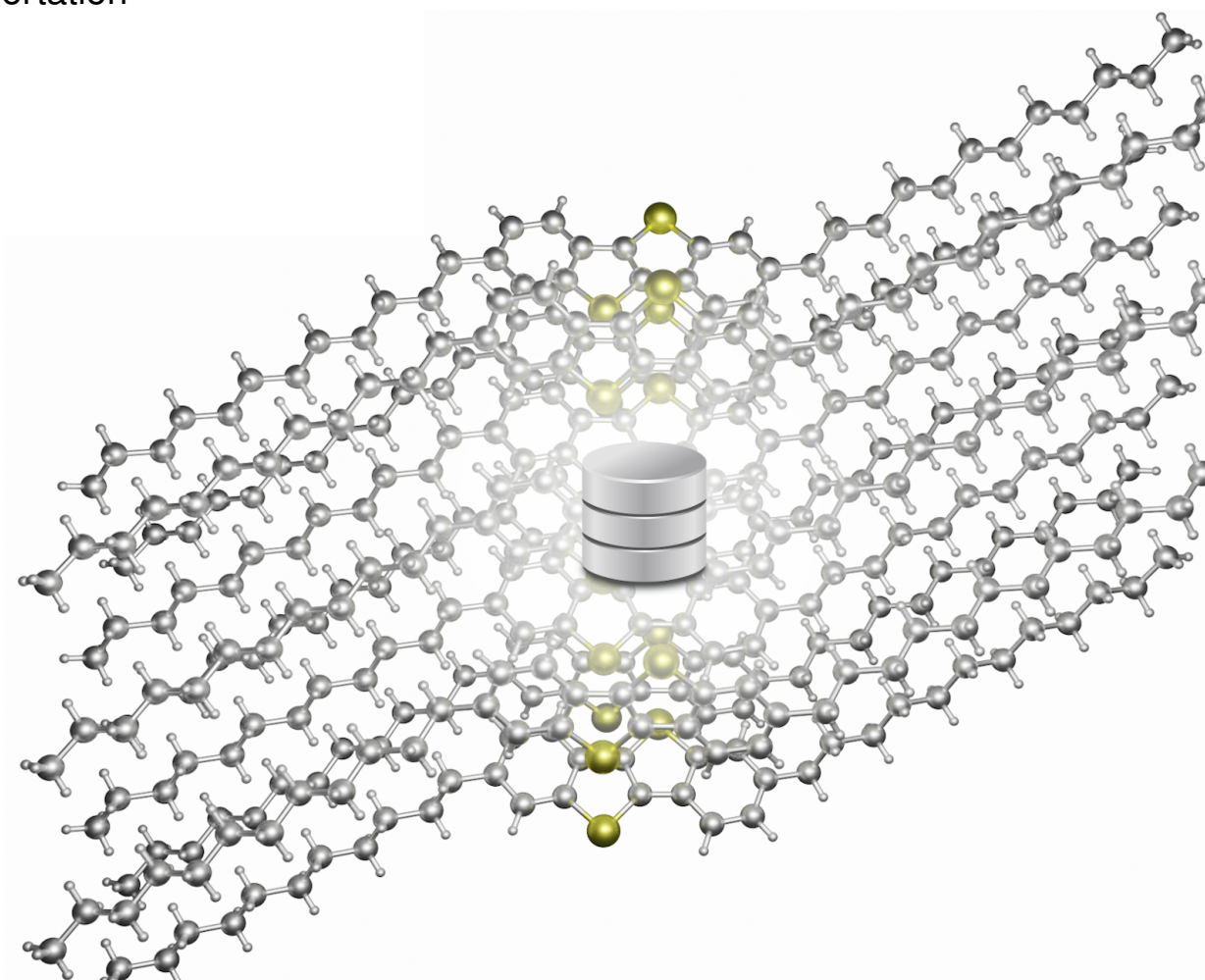
Technische Universität München
Lehrstuhl für Theoretische Chemie
Fakultät für Chemie

Data-driven Organic Semiconductor Discovery

Christian Kunkel

Dissertation



Technische Universität München
Fakultät für Chemie
Lehrstuhl für Theoretische Chemie

Data-driven Organic Semiconductor Discovery

Christian Kunkel

Vollständiger Abdruck der von der Fakultät für Chemie der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender: Priv.-Doz. Dr. Harald Oberhofer

Prüfer der Dissertation:

1. Prof. Dr. Karsten Reuter
2. Prof. Dr. Frank Ortmann
3. Prof. Dr. Egbert Zojer

Die Dissertation wurde am 21.04.2021 bei der Technischen Universität München eingereicht und durch die Fakultät für Chemie am 17.06.2021 angenommen.

Für meine Eltern und Großeltern,
die es alles möglich gemacht haben.

Preface

This dissertation is publication-based, meaning its scientific content is published in a series of related, but independent articles, all of which have undergone the scientific peer-review process in international scientific journals. The first chapters therefore mainly serve as an introduction to methods and relevant literature. Summaries for each article are then provided in chapter 6. The main part of the presented work has been carried out at the Chair of Theoretical Chemistry of the Technical University of Munich (TUM) between March 2017 and September 2020, under the supervision of Prof. Dr. Karsten Reuter and it has been completed between October 2020 and February 2021 at the Fritz Haber Institute of the Max Planck Society in Berlin. A research stay in May 2018 hosted by Prof. Dr. Patrick Rinke at Aalto University complemented this work.

Abstract

Organic electronics have a low ecological- and economic footprint and are versatile in their application. Progress in research and stronger commercialization have also raised high expectations for their continued market-success. Improving device parameters and materials properties such as electrical conductivity however remains important, so far usually tackled by laborious empirical structural tuning of a promising compound and device architecture. The advent of molecular machine learning and data-driven design techniques has lead to high hopes, as such methods can potentially enable a more efficient, computationally guided improvement of important OSC material properties. In this dissertation such strategies are explored in a series of related but independent publications. The first part of the dissertation was based on a previously established in-house dataset of > 64.000 organic small-molecule organic crystals – the 64k-dataset –, annotated with charge-transport related descriptors (electronic coupling and the reorganization energy) computed from first-principles. The virtual screening effort from which it had originated was able to recover known and well-performing materials, while it could also uncover many additional promising candidates. Building on this well-suited data-source, we first provide a more in-depth analysis of the encoded design space. To arrive at design principles we evaluated the relative performance of molecular scaffold and side group clusters occurring in the compounds, finding certain scaffolds and side groups to consistently improve charge-transport properties. Functionalizing promising scaffolds with favorable side groups can then result in molecular crystals with improved charge-transport properties. In a subsequent study, we further analyzed this design space by a chemical space network, whose visualization hints at already covered- as well as promising new regions of the design space. In a next step and in order to complement the workhorse-method of density functional theory (DFT), – still relatively expensive for these large-scale screening studies, efficient molecular machine learning (ML) methods were tested that can greatly accelerate the molecular design workflow in vast molecular spaces. In collaboration and starting from the 64k-dataset , the OE62-dataset was assembled and made public, allowing ML method development for the prediction of molecular electronic properties. Composed of large, technologically relevant molecules from a sparsely and unevenly sampled chemical space, the dataset is complementary to the QM9-dataset commonly used to assess the performance of new ML methods. While working on the dissertation, the new OE dataset was already used by our colleagues and us to test and extend the predictive capacity of common molecular ML methods to larger systems. In a last step, we employ molecular machine learning for the OSC design task, devising an active machine learning (AML) framework that explores an unlimited search space of π-conjugated molecules along consecutively applied molecular transformation operations. The dissertation thereby highlights the usefulness of data-based approaches for a targeted design of organic electronics materials, while further work should extend the approaches in scope and accuracy, as well as include additional important design parameters.

Zusammenfassung

Organische Elektronik ist ökonomisch und ökologisch attraktiv und vielseitig einsetzbar. Erzielte Fortschritte in der Forschung und zunehmende Kommerzialisierung lassen außerdem auf anhaltenden Erfolg hoffen. Bauteilparameter und Eigenschaften der aktiven Materialschichten wie elektrische Leitfähigkeit müssen jedoch weiter verbessert werden. Häufig geschieht dies bisher durch arbeitsintensive empirische (Struktur-) Verbesserung einer vielversprechenden Verbindungsklasse oder Bauteilarchitektur. Das Aufkommen von maschinellem Lernen (ML) und datengestütztem Design lässt jedoch erhoffen, dass eine computerunterstützte Verbesserung von wichtigen Materialeigenschaften möglich wird. In dieser Dissertation wird die Anwendbarkeit entsprechender Strategien in einer Reihe zusammenhängender, aber eigenständiger Publikationen untersucht. Der erste Teil basiert dabei auf einer vorhandenen Datenbank von > 64.000 organischen molekularen Kristallen für die Deskriptoren für Ladungstransport durch ab-initio Simulationsmethoden berechnet wurden (elektronische Kopplungselemente und Reorganisationsenergien) – der 64k-Datensatz. Das virtuelle Datenbankscreening aus dem der Datensatz hervorging konnte dabei bereits bekannte und leistungsfähige Materialien wiederfinden, während eine Vielzahl weiterer vielversprechende Kandidaten entdeckt wurde. Aufbauend auf dieser gut geeigneten Datenquelle geben wir im ersten Schritt dieser Dissertation einen Überblick über den Designraum organischer Halbleiterkristalle. Um Designregeln abzuleiten, werteten wir die Eignung von molekularen Gerüsten (scaffolds) und Seitengruppen aus. Dabei zeigte sich, dass bestimmte molekulare Gerüste und Seitengruppen die Ladungstransporteigenschaften beständig verbessern. Die Kombination von vielversprechenden Gerüsten mit vorteilhaften Seitengruppen kann dann zu molekularen Kristallen mit verbesserten Ladungstransporteigenschaften führen. Darauf aufbauend analysierten wir den Designraum mithilfe eines "Chemical Space Network", dessen Visualisierung auf bereits untersuchte sowie auf neue vielversprechende Regionen im Designraum hinweist. In einem weiteren Schritt und um die bis dahin verwendete wichtigste Simulationsmethode "Dichtefunktionaltheorie" (DFT) zu erweitern –die für große Screening-Studien viel Rechenzeit benötigt– wurden effiziente Methoden des maschinellen Lernens für die Anwendung an Molekülen getestet, welche den Workflow des molekularen Designs in umfangreichen Designräumen signifikant beschleunigen können. In Kooperation und ausgehend vom 64k-Datensatz bauten wir daher den OE62-Datensatz auf. Dieser publizierte Datensatz erlaubt die Entwicklung von ML-Methoden zur Vorhersage von molekularen elektronischen Eigenschaften. Zusammengesetzt aus großen, technologisch relevanten Molekülen aus einem ungleichmäßig abgedeckten chemischen Raum ist dieser Datensatz komplementär zum QM9-Datensatz, welcher häufig genutzt wird, um die Performance neuer ML-Methoden zu bewerten. Der neue Datensatz wurde im Laufe der Dissertation bereits von unseren Kollegen, sowie von uns verwendet, um die Anwendbarkeit von molekularem ML für größere Moleküle zu testen und zu erweitern. Im letzten Schritt wenden wir molekulares ML für OSC-Design an. Wir verwenden dabei aktives (maschinelles) Lernen (AML) um in einem virtuell unbegrenzten Suchraum der durch konsekutive Anwendung von molekularen Transformationsregeln erzeugt wird nach vorteilhaften π-konjugierten Molekülen zu suchen. Insgesamt hebt die Dissertation damit die Nützlichkeit von datenbasierten Ansätzen für ein gezieltes Design von organischen elektronischen Materialien hervor. Weitere Arbeiten sollten diese Ansätze in Umfang und Genauigkeit erweitern und durch zusätzliche Kenngrössen ergänzen.

Abbreviations

OSC	Organic semiconductor
HOMO	Highest occupied molecular orbital
LUMO	Lowest unoccupied molecular orbital
IP	Ionization potential
EA	Electron affinity
CB	Conduction band
VB	Valence band
HTL	Hole transport layer
HTL	Electron transport layer
EML	Emitting material layer
OLED	Organic light-emitting diode
OPV	Organic photovoltaic (device)
OFET	Organic field-effect transistor
AML	Active machine learning
ML	Machine learning
SOAP	Smooth overlap of atomic positions
DFT	Density functional theory
FO-DFT	Fragment-orbital DFT
KRR	Kernel Ridge Regression
GPR	Gaussian Process Regression
CM	Coulomb matrix
MBTR	Many-body tensor representation
AI	Artificial Intelligence
QML	Quantum machine learning
SMILES	Simplified Molecular-Input Line-Entry System
CSD	Cambridge Structural Database
CSN	Chemical space network
BM	Bemis-murcko (scaffold)
PCA	Principal component analysis
KPCA	Kernel PCA
TST	Transition state theory

Contents

Preface v

Abstract vii

Zusammenfassung ix

Abbreviations xi

1 Introduction 1

2 Theoretical models for charge conductivity in organic semiconductors 5
2.1 Organic semiconductor devices 5
2.2 Conductivity in OSC materials 7
2.3 Charge carrier injection 8
2.4 Models for charge mobility 9

3 The 64k- and OE62-datasets 17

4 Molecular analysis and machine learning 21

5 Visualization of high-dimensional chemical space 27

6 Publications 31
6.1 Finding the Right Bricks for Molecular Legos: A Data Mining Approach to Organic Semiconductor Design 32
6.2 Knowledge discovery through chemical space networks: the case of organic electronics 33
6.3 Atomic structures and orbital energies of 61,489 crystal-forming organic molecules 34
6.4 Mapping materials and molecules 36
6.5 Active Discovery of Organic Semiconductors 38
6.6 Further work 39

7 Conclusion and Outlook 41

8 Acknowledgments / Danksagung 43

Bibliography 45

1 | Introduction

Inorganic semiconducting materials (such as silicon or GaAs) are an empowering factor for whole industries, deeply rooted in the information technology- or solar energy sectors. This is not surprising, as the devices made from such materials often display high performance. They can however turn out to be too brittle or expensive for certain applications.[1] Considerable technological interest in new materials therefore exists, and these could extend the application-spectrum of electronic devices by the novel properties they potentially bring along.

Organic semiconductors (OSCs)[2] are an interesting class of materials, displaying clear-cut benefits such as a low ecological- and economic footprint, while allowing for a straightforward production of transparent, lightweight, and mechanically flexible large-area devices, that are adaptable to a customers' needs (see Figure for an example 1.1). Device performances achieved over the last decades have thereby already led to some commercially successful applications, such as the routine use of organic light-emitting diodes (OLEDs) in high-resolution self-luminous displays[3] which created a multi-billion dollar industry.[4] Organic photovoltaic devices[5] have followed and further innovative products are expected to emerge, such as lasers,[6] organic radio frequency identification devices (RFID),[7] nanoscale memory- or sensing devices,[8,9] possibly embedded in smart textiles[10] or used on skin.[11]

Obstacles for widespread use of OSC materials however remain, such as their limited longevity or low electrical conductivity.[15] This is true for polymers and small molecular materials alike, both of which have been intensively researched in the OSC community. In fact, a smaller number of ordered, crystalline materials made from small molecules now reproducibly yield high charge carrier mobilities and conductivities,[13,16–20] see examples in Figure 1.2. The discovery of further improved molecular materials can however take years of intensive research, involving labor-intensive cycles of iterative improvement.

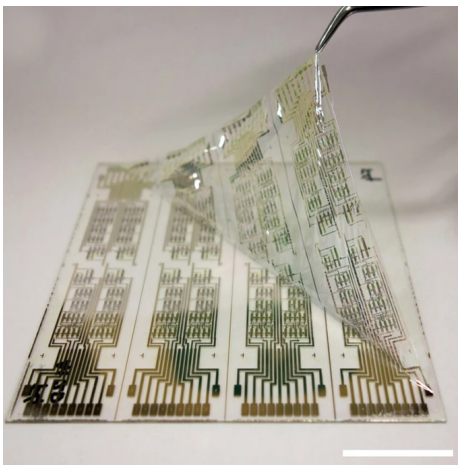


Figure 1.1 Organic CMOS logic circuit with a total thickness of less than 3 m. Scale bar: 25 mm. Reproduced from reference 12 under a Creative Commons Attribution 4.0 International License.

These discovery campaigns are often based on a selected molecular family, trying to identify the best-performing candidate based on small modifications. Such efforts are then usually guided by experimental results, relying also on empirically derived knowledge,[16,21] and chemical intuition. Moving beyond such local exploration could well-extend the number and versatility of known materials. In fact, the well-performing materials known to date might be only the tip of the iceberg, as the following argument illustrates. Even from a few building blocks that regularly occur in typical π-conjugated molecular systems,

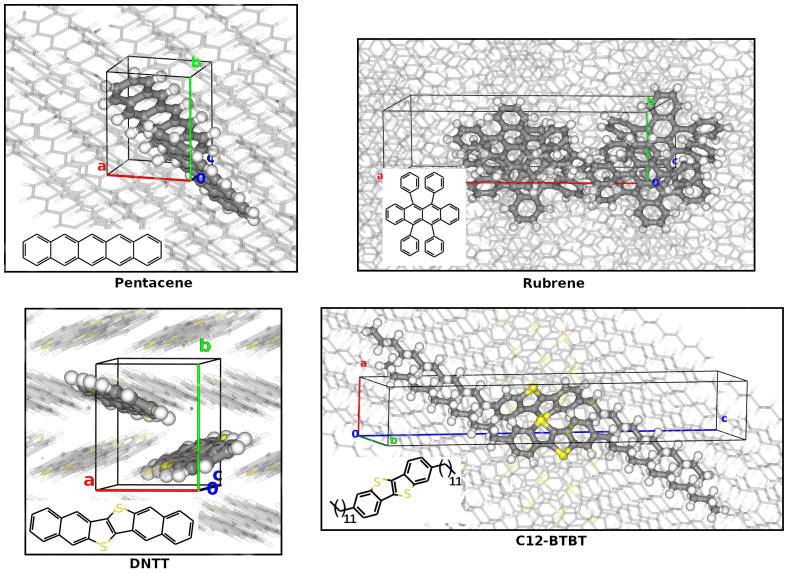


Figure 1.2 Examples of crystalline molecular materials exhibiting state-of-the-art charge-carrier mobilities.[13] Constituent molecules are usually made of extended π-conjugated systems, allowing for the (partial) accommodation of charge-carriers. The exhibited crystalline arrangements originate from an intricate balance of intermolecular interactions (e.g. van-der-Waals, exchange- correlation and electrostatic) among neighboring molecules, dictating also the charge-transport properties of the structure. The displayed experimentally resolved crystals are PENCEN02, QQQCIG04, NICLAN and PIVBAY as identified by their Cambridge Structural Database[14] (CSD) reference codes.

a vast number of new candidates can already be obtained – a combinatorial explosion. An estimate for small molecules of molecular weight < 500 Da that are composed of the most relevant elements in organic chemistry (C, N, O, S and halogens) reaches 10^{33}.[22] Adding to that, molecular properties can be highly sensitive to small changes[23,24] and are consequently also highly tuneable. It is therefore expectable that hitherto unknown, but highly favorable materials can be found in such vast materials spaces. Bespoke vastness however also makes molecular discovery by data-efficient search strategies crucial.

