Technische Universität München
TUM School of Natural Sciences

# Molecular Machine Learning for Complex Electronic Properties

## Ke Chen

Vollständiger Abdruck der von der TUM School of Natural Sciences der Technischen Universität München zur Erlangung des akademischen Grades einer

**Doktorin der Naturwissenschaften (Dr. rer. nat.)**

genehmigten Dissertation.

**Vorsitz:** Prof. Dr. Hubert A. Gasteiger

**Prüfer der Dissertation:**

1. Prof. Dr. Karsten Reuter
2. Prof. Dr. Alessio Gagliardi

Die Dissertation wurde am 31.07.2023 bei der Technischen Universität München eingereicht und durch die TUM School of Natural Sciences am 12.09.2023 angenommen.

# Abstract

Machine learning (ML) has been widely used for chemical property prediction. In the context of molecular sciences, the prediction of simple electronic properties (e.g. the atomization energy) has reached extremely high accuracy for small and rigid molecules. However, the accurate prediction of more complex electronic properties, which are important targets for molecular/materials design, still poses a challenge. This is especially true for highly flexible molecules, due to their vast conformational variations and the intricacies of the underlying structure-property relationships. Examples of such properties are orbital energy levels, electronic couplings and molecular reorganization energies, all of which display markedly different properties than more common chemical ML targets like atomization energies. There is thus significant demand for methodological developments in chemical ML, in order to leverage its potential for molecular and materials design, for example in the context of organic semiconductors (OSCs).

In this work, we specifically focus on two molecular properties which are known to influence the performance of OSCs, namely the energy of the highest occupied molecular orbital (HOMO) as well as the internal reorganization energy ($\lambda$). First, we explore strategies to improve the performance of ML models for $\lambda$ prediction by using semi-empirical methods for conformer sampling and as a baseline. The obtained models are then evaluated in a diverse chemical space to discover novel low-$\lambda$ structures. This also enables the discovery of general chemical design rules by substructure searching among favourable candidates. Second, we consider the suitability of state-of-the-art ML models for predicting the HOMO energy. Being an intensive and sometimes localized property, we show that the common pooling strategies in atomistic neural networks are ill-suited for this target. To overcome this issue, we propose a series of physically motivated pooling functions. Among these, the novel orbital weighted average (OWA) approach is developed. OWA enables accurately predicting the orbital energies and distributions simultaneously. The underlying approach is also promising for other intensive properties such as excitation energies.

# Zusammenfassung

Machine Learning (ML) wird umfassend für die Vorhersage chemischer Eigenschaften verwendet. Im Kontext Molekularer Wissenschaften hat die Vorhersage einfacher elektronischer Eigenschaften (z.B. der Atomisierungsenergie) für kleine und starre Moleküle eine äußerst hohe Genauigkeit erreicht. Die genaue Vorhersage komplexerer elektronischer Eigenschaften, die wichtig für das Design von Molekülen oder Materialien sind, stellt jedoch immer noch eine Herausforderung dar. Dies gilt insbesondere für sehr flexible Moleküle aufgrund der großen Anzahl konformationeller Variationen und der Komplexität der zugrunde liegenden Struktur-Eigenschafts-Beziehungen. Beispiele solcher Eigenschaften sind Energieniveaus, elektronische Kopplungen und Molekülreorganisationsenergien, die sich alle deutlich von häufigeren mit ML vorhergesagten Eigenschaften in der Chemie, wie Atomisierungsenergien, unterscheiden. Es besteht daher ein erheblicher Bedarf an methodischen Entwicklungen in ML in der Chemie, um dessen Potenzial für das Design von Molekülen und Materialien zu nutzen, beispielsweise im Zusammenhang mit organischen Halbleitern.

Der Fokus dieser Arbeit liegt auf zwei molekularen Eigenschaften, die bekanntermaßen die Leistung von organischen Halbleitern beeinflussen, nämlich die Energie des höchsten besetzten Molekülorbitals (HOMO) sowie die interne Reorganisationsenergie ($\lambda$). Zunächst untersuchen wir Strategien zur Verbesserung der Leistung von ML-Modellen für die Vorhersage von $\lambda$, indem wir semi-empirische Methoden für das Sampling von Konformeren und als Baseline verwenden. Die erhaltenen Modelle werden dann in einem vielfältigen chemischen Raum bewertet, um neuartige Strukturen mit geringer $\lambda$ zu entdecken. Dadurch können auch allgemeine Designregeln entworfen werden, um die Suche nach chemischen Teilstrukturen unter den vielversprechenden Kandidaten zu erleichtern. Zudem wird die Eignung moderner ML-Modelle zur Vorhersage der HOMO-Energie überprüft. Da es sich bei der HOMO-Energie um eine intensive und manchmal lokalisierte Eigenschaft handelt, wird gezeigt, dass die gängigen Pooling-Strategien in atomistischen neuronalen Netzwerken zur Vorhersage von HOMO-Energien ungeeignet sind. Um dieses Problem zu überwinden, werden eine Reihe von physikalisch motivierten Pooling-Funktionen vorgeschlagen. Dazu wird der neuartige Ansatz des orbitalgewichteten Durchschnitts benutzt. Dies ermöglicht eine genaue Vorhersage der Orbitalenergien und -verteilungen gleichzeitig. Der zugrunde liegende Ansatz zeigt auch vielversprechende Ergebnisse für andere intensive Eigenschaften wie Anregungsenergien.

# Abbreviations

| | |
|---|---|
| **ML** | Machine Learning |
| **HTVS** | High-Throughput Virtual Screening |
| **DFT** | Density Functional Theory |
| **SOAP** | Smooth Overlap of Atomic Position |
| **MBTR** | Many-Body Tensor Representation |
| **CM** | Coulomb Matrix |
| **GPR** | Gaussian Process Regression |
| **OSC** | Organic Semiconductor |
| **GAP** | Gaussian Approximation Potential |
| **BP-NN** | Behler-Parrinello Neural Network |
| **HOMO** | Highest Occupied Molecular Orbital |
| **LUMO** | Lowest Unoccupied Molecular Orbital |
| **OFET** | Organic Field Effect Transistor |
| **OPV** | Organic Photovoltaic |
| **OLED** | Organic Light Emitting Diode |
| **IP** | Ionization Potential |
| **EA** | Electron Affinity |
| **MD** | Molecular Dynamics |
| **DG** | Distance Geometry |
| **ETKDG** | Experimental-Torsion-Knowledge Distance Geometry |
| **CREST** | Conformer-Rotamer Ensemble Sampling Tool |
| **RMSD** | Root-Mean-Square Deviation |
| **MTD** | Meta-Dynamics |
| **GC** | Genetic Crossing |
| **FF** | Force Field |
| **PES** | Potential Energy Surface |
| **DFTB** | Density Functional based Tight Binding |
| **SCC** | Self-Consistent-Charge |
| **ACE** | Atomic Cluster Expansion |
| **FCHL** | Faber-Christensen-Huang-Lilienfeld |
| **BoB** | Bag of Bonds |
| **NN** | Neural Network |
| **HIP-NN** | Hierarchically Interacting Particle Neural Network |
| **RBF** | Radial Basis Function |
| **ACSF** | Atom-Centered Symmetry Function |
| **SMILES** | Simplified Molecular-Input Line-Entry System |
| **AML** | Active Machine Learning |

# Contents

# 1. Introduction

With the success of DFT and fast development of computational power as well as application of advanced instruments for experiments, a huge number of theoretical as well as experimental data can be obtained[1]. In the wake of this, data-driven molecular and materials science has been a subject of intensive research. In particular, machine learning (ML) has been widely used for high-throughput virtual screening (HTVS)[2,3], accelerating theoretical simulations[4,5], predicting new materials[6–8] and other applications. This success is due to the fact that ML models are highly effective in learning the relationships between materials' structures and properties from data[9–19]. Compared with high-cost quantum chemical calculations as well as expensive experiments, ML approaches can dramatically accelerate the prediction of properties. Once trained, these models allow the inexpensive prediction for a large range of materials. By providing fast and accurate prediction of molecular and materials properties, ML thus has a huge potential to increase the speed and scope of molecular/materials discovery in vast chemical spaces.

In the context of molecular science, ML has been applied to learn a variety of molecular properties that are directly available from single-point electronic structure calculations (e.g. density functional theory, DFT), such as atomization energies[20], dipole moments[21], band gaps[22], excitation energies[23] and ionization energies[24]. The typical workflow of chemical ML models for such properties predictions is shown in Fig. 1. For established benchmark datasets of small molecules (such as the QM9[25] set), state-of-the-art ML models already achieve very high accuracy, comparable to the intrinsic error of the underlying DFT data. Despite this success, there remains a gap between the small, rigid molecules in QM9 and larger, much more flexible molecules (e.g. pharmaceutical compounds, organic polymers) which are of high technical and scientific interest. Indeed, molecular flexibility can significantly influence the quality of ML predictions when the target property sensitively depends on the molecular geometry. Due to the large conformational variety of flexible molecules, multiple low-energy conformers are usually accessible. This makes an accurate mapping between structures and properties challenging, as state-of-the-art representations of molecular structures in ML

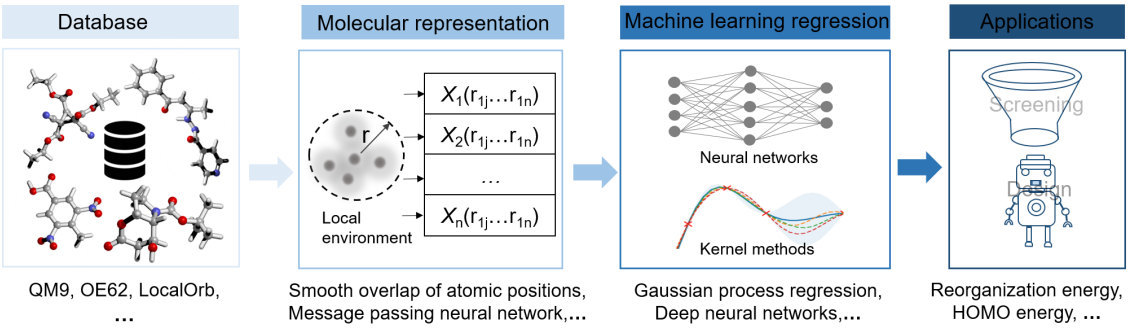



**Figure 1 Components of a chemical machine learning workflow.** Key aspects of chemical machine learning are training databases, structural representations, the machine learning models themselves and the target applications. All of these aspects need to be matched in order to obtain optimal performance.

(e.g. the Smooth Overlap of Atomic Position, SOAP,[26] or Many-Body Tensor Representation, MBTR[27]) are based on the full 3D molecular geometry. Accurate conformer sampling is therefore crucial for reliable ML predictions. However, this has mostly been overlooked in ML model development so far. Predicting complex properties for highly flexible molecules thus remains an open challenge in chemical ML.

In this work, ML models for two distinct complex molecular properties are developed. These are the internal reorganization energy $\lambda$ and the energy of the highest occupied molecular orbital (HOMO), which are both of great importance for organic semiconductors (OSCs). In Ref. [28], a novel dataset containing highly flexible organic molecules was generated. This allows for exploring the influence of molecular flexibility in ML models[29]. Specifically, the internal reorganization energy ($\lambda$), which measures the energetic cost for charge carriers to move between molecular sites in OSCs is considered. $\lambda$ is one of the most important factors determining the charge-carrier mobility in crystalline and amorphous organic semiconductors[30]. It has been widely used as a charge mobility descriptor to virtually screen promising candidates of OSCs by using DFT calculations[31–33]. Different from the properties that are directly available from ground-state DFT calculations, $\lambda$ depends on two potential energy surfaces and is therefore very sensitive to small variations of the molecular structure. Although some ML approaches have been proposed for predicting $\lambda$ in rigid molecules[34], the prediction of $\lambda$ for flexible molecules remains an unsolved challenge[35]. In order to address this issue, conformer sampling methods for flexible molecules are investigated and several strategies for how to improve the predictive performance of $\lambda$ using neutral geometries alone are explored in this thesis[29].

Chemical ML so far has been focusing on predicting energies and forces in high-dimensional systems. In this context, a chemical system is usually described as a set of atomic environments and the total energy of a molecule is computed as a sum over atomic contributions. This ensures size-extensivity, which is essential for transferable and extensible models, and makes the computational cost of the models scale linearly with the system size. However, some electronic properties (such as HOMO energies, excitation energies and ionization energies) do not scale linearly with system size. In such case, summing over atomic contributions however yields unphysical results, leading to large errors. This raises a question whether there are more suitable approaches for the localized size-intensive properties prediction. In this thesis, this question is addressed by proposing a series of pooling functions and benchmarking their performance on the novel 'LocalOrb' dataset of orbital energies featuring a wide range of localization degrees[36]. In this context, a new approach, incorporating physical information into the network architecture, is developed to provide physical understanding and to accurately predict localized HOMO energies as well as orbital information simultaneously.

In order to provide an in-depth understanding of this thesis, chapter 2 introduces the electronic properties of interest studied in this thesis. Chapter 3 describes state-of-the-art conformer sampling approaches and the conformer sampling workflows developed in this thesis. Chapters 4 provides an overview of the ML methods employed in this thesis. Finally, summaries of the published articles related to this thesis are provided in chapter 5.

# 2. Organic Semiconductors

Organic semiconductors (OSCs), with the advantages of high mechanical flexibility, chemical tunability and light weight, are very promising and attractive for electronic devices application [37–44]. OSCs have been widely used in devices such as organic field effect transistors (OFETs)[45], organic photovoltaics (OPVs)[46] or organic light emitting diodes (OLEDs)[47]. For electronic device applications, one of the most important figures-of-merit is the electric conductivity ($\sigma$), which is determined by the density of charge carriers ($\rho$) and the charge carrier mobility ($\mu$) with the electronic charge $\mathrm{e}$ ($\sigma = \mathrm{e}\rho\mu$)[30]. Unfortunately, the relatively low $\sigma$ of current OSCs hinders their adoption in many commercial applications. Improvements of $\sigma$ are therefore highly desirable. This chapter describes some important molecular electronic properties that affect the charge carrier mobility and density in OSCs and are thus of great interest as molecular design targets. Note that, depending on the majority charge carrier, OSCs can be distinguished into hole-transporting (p-type) and electron-transporting (n-type) semiconductors. Since the vast majority of OSCs are p-type[48], this thesis exclusively investigates p-type OSCs, though generalization to n-type OSCs is straightforward.

## 2.1. Reorganization Energy

As mentioned above, the charge carrier mobility $\mu$ is one of the most important factors for the efficient operation of OSC devices. Intense research interest has thus focused on understanding the mechanism of charge transport in OSCs. Two types of charge carrier transport mechanism are considered for OSCs: the coherent band mechanism and the incoherent hopping mechanism[30,49]. At low temperatures, charge carriers are delocalized and their mobility can be described according to the band mechanism. Here, $\mu$ is determined by the effective mass of the charge carrier and the relaxation time of the band[50]. With increasing temperature (i.e. around room temperature), the vibration of the crystal lattice and the associated scattering become stronger, leading to a narrowing of the bands[43]. Consequently, instead of being delocalized over the whole system, the charge carriers become localized on individual sites. This site can be a single molecule or part of a larger molecule/polymer within the solid film. As shown in Fig. 2a, the charge carriers then move from one site to another by a series of discrete jumps. This is generally described as a hopping mechanism. In the hopping model, the charge transfer rate between two adjacent molecules at the non-adiabatic limit is typically expressed via Marcus theory[51,52] as follows,

$$k_{\text{non-adiabatic},ab} = \frac{2\pi}{\hbar}\frac{1}{\sqrt{4\pi\lambda k_{\mathrm{B}}T}}|H_{ab}|^2 e^{-\beta\Delta G^{\ddagger}}, \tag{2.1}$$

where

$$\Delta G^{\ddagger} = \frac{(\lambda + \Delta G^0)^2}{4\lambda}, \tag{2.2}$$

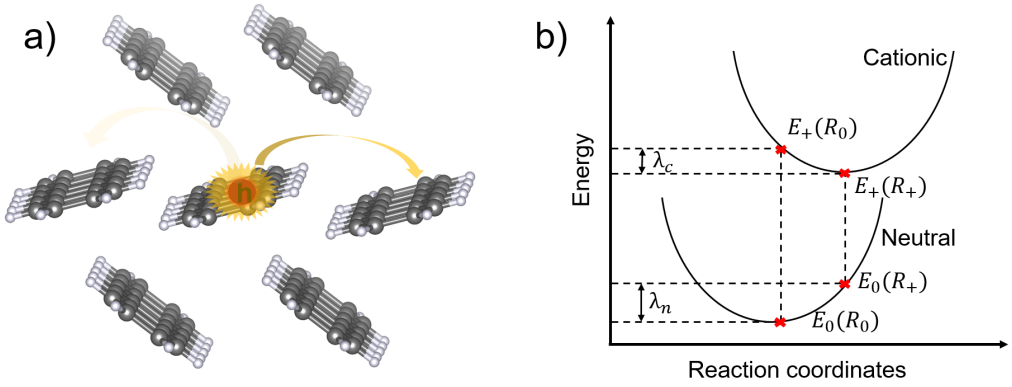



**Figure 2 Schematic depiction of the hopping transport**. a) A charge carrier localized at a molecule moves to the neighboring molecules by a thermally activated hopping process. b) Illustration of the adiabatic potential energy surfaces of neutral and cationic molecular states for hole transfer.

$\Delta G^0$ is the driving force, $T$ is the temperature, $k_\mathrm{B}$ is the Boltzmann constant, $H_{ab}$ is the electronic coupling, and $\lambda$ is the reorganization energy. $H_{ab}$ is related to the overlap of adjacent molecular orbitals and is rather sensitive to molecular packing arrangements. It varies significantly with the intermolecular distance as well as orientation[53]. $\lambda$ describes the energy change due to the geometry relaxation upon the charge transfer.

From equation 2.1, it can be observed that the charge transfer rate is mainly determined by $H_{ab}$ and $\lambda$, especially in ordered crystals where $\Delta G^0$ has a negligible contribution. To be precise, small $\lambda$ and large $H_{ab}$ are favorable for high charge mobility OSCs[31,33]. These two quantities are thereby suitable descriptors for charge mobility, for example in the context of HTVS, to find promising OSC candidates. In fact, since $\lambda$ enters exponentially in this equation, it is particularly crucial.

In full rigor, $\lambda$ accounts for internal and external contributions, $\lambda = \lambda_\mathrm{int} + \lambda_\mathrm{ext}$. $\lambda_\mathrm{int}$ originates from molecular deformations upon charge transfer. $\lambda_\mathrm{ext}$ is associated with the response of the surrounding medium during the charge transfer, which is more challenging to calculate or measure. Since $\lambda_\mathrm{ext}$ represents a relatively small contribution, it can often be neglected to a good approximation[49,54]. This leaves $\lambda_\mathrm{int}$ as a robust and straightforwardly computable descriptor of charge mobility for small molecule OSCs. As such, it has been used in several recent studies[28,29,31,33,34]. Our work only focuses on $\lambda_\mathrm{int}$. To simplify the notation, $\lambda_\mathrm{int}$ will be abbreviated as $\lambda$ in the following, unless noted otherwise.

The standard procedure to calculate $\lambda$ is called the four-point scheme[55]. For a p-type semiconductor, $\lambda$ is expressed as follows:

$$\lambda = E_0(\mathrm{R}_+) - E_0(\mathrm{R}_0) + E_+(\mathrm{R}_0) - E_+(\mathrm{R}_+). \tag{2.3}$$

As illustrated in Fig. 2b, $E_0$ and $E_+$ are the total energies of the neutral and cationic molecular states, evaluated at the equilibrium geometries $\mathrm{R}_0$ and $\mathrm{R}_+$ of the respective states. $E_0(\mathrm{R}_0)$ and $E_+(\mathrm{R}_+)$ represent the energies of the neutral and cation states in their lowest energy geometries, respectively. Similarly, $E_0(\mathrm{R}_+)$ and $E_+(\mathrm{R}_0)$ are the energies of the cationic and neutral states with the geometries of the neutral and cation state, respectively. In practice,

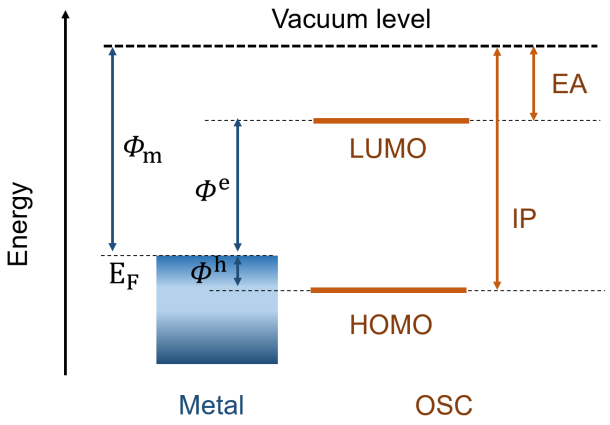



**Figure 3 Schematic energy level diagram of the Schottky-Mott limit vacuum level alignment at a metal-organic interface.** The electron and hole can be injected from a metal electrode into the organic semiconductor (OSC). $\Phi^{\mathrm{h}}$ and $\Phi^{\mathrm{e}}$ denote the injection barriers for holes and electrons. $\mathrm{E_F}$ is the Fermi level of the metal. IP is the ionization potential and EA is the electron affinity. $\Phi_m$ is the work function of the metal electrode.

two equilibrium geometries and four different energies therefore need to be obtained. This poses a challenge for $\lambda$ prediction in ML because $\lambda$ simultaneously depends on two potential energy surfaces, while typical target properties for chemical ML (i.e. atomization energies) are pure ground state properties.

## 2.2. HOMO Energies and Orbital Locations

The HOMO energy is another important electronic property for OSCs. From the perspective of the charge carrier mobility $\mu$, the HOMO energy can be used to calculate the static and dynamic energetic disorder of amorphous p-type OSCs[56]. From the perspective of the charge carrier density $\rho$, the HOMO energy influences the efficiency of the charge injection process from an electrode. In the ideal Schottky–Mott limit, the efficient injection of holes (electrons) from a metal electrode to an OSC layer depends on the difference between work function ($\Phi_m$) of the electrode and the ionization potential (IP) (electron affinity, EA) of the semiconductors in the vacuum level alignment at the interface[57–61], as shown in Fig. 3. In the OSC field, the HOMO energy is used as a common approximation for a material's IP. Indeed, gas-phase HOMO energies calculated at the hybrid B3LYP level of DFT correlate remarkably well with solid state IPs from experimental measurements[62]. The energetic mismatch between the electrode work function ($\Phi_m$) and the OSC HOMO energy is therefore an important descriptor for charge injection efficiency[28].

Intensive studies have recently been carried out for the estimation of HOMO energies[28,63,64], while the orbital locations have been rarely simultaneously discussed. The HOMO orbital location is nevertheless important to understand the charge transfer processes and it also influences chemical reactivity[65–67]. For example, the different localization of HOMO orbitals leads to the different site reactivity between Tetrahymena and Oxytricha telomeric quadruplex DNA[68].

Even disregarding orbital locations, ML models for HOMO energies are still relatively less accurate than for other electronic properties like atomization energies[20]. The current state-of-the-art atomistic ML methods are mainly based on local atomic interactions and assume that the target properties can be obtained by summation over atomic contributions. However, the HOMO energies are size-intensive and the HOMO orbitals may be spatially localized in the system so that summing or averaging over all atoms may be inappropriate. More efficient aggregation methods in ML models are highly demanded for localized HOMO energies prediction especially in systems with low symmetry. Thus, in this thesis, the HOMO energy and location are studied as a representative localized intensive property.

# 3. Conformer Sampling

To build efficient data-driven molecular ML models that use 3D geometries as input, accurate conformations of an organic molecule are crucial for reliable prediction of its properties, since the properties of interest depend on the precise atomic arrangement in a molecule[69–73]. For quantum properties prediction, the lowest-energy conformation is usually considered as the decisive structure[74]. However, finding the lowest energy conformer or all thermally-accessible conformers is challenging due to the high dimensionality of the search space for conformer generation (as shown in Fig. 4) and the computational cost of evaluating the relative energies of different configurations. When screening vast chemical spaces, efficient conformer sampling is thus a crucial but often overlooked part of ML workflows. In this chapter, different conformer generation approaches are briefly introduced and the effective conformer sampling workflows developed in this dissertation are illustrated.

## 3.1. Conformer Generation Algorithms

As shown in Fig. 5, classical conformer generation algorithms can be broadly classified as systematic and stochastic[72,75]. ML has also recently been applied to conformer generations [76–78]. These conformer generation algorithms have already made some achievements and are introduced below to broaden the view.

- **Systematic algorithms**. The conformational space is exhaustively sampled by complete torsional scans of all the rotatable bonds in a molecule using systematic algorithms[79]. This approach ensures that all the conformers are enumerated but is limited to fairly small or rigid molecules, since the dimensionality as well as the computational effort scale exponentially with the number of rotatable bonds (combinatorial explosion).
- **Stochastic algorithms**. The conformational space is sampled using methods such as molecular dynamics (MD)[80], Monte Carlo[81], distance geometry (DG)[75] or ge-

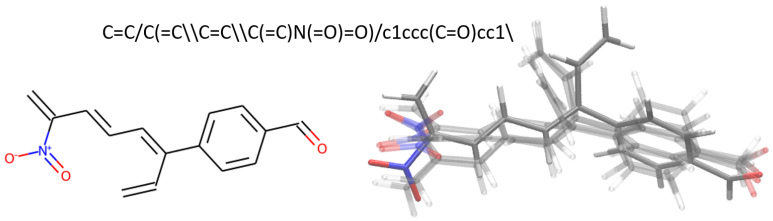



**Figure 4 The conformations of an example organic molecule.** The top part of this figure shows a simplified molecular-input line-entry system (SMILES) string, the left shows a stereochemical formula as a 2D graph, the right shows a superposition of 3D conformers.

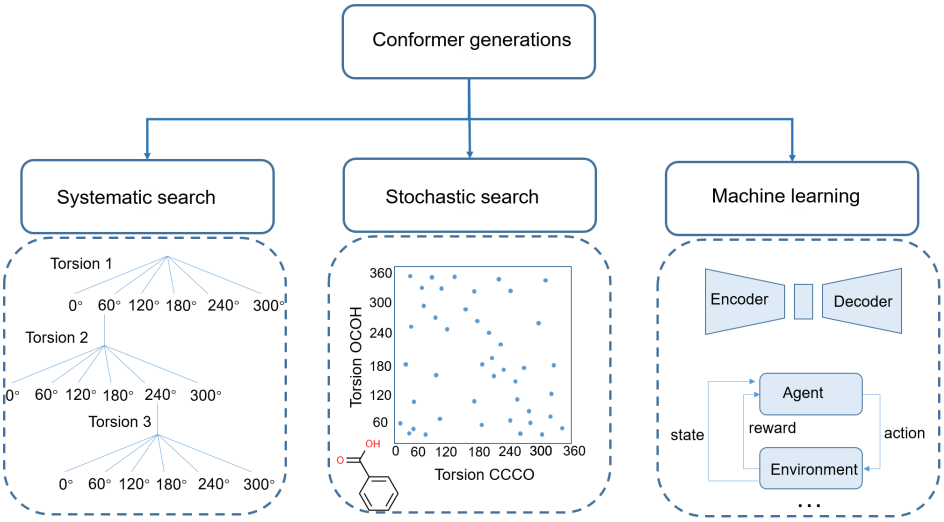



**Figure 5 Algorithms to search conformational space.** Systematic search, stochastic search and ML sampling are listed. The block for systematic search represents the enumeration of all rotatable bonds, the block for stochastic search represents random sampling of two torsion angles, the block for machine learning represents popular approaches such as generative models and reinforcement learning.

netic algorithms[82]. These approaches in principle can be applied to molecules with an arbitrary number of rotatable bonds. DG is a popular stochastic approach, which is widely applied for example in the RDKit free cheminformatic package as well as commercial conformer sampling packages[83,84]. In the DG method, conformational space is searched by randomly generating a large number of atomic coordinates based on upper and lower distance constraints. Furthermore, knowledge based methods have also been developed for conformer generation, by using predefined libraries of torsion angles and ring conformations[85–88]. These libraries are built mainly based on experimental structures from databases such as the Cambridge Structural Database[89] and Protein Data Bank[90]. A successful example of this is the Experimental-Torsion-Knowledge Distance Geometry (ETKDG), which utilizes the DG approach together with knowledge derived from experimental crystal structures to reliably generate conformers[83]. Nonetheless, stochastic methods can in some cases miss conformers. For example, MD simulations can become stuck in local minimal even with long simulations[70]. Recently, efforts were made to avoid this issue by using the metadynamics based sampling such as in the Conformer-Rotamer Ensemble Sampling Tool (CREST), which can sample low-energy conformational space more efficiently[91,92]. Both ETKDG and CREST are extensively used in this thesis.

- **Machine learning algorithms**. With increasing application of ML in chemistry and materials science, ML-based conformer generation methods have emerged as a new direction. ML-based conformer generation methods have been developed by using generative models[76,77] or reinforcement learning[78].

Following the conformer generation, geometry optimization and energetic ranking are re-

quired, in order to obtain accurate geometries and energies. In terms of energy estimation, classical force fields (FFs) and quantum chemistry methods are commonly used. Although classical FFs are widely used to identify low energy conformers, recent studies have shown that they are not reliable to assign an accurate energy ranking order, sometimes leading to missing the true low-energy conformers[93]. For example, the commonly used MMFF94[94] can reliably generate reasonable molecular geometries of organic compounds, but it fails at accurately ranking and identifying geometrically diverse conformers compared with quantum chemistry methods[93]. Meanwhile, quantum chemical methods are accurate but computationally inefficient.

To balance efficiency and accuracy, fast and cheap computational methods with lower accuracy are usually employed to sample the configuration space, while more costly methods with higher accuracy are subsequently employed to refine the conformer structures and energies[73]. In principle, high level quantum chemistry methods can undoubtedly yield very accurate energies of molecules. However, methods such as DFT (or even higher level methods like CCSD(T)) are too expensive and time-consuming to calculate energies for all conformers in the conformational space (especially for large molecules). Using semiempirical quantum chemical methods represents a good alternative in this context, as they are considered to bridge the gap between FFs and quantum chemistry methods. Methods from the recently developed GFN family are promising candidates since they are designed to yield good geometries, frequencies and noncovalent interaction energies[95–98]. Nonetheless, a prior benchmark test between low-level and high-level computational methods is always recommended to determine the efficient method in terms of computational cost and accuracy.

The efficient conformer generation methods, geometry optimization as well as energy reranking to obtain lowest energy conformers are explored in this thesis. ETKDG and CREST are mainly used and are therefore introduced below for a better understanding of the conformer sampling mechanism.

### 3.1.1. Distance Geometry Methods in RDKit

RDKit[99], a popular cheminformatic software in the chemical community, is widely utilized in molecular conformation generation. The conformer generation algorithms DG and ETKDG as implemented in RDKit emerge as preferred algorithms, being under most scenarios the best performing free available alternatives[83]. The dominance of accuracy and speed for RDKit make it a distinguished tool for conformer sampling in computational materials discovery projects.

In the standard DG approach, a distance bounds matrix, which represents the minimum and maximum interatomic distances for all pair of atoms in a molecule, is defined to effectively describe the whole conformational space of a molecule[100]. The distance bounds matrix is constructed mainly based on the connection table and a set of rules[75,83,100]. Based on this, random distance matrices can be generated by sampling distances between the corresponding upper and lower bounds. This distance matrix, which is consistent with the distance bounds constraints, is subsequently used to produce atom coordinates. After generating the

initial coordinates, a subsequent energy minimization is crucial in order to clean up unreasonable structures. Here, classical FF methods such as MMFF94 and UFF are typically employed to optimize the generated conformers and evaluate their energies. Combining the DG algorithm with a reasonable energy minimization in this manner has proven to yield diverse and representative conformers, which are often close to the experimentally or theoretically known structures[72,101–104].

Although the DG approach performs well, sometimes the generated conformers have unphysical structures such as distorted aromatic rings, unreasonable $\mathrm{sp}^2$ centers or torsion-angle values. To avoid this, the ETKDG approach has been developed as an alternative strategy. This is based on the DG approach but additionally utilizes torsional-angle preferences obtained from experimental crystal data and 'basic knowledge' such as 'aromatic rings are flat' or 'bonds connected to triple bonds are linear'[83]. ETKDG is currently the default conformer ensemble generator in RDKit. Importantly, benefiting from chemical knowledge, the ETKDG can generate reasonable structures even without energy minimization. However, some studies show that the subsequent energy minimization in ETKDG could still improve conformational sampling performance further, which is especially beneficial for obtaining the lowest-energy conformer[74,105]. Generally, it is recommended to test the performance of DG or ETKDG for a given application, due to the their different advantages.

One issue with the DG based conformer sampling methods is that the number of generated conformers must be specified beforehand. If a molecule is rigid, a large number of generated conformers will lead to many redundant structures. Meanwhile for very flexible molecules with a large number of rotatable bonds, it is possible to miss the low-energy conformer when sampling insufficient structures. Considering the processing time for conformer generation (particularly related to the cost of energy minimization), setting a suitable number of conformers is thus essential. Here, an extensive benchmark paper[101] suggested setting the number of generated conformers based on the number of rotatable bonds ($n_{\mathrm{rot}}$). Specifically, the number of conformers $(n)$ to generate is recommended as follows,

$$n = \begin{cases} 50, & \text{if } n_{\mathrm{rot}} \leq 7 \\ 200, & \text{if } 8 \leq n_{\mathrm{rot}} \leq 12 \\ 300, & \text{otherwise} \end{cases} , \tag{3.1}$$

To remove very similar structures, the conformers can be differentiated based on root-mean-square deviation (RMSD) values. For example, only conformers that are a certain RMSD threshold apart are kept. The conformer energy is nevertheless traditionally evaluated by FFs after the energy minimization when using the DG or ETKDG approaches. As mentioned above, the inaccuracy of classical FF energies may then lead to overlooking the real low-energy conformers. Therefore, higher level conformer energy calculation by using semi-empirical or first principles methods is recommended. For example, the semi-empirical GFN-xTB methods, which are introduced in the next subsection, are promising candidates as FF alternatives to efficiently obtain reliable conformer energies.

### 3.1.2. CREST Conformer Generation

The Conformer-Rotamer Ensemble Sampling Tool (CREST)[92], is a recently developed software which aims to strike a balance between speed and accuracy for molecular conformation sampling. It combines semiempirical tight-binding methods of the GFN-xTB family or the GFN-FF with RMSD-based metadynamics (MTD) simulation for conformer search. For the automatic exploration of the conformational space, a complex workflow called iMTD-GC[91] was developed, consisting of MTD, regular MD sampling and genetic structure crossing (GC). A history-dependent biasing potential that utilizes the atomic Cartesian RMSD as a collective variable is applied to accelerate the exploration of the potential energy surface (PES), where the biasing contribution is expressed by a Gaussian-type potential,

$$V_{\mathrm{bias}} = \sum_{i}^{n} \mathrm{k}_i \exp(-\alpha_i \Delta_i^2), \tag{3.2}$$

where $\Delta_i$ is the collective variable, $n$ is the number of reference structures, $\mathrm{k}_i$ and $\alpha_i$ are empirical parameters. During the evolution of the simulation, this bias adds up to prevent the system from returning to previously explored parts of the conformational space. By continuously expanding the reference structure list, the PES can be efficiently explored in this manner, making the MTD approach much more efficient for conformer sampling than regular MD simulation.

