JCTC Journal of Chemical Theory and Computation

Article

pubs.acs.org/JCTC

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

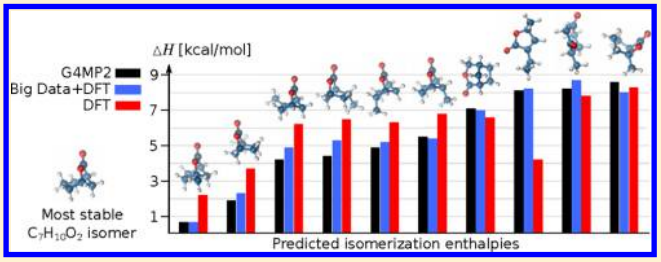
[¶]Max-Planck-Institut für Kohlenforschung, Kaiser-Wilhelm-Platz 1, 45470 Mülheim an der Ruhr, Germany

[‡]Computer-Chemie-Centrum and Interdisciplinary Center for Molecular Materials, Department Chemie und Pharmazie, Friedrich-Alexander-Universität Erlangen-Nürnberg, Nägelsbachstraße 25, 91052 Erlangen, Germany

[§]Argonne Leadership Computing Facility, Argonne National Laboratory, 9700 S. Cass Avenue, Lemont, Illinois 60439, United States

Supporting Information

ABSTRACT: Chemically accurate and comprehensive studies of the virtual space of all possible molecules are severely limited by the computational cost of quantum chemistry. We introduce a composite strategy that adds machine learning corrections to computationally inexpensive approximate legacy quantum methods. After training, highly accurate predictions of enthalpies, free energies, entropies, and electron correlation energies are possible, for significantly larger molecular sets than used for training. For thermochemical properties of up to 16k isomers of $C_7H_{10}O_2$ we present numerical evidence that chemical accuracy can be reached. We also predict electron correlation energy in post Hartree−Fock methods, at the computational cost of Hartree−Fock, and we establish a qualitative relationship between molecular entropy and electron correlation. The transferability of our approach is demonstrated, using semiempirical quantum chemistry and machine learning models trained on 1 and 10% of 134k organic molecules, to reproduce enthalpies of all remaining molecules at density functional theory level of accuracy.



INTRODUCTION

Designing new molecular materials is one of the key challenges in chemistry, and a major obstacle in solving many of the pressing issues that today's society faces, such as clean and cheap water, advanced energy materials, or novel drugs to fight antibiotic-resistant bacteria. Unfortunately, the number of potentially interesting small molecules is too large for exhaustive screening,[1−3] even when relying on automated synthesis and combinatorial high-throughput "click-chemistry".[4,5] Virtual screening strategies, made feasible by ever increasing compute power, advanced atomistic simulation software, and quantitative structure−property relationships have already helped in the discovery of new materials, and provided crucial guidance for subsequent experimental characterization and fabrication.[6−11] To achieve the overall goal of *de novo in silico* molecular and materials design[12−15] however, substantial progress is still necessary,[16] especially regarding prediction accuracy, computational speed, and transferability of the employed models.

For quantum chemistry models to attain "chemical accuracy" (≈1 kcal/mol) in the prediction of covalent binding is crucial in many scientific domains. Examples include the understanding of combustion processes,[17−19] questions relevant to interstellar chemistry,[20] and prediction of reaction rates essential for catalysis. The latter depend exponentially on energy differences, implying that small errors on the order of k_BT propagate dramatically. More generally, reaching chemical accuracy can be crucial for the detection of new structure−property relationships, trends, or patterns in Big Data, the design of new molecular materials with sensitive property requirements, or the energetics of competing reactants and products determining mechanisms and reaction rates. Control over the accuracy of important thermochemical properties of molecules can be achieved through application of well-established hierarchies in quantum chemistry.[21] Calibrated composite methods such as John Pople's Gaussian model chemistry exploit the inherent transferability of corrections to electronic correlation, the Born−Oppenheimer approximation, or basis-set deficiencies.[22,23] This has enabled chemists to routinely achieve chemical accuracy for *any* nonexotic and medium-sized organic molecule at substantial yet manageable computational costs.[24,25]

Unfortunately, such calculations are too demanding for the routine investigation of larger subsets of chemical space. Note, however, that the computationally most demanding task in a quantum chemistry calculation corresponds to an energy contribution that constitutes only a minor fraction of the

Received: February 3, 2015
Published: April 10, 2015

DOI: 10.1021/acs.jctc.5b00099
J. Chem. Theory Comput. 2015, 11, 2087−2096

total energy, while most of the relevant physics can already be accounted for through computationally very efficient approximate legacy quantum chemistry, such as the semiempirical theory PM7, Hartree−Fock (HF), or even density functional theory (DFT). For the water molecule H_2O, for example, HF predicts 90% of the experimental ionization potential.[26] Calculating the remaining Δ with chemical accuracy using correlated electronic structure methods requires a disproportionate amount of computational effort because of unfavorable prefactors and scaling with number of electrons. In this study, we introduce an alternative *Ansatz* to model the expensive Δ using a statistical model trained on reference data requiring only a fraction of the computational cost. The observed speedup, up to several orders of magnitude, is due to the computational efficiency of machine learning (ML) models. We have validated this idea for several molecular properties, combining quantum chemistry results at several levels of theory with Δ-ML models trained over comprehensive molecular data sets drawn from 134k organic molecules published in ref 27. While the basic idea to augment approximate models with ML is not new,[28−30] we present a generalized Δ-ML-model that achieves unprecedented chemical accuracy and transferability.

We present numerical evidence for predicted atomization enthalpies, free energies, and electron correlation in many thousands of organic molecules (reaching molecular weights of up to 150 Da) with an accuracy of ≈1 kcal/mol at the computational cost of DFT or PM7. We validate the Δ-ML model for entirely new subsets of chemical space, up to 2 orders of magnitude larger than the set used for training. Using Δ-ML-based screening, we find that within the constitutional isomers of $C_7H_{10}O_2$, molecular entropy and electron correlation energy of atomization are not entirely independent from each other. This suggests not only significant correlation between electronic and vibrational eigenstates but also the existence of Pareto fronts that can impose severe limitations with respect to simultaneous property optimization. Finally, we establish transferability by accurately predicting properties for a much larger molecular data set comprising 134k molecules.

THE Δ-ML APPROACH

The Δ_b^t-model of a molecular property corresponds to a baseline (b) value plus a correction, toward a targetline (t) value, modeled statistically. More specifically, given a property P'_b, such as the energy E_b, for the relaxed geometry R_b of a new query molecule, calculated using an approximate baseline level of theory, another related property P_t, such as the enthalpy H_t, corresponding to a more accurate and more demanding target level of theory can be estimated as

$$P_t(R_t) \approx \Delta_b^t(R_b) = P'_b(R_b) + \sum_{i=1}^{N} \alpha_i k(R_b, R_i) \quad (1)$$

The sum represents an ML-model, here a linear combination of Slater-type basis functions, $k(R_b, R_i) = e^{-|R_i - R_b|/\sigma}$, centered on N training molecules, and with global hyperparameter σ—the kernel's width. The regression coefficients $\{\alpha_i\}$ are obtained through kernel ridge regression, a regularized nonlinear regression model[31] that limits the norm of regression coefficients, thereby reducing overfitting and improving the transferability of the model to new molecules. $|R_i - R_b|$ is a quantitative measure of similarity between query molecule and training molecule i, using the Manhattan-norm (L_1) between sorted Coulomb matrix representations.[32,33] The latter uniquely encodes (except among enantiomers) the external potential of any given molecule in a way that is invariant with respect to molecular translation, rotation, or atom-indexing. We note that while atom sorting can lead to property differentiability issues with respect to geometry changes, here we only study molecules in well separated potential energy minima. For the sorted Coulomb matrix (CM) descriptor, the combination of Slater-type kernel basis functions with L_1 norm has been shown to yield the most accurate ML model for molecular atomization energies.[33] As such, P_t of a new molecule, consistent with its minimum geometry R_t at the target level of theory, is estimated using exclusively R_b and P'_b as input. Thus, the Δ-model accounts for differences in (i) definition of property observable, for example, energy → enthalpy; (ii) level of theory, for example, PM7 → G4MP2; and (iii) changes in geometry (illustrated in Figure 1).

