldots, f_{i_k}))]$$
More on model-based approaches

Distance matrix for composition

Graph Convolution for composition
Applications in materials search

Elpasolites (ABC$_2$D$_6$)
Formation Energy prediction

train/eval on $10^4$ elpasolites examples

prediction error $\sim 0.15\text{eV}/\text{atom}$
on par with density functional computation
Applications in materials search

Half-heusler (XYZ)
Formation Energy prediction
Metal/Insulator classification

reasonable clustering of transition metals

similar classification accuracy as 18-electron rules

Strong covalent bonding indicates band gap

\[
\begin{align*}
X & \quad Y & \quad Z \\
Sc & \quad Al & \quad Sb \\
3 & + & 10 & + & 5 & = & 18 \\
Sc & \quad Co & \quad Sb \\
3 & + & 9 & + & 5 & \neq & 18
\end{align*}
\]
The Era of AI