Electronic Supplementary Material (ESI) for Physical Chemistry Chemical Physics.

Supporting Information

COMPAS-3: A Computational Dataset of *peri*-Condensed Polybenzenoid Hydrocarbons

Alexandra Wahab[a] and Renana Gershoni-Poranne*[b]

[a] *The Laboratory for Organic Chemistry, Department of Chemistry and Applied Biosciences, ETH Zürich, 8093 Zürich, Switzerland*
[b] *The Schulich Faculty of Chemistry and the Resnick Sustainability Center for Catalysis, Technion – Israel Institute of Technology, Haifa 32000, Israel*

e-mail: rporanne@technion.ac.il

Contents

Abbreviations and Acronyms

E_{rel}	relative single-point energy
aEA	adiabatic electron affinity
aIP	adiabatic ionization potential
CaGe	Chemical & abstract Graph environment
COMPAS	computational database of polycylic aromatic systems
DFT	density functional theory
HLG	HOMO-LUMO gap
HOMO	highest occupied molecular orbital
LUMO	lowest unoccupied molecular orbital
MARE	mean absolute relative error
PCD	pivoted Cholesky decomposition
RMSD	root-mean-square deviation
TD-DFT	time-dependant density functional theory
xTB	extended tight-binding

S1 Ground-state calculations

All GFN2-xTB calculations were performed with xTB[S1] version 6.2. All DFT calculations were performed with ORCA 5.0.3[S2,S3] using Grimme's D3[S4] dispersion correction and the Becke-Johnson damping scheme.[S5,S6]

S1.1 Calculation input template

In this section, we provide input templates for all types of calculations performed to generate COMPAS-3x and COMPAS-3D.

S1.1.1 xTB input templates

The xTB calculations were performed identically to COMPAS-1:[S7] the CaGe[S8]-generated xyz-coordinates were optimized with the following command, where `NAME` is a placeholder for the identifying name given to each molecule.

```
xtb NAME.xyz --ohess vtight --molden > NAME.out
```

The xTB-optimized geometry of the neutral form was then used to calculate the anionic and cationic forms as follows.

```
xtb xtbopt.xyz --ohess vtight --molden --chrg -1 --uhf 1 > NAME_-1.out
xtb xtbopt.xyz --ohess vtight --molden --chrg +1 --uhf 1 > NAME_1.out
```

S1.1.2 DFT input template

The following input was used to calculate the COMPAS-3D dataset. We used the CAM-B3LYP[S9–S13] functional and the def2-SVP[S14] basis set for optimization, followed by a single point with a larger basis set (aug-cc-pVDZ)[S15–S17] with Grimme's D3[S4] dispersion correction and the Becke-Johnson damping scheme.[S5,S6] The choice functional and basis set was done on the basis of having the best balance between accuracy and efficiency for TD-DFT calculations as we intend to add excited state properties to our database and want to keep the method consistent overall to enable comparison. The selection of functional/basis for TD-DFT calculations is detailed in the next Section S2.

In the template below, `CHARGE` is a placeholder for the charge $(-1, 0, +1)$ of the molecule and `MULTIPLICITY` is a placeholder for the respective multiplicity. The starting xyz coordinates used for optimization are xTB-optimized coordinates. The single point uses the DFT-optimized coordinates to calculated the properties with a larger basis set.

```
#### Optimization ####

# functional and basis set
! cam-b3lyp def2-svp

# accuracy, approximations, and dispersion corrections
! tightscf rijcosX def2/j d3bj

# type of calculation
! opt

# output control
```

```
! miniprint

%base"cam-b3lyp_def2-svp_opt_0"

# type of input; charge; multiplicity; input
*xyzfile CHARGE MULTIPLICITY xtbopt.xyz

$new_job
#### Single Point ####

# functional and basis set
! cam-b3lyp cc-pVDZ

# accuracy, approximation, and dispersion corrections
! tightscf rijcosx d3bj

# type of calculation
! sp PModel NoTrah

# output control
! normalprint printgap

# basis set block
%basis
newgto C "aug-cc-pVDZ" end   # set aug-cc-pVDZ as the basis set for C atoms
auxj "aug-cc-pVTZ/JK"        # auxiliary basis set
end

# scf block
%scf
sthresh 1e-6                 # set threshold for smallest eigen value
end

%base"cam-b3lyp_aug-cc-pvdz_sp_0"

# type of input; charge; multiplicity; input
*xyzfile CHARGE MULTIPLICITY cam-b3lyp_def2-svp_opt_0.xyz
```

S2 Benchmarking for TD-DFT

We performed a benchmarking study to identify a cost-effective and accurate functional/basis set combination for TD-DFT calculations of the T_1 and S_1 excited states. Although COMPAS-3 contains only closed-shell ground-state molecules by design, future installments of the COMPAS Project are focused on excited states. To ensure compatibility for future analyses comparing the ground and excited states, we needed to choose a level of theory that would be suitable for both cases. Since excited-state calculations are much more sensitive to method choice, we benchmarked the methods for excited states to calculate COMPAS-3. The selected methods have been extensively used for ground-state PASs and are generally considered appropriate for them.

We selected CAM-B3LYP[S13] as our functional as it was found to be one of the best performing functionals in a recent benchamrking study conducted by Head-Gordon and coworkers.[S18] The choice of which basis set to test was also based on the benchmarking paper, where the autors compared multiple basis set families – aug-cc-pVXZ ($X = D, T, Q$),[S15–S17] d-aug-cc-pVXZ,[S19] and the def2 series[S14,S20] – on a select few functionals. From their results, we selected the best performing and less resource consuming basis set family—aug-cc-pVXZ. We compared the performances of aug-cc-pVDZ (aDZ) and aug-cc-pVTZ (aTZ) on the same benchmarking set previously use in our DFT benchmarking for COMPAS-1 (see Figure S1).

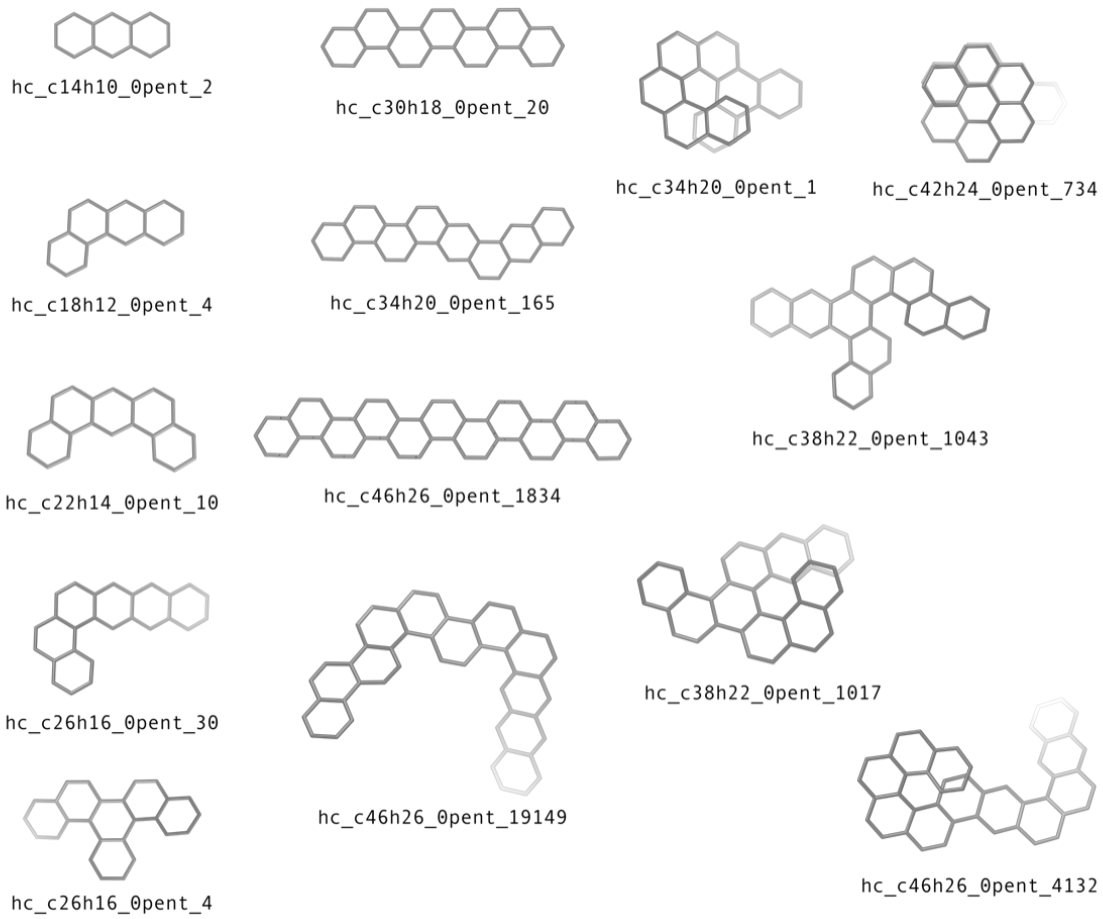

Figure S1. Sample of 14 molecules from COMPAS-1 with 3 to 11 rings used for benchmarking.

aTZ is significantly more costly than aDZ, especially as the calculated system gets larger.

Nonetheless, it has the advantage of having a dedicated auxiliary basis set implemented in ORCA 5.0.3, which is not the case for aDZ. Auxiliary basis sets are important for the simultaneous fitting of Coulomb and exchange integrals. In other words, it is essential for reducing calculation run times. Thus, we tested three solutions to remedy this problem: (1) using the def2/J basis set (test 1); (2) using a pivoted Cholesky decomposition (PCD) with 'AutoAux', meaning that the adequate auxiliary basis set is chosen automatically and the PCD will deal with linear dependencies that may arise (test 2); (3) using aug-cc-pVTZ/JK auxiliary basis set (test 3). We also added a fourth option (final), which speeds up the calculations significantly (see Figure S2). All five set-ups are described in Table S2.

Table S2. The different set-ups compared for the benchmarking.