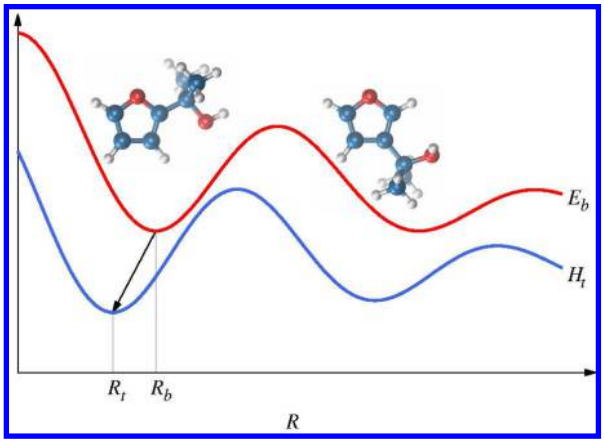


Figure 1. Two hypothetical property profiles connecting two constitutional isomers of $C_7H_{10}O_2$. The Δ-model, eq 1, estimates the difference between baseline and targetline properties (arrow) which differ in level of theory ($b \rightarrow t$), geometry ($R_b \rightarrow R_t$), and property ($E_b \rightarrow H_t$).

As a first test of our *Ansatz*, we have trained Δ_b^t models for HOMO and LUMO eigenvalues calculated at various levels of theory[34] for the smallest 7k organic molecules in the GDB-data set introduced by Reymond and co-workers.[35] After training on calculated data for 1k molecules, the resulting "1k Δ-model" reduces the mean absolute error (MAE) in the prediction of GW HOMO eigenvalues for the remaining 6k molecules from 0.78 to 0.23 eV for the semiempirical ZINDO baseline method. Interestingly, while the less empirical DFT hybrid (PBE0) baseline method has an MAE of more than 2 eV, this reduces to less than 0.1 eV when combined with the 1k $\Delta_{\mathrm{PBE0}}^{\mathrm{GW}}$-model. Correspondingly, MAEs for predicting GW LUMO eigenvalues reduce from 0.91 to 0.16 eV and from 1.3 to 0.13 eV for $\Delta_{\mathrm{ZINDO}}^{\mathrm{GW}}$ and $\Delta_{\mathrm{PBE0}}^{\mathrm{GW}}$, respectively. This observation suggests that more sophisticated baseline models, albeit occasionally leading to more substantial errors than simpler models, overall are smoother in chemical space, and therefore easier to learn.

RESULTS AND DISCUSSION

Chemically Accurate Prediction of Covalent Bonding.

To demonstrate that the Δ-ML model can reach chemical accuracy, we have investigated the covalent binding energies of the 6k constitutional isomers of $C_7H_{10}O_2$ in the GDB.[35] Note that any other molecular set could have been used just as well. We relied on previously calculated highly accurate target level

DOI: 10.1021/acs.jctc.5b00099
J. Chem. Theory Comput. 2015, 11, 2087−2096

atomization energies (E_t in Figure 1) at the G4MP2 level[25] for these isomers.[27] G4MP2 is widely considered to be on par with experimental uncertainties.[36] While structurally highly diverse, this data set exhibits many near-degeneracies with high energy densities of up to ∼100 molecules per kcal/mol in atomization enthalpy H (inset of Figure 2).

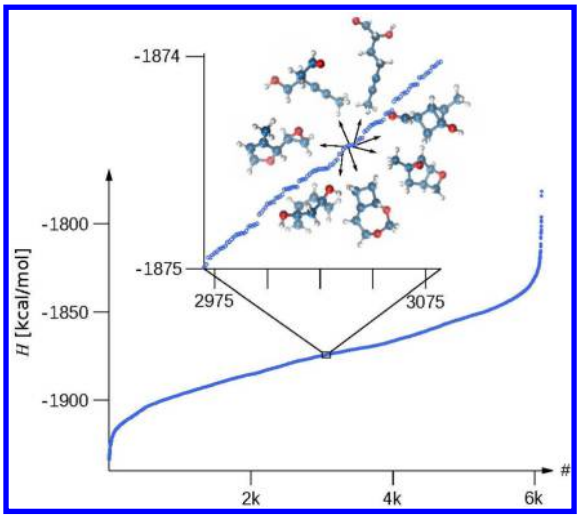


Figure 2. Illustration of chemical diversity and data density of up to ∼100 molecules per kcal/mol of atomization enthalpy H (G4MP2 level of theory[24,25]), shown in ascending order for all 6k constitutional isomers of $C_7H_{10}O_2$ in the GDB-17 data set.[27,35] Seven near degenerate (within ∼0.01 kcal/mol) molecules in the inset exemplify the chemical diversity.

The Δ-ML model's systematic improvement of accuracy with increasing training set size is shown as a log−log plot in Figure 3 for the potential energy of atomization. Starting at different offsets, corresponding to the respective error of the pure baseline methods, the MAE, measured out-of-sample on the remaining molecules in the 6k set, rapidly decreases. Note the constant decay rate for training set sizes larger than 1000 for all Δ_b^{G4MP2}-ML model errors. This suggests that for all models the error could be lowered even further if more training data were used.

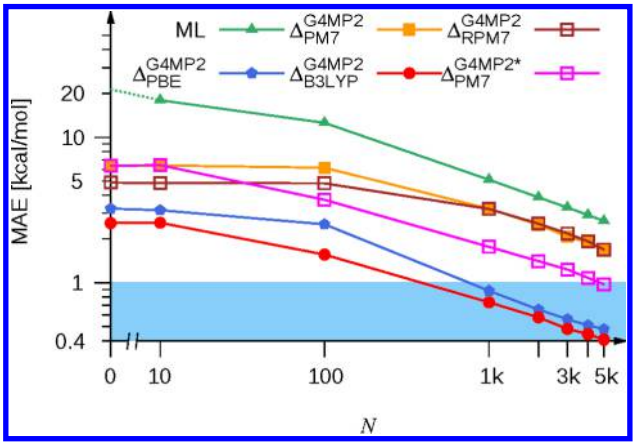


Figure 3. Mean absolute error (MAE) [kcal/mol] of ML predicted atomization energies compared to G4MP2 reference values as a function of training set size N for out-of-sample predictions. The lines correspond to various baselines in the Δ_b^{G4MP2}-model (eq 1). The MAE at $N = 0$ represents the baseline's error. For comparison, the baseline-free ML model is shown as well (green); its $N = 0$ value is the standard deviation in the G4MP2 atomization energies of the data set. The "chemical accuracy" target of 1 kcal/mol is highlighted in blue. Δ^{G4MP2}_{RPM7} (brown) and Δ^{G4MP2*}_{PM7} (pink) are variants of Δ^{G4MP2}_{PM7} (yellow) using reparameterized PM7 as baseline or an alternative molecular representation, respectively.

While the error of the GGA DFT baseline model (Δ^{G4MP2}_{PBE}) in Figure 3 starts off slightly higher than the more accurate hybrid DFT analogue (Δ^{G4MP2}_{B3LYP}), both rapidly converge to chemical accuracy (<1 kcal/mol) for less than 1k training molecules. Δ^{G4MP2}_{B3LYP} reaches ∼0.4 kcal/mol for a 5k training set. For the substantially faster, yet more approximate PM7 baseline model (Δ^{G4MP2}_{PM7}) an accuracy similar to pure hybrid DFT (B3LYP) can be achieved with 2k training molecules, and for 5k training molecules the error has been quenched to less than 2 kcal/mol. All Δ-ML models outperform an ML model trained on the absolute value of E_t directly,[32,33] without a baseline.

