scientific data

Check for updates

OPEN

DATA DESCRIPTOR

COMPAS-2: a dataset of *cata*-condensed hetero-polycyclic aromatic systems

Eduardo Mayo Yanes, Sabyasachi Chakraborty & Renana Gershoni-Poranne ✉

Polycyclic aromatic systems are highly important to numerous applications, in particular to organic electronics and optoelectronics. High-throughput screening and generative models that can help to identify new molecules to advance these technologies require large amounts of high-quality data, which is expensive to generate. In this report, we present the largest freely available dataset of geometries and properties of *cata*-condensed poly(hetero)cyclic aromatic molecules calculated to date. Our dataset contains ~500k molecules comprising 11 types of aromatic and antiaromatic building blocks calculated at the GFN1-xTB level and is representative of a highly diverse chemical space. We detail the structure enumeration process and the methods used to provide various electronic properties (including HOMO-LUMO gap, adiabatic ionization potential, and adiabatic electron affinity). Additionally, we benchmark against a ~50k dataset calculated at the CAM-B3LYP-D3BJ/def2-SVP level and develop a fitting scheme to correct the xTB values to higher accuracy. These new datasets represent the second installment in the COMputational database of Polycyclic Aromatic Systems (COMPAS) Project.

Background & Summary

Polycyclic aromatic systems (PASs) are molecules composed of fused aromatic rings. They are an important and pervasive class of molecules, found in both the natural and man-made worlds, that has captivated researchers across many scientific disciplines, thanks to their remarkable structural and functional diversity. To date, PASs have been employed in a wide variety of uses, including as highly tunable fluorescent emitters[1–3], catalysts[4,5], organic semiconductors[6–8], light-emitting diodes[9], field effect transistors[10–12], organic photovoltaics[13–15], synthetic metals[16], chemical sensors[17], and even medicines[18,19].

To design new molecules that can fully harness the potential functionality of PASs, it is necessary to understand their underlying structure-property relationships. However, due to the structural diversity of these compounds, uncovering such relationships is not straightforward. Data-driven approaches can accelerate the discovery and design of new functional molecules. Indeed, such approaches have already allowed the exploration of new swaths of chemical space[20–25]. However, the same tools have seen limited application for PASs[26–28], because the large amounts of data that are necessary to apply them are not readily available. Indeed, despite the importance of PASs to many fields, this chemical space is under-represented in many existing databases, likely due to their molecular size and the computational cost associated with their characterization. Until recently, there were only a few examples of publicly accessible repositories containing appreciable numbers of polycyclic aromatic hydrocarbons (PAHs)[27,29,30] and PASs[20]. Nevertheless, the more recently established OCELOT[31] and PAH335[32] datasets reiterate the importance of and the growing interest in this type of data.

To address the paucity of data for PASs in a methodical and organized manner, we conceptualized and initiated the COMPAS Project (COMputational database of PASs)[33]. Motivated by the understanding that big-data endeavors are crucial to guiding experimental efforts and advancing our chemical understanding[34], we designed the COMPAS Project to enable data-driven investigations of PASs. Among the key features of the COMPAS database are: a) each dataset is generated at a uniform and suitable level of theory, which is necessary to allow the use of data-driven approaches and extraction of chemical insight; b) the data are curated and stored in a manner that is optimal for use with data science tools; c) inexpensive computational methods are benchmarked and fit to higher levels of accuracy, which enables rapid and affordable expansion of the database; d) all data is freely and openly accessible, in compliance with the FAIR principles[35].

Schulich Faculty of Chemistry, Technion - Israel Institute of Technology, Haifa, 32000, Israel. ✉e-mail: rporanne@technion.ac.il

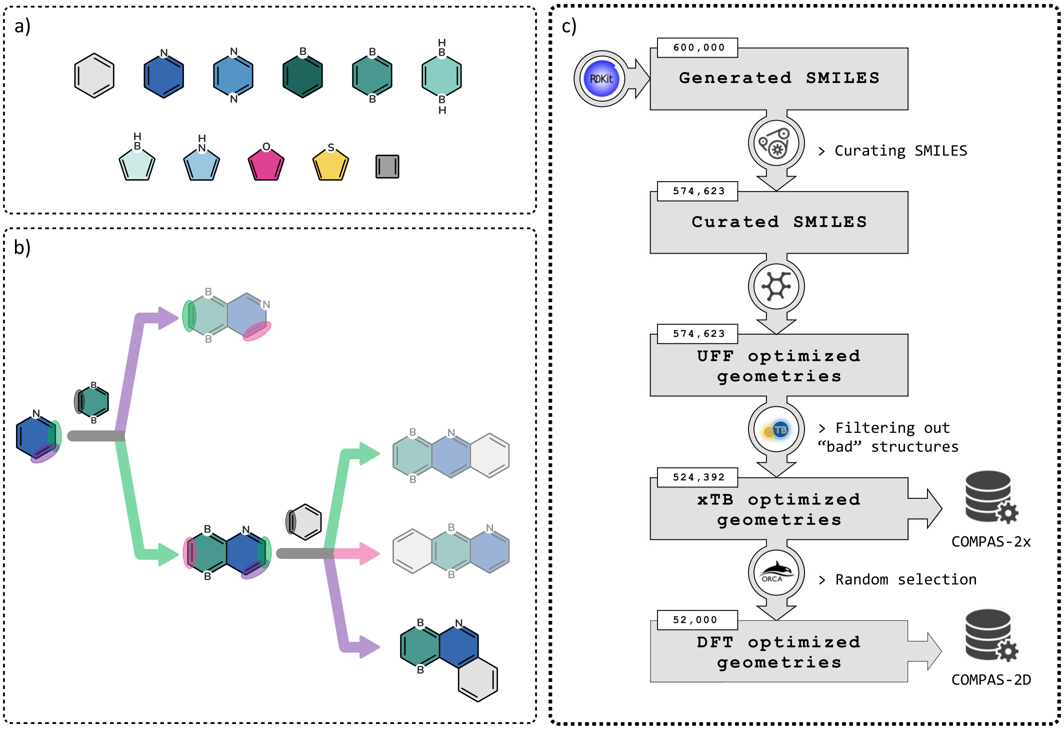

Fig. 1 Various aspects of the COMPAS-2 generation protocol: (**a**) Library of cyclic building blocks used in COMPAS-2; (**b**) An example of an enumeration pathway for generating a tricyclic cc-hPAS molecule; (**c**) The data generation workflow, from structure enumeration to high-throughput calculations to obtain optimized structures and molecular properties.

Herein, we present the second installment of the COMPAS Project, focused on *cata*-condensed heterocycle-containing PASs (cc-hPASs). Such molecules are especially promising as organic semiconductors[36–41]. We describe the construction of two datasets: COMPAS-2x and COMPAS-2D. The former contains the optimized geometries of 524,392 unique cc-hPASs calculated at the GFN1-xTB level[42]; the latter contains the optimized geometries of 52,000 cc-hPASs calculated at the CAM-B3LYP-D3BJ/def2-SVP level[43–48]. The molecules in both datasets range in size from 2 to 10 rings and are constructed from a library of 11 building blocks of diverse size, composition, and aromatic character. To our knowledge, these represent the largest and most structurally diverse datasets of cc-hPASs prepared to date. At the same time, we emphasize that the COMPAS Project is under constant expansion and future installments are already underway.

