COMPAS-4: a Dataset of $(BN)_1$ Substituted *Cata*-Condensed Polybenzenoid Hydrocarbons – Data Analysis and Feature Engineering

Sabyasachi Chakraborty, Itay Almog, and Renana Gershoni-Poranne*

The Schulich Faculty of Chemistry and the Resnick Sustainability Center for Catalysis, Technion – Israel Institute of Technology, Haifa 32000, Israel

E-mail: rporanne@technion.ac.il

Abstract

Incorporation of a BN pair into polycyclic aromatic hydrocarbons is a common approach for modulating their electronic properties. However, a conceptual and quantitative framework rationalizing the observed effects has not been developed, and general structure-property relationships remain elusive. In this work, we perform a data-driven investigation that leads to concrete principles for rational design of $(BN)_1$-PBHs with targeted properties. We construct a new chemical database, COMPAS-4, which contains the geometries and properties of all possible $(BN)_1$-PBH isomers up to 6 rings, calculated at both the GFN1-xTB and DFT (CAM-B3LYP/def2-SVP) levels of theory. We investigate the influence of BN-substitution on various molecular properties, including their molecular orbital energies and aromaticity, and define specific structural features that determine these properties. Notably, all of these features are chemically intuitive and simple to extract from the structure of the molecule, without any prior computation. We find that the most influential feature is the number of rings whose cyclic delocalization is disturbed as a result of the substitution.

Introduction

Polycyclic aromatic systems (PASs) are a class of conjugated molecules that are prevalent in many areas of chemistry and materials sciences.[1,2] They are often considered the 'workhorse' of organic electronics and many examples exist of PASs as organic semiconductors,[3,4] light-emitting diodes,[5] field-effect transistors,[6] organic photovoltaics,[7–9] and fluorescent emitters.[10,11] Nevertheless, there is an ongoing search for new functional PASs with tailored properties, which can potentially enable enhanced device performance and new technologies.

One promising direction is embedding boron-nitrogen (BN) pairs into polybenzenoid hydrocarbons (PBHs; PAS comprising only benzene rings) by replacing any two carbon atoms. Though the resulting BN-PBHs are isoelectronic to their respective parent PBHs, the complementary electron-accepting and electron-donating properties of the B and N atoms have a dramatic effect on the (opto)electronic property space.[12–14]

The earliest attempts to synthesize BN-PBHs were by Dewar and co-workers in the 1960s.[15–17] For several decades afterward, the field lay dormant, until it recently experienced a renaissance led by Paetzold,[18,19] Ashe,[20–23] Piers,[24–27] Liu groups,[28–31] and others.[32–41] These efforts led to the recent identification of function-

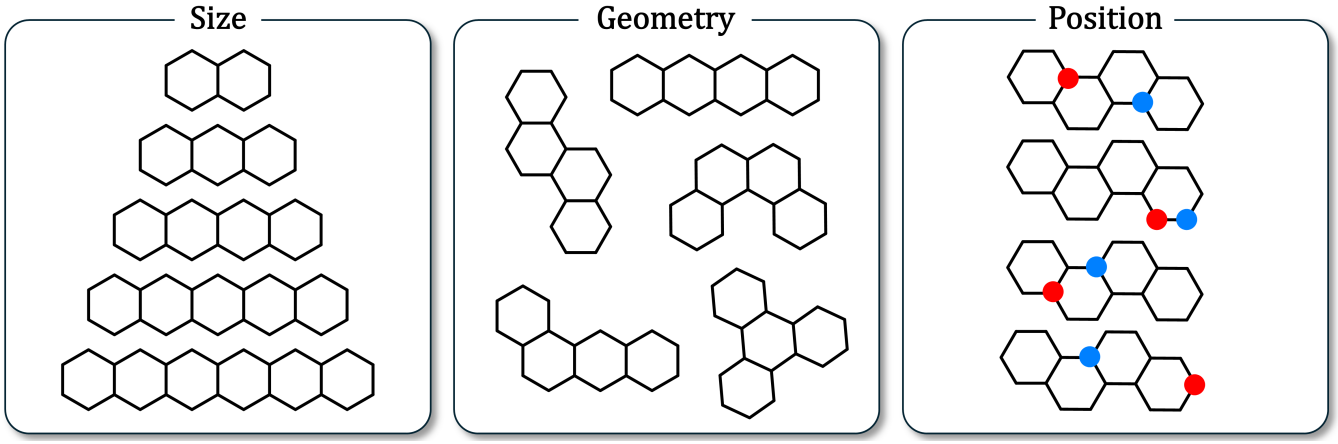

Figure 1: Schematic illustration of the three aspects that provide the structural diversity of $(BN)_1$-PBHs. Explicit Hs and double bonds are omitted for clarity. Red: boron, blue: nitrogen.

ally important systems—molecular solar thermal systems,[42] air-stable organo-electronics,[43] and circularly polarized light emitters.[44] As experimental interest in these molecules surged, so did computational efforts to characterize and understand the underlying structure-property relationships of BN-PBHs. Yet, this remains a daunting task due to the vastness of the chemical space. The structural diversity stems from three core aspects (Figure 1): the number of rings, the geometry in which the rings are fused, and the placement of the B and N atoms. To provide a sense of scale: enumeration of the chemical space of $(BN)_x$-substituted PBHs containing $n_{\text{rings}} \leq 6$ (where $1 \leq x \leq 0.5 n_{\text{C}}$, n_{C} is the number of carbons in the parent PBH, and n_{rings} is the number of rings in the molecule) yields $7.4 \cdot 10^{12}$ molecules.[45]

Both computational and experimental efforts implemented so far have generally focused on studying the positional isomers of a single scaffold. For example, Baranac-Stojanović studied the effect of BN substitution on benzene, naphthalene, and coronene.[46–49] Similar investigations were carried out into the chemical spaces formed from BN-substitution of benzene,[50] naphthalene, anthracene,[51] phenanthrene,[52] tetracene,[13] picene,[53] and perylene.[54] While such investigations do lead to a deeper understanding of the behavior of various subsets of molecules, they are not conducive to defining general design principles. To do so, it is necessary to explore two aspects concurrently: variations in BN-substitution patterns as well as changes to the structure of the PBH scaffold. Only a comprehensive study of this type may lead to the identification of structure-property relationships that apply to the broad chemical space of BN-PBHs.

In this work, we begin to address this gap by constructing, analyzing, and feature-engineering a new dataset, COMPAS-4, which contains all possible *cata*-condensed (cc) $(BN)_1$-PBH isomers comprising up to 6 rings. The COMPAS-4 dataset contains molecules that differ in their n_{rings}, their positional BN isomerization, and their cc-PBH scaffold. Investigating this complete and well-defined chemical space ensures the identification of transferable trends. We define a small set of chemically intuitive structural features and demonstrate that these domain-informed features capture many chemical trends, enabling the prediction of various electronic molecular properties, such as the HOMO-LUMO gap ($\Delta E_{\text{H-L}}$), without the need for prior quantum chemical calculations.

The newly developed dataset is the most recent installment of the COMPAS Project (COMputational database of Polycyclic Aromatic Systems), an open-access database established by our group. The COMPAS database already houses COMPAS-1 (cc-PBHs),[55] COMPAS-2 (heterocyclic cc-PBHs),[56,57] and COMPAS-3 (*peri*-condensed PBHs),[58] which have been used to train both interpretable[59,60] and generative models.[61] COMPAS joins other

databases, such as NASA's Ames,[62] PAH335,[63] FORMED,[64] OE62,[65] and OCELOT,[66] to provide the necessary foundation for data-driven investigations into the important chemical space of PASs.

Data Generation Workflow

The process of generating the COMPAS-4 datasets involved several steps: enumeration, geometry optimization, data filtration, and curation of desired properties. The following subsections detail these steps.

Step 1: Structure Enumeration

Our chemical space of $(BN)_1$-PBHs was based on the collection of 57 parent scaffolds, i.e., all possible unsubstituted cc-PBHs containing $n_{\text{rings}} \leq 6$. By replacing any two carbons with a BN pair, a total of 23,894 unique $(BN)_1$-PBHs can be generated. All of the $(BN)_1$-PBHs included in the dataset were isoelectronic to their parent cc-PBH, meaning that the number of hydrogen atoms was not varied, regardless of the position of substitution. The numbers of scaffolds and $(BN)_1$-PBH isomers available for each n_{rings} are provided in Table 1.