We also considered the effect of reparameterizing the baseline method on the training set before applying Δ-ML. The arguably simplest model, Benson's thermochemical bond additivity model (bond counting), has a prediction error of MAE ≈ 100 kcal/mol for the 6k isomers. After reparameterization of all bond energies to fit the training data, its MAE reduces to ∼30 kcal/mol, which is worse than direct ML. Along the same line, we optimized all semiempirical parameters in PM7 in order to reproduce G4MP2 atomization enthalpies of up to 128 $C_7H_{10}O_2$ isomers, drawn at random. Figure 3 (brown line) shows that when using this reparameterized (RPM7) baseline model in Δ^{G4MP2}_{RPM7} it does allow for an improved offset; however, the advantage vanishes when the training set size is increased beyond 1k (Figure 3). Using the semiempirical model OM2,[37] instead of PM7, we found a similarly vanishing effect. These findings indicate the severe limitations inherent in fixed functional forms of electronic semiempirical model Hamiltonians used in combination with globally optimized parameters—no matter the actual combination of parameters. Statistically learned Δ-corrections, inferred from large numbers of example molecules, however, seem capable of capturing the more delicate energy contributions in the G4MP2 energy. We have also tested the effect of using an alternative molecular representation. The Δ^{G4MP2*}_{PM7} model in Figure 3 (pink line) shows the improvement of performance of the Δ^{G4MP2}_{PM7} model when replacing the above-mentioned Coulomb-matrix representation by the bag-of-bond descriptor recently introduced by two of us.[38] Encouragingly, also for this descriptor one observes similar decay rates and an even better performance than for the Coulomb-matrix based Δ-model, reaching chemical accuracy for a training set size of 5k.

Chemically Accurate Thermochemistry. Prediction accuracy for thermochemical properties, such as enthalpies, and free energies of atomization at 298.15 K, all trained to reproduce the G4MP2 target level of theory for the same set of 6k constitutional isomers of $C_7H_{10}O_2$ were investigated. Δ-ML models have been trained for three baselines, Δ^{G4MP2}_{PM7}, Δ^{G4MP2}_{PBE}, and Δ^{G4MP2}_{B3LYP}, on subsets of varying sizes. The baseline properties corresponded in this case simply to the potential energy of atomization, with the ML model accounting for differences in level of theory, in geometry, as well as for the respective thermodynamic effects. Table 1 lists resulting errors and standard deviations of predicted enthalpies of atomization at 298.15 K for various training set sizes. As before, Δ-ML models display rapid error decay with increasing training set size. Encouragingly the standard deviation also decays rapidly with

DOI: 10.1021/acs.jctc.5b00099
J. Chem. Theory Comput. 2015, 11, 2087−2096

Table 1. Mean Absolute Errors ± Standard Deviations for Predicted out-of-Sample Enthalpies of Atomization H (T = 298.15 K) at G4MP2 Level of Theory Using the Δ_b^{G4MP2}-ML Model for Increasing Training Set Size N in eq 1[a]

N	$\triangle^{G4MP2}_{PM7}$	$\triangle^{G4MP2}_{PBE}$	$\triangle^{G4MP2}_{B3LYP}$
0	6.4 ± 8.6	3.0 ± 4.1	2.5 ± 3.1
0.1k	5.7 ± 7.6	2.2 ± 2.9	1.5 ± 1.9
1k	3.9 ± 4.1	0.8 ± 1.1	0.7 ± 0.9
2k	2.4 ± 3.1	0.6 ± 0.8	0.6 ± 0.7
3k	2.2 ± 2.8	0.5 ± 0.7	0.5 ± 0.6
4k	1.9 ± 2.4	0.5 ± 0.6	0.4 ± 0.6
5k	1.7 ± 2.2	0.5 ± 0.6	0.4 ± 0.5

[a]All values in kcal/mol. Training and test set sizes always add up to 6095 constitutional isomers of $C_7H_{10}O_2$.

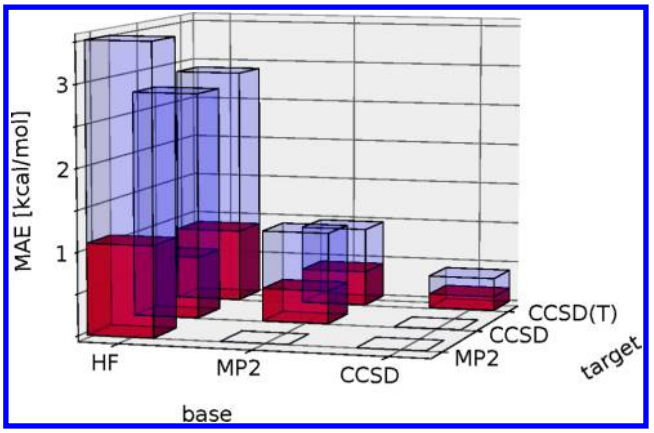


Figure 4. Mean absolute error (MAE) [kcal/mol] of E_b (blue bars) and 1k Δ_b^t-ML model predictions (red bars) of atomization energies for various combinations of increasingly correlated post-Hartree−Fock methods as target and baseline methods. These values are obtained after subtracting the systematic average shift that exists between methods for the 6k isomers. See Supporting Information for details.

training set size. Again, the DFT baseline models yield MAEs of less than 1 kcal/mol already at 1k training set size, and the error of the 5k-Δ^{G4MP2}_{B3LYP}-ML model remains below even after the addition of the standard deviation. The computationally less expensive PM7 baseline model performs slightly worse than the DFT-based models. The 1k-Δ^{G4MP2}_{PM7}-ML model decreases the pure PM7 prediction error and standard deviation by more than ∼50%, and converges to near chemical accuracy (1.7 kcal/mol) for a 5k training set. Computational effort for out-of-sample predictions is dominated by baseline evaluations. For free energies of atomization we have observed nearly identical convergence and baseline trends. All these results indicate that the Δ-ML approach represents an inexpensive strategy to accurately estimate not only differences in potential energies due to different electronic structure models, but to also account for thermal contributions to thermodynamic state functions *without* having to calculate the corresponding partition functions. Note that the latter can be prohibitively expensive when using more accurate theories.

Electron Correlation. To further assess the applicability of the Δ-ML *Ansatz*, we modeled electron correlation energies, essential for achieving chemical accuracy.[21] Within post-HF theory, electron correlation energy can be defined as the difference between converged basis-set HF potential energy and its corresponding nonrelativistic exact counterpart.[26] Evaluating the many-electron correlation energy at the post-HF level of theory requires substantial computational effort. The computational complexity of the simplest post-HF method, second order perturbation theory (MP2), scales as N_e^5, where N_e is the number of electrons. The "gold standard" of quantum chemistry, CCSD(T), even scales as N_e^7. Revisiting the 6k constitutional isomers of $C_7H_{10}O_2$ (Figure 2), we have calculated the difference in the correlation energy part of molecular atomization energies for various correlated methods.

For the 6k isomers, MAEs, after accounting for systematic shifts, are shown in Figure 4 for various combinations of HF, MP2, CCSD, and CCSD(T), along with the reduction of error due to 1k-Δ_b^t models. Particularly noteworthy is the result for the $\Delta_{HF}^{CCSD(T)}$ model giving the total correlation energy (defined as the difference between HF and the exact result as approximated by CCSD(T)) between the least and most expensive method: The MAE is reduced from ∼2.9 to less than 1 kcal/mol. Note that the 1k-Δ_{HF}^{MP2} model has a larger MAE from MP2 than the MAE of the 1k-Δ_{HF}^{CCSD} from CCSD, even though MP2 is more approximate in nature than CCSD. As such, the MP2 correlation energy, that is, the difference between HF and MP2, albeit being less accurate and more approximate than the CCSD correlation energy, appears to be a more complex function in chemical space.