The current contribution joins the first installment, datasets COMPAS-1x and COMPAS-1D, which contain the structures and properties of *cata*-condensed polybenzenoid hydrocarbons (cc-PBHs) ranging in size from 1 to 11 rings (at the GFN2-xTB level) or from 1 to 10 rings (at the B3LYP-D3BJ/def2-SVP level), respectively[33]. These data can assist in guiding the synthesis of novel molecules, in screening for structures or substructures of interest, in probing fundamental properties (e.g., aromaticity, reactivity), or in training machine learning and deep learning models for various tasks. Indeed, we have recently reported on interpretable models for extracting chemical insight trained on COMPAS-1x[49,50], as well as on a novel guided diffusion model for generating cc-hPASs with targeted properties, trained on some of the data described in this report[51].

In the present report, we describe and discuss the following: a) the composition of the datasets; b) the workflow employed for data generation; c) benchmarking of the data against higher-level calculations and a fitting scheme for obtaining density functional theory (DFT)-level properties from GFN1-xTB calculations.

Methods

In this section, we discuss our protocol for the enumeration of a random subset of the chemical space of *cata*-condensed heterocycle-containing PASs (cc-hPASs) and the high-throughput computations employed to obtain optimized geometries and molecular properties with different methods.

Building-block library. For the construction of the cc-hPAS molecules in COMPAS-2, we used a library of 11 cyclic building blocks, varying in size (from four- to six-membered rings), composition (B, N, O, and S mono- and di-substitution), and aromatic character (aromatic and antiaromatic). Namely, these building blocks are: benzene, pyridine, pyrazine, borinine, 1,4-diborinine, 1,4-dihydro-1,4-diborinine, borole, pyrrole, furan, thiophene, and cyclobutadiene (shown in Fig. 1a). These specific moieties were chosen due to their prevalence and importance in various functional PASs, in particular in the field of organic electronics[36,39–41]. The number of building blocks was limited to 11, which allows us to sample a broad diversity of structures and properties within a feasible number of molecules.

Fragment attached	SMARTS encoding
Benzene	[#6;R1:1]~[#6;R1:2]»[c:2]:1:[c:1]:[c:3]:[c:4]:[c:5]:[c:6]:1
Pyridine	[#6;R1:1]~[#6;R1:2]»[c:6]:1:[c:2]:[c:1]:[n:3]:[c:4]:[c:5]:1
Borinine	[#6;R1:1]~[#6;R1:2]»[#5;a:3]:1:[c:1]:[c:2]:[#6;a:6]:[c:5]:[c:4]:1
Pyrazine	[#6;R1:1]~[#6;R1:2]»[c:2]:1:[c:1]:[n:3]:[c:4]:[c:5]:[n:6]:1
1,4-diborinine	[#6;R1:1]~[#6;R1:2]»[#5;H0;a:6]:1:[#6:2]:[#6:1]:[#5;H0;a:3]:[#6:4]:[#6:5]:1
1,4-dihydro-1, 4-diborinine	[#6;R1:1]~[#6;R1:2]»[H][#5:3]-1-[c:1][c:2]-[#5:6]([H])-[c:5][c:4]-1
Pyrrole	[#6;R1:1]~[#6;R1:2]»[c:2]:1:[c:5]:[c:4]:[n:3]([H]):[c:1]:1
Borole	[#6;R1:1]~[#6;R1:2]»[c:2]:1:[c:5]:[c:4]:[b:3]([H]):[c:1]:1
Thiophene	[#6;R1:1]~[#6;R1:2]»[c:2]:1:[c:5]:[c:4]:[s:3]:[c:1]:1
Furan	[#6;R1:1]~[#6;R1:2]»[c:2]:1:[c:5]:[c:4]:[o:3]:[c:1]:1
Cyclobutadiene	[#6;R1:1]~[#6;R1:2]»[c:1]1[c:2][c:4][c:3]1

Table 1. Table of fragments and their SMARTS encodings.

Enumeration Protocol. To generate cc-hPASs from the building blocks detailed in Fig. 1a, we designed and implemented an enumeration pipeline using SMARTS (SMILES arbitrary target specification language[52]; SMILES - Simplified Molecular Input Line Entry System)[53,54]. Using the SMARTS representation, we encoded different 'reactions' of fusing two rings together in a *cata*-condensed fashion (*cata*-condensation refers to a manner of ring fusion whereby each atom is shared by, at most, two rings). Each 'reaction' creates a fused bond that is shared by two adjoining rings, i.e., two neighboring atoms that are endocyclic to both rings (see Section S2 in the Supporting Information for further details on the SMARTS representation). By performing sequential 'reactions', we generated 600,000 polycyclic compounds.

In generating the structures, we imposed several (arbitrary) constraints. First, to simplify the generation rules and the resulting structures, we opted to allow only carbon atoms on the fused bonds. In other words, heterocyclic moieties can only fuse at their C-C bonds; heteroatoms remain on unfused bonds (the SMARTS formalism for this 'reaction' is shown in Table 1). Second, to ensure only *cata*-condensation is achieved, both carbons in the C-C bond chosen for fusion must belong to only one ring of the nascent cc-hPAS. Third, in cases where more than one C-C bond is suitable for fusion, the choice of which bond to use as the fusion site is random. Fourth, in the ring choice step, we invoked a bias of 10:1 favoring benzene over all other rings (see Section S2 in the Supporting Information for further details). This was done to ensure a more realistic distribution of structures. Fifth, we biased the generation towards molecules of intermediate size (8 and 9 rings) and limited the size of generated molecules to 10-ring systems. The rationale behind this choice was that these sizes provide large structural diversity, and any structure-property relationships should already become obvious in systems of this size (as we previously showed for the cc-PBHs)[33,49,50]. Thus, there was no need to perform calculations of larger systems, which would be substantially more resource-consuming.

The structure generation workflow consists of the following steps:

- Step 1: Generate a random integer (*n*) between 1 and 10.
- Step 2: Randomly select *n* building blocks from the library and store them in a list. The order of the list will be the order of addition of the building blocks to the nascent molecule.
- Step 3: Initialize the nascent molecule with the first building block in the list.
- Step 4: Join the next building block in the list to each of the available C-C bonds in the molecule in turn, each time creating a new structure.
- Step 5: Check all of the resulting structures for chemical validity. Correct errors (e.g., double-bond placement) and remove duplicates.
- Step 6: Randomly select one of the structures. This is now the nascent molecule.
- Step 7: Repeat steps 4–6 until all building blocks in the list have been added.