Hope thereby rests on efficient, data-driven materials design methods,[25] which are the topic of this dissertation. As a matter of fact, while the available theories for charge-conductivity (see chapter 2) already early on helped to explain performance trends in experimentally characterized OSC materials, they can also guide an in-silico discovery process based on the computable descriptors they expose. In this spirit, and in limited materials spaces, this approach led to the successful and experimentally verified discovery of a high-performance OSC material,[26,27] while subsequent virtual discovery efforts have more recently been scaled to larger molecular databases,[23,28–35] see also.[36–38]

A study by Schober et al.[30,39] is representative of this development and played a central role at the outset of this dissertation. Being among the first large-scale studies that screened a large database for potential OSC materials, it recovered many known and well-performing materials, while also uncovering many more promising ones, not yet considered for organic electronics applications. Based on the resulting 64k-dataset of $>$ 64.000 experimentally known organic molecular crystals annotated with computed charge-transport descriptors (see chapter 3), we in this dissertation first applied methods of knowledge-discovery. We thereby arrive at general principles for a "molecular Lego" design approach[40] as well as at an intuitive visualization of the available data.[41] The corresponding methods are introduced mainly in chapters, 4 and 5, while the idea is illustrated in Figure 1.3.

Such exhaustive materials screening approaches however relied on the expensive computational property evaluation for a large number of materials to produce the underlying data. We, therefore, researched machine-learning methods, which allow for the cheap interpolation of relevant molecular properties among thousands of candidate materials, see chapter 4. For this purpose, the methods simply establish surrogate models of the underlying physical relationships, learning from data-sets that are representative of the problem. Such methods were first established in a joint effort with colleagues from Aalto University, starting from the 64k-dataset and leading up to the new OE62-dataset[42] (see chapter 3), a specialized and challenging benchmark for molecular machine-learning of quantum properties on technologically relevant molecules. Studies on its application with established and new models were also undertaken.[43,44]

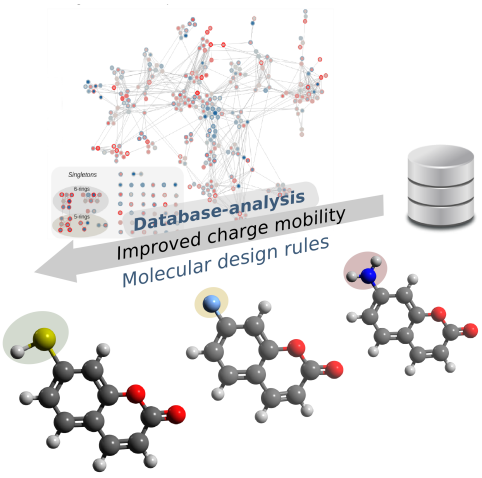


Figure 1.3 Idea of data-driven molecular design explained. By means of data-analysis or machine-learning, general molecular design rules are derived from existing data and subsequently used to propose new materials of high performance. Figure adapted from references 40 and 41 with permissions from American Chemical Society and the Springer Nature Customer Service Centre GmbH.

In further collaboration with colleagues from Oxford, Cambridge and Washington we then extended our capabilities, in the visualization of or materials design spaces, leading to an easy-to-read overview of the developing field[45] (see chapter 5). At last, we returned to the OSC design problem again, approaching it with the help of molecular machine learning and visualization. As mentioned, a drawback of the above-mentioned resource intensive virtual screening or data-driven methods had been the reliance on an exhaustive (staged) screening. We thus employed an active machine-learning (AML) feedback loop, training a machine learning model on available data, while using it to guide the search to the next prospects. Feedback from these subsequently accumulating computational results can then be used for an ever better-informed selection strategy and discovery success (see chapter 4). Since the dissertation is publication based, all results have been published in peer-reviewed articles. The following chapters therefore discuss the most important concepts necessary for a broader understanding of this line of research, while summaries of the single articles are provided in chapter 6.

2 | Theoretical models for charge conductivity in organic semiconductors

The dissertation is mainly concerned with the optimization of conductivity in small molecule crystalline organic semiconductors. This chapter thus provides the physical background for OSC molecular- and material optimization. It starts with an introduction of the main devices which employ OSCs as active charge transport layers. This introduction is followed by a review of theoretical models for charge conductivity. Due to their relevance to molecular design, a focus will thereby lie on the most decisive descriptors that these incorporate.

2.1 Organic semiconductor devices

Thin OSC films are often employed as active material layers of OLEDs (organic light emitting diodes), OPVs (organic photovoltaic devices) and OFETs (organic field effect transistors). Since excellent in-depth summaries of employed materials[5,17,18] and physical principles[2,46] are available, I will only provide an introductory overview of their operational principles, schematically depicted in Figure 2.1. OLEDs and OPVs will thereby be briefly introduced. The focus in this work however lies on computational discovery of organic materials with high electrical conductivity. In a direct way these are most relevant to OFETs, and hence these will be discussed in more detail.

OLEDs[47,48] produce photon emission (light) upon application of an electric current. These photons are internally created through recombination of electrons and holes under radiative (fluorescence or phosphorescence) decay. In these often multi-layered devices, the latter process usually takes place in an emitting material layer (EML) with suitable luminescent properties. This layer is usually sandwiched between two additional organic layers with high and balanced charge mobility, which transport injected holes and electrons from anode and cathode to the EML. An

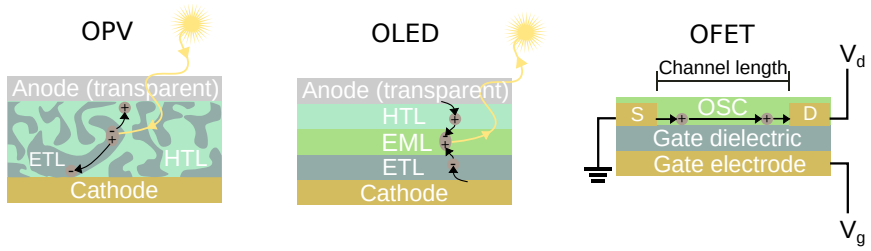


Figure 2.1 OSC-based device architectures discussed in this chapter. For the OFET device a bottom gate, bottom contact architecture is shown, while different variants exist. For the OPV device, the process of photon absorption takes place in a bulk-heterojunction (blend of donor and acceptor materials).

efficient injection of charge-carriers, as well as a balanced performance of these latter hole- and electron transport layers (HTL and ETL) is hence necessary to allow for equally efficient transport to and subsequent radiative decay in the EML. Note in this context, that blocking layers that prevent injected charges from leaving the EML, or hole or electron injection layers are here omitted for brevity. For light to be able to escape from such a device, one of the electrodes needs to be transparent for the emitting photon wavelength. Indium tin oxide (ITO) is thereby often used for this purpose.

By reversing the process, **OPVs**[49–51] absorb light in the form of photons, converting them to electrical energy. In devices, a hole and an electron conducting layer are brought in contact, and form an interface. Upon illumination, excitons –Coulomb-bound, electrically neutral electron-hole pairs– form in the organic layers, subsequently diffusing through the material. Exciton diffusion to and dissociation at the material interface then leads to separate holes and electrons. Once separated, they are transported to the respective electrodes, generating a current. The efficiency of this latter step thereby crucially depends on the mobility in the respective organic layers. While details are beyond the scope, it is worth mentioning an important breakthrough for OPVs –the bulk heterojunction. A mixed phase of hole- and electron conducting semiconductors thereby enhances the interfacial area, positively influencing the efficiency of charge-separation and subsequent transport to the electrodes.

Apart from these optoelectronic devices, classic electronic devices such as transistors can be made with organic materials as well. Such **OFETs**[1,15,52,53] can also be used to make complex logical circuits of electronic devices, i.e. to drive the operation of OLED displays.[1] In an OFET, two electrodes (source and drain) are directly connected through a thin OSC layer (the channel). A third electrode (gate) is located close to the channel as well but is kept in spatial separation from it by an insulating dielectric layer – forming a capacitor with it. A controlled current between the source- and the drain electrode can then flow when applying potentials at the gate and the drain electrodes simultaneously: The potential at the gate electrode thereby controls the flow of current by the induction of charge carriers in the OSC layer ("field-effect doping"). The charge carriers can then be transported through the OSC layer along the direction of source to drain potential difference. In this way, the potential applied at the gate electrode acts as a switch that can be used to allow for, or amplify an electric signal through the device. For OFETs a high charge carrier mobility in the OSC layer is crucially affecting device performance, i.e. lowering the response times as well as the signal-to-noise ratio of the OFET and thereby determining the frequency and accuracy with which the logic circuit can be operated. Achieving high charge conductivities in the OSC active layers thus allows for faster switching times, leading to reduced calculation times and energy savings. A more fundamental application of the OFET architecture is in the experimental evaluation of a materials charge carrier mobility, care must be exerted to arrive at reproducible results.[54] Nevertheless, OFETs achieve the highest mobilities when made from single crystal OSC layers, linked to their high chemical purity and high structural order (in particular the absence of grain boundaries).

Technological interest for single-crystal applications is thus considerable.[19,55,56] In the herein described materials discovery efforts related to the 64k-dataset , we hence mainly focused on improved crystalline materials applicable e.g. in OFET devices. The use case is however not exclusive, as OPV and OLED devices could also benefit from improved charge-transport characteristics, additionally considering their optical properties. On the other hand, the molecular properties we assess are not exclusively relevant for charge-transport in molecular crystals, but also in amorphous films, albeit charge-mobility is here often by an order of magnitude lower.[57] While we mainly focus on the computational discovery of favorable small molecules tailored for an envisioned electronic OSC application, other factors that can significantly influence performance should be mentioned, such as method and conditions of material deposition (i.e. spin-coating, printing from solution, or vapor deposition), which influence the degree of order present in the de-

posited amorphous or crystalline material film. This also includes structural defects and impurities often present in the real material layers.

2.2 Conductivity in OSC materials

The presence and efficient migration of charge carriers through an organic material layer is important for the operation of the introduced devices. The theoretical modeling of this process will therefore here be reviewed, focussing on the descriptors that mostly influence charge carrier conductivity σ in such material layers. The literature on this topic is however vast and numerous models are available. I will hence provide a perspective and reference the most important developments pertaining to the goals of this dissertation.

By Ohms law, σ relates the voltage V applied to a material with the electrical current I flowing through it as

$$I = \sigma V \tag{2.1}$$

In a simplified picture, σ can be decomposed into charge mobility, μ and charge carrier concentration ρ as

$$\sigma = \mathrm{e}\rho\mu \tag{2.2}$$

with an elementary charge e. It should be noted, that conductivity or mobility in organic crystals are generally anisotropic,[58] from here on illustrated by the use of rank-2 tensors (bold, underlined).

$\underline{\boldsymbol{\mu}}$ thereby characterizes the intrinsic ability of a charge carrier to move in the bulk of an organic layer, making it an essential parameter for a material's performance in its current morphology. Different theoretical models for $\underline{\boldsymbol{\mu}}$ exist, related to a variety of proposed and observed mechanisms and transport regimes, see below. Charge carriers in OSC materials on the other hand can arise from different sources, in total giving rise to a charge carrier concentration ρ. Here, a significant difference between OSCs and inorganic semiconductors becomes evident. In inorganic semiconductors, mobile charge carriers are intrinsically available at room temperature, simply due to a small gap (usually < 1.5 eV) between the valence- and conduction band which can be overcome by thermal excitation. Common OSC materials on the other hand show a wide band gap (> 2 eV), rendering such intrinsic charge carriers a minority species. Extrinsic sources for charge carriers such as (unintentional) doping,[59,60] photogeneration by absorption of light or especially efficient charge injection at electrodes (see below) are thus important.

To tackle the materials design task for high conductivity $\underline{\boldsymbol{\sigma}}$ one can therefore target $\underline{\boldsymbol{\mu}}$ or ρ, and theoretical models for both quantities are described in more detail below. In this dissertation we thereby focused on the improvement of hole (p-type) conductivity of small molecule organic semiconductors,[17,61] mainly used in OFETs to date. Electron-conductive or ambipolar materials could however also be used[62] and the design problem could be tackled by simply changing to the respective electron-conduction related descriptors.

2.3 Charge carrier injection

In a simple Mott-Schottky model, the efficient injection of holes (electrons) from a metal electrode to an OSC layer is dependent on the energy difference between the work function Φ of the metallic electrode (energy to eject an electron into vacuum) and the respective molecular states in the organic layer, see Figure 2.2. Injection is thereby barrierless if Φ is perfectly aligned with the relevant molecular solid state ionization potential (IP) and electron affinity (EA). In practice, this is hardly achieved and an injection barrier arises from the energy mismatch.[63] It should however be mentioned that this simplified picture neglects interface dipole effects that can arise at the surface of electrodes, and can significantly shift respective energy levels and alter their relative alignment.[63,64] Interfacial geometric effects or charge redistribution can add to this effect, e.g. brought about upon chemical reaction between electrode and organic layer.[65] Contact doping can be used to lower barriers.[66]

Nevertheless, the simple guideline provided by the Mott-Schottky model can be used in a computational screening, choosing useful molecules that satisfy an Ohmic contact condition with a small energy mismatch, while those with an expected higher injection barrier are excluded. In devices, hole injecting electrodes such as indium tin oxide or gold are common, while calcium or aluminum electrodes are used to inject electrons. Experimental values for Φ are thereby tabulated[67] for different metals (and their specific surfaces). Solid-state IPs or EAs on the other hand can be gauged by DFT-methods employing e.g. tuned range-corrected functionals and implicit solvation models that mimic the solid state environment.[68,69] To avoid confusion, it should here be noted that bespoke solid state IPs/EAs differ significantly from values found for the isolated molecules in vacuum, simply due to polarization in the solid state. As an example, common IPs measured in vacuum are on the order of 8 eV, reduced by roughly 2 eV in the solid state.[69]

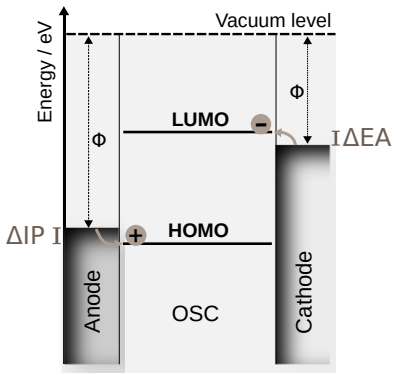


Figure 2.2 Schematic depiction of hole- and electron injection from metal electrodes into a single organic semiconductor layer. While overcoming a small Mott-Schottky barrier $\Delta\mathrm{IP}$ ($\Delta\mathrm{EA}$), holes (electrons) are injected to an energy level in the OSC layer, represented by the molecular HOMO (LUMO) level. Note, a common vacuum level is here assumed, while detailed discussions on vacuum levels at finite distance to the material surfaces can be found in references 63, 64. Figure adapted from Koehler.[2]

In the OSC field, it is however common practice to substitute IP/EA values by HOMO (LUMO) energies. On the one hand this is justified when assuming an effective one-electron picture in which hole (electron) injection takes place into these energy levels. On the other hand, DFT calculations on molecules in vacuum and at the B3LYP level of theory[70–72] yield HOMO or LUMO energies that fortunately match well with experimental data of solid state IP or EA,[73] rendering this computational level a well established standard for this task.[31,34,58,74]

2.4 Models for charge mobility

Following Oberhofer et al.[75] charge mobility is defined as the response of a charge carrier's velocity $\mathbf{v}$ within the material to an applied electric field $\mathbf{F}$.

$$\mu_{ij} = \frac{\langle v \rangle_i}{F_j} \tag{2.3}$$

Alternatively a diffusion equation is often used, relating mobility to the charge carriers diffusion coefficient $\underline{\boldsymbol{D}}$ as

$$\mu_{ij} = \frac{qD_{ij}}{k_b T} \tag{2.4}$$

with k_b and T denoting Boltzmann-constant and temperature, respectively. This latter equation is also known as the Einstein-Smoluchowski equation often used to treat a purely diffusive mobility in the limit of a vanishing electric field.