Remarkably, when using the 1k-Δ_{MP2}^{CCSD} or 1k-$\Delta_{MP2}^{CCSD(T)}$ models the MAE amounts in *both* cases to less than 0.5 kcal/mol. This suggests the possibility of the proverbial free lunch modeling the more accurate and computationally more expensive CCSD(T) level of theory with the same training set size and precision as the less accurate and less expensive CCSD method. The least approximate baseline, encoded by the 1k-$\Delta_{CCSD}^{CCSD(T)}$ model, yields a chemically nearly negligible MAE from CCSD(T), ∼0.1 kcal/mol. For larger training set sizes we have observed correlation energy errors to decay similarly to the error of DFT baseline models for thermodynamic properties.

Applicability: Diastereomers of $C_7H_{10}O_2$. We have tested the applicability of the 1k-Δ-ML model, trained on 1k out of the 6k $C_7H_{10}O_2$ isomers in the GDB database, for the identification of the most stable diastereomers that can be generated from the parent isomers. Such screening applications are highly relevant for spectroscopic or computational experiments aimed at the discovery and characterization of competing reaction pathways, recently discussed for an "*ab initio* nanoreactor".[39] More specifically, we applied the 1k-Δ^{G4MP2}_{B3LYP} model of atomization enthalpy at 298.15 K (Table 1), to screen all the 9868 unique and stable diastereomers resulting from inversion of atomic stereocenters in the original GDB set of 6095 constitutional isomers of $C_7H_{10}O_2$ (see Methods section). For validation, we have randomly drawn 3k diastereomers and calculated their computationally demanding G4MP2 enthalpies of atomization. The 1k-Δ^{G4MP2}_{B3LYP} model yields a MAE of 0.8 kcal/mol for these 3k diastereomers. We have chosen the DFT baseline for this exercise because cheaper baseline models, such as 5k Δ^{G4MP2}_{PM7} and Δ^{G4MP2*}_{PM7} (Figure 3), exhibit less transferability when validated on the G4MP2 results for the 3k diastereomers, namely MAEs of 3.5 and 2.8 kcal/mol, respectively.

Out of all the 10k diastereomers, the 1k-Δ^{G4MP2}_{B3LYP} model predicts 6-oxabicyclooctan-7-one, which is caprolactone with a methyl bridge between positions 1 and 5, with an estimated atomization enthalpy H of −1933.5 kcal/mol, to be the most stable isomer at ambient conditions. A validating G4MP2

DOI: 10.1021/acs.jctc.5b00099
J. Chem. Theory Comput. 2015, 11, 2087−2096

calculation yielded the same number. Figure 5 shows this molecule along with its 10 enthalpically closest isomers. These

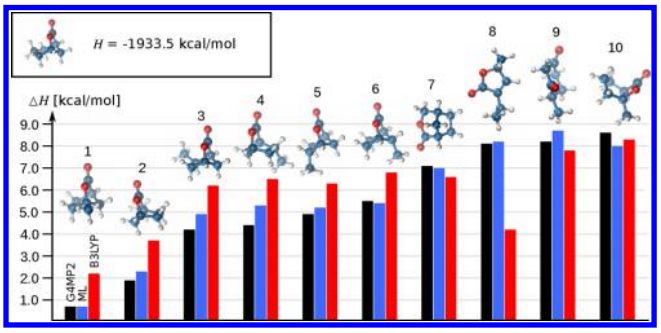


Figure 5. Calculated reaction enthalpies at 298.15 K between the most stable molecule with $C_7H_{10}O_2$ stoichiometry (6-oxabicyclooctan-7-one, in inset, with atomization enthalpy, −1933 kcal/mol), and its 10 energetically closest isomers in increasing order, according to targetline method G4MP2 calculated *a posteriori* (black). 1k Δ^{G4MP2}_{B3LYP} ML model predictions are given in blue. Baseline method B3LYP predictions are shown for comparison (red).

span a narrow energetic window of 9 kcal/mol, which is sparse in comparison to the aforementioned 100 molecules/kcal/mol energy density. The six isomers, for which $\Delta H < 6$ kcal/mol, correspond to diastereomers of oxabicyclo[2.2.1]heptan-3-one, methylated at the 1,4,5,5,6,7 positions. Isomers 3 (ΔH = 4.3 kcal/mol) and 4 (ΔH = 4.5 kcal/mol) differ only by the chirality of the carbon atom at position 1. The next four high-lying isomers, although populating only a narrow ΔH range of 7−9 kcal/mol, exhibit very diverse chemical structures: They include a cyclopentane fused with γ-valerolactone, a methyl, ethyl-substituted furanone, a methylated cyclohexanedione, and a cyclopentane fused with β-propiolactone and methylated bridge atom framework. After identification of these isomers, we calculated the validation of G4MP2 enthalpies (Figure 5). The 1k-Δ^{G4MP2}_{B3LYP} ML model estimates the isomerization enthalpy of these products with a maximal error of 0.6 kcal/mol for product 10. The ML-model predictions agree with G4MP2 results calculated *a posteriori*, and never exceed the threshold of chemical accuracy (1 kcal/mol).

For comparison, Figure 5 also features estimates obtained from the DFT baseline method B3LYP, which is popular among many computational as well as experimental chemists. While B3LYP would have predicted the same global minimum, its reaction enthalpies can deviate substantially, and, sometimes fail spectacularly (isomer 8). It is interesting to note that the ML-model is apparently able to reduce or increase the estimate depending on its baseline overshooting (isomers 1−6) or underestimating (isomers 7−9). Only in the case of isomer 10, use of the ML-model would deteriorate the baseline's prediction error, albeit only from 0.3 to 0.5 kcal/mol. We believe that such overall agreement of predicted reaction enthalpies with G4MP2 results obtained *a posteriori* strongly indicates that the Δ-ML *Ansatz* is capable of accounting for subtle errors made in the prediction of competitive chemical bonding—at the baseline's computational cost (in this case DFT).

Interpretation of the Δ-Model. One can understand the trained corrections as follows: The $\Delta^{CCSD(T)}_{HF}$ ML model of atomization energies can be viewed as a ML model of the correlation energy of atomization. Likewise, when using atomization energies as baseline properties for free energies, and enthalpies, the difference in the resulting ML models cancels the baseline energy and corresponds, after division by T, to the entropy of atomization

$$(\Delta^H_E - \Delta^G_E)/T = S \quad (2)$$

Using a random 1k subset of the 6k $C_7H_{10}O_2$ isomers, we have trained two ML models, one on S of atomization at G4MP2 level of theory, taken as $(H - G)/T$ from ref 27, the other on correlation energy of atomization, E_c, (i.e., HF energy of atomization subtracted from the CCSD(T) counterpart), also from ref 27. Computationally efficient PM7 equilibrium geometries have been used for training, testing, and predictions.

We have reapplied the resulting 1k models to screen the aforementioned 10k diastereomers for those molecular pairs which exhibit maximal isomerization entropies ΔS and correlation energies ΔE_c. Structures of the molecules with extreme S and E_c are shown in Figure 6. The molecular pair

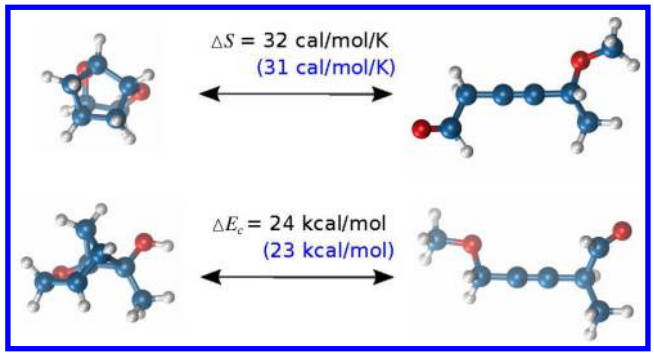


Figure 6. Molecular pairs with maximal reaction entropy (top) and electron correlation energy (bottom) out of the 10k stable diastereomers of $C_7H_{10}O_2$. Reaction properties have been estimated using a 1k-Δ^{G4MP2}_{B3LYP}-ML model for the entropy and a 1k-$\Delta^{CCSD(T)}_{HF}$-ML model for the correlation energy (black). Corresponding reference values calculated for validation are given in parentheses (blue).