The CREST program deeply relies on the semiempirical GFNn-xTB methods. The semiempirical GFNn-xTB and force-field GFN-FF approaches have been successfully developed and are parameterized for almost the whole periodic table up to Radon ($\mathrm{Z} \leq 86$)[95]. These methods are all implemented in the efficient and freely available xtb program. The GFNn-xTB methods (n=0,1,2) are represented by GFN0-xTB, GFN1-xTB as well as GFN2-xTB, respectively. The first GFN family member, GFN1-xTB[96] is a density functional based tight binding (DFTB) variant, which utilizes almost the same approximations for the Hamiltonian and electrostatic energies as DFTB but avoids element-pairwise parameters in order to achieve full coverage of the periodic table. Instead, only global and element-specific parameters are employed and optimized by fitting reference data at the hybrid density functional theory level. The successor GFN2-xTB[97] is further developed including electrostatic interactions and exchange–correlation terms beyond the monopole approximation.

Although GFN1-xTB and GFN2-xTB have been widely used for a broad range of applications, there are sometimes problematic convergence issues during the self-consistent-charge (SCC) procedure. Furthermore, the Hamiltonian matrix must be diagonalized at each step of the SCC cycle, leading to significant computational effort, especially for very large systems. In order to avoid SCC iterations and to accelerate the calculations, a non-self-consistent method entitled GFN0-xTB[98] was developed by truncating the Taylor expanded DFT energy $E[\rho]$ in terms of electron-density fluctuation $\delta\rho$ after the first-order term. The electrostatic terms are treated by an electronegativity equilibration atomic charge model[106] and only a single Hamiltonian matrix is diagonalized. This non-self-consistent approach accelerates the calculations but in some cases leads to somewhat worse performance than GFN1- and GFN2-xTB as a

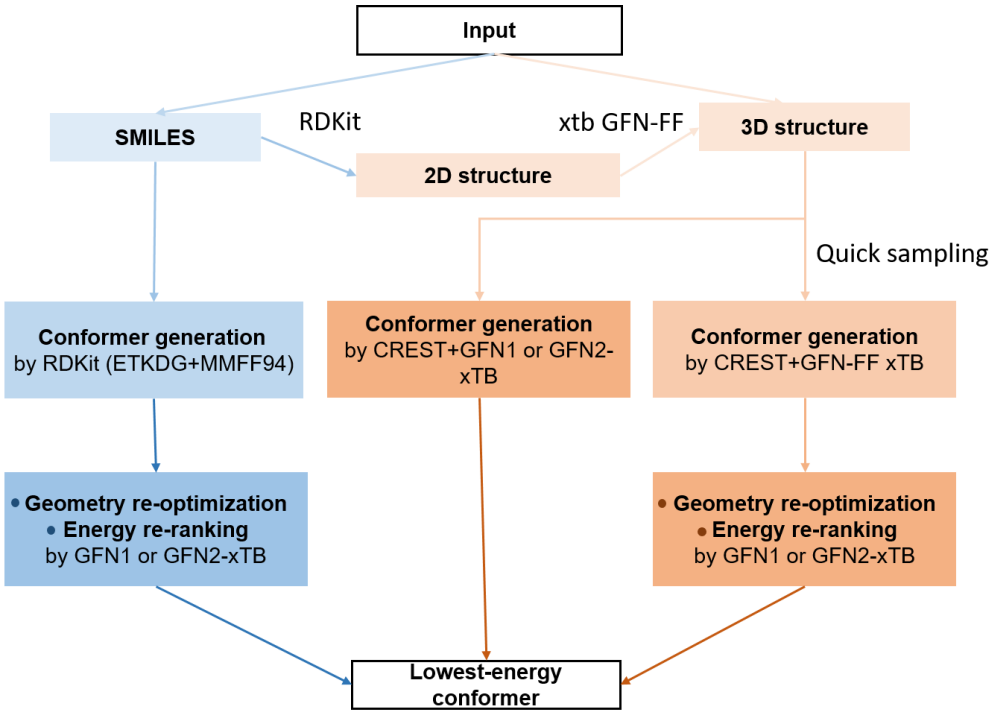



**Figure 6 The conformer generation workflows used in this thesis**. The RDKit input is usually a 2D graph converted from a SMILES string, while CREST requires an initial 3D geometry as input. The CREST input can also be generated via GFN-FF from a 2D graph.

trade-off.

Many studies show that the GFNn-xTB methods outperform other semiempirical methods and even approach high-level QC methods in many cases[107–113]. The GFN-xTB based CREST conformer generation and energy ranking is thus very reliable[114]. Furthermore, a force-field method named GFN-FF[115] was also developed in Grimme's group to effectively handle very large systems beyond thousands of atoms. GFN-FF has the most favorable cost-accuracy ratio for conformational search[109,111,116,117]. However, the main limitation of GFN-FF is that the input structures must be reasonable and all GFN-FF generated conformers are recommended to be re-optimized at a higher level such as GFN1- or GFN2-xTB or even DFT to reliably obtain lowest-energy conformers[118]. Overall, choosing the appropriate method from the GFN family within CREST significantly affects the speed and reliability of the conformer search.

### 3.1.3. Hybrid Conformer Sampling Workflows

For the conformer generation in this dissertation, ETKDG and CREST were used in different projects depending on an acceptable time/cost ratio. This workflow evolved according to the requirements of each project, as shown in Fig. 6. For the first project[28] (see chapter 5.1), the ETKDG in RDKit was used to generate the conformers, followed by an initial relaxation with the MMFF94 force field. In order to reliably obtain lowest-energy conformers, GFN1-xTB, which approaches the DFT level (B3LYP) accuracy based on our prior benchmark test in this

work, was utilized to re-optimize the geometries and select the lowest-energy conformers. Combining RDKit and GFN-xTB reaches a high cost-accuracy ratio for conformer sampling and is therefore recommended for RDKit users. For the second project[29] (see chapter 5.2), a large amount of molecules (130k) was sampled. To this end, the CREST approach with GFN-FF energies was thus applied for very fast conformer ensemble sampling. The GFN-FF generated structures were subsequently re-optimized at GFN1-xTB level in order to yield good geometries and energies. As the 'LocalOrb' dataset of the third project was somewhat smaller in size (21k)[36], the CREST with GFN2-xTB was directly employed to provide a more consistent generation of conformers compared to GFN-FF (see chapter 5.3). The lowest-energy conformers were obtained directly from CREST since GFN2-xTB yields acceptable accuracy for energy calculations. Choosing GFN1 or GFN2-xTB for conformer generation or geometry optimization however depends on the performance compared with higher-level calculations for each specific application.

# 4. Molecular Machine Learning

As a subset of the field of 'artificial intelligence', ML has been widely adopted in physics, chemistry and material science[16,119,120]. Example applications are the assistance of theoretical simulations, the prediction of physicochemical properties, the de novo design of new materials, as well as the optimization of experimental synthesis[11,14,15,121,122]. Regarding the main focus of this thesis, ML has also been applied to materials design and screening in recent years, but there remain significant challenges[18,123–125]. For example, ML models with high accuracy have been reported for small, rigid organic molecules, but these cannot in general be transferred to larger, more flexible molecules. The accuracy of chemical ML models that predict structure-property relationships crucially depends on the choice of the structural representations used as inputs to the models. Furthermore, the architecture of the model itself must also match the requirements of the target properties. The structural representations as well as the ML algorithms used in the thesis are therefore introduced in this chapter.

## 4.1. Structural Representations

Choosing an appropriate representation of a molecular structure is perhaps the most crucial decision when building a chemical ML model[10,16,126–131]. To this end, the molecular structure must to be converted from its chemical compositions and atomic coordinates into invariant representations under translation, rotation and permutation. Furthermore, the representation should ideally be unique, so that no two different structures are mapped to the same representation. In recent years, numerous molecular representations have been developed, which can broadly be classified into two categories. On one hand, there are local structural representations (or atomic representations), such as the atomic cluster expansion (ACE)[132], Faber-Christensen-Huang-Lilienfeld (FCHL)[133], or the Smooth Overlap of Atomic Position (SOAP)[26]. On the other hand, there are global representations, such as the Bag-of-Bonds (BoB)[134], Coulomb Matrix (CM)[20] and Many Body Tensor Representation (MBTR)[27]. Typically, local representations are centered on each atom and encode information about its neighbors[122], describing atoms in their environments. Meanwhile, global representations describe the whole system in terms of internal coordinates. Put differently, in local representations, a chemical system is described as a set of atomic environments. Therefore, a global representation can be generated by combining the set of atomic representations of all environments, e.g. by taking a sum or average the atomic representations.

As the name indicates, local representations are based on the assumption of locality, meaning that the properties of a compound (e.g. the total or atomization energy) can be approximated as a function of atomic contributions[16,129]. This assumption is widely employed in ML interatomic potentials (e.g. Gaussian Approximation Potentials (GAP)[135] or Behler-Parrinello neural networks (BP-NN)[136]), where the total energy is typically obtained as a sum of local atomic energy contributions. Notably, some properties are less suitable to decomposing into

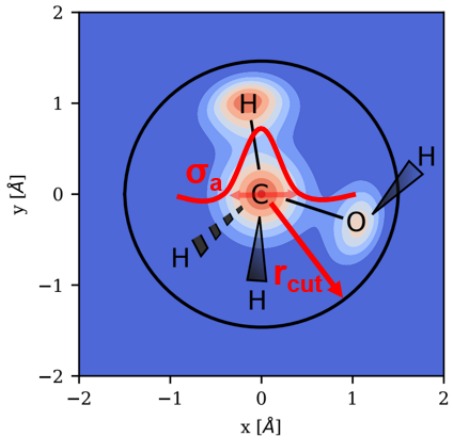



**Figure 7 Schematic illustration of the smooth overlap of atomic positions (SOAP).** The main idea of the SOAP representation is that the neighborhood density of a given atom (within a cutoff $\mathrm{r_{cut}}$) is expressed through a superposition of atom-centered Gaussian functions of width $\sigma_\mathrm{a}$. Figure adapted from reference [137].

local atomic contributions and might be better described by global models. Recently, both local and global representations have been reviewed extensively[16,126,129]. This subsection will therefore focus on the representations used in this dissertation.

### 4.1.1. The Smooth Overlap of Atomic Position

The Smooth Overlap of Atomic Position[26] is an atomic density based representation. An atomic environment $\mathcal{X}$ around a central atom is represented by the local density $\rho$, which is constructed by the summation of Gaussian functions placed on each atom $i$ inside the cutoff sphere, as shown in Fig. 7:

$$\rho_{\mathcal{X}}(\mathbf{r}) = \sum_{\mathrm{i}\in\mathcal{X}} \exp\left(-\frac{|\mathbf{r}_i - \mathbf{r}|^2}{2\sigma_\mathrm{a}^2}\right) \cdot f_\mathrm{cut}(|\mathbf{r}|)) \quad , \tag{4.1}$$

where the smoothness of the density is controlled by the width of the Gaussian $\sigma_\mathrm{a}$ and the number of neighbors that are considered in the summation is determined by the cutoff radius $r_\mathrm{cut}$ in the cutoff function $f_\mathrm{cut}$.
This density $\rho$ exhibits translational and permutational invariance by construction, but no invariance to rotations. To obtain a rotationally invariant representation, the density is thus first expanded in a basis of spherical harmonics and orthogonal radial basis functions. Subsequently, the rotationally invariant power spectrum $p_{(nn'l)}$ is computed from the expansion coefficients $c_{nlm}$:

$$p_{(nn'l)}(\mathcal{X}) = \pi\sqrt{\left(\frac{8}{2l+1}\right)}\sum_m (c_{nlm})^\dagger c_{n'lm} \quad , \tag{4.2}$$

Here, the indices $n$ and $n^{'}$ label radial basis functions while $l$ and $m$ label the spherical harmonics. The radial and angular resolution of the atomic density is thus limited by the maximum values for $l$ and $n$. Further details about the mathematical transformation from the

atomic density to the power spectrum can be obtained from the literature[26,138].
As indicated above, a series of hyperparameters must be defined when using SOAP. In this dissertation, these were either chosen by grid search (optimizing validation errors of the corresponding models) or automatically selected according to universal heuristics. These universal heuristics were introduced by Cheng et al.[139] to choose SOAP hyperparameters based on the characteristic bond lengths of arbitrary chemical species in a system.
In general, the local SOAP representation can easily be transformed into a global representation. For example, for the work presented in chapter 5.2, a SOAP-based global representation was built using the Auto-Bag method[140]. The SOAP representation is also used for data visualization based on kernel principal component analysis in this thesis.

### 4.1.2. Electronic Properties as Representations

Beyond structural representations, which are based on the Cartesian coordinates of atomic positions, molecules can also be represented by electronic properties computed from computationally inexpensive electronic structure calculations (e.g. the semiemprical xTB methods) [129,130]. As an example, molecular orbital based representations (e.g. electron densities, density matrices or orbital coefficients) have been used as input features for electronic properties prediction, yielding high learning efficiency and transferability[141–143].
For simpler models, electronic descriptors which are correlated to the target of interests can also be used. For example, Terrones et al. selected HOMO, LUMO, IP, EA, and partial charges from xTB calculations as feature sets to efficiently predict excited state properties[144]. In chapter 5.2, we used frontier orbital energies, gaps, Fermi levels, total energies and vertical energy differences of the neutral and cationic system of molecules computed at the GFN1-xTB level to construct electronic properties representations for $\lambda$ prediction[29]. Nevertheless, the ML performance can be limited by the expressiveness of the respective properties.

### 4.1.3. End-to-End Neural Network Representations

Predefined representations like SOAP can be used in combination with kernel-based ML methods or shallow neural networks. In contrast, deep neural networks (NNs) are flexible enough that they can learn efficient representations of molecules and materials from data[16,129,145]. Specifically, deep NNs can construct atomic representations in an end-to-end fashion, only using atom types and their positions as input.
SchNet[146] is a prototypical end-to-end deep neural network architecture, based on continuous filter convolutions and designed to learn representations of the local environment that satisfy all the required invariances. In this thesis, SchNet was employed to predict the HOMO energies of organic molecules (see section 5.3). A detailed discussion of the SchNet architecture is provided in section 4.2.2.

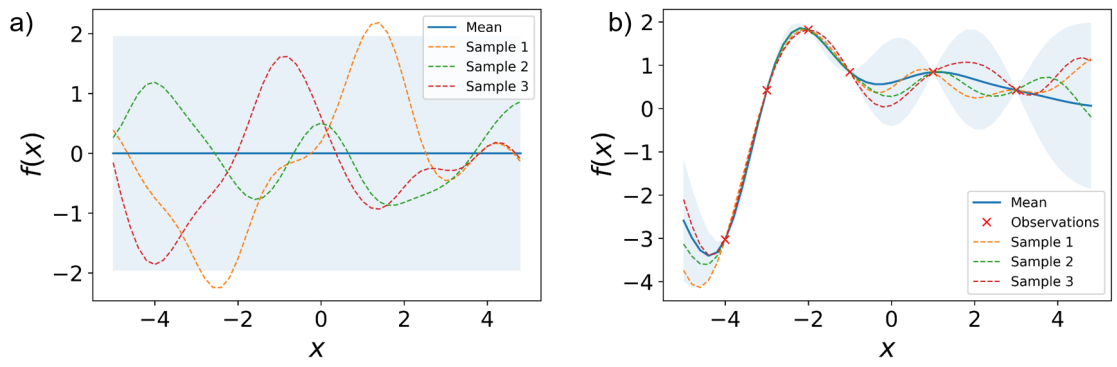



**Figure 8 Functions drawn from a Gaussian Process prior and posterior distribution**. a) Three samples drawn from a Gaussian Process prior. b) Three samples drawn from a Gaussian process posterior conditioned on observations. The blue regions represent the uncertainty range, the true underlying function is $f(x) = x\sin(x)$.

## 4.2. Regression Techniques

The accuracy of a ML model based on structure-property relationships is determined by the structural representations and the regression model employed to associate them with the target properties. After defining the representations, it is therefore necessary to define the regression method. Numerous types of regression models are available,[129] which can be used to approximate nonlinear and high-dimensional functions. The most common examples are kernel-based methods and deep neural networks. In this section, Gaussian Process Regression (GPR) and SchNet, both of which were used in this thesis, are described in detail.

### 4.2.1. Gaussian Process Regression

Gaussian Process Regression is a kernel-based, probabilistic, and non-parametric machine learning method. GPR models can efficiently approximate the underlying structure-property relationships for a training set $D = \{\mathbf{X}, \mathbf{y}\}$. Here $\mathbf{X}$ is a set of molecular representation vectors $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ and $\mathbf{y}$ is the column vector of the target values $\mathbf{y} = [y_1, ..., y_n]$. In GPR, we assume that the observed target value $y_i$ can be expressed as a function of the input vector $\mathbf{x}_i$,

$$y_i = f(\mathbf{x}_i) + \epsilon_i, \tag{4.3}$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma_{\mathrm{n}}^2)$ is a noise term. The function $f(\mathbf{x})$ is typically assumed to be distributed as a Gaussian process, where a Gaussian process is a statistical distribution over functions and is defined by a mean and a covariance function[147]. A Gaussian process prior over function $f$ is typically assumed with zero mean and covariance function without training information. From the prior distribution, a posterior distribution is obtained by conditioning the joint Gaussian prior distribution on the observations $D$, as shown in Fig. 8. The posterior distribution can then be used to predict the distribution of the target property for previously

unseen molecular representations $\mathbf{X}^{'}$, where

$$f(\mathbf{X}^{'}) \sim \mathcal{N}(\mu(\mathbf{X}^{'}), \sigma^2(\mathbf{X}^{'})). \tag{4.4}$$

Point estimates of the predicted values thus can be obtained as the predictive mean,

$$\mu(\mathbf{X}^{'}) = K(\mathbf{X}^{'}, \mathbf{X})[K(\mathbf{X}, \mathbf{X}) + \sigma_n^2 \mathrm{I}]^{-1}\mathbf{y}. \tag{4.5}$$

Additionally, the variance of the predicted value can also be obtained, which is usually used as statistical error estimate to evaluate the reliability of the predicted value.

$$\sigma^2(\mathbf{X}^{'}) = K(\mathbf{X}^{'}, \mathbf{X}^{'}) - K(\mathbf{X}^{'}, \mathbf{X})[K(\mathbf{X}, \mathbf{X}) + \sigma_n^2 \mathrm{I}]^{-1}K(\mathbf{X}, \mathbf{X}^{'}). \tag{4.6}$$

Here, $\mathrm{I}$ is the identity matrix. $\sigma_n$ is an assumed Gaussian noise from observations. $K$ is the covariance matrix with $K_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$, where the kernel $k(\mathbf{x}, \mathbf{x}^{'})$ is used to measure the similarity between different representations. Several kernel functions are available to deal with different types of data[148], such as the Tanimoto-kernel, which is typically used for molecular fingerprint input in cheminformatics[149].
The radial basis function (RBF) kernel is one of the most popular choices, defined as:

$$k(\mathbf{x}_\mathrm{i}, \mathbf{x}_\mathrm{j}) = \sigma_\mathrm{f}^2 \exp\left(-\frac{\mathrm{d}(\mathbf{x}_\mathrm{i}, \mathbf{x}_\mathrm{j})^2}{2l^2}\right), \tag{4.7}$$

where $\mathrm{d}(.,.)$ is the Euclidean distance. The length-scale $l$ and the vertical scale $\sigma_\mathrm{f}$ as well as the noise $\sigma_\mathrm{n}$ are usually referred to as hyperparameters and can be optimized or determined during model training. Additionally, one can also build more complex kernels by combining simpler kernels through addition or multiplication operations to make the model more powerful. Indeed, a GPR model that combines a structure-based kernel and an electronic property-based kernel is used in one of the works of this thesis (see chapter 5.2). Like the representations, kernels also have hyperparameters. One of the advantages of GPR is that the hyperparameters can be obtained by optimizing the log-marginal likelihood. Therefore, GPR hyperparameters are usually optimized by maximizing the log marginal likelihood using the L-BFGS algorithm.
GPR is a powerful method for fitting interatomic potentials (e.g. in GAP) and to predict a wide range of atomic-scale properties (e.g. NMR chemical shielding or electron densities)[150]. Furthermore, the ability to estimate the uncertainty of the predicted values further increases the usefulness of GPR models. In particular, the predictive uncertainty (or variance) provides the foundation of many acquisition functions, which can be used to guide the exploration–exploitation trade-off in active learning and global optimization strategies[28,151].

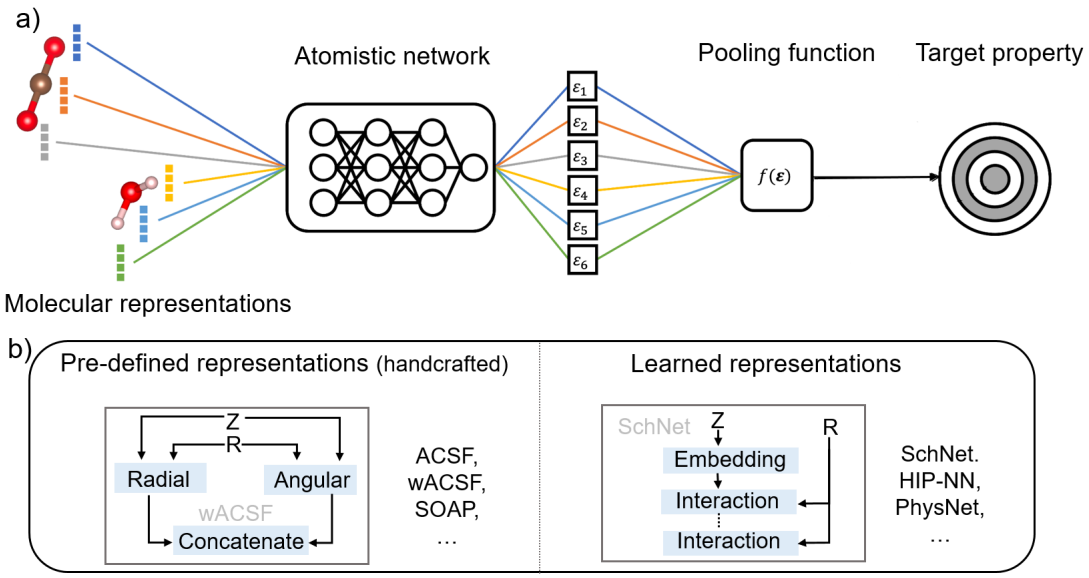



**Figure 9 Illustration of atomistic neural network architectures.** a) Schematic depiction of an atomistic neural network. b) Two categories of representations. The left part represents pre-defined representations, the right part represents end-to-end representations which can be learned from the neural network during the training process. Figures a) and b) adapted from references [36, 152], respectively.

### 4.2.2. Atomistic Neural Networks

A general scheme of an atomistic NN is illustrated in Fig. 9a. In brief, each atom $i$ in a given system with N atoms corresponds to a chemical environment representation[36]. This representation is passed through the NN to yield a scalar output $\epsilon_i$. The target property is finally obtained by combining the output of all atom contributions through a pooling function. Atomistic NNs have been widely used to build ML interatomic potentials (e.g. the BP-NN[136], SchNet and hierarchically interacting particle neural network (HIP-NN)[153] approaches) and to predict the physico-chemical properties of atomistic systems (e.g. atomization energies or dipole moments)[145,154]. This family of chemical ML methods was pioneered by Behler and Parrinello and is based on the already mentioned idea that the properties of an atomistic system can be decomposed into local contributions[136]. The target property is thus reconstructed by aggregating atomic contributions via a physically motivated aggregation layer (a pooling function) at the last step of the NNs. For example, in the most common case, the system's total energy is assumed to be obtained by summing over atomic energies. As a result, atomistic NN models ensure size-extensivity and enable linear scaling of computational resources with the system size.

As for GPR models, suitable structural representations play a crucial role in determining the atomistic NNs accuracy. Depending on the strategy used to obtain the representations, two categories of atomistic NN models can be distinguished, as depicted in Fig. 9b. The first category employs representations which have been defined before training. In BP-NN, this is achieved via atom-centered symmetry functions (ACSFs)[136] or modified symmetry functions as in the ANI potential[155]. Similarly, SOAP can be used as a local atomic representation, as shown in chapter 5.3. The second category is to learn an efficient representation end-to-end

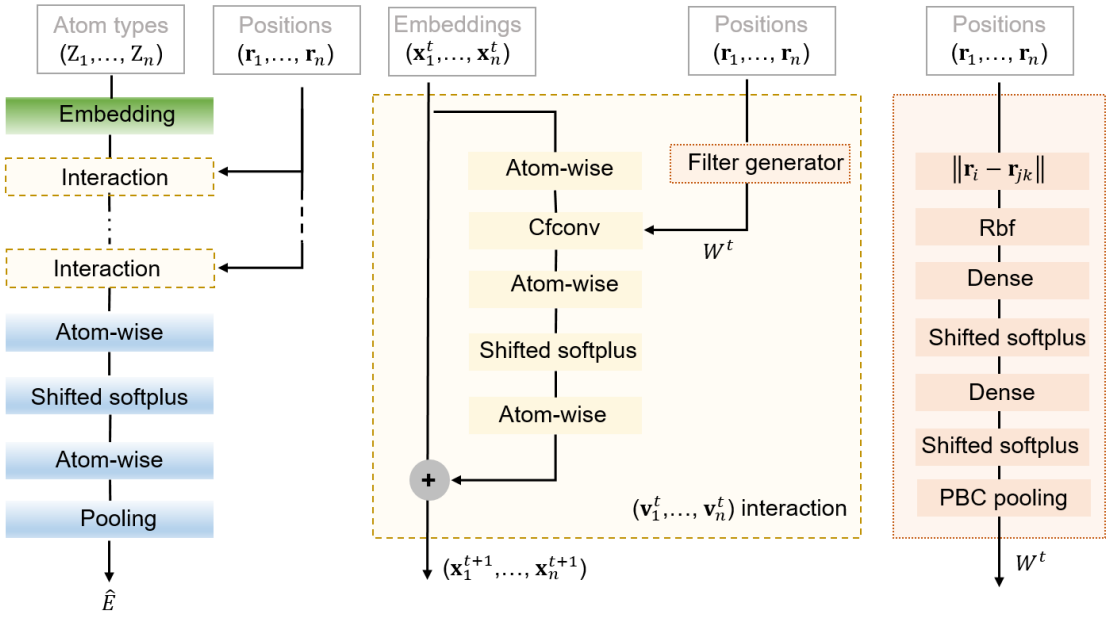



**Figure 10 The SchNet architecture**. In SchNet, only atom types as well as positions are needed as input data. The most important component is the interaction layer, in which atoms interact via continuous convolution functions with a filter generator. Figure adapted from reference [146].

during the training process using a deep NN. For example, SchNet and HIP-NN construct the atom-wise representations using interaction layers, which yield an increasingly complex and complete description of atomic neighborhoods as the number of layers increases. More details on end-to-end neural network representations can be found in recent reviews[16, 129, 145]. Below, we focus on SchNet as a representative end-to-end NN approach.

SchNet is an end-to-end deep convolutional neural network. It is able to learn representations of molecules and materials by only using atom types and positions of a system with minimal hyper-parameter tuning[146,152]. This end-to-end network can automatically adapt to atom-wise representations and the target properties for the given data. An overview of the SchNet architecture is depicted in Fig. 10. It shows that the SchNet network, first maps the atom types to $n$-dimensional embeddings to obtain the initial features. These features are then processed by several interaction blocks to encode the atomic neighborhood. The updated representations can subsequently be passed to a fully connected prediction network. The interactions are modeled using continuous-filter convolutional layers with a filter-generating network. These components are described in detail as follows.

***Atom Embedding***. Given atom types $\mathrm{Z}_1, .., \mathrm{Z}_n$ for an atomistic system containing $n$ atoms, the feature of atom $i$ is initialized by using an embedding layer which depends on the atom type $\mathrm{Z}_i$.

$$\mathbf{x}_i^{(0)} = \mathbf{A}_{\mathrm{Z}_i}. \tag{4.8}$$

These randomly initialized features currently without any information about surrounding environment will be optimized during training. Each atomic type corresponding to each atom is mapped to a vector to generate the initial representation in a given system.

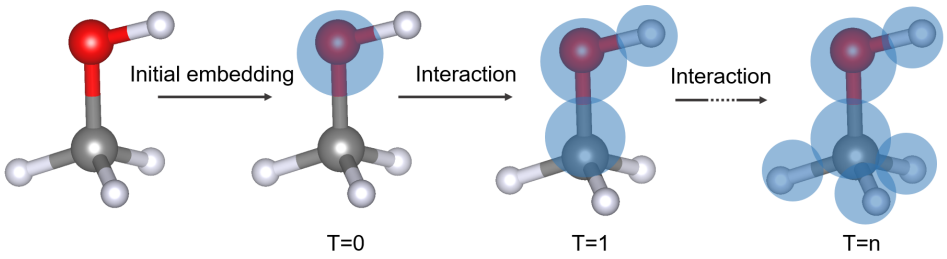



**Figure 11 Schematic illustration of the interaction blocks.** Following the initial embedding of the molecule, each atomic environment takes into account more interaction information between the neighboring environments by multiple times interactions ($\mathrm{T} = \mathrm{n}$) refinement. Figure adapted from reference [129].

***Interaction blocks***. Following the initial embedding, the features gain spatial information of the chemical environment by multiple pair-wise interaction corrections $\mathbf{v}^t$ with the surrounding atoms in interaction blocks[154] (as shown in Fig. 11), which are also known as message-passing steps.

$$\mathbf{x}_i^{t+1} = \mathbf{x}_i^t + \sum_{j \neq i} \mathbf{v}^t(\mathbf{x}_j^t, r_{ij}). \tag{4.9}$$

In SchNet, the interaction corrections $\mathbf{v}^t$ are modeled by utilizing continuous-filter convolutions with a smooth filter-generating network. The interactions of the atom $i$ can be obtained as the convolution with all neighboring atoms,

$$\mathrm{cfconv}((\mathbf{x}_i, r_i)) = \sum_{j \in \mathrm{nbh}(i)} \mathbf{x}_j \circ W_{\mathrm{filter}}(r_{ij}), \tag{4.10}$$

where '$\circ$' denotes the element-wise multiplication, $\mathrm{nbh}(i)$ is the neighborhood of atom $i$ (defined via a cutoff radius, as in SOAP). $W_{\mathrm{filter}}$ is the smooth filter-generating network, where the filter value $W_{\mathrm{filter}}(r_{ij})$ can be obtained in a fully-connected neural network depending on the pair-wise distance $r_{ij}$ to include rotational invariance in the model. Many-body terms thereby are achieved by the successive interactions. Notably, each interaction also increases the receptive field of the network beyond the original cutoff distance.

Besides convolution layers, the interaction blocks also contain the atom-wise, fully-connected layers to mix feature maps. These are defined as,

$$\mathrm{linear}(\mathbf{x}_i) = W^{\mathrm{T}}\mathbf{x}_i + \mathbf{b}, \tag{4.11}$$

where the atom-wise layers are applied to each atom $i$, while the weight $W$ and the bias $\mathbf{b}$ are independent of the atom $i$. The shifted softplus activation function $\mathrm{ssp}(x) = \ln(0.5e^x + 0.5)$ is usually used for non-linearities throughout the network.

***Aggregation and readout***. After $T$ interaction refinements, the final atom-wise representation $\mathbf{x}_i^{(T)}$ is obtained. This representation can be subsequently passed to a property-specific output network (i.e, a fully connected prediction network) to predict the property of interest. This output neural network usually consists of several atom-wise layers with non-linearities and an aggregation layer with a pooling function that recombines the atomic con-

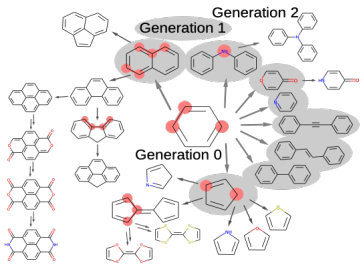
tributions to the final target. Different aggregation layers can be formulated for different target properties[145]. Typically, electronic properties can be distinguished into two types: extensive and intensive properties. The magnitude of extensive properties (e.g. the total energy) is additive with trivial extension of the system size, while the magnitude of intensive properties (e.g. the HOMO energy) is independent of the system size[10]. Depending on whether the property is intensive or extensive, the sum pooling is usually used for extensive properties, while it is common to use average pooling for intensive properties.

However, average pooling in some cases yields unphysical results for intensive properties prediction when the target properties are spatially localized or the system has low symmetry. More aggregation functions are available, such as max, softmax, set2set as well as self-attention[156–158]. As described in chapter 2, the HOMO energy is of great importance in OSCs and can easily be spatially localized, especially in molecular solids or polymers with low symmetries. Thus, the pooling functions for HOMO energies prediction for organic molecules were explored and discussed in chapter 5.3.

After building the end-to-end network, the model can then be trained to optimize the initial embedding vectors, the interaction blocks' parameters as well as the output network. It should be noted that the specific form of the loss function, the selected optimizer, the learning rate and the dimensions of the NN significantly influence the accuracy of the models. As for GPR, such hyperparameters should be judiciously selected.

# 5. Publications

## 5.1. Active Discovery of Organic Semiconductors



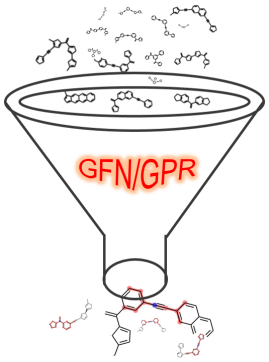

Christian Kunkel, Johannes T. Margraf, Ke Chen, Harald Oberhofer and Karsten Reuter

*Nature Communications* **2021**, 12, 2422.

**Summary:** Screening large design spaces for high charge mobility OSCs could help to uncover new materials for electronic device development. Here, efficient ML strategies can enable an unprecedented depth of design space exploration. To this end, we first design a scheme to generate an in principle unlimited space of potential OSC molecules. Inspired by the common building block strategies in materials design, a set of well-performing $\pi$-conjugated OSC molecules is analyzed and underlying molecular-construction rules (morphing operations) are derived. Through the successive application of these morphing operations (i.e. ring contraction, biphenyl addition, etc.) starting at benzene, a wide range of diverse molecules can be generated step-by-step going from small to large systems. This space is then explored by the active machine learning (AML) discovery strategy. Two descriptors (reorganization energy $\lambda$ and HOMO energy) which have been introduced in Chapter 2 are used to evaluate the qualification of candidate molecules for OSCs application. To obtain the descriptor values, 3D conformers are generated from 2D molecular graphs and the lowest-energy conformer of each molecule is selected to compute the descriptor values using DFT. The AML algorithm based on Gaussian Process Regression surrogate model is optimized in a limited test space. This optimized AML approach can then be applied to rapidly identify well-known, as well as hitherto unknown OSC molecules with prominent performance.

**Individual Contributions:** The project idea was conceived by Christian Kunkel, Harald Oberhofer and Karsten Reuter. Christian Kunkel was the main developer of this project and carried out the molecular space design as well as AML discovery modeling, with input from Johannes T. Margraf. Ke Chen benchmarked the accuracies of predicted geometries and reorganization energies of OSC molecules, comparing the accuracy of semi-empirical GFN1-xTB and first-principles B3LYP calculations. Furthermore, Ke Chen also tested the different AML query strategies such as bootstrapping and random selection. Christian Kunkel, Johannes T. Margraf, Harald Oberhofer and Karsten Reuter jointly wrote the manuscript. All authors discussed and revised the manuscript.

