Machine Learning in Computational Chemistry
Do you think machine learning is overhyped?

- Agree: 31.8%
- Strongly agree: 13.9%
- Disagree: 24.5%
- Strongly disagree: 7.9%
- Not sure: 21.2%
- Other: 0.7%

*Chemical & Engineering News, August 2018*
Despite all this promise—or perceived promise—one thing that machine learning isn’t is magic. “Let’s be realistic,” says **George Dahl**, a computer scientist at Google. “Machine learning is nonlinear regression,” a simple type of statistical analysis in which collected data are “fit” with model parameters. Dahl won a Merck & Co. machine-learning competition while a graduate student in **Geoffrey Hinton**’s group at the University of Toronto.

> “Since machine learning simply is interpolation within big data sets, it will remain difficult or impossible to use in areas where it is hard to generate large sets of reliable data, because extrapolation by machine learning can and will produce wildly wrong answers.”—**Bernd Hartke**, professor, University of Kiel

Deep neural networks’ ability to learn even from very complex data makes them especially attractive in pharmaceutical chemistry. Abraham Heifets is cofounder and CEO of **Atomwise**, which makes deep-neural-network-based software for predicting binding affinities of drug candidates to targets in the body. He calls the introduction of deep neural networks a fundamental change in machine learning.

Some have accused Atomwise of overhyping machine learning’s capabilities. A **2015 TechCrunch article** quoted the firm’s cofounder Alexander Levy as saying the Atomwise software allowed him to predict a cure for measles from his living room.
Z-Score Sums for Groups in CASP14

mean Z-score for AlphaFold 2 = 2.5
for "hard" targets = 3.8 ("IQ" > 160)
Historical CASP Results

GDT=60  Overall Fold Correct  GDT=80  Sidechains Correct
Sample CASP14 Results from AlphaFold 2

Bacteriophage T4 SPACKLE
(total RMSD of 0.48 Ang)

SARS-CoV-2 ORF8
(“poor” result for AlphaFold2)
Comparison for SARS-CoV-2 ORF8

Expt
Baker
AlphaFold 2
Zhang
Comparison for Bacteriophage T4 SPACKLE

- **Expt**
- **Baker**
- **AlphaFold 2**
- **Zhang**
AlphaFold 2 Side Chain Accuracy
Sequence Covariance is a Major Data Source Used Effectively by AlphaFold 2
the predictive process proceeds in iterative fashion by "passing information back and forth between the MSA and the internal representation of the protein"
Big Data Meets Quantum Chemistry Approximations: The Δ-Machine Learning Approach

Raghunathan Ramakrishnan,† Pavlo O. Dral,¶,‡ Matthias Rupp,† and O. Anatole von Lilienfeld*,†,§

†Institute of Physical Chemistry and National Center for Computational Design and Discovery of Novel Materials, Department of Chemistry, University of Basel, Klingelbergstraße 80, CH 4056 Basel, Switzerland
¶Max Planck Institut für Kohlenforschung, Kaiser Wilhelm Platz 1, 45470 Mülheim an der Ruhr, Germany
‡Computer Chemie Centrum and Interdisciplinary Center for Molecular Materials, Department Chemie und Pharmazie, Friedrich Alexander Universität Erlangen Nürnberg, Nägeelsbachstraße 25, 91052 Erlangen, Germany
§Argonne Leadership Computing Facility, Argonne National Laboratory, 9700 S. Cass Avenue, Lemont, Illinois 60439, United States

JCTC, 11, 2087-2096 (2015)
Figure 1. Two hypothetical property profiles connecting two constitutional isomers of C_7H_{10}O_2. The Δ model, eq 1, estimates the difference between baseline and targetline properties (arrow) which differ in level of theory (b → t), geometry (R_b → R_t), and property (E_b → H_t).
Figure 4. Mean absolute error (MAE) [kcal/mol] of $E_b$ (blue bars) and 1k $\Delta^t_b$ ML model predictions (red bars) of atomization energies for various combinations of increasingly correlated post Hartree–Fock methods as target and baseline methods. These values are obtained after subtracting the systematic average shift that exists between methods for the 6k isomers. See Supporting Information for details.
Figure 5. Calculated reaction enthalpies at 298.15 K between the most stable molecule with $C_7H_{10}O_2$ stoichiometry (6-oxabicyclooctan-7-one, in inset, with atomization enthalpy, $-1933$ kcal/mol), and its 10 energetically closest isomers in increasing order, according to targetline method G4MP2 calculated \textit{a posteriori} (black). 1k $\Delta_{\text{B3LYP}}^{\text{G4MP2}}$ ML model predictions are given in blue. Baseline method B3LYP predictions are shown for comparison (red).
Figure 6. Molecular pairs with maximal reaction entropy (top) and electron correlation energy (bottom) out of the 10k stable diastereomers of C\textsubscript{7}H\textsubscript{10}O\textsubscript{2}. Reaction properties have been estimated using a 1k $\Delta_{\text{B3LYP}}^{\text{G4MP2}}$ ML model for the entropy and a 1k $\Delta_{\text{HF}}^{\text{CCSD(T)}}$ ML model for the correlation energy (black). Corresponding reference values calculated for validation are given in parentheses (blue).
Figure 7. Scatter plot of ML model predicted entropy of atomization at $T = 298.15$ K versus ML model predicted electron correlation contribution to atomization energy for the 16k stable out of sample diastereomers of $C_7H_{10}O_2$. All predictions are made using ML models trained on 1k molecules, randomly drawn from a set of 6k parent isomers. Training data are shown in red.
Transferable Dynamic Molecular Charge Assignment Using Deep Neural Networks

