The COMPAS Project: A Computational Database of Polycyclic Aromatic Systems. Phase 1: *cata*-condensed Polybenzenoid Hydrocarbons

Alexandra Wahab,[†] Lara Pfuderer,[†] Eno Paenurk,[‡] and Renana Gershoni-Poranne[*,¶,†]

[†]*Laboratory for Organic Chemistry, Department of Chemistry and Applied Biosciences, ETH Zurich, 8093 Zurich, Switzerland*
[‡]*Laboratory for Organic Chemistry, Department of Chemistry nd Applied Biosciences, ETH Zurich, 8093 Zurich, Switzerland*
[¶]*Schulich Faculty of Chemistry, Technion - Israel Institute of Technology, Haifa 32000, Israel - Israel Institute of Technology*

E-mail: rporanne@technion.ac.il

Abstract

Chemical databases are an essential tool for data-driven investigation of structure-property relationships and design of novel functional compounds. We introduce the first phase of the COMPAS Project – a COMputational database of Polycyclic Aromatic Systems. In this phase, we have developed two datasets containing the optimized ground-state structures and a selection of molecular properties of ∼34k and ∼9k *cata*-condensed polybenzenoid hydrocarbons (at the GFN2-xTB and B3LYP-D3BJ/def2-SVP levels, respectively), and have placed them in the public domain. Herein we describe the process of the dataset generation, detail the information available within the datasets, and show the fundamental features of the generated data. We analyze the correlation between the two types of computation as well as the structure-property relationships of the calculated species. The data and the insights gained from them can inform rational design of novel functional aromatic molecules for use in, e.g., organic electronics, and can provide a basis for additional data-driven machine- and deep-learning studies in chemistry.

Introduction

Polycyclic aromatic systems (PASs) – molecules comprising multiple aromatic rings – are one of the most prevalent classes of compounds in both nature and man-made materials. They are important in many fields of chemistry, but their popularity in recent decades is largely due to their performance as the semiconducting components in organic electronics.[1,2] They are uniquely suited to this role, as their characteristic π-conjugation enables high conductance[1,3] and their rigid structure enables close packing,[4,5] which allows for good charge mobility. Moreover, their specific electronic and physical properties (e.g., band-gap, solubility) can be tuned through changes in annulation or substitution with functional groups.[6–10] In addition, PASs can function as platforms for electrocatalysis,[11] redox-active materials,[12] and organic electrode materials.[13,14]

The development of improved organic electronic devices, such as light emitting diodes (OLEDs),[15] field effect transistors (OFETs),[5,16,17] solar cells (OSCs),[2,8] sensors,[18] and semiconductors[10,16,17,19] hinges on the design of new functional PASs. With the decreasing cost of computational resources and the concurrent advent of machine-learning (ML) and deep-learning (DL) techniques, data-driven design of molecules and materials for various uses has become an increasingly promising approach. Such tools may efficiently map the chemical space and allow discovery of new molecular motifs. However, they require suitable databases, and though many new chemical databases have been constructed and curated in recent years, a dedicated PAS database is not available, to the best of our knowledge.

To date, the largest structured PAS-dedicated database is the NIST Polycyclic Aromatic Hydrocarbon (PAH) database,[20] which was first published in 1997 and later revised and corrected in 2020.[21] The NIST PAH database houses 660 PAHs – molecules comprising only carbon and hydrogen – and their various experimentally and computationally obtained properties. Notably, not all properties are provided for every entry in the database, which makes the database sparse and less suitable for ML and DL applications. Recently, Alvarez-Ramírez and Ruiz-Morales[22] used the enumerated structures of this database to generate the FAR-database, in which they provide several types of nucleus-independent chemical shift (NICS)[23–25] values for each molecule (namely, NICS(0), NICS(1), NICS(0)$_{ZZ}$, and NICS(1)$_{ZZ}$). The recent revision of the database, as well as the data expansion of Alvarez-Ramírez and Ruiz-Morales, highlight both the ongoing interest in these molecules and the need for relevant data.

For the past number of years, our group has been studying PASs with the aim of obtaining a deeper understanding of their properties and, specifically, how the global and local molecular properties map to individual structural features. This knowledge can inform the design of novel materials with enhanced features and improved functionality. Our research goals, combined with the community's need of a better PAS database of those compounds, led us to embark on the COMPAS (COMputational database of Polycyclic Aromatic Systems) Project (`https://gitlab.com/porannegroup/compas`). The COMPAS Project undertakes the construction of a curated, computationally-generated, freely-accessible database of PAS structures and properties. To methodically and effectively tackle the challenge of mapping the large and diverse chemical space of PASs, we have divided the project into phases, according to subclasses of PASs.

Herein, we report on Phase 1 of the COMPAS project, which focuses on the subclass of *cata*-condensed polybenzenoid hydrocarbons (PBHs, sometimes also referred to as polycyclic aromatic hydrocarbons, PAHs, or catafusenes) in the ground state. For this subclass of compounds we have generated two computational datasets: (1) COMPAS-1D – 8,678 *cata*-condensed PBHs comprising 1–10 rings, calculated with density functional theory (DFT) at the B3LYP-D3BJ/def2-SVP level of theory; (2) COMPAS-1x – 34,074 *cata*-condensed PBHs comprising 1–11 rings, calculated with xTB using GFN2-xTB. In this manuscript, we detail the technical aspects of constructing this new database and describe the main features of the generated data. We glean insight into the structure-property relationships of the *cata*-condensed PBHs and delineate future directions for investigation.

Data Generation Workflow

The first phase of the COMPAS project focuses on the family of *cata*-condensed polybenzenoid hydrocarbons (PBHs), which comprise only benzene rings (also known as PAHs or catafusenes). Several examples of such molecules are depicted in Figure 1. The flowchart in Figure 2 illustrates the steps taken to generate these datasets. In the following sections, we detail and rationalize the individual steps.

Step 1. Structure Enumeration

The chemical space of *cata*-condensed PBHs containing up to 11 rings was fully enumerated (see Table 1) with the CaGe (the Chemical &

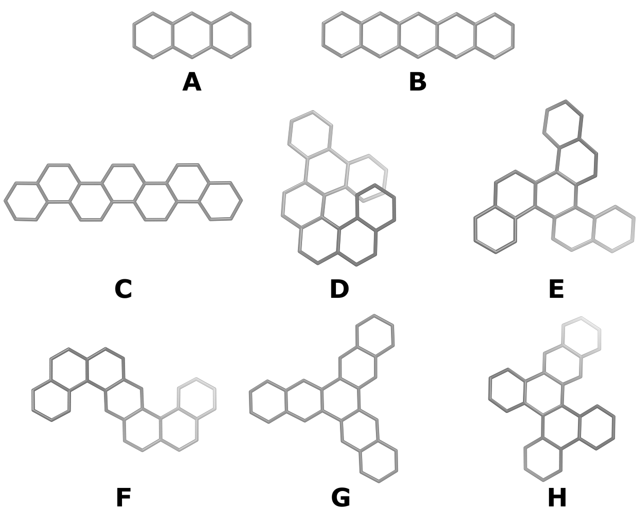

Figure 1: Examples of *cata*-condensed polycyclic benzenoid hydrocarbons from the COMPAS database. Double bonds and hydrogens are omitted for clarity.

abstract Graph environment)[26] software, which is an open source package for generating mathematical and chemical graphs. Using CaGe, we obtained initial (unoptimized) *xyz*-coordinates of all possible 36,043 molecules in this chemical space (Figure 2, step 1).

Table 1: Overview of the data set.

