Modeling of Molecular Atomization Energies using Machine Learning

Matthias Rupp,[1,2,3] Alexandre Tkatchenko,[4,2] Klaus-Robert Müller,[1,2] O. Anatole von Lilienfeld[5,2]

[1] Machine Learning Group, Technical University of Berlin, Franklinstr. 28/29, 10587 Berlin, Germany. mrupp@mrupp.info [2] Institute for Pure and Applied Mathematics, University of California, Los Angeles, California 90095, USA [3] Computer-Assisted Drug Design, Eidgenössische Technische Hochschule Zürich, Wolfgang-Pauli-Str. 10, 8093 Zürich, Switzerland [4] Fritz Haber Institute of Max Planck Society, Faradayweg 4-6, 14195 Berlin, Germany [5] Argonne Leadership Computing Facility, Argonne National Laboratory, Argonne, Illinois 60439, USA

1 Introduction

Summary

Machine learning is used to predict atomization energies of a large set of diverse organic molecules, based on nuclear charges and atomic positions only, with a mean absolute error of 10 kcal/mol.

Key facts:

- Data set: 7165 small organic compounds, DFT PBE0 reference energies
- Descriptor: Eigenvalues of Coulomb matrix
- Prediction: Kernel ridge regression, Gaussian kernel
- Accuracy: 10 kcal/mol mean absolute error

For details, see arXiv preprint [1].

Relevance

- Close to electronic structure theory accuracy at a fraction of the cost
- Potentially allows large scale applications
- Figure 1 presents energy curves beyond equilibrium geometries to demonstrate transferability and applicability

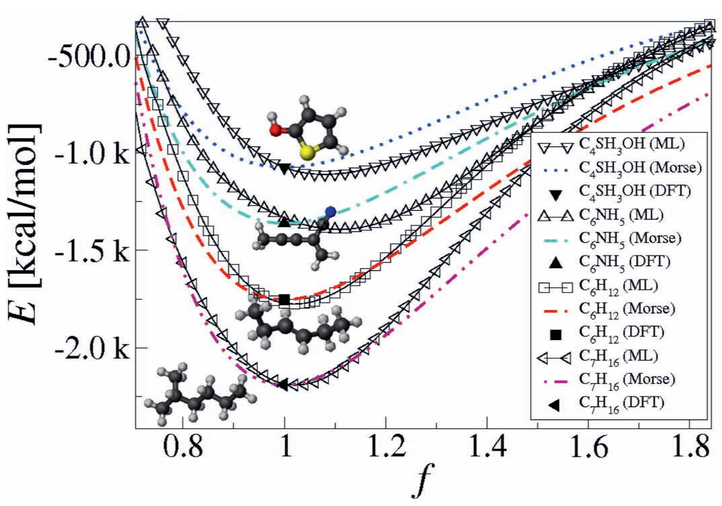


Figure 1: Energies of atomization curves of four molecules. Training set (n=1000) was augmented with scaled molecules representing repulsive wall, dissociative limit, and energy minimum, without computing additional reference energies. Resulting model distinguishes molecules with different bond types.

2 Data

Data set

- All molecules of GDB-13 database [2] with up to 7 C, N, O, S atoms (n=7165)
- Low energy geometries by OpenBabel [3]
- Constitutional, but not conformational isomers
- Density functional theory (DFT) level reference energies (PBE0 functional [4]) by FHI-aims [5] (tight settings/tier2 basis set)
- Energies range from -400 to -2200 kcal/mol

Representation

- Uses same information as Hamiltonian in electronic structure calculation: Cartesian atom coordinates $\{\boldsymbol{R}_i\}$ and nuclear charges $\{Z_i\}$
- Define Coulomb matrix as

$$\mathbf{M}_{ij} = \begin{cases} 0.5 Z_i^{2.4} & \text{for } i = j \\ \frac{Z_i Z_j}{|\mathbf{R}_i - \mathbf{R}_j|} & \text{for } i \neq j \end{cases}, \quad \text{where } i, j \text{ are atom indices}$$

- Distance between molecules **M** and **M'** is Euclidean distance between their eigenvalue vectors **ε** and **ε'**, $d(\mathbf{M}, \mathbf{M}') = \sqrt{\sum_i |\epsilon_i - \epsilon'_i|^2}$
- Nuclear charges strongly influence distance (Figure 2)
- Distance *d* is invariant against translation, rotation, re-indexing of atoms