## 5.2. Reorganization Energies of Flexible Organic Molecules as a Challenging Target for Machine Learning Enhanced Virtual Screening




Ke Chen, Christian Kunkel, Karsten Reuter and Johannes T. Margraf

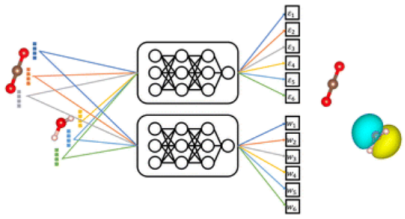
*Digital Discovery* **2022**, 1, 147-157.

**Summary:** The molecular reorganization energy $\lambda$ is an important target for high performance OSC molecules screening and design, since the charge carrier mobility of OSCs is highly influenced by $\lambda$. Unfortunately, $\lambda$ is a very complicated property, simultaneously depending on two potential energy surfaces. It is therefore sensitive to small geometrical changes and different molecular conformations. While ML models have been used to predict $\lambda$ for rigid molecules, larger and more flexible molecules still pose a significant challenge due to the conformational flexibility and the general difficulty of predicting $\lambda$ from the equilibrium geometries alone. In this contribution, we generate a set of highly flexible $\pi$-conjugated hydrocarbon molecules based on the procedure we developed in chapter 5.1. State-of-the-art conformer sampling methods are intensively explored and an efficient conformer search workflow is constructed by using semi-empirical electronic methods. Structure and electronic-property based GPR models are developed, respectively. We observe that the performance of the ML model is significantly influenced by the conformer sampling methods, revealing the importance of accurate conformer sampling, since incorrect conformers can introduce notable noise to the model. The performance of both models are significantly improved by adopting a $\Delta$-ML approach using a semiempirical baseline. In particular, combining structural and electronic properties kernels with the aid of $\Delta$-ML yielded the best performing models.
After obtaining a series of ML models, the usefulness of these models is evaluated by high throughput virtual screening (HTVS) in a diverse chemical space. Compared with a semiempirical screening, we find that the ML enhanced models are more efficient in identifying promising candidates, while the semiempircial model exhibits higher structural diversity. This reflects the fact that GPR model's predictions are based on feature similarity. Finally, we use the low $\lambda$ structures obtained from HTVS results to perform a substructure analysis so that general design rules can be derived to reveal promising building blocks.

**Individual Contributions:** As a follow up project of the project described in chapter 5.1, the project idea was initially conceived by Ke Chen, Johannes T. Margraf and Christian Kunkel.

The original dataset was provided by Christian Kunkel. Ke Chen built the conformer sampling workflow and carried out all the semi-empirical as well as DFT-based calculations. Ke Chen built and trained all ML models. Methodological details were worked out with Ke Chen, Christian Kunkel and Johannes T. Margraf. Ke Chen wrote the first version of the manuscript. Christian Kunkel, Johannes T. Margraf and Karsten Reuter jointly revised the manuscript.

## 5.3. Physics-Inspired Machine Learning of Localized Intensive Properties



Ke Chen, Christian Kunkel, Bingqing Cheng, Karsten Reuter and Johannes T. Margraf

*Chemical Science* **2023**, 14, 4913-4922.

**Summary:** As detailed above, most state-of-the-art molecular ML approaches are based on the idea of atom-centered local chemical environment representations. For size-extensive properties, the summation over atomic contributions is physically motivated and yields good accuracies in this context. For intensive properties, which do not scale linearly with trivial extensions of the system size, using size-extensive ML models can lead to large errors. This is particularly true when the property may be localized and the system has low symmetry. In order to address this question, we analyze the pooling functions that atomistic NNs use to aggregate atomic contributions and propose a series of physically motivated pooling functions for localized intensive properties, using the HOMO energy as a representative test case.

We build a novel dataset consisting of highly flexible organic molecules with a wide range of localization degrees. The diverse orbital distributions in the dataset allows us to study localized and delocalized orbitals in depth. This dataset is subsequently used to extensively benchmark different pooling functions. The newly developed orbital weighted average (OWA) approach, which can efficiently predict the HOMO energies as well as corresponding orbital locations, is shown to outperform the alternatives. In a nutshell, the OWA approach is based on two NNs, where one is used to predict the HOMO energies, while the second network is used to predict the atomic weights. The final target property is obtained via a weighted average operation of the two networks outputs. In particular, the NN model is trained using a joint loss function that depends both on the orbital locations and energies. The OWA methodology is subsequently applied to the highly challenging OE62 dataset, which consists of experimentally reported organic molecules with large structural diversity. The results show that the OWA model displays state-of-the-art performance for HOMO energy prediction on OE62, providing orbital localization information without extra computational cost. This is crucial for charge transfer analysis in OSCs. Overall, OWA can thus be recommended as a robust and physically motivated pooling function for orbital energy prediction as well as other localized intensive property predictions.

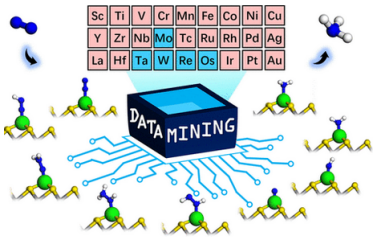
**Individual Contributions:** This project was conceptualized by Ke Chen and Johannes T. Margraf. Ke Chen implemented the concept and built the novel dataset as well as conducted all the NN models constructions and the models training. Methodological details were worked out by Ke Chen, Christian Kunkel, Bingqing Cheng and Johannes T. Margraf. Ke Chen wrote the first version of manuscript, which was then revised by Christian Kunkel, Bingqing Cheng,

Johannes T. Margraf and Karsten Reuter.

## 5.4. Further Work

The following article was published during my time working at the Chair of Theoretical Chemistry at TUM as a side project. This article is not an essential part of the dissertation, and it is mentioned here for the sake of completeness.

**Subgroup Discovery Points to the Prominent Role of Charge Transfer in Breaking Nitrogen Scaling Relations at Single-Atom Catalysts on $VS_2$.**




Haobo Li, Yunxia Liu, Ke Chen, Johannes T. Margraf, Youyong Li and Karsten Reuter

*ACS Catalysis* **2021**, 11, 7906–7914.

Colleagues from our catalysis research subgroup generated a database of DFT-computed adsorption energies for different intermediates of the nitrogen reduction reaction across 27 single-atom transition metals on a vanadium disulfide support. In this context, they found strongly broken scaling relations between calculated adsorption energies of different intermediates. Subsequently, I conducted a data-driven analysis by means of outlier detection and subgroup discovery to analyze these broken scaling relations. The data-driven analysis revealed that this breaking is restricted to early transition metals, which is in agreement with the subsequent electronic properties analysis revealing that the charge mainly transfers to the support for early transition metals.

# 6. Conclusion and Outlook

ML has been widely used in molecular and material science for accelerating simulations, predicting electronic properties, and designing new materials. In this work, we focus on two distinct complex electronic properties: the reorganization energy $\lambda$ and the HOMO energy, both of which are crucial for designing OSCs.

First, we explored the potential benefits of using ML models to enhance virtual screening for low $\lambda$ molecules. We find that $\lambda$ is challenging to predict due to the conformational flexibility and the difficulty to map the structure-property relationships from equilibrium geometry alone. To address this issue, an efficient conformer search workflow was constructed using semi-empirical electronic structure methods. Strategies to improve the ML performance for $\lambda$ prediction were explored. Here, the model performance was significantly improved by adopting a $\Delta$-ML strategy using a semiempirical baseline. However, we also found that ML is not necessarily the best choice for virtually screening for low-$\lambda$ molecules. Quantitatively, the data-driven ML models outperformed a semi-empirical method in identifying promising candidates, but they also yielded a less diverse sample.

In the second main project of this thesis work, ML models for HOMO energy prediction were considered. Since this is a size-intensive property, which can furthermore be spatially localized in the system, using size-extensive models here can be problematic. In this context, we proposed a series of potentially suitable pooling functions and tested them on a diverse orbital distribution dataset. A novel OWA approach was developed by joining the physical orbital information and HOMO energies in the loss function. This enabled the efficient and accurate prediction of HOMO energies and locations simultaneously. The OWA approach was furthermore successfully applied to the challenging OE62 dataset consisting of diverse experimentally reported molecules, indicating the robustness and high transferability of the method.

To conclude, this thesis explores strategies to improve the performance of molecular ML models for two challenging electronic properties. The conformer sampling workflows we built can reliably obtain conformers and could also contribute to other applications such as the design of molecular photoswitches[159]. Our methodological improvements are promising for accelerating and improving the virtual screening for low $\lambda$ molecules and assisting molecular design of OSCs more generally. The developed OWA approach can accurately predict localized intensive properties which are however challenging for the state-of-the-art ML models.

An ongoing project is the application of the OWA approach to investigating static and dynamic energetic disorder in amorphous organic semiconductors. This opens the door towards the multiscale modeling of realistic OSCs, as they are typically found in devices. Current ML models for energetic disorder calculation are limited to predicting HOMO energies[56]. Using the OWA approach, the energy disorder can be efficiently analyzed by predicting HOMO energies and locations, which will allow the visualization of its evolution along time and space.

Furthermore, the OWA approach is not only designed for orbital energies but also suitable

for other intensive properties, such as excitation energies, ionization energies and defect formation energies. The corresponding physical information incorporated into the model should however be adapted for each target of interest. In future work, we also plan to apply this approach to recent advanced neural network architectures (e.g., the Allegro[160] or MACE[161] approaches), in order to further improve the predictive performance of the localized intensive properties.

# Acknowledgements/Danksagung

First of all, I would like to thank Prof. Dr. Karsten Reuter to give me the chance to join this excellent family to carry out my research projects. I am always proud to share to other people how knowledgeable and friendly our professor is. Thanks for joining our birthday celebrations, festival celebrations as well as after-work outing activities. Thanks for offering nice work equipment as well as creating such a comfortable environment with a strong scientific as well as social atmosphere. Thanks for organizing the yearly workshop to help us to get to know our colleagues work, to spend time with us as well as invite so many excellent experts from the world to give seminar talks to gain a deeper scientific knowledge for us.

Next, I would like to thank my subgroup leader Dr. Johannes.T. Margraf. I am grateful to have met such an outstanding and responsible group leader that always has so many excellent ideas, which broaden my scientific horizon. Thanks for always providing valuable feedback and encouraging me. I sincerely appreciate your fast replies to my messages and emails regarding scientific questions or updates of my projects. It is a great pleasure to work in this subgroup. Thanks for helping me to organize an internship abroad to gain more research experience. I would also like to thank Dr. Christian Kunkel for these four amazing years of working together, which guided me a lot. Thanks for helping me to quickly understand and solve many scientific or technical issues, by which I quickly improved my scientific skills. I am grateful to have the chance to stand on the shoulders of giants.

Furthermore, I would like to thank all my colleagues from the whole theory department as well. Thanks Ruth, Julia, Steffen, Christoph for administrative and IT support. I have also so many good memories with Hanna, Christian, Carsten, Thorben, Hendrik and more colleagues during our two-days Hamburg stay. Thanks Hanna and Robert for visiting me when I was in Vienna, where we had good time exploring Vienna. Thanks Jakob for bringing us to experience the amazing Berlin techno music and the bar culture. Thanks Simiam for sharing the nice music and inviting me to attend a lot of activities together. Thanks Wenbin for helping me to quickly settle down after I arrived Germany. I would also like to thank my subgroup members Mengnan, Martin, Hyunwook, Elisabetha and more colleagues for the nice time of monthly subgroup outing events and regular lunch gatherings. It is awesome that we build such nice relationships. Great thanks to Simeon to help me to set up my workstations when we were in Munich. I would also thank many colleagues for the time we spent together during many conferences and workshops. Here I would also like to thank Markus in TUM, Bingqing and Felix in ISTA for helping me with the internship. I would like to give special thanks to Frau Kim for the great time we spent together. It was awesome that we had vacation in Bodensee and I am looking forward to the travel together as well as experiencing more in the future.

Last, I would like to thank my parents. Thanks for their supporting for my decision to study abroad. In these four years I suffered the biggest pain in my life. My father has left me forever last year in March after two and a half years torment. It was extremely unlucky that he suffered complications after cancer surgery. Thanks that he tried his best to fight with this sickness for my mother and me. I know that at last he was too painful to live. Thanks for my

mother who took care of my father and and stayed with him in the hospital for these whole two and a half years. I sincerely thank the people who helped my parents.

# Bibliography


[1] D. Jha, K. Choudhary, F. Tavazza, W.-k. Liao, A. Choudhary, C. Campbell, A. Agrawal, *Nat. Commun.* **2019**, *10*, 5316.

[2] J. Yang, S. De, J. E. Campbell, S. Li, M. Ceriotti, G. M. Day, *Chem. Mater.* **2018**, *30*, 4361–4371.

[3] S. Nagasawa, E. Al-Naamani, A. Saeki, *J. Phys. Chem. Lett.* **2018**, *9*, 2639–2646.

[4] R. Pederson, B. Kalita, K. Burke, *Nat. Rev. Phys.* **2022**, *4*, 357–358.

[5] L. Fiedler, K. Shah, M. Bussmann, A. Cangi, *Phys. Rev. Mater.* **2022**, *6*, 040301.

[6] J. Gubernatis, T. Lookman, *Phys. Rev. Mater.* **2018**, *2*, 120301.

[7] S. M. Moosavi, K. M. Jablonka, B. Smit, *J. Am. Chem. Soc.* **2020**, *142*, 20273–20287.

[8] E. Mazhnik, A. R. Oganov, *J. Appl. Phys.* **2020**, *128*, 075102.

[9] D. Packwood, L. T. H. Nguyen, P. Cesana, G. Zhang, A. Staykov, Y. Fukumoto, D. H. Nguyen, *Mach. Learn. Appl.* **2022**, *8*, 100265.

[10] M. F. Langer, A. Goeßmann, M. Rupp, *Npj Comput. Mater.* **2022**, *8*, 41.

[11] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, A. Walsh, *Nature* **2018**, *559*, 547–555.

[12] R. Vasudevan, G. Pilania, P. V. Balachandran, *J. Appl. Phys.* **2021**, *129*, 070401.

[13] J. Westermayr, M. Gastegger, K. T. Schütt, R. J. Maurer, *J. Chem. Phys.* **2021**, *154*, 230903.

[14] M. Ceriotti, C. Clementi, O. Anatole von Lilienfeld, *J. Chem. Phys.* **2021**, *154*, 160401.

[15] S. Axelrod, D. Schwalbe-Koda, S. Mohapatra, J. Damewood, K. P. Greenman, R. Gómez-Bombarelli, *Acc. Mater. Res.* **2022**, *3*, 343–357.

[16] J. A. Keith, V. Vassilev-Galindo, B. Cheng, S. Chmiela, M. Gastegger, K.-R. Müller, A. Tkatchenko, *Chem. Rev.* **2021**, *121*, 9816–9872.

[17] A. C. Mater, M. L. Coote, *J. Chem. Inf. Model.* **2019**, *59*, 2545–2559.

[18] D. Morgan, R. Jacobs, *Annu. Rev. Mater. Res.* **2020**, *50*, 71–103.

[19] D. M. Dimiduk, E. A. Holm, S. R. Niezgoda, *Integr. Mater. Manuf. Innov.* **2018**, *7*, 157–172.

[20] M. Rupp, A. Tkatchenko, K.-R. Müller, O. A. Von Lilienfeld, *Phys. Rev. Lett.* **2012**, *108*, 058301.

[21] C. G. Staacke, S. Wengert, C. Kunkel, G. Csanyi, K. Reuter, J. T. Margraf, *Mach. learn.: sci. technol.* **2022**, *3*, 015032.

[22] B. Mazouin, A. A. Schöpfer, O. A. von Lilienfeld, *Mater. Adv.* **2022**, *3*, 8306–8316.

[23] J. Westermayr, P. Marquetand, *Chem. Rev.* **2020**, *121*, 9873–9926.

[24] Y. Liu, Z. Li, *J. Chem. Inf. Model.* **2023**, *63*, 806–814.

[25] R. Ramakrishnan, P. O. Dral, M. Rupp, O. A. Von Lilienfeld, *Sci. Data* **2014**, *1*, 1–7.

[26] A. P. Bartók, R. Kondor, G. Csányi, *Phys. Rev. B* **2013**, *87*, 184115.

[27] H. Huo, M. Rupp, *Mach. learn.: sci. technol.* **2022**, *3*, 045017.

[28] C. Kunkel, J. T. Margraf, K. Chen, H. Oberhofer, K. Reuter, *Nat. Commun.* **2021**, *12*, 2422.

[29] K. Chen, C. Kunkel, K. Reuter, J. T. Margraf, *Digital Discovery* **2022**, *1*, 147–157.

[30] H. Oberhofer, K. Reuter, J. Blumberger, *Chem. Rev.* **2017**, *117*, 10319–10357.

[31] C. Schober, K. Reuter, H. Oberhofer, *J. Phys. Chem. Lett.* **2016**, *7*, 3973–3977.

[32] C. O. Schober, PhD thesis, Technische Universität München, **2017**.

[33] C. Kunkel, C. Schober, J. T. Margraf, K. Reuter, H. Oberhofer, *Chem. Mater.* **2019**, *31*, 969–978.

[34] S. Atahan-Evrenk, F. B. Atalay, *J. Phys. Chem. A* **2019**, *123*, 7855–7863.

[35] O. D. Abarbanel, G. R. Hutchison, *J. Chem. Phys.* **2021**, *155*.

[36] K. Chen, C. Kunkel, B. Cheng, K. Reuter, J. T. Margraf, *Chem. Sci.* **2023**, *14*, 4913–4922.

[37] J. E. Anthony, A. Facchetti, M. Heeney, S. R. Marder, X. Zhan, *Adv. Mater.* **2010**, *22*, 3876–3892.

[38] P. Friederich, A. Fediai, S. Kaiser, M. Konrad, N. Jung, W. Wenzel, *Adv. Mater.* **2019**, *31*, 1808256.

[39] D. Yang, D. Ma, *Adv. Opt. Mater.* **2019**, *7*, 1800522.

[40] A. Mishra, P. Bäuerle, *Angew. Chem. Int. Ed.* **2012**, *51*, 2020–2067.

[41] Y. Zhang, A. Chen, M.-W. Kim, A. Alaei, S. S. Lee, *Chem. Soc. Rev.* **2021**, *50*, 9375–9390.

[42] Y. Zhang, Y. Wang, C. Gao, Z. Ni, X. Zhang, W. Hu, H. Dong, *Chem. Soc. Rev.* **2023**, *52*, 1331–1381.

[43] C. Wang, H. Dong, L. Jiang, W. Hu, *Chem. Soc. Rev.* **2018**, *47*, 422–500.

[44] I. D. W. Samuel, G. A. Turnbull, *Chem. Rev.* **2007**, *107*, 1272–1295.

[45] S. Allard, M. Forster, B. Souharce, H. Thiem, U. Scherf, *Angew. Chem. Int. Ed.* **2008**, *47*, 4070–4098.

[46] B. Kippelen, J.-L. Brédas, *Energy Environ. Sci.* **2009**, *2*, 251–261.

[47] B. Geffroy, P. Le Roy, C. Prat, *Polym. Int.* **2006**, *55*, 572–582.

[48] G. R. Hutchison, M. A. Ratner, T. J. Marks, *J. Am. Chem. Soc.* **2005**, *127*, 2339–2350.

[49] R. Oshi, S. Abdalla, M. Springborg, *Eur. Phys. J. D* **2019**, *73*, 124.

[50] G. Gryn'ova, K.-H. Lin, C. Corminboeuf, *J. Am. Chem. Soc.* **2018**, *140*, 16370–16386.

[51] R. A. Marcus, *J. Chem. Phys.* **1956**, *24*, 966–978.

[52] R. A. Marcus, *Rev. Mod. Phys.* **1993**, *65*, 599.

[53] T. Tan, D. Wang, *J. Chem. Phys.* **2023**, *158*, 094102.

[54] J. Nelson, J. Kwiatkowski, J. Kirkpatrick, J. Frost, *Acc. Chem. Res* **2009**, *42*, 1768–1778.

[55] S. F. Nelsen, S. C. Blackstock, Y. Kim, *J. Am. Chem. Soc.* **1987**, *109*, 677–682.

[56] P. Reiser, M. Konrad, A. Fediai, S. Léon, W. Wenzel, P. Friederich, *J. Chem. Theory Comput.* **2021**, *17*, 3750–3759.

[57] H. Ishii, K. Sugiyama, E. Ito, K. Seki, *Adv. Mater.* **1999**, *11*, 605–625.

[58] P. Li, Z.-H. Lu, *Small Science* **2021**, *1*, 2000015.

[59] M. Waldrip, O. D. Jurchescu, D. J. Gundlach, E. G. Bittle, *Adv. Funct. Mater.* **2020**, *30*, 1904576.

[60] A. Franco-Cañellas, S. Duhm, A. Gerlach, F. Schreiber, *Rep. Prog. Phys.* **2020**, *83*, 066501.

[61] M.-C. Lu, R.-B. Wang, A. Yang, S. Duhm, *J. Condens. Matter Phys.* **2016**, *28*, 094005.

[62] P. E. Schwenn, P. L. Burn, B. J. Powell, *Org. Electron.* **2011**, *12*, 394–403.

[63] A. Stuke, M. Todorović, M. Rupp, C. Kunkel, K. Ghosh, L. Himanen, P. Rinke, *J. Chem. Phys.* **2019**, *150*, 204121.

[64] F. Pereira, K. Xiao, D. A. Latino, C. Wu, Q. Zhang, J. Aires-de-Sousa, *J. Chem. Inf. Model.* **2017**, *57*, 11–21.

[65] Z. Shuai, H. Geng, W. Xu, Y. Liao, J.-M. André, *Chem. Soc. Rev.* **2014**, *43*, 2662–2679.

[66] L. Xiang, J. L. Palma, C. Bruot, V. Mujica, M. A. Ratner, N. Tao, *Nat. Chem.* **2015**, *7*, 221–226.

[67] K. Fukui, *science* **1982**, *218*, 747–754.

[68] M. Morikawa, K. Kino, T. Oyoshi, M. Suzuki, T. Kobayashi, H. Miyazawa, *Bioorganic Med. Chem. Lett.* **2015**, *25*, 3359–3362.

[69] S. Axelrod, R. Gomez-Bombarelli, *Sci. Data* **2022**, *9*, 185.

[70] S. Axelrod, R. Gomez-Bombarelli, *arXiv preprint arXiv:2012.08452* **2020**.

[71] P. C. Hawkins, S. Wlodek, *J. Chem. Inf. Model.* **2020**, *60*, 3518–3533.

[72] P. C. Hawkins, *J. Chem. Inf. Model.* **2017**, *57*, 1747–1756.

[73] L. Fang, E. Makkonen, M. Todorović, P. Rinke, X. Chen, *J. Chem. Theory Comput.* **2021**, *17*, 1955–1966.

[74] G. Zhou, Z. Gao, Z. Wei, H. Zheng, G. Ke, *arXiv preprint arXiv:2302.07061* **2023**.

[75] D. C. Spellmeyer, A. K. Wong, M. J. Bower, J. M. Blaney, *J. Mol. Graph. Model.* **1997**, *15*, 18–36.

[76] E. Mansimov, O. Mahmood, S. Kang, K. Cho, *Sci. Rep.* **2019**, *9*, 20381.

[77] M. Xu, L. Yu, Y. Song, C. Shi, S. Ermon, J. Tang, *arXiv preprint arXiv:2203.02923* **2022**.

[78] T. Gogineni, Z. Xu, E. Punzalan, R. Jiang, J. Kammeraad, A. Tewari, P. Zimmerman, *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 20142–20153.

[79] D. D. Beusen, E. B. Shands, S. Karasek, G. R. Marshall, R. A. Dammkoehler, *J. Mol. Struct.: THEOCHEM* **1996**, *370*, 157–171.

[80] H. Tsujishita, S. Hirono, *J. Comput. Aided Mol. Des.* **1997**, *11*, 305–315.

[81] S. R. Wilson, W. Cui, J. W. Moskowitz, K. E. Schmidt, *J. Comput. Chem.* **1991**, *12*, 342–349.

[82] M. J. Vainio, M. S. Johnson, *J. Chem. Inf. Model.* **2007**, *47*, 2462–2474.

[83] S. Riniker, G. A. Landrum, *J. Chem. Inf. Model.* **2015**, *55*, 2562–2574.

[84] P. C. Hawkins, A. G. Skillman, G. L. Warren, B. A. Ellingson, M. T. Stahl, *J. Chem. Inf. Model.* **2010**, *50*, 572–584.

[85] J. C. Cole, O. Korb, P. McCabe, M. G. Read, R. Taylor, *J. Chem. Inf. Model.* **2018**, *58*, 615–629.

[86] S. Kothiwale, J. L. Mendenhall, J. Meiler, *J. Cheminform.* **2015**, *7*, 47.

[87] C. Schärfer, T. Schulz-Gasch, J. Hert, L. Heinzerling, B. Schulz, T. Inhester, M. Stahl, M. Rarey, *ChemMedChem* **2013**, *8*, 1690–1700.

[88] W. Guba, A. Meyder, M. Rarey, J. Hert, *J. Chem. Inf. Model.* **2016**, *56*, 1–5.

[89] C. R. Groom, I. J. Bruno, M. P. Lightfoot, S. C. Ward, *Acta Crystallogr. B: Struct. Sci. Cryst. Eng. Mater.* **2016**, *72*, 171–179.

[90] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, *Nucleic Acids Res.* **2000**, *28*, 235–242.

[91] S. Grimme, *J. Chem. Theory Comput.* **2019**, *15*, 2847–2862.

[92] P. Pracht, F. Bohle, S. Grimme, *Phys. Chem. Chem. Phys.* **2020**, *22*, 7169–7192.

[93] I. Y. Kanal, J. A. Keith, G. R. Hutchison, *Int. J. Quantum Chem.* **2018**, *118*, e25512.

[94] P. Tosco, N. Stiefl, G. Landrum, *J. Cheminform.* **2014**, *6*, 37.

[95] C. Bannwarth, E. Caldeweyher, S. Ehlert, A. Hansen, P. Pracht, J. Seibert, S. Spicher, S. Grimme, *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2021**, *11*, e1493.

[96] S. Grimme, C. Bannwarth, P. Shushkov, *J. Chem. Theory Comput.* **2017**, *13*, 1989–2009.

[97] C. Bannwarth, S. Ehlert, S. Grimme, *J. Chem. Theory Comput.* **2019**, *15*, 1652–1671.

[98] P. Pracht, E. Caldeweyher, S. Ehlert, S. Grimme, **2019**, 1–19.

[99] The RDKit: Open-Source Cheminformatics Software, version 2019.09.3., 2019, `http://www.rdkit.org`.

[100] J. M. Blaney, J. S. Dixon, *Rev. Comput. Chem.* **1994**, 299–335.

[101] J.-P. Ebejer, G. M. Morris, C. M. Deane, *J. Chem. Inf. Model.* **2012**, *52*, 1146–1158.

[102] N.-O. Friedrich, F. Flachsenberg, A. Meyder, K. Sommer, J. Kirchmair, M. Rarey, *J. Chem. Inf. Model.* **2019**, *59*, 731–742.

[103] N.-O. Friedrich, A. Meyder, C. de Bruyn Kops, K. Sommer, F. Flachsenberg, M. Rarey, J. Kirchmair, *J. Chem. Inf. Model.* **2017**, *57*, 529–539.

[104] N.-O. Friedrich, C. de Bruyn Kops, F. Flachsenberg, K. Sommer, M. Rarey, J. Kirchmair, *J. Chem. Inf. Model.* **2017**, *57*, 2719–2728.

[105] A. R. Romero, PhD thesis, University of Groningen, **2021**.

[106] S. A. Ghasemi, A. Hofstetter, S. Saha, S. Goedecker, *Phys. Rev. B* **2015**, *92*, 045131.

[107] P. Pracht, D. F. Grant, S. Grimme, *J. Chem. Theory Comput.* **2020**, *16*, 7044–7060.

[108] S. Dohm, M. Bursch, A. Hansen, S. Grimme, *J. Chem. Theory Comput.* **2020**, *16*, 2002–2012.

[109] S. Ehlert, S. Grimme, A. Hansen, *J. Phys. Chem. A* **2022**, *126*, 3521–3535.

[110] S. Schmitz, J. Seibert, K. Ostermeir, A. Hansen, A. H. Göller, S. Grimme, *J. Phys. Chem. B* **2020**, *124*, 3636–3646.

[111] S. Spicher, M. Bursch, S. Grimme, *J. Phys. Chem. C* **2020**, *124*, 27529–27541.

[112] S. Spicher, S. Grimme, *J. Phys. Chem. Lett.* **2020**, *11*, 6606–6611.

[113] M. Bursch, H. Neugebauer, S. Grimme, *Angew. Chem. Int. Ed.* **2019**, *58*, 11078–11087.

[114] M. Bursch, A. Hansen, P. Pracht, J. T. Kohn, S. Grimme, *Phys. Chem. Chem. Phys.* **2021**, *23*, 287–299.

[115] S. Spicher, S. Grimme, *Angew. Chem. Int. Ed.* **2020**, *59*, 15665–15673.

[116] J. Gorges, S. Grimme, A. Hansen, P. Pracht, *Phys. Chem. Chem. Phys.* **2022**, *24*, 12249–12259.

[117] P. Pracht, S. Grimme, *Chem. Sci.* **2021**, *12*, 6551–6568.

[118] S. Grimme, F. Bohle, A. Hansen, P. Pracht, S. Spicher, M. Stahn, *J. Phys. Chem. A* **2021**, *125*, 4039–4054.

[119] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, L. Zdeborová, *Rev. Mod. Phys.* **2019**, *91*, 045002.

[120] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, C. Kim, *Npj Comput. Mater.* **2017**, *3*, 54.

[121] P. Friederich, F. Häse, J. Proppe, A. Aspuru-Guzik, *Nat. Mater.* **2021**, *20*, 750–761.

[122] V. L. Deringer, M. A. Caro, G. Csányi, *Adv. Mater.* **2019**, *31*, 1902765.

[123] B. Dou, Z. Zhu, E. Merkurjev, L. Ke, L. Chen, J. Jiang, Y. Zhu, J. Liu, B. Zhang, G.-W. Wei, *Chem. Rev.* **2023**, *123*, 8736–8780.

[124] G. Chen, Z. Shen, A. Iyer, U. F. Ghumman, S. Tang, J. Bi, W. Chen, Y. Li, *Polymers* **2020**, *12*, 163.

[125] V. Vassilev-Galindo, G. Fonseca, I. Poltavsky, A. Tkatchenko, *J. Chem. Phys.* **2021**, *154*, 094119.

[126] F. Musil, A. Grisafi, A. P. Bartók, C. Ortner, G. Csányi, M. Ceriotti, *Chem. Rev.* **2021**, *121*, 9759–9815.

[127] B. Huang, N. O. Symonds, O. A. von Lilienfeld, *Handbook of Materials Modeling: Methods: Theory and Modeling* **2020**, 1883–1909.

[128] M. Ceriotti, M. J. Willatt, G. Csányi, *Handbook of Materials Modeling: Methods: Theory and Modeling* **2020**, 1911–1937.

[129] K. M. Jablonka, D. Ongari, S. M. Moosavi, B. Smit, *Chem. Rev.* **2020**, *120*, 8066–8129.

[130] B. Huang, O. A. Von Lilienfeld, *Chem. Rev.* **2021**, *121*, 10001–10036.

[131] J. Damewood, J. Karaguesian, J. R. Lunger, A. R. Tan, M. Xie, J. Peng, R. Gómez-Bombarelli, *Annu. Rev. Mater. Res.* **2023**, *53*.

[132] R. Drautz, *Phys. Rev. B* **2019**, *99*, 014104.

[133] F. A. Faber, A. S. Christensen, B. Huang, O. A. Von Lilienfeld, *J. Chem. Phys.* **2018**, *148*, 241717.

[134] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. Von Lilienfeld, K.-R. Muller, A. Tkatchenko, *J. Phys. Chem. Lett.* **2015**, *6*, 2326–2331.

[135] A. P. Bartók, M. C. Payne, R. Kondor, G. Csányi, *Phys. Rev. Lett.* **2010**, *104*, 136403.

[136] J. Behler, M. Parrinello, *Phys. Rev. Lett.* **2007**, *98*, 146401.

[137] S. Stocker, G. Csányi, K. Reuter, J. T. Margraf, *Nat. Commun.* **2020**, *11*, 5505.

[138] L. Himanen, M. O. Jäger, E. V. Morooka, F. F. Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke, A. S. Foster, *Comput. Phys. Commun.* **2020**, *247*, 106949.

[139] B. Cheng, R.-R. Griffiths, S. Wengert, C. Kunkel, T. Stenczel, B. Zhu, V. L. Deringer, N. Bernstein, J. T. Margraf, K. Reuter, et al., *Acc. Chem. Res.* **2020**, *53*, 1981–1991.

[140] S. A. Meldgaard, E. L. Kolsbjerg, B. Hammer, *J. Chem. Phys.* **2018**, *149*, 134104.

[141] Z. Qiao, M. Welborn, A. Anandkumar, F. R. Manby, T. F. Miller, *J. Chem. Phys.* **2020**, *153*, 124111.

[142] J. T. Margraf, K. Reuter, *Nat. Commun.* **2021**, *12*, 344.

[143] K. Karandashev, O. A. von Lilienfeld, *J. Chem. Phys.* **2022**, *156*, 114101.

[144] G. G. Terrones, C. Duan, A. Nandy, H. J. Kulik, *Chem. Sci.* **2023**, *14*, 1419–1433.

[145] K. T. Schütt, S. Chmiela, O. A. von Lilienfeld, A. Tkatchenko, K. Tsuda, K.-R. Müller, *Lecture Notes in Physics* **2020**.

[146] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, K.-R. Müller, *J. Chem. Phys.* **2018**, *148*, 241722.

[147] E. Schulz, M. Speekenbrink, A. Krause, *J. Math. Psychol.* **2018**, *85*, 1–16.

[148] T. Hofmann, B. Schölkopf, A. J. Smola, *Ann. Stat.* **2008**, *36*, 1171–1220.

[149] L. Ralaivola, S. J. Swamidass, H. Saigo, P. Baldi, *Neural networks* **2005**, *18*, 1093–1110.

[150] V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti, G. Csányi, *Chem. Rev.* **2021**, *121*, 10073–10141.

[151] J. Vandermause, S. B. Torrisi, S. Batzner, Y. Xie, L. Sun, A. M. Kolpak, B. Kozinsky, *Npj Comput. Mater.* **2020**, *6*, 20.

[152] K. Schütt, P. Kessel, M. Gastegger, K. Nicoli, A. Tkatchenko, K.-R. Müller, *J. Chem. Theory Comput.* **2018**, *15*, 448–455.

[153] N. Lubbers, J. S. Smith, K. Barros, *J. Chem. Phys.* **2018**, *148*, 241715.

[154] K. T. Schütt, M. Gastegger, A. Tkatchenko, K.-R. Müller, *Explainable AI: Interpreting Explaining and Visualizing Deep Learning* **2019**, 311–330.

[155] J. S. Smith, O. Isayev, A. E. Roitberg, *Chem. Sci.* **2017**, *8*, 3192–3203.

[156] A. M. Schweidtmann, J. G. Rittig, J. M. Weber, M. Grohe, M. Dahmen, K. Leonhard, A. Mitsos, *Comput. Chem. Eng.* **2023**, *172*, 108202.

[157] O. Vinyals, S. Bengio, M. Kudlur, *arXiv preprint arXiv:1511.06391* **2015**.

[158] J. Lee, I. Lee, J. Kang in International conference on machine learning, PMLR, **2019**, pp. 3734–3743.

