Revealing Structure-Property Relationships in Polybenzenoid Hydrocarbons with Interpretable Machine-Learning

Shachar Fite,[a] Alexandra Wahab,[b] Eno Paenurk,[b] Zeev Gross,[a] and Renana Gershoni-Poranne* [a, b]

[a] Schulich Faculty of Chemistry, Technion – Israel Institute of Technology, Haifa 32000, Israel

[b] Laboratory for Organic Chemistry, Department of Chemistry and Applied Biosciences, ETH Zurich, 8093 Zurich, Switzerland

E-mail: rporanne@technion.ac.il

Abstract

The structure-property relationships of polybenzenoid hydrocarbons (PBHs) were investigated with interpretable machine learning, for which two new tools were developed and applied. First, a novel textual molecular representation, based on the annulation sequence of PBHs was defined and developed. This representation can be used either in its textual form or as a basis for a curated feature-vector; both forms show improved interpretability over the standard SMILES representation, and the former also has increased predictive accuracy. Second, the recently-developed model, CUSTODI, was applied for the first time as an interpretable model and identified important structural features that impact various electronic molecular properties. The resulting insights not only validate several well-known "rules of thumb" of organic chemistry but also reveal new behaviors and influential structural motifs, thus providing guiding principles for rational design and fine-tuning of PBHs.

Introduction

In recent years, machine learning (ML) has been increasingly used in chemistry to address a wide variety of challenges, ranging from drug design[1,2] to automatic synthesis,[3–5] to accelerating traditional computations.[6–8] Whereas the success of earlier models was measured by efficiency and accuracy in prediction, current models are often aimed towards better "interpretability" – i.e., an ability to provide guiding principles and insight into domain relationships.[9] In other words, scientists wish to understand "what the model has learned", which may serve to validate existing chemical laws and intuitions,[10,11] or, hopefully, even lead to the discovery of new physical and chemical insights.[9,12,13]

Recent reports have demonstrated that ML can "rediscover" concepts and conventional wisdom in chemistry and physics. Examples include: the effect of specific functional groups on solubility and HOMO level,[11] the hard and soft acids and bases (HSAB) principle for stability of inorganic complexes, and the identification of important normal modes for molecular dissociation[14]. Alongside these, there is discussion of how ML can lead to entirely new discoveries.[12] It should be mentioned, however, that in all of these cases, domain expertise is required, either to engineer the features given to the model or to place the "understanding" of the model in a domain-appropriate context.

In this work, we apply interpretable machine learning to the question of structure-property relationships in the family of compounds known as *cata*-condensed polybenzenoid hydrocarbons (PBHs; sometimes also referred to as catafusenes or as polycyclic aromatic hydrocarbons, PAHs). These molecules are impactful in many areas, in particular in human and environmental health[15,16] and in organic electronics.[17–19] Due to their importance, these compounds have been extensively studied for many decades, both computationally and experimentally. They continue to garner attention for their potential to be used as organic semiconductors[20] and because they are precursors to nano-graphene sheets.[21] Understanding the properties of PBHs is crucial to both understanding their reactivity and designing new functional materials and new pathways for safe disposal of harmful ones. Thus, obtaining a deeper understanding of structure-property relationships governing the behavior of PBHs is of interest both from the conceptual aspect and from the practical one.

Beyond these reasons to study PBHs, there is also a fundamental issue. To paraphrase Randic:[22] in order to understand the behavior of polycyclic aromatic systems (PASs) in a general way, one must first understand the systems comprising the archetypal aromatic unit – benzene. We envision that the current study is the necessary foundation for future investigations of broader swaths of the PAS chemical space. The prevalence of PASs in both natural and man-made materials entails that factors affecting their molecular properties are important to consider in designing new functional molecules and materials.

We approach the subject of interpretable ML in the context of aromatic molecules from two directions: a) the introduction of a new type of molecular representation specifically suited to this kind of molecules and b) the application of a novel interpretable ML method, named CUSTODI,[23] which does not require any human-aided feature selection. We show that our new representation is suitable for extracting chemically meaningful insight and has similar performance to state-of-the-art techniques, but with shorter training times and fewer data required for training. The combination of these two new tools allows us both to validate structure-property relationships previously revealed using electronic-structure investigations and also to uncover additional relationships. These can then inform the rational design and/or fine-tuning of properties.

Methods

The LALAS Representation

Our group has demonstrated in a series of reports over the past few years that *cata*-condensed PASs can be broken down into their smaller components (monocyclic, bicyclic, and tricyclic), and the magnetic properties of the larger molecules can be predicted by summing the contributions of these smaller subunits using an additivity scheme.[24–26]

For the particular case of the PBHs, molecular properties can be predicted by the type and order of the tricyclic components themselves, where the two tricyclic subunits differ only in their annulation: linear or angular, i.e., anthracene or phenanthrene, respectively. This conclusion allows for a reduction of the molecular structure to the sequence of tricyclic subunits (i.e., the annulation sequence). We have formulated this sequence as a textual representation of the molecule (Figure 1a), containing only the characters "L", "A", "(" and

“)” (parentheses are used to denote branching points, where applicable; see Figure 1b for a selection of PBHs and their respective annulation sequences). The resulting names are strings of varying lengths comprising the letters “L” and “A”, which we have accordingly named “LALA Strings” or “LALAS” (the terms “LALAS representations”, “LALAS”, and “annulation sequences” are interchangeable).

The annulation sequence, or LALAS, has been clearly demonstrated to be linked to molecular properties: molecules sharing the same annulation sequence are equiaromatic (i.e., the same aromatic behavior) in both the ground state and the lowest excited triplet state.[27] In addition, we have shown that the annulation sequences themselves demonstrate a clear connection to and enable prediction of numerous molecular properties, including relative stability, aromatic character, singlet-triplet energy gaps, and location of spin density in the triplet excited state.[27]

The generation of a LALAS for a given molecule proceeds according to the following protocol (similar to IUPAC rules for naming branched alkanes), which we have automated in a modified version of Predi-XY.[28] The modified code for generating the LALAS is freely accessible online.

a. For unbranched molecules, each tricyclic subcomponent is denoted as a letter “L” or “A”, depending on the type of annulation. The choice of “left-to-right” or “right-to-left” is random, i.e., each molecule has (at least) two valid LALAS. E.g., the molecule LLA (Figure 1B) can also be read as ALL.
b. For branched molecules, we search for the longest possible path through the molecule, and denote this the “main branch”. E.g., the main branch of molecule LLA()is a chain of 5 rings.
c. If there are branching points, they are denoted with “()” (e.g., LLA() in Figure 1B). Note that branching points will always follow an “A”, as they are by necessity connected to the middle ring of an angular annulation. Note, also, that the notation “()” implies a branch containing a single ring.
d. Branches longer than a single ring will have their own sequence, which will be detailed within the parentheses (e.g., LAA(L)LL in Figure 1B).
e. If there are two different paths of similar length, the one with more branching points is chosen as the main branch.

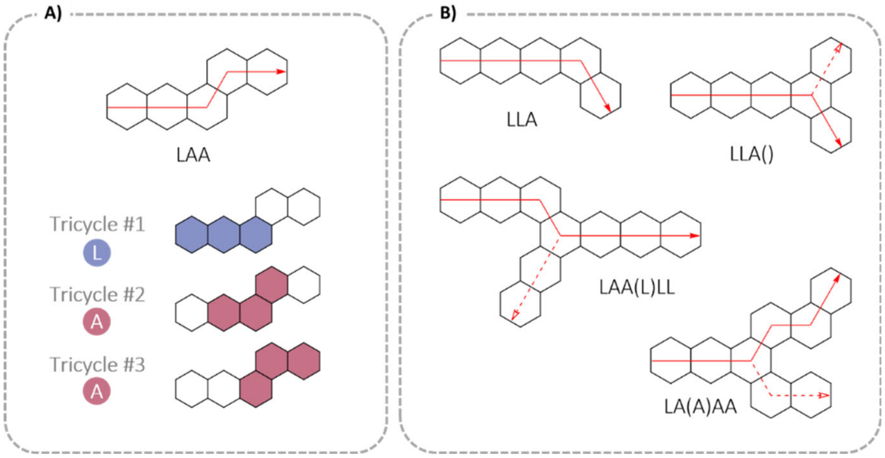


Figure 1: A) Illustration of the LALAS for a simple pentacyclic PBH (left). B) examples of selected PBHs and their LALAS (right). The direction of reading along the main branches that is consistent with the given name is shown with solid red arrows; additional branches are shown with dashed red arrows. Double bonds were omitted for clarity.