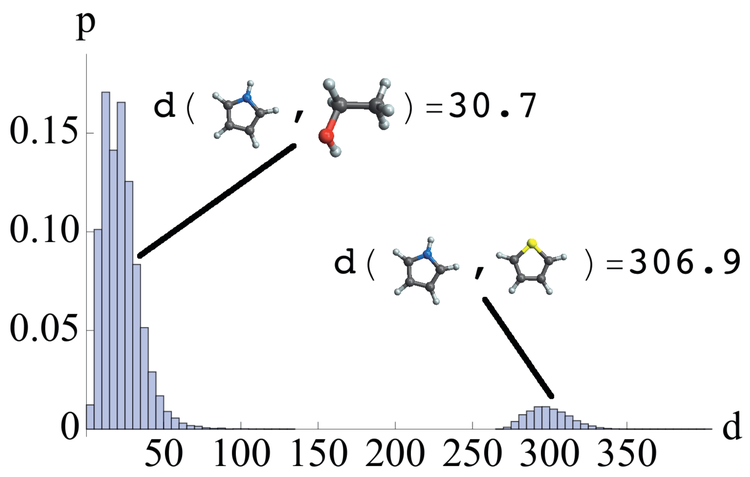


Figure 2: Distribution of distances d(**M**,**M'**) between all molecules in the data set. As an example, distances between pyrrol/ethanol and pyrrol/thiophene are shown.

3 Machine learning

Regression

- Kernel ridge regression [6], a non-linear regularized form of regression
- Gaussian kernel $k(\mathbf{M}, \mathbf{M}') = \exp\left(-d(\mathbf{M}, \mathbf{M}')^2 / 2\sigma^2\right)$ provides non-linearity. Length scale σ determines locality
- Model has form $E^{\text{est}}(\mathbf{M}) = \sum_{i=1}^{n} \alpha_i k(\mathbf{M}_i, \mathbf{M})$
- Regression weights **α** are the solution to the optimization problem

$$\min_{\alpha} \sum_i \left(E^{\text{est}}(\mathbf{M}_i) - E^{\text{ref}}(\mathbf{M}_i)\right)^2 + \lambda \sum_i \alpha_i^2$$

where λ controls strength of regularization.
Regularization keeps weights small to prevent over-fitting.

- Analytic solution has form $\alpha = (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{E}^{\text{ref}}$, where $\mathbf{K}_{ij}$ = k($\mathbf{M}_i$,$\mathbf{M}_j$), **I** is the identity matrix, and $\mathbf{E}^{\text{ref}}$ are the reference energies.

Validation

- Performance estimation via stratified 5-fold cross-validation [6,7]
- Stratification ensures training set covers energy range
- Hyper-parameters λ, σ are optimized in an inner (nested) loop of 5-fold cross-validation using a grid search
- For larger n, length scale σ gets smaller; regularization strength λ increases slightly (Figure 3)

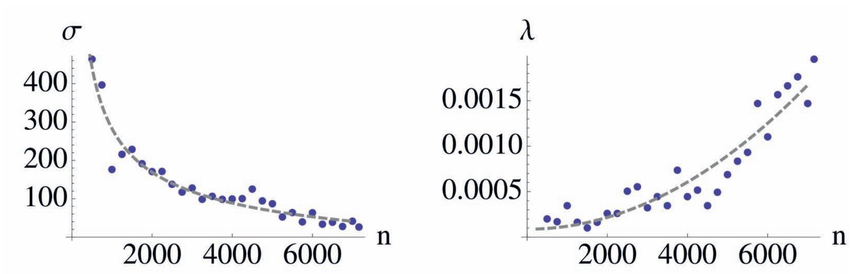


Figure 3: Optimal hyper-parameters over training set size.

