



PCCP

**Statistically Representative Databases for Density  
Functional Theory via Data Science**

Journal:	<i>Physical Chemistry Chemical Physics</i>
Manuscript ID	CP-ART-06-2019-003211.R1
Article Type:	Paper
Date Submitted by the Author:	15-Aug-2019
Complete List of Authors:	Morgante, Pierpaolo; Florida Institute of Technology, Chemistry Program Peverati, Roberto; Florida Institute of Technology, Chemistry Program

SCHOLARONE™  
Manuscripts

## ARTICLE

# Statistically Representative Databases for Density Functional Theory via Data Science

Pierpaolo Morgante,<sup>a</sup> Roberto Peverati<sup>\*a</sup>Received 00th January 20xx,  
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

The number of data and databases for the assessment and parametrization of density functional theory methods has grown substantially in the past two decades. In this work, we introduce a novel cluster analysis technique for density functional theory calculations of the electronic structure of atoms and molecules with the goal of creating new statistically significant databases with broad chemical scope, and a manageable number of data-points. By analyzing without *a priori* chemical assumptions a population of almost 350k data-points, we create a new database called ASCDB containing only 200 data-points. This new database holds the same chemical information as the larger population of data from which it is obtained, but with a computational cost that is reduced by several orders of magnitude. The labelling of the significant chemical properties is performed *a posteriori* on the resulting 16 subsets, classifying them into four areas of chemical importance: non-covalent interactions, thermochemistry, non-local effects, and unbiased calculations. The analysis of the results and their transferability shows that ASCDB is capable of providing the same information as that of the larger collection of data—such as GMTKN55, MGCD84, and Minnesota 2015B—for several density functional theory methods and basis sets. In light of these results, we suggest the use of this new small database as a first inexpensive tool for the evaluation and parametrization of electronic structure theory methods.

## 1. Introduction

Several large databases have been published and extensively used for the evaluation and parametrization of approximated electronic structure theory methods in the past years. Some of the most recent comprehensive efforts in the development of databases with broad chemical scope have originated in the context of density functional theory (DFT), and include the GMTKN55,<sup>1</sup> MGCD84,<sup>2</sup> and Minnesota 2015<sup>3-5</sup> databases. Their importance for the development<sup>6,7</sup> and evaluation<sup>8-10</sup> of new exchange–correlation functionals cannot be overstated, and their continuous expansion is at the basis of the development and assessment of methods with broad applicability in chemistry and physics. Quite recently, we collected all the data in those three major databases as the core of the ACCDB collection,<sup>11</sup> and then we expanded it by adding several smaller databases of under-represented chemical properties. The ACCDB collection is currently composed of 8,656 unique data-points (6,953 of which are from the three core databases) representing several properties of interest throughout chemistry, as well as of an open-source central repository that is easy to access, share, and expand.<sup>12</sup> However, the inclusion of as many properties and as many data as

possible caused three significant drawbacks: a) the extensive, and often unclear, overlap between several data, b) the unbalance in the number of data for several subsets, and c) the overall large computational cost.

The first important drawback of the ACCDB collection—mostly inherited from its core databases—is that the choice of how to divide the data into representative subsets of each property is necessarily made *a priori*, and is based on chemical definitions whose boundaries are inherently fuzzy.<sup>13-15</sup> For example, while it is relatively straightforward to distinguish data that represent non-covalent interactions from those that represent dissociation energies of covalent bonds—they clearly belong to two distinct classes—the subdivision becomes questionable when trying to differentiate between more complex properties, such as barrier heights and systems that are dominated by strong correlation effects (most transition states include stretched bonds whose electronic structures are likely to be dominated by strong correlation effects). Several other not-so-clear cases are undoubtedly present. Korth and Grimme were the first to introduce the idea of “mindless” benchmarking as a new approach to generate new subsets.<sup>16</sup> In doing so, they pointed out that the subdivisions of data-points into different subsets is often biased because of chemical intuition and other factors,<sup>17-19</sup> including what they later called a “developers’ bias”,<sup>1</sup> which in turn brings to a reduction of the dimension of the chemical space spanned by each database. The second drawback of ACCDB and its core databases is that the properties are not well balanced in terms of number of data-points per property (i.e. the subsets have significant data size bias). In fact, 48% of the total data-points in ACCDB are for non-

<sup>a</sup> Chemistry Program, Florida Institute of Technology, 150 W. University Blvd., Melbourne, Florida, 32901, United States; E-mail: rpeverati@fit.edu

†Electronic Supplementary Information (ESI) available: [Details of all xc functionals, basis sets, calculations, and results (SI.pdf); Excel file to calculate MUE and estimated MUE from ASCDB database (ASCDB.xlsx); structures of all input files for ASCDB in xyz format (ASCDB.zip)]. See DOI: 10.1039/x0xx00000x

covalent interactions, while only 8.3% are for bond energies, and less than 1% of points include transition metals. The under-representation of transition metal-containing systems is particularly worrisome, especially because countless applications rely on DFT calculations of metals,<sup>20</sup> and their number is destined to grow even further in the near future. The situation is further complicated by the intrinsic difficulties in obtaining accurate (high-level) reference data across properties: while calculating a large number of high-level references for main-group chemical properties is straightforward (though computationally expensive), obtaining references with similar accuracy for complex cases might be more challenging, or even impossible. The final drawback of our collection is a direct result of the 10,049 single-point calculations currently required to calculate all its data-points: ACCDB is computationally expensive. Based on our own experience, the calculation of all data-points, using a modern quantum chemistry program on a recently acquired machine with a hybrid-GGA functional and a large quadruple- $\zeta$  basis set, requires around  $10^7$  single-core-hours. Hence, it would be beneficial to reduce the number of single-point calculations by a few orders of magnitudes, especially if this can be done without losing statistical significance. Several strategies for the reduction of the number of data-points of a large chemical database have been used in the past, starting with the pioneering work of Truhlar and coworkers,<sup>21–25</sup> who presented several reduced databases obtained from significantly larger ones (e.g. AE6 and BH6,<sup>21</sup> MLBE4/05,<sup>22</sup> BH24,<sup>23,24</sup> V4 and R4<sup>25</sup>). Recently, Chan<sup>26</sup> and Gould<sup>27</sup> used modern statistical analysis techniques to reduce the number of data-points in MGCD84 and GMTKN55, respectively. Despite being statistically solid (see Supporting Information for a more detailed discussion), their databases still inherit the limitations of their respective parent, as they do not contain transition metal systems, and the description of chemical properties is still unbalanced.

To overcome the limitations listed above, a modern collection of data for development and evaluation of electronic structure methods should have the following four characteristics: 1) be representative of various chemical properties, 2) have the data uniquely and impartially divided into subsets that include the largest accessible chemical space, 3) have a balanced number of data for each subset, so that no particular property is either overemphasized or downplayed, and 4) contain a number of data that is statistically significant, but computationally tractable. MGCD84, GMTKN55, and Minnesota 2015 are certainly successful in addressing point 1. Moreover, it is possible to find *ad hoc* weighting schemes to balance the results of large databases to account for point 3. Finally, previous work to reduce the number of data in significant databases appeared in the literature to address point 4.<sup>26,27</sup> However, to the best of our knowledge, there is currently no database that is specifically designed to address all four points simultaneously. Here, we present a new database, obtained using statistical analysis tools borrowed from data science, that will fill this gap. The resulting interpretation of the unbiased statistical results in our new database will also translate into insights on the validity of fuzzy chemical concepts,

and in the definition of new problems that must be addressed in the future.

## 2. Cluster Analysis

In order to obtain a large representative population of chemically relevant data that could be treated with statistical tools, we selected all data-points in the three core databases of ACCDB, and calculated the mean unsigned error (MUE) of 50 DFT exchange–correlation functional approximations (*xc* functionals, or just simply functionals), with basis sets that are close to the complete basis set limit. The resulting population is composed of a total of 347,650 data-points. When putting our ACCDB collection together,<sup>11</sup> we noted that there is overlap between some subsets in the MGCD84, GMTKN55 and Minnesota 2015B databases, for a total of  $\sim 900$  data-points. After a more careful analysis, we saw that 90 data-points coming from Minnesota 2015B are present in either MGCD84 or GMTKN55, while the overlap is more substantial when considering the other two databases. In fact, MGCD84 takes subsets from GMTKN30<sup>28</sup> which is an older version of GMTKN55, for a total of  $\sim 810$  data-points. However, some subsets have updated reference values in the latter. Overall, the data-points that actually overlap are only 563 (8%), and the average difference between the old and the new value is about 0.4 kcal/mol, taken over 337 data-points. We think that it is enough to consider the old and the new data-points as different, at least from the cluster analysis point of view. After we analysed the data (see below) we found out that the statistical tool does not favour neither the new nor the old references. The selection of the *xc* functionals was performed to cover a heterogeneous mix of approximations, including local and non-local functionals, historically significant (B3LYP,<sup>29–31</sup> PBE,<sup>32</sup> BP86<sup>33,34</sup>) and modern functionals (SCAN,<sup>35</sup> MN15,<sup>5</sup>  $\omega$ B97M-V<sup>36</sup>), functionals that are obtained from constraints-satisfaction (PBE, SCAN) and functionals that are obtained via curve-fittings ( $\omega$ B97<sup>36–40</sup> and Minnesota<sup>4,5,41–50</sup> families), as well as different flavors of dispersion-corrected functionals (D3(0),<sup>51</sup> D3(BJ),<sup>52</sup> and VV10<sup>53</sup> corrections). The goal of this selection was to reduce the bias by broadly sampling the functional space. For this reason, we selected functionals from the first four rungs of Perdew's Jacob's ladder of functionals approximations,<sup>54</sup> as well as all the previous three decades of *xc* functionals development. The only notable exclusions are functionals that contain terms that depend on the virtual orbitals (doubly hybrid, or double-hybrid, or fifth-rung functionals). In fact, we were unable to collect data for all the structures in the parent databases because of the inherent limitations (e.g. unclear basis-set convergence,<sup>55,56</sup> erroneous dissociation limits,<sup>57</sup> divergent behaviors of the virtual orbital-dependent terms,<sup>58</sup> uncertainty in the definition of orbitals<sup>59</sup>) of the PT2-like correlation for some of the datapoints. The transition-metal containing systems of the Minnesota database are the most notable example in this sense. If we had to use incomplete results, the cluster analysis would have automatically excluded the doubly-hybrids, therefore creating a bias against them. Instead, in order to avoid this issue, we decided to use them as a tool for

**Table 1.** Summary of the clusters in the ASCDB database with the “*a posteriori*” assignment of chemical properties. The average relative absolute energies  $|\overline{\Delta E}|$ , and the average MUE for the 50 functionals, are also reported in the last two columns (in kcal/mol).