We emphasize that, in contrast to SMILES or SELFIES, which describe molecules on an atom-connectivity basis, LALAS describe molecules using ring-based subunits. As such, they reduce the dimensionality of the molecular representation, while retaining important structural information. This trait could allow for significant improvement in efficiency of training ML models –in reducing both the training time and the required training set size. We also note that several graph-theoretical based notations for PBHs have been previously proposed, most notably by Gutman,[29] Balaban,[30–32] and Cyvin.[33] To the best of our knowledge, these have been used mostly for enumeration of isomers of various types of PBHs. The 3-digit code developed by Balaban in the 1960s, which is the most similar in approach to our own formulation, has been also used to identify correlations to molecular properties (e.g., ionization potential, IP, and electron affinity, EA).[34,35] In this work, LALAS representations were generated using a modified version of the Predi-XY code developed in our group[28] and were used in two ways: a) tokenization directly from the string format (LALAS) and b) as a basis for generating a LALA-based feature vector (LFV) for each molecule (*vide infra*).

Data Sources

With the advent of more efficient computational techniques, data-driven investigations have become increasingly common; however, it has been difficult to apply such methods to PBHs, as there is a paucity of suitable data. Recently, our group reported on the COMPAS Project: the construction of a novel COMputational database of Polycyclic Aromatic Systems.[36] The first instalment of the database, denoted COMPAS-1D, contains data on ~8,600 *cata*-condensed PBHs, including their optimized structures and a selection of electronic properties (calculated with DFT at the B3LYP-D3BJ/def2-svp level), as well as their respective SMILES representations and LALAS representations.

For the current study, we removed benzene and naphthalene from the dataset, as they are too short to have a LALAS. Both LALAS and SMILES representations were tokenized using two methods: one-hot[37] and CUSTODI.[23]

The properties we extracted from the COMPAS-1D database for this study were: a) HOMO energy; b) LUMO energy; c) HOMO-LUMO gap; d) adiabatic ionization potential (AIP); e) adiabatic electron affinity (AEA); f) relative single-point energy.

LALA Feature-Vector (LFV)

In addition to the LALAS, we generated for each molecule a feature-vector based on curated structural features derived from the LALAS, denoted LFV. This set of chemically intuitive features (detailed in Table 1) was inspired by our collective experience studying PBHs and by structure-property relationships previously found in smaller datasets.[24,27] The purpose of using the LFV as input was threefold: (1) to validate the intuition we developed previously, (2) to check the predictive power of these descriptors, and (3) to compare the conclusions derived from this set of PBH-specific features to those derived from more general chemical representation.

Table 1. List of curated structural features extracted from the LALAS of each molecule.

feature #	Description
1	Longest linear stretch
2	Number of rings
3	Ratio of "L" tokens in total LALAS
4	Number of branching points
5	Longest linear stretch degeneracy
6	Longest angular stretch
7	Second longest linear stretch
8	Number of "LAL" subsequences

The CUSTODI Framework

CUSTODI is a recently developed tokenization technique for text-based molecular representations. A full description of the method is beyond the scope of this report. In brief, the approach of CUSTODI is to find, using linear regression, the best fitting tokenization dictionary for a given target property. The resulting dictionary can be used for tokenization (CUSTODI representation) or for prediction (CUSTODI model), as shown in Figure 2. Both the CUSTODI representation and the CUSTODI model were used in this work. For further details on the method, the reader is referred to reference [23].

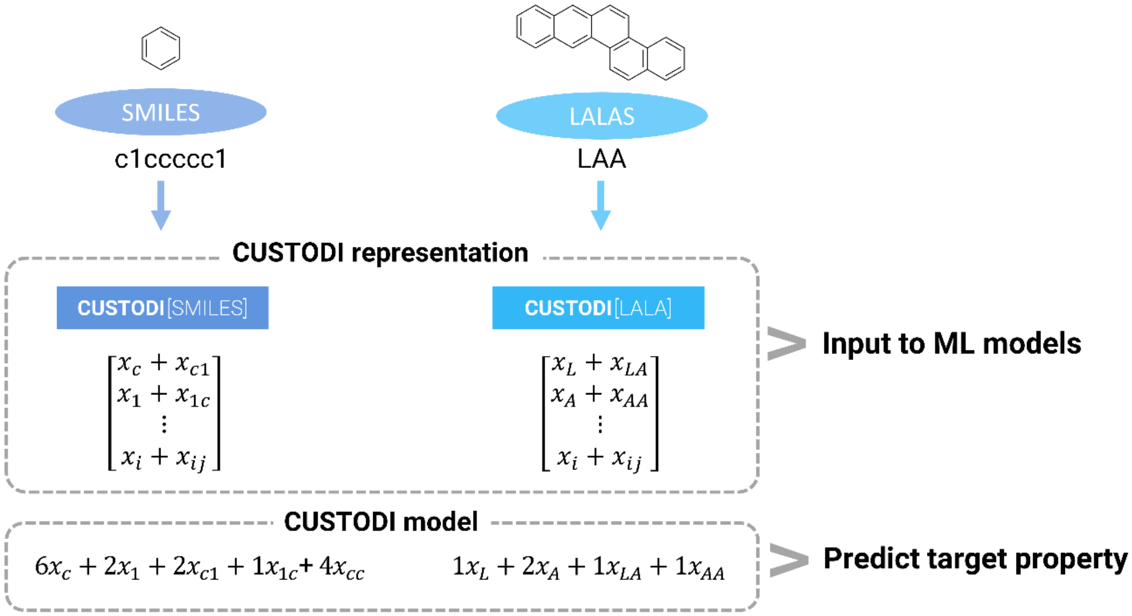

SMILES
c1ccccc1
LALAS
LAA
CUSTODI representation
CUSTODI[SMILES]
CUSTODI[LALA]
$\begin{bmatrix} x_c + x_{c1} \\ x_1 + x_{1c} \\ \vdots \\ x_i + x_{ij} \end{bmatrix}$
$\begin{bmatrix} x_L + x_{LA} \\ x_A + x_{AA} \\ \vdots \\ x_i + x_{ij} \end{bmatrix}$
Input to ML models
CUSTODI model
$6x_c + 2x_1 + 2x_{c1} + 1x_{1c} + 4x_{cc}$
$1x_L + 2x_A + 1x_{LA} + 1x_{AA}$
Predict target property

Figure 2: Schematic illustration of the use of SMILES and LALAS to create a CUSTODI representation and with the CUSTODI model.

To perform a methodical comparison, four supervised learning models were used, ranging from standard to state-of-the-art (for further details on each model, please refer to Section S4 in the Supporting Information). Kernel ridge regression (KRR) and random forest (RF) were used in conjunction with CUSTODI tokenization and LALA features as input (denoted as CUSTODI[LALAS]); a recurrent neural network (RNN) was trained on one-hot tokenization of the LALAS and SMILES (denoted One-Hot[SMILES] and One-Hot[LALAS], respectively); and a state-of-the-art[38,39] graph convolution (GC) model was used with the MolConv input.[40] The KRR and RF models were implemented using scikit-learn,[41] the RNN model by using tensorflow,[42] and the GC model by using DeepChem[43] with an identical architecture as the MoleculeNet benchmark.[38]

The data was split into training and testing sets and hyperparameter optimization was performed using Bayesian optimization algorithm (Gaussian process) as implemented in the scikit-optimize python package.[44] Model selection was done using 5-fold cross validation. Exact details on the hyperparameters of each model are in Section S4.2 in the Supporting Information. The best model was retrained on the whole training set and used to estimate the model's performance. All properties were normalized using z-score normalization (0 mean and 1 standard deviation) and all tokenized strings were padded before insertion into the models.