[159] R.-R. Griffiths, J. L. Greenfield, A. R. Thawani, A. R. Jamasb, H. B. Moss, A. Bourached, P. Jones, W. McCorkindale, A. A. Aldrick, M. J. Fuchter, et al., *Chem. Sci.* **2022**, *13*, 13541–13551.

[160] A. Musaelian, S. Batzner, A. Johansson, L. Sun, C. J. Owen, M. Kornbluth, B. Kozinsky, *Nat. Commun.* **2023**, *14*, 579.

[161] I. Batatia, D. P. Kovacs, G. Simm, C. Ortner, G. Csányi, *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 11423–11436.

# List of Figures

# Appendix

# Paper 1

**Active Discovery of Organic Semiconductors**

Christian Kunkel, Johannes T. Margraf, Ke Chen, Harald Oberhofer and Karsten Reuter

*Nature Communications* **2021**, *12*, 2422.

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-021-22611-4 OPEN

# Active discovery of organic semiconductors

Christian Kunkel [1], Johannes T. Margraf [1], Ke Chen [1], Harald Oberhofer [1] & Karsten Reuter [1,2]✉

The versatility of organic molecules generates a rich design space for organic semiconductors (OSCs) considered for electronics applications. Offering unparalleled promise for materials discovery, the vastness of this design space also dictates efficient search strategies. Here, we present an active machine learning (AML) approach that explores an unlimited search space through consecutive application of molecular morphing operations. Evaluating the suitability of OSC candidates on the basis of charge injection and mobility descriptors, the approach successively queries predictive-quality first-principles calculations to build a refining surrogate model. The AML approach is optimized in a truncated test space, providing deep methodological insight by visualizing it as a chemical space network. Significantly outperforming a conventional computational funnel, the optimized AML approach rapidly identifies well-known and hitherto unknown molecular OSC candidates with superior charge conduction properties. Most importantly, it constantly finds further candidates with highest efficiency while continuing its exploration of the endless design space.

[1] Chair for Theoretical Chemistry and Catalysis Research Center, Technische Universität München, Garching, Germany. [2] Fritz-Haber-Institut der Max-Planck-Gesellschaft, Berlin, Germany. ✉email: reuter@fhi-berlin.mpg.de

The sheer vastness of chemical spaces[1] has long motivated prior-to-synthesis virtual discovery. In corresponding work, promising candidate molecules or materials for refined study are often searched and identified on the basis of a small number of quantities that are deemed representative for the targeted application[2–4]. Prevalent for first-principles computational screening approaches is to calculate such descriptors at predictive quality through electronic structure theory for every candidate in a somehow enumerated chemical space or otherwise given database. Initially performed for small focused libraries, the screening is now extended to search spaces of ever increasing size and—since discovery is limited to the explicitly considered molecules or materials—to ever more systematic and exhaustive enumerations within these spaces.

Unfortunately, the combinatorial explosion characteristic for chemical versatility quickly leads to intractable numbers of candidates for such exhaustive first-principles screenings, even if based on computationally comparably undemanding descriptors. A common strategy to tackle this problem is a computational funnel[5]. Here, the exhaustive screening is only performed for computationally least-demanding descriptors or even less demanding estimates thereof. Subsequently, the large candidate set is narrowed in staged filtering and the calculation of other descriptors is only performed for smaller and smaller subsets which appear promising in terms of the previously calculated descriptors. Unfortunately, chemical diversity suggests the multi-objective (descriptor) landscape spanned over the search space to be quite rugged[6], with molecular or materials sub-classes likely constituting separate funnels and related analogs leading to multiple local minima. This raises concerns whether the true optimum candidates can reliably be identified through such computational funneling.

An ever more appealing alternative is therefore to completely abandon the original idea to exhaustively screen a once defined chemical space or database. Instead, the explicit first-principles computation of the descriptors is restricted to candidates emerging in an iteratively refining search[7–9]. In the context of data science, this is afforded by several learning concepts, which additionally allow to even avoid predefining or a priori enumerating the search space itself. Examples include (semi-)supervised learning, meta-, transfer-, or few-shot learning and generative models[10,11]. For drug-discovery tasks[12,13], such concepts have already been successfully employed to further accelerate molecular de novo design[14] and drive autonomous discovery[15]. For materials discovery based on first-principles descriptors, in particular active machine learning (AML)[16] has been explored as a most data-efficient method[17–22].

In AML, the acquired knowledge in form of explicitly calculated descriptors is used to successively establish a surrogate model of larger and larger regions of the rugged descriptor landscape. In an iterative procedure, the predictive-quality calculations for new candidates can then also be balanced between exploitation and exploration. In exploitation, the global insight provided by the current surrogate model is used for a targeted identification of new promising candidates. In exploration, descriptors for new candidates are specifically calculated to refine and extend the surrogate model. For this, we here employ Gaussian Process Regression (GPR) and use high values of its inherent Bayesian uncertainty estimate to flag candidates (or regions in chemical space) for which an explicit descriptor calculation will maximally contribute new information.

We pursue this concept for the efficient virtual discovery of organic semiconductors (OSCs) for electronic applications. Used in organic field effect transistors (OFETs),[23] photovoltaics (OPVs),[24] or light emitting diodes (OLEDs),[25] OSCs offer great versatility and novel materials' properties, paired with a low ecologic and economic footprint. Typical OSC-constituting molecules are, however, of considerable size (e.g., 22 or 42 non-hydrogen atoms in the classic examples pentacene or rubrene, respectively) and the spanned electronic property landscapes are known to be highly sensitive even to small molecular substitutions.[26–28] A vast number of ~$10^{33}$ similar-sized molecules is estimated to be synthesizable[1], raising the suspicion that presently known well-performing OSC molecular materials are not even the tip of the iceberg. This has motivated a number of preceding exhaustive screening or virtual discovery studies in more or less restricted closed subspaces.[3,5,29–34].

In this work we first analyze a diverse set of OSC molecules to derive clear molecular-construction rules that allow to generate an in principle unlimited OSC chemical space. This space is then successively explored by the AML discovery strategy, rapidly identifying molecular candidates that are superior to well-known OSC materials in terms of their molecular electronic descriptors assessing efficient charge injection and charge mobility. Deep methodological insight is gained by analyzing and visualizing the AML exploration inside a chemical space network (CSN) containing only a subset of the design space, limited to allow its full enumeration. Even inside this truncated chemical space the AML-discovery clearly outperforms a conventional funnel approach.

## Results

**Morphing based generation of an unlimited OSC search space.** The basis for our efficient AML exploration of an a priori unlimited molecular search space is the development of a concise set of molecular construction rules that allow to generate this space by iterative application. To establish a diverse, but problem-specific chemical space, we resort to existing domain knowledge and analyze the building blocks and motives contained in molecules constituting a number of well-performing crystalline OSC molecular materials. For this analysis, we exploit the fact that most functionalized organic molecules can be unambiguously fragmented into a molecular backbone (of one or more cores), linkers (that connect cores) and side groups (attached to cores) as illustrated in Fig. 1. Without loss of generality, we correspondingly fragment 30 prominent π-conjugated molecules that belong to a variety of important molecular families[23] (Acenes, Thienoacenes, TTF-derivatives, Carbazoles, Triphenylamines, Diimides, Quinacridones and Azaacenes) and consist of the most common organic elements C, H, N, O and S. Figure 1 highlights some of these peer molecules and the full set is given in the SI in Supplementary Fig. 1. Intriguingly, the richness of chemical building blocks identified in this way can be exhaustively generated by a set of only 22 simple molecular morphing operations starting from the smallest aromatic building block benzene. As illustrated in Fig. 1 these morphing operations each act on a molecule's individual atomic sites or fragments, each time adding, modifying or removing fragments. These morphing operations should be seen as alchemical transformations to navigate between molecules, while applying organic synthesis steps could be a viable alternative.[35] Even though at a first glance rather unintuitive for the generation of successively larger or complex molecules, we also note that the inclusion of every morphing operation in a backwards step, i.e., resubstituting a fragment substructure, is crucial to increase the interconnectivity of the forming chemical space, see Supplementary Fig. 3.

The generic nature of the morphing operations identified through the fragmentation ansatz is not only a stepping stone for the efficient AML exploration. It also provides a blueprint for future variations of the present search space or the generation of different search spaces for other applications. Additional morphing operations will lead to more general search spaces and could be automatically extracted from a diverse chemical database[36],

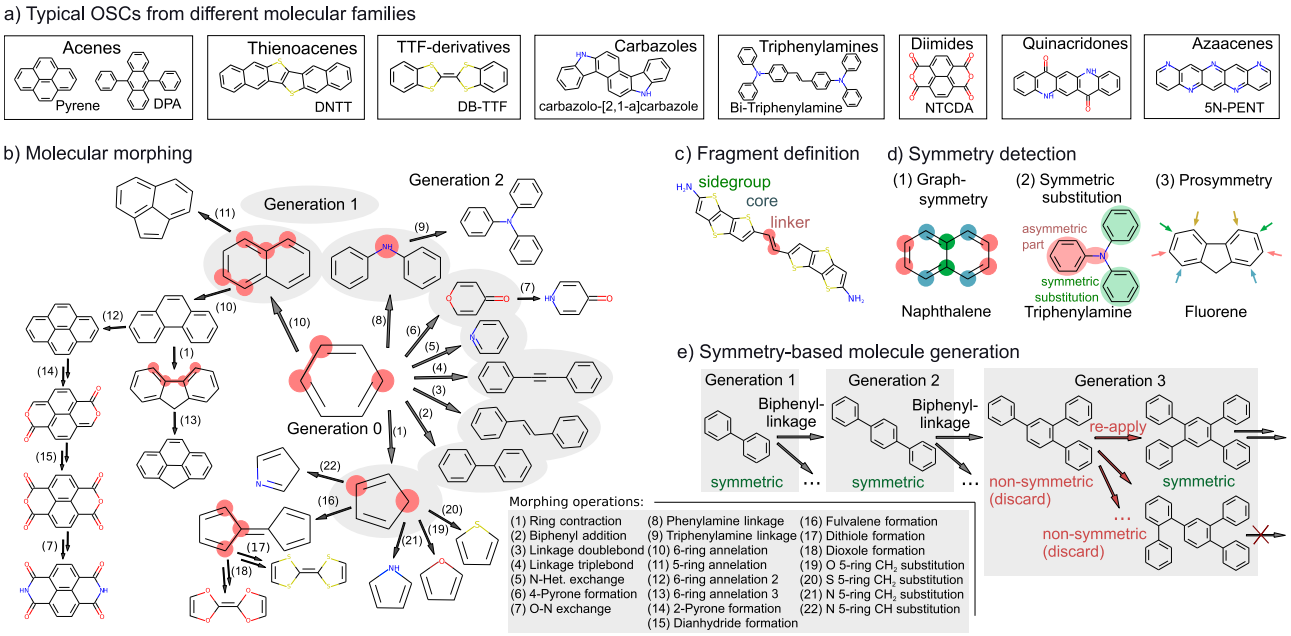



**Fig. 1 Molecular construction approach to generate an unlimited OSC chemical space. a** Important π-conjugated molecular families and examples of well-performing OSC-molecules therein. Molecular morphing operations are designed such that the generated OSC space includes these families. **b** Schematic overview of the molecular generation process. Starting from benzene, diverse molecules are created by iterative application of up to 22 morphing operations. The first generation resulting from the 8 morphing operations applicable to benzene is fully shown. Molecules in further generations are only shown as examples, but every operation type is depicted at least once, see also Supplementary Fig. 2 for an extended depiction. **c** Fragment-definitions used throughout the text exemplified for the molecule BDTTE. Connected aromatic ring structures are cores. Linkers and sidegroups both branch from a core structure with a single bond, but are either connecting to at least two core structures or only bonded to one core fragment. **d** Concepts for symmetry detection used throughout the molecular generation process. **e** Modified molecular morphing step, adapted to the symmetry constraints imposed on candidate molecules.

while deliberate suppression of morphing operations can be used to focus on molecular sub-classes. Ring-annelation type morphing operations as well as biphenylic addition are for example essential for the iterative construction of core Acene fragments, such as in Pyrene or DPA. To build structures like Thienoacenes, Azaacenes or Carbazoles, ring contractions that lead to 5-membered rings are included as intermediates for heteroaromatic ring construction. This, though, comes at the cost of potentially yielding pericyclically reactive molecules, as discussed further below. Similarly, two types of linker operations are included to access the family of Triphenylamines. Further examples together with a detailed description of every morphing operation are provided in Supplementary Note 1. Considering their known OSC tuning potential,[28,37,38] we note that in particular the augmentation of the present backbone-oriented set of construction rules by specific morphing operations for side groups or additional functional groups is expected to lead to an important extension of the here showcased search space.

The construction rules may also be modified to incorporate further prior knowledge about the OSC design problem. Here, we notably include constraints on molecular symmetry. Molecular symmetry may be beneficial for synthetic accessibility. Furthermore, it can mitigate mobility reducing charge localization[27] and in particular in monomolecular crystals often favors charge percolation pathways[3,39,40] (albeit its role can be intricate[41]). We correspondingly prune the construction rules for the present OSC context to enforce 2D graph symmetries expected to provide a prosymmetry for the 3D case. Specifically, generated molecules are only considered for further morphing, if they fall into three types of symmetry classes as explained in Fig. 1d, e: They (1) exhibit a full graph-symmetry, with all atomic environments appearing at least twice. (2) An asymmetric part in the molecule made of one or more fragments is symmetrically substituted by an even number of similar fragments, or (3) a molecule is prosymmetric such that it has atomic sites on which a single substitution operation could lead to a molecule of class (1) or (2). Further details on symmetry detection are provided in Supplementary Note 2. As always, incorporation of any such domain-specific heuristics like symmetry is thereby a double-edged sword, possibly generating more meaningful search spaces as much as introducing a limiting bias. AML is particularly appealing in this respect. Any such rules can readily be added or dropped without incurring excessive computational costs as in exhaustive screenings of predefined search spaces.

**Charge-conduction based fitness**. In the spanned search space, we assess the suitability of candidate molecules for OSC applications by two descriptors known to probe two important and complementary aspects related to the conduction of charge. One concerns the efficient injection of charge from a contacting electrode into the OSC material. The other assesses the required high charge mobility inside the OSC bulk. For predominantly *p*-type OSC materials[23] a detrimentally high barrier for a corresponding hole injection from a standard gold electrode is readily probed by a level-alignment descriptor $\epsilon_{\mathrm{align}} = |\epsilon_{\mathrm{HOMO}} - \Phi_{\mathrm{Au}}|$,[42] which evaluates the energetic mismatch between the Au work function $\Phi_{\mathrm{Au}} = -5.1\ \mathrm{eV}$[43] and the energetic position of the highest occupied molecular orbital (HOMO) $\epsilon_{\mathrm{HOMO}}$ as a common approximation of the material's ionization potential.[44,45] Adapting this descriptor to other electrode materials or to *n*-type OSC materials (then involving the energetic position of the lowest unoccupied molecular orbital, LUMO) is straightforward. As an equally established descriptor for the bulk charge mobility we employ the

intra-molecular (hole) reorganization energy $\lambda_h$, which measures the cost of accommodating a new charge state after the carrier has moved to the next molecular site.[46,47] As molecular properties, both $\epsilon_{HOMO}$ and $\lambda_h$ can be determined by efficient first-principles calculations as detailed in Supplementary Note 2, where the density-functional theory (DFT) B3LYP[48–50] level of theory constitutes a well established accuracy standard[27,31,39,40,51], matching experimental data[44,52]. We emphasize though that using the lowest-energy gas-phase conformer for the descriptor calculation disregards packing-effects in the molecular crystal[53–55] and we further discuss the influence of conformers on descriptor values in Supplementary Note 3.

To evaluate molecular fitness and prioritize candidates during AML discovery, both objectives are combined in a scalarized fitness function

$$F = -\left\| \begin{pmatrix} \lambda_h \\ \epsilon_{align} \end{pmatrix} \cdot \mathbf{w} \right\|_2 , \quad (1)$$

which an ideal candidate molecule will maximize.[56] Here, the weight vector $\mathbf{w} = (1.0, 0.7)^\top$ accommodates the generally different absolute scales of the two descriptors, with the value of 0.7 chosen to yield an essentially Ohmic alignment with the electrode of $|\epsilon_{align}| < 0.3$ eV if $\lambda_h$ falls into the range of commonly known OSCs. We note, though, that the exact choice of weights is rather unimportant for the performance of the AML search, as it only linearly biases $F$ towards either of the descriptors, as further detailed below. With the currently chosen weight and at the DFT-B3LYP level of theory, pentacene and rubrene – materials that have been contacted by gold electrodes before[57,58] – will feature $F$ values of −0.16 and −0.2, respectively. A threshold $F \geq -0.2$ will therefore later on be used to measure discovery success of the AML.

**AML: design and search strategy**. By successively querying the explicit first-principles calculation of the descriptors for identified candidate molecules, the AML algorithm establishes an ever improving surrogate model of the fitness function $F$ over the search space. Out of a manifold of in principle possible surrogate models, we found GPR to already achieve outstanding performance at very moderate amounts of data. In brief, the employed model uses circular Morgan fingerprints[59] to compare the structural similarity of not yet explicitly calculated molecules with the hitherto acquired ones. Specifically, counts of substructures that can be extracted by moving up to two bonds away from each central atom are generated. The similarity between two molecules is then measured with a substructure count kernel. A full account of the GPR learning through log-marginal likelihood maximization is provided in Supplementary Note 2. A central advantage of GPR for the AML context is that it not only provides a prediction for the targeted fitness function $F$, but also the corresponding predictive uncertainty $\sigma$ from the Gaussian variance. Balancing between exploitation and exploration, the AML algorithm can thus query new candidate molecules either because they are highly promising in terms of a maximum predicted fitness $F$ or because they exhibit a high uncertainty $\sigma$ such that their explicit calculation will maximally improve the surrogate model. Practically, molecules are thereby chosen according to an upper confidence bound acquisition function

$$F_{acq} = F + \kappa\sigma. \quad (2)$$

This represents a simple, well-tested strategy in Bayesian optimization[60–62] or active-search[63,64] with GPRs, which contains only one hyperparameter $\kappa$ to balance exploration and exploitation.

Multiple possibilities arise how to actually execute the iterative AML process. After initializing the surrogate model by training on a defined number $N_{initial}$ of molecules, central questions concern the acquisition of new data before the surrogate model is retrained. Compatible with super-computing resources that encourage a parallel first-principles evaluation of the descriptors for multiple molecules, we opt for a batch-based learning where $N_{batch}$ molecules with maximum $F_{acq}$ are queried and the model is then retrained on the basis of the accumulated new descriptor data. Future improvements could include an additional enforcement of diversity in the prioritized batch.[18,21,65,66] In an in principle infinite chemical space, another central AML design choice regards the extent over which new molecules are practically assessed with the established, conceptually global surrogate model. Aiming for high-performance OSC molecules of tractable size and complexity, we here opt for a single tree expansion that limits the candidates to those in the vicinity of already sampled ones[67].

In a most straightforward realization and if all molecules for which first-principles descriptors have already been computed define the current population at step $n$ of the AML search, then the $N_{batch}$ molecules for the next step $n+1$ are identified in the search space formed by all molecules that can be generated by one-time application of any of the morphing operations to every molecule in the current population. While this nicely exploits the evolutionary pressure contained in the current population of size $N_{pop} = N_{initial} + n \times N_{batch}$, the search space for step $n+1$ could also be systematically increased by exhaustive multiple-time application of the morphing operations. As illustrated below by comparing a corresponding search depth of one- or two-time application, this may help to overcome local funnels and navigate more efficiently through chemical space. On the other hand and regardless of the actual search depth $d_{search}$, the continuously growing population size will at later learning steps $n$ inevitably lead to a combinatorial explosion of new candidates for any such exhaustive enumeration. Eventually, this requires to decrease the resolution in the ever increasing search space. Note that precisely this combinatorial explosion also precludes popular supervised machine learning approaches that exhaustively learn molecular properties in a closed chemical space, possibly followed by some form of data mining[3].

A decreasing resolution in the AML search space can for instance be achieved by imposing additional heuristic selection criteria, e.g., selectively suppressing certain morphing operations for increasing search depths, or other more sophisticated tree-search policies[68] also employed in reinforcement learning[35,69]. Here, we realize deeper partial expansions of the search tree up to a search depth $d_{search}$ by applying the molecular morphing operations only to a fixed number of $N_{deep}$ molecules selected first from the current population and then subsequently from those molecules that were created by the previous morphing operations. By each time selecting the $N_{deep}$ molecules through fitness-rank based roulette-wheel selection, i.e., by assigning higher selection probabilities to molecules with high $F_{acq}$ values, the search tree is thus preferentially expanded into regions of the OSC space that the surrogate model anticipates to be rewarding (either in terms of exploitation or exploration).

**Hyperparameter optimization**. The thus defined AML approach contains a number of hyperparameters that may critically affect its performance. Most notably, these are $\kappa$ that balances exploration and exploitation in the acquisition function, $N_{batch}$ the size of the prioritized batch in each learning step, as well as $d_{search}$ the depth of the search space in terms of the number of applied morphing operations. The decreased resolution strategy additionally requires the specification of the fixed subset size of $N_{deep}$ molecules to which morphing operations are applied. Less

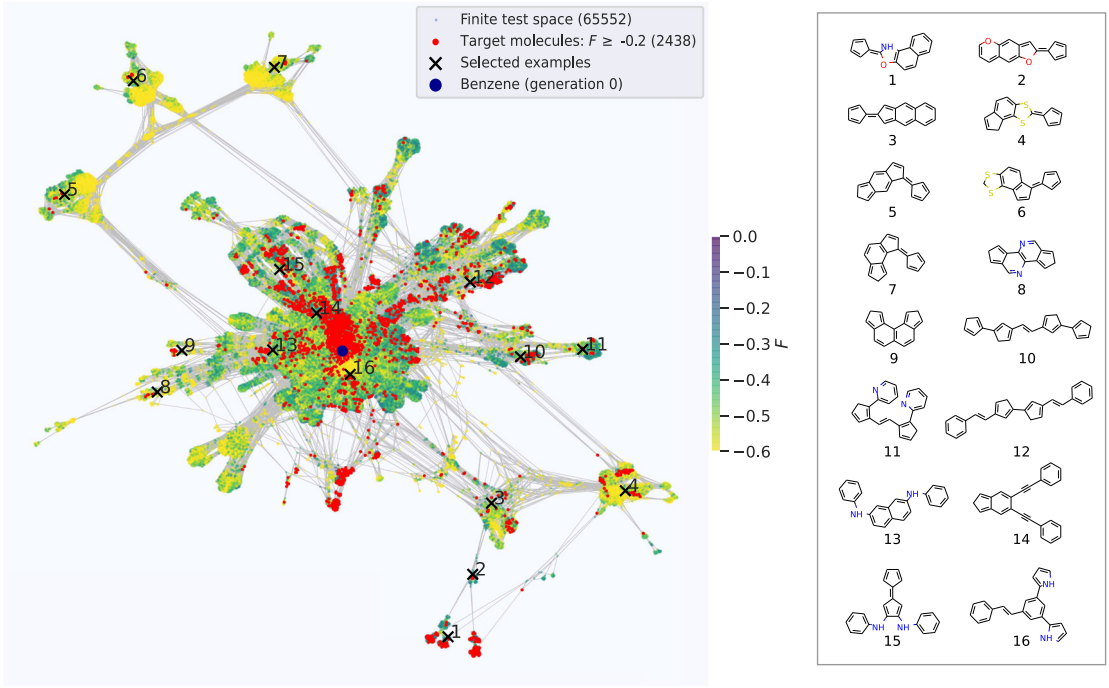



**Fig. 2 Finite OSC test space.** Left panel: Chemical space network (CSN) representation of the finite OSC test space of 65.552 unique molecules generated by exhaustive application of all morphing operations up to 14 times. Each molecule is surrounded by morphing-related analogs (see text). Benzene as the smallest base molecule is colored in blue. All other molecular nodes are colored according to their fitness function $F$ as calculated at the semi-empirical density-functional tight-binding level. 2438 red nodes form the target discovery group of top-performing molecules with high fitness $F \geq -0.2$. Right panel: Example molecules from the top-performing group, chosen randomly from different areas of the CSN to illustrate the structural diversity contained in the test space.

decisive is the initial number of molecules $N_{\text{initial}}$ used for the first training of the surrogate model, which defines only an insignifiant part of the total executed first-principles calculations and which should only be large enough to somehow kick-start the AML process. Here, we suitably set $N_{\text{initial}}$ to the 179 unique molecules that result in the first two generations when applying all morphing operations up to two times starting from the simplest building block benzene, cf. Fig. 1.

In order to explore the effect of the other hyperparameters and optimize them for first-principles OSC discovery, we consider the finite subspace formed of all molecules up to a maximum size of 4 rings, 4 heteroatoms and 2 linkers that are generated by exhaustive application of all morphing operations up to 14 times, see Supplementary Note 2. With 65.552 unique molecules this subspace is already representative for the design problem and contains many and diversely structured high-performing molecules as illustrated in Fig. 2. At the same time, the still tractable size of the finite test space allows for the exhaustive calculation of all molecular descriptors with van der Waals (vdW) corrected density functional tight-binding (DFTB).[70] While this semi-empirical level of theory is not fully quantitative, it provides a sufficiently realistic account of the descriptor landscape for the intended method testing as analyzed in detail in Supplementary Fig. 4. Further details on molecular test space generation and descriptor calculation are provided in the Supplementary Note 2.

The finite test space contains a total of 2438 top-performing molecules with a high fitness $F \geq -0.2$. As a quantitative benchmark, we thus measure the discovery success $S(N)$ as the fraction of these molecules that are identified after the descriptors of $N$ molecules have been queried. With 179 queries used for the initialization, see above, the final measure $S(5179)$ thus evaluates the discovery success after $n = 50$ learning steps when using $N_{\text{batch}} = 100$. Supplementary Fig. 6 compiles the corresponding success curves $S(N)$, when systematically combining $N_{\text{batch}} = 50$, 100, or 200 with $\kappa$ values in half-integer steps between 0 and 5, as well as for a search depth of one- or two-time exhaustive application of all morphing operations. Fortunately, we find the AML search to be highly robust with respect to the choice of $N_{\text{batch}}$ and $\kappa$. Only a small variation of $0.71 < S(N = 5179) < 0.80$ is obtained over all tested combinations for a search depth of one, meaning that 70–80% of the top-performing molecules are consistently found after descriptors for less than 8% of the entire test space have actually been computed. For a search depth of two, this success rate becomes slightly higher, reaching up to 85% as compiled in Supplementary Fig. 7. Generally, larger batch sizes seem to implicitly increase the explorative behavior, such that an almost indistinguishably optimum performance is obtained for larger $N_{\text{batch}}$ in combination with successively smaller exploration weights $\kappa$ in the acquisition function, cf. Eq. (2). For too small $\kappa$, the success curves become stepped though, indicating that temporarily the mainly exploitative algorithm then only meanders through identified sub-pockets of the test space. Too large $\kappa$, on the other hand, diminish the initial success of a then too explorative algorithm in the first learning steps. Overall, an intermediate value pair $(N_{\text{batch}}, \kappa) = (100, 2.5)$ thus provides a robust setting and is henceforth employed in all AML runs. For these values of $(N_{\text{batch}}, \kappa)$, we also performed a sensitivity analysis with regard to the employed weight vector $\mathbf{w}$ in Eq. (1) and the bond radius in the Morgan fingerprints used to assess molecular similarity. The results are summarized in Supplementary Figs. 8 and 9, respectively, and again demonstrate a high robustness with respect to these parameters.

The higher success rate for $d_{\text{search}} = 2$ indicates that it is generally advantageous to further expand the search space away from the known topologies of the current population. Assessing the dependence of the decreased resolution AML algorithm on its

two additional hyperparameters, Supplementary Table 1 summarizes the corresponding discovery successes when systematically combining a varying subset size $N_{deep} = 100, 250, 500$ and 1000 with search depths $d_{search} = 1, 2, 3, 4, 5$ and 10. Again, we find the algorithm to be quite robust, with higher $d_{search}$ compensating smaller $N_{deep}$. Within the finite test space, many combinations thus saturate at success rates around 82–83%. This is essentially as good as the best performance of the previous exhaustive enumerations, but comes at the advantage of a controlled growth of the search space at later learning steps. For the first-principles AML discovery in the virtually unlimited OSC space below we correspondingly employ this decreased resolution search strategy with a top-performing hyperparameter combination $(d_{search}, N_{deep}) = (3, 500)$.

**Visualizing AML at work**. The finite test space can also be viewed as a chemical space network (CSN), in which the morphing operations establish a total of 315.451 directed connections between the constituting molecules. This allows us to visualize the space in form of a 2D graph structure, in which the molecules are mutually repelling nodes, while morphing relationships between them lead to attractive edges[71], see Supplementary Note 1 for details. In such a representation each molecule is thus spatially surrounded by morphing-related analogs. Figure 2 shows the resulting graph, in which the individual nodes are colored according to their DFTB calculated fitness. As expected, the target group for discovery in form of the 2438 top-performing molecules is widely scattered over disjoint parts of chemical space, with ensembles of related molecules often clustered in sub-pockets.

Apart from providing a bird's eye view of the design problem, the CSN representation also affords a direct visual access to the AML process. Plotting the evolving population $N$ over subsequent learning steps $n$ reveals how much a chosen AML strategy is able to focus its exploration onto the interesting regions of chemical space and how efficiently it prioritizes OSC molecules with desired properties. Figure 3 illustrates this for the determined optimum hyperparameters and contrasts the learning for exhaustive searches with depths of one or two, with the decreased resolution strategy where the searches partially expand subsets of

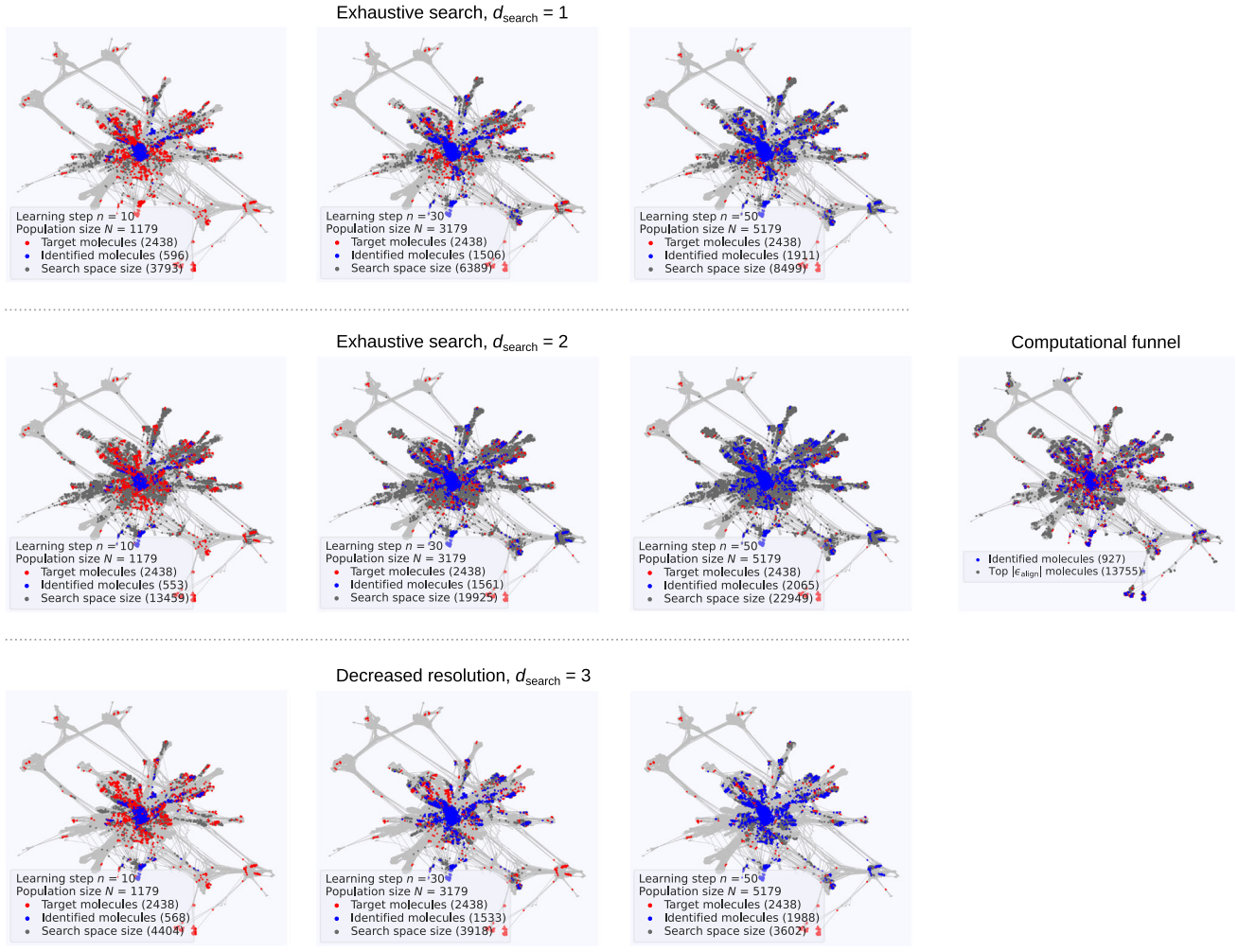



**Fig. 3 AML exploration of the finite test space.** The same CSN representation of the OSC test space as in Fig. 2 is shown in gray. Superimposed are the target group of 2438 top-performing molecules in red. Each panel shows the discovery success after $n$ learning steps with the color of all identified top-performing molecules changed to blue and the search space for the next learning step $n + 1$ colored in dark gray. Left upper panels: Steps $n = 10$, 30, 50 for an exhaustive search with search depth of one. Left middle panels: Steps $n = 10$, 30, 50 for an exhaustive search with search depth of two. Left lower panels: Steps $n = 10$, 30, 50 for a decreased resolution search ($N_{deep} = 500$) with search depth of three (see text). Supplementary Movies 1–3 provide the detailed, full trajectory of all three AML discovery runs over learning steps 1–50. Right centered panel: Discovery success of a conventional computational funnel after computing an equal number of descriptors (5179) as after 50 learning steps, and anticipating that knowledge of 13.755 molecules with optimum $|\epsilon_{align}| < 0.3$ eV is present (see text).

$N_{deep} = 500$ molecules at search depth three. For the exhaustive search with $d_{search} = 1$, the discovery is centered to more morphing-related top-performing molecules all more or less located in the core region of the CSN. In contrast, for the deeper exhaustive search, the algorithm also successfully identifies top-performing molecules in the periphery of the network that are topologically quite disconnected from the initial population. The downside is a rapidly increasing size of the search space that in the present case is only bounded by the finiteness of the considered test space. This is largely mitigated by the decreased resolution search, which nevertheless equally successfully identifies top-performing molecules at the CSN periphery.

To put this performance of the AML searches into perspective, we also contrast them in Fig. 3 with the result of a conventional computational funnel. For the latter we pretend that the calculation of $\epsilon_{HOMO}$ has a negligible computational cost and the value of this descriptor is known for every molecule in the test space. This allows to identify a subset of 13.755 promising molecules for which $|\epsilon_{align}| < 0.3$ eV and which contains all previously considered 2438 top-performing molecules. The computational funnel approach would then focus the explicit calculation of the more demanding $\lambda_h$ descriptor to molecules in this subset. To enable a direct comparison with the preceding AML assessment, a random selection of 5179 molecules out of this subset would then lead to a success rate of $S(5179) \approx 0.4$. Even in this finite test space, where the AML algorithm can not even unfold its real strength, less than half of the top-performing molecules are thus found by this prevalent computational screening strategy after spending the same amount of CPU time (assuming that the exhaustive calculation of 65.552 $\epsilon_{HOMO}$ descriptors for the entire test space would constitute an insignificant computational effort).