Figure 1b presents a schematic illustration of an enumeration process leading to a tricyclic product. In the scheme, we show all of the possible resulting structures (which are constitutional isomers), however, in practice, a deterministic choice was made at each step, leading to a single product at the end of each enumeration process. In our example, the process began by randomly choosing $n = 3$ and a list of building blocks comprising pyridine, 1,4-diborinine, and benzene (in that order). In principle, pyridine (shown in blue) has four C-C bonds that can serve as fusion sites. However, only two of them are unique, due to the symmetry of the molecule (these are circled in purple and green, respectively). The next building block that was randomly chosen was 1,4-diborinine (shown in turquoise), which has only one type of fusion site (circled in gray). Joining this new building block to the nascent molecule (pyridine) at the bond circled in purple led to the bicyclic product shown on top (following the purple arrow); joining the new building block at the C-C bond circled in green led to the bottom bicyclic product (following the green arrow). The algorithm then randomly chose to continue with the bottom product (hence, the top one is faded out and there was no continuation of molecular construction). This nascent molecule had three potential fusion sites (circled in pink, green, and purple, respectively). Following the similarly-colored arrows led to each of the three tricyclic products that were obtained through the fusion of the third building block, benzene (shown in gray). At this point, the algorithm once again randomly selected only one of the products (in this case, the bottom one; the other two are faded out). Having reached the

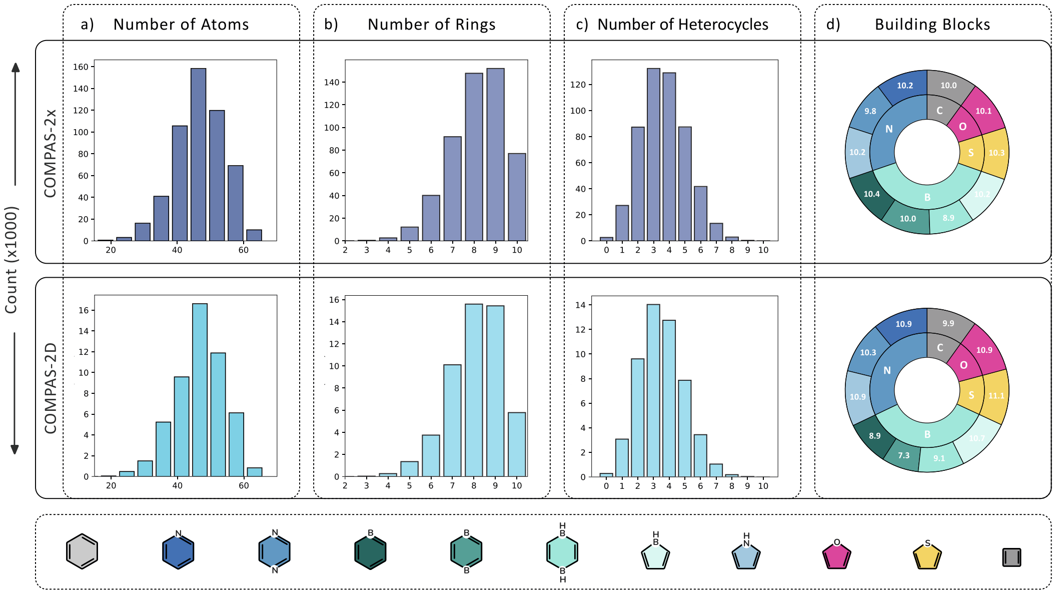

Fig. 2 Overview of the data distribution in the COMPAS-2 datasets: top: COMPAS-2x; bottom: COMPAS-2D. (**a**) histogram of the number of atoms in each molecule; (**b**) histogram of the number of rings in each molecule; (**c**) histogram of the number of heterocycles contained in each molecule; (**d**) doughnut charts of the frequency of different atoms (inner ring) and building blocks (outer rings) present in the dataset. The corresponding color-coded legend of the individual building blocks is also provided.

end of the building block list, the algorithm recognized that the construction had been completed and entered the selected molecule into the dataset. All other structures generated in the process were discarded.

We performed this generation process 600k times and, following each generation process, the resulting cc-hPAS was annotated with its canonical SMILES and InChI[55,56] representations using RDKit[57]. The InChI representation was used to identify and remove duplicate entries. We note that the current enumeration protocol is not memory efficient and may be improved using graph-theoretical methods. Nevertheless, it ensures an exhaustive exploration of the constitutional isomer chemical space (within the described constraints) and generates unique cc-hPASs. The histograms of the various structural features present in the dataset (Fig. 2) show that the distribution of molecular sizes and compositions is well sampled.

High-throughput data generation. Using the protocol described above, we enumerated the InChI representations for a diverse set of 600,000 cc-hPASs. These molecules were then put through a high-throughput computational pipeline to obtain optimized geometries and molecular properties. The steps of the workflow (shown schematically in Fig. 1c) are as follows:

- Step 1: Embed the molecule in 3D space using the Experimental-Torsion Distance Geometry (ETDKG) method with additional "basic knowledge"[58,59], as implemented in RDKit.
- Step 2: Pre-optimize the structure with the universal force field (UFF)[60], as implemented in RDKit. (We note that because sp^3 hybridized B parameters are unavailable in UFF (RDKit), we used sp^2 B parameters. Although this led to suboptimal pre-optimized structures, the subsequent steps ensured close approximations to the ground truth.)
- Step 3: Optimize the structure at the GFN1-xTB level using the xTB software[61] (see Section S3 in the Supporting Information for further details on benchmarking and choice of method).
- Step 4: Calculate harmonic vibrational frequencies to ensure the geometry is a minimum on the potential energy surface (i.e., $N_{imag} = 0$).
- Step 5: Filter out molecules that did not optimize correctly (i.e., optimization did not converge, presence of imaginary frequencies, presence of bond lengths greater than 2.0 Å, or presence of atom-atom distances shorter than 0.1 Å.)
- Step 6: If the obtained structure passes the validity check, optimize the geometries and calculate the frequencies of the anionic and cationic forms of the molecule at the GFN1-xTB level.
- Step 7: Repeat Step 5 for the cationic and anionic forms.

With this pipeline, we obtained the optimized geometries and molecular properties of 524,392 cc-hPASs (corresponding to 22,735 molecular formulae), calculated at the GFN1-xTB level–these comprise the COMPAS-2x dataset. Figure 2a shows the structural diversity of the molecules contained in the COMPAS-2x dataset in terms of molecular size and the distribution of heterocyclic moieties among the molecules.

We note that the majority of COMPAS-2x molecules are medium-sized molecules with ~50 atoms, comprising 8 or 9 rings (Fig. 2a,b, top row). This is because we used a quasi-Poisson distribution to bias the size of the generated molecules towards 8 and 9 rings. Hence, the 10-ring family is smaller even though the number of possible structures increases substantially with the increase in the number of rings. We observe a distribution of the number of heterocycles per molecule, with most of the compounds containing 3 or 4 heterocycles (Fig. 2c, top row). The distribution of the different heterocycles is uniform (Fig. 2d, top row). Overall, the histograms show that the enumeration protocol successfully generates a random and broad sampling of the chemical space.

From COMPAS-2x, we randomly chose 52,000 molecules (approximately 10%), for which we performed geometry optimizations and property calculations with density functional theory (DFT) using the ORCA software[62,63]. For these calculations, we employed the CAM-B3LYP functional[43] with the def2-SVP basis set[47], using Grimme's D3 dispersion correction[44,45] with Becke-Johnson damping[46]. The DFT-optimized geometries and molecular properties of these 52,000 cc-hPASs comprise the COMPAS-2D dataset (corresponding to 9,776 molecular formulae). The assessment of structural diversity for the COMPAS-2D dataset (Fig. 2, bottom row) shows that the distribution of COMPAS-2D is similar to that of COMPAS-2x, indicating that the selection was successfully random and that this dataset is a good sampling of the chemical space, as well.