Benjamin Nebgen, Nicholas Lubbers, Justin S. Smith, Andrew E. Sifain, Andrey Lokhov, Olexandr Isayev, Adrian E. Roitberg, Kipton Barros, and Sergei Tretiak

†Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87545, United States
§Department of Chemistry, University of Florida, Gainesville, Florida 32611, United States
§Department of Physics and Astronomy, University of Southern California, Los Angeles, California 90037, United States
||UNC Eshelman School of Pharmacy University of North Carolina Chapel Hill, Chapel Hill, North Carolina 27599, United States
Figure 1. A diagrammatic representation of HIP NN. Green bars represent interaction layers and red squares represent on site layers for atoms. The blue squares are the output layers which return a series of corrections to the atomic charge. The molecule on top illustrates how information can be passed from one atom to another. This includes information being indirectly passed from distance atoms through an intermediate interaction layer.
**Figure 6.** IR spectrum generated completely by machine learning (red line) and from a normal mode analysis (black line). The ML spectrum was generated with an identical method as in **Figure 5**. Different charge assignment schemes produce significantly different IR spectra with Hirshfeld charges being the most reliable for larger molecules.
Molecular Dynamics Fingerprints (MDFP): Machine Learning from MD Data To Predict Free-Energy Differences

Sereina Riniker*

Laboratory of Physical Chemistry, ETH Zürich, Vladimir Prelog Weg 2, 8093 Zürich, Switzerland

Supporting Information

abstract: While the use of machine learning (ML) techniques is well established in cheminformatics for the prediction of physicochemical properties and binding affinities, the training of ML models based on data from molecular dynamics (MD) simulations remains largely unexplored. Here, we present a fingerprint termed MDFP which is constructed from the distributions of properties such as potential energy components, radius of gyration, and solvent accessible surface area extracted from MD simulations. The corresponding fingerprint elements are the first two statistical moments of the distributions and the median. By considering not only the average but also the spread of the distribution in the fingerprint, some degree of entropic information is encoded. Short MD simulations of the molecules in water (and in vacuum) are used to generate MDFP. These are further combined with simple counts based on the 2D structure of the molecules into MDFP+. The resulting information rich MDFP+ is used to train ML models for the prediction of solvation free energies in five different solvents (water, octanol, chloroform, hexadecane, and cyclohexane) as well as partition coefficients in octanol/water, hexadecane/water, and cyclohexane/water. The approach is easy to implement and computationally relatively inexpensive. Yet, it performs similarly well compared to more rigorous MD based free energy methods such as free energy perturbation (FEP) as well as end state methods such as linear interaction energy (LIE), the conductor like screening model for realistic solvation (COSMO RS), and the SMx family of solvation models.
Solvation via FEP and MDFP+ Machine Learning

JCTC, Vol 13 (2017)
Atomic Forces from Machine Learning

Learning scheme to predict atomic forces and accelerate materials simulations

V. Botu
Department of Chemical and Biomolecular Engineering, University of Connecticut, Storrs, Connecticut 06269, USA

R. Ramprasad*
Department of Materials Science and Engineering, University of Connecticut, Storrs, Connecticut 06269, USA
(Received 5 May 2015; published 25 September 2015)

The behavior of an atom in a molecule, liquid, or solid is governed by the force it experiences. If the dependence of this vectorial force on the atomic chemical environment can be learned efficiently with high fidelity from benchmark reference results—using “big-data” techniques, i.e., without resorting to actual functional forms—then this capability can be harnessed to enormously speed up in silico materials simulations. The present contribution provides several examples of how such a force field for AI can be used to go far beyond the length-scale and time-scale regimes presently accessible using quantum-mechanical methods. It is argued that pathways are available to systematically and continuously improve the predictive capability of such a learned force field in an adaptive manner, and that this concept can be generalized to include multiple elements.
FIG. 1. (Color online) Comparison of the forces predicted using the ML force field with reference DFT results, for (a) the trained model (light blue) and the validation data set (dark blue), (b) a test unit cell containing over 800 Al atoms in the fcc phase, and (c) a test unit cell containing over 160 atoms in a hypothetical bcc phase. In (b) and (c), atoms were randomly perturbed from their equilibrium positions. Insets show the distribution of the prediction errors (defined as the difference between the predicted and reference DFT forces) leading to respectable mean absolute errors (MAEs).
“Newton-in-a-Box”

$F = ma$

Deep Learning ??
ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost†

J. S. Smith a, O. Isayev b and A. E. Roitberg c

aUniversity of Florida, Department of Chemistry, PO Box 117200, Gainesville, FL, USA 32611-7200. E-mail: roitberg@ufl.edu
bUniversity of North Carolina at Chapel Hill, Division of Chemical Biology and Medicinal Chemistry, UNC Eshelman School of Pharmacy, Chapel Hill, NC, USA 27599. E-mail: olexandr@olexandrisayev.com

Received 31st December 2016, Accepted 7th February 2017

First published on 8th February 2017
Fig. 1 Behler and Parrinello’s HDNN or HD-atomic NNP model. (A) A scheme showing the algorithmic structure of an atomic number specific neural network potential (NNP). The input molecular coordinates, $\vec{q}$, are used to generate the atomic environment vector, $G_i^X$, for atom $i$ with atomic number $X$. $G_i^X$ is then fed into a neural network potential (NNP) trained specifically to predict atomic contributions, $E_i^X$, to the total energy, $E_T$. Each $l_k$ represents a hidden layer of the neural network and is composed of nodes denoted $a_j^k$ where $j$ indexes the node. (B) The high-dimensional atomic NNP (HD-atomic NNP) model for a water molecule. $G_i^X$ is computed for each atom in the molecule then input into their respective NNP ($X$) to produce each atom’s $E_i^X$, which are summed to give $E_T$. 