No. Rings	Molecular Formula	No. Isomers	No. Valid
1	C_6H_6	1	1
2	$C_{10}H_8$	1	1
3	$C_{14}H_{10}$	2	2
4	$C_{18}H_{12}$	5	5
5	$C_{22}H_{14}$	12	12
6	$C_{26}H_{16}$	37	37
7	$C_{30}H_{18}$	123	121
8	$C_{34}H_{20}$	446	440
9	$C_{38}H_{22}$	1,689	1,651
10	$C_{42}H_{24}$	6,693	6,408
11	$C_{46}H_{26}$	27,034	25,394
		36,043	**34,072**

Step 2. xTB Optimization

The *xyz*-coordinates obtained from CaGe for the 36,043 molecules enumerated were optimized with xTB[27] version 6.2, using GFN2-xTB, a tight-binding quantum chemical method to perform fast calculations of molecular geometries at the semi-empirical level. Following optimization of the structures, harmonic vibrational frequencies were calculated to ensure true minima on the potential energy surface (i.e., $N_{\text{imag}} = 0$; Figure 2, step 2). After a subsequent filtration step (see subsequent section, Step 3. Data Filtration), 34,072 molecules were retained. For each molecule, the cationic and anionic forms were optimized with xTB as well (with subsequent frequency calculations). A total of ∼110k species overall were optimized. In summary, 34,072 *cata*-condensed PBH structures containing up to 11 rings make up the dataset denoted as COMPAS-1x (see Table 1).

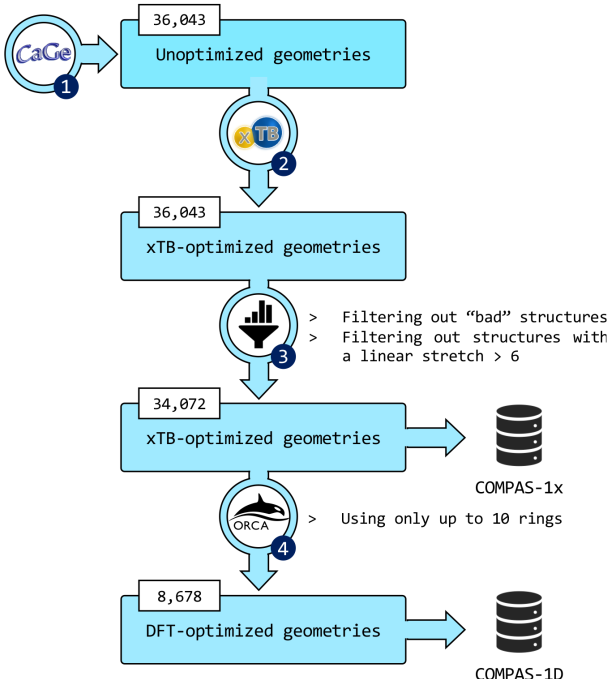

Figure 2: Flowchart of the data-generation process. (1) CaGe[26] was used to generate unoptimized geometries of all *cata*-condensed structures containing up to 11 rings. (2) xTB was used to optimize all geometries. (3) The data were filtered to remove invalid structures or those having more than six rings annulated linearly. The geometries of the remaining molecules and their electronic properties make up the COMPAS-1x dataset (34,072 molecules). (4) DFT was used to further optimize the molecules containing up to 10 rings. The geometries of these 8,678 molecules and their electronic properties make up the COMPAS-1D dataset.

Step 3. Data Filtration

Because the initial coordinates generated by CaGe were sometimes quite far from the optimal geometries (especially for non-planar molecules), some structures did not optimize properly. The main difficulty arose from CaGe placing atoms closer to each other than they should be. Specifically, we identified two types of cases where this led to an incorrectly optimized structure: a) when two carbons were too close in space, they would form a bond, even though they were not supposed to be bonded, resulting in various-sized rings or sp^3-hybridized carbons; b) when a hydrogen atom was too close to two carbons, it might shift during the optimization process to the wrong carbon, leading to the wrong final structure. In some cases, we encountered both of these issues in the same molecule. To check for and filter out the incorrectly-optimized structures, we performed two individual tests. The first test used a modified version of the Predi-XY program[28] to check that all molecules contained only six-membered rings. The second test used the `xyz2mol` script developed by the Jensen group,[29] which converts *xyz*-coordinates to SMILES formats, to check for undesired motifs, such as saturated carbons. The "bad" structures were discarded (Figure 2, step 3). We then filtered the data further by discarding all molecules containing a linear stretch longer than six rings. A "linear stretch" is a substructure within the PBH, in which the consecutive rings are annulated linearly. Examples **A** and **B** in Figure 1 depict linear stretches of three and five rings, respectively (i.e., anthracene and pentacene). Linear stretches longer than six rings are known to have non-negligible open-shell character[30–32] in the ground-state and such molecules are relatively unstable.

Step 4. Further Optimization with DFT

Of the 34,072 retained xTB-optimized structures, 8,678 (all valid structures containing up to 10 rings, Figure 2, step 4) were further optimized with DFT calculations, performed with ORCA version 4.2.0,[33,34] using the B3LYP[35–38] functional and the def2-SVP[39] basis set, with Grimme's D3[40] dispersion correction and the Becke-Johnson damping scheme.[41,42] This level of theory was chosen following a benchmarking procedure (see Supporting Information, Section S1.2, for further details). DFT is considered to be a more accurate computational method than semi-empirical methods such as tight-binding (for both geometries and molecular properties) and, in particular, B3LYP has been shown to perform well with PAHs.[43–45] Nevertheless, it is also a more computationally costly method, which factored into our decision to reduce the dataset at this stage. The main rationale behind reducing the dataset was that the COMPAS Project is aimed at enabling investigations of structure-property relationships in PASs and molecular design of novel PASs. From the data of the 1–10-ring isomers, we could already grasp the important insight needed (see Data Analysis section), without adding the data of the 11-ring isomers. Moreover, the results of the 1–10-ring isomers indicated that there is a good linear correlation between xTB and DFT results (see Data Analysis section). Therefore, if needed, it is possible to obtain close to DFT-level accuracy from xTB-level calculations. Thus, we elected to forgo the more computationally expensive DFT calculations on the large family of 11-ring isomers. The DFT-optimized geometries and properties form the dataset denoted as COMPAS-1D.

Representations and Properties

As mentioned above, to differentiate the two datasets, we denote the dataset containing xTB-optimized structures and xTB-calculated properties as COMPAS-1x, and the dataset containing DFT-optimized structures and DFT-calculated properties as COMPAS-1D.

The properties contained in each of the two datasets are detailed in Table 2, where HOMO and LUMO are the highest occupied and lowest unoccupied molecular orbitals, respectively; SPE is the dispersion-corrected single-point energy (i.e., the energy of the optimized structure without zero-point corrections); SCF energy is the energy of the optimized structure without dispersion correction; ZPE is the zero-point energy; aIP is the adiabatic ionization potential; and aEA is the adiabatic electron affinity.

Table 2: Properties available in the COMPAS-1x and the COMPAS-1D databases, respectively. All energies are reported in eV and the dipole moment is reported in Debye.