Interpretation of CUSTODI

The interpretation of CUSTODI is relatively straightforward: each tokenization value corresponds to a substring (e.g., atom or functional group), and these values are used to make the model's prediction (Eq. 1). In this work, each tokenization value corresponds to a tricyclic substructure within the PBH.

$$\text{bias} + \sum_{c_i \in s_i} \sum_{k=1}^{\text{degree}} n_{c_i \dots c_{i+k-1}} x_{c_i \dots c_{i+k-1}} = p_i \tag{1}$$

Where s_i is the string representation of molecule i in the database, c_i is the ith character in s_i, $n_{c_i \dots c_{i+k-1}}$ is the number of occurrences of the substring $c_i \dots c_{i+k-1}$ in s_i, $x_{c_i \dots c_{i+k-1}}$ is the substring's tokenization value and p_i is the target property. From Eq. 1, the tokenization value is proportional to the significance of the represented substructure for a given property. The tokenization value x is not independent of the number of occurrences n, and there is actually an inverse proportion between them. To account for this proportion, the importance of each substring is given by

$$\beta_{c_i \dots c_{i+k-1}} = \frac{n_{c_i \dots c_{i+k-1}} x_{c_i \dots c_{i+k-1}}}{\text{bias} + \sum_{c_i \in s_i} \sum_{k=1}^{\text{degree}} n_{c_i \dots c_{i+k-1}} x_{c_i \dots c_{i+k-1}}} \tag{2}$$

So that the sum of all the importance terms is 1.

We emphasize that the analysis made here can be repeated for many chemical compounds and can produce similar intuition on the effects of various functional groups on properties. Unlike previous reports (a few are detailed in the Introduction, *vide supra*), CUSTODI does not rely on hand-crafted features. By iterating over all possible substrings, CUSTODI in essence performs its own data-driven feature-engineering. The main advantages are that this does not require any chemical intuition and tests all substructures in the dataset simultaneously. As a result, this reduces possible sources of bias and allows for identification of features that might not be obvious to experts. The disadvantages are that CUSTODI cannot search for varying-length substrings and will likely not identify features that involve non-adjacent structural components.

Results and Discussion

The two main aspects of the work are presented and discussed in the following sections: a) the performance of LALAS as a molecular representation and b) the use of LALAS in conjunction with interpretable ML models (CUSTODI and RF) to gain new chemical insights into PBHs.

The Performance of LALAS

As mentioned above, LALAS are specifically tailored to describe PBH compounds. To test the added value of this dedicated representation versus commonly used general-purpose representations, the performance of several models trained on LALAS was compared to the same models trained on other types of input (see Methods for further details on the selected models for comparison). The models employed are detailed in the Methods section and in Figure 3, which provides an illustration of all input+model combinations.

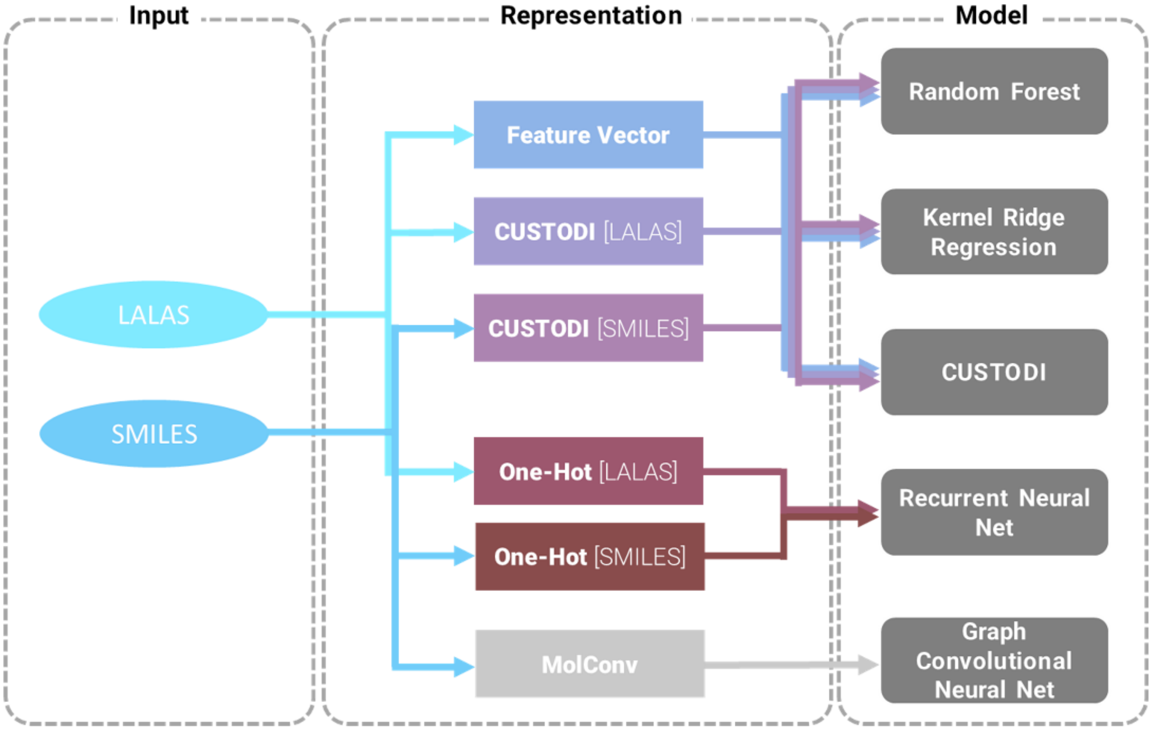


Figure 3. Graphical illustration of the various molecular representations and model combinations used in this work.

The results obtained with a training set containing 7,674 molecules (90%) are illustrated in Figure 4 (the full fit results on the database are in Section S6 in the Supporting Information).