	Set up details
aTZ	CAM-B3LYP functional / aug-cc-pVTZ basis set / aug-cc-pVTZ/JK auxiliary
Test 1	CAM-B3LYP functional / aug-cc-pVDZ basis set / Def2/JK auxiliary

Test 2	CAM-B3LYP functional / aug-cc-pVDZ basis set
	`%basis`
	`auxj "AutoAux"`
	`PCDTrimAuxJ  Coulomb  # Trim the AuxJ basis in the`
	`# Coulomb metric`
	`PCDThresh -1  # Threshold for the PCD:`
	`# chosen automatically if >0`
	`end`

Test 3	CAM-B3LYP functional with aug-cc-pVDZ basis set and aug-cc-pVTZ/JK auxiliary

Final	CAM-B3LYP functional with aug-cc-pVDZ basis set
	`%basis`
	`newgto C "aug-cc-pVDZ" end  # set aug-cc-pVDZ as the`
	`# basis set for C atoms`
	`auxj "aug-cc-pVTZ/JK"  # auxiliary basis set`
	`end`

Some of the calculations were harder to converge, notably 8 out of the 14 molecules required adjustment when calculated with the aTZ basis set. We gradually increased 'DIISMaxEq' value until convergence. This keyword indicates the number of Fock matrices to be remembered for the DIIS extrapolation. The default value is 5, but difficult systems might require a value between 15-40. If convergence was not reached, we lowered the 'DirectResetFreq' value to 5, then to 1. This keyword controls how often the full Fock matrix is rebuilt. The default value is 15. This keyword greatly affects the run time of the computations as building the Fock matrix is a costly computation. The molecules that required these changes are show in Table S3.

To check that the more cost effective choices would not lead to large deviations of the calculated properties, we compared the vertical T_1 (Table S4) and vertical S_1 (Table S5) excitation energies to aTZ, which should be the most accurate from the set-ups compared.

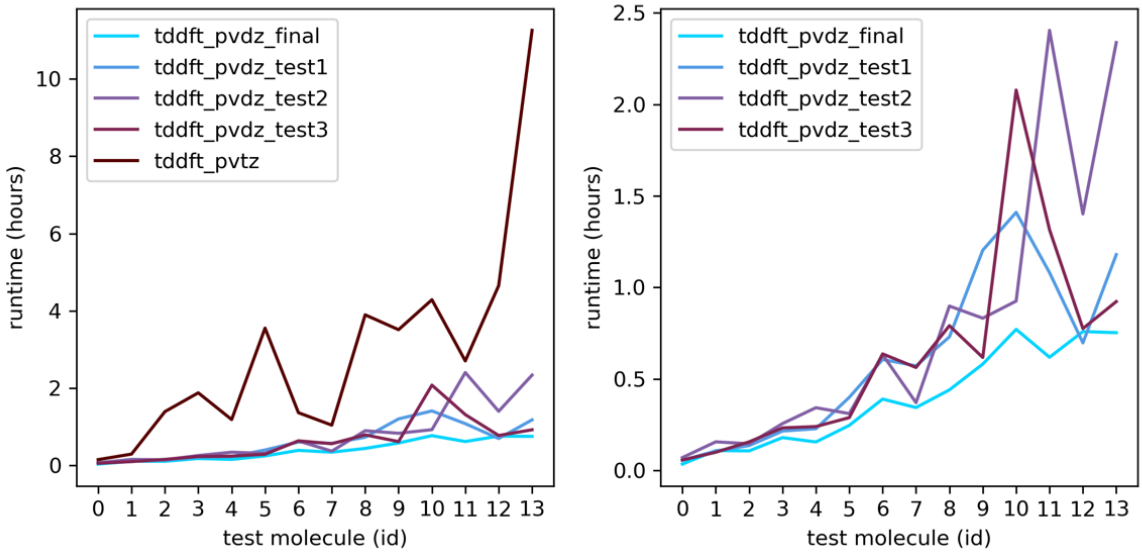

Figure S2. Total run times in hours obtained from jobs performed with 12 processors each. On the left: the five set-ups we compared. On the right: removed aTZ for better comparison of faster set-ups.

Table S3. Changes made to the %scf block.

ID	Molecule	aTZ	Test 1	Test 2	Test 3	Final
1	hc_c14h10_0pent_2	-	-	-	-	-
2	hc_c18h12_0pent_4	-	-	DirectResetFreq 5	-	-
3	hc_c22h14_0pent_10	DIISMaxEq 20	-	-	-	-
4	hc_c26h16_0pent_30	-	-	-	-	-
5	hc_c26h16_0pent_4	-	-	-	-	-
6	hc_c30h18_0pent_20	DIISMaxEq 20	-	-	-	-
7	hc_c34h20_0pent_1	-	-	-	-	-
8	hc_c34h20_0pent_165	-	-	DirectResetFreq 1	-	-
9	hc_c38h22_0pent_1017	DIISMaxEq 20	-	-	-	-
10	hc_c38h22_0pent_1043	DIISMaxEq 15	-	-	-	-
11	hc_c42h24_0pent_734	DirectResetFreq 1	-	-	-	-
12	hc_c46h26_0pent_1834	DIISMaxEq 20	-	-	-	-
13	hc_c46h26_0pent_19149	DirectResetFreq 1	-	-	-	-
14	hc_c46h26_0pent_4132	DIISMaxEq 15	-	-	-	-

All set-ups showed extremely good correlations with the aTZ energies with slopes equal to 1 and intercepts going through the origin, and R^2 varying from 0.9999 to 1.0000.

The set-up we converged on (referred to as final in the tables and plots) showed good accuracy when compared to aTZ, while displaying the shortest run times (Figure S2). Additionally, the initial input did not require any modifications for all 14 test molecules and converged without issues, which is an important quality for high-throughput calculations that we want to perform.

Table S4. T_1 excitation energies for the five set-up that were tested. All values reported are in eV.

ID	Molecule	aDZ				aTZ
		Final	**Test 1**	**Test 2**	**Test 3**	
1	hc_c14h10_0pent_2	2.128	2.129	2.128	2.128	2.127
2	hc_c18h12_0pent_4	2.384	2.384	2.384	2.384	2.382
3	hc_c22h14_0pent_10	2.605	2.605	2.604	2.604	2.603
4	hc_c26h16_0pent_30	1.613	1.614	1.614	1.614	1.612
5	hc_c26h16_0pent_4	2.625	2.626	2.626	2.626	2.624
6	hc_c30h18_0pent_20	2.748	2.749	2.749	2.749	2.748
7	hc_c34h20_0pent_1	2.538	2.539	2.539	2.539	2.536
8	hc_c34h20_0pent_165	2.370	2.371	2.370	2.370	2.369
9	hc_c38h22_0pent_1017	2.113	2.114	2.114	2.114	2.112
10	hc_c38h22_0pent_1043	2.208	2.209	2.208	2.208	2.206
11	hc_c42h24_0pent_734	2.190	2.191	2.191	2.191	2.189
12	hc_c46h26_0pent_1834	2.697	2.698	2.697	2.697	2.697
13	hc_c46h26_0pent_19149	1.644	1.644	1.644	1.644	1.643
14	hc_c46h26_0pent_4132	2.038	2.039	2.039	2.039	2.037

Table S5. S_1 excitation energies for the five set-up that were tested. All values reported are in eV.

ID	Molecule	aDZ				aTZ
		Final	**Test 1**	**Test 2**	**Test 3**	
1	hc_c14h10_0pent_2	3.726	3.726	3.725	3.725	3.715
2	hc_c18h12_0pent_4	3.825	3.824	3.824	3.824	3.816
3	hc_c22h14_0pent_10	3.749	3.748	3.749	3.749	3.744
4	hc_c26h16_0pent_30	2.981	2.981	2.981	2.981	2.972
5	hc_c26h16_0pent_4	3.653	3.653	3.653	3.653	3.648
6	hc_c30h18_0pent_20	3.771	3.771	3.771	3.771	3.766
7	hc_c34h20_0pent_1	3.431	3.431	3.431	3.431	3.426
8	hc_c34h20_0pent_165	3.494	3.493	3.493	3.493	3.489
9	hc_c38h22_0pent_1017	3.310	3.310	3.310	3.309	3.304
10	hc_c38h22_0pent_1043	3.410	3.410	3.410	3.410	3.401
11	hc_c42h24_0pent_734	3.230	3.230	3.230	3.230	3.225
12	hc_c46h26_0pent_1834	3.675	3.675	3.675	3.675	3.670
13	hc_c46h26_0pent_19149	2.999	2.999	2.998	2.998	2.990
14	hc_c46h26_0pent_4132	3.278	3.278	3.278	3.278	3.270

S3 Mean absolute relative errors of correction schemes

As shown in the main text, the values obtained with xTB are quite far from those calculated with DFT. Nonetheless, there is a good linear correlation between the two methods. Thus, we used a linear regression for each property to "correct" the xTB values to DFT. These corrected values are included in the COMPAS-3x dataset. We report here the mean absolute relative error (MARE) for each of the corrected properties. The MARE was calculated as follows

$$\mathrm{MARE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_{\mathrm{true},i} - y_{\mathrm{pred},i}}{y_{\mathrm{true},i}}\right| * 100 \tag{S1}$$

where $y_{\mathrm{true},i}$ is the DFT-calculated value and $y_{\mathrm{pred},i}$ is the xTB-to-DFT predicted value. N is the dataset size (8,844).

Table S6. MARE of the corrections of xTB toward DFT-level values.

Property	**MARE**
HOMO	0.462%
LUMO	2.628%
HOMO-LUMO gap	0.893%
aIP	1.128%
aEA	5.744%

S4 COMPAS-1D functional/basis set comparison

As discussed in the main text, we initially calculated the properties of the COMPAS-1D dataset using a different functional and basis set: B3LYP-D3BJ/def2-SVP. We recalculated the complete COMPAS-1D dataset at the same level as the COMPAS-3D, CAM-B3LYP-D3BJ/aug-cc-pVDZ. Here, we compare the two different functional/basis set combos and show that the values obtained by both combos are very similar in the case of aEA and E_{rel}, while the other properties (HOMO, LUMO, HLG, and aIP) seem to mostly be affected by an offset. Thus, the analysis we conducted in reference S7 still holds.

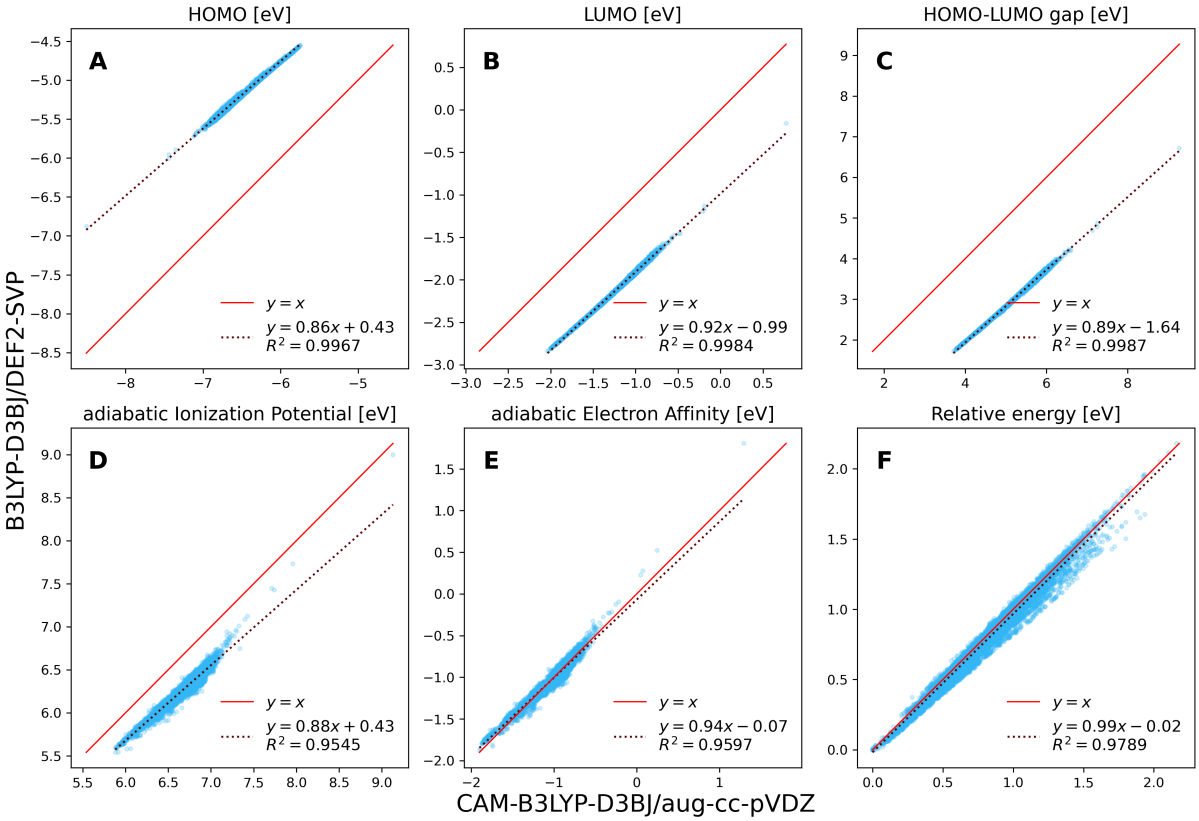

Figure S3. Scatter plots of the various molecular properties, calculated with B3LYP-D3BJ/def2-SVP versus CAM-B3LYP-D3BJ/aug-cc-pVDZ for COMPAS-1D: A) HOMO; B) LUMO; C) HLG; D) aIP; E) aEA; F) E_{rel}. The $x = y$ line is shown in red.

In the main text, we discuss the presence of an outliers-"island" in the xTB versus DFT regression plots for the aEA for both COMPAS-3 and COMPAS-1 (see Figure 5, in the main text). The following plot highlights (burgundy) the outliers found in COMPAS-1 at B3LYP-D3BJ/def2-SVP level in both subplots, and shows that these outliers are exactly the same with both DFT-methods used, *i.e.*, B3LYP-D3BJ/def2-SVP and CAM-B3LYP-D3BJ/aug-cc-pVDZ.

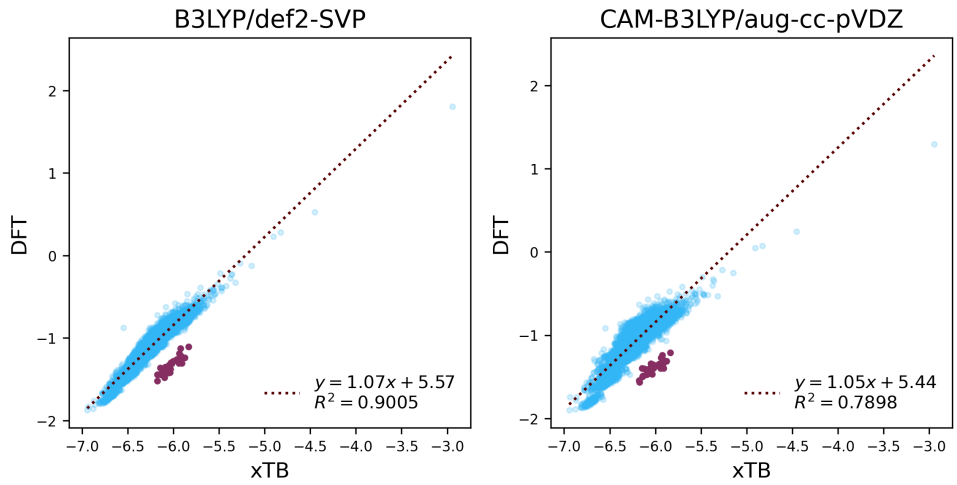

Figure S4. Scatter plots of DFT- versus xTB-calculated values of aEA at B3LYP-D3BJ/def2-SVP (left) and CAM-B3LYP-D3BJ/aug-cc-pVDZ level of theory. Outliers are colored in burgundy.

S5 Method Comparison

S5.1 xTB- vs DFT-optimized geometries

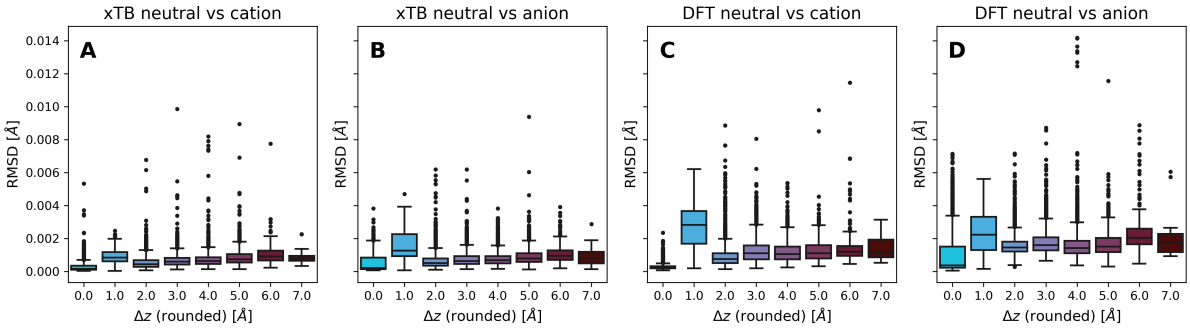

Figure S5. Boxplots of RMSD of: (**A**) xTB-optimized geometries of the neutral versus cationic species; (**B**) xTB-optimized geometries of the neutral versus anionic species; (**C**) DFT-optimized geometries of the neutral versus cationic species; (**D**) DFT-optimized geometries of the neutral versus anionic species. For each subplot, the full 8,844 dataset is separated by rounded Δz obtained from DFT-optimized geometries.

S5.2 Analysis of outliers in the aIP and aEA plots

As mentioned in the main text, in the case of aIP and aEA, we observe two groupings of molecules deviating from the regression line. In this section we discuss the two populations and look deeper into possible sources for these discrepancies.

S5.2.1 Outlier Group 1

In this section we analyze the outlier molecules referred to as "Group 1", which are highlighted in Figure S6. For these molecules, the DFT values are lower than would be expected, based on the fitting equation.

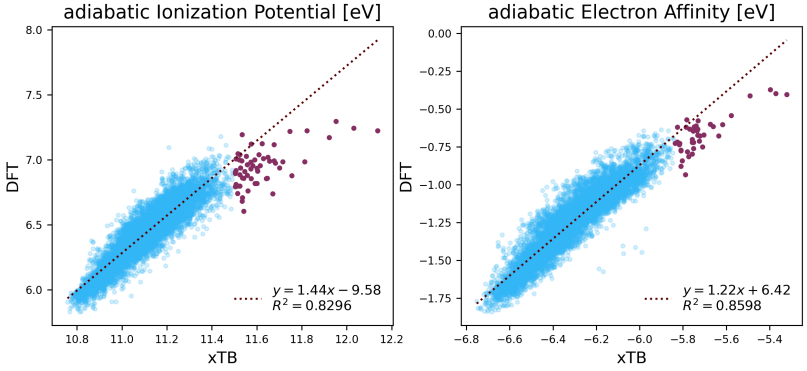

Figure S6. xTB vs DFT scatter plots of the aIP (left) and aEA (right). Outliers are colored in purple.

We counted 59 outlier molecules in the aIP plot, with aIP(xTB) > 11.5eV, and 42 outlier molecules in the aEA plot, with aEA(xTB) > −5.84eV. We then extracted the

structures of these outliers from the dataset and identified the specific molecules. Although these molecules represent only a very small fraction of the ∼9k dataset (6.67%) and are not expected to substantially affect any data-driven conclusions, we envisioned that they might prove instructional. Specifically, they may point to particular boundary conditions for which the xTB-to-DFT correction fails.

It was previously shown that aIP and aEA are affected by the molecular size,[S21,S22] which we showed was indeed the case for COMPAS-1.[S7,S23] Similarly, xTB-calculated aIP and aEA are more sensitive to system size compared to DFT-calculated ones (see Figure S7). Therefore, it is possible that this deviation from the linear regression line simply comes down to the difference in sensitivity to system size.

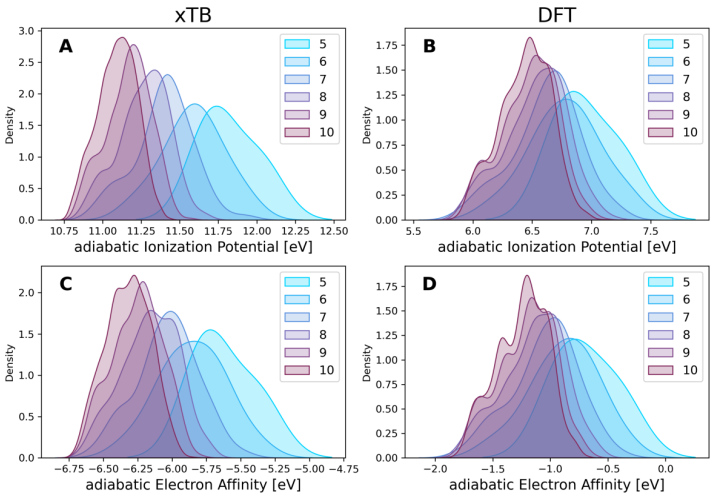

Figure S7. KDE plots per number of rings for xTB-calculated (left) and DFT-calculated (right) aIP (top) and aEA (bottom).

We first checked whether the difference in size-dependency may lead to a discrepancy in optimized geometries. Therefore, we looked at the RMSDs between the two methods for all species (see Figure S8). We observed a small increase in the RMSD between xTB- and DFT-optimized geometries with the number of rings. However, the RMSD values all remained extremely low. Thus, although the values of the properties themselves seem to be affected by the system size, the geometries are not. We could rule this out as a possible source of the discrepancy.

To further verify whether the energetic discrepancy may be due to an underlying geometric one that specifically affects the outliers, we also generated boxplots of the RMSDs between xTB- and DFT-optimized geometries of the full dataset without the outliers and of only the outliers. This was aimed to show if there is a noticeable difference in the optimized geometries of the outlying molecules between the two methods. In the case of the common outliers and the aIP-specific outliers, we did not observe any significant differences between the RMSDs of the full dataset and the outliers sets (see Figures S9 and S10, **E-G**). In contrast, the RMSDs of the four aEA-specific outliers do show an increase between the two methods for all species (see Figure S14**E-G**).

As these four outliers are highly non-planar, this may hint at the differences between xTB- and DFT-optimized geometries being linked with the deviation from planarity. However, the overall conclusion from our RMSD analysis suggests that the issue is indeed at the level of energy calculations and not just geometry optimization.

We then hypothesized about alternative origins for the difference. Considering that

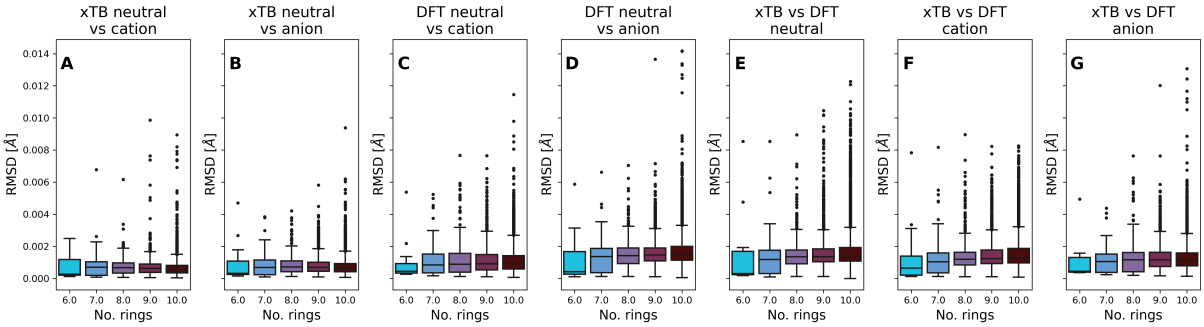

Figure S8. Boxplots of RMSD of xTB-optimized geometries of the neutral versus cationic (**A**), and anionic (**B**) forms; of DFT-optimized geometries of the neutral versus cationic (**C**), and anionic (**D**) forms; and of xTB- versus DFT-optimized geometries of the neutral (**E**), cationic (**F**), and anionic (**G**) species. For each subplot, the full 8,844 dataset is separated by number of rings. Families 4 and 5 were omitted as they contain 1 and 3 molecules, respectively.

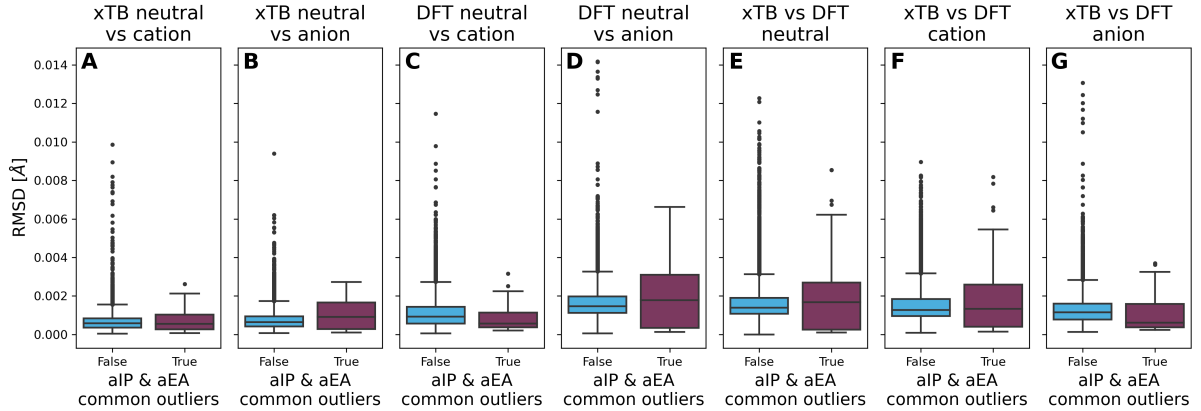

Figure S9. Boxplots of RMSD of xTB-optimized geometries of the neutral versus cationic (**A**), and anionic (**B**) forms; of DFT-optimized geometries of the neutral versus cationic (**C**), and anionic (**D**) forms; and of xTB- versus DFT-optimized geometries of the neutral (**E**), cationic (**F**), and anionic (**G**) species. For each subplot, boxplot of the complete dataset up to 10 rings without the 38 common outlier-molecules is on the left (False), and the boxplot of the 38 common outlier-molecules is on the right (True).

the outliers appear in the aEA and aIP plots, but not in the HOMO and LUMO plots, we concluded that the deviations originate in the calculations of the ionic species, rather than the neutral ones. If this is the case, then the discrepancies could be due to the difficulty of adequately treating the presence of charge on specific structural motifs. In other words, we may be able to pinpoint specific structural motifs for which xTB and DFT do not agree on how charge is stabilized. Initial investigation seemed to support the idea that specific motifs were to blame. Indeed, we found 38 molecules that are present in both outlier sets (see Figure S11), suggesting some degree of similarity between the outliers sets. However, after further investigation, this did not seem to be the case. The 21 aIP-specific outliers seem to be mostly planar (see Figure S12), while the 4 aEA-specific ones are highly non-planar (see Figure S13). The 4 aEA-specific outliers are, however, structurally extremely similar—according to visual inspection—with the largest common substructure being a pyrene unit with a helix annulated to one of its *b* rings. Thus, we could not identify a specific recurring motif for all the outliers based on visual inspection.

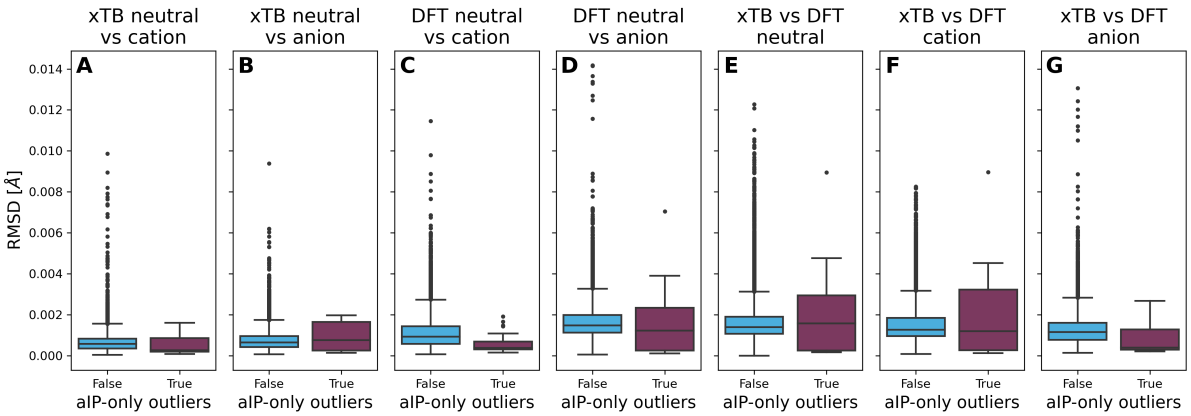

Figure S10. Boxplots of RMSD of xTB-optimized geometries of the neutral versus cationic (**A**), and anionic (**B**) forms; of DFT-optimized geometries of the neutral versus cationic (**C**), and anionic (**D**) forms; and of xTB- versus DFT-optimized geometries of the neutral (**E**), cationic (**F**), and anionic (**G**) species. For each subplot, boxplot of the complete dataset up to 10 rings without the 21 aIP outlier-molecules is on the left (False), and the boxplot of the 21 aIP outlier-molecules is on the right (True).

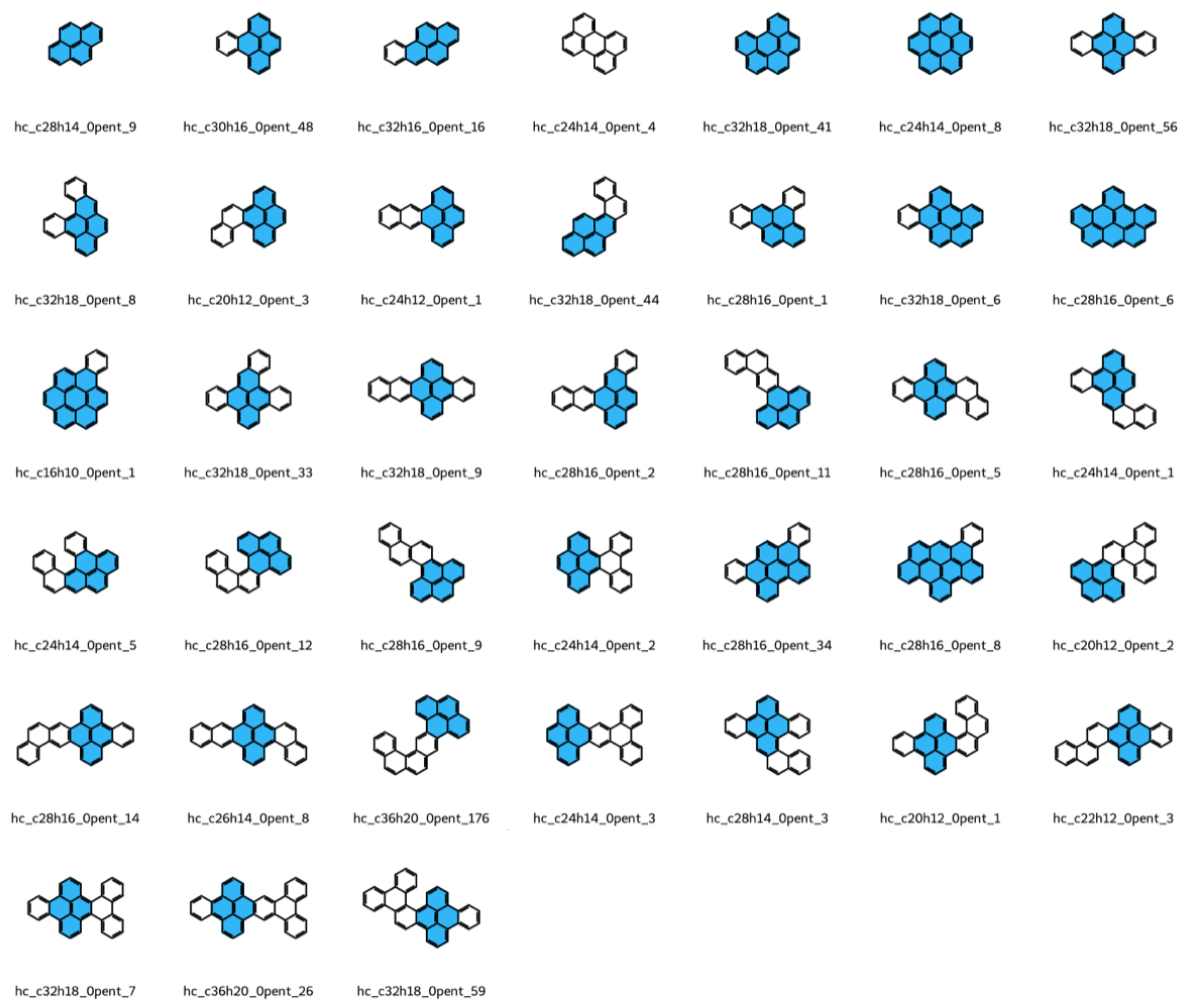

Figure S11. Outlier molecules for both the aIP and aEA. The most common substructure, pyrene, is colored in blue in each molecule where it is present.

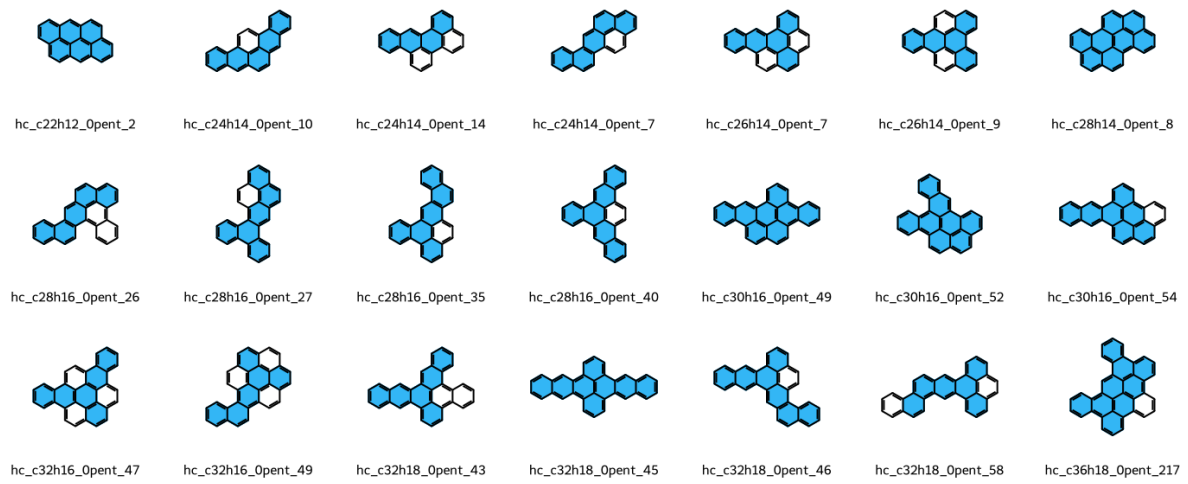

Figure S12. Outlier molecules of aIP. The most common substructure, *LA*, is colored in blue in each molecule where it is present.

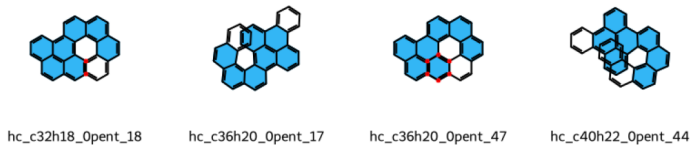

Figure S13. Outlier molecules of aEA. The most common substructure, a pyrene-unit with a helix emerging from its side, is colored in blue in each molecule where it is present.

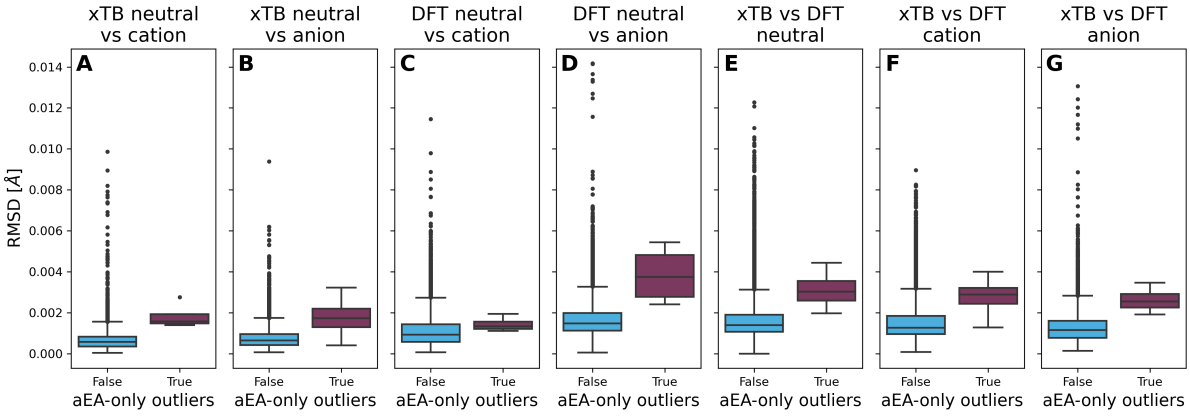

Figure S14. Boxplots of RMSD of xTB-optimized geometries of the neutral versus cationic (**A**), and anionic (**B**) forms; of DFT-optimized geometries of the neutral versus cationic (**C**), and anionic (**D**) forms; and of xTB- versus DFT-optimized geometries of the neutral (**E**), cationic (**F**), and anionic (**G**) species. For each subplot, boxplot of the complete dataset up to 10 rings without the 4 aEA outlier-molecules is on the left (False), and the boxplot of the 4 aEA outlier-molecules is on the right (True).

S5.2.2 Outlier Group 2

In this section we analyze the outlier molecules referred to as "Group 2", which are highlighted in S15. We note that we also observed such an "island" of outliers in the aEA plots of the COMPAS-1 datasets.

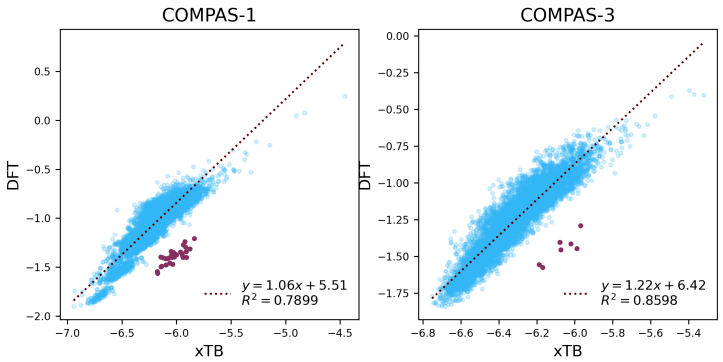

Figure S15. xTB- vs DFT-calculated values of aEA for COMPAS-1 (left) and COMPAS-3 (right). The "island" of outliers is highlighted in purple.

Because we calculated COMPAS-1 at two different levels of theory (with B3LYP-D3BJ/def2-SVP[S7] and subsequently with CAM-B3LYP-D3BJ/aug-cc-pVDZ), we were able to rule out that the presence of such outlier is method-dependent. Indeed, in both cases the outlier data points appear, and we identified these as being the exact same molecules (see Section S4). Furthermore, we see the presence of the same island when plotting the LUMO versus the aEA for DFT but not xTB (see Figure S16**B**&**D**). Thus, these observations strongly suggest that the discrepancy stems from the DFT calculations, and that it is not method-dependent.

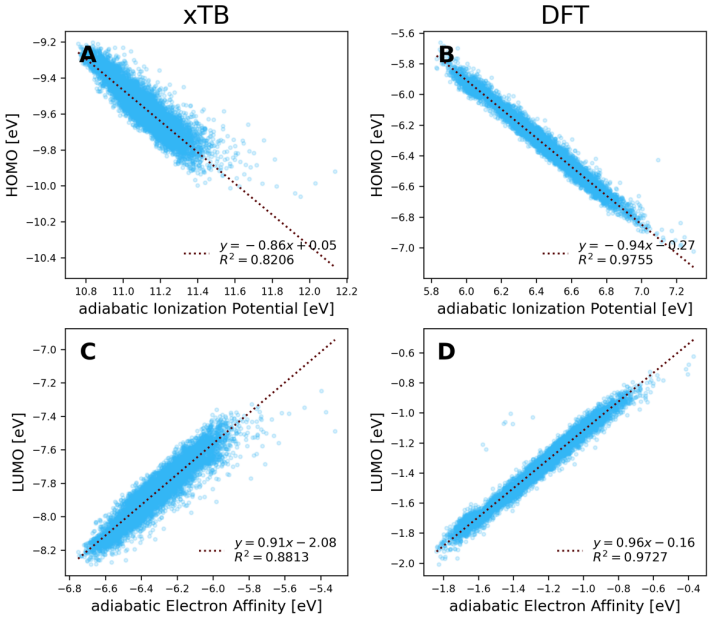

Figure S16. Scatter plots for xTB-calculated (left) and DFT-calculated (right) HOMO versus aIP (top) and LUMO versus aEA (bottom).

When comparing the COMPAS-1 and COMPAS-3 plots, we notice is that there are

much fewer molecules present in the COMPAS-3 outlier island than in the COMPAS-1 outlier island: 7 molecules for COMPAS-3 (0.08% of the dataset) as opposed to 32 for COMPAS-1 (0.37% of the dataset). This is likely due to specific structural features that are more present in cc-PBHs than in pc-PBHs.

A closer look at these 7 molecules shows a strong presence of cove motifs (see Figure S17), which is also the case for the molecules found in the aEA island in COMPAS-1 (see Figure S18).

Figure S17. Molecules found in the "island" in the xTB vs DFT scatter plot of aEA of the COMPAS-3 dataset. The cove motif is highlighted in blue.

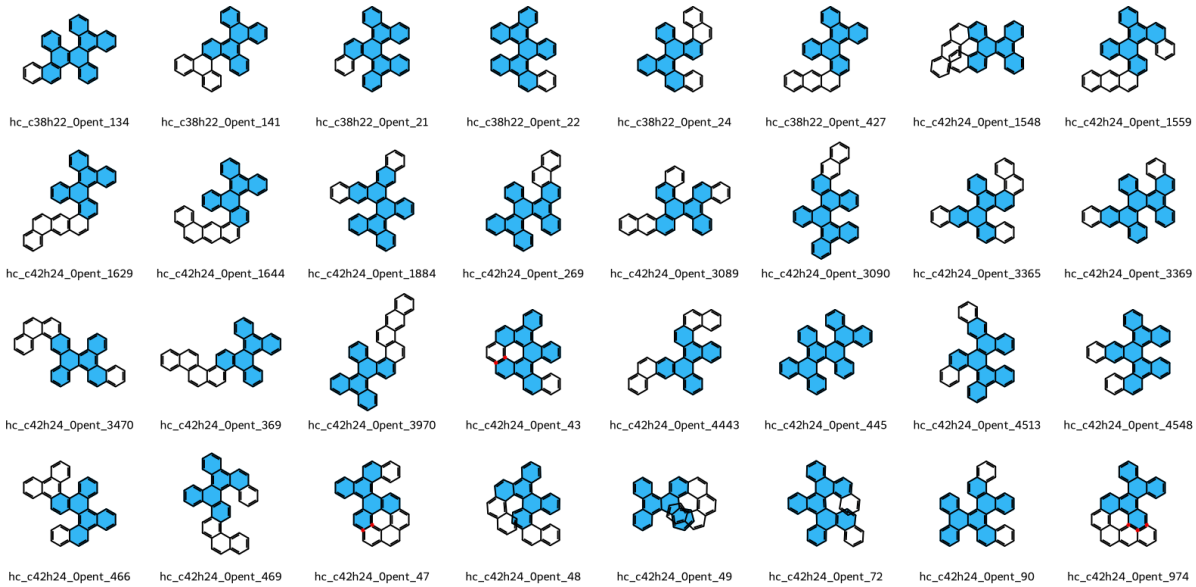

Figure S18. Molecules found in the "island" in the xTB vs DFT scatter plot of aEA of the COMPAS-1 dataset. The largest common motif is dibenzo[*g,p*]chrysene, which can be described as two 'cove' substructures linked at the center, and is highlighted in blue.

As both COMPAS-1D and COMPAS-3D contain up to 10 rings, molecules in COMPAS-1D are more likely to contain these cove/helical motifs as they do not have the restrictions that comes with *peri*-condensation.

To try to understand why these specific molecules where found in the outlying "island", we investigated the differences between the anionic and neutral forms of these molecules. We compared the geometries of both forms by calculating the RMSD between them for both xTB- and DFT (see Figure S19**B**&**D**). We found that in the case of the DFT-optimized geometries, the anionic form of the molecules clearly deviates from its neutral form (Figure S19**D**), indicating a change in geometry due to the need to stabilize excess charge. Surprisingly, the same difference between the anionic and neutral forms was not observed in xTB-optimized geometries (Figure S19**B**). This could imply that xTB underestimates the destabilization incurred by the added charge, or that DFT overestimates it. For

control, we also compared the geometries of the cationic and neutral forms (Figure S19**C**) and do not observe major changes in geometry, for either of the computational methods, confirming that the difference in geometries is specific to the anionic forms.

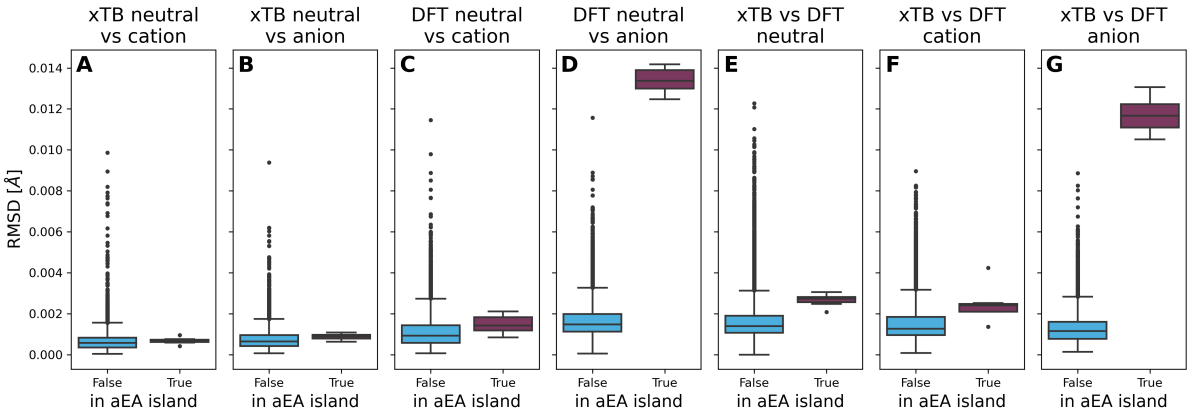

Figure S19. Boxplots of RMSD of xTB-optimized geometries of the neutral versus cationic (**A**), and anionic (**B**) forms; of DFT-optimized geometries of the neutral versus cationic (**C**), and anionic (**D**) forms; and of xTB- versus DFT-optimized geometries of the neutral (**E**), cationic (**F**), and anionic (**G**) species. For each subplot, boxplot of the complete dataset up to 10 rings without the 7 outliers is on the left (False), and the boxplot of the 7 outliers is on the left (True).

To better pinpoint the source of the geometric discrepancy, and because these molecules all contain non-planar motifs, we focused on the agreement for Δz between xTB- and DFT-optimized geometries. We found that in the case of the neutral and cationic forms, there is a good agreement between Δz(xTB) and Δz(DFT) (see Table S7 and Figure S20). On the other hand, the Δz of the anionic forms strongly disagrees between the two methods. Therefore, it appears that the disagreement between xTB- and DFT-calculated aEA values of the Group 2 outliers can be traced back to different extents of non-planarity in the optimized geometries of the anionic species (in contrast to Group 1 outliers). It is possible that the inclusion of diffuse functions might increase the accuracy of the calculation of these molecules. Nevertheless, we reiterate that this issue affects only a small fraction of the dataset. The majority of molecules follows a very good linear fit.

Table S7. Deviation from planarity (Δz) in Å, calculated for xTB-optimized and DFT-optimized geometries. $\Delta\Delta z$ indicates the difference between DFT and xTB. $\Delta\Delta z = \Delta z(\text{DFT}) - \Delta z(\text{xTB})$.

Molecule ID	Anionic form			Cation form			Neutral form		
	xTB	DFT	$\Delta\Delta z$	xTB	DFT	$\Delta\Delta z$	xTB	DFT	$\Delta\Delta z$
hc_c36h20_0pent_125	3.39	4.72	1.32	3.38	3.60	0.21	3.35	3.64	0.29
hc_c40h22_0pent_3389	3.46	4.93	1.47	3.43	3.55	0.12	3.42	3.68	0.26
hc_c40h22_0pent_413	3.93	5.13	1.20	3.85	3.97	0.12	3.83	4.00	0.17
hc_c40h22_0pent_584	3.67	4.44	0.77	3.60	3.81	0.20	3.53	3.74	0.22
hc_c40h22_0pent_587	4.26	6.10	1.84	4.26	4.76	0.50	4.19	4.45	0.26
hc_c40h22_0pent_654	3.58	5.51	1.93	3.56	3.78	0.22	3.50	3.73	0.23
hc_c40h22_0pent_750	3.42	5.13	1.71	3.44	3.51	0.07	3.35	3.58	0.22

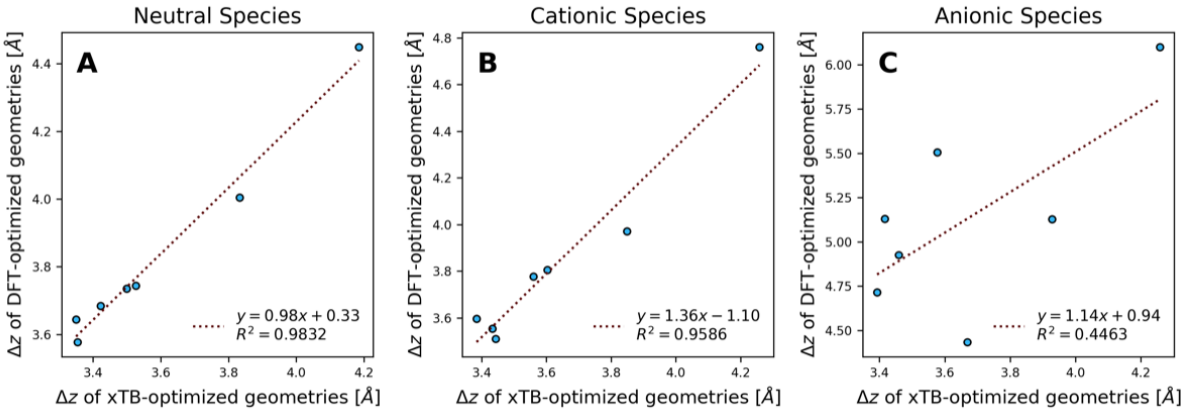

Figure S20. Scatter plots of Δz from the DFT-optimized geometries versus the xTB-optimized geometries of the neutral (left), cationic (middle), and anionic (right) forms.

S5.3 Additional analysis of the role of dispersion corrections

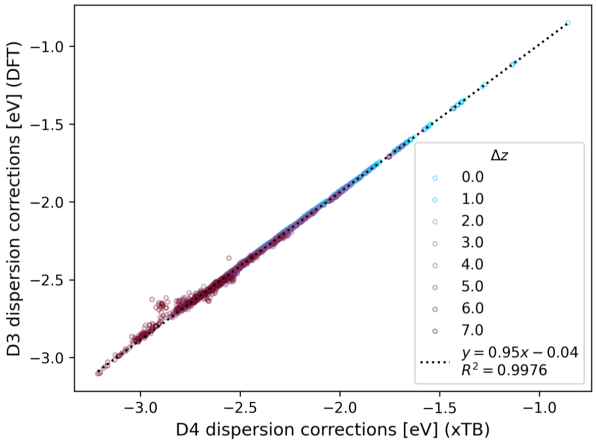

Figure S21. D3 dispersion corrections (DFT) versus D4 dispersion corrections (xTB) colored by rounded Δz obtained from DFT-optimized geometries. Values are in eV.

S5.4 Additional analysis on the relative energy

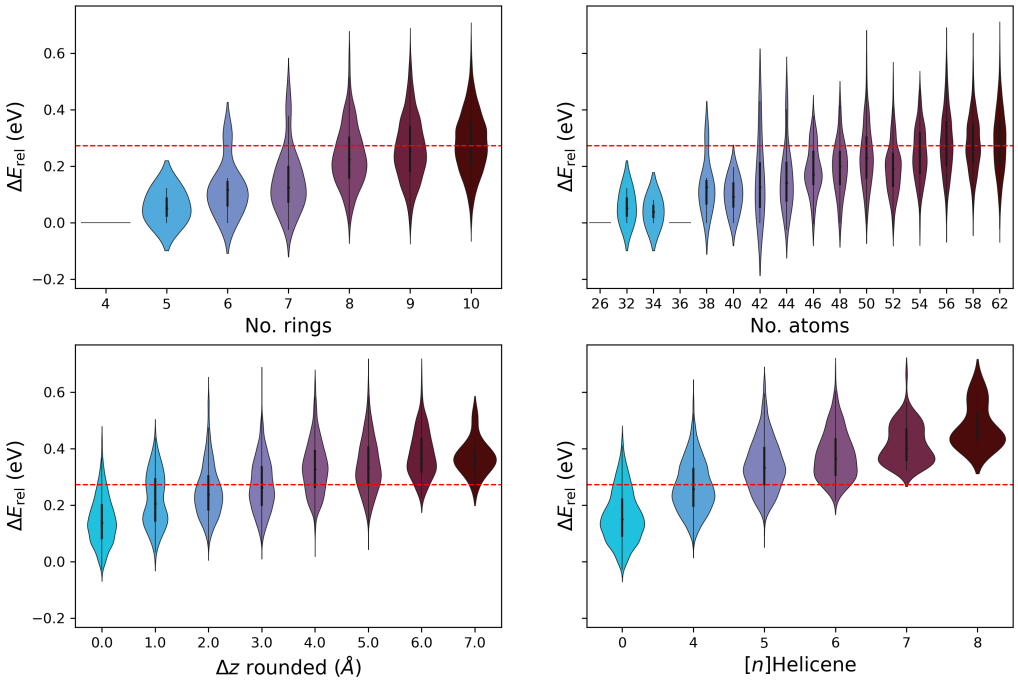

Figure S22. Violin plots of the difference in relative energy between xTB and DFT (ΔE_{rel}) in function of the number of rings (upper left), total number of atoms (upper right), rounded values of Δz—a measure of the deviation from planarity—(lower left), and $[n]$Helicene—where n stand for the number of rings forming the helical structure of helicene—(lower right). The red line indicates the median value of the distribution of ΔE_{rel}.

S6 Additional figures for trends analysis

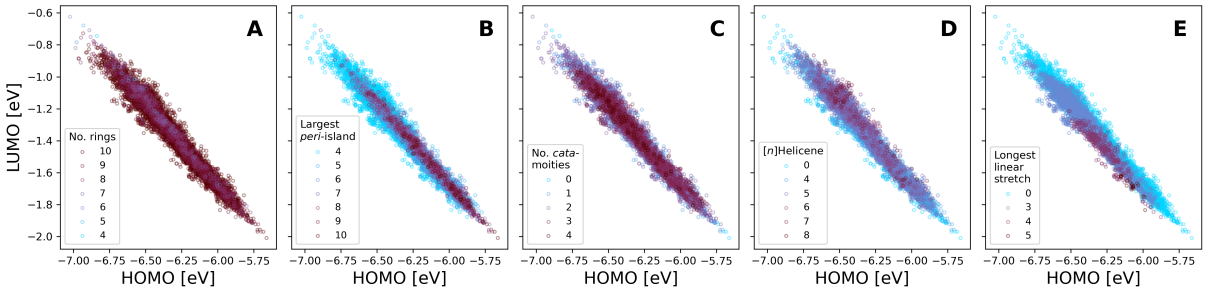

Figure S23. Scatter plots of the DFT-calculated values of the LUMO versus HOMO, colored by: A) number of rings, B) number of rings in the largest *peri*-island, C) number of *cata*-moieties, D) longest contained [*n*]Helicene, and E) longest stretch of linearly annulated rings.

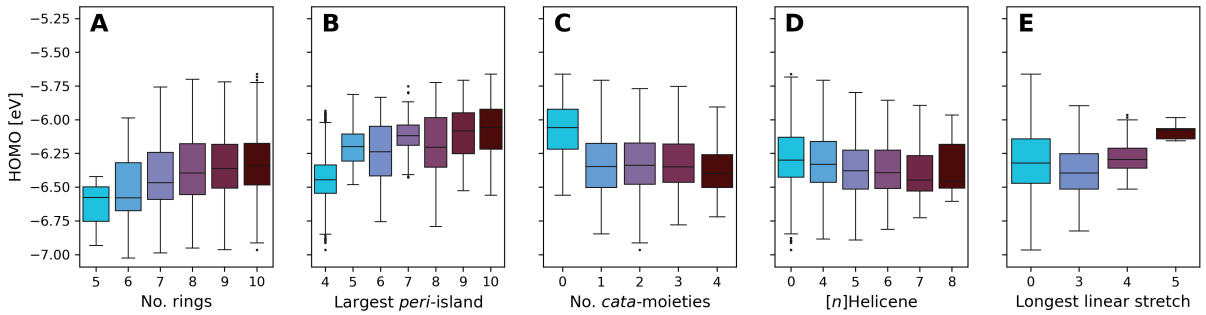

Figure S24. Box plots of the DFT-calculated values of the HOMO, colored by: A) number of rings, B) number of rings in the largest *peri*-island, C) number of *cata*-moieties, D) longest contained [*n*]Helicene, and E) longest stretch of linearly annulated rings. Plot A presents the data from all molecules in families 5–10. Plots B–E present data from family 10 only.

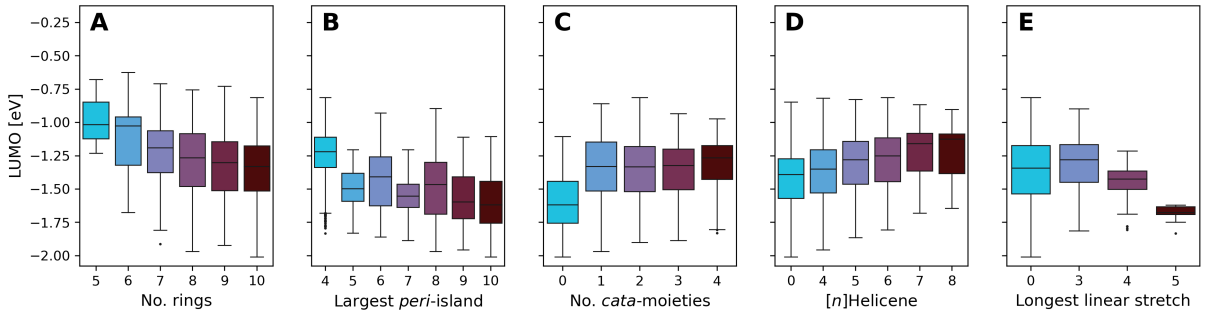

Figure S25. Box plots of the DFT-calculated values of the LUMO, colored by: A) number of rings, B) number of rings in the largest *peri*-island, C) number of *cata*-moieties, D) longest contained [*n*]Helicene, and E) longest stretch of linearly annulated rings. Plot A presents the data from all molecules in families 5–10. Plots B–E present data from family 10 only.

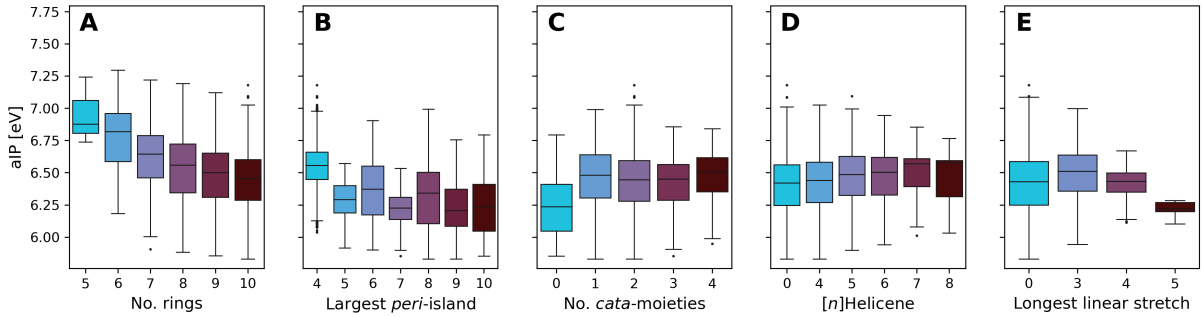

Figure S26. Box plots of the DFT-calculated values of the aIP, colored by: A) number of rings, B) number of rings in the largest *peri*-island, C) number of *cata*-moieties, D) longest contained [*n*]Helicene, and E) longest stretch of linearly annulated rings. Plot A presents the data from all molecules in families 5–10. Plots B–E present data from family 10 only.

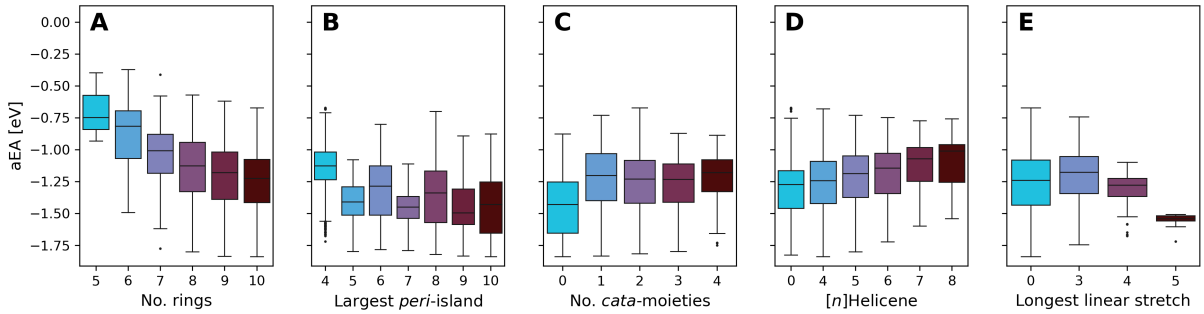

Figure S27. Box plots of the DFT-calculated values of the aEA, colored by: A) number of rings, B) number of rings in the largest *peri*-island, C) number of *cata*-moieties, D) longest contained [*n*]Helicene, and E) longest stretch of linearly annulated rings. Plot A presents the data from all molecules in families 5–10. Plots B–E present data from family 10 only.

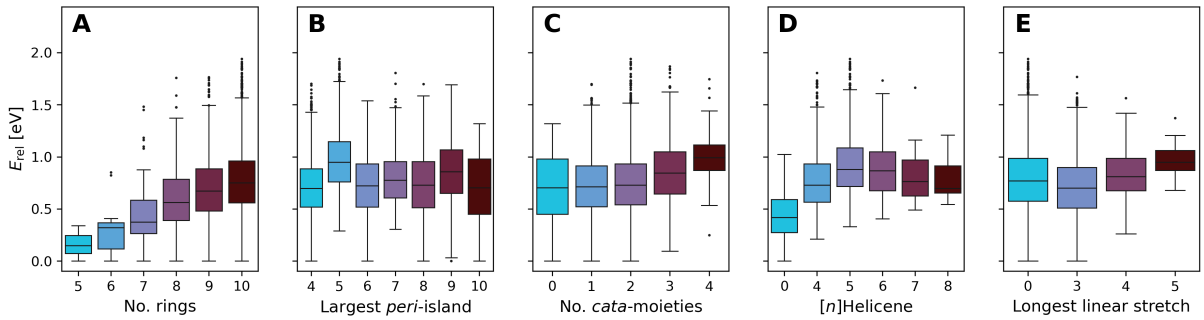

Figure S28. Box plots of the DFT-calculated values of the E_{rel}, colored by: A) number of rings, B) number of rings in the largest *peri*-island, C) number of *cata*-moieties, D) longest contained [*n*]Helicene, and E) longest stretch of linearly annulated rings. Plot A presents the data from all molecules in families 5–10. Plots B–E present data from family 10 only.

The following tables contain the number of molecules plotted in each boxplot.

Table S8. Number of molecules in boxplots separated by number of rings.

No. rings	**No. molecules**
5	3
6	15
7	68
8	329
9	1537
10	6891

Table S9. Number of molecules in boxplots separated by largest *peri*-island size. Only for molecules with 10 rings.

Largest *peri*-island	**No. molecules**
4	3728
5	1273
6	1047
7	309
8	383
9	111
10	40

Table S10. Number of molecules in boxplots separated by number of *cata*-moieties. Only for molecules with 10 rings.

No. *cata*-moieties	**No. molecules**
0	40
1	1746
2	3480
3	1482
4	143

Table S11. Number of molecules in boxplots separated by largest [n]Helicene. Only for molecules with 10 rings.

[n]Helicene	**No. molecules**
0	883
4	3600
5	1943
6	377
7	78
8	10

Table S12. Number of molecules in boxplots separated by the longest linear stretch. Only for molecules with 10 rings.

Longest linear stretch	**No. molecules**
0	4417
3	2155
4	301
5	18

References

(S1) Bannwarth, C.; Ehlert, S.; Grimme, S. *J. Chem. Theory Comput.* **2019**, *15*, 1652–1671.

(S2) Neese, F. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2012**, *2*, 73–78.

(S3) Neese, F. *Wiley Interdiscip. Rev. Comput. Mol.* **2022**, *12*, e1606.

(S4) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. *J. Chem. Phys* **2010**, *132*, 154104.

(S5) Johnson, E. R.; Becke, A. D. *J. Chem. Phys* **2005**, *123*, 024101.

(S6) Grimme, S.; Ehrlich, S.; Goerigk, L. *J Comput Chem* **2011**, *32*, 1456–1465.

(S7) Wahab, A.; Pfuderer, L.; Paenurk, E.; Gershoni-Poranne, R. *J. Chem. Inf. Model.* **2022**, *62*, 3704–3713.

(S8) Brinkmann, G.; Friedrichs, O. D.; Lisken, S.; Peeters, A. *Commun. Math. Comput. Chem.* **2009**, *63*, 533–552.

(S9) Becke, A. D. *J. Chem. Phys* **1993**, *98*, 5648–5652.

(S10) Lee, C.; Yang, W.; Parr, R. G. *Phys. Rev. B* **1988**, *37*, 785–789.

(S11) Miehlich, B.; Savin, A.; Stoll, H.; Preuss, H. *Chem. Phys. Lett.* **1989**, *157*, 200–206.

(S12) Hertwig, R. H.; Koch, W. *Chem. Phys. Lett.* **1997**, *268*, 345–351.

(S13) Yanai, T.; Tew, D. P.; Handy, N. C. *Chem. Phys. Lett.* **2004**, *393*, 51–57.

(S14) Weigend, F.; Ahlrichs, R. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297.

(S15) Dunning, T. H. *J. Chem. Phys.* **1989**, *90*, 1007–1023.

(S16) Kendall, R. A.; Dunning, T. H.; Harrison, R. J. *J. Chem. Phys.* **1992**, *96*, 6796–6806.

(S17) Woon, D. E.; Dunning, T. H. *J. Chem. Phys.* **1993**, *98*, 1358–1371.

(S18) Liang, J.; Feng, X.; Hait, D.; Head-Gordon, M. *J. Chem. Theory Comput.* **2022**, *18*, 3460–3473.

(S19) Woon, D. E.; Dunning Jr., T. H. *J. Chem. Phys.* **1994**, *100*, 2975–2988.

(S20) Rappoport, D.; Furche, F. *J. Chem. Phys.* **2010**, *133*, 134105.

(S21) Modelli, A.; Mussoni, L. *Chem. Phys.* **2007**, *332*, 367–374.

(S22) Khatymov, R. V.; Muftakhov, M. V.; Shchukin, P. V. *Rapid Commun. Mass Spectrom.* **2017**, *31*, 1729–1741.

(S23) Fite, S.; Wahab, A.; Paenurk, E.; Gross, Z.; Gershoni-Poranne, R. *Journal of Physical Organic Chemistry* **2023**, *36*, e4458.