A larger number of theories estimate $\underline{\boldsymbol{D}}$ or $\mathbf{v}$, usually making assumptions about the underlying transport regime and mechanism, see extensive reviews on the topic[15,16,57,75–81] or monographs.[2,46,82–84] The reason for the wealth of theoretical descriptions is simple and complicated at the same time: *"In essence, the problem consists in the prediction of the quantum dynamics in a system with strong coupling between electronic and nuclear degrees of freedom where it is not easy to introduce the standard approximations because all the relevant time/energy scales coincide."* as Nematiaram and Troisi aptly note.[81]

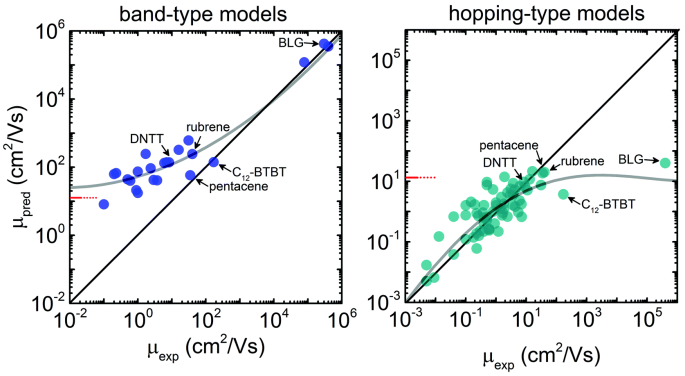


Figure 2.3 Experimental room-temperature organic crystal charge carrier mobilities compared to the respective theoretical predictions. A typical upper/lower limit of the respective theory is indicated by a red line. Gray lines indicate the applicability limits of the theories. Short forms of known materials have already been introduced in Figure 1.2. Mobility values above $10^4 \mathrm{cm}^2/\mathrm{Vs}$ amount to high-mobility reference systems, among them bi-layer graphene (BLG). Figure reproduced from reference 13 (I. Yavuz) with permission from the PCCP Owner Societies.

This was especially realized when reliable experimental mobility values became available that were measured in high-quality single-crystal OFETs.[85] The measured mobilities were neither fully consistent with predictions of widely employed localized charge carrier hopping nor with delocalized band-transport models as illustrated by Yavuz,[13] see Figure 2.3. In contrast to inorganic semiconductors where band models are in fact widely applicable, the ambiguity in OSCs results from the charge carrier's interactions with the surroundings, experienced when passing through the "soft" material. A quantitative description of charge carrier mobility in such single-crystal OSCs will therefore likely rely on an accurate description of an intermediate regime transport

mechanism in the crossover between typical hopping and band transport pictures. For a more systematic overview, the predominant interactions are here summarized in a minimal Hamiltonian, describing the charge carrier in an organic layer (following reference 16).

$$\hat{H} = \sum_{a} \epsilon_a \hat{c}_a^\dagger \hat{c}_a + \sum_{\substack{ad \\ a \neq d}} H_{ad} \hat{c}_a^\dagger \hat{c}_d + \\ \sum_{\mathbf{Q}} \hbar\omega_{\mathbf{Q}} \left(\hat{b}_{\mathbf{Q}}^\dagger \hat{b}_{\mathbf{Q}} + \frac{1}{2} \right) + \sum_{a\mathbf{Q}} \hbar\omega_{\mathbf{Q}} g_{aa\mathbf{Q}} \left(\hat{b}_{\mathbf{Q}}^\dagger + \hat{b}_{-\mathbf{Q}} \right) \hat{c}_a^\dagger \hat{c}_a + \sum_{\substack{ad\mathbf{Q} \\ a \neq d}} \hbar\omega_{\mathbf{Q}} g_{ad\mathbf{Q}} \left(\hat{b}_{\mathbf{Q}}^\dagger + \hat{b}_{-\mathbf{Q}} \right) \hat{c}_a^\dagger \hat{c}_d \tag{2.5}$$

In the purely electronic part (line 1), ϵ_a is the on-site energy at the molecular site a, while H_{ad} are the electronic coupling elements between site a with neighboring site d. $\hat{c}_a^{(\dagger)}$ denote the annihilation (creation) operators for a respective charge carrier. Adding to this electronic description of the charge carrier, in line 2 the presence and interaction with (harmonically approximated) lattice phonons modes $\mathbf{Q}$ of frequency $\omega_{\mathbf{Q}}$ is modeled, giving rise to "dynamic disorder".[86] Phonon annihilation (creation) operators $\hat{b}_{\mathbf{Q}}^{(\dagger)}$ occur respectively, while the strength of the coupling of charge carrier and phonons is (in linear approximation) described by the coupling constants $g_{aa\mathbf{Q}}$ and $g_{ad\mathbf{Q}}$. These respectively capture the local modulation of on-site energy and the non-local modulation of electronic couplings.

Depending on the strength of these couplings, a charge carrier is expected to be either fully localized (hopping from site to site), weakly (transiently) localized spanning multiple sites, or fully delocalized (forming a band) as schematically depicted in Figure 2.4. So far, these regimes are often treated with different theoretical models, while especially theory-development in the crossover regime of transient localization[79] as well as insights into the actual charge carrier dynamics are of current interest,[87] and have been reviewed recently.[15,75,81] In light of the results, the discussion about a realistic limit for charge-mobility in organic small molecule crystals is also still ongoing, but estimates range between 70 - 100 cm^2/Vs,[35,50] albeit an exceptionally-high value of 170 cm^2/Vs has been reported[88] in an OFET device at room temperature. In light of these results, mobilities of OSCs therefore easily surpass typical values found for amorphous silicon, but fall way behind crystalline silicon.[89]

From a materials design perspective, it is however often sufficient to rely on a simple hopping model that already incorporates the factors that should first and foremost be fulfilled to arrive at a well-performing material. I will therefore first introduce the main ideas behind classic hopping models. This also includes a summary of the most decisive factors they incorporate, for which I point out their role in other models and regimes. Nevertheless, the limitations of the discussed hopping model should be kept in mind and will be discussed as well.

Hopping regime

Hopping models operate under the assumption that a charge carrier intermittently localizes on molecular sites while traveling through the material. Pictorially speaking, this localization is caused by induced deformations that the charge carrier drags with it through the material (in combination a so-called "small polaron"). This polaron is then considered to move between the molecular sites in discrete jumps. For now, it is assumed that a molecular site is here a single molecule in the molecular crystal or amorphous film, but recalling Figure 2.4 and the associated discussion, this is not always given. Nevertheless, simple hopping models have been highly successful and often empirically reproduced trends among measured mobilities.[13,58,87] In the models,

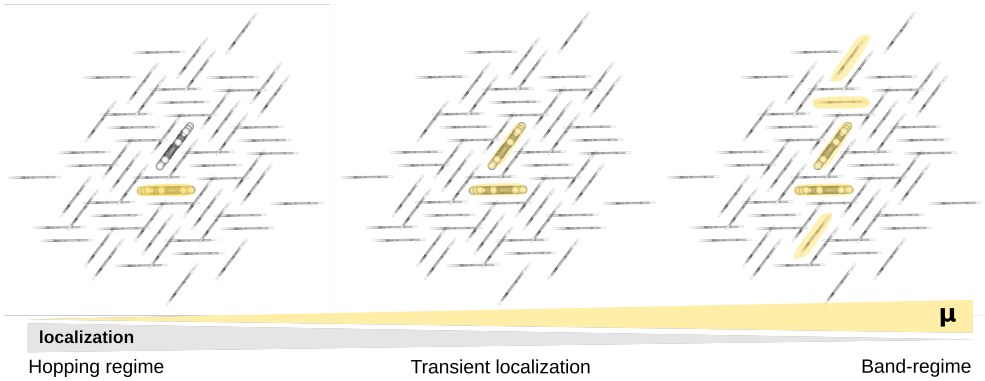


Figure 2.4 Charge carrier localization and mobility expected in different regimes. Note, that these are schematic depictions, for explicit simulations using see e.g. references 87, 90.

a charge carrier jumps from a molecular site a to the other sites d, quantifiable by charge transfer reaction rates k_{ad} which can be used to arrive at a mobility of a charge carrier at the site

$$\mu_a = \frac{q}{k_B T} \frac{1}{2n} \sum_d p_d k_{ad} r_{ad}^2 \tag{2.6}$$

with $p_d = k_{ad} / \sum_d k_{ad}$ being the probability of hopping from site a along the percolation path to site d. An important underlying assumption is thereby that of an incoherent transport, meaning hopping events are not influenced by preceding steps. Note however, that this simple weighted average can fail when a single, high coupling is present in the network where the charge carrier jumps back and forth, not contributing to the overall mobility. An alternative are kinetic Monte Carlo simulations that can be used to obtain such hopping mobilities.[58]

Based on semiclassical transition state theory, the charge-transfer rate is given by[75]

$$k_{\mathrm{TST}} = v_{\mathrm{eff}} \kappa_{\mathrm{el}} \Gamma_{\mathrm{n}} e^{-\beta(\Delta G^{\ddagger} - \Delta^{\ddagger})} \tag{2.7}$$

with the effective vibrational frequency v_{eff}, the electronic transmission coefficient κ_{el}, the nuclear tunneling factor Γ_{n} and $\beta = 1/\mathrm{k_b}T$. The activation energy $\Delta G^{\ddagger}$ for the transition between the diabatic donor and acceptor state is thereby corrected by an adiabatic correction $\Delta^{\ddagger}$ which captures the electronic coupling between these states, see below. I will here focus on the so-called non-adiabatic charge-transport regime. A number of models have here again been developed, working with different assumptions and levels of approximation.[13,91] Prototypical and widely used is the model devised by Marcus for electron transfer reactions in solvents.[92,93]

$$k_{ad} = \frac{2\pi}{\hbar} |H_{ad}|^2 \frac{1}{\sqrt{4\pi\lambda k_B T}} e^{-\beta \Delta G^{\ddagger}} \qquad \Delta G^{\ddagger} = \frac{(\lambda + \Delta G^0)^2}{4\lambda} \tag{2.8}$$

The determining descriptors are thereby the driving force ΔG^0, the **electronic coupling** H_{ad} between acceptor and donor, and the **reorganization energy** λ. In ordered, mono-molecular periodic crystals it is often the case that ΔG^0 is 0, simply due to the equivalence of crystal sites. We assume the latter approximation in our database analysis and λ and H_{ad} were therefore the main mobility-related descriptors. While introduced here in the context of the hopping regime, It is important to stress that λ and H_{ad} also play a general role in the modeling of mobility in other regimes,[75,81,87] see further discussion below.

As mentioned above, Marcus theory describes non-adiabatic charge-transport. In essence, the electronic coupling is here considered to be significantly smaller than the reorganization energy and it is assumed that the geometric relaxation of and around the molecular site upon charge localization is fast compared to the charge transfer itself. Clearly, not every system should be

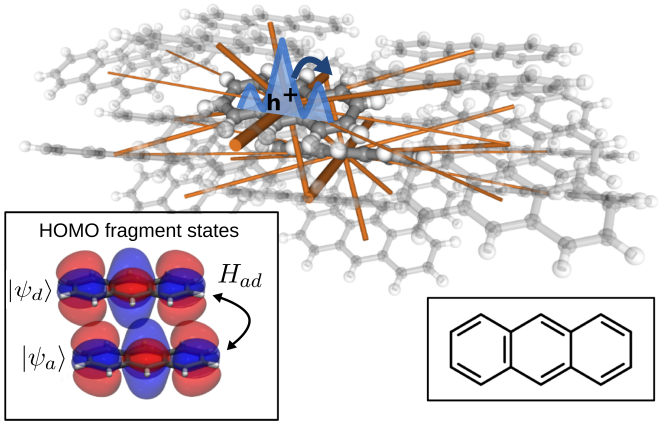


Figure 2.5 Hole hopping along the main charge transport pathways in the anthracene crystal. In the main part of the image the network of electronic couplings between molecular dimers forming transport pathways are illustrated by orange sticks, scaled by the respective size of $|H_{ad}|$. Note, that molecular sites of a single unit-cell are emphasized as ball-and-stick models, while the surrounding periodically repeated environment is represented with opacity. In the left inset, the FO-DFT method to obtain $|H_{ad}|$ values is illustrated. The ground-state highest occupied molecular orbitals obtained in isolated vacuum DFT-calculations form the approximate diabatic states. Here they are arranged according to a dimer geometry occurring in the crystal. As can be deduced from the nodal structure, the specific arrangement significantly influences electronic coupling elements. In the right inset, the molecular graph representation of anthracene is reproduced.

treated in this regime. Especially for organic crystals with good transport properties, e.g. a large electronic coupling and small reorganization energies (routinely reaching 100 meV in both cases), $\Delta^{\ddagger}$ becomes substantially large and the non-adiabatic picture of localized charge carriers and slow transfer breaks down, up to the point where the concept of a hopping-rate becomes ill-defined.[13,75] It is further noted that small polaron hopping is thermally activated as seen from equation 2.8, but it can not always be conclusively inferred from the temperature dependence of the charge carrier mobility.[94]

Electronic coupling

The modeling of charge transfer processes between a donor and an acceptor fragment often relies on diabatic (localized) electronic states. The electronic couplings H_{ad} among these states then influence the hopping rate (or in the band picture the degree of charge delocalization over the states). Unfortunately most electronic structure methods –including DFT– provide adiabatic states of the combined donor-acceptor system. Deriving diabatic states from this adiabatic representation, followed by an accurate calculation of their electronic coupling is therefore an important part in the accurate modeling of a charge-transfer process.

A number of methods exist, among them constrained density functional theory (CDFT), the generalized Mulliken-Hush method, or Block-Diagonalization, see Oberhofer et al.[75] for an overview. We here mainly relied on the so-called fragment-orbital density functional theory (FO-DFT), a standard method in the OSC field.[95–97] The H_{ad} values contained in the 64k-dataset [30,39] were obtained by the implementation of this method in the FHI-aims DFT code.[97]

$$H_{ad} = \langle \psi_a | \hat{H} | \psi_d \rangle \tag{2.9}$$

with the Hamiltonian $\hat{H}$ of the donor-acceptor system and the single diabatic states of donor and acceptor $|\psi_d\rangle$ and $|\psi_a\rangle$.

In FO-DFT these diabatic states are approximated from isolated monomer calculations, carried out for each of the neighboring molecules (e.g. extracted from the crystal) as Figure 2.5 illustrates for anthracene. From the obtained fragment Kohn-Sham states, the combined Hamiltonian $\hat{H}$ of the donor-acceptor system can then be constructed and the coupling element H_{ad} evaluated as $\langle\phi_a|\hat{H}|\phi_d\rangle$ between the (orthogonalized) monomer Kohn-Sham orbitals. For hole (electron) transport, HOMOs (LUMOs) are thereby decisive. Even when relying on a computationally cheap non-selfconsistent construction of the Hamiltonian, the FO-DFT approach yields electronic couplings of high accuracy as benchmark results demonstrated.[97] Specifically, the best-performing $H^{2n-1}@D^+A$ method was used in the database screening of Schober et al.[30] for the construction of the 64k-dataset . A downside of the method is that polarization effects between donor and acceptor are in this way not taken into account. Our recently developed Block-Diagonalization methods that yield well-localized diabatic states could be used in that case.[98]

The nodal structure of the molecular orbitals (HOMO or LUMO) can lead to highly modulated electronic coupling values already at small geometric displacements.[57,99,100] Even in well-defined crystalline materials the values can thus be heavily influenced by the application of external pressure[101] or strongly modulated by the intrinsic intermolecular vibrations.[102] As mentioned above, the treatment of the latter contribution ("dynamic disorder") is considered an important factor in the accurate modeling of charge mobility, while strategies to reduce it are being investigated.[103] From a molecular design perspective, the highly non-linear behavior of electronic couplings however also makes molecular design or discovery of novel organic semiconductors an extremely challenging task. This can be easily seen from the fact, that polymorphs –different, stable experimental crystal structures of the same molecule– often show differing H_{ad} values[88,100,104,105] and charge mobilities related to the differing molecular arrangement. To accelerate the computation of this highly sensitive property during multiscale simulations, machine-learning based predictions of electronic coupling elements are now entering the field.[106–108] From an experimental perspective, controlling the molecular arrangement in the solid-state thin-film by a well-defined deposition technique is thus crucial for high device performance. One can also conclude that a successful in-silico design-problem thus heavily relies on obtaining accurate structural models of it, see discussion in chapter 7.

Reorganization energy

The reorganization energy λ provides a measure for charge-carrier stabilization due to its interaction with molecular site and surroundings. It is closely related to the more descriptive concept of the "polaron binding energy".[57,109] Accordingly partitioning λ into a local contribution of the molecular site and a nonlocal contribution of the molecular surroundings, the former is in fact often dominant.[110–112] In a good approximation, we therefore here focused on the local contribution to judge a molecules prospective applicability in the charge-transport context, simply denoting it as λ.

To compute λ, the 4-point method by Nelsen[113] is often used, in which separate total energy calculations are carried out in vacuum:

$$\lambda = \lambda_+ + \lambda_0 = [E^0(R_+) - E^0(R_0)] + [E^+(R_0) - E^+(R_+)] \tag{2.10}$$

The four occurring total energies E thereby result from the combination of the charged (E^+) or neutral (E^0) electronic state energies at equilibrium geometries for charged (R_+) or neutral states (R_0) respectively, see Figure 2.6 a) for a schematic depiction of the involved potential energy surfaces. Arrows correspondingly illustrate how λ measures the energetic cost to convert the donor state nuclear configuration to the respective acceptor state configuration,[76] while keeping the electronic configuration fixed.

