Machine Learning in Chemical Space: Predicting Electronic Structure Properties

Matthias Rupp, Grégory Montavon, Vivekanand Gobey, AlyزارVazquez-Mayagoitia, Katja Hansen, Alexandre Tkatchenko, Klaus-Robert Müller, O. Anatole von Lilienfeld

1 Introduction

Machine learning models for the prediction of 14 ground and excited state properties of 7000 small organic molecules are presented. Cross-validated model accuracies are on the order of the used reference methods. [1-6]

Key facts:
- Data set: 7211 small organic molecules; 14 properties [1,2]
- Representation: “Coulomb matrix” [1,5]
- ML methods: Gaussian process regression [7], artificial neural networks [11]
- Errors: Same order of magnitude as reference methods [2]

Background:
- Electronic structure calculations for many similar molecules contain redundant information, which is currently not exploited. Examples: structural relaxation, molecular dynamics, high-throughput calculations of QM properties
- Machine learning can be used to interpolate between reference calculations. Idea: Map the problem of solving the electronic Schrödinger equation onto an non-linear statistical regression problem
- ML models are entirely empirical, i.e., applicable only to the “chemical space” covered by training data. Hybrid QM/ML models: Use general, but slow QM for reference calculations, and fitted, but fast ML to interpolate
- Substantial savings in computing time. Speed: hours (QM) versus ms (ML)

2 Data

- 7211 small organic molecules, elements H, N, O, S, Cl, from the GDB [8]
- Single, double, triple bonds, (hetero-)cycles, carboxy, cyano, amide, amines, alcohol, epoxide, sulfide, ether, ester, chloride, aliphatic, aromatic groups, ...
- Geometries: Relaxation by universal force field, then DFT-PBE
- Data to be publicly available on quantum-machine.org and mruupp.info

Figure 2: Overview of data set used to train and test ML models. Shown are QM results for 14 properties of 7211 molecules (scatter plots, top right), and a cartoon of a QM/ML model (boxed, bottom left) that correlates input structures with two seemingly unrelated properties, HOMO eigenvalue and atomization energy.

3 Method

- Gaussian process regression (GP) [7] is a non-parametric Bayesian regression method. A major advantage is that it provides confidence estimates for its predictions, i.e., it has built-in domain of applicability [9].

- Deep architecture unfolds complex input into simple representation
- Multi-task ANN learns several properties simultaneously

Figure 3: Gaussian process regression. Starting from the prior distribution (left), one conditions on the observed samples (red crosses). Mean (solid line) and variance (grey area) of the posterior (right) serve as predictor and confidence estimates, respectively.

4 Results

- Four layer ANN with sigmoidal activation using stochastic “Coulomb matrices”:

<table>
<thead>
<tr>
<th>Property</th>
<th>Method</th>
<th>MAE</th>
<th>RMSE</th>
<th>Reference MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atomization energy $E$</td>
<td>PBE0</td>
<td>0.16</td>
<td>0.36</td>
<td>0.15, 0.23, 0.09-0.22</td>
</tr>
<tr>
<td>Polarizability $\alpha$</td>
<td>PBE0</td>
<td>0.11</td>
<td>0.18</td>
<td>0.05-0.27, 0.04-0.14</td>
</tr>
<tr>
<td>Polarizability $\alpha$</td>
<td>SC5</td>
<td>0.08</td>
<td>0.12</td>
<td>0.05-0.27, 0.04-0.14</td>
</tr>
<tr>
<td>HOMO eigenvalues</td>
<td>GW</td>
<td>0.16</td>
<td>0.22</td>
<td></td>
</tr>
<tr>
<td>HOMO eigenvalues</td>
<td>PBE0</td>
<td>0.15</td>
<td>0.21</td>
<td>2.08</td>
</tr>
<tr>
<td>HOMO eigenvalues</td>
<td>ZINDO</td>
<td>0.15</td>
<td>0.22</td>
<td>0.79</td>
</tr>
<tr>
<td>LUMO eigenvalues</td>
<td>GW</td>
<td>0.13</td>
<td>0.21</td>
<td></td>
</tr>
<tr>
<td>LUMO eigenvalues</td>
<td>PBE0</td>
<td>0.12</td>
<td>0.20</td>
<td>1.30</td>
</tr>
<tr>
<td>LUMO eigenvalues</td>
<td>ZINDO</td>
<td>0.11</td>
<td>0.18</td>
<td>0.13</td>
</tr>
<tr>
<td>Ionization potential $IP$</td>
<td>ZINDO</td>
<td>0.17</td>
<td>0.26</td>
<td>0.20, 0.15</td>
</tr>
<tr>
<td>Electron affinity $EA$</td>
<td>ZINDO</td>
<td>0.11</td>
<td>0.18</td>
<td>0.16, 0.11</td>
</tr>
<tr>
<td>First exc. energy $E_{1}$</td>
<td>ZINDO</td>
<td>0.13</td>
<td>0.31</td>
<td>0.18, 0.21</td>
</tr>
<tr>
<td>Exc. freq. max. abs. $E_{\infty}$</td>
<td>ZINDO</td>
<td>1.06</td>
<td>1.76</td>
<td></td>
</tr>
<tr>
<td>Max. abs. intensity $I_{\infty}$</td>
<td>ZINDO</td>
<td>0.07</td>
<td>0.12</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Prediction error of ANN model. For each property, mean absolute error (MAE), root mean squared error (RMSE), and, where available, MAE of the reference method are shown. [2] Errors are for out-of-sample predictions of 2211 molecules (5000 training molecules). Energies, polarizabilities, and intensity are in eV, Å3 and arbitrary units.

- Molecular representation is critical; “Coulomb matrix” is simple and based only on positions and nuclear charges
- Training GPS scales cubically; training ANNs requires tricks of the trade [11]
- For 5000 training molecules, models reach accuracy of reference method

References