Cite this: DOI: 10.1039/xxxxxxxxxx

Poisoning density functional theory with benchmark sets of difficult systems†

Tim Gould and Stephen J. Dale[a]

Large benchmark sets like GMTKN55 [Goerigk *et al., Phys. Chem. Chem. Phys.*, 2017, 19, 32184] let us analyse the performance of density functional theory over a diverse range of systems and bonding types. However, assessing over a large and diverse set can miss cases where approaches fail badly, and can give a misleading sense of security. To this end we introduce a series of 'poison' benchmark sets, P30-5, P30-10 and P30-20, comprising systems with up to 5, 10 and 20 atoms, respectively. These sets represent the most difficult-to-model systems in GMTKN55. We expect them to be useful in developing new approximations, identifying weak points in existing ones, and to aid in selecting appropriate DFAs for computational studies involving difficult physics, e.g. catalysis.

It is becoming rare to find a physical chemistry paper that does not contain results from a density functional theory (DFT) calculation. DFT provides quantum chemical insights at low cost, by using density functional approximations (DFAs) to capture difficult quantum mechanics. [note, many authors erroneously use DFT to refer to DFAs] Popularity has brought with it diversity – there is now a 'zoo' of hundreds, if not thousands of DFAs to choose from.[1–4] The zoo continues to grow, as density functional developers produce new DFAs to solve outstanding problems and to cook new DFAs using new ingredients.

Selecting the best DFA for a given study has thus become an onerous task, which has given rise to benchmarking studies that seek to make the task less onerous. Benchmarking helps to identify useful DFAs by scrutinizing their performance on relevant reactions. In broad terms, the process of benchmarking first requires computing,

$$\mathrm{Err}(d,I) = |\Delta E_I^d - \Delta E_I^{\mathrm{ref}}| \tag{1}$$

for DFA d on 'reaction' I, which might be a traditional reaction or another important energy difference like an ionisation potential. ΔE_I^d is the energy difference calculated using the DFA. $\Delta E_I^{\mathrm{ref}}$ is a reference value for the reaction computed using a high-level theory or obtained from appropriately modified experimental data. Their absolute difference is thus a reasonable metric for the error of the given DFA on the given reaction.

The "general main group thermochemistry, kinetics, and noncovalent interactions" (GMTKN55) benchmark database[5], consisting of ~ 1500 reactions and ~ 2500 individual energies, goes one step further than traditional benchmarking approaches. It condenses performance on a large benchmark set into a single number (WTMAD-2) for a given DFA, which is designed to assess a variety of predictions with different energy scales. GMTKN55 may thus be used to assess performance of existing DFAs, and also aid in optimising and scrutinising new approaches. Other sets, like MGCDB84[6] and MB16-43,[7] serve similar roles.

The WTMAD-2 scheme (see Sec 4 of GMTKN55[5] for details) defines an overall quality metric,

$$\mathrm{WTMAD}(d) := \frac{1}{N_{\mathrm{GMTKN55}}} \sum_I W_I \mathrm{Err}(d,I) \tag{2}$$

using eq. (1) and weights, W_I, that depend only on the benchmark set $B(I)$ containing I. These weights normalise the deviations to ensure that sets with large energies, yet small relative errors, do not dominate over sets with small energies, yet relatively large and important errors. Thus, WTMAD-2 can be used to assess the effectiveness of a DFA, in practice.

However, a problem with any ranking protocol defined on a very large number of systems, such as WTMAD-2, is that the single number can mask systematic deficiencies of DFAs. For example, a DFA that very accurately predicts 1450 reactions can fail rather badly on the remaining 50 outliers without much statistical impact. One of the authors[8] previously showed that WTMAD-2 statistics could be reproduced imperfectly by just 50 carefully chosen reactions, and nearly perfectly by 150 reactions, using 'Diet-GMTKN55' benchmark sets. The ability to reduce the benchmark set size without impacting statistics might suggest that outliers are unlikely to be disastrous, since they are likely to have a greater impact on the smaller set. A more troubling possibility is that outliers simply did not make it into the 'diet' sets due to their low statistical importance.