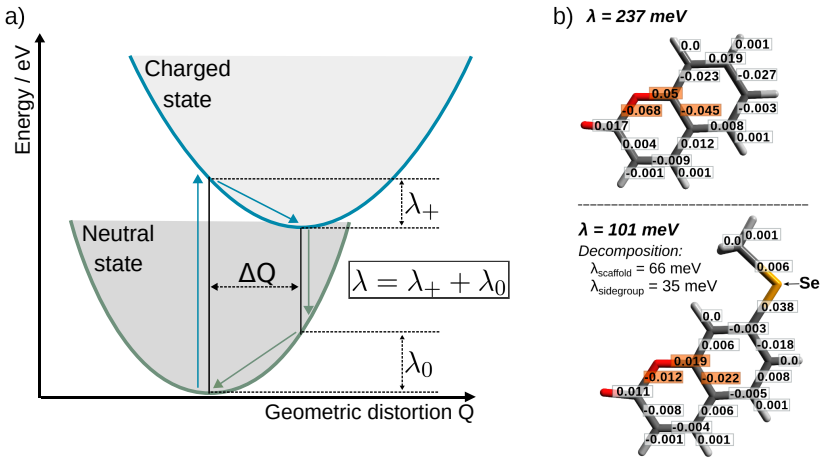


Figure 2.6 a) State-diagram of charged and neutral state. Using the 4-point method, λ can be derived from single-point energies obtained for the two potential energy surfaces. b) Effect of molecular modification on reorganization energies. At the top, a simple scaffold is shown, together with bond-length variations (in Å) between neutral and charged state geometries. Below, the scaffold is modified with a favorable side group, lowering its reorganization energy. A simple decomposition scheme can be used to arrive at contributions, see text. Figure reproduced with permission from reference 40 under the terms of American Chemical Society's Policy on Theses and Dissertations.

Aside from the 4-point method, and especially for rigid, fused organic systems, contributions to λ can often be reliably computed from vibrational harmonic normal mode contributions as[76,114]

$$\lambda_{+/0} = \sum_M \lambda_M = \sum_M \frac{1}{2}\omega_M \Delta Q_M^2 \tag{2.11}$$

where ω_M is the frequency of normal mode M, and ΔQ_M the mode resolved displacement between the involved state geometries (neutral and charged states). Alternatively, the normal-mode resolved electron-vibrational (vibronic) coupling can be assessed from forces arising in the charged state at the neutral state geometry.[24,74,115] It is now easily seen, that the (local) reorganization energy is directly related to the local electron-phonon coupling of charge carrier and molecular site, which already appeared in equation 2.5.

I will now mostly focus on holes as charge-carriers, as these were the main target of the work, denoting the corresponding reorganization energy as λ_h . DFT has been established as the standard method for the practical computation of molecular reorganization energies. It should however be considered, that charge-carrier localization is highly dependent on the employed exchange-correlation functional. Global hybrid functionals like the routinely employed B3LYP functional might thereby still underestimate local reorganization energies, albeit reproducing trends well.[116] Experimental reorganization energies measured by UPS for a few common oligoacene OSC materials have been found to match values obtained at the B3LYP.[117]

From a molecular design perspective, the decomposition of λ_h is valuable as an analytic tool.[118] As especially the low-energy normal modes are often delocalized over the whole molecule it is however difficult to derive insight, let alone to propose changes to the molecular structure based on them. Other more local mode decomposition schemes have therefore been applied, allowing one to focus the analysis on the contribution of specific molecular parts. A good example is an investigation of chemical substitution effects of an indolocarbazole framework by means of an internal coordinate based decomposition.[24] In our systematic investigation of side-group effects,[40] we found that evaluating λ_h contributions on two geometries in which a molecular framework geometry in its charged state is combined with the neutral side group geometry and vice versa

already yielded a quantitative decomposition for a single side-group, see Figure 2.6 b). A more recent decomposition into larger more localized fragment contributions was further proposed, deriving from them a successful molecular redesign that adds a non-covalent lock to the molecular backbone.[119] Machine learning[33] or evolutionary strategies[120] have also been employed to uncover useful compound designs, including a recent study using AML.[121]

As according to Koopmans-theorem the HOMO is the relevant frontier orbital to accommodate a hole as a charge carrier, design strategies that focus on the tuning of the HOMO wavefunction have been proposed.[118,122–124] Similarly, vibronic coupling densities have also been proposed as an analytical tool.[74] Apart, the literature also contains a set of empirical rules and structure-property relationships for molecular modification to improve the reorganization energy, such as heteroatom replacement,[125,126] introduction of side groups,[24,40,127] or enlarging the system by ring fusion,[123,124,128] the latter also exploiting the tendency of λ_{h} to decrease with increasing molecular size.[30] These strategies are further applicable to the tuning of HOMO and LUMO energies,[23,129] relevant to charge-injection from electrodes, see above. We incorporated such strategies in our AML method based on chemical transformations of the molecular graph, see section 6.5.

3 | The 64k- and OE62-datasets

Parts of the presented work are based on the "64k-dataset", assembled by Christoph Schober, Harald Oberhofer and Karsten Reuter.[30,39] It was choosen as a starting point for this dissertation to extract structure-property relationships for molecular design from a diverse dataset. In this respect, it provided the necessary large data-source and is here described in more detail.

The dataset was originally assembled in one of the first large-scale virtual screening efforts for high charge-mobility in OSCs. Accordingly, the study leveraged computational methods to evaluate a large pool of candidate materials and to filter for the most promising ones. Similar concepts mostly originated in the field of drug discovery,[130,131] and were later adapted to materials discovery,[132] among them organic functional materials[133] or organic photovoltaics.[28,29,134,135]

Experimentally well-resolved organic crystal structures were used as a starting-point for the undertaken screening. These crystal structures were originally retrieved from the Cambridge Structural Database (CSD) – to date the largest repository of crystallographic data for organic crystals.[14,136] Among 750.000 experimental structures, an initial library of 95.445 was built, focusing on well resolved organic crystals without metal-organic components, structural disorder or polymeric parts, containing only one type of molecular species in each crystal ("monomolecular crystal"). Using a Python-based workflow, the library was then screened for descriptors of high charge-carrier mobility by employing a computational funnel concept, see Figure 3.1. This included extraction of molecular environments around a single molecule and the computation of electronic coupling elements H_{ad} for all molecular pairs formed between central molecule and neighbor using FO-DFT. At this stage, 64.725 organic crystals had been successfully processed and finally entered the "64k-dataset". A crystal was now only passed on for further processing when an H_{ad} value exceeded a minimal threshold value of 50 meV, as confirmed at higher accuracy computational settings. For a confirmed favorable crystal λ_{h} finally was calculated, preserving a realistic solid-state environment, finally leading to 10.214 annotated crystals.

The outcome of this screening is presented in a 2D histogram, see Figure 3.2. The final selection already recovered known and well-performing materials and examples are indicated by orange markers –as expected– appearing in the region of high maximal $|H_{ad}^{\mathrm{max}}|$ and low λ_{h}. The screening however also uncovered many promising candidates, not yet considered for organic electronics applications. In the histogram, four cases are highlighted in red and molecular structures are provided. A closer analysis of their connected charge-transport networks is provided in the original work.[30]

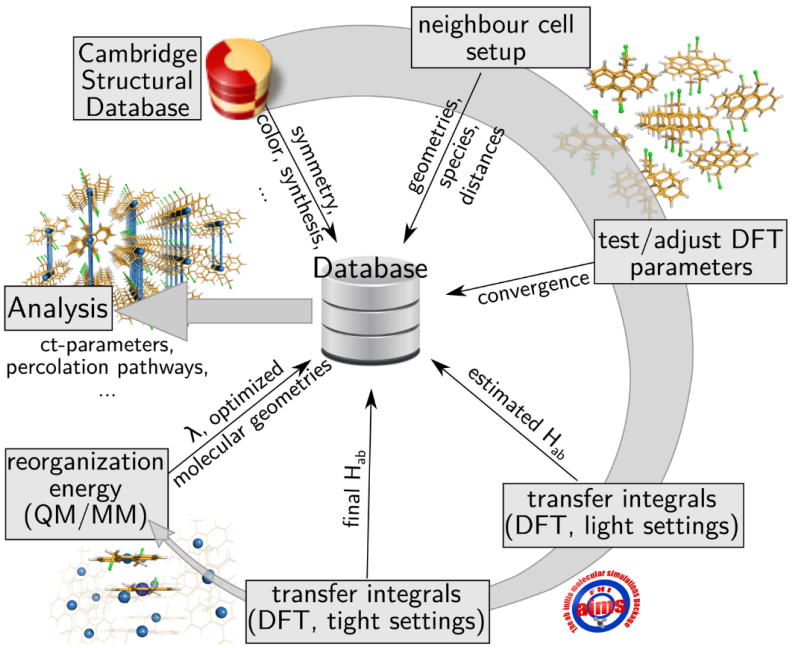


Figure 3.1 Overview of the virtual screening approach that led to the generation of the 64k-dataset. Reproduced from reference 30 with permission from the American Chemical Society.

The 64k-dataset was stored in an SQLite database from where it could be further analyzed. In short, the database contained the xyz-coordinates of the central molecule originally extracted from the crystal, its molecular graph stored in a unique SMILES string, metadata and crystallographic information from the CSD, cheminformatics descriptors derived from the RDKit,[137] as well as quantum chemical data for the respective entries ($|H_{ad}|$, λ_{h} and HOMO-/LUMO energy). A full description is given in.[39] Stored in this form, the 64k-dataset has formed the basis of work carried out in this dissertation. In a first step, we performed a molecular analysis to uncover structure-property relationships present in the dataset. We thereby relied on clustering and data-mining as described in[40] or summarized in section 6.1. We further provide a visually understandable representation of this analysis,[41] see section 6.2.

During our work on the dataset, we also investigated the diversity of molecular structures and crystals contained in it. This analysis revealed a high structural diversity. The molecules in the dataset are composed of up to 174 (or 92 non-hydrogen) atoms, covering all elements commonly occurring in organic chemistry, adding some less commonly occurring ones, see Figure 3.3 a). These molecules are distributed over more than 20 molecular point groups, while for their crystals, 83 space groups were found to occur at least three times in the dataset. Further, a variety of molecular scaffolds are contained with > 800 side chains occurring in more than three different crystals.[40] While an absolute attribution of "diversity" is challenging,[138] these numbers indicate a high diversity. In fact, in a short survey, we found common pharmaceuticals, intermediates of chemical synthesis, pigments, fungicides, antioxidants or photoinitiators occurring in the 64k-dataset. As has also been pointed out by others, the underlying CSD is in fact considered to be composed of highly diverse molecules[136] with organic crystal structures originating from a wide background of chemical applications.

Having established that the 64k-dataset is composed of highly diverse molecules, we envisioned it to be useful for the testing and benchmarking of quantum machine learning methods on realistically sized and technologically relevant molecules. In collaboration with our colleagues from Aalto

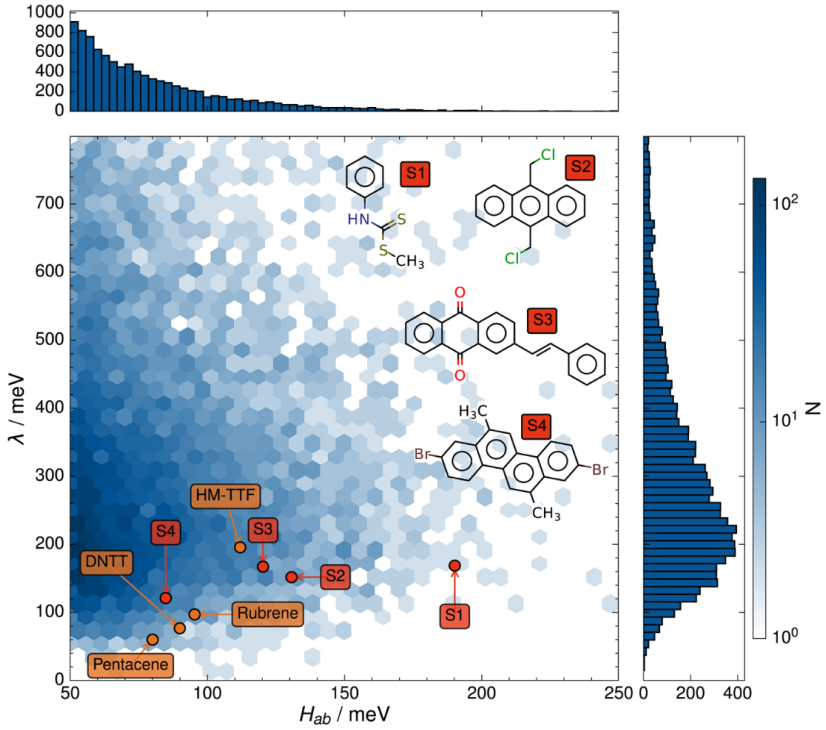


Figure 3.2 Final selection from the 64k-dataset . The result is presented as a histogram (middle 2D, flanked with 1D versions). Reproduced from reference 30 with permission from the American Chemical Society.

University we therefore decided to base a new dataset for molecular quantum machine learning on it – the OE62-dataset. To provide some background on this idea, I here want to compare molecular composition and size-distribution to the often cited and widely used QM9-dataset[139] which contains 133.885 small organic molecules. As seen in Figure 3.3 b), molecules in QM9 are significantly smaller and composed of H and up to nine heavy atoms (C, N, O, and F only). An additional characteristic of QM9 is the (by construction) dense coverage of chemical space, clearly distinguishing it from OE62. This is due to the fact that molecules in QM9 originate from the exhaustively enumerated GDB-9 subset of the GDB-17 dataset,[140] and were thus derived from a virtual library enumeration approach. An overview of QM9 is also provided in our work on the visualization of molecular- and crystal structures,[45] see also chapter 5. QM9 then provides equilibrium geometries and 13 quantum chemical properties computed at a hybrid DFT level of theory. Based on this data, quantum machine learning models have been developed that can accurately predict these properties, ever improving in accuracy.[141] As others have also noted, a general lack of diversity could hamper the further development of machine learning models,[142] while the advent of new and challenging datasets could spur its further spread.

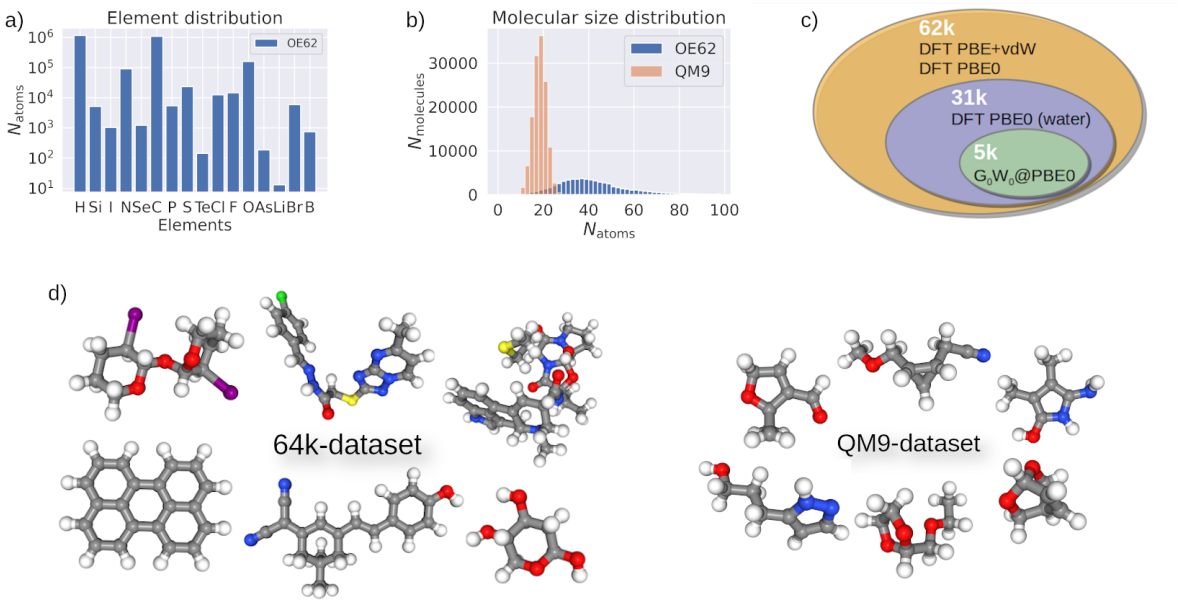


Figure 3.3 a) Composition (by element) of the unique molecules contained in the 64k-dataset . b) Molecular size distribution of OE62- and QM9-datasets. c) Overview of the computational results available in the OE62 dataset and its three subsets. d) Comparison of typical structures found in the OE62- and QM9-datasets. A combinatorally rich chemical space arises by the combination of a multitude of chemical scaffolds, side groups and chemical elements of which these molecules are composed of.

The published OE62-dataset[42] provides a highly diverse and challenging dataset that can be used for the further development of machine learning models. It contains equilibrium gas-phase geometries for 61.489 unique molecules taken from the 64k-dataset , calculated at the DFT GGA+vdW level of theory. In addition, smaller subsets provide computational data at other levels of theory, see Figure 6.3 c) for a Venn-diagram of the structure. Since our colleagues from Aalto-University mainly pursued the development of machine learning models for molecular energy level prediction, these are available in the dataset at different levels of theory. A full summary is provided in section 6.3. In two studies with my involvement, we already saw how the dataset poses new challenges for common molecular representations and machine learning models. For completeness, I want to point out that additional datasets for quantum machine learing are available, see a summary.[143]

4 | Molecular analysis and machine learning

Data-driven approaches[25] and artificial intelligence (AI)[144] have regained popularity in all sciences. With early approaches e.g. undertaken in cheminformatics,[145,146] this includes the chemical sciences as well. A number of recent specialized review articles now cover this broad array of methods, especially focusing on the most important subfield of machine learning.[141,147–154] These methods now also partially fulfill the promise of a more efficient *in silico* drug-discovery[155] in vast chemical spaces, or use of these methods for advanced modeling.[156,157] Following this general trend, and especially in the years of working on this dissertation, such methods are also entering the field of OSC materials design,[32,33,37,38,43,158] see also.[36,38] This is not surprising: As mentioned before, the chemical spaces under scrutiny are vast, and potentially not enumerable. An exhaustive screening for desired but highly nonlinear electronic properties thus benefits greatly from the use of efficient, data-driven methods that can navigate such spaces.

