UNIVERSITY OF LIVERPOOL

DOCTORAL THESIS

Accelerating Molecular Materials Discovery Following Data-Driven Approaches

PhD Student:

Aikaterini Vriza

Supervisors:

Dr Matthew Dyer

Dr Vitaliy Kurlin

Dr Dmytro Antypov

Dr Pete Wood

This thesis is submitted in fulfilment of the requirements

for the degree of Doctor of Philosophy

in the

Department of Chemistry

April 20, 2022

DECLARATION OF ACADEMIC INTEGRITY

NAME	AIKATERINI VRIZA
STUDENT NUMBER	201373684
MODULE TITLE/CODE	PHD IN CHEMISTRY
TITLE OF WORK	ACCELERATING MOLECULAR MATERIALS DISCOVERY FOLLOWING DATA-DRIVEN APPROACHES

This form should be completed by the student and appended to any piece of work that is submitted for summative assessment.

Students should familiarise themselves with Section 9 of the Code of Practice on Assessment and Appendix L of the University's Code of Practice on Assessment which provide the definitions of academic malpractice and the policies and procedures that apply to the investigation of alleged incidents.

Students found to have committed academic malpractice are liable to receive a mark of zero for the assessment or the module concerned. Unfair and dishonest academic practice will attract more severe penalties, including possible suspension or termination of studies.

STUDENT DECLARATION

I confirm that I have read and understood the University's Academic Integrity Policy.

I confirm that I have acted honestly, ethically and professionally in conduct leading to assessment for the programme of study.

I confirm that I have not copied material from another source nor committed plagiarism nor fabricated data when completing the attached piece of work. I confirm that I have not previously presented the work or part thereof for assessment for another University of Liverpool module. I confirm that I have not copied material from another source, nor colluded with any other student in the preparation and production of this work.

I confirm that I have not incorporated into this assignment material that has been submitted by me or any other person in support of a successful application for a degree of this or any other University or degree awarding body.

SIGNATURE……….…………………………………………………….…………
DATE…………………………………........20/04/2022...

Abstract

Designing new materials with desired properties is one of the main challenges for the current industrial and academic research, in the attempt to cover the societal demands. The 'utopia' would be, not only to find more reliable methodologies but also develop smarter ways for accelerating their discovery. Data-driven approaches are gaining ground as a tool for detecting patterns in known datasets and perform straightforward predictions.

In this work, both computational chemistry approaches and machine learning were employed to investigate two different types of molecular materials: 1) metal-doped polyaromatic hydrocarbons and 2) co-crystals.

The first Chapter provides a wide overview of the developments the data science tools have brought to the molecular world in the past years and covers the main theoretical aspects of the studied materials.

In Chapter two, a broad overview of the methods used to support this work is given covering both data science and computational chemistry aspects.

Chapter three is about the study of the relations between electronic properties and molecular structure in polyaromatic hydrocarbons, which are the building blocks of the materials studied herein.

Chapter four is diving more into the metal-polyaromatic hydrocarbon systems, starting from the extraction of all the available information regarding the currently known systems and further on developing strategies on how to guide the selection of the next most interesting systems.

Chapters five and six are related to co-crystals and how machine learning can be effectively used to provide an in-silico screening tool to prioritize molecular pairs that have high probability to form co-crystals. Chapter five is focused on the formation of molecular crystals, that consist of two components (co-crystals) connected via π-π interactions that might have electronic functionalities, *i.e.*, conductivity, whereas in chapter six the methodology developed for π-π co-crystals is scaled-up to cover all the co-crystal types. In both chapters, computational and machine learning approaches are implemented to detect promising coformers. Cambridge Structural Database (CSD), which is the world's repository for small-molecule organic crystal structures is the knowledge source for extracting the crystal structures of interest and then trying to understand the rules that guide their existence in terms of their conformer combinations.

Overall, this work is an attempt to combine predictive approaches using various machine learning algorithms with high-throughput computational modelling to guide the synthesis of new functional organic crystals. It is postulated that the complementarity of these tools will enable us to gain better insight into the materials discovery problems and thus drive to innovative and creative solutions.

Acknowledgments

First of all, I would like to thank both my supervisors Dr Matthew Dyer and Dr Vitaliy Kurlin for giving me the opportunity to start my PhD at the University of Liverpool and become a part of this interdisciplinary research.

I owe special thanks to my primary supervisor Dr Matthew Dyer, for our fruitful discussions, his great guidance around the projects I was involved in as well as for reading all my reports and giving me valuable feedback. Many thanks to my second supervisor, Dr Vitaliy Kurlin for introducing me to Data Science and given me a better insight into the mathematical part of data analysis approaches. Moreover, I would like to thank my third supervisor, Dr Dmytro Antypov for his directions through various computational software and his advice on my thesis drafts.

A big thank you to Professor Matthew Rosseinsky and Professor Neil Berry for their valuable feedback and guidance during our group meetings. This work wouldn't have been able without the collaboration with the members of the Rosseinsky group, Angelos, Rebecca and Rhian. Furthermore, I would like to thank our external collaborators from the Cambridge Structural Datacenter (CSD), Dr Ioana Sovago, Dr Peter Wood and Rob Willacy, for all their feedback and great suggestions regarding the co-crystals work and also for giving me the opportunity to present my work during the students science day organized by CSD.

I also owe great thanks to all my friends in Liverpool, Alice, Alexandra, Alex, Elpida, Marta, Pamela, Teo, Tonia, Makis, without them life here wouldn't have been so enjoyable.

I am extremely grateful to the Academy of Athens as without their support and help I would have never managed to move to the UK and start my post-graduate studies.

Finally, I would like to thank my family and especially my partner Tobenna for being next to me all the time even when he was physically miles away.

To my grandad, Vangelis, whose wise advice following me throughout my life

Che Fece ... Il Gran Rifiuto

For some people the day comes
when they have to declare the great Yes
or the great No. It’s clear at once who has the Yes
ready within him; and saying it,

he goes from honor to honor, strong in his conviction.
He who refuses does not repent. Asked again,
he’d still say no. Yet that no—the right no—
drags him down all his life.

BY C. P. CAVAFY
TRANSLATED BY EDMUND KEELEY

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AI	**A**rtificial **I**ntelligence
ANN	**A**rtificial **N**eural **N**etworks
API	**A**ctive **P**harmaceutical **I**ngredient
AUC	**A**rea **U**nder **C**urve
CIF	**C**rystallographic **I**nformation **F**ile
CSD	**C**ambridge **S**tructural **D**atabase
CSP	**C**rystal **S**tructure **P**rediction
DFT	**D**ensity **F**unctional **T**heory
ECFP	**E**xtended **C**onnectivity **F**ingerprint
GNN	**G**raph **N**eural **N**etwork
ML	**M**achine **L**earning
MO	**M**olecular **O**rbital
PAHs	**P**olyaromatic **H**ydrocarbons
QSAR	**Q**uantitative **S**tructure **A**ctivity **R**elationship
SMILES	**S**implified **M**olecular **I**nput **L**ine **E**ntry **S**ystem
SOAP	**S**mooth **O**verlap of **A**tomic **P**ositions
VASP	**V**ienna **A**b initio **S**imulation **P**ackage
vdW	**n**an **d**er **W**aals

1. Introduction

Discovering new materials is one of the main drivers of technological progress. Over the last hundred years, new materials were usually found accidentally with trial-and-error experiments guided by human expertise. More recently, with the establishment of computational science, experimental work has been accelerated through computational screening. In the era of the fourth industrial revolution, materials design has changed its shape and priorities following the rise of artificial intelligence, high-performance computing and open-data regulations.[1]

Current breakthroughs in the application of Artificial Intelligence (AI) and Machine Learning (ML) in life sciences have set the ground for systematizing materials design and discovery. Alphafold enables the accurate prediction of the three dimensional protein structure given only one dimensional information *i.e.*, the sequence of amino-acids.[2] Molecular transformer has been successfully used in synthesis planning for predicting the products of a reaction given the reactants and reagents.[3] It was also proven that the model can learn the language of chemical reactions and correctly classify them based on organic chemistry rules.[4] Message Passing Neural Networks (MPNN) were used for screening a large molecular database to identify novel compounds with antibiotic activity[5], to name just a few successful applications of data-driven models which not only accelerate materials discovery but also aid in gaining a deeper understanding of the existing data.

Following these advances, the main areas where AI/ML has been established include: *i)* property prediction, *i.e.*, models trained on curated datasets with known properties and are able to predict the desirable properties of any new material.[6–8] Using the trained models, we can now rapidly screen materials for desirable properties by searching materials databases. *ii)* materials classification, *i.e.*, being able to categorize materials based on their similarities. In that way materials that belong to the same class are expected to have similar properties[9,10] *iii)* The systematic design of novel materials to expand our search beyond the structures stored in databases using inverse design. In this regard, new materials with optimal properties can be generated.[11,12]

Some of the main problems the current AI-based systems encounter is the lack of negative data, the extremely biased datasets, unstructured data, and lack of explainability and physical understanding of the machine learning predictions.[13–16] Moreover, most of the ML models are based on the available data, and it is hard to extrapolate to unseen and novel materials. These challenges and opportunities for the important domain of molecular materials are going to be discussed in this thesis.

1.1 Molecular materials

Ranging from pharmaceuticals to electronic materials (Figure 1.1), the discovery of functional compounds is recognized as one of the fundamental pillars for the development of advanced technologies which are needed to face the challenges in clean energy, sustainability and global health.[17] Drug discovery is a priority task, aiming towards not only finding new drug candidates but also being able to modify their physicochemical properties, *e.g.*, making the drug more soluble to improve its delivery efficiency or improve its binding to proteins.[18]

Furthermore, the interest towards organic electronics has grown since many successful organic photovoltaic cells (OPC), light emitting diodes (LED) and thin film transistors (TFT) have been reported.[19,20] Organic electronics offer an alternative option to inorganic materials for applications that require low-cost, large-area and flexible electronic devices. In this regard, isotropic polymers, *i.e.*, amorphous materials with identical properties in all their directions, are more used in the LED development, while highly ordered (anisotropic) compounds are considered more suitable for TFTs.[21] In the context of energy storage, great focus is also put on the design of batteries from sustainable and cheap resources to substitute toxic lithium-ion counterparts. Metal-organic frameworks (MOFs) have gained significant attention from the academic community as gas separation and storage materials due to their unprecedented chemical and structural tunability.[22]

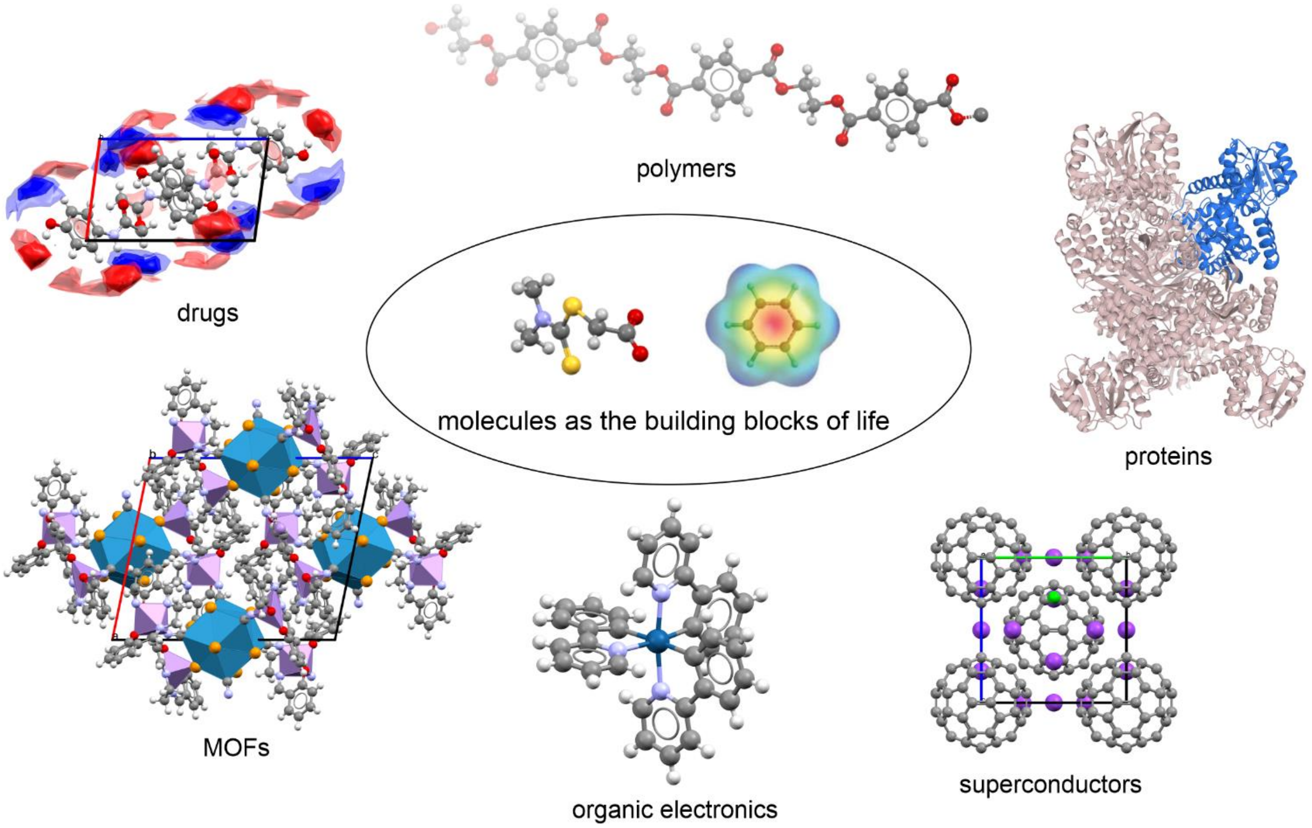


Figure 1.1 *Examples of molecular materials, including drugs (csd id: COTZAN02),*[23] *Metal Organic Frameworks (csd id: ACAHAN),*[24] *polymers (*csd id: WIMZEX*),*[25] *proteins (Uniprot id P00370),*[26] *organic electronics (csd id: BORCIW),*[27] *superconductors (csd id: QUHYOH).*[28] *(Central molecules csd ids: YUFMAN, BENZEN)*

It is evident that the molecular world is vast and current technologies are aimed towards finding clever approaches to navigate this world and investigate the huge possibilities it can offer. A synergy between AI/ML-driven models, computational chemistry and experimental realization is supposed to greatly improve the pace new discoveries are made and enable for a better understanding of molecular materials.

1.2 Data-driven approaches for the discovery of new functional materials

Machine learning has been incorporated in many fields of science and technology, ranging from medical diagnostics to materials design.[29–31] Computational models are built to identify hidden patterns in data for the automated generation of information, often with a strong focus on making predictions of future data.

The advances in machine learning algorithms, the vast amount of open-source chemical data and the availability of powerful computational resources have given rise to the development of different types of mathematical models that once trained on a dataset can infer the hidden patterns on the data and map input data to output values. For a given field of research, the success of data-driven materials discovery is often contingent on the availability of a large and diverse set of chemical data that display patterns according to structure-property relationships that are associated with that field.[1] All the available chemical information makes the 'chemical space' of each subdomain of materials. The shape and size of the chemical space define the type of machine learning models that could be applied to the discovery of new materials.

The goal of machine learning is to use algorithms to learn from data, in order to build generalizable models that give accurate classifications or predictions, or to find (useful) patterns, particularly with new and previously unseen data. The concept of deriving structure-property relationships is not new. This started many years ago with a method known as Quantitative Structure Activity Relationship (QSAR) modelling.[32] Nowadays with the increase on the available data and the speed of calculations, QSAR has been substituted by up-to-date neural networks that are capable of modelling non-linear relationships in the data. Current research aims to bridge the gap between experiment and theory, and to promote a more data-intensive and systematic research approach.

Applications of data-driven approaches can be found in various domains, *e.g.*, for extracting important information of the electronic features space and understanding which of them have an important role in predicting some specific properties relevant to material's performance. Sahu *et al.*. managed to estimate the power conversion efficiency (PCE) of organic photovoltaics using 13 important microscopic properties.[33] Padula *et al.* combined both electronic properties with structural similarities of organic molecules to assess the power efficiency of similar molecules.[34] In another work, researchers were able to simultaneously predict multiple electronic properties, including static polarizabilities, excitation energies and intensities, based only on stoichiometry and configurational information of small organic molecules[35].

ML complements and can even be combined with established theoretical chemistry techniques, such as Density Functional Theory (DFT), wave function theory, force fields, and molecular dynamics. These sophisticated, physics inspired methods have proven to be highly valuable *post hoc* to understand specific systems. However, their prospective use is less common, due to their significant computational cost. With appropriate training data in hand, an ML model can learn on its own to generate such predictions, also for data points which the model has never encountered before – independent of current (human) knowledge.

Machine learning methods bear the potential to change, or at least to strongly impact, the way chemical challenges will be approached in the future – guiding and complementing the skill set of synthetic chemists. With increasing amounts of well-curated data and algorithmic advances, the prime time for applying machine learning in chemistry is yet to come. The focus of this thesis is on developing data-driven strategies for accelerating the discovery of new materials that can arise by combining two different molecules, or a molecule with a metal.

1.3 Co-crystals: materials based on molecular combinations

A co-crystal can be defined as a crystalline material consisting of two or more different molecules in specific stoichiometries. The basic requirements of those structures for being considered as co-crystals are provided by the crystal engineering field and could be summarized as following[36]:
i) Co-crystals differ from salts, as none of their components are charged; in the co-crystal lattice, the components co-exist with a defined stoichiometry and interact non-ionically, whereas salts consist of charged molecules.
ii) All co-crystal components are organic species, ruling out inorganics and organometallics.
iii) Water and solvents are excluded as components, otherwise the crystal structures are characterized as hydrates or solvates, respectively.
Co-crystals could be broadly categorized in those of pharmaceutical and those of electronic interest. They offer great opportunities in materials science as their solid form properties can be easily modified by the combination of different molecular species instead of modifying the original molecules.

1.3.1 Co-crystals for pharmaceutical applications

Co-crystallization has emerged as an important process for drug development. According to the regulatory classification of pharmaceutical co-crystals, produced by the Food and Drug Administration (FDA), one of the constitutional components of a co-crystal is considered as the API (active pharmaceutical ingredient) and the other coformers are selected such that they comply with the above mentioned requirements[37]. The importance of cocrystals in the pharmaceutical industry lies in the fact that they can change the physical properties of an API, whereas the chemical features are preserved[38]. Some of the physical properties that can be effectively tuned by co-crystallization are dissolution rate, compressibility and physical stability[36,39].

1.3.2 Co-crystals as organic conductors

Organic conductors are extended conjugated π-systems that have the ability to transport charge when an electrical bias is applied. Research has shown that the electronic properties of interfaces between two different solids might significantly differ from those of the constituent materials. As an example, interfaces formed by insulating transition-metal oxides have shown metallic conductivity[40,41] and even, under some conditions, superconductivity.[42] Likewise, there are reports of metallic systems being created from conjugated organic molecules that are insulators with the first metallic organic charge-transfer co-crystal being synthesized by tetrathiofulvalene (TTF) and 7,7,8,8-tetracyanoquinodimethane (TCNQ) in 1:1 ratio.[43]

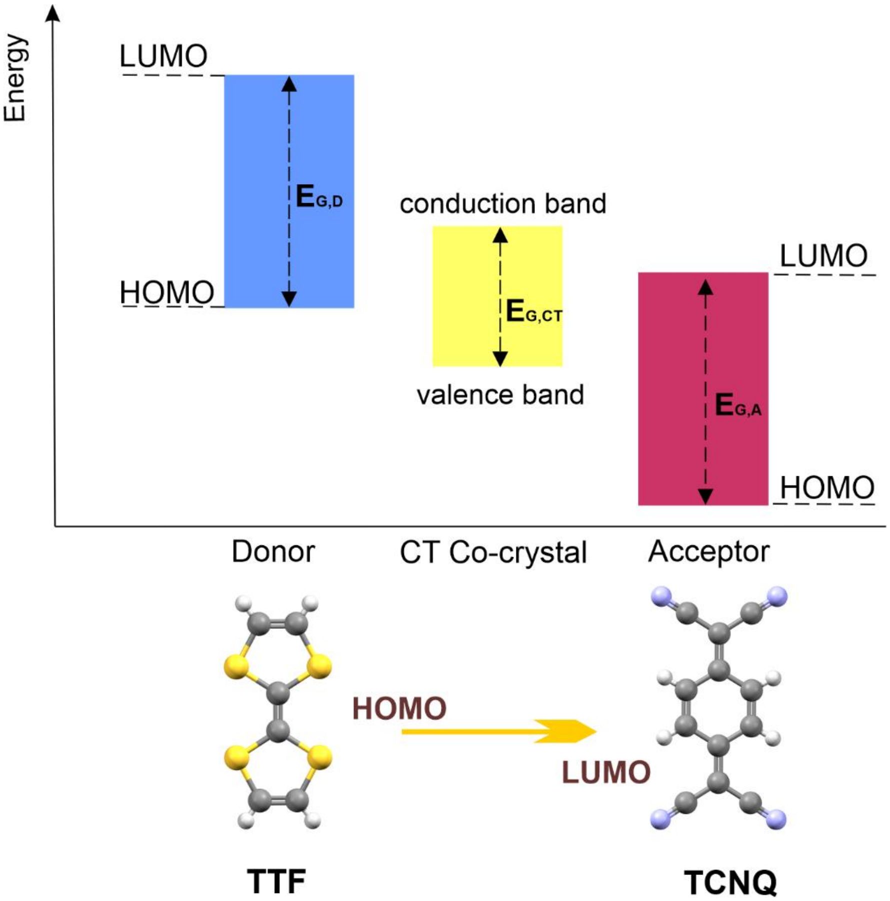


Figure 1.2. *Charge transfer in the TTF–TCNQ system. The TTF and TCNQ molecules are well known since their use in the synthesis of the first metallic charge-transfer compound. In TTF–TCNQ crystals, electrons from the HOMO of the TTF molecules are transferred into the LUMO of the TCNQ molecules, leading to a stable charge-transfer state.*

This co-crystal behaves as a metal over a large temperature range and has a large maximum electrical conductivity σ_{max}=1.47x10^4 cm^{-1} at 66 K[44], although single TTF and TCNQ crystals are semiconductors showing very low conductivity with HOMO-LUMO gap larger than 2eV.[45] In the TTF–TCNQ system, the electronic transport is achieved between the highest occupied molecular orbital (HOMO) of TTF (donor) and the lowest unoccupied molecular orbital (LUMO) of TCNQ (acceptor), as shown in Figure 1.2. The constituent molecules are arranged in linear chains and the material is highly conducting at room temperature when these chains behave as decoupled, one-dimensional electronic systems. At low temperatures, the compound becomes an insulator owing to two Peierls transitions, *i.e.*, rearrangement of electrons due to lattice distortion, occurring independently on the TTF and TCNQ chains (at T= 54 K for the TCNQ chain and at T=38 K for the TTF chain).

1.3.3 Driving forces of co-crystal formation

Various knowledge-based approaches have been implemented for understanding the powers that affect co-crystallization. The fundamental know-how around their formation includes the selection of the constitutional molecules and investigates their connection to each other in a way that stable crystal structures will be formed. Considering the types of bonding in various co-crystal structures, the most commonly found are those relevant to the functional groups of the molecules (hydrogen-bond donors and acceptors, halogen atoms)[46] or refer to weakly bound co-crystals with no functional groups (π-π stacking or other weak interactions)[47].

A popular virtual screening method is calculating the electrostatic potential surfaces for hydrogen bonded two-component cocrystals and the energy difference between the two pure solids and cocrystals is used as a probability measurement for a cocrystal formation. It was shown that for the experimentally observed structures this measurement is higher, indicating that this metric can be useful for quick assessment of a cocrystal formation. However, it is only limited to cocrystals with H-bond interactions[48]. Machine learning approaches can also predict quite accurately if two components will form a co-crystal or not, based on the complementarity of their functional groups (supramolecular synthons)[49] or their ability to form hydrogen bonds (hydrogen-bond forming moieties) [50]. However, those methods cannot be appropriately fitted to structures where π-π interactions dominate. Thus, more general approaches that consider a wider range of molecular properties have been developed, which propose that shape and polarity are the most important descriptors that can influence the co-crystal formation.[51]

Although the selection of the appropriate model depending on the candidate molecules enables for a quick *in-silico* screening for prioritizing molecular pairs, that does not guarantee the formation of a co-crystal. Some other approaches for defining a general rule for co-crystal formation are considering the energetic profile of the coformers, proving that a multicomponent crystal is expected to form only if it is thermodynamically more stable than the crystals of its constituents.[52,53] In this approach all the possible types of bonding were taken into consideration and the changes in energy that might happen when a cocrystal is formed are calculated. It is suggested that the co-crystallization is almost always the thermodynamically favoured process, but it is still difficult to extract a general rule for guiding the synthesis[53]. For this type of evaluation to be feasible, the possible crystal structure should be first determined. A crystal structure can be predicted from the chemical diagram as was shown in the most recent international blind test organized by the Cambridge Crystallographic Datacentre which includes a co-crystal.[54] However, one of the highest concerns when working with crystals is the polymorphism they present. Polymorphism has been declared as the nemesis of crystal design as the physical properties of the several polymorphs cannot be easily controlled.[55] Consequently, the stable polymorphs should also be considered with the Crystal Structure Prediction (CSP) methods and evaluated in terms of their properties[56]. In addition to the energetic stability considerations, there are also some experimental considerations that might prohibit co-crystal formation, *e.g.*, in solution crystallization methods if one of the two co-crystal components crystallized first then a co-crystal in infeasible.[57]

Overall, the co-crystallization prediction task is very challenging as many parameters should be taken into consideration. The CSP methods show promise, but they are still quite time consuming and thus cannot be effectively used to evaluate a large amount of possible molecular pairs. Moreover, the CSP predicted structures cannot incorporate experimental considerations such as the solubility of the coformers. For that reason, a virtual co-crystal screening is an important first step before the application of a CSP method such that only the highly promising pairs are further evaluated.

1.4 Metal-Polyaromatic Hydrocarbon (PAH) systems of electronic importance

Focusing on novel materials with electronic applications, systematic research in the field of metal-intercalated PAHs has been performed by several research groups.[58–63] These systems are comprised of a polyaromatic hydrocarbon dopped with alkali metals (Figure 1.3). The electronic properties that arise show promise and thus these materials can serve as alternatives to expensive metals and inorganic optoelectronics.

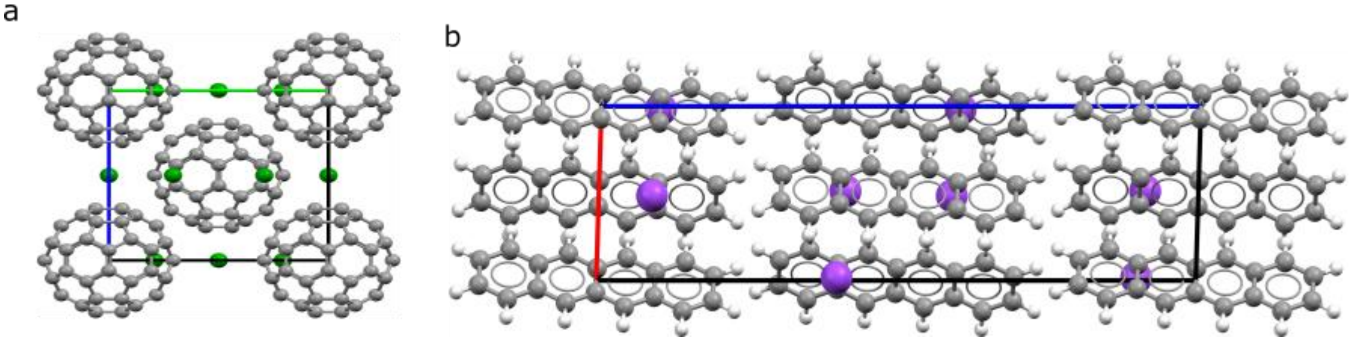


Figure 1.3. *Metal-intercalated PAHs of a) C_{60} dopped with caesium (Cs) b) tetracene dopped with potassium (K).*

Further on, an overview of the important electronic characteristics of organic molecules is given, followed by the properties than could be achieved after the metal insertion.

1.4.1 The electronic characteristics of organic materials

Organic materials have recently been the object of intense studies due to their opto-electronic properties which follow from the behaviour of the outer-shell electrons. For designing materials for electronic applications, it is essential to first identify the functionality we are interested in and then the important characteristics that might control that functionality. In this work, we are aiming at identifying the ways to design materials with conducting properties, thus the features that affect conductivity in a structure are going to be further discussed. As conductivity in molecular materials is highly related to the electronic structure, the connection between the molecular orbital theory and the band theory is first discussed below.

From molecular orbital energies to the band theory of solids

Molecular orbitals can be defined as a set of energy levels that describe the motion of electrons in molecules. Molecular orbitals encode the distribution of electrons in a molecule, thus offering direct insights into its underlying electronic structure. Especially the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO) could play significant roles as they are involved in exciton formation, exciton dissociation, and hole transport processes influencing the macroscopic conductivity[33]. The HOMO orbital is often associated with the ionization potential of the molecule, whereas the LUMO with the electron affinity, *i.e.*, the ease with which the molecule may accept an electron. To date, the design and selection of organic electronics was based primarily on the HOMO and LUMO of organic molecules. However, current research has shown that LUMO, LUMO_{+1} gap could play a very important role in the efficiency of those materials[33].

When zooming out of a single molecule, we observe that molecules are found in periodic arrangements, if crystalline, comprising a solid material. The electronic structure of the solid is described by band theory instead of molecular orbital theory. The electronic structure of solids can be regarded as an extension of molecular orbital theory to aggregates consisting of virtually infinite numbers of atoms. When talking about the band theory in solids, *e.g*, crystal structures, we refer to the formation of continuous bands of energy levels instead of discrete levels. Moreover, the translational symmetry of the lattice has a key role to play to the energy levels.

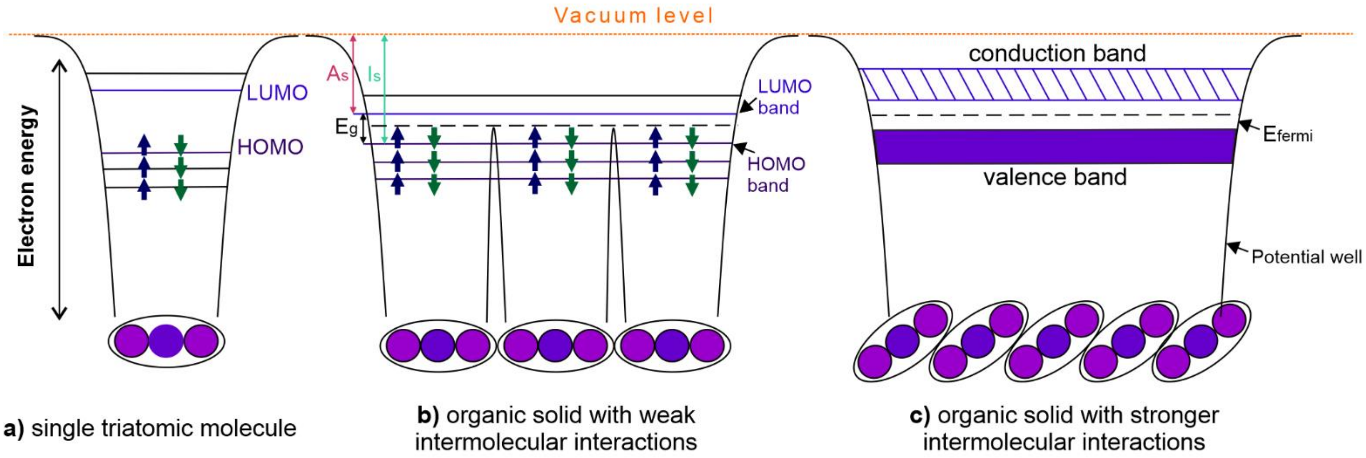

Vacuum level
Electron energy
LUMO
HOMO
A_s
I_s
E_g
LUMO band
HOMO band
conduction band
E_{fermi}
valence band
Potential well
a) single triatomic molecule
b) organic solid with weak intermolecular interactions
c) organic solid with stronger intermolecular interactions

Figure 1.4. *Evolution of electronic structure represented by the potential well of a) single molecule (b) solid with weak and (c)solids with stronger intermolecular interactions. When intermolecular interactions are weak, the width of energy bands is very narrow. With increase in the intermolecular interaction the bandwidth becomes larger. E_g: bandgap, A_s: electron affinity of solid, I_s: ionization energy of solid.*

Figure 1.4 illustrates the concept of the evolution of the electronic structure from molecular orbitals to the bands of solids. Starting from a single polyatomic molecule (Figure 1.4a) or else regarded as gas phase, the potential well is formed by the Coulombic potential of each atomic nucleus. The wells of the nuclei are merged to form a broad well where various molecular orbitals (MOs) exist and produce discrete energy levels that are different from the atomic energy levels. The upper horizontal part of the potential well is the vacuum level at which an electron that is taken apart from the molecule stops moving and its kinetic energy is zero.[64] When molecules come together connected by weak interactions to form an organic solid (Figure 1.4b), the wave functions of the occupied valence states and the lower unoccupied states are mainly localized in each molecule, yielding narrow intermolecular energy band of the bandwidth approximately. When intermolecular interaction becomes larger (Figure 1.4c), both the occupied (valence) and unoccupied (conduction) bands become wider because of larger overlapping of relevant MOs of adjacent molecules, forming the bands of solid materials[64]

Band theory and the Fermi level

The electronic structure of a solid consists therefore of three main parts, the conduction band, the valence band and the band gap, which govern the electronic conductivity properties of a material and the following classification arises (Figure 1.5):[65]

i) Insulators: Materials that cannot conduct electricity due to a large band gap (E_g) that separated the valence band containing an even number of valence electrons per unit cell for the empty conduction band.

ii) Semiconductors: Materials that can conditionally conduct electricity due to the even number of valence electrons per unit cell. They possess a full valence band separated by a small band gap (E_g) from an empty conduction band. Their conductivity arises from the excitation of electrons from the valence band into the conduction band. The number of electrons promoted to the conduction band increases with an increase in temperature.

iii) Metallic conductors: Materials which have partially filled conduction bands and thus electric conductivity which increases as the temperature is decreasing. Metals are regarded as the most well-known conducting materials as the presence of charge-carrying electrons is most obvious. A metal can be described as an array of widely spaced, small ionic cores, with the mobile valence electrons spread through the volume between and thus conduct electricity. The metallic state is favored by most elements, especially those that belong to the left-hand side of the periodic table.

It should be noted that another category that arises from conductivity considerations is superconductivity. This category is further discussed in a separate section (Section 1.4.2).

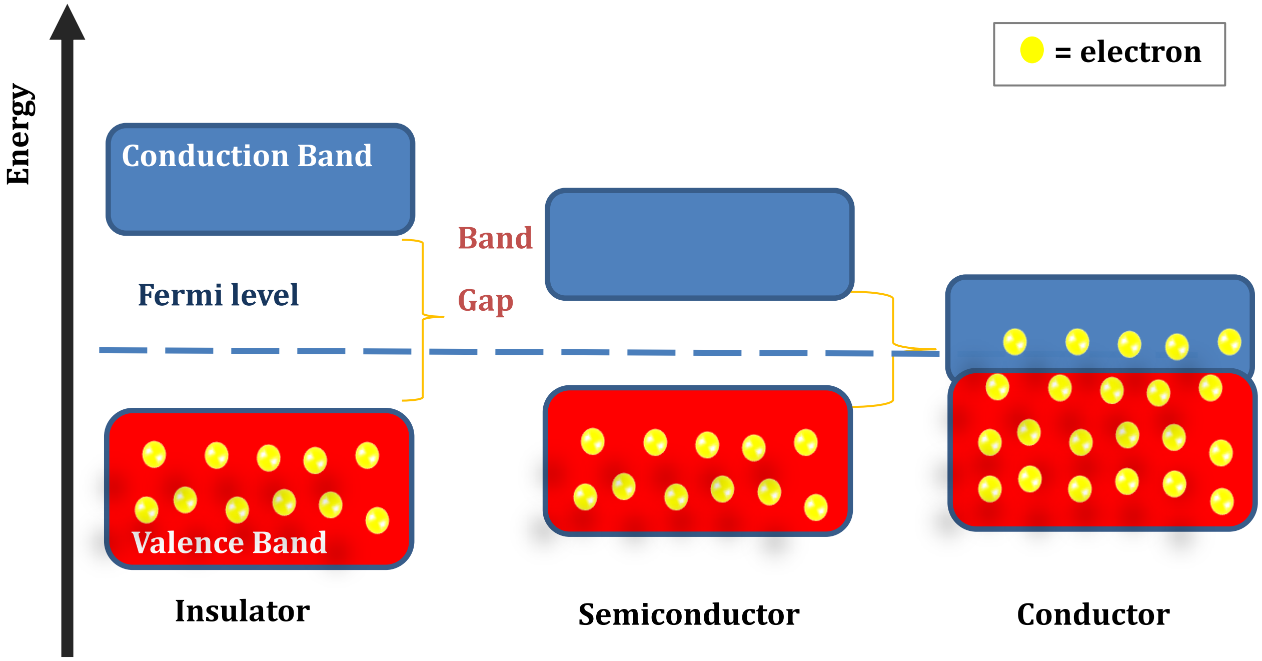


Figure 1.5. *Energy level diagram for an insulator, a semiconductor, and a metal. The band theory of solids gives the picture that there is a sizable gap between the Fermi level and the conduction band of the semiconductor or an insulator. At higher temperatures, a larger fraction of the electrons can bridge this gap and participate in electrical conduction. In conductors, the valence band and the conduction band overlap. Hence there is no bandgap and the valence electrons can move to the conduction band easily.*

Band theory is used to explain the behaviour of conductors, semiconductors and insulators. The bands of orbitals take their name depending on the type of orbitals they are formed by. For instance, we have s-band when the band of orbitals is formed by the linear combination of s-orbitals, or p-band when it is a linear combination of p-orbitals. In a typical semiconductor/insulator, the energy separation of the s- and p-orbitals of the free atoms will be quite large and as a result the two bands will not overlap.

Another important concept related to the band theory of solids is the Fermi energy (E_{Fermi}). The Fermi energy is the energy of the highest occupied electronic state at 0 K temperature.[66] At absolute zero, the electrons pack into the lowest available energy states and build up a "Fermi sea" of electron energy states. The Fermi level is the surface At higher temperatures a certain fraction, characterized by the Fermi function, will exist above the Fermi level.

In metals, the Fermi energy is equal to the energy difference between the highest and the lowest electron energy states of the conductions electrons at absolute zero.[67] Consequently, there are electronic states just above the Fermi energy that can be populated by electrons which are accelerated from an electric field. As a result, the material can readily conduct electricity.[68]

In semiconductors and insulators, the Fermi function *f(E)* gives the probability that a given available electron energy state will be occupied at a given temperature. The Fermi function comes from Fermi-Dirac statistics and has the form:

$$f(E) \ = \frac{1}{e^{(E-E_F)/kT}+1} \qquad (1.1)$$

The basic nature of this function dictates that at finite temperatures, most of the levels up to the Fermi level E_F are filled, and relatively few electrons have energies above the Fermi level. Note that although the Fermi function has a finite value in the gap, there is no electron population at those energies. The population depends upon the product of the Fermi function and the electron density of states. Consequently, in the gap there are no electrons because the density of states is zero. In the conduction band at 0K, there are no electrons even though there are plenty of available states, but the Fermi function is zero. At high temperatures, both the density of states and the Fermi function have finite values in the conduction band, so there is a finite conducting population.[69]

Overall, in metals the Fermi energy falls into the conduction band, whereas for semiconductors and insulators is within the band gap, as shown in Figure 1.5. In doped semiconductors, p-type and n-type, the Fermi level is shifted by the impurities and is close to the band edge or falls inside the conduction band.[70]

Other important electronic characteristics

Some of the important features that have been used up to date in research on organic electronic materials are presented below:

i) Electron-electron coupling (Hubbard term): The Hubbard Hamiltonian offers insight on how the interactions between electrons give rise to insulating, magnetic and even novel superconducting effects in a solid.[71]

ii) Electron-phonon coupling (reorganization energy): The electron-phonon interaction is very important both in creating the phonon scattering of the electrons but also in the formation of Cooper pairs. This interaction is indeed the cause of superconductivity.[72]

iii) Polarizability: A large polarizability is expected to stabilize the charge separated states and thus reduce the exciton binding energy.[73]

1.4.2 Superconductivity in PAHs

Superconductivity is the phenomenon present in a material where electricity flows though it with no resistance when the material is cooled below a transition temperature (Tc). The phenomenon was discovered in 1911 by Dutch scientist Heike Kamerlingh Onnes, who demonstrated the lack of resistance by creating an electrical current in a closed loop of a mercury superconductor[74]. Superconductivity is a state that usually exists either at very low temperatures or at higher temperatures with high pressure involved.

The highest transition temperature superconductors at ambient pressure belong to the cuprates family, *i.e.*, cuprates of mercury, barium and calcium at 133K.[75,76] In general, superconductivity used to be a phenomenon observed mainly in inorganic materials. However, superconductivity has been observed in organic chemistry in structures that involve graphite or fullerenes (C_{60}). Graphite has a layered structure composed of infinite benzene-fused π-planes (graphenes) with sp^2 character, whereas fullerene has a soccer-ball-like structure with 12 pentagonal and 20 hexagonal symmetrically arrayed faces that belongs to a high symmetry group, the icosahedral point group I_h. Superconductivity in these molecules was first reported in 1990 after alkali metal doping.[77,78] Only recently, in 2018, superconductivity in pure graphene was reported in a sandwich of two graphene layers when they are twisted at a ‘magic’ angle of 1.1^o.[79]

Going beyond graphite and C_{60}, alkali-doped polyaromatic hydrocarbons have shown promise for their potential superconducting behaviour. PAHs are currently one of the most interesting and challenging research subjects due to their high stability, their rigid planar structure, and their characteristic optical spectra.[80] Polyaromatic hydrocarbons are of interest for their structural relationship to fullerenes, as they include fused benzene rings and thus a conjugated π-system. The intermolecular interactions, in particular the π–π interactions, depend on the packing of the molecules in the crystal structure, and there are a few typical arrangements which favour the π–π interactions and, therefore, the electronic properties.

With superconductivity being a highly sought-after property, there is a rich literature both on theoretical and experimental studies of different PAHs showing superconductivity when intercalated with alkali-metals. Once the crystal structure of the material is determined, it is possible to obtain information about the electronic band structure. The occupation of the low energy orbitals of the PAHs by the electrons given by the potassium atoms will affect the Fermi level and consequently impact the charge transport in the material. Some examples are K_3Picene, with a critical temperature of 18 K, and subsequently phenanthrene-, dibenzopentacene-, and coronene-based materials. The maximum superconducting temperature of 33 K was reported in potassium-doped 1,2:8,9-dibenzopentacene[81]. In addition to using the alkali metals dopants, there are references for superconductivity with rare earth doping in phenanthrene with the T_c approaching 6 K for La and Sm and in chrysene with T_c around 5 K for Sm.[82]

Despite the vast reports of superconductivity in PAHs, the reproducibility of those products and a lack of detailed characterization inhibits the understanding of the properties of these materials.[83,84] However, some insights about the mechanism of superconductivity have been gained through theoretical models, which suggest that both electron-phonon interactions as well as electron correlations might play an important role.[85]

1.5 Scope and structure of this thesis

The aim of this thesis is to provide computational tools and workflows for enabling functional materials discovery. The main tools used are machine learning models, computational chemistry software and databases. Having polyaromatic hydrocarbons as the main building blocks for electronic materials, we are investigating their behaviour in metal:PAHs systems and in co-crystals.

The thesis is outlined as follows:

- **Chapter 2** presents the main methodologies and tools used to support this work, ranging from machine learning algorithms to density functional theory.

- **Chapter 3** is related to a general understanding of the structure-property relations in PAHs. Starting from a small dataset of 210 PAHs we explore the important correlations between structure and orbital energies, testing on different 2D and 3D representations. Using these findings, we extrapolate on a large and diverse dataset of 7,000 PAHs. Two main datasets are constructed which are the base for the analysis in the next chapters: *i*) 210 PAHs dataset for selecting pairs for co-crystallization and *ii*) 7,000 PAHs to select promising molecules for intercalation.

- **Chapter 4** is focusing on metal:PAHs systems. A comprehensive workflow is built to drive the selection of the best PAHs candidates to be experimentally intercalated. A Crystal Structure Prediction (CSP) study is performed on the most promising candidates for identifying energetically stable configurations and comparing the relative energies with the currently known metal:PAHs systems. The calculations demonstrate that all intercalated structures are more stable than their constituent parts, indicating that the energetic stability might be a driving force for the formation of these systems.

- **Chapter 5** aims to establish a computationally efficient methodology, namely the one class classification, to identify promising molecular pairs for designing novel co-crystal structures with important electronic features, based on electronic, molecular and topological properties of the monomers. In this chapter, machine learning approaches are implemented to learn the boundaries of the area in which all the known PAHs co-crystals belong and then apply this knowledge to rank promising combinations of molecules according to their similarity to the known instances. The selection of the high-ranking pairs is further optimized based on electronic considerations, *i.e.,* similarity to TCNQ, an electronically active molecule. Experimental verification of the method shows promising results as two novel PAHs co-crystals were synthesized with bandgaps in the range of semi-conducting materials.

- **Chapter 6** extends chapter 5 and explores how the one class classification approach can be extended to cover the whole known co-crystal space. The approach is validated on extended benchmark datasets including both successful and unsuccessful co-crystal screening results that have been gathered from literature. The methodology is significantly updated by investigating different molecular representations, tuning the network hyperparameters and including a measure for the uncertainty of predictions. The best performing model is used for ranking possible molecular pairs from the ZINC20 database including pharmaceutical and electronic co-crystals. A web application is also built for enabling for in-silico co-crystal screening by the user.

- **Chapter 7** concludes the thesis, summarises the main contributions and gives an outlook of the future of materials design.

2 Methods

2.1 Introduction to machine learning

Moving towards the 4th industrial revolution, new ways to make better use of the wealth of data are sought after. Machine Learning (ML) stands in the forefront of the current developments and has found several applications in different fields, ranging from economics to materials science.[86,87] The four core components to be taken into consideration for any machine learning model include:

1. The data to be learned from and how they are represented.

2. A model to transform the data into a format such that training can be performed.

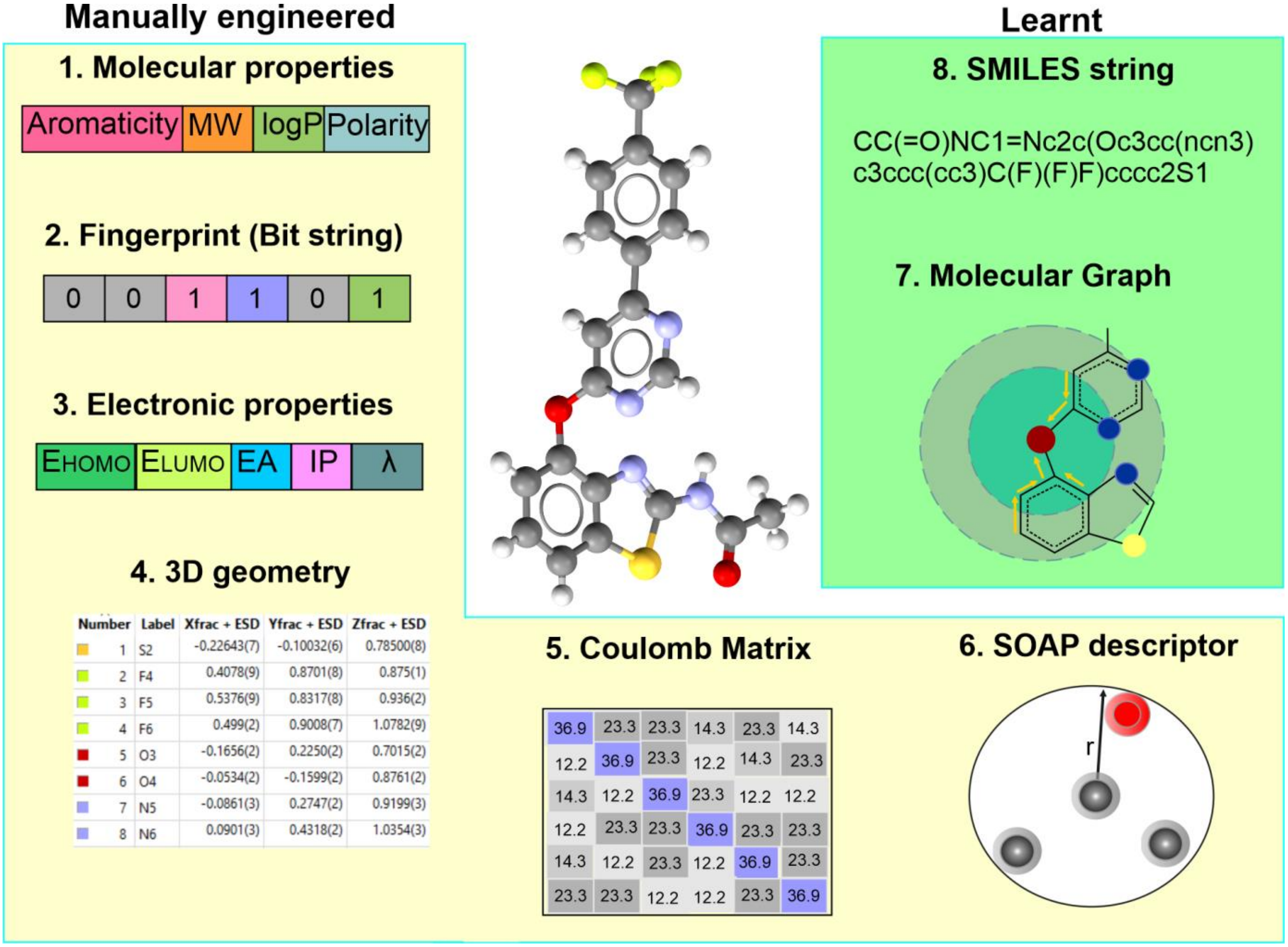
3. An objective function that quantifies how well the model is doing based on the selected evaluation metrics.

4. An algorithm to adjust the model's parameters to optimize the objective function.

ML is concerned with finding methods with which computers can extract useful patterns from data, transforming them into a model capable of performing a task without being explicitly programmed to do so. ML models usually fall under one of two groups: generative models, which are capable of generating similar data to what the trained model has seen; and discriminative models, which are capable of making predictions of properties of interest given the data.[88,89] This thesis would be mainly focused on making use of discriminative models.

2.1.1 Molecular Materials Representation

A necessary step before performing any machine learning analysis on materials science data is to represent the material under consideration in a machine-readable format. This representation termed "descriptor" should contain all the relevant information on the system needed for the desired learning task.[90–92] In the case of molecular materials, the critical choice of the representation will affect the accurate modelling and prediction of molecular properties. The most widely used molecular representation techniques are introduced in Figure 2.1. Herein, the categorization refers to two distinct types of representation techniques: *i*) the expert-designed molecular descriptors, *i.e.*, obtained with rule-based algorithms and *ii*) the learnt molecular representations directly from data.[93,94] The expert-encoded descriptors refer to features that describe the molecule based on structure or relevant molecular properties and are selected by experts. Examples in this category involve the molecular descriptors which encode a wide variety of molecular information, e.g., shape, geometry, atomic properties, pharmacophores.[90] Using the molecular features in predictive models has shown not only a promising way for accurate molecular property predictions but also for a straightforward way to understand the factors that contributed to the predictions.[95–97] Morgan Fingerprint or else extended connectivity fingerprint (ECFP), which is a bit string with each bit denoting the presence or absence of a molecular feature or substructure, is another well-established representation technique.[98] The ECFP algorithm effectively encodes each molecule as a “bag-of- fragments” based on local atomic environments, generating unique integer identifiers that are subsequently hashed into a fixed-length representation. As a result, each fragment is necessarily and completely distinct. Despite their simplicity and sparseness (many 0s) Morgan fingerprints have shown great predictive capabilities for various molecular properties, showcasing their usefulness.[99,100] The use of these fingerprints have also resulted in a fast method for bit-wise comparison of molecular features, namely Tanimoto similarity, which allowed for rapid filtering and search in chemical databases.[101,102] Except from the structure-based descriptors, the electronic descriptors (*e.g.*, orbital energies, reorganization energy) have been successfully used for predicting electronic related properties, *e.g.*, photovoltaic efficiency or photocatalytic activities.[34,103,104]



Number		Label	Xfrac + ESD	Yfrac + ESD	Zfrac + ESD
🟧	1	S2	-0.22643(7)	-0.10032(6)	0.78500(8)
🟩	2	F4	0.4078(9)	0.8701(8)	0.875(1)
🟩	3	F5	0.5376(9)	0.8317(8)	0.936(2)
🟩	4	F6	0.499(2)	0.9008(7)	1.0782(9)
🟥	5	O3	-0.1656(2)	0.2250(2)	0.7015(2)
🟥	6	O4	-0.0534(2)	-0.1599(2)	0.8761(2)
🟦	7	N5	-0.0861(3)	0.2747(2)	0.9199(3)
🟦	8	N6	0.0901(3)	0.4318(2)	1.0354(3)

36.9	23.3	23.3	14.3	23.3	14.3
12.2	36.9	23.3	12.2	14.3	23.3
14.3	12.2	36.9	23.3	12.2	12.2
12.2	23.3	23.3	36.9	23.3	23.3
14.3	12.2	23.3	12.2	36.9	23.3
23.3	23.3	12.2	12.2	23.3	36.9

CC(=O)NC1=Nc2c(Oc3cc(ncn3)
c3ccc(cc3)C(F)(F)F)cccc2S1

Figure 2.1 *Different types of molecular representations using as an example a known drug TEGFIW, which is a molecule having several functional groups and used in several co-crystals; 1) Molecular descriptors based on the molecular structure, 2) Morgan Fingerprint as a bit-like vector, 3) Electronic properties, 4) 3D geometry, 5) Coulomb matrix, 6) SOAP descriptor, 7) Graph with atoms as vertices and bonds as edges, 8) SMILES string.*

Other categories of manually engineered features involve the 3D geometry of the molecule as represented from the atomic coordinates, x,y,z, and the atomic structure descriptors (coulomb matrix, SOAP) describing the neighbourhood around each atom.[35,105] More complex representations have been also implemented as input in deep neural networks, where molecular structures are represented through a vector of nuclear charges and a matrix of inter-atomic distances[106]. Another approach, inspired from Natural Language Processing (NLP), is to represent molecules as word vectors (mol2vec) where molecules are represented as sentences with the subsequent functional groups as words[107].

Large scale investigations have shown that the existing expert encoded molecular descriptors are insufficiently expressive for many applications. Consequently, there is a necessity for general-purpose molecular representations that can capture the rich diversity of chemical space. Deep learning models can efficiently learn compact molecular representations and provide an alternative way for describing the molecules. String- and graph-based formats are extensively used in deep neural networks for encoding the complete composition and bonding of molecules in continuous vectors as opposed to the discrete vectors of the hand-crafted features. The Simplified Molecular Input Line Entry System (SMILES) is a string-based representation that follows a formal grammar system allowing the direct adaptation of methods and architectures from natural language processing and neural machine translation. Another way for handling molecules is by representing them as graphs, with atoms as nodes and the bonds and the edges. Graph learning proceeds in several steps. First, existing molecular features, such as atom type and hybridization, are directly encoded to each node representation. Throughout the layers in a GNN, node representations are updated with information passed from their surrounding neighbourhoods in a framework known as message passing. This process of iterative message passing, and updates allows information to flow across the graph to create a continuous and dense representation of each node.[94,108]

Overall, it should be noted that the representations should be complete and invariant to be effective and that the selection of the representation is mainly based on the property or functionality that needs to be predicted. For instance, molecular descriptors such as molecular weight or polarity might correlate well to a property such as boiling point but might suffer in more complex tasks but as protein binding where aspects of geometry and structure provide crucial information. A representation should also be interpretable to ensure that the model performance derives from learning relevant patterns instead of by exploiting experimental noise or other possible artifacts.

2.1.2 Types of Machine Learning Algorithms

The three broad categories of machine learning algorithms are supervised learning, unsupervised learning, and reinforcement learning. Semi-supervised learning is regarded as a subcategory which falls between unsupervised and supervised learning.

Supervised learning: In supervised learning, we are given a dataset in which we already know that there is a relationship between the input and output and how the correct output should look like. The main goal is to use labelled data to 'teach' a model (function) to predict the desired output on unseen or future data. Supervised learning problems can be broadly categorized into "regression" and "classification" problems. In a regression problem, we are trying to predict results within a continuous output space, meaning that we are trying to map input variables to some continuous function. Supervised regression models are often used for molecular properties prediction tasks, where a large amount of experimental or computational data exist together with known or calculated properties. On the other hand, in a classification problem the goal is to predict categorical class labels of new instances based on a given input. The set of class labels does not have to be binary in nature and can be of arbitrary size usually dictated by the number of class labels present in the training dataset.[109] As a concrete example in computer vision, supervised machine learning can be used to classify dogs and cats. The model is trained using labelled images of cats and dogs and then is able to assign a label to any image. An important consideration of supervised learning approaches is how well they can generalize to out-of-distribution examples.

Reinforcement learning: In reinforcement learning the goal is to develop an autonomous agent that learns to perform a task by acting in an environment. The agent receives a reward signal for each action and is trying to maximize the cumulative reward.[110] Reinforcement learning requires a trade-off between exploration and exploitation, with exploration referring to taking actions in order to obtain new training data and exploitation to taking actions that we know will achieve a high reward. DeepMind demonstrated that a reinforcement learning system based on deep learning was able to learn playing Atari video games reaching human-level performance, without being trained on past games, having as goal to maximize the game score.[111]

Unsupervised learning: Unlike supervised learning, where the ground truth is known, or reinforcement learning, where a proxy of the label can be achieved by querying the environment, in unsupervised learning we are dealing with unlabelled data or data with unknown structure. This structure can be derived based on the relationships among the variables, by a technique called clustering. Clustering is an exploratory data analysis technique that allows us to organise a pile of meaningful subgroup (clusters) without having any prior knowledge of their group memberships.[110] Each cluster defines a group of objects that share a certain degree of similarity and simultaneously they are more dissimilar to objects in other clusters. Clustering is a technique for structuring information and deriving meaningful relationships from data. Unsupervised learning has found significant applications in medical imaging, where a quicker categorization of the patients based on disorders can be done.[112]

Semi-supervised learning: The intersection of supervised and unsupervised learning is semi-supervised learning. This type of learning is used when the training set has both labelled and unlabelled data.[113–115] Semi-supervised learning can be particularly useful for medical images. For instance, a radiologist can label a small subset of the scans for tumours or diseases and that will improve the predictions of which patients require more medical attention.[116] A particular subclass of this method which is of interest and studied in this thesis is that of one-class classification. This is a scenario where there are two classes (positive and negative), however only the labels of one class are available for some of the data points. One class classification method has been well-studied with various algorithms implemented for tackling problems as anomaly or novelty detection.[117]

2.1.3 Traditional machine learning models and dimensionality reduction

Any machine learning model can be described by the following function:[118]

$$F(x, w) = y \qquad (2.1)$$

where x is the input, w the learnable parameters and y the predicted output. The performance of the model is then evaluated using a loss/error function, which is a function that maps an event or the values of one or more variables onto a real number representing some "cost" associated with the event. In the case of supervised learning the loss function becomes:

$$Loss \ = \ mean((y - y_{true})^2) \qquad (2.2)$$

where y_{true} represents the true label of a datapoint. ML algorithms seek to minimize the loss function or else solve an optimization problem. As traditional machine learning models, we usually refer to logistic regression, SVMs, random forests, k-nearest neighbours.[119,120] A detailed description of the traditional ML models used for one class classification is given in Chapter 5, Section 5.2.

With regards to the features used as input, the application of traditional machine learning models requires manually discovering and creating relevant features with a process known as feature engineering.[121] The need for feature engineering is often related to the 'curse of dimensionality',[122] meaning that the amount of data needed to estimate a function with a given level of accuracy grows exponentially with respect to the dimensionality of the input variables of the function. Dimensionality reduction is important for data science as a technique for both visualization and as pre-processing before a ML algorithm is applied. Several methods can be used for reducing the dimensionality of the features, e.g. Principal Component Analysis (PCA)[102] and UMAP (Uniform Manifold Approximation and Projection).[123] PCA is used for reducing the dimensionality of a dataset when there are significant correlations between some or all the features. In that way the variation of the data can be explained by a small number of principal components and the global structure of the data is preserved. On the other hand, UMAP uses a more complex technique trying to preserve both the local and the global distances, *i.e.*, if two datapoint are close in the high dimensions, then they will be also close in the lower dimensions, whilst the distance structure in the data is well preserved.

2.1.4 Artificial Neural Networks

Artificial Neural Networks (ANNs) are a subclass of machine learning models that try to model the way in which brain performs a particular task.[124] In its simplest form, an ANN is called a perceptron and as shown in Figure 2.2 it is composed of three layers, an input layer, a single hidden layer, and an output layer. The perceptron accepts the inputs $x_1,x_2,\ldots x_n$ and moderates them by multiplication with the weight values $w_1,w_2,\ldots,w_n$. The summation function Σ is adding the adjusted weights and then an activation functions *e.g.,* a sigmoid function, converts the numerical output to +1 or -1. The predicted output is compared with the known label and the error is backpropagated to allow for a further weight adjustment.

What is specific to neural networks is how the weights are updated using the loss function. Neural networks must be differentiable *i.e.*, they are composed of smooth continuous functions, so the weights (learnable parameters) can be updated with the following function:

$$w_i = w_i - \eta \frac{dloss}{dw_i} \qquad (2.3)$$

with η being the learning rate and $\frac{dloss}{dw_i}$ being the gradient across every single parameter of the network. This process continues to iterate until the model converged on the training data.

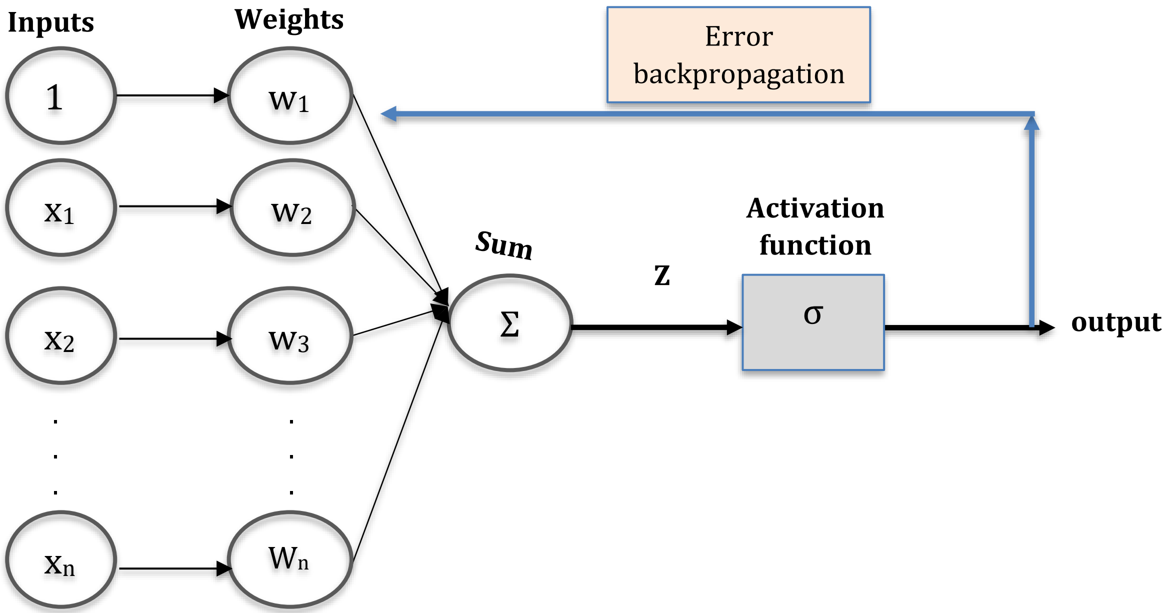


Figure 2.2 *Perceptron architecture.*

Deep learning models are neural networks, based on ANNs, that possess multiple hidden layers and thus the network is considered "deep". The information is propagated in a layer-wise fashion, as each layer receives the output of the previous layer. A trained model refers to a neural network architecture along with learned weights connecting all its neurons.

The advantage of using deep learning models instead of traditional machine learning techniques, is mainly that the input features are not hand-engineered but learnt from the data. In other words, deep learning algorithms perform a type of feature learning, *i.e.*, representation learning instead of feature engineering used in the traditional ML models.

2.1.5 Autoencoders

An autoencoder is a neural network that is trained with the task to recreate its input.[110] The network consists of two parts, an encoder that maps the input to a hidden layer *h by creating a compressed feature set* and a decoder that given h tries to reconstruct the input. Autoencoders are designed to copy approximately and only the input that resembles the training data. For that reason, the model is able to prioritize the aspects of the input that are important, learn useful properties of the data and being used as a dimensionality reduction or feature learning technique. The autoencoder is trained with a purpose to minimize its reconstruction loss:

$$Loss = mean((g(f(x)) - x)^2) \qquad (2.4)$$

where $g(f(x))$ is the reconstruction of the input, *x* the input and Loss is a loss function penalizing $g(f(x))$ for being dissimilar from *x*. After being trained, the part of the autoencoder that is more useful is the latent dimension *h*, as it can be used as a reduced-dimensions representation of the data.

2.1.6 Transformers

Transformers are a type of deep neural networks which have been primarily designed for language translations. The network learns word embeddings from their contextual usage proving an expressive dense representation that captures relationships between words.

The building block of a transformer is the attention mechanism.[125] Attention is described from the following formula:

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (2.5)$$

where Q is the query, K is the key and V is the value matrices of dimension d_k. In a machine translation task, the encoder computes one vector from every input word resulting in a context matrix. The context matrix information is used for K and V. During the decoding process, the decoder queries the context matrix using Q for getting the most relevant information to predict the next words. The attention function returns the values weighted by how aligned keys and queries are. The output of the softmax function, namely attention weights, can be visualized to show where the decoder is focusing on to predict the output words.

Although the transformer models were designed for natural language problems, they have impacted significantly in a variety of fields. Several types of transformers have been recently developed for tacking a wide range of problems.[126,127] SetTransformer is a type of transformers used for problems involving sets and is described in detail in Chapter 6, Section 6.2.2.

2.1.7 Explainable AI

For proving that a model is useful in materials science, a high accuracy is not enough. Knowing the physical meaning of the predictions is advantageous for ML models. Several methods for model interpretability have been introduced to shed light on the internal decision processes of neural networks showing which features are salient to final predictions. The two broad computational approaches towards explainable AI that have been used in materials science are the feature attributions and the graph convolution based.[128] Feature attributions are quantifying the impact of removing features on the predictive performance. In contrast, graph convolution-based methods are mainly using attention to highlight the parts on the input that are most importance in the performance of the model. Interpretable graph neural networks have shown promise in better understanding the chemistry behind the predictions.[129] This work is using feature attribution techniques to learn the important motifs without expert-encoded knowledge.

A widely used feature attributions methods is LIME (Local Interpretable Model-agnostic Explanations), which is locally approximating the model around a given prediction.[130] SHAP (Shapley Additive exPlanations) is an extension of LIME which was employed in this work as a model interpretation framework. SHAP is a model independent method, meaning that it does not take into consideration the feature weights but measures the influence each feature change has on the final decision of the model.[131] In other words, by calculating Shapley values, the contribution of each feature to the final score is estimated. The overall SHAP formula is shown in equation (2.6), where g is the explanation model, M is the number of simplified input features, $\varphi_i \epsilon \mathbb{R}$ is the feature attribution for a feature i, $z' \epsilon \{0,1\} M$, and φ_0 represents the model output with all the simplified inputs missing.

$$g(z') = \varphi_0 + \sum_{\iota=1}^{M} \varphi_i z'_i \qquad (2.6)$$

To obtain the contribution of a feature *i*, all operations by which a feature might have been added to the set (N!) and a summation over all possible sets (S) is considered. For any feature sequence, the marginal contribution through addition of feature *i* is given by [f(S∪{i}) − f(S)], where f(S) corresponds to the output of the ML model. The resulting quantity is weighted by the different possibilities the set could have been formed prior to feature i's addition (|S|!) and the remaining features could have been added ((|N| − |S| − 1)!). Hence, the importance of a given feature is defined by equation (2.7):

$$\varphi_i = \frac{1}{N!}\sum_{S\subseteq \mathbf{N}\setminus\{\mathbf{i}\}} |S|!\,(|N| - |S| - 1)!\,[f(S \cup \{i\}) - f(S)] \qquad (2.7)$$

It follows that Shapley values represent a unique way to divide a model's output among feature contributions satisfying three axioms: local accuracy (or additivity), consistency (or symmetry), and nonexistence (or null effect).[132]

2.1.8 Evaluation Metrics

ML models should always be tested on previously unseen data to ensure their generalization and extrapolation ability. The ML problem encountered in this work is clustering , *i.e.*, building a model for deciding if a pair of molecules can form a stable crystal structure or not. Several evaluation metrics exist for clustering models, which are based mainly on the size and the balance of the evaluation data. For the development of our ML model for screening a small subset of co-crystals, namely π-π co-crystals, only positive data for both training and validation were available. Consequently, the selected evaluation metric is the True Positive Rate (TPR):

$$Rate\ (TPR)\ =\ \frac{1}{K}\sum\frac{correctly\ predicted\ inliers}{size\ of\ training\ set\ in\ each\ fold} \qquad (2.8)$$

where K is the number of folds in cross validation and as inliers, we define the positive datapoints. The training dataset was split into K groups (K=5) and for each one of k iterations a unique group is considered as the test set, whereas the remaining groups comprise the training set. Each time a model is fitted on the training set and then is evaluated on the test set.

After the scaling-up on the ML model to cover all the existing types of co-crystals, both positive and negative validation data exist and thus the common evaluation metrics for balanced binary data were used as describe in Chapter 6, Section 6.2.4.

2.1.9 Databases and Cheminformatics software

The successful application of ML models depends mainly on the amount and quality of data that is available. Manually curated databases have largely grown according to the need for technological innovation, understanding life, characterizing structures and synthetic chemistry.[1] The two main databased used in this work are:

1) CSD Database: The Cambridge Structural Database (CSD) is the world's repository for small-molecule organic and metal-organic crystal structures, containing over 1 million structures acquired from X-ray and neutron diffraction analyses[133]. CSD was the main source of knowledge for the further investigation of the up to date known cocrystals (https://ccdc.cam.ac.uk/).

2) ZINC Database: ZINC15 and the most recent version of ZINC20 is a free public access database that contains over millions of purchasable organic compounds (https://zinc15.docking.org/substances/home/).

The basic cheminformatics software and computational tools used for this work is presented below:

1) Pipeline Pilot: Pipeline pilot (version 2017) is a BIOVIA's graphical scientific authoring application that offers advanced analytic tools. Herein, it was implemented for reading the SMILES of the organic molecules acquired from ZINC Database and drop the duplicate structures. Moreover, a protocol was used to keep only the organic molecules that do not have acidic hydrogens.

2) Dragon software: Dragon version 6.0/2012 was implemented for molecular descriptors calculation. Dragon descriptors can be used to evaluate molecular structure-activity or structure-property relationships, as well as for similarity analysis and high throughput screening of molecule databases. Dragon provides almost 5,000 molecular descriptors that are divided into 29 logical blocks, each in turn divided into a number of sub-blocks to allow easy retrieval of the molecular descriptors of interest. The user can calculate not only the simplest atom types, functional group and fragment counts, but also several topological and geometrical descriptors. Some molecular properties such as logP, molar refractivity, number of rotatable bonds, H-donors, H-acceptors, molecular volume and surface areas are also calculated by using some common models taken from the literature[134].

3) Mordred library: A freely available python library, which can calculate more than 1800 numerical representations of molecular properties and/or structural features using predefined algorithmic rules.[135] Disadvantage of this approach is that the library is not further updated and as a result many packages start deprecating producing many nan (non a number) values.

4) CCDC software: Basic utilities from CCDC software that were used in this thesis are:[136,137]

i) Mercury - Software providing tools for visualizing 3D structures and running calculations.

ii) CCDC Python API – Allows for writing code in Python capable of searching the Crystal Structural Database (CSD)

iii) Conquest – A search software enabling advanced searching of the CSD data after applying user-defined search constraints.

iv) Isostar – A library providing information regarding intermolecular interactions from CSD.

5) Crystal Structure Predictions – USPEX software: Crystal structure prediction is an optimization task, which involves the identification of the positions of the atoms in the unit cell such that the system of interest yields a desired response, *i.e.*, the lattice energy of the crystal is minimized. The energetically favoured solid forms are considered as the most stable and possible experimental observations. USPEX code is using an evolutionary algorithm coupled with *ab initio* structure relaxations. USPEX is currently the only CSP software that can handle both organic and inorganic crystals and thus was used for CSP calculations on the metal-PAHs systems.[138]

6) Zeo++ software: Void space analysis of crystal structures was performed with ZEO++. The tools implemented by Zeo++ software are based on the Voronoi decomposition, which for a given arrangement of atoms in a periodic domain provides a graph representation of the void space and in that way the atomic connectivity is determined. In more detail, for each atom in the lattice, the Voronoi cell is constructed around that atom. Consequently, the material space is divided into irregular polyhedral cells which are analysed to determine the pore topology.[139,140]

2.2 Introduction to computational chemistry

Computational chemistry is the study of chemical systems through computer simulations, typically based on a theoretical framework describing the behaviour of electrons, atoms and molecules. Running simulations allows users to analyse chemical systems at an atomistic level, not easily accessible through experiment. As such, simulations can provide insight into fundamental processes occurring in bulk environments, which may be difficult to probe experimentally. Assuming simulations are performed to a sufficient degree of accuracy, they also facilitate property prediction for simulated materials. This has become an important tool in the field of materials discovery and design, as simulations are usually much cheaper, faster and easier to perform at scale than experiments. This means that materials can undergo an initial computational screening, after which synthetic resources are focused only on promising candidates, thereby increasing the rate and decreasing the cost of materials discovery.

The two general methods of computational chemistry are molecular mechanics and electronic structure methods, with the latter including *ab initio* and semi-empirical methods. An overview of the basic concepts regarding Density Functional Theory (*ab initio* method) and semi-empirical methods is given below:

2.2.1 *Ab Initio* methods: Density Functional Theory

In quantum systems, a particle such as an electron does not have an exact location. Instead, its position is described by a probability density. Despite the complexity of the problem, the basis of the theory can be reduced down to a few straightforward equations. These equations are sufficient to describe the behaviour of all the familiar matter we see around us at the level of atoms and their nuclei. Their counterintuitive nature leads to all sorts of exotic phenomena: superconductors, super fluids, lasers, and semiconductors that are only possible because of the quantum effects. But even the covalent bond, which is the basic building block of organic chemistry, is a consequence of the quantum interactions of electrons. Once these rules were worked out in the 1920s, scientists realised that, for the first time, they had a detailed theory of how chemistry works. In principle, they could just set up these equations for different molecules, solve for the energy of the system, and figure out which molecules were stable and which reactions would happen spontaneously. However, the actual calculation of the solution to these equations was only possible for the simplest atom (hydrogen) as everything else was too complicated.[141,142] The most famous of these equations, the Schrödinger equation, describes the behaviour of particles at the quantum scale and thus the electronic structure of a material according to the following equation:

$$\widehat{H}(r;\ R)\,\Psi\ (r;\ R) =\ \ E\ (R\ )\Psi\ (r;\ R) \quad (2.9)$$

Where $\widehat{H}$ is the Hamiltonian operator, Ψ is the electronic wavefunction dependent on r (electron coordinates), E is the energy and *R* the nuclei coordinates. The conceptual framework of density functional theory is starting from the problem of solving the Schrödinger equation in a many particles system.

Although the many-body Schrödinger equation is unsolvable due to the high dimensionality of the problem, several approximations have been introduced to reduce the number of variables and find accurate solutions. The first approximation is known as the Born-Oppenheimer approximation followed by the Hohenberg-Kohn and the Kohn-Sham approximations which are explained in the following paragraphs.

Born-Oppenheimer approximation

The Born-Oppenheimer approximation neglects the motion of the atomic nuclei when describing the electrons in a molecule, as the mass of an atomic nucleus is much larger than the mass of an electron (approximately 1000s of times) and thus the nuclei move much slower than the electrons.

The many-body Schrödinger equation using the Born-Oppenheimer approximation can be written as:

$$\widehat{H}\Psi = \left[\widehat{T} + \widehat{V}_{ext} + \widehat{V}_{int} + E_{II}\right]\Psi =\ \ E\Psi \qquad (2.10)$$

where $\widehat{T}$ is the kinetic energy operator, $\widehat{V}_{ext}$ is potential energy associated with the nuclei, $\widehat{V}_{int}$ describes the electron-electron interactions, E_{II} is the classical interaction between nuclei, and E is the energy.[143]

Equation (2.10) is computationally infeasible to solve for more than tens of particles because its complexity scales exponentially with the number of particles.

Hohenberg-Kohn theorems

For that reason, the two Hohenberg-Kohn theorems apply to enable an easier computed form:[144]

i) for a given system $\left\{\{\widehat{V}\}\right\}_{\{ext\}}(r)$ and thus the total energy of a system is a functional of the ground state charge density, n(r) descripted as:

$$E[n(r)] = \int dr\,\widehat{V}_{ext}(r)n(r) + F[n(r)] + E_{II} \qquad (2.11)$$

where F$[n(r)]$ includes kinetic energy and electron-electron interactions, and

ii) the density that minimizes the total energy is the ground state charge density.

These two theorems combined imply that the ground state total energy of a system of interacting electrons and nuclei is determined by the ground state electron density. We only need to know the energy functional $E[n(r)]$ to solve for the ground state charge density variationally. While a wavefunction has 3^N variables, a charge density has only three (one for each spatial dimension). The greatly reduced number of degrees of freedom needed to describe the total energy and density of interacting many-body systems makes DFT a potentially powerful approach compared to more expensive wavefunction based methods.

Kohn-Sham equation

While the Hohenberg-Kohn theorem proves the existence of a universal functional, it does not provide a way to determine this functional. Soon after the appearance of the Hohenberg-Kohn theorem in 1964, Kohn and Sham found a way to map the many-body problem to that of a single electron in an external potential of non-interacting electrons in 1965. In this formulation, solving for the ground state charge density of this effective non-interacting system leads to the same ground state charge density of the true interacting many-body system; the ground state wavefunctions of the effective non-interacting system, obtained from the Kohn-Sham equations are used to determine ground state charge density of the many-body system. More explicitly, the Kohn-Sham approach reformulates the Hohenberg-Kohn expression for the ground state functional as:

$$E_{\{KS\}} = T_{S[n]} + \int d\, r \hat{V}_{ext}(r)\mathrm{n}(r) + E_{\{Hartee\}[\mathrm{n}]} + E_{II} + E_{XC}[n] \qquad (2.12)$$

where $T_{S[n]}$ is the kinetic energy of non-interacting electrons, $E_{\{Hartee\}[\mathrm{n}]}$ is the mean-field Coulomb interaction energy of the electron density, and $E_{XC[\mathrm{n}]}$ is the exchange-correlation functional. Modern DFT relies on the Kohn-Sham equations which are solved self-consistently. The Kohn-Sham Schrödinger-like equations are given as:[143]

$$(H_{KS}^{\sigma} - \varepsilon_i^{\sigma})\psi_i^{\sigma}(r) = 0 \qquad (2.13)$$

where the ε_i are the eigenvalues, and H_{KS} is the effective Hamiltonian

$$H_{KS}^{\sigma} = -\frac{1}{2}\nabla^2 + V_{KS}^{\sigma}(r) \qquad (2.14)$$

where $V_{KS}^{\sigma}(r) = \hat{V}_{ext}(r) + \frac{\delta E_{Hartee}}{\delta n(r,\sigma)} + \frac{\delta E_{XC}}{\delta n(r,\sigma)} = \hat{V}_{ext}(r) + \hat{V}_{Hartee}(r) + V_{XC}^{\sigma}(r) \qquad (2.15)$

In equation (2.12) the first three terms are known and can be straightforwardly solved for. The $E_{\{XC\}[\mathrm{n}]}$ term is in general unknown and expresses the difference in kinetics and potential of an interacting versus a non-interacting system and is given from:

$E_{\{XC\}[\mathrm{n}]} = \langle\widehat{T}\rangle - T_S[n] + \langle\hat{V}_{int}\rangle - E_{Hartee}[n]$ (2.16)

Density functionals - XC Energy Term Approximation

The quality of the density functional approach relies on the accuracy of the chosen approximation to $E_{\{XC\}}$. Each density functional approximates the $E_{\{XC\}}$ energy term by adding different variables. An overview of the most important functionals is going to be given below.

Basic functionals: The two main types of exchange correlation functionals used in DFT are the local density approximation (LDA) and the generalized gradient approximation (GGA).

The local density approximation (LDA) functional can be regarded as the basis system of all exchange-correlation functionals. The central idea of this model is the existence of a hypothetical uniform electron gas, which is a system of electrons moving on a positive background charge distribution forming an electrically neutral environment. The main characteristics of that environment are the number of electrons N and the volume of the gas V. Whilst both these parameters approach infinity, their ratio N/V, which represents the electron density remains finite and attains a constant value ρ everywhere. Based on that, the $E_{\{XC\}}$ for the LDA is linearly dependant to the charge density and is given from the following equation:

$E_{\mathrm{XC}}^{LDA}[\rho] = \int d\,\vec{r}\rho(\vec{r})\varepsilon_{XC}^{homogen}\big(\rho(\vec{r})\big)$ (2.17)

where $\varepsilon_{XC}^{homogen}\big(\rho(\vec{r})\big)$ is the exchange-correlation energy per particle of a uniform electron gas of density $\rho(\vec{r})$. The energy per particle is then weighted with the probability $\rho(\vec{r})d\vec{r}$ of finding an electron at this position in space. The idea of a uniform electron gas could be an appropriate model for simple metals. However, it cannot be representative for atoms and molecules which are characterized by rapidly varying electron densities. Nonetheless, LDA is not accurate enough for chemical applications as it tends to overbind resulting in structures that have smaller lattice parameters than experiment and generally underestimates band gaps. Hence, more sophisticated approximations were developed. The generalized gradient approximation (GGA) is often implemented as a corrective function of the LDA and includes corrections for gradients in the electron density.

$E_{\{XC\}}^{\{GGA\}[\mathrm{n}]} = \int d\,r\varepsilon_{XC}^{GGA}\big(\rho(\vec{r}),\ \nabla\rho(\vec{r})\big)$ (2.18)

The functional by Perdew, Burke and Ernzerhof (PBE) is a specific functional based on the GGA. PBE does not treat van der Waals (vdW) dispersion interactions which is a non-local correlation effect. PBE tends to underbind resulting in structures that tend to have larger lattice parameters than experiment. PBE also generally underestimates band gaps.

van der Waals dispersion and *π-π* interactions: Before referring to the functionals that include dispersion corrections, we will introduce this type of interactions with a focus on the systems with π-π bonding. vdW interactions are essential for the description of the structure, stability and properties of many molecular systems.[145] These distance-dependent forces arise from electrostatic interactions between fluctuations in the electron charge density. Although vdW forces are relatively weak and they are considered to have a small contribution to the total energy of a system, they play a key role in accurate description of molecular systems and materials.

In π-π systems that involve aromatic molecules, which are of interest in this work, there are theories that support that the vdW interactions between the electron clouds around the molecules enforce their stabilization. The aromatic ring is seen as such there is a partial negative electrostatic potential above the two aromatic faces and a partial positive electrostatic potential around the periphery of the aromatic molecule.[146] Consequently, the π-π interactions take place not due the attractive electronic interactions between two π-systems but because π-σ attractions outweigh unfavourable contributions such as π-electron repulsion.[147]

Functionals that include dispersion corrections: Although the development of new functionals have greatly improved the accuracy of DFT calculations, systems in which van der Waals dispersion interactions play a significant role cannot be effectively modelled. For that reason, developing vdW-inclusive methods has been one of the important fields of development in DFT in the last decade. The methods of dispersion correction led to further improvements in accuracy and broader applicability in more complex systems.

The van der Waals density functional (vdW-DF) was first developed for including dispersion in approximate density functional theory exchange-correlation functionals.[148] Using the vdW-DF method a broad range of systems, *e.g.*, metals, insulators, ionic compounds, held by dispersion forces can be effectively described. However, it was found that vdW-DF overestimates lattice constants and also be inferior for a range of systems, *e.g.*, systems with hydrogen bonds. To overcome these limitations, new methods were developed such as the PBE+D3 and the SCAN+rVV10.

PBE+D3: Perdew–Burke–Ernzerhof (PBE) approximation. The new parameters introduced on PBE+D3 are the atom-pairwise specific dispersion coefficients and a cutoff radii that are both calculated from first principles.[149] In the D3 correction of Grimme *et al.*, the following vdW-energy expression is used:

$$E_{disp} = -\frac{1}{2} \sum_{i=1}^{Nat} \sum_{j=1}^{Nat} \sum_{L} (f_{d,6}(r_{ij,L}\,) \frac{C_{6ij}}{r_{1j,L}^{6}} + f_{d,8}(r_{ij,L}) \frac{C_{8ij}}{r_{1j,L}^{8}}) \qquad (2.19)$$

where $f_{d,n}$ are damping functions used for determining the range of the dispersion correction, C_{6ij} are the dispersion coefficients, C_{8ij} are the dipole-quadrupole interactions, *Nat* is the number of atoms and $r_{ij,L}$ the atomic distances.

SCAN+rVV10: The 'strongly constrained and appropriately normed' (SCAN) meta-generalized gradient approximation (meta-GGA) can generally improve over the non-empirical Perdew-Burke- Ernzerhof (PBE) GGA not only for strong chemical bonding, but also for the intermediate-range van der Waals (vdW) interaction.[148]

K-points and reciprocal space

DFT calculations are applied to atoms that are located within the volume of a solid and can be specified as point positions in a three-dimensional Euclidean space called direct or real space. The atoms belong to a unit cell which is periodically repeated in the space forming an infinite crystal. Given this periodic system defined by lattice vectors **a**$_1$, **a**$_2$, **a**$_3$ the solution of the Schrödinger equation must satisfy Bloch's theorem which can be expressed as:

$$\varphi_k(r) = exp(ikr)u_k(r) \qquad (2.20)$$

where $u_k(r) = u_k(r + n_1a_1 + n_2a_2 + n_3a_3)$ for any integers $n_{1,}$ $n_{2,}$ n_3. The space of vectors **r** is called real space and the space of vectors **k** is called reciprocal space. Many parts of the mathematical problems posed by DFT are more convenient to solve in terms of **k** than in terms of **r**. Thus, the reciprocal k vectors can be expressed as a mapping from the real vectors **a**$_1$, **a**$_2$, **a**$_3$ as:

$$b_1 = 2\pi \frac{a_2 \times a_3}{a_1\,(a_2 \times a_3)}, \quad b_2 = 2\pi \frac{a_3 \times a_1}{a_2\,(a_3 \times a_1)}, \quad b_3 = 2\pi \frac{a_1 \times a_2}{a_3\,(a_1 \times a_2)} \quad (2.21)$$

A general vector of the form

G= h**b**$_1$ + k**b**$_2$ +l**b**$_3$ $\qquad$ h,k,l = integers $\qquad$ (2.22)

generates a reciprocal lattice. The integers h,k,l are called the Miller indices of a lattice plane and define lattice planes and directions in the lattice.

Plane waves and pseudopotentials

To represent wavefunctions, one needs a basis set. As the materials we are working on are crystal systems, we use plane-waves. The plane-wave basis depends on the crystallographic lattice parameters of the input unit cell and an energy cutoff (E_{cutoff}). For a given E_{cutoff}, all planewaves satisfying the following equation are included:

$$\frac{h^2}{2m_e}|G + k| < E_{cutoff}, \qquad (2.23)$$

where G is a reciprocal lattice vector and k is a vector in reciprocal space within the first Brillouin zone. For many systems, including those described in this thesis, it is computationally expensive to treat all electrons independently. Like for all electrons, the wavefunctions of core electrons must be orthogonal to one another. Because core electrons exist in a rather confined region near the nucleus, this requires core electron wavefunctions to oscillate and defines their nodal structure. To accurately describe these core wavefunctions, one must use a basis that has a resolution comparable to these oscillations, which can be hundredths of Angstroms. For example, for Ecuttoff = 520 eV, the resolution is approximately a tenth of an Angstrom.

$$\frac{h^2}{2m_e}\left|\frac{1}{1.2\ x\ 10^{-11}\ meters}\right|^2 \cong 520\ eV \qquad (2.24)$$

To achieve a resolution of approximately a hundredth of an Angstrom, E_{cutoff} would have to be increased a hundred-fold. Moreover, core electrons are highly localized and well-separated in energy from valence electrons, which are crucial to determining structural and electronic properties. Therefore, it is a good approximation to freeze them into an effective core and neglect core degrees of freedom in solving the Kohn-Sham equations. Thus, rather than treating core electrons directly, we use pseudopotentials which combine the nuclear and core electron contributions and create a smooth potential for valence electrons. Current DFT codes provide a library of potentials of different elements. The Projector-Augnented-Wave (PAW) method is a technique used for calculating the pseudopotentials. Following this approach, the rapidly oscillating wavefunctions are transformed into smooth wavefunctions which are more computationally convenient.

2.2.2 Hartree-Fock approximation

Hartree-Fock theory is one of the simplest approximate theories for solving the many-body Schrödinger equation (eq. 2.9), requiring that the electrons are independent particles. Herein, the motions of the electrons in the molecular orbitals are approximated by a sum of the motions of electrons in the atomic orbitals. The electronic wavefunction is expressed by combining one-electron wavefunctions in a way that satisfies the antisymmetry principle. That can be achieved by using a single Slater determinant, *i.e.*, the determinant of a matrix of single electron wavefunctions. The Slater determinant for the case of two electrons is given as:

$$\psi(x_1, x_2) = \frac{1}{\sqrt{2}} det \begin{bmatrix} \chi_1(x_1) & \chi_2(x_1) \\ \chi_1(x_2) & \chi_2(x_2) \end{bmatrix} = \frac{1}{\sqrt{2}} [\chi_1(x_1)\chi_2(x_2) \quad \chi_2(x_1)\chi_1(x_2)] \qquad (2.25)$$

where the coefficient $\frac{1}{\sqrt{2}}$ is a normalization factor and $\chi_1(x_1)$ is a spin orbital with x_1 being a vector of coordinates that defines the position of the first electron and its spin state. In general, the rows of the Slater determinant correspond to the electron and the columns to the spin orbital. The Slater determinant can be generalized to a system of *N* electrons by forming an *N x N* matrix of single electron spin orbit. It can be written in a conventional form by listing the spin orbitals χ as:

$$\psi = |\chi_i \chi_j \ldots\ldots \chi_k\rangle \qquad (2.26)$$

where *i,j,k* are the indices of the spin orbitals. The Hartree-Fock method is then trying to solve the wavefunction for those orbitals that minimize the electronic energy, which is mathematically equivalent to assuming each electron interacts only with the average charge cloud of the other electrons. Each spin orbital $\chi(x)$ is a function of four coordinates χ(x, y, z, ω) and can be written as a product of a spatial part *φ(**r**)* and a spin part *σ(ω)*:

$$\chi(x) = \varphi(\boldsymbol{r})\, \sigma(\omega) \qquad (2.27)$$

The Hartree-Fock energy expression that should be minimized can be written in terms of integrals of one- and two-electron operators:

$$E_{HF} = \underbrace{\sum_i^{elec} \langle i|\widehat{h}|i\rangle}_{\text{one-electron term}} + \underbrace{\sum_{i>j}^{elec} [ii|jj] - [ii|jj]}_{\text{two-electrons term}} \qquad (2.28)$$

The one electron integral refers to the electron kinetic energy and the electron and nuclei attraction, whereas the two-electron integral represents the Coulomb repulsion between electron 1 in orbital *i* and electron 2 in orbital *j*.

The two-electrons term may be called one-, two-, three- and four-centre integral depending on the values of the indices. The most expensive task of Hartree-Fock is evaluating and transforming the two-electron integrals and thus several semiempirical models were developed to approximate that task.

2.2.3 Semi-Empirical methods

Semiempirical (SE) methods can be derived by applying systematic approximations either on Hartee-Fock (HF) or density functional theory (DFT) level, resulting in calculations that are several orders of magnitude faster than the *ab initio* computational schemes.[150] The most prevalent SE methods based on the approximations to HF theory are AM1, PM3, PM6 MNDO/d and OMx. An approach focusing on approximating DFT that has become popular in the past decade in the density functional tight binding (DFTB) method which is based on a Taylor expansion of the energy with respect to a reference density. In this work, PM6 method was used for geometry optimization and orbital energy calculations. A brief overview of the evolution of the main HF-based semiempirical methods which drove to the development of PM6 is given below.

The HF-based semiempirical methods are trying to simplify the Hartree-Fock energy calculation by approximating the two-electron integrals of equation (2.28). These methods treat explicitly the valence electrons, and the names of the various methods are suggestive of which two-electron integrals are set as zero in the treatment.[65] The most primitive approach was the complete neglect of differential overlap (CNDO), where the two-electron integral is set to zero. The next level of approximation is the intermediate neglect of differential overlap (INDO), in which the $(ii|jj)$ is retained if χ_i and χ_j belong to the same atom. Following INDO, the neglect of diatomic differential overlap (NDDO) is introduced in which the differential overlap is neglected only when the basis functions belong to different atoms. According to the NDDO formalism, all one-centre two-electron integrals and not just the one-centre exchange integrals are retained.[65] Based on the NDDO, the modified neglect of differential overlap (MNDO) method was developed, where two main approximations exist: *i)* the two-centre two-electron integrals are replaced by approximate integrals derived from multipole interactions *ii)* there are improved core-core interaction terms in the one-electron operator. Although MNDO was a significant improvement, there were still deficiencies in particular regarding systems with hydrogen bonds. For that reason, an improved version of MNDO was developed namely Austin model 1 (AM1), which added up to four Gaussian functions to the core-core repulsion term to alleviate problems with short-range interactions. The continuous development and improvement of semiempirical methods as well as the incorporation of experimental data brought up a new more efficient method, called parametric model 3 (PM3) which as a reparameterization of AM1 using a different parametrization strategy and only two Gaussian functions to correct the core-core repulsion. Finally, after PM3 and due to the availability of an increasing amount of reference data to fit the parameters, the parametric model 6 (PM6) was introduced. Besides using a much larger set of reference data, PM6 also involves several improvements in the core-core terms by employing pairwise parameters rather than element-specific parameters. PM6 further uses different core-core repulsion potentials for N-H, O-H, C-C, and Si-O pairs to correct for specific weaknesses in the parametrization. Lastly PM6 method also adds d-orbitals in the atomic basis to certain elements.

3 Uncovering the structure-property relationships in polyaromatic hydrocarbons (PAHs)

3.1 Introduction

Since the discovery of graphene, which can be regarded as a giant PAH, the interest surrounding those materials has grown significantly. PAHs can be defined as a uniform class of very similar molecules built up by six-membered rings of sp^2-hybridized carbon atoms and hydrogens.[151] The molecular sizes covered can range from the simplest case of benzene with its six carbons up to disk-like molecules containing as many as 96 carbon atoms. Despite their similar atomic composition, PAHs dramatically differ in terms of optical and chemical properties depending on their size and geometry.[80,152]

Considering the size of the molecular space, which is estimated to contain about 10^{60} compounds,[153] the number of PAHs is far too large to be screened by a human. For that reason, computational screening is increasingly becoming a choice for an initial screening of large sets of compounds before any attempts of experimental realization. With an effective virtual screening methodology, the enormous molecular space is narrowed down, and the most promising targets could be identified.

In this work, we screened two molecular databases, *i.e.*, ZINC15 and ZINC20, for identifying PAHs which can *i*) serve as hosts for metal insertion to build materials showing superconductivity or other interesting electronic phenomena and *ii*) become co-crystal components to design semiconducting materials. Our main focus in this chapter was to examine how the structural properties of PAHs can be related to their electronic properties and identify the best ways to categorize them based on their similarity.

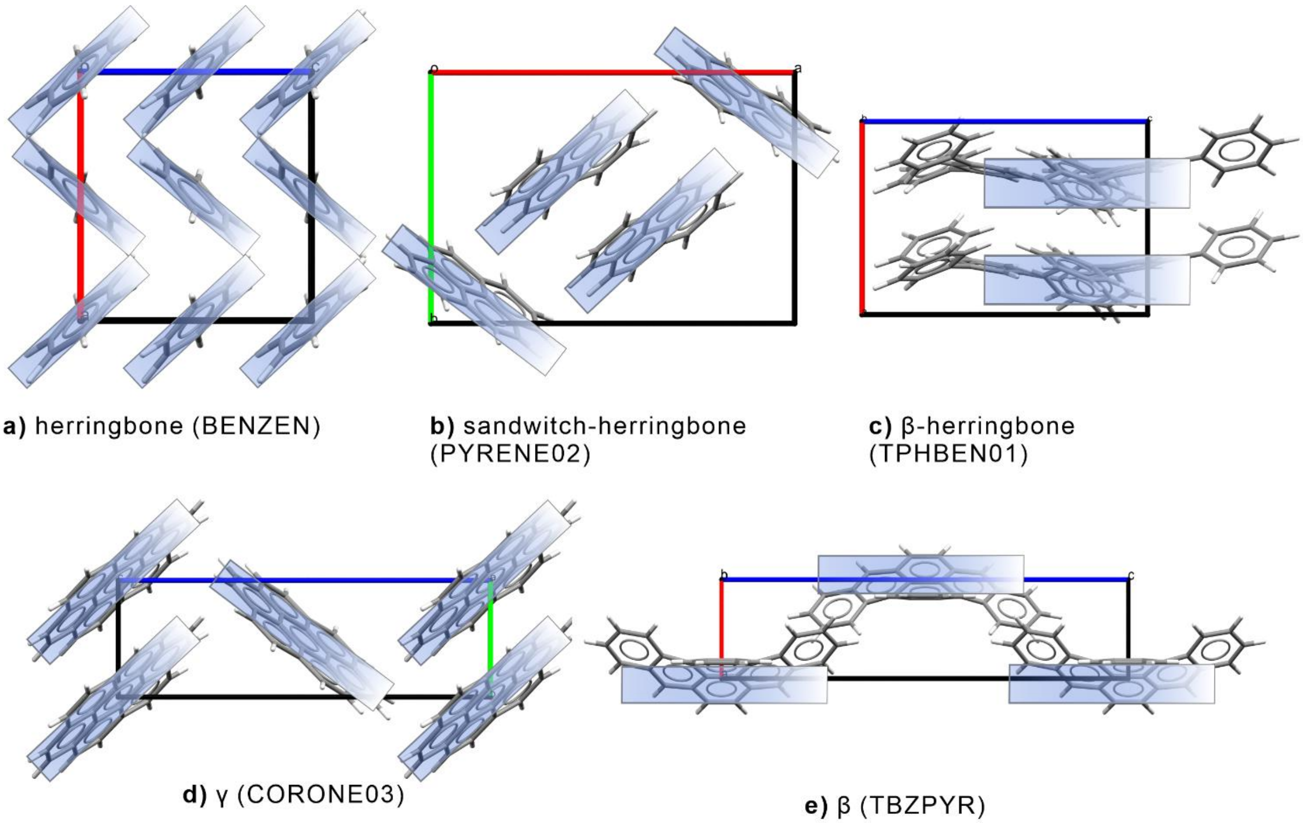
Starting from a small subset of 210 PAHs, we computed their equilibrium geometries and orbital energies (PM6 semiempirical model) and tested several representation techniques for describing the molecules. It was found that the representations that incorporate 3D information have higher correlation with the electronic properties (orbital energies). Using that representation, we divided the PAHs into classes of structurally similar molecules and investigated the electronic properties shared among these classes. Further on, we categorized a larger dataset of more than 7,000 PAHs based on their molecular shape and orbital energies distribution and further relations between the shape and electronic characteristics were extracted. The more structurally and electronically interesting readily available PAHs were listed to be further evaluated in the following chapter.

3.1.1 Structural and electronic analysis of PAHs

PAHs can be classified based on their topology to *i)* linear (comprised of homologous groups of oligoacenes, phenacenes, and oligorylenes), *ii)* circular flakes/discs (K-region PAHs and circumacenes), and *iii)* triangular.[154]

Within each topology two periphery types exist: (1) zigzag and (2) armchair.[154] Known relationships between the topological characteristics of PAHs and their electronic properties involve the observations that the band gap (Eg) decreases as the number of aromatic rings (or carbon atoms) increases, where arm-chair edge PAHs have larger band gaps and enthalpies of formation than their zigzag counter parts.[155] It is then understood that the Ionization Potential (IP) decreases, and the Electron Affinity (EA) increases with increasing number of rings in a homologous class.

Traditionally, PAHs can be either heterocyclic or only carbon containing and assemble in molecular crystalline arrays under ambient conditions. Those PAHs containing only hydrogen and aromatic carbon can be classified into five crystalline motifs as shown in Figure 3.1: *i)* herringbone, characterised by tilted edge-to-face C-H····π interactions, *ii)* sandwich-herringbone, where pairs of co-facial molecules make up the herringbone motif, *iii)* β-herringbone, observed in PAHs with σ-bound aromatic groups *iv)* γ, a flattened herringbone featuring stacks of parallel translationally related molecules and *v)* β, sheet-like packing of molecules.[156,157]


a) herringbone (BENZEN)

b) sandwitch-herringbone (PYRENE02)

c) β-herringbone (TPHBEN01)

d) γ (CORONE03)

e) β (TBZPYR)

Figure 3.1. *Motifs in PAHs. a) herringbone which is dependent on* C····H interactions, *b) sandwich-herringbone which is stabilized by both* C····H and C····C interactions, *c) β-herringbone, which is dependent on* C····C interactions, *d) γ, which is dependent on* C····C interactions, *e) β, which is dependent on* C····C interactions.

A famous theory incorporating structure-property relationships is Clar's rule, which refers to the effect of sextets in the stability and reactivity of PAHs.[80] A sextet is a grouping of the π-electrons within the aromatic ring, which is usually indicated by drawing a circle inside the ring. In cases where we have two consecutive aromatic rings, only one sextet can be formed. For instance, in anthracene (Figure 3.2a), only one sextet can be assigned to one of the rings whereas the remaining 8 π-electrons remain ungrouped. However, in triphenylene, all 18 electrons can be grouped into sextets and be assigned to each ring (Figure 3.2b). According to Clar's sextet rule, the electron sextets possess strong aromatic stabilization, whereas the bonds not included in the sextet are more susceptible to chemical reactions. It is concluded that for optical properties a balance among stability and reactivity should be found.[158]

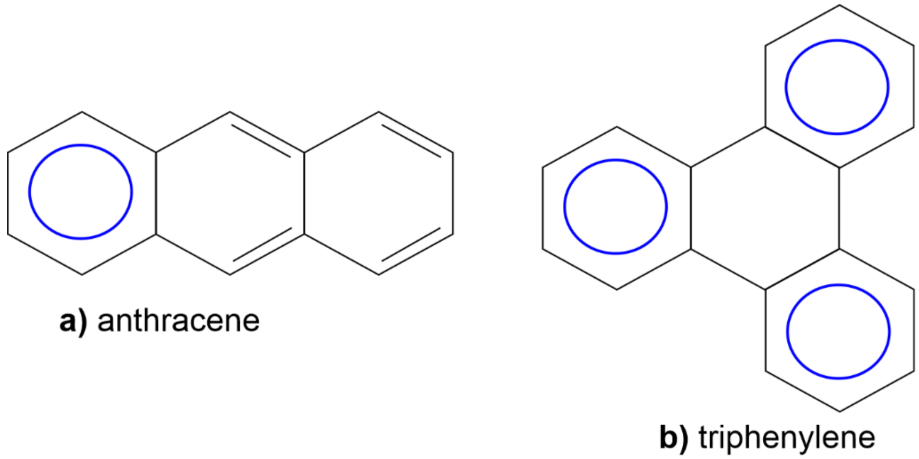


Figure 3.2. *The Clar structure in a) anthracene and b) triphenylene. The blue rings indicate the sextets, whereas the electrons not in a sextet remain as double bond.*

On another note, as PAHs have been reported as substrates in hydrocarbon-based superconductors,[74,77,159] the electronic properties that might be related to that phenomenon are highly sought after. There are several reports claiming that a near of exact degeneracy on the LUMO, LUMO$_{+1}$ orbitals play an important role for observing exotic electronic properties in PAHs.[160] That is because, when we gradually insert free electrons into a material, the first electron will occupy the lowest energy states (LUMO orbital), whereas the subsequent electrons will be forced to occupy higher energy states. According to the Pauli's rule the electrons of different spin pair together. Thus, the first electron will occupy the LUMO orbital, the second electron will again occupy the LUMO with different spin to form a pair. However, in the cases where Δ (LUMO$_{+1}$ – LUMO) is close to zero, the second electron will occupy the LUMO$_{+1}$ orbital and the system is going to have two unpaired electrons which makes the material able to conduct current because of the holes that are going to be formed. The effect of orbital degeneracy is more apparent in the C_{60} case, where their unique spherical shape is responsible for a triple degeneracy of the LUMO orbital and the unique distribution of the electronic potential[161]. That is the main reason why those molecules are more prone to reach the exotic states, such as superconductivity[162]. Research has also shown that high performing non-fullerene electron acceptors are characterized by very low gap between LUMO and $LUMO_{+1}$ $LUMO_{+2}$ orbitals, whereas the non-planarity of the acceptor might be beneficial for some classes of acceptors.[163] Of course, it is evident that for determining the electronic properties not only the type and shape of each single PAHs is important but also their configuration in the crystal lattice and their connectivity to their neighbouring molecules. The crystal structure of the PAHs will be considered in Chapter 4.

3.1.2 High-throughput screening for materials discovery

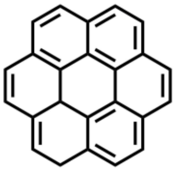
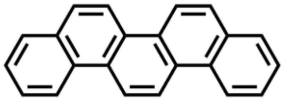
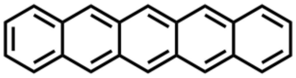
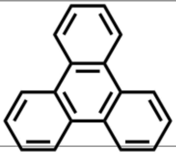
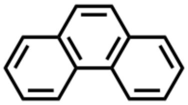
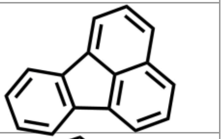
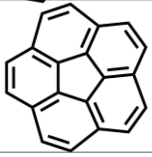
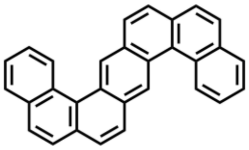
When searching for compounds with targeted properties for certain applications, large databases should be effectively used. That brings the development of screening workflows widely known as high-throughput virtual screening (HTVS). HTVS can be defined as the computational investigation of a large set of materials to assess their suitability for a particular function. The term 'large' is relevant and can range from hundreds to millions of materials. Taking into consideration the size of molecular space, which has been estimated to 10^{60}, a rational global search is extremely challenging. Starting from large and reliable databases containing a sufficient large number of known structures, HTVS can be applied by: *i*) using low-cost computational infrastructure, *ii*) applying cheminformatics tools, *iii*) employing robust quantum chemical methods, *iv*) following data science methods. HTVS of material databases has been so far increasingly successful in the discovery of new functional materials, with the most interesting finding being the identification of new patterns on the datasets and structure-property relations. Moreover, a list of top candidates based on the desired application can be easily extracted following these routes.

3.2 Methods

Dataset construction

An initial search was performed on ZINC15 database for purchasable molecules similar to the eight initial molecules shown in Table 3.1 on the basis of molecular fingerprints with a Tanimoto similarity threshold of > 0.35. The similarity search in ZINC15 is based on 512 bit ECFP4 fingerprints[164], meaning that the atomic environment between two under comparison molecules is four bonds long with size of fingerprint is 512 bits. Further on, Pipeline Pilot[165] was used for filtering out the incompatible groups, *i.e.*, acidic hydrogens, affording a library of 210 candidate molecules. For extending the initial PAHs dataset, ZINC20,[166] a new version of ZINC, which includes billions of new molecules and new search methodologies was implemented. Starting again from the same eight representative molecules as before, SmallWorld algorithm was used for similarity search in the whole 'in-stock' ZINC20 database. SmallWorld algorithm is reported as a graph-edit distance and maximum common subgraph method, meaning that the algorithm is first indexing the topological space of organic molecules into anonymous graphs and then connects each graph to its neighbours by elementary steps in graph- edit-distance space by adding/deleting a terminal atom, ring opening/closure, inserting/deleting a linker atom.[166] Following the same process as before, Pipeline Pilot was used to filter out the acidic groups affording an extended PAHs dataset of 7,060 molecules.

Table 3.1 *Initial Polyaromatic Hydrocarbons (PAHs) which differ according to shape and symmetry.*

CCDC Search Identifier	Zinc Search Identifier	Actual Name	Molecular structure
CORONE	ZINC0000001580987	CORONENE	
ZZZOYC	ZINC000001598876	PICENE	
PENCEN	ZINC000001581013	PENTACENE	
TRIPHE	ZINC000001688068	TRIPHENYLENE	
PHENAN	ZINC000000967819	PHENANTHRENE	
FLUANT	ZINC000008585874	FLUORANTHENE	
CORANN01	ZINC0000079045456	CORANNULENE	
DNAPAN	ZINC0000167079286	DINAPHTHO,(1,2 a:1',2'-h) ANTHRACENE	

Data representation

The molecules have been encoded using a diverse selection of molecular representations. Molecular descriptors incorporating several molecular properties have been extracted using Mordred library (Methods Section 2.1.9).[135]

Morgan fingerprint with vector length 1,024 bits was extracted from RDKit.[167] SOAP (Smooth Overlap of Atomic Positions) descriptors were also tested for transforming the atomistic structures into fixed-sized numerical fingerprints. The SOAP descriptor describes the neighbour density around each atom using radial and angular basis functions.[105] Molecular orbitals have been calculated using PM6 method and Spartan software. The electronic descriptor used is a vector of the orbital energies from $HOMO_{-1}$ to $LUMO_{+4}$.

Principal moments of inertia (PMI)

The shape analysis was performed using the principal moments of inertia as described by the NPR1 and NPR2 distributions. Here, we define NPR1 = I_1/I_3 and NPR2 = I_2/I_3, where I_1 is the first (smallest) Principal Moment of Inertia (PMI), I_2 is the second PMI, and I_3 is the third (largest) PMI. PMI descriptors assess the extent to which a given 3D molecular structure is rod-shaped, disc-shaped and sphere-shaped. Ternary plots are usually used for visualizing the PMIs with the top-left corner representing the purely rod-shaped, the top-right corner the structures that are entirely sphere-shaped and the bottom corner structures that are completely disc-shaped.[168] The points inside the ternary plot are represented as hexagonal binning, which is a technique of data aggregation for grouping a dataset of N values into less than N discrete groups. Each hexagon might represent one single molecule or a grouping of overlapping molecules.

Python API CCDC

For the calculation of the SOAP descriptor and the construction of the PMI plots, the 3D structure of the molecules is needed. As the information provided from the ZINC database is two-dimensional, the simplified molecular-input line-entry system (SMILES) strings were converted to a 3D structure using the CSD Python API 2021.1 release. The tool is based on the CSD Conformer Generator which uses knowledge from more than 1 million experimentally derived structures to predict and generate appropriate conformers with bond lengths and angles based on known data. The generated 3D structures were relaxed, using the Spartan Software and PM6 semiempirical model, until the geometric minimum in found.

Theoretical model

PM6 semi-empirical method was applied through Spartan software for the orbital energy calculations. As PM6 is a low-cost computation with lower accuracy than DFT, it was benchmarked with a more accurate but time consuming B3LYP (B3LYP/3-21G*) model by identifying the linear relation between the PM6 and the B3LYP calculated HOMO, LUMO, $LUMO_{+1}$ orbital energies. The calibration curves are presented in Figure 3.3, showing a very strong correlation between the two models and thus allowing for a reliable estimate.

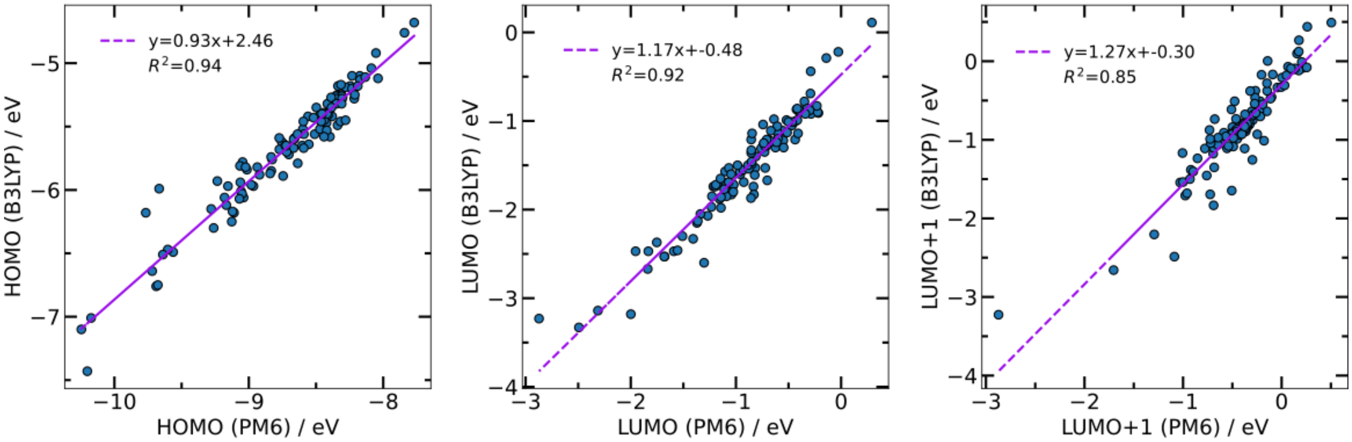


Figure 3.3. *Calibration curves of the lower (PM6) versus the higher level of theory (B3LYP/3-21G*) on the orbital energies of 210 PAHs. The comparison between semi-empirical and DFT theory is quite satisfactory, based on the squared correlation coefficient R^2 and the m and b values for the liner fit y=mx+b.*

3.3 Results

3.3.1 Measuring similarity between PAHs: Extracting structure-property relationships

A detailed analysis of the molecular and electronic structure of PAHs is very important with regards to their application as organic electronic materials. Similarity measures are broadly used in cheminformatics and drug discovery to help uncovering relationships between different instances. Simplistic predictive tools are based on the general theory that if two molecules are similar, they will likely show similar behaviour. Reflecting on this logic, we want to examine various ways to identify molecules that are similar to the ones with desired electronic functionalities.

Our starting PAHs dataset is composed of 210 molecules which are structurally similar, based on the Tanimoto similarity, to eight representative PAHs (See Methods 3.2) and C_{60} which is so far the PAH with the most interesting electronic properties that arise from its unique shape. To visualize the differences in shape between the PAHs, the PMI ratios (NPR1 and NPR2) were computed using RDKit. In Figure 3.4a these results are visualized, with the longer molecule being p-quaterphenyl, the most spherical being C_{60} and the most circular being coronene. The elemental composition of the molecules in the dataset is also shown in Figure 3.4b.

The functional molecule we are mainly focused on is fullerene (C_{60}) as it is currently the most studied among PAHs being both electronically and structurally interesting. C_{60} owes its electronic properties to its triple degenerate LUMO orbitals as well as to its non-planar circular shape. The question to answer is how we could identify molecules similar to C_{60} and which similarity metric could be more informative for enabling the more efficient categorization of PAHs. It is evident that there are several methods to measure similarity as there are several ways to describe the molecules. For the structural representation we are using well known molecular features which take into account 2-dimension or 3-dimensional characteristics, *i.e.*, the Morgan fingerprint, the molecular descriptors provided from Mordred library and the atomic environment descriptors using the SOAP method. The electronic description of PAHs is stated by measuring the band energy structures of each molecule (See Methods 3.2).

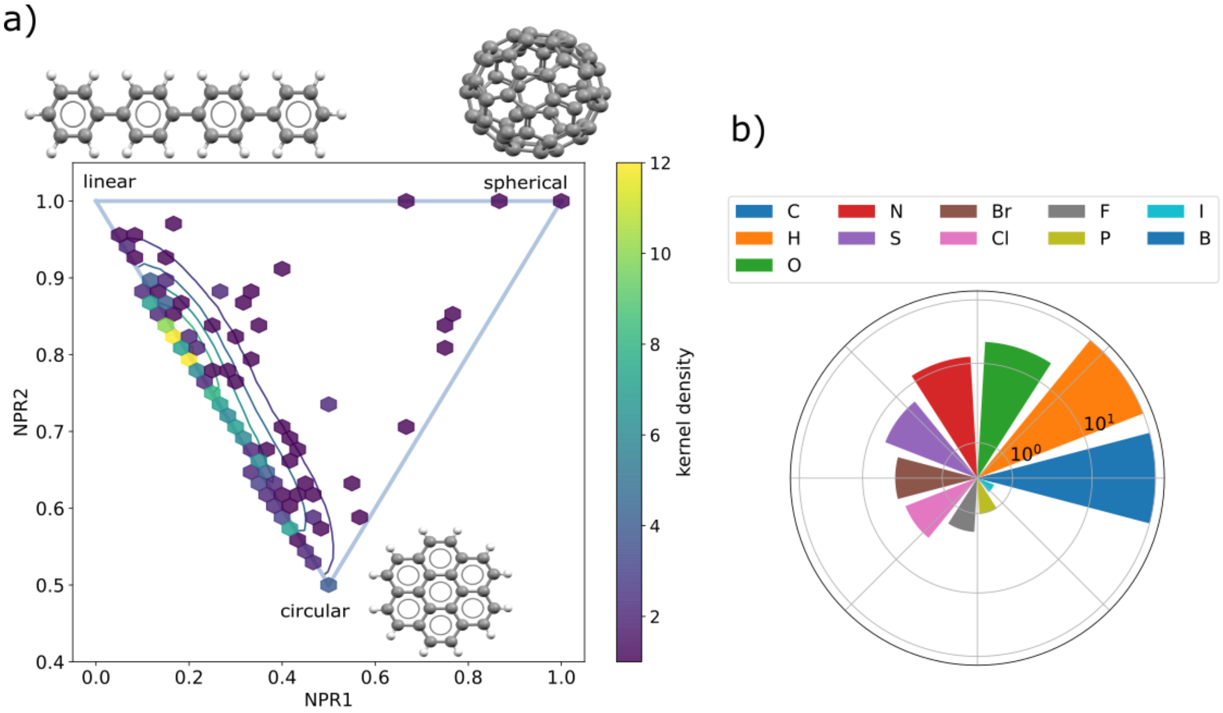
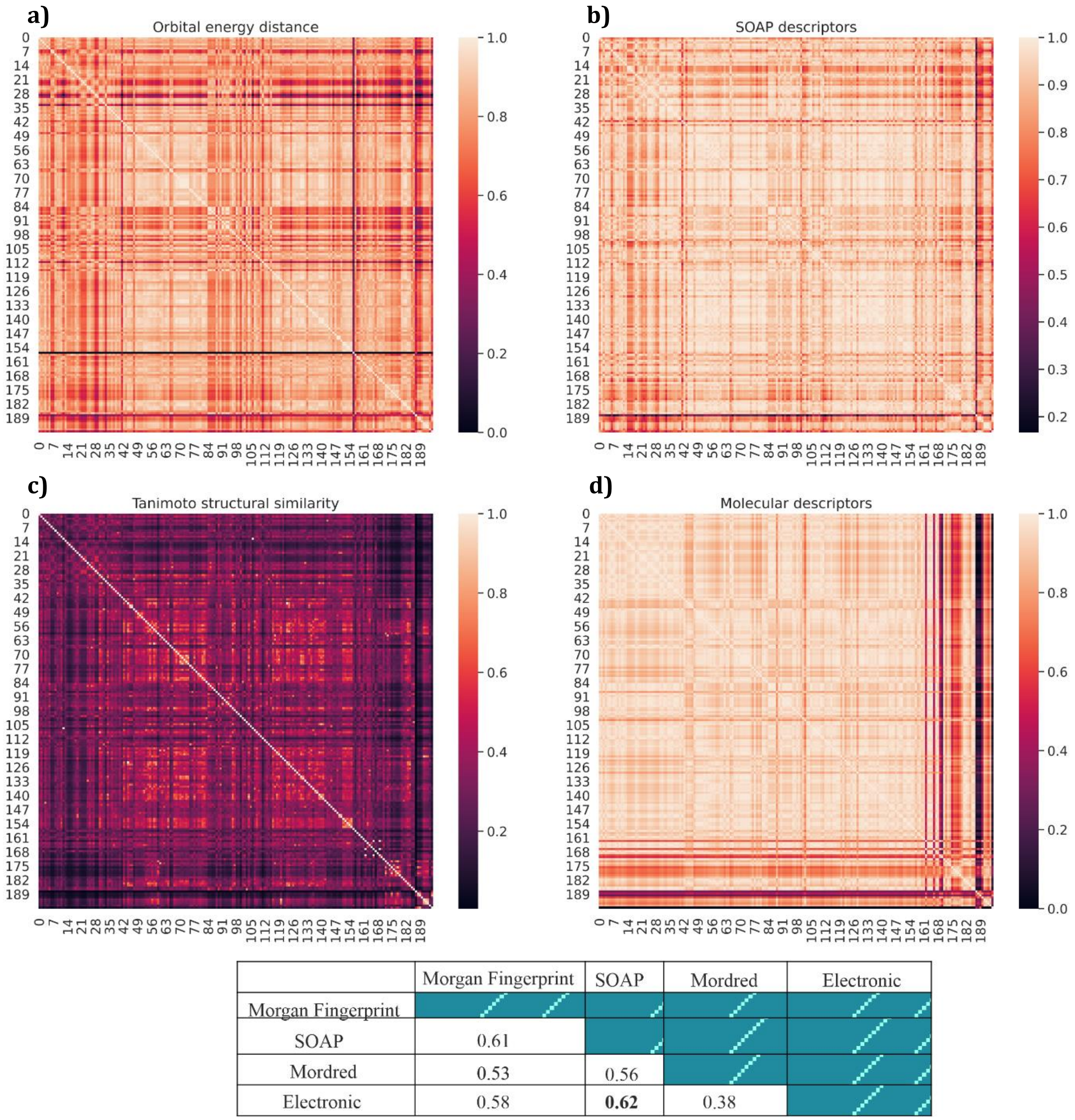


Figure 3.4. *A library of 210 polyaromatic hydrocarbons (PAHs). a) PMI plots for the PAHs dataset. C_{60}, coronene and p-quaterphenyl were found as the corner molecules. The contour is generated from kernel density estimator on the data with the colourbar indicating the density of the data, which lies between the linear and circular molecules. A hexagon is present if at least one structure belongs to that regime. b) Bar chart showing the composition analysis of the PAHs in our library: 100% contain C and H, whereas O, N and S atoms can be found in 20, 15, 8% respectively. The radial coordinates are on a logarithmic scale.*

Starting from the ZINC15 dataset of 210 PAHs, the molecular similarity is measured using different distance metrics according to each representation. For the comparison based on the Morgan fingerprint, the Tanimoto coefficient (T_c) was used, $T_c = \frac{c}{a+b-c}$ where *a* and *b are* the bit vectors of the two molecules under comparison and *c* the bits the two molecules have in common. For the comparisons based on the orbital energy distance and the molecular descriptors distance the Euclidean distance was used. The SOAP descriptors distance was calculated using the "regularized-entropy match" (REMatch) kernel.[105] The selection of the distance metrics is in accordance with the type of the representations and was inspired by similar works where researchers were analysing libraries of electronically active organic molecules.[34,103] These four distance matrices are presented in Figure 3.5 a,b,c&d.



	Morgan Fingerprint	SOAP	Mordred	Electronic
Morgan Fingerprint				
SOAP	0.61			
Mordred	0.53	0.56		
Electronic	0.58	**0.62**	0.38	

***Figure 3.5**. a) The electronic structure comparison of the PAHs based on the orbital energy vector (HOMO$_{-1}$ to LUMO$_{+4}$) b) Similarity based on the chemical topology using SOAP descriptor of the local environment (3D structure). c) Tanimoto similarity d) Molecular features similarity with Mordred descriptors. Distances are multiplied by -1 and scaled from 0 to 1 to be equivalent to similarity measures. e) Table showing the Pearson correlation between the matrices a,b,c,d. It is evident that there is a high positive correlation between the SOAP descriptor and the orbital energies descriptor. We can observe that molecules with similar 3D structures are expected to have similar electronic properties.*

According to the similarity techniques shown in the table in Figure 3.5e, there is an important correlation between the molecular representation and the orbital energy representation, derived after calculating the Spearman correlation coefficient between the four matrices. That could indicate that molecules with similar molecular structure are expected to have similar electronic properties and thus the structural categorization could provide a sensible choice for selecting electronically similar molecules. Although 2D molecular representation using the Morgan fingerprint, shows a considerable correlation with the orbital energies, the 3D method, namely SOAP descriptors, was found as more informative and correlated.

3.3.2 Categorizing PAHs with unsupervised machine learning

Clustering molecules based on structural similarity

Unsupervised learning was applied to identify the distinct clusters with structurally similar molecules in the dataset. The three different representations were tested *i.e.*, Morgan fingerprint paired with Tanimoto similarity, molecular features with Euclidean distance and SOAP descriptors with REMatch kernel, alongside with three different clustering techniques, *i.e.*, k-means clustering, affinity propagation and Gaussian Mixtures. The selection of the optimal number of clusters was performed using the elbow method. The clustering performance was evaluated considering the optimal cluster separation, using the Silhouette Coefficient[169] and the Davies-Bouldin Index[170] as described in the Appendix Table A1.1 & Figure A1.1.

SOAP descriptors with REMatch kernel have previously shown (Figure 3.5b) the best correlation with the orbital energies similarity matrix and also the best cluster separation in comparison to the other representations. As such this encoding is used to quantify the structural similarity between all PAHs in the dataset. The SOAP-based similarity matrix is projected onto a 2D space by a UMAP (Uniform Manifold Approximation and Projection) embedding used for dimensionality reduction, as shown in Figure 3.6a, where each point represents a molecule.

K-means clustering on the 2D UMAP coordinates was found to better separate the clusters and was used to categorize the PAHs datasets into five groups of structurally similar molecules. For extracting the electronic property trends in each molecular group, the average orbital energies of each cluster are plotted in Figure 3.6b. The averaging of the orbital energies was performed by initially applying a gaussian function to the orbital energies from HOMO-1 to LUMO+4 for each molecule. That results in a continuous function representing each molecule. The molecules that belong to the same cluster were then grouped together and their orbital energy functions were averaged.

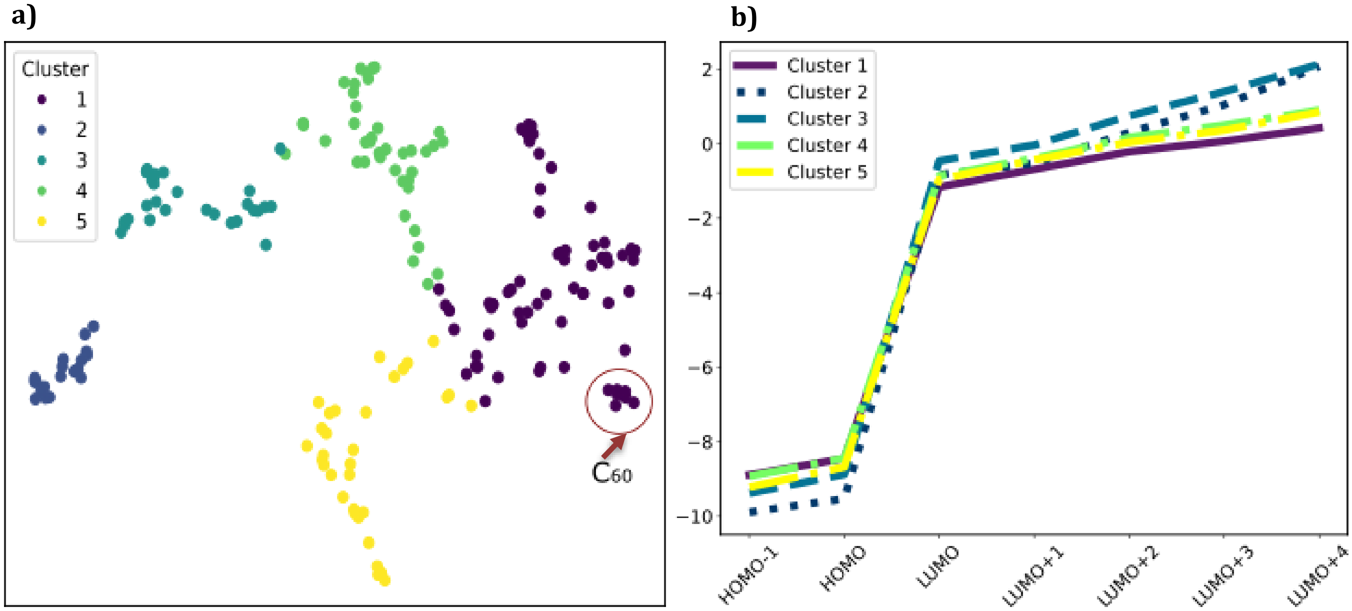


Figure 3.6. *Structure-property map of the PAHs library extracted from ZINC15. a) Unsupervised clustering of the starting PAHs based on their SOAP representation using k-means algorithm. b) Average value of orbital energies on each cluster.* C_{60} *belongs to Cluster 1, where a small HOMO-LUMO gap and the lowest LUMO orbital energies can be observed.*

That resulted in five distinctive orbital energy-based categories which can be analysed as follows: *i*) cluster 1, with small HOMO-LUMO gap and low LUMO energies. Moreover, the slope of the line connecting the LUMO orbitals is the smallest in comparison to the other clusters, thus indicating several molecules with LUMO orbital degeneracy, *ii*) cluster 2, with large HOMO-LUMO gap, and high LUMO energies, *iii*) cluster 3 with a small HOMO-LUMO gap and high HOMO energy, *iv*) cluster 4, with small HOMO-LUMO gap and high HOMO energies *v*) cluster 5, with small HOMO-LUMO gap high HOMO and high LUMO energies. It can be observed that the five different structural categories have also distinctive electronic features, proving that the clustering technique is also electronically meaningful. C_{60} belongs to the first cluster with the most obvious LUMO orbital degeneracy and small HOMO-LUMO gap. Some characteristic molecules that belong to cluster 1 are shown in Figure 3.7. Among them coronene and bezanthracene were considered in more detail in Chaper 4.

It can be seen that a wide range of structural characteristics were found to belong to the same cluster, *e.g.*, molecules with curvature achieved with pentagon integration, molecules crosslinked via an aliphatic bond, twisted molecules and circular molecules.

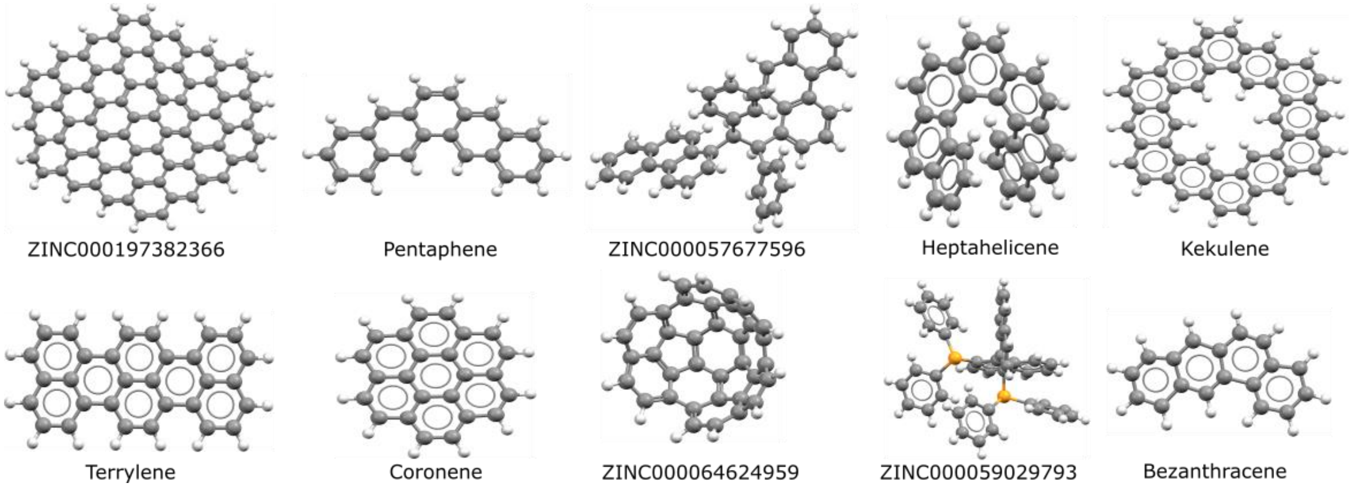


Figure 3.7. *Representative molecules from cluster 1 which were identified as structurally similar to C_{60}.*

Clustering molecules based on the electronic similarity

Hierarchical clustering based on Euclidean distance was further implemented as an alternative way to directly categorize the PAHs based on their orbital energies vectors. As before, a PAH was represented by the vector containing the orbital energies from $HOMO_{-1}$ to $LUMO_{+4}$. As shown in Figure 3.8 five main clusters were identified with C_{60} belonging to the golden one.

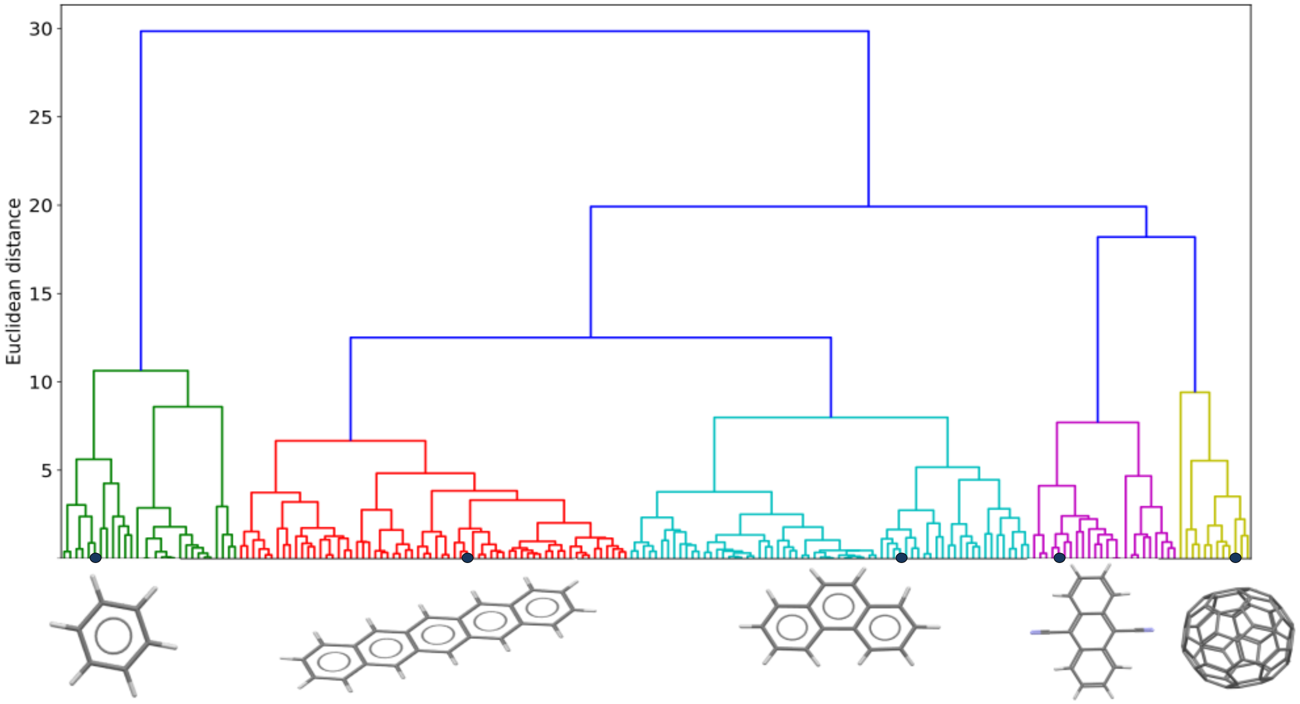


Figure 3.8. *Hierarchical clustering of the PAHs dataset based on the orbital energies vector representation. Some representative molecules of each class are shown on the x axis.*

The green cluster includes all the solvent-like polyaromatic molecules, *e.g.*, naphthalene, benzene and is the cluster with the largest distance to the C_{60}-containing cluster. Most of the molecules shown in Figure 3.7 fall inside the golden cluster confirming their electronic similarity alongside with their structural similarity. The way hierarchical clustering based on orbital energies works is by capturing the actual energy levels and not the degeneracy, *i.e.*, one molecule, such as benzene, might have degenerate LUMO, $LUMO_{+1}$ orbitals however the energy level is quite higher than the LUMO energy of C_{60}, consequently benzene is going to belong to a different cluster from C_{60}.

Overall, it can be concluded that there are several ways to measure similarity between polyaromatic hydrocarbons for identifying molecules similar to C_{60}. Similarity can be structural or electronic, but it mostly depends on the way we are going to represent the molecules. Unsupervised clustering is a useful technique for exploratory data analysis to get better insights from the dataset at hand and enables measuring similarity through different viewpoints. However, one of the major limitations of unsupervised clustering techniques is that a visual inspection of the identified clusters is essential for proving its usefulness.

In this work, the clustering was important for understanding the PAHs dataset, proving that there is a strong relationship between structure and electronic properties in PAHs and identifying molecular families with similar trends. Through our analysis, the uniqueness of C_{60} with its triply degenerate orbitals and spherical shape was proven. Further on, we are going to search for similar molecules in a larger dataset focusing mainly in the LUMO orbital degeneracy.

3.4 Scaling-up the screening on ZINC20

After the initial investigation on the smaller PAHs dataset, the search was expanded towards a wider range of PAHs. ZINC20[166] was screened using the new search functionality resulting in an extended dataset of 7,060 molecules (see Methods Section 3.2). The new dataset covers larger area on the PMI plot (Figure 3.9a) and also involves heteroatoms in higher percentage (Figure 3.9b). C_{60}, decacyclene and 4,4'-bis({[1,1'-biphenyl]-4-yl})-1,1'-biphenyl occupy the corners of the ternary plot indicating the most spherical, most circular and most linear molecules in the extended dataset. The highest density of the plot, as indicated in yellow colour, lies on the area with linear molecules, whereas the ‘spherical’ corner is occupied only by C_{60}.

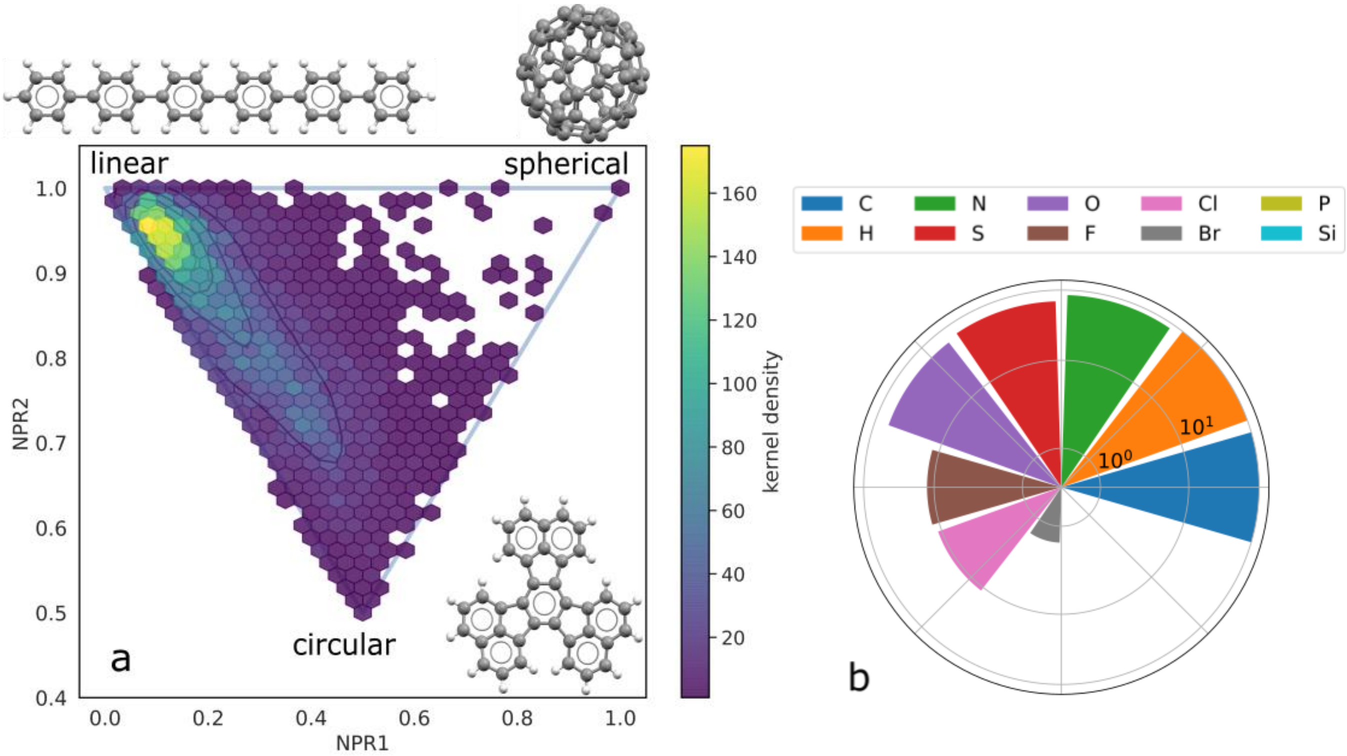


Figure 3.9. a) *PMI plot for the extended ZINC20 dataset showing high space coverage in terms of shape distributions.* C_{60}*, decacyclene and 4,4'-bis({[1,1'-biphenyl]-4-yl})-1,1'-biphenyl are located on the corners of the triangle.* ***b)*** *Barchart of the compositional analysis of the extended PAHs dataset showing that the considered PAHs are rich in heteroatoms. The radial coordinates are on a logarithmic scale.*

Further on, the same PMI plot is colour-coded based on the orbital energy differences across the PAHs space (Figure 3.10). These maps provide a visual representation of the spread in some important electronic properties within the selected molecular space. The three most important energy differences that were taken into consideration are *i*) the HOMO-LUMO gap, *ii*) the LUMO-LUMO$_{+1}$ degeneracy and *iii*) the LUMO$_{+1}$-LUMO$_{+2}$ degeneracy.

The HOMO-LUMO gap describes the energy difference between the highest occupied and lowest unoccupied molecular orbital and is an important property for designing organic semiconductors. The lower the HOMO-LUMO gap, the higher the chance to find an organic semiconductor are, as the electrons that will be excited from the HOMO orbital need less energy to reach and occupy the LUMO orbital. Most of the research related to the high-throughput identification of organic molecules with electronic interest is focusing on screening molecules based on the HOMO-LUMO gap.[171–173]

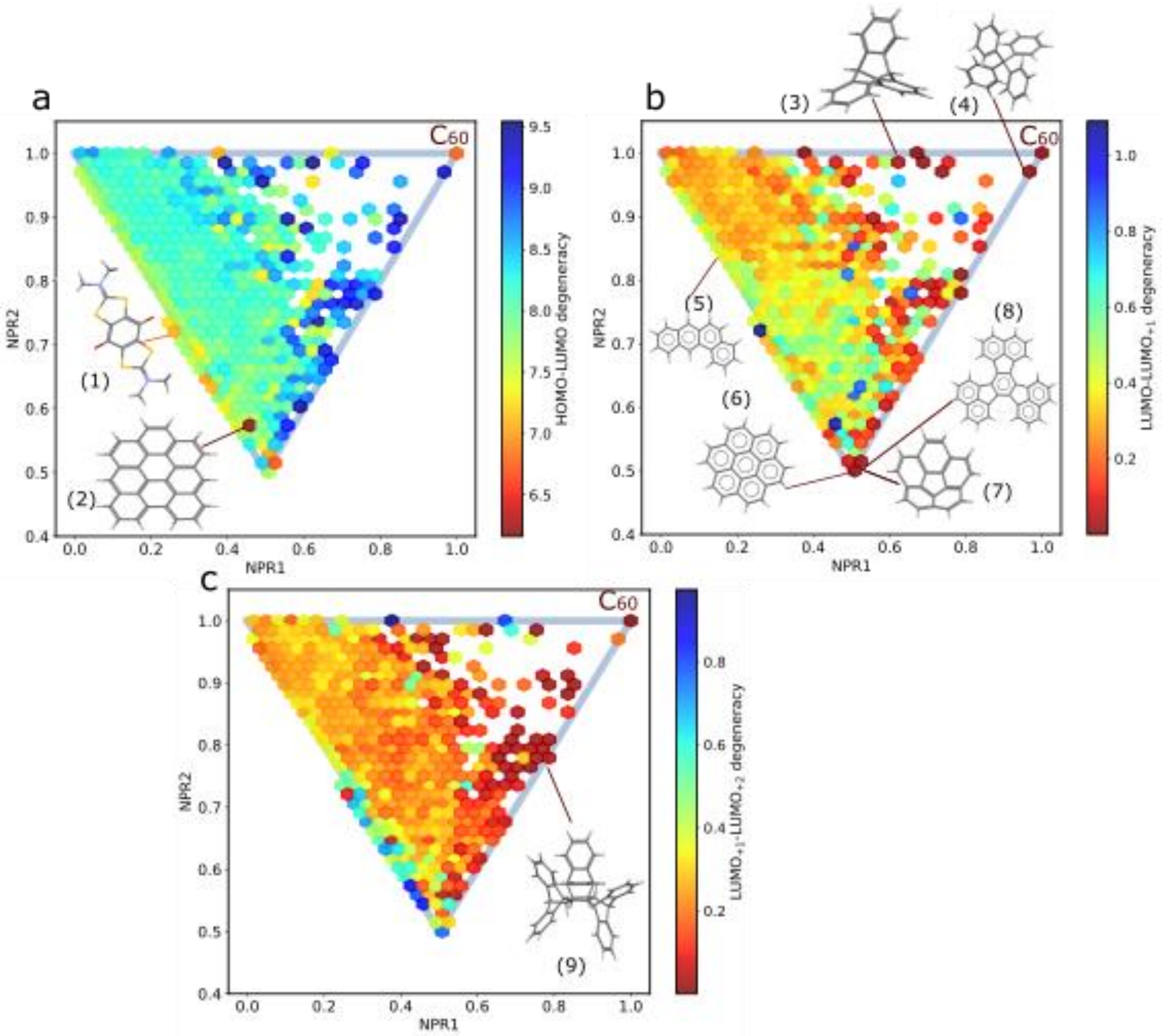


Figure 3.10. *PMI plots for the extended ZINC20 dataset showcasing the shape vs orbital energy relationship in PAHs. The plots are colour-coded based on the* ***a)*** *HOMO-LUMO gap,* ***b)*** *LUMO-LUMO$_{+1}$ and* ***c)*** *LUMO$_{+1}$-LUMO$_{+2}$ degeneracies. C_{60} is the only molecule with the lowest orbital energy differences in all three cases. It can also be seen that the near degeneracy in the LUMO orbitals is favoured for circular and spherical molecules.*

In Figure 3.10a, the distribution of the HOMO-LUMO gap across the PAHs reveals that the area with the lowest gap lies between the linear and circular structures. This area is dominated by sulphur (S)-containing PAHs, *e.g.*, molecule (1), which are known for their optoelectronic applications and have been reported to display photophysical and hole transport properties.[174,175] Interestingly, the PAH with the smallest bandgap in the dataset contains no heteroatoms and is dibenzo(bc,ef)coronene (2). Among the spherical molecules only C_{60} shows a small HOMO-LUMO gap.

The LUMO-LUMO$_{+1}$ degeneracy describes the energy difference between the first and the second lowest unoccupied molecular orbital and has been reported as an important measure for designing organic molecular magnets as the free electrons will occupy the orbitals in an open shell distribution. According to Figure 3.10b, the molecules that show exact degeneracy are mainly around the spherical and circular corners, *e.g.*, C_{60}, tetraphenyl methane (3), coronene (6), corannulene (7) and decacyclene (8).

The LUMO$_{+1}$-LUMO$_{+2}$ degeneracy describes the energy difference between the second and the third lowest unoccupied molecular orbital and can also play an important role in organic electronics. The molecules with that type of degeneracy lie to the area between the circular and spherical. Molecules can have LUMO-LUMO$_{+2}$ degeneracy without showing LUMO-LUMO$_{+1}$ degeneracy. C_{60} is one of the molecules showing exact degeneracy in these three orbitals (LUMO-LUMO$_{+1}$-LUMO$_{+2}$)

The uniqueness of C_{60} both in terms of structure and electronic characteristics can be seen in the PMI maps, as it is the only molecule with all three orbital energy differences being the lowest. Its icosahedral symmetry is causing the triple degenerate LUMO orbitals that play a key role in its exotic electronic properties, *i.e.*, high temperature superconductivity after metal insertion. As the orbital degeneracy in the LUMO orbitals plays a significant role for observing electronic properties after electron injection, we are focusing on identifying molecules that have this exact degeneracy in their LUMO orbitals as candidates for metal insertion in Chapter 4.

According to Figure 3.10, LUMO degeneracies are observed in molecules in the ‘spherical’ and ‘circular’ area. This finding indicates the role of molecular shape in the orbitals. Molecules in these areas are more symmetrical and thus they have near-orbital degeneracies. We can observe that molecular symmetry is highly related to orbital energies, as highly symmetric structures tend to have degenerate orbitals. It can also be seen that molecules with S4 symmetry, *e.g.* molecule (9) can have exact triple degeneracy similarly to C_{60}.

3.5 Future work

After observing the significant correlation between structure and property in PAHs, it is evident that the large and diverse PAHs dataset could further be used for developing machine learning models that are able to predict orbital energies by learning this correlation. Developing such a model could enable the high throughput screening of large databases (*e.g.,* ChEMBL, CSD), where given the molecular structure, molecules with the desired orbital energies or orbital energy differences can be directly identified. The different 2D and 3D molecular representations could become the input to the model and several deep learning or conventional machine learning techniques could be tested for their ability to predict the desired electronic properties, *e.g.*, orbital degeneracies. By creating models which can effectively predict targeted properties we can accelerate materials discovery by quickly identifying the most interesting materials to be experimentally tested.

3.6 Conclusion

The main target of this introductory chapter is to construct the PAHs datasets to be further used in search of novel materials for electronic applications. The two main datasets derived from this work are:

i) A molecular dataset with 210 PAHs which is a starting point for the evaluation of structure-property relations and is further used on Chapter 5 for extracting molecular pairs that could form co-crystals. This dataset was analysed in terms of structural and electronic characteristics and unsupervised learning clustering enabled the categorization of the molecules into classes based on their similarities.

ii) An extended molecular dataset of more than 7,000 PAHs which is used for selecting the most promising molecules for metal intercalation. Although the majority of research for identifying electronically interesting (conducting) PAHs is focusing on the search for a small HOMO-LUMO gap, in this work, the main aim is to identify PAHs among the extended dataset with degeneracy on the LUMO, LUMO$_{+1}$, LUMO$_{+2}$ orbitals. We provide the full dataset of PAHs with the orbital energies as calculated from the semi-empirical model (PM6).

Overall, we tried to understand the properties of PAHs on the molecular level and categorize them in terms of similarity to C_{60} employing unsupervised machine learning techniques. It was found that structurally similar molecules are also electronically similar in terms of their orbital energies, which makes the use of machine learning models for property prediction a powerful tool for exploring these materials. We can also conclude that for different applications we need to focus more on different properties, *e.g.,* for a semiconductor HOMO-LUMO should be small but for the intercalation chemistry the LUMO degeneracy is more important. The PMI plots provided a sensible way to visualize the distribution of the structural characteristics across a large dataset of molecules. Of course, there are many more structural categorizations that can be made, *e.g.*, twisting, curvature. However, these can also be regarded as subparts in the PMI plots. On another note, it should be reported that searching the ZINC database for available molecules can be sometimes misleading as although there seems to be a vendor for a molecule considered as ‘purchasable’ in reality the molecule might not be available or the vendor to be disconnected. As a result, for selecting the most interesting molecules for the metal insertion part we considered both LUMO degeneracy and real availability. The selected molecules are shown in the Appendix, Table A1.2.

Herein, we investigated properties of PAHs on the molecular level. However, it is evident that the a priori design of functional molecular organic crystals with desirable properties is one of the most challenging cases, since they rarely obey simple geometric principles, like framework-based materials *e.g.*, zeolites and MOFs, which could be exploited for rational design. Even very small changes to the molecular structure might result in considerable effects on crystal packing and hence on the resultant solid-state properties. Molecular crystal packing is often dictated by weak, competing intermolecular interactions. For that reason, crystal structure prediction in Chapter 4 is further used as an exploratory tool for assessing the different possible positions the molecules might have in the crystal structure. A mechanistic understanding of the of the molecular function can only be facilitated by understanding their structure.

4. Strategies for identifying PAHs systems as hosts for metal intercalation

4.1 Introduction

Carbon-based materials of different structural topologies showing exotic electronic properties have attracted significant attention. In this work we are investigating the possible metallic and superconducting properties of a certain category of carbon-rich materials, the alkali-doped aromatic compounds.

The tantalizing perspective of designing flexible, large area, low-cost electronic materials made from abundant and simple components, such as polyaromatic hydrocarbons and metals, has sparked considerable research interest in that field. It has been shown that the alkali metal doping of organic molecules causes the activation of their electrical conductivity. Alkali metal atoms have a ‘noble gas-like’ ionic core surrounded by one loosely bound valence electron. When the alkali metal atoms lose the outer electron, they get a stabler noble gas-like configuration, whilst the aromatic molecule is activated by incorporating an extra electron to the LUMO orbitals.

Although there are several metal-PAH structures reported in literature for having extraordinary electronic properties, including high temperature superconductivity and quantum magnetism,[83,176] the nature of these properties still remains mysterious. For instance, several studies report superconductivity in a number of potassium- and rubidium-intercalated materials.[177] However, the incomplete structural characterization of these materials hinders the understanding of the underlying chemistry and physics of these systems. Nevertheless, the electrical conductivity of most doped organic structures is still orders of magnitude lower than the best inorganic conductors.

Although significant progress has been made regarding the theoretical and experimental investigation of these materials a systematic way to predict basic elements such as stoichiometry, crystal structure and electronic bands is not yet established.[178] I hope that the presented methodology could afford as a starting point for the determination of viable synthetic pathways.

4.1.1 Key points of this chapter

a. Investigation of metal-π landscapes in the Cambridge Structural Database. The extracted geometric parameters are further used in CSP calculations as initial constraints to aid in the generation of more sensible structures.

b. Analysis of the existing fully structurally characterized pure metal-PAHs structures. The findings of this analysis are driving towards the development of a strategy for identifying the next most promising candidates.

c. Evaluation of the new proposed candidates in terms of void space, orbital degeneracy, metal capacity and energetic stability.

d. Crystal structure identification of selected metal-PAHs systems. Furthermore, electronic structure analysis is performed for the energetically favourable structures.

4.2 Motivation

It is well known that the electrical conductivity (σ) is directly proportional to the carrier concentration (*n*), the charge of the carrier (q) and the charge-carrier mobility (μ), following equation:

$\sigma = nq\mu$ (4.1)

This relation indicates that n and μ should be increased to achieve larger conductivity and electrical current. For increasing *n*, charge carriers should be effectively injected, whereas μ is mostly related to the configuration of the molecules. In this regard, in the making of an open shell conducting material, the alkali metals play the role of the charge carrier (*n*) and the molecular packing of the final system controls the μ parameter.[179]

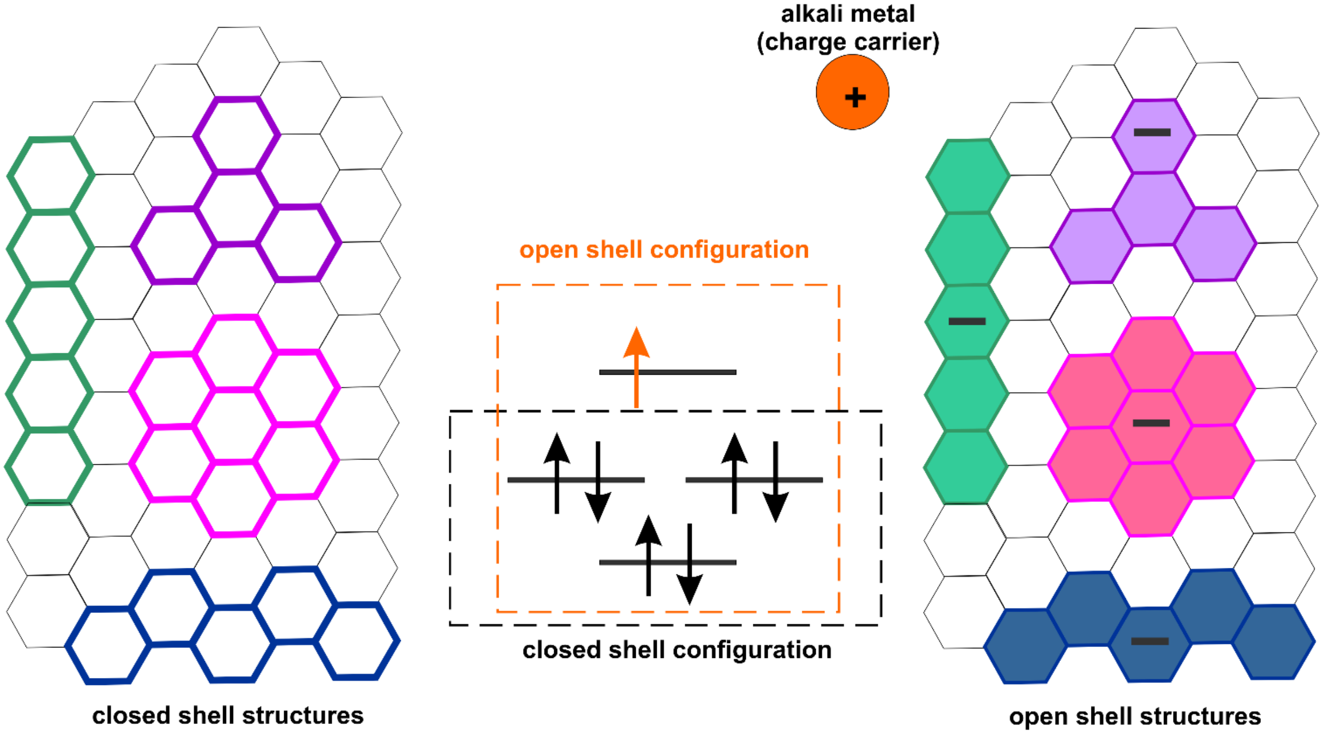


Figure 4.1. *PAHs are regarded as closed-shell structures with no electron mobility around them, whereas the metals serve as reducing agents and thus have the ability to transfer electrons to the system and drive to the formation of open-shell molecular units.*

Reported examples where an alkali metal acting as a charge carrier gave rise to extraordinary electronic properties in PAHs are C_{60}, picene and triphenylene, to name just a few.[85] However, C_{60} remains the only PAH superconductor that is fully characterised with single crystal diffraction. The C_{60} alkali-doped structures and the reported superconducting temperatures (T_c) involve K_3C_{60} at 18°C[77], Rb_3C_{60} at 29K[180], $RbCs_2C_{60}$ at 33K[181] and Cs_3C_{60} at 38K.[182]

In fullerene systems, the alkali metal cations occupy the interstitial voids that already exist in the host structure leaving the crystal unchanged. As the empty voids are adjacent to the electron densest area in the structure, consequently the conjugated π-electron system enables the strong interaction among the inserted cation and the fullerene molecules. On the other hand, in PAHs systems the metal insertion affects the arrangement of PAHs in the crystal structure. Consequently, the derived electronic propertied are highly dependent on the crystal packing the PAHs will afford after the metal insertion. Regarding some of the known metal-PAHs systems, pristine picene, pentacene, phenanthracene and tetracene crystals all adopt the herringbone packing motif with edge-to-face (σ−π) interactions dominating over any potential π−π interactions. The largest voids are located between the molecular layers, adjacent to the saturated C-H bonds and far from the electron density of the PAH π systems (Figure 4.2a). The rearrangement of the packing after the metal insertion is performed in a way that the K sites are now closer to the aromatic π systems and strengthens the C-H…..π contacts whilst a single void per molecule is created (Figure 4.2b).[183] However, in these structures the interaction among the cation and the PAH is still weak.

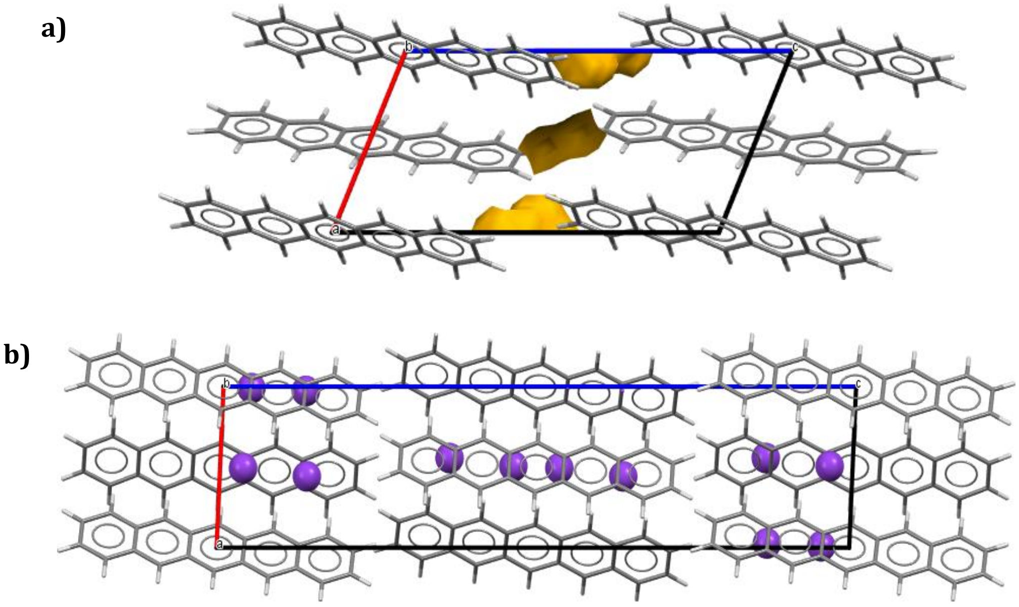


Figure 4.2. *Voids modification after metal insertion in pentacene, viewing along b axis. a) Pristine pentacene crystal structure, containing 2 molecules/unit cell with lattice parameters **a** 7.90, **b** 6.06, **c** 16.01, **a** 101.90 **β** 112.60 **γ** 85.80, displaying the available void spaces in yellow. b) Modified pentacene structure after potassium insertion (purple balls), containing 4 molecules/unit cell with lattice parameters **a** 7.20, **b** 7.22, **c** 30.44, **α** 90, **β** 92.66 **γ** 90. Herein, it can be observed a significant lattice parameters modification.*

The rubrene case reveals a different scenario, in which the dominating interactions are π−π intermolecular interactions between neighbouring tetracene cores of parallel rubrene molecules (Figure 4.3a). It was observed that the insertion of the K metal completely disrupts the intermolecular interactions of pristine rubrene. In this regard, two large voids per molecule are created where the accommodated K^+ cations strongly interact with both the tetracene core and the phenyl groups of rubrene (Figure 4.3b).

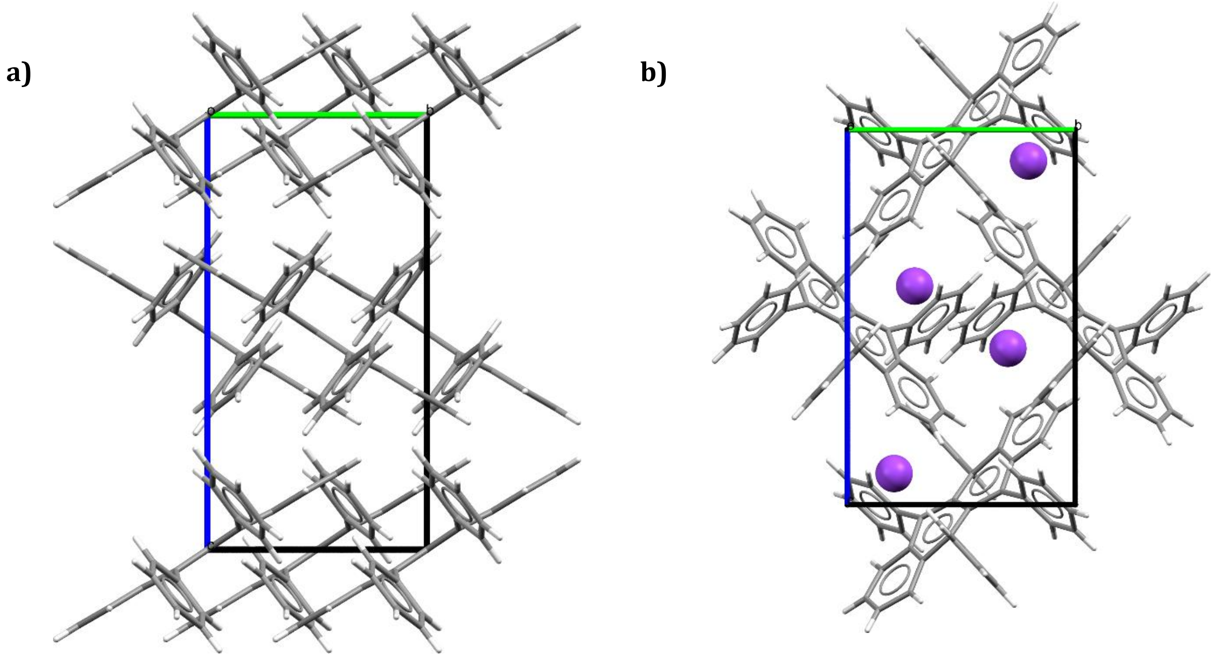


Figure 4.3. *Structure modification after metal insertion in rubrene, viewing along a axis. a) Pristine rubrene crystal structure, containing 8 molecules/unit cell with lattice parameters* ***a*** *26.86,* ***b*** *7.19,* ***c*** *14.43,* ***β*** *90. The main interactions are π-π intermolecular interactions highlighted by red dotted lines However, σ-π intermolecular interactions are also present (blue dotted line). b) Modified rubrene structure after potassium insertion (purple balls), containing 2 molecules/unit cell with lattice parameters* ***a*** *12.85,* ***b*** *8.36,* ***c*** *14.53,* ***β*** *71.26.*

4.3. Methods

4.3.1 Isostar library

The Isostar library provided by CSD is a computerized library containing crystallographic and theoretical (*ab inito*) data on intermolecular nonbonded interactions[184]. Herein, it was used for extracting valuable insights about chemical groupings between polyaromatic rings and metals aiming to display visually the most possible space where the metals can be found. The screening for the detection of the existing metal-PAHs systems was done using

Conquest software. The unsaturated hydrocarbons were defined as a benzene fused to aromatic ring (central group), which is designed as a benzene ring connected with two aromatic bond which can indicate a polyaromatic structure. In that way, it is ensured that all the polyaromatics are included to the search. The metals in contact (contact groups) that were selected are all the alkali metals. The enquiry for the fragments search is shown below (Figure 4.4).

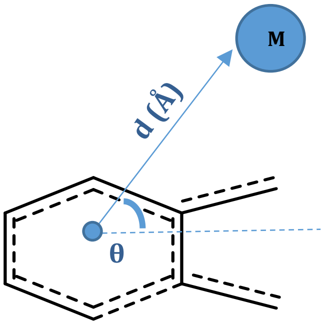


Figure 4.4: *Metal-π bond configuration used in the CSD Conquest, representing the benzene fused to aromatic ring and the alkali metals(M).*

The distance d is set between 0 to 5 Å without constraints to the angles. The geometric parameters that were taken into consideration are the distance between the metal atom/ion and the centroid of the aromatic ring as well as the angle formed. Scatterplots with the distribution of metals around the aromatic fused ring were generated using the Isogen library and are presented in the Appendix Figure B1.1.

4.3.2 Zeo++ software

The void space analysis was performed using the well-established Zeo++ software[140] widely used for analysis of porous materials such as zeolites and metal organic frameworks (MOFs). The arrangement of polyaromatic hydrocarbons in certain crystal structures is generating void spaces that are appropriate for metal insertion. As the structural characterization is a key part of computer-aided design of porous materials, the investigation of the geometrical parameters describing pores is essential and will enable a better prioritization of candidates.

The void space analysis using Zeo++ is based on the Voronoi decomposition. For a given arrangement of atoms in a periodic domain it provides a graph representation of the void space and in that way the atomic connectivity is determined. In more detail, for each atom in the lattice the Voronoi cell is constructed around that atom. Consequently, the material space is divided into irregular polyhedral cells which are analysed to determine the pore topology (Figure 4.5a). The resulting Voronoi network is analysed to obtain the diameter of the largest included sphere and the largest free sphere, which are two geometrical parameters that are frequently used to describe pore geometry (Figure 4.5b).

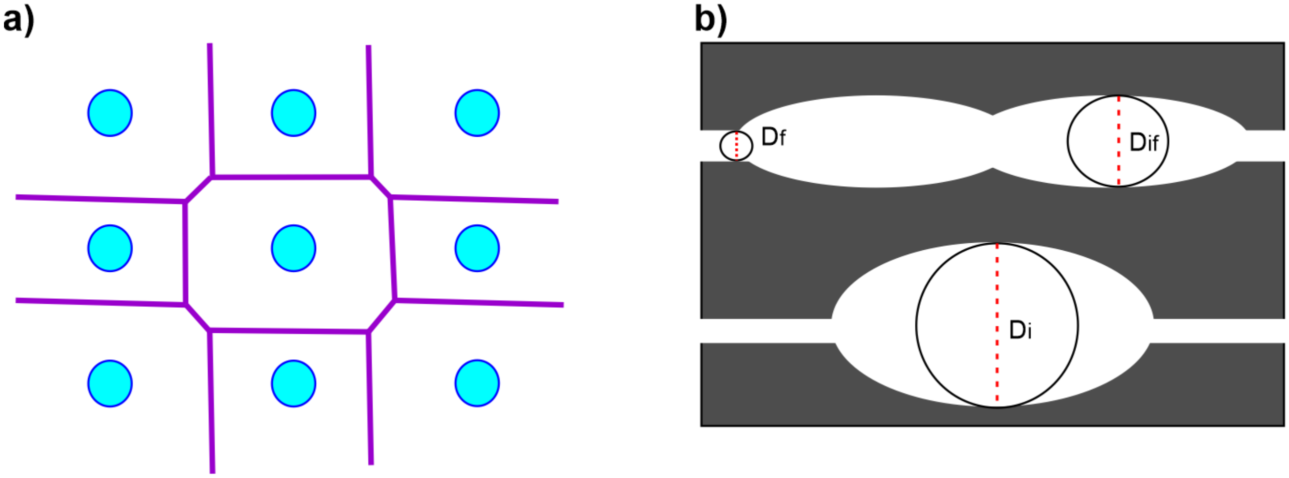


Figure 4.5. a) 2D Voronoi diagram of nine atoms. b) The typical parameters describing pore sizes are the diameters describing: (1) the largest included sphere (Di), (2) the largest free sphere (Df), and (3) the largest included sphere along the free sphere path (Dif).

4.3.3 DFT calculations for modelling the known metal-PAHs systems

Dispersion inclusive density functional theory was implemented to understand the electronic structure of the experimentally known materials, evaluate the energetic stability and investigate the effect of metal intercalation to them. Plane-wave-based DFT calculations were performed using the VASP (version 5.4.1) programme. The optB86b-vdW functional was used to improve the description of van der Waals interactions over other semilocal DFT functionals,[185] with a plane-wave cutoff energy of 520 eV. Core electrons were treated using the projector augmented-wave method. Unit-cell parameters and atomic positions were relaxed until all the forces were reduced to below 10^{-3} eV Å^{-1}. Calculations on all the metal-PAHs systems were performed using the unit cells which correspond to the refined crystal structures, setting k-space parameter to 0.2. For the calculations of the potassium crystal the body-centered cubic (bcc) structure with Im-3m space group was extracted from ICSD and the same DFT parameters were applied. The k-point strings used to plot band structures were generated using the AFLOW framework.[186] The benchmarking of the selected VASP parameters is described in the Appendix Table B1.1, Table B1.2, Figure B1.2.

4.3.4 Convex hull construction

The convex hull construction is a methodology used to identify relative stability of structures with different stoichiometries. In the absence of kinetic effects, a convex hull construction can be used to identify structures and compounds that are stable with respect to decomposition into two or more parent structures at fixed thermodynamic conditions. In this work we make use of a compositional based convex hull, which could be interpreted as the following: If two structures A and B with compositions C(A) and C(B) and free energies G(A) and G(B) are part of the hull, then any structure C with composition C (A) < C (C) < C (B) and a free energy G(C) that lies above the line joining A and B on the hull will spontaneously decompose at constant volume into a mixture of A and B.[187] Herein, the structures A and B correspond to the polyaromatic hydrocarbon and the metal atom. The steps followed to construct the convex hull for the metal-PAHs systems are the following: *i*) starting from the known structure for a single PAH, the empty voids are identified and ranked by their size from the highest to the lowest, *ii*) inserting metals in the voids after selecting the desired metal content in the structure, *e.g.*, $1 \leq x \leq 4$ in most of the cases, *iii*) structural relaxation until full convergence with both atomic positions and cell parameters (a, b, c, α, β, γ) being free to change to minimize the forces on atoms and stresses on the unit cell, and *iv*) the construction of the convex hull plot showing on the x axis the metal concentration in the structure as K/(K+molecule) and on the y axis the energy difference between the intercalated structure and their constituent single-component structures. The void space detection and theoretical metal insertion (steps *i* and *ii*) were performed using an in-house script employing the Pymatgen library (version 4.7.5) and the incorporated Zeo++ library.[139,188] Step *iii* was performed using the same DFT model as for modelling the known metal-PAHs systems (section 4.3.3). Moreover, spin polarized calculations were also performed for some theoretically intercalated systems. These calculations are the basis of theoretical determination of spin magnetic moments. In addition, they can be used to understand the basic mechanisms which might lead to the occurrence of magnetism in solid state materials. The approach described above will be referred in the text as 'simple intercalation approach' as only the parent crystal structure of the single molecules was used for metal insertion and no crystal structure was generated from scratch. The term 'simple intercalation approach' is used for discriminating this approach from the crystal structure predictions (section 4.3.5) where new structures were generated given only the molecular diagram and no information about the crystal structure.

4.3.5 Crystal structure prediction

When the most electronically and structurally promising combinations of molecules have been identified, their directed assembly into crystalline materials with targeted properties is the next significant step. Various successful computational methods have been developed for Crystal Structure Prediction (CSP) of molecular materials. According to these methods, a set of thermodynamically feasible crystal structures can be generated and a better insight into the range of polymorphs can be provided. The most widely used approach for predicting organic crystal structures is to minimize the lattice energy of a large number of systematically or randomly generated candidate structures. In this approach, the experimentally observed structure is assumed to correspond to the most stable packing arrangement.

In this work, USPEX 10.4 software for crystal structure prediction was implemented as it is currently the only CSP software that can afford modelling both organic and inorganic systems.[138] USPEX code employs an evolutionary algorithm (EA) which is an iterative and stochastic search method inspired by Darwinian evolution. Herein, each crystal structure is described by the atom coordinates and the lattice vectors (a, b, c, α, β, γ). During the first generation a set of structures is randomly generated which could optionally satisfy some constraints, *e.g.*, distance of molecular centres. The quality of each member of the population is assessed by a so-called fitness function which in our case is the energy value after the *ab initio* calculation. The best members (structures with the lowest energy) are selected as the parent structures and the next generation is created by applying specially designed variation operators, *i.e.*, heredity, mutation or randomly. The process continues until some stopping criteria are achieved. In our cases the stopping criteria are either reaching 1,000 generated structures or if the same best structure remains unchanged for eight consecutive generations.

The CSP study was carried out using an evolutionary algorithm, as implemented in the USPEX code, in conjunction with *ab initio* structure relaxations based on density functional theory (DFT) within the dispersion inclusive Perdew-Burke-Ernzerhof (PBE+D3) generalized gradient approximation as implemented in VASP (Vienna ab initio simulation package). It should be noted that even modest changes in the orientation and deformation of the PAHs molecules and in the relative position and distance of alkali-metal atoms with respect to the PAHs and among themselves can alter the bonding nature of the doped system and its electronic structure. Moreover, even by knowing the metal composition in the structure, several competing and coexisting crystalline structures with the same chemical composition but different symmetries and physicochemical properties might be possible. To explore the variety of low energy structural polymorphs an extensive search over the energy landscape of each compound is essential, according to the following steps:

1. First, we need to set the number of molecules and metals per unit cell, *e.g.*, 4 molecules + 8 K for 1:2 stoichiometry. To cover all the possible ratios and directly compare with the simple intercalation approach, we selected four stoichiometries 1:1, 1:2, 1:3, 1:4. For both the examined systems, *i.e.*, $K_x$bezanthracene and $K_x$coronene, two aromatic molecules and two, four, six and eight alkali atoms were used respectively to achieve the desirable stoichiometry. We also assumed integrity of molecules, excluding all possibilities of structural decomposition or polymerization.
2. Calculations were performed using DFT supplemented by van der Waals (vdW) corrections by using the PBE+D3 method at zero pressure and temperature. The relaxation is performed in four stages, starting from stage 1 and 2 with low and medium precision respectively. In these steps very crude structure relaxations of both atomic positions and cell parameters keeping the volume fixed are performed. In stages 3 and 4 a full relaxation with medium and normal precision respectively takes place with unit-cell parameters and atomic positions being relaxed.
3. The energy cutoff for the plane-wave basis was set to 520 eV to ensure full convergence, and zero-point energies were not included. The Brillouin zone was sampled by Monkhorst-Pack meshes with the resolution of $2\pi\times0.05A^{-1}$.
4. For each system, 4000 structures were explored within 20 generations for each stoichiometry.

5. The final most stable CSP generated structures of each stoichiometry were recalculated at the same accuracy as the simple intercalation approach for direct comparisons. The USPEX workflow was also benchmarked to a known system, namely K_2tetracene (csd id: MURLIX) in which the CSP started given the known composition (4 tetracene molecules and 8K). The ability of the software to generate sensible structures as well as the structure-energy maps are discussed in the Appendix B3.

4.3.6 Electronic structure calculations

To understand the origin of the electronic properties of the metal-PAHs systems, we performed density functional theory (DFT) calculations with the VASP code using the optimized lattice parameters and atomic positions as derived from the lowest energy predicted structures. The position of the Fermi level in relation to the band energy levels is a crucial factor in determining electrical properties. In highly conducting materials, *e.g.,* K_3C_{60}, the Fermi level intersects the LUMO band, whereas LUMO and LUMO$_{+1}$ bands are connected (no band gap between them).[176]

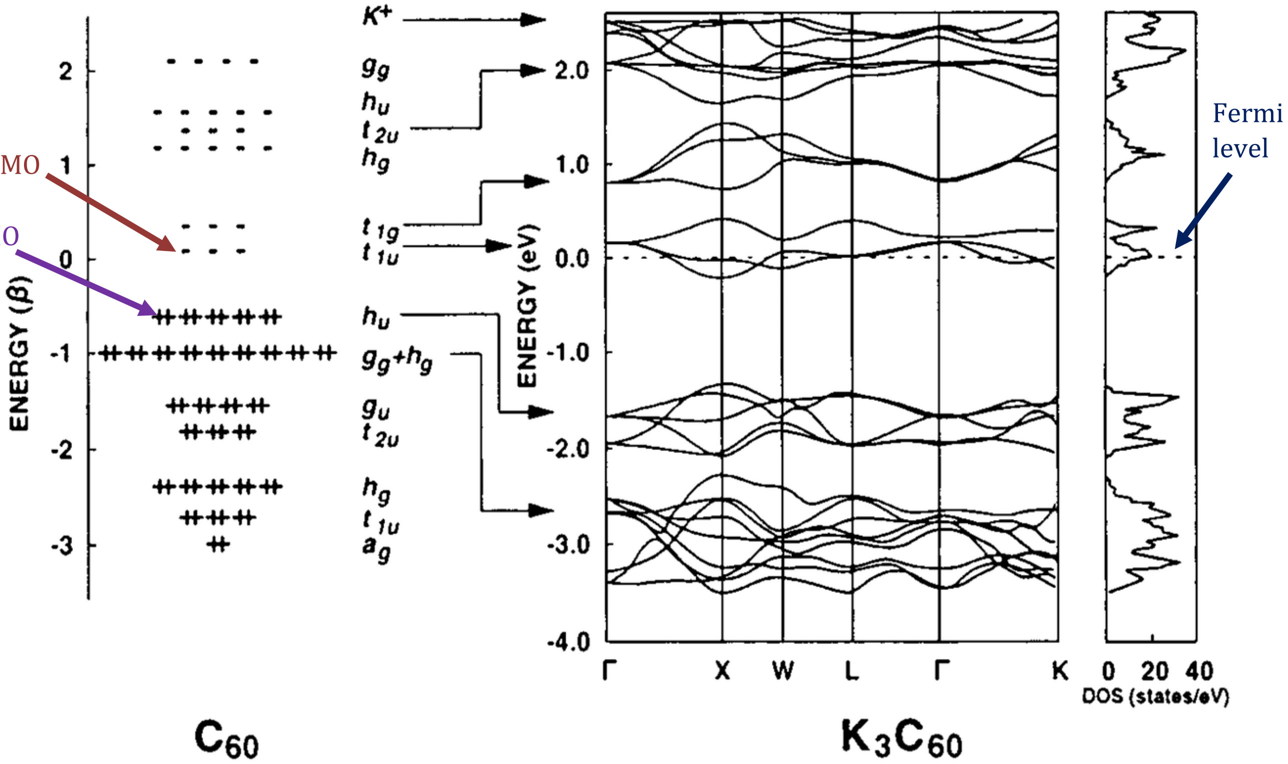


Figure 4.6. *The electronic structure of C_{60} and the band structure of the metallic K_3C_{60}. Reproduced from ref. 191 (https://journals.aps.org/rmp/pdf/10.1103/RevModPhys.68.855). Copyright (1996) by the American Physical Society.*

The electronic structure of the undoped insulating C_{60} and the band structure of the metallic K_3C_{60} is presented in Figure 4.6. In the undoped C_{60}, the HOMO band is filled and the triply degenerate LUMO band is empty, hence

C_{60} is an insulator. On the other hand, in K_3C_{60}, the alkali metal is donating its electron to the LUMO orbitals. Hence K_3C_{60} is expected to be a metal with a half-filled LUMO band.[189]

4.4. Results

4.4.1 Existing Metal-π bond landscapes in the Cambridge Structural Database

Using Conquest and the Isostar library with the parameters defined in the Methods 4.3.1 section, the search for metal-PAHs contacts revealed 142 structures and 1,817 molecular fragments. Potassium (K), lithium (Li) and sodium (Na) were found to be the dominant metals in these systems as shown in Table 4.1.

Table 4.1. *Distribution of the geometric parameters in the metal-PAHs systems, based on the alkali metals. The main geometric parameters are the distance between the metal atom/ion and the centroid of the aromatic and the angle formed, as shown in Figure 4.4.*

Metal	**Radius**	**Total Number of structures**	**Number of fragments**	**Distance (Å)**		**Angle(θ◦)**	
				Range	Median	Range	Median
Li	1.28	26	497	1.92-4.99	3.67	16.31-161.91	88.45
Na	1.66	32	95	2.74-4.95	4.36	37.47-165.71	90.02
K	2.03	49	683	2.64-4.99	3.84	33.51-151.09	90.16
Rb	2.2	11	203	2.78-4.98	3.89	36.99-148.08	84.67
Cs	2.44	24	339	3.02-4.94	3.81	33.91-146.18	84.57

From the total number of fragment found in the CSD, it is evident that there are many interactions among a fused polyaromatic ring and metals. However, very limited structures contain only metal and a polyaromatic hydrocarbon in their structure. These pure crystal structures are summarized in Table 4.2. In the majority of the cases examined with Isostar, a solvent was present in the structure which can be prohibitive for electron correlation effects.

Nevertheless, for the extracted structures that contain a metal-π interaction, it was observed that as the radius of the metal gets bigger they tend to be more concentrated around 90^o (perpendicular to the center of the aromatic ring). Moreover, our statistical analyis indicate that lithium, which has the lower radius, is the only alkali metal that can be found closer to the aromatic ring, with distances even lower than 2 Å (Figure 4.7).

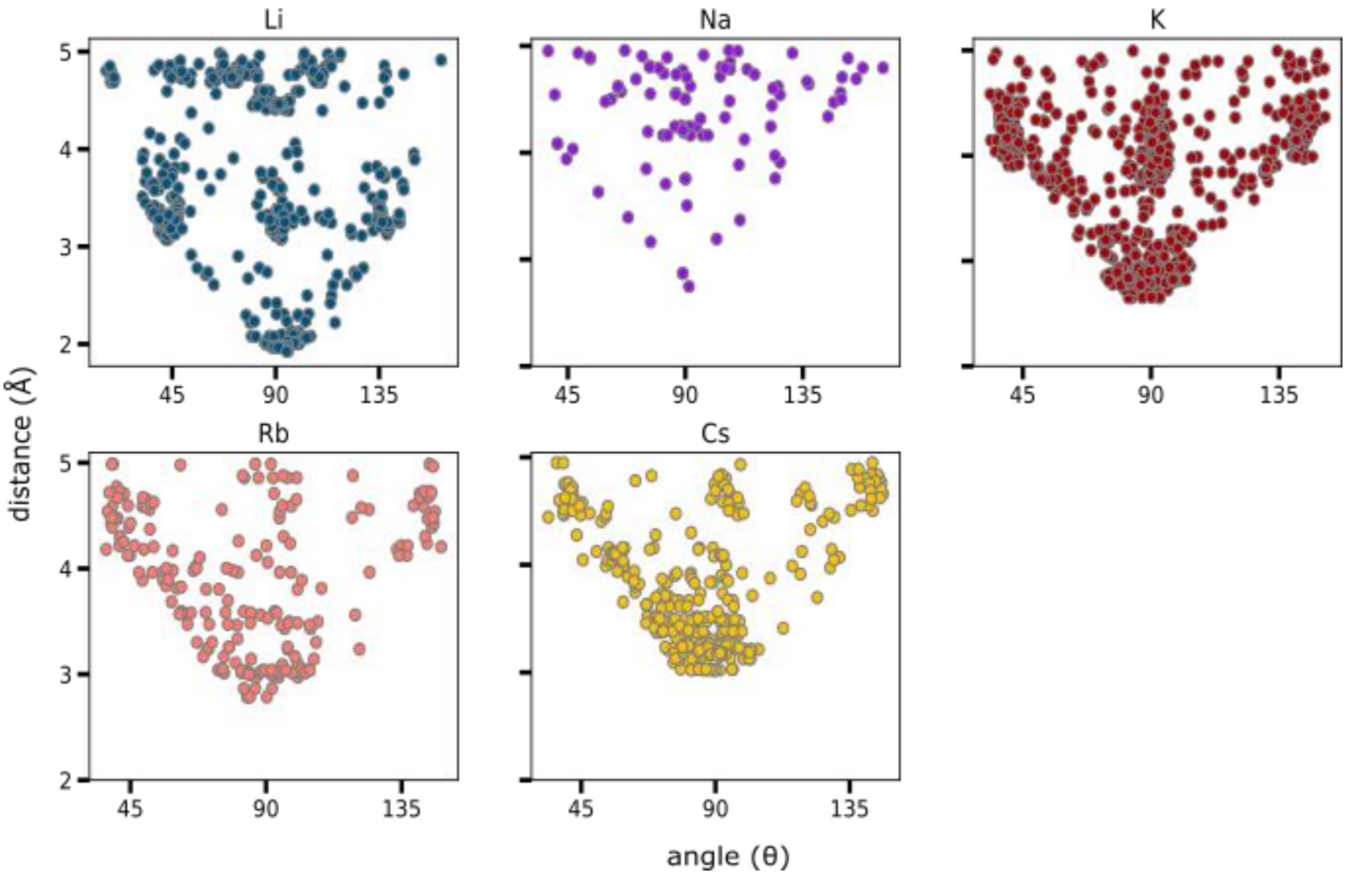


Figure 4.7: *Scatterplots showing the distribution of the geometric parameters according to the radius of each metal. The alkali metals are sorted based on increasing radius, starting from Li with the smallest to Cs with the highest.*

4.4.2 Analysing the existing Metal-PAHs systems

From the statistical analysis of CSD, it can be concluded that the majority of currently existing metal-PAHs landscapes refer to coordination compounds or systems which include a solvent in the structure. All the metal-PAH systems that were fully structurally characterized are summarized in Table 4.2. The energetic stability of the doped structures is measured by the energy difference between the final energy of the doped form and the sum of the energies of the un-doped molecular crystals and the crystal of the single metal, following the equation:

$$\text{Energy difference (eV)} = \text{energy of KxPAH} - (\text{energy of PAH} + \text{energy of single K crystal}) \qquad (4.1)$$

Table 4.2 shows the lattice energy calculations on the existing pure phase and fully characterized metal-PAHs systems. The calculations were performed used the optB86b-vdW functional to include dispersion corrections. The final energy difference is given in eV per Functional Unit (FU), *i.e.*, the number of PAHs in the structure.

Table 4.2. *Evaluation of energetic stability in metal-PAHs systems using Wan der Walls corrections (optB86b functional).*

Known doped structures in CSD	Energy of PAH without K (eV)	Energy of single K (eV)	Energy of K$_x$PAH (eV)	Energy difference (eV)	Energy difference (eV/FU)	Density (g/cm^3)	Ref
K2tetracene (MURLIX)	-741.386 4 molecules	9.289 8 K atoms	-742.086	-9.989	-2.50	1.497	190
PIWHUB (K2rubrene)	-861.11 2 molecules	4.6447 4 K atoms	-861.603	-5.138	-2.57	1.370	191
K2pentacene (YASTOE)	-896.483 4 molecules	9.289 8 K atoms	-898.725	-11.531	- 2.88	1.824	183
K2picene (YASTUK)	-898.729 4 molecules	9.289 8 K atoms	-894.262	-4.822	-1.20	1.604	183
Csphenanthrene	-1173.627 8 molecules	-2.42 8 atoms	-1182.724	-7.885	-0.99	1.964	59
Cs2phenanthracene	-586.813 4 molecules	-2.42 8 atoms	-594.829	-5.595	-1.40	2.490	59

Our findings indicate that in all cases the formation energy of the intercalated crystals is significantly lower than the energy of the constituent molecules by more than -1 eV/FU. It can also be observed that all structures have high density above 1.3 g/cm^3. Regarding the volume expansion (see Figure 4.8) after the metal doping we can conclude the following: *i*) potassium (K^+) insertion to pentacene to afford K_2pentacene resulted in an expansion of the original unit cell by 15% per pentacene moiety *ii*) potassium (K^+) insertion to picene to afford K_2picene led to an expansion by 11.4% of the original unit cell *iii*) caesium (Cs^+) intercalation to phenanthrene led to a ~16% or a ~27% volume expansion for Csphenathracene and Cs_2phenathracene, respectively[59] *iv*) potassium (K^+) insertion to rubrene to afford K_2rubrene resulted in 6.4% unit cell expansion.

In Figure 4.8, we investigate the modification of the voids after metal insertion. We report the void parameter of both the single molecule structures and the intercalated structures (after theoretically removing the metals). It can be seen that in all cases the metal insertion drove to a significant void expansion which is in accordance with the unit cell expansion reported in literature for the above-mentioned cases. The structures are represented as circles and are colour-coded based on the LUMO orbital degeneracy. The only two fully structurally characterized materials found with metallic character are C_{60} and phenanthrene, which are also the two molecules with the smallest

LUMO, $LUMO_{+1}$ orbital energy difference. Picene, which also shows significant LUMO degeneracy, has been reported as superconductor in the K_3picene ratio. However, its structure has not been verified.

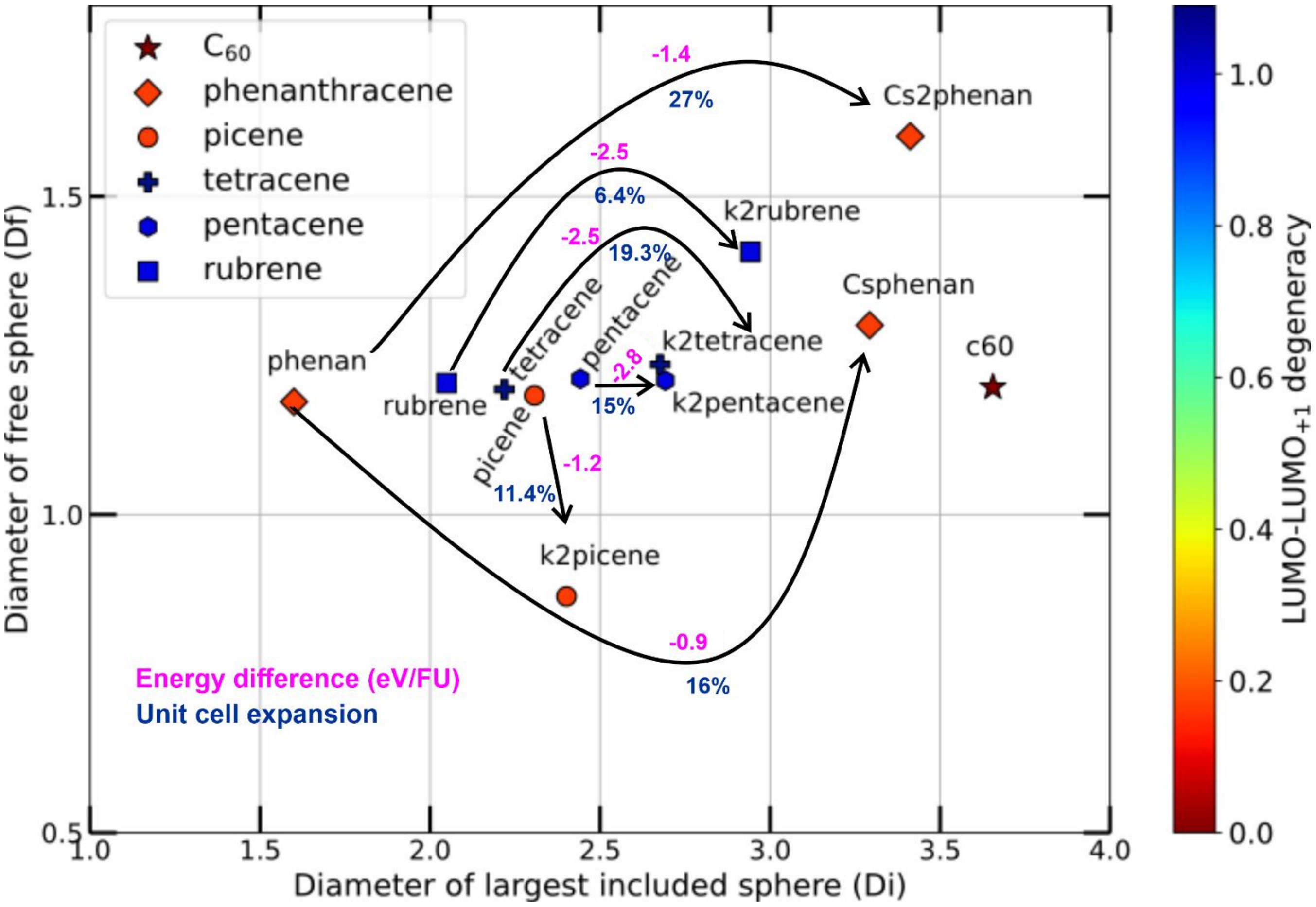


Figure 4.8. *Mapping the voids modification and energetic stability after metal insertion in well-studied PAHs. Both the undoped and doped structures are represented, with the curly arrows indicating the changes in energy and volume. The structures are colour-coded based on the LUMO degeneracy (in eV) of the PAH.*

After observing the trends on the existing doped PAHs systems in Figure 4.8, we can conclude that*: i)* PAHs that have been intercalated have diameter of largest included sphere Di > 2 Å *ii)* The metal insertion leads to an energetically favourable conformation in all the studied cases *iii)* Superconductivity was only observed in C_{60} – Metal systems which afford exact triple degeneracy, *iv*) Metal insertion leads to significant unit cell expansion in all the cases, except the C_{60} systems *v*) Finally, from several experimental studies and reported observations arises that decomposition of the structures is a possible phenomenon.[191,192]

4.4.3 Identifying possible PAHs as metal hosts

For the selection of new molecules as substrates for metal insertion, we are focusing only on aromatic molecules, based on the assumption that the presence of a π-conjugated system is necessary for electron mobility. For the computational screening of the PAHs, we are considering three major parameters: *i)* void space *ii)* orbital degeneracy *iii)* metal capacity. Based on the investigation of these key points of interest, we are going to propose good candidates for experimental consideration using as a fourth criterion the availability in the lab/purchasability.

The PAHs dataset, consisting of 210 molecules, used as a starting point for the analysis is previously reported in Chapter 3 and was extracted from the ZINC15 database after searching for Tanimoto similarity to the eight representative PAHs (Chapter 3, Methods). The Cambridge structural database was searched using Conquest tool for identifying 84 molecules out of the 210 having a reported crystal structure. The void diameters of these structures are reported in the Appendix, Figure B2.2.

Considering the effect of orbital degeneracy and void space

C_{60} is the only molecule with reported intercalated structure in which unarguably high T_c superconductivity was observed experimentally. The indicative triple degeneracy in the LUMO orbital of C_{60} as was discussed in Chapter 3 is regarded as the top important characteristic for observing electron correlation. Moreover, several theoretical studies have identified that the near degeneracy between the LUMO, $LUMO_{+1}$ half-filled conduction bands close to the Fermi level might be a driving element of phonon-driven superconductivity.

Following the analysis and comparison of orbital degeneracies, we selected available molecules with exact double degeneracy for further analysis. In this regard, if the original double LUMO degeneracy is retained, then a single electron transferred from the alkali metal would lead to a 1/4 filled band and the PAH should behave as a metal. However, in some cases it is experimentally observed that PAHs lose the 2-fold LUMO degeneracy and therefore are easily driven to the Mott insulating state.

As described in the methods section, Zeo++ was used for the calculations of the void space in the identified PAHs crystals. Although high porosity might be desirable to increase the pore surface and let the structure afford higher metal concentration, it might also reduce the carrier mobility. We need both degeneracy and high void space such that the degeneracy will remain low after the metal insertion. The six most promising structures in addition to C_{60} in terms of large void space, degeneracy and availability are shown in Figure 4.9.

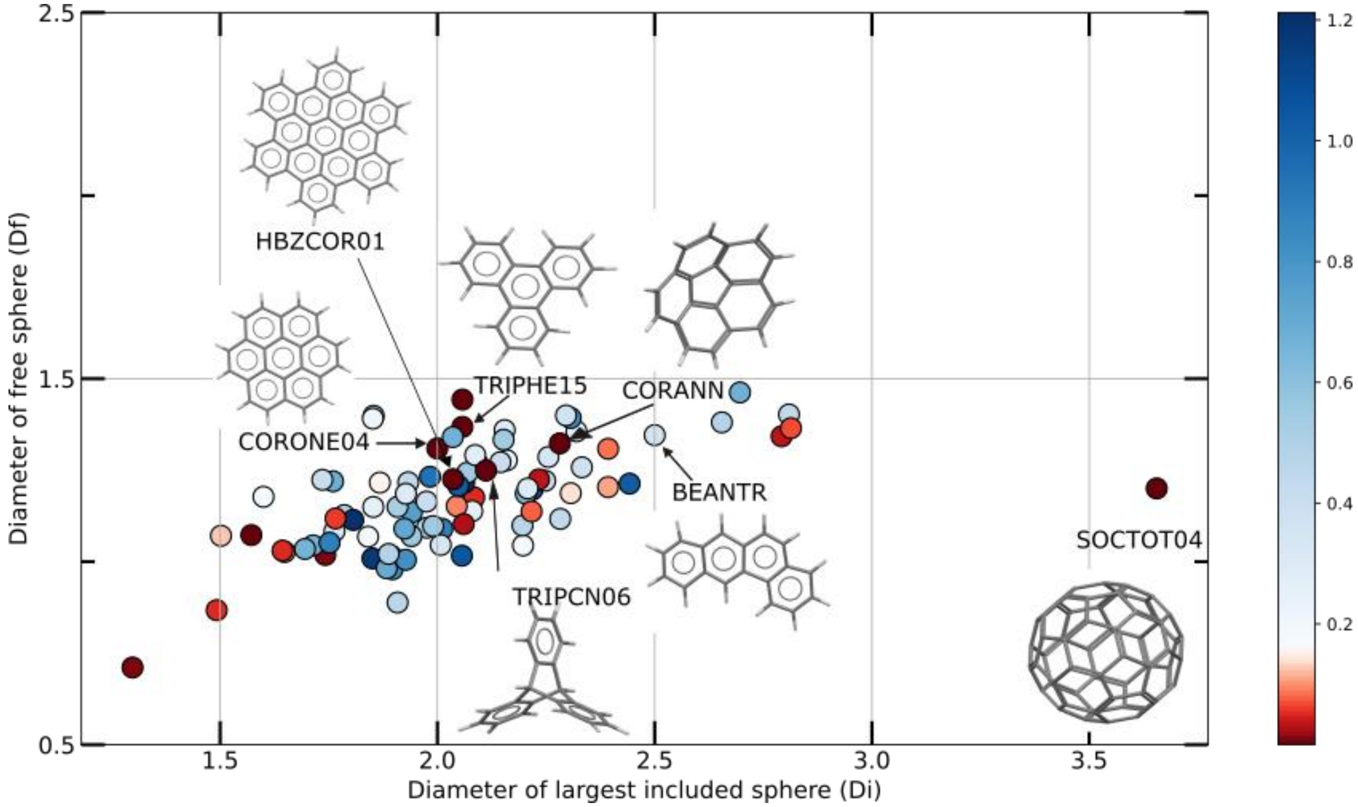


Figure 4.9. *Pore size parameters and LUMO degeneracy of the ZINC single molecules with known structures. Each point represents a crystal structure and is colour-coded based to the LUMO- $LUMO_{+1}$ orbital energy difference (eV). Bezanthracene (BEANTR) is shown as a promising candidate for metal insertion in terms of available void space. Coronene (CORONE),triphenylene (TRIPHE15) and corannulene (CORANN) are promising in terms of exact LUMO orbital degeneracy.*

Considering the effect of metal capacity

It has been observed that the metal ratio plays an important role for the electronic properties of the materials. For instance, in the picene case, K_3picene is reported as a high Tc superconductor, whereas the K_2picene was an insulator. For the investigation of metal capacity in the detected PAHs systems, the convex hull was constructed for each structure as described in Methods. Similar analysis was performed for all the PAHs highlighted in Figure 4.9 and the overall convex hull is shown below. The known potassium intercalated structures are represented as stars.

It can be observed that all the known potassium intercalated structures, represented as stars, can afford 1:2 stoichiometry.

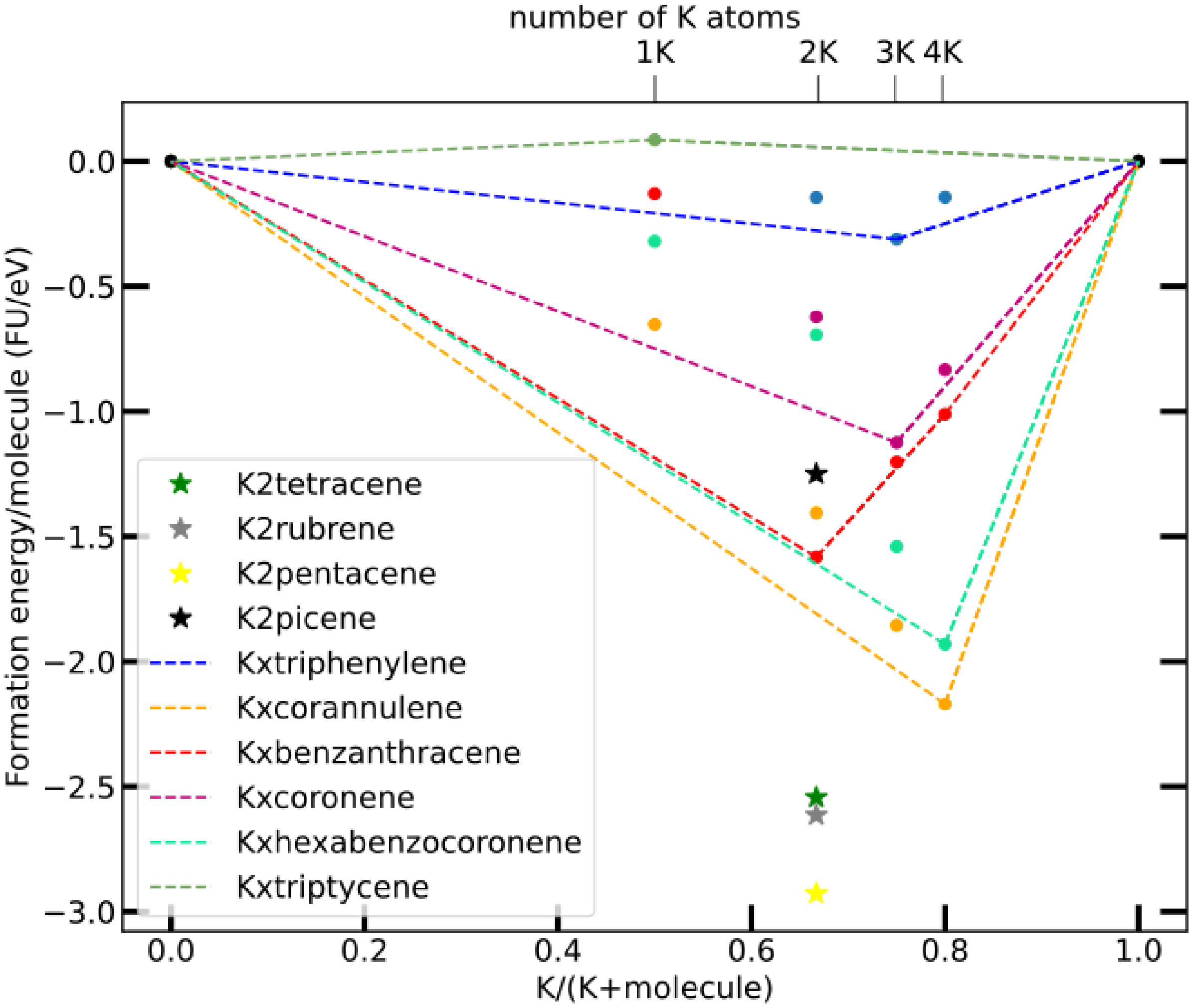


Figure 4.10. *Convex hull summary of the selected PAHs showing the most stable ratios for the different intercalated structures. Structures with low energy, promising metal ratio (open shell ratio), high metal capacity are prioritized as synthetic targets.*

Spin-polarized calculations for the K1 and K3 ratios were further used for the promising systems for describing the magnetism of itinerant electrons. Detailed tables with all the *ab initio* calculations are given in the Appendix B2.

Considering the effect of metal type

Another important factor to be taken into consideration regarding metal-PAH systems is the type of metal. potassium, sodium and caesium are the three metals usually found in analogous systems and thus their effect was also investigated for two systems, *i.e.*, metal-Coronene and metal-Triphenylene. These systems were selected as both have exact LUMO degeneracy and an open cell ratio for potassium x=3.

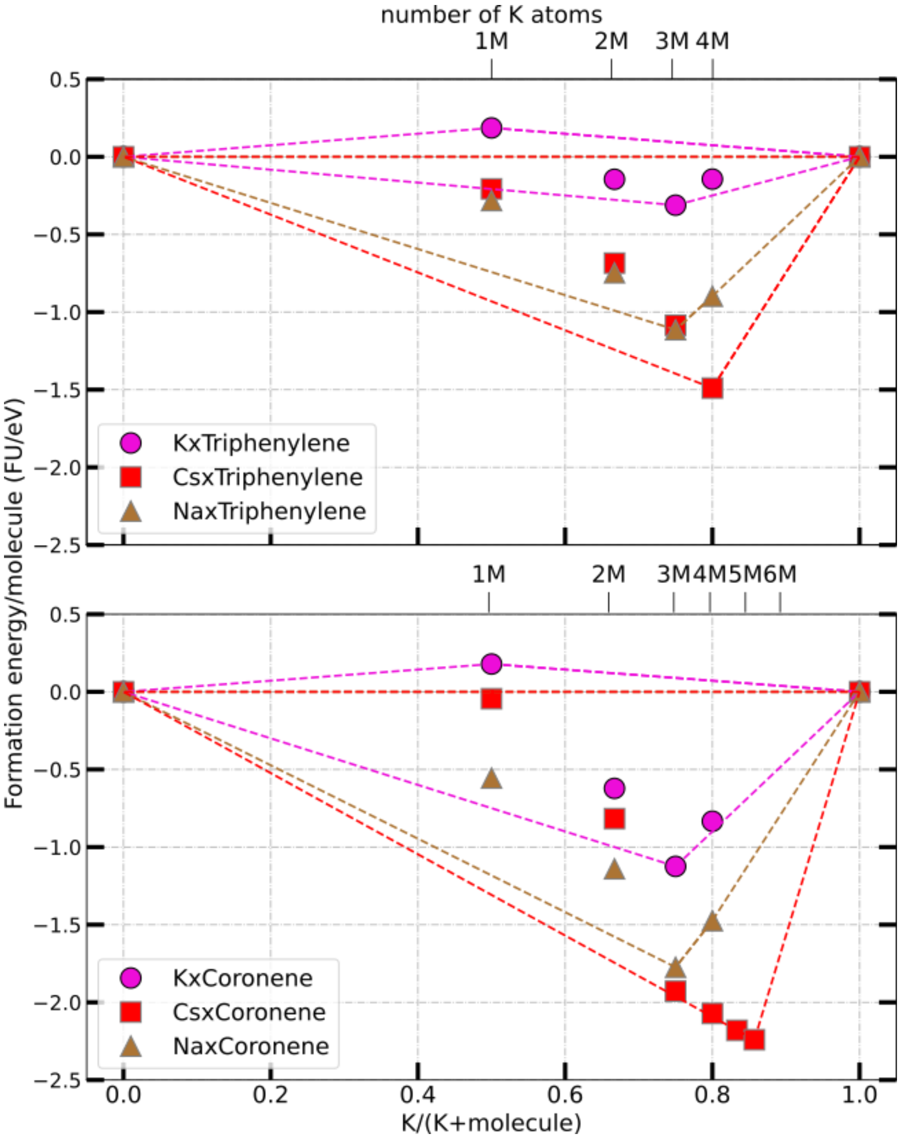


Figure 4.11. *Convex hull comparisons based on the metal type, i.e., potassium (K), Caesium (Cs) and sodium (Na). (top) triphenylene (bottom) coronene case.*

Our findings indicate that sodium has similar behaviour as potassium for both structures, accommodating the same 1:3 stoichiometry. Cs, as a more aggressive reducing agent than potassium and sodium, seems to be able to afford higher ratios and more stable structures. It was also found that Cs is causing significant deformation to the starting crystal structure in both coronene and triphenylene as opposing to K and Na (Appendix Figure B2.3, Table B2.14). It is evident that as Cs is a larger alkali metal, when inserted in structures of small size PAHs where direct intermolecular forces are weak, the addition of high metal quantities is a major perturbation that impacts the pristine herringbone structure. Based on the spin polarized calculations for the metal-coronene open shell ratio (x=3) (Appendix tables B2.2, B2.3, B2.4), the ferromagnetic (FM) K_3coronene was found more stable in comparison to its non-spin polarized solution indicating a magnetic character.

Selecting the candidates for further investigation

Alongside the *ab initio* method for the convex hull determination, a well-established Crystal Structure Prediction (CSP) software coupled with DFT theory was used to identify the lowest energy structure for each metal ratio. CSP is a key to computational materials discovery as the properties of a material depend sensitively on its structure. Herein, we use CSP in search of evidence of metallic and superconducting phases for the promising metal-PAH systems identified in the previous chapters. Due to the time demanding nature of CSP, we performed these calculations only for two systems, namely K_xBezanthracene and K_xCoronene. The selection of these two candidate systems was based on large void space (K_xBezanthracene), exact LUMO orbital degeneracy (K_xCoronene) and immediate availability in our lab (both). Both molecules were also classified in the same cluster as C60 in the analysis performed in Chapter 3 (Figure 3.6). Bezanthracene is a molecule with high structural similarity to tetracene and phenanthracene, which both have been successfully intercalated with potassium and caesium respectively. Coronene is a very well-studied material for metal intercalation. Although it has been reported several times as a superconductor, an experimental structure has not yet been determined nor a detailed first principles study has been yet reported to the best of our knowledge.

4.4.4 The atomic structure of potassium-doped bezantrhacene from a first-principles study

Benzanthracene (BEANTR) is the first structure for further computational investigation as it has a large enough void space for metal insertion and was also grouped together with C_{60} in the analysis performed in Chapter 3, Figure 3.6. Moreover, it shows interesting structural properties as it is the first molecule that could be regarded as an intersection of an acene and phenacene, *i.e* includes both linear fusion of and zig-zag benzene rings respectively. That could be of importance as the zig-zag molecules are more stable, whilst the linear fusion rings absorb more energy. Thus, benzanthracene as an intersection could have both stability and electronic properties. Bezanthracene is reported as an insulator with band gap of 2.02 eV.

In the following paragraphs, we will present the structures, their energies, and their electronic structures followed by a conclusive discussion. The structures generated with the simple intercalation approach, *i.e.*, theoretically intercalating metals in the largest free void spaces, are named as K_xPAH_simple, with x=1,2,3,4. The structures generated with USPEX are indicated as K_xPAH_(EAy) with with x=1,2,3,4 and y being the id number of the generated structure as defined by USPEX.

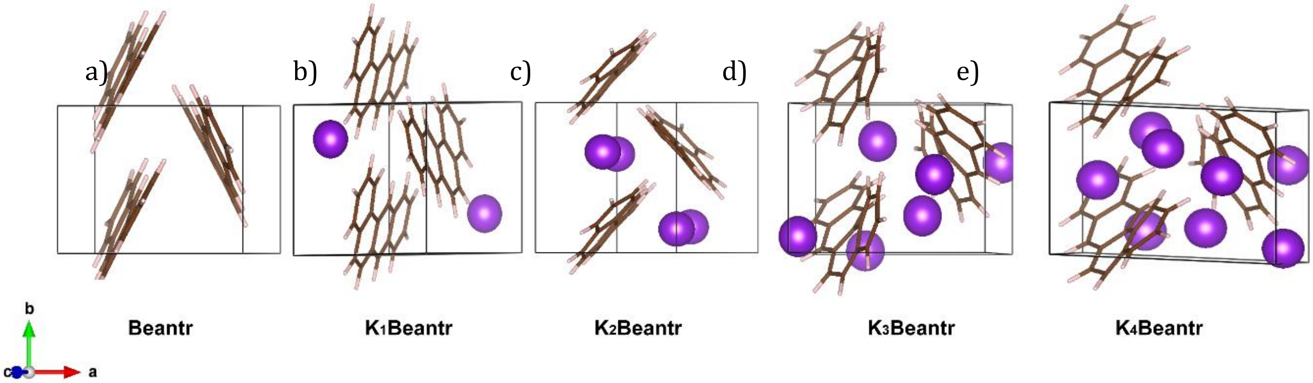


***Figure 4.12.** Structures of potassium-doped benzanthracene (KxBeantr, x = 1, 2, 3, 4) comparing with undoped case. (a)Bezanthracene, (b)K_1bezanthracene, (c) K_2bezanthracene, (d) K_3bezanthracene, (e) K_4bezanthracene. The purple balls represent K atoms. In K_2Beantr, which is the most stable configuration, the insertion of K atoms seems to promote the σ-π bonds (C-H…..π contacts).*

After the doping of the parent bezanthracene structure with different amount of potassium, the energetically favoured configuration was found to be that of K_2beantr with a unit cell volume expansion of **13.9%.** The void space in the parent herringbone structure lies between the PAH layers and is lined with C-H σ bonds (Figure 4.12 a). However, the insertion of potassium atoms seems to contribute to a structural rearrangement of the bezanthracene molecules, empowering the σ-π bonds (C-H…..π contacts) (Figure 4.12 b).

The convex hull of the USPEX generated structures for the $K_x$bezanthracene system is presented in Figure 4.13 a. Each point represents one of the generated compounds, described by the stoichiometric ratio and the energy difference as derived from equation 4.1. The most stable stoichiometry is the divalent (K_2bezanthracene), which is in agreement with the outcome of the simple intercalation approach (Figure 4.13b). For direct comparison of the energetic stability, the energies of the lowest USPEX generated structures (EA1046, EA602, EA979, EA1077) were recalculated in the same level of accuracy with the more accurate functional optB86b-vdW[185] and projected in the same convex hull (Figure 4.13b). It can be observed that the CSP results are comparable to the simple intercalation approach outcome in terms of the identified stable ratio, namely K_2bezanthracene.

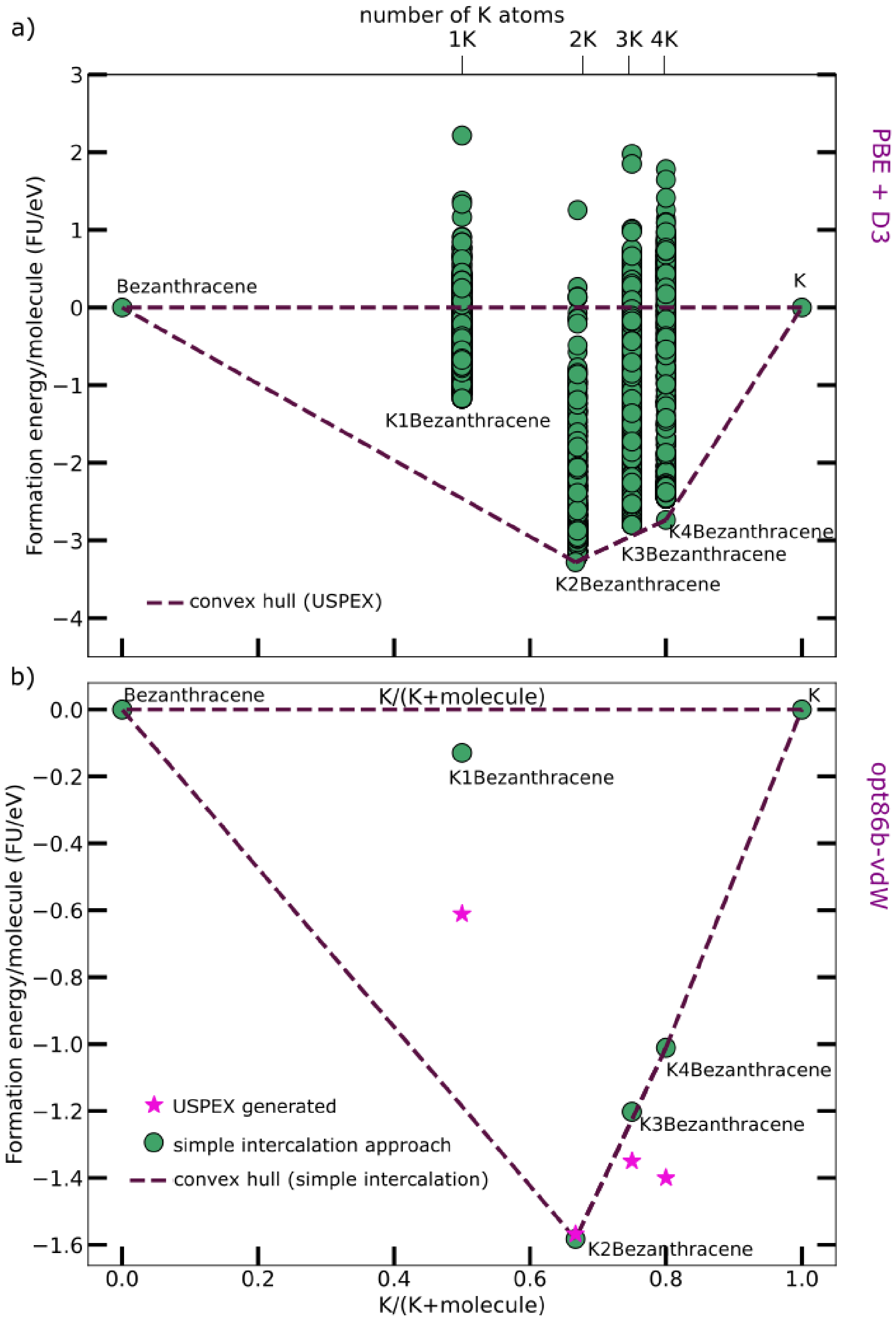


Figure 4.13. *Convex hulls of the KxBezantracene system. a) constructed with USPEX b) derived from the simple intercalation approach. The stable composition in both cases is K_2bezanthracene. The pink stars in b represent the lowest energy structures from USPEX reoptimized with higher accuracy with optB86b.*

Table 4.3. *The lattice parameters for various phases of $K_x$bezanthracene with x=1,2,3,4. The energies of all the systems were calculated with the same level of accuracy for direct comparisons. The most stable structure overall is highlighted in bold.*

Phases	**a**	**b**	**c**	**α**	**β**	**γ**	**Energy difference (eV/FU)**
K_1beantr_simple	7.054	7.378	12.101	90.153	114.311	88.968	-0.129
K_2beantr_simple	7.236	7.997	11.795	89.992	110.734	90.004	**-1.582**
K_3beantr_simple	9.516	6.834	11.046	89.646	96.979	90.012	-1.202
K_4beantr_simple	9.937	7.184	10.962	90.405	97.548	92.621	-1.011
K_1beantr_uspex (EA1046)	14.683	6.822	6.835	108.740	79.034	101.073	-0.611
K_2beantr_uspex (EA602)	12.248	6.222	9.802	89.988	89.953	63.323	-1.568
K_3beantr_uspex (EA979)	10.585	13.371	5.360	87.393	89.910	90.211	-1.328
K_4beantr_uspex (EA1077)	12.865	11.584	5.473	90.192	75.076	96.451	-1.407
Pristine bezanthracene	7.958	6.50	12.121	90.000	100.500	90.000	-

The lowest energy structures for each stoichiometry as obtained with both the simple intercalation and USPEX methods are listed in Table 4.3. Overall, the divalent stoichiometry K_2beantr_simple is the energetically favourable structure. The alkali metal atoms are found in intralayer positions, *i.e.*, between planes defined by PAH molecules. The fully optimized values for the unit cell length and angles are a=7.236, b=7.997, c=11.795, α=89.9°, β=110.734°, γ=90.0°, whereas the herringbone pattern of pristine bezanthracene is preserved (Figure 4.14).

The doping of bezanthracene results in a very small volume expansion from 615.814 to 638.436 Å^3 whilst the unit cell parameters remain similar to pristine bezanthracene. The distance between the alkali metals was found to be 4.141 Å, whereas the distance between the metal and the centre of the aromatic ring is 2.901 Å, which is in agreement with the extracted statistics from metal-π contacts.

The synthesis of the $K_x$bezanthracene system was performed by Dr Angelos Tsanai and the PXRD data were analyzed by Dr Rhian Patterson and Dr Rebecca Vismara. According to the experimental findings a new intercalated phase has been identified, which most probable is a mixture of two phases. Different compositions and temperatures have been tried to identify those in which the pattern remains the same. There is a high possibility that the intercalated phase consists of two different stoichiometries, however the PXRD patterns have not been solved and there is no exact match between the simulated low energy structures and the experimental pattern. It is most likely that the experimental structure involves four bezanthracene molecules, considering the possible unit cell dimensions suggested from the crystal structure analysis software used.

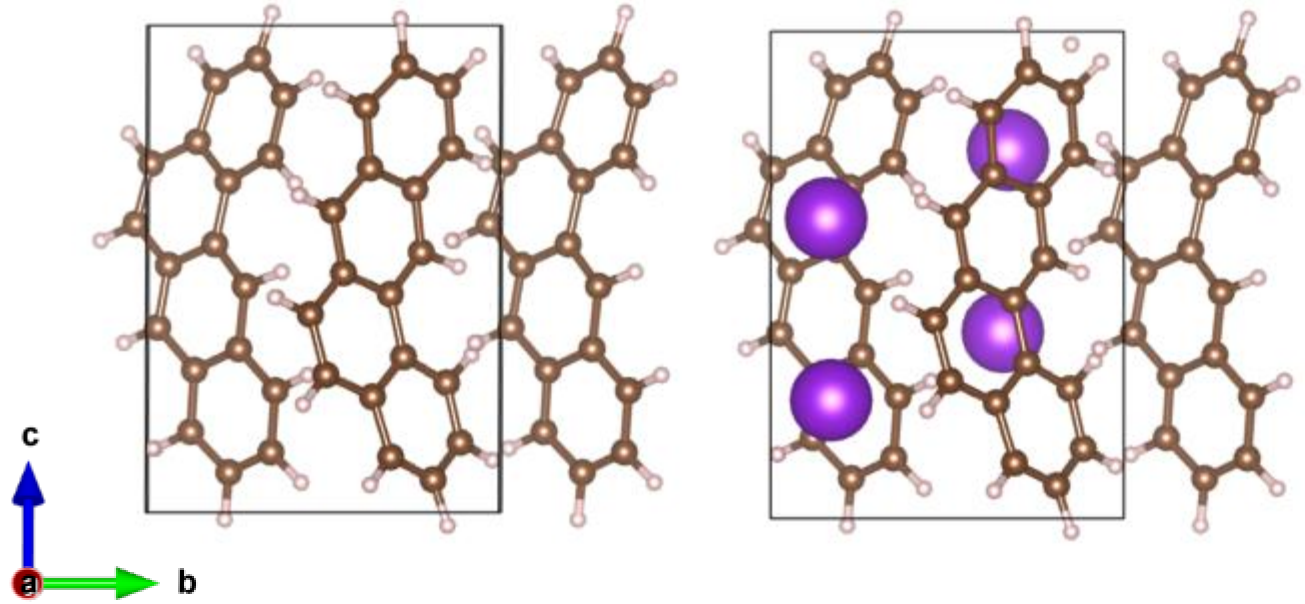


Figure 4.14. Crystal structures viewing along the a axis of a) pristine bezanthracene. The unit cell contains two molecules and b) K_2Bezanthracene. The unit cell contains two molecules and four K atoms in the intralayer space.

4.4.5 The electronic properties of potassium-doped bezanthracene

Herein, we have obtained the electronic structure of solid bezanthracene, both doped and undoped. As the K_2Bezanthracene structure obtained with the simple intercalation approach was found to be the energetically favourable one it was used as representative for the doped case. The nonmagnetic, the ferromagnetic (FM) and the antiferromagnetic (AFM) structures were considered for the doped case, with lattice energies -369.519 eV, -369.320 eV, -369.519 eV, respectively. The nonmagnetic solution was the most stable, and thus was used for further electronic structure analysis.

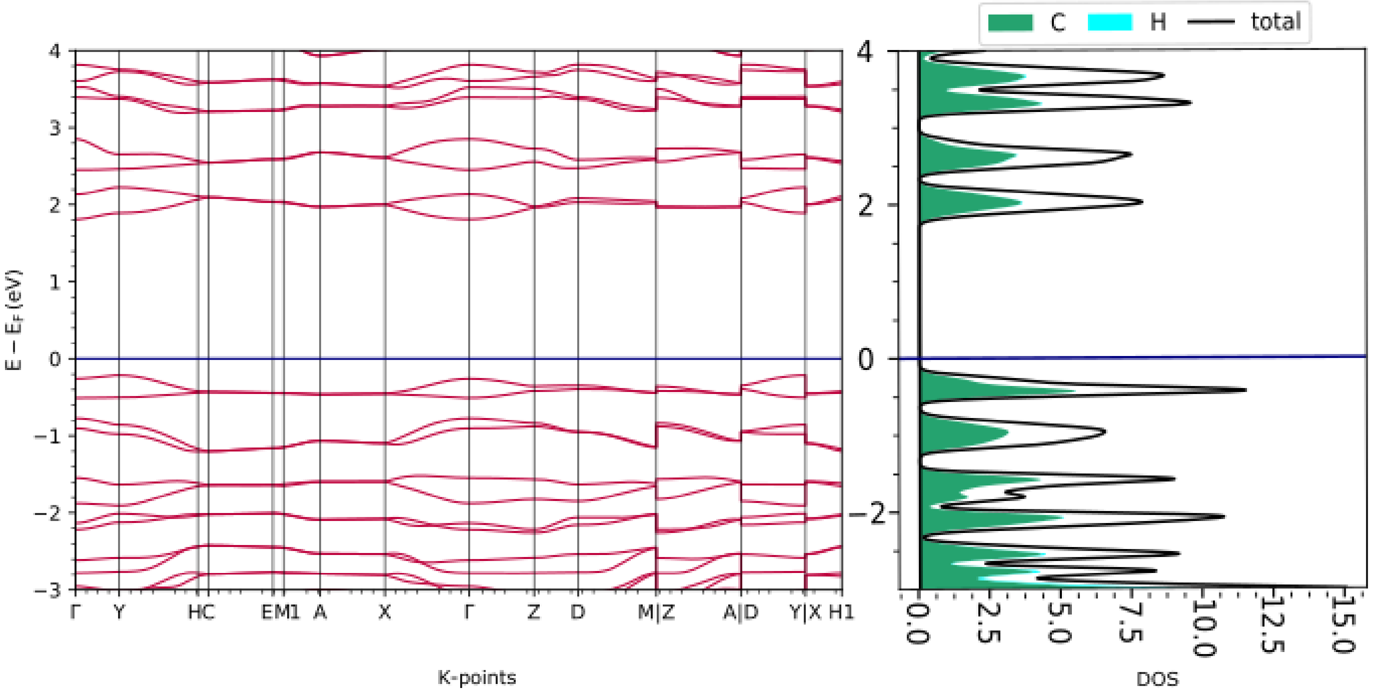


Figure 4.15. *The electronic structure of pristine benzanthracene. The band structure (left) and density of states (right) are shown for the relaxed structure of bezanthracene(BEANTR). The horizontal blue line indicates the Fermi level.*

In the case of the undoped structure (Figure 4.15), the molecules take a herringbone pattern, and the first unoccupied band above the Fermi level (conduction band) is a mixture of the LUMO states of the two molecules in the unit cell forming two entangled bands. The second unoccupied band is a mixture of the LUMO$_{+1}$ states and is separated from the first by 0.2 eV.

It is expected that in the case of doped bezanthracene with two potassium atoms per molecule the two electrons per bezanthracene molecule would fill the empty bands corresponding to the LUMO states resulting in two filled bands and no metallic character is expected. Indeed, the electronic structure analysis of the K$_2$bezanthracene system (Figure 4.16) showed that the final structure consists of a filled LUMO derived band and an unoccupied LUMO$_{+1}$ derived band, with the gap between these two gaps being calculated to be 0.3 eV. The Fermi level is not intersecting any of the bands, so there are not any partly filled bands in the doped structure.

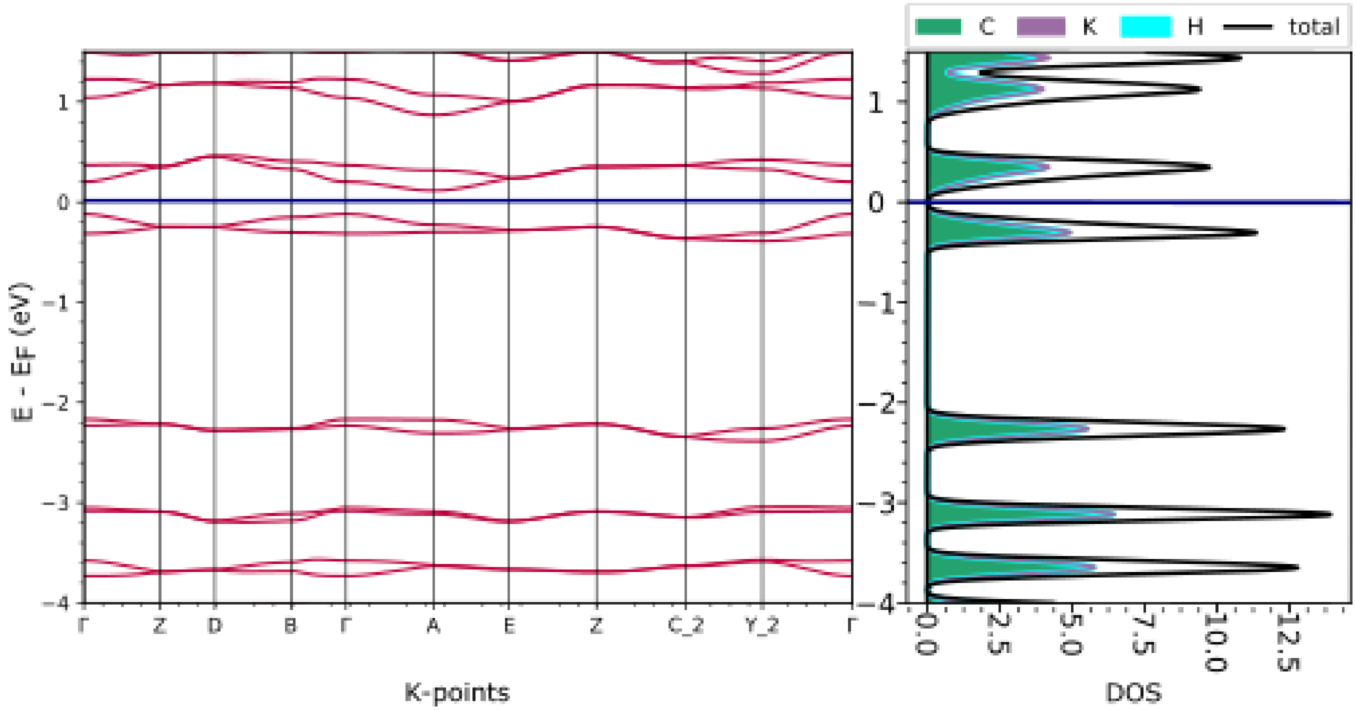


Figure 4.16. *The electronic structure of K_2benzanthracene. The band structure (left) and density of states (right) are shown for the relaxed structure of the K_2bezanthracene. Based on the position of the Fermi level (blue line), the LUMO band is filled with electrons whereas the $LUMO_{+1}$ is unoccupied.*

4.4.6 The atomic structure of potassium-doped coronene from a first-principles study

Similar analysis as for the potassium-doped benzanthracene was performed for the potassium-doped coronene. A full determination of the doped coronene structure based on x-rays diffraction data has not been possible so far because of the large background and the small number of useful measured peaks. As a result, the crystal structure identification can be only performed by simulating existing crystal structures of single coronene after metal insertion. Coronene consists of six benzene rings arranged in a ring-like manner. The unit cell of pristine coronene contains two molecules arranged in herringbone pattern. Pristine coronene has a monoclinic structure (space group of $P2_1/\alpha$) and the lattice parameters are a = 16.094 Å, b = 4.690 Å, c = 10.049 Å, and β = 110.79°. All the structures generated with USPEX as well as the structures resulted from the simple metal insertion approach were compared in terms of relative lattice energy as shown in Figure 4.17. Our results indicate that the possible stoichiometric content of potassium in $K_x$coronene compounds is x=3, as verified from both the USPEX generated structures and the simple intercalation approach.

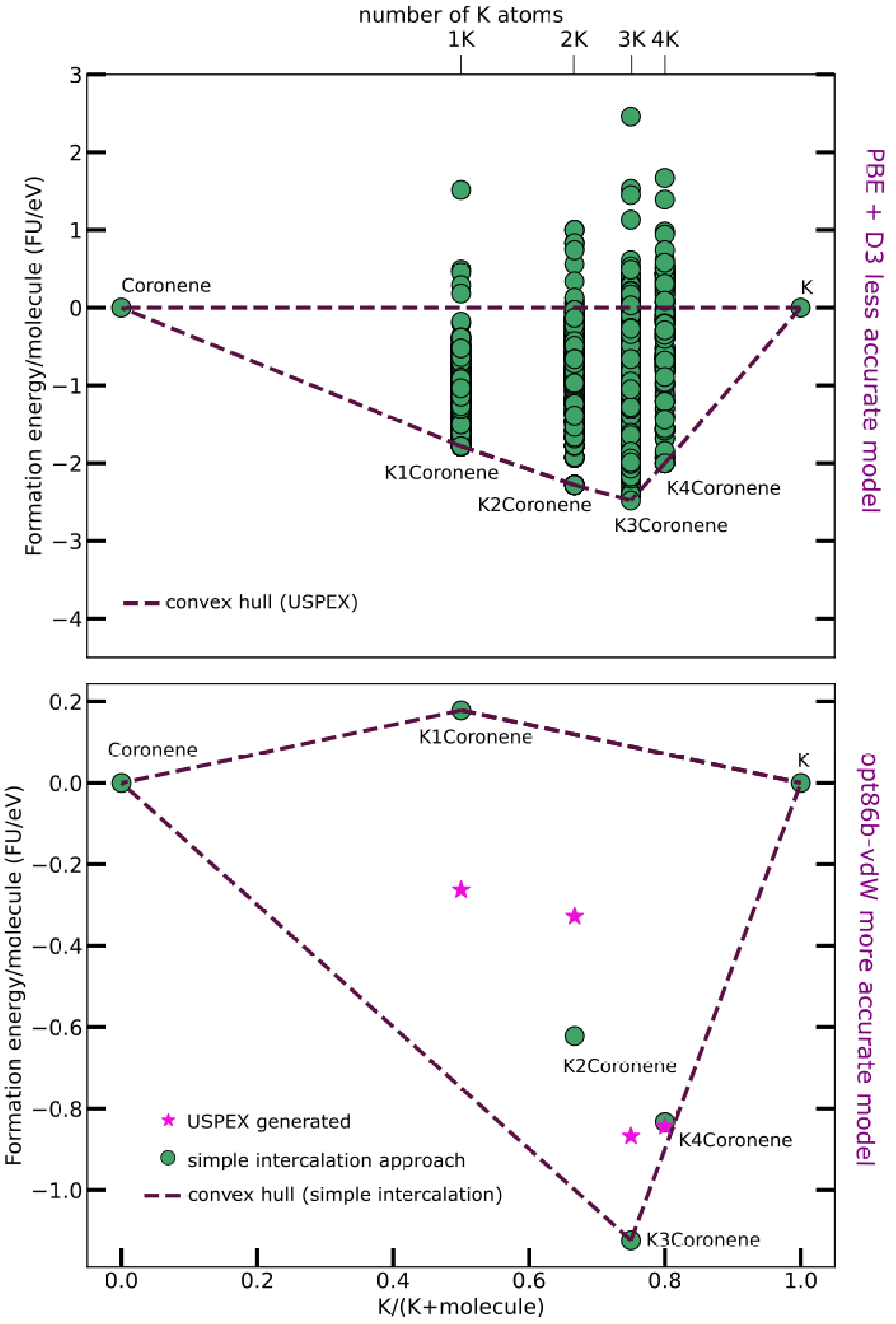


Figure 4.17. *Convex hull constructed for $K_x$coronene with x=1,2,3,4 a) using USPEX software for crystal structure prediction b) following the simple intercalation approach. The most stable composition identified from both methods was for x=3 (K_3coronene). The pink stars in b represent the lowest energy structures from USPEX reoptimized with higher accuracy with optB86b.*

The lattice parameters of the most stable structures of each method and stoichiometric ratio are shown on Table 4.4 and the detailed calculations on the lattice energies are on the SI, Table A.1. For the cases of x=4 USPEX identified a more stable structure whereas for all the other ratios the simple intercalation approach resulted in the energy minima. In the USPEX generated K_2Coronene, although around 1,000 structures were generated, the structure EA420 was the best found and remained the same for 8 generations.

Table 4.4. *The lattice parameters for various phases of KxCoronene with x=1,2,3,4. The formation energy is measured in eV per functional unit (FU).*

Phases	**a**	**b**	**c**	**α**	**β**	**γ**	**Formation energy (eV/FU)**
K_1Coronene_simple	16.582	4.596	10.720	89.999	116.944	90.001	0.178
K_2Coronene_simple	14.636	6.639	7.914	90.011	95.046	90.023	-0.622
K_3Coronene_simple	11.627	7.791	9.743	91.639	108.042	92.900	**-1.124**
K_4Coronene_simple	17.710	5.864	10.113	90.103	118.215	89.953	-0.833
K_1Coronene_uspex (EA1056)	7.519	7.017	14.145	89.055	90.640	86.311	-0.263
K_2Coronene_uspex (EA420)	9.916	4.924	19.309	95.360	84.256	90.950	-0.328
K_3Coronene_uspex (EA1305)	10.024	5.341	17.567	90.123	90.074	74.002	-0.868
K_4Coronene_uspex (EA950)	15.547	5.681	10.543	90.128	90.098	81.966	-0.844
Pristine Coronene	10.014	4.662	15.575	90	106.53	90	-

It was found that the structure derived from the simple intercalation approach with a doping level of x=3 is the most stable with the largest formation energy (-1.1244 eV/FU) in all the above structural phases. The alkali metal atoms are found in intralayer positions, *i.e.*, between planes defined by PAH molecules. The fully optimized values for the unit cell length and angles are a=11.627, b=7.791, c=9.743, α=91.639°, β=108.042°, γ=92.90°, whereas the herringbone pattern of pristine coronene is preserved (Figure 4.18). The doping of coronene results in a volume expansion from 715.27 to 837.239 $Å^3$ and significant changes of the *b* and *c* unit cell lengths. The distance between the alkali metals was found to be 5.64 Å, whereas the distance between the metal and the centre of the aromatic ring is 3 Å, which is in agreement with the extracted statistics from metal-π contacts.

Experimental work on the $K_x$coronene system was performed by Dr Angelos Tsanai and the PXRD data were analyzed by Dr Rhian Patterson. According to the experimental findings, a new intercalated phase has been identified. Different compositions and temperatures have been tried and it is postulated from the crystal structure analysis software used that the experimental structure affords 1:3 ratio with 8 coronene molecules and 24 K in the unit cell. The final structure has not been found yet for enabling direct comparisons with the theoretical ones.

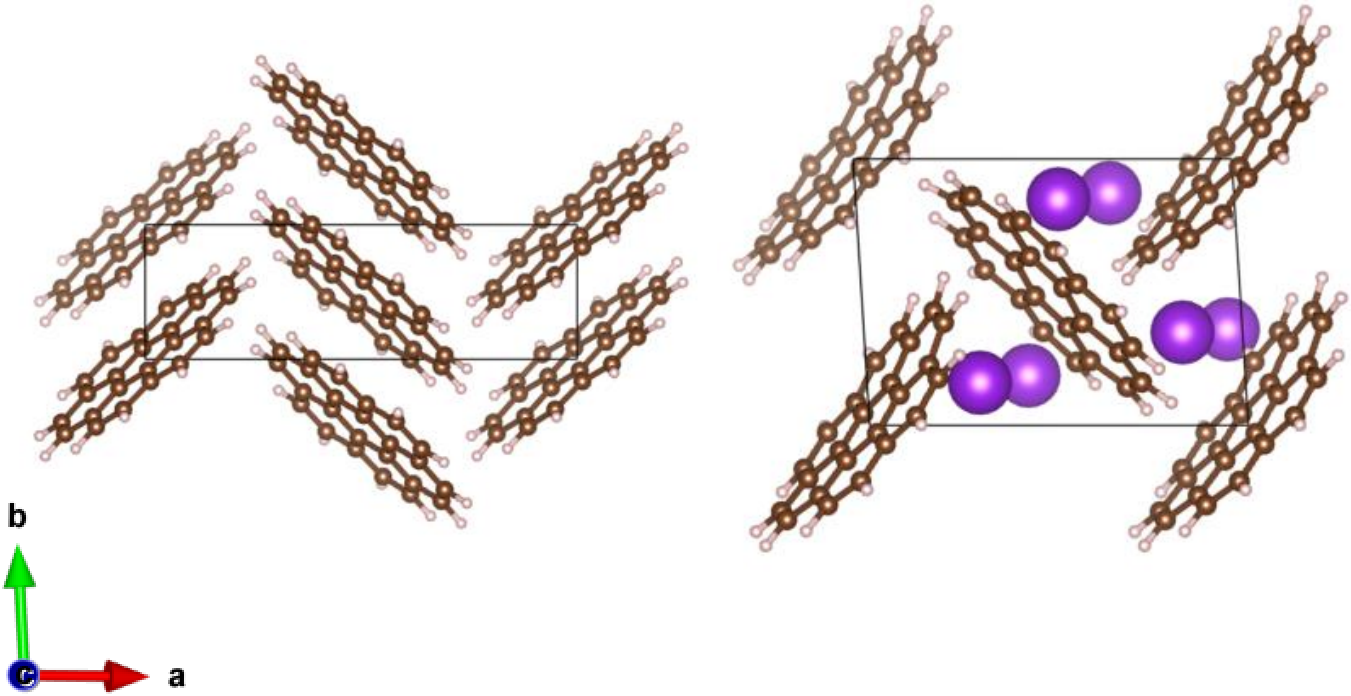


Figure 4.18. *Crystal structures viewing along the c axis of a) pristine coronene. The unit cell contains two molecules and b) K_3Corone. The unit cell contains two molecules and six K atoms in the intralayer space.*

4.4.7 The electronic properties of potassium-doped coronene

For having accurate electronic properties calculations, getting the magnetic structure is a critical step. As the structure of K_3Coronene_simple was found to be the most stable, further analysis for the electronic structure is performed using that structure. The ferromagnetic (FM) and antiferromagnetic (AFM) structures were obtained by allowing spin polarization in the initial K_3coronene structure. The lattice energies of the nonmagnetic, FM and AFM structures are -464.146 eV, -464.173 eV, -464.169 eV respectively, indicating that the FM solution is the most energetically favourable. Consequently, the relaxed ferromagnetic structure was used for further electronic structure analysis.

Both the undoped and doped coronene were analysed and with the band structures and density of states plots shown in Figures 4.18 and 4.19, respectively. In the undoped crystalline coronene, the two molecules in the unit cell take a herringbone structure. The conduction band consists of four entangled bands originating from the lowest two unoccupied molecular orbitals (LUMO and $LUMO_{+1}$) of each of the two molecules in the unit cell. The LUMO and $LUMO_{+1}$ orbitals are degenerate.

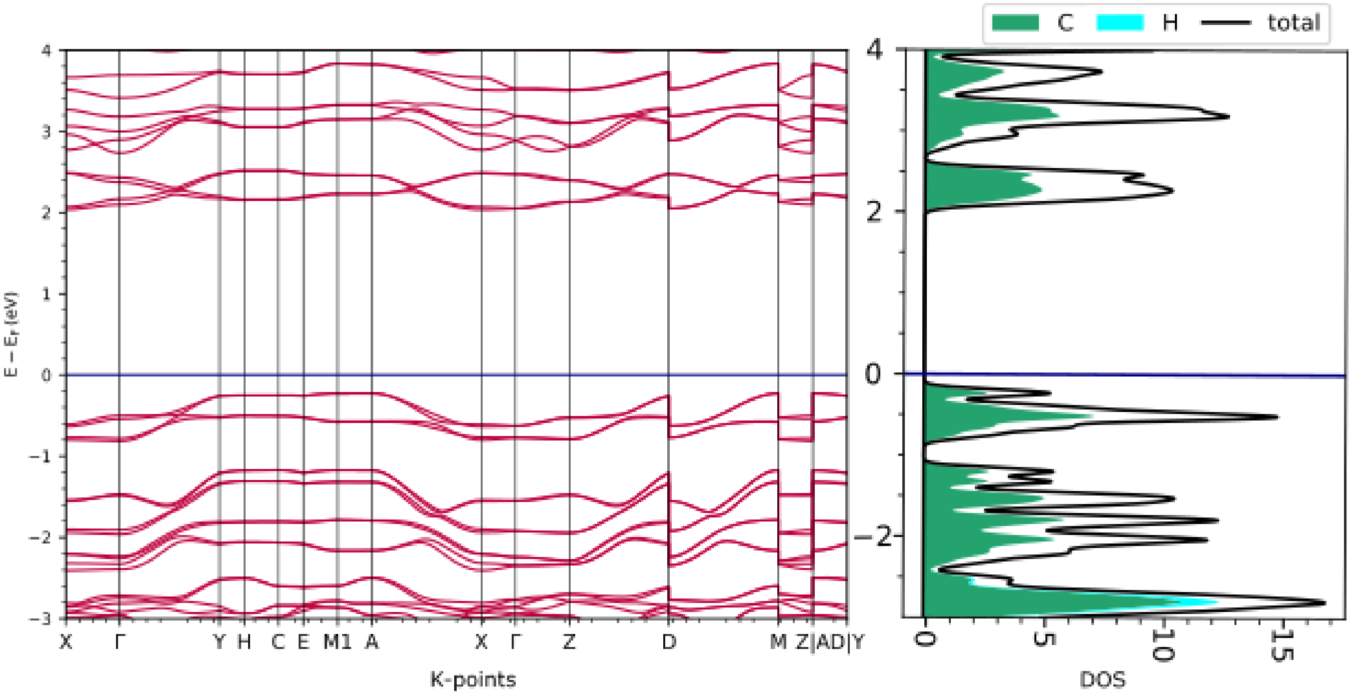


Figure 4.19. *Calculated electronic structure of undoped solid coronene. The origins of energy are set to their Fermi levels. The band structure (left) and density of states (right) are shown for pristine solid coronene. The first unoccupied band above the Fermi level (solid blue line) is a mixture of LUMO and LUMO$_{+1}$ states of the two coronene molecules in the structure.*

When coronene is doped with potassium atoms, the original double degeneracy of pristine coronene remains in the simulated structure. This leads to a metallic character after the two first unoccupied bands (which are also degenerate) of pristine coronene are filled with the electrons provided by the potassium metal, resulting in a ¾-filled two-band system. Although the metallicity of coronene containing open shell molecular ions has been already claimed from other theoretical studies,[81] the experimental outcomes are contradictory. Potassium intercalated coronene films have been studied using photoemission spectroscopy. However, no emissions from the Fermi level were experimentally measured, ruling out the possibility for a metallic ground state.[193]

As there are not yet any published data that allow a detailed structure refinement and thus enabling the determination of the real crystal structure, we can only speculate regarding the existence of any metallicity in K_3coronene. Nonetheless, even a small structural difference could be responsible for the different metallic or insulating ground state at the same doping level, since this small difference might change the balance between the bandwidth (kinetic energy gain) and the Coulomb repulsion in compounds with an integer doping level.[193] Reports for these metal insulator transitions exist both in theoretical and experimental studies for K_3phenanthracene[58] and K_3C_{60}, respectively.[71] In the case of doped phenanthrene among the lowest energy structures generated using USPEX, one is a band insulator and the other is metallic. Whereas in the case of doped fullerene it was found that a lattice expansion and symmetry lowering in the metallic K_3C_{60} or a change of the lattice symmetry transiting from K_3C_{60} to K_4C_{60} results in an insulating ground state.[194]

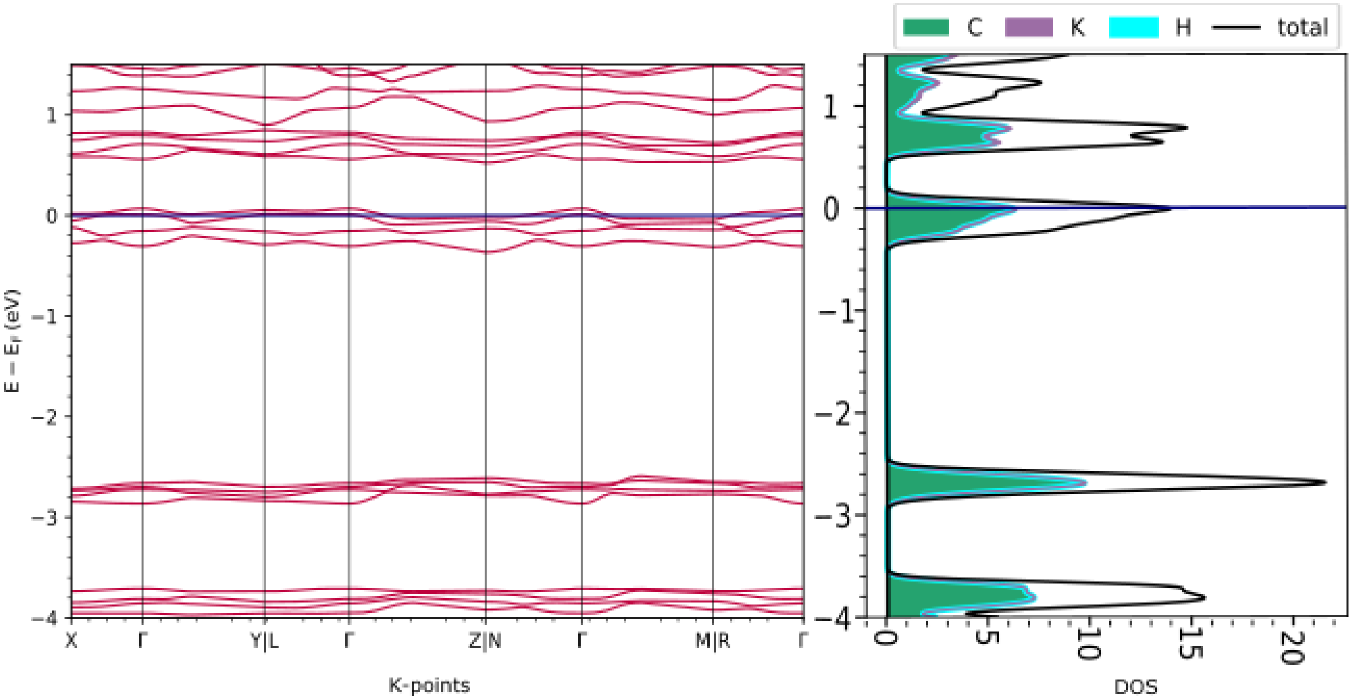


Figure 4.20. *Calculated electronic structure of the potassium-doped solid coronene. The origins of energy are set to their Fermi levels. The band structure (left) and density of states (right) are shown for the simulated K_3coronene. The first unoccupied band above the Fermi level (solid blue line) is a mixture of LUMO and $LUMO_{+1}$ states of the two coronene molecules in the structure, indicating that the original double degeneracy of LUMO is retained after the metal insertion. This band is filled with three electrons (provided by potassium) per coronene molecule resulting in a ¾ filled band and indicating a metallic character.*

4.4.8 Overall proposed strategy

The overall proposed process for the theoretical investigation of the metal-PAH systems is summarized in Figure 4.21. Starting from large molecular databases, the molecules with exact double degeneracy are found to have a better perspective for keeping this degeneracy in the crystal structure and accommodate half-filled bands after the electrons insertion. The next parameter to be considered is if these molecules have a known crystal structure such that the available void space can be calculated. A large enough void space could indicate that the structure will not change significantly after the metal insertion and thus the orbital degeneracy will not be affected. Given the crystal structure, it was found that by applying a simple intercalation approach can give a very good estimation about the metal capacity, although the experimental crystal structure might not be found. Previous works[195,196] performing an *ab initio* workflow for metal-PAHs similar to ours are mainly based on the assumption that the unit cell parameters do not change significantly upon metal insertion. However, the recent well-characterized structures reveal a different scenario, where the lattice is being modified significantly *e,g.*, K_2tetracene. For this reason, we tried to go further from these assumptions and used a CSP method to generate from scratch structures using a genetic algorithm.

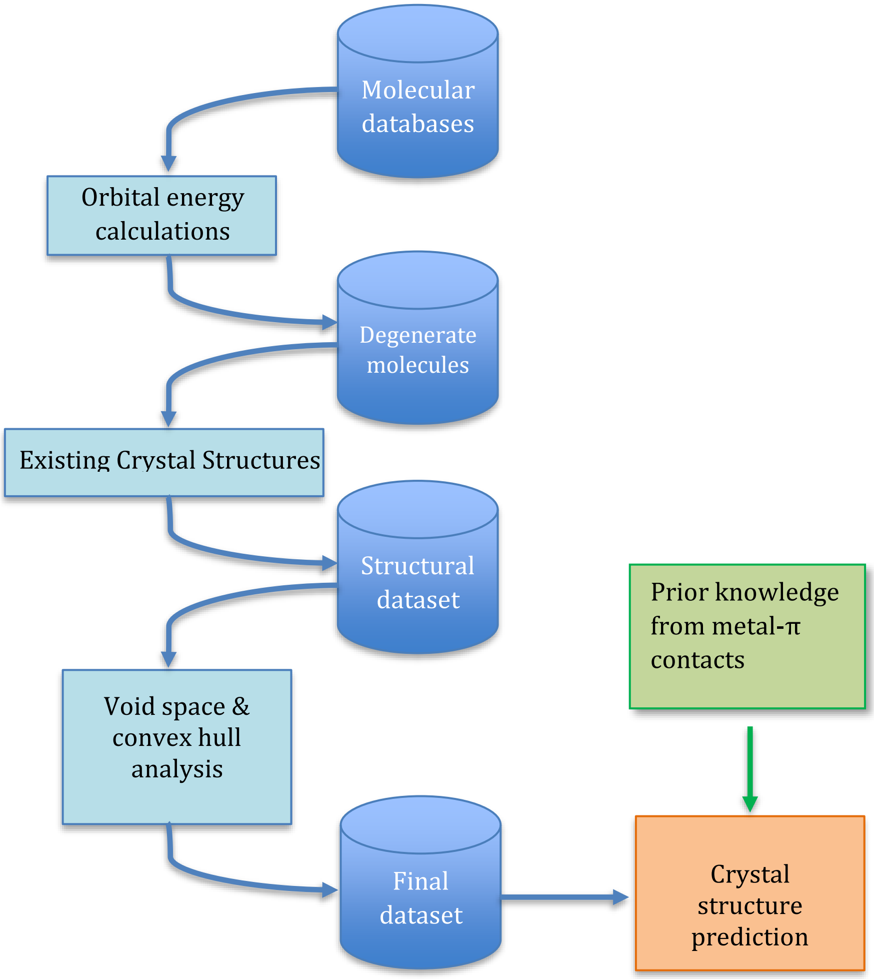

Molecular databases
Orbital energy calculations
Degenerate molecules
Existing Crystal Structures
Structural dataset
Prior knowledge from metal-π contacts
Void space & convex hull analysis
Final dataset
Crystal structure prediction

Figure 4.21. *Overall proposed strategy for rationalizing the selection of PAHs for intercalation.*

CSP could be beneficial in the cases where some of the initial parameters are known. Firstly, the distance of the molecular centers was found to be very important for generating structures with high density as demonstrated in the Appendix, Figures B3.1 & B3.2 Moreover, the knowledge of the number of compounds in the unit cell is very important to generate structures which are comparable to the synthesized ones. Another limitation of USPEX workflow is that the dispersion correction method used (PBE + D3) describes metal-PAHs systems less accurately than the more computationally expensive optB86b-vdW functional and the ranking of the structures may differ according to which functional in used.

4.5 Discussion and perspectives

The scope of this chapter was to investigate the metal-unsaturated hydrocarbon salt chemistry. To this date, the metallic and superconducting behaviour of many of these exciting systems is proving difficult to reproduce and there is considerable uncertainty, ranging from stoichiometry, mechanism and precise crystal structure. For that reason, simultaneous efforts of experiment and theory are necessary. This work is aiming to propose new strategies for the identification of the most appropriate candidates for metal insertion and can be divided into four major parts:

1) Statistical analysis of the existing metal-polyaromatic hydrocarbons (PAHs) interactions in the Cambridge Structural Database to extract important parameters that have been observed in these systems, *e.g.* minimum distance between the alkali metals and the molecular centres.

2) Computational analysis of all the known pure phase metal-PAHs structures, measuring the energetic stability, the void space modification, and the orbital degeneracy. Our calculations demonstrate that in all the known cases the doped structures were on average -1.92 eV/FU more stable than their constituents. However, although the energetic stability is the driving force for the formation of metal-PAHs systems, there are many prohibitive parameters related to the experimental conditions used that might led to unwanted decomposition of molecules.

3) Following the observations on the known data, we develop new strategies for selecting the next most promising systems having as a starting point the PAHs datasets created in Chapter 3. The selection criteria involve large void space, degenerate LUMO orbitals and high metal capacity. An *ab initio* method has been developed for measuring the stability of PAHs after doping with alkali metals.

4) Crystal structure identification. As all the properties are derived from the crystal structure, two different methods were tested for the theoretical determination of the crystal structure of metal-PAHs systems. The first method employs Zeo++ to identify the appropriate positions in the structure to insert metal ions and then density functional theory is used to optimize the structures allowing for complete variational freedom of the crystal structure parameters and the molecular atomic positions. Several PAHs were explored within that framework after being theoretically intercalated mainly with potassium. The second method tested is a well-established software, namely USPEX which is able to handle systems containing both organic and inorganic components. As this approach is more time and resource consuming only two detailed studies were performed for two interesting candidate systems, K_xCoronene and K_xBezanthracene. The structures derived from USPEX are comparable with the structures of the simple intercalation approach. Both methods agree on the most stable stoichiometry, x=2 for K_xBezanthracene and x=3 for K_xCorone. From the calculated electronic structure, we expect insulator behaviour for potassium-doped bezanthracene and metallic behaviour for potassium-doped coronene. It can be concluded that Crystal Structure Prediction for the metal-PAHs systems can only be beneficial in the cases where some starting information about the structure exists, e.g. the number of molecules in the structure or the possible unit cell parameters. Otherwise, it

is a very time-consuming method and the only reliable information that we can get is the stoichiometry that could also be obtained from the simple intercalation approach. For CSP the initial constraints such as the distance of the molecular centers are very important for generating more sensible structures. After testing and optimizing the USPEX workflow in a benchmark system, *i.e*., K_2Tetracene, where the structure is known, it was found that starting the structure generation given lattice parameters close the experimental ones (which could be derived from the analysis of a PXRD pattern) resulted in structures very similar to the experimental (as demonstrated from the PXRD patterns comparison in the Appendix Figure B3.3 and the unit cell parameters comparison in Table B3.2).

Our main conclusions in this chapter can be summarized as following; if a metal-PAH system is formed then that is going to be more stable than its components. However, if a theoretical system is calculated as more stable than the constituents that does not automatically mean that the crystal will be formed as some experimental restrictions might apply, *e.g.*, decomposition of the PAH. The formation of a metal-PAH system is a thermodynamically driven process which was quantified by calculating the formation energies on all the currently known systems. For the identification of the most likely metal ratio in the structure a simple intercalation approach, *i.e.*, void space analysis and manual insertion of K atoms, gives comparable results with the time consuming CSP method. A list with the molecules identified as good hosts for metal insertion is given in the Appendix, Table A1.2. Further work in this field will require faster crystal structure prediction approaches, such as coupling CSP software with low cost and relatively accurate models (*e.g.* DFTB+). For the crystal structure identification, close collaboration with crystallographers to extract important information related to lattice parameters will be beneficial for initiating the CSP given some constraints.

Having established the protocol for identifying promising PAH molecules and stable compositions for their intercalation with alkali metals, the future of work in this area will focus on the simultaneous improvement of the computational and synthetic approaches to design and realise more materials of this type.

5 Accelerating π-π co-crystal discovery with One Class Classification

This work is reproduced from Ref. 253 with permission from the Royal Society of Chemistry.

5.1 Introduction

Machine learning approaches are being increasingly incorporated into the design workflows to explore and better understand the materials space.[13,197,198] The ultimate goal is to identify more reliable methodologies and to develop smarter ways to accelerate the discovery of new materials with novel properties. Following the rapidly growing data availability, data-driven approaches have taken hold as a tool for detecting patterns in known datasets and performing straightforward predictions. However, they still suffer from many limitations in terms of defining the appropriate representations of the target materials and/or achieving reliable predictions based solely on known instances or otherwise biased datasets. One matter of concern for the data-driven approaches is the lack of negative data from unsuccessful synthetic attempts, which might generate inherently imbalanced datasets. In this chapter we introduce a data-driven workflow based on one-class classification, which is a method specifically designed to address the issue of 'positive-only' data. An extensive study on the different one-class classification algorithms was performed in order to identify the most appropriate workflow for guiding the discovery of the weakly bound polyaromatic hydrocarbon co-crystals. The two-step approach presented in this study first trains the model using all the known molecular combinations that form this class of co-crystals extracted from the Cambridge Structural Database (1722 molecular combinations), followed by scoring possible yet unknown pairs from the ZINC15 database (21736 possible molecular combinations). Focusing on the highest-ranking pairs predicted to have higher probability of forming co-crystals, materials discovery can be accelerated by reducing the vast molecular space and directing the synthetic efforts of chemists. Furthermore, a more detailed understanding of the molecular properties which lead to co-crystallization is sought after with the use of interpretability techniques. The applicability of the current methodology is demonstrated with the discovery of two novel co-crystals, namely pyrene-6H-benzo[c]chromen-6-one (1) and pyrene-9,10-dicyanoanthracene (2). The electronic structure analysis of the two synthesized co-crystals reveals that (2) has a band gap in the range of semiconducting materials.

5.1.1 Co-crystal definition

As discussed in the Introduction, a co-crystal is a crystalline single-phase material composed of two or more different molecular compounds in a specific stoichiometry.[199–201] These compounds are neither solvates/hydrates nor simple salts and they are connected via one or more non-covalent interactions, such as hydrogen bonding, π-π stacking, halogen bonds and charge transfer(C-T) interactions.[202] Co-crystal design has undoubtedly received a lot of attention from the Pharmaceutical Industry. These compounds may offer the advantage of preserving the pharmacological properties of the Active Pharmaceutical Ingredient (API) whilst improving the physicochemical properties of the potential drug. Consequently, this attention stimulated the development of various theoretical and experimental studies for designing pharmaceutical co-crystals by selecting effective coformers which are suitable with the API.[203] Hydrogen bond propensity (HBP), pKa rule, Fabian's method for molecular complementarity and Hansen solubility parameters are some of the most effective designing approaches.[203] The selection of the appropriate method is based mainly on the nature of the molecules and the way these molecules are interconnected.[199,204]

5.1.2 Co-crystals with electronic properties

Co-crystals are gaining emerging interest in other cutting-edge research fields, ranging from photonic, to optical and electronic materials.[205–207] It is well-known that most organic molecular crystals are insulators as there is no electronic interaction between the molecules.[208] However, molecules with electron rich π-orbitals could possibly overcome this barrier, thus enabling the electron mobility in cases where there is a favourable overlap of π-orbitals in adjacent molecules.[209] π-π stacking is a common motif for getting electronic communication between the molecules and has been proven to be an important characteristic of organic electronics (*e.g.* in conjugated polymers).[210,211] A special category of molecules which self-assemble via π-π interactions are the polycyclic aromatic hydrocarbons (PAHs), which are regarded as two-dimensional graphite segments.[78] Hence, PAHs are possibly considered promising candidates for electronic materials and have been extensively used for designing co-crystals with desirable electron mobilities.[205,212,213] Most of the research on electronic co-crystals is focused on the charge-transfer complexes between a good electron donor and a poor electron acceptor.[212,214,215] This work suggests a promising pathway to expand the investigation on PAHs based co-crystals where the π-π interactions are the dominant structure-defining forces.

In this context, the strong structure directing groups such as hydrogen-bonding are eliminated, and targeted exploration of carbon-based π-electron systems is performed. Although π-π interactions are desirable for designing electronic functional co-crystals, they are relatively weak compared to stronger interactions such as hydrogen or halogen bonding. In a recent computational work, Taylor *et al.* emphasized the difficulty in evaluating the thermodynamic stability of weakly-bound co-crystals without any additional group that can form charge transfer systems.[53] The lack of strong energetic driving forces for co-crystallization makes the formation less favourable, thus these co-crystals are rare. In addition, the weak interactions give rise to shallow energy landscapes associated with multiple configurations of similar energy, hindering the structure prediction. The synthesis of weakly-bound co-crystal materials still remains a challenging task, albeit interaction between aromatic hydrocarbon systems have been suggested as a viable synthetic way on first principle calculations.[216] All these evidences bring to light the challenging prediction of π-π co-crystallization.

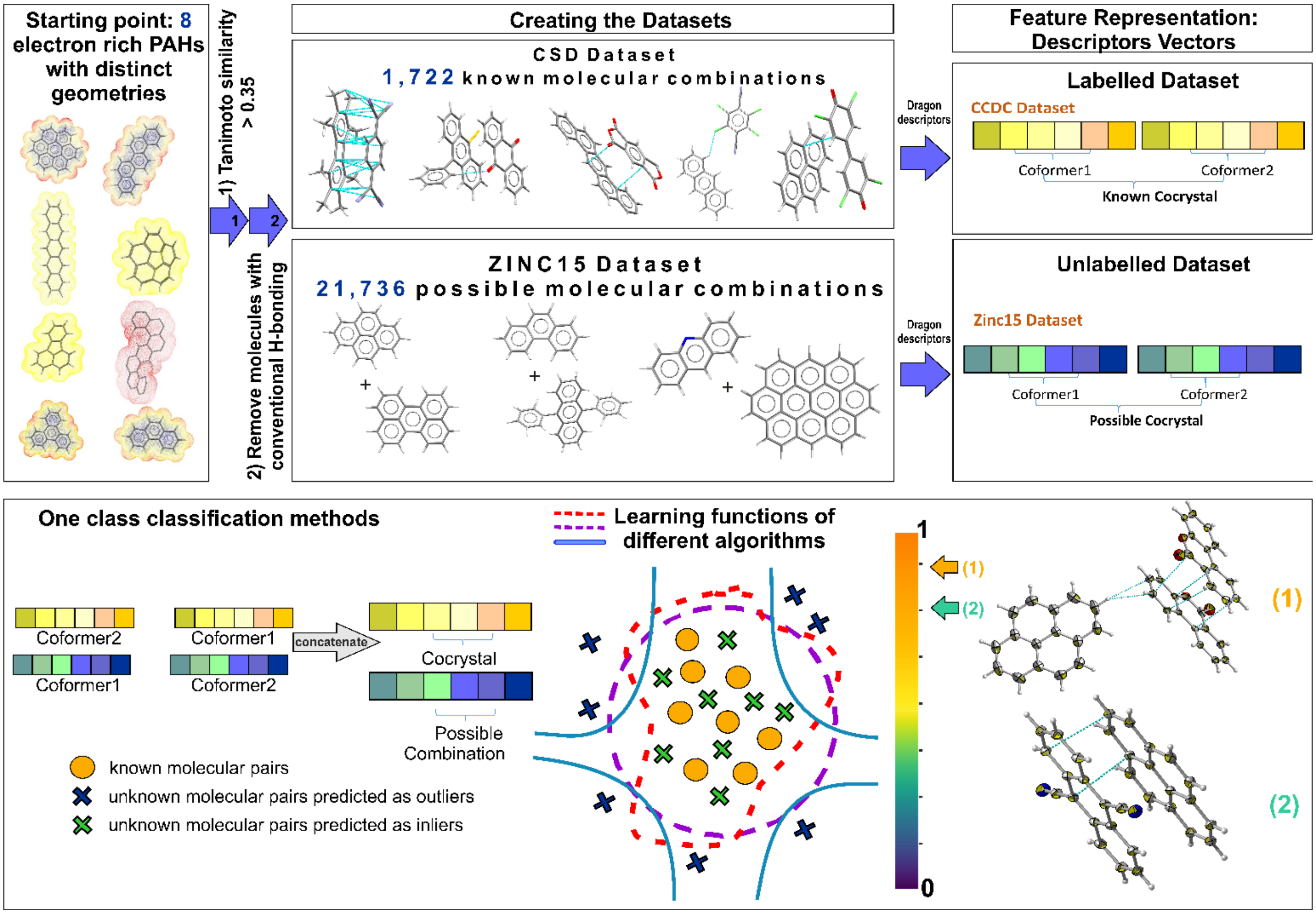

Starting point: 8 electron rich PAHs with distinct geometries
1) Tanimoto similarity > 0.35
2) Remove molecules with conventional H-bonding
1
2
Creating the Datasets
CSD Dataset
1,722 known molecular combinations
ZINC15 Dataset
21,736 possible molecular combinations
+
+
+
Dragon descriptors
Dragon descriptors
Feature Representation: Descriptors Vectors
Labelled Dataset
CCDC Dataset
Coformer1
Coformer2
Known Cocrystal
Unlabelled Dataset
Zinc15 Dataset
Coformer1
Coformer2
Possible Cocrystal
One class classification methods
Coformer2
Coformer1
Coformer1
Coformer2
concatenate
Cocrystal
Possible Combination
known molecular pairs
unknown molecular pairs predicted as outliers
unknown molecular pairs predicted as inliers
Learning functions of different algorithms
1
0
(1)
(2)
(1)
(2)

Figure 5.1. *Proposed one class classification workflow. Starting from eight representative PAHs, two molecular pairs datasets were constructed, i) the labelled, including all the known co-crystal forming pairs from CSD ii) the unlabelled, including all the possible pairs extracted from ZINC15 database. Various one class classification algorithms were trained only on the positive data and were further used to provide a ranking for the unlabelled pairs. The reliability of the presented procedure is tested and supported by experimental data.*

Reproduced from Ref. 253 with permission from the Royal Society of Chemistry.

5.2 Methods

5.2.1 One-class classification/outlier detection algorithms

Our method is based on one-class classification, a well-known method that has been applied under many research themes, such as novelty/outlier detection, concept learning or single class classification.[117] It is imbalance tolerant, so no specific distribution of the target class has to be assumed. The objective of one-class classification approaches is to accurately describe the 'normality', namely the distribution of the known dataset. It is assumed that the majority of the training dataset consists of ‘normal’ data.[217] Thus, the one class classification algorithms learn to accurately describe the positive/known data. Deviations from this description are seen as anomalies and thus belong to a different class. The known data class is well characterized, and these instances are used as the training set. In this way the classifiers are focused on the deviations from the known distribution rather than focusing on the discrimination task between the data. The existing algorithms for one-class classification/outlier detection are discussed below:

Distribution based. Methods in this category are basically inspired from statistical modelling, that deploy some standard distribution model and flag as outliers the instances that deviate from the model, whereas inliers are those that follow the same distribution.[218] Typical examples are the Autoencoders and the Gaussian Mixture models.

Density based. These methods assume that normal data points occur around a dense neighborhood. The local outlier factor (LOF) approach is one of the well-known algorithms in this category, where normal points get low LOF values as they belong to a local dense neighborhood. The neighborhood is defined by the distance to the MinPts-th nearest neighbor, with MinPts being the minimum number of neighbors used for defining the local neighbourhood.[219]

Distance based. Among other distance based methodologies, k-nearest neighbour algorithm is ranking each point on the basis of its distance to its kth nearest neighbor.[2.7] The lower the distance the closest to the normal data is the point.

Clustering based. Clustering Based Outlier Factor (CBLOF) is an algorithm developed for considering both the size of clusters and the distance between points and the closest cluster. Each datapoint is then assigned a score/outlier factor based on these considerations.[221]

Support Vector Machine. One class support vector machine algorithm (OC-SVM) is an extension on the well-known support vector machine technique. The planar approach of OCSVM is about finding a linear boundary to maximally separate all the data points from the origin, whereas the spherical approach designs a spherical boundary in feature space around the data and the algorithm tries to minimize the volume of the hypersphere.[117]

Histogram-based. This method assumes that all the features are independent from each other. For each single feature a univariate histogram is constructed where the height of the bins gives an estimation of the density. Then the score of each point is calculated by combining all the histograms using the corresponding height of the bins where the point in located.[222]

Forest-based. Whilst most of the afore-mentioned models are basically used to profile the normal labelled data, this model is focused on isolating anomalous instances. The isolation forest algorithm is recursively randomly partitioning a randomly selected feature between its minimum and maximum values, with the partitions represented as a tree structure. The number of recursive partitions required to isolate an instance is equivalent to the path length from the root node to the terminating node. The instances with short path lengths are regarded as anomalies with the anomaly score being computed by the mean anomaly score of the trees in the forest.[223]

Ensemble-based. The ensemble technique is highly suggested for one-class classification tasks. In that approach a number of base detectors is fitted to different sets of features on the dataset and identifies outliers based on the probability of each point to be an anomaly. Representative model of this category is the feature bagging algorithm.[224]

Deep One Class. In contrast to traditional approaches which make use of heuristics or statistical methods, deep learning approaches stack multiple processing layers one above another with each layer providing higher order interactions among the features. The success of deep learning is rooted in the ability of deep neural networks to learn descriptors of data with different level of abstraction without human intervention. Deep learning approaches specifically designed for one class classification are not yet very widespread. The majority of the existing models involve neural networks being trained to perform tasks other than one class classification which are then adapted for use in the one class problems. Deep networks designed for one class (anomaly detection) involve the objective function of a traditional one class approach. However, they are trained deeper *i.e.,* using more layers and in higher dimensions for fitting the appropriate function to the normal data. Deep learning models could easily handle more complex molecular representations as inputs, *e.g*., SMILES strings or 3D molecular configurations (See section 2.1.1 for molecular materials representations).[32]

5.2.2 Configuring the appropriate co-crystal datasets

The approach we followed was to build models for the class corresponding to the normal behavior and use this model to identify normal and abnormal points on the test set. For that reason, we had to construct two datasets, one extracted from CSD containing all the stable structures in which acene-like molecules can be found and one manually constructed with unknown but possible combinations of the same type of molecules. At that point we are focusing only on binary co-crystals, composed of two different molecules as they can be easier tested experimentally.

Extracting the labelled dataset. The labelled dataset of existing co-crystals in the CSD database was extracted using the CSD Python API (Application Programming Interface), version 2.0 (December 2018). As a starting point, eight molecules (Table 3.1) with extended polyaromatic systems are used as a representative set for searching the CSD and generating the co-crystal space of interest (> 1700 molecular combinations). The selection of these representative eight initial molecules is performed on the basis of promising electronic properties (*e.g.,* known organic semiconductors) and distinct geometry (*i.e.,* the set is diverse in shape and symmetry). The names of the initial molecules as well as their 6 letter CSD Refcode were: Coronene (CORONE), Picene (ZZYOC04), Pentacene (PENCEN), Triphenylene (TRIPHE), Phenanthrene (PHENAN), Fluoranthene (FLUANT), Corannulene

(CORANN01), Dinaphthol-anthracene (DNAPAN). The similarity search function of the CSD Python API is applied to those molecules, using the standard CSD fingerprint similarity search with a Tanimoto similarity threshold of > 0.35[225] and accepting only neutral organic molecules with known SMILES identifiers. The 1722 entries in the resulting list are crystal structures that include either one of these molecules or molecules that are structurally similar to them (based on CSD molecular fingerprint similarity). The search aims to identify all the co-crystals that have as co-formers PAHs whilst the main interaction between them is π-π stacking. Each co-crystal in CSD can be represented as a combination of Simplified Molecular Input Line Entry System (SMILES)[226] separated with a full stop *e.g.*, 'c1cc2ccc3cccc4ccc(c1)c2c34. N#CC(C#N)=C1C=CC(C=C1)=C(C#N)C#N representing pyrene-TCNQ'. Using this form we can count the number of different molecules in the asymmetric unit and take into consideration the molecular stoichiometry of the co-formers. Combinations including common non-aromatic solvents are excluded. However, aromatic solvents are accepted *e.g.*, benzene, as the interactions in this case are only π-π stacking and these combinations might hold important information about the predictions this work is interested in. Finally, the molecular combinations are filtered using Pipeline Pilot (version 2017)[165] by applying a SMARTS[227] filter that removes molecules with acidic hydrogens, making sure that the main interaction among the co-crystals is π-π stacking (Appendix Figures C1.2 & C1.3). The whole process for the extraction of the labelled dataset is schematically described in the Supporting Information (Appendix Figure C1.1).

Designing the unlabelled dataset. The dataset with the promising combinations of molecules is constructed using the ZINC15 database,[228] which includes all the purchasable organic molecules. The molecules were taken from a version downloaded in August 2018. The same initial molecules used for the CSD search were used and the database was searched based on molecular Extended Connectivity Fingerprints (ECFP4) with a Tanimoto similarity threshold of > 0.35[101] After filtering out the molecules with acidic hydrogens using Pipeline Pilot, the ZINC database reveals 209 molecules with calculated Dragon descriptors that match the selected similarity criteria with the initial molecules. All the possible combinations of these 209 molecules are taken into consideration, resulting in a dataset with 21736 unique pairs.

Data representation. Each molecule is represented as an *n*-dimensional vector with *n* being the number of the available descriptors calculated with Dragon software,[229] version 6.0/2012. Although the deep one class approach doesn't require any manual feature engineering, for the traditional one-class classification approaches it is desirable to reduce the dimensions of the problem before the analysis. The dimensionality reduction is performed following the standard good practices for removing descriptors that are highly correlated to each other or describe similar properties.[230] Features that are correlated more than 0.92 as well as those that have variance lower than 0.4 were removed from each co-former's dataset. The feature selection process was performed according to the molecular complementarity approach.[51] All the pairwise correlations between the molecular pairs were calculated, after removing co-crystals contain benzene-like solvents to avoid possible bias on the feature importance. The pairwise correlations were calculated with both Pearson and Spearman methods[51] and the p-values were used to verify that the correlations are statistically significant. We regard as important and unbiased features those with both Pearson's and Spearman's correlations above 0.4 and p-values below 10^{-3}. Finally, each single molecule is represented by a 24-dimensional space of the highly pairwise-correlated descriptors (Table C1.2). Thus, the molecular pairs are the concatenation of the individual vectors of each single molecule. As each molecular pair is order invariant, we need to find a way to denote the combination of molecules. Consequently, the training was performed using both orders (a,b) and (b,a) for each molecular combination. All the labelled molecules were standardized to [0,1] using the scaling methods provided from sci-kit learn, such that all the numerical features will belong to the same range. The scaler is fitted to the known molecules that form co-crystals. Then the trained scaler is implemented to transform each molecule in the molecular pairs in both the labelled and the unlabelled datasets, such that there will be a consistency among them and the same molecules will get the same representation independent of which pair they belong to.

5.2.3 Designing the One Class Models

Traditional one class classification. Eight different algorithms were selected from the PyOD and sklearn library representing the wide range of the one-class classification (anomaly detection) categories as described above: Gaussian Mixture Models (GMM), Local Outlier Factor (LOF), k-nearest neighbors (kNN), Isolation Forest (Iforest), One Class SVM (OCSVM), Histogram Based Outlier Score (HBOS), Cluster-based Local Outlier Factor (CBLOF) and Feature Bagging.[231] Each algorithm has its internal scoring function, depending on the cost function it tries to minimize. For achieving better predictive performance and ensuring the robustness of our method the models were combined in an ensemble way. For consistency with the GMM model from the scikit-learn library,[232] the scores from the PyOD library were multiplied by -1 to have higher scores for the inliers and lower for the outliers. Each model was initially trained and optimized separately to provide an anomaly score to the input data. Then the scores of the pretrained models were normalised between [0,1] and averaged, following the methodology from the combo library[233] so that the outputs become comparable.

Deep Learning Approach. Using the traditional one class classification algorithms as baselines, the application of a deep learning method was investigated for extending the dataset to the whole *n*-dimensional space ($n = 3714$, *i.e.*, 1857 descriptors for each molecule in the pair). In that way the predictions are not only dependent on the selected pairwise correlated descriptors. That is very important as the co-crystal design problem is complex and thus higher-order interacting features might have a key role in the co-crystal formation. The main advantage of using a neural network in this context is that the extensive feature engineering part can be omitted, as the network can learn relevant features automatically. The most broadly used deep learning approaches for one class classification rely mainly on Autoencoders. An Autoencoder is a neural network that learns a representation of the input data by trying to accurately reconstruct the input with minimum error. It is considered to be an effective measure for separating inlier and outlier points.[234] Autoencoders are used for learning the representation of the labelled data and then the unlabelled data are reconstructed using the same weights from the target class. The decision of whether a new datapoint is an inlier or an outlier is made based on the reconstruction error. High reconstruction error indicates that a sample is most probably an outlier, whereas when we have low reconstruction error the samples most probably belong to the same distribution as the labelled data. Autoencoders have the objective of minimizing the reconstruction error, but do not target one class classification directly. For designing a more compact methodology, the adapted approach incorporates both an Autoencoder for representational learning which is jointly trained with a Feed Forward Network targeting one-class classification.

Deep One Class Architecture. The Deep Support Vector Data Description (DeepSVDD) architecture used in this paper is adapted from the work of Ruff *et al.*.[217,235] The aim of DeepSVDD is to find a data-enclosing hypersphere of smaller size, such that the majority of the normal data will be found there, whereas the anomalous data will be outside. The objective of DeepSVDD is to jointly learn the network parameters together with minimizing the volume of the hypersphere. Using these settings, we expect the normal data points to get mapped near the hypersphere center whereas anomalous data are mapped further away. The hypersphere center is calculated with a pretraining step and is fixed as the mean of the network representations of the known data.[217] Each pair of molecules is scored based on its distance from the center of the hypersphere. The DeepSVDD network consists of a Convolutional Autoencoder, where the output of the Encoder is connected with a Feed Forward Neural Network with the specific task of minimizing the loss function (distance from the center of the hypersphere). The same pretraining and training steps as in the DeepSVDD method were used for our known dataset, whereas the Convolutional Autoencoder was substituted with the Set Transformer Autoencoder adapted from Lee *et al.*.[236] The implemented set-input architecture uses a self-attention mechanism that allows the encoding of higher-order interactions among pairs. We use a batch size of 200 and set the weight decay hyperparameter to $\lambda = 10^{-6}$. All the known data are considered to belong to the hypersphere and they are scored based on their distance from the center, thus the lower the score the closer to the center and the more of an inlier is the data point. Likewise, the unlabelled data are assigned scores based on their distance from the pre-defined center. All the scores are multiplied by -1 and normalized from 0 to 1 so that they are comparable to the other models and give scores close to 1 for the inliers, whereas the points scored close to 0 are the anomalies.

Hyperparameter tuning. As the performance of the algorithms is highly dependent on the choice of the hyperparameters*, i.e.*, algorithm variables, the optimization step is crucial for achieving the highest possible accuracy. For the machine learning models, the optimization step is about searching for the hyperparameters with the lowest validation loss. Bayesian optimization was used via the Hyperopt library.[237] The main idea behind Hyperopt is to get more points from the regions with high probability of yielding good results and less points from elsewhere. Hyperopt library was implemented for each of the eight algorithms from the PyOD&scikitlearn library,[231,232] to find the best set of parameters to maximize the average accuracy of the k-fold cross-validation.

Model evaluation. The evaluation of the classification performance for one-class classifiers differs from multi-class classification as only the probability density of the positive class is known. That means that the model can only be optimized and validated by minimizing the number of positive class instances that are not accepted by the one-class classifier (false negatives).[117] Opposed to the binary classifiers, where the decision of the class is made based on a set threshold, usually 0.5 (if a point scores below 0.5 it belongs to the first class else to the second), in one class classification the threshold is defined only from the known class. That is set using a parameter (here refered as contamination), which defines the amount of noise we expect to have in our known class. Herein, we accept that parameter as 0.05, meaning that 95% of the known data are inliers and only a very small part of them that deviated from the rest can be regarded as outliers. The evaluation of the models was performed using five-fold cross validation on the labelled dataset. The labelled dataset is split into five parts (folds) where 4/5 are used for the training and the remaining part is used for the validation. The process is repeated five times, each time selecting a different fold and the evaluation is performed using accuracy metrics from version 0.22 of the scikit-learn package. The final accuracy is calculated by taking the mean of the five accuracy scores of the validation set.

5.2.4 Electronic structure calculations

Calculations of the orbital energies of the single molecule were carried out at the B3LYP/6-31g* level of theory using the SPARTAN'18 software package (Spartan, Wave Function Inc. CA). The electronic structure analysis of the co-crystal was performed with plane-wave-based DFT calculations using the VASP programme.[238] The SCAN+rVV10 functional was used to improve the description of van der Waals interactions over other semilocal DFT functionals, with a plane-wave cutoff energy of 600 eV. The KSPACING parameter, the functional for van der Waals corrections as well as the cut-off energy were selected after convergence check on a known co-crystal system as shown in Appendix C3 Figure C3.1, Table C3.1.

5.3 Results

5.3.1 Models evaluation and comparison

The two different one class classification workflows followed involve *i*) the application of traditional algorithms designed for one class classification after extensive feature engineering to reduce the dimensionality of the problem and *ii*) the design of a deep learning methodology for handling the specific co-crystals dataset, considering them as pairs of data, and avoiding feature engineering by solving the problem in higher dimensions. As traditional algorithms we are referring to the provided algorithms from PyOD/scikit-learn libraries and as Deep One Class to the deep learning model that was built by combining an Attention-based Encoder and deepSVDD network. In both workflows a two-step process is employed. First the algorithms were trained and optimized on the known data and then they were used for scoring both the labelled and unlabelled molecular combinations. High scores are an indication for inliers, whereas the lower the score the higher the probability for a point to be an outlier.

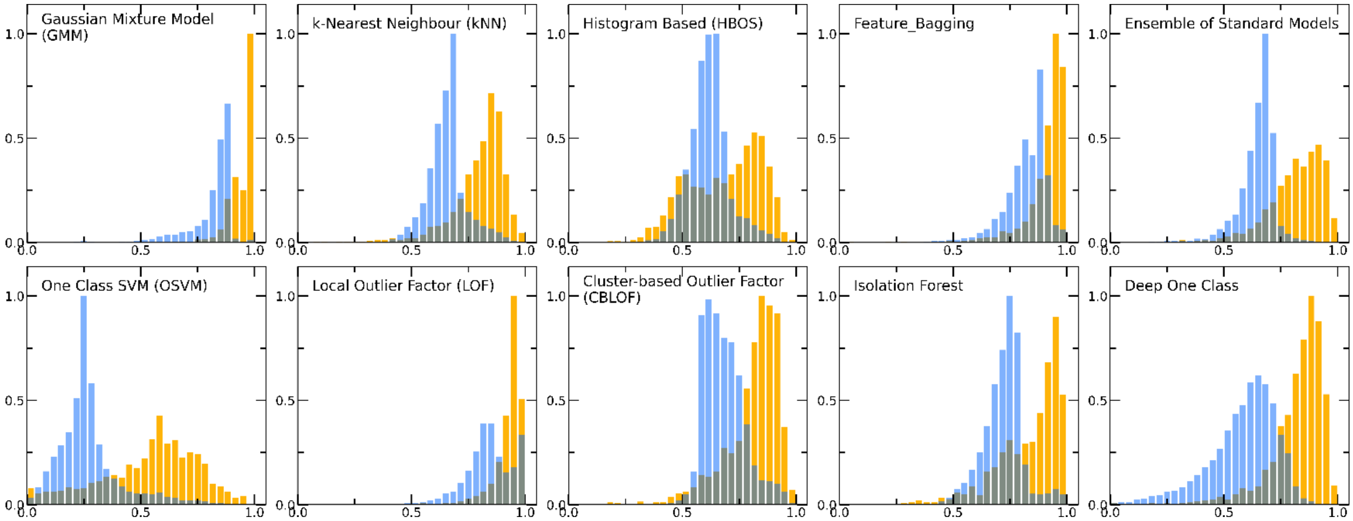


Figure 5.2. *Score distributions of the labelled (orange) and unlabelled (light blue) data using all the discussed one-class classification algorithms. Each algorithm employs a different scoring function to assign scores to the molecular combinations, giving in all the cases higher scores to the labelled combinations (training set) whereas only a certain part of the unlabelled combinations (test set) receives high scores and can be regarded as inliers. As the number of unlabelled data is significantly higher than the number of known data, the y axis (showing the frequency) is normalized to [0,1] (for visualization purposes). The output scores of all the models are also normalized to [0,1].*

Reproduced from Ref. 253 with permission from the Royal Society of Chemistry.

The score distribution of both the labelled and unlabelled data for all the implemented algorithms is presented in Figure 5.2. It can be observed that the labelled and unlabelled data form two overlapping classes. The unlabelled data consist of both positive and negative examples in an unknown proportion. Consequently, a certain part of the unlabelled data is expected to belong to the known class *i.e.*, are inliers. Moreover, in the labelled data there is a small proportion of examples that significantly differ from the rest of the data and is regarded as noise of the normal class, *i.e.*, outlier examples. The impact of the class noise is mitigated using one class classification, as a percentage of the labelled data are regarded as outliers during the hyperparameter optimization process (see Methods). In general, for both the traditional and deep one class classification workflows, *i*) the labelled data show higher scores with all the methods, *ii*) each method has a different way of scoring the samples and deciding for whether a point is a normality or anomaly and *iii*) only a certain part of the unlabelled data receives high scores. Differences arise between the algorithms because each is based on different definitions on what an oulier/inlier means, *i.e.*, an outlier is a point far from other points (kNN), an easily splittable point (Isolation forest), not part of a large cluster (Cluster-based outlier detection) or a point far away from the center of a hypersphere (deepSVDD). Moreover, the traditional approaches differ from the deep approach in terms of the dimensionality of the features and the way the molecular pairs are perceived by the models. For achieving more reliable and robust predictions, the eight traditional one class classification algorithms were combined in an ensemble way by averaging their output. Thus the final scores of both the labelled and unlabelled data were calculated by the ensemble. The distribution of the ensemble scores, after being normalized to [0,1], are shown in Figure 5.2. It is observed that the ensemble separates better the labelled from the unlabelled data in comparison to the individual traditional algorithms. That is an indication that the ensemble is a better classifier as the balance point above which the amount of labelled data is maximum and the number of unlabelled data is minimum is easier found.[239]

The performance of each algorithm was calculated by the True Positive Rate (TPR), meaning the average of correctly predicted inliers resulting from five-fold cross validation. As illustrated in Figure 5.3, all the algorithms achieve a high accuracy on the True Positive Rate and perform quite well on unseen data. However, the Gaussian Mixture Model (GMM) and the Histogram-based Model (HBOS) are less robust as indicated by the higher variation in the total accuracy. The effect that the addition of data in the training set has on the accuracy is also investigated after calculating the learning curves of each algorithm. For the correct sampling of the bidirectional dataset in the different training set sizes, it should be ensured that equivalent pairs exist in each subset.

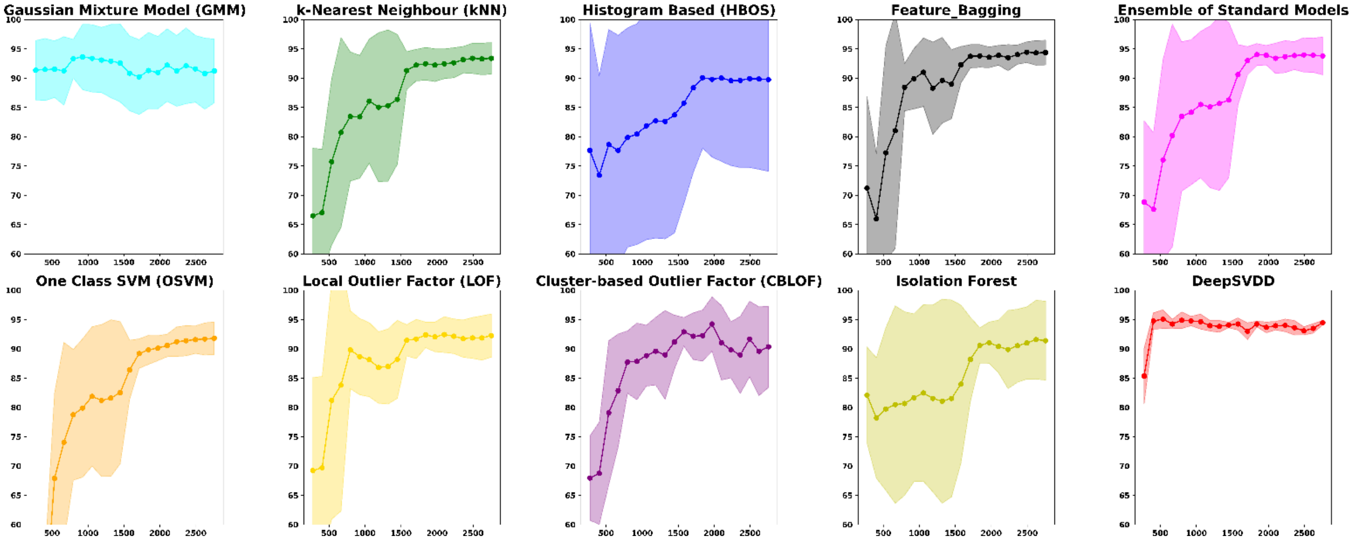


Figure 5.3. *Learning curves of all the implemented algorithms showing the performance of the models while the size of the training set increases. The highlighted grey area represents the standard deviation of each model. The validation metric used is the True Positive Rate (TPR), i.e., number of correctly predicted inliers/total size of the training set in each fold of the k-fold (k = 5) cross validation. It is observed that the Deep learning model (DeepSVDD) outperforms the traditional algorithms as it has higher accuracy and low standard deviation.*

The two workflows are also compared scores-wise (Figure 5.4). It can be seen that there is a good agreement (correlation) in high scores, whereas in the lower scores area there is not a clear correlation as the ensemble method gives a narrower range of scores and higher scores for low-scoring examples in the Deep case.

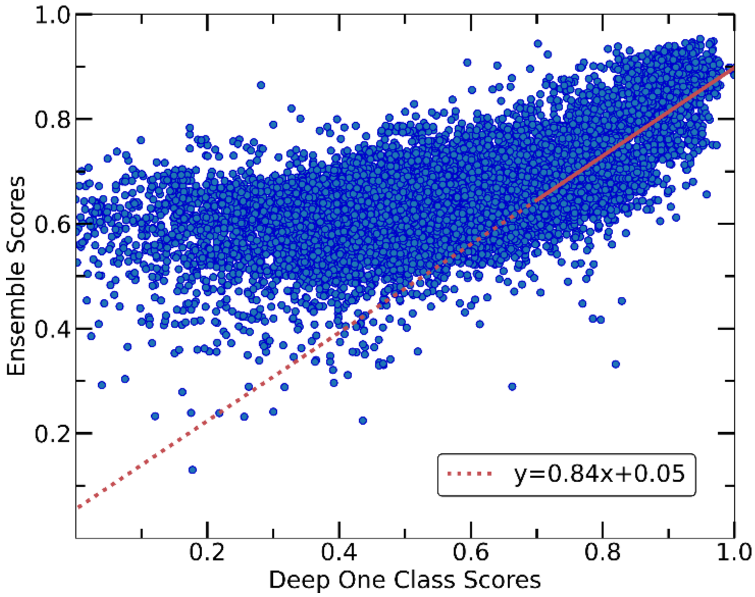


Figure 5.4. *Correlation between the scores of the Ensemble and Deep One Class methods. Both workflows show a good correlation in the general distribution of scores, with Deep One Class covering a wider range of scores and enabling in that way a better separation between inliers and outliers. A significant correlation exists for the high score pairs, showing that both methods could be reliable in the high-score region.*

Reproduced from Ref. 253 with permission from the Royal Society of Chemistry.

In every classification problem, a threshold should be specified above which the datapoints that belong to the normal class can be found. We set that threshold at 0.7 and thus all the molecular pairs with scores higher than 0.7 are regarded as reliable inliers with a high probability to exist. That threshold was selected based on the good agreement between both workflows for scores above 0.7. Moreover, it is a good balance point as the majority of the labelled data receive scores above that threshold whereas only the top quartile of the unlabelled data can be found in that area. In cases were a better separability is achieved,[240] the amount of misclassified data (FP: False Positives) is minimized significantly, thus the selection of the threshold (on 0.7) could be regarded a reasonable decision boundary.

5.3.2 Understanding the predictions

The reliability of any machine learning model is improved when the models' decisions are related to physical properties. Following the traditional one class classification workflow, the features associated with the final predictions are already known after the extensive feature engineering process. On the other hand, an understanding about the features that played a key role in the deep learning approach is a more challenging task, as the complexity of the model is higher. To better understand the features that are important for the neural network categorization of the molecular pairs in one class, we used SHAP (SHapley Additive exPlanations).[131] This interpretability method is based on the calculation of the game theoretically optimal Shapley values, which are indicative of the contribution of each feature to the final prediction. To this end, features that play a key role in the scoring for the deep learning approach are retrieved and analysed. The aim of this process is to identify molecular properties or characteristics that might provide a chemical understanding to the models' decisions and assist the experimental screening process. As for many of the Dragon descriptors it is hard to extract a physical meaning, the correlations among the most significant descriptors with those that are more general and understandable are calculated .

According to Shapley analysis, the most important features that the inliers have in common and dominate the decisions are related to the descriptors B06[C-C], ATS6i, B08[C-C], ChiA_Dz(p), Eig06_AEA(dm) and SpMin5_Bh(s). The physical meaning of these descriptors is extracted after calculating the correlations between them and the other Dragon descriptors, that are higher than 75% (Table C2.1). Interestingly, except for the B06[C-C] and B08[C-C], which are related to the topological distance between two carbon atoms, *i.e.,* the presence of connected carbon atoms at specific positions on a molecular graph, the other descriptors are highly correlated with easily understandable molecular properties. These chemically meaningful descriptors refer to *i*) electronic properties, such as the sum of first ionization potentials (Si), sum of atomic Sanderson electronegativities (Se), sum of atomic polarizabilities (Sp) *ii*) molecular size, such as McGowan volume (Vx), sum of atomic van der Waals volumes (Sv), *iii*) molecular shape, regarding the molecular branching (Ram, eta_B), *iv*) polarity (Pol, SAtot) and *v*) molecular weight (MW).

The relationship among some of the important interpretable descriptors in the molecular pairs is illustrated in Figure 5.5 for both the labelled and the unlabelled datasets. The distribution of the property values in the high scoring pairs (inliers) in the unlabelled dataset (Figure 5.5b) are predicted to follow the same patterns at the labelled dataset indicating that the deep learning model effectively learnt the trends of the labelled dataset and was able to score the unlabelled dataset based on those trends. The dominating trends on the labelled dataset can be observed with darker orange colour indicating the densest area with more molecular combinations. Two main areas are extracted from the labelled dataset. The first area includes molecular pairs where both molecules have low values of the same property, *e.g.*, in the Polarity plot the area $0<Pol<60$, where both molecules could have similar values. The second area includes molecular pairs with higher difference on their property values, *i.e.*, when one molecule has a low value of one descriptor then the pairing molecule has a higher value for the same descriptor, complementing the first molecule. These observations are also compared with a previous study by Fabian that focused on the CSD co-crystal dataset.[51] Fabian's statistical analysis of the data at that time concluded that the majority of co-crystals in CSD (CSD, version 5.29, November 2007) are formed by molecules of similar size and polarity.[51] Our analysis shows a more complex scenario. Size, shape and polarity, identified as important factors of co-crystallization, have similar property values only in the low value region, in agreement with Fabian's conclusions. However, in the high value regions the trend drastically changes; molecules having high size, shape and polarity values tend to pair with molecules having low values of these parameters.

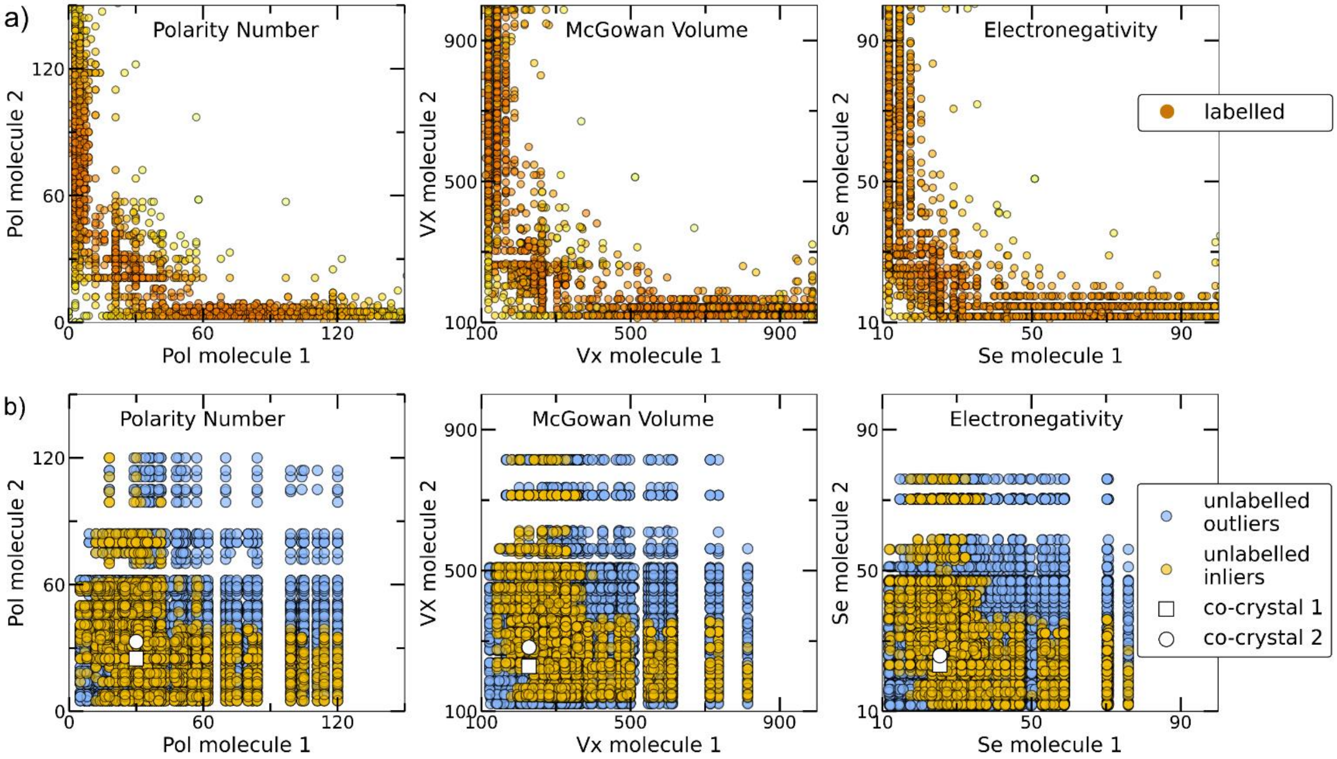


Figure 5.5. *a) Scatterplots showing the distribution of representative descriptors among the molecular pairs on the labelled dataset. The plotted descriptors are those identified as the most general and highly correlated to the descriptors extracted using the Shapley analysis. b) The distribution of the same descriptors for the unlabelled data. Blue circles represent the whole unlabelled dataset extracted from ZINC15 (21736 points) and yellow-orange represent the top quartile of the unlabelled data having scores above 0.7 and are regarded as inliers. It can be clearly seen that the predicted inliers follow the distribution of the labelled dataset, especially in the densest area.*

The dominating features as expressed with global Shapley values can give a general picture of the dataset. However, it should be noted that a better understanding for specific groups of pairs that might be of interest can be attained when focusing on them explicitly. The advantage of using Shapley values is that local explanations are given to each individual molecular pair or to a subset of interest among the molecular pairs. As a case study, the pyrene-cocrystal family is investigated, aiming to extract some general patterns about the important molecular characteristics that drive a good match for co-crystal formation with pyrene. The dominating features in the known co-crystals with pyrene are presented in Figure 5.6. It was found that the existence of heteroatoms such as oxygen and/or nitrogen groups on various topological distances, as indicated by the B03[C-O], B02[C-O], B02[C-N] and

B05[C-N] descriptors or the existence of halogen atoms as indicated by the X% descriptor play a key role in the assignment of high scores in these combinations. Furthermore, the aromaticity as represented by the ARR descriptor was a factor that contributed to high scores.

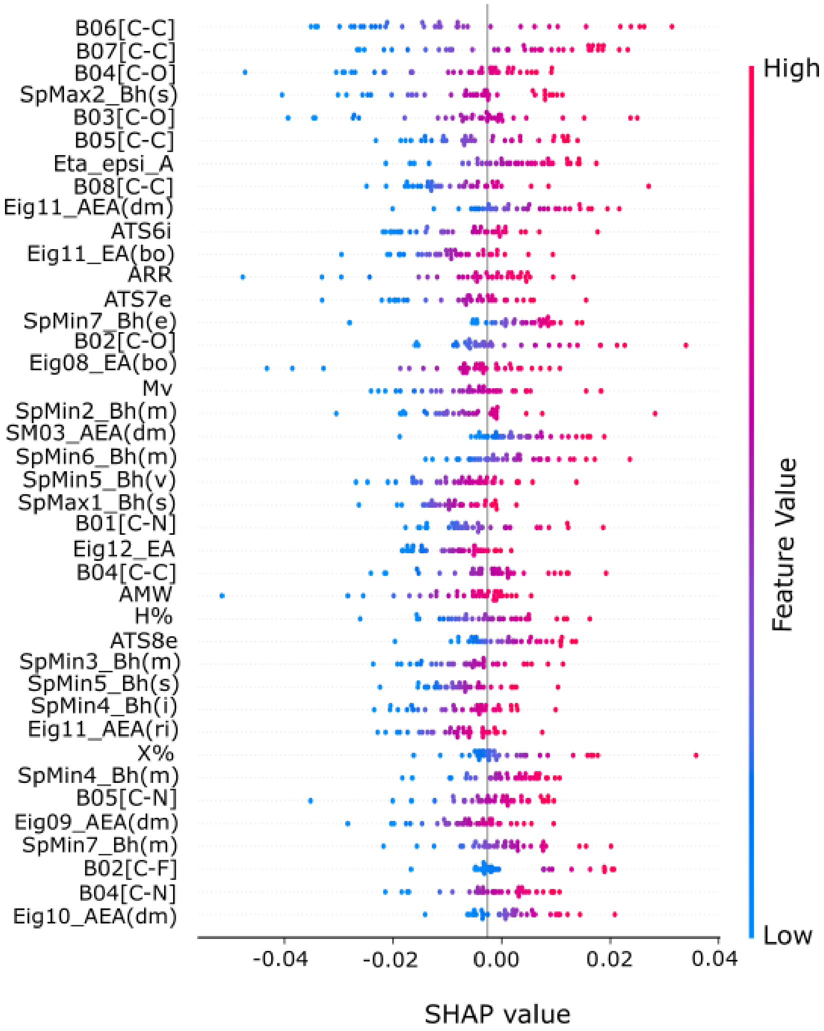


Figure 5.6. *Shapley values showing the important descriptors that molecules pairing with Pyrene in the labelled dataset have. Only the contributions of the second co-formers are shown here. The presence of heteroatoms in several topological distances in the molecule are those that seem to contribute more. The notable elements are N and O. It is expected that molecules with these groups in the certain topological distances and high scores are good candidates for forming co-crystals with pyrene.*

The key findings from the model interpretation and feature analysis can be summarized below:

i) Shape, Size and Polarity were detected as important factors for co-crystallization, which is in accordance with previous understanding about the co-crystals of CSD. However, Fabian's observations are relevant only for low values of these properties. We observe that there are no cases in the labelled data and in the inlier part of the unlabelled dataset where both molecules have very high values of polarity and/or volume. This could be an indication for factors prohibiting co-crystallization. In cases, where high polarity or volume values are assigned to one molecule the pairing molecule usually has a low value of that property.

ii) PAH co-crystals seem to deviate from empirically established rules and trends observed for organic co-crystals in general. Thus, a deeper understanding of their properties can only be gained when they are studied separately. As PAHs lack hydrogen bonding, other types of interactions appear as stabilizing factors for co-crystallization. For instance, in the pyrene-based co-crystals the existence of O and/or N groups has been identified as a key parameter as the majority of molecules that form co-crystals with pyrene contain these groups. The existence of these groups can drive the formation of C-H···N, C-H···O and C-H···X (X= halogen groups) which will probably stabilize the co-crystal formation.

iii) There is not a 'magic' descriptor or set of some descriptors that can directly predict co-crystallization. The synergy among many descriptors will led to a successful combination. The more parameters, and the more the relationships among them, that are taken into consideration, the more reliable the predictions and the more accurate the results we can attain. For instance, it is not enough that a molecule in a pair has a polar group (*e.g.*, the -CN group), as many other driving forces (*i.e.*, significant descriptors) should be in line to get a successful molecular combination. This is the reason the implementation of the appropriate ML tools could save significant amount of time and guide the synthetic work, as this is the only way where the relationship among a large number of properties is simultaneously considered. As seen in Figure 5.5 (a and b), the current model is able to extract the descriptors' trends from the labelled dataset and learn the dominating patterns. In this way the model gives a score and suggests molecular pairs that look feasible based on the known co-crystals.

5.3.3 Molecular Ratios Prediction

Herein, we showcase that the representation learned for the Attention-based autoencoder can be effectively used for predicting the stoichiometry of the molecules in the co-crystal. An important parameter that should be taken into consideration in co-crystals design is the stoichiometry of the co-formers. The molecular ratio is going to affect the crystal packing and thus contribute to possible materials properties. To this end, the labelled co-crystals dataset was further tested for molecular ratio prediction. The molecular ratio of all the combinations was extracted during the labelled dataset construction (See Methods). The dominating ratio in the dataset is 1:1 as shown in Figure 5.7, resulting in a highly biased dataset towards the molecular ratios. The problem setting was adjusted for performing binary classification and investigated whether the molecular ratio is going to be 1:1 or higher. We assigned label '0' to all the molecular pairs having 1:1 ratio and '1' otherwise. The problem was solved using SMOTE technique for balancing the two classes of the dataset such that they have equivalent amount of data having 1:1 ratios and data having ratios different to 1:1.

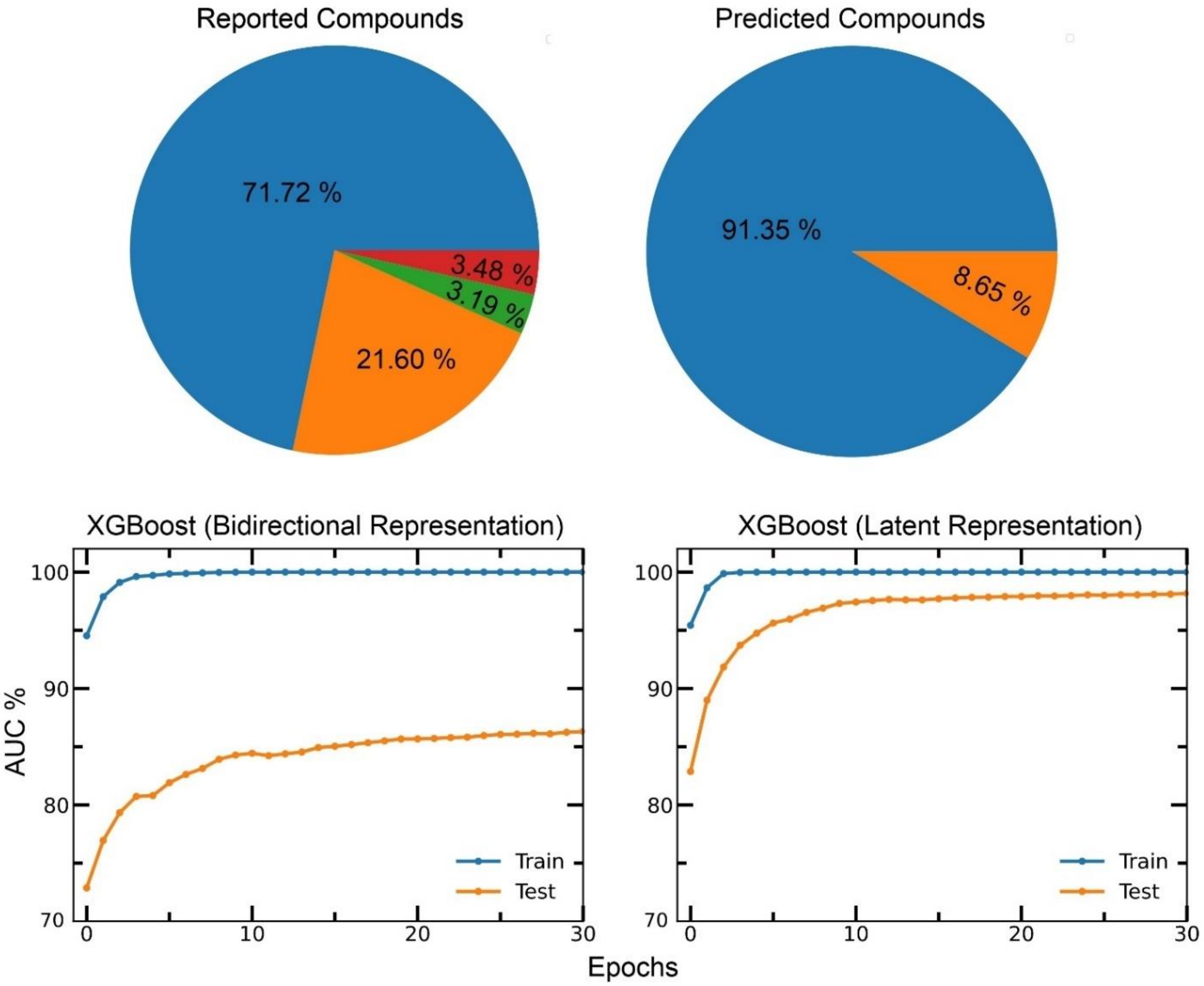


Figure 5.7. *Piecharts illustrating the molecular stoichiometry on the reported (left, labelled dataset) and on the predicted (right, inliers) compounds of the co-crystal dataset. The blue area represents the 1:1,, the orange area the 1:2, the green area the 1:3 and the red area the 1:4 molecular pairs stoichiometry. It can be observed that the dominating ratio is 1:1, resulting in a highly imbalanced dataset towards molecular ratios. Significant improvement when using the Set Transformer latent representation.*

Reproduced from Ref. 253 with permission from the Royal Society of Chemistry.

The labelled dataset was split into a training and a test set with the latent representation being the input to a binary classifier. The model showed strong predictive power, with accuracy on both the training and test sets of about 94 % and no overfitting on the training data (Figure 5.7). The same model was then implemented for predicting the molecular ratios in the inlier pairs.

5.3.4 Pareto front optimization on the predictions

To narrow down the selection of potential co-formers from those identified using the single class classifier model, we chose pyrene as a fixed component because both the existing data (i.e., CSD database) and the model output reveal its popularity and versatility as a co-former, i.e., pyrene can co-crystalize with a diverse range of molecules forming high score pairs. In total, 207 possible pyrene-containing co-crystals were identified by the single class classifier model which were narrowed down to a subset of 29 pairs where the second co-former has zero examples of known co-crystals with any other molecule (blue points in Figure 5.8). Pareto optimization was used to identify the most suitable candidate co-formers for experimental investigation. Pareto optimization simultaneously identifies the optimal values in a set of parameters and was used to select and prioritise the co-formers to be experimentally tested. In our case the parameters that were optimised are the score from the model and the similarity to 7,7,8,8-Tetracyanoquinodimethane (TCNQ). This two-parameter optimization was implemented to drive the decision making for the experimental screening. From the Pareto front (Figure 5.8 green line) 1-4 are identified as the optimal candidates and 5 is the highest scoring co-former off the Pareto front.

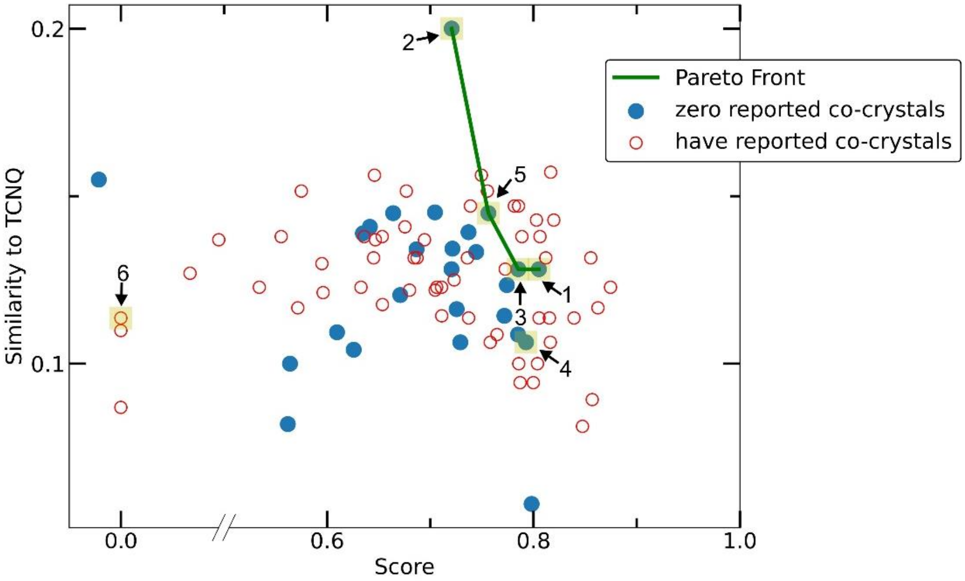


__Figure 5.8.__ Scatterplot illustrating the selection criteria for the experimental screening process. Pareto optimization was implemented having as the main task the optimization of two objectives, i) the score of the deep learning model and ii) the Tanimoto similarity to TCNQ. Each point represents a molecule that could be used as the second co-former in pyrene co-crystals. Red empty circles stand for molecules that are already known to form co-crystals in the CSD, whereas molecules represented with filled blue circles have zero reported co-crystals. The molecules selected and experimentally tested are highlighted in green circles.

Reproduced from Ref. 253 with permission from the Royal Society of Chemistry.

The experimental realization of the one class classification approach led to the synthesis of two novel co-crystals, namely pyrene-6H-benzo[c]chromen-6-one (**1**) and pyrene-9,10-dicyanoanthracene (**2**).

5.3.5 Electronic structure of the experimentally validated predictions

So far, only the structural similarity to TCNQ on the molecular level was taken into consideration to drive the selection of the coformers. In this section, the electronic characteristics of the two co-formers (1) and (2) are also calculated and directly compared to TCNQ. Moreover, the relationship between the crystal structures and the density of states for the two new co-crystals were investigated and compared with a TCNQ-based semiconducting co-crystal.

7,7,8,8-Tetracyanoquinodimethane (TCNQ), has been extensively studied for its interesting electronic properties both in the crystalline form and as a co-crystal.[212,216,241–247] TCNQ is one of the most widely used electron acceptors in organic electronics having four strong electron-accepting cyano groups (-C≡N).[248]

Resembling TCNQ, 9,10-dicyanoanthracene has two strong electron-accepting cyano groups. On the other hand, 6H-benzo[c]chromen-6-one has a lactone (cyclic carboxylic ester) motif which has been reported as an electron deficient building block used into conjugated polymers.[249]

Both the cyano (-CN) and ester (-COOR) groups are strong electron withdrawing groups (EWG) that reduce the electron density in a molecule through the carbon atom it is bonded to. In all three molecules the lowest-energy empty molecular orbital (LUMO) is delocalized onto several atoms and as shown in Figure 5.9 it is more concentrated on the carbons than on the nitrogen or oxygen atoms. All three molecules, namely TCNQ, (1) and (2) have low-lying LUMO levels complying with their electron acceptor nature.

Following similar trend as for the structural similarity, 9,10-dicyanoanthracene is electronically more similar to TCNQ, as both a have lower LUMO energy values and in higher HOMO orbital values in comparison to 6H-benzo[c]chromen-6-one. In this regard, 9,10-dicyanoanthracene could be a strong candidate for co-crystals of electronic interest, although its HOMO-LUMO gap is larger by 0.5 eV from TCNQ. Consequently, both (1) and (2) can be regarded as Donor-Acceptor (DA) co-crystals with pyrene playing the role of a π-electron rich donor and the two molecules containing the lactone and cyano groups, respectively, serving as the electron acceptors.

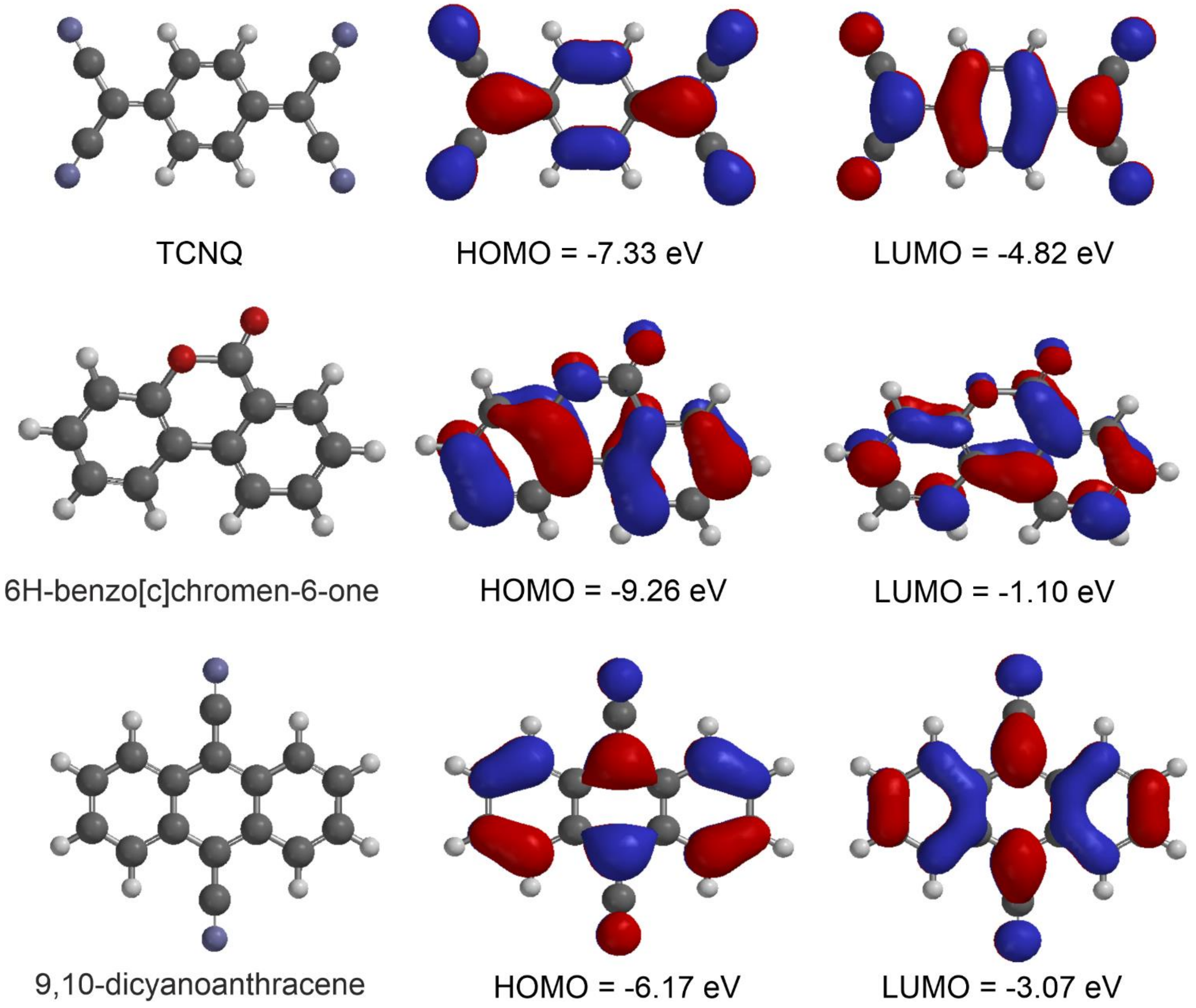


Figure 5.9. *The shapes of the HOMO and LUMO orbitals of TCNQ (top),* 6H-benzo[c]chromen-6-one *(middle) and* 9,10-dicyanoanthracene *(lower). The blue and red regions correspond to positive and negative values of the orbital (the absolute sign is arbitrary).*

As the final electronic properties of the materials are defined from their 3D conformation, the electronic structure analysis of the structure was performed using Density Functional Theory. DA co-crystals with 1:1 D:A stoichiometric ratios generally feature the D and A molecules packed in segregated- (*i.e.*, a column of D molecules, ⋯D−D−D−D⋯, aligned next to a column of A molecules, ⋯A−A−A−A⋯) or mixed-stack (*i.e.*, a column of D and A molecules stacked in a regular ⋯D−A−D−A⋯ pattern) arrangements. Cocrystals formed by varying the D:A stoichiometric ratios, for example, 2:1 and 3:1, have also been created and studied.[250] The two synthesized co-crystals have different stoichiometries and molecular arrangement, as shown in Figure 5.10. (1) was correctly predicted from the one class classifier having a 1:2 ratio. D and A molecules are packed in segregated-stack, *i.e.*, a column of D molecules, -D−D−D−D-, aligned next to a column of A molecules, -A−A−A−A-. On the other hand, (2) was correctly predicted to afford a 1:1 stoichiometry and it was experimentally found to form a mixed-stack arrangement, *i.e.*, D-A-D-A pattern.

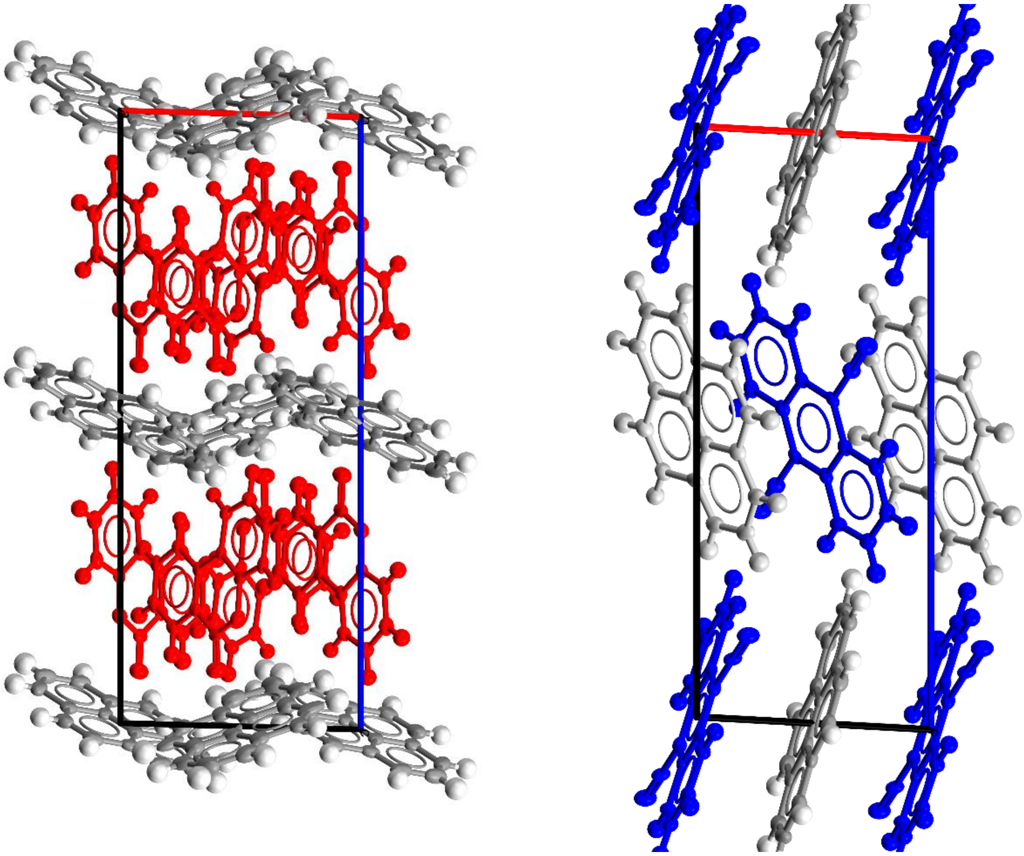


Figure 5.10. *Crystal structures of (1) (left) and (2) (right), viewed along the b axis. The electron acceptor molecules are coloured in red (*6H-benzo[c]chromen-6-one*) and blue (*9,10-dicyanoanthracene*).*

The density of states of the two synthesized co-crystals (1) and (2) was compared with single pyrene crystal and the pyrene-TCNQ co-crystal as shown in Figure 5.11. From the partial density of states (PDOS) of (1) can be observed that both pyrene and 6H-benzo[c]chromen-6-one contribute to the conduction band and the valence band. On the other hand, from the PDOS of (2) we can observe that in the conduction band 9,10-dicyanoanthracene is predominant (based on the carbon and nitrogen content), whereas the valence band has significantly higher pyrene character. A similar behaviour is found in the pyrene-TCNQ complex, where TCNQ contributes to the valence and pyrene to the valence band. Noteworthy, it was found that (2) has a small bandgap of 1.49 eV which belongs to the region of currently known organic semiconductors and is comparable to the pyrene-TCNQ co-crystal.[251,252] As discussed in Chapter 3 the electronic structure is correlated with the structural characteristics thus by searching for molecules that are structurally similar to an electronically interesting compound can lead in the formation of semiconducting materials with important electronic properties.

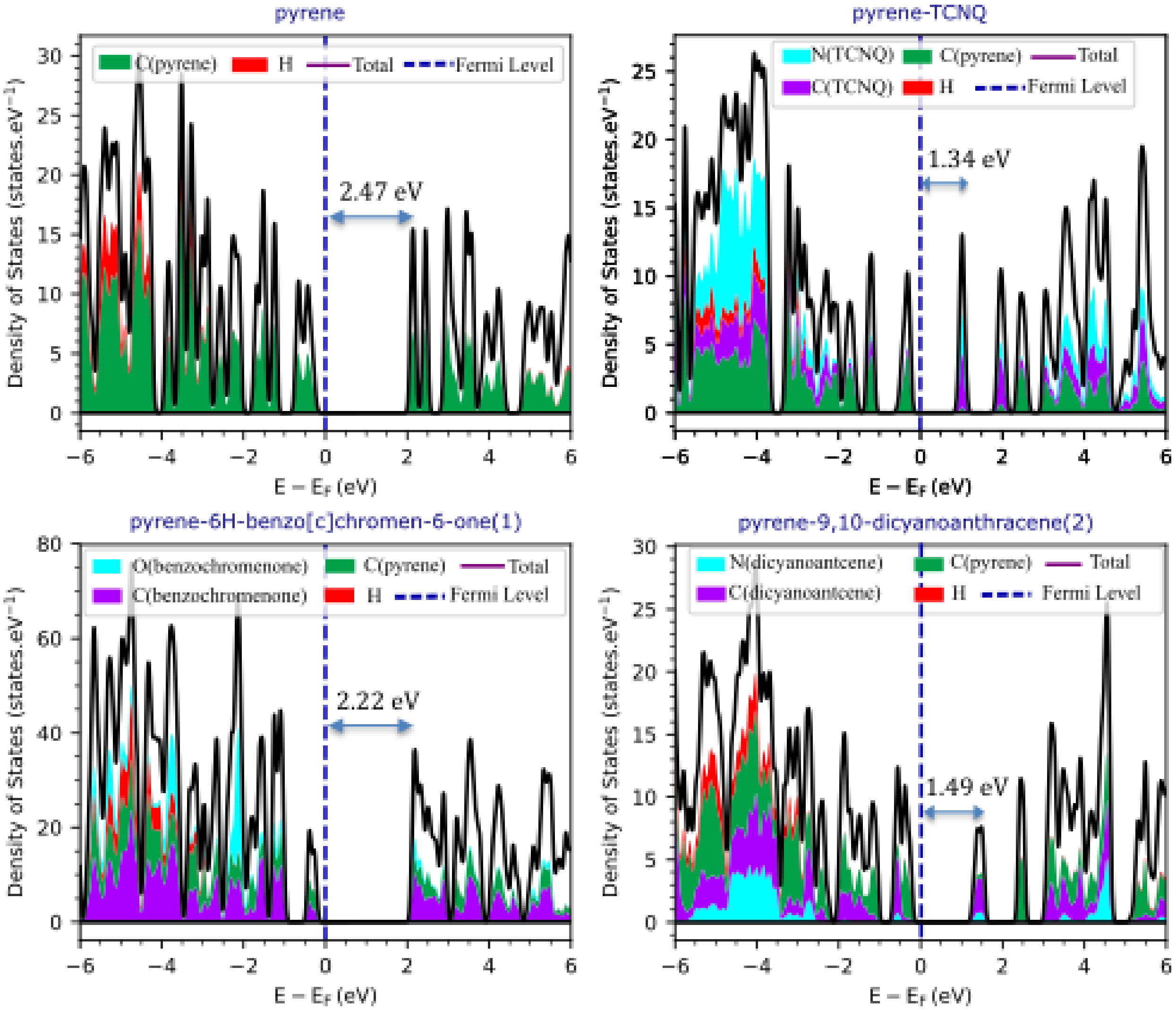


Figure 5.11. *Density of states and bandgap of a) pyrene single crystal (PYRENE) b) pyrene-TCNQ (PYRCBZ04) known semiconductor, c) EHUFIZ(1) and d) EHUFEV(2).*

5.4 Discussion

A machine learning tool is developed which is able to extract the patterns from only positive known co-crystal data and rank novel molecular pairs based on their probability to form new crystalline materials. The major research challenge that sparked this work is the lack of densely and uniformly sampled data in materials chemistry, which result in inherently imbalanced datasets and unreliable negative counterexamples. The existing databases constructed of published literature typically only include positive results, with scientists very rarely publishing such clear details of experiments that did not work. From a machine learning perspective, this means that only one class (*i.e.*, the positive outcome) is well defined by the data.

The implementation of one class classification as a methodology for dealing with the ‘only positive data’ challenge was highlighted. We report as a case study the prediction of new molecules which have not previously been recognised as co-formers in the unique and limited class of materials, the π-π interconnected co-crystals. In the attempt to improve our understanding about one class classification, a broad overview of the current methods and concepts is given. The problem is initially investigated using traditional one class classification algorithms in lower dimensions after extensive feature engineering. Further on, we demonstrate that by using a Deep One Class approach, the manual feature engineering could be avoided, and we can not only achieve higher accuracy, but also the incorporation of more feature interactions among the co-formers. In this way, all the features that might lead to the formation of stable co- crystals are taken into consideration and the relationships among them are extracted. Co-crystallization emerges as a difficult task for both computational predictions and experimental screening, particularly for cases of limited strong directional forces that could give a strong indication for a successful outcome. In our contribution, we show that the implementation of the appropriate data mining strategy combined with the extraction of a reliable dataset can leverage the synthetic attempts and lead to the successful discovery of new materials. Moreover, an in-depth understanding of the machine learning model with a rationale about the predictions is sought after for advancing our knowledge on the chemical factors that favour co-crystal formation. Currently, many steps towards explainability of machine learning models have been made. Therefore, for a computational strategy to be reliable it is important to incorporate interpretability for rationalizing the predictions. SHAP calculations were carried out for interpreting the scoring of the deep learning model by assigning feature weights. Consequently, a better understanding of the features that dominate the known molecular pairs is gained and meaningful information regarding the characteristics of the molecules that can relate to π-π stacking is extracted. Shape, size and polarity were detected as important factors for co-crystallization, which is in accordance with previous understanding about the co- crystals of CSD. However, our analysis reveals a more complex scenario, where co-crystallization is feasible for molecules having similarly low values of these properties or coupling molecules with low and high values of the same feature. Overall, it can be concluded that the rules that dominate the co-crystal formation are far more complex than just some general properties and many parameters should be taken into consideration.

The computational strategy followed is able to successfully extract the patterns that dominate the known co-crystals and predict a range of potential combinations showing similar trends with the labelled data. Therefore, the number of experiments as well as the time frame required to obtain new compounds can be significantly reduced by focusing on co-formers with high scores and possible interesting properties. A realistic picture of the single class applicability is demonstrated by the discovery of two co-crystals (pyrene-6h-benzo[c]chromen-6-one (1) and pyrene-9,10-dicyanoanthracene (2)), both containing molecules which have not previously been reported as co-formers in the CSD. The co-formers of 1 and 2 are characterized by similar shape/size, polarity and electronic characteristics, confirming the ability of the model to learn and reproduce the key-features of the labelled dataset. The electronic analysis of the two newly synthesized structures revealed that both structures have bandgaps in the range of known organic semi-conductors, pointing out the power of our model in exploring, understanding and expanding the targeted labelled dataset.

5.5 Further work

The one class classification workflow proposed herein is a promising way to tackle imbalanced datasets and prioritising synthetic experiments. However, that was only the first step towards the design of a practical tool for in-silico co-crystal screening. The focus of this current work lies on the π-π co-crystals discovery, consequently the models were only trained on a very limited category of materials. As there are not any available validation datasets, the approach was only validated towards the true positive rate which can be misleading in some instances. Moreover, only one type of representation was tested as the input to the models, the 2D molecular descriptors. The resulting workflow can rank any possible molecular pair based on the extracted patterns from the training set. However, it is not providing any information about how certain the model is for the ranking of the pairs. Keeping these pitfalls in mind, we are going to extend the current methodology trying to incorporate predictions for the whole range of co-crystals in CSD. Covering a larger and more diverse dataset, the validation can be performed more effectively using the results from publicly available experimental data. The machine learning method expansion is discussed in the following chapter.

6 Molecular Set Transformer: Attending to co-crystals in the Cambridge Structural Database

6.1 Introduction

The tendency of various molecules to form multi-component crystal structures has been linked to the observation of several new properties in organic materials. Understanding the molecular basis of co-crystallization and predicting whether two molecules will form a co-crystal or not can have a significant impact in the design of functional materials and especially in the drug discovery process. Although the crystal structure determines the properties of the material and is the most trustworthy indicator that a co-crystal can indeed exist, crystal structure prediction is a time consuming method and thus prohibitive for quick co-former screening.

The aim of this work is to develop predictive models for co-crystal formation that can generalize to all types of currently known co-crystals, ranging from pharmaceutical to electronic co-crystals. For that reason, the workflow proposed in our previous work,[253] *i.e.,* training using only the 'positive data', will be adjusted and scaled-up to cover all the existing co-crystals in the Cambridge Structural Database (~7,500 molecular combinations). Key improvements of this framework include the consideration of various molecular representation techniques, extensive hyperparameter tuning, uncertainty estimation and extensive validation. Feature representation has a major impact on the effectiveness of Machine Learning (ML) models especially on imbalanced datasets. In this context, if both the positive and negative or unknown classes with high amount of disproportionality are well-represented with non-overlapping distributions, good classification rates can be obtained by the ML classifiers.

Molecular Set Transformer, which is an attention based autoencoder designed for sets, is the building block of our classifier. The training of our model was performed in a way such that the reconstruction error is minimised and also an uncertainty aware component can be added. The uncertainty estimates of each prediction can mitigate the effect of out-of-distribution examples and provide a degree of confidence with which the classifier ranks every new datapoint. The final models were tested in real case scenarios using several independent external co-crystal screening datasets collected from literature. To showcase the applicability of the methodology, the best performing model was used for ranking an independent molecular pairs datasets extracted from ZINC20, considering the drug delivery and solubility of the co-formers.

To help visualize and get further insights of all the CCDC co-crystals, we developed an interactive browser-based explorer (https://csd-cocrystals.herokuapp.com/). An online app has also been designed for enabling the wider use of our models for in-silico co-crystal screening (https://github.com/katerinavr/streamlit).

6.1.1 Trends in co-crystal research

Co-crystals are crystalline materials composed of two or more different uncharged molecular compounds in a particular stoichiometry. Over the past years significant attention has been received both from academia and industry due to their possible applications in the pharmaceutical and electronic materials industries. This can be verified by the exponential increase in deposited co-crystals in the Cambridge Structural Database over the recent years (Figure 6.1). Looking at the timeline, it can be observed that the first co-crystals were composed of smaller molecules, as indicated by the average length of their SMILES (Simplified Molecular Input Line Entry System). The highest complexity among the molecular pairs is observed around the early 90's with the discovery of the fullerene (C_{60}) co-crystals.[254] Moreover, an increasing interest in co-crystals with electronic properties is also observed, based on the statistics extracted from Web of Science using as key words 'co-crystal' AND 'electronic'.

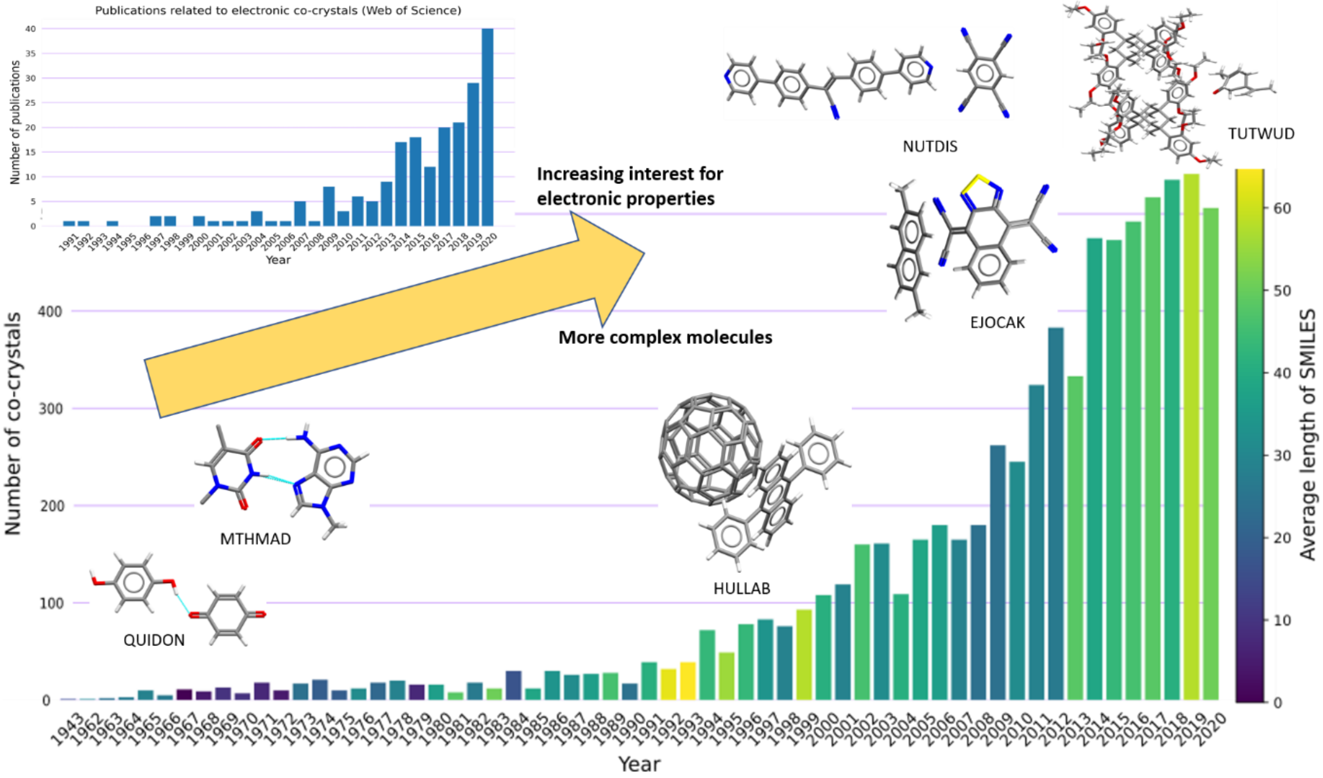


Figure 6.1. *Bar chart with the timeline of co-crystals structures deposited in CSD. The bars are colour-coded based on the complexity of the molecules that form the co-crystals, as indicated from the average length of the SMILES strings. The increase of publications regarding electronic co-crystals is shown in the inset. It can be observed that there are two significant trends, i.e., for designing more complex and electronically interesting co-crystals.*

In pharmaceutical co-crystals at least one of the components is an Active Pharmaceutical ingredient (API), whereas the co-crystals of electronic interest are mainly composed from polyaromatic hydrocarbons (PAHs) which are π-electron rich molecules. For pharmaceutical applications, co-crystallization is an important technique for improving the physicochemical properties of the API without interfering with the chemical behaviour. For example, many pharmaceutical compounds do not make it to the commercial market due to their low solubility. The incorporation of a co-former into the API can result in significantly higher solubility levels in comparison to any crystal form of the API itself. In electronic co-crystals, the existence of a second molecule in the structure might generate a charge-transfer complex where electrons or holes can freely be exchanged between the compounds and thus electrical conductance can be enabled.

6.1.2 Interactions between molecular pairs

Co-crystallization relies purely on intermolecular interactions, and it therefore opens a new range of potential combinations of building blocks to be investigated. If the two building blocks contain only one binding site each and if there is only one way in which those two moieties can be connected a heteromeric synthon can be formed. However, synthetic predictability deteriorates quickly when the number of potentially interacting moieties on each reactant is increased or in cases where one or both reactants lack strong directional moieties. The intermolecular interactions that are present in co-crystals are largely dominated by hydrogen bonds. Hydrogen bonds are formed when a hydrogen atom is covalently bonded to an electronegative atom (A), such that the hydrogen becomes partially positively charged ($H^{\delta+}$). This hydrogen atom can then go on to form an attractive interaction with a second atom (B) which possesses either a lone pair of electrons or polarizable π-electrons. Within the crystal structure the molecules with appropriate functionalities will arrange themselves in a packing arrangement in an attempt to maximize the number and strength of the hydrogen bonding interactions within the solid-state crystal.[255] Alongside H-bonding, other interactions appear to play a significant role in the formation of stable structures, *i.e.*, halogen bonding and π-π stacking (Table D1.1).

Halogen bonds are another type of non-covalent bonding which is typically formed between iodine- or bromine atoms (the halogen-bond donor) and an appropriate halogen-bond acceptor (electron-pair donor) such as an N-heterocycle.[256] Hydrogen and halogen bonds display important strength and directionality and thus offer a good starting point for supramolecular strategies that simultaneously encompass two different non-covalent interactions.

The π-π interactions play a key role in the electronic structure of the materials and refer to the attractive interactions between adjacent π systems such as aromatic rings, arising from attractive interactions between π-electrons and the σ-framework outweighing the repulsive forces between π-electrons. Aromatic rings of neighbouring molecules can arrange themselves in a variety of different orientations, each of which can allow for π-π stacking interactions to form. The way in which the aromatic rings arrange themselves with respect to one another can be influenced by the substituents on the rings, due to the resultant polarisation of the electron cloud. For example, species with unsubstituted aromatic rings tend to form edge-to-face stacking, whereas rings with large substituents form parallel stacking arrangements such as offset π-stacking – face-to-face stacking is rarely observed.[257]

6.1.3 Data-driven approaches for in-silico co-crystal screening

Following the trend of increasing interest in co-crystal synthesis, data-driven methods aimed towards reducing the time needed to screen co-crystals are being actively developed. The first such data-driven method was proposed back in 2009 by Fabian, who first analysed the co-crystals in the Cambridge Structural Database and extracted important statistics that drive co-crystallization. Since then, several other data-driven workflows have been developed, either focusing on a co-crystal subset[204,253,258,259] or on the whole co-crystal dataset.[260,261] In more detail, Wicker *et al.* used a binary classifier trained on an inhouse co-crystal screening dataset composed from both successful and unsuccessful experiments.[204] Przybylek *et al.* are focused on a co-crystal subset based on the co-formers instead of the APIs, showcasing the importance of phenolic and dicarboxylic acids.[258] Devogelaer *et al.* extracted the network of the whole CSD co-crystals and uncover the relations between the molecules.[261,262] Wang *et al.* performed in-silico screening by training a Random Forest binary classifier after generating possible negative pairs based on Tanimoto similarity to the already known molecular pairs that form co-crystals.[260] The common ground in the aforementioned approaches is that they all use a negative dataset and focus on training binary classifiers. Labels in chemistry can be expensive (more experiments), unsustainable (solvents) or in some cases unreliable (different experimentalist and/or different conditions might enable the synthesis of a materials that was previously labelled as negative). For that reason, we want to focus only on the information we have at hand and try to make better use of it. Initially, we started with a small co-crystal dataset, referring to π-π interconnected polyaromatic hydrocarbons (PAHs) co-crystals. That type of co-crystal is interesting in terms of the electronic properties that the materials might possess. We implemented and compared several one class classification approaches and designed a neural network for one class classification which outperformed the standard anomaly detection algorithms. Indeed, we managed to synthesize two new co-crystals based on the pareto optimal co-formers which had the highest similarity to TCNQ, well known for its application in electronically active co-crystals. One major problem we came across, was the complete lack of negative data, even for evaluating the performance of our algorithms which was limited to the evaluation based on the true positive rate (TPR). This evaluation involves the split of the training dataset in five folds, use the four folds for training and the one-fold for evaluating based on how many positives were indeed identified as positives. A dummy classifier would of course have a very high TPR if all the data were identified as inliers. To ensure that we are not facing this problem we used another dataset of unlabelled data with possible molecular pairs that have not yet reported as co-crystals from ZINC15 database and verified that only a part of these data was found to be high-scoring and thus enabled us to draw a reliable threshold above which we assumed that a new pair can be formed.

6.2. Methods

A schematic representation of the workflow followed in this chapter is shown in Figure 6.2.

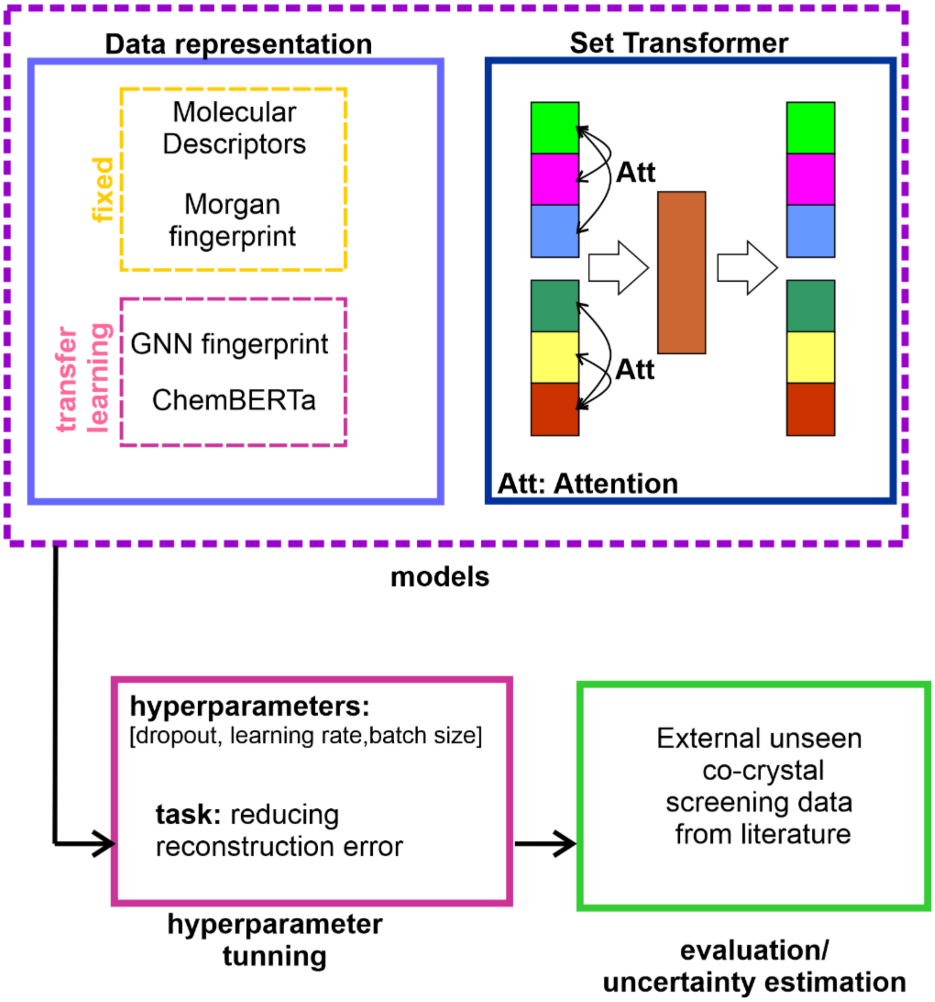


Figure 6.2. *Training/evaluation pipeline and task description. Simplified schematic of Molecular Set Transformer with bidirectional loss architecture.*

6.2.1 Creating the datasets

Training dataset. A key part of the development of a data-driven approach is the creation of a curated dataset that is reliable and can be used for training. The co-crystal dataset used for training the models was extracted from CSD 2020 using the CSD Python API and an in-house python script. The CSD database contains more than one million crystal structures of small molecules and metal-organic molecular crystals resolved by X-ray and neutron diffraction experiments. The whole database was screened with the following constraints:

i) The structures should be only organic, not polymeric, not ionic and should not contain metals.

ii) The structures should have 3D coordinates and no disorder to ensure the high quality of the structures.

iii) Polymorphs are ruled out based on the CSD identifier by dropping out structures that have the first 6 letters the same.

iv) The structures should have exactly two distinct molecules independent of the stoichiometry, *i.e.*, the csd entry *CSATBR* with SMILES string: *OC(=O)c1cc(Cl)ccc1O.OC(=O)c1cc(Cl)ccc1O.CN1C=NC2=C1C(=O)NC(=O)N2C*, has three molecules in the asymmetric unit, however there are only two different co-formers with 1:2 stoichiometry. Given the CSD refcode identifier, the SMILES string representation is extracted and split into the subsequent strings (one SMILES string for each molecule in the structure). A structure is proceeded only if after removing the duplicate strings in each structure, only two different strings remain. In that way we incorporate to the co-crystals dataset structures that belong to different molecular stoichiometries.

v) Neither of the two different molecules in the extracted structure should be a solvent or single atom, as listed in the Appendix Table D1.2.

This process resulted in a training dataset of 7470 molecular pairs.

External validation datasets. As the interest around co-crystals is rising, several studies report both the successful and unsuccessful results from the synthetic attempts. However, the results are not reported in a consistent manner and an extensive literature screening is unavoidable. For the validation and comparison of our models, a benchmark database was created in collaboration with Dr Ioana Sovago from CCDC. This was a time-consuming process that took over 2 months to screen all the related literature, collect the experimental data and then convert them in machine readable files (csv files). In most of the papers the overall screening experiments were reported as supporting information and only the names of the molecules as well as the outcome, *i.e.*, successful or unsuccessful co-crystallization, were given. We had to identify the correct SMILES strings given the names and then assign the label '1' for successful and '0' for unsuccessful experiments. It should be also noted that the experimental validation of a successful co-crystal was not always performed with a detailed crystal structure determination process, but in many cases IR or PXRDs observations were enough for categorizing a molecular pair as a positive or negative example. As Wang *et al.* have already pointed out,[260] there are cases where a molecular pair has been reported as negatives, however after some years the structures were experimentally proven to be positive. In this work the labels have been corrected in a similar way as proposed from Wang *et al.*.[260]

The wide range of diverse categories containing both positive and negative outcomes are listed based on chronological order in Table 6.1.

Table 6.1. *Publicly available co-crystal screening datasets in total consisting of 1,057 negative and 1,320 positive examples.*

Dataset name	**Dataset description**	**Year**	**Number or data**	**Reference**
MEPS dataset	18 APIs against different co-formers	2014	**432** (300 negatives + 132 positives)	263
Artemisinin dataset	Artemisinin + coformers	2014	**38** (36 negatives + 2 positives)	36264
Cooper dataset	20 APIs + 34 coformers (always the same)	2017	**680** (408 negatives + 272 positives)	204
Propyphenazone dataset	Propyphenazone + coformers	2017	**89** (81 negatives + 8 positives)	47
Phenolic acids dataset	Phenolic acids as co-formers	2018	**226** (58 negatives + 168 positives)	259
Dicarboxylic acids dataset	Dicarboxylic acids as co-formers	2019	**712** (104 negatives + 608 positives)	258
Aakeröy dataset	Desloratadine & loratadine + coformers	2020	**82** (17 negatives + 65 positives)	265
Linezolid dataset	Linezolid + coformers	2021	**19** (9 negatives and 10 positives)	266
Pyrene dataset	Pyrene + coformers with electronic similarity to TCNQ	2021	**6** (4 negatives + 2 positives)	253
Praziquantel dataset	Praziquantel + coformers	2021	**30** (18 negatives + 12 positives)	267
MOP dataset	2-amino-4,6-dimethoxypyrimidine (MOP) + 63 co-formers	2021	**63** (22 negative + 41 positives)	268

6.2.2 Data representation

In machine learning for chemistry applications, molecules are translated into a numerical vector of a fixed length, namely the molecular representation or molecular fingerprint. A molecular fingerprint can be either fixed or learned, depending on whether the algorithm will always return the same vector for a molecule (Morgan fingerprint, molecular descriptors) or will learn a task-specific, database dependant vector (neural fingerprint, message passing fingerprint).[93,94]

6.2.2.1 Fixed molecular features

Molecular descriptors. The first case study was on the use of molecular descriptors extracted from a freely available library, namely Mordred.[135] Mordred can calculate more than 1800 numerical representations of molecular properties and/or structural features using predefined algorithmic rules. The disadvantage of this approach is that the library is not further updated and as a result many packages start deprecating, which can result in many NaN (Not a Number) values.

Morgan fingerprint. Morgan Fingerprint (MF) or else extended connectivity fingerprint (ECFP) is generated by assigning unique identifiers, *i.e.*, Morgan identifiers, to all the substructures within a defined radius around all heavy atoms in a molecule.[269] These identifiers are afterwards hashed to a vector with a fixed length. In this work we used the MF with lengths 2048 and 4096 extracted from RDKit library.[167]

6.2.2.2 Learned molecular fingerprints with pretraining

Deep learning models usually require a large amount of data to be trained efficiently. However, not all tasks have enough data to train on. One approach to help achieving better results is pretraining, *i.e.,* a model is first trained on an auxiliary task for which more data exist and then the pretrained model starts with more favourable weights than randomly initialized ones to learn the actual task.[270] For attaining a learned vector, a large, labelled dataset is needed, such that the algorithm will learn the best representation based on the task to be predicted. As in our case no training labels are available, a transfer learning approach was followed by using pretrained models in different tasks where labelled large datasets exist. Transfer learning is supposed to be an effective way for reducing the training bias. We used two different models pretrained in very different tasks, i) a graph-neural network fingerprint pretrained in a self-supervised manner with masking on 2 million unlabelled molecules from ZINC15 database.[271] Each molecule is represented as a 300-dimensional vector after applying the pretrained model. ii) an NLP based fingerprint which is learning the molecular fingerprint by translating the SMILES string to the chemical name trained in 1 million molecules from ChEMBL.

Using pre-trained Graph Neural Networks with transfer learning. Graph neural networks (GNNs) have found many applications in chemistry data as molecules can be easily represented as graphs with the atoms being the nodes and the bonds being the edges. GNNs learn parametrized mappings from graph-structures objects to continuous feature vectors and have achieved state-of-the art performance in a wide variety of problems for property prediction or materials classification. Common feature in these cases was that the training data were labelled and thus the graph neural network is training to extract the molecular representation having a downstream task to achieve. There are numerous cases in chemistry, where labels are not available or are very costly to attain. For that reason, the combination of self-supervised learning with transfer learning for a downstream task is an approach very useful in these situations. In the present work I focused on GNNs which are pretrained with self-supervised methods for learning useful local and global representations simultaneously.[271] Then using transfer learning, I want to examine if the representation learnt in the self-supervised task can be applied to for the pairs representation. The attribute masking as the pretraining step was used, where node/edge attributes of molecules in a large unlabelled dataset are masked and then the GNN tries to predict those attributes based on the neighbourhood structure. We adapted the trained model released by Hu et al.[271] for computing the molecular embeddings of our molecules on the co-crystal pairs and used that representation as the input to Set Transformer, as an alternative fingerprint.

Using Natural Language Processing (NLP) based models and transfer learning. One NLP-based pretrained model, namely ChemBERTa[272] was tested for encoding the molecular SMILES in a learned vector. The vital part for processing text-based chemical representations for deep learning models is the tokenization, i.e., how to break SMILES strings into a sequence of standard units, known as tokens. The tokens are supposed to contain the essential structural information that can reliably and consistently characterize the molecules. ChemBERTA is transformer model that learns molecular fingerprints through semi-supervised pre-training of the sequence-to-sequence language model, using masked-language modelling of a large corpus of 10 million SMILES strings from PubChem. The raw SMILES were tokenized using a Byte-Pair Encoder (BPE) from the Hugging face tokenizers library.

6.2.3 Molecular Set Transformer

Traditional ML approaches usually operate on fixed dimensional vectors or matrices. However, there are several problems that demand the inputs to be order invariants, *i.e.,* sets. Deep learning tasks defined on sets usually require learning functions to be permutation invariant. The Set Transformer architecture was adapted from the work of Lee et al.[236] and was used as building block for the One Class Classifier reported in previous work.[253] For our Molecular Set Transformer, we utilize an Attention Based Autoencoder.

In its simplest form the Autoencoder has two components: an encoder and a decoder. The encoder takes an input and transforms it into a latent representation which is usually a more compact representation than the original datapoint. On the other hand, the decoder is trying to reconstruct the original input from the latent dimension. Mathematically, for a given datapoint x, the encoder compresses the information to a vector z, and the decoder decompresses the data into a reconstructed sample $\hat{x}$. To learn these transformations, neural networks are used as computational and optimizable building blocks for the encoder and decoder. The encoder and decoder are then optimized according to a loss, which is a low reconstruction error ($||x- \hat{x}||$).[110] Set Transformer captures the input in a permutation invariant way. However, to ensure that the output is order invariant as well, a permutation invariant training technique was applied by integrating a bidirectional reconstruction loss function to the original model.[273]

The way the Set Transformer extracts the features is key for capturing the complexity of the problem. Set Transformer 'looks' in all the features across a single molecule as well as in all the features of the pairing molecule. In that way the latent dimension holds information for the relation between the features for each molecular pair. Set Transformer uses a learnable pooling operation, instead of a fixed pooling operation such as mean, to combine the set input such that most of the information is preserved after compressing the data. The pooling operation is the dot-product attention with *SoftMax* (*i.e.*, the self-attention mechanism). In this way, a richer representation of the input data is ensured, that captures higher-order interactions which are important for co-crystal design. The main architectural differences with the previous workflow we implemented for co-crystal screening is that the feed forward neural network was completely removed, and it is now fully based on the attention mechanism with a bidirectional reconstruction loss function.

6.2.4 Hyperparameter tuning.

As the performance of the neural network is highly dependent on the choice of the hyperparameters*, i.e*., algorithm network variables, the hyperparameters were tuned using Weights and Biases software.[274] The model was trained on all 'positive' co-crystal data, excluding those molecular pairs that belong to the validation sets. The traininig was performed without labels and with a different set of parameters each time, having as the final goal to minimize the bidirectional reconstruction loss. After the identification of the optimalset of parameters for each model, the models were retrained using the selected hyperparameters and used for the evaluation on the external validation datasets.

6.2.5 Evaluation metrics

The evaluation of the Molecular Set transformer inspired models is performed in the external datasets containing experimental results from co-crystal screening data. The datasets are balanced between the two classes of co-crystal and not observing a co-crystal, with 1,320 positives and 1,057 negatives assigned as 1 and 0 respectively (Table 6.1). The evaluation metrics used are described below.

The Area Under Curve (AUC) is defined as the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one.

The F1 score is defined as the harmonic mean of precision and recall, where TP stands for true positives, TN for true negatives, FP for false positives, and FN for false negatives predicted by the classifier.

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\,score = 2\ \frac{(Precision \times Recall)}{(Precision + Recall)}$$

Sensitivity or else True Negative Rate (TNR) is an indicator of how correctly the model is predicting the negative class

$$Sensitivity = \frac{TN}{TN + FP}$$

6.2.6 Adding an uncertainty aware component

Machine learning techniques can be used as a powerful and cost-effective strategy to learn from existing datasets and perform predictions on new unseen data. The standard approach is to train the network to minimize a prediction loss. However, the resultant model remains ignorant to its prediction confidence. Herein, we demonstrate the use of Monte Carlo Dropout Ensembling as a Bayesian approximation technique to provide uncertainty estimates on the network's scores.

Dropout is a well-established technique for training neural networks by stochastically setting the weight of each node in the network to zero with probability *p* at every training step. Dropout was initially introduced as a way to avoid overfitting, however, it has been applied is several other works as a strategy to approximate Bayesian inference.[3,275,276]

6.3. Results

6.3.1 Co-crystal space exploration

In order to get insights from the existing co-crystals in the CSD, we initially categorize them in terms of the type of bonding that connects the molecules in the crystal structure. The three main bonding types involve hydrogen bonding, halogen bonding and weak interactions (π-π stacking). The distance between all the existing co-crystals was calculated by the Average Minimum distance metric using the crystal structure of each co-crystal as provided from the Crystallographic Information File (CIF).[277]

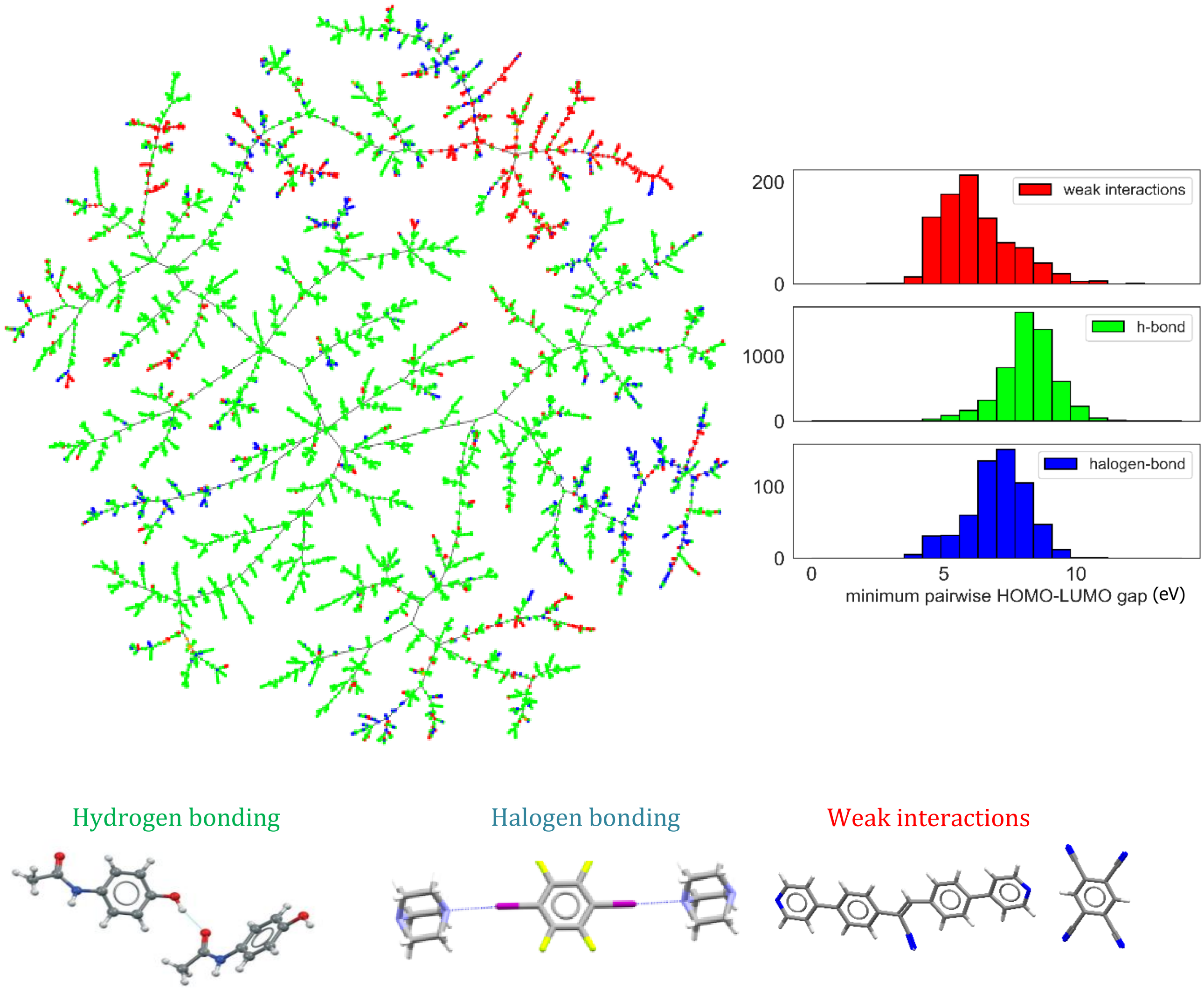


Figure 6.3. *Co-crystal space representation based on the average minimum distance metric (AMD) of the crystal structures, generated by Daniel Widdowson (University of Liverpool) after providing him with the co-crystal data.*[277] *The co-crystals are colour coded based on the main interactions between the two different molecules. Hydrogen bonding is the dominating interaction, whereas the interesting electronic properties arise in the area of the weak interactions where the pairwise HOMO-LUMO difference enables charge transfer interactions.*

The construction of the two-dimensional map, as shown in Figure 6.3, was performed using TMAP algorithm with the structures being colour-coded based on the interactions group they belong to. TMAP is as a dimensionality reduction and data visualization technique capable of representing large high-dimensional datasets as a two-dimensional tree. The local and global structure of the data is preserved, meaning that datapoints that are close in the high dimensional space will also be found close in the lower dimensions.[278] As shown in Figure 6.3, the co-crystal space is dominated by molecules connected via hydrogen-bonds. For getting a further insight regarding the electronic characteristics of the molecular pairs that form the co-crystals, the HOMO-LUMO gap between the two molecules was calculated using PM6 semiempirical method.[279] The calculation was performed by taking the minimum HOMO-LUMO difference between the two isolated co-formers as min($\mathrm{LUMO}_{\mathrm{mol2}}$-$\mathrm{HOMO}_{\mathrm{mol1,}}$ $\mathrm{LUMO}_{\mathrm{mol1}}$-$\mathrm{HOMO}_{\mathrm{mol2}}$). Apparently, the HOMO-LUMO gap is smaller in the area of the weak interaction (red data in Figure 6.3). This was expected as the molecules that participate in charge transfer complexes have small LUMO orbital energies.

Further on, the shape of the individual molecules that form the co-crystal pairs is also investigated. Molecular pairs are sorted such that the first co-former has larger molecular weight than the second co-former. In the PMI plots presented in Figure 6.4 we visualize the shape distribution of the two sets of co-formers.

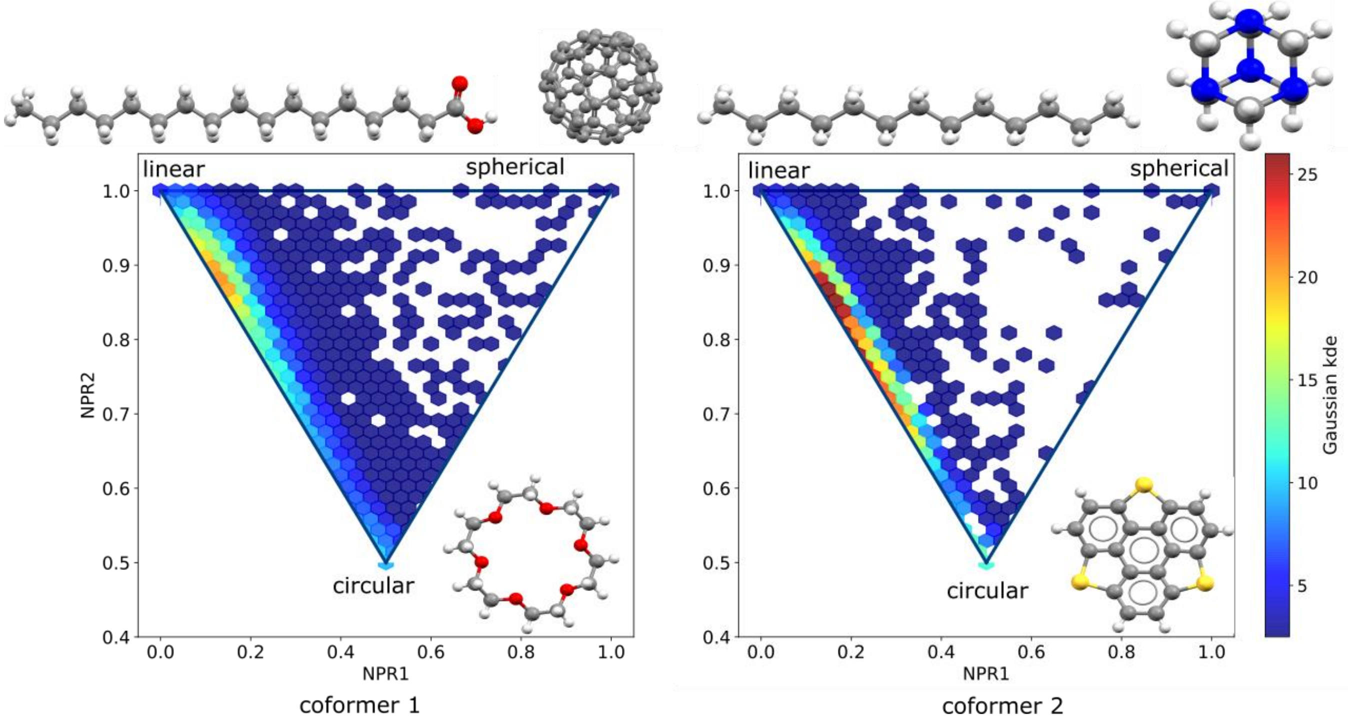


Figure 6.4. *PMI plots of the two co-crystal components, sorted such that the first molecule in the pair is the one with the highest molecular weight. The corners of the triangle show the most linear, most circular and most spherical molecules in the dataset. On the left, the shape distribution of the molecules found as the first co-former is shown, covering a wide are of the triangle. On the right plot, the molecules found as the second co-former covering a smaller area of the triangle. The plots are colour-coded according to the kernel-density estimate (kde) using Gaussian kernels.*

It can be seen that the molecules used as the first co-former (typically the API for in the pharmaceutical co-crystals) cover a wider area on the plot indicating that the molecules are more shape-diverse than those used as second co-formers (or known as excipients for the pharmaceutical co-crystals case). The frequency of the molecules appearing as first and second co-formers was counted, with the top ten molecules of each category being visualized in Figures 6.5 & 6.6.

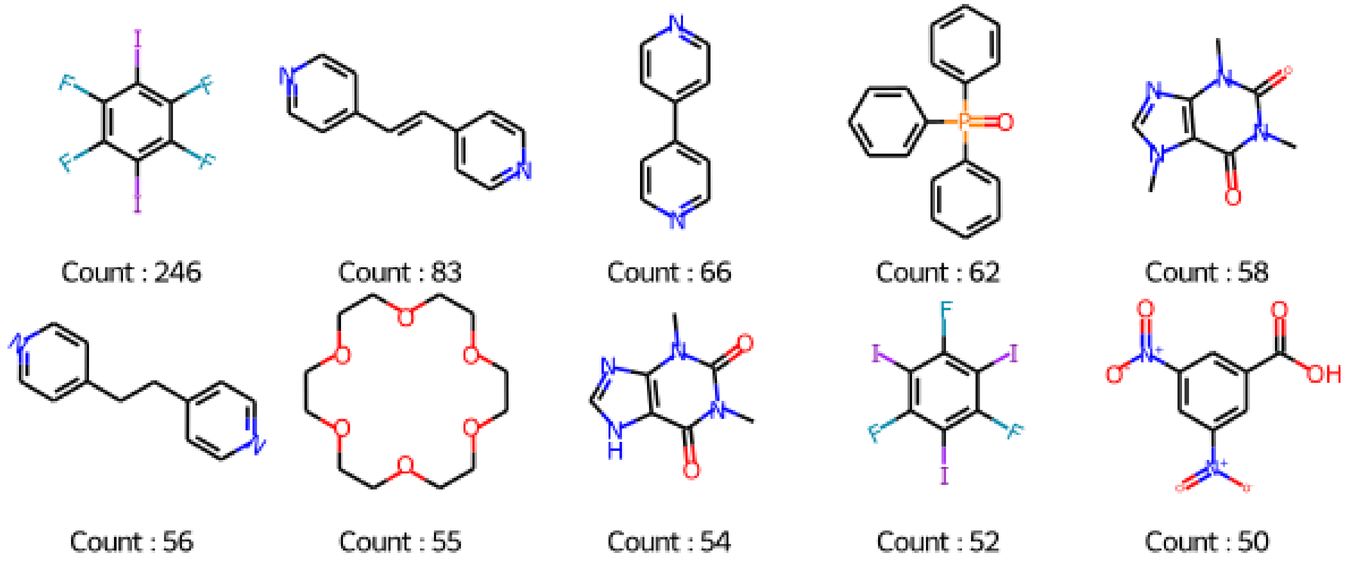


Figure 6.5. *The ten most popular molecules appear as the first co-former and their frequency on the co-crystals dataset.*

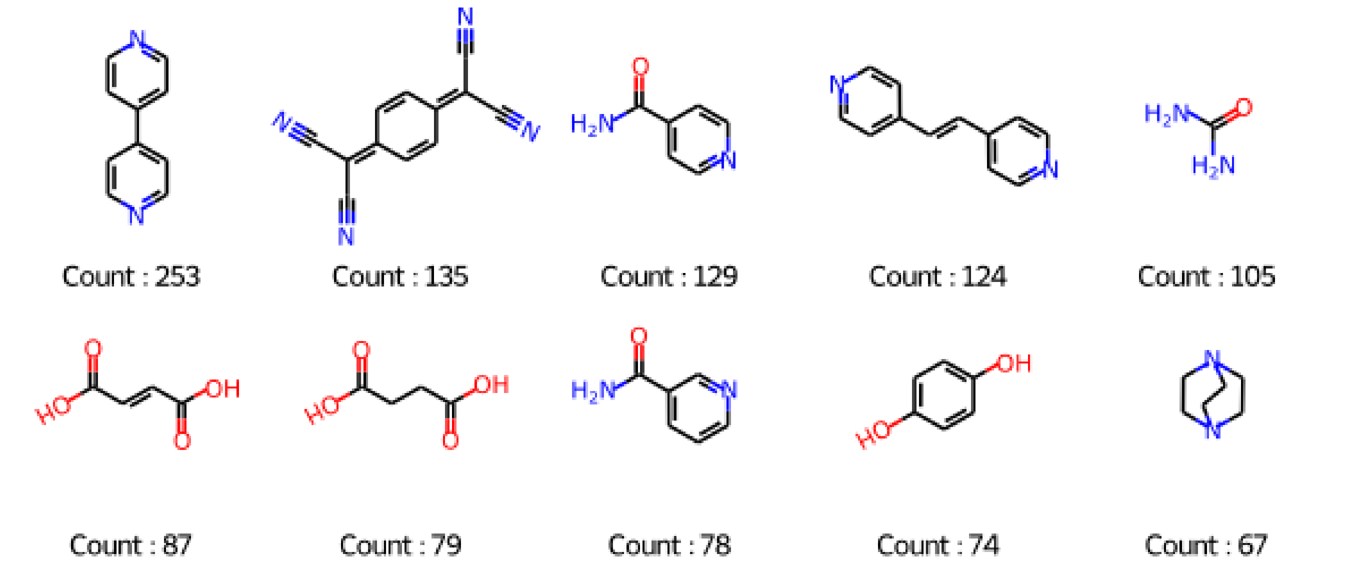


Figure 6.6. *The ten most popular molecules appear as the second co-former and their frequency on the co-crystals dataset.*

6.3.2 Model comparisons

As we have established the one-class approach, based on Set-Transformer, for dealing with the co-crystallization problem, what remains is to identify the most effective representation of our molecules. Herein, we compare four different representation strategies that make use of the 2D molecular structure. Based on each molecular pair's representation method, we developed four different workflows. In addition, two traditional one class classification algorithms (see section 5.2.3), *i.e.*, kNN and Iforest with the Morgan fingerprint as the molecular representation, were trained and tested on the same data as the Molecular Set Transformer (Table 6.3).

The four different models based on the diverse representation techniques were trained on the 'positive' co-crystal data. A dataset collection containing 11 different experimental co-crystal screening datasets was used for the validation and comparison of the models. It should be highlighted that for fair comparisons all the overlapping molecular pairs between the training and the validation datasets were removed from the training set, such that the models haven't previously 'seen' any of the molecular pairs they are validated on. As there are no labels on the training data, the training task of all the models is to minimize the reconstruction loss of the Autoencoder, which is the building block of the Molecular Set Transformer. We explored the relation between the network parameters and the final accuracy on the external data by performing grid search on the learning rate, the batch size, the weight decay, the number of epochs and the dropout rate. The range of the hyperparameters is presented in Table 6.2.

Table 6.2. *Hyperparameters optimization.*

Hyperparameters	Values range
Learning rate	[10^{-3}, 10^{-4}, $3*10^{-4}$, 10^{-5}]
Batch size	[32, 64, 128, 256]
Number of epochs	[50, 100]
Weight decay	[0.0005, 0.005, 0.0001]
Dropout	[0.1, 0.2, 0.3, 0.4]

All the hyperparameter contributions for the GNN-based model are plotted on the parallel coordinates plot as shown in Figure 6.7a. The range of the hyperparameters, the reconstruction loss of the network and the total validation accuracy on the 'unseen' data are shown in the parallel axes. A visual inspection of the relations among the parallel coordinates reveals that there is a strong correlation between the reconstruction loss and the validation accuracy (Figure 6.7b).

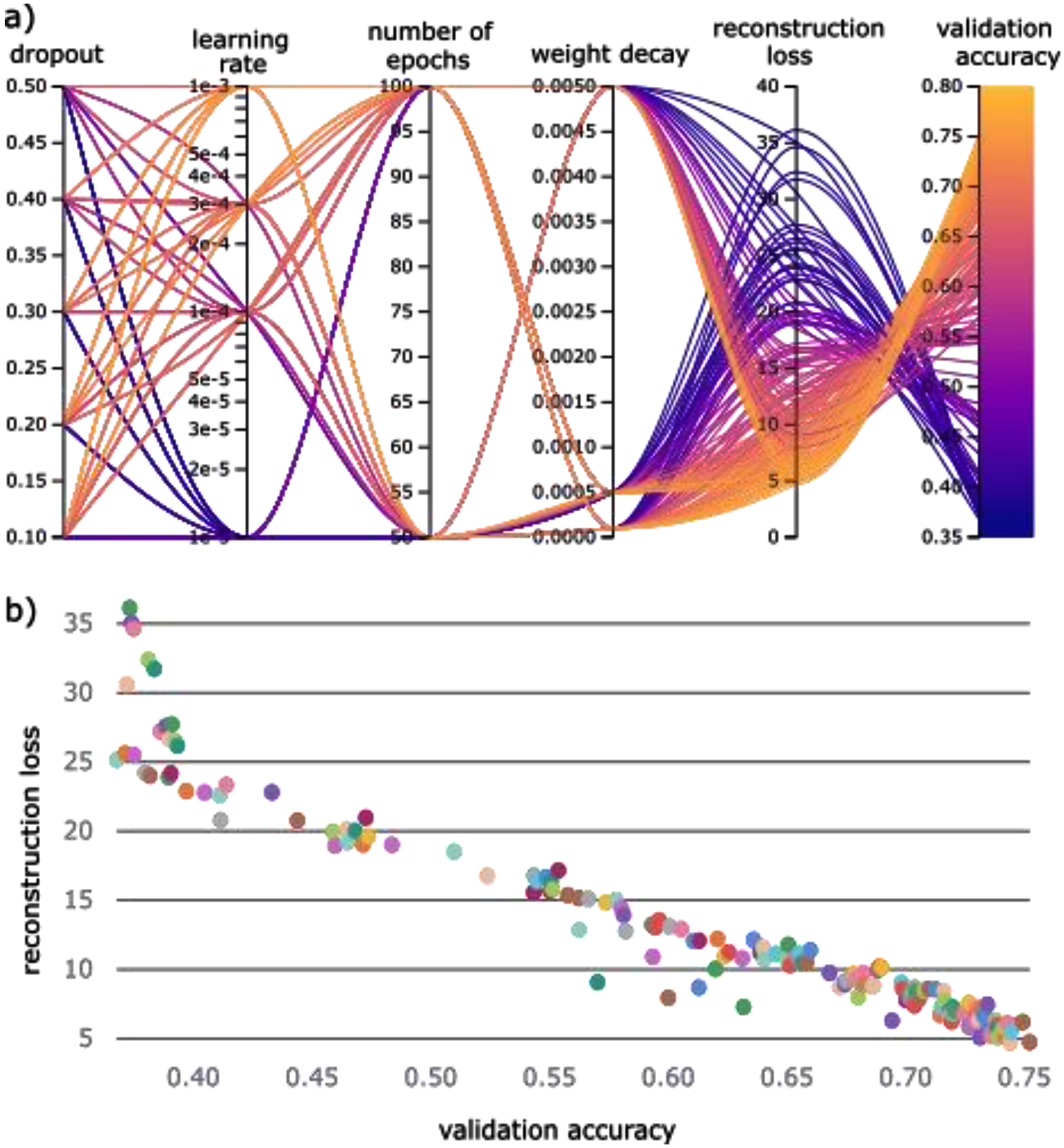


Figure 6.7. *a) Parallel coordinates plot showing the hyperparameters contribution towards the final task, i.e., the minimization of the reconstruction loss. Importantly, it can be observed that as the reconstruction loss decreases, the validation accuracy on previously unseen data increases. b) Scatterplot visualizing the correlation between the validation accuracy and the reconstruction loss. Each run with the different parameters is shown in a different colour. The plots were generated using wandb library (https://wandb.ai/).*

After the selection of the best performing hyperparameters, the models were retrained and their performance of the unseen data is reported in Table 6.3. The performance measures include the total accuracy, Sensitivity and F1.

Table 6.3 *Evaluating Molecular Set Transformer with different input representations on the external benchmark dataset. The metrics include accuracy,* Sensitivity (TNR) *and F1 score. The performance of two traditional one-class classification algorithms is also reported as baseline performance.*

	Accuracy	**Sensitivity**	**F1**
Molecular Set Transformer + Mordred descriptors [a]	0.74 ± 0.007	0.68 ± 0.005	0.7 ± 0.005
Molecular Set Transformer + ECFP4 (2048) [b]	0.73 ± 0.004	0.71 ± 0.005	0.71 ± 0.005
Molecular Set Transformer + ECFP4 (4096) [b]	0.75 ± 0.005	0.71 ± 0.006	0.72 ± 0.005
Molecular Set Transformer + GNN [c]	0.76 ± 0.001	0.69 ± 0.004	0.73 ± 0.005
Molecular Set Transformer + ChemBERTa [d]	0.66 ± 0.005	0.65 ± 0.005	0.63 ± 0.005
Isolation forest + ECFP4 (2048) [b]	0.65	0.58	0.64
kNN+ ECFP4 (2048) [b]	0.62	0.56	0.61

[a] 2D molecular descriptors, 1023 dimensions

[b] 2048 dimensions

[c] 600 dimensions

[d] 354 dimensions

The validation dataset is balanced so the standard metrics can be used to evaluate the performance of the different models. Finding an accuracy of 75% is a significant result considering the fact that the validation data are not extremely reliable especially concerning the negatives cases. The experimental validation of the reported successful or unsuccessful co-crystals was not always performed with a detailed crystal structure determination process, but with IR or PXRDs observations. There are several cases that a molecular pair was reported as unable to form a crystal structure and afterwards trying different conditions from a different researcher gave a successful result (See Methods, External validation datasets).

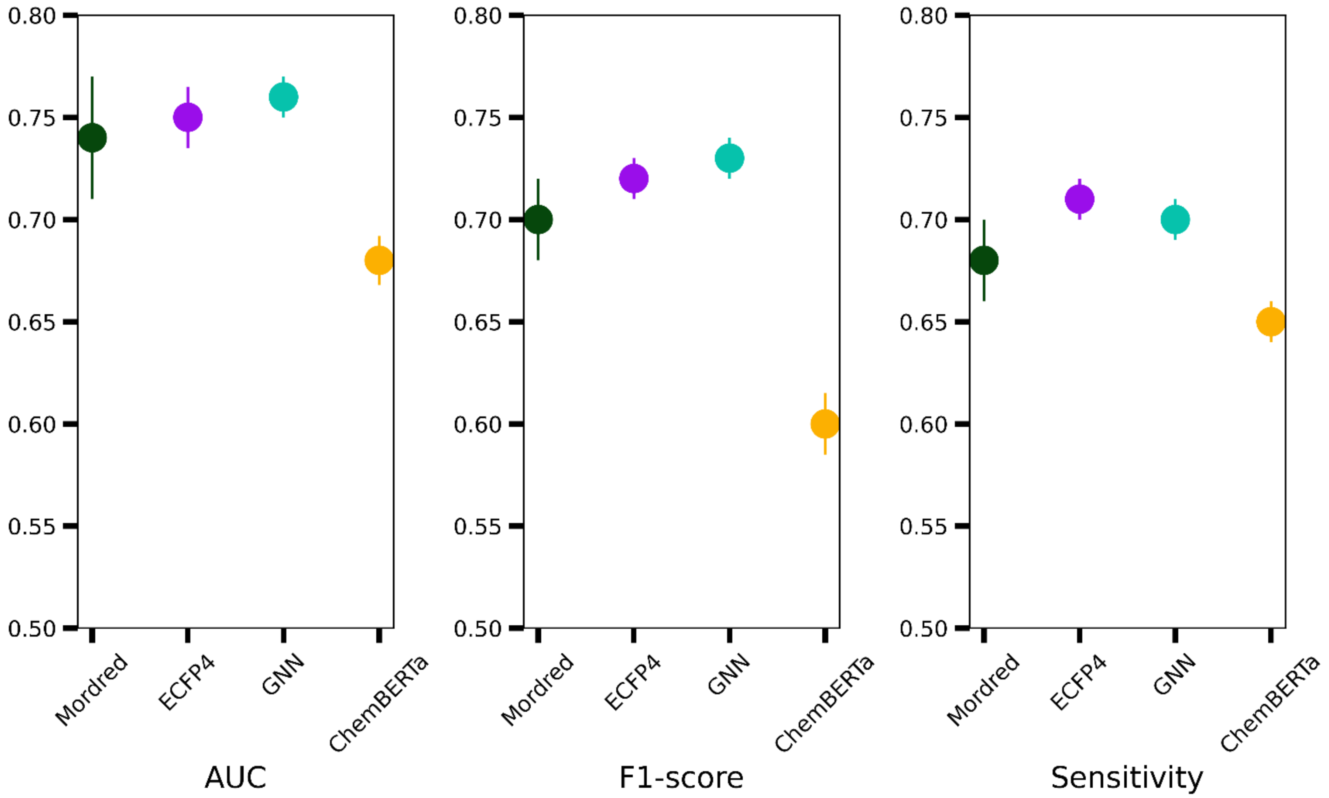


Figure 6.8. *Evaluation metrics and standard deviation of the four different models. A naïve classifier would have 0.5 accuracy. All four different models perform better than a random classifier with the Molecular Set Transformer using GNN or Morgan fingerprint to outperform the other cases.*

From the above plots it can be seen that the Molecular Set Transformer using either Morgan (ECFP4) or GNN fingerprints perform quite well with unseen data. Figure 6.9 shows the probability ranking of the list of co-formers on the validation data. The scores distribution between the true positives and true negatives for each model as well as the confusion matrices are presented in the Appendix Figure D2.1 and D2.2 respectively. We can see that in all models the true positive data points tend to stack on the top of the ranking scatterplot and getting scores close to 1. The experimentally observed hits are significantly enriched at the top, indicating that virtual screening is a promising tool for focusing experimental efforts and reducing the number of experiments required to identify successful molecular pairs. The selection of the 'best' representation is dependent on the domain of application. Numerous studies have shown that GNN fingerprint could yield more promising results, whereas other studies claim that there is not much difference.[108] We should also consider the fact that a GNN representation is not as easily interpretable as the molecular descriptors of the Morgan fingerprint.

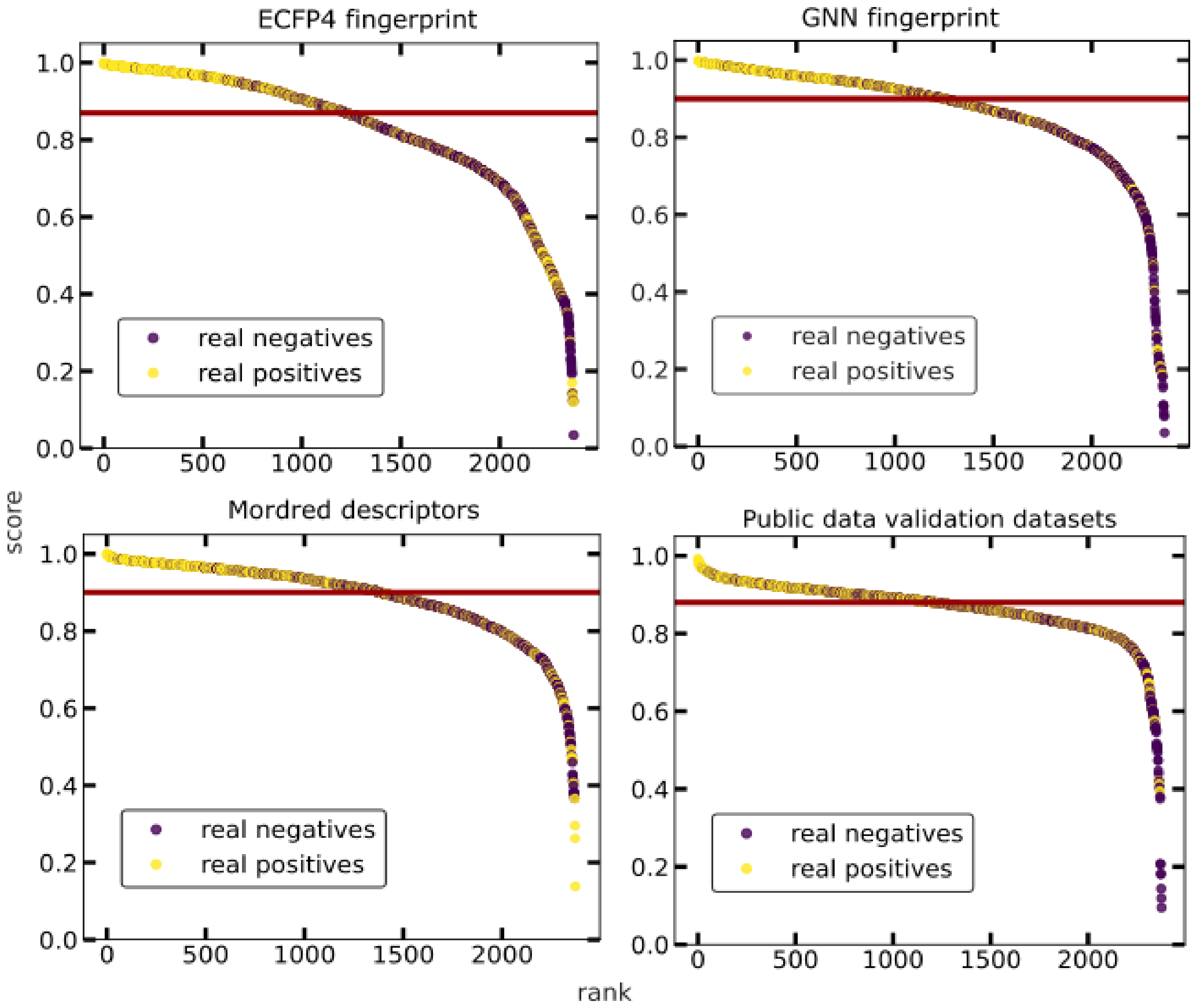


Figure 6.9. *Probability ranking by the four different models used in this work for the external validation sets. The external data consist of two balanced classes of positive and negative data. Yellow dots indicate known co-crystals, whereas unsuccessful co-former pairs are represented as purple dots. The red line is the selected threshold just that the better discrimination between the classes can be achieved. The experimentally observed hits are significantly enriched at the top-ranking percentile, indicating that virtual screening is a promising tool for focusing experimental efforts.*

6.3.3 Benchmarking with currently available methods

The importance of developing accurate and time-efficient co-crystal screening models is showcased by the number of approaches that have been released for this task in the past years.[268,280,281] Most of these approaches are targeting pharmaceutical co-crystals, *i.e.*, pre-screening co-formers against several APIs, due to the importance of making the API more soluble such that is could be easier delivered to the body. To prove the effectiveness of our method, we compared our two best models, *i.e.*, Molecular Set Transformer using either GNN or Morgan fingerprints, with other screening approaches that are currently used and report their performance on publicly available data.

The comparisons are performed against two physical modelling methods and two machine learning methods on single APIs versus the co-formers. As shown in Figure 6.10, the evaluation metric is the AUC per each API. The two physical modelling methods are COSMO-RS[282] and the method based on calculated gas phase molecular electrostatic potential surfaces (MEPS).[263] The two ML models refer to a screening tool developed from Wang *et el* and CCGnet[280] developed from Jiang *et al.*.

COSMO-RS relies on the observation that if the enthalpy between an API-coformers mixture is more negative than the enthalpy of the pure components, then the formation of a co-crystal between the two components is highly possible. The method assesses the miscibility of two components in a super cooled liquid phase according to their excess enthalpy, ΔH_{ex}, which is the difference between enthalpy of the mixture and those of the pure components. The more negative the ΔH_{ex} the more likely the components are to form a stable structure.

On the other hand, MEPS is based on an electrostatic model that treats intermolecular interactions as point contacts between specific polar interaction sites on molecular surfaces. The MEPS of a molecule is calculated in the gas phase, and this is used to identify a discrete set of surface site interaction points (SSIPs), which are described by H-bond donor and H-bond acceptor parameters α and β. SSIPs identify conventional H-bond donor and acceptor sites as well as less polar sites that make weak electrostatic interactions, so they completely describe the surface properties of a molecule and can be used to calculate the total interaction of a molecule with its environment.[263]

The large-scale machine learning model, indicated as Wang method, based on random forest and Morgan fingerprint have been previously tested on most of the twelve APIs shown in Figure 6.10.

Another recent data-driven method, namely CCGnet combining 3D molecular structures and some important molecular fingerprints in a graph neural network reports 97% accuracy on external validation sets. CCGnet is trained on labelled data, both positive and negative, derived from literature screening. Herein, CCGnet was tested on the MOP and ibuprofen external datasets (Figure 6.10 grey bars) as these were the only two APIs that were not part in their training set and a reliable out-of-distribution evaluation score can be calculated. The accuracy in the two previously unseen from their model cases is smaller than any of the other models tested in this work indicating an overinflating reported accuracy.

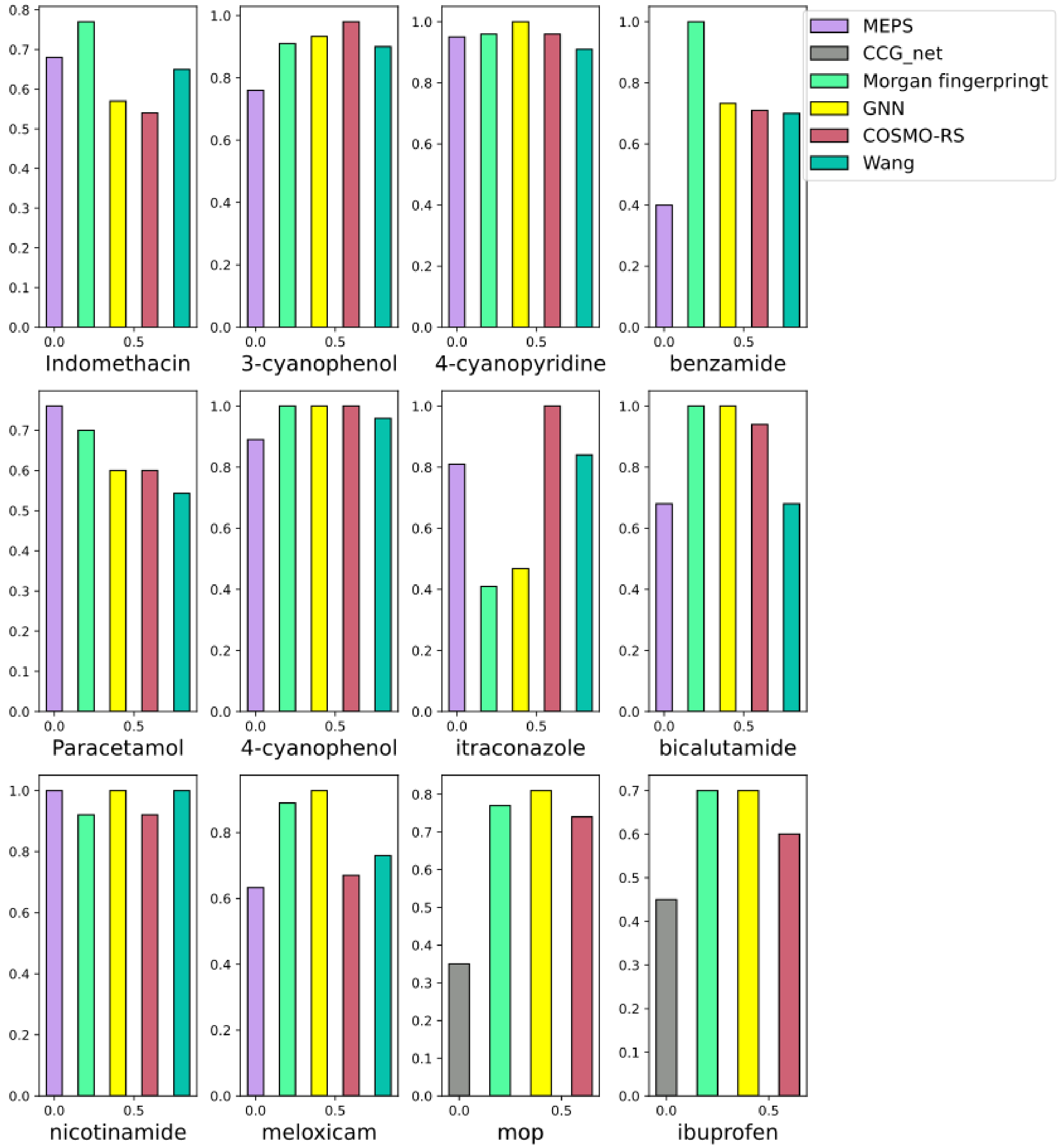


Figure 6.10. *Head-to-head comparison of our best models (green and yellow bars) on individual APIs with other models and methods reported in literature. The evaluation metric is the AUC accuracy (y axis).*

Note that the majority of methods we are comparing with are either computational chemistry models, or ML binary classifiers. Our methodology is only based on neural networks and only positive data were used due to the lack of reliable negative data points within our training set. It is noteworthy that the Molecular Set Transformer was able to have comparable accuracy to computational chemistry models whereas it was based only on the molecular fingerprint. The ML model (CCGnet)[280] claiming 97% accuracy on external data showed the lowest performance on previously 'unseen' data, with a 35% AUC. That is a strong indication that this model was overfit on the training data and that the incorporation of 3D representations did not resulting in a significant advancement of the in-silico screening process. Both our models show the lowest performance in the itraconazole dataset which is the smallest one containing only 8 entries. Itraconazole (Figure 6.11) is a large molecule containing many functional groups and branching. Although four itraconazole co-crystals are reported as a hit in the literature extracted dataset, there is only one itraconazole co-crystal deposited in CSD, *i.e.*, itraconazole:succinic acid (csd id: REWTUK). Consequently, our models have only been trained in one itraconazole co-crystal and were not able to perform well.

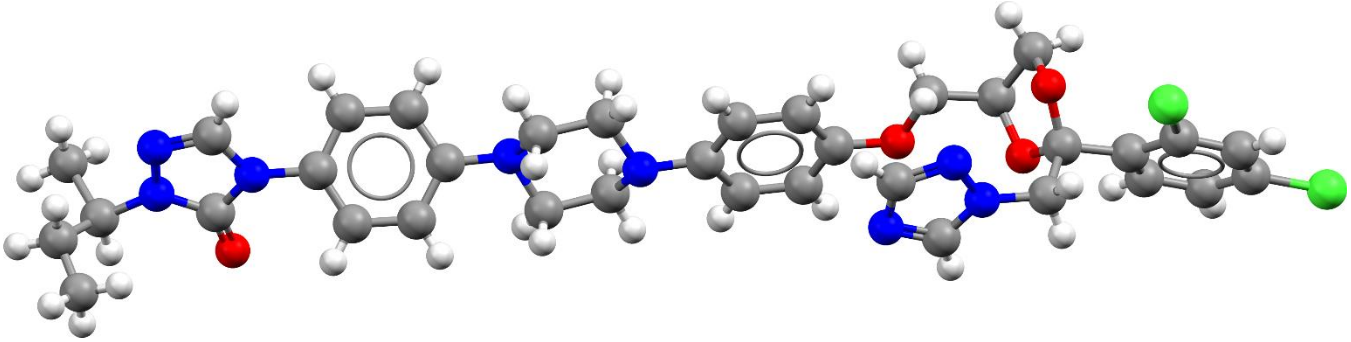


Figure 6.11. *Itraconazole molecular structure.*

6.3.4 Rationalizing the predictions

As the key goal is to generate both predictive models and to gain physical insights for the co-crystallization driving forces, an explainable AI technique was applied. Shapley additive explanations (SHAP) is implemented for rationalizing the scoring of each molecular pair by using feature weights represented as Shapley values from game theory. SHAP is a model-agnostic method where sensitivity analysis is used to investigate the influence of systematic feature values changes on the model output. SHAP-generated explanations can be categorized as global, *i.e*., summarizing the relevance of input features in the model or local, *i.e.,* based on individual predictions.

Of course, the choice of the molecular 'representation model' is an important factor governing the explainability and performance of the AI model as it determines the content and type of the obtained interpretability, *i.e.,* physicochemical properties, functional groups. The features of the input vector are randomly set on and off, thereby examining feature influence in the final scoring. In that way we can get better insights about which features played an important role in the ranking. The advantage of using Shapley values is that we can get local interpretations, meaning that for any single pair or subset of molecular pairs we can 'see' which where the molecular characteristics that played an important role.

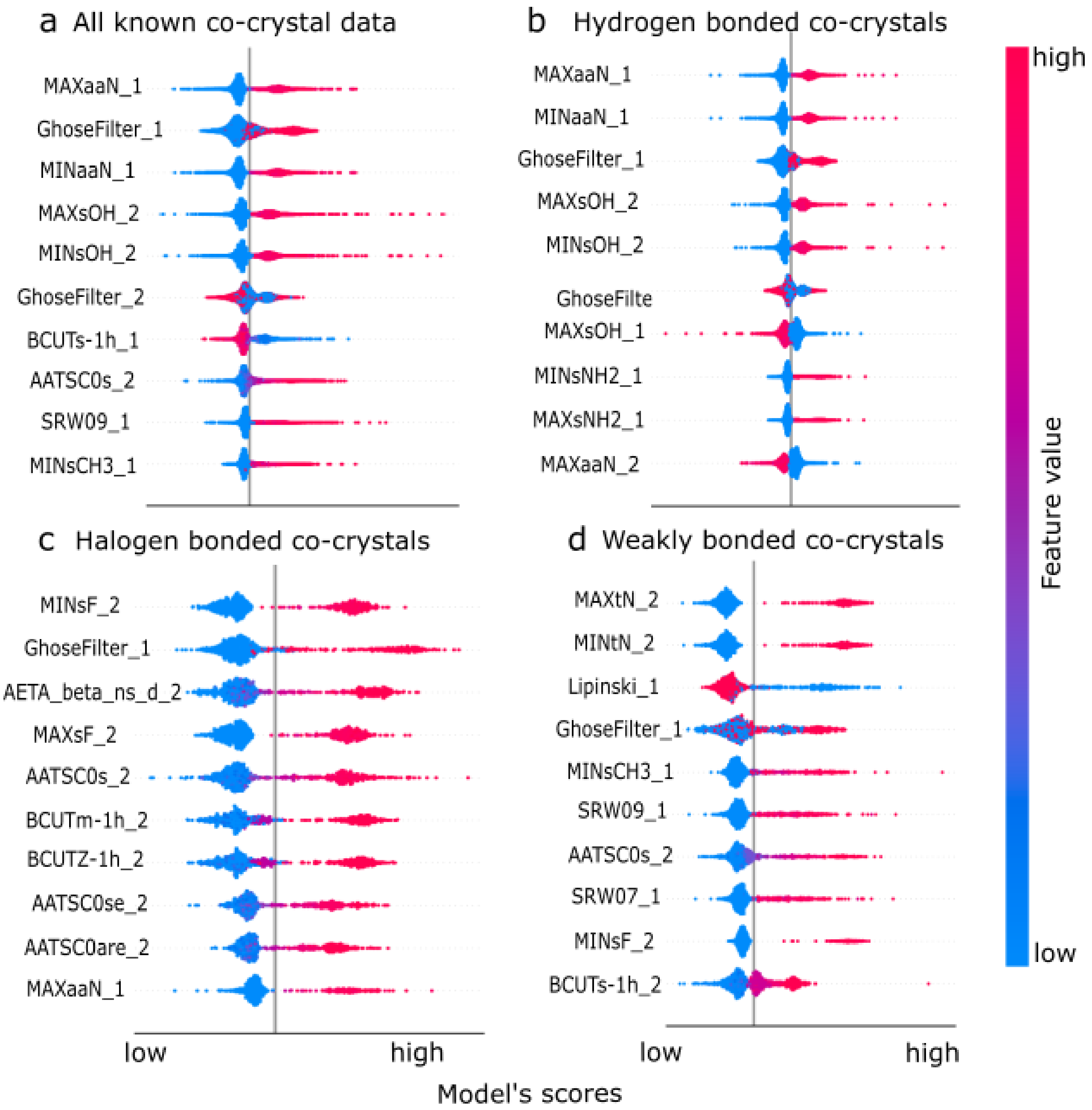


Figure 6.12. Shapley additive explanations categorized according to the type of interactions between the molecular pairs. Each molecule is represented as a vector containing the Mordred descriptors. The notation _1 and _2 indicate the first or second molecule in the pair. a) global interpretation of the whole co-crystals dataset and local interpretations of b) hydrogen bonded pairs, c) halogen bonded pair, d) weakly bonded molecular pairs. The pink colour refers to high values of the molecular features and the blue to low values, whereas the x axis refers to the model's scores being high or the right and low on the left.

As for the co-crystal formation the type of interactions among the molecular pairs plays a crucial role, we got insights for what affected co-crystallization based on which bonding group the pairs belong to. Molecular descriptors representation is straight forward and the weight of each feature can be directly extracted with SHAP. According to Figure 6.12a representing the Shapley global interpretation, we can observe that as the dataset is dominated by H-bond interactions the most important features are related to the OH group (MAXsOH,MINsOH) and the N group (MAXaaN, MinaaN). According to the Shapley local interpretations we can derive *i*) the important features for hydrogen bonded pairs (Figure 6.12b) where the existence of OH (MAXsOH,MINsOH), NH2 (MAXsNH2, MINsNH2) and N groups (MAXaaN, MinaaN) are highlighted as the most important contributing factors, *ii*) the dominating features for the halogen bonded co-crystals (Figure 6.12c) are those related to the existence of F groups (MINsF, MAXsF) and *iii*) in the case of weak interactions the existence of electronegative groups such as terminal triple bonded N (≡N) groups (MAXtN, MINtN) or F groups (MINsF) was found to be the most important for the formation of that type of weak bonding. It can be concluded that the top important descriptors of each category are mostly related to the existence of some functional groups in the molecules that form the pairs and not a physical property. That could be the reason why using the molecular fingerprints for the co-crystallization prediction shown a good performance in the tested systems. Shapley local explanations can also be directly used to highlight the important functional groups of high-scoring pairs, when molecules are represented as bit strings (Morgan fingerprint) as shown in the Appending Figure D2.3.

6.3.5 Dataset of suggested experiments - ZINC20

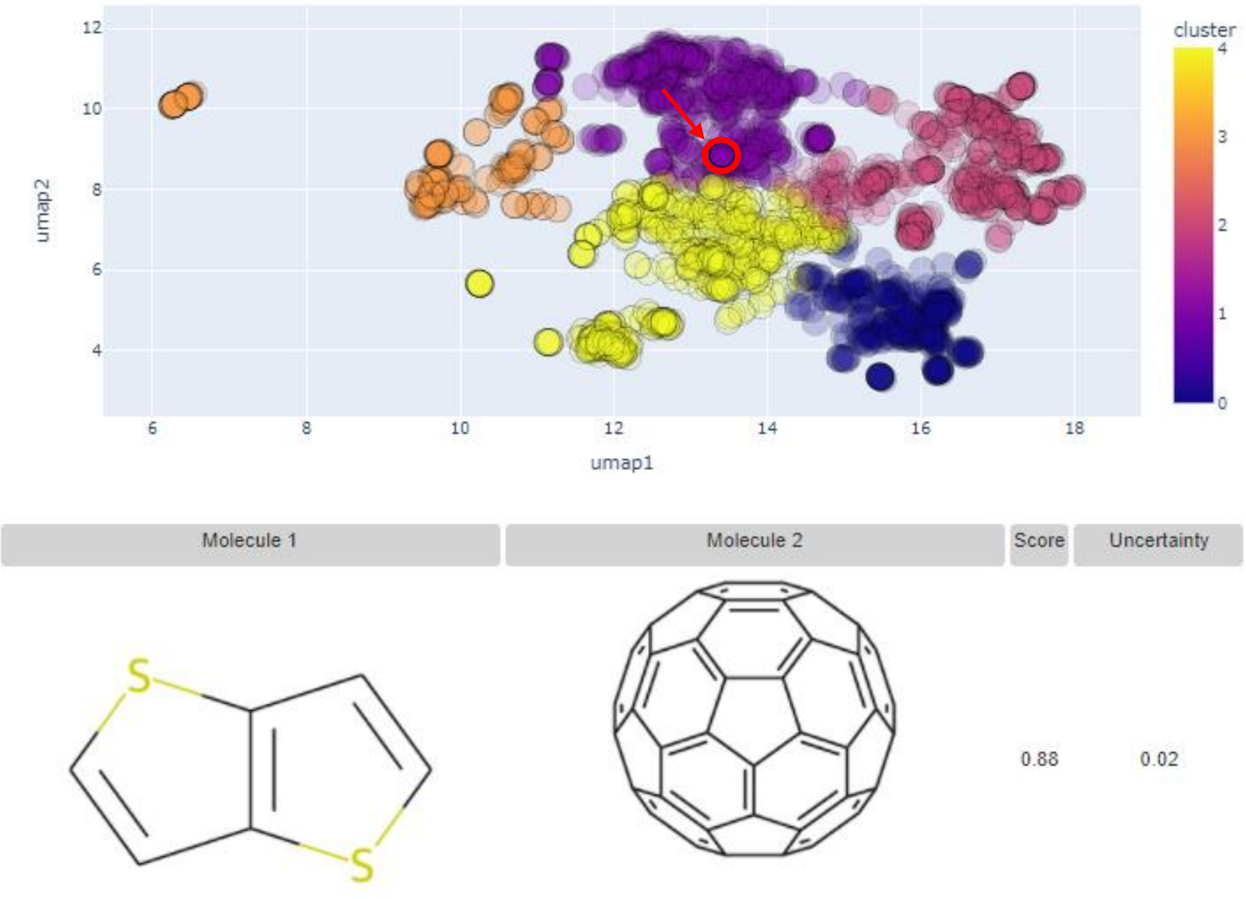
To further demonstrate the applicability of the methodology, one of our best performing model, *i.e.*, the fingerprint model, was used for predicting high-probability molecular pairs from a freely available database with purchasable molecules, namely ZINC20[166]. We extracted all the neutral in-stock molecules getting a starting dataset of 6,883,326 molecules represented as SMILES strings. Out of them we selected only those that have Tanimoto similarity > 0.8 with the molecules that form all the known co-crystals in CSD. That process limited the dataset to 3,119 single molecules.

Solubility and lipophilicity are key parameters that can dictate the success or failure rate of drug discovery and development. Successful drug compounds should have lipophilicity optimal values between 2 and 3 to achieve the optimal bioavailability resistance to metabolism solubility and toxicity. Their measurement is vital for both in-vivo and in-silico evaluation of drug properties. We followed similar approach to Zhao *et al.*. using SwissADME[283] for the calculation of the loqP values as an indicator for lipophilicity.[284] By limiting the selected molecules based on lipophilicity, the molecular dataset was reduced to 300 molecules that pass all these constraints.

All the possible pairs between these molecules were generated and ranked based on our model. Those pairs that scored above 0.8 are plotted in a 2D map and unsupervised clustering was used to cluster them into similarity groups. The representation used is fingerprint and the distance metric is the Tanimoto distance of the pairs. An interactive plot of the high scoring pairs is provided (https://zinc20.herokuapp.com/) as demonstrated in Figure

6.13. The molecular pairs identified from the screening were projected into a two-dimensional map and were grouped into chemical families using the kmeans clustering algorithm. By selecting one point in the interactive map a table is printed which displays the SMILES strings of the two molecules. the molecular diagrams as well as the score and uncertainty of each molecular combination.

Overall, we identified ~2,000 high-scoring potential molecular pairs with low uncertainty, which cover a diverse set of shapes in the molecular space. These pairs could be good possible synthetic targets for achieving novel co-crystals.



Molecule 1	Molecule 2	Score	Uncertainty
		0.88	0.02

Figure 6.13. *2D UMAP embedding of the chemical space of the high scoring co-crystal pairs, colour-coded by k-means clusters identified using the 2D UMAP coordinates. For each selected point a table is displayed showing the images of the molecular pairs, the score of the model and the uncertainty of the prediction.*

6.4. Discussion

Data scarcity remains a fundamental challenge for supervised learning in the materials science domain in which each new labelled data point requires costly and time-consuming laboratory testing. Determining effective ways to make use of large amounts of unlabelled data remains an important unsolved challenge. Herein, we propose the use of the Molecular Set Transformer for learning how to represent molecular sets with high probability to form co-crystals. The machine learning framework has three main parts: (1) data representation (2) machine-learning algorithm and hyperparameter tuning (3) validation on external literature data and uncertainty estimation.

In terms of molecular representations, both fixed (Mordred descriptors, Morgan fingerprint) and learned representations (GNN, ChemBERTa) have been tested. For the learnt representations, pretraining coupled with task-specific finetuning provides substantial gains. We used self-supervised pretraining strategies for GNNs and ChemBerta to assess the viability of these architectures for co-crystal screening. We demonstrated that pre-trained models can be effectively used as 'encoders' for molecules to generate structural features. These features can then be used as input to Set transformer to predict molecular pairs for co-crystallization.

Some of the key findings include that using the Molecular Set Transformer with either the Morgan fingerprint or Graph Neural Network representations perform well on previously unseen data. However, the advantage of using pretrained models (self-supervised training coupled with transfer learning) in the scenarios that only a small amount of training data exists lies to the fact that these models can perform better in data outside the confidence area and provide lower uncertainty with molecules that differ from the training set. Previous approaches have been reported to use graph neural network representations for co-crystals. However, they are only trained with a small amount of data, and usually these types of networks need more than 1M data points to extract the underlying trends. The existing ML models are focusing on solving the co-crystal screening problem by using either sparse or somewhat unreliable negative data from alternative sources to produce a trained model. Our work illustrates that one class classification can overcome these limitations and learn how to effectively describe a certain class of interest, showing the potential to significantly advance many areas of chemical research. As such, we highlight the implementation of one class classification as a methodology for dealing with the 'only positive data' challenge in materials design.

One important observation from our co-crystal prediction approach is that the better and in more detail, we describe the substructure of a molecule, the better we can reconstruct the molecular pairs and extract the interactions among them. Importance of interpreting and getting insights from predictions. Molecular descriptors and fingerprint can offer a better understanding of the characteristics that contribute to the final scoring.

Overall, this work is aiming towards contributing to the co-crystal design field by addressing the major challenges current data-driven for materials discovery face. The problems addressed herein are the lack of negative data, the representation selection, the uncertainty calibration of the model's predictions, the extrapolation on previously unseen data and the interpretability of the models. A solution to these problems is given by providing models that can evaluate diverse molecular pairs in their possibility to form co-crystals, not limited to pharmaceutical co-crystals but also co-crystals of electronic interest. The usefulness of the proposed approach is further demonstrated by ranking combinations from ZINC20 and providing an interactive map of high-ranking high-certainty combinations.

7 Conclusion

7.1 Summary of thesis

This thesis is an attempt to explore the enormous potential data-driven methods have towards accelerating molecular materials discovery. Machine learning models can not only be useful for guiding the research of suitable molecular candidates but also for avoiding prohibitively time consuming calculations or experiments. The two main categories of materials that were investigated are the metal-polyaromatic hydrocarbons systems and co-crystals.

Metal-polyaromatic hydrocarbons systems are unique materials with proven extraordinary electronic properties. Despite their theoretical interest, there are only a few studies reporting their actual structure and related properties. A periodic density functional theory study for all known pure metal-PAHs structures was performed and showed that the energetic stability achieved after the metal insertion could be the driving force for the formation of these materials. To the best of our knowledge this is the first time a systematic study on the quantification of the thermodynamic stability of these systems has been done. Further on, a workflow is proposed for the establishment of some general rules and guidelines to select the most promising PAHs for metal loading. The candidates are selected based on void space availability, orbital degeneracy, metal loading capacity and experimental accessibility. Alongside with that, a complete crystal structure prediction study was performed for identifying and analysing possible crystal structures for some selected metal-PAH combinations.

PAHs were further investigated as potential substrates for the formation of multicomponent crystals dominated by weak interactions, namely π-π co-crystals. A machine learning framework based on one class classification for sets was developed for enabling high-throughput in-silico co-crystal screening. This tool was able to rank more than 20,000 possible polyaromatic hydrocarbon pairs which were further screened based on their similarity to an electronically active molecule, TCNQ. As an outcome two novel PAHs co-crystals were synthesized showing semi-conducting behaviour.

Given the successful application of the machine learning model in the small π-π co-crystals subset, the integration of more co-crystal cases was further tested. Data quality, molecular representations and uncertainty estimations are the basic improvements of the extended model, namely Molecular Set Transformer. Molecular Set Transformer is an autoencoder designed for sets, *i.e.*, inputs that should be order invariant. The models were trained on the whole existing co-crystal data and their reliability is evaluated on external co-crystal screening experimental data. Our findings indicate that even without any labels given, our model outperforms other data-driven or physical modelling co-crystal screening methodologies.

7.2 The future of materials design: challenges and opportunities

The last part of the conclusion is dedicated to open challenges and opportunities in machine learning for materials chemistry. Some of the fundamental assets for the future of materials design are discussed below.

7.2.1 Data and databases

Machine learning models benefit from data volume and data integrity. The quality of the data used to train machine learning algorithms is crucial for the outcome. Although open data availability and databases that follow the FAIR (Findability, Accessibility, Interoperability, and Reuse) principles[285] have significantly contributed to the development and benchmarking of newly developed algorithms and methodologies, there is still room for improvement.

The existing databases continue to suffer from missing data, publication bias and lack of negative data points. Moreover, many existing benchmarks are small, not diverse, and can easily be overfit. Some recent attempts to overcome these limitations include the development of the Open Reaction Database (ORD) for structuring and sharing organic reaction data[286] and the development of the High Throughput Experimental Materials (HTEM) Database for sharing experimental data for synthesizing inorganic materials.[287] Another critical bottleneck is the lack of large databases that report materials properties, *i.e.*, ICSD contains more than 100,000 entries but provides only information of the composition and structure of the inorganic materials and not any related properties. Although there are some smaller databases providing different properties for inorganic materials, *e.g.*, AtomWorks,[288] the availability of datasets referring to molecular properties is still very limited. So far, the most widely used dataset for organic molecules is QM9[289] which reports computed geometric, energetic, electronic, and thermodynamic properties, however it only refer to small molecules with up to nine heavy atoms.

Following the limitation in the existing databases, materials datasets that derive from them are inhomogeneous and heavily biased towards materials frequently used and successful experiments. The publication of failed experiments would enable the development of models that could learn from negative examples. In this work, an alternative methodology for dealing with the underrepresentation of negative data points is proposed, using as a case study the co-crystals deposited in CSD (see Chapters 5 and 6). In addition, we created a benchmark dataset consisting of 2,377 experimental co-crystal screening results including both positive and negative outcomes. This benchmark dataset could be further used for training or validating ML models tailored for *in-silico* co-crystal screening.

Another important consideration is that ML models show great promise when dealing with large materials labelled data. However, they suffer in the small data regime which is common in materials science. For that reason, the development of ML models that can work efficiently given only limited data points is essential.

7.2.2 Molecular representations

In this thesis, several approaches for molecular representation have been tested, mostly based on 2D graph structures or text-based representation (SMILES). However, the 3-dimensional shape of molecules, which may play a crucial role both in material structures and properties, is poorly captured by these representations. In Chapter 3, it was shown that descriptors which can capture the 3D structure of the molecules are more correlated with electronic properties such as the molecular orbital energies and thus are expected to show better predictive performance in comparison to 2D descriptors in that task.

Future studies should focus on better molecular representations including 3D information as this is a promising future direction for property predictions.[290] Although for several applications the 2D graph representation has been proven powerful,[101,291,292] there are cases where the knowledge of the 3D shape is crucial. Using only the 2D graph molecular notation, some important aspects of three-dimensional shape, *e.g.,* the chirality, and relevant conformational dynamics for flexible molecules are neglected. Representative examples include the search for drug-like molecules that can attach to disease-causing proteins and change their functionality[293] and the recent ground-breaking results in protein structure prediction using 3D-rotoequivariant neural networks.[127] The ongoing field of geometrical deep-learning will allow researchers to leverage the symmetries of the molecular representations and thus increase the versatility of computational models for molecular structure generation and property prediction.[294]

While for non-periodic chemical systems, such as molecules, several representations have been proven powerful, *e.g.*, fingerprint, SMILES, graphs, defining a representation for periodic crystalline molecular materials still remains as a big challenge. Recent developments of invariants for crystals inspired from periodic geometry open up a new field on the representation of these systems.[277,295,296] As demonstrated in Chapter 6 Section 6.3.1, an invariant geometric representation of the whole co-crystals space can effectively capture the diversity of the structures and enable for better visualizations.

7.2.3 Models to deal with limited data

Small datasets are ubiquitous in materials science as the data generation process might be expensive or time and resource consuming. For that reason, it is crucial to identify methodologies that can alleviate this problem and be able to perform sensible predictions given a limited amount of data. Some recently developed models for tackling the small data regime problem employ a method called few-shot learning. This type of learning, characterized as a type of meta-learning, involves a pre-training step in a substantially large corpus of data, such that after the pretraining, predictions for smaller and related datasets can be given.[297] The knowledge transfer between the low-data tasks has shown promising results so far in drug discovery, making it an attractive way to go forward in these problems.[298] In our work we demonstrated how pre-trained models can be used for learning the representation of a small set of molecular pairs and improve the predictive ability of our neural network.

Future directions in this field and especially for coupling ML with autonomous platforms require the development of more models and techniques for limited data point. Some recent examples include one-shot learning methods such as Siamese Neural Networks have shown good potential to perform well with low data amount. This type of networks has shown strong capability in image recognition with limited examples and high predictive ability in the low data regime in drug discovery tasks with molecular structure information as an input.[299,300] The coupling of Siamese neural networks with graph convolutional neural networks has also enabled the prediction of synthetic conditions for golden nanoparticles after being trained with only 54 examples.[301]

7.2.4 Explainability

One of the major drawbacks of using machine learning in natural sciences is the lack of physical understanding of the predictions due to the inherent black box character of ML models. However, this information is important for capturing meaningful relationships between features and properties. Recently, Explainable Artificial Intelligence (XAI) methods have become popular for rationalizing the predictions of ML models. Explainability methods that currently exist range from model agnostic (SHAP) and model specific (LIME) libraries. The SHAP library was used in this work to extract those features that are more related to co-crystallization and provide better guidance on the selection of the molecular pairs. Our study demonstrates that neural network-based approaches using molecular descriptors or the Morgan fingerprint to represent the molecules in the pair can be effectively interpreted. On the other hand, using learnt fingerprints though transfer learning requires more complex handling and thus appropriate interpretation techniques has not yet been reported. Further work in this field is required as the more complex the ML models become, the hardest to derive a physical meaning of the predictions.

7.2.5 Inverse design

Inverse design is the concept of designing a material based on the desired functionality. As opposed to property predicting algorithms, which aim to predict a property y given a datapoint x, inverse design is far more challenging. Here the input is the functionality, and the output is a distribution of possible structures.[11] Consequently, the input features which are usually high-dimensional are difficult to predict from the outputs which are low dimensional. However, there are successful cases where inverse design techniques were applied to materials science, *e.g.*, for synthesizing polymers with desired phase behaviour employing particle swarm optimization.[302] In general, the most widely used approaches for inverse design are the generative models. VAEs have been increasingly proposed as appropriate generative frameworks for property targeted molecular generation.[11] Until now the majority of generative models for molecules are focusing on the creation novel valid SMILES strings. However, they are restricted by a lack of spatial information which can only be enabled by the 3D representation of molecules. A recently developed generative network, namely G-SchNet,[12] is able to generate new molecules with tailored properties by their 3D coordinates and thus capturing the relationship between 3D geometry and electronic properties. The use of such a network could be further extent our work on the detection of the most promising PAHs with the desired electronic properties. As our interest was to identify molecules with some certain structural and electronic characteristics that resemble C_{60} and because of the fact that there are limited such molecules in the current databases, a generative model like G-Schnet could be beneficial for generating entirely new molecules.

7.2.6 Uncertainty aware AI

Reliable uncertainty estimates are important for assessing confidence in predictions and enabling decision making and automation. Bayesian optimization is widely used in materials discovery, as it includes the uncertainly estimation which is an indispensable part of every materials system. Using Bayesian optimization models, the aim is not only to predict the property of a given material, but also propose a material to simulate in a next iteration step in a way that minimizes the total number of simulations needed. Bayesian optimizers coupled with automation platforms are going to advance the way chemistry is currently performed. Any predictive ML model should be able to provide some uncertainty estimates for being trustworthy. In our case, the co-crystal data which are currently deposited in the databases might significantly differ from the molecular pairs which are confidential to pharmaceutical companies as such the ability of the model to provide the uncertainty of its predictions is crucial for developing a useful tool with broader applicability.

7.2.6 Automation in chemistry

Automation and robotic platforms are changing the way materials design and synthesis is performed. Automated processes are more efficient, less error-prone than human labour and more reproducible. A single experiment can be considered as a point in a multidimensional space, the parameter space, which is defined by the combination of several experimental conditions, *e.g.*, temperature, reagent stoichiometry, reaction time. Instead of navigating in the parameter space in an orthogonal way, *i.e.*, optimizing one factor as a time by fixing all process factors except for one, automation platforms coupled with optimization algorithms enable the simultaneous optimization of several parameters. Automation processes enable for an efficient navigation on the parameter space by not only predicting a property of a given material but also proposing a material to synthesize or simulate in the next iteration step in a way that minimizes the total number of experiments needed. The less constrained the search space and more flexible the automation platform, the more extensive the possibility for new discoveries. Pairing automation platforms or mobile robotic chemists with data-driven models is going to revolutionise the way chemistry is done today. The research interest is going to be shifted from being only based on data-driven models to the design of rule-based models, which could be taught the rules of chemistry to be able to acquire a ‘chemical intuition’.

Keeping these challenges and opportunities in mind, machine learning methods bear the potential to change, or at least to strongly impact, the way chemical challenges will be approached in the future – guiding and complementing the skill set of a synthetic chemist. With increasing amounts of well-curated data, algorithmic advances, robotics, and cloud computing the prime time for applying machine learning in chemistry is yet to come.

Overall, it is evident that complicated problems and novel materials acceleration require novel methodologies and a highly interdisciplinary way of thinking and approaching each problem. Joined interdisciplinary research between ML and experimental chemistry will unfold the full potential.

Developing a closed loop system between automation platforms or mobile robotic chemists and data-driven models has enormous potential and will likely revolutionise the way chemistry is done today. The field of chemistry requires new methods leveraging all emerging technologies, *i.e.*, AI, cloud computing, robotics and quantum computers.

We should welcome the era of digitalization in chemistry and of course the end goal of all these attempts should be the transition towards to a more sustainable and greener society.

8 Bibliography

1. Cole, J. M. How the Shape of Chemical Data Can Enable Data-Driven Materials Discovery. *Trends Chem.* **3**, 111–119 (2021).

2. Senior, A. W. *et al.* Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).

3. Schwaller, P. *et al.* Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent. Sci.* **5**, 1572–1583 (2019).

4. Schwaller, P., Hoover, B., Reymond, J. L., Strobelt, H. & Laino, T. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. *Sci. Adv.* **7**, (2021).

5. Stokes, J. M. *et al.* A Deep Learning Approach to Antibiotic Discovery. *Cell* **180**, 688-702.e13 (2020).

6. Pinheiro, G. A. *et al.* Machine Learning Prediction of Nine Molecular Properties Based on the SMILES Representation of the QM9 Quantum-Chemistry Dataset. *J. Phys. Chem. A* **124**, 9854–9866 (2020).

7. Shen, J. & Nicolaou, C. A. Molecular property prediction: recent trends in the era of artificial intelligence. *Drug Discovery Today: Technologies* vols 32–33 29–36 (2019).

8. Turcani, L., Greenaway, R. L. & Jelfs, K. E. Machine Learning for Organic Cage Property Prediction. *Chem. Mater.* **31**, 714–727 (2019).

9. Cheng, Z., Zhang, Y., Zhou, C., Zhang, W. & Gao, S. Classification of 5-HT1A receptor ligands on the basis of their binding affinities by using PSO-Adaboost-SVM. *Int. J. Mol. Sci.* **10**, 3316–3337 (2009).

10. Schwaller, P. *et al.* Mapping the space of chemical reactions using attention-based neural networks. *Nat. Mach. Intell.* **3**, 144–152 (2021).

11. Sanchez-Lengeling, B. & Aspuru-Guzik, A. Inverse molecular design using machine learning:Generative models for matter engineering. *Science* vol. 361 360–365 (2018).

12. Gebauer, N. W. A., Gastegger, M. & Schütt, K. T. Symmetry-adapted generation of 3D point sets for the targeted discovery of molecules. in *Advances in Neural Information Processing Systems* vol. 32 (2019).

13. Schmidt, J., Marques, M. R. G., Botti, S. & Marques, M. A. L. Recent advances and applications of machine learning in solid-state materials science. *npj Computational Materials* vol. 5 1–36 (2019).

14. Mervin, L. H., Johansson, S., Semenova, E., Giblin, K. A. & Engkvist, O. Uncertainty quantification in drug design. *Drug Discovery Today* (2020) doi:10.1016/j.drudis.2020.11.027.

15. Chen, L. *et al.* Hidden bias in the DUD-E dataset leads to misleading performance of deep learning in structure-based virtual screening. *PLoS One* **14**, (2019).

16. Kovács, D. P., McCorkindale, W. & Lee, A. A. Quantitative interpretation explains machine learning models for chemical reaction prediction and uncovers bias. *Nat. Commun.* **12**, 1–9 (2021).

17. Correa-Baena, J. P. *et al.* Accelerating Materials Development via Automation, Machine Learning, and High-Performance Computing. *Joule* **2**, 1410–1420 (2018).

18. Savjani, K. T., Gajjar, A. K. & Savjani, J. K. Drug Solubility: Importance and Enhancement Techniques. *ISRN Pharm.* **2012**, 1–10 (2012).

19. Kato, K., Hagi, S., Hinoshita, M., Shikoh, E. & Teki, Y. Photoconductivity and magnetoconductance effects on vacuum vapor deposition films of weak charge-transfer complexes. *Phys. Chem. Chem. Phys.* **19**, 18845–18853 (2017).

20. Petty, M. C., Nagase, T., Suzuki, H. & Naito, H. Molecular electronics. in *Springer Handbooks* 1 (Springer, 2017). doi:10.1007/978-3-319-48933-9_51.

21. Troisi, A. Theories of the Charge Transport Mechanism in Ordered Organic Semiconductors. 213–258 (2009)

doi:10.1007/12_2009_10.

22. Baumann, A. E., Burns, D. A., Liu, B. & Thoi, V. S. Metal-organic framework functionalization and design strategies for advanced electrochemical energy storage devices. *Commun. Chem. 2019 21* **2**, 1–14 (2019).

23. Stone, K. H., Lapidus, S. H. & Stephens, P. W. Implementation and use of robust refinement in powder diffraction in the presence of impurities. *J. Appl. Crystallogr.* **42**, 385–391 (2009).

24. Kim, Y., Park, S. M., Nam, W. & Kim, S. J. Crystal structure of the two-dimensional framework [Mn(salen)]4n[Re6Te8(CN)6]n [salen = N, N'-ethylenebis(salicylideneaminato)]. *Chem. Commun.* **1**, 1470–1471 (2001).

25. Tse, J. S. & Mak, T. C. W. Refinement of the crystal structure of polyethylene terephthalate. *J. Cryst. Mol. Struct.* **5**, 75–80 (1975).

26. Engel, P. C. Glutamate dehydrogenases: the why and how of coenzyme specificity. *Neurochem. Res.* **39**, 426—432 (2014).

27. Younker, J. M. & Dobbs, K. D. Correlating experimental photophysical properties of iridium(III) complexes to spin-orbit coupled TDDFT predictions. *J. Phys. Chem. C* **117**, 25714–25723 (2013).

28. Margadonna, S. *et al.* Crystal structure of superconducting K3Ba3C60: A combined synchrotron X-ray and neutron diffraction study. *Chem. Mater.* **12**, 2736–2740 (2000).

29. Lo, Y.-C., Rensi, S. E., Torng, W. & Altman, R. B. Machine learning in chemoinformatics and drug discovery. *Drug Discov. Today* **23**, 1538–1546 (2018).

30. Liu, Y. *et al.* Materials discovery and design using machine learning. *Journal of Materiomics* vol. 3 159–177 (2017).

31. Richens, J. G., Lee, C. M. & Johri, S. Improving the accuracy of medical diagnosis with causal machine learning. *Nat. Commun.* **11**, 1–9 (2020).

32. Winter, R., Montanari, F., Noé, F. & Clevert, D. A. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chem. Sci.* **10**, 1692–1701 (2019).

33. Sahu, H., Rao, W., Troisi, A. & Ma, H. Toward Predicting Efficiency of Organic Solar Cells via Machine Learning and Improved Descriptors. *Adv. Energy Mater.* **8**, 1801032 (2018).

34. Padula, D., Simpson, J. D. & Troisi, A. Combining electronic and structural features in machine learning models to predict organic solar cells properties. *Mater. Horizons* **6**, 343–349 (2019).

35. Montavon, G. *et al.* Machine learning of molecular electronic properties in chemical compound space. *New J. Phys.* **15**, (2013).

36. Wood, P. A. *et al.* Knowledge-based approaches to co-crystal design. *CrystEngComm* **16**, 5839–5848 (2014).

37. Fda, Cder, Stewart & Felicia. *Regulatory Classification of Pharmaceutical Co-Crystals Guidance for Industry*. http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/default.htm (2018).

38. Schultheiss, N. & Newman, A. ReViews Pharmaceutical Cocrystals and Their Physicochemical Properties. doi:10.1021/cg900129f.

39. Duggirala, N. K., Perry, M. L., Almarsson, Ö. & Zaworotko, M. J. Pharmaceutical cocrystals: along the path to improved medicines. *Chem. Commun.* **52**, 640–655 (2016).

40. Ohtomo, A. & Hwang, H. Y. A high-mobility electron gas at the LaAlO3/SrtiO3 heterointerface. *Nature* **427**, 423–426 (2004).

41. Brinkman, A. *et al.* Magnetic effects at the interface between non-magnetic oxides. *Nat. Mater.* **6**, 493–496 (2007).

42. Reyren, N. *et al.* Superconducting interfaces between insulating oxides. *Science (80-.).* **317**, 1196–1199 (2007).

43. Alves, H., Molinari, A. S., Xie, H. & Morpurgo, A. F. Metallic conduction at organic charge-transfer interfaces. *Nat.*

Mater. **7**, 574–580 (2008).

44. Ferraris, J., Cowan, D. O., Walatka, V. & Perlstein, J. H. Electron Transfer in a New Highly Conducting Donor-Acceptor Complex. *J. Am. Chem. Soc.* **95**, 948–949 (1973).

45. Goetz, K. P. *et al.* Charge-transfer complexes: new perspectives on an old class of compounds. *J. Mater. Chem. C* **2**, 3065–3076 (2014).

46. Galek, P. T. A., Fábián, L., Samuel, W. D., Allen, F. H. & Feeder, N. Structural Science Knowledge-based model of hydrogen-bonding propensity in organic crystals. doi:10.1107/S0108768107030996.

47. Mapp, L. K., Coles, S. J. & Aitipamula, S. Design of cocrystals for molecules with limited hydrogen bonding functionalities: Propyphenazone as a model system. *Cryst. Growth Des.* **17**, 163–174 (2017).

48. Musumeci, D., Hunter, C. A., Prohens, R., Scuderi, S. & McCabe, J. F. Virtual cocrystal screening. *Chem. Sci.* **2**, 883–890 (2011).

49. Wicker, J. G. P. *et al.* CrystEngComm COMMUNICATION Will they co-crystallize? †. **19**, 5336 (2017).

50. Zhou, L. *et al.* Co-crystal formation based on structural matching. (2016) doi:10.1016/j.ejps.2016.02.017.

51. Fábián, L. Cambridge structural database analysis of molecular complementarity in cocrystals. *Cryst. Growth Des.* **9**, 1436–1443 (2009).

52. Issa, N., Karamertzanis, P. G., Welch, G. W. A. & Price, S. L. Can the formation of pharmaceutical cocrystals be computationally predicted? I. Comparison of lattice energies. *Cryst. Growth Des.* **9**, 442–453 (2009).

53. Taylor, C. R. & Day, G. M. Evaluating the Energetic Driving Force for Cocrystal Formation. *Cryst. Growth Des.* **18**, 892–904 (2018).

54. Reilly, A. M. *et al.* Report on the sixth blind test of organic crystal structure prediction methods. *Acta Crystallogr. Sect. B Struct. Sci. Cryst. Eng. Mater.* **72**, 439–459 (2016).

55. Corpinot, M. K. & Bučar, D. K. A Practical Guide to the Design of Molecular Crystals. *Cryst. Growth Des.* **19**, 1426–1453 (2019).

56. Prohens, R. *et al.* Polymorphism of Cocrystals: The Promiscuous Behavior of Agomelatine. (2015) doi:10.1021/acs.cgd.5b01628.

57. Chiarella, R. A., Davey, R. J. & Peterson, M. L. Making co-crystals - The utility of ternary phase diagrams. *Cryst. Growth Des.* **7**, 1223–1226 (2007).

58. Naghavi, S. S. & Tosatti, E. Crystal structure search and electronic properties of alkali-doped phenanthrene and picene. *Phys. Rev. B - Condens. Matter Mater. Phys.* **90**, 75143 (2014).

59. Takabayashi, Y. *et al.* π-electron S = 1/2 quantum spin-liquid state in an ionic polyaromatic hydrocarbon. *Nat. Chem.* **9**, 635–643 (2017).

60. Valentí, R. & Winter, S. M. Polycyclic aromatic hydrocarbons: Synthesis successes. *Nature Chemistry* vol. 9 608–609 (2017).

61. Zabula, A. V., Spisak, S. N., Filatov, A. S., Rogachev, A. Y. & Petrukhina, M. A. Record Alkali Metal Intercalation by Highly Charged Corannulene. *Acc. Chem. Res.* **51**, 1541–1549 (2018).

62. Yoon, T., Koo, J. Y. & Choi, H. C. High Yield Organic Superconductors via Solution-Phase Alkali Metal Doping at Room Temperature. *Nano Lett.* **20**, 612–617 (2020).

63. Gadjieva, N. A. *et al.* Intermolecular Resonance Correlates Electron Pairs down a Supermolecular Chain: Antiferromagnetism in K-Doped p-Terphenyl. *J. Am. Chem. Soc.* **142**, 20624–20630 (2020).

64. Ueno, N. Electronic Structure of Molecular Solids: Bridge to the Electrical Conduction. in *Physics of Organic Semiconductors: Second Edition* 65–89 (Wiley-VCH, 2013). doi:10.1002/9783527654949.ch3.

65. *Molecular Quantum Mechanics*. *Molecular Quantum Mechanics* (2004). doi:10.1201/9781482265545.

66. Dreizler, E. E. · R. M. *Density Functional Theory Theoretical and Mathematical Physics*. *Theoretical and Mathematical Physics* (2011).

67. Lang, P. F. Fermi energy, metals and the drift velocity of electrons. *Chem. Phys. Lett.* **770**, 138447 (2021).

68. Sholl, David S , Stecke;, J. A. Density functional theory. A practical introduction. in *Density functional theory* doi:10.1201/9781420045451.

69. Paufler, P. Introductory Solid State Physics. *Zeitschrift für Krist.* **195**, 160–160 (1991).

70. Transport, E. Chapter 11 Density of States , Fermi Energy and Energy Bands. 1–23 (1854).

71. Durand, P., Darling, G. R., Dubitsky, Y., Zaopo, A. & Rosseinsky, M. J. The Mott-Hubbard insulating state and orbital degeneracy in the superconducting C603- fulleride family. *Nat. Mater.* **2**, 605–610 (2003).

72. Mas-Torrent, M. *et al.* Correlation between crystal structure and mobility in organic field-effect transistors based on single crystals of tetrathiafulvalene derivatives. *J. Am. Chem. Soc.* **126**, 8546–8553 (2004).

73. von Lilienfeld, O. A., Müller, K. R. & Tkatchenko, A. Exploring chemical compound space with quantum-based machine learning. *Nat. Rev. Chem.* **4**, 347–358 (2020).

74. Mitsuhashi, R. *et al.* Superconductivity in alkali-metal-doped picene. *Nature* **464**, 76–79 (2010).

75. Schilling, A., Cantoni, M., Guo, J. D. & Ott, H. R. Superconductivity above 130 K in the Hg-Ba-Ca-Cu-O system. *Nature* **363**, 56–58 (1993).

76. Lee, P. A. From high temperature superconductivity to quantum spin liquid: progress in strong correlation physics. *Reports Prog. Phys.* **71**, 012501 (2007).

77. Hebard, A. F. *et al.* Superconductivity at 18 K in potassium-doped C60. *Nature* **350**, 600–601 (1991).

78. Wu, J., Pisula, W. & Müllen, K. Graphenes as potential material for electronics. *Chem. Rev.* **107**, 718–747 (2007).

79. Cao, Y. *et al.* Unconventional superconductivity in magic-angle graphene superlattices. (2018) doi:10.1038/nature26160.

80. Rieger, R. & Müllen, K. Forever young: Polycyclic aromatic hydrocarbons as model cases for structural and optical studies. *Journal of Physical Organic Chemistry* vol. 23 315–325 (2010).

81. Kubozono, Y. *et al.* Metal-intercalated aromatic hydrocarbons: a new class of carbon-based superconductors. *Phys. Chem. Chem. Phys.* **13**, 16476 (2011).

82. Artioli, G. A. *et al.* Superconductivity in Sm-doped [n]phenacenes (n = 3, 4, 5). *Chem. Commun.* **51**, 1092–1095 (2015).

83. Takabayashi, Y. *et al.* π-electron S = 1/2 quantum spin-liquid state in an ionic polyaromatic hydrocarbon. *Nat. Chem.* **9**, 635–643 (2017).

84. Romero, F. D. *et al.* Redox-controlled potassium intercalation into two polyaromatic hydrocarbon solids. *Nat. Chem.* **9**, 644–652 (2017).

85. Zhong, G. H., Chen, X. J. & Lin, H. Q. Superconductivity and its enhancement in polycyclic aromatic hydrocarbons. *Front. Phys.* **7**, 52 (2019).

86. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* vol. 559 547–555 (2018).

87. Gogas, P. & Papadimitriou, T. Machine Learning in Economics and Finance. *Comput. Econ. 2021 571* **57**, 1–4 (2021).

88. Agrawal, A. & Choudhary, A. Deep materials informatics: Applications of deep learning in materials science. *MRS Communications* vol. 9 779–792 (2019).

89. Jain, S., White, M. & Radivojac, P. *Estimating the class prior and posterior from noisy positives and unlabeled data*.

90. Grisoni, F., Consonni, V. & Todeschini, R. Impact of Molecular Descriptors on Computational Models. *Methods Mol. Biol.* **1825**, 171–209 (2018).

91. Seko, A., Togo, A. & Tanaka, I. Descriptors for Machine Learning of Materials Data. in *Nanoinformatics* 3–23 (Springer Singapore, 2018). doi:10.1007/978-981-10-7617-6_1.

92. Goldsmith, B. R., Boley, M., Vreeken, J. & -, al. Learning physical descriptors for materials science by compressed sensing. (2017) doi:10.1088/1367-2630/aa57bf.

93. Chuang, K. V., Gunsalus, L. M. & Keiser, M. J. Learning Molecular Representations for Medicinal Chemistry. *Journal of Medicinal Chemistry* vol. 63 8705–8722 (2020).

94. Yang, K. *et al.* Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **59**, 3370–3388 (2019).

95. Mikulskis, P., Alexander, M. R. & Winkler, D. A. Toward Interpretable Machine Learning Models for Materials Discovery. *Adv. Intell. Syst.* **1**, 1900045 (2019).

96. Joudaki, D. & Shafiei, F. QSPR Models to Predict Thermodynamic Properties of Cycloalkanes Using Molecular Descriptors and GA-MLR Method. *Curr. Comput. Aided. Drug Des.* **16**, 6–16 (2019).

97. Falcón-Cano, G., Molina, C. & Cabrera-Pérez, M. A. ADME prediction with KNIME: A retrospective contribution to the second “Solubility Challenge”. *ADMET DMPK* **9**, 209–218 (2021).

98. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).

99. Zagidullin, B., Wang, Z., Guan, Y., Pitkänen, E. & Tang, J. Comparative analysis of molecular fingerprints in prediction of drug combination effects. *Brief. Bioinform.* **22**, 1–15 (2021).

100. Pattanaik, L. & Coley, C. W. Molecular Representation: Going Long on Fingerprints. *Chem* **6**, 1204–1207 (2020).

101. Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discov. Today* **11**, 1046–1053 (2006).

102. Leach, A. R. & Gillet, V. J. *An introduction to chemoinformatics*. *An Introduction To Chemoinformatics* (Springer Netherlands, 2007). doi:10.1007/978-1-4020-6291-9.

103. Li, X. *et al.* Combining machine learning and high-throughput experimentation to discover photocatalytically active organic molecules. *Chem. Sci.* **12**, 10742–10754 (2021).

104. Zhao, Z. W., Del Cueto, M., Geng, Y. & Troisi, A. Effect of Increasing the Descriptor Set on Machine Learning Prediction of Small Molecule-Based Organic Solar Cells. *Chem. Mater.* **32**, 7777–7787 (2020).

105. De, S., Bartók, A. P., Csányi, G. & Ceriotti, M. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.* **18**, 13754–13769 (2016).

106. Schütt, K. T., Arbabzadah, F., Chmiela, S., Müller, K. R. & Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.* **8**, 13890 (2017).

107. Jaeger, S., Fulle, S. & Turk, S. Mol2vec: Unsupervised Machine Learning Approach with Chemical Intuition. (2017) doi:10.1021/acs.jcim.7b00616.

108. Jiang, D. *et al.* Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. *J. Cheminformatics 2021 131* **13**, 1–23 (2021).

109. Leke, C. A. & Marwala, T. Introduction to Deep Learning. in 21–40 (2019). doi:10.1007/978-3-030-01180-2_2.

110. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning*. (MIT Press, 2016).

111. Mnih, V. *et al.* Human-level control through deep reinforcement learning. *Nat. 2015 5187540* **518**, 529–533 (2015).

112. Zhao, Y., Xia, X. & Togneri, R. Applications of Deep Learning to Audio Generation. *IEEE Circuits Syst. Mag.* **19**, 19–

38 (2019).

113. Jain, S., White, M., Trosset, M. W. & Radivojac, P. *Nonparametric semi-supervised learning of class proportions*. *Journal of Machine Learning Research* (2015).

114. Sechidis, K., Calvo, B. & Brown, G. *LNAI 8726 - Statistical Hypothesis Testing in Positive Unlabelled Data*. www.cs.man.ac.uk/.

115. Ruff, L. *et al.* Deep Semi-Supervised Anomaly Detection. in *ArXiv* (2020).

116. Ge, C., Gu, I. Y. H., Jakola, A. S. & Yang, J. Deep semi-supervised learning for brain tumor classification. *BMC Med. Imaging* **20**, 1–11 (2020).

117. Khan, S. S. & Madden, M. G. One-class classification: Taxonomy of study and review of techniques. *Knowledge Engineering Review* vol. 29 345–374 (2014).

118. Cutler, J. & Dickenson, M. Introduction to Machine Learning with Python. in *O'Reilly Media, Inc.* 129–142 (2020). doi:10.1007/978-3-030-36826-5_10.

119. Wang, P., Fan, E. & Wang, P. Comparative analysis of image classification algorithms based on traditional machine learning and deep learning. *Pattern Recognit. Lett.* **141**, 61–67 (2021).

120. Rajan, K. Materials informatics. *Mater. Today* **8**, 38–45 (2005).

121. Zheng, A. & Casari, A. *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. (2018).

122. Bellman, R. E. *Adaptive Control Processes: A Guided Tour*. (1961).

123. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **3**, 861 (2018).

124. Haykin, S. *Neural Networks: A Comprehensive Foundation*. (Prentice Hall, 1999).

125. Vaswani, A. *et al.* Attention is all you need. in *Advances in Neural Information Processing Systems* vols 2017-December 5999–6009 (Neural information processing systems foundation, 2017).

126. Tay, Y., Dehghani, M., Bahri, D. & Metzler, D. Efficient Transformers: A Survey. (2020).

127. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

128. Jiménez-Luna, J., Grisoni, F. & Schneider, G. Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence* vol. 2 573–584 (2020).

129. Qi Yuan, Filip T. Szczypiński, and K. E. J. Explainable Graph Neural Networks for Organic Cages. *preprint* (2021).

130. Ribeiro, M. T., Singh, S. & Guestrin, C. 'Why should i trust you?' Explaining the predictions of any classifier. in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* vols 13-17-Augu 1135–1144 (2016).

131. Lundberg, S. M. & Lee, S. I. A unified approach to interpreting model predictions. in *Advances in Neural Information Processing Systems* vols 2017-Decem 4766–4775 (2017).

132. Rodríguez-Pérez, R. & Bajorath, J. Interpretation of compound activity predictions from complex machine learning models using local approximations and shapley values. *J. Med. Chem.* **63**, 8761–8777 (2020).

133. Cambridge. Cambridge Crystallographic Data Centre (CCDC), The Cambridge Structural Database. https://www.ccdc.cam.ac.uk/solutions/csd-system/components/csd/ (2020).

134. Molecular descriptors, QSAR, chemometrics and chemoinformatics - Talete srl. http://www.talete.mi.it/.

135. Moriwaki, H., Tian, Y. S., Kawashita, N. & Takagi, T. Mordred: A molecular descriptor calculator. *J. Cheminform.* **10**, (2018).

136. Groom, C. R., Bruno, I. J., Lightfoot, M. P. & Ward, S. C. The Cambridge Structural Database. *Acta Crystallogr. Sect. B Struct. Sci. Cryst. Eng. Mater.* **72**, 171–179 (2016).

137. *Mercury User Guide and Tutorials 2018 CSD Release*. http://www.ccdc.cam.ac.uk (2017).

138. Glass, C. W., Oganov, A. R. & Hansen, N. USPEX-Evolutionary crystal structure prediction. *Comput. Phys. Commun.* **175**, 713–720 (2006).

139. Willems, T. F., Rycroft, C. H., Kazi, M., Meza, J. C. & Haranczyk, M. Algorithms and tools for high-throughput geometry-based analysis of crystalline porous materials. *Microporous Mesoporous Mater.* **149**, 134–141 (2012).

140. Haranczyk, M. & Martin, R. L. Mathematical Tools for Discovery of Nanoporous Materials for Energy Applications. *J. Phys. Conf. Ser. OPEN ACCESS* doi:10.1088/1742-6596/574/1/012103.

141. Pfau, D., Spencer, J. S., Matthews, A. G. D. G. & Foulkes, W. M. C. Ab initio solution of the many-electron Schrödinger equation with deep neural networks. *Phys. Rev. Res.* **2**, 033429 (2020).

142. Troyer, M. & Wiese, U. J. Computational complexity and fundamental limitations to fermionic quantum Monte Carlo simulations. *Phys. Rev. Lett.* **94**, 170201 (2005).

143. Probert, M. Electronic Structure: Basic Theory and Practical Methods, by Richard M. Martin. *Contemp. Phys.* **52**, 77–77 (2011).

144. Hohenberg, P. & Kohn, W. Inhomogeneous electron gas. *Phys. Rev.* **136**, B864 (1964).

145. Hermann, J., DiStasio, R. A. & Tkatchenko, A. First-Principles Models for van der Waals Interactions in Molecules and Materials: Concepts, Theory, and Applications. *Chemical Reviews* vol. 117 4714–4758 (2017).

146. Thakuria, R., Nath, N. K. & Saha, B. K. The Nature and Applications of π−π Interactions: A Perspective Published as part of a Crystal Growth and Design virtual special issue on π−π Stacking in Crystal Engineering: Fundamentals and Applications. *Cryst. Growth Des* **19**, 13 (2019).

147. Hunter, C. A. & Sanders, J. K. M. The Nature of π-π Interactions. *J. Am. Chem. Soc.* **112**, 5525–5534 (1990).

148. Peng, H., Yang, Z. H., Perdew, J. P. & Sun, J. Versatile van der Waals density functional based on a meta-generalized gradient approximation. *Phys. Rev. X* **6**, 041005 (2016).

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

150. Christensen, A. S., Kubař, T., Cui, Q. & Elstner, M. Semiempirical Quantum Mechanical Methods for Noncovalent Interactions for Chemical and Biochemical Applications. *Chem. Rev.* **116**, 5301–5337 (2016).

151. Wang, X. Y., Yao, X. & Müllen, K. Polycyclic aromatic hydrocarbons in the graphene era. *Science China Chemistry* vol. 62 1099–1144 (2019).

152. Nguyen, L. H. & Truong, T. N. Quantitative Structure-Property Relationships for the Electronic Properties of Polycyclic Aromatic Hydrocarbons. *ACS Omega* **3**, 8913–8922 (2018).

153. Kirkpatrick, P. & Ellis, C. Chemical space. *Nature* **432**, 823 (2004).

154. Schatschneider, B., Monaco, S., Liang, J. J. & Tkatchenko, A. High-throughput investigation of the geometry and electronic structures of gas-phase and crystalline polycyclic aromatic hydrocarbons. *J. Phys. Chem. C* **118**, 19964–19974 (2014).

155. Kuc, A., Heine, T. & Seifert, G. Graphene nanoflakes - structural and electronic properties. *Physical Review B* 1–18 (2013).

156. Schatschneider, B., Phelps, J. & Jezowski, S. A new parameter for classification of polycyclic aromatic hydrocarbon crystalline motifs: A Hirshfeld surface investigation. *CrystEngComm* **13**, 7216–7223 (2011).

157. Desiraju, G. R. & Gavezzotti, A. Crystal structures of polynuclear aromatic hydrocarbons. Classification,

rationalization and prediction from molecular structure. *Acta Crystallogr. Sect. B* **45**, 473–482 (1989).

158. Solà, M. Forty years of Clar's aromatic π-sextet rule. *Front. Chem.* **1**, 22 (2013).

159. Xue, M. *et al.* Superconductivity above 30 K in alkali-metal-doped hydrocarbon. *Sci. Rep.* **2**, (2012).

160. Kim, M., Choi, H. C., Shim, J. H. & Min, B. I. Correlated electronic structures and the phase diagram of hydrocarbon-based superconductors. *New J. Phys.* **15**, 113030 (2013).

161. Böhme, D. K. Fullerene ion chemistry: a journey of discovery and achievement. *Philos. Trans. A. Math. Phys. Eng. Sci.* **374**, (2016).

162. Zadik, R. H. *et al.* Optimized unconventional superconductivity in a molecular Jahn-Teller metal. *Sci. Adv.* **1**, 1–10 (2015).

163. Kuzmich, A., Padula, D., Ma, H. & Troisi, A. Trends in the electronic and geometric structure of non-fullerene based acceptors for organic solar cells. *Energy Environ. Sci.* **10**, 395–401 (2017).

164. Sterling, T. & Irwin, J. J. ZINC 15 – Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **55**, 2324–2337 (2015).

165. Pipeline Pilot. http://accelrys.com/products/pipeline-pilot/.

166. Irwin, J. J. *et al.* ZINC20—A Free Ultralarge-Scale Chemical Database for Ligand Discovery. *J. Chem. Inf. Model.* **0**, null.

167. Landrum, G. *et al.* rdkit/rdkit: 2020_03_1 (Q1 2020) Release. https://zenodo.org/record/3732262 (2020) doi:10.5281/ZENODO.3732262.

168. Meyers, J., Carter, M., Mok, N. Y. & Brown, N. On the origins of three-dimensionality in drug-like molecules. *Future Med. Chem.* **8**, 1753 (2016).

169. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).

170. Davies, D. L. & Bouldin, D. W. A Cluster Separation Measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **PAMI-1**, 224–227 (1979).

171. Gebauer, N. W. A., Gastegger, M. & Schütt, K. T. Symmetry-adapted generation of 3D point sets for the targeted discovery of molecules. in *Advances in Neural Information Processing Systems* vol. 32 (2019).

172. Jørgensen, P. B. *et al.* Machine learning-based screening of complex molecules for polymer solar cells. *J. Chem. Phys.* **148**, 241735 (2018).

173. Menon, A. *et al.* Optical band gap of cross-linked, curved, and radical polyaromatic hydrocarbons. *Phys. Chem. Chem. Phys.* **21**, 16240–16251 (2019).

174. Mishra, A., Ma, C. Q. & Bäuerle, P. Functional oligothiophenes: Molecular design for multidimensional nanoarchitectures and their applications. *Chemical Reviews* vol. 109 1141–1176 (2009).

175. Delaunay, W. *et al.* Synthesis and electronic properties of polycyclic aromatic hydrocarbons doped with phosphorus and sulfur. *Dalt. Trans.* **45**, 1896–1903 (2016).

176. Hebard, A. F. Superconductivity in Doped Fullerenes. *Phys. Today* **45**, 26–32 (1992).

177. Kubozono, Y. *et al.* Metal-intercalated aromatic hydrocarbons: A new class of carbon-based superconductors. *Physical Chemistry Chemical Physics* vol. 13 16476–16493 (2011).

178. Naghavi, S. S. & Tosatti, E. Crystal structure search and electronic properties of alkali-doped phenanthrene and picene. *Phys. Rev. B - Condens. Matter Mater. Phys.* **90**, 075143 (2014).

179. Ueno, N. Electronic Structure of Molecular Solids: Bridge to the Electrical Conduction. in *Physics of Organic Semiconductors: Second Edition* 65–89 (2013). doi:10.1002/9783527654949.ch3.

180. Fleming, R. M. *et al.* Relation of structure and superconducting transition temperatures in A3C60. *Nature* **352**, 787–788 (1991).

181. Tanigaki, K. *et al.* Superconductivity at 33 K in CsxRbyC60. *Nature* **352**, 222–223 (1991).

182. Ganin, A. Y. *et al.* Bulk superconductivity at 38 K in a molecular system. *Nat. Mater.* **7**, 367–371 (2008).

183. Romero, F. D. *et al.* Redox-controlled potassium intercalation into two polyaromatic hydrocarbon solids. *Nat. Chem.* **9**, 644–652 (2017).

184. Bruno, I. J. *et al. IsoStar: A library of information about nonbonded interactions**. *Journal of Computer-Aided Molecular Design* vol. 11 https://link.springer.com/content/pdf/10.1023%2FA%3A1007934413448.pdf (1997).

185. Klime, J., Bowler, D. R. & Michaelides, A. Van der Waals density functionals applied to solids. *Phys. Rev. B - Condens. Matter Mater. Phys.* **83**, 195131 (2011).

186. Curtarolo, S. *et al. AFLOW: an automatic framework for high-throughput materials discovery*. www.aflowlib.org.

187. Anelli, A., Engel, E. A., Pickard, C. J. & Ceriotti, M. *Generalized convex hull construction for materials discovery*.

188. Ong, S. P. *et al.* Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319 (2013).

189. Pennington, C. H. & Stenger, V. A. Nuclear magnetic resonance of C60 and fulleride superconductors. *Rev. Mod. Phys.* **68**, 855–910 (1996).

190. Hiley, C. I. *et al.* Crystal Structure and Stoichiometric Composition of Potassium-Intercalated Tetracene. *Inorg. Chem.* **59**, 12545–12551 (2020).

191. Zhang, J. *et al.* Reactivity of Solid Rubrene with Potassium: Competition between Intercalation and Molecular Decomposition. *J. Am. Chem. Soc.* **140**, 18162–18172 (2018).

192. Štefančič, A. *et al.* Triphenylide-Based Molecular Solid - A New Candidate for a Quantum Spin-Liquid Compound. *J. Phys. Chem. C* **121**, 14864–14871 (2017).

193. Mahns, B., Roth, F. & Knupfer, M. Absence of photoemission from the Fermi level in potassium intercalated picene and coronene films: Structure, polaron, or correlation physics. *J. Chem. Phys* **136**, 134503 (2012).

194. Fleming, R. M. *et al.* Preparation and structure of the alkali-metal fulleride A4C60. *Nature* **352**, 701–703 (1991).

195. Yan, X. W., Zhang, C., Zhong, G., Ma, D. & Gao, M. The atomic structures and electronic properties of potassium-doped phenanthrene from a first-principles study. *J. Mater. Chem. C* **4**, 11566–11571 (2016).

196. De Andres, P. L., Guijarro, A. & Vergés, J. A. Crystal structure and electronic states of tripotassium picene. *Phys. Rev. B* **83**, 245113 (2011).

197. Raccuglia, P. *et al.* Machine-learning-assisted materials discovery using failed experiments. *Nature* **533**, 73–76 (2016).

198. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).

199. Wood, P. A. *et al.* Knowledge-based approaches to co-crystal design. *CrystEngComm* **16**, 5839–5848 (2014).

200. Aitipamula, S. *et al.* Polymorphs, salts, and cocrystals: What's in a name? *Crystal Growth and Design* vol. 12 2147–2152 (2012).

201. Desiraju, G. Co-crystals. Preparation, Characterization and Applications . Edited by C. B. Aakeröy and A. S. Sinha. Royal Society of Chemistry, Monographs in Supramolecular Chemistry No. 24, 2018, Hardcover, pp. 342. Price GBP 159.00. ISBN 978-1-78801-115-0. *Acta Crystallogr. Sect. B Struct. Sci. Cryst. Eng. Mater.* **75**, (2019).

202. Zhang, X. & Hu, W. Molecular cocrystals: design, charge-transfer and optoelectronic functionality. *Phys. Chem. Chem. Phys* **20**, 6009 (2018).

203. Kumar, S. & Nanda, A. Approaches to Design of Pharmaceutical Cocrystals: A Review. *Mol. Cryst. Liq. Cryst.* **667**, 54–77 (2018).

204. Wicker, J. G. P. *et al.* Will they co-crystallize? *CrystEngComm* **19**, 5336–5340 (2017).

205. Zhu, W. *et al.* Revealing the charge-transfer interactions in self-assembled organic cocrystals: Two-dimensional photonic applications. *Angew. Chemie - Int. Ed.* **54**, 6785–6789 (2015).

206. Park, S. K. *et al.* Tailor-made highly luminescent and ambipolar transporting organic mixed stacked charge-transfer crystals: An isometric donor-acceptor approach. *J. Am. Chem. Soc.* **135**, 4757–4764 (2013).

207. Huang, Y. *et al.* Organic Cocrystals: Beyond Electrical Conductivities and Field-Effect Transistors (FETs). doi:10.1002/ange.201900501.

208. Sano, M. Foreword. *Mol. Cryst. Liq. Cryst. Inc. Nonlinear Opt.* **171**, v–vi (1989).

209. Anthony, J. E. Functionalized acenes and heteroacenes for organic electronics. *Chemical Reviews* vol. 106 5028–5048 (2006).

210. Huang, Y. & Egap, E. Open-shell organic semiconductors: an emerging class of materials with novel properties. *Polym. J.* **50**, 603–614 (2018).

211. Nakano, T. Synthesis, structure and function of π-stacked polymers. *Polymer Journal* vol. 42 103–123 (2010).

212. Mandal, A., Choudhury, A., Kumar, R., Iyer, P. K. & Mal, P. Exploring the semiconductor properties of a charge transfer cocrystal of 1-aminopyrene and TCNQ. *CrystEngComm* **22**, 720–727 (2020).

213. Wang, Y. *et al.* Co-crystal engineering: a novel method to obtain one-dimensional (1D) carbon nanocrystals of corannulene-fullerene by a solution process †. *Nanoscale* **8**, 22 (2016).

214. Usman, R. *et al.* Investigation of charge-transfer interaction in mixed stack donor− Acceptor cocrystals toward tunable solid-state emission characteristics. *Cryst. Growth Des.* **18**, 6001–6008 (2018).

215. Khan, A. *et al.* Solid emission color tuning of organic charge transfer cocrystals based on planar π-conjugated donors and TCNB. *J. Solid State Chem.* **272**, 96–101 (2019).

216. Colombo, V., Lo Presti, L. & Gavezzotti, A. Two-component organic crystals without hydrogen bonding: Structure and intermolecular interactions in bimolecular stacking. *CrystEngComm* **19**, 2413–2423 (2017).

217. Ruff, L. *et al.* Deep one-class classification. in *35th International Conference on Machine Learning, ICML 2018* vol. 10 6981–6996 (International Machine Learning Society (IMLS), 2018).

218. Papadimitriou, S., Kitagawa, H., Gibbons, P. B. & Faloutsos, C. LOCI: fast outlier detection using the local correlation integral. *Proc. 19th Int. Conf. Data Eng. (Cat. No.03CH37405)* 315–326 (2003).

219. Breunig, M. M., Kriegel, H.-P., Ng, R. T. & Sander, J. LOF. in *Proceedings of the 2000 ACM SIGMOD international conference on Management of data - SIGMOD '00* 93–104 (ACM Press, 2000). doi:10.1145/342009.335388.

220. Ramaswamy, S., Rastogi, R. & Shim, K. Efficient algorithms for mining outliers from large data sets. in 427–438 (Association for Computing Machinery (ACM), 2000). doi:10.1145/342009.335437.

221. He, Z., Xu, X. & Deng, S. Discovering cluster-based local outliers. *Pattern Recognit. Lett.* **24**, 1641–1650 (2003).

222. Goldstein, M., Goldstein, M. & Dengel, A. Histogram-based Outlier Score (HBOS): A fast Unsupervised Anomaly Detection Algorithm.

223. Liu, F. T., Liu, F. T., Ting, K. M. & Zhou, Z. Isolation Forest. *ICDM '08 Proc. 2008 EIGHTH IEEE Int. Conf. DATA MINING. IEEE Comput. Soc.* 413--422.

224. Lazarevic, A. & Kumar, V. Feature bagging for outlier detection. in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 157–166 (ACM Press, 2005). doi:10.1145/1081870.1081891.

225. Thomas, I. R. *et al.* WebCSD: the online portal to the Cambridge Structural Database. *J. Appl. Cryst* **43**, 362–366

(2010).

226. Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **281413**, 31–36 (1988).

227. Daylight Theory: SMARTS - A Language for Describing Molecular Patterns. https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html.

228. Sterling, T. & Irwin, J. J. ZINC 15 − Ligand Discovery for Everyone. (2015) doi:10.1021/acs.jcim.5b00559.

229. Mauri, A., Consonni, V., Pavan, M. & Todeschini, R. DRAGON software: An easy approach to molecular descriptor calculations. *Match* **56**, 237–248 (2006).

230. Leach, A. R. & Gillet, V. J. Representation And Manipulation Of 3d Molecular Structures. in *An Introduction To Chemoinformatics* 27–52 (Springer Netherlands, 2007). doi:10.1007/978-1-4020-6291-9_2.

231. Zhao, Y., Nasrullah, Z. & Li, Z. PyOD: A python toolbox for scalable outlier detection. *J. Mach. Learn. Res.* **20**, (2019).

232. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. P. and É. D. *Scikit-learn: Machine Learning in Python*. *Journal of Machine Learning Research* vol. 12 http://scikit-learn.sourceforge.net. (2011).

233. Zhao, Y., Wang, X., Cheng, C. & Ding, X. Combining Machine Learning Models using combo Library. (2019) doi:10.1609/aaai.v34i09.7111.

234. Sabokrou, M., Khalooei, M., Fathy, M. & Adeli, E. Adversarially Learned One-Class Classifier for Novelty Detection. *2018 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.* 3379–3388 (2018).

235. Ruff, L. *et al.* Deep Semi-Supervised Anomaly Detection. in *ArXiv* vol. abs/1906.0 (2020).

236. Lee, J. *et al.* Set Transformer: A Framework for Attention-based Permutation-Invariant Neural Networks. in *ICML* (2019).

237. Bergstra, J., Pinto, N. & Cox, D. D. Computational Science &amp; Discovery Hyperopt: a Python library for model selection and hyperparameter optimization Related content SkData: data sets and algorithm evaluation protocols in Python. (2015) doi:10.1088/1749-4699/8/1/014008.

238. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B - Condens. Matter Mater. Phys.* **54**, 11169–11186 (1996).

239. Han, J., Zuo, W., Liu, L., Xu, Y. & Peng, T. Building text classifiers using positive, unlabeled and 'outdated' examples. *Concurr. Comput.* **28**, 3691–3706 (2016).

240. Vitale, R., Marini, F. & Ruckebusch, C. SIMCA Modeling for Overlapping Classes: Fixed or Optimized Decision Threshold? *Anal. Chem.* **90**, 10738–10747 (2018).

241. Menard, E. *et al.* High-Performance n- and p-Type Single-Crystal Organic Transistors with Free-Space Gate Dielectrics. *Adv. Mater.* **16**, 2097–2101 (2004).

242. Wang, Y. *et al.* Organic Cocrystals: New Strategy for Molecular Collaborative Innovation. *Top Curr Chem* **374**, 83 (2016).

243. Zhang, J. *et al.* Sulfur-Bridged Annulene-TCNQ Co-Crystal: A Self-Assembled '"Molecular Level Heterojunction"' with Air Stable Ambipolar Charge Transport Behavior. *Adv. Mater.* **24**, 2603–2607 (2012).

244. Salzillo, T. *et al.* Structure, Stoichiometry, and Charge Transfer in Cocrystals of Perylene with TCNQ-Fx. *Cryst. Growth Des.* **16**, 3028–3036 (2016).

245. Fujisue, C. *et al.* Air-stable ambipolar organic transistors based on charge-transfer complexes containing dibenzopyrrolopyrrole. *RSC Adv.* **6**, 53345–53350 (2016).

246. Wu, H. Di, Peng, H. D. & Pan, G. B. Precise growth of low-dimensional pyrene·perylene·TCNQ co-crystals and

structure-property related optoelectronic properties. *RSC Adv.* **6**, 78979–78983 (2016).

247. Qin, Y. *et al.* Efficient ambipolar transport properties in alternate stacking donor-acceptor complexes: From experiment to theory. *Phys. Chem. Chem. Phys.* **18**, 14094–14103 (2016).

248. Chen, W., Qi, D., Gao, X. & Wee, A. T. S. Surface transfer doping of semiconductors. *Progress in Surface Science* vol. 84 279–321 (2009).

249. Wang, X.-Y., Zhang, W., Zhuang, D., Wang, J.-Y. & Pei, J. Lactone-fused electron-deficient building blocks for n-type polymer field-effect transistors: synthesis, properties, and impact of alkyl substitution positions. *Polym. Chem.* **7**, 2264 (2016).

250. Ai, Q. *et al.* Unusual Electronic Structure of the Donor-Acceptor Cocrystal Formed by Dithieno[3,2-a:2′,3′-c]phenazine and 7,7,8,8-Tetracyanoquinodimethane. *J. Phys. Chem. Lett.* **8**, 4510–4515 (2017).

251. Nematiaram, T., Padula, D., Landi, A. & Troisi, A. On the Largest Possible Mobility of Molecular Semiconductors and How to Achieve It. *Adv. Funct. Mater.* **30**, 2001906 (2020).

252. Costa, J. C. S., Taveira, R. J. S., Lima, C. F. R. A. C., Mendes, A. & Santos, L. M. N. B. F. Optical band gaps of organic semiconductor materials. *Opt. Mater. (Amst).* **58**, 51–60 (2016).

253. Vriza, A. *et al.* One class classification as a practical approach for accelerating π-π co-crystal discovery. *Chem. Sci.* **12**, 1702–1719 (2021).

254. Nadtochenko, V. A., Gritsenko, V. V., D'yachenko, O. A., Shilov, G. V. & Moravskii, A. P. Synthesis of a C60 complex withN,N,N′,N′-tetramethyl-p-phenylenediamine and its crystal structure. *Russ. Chem. Bull. 1996 455* **45**, 1224–1225 (1996).

255. Lunt, R. Solid Form Selectivity in Multi-Component Molecular Crystals: from Batch to Continuous. (2019).

256. Aakeröy, C. B., Schultheiss, N. C., Rajbanshi, A., Desper, J. & Moore, C. Supramolecular synthesis based on a combination of hydrogen and halogen bonds. *Cryst. Growth Des.* **9**, 432–441 (2009).

257. Hunter, C. A., Lawson, K. R., Perkins, J. & Urch, C. J. Aromatic interactions. *J. Chem. Soc. Perkin Trans. 2* 651–669 (2001) doi:10.1039/B008495F.

258. Przybyłek, M. *et al.* Application of Multivariate Adaptive Regression Splines (MARSplines) Methodology for Screening of Dicarboxylic Acid Cocrystal Using 1D and 2D Molecular Descriptors. *Cryst. Growth Des.* **19**, 3876–3887 (2019).

259. Przybyłek, M. & Cysewski, P. Distinguishing Cocrystals from Simple Eutectic Mixtures: Phenolic Acids as Potential Pharmaceutical Coformers. *Cryst. Growth Des.* **18**, 3524–3534 (2018).

260. Wang, D., Yang, Z., Zhu, B., Mei, X. & Luo, X. Machine-Learning-Guided Cocrystal Prediction Based on Large Data Base. *Cryst. Growth Des.* **20**, 6610–6621 (2020).

261. Devogelaer, J., Meekes, H., Tinnemans, P., Vlieg, E. & Gelder, R. Co-crystal Prediction by Artificial Neural Networks**. *Angew. Chemie Int. Ed.* **59**, 21711–21718 (2020).

262. Devogelaer, J. J., Meekes, H., Vlieg, E. & de Gelder, R. Cocrystals in the cambridge structural database: A network approach. *Acta Crystallogr. Sect. B Struct. Sci. Cryst. Eng. Mater.* **75**, 371–383 (2019).

263. Grecu, T., Hunter, C. A., Gardiner, E. J. & McCabe, J. F. Validation of a computational cocrystal prediction tool: Comparison of virtual and experimental cocrystal screening results. *Cryst. Growth Des.* **14**, 165–171 (2014).

264. Karki, S., Friić, T., Fábián, L. & Jones, W. New solid forms of artemisinin obtained through cocrystallisation. *CrystEngComm* **12**, 4038–4041 (2010).

265. Sarkar, N., Mitra, J., Vittengl, M., Berndt, L. & Aakeröy, C. B. A user-friendly application for predicting the outcome of co-crystallizations. *CrystEngComm* **22**, 6776–6779 (2020).

266. Khalaji, M., Potrzebowski, M. J. & Dudek, M. K. Virtual Cocrystal Screening Methods as Tools to Understand the

Formation of Pharmaceutical Cocrystals - A Case Study of Linezolid, a Wide-Range Antibacterial Drug. *Cryst. Growth Des.* **21**, 2301–2314 (2021).

267. Devogelaer, J.-J. *et al.* Cocrystals of Praziquantel: Discovery by Network-Based Link Prediction. *Cryst. Growth Des.* **21**, 3428–3437 (2021).

268. Wu, D. *et al.* Evaluation on Cocrystal Screening Methods and Synthesis of Multicomponent Crystals: A Case Study. *Cryst. Growth Des.* **21**, 4531–4546 (2021).

269. Gardiner, E. J., Holliday, J. D., O'Dowd, C. & Willett, P. P reliminary C ommunication S pecial F ocus : C omputational C hemistry Effectiveness of 2D fingerprints for scaffold hopping. *Future Med. Chem.* **3**, 405–414 (2011).

270. Dai, A. M. & Le, Q. V. Semi-supervised Sequence Learning. *Adv. Neural Inf. Process. Syst.* **28**, (2015).

271. Hu, W. *et al.* Strategies for Pre-training Graph Neural Networks. (2019).

272. Chithrananda, S., Grand, G. & Ramsundar, B. ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction. (2020).

273. Yu, D., Kolbaek, M., Tan, Z. H. & Jensen, J. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings* 241–245 (2017). doi:10.1109/ICASSP.2017.7952154.

274. Biewald, L. Experiment Tracking with Weights and Biases. (2020).

275. Janet, J. P., Duan, C., Yang, T., Nandy, A. & Kulik, H. J. A quantitative uncertainty metric controls error in neural network-driven chemical discovery. *Chem. Sci.* **10**, 7913–7922 (2019).

276. Hirschfeld, L., Swanson, K., Yang, K., Barzilay, R. & Coley, C. W. Uncertainty Quantification Using Neural Networks for Molecular Property Prediction. *J. Chem. Inf. Model.* **60**, 3770–3780 (2020).

277. Widdowson, D., Mosca, M., Pulido, A., Kurlin, V. & Cooper, A. I. The asymptotic behaviour and a near linear time algorithm for isometry invariants of periodic sets. (2020).

278. Probst, D. & Reymond, J. L. Visualization of very large high-dimensional data sets as minimum spanning trees. *J. Cheminform.* **12**, 12 (2020).

279. Saito, T., Kitagawa, Y. & Takano, Y. Reparameterization of PM6 Applied to Organic Diradical Molecules. (2016) doi:10.1021/acs.jpca.6b08530.

280. Jiang, Y. *et al.* Coupling complementary strategy to flexible graph neural network for quick discovery of coformer in diverse co-crystal materials. *Nat. Commun.* **12**, (2021).

281. Wang, D., Yang, Z., Zhu, B., Mei, X. & Luo, X. Machine-Learning-Guided Cocrystal Prediction Based on Large Data Base. *Cite This Cryst. Growth Des* **20**, 6621 (2020).

282. Abramov, Y. A., Loschen, C. & Klamt, A. Rational coformer or solvent selection for pharmaceutical cocrystallization or desolvation. *J. Pharm. Sci.* **101**, 3687–3697 (2012).

283. Daina, A., Michielin, O. & Zoete, V. SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Sci. Reports 2017 71* **7**, 1–13 (2017).

284. Zhao, Z. W., Omar, Ö. H., Padula, D., Geng, Y. & Troisi, A. Computational Identification of Novel Families of Nonfullerene Acceptors by Modification of Known Compounds. *J. Phys. Chem. Lett.* **12**, 5009–5015 (2021).

285. Wilkinson, M. D. *et al.* Comment: The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, (2016).

286. Kearnes, S. M. *et al.* The open reaction database. *J. Am. Chem. Soc.* **143**, 18820–18826 (2021).

287. Zakutayev, A. *et al.* An open experimental database for exploring inorganic materials. *Sci. Data* **5**, 1–12 (2018).

288. Xu, Y., Yamazaki, M. & Villars, P. Inorganic materials database for exploring the nature of material. in *Japanese*

Journal of Applied Physics vol. 50 11RH02 (IOP Publishing, 2011).

289. Ramakrishnan, R., Dral, P. O., Rupp, M. & Von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **1**, (2014).

290. Luo, S., Research, H., Guan, J., Ma, J. & Peng, J. A 3D Generative Model for Structure-Based Drug Design. *Adv. Neural Inf. Process. Syst.* **34**, (2021).

291. Xin, D., Gonnella, N. C., He, X. & Horspool, K. Solvate Prediction for Pharmaceutical Organic Molecules with Machine Learning. *Cryst. Growth Des.* **19**, 1903–1911 (2019).

292. Wang, D., Yang, Z., Zhu, B., Mei, X. & Luo, X. Machine-Learning-Guided Cocrystal Prediction Based on Large Data Base. *Cryst. Growth Des.* **20**, 6610–6621 (2020).

293. Ganea, O.-E. *et al.* GeoMol: Torsional Geometric Generation of Molecular 3D Conformer Ensembles. (2021).

294. Atz, K., Grisoni, F. & Schneider, G. Geometric deep learning on molecular representations. *Nat. Mach. Intell. 2021* 1–10 (2021) doi:10.1038/s42256-021-00418-8.

295. Ropers, J., Mosca, M. M., Anosova, O., Kurlin, V. & Cooper, A. I. Fast predictions of lattice energies by continuous isometry invariants of crystal structures. (2021).

296. Anosova, O. & Kurlin, V. Introduction to Periodic Geometry and Topology.

297. Snell, J., Swersky, K. & Zemel, R. Prototypical networks for few-shot learning. in *Advances in Neural Information Processing Systems* vols 2017-Decem 4078–4088 (2017).

298. Stanley, M. *et al.* FS-Mol: A Few-Shot Learning Dataset of Molecules. (2021).

299. Koch, G., Zemel, R. & Salakhutdinov, R. Siamese Neural Networks for One-shot Image Recognition.

300. Altae-Tran, H., Ramsundar, B., Pappu, A. S. & Pande, V. Low Data Drug Discovery with One-Shot Learning. *ACS Cent. Sci.* **3**, 283–293 (2017).

301. Li, J. *et al.* Deep Learning Accelerated Gold Nanocluster Synthesis. *Adv. Intell. Syst.* **1**, 1900029 (2019).

302. Kumar, J. N. *et al.* Machine learning enables polymer cloud-point engineering via inverse design. *npj Comput. Mater. 2019 51* **5**, 1–6 (2019).

303. Stachulski, A. V. *et al.* Thiazolides as novel antiviral agents. 1. Inhibition of hepatitis B virus replication. *J. Med. Chem.* **54**, 4119–4132 (2011).

APPENDIX A - Chapter 3

A1. Categorizing PAHs with unsupervised clustering

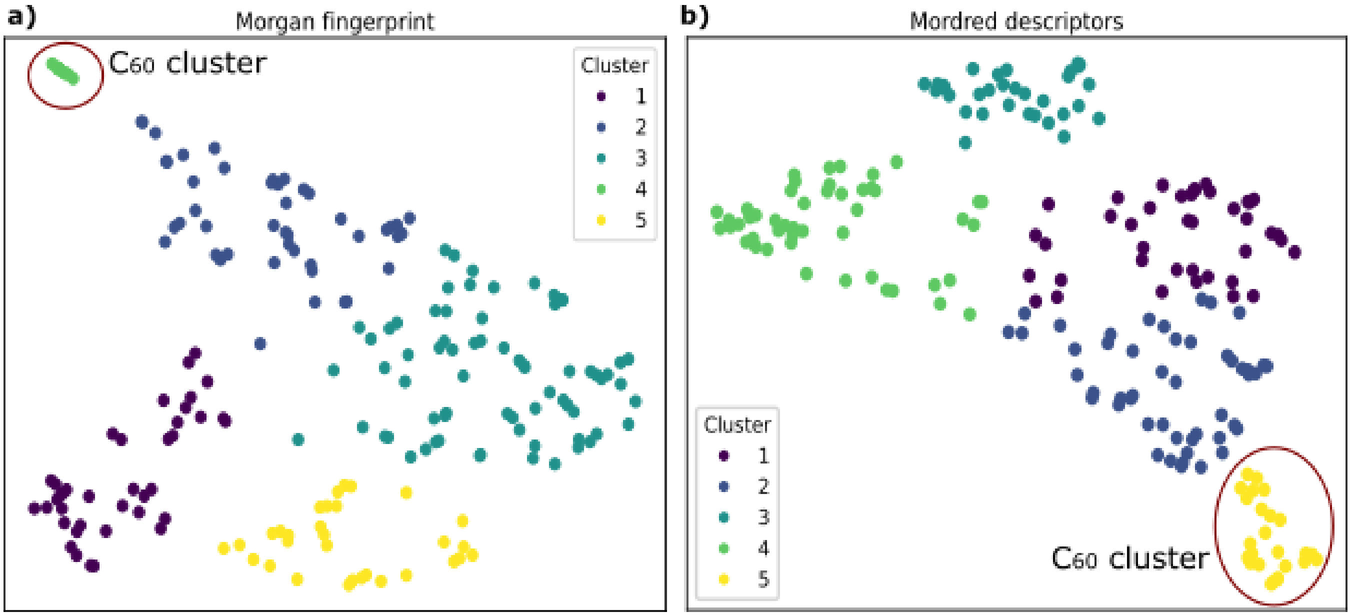


Figure A1.1 Unsupervised clustering based on a) the Morgan fingerprint with Tanimoto index and b) Mordred fingerprint with Euclidean distance as similarity measure. The clusters are colour-coded by Gaussian Mixture clusters identified using the 2D UMAP coordinates.

Table A1.1 Evaluation of the unsupervised clustering techniques tested in this work. The best performing clustering algorithm for each representation is highlighted in bold.

Clustering algorithm	Representation	Silhouette score*	Davies-Bouldin index**
k-means	Morgan fingerprint	0.46	0.67
Affinity propagation		0.43	0.80
Gaussian Mixture		**0.47**	**0.59**
k-means	SOAP descriptor	**0.56**	**0.53**
Affinity propagation		0.55	0.55
Gaussian Mixture		0.53	0.59
k-means	Mordred descriptors	0.51	0.66
Affinity propagation		0.44	0.82
Gaussian Mixture		**0.51**	**0.65**

***** The Silhouette Coefficient for a set of samples is given as:

$S = mean\left(\frac{b-a}{max(a,b)}\right)$ **,** where a is the mean distance between a sample and all other points in the same class and b the mean distance between a sample and all other points in the next nearer cluster. The silhouette coefficient ranges from -1 to 1 and higher the Silhouette coefficient the better the cluster separation.[169]

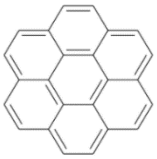
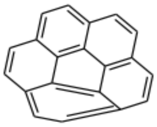
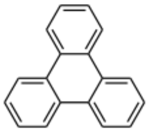
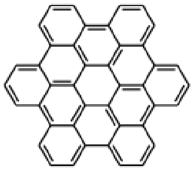
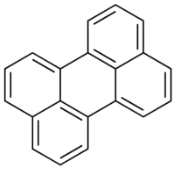
****** The Davies-Bouldin index is given from:[170]

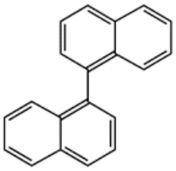
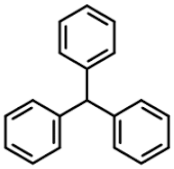
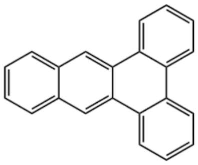
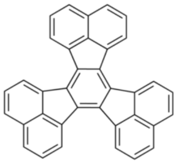
$DB = \frac{1}{k}\sum_{i=1}^{k} max\ R_{ij}$,

where $R_{ij} = \frac{s_i + s_j}{d_{ij}}$, with s_i being the average distance between each point of cluster i and the centroid of that cluster and d_{ij} being the distance between cluster centroids i and j.

The lower the Davies-Bouldin index, the better the separation with the minimum score being zero.

Table A1.2. Single PAHs identified as good hosts for metal insertion

PAH	Reasoning
Coronene 	• Exact double degeneracy in LUMO, LUMO+1 • Most stable ratio in the convex hull is an open shell ratio: K3Coronene • Magnetic character found with DFT calculations
Corannulene 	• Exact double degeneracy in LUMO, LUMO+1 • Most stable ratio in the convex hull is a high metal content ratio: K4Corannulene
Triphenylene 	• Exact double degeneracy in LUMO, LUMO+1 • Most stable ratio in the convex hull is an open shell ratio: K3Triphenylene or a high metal content ratio Cs4triphenylene
Hexabenzo[bc,ef,hi,kl,no,qr] coronene 	• Exact double degeneracy in LUMO, LUMO+1 • Most stable ratio in the convex hull is a high metal content ratio: K6hexabenzocoronene
Perylene 	• Electronic similarity to C60 found by measuring the Euclidean distance of the energy orbitals • Near double degeneracy in LUMO+1, LUMO+2 • Most stable ratio in the convex hull is an open shell ratio: K3Perylene

Binapthalene 	• Large void space • Most stable ratio in the convex hull is a high metal content ratio: K4binapthalene
Triphenylmethane 	• Near double degeneracy • Not stable in the protonated form • Becomes stable after deprotonation on the K2triphenylmethane ratio
Dibezanthracene 	• Large void space • Most stable ratio in the convex hull is a high metal and open-shell content ratio: K5dibenzanthracene
Decacyclene 	• Large void space • Exact LUMO degeneracy • Open shell content ratio

APPENDIX B - Chapter 4

B1. Intercalated PAHs systems

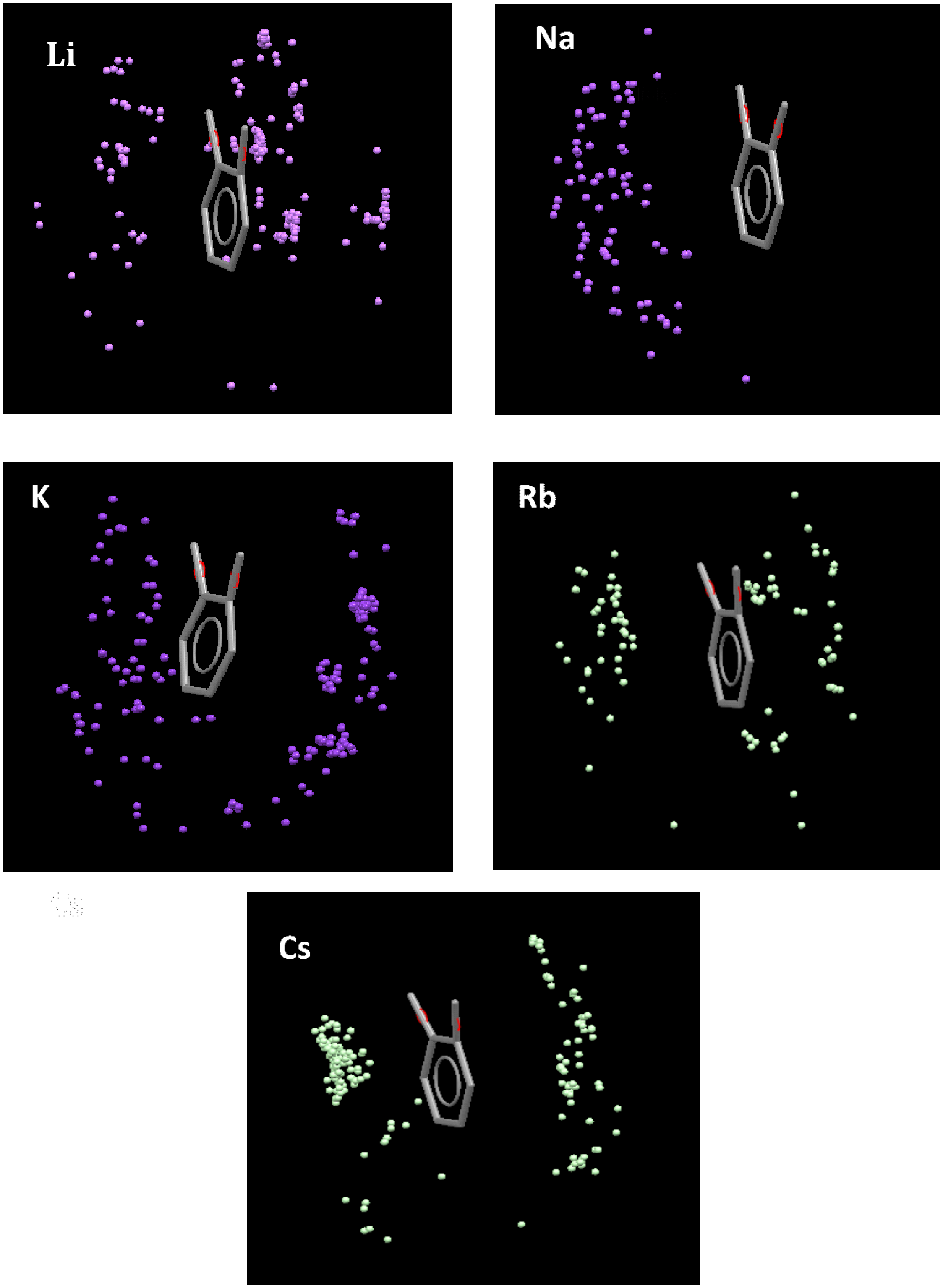


Figure B1.1 Scatterplots showing the distribution of the metals around the polyaromatic structure.

Benchmarking the VASP system

Identifying the optimal VASP parameters tailored for metal-polyaromatic hydrocarbons systems. The benchmarking system is K_2Tetracene (CSD id: MURLIX).

- An E_{cutoff} of 520eV was used in all the calculations based on previous work on tetracene.[190]
- KSPACING selection: KSPACING values = 0.1, 0.2, 0.3, 0.4, 0.5

Table B1.1. KSPACING selection

	KSPACING									
	0.1		0.2		0.3		0.4		0.5	
	Energy (eV)	kpts	Energy (eV)	kpts	Energy (eV)	kpts	Energy (eV)	kpts	Energy (eV)	kpts
K2 tetracene	-742.086	9 9 3	-742.086	5 5 2	-742.082	3 3 1	-742.082	3 3 1	-742.077	2 2 1
K	2.375	14 14 10	2.366	7 7 5	2.369	5 5 4	2.431	4 4 3	2.322	3 3 2
tetracene	-370.693	9 11 6	-370.693	5 6 3	-370.693	3 4 2	-370.693	3 3 2	-370.693	2 3 2
energy difference (eV/TU)	-2.550		-2.541		-2.543		-2.605		-2.495	

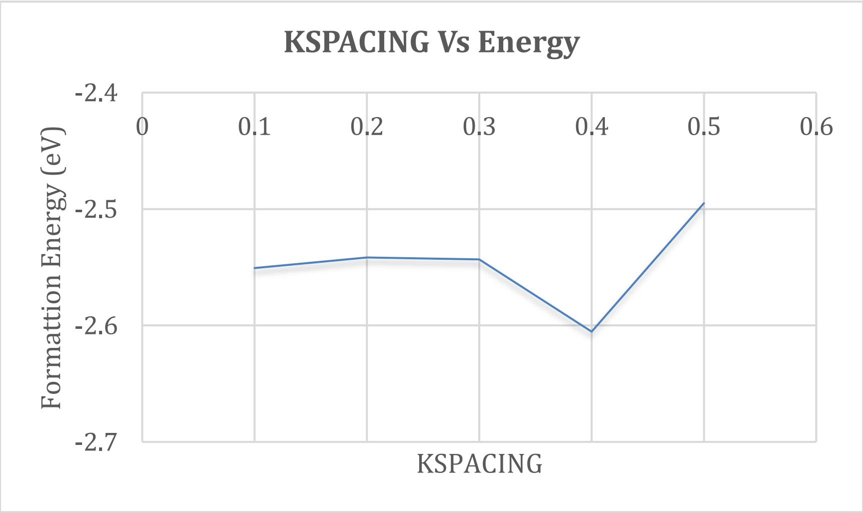


Figure B1.2. KSPACING versus Formation energy for the benchmark K_2tetracene system. The selected KSPACING for further investigating the alkali metal intercalated PAHs systems is **0.2**.

Lattice parameters comparison between simulated and experimental structures

We have simulated the crystal structure of pristine tetracene and found that the optimized lattice parameters with a have a perfect agreement with the experimental ones. The consistency indicates that the C and H pseudopotentials and the parameters selected in our calculations are reasonable, which is a reliable basis for exploring the doped PAHs structures.

Table B1.2. Testing the selected VASP on the experimental structure.

	Volume (Å^3)	Unit cell params
Experimental	1359.66	**a** 7.259 **b** 7.274 **c** 25.756 α 90 β 91.783 γ 90
Simulated	1303.036	**a** 7.13826 **b** 7.19650 **c** 25.38743 **α** 90 **β** 92.384 **γ** 90

B2 Convex hull analysis (simple intercalation approach)

Table B2.1. K$_x$bezanthracene calculations

a) structure	**Energy (E$_H$)**	**E$_K$**	**K ratio**	**Formation energy/molecule (eV/FU)**
Bezanthracene (2 molecules/unit cell)	-371.090		0	
K1 bezanthracene (2 molecules + 2 K/unit cell)	-368.981	2.368	0.5	-0.1295
K2 bezanthracene (2 molecules + 4 K/unit cell)	-369.519	4.736	0.667	-1.5825
K3 bezanthracene (2 molecules + 6 K/unit cell)	-366.391	7.104	0.75	-1.2025
K4 bezanthracene (2 molecules + 8 K/unit cell)	-363.640	9.472	0.8	-1.011
Potassium (K) (2 atoms/unit cell)	2.368	-	1	0

Table B2.2. K_xCoronene calculations

a) structure	Energy (E_H)	E_K	K ratio	Formation energy/molecule (eV/FU)
Pristine Coronene (2 molecules/unit cell)	-469.0016	-	0	0
K1coronene (2 molecules + 2 K/unit cell)	-466.2758	2.368	0.5	0.1789
K2coronene (2 molecules + 4K/unit cell)	-465.5109	4.736	0.667	-0.622
K3corone (2 molecules + 6K/unit cell)	-464.1465	7.104	0.75	-1.1244
AFM K3corone (2 molecules + 6K/unit cell)	-464.1696	7.104	0.75	-1.1364
FM K3corone (2 molecules + 6K/unit cell)	-464.1731	7.104	0.75	-1.1377
K4coronene (2 molecules + 8K/unit cell)	-461.1967	9.472	0.8	-0.8335

Table B2.3. Cs_xCoronene calculations

a) structure	Energy (E_H)	E_K	K ratio	Formation energy/molecule (eV/FU)
Cs1coronene (2 molecules + 2 Cs/unit cell)	-469.696	-0.6056	0.5	-0.0446
Cs2coronene (2 molecules + 4Cs/unit cell)	-471.84653	-1.2112	0.667	-0.8168
Cs3corone (2 molecules + 6Cs/unit cell)	-474.67504	-1.8168	0.75	-1.928
AFM Cs3corone (2 molecules + 6Cs/unit cell)	-474.670	-1.8168	0.75	-1.928
FM Cs3corone (2 molecules + 6Cs/unit cell)	-474.6744	-1.8168	0.75	-1.928
Cs4coronene (2 molecules + 8Cs/unit cell)	-475.567	-2.422	0.8	-2.0717
Cs5coronene (2 molecules + 10Cs/unit cell)	-476.413	-3.028	0.8333	-2.1802
Cs6coronene (2 molecules + 12Cs/unit cell)	- 477.117	-3.633	0.8571	-2.2409
Caesium (Cs) (4 atoms/unit cell)	-1.211	-	1	0

Table B2.4. NaxCoronene calculations

a) structure	Energy (E_H)	E_K	K ratio	Formation energy/molecule (eV/FU)
Na1coronene (2 molecules + 2 Na/unit cell)	-464.2308	5.881	0.5	-0.556
Na2coronene (2 molecules + 4Na/unit cell)	-459.5179	11.7638	0.667	-1.14
Na3corone (2 molecules + 6Na/unit cell)	-454.9041	17.6457	0.75	-1.7741
AFM Na3corone (2 molecules + 6Na/unit cell)	-454.904	17.6457	0.75	-1.7741
FM Na3corone (2 molecules + 6Na/unit cell)	-454.7670	17.6457	0.75	-1.7055
Na4coronene (2 molecules + 8Na/unit cell)	-448.42190	23.5276	0.8	-1.47395
Sodium (Na) (2 atoms/unit cell)	5.881	-	1	0

Table B2.5. K_xTriphenylene calculations

a) structure	Energy (E_H)	E_K	K ratio	Formation energy/molecule (eV/FU)
Pristine Triphenylene (4 molecules/unit cell)	-742.596	-	0	0
K1triphenylene (4 molecules + 4 K/unit cell)	-737.117	4.736	0.5	0.18575
K2triphenylene (4 molecules + 8K/unit cell)	-733.704	9.472	0.667	-0.145
K3triphenylene (4 molecules + 12K/unit cell)	-729.635	14.208	0.75	-0.3117
AFM ↑↓↑↓ K3triphenylene (4 molecules + 12K/unit cell)	-729.635	14.208	0.75	-0.3117
AFM ↑↓↓↑ K3triphenylene (4 molecules + 12K/unit cell)	-729.637	14.208	0.75	-0.3117
AFM ↑↑↓↓ K3triphenylene (4 molecules + 12K/unit cell)	-729.635	14.208	0.75	-0.3117
FM K3triphenylene (4 molecules + 12K/unit cell)	-729.480	14.208	0.75	-0.273
K4triphenylene (4 molecules + 16K/unit cell)	-724.227	18.944	0.8	-0.1437

Table B2.6 Cs_xTriphenylene calculations

a) structure	Energy (E_H)	E_K	K ratio	Formation energy/molecule (eV/FU)
Cs1triphenylene (4 molecules + 4 Cs/unit cell)	-744.6347	-1. 2112	0.5	-0.2068
Cs2triphenylene (4 molecules + 8 Cs/unit cell)	-747.7577	-2.4224	0.667	-0.6848
Cs3triphenylene (4 molecules + 12 Cs/unit cell)	-750.5702	-3.6337	0.75	-1.0851
Cs4triphenylene (4 molecules + 16 Cs/unit cell)	-753.4007	-4.8449	0.8	-1.4899

Table B2.7 Na_xTriphenylene calculations

a) structure	Energy (E_H)	E_K	K ratio	Formation energy/molecule (eV/FU)
Na1triphenylene (4 molecules + 4 Na/unit cell)	-731.962	11.764	0.5	-0.282
Na2triphenylene (4 molecules + 8 Na/unit cell)	-722.059	23.528	0.667	-0.747
Na3triphenylene (4 molecules + 12 Na/unit cell)	-711.761	35.292	0.75	-1.114
Na4triphenylene (4 molecules + 16 Na/unit cell)	-699.133	47.056	0.8	-0.898

Table B2.8. Corannulene calculations

structure	Energy (E_H)	E_K	K ratio	Formation energy/molecule (eV/FU)
Corannulene (8molecules /unit cell)	-1543.0703		0	
K1corann (8 molecules + 8 Na/unit cell)	-1538.8109	9.472	0.5	-0.6515
K2corann (8 molecules + 26 Na/unit cell)	-1535.377	18.944	0.667	-1.406
K3corann (8 molecules + 24 Na/unit cell)	-1529.818	28.416	0.75	-1.895
AFM ↑↑↓↓↑↑↓↓ K3corann (8 molecules + 24 Na/unit cell)	-1529.8186	28.416	0.75	-1.895
AFM ↑↓↑↓↑↓↑↓ K3corann (8 molecules + 24 Na/unit cell)	-1529.8186	28.416	0.75	-1.895
NUPDOWN=8 FM K3corann (8 molecules + 24 Na/unit cell)	-1529.7290	28.416	0.75	-1.884
K4corann (8 molecules + 32 Na/unit cell)	-1522.549	37.888	0.8	-2.170
K5corann (8 molecules + 40 Na/unit cell)	-1511.235	47.36	0.8333	-1.9407

Table B2.9. Hexabenzocoronene non-spin polarized calculations (2 molecules in the structure)

structure	Energy (E_H)	E_K	K ratio	Formation energy/molecule (eV/FU)
Hbc (2 molecules/unit cell)	-799.6850		0	
K1hbc (2 molecules + 2 K/unit cell)	-797.9577	2.368	0.5	-0.3203
K2hbc (2 molecules + 4 K/unit cell)	-796.3361	4.736	0.667	-0.6935
K3hbc (2 molecules + 4 K/unit cell)	-795.6617	7.104	0.75	-1.5403
AFM ↑↓ K3hbc (2 molecules + 4 K/unit cell)	-795.6635	7.104	0.75	-1.5403
K4hbc (2 molecules + 4 K/unit cell)	-794.0737	9.472	0.8	-1.9303
K5hbc (2 molecules + 4 K/unit cell)	-792.5760	11.84	0.8333	-2.3655
K6hbc (2 molecules + 4 K/unit cell)	-791.3432	14.208	0.8571	-2.9332

Table B2.10. Decacyclene calculations

structure	Energy (E_H)	E_K	K ratio	Formation energy/molecule (eV/FU)
Decacyclene (4 molecules/unit cell)	-1395.6204		0	
K1 decacyclene (4 molecules + 4 K/unit cell)	-1392.5894	4.736	0.5	-0.4262
K2 decacyclene (4 molecules + 8 K/unit cell)	-1391.3485	9.472	0.667	-1.30004
K3 decacyclene (4 molecules + 12 K/unit cell)	-1391.1404	14.208	0.75	-2.43200
K4 decacyclene (4 molecules + 16 K/unit cell)	-1387.4909	18.944	0.8	-2.70362

Table B2.11. perylene calculations

structure	Energy (E_H)	E_K	K ratio	Formation energy/molecule (eV/FU)
Perylene (4 molecules/unit cell)	-806.5425		0	
K1 perylene (4 molecules + 4 K/unit cell)	-805.0967	4.736	0.5	-0.8225
K2 perylene (4 molecules + 8 K/unit cell)	-803.0941	9.472	0.667	- 1.50588
K3 perylene (4 molecules + 12 K/unit cell)	-799.72574	14.208	0.75	-1.855
K4 perylene (4 molecules + 16 K/unit cell)	-793.96904984	18.944	0.8	-1.592

Table B2.12. dibenz[a,c]anthracene non-spin polarized calculations (4 molecules in the structure)

structure	Energy (E_H)	E_K	K ratio	Formation energy/molecule (eV/FU)
dibenz[a,c]anthracene (4 molecules/unit cell)	-897.9260293		0	
K1 dibenz[a,c]anthracene (4 molecules + 4 K/unit cell)	-893.954	4.736	0.5	-0.087
K2 dibenz[a,c]anthracene (4 molecules + 8 K/unit cell)	-892.0956329	9.472	0.667	-0.9104
K3 dibenz[a,c]anthracene (4 molecules + 12 K/unit cell)	-887.7189869	14.208	0.75	-1.000
K4 dibenz[a,c]anthracene (4 molecules + 16 K/unit cell)	-883.9205887	18.944	0.8	-1.2346
K5 dibenz[a,c]anthracene (4 molecules + 20 K/unit cell)	-879.3438252	23.68	0.8333	-1.274

Table B2.13. 1,1'-Binaphthalene

structure	Energy (E_H)	E_K	K ratio	Formation energy/molecule (eV/FU)
1,1'-Binaphthalene (4 molecules/ unit cell)	-833.1501		0	
K1 1,1'-Binaphthalene (4 molecules + 4 K/unit cell)	-829.9362	4.736	0.5	-0.3805
K2 1,1'-Binaphthalene (4 molecules + 8 K/unit cell)	-826.9562	9.472	0.667	-0.8195
K3 1,1'-Binaphthalene (4 molecules + 12 K/unit cell)	-821.4434	14.208	0.75	-0.62533
K4 1,1'-Binaphthalene (4 molecules + 16 K/unit cell)	-817.6482	18.944	0.8	-0.86054

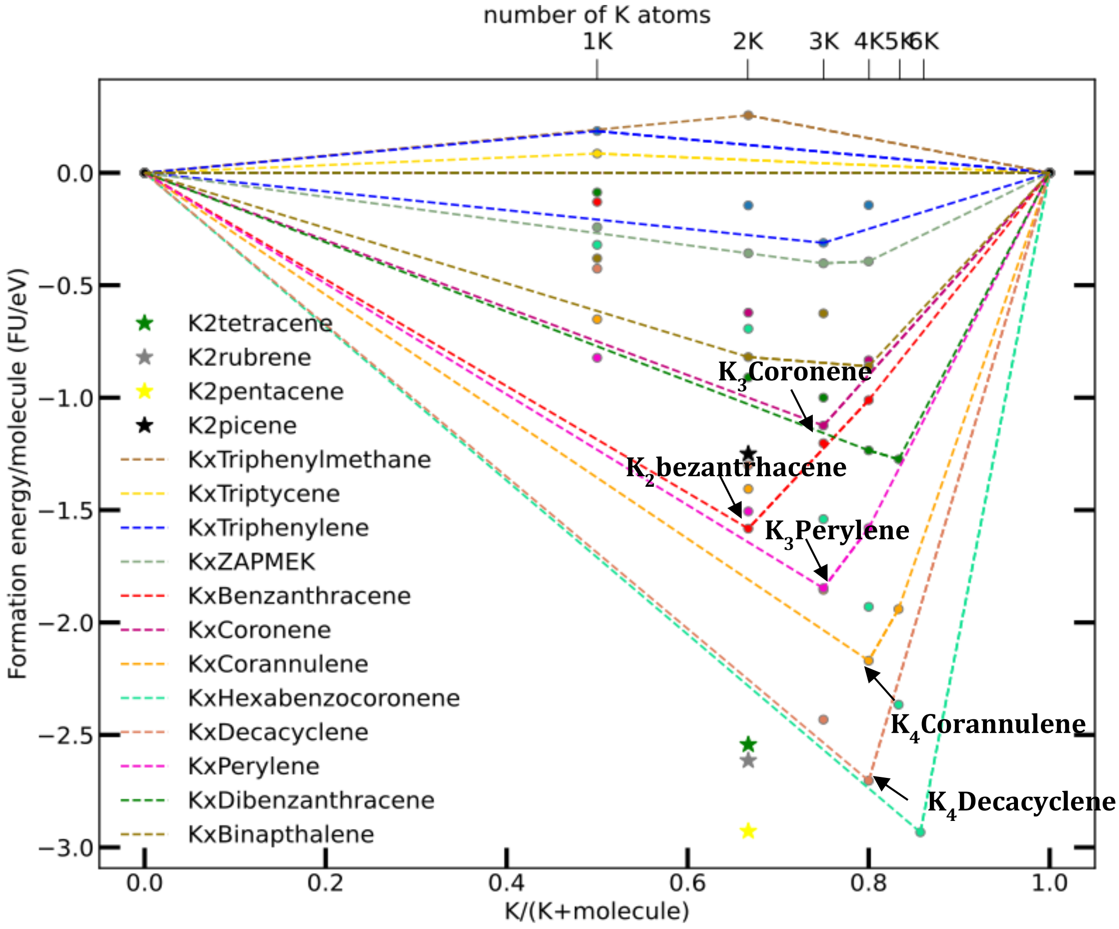


Figure B2.1. Detailed convex hull of all the PAHs structures that have been analysed with the simple intercalation approach. The most interesting ratios are highlighted in the plot.

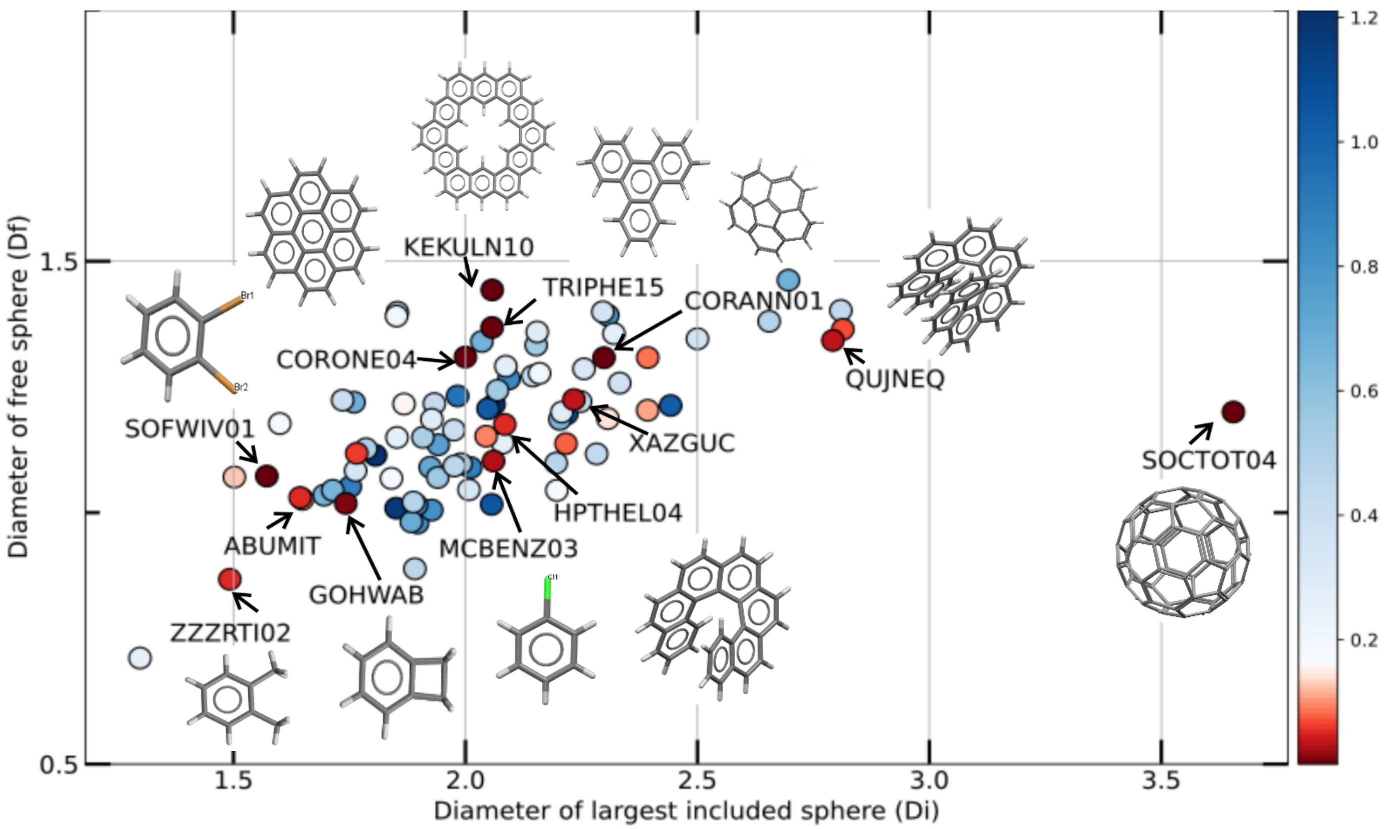
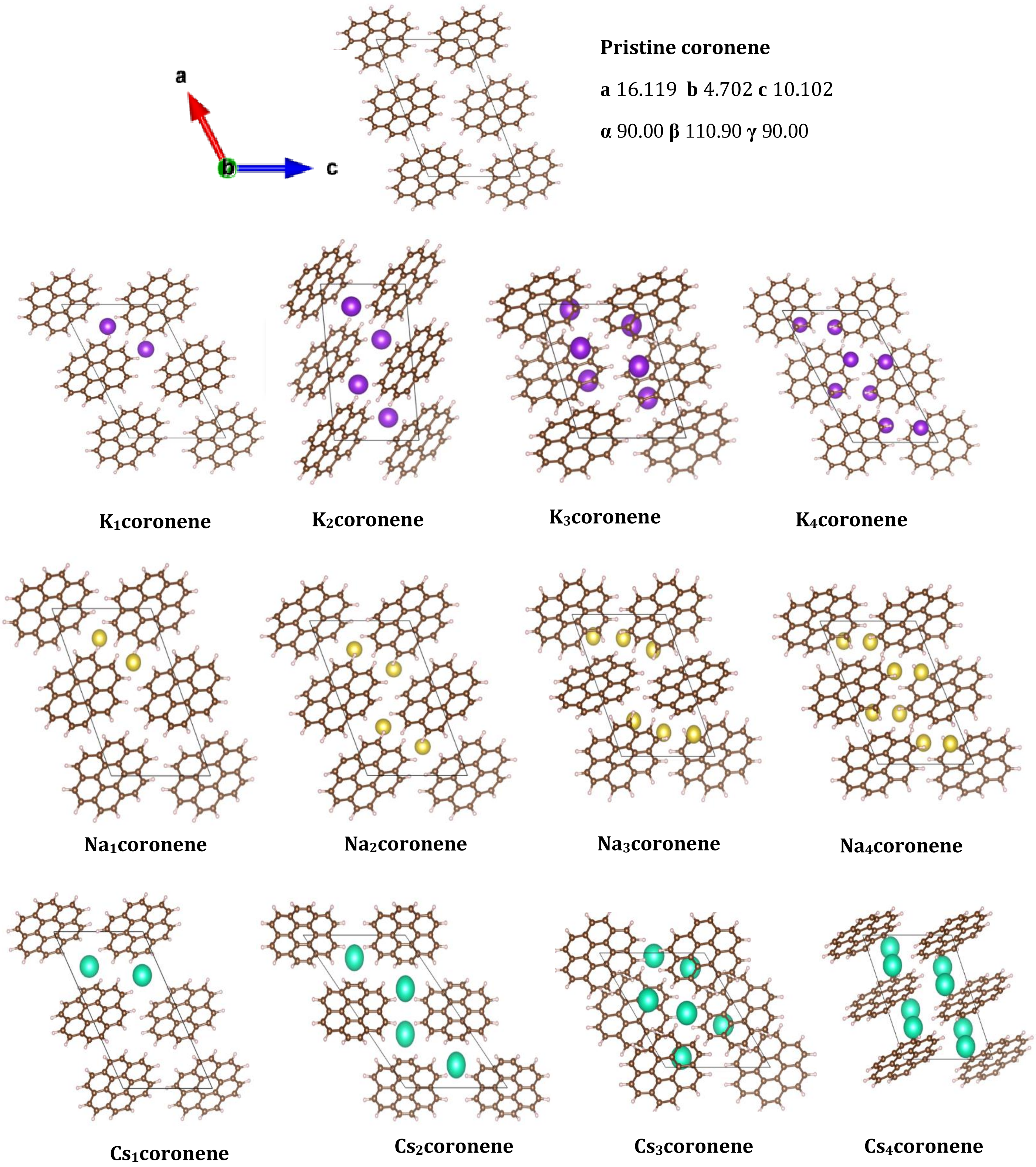


Figure B2.2. Scatterplot showing all the molecules identified with exact LUMO-LUMO+1 degeneracy.

Table B2.14. Unit cell parameters for the intercalated coronene phases

	x=1	**x=2**	**x=3**	**x=4**
K$_x$coronene	**a** 16.58 **b** 4.59 **c** 10.72 **α** 89.99 **β** 116.94 **γ** 90.00	**a** 14.63 **b** 6.63 **c** 7.91 **α** 90.01 **β** 95.04 **γ** 90.02	**a** 11.62 **b** 7.79 **c** 9.74 **α** 91.63 **β** 108.04 **γ** 92.90	**a** 17.71 **b** 5.86 **c** 10.11 **α** 90.10 **β** 118.21 **γ** 89.95
Na$_x$coronene	**a** 15.33 **b** 4.26 **c** 9.84 **α** 89.99 **β** 112.58 **γ** 90.00	**a** 15.23 **b** 4.25 **c** 9.45 **α** 90.00 **β** 111.02 **γ** 90.00	**a** 12.70 **b** 5.15 **c** 9.53 **α** 87.60 **β** 113.37 **γ** 89.84	**a** 13.60 **b** 4.84 **c** 9.54 **α** 90.01 **β** 115.27 **γ** 89.99
Cs$_x$coronene	**a** 16.84 **b** 5.11 **c** 9.98 **α** 89.99 **β** 117.42 **γ** 89.99	**a** 16.95 **b** 5.51 **c** 11.63 **α** 89.88 **β** 133.41 **γ** 90.06	**a** 12.22 **b** 8.12 **c** 10.86 **α** 89.99 **β** 123.92 **γ** 90.02	**a** 12.49 **b** 9.08 **c** 10.00 **α** 89.96 **β** 111.83 **γ** 90.00



Pristine coronene

a 16.119 **b** 4.702 **c** 10.102

α 90.00 **β** 110.90 **γ** 90.00

K_1coronene **K_2coronene** **K_3coronene** **K_4coronene**

Na_1coronene **Na_2coronene** **Na_3coronene** **Na_4coronene**

Cs_1coronene **Cs_2coronene** **Cs_3coronene** **Cs_4coronene**

Figure B2.3. Coronene theoretical structure modification after the insertion of K, Na and Cs. It can be observed that Cs causes the higher deformation of the initial structure. The structures are viewed along the b axis.

B3. Crystal Structure Prediction – Benchmarking system K_2tetracene (csd id: MURLIX)

The CSP method was first tested in a known system, namely K_2tetracene (csd id: MURLIX) where the experimental crystal structure is determined. USPEX software with VASP and PBE + D3 corrections was implemented trying to predict the experimental structure, given the correct number of compounds in the structure, *i.e.*, 4 tetracene molecules and 8 K atoms. The relaxation consists of five steps with increasing accuracy and kspacing.

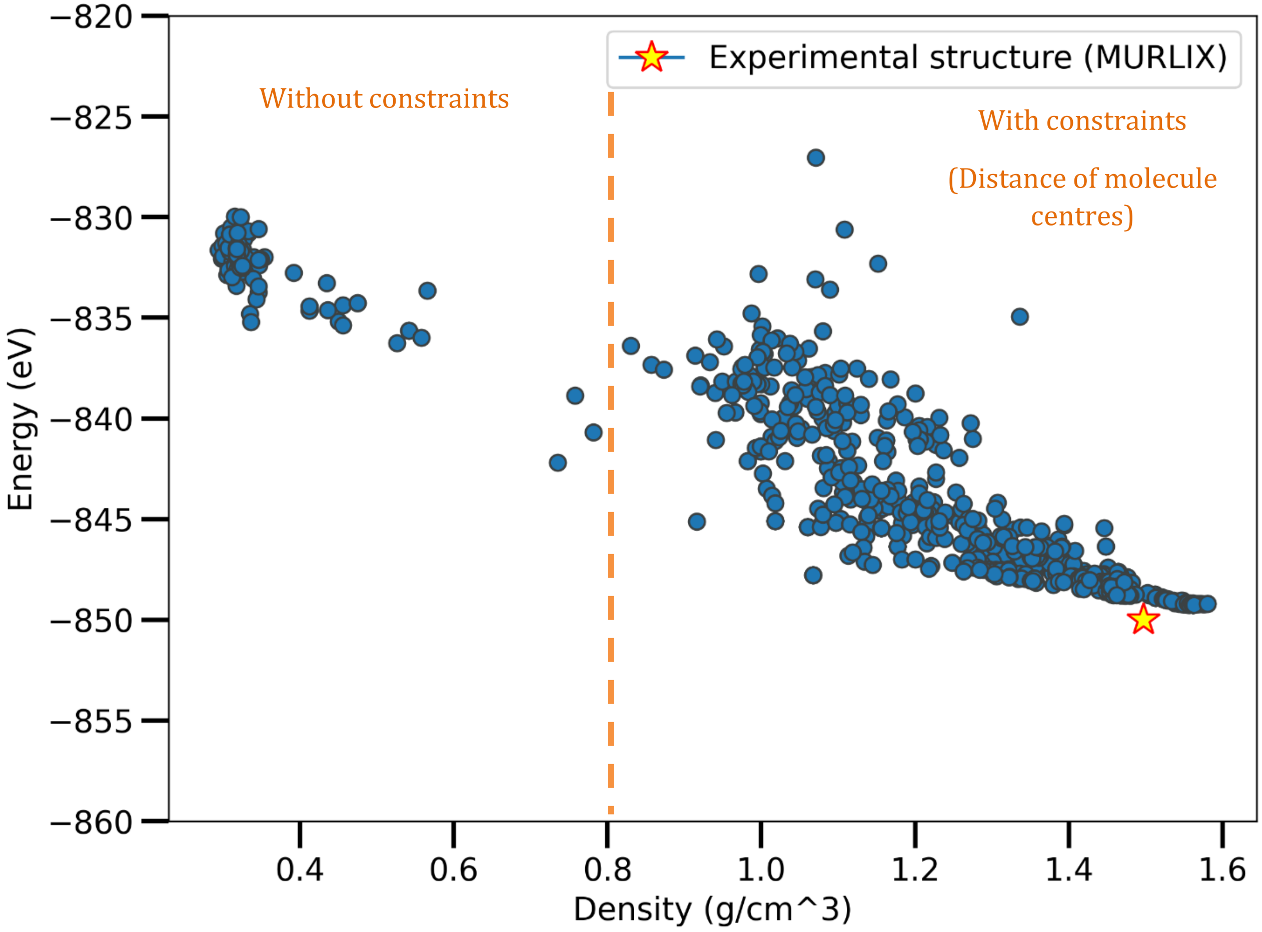


Figure B3.1. Energy-density scatterplot showing the crystal structures generated using USPEX with VASP. Herein, the only initial constraints used were the distance of the molecular centres, which was set according to the observations on the known metal-PAHs systems. The effect of the constraints in the generation of more sensible structures is demonstrated, as without the constraints the generated structures have a very small density and it will take much computational time until reaching the lower density configurations. On the other hand, when starting with the constraints all the generated structures have densities above 0.8. The generated structures are close to the experimental. However, the experimental is still lower in energy. This process required 6 weeks' time on a supercomputer for the generation of 1,000 structures.

To overcome the limitations of time, semi-empirical DFT-based method was next used, namely DFTB+ for the structural relaxations and energy calculations. Although the calculations can be completed in a significantly lower time, *i.e.*, 1,500 structures in two weeks' time, the exact known experimental structure was not found.

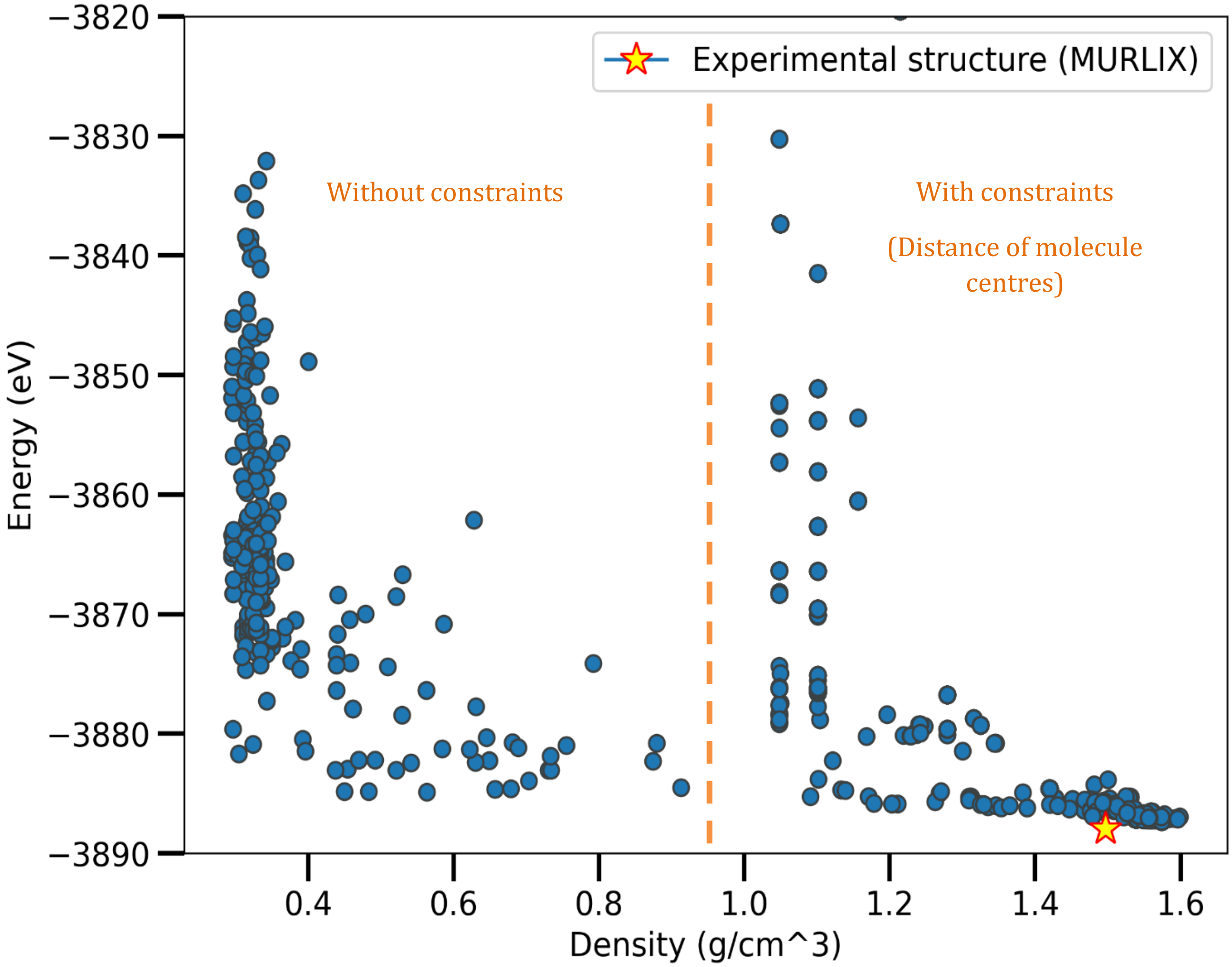


Figure B3.2. Energy-density scatterplot showing the crystal structures that were generated using USPEX with DFTB+. The effect of the initial constraints in the generation of more sensible structures is shown here. The generated structures are close to the experimental. However, the experimental is still lower in energy.

The final approach that was tested regarding K2etracene was to start generating structures given the known unit cell parameters.

Table B3.1. Testing starting parameters for reproducing K2tetracene experimental structure.

Parameters		Num of generations
- 8 K + 4 tetracene - Molcenters distance	- Correct unit cell params: **a** 7.25 **b** 7.27 **c** 25.75 **α** 90 **β** 91.78 **γ** 90 - Space group: P21/c	No structure was generated
	- Correct unit cell params: **a** 7.25 **b** 7.27 **c** 25.75 **α** 90 **β** 91.78 **γ** 90 - No spacegroup	No structure was generated
	- Correct unit cell params: **a** 7.25 **b** 7.27 **c** 25.75 - No angles - No spacegroup	No structure was generated
	- Unit cell lengths rounded up: **a** 8 **b** 8 **c** 26 - No angles - No spacegroup	Experimental structure found on the second generation

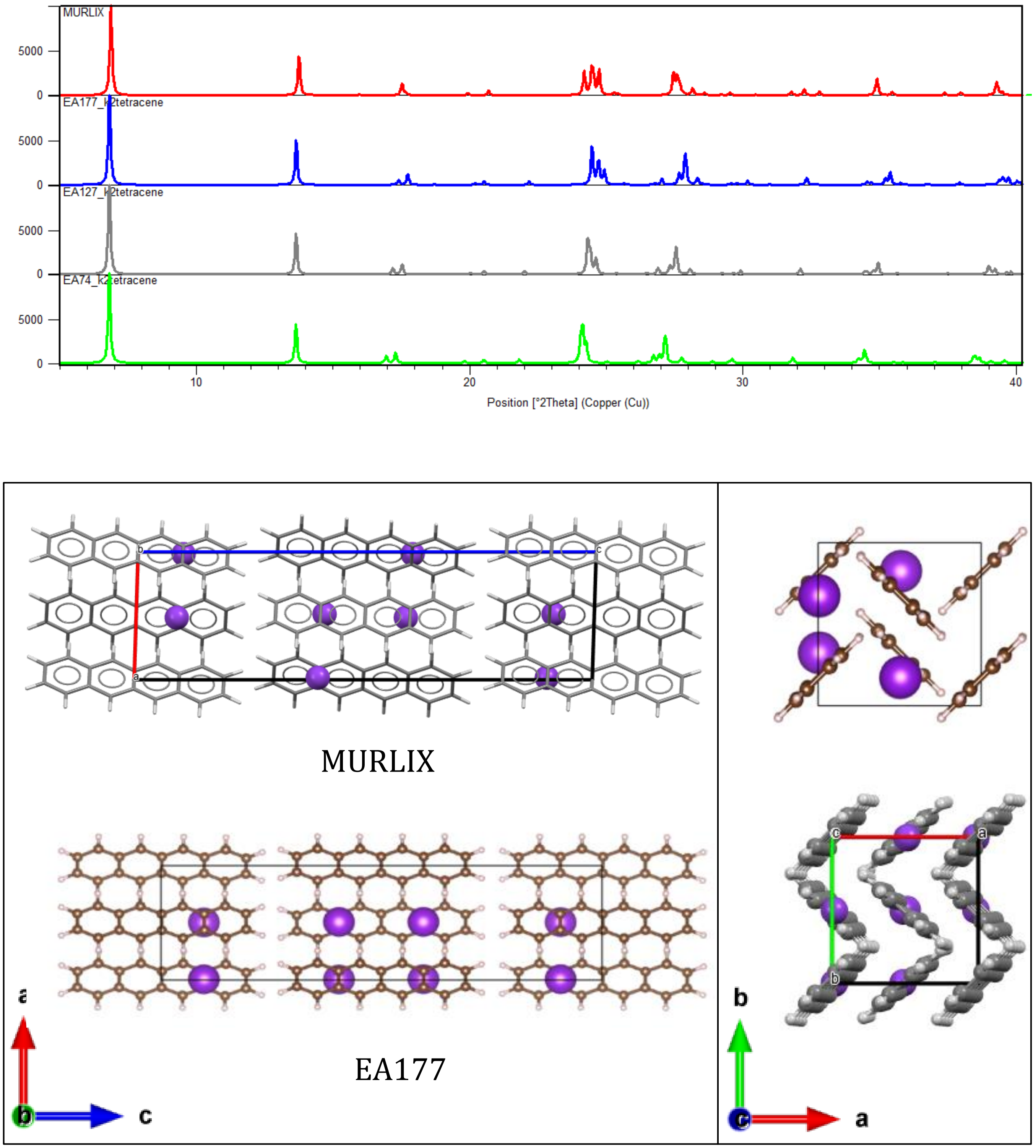
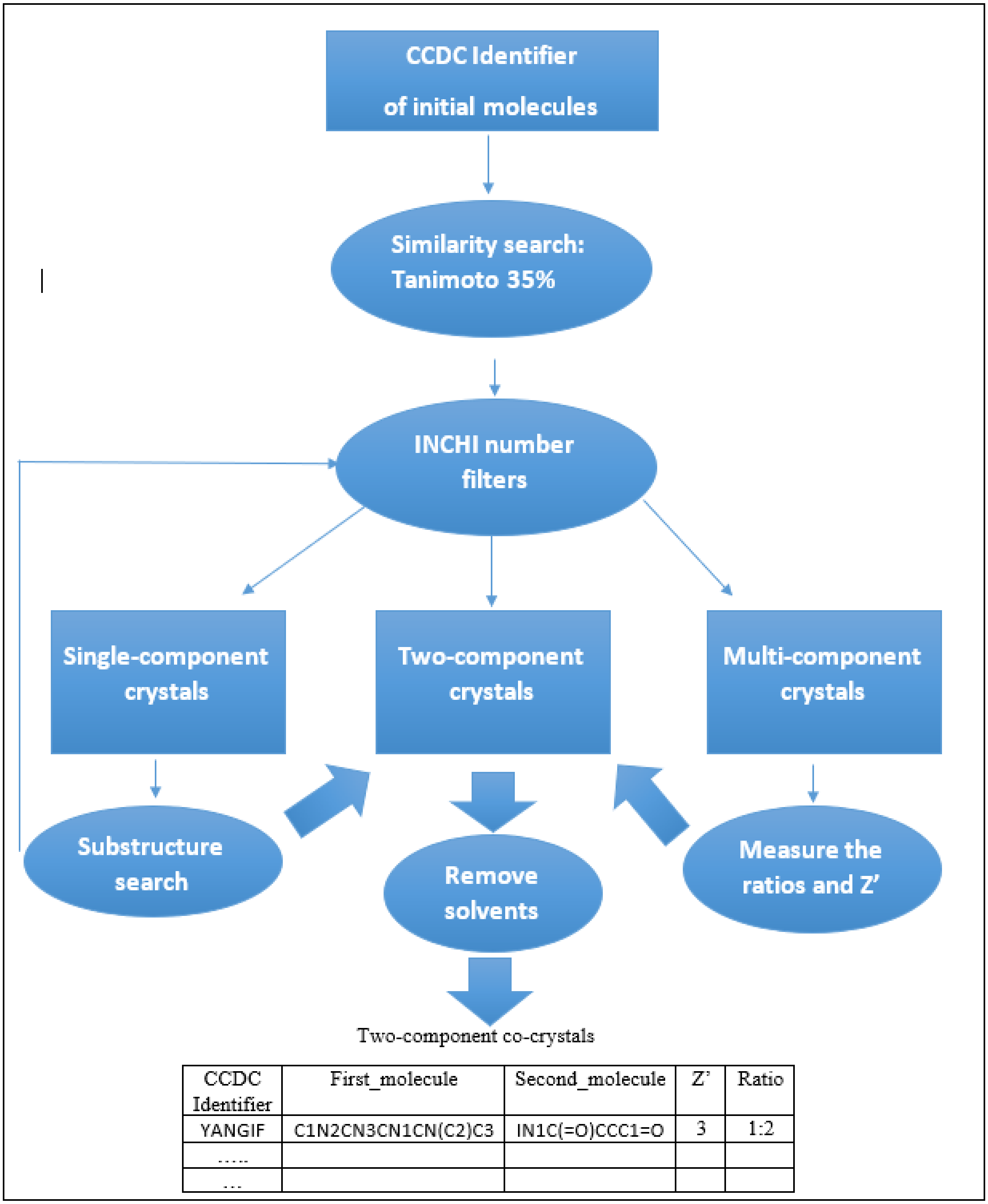


Figure B3.3. PXRD pattern comparison of the experimental structure (MURLIX) and the USPEX generated structures given as initial constraints the distance between the molecule centers and the correct unit cell lengths.

Table B3.2. Comparison of the known experimental structure (MURLIX) with the USPEX generated structures.

Structures	**Unit cell**	**Volume**	**Spacegroup**	**Density (g cm-3)**	**Energy (eV)**
MURLIX	**a** 7.25 **b** 7.27 **c** 25.75 **α** 90 **β** 91.78 **γ** 90	1359.66	P 21/c	1.497	-742.086
EA177 (given the lattice params)	**a** 7.19 **b** 7.21 **c** 25.94 **α** 90 **β** 90 **γ** 90	1346.337	P 21/c	1.512	-742.063
EA1280_dftb+ **(no initial lattice params)**	**a** 9.63 **b** 12.68 **c** 10.59 **α** 89.99 **β** 89.93 **γ** 90.00	1295.42	P1	1.574	-741.284
EA845_vasp **(no initial lattice params)**	**a** 11.12 **b** 7.42 **c** 15.07 **α** 89.96 **β** 90.01 **γ** 90.0	1245.56	P1	1.561	-741.672

APPENDIX C - Chapter 5



CCDC Identifier	First_molecule	Second_molecule	Z’	Ratio
YANGIF	C1N2CN3CN1CN(C2)C3	IN1C(=O)CCC1=O	3	1:2
…..				
…				

Figure C1.1. Flow diagram for PAH co-crystals extraction. The search starts with 8 representative PAHs and Python API CCDC is employed for extracting all the co-crystals that are formed from these 8 molecules or molecules that are similar to them on the basis of molecular fingerprints (ECFP4 > 0.35 Tanimoto Similarity). The extracted dataset was further filtered for removing co-crystals containing molecules with acidic parts.

Table C1.2. Initial Polyaromatic Hydrocarbons (PAHs) for co-crystals extraction.

Dragon Descriptor	Description	Pearson Correlation	Spearman Correlation	p-value
nBT	molecular weight	0.403	0.620	$< 10^{-5}$
nHet	number of heteroatoms	0.515	0.685	
ZM1V	first Zagreb index by valence vertex degrees	0.528	0.729	
DBI	Dragon branching index	0.548	0.654	
ICR	radial centric information index	0.546	0.422	
MAXDN	maximal electrotopological negative variation	0.440	0.600	
MAXDP	maximal electrotopological positive variation	0.426	0.626	
DELS	molecular electrotopological variation	0.414	0.629	
CIC0	Complementary Information Content index (neighborhood symmetry of 0-order)	0.298	0.515	
J_D/Dt	Balaban-like index from distance/detour matrix	0.323	0.424	
SM1_Dz(Z)	spectral moment of order 1 from Barysz matrix weighted by atomic number	0.551	0.627	
SM1_Dz(v)	spectral moment of order 1 from Barysz matrix weighted by van der Waals volume	0.404	0.479	
SM1_Dz(e)	spectral moment of order 1 from Barysz matrix weighted by Sanderson electronegativity	0.480	0.558	
HyWi_B(s)	hyper-Wiener-like index (log function) from Burden matrix weighted by I-State	0.744	0.682	
SpMax4_Bh(m)	largest eigenvalue n. 4 of Burden matrix weighted by mass	0.541	0.571	
SpMax3_Bh(s)	largest eigenvalue n. 3 of Burden matrix weighted by I-state	0.422	0.482	
SpMax7_Bh(s)	largest eigenvalue n. 7 of Burden matrix weighted by I-state	0.439	0.542	
P_VSA_v_2	P_VSA-like on van der Waals volume, bin 2	0.501	0.684	
P_VSA_s_6	P_VSA-like on I-state, bin 6	0.522	0.704	
Eta_F_A	eta average functionality index	0.434	0.438	
Eig02_AEA(dm)	eigenvalue n. 2 from augmented edge adjacency mat. weighted by dipole moment	0.530	0.539	
Eig03_AEA(dm)	eigenvalue n. 3 from augmented edge adjacency mat. weighted by dipole moment	0.609	0.572	
nHAcc	number of acceptor atoms for H-bonds (N,O,F)	0.449	0.620	
Uc	unsaturation count	0.520	0.551	

Filtering with Pipeline Pilot

The filtering for incompatible functional groups in both the labelled and unlabelled dataset was performed using Pipeline Pilot[165] with the following workflow.

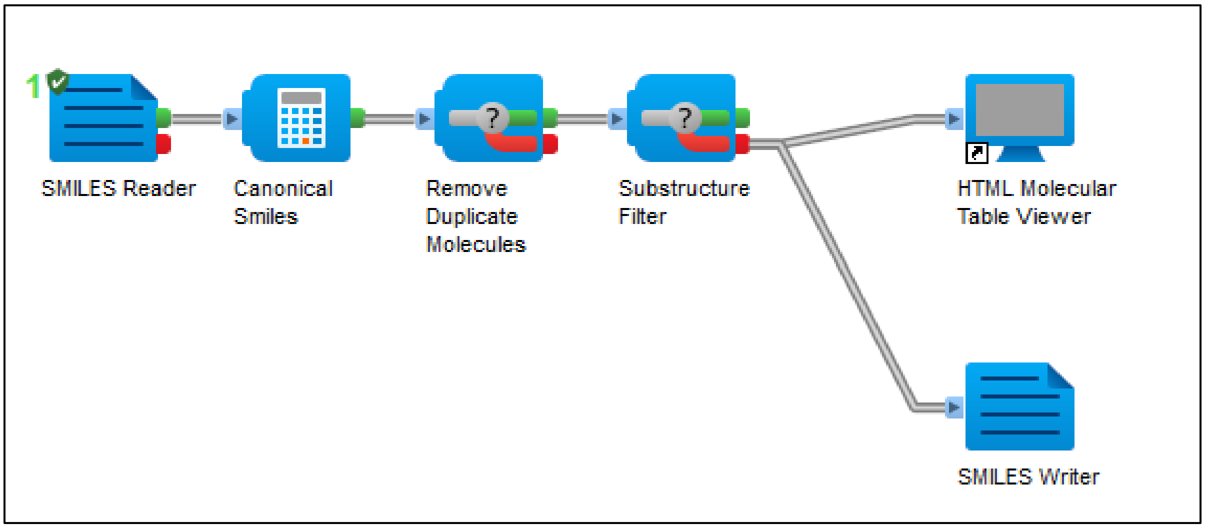
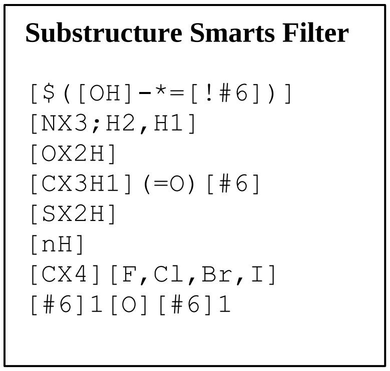


Figure C1.2. Pipeline Pilot workflow.



Substructure Smarts Filter

```
[$([OH]-*=[!#6])]
[NX3;H2,H1]
[OX2H]
[CX3H1](=O)[#6]
[SX2H]
[nH]
[CX4][F,Cl,Br,I]
[#6]1[O][#6]1
```

Figure C1.3. Substructure SMARTS[227] filter for detecting the molecular combinations with at least one molecule with acidic hydrogens.

C2. Results

Table C2.1 Descriptors correlated to the descriptors identified as important for the decisions of the deep learning model. The correlation between the descriptors follows a previously reported method.[303]

Descriptor	Correlated Descriptors	Correlation	Description	Related Physical Meaning
B06[C-C]	B07[C-C]	0.857434	Presence/absence of C - C at topological distance 7	atom pairs descriptors that describe pairs of atoms and bond types connecting them in 2D space
	B05[C-C]	0.812225	Presence/absence of C - C at topological distance 5	atom pairs descriptors that describe pairs of atoms and bond types connecting them in 2D space
ATS6i	ATS6e	0.998216	Broto-Moreau autocorrelation of lag 6 (log function) weighted by Sanderson electronegativity	electronegativity
	ATS5e	0.983335	Broto-Moreau autocorrelation of lag 5 (log function) weighted by Sanderson electronegativity	electronegativity
	ATS5i	0.981890	Broto-Moreau autocorrelation of lag 5 (log function) weighted by ionization potential	ionization potential
	SpMax8_Bh (i)	0.928269	largest eigenvalue n. 8 of Burden matrix weighted by ionization potential	Ionization potential
	SpMax8_Bh (p)	0.923641	largest eigenvalue n. 8 of Burden matrix weighted by polarizability	polarizability
	ATS8e	0.927747	Broto-Moreau autocorrelation of lag 8 (log function) weighted by Sanderson electronegativity	electronegativity
	Vx	0.913402	McGowan volume	shape
	Si	0.945914	sum of first ionization potentials (scaled on Carbon atom)	Ionization potential
	Se	0.940544	sum of atomic Sanderson electronegativities (scaled on Carbon atom)	electronegativity
	nBT	0.934793	number of bonds	general
	Sp	0.923744	sum of atomic polarizabilities (scaled on Carbon atom)	polarizability
	Sv	0.913610	sum of atomic van der Waals volumes (scaled on Carbon atom)	shape

	IAC	0.900917	total information index on atomic composition	composition
	S1K	0.887118	1-path Kier alpha-modified shape index	Shape
	Eta_epsi	0.875800	eta electronegativity measure	electronegativity
	SAtot	0.871258	total surface area from P_VSA-like descriptors	polarity
	Pol	0.863927	polarity number	polarity
	nSK	0.853433	number of non-H atoms	general
	MW	0.828710	Molecular weight	general
Eig06_AEA (dm):	Eig05_AEA(dm)	0.956601	eigenvalue n. 5 from augmented edge adjacency mat. weighted by dipole moment	dipole moment
	Eig7_AEA(dm)	0.938136	eigenvalue n. 7 from augmented edge adjacency mat. weighted by dipole moment	dipole moment
	Eig08_AEA(dm)	0.918267	eigenvalue n. 8 from augmented edge adjacency mat. weighted by dipole moment	dipole moment
	Ram	0.792930	Ramification	branching
	Eta_B	0.778573	eta branching index	Shape
ChiA_Dz(p)	SpMaxA_B(p)	0.910006	normalized leading eigenvalue from Burden matrix weighted by polarizability	polarizability
	WiA_B(p)	0.908640	average Wiener-like index from Burden matrix weighted by polarizability	polarizability
	ChiA_Dz(e)	0.901665	average Randic-like index from Barysz matrix weighted by Sanderson electronegativity	electronegativity
	UNIP	0.933653	unipolarity	Polarity
	Sv	0.822757	sum of atomic van der Waals volumes (scaled on Carbon atom)	shape
	MW	0.822103	Molecular weight	molecular weight
	VvdwMG	0.819518	van der Waals volume from McGowan volume	Shape
	Vx	0.819518	McGowan volume	shape
	Si	0.815686	sum of first ionization potentials (scaled on Carbon atom)	Ionization potential
	Pol	0.805521	polarity number	polarity
	Sp	0.795808	sum of atomic polarizabilities (scaled on Carbon atom)	polarizability

SpMin5_Bh (s)	ATS3i	0.921903	Broto-Moreau autocorrelation of lag 3 (log function) weighted by ionization potential	ionization potential
	ATS3e	0.917570	Broto-Moreau autocorrelation of lag 3 (log function) weighted by Sanderson electronegativity	electronegativity
	SpMin5_Bh (e)	0.915201	smallest eigenvalue n. 5 of Burden matrix weighted by Sanderson electronegativity	electronegativity
	Sv	0.898829	sum of atomic van der Waals volumes (scaled on Carbon atom)	shape
	Sp	0.895652	sum of atomic polarizabilities (scaled on Carbon atom)	polarizability
	Si	0.882950	sum of first ionization potentials (scaled on Carbon atom)	Ionization potential
	Se	0.881810	sum of atomic Sanderson electronegativities (scaled on Carbon atom)	electronegativity
	Vx	0.878079	McGowan volume	shape
	VvdwMG	0.878079	van der Waals volume from McGowan volume	shape
	MW	0.803832	Molecular weight	molecular weight
	Ram	0.800056	Ramification	shape
Eig06_EA(bo)	Pol	0.888838	Polarity number	polarity
	CSI	0.887028	eccentric connectivity index	shape
	UNIP	0.871951	unipolarity	polarity
	Sv	0.859414	sum of atomic van der Waals volumes (scaled on Carbon atom)	shape
	MW	0.834828	Molecular weight	general
	Ram	0.831023	Ramification	branching
	Vx	0.818124	van der Waals volume from McGowan volume	shape
	VvdwMG	0.818124	van der Waals volume from McGowan volume	Shape
	Sp	0.811851	sum of atomic polarizabilities (scaled on Carbon atom)	polarizability

C3. Electronic Structure Calculations

- **C60 co-crystals: C60-triphenylphosphineoxide**

C60-triphenylphosphineoxide was used as the benchmarking system for selecting the vasp parameters.

➢ Energy cut-off selection: As we have carbon in our structure, the default cutoff will be 400 eV and usually we need to multiply this by at least 4/3 for unit cell optimisations. So, the testing cut-offs are 550, 600, 650.

➢ KSPACING selection: The selected KSPACING parameters are 0.5, 0.4, 0.3, 0.2, 0.1. We should start from the faster ones (KSPACING=0.5) and then do the slower ones starting from the end point of the other calculations. So, for the KSPACING=0.4 we are going to use the CONTCAR file from the final converged structure with KSPACING=0.5 and so on when progressing to lower KSPACING parameters.

KSPACING	Ecut	Energy
0.5 (2 1 1)	550	-677.509
	600	-677.865
	650	-678.305
0.4 (2 2 2)	550	-677.531
	600	-677.88
	650	-678.305
0.3 (3 2 2)	550	-677.516
	600	-677.864
	650	-678.287
0.2 (4 3 3)	550	-677.521
	600	-677.867
	650	-678.291
0.1 (7 5 5)	550	-677.52
	600	-677.867
	650	-678.29

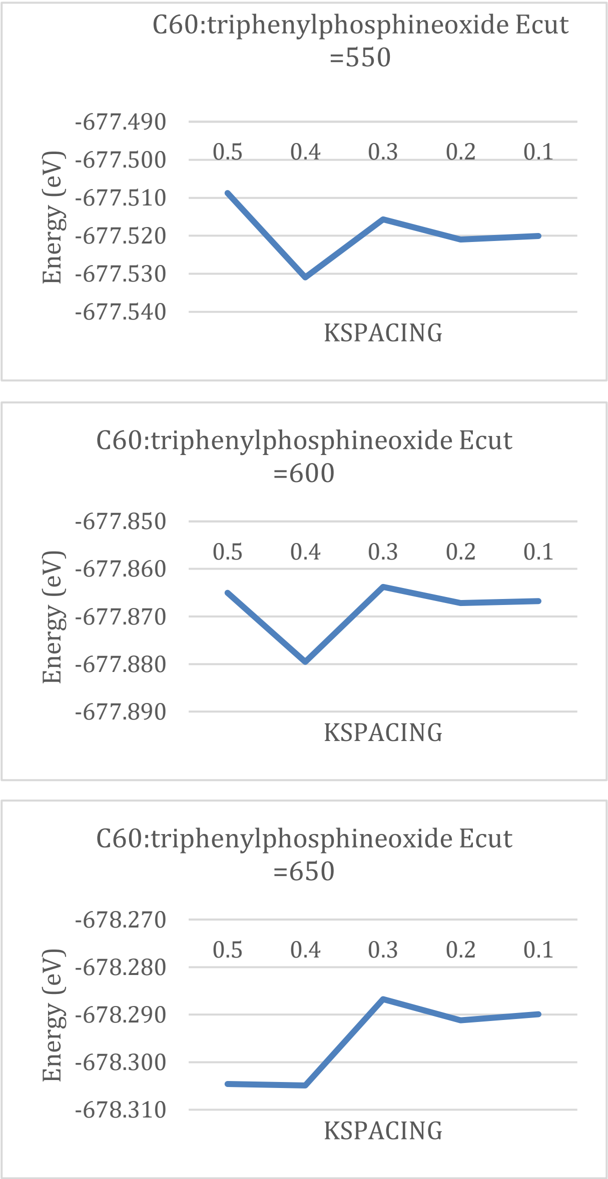
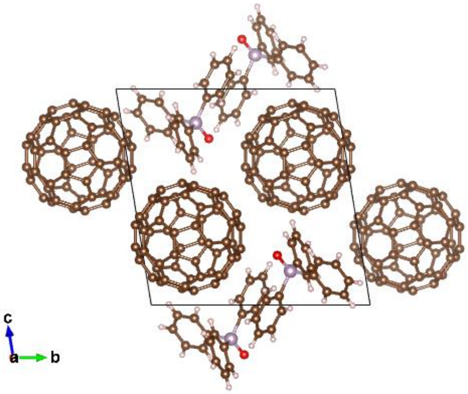
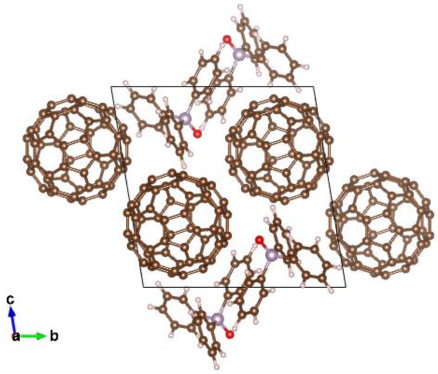
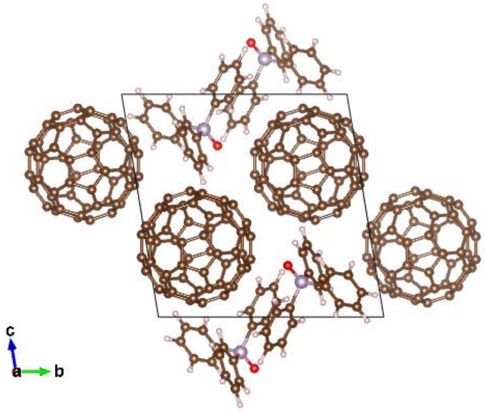


Figure C3.1. Convergence check for the C60 co-crystal systems using SCAN + rVV10. The parameters that gave the best convergence were KSPACING=0,2 and Ecut=600. These are going to be further used when studying similar systems.

Table C3.1: Comparing the experimental structure with the VASP optimized

	Volume (Å^3)	Unit cell params	Shape
Experimental	2012.3840	**lengths** 9.993 14.276 14.719 **angles** 99.002 103.842 90.015	
SCAN + rVV10 **Energy:** -677.910 eV	1864.4585	**lengths** 9.763 13.877 14.362 **angles** 98.976 103.935 89.734	
PBE + D3 **Energy:** -1520.3896	2004.765070	10.001 14.250 14.687 99.148 103.884 89.960	

Appendix D - Chapter 6

D1. Data preparation

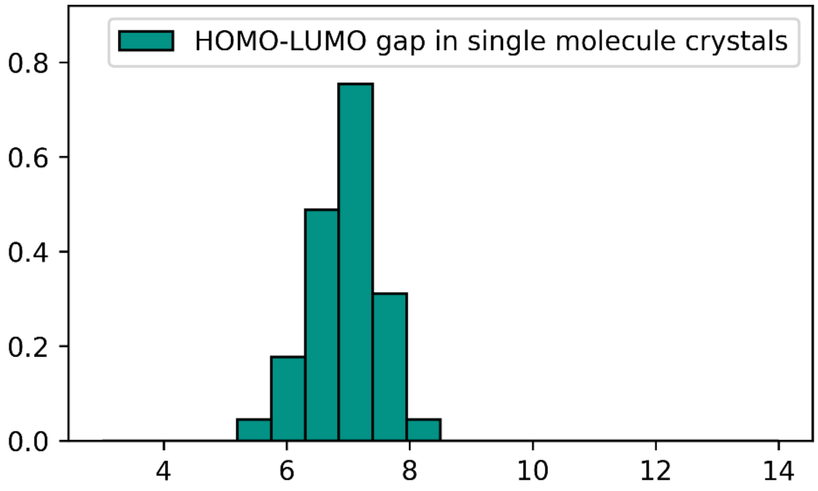


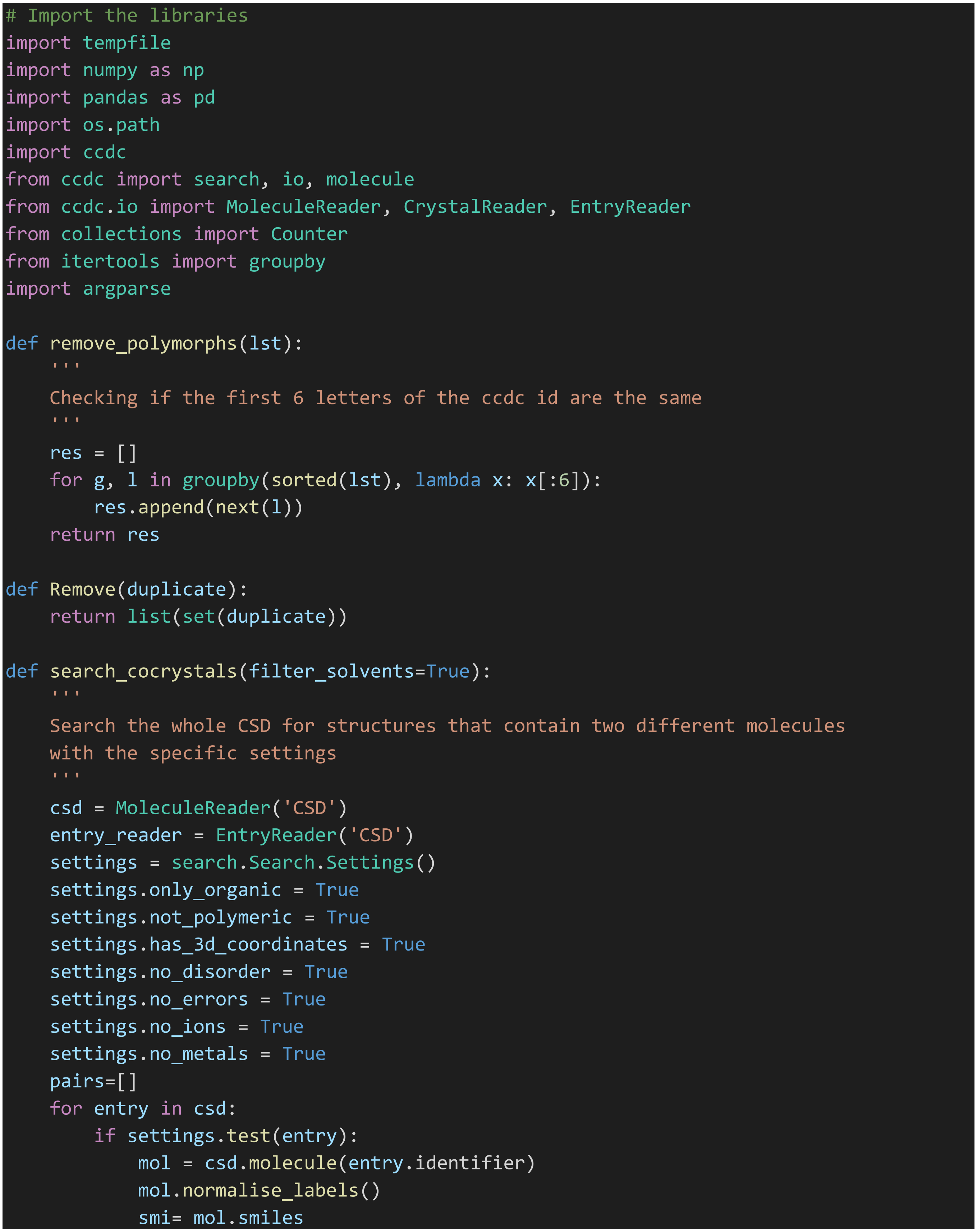
Figure D1.1. HOMO-LUMO gap in single molecule semiconductors. The orbital energies using PM6 were calculated for the list of the top 40 molecules reported in the SI of Nematiaram *et al.*.[251]

Table D1.1. Co-crystals categorized based on the types of bonding.

Type of bonding	Functional groups	Comments
Hydrogen bonding	Both molecules have OH or NH or SH	the donor atom D is any of N, O, or S, and the acceptor atom A is any of N, O, or S
Halogen bonded	One molecule should have a halogen and the other a heteroatom	D···X-A, where D is one of N, O, S, or Cl; X is either Br or I
Weakly bound (π-π stacking)	At least one molecule of the pair has one aromatic ring without heteroatoms	interactions that do not belong to any other category, mainly π-π interconnected

Table D1.2. Solvents and single atoms that were excluded from the molecular pairs during the co-crystal extraction.

CC(Cl)(Cl)Cl	**NC=O**	**CCNCC**
OCC(F)(F)F	OC=O	F
ClC=C(Cl)Cl	CCCCCCC	Br
ClC(Cl)=C	CCCCCC	BrBr
CCOC(CC)OCC	CC(C)COC(C)=O	[F]
COCOC	CCCCCC(C)C	[O]
ClCCCl	CC(C)O	[C]
ClC=CCl	CC(C)OC(C)=O	[Cl]
COCCOC	CC(C)OC(C)C	[Br]
C1COCCO1	Cc1cccc(C)c1	[Xe]
CCCCO	CO	[N]
CCCCCO	COc1ccccc1	[H]
CCCO	COC(C)=O	[I]
COC(C)(C)OC	CCCCC(C)=O	[He]
CCC(C)O	CC1CCCCC1	Cl
CCOCCO	CCC(C)=O	ClCl
COCCO	CC(C)CC(C)=O	I
CC(C)CO	CC(C)C(C)=O	II
CC1CCCO1	C1COCCN1	IIII
CC(C)CCO	CN(C)C(C)=O	IC(I)I
CC(O)=O	CN1CCCC1=O	ICl
CC(C)=O	CN([O])=O	C=O
CC#N	Cc1ccccc1C	C#C
c1ccccc1	Cc1ccc(C)cc1	ClCl
CCCCOC(C)=O	CCCCC	ClI
ClC(Cl)(Cl)Cl	CCCOC(C)=O	COC
Clc1ccccc1	c1ccncc1	OB(O)O
ClC(Cl)Cl	O=S1(=O)CCCC1	S=C=S
CC(C)c1ccccc1	COC(C)(C)C	O=S=O
C1CCCCC1	C1CCc2ccccc2C1	O=C=O
ClCCl	C1CCOC1	N#N
CCOCC	Cc1ccccc1	C#C
CC(C)NC(C)C	OC(=O)C(Cl)(Cl)Cl	CC#CC
CN(C)C=O	OC(=O)C(F)(F)F	I[As](I)I
CS(C)=O	O	NCCN
CCO	OO	IC#CI
CCOC(C)=O	C	CBr
OCCO	S	BrI
CCOC=O	N	

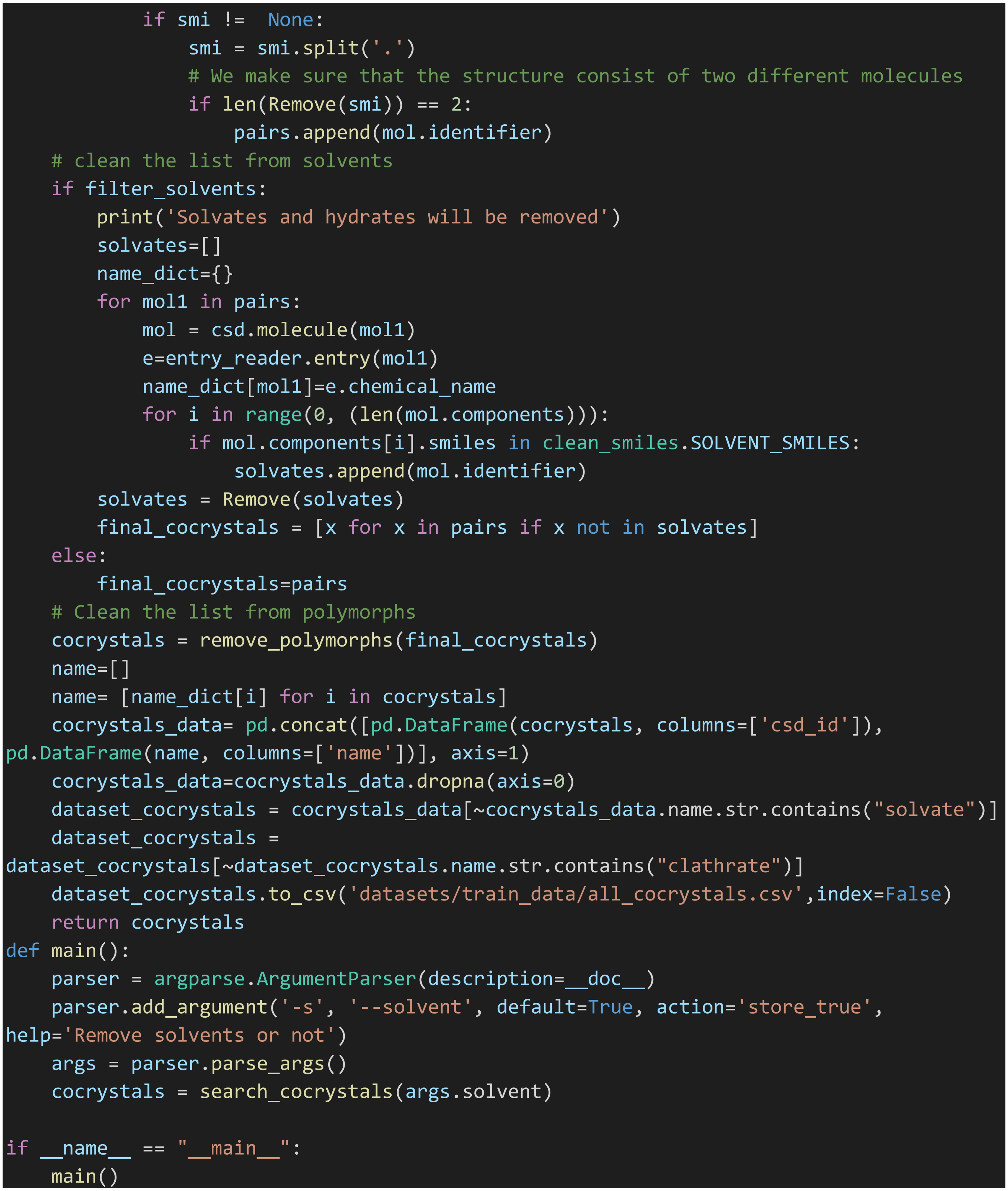


```
# Import the libraries
import tempfile
import numpy as np
import pandas as pd
import os.path
import ccdc
from ccdc import search, io, molecule
from ccdc.io import MoleculeReader, CrystalReader, EntryReader
from collections import Counter
from itertools import groupby
import argparse

def remove_polymorphs(lst):
    '''
    Checking if the first 6 letters of the ccdc id are the same
    '''
    res = []
    for g, l in groupby(sorted(lst), lambda x: x[:6]):
        res.append(next(l))
    return res

def Remove(duplicate):
    return list(set(duplicate))

def search_cocrystals(filter_solvents=True):
    '''
    Search the whole CSD for structures that contain two different molecules
    with the specific settings
    '''
    csd = MoleculeReader('CSD')
    entry_reader = EntryReader('CSD')
    settings = search.Search.Settings()
    settings.only_organic = True
    settings.not_polymeric = True
    settings.has_3d_coordinates = True
    settings.no_disorder = True
    settings.no_errors = True
    settings.no_ions = True
    settings.no_metals = True
    pairs=[]
    for entry in csd:
        if settings.test(entry):
            mol = csd.molecule(entry.identifier)
            mol.normalise_labels()
            smi= mol.smiles
```



```
            if smi !=  None:
                smi = smi.split('.')
                # We make sure that the structure consist of two different molecules
                if len(Remove(smi)) == 2:
                    pairs.append(mol.identifier)
    # clean the list from solvents
    if filter_solvents:
        print('Solvates and hydrates will be removed')
        solvates=[]
        name_dict={}
        for mol1 in pairs:
            mol = csd.molecule(mol1)
            e=entry_reader.entry(mol1)
            name_dict[mol1]=e.chemical_name
            for i in range(0, (len(mol.components))):
                if mol.components[i].smiles in clean_smiles.SOLVENT_SMILES:
                    solvates.append(mol.identifier)
        solvates = Remove(solvates)
        final_cocrystals = [x for x in pairs if x not in solvates]
    else:
        final_cocrystals=pairs
    # Clean the list from polymorphs
    cocrystals = remove_polymorphs(final_cocrystals)
    name=[]
    name= [name_dict[i] for i in cocrystals]
    cocrystals_data= pd.concat([pd.DataFrame(cocrystals, columns=['csd_id']),
pd.DataFrame(name, columns=['name'])], axis=1)
    cocrystals_data=cocrystals_data.dropna(axis=0)
    dataset_cocrystals = cocrystals_data[~cocrystals_data.name.str.contains("solvate")]
    dataset_cocrystals =
dataset_cocrystals[~dataset_cocrystals.name.str.contains("clathrate")]
    dataset_cocrystals.to_csv('datasets/train_data/all_cocrystals.csv',index=False)
    return cocrystals
def main():
    parser = argparse.ArgumentParser(description=__doc__)
    parser.add_argument('-s', '--solvent', default=True, action='store_true',
help='Remove solvents or not')
    args = parser.parse_args()
    cocrystals = search_cocrystals(args.solvent)

if __name__ == "__main__":
    main()
```

Figure D1.2. Python script for extracting co-crystals

D2. Results

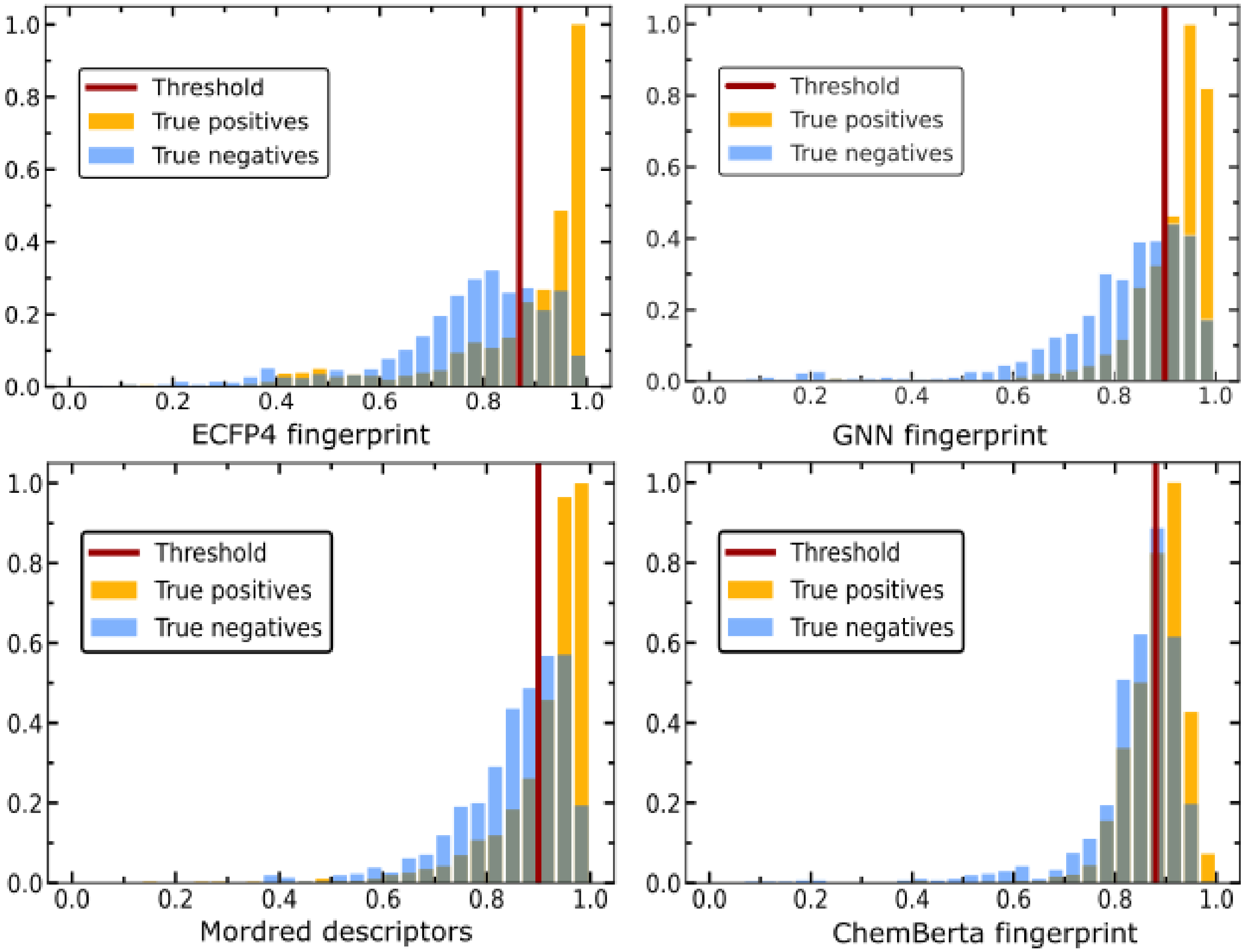
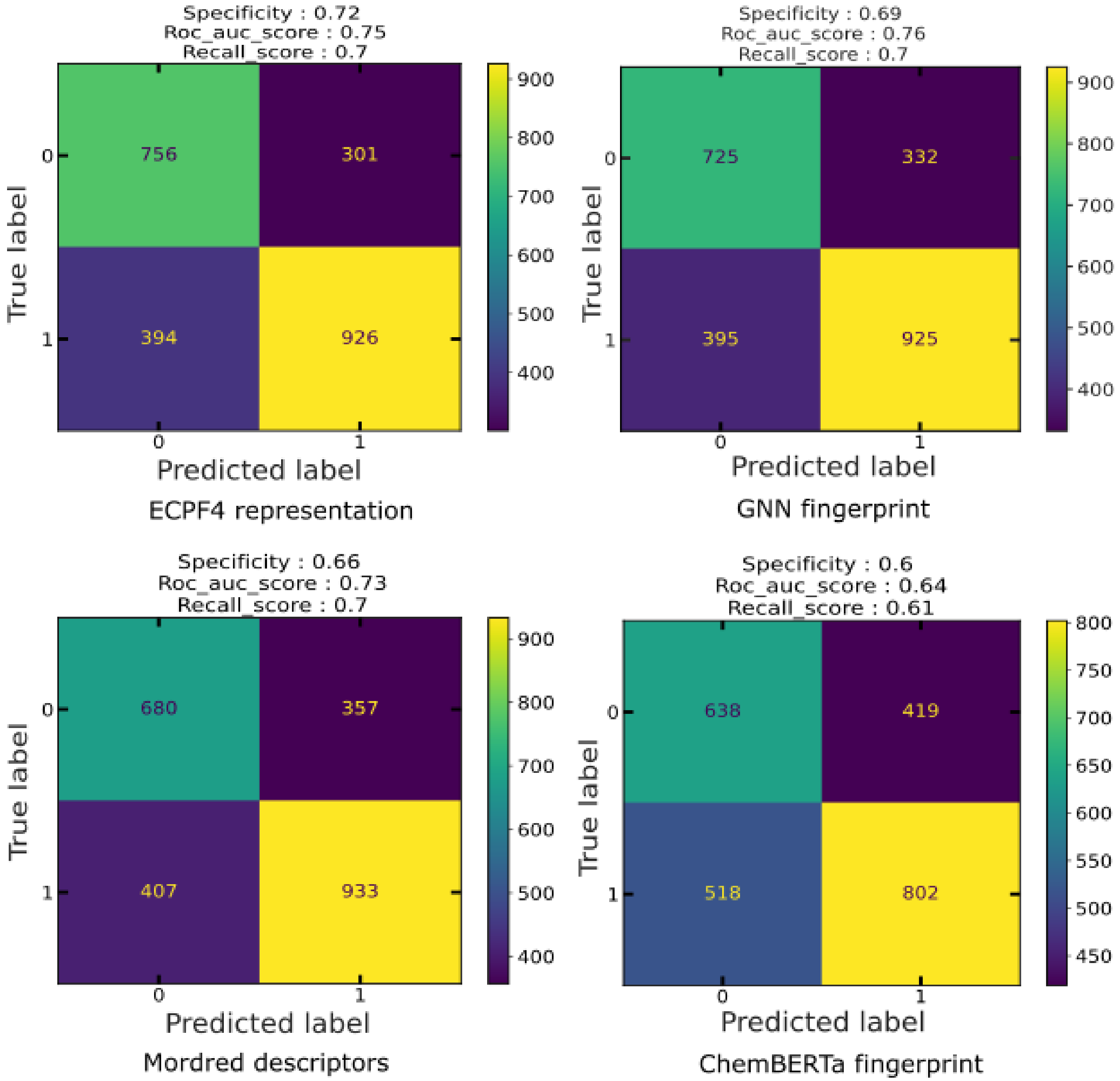


Figure D2.1. Scores distribution of the different models on the external validation sets. The real positives (orange bars) have higher scores than the true negatives (blue bars) for all four models. A better discrimination between the two classes is achieved for the ECFP4 and GNN models.



ECPF4 representation

GNN fingerprint

Mordred descriptors

ChemBERTa fingerprint

Figure D2.2. Confusion matrices of the four different models based on the representation techniques.

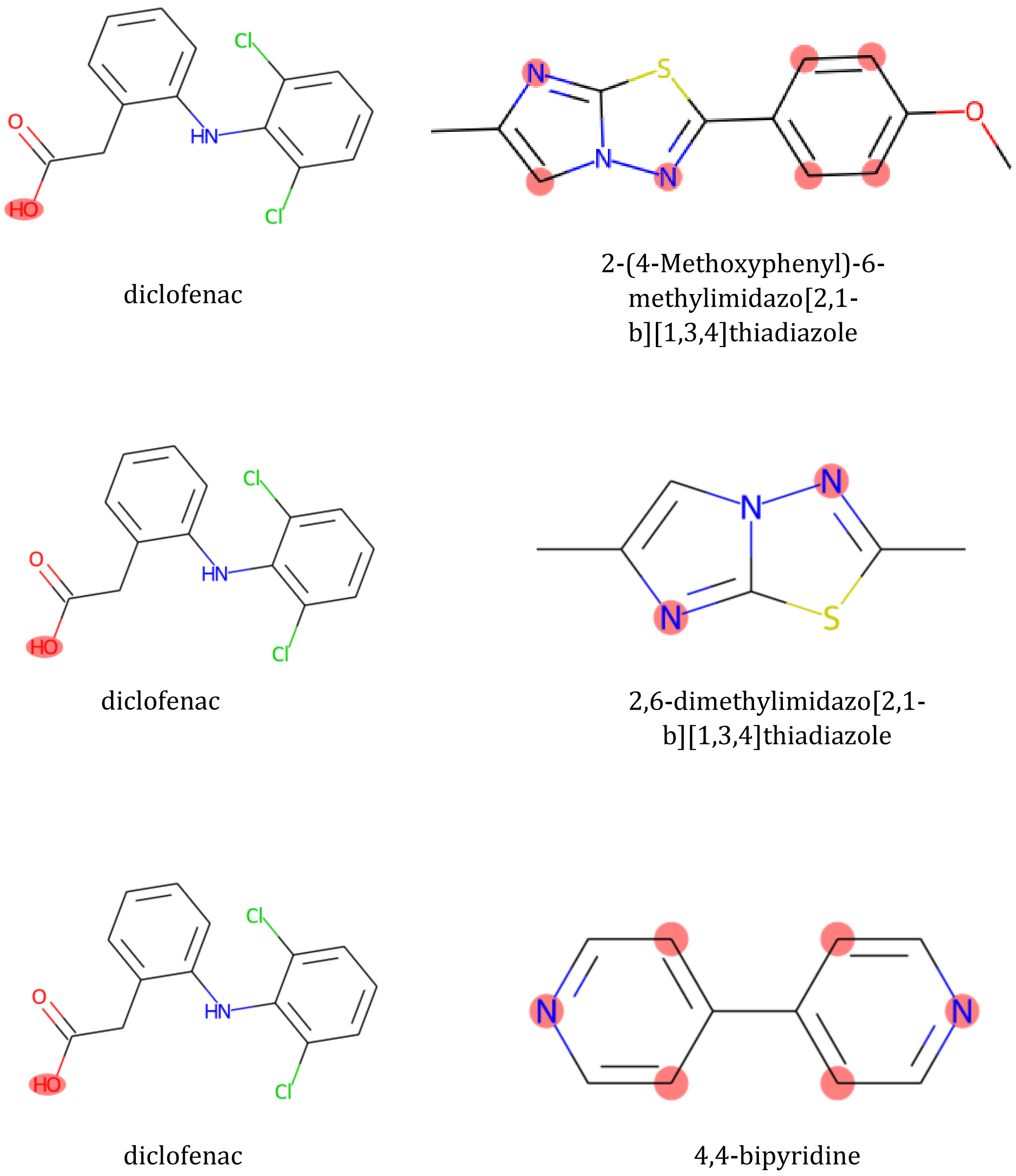

diclofenac
2-(4-Methoxyphenyl)-6-methylimidazo[2,1-b][1,3,4]thiadiazole
diclofenac
2,6-dimethylimidazo[2,1-b][1,3,4]thiadiazole
diclofenac
4,4-bipyridine

Figure D2.3. Three examples of diclofenac co-crystals when using Shapley local explanations to visualize the important bits of the molecular graph that drove to high scores of the Molecular Set Transformer. The bits with the highest importance are highlighted with red circles. It can be observed that the two most important groups are the -OH group of the API (diclofenac) and the N group of the co-former which can form H-bonding.