Table 1: Numbers of cc-PBHs scaffolds and $(BN)_1$-PBHs resulting from different n_{rings}.

n_{rings}	No. cc-PBH	No. $(BN)_1$-PBHs
2	1	23
3	2	137
4	5	741
5	12	3,813
6	37	19,180
Total	57	23,894

Step 2: Structure Optimization

The initial Cartesian coordinates of the 23,894 enumerated $(BN)_1$-PBHs were extracted from Ref. 45. The structures were subjected to geometry re-optimization performed with two different computational methods: density functional theory (DFT) and semi-empirical. In both cases, the specific level of theory was chosen so as to maintain consistency and uniformity with the other COMPAS datasets.[55,56,58] These methods were selected following benchmarking procedures that have been detailed in our previous reports.

DFT Optimization. DFT optimization was performed with the CAM-B3LYP functional[67] and the def2-SVP[68] basis set, using Grimme's D3 dispersion correction[69] with Becke-Johnson damping,[70,71] as implemented in Orca 5.0.3 and 5.0.4.[72,73] The resolution-of-identity and the chain-of-spheres approximations (RIJCOSX) were invoked for the Coulomb and exchange integrals, respectively. The AuxJ keyword was used to call the required auxiliary basis sets.[74] Harmonic vibrational frequencies were calculated at the same level of theory to confirm that all geometries were minima on their respective potential energy surfaces. For each molecule, the respective cation and anion were optimized at the same level of theory including subsequent frequency calculations, as well. Altogether, the geometries of 72k species were optimized.

xTB Optimization. The same initial *xyz* coordinates were subjected to optimization with xTB, using GFN1-xTB.[75] Following optimization, harmonic vibrational frequencies were calculated to ensure true minima on the potential energy surface. For each molecule, the respective cation and anion were also optimized with GFN1-xTB, including subsequent frequency calculations. Altogether, the geometries of 72k species were optimized.

Step 3: Data Filtration

We observed that some molecules underwent rearrangements during optimization. In particular, molecules with helical structures tended to undergo undesired bond formation (i.e., cyclization), leading to the wrong structures. These were identified by comparing the InChI descriptors[76] of the optimized structures (generated using OpenBabel[77]) to the input geometries. To ensure both datasets contain the same

molecules, any molecule that rearranged in any of the six optimizations (two computational methods multiplied by three charge states each) was discarded. Altogether, 38 molecules were removed. The geometries and properties of the remaining 23,856 molecules calculated with GFN1-xTB and with CAM-B3LYP/def2-SVP comprise the COMPAS-4x and the COMPAS-4D datasets, respectively.

All neutral species were found to be minima. However, 1,574 DFT-optimized structures (COMPAS-4D) showed imaginary frequencies in the cationic or anionic states (or both). These molecules are included in the datasets but are not considered in the analysis of the aIP and aEA (see the Supporting Information). The data files include a notation identifying these structures.

Step 4: Property Curation

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; $\Delta E_{\mathrm{H-L}}$ is the HOMO-LUMO energy gap; SPE is the dispersion-corrected single-point energy (i.e., the energy of the optimized structure without zero-point corrections); E_{rel} is the relative SPE (*vide infra*); aIP is the adiabatic ionization potential; and aEA is the adiabatic electron affinity. The Gibbs free energy and enthalpy were calculated at 273.15 K.

E_{rel} (calculated only for the neutral species) was obtained by calculating the SPE difference between each molecule and the most stable molecule of the same size (i.e., the same n_{rings}). Accordingly, for every n_{rings}, the lowest value is zero, with all molecules belonging to that subset exhibiting positive E_{rel} with respect to the reference isomer.

For a detailed account of the structural and property range distributions for each of the datasets, see Section S2 in the Supporting Information.

Table 2: Molecular properties included in the COMPAS-4 datasets.

Property	COMPAS-4x	COMPAS-4D
HOMO	✓	✓
LUMO	✓	✓
HLG	✓	✓
SPE (all species)	✓	✓
$E_{\mathrm{rel}}^{\mathrm{SPE}}$ (neutral)	✓	✓
aEA	✓	✓
aIP	✓	✓
Dipole moment	✓	✓
ZPE (all species)	✓	✓
Enthalpy	✓	✓
Gibbs Free Energy	✓	✓

Results and Discussion

This section contains three subsections: domain-informed feature design, structure property relationships, and predictive performance of our feature set. All results in this section are based on the COMPAS-4D dataset.

Domain-Informed Feature Design

In recent years, our group has developed new chemical representations tailored for PASs by identifying which structural features are most dominant in determining molecular properties. Although many other types of representations exist—from very simple (e.g., molecular formula) to complex (e.g., physics-based methods such as Coulomb Matrix,[78] SOAP,[79] SLATM,[80] FCHL,[81,82] and MAOC)[83]—one of the defining characteristics of our representations is that they are based solely on the connectivity of the molecular structure. As a result, they can be extracted simply by visual inspection of a structure and do not require any quantum-chemical calculations. We have demonstrated that our representations allow faster and more efficient training and, more importantly, that they are *interpretable*.[59,60] Chemical insight can be easily extracted from models trained on these representations because they are based on intuitive and clear structural features.

Continuing in the same vein, we sought to identify intuitive and easily extracted structural features that determine the molecular proper-

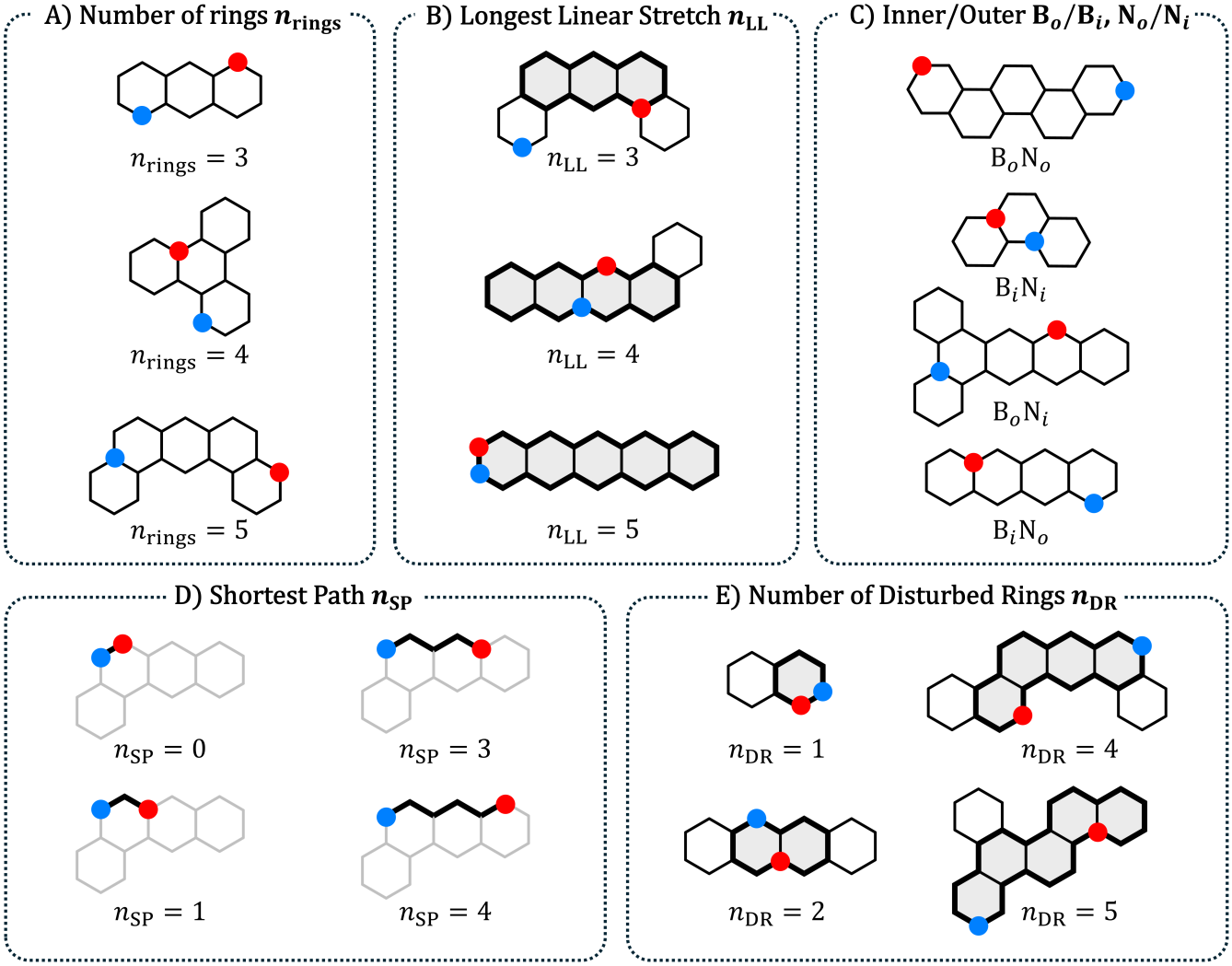

Figure 2: Illustration and representative examples of the five structural descriptors defined in this work. Explicit Hs and double bonds are omitted for clarity. Boron is denoted with a red circle, Nitrogen is denoted with a blue circle.

ties of $(BN)_1$-PBHs. To account for the polycyclic scaffold as well as the positional isomerism of the B and N atoms, we defined five features of two types: two scaffold-dependent (i.e., based on the PBH parent structure) and three BN-location dependent (Figure 2). The five features are:

1. n_{rings}—the number of rings in the molecule (Figure 2A).
2. n_{LL}—the length of the longest linear stretch (Figure 2B).
3. B_i/B_o, N_i/N_o—a binary classification for each heteroatom that indicates whether it is located on a fused bond or not (Figure 2C).
4. n_{SP}—the number of carbons situated between the B and N atoms along the *Shortest Path* that connects them (Figure 2D).
5. n_{DR}—the number of *Disrupted Rings* in the $(BN)_1$-PBH (Figure 2E).

Each of these is discussed further in the next subsection.

Structure-Property Relationships

Each of the features introduced in the previous subsection captures some aspect of the structure of a given $(BN)_1$-PBH. In this subsection, we explain the rationale behind the choice of each feature and demonstrate to what extent it

is relevant to the molecular properties of $(BN)_1$-PBHs. For conciseness, we limit the discussion here to two properties, ΔE_{H-L} and E_{rel}; the analyses performed for three additional properties (aIP, aEA, and the dipole moment) are presented and discussed in Section S4 of the Supporting Information.

Feature #1: n_{rings}

Our previous investigations of PBHs and PASs[55–57] showed that ΔE_{H-L} decreases as n_{rings} increases, which aligns with the well-known behavior of conjugated systems, whereby increasing the size of the conjugated systems raises the HOMO and lowers the LUMO. Unsurprisingly, a similar trend was observed for the $(BN)_1$-PBHs (Figure 3A).

The E_{rel} plots (Figure 3B) showed an opposite relationship with size: as n_{rings} increases, so do the maximal E_{rel} values. It was not surprising to observe opposite trends between ΔE_{H-L} and E_{rel}; The larger ΔE_{H-L} values in PASs generally indicate greater aromaticity and therefore enhanced stability.[84] In addition, there is the effect of the σ-framework. As n_{rings} increases, so do the various geometries that can be formed, including non-planar substructures that incur torsional strain. As we have previously shown, there is a direct link between such non-planar motifs (e.g., fjord, helix) and an increase in $E_{rel}$of cc-PBHs.[59,60]

We also note that, for all molecular sizes, the majority of isomers were located in the higher E_{rel} values (that is, less thermodynamically stable), with trailing edges toward the lower values. Visual inspection of the molecules at different areas of the distribution showed that the $(BN)_1$-PBHs that appeared in the lower E_{rel} regions were those in which B and N shared a bond. Their greater stability could be attributed to their additional electrostatic stabilization.

Feature #2: n_{LL}

In our previous analyses of cc-PBHs, we demonstrated that larger cc-PBHs can be described as sequences of angular (i.e., phenanthrene) and linear (i.e., anthracene) annulations. We

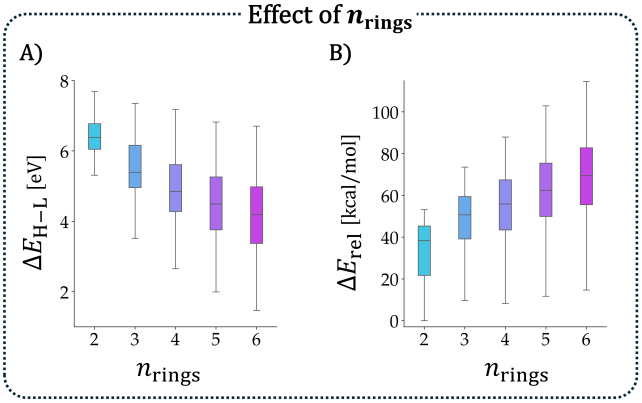

Figure 3: The effect of n_{rings}on A) ΔE_{H-L} (in eV) and B) E_{rel}(in kcal/mol). All molecules in COMPAS-4D were plotted.

showed that several molecular electronic properties are dominated by the longest consecutive sequence of linearly annelated tricycles,[59,60,85] which we represent numerically with the n_{LL} feature. Each 'L' represents three linearly annulated rings; thus, $n_{LL} \leq n_{rings}-2$ ($n_{LL}=n_{rings}-2$ only for a fully-linear isomer).

Put simply, n_{LL} describes the largest polyacene contained within the studied cc-PBH. To study the effect of this motif in $(BN)_1$-PBHs, we plotted the distributions of the two properties, ΔE_{H-L} and E_{rel}, as a function of n_{LL}(Figure 4; to circumvent size dependency, only $n_{rings}=6$ isomers were plotted). Figure 4A showed that ΔE_{H-L} did indeed decrease as n_{LL} increased, indicating that the underlying trend was retained despite BN-substitution. In contrast, n_{LL} had a very modest influence on E_{rel} (Figure 4B). Close visual inspection of the distributions revealed that, on average, molecules with higher n_{LL} values (i.e., longer linear stretches) were slightly less stable. However, all subsets had similar ranges of values. This suggested that, except for the overall extent of conjugation (assessed via n_{rings}) the thermodynamic stability of $(BN)_1$-PBHs was mainly affected by effects such as strain and electrostatics, rather than by π-effects.

Feature #3: B_o/B_i, N_o/N_i

To describe the location of BN-substitution within the PBH scaffold, we introduced the *inner/outer* classification, which indicates

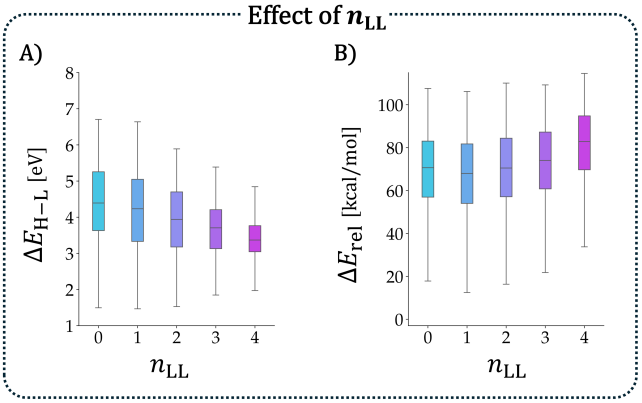

Figure 4: The effect of n_{LL}on A) ΔE_{H-L} (in eV) and B) E_{rel} (in kcal/mol). Only n_{rings}=6 isomers are included.

whether the heteroatom is located on a fused bond (i.e., *inner*: N_i/B_i) or elsewhere (i.e., *outer*: N_o/B_o). Our underlying rationale was that the molecular properties may vary when heteroatoms are on fused bonds, due to, e.g., the absence of bonded hydrogen atoms, changes in geometric strain, or their participation in two separate conjugated cycles.

To study this, we plotted the property distributions for various combinations (Figure 5; to avoid size-dependence, only n_{rings}=6 isomers were included): a) B_oN_o, b) B_iN_i, c) B_oN_i, and d) B_iN_o. Somewhat disappointingly, we did not observe a strong effect, suggesting that this feature had only a mild impact on the molecular properties. For ΔE_{H-L}, the (mild) effect was apparently important only when both heteroatoms were in the same type of position. The clearest difference was seen between the B_oN_o cases (lowest values) and B_iN_i cases (highest values), while essentially no differences were observed between B_oN_i and B_iN_o.

In contrast, for E_{rel}, the effect appeared to be more dependent on which heteroatom was in which position. The highest E_{rel} values were obtained for B_oN_i and the lowest for B_iN_o. The distribution of B_iN_i was closer to B_iN_o, and that of B_oN_o was closer to B_oN_i; this suggested that the position of B played a more dominant role in determining E_{rel}. We surmised that this was due to the number and types of bonds created in each position. When B is in an inner position, it is involved in three B-C bonds; when it is in an outer position, it is involved in two B-C bonds, and another C-C bond is formed. The bond dissociation energies (BDEs) of B-C and C-C bonds are 448 and 618 kJ/mol, respectively.[86] In contrast, B-H and C-H all have comparable BDEs (approx. 340 kJ/mol). As a result, there is a preference for C to be situated in the inner position. We note that this is just a semi-quantitative analysis, based on the most rudimentary BDEs, which do not reflect the complex nature of the conjugated systems under study.

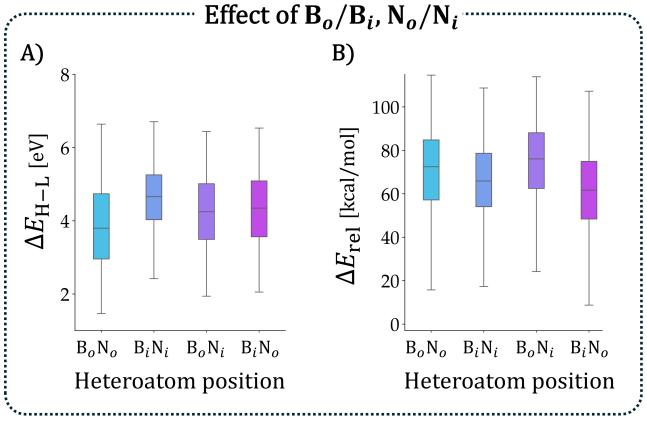

Figure 5: The effect of inner/outer positions on A) ΔE_{H-L} (in eV) and B) E_{rel} (in kcal/mol). Only n_{rings}=6 isomers are included.

Feature #4: n_{SP}

If one considers a parent PBH as a conjugated system, then substitution with a BN pair can be seen as a perturbation of the system. Even though the B and N atoms are sp^2-hybridized and the overall conjugation is formally retained, we hypothesized that the difference in electronegativity creates a disturbance in the original polyene conjugation. Simply put, the dominant substructure in the molecule becomes the polyene-like conjugated path between the B and N atoms. Hence, the molecular properties should be directly dependent on the distance between B and N. This led us to define the n_{SP} feature, which is the number of carbons between the B and N atoms, along the shortest possible acyclic path (several examples are shown in Figure 6).

The distributions of ΔE_{H-L} and E_{rel} against n_{SP} showed that this feature did indeed have a

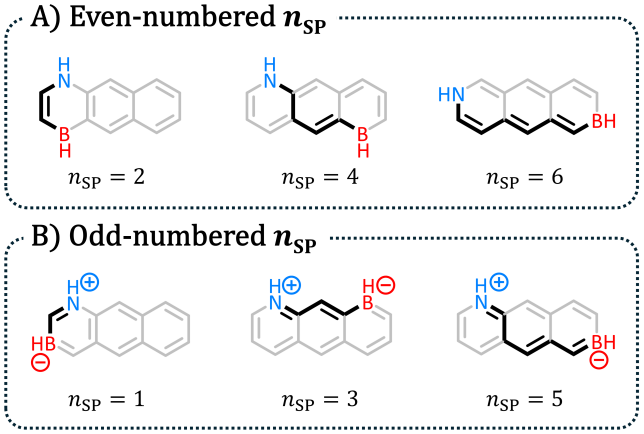

Figure 6: Representative examples of molecules with A) even-numbered and b) odd-numbered n_{SP} values.

strong effect on the molecular properties (Figure 7; to avoid size-dependence, only n_{rings}=6 isomers were included). The ΔE_{H-L} decreased and the E_{rel} increased with greater n_{SP} values. Interestingly, the ΔE_{H-L} plot was reminiscent of the well-known $1/n$ behavior of polyenes, which supported the idea that the path between B and N is indeed the operative polyene structure, as we hypothesized. In both plots, we noticed a minor 'zig-zag' effect between odd- and even-numbered n_{SP} values, which suggested that they were actually two separate series, which could be explained with resonance structures (RSs). Molecules with even n_{SP} (Figure 6A) had homologous vinylic configurations between the B and N, while molecules with odd n_{SP} (Figure 6B) had homologous allylic configurations. Similar trends were observed for aIP and aEA (see Section S4 in the Supporting Information).

We also noted that the E_{rel} values for n_{SP}=0 were substantially lower than the rest of the series, which could be explained by the strong electrostatic stabilization that occurred when B and N shared a bond. This important characteristic was not captured directly by any of the other features. Indeed, this substantial effect has led some researchers to refer to the B-N bond as "the smallest p-n junction".[87]

Feature #5: n_{DR}

In addition to the perturbation of the polyene

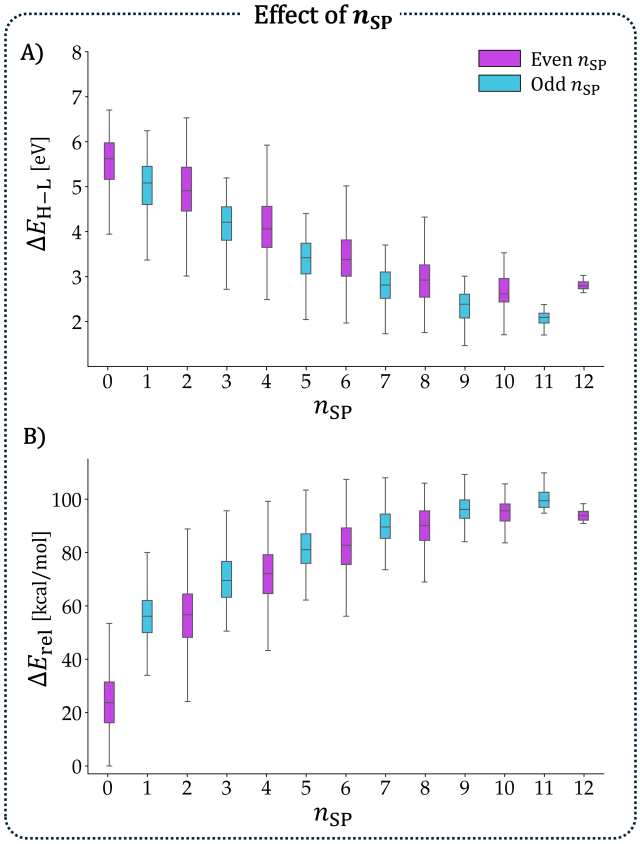

Figure 7: The effect of n_{SP} on A) ΔE_{H-L} (in eV) and B) E_{rel} (in kcal/mol). Only n_{rings}=6 isomers are included.

conjugation, there was also another type of perturbation to consider: the disruption of cyclic conjugation (i.e., aromaticity) in the polycyclic scaffold. Aromaticity in PBHs is often evaluated by the number of Clar sextets (i.e., disjoint sets of 6 π-electrons) in the molecule—the greater the number of sextets, the "more aromatic" the molecule.[88]

As shown in Figure 8A, it is possible to draw two types of RSs for a ring containing B/N: quinoidal or Clar. In the former, there is no Clar sextet (i.e., no aromatic character) and no formal charges on either heteroatom. In the latter, a Clar sextet indicates the existence of aromatic character but comes at the cost of forming formal charges on the B and N atoms. Despite the existence of a Clar sextet, the aromaticity of such rings is attenuated due to the strong localization of π-electrons, as is known from the example of borazine.[89–91] Thus, incorporation of B/N into a ring necessarily disrupts its aromaticity.

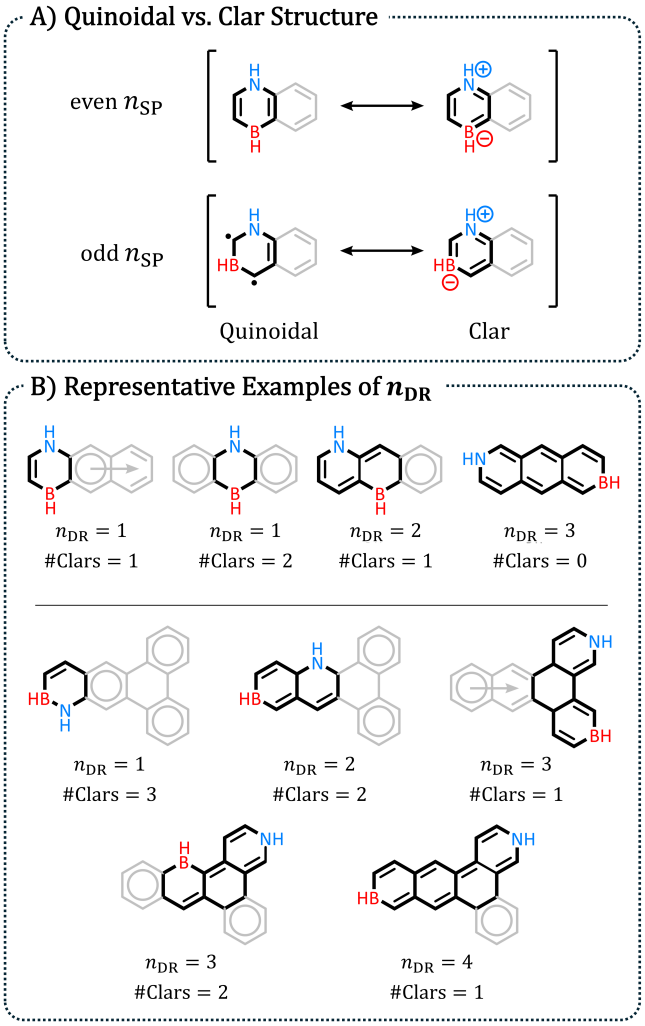

Figure 8: A. Quinoidal and Clar resonance structures for $(BN)_1$-PBHs with an even (top) and odd (bottom) n_{SP} value. B) Two examples of different n_{DR} values for different $(BN)_1$-PBH isomers stemming from the same scaffold.

The disruption is exacerbated when B and N are located in different rings, because any rings situated between them will also adopt a quinoidal structure, precluding them from forming Clar sextets (see examples in Figure 8B). Therefore, the positions of B and N within the scaffold are expected to have a direct effect on the number and size of conjugated circuits, which in turn have a direct impact on the stability (E_{rel}) and ΔE_{H-L} of the $(BN)_1$-PBH. To represent this as a numerical feature, we defined n_{DR}, which is the number of disrupted rings. We note that n_{DR} is not the sole parameter affecting the number of Clar sextets formed; the geometry of the scaffold also plays an important role. Nevertheless, there is a qualitative correspondence, which led us to include n_{DR} in our feature set.

This choice was validated by the plots of ΔE_{H-L} and E_{rel} versus n_{DR}, which revealed strong relationships between the two properties and this structural feature. Figure 9A showed that ΔE_{H-L} steadily decreased with n_{DR}. This result aligned with our rationalization because the "more aromatic" (i.e., less disrupted) a system was, the greater its ΔE_{H-L} was expected to be. Indeed, it has been experimentally determined that the absorption wavelength decreases for isomers with higher numbers of Clar sextets.[92] For the same reason, the decrease in E_{rel} with greater n_{DR} values was unsurprising (Figure 9B). Aromaticity is a stabilizing property, hence, the more it is disrupted, the less thermodynamically stable the molecule should be.

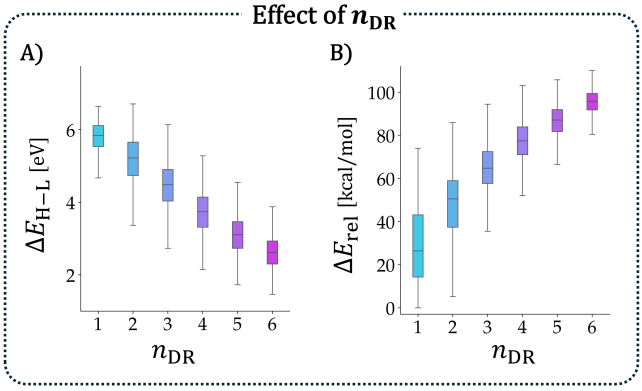

Figure 9: The effect of n_{DR} on A) ΔE_{H-L} (in eV) and B) E_{rel} (in kcal/mol). Only n_{rings}=6 isomers are included.

Predictive Performance of Feature Set

Having defined our set of simple yet meaningful structural features, we turned to the next phase: using these features as input for predictive models. Though each of the features showed a certain relationship to the properties of the $(BN)_1$-PBHs, none of them was sufficient on its own to provide a quantitative prediction of those same properties. However, the *combination* of these features could enable an accurate prediction of molecular properties. If so,

this would validate our selection of structural motifs. Moreover, a successfully trained model could be interrogated to reveal the relative importance of the different features.

To this end, we trained several regression models to predict either the $\Delta E_{\mathrm{H-L}}$ or the E_{rel}. In all cases, the molecular structures were inputted as a vector comprising the five features detailed above. No other structural or property data were used to train the models. A 75:25 (train:test) data split and 5-fold cross-validation were used. The results of the best-performing model, the Light Gradient-Boosting Method (LGBM), are presented in Figure 10.

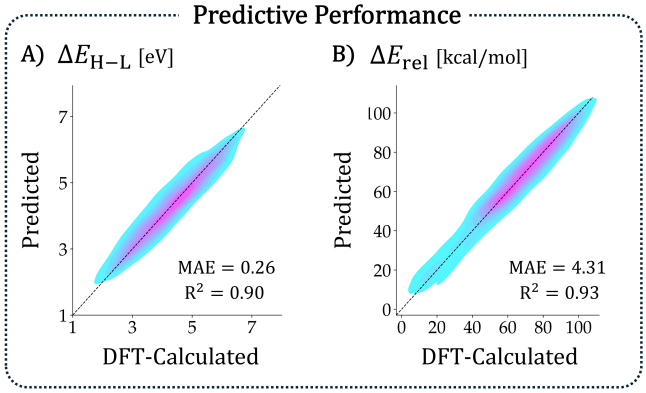

Figure 10: Performance of the trained LGBM regressor models for prediction of A) $\Delta E_{\mathrm{H-L}}$ and B) E_{rel}. Each plot contains the predicted vs. calculated values of the 25% of data reserved for testing. MAE and R^2 are noted on each plot. MAE for $\Delta E_{\mathrm{H-L}}$ is in eV and for E_{rel} in kcal/mol.

Considering the simplicity of our feature set (including the fact that none of the features required any preliminary calculations), the performance of the LGBM model was remarkable. The mean absolute error (MAE) for $\Delta E_{\mathrm{H-L}}$ was MAE=0.26 eV (with R^2=0.90) and for E_{rel} was MAE=4.31 kcal/mol (R^2 = 0.93) Figure 10). The other models performed comparably well, providing further evidence of this feature set's utility. Those results, as well as an ablation study demonstrating the importance of each feature, are provided in Section S5 of the Supporting Information.

Conclusions

The chemical space of BN-PBHs is vast and complex. To date, only a small number of molecules have been synthesized, yet these have already shown promise for (opto)electronics. A methodical and comprehensive exploration is needed to identify additional promising candidates but, more importantly, to define principles for the molecular design of new functional BN-PBHs.

In this study, we reported on the data-driven exploration of the chemical compound space of $(BN)_1$-PBHs containing up to six rings—a collection of 24k molecules. To this end, we computationally generated two new datasets, COMPAS-4D and COMPAS-4x, which contain the geometries and properties of these molecules obtained at the CAM-B3LYP/def2-SVP and GFN1-xTB levels of theory, respectively.

To establish principles for the design of new functional $(BN)_1$-PBHs with specific properties, we sought to identify dominant structure-property relationships. Based on our domain expertise in PASs, we defined a small set of structural motifs: Two features described the size and geometry of the scaffold (n_{rings} and n_{LL}) and the other three addressed the location of B and N within the scaffold ($\mathrm{B}_o/\mathrm{B}_i$, $\mathrm{N}_o/\mathrm{N}_i$, n_{SP}, and n_{DR}). An important aspect of our defined features was that none of them required any quantum chemical calculations, including geometry optimization. In other words, they were truly "back-of-the-envelope".

We investigated the effect of each feature on a set of molecular properties ($\Delta E_{\mathrm{H-L}}$, E_{rel}, aIP, aEA, and dipole moment) and found clear trends that shed light on the underlying relationships. Importantly, the observed trends could be rationalized using fundamental chemical concepts, such as RSs and aromaticity. Notably, the strongest effect was observed for n_{DR}. This result highlighted the importance of considering PASs through the lens of intuitive, organic chemistry-based concepts.

Finally, the defined features were used as input to train various models. The models showed remarkable predictive performance, es-

pecially considering the small number of features and their simplicity. This success demonstrated the power of chemically-informed feature engineering. Most importantly, the intuitive, simple, and interpretable nature of these features allows one to rationally design new $(BN)_1$-PBHs based on an understanding of the underlying structure-property relationships.

Author Information

Corresponding author

- Renana Gershoni-Poranne—Schulich Faculty of Chemistry and the Resnick Sustainability Center for Catalysis, Technion—Israel Institute of Technology, Haifa 32000, Israel; orcid.org/0000-0002-2233-6854; Email: rporanne@technion.ac.il

Authors

- Sabyasachi Chakraborty—Schulich Faculty of Chemistry, Technion – Israel Institute of Technology, Haifa 32000, Israel; orcid.org/0000-0003-4183-811X
- Itay Almog—Schulich Faculty of Chemistry, Technion – Israel Institute of Technology, Haifa 32000, Israel; orcid.org/0009-0004-6912-0470

Author contributions

- Sabyasachi Chakraborty—Methodology, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review and Editing, Visualization.
- Itay Almog—Methodology, Validation, Formal analysis, Investigation, Data curation, Writing - Review and editing, Visualization. item Renana Gershoni-Poranne—Conceptualization, Methodology, Formal analysis, Resources, Writing - Original Draft, Writing - Review and editing, Visualization, Supervision, Project Administration, Funding Acquisition.

Acknowledgement The authors gratefully acknowledge the financial support of the Branco Weiss Fellowship via a Society in Science grant to R.G.P. and the Israel Science Foundation for a Personal Research Grant (#1745/23) to R.G.P. The authors express their deepest appreciation to Prof. Dr. Peter Chen for his ongoing support. The authors are immensely grateful to Prof. Roi Poranne for his valuable assistance and insights with data analysis. S.C. is thankful for ChatGPT and GitHub Co-Pilot.

Data Availability Statement

The data generated in the course of this study and the jupyter notebooks used for analysis and plot generation are all available at https://gitlab.com/porannegroup/compas.

Supporting Information Available

Computational details, input templates, overview of the COMPAS-4D and COMPAS-4x datasets, feature correlation analysis, analysis of three more properties (aIP, aEA, dipole moment), regression analyses for additional models.

References

(1) Mastral, A. M.; Callen, M. S. A review on polycyclic aromatic hydrocarbon (PAH) emissions from energy generation. *Environ. Sci. Technol.* **2000**, *34*, 3051–3057, DOI: 10.1021/es001028d.

(2) Tielens, A. G. Interstellar polycyclic aromatic hydrocarbon molecules. *Annu. Rev. Astron. Astrophys.* **2008**, *46*, 289–337, DOI: 10.1146/annurev.astro.46.060407.145211.

(3) Sergeyev, S.; Pisula, W.; Geerts, Y. H. Discotic liquid crystals: a new generation of organic semiconductors. *Chem. Soc. Rev.* **2007**, *36*, 1902–1929, DOI: 10.1039/B417320C.

(4) Chen, C.; Zhang, Y.; Wang, X.-Y.; Wang, J.-Y.; Pei, J. Boron-and nitrogen-embedded polycyclic arenes as an emerging class of organic semiconductors. *Chem. Mater.* **2023**, *35*, 10277–10294, DOI: 10.1021/acs.chemmater.3c02106.

(5) Xu, Y.; Xu, P.; Hu, D.; Ma, Y. Recent progress in hot exciton materials for organic light-emitting diodes. *Chem. Soc. Rev.* **2021**, *50*, 1030–1069, DOI: 10.1039/D0CS00391C.

(6) Lakshminarayana, A. N.; Ong, A.; Chi, C. Modification of acenes for n-channel OFET materials. *J. Mater. Chem. C* **2018**, *6*, 3551–3563, DOI: doi.org/10.1039/C8TC00146D.

(7) Li, C.; Liu, M.; Pschirer, N. G.; Baumgarten, M.; Mullen, K. Polyphenylene-based materials for organic photovoltaics. *Chem. Rev.* **2010**, *110*, 6817–6855, DOI: 10.1021/cr100052z.

(8) Kumar, M.; Kumar, S. Liquid crystals in photovoltaics: a new generation of organic photovoltaics. *Polym J* **2017**, *49*, 85–111, DOI: 10.1038/pj.2016.109.

(9) 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.

(10) 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**, *61*, e202200337, DOI: 10.1002/anie.202200337.

(11) Hu, Y.; Huang, M.; Liu, H.; Miao, J.; Yang, C. Narrowband Fluorescent Emitters Based on BN-Doped Polycyclic Aromatic Hydrocarbons for Efficient and Stable Organic Light-Emitting Diodes. *Angew. Chem. Int. Ed.* **2023**, *62*, e202312666, DOI: 10.1002/anie.202312666.

(12) Zeng, T.; Mellerup, S. K.; Yang, D.; Wang, X.; Wang, S.; Stamplecoskie, K. Identifying (BN) 2-pyrenes as a new class of singlet fission chromophores: significance of azaborine substitution. *J. Phys. Chem. Lett.* **2018**, *9*, 2919–2927, DOI: 10.1021/acs.jpclett.8b01226.

(13) Pinheiro, M.; Machado, F. B.; Plasser, F.; Aquino, A. J.; Lischka, H. A systematic analysis of excitonic properties to seek optimal singlet fission: the BN-substitution patterns in tetracene. *J. Mater. Chem. C* **2020**, *8*, 7793–7804, DOI: 10.1039/C9TC06581D.

(14) Patra, R.; Das, M. Designing an Efficient Singlet Fission Material with B- N Substitution in Pyrene: A Model Exact Study. *J. Phys. Chem. A* **2024**, *128*, 7375–7383, DOI: 10.1021/acs.jpca.4c03346.

(15) Dewar, M.; Kubba, V. P.; Pettit, R. 624. New heteroaromatic compounds. Part I. 9-Aza-10-boraphenanthrene. *J. Chem. Soc.* **1958**, 3073–3076, DOI: 10.1039/JR9580003073.

(16) Davies, K. M.; Dewar, M. J.; Rona, P. New heteroaromatic compounds. XXVI. synthesis of borazarenes. *J. Am. Chem. Soc.* **1967**, *89*, 6294–6297, DOI: 10.1021/ja01000a054.

(17) Davis, F. A.; Dewar, M. J.; Jones, R.; Worley, S. D. Heteroaromatic compounds. XXXII. Properties of 10, 9-borazaronaphthalene and 9-aza-10-boradecalin. *J. Am. Chem. Soc.* **1969**, *91*, 2094–2097, DOI: 10.1021/ja01036a038.

(18) Paetzold, P. New perspectives in boron-nitrogen chemistry-I. *Pure & Appl. Chem.* **1991**, *63*, 345–350, DOI: 10.1351/pac199163030345.

(19) Paetzold, P.; Stanescu, C.; Stubenrauch, J. R.; Bienmüller, M.; Englert, U. 1-Azonia-2-boratanaphthalenes. *Z. anorg. allg. Chem.* **2004**, *630*, 2632–2640, DOI: 10.1002/zaac.200400333.

(20) Ashe, A. J.; Fang A Synthesis of Aromatic Five-and Six-Membered B- N Heterocycles via Ring Closing Metathesis. *Org. Lett.* **2000**, *2*, 2089–2091, DOI: 10.1021/ol0001113.

(21) Ashe, A. J.; Fang, X.; Fang, X.; Kampf, J. W. Synthesis of 1, 2-dihydro-1, 2-azaborines and their conversion to tricarbonyl chromium and molybdenum complexes. *Organometallics* **2001**, *20*, 5413–5418, DOI: 10.1021/om0106635.

(22) Ashe, A. J.; Yang, H.; Fang, X.; Kampf, J. W. Synthesis and Coordination Chemistry of 3a, 7a-Azaborindenyl, a New Isoelectronic Analogue of the Indenyl Ligand. *Organometallics* **2002**, *21*, 4578–4580, DOI: 10.1021/om0204944.

(23) Pan, J.; Kampf, J. W.; Ashe, A. J. 1, 2-Azaboratabenzene: A heterocyclic π-ligand with an adjustable basicity at nitrogen. *Organometallics* **2004**, *23*, 5626–5629, DOI: 10.1021/om049399g.

(24) Jaska, C. A.; Emslie, D. J.; Bosdet, M. J.; Piers, W. E.; Sorensen, T. S.; Parvez, M. Triphenylene analogues with B2N2C2 cores: synthesis, structure, redox behavior, and photophysical properties. *J. Am. Chem. Soc.* **2006**, *128*, 10885–10896, DOI: 10.1021/ja063519p.

(25) Bosdet, M. J.; Jaska, C. A.; Piers, W. E.; Sorensen, T. S.; Parvez, M. Blue fluorescent 4a-aza-4b-boraphenanthrenes. *Org. Lett.* **2007**, *9*, 1395–1398, DOI: 10.1021/ol070328y.

(26) Bosdet, M.; Piers, W. E.; Sorensen, T. S.; Parvez, M. 10a-Aza-10b-borapyrenes: heterocyclic analogues of pyrene with internalized BN moieties. *Angew. Chem. Int. Ed.* **2007**, *46*, 4940, DOI: 10.1002/anie.200700591.

(27) Bosdet, M. J.; Piers, W. E. BN as a CC substitute in aromatic systems. *Can. J. Chem.* **2009**, *87*, 8–29, DOI: 10.1139/v08-110.

(28) Marwitz, A. J.; Lamm, A. N.; Zakharov, L. N.; Vasiliu, M.; Dixon, D. A.; Liu, S.-Y. BN-substituted diphenylacetylene: A basic model for conjugated π-systems containing the BN bond pair. *Chem. Sci.* **2012**, *3*, 825–829, DOI: 10.1039/C1SC00500F.

(29) Campbell, P. G.; Marwitz, A. J.; Liu, S.-Y. Recent advances in azaborine chemistry. *Angew. Chem. Int. Ed.* **2012**, *51*, 6074–6092, DOI: 10.1002/anie.201200063.

(30) Liu, Z.; Ishibashi, J. S.; Darrigan, C.; Dargelos, A.; Chrostowska, A.; Li, B.; Vasiliu, M.; Dixon, D. A.; Liu, S.-Y. The least stable isomer of BN naphthalene: toward predictive trends for the optoelectronic properties of BN acenes. *J. Am. Chem. Soc.* **2017**, *139*, 6082–6085, DOI: 10.1021/jacs.7b02661.

(31) Ishibashi, J. S.; Dargelos, A.; Darrigan, C.; Chrostowska, A.; Liu, S.-Y. BN Tetracene: extending the reach of BN/CC isosterism in acenes. *Organometallics* **2017**, *36*, 2494–2497, DOI: 10.1021/acs.organomet.7b00296.

(32) Hatakeyama, T.; Hashimoto, S.; Seki, S.; Nakamura, M. Synthesis of BN-fused polycyclic aromatics via tandem intramolecular electrophilic arene borylation. *J. Am. Chem. Soc.* **2011**, *133*, 18614–18617, DOI: 10.1021/ja208950c.

(33) Müller, M.; Maichle-Mössmer, C.; Bettinger, H. F. BN-Phenanthryne: Cyclotetramerization of an 1, 2-Azaborine Derivative. *Angew. Chem. In. Ed.* **2014**, *53*, 9380–9383, DOI: 10.1002/anie.201403213.

(34) Wang, X.-Y.; Wang, J.-Y.; Pei, J. BN heterosuperbenzenes: synthesis and proper-

ties. *Chem. Eur. J.* **2015**, *21*, 3528–3539, DOI: 10.1002/chem.201405627.

(35) Edel, K.; Brough, S. A.; Lamm, A. N.; Liu, S.-Y.; Bettinger, H. F. 1, 2-Azaborine: The Boron-Nitrogen Derivative of ortho-Benzyne. *Angew. Chem.* **2015**, *127*, 7930–7933, DOI: 10.1002/ange.201502967.

(36) Huang, H.; Pan, Z.; Cui, C. The synthesis of BN-embedded tetraphenes and their photophysical properties. *Chem. Commun.* **2016**, *52*, 4227–4230, DOI: 10.1039/C6CC00161K.

(37) Zhuang, F.-D.; Han, J.-M.; Tang, S.; Yang, J.-H.; Chen, Q.-R.; Wang, J.-Y.; Pei, J. Efficient modular synthesis of substituted borazaronaphthalene. *Organometallics* **2017**, *36*, 2479–2482, DOI: 10.1021/acs.organomet.6b00811.

(38) Chen, Y.; Chen, W.; Qiao, Y.; Zhou, G. B2N2-Embedded Polycyclic Aromatic Hydrocarbons with Furan and Thiophene Derivatives Functionalized in Crossed Directions. *Chem. Eur. J.* **2019**, *25*, 9326–9338, DOI: 10.1002/chem.201901782.

(39) Tasseroul, J.; Lorenzo-Garcia, M. M.; Dosso, J.; Simon, F.; Velari, S.; De Vita, A.; Tecilla, P.; Bonifazi, D. Probing Peripheral H-Bonding Functionalities in BN-Doped Polycyclic Aromatic Hydrocarbons. *J. Org. Chem.* **2020**, *85*, 3454–3464, DOI: 10.1021/acs.joc.9b03202.

(40) Ouadoudi, O.; Kaehler, T.; Bolte, M.; Lerner, H.-W.; Wagner, M. One tool to bring them all: Au-catalyzed synthesis of B, O-and B, N-doped PAHs from boronic and borinic acids. *Chem. Sci.* **2021**, *12*, 5898–5909, DOI: 10.1039/D1SC00543J.

(41) Zhao, K.; Yao, Z.-F.; Wang, Z.-Y.; Zeng, J.-C.; Ding, L.; Xiong, M.; Wang, J.-Y.; Pei, J. "Spine Surgery" of Perylene Diimides with Covalent B–N Bonds toward Electron-Deficient BN-Embedded Polycyclic Aromatic Hydrocarbons. *J. Am. Chem. Soc.* **2022**, *144*, 3091–3098, DOI: 10.1021/jacs.1c11782.

(42) Richter, R. C.; Biebl, S. M.; Einholz, R.; Walz, J.; Maichle-Mössmer, C.; Ströbele, M.; Bettinger, H. F.; Fleischer, I. Facile Energy Release from Substituted Dewar Isomers of 1, 2-Dihydro-1, 2-Azaborinines Catalyzed by Coinage Metal Lewis Acids. *Angew. Chem. Int. Ed.* **2024**, *63*, e202405818, DOI: 10.1002/anie.202405818.

(43) Chen, C.; Du, C.-Z.; Wang, X.-Y. The Rise of 1, 4-BN-Heteroarenes: Synthesis, Properties, and Applications. *Advanced Science* **2022**, *9*, 2200707, DOI: 10.1002/advs.202200707.

(44) Appiarius, Y.; Míguez-Lago, S.; Puylaert, P.; Wolf, N.; Kumar, S.; Molkenthin, M.; Miguel, D.; Neudecker, T.; Juríček, M.; Campaña, A. G.; others Boosting quantum yields and circularly polarized luminescence of penta-and hexahelicenes by doping with two BN-groups. *Chemical Science* **2024**, *15*, 466–476, DOI: 10.1039/d3sc02685j.

(45) Chakraborty, S.; Kayastha, P.; Ramakrishnan, R. The chemical space of B, N-substituted polycyclic aromatic hydrocarbons: Combinatorial enumeration and high-throughput first-principles modeling. *J. Chem. Phys.* **2019**, *150*, 114106, DOI: 10.1063/1.5088083.

(46) Baranac-Stojanović, M. Aromaticity and stability of azaborines. *Chem. Eur. J.* **2014**, *20*, 16558–16565, DOI: 10.1002/chem.201402851.

(47) Stojanović, M.; Baranac-Stojanović, M. Mono BN-substituted analogues of naphthalene: a theoretical analysis of the effect of BN position on stability, aromaticity and frontier orbital energies. *New J. Chem.* **2018**, *42*, 12968–12976, DOI: 10.1039/c8nj01529e.

(48) Baranac-Stojanović, M. Triplet-state structures, energies, and antiaromaticity of BN analogues of benzene and their benzo-fused derivatives. *J. Org. Chem.* **2019**, *84*, 13582–13594.

(49) Baranac-Stojanović, M.; Stojanović, M.; Aleksić, J. A theoretical study on application of BN/CC isosterism to modify topology of coronene aromaticity and HOMO–LUMO energy gaps. *New J. Chem.* **2024**, *48*, 14277–14291.

(50) Gupta, D.; Bettinger, H. F. Reactions of 1, 2-Azaborinine, a BN-Benzyne, with Organic π Systems. *J. Org. Chem.* **2023**, *88*, 8369–8378, DOI: 10.1021/acs.joc.3c00401.

(51) Shiraogawa, T.; Krug, S. L.; Ehara, M.; von Lilienfeld, O. A. Antisymmetry rules of response properties in certain chemical spaces. *arXiv preprint arXiv:2502.12761* **2025**,

(52) Shiraogawa, T.; Hasegawa, J.-y. Exploration of chemical space for designing functional molecules accounting for geometric stability. *J. Phys. Chem. Lett.* **2022**, *13*, 8620–8627, DOI: 10.1021/acs.jpclett.2c02355.

(53) von Rudorff, G. F.; von Lilienfeld, O. A. Simplifying inverse materials design problems for fixed lattices with alchemical chirality. *Sci. Adv.* **2021**, *7*, eabf1173, DOI: 10.1126/sciadv.abf1173.

(54) Walia, R.; Yang, J. Exploring optimal multimode vibronic pathways in singlet fission of azaborine analogues of perylene. *Photochem Photobiol Sci* **2022**, *21*, 1689–1700, DOI: 10.1007/s43630-022-00251-x.

(55) 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.* **2022**, *62*, 3704–3713, DOI: 10.1021/acs.jcim.2c00503.

(56) Mayo Yanes, E.; Chakraborty, S.; Gershoni-Poranne, R. COMPAS-2: a dataset of cata-condensed hetero-polycyclic aromatic systems. *Sci Data* **2024**, *11*, 97, DOI: 10.1038/s41597-024-02927-8.

(57) Chakraborty, S.; Yanes, E. M.; Gershoni-Poranne, R. Hetero-polycyclic aromatic systems: A data-driven investigation of structure–property relationships. *Beilstein J. Org. Chem.* **2024**, *20*, 1817–1830, DOI: 10.3762/bjoc.20.160.

(58) Wahab, A.; Gershoni-Poranne, R. COMPAS-3: a dataset of peri-condensed polybenzenoid hydrocarbons. *Phys. Chem. Chem. Phys.* **2024**, *26*, 15344–15357, DOI: 10.1039/d4cp01027b.

(59) 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* **2023**, *36*, e4458, DOI: 10.1002/poc.4458.

(60) Weiss, T.; Wahab, A.; Bronstein, A. M.; Gershoni-Poranne, R. Interpretable deep-learning unveils structure–property relationships in polybenzenoid hydrocarbons. *J. Org. Chem.* **2023**, *88*, 9645–9656, DOI: 10.1021/acs.joc.2c02381.

(61) Weiss, T.; Mayo Yanes, E.; Chakraborty, S.; Cosmo, L.; Bronstein, A. M.; Gershoni-Poranne, R. Guided diffusion for inverse molecular design. *Nat Comput Sci* **2023**, *3*, 873–882, DOI: 10.1038/s43588-023-00532-0.

(62) Bauschlicher, C.; Boersma, C.; Ricca, A.; Mattioda, A.; Cami, J.; Peeters, E.; de Armas, F. S.; Saborido, G. P.; Hudgins, D.; Allamandola, L. The NASA

Ames polycyclic aromatic hydrocarbon infrared spectroscopic database: the computed spectra. *ApJS* **2010**, *189*, 341, DOI: 10.1088/0067-0049/189/2/341.

(63) Karton, A.; Chan, B. PAH335–A diverse database of highly accurate CCSD (T) isomerization energies of 335 polycyclic aromatic hydrocarbons. *J. Chem. Phys. Letters* **2023**, *824*, 140544, DOI: 10.1016/j.cplett.2023.140544.

(64) Blaskovits, J. T.; Laplaza, R.; Vela, S.; Corminboeuf, C. Data-Driven Discovery of Organic Electronic Materials Enabled by Hybrid Top-Down/Bottom-Up Design. *Adv. Mater.* **2024**, *36*, 2305602, DOI: 10.1002/adma.202305602.

(65) Stuke, A.; Kunkel, C.; Golze, D.; Todorović, M.; Margraf, J. T.; Reuter, K.; Rinke, P.; Oberhofer, H. Atomic structures and orbital energies of 61,489 crystal-forming organic molecules. *Sci. Data* **2020**, *7*, 58, DOI: 10.1038/s41597-020-0385-y.

(66) Ai, Q.; Bhat, V.; Ryno, S. M.; Jarolimek, K.; Sornberger, P.; Smith, A.; Haley, M. M.; Anthony, J. E.; Risko, C. OCELOT: An infrastructure for data-driven research to discover and design crystalline organic semiconductors. *J. Chem. Phys.* **2021**, *154*, DOI: 10.1063/5.0048714.

(67) Yanai, T.; Tew, D. P.; Handy, N. C. A new hybrid exchange–correlation functional using the Coulomb-attenuating method (CAM-B3LYP). *J. Chem. Phys. letters* **2004**, *393*, 51–57, DOI: 10.1016/j.cplett.2004.06.011.

(68) 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*, 3297–3305, DOI: 10.1039/b508541a.

(69) 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.* **2010**, *132*, DOI: 10.1063/1.3382344.

(70) 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.

(71) Johnson, E. R.; Becke, A. D. A post-Hartree-Fock model of intermolecular interactions: Inclusion of higher-order corrections. *J. Chem. Phys.* **2006**, *124*, DOI: /10.1063/1.2190220.

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

(73) Neese, F. Software update: The ORCA program system—Version 5.0. *WIREs Comput Mol Sci* **2022**, *12*, e1606, DOI: 10.1002/wcms.1606.

(74) Weigend, F. Accurate Coulomb-fitting basis sets for H to Rn. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1057–1065, DOI: 10.1039/b515623h.

(75) 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.* **2017**, *13*, 1989–2009, DOI: 10.1021/acs.jctc.7b00118.

(76) Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S.; Tchekhovskoi, D. InChI, the IUPAC international chemical identifier. *J. Cheminform.* **2015**, *7*, 1–34, DOI: s13321-015-0068-4.

(77) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical

toolbox. *J. Cheminform.* **2011**, *3*, 1–14, DOI: 10.1186/1758-2946-3-33.

(78) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; Von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **2012**, *108*, 058301, DOI: 10.1103/PhysRevLett.108.058301.

(79) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B* **2013**, *87*, 184115, DOI: 10.1103/PhysRevB.87.184115.

(80) Huang, B.; von Lilienfeld, O. A. Quantum machine learning using atom-in-molecule-based fragments selected on the fly. *Nat. Chem.* **2020**, *12*, 945–951, DOI: 10.1038/s41557-020-0527-z.

(81) Faber, F. A.; Christensen, A. S.; Huang, B.; Von Lilienfeld, O. A. Alchemical and structural distribution based representation for universal quantum machine learning. *J. Chem. Phys.* **2018**, *148*.

(82) Christensen, A. S.; Bratholm, L. A.; Faber, F. A.; Anatole von Lilienfeld, O. FCHL revisited: Faster and more accurate quantum machine learning. *J. Chem. Phys.* **2020**, *152*.

(83) Llenga, S.; Gryn'ova, G. Matrix of orthogonalized atomic orbital coefficients representation for radicals and ions. *J. Chem. Phys.* **2023**, *158*, DOI: 10.1063/5.0151122.

(84) Gershoni-Poranne, R.; Rahalkar, A. P.; Stanger, A. The predictive power of aromaticity: quantitative correlation between aromaticity and ionization potentials and HOMO–LUMO gaps in oligomers of benzene, pyrrole, furan, and thiophene. *Phys. Chem. Chem. Phys.* **2018**, *20*, 14808–14817, DOI: 10.1039/C8CP02162G.

(85) 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.

(86) Luo, Y.-R. *Comprehensive handbook of chemical bond energies*; CRC press, 2007.

(87) Liu, Z.; Marder, T. B. B–N versus C–C: How Similar Are They? *Angew. Chem. Int. Ed.* **2008**, *47*, 242–244, DOI: 10.1002/anie.200703535.

(88) Solà, M. Forty years of Clar's aromatic π-sextet rule. *Frontiers in chemistry* **2013**, *1*, 22, DOI: 10.3389/fchem.2013.00022.

(89) Kiran, B.; Phukan, A. K.; Jemmis, E. D. Is Borazine aromatic? Unusual parallel behavior between hydrocarbons and corresponding B- N analogues. *Inorg. Chem.* **2001**, *40*, 3615–3618, DOI: 10.1021/ic001394y.

(90) Islas, R.; Chamorro, E.; Robles, J.; Heine, T.; Santos, J. C.; Merino, G. Borazine: to be or not to be aromatic. *Structural Chemistry* **2007**, *18*, 833–839, DOI: 10.1007/s11224-007-9229-z.

(91) Merino-García, M. d. R.; Soriano-Agueda, L. A.; Guzmán-Hernández, J. d. D.; Martínez-Otero, D.; Landeros Rivera, B.; Cortés-Guzmán, F.; Barquera-Lozada, J. E.; Jancik, V. Benzene and borazine, so different, yet so similar: insight from experimental charge density analysis. *Inorg. Chem.* **2022**, *61*, 6785–6798, DOI: 10.1021/acs.inorgchem.1c03923.

(92) Rieger, R.; Müllen, K. Forever young: polycyclic aromatic hydrocarbons as model cases for structural and optical studies. *J. Phys. Org. Chem.* **2010**, *23*, 315–325, DOI: 10.1002/poc.1644.

TOC Graphic