We herein employed such data-driven methods for the identification of OSC design principles. On the one hand, we therefore relied on methods of **data-mining** ("knowledge discovery from data") to derive chemical insight from our large 64k-dataset , see Figure 4.1 a) for a schematic depiction. We thereby focused on molecular substructures and their relation to properties, decomposing a molecular graph along specific bonds. Different decomposition schemes can be employed for such tasks, including decomposition along synthetically reasonable disconnections (e.g. RECAP[159] or BRICS[160]), or (rotatable) non-ring bonds.[161] Such decomposition schemes are often also stepping-stones for combinatorial, targeted molecular library creation[162,163] and de-novo design of molecules,[164–166] see also our active-learning discovery of molecules (section 6.5). For our data-mining approach,[40] we relied on a fragment-definition that incorporates the chemically intuitive notion of a molecular scaffold (or backbone). This was afforded by the scaffold-definition of Bemis and Murcko,[167,168] in which a molecular scaffold is defined as an (aromatic) core composed of connected ring systems and conjugated linkers, resulting after removal of all side group atoms that branch off from it, see Figure 4.2. This often extracts the largest and most decisive part of the molecule – largely determining electronic structure, shape, and conformational flexibility. Vice versa, and based on this definition, the influence of side groups can be investigated as well.

In a second step, privileged structures among these molecular scaffolds or side groups were identified. Using e.g. a statistical assessment, a fragment can be evaluated for a significant structure-property relation. We here used the nonparametric Mann-Whitney U test[169] to verify that the median of a property-distribution of all compounds with a certain substructure is significantly different from the background distribution. In fact, statistically significant relationships could be discovered in the scaffold and side group clustered data originating from the 64k-dataset , see the summary in 6.1. Similar ideas have been applied to interpret gene expression data (gene-

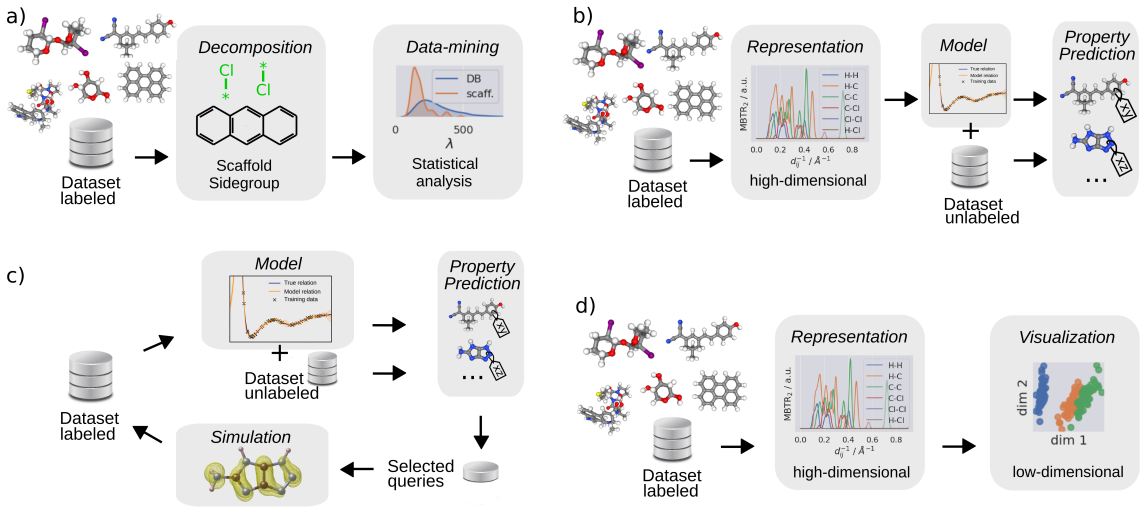


Figure 4.1 Schematic summary of data-driven methods used in this work. a) Data-mining strategies applied to derive structure-property relationships from a chemical database. b) Supervised machine-learning for molecular property prediction. c) Active-learning cycle allowing a gradually improving machine-learning model to propose an approximate strategy to an optimal exploration-exploitation tradeoff. d) Unsupervised visualization of a high-dimensional data in a low dimensional space, see chapter for a discussion 5.

set enrichment analysis),[170] to derive structural alerts for chemical compound toxicity,[171] or to identify biologically active chemical series from screening data on drug-targets.[172]

While these data-mining methods could easily provide chemical insight, they relied on a specific type of molecular decomposition. We thus additionally employed **supervised machine-learning** methods to derive surrogate models that can more directly interpolate the molecular property space. These surrogate models can thus be used to infer properties of so far unlabeled structures –allowing for a computationally cheap molecular property prediction–, while being trained on a limited amount of data e.g. generated from expensive *ab initio* computations, see Figure 4.1 b). By making use of the underlying correlations present in the dataset these methods therefore hold the promise that redundant calculations can be avoided.

A certain amount of expensive training data is still needed and used to derive the ML models. If not available, this data needs to be generated first, usually requiring some cost for annotation with labels (e.g. by DFT computations on molecular structures). By resorting to more efficient and smart data sampling methods such as **active machine learning**[173] (AML), the cost of labeling can be minimized significantly. Smart sampling can e.g. mean that a successively improving machine-learning model subsequently infers useful training examples for which properties are then simulated or obtained. Figure 4.1 c) depicts a respective design cycle. The process can also be designed in such a way that a significant number of desired molecules is discovered along the way – exploiting the knowledge the models gradually gain, while exploring the space further. Such strategies have found use for drug-discovery[174] and later-on also in materials science.[175]

A fourth strategy is geared towards an intuitive understanding of structure-property relationships in molecular datasets by means of **visualization**, see Figure 4.1 d). This is more akin to unsupervised machine learning and will be discussed in detail in chapter 5. In the remaining chapter I will now provide a brief overview of ML methods used while performing the work for this dissertation.

With the completion of our OE62-dataset for molecular machine learning (chapter 3), different machine learning methods for molecular property prediction were applied and further developed. An important first step is thereby to transform the molecular structures into a suitable repre-

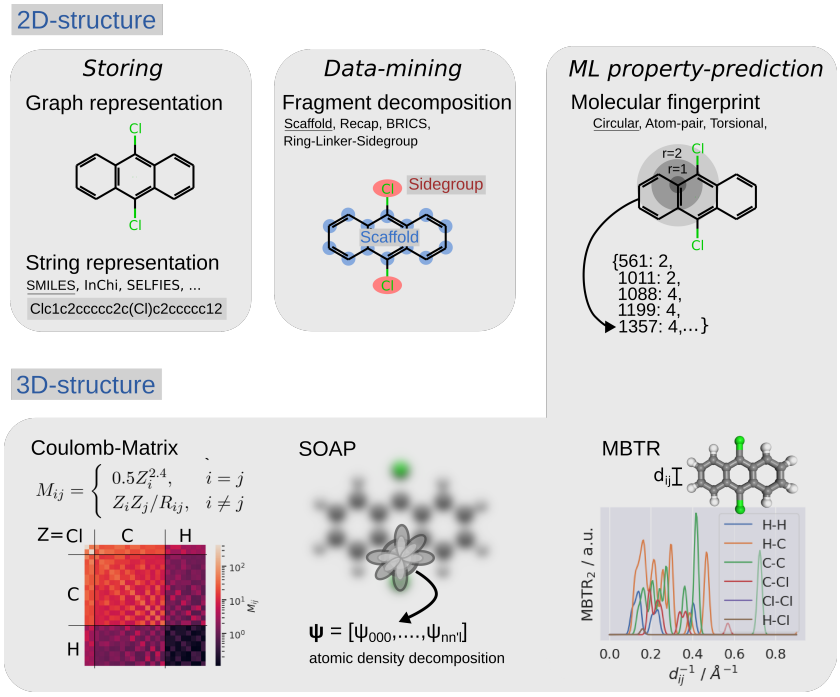


Figure 4.2 Summary of molecular representations used throughout the work.

sentation (e.g. a vector) that encodes the relation of their constituent atoms (either in 2D or 3D). These representations are passed to the learning algorithms, that can infer the underlying structure-property relationships from a labeled dataset, while the respective properties for the hitherto unlabeled molecules of interest can be predicted. The process of creating a representation is called molecular featurization and a larger number of algorithms for it has been proposed in different communities.[176–179] The featurization process thereby either starts from the molecular graph, the molecular 3D structure or quantum chemical data, such as the molecular density or electrostatic potential. The resulting representations are then stored as (fixed-length) vectors or matrices, encoding structural counts, 2D graph representations, grid representations of spatial data. Representations used within this work are summarized in Figure 4.2.

A common type of representation are **"molecular fingerprints"**, often put to use in virtual screening and similarity searching.[180,181] The generation of such fingerprints usually relies on a rule-based partitioning of the 2D molecular graph into linear or branched subgraphs, while storing the occurrence (counts) of the latter. In our work on an active machine-learning exploration of a molecular OSC design space, we here mostly relied on the "extended connectivity fingerprints",[182] see Figure 4.2. To derive the fingerprint of a particular molecule, all available circular subgraphs around the atoms are produced – starting from each atom and moving outwards up to a predefined diameter. Iterative updating and hashing then generate unique identifiers for each type of subgraph. Fixed-length bit- or count vectors can also be generated from this data, e.g. by the application of a hashing function and a subsequent folding to a fixed-length vector representation. This particular, as well as other similar molecular fingerprints are still in heavy use. It should however be noted that the idea of a more flexible, data-driven subgraph extraction is being actively developed, e.g. employing graph-neural networks that directly infer the input representation.[177,183,184]

On the other hand, and more often based on 3D-structures, significant progress has been made in applying machine learning for the prediction of quantum mechanical properties.[141,185] Again, a variety of suitable molecular representations have been developed. I will review those represen-

tations relevant to the work on molecular machine learning I was involved in[43,44] as well as the one used in the work on visualization.[45] A summary of these is shown in Figure 4.2.

- The **coulomb matrix (CM)** is an early example.[186] Essentially it is an inverse distance matrix, encoding the internuclear coulomb repulsion between atoms. The underlying idea is thereby to mimic the structure of the molecular Hamiltonian, providing an analogous mapping between coordinates and properties. The CM does however not respect atom-permutational symmetry and this is a clear drawback of the representation. It is therefore now mainly used as a baseline model and for illustrative purposes. In this spirit it was also employed in the work of Stuke et al.,[43] that dealt with the prediction of molecular orbital energies of our OE62-dataset, see section 6.3.

- We have also made use of a global descriptor for molecular or crystal structures – **the many-body tensor representation (MBTR)**.[187] The descriptor thereby collects the characteristic geometric features occurring in a structure (atom counts, distances, angles or even higher body-order terms) in broadened, discretized (fixed-length) distribution functions, thereby achieving permutational and rotational invariance. The MBTR representation should be a natural choice if global properties of structures are to be predicted as the representation initially does not assume a conceptual decomposition into local (atomic) properties. It was therefore employed to predict HOMO energies as well as atomization energies for molecules contained OE62-dataset, see section 6.3.

- The **smooth overlap of atomic positions (SOAP)**[179,185] representation provides a local descriptor for atomic environments. An atomic density function is first created by placing a Gaussian density on each atom. This density is then expanded in atom-centered spherical harmonics and orthogonal radial functions, within a cutoff. The expansion-coefficients can be used to form a rotationally invariant descriptor. In the context of this thesis, we mainly used this concept to visualize the different types of atomic environments present in it. Note however, that the descriptor is also widely employed to fit interatomic potentials.[157,188] While the concept fully builds on local structural representation of a single atomic environment, it can easily be extended to a global descriptor (for molecules or periodic structures alike) by averaging over structures.[189] In our work on visualization, we also employed this concept in the context of visualizing the relationships between molecules.[45]

Having introduced the representations, I now want to focus on machine-learning models. Again, numerous methods[190–192] have been popularized,[193] with significant recent development especially taking place in the deep-learning community.[194] While the discussed representations are quite general in their construction, at least some of them have been developed and used in combination with kernel-based machine learning methods[195,196] such as Kernel Ridge Regression (KRR) and Gaussian Process Regression (GPR) – now highly popular in the chemical sciences. Since we were mainly concerned with scalar property prediction (e.g. HOMO orbital energy, λ) we here used these models to map multidimensional inputs to respective scalar outputs, while the mapping is inferred from the labeled training data first. While KRR has also been used in our work,[43,44] I will focus on GPR,[197,198] easily employable for supervised- and active machine learning tasks.

For a molecular machine learning task, a training set $D = \{X, \boldsymbol{y}\}$ is assumed, with $X = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_N\}$ denoting a set of molecular descriptor vectors, e.g. derived by one of the above-discussed representations of Figure 4.2. The associated property values $\boldsymbol{y} = \{y_1, ..., y_N\}$ on the other hand are also available, e.g. from previously performed DFT simulations. In the GPR-

framework, a property prediction for an unlabeled structure $\boldsymbol{x}'$ can then be obtained, it is however not simply a scalar, but follows the predictive Gaussian distribution

$$f(\boldsymbol{x}') \sim \mathcal{N}(\mu(\boldsymbol{x}'), \sigma^2(\boldsymbol{x}')) \tag{4.1}$$

The scalar property-prediction can then be obtained as the mean $\mu(\boldsymbol{x}')$ of this distribution as

$$\mu(\boldsymbol{x}') = k(\boldsymbol{x}', X)[K + \sigma_n^2 I]^{-1}\boldsymbol{y} \tag{4.2}$$

where I is the identity. In addition, a predictive variance $\sigma(\boldsymbol{x}')^2$ can be obtained, its magnitude indicating, whether $\boldsymbol{x}'$ is coming from a densely- or loosely sampled region of feature space, about which the model is more or less uncertain in its prediction

$$\sigma^2(\boldsymbol{x}') = k(\boldsymbol{x}', \boldsymbol{x}') - k(\boldsymbol{x}', X)[K + \sigma_n^2 I]^{-1}k(X, \boldsymbol{x}') \tag{4.3}$$

In both cases, $k(\boldsymbol{x}, \boldsymbol{x}')$ is the covariance- or kernel function, which measures the similarity between two molecular representations $\boldsymbol{x}$ and $\boldsymbol{x}'$. K then correspondingly known as the covariance or kernel matrix of the training set, with entries defined as $K_{ij} = k(x_i, x_j)$. A noise level σ_n can thereby be used to model intrinsic noise in the property values. As an example I want to mention the stationary Gaussian kernel

$$k(\boldsymbol{x}, \boldsymbol{x}') = \sigma_v^2 \exp\left(-\frac{d^2}{2l^2}\right) \tag{4.4}$$

with euclidean distances $d = \|\boldsymbol{x} - \boldsymbol{x}'\|_2$ between datapoints, often applied for CM- or MBTR representations. Noise level σ_n, vertical scale σ_v and kernel bandwith l are so-called hyperparameters that need to be determined during model fitting, and can be inferred from the training-set[197] e.g. by log marginal likelihood maximization. Commonly employed in cheminformatics are count-based kernels,[199] take e.g. the Tanimoto-kernel, which measures the similarity between molecular fingerprints. In principle many possibilities for valid kernel construction exist. For a more technical introduction on the underlying prerequisites for a valid (Mercer) kernel, I refer to references 191, 197 and 199. It should however be noted that publications describing molecular representations often also suggest an employable kernel measure.

As the Gaussian Process Regression (GPR) model provides property predictions $\mu(\boldsymbol{x}')$ it can be employed for supervised machine learning to predict molecular properties. The inherently provided uncertainty estimates $\sigma^2(\boldsymbol{x}')$ however makes the model more versatile. In the spirit of an active-learning strategy, $\sigma^2(\boldsymbol{x}')$ can e.g. be used to identify candidates of low predictability, and requesting an explicit descriptor calculation on such molecules can thus provide significant new information, finally increasing the applicability of the model (uncertainty sampling). Combining both objectives in an acquisition function, a balance between exploratory and exploitative queries can be achieved.[200]

5 | Visualization of high-dimensional chemical space

With growing dataset sizes, advanced visualization techniques have become an important area of machine learning research.[201,202] Such techniques can be used to visualize the underlying trends in the highly non-linear property-distributions of a dataset.[201,203–205] A deeper understanding of a new dataset can thereby often be gained, or opaque machine learning predictions can be rationalized, resembling ideas of explorative data analysis and explainable AI.[206]

The success of such "unsupervised machine learning" methods rests on the assumption that a dataset can be represented and visualized over two- or three dimensions while (qualitatively) preserving the underlying relationships among high-dimensional (molecular) representations. To arrive at such visualizations, sophisticated dimensionality reduction techniques often need to be applied. An important linear method among them is principal component analysis (PCA). PCA extracts representative linear combinations of features to capture variance in as few dimensions as possible. In high-dimensional spaces, this linear technique can however reach its limitations and non-linear projections are used instead. A reformulation of PCA – Kernel PCA (KPCA) can easily incorporate such non-linearity by employing a non-linear kernel similarity between (molecular) representations (see chapter 4). In addition, other methods have gained popularity in the chemical sciences as well, including t-SNE,[207] sketch-map[203] or UMAP[208] and could be applied for "materials cartography".[209]

In this dissertation we relied on PCA and KPCA to illustrate how dimensionality reduced visualizations can be helpful in the analysis of crystalline or amorphous materials, as well as of molecular datasets. PCA and KPCA were thereby used in conjunction with the SOAP descriptor for 3D structures, already discussed in section 4. A summary of this extensive perspective article is provided in section 6.4, while Figure 5.1 reproduces an example map. While the article points out why and how such maps are highly useful, it should be noted that their appearance is dependent on the specific method used for dimensionality reduction, as well as on the representation used to encode the structures.