Cluster:	Total Data-Points	Reduced Data-Points	Name of subset:	$ \overline{\Delta E} $	Average MUE
1	2582	15	Non-Covalent: A (NCA15)	17.63	0.43
2	2295	21	Non-Covalent: B (NCB21)	94.78	2.10
3	610	21	Non-Covalent: Cluster (NCC21)	42.45	4.37
4	18	3	Non-Covalent: Water (NCW3)	205.33	26.16
5	374	19	Thermochemistry: Atomization and Reaction Energies (TARE19)	116.83	3.97
6	531	20	Thermochemistry: Barrier Heights A (TBHA20)	165.88	5.37
7	139	16	Thermochemistry: Barrier Heights B (TBHB16)	133.84	9.95
8	22	5	Thermochemistry: Hydrocarbons Reactions (THR5)	2204.76	21.43
9	242	15	Non-Local effects: Mixed A (NLMA15)	1701.33	7.24
10	34	7	Non-Local effects: Mixed B (NLMB7)	348.46	10.49
11	15	11	Non-Local effects: Multi-Reference (NLMR11)	147.24	24.57
12	33	9	Non-Local effects: Self-Interaction-Error (NLSIE9)	123.97	18.92
13	22	8	Unbiased calculations: Mindless Benchmarks A (UMBA8)	445.17	30.92
14	8	5	Unbiased calculations: Mindless Benchmarks B (UMBB5)	352.91	36.38
15	10	7	Unbiased calculations: Mindless Benchmarks C (UMBC7)	726.35	53.06
16	18	18	Unbiased calculations: Atomic Energies per electron (UAE18)	8201.08	1.77
TOTAL:	6953	200		15028.01	

validation and proof of reliability of our database. The complete list of all considered functionals can be found in the Supporting Information.<sup>1-5,19,28-34,36,37,39-53,60-156</sup> The data for the MGCD84 database have been taken from the supporting information of ref. <sup>2</sup>, while some data relative to GMTKN55 have been taken from Prof. Grimme’s website associated with the work in ref. <sup>1</sup>. We have calculated the remainder of the data—including those for the Minnesota database (for which we used the 2015B variant<sup>5</sup>)—with Q-Chem 5.1.<sup>157</sup> We used the same basis sets used in the parent articles, which are mostly of quadruple- $\zeta$  quality, so that results are close to the complete basis set (CBS) limit, but for several molecules in the Minnesota database, a triple- $\zeta$  basis set was used (a detailed description of the basis sets used for each data-point is also reported in the Supporting Information). Stability analysis of all the converged solutions—and re-optimization to the most stable one, when necessary—has been performed. When necessary, symmetry-breaking was also allowed. In all cases (but HF), a Lebedev grid of 99 radial and 590 angular points was used for the integration of the xc functional, while the coarser SG-1 grid<sup>158</sup> was used for the integration of the non-local VV10 contribution for functionals that include it. The entirety of the statistical analysis has been performed with the JMP Pro 14 program.<sup>159</sup> All energies given below are in kcal/mol.

**A. Unbiased cluster selection.** The goal of the cluster analysis tool that we used to treat our population was to divide the data into different groups, or *clusters*, without prior knowledge of the nature of the data-points, and without biases on the number and nature of the clusters themselves. The starting point for the analysis is the extreme situation where each data-point is in its own cluster (347,650 initial clusters), and then we proceeded by locating the two clusters with the smallest Euclidean distance in the hyper-space of all functionals, replacing them with a new cluster of two data-points, characterized by its mean. The procedure is iterated until the number of clusters is sufficiently small. The stopping point

emerges quite clearly from the data analysis at a total of 16 clusters. If we attempt further reduction to a number of clusters smaller than 16, the cost function grows indiscriminately, indicating that the procedure is merging clusters that are not independent of each other.

**B. Reduction of the data-points.** Since each point in every cluster has a high similarity with the other points in the same cluster, the reduction of the total number of data-points should be simple. Following similar routes as those used by Chan and Gould, we explored the use of both least absolute shrinkage and selection operator (LASSO),<sup>160</sup> and stepwise regression<sup>161</sup> methods for the overall reduction of the data within each cluster, and we ultimately used the method that allowed us to maintain a coefficient of determination  $R^2 \geq 0.99$  for each cluster. The relationship between the reduced data-points cluster and the original one is linear, and a simple regression formula recovers the full cluster performance. Overall, we can calculate the estimated MUE (eMUE) of each individual parent database using a set of linear regression formulas containing 57 coefficients. The 200 data-points in the reduced database represent less than 3% of the total original data-points, but the regression formulas retain a striking statistical significance, with values of  $R^2 \geq 0.92$ . The 16 resulting clusters are summarized in Table 1, and all our regression formulas and coefficients are reported in the Supporting Information.

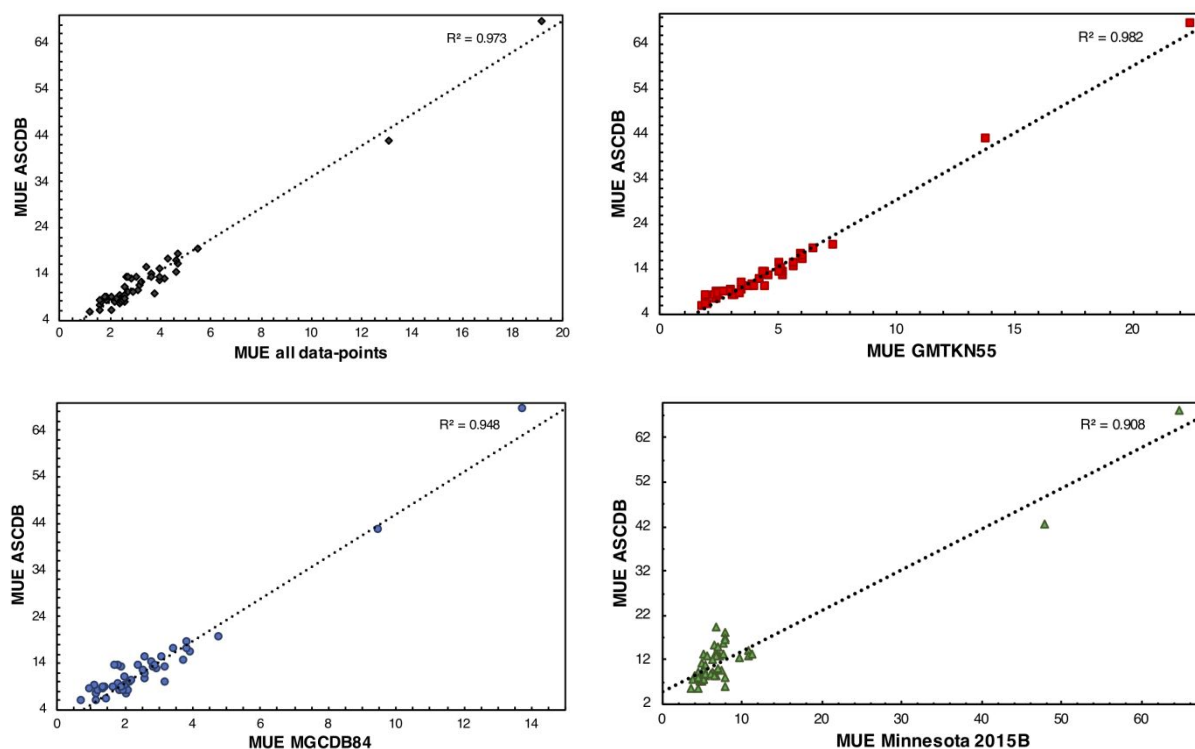
### 3. A Smaller Chemistry Database of Broad Purpose