Properties	COMPAS-1x	COMPAS-1D
HOMO	✓	✓
LUMO	✓	✓
HOMO-LUMO gap	✓	✓
SPE (neutral)	✓	✓
SPE (cation)	✓	✓
SPE (anion)	✓	✓
Rel. SPE (neutral)	✓	✓
SCF energy (neutral)		✓
ZPE (neutral)	✓	
ZPE (cation)	✓	
ZPE (anion)	✓	
aIP	✓	✓
aEA	✓	✓
Dipole moment	✓	✓

COMPAS-1x contains the ZPEs for the neutral and charged (-1 and $+1$) forms for all 34,072 structures. COMPAS-1D does not contain ZPE data, because we did not perform frequency calculations at the DFT level. The xTB-calculated ZPE values can be used to correct the aIP, the aEA, and the relative energy for both the xTB and the DFT calculated properties if desired (ZPE corrections are not highly method-dependent,[46] thus can often be used across methods). The relative single-point energy (only for the neutral forms) was obtained by calculating the difference in single-point energy between each molecule and its lowest-energy isomer. Accordingly, for each molecular formula, the lowest value is zero, with all isomers having positive relative energy with respect to the reference isomer.

In addition to these properties, we include for each molecule three types of identifiers/representations: a) a given name that includes its molecular formula and a serial number; b) its SMILES representation;[47,48] and c) its annulation sequence.[49]

Data Analysis

The newly-constructed COMPAS-1D and COMPAS-1x are the first datasets of their type, to the best of our knowledge, and offer a unique opportunity to probe the properties of the *cata*-condensed PBHs and their distribution within the chemical space. In this section we provide an overview of the data and discuss some of the trends that are uncovered.

Agreement between xTB and DFT

Figure 3 shows the distributions of the various properties, computed with both xTB and DFT (note: to enable comparison between the two methods, only the data for the PBHs containing up to 10 rings are displayed). At first glance, the plots in Figures 3**A**-**E** appear to show seemingly consistent offsets between the distributions of the two methods, with the xTB values always being more negative than the DFT ones. For the HOMO level, this offset is ca. 4 eV; for the LUMO level it is ca. 6 eV; and, accordingly, for the HOMO-LUMO gap it is ca. 2 eV. Similarly, there is an offset seen in Figures 3**D**-**E**, with DFT giving ca. 5 eV lower aIP values and ca. 5 eV higher aEA values, which is in accordance with the HOMO and LUMO energies (as expected from analogy to Koopmans' theorem and its DFT counterpart[50,51]). Despite these shifts, the distribution shapes seem to be quite similar between the two datasets for these five properties. Plotting the individual data points of the xTB calculations versus the DFT calculations (Figure 4) shows that the results of the two methods are linearly correlated, confirming that, though the two methods do not quantitatively agree, the trends are similar. Thus, it is possible to obtain chemical insight relating to trends in the data from each of them. This also demonstrates that it is possible to obtain properties at DFT-level accuracy for PBHs from the much faster and less expensive xTB calculations, using a correction/fitting scheme. However, we note that for each property, the slope is not equal to 1, which means that the difference is not simply a constant offset. Rather, the difference between the methods increases with the

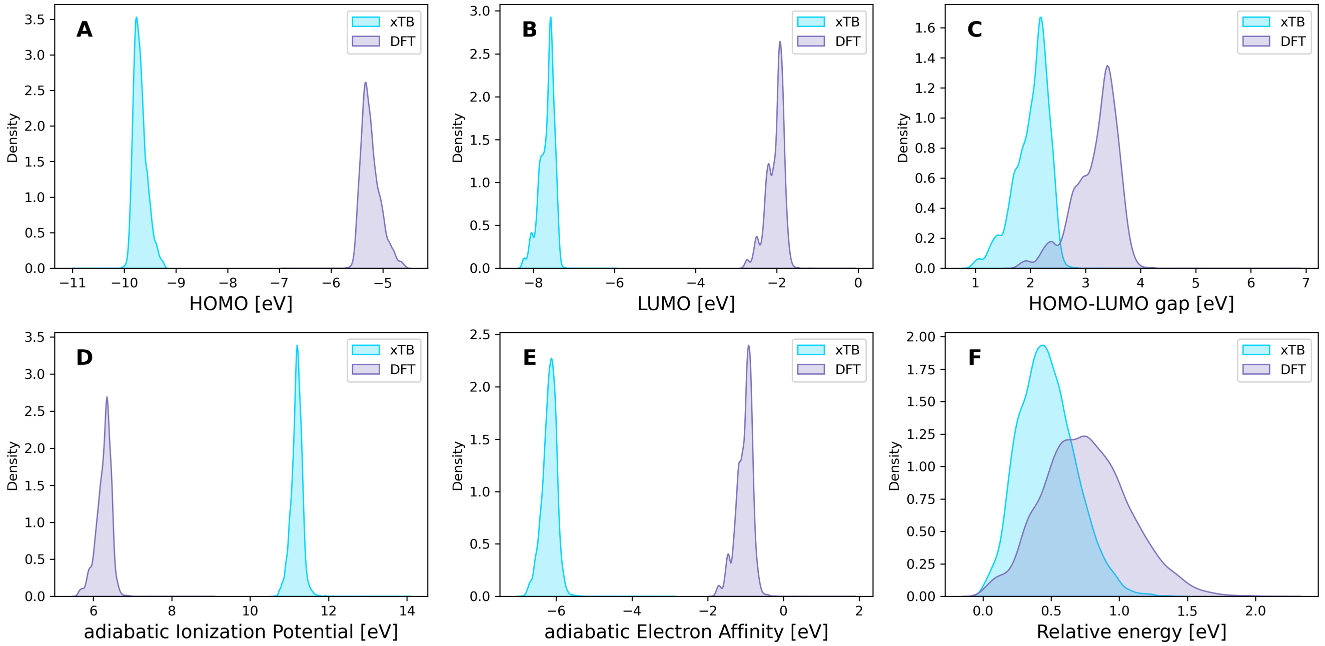

Figure 3: KDE plots of the distribution of xTB-calculated properties vs DFT-calculated properties: **A)** HOMO; **B)** LUMO; **C)** HOMO-LUMO gap; **D)** adiabatic ionization potential; **E)** adiabatic electron affinity; **F)** relative single-point energy. All values are reported in eV.

value of the property. We also note that, unsurprisingly, the single data point that appears separate from the rest of the data is benzene.

We observe that the aEA plot has a small "island" of data points that are not on the main correlation line. A closer inspection of this area reveals that it contains ca. 30 data points. Plotting the aEA against the LUMO for both the xTB and DFT data shows that only the DFT plot has this "island", which suggests that the deviation is attributed to the DFT calculation (Supporting Information, Figure S4). The well-behaved data of the DFT-calculated LUMO versus the HOMO (Figure 7) indicate that the neutral forms are treated well and the deviation likely stems from the DFT calculation of the anionic forms of the molecules. The molecules contained in this "island" have certain structural similarities: they all comprise long consecutive angularly annulated stretches and multiple branching points (Supporting Information, Figures S7**E** and S8**E**). The combination of these structural features often leads to curvature and deviation from planarity, and it is possible that our chosen functional/basis set combination is less appropriate for such cases. In particular, for anionic species, it is considered important to include diffuse functions in the basis set, to allow for better treatment of the electron delocalization. Though previous reports have indicated that addition of diffuse functions does not necessarily lead to more accurate geometries and energies for planar PAHs,[43,52] it is not clear how this affects non-planar systems. At the same time, considering the small number of molecules (ca. 0.3%) that appear to be affected, we believe the choice of a more cost-effective basis set is justified.

The only property which shows a noticeable difference between the distributions of two types of calculations is the relative energy (Figure 3**F**). Because the relative energy is calculated as the difference between the SPE of each molecule and the SPE of its lowest-energy isomer (similar to a homodesmotic equation), some of the method-dependent variances are expected to cancel out. Yet, xTB shows a distribution ranging between 0 and 1.4 eV and DFT shows a distribution ranging between 0 and 2.2 eV. This is not surprising, as the two methods are inherently different. Despite these differences, the scatter plot (Figure 4**F**) shows

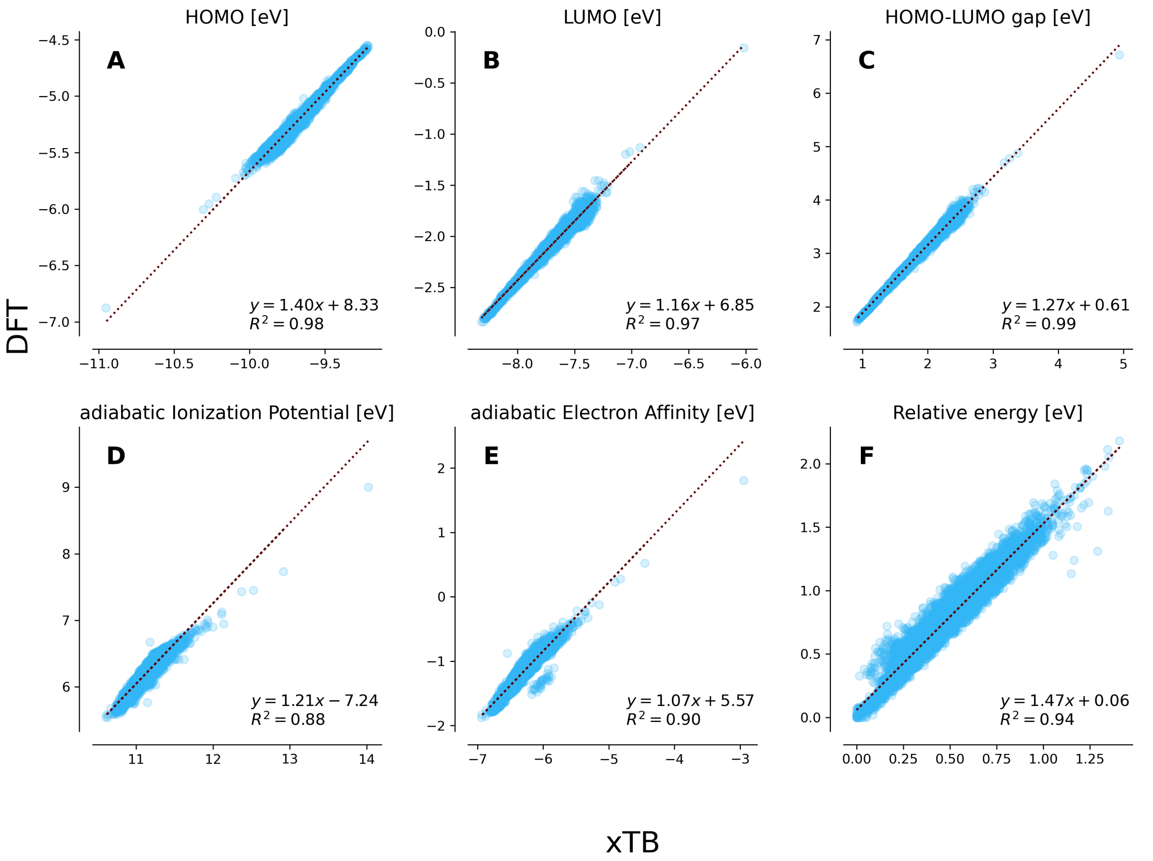

Figure 4: Scatter plots of the various molecular properties, calculated with DFT versus calculated with xTB: **A)** HOMO; **B)** LUMO; **C)** HOMO-LUMO gap; **D)** adiabatic ionization potential; **E)** adiabatic electron affinity; **F)** relative single-point energy. All values reported in eV.

a good agreement between the two methods ($R^2 = 0.94$), with a slope of ca. 1.5. We initially suspected that the quantitative difference between the results might stem from the two methods accounting differently for dispersion interactions: the xTB calculations were performed with the D4[53] correction whereas the DFT calculations were performed with the D3[40] correction (with Becke-Johnson damping). To further investigate this, we studied the correlations of the the self-consistent field (SCF) energies and dispersion corrections separately (Section 2.2 in the Supporting Information). Indeed, the plot of the D3 corrections versus the D4 corrections revealed a size-dependent relationship, whereby there is a slightly different linear correlation for each family of isomers, which could lead to variance in the relative energy (Figure S10). In addition, we found a slightly better agreement between the DFT-calculated and xTB-calculated SCF relative energies (i.e., when the dispersion correction was omitted, Figure 5). However, the overall trend of a slope > 1 remained, which stems from the intrinsic differences between the two methods.

We then hypothesized that the quantitative differences may be attributed be the geometric deformation from planarity of the PBHs. While many of the PBH structures are planar (or nearly planar), a large number are curved or contain curved regions, which introduce torsional and/or helical strain (e.g., structures **D**, **E**, **F**, and **H** in Figure 1). We surmised that the two computational methods may account for this strain differently, resulting in their lack of agreement. To validate this hypothesis, we colored the individual data points according to the deviation along the z-axis (i.e., Δz, calculated as the largest difference between z-coordinates after placing the DFT-optimized molecules in the xy-plane). Both plots in Figure 5 show a stratification of the data, whereby there are distinct individual linear correlations corresponding to the different extents of deviation from planarity (Δz). This indicates that, independent from the dispersion effects, the deviation from planarity plays a role in the two method affording different values of relative energy.

The link to deviation from planarity led us to question whether the two methods might have significant differences in the optimized structures of the curved PBHs. However, we found that the methods agree rather well on the Δz values (Figure S13 in the Supporting Information shows a good linear correlation between the Δz values of the xTB-optimized structures and the DTF-optimized structures, with a slope of 1.03). This implies that the optimized geometry is not significantly method-dependent. In addition, as mentioned above, the excellent agreements observed in the other five plots indicate that the other electronic properties are treated quite well. Yet, there is a clear difference in calculation of the energies, which is related to non-planarity, as seen from Figure 5. Thus, it is possible that this discrepancy stems from the GFN2-xTB and DFT methods treating curved aromatic systems differently; in particular, we believe it is caused by the torsional (helical) strain being treated differently by these methods.

Trends within the data

The distribution plots in Figures 3**A**-**E** indicate the likelihood of locating a PBH molecule with a property in a specific range, but do not provide any information on the connection between structural features and the location of a given molecule within the distribution. To probe the structure-property relationships further, we first divided the dataset into "families", where each family contains all of the isomers of the same number of rings (e.g., Family 5 is the set of molecules containing five rings, and so forth). For *cata*-condensed PBHs, all molecules with the same number of rings also have the same molecular formula, therefore each family can also be defined by a unique molecular formula (see Table 1 for the molecular formula corresponding to each family). We then plotted the kernel-density estimate distributions (KDEs) of the DFT-calculated properties for Families 5–10 (Families 1–4 were omitted because they contain too few molecules to form meaningful distributions), with each individual family indicated by a different color (Figure 6). We caution that the KDE visualization tends to exaggerate the range of distributions, i.e., the extremities of the density plots are not entirely accurate. This can be seen clearly in the plot of the relative energy (Figure 6**F**) where the values are all positive, but the KDE visualization shows distributions spreading into the negative