When assessing DFAs, it would be useful to know how well they work for difficult cases, to test their overall robustness and reliability. The present work thus seeks to answer the following questions:

[a] *Qld Micro- and Nanotechnology Centre, Griffith University, Nathan, Qld 4111, Australia; E-mail: t.gould@griffith.edu.au*

1. What reactions in GMTKN55 are most difficult for DFA to reproduce?

2. How well does performance on WTMAD-2 predict performance on outliers?

3. Can we better understand density functional approximations by using outliers?

Answering these questions will be the subject of the rest of this manuscript.

What reactions in GMTKN55 are most difficult for DFA to reproduce?

Answering the first question requires identifying what makes a reaction difficult. Any given DFA will struggle with some reactions – statistics basically guarantee it. A difficult reaction is therefore one which is *difficult for a large number of DFAs to reproduce* and which therefore poisons the statistics for a large number of DFAs. One therefore seeks systems that are outliers across many DFAs.

Identifying these reactions involves turning the usual benchmarking problem on its head. Eq. (2) assigns each given DFA, d, a (weighted) average of its performance across all reactions, I, which can be used to rank it – lower values indicated better quality. By contrast, the expression,

$$\text{Poison}(I) = \frac{1}{N_{\text{DFA}}} \sum_{d \in \text{DFA}} \text{Err}(d,I) \,, \tag{3}$$

yields higher value for any reaction that is difficult for all N_{DFA} DFAs to reproduce, and which thus poisons the most DFAs. The reactions may then be sorted from worst (largest Poison(I) value) to best (smallest Poison(I) value).

Selecting reactions with only the largest values of Poison(I) therefore defines a new benchmark set that is, by design, difficult for DFAs to reproduce. The smaller this set, the more difficult it will be for DFAs to reproduce the systems (by contrast, the entire set simply reproduces GMTKN55). However, smaller sets expose a potential deficiency in eq. (3) – it is based on absolute energies only and so may be biased to large systems with more bonds, since each bond can introduce error when computed using a DFA.

Maximal discriminatory power is therefore obtained by selecting a sufficiently small set of poison species, and ensuring it is not biased toward large systems. **Poison30-N_{atom} (P30-N)** benchmark sets achieve this task, by using the thirty most difficult systems, as ranked by eq. (3), whose reactions contain at most N_{atom} atoms. Thirty reactions (2% of the total GMTKN55 set) is chosen as a manageable number, with enough reactions for variety but few enough reactions to contain only outliers. The issue of bonds is avoided by using only reactions with up to a given number of atoms in its largest molecule. The present work settles on **P30-5**, **P30-10** and **P30-20**, with five, ten and 20 atoms respectively. [One may also weight eq. (3) by W_I before sorting, but defining a useful weight is not a trivial task.]

Figure 1 shows the molecules (but not atoms) used in **P30-5**. The other two sets are shown as Supplementary Figure 1 in the Supporting Information. Reactions in all poison sets come from just eight of the 55 sets in GMTKN55: (1) ALK8[5]; (2) ALKBDE10[9]; (3) BH76[10–12]; (4) DC13[12–23]; (5) G2RC[12,24];

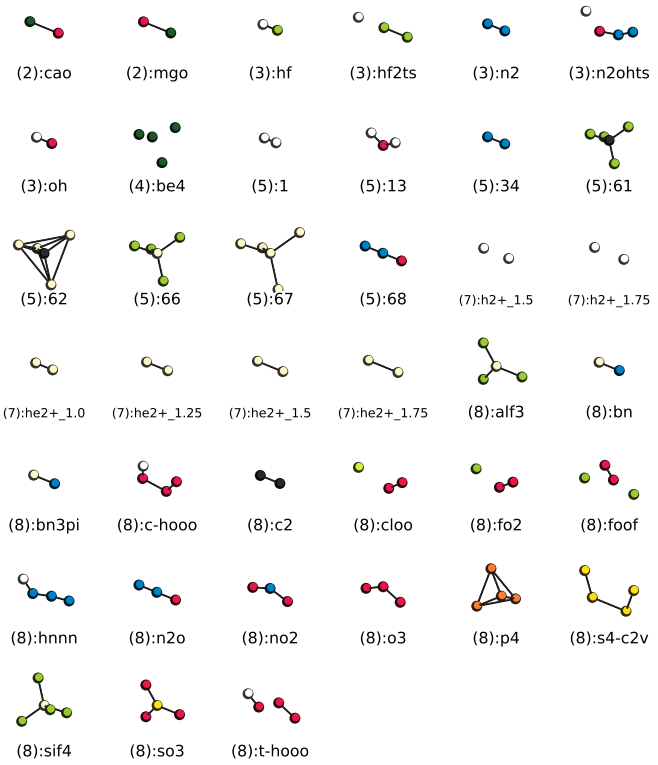

Fig. 1 The 39 molecules of up to five atoms (atoms not shown) used to form the 30 'poison' reactions in the **P30-5** benchmark set.