Data Records

The COMPAS Project is hosted on the Poranne Group's GitLab repository (https://gitlab.com/porannegroup/compas) and is openly and freely available. A minted version of the data reported in this manuscript is available on Figshare (https://doi.org/10.6084/m9.figshare.24347152) (dataset posted on 2023-10-19)[64]. Furthermore, a website (https://compas.net.technion.ac.il/) has been developed and deployed to facilitate sub-structure and property-based queries. The current contribution expands the existing database with two datasets: COMPAS-2x (524,392 cc-hPASs; geometries and properties calculated with GFN1-xTB) and COMPAS-2D (52,000 cc-hPASs; geometries and properties calculated at the CAM-B3LYP-D3BJ/def2-SVP level). Additionally, we include in COMPAS-2x properties of the neutral compounds in COMPAS-2x, which have been corrected from the GFN1-xTB to the CAM-B3LYP-D3BJ/def2-SVP level, using a multi-linear regression correction scheme (see below for further details). Jupyter notebooks used to perform data analyses and multi-linear regressions are also available on the GitLab repository.

File Format. All molecular geometries optimized at the GFN1-xTB and CAM-B3LYP-D3BJ/def2-SVP levels are publicly available for download as compressed sdf files *COMPAS-2x.sdf.gz* and *COMPAS-2D.sdf.gz* files, respectively, from https://gitlab.com/porannegroup/compas. These files contain the optimized geometries (Cartesian coordinates and connectivity information) of 524,392 and 52,000 molecules, respectively. All molecular properties computed at the GFN1-xTB and CAM-B3LYP-D3BJ/def2-SVP level for these optimized geometries in their neutral, cationic, and anionic forms are publicly available for download as *COMPAS-2x.csv*, and *COMPAS-2D.csv* files, respectively, from https://gitlab.com/porannegroup/compas.

Properties. The columns of the *.csv* files correspond to the properties described in Table 2. For every molecule in COMPAS-2x and COMPAS-2D, the respective dataset contains the molecular formula, number of atoms, types of atoms, InChI, SMILES, charge, energy of the highest occupied molecular orbital (HOMO), energy of HOMO–1, energy of the lowest unoccupied molecular orbital (LUMO), energy of LUMO + 1, energy of the HOMO-LUMO gap (Gap, Eq. 1), adiabatic ionization potential (AIP, Eq. 2), and adiabatic electron affinity (AEA, Eq. 3), along with several structural properties, as listed in Table 2. In addition, COMPAS-2x contains the zero-point correction to the energy (ZPE) and the total GFN1-xTB energy (E_{tot}(xTB)), which is the sum of the electronic energy calculated with the self-consistent-charge method (this includes the D4 dispersion correction). COMPAS-2D contains the DFT total energy (E_{tot}(DFT)), which is the sum of the self-consistent field electronic energy, the nuclear repulsion, and the D3-BJ dispersion correction. *COMPAS-2x.csv* also contains "corrected" properties for the neutral state of the 524,392 molecules in COMPAS-2x; i.e., these are values that have been linearly regressed with respect to the *COMPAS-2D.csv* (corrected from xTB to DFT level). The input templates used to compute the properties for COMPAS-2x and COMPAS-2D are provided in Section S1 of the Supporting Information.

The Gap, AIP, and AEA are calculated as follows:

$$\mathrm{Gap} = \mathrm{LUMO} - \mathrm{HOMO} \tag{1}$$

$$\mathrm{AIP} = \mathrm{E_{tot}}^{\mathrm{cation}} - \mathrm{E_{tot}}^{\mathrm{neutral}} \tag{2}$$

$$\mathrm{AEA} = \mathrm{E_{tot}}^{\mathrm{anion}} - \mathrm{E_{tot}}^{\mathrm{neutral}} \tag{2}$$

where $E_{tot}^{neutral}$ is the dispersion-corrected total energy of the optimized molecule in the neutral form, and E_{tot}^{cation} and E_{tot}^{anion} are the dispersion-corrected total energies of the molecule in the cationic and anionic forms, respectively.

Properties	Units	Description	COMPAS-2x	COMPAS-2D
name	—	9-character alpha-numeric name	✓	✓
charge	e	Charge on the molecule	✓	✓
formula	—	Molecular formula	✓	✓
inchi	—	InChI descriptor	✓	✓
smiles	—	SMILES descriptor	✓	✓
rings	—	Number of rings	✓	✓
aromatic_rings	—	Number of aromatic rings	✓	✓
atoms	—	Total number of atoms	✓	✓
heteroatoms	—	Number of heteroatoms	✓	✓
heterocycles	—	Number of heterocycles	✓	✓
branch	—	Number of branches	✓	✓
cyclobutadiene	—	Number of cyclobutadiene rings	✓	✓
pyrrole	—	Number of pyrrole rings	✓	✓
borole	—	Number of borole rings	✓	✓
furan	—	Number of furan rings	✓	✓
thiophene	—	Number of thiophene rings	✓	✓
dhdiborinine	—	Number of 1,4-dihydro-1,4-diborinine rings	✓	✓
14diborinine	—	Number of 1,4-diborinine rings	✓	✓
pyrazine	—	Number of pyrazine rings	✓	✓
pyridine	—	Number of pyridine rings	✓	✓
borinine	—	Number of borinine rings	✓	✓
benzene	—	Number of benzene rings	✓	✓
h	—	Number of hydrogen atoms	✓	✓
c	—	Number of carbon atoms	✓	✓
b	—	Number of boron atoms	✓	✓
s	—	Number of sulfur atoms	✓	✓
o	—	Number of oxygen atoms	✓	✓
n	—	Number of nitrogen atoms	✓	✓
homo	eV	Energy of the HOMO	✓	✓
lumo	eV	Energy of the LUMO	✓	✓
lumo + 1	eV	Energy of the LUMO + 1	✓	✓
homo - 1	eV	Energy of the HOMO - 1	✓	✓
gap	eV	Energy of the LUMO – Energy of the HOMO	✓	✓
nfod	—	Fractional occupation density	✓	
zero_point_energy	Eh	Zero point energy of molecule	✓	
dispersion	Eh	Dispersion correction	✓	✓
energy	Eh	Final energy of molecule	✓	✓
aip	eV	Adiabatic ionization potential	✓	✓
aea	eV	Adiabatic electron affinity	✓	✓
dipole	D/a.u.	Dipole vector of the molecule	✓	✓
homo_corr	eV	xTB-level HOMO corrected to DFT-level	✓	
lumo_corr	eV	xTB-level LUMO corrected to DFT-level	✓	
gap_corr	eV	xTB-level Gap corrected to DFT-level	✓	
energy_corr	Eh	xTB-level E_{tot} corrected to DFT-level	✓	
aip_corr	eV	xTB-level AIP corrected to DFT-level	✓	
aea_corr	eV	xTB-level AEA corrected to DFT-level	✓	
rmsd	Å	Root mean square deviation between xTB- and DFT-optimized structures		✓

Table 2. Property keys, units of the respective quantities, and description of the molecular data present in *COMPAS-2x.csv* and *COMPAS-2D.csv* files. HOMO, LUMO, Gap, AIP, and AEA are provided only for neutral systems. Energies are provided in electron volts (eV) or Hartree (Eh) units. Charge is reported in atomic units (a.u.). The dipole vector for COMPAS-2x is in Debye (D) and in atomic units (a.u.) for COMPAS-2D.