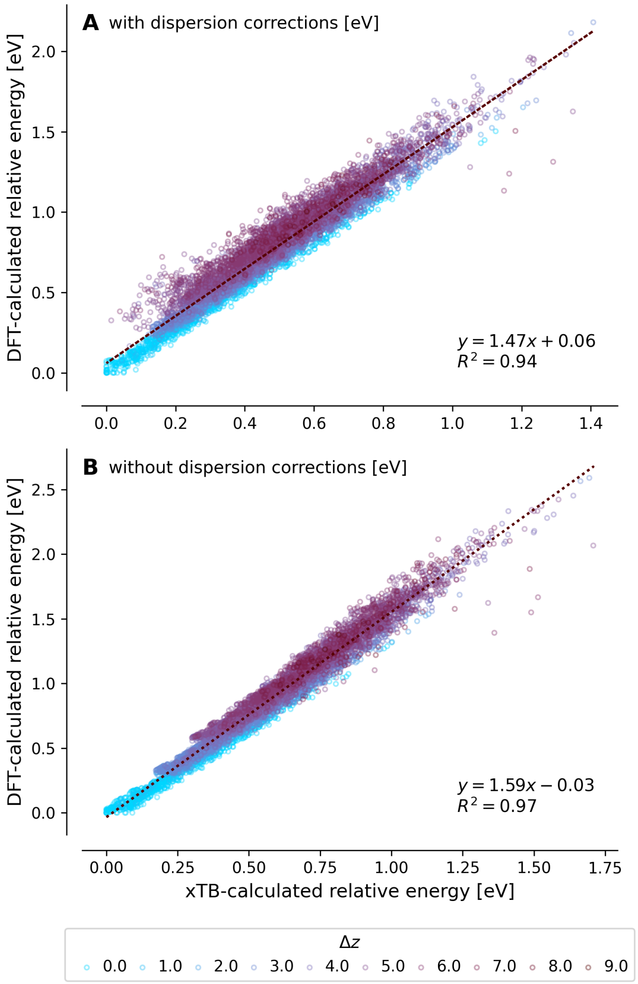

Figure 5: Scatter plot of the DFT-calculated vs the xTB-calculated relative energy with (**A**) and without (**B**) the dispersion corrections in eV, colored based on the Δz values.

region. The exact histograms for all families and all properties are provided in the Supporting Information (sections 2.3 and 2.4).

It can be seen from all plots that, though there are overlaps between the families, there also clear trends based on the number of rings contained in the system. We observe that the median HOMO level becomes higher and the median LUMO level becomes lower as the number of rings grows. This is in line with the known trend of increased conjugation "squeezing" the frontier molecular orbitals together. Accordingly, the average HOMO-LUMO gap becomes smaller as the number of rings increases. The aIP decreases and the aEA increases (in absolute value) with the number of rings. Both of these trends are in accordance with the respective behaviors of the HOMO and LUMO levels (which, within Koopmans' theorem, give a good approximation of the aIP and aEA). In addition, the aEA is known to be size-dependent,[54] as larger molecules stabilize the negative charge more efficiently through delocalization. However, we note at the same time the similarities in range for all the families. One might expect that larger structures, which have more structural diversity, might also display a wider range of properties. Yet, the range of property values does not change significantly with the increase in molecular size, starting from Family 5 (mean, median, minimum, and maximum values of the distributions for each family are shown in the Supporting Information, section 2.5). This suggests that structural diversity, in and of itself, does not lead to substantial changes in the property values. This observation can be rationalized by conclusions from previous work from our group. As part of our ongoing research into structure-property relationships of PASs, we recently delineated several guidelines for molecular design of triplet-state PBHs, based on their decomposition into series of tricyclic subcomponents.[49] We observed that a specific structural motif – the longest linear stretch – determines several molecular properties, including the singlet-triplet energy gap (which is dependent on the HOMO and LUMO energies) and the location of spin density in the triplet state (note: the longest linear stretch refers to the longest series of consecutive laterally annulated benzene rings in the molecule). In other words, if the longest linear substructure determines the molecular properties, regardless of the overall molecular size, this can explain how molecules of varying sizes have similar property values (overlapping distributions). It also explains the plateau-like behavior observed starting at Family 6 in the minimum values of the LUMO, HOMO-LUMO Gap, aEA, and aIP and in the maximum value of the HOMO (section 2.5 in the Supporting Information). Because all linear stretches longer than six rings were excluded from the datasets, the longest linear stretch in all of the Families 6–10 is a six-ring stretch. Accordingly, they all have similar min/max values.

Finally, we see that the maximum relative energy increases with the number of rings. Our understanding of this trend is that, the larger the molecules are, the more opportunities there are for introducing destabilizing effects. Namely, we anticipate that helical structures or multiple curved structures can begin to emerge as the number of rings increases, which generate torsional/helical strain.

To further investigate the effects of structural motifs on the molecular properties of most interest to us, we plotted the HOMO versus the LUMO values and colored the individual data-points according to various structural motifs. The motifs we selected for visualization were: a) the number of rings; b) longest linear stretch, which is the longest consecutive sequence of linearly annulated rings; c) longest angular stretch, which is the longest consecutive sequence of angularly annulated rings; and d) the number of branching points in the molecule. These visualizations are depicted in Figure 7**A**-**D**, respectively).

From Figure 7**A** it is clear that both the HOMO and the LUMO are not directly affected by the number of rings in the molecule. Meaning, the family-dependent trends observed in the distribution plots (Figure 6) are not trivially linked to the number of rings contained in the molecules of each family. Rather, there is a more subtle relationship between the molecular electronic property and the structure of

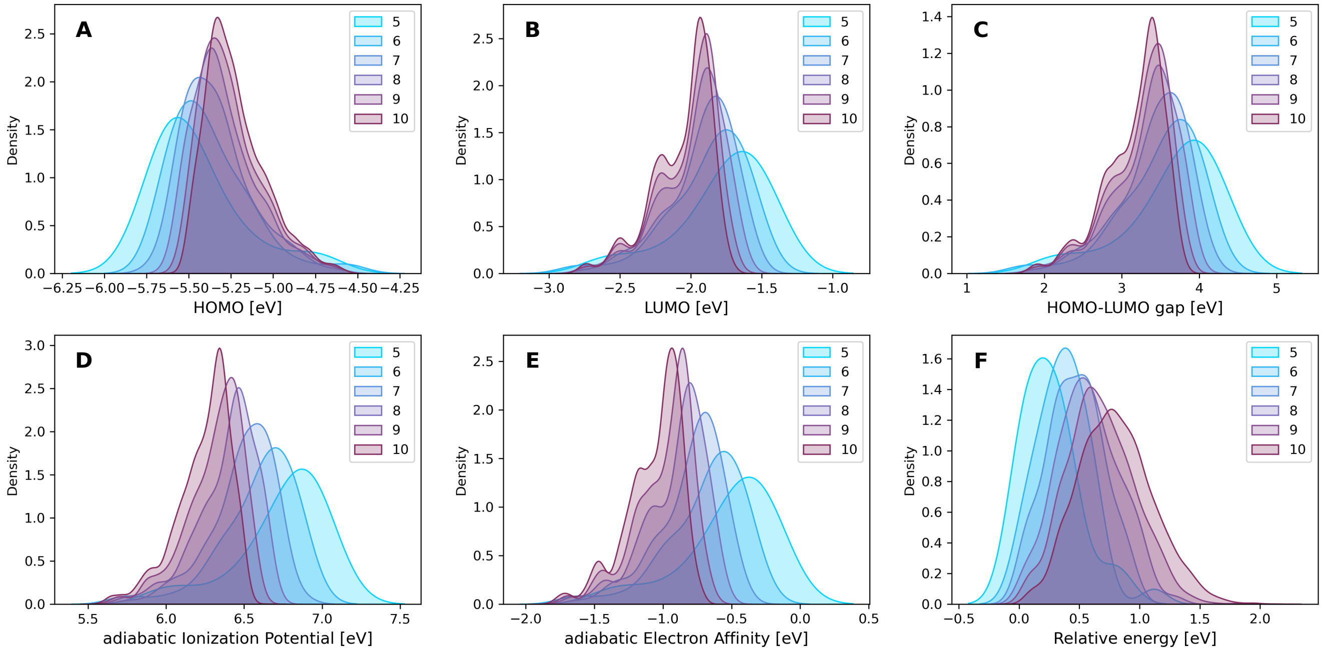

Figure 6: KDE plots of the distribution of DFT-calculated values, separated into families: **A)** HOMO; **B)** LUMO; **C)** HOMO-LUMO gap; **D)** adiabatic ionization potential; **E)** adiabatic electron affinity; **F)** relative single-point energy. All values reported in eV. Families containing molecules with fewer than 5 rings are not shown.

the molecule, that is indirectly connected to the size of the molecules. The structural motif that shows the clearest stratification of the data is the longest linear stretch, which is consistent with our previous conclusions regarding the importance of this feature (*vide supra*). Figure 7**B** shows clearly that the HOMO level becomes less negative and the LUMO level becomes more negative as the linear stretch elongates. This is in line with the family-dependent trends seen above: the greater the number of rings, the more isomers can be made with longer linear stretches. Conversely, it appears that the HOMO becomes more negative and the LUMO becomes less negative with the elongation of the longest angular stretch. However, this behavior is not as clear-cut as with the linear stretch; one can see from Figure 7**C** that there are molecules with varying longest angular stretches distributed throughout the scatter plot. Thus, this general relationship might be better rationalized as the lack of a long linear stretch motif. In other words, given that the molecules have a finite size, molecules that have long consecutive angular stretches will naturally have shorter consecutive linear stretches, which can explain the apparent trend. The effect of the presence of branching points is visualized in Figure 7**D**. In general, it appears that increasing the number of branching points leads to a lower HOMO and a higher LUMO. Again, this can be interpreted in light of the linear stretches: an increase in branching points by necessity precludes the existence of long linear stretches. The same behaviors were observed for the aIP and aEA (see Supporting Information, section S2.1).

The relative energy shows similar trends (see Supporting Information, Figure S11), though the behavior is more complex: while the relative energy generally appears to increase with the elongation of the longest linear stretch, there are also several data points with very short longest linear stretches that have high relative energy values. These data points have a high number of branching points and/or a long consecutive angular stretch, which indicates that the relative energy has an additional structural dependencies, as was implied earlier (*vide supra*).

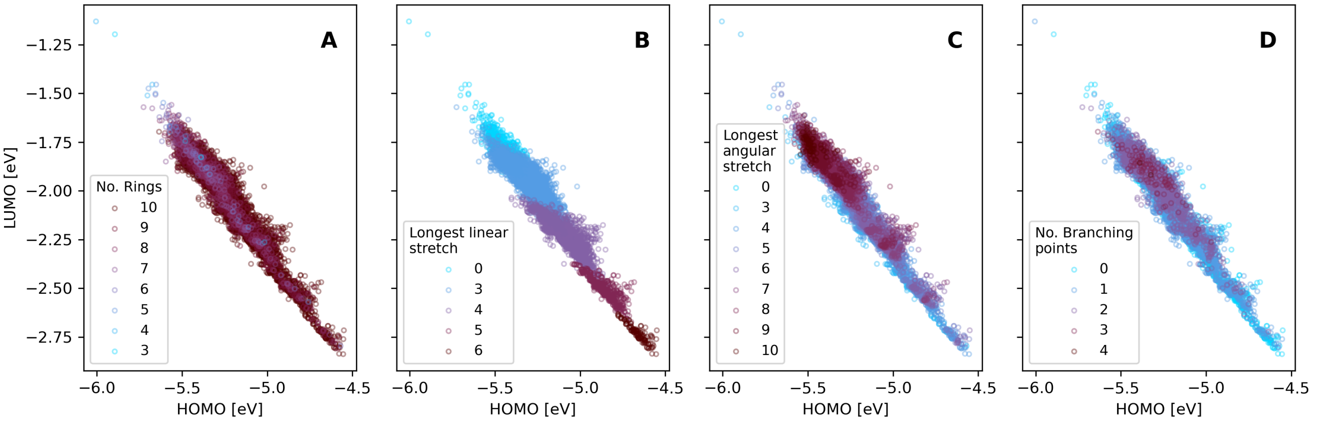

Figure 7: Scatter plots of the DFT-calculated HOMO versus LUMO, colored based on different structural features: **A)** Number of rings; **B)** Longest linear stretch; meaning longest chain of rings connected in a linear fashion; **C)** Longest angular stretch; meaning the longest chain of rings connected in an angular fashion; **D)** Number of branching points. Benzene and naphthalene are not shown.

Conclusions

In this report, we introduced the COMPAS Project, a new computationally-generated database of polycyclic aromatic systems, and described the first phase of the database construction. In Phase 1, we focused on the family of *cata*-condensed PBHs and generated two separate datasets: (1) COMPAS-1x, and (2) COMPAS-1D. The former contains ∼34k PBHs consisting of 1–11 rings, with structures and properties calculated with xTB (using GFN2-xTB). The latter contains ∼9k PBHs consisting of 1–10 rings, with structures and properties calculated with DFT (namely, at the B3LYP-D3BJ/def2-SVP level of theory). Both datasets are freely available at `https://gitlab.com/porannegroup/compas`.

In addition, we performed an analysis of the data. Our results corroborated trends reported in previous experimental and computational work and provided further support for structure-property relationships we previously revealed with electronic-structure investigations. Specifically, we showed that the longest linear annulation stretch is the determining structural feature for several electronic molecular properties (HOMO, LUMO, aEA, aIP), whereas molecular size in and of itself is not an important factor for these properties. The relative energy of PBH isomers appears to show a size-dependency, but a more careful inspection revealed that this is likely due to larger molecules having more possibilities to generate structures that are more distorted from planarity. Similarly, the apparent relationship between the number of branching points and relative energy can also be explained in terms of molecular non-planarity, as branching introduces more curved regions. In other words, we believe the increase in relative energy can be, in large part, attributed to greater torsional/helical strain. Comparison of the xTB and DFT results showed a linear relationship between the two methods for HOMO, LUMO, HOMO-LUMO gap, aEA, and aIP, providing a basis for obtaining close to DFT-level properties with xTB calculations via a correction scheme. For the relative energy, we observed a lack of agreement between the xTB and DFT values that is related to the devitation from planarity. We ascribe this lack of agreement to xTB and DFT calculating the energies of non-planar PBHs differently, specifically their torsional/helical strain energy. In light of this finding, we can further conclude that the other molecular properties are robust to minor structural variability. Additionally, we conclude that all properties are treated adequately by both methods and can be used to glean insights into structure-property trends, which can in turn inform design strategies for preparation of PBHs with tailored properties.

Our group is currently performing additional investigations of the data contained in the COMPAS-1x and COMPAS-1D datasets. We are probing the structure-property relationships of these molecules using interpretable ML and DL techniques, as well as analyzing the aromatic character of these molecules using a wide variety of aromaticity indices. The results of these studies will be communicated in due course. In addition, we are have commenced work on the next phases of the COMPAS Project, which focus on *peri*-condensed PBHs and heterocycle-containing PASs. These subsequent datasets will also be made freely available to the scientific community upon completion.

Acknowledgement The authors thank Prof. Dr. Peter Chen for his unflagging scientific and financial support and Dr. Eno Paenurk, Dr. Alexandra Tsybizova, and Prof. Amnon Stanger for fruitful discussions and helpful commentary on the manuscript. The authors also express their deep appreciation to the Branco Weiss Fellowship for supporting this research as part of a Society in Science grant. R.G.P. is a Branco Weiss Fellow and a Horev Fellow.

Data and Software Availability

The data underlying this study are openly available on Gitlab at `https://gitlab.com/porannegroup/compas`.

Supporting Information Available

General computational details, description of benchmarking procedure, histograms of data distribution, color-coded plots for all studied structural features, further analysis on D3 versus D4 corrections.

References

(1) Anthony, J. E. Functionalized Acenes and Heteroacenes for Organic Electronics. *Chem. Rev.* **2006**, *106*, 5028–5048, DOI: 10.1021/cr050966z.

(2) Aumaitre, C.; Morin, J.-F. Polycyclic Aromatic Hydrocarbons as Potential Building Blocks for Organic Solar Cells. *Chem. Rec.* **2019**, *19*, 1142–1154, DOI: 10.1002/tcr.201900016.

(3) Grant, P. M.; Batra, I. P. Electronic structure of conducting π-electron systems. *Synth. Met.* **1980**, *1*, 193–212, DOI: 10.1016/0379-6779(80)90010-7.

(4) Loots, L.; Barbour, L. J. A simple and robust method for the identification of π–π packing motifs of aromatic compounds. *CrystEngComm* **2011**, *14*, 300–304, DOI: 10.1039/C1CE05763D.

(5) Sutton, C.; Risko, C.; Brédas, J.-L. Noncovalent Intermolecular Interactions in Organic Electronic Materials: Implications for the Molecular Packing vs Electronic Properties of Acenes. *Chem. Mater.* **2016**, *28*, 3–16, DOI: 10.1021/acs.chemmater.5b03266.

(6) Thiessen, A.; Wettach, H.; Meerholz, K.; Neese, F.; Höger, S.; Hertel, D. Control of electronic properties of triphenylene by substitution. *Org. Electron.* **2012**, *13*, 71–83, DOI: 10.1016/j.orgel.2011.10.005.

(7) Dou, J.-H.; Zheng, Y.-Q.; Yao, Z.-F.; Yu, Z.-A.; Lei, T.; Shen, X.; Luo, X.-Y.; Sun, J.; Zhang, S.-D.; Ding, Y.-F.; Han, G.; Yi, Y.; Wang, J.-Y.; Pei, J. Fine-Tuning of Crystal Packing and Charge Transport Properties of BDOPV Derivatives through Fluorine Substitution. *J. Am. Chem. Soc.* **2015**, *137*, 15947–15956, DOI: 10.1021/jacs.5b11114.

(8) Wang, Y.; Liu, B.; Koh, C. W.; Zhou, X.; Sun, H.; Yu, J.; Yang, K.;

Wang, H.; Liao, Q.; Woo, H. Y.; Guo, X. Facile Synthesis of Polycyclic Aromatic Hydrocarbon (PAH)–Based Acceptors with Fine-Tuned Optoelectronic Properties: Toward Efficient Additive-Free Nonfullerene Organic Solar Cells. *Adv. Energy Mater.* **2019**, *9*, 1803976, DOI: 10.1002/aenm.201803976.

(9) Kovalenko, A.; Yumusak, C.; Heinrichova, P.; Stritesky, S.; Fekete, L.; Vala, M.; Weiter, M.; Sariciftci, N. S.; Krajcovic, J. Adamantane substitutions: a path to high-performing, soluble, versatile and sustainable organic semiconducting materials. *J. Mater. Chem. C* **2017**, *5*, 4716–4723, DOI: 10.1039/C6TC05076J.

(10) Lee, J.; Park, S. A.; Ryu, S. U.; Chung, D.; Park, T.; Son, S. Y. Green-solvent-processable organic semiconductors and future directions for advanced organic electronics. *J. Mater. Chem. A* **2020**, *8*, 21455–21473, DOI: 10.1039/D0TA07373C.

(11) Mahmood, J.; Anjum, M. A. R.; Baek, J.-B. Fused Aromatic Network Structures as a Platform for Efficient Electrocatalysis. *J. Adv. Mater.* **2019**, *31*, 1805062, DOI: 10.1002/adma.201805062.

(12) Jang, J. H.; Ahn, S.; Park, S. E.; Kim, S.; Byon, H. R.; Joo, J. M. Synthesis of Redox-Active Phenanthrene-Fused Heteroarenes by Palladium-Catalyzed C–H Annulation. *Org. Lett.* **2020**, *22*, 1280–1285, DOI: 10.1021/acs.orglett.9b04545.

(13) Das, S.; Bhauriyal, P.; Pathak, B. Polycyclic Aromatic Hydrocarbons as Prospective Cathodes for Aluminum Organic Batteries. *J. Phys. Chem. C* **2021**, *125*, 49–57, DOI: 10.1021/acs.jpcc.0c07853.

(14) Kong, D.; Cai, T.; Fan, H.; Hu, H.; Wang, X.; Cui, Y.; Wang, D.; Wang, Y.; Hu, H.; Wu, M.; Xue, Q.; Yan, Z.; Li, X.; Zhao, L.; Xing, W. Polycyclic Aromatic Hydrocarbons as a New Class of Promising Cathode Materials for Aluminum-Ion Batteries. *Angew. Chem. Int. Ed.* **2022**, *61*, e202114681, DOI: 10.1002/anie.202114681.

(15) Cai, X.; Xue, J.; Li, C.; Liang, B.; Ying, A.; Tan, Y.; Gong, S.; Wang, Y. Achieving 37.1% Green Electroluminescent Efficiency and 0.09 eV Full Width at Half Maximum Based on a Ternary Boron-Oxygen-Nitrogen Embedded Polycyclic Aromatic System. *Angew. Chem. Int. Ed.* **2022**, *n/a*, e202200337, DOI: 10.1002/anie.202200337.

(16) Yamashita, Y. Organic semiconductors for organic field-effect transistors. *Sci. Technol. Adv. Mater.* **2009**, *10*, 024313, DOI: 10.1088/1468-6996/10/2/024313.

(17) Chen, M.; Yan, L.; Zhao, Y.; Murtaza, I.; Meng, H.; Huang, W. Anthracene-based semiconductors for organic field-effect transistors. *J. Mater. Chem. C* **2018**, *6*, 7416–7444, DOI: 10.1039/C8TC01865K.

(18) Bachar, N.; Liberman, L.; Muallem, F.; Feng, X.; Müllen, K.; Haick, H. Sensor Arrays Based on Polycyclic Aromatic Hydrocarbons: Chemiresistors versus Quartz-Crystal Microbalance. *ACS Appl. Mater. Interfaces* **2013**, *5*, 11641–11653, DOI: 10.1021/am403067t.

(19) Ando, S.; Nishida, J.-i.; Fujiwara, E.; Tada, H.; Inoue, Y.; Tokito, S.; Yamashita, Y. Novel p- and n-Type Organic Semiconductors with an Anthracene Unit. *Chem. Mater.* **2005**, *17*, 1261–1264, DOI: 10.1021/cm0478632.

(20) Sander, L. C.; Wise, S. A. Polycyclic Aromatic Hydrocarbon Structure Index. *NIST Special Publication 922* **1997**,

(21) Sander, L. C.; Wise, S. A. *Polycyclic Aromatic Hydrocarbon Structure Index*; 2020; DOI: 10.6028/NIST.SP.922e2020.

(22) Alvarez-Ramírez, F.; Ruiz-Morales, Y. Database of Nuclear Independent Chemical Shifts (NICS) versus NICSZZ of Polycyclic Aromatic Hydrocarbons (PAHs). *J.*

Chem. Inf. Model. **2020**, *60*, 611–620, DOI: 10.1021/acs.jcim.9b00909.

(23) Schleyer, P. v. R.; Maerker, C.; Dransfeld, A.; Jiao, H.; van Eikema Hommes, N. J. R. Nucleus-Independent Chemical Shifts: A Simple and Efficient Aromaticity Probe. *J. Am. Chem. Soc.* **1996**, *118*, 6317–6318, DOI: 10.1021/ja960582d.

(24) Chen, Z.; Wannere, C. S.; Corminboeuf, C.; Puchta, R.; Schleyer, P. v. R. Nucleus-Independent Chemical Shifts (NICS) as an Aromaticity Criterion. *Chem. Rev.* **2005**, *105*, 3842–3888, DOI: 10.1021/cr030088+.

(25) Gershoni-Poranne, R.; Stanger, A. In *Aromaticity*; Fernandez, I., Ed.; Elsevier, 2021; pp 99–154, DOI: 10.1016/B978-0-12-822723-7.00004-2.

(26) Brinkmann, G.; Friedrichs, O. D.; Lisken, S.; Peeters, A. CaGe – a Virtual Environment for Studying Some Special Classes of Plane Graphs – an Update. *Commun. Math. Comput. Chem.* **2009**, *63*, 533–552.

(27) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *J. Chem. Theory Comput.* **2019**, *15*, 1652–1671, DOI: 10.1021/acs.jctc.8b01176.

(28) Wahab, A.; Fleckenstein, F.; Feusi, S.; Gershoni-Poranne, R. Predi-XY: a python program for automated generation of NICS-XY-scans based on an additivity scheme. *Electron. Struct.* **2020**, *2*, 047002, DOI: 10.1088/2516-1075/abd081.

(29) Jensen, J. H. "xyz2mol". https://github.com/jensengroup/xyz2mol.

(30) Bendikov, M.; Duong, H. M.; Starkey, K.; Houk, K. N.; Carter, E. A.; Wudl, F. Oligoacenes: Theoretical Prediction of Open-Shell Singlet Diradical Ground States. *J. Am. Chem. Soc.* **2004**, *126*, 7416–7417, DOI: 10.1021/ja048919w.

(31) Knippenberg, S.; Starcke, J. H.; Wormit, M.; Dreuw, A. The low-lying excited states of neutral polyacenes and their radical cations: a quantum chemical study employing the algebraic diagrammatic construction scheme of second order. *Mol. Phys.* **2010**, *108*, 2801–2813, DOI: 10.1080/00268976.2010.526643.

(32) Tönshoff, C.; Bettinger, H. F. Pushing the Limits of Acene Chemistry: The Recent Surge of Large Acenes. *Chem. Eur. J.* **2021**, *27*, 3193–3212, DOI: 10.1002/chem.202003112.

(33) Neese, F. The ORCA program system. *WIREs Comput. Mol. Sci.* **2012**, *2*, 73–78, DOI: 10.1002/wcms.81.

(34) Neese, F. Software update: the ORCA program system, version 4.0. *WIREs Comput. Mol. Sci.* **2018**, *8*, e1327, DOI: 10.1002/wcms.1327.

(35) Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *Chem. Phys.* **1993**, *98*, 5648–5652, DOI: 10.1063/1.464913.

(36) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **1988**, *37*, 785–789, DOI: 10.1103/PhysRevB.37.785.

(37) Miehlich, B.; Savin, A.; Stoll, H.; Preuss, H. Results obtained with the correlation energy density functionals of becke and Lee, Yang and Parr. *Chem. Phys. Lett.* **1989**, *157*, 200–206, DOI: 10.1016/0009-2614(89)87234-3.

(38) Hertwig, R. H.; Koch, W. On the parameterization of the local correlation functional. What is Becke-3-LYP? *Chem. Phys. Lett.* **1997**, *268*, 345–351, DOI: 10.1016/S0009-2614(97)00207-8.

(39) Weigend, F.; Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297, DOI: 10.1039/b508541a.

(40) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate *ab initio* parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *Chem. Phys.* **2010**, *132*, 154104, DOI: 10.1063/1.3382344.

(41) Johnson, E. R.; Becke, A. D. A post-Hartree–Fock model of intermolecular interactions. *Chem. Phys.* **2005**, *123*, 024101, DOI: 10.1063/1.1949201.

(42) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* **2011**, *32*, 1456–1465, DOI: 10.1002/jcc.21759.

(43) Modelli, A.; Mussoni, L.; Fabbri, D. Electron Affinities of Polycyclic Aromatic Hydrocarbons by Means of B3LYP/6-31+G* Calculations. *J. Phys. Chem. A* **2006**, *110*, 6482–6486, DOI: 10.1021/jp0605911.

(44) Hammonds, M.; Pathak, A.; Sarre, P. J. TD-DFT calculations of electronic spectra of hydrogenated protonated polycyclic aromatic hydrocarbon (PAH) molecules: implications for the origin of the diffuse interstellar bands? *Phys. Chem. Chem. Phys.* **2009**, *11*, 4458–4464, DOI: 10.1039/B903237A.

(45) Allison, T. C.; Donald R. Burgess, J. High-Quality Thermochemistry Data on Polycyclic Aromatic Hydrocarbons via Quantum Chemistry. *Polycycl Aromat Compd* **2015**, *35*, 16–31, DOI: 10.1080/10406638.2014.892890.

(46) Bauschlicher, C. W. A comparison of the accuracy of different functionals. *Chem. Phys. Lett.* **1995**, *246*, 40–44, DOI: 10.1016/0009-2614(95)01089-R.

(47) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Model.* **1988**, *28*, 31–36, DOI: 10.1021/ci00057a005.

(48) Daylight. https://daylight.com/.

(49) Markert, G.; Paenurk, E.; Gershoni-Poranne, R. Prediction of Spin Density, Baird-Antiaromaticity, and Singlet–Triplet Energy Gap in Triplet-State Polybenzenoid Systems from Simple Structural Motifs. *Chem. Eur. J.* **2021**, *27*, 6923–6935, DOI: 10.1002/chem.202005248.

(50) Luo, J.; Xue, Z. Q.; Liu, W. M.; Wu, J. L.; Yang, Z. Q. Koopmans' Theorem for Large Molecular Systems within Density Functional Theory. *J. Phys. Chem. A* **2006**, *110*, 12005–12009, DOI: 10.1021/jp063669m.

(51) Salzner, U.; Baer, R. Koopmans' springs to life. *J. Chem. Phys.* **2009**, *131*, 231101, DOI: 10.1063/1.3269030.

(52) Treitel, N.; Shenhar, R.; Aprahamian, I.; Sheradsky, T.; Rabinovitz, M. Calculations of PAH anions: When are diffuse functions necessary? *Phys. Chem. Chem. Phys.* **2004**, *6*, 1113–1121, DOI: 10.1039/B315069K.

(53) Caldeweyher, E.; Ehlert, S.; Hansen, A.; Neugebauer, H.; Spicher, S.; Bannwarth, C.; Grimme, S. A generally applicable atomic-charge dependent London dispersion correction. *J. Chem. Phys.* **2019**, *150*, 154122, DOI: 10.1063/1.5090222.

(54) Khatymov, R. V.; Muftakhov, M. V.; Shchukin, P. V. Negative ions, molecular electron affinity and orbital structure of cata-condensed polycyclic aromatic hydrocarbons. *RCM* **2017**, *31*, 1729–1741, DOI: 10.1002/rcm.7945.

TOC Graphic