In cheminformatics,[211] methods based on neural networks (self-organizing maps, or generative topographic mapping[212]) or network structures are also widely used.[168] The latter type can e.g. be used to visualize structure–activity relationships in biologically relevant molecular space.[213] For a visualization of an organic semiconductor design space, we also relied on this technique and provided a chemical space network (CSN) visualizing similarities between molecular scaffolds contained in the 64k-dataset . Annotated with performance-related electronic descriptors, a layout of the representation places related structures in close proximity and can easily be generated. Again, two major factors that influence the final appearance of these CSNs are measures of molecular similarity and the specific layout algorithm employed. A summary of this approach is provided in section 6.2.

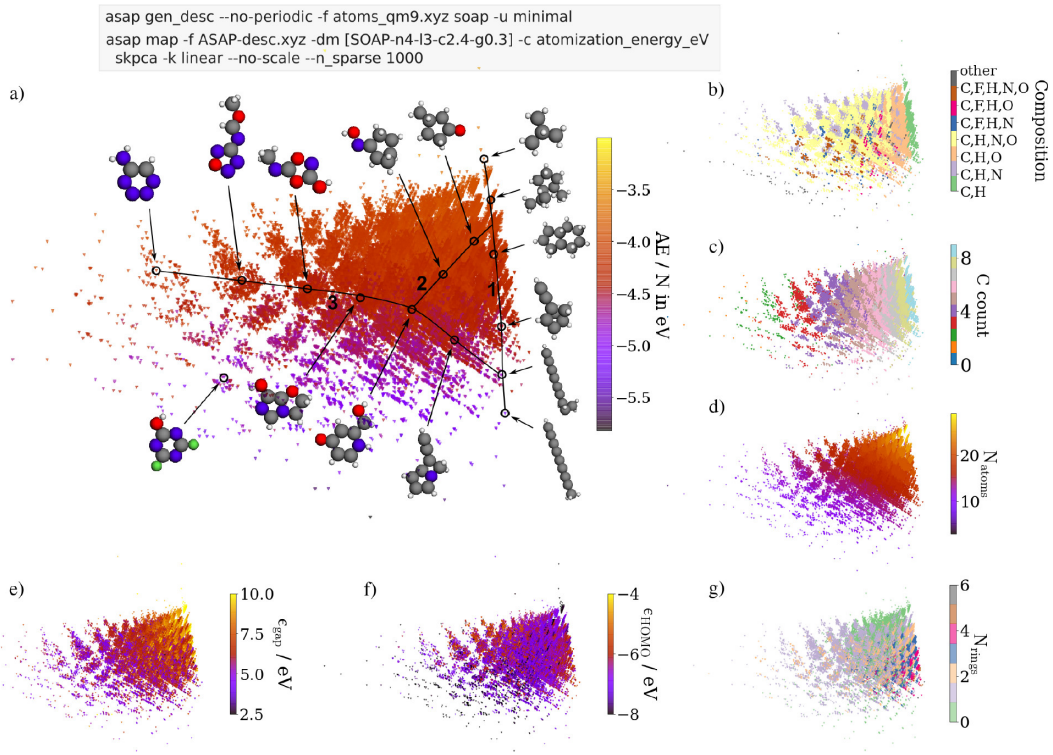


Figure 5.1 KPCA map of the QM9-dataset of small molecules (see also chapter 3). A global SOAP kernel was used for structural representation and similarity comparison. The composition and topology of molecules contained in the finalized map can then be navigated along various paths through the chemical space. During production of the map, these were compiled using the interactive viewer, see Figure 5.2. Color-coded structural descriptors (b, c, d, g) and quantum mechanical properties (a, e, f) are shown in different frames. A command for the ASAP code[210] is provided in a gray box above, with which the layout can be readily reproduced. A full discussion is given in reference 45 from which the figure was also reprinted with permission under the terms of American Chemical Society's Policy on Theses and Dissertations.

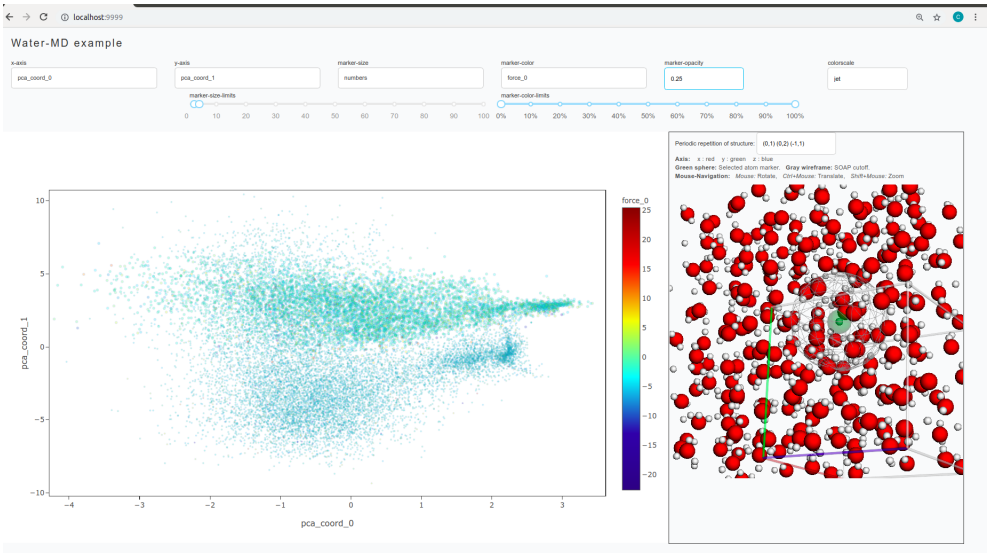


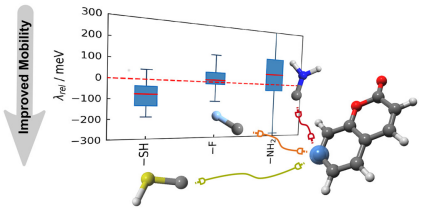
Figure 5.2 Screenshot of the browser-based projection viewer[216] for interactive exploration of dimensionality-reduced structural datasets. In the top bar, fields (pertaining to properties or coordinates in the dimensionality reduced projection) can be selected to be displayed on the x- or y-axis, or used for coloring or scaling of points in the scatter plot (displayed in the lower left part of the window). On the right hand side, a structural viewer is implemented, that visualizes respective structures selected upon clicking on a point in the scatterplot. This mechanism allows for interactive exploration of the map.

An important factor that made these dimensionality-reduction techniques useful to us are interactive visualization tools. These allow for an interactive exploration of large datasets. We here mostly relied on general-purpose libraries for interactive visualization to easily create interactive plots or dashboards. These included the python libraries "bokeh"[214] or "plotly"[215] and the associated "dash" library. Their close integration with powerful javascript libraries allowed for a smooth visualization in web browsers. When coupled with javascript-based interactive chemical structure viewers, exploration of chemical datasets can easily be performed by mouse navigation. We used this principle in our implementation of a viewer tool,[216] see Figure 5.2, designed to work with the "ASAP" code that provides a framework to generate structure maps.[210] It can visualize data of periodic materials as well as molecular structures, allowing also to focus on the visualization of specific atomic environments in these structures. For completeness, I want to reference some examples of other (now) available visualization tools here,[217–219] while some powerful workflow tools (i.e. freely accessible ones[220,221]) also integrate functionality.

6 | Publications

This chapter provides an overview of articles that have been published by my coauthors and me during the work on this dissertation. I thereby restrict myself to those that are of main relevance to this dissertation. For each publication a short summary is provided with a listing of my detailed contributions. The publications are ordered thematically.

6.1 Finding the Right Bricks for Molecular Legos: A Data Mining Approach to Organic Semiconductor Design



Christian Kunkel, Christoph Schober, Johannes T. Margraf, Karsten Reuter, and Harald Oberhofer

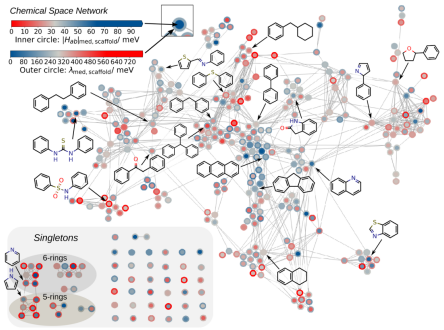
Chemistry of Materials **2019**, *31*, 969–978

Summary: The project of data-driven organic semiconductor discovery resulted from Christoph Schober's work on organic semiconductor virtual screening, see chapter 3. This article is a direct follow-up, analyzing the molecular structures, contained in the 64k-dataset , with the aim of establishing guiding principles that potentially improve upon charge carrier mobilities. Topically the article fitted the time, as other fields of materials science also started to extract general design criteria by data mining or even machine learning approaches from their available large-scale datasets. Approaches targeting organic semiconductors (OSCs) had however been limited to a much smaller scale and molecular diversity.

We therefore combined statistical tools of data mining and molecular analysis to uncover guiding principles for organic semiconductor materials design. Our approach focused on commonalities among the contained molecular structures, and we herein investigate how they relate to the charge-transport descriptors contained in the dataset. One design aspect of the study was thereby to make the analysis chemically intuitive and comprehensible. The analysis was hence undertaken with a rule-based molecular decomposition scheme, partitioning a molecule into scaffold and side groups — concepts well-known from organic chemistry. In this context, we rely on the established scaffold definition by Bemis and Murcko (BM)[167] where the molecular scaffold is defined as the molecular core comprised of connected ring systems and their linkers, while side group atoms are defined as branching off from it. Being able to carry out the decomposition in an automated fashion, we then clustered molecules by common scaffolds and side groups. For the 195 considered scaffold clusters comprising close to 7000 molecules and crystals, we obtained statistically significant performance differences, identifying scaffolds that are favorable for charge transport. A range of identified side groups then generally lowers the reorganization energy, meaning that functionalizing promising scaffolds with favorable side groups can result in improved charge-transport properties, directly suggesting a promising design criterion.

Individual contributions: The 64k-dataset was kindly provided by Christoph Schober. Based on the dataset I carried out the BM-scaffold analysis using the contained $|H_{\mathrm{ad}}|$ and λ_h descriptor data. Due to the staged screening-workflow applied by Christoph Schober, not all λ_h values were directly available for the BM-clustered data-subset analyzed here, and I carried out the necessary additional descriptor calculations. The final manuscript was then jointly written by Johannes T. Margraf, Harald Oberhofer and Karsten Reuter and myself.

6.2 Knowledge discovery through chemical space networks: the case of organic electronics



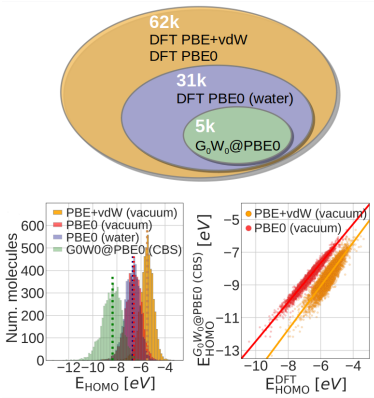
Christian Kunkel, Christoph Schober, Harald Oberhofer, and Karsten Reuter

Journal of Molecular Modeling **2019**, *25:87*

Summary: Following up on our previous work, we now provided a visually comprehensive analysis of the design space in the 64k-dataset. As before, we focused on a scaffold-clustered dataset. As a visualization technique, we then used a so-called chemical space network (CSN), treating every BM scaffold cluster as a node, while edges are inserted between nodes based on detected pairwise scaffold similarity. For pairwise similarity detection, four different methods were employed that capture different aspects of similarity between scaffolds. With an algorithmic layout, a map of the chemical landscape is created, which embeds molecular scaffold clusters into an environment of similar topologies. To analyze this global map as well as the local communities of closely related analogues occurring in it, we included the median descriptor values of each cluster in the CSN visualization by a color-code. From a close inspection of the generated map we find that it not only visually reproduced known trends for good organic semiconductors, but further allowed for a visual extraction of design rules, as we illustrate in selected examples. A crowded local cluster was e.g. found to represent the acene family of structures, well-known as organic semiconductors and correspondingly enriched in favorable charge-transport descriptor values. On a narrower scale, effects of scaffold extension or hetero-atom exchange are directly visible when following along the gradual progression in design space that is by construction integrated in the CSN representation. We further identify local environments of scaffolds where clusters with promising descriptor values reside. Some of these show little or no connection to the sampled chemical space, and potential for further exploration. In combination with a created browser-based tool that we extensively used for interactive visualization during the writing of the manuscript, we found CSNs to be a useful tool for materials design and organic semiconductor discovery.

Individual contributions: Following the initial idea of Karsten Reuter, I developed the chemical space network visualization and performed the analysis. The dataset was again based on the 64k-dataset of Christoph Schober. I also carried out additionally necessary λ_{h} descriptor calculations. The manusscript was jointly written and edited by all authors.

6.3 Atomic structures and orbital energies of 61,489 crystal-forming organic molecules



Annika Stuke*, Christian Kunkel*, Dorothea Golze, Milica Todorović, Johannes T. Margraf, Karsten Reuter, Patrick Rinke and Harald Oberhofer

Scientific Data **2020**, *7*, 58

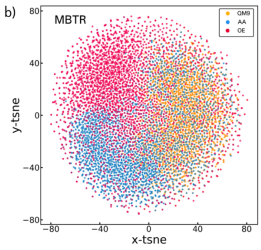
* These authors contributed equally to the work.

Summary: While working on this dissertation, methods for molecular property prediction by machine learning have seen a significant development. This is not surprising as they generally promise a simple and computationally cheap screening of vast molecular candidate spaces during molecular discovery projects. Methodological development of quantum chemical property prediction had however often relied on smaller molecules, using e.g. the well-established QM9-dataset, containing small molecules composed of H and up to nine heavy atoms (C, N, O, and F only). We noticed, that the technologically relevant molecular structures contained in the 64k-dataset are larger and of much higher diversity – being composed of up to 92 non-hydrogen atoms and 16 different elements, while containing extended hetero aromatic backbones and attached functional groups. Being mostly interested in such larger structures, we started a collaboration with colleagues from Aalto-University to work on this problem. In this contribution, we hence used the 64k-dataset as a starting point to produce a high-quality reference data set for quantum chemical property prediction of large molecules. For this so-called OE62-dataset, 61,489 unique organic molecular structures were extracted from the respective organic crystals of the 64k-dataset . These geometries were then first relaxed in vacuum using van-der Waals corrected density-functional theory (DFT) at the PBE level of theory. Based on these equilibrium structures, the OE62-dataset then supplies vacuum total energies, partial charges and orbital eigenvalues at the PBE and the computationally more demanding PBE hybrid (PBE0) level of DFT for all 62 k molecules. Further, the PBE0 level values in (implicit) water are included for a subset of 30,876 molecules. At the most expensive computational level, the dataset provides quasi-particle energies for 5,239 molecules in vacuum computed with many-body perturbation theory in the G0W0 approximation extrapolated to the complete basis set (CBS) limit, while using PBE0 as a starting point. For the different levels of theory we provide a technical validation including information on numerical accuracy. The multi-level computational results summarized in this new and freely available dataset can now be used to develop and evaluate machine learning algorithms.

Individual contributions: The initial idea was conceived by Karsten Reuter, Patrick Rinke and Harald Oberhofer. Annika Stuke and me jointly curated the data and later postprocessed the results. I thereby performed the calculations at the DFT-levels of theory. Calculations at the G0W0 level of theory were conducted by Annika Stuke and Dorothea Golze. The manuscript was cowritten by all authors.

Additional remarks: To illustrate use-cases of the dataset, I here want to mention publications that already applied it in molecular machine learning, limited to those two publications I was personally involved in:

- ***Chemical diversity in molecular orbital energy predictions with kernel ridge regression***

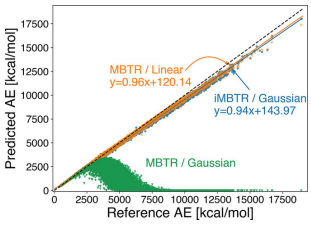


Annika Stuke, Milica Todorović, Matthias Rupp, Christian Kunkel, Kunal Ghosh, Lauri Himanen, and Patrick Rinke

The Journal of Chemical Physics **2019**, *150*, 204121

Summary: Our colleagues from Aalto University tested the performance of kernel ridge regression (KRR) for molecular machine learning for highest occupied molecular orbital (HOMO) prediction. In detail, the study provides a comparative benchmark on three large datasets of different composition and scope: Small organic molecules are included in the QM9-dataset. A second dataset contains amino acid and dipeptide conformers. Thirdly, the PBE-subset of OE62 supplied large and diverse molecules. In addition, two different representations that encode the molecular structure were compared. While finding that HOMO energy predictions on unseen molecules are possible for any dataset, the overall accuracy varied among them, influenced also by the chosen molecular representation. Remarkably, the justifiably most diverse and unevenly sampled OE62-dataset posed the biggest challenge to accuracy, rendering it a suitable dataset to develop molecular machine learning methods further.

- ***Size-Extensive Molecular Machine Learning with Global Representations***



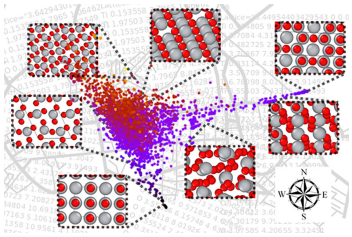
Hyunwook Jung*, Sina Stocker*, Christian Kunkel, Harald Oberhofer, Byungchan Han, Karsten Reuter, and Johannes T. Margraf

ChemSystemsChem **2020**, *2*, e1900052

* These authors contributed equally to the work.