Technical Validation

Comparison between GFN1-xTB and CAM-B3LYP-D3BJ results. The speed and low computational cost of GFN1-xTB make it ideal for high-throughput exploration of large chemical spaces. Naturally, this comes at the expense of accuracy; although xTB is considered to give good energies for reactions, as a semi-empirical method it is less accurate than higher-level *ab initio* and most modern DFT methods. Nevertheless, it is possible

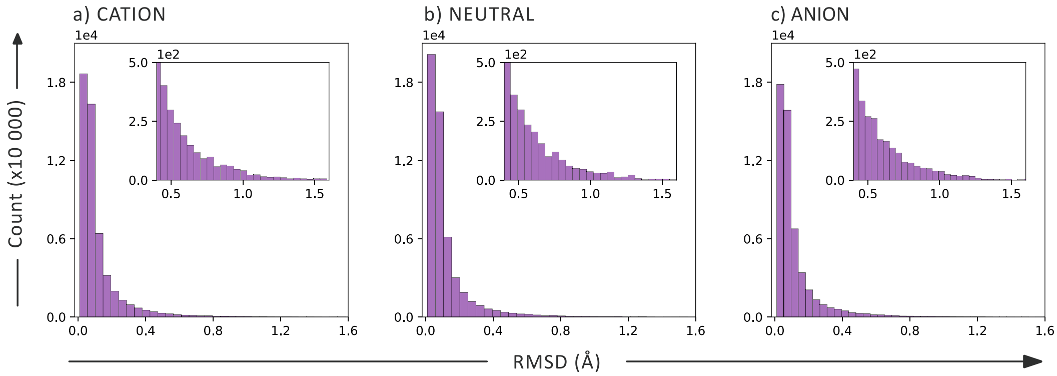

Fig. 3 RMSD (Å) between the structures optimized at GFN1-xTB and CAM-B3LYP/def2-SVP level of theory for the: (**a**) Cation, (**b**) Neutral, and (**c**) Anion species.

to leverage the advantages of the less expensive calculations if a robust scheme can be constructed to correct the GFN1-xTB calculated properties towards a higher accuracy level. In this section, we compare the results obtained with GFN1-xTB to those obtained with CAM-B3LYP-D3BJ/def2-SVP and implement such a correction scheme.

We first compare the geometries obtained by the two levels of theory. The root mean square deviations (RMSDs) for the cationic, neutral, and anionic species are shown in Fig. 3. We observe that the agreement between the geometries obtained with the two methods is satisfactory, with an RMSD < 0.2 Å for 83%, 88%, and 87% of the cationic, neutral, and anionic molecules, respectively. This demonstrates that the xTB method manages to provide similar geometries to the more expensive DFT method. Moreover, it shows that the charge of the species does not affect the agreement between the methods.

Next, we compare the values of various properties in the datasets. Figure 4 displays contour plots of the HOMO, LUMO, Gap, AIP, AEA, and E_{tot} data calculated with the two methods. From the axis values, it is immediately apparent that the values calculated with GFN1-xTB span an entirely different range than those calculated with DFT. For DFT, the orbital energy values are almost always less negative (for HOMO by 4 eV, for LUMO by 7 eV, and accordingly for Gap by 3 eV. For AIP, DFT values are 5 eV lower, consistent with a less negative HOMO. Similarly, for AEA, DFT gives values that are 5 eV less negative, which agrees with a higher-lying LUMO. Despite the offsets in the property value ranges, the plots of the data show that the two methods are reasonably well correlated. The Pearson correlation coefficients (noted for plots (a–e) on each respective plot) range from 0.78 to 0.93, implying that the individual correlations are largely linear. Taking into account only the 80% of the data present in the densest regions increases the coefficients to 0.93–0.97 (see Section S4 in the Supporting Information for scatter plots of the data). We note that the agreement for the AEA is the poorest of all the properties, just as it was for the cc-PBHs we studied previously[33]. We hypothesize that this is because the basis set used for the DFT calculations (def2-SVP, which does not contain diffuse functions) is not optimal for anionic systems, especially non-planar ones. Despite this, the overall satisfactory agreement suggests that the choice of this inexpensive basis set is justified. An interesting phenomenon is observed in the plot of the E_{tot} (Fig. 4f): a series of separate linear correlations is obtained, rather than one main grouping. By examining the structural features of the molecules contained within each grouping, we determined that the differentiation stems from the number of sulfur atoms in the molecule (for further information, see Section S5 in the Supporting Information). We note that a similar issue was reported for organosilicon compounds[65], indicating that there may be a general discrepancy between xTB and DFT in treating third-period atoms. This means that any correction scheme must take this structural information into account. Nevertheless, the remarkably good linear correlation between the two methods suggests that a suitable regression can be constructed to correct this behavior.

GFN1-xTB Corrected Towards CAM-B3LYP-D3BJ. The high Pearson coefficients of the scatter plots (Fig. 4) suggest that linear regressions may be sufficient to correct the GFN1-xTB data towards the CAM-B3LYP-D3BJ/def2-SVP level. Hence, we employed a multi-linear regression, using the GFN1-xTB calculated property value as the baseline and the molecular formula of the molecule as the feature set (for further details see Section S6 of the Supporting Information, which also describes additional regression models that were tested). The advantage of this model is its simplicity–it does not require any knowledge of the specific molecular structure beyond the atomic composition. This model is reminiscent of the quasi-atom corrections[66–68] often used in correcting DFT-level properties, such as formation enthalpies, with respect to composite wavefunction theories[69].

We used the COMPAS-2D molecules as the benchmarking dataset and extracted the property values of the same 52,000 molecules from the COMPAS-2x dataset. We then separated the 52,000 molecules into training (80%) and test (20%) sets and used the training set to optimize the coefficients of the multi-linear regression for each property with respect to the individual features (i.e., numbers of atoms of each type). The coefficients and intercepts obtained from the multi-linear regression are detailed in Table 3. We specifically note the anomalously high coefficient for sulfur atoms in the regression for E_{tot}, which relates to our previous observation regarding the dependence of the energy on the number of sulfurs, as described above.

The resulting fitting equations were then used to correct the GFN1-xTB calculated properties of the test set. The agreement between the values predicted by the corrected scheme and the DFT-calculated values was

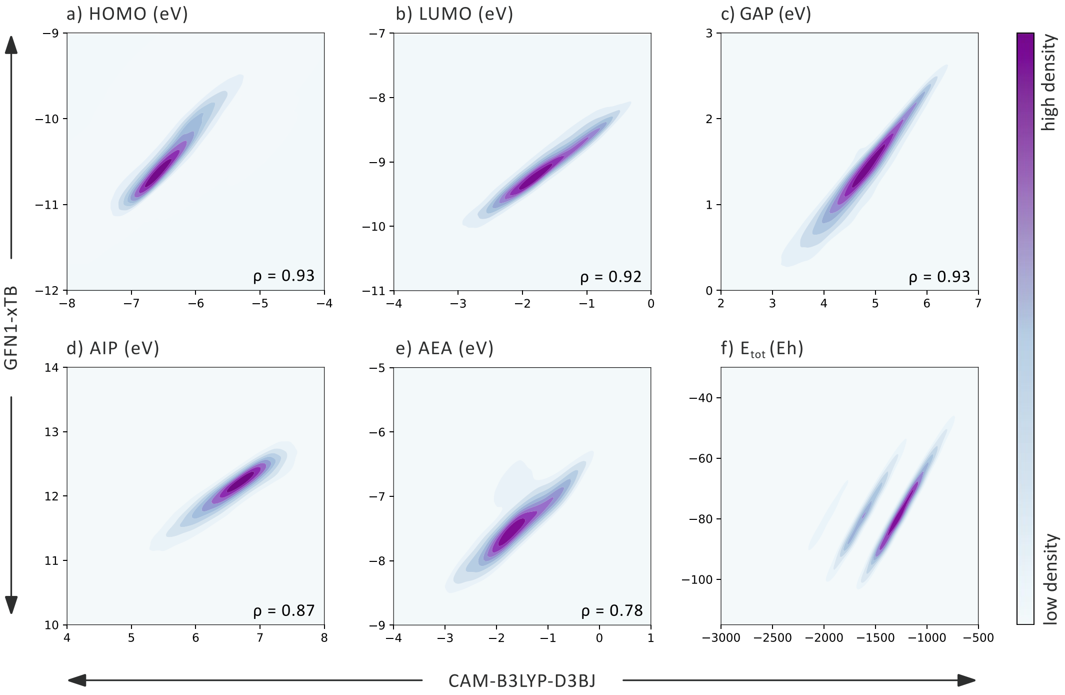

Fig. 4 Comparison of GFN1-xTB and CAM-B3LYP-D3BJ/def2-SVP calculated values for: (**a**) HOMO (eV), (**b**) LUMO (eV), (**c**) Gap (eV), (**d**) AIP (eV), (**e**) AEA (eV), and (**f**) E_{tot} (Eh). The colors of the contour plots indicate the density of points in the region: darker shades indicate high density, lighter shades indicate low density.