(6) MB16-43[7]; (7) SIE4x4[5]; and (8) W4-11[25], with MB16-43, SIE4x4 and W4-11 making up the majority of the P30 sets in which they feature. Text descriptors in Figure 1 and its supporting counterparts show the original database, using these labels, and descriptions of each molecule from each original database.

The reactions identified for each P30 set can be rationalised by a combination of chemical and density-functional theorist intuition. The requirements to be selected in a poison set is to possess the highest error below a certain system size. The highest errors will therefore come from cases where (a) DFT performs particularly poorly, and/or (b) bond energies (and by extension errors) are particularly high, preferably both.

The three benchmark sets that feature prominently in the P30 sets match this intuition. Self-interaction error[26–34] is a well known DFA error which is often tested using dissociation of homodiatomic charged molecules. The SIE4x4 set is specifically designed to test for self-interaction error, so clearly satisfies (a), with charge imbalances common in these systems driving up energy differences satisfying (b) also. W4-11 is a set of total atomisation energies which satisfy requirement (b). Taking a closer look at some of the specific systems reveals a preference for homoneuclear molecules or multiple double bonds, systems known to exacerbate static/strong correlation[35–38] which is poorly described by DFAs and thus satisfy (a). MB16-43 is a set of randomly generated 'artificial molecules' and in not following the "narrow structural space of chemical intuition"[7] so it is not hard to image satisfying requirements (a) and (b). The remaining contributors to the P30

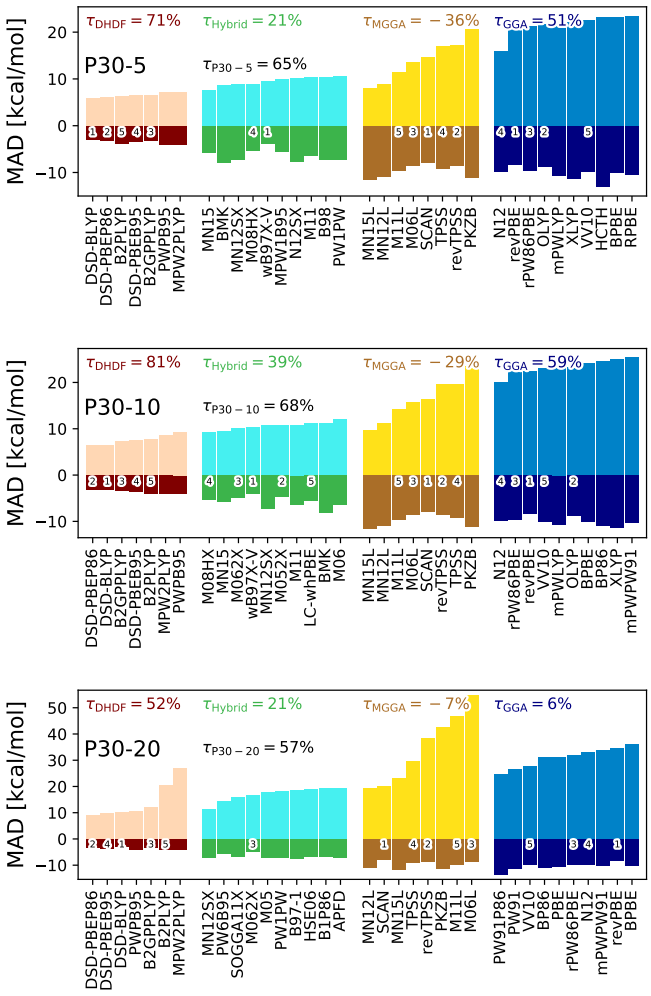

Fig. 2 Errors found in **P30-5** (top), **P30-10** (middle) and **P30-20** (bottom). Bars above zero indicate the MAD across the 'poison' sets, while bars below zero indicate WTMAD-2. Results are divided into four rungs: double-hybrid DFAs (reds), hybrid DFAs (greens), meta-GGAs (earth tones) and GGAs (blues). Numbers below zero indicate top-ranked approximations on WTMAD-2. Only the ten highest ranked DFAs are shown. Kendall rank correlation coefficients, τ, are also reported for the four DFA rungs, as well as across all DFA.

sets from other benchmark sets also fit into the arguments given here.