In the previous section, we outlined the procedure that we used to generate a statistically significant database that is substantially smaller than its parent collection, as well as the regression formulas that can be used to relate its errors with the MUEs of the parent databases. We stress again here that the subdivisions and reductions have been generated only by

statistical analysis, without any bias or chemical assumption. We now analyze the data in each cluster more closely, to check if we can find some *a posteriori* chemical significance that justifies the subdivision performed by the statistical tool. In fact, to our surprise, this can be done in a fairly straightforward manner, illustrated here using cluster 1. The first cluster is composed of 2,582 data-points coming from the following subsets of the parent core databases: A21x12, BzDC215, RG10 and S66x8 from MGCD84, Amino20x4, BUT14DIOL, HAL59, MCONF, and SCONF from GMTKN55, and NCCE30, NGDWI21, and a small number from SR-MGN-BE107 from Minnesota 2015B (note: for detailed information on these subsets and full explanation of their names we refer the reader to the original publications,<sup>28,42,95,152,162-175</sup> to ref. 1-5, or alternatively to our ACCDB paper<sup>11</sup>). In fact, the vast majority of data in this subset is easily attributable to the class of non-covalent interactions, with only a small number of exceptions. The presence of such apparent outliers can be explained by the fact that some data that are not intrinsically attributable to non-covalent interactions do indeed behave like a non-covalent interaction on a *xc* functional standpoint. In other words, functionals that provide a good description of systems dominated by non-covalent interactions, do also provide good descriptions for such cases. This should not come as a surprise: in fact, during the optimization of a *xc* functional it is common to notice that a new set of parameters might improve the performance not only for property *A*, but also for molecules dominated by property *B*. At the same time, it might also systematically worsen some other property *Z*, even when no data-points for property *B*, or

*Z*, are included in the optimization function. Properties that seem chemically disconnected might be correlated on a density standpoint.

A similar *a posteriori* analysis can be repeated for the other 15 clusters, with the following results: three more clusters (cluster 2, 3, and 4) contain mostly non-covalent interactions data; four clusters (clusters 5–8) contain mostly thermochemistry data; four other clusters (clusters 9–12) contain mostly systems that are dominated by strong-correlation effects (which we decided to call non-local effects, *vide infra*); three clusters (cluster 13–15) prevalently contain data from the “mindless benchmark” subset; and the last cluster (cluster 16) contains all the atomic energies (which we decided to report on a per-electron base, to avoid biasing the error towards the heavier elements). We used labels such as “A”, “B”, and “C” when it is difficult to differentiate between clusters within the same group, while we used more descriptive labels for clearer cases. In the non-covalent group, for example, we have the “non-covalent A” and “non-covalent B” subsets with mixed cases (dimers and trimers), as well as the “non-covalent clusters” subset including cluster of molecules (where the term cluster is now used in the chemical acceptance of the term, borrowing the definition from Mardirossian and Head-Gordon’s work<sup>2</sup>), and the “non-covalent water” subset that contains only data for clusters of water molecules. The subsets in the “thermochemistry” group include data for “atomization and reaction energies” (cluster 5), “barrier heights A” and “barrier heights B” (cluster 6 and 7), and “alkanes reactions” (cluster 8). The next group mainly deals with systems



**Fig. 1.** Correlation plots between the MUE calculated on ASCDB and the total MUE calculated using all the data-points from the three major databases (top-left panel, black diamonds), the MUE calculated with GMTKN55 (top-right panel, red squares), MGCD84 (bottom-left panel, blue circles), Minnesota 2015B (bottom-right, green triangles). Units on the axis are kcal/mol.

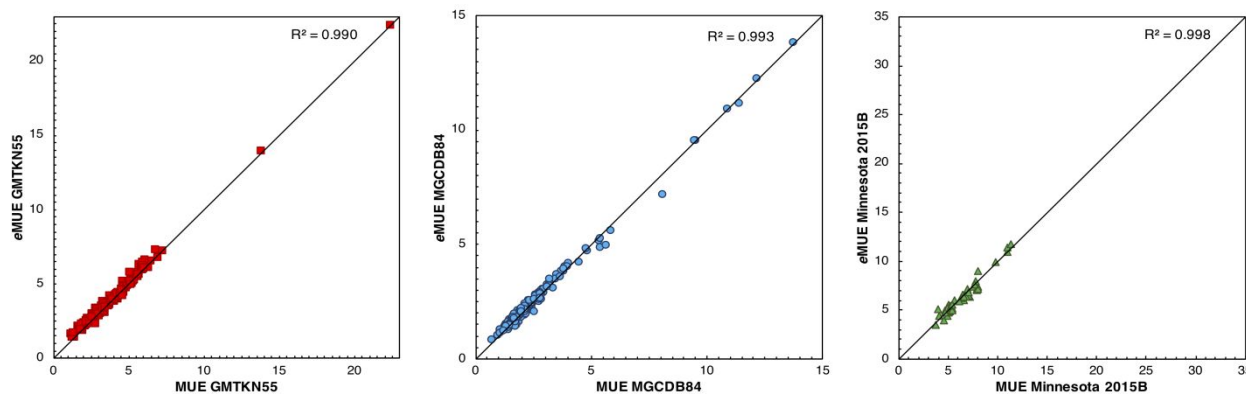


Fig. 2. Correlation plots between the estimated MUE (eMUE) calculated through regression from the data-points in ASCDB and the MUE calculated with GMTKN55 (left panel, red squares), MGCDB84 (middle panel, blue circles), and Minnesota 2015B (right panel, green triangles). Units on the axis are kcal/mol.

that suffer from errors in the description of non-local effects in the exchange, *e.g.* cases with significant “self-interaction error” (SIE), or in the correlation, *e.g.* cases with significant “multi-reference” (MR) character, or both. For clusters 9 and 10, it is not clear (at least to us) what is the dominating correlation effect, and therefore we labeled them “mixed A” and “mixed B”. We assigned the label MR to cluster 11 because it includes species that are not properly described by a single-determinant treatment, as exemplified by a small HOMO-LUMO gap. In general, *xc* functionals that contains a high percentage of non-local exact (“HF-like”) exchange fail badly for such systems. We assigned the label SIE to cluster 12 to denote systems that require a non-local exchange functional for their accurate description. For these systems *xc* functionals with a high percentage of exact exchange solve most issues providing reasonably satisfactory results. Lastly, the “unbiased calculations” group is composed of data from the last four clusters (13 to 16). Since clusters 13, 14, and 15 mainly collect data from the MB16-43 subset of Grimme, Goerigk and co-workers,<sup>1</sup> which includes 43 reactions of randomly-generated molecules of 16 atoms each, we decided to keep the original name in these cases. Because of their random nature, it is indeed difficult to classify these molecules using a chemical rationale,<sup>16</sup> and it is very interesting to notice that our cluster analysis put most of these systems within three separate and independent clusters. This also suggests that these molecules might include interactions that sample regions of the exact *xc* functionals that are usually not covered by parameterization based on conventional chemical systems, and only highly transferrable *xc* functional approximations will provide accurate descriptions for them. In fact, it is highly unlikely that error cancellations can play a role on these systems, as also exemplified by the highest average MUE of the sampled 50 functionals among all of our clusters. Hence, we chose the “unbiased calculation” label for this group, borrowing the definition, once again, from Goerigk et al.<sup>1</sup> The atomic energies cluster was also included in the “unbiased calculations” group because cancellation of errors is impossible by definition for these systems as well.

By analyzing these final assignments, we noticed that the 200 data-points, divided into 4 subsets and 16 clusters (4 per

subset), are a reasonably small collection of a diverse database on their own. We name our new reduced database “ASCDB: A Smaller Chemistry DataBase” (as compared to the much larger ACCDB collection). ASCDB fulfils all the requirements for a useful small database for chemistry. In fact, it covers a vast chemical space, and its data-points are well divided among the four main properties that we introduced above, with each property being well represented (the “non-covalent” group has 60 data-points, the “thermochemistry” has 60 as well, while the “non-local effects” and the “unbiased calculations” have 42 and 38 data-points each, respectively). Additionally, the cluster analysis selected a heterogeneous spectrum of molecular sizes including atoms, small molecules, and large systems of up to 80 atoms. The majority (about 94%) of systems have a number of atoms that varies from 1 to 30, extensively covering the average molecular size studied in routine calculations, but molecules with more than 30 atoms are also well represented, being about 6% of the database. As a comparison, the percentage of molecules with more than 30 atoms is 11% in GMTKN55, about 7% for MGCDB84, and less than 1% for the Minnesota 2015B database. Before recommending ASCDB for general use in chemistry, we show some of its most important results in the next section. In addition, we also test its robustness by using

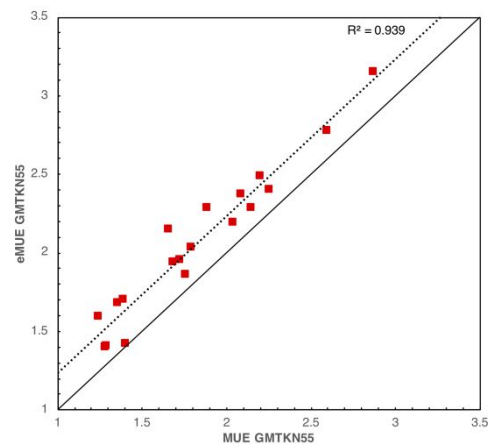
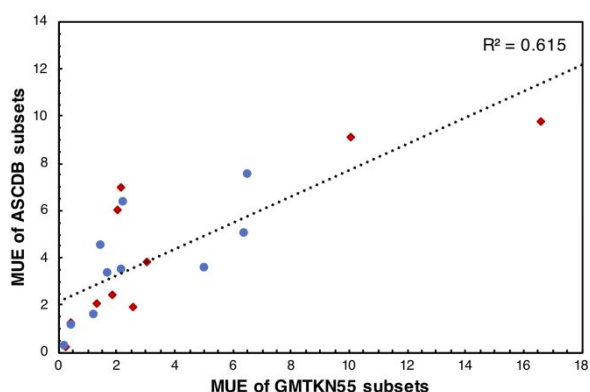


Fig. 3. Correlation plots between the estimated MUE (eMUE) calculated through regression from the data-points in ASCDB and the MUE calculated with GMTKN55 for doubly hybrid functionals. Units on the axis are kcal/mol.