**First-principles AML discovery in a virtually unlimited OSC chemical space**. Based on the gathered methodological understanding and optimized algorithmic settings ($N_{batch} = 100$, $\kappa = 2.5$, $d_{search} = 3$, $N_{deep} = 500$) we now proceed to first-principles AML discovery at the vdW-corrected DFT-B3LYP level of theory. This is a truly challenging endeavor, considering the vastness of the OSC design space. While the space of molecules that can be generated through the morphing operations is in principle unbounded, we here restrict it to the realm of "small molecules" containing a maximum of 100 atoms (including H atoms). This realm appears as a first, more practical target for synthesis and crystallization, also considering that essentially all known top-performing OSC molecules to date fall into this size range. Estimated to surpass a size of $10^{30}$ molecules, see Supplementary Note 2, the corresponding chemical space is nevertheless virtually unlimited for all practical purposes and would defy any conventional exhaustive computational screening. While an iterative search as with AML is thus the only tractable means to explore this space at predictive quality, an additional technical aspect emerges that did not yet play a role in the analysis of the finite test space at the semi-empirical level before. It concerns the typically massively parallel processing on the required high-performance computing (HPC) infrastructure. As a result of queuing or downtimes, as well as convergence behavior of the first-principles calculations, the results for the $N_{batch}$ descriptor calculations can become available at quite different times (or in rare cases of failed convergence or system instabilities may not become available at all). A practical way to avoid long waiting times before the last calculations are ready is to initially select a larger batch size for descriptor calculation and then continue with the forthcoming learning steps whenever the desired number of $N_{batch}$ molecules has been processed (successfully or unsuccessfully). We found this strategy to afford an efficient and continuous HPC workflow, here initially submitting the 200 molecules with highest $F_{acq}$ values for descriptor calculations. These are continuously processed on the HPC system by 40–100 parallel worker processes, to reach the targeted batch size $N_{batch} = 100$, while for a retraining of the surrogate model only successfully processed cases are included. In this respect, the above determined robustness of the AML performance with regard to the exact batch size also constitutes an important asset for such HPC operation.

Figure 4 summarizes the results of the AML discovery run over its first 15 learning steps. Gratifyingly, the algorithm quickly stabilizes into a highly efficient mode of operation while simultaneously meandering deep into unknown chemical space. Already after five learning steps even the median fitness of the entire prioritized batch exceeds the threshold value $F \geq -0.2$ for the first time, reflecting top-performing molecules. However, as clearly seen from the violin plots of the $F$ distribution over the batches in Fig. 4b, this high efficiency does not simply result from the algorithm just exploiting its established knowledge. Even at later learning steps, the algorithm steadily queries quite unfavorable molecules with a fitness worse than $F < -0.3$. While such exploratory queries can either be based on high model uncertainty or induced by model prediction errors, they serve to continuously improve the surrogate model also outside the already considered search space. As a result, at each later learning step, the algorithm keeps on identifying top-performing molecules at a stable, high rate.

After 15 learning steps and a corresponding calculation of first-principles descriptors for 1680 molecules (and only 35 unsuccessfully terminated calculations), a total of 900 molecules with molecular fitness $F \geq -0.2$ have been found. A relative success rate of 54%, i.e., essentially every second first-principles calculation yields a promising molecule and this without any a priori knowledge of the vast OSC space. A second AML discovery run described in Supplementary Note 4 confirms the robustness of this high performance. Notably, due to the random nature in our search strategy, significantly different, but equally favorable molecules are identified in this run. This performance becomes even more impressive from the viewpoint that these molecules are true discoveries, as essentially none of them are contained in existing focused libraries assembled in previous screening studies[3,31–34]. With typically $\sim 10^5 - 10^6$ entries, these data sets reflect the wealth of our existing knowledge and synthesis efforts, but simply do not even scratch the surface of the true OSC design possibilities. To this end, the negligible overlap with the top-performing molecules identified in these previous studies also has to do with molecular size. Within the first learning steps, the average size in the prioritized batch quickly rises to around 90 atoms, which is at the edge of the limit currently imposed on our search and in a size regime that could barely be addressed by the previous exhaustive enumeration studies. At the same time, even archetypical and acclaimed molecular OSC materials like DNTT ($C_{22}H_{12}S$) or rubrene ($C_{42}H_{28}$) approach this size regime, with many other experimentally tested candidates falling right into it[23]. The preferred prioritization of such larger molecules is thereby to some extent likely simply a result of the combinatorially exploding phase space. On the other hand, another physical factor could be that the AML algorithm learns and exploits the tendency of $\lambda_h$ to decrease with increasing molecular size[3] as a consequence of a larger hole delocalization (which even at the hybrid DFT-B3LYP level of theory may be slightly overestimated[72]). The inclusion of molecular coupling-sensitive descriptors into the fitness function is therefore certainly a promising topic for future studies.

The discovered molecules exhibit a diverse set of structures, incorporating distinct core fragments and the full set of allowed

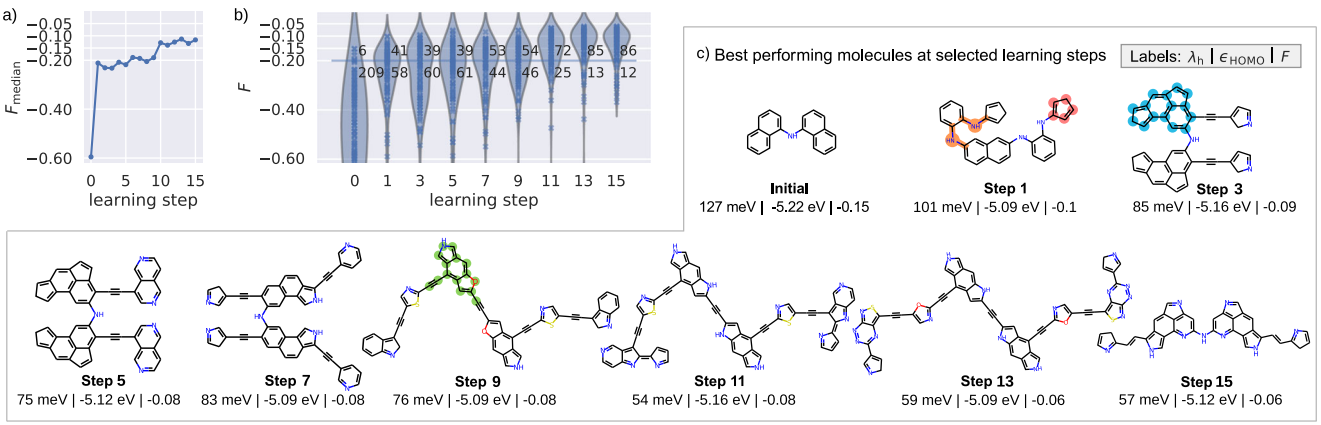



**Fig. 4 First-principles AML discovery in a virtually unlimited space. a** Median values of molecular fitness $F$ over the prioritized $N_{batch}$ molecules at the different learning steps (step 0 shows the median of the initial population $N_{initial}$). **b** Corresponding violin plot showing the (kernel-density estimated) distribution of molecular fitness $F$ over the batch. These smooth kernel-density estimated distributions can slightly extend beyond the true range of $F$ values as indicated by the explicit values marked by blue crosses. The number of queries leading to favorable and unfavorable molecules is indicated next to each violin. Due to descriptor calculation failures (see text) these numbers do not always add up to $N_{batch} = 100$. **c** Examples of top-performing molecules identified at various learning steps (see text for an explanation of the different color-highlighted geometric motifs). An extended list of the 4 top-performing molecules of each learning step is shown in Supplementary Fig. 10.

heteroatoms and linkers. Figure 4c illustrates this with the best-performing molecules identified at selected learning steps, and an extended list being compiled in Supplementary Fig. 11. This diversity indicates that the AML algorithm successfully explored topologically widely differing areas of the OSC space and did not get stuck in one or a few subpockets. Nevertheless, some commonalities can be spotted, like the recurrent presence of phenylamine linker motifs (marked in orange in the best-performing molecule of learning step 1 in Fig. 4c). Similarly, more complex ring systems emerged at later learning stages (marked in blue and green in the most favorable molecule of step 3 and 9, respectively) and are from thereon quite pronounced among well-performing molecules. While a diverse molecular space is searched, the AML discovery thus automatically identifies and prioritizes privileged design motifs. After harvesting a larger number of molecules in further learning steps, an exciting prospect for future studies is therefore to mine the accumulating data set and systematically extract this implicit knowledge for rational design. To this end, the trained surrogate model can also be used to quickly assess the suitability of such manually constructed molecules or of deliberate modifications of the here identified ones. The latter could be particularly appealing in view of long-term device-stability or synthetic accessibility. We note that certainly not all identified molecules are suitable in this regard. For instance, the 5-membered unsaturated rings of the displayed compound of learning step 1 (marked in red) in Fig. 4c could be problematic as they might undergo Diels-Alder type reactions, and we attribute the appearance of such ring motives as the algorithm's intent to provide intermediates on the way to the later explored, more stable 5-membered heterocycles. Nonetheless, multiple of the favorable molecules are symmetric and composed of standard building blocks that should be easily accessible through short and reliable synthesis routes, with the surrogate model furthermore available to gauge the effect of stabilizing modifications.

## Discussion

In our view, active machine learning based on first-principles descriptors constitutes a most promising route to prior-to-synthesis virtual discovery. Its iterative refinement allows to most efficiently focus the data-generating calculations and meaningfully explore the vastness of chemical spaces at predictive quality and without a priori specifications, enumeration or reliance on empirical descriptors with limited validity range. In this work we have established such an AML discovery approach for molecular OSC materials through versatile molecular morphing operations and based on charge injection and conduction querying descriptors. Fortunately and with a view on explainable ML models, our systematic assessment within a finite test space suggests the approach to be quite robust with respect to the algorithmic hyperparameters. Most promising to further increase its already high efficiency and prevent an over-exploitation of particular structural motifs, is likely to additionally enforce structural diversity among the $N_{batch}$ molecules selected at each learning step, instead of the present purely fitness-ranked roulette-wheel selection.

Central to assess this performance and enable an unbiased and systematic comparability of different AML approaches will be the establishment of well-designed, balanced and freely available benchmark platforms for unlimited search spaces. As clear from the present work, already within the here pursued single-tree expansion there are multiple design strategies and concomitant algorithmic parameters. While we have explored these in a truncated test space, AML only unfolds its full potential in the exploration of unlimited spaces. Representative and standardized benchmark platforms as already available for drug-design tasks[13] will therefore be pivotal to truly compare various learning concepts that work without a priori enumeration or pre-definition of the search problem.

Further challenges and advancements in the physico-chemical domain comprise the adaption and extension of the molecular morphing operations to tailor the OSC search space. The present set derived from literature domain knowledge spans a design space geared towards flexible, π-conjugated molecules. Ultimately, a generic, but chemically-valid creation of morphing operations could drive discovery of many novel structural motives. Heavier requirements on the surrogate GPR-model in such cases could then be tackled with improved covariance functions for 2D molecular graphs[73] or conformer-specific 3D coordinates[74], while alleviating the limited scaling by sparse approximations[75], or application of alternative models[76–79].

Another major area for development concerns the first-principles descriptors entering the employed multi-objective fitness function. Devising such suitable descriptors has evolved into

an important research area of its own[80–83], independent of the present AML and OSC context. With the presently employed level-alignment descriptor $\epsilon_{\mathrm{align}}$ and the hole reorganization energy $\lambda_{\mathrm{h}}$ our search readily identified a diverse range of hitherto unknown molecular candidates. Just as in conventional computational screening, there are numerous possibilities to refine the underlying candidate evaluation through additional (or alternative) descriptors. In the exemplified OSC context, obvious avenues could be to explicitly consider synthetic accessibility[84], electronic coupling and charge-transport networks in the molecular solid[46,51,85,86] or electron-phonon coupling[87]. In view of the high data efficiency of the AML approach, one may also drop the present focus on computationally least-demanding descriptors, originally dictated by the excessive queries in conventional exhaustive screening work. More elaborate descriptors like structural interfacing with electrode materials[88] could therefore routinely (or at least occasionally) be requested. Eventually, one could even think of incorporating experimental feedback from self-driving laboratories[89]. The prospects are thus as manifold as exciting. Regardless of the specific road chosen, it is conceptually clear that autonomously operating workflows like the present AML approach offer an unparalleled means to accelerate the discovery and design of viable future materials like the high-mobility organic semiconductors featured in this work.

## Data availability

The source data necessary to reproduce the main figures of the manuscript is provided in the supplementary materials of this article. Source data are provided with this paper.

## Code availability

The code used to run AML discovery is available at https://doi.org/10.5281/zenodo.4554331

Received: 29 September 2020; Accepted: 15 March 2021;
Published online: 23 April 2021

## References

1. Polishchuk, P. G., Madzhidov, T. I. & Varnek, A. Estimation of the size of drug-like chemical space based on gdb-17 data. *J. Comput. Aided Mol. Des.* **27**, 675–679 (2013).
2. Jorgensen, W. L. The many roles of computation in drug discovery. *Science* **303**, 1813–1818 (2004).
3. Schober, C., Reuter, K. & Oberhofer, H. Virtual screening for high carrier mobility in organic semiconductors. *J. Phys. Chem. Lett.* **7**, 3973–3977 (2016).
4. Pulido, A. et al. Functional materials discovery using energy–structure–function maps. *Nature* **543**, 657–664 (2017).
5. Pyzer-Knapp, E. O., Suh, C., Gómez-Bombarelli, R., Aguilera-Iparraguirre, J. & Aspuru-Guzik, A. What is high-throughput virtual screening? a perspective from organic materials discovery. *Annu. Rev. Mater. Res.* **45**, 195–216 (2015).
6. Gaspar, H. A., Baskin, I. I., Marcou, G., Horvath, D. & Varnek, A. Chemical data visualization and analysis with incremental generative topographic mapping: big data challenge. *J. Chem. Inf. Model.* **55**, 84–94 (2015).
7. Reymond, J.-L., van Deursen, R., Blum, L. C. & Ruddigkeit, L. Chemical space as a source for new drugs. *Med. Chem. Commun.* **1**, 30–38 (2010).
8. Devi, R. V., Sathya, S. S. & Coumar, M. S. Evolutionary algorithms for de novo drug design – a survey. *Appl. Soft Comput.* **27**, 543–552 (2015).
9. Le, T. C. & Winkler, D. A. Discovery and optimization of materials using evolutionary approaches. *Chem. Rev.* **116**, 6107–6132 (2016).
10. Sanchez-Lengeling, B. & Aspuru-Guzik, A. Inverse molecular design using machine learning: generative models for matter engineering. *Science* **361**, 360–365 (2018).
11. Elton, D. C., Boukouvalas, Z., Fuge, M. D. & Chung, P. W. Deep learning for molecular design – a review of the state of the art. *Mol. Syst. Des. Eng.* **4**, 828–849 (2019).
12. Reker, D. & Schneider, G. Active-learning strategies in computer-assisted drug discovery. *Drug Discov. Today* **20**, 458–465 (2015).
13. Brown, N., Fiscato, M., Segler, M. H. & Vaucher, A. C. Guacamol: benchmarking models for de novo molecular design. *J. Chem. Inf. Model.* **59**, 1096–1108 (2019).
14. Schneider, G. *De novo Molecular Design* (Wiley, 2013) https://books.google.de/books?id=ZxlrmwEACAAJ.
15. Jensen, K. F., Coley, C. W. & Eyke, N. S. Autonomous discovery in the chemical sciences part i: progress. *Angew. Chem. Int. Ed.* **59**, 22858–22893 (2019).
16. Settles, B. *Active learning literature survey* (University of Wisconsin, Madison, 2010).
17. Lookman, T., Balachandran, P. V., Xue, D. & Yuan, R. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *Npj Comput. Mater.* **5**, 21 (2019).
18. Smith, J. S., Nebgen, B., Lubbers, N., Isayev, O. & Roitberg, A. E. Less is more: sampling chemical space with active learning. *J. Chem. Phys.* **148**, 241733 (2018).
19. Häse, F., Roch, L. M., Kreisbeck, C. & Aspuru-Guzik, A. Phoenics: a bayesian optimizer for chemistry. *ACS Cent. Sci.* **4**, 1134–1145 (2018).
20. Vandermause, J. et al. On-the-fly active learning of interpretable bayesian force fields for atomistic rare events. *Npj Comput. Mater.* **6**, 20 (2020).
21. Janet, J. P., Ramesh, S., Duan, C. & Kulik, H. J. Accurate multiobjective design in a space of millions of transition metal complexes with neural-network-driven efficient global optimization. *ACS Cent. Sci.* **6**, 513–524 (2020).
22. Bisbo, M. K. & Hammer, B. Efficient global structure optimization with a machine-learned surrogate model. *Phys. Rev. Lett.* **124**, 086102 (2020).
23. Wang, C., Dong, H., Hu, W., Liu, Y. & Zhu, D. Semiconducting $\pi$-conjugated systems in field-effect transistors: a material odyssey of organic electronics. *Chem. Rev.* **112**, 2208–2267 (2012).
24. Lin, Y., Li, Y. & Zhan, X. Small molecule semiconductors for high-efficiency organic photovoltaics. *Chem. Soc. Rev.* **41**, 4245–4272 (2012).
25. Xu, R.-P., Li, Y.-Q. & Tang, J.-X. Recent advances in flexible organic light-emitting diodes. *J. Mater. Chem. C* **4**, 9116–9142 (2016).
26. Geng, H. et al. Theoretical study of substitution effects on molecular reorganization energy in organic semiconductors. *J. Chem. Phys.* **135**, 104703 (2011).
27. Uejima, M., Sato, T., Tanaka, K. & Kaji, H. Vibronic coupling density analysis for the chain-length dependence of reorganization energies in oligofluorenes: a comparative study with oligothiophenes. *Phys. Chem. Chem. Phys.* **15**, 14006–14016 (2013).
28. Wilbraham, L., Smajli, D., Heath-Apostolopoulos, I. & Zwijnenburg, M. A. Mapping the optoelectronic property space of small aromatic molecules. *Commun. Chem.* **3**, 14 (2020).
29. Gryn'ova, G., Lin, K.-H. & Corminboeuf, C. Read between the molecules: computational insights into organic semiconductors. *J. Am. Chem. Soc.* **140**, 16370–16386 (2018).
30. Saeki, A. & Kranthiraja, K. A high throughput molecular screening for organic electronics via machine learning: present status and perspective. *Jpn. J. Appl. Phys.* **59**, SD0801 (2019).
31. Matsuzawa, N. N. et al. Massive theoretical screen of hole conducting organic materials in the heteroacene family by using a cloud-computing environment. *J. Phys. Chem. A* **124**, 1981–1992 (2020).
32. Nematiaram, T., Padula, D., Landi, A. and Troisi, A. On the largest possible mobility of molecular semiconductors and how to achieve it. *Adv. Funct. Mater.* **30**, 2001906 (2020).
33. Atahan-Evrenk, S. & Atalay, F. B. Prediction of intramolecular reorganization energy using machine learning. *J. Phys. Chem. A* **123**, 7855–7863 (2019).
34. Cheng, C. Y., Campbell, J. E. & Day, G. M. Evolutionary chemical space exploration for functional materials: computational organic semiconductor discovery. *Chem. Sci.* **11**, 4922–4933 (2020).
35. Segler, M. H. S., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604–610 (2018).
36. Segler, M. H. S. & Waller, M. P. Neural-symbolic machine learning for retrosynthesis and reaction prediction. *Chem. Eur. J.* **23**, 5966–5971 (2017).
37. Yao, Z.-F., Wang, J.-Y. & Pei, J. Control of $\pi$–$\pi$ stacking via crystal engineering in organic conjugated small molecule crystals. *Cryst. Growth Des.* **18**, 7–15 (2018).
38. Kunkel, C., Schober, C., Margraf, J. T., Reuter, K. & Oberhofer, H. Finding the right bricks for molecular legos: a data mining approach to organic semiconductor design. *Chem. Mater.* **31**, 969–978 (2019a).
39. Stehr, V., Pfister, J., Fink, R. F., Engels, B. & Deibel, C. First-principles calculations of anisotropic charge-carrier mobilities in organic semiconductor crystals. *Phys. Rev. B* **83**, 155208 (2011).
40. Li, P., Cui, Y., Song, C. & Zhang, H. Electronic and charge transport properties of dimers of dithienothiophenes: effect of structural symmetry and linking mode. *RSC Adv.* **5**, 50212–50222 (2015).
41. Ren, L. et al. Critical role of molecular symmetry for charge transport properties: a paradigm learned from quinoidal bithieno[3,4-b]thiophenes. *Chem. Mater.* **29**, 4999–5008 (2017).

42. Ishii, H., Sugiyama, K., Ito, E. & Seki, K. Energy level alignment and interfacial electronic structures at organic/metal and organic/organic interfaces. *Adv. Mater.* **11**, 605–625 (1999).
43. Michaelson, H. B. The work function of the elements and its periodicity. *J. Appl. Phys.* **48**, 4729–4733 (1977).
44. Schwenn, P., Burn, P. & Powell, B. Calculation of solid state molecular ionisation energies and electron affinities for organic semiconductors. *Org Electron.* **12**, 394 – 403 (2011).
45. Bhandari, S., Cheung, M. S., Geva, E., Kronik, L. & Dunietz, B. D. Fundamental gaps of condensed-phase organic semiconductors from single-molecule calculations using polarization-consistent optimally tuned screened range-separated hybrid functionals. *J. Chem. Theory Comput.* **14**, 6287–6294 (2018).
46. Oberhofer, H., Reuter, K. & Blumberger, J. Charge transport in molecular materials: an assessment of computational methods. *Chem. Rev.* **117**, 10319–10357 (2017).
47. Nelsen, S. F., Blackstock, S. C. & Kim, Y. Estimation of inner shell marcus terms for amino nitrogen compounds by molecular orbital calculations. *J. Am. Chem. Soc.* **109**, 677–682 (1987).
48. Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A* **38**, 3098–3100 (1988).
49. Lee, C., Yang, W. & Parr, R. G. Development of the colle-salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **37**, 785–789 (1988).
50. Stephens, P. J., Devlin, F. J., Chabalowski, C. F. & Frisch, M. J. Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields. *J. Phys. Chem.* **98**, 11623–11627 (1994).
51. Moral, M., Garzón-Ruiz, A., Castro, M., Canales-Vázquez, J. & Sancho-García, J. C. Virtual design in organic electronics: screening of a large set of 1,4-bis (phenylethynyl)benzene derivatives as molecular semiconductors. *J. Phys. Chem. C* **121**, 28249–28261 (2017).
52. Kera, S. et al. Experimental reorganization energies of pentacene and perfluoropentacene: effects of perfluorination. *J. Phys. Chem. C* **117**, 22428–22437 (2013).
53. Stuke, A. et al. Atomic structures and orbital energies of 61,489 crystal-forming organic molecules. *Sci. Data* **7**, 58 (2020).
54. Mitzel, N. W. & Rankin, D. W. H. Saracen – molecular structures from theory and experiment: the best of both worlds. *Dalton Trans.* 3650–3662 (2003).
55. Blomeyer, S. et al. Intramolecular $\pi$–$\pi$ interactions in flexibly linked partially fluorinated bisarenes in the gas phase. *Angew. Chem. Int. Ed.* **56**, 13259–13263 (2017).
56. Besnard, J. et al. Automated design of ligands to polypharmacological profiles. *Nature* **492**, 215–220 (2012).
57. Takeya, J. et al. Very high-mobility organic single-crystal transistors with in-crystal conduction channels. *Appl. Phys. Lett.* **90**, 102120 (2007).
58. Jurchescu, O. D., Baas, J. & Palstra, T. T. M. Effect of impurities on the mobility of single crystal pentacene. *Appl. Phys. Lett.* **84**, 3061–3063 (2004).
59. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).
60. Srinivas, N., Krause, A., Kakade, S. M. & Seeger, M. W. Information-theoretic regret bounds for gaussian process optimization in the bandit setting. *IEEE Trans. Inf. Theory* **58**, 3250–3265 (2012).
61. Auer, P. Using confidence bounds for exploitation-exploration trade-offs. *J. Mach. Learn. Res.* **3**, 397–422 (2002).
62. Srinivas, N., Krause, A., Kakade, S. & Seeger, M. Gaussian process optimization in the bandit setting: no regret and experimental design. In *Proceedings of the 27th International Conference on International Conference on Machine Learning, ICML'10,* 1015–1022 (Omnipress, Madison, WI, USA, 2010).
63. Vanchinathan, H. P., Marfurt, A., Robelin, C.-A., Kossmann, D. & Krause, A. Discovering valuable items from massive data. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '15*, 1195–1204 (Association for Computing Machinery, New York, NY, USA, 2015) https://doi.org/10.1145/2783258.2783360.
64. Ma, Y., Huang, T.-K. & Schneider, J. Active search and bandits on graphs using sigma-optimality. In *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence, UAI 2015*, 542–551 (2015).
65. Pinsler, R., Gordon, J., Nalisnick, E. & Hernández-Lobato, J. M. Bayesian batch active learning as sparse subset approximation. In *Advances in Neural Information Processing Systems 32*, (eds Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. & Garnett, R.) 6359–6370 (Curran Associates, Inc., 2019).
66. Ash, J. T., Zhang, C., Krishnamurthy, A., Langford, J. & Agarwal, A. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations* (2020).
67. Madhawa, K. & Murata, T. A multi-armed bandit approach for exploring partially observed networks. *Appl. Netw. Sci.* **4**, 26 (2019).
68. Browne, C. et al. A survey of monte carlo tree search methods. *IEEE Trans. Comput. Intell. AI Games* **4**, 1–43 (2012).
69. Zhou, Z., Kearnes, S., Li, L., Zare, R. N. & Riley, P. Optimization of molecules via deep reinforcement learning. *Sci. Rep.* **9**, 10752 (2019).
70. Grimme, S., Bannwarth, C. & Shushkov, P. A robust and accurate tight-binding quantum chemical method for structures, vibrational frequencies, and noncovalent interactions of large molecular systems parametrized for all spd-block elements (z = 1–86). *J. Chem. Theory Comput.* **13**, 1989–2009 (2017).
71. Kunkel, C., Schober, C., Oberhofer, H. & Reuter, K. Knowledge discovery through chemical space networks: the case of organic electronics. *J. Mol. Model.* **25**, 87 (2019b).
72. Brückner, C. & Engels, B. A theoretical description of charge reorganization energies in molecular organic p-type semiconductors. *J. Comput. Chem.* **37**, 1335–1344 (2016).
73. Ralaivola, L., Swamidass, S. J., Saigo, H. & Baldi, P. Graph kernels for chemical informatics. *Neural Netw.* **18**, 1093 – 1110 (2005).
74. Himanen, L. et al. Dscribe: Library of descriptors for machine learning in materials science. *Comput. Phys. Commun.* **247**, 106949 (2020).
75. Quinonero-Candela, J. A unifying view of sparse approximate gaussian process regression. *J. Mach. Learn. Res.* **6**, 1939–1959 (2005).
76. Beluch, W. H., Genewein, T., Nürnberger, A. & Köhler, J. M. The power of ensembles for active learning in image classification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018) pp. 9368–9377.
77. Kendall, A. & Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems* (2017).
78. Zhang, Y. & Lee, A. A. Bayesian semi-supervised learning for uncertainty-calibrated prediction of molecular properties and active learning. *Chem. Sci.* **10**, 8154–8163 (2019).
79. Wenzel, F. et al. How good is the bayes posterior in deep neural networks really? In *International Conference on Machine Learning* (2020).
80. Curtarolo, S., Hart, G. L. W., Nardelli, M. B., Mingo, N., Sanvito, S. & Levy, O. The high-throughput highway to computational materials design. *Nat. Mater.* **12**, 191–201 (2013).
81. Pracht, P., Bauer, C. A. & Grimme, S. Automated and efficient quantum chemical determination and energetic ranking of molecular protonation sites. *J. Comput. Chem.* **38**, 2618–2631 (2017).
82. Andersen, M., Levchenko, S. V., Scheffler, M. & Reuter, K. Beyond scaling relations for the description of catalytic materials. *ACS Catal.* **9**, 2752–2759 (2019).
83. Cubuk, E. D., Sendek, A. D. & Reed, E. J. Screening billions of candidates for solid lithium-ion conductors: a transfer learning approach for small data. *J. Chem. Phys.* **150**, 214701 (2019).
84. Coley, C. W., Rogers, L., Green, W. H. & Jensen, K. F. Scscore: synthetic complexity learned from a reaction corpus. *J. Chem. Inf. Model.* **58**, 252–261 (2018).
85. Ishii, H. et al. Charge mobility calculation of organic semiconductors without use of experimental single-crystal data. *Sci. Rep.* **10**, 2524 (2020).
86. Friederich, P. et al. Molecular origin of the charge carrier mobility in small molecule organic semiconductors. *Adv. Funct. Mater.* **26**, 5757–5763 (2016).
87. Landi, A. & Troisi, A. Rapid evaluation of dynamic electronic disorder in molecular semiconductors. *J. Phys. Chem. C* **122**, 18336–18345 (2018).
88. Egger, A. T. et al. Charge transfer into organic thin films: a deeper insight through machine-learning-assisted structure search. *Adv. Sci.* **7**, 2000992 (2020).
89. MacLeod, B. P. et al. Self-driving laboratory for accelerated discovery of thin-film materials. *Sci. Adv.* **6**, eaaz8867 (2020).


## Acknowledgements

C.K. and J.T.M are grateful for support by Deutsche Forschungsgemeinschaft (DFG) through TUM International Graduate School of Science and Engineering (IGSSE), GSC 81. C.K., H.O. and K.R. gratefully acknowledge support from the Solar Technologies Go Hybrid initiative of the State of Bavaria. K.C. acknowledges funding from the China Scholarship Council. We also thankfully acknowledge computational resources provided by the Leibniz Supercomputing Centre. J.T.M. would like to acknowledge illuminating discussions on synthesizability and chemical stability with Tobias Schaub.

## Author contributions

C.K, H.O., J.T.M and K.R. conceived the idea. C.K. implemented the algorithms in code and carried out the calculations. Methodological details were thereby worked out by C.K., K.C. and J.T.M. C.K, H.O., J.T.M and K.R. wrote the manuscript.

## Funding

Open Access funding enabled and organized by Projekt DEAL.

## Competing interests
The authors declare no competing interests.

## Additional information
**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-22611-4.

**Correspondence** and requests for materials should be addressed to K.R.

**Peer review information** *Nature Communications* thanks Graeme Day and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

**Reprints and permission information** is available at http://www.nature.com/reprints

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

Paper 2

**Reorganization Energies of Flexible Organic Molecules as a Challenging Target for Machine Learning Enhanced Virtual Screening**

Ke Chen, Christian Kunkel, Karsten Reuter and Johannes T. Margraf

*Digital Discovery* **2022**, *1*, 147-157.

Digital Discovery

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue



Cite this: *Digital Discovery*, 2022, **1**, 147

# Reorganization energies of flexible organic molecules as a challenging target for machine learning enhanced virtual screening†

Ke Chen, [ab] Christian Kunkel, [ab] Karsten Reuter [ab] and Johannes T. Margraf *[ab]

Received 16th November 2021
Accepted 3rd February 2022

DOI: 10.1039/d1dd00038a

rsc.li/digitaldiscovery

The molecular reorganization energy $\lambda$ strongly influences the charge carrier mobility of organic semiconductors and is therefore an important target for molecular design. Machine learning (ML) models generally have the potential to strongly accelerate this design process (*e.g.* in virtual screening studies) by providing fast and accurate estimates of molecular properties. While such models are well established for simple properties (*e.g.* the atomization energy), $\lambda$ poses a significant challenge in this context. In this paper, we address the questions of how ML models for $\lambda$ can be improved and what their benefit is in high-throughput virtual screening (HTVS) studies. We find that, while improved predictive accuracy can be obtained relative to a semiempirical baseline model, the improvement in molecular discovery is somewhat marginal. In particular, the ML enhanced screenings are more effective in identifying promising candidates but lead to a less diverse sample. We further use substructure analysis to derive a general design rule for organic molecules with low $\lambda$ from the HTVS results.



## I. Introduction

By providing fast and accurate predictions of molecular properties, chemical machine learning (ML) has the potential to significantly increase the speed and scope of molecular discovery.[1–3] In this context, much attention has been paid on properties that are directly available from single-point electronic structure (*e.g.* density functional theory, DFT) calculations, such as atomization energies[4–6] or molecular orbital energies.[7,8] For established benchmark sets of small molecules like QM9,[9] state-of-the-art ML models now reach extremely high accuracies for such properties, often surpassing the intrinsic error of the reference electronic structure methods.

Despite this success, there remains a gap between the small, rigid molecules in QM9 and technologically or pharmaceutically relevant compounds, which are often larger and much more flexible. Furthermore, the target properties of molecular discovery are in practice seldom simple electronic properties that are directly accessible through single-point DFT calculations. Instead, complex properties like the bulk electronic conductivity, pharmacological or catalytic activity of a molecule are ultimately of interest.[10] Unfortunately, these are extremely complicated to rigorously simulate even for a single molecule. In high-throughput virtual screening (HTVS) studies, it has therefore become common to focus on simplified descriptors that are known to correlate with the property of interest.[11–13] Such descriptors include, *e.g.*, the binding energy of a key intermediate in catalysis or the internal reorganization energy ($\lambda$) in molecular electronics.

Measuring the energetic cost for charge-carriers to move between molecular sites,[14,15] $\lambda$ provides an important contribution to the charge-carrier mobility in crystalline and amorphous organic semiconductors.[16,17] While computational screening for low-$\lambda$ molecular structures has successfully guided discovery,[18] its sensitivity to small variations in molecular structure[19] renders a targeted molecular design challenging. Fragment[19–21] or rule-based[22,23] design strategies have been proposed to tackle this problem, while virtual screening[24–29] or data-efficient[30,31] discovery were used to assess large molecular candidate spaces, albeit without fully capturing the underlying structure–property relationships.

A reliable ML-based prediction of $\lambda$ could fill exactly this gap—providing significant speed-ups for the assessment of thousands of molecules while potentially allowing for the extraction of robust chemical rules by explainable AI.[32] ML-based approaches were indeed recently successful for the prediction of $\lambda$ for rigid molecules,[33] while flexible molecules still pose a significant challenge,[34] likely because $\lambda$ simultaneously depends on two potential energy surfaces (see Fig. 1).

[a]*Chair for Theoretical Chemistry, Catalysis Research Center, Technische Universität München, Lichtenbergstraße 4, D-85747 Garching, Germany. E-mail: margraf@fhi-berlin.mpg.de*

[b]*Fritz-Haber-Institut der Max-Planck-Gesellschaft, Faradayweg 4-6, D-14195 Berlin, Germany*

† Electronic supplementary information (ESI) available: Details on structure generation, electronic properties, hyperparameters and substructure analysis. See DOI: 10.1039/d1dd00038a

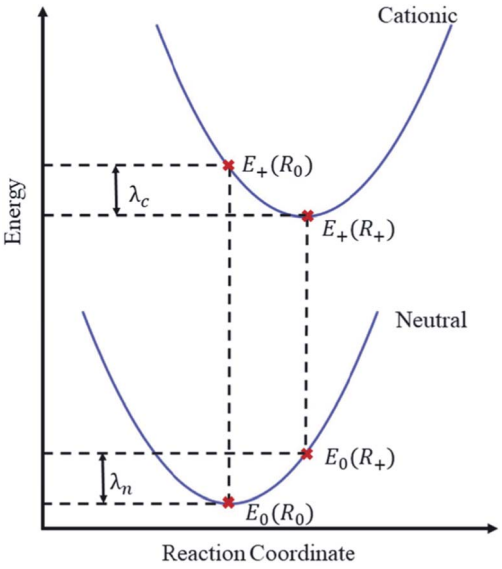



Fig. 1 Illustration of the adiabatic potential energy surfaces of neutral and cationic molecular states. The reorganization energy $\lambda$ is here calculated from the four indicated points[38] as $\lambda = E_0(R_+) - E_0(R_0) + E_+(R_0) - E_+(R_+)$. Focusing on holes as charge carriers, $E_0$ and $E_+$ are the total energies of the neutral and cationic molecular states, evaluated at the equilibrium geometries $R_0$ and $R_+$ of the respective states. In practice, two equilibrium geometries thus need to be obtained.