4 Results

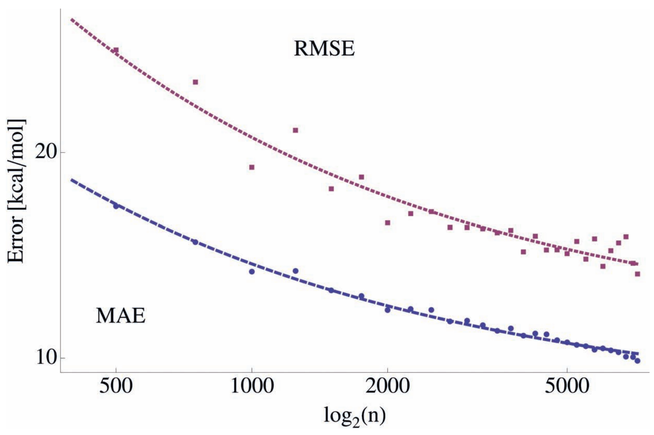
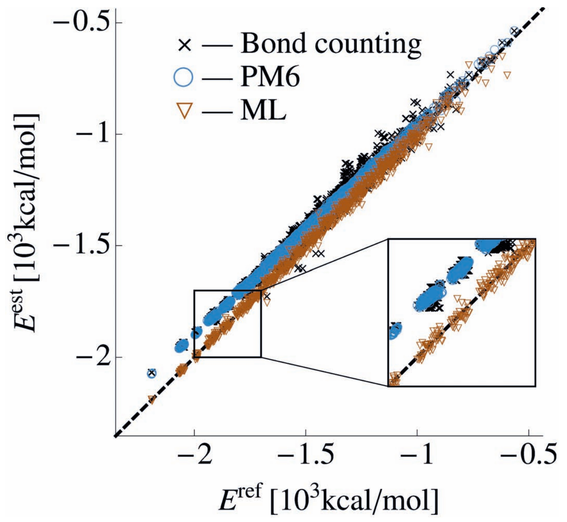


Figure 4: Cross-validated prediction error as a function of training set size.

- Performance depends on training set size (Figure 4)
- For n = 1000, RMSE = 20.4±1.6 kcal/mol, MAE = 14.6±0.8 kcal/mol, and R^2 = 0.996±0.002
- For n = 7000, MAE < 10 kcal/mol
- Statistical learning theory [6,8] suggests an error of ~7.6 kcal/mol for n→∞
- Figure 5 presents a scatter plot of DFT-PBE0 versus predicted energies
- Largest errors for small compounds due to fewer training samples
- Smaller systematic error than bond counting and semi-empirical PM6

Figure 5: Scatter plot of DFT PBE0 energies (E^{ref}) versus predicted energies (E^{est}). Bond counting and semi-empirical PM6 method shown for comparison.



References

[1] Rupp, M., Tkatchenko, A., Müller, K.-R., von Lilienfeld, O.A.: Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning, *arXiv* 1109.2618, 2011.

[2] Blum, L.C., Reymond, J.L.: 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13, *J. Am. Chem. Soc.* 131(25): 8732-8733, 2009.

[3] Guha, R., Howard, M.T., Hutchison, G.R., Murray-Rust, P., Rzepa, H., Steinbeck, C., Wegner, J., Willighagen, E.L.: The Blue Obelisk—Interoperability in Chemical Informatics, *J. Chem. Inf. Model.* 46(3): 991-998, 2006.

[4] (a) Perdew, J.P., Ernzerhof, M., Burke, K.: Rationale for Mixing Exact Exchange with Density Functional Approximations, J. Chem. Phys. 105(22): 9982-9985, 1996. (b) Ernzerhof, M., Scuseria, G.E.: Assessment of the Perdew–Burke–Ernzerhof Exchange-Correlation Functional, J. Chem. Phys. 110(11): 5029-5036, 1999.

[5] Blum, V., Gehrke, R., Hanke, F., Havu, P., Havu, V., Ren, X., Reuter, K., Scheffler, M.: *Ab initio* Molecular Simulations with Numeric Atom-Centered Orbitals, *Comput. Phys. Comm.* 180(11): 2175-2196, 2009.

[6] Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Data Mining, Inference, and Prediction, Springer, New York, 2009.

[7] Müller, K.-R., Mika, S., Rätsch, G., Tsuda, K.: An Introduction to Kernel-Based Learning Algorithms, *IEEE Trans. Neural. Network.* 12(2), 181-201, 2001.

[8] Müller, K.-R., Finke, M., Murata, N., Schulten, K., Amari, S.: A Numerical Study on Learning Curves in Stochastic Multilayer Feedforward Networks, *Neural Comput.* 8(5): 1085-1106, 1996.

IPAM · berlin · Max-Planck-Gesellschaft · Argonne National Laboratory · Marie Curie Actions · ETH Eidgenössische Technische Hochschule Zürich Swiss Federal Institute of Technology Zurich