How well does performance on WTMAD-2 predict performance on outliers?

With the P30-N sets established, the second question may now be addressed. First, the unweighted MAD,

$$\mathrm{MAD}_{\mathrm{P30}\text{-}N}(d) = \frac{1}{30}\sum_{I\in \mathrm{P30}\text{-}N} \mathrm{Err}(d,I) \tag{4}$$

gives an error metic for each DFA, d, on poison set P30-N. Then, the error of each DFA on P30-N may be compared against the WTMAD-2 value for the same DFA using eq. (2).

Following GMTKN55, DFAs are divided into different rungs of "Jacob's ladder".[39] The lowest rung is GGAs, which are functionals of the density, n and its gradient, $|\nabla n|$, only. The next rung is meta-GGAs (MGGAs) which introduce the kinetic energy density, τ, and Laplacians of the density, $\nabla^2 n$. The next rung is hybrids, which introduce some Hartree-Fock (HF) exchange into the functional – we do not distinguish between range-separated and traditional hybrids. The highest rung is double hybrids, (DHDF) which add a wave-function based correlation energy (usually second order Møller Plesset perturbation theory) into the mixture. As a rule of thumb, higher rungs are expected to be more accurate than lower ones, but also take longer to compute. A D3(BJ) dispersion correction is used in almost all cases due to its ability to improve most DFAs in GMTKN55, with exceptions being approaches where D3(0) or another dispersion correction are more appropriate. Choices are made consistent with Sec 3.2 of Sec 4 of GMTKN55[5].

Figures 2 shows average errors for **P30-5**, **P30-10** and **P30-20**. Only the ten best methods of each rung are reported, when there are more than ten available. The reported hybrids vary significantly with the number of atoms in the poison set. However, this most likely reflects the fact that there are 47 hybrids to choose from, many with similar performance, versus 17 GGAs in the DFA suite. All 7 double hybrids and 8 meta-GGAs are reported. Supplementary Figure 2 shows errors for all DFAs broken down by set and rung.

To go beyond visual comparisons, a Kendall rank correlation coefficient[40] is computed for each of the rungs (including DFAs not shown), using,

$$\tau^N_{\mathrm{rung}} = \frac{\sum_{d\neq d'\in \mathrm{rung}} S_N(d,d')S_W(d,d')}{n_{\mathrm{rung}}(n_{\mathrm{rung}}-1)}\,. \tag{5}$$

Here, $S_N(d,d') = \mathrm{sgn}[\mathrm{MAD}_{\mathrm{P30}\text{-}N}(d) - \mathrm{MAD}_{\mathrm{P30}\text{-}N}(d')]$ $S_W(d,d') = \mathrm{sgn}[\mathrm{WTMAD}(d) - \mathrm{WTMAD}(d')]$ and n_{rung} is the number of DFA in the given rung. τ reveals how similar the poison sets rank DFAs compared to WTMAD-2. A value of 100% indicates perfect agreement between the two rankings, -100% indicates a perfect reversal of ranking, and 0% indicates no relationship between the poison and WTMAD-2 rankings. Values are shown in the relevant figures. A Kendall coefficient, $\tau_{\mathrm{P30}\text{-}N}$, for all DFAs is also provided.

Four main conclusions may be drawn from the Kendall coefficients and Figure 2: i) hybrids and GGAs have minimal correlation between the poison sets and WTMAD-2; ii) double hybrids are reliable in the sense that they performance on difficult (poison) systems reflects their overall performance, as indicated by relative large values of τ; iii) meta-GGAs tend to sacrifice quality on poison reactions for better overall accuracy, as indicated by negative values of τ; iv) correlation is much lower in **P30-20** than the other two sets, as indicated by $|\tau|$ approaching zero.

Can we better understand density functional approximations by using outliers?

Finally, we address the question of what we can learn from the poison sets. Many of the most accurate DFAs on GMTKN55 involve multiple empirical numbers of parameters that are optimized on large benchmark sets. Unsurprisingly, these also perform well on the poison sets, albeit less well and less consistently than on the complete set. Unfortunately, the large number of pa-

rameters make it difficult to understand sources of errors.

PBE0[41,42] and B3LYP[43,44] offer a simpler level of empiricism, which do allow analysis. PBE0 may be considered to have one empirical parameter being the fraction of Hartree-Fock exchange used in the hybrid,

$$\mathrm{PBE0}_\alpha := \alpha \times \mathrm{xHF} + (1-\alpha) \times \mathrm{xPBE} + \mathrm{cPBE}\,, \qquad (6)$$

and which is set to 0.25 in regular PBE0. B3LYP enhances this with a second parameter for the mixture of LDA and GGA and uses $\alpha \approx 0.2$. Since PBE0 has just one parameter we investigate it further. Barring exceptional cases, each reaction will have a value, α_0, such that the errors of HF and PBE cancel out. That is, we can find $\alpha_0(I)$ for each reaction I such that $\mathrm{Err}(\mathrm{PBE0}_{\alpha_0(I)}, I) = 0$.

We compute errors as a function of α for all reactions in **P30-5**. This lets us determine α_0, which we plot in the left panel of Figure 3. The right plot shows errors as a function of α, broken down by subset (i.e. MAD within each subset) and over the full poison set – we restrict to subsets with more than two reactions present in P30-5. In some cases, calculations with $0 \leq \alpha \leq 1$ did not contain the optimal value, so we extrapolated α_0 – see Supporting Information. G2RC:11 (exchange of Si and C in SiF_4 and CCl_4) is particularly bad, with a predicted and highly unphysical $\alpha_0 \approx -2$.

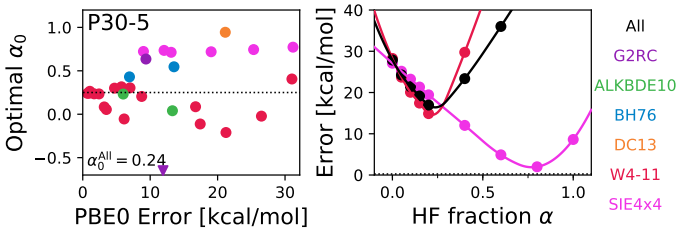

Fig. 3 Optimal HF exchange mixing parameter, α_0, for the thirty reactions in **P30-5**. The left plot shows α_0 as a function of the PBE0 error for each reaction. The right plot shows the MAD for each subset as a function of α, and across the full poison set. Colours indicate the subset. The arrow indicates α_0 outside of the reported range.

We return to the prominent benchmark sets SIE4x4 and W4-11 in order to discuss patterns in α_0 revealed by Figure 3. SIE4x4 consistently gives $\alpha_0 \approx 0.7$. Self-interaction errors are not present in HF theory and so these larger α values effectively offset errors seen in SIE4x4. W4-11 is an atomisation benchmark set yielding α_0 scattered around 0.25. This is consistent with the ideal α observed in most hybrid studies which focus on atomization energies as the benchmark, especially the earliest hybrid works.[42,43,45]

Remarkably, the optimal α for all full poison sets, that is **P30-5**, **P30-10** and **P30-20**, is also close to the PBE0 value of 0.25. This can be seen from the right plot of Figure 3 and Supplementary Figure 3. However, we also note that the optimal value of α is clearly an average, and that α_0 is very system specific. Indeed, much has been made of the seemingly fortuitous one quarter balance of exact exchange, but with little consensus emerging in the last two decades. Optimally tuned hybrid functionals,[46–68] which identify an optimal α for a given reaction using accessible criteria,

Table 1 DM21 results compared against various global hybrid DFA. All energies in kcal/mol.

Set	DM21	ωB97X-V	M052X	PBE0	B3LYP
WTMAD-2	3.98	3.93	4.62	6.59	6.39
P30-5	5.60	9.46	11.15	11.75	11.94
P30-10	6.53	10.22	10.72	12.97	13.48
P30-20	7.32	38.07	25.59	19.70	31.73

are an increasingly common way to avoid the 'scatter' in α_0.

This analysis may offer insights into the good behaviour of the recently introduced DM21 DFA,[69] which is based on a machine-learned local-hybrid ansatz in which α varies in space. Because α can vary in space, DM21 is able to 'see' local conditions and adjust the weight accordingly. This feature cannot be replicated by global hybrids where α is a unique constant – all hybrid DFA reported except DM21 are global in nature. As is revealed in Table 1, DM21 out-performs global hybrids (by a factor of five on **P30-20**, versus ωB97x-V[70]) for difficult 'poison' cases, without sacrificing its strong overall performance (here illustrated by WTMAD-2) which is as good as top-performing (on WTMAD-2) ωB97x-V and better than second-ranked M052X.

Interestingly, PBE0 (one parameter) performs much better on the largest poison set than the highly-empirical and accurate ωB97x-V (around 15 parameters) and M052X[71] (around 20 parameters), and the popular B3LYP (two parameters). [Note, DM21 has effectively thousands of parameters] The reason for this unusual success should be investigated thoroughly as it may offer strategies to improve robustness of global hybrid DFAs.

To conclude, this work introduced several new benchmark sets of difficult 'poison' reactions. The three sets, **P30-5**, **P30-10** and **P30-20**, consist of thirty reactions involving systems with up to five, ten and 20 atoms, respectively. Each set is composed of the reactions in GMTKN55 that are most difficult for DFAs to reproduce, and which likely offer the most challenging test for any DFA. They thus provide benchmark sets that are particularly difficult to model, and which serve as a stringent test of any DFA.

With the exception of double hybrid DFAs, which are quite expensive to compute, a high overall ranking on WTMAD-2 did not necessarily mean good success on P30-*N*. This means that overall success is only weakly correlated with success on difficult systems, especially as they become larger and especially for meta-GGAs, which should therefore be used with caution. Selection of a DFA using WTMAD-2 (or, we expect, similar metrics) may therefore be of little predictive value when studying difficult reaction processes, such as those involved in complex catalysis.

Acknowledgement

This work was supported by the Australian Research Council (DP200100033).

References

1 K. Burke, *J. Chem. Phys.*, 2012, **136**, 150901.
2 A. D. Becke, *J. Chem. Phys.*, 2014, **140**, 18A301.
3 A. Pribram-Jones, D. A. Gross and K. Burke, *Annu. Rev. Phys. Chem.*, 2015, **66**, 283–304.
4 S. Grimme and P. R. Schreiner, *Angew. Chem.*, 2018, **57**, 4170–4176.
5 L. Goerigk, A. Hansen, C. Bauer, S. Ehrlich, A. Najibi and S. Grimme, *Phys. Chem. Chem. Phys.*, 2017, **19**, 32184–32215.
6 N. Mardirossian and M. Head-Gordon, *Mol. Phys.*, 2017, **115**, 2315–2372.
7 M. Korth and S. Grimme, *J. Chem. Theory Comput.*, 2009, **5**, 993–1003.
8 T. Gould, *Phys. Chem. Chem. Phys.*, 2018, **20**, 27735–27739.
9 H. Yu and D. G. Truhlar, *J. Chem. Theor. Comput.*, 2015, **11**, 2968–2983.
10 Y. Zhao, B. J. Lynch and D. G. Truhlar, *Phys. Chem. Chem. Phys.*, 2005, **7**, 43–52.
11 Y. Zhao, N. González-García and D. G. Truhlar, *J. Phys. Chem. A*, 2005, **109**, 2012–2018.
12 L. Goerigk and S. Grimme, *J. Chem. Theor. Comput.*, 2010, **6**, 107–126.
13 Y. Zhao and D. G. Truhlar, *Theor. Chem. Acc.*, 2008, **120**, 215–241.
14 S. Grimme, *J. Chem. Phys.*, 2006, **124**, 034108.
15 S. Grimme, C. Mück-Lichtenfeld, E.-U. Würthwein, A. W. Ehlers, T. Goumans and K. Lammertsma, *J. Phys. Chem. A*, 2006, **110**, 2583–2586.
16 M. Piacenza and S. Grimme, *J. Comput. Chem.*, 2004, **25**, 83–99.
17 H. L. Woodcock, H. F. Schaefer and P. R. Schreiner, *J. Phys. Chem. A*, 2002, **106**, 11923–11931.
18 P. R. Schreiner, A. A. Fokin, R. A. Pascal and A. de Meijere, *Org. Lett.*, 2006, **8**, 3635–3638.
19 C. Lepetit, H. Chermette, M. Gicquel, J.-L. Heully and R. Chauvin, *J. Phys. Chem. A*, 2007, **111**, 136–149.
20 J. S. Lee, *J. Phys. Chem. A*, 2005, **109**, 11927–11932.
21 A. Karton and J. M. Martin, *Mol. Phys.*, 2012, **110**, 2477–2491.
22 Y. Zhao, O. Tishchenko, J. R. Gour, W. Li, J. J. Lutz, P. Piecuch and D. G. Truhlar, *J. Phys. Chem. A*, 2009, **113**, 5786–5799.
23 D. Manna and J. M. Martin, *J. Phys. Chem. A*, 2016, **120**, 153–160.
24 L. A. Curtiss, K. Raghavachari, P. C. Redfern and J. A. Pople, *The Journal of Chemical Physics*, 1997, **106**, 1063–1079.
25 A. Karton, S. Daon and J. M. Martin, *Chem. Phys. Lett.*, 2011, **510**, 165–178.
26 Y. Zhang and W. Yang, *J. Chem. Phys.*, 1998, **109**, 2604.
27 P. Mori-Sánchez, A. J. Cohen and W. Yang, *J. Chem. Phys.*, 2006, **125**, 201102.
28 A. J. Cohen, P. Mori-Sánchez and W. Yang, *Science*, 2008, **321**, 792.
29 X. Zheng, M. Liu, E. R. Johnson, J. Contreras-García and W. Yang, *J. Chem. Phys.*, 2012, **137**, 214106.
30 E. R. Johnson, A. Otero-de-la Roza and S. G. Dale, *J. Chem. Phys.*, 2013, **139**, 184116.
31 E. R. Johnson, M. Salamone, M. Bietti and G. A. DiLabio, *J. Phys. Chem. A.*, 2013, **117**, 947–952.
32 A. Otero-De-La-Roza, E. R. Johnson and G. A. DiLabio, *J. Chem. Theor. Comput.*, 2014, **10**, 5436–5447.
33 S. R. Whittleton, X. A. Sosa Vazquez, C. M. Isborn and E. R. Johnson, *J. Chem. Phys.*, 2015, **142**, 184106.
34 D. R. Lonsdale and L. Goerigk, *Phys. Chem. Chem. Phys.*, 2020, **22**, 15805–15830.
35 W. Yang, Y. Zhang and P. W. Ayers, *Phys. Rev. Lett.*, 2000, **84**, 5172.
36 A. J. Cohen, P. Mori-Sánchez and W. Yang, *J. Chem. Phys.*, 2008, **129**, 121104.
37 A. D. Becke, *J. Chem. Phys.*, 2013, **138**, 074109.
38 A. D. Becke, *J. Chem. Phys.*, 2013, **138**, 161101.