In this contribution, we illustrate the issue of "size-extensivity" in so-called global molecular representations, used to encode molecules for machine learning property prediction. "Size-extensivity" thereby refers to a dependence of the target property on an (increasing) molecular size, and should be accordingly treated in a devised molecular machine-learning model. To illustrate this point we exploited, that the QM9- and OE62-datasets show a widely differing size-distribution among constituent molecules. We here again use KRR, but now focus on size-extensive molecular atomization energies. We show that non size-extensive models are only useful in the range of their training set, meaning a model trained on QM9 data can hardly predict values for the larger molecules of OE62. Building size-extensivity into the global molecular representations then already provides reasonable predictions across large size differences, allowing for model training on QM9- and prediction on OE62 molecules. Sources of error for extensive models however remain as discussed in the article.

6.4 Mapping materials and molecules



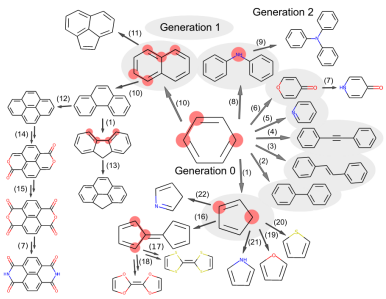
Bingqing Cheng, Ryan-Rhys Griffiths, Simon Wengert, Christian Kunkel, Tamas Stenczel, Bonan Zhu, Volker L. Deringer, Noam Bernstein, Johannes T. Margraf, Karsten Reuter, and Gábor Csányi
Accounts of Chemical Research **2020**, *53*, 1981-1991

Summary: In our earlier work on the 64k-dataset , we had already seen how a well-performed large-dataset visualization could contribute to an understanding of its inner structure. We now teamed up with colleagues from Oxford, Cambridge and, Washington to provide an easy-to-read perspective of the developing field of molecular or materials visualization. For the contribution, we mainly relied on a description of atomic environments present in a dataset by a Smooth Overlap of Atomic Positions (SOAP) descriptor. Universally applicable hyperparameters for the SOAP descriptor are thereby also supplied with the article, creating a standard for visualization. Using this high-dimensional descriptor, atomic environments can be easily compared and a continuous structural similarity can be assigned. The concept can be easily extended to entire molecules or crystal structures, e.g. by computing an averaged descriptor from it. A second ingredient to the visualization of a respectively encoded structural dataset is then the computation of a low dimensional representation. To visualize the relationship between structures, this representation should well-preserve local structural similarities among the single data-points in a low-dimensional map, while providing a global overview of dataset structure as well. While multiple techniques can be applied, we here mainly relied on principal component analysis (PCA) and a kernelized version of it (KPCA). Examples illustrate the outcome of the workflow, among others comprising mappings of amorphous carbon local environments, liquid water structure, crystal structures of titanium dioxide, organic crystal structures, or of organic molecular databases. The use cases for the different mappings are thereby as diverse as the corresponding datasets. They include analysis of differing atomic environments present in a dataset and their relation to chemical reactivity and stability, classification of differing crystal structures resulting from automated structure search, diversity and compositional analysis of molecular datasets, or the analysis of molecular dynamics trajectories. Upon analyzing such diverse datasets, we thereby found that valuable information can often readily be extracted by navigating interactively in the corresponding maps. This renders them a valuable standard tool for the computational chemistry and materials science community in the future. Finally, all examples discussed in the article can be easily reproduced by the community, as data and an automated-framework for visualizing and analyzing such structural data sets (the ASAP package) are provided in repositories stored on github. We also implemented a complementary browser-based visualization tool for interactive exploration of the generated maps.

Individual contributions: The idea for this article was a result of intense discussion with our collaborators at workshops, and also within our group. All people involved in the presented article work with machine learning on larger data sets and the idea of working on scientific visualization thereby mostly grew out of the need to perform analyses of the various datasets that were around by the time in the different groups. I hence want to stress, that everyone involved contributed fruitful ideas, datasets, code or sections of the article, and we jointly improved on them. This process cannot and should not fully be delimited. Nevertheless, my contributions were on the

one hand a browser-based visualizer tool for chemical structures and atomic environments that I had started to work on quite early and that was then finalized together with Simon Wengert and Tamas Stenczel. Its published code is referenced at the end of the article and the tool was used by Simon Wengert and me to directly produce parts of the figures in the article. I also contributed a section on organic molecular visualization using the example of the QM9-dataset, writing the corresponding part in the article.

6.5 Active Discovery of Organic Semiconductors



Christian Kunkel, Johannes T. Margraf, Ke Chen, Harald Oberhofer, and Karsten Reuter
Nature Communications, ***2021****, 12, 2422*

Summary: As we have seen before, virtual screening approaches have already uncovered a larger number of favorable candidates for high-performing organic semiconductors. Data-mining strategies applied to the resulting databases then unveiled interesting design principles. Nevertheless, the design space that could be spawned by the combination of even a limited number of fragments is virtually unlimited. This clearly dictates design strategies that go beyond exhaustive virtual screening of candidates, relying on their smart selection for effective use of computational resources. To devise such a strategy we apply active machine learning (AML) concepts for the efficient discovery of p-type organic semiconductor candidates. A versatile set of molecular construction rules thereby spanned a searchable molecular design space of flexible, π-conjugated candidates. The usefulness of candidates can then be assessed by a combination of charge-injection and mobility related molecular descriptors. The multi objective problem of molecular discovery in this design space was then efficiently solved by AML, in which an ever improving surrogate model of the underlying property surface is used to judge the usefulness of molecular candidates and advance the search. During runtime, the devised algorithm can thereby actively exploit knowledge about favorable candidates and query explicit descriptor calculation on them, balanced with explorative queries that are used to gain knowledge about molecules with an uncertain outcome. Computational resources were thereby employed exceptionally well, leading to a significant acceleration of discovery over random searches or commonly employed computational funnels. This scheme is further applicable even in a virtually unlimited molecular space, not requiring the full exhaustive enumeration of all structures. Methodological insight was first gained on a fully enumerated, but limited molecular test space of > 65.000 molecules annotated with cheaply computed molecular descriptors. In this space, a simple realization of the algorithm discovered up to 85 % of the known 2438 favorable candidates after querying descriptor calculation on 5179 molecules. Application of this algorithm to a virtually unlimited molecular design space, evaluated with DFT-B3LYP calculations lead to the discovery of 900 favorable and diverse molecules among 1680 processed ones, a relative success rate of > 50 %. The AML-discovery strategy was, therefore, able to drive efficient and continuous autonomous discovery of promising OSC materials in a design space with initially unknown structure-property relationships.

Individual contributions: The idea was jointly conceived by Johannes T. Margraf, Harald Oberhofer, Karsten Reuter and myself. I implemented and executed the algorithms for molecular space enumeration and AML discovery on the high-performance computing hardware. Methodological details were worked out in conjunction with Ke Chen and Johannes T. Margraf. Harald Oberhofer, Johannes T. Margraf and Karsten Reuter and me wrote the manuscript.

6.6 Further work

The following articles have appeared during my time of working at the Chair of Theoretical Chemistry or are in preparation. The two first ones are topically related to our work on OSCs, containing methodology on the computation and use of electronic coupling values. Others have no direct relation to the topic but are nevertheless related to the field of computational chemistry. Since these articles don't form essential parts of the dissertation, they haven't been included, and I only want to mention them here for completeness and future reference.

- ***Electronic property trends of single-component organic molecular crystals containing C, N, O, and H.***
  Steven Monaco, Ryan Baer, Ryan Giernackya, Miguel Villalbaa, Taylor Garcia, Carlos Mora-Perez, Spencer Brady, Kris Erlitz, Christian Kunkel, Sebastian Jezowskia, Harald Oberhofer, Carsten Lange, and Bohdan Schatschneider
  submitted

- ***Improved projection-operator diabatization schemes for the calculation of electronic coupling values***
  Simiam Ghan, Christian Kunkel, Karsten Reuter, and Harald Oberhofer
  Journal of Chemical Theory and Computation, **2020**, *16*, 7431–7443

- ***Anomalous pressure dependence of the electronic properties of molecular crystals explained by changes in intermolecular electronic coupling***
  Maituo Yu, Xiaopeng Wang, Xiong-Fei Du, Christian Kunkel, Taylor M. Garcia, Stephen Monaco, Bohdan Schatschneider, Harald Oberhofer, and Noa Marom
  Synthetic Metals, **2019**, *253*, 9-19

- ***Towards Density Functional Approximations from Coupled Cluster Correlation Energy Densities***
  Johannes T. Margraf, Christian Kunkel, and Karsten Reuter
  The Journal of Chemical Physics, **2019**, *150*, 244116

- ***Generalized molecular solvation in non-aqueous solutions by a single parameter implicit solvation scheme***
  Christoph Hille, Stefan Ringe, Martin Deimel, Christian Kunkel, William E. Acree, Karsten Reuter, and Harald Oberhofer
  The Journal of Chemical Physics, **2019**, *150*, 041710

- ***Surface Activity of Early Transition Metal Oxycarbides: CO_2 Adsorption Case Study***
  Christian Kunkel, Francesc Viñes, and Francesc Illas
  The Journal of Physical Chemistry C, **2019**, *123*, 3664-3671

7 | Conclusion and Outlook

Organic semiconducting devices have undergone significant development over the last years, leading to their commercialization. Nevertheless, performance increases could still lead to a further boost in market penetration and allow for new applications. Especially in electronic devices, charge conductivity is the decisive property, determining e.g. the performance of organic field-effect devices (OFETs). While materials have been gradually improving, the richness of possible molecular structures and their complex structure-property relationships make it difficult to identify optimal ones solely based on labor-intensive experimentation. Guided by well-characterized theory-derived charge-conductivity descriptors, in-silico methods relying on computational simulation have a large potential in accelerating this process. In fact, the theoretical description of charge transport in OSC materials has also evolved significantly. New insight suggests that improved OSC materials, with charge-transport properties less dominated by disorder, could in the future reach the suspected upper-limits to materials performance.

In this work, we, therefore, explored the optimization of charge-transport related properties through a data-driven approach. In the first part, we examined a large-scale 64k-dataset of organic crystals available to us through the work of Schober et al.[30,39] and annotated with charge-transport descriptors. The descriptors were the electronic coupling $|H_{\mathrm{ad}}|$ related to molecular dimer geometries occurring in the crystal, as well as the reorganization energy λ, related to molecular structures. Data-mining and visualization strategies then revealed favorable relations of certain molecular structures with charge-transport descriptors, and these could be used to propose new combinations and hence molecules. So far limited to and biased by the availability of experimental organic crystals and their computational annotation, correlations between structural elements (i.e. functionalization of certain scaffolds with side groups at selected positions) are not yet fully captured in the data-mining approach, influenced here also by the specific molecular decomposition scheme chosen for analysis.

Extending the search to even larger organic materials spaces, the workhorse method – density functional theory (DFT) seems computationally too expensive for an exhaustive screening. Machine learning applied to the prediction of molecular properties hence emerged as a computationally more tractable method to (partially) tackle this problem. By the time of writing, it had however routinely been tested on small molecules and rarely been applied to molecular sizes of technological interest. We, therefore, made the 64k-dataset publicly available as a complicated benchmark dataset for molecular machine learning – the OE62-dataset. The work of our collaborators on this dataset demonstrated that the performance of a common machine learning technique for the prediction of orbital energies then inherently decreases with the increasing complexity represented by this dataset as compared to other datasets. Nevertheless, switching the property to be predicted to the atomization energy and adjusting the underlying machine learning method to incorporate "size-extensivity" helped to greatly improve predictions across different

datasets and large size differences as was shown by us. This part of the work thus demonstrated that molecular machine learning methods can still be significantly improved to generalize to the vastness of organic chemistry. On the other hand and if possible for the molecular design problem at hand, the narrowing of chemical space to a more targeted, less diverse design space can make the problem easier to tackle.

From thereon, we revisited the OSC design problem with the help of molecular machine learning and visualization. This time, and to circumvent the reliance on an exhaustively screened dataset we finally employed an active machine learning (AML) feedback loop, training a machine learning model on available data, while using it to guide the search to new prospective candidates. The subsequently accumulating computational results then lead to an ever-improving selection strategy of the oncoming computations. Accordingly, the method is here employed to search a combinatorially vast molecular space, determined from a set of chemical transformations from which typical OSC candidate molecules can be generated. Molecules exhibiting a favorable balance of charge-conductivity related descriptors are then automatically searched for with a continuous refinement allowing to most efficiently focus on promising parts of the chemical space.

The challenges for the ongoing development of autonomous OSC discovery seem to be an extension to more diverse chemical spaces, and a search by refined models. A third avenue is the improvement of OSC candidate fitness evaluation, which was here so far limited to two molecular electronic properties, neglecting the solid-state arrangement. Feedback on actual carrier dynamics[87] and solid-state properties such as charge-transport networks should be incorporated, probably employing refined models e.g. derived from dimer, crystalline- or amorphous phase models.[31,32,34,222,223] The computational savings of the AML discovery strategy could correspondingly be invested in a finer candidate evaluation. Adding a solid-state description in the evaluation function might then also influence the molecular selection made by the current AML methodology – so far strongly prioritizing larger molecules. Their larger combinatorial availability, as well as a favorable scaling of λ with this larger size, might be a cause. The influence of dynamic disorder[81,86] or interfaces with electrode materials could additionally (and occasionally) be requested[224] during AML, augmented by (automated) occasional experimental evaluation.[225]

8 | Acknowledgments / Danksagung

First, I want to thank my supervisors Prof. Dr. Karsten Reuter and Dr. Harald Oberhofer for the opportunity to carry out this work. I especially want to thank them for academic and strategic guidance, scientific discussion and possibilities to visit (inter)national conferences and collaborators. I also want to thank Dr. Johannes Margraf for our ongoing exchange about projects and science in general.

Moreover, I want to thank the whole group for the educative and enjoyable experience we had together. Before Corona, many things happened, including group retreats (special thanks to the organizers Sina, David, Mitch and Cristina), DPG spring meetings, international food evenings, sailing trips at Ammersee, and PhD parties. Coronavirus changed many things, but virtual get-togethers, journal clubs and conferences made it worthwhile. I also want to thank the admin-team - Martin, Simon, Simeon, Simiam, Xhristoph, Matthias and David for their hard work and interesting discussion about computing. This, of course, extends to Christoph Scheurer who kept impressing us with his broad knowledge. Teaching the legendary python course was also great fun and I especially want to thank my colleagues Xristoph and Simeon. Ruth Mösch and Julia Pach also deserve a big thank you from everyone for their help in battling bureaucracy.

I am especially grateful for our welcoming, international and progressive group spirit. I want to thank Cristina for making pasta for us, Jakob T. for being a great guide to Munich- and Berlin nightlife, and Markus and Mitch for karaoke parties at the Ismaning-mansion. I also want to thank Simiam, for shaping Munich's cultural scene during Corona and for being a great companion for occasional weekends at the office. Last, not least I want to thank Patrick Gütlein and David Egger for our Jazz evenings, tennis and other events!

I am also happy to have worked with so many collaborators outside of the Chair of Theoretical Chemistry. I sincerely want to thank Annika Stuke, Kunal Ghosh, Dorothea Golze, Milica Todorovic, Patrick Rinke, Bingqing Cheng, Tamas Stenczel, Gábor Csányi, Noa Marom and Bohdan Schatschneider for great work, advice and enlightening discussions. I especially want to thank Prof. Dr. Patrick Rinke for hosting an interesting and motivating visit at Aalto University in Mai 2018 and for introducing me to floorball. I'm also grateful to Gábor Csányi for hosting an intended visit to the University of Cambridge in Mai/June 2020, that due to Corona we sadly had to cancel. Further, I want to thank the DAAD for a Travel Grant to the ACS National Meeting in San Diego 2019. The whole trip was one of the enduring and motivating experiences during the PhD and I want to thank the whole crew again (Xristoph, Jakob F., Cristina, Matthias, Simiam, David and Patrick) for interesting talks, good vibes, nice cars and lasting memories.

The Solar Technologies Go Hybrid (SolTech) initiative is acknowledged for providing an interesting scientific environment. In this regard, I want to thank Sebastian Grott, my fellow TUM.Solar PhD representative who loved the SolTech PhD events as much as I did. Computing time at the LRZ,

the MPCDF and the ALCF was highly appreciated. Support from the International Graduate School of Science and Engineering (IGSSE) of TUM is further acknowledged.

Good friends made living in Munich even more worthwhile. Thanks, Ulrich for motivating me to go to Peru, Bolivia and Malaysia. Thank you, Felix and Dominik, for techno parties and rooftop-meetings. Thank you, Kristof, for the nice atmosphere we had in the flat at Münchner Freiheit. Thanks to Timo and Martine for exploring many restaurants with us. Greetings also to Würzburg (Tobias, Sebbo G., Sebbo S. and Sophia), and to Simon and Sina in Berlin, and sorry, that I was "so busy" at times. I also want to thank my favourite bars around Münchner Freiheit and the coffee shops I misused as an office while clinging to my cup of coffee. Greetings to all the other regulars that did the same.

In the end, I sincerely want to thank my family. My parents, who always did a great job. They believed in us, made everything possible, and were always there when a helping hand or advice was needed. I also want to thank my brother and my grandparents who were always encouraging, supporting and in general an important part of my life. Finally, I want to thank Marta for her love, patience and support.

Bibliography

[1] Li, F.; Nathan, A.; Wu, Y.; Ong, B. *Organic Thin Film Transistor Integration: A Hybrid Approach*; Wiley, 2011.

[2] Köhler, A.; Bässler, H. *Electronic Processes in Organic Semiconductors: An Introduction*; Wiley, 2015.

[3] Xu, R.-P.; Li, Y.-Q.; Tang, J.-X. Recent advances in flexible organic light-emitting diodes. *Journal of Materials Chemistry C* **2016**, *4*, 9116–9142.

[4] Das, R.; He, X.; Ghaffarzadeh, K. Flexible, Printed and Organic Electronics 2020-2030: Forecasts, Technologies, Markets.
https://www.idtechex.com/de/research-report/flexible-printed-and-organic-electronics-2020-2030-forecasts-technologies-markets/687,
[Online; accessed 21-October-2020].

[5] Lin, Y.; Li, Y.; Zhan, X. Small molecule semiconductors for high-efficiency organic photovoltaics. *Chemical Society Reviews* **2012**, *41*, 4245–4272.

[6] Liu, D. et al. Organic Laser Molecule with High Mobility, High Photoluminescence Quantum Yield, and Deep-Blue Lasing Characteristics. *Journal of the American Chemical Society* **2020**, *142*, 6332–6339.

[7] Myny, K.; Steudel, S.; Smout, S.; Vicca, P.; Furthner, F.; van der Putten, B.; Tripathi, A.; Gelinck, G.; Genoe, J.; Dehaene, W.; Heremans, P. Organic RFID transponder chip with data rate compatible with electronic product coding. *Organic Electronics* **2010**, *11*, 1176 – 1179.

[8] Gelinck, G.; Heremans, P.; Nomoto, K.; Anthopoulos, T. D. Organic Transistors in Optical Displays and Microelectronic Applications. *Advanced Materials* **2010**, *22*, 3778–3798.

[9] Zhang, C.; Chen, P.; Hu, W. Organic field-effect transistor-based gas sensors. *Chem. Soc. Rev.* **2015**, *44*, 2087–2107.

[10] Tebyetekerwa, M.; Marriam, I.; Xu, Z.; Yang, S.; Zhang, H.; Zabihi, F.; Jose, R.; Peng, S.; Zhu, M.; Ramakrishna, S. Critical insight: challenges and requirements of fibre electrodes for wearable electrochemical energy storage. *Energy and Environmental Science* **2019**, *12*, 2148–2160.

[11] Wang, S. et al. Skin electronics from scalable fabrication of an intrinsically stretchable transistor array. *Nature* **2018**, *555*, 83–88.

[12] Takeda, Y.; Hayasaka, K.; Shiwaku, R.; Yokosawa, K.; Shiba, T.; Mamada, M.; Kumaki, D.; Fukuda, K.; Tokito, S. Fabrication of Ultra-Thin Printed Organic TFT CMOS Logic Circuits Optimized for Low-Voltage Wearable Sensor Applications. *Scientific Reports* **2016**, *6*, 25714.

[13] Yavuz, I. Dichotomy between the band and hopping transport in organic crystals: insights from experiments. *Physical Chemistry Chemical Physics* **2017**, *19*, 25819–25828.

[14] Allen, F. H. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallographica Section B* **2002**, *58*, 380–388.

[15] Schweicher, G.; Garbay, G.; Jouclas, R.; Vibert, F.; Devaux, F.; Geerts, Y. H. Molecular Semiconductors for Logic Operations: Dead-End or Bright Future? *Advanced Materials* **2020**, *32*, 1905909.

[16] Wang, L.; Nan, G.; Yang, X.; Peng, Q.; Li, Q.; Shuai, Z. Computational methods for design of organic materials with high charge mobility. *Chemical Society Reviews* **2010**, *39*, 423–434.

[17] Wang, C.; Dong, H.; Hu, W.; Liu, Y.; Zhu, D. Semiconducting -Conjugated Systems in Field-Effect Transistors: A Material Odyssey of Organic Electronics. *Chemical Reviews* **2012**, *112*, 2208–2267.

[18] Ostroverkhova, O. Organic optoelectronic materials: mechanisms and applications. *Chemical Reviews* **2016**, *116*, 13279–13412.

[19] Wang, C.; Dong, H.; Jiang, L.; Hu, W. Organic semiconductor crystals. *Chemical Society Reviews* **2018**, *47*, 422–500.

[20] Wang, Y.; Sun, L.; Wang, C.; Yang, F.; Ren, X.; Zhang, X.; Dong, H.; Hu, W. Organic crystalline materials in flexible electronics. *Chemical Society Reviews* **2019**, *48*, 1492–1530.

[21] Mei, J.; Diao, Y.; Appleton, A. L.; Fang, L.; Bao, Z. Integrated Materials Design of Organic Semiconductors for Field-Effect Transistors. *Journal of the American Chemical Society* **2013**, *135*, 6724–6746.

[22] Polishchuk, P. G.; Madzhidov, T. I.; Varnek, A. Estimation of the size of drug-like chemical space based on GDB-17 data. *Journal of Computer-Aided Molecular Design* **2013**, *27*, 675–679.

[23] Wilbraham, L.; Smajli, D.; Heath-Apostolopoulos, I.; Zwijnenburg, M. A. Mapping the optoelectronic property space of small aromatic molecules. *Communications Chemistry* **2020**, *3*, 14.

[24] Geng, H.; Niu, Y.; Peng, Q.; Shuai, Z.; Coropceanu, V.; Brédas, J.-L. Theoretical study of substitution effects on molecular reorganization energy in organic semiconductors. *Journal of Chemical Physics* **2011**, *135*, 104703.

[25] Agrawal, A.; Choudhary, A. Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science. *APL Materials* **2016**, *4*, 53208.

[26] Sokolov, A. N.; Atahan-Evrenk, S.; Mondal, R.; Akkerman, H. B.; Sánchez-Carrera, R. S.; Granados-Focil, S.; Schrier, J.; Mannsfeld, S. C. B.; Zoombelt, A. P.; Bao, Z.; Aspuru-Guzik, A. From computational discovery to experimental characterization of a high hole mobility organic crystal. *Nature Communications* **2011**, *2*, 437.

[27] Gómez-Bombarelli, R. et al. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nature Materials* **2016**, *15*, 1120.

[28] Olivares-Amaya, R.; Amador-Bedolla, C.; Hachmann, J.; Atahan-Evrenk, S.; Sanchez-Carrera, R. S.; Vogt, L.; Aspuru-Guzik, A. Accelerated computational discovery of high-performance materials for organic photovoltaics by means of cheminformatics. *Energy and Environmental Science* **2011**, *4*, 4849–4861.

[29] Kanal, I. Y.; Owens, S. G.; Bechtel, J. S.; Hutchison, G. R. Efficient Computational Screening of Organic Polymer Photovoltaics. *The Journal of Physical Chemistry Letters* **2013**, *4*, 1613–1623.

[30] Schober, C.; Reuter, K.; Oberhofer, H. Virtual Screening for High Carrier Mobility in Organic Semiconductors. *The Journal of Physical Chemistry Letters* **2016**, *7*, 3973–3977.

[31] 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. *The Journal of Physical Chemistry C* **2017**, *121*, 28249–28261.

[32] Yang, J.; De, S.; Campbell, J. E.; Li, S.; Ceriotti, M.; Day, G. M. Large-Scale Computational Screening of Molecular Organic Semiconductors Using Crystal Structure Prediction. *Chemistry of Materials* **2018**, *30*, 4361–4371.

[33] Atahan-Evrenk, S.; Atalay, F. B. Prediction of Intramolecular Reorganization Energy Using Machine Learning. *The Journal of Physical Chemistry A* **2019**, *123*, 7855–7863.

[34] Matsuzawa, N. N.; Arai, H.; Sasago, M.; Fujii, E.; Goldberg, A.; Mustard, T. J.; Kwak, H. S.; Giesen, D. J.; Ranalli, F.; Halls, M. D. Massive Theoretical Screen of Hole Conducting Organic Materials in the Heteroacene Family by Using a Cloud-Computing Environment. *The Journal of Physical Chemistry A* **2020**, *124*, 1981–1992.

[35] Nematiaram, T.; Padula, D.; Landi, A.; Troisi, A. On the Largest Possible Mobility of Molecular Semiconductors and How to Achieve It. *Advanced Functional Materials* **2020**, 2001906.

[36] Gryn'ova, G.; Lin, K.-H.; Corminboeuf, C. Read between the Molecules: Computational Insights into Organic Semiconductors. *Journal of the American Chemical Society* **2018**, *140*, 16370–16386.

[37] Friederich, P.; Fediai, A.; Kaiser, S.; Konrad, M.; Jung, N.; Wenzel, W. Toward Design of Novel Materials for Organic Electronics. *Advanced Materials* **2019**, *31*, 1808256.

[38] Saeki, A.; Kranthiraja, K. A high throughput molecular screening for organic electronics via machine learning: present status and perspective. *Japanese Journal of Applied Physics* **2019**, *59*, SD0801.

[39] Schober, C. O. Ab Initio Charge Carrier Mobility and Computational Screening of Molecular Crystals for Organic Semiconductors. Dissertation, Technische Universität München, München, 2017.

[40] 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. *Chemistry of Materials* **2019**, *31*, 969–978.

[41] Kunkel, C.; Schober, C.; Oberhofer, H.; Reuter, K. Knowledge discovery through chemical space networks: the case of organic electronics. *Journal of Molecular Modeling* **2019**, *25*, 87.

[42] Stuke, A.; Kunkel, C.; Golze, D.; Todorovic, M.; Margraf, J. T.; Reuter, K.; Rinke, P.; Oberhofer, H. Atomic structures and orbital energies of 61,489 crystal-forming organic molecules. *Scientific Data* **2020**, *7*, 58.

[43] Stuke, A.; Todorović, M.; Rupp, M.; Kunkel, C.; Ghosh, K.; Himanen, L.; Rinke, P. Chemical diversity in molecular orbital energy predictions with kernel ridge regression. *The Journal of Chemical Physics* **2019**, *150*, 204121.

[44] Jung, H.; Stocker, S.; Kunkel, C.; Oberhofer, H.; Han, B.; Reuter, K.; Margraf, J. T. Size-Extensive Molecular Machine Learning with Global Representations. *ChemSystemsChem* **2020**, *1900052*, syst.201900052.

[45] Cheng, B.; Griffiths, R.-R.; Wengert, S.; Kunkel, C.; Stenczel, T.; Zhu, B.; Deringer, V. L.; Bernstein, N.; Margraf, J. T.; Reuter, K.; Csanyi, G. Mapping Materials and Molecules. *Accounts of Chemical Research* **2020**, *53*, 1981–1991.

[46] Brütting, W.; Adachi, C. *Physics of Organic Semiconductors*; Wiley, 2012.

[47] Kordt, P.; van der Holst, J. J. M.; Al Helwi, M.; Kowalsky, W.; May, F.; Badinski, A.; Lennartz, C.; Andrienko, D. Modeling of Organic Light Emitting Diodes: From Molecular to Device Properties. *Advanced Functional Materials* **2015**, *25*, 1955–1971.

[48] Paterson, L.; May, F.; Andrienko, D. Computer aided design of stable and efficient OLEDs. *Journal of Applied Physics* **2020**, *128*, 160901.

[49] Kippelen, B.; Brédas, J.-L. Organic photovoltaics. *Energy and Environmental Science* **2009**, *2*, 251–261.

[50] Zhugayevych, A.; Tretiak, S. Theoretical Description of Structural and Electronic Properties of Organic Photovoltaic Materials. *Annual Review of Physical Chemistry* **2015**, *66*, 305–330.

[51] Hedley, G. J.; Ruseckas, A.; Samuel, I. D. W. Light Harvesting for Organic Photovoltaics. *Chemical Reviews* **2017**, *117*, 796–837.

[52] Braga, D.; Horowitz, G. High-Performance Organic Field-Effect Transistors. *Advanced Materials* **2009**, *21*, 1473–1486.

[53] Lamport, Z. A.; Haneef, H. F.; Anand, S.; Waldrip, M.; Jurchescu, O. D. Tutorial: Organic field-effect transistors: Materials, structure and operation. *Journal of Applied Physics* **2018**, *124*, 71101.

[54] Uemura, T.; Rolin, C.; Ke, T.-H.; Fesenko, P.; Genoe, J.; Heremans, P.; Takeya, J. On the Extraction of Charge Carrier Mobility in High-Mobility Organic Transistors. *Advanced Materials* **2016**, *28*, 151–155.

[55] Podzorov, V. Organic single crystals: Addressing the fundamentals of organic electronics. *MRS Bulletin* **2013**, *38*, 15–24.

[56] Sawatzki, M. F.; Kleemann, H.; Boroujeni, B. K.; Wang, S.-J.; Vahland, J.; Ellinger, F.; Leo, K. Doped Highly Crystalline Organic Films: Toward High-Performance Organic Electronics. *Advanced Science n/a*, 2003519.

[57] Coropceanu, V.; Cornil, J.; da Silva Filho, D. A.; Olivier, Y.; Silbey, R.; Brédas, J. L. Charge transport in organic semiconductors. *Chemical Reviews* **2007**, *107*, 926–952.

[58] Stehr, V.; Pfister, J.; Fink, R. F.; Engels, B.; Deibel, C. First-principles calculations of anisotropic charge-carrier mobilities in organic semiconductor crystals. *Physical Review B* **2011**, *83*, 155208.

[59] Walzer, K.; Maennig, B.; Pfeiffer, M.; Leo, K. Highly Efficient Organic Devices Based on Electrically Doped Transport Layers. *Chemical Reviews* **2007**, *107*, 1233–1271.

[60] Schwarze, M. et al. Molecular parameters responsible for thermally activated transport in doped organic semiconductors. *Nature Materials* **2019**, *18*, 242–248.

[61] Stolar, M.; Baumgartner, T. Organic n-type materials for charge transport and charge storage applications. *Physical Chemistry Chemical Physics* **2013**, *15*, 9007–9024.

[62] Zaumseil, J.; Sirringhaus, H. Electron and Ambipolar Transport in Organic Field-Effect Transistors. *Chemical Reviews* **2007**, *107*, 1296–1323.

[63] Ishii, H.; Sugiyama, K.; Ito, E.; Seki, K. Energy Level Alignment and Interfacial Electronic Structures at Organic/Metal and Organic/Organic Interfaces. *Advanced Materials* **1999**, *11*, 605–625.

[64] Zojer, E.; Taucher, T. C.; Hofmann, O. T. The Impact of Dipolar Layers on the Electronic Properties of Organic/Inorganic Hybrid Interfaces. *Advanced Materials Interfaces* **2019**, *6*, 1900581.

[65] Heimel, G.; Romaner, L.; Zojer, E.; Bredas, J.-L. The Interface Energetics of Self-Assembled Monolayers on Metals. *Accounts of Chemical Research* **2008**, *41*, 721–729, PMID: 18507404.

[66] Günther, A. A.; Sawatzki, M.; Formánek, P.; Kasemann, D.; Leo, K. Contact Doping for Vertical Organic Field-Effect Transistors. *Advanced Functional Materials* **2016**, *26*, 768–775.

[67] Michaelson, H. B. The work function of the elements and its periodicity. *Journal of Applied Physics* **1977**, *48*, 4729–4733.

[68] Sun, H.; Ryno, S.; Zhong, C.; Ravva, M. K.; Sun, Z.; Körzdörfer, T.; Brédas, J.-L. Ionization Energies, Electron Affinities, and Polarization Energies of Organic Molecular Crystals: Quantitative Estimations from a Polarizable Continuum Model (PCM)-Tuned Range-Separated Density Functional Approach. *Journal of Chemical Theory and Computation* **2016**, *12*, 2906–2916.

[69] 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. *Journal of Chemical Theory and Computation* **2018**, *14*, 6287–6294.

[70] Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Physical Review A* **1988**, *38*, 3098–3100.

[71] Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Physical Review B* **1988**, *37*, 785–789.

[72] 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. *The Journal of Physical Chemistry* **1994**, *98*, 11623–11627.

[73] Schwenn, P.; Burn, P.; Powell, B. Calculation of solid state molecular ionisation energies and electron affinities for organic semiconductors. *Organic Electronics* **2011**, *12*, 394 – 403.

[74] 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. *Physical Chemistry Chemical Physics* **2013**, *15*, 14006–14016.

[75] Oberhofer, H.; Reuter, K.; Blumberger, J. Charge Transport in Molecular Materials: An Assessment of Computational Methods. *Chemical Reviews* **2017**, *117*, 10319–10357.

[76] Brédas, J.-L.; Beljonne, D.; Coropceanu, V.; Cornil, J. Charge-Transfer and Energy-Transfer Processes in π-Conjugated Oligomers and Polymers: A Molecular Picture. *Chemical Reviews* **2004**, *104*, 4971–5004.

[77] Ortmann, F.; Bechstedt, F.; Hannewald, K. Charge transport in organic crystals: Theory and modelling. *Physica Status Solidi (B) Basic Research* **2011**, *248*, 511–525.

[78] Troisi, A. Charge transport in high mobility molecular semiconductors: Classical models and new theories. *Chemical Society Reviews* **2011**, *40*, 2347–2358.

[79] Fratini, S.; Mayou, D.; Ciuchi, S. The Transient Localization Scenario for Charge Transport in Crystalline Organic Materials. *Advanced Functional Materials* **2016**, *26*, 2292–2315.

[80] Groves, C. Simulating charge transport in organic semiconductors and devices: a review. *Reports on Progress in Physics* **2016**, *80*, 26502.

[81] Nematiaram, T.; Troisi, A. Modeling charge transport in high-mobility molecular semiconductors: Balancing electronic structure and quantum dynamics methods with the help of experiments. *The Journal of Chemical Physics* **2020**, *152*, 190902.

[82] Shuai, Z.; Wang, L.; Song, C. *Theory of Charge Transport in Carbon Electronic Materials*; SpringerBriefs in Molecular Science; Springer Berlin Heidelberg, 2012.

[83] Stehr, V. Prediction of charge and energy transport in organic crystals with quantum chemical protocols employing the hopping model. