Machine learning materials science from experimental data

S.J.L. Billinge
Columbia University, Brookhaven National Laboratory

George Box: All models are wrong, some models are useful

Andy Millis: I want to be able to work on some classical problems
Columbia University in the City of New York
Brookhaven National Laboratory

National Synchrotron Light Source-II (NSLS-II) =>

• XPD beamline
• Coherence
• Small beams
• High energy resolution
• Resonant scattering
Materials Genomics

1. Structure prediction

2. Structure solution aka phase problem

3. Materials Discovery

DFT/MD

Synthesis Material Diffraction Structure Properties Understanding
The Crystal Structure Problem

Problem:
- Here is a crystal, what is its structure?

Solution:
1. Give it to your grad student
2. She puts it on the x-ray machine
3. …Pushes the button
   1. Machine tells you the structure
   2. Or Machine gets stuck
      1. Throw away the crystal
      2. Make it the subject of her thesis

Crystallography is largely a solved problem

From LiGaTe2: A New Highly Nonlinear Chalcopyrite Optical Crystal for the Mid-IR
Quantum Dot solar cells
The Nanostructure Problem

• **Problem:**
  – Here is a nanoparticle, what is its structure?

• **Solution:**
  1. Give it to your grad student
  2. She puts it on the x-ray machine
  3. ... Pushes the button
The Challenge with Real Material Structure Determination

(a) $\alpha$-ZrP

(b) H-Zr
Nanostructure inverse problems

... are often ill-posed
The atomic Pair Distribution Function

Structure function

Raw data

PDF

$$G(r) = \frac{2}{\pi} \int_0^\infty Q[S(Q) - 1] \sin QrdQ$$
\[ G(r) = \frac{2}{\pi} \int_{Q_{\text{min}}}^{Q_{\text{max}}} F(Q) \sin(Qr) \, dQ \]
$G_{\text{calc}}(r) = \frac{1}{N r} \sum_i \sum_{j \neq i} \left( \frac{b_i b_j}{\langle b \rangle^2} \delta(r - r_{ij}) \right) - 4\pi r \rho_0$
\[ R_w = 0.019 \]

\[ R_w = \sqrt{\frac{\sum_{i=1}^{n} [G_{\text{obs}}(r_i) - G_{\text{calc}}(r_i, P)]^2}{\sum_{i=1}^{n} G_{\text{obs}}(r_i)^2}} \]
$R_w = 0.019$

Fig. S1: Pd$_{N}$ 3.0 ± 0.3 nm

JCPDS # 05-0681
$G(r,d)_s = G(r)f(r,d)$

$f(r,d) = \left[1 - \frac{3r}{2d} + \frac{1}{2}\left(\frac{r}{d}\right)^3\right] \Theta(d-r)$
AC fits to nanoparticle data are good
....but not that good!

• Excuses excuses excuses!
• Can we do better?
<table>
<thead>
<tr>
<th>Name</th>
<th>Composition</th>
<th>D (nm)</th>
<th>Surfact.</th>
<th>Beamline</th>
<th>Morph.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pd$^P$</td>
<td>Pd</td>
<td>3.0$^a$</td>
<td>OAm</td>
<td>XPD</td>
<td>P</td>
</tr>
<tr>
<td>CoPd$^P$</td>
<td>Co$<em>{0.2}$Pd$</em>{0.8}$</td>
<td>8.6$^a$</td>
<td>OAm, TOP</td>
<td>XPD</td>
<td>P</td>
</tr>
<tr>
<td>PdNi$^W$</td>
<td>Pd$<em>{0.83}$Ni$</em>{0.17}$</td>
<td>2.3$^a$</td>
<td>ODA</td>
<td>X17A</td>
<td>W</td>
</tr>
<tr>
<td>Ag$^P$</td>
<td>Ag</td>
<td>38.0$^b$</td>
<td>PVP</td>
<td>XPD</td>
<td>P</td>
</tr>
<tr>
<td>AgAu$^P$</td>
<td>Ag$<em>{0.5}$Au$</em>{0.5}$</td>
<td>5.0$^b$</td>
<td>PVP</td>
<td>XPD</td>
<td>P</td>
</tr>
<tr>
<td>PdAu$^W$</td>
<td>Pd$<em>{0.9}$Au$</em>{0.1}$</td>
<td>2.1$^a$</td>
<td>ODA</td>
<td>X17A</td>
<td>W</td>
</tr>
<tr>
<td>Pd$^W$</td>
<td>Pd</td>
<td>2.0$^a$</td>
<td>ODA</td>
<td>X17A</td>
<td>W</td>
</tr>
<tr>
<td>Au$^P$</td>
<td>Au</td>
<td>9.0$^b$</td>
<td>PVP</td>
<td>XPD</td>
<td>P</td>
</tr>
<tr>
<td>PtRu$^{P,C}$</td>
<td>Pt$<em>{0.5}$Ru$</em>{0.5}$</td>
<td>3.5$^a$</td>
<td>PFSA</td>
<td>XPD</td>
<td>P,C</td>
</tr>
<tr>
<td>Pt$^P$</td>
<td>Pt</td>
<td>3.0$^a$</td>
<td>OAm, TOP</td>
<td>XPD</td>
<td>P</td>
</tr>
<tr>
<td>Au$^{C1}$</td>
<td>Au</td>
<td>1-2</td>
<td>p-MBA</td>
<td>ID11</td>
<td>Cl1</td>
</tr>
<tr>
<td>Au$^{C2}$</td>
<td>Au</td>
<td>1-2</td>
<td>SC6</td>
<td>11IDB</td>
<td>Cl2</td>
</tr>
<tr>
<td>Ni$^B$</td>
<td>Ni</td>
<td>Bulk</td>
<td>None</td>
<td>XPD</td>
<td>B</td>
</tr>
</tbody>
</table>

$^a$ Average size by transmission electron microscopy (TEM)

$^b$ Average size by differential centrifugal sedimentation (DCS)

Fig. S2: CoPd$^P$ 8.6 ± 1.7 nm
Residuals

![Graph with Residuals](image)
Residuals
Residuals

![Graphs showing residual plots for different materials.](HTTP://thebillingegroup.com)
Residuals
Testing discrete models
Testing discrete models
Testing discrete models
Testing discrete models
Testing discrete models

Automating cluster finding
Standard approach

- 1 PDF
- 1 model
- 20 parameters
- Vary the parameters until the model agrees as well as possible with the data.
- Emphasis on parameter estimation, not on model selection
- Challenge is finding the right model
New approach: Cluster-Mining

- 1 PDF
- many models
- few parameters
- Iterate over large numbers of models
- Emphasis on model selection
- Advantage: find multiple nearby models!
- Challenge: structure must be in the structure-mine
Structure-mining: An automated screening of large numbers of candidate structures to the atomic pair distribution function (PDF)

Long Yang, Pavol Juhas, Maxwell W. Terban, Simon J. L. Billinge
Acta Crystallographica Section A, in press
Introduction

Crystal structural database

Materials Project Database
Crystallography Open Database

PDF auto-refinement

DiffPy-CMI

Example result of BaTiO3 nanoparticle heuristic - 1

<table>
<thead>
<tr>
<th>No.</th>
<th>$R_w$</th>
<th>formula</th>
<th>space_group</th>
<th>$U_{eq,r}$</th>
<th>$O_{eq,r}$</th>
<th>$Tl_{eq,r}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>0.143308</td>
<td>Ba2O6Ti2</td>
<td>Amm2</td>
<td>0.02085</td>
<td>0.018062</td>
<td>0.008006</td>
</tr>
<tr>
<td>21</td>
<td>0.144425</td>
<td>Ba2O6Ti2</td>
<td>Amm2</td>
<td>0.02027</td>
<td>0.017803</td>
<td>0.008684</td>
</tr>
<tr>
<td>22</td>
<td>0.145316</td>
<td>Ba2O6Ti2</td>
<td>Amm2</td>
<td>0.01990</td>
<td>0.017465</td>
<td>0.009089</td>
</tr>
<tr>
<td>26</td>
<td>0.146433</td>
<td>Ba2O6Ti2</td>
<td>Amm2</td>
<td>0.01973</td>
<td>0.016750</td>
<td>0.009420</td>
</tr>
<tr>
<td>30</td>
<td>0.148121</td>
<td>Ba2O6Ti2</td>
<td>Amm2</td>
<td>0.01957</td>
<td>0.017870</td>
<td>0.009978</td>
</tr>
<tr>
<td>33</td>
<td>0.151343</td>
<td>Ba2O6Ti2</td>
<td>Amm2</td>
<td>0.02535</td>
<td>0.006099</td>
<td>0.008263</td>
</tr>
<tr>
<td>2</td>
<td>0.155847</td>
<td>Ba1O3Ti1</td>
<td>P4mm</td>
<td>0.002735</td>
<td>0.011571</td>
<td>0.006421</td>
</tr>
<tr>
<td>8</td>
<td>0.161942</td>
<td>Ba1O3Ti1</td>
<td>P4mm</td>
<td>0.002634</td>
<td>0.016226</td>
<td>0.008579</td>
</tr>
<tr>
<td>9</td>
<td>0.161942</td>
<td>Ba1O3Ti1</td>
<td>P4mm</td>
<td>0.002634</td>
<td>0.016226</td>
<td>0.008579</td>
</tr>
<tr>
<td>31</td>
<td>0.163028</td>
<td>Ba1O3Ti1</td>
<td>R3m</td>
<td>0.002704</td>
<td>0.016278</td>
<td>0.007618</td>
</tr>
<tr>
<td>29</td>
<td>0.163386</td>
<td>Ba1O3Ti1</td>
<td>Amm2</td>
<td>0.02027</td>
<td>0.006292</td>
<td>0.005347</td>
</tr>
<tr>
<td>28</td>
<td>0.163120</td>
<td>Ba1O3Ti1</td>
<td>R3m</td>
<td>0.002699</td>
<td>0.016083</td>
<td>0.007675</td>
</tr>
<tr>
<td>27</td>
<td>0.163405</td>
<td>Ba1O3Ti1</td>
<td>R3m</td>
<td>0.002681</td>
<td>0.015975</td>
<td>0.007887</td>
</tr>
<tr>
<td>25</td>
<td>0.163844</td>
<td>Ba1O3Ti1</td>
<td>R3m</td>
<td>0.002663</td>
<td>0.015968</td>
<td>0.008149</td>
</tr>
<tr>
<td>17</td>
<td>0.163934</td>
<td>Ba1O3Ti1</td>
<td>R3m</td>
<td>0.002649</td>
<td>0.015838</td>
<td>0.008251</td>
</tr>
<tr>
<td>24</td>
<td>0.164318</td>
<td>Ba1O3Ti1</td>
<td>R3m</td>
<td>0.002630</td>
<td>0.015253</td>
<td>0.008431</td>
</tr>
<tr>
<td>16</td>
<td>0.164382</td>
<td>Ba1O3Ti1</td>
<td>R3m</td>
<td>0.002639</td>
<td>0.015667</td>
<td>0.008449</td>
</tr>
<tr>
<td>18</td>
<td>0.164820</td>
<td>Ba1O3Ti1</td>
<td>R3m</td>
<td>0.002614</td>
<td>0.015209</td>
<td>0.008652</td>
</tr>
<tr>
<td>32</td>
<td>0.165866</td>
<td>Ba1O3Ti1</td>
<td>R3m</td>
<td>0.002599</td>
<td>0.014932</td>
<td>0.009055</td>
</tr>
<tr>
<td>3</td>
<td>0.165928</td>
<td>Ba1O3Ti1</td>
<td>P4mm</td>
<td>0.002610</td>
<td>0.015074</td>
<td>0.009552</td>
</tr>
<tr>
<td>9</td>
<td>0.166332</td>
<td>Ba1O3Ti1</td>
<td>P4mm</td>
<td>0.002616</td>
<td>0.015758</td>
<td>0.009675</td>
</tr>
<tr>
<td>4</td>
<td>0.167933</td>
<td>Ba1O3Ti1</td>
<td>Pnmm2</td>
<td>0.002845</td>
<td>0.015098</td>
<td>0.009475</td>
</tr>
<tr>
<td>23</td>
<td>0.169423</td>
<td>Ba1O3Ti1</td>
<td>P4mm</td>
<td>0.002625</td>
<td>0.016364</td>
<td>0.010359</td>
</tr>
<tr>
<td>11</td>
<td>0.170190</td>
<td>Ba1O3Ti1</td>
<td>P4/mmm</td>
<td>0.002643</td>
<td>0.017426</td>
<td>0.010542</td>
</tr>
<tr>
<td>10</td>
<td>0.170190</td>
<td>Ba1O3Ti1</td>
<td>P4/mmm</td>
<td>0.002643</td>
<td>0.017426</td>
<td>0.010542</td>
</tr>
<tr>
<td>15</td>
<td>0.201176</td>
<td>Ba1O3Ti1</td>
<td>Pn-3m</td>
<td>0.004582</td>
<td>0.017209</td>
<td>0.012186</td>
</tr>
<tr>
<td>1</td>
<td>0.201176</td>
<td>Ba1O3Ti1</td>
<td>Pn-3m</td>
<td>0.004582</td>
<td>0.017209</td>
<td>0.012186</td>
</tr>
<tr>
<td>13</td>
<td>0.201176</td>
<td>Ba1O3Ti1</td>
<td>Pn-3m</td>
<td>0.004582</td>
<td>0.017209</td>
<td>0.012186</td>
</tr>
<tr>
<td>12</td>
<td>0.201176</td>
<td>Ba1O3Ti1</td>
<td>Pn-3m</td>
<td>0.004582</td>
<td>0.017209</td>
<td>0.012186</td>
</tr>
<tr>
<td>5</td>
<td>0.201176</td>
<td>Ba1O3Ti1</td>
<td>Pn-3m</td>
<td>0.004582</td>
<td>0.017209</td>
<td>0.012186</td>
</tr>
<tr>
<td>14</td>
<td>0.201176</td>
<td>Ba1O3Ti1</td>
<td>Pn-3m</td>
<td>0.004582</td>
<td>0.017209</td>
<td>0.012186</td>
</tr>
<tr>
<td>6</td>
<td>0.266677</td>
<td>Ba1O3Ti1</td>
<td>Pm-3m</td>
<td>0.005790</td>
<td>0.009941</td>
<td>0.012610</td>
</tr>
<tr>
<td>7</td>
<td>0.273221</td>
<td>Ba1O3Ti1</td>
<td>Pn-3m</td>
<td>0.006070</td>
<td>0.046913</td>
<td>0.004137</td>
</tr>
<tr>
<td>19</td>
<td>0.070576</td>
<td>Ba1O3Ti1</td>
<td>P4mm</td>
<td>0.004189</td>
<td>0.047858</td>
<td>0.248950</td>
</tr>
</tbody>
</table>

HTTP://thebillingegroup.com
• Structure-mining found the same model as in prior work, MPD No. 1003 (NaFeSi$_2$O$_6$) and COD No. 2983 (NaFeSi$_2$O$_6$), s.g.: C 2/c.
• It also returns some structures with space group C 2, such as MPD No. 998 (Na$_{0.83}$FeSi$_2$O$_6$), which may be viewed as a very similar structure but with a lowered symmetry and deficient atoms at some sites
• It also returns some structures substituting at Na or Fe sites by other elements. For example, MPD No. 1021 (NaGaSi$_2$O$_6$).
To the cloud!!!!!
structureMining

Given a measured (or calculated) PDF, structureMining will search databases to find the best structures to fit it.

Please see the structureMining paper for more information. Please cite the structureMining paper if this helps you get a publishable result.

Upload Data

PDF file: Browse... No file selected.

- X-ray
- Neutron

Composition: 

Optional Parameter: 

Submit
structureMiner

Results

SM found total 3 structures and 3 structures with weighted profile agreement factor, $R_w < 0.5$.

<table>
<thead>
<tr>
<th>rw</th>
<th>formula</th>
<th>space_group</th>
<th>db</th>
<th>db_id</th>
<th>ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>As8 Fe8 Sr4</td>
<td>Fmmm</td>
<td>COD</td>
<td>9016523</td>
<td>Tegel. Journal of Physics: Condensed Matter. 20 (2008) 452201_1--452201_5</td>
</tr>
<tr>
<td>2</td>
<td>As2 Fe2 Sr1</td>
<td>l4/mmm</td>
<td>MPD</td>
<td>mp-4488</td>
<td>Saha et al. Journal of Physics: Conference Series. 273 (2011) 012104/1-012104/4</td>
</tr>
</tbody>
</table>
The Chemistry student laptop/notebook problem
The analyzed structured labelled data from everyone in a database problem
and all this with spatial and time dependence
This new approach to modeling opens the door for Machine Learning/Deep learning

- I don’t know an person who can look at a PDF and tell the s.g. (except perhaps if it is an fcc material or something)

Question:
- Can we train a Machine Learning model to recognize it?

Answer:
- Yes!

Details

• Work of Yunzhe Tao and Chia-Hao Liu
• PDF itself is the input feature vector
• 101,802 structures in 45 space-groups
• Train on 80% of data, 20% retained for testing
Space Groups from PDFs

- First try, Logistic Regression
- Second try, Convolution Neural Net
  - Classification ratio’s are for finding the right s.g. in the top 6
**Confusion Matrix**

- **Rows:** true label
- **Columns:** predicted label

### Table

<table>
<thead>
<tr>
<th></th>
<th>P1</th>
<th>P2_1</th>
<th>P2_2</th>
<th>C2_1</th>
<th>C2_2</th>
<th>C2/c</th>
<th>P2_1/c</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>0.10</td>
<td>0.20</td>
<td>0.30</td>
<td>0.00</td>
<td>0.10</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>P2</td>
<td>0.00</td>
<td>0.10</td>
<td>0.20</td>
<td>0.30</td>
<td>0.00</td>
<td>0.10</td>
<td>0.00</td>
</tr>
<tr>
<td>C2</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.10</td>
<td>0.20</td>
<td>0.30</td>
</tr>
</tbody>
</table>

### Diagram

- **Pbcn**
- **Pbca**
- **P2_1/c**
- **P2_2 2_1**
- **C2/c**
- **P2_1**

---

[HTTP://thebillingegroup.com](HTTP://thebillingegroup.com)

**COLUMBIA UNIVERSITY**
**IN THE CITY OF NEW YORK**

---
Acknowledgements

• A special thank you to all my current and former students and post-docs
• Also my many wonderful collaborators, mentioned during the talk
• Facilities:
  – APS, CHESS, NSLS, NSLS-II (and people therein)
  – MLNSC, ISIS, SNS (and people therein)
• Funding: DOE-BES and NSF-DMR