Abstract Submitted
for the MAR17 Meeting of
The American Physical Society

**The role of representation and training set selection for improved machine learning models of matter** BING HUANG, O. ANATOLE VON LILIENFELD, University of Basel — Choice of representation and training set are fundamentally important in machine learning (ML) models of chemical and physical properties of matter. Based on the postulates of quantum mechanics we have developed a hierarchy of representations which meet uniqueness and target similarity criteria. To systematically control target similarity, we rely on interatomic many body expansions, as implemented in universal force-fields, including bags of sorted Bonding, Angular, and higher order terms (BA). Addition of higher order contributions systematically increases the predictive accuracy of the resulting BAML models. BAML predicts properties of out-of-sample molecules with unprecedented accuracy and speed.[1] To select optimal training sets we have developed a rational approach which results in ML models with very rapid error decay.[2] In combination with BAML based atomic representations, these ML models reach chemical accuracy for atomization energies (∼1 kcal/mol) after training on reference results for only hundreds of chemical compounds. Our findings suggest a dramatic reduction in need for data.

[1]Huang and von Lilienfeld, *J. Chem. Phys.* Comm. **145**, 161102 (2016)
[2]Huang and von Lilienfeld, in preparation (2016)

Bing Huang
University of Basel

Date submitted: 10 Nov 2016 Electronic form version 1.4