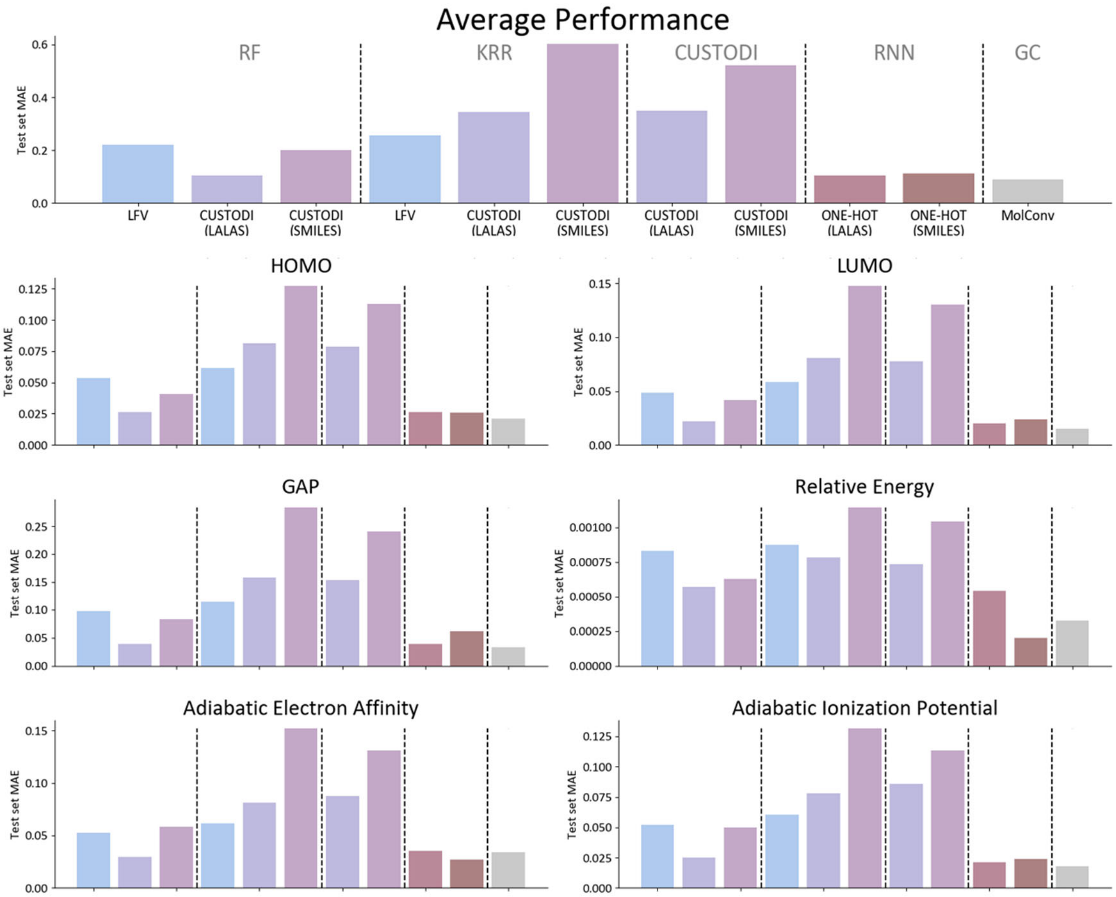


Figure 4: top: average test set MAE (average of MAE/standard deviation) for all the tested model-representation pairs. Bottom: test set MAE of all model-representation

As seen in the averaged performance plot (Figure 4, top row), the RNN model had the best performance of all models trained on the various LALA-based representations. We also observe that the RNN model is the only one in which both One-Hot[LALAS] and One-Hot[SMILES] performed comparably well (and, surprisingly, almost as well as the GC model).

In all other cases, the CUSTODI[LALAS] representation performed markedly better than the CUSTODI[SMILES] representation. What is more, when using CUSTODI[LALAS] as input, the RF model performed similarly to the best-performing RNN and GC models, which are considered much more sophisticated. A possible explanation is that the simpler syntax of LALAS better suits the linear approximation in CUSTODI, thus allowing for better tokenization dictionaries to be generated for LALA, as compared to SMILES.

The LFV did not show consistent behavior as an input: the RF model trained on LFVs performed the poorest; however, the KRR model trained on LFVs performed better than both CUSTODI[SMILES] and CUSTODI[LALAS]. This variance in performance is not surprising, as the RF and KRR models work best on significantly different latent spaces.

A visual inspection of the individual plots in Figure 4 indicates that the relative energy is the only property with a qualitatively different picture. For this property, it appears that there is a dramatic difference in the performance of the two RNN models, and the performance of RNN model trained on the One-Hot[LALAS] representation appears to be noticeably poorer than the model trained on the One-Hot[SMILES] representation. We must emphasize, however, that such an interpretation is misleading, considering that, in fact, all the models show very satisfactory accuracy: they predict the relative electronic energy with MAE < 0.002 eV, which is smaller than the margin of error of the DFT calculations.

Nevertheless, this apparent shift in performance (relative to the other molecular properties) led us to consider possible differences between the representations, which might affect the prediction of relative energy. One important difference is the way LALAS treat angular annulations. Angular annulations can have two types of direction – clockwise and counter-clockwise. Consecutive angular annulations in opposite directions create a zig-zag type of topology, which is planar in the ground state.[45,46] However, consecutive angular annulations in the same direction create cove, fjord, and eventually helix formations (for two, three, and four consecutive A annulations, respectively). These differences do not necessarily affect electronic properties (e.g., molecules with similar annulation sequences are equiaromatic – i.e., have similar aromaticity patterns – regardless of the direction of the A annulations), however such substructures can affect relative energy as they have an increasing degree of curvature, which introduces helical strain, i.e., higher relative energy. Whereas SMILES representations include this information, LALAS do not differentiate between the types of angular annulations. Therefore, in principle, there could be performance discrepancies between the two; in practice, we observe that both perform exceedingly well on the given data.

Having analyzed the performance of the individual models in terms of prediction accuracy, we now turn to comparing the training time required for each of the models. Table S2 (Supporting Information, Section S2) gives the average time per molecule for each of the models. In general, we find that using the LALAS (or representations derived from LALAS) markedly decreases training time for the RF and RNN models (by factors of ~6 and ~5, respectively), and moderately decreases training time for the KRR and CUSTODI models (by a factor of ~2 for both).

Finally, a major advantage of LALAS is revealed when comparing the performance of models trained on smaller training sets. The top four best-performing input+model combinations were identified from Figure 4, and new models were trained on varying training-set sizes. Both LALAS and LFVs indeed show superior performance in small datasets compared to other tested methods. Using only 10% of the data for training, the RNN model trained on One-Hot[LALAS] achieved a normalized test set MAE of 0.12 eV, which is markedly lower than the other three models. At a training-set size of 40%, GC achieved similar results as the RNN, MAE = 0.11 eV, and at 70% all four models showed comparable results. This may be attributed to the concise nature of the LALAS: the LALAS of a given molecule is, on average, shorter by 86.5% (55 characters) than the SMILES of the same molecule. In addition, the complexity of the language is substantially reduced – only four types of characters. Thus, lower variance is expected for models trained on LALAS.

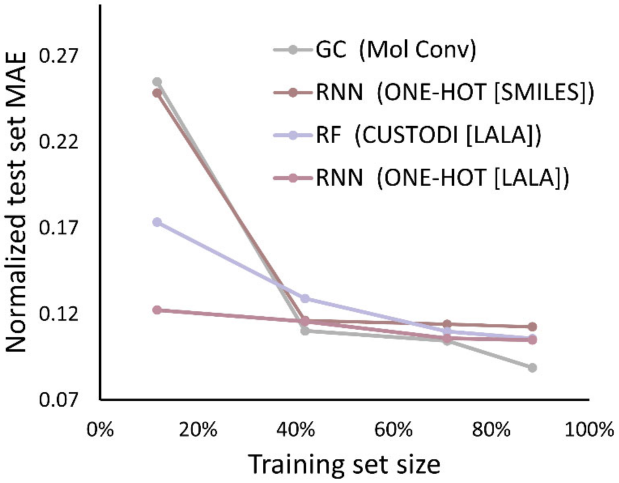


Figure 5: Training set size dependence for the top-4 performing models.

Interpretation Based on Annulation Sequence

As mentioned above, the simplification that is inherent in LALAS affects not only model performance, but also interpretability, which is a main goal of this work. Whereas single atoms or atom-pairs can have meaning as functional groups in many organic molecules, in PBHs individual carbon atoms often do not carry significant chemical meaning. LALAS connect textual characters with chemically meaningful subunits, i.e., specific ring annulation patterns. This makes it amenable to interpretation when used for training text-based models such as CUSTODI (the methodology for interpreting the CUSTODI model dictionary is presented in the Methods section).

To extract the most meaningful insights from a given model, one should first ensure that the model shows good and reliable performance. Therefore, we initially performed a benchmarking procedure, to determine the optimal degree of CUSTODI for these data. This procedure is included in the hyperparameter optimization of the CUSTODI model, as the degree of CUSTODI is a hyperparameter of the model (see Methods for details). In other words: CUSTODI-1 was trained on subsequences of a single character (e.g., "L", ")"), CUSTODI-2 was trained subsequences containing either one or two characters (e.g., "L", "LA"), and CUSTODI-3 was trained on subsequences containing either one, two or three characters (e.g., "A", "(L", "ALA"). The best-performing model was found to be CUSTODI-2. The importance terms of the trained CUSTODI-2 model are presented in Figure 6. We emphasize that, while these terms can help assign the importance of the various structural features, they do not tell us in which way the features impact each property, i.e., increasing or decreasing the value of the predicted property. Such an analysis requires different treatment, which is the subject of ongoing work and will be disclosed in due course.