Property	H	C	B	S	O	N	xTB	Intercept	R^2	RMSE	MAE
HOMO	0.0124	0.0065	−0.0423	−0.0370	0.0258	−0.0253	1.2397	6.2841	0.89	0.16	0.11
LUMO	0.0160	−0.0079	−0.0862	−0.0303	0.0238	−0.0093	1.0648	8.1821	0.94	0.14	0.10
Gap	0.0263	−0.0245	−0.0358	0.0222	0.0190	0.0331	1.1815	3.4427	0.89	0.23	0.16
AIP	0.0061	0.0088	0.0608	0.0824	0.0042	0.0538	1.3537	−10.3931	0.84	0.22	0.14
AEA	0.0504	−0.0103	−0.1113	0.0012	0.0469	0.0164	0.8820	4.6421	0.74	0.32	0.23
E_{tot}	−0.2102	−36.3793	−23.6932	−395.1947	−71.4597	−52.1923	0.7880	0.1285	1.00	0.02	0.02

Table 3. Statistical data for correction schemes from xTB-calculated properties to DFT-level properties, for the COMPAS-2D molecules. For all multi-linear regressions, the coefficients of atomic features, R^2, RMSE, and MAE are reported. RMSEs and MAEs for all properties are reported in eV, except for E_{tot}, which is reported in Eh.

evaluated (Table 3). Remarkably good correlations were obtained with this very simple regression method and the mean absolute errors (MAEs) indicate that the properties are calculated with satisfying accuracy (especially considering the low computational cost): 0.11 eV, 0.10 eV, 0.16 eV, 0.14 eV, 0.23 eV, and 0.02 Eh for the HOMO, LUMO, Gap, AIP, AEA, and E_{tot}, respectively. The coefficients of determination (ranging between 0.74–1.00) indicate a high measure of linear correlation and good prediction performance. We note that the property with the highest error is the AEA. This is not surprising, given that this property also showed the lowest linearity in Fig. 4e, as we discussed above. Although slightly better agreements can be achieved with more sophisticated models (see Section S6 in the Supporting Information), the simplicity, transparency, and interpretability of the multi-linear regression make it an attractive choice.

To evaluate the performance of our correction scheme, we plotted superimposed histograms of the corrected properties and the DFT-calculated properties for the molecules in our test set (20% of the molecules in COMPAS-2D, which were not used in training the models, *vide supra*). These histograms are shown in Fig. 5 (the values predicted with our correction scheme are shown in light blue; the values calculated for the same molecules with DFT are shown in darker blue). Across all properties, we note a very high degree of overlap, suggesting that the models capture not only the average values (see Table 3), but also the distribution of the data well, and our correction scheme is therefore transferable to other cc-hPASs. This allows us to generate additional datasets with rapid and inexpensive calculations, and easily correct the values towards the more expensive and more accurate DFT level (as we have already done for COMPAS-2x). Although higher accuracy is not always necessary to gain insight and learn structure-property relationships (as we have recently shown[51]), it can be crucial in certain cases. For example, for calculation of properties such as oxidation potential and power conversion efficiency, which are parameterized against experimentally obtained reference states.

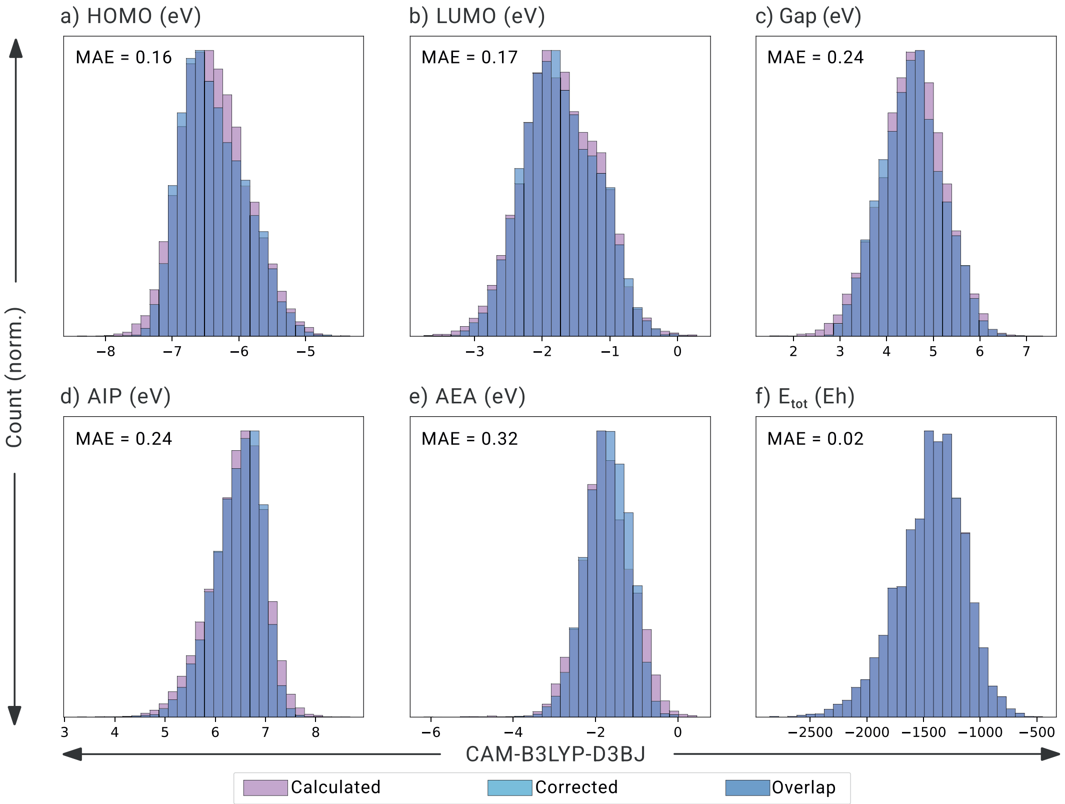

Fig. 5 Comparison of electronic properties obtained with the multi-linear correction scheme (Corrected) against DFT-calculated properties (Calculated) on the test set (20% of COMPAS-2D).

In summary, the COMPAS-2 datasets represent the largest freely available dataset of PASs to date. As the second installment in the still-growing COMPAS Project, COMPAS-2 promises discoveries of hitherto unknown chemical trends in the cc-hPAS chemical space. Considering the importance and prevalence of PASs in chemistry and materials science, these new data can be used to advance a wide variety of disciplines with new opportunities for data-driven investigations to enable the identification of novel functional molecules that may find applications in organic semiconductors and optoelectronics.

Usage Notes

The Python 3.10[70] code used to enumerate the molecular structures, to perform geometry optimization, and to analyze the data are available on GitLab. The repository contains several Python scripts and Jupyter notebooks:

- Scripts to generate molecular structures.
- Scripts to perform semi-empirical calculations and compile the results.
- Scripts to perform DFT calculations and compile the results.
- Notebooks to walk through the chemical library enumeration, data curation, and annotation and reproduce the figures.
- Notebooks to perform the xTB-to-DFT correction.

Code availability

All code is available on the Poranne Group repository on GitLab: https://gitlab.com/porannegroup/compas, licensed under a CC-BY license. Further details are provided in the repository's online README.md file.

Received: 19 October 2023; Accepted: 5 January 2024;
Published online: 19 January 2024

References

1. Boens, N., Leen, V. & Dehaen, W. Fluorescent indicators based on bodipy. *Chem. Soc. Rev.* **41**, 1130–1172 (2012).
2. Cao, D. *et al*. Coumarin-based small-molecule fluorescent chemosensors. *Chem. Rev.* **119**, 10403–10519 (2019).
3. Yang, M., Park, I. S. & Yasuda, T. Full-color, narrowband, and high-efficiency electroluminescence from boron and carbazole embedded polycyclic heteroaromatics. *J. Am. Chem. Soc.* **142**, 19468–19472 (2020).
4. Herrmann, W. A. N-heterocyclic carbenes: a new concept in organometallic catalysis. *Angew. Chem. Int. Ed.* **41**, 1290–1309 (2002).
5. Wang, M. H. & Scheidt, K. A. Cooperative catalysis and activation with n-heterocyclic carbenes. *Angew. Chem. Int. Ed.* **55**, 14912–14922 (2016).

6. Chen, Z. *et al*. Evolution of the electronic structure in open-shell donor-acceptor organic semiconductors. *Nat. Commun.* **12**, 5889 (2021).
7. Lopez, S. A. *et al*. The Harvard organic photovoltaic dataset. *Sci. data* **3**, 1–7 (2016).
8. Jiang, W., Li, Y. & Wang, Z. Heteroarenes as high performance organic semiconductors. *Chem. Soc. Rev.* **42**, 6113–6127 (2013).
9. Guo, J. *et al*. Achieving high-performance nondoped oleds with extremely small efficiency roll-off by combining aggregation-induced emission and thermally activated delayed fluorescence. *Adv. Funct. Mater.* **27**, 1606458 (2017).
10. Kono, T. *et al*. High-performance and light-emitting n-type organic field-effect transistors based on dithienylbenzothiadiazole and related heterocycles. *Chem. Mater.* **19**, 1218–1220 (2007).
11. Chini, M. K., Mahale, R. Y. & Chatterjee, S. Effect of heterocycles on field-effect transistor performances of donor-acceptor-donor type small molecules. *Chem. Phys. Lett.* **661**, 107–113 (2016).
12. Zhao, Z. *et al*. High-performance, air-stable field-effect transistors based on heteroatom-substituted naphthalenediimide-benzothiadiazole copolymers exhibiting ultrahigh electron mobility up to 8.5 cm v- 1 s- 1. *Adv. Mater.* **29**, 1602410 (2017).
13. Chai, G. *et al*. Deciphering the role of chalcogen-containing heterocycles in nonfullerene acceptors for organic solar cells. *ACS Energy Lett.* **5**, 3415–3425 (2020).
14. Yu, H. *et al*. Tailoring non-fullerene acceptors using selenium-incorporated heterocycles for organic solar cells with over 16% efficiency. *J. Mater. Chem. A.* **8**, 23756–23765 (2020).
15. Zhu, E. *et al*. NIR-absorbing electron acceptor based on a selenium-heterocyclic core attaching to phenylalkyl side chains for polymer solar cells with 17.3% efficiency. *ACS Appl. Mater. Interfaces* **14**, 7082–7092 (2022).
16. Cameron, J., Kanibolotsky, A. L. & Skabara, P. J. Lest we forget–the importance of heteroatom interactions in heterocyclic conjugated systems, from synthetic metals to organic semiconductors. *Adv. Mater*. 2302259 (2023).
17. Horak, E., Kassal, P. & Murković Steinberg, I. Benzimidazole as a structural unit in fluorescent chemical sensors: the hidden properties of a multifunctional heterocyclic scaffold. *Supramol. Chem.* **30**, 838–857 (2018).
18. Baumann, M. & Baxendale, I. R. An overview of the synthetic routes to the best selling drugs containing 6-membered heterocycles. *Beilstein J. Org. Chem.* **9**, 2265–2319 (2013).
19. Taylor, A. P. *et al*. Modern advances in heterocyclic chemistry in drug discovery. *Org. Biomol. Chem.* **14**, 6611–6637 (2016).
20. Hachmann, J. *et al*. The Harvard clean energy project: large-scale computational screening and design of organic photovoltaics on the world community grid. *J. Phys. Chem. Lett.* **2**, 2241–2251 (2011).
21. Ramakrishnan, R., Dral, P. O., Rupp, M. & Von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **1**, 1–7 (2014).
22. Kirklin, S. *et al*. The Open Quantum Materials Database (OQMD): assessing the accuracy of dft formation energies. *Npj Comput. Mater.* **1**, 1–15 (2015).
23. Montoya, J. H. & Persson, K. A. A high-throughput framework for determining adsorption energies on solid surfaces. *Npj Comput. Mater.* **3**, 14 (2017).
24. Gallarati, S. *et al*. OSCAR: an extensive repository of chemically and functionally diverse organocatalysts. *Chem. Sci.* **13**, 13782–13794 (2022).
25. Stuyver, T., Jorner, K. & Coley, C. W. Reaction profiles for quantum chemistry-computed [3 + 2] cycloaddition reactions. *Sci. Data* **10**, 66 (2023).
26. Schatschneider, B., Monaco, S., Liang, J.-J. & Tkatchenko, A. High-throughput investigation of the geometry and electronic structures of gas-phase and crystalline polycyclic aromatic hydrocarbons. *J. Phys. Chem. C* **118**, 19964–19974 (2014).
27. Bauschlicher, C. *et al*. The NASA ames polycyclic aromatic hydrocarbon infrared spectroscopic database: the computed spectra. *The Astrophysical Journal Supplement Series* **189**, 341 (2010).
28. Allamandola, L. J. *et al*. The NASA Ames PAH IR Spectroscopic database. astrobiology habitable environment database. Accession date: Jun 21, (2023).
29. Sander, L. C. & Wise, S. A. Polycyclic Aromatic Hydrocarbon Structure Index. *NIST Special Publication 922* (1997).
30. Sander, L. C. & Wise, S. A. Polycyclic Aromatic Hydrocarbon Structure Index. Tech. Rep., National Institute of Standards and Technology (2020).
31. Ai, Q. *et al*. Ocelot: An infrastructure for data-driven research to discover and design crystalline organic semiconductors. *J. Chem. Phys.* **154**, 174705 (2021).
32. Karton, A. & Chan, B. Pah335–a diverse database of highly accurate CCSD (T) isomerization energies of 335 polycyclic aromatic hydrocarbons. *Chemical Physics Letters* **824**, 140544 (2023).
33. Wahab, A., Pfuderer, L., Paenurk, E. & Gershoni-Poranne, R. The compas project: A computational database of polycyclic aromatic systems. phase 1: cata-condensed polybenzenoid hydrocarbons. *J. Chem. Inf. Model.* **62**, 3704–3713 (2022).
34. Yano, J. *et al*. The case for data science in experimental chemistry: examples and recommendations. *Nat. Rev. Chem.* **6**, 357–370 (2022).
35. Draxl, C. & Scheffler, M. Nomad: The fair concept for big data-driven materials science. *MRS Bull.* **43**, 676–682 (2018).
36. Anthony, J. E. Functionalized acenes and heteroacenes for organic electronics. *Chemical reviews* **106**, 5028–5048 (2006).
37. Lin, Y., Li, Y. & Zhan, X. Small molecule semiconductors for high-efficiency organic photovoltaics. *Chem. Soc. Rev.* **41**, 4245–4272 (2012).
38. Sirringhaus, H. 25th anniversary article: organic field-effect transistors: the path beyond amorphous silicon. *Advanced materials* **26**, 1319–1335 (2014).
39. Marques, G. *et al*. De novo design of molecules with low hole reorganization energy based on a quarter-million molecule dft screen. *J. Phys. Chem. A* **125**, 7331–7343 (2021).
40. Staker, J. *et al*. De novo design of molecules with low hole reorganization energy based on a quarter-million molecule dft screen: Part 2. *J. Phys. Chem. A* **126**, 5837–5852 (2022).
41. Wang, C., Zhang, X. & Hu, W. Organic photodiodes and phototransistors toward infrared detection: materials, devices, and applications. *Chem. Soc. Rev.* **49**, 653–670 (2020).
42. Grimme, S., Bannwarth, C. & Shushkov, P. A robust and accurate tight-binding quantum chemical method for structures, vibrational frequencies, and noncovalent interactions of large molecular systems parametrized for all spd-block elements (z = 1–86). *J. Chem. Theory Comput.* **13**, 1989–2009 (2017).
43. Yanai, T., Tew, D. P. & Handy, N. C. A new hybrid exchange–correlation functional using the coulomb-attenuating method (cam-b3lyp). *Chem. Phys. Lett.* **393**, 51–57 (2004).
44. 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. *J. Chem. Phys* **132**, 154104 (2010).
45. Grimme, S., Ehrlich, S. & Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *J Comput Chem* **32**, 1456–1465 (2011).
46. Johnson, E. R. & Becke, A. D. A post-Hartree–Fock model of intermolecular interactions. *J. Chem. Phys* **123**, 024101 (2005).
47. 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.* **7**, 3297–3305 (2005).
48. Weigend, F. Accurate coulomb-fitting basis sets for h to rn. *Phys. Chem. Chem. Phys.* **8**, 1057–1065 (2006).
49. Fite, S., Wahab, A., Paenurk, E., Gross, Z. & Gershoni-Poranne, R. Text-based representations with interpretable machine learning reveal structure–property relationships of polybenzenoid hydrocarbons. *J. Phys. Org. Chem* **36**, e4458 (2023).

50. Weiss, T., Wahab, A., Bronstein, A. M. & Gershoni-Poranne, R. Interpretable deep-learning unveils structure & property relationships in polybenzenoid hydrocarbons. *J. Org. Chem.* **88**, 9645 (2023).
51. Weiss, T. *et al*. Guided diffusion for inverse molecular design. *Nat. Comput. Sci.* **3**, 873 (2023).
52. Daylight Chemical Information Systems, I. SMARTS-a language for describing molecular patterns https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (2007).
53. Weininger, D. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Model.* **28**, 31–36 (1988).
54. Weininger, D., Weininger, A. & Weininger, J. L. SMILES. 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Model.* **29**, 97–101 (1989).
55. Heller, S., McNaught, A., Stein, S., Tchekhovskoi, D. & Pletnev, I. InChI-the worldwide chemical structure identifier standard. *J. Cheminformatics* **5**, 1–9 (2013).
56. Heller, S. R., McNaught, A., Pletnev, I., Stein, S. & Tchekhovskoi, D. InChI, the IUPAC international chemical identifier. *J. Cheminformatics* **7**, 1–34 (2015).
57. Landrum, G. *et al*. RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum* **8** https://www.rdkit.org/RDKit_Overview.pdf. (2013).
58. Riniker, S. & Landrum, G. A. Better informed distance geometry: using what we know to improve conformation generation. *J. Chem. Inf. Model.* **55**, 2562–2574 (2015).
59. Wang, S., Witek, J., Landrum, G. A. & Riniker, S. Improving conformer generation for small rings and macrocycles based on distance geometry and experimental torsional-angle preferences. *J. Chem. Inf. Model.* **60**, 2044–2058 (2020).
60. Rappé, A. K., Casewit, C. J., Colwell, K., Goddard, W. A. III & Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.* **114**, 10024–10035 (1992).
61. Bannwarth, C. *et al*. Extended tight-binding quantum chemistry methods. *WIREs Comput. Mol. Sci.* **11**, e1493 (2021).
62. Neese, F. The ORCA program system. *WIREs Comput. Mol. Sci.* **2**, 73–78 (2012).
63. Neese, F. Software update: the ORCA program system, version 4.0. *WIREs Comput. Mol. Sci.* **8**, e1327 (2018).
64. Mayo, E., Chakraborty, S., & Gershoni-Poranne, R. The COMPAS Project, Phase 2: Cata-Condensed Hetero-Polycyclic Aromatic Systems (COMPAS-2)., *Figshare*, https://doi.org/10.6084/m9.figshare.24347152 (2023).
65. Komissarov, L. & Verstraelen, T. Improving the silicon interactions of gfn-xtb. *J. Chem. Inf. Model.* **61**, 5931–5937 (2021).
66. Winget, P. & Clark, T. Enthalpies of formation from b3lyp calculations. *J. Comp. Chem.* **25**, 725–733 (2004).
67. Grimme, S. Accurate calculation of the heats of formation for large main group compounds with spin-component scaled mp2 methods. *J. Phys. Chem. A* **109**, 3067–3077 (2005).
68. Das, S. K., Chakraborty, S. & Ramakrishnan, R. Critical benchmarking of popular composite thermochemistry models and density functional approximations on a probabilistically pruned benchmark dataset of formation enthalpies. *J. Chem. Phys.* **154** (2021).
69. Karton, A. A computational chemist's guide to accurate thermochemistry for organic molecules. *WIREs Comput. Mol. Sci.* **6**, 292–310 (2016).
70. Van Rossum, G. & Drake, F. L. *Python 3 Reference Manual* (CreateSpace, Scotts Valley, CA, 2009).

Acknowledgements

The authors thank Ms. Alexandra Wahab for fruitful discussions and Prof. Dr. Peter Chen for his support. 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, a Horev Fellow, and an Alon Scholarship recipient.

Author contributions

R.G.P. conceptualized the idea and supervised the preparation, curation, and validation of the data. E.M.Y. implemented the structure enumeration and computational pipelines. S.C. performed the benchmarking experiments. E.M.Y. and S.C. performed statistical analyses. E.M.Y., S.C., and R.G.P. wrote the manuscript. E.M.Y. and R.G.P. prepared the figures. All authors contributed to and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-024-02927-8.

Correspondence and requests for materials should be addressed to R.G.-P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.