39 J. P. Perdew and K. Schmidt, AIP Conference Proceedings, 2001, pp. 1–20.
40 M. G. Kendall, *Biometrika*, 1938, **30**, 81–93.
41 J. P. Perdew, K. Burke, and M. Ernzerhof, *Phys. Rev. Lett.*, 1996, **77**, 3865–68.
42 C. Adamo and V. Barone, *J. Chem. Phys.*, 1999, **110**, 6158–69.
43 A. D. Becke, *J. Chem. Phys.*, 1993, **98**, 1372–77.
44 C. Lee, W. Yang and R. G. Parr, *Phys. Rev. B*, 1988, **37**, 785.
45 A. D. Becke, *J. Chem. Phys.*, 1993, **98**, 1372.
46 U. Salzner and R. Baer, *J. Chem. Phys.*, 2009, **131**, 231101.
47 R. Baer, E. Livshits and U. Salzner, *Annual review of physical chemistry*, 2010, **61**, 85–109.
48 T. Stein, H. Eisenberg, L. Kronik and R. Baer, *Phys. Rev. Lett.*, 2010, **105**, 266802.
49 N. Sai, P. F. Barbara and K. Leung, *Phys. Rev. Lett.*, 2011, **106**, 226403.
50 S. Refaely-Abramson, R. Baer and L. Kronik, *Phys. Rev. B*, 2011, **84**, 075144.
51 L. Kronik, T. Stein, S. Refaely-Abramson and R. Baer, *J. Chem. Theor. Comput.*, 2012, **8**, 1515–1531.
52 S. Refaely-Abramson, S. Sharifzadeh, M. Jain, R. Baer, J. B. Neaton and L. Kronik, *Phys. Rev. B*, 2013, **88**, 081204.
53 S. Zheng, E. Geva and B. D. Dunietz, *J. Chem. Theor. Comput.*, 2013, **9**, 1125–1131.
54 T. B. de Queiroz and S. Kümmel, *J. Chem. Phys.*, 2014, **141**, 084303.
55 H. Phillips, Z. Zheng, E. Geva and B. D. Dunietz, *Org. Electron.*, 2014, **15**, 1509–1520.
56 S. Refaely-Abramson, M. Jain, S. Sharifzadeh, J. B. Neaton and L. Kronik, *Phys. Rev. B*, 2015, **92**, 081204.
57 T. B. de Queiroz and S. Kümmel, *J. Chem. Phys.*, 2015, **143**, 034101.
58 Z. Zheng, J.-L. Bredas and V. Coropceanu, *J. Phys. Chem. Lett.*, 2016, **7**, 2616–2621.
59 D. Neuhauser, E. Rabani, Y. Cytter and R. Baer, *J. Phys. Chem. A*, 2016, **120**, 3071–3078.
60 H. Sun, S. Zhang, C. Zhong and Z. Sun, *J. Comput. Chem.*, 2016, **37**, 684–693.
61 H. Sun, S. Ryno, C. Zhong, M. K. Ravva, Z. Sun, T. Körzdörfer and J.-L. Bredas, *J. Chem. Theor. Comput.*, 2016, **12**, 2906–2916.
62 M. Rubesova, E. Muchova and P. Slavicek, *J. Chem. Theor. Comput.*, 2017, **13**, 4972–4983.
63 A. Boruah, M. P. Borpuzari, Y. Kawashima, K. Hirao and R. Kar, *J. Chem. Phys.*, 2017, **146**, 164102.
64 M. Alipour and Z. Safari, *J. Phys. Chem. C*, 2018, **123**, 746–761.
65 M. Alipour, *J. Comput. Chem.*, 2018, **39**, 1508–1516.
66 A. J. Lee, M. Chen, W. Li, D. Neuhauser, R. Baer and E. Rabani, *Phys. Rev. B*, 2020, **102**, 035112.
67 S. G. Dale and E. R. Johnson, *J, Chem. Phys.*, 2015, **143**, 184112.
68 L. O. Hemmingsen, O. A. Hervir and S. G. Dale, *J. Chem. Phys.*, 2022, **156**, 014106.
69 J. Kirkpatrick, B. McMorrow, D. H. P. Turban, A. L. Gaunt, J. S. Spencer, A. G. D. G. Matthews, A. Obika, L. Thiry, M. Fortunato, D. Pfau, L. R. Castellanos, S. Petersen, A. W. R. Nelson, P. Kohli, P. Mori-Sánchez, D. Hassabis and A. J. Cohen, *Science*, 2021, **374**, 1385–1389.
70 J.-D. Chai and M. Head-Gordon, *J. Chem. Phys.*, 2008, **128**, 084106.
71 Y. Zhao, N. E. Schultz, and D. G. Truhlar, *J. Chem. Theory and Comput.*, 2006, **2**, 364–82.
72 D. G. Smith, L. A. Burns, A. C. Simmonett, R. M. Parrish, M. C. Schieber, R. Galvelis, P. Kraus, H. Kruse, R. Di Remigio, A. Alenaizan and et al, *J. Chem. Phys.*, 2020, **152**, 184108.
73 F. Weigend and R. Ahlrichs, *Phys. Chem. Chem. Phys.*, 2005, **7**, 3297–3305.