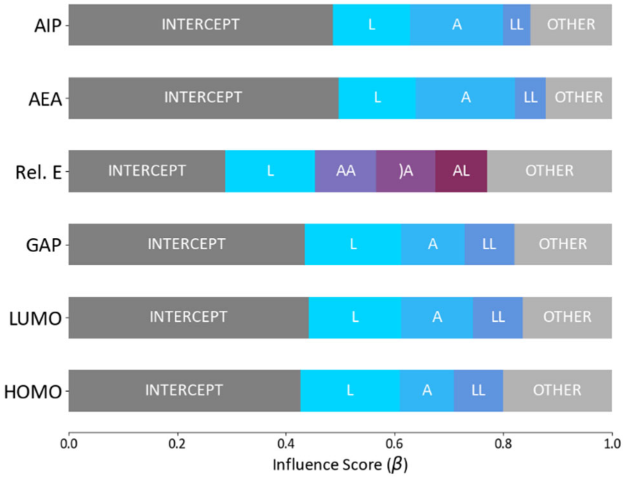


Figure 6: Importance terms (β) for substrings of LALAS for each property in the database. Only substrings with importance greater than 5% are displayed. Importance is based on CUSTODI model trained on 88% of the data.

Figure 6 shows that the properties HOMO, LUMO, and HOMO-LUMO gap have a similar dependence on particular substring sequences, which is not surprising. In addition, we observe a marked difference between the relative importance of the factors governing these three properties and those determining the relative energy of each molecule (note: the relative energy is calculated with respect to the respective lowest-energy isomer; for further details see [36]). The adiabatic ionization potential (AIP) and adiabatic electron affinity (AEA) have some similarity to the three aforementioned electronic properties, which is in accordance with Koopman's theorem.[47] Yet, there are also dissimilarities, which demonstrate that the model is capable of distinguishing between the property types.

The main factor influencing the HOMO, LUMO, and HOMO-LUMO gap is the presence of linear annulations (L, $\bar{\beta} = 17.7\%$) and stretches of two consecutive linear annulations (LL, $\bar{\beta} = 9.1\%$; i.e., four benzene rings annulated linearly, akin to naphthacene). These properties are affected by the presence of angular annulations to a lesser extent (A, $\bar{\beta} = 11.6\%$), while the existence of branching points does not seem to be important. Our recent analysis of the COMPAS-1D dataset showed that the HOMO, LUMO, and HOMO-LUMO gap all depend on the length of the Longest L subsequence. Because CUSTODI-2 only looks as subsequences up to two letters long, we cannot see here the importance of longer Longest L subsequences. Nevertheless, all of these observations are in line with our previous observations on these compounds.[36]

In contrast, the main factors influencing the relative energy are different. We observe the following dependencies: linear annulations (L, $\beta = 16.5\%$), consecutive series of angular annulations (AA, $\beta = 11.3\%$), and branching points following an angular annulation (")A", $\beta = 10.8\%$). The subsequence "AL" also appears, which implies that it is not only the presence of angular annulations that matters, but also what surrounds them, or at what point the A sequence is broken. These results are in line with our previous observations pertaining to prediction of the relative energy, which we attributed to the strain that is incurred by sequences of consecutive A motifs. Specifically, we noted that consecutive sequences of A annulations can be either helical or planar, depending on the direction of the consecutive As. While A annulations in opposing directions lead to “zig-zag” formation that is planar, stretches of A annulations in the same direction lead to the formation of helical structures (known as cove, fjord, and helix). The features entail helical strain which raises the relative energy. Therefore, it is not surprising to find them among the main influencers in the prediction of relative energy. Corroboration for this interpretation can be found in our analysis of the COMPAS-1D dataset, which has shown that the increase in relative electronic energy is correlated to the deviation from planarity.[36]

In this context, we note that the relationship between angular annulations and stability has also been investigated with other computational and conceptual tools. For example, the same observations can be interpreted in the context of Clar’s rule,[48,49] which states that isomers with a larger number of Clar sextets are more stable than those with fewer Clar sextets. In general, angular annulations and branching points allow for more Clar sextets to be generated, which can therefore influence the relative energy. We are currently investigating the link between Clar structures, aromaticity indices, and the relative energy, to see if this interpretation can be substantiated. Other computational analyses have also rationalized the greater stability of angular isomers in the ground state via graph-theory,[50] additional π-bonding,[51,52] and a greater number of resonance structures.[53]

Though the L motifs are predicted by the model to have an importance effect, the direction of this effect is unknown. Hence, it can, in principle, be perceived in two ways: a) following the previous rationalization, the presence of L motifs can be seen as precluding the formation of such non-planar motifs and therefore contributing to stabilization; or b) the L motifs may contribute to destabilization, not via geometric deformation but rather through an electronic effect. Since it is well-established that the most stable isomers are the phenacenes (i.e., the “zig-zag” PBHs),[51,52] one may conclude based on this previous knowledge that the operative case is (b). Nevertheless, we are currently working on implementation of more sophisticated DL models that also reveal the direction of each feature’s influence.

As mentioned above, the AEA and AIP mostly show similarity to the HOMO, LUMO, and HOMO-LUMO gap analyses, with some exceptions. The main difference is that for both AIP and AEA the angular annulation (“A”, $\bar{\beta} = 17.7\%$) shows slightly greater importance than the linear annulation (“L”, $\bar{\beta} = 14.2\%$). One possible explanation can be found in the work of Khatymov et al., who found that the stabilization of the LUMO is hampered due to specific symmetry features in the angular phenanthrene, which may be generalized to homologous series of angularly annulated PBHs.[54] As a result, within Koopmans’ theorem (though just a crude approximation for our DFT-calculated values), the EA is expected to decrease in magnitude. An alternative, or complementary, explanation is that many of the molecules containing multiple A annulations have some degree of helicity, which may affect the charge delocalization. Therefore, the presence of As becomes an important factor for the predictive model.

We note that, for all properties, the intercept has a large importance value, i.e., a large influence on the predicted value. As described in the Methods section, the intercept is a constant value that describes the bias of the CUSTODI model. In cases where the bias itself has a large value, relative to the individual tokenization values, the intercept has a strong influence. This can be understood in the following way: the CUSTODI model learns the “average value” of a property and the importance assigned to each of the subsequences represents the effect of the respective subsequence on that relative value.

Interpretation Based on the LFV

The RF model has an inherent way of finding feature importance.[55] Our analysis focuses on the RF model trained on LFVs (Figure 7). The results show very similar patterns to those obtained with CUSTODI[LALAS]. Considering that LFVs are essentially domain expertise-based features which we extracted from LALAS, this implies that the CUSTODI model successfully captures the features directly from the textual representations, without the need for human intervention.

The RF model shows that the HOMO, LUMO, and HOMO-LUMO gap are mainly affected by the length of the longest stretch of linear annulations (“Longest L”, 87%). Unsurprisingly, the AIP and AEA are also mainly influenced by the linear annulations (“Longest L”, 80.2%). However, AEA and AIP are also affected by the number of rings, which is in line with previous reports of a size-dependency for these properties.[56] It is generally considered that the larger a conjugated system is, the better it is expected to stabilize excess charge through delocalization.

The relative energy displays a very different set of dependencies, chief among them are the longest stretch of angular annulations (“Longest A”, 27.1%), the number of branches in the molecule (“No. Branching points”, 21.2%), and the degeneracy of the longest linear sequence (“Longest L Degeneracy”, 20.6%). The ratio of L motifs, the longest linear sequence, and the number of LAL sequences also have non-negligible influences (10%, 12.1%, and 6.4%, respectively). As mentioned above, we believe that the impact of the angular annulations can be attributed either to variations in helical strain or to the possible number of Clar sextets that can be formed. Similarly, the number of branches is influential because it is related to the tendency to form helical structures (an increase in branches precludes linear stretches and increases the likelihood of angular stretches in similar directions).

As we explained above, we hypothesize that, while the A motifs appear to raise the energy through geometrical deformation, the L motifs raise it via electronic effects. Thus, we observe a dependence also on several features describing the presence of L motifs. As opposed to the other properties, where only the longest linear stretch was important, here also the degeneracy (i.e., the longest stretch that appears more than once) is important. This indicates that the effect of individual linear stretches on the relative energy may be additive, while on other properties it is exclusive. Interestingly, the relative energy also shows a dependence on a specific substructure, “LAL”. This particular subsequence was previously noted as behaving in an anomalous manner[24] in the prediction of magnetic behavior in PBHs.

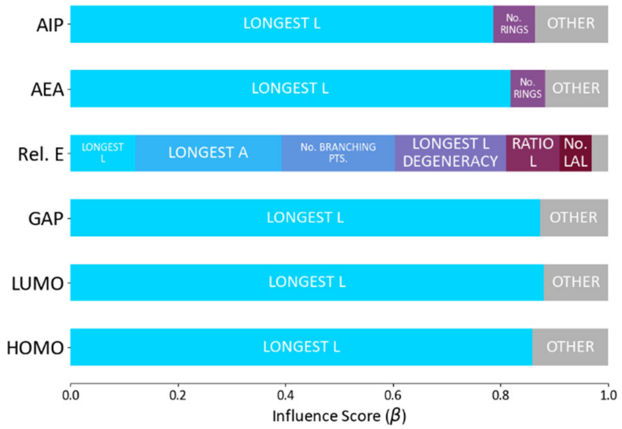


Figure 7: Feature importance for the LALA features. Derived from RF model trained on 88% of the data.

A similar analysis using the CUSTODI model trained on SMILES strings yielded no meaningful results, as the substrings used in CUSTODI models are short (this results from the hyperparameter optimization; see Section S4.2 in the Supporting Information for more details). The results of the influence analysis on SMILES strings are also provided in the Supporting Information (Section S5, Figure S1). Similarly, RF trained on CUSTODI[SMILES] did not afford any interpretable results.

Conclusions

In this work, we applied interpretable ML tools to investigate the structure-property relationships in the family of PBHs, which are archetypal polycyclic species. We introduced a new type of textual molecular representation, which is specifically suited for these molecules. This representation can be used either in string form (LALAS) or as the basis for a feature-vector (LFV). In addition, we applied a new type of interpretable ML method, CUSTODI. Comparison to standard models and input types demonstrated the added value of LALAS to both efficiency and interpretability.

The application of these two new tools to the newly reported database, COMPAS-1D, [ref: database paper] allowed us to gain chemical insight into the structure-property relationships of PBHs. Four main conclusions were reached:

(1) most of the electronic properties of PBHs we studied are primarily influenced by the presence and length of consecutive linear annulations in the molecule;
(2) the relative energy of isomeric PBHs is mainly affected by the presence of angular annulations and branching points in the molecule;
(3) as expected from Koopmans' theorem, AIP and AEA have similar dependencies as HOMO, LUMO, and HOMO-LUMO gap, however, the former two are also size-dependent while the latter appear not to be;
(4) there are "privileged" subsequences, one of which we identified – "LAL".

To a certain extent, (some of) these insights may be considered well-known "rules of thumb" or "conventional wisdom" in the chemical community. However, to the best of our knowledge, have never been demonstrated in a data-driven manner. Indeed, the agreement between the ML interpretation and generally accepted chemical behavior indicates that the models performed reliably well, and we have validated these rules of thumb with an unprecedented dataset containing ~8,700 PBHs. Nevertheless, there are also new insights, such as factors influencing relative energy of PBH isomers and the existence of "privileged" subsequences. We also emphasize that the importance analysis presented here indicated that the relationship between linear sequences and the various molecular properties is different. Specifically, for all of the properties except the relative energy, it appears that only the single longest linear stretch is important and how many times such a sequence appears does not matter; in contrast, for the relative energy, the degeneracy of these sequences does matter, which suggests that they might contribute cumulatively to destabilization.

Importantly, similar conclusions were obtained using the CUSTODI model, which was trained on LALAS without any preprocessing, and the RF model, which was trained on LFVs – domain-expert curated features. This serves to indicate that the CUSTODI model is capable of extracting the important structural features from this new representation automatically, without expert intervention. We emphasize that CUSTODI can be used in a similar manner on different string representations to derive structure-property relationships.

Both the RF and CUSTODI models describe the relative importance of various structural features/subunits, but they could not describe their effect – i.e., increase or decrease in magnitude. Our group is currently exploring the use of additional interpretable algorithms to provide further insight into this, as well as other, aspects. In particular, we are investigating the direct impact of individual structural motifs on different aromaticity indices. Additional emphasis is on the expansion of the LALAS representation concept to include *peri*-condensed and poly(hetero)cyclic aromatic systems and on generating the relevant data to enable further exploration and analysis of this chemical space.

Data and Code Availability

The full code used in this paper appear in our GitLab repository at https://gitlab.com/porannegroup/lalas. The data was taken from the COMPAS Project repository at https://gitlab.com/porannegroup/compas.

Acknowledgments

The authors thank Prof. Dr. Peter Chen for his continued scientific and financial support and Mr. Felix Fleckenstein for technical assistance with developing the modified code for generating LALAS. R.G.P. and A.W. are grateful to the Branco Weiss Fellowship for supporting this research under a Society in Science grant. R.G.P. is a Branco Weiss Fellow and a Horev Fellow.

Conflict of Interest

The authors declare no conflict of interest.

References

(1) Brown, N.; Ertl, P.; Lewis, R.; Luksch, T.; Reker, D.; Schneider, N. Artificial Intelligence in Chemistry and Drug Design. *J. Comput. Aided. Mol. Des.* **2020**, *34* (7), 709–715. https://doi.org/10.1007/s10822-020-00317-x.

(2) Harren, T.; Matter, H.; Hessler, G.; Rarey, M.; Grebner, C. Interpretation of Structure–Activity Relationships in Real-World Drug Design Data Sets Using Explainable Artificial Intelligence. *J. Chem. Inf. Model.* **2022**, *62* (3), 447–462. https://doi.org/10.1021/acs.jcim.1c01263.

(3) Granda, J. M.; Donina, L.; Dragone, V.; Long, D.-L.; Cronin, L. Controlling an Organic Synthesis Robot with Machine Learning to Search for New Reactivity. *Nature* **2018**, *559* (7714), 377–381. https://doi.org/10.1038/s41586-018-0307-8.

(4) Eyke, N. S.; Koscher, B. A.; Jensen, K. F. Toward Machine Learning-Enhanced High-Throughput Experimentation. *Trends in Chemistry* **2021**, *3* (2), 120–132. https://doi.org/10.1016/j.trechm.2020.12.001.

(5) Tao, H.; Wu, T.; Aldeghi, M.; Wu, T. C.; Aspuru-Guzik, A.; Kumacheva, E. Nanoparticle Synthesis Assisted by Machine Learning. *Nat. Rev. Mater.* **2021**, *6* (8), 701–716. https://doi.org/10.1038/s41578-021-00337-5.

(6) Pollice, R.; dos Passos Gomes, G.; Aldeghi, M.; Hickman, R. J.; Krenn, M.; Lavigne, C.; Lindner-D'Addario, M.; Nigam, A.; Ser, C. T.; Yao, Z.; Aspuru-Guzik, A. Data-Driven Strategies for Accelerated Materials Design. *Acc. Chem. Res.* **2021**, *54* (4), 849–860. https://doi.org/10.1021/acs.accounts.0c00785.

(7) Nandy, A.; Duan, C.; Taylor, M. G.; Liu, F.; Steeves, A. H.; Kulik, H. J. Computational Discovery of Transition-Metal Complexes: From High-Throughput Screening to Machine Learning. *Chem. Rev.* **2021**, *121* (16), 9927–10000. https://doi.org/10.1021/acs.chemrev.1c00347.

(8) Keith, J. A.; Vassilev-Galindo, V.; Cheng, B.; Chmiela, S.; Gastegger, M.; Müller, K.-R.; Tkatchenko, A. Combining Machine Learning and Computational Chemistry for Predictive Insights Into Chemical Systems. *Chem. Rev.* **2021**, *121* (16), 9816–9872. https://doi.org/10.1021/acs.chemrev.1c00107.

(9) Kovalerchuk, B.; Ahmad, M. A.; Teredesai, A. Survey of Explainable Machine Learning with Visual and Granular Methods Beyond Quasi-Explanations. In *Interpretable Artificial Intelligence: A Perspective of Granular Computing*; Pedrycz, W., Chen, S.-M., Eds.; Studies in Computational Intelligence; Springer International Publishing: Cham, 2021; pp 217–267. https://doi.org/10.1007/978-3-030-64949-4_8.

(10) George, J.; Hautier, G. Chemist versus Machine: Traditional Knowledge versus Machine Learning Techniques. *Trends in Chemistry* **2020**. https://doi.org/10.1016/j.trechm.2020.10.007.

(11) Friederich, P.; Krenn, M.; Tamblyn, I.; Aspuru-Guzik, A. Scientific Intuition Inspired by Machine Learning Generated Hypotheses. *arXiv:2010.14236 [physics, physics:quant-ph]* **2020**.

(12) Roscher, R.; Bohn, B.; Duarte, M. F.; Garcke, J. Explainable Machine Learning for Scientific Insights and Discoveries. *IEEE Access* **2020**, *8*, 42200–42216. https://doi.org/10.1109/ACCESS.2020.2976199.

(13) Rodríguez-Pérez, R.; Bajorath, J. Explainable Machine Learning for Property Predictions in Compound Optimization. *J. Med. Chem.* **2021**, *64* (24), 17744–17752. https://doi.org/10.1021/acs.jmedchem.1c01789.

(14) Häse, F.; Fdez. Galván, I.; Aspuru-Guzik, A.; Lindh, R.; Vacher, M. How Machine Learning Can Assist the Interpretation of *Ab Initio* Molecular Dynamics Simulations and Conceptual Understanding of Chemistry. *Chem. Sci.* **2019**, *10* (8), 2298–2307. https://doi.org/10.1039/C8SC04516J.

(15) Straif, K.; Baan, R.; Grosse, Y.; Secretan, B.; Ghissassi, F. E.; Cogliano, V.; Drewski, D.; Partanen, T.; Vähäkangas, K.; Stücker, I.; Borlak, J.; Feron, V. J.; Marques, M. M.; Gustavsson, P.; Fletcher, A.; Arey, J.; Beland, F. A.; Burchiel, S.; Flowers, L.; Herbert, R. A.; Mukhtar, H.; Nesnow, S.; Penning, T. M.; Sinha, R.; Shimada, T. Carcinogenicity of Polycyclic Aromatic Hydrocarbons. *The Lancet Oncology* **2005**, *6* (12), 931–932. https://doi.org/10.1016/S1470-2045(05)70458-7.

(16) Abdel-Shafy, H. I.; Mansour, M. S. M. A Review on Polycyclic Aromatic Hydrocarbons: Source, Environmental Impact, Effect on Human Health and Remediation. *Egyptian Journal of Petroleum* **2016**, *25* (1), 107–123. https://doi.org/10.1016/j.ejpe.2015.03.011.

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

(18) Figueira-Duarte, T. M.; Muellen, Klaus. Pyrene-Based Materials for Organic Electronics. *Chem. Rev. (Washington, DC, U. S.)* **2011**, *111*, 7260–7314. https://doi.org/10.1021/cr100428a.

(19) Al Ruzaiqi, A.; Okamoto, H.; Kubozono, Y.; Zschieschang, U.; Klauk, H.; Baran, P.; Gleskova, H. Low-Voltage Organic Thin-Film Transistors Based on [n]Phenacenes. *Organic Electronics* **2019**, *73*, 286–291. https://doi.org/10.1016/j.orgel.2019.06.021.

(20) Tönshoff, C.; Bettinger, H. F. Pushing the Limits of Acene Chemistry: The Recent Surge of Large Acenes. *Chem. Eur. J. 27, 3193*. https://doi.org/10.1002/chem.202003112.

(21) Drummer, M. C.; Singh, V.; Gupta, N.; Gesiorski, J. L.; Weerasooriya, R. B.; Glusac, K. D. Photophysics of Nanographenes: From Polycyclic Aromatic Hydrocarbons to Graphene Nanoribbons. *Photosynth. Res.* **2022**, *151* (2), 163–184. https://doi.org/10.1007/s11120-021-00838-y.

(22) Randić, M. Benzenoid Rings Resonance Energies and Local Aromaticity of Benzenoid Hydrocarbons. *J. Comp. Chem.* **2019**, *40* (5), 753–762. https://doi.org/10.1002/jcc.25760.

(23) Fite, S.; Nitecki, O.; Gross, Z. Custom Tokenization Dictionary, CUSTODI: A General, Fast, and Reversible Data-Driven Representation and Regressor. *J. Chem. Inf. Model.* **2021**, *61* (7), 3285–3291. https://doi.org/10.1021/acs.jcim.1c00563.

(24) Paenurk, E.; Feusi, S.; Gershoni-Poranne, R. Predicting Bond-Currents in Polybenzenoid Hydrocarbons with an Additivity Scheme. *J. Chem. Phys.* **2021**, *154* (2), 024110. https://doi.org/10.1063/5.0038292.

(25) Gershoni-Poranne, R. Piecing It Together: An Additivity Scheme for Aromaticity Using NICS-XY-Scans. *Chem. Eur. J.* **2018**, *24* (16), 4165–4172. https://doi.org/10.1002/chem.201705407.

(26) Finkelstein, P.; Gershoni-Poranne, R. An Additivity Scheme for Aromaticity: The Heteroatom Case. *ChemPhysChem* **2019**, *20*, 1508–1520. https://doi.org/10.1002/cphc.201900128.

(27) 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*, 1–14. https://doi.org/10.1002/chem.202100464.

(28) Wahab, A.; Fleckenstein, F.; Feusi, S.; Gershoni-Poranne, R. Predi-XY: A Python Program for Automated Generation of NICS-XY-Scans Based on an Additivity Scheme. *Electron. Struct.* **2020**, *2*, 047002. https://doi.org/10.1088/2516-1075/abd081.

(29) Gutman, I.; BokoviC', R. Topological Properties of Benzenoid Systems. 7.

(30) Balaban, A. T.; Harary, F. Chemical Graphs—V: Enumeration and Proposed Nomenclature of Benzenoid Cata-Condensed Polycyclic Aromatic Hydrocarbons. *Tetrahedron* **1968**, *24* (6), 2505–2516. https://doi.org/10.1016/S0040-4020(01)82523-0.

(31) Balaban, A. T. Chemical Graphs—VII: Proposed Nomenclature of Branched Cata-Condensed Benzenoid Polycyclic Hydrocarbons. *Tetrahedron* **1969**, *25* (15), 2949–2956. https://doi.org/10.1016/S0040-4020(01)82827-1.

(32) Balaban, A. T. Challenging problems involving benzenoid polycyclics and related systems. *Pure and Applied Chemistry* **1982**, *54* (5), 1075–1096. https://doi.org/10.1351/pac198254051075.

(33) Cyvin, S. J.; Brunvoll, J.; Cyvin, B. N. Formulas and Numbers of Isomers for Benzenoid Hydrocarbons. *Polycyclic Aromatic Compounds* **1997**, *12* (3), 201–212. https://doi.org/10.1080/10406639708233836.

(34) Balaban, A. T.; Pompe, M. QSPR for Physical Properties of Cata-Condensed Benzenoids Using Two Simple Dualist-Based Descriptors. *J. Phys. Chem. A* **2007**, *111* (12), 2448–2454. https://doi.org/10.1021/jp068743f.

(35) Randić, M.; Balaban, A. T. Ring Signatures for Benzenoids with up to Seven Rings, Part 1: Catacondensed Systems. *Int. J. Quant. Chem.* **2008**, *108* (5), 865–897. https://doi.org/10.1002/qua.21578.

(36) 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. **2022**. https://doi.org/10.26434/chemrxiv-2022-2l1m9.

(37) Brownlee, J. Ordinal and One-Hot Encodings for Categorical Data. *Machinelearningmastery*.

(38) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: A Benchmark for Molecular Machine Learning. *Chem. Sci.* **2018**, *9* (2), 513–530. https://doi.org/10.1039/C7SC02664A.

(39) Faber, F. A.; Hutchison, L.; Huang, B.; Gilmer, J.; Schoenholz, S. S.; Dahl, G. E.; Vinyals, O.; Kearnes, S.; Riley, P. F.; von Lilienfeld, O. A. Prediction Errors of Molecular Machine Learning Models Lower than Hybrid DFT Error. *J. Chem. Theory Comput.* **2017**, *13* (11), 5255–5264. https://doi.org/10.1021/acs.jctc.7b00577.

(40) Duvenaud, D.; Maclaurin, D.; Aguilera-Iparraguirre, J.; Gómez-Bombarelli, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. *arXiv:1509.09292 [cs, stat]* **2015**.

(41) *Scikit-Learn: Machine Learning in Python, Pedregosa et al., JMLR 12, Pp. 2825-2830, 2011.*

(42) *(2016). Tensorflow: A System for Large-Scale Machine Learning. In 12th Symposium on Operating Systems Design and Implementation (Pp. 265–283).*

(43) *DeepChem: Deep-Learning Models for Drug Discovery and Quantum Chemistry; Http://Github.Com/Deepchem/ Deepchem; Accesss 2021-04-07.*

(44) Head, T.; MechCoder; Louppe, G.; Iaroslav Shcherbatyi; Fcharras; Zé Vinícius; Cmmalone; Schröder, C.; Nel215; Campos, N.; Young, T.; Cereda, S.; Fan, T.; Rene-Rex; Kejia (KJ) Shi; Schwabedal, J.; Carlosdanielcsantos; Hvass-Labs; Pak, M.; SoManyUsernamesTaken; Callaway, F.; Estève, L.; Besson, L.; Cherti, M.; Karlson Pfannschmidt; Linzberger, F.; Cauet, C.; Gut, A.; Mueller, A.; Fabisch, A. *Scikit-Optimize/Scikit-Optimize: V0.5.2*; Zenodo, 2018. https://doi.org/10.5281/ZENODO.1207017.

(45) Portella, G.; Poater, J.; Bofill, J. M.; Alemany, P.; Sola, M. Local Aromaticity of [n]Acenes, [n]Phenacenes, and [n]Helicenes (n = 1-9). *J. Org. Chem.* **2005**, *70* (7), 2509–2521. https://doi.org/10.1021/jo0480388.

(46) Pino-Rios, R.; Báez-Grez, R.; Solà, M. Acenes and Phenacenes in Their Lowest-Lying Triplet States. Does Kinked Remain More Stable than Straight? *Phys. Chem. Chem. Phys.* **2021**. https://doi.org/10.1039/D1CP01441B.

(47) Koopmans, T. Über die Zuordnung von Wellenfunktionen und Eigenwerten zu den Einzelnen Elektronen Eines Atoms. *Physica* **1934**, *1* (1–6), 104–113. https://doi.org/10.1016/S0031-8914(34)90011-2.

(48) Clar, E. *Aromatic Sextet*; New York, Wiley, 1972.

(49) Solà, M. Forty Years of Clar's Aromatic π-Sextet Rule. *Front. Chem.* **2013**, *1*. https://doi.org/10.3389/fchem.2013.00022.

(50) Aihara, J. Reduced HOMO-LUMO Gap as an Index of Kinetic Stability for Polycyclic Aromatic Hydrocarbons. *J. Phys. Chem. A, 103* (37), 7487–7495. https://doi.org/10.1021/jp990092i.

(51) Poater, J.; Visser, R.; Solà, M.; Bickelhaupt, F. M. Polycyclic Benzenoids: Why Kinked Is More Stable than Straight. *J. Org. Chem.* **2007**, *72* (4), 1134–1142. https://doi.org/10.1021/jo061637p.

(52) Poater, J.; Duran, M.; Solà, M. Aromaticity Determines the Relative Stability of Kinked vs. Straight Topologies in Polycyclic Aromatic Hydrocarbons. *Front. Chem.* **2018**, *6*. https://doi.org/10.3389/fchem.2018.00561.

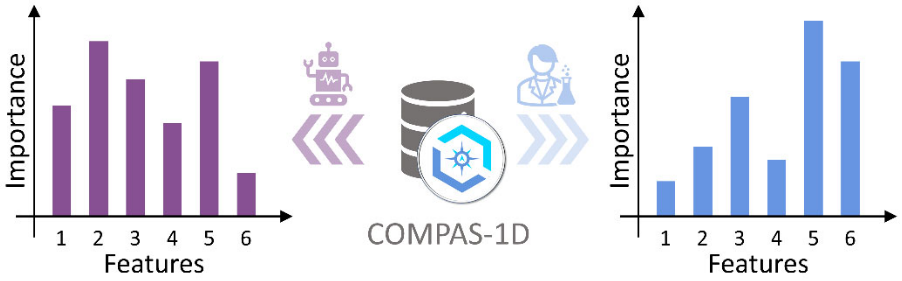
(53) Dias, J. R. Conjugation, Number of Dewar Resonance Structures (DSs) in Homologous Polyzethrene and Related Conjugated Polycyclic Hydrocarbon Series, and Kinked versus Straight. *Mol. Phys.* **2015**, *113* (22), 3389–3394. https://doi.org/10.1080/00268976.2015.1025882.

(54) Khatymov, R. V.; Muftakhov, M. V.; Shchukin, P. V. Negative Ions, Molecular Electron Affinity and Orbital Structure of Cata-Condensed Polycyclic Aromatic Hydrocarbons. *Rapid Comm. in Mass Spec.* **2017**, *31* (20), 1729–1741. https://doi.org/10.1002/rcm.7945.

(55) Rogers, J.; Gunn, S. Identifying Feature Relevance Using a Random Forest. In *Subspace, Latent Structure and Feature Selection*; Saunders, C., Grobelnik, M., Gunn, S., Shawe-Taylor, J., Eds.; Lecture Notes in Computer Science; Springer Berlin Heidelberg: Berlin, Heidelberg, 2006; Vol. 3940, pp 173–184. https://doi.org/10.1007/11752790_12.

(56) Modelli, A.; Mussoni, L. Rapid Quantitative Prediction of Ionization Energies and Electron Affinities of Polycyclic Aromatic Hydrocarbons. *Chem. Phys.* **2007**, *332* (2), 367–374. https://doi.org/10.1016/j.chemphys.2007.01.004.

TOC Graphic and Synopsis



A text-based molecular representation was designed for polybenzenoid hydrocarbons, enabling automatic feature extraction by interpretable machine learning models. New structure-property relationships were found.