with maximal ΔS is consistent with chemical intuition: The lowest entropy isomer, 2,5-dioxatricyclononane, has a cage-like structure and is very compact, bearing some resemblance to adamantane, and void of any conformational degree of freedom. By contrast, the molecule with largest entropy, 5-methoxyhex-3-ynal, possesses multiple conformational degrees of freedom, made possible through the occurrence of a double and a triple bond that consume the valences otherwise accessible for ring or cage formation. The resulting ΔS estimated by the ML model deviates from the reference G4MP2 value by only 1 cal/mol/K. Quantitative rationalization of the trend in E_c is less obvious, owing to its origin in electronic many-body effects. However, E_c can be intuitively related to the number of interacting electron pairs. This number is small when the molecule is long since, according to the nearsightedness principle,[40] electrons localized on one end of the molecule interact less with those from the other end. In compact molecules by contrast, more electrons can interact, hence the number of interacting electron pairs, and consequently the magnitude of E_c, are large. After screening the diastereomers using our E_c model we found among 10k diastereomers that 6-methyl-2-oxatricycloheptan-6-ol and 5-methoxy-2-methylpent-3-ynal have the maximal reaction electron correlation energy, see Figure 6. The maximal electron correlation energy difference, 24 kcal/mol, deviates from validating CCSD(T) reference results by 1 kcal/mol. Both

DOI: 10.1021/acs.jctc.5b00099
J. Chem. Theory Comput. 2015, 11, 2087−2096

molecules confirm intuition: The most compact molecule with few degrees of freedom exhibits maximal correlation while the most elongated corresponds to the least amount of correlation energy.

In the above discussion we notice that for both pairs with maximal difference in electron correlation and in entropy, respectively, similar observations hold: Compactness/extension appears to maximize the difference in both cases. This observation raises the question whether S, which arises from the vibrational partition function, and E_c, due to electronic many-body effects, are interdependent. An underlying relationship could aid not only in pin-pointing molecules that pose interesting challenges for benchmarking approximate electronic theories but might even lead to semiquantitative estimations of E_c via S. Thermal molecular properties, such as heat-capacities, could be linked directly to their electronic structure. To elucidate their potential relationship, Figure 7 shows a scatter plot of the model-predicted atomization entropy and correlation energies of the 10k diastereomers, as well as the 6k parent isomers of $C_7H_{10}O_2$.

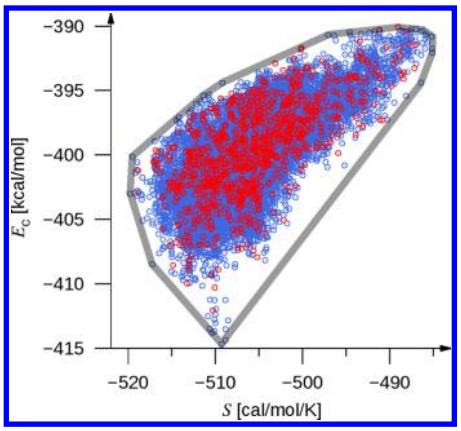


Figure 7. Scatter plot of ML-model predicted entropy of atomization at T = 298.15 K versus ML-model predicted electron correlation contribution to atomization energy for the 16k stable out-of-sample diastereomers of $C_7H_{10}O_2$. All predictions are made using ML models trained on 1k molecules, randomly drawn from a set of 6k parent isomers. Training data are shown in red. Linear regression yields $E_c \approx 0.445 \times TS - 175.211$ [kcal/mol]. Pareto fronts are indicated by convex hull shown in gray.

Such a dependency might serve the construction of rough structure−property relationships for the filtering of compounds using one property as a mutual descriptor for the other. Furthermore, this relationship could possibly impose severe constraints on how freely S and E_c can be varied independently within multiobjective property optimization procedures in chemical compound space. To further illustrate this point, Figure 7 also highlights corresponding Pareto fronts. Note, for example, that while Figure 6 displays the pairs that maximize the vertical (E_c) or horizontal (S) axis in Figure 7, the molecular pair that simultaneously maximizes both differs. Other molecules, such as bullvalene, also happen to fall onto the same linear relationship. However, for organic molecules with very different sizes, taken from the 134k GDB-9 data set, this linear trend breaks down. As such, it might still require normalization by number of atoms or electrons to hold in general.

We finally note that arriving at these observations exclusively via high-throughput *ab initio* computations would have required N_e^7-scaling G4MP2 calculations for all the 10k diastereomers with an estimated need for compute time of ∼20 CPU years. The PM7 baseline predictions, by contrast, required only ∼1 CPU day for all geometry relaxations, and the remaining deviation from target properties G4MP2-S and CCSD(T)-E_c is given instantaneously by the ML correction.

Thermochemistry for 134k Organic Molecules. When dealing with hundreds of thousands of molecules one typically assumes that it is not necessary to achieve chemical accuracy for all of them. Instead, hierarchical procedures where less accurate but computationally more efficient methods, such as DFT, filter out the most relevant compounds which subsequently can be studied using more accurate and computationally more demanding methods, such as G4MP2. DFT calculations, however, are ordinarily too expensive to be used for filtering hundreds of thousands of molecules. We have investigated whether the Δ-approach can be used for DFT-quality filtering at the computational cost of a semiempirical quantum chemistry calculation. Specifically, based on PM7 baselines we have predicted DFT targetline enthalpies and entropies of atomization. To more systematically assess transferability, we have trained on a 1k and 10k training set drawn at random from the nearly 134k organic molecules containing up to nine C, N, O, or F atoms (published as GDB-9 in ref 27). For subsequent validation, we have used the remaining 133k and 124k molecules, respectively.

On average, PM7 enthalpies of atomization deviate from B3LYP by 7.2 kcal/mol. For a randomly drawn training set of 1k molecules, 1k-Δ_{PM7}^{B3LYP}-ML predicts B3LYP enthalpies of the 133k additional (out-of-sample) molecules with an MAE of 4.8 kcal/mol. Increasing the number of training molecules to 10k leads to an improved MAE of 3.0 kcal/mol, as measured for the remaining 124k out-of-sample molecules. We note that such a predictive accuracy places the 10k-Δ_{PM7}^{B3LYP}-ML model on par with generalized gradient approximated (GGA) or even hybrid DFT[36,41]—at the computational cost of PM7. Figure 8 features the corresponding scatter plot of actual versus predicted B3LYP enthalpies of atomization. The lower right inset shows that the baseline's systematic underestimation, as well as its skew, has been removed already by the 1k-Δ_{PM7}^{B3LYP}-ML model. The error distribution contracts further as the training set size is increased to 10k. The upper left inset plots ML estimated deviations of PM7 from B3LYP atomization energies versus reference values, revealing improvement in correlation with an increase in training data. The molecular structure on the top right is the most extreme outlier, PM7 underestimates its stability by 86.0 kcal/mol. Encouragingly, 1k and 10k-Δ_{PM7}^{B3LYP}-ML models reduce the error for this outlier to 73.9 and 58.0 kcal/mol, respectively.