In this contribution we therefore critically study the ML prediction of $\lambda$ (specifically for hole conduction) as a challenging problem for chemical machine learning. To this end, we present a new dataset of hybrid DFT-level reorganization energies for 10 900 carbon and hydrogen containing molecules consisting of up to sixty atoms and five rotatable bonds. A series of Gaussian Process Regression (GPR)[35,36] models are developed for this dataset, both for straightforward structure/property mapping and Δ-ML[37] using a semiempirical baseline. We find that the conformational freedom of these molecules can introduce significant noise to this inference task, so that the performance of the models is strongly influenced by the conformer sampling method. We further show that significant improvements in the predictive performance are achieved by adopting the Δ-learning strategy. Finally, we critically evaluate the usefulness of the obtained ML methods for the discovery of low-$\lambda$ structures in a diverse chemical space and for deducing molecular design rules.

# II. Methods

### Dataset

A set of flexible π-conjugated hydrocarbon molecules was generated by successively applying a series of molecular transformation operations to benzene (see Fig. S1†), similar to the procedure used in ref. 30. At each step, these operations modify structural elements in the parent molecule or add additional ones. The set of operations used herein includes biphenyl-conjugation, annelation (5/6-ring) and ring-contraction, among others (see ESI† for details). Based on these transformations, molecular structures with up to four rings and two linker atoms were randomly generated, leading to 131 810 unique structures. This set forms the virtual screening space for this study. DFT calculations were performed for a subset of 10 900 structures as detailed in the section on Structure-based ML models.

While these molecules thus purposely cover a diverse molecular and conformational space, we note that—as with any enumerated chemical dataset—unstable and reactive systems could be contained and synthesizability should be assessed separately. All chemoinformatics-related tasks were carried out using RDKit 2019.09.03.[39]

### Reorganization energies

Reorganization energies were calculated for the lowest-energy conformer of each molecule. To determine this conformer, RDKit is first used to compute 2D coordinates for the molecular graph, while an initial 3D structural guess is obtained and relaxed at the GFN2-xTB level using the xTB program (v6.3.0).[40] Conformational search is then carried out using the iterative meta-dynamics sampling and genetic crossover (iMTD-GC) approach, as implemented in the "Conformer-Rotamer Ensemble Sampling Tool" (CREST).[41] Here, three different settings were compared as fully detailed in the Results section.

For the lowest-energy conformers, reorganization energies were computed at the GFN1-xTB level ($\lambda_{\mathrm{GFN1}}$). Note that GFN1-xTB was chosen instead of its successor (GFN2-xTB) because we found the former to be slightly more reliable in terms of predicting $\lambda$ and molecular geometries for the systems considered herein (see Fig. S2 and S3†). Electronic descriptor values entering property-based ML models (as detailed in the Results section) were also extracted from results of these calculations. These include frontier orbital energies and their gaps, Fermi levels, total energies and vertical energy differences. Final target $\lambda_{\mathrm{DFT}}$ values were calculated at the B3LYP[42–44] level of theory using the FHI-AIMS[45] code, including the TS dispersion correction.[46] Electronic wave functions were expanded in an extended "tier 1" basis set using "light" integration settings. Note that this level of theory is commonly employed for characterizing organic semiconductors, thus forming a good reference method for this study.[19,25,28,47]

### ML models

All models presented herein use GPR, a probabilistic machine learning method that allows for the smooth interpolation of property values from data. Specifically, these models infer the underlying relationship between different molecular representations and $\lambda$, based on a training set $D = \{\mathbf{X}, \mathbf{y}\}$. Here, $\mathbf{X}$ is a matrix consisting of molecular representation vectors $\mathbf{x}^{(i)}$ and $\mathbf{y}$ is a vector of target properties for the training molecules, with elements $y^{(i)}$. Predictions for a set of unseen molecular representations $\mathbf{X}^*$ can then be obtained as the predictive mean

$$\bar{\mathbf{y}}(\mathbf{X}^*) = \alpha \mathbf{K}(\mathbf{X}^*, \mathbf{X}), \quad (1)$$

where the covariance (or kernel) matrix $\mathbf{K}$ with elements $K_{ij} = K(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$ quantifies the similarity between molecular representations. The coefficients $\alpha$ minimize a regularized least-squares error between property predictions and reference values and can be calculated as

$$\alpha = (\mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma_n{}^2\mathbf{1})^{-1}\mathbf{y} \quad (2)$$

where $\mathbf{K}(\mathbf{X}, \mathbf{X})$ is again a covariance matrix. The hyperparameter $\sigma_n$ incorporates observation noise, in this case, *e.g.* related to uncertainty due to conformational sampling (as detailed in the section on Conformer sampling).

In all models reported herein, the commonly used radial basis function (RBF) kernel is employed:

$$k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \sigma_f{}^2 \exp\left(-\frac{d(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})^2}{2l^2}\right) \quad (3)$$

where the $l$ is the kernel length-scale, $\sigma_f{}^2$ is the signal variance and $d(.,.)$ is the Euclidean distance.

A series of GPR models are presented herein, which differ in the type of representation and in how the covariance matrix is constructed. The most straightforward of these uses a representation of the molecular geometry of the lowest-energy conformer in the neutral charge state. This representation $\mathbf{x}_s^{(i)}$ is constructed in two steps. First, each atomic environment is encoded into a rotationally invariant local representation using the smooth overlap of atomic positions (SOAP)[48] as implemented in Dscribe[49] (see Fig. S4† for details). These atomic representations are then combined into molecular representations using the auto-bag method,[50] which partitions the local feature vectors into $k_{max}$ clusters using the $k$-means algorithm.[51] Each molecular structure can then be encoded by a $k_{max}$-dimensional global feature vector that counts the occurrence of local environments that are assigned to each cluster. The effect of the hyperparameter $k_{max}$ on the predictive performance is shown in Fig. S5,† arriving at a converged value of 500. Here, SOAP is only one of the possible choices for representing atomic environments. In fact, there is a range of modern many-body representations, which are closely related to each other and typically display comparable accuracy.[52] To illustrate this we also considered the Many-Body Tensor Representation of Huo and Rupp.[53] This indeed yields very similar predictive performance for structure based models (see Fig. S6†).

Note that above we introduced the subscript s to refer to the use of structure-based molecular representations and the corresponding baseline ML model is denoted with $K_s$. Furthermore, a model termed $K_p$ based on electronic properties computed at the semiempirical GFN1-xTB level was developed, with the corresponding representation $\mathbf{x}_p^{(i)}$ (see below for details). Finally, a model $K_{sp}$ is explored, that combines the two kernel functions as $K_{sp}(i,j) = K_s(\mathbf{x}_s^{(i)},\mathbf{x}_s^{(j)}) + K_p(\mathbf{x}_p^{(i)},\mathbf{x}_p^{(j)})$.

The hyperparameters $\theta_s = (\sigma_{fs}, l_s, \sigma_n)$, $\theta_p = (\sigma_{fp}, l_p, \sigma_n)$, and $\theta_{sp} = (\sigma_{fs}, \sigma_{fp}, l_s, l_p, \sigma_n)$ for the respective models are determined by maximizing their log-marginal likelihood over $D$ using the L-BFGS algorithm with randomly sampled initial values. Our custom GPR model is based on respective code from the scikit-learn[54] implementation.

It should be noted that the choice of the ML method can in principle have a strong influence on the predictive accuracy. For the case of molecular reorganization energies, Abarbanel and Hutchison therefore performed an extensive comparison of different regression approaches (*e.g.* using kernel, decision tree and neural network based methods), finding little difference between different ML approaches.[34] To confirm this insensitivity, we also trained a decision tree based AdaBoost[55] model on the current data set and indeed found little difference to the GPR approach used herein (see Fig. S7†).

# III. Results

### Conformer sampling

The hydrocarbon dataset presented herein contains molecules with diverse structural elements (see Fig. 2a for 10 randomly selected examples). While the enumerated 2D molecular graphs contain information on molecular bonding, they do not fully determine the molecular geometry, *e.g.* with respect to relative configurations around rotatable single bonds. As an example, 115 888 (53 046) of the contained molecules incorporate at least 2 (4) rotatable bonds, with a maximum of 5 rotatable bonds occurring overall. We thus expect a significant conformational flexibility for these molecules.

This flexibility can influence the ML predictions of $\lambda$ in two ways. First, the reference $\lambda$ values may depend on the conformer, and flexible molecules display much larger conformational variety. Second, the ML prediction of $\lambda$ is based on a representation derived from a 3D molecular geometry. For highly flexible molecules, we can expect significantly larger deviations between the geometries predicted with more approximate levels of theory and high-level references. This is known to impact the accuracy of ML models adversely.[56] To arrive at an internally consistent procedure when comparing among different molecular systems, we therefore focus on the lowest energy conformers that we can identify for each molecular system.

Unfortunately, a full conformer search at the DFT level is prohibitively expensive. This means that we require a robust and efficient protocol for the search of low-energy conformers. To this end we rely on semiempirical and force-field methods from the GFN family, which have recently been established for this purpose. These are used in combination with CREST, which implements a purpose-built workflow for conformational search.[41] Depending on the underlying energy function, the accuracy and computational cost of this search can vary significantly, however. We therefore tested three different workflows, denoted as conf1-3.

In our reference method (conf1), we employ CREST in combination with the density functional tight-binding method GFN1-xTB.[40] Performing conformer searches for the 10 molecules of Fig. 2a, we find that between 3 and 90 conformers are identified within the default energy window of 6 kcal mol$^{-1}$ (260 meV) above the lowest energy one, underscoring the conformational flexibility of molecules in our dataset. For these

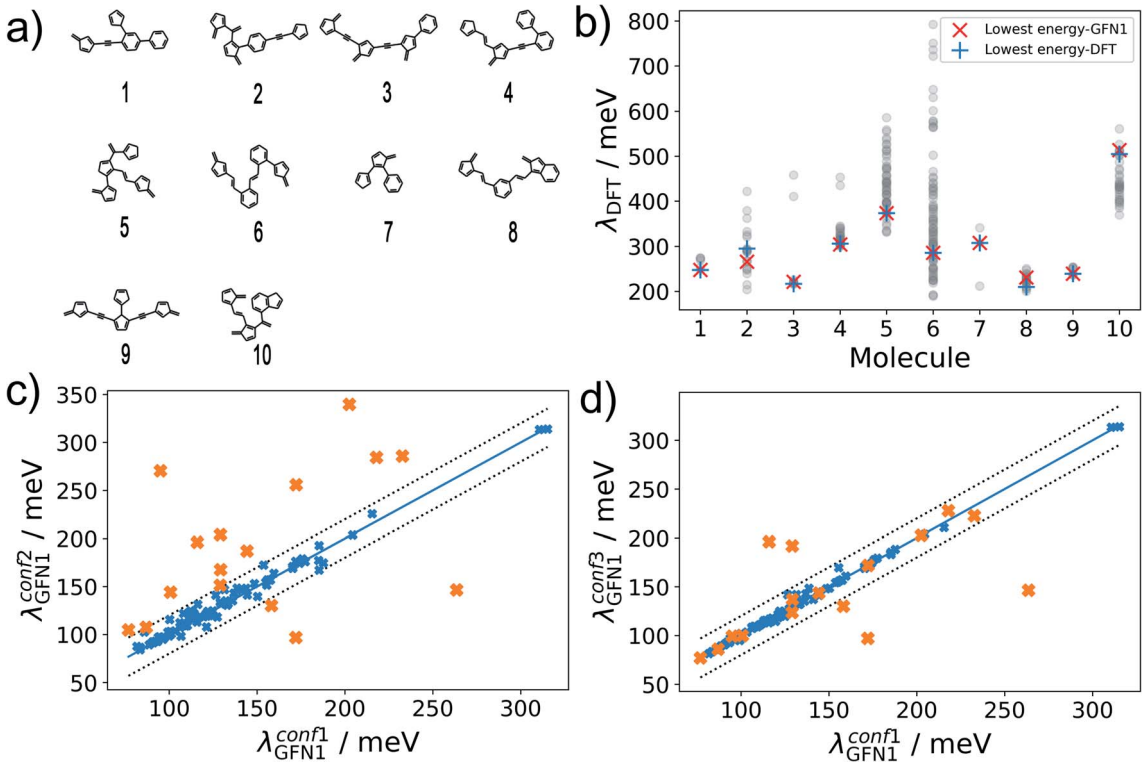



Fig. 2 Conformational diversity of the dataset. (a) Random molecules contained in the dataset. (b) Variability of $\lambda_{DFT}$ obtained for full conformer ensembles derived from conf1 searches. Respective values obtained for the lowest-energy DFT or GFN1-xTB conformers are marked. (c) Correlation between $\lambda_{GFN1}$ for the lowest energy conformers obtained by with conf1 and conf2. Outliers are marked in orange. (d) Improved correlation is obtained for conf3, while outliers of (c) are again marked in orange.

conformer ensembles, we show the wide range of encountered $\lambda_{DFT}$ values in Fig. 2b. Importantly, there is little variation between the values of $\lambda_{DFT}$ calculated for the lowest-energy conformers at the GFN1-xTB and DFT level, which suggests that GFN1-xTB conformers are a reliable proxy for the true first-principles ground state geometry. Note that the excellent agreement in Fig. 2b only reflects the quality of GFN1-xTB conformers, while all reorganization energies in this subfigure were calculated at the DFT level. Unfortunately, performing the full conformer search at the GFN1-xTB level is still computationally prohibitive for hundreds of thousands of molecules, however.

Alternatively, the significantly more efficient force-field method GFN-FF[57] can be used, and the conformer search be accelerated using the 'quick' setting in CREST (herein termed conf2). For 100 randomly selected molecules, Fig. 2c shows a comparison of $\lambda_{GFN1}$ values for the lowest-energy conformers obtained with conf1 and conf2. While the bulk of the predictions falls within the error margins of $\pm$20 meV, we also find 16 outliers – marked in orange. These can be attributed to an incomplete coverage of conformational space in the conf2 ensemble and to differences in the energetic ranking between GFN1-xTB and GFN-FF.

To address the latter point, in conf3 we therefore combine the higher accuracy of GFN1-xTB and the computational speed of GFN-FF: a conformer ensemble is generated with CREST at the GFN-FF level, while a subsequent local relaxation and energetic re-ranking is carried out using GFN1-xTB. Comparing again to conf1, we see a significantly better agreement between the methods (see Fig. 2d), with 5 remaining outliers falling beyond the error margins of $\pm$20 meV. It should be noted, that conformer searches are in general a difficult global optimization problem, which cannot be solved deterministically in an efficient manner. Therefore, some amount of uncertainty is unavoidable and will affect the ML models in all cases. As discussed in the following, achieving lower uncertainty at this stage leads to significantly lower predictive errors, however.

### Structure-based ML models

Having established an efficient conformer search workflow, we now turn to structure based ML models for predicting $\lambda$ ($K_s$). As these models require 3D geometries as inputs, they are well suited to investigate the effect of the conformer search protocols on the ML models themselves, see Fig. 3. Here, learning curves for $\lambda_{GFN1}$ and $\lambda_{DFT}$ are shown. While all models improve with more data, two striking differences can be seen. First, the models using the more accurate conformer search conf3 are consistently better than the ones using conf2. Second, the predictive error is consistently lower for $\lambda_{GFN1}$ than for $\lambda_{DFT}$.

In part, this can be explained by the smaller range of $\lambda_{GFN1}$ values (see next section). However, a fundamental difference between the two targets also exists: While we predict $\lambda_{GFN1}$ on

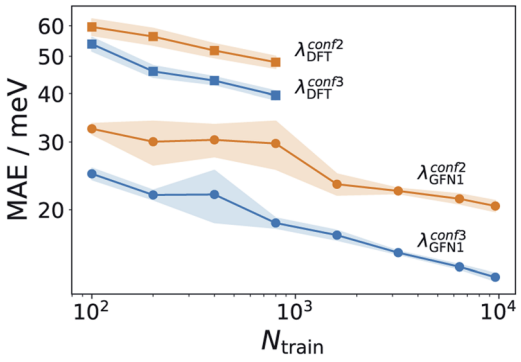



Fig. 3 Effect of improved conformer searches on learning behavior. Learning curves for $\lambda_{DFT}$ and $\lambda_{GFN1}$ using the conf2 and conf3 conformer search protocols. Training sets for $\lambda_{GFN1}$ ($\lambda_{DFT}$) consist of up to 9900 (880) molecules, with 1000 (100) unseen data points used to evaluate the predictive errors. Shaded errors indicate the standard deviation for five randomly draw training sets of each size. Note that the DFT assessment was stopped earlier due to the significantly higher computational cost of the method.

the basis of the corresponding neutral state molecular equilibrium structures, this does not hold for $\lambda_{DFT}$. In, the latter case, the differing neutral state equilibrium geometries (between GFN1-xTB and DFT) further complicate the learning task.

It should be noted here that learning $\lambda_{GFN1}$ is itself only of methodological interest, however. Indeed, the conf3 search requires GFN1-xTB for energy ranking, which has a similar computational effort to calculating $\lambda_{GFN1}$. In the following, we therefore exclusively focus on predicting $\lambda_{DFT}$, using conf3 for structure generation. To this end we extended our DFT annotated dataset to cover in total 10 900 molecules, randomly drawn from the full hydrocarbon database. The distribution of obtained $\lambda_{DFT}$ values is shown in Fig. S8.† 1000 molecules served as an external test set for model validation, while at maximum 9600 of the remaining 9900 entered the respective training sets.

### Beyond structure-based models

While the above results show that $\lambda_{DFT}$ can be learned from the structure, the accuracy of the models leaves something to be desired, given that the intrinsic standard deviation of the dataset is *ca.* 80 meV. To explore how this performance is impacted by molecular flexibility, additional $\Delta K_s$ models were trained on different subsets of 1000 molecules with a fixed number of rotatable bonds ($N_{rb}$ = 2,3,4,5). These models were then evaluated on test sets with the corresponding $N_{rb}$ (see Fig. S9†). We find that models for less flexible molecules are indeed significantly more accurate than those for more flexible molecules. This confirms the notion that molecular flexibility poses a challenge for molecular ML models and underscores our previous point on the highly challenging nature of $\lambda$ as a target property, *e.g.* compared to the atomization energy.

Since robust models already require the use of GFN1-xTB for conformer ranking, it is natural to ask whether electronic properties at the GFN1-xTB level could be used to improve them. The most straightforward way to do this is *via* a Δ-learning[37] strategy, *i.e.* by learning a correction to $\lambda_{GFN1}$. To this end, we first use a simple linear regression to describe systematic differences between $\lambda_{DFT}$ and $\lambda_{GFN1}$:

$$\lambda_{lin} = a\lambda_{GFN1} + b \quad (4)$$

This linear model alone yields a stable MAE of 40 meV, independent of the training set size. It thus outperforms the structure based $K_s$ models for all but the largest training sets (see Fig. 4). This means that, contrary to the findings of ref. 34 we find a reasonably good correlation between GFN and DFT based reorganization energies ($R^2 = 0.54$, see Fig. S10†). This is likely due to the different class of molecules (thiophene oligomers) considered therein. Defining as a new target property:

$$\lambda_\Delta = \lambda_{DFT} - \lambda_{lin}, \quad (5)$$

we can now build Δ-learning models that further improve on the linear approach. As expected, the Δ-learning variant of $K_s$ (termed $\Delta K_s$) indeed performs significantly better than both the linear and the baseline model, approaching an MAE of 30 meV at the largest training set size.

The GFN1-xTB calculations required for obtaining $\lambda_{GFN1}$ can also be exploited in a different way. One challenge for the structure-based models is the indirect relationship between the neutral GFN1-xTB geometry and $\lambda_{DFT}$. We therefore also explored property-based models (termed $K_p$) which use frontier orbital energies and gaps, Fermi levels, total energies and vertical energy differences of the neutral and cationic system to construct a representation, as fully detailed in Table S2.† The respective $\Delta K_p$ model is actually slightly better than the corresponding structure-based model $\Delta K_s$, despite not including any structural information. Finally, a combined model incorporating the structural and property kernels (termed $\Delta K_{sp}$), performs better still, reaching an MAE of 25 meV at the largest training set size.

Please note that no optimization of the feature selection was performed for the property based models, other than checking

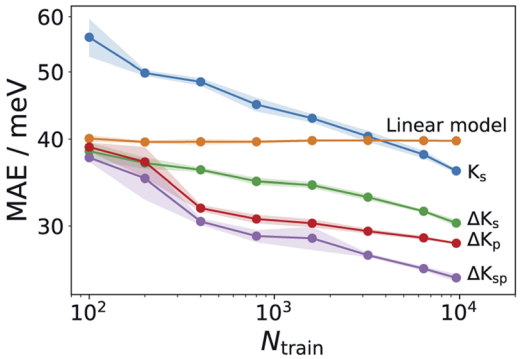



Fig. 4 Learning curves for various ML models. Comparison of $K_s$ models with various Δ-learning approaches. Shadings analogous to Fig. 3. The $K_s$ model corresponds to the curve labeled $\lambda_{DFT}^{conf3}$ in that figure.

that there were no strong linear dependencies between different properties. However, a more systematic feature selection procedure can provide physical insight and potentially improve the models. To explore this, we performed permutational feature importance (PFI) analysis for the $\Delta K_p$ model (see Fig. S11†).[59] This indicates that some features are particularly relevant for the model, *e.g.* the HOMO energy of the cationic state in the neutral geometry, the Fermi energy of the neutral state in the cation geometry and the individual contributions to the GFN1 reorganization energy. Based on this, we constructed additional models which only used subsets of the most important features. However, these sparse models displayed somewhat worse performance than the full model, indicating that all features ultimately contribute to the prediction accuracy. Nonetheless, more sophisticated feature engineering (*e.g.* using recursive selection or nonlinear transformations) may be able to achieve better performance with sparse models.

### ML-assisted virtual screening

So far, we have seen that in a Δ-ML setting, the presented GPR models can lead to a modest increase in predictive performance relative to a semiempirical baseline method. This raises the question of whether this improvement has a tangible effect on the results of a HTVS for low-$\lambda_{DFT}$ molecules. To address this issue, we applied $\Delta K_s$, $\Delta K_{sp}$ (each trained on 9600 molecules) and GFN1-xTB to screen 120 910 previously unseen molecules for promising candidates. For each model, we extracted 500 candidates with the lowest predicted $\lambda$ and calculated their actual $\lambda_{DFT}$ values.

As illustrated in Fig. 5a, all three methods are quite successful in identifying promising candidates: from the 500 selected systems, GFN1-xTB identifies 436 molecules that display $\lambda_{DFT} < 200$ meV, compared to the somewhat higher numbers for the $\Delta K_s$ and the $\Delta K_{sp}$ models (where 487 and 492 are respectively identified). Narrowing the range to $\lambda_{DFT} < 140$ meV, the $\Delta K_{sp}$ still performs best and identifies 251 structures, while the $\Delta K_s$ and the GFN1-xTB identify 217 and 118 such cases, respectively.

The 20 lowest-$\lambda$ structures from all three screenings are shown in Fig. 6. Interestingly, 15 compounds in this subset were identified by the GFN1-xTB screening, while the $\Delta K_s$ and $\Delta K_{sp}$ models identified 9 and 11, falling slightly behind. In other words, the GFN1-xTB model actually has an edge over the ML model when considering the extreme low end of the distribution, although it is in general less effective in identifying low-$\lambda$ structures. It is also notable that, although some overlap between the methods is observed (*i.e.* from the 1500 molecules selected by the three screenings only 1131 are unique candidates), many structures are exclusively identified by one method, in particular by GFN1-xTB. This is illustrated by the Kernel principal component analysis map[58] shown in Fig. 5b, which places similar molecular structures close to each other. Clearly, the semiempirical GFN1-xTB model overall exhibits the highest diversity, while the candidates selected by the data-driven models appear somewhat more concentrated. This reflects the fact that GPR models use metrics of molecular similarity in their predictions.

However, this is not primarily just a problem of the chosen models, since other ML approaches also (implicitly) work with feature similarity. It is rather that ML models are by definition most strongly influenced by those types of molecules which occur most frequently in the dataset. The HTVS setting does not necessarily require a good description of an *average* molecule, however. Instead, it requires a good description of the small percentage of unusual molecules that we are interested in. This implies that a non-uniform sampling strategy for training set construction might be helpful in this context. This will be explored in future work.

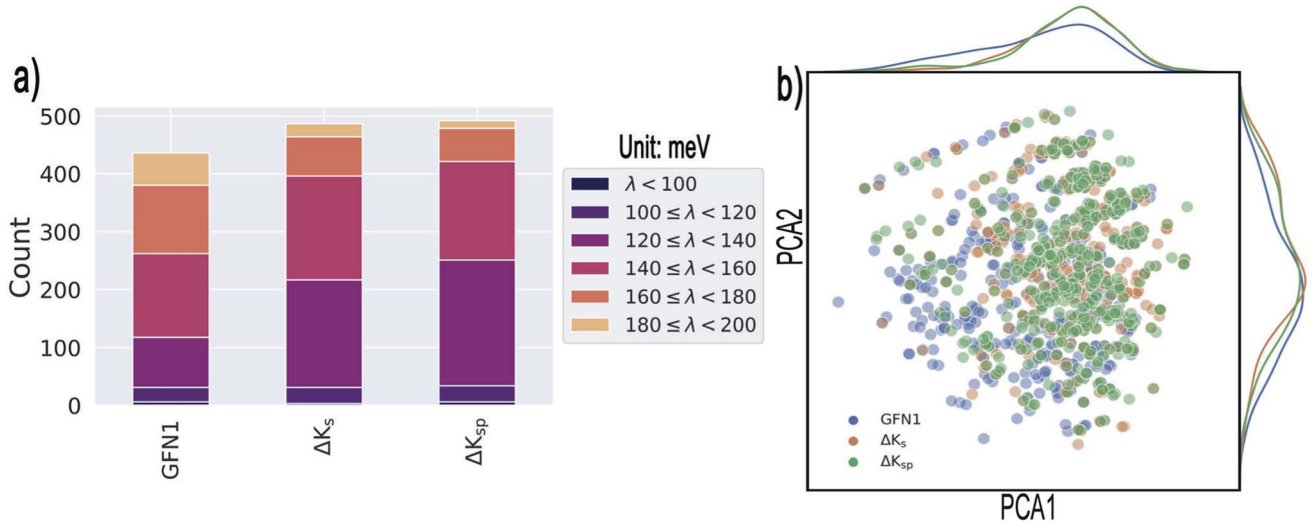



Fig. 5 Results of the targeted identification of low-$\lambda$ structures. (a) Distribution of $\lambda_{DFT}$ values in the final selections derived from three different methods (see text). We only consider compounds that satisfy $\lambda_{DFT} < 200$ meV. (b) Kernel principal component analysis map of the identified structures (generated with the ASAP[58] code). Kernel-density estimates are shown along the principal components.

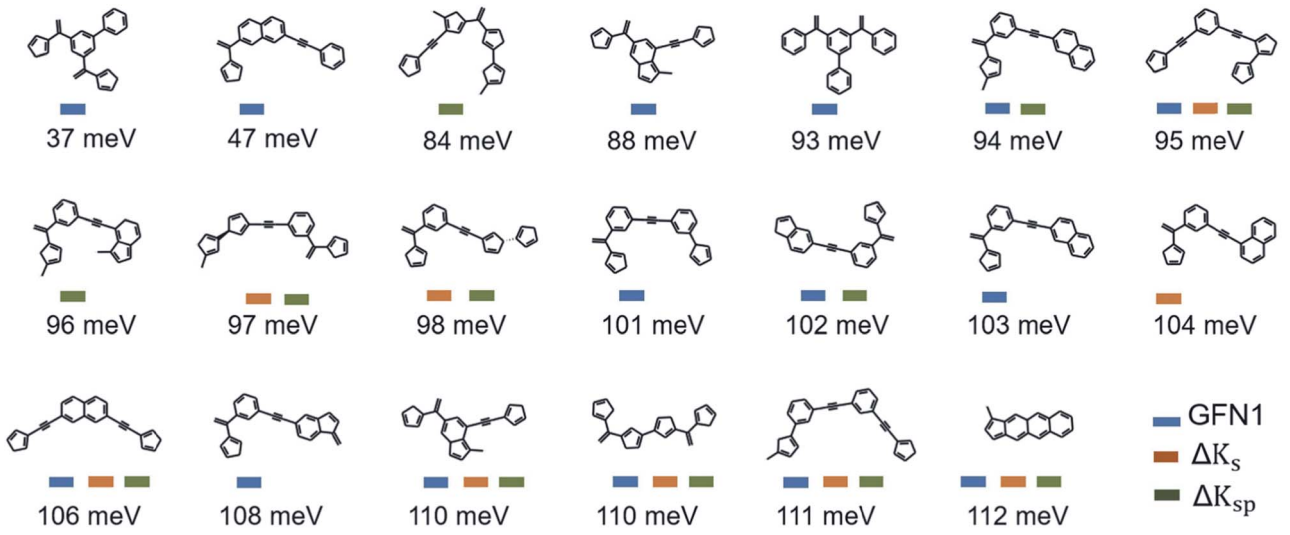



Fig. 6 Lowest-$\lambda_{DFT}$ candidates. Shown are the best candidates identified among 120k molecules in the three virtual screening campaigns. The corresponding $\lambda_{DFT}$ values are listed below.

At the suggestion of a reviewer, the virtual screening was also performed with the $\Delta K_p$ approach (see Fig. S12 and S13†). This model shows comparable performance to $\Delta K_s$ for systems with $\lambda < 140$ meV, but is considerably worse for the range 140 meV $< \lambda <$ 200 meV. This indicates that the structural information in $\Delta K_s$ and $\Delta K_{sp}$ helps the models to reliably identify systems that are structurally similar to low-$\lambda$ training set molecules, thus increasing their screening accuracy.

### Substructure analysis

Given a set of candidates from HTVS like the one in Fig. 6, it is natural to ask what makes these systems such good candidates. If general design rules could be obtained from this set, this would arguably be even more useful than the candidates themselves. Visual inspection indeed points to certain structural motifs that are fairly common, such as cyclopentadiene moieties and acetylene-bridged aromatic rings.

A more quantitative understanding of this can be obtained from a substructure analysis. To this end, we analysed whether certain structural motifs are significantly more likely to be found in the low-$\lambda$ subset than in the full dataset. This can be quantified *via* the *enrichment* of a given substructure, defined as

$$\chi_i = \frac{(n_{i,low}/N_{low})}{(n_{i,all}/N_{all})}, \quad (6)$$

where $n_{i,low}$ and $n_{i,all}$ are the number of times substructure i is found in the low-$\lambda$ and full datasets, while $N_{low}$ and $N_{all}$ are the total number of molecules in each dataset. We complement this metric with the *frequency* of a given substructure in the dataset, defined as

$$f_i = (n_{i,all}/N_{all}). \quad (7)$$

To obtain a general design rule, we search for substructures with both high enrichment and reasonably high frequency. This allows balancing between overly specific substructures that only occur in very few molecules to begin with (high enrichment/low frequency) and overly simple motifs that occur in many molecules, independent of $\lambda$ (low enrichment/high frequency).

As a preliminary screening, potential substructures were defined *via* Morgan-fingerprints[60] of different bond-radii (see Fig. 8). As illustrated in Fig. S14,† this revealed a number of highly enriched substructures, which confirmed the initial impression that acetylene-bridged and cyclopentadiene containing structures are highly favourable. However, the substructures obtained in this fashion are often redundant and chemically unintuitive (*i.e.* by only containing parts of aromatic rings). We therefore manually derived a number of reasonable substructures from this analysis, in order to elucidate a robust and general design rule for low-$\lambda$ molecules (see Fig. 7). Here, we focused on acetylene-bridged benzene rings, as cyclopentadiene is prone to dimerize in Diels–Alder reactions, pointing to potential stability issues with these molecules.

In Fig. 7a, we plot the enrichment and frequency of each substructure. This reveals a contravening trend: The simplest structure (1) is very common in the full dataset, but also displays very low enrichment in the low-$\lambda$ set. In contrast, the more elaborate structures (8) and (9) are highly enriched, but very rare overall. Meanwhile substructure (5) (two meta-substituted acetylene-bridged benzene rings) features a quite high enrichment and is also fairly common in the database. As a consequence, ten further molecules with this motif can be found in the previously computed set of 10 900 $\lambda_{DFT}$-values. This allows us to confirm that the corresponding molecules indeed display significantly lower reorganization energies than the full training set (Fig. 7c).

The distributions of $\lambda_{DFT}$-values for all substructures are shown in Fig. 7d. This confirms the impression obtained from the enrichment plots. Simple substructures like (1) are generally unspecific and can be found in both high- and low-$\lambda$ molecules.

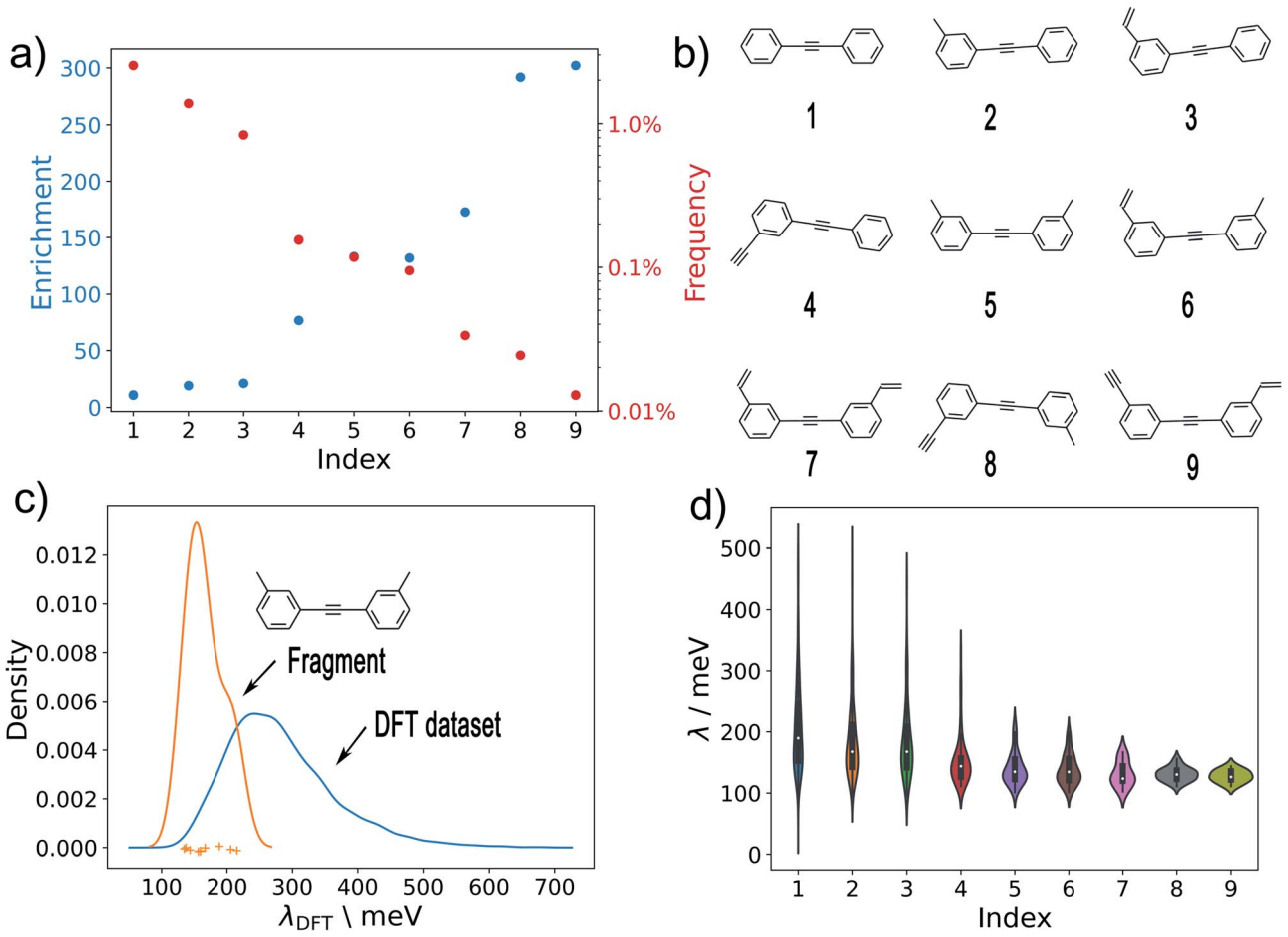



Fig. 7 Substructure analysis. (a) The enrichment and frequency of different substructures in the low-$\lambda$ and full datasets, respectively. (b) Analysed substructures. (c) The kernel density estimated $\lambda_{DFT}$ distributions of substructure 5 (shown in (b)) in the full training and validation sets (*i.e.* in 10 900 DFT datapoints). The individual $\lambda$-values of the ten molecules containing substructure 5 are shown as crosses. (d) Violin plots of $\lambda_{DFT}$ for all substructures in all $\lambda_{DFT}$ data.

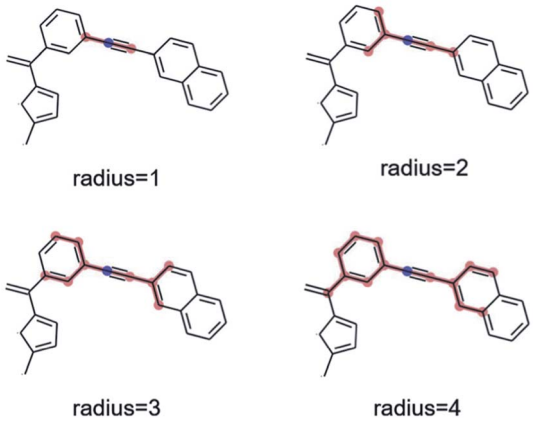



Fig. 8 Graphical illustration of Morgan fingerprints with various radii. Fingerprints allow highlighting common structural motifs but also produce redundant results and may unintuitively cut through aromatic rings or functional groups.

Meanwhile, highly enriched substructures indeed robustly predict high quality candidates, and can thus be used to define general design rules.

It should be noted that the above analysis is ultimately limited by the biases of the underlying dataset. For example, heteroatomic substituents could affect the suitability of certain motifs quite strongly due to electronic push–pull effects, which are largely absent in the hydrocarbon dataset used herein. Nonetheless, the methodology we apply could of course also be applied to other datasets.

## IV. Conclusion

In this work we have explored the potential benefits of using ML models to enhance virtual screening studies for molecules with low reorganization energies $\lambda$. We find that this is a challenging setting for molecular ML, both because of the conformational flexibility of the studied hydrocarbon molecules and the intrinsic difficulty of predicting $\lambda$ from the equilibrium geometry alone. Both aspects can be mitigated by using a semi-empirical electronic structure method for conformer searching and as a baseline model (provided there is at least a moderate correlation with the target property).

While this leads to a significant improvement of the predictive performance compared to the baseline, we find that

the benefits of this are actually somewhat marginal in the context of virtual screening. Specifically, ML enhanced screening is more effective in identifying promising candidates, but the semiempirical model actually has some advantages in terms of candidate diversity. This calls into question whether the cost of building the ML models (in particular the generation of training data) is actually justified. In particular, computing $\lambda_{\mathrm{DFT}}$ for a single molecule takes on average 28 CPU hours on our hardware. In contrast, the generation of conformer ensembles (*ca.* 1 CPU hour per molecule) and the training of the ML models (one-time cost of 20 CPU hours for the largest training sets) are reasonably affordable. To obtain a clear advantage, more accurate and/or data-efficient ML models are thus required.

One way to achieve this would be to work with full conformer ensembles rather than single conformers to construct the representations.[61] It should also be noted that packing and contact effects occurring in molecular crystals or amorphous structures are known to influence the encountered solid-state conformation and flexibility for geometrical relaxation.[26,62,63] Potentially, generative ML models trained on condensed phase data could therefore help producing more realistic conformer ensembles.

## Data availability

Data and code for this paper is publicly available at https://gitlab.mpcdf.mpg.de/kchen/oscs.

## Conflicts of interest

The authors declare no competing financial interests.

## Acknowledgements

KC acknowledges funding from the China Scholarship Council. CK and JTM are grateful for support by Deutsche Forschungsgemeinschaft (DFG) through TUM International Graduate School of Science and Engineering (IGSSE), GSC 81. CK and KR gratefully acknowledge support from the Solar Technologies Go Hybrid initiative of the State of Bavaria. We also thankfully acknowledge computational resources provided by the Leibniz Supercomputing Centre.

## References

1 K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev and A. Walsh, Machine learning for molecular and materials science, *Nature*, 2018, **559**, 547–555.
2 O. A. von Lilienfeld, K.-R. Müller and A. Tkatchenko, Exploring chemical compound space with quantum-based machine learning, *Nat. Rev. Chem.*, 2020, **4**, 347–358.
3 E. O. Pyzer-Knapp, K. Li and A. Aspuru-Guzik, Learning from the harvard clean energy project: the use of neural networks to accelerate materials discovery, *Adv. Funct. Mater.*, 2015, **25**, 6495–6502.
4 M. Rupp, A. Tkatchenko, K.-R. Müller and O. A. Von Lilienfeld, Fast and accurate modeling of molecular atomization energies with machine learning, *Phys. Rev. Lett.*, 2012, **108**, 058301.
5 G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller and O. A. Von Lilienfeld, Machine learning of molecular electronic properties in chemical compound space, *New J. Phys.*, 2013, **15**, 095003.
6 H. Jung, S. Stocker, C. Kunkel, H. Oberhofer, B. Han, K. Reuter and J. T. Margraf, Size-extensive molecular machine learning with global representations, *ChemSystemsChem*, 2020, **2**, e1900052.
7 A. Stuke, M. Todorović, M. Rupp, C. Kunkel, K. Ghosh, L. Himanen and P. Rinke, Chemical diversity in molecular orbital energy predictions with kernel ridge regression, *J. Chem. Phys.*, 2019, **150**, 204121.
8 M. Glavatskikh, J. Leguy, G. Hunault, T. Cauchy and B. Da Mota, Dataset's chemical diversity limits the generalizability of machine learning predictions, *J. Cheminf.*, 2019, **11**, 1–15.
9 R. Ramakrishnan, P. O. Dral, M. Rupp and O. A. von Lilienfeld, Quantum chemistry structures and properties of 134 kilo molecules, *Sci. Data*, 2014, **1**, 1–7.
10 A. R. Thawani, R.-R. Griffiths, A. Jamasb, A. Bourached, P. Jones, W. McCorkindale, A. A. Aldrick, and A. A. Lee, The photoswitch dataset: a molecular machine learning benchmark for the advancement of synthetic chemistry, 2020, arXiv preprint arXiv:2008.03226.
11 J. Hachmann, R. Olivares-Amaya, S. Atahan-Evrenk, C. Amador-Bedolla, R. S. Sánchez-Carrera, A. Gold-Parker, L. Vogt, A. M. Brockway and A. Aspuru-Guzik, The harvard clean energy project: large-scale computational screening and design of organic photovoltaics on the world community grid, *J. Phys. Chem. Lett.*, 2011, **2**, 2241–2251.
12 E. O. Pyzer-Knapp, C. Suh, R. Gómez-Bombarelli, J. Aguilera-Iparraguirre and A. Aspuru-Guzik, What is high-throughput virtual screening? a perspective from organic materials discovery, *Annu. Rev. Mater. Res.*, 2015, **45**, 195–216.
13 Ö. H. Omar, M. del Cueto, T. Nematiaram and A. Troisi, High-throughput virtual screening for organic electronics: a comparative study of alternative strategies, *J. Mater. Chem. C*, 2021, **9**, 13557–13583.
14 V. Coropceanu, J. Cornil, D. A. da Silva Filho, Y. Olivier, R. Silbey and J.-L. Bredas, Charge transport in organic semiconductors, *Chem. Rev.*, 2007, **107**, 926–952.
15 D. P. McMahon and A. Troisi, Evaluation of the external reorganization energy of polyacenes, *J. Phys. Chem. Lett.*, 2010, **1**, 941–946.
16 C. Wang, H. Dong, W. Hu, Y. Liu and D. Zhu, Semiconducting π-conjugated systems in field-effect transistors: a material odyssey of organic electronics, *Chem. Rev.*, 2012, **112**, 2208–2267.
17 J. Mei, Y. Diao, A. L. Appleton, L. Fang and Z. Bao, Integrated materials design of organic semiconductors for field-effect transistors, *J. Am. Chem. Soc.*, 2013, **135**, 6724–6746.

18 A. N. Sokolov, S. Atahan-Evrenk, R. Mondal, H. B. Akkerman, R. S. Sánchez-Carrera, S. Granados-Focil, J. Schrier, S. C. Mannsfeld, A. P. Zoombelt, Z. Bao and A. Aspuru-Guzik, From computational discovery to experimental characterization of a high hole mobility organic crystal, *Nat. Commun.*, 2011, **2**, 1–8.

19 H. Geng, Y. Niu, Q. Peng, Z. Shuai, V. Coropceanu and J.-L. Bredas, Theoretical study of substitution effects on molecular reorganization energy in organic semiconductors, *J. Chem. Phys.*, 2011, **135**, 104703.

20 C. Kunkel, C. Schober, J. T. Margraf, K. Reuter and H. Oberhofer, Finding the right bricks for molecular legos: A data mining approach to organic semiconductor design, *Chem. Mater.*, 2019, **31**, 969–978.

21 K.-H. Lin and C. Corminboeuf, Fb-reda: fragment-based decomposition analysis of the reorganization energy for organic semiconductors, *Phys. Chem. Chem. Phys.*, 2020, **22**, 11881–11890.

22 W.-C. Chen and I. Chao, Molecular orbital-based design of π-conjugated organic materials with small internal reorganization energy: generation of nonbonding character in frontier orbitals, *J. Phys. Chem. C*, 2014, **118**, 20176–20183.

23 W. Huang, W. Xie, H. Huang, H. Zhang and H. Liu, Designing organic semiconductors with ultrasmall reorganization energies: insights from molecular symmetry, aromaticity and energy gap, *J. Phys. Chem. Lett.*, 2020, **11**, 4548–4553.

24 G. R. Hutchison, M. A. Ratner and T. J. Marks, Hopping transport in conductive heterocyclic oligomers: reorganization energies and substituent effects, *J. Am. Chem. Soc.*, 2005, **127**, 2339–2350.

25 M. Misra, D. Andrienko, B. Baumeier, J.-L. Faulon and O. A. von Lilienfeld, Toward quantitative structure-property relationships for charge transfer rates of polycyclic aromatic hydrocarbons, *J. Chem. Theory Comput.*, 2011, **7**, 2549–2555.

26 C. Schober, K. Reuter and H. Oberhofer, Virtual screening for high carrier mobility in organic semiconductors, *J. Phys. Chem. Lett.*, 2016, **7**, 3973–3977.

27 J. Yang, S. De, J. E. Campbell, S. Li, M. Ceriotti and G. M. Day, Large-scale computational screening of molecular organic semiconductors using crystal structure prediction, *Chem. Mater.*, 2018, **30**, 4361–4371.

28 E. Antono, N. N. Matsuzawa, J. Ling, J. E. Saal, H. Arai, M. Sasago and E. Fujii, Machine-learning guided quantum chemical and molecular dynamics calculations to design novel hole-conducting organic materials, *J. Phys. Chem. A*, 2020, **124**, 8330–8340.

29 T. Nematiaram, D. Padula, A. Landi and A. Troisi, On the largest possible mobility of molecular semiconductors and how to achieve it, *Adv. Funct. Mater.*, 2020, **30**, 2001906.

30 C. Kunkel, J. T. Margraf, K. Chen, H. Oberhofer and K. Reuter, Active discovery of organic semiconductors, *Nat. Commun.*, 2021, **12**, 1–11.

31 C. Y. Cheng, J. E. Campbell and G. M. Day, Evolutionary chemical space exploration for functional materials: computational organic semiconductor discovery, *Chem. Sci.*, 2020, **11**, 4922–4933.

32 J. Jiménez-Luna, F. Grisoni and G. Schneider, Drug discovery with explainable artificial intelligence, *Nat. Mach. Intell.*, 2020, **2**, 573–584.

33 S. Atahan-Evrenk and F. B. Atalay, Prediction of intramolecular reorganization energy using machine learning, *J. Phys. Chem. A*, 2019, **123**, 7855–7863.

34 O. Abarbanel and G. Hutchison, Machine learning to accelerate screening for marcus reorganization energies, *J. Chem. Phys.*, 2021, **155**, 054106.

35 C. Williams and C. Rasmussen, Gaussian processes for regression, in *Advances in neural information processing systems*, Max-Planck-Gesellschaft, MIT Press, Cambridge, MA, USA, 1996, pp. 514–520.

36 C. Rasmussen and C. Williams, *Gaussian processes for machine learning, Adaptative computation and machine learning series*, University Press Group Limited, 2006.

37 R. Ramakrishnan, P. O. Dral, M. Rupp and O. A. von Lilienfeld, Big data meets quantum chemistry approximations: The δ-machine learning approach, *J. Chem. Theory Comput.*, 2015, **11**, 2087–2096.

38 S. F. Nelsen, S. C. Blackstock and Y. Kim, Estimation of inner shell marcus terms for amino nitrogen compounds by molecular orbital calculations, *J. Am. Chem. Soc.*, 1987, **109**, 677–682.

39 *The RDKit: Open-Source Cheminformatics Software, version 2019.09.3.*, 2019, http://www.rdkit.org.

40 S. Grimme, C. Bannwarth and P. Shushkov, A robust and accurate tight-binding quantum chemical method for structures, vibrational frequencies, and noncovalent interactions of large molecular systems parametrized for all spd-block elements ($z = 1$–86), *J. Chem. Theory Comput.*, 2017, **13**, 1989–2009.

41 P. Pracht, F. Bohle and S. Grimme, Automated exploration of the low-energy chemical space with fast quantum chemical methods, *Phys. Chem. Chem. Phys.*, 2020, **22**, 7169–7192.

42 A. D. Becke, Density-functional exchange-energy approximation with correct asymptotic behavior, *Phys. Rev. A: At., Mol., Opt. Phys.*, 1988, **38**, 3098.

43 C. Lee, W. Yang and R. G. Parr, Development of the colle-salvetti correlation-energy formula into a functional of the electron density, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1988, **37**, 785–789.

44 P. J. Stephens, F. J. Devlin, C. F. Chabalowski and M. J. Frisch, Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields, *J. Phys. Chem.*, 1994, **98**, 11623–11627.

45 V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter and M. Scheffler, Ab initio molecular simulations with numeric atom-centered orbitals, *Comput. Phys. Commun.*, 2009, **180**, 2175–2196.

46 A. Tkatchenko and M. Scheffler, Accurate molecular van der Waals interactions from ground-state electron density and free-atom reference data, *Phys. Rev. Lett.*, 2009, **102**, 073005.

47 N. E. Gruhn, D. A. da Silva Filho, T. G. Bill, M. Malagoli, V. Coropceanu, A. Kahn and J.-L. Bredas, The vibrational

reorganization energy in pentacene: molecular influences on charge transport, *J. Am. Chem. Soc.*, 2002, **124**, 7918–7919.

48 A. P. Bartók, R. Kondor and G. Csányi, On representing chemical environments, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2013, **87**, 184115.

49 L. Himanen, M. O. Jäger, E. V. Morooka, F. F. Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke and A. S. Foster, Dscribe: library of descriptors for machine learning in materials science, *Comput. Phys. Commun.*, 2020, **247**, 106949.

50 S. A. Meldgaard, E. L. Kolsbjerg and B. Hammer, Machine learning enhanced global optimization by clustering local environments to enable bundled atomic energies, *J. Chem. Phys.*, 2018, **149**, 134104.

51 S. Lloyd, Least squares quantization in pcm, *IEEE Trans. Inf. Theory*, 1982, **28**, 129–137.

52 F. Musil, A. Grisafi, A. P. Bartók, C. Ortner, G. Csányi and M. Ceriotti, Physics-Inspired Structural Representations for Molecules and Materials, *Chem. Rev.*, 2021, **121**, 9759–9815.

53 H. Huo and M. Rupp, Unified Representation of Molecules and Crystals for Machine Learning, 2017, arXiv:1704.06439.

54 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. Perrot and E. Duchesnay, Scikit-learn: Machine learning in Python, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.

55 Y. Freund and R. E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, *J. Comput. Syst. Sci.*, 1997, **55**, 119–139.

56 A. P. Bartók, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi and M. Ceriotti, Machine learning unifies the modeling of materials and molecules, *Sci. Adv.*, 2017, **3**, e1701816.

57 S. Spicher and S. Grimme, Robust atomistic modeling of materials, organometallic, and biochemical systems, *Angew. Chem., Int. Ed.*, 2020, **132**, 15795–15803.

58 B. Cheng, R.-R. Griffiths, S. Wengert, C. Kunkel, T. Stenczel, B. Zhu, V. L. Deringer, N. Bernstein, J. T. Margraf, K. Reuter and G. Csanyi, Mapping materials and molecules, *Acc. Chem. Res.*, 2020, **53**, 1981–1991.

59 A. Altmann, L. Toloşi, O. Sander and T. Lengauer, Permutation importance: a corrected feature importance measure, *Bioinformatics*, 2010, **26**, 1340–1347.

60 D. Rogers and M. Hahn, Extended-connectivity fingerprints, *J. Chem. Inf. Model.*, 2010, **50**, 742–754.

61 S. Axelrod and R. Gomez-Bombarelli, Molecular machine learning with conformer ensembles, 2020, arXiv preprint arXiv:2012.08452.

62 G. Barbarella, M. Zambianchi, L. Antolini, P. Ostoja, P. Maccagnani, A. Bongini, E. A. Marseglia, E. Tedesco, G. Gigli and R. Cingolani, Solid-state conformation, molecular packing, and electrical and optical properties of processable β-methylated sexithiophenes, *J. Am. Chem. Soc.*, 1999, **121**, 8920–8926.

63 J. T. Blaskovits, K.-H. Lin, R. Fabregat, I. Swiderska, H. Wu and C. Corminboeuf, Is a single conformer sufficient to describe the reorganization energy of amorphous organic transport materials?, *J. Phys. Chem. C*, 2021, **125**, 17355–17362.

Paper 3

**Physics-Inspired Machine Learning of Localized Intensive Properties**

Ke Chen, Christian Kunkel, Bingqing Cheng, Karsten Reuter and Johannes T. Margraf
*Chemical Science* **2023**, *14*, 4913-4922.

Chemical Science

ROYAL SOCIETY OF CHEMISTRY

EDGE ARTICLE

View Article Online
View Journal | View Issue



Cite this: *Chem. Sci.*, 2023, **14**, 4913

All publication charges for this article have been paid for by the Royal Society of Chemistry

Received 14th February 2023
Accepted 10th April 2023

DOI: 10.1039/d3sc00841j

rsc.li/chemical-science



# Physics-inspired machine learning of localized intensive properties†

Ke Chen, [abc] Christian Kunkel, [a] Bingqing Cheng, [c] Karsten Reuter [ab] and Johannes T. Margraf *[a]

Machine learning (ML) has been widely applied to chemical property prediction, most prominently for the energies and forces in molecules and materials. The strong interest in predicting energies in particular has led to a 'local energy'-based paradigm for modern atomistic ML models, which ensures size-extensivity and a linear scaling of computational cost with system size. However, many electronic properties (such as excitation energies or ionization energies) do not necessarily scale linearly with system size and may even be spatially localized. Using size-extensive models in these cases can lead to large errors. In this work, we explore different strategies for learning intensive and localized properties, using HOMO energies in organic molecules as a representative test case. In particular, we analyze the pooling functions that atomistic neural networks use to predict molecular properties, and suggest an orbital weighted average (OWA) approach that enables the accurate prediction of orbital energies and locations.

## 1. Introduction

Due to their great potential for accelerating materials discovery and design, there has been significant interest in machine learning (ML) models that enable the fast and accurate prediction of molecular and materials properties.[1–5] Consequently, a wide range of neural network (NN) and Kernel ML methods have been developed and applied to systems ranging from isolated molecules to complex amorphous solids.[6–14]

In this context, many state-of-the-art approaches exploit the approximately local nature of chemical interactions. This is achieved by representing chemical structures in terms of the element of each atom and the types and positions of the atoms in its immediate surrounding (the chemical environment).[15–17] This is, *e.g.*, commonly used when developing ML interatomic potentials, where the total energy is then obtained as a sum of local atomic contributions (see Fig. 1).

There are two distinct but related advantages to this approach. On one hand, locality ensures that the computational cost of the model asymptotically displays linear scaling with the size of the system, allowing for instance the routine application of ML potentials to systems with a thousand atoms or more. On the other hand, the summation of atomic contributions ensures size-extensivity, which is often desirable, if not a key requirement as in the case of interatomic potentials.

Simply put, size-extensivity means that predicted properties (*e.g.* energies) scale linearly upon trivial extensions of the system size, *e.g.* when describing ideal crystals in larger periodic supercells or replicating non-interacting molecules. This allows size-extensive ML models to be trained on small molecules or simulation cells and later applied to large systems.[1,16,18] However, size extensivity is not necessarily always a good assumption. Indeed, many electronic properties like excitation energies,[19] orbital energies[20] or ionization potentials[21] are intensive, meaning that they remain constant for such trivial scalings of the system size. In this case summing over atomic contributions therefore yields unphysical results, in particular when extrapolating to systems that are larger than the ones contained in the training set.

From an ML perspective, the summation of atomic contributions is simply one of many possible pooling functions.[22–24] For example, when taking the average instead of the sum, predictions remain constant as the system size is scaled.[18,25] Average pooling is therefore often used as the default pooling function for intensive properties. Unfortunately, average pooling can still yield unphysical results, particularly when the target property is localized and the system has low symmetry.

To illustrate this, consider a model trained on the ionization energies (IEs) of isolated monomers of water (12.6 eV) and $CO_2$ (13.8 eV). An average pooling model will correctly predict that the IE remains constant for a non-interacting supersystem consisting of two separated water molecules. However, for

[a]Fritz-Haber-Institut der Max-Planck-Gesellschaft, Faradayweg 4-6, D-14195 Berlin, Germany. E-mail: margraf@fhi-berlin.mpg.de

[b]Chair for Theoretical Chemistry and Catalysis Research Center, Technische Universität München, Lichtenbergstraße 4, D-85747 Garching, Germany

[c]Institute of Science and Technology, Am Campus 1, 3400 Klosterneuburg, Austria

† Electronic supplementary information (ESI) available: Details on structure generation, model hyperparameters, additional learning curves, and further details on the LocalOrb dataset. See DOI: https://doi.org/10.1039/d3sc00841j

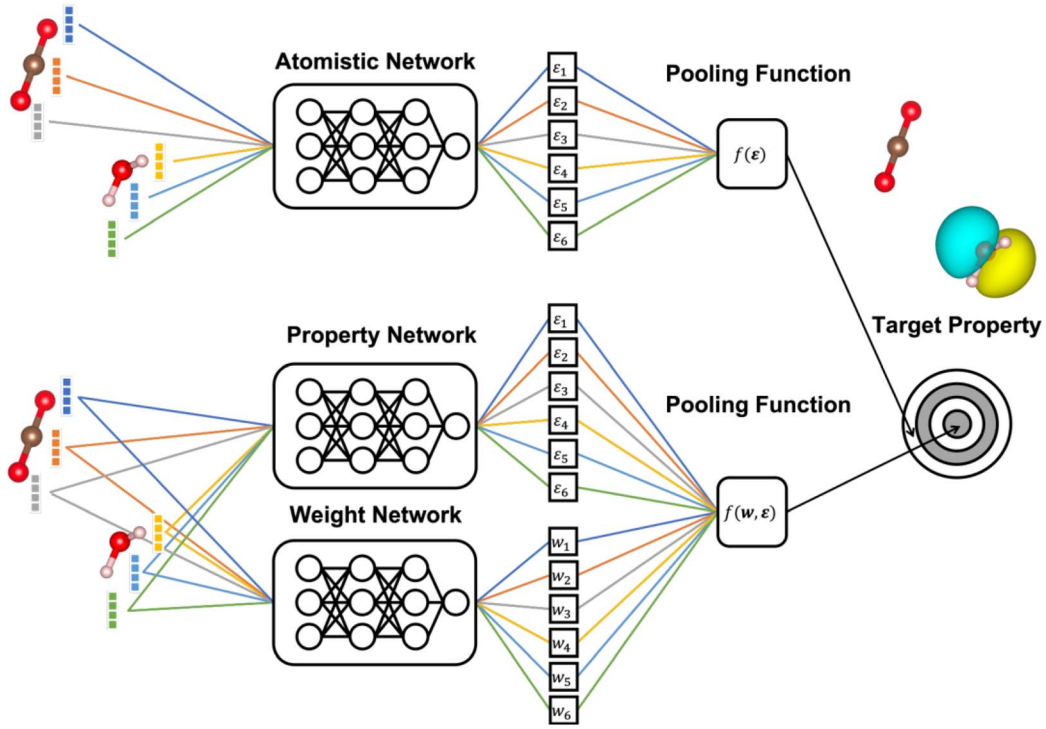



Fig. 1 Schematic illustration of atomistic neural networks. In a conventional atomistic neural network (top), the representation of each atomic environment is converted to a scalar output $\varepsilon_i$. These outputs are aggregated to the target property using a pooling function. The (orbital) weighted average models introduced herein ((O)WA) additionally predict the weight of each atom in the pooling function, using a second neural network (bottom). This is beneficial in the depicted example case of water and $CO_2$, where the target property (in this case an orbital energy) is localized on a part of the system.

a non-interacting supersystem consisting of one water and one $CO_2$ molecule, this model would predict that the IE is the average of the corresponding water and $CO_2$ values, which is clearly incorrect. The problem here is that the model fails to take into account that an ionization of this supersystem is localized on the water molecule, since it has the lower IE.

While this is a somewhat artificial example, many real chemical systems also display ionizations, excitations or orbitals that are spatially localized. Examples include disordered, defected or doped solids,[26,27] functionalized organic molecules and polymers,[28] as well as complex biomolecules like DNA and RNA.[29] This raises the question whether there are more appropriate pooling functions for electronic properties with a (potentially) localized nature.

In this contribution, we address this question by proposing a series of pooling functions that are formally able to treat localized (electronic) properties correctly. We then develop a new dataset of organic molecules, which is purposefully designed to contain both systems with localized and delocalized highest occupied molecular orbitals (HOMOs). This allows us to extensively benchmark the proposed pooling functions, and analyze their ability to predict the location of the orbital, as well as the energy. Finally, the most reliable methodology is applied to predict the orbital energies of the general OE62 dataset,[30] consisting of experimentally reported organic molecules with large structural diversity.

## 2. Methods

### 2.1 Atomistic neural networks

The general structure of an atomistic NN is shown in Fig. 1. Briefly, the chemical environment of an atom $i$ in a given system with $N$ atoms is represented by a vector or tensor $\chi_i$. This representation is passed through the NN to yield a scalar output $\varepsilon_i$. In a final step, the outputs of all atoms are combined to the global target property $P$ through a pooling function $f(\varepsilon_1,\ldots,\varepsilon_N)$, to be specified below.

Two classes of atomistic NNs are in common use. The original approach of Behler and Parinello uses a predefined set of radial and angular basis functions to generate the representation of the chemical neighborhood within a fixed cutoff radius around each atom.[15] Common choices for these predefined representations are the Atomic Symmetry Functions (ASFs) of Behler and Parinello, and the Superposition of Atomic Positions (SOAP) of Bartók and Csányi.[31,32] More recently, Message-Passing Neural Networks (MPNNs) have been proposed as an alternative.[33,34] These replace predefined representations with an end-to-end deep NN architecture that learns a data-driven representation during training.

The current paper is focused on the nature of the pooling function and not on the structural representation. For generality, we will therefore consider both approaches in the following. Specifically, the SOAP representation will be used as

implemented in Dscribe,[35] using the universal length scale hyperparameters defined in ref. 36. As a prototypical MPNN, the SchNet architecture is used.[16] For consistency, both SOAP and SchNet models are implemented with the PyTorch based SchNetPack library,[37] using default hyperparameters unless noted otherwise (see ESI† for details).

### 2.2 Pooling functions

In the following we focus on learning HOMO energies ($E_{\mathrm{HOMO}}$) as a prototypical localized intensive property. While the concepts we introduce below are generally applicable to all intensive properties, the concrete shape of the pooling function can vary depending on the target property. Any property-specific aspects will be highlighted when necessary.

The two most commonly used pooling functions in atomistic NNs are sum and average pooling, defined as

$$f_{\mathrm{sum}}(\varepsilon_1, \ldots, \varepsilon_N) = \sum_{i=1}^{N} \varepsilon_i, \tag{1}$$

and

$$f_{\mathrm{avg}}(\varepsilon_1, \ldots, \varepsilon_N) = \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i, \tag{2}$$

respectively. As discussed above, both of these yield unphysical results for localized intensive properties, however.

The simplest pooling function that potentially shows the correct behavior for such localized properties is max pooling, expressed as:

$$f_{\mathrm{max}}(\varepsilon_1,\ldots,\varepsilon_N) = \max(\{\varepsilon_1,\ldots,\varepsilon_N\}) \tag{3}$$

Note that here we are assuming that the target property is the energy of the highest occupied molecular orbital (HOMO). In other cases the min function would be appropriate, *e.g.* for the IE or the lowest unoccupied molecular orbital (LUMO) energy.

While $f_{\mathrm{max}}$ may have the desired formal properties, it arguably takes things too far since it ultimately makes the predicted molecular or materials property a function of a single atomic contribution. In real interacting systems, even fairly localized orbitals will typically extend over several atoms, however. More importantly, it would be desirable to have a pooling function that is simultaneously adequate both for localized and delocalized properties. A simple way to achieve this is *via* softmax pooling:

$$f_{\mathrm{softmax}}(\varepsilon_1, \ldots, \varepsilon_N) = \sum_{i=1}^{N} \frac{\exp(\varepsilon_i)}{\sum_{j=1}^{N} \exp(\varepsilon_j)} \varepsilon_i. \tag{4}$$

In a fully symmetrical system where each atom has an identical chemical environment this function behaves like average pooling, whereas it behaves more like max pooling in strongly unsymmetric cases like the above mentioned non-interacting water-$CO_2$ toy system.

More generally speaking, softmax pooling is just one example of a weighted average, with weights defined as $\frac{\exp(\varepsilon_i)}{\sum_{j=1}^{N} \exp(\varepsilon_j)}$. This assumes that both the target property and its localization can be simultaneously predicted from the scalar outputs $\varepsilon_i$. As a more flexible approach, the weights could also be predicted by a second NN, as shown on the bottom of Fig. 1. This leads to the general weighted average (WA) pooling:

$$f_{\mathrm{WA}} = \sum_{i=1}^{N} w_i \varepsilon_i, \tag{5}$$

Note that herein the softmax function (see eqn (4)) is used to normalize the outputs of the second NN, so that $\sum_{i}^{N} w_i = 1$ (see ESI†). This step rigorously enforces size-intensivity of the resulting models.

From a physical perspective it is interesting to consider what the ideal weights in WA pooling should be. For HOMO energy prediction it stands to reason that they should be related to the localization of the orbital. When the HOMO is expressed as a linear combination of atomic orbitals (indexed with $\mu$, $\nu$), the fraction $l_i$ of the orbital that is localized on a given atom $i$ can be obtained as:[38]

$$l_i = \left( \frac{\sum_{\mu \in i} {c_\mu}^2}{\sum_{\nu} {c_\nu}^2} \right), \tag{6}$$

where $c_\mu$ are the orbital coefficients in the atomic basis and the upper sum is restricted to all basis functions localized on atom $i$. Based on this, we can define an orbital coefficient based pooling function:

$$f_{\mathrm{coeff}}(\varepsilon_1, \ldots, \varepsilon_N) = \sum_{i=1}^{N} l_i \varepsilon_i. \tag{7}$$

Clearly, this function is of limited practical value for predicting orbital energies though. If the orbital coefficients were known, so would be the corresponding energies. Nonetheless we apply this coefficient pooling function below as a benchmark. In principle, it could also be applied with orbital coefficients from lower level methods, but this is beyond the scope of the current work.

As a practically tractable and computationally efficient approximation to $f_{\mathrm{coeff}}$, we explore including $l_i$ in the training procedure of WA models. In the resulting Orbital Weighted Average (OWA) approach, the loss function is augmented so that the weights reproduce the orbital localization fractions $l_i$ as closely as possible:

$$\mathscr{L}_{\mathrm{OWA}} = \frac{1}{N_{\mathrm{train}}} \left[ \alpha \sum_{A=1}^{N_{\mathrm{train}}} \left( E_{\mathrm{HOMO},A} - \sum_{i=1}^{N_A} w_{A,i} \varepsilon_{A,i} \right)^2 + \beta \sum_{A=1}^{N_{\mathrm{train}}} \times \sum_{i=1}^{N_A} \left( l_{A,i} - w_{A,i} \right)^2 \right] \tag{8}$$

Here, the loss is computed as an average over all $N_{\mathrm{train}}$ systems $A$ in the training set or batch. To clarify this, each of the previously

used variables is augmented with an additional index $A$ in this equation. The global parameters $\alpha$ and $\beta$ determine the relative contributions of orbital energies and localizations to the loss. The latter are optimized for orbital energy prediction on a separate validation set (see ESI†). In contrast, WA models are trained on the same purely orbital energy based loss function as the other models (see ESI†).

It should be noted that sum, average and max pooling have previously been used in the literature, *e.g.* in ref. 24, while the other approaches discussed herein are to the best of our knowledge used for the first time for molecular property prediction. We also note that the simple pooling functions used herein can in principle be replaced by separate neural network components, which try to learn appropriate pooling behaviour from data.[39] In this case, correct scaling with system size is not rigorously enforced, however.

### 2.3 LocalOrb dataset

Having established a series of pooling functions with desirable formal properties, our next goal is to benchmark how accurately the corresponding models can predict localized electronic properties. As a challenging test case we set out to predict HOMO energies in flexible organic molecules, which span a wide range of localization degrees. Specifically, a set of candidate molecules was generated by substituting 41 functional groups[40] at predefined positions of alkane or alkene backbones as illustrated in Fig. 2a. The chain length of these backbones varies from two to eight carbon atoms (see ESI† for a definition of all sidegroups and backbones, as well as further details on the dataset). All molecules in this chemical space were enumerated as SMILES strings, using the RDKit package.[41] Duplicated SMILES were detected and removed from the dataset, resulting in 21 081 unique 2D structures with a maximum of 11 rotatable bonds.

Initial 3D structures were generated from the SMILES strings using the ETKDG method[42] as implemented in RDKit. Based on these geometries, the CREST[43] package was used to explore the conformational space of each molecule at the semi-empirical GFN2-xTB level.[44] Default values were used for all CREST hyperparameters. Final geometries were obtained using the efficient *meta*-GGA composite density functional theory (DFT) method r2SCAN-3c[45] as implemented in ORCA 5.0.2.[46] To avoid the well known delocalization errors of semi-local density functionals, accurate orbital energies and coefficients were finally obtained with the range-separated hybrid wB97X-D3 (ref. 47) functional and def2-TZVP[48] basis set.

Note that the choice of saturated and conjugated backbones and the wide range of electron withdrawing and donating functional groups considered herein ensures a high diversity in the localization of the HOMO for these molecules (see Fig. 2b). This is further exacerbated by their high flexibility, which leads to an additional influence of the specific conformer configurations on orbital localization and energetics.[49]

For training and model evaluation, the 21 081 unique molecules were separated into two categories: to generate the training set, 4000 unique molecules were used. After the corresponding CREST runs, the lowest energy conformer and up to five further randomly selected conformers were used for DFT refinement, yielding 18 784 structures overall. To generate an independent test set, 15 462 of the remaining unique molecules were used. Here only the most stable conformer was refined with DFT for each molecule. This choice was made to maximize the chemical diversity in the test set, since we expect orbital locality to be more strongly influenced by the molecular structure than by the conformation.

### 2.4 Orbital localization index

As we are interested in the performance of the proposed pooling functions for both localized and delocalized HOMOs, a metric

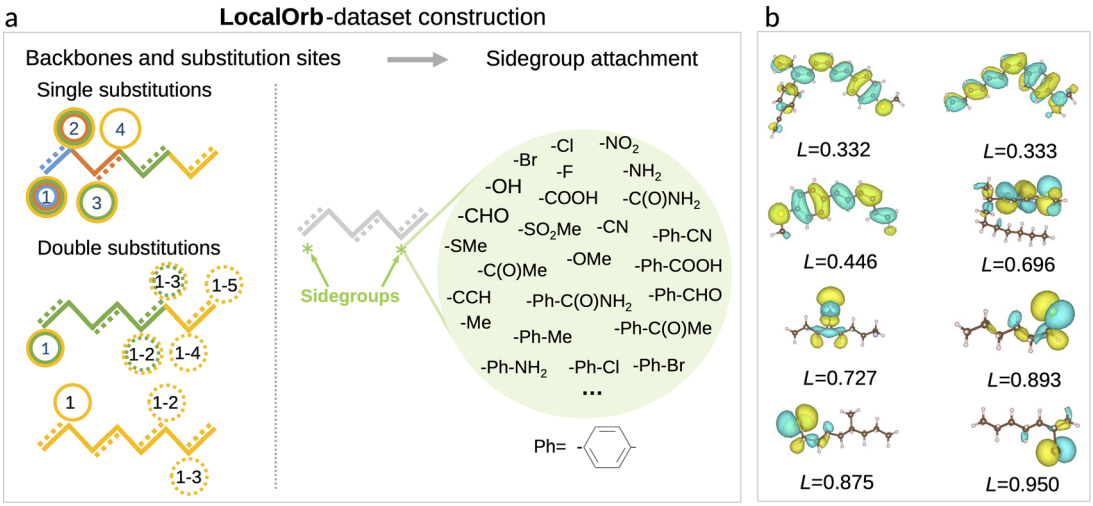



Fig. 2 LocalOrb dataset. (a) Illustration of the dataset construction principle, with alkane and conjugated alkene backbones of different length being decorated with one or two sidegroups. Note that only a representative subset of the 41 sidegroups is shown. Substitution sites are separated by at least three carbon atoms to avoid steric clashes. (b) Example molecules from the LocalOrb dataset with HOMO isosurfaces showing the diversity of localized and delocalized orbitals. This is quantified by the orbital localization index $L$, defined in the main text.

for orbital localization in a given molecule is needed. To this end, we can use the orbital localization fractions $l_i$ defined in eqn (6). Specifically, we define the orbital localization index $L$ as:

$$L = \sqrt{\max(\{l_1, \ldots, l_n\}) - \min(\{l_1, \ldots, l_n\})}. \quad (9)$$

If the HOMO is fully localized on a single atom this yields $L = 1$, whereas $L = 0$ if the HOMO is evenly distributed across all atoms.

While this definition is admittedly somewhat arbitrary, the metric matches our intuitive concept of localization and delocalization rather well, as shown in Fig. 2b. This also illustrates that the LocalOrb dataset indeed covers a highly diverse range of orbital distributions. Based on this we define highly localized orbitals as those with $L \geq 0.8$ and highly delocalized ones as those with $L < 0.4$.

# 3. Results

## 3.1 Pooling function performance

Fig. 3 collects learning curves for SchNet and SOAP based models using the pooling functions defined above. Here, subsets of the test set are shown, emphasizing molecules with particularly delocalized ($L < 0.4$, 3867 systems) and localized ($L \geq 0.8$, 539 systems) orbitals. Learning curves for the full test are shown in Fig. S5.† Directly at first glance this already reveals that localized orbitals are more challenging to predict, though this may be related to the fact that they are less frequent in the training set. Indeed, the performance for localized orbitals is quite sensitive to the number of localized configurations in the training set, as shown in Fig. S6.†

More importantly, the pooling functions are found to have a substantial influence on performance. In all cases, sum pooling displays very large errors. This underscores the importance of using properly intensive pooling functions when predicting orbital energies that has previously been reported.[18,24] Among the intensive pooling functions the differences are more subtle but still significant. Max pooling performs worst for delocalized systems with softmax being a slight improvement. Meanwhile, the commonly used average pooling tends to perform somewhat better than max and softmax for delocalized systems but worse for localized ones. This is basically in line with our expectations, since average and max are by construction suited for highly delocalized and highly localized orbitals, respectively. Though softmax should in principle represent a compromise between these extremes, it performs quite similarly to max in our tests.

To improve further, we turn to the more sophisticated weighted average approaches. As discussed in the Methods section, coefficient pooling represents a benchmark method in this context, as it incorporates exact information about orbital localization. We find that it indeed yields a significant improvement over average pooling and is among the best

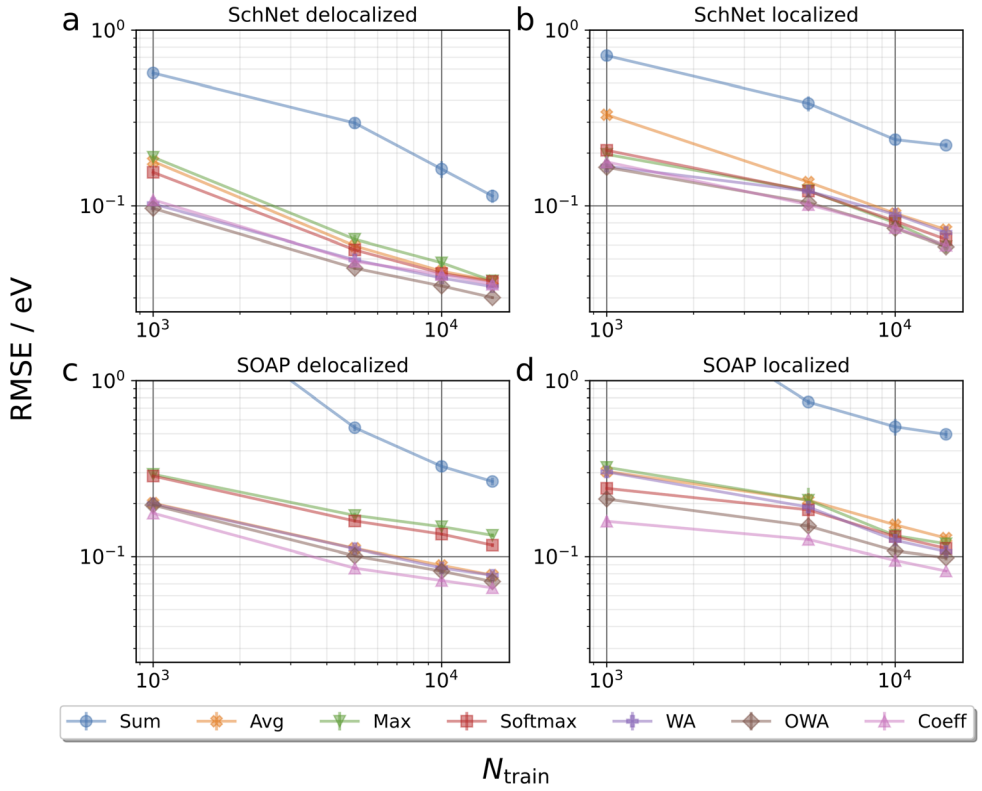



Fig. 3 Learning curves for HOMO energy prediction on LocalOrb. The root mean squared errors (RMSEs) of atomistic neural networks based on the SchNet and SOAP representations are shown for test set molecules with particularly delocalized or localized orbitals, as a function of the training set size $N_{\text{train}}$. Error bars indicate standard deviations over five randomly drawn training sets of the respective size. Note that the WA curve in frame c nearly overlaps with the Avg curve.

methods overall. Perhaps surprisingly, OWA pooling is even better in some cases, although it is formally designed to approximate coefficient pooling. To verify that the improved performance of OWA and WA is not merely due to the larger number of trainable parameters in the pooling function, additional SchNet results for average pooling models with increased embedding dimensions are shown in the ESI.† This reveals that simply increasing the capacity of the networks does not improve the test performance in this case.

As noted above, the OWA model predicts orbital localization with a second neural network, trained on the orbital fractions used in coefficient pooling. Its superior performance is likely due to the fact that both NNs in the model are trained using a joint loss function that depends both on the orbital locations and energies. Consequently, the model can in principle improve the predictive accuracy on energies by deviating from the reference orbital localizations. This additional flexibility is missing in the case of coefficient pooling.

Nevertheless, the orbital fractions provide an important inductive bias for the model. This is illustrated by the fact that WA pooling (which lacks this information) performs somewhat worse than both the OWA and coefficient pooling methods. Overall, OWA is found to be at least as accurate as the coefficient pooling benchmark and much more efficient from a computational perspective. It thus emerges as the pooling function of choice for localized intensive properties.

While not being the main focus of this paper, it is also interesting to compare the performance of the SchNet and SOAP based models. Overall, the SchNet models are found to be somewhat more accurate. This is in contrast to other benchmarks, *e.g.* for atomization energies, where SOAP-based models usually outperform SchNet (particularly for small training sets).[2] However, it should be emphasized that no hyperparameter optimization of the SOAP representation has been performed herein and that there is no reason to believe that the defaults we used are optimal for orbital energy prediction. A more detailed comparison of SchNet and SOAP is beyond the scope of this paper, however.

It is also notable that the spread among different pooling functions is somewhat larger for SOAP than for SchNet. This is likely due to the fact that the message passing mechanism in SchNet gives some additional flexibility to compensate inadequacies of the pooling functions. In particular, the scalar atomic quantities that are passed to the pooling function are much less local in SchNet than in SOAP. In other words, the message passing scheme performs some preliminary pooling among neighboring atoms. For conciseness we focus on the SchNet models in the following.

### 3.2 Predicting orbital locations

An added benefit of pooling functions like softmax, WA and OWA is that their weights can in principle be interpreted as approximate orbital localization fractions $l_i$. This is particularly pertinent for the OWA approach, where the weights should approximate $l_i$ by design. However, it is also interesting to consider if methods like softmax and WA implicitly learn to predict orbital locations when training on orbital energies alone.

To quantify this, Pearson correlation coefficients between the learned weights and the DFT-based $l_i$-values were calculated for all molecules in the test subsets used in Fig. 3. The corresponding histograms are shown in Fig. 4a. This confirms that OWA weights indeed represent excellent approximations to the true $l_i$-values, with all correlations being close to 1. The WA method also displays moderate to high correlations, in particular for localized states. In the delocalized case, the spread is somewhat larger but nearly all correlations lie above 0.5. Finally, the softmax method shows the weakest correlations and is particularly bad for the localized cases.

The high correlations between OWA weights and orbital distributions are also shown in Fig. 4b, where the weights are illustrated as semitransparent spheres forming phase-less pseudoorbitals. The OWA NN is thus a bona fide multi-property network that can be used to predict orbital energies and locations on the same footing, with potential applications for organic semiconductors.[50] The surprisingly good performance of WA in predicting orbital locations (particularly for localized orbitals) also underscores that $l_i$ is the right physical prior for the pooling function in this context. Even if they are not included in the training, the model indirectly (and imperfectly) infers them from the orbital energies.

### 3.3 Application to organic semiconductors

So far we have focused on the intentionally artificial LocalOrb set, which allowed us to study particularly localized and delocalized orbitals in depth. To test whether these insights are transferable to a real chemical application, we now turn to the OE62 dataset.[30] This set consists of >62 000 organic molecules extracted from crystal structures reported in the Cambridge Crystal Structure Database and was originally composed to screen for potential organic semiconductors.

This dataset is significantly more challenging than LocalOrb, with more structural diversity, a broader size distribution and more chemical elements. This is illustrated *via* a Kernel Principal Component Analysis plot in Fig. 5a.[36] Here, the LocalOrb set can be seen to cover a subset of the space covered by the OE62 set. Fig. 5b shows four representative molecules from OE62 and the corresponding HOMOs. This confirms that orbital localization is also an important aspect in real organic molecules. Note that since the original OE62 dataset lacks orbital coefficients, these were recomputed for this study (see ESI†).

Because the OE62 dataset has previously been used to train models for HOMO energy prediction, it also allows us to compare the methodology presented herein with the recent literature. To this end, SchNet models with average and OWA pooling were trained on randomly drawn training sets of 32 000 molecules. For robust statistics, this process was repeated ten times for each model and the performance was checked on an unseen test set of 10 000 molecules (see Fig. 5c). This procedure is analogous to the one used in ref. 51, with the best performing model from that paper (using Kernel Ridge Regression and the

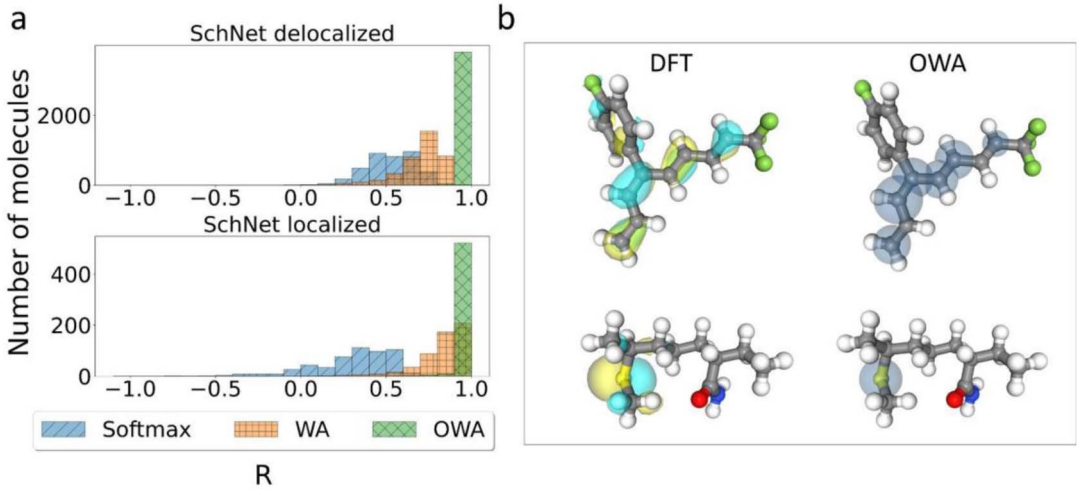



Fig. 4 Predicting Orbital Locations. (a) Pearson correlation coefficients $R$ between DFT-based orbital localization fractions $l_i$ and machine-learned weights obtained with different pooling functions. The two panels show correlations for particularly delocalized and localized systems, respectively. (b) Visual comparison of DFT orbitals and machine-learned pseudoorbitals obtained with the OWA approach. In the latter, learned weights are visualized as semitransparent spheres.

Many-Body Tensor Representation, MBTR[53]) also shown in Fig. 5c. Both the average and OWA models significantly outperform this baseline (RMSE = 0.24 eV) with RMSEs of 0.18 and 0.15 eV, respectively. Here, the improved performance of OWA is consistent with what we observed for the LocalOrb dataset. We also compare with two more recent graph neural network (GNN) based models from ref. 52, with RMSEs of 0.21 and 0.18 eV, respectively.

This shows that the OWA model displays state-of-the-art performance for HOMO energy prediction on OE62, while also providing orbital localization information, which the other models lack. Importantly, the benefits of the physically motivated OWA pooling function are not restricted to the artificial LocalOrb dataset, but also show up for the realistic and diverse molecules in the OE62 set. As shown in the ESI,† OWA outperforms average pooling across all molecule sizes in OE62, with the biggest improvement for the largest molecules. Overall,

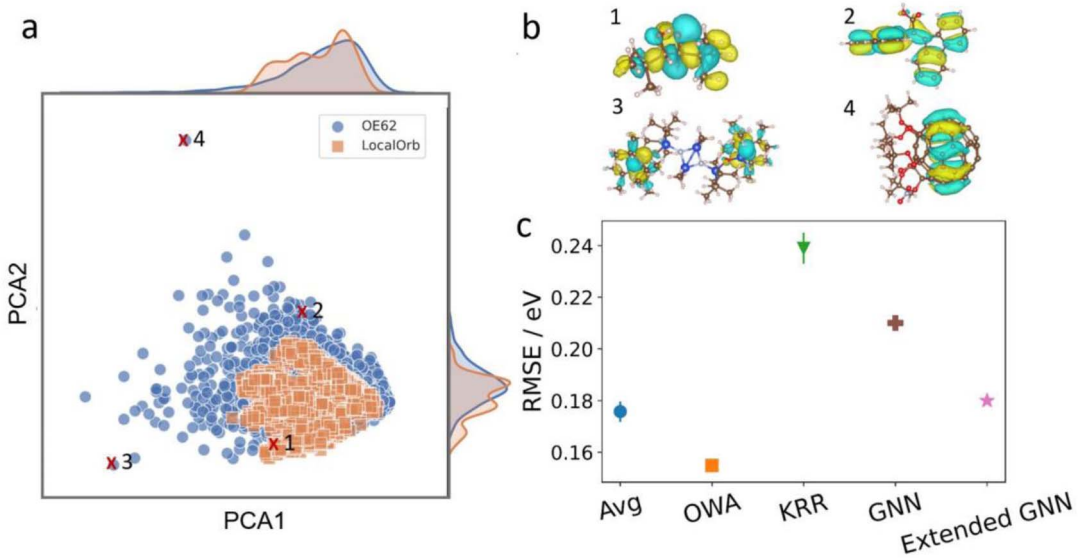



Fig. 5 Performance on the OE62 dataset. (a) SOAP-based Kernel principal component analysis plot showing 3000 randomly drawn molecules from the LocalOrb and OE62 datasets. This illustrates the significantly greater structural diversity of OE62. (b) Example molecules from OE62 with HOMO isosurfaces showing different levels of localization. (c) RMSEs of SchNet models using average and OWA pooling compared with previously reported models using Kernel Ridge Regression (KRR),[51] and Graph Neural Networks (GNNs).[52] In all cases, 32 000 molecules were used for training, and 10 000 molecules were used as a test set. Where shown, error bars reflect standard deviations over ten randomly drawn training sets.

OWA can thus be recommended as a robust and physically motivated pooling function for orbital energy prediction.

It should be noted that a series of other orbital energy prediction models have been proposed in the literature, which cannot directly be compared to these results. Most notably, several models were developed to predict machine-learned Hamiltonians, which yield both orbital energies and coefficients upon diagonalization.[20,38,54] These often focus on a range of occupied and unoccupied orbitals at once, so that they usually do not report HOMO prediction accuracies alone, even when they are tested on OE62.[20]

ML Hamiltonians in many ways are the most physically sound approach to predicting orbital energies and other intensive electronic properties. However, they also represent a significant computational overhead compared to OWA. In particular, their inference costs do not scale linearly with system size, due to the required diagonalization step. To overcome this, ref. 20 uses a constant-size ML Hamiltonian. Here, the correct treatment of isolated supersystems is not guaranteed, however. In our view, pooling functions like OWA therefore fill an important niche, providing physically sound and computationally efficient predictions of localized intensive properties.

## 4. Conclusions

In this contribution, the role of the final aggregation step in predicting localized intensive properties with atomistic neural networks was analyzed. Based on this analysis, a series of physically motivated pooling functions was proposed. To test these functions empirically, we generated the novel LocalOrb dataset, consisting of organic molecules with highly diverse orbital distributions. In this context, the OWA approach, which relies on predicting orbital locations along with their energies was found to be an optimal choice.

The physics-based approach proposed herein has two main advantages over purely data-driven ones. Firstly, it is useful whenever information about the localization of a property is of interest. This is, *e.g.*, the case when modelling organic semiconductors, where orbital locations are relevant for predicting electronic couplings between molecules.[55] Secondly, rigorously enforcing correct scaling with system size is essential whenever a ML model should be trained on small systems and applied to larger ones, *e.g.* to molecular clusters, crystals or polymers.

More broadly, the current study shows that a physical analysis of the target property based on interesting edge cases like non-interacting subsystems pays real dividends in chemical machine learning. We expect that combining these insights with recent advances in neural network architectures (*e.g.* the NequIP,[56] GemNet,[57] or MACE[58] models) can lead to further improvement in predicting orbital or ionization energies for complex systems.

Finally, the scope of localized intensive properties is in principle much wider than orbital energies and the related quantities discussed herein. For example, defect formation energies, catalytic activities or drug binding affinities display similar characteristics. In future work, we aim to generalize the findings of this study in these directions. In this context, it should be emphasized that localization is a property specific concept. Multi-property networks will thus require multiple weight networks. Furthermore, physical reference values for localization are not always as straightforward to define.

## Data availability

Data and code for this paper are publicly available at **https://gitlab.mpcdf.mpg.de/kchen/localized-intensive-property-prediciton.git**.

## Author contributions

This project was conceptualized by K. C. and J. T. M. K. C. implemented the concept and conducted the corresponding calculations. Methodological details were worked out by K. C., C. K., B. Q. C., and J. T. M. K. C., K. R. and J. T. M. wrote the manuscript. All authors discussed and revised the manuscript.

## Conflicts of interest

The authors declare no competing financial interests.

## Acknowledgements

KC acknowledges funding from the China Scholarship Council. KC is grateful for the TUM graduate school finance support to visit Bingqing Cheng's group in IST for two months. We also thankfully acknowledge computational resources provided by the MPCDF Supercomputing Centre.

## References

1 J. Behler, Four generations of high-dimensional neural network potentials, *Chem. Rev.*, 2021, **121**, 10037–10072.
2 V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti and G. Csányi, Gaussian process regression for materials and molecules, *Chem. Rev.*, 2021, **121**, 10073–10141.
3 N. Fedik, R. Zubatyuk, M. Kulichenko, N. Lubbers, J. S. Smith, B. Nebgen, R. Messerly, Y. W. Li, A. I. Boldyrev, K. Barros, *et al.*, Extending machine learning beyond interatomic potentials for predicting molecular properties, *Nat. Rev. Chem.*, 2022, **6**, 653–672.
4 M. Staszak, K. Staszak, K. Wieszczycka, A. Bajek, K. Roszkowski and B. Tylkowski, Machine learning in drug design: Use of artificial intelligence to explore the chemical structure–biological activity relationship, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2022, **12**, e1568.
5 J. Margraf, Science-driven atomistic machine learning, *Angew. Chem., Int. Ed.*, 2023, e202219170.
6 P. Reiser, M. Neubert, A. Eberhard, L. Torresi, C. Zhou, C. Shao, H. Metni, C. van Hoesel, H. Schopmans, T. Sommer, *et al.*, Graph neural networks for materials science and chemistry, *Commun. Mater.*, 2022, **3**, 1–18.

7 W. P. Walters and R. Barzilay, Applications of deep learning in molecule generation and molecular property prediction, *Acc. Chem. Res.*, 2020, **54**, 263–270.
8 P. Reiser, M. Konrad, A. Fediai, S. Léon, W. Wenzel and P. Friederich, Analyzing dynamical disorder for charge transport in organic semiconductors via machine learning, *J. Chem. Theory Comput.*, 2021, **17**, 3750–3759.
9 T. Morawietz, A. Singraber, C. Dellago and J. Behler, How van der waals interactions determine the unique properties of water, *Proc. Natl. Acad. Sci. U. S. A.*, 2016, **113**, 8368–8373.
10 B. Cheng, G. Mazzola, C. J. Pickard and M. Ceriotti, Evidence for supercritical behaviour of high-pressure liquid hydrogen, *Nature*, 2020, **585**, 217–220.
11 V. L. Deringer, N. Bernstein, G. Csányi, C. Ben mahmoud, M. Ceriotti, M. Wilson, D. A. Drabold and S. R. Elliott, Origins of structural and electronic transitions in disordered silicon, *Nature*, 2021, **589**, 59–64.
12 V. Kapil, C. Schran, A. Zen, J. Chen, C. J. Pickard and A. Michaelides, The first-principles phase diagram of monolayer nanoconfined water, *Nature*, 2022, **609**, 512–516.
13 S. Stocker, G. Csányi, K. Reuter and J. T. Margraf, Machine learning in chemical reaction space, *Nat. Commun.*, 2020, **11**, 227.
14 S. Stocker, J. Gasteiger, F. Becker, S. Günnemann and J. T. Margraf, How robust are modern graph neural network potentials in long and hot molecular dynamics simulations?, *Mach. Learn.: Sci. Technol.*, 2022, **3**, 045010.
15 J. Behler and M. Parrinello, Generalized neural-network representation of high-dimensional potential-energy surfaces, *Phys. Rev. Lett.*, 2007, **98**, 146401.
16 K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko and K.-R. Müller, SchNet–A deep learning architecture for molecules and materials, *J. Chem. Phys.*, 2018, **148**, 241722.
17 N. Lubbers, J. S. Smith and K. Barros, Hierarchical modeling of molecular energies using a deep neural network, *J. Chem. Phys.*, 2018, **148**, 241715.
18 W. Pronobis, K. T. Schütt, A. Tkatchenko and K.-R. Müller, Capturing intensive and extensive DFT/TDDFT molecular properties with machine learning, *Eur. Phys. J. B*, 2018, **91**, 178.
19 A. E. Sifain, L. Lystrom, R. A. Messerly, J. S. Smith, B. Nebgen, K. Barros, S. Tretiak, N. Lubbers and B. J. Gifford, Predicting phosphorescence energies and inferring wavefunction localization with machine learning, *Chem. Sci.*, 2021, **12**, 10207–10217.
20 J. Westermayr and R. J. Maurer, Physically inspired deep learning of molecular excitations and photoemission spectra, *Chem. Sci.*, 2021, **12**, 10755–10764.
21 R. Zubatyuk, J. S. Smith, B. T. Nebgen, S. Tretiak and O. Isayev, Teaching a neural network to attach and detach electrons from molecules, *Nat. Commun.*, 2021, **12**, 4870.
22 D. Grattarola, D. Zambon, F. M. Bianchi and C. Alippi, Understanding pooling in graph neural networks, *IEEE Trans. Neural Netw. Learn. Syst.*, 2022, 1–11.
23 A. Zafar, M. Aamir, N. Mohd Nawi, A. Arshad, S. Riaz, A. Alruban, A. K. Dutta and S. Almotairi, A comparison of pooling methods for convolutional neural networks, *Appl. Sci.*, 2022, **12**, 8643.
24 A. M. Schweidtmann, J. G. Rittig, J. M. Weber, M. Grohe, M. Dahmen, K. Leonhard and A. Mitsos, Physical pooling functions in graph neural networks for molecular property prediction, *Comput. Chem. Eng.*, 2023, **172**, 108202.
25 H. Jung, S. Stocker, C. Kunkel, H. Oberhofer, B. Han, K. Reuter and J. T. Margraf, Size-extensive molecular machine learning with global representations, *ChemSystemsChem*, 2020, **2**, e1900052.
26 H. Qiu, T. Xu, Z. Wang, W. Ren, H. Nan, Z. Ni, Q. Chen, S. Yuan, F. Miao, F. Song, *et al.*, Hopping transport through defect-induced localized states in molybdenum disulphide, *Nat. Commun.*, 2013, **4**, 2642.
27 M. Nolan, S. D. Elliott, J. S. Mulley, R. A. Bennett, M. Basham and P. Mulheran, Electronic structure of point defects in controlled self-doping of the $TiO_2$(110) surface: Combined photoemission spectroscopy and density functional theory study, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2008, **77**, 235424.
28 C. Wang, G. Zhou, H. Liu, J. Wu, Y. Qiu, B.-L. Gu and W. Duan, Chemical functionalization of carbon nanotubes by carboxyl groups on stone-wales defects: A density functional theory study, *J. Phys. Chem. B*, 2006, **110**, 10266–10271.
29 I. Kratochvílová, M. Vala, M. Weiter, M. Špérová, B. Schneider, O. Páv, J. Šebera, I. Rosenberg and V. Sychrovský, Charge transfer through DNA/DNA duplexes and DNA/RNA hybrids: Complex theoretical and experimental studies, *Biophys. Chem.*, 2013, **180**, 127–134.
30 A. Stuke, C. Kunkel, D. Golze, M. Todorović, J. T. Margraf, K. Reuter, P. Rinke and H. Oberhofer, Atomic structures and orbital energies of 61,489 crystal-forming organic molecules, *Sci. Data*, 2020, **7**, 58.
31 J. Behler, Atom-centered symmetry functions for constructing high-dimensional neural network potentials, *J. Chem. Phys.*, 2011, **134**, 074106.
32 A. P. Bartók, R. Kondor and G. Csányi, On representing chemical environments, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2013, **87**, 184115.
33 J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *Proceedings of the 34th International Conference on Machine Learning*, ed. D. Precup and Y. W. Teh, PMLR, Proceedings of Machine Learning Research, 2017, vol. 70, pp. 1263–1272.
34 K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller and A. Tkatchenko, Quantum-chemical insights from deep tensor neural networks, *Nat. Commun.*, 2017, **8**, 190.
35 L. Himanen, M. O. Jäger, E. V. Morooka, F. F. Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke and A. S. Foster, Dscribe: library of descriptors for machine learning in materials science, *Comput. Phys. Commun.*, 2020, **247**, 106949.
36 B. Cheng, R.-R. Griffiths, S. Wengert, C. Kunkel, T. Stenczel, B. Zhu, V. L. Deringer, N. Bernstein, J. T. Margraf, K. Reuter

and G. Csanyi, Mapping materials and molecules, *Acc. Chem. Res.*, 2020, **53**, 1981–1991.

37 K. T. Schütt, P. Kessel, M. Gastegger, K. A. Nicoli, A. Tkatchenko and K.-R. Müller, SchNetPack: A deep learning toolbox for atomistic systems, *J. Chem. Theory Comput.*, 2019, **15**, 448–455.

38 T. Zubatiuk, B. Nebgen, N. Lubbers, J. S. Smith, R. Zubatyuk, G. Zhou, C. Koh, K. Barros, O. Isayev and S. Tretiak, Machine learned hückel theory: Interfacing physics and deep neural networks, *J. Chem. Phys.*, 2021, **154**, 244108.

39 D. Buterez, J. P. Janet, S. J. Kiddle, D. Oglic, and P. Liò, Graph neural networks with adaptive readouts, *arXiv*, 2022, DOI: **10.48550/arXiv.2211.04952**.

40 M. Koerstz, A. S. Christensen, K. V. Mikkelsen, M. B. Nielsen and J. H. Jensen, High throughput virtual screening of 230 billion molecular solar heat battery candidates, *PeerJ Phys. Chem.*, 2021, **3**, e16.

41 "*The RDKit: Open-Source Cheminformatics Software, version*2021.03.4, 2021, **http://www.rdkit.org**.

42 S. Riniker and G. A. Landrum, Better informed distance geometry: Using what we know to improve conformation generation, *J. Chem. Inf. Model.*, 2015, **55**, 2562–2574.

43 P. Pracht, F. Bohle and S. Grimme, Automated exploration of the low-energy chemical space with fast quantum chemical methods, *Phys. Chem. Chem. Phys.*, 2020, **22**, 7169–7192.

44 C. Bannwarth, S. Ehlert and S. Grimme, GFN2-xTB-an accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions, *J. Chem. Theory Comput.*, 2019, **15**, 1652–1671.

45 S. Grimme, A. Hansen, S. Ehlert and J.-M. Mewes, r2SCAN-3c: A "swiss army knife" composite electronic-structure method, *J. Chem. Phys.*, 2021, **154**, 064103.

46 F. Neese, F. Wennmohs, U. Becker and C. Riplinger, The orca quantum chemistry program package, *J. Chem. Phys.*, 2020, **152**, 224108.

47 J.-D. Chai and M. Head-Gordon, Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections, *Phys. Chem. Chem. Phys.*, 2008, **10**, 6615–6620.

48 F. Weigend and R. Ahlrichs, Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy, *Phys. Chem. Chem. Phys.*, 2005, **7**, 3297–3305.

49 K. Chen, C. Kunkel, K. Reuter and J. T. Margraf, Reorganization energies of flexible organic molecules as a challenging target for machine learning enhanced virtual screening, *Digit. Discov.*, 2022, **1**, 147–157.

50 P. Friederich, A. Fediai, S. Kaiser, M. Konrad, N. Jung and W. Wenzel, Toward design of novel materials for organic electronics, *Adv. Mater.*, 2019, **31**, 1808256.

51 A. Stuke, M. Todorović, M. Rupp, C. Kunkel, K. Ghosh, L. Himanen and P. Rinke, Chemical diversity in molecular orbital energy predictions with kernel ridge regression, *J. Chem. Phys.*, 2019, **150**, 204121.

52 O. Rahaman and A. Gagliardi, Deep learning total energies and orbital energies of large organic molecules using hybridization of molecular fingerprints, *J. Chem. Inf. Model.*, 2020, **60**, 5971–5983.

53 H. Huo and M. Rupp, Unified representation of molecules and crystals for machine learning, *Mach. Learn.: Sci. Technol.*, 2022, **3**, 045017.

54 K. T. Schütt, M. Gastegger, A. Tkatchenko, K.-R. Müller and R. J. Maurer, Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions, *Nat. Commun.*, 2019, **10**, 5024.

55 H. Oberhofer, K. Reuter and J. Blumberger, Charge transport in molecular materials: an assessment of computational methods, *Chem. Rev.*, 2017, **117**, 10319–10357.

56 S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt and B. Kozinsky, E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials, *Nat. Commun.*, 2022, **13**, 1.

57 J. Gasteiger, F. Becker, and S. Günnemann, Gemnet: Universal directional graph neural networks for molecules, in *Adv Neural Inf Process*, 2021.

58 I. Batatia, D. P. Kovacs, G. N. C. Simm, C. Ortner, and G. Csanyi, "MACE: Higher order equivariant message passing neural networks for fast and accurate force fields," in *Advances in Neural Information Processing Systems*, ed. A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, 2022.