**Fig. 4.** Correlation between the MUEs of subsets of ASCDB and the corresponding ones of GMTKN55 with two doubly hybrid functionals: B2-PLYP-D3(BJ) (red, diamonds) and DSD-PBEP86-D3(BJ) (blue, circles). Details on the corresponding pairs of subsets are reported in the supporting information. Units on the axis are kcal/mol.

methodologies that are outside those that we used to create it, such as different basis sets, different *xc* functionals, and also different statistical indicators.

## 4. Results and Validation

**A. Results.** The first step to assess the effectiveness of ASCDB as a benchmarking tool is to quantify the amount of information that was lost in going from the large parent databases to the reduced one. This is achieved by correlating the mean unsigned errors (MUEs) calculated for the 200 data-points in ASCDB with the total MUE of all 6,953 data-points, as well as with those of the individual parent databases. Figure 1 shows the remarkable correlation obtained from such data. In Figure 2, we also report the correlation plots between the MUE of each parent database and the estimated-MUE (*e*MUEs) obtained with the regression formulas. The latter plots provide a straightforward visualization of the strength of the data-science technique, because points align well on the diagonal. In all cases, the correlation is excellent, with an  $R^2$  always greater than 0.900, and the worst correlation coefficient (for the GMTKN55 database) being 0.990. The obvious advantage of ASCDB is that it provides information on all three parent databases

**Table 2.** Estimated RMSEs from ASCDB compared to the calculated RMSEs (in parenthesis) for the main subsets of MGCD84<sup>a</sup> for three *xc* functionals and three basis sets that are different than the one used for the data analysis. All data are in kcal/mol.

Functional:	Basis set:	NC	ISO	TC	BH
PBE-D3(BJ)	6-31G*	7.38 (8.35)	1.43 (2.77)	14.33 (14.92)	8.26 (15.88)
	def2-SVPD	1.63 (2.9)	1.16 (1.97)	13.72 (15.23)	10.35 (11.28)
	def2-TZVPPD	1.23 (1.27)	1.57 (1.48)	11.61 (11.42)	10.14 (10.09)
B3LYP-D3(BJ)	6-31G*	6.77 (7.47)	1.58 (11.96)	6.37 (9.45)	5.02 (10.96)
	def2-SVPD	2.41 (2.46)	1.3 (1.83)	6.72 (8.49)	7.78 (6.29)
	def2-TZVPPD	0.77 (0.75)	1.46 (1.84)	4.32 (4.32)	6.75 (5.71)
$\omega$ B97M-V	6-31G*	10.55 (6.82)	2.74 (2.07)	12.83 (9.63)	2.8 (8.84)
	def2-SVPD	2.15 (2.14)	1.09 (0.87)	9.31 (9.65)	4.09 (3.21)
	def2-TZVPPD	0.31 (0.3)	0.57 (0.61)	2.79 (3.04)	3.56 (1.78)

<sup>a</sup>NC: non-covalent interactions (most data from MGCD84); ISO: isomerization energies (most data from GMTKN55); TC: thermochemistry (equally divided); BH: barrier heights (most data from Minnesota 2015B).

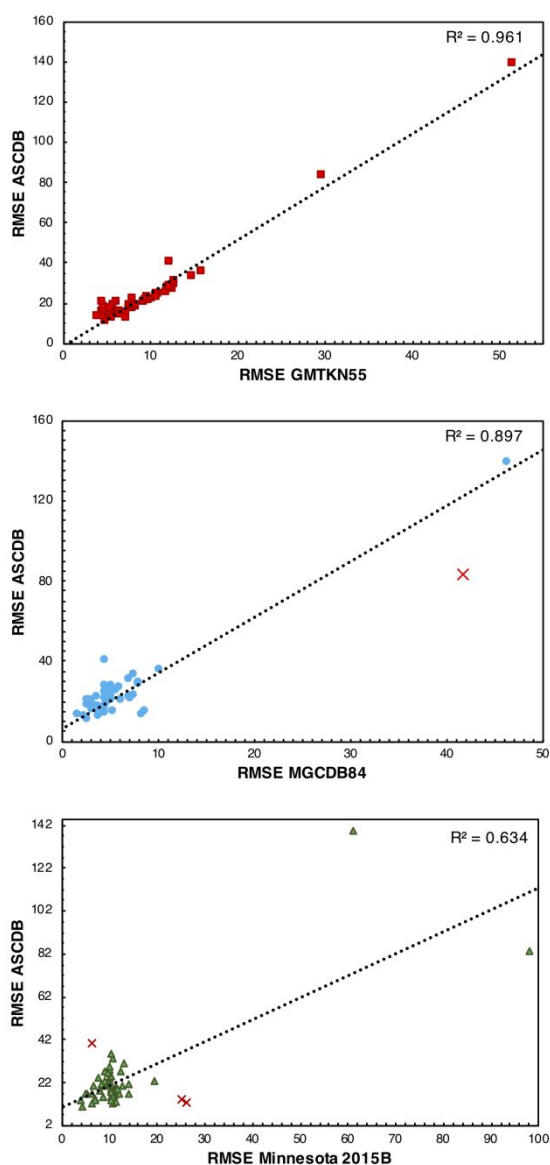
**Table 3.** Comparison of errors between significant chemical properties of ASCDB and GMTKN55 using doubly hybrid functionals. All data are MUE in kcal/mol.

Chemical Property:	B2-PLYP-D3(BJ)	DSD-PBEP86-D3(BJ)
Non-Covalent (ASCDB)	1.52	1.28
Total non-covalent (GMTKN55)	0.34	0.36
Thermochemistry (ASCDB)	3.44	3.01
Basic Properties (GMTKN55)	2.17	2.00
Nonlocality Error Dominated (ASCDB)	5.22	5.77
W4-11MR (GMTKN55)	3.04	6.55
Unbiased Calculations (ASCDB)	9.79	5.05
Mindless benchmark (GMTKN55)	16.62	6.46
TOTAL MUE (ASCDB)	4.60	3.51
TOTAL MUE (GMTKN55)	1.79	1.29

simultaneously, at a fraction of their computational cost. For example, only Minnesota 2015B contains transition metals, but by running ASCDB, we can obtain the same information on transition metals that is in the Minnesota database, as well as all the chemical information contained in MGCD84 and GMTKN55.

**B. Validation.** In order to validate our results and to prove the robustness of ASCDB as a standalone tool for the evaluation and development of new electronic structure methods, we explore the results of several methods that were not used to create the database. Our first validation is with respect to the choice of basis set. To generate our data, we used the largest basis set that we could afford (mostly quadruple- $\zeta$ , see the Supporting Information for details), to be as close as possible to the basis set limit. It is important to notice though that some unwanted error might still have been present in the resulting data. Possible sources of such errors are, for example, the triple- $\zeta$  basis set that we used for some large or problematic case, the residual incompleteness error that could have been present even at the quadruple- $\zeta$  level,<sup>176</sup> or the inability of some functionals to reach the basis set limit itself.<sup>177</sup> To validate the stability of the data analysis with respect to basis set size, we selected three basis set that are different than those used to generate the original data and compared the MUEs on ASCDB and the parent databases using three sample functionals. The three basis sets are 6-31G\*,<sup>178,179</sup> def2-SVPD,<sup>60</sup> and def2-TZVPPD,<sup>60</sup> while the three functionals are PBE-D3(BJ),<sup>32</sup> B3LYP-D3(BJ),<sup>29-31,52</sup> and  $\omega$ B97M-V.<sup>36</sup> Results on four chemical categories of the parent databases—non-covalent interactions (NC), isomerization energies (ISO), thermochemistry (TC), and barrier heights (BH)—are reported in Table 2, with detailed correlation plots and data for the other parent databases reported in the supporting information. The trends in MUEs of the parent databases are all very well reflected from the ASCDB results, with the exception of a very limited numbers of outliers (only one outlier is clear in table 2 for the  $\omega$ B97M-V/def2-TZVPPD results). In general, we can observe a trend in the stability of the ASCDB results with respect to changes in the basis set that is similar to the one observed in the parent databases.

A second validation of the robustness of ASCDB comes from expanding the functional space to include doubly hybrid functionals. We originally excluded fifth-rung functionals from



**Fig. 5.** Correlation plots between the root mean squared error (RMSE) calculated on ASCDB and the RMSE calculated with GMTKN55 (top panel, red squares), MGCD84 (middle panel, blue circles), and Minnesota 2015B (bottom panel, green triangles). Some points have been excluded from the calculation of the correlation coefficients and are reported with a red cross. Units on the axis are kcal/mol.

our analysis because it was not possible to obtain reliable PT2 correlation energies for several small-gap (multi-reference) systems in the Minnesota database (see the Supporting Information for details on the problematic systems). However, we can use doubly hybrid functionals as a good validation tool by comparing our results to GMTKN55 only, since it does not include any transition metal. To do so, we first extracted the doubly hybrid data of ref. <sup>1</sup> (21 functionals, including different flavors of dispersion corrections; all data are reported in Table S7 of the Supporting Information) and we calculated the correlation plot between the MUE from GMTKN55 and the eMUE from ASCDB. Results are in Figure 3. The plots show very good correlation ( $R^2 = 0.94$ ), and a slight systematic overestimation of about 0.2 kcal/mol. To expand the doubly hybrid results even further, we calculated the MUEs of all molecules in ASCDB using two of the most popular functionals: B2-PLYP-D3(BJ)<sup>135,138</sup> and DSD-PBEP86-D3(BJ)<sup>142</sup>. The MUEs obtained from these additional calculations are calculated without using the regression formulas, and are collected in Table 3, and in Figure 4. A direct correlation can be observed in Table 2 between each of the four chemical areas of ASCDB (non-covalent interactions, thermochemistry, non-local effects, unbiased calculations), and a corresponding area of GMTKN55 (total non-covalent, basic properties, multi-reference cases from W4-11, and mindless benchmarks). Moreover, good correlation between the individual subsets of ASCDB and GMTKN55 can be observed in the plot of Figure 4 (the corresponding numerical data with details on each corresponding pairs of subsets are reported in the Supporting Information). The calculations using doubly hybrid functionals reported above validate the usefulness of the information that can be obtained from ASCDB for methods that are outside the scope of the initial data analysis.



## ARTICLE

As an additional validation, we expanded the statistical analysis beyond the mean unsigned error. We report in figure 5 the correlation of the root mean square error (RMSE) of ASCDB with those of the parent databases. These plots show—once again—reasonably good correlation, with the only exception of a couple of outliers that can be easily identified and excluded from the trend. Another useful statistical indicator for the evaluation of the results is the mean signed error (MSE). Specifically, the MSE is convenient to evaluate the systematic under or overestimation trends of methods for predicting some specific property: a method that systematically underestimates the reference data will present MSE and MUE with very similar magnitudes for one specific subset, but opposite signs. On the contrary, methods that systematically overestimate the reference data will have MUE and MSE that are comparable in both magnitude and sign. Finally, for cases with no systematic errors, the MSE will be close to zero, regardless of the magnitude of the MUE. The striking majority of cases where a systematic error happens involves three properties: non-covalent interactions (NC), non-local effects (NL), and unbiased calculations (UC). In Figure 6 we report a comparison between the results obtained with ASCDB and those obtained from the parent functionals for NC and NL, using 10 methods that present systematic errors for either case. The corresponding plot for UC does not provide useful information, since all data comes from the mindless benchmark subset of GMTKN55, and the MSEs are trivially overlapping. We clearly notice that all functionals that underestimate a property using the parent database will also present a negative sign of the corresponding subsets of ASCDB, with magnitudes of the errors that are also very comparable. Similarly, all functionals that are systematically overestimating a property will have a consistently positive MSE for the corresponding ASCDB subset, with magnitudes that are also equally well correlated.

Overall, our validation data show a very encouraging correspondence between the results obtained with ASCDB and those obtained from the parent databases, supporting the transferability of the results obtained with the new database to basis sets, methods, and statistical indicators that were not included in the cluster analysis that generated it.

#### 4. Conclusions

As a first attempt to introduce modern data science tools into chemistry, we used standard cluster analysis techniques to create a new database that is: a) representative (i.e. with broad chemical purpose); b) unbiased (i.e. with subdivisions into subsets of properties that are performed *a posteriori*, using impartial methodologies); c) well-balanced (i.e. with a comparable number of data for each property); and d) computationally affordable (i.e. with a small number of data).

The database is called ASCDB (A Smaller Chemistry DataBase) and is an unbiased, statistically representative subset of three large core chemistry databases in our ACCDB collection. ASCDB contains 200 data-points (from 350 unique single-point energy calculations) that are impartially divided into 16 subsets

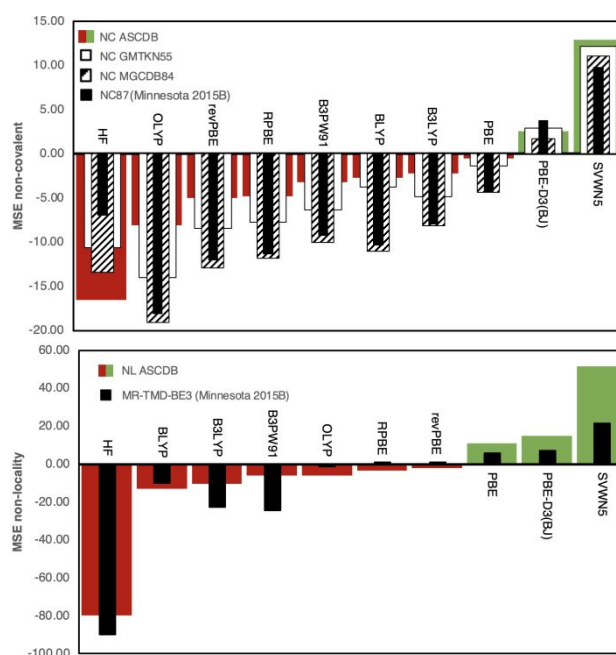


Fig. 6. Comparison of the mean signed error (MSE) for 10 methods that present a systematic error for non-covalent interactions (NC, top panel) and non-local effects (NL, bottom panel), using ASCDB and selected subsets of the parent functionals. Units on the axis are kcal/mol.

representative of four different chemical properties. It does not suffer from size-bias, since every subset contains a similar amount of data. It can be also used to estimate the mean unsigned errors for the larger core databases—with a remarkable  $R^2$  accuracy—via a set of provided linear regression formulas and coefficients.

The advantage of our new database is that by performing a reasonable number of single-point energy calculations, one can obtain two complementary types of information for the evaluation or parametrization of approximated electronic structure theory methods: 1) a statistical estimate of the performance of the method on much larger databases, via regression formulas—representative for GMTKN55, MGCB84, and Minnesota 2015B—as well as, 2) a new unbiased and well-balanced tool, via its subsets. The reliability of this information has been validated on several methods that are outside the scope of the data analysis that was used to generate the database itself. In light of these results, we recommend the use of ASCDB as the first tool for evaluation and parametrization of electronic structure theory methods. Larger collection, such as GMTKN55, MGCB84, and Minnesota 2015, can then be used for further validation and transferability studies, when more granular data are needed.

#### Conflicts of interest

There are no conflicts to declare.

## Acknowledgments

Some of the calculations have been performed on the Florida Tech Blueshark cluster, which is supported by the National Science Foundation under Grant No. CNS 09-23050. The authors are also thankful to Prof. Nasri Nesnas and Kelly Hambel for proofreading the manuscript.

## References

- 1 L. Goerigk, A. Hansen, C. Bauer, S. Ehrlich, A. Najibi and S. Grimme, *Phys. Chem. Chem. Phys.*, 2017, **19**, 32184–32215.
- 2 N. Mardirossian and M. Head-Gordon, *Mol. Phys.*, 2017, **115**, 2315–2372.
- 3 H. S. Yu, W. Zhang, P. Verma, X. He and D. G. Truhlar, *Phys. Chem. Chem. Phys.*, 2015, **17**, 12146–12160.
- 4 H. S. Yu, X. He and D. G. Truhlar, *J. Chem. Theory Comput.*, 2016, **12**, 1280–1293.
- 5 H. S. Yu, X. He, S. L. Li and D. G. Truhlar, *Chem. Sci.*, 2016, **7**, 5032–5051.
- 6 N. Mardirossian and M. Head-Gordon, *J. Chem. Phys.*, 2018, **148**, 241736-1–241736-14.
- 7 G. Santra, N. Sylvetsky and J. M. L. Martin, *J. Phys. Chem. A*, 2019, **123**, 5129–5143.
- 8 N. Mehta, M. Casanova-Páez and L. Goerigk, *Phys. Chem. Chem. Phys.*, 2018, **20**, 23175–23194.
- 9 A. Najibi and L. Goerigk, *J. Chem. Theory Comput.*, 2018, **14**, 5725–5738.
- 10 M. Wang, D. John, J. Yu, E. Proynov, F. Liu, B. G. Janesko and J. Kong, *J. Chem. Phys.*, 2019, **150**, 204101-1–204101-8.
- 11 P. Morgante and R. Peverati, *J. Comput. Chem.*, 2019, **40**, 839–848.
- 12 Available at: <https://github.com/peverati/ACCDB>, accessed Aug. 2019.
- 13 F. Akeroyd, *HYLE - Int. J. Phil. Chem.*, 2000, **6**, 161–173.
- 14 J. F. Gonthier, S. N. Steinmann, M. D. Wodrich and C. Corminboeuf, *Chem. Soc. Rev.*, 2012, **41**, 4671–4687.
- 15 J. Grunenberg, *Int. J. Quantum Chem.*, 2017, **117**, e25359–11.
- 16 M. Korth and S. Grimme, *J. Chem. Theory Comput.*, 2009, **5**, 993–1003.
- 17 S. Grimme, M. Steinmetz and M. Korth, *J. Org. Chem.*, 2007, **72**, 2118–2126.
- 18 G. B. I. Csonka, A. Ruzsinszky, J. Tao and J. P. Perdew, *Int. J. Quantum Chem.*, 2005, **101**, 506–511.
- 19 S. Grimme, *J. Phys. Chem. A*, 2005, **109**, 3067–3077.
- 20 C. J. Cramer and D. G. Truhlar, *Phys. Chem. Chem. Phys.*, 2009, **11**, 10757–10816.
- 21 B. J. Lynch and D. G. Truhlar, *J. Phys. Chem. A*, 2003, **107**, 8996–8999.
- 22 N. E. Schultz, Y. Zhao and D. G. Truhlar, *J. Phys. Chem. A*, 2005, **109**, 11127–11143.
- 23 J. Zheng, Y. Zhao and D. G. Truhlar, *J. Chem. Theory Comput.*, 2007, **3**, 569–582.
- 24 J. Zheng, Y. Zhao and D. G. Truhlar, *J. Chem. Theory Comput.*, 2009, **5**, 808–821.
- 25 K. Yang, R. Peverati, D. G. Truhlar and R. Valero, *J. Chem. Phys.*, 2011, **135**, 044118-1–044118-22.
- 26 B. Chan, *J. Chem. Theory Comput.*, 2018, **14**, 4254–4262.
- 27 T. Gould, *Phys. Chem. Chem. Phys.*, 2018, **20**, 27735–27739.
- 28 L. Goerigk and S. Grimme, *J. Chem. Theory Comput.*, 2011, **7**, 291–309.
- 29 A. D. Becke, *J. Chem. Phys.*, 1993, **98**, 5648–5652.
- 30 P. Stephens, F. Devlin, C. Chabalowski and M. J. Frisch, *J. Phys. Chem.*, 1994, **98**, 11623–11627.
- 31 C. Lee, W. Yang and R. G. Parr, *Phys. Rev. B*, 1988, **37**, 785–789.
- 32 J. P. Perdew, K. Burke and M. Ernzerhof, *Phys. Rev. Lett.*, 1996, **77**, 3865–3868.
- 33 A. D. Becke, *Phys. Rev. A*, 1988, **38**, 3098–3100.
- 34 J. P. Perdew, *Phys. Rev. B*, 1986, **33**, 8822–8824.
- 35 J. Sun, A. Ruzsinszky and J. P. Perdew, *Phys. Rev. Lett.*, 2015, **115**, 64.
- 36 N. Mardirossian and M. Head-Gordon, *J. Chem. Phys.*, 2016, **144**, 214110-1–214110-23.
- 37 N. Mardirossian and M. Head-Gordon, *Phys. Chem. Chem. Phys.*, 2014, **16**, 9904–9924.
- 38 N. Mardirossian and M. Head-Gordon, *J. Chem. Phys.*, 2014, **140**, 18A527-1–18A527-15.
- 39 N. Mardirossian and M. Head-Gordon, *J. Chem. Phys.*, 2015, **142**, 074111-1–074111-32.
- 40 N. Mardirossian, L. R. Pestana, J. C. Womack, C.-K. Skylaris, T. Head-Gordon and M. Head-Gordon, *J. Phys. Chem. Lett.*, 2016, **8**, 35–40.
- 41 Y. Zhao, N. E. Schultz and D. G. Truhlar, *J. Chem. Phys.*, 2005, **123**, 161103.
- 42 Y. Zhao, N. E. Schultz and D. G. Truhlar, *J. Chem. Theory Comput.*, 2006, **2**, 364–382.
- 43 Y. Zhao and D. G. Truhlar, *J. Chem. Phys.*, 2006, **125**, 194101-1–194101-18.
- 44 Y. Zhao and D. G. Truhlar, *J. Phys. Chem. A*, 2006, **110**, 13126–13130.
- 45 Y. Zhao and D. G. Truhlar, *Theor. Chem. Acc.*, 2008, **120**, 215–241.
- 46 Y. Zhao and D. G. Truhlar, *J. Chem. Theory Comput.*, 2008, **4**, 1849–1868.
- 47 R. Peverati and D. G. Truhlar, *J. Phys. Chem. Lett.*, 2011, **2**, 2810–2817.
- 48 R. Peverati and D. G. Truhlar, *J. Phys. Chem. Lett.*, 2012, **3**, 117–124.
- 49 R. Peverati and D. G. Truhlar, *Phys. Chem. Chem. Phys.*, 2012, **14**, 13171–13174.
- 50 R. Peverati and D. G. Truhlar, *Phys. Chem. Chem. Phys.*, 2012, **14**, 16187–16191.
- 51 S. Grimme, J. Antony, S. Ehrlich and H. Krieg, *J. Chem. Phys.*, 2010, **132**, 154104-1–154104-19.
- 52 S. Grimme, S. Ehrlich and L. Goerigk, *J. Comput. Chem.*, 2011, **32**, 1456–1465.
- 53 O. A. Vydrov and T. Van Voorhis, *J. Chem. Phys.*, 2010, **133**, 244103.
- 54 J. P. Perdew and K. Schmidt, *AIP Conf. Proc.*, 2001, **577**, 1–20.
- 55 A. Karton and J. M. L. Martin, *J. Chem. Phys.*, 2011, **135**, 144119-1–144119-8.
- 56 J. C. Sancho-Garcia and C. Adamo, *Phys. Chem. Chem. Phys.*, 2013, **15**, 14581.
- 57 D. Hait and M. Head-Gordon, *J. Chem. Phys.*, 2018, **148**, 171102-1–171102-4.
- 58 J.-D. Chai and M. Head-Gordon, *J. Chem. Phys.*, 2009, **131**, 174105-1–174105-13.


- 59 R. Peverati and M. Head-Gordon, *J. Chem. Phys.*, 2013, **139**, 024110-1–024110-6.
- 60 F. Weigend and R. Ahlrichs, *Phys. Chem. Chem. Phys.*, 2005, **7**, 3297–3305.
- 61 B. J. Lynch, Y. Zhao and D. G. Truhlar, *J. Phys. Chem. A*, 2003, **107**, 1384–1388.
- 62 P. L. Fast, M. Sanchez and D. G. Truhlar, *Chem. Phys. Lett.*, 1999, **306**, 407–410.
- 63 L. Curtiss, P. Redfern, K. Raghavachari, V. A. Rassolov and J. A. Pople, *J. Chem. Phys.*, 1999, **110**, 4703–4709.
- 64 A. K. Wilson, D. E. Woon, K. A. Peterson and T. H. Dunning Jr., *J. Chem. Phys.*, 1999, **110**, 7667–7676.
- 65 N. B. Balabanov and K. A. Peterson, *J. Chem. Phys.*, 2005, **123**, 064107-1–064107-15.
- 66 N. B. Balabanov and K. A. Peterson, *J. Chem. Phys.*, 2006, **125**, 074110-1–074110-10.
- 67 R. A. Kendall, T. H. Dunning and R. Harrison, *J. Chem. Phys.*, 1992, **96**, 6796–6806.
- 68 D. E. Woon and T. H. Dunning Jr., *J. Chem. Phys.*, 1993, **98**, 1358–1371.
- 69 K. A. Peterson, D. Figgen, M. Dolg and H. Stoll, *J. Chem. Phys.*, 2007, **126**, 124101-1–124101-12.
- 70 J. Zheng, X. Xu and D. G. Truhlar, *Theor. Chem. Acc.*, 2010, **128**, 295–305.
- 71 B. P. Prascher, D. E. Woon, K. A. Peterson, T. H. Dunning and A. K. Wilson, *Theor. Chem. Acc.*, 2011, **128**, 69–82.
- 72 J. Koput and K. A. Peterson, *J. Phys. Chem. A*, 2002, **106**, 9595–9599.
- 73 D. E. Woon and T. H. Dunning Jr., *J. Chem. Phys.*, 1995, **103**, 4572–4585.
- 74 R. Krishnan, J. Binkley and R. Seeger, *J. Chem. Phys.*, 1980, **72**, 650.
- 75 J. C. Slater, *Phys. Rev.*, 1951, **81**, 385–390.
- 76 S. H. Vosko, L. Wilk and M. Nusair, *Can. J. Phys.*, 1980, **58**, 1200–1211.
- 77 J. P. Perdew, J. A. Chevary, S. H. Vosko, K. A. Jackson, M. R. Pederson, D. J. Singh and C. Fiolhais, *Phys. Rev. B*, 1992, **46**, 6671–6687.
- 78 S. Grimme, *J. Comput. Chem.*, 2006, **27**, 1787–1799.
- 79 Y. Zhang and W. Yang, *Phys. Rev. Lett.*, 1997, **80**, 890–890.
- 80 B. Hammer, L. Hansen and J. Norskov, *Phys. Rev. B*, 1999, **59**, 7413–7421.
- 81 A. D. Boese and N. C. Handy, *J. Chem. Phys.*, 2000, **114**, 5497–5503.
- 82 N. C. Handy and A. J. Cohen, *Mol. Phys.*, 2001, **99**, 403–412.
- 83 R. Peverati and D. G. Truhlar, *J. Chem. Theory Comput.*, 2012, **8**, 2310–2319.
- 84 J. Sun, A. Ruzsinszky and J. P. Perdew, *Phys. Rev. Lett.*, 2015, **115**, 036402-1–036402-6.
- 85 R. Sabatini, T. Gorni and S. de Gironcoli, *Phys. Rev. B*, 2013, **87**, 041108-1–041108-4.
- 86 A. D. Boese and N. C. Handy, *J. Chem. Phys.*, 2002, **116**, 9559–9569.
- 87 J. Tao, J. P. Perdew, V. N. Staroverov and G. E. Scuseria, *Phys. Rev. Lett.*, 2003, **91**, 146401-1–146401-4.
- 88 J. P. Perdew, A. Ruzsinszky, G. I. Csonka, L. A. Constantin and J. Sun, *Phys. Rev. Lett.*, 2009, **103**, 026403-1–026403-4.
- 89 D. R. Hartree, *Math. Proc. Cambridge Phil. Soc.*, 1928, **24**, 89–132.
- 90 V. Fock, *Z. Physik*, 1930, **61**, 126–148.
- 91 J. P. Perdew and Y. Wang, *Phys. Rev. B*, 1992, **45**, 13244–13249.
- 92 C. Adamo and V. Barone, *J. Chem. Phys.*, 1999, **110**, 6158–6170.
- 93 M. Ernzerhof and G. E. Scuseria, *J. Chem. Phys.*, 1999, **110**, 5029–5036.
- 94 P. Wilson, T. Bradley and D. J. Tozer, *J. Chem. Phys.*, 2001, **115**, 9233–9242.
- 95 R. Peverati and D. G. Truhlar, *J. Chem. Phys.*, 2011, **135**, 191102-1–191102-4.
- 96 F. Hamprecht, A. J. Cohen, D. J. Tozer and N. C. Handy, *J. Chem. Phys.*, 1998, **109**, 6264–6271.
- 97 A. D. Boese and J. M. L. Martin, *J. Chem. Phys.*, 2004, **121**, 3405–3416.
- 98 Y. Zhao and D. G. Truhlar, *J. Phys. Chem. A*, 2005, **109**, 5656–5667.
- 99 V. N. Staroverov, G. E. Scuseria, J. Tao and J. P. Perdew, *J. Chem. Phys.*, 2003, **119**, 12129–12137.
- 100 A. V. Krukau, O. A. Vydrov, A. F. Izmaylov and G. E. Scuseria, *J. Chem. Phys.*, 2006, **125**, 224106-1–224106-5.
- 101 T. M. Henderson, B. G. Janesko and G. E. Scuseria, *J. Chem. Phys.*, 2008, **128**, 194105-1–194105-9.
- 102 H. Schröder, A. Creon and T. Schwabe, *J. Chem. Theory Comput.*, 2015, **11**, 3163–3170.
- 103 D. G. A. Smith, L. A. Burns, K. Patkowski and C. D. Sherrill, *J. Phys. Chem. Lett.*, 2016, **7**, 2197–2203.
- 104 W. Hujo and S. Grimme, *J. Chem. Theory Comput.*, 2011, **7**, 3866–3871.
- 105 A. Karton, D. Gruzman and J. M. L. Martin, *J. Phys. Chem. A*, 2009, **113**, 8434–8447.
- 106 T. W. Keal and D. J. Tozer, *J. Chem. Phys.*, 2005, **123**, 121103-1–121103-4.
- 107 L. A. Burns, A. V. Mayagoitia, B. G. Sumpter and C. D. Sherrill, *J. Chem. Phys.*, 2011, **134**, 084107-1–084107-25.
- 108 A. D. Becke, *J. Chem. Phys.*, 1997, **107**, 8554–8560.
- 109 L. A. Constantin, E. Fabiano and F. Della Sala, *J. Chem. Theory Comput.*, 2013, **9**, 2256–2263.
- 110 T. Tsuneda, T. Suzumura and K. Hirao, *J. Chem. Phys.*, 1999, **110**, 10664–10678.
- 111 T. Yanai, D. Tew and N. C. Handy, *Chem. Phys. Lett.*, 2004, **393**, 51–57.
- 112 A. D. Boese, N. Doltsinis, N. C. Handy and M. Sprick, *J. Chem. Phys.*, 1999, **112**, 1670–1678.
- 113 E. Weintraub, T. M. Henderson and G. E. Scuseria, *J. Chem. Theory Comput.*, 2009, **5**, 754–762.
- 114 M. A. Rohrdanz, K. M. Martins and J. M. Herbert, *J. Chem. Phys.*, 2009, **130**, 054112-1–054112-9.
- 115 L. Goerigk, *J. Phys. Chem. Lett.*, 2015, **6**, 3891–3896.
- 116 J. P. Perdew, A. Ruzsinszky, G. I. Csonka, O. A. Vydrov, G. E. Scuseria, L. A. Constantin, X. Zhou and K. Burke, *Phys. Rev. Lett.*, 2008, **100**, 136406-1–136406-4.
- 117 J. Wellendorff, K. T. Lundgaard, K. W. Jacobsen and T. Bligaard, *J. Chem. Phys.*, 2014, **140**, 144107-1–144107-11.
- 118 J. Sun, R. Haunschild, B. Xiao, I. W. Bulik, G. E. Scuseria and J. P. Perdew, *J. Chem. Phys.*, 2013, **138**, 044113-1–044113-8.
- 119 J. Sun, J. P. Perdew and A. Ruzsinszky, *Proc. Natl. Acad. Sci. USA*, 2015, **112**, 685–689.
- 120 C. Adamo and V. Barone, *J. Chem. Phys.*, 1997, **108**, 664–675.
- 121 S. Grimme, J. G. Brandenburg, C. Bannwarth and A. Hansen, *J. Chem. Phys.*, 2015, **143**, 054107-1–054107-20.
- 122 J. P. Perdew, S. Kurth, A. Zupan and P. Blaha, *Phys. Rev. Lett.*, 1999, **82**, 2544–2547.

- 123 G. I. Csonka, J. P. Perdew and A. Ruzsinszky, *J. Chem. Theory Comput.*, 2010, **6**, 3688–3703.
- 124 E. D. Murray, K. Lee and D. C. Langreth, *J. Chem. Theory Comput.*, 2009, **5**, 2754–2762.
- 125 K. Hui and J.-D. Chai, *J. Chem. Phys.*, 2016, **144**, 044114-1–044114-8.
- 126 J. G. Brandenburg, J. E. Bates, J. Sun and J. P. Perdew, *Phys. Rev. B*, 2016, **94**, 115144-1–115144-11.
- 127 H. Peng, Z.-H. Yang, J. Sun and J. P. Perdew, *Phys. Rev. X*, 2016, **6**, 041005-1–041005-15.
- 128 Y. Zhao and D. G. Truhlar, *J. Chem. Phys.*, 2008, **128**, 184109-1–184109-8.
- 129 J. Tao and Y. Mo, *Phys. Rev. Lett.*, 2016, **117**, 073001-1–073001-6.
- 130 J.-D. Chai and M. Head-Gordon, *Phys. Chem. Chem. Phys.*, 2008, **10**, 6615–6620.
- 131 Y.-S. Lin, G.-D. Li, S.-P. Mao and J.-D. Chai, *J. Chem. Theory Comput.*, 2012, **9**, 263–272.
- 132 Y.-S. Lin, C.-W. Tsai, G.-D. Li and J.-D. Chai, *J. Chem. Phys.*, 2012, **136**, 154109-1–154109-12.
- 133 A. Austin, G. A. Petersson, M. J. Frisch, F. J. Dobek, G. Scalmani and K. Throssell, *J. Chem. Theory Comput.*, 2012, **8**, 4989–5007.
- 134 A. D. Becke, *J. Chem. Phys.*, 1996, **104**, 1040–1046.
- 135 L. Goerigk and S. Grimme, 2011, **13**, 6670–6688.
- 136 B. Miehlisch, A. Savin, H. Stoll and H. Preuss, *Chem. Phys. Lett.*, 1988, **157**, 200–206.
- 137 A. Karton, A. Tarnopolsky, J.-F. Lamere, G. C. Schatz and J. M. L. Martin, *J. Phys. Chem. A*, 2008, **112**, 12868–12886.
- 138 S. Grimme, *J. Chem. Phys.*, 2006, **124**, 034108-1–034108-16.
- 139 H. Schmider and A. D. Becke, *J. Chem. Phys.*, 1998, **109**, 8188–8199.
- 140 A. D. Becke, *J. Chem. Phys.*, 1993, **98**, 1372–1377.
- 141 S. Kozuch, D. Gruzman and J. M. L. Martin, *J. Phys. Chem. C*, 2010, **114**, 20801–20808.
- 142 S. Kozuch and J. M. L. Martin, *Phys. Chem. Chem. Phys.*, 2011, **13**, 20104–20107.
- 143 T. M. Henderson, A. F. Izmaylov, G. E. Scuseria and A. Savin, *J. Chem. Theory Comput.*, 2008, **4**, 1254–1262.
- 144 J. Heyd, G. E. Scuseria and M. Ernzerhof, *J. Chem. Phys.*, 2003, **118**, 8207–8215.
- 145 J. Moellmann and S. Grimme, *J. Phys. Chem. C*, 2014, **118**, 7615–7621.
- 146 T. M. Henderson, A. F. Izmaylov, G. Scalmani and G. E. Scuseria, *J. Chem. Phys.*, 2009, **131**, 044108-1–044108-9.
- 147 Y. Zhao and D. G. Truhlar, *J. Phys. Chem. A*, 2004, **108**, 6908–6918.
- 148 Y. Zhao, N. González-García and D. G. Truhlar, *J. Phys. Chem. A*, 2005, **109**, 2012–2018.
- 149 T. Schwabe and S. Grimme, *Phys. Chem. Chem. Phys.*, 2006, **8**, 4398–4401.
- 150 J. P. Perdew, *Electronic Structure of Solids 1991*, Akademie Verlag, Berlin, 1991.
- 151 W. M. Hoe, A. J. Cohen and N. C. Handy, *Chem. Phys. Lett.*, 2001, **341**, 319–328.
- 152 Y. Zhao and D. G. Truhlar, *J. Chem. Theory Comput.*, 2005, **1**, 415–432.
- 153 M. Ernzerhof and J. P. Perdew, *J. Chem. Phys.*, 1998, **109**, 3313–3320.
- 154 L. Goerigk and S. Grimme, *J. Chem. Theory Comput.*, 2010, **6**, 107–126.
- 155 Y. Zhao, B. J. Lynch and D. G. Truhlar, *Phys. Chem. Chem. Phys.*, 2005, **7**, 43–52.
- 156 X. Xu and W. A. Goddard, *Proc. Natl. Acad. Sci. USA*, 2004, **101**, 2673–2677.
- 157 Y. Shao, Z. Gan, E. Epifanovsky, A. T. B. Gilbert, M. Wormit, J. Kussmann, A. W. Lange, A. Behn, J. Deng, X. Feng, D. Ghosh, M. Goldey, P. R. Horn, L. D. Jacobson, I. Kaliman, R. Z. Khaliullin, T. Kus, A. Landau, J. Liu, E. Proynov, Y. M. Rhee, R. M. Richard, M. A. Rohrdanz, R. P. Steele, E. J. Sundstrom, H. L. Woodcock, P. M. Zimmerman, D. Zuev, B. Albrecht, E. Alguire, B. Austin, G. J. O. Beran, Y. A. Bernard, E. Berquist, K. Brandhorst, K. B. Bravaya, S. T. Brown, D. Casanova, C. M. Chang, Y. Chen, S. H. Chien, K. D. Closser, D. L. Crittenden, M. Diedenhofen, R. J. DiStasio, H. Do, A. D. Dutoi, R. G. Edgar, S. Fatehi, L. Fusti-Molnar, A. Ghysels, A. Golubeva-Zadorozhnaya, J. Gomes, M. W. D. Hanson-Heine, P. H. P. Harbach, A. W. Hauser, E. G. Hohenstein, Z. C. Holden, T. C. Jagau, H. Ji, B. Kaduk, K. Khistyayev, J. Kim, J. Kim, R. A. King, P. Klunzinger, D. Kosenkov, T. Kowalczyk, C. M. Krauter, K. U. Lao, A. Laurent, K. V. Lawler, S. V. Levchenko, C. Y. Lin, F. Liu, E. Livshits, R. C. Lochan, A. Luenser, P. Manohar, S. F. Manzer, S. P. Mao, N. Mardirossian, A. V. Marenich, S. A. Maurer, N. J. Mayhall, E. Neuscamman, C. M. Oana, R. Olivares-Amaya, D. P. O'Neill, J. A. Parkhill, T. M. Perrine, R. Peverati, A. Prociuk, D. R. Rehn, E. Rosta, N. J. Russ, S. M. Sharada, S. Sharma, D. W. Small, A. Sodt, T. Stein, D. Stuck, Y. C. Su, A. J. W. Thom, T. Tsuchimochi, V. Vanovschi, L. Vogt, O. Vydrov, T. Wang, M. A. Watson, J. Wenzel, A. White, C. F. Williams, J. Yang, S. Yeganeh, S. R. Yost, Z. Q. You, I. Y. Zhang, X. Zhang, Y. Zhao, B. R. Brooks, G. K. L. Chan, D. M. Chipman, C. J. Cramer, W. A. Goddard, M. S. Gordon, W. J. Hehre, A. Klamt, H. F. Schaefer, M. W. Schmidt, C. D. Sherrill, D. G. Truhlar, A. Warshel, X. Xu, A. Aspuru-Guzik, R. Baer, A. T. Bell, N. A. Besley, J.-D. Chai, A. Dreuw, B. D. Dunietz, T. R. Furlani, S. R. Gwaltney, C. P. Hsu, Y. Jung, J. Kong, D. S. Lambrecht, W. Liang, C. Ochsenfeld, V. A. Rassolov, L. V. Slipchenko, J. E. Subotnik, T. Van Voorhis, J. M. Herbert, A. I. Krylov, P. M. W. Gill and M. Head-Gordon, *Mol. Phys.*, 2015, **113**, 184–215.
- 158 P. M. W. Gill, B. G. Johnson and J. A. Pople, *Chem. Phys. Lett.*, 1993, **209**, 506–512.
- 159 B. Jones and J. Sall, *WIREs Comp Stat*, 2011, **3**, 188–194.
- 160 R. Tibshirani, *J. R. Statist. Soc. B*, 1996, **58**, 267–288.
- 161 M. A. Efronson, *Multiple regression analysis, Mathematical Methods for Digital Computers*, Wiley, New York, 1960.
- 162 J. Witte, M. Goldey, J. B. Neaton and M. Head-Gordon, *J. Chem. Theory Comput.*, 2015, **11**, 1481–1492.
- 163 D. L. Crittenden, *J. Phys. Chem. A*, 2009, **113**, 1663–1669.
- 164 K. T. Tang and J. P. Toennies, *J. Chem. Phys.*, 2003, **118**, 4976–4983.
- 165 J. Řezáč, K. E. Riley and P. Hobza, *J. Chem. Theory Comput.*, 2011, **7**, 2427–2438.
- 166 M. K. Kesharwani, A. Karton and J. M. L. Martin, *J. Chem. Theory Comput.*, 2015, **12**, 444–454.
- 167 S. Kozuch, S. M. Bachrach and J. M. L. Martin, *J. Phys. Chem. A*, 2014, **118**, 293–303.
- 168 S. Kozuch and J. M. L. Martin, *J. Chem. Theory Comput.*, 2013, **9**, 1918–1931.
- 169 U. R. Fogueri, S. Kozuch, A. Karton and J. M. L. Martin, *J. Phys. Chem. A*, 2013, **117**, 2269–2277.

## ARTICLE


## Physical Chemistry Chemical Physics

- 170 G. I. Csonka, A. D. French, G. P. Johnson and C. A. Stortz, *J. Chem. Theory Comput.*, 2009, **5**, 679–692.
- 171 M. S. Marshall, L. A. Burns and C. D. Sherrill, *J. Chem. Phys.*, 2011, **135**, 194102-1–1941092-11.
- 172 R. Peverati and D. G. Truhlar, *Phil Trans Roy Soc A*, 2014, **372**, 20120476.
- 173 O. A. Vydrov and T. van Voorhis, *J. Chem. Theory Comput.*, 2012, **8**, 1929–1934.
- 174 K. M. de Lange and J. R. Lane, *J. Chem. Phys.*, 2011, **134**, 034301-1–034301-10.
- 175 J. D. McMahon and J. R. Lane, *J. Chem. Phys.*, 2011, **135**, 154309-1–154309-10.
- 176 D. Hait and M. Head-Gordon, *Phys. Chem. Chem. Phys.*, 2018, **20**, 19800–19810.
- 177 N. Mardirossian and M. Head-Gordon, *J. Chem. Theory Comput.*, 2013, **9**, 4453–4461.
- 178 R. Ditchfield, *J. Chem. Phys.*, 1971, **54**, 724–6.
- 179 W. J. Hehre, R. Ditchfield and J. A. Pople, *J. Chem. Phys.*, 1972, **56**, 2257–2261.



**ASCDB**

Cluster	Chemical description of properties	Cluster ID	Number of molecules	Number of atoms	Number of bonds
Cluster 1	Hydrocarbons	AL010	6,000	60	100
Cluster 2	Non-Covalent H	AL021	2,000	20	30
Cluster 3	Non-Covalent Oxygen	AL031	400	4	6
Cluster 4	Non-Covalent Water	AL041	10	1	2
Cluster 5	Thermochemistry: Equilibrium and Reaction Energies	TR051	200	2	3
Cluster 6	Thermochemistry: Bond Lengths &	TR061	100	10	15
Cluster 7	Thermochemistry: Bond Angles &	TR071	100	10	15
Cluster 8	Thermochemistry: High-Pressure Reaction	TR081	20	2	3
Cluster 9	Non-Covalent Hydrogen	AL091	100	10	15
Cluster 10	Non-Covalent Hydrogen	AL101	10	1	2
Cluster 11	Non-Covalent Hydrogen	AL111	10	1	2
Cluster 12	Non-Covalent Hydrogen	AL121	10	1	2
Cluster 13	Non-Covalent Hydrogen	AL131	10	1	2
Cluster 14	Non-Covalent Hydrogen	AL141	10	1	2
Cluster 15	Non-Covalent Hydrogen	AL151	10	1	2
Cluster 16	Non-Covalent Hydrogen	AL161	10	1	2
Cluster 17	Non-Covalent Hydrogen	AL171	10	1	2
Cluster 18	Non-Covalent Hydrogen	AL181	10	1	2
Cluster 19	Non-Covalent Hydrogen	AL191	10	1	2
Cluster 20	Non-Covalent Hydrogen	AL201	10	1	2



Cluster analysis applied to quantum chemistry: A new broad database of chemical properties with a reasonable computational cost.