We have also analyzed the effect of molecular shape and topology on the performance of both the baseline theory, and the Δ-ML models. It is well-known that the faithfulness of common quantum chemical methods can alter drastically when the geometry of the molecule is changed. Straining chemical bonds in cycles, for instance, can lead to severe errors, even for methods that predict the energy minimum perfectly well. To systematically assess the effect of geometry, we compare the predicted B3LYP atomization enthalpies for PM7 and 10k-Δ_{PM7}^{B3LYP} for all 134k molecules, as a function of their normalized principal moments of inertia. Figure 9 displays the resulting deviation from B3LYP, spanned by molecular geometry (rod, disk, or sphere-like). While PM7 has particularly strong deviations (∼20 kcal/mol) on the linear to planar edge, as well as close to the lower part of the linear to spherical edge,

DOI: 10.1021/acs.jctc.5b00099
J. Chem. Theory Comput. 2015, 11, 2087−2096

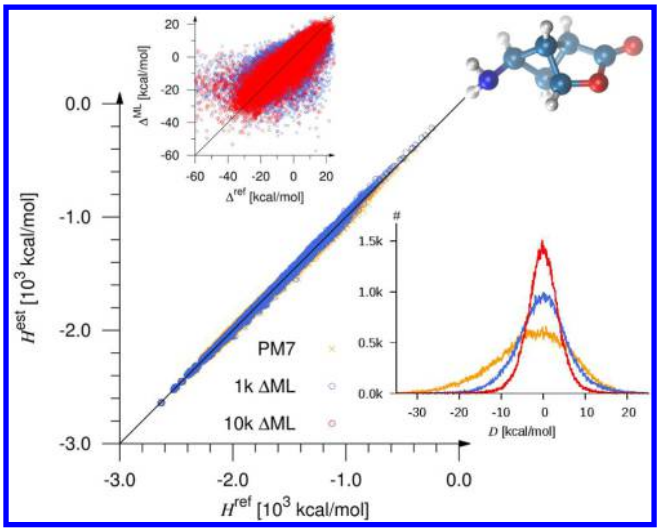


Figure 8. Scatter plot of predicted atomization enthalpies H at 298.15 K for 124k out-of-sample GDB-9 molecules.[35] Estimated H is plotted for PM7 (yellow, MAE = 7.2 kcal/mol) and 1k (blue) and 10k (red) Δ^{B3LYP}_{PM7}-ML models (MAE = 4.8 and 3.0 kcal/mol, respectively) versus reference B3LYP values. The left side inset shows the ML contribution to the estimated energy differences between PM7 and reference B3LYP enthalpies, Δ^{ML}, for 1k (blue) and 10k (red) models versus reference difference, Δ^{ref}. The right side inset shows the error distribution around B3LYP enthalpies for PM7, 1k, and 10k models, respectively. The most extreme outlier (top, right), 7-amino-3-oxatricycloheptan-4-one, has error 86, 74, and 58 kcal/mol for PM7, 1k, and 10k models, respectively.

use of the ML correction homogeneously quenches the error throughout the triangle into the 5 kcal/mol error window, with very few 20 kcal/mol outliers persisting on the rod−disk edge. Note that due to the nonuniqueness of the moments of inertia, error heatmaps in Figure 9 of many molecules superimpose each other in increasing order. To avoid possible misinterpretations, the inset with a heat-map of data density provides a means to visually normalize the error heatmaps.

Regarding the computational speed-up, we note that on a typical CPU, a single Δ^{B3LYP}_{PM7}-ML evaluation requires no more than 10 s for the largest molecule in GDB-9. Thus, screening of the entire set of 134k molecules has consumed less than 2 CPU weeks. By contrast, the average computational cost for obtaining a B3LYP atomization enthalpy amounts to roughly 1 CPU hour per molecule, implying 15 CPU years for DFT-based screening of the 134k molecules.

Conclusions. We have introduced a composite quantum chemistry/machine learning approach. It combines approximate but fast legacy quantum chemical approximations with modern big data-based machine learning estimates trained on expensive and accurate reference results throughout chemical space. We have shown that the Δ-ML model can be used to study other out-of-sample molecules, not part of training. Effectively one can reach the accuracy of high-level quantum chemistry at a dramatically lower computational burden which is dominated by the employed baseline method, such as semiempirical quantum chemistry (PM7), HF, or DFT. Mere reparameterization of the baseline method's global parameters for a given training set does not suffice, yielding measurable advantage only for very small and selected training and test sets. Alternative molecular representations, however, could still lead to faster convergence to chemical accuracy. Similar learning rates with respect to training set size among all model-combinations, merely differing by offset, suggest that even very approximate and computationally inexpensive baseline models can be used, provided there is access to sufficiently large training sets. For chemically diverse sets of organic molecules we have presented numerical evidence that chemically accurate molecular thermochemistry predictions can be made at a computational cost reduced by several orders of magnitude when compared to the reference method for new out-of-sample molecules.

For the most stable isomer in the set of 10k diastereomers generated from all 6k molecules with $C_7H_{10}O_2$ stoichiometry in GDB-17,[35] we have demonstrated how to identify the 10 most competitive reaction isomers. For the same diastereomers we also identified a qualitative dependency between entropy and correlation energy of atomization. Finally, we have presented evidence for the transferability of the Δ-ML model by reducing the error of semiempirical quantum chemistry method from 7.2 kcal/mol to the error of generalized gradient approximated (∼5 kcal/mol) or hybrid density functional theory (∼3 kcal/mol) for over a hundred thousand organic molecules using less than 1 and 10% of them for training, respectively.

We believe the high predictive accuracy to be due to the fact that approximate theories already capture the most important contributions to chemical energetics. The remaining deviations from the reference results are typically smaller, possibly also smoother, and prove to be more amenable to statistically trained ML models. Overall, our results suggest that the Δ-ML model represents an attractive strategy for augmenting legacy quantum chemistry with modern big data driven ML models. It would be interesting to investigate the proposed strategy in the context of accelerated first-principles predictions of various properties such as heat capacities, nonadiabatic energy corrections, barriers of elementary reaction steps, optical properties, atomic forces for molecular dynamics calculations, molecule specific parameters for semiempirical theories, or electronic excitations.

METHODS

Molecular Data Sets. We have considered four sets of organic molecules. The first set has been used for preliminary testing of the *Ansatz*, and consists of the 7211 (7k) organic molecules and HOMO/LUMO eigenvalues and molecular polarizabilities at different levels of theory as published in ref 34. The second set contains 133885 (134k) molecules with up to 9 heavy atoms (C, O, N, F, not counting H) in the universe of small organic molecules "GDB"[35] for which we calculated and published semiempirical (PM7) and density functional theory (B3LYP)-based thermochemical properties such as enthalpies and entropies of atomization.[27] The diversity of this set is shown in Figure 9. We note at this point that in Δ-ML models other baseline methods, such as extended Hückel, tight-binding DFT,[42] OM2,[37] or AM05[43] could have been used just as well. The third set corresponds to a subset of the second set: For 6095 (6k) constitutional isomers of $C_7H_{10}O_2$, we calculated the same thermochemical properties at a significantly more sophisticated and computationally demanding level of theory, widely considered to be of "chemical accuracy" (∼1 kcal/mol). Also this set has been published in ref 27. Finally, the versatility of this method is assessed for a fourth set of molecules, consisting of 9868 (10k) stable diastereomers that are not part of the GDB universe, and have been obtained by inverting all atomic stereocenters in the aforementioned third

DOI: 10.1021/acs.jctc.5b00099
J. Chem. Theory Comput. 2015, 11, 2087−2096

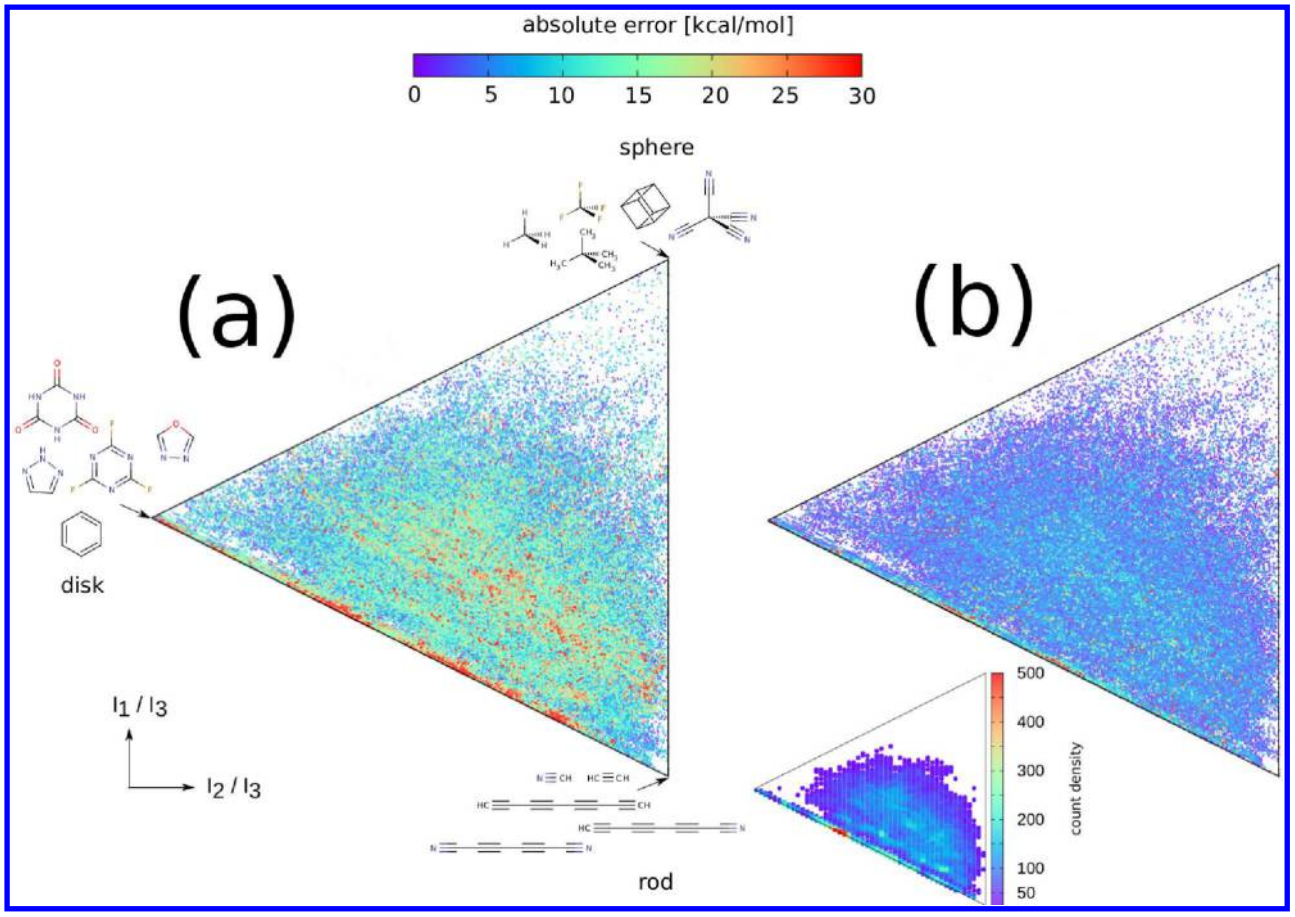


Figure 9. Shape distribution of color-coded absolute deviations between predicted and B3LYP reference atomization enthalpies of the 134k organic molecules with up to nine atoms (not counting hydrogens) in the GDB-17 data set.[35] Vertical and horizontal axes correspond to normalized principal moments of inertia I_1/I_3 and I_2/I_3, respectively, with $I_1 \leq I_2 \leq I_3$. (a) Large PM7 errors (>25 kcal/mol) are predominantly present for geometries on the rod−disk edge, and in the center of the triangle. Corners indicate the geometrical shape of molecules with molecular drawings corresponding to examples of linear ($I_1 = 0$, $I_2 = I_3$), planar ($2I_1 = 2I_2 = I_3$), and spherical ($I_1 = I_2 = I_3$) cases. (b) 10k-Δ_{PM7}^{B3LYP}-ML errors are systematically smaller, some outliers at the rod−disk edge persist. A heatmap of the molecular data density is shown for the same coordinate system in the inset below part b.

set of 6k $C_7H_{10}O_2$-isomers. This data set is a part of this publication, and is available on the authors' homepage.

Computational Details. From ref 35, we obtained all SMILES[44] strings for molecules with up to nine heavy atoms. We then excluded cations, anions, and molecules containing S, Br, Cl, or I, arriving at 133885 molecules. These data are presented and analyzed in more depth in ref 27. Cartesian coordinates for the subset of 6095 isomers of $C_7H_{10}O_2$ were determined by parsing the corresponding SMILES strings using Openbabel software,[45] followed by a consistency check using the CORINA code.[46] Structures of 9868 nonenantiomeric stable diastereomers were obtained through inversions of chiral C atoms in the SMILES strings followed by conversion to Cartesian coordinates using CORINA. To verify that all theoretical methods preserved topology and chirality, we transformed the Cartesian coordinates back to SMILES, and InChI strings using Openbabel. Using these initial structures, we carried out geometry relaxations at the PM7[47] semiempirical level of theory using MOPAC.[48] We used the PM7 equilibrium coordinates as initial geometries and performed DFT and G4MP2 geometry calculations using Gaussian 09.[49] For DFT calculations, we employed the Gaussian basis set 6-31G(2df,p) which is also used in the G4MP2 calculations in combination with the DFT method B3LYP,[50] for geometry relaxation and frequency computations. We used the same basis set also in the GGA-PBE[51] calculations. G4MP2 employs harmonic oscillator and rigid rotor approximations to estimate the entropy of nuclear degrees of freedom.[25] At all levels of theory, we performed harmonic vibrational analysis for all molecules to confirm that the predicted equilibrium structures were local minima on the potential energy surface. HF, MP2, CCSD, CCSD(T) energies have been computed with the basis set 6-31G(d) as a part of G4MP2. Further technical details regarding all quantum chemistry data, including convergence thresholds employed, are given in ref 27.

■ ASSOCIATED CONTENT

Supporting Information

Summary of the technical details of the ML methodology and systematic average errors between various combinations of theoretical levels. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: anatole.vonlilienfeld@unibas.ch.

DOI: 10.1021/acs.jctc.5b00099
J. Chem. Theory Comput. 2015, 11, 2087−2096

Funding
This work was funded by the Swiss National Science Foundation (No. PP00P2_138932).

Notes
The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank J.-L. Reymond for discussions and GDB SMILES, J. Stewart and C. H. Schwab for providing trial licenses for the packages MOPAC and Corina, respectively. A. De Vita, K-R. Müller, A. Tkatchenko, and P. W. Ayers are acknowledged for discussions. Most calculations were performed at sciCORE (http://scicore.unibas.ch/) scientific computing core facility at University of Basel. This research also used resources of the Argonne Leadership Computing Facility at Argonne National Laboratory, which is supported by the Office of Science of the U.S. DOE under Contract DE-AC02-06CH11357.

■ REFERENCES

(1) Kirkpatrick, P.; Ellis, C. Chemical space. *Nature* **2004**, *432*, 823.
(2) Virshup, A. M.; Contreras-García, J.; Wipf, P.; Yang, W.; Beratan, D. N. Stochastic voyages into uncharted chemical space produce a representative library of all possible drug-like compounds. *J. Am. Chem. Soc.* **2013**, *135*, 7296−7303.
(3) von Lilienfeld, O. A. First principles view on chemical compound space: Gaining rigorous atomistic control of molecular properties. *Int. J. Quantum Chem.* **2013**, *113*, 1676−1689.
(4) Ellman, J.; Stoddard, B.; Wells, J. Combinatorial thinking in chemistry and biology. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 2779−2782.
(5) Kolb, H. C.; Finn, M. G.; Sharpless, K. B. Click chemistry: Diverse chemical function from a few good reactions. *Angew. Chem., Int. Ed.* **2001**, *40*, 2004−2021.
(6) Nørskov, J. K.; Bligaard, T.; Rossmeisl, J.; Christensen, C. H. Towards the computational design of solid catalysts. *Nat. Chem.* **2009**, *1*, 37−46.
(7) Jorgensen, W. L. The many roles of computation in drug discovery. *Science* **2004**, *303*, 1813−1818.
(8) Kutchukian, P. S.; Shakhnovich, E. I. *De novo* design: Balancing novelty and confined chemical space. *Expert Opin. Drug Discovery* **2010**, *5*, 789−812.
(9) Hachmann, J.; Olivares-Amaya, R.; Atahan-Evrenk, S.; Amador-Bedolla, C.; Sánchez-Carrera, R. S.; Gold-Parker, A.; Vogt, L.; Brockway, A. M.; Aspuru-Guzik, A. The Harvard clean energy project: Large-scale computational screening and design of organic photovoltaics on the world community grid. *J. Phys. Chem. Lett.* **2011**, *2*, 2241−2251.
(10) Jain, A.; Ong, S. P.; Hautieri, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A. The materials project: A materials genome approach to accelerating materials innovation. *APL Mater.* **2013**, *1*, 011002.
(11) Yang, L.; Ceder, G. Data-mined similarity function between material compositions. *Phys. Rev. B* **2013**, *88*, 224107.
(12) Kuhn, C.; Beratan, D. N. Inverse strategies for molecular design. *J. Phys. Chem.* **1996**, *100*, 10595−10599.
(13) Franceschetti, A.; Zunger, A. The inverse band-structure problem of finding an atomic configuration with given electronic properties. *Nature* **1999**, *402*, 60−63.
(14) von Lilienfeld, O. A.; Lins, R.; Rothlisberger, U. Variational particle number approach for rational compound design. *Phys. Rev. Lett.* **2005**, *95*, 153002.
(15) Wang, M.; Hu, X.; Beratan, D. N.; Yang, W. Designing molecules by optimizing potentials. *J. Am. Chem. Soc.* **2006**, *128*, 3228−3232.
(16) Schneider, G. Virtual screening: An endless staircase? *Nat. Rev.* **2010**, *9*, 273−276.
(17) Sarathy, S. M.; Vranckx, S.; Yasunaga, K.; Mehl, M.; Osswald, P.; Metcalfe, W. K.; Westbrook, C. K.; Pitz, W. J.; Kohse-Höinghaus, K.; Fernandes, R. X.; Curran, H. J. A comprehensive chemical kinetic combustion model for the four butanol isomers. *Combust. Flame* **2012**, *159*, 2028−2055.
(18) Merchant, S. S.; Zanoelo, E. F.; Speth, R. L.; Harper, M. R.; Van Geem, K. M.; Green, W. H. Combustion and pyrolysis of iso-butanol: Experimental and chemical kinetic modeling study. *Combust. Flame* **2013**, *160*, 1907−1929.
(19) Tsang, W.; Hampson, R. F. Chemical kinetic data base for combustion chemistry. Part I. Methane and related compounds. *J. Phys. Chem. Ref. Data* **1986**, *15*, 1087−1279.
(20) Duley, W. W.; Williams, D. A. *Interstellar Chemistry*; Academic Press: London, England, 1984.
(21) Friesner, R. A. Ab initio quantum chemistry: Methodology and applications. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 6648−6653.
(22) Pople, J. A.; Head-Gordon, M.; Fox, D. J.; Raghavachari, K.; Curtiss, L. A. Gaussian-1 theory: A general procedure for prediction of molecular energies. *J. Chem. Phys.* **1989**, *90*, 5622−5629.
(23) Curtiss, L. A.; Raghavachari, K.; Trucks, G. W.; Pople, J. A. Gaussian-2 theory for molecular energies of first- and second-row compounds. *J. Chem. Phys.* **1991**, *94*, 7221−7230.
(24) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. Gaussian-4 theory. *J. Chem. Phys.* **2007**, *126*, 084108.
(25) Curtiss, L. A.; Redfern, P. C.; Raghavachari, K. Gaussian-4 theory using reduced order perturbation theory. *J. Chem. Phys.* **2007**, *127*, 124105.
(26) Szabo, A.; Ostlund, N. S. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*; Dover: Mineola, NY, 1996.
(27) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **2014**, *1*, 140022.
(28) Lomakina, E. I.; Balabin, R. M. Neural network approach to quantum-chemistry data: Accurate prediction of density functional theory energies. *J. Chem. Phys.* **2009**, *131*, 074104.
(29) LiHong, H.; Wang, X.; Wong, L.; Chen, G. Combined first-principles calculation and neural-network correction approach for heat of formation. *J. Chem. Phys.* **2003**, *119*, 11501.
(30) Gillan, M. J.; Alfè, D.; Bartók, A. P.; Csányi, G. First-principles energetics of water clusters and ice: A many-body analysis. *J. Chem. Phys.* **2013**, *139*, 244504.
(31) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*, 2nd ed.; Springer: NY, 2009.
(32) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.
(33) Hansen, K.; Montavon, G.; Biegler, F.; Fazli, S.; Rupp, M.; Scheffler, M.; von Lilienfeld, O. A.; Tkatchenko, A.; Müller, K.-R. Assessment and validation of machine learning methods for predicting molecular atomization energies. *J. Chem. Theory Comput.* **2013**, *9*, 3404−3419.
(34) Montavon, G.; Rupp, M.; Gobre, V.; Vazquez-Mayagoitia, A.; Hansen, K.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Machine learning of molecular electronic properties in chemical compound space. *New J. Phys.* **2013**, *15*, 095003−095020.
(35) Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864−2875.
(36) Cohen, A. J.; Mori-Sánchez, P.; Yang, W. Challenges for density functional theory. *Chem. Rev.* **2012**, *112*, 289−320.
(37) Weber, W.; Thiel, W. Orthogonalization corrections for semiempirical methods. *Theor. Chim. Acta* **2000**, *103*, 495−506.
(38) Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A. Interaction Potentials in Molecules and Non-Local Information in Chemical Space. Submitted for publication.

DOI: 10.1021/acs.jctc.5b00099
J. Chem. Theory Comput. 2015, 11, 2087−2096

(39) Wang, L.-P.; Titov, A.; McGibbon, R.; Liu, F.; Pande, V. S.; Martínez, T. J. Discovering chemistry with an ab initio nanoreactor. *Nat. Chem.* **2014**, *6*, 1044−1048.
(40) Prodan, E.; Kohn, W. Nearsightedness of electronic matter. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 11635−11638.
(41) Koch, W.; Holthausen, M. C. *A Chemist's Guide to Density Functional Theory*; Wiley-VCH: Weinheim, Germany, 2000.
(42) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, Th.; Suhai, S.; Seifert, G. Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties. *Phys. Rev. B* **1998**, *58*, 7260−7268.
(43) Armiento, R.; Mattsson, A. E. Functional designed to include surface effects in self-consistent density functional theory. *Phys. Rev. B* **2005**, *72*, 085108.
(44) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comp. Sci.* **1988**, *28*, 31−36.
(45) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Openbabel: An open chemical toolbox. *J. Cheminf.* **2011**, *3*, 33.
(46) Sadowski, J.; Gasteiger, J. From atoms and bonds to 3-dimensional atomic coordinates—Automatic model builders. *Chem. Rev.* **1993**, *93*, 2567−2581.
(47) Stewart, J. J. P. Optimization of parameters for semiempirical methods. VI: More modifications to the NDDO approximations and re-optimization of parameters. *J. Mol. Modeling* **2013**, *19*, 1.
(48) Stewart, J. J. P. *MOPAC2012*, version 13.136L; Stewart Computational Chemistry: Colorado Springs, CO, 2012.
(49) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A.Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. *Gaussian 09*, revision D.01; Gaussian Inc.: Wallingford, CT, 2009.
(50) Stevens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields. *J. Phys. Chem.* **1994**, *98*, 11623−11627.
(51) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **1996**, *77*, 3865.

DOI: 10.1021/acs.jctc.5b00099
J. Chem. Theory Comput. 2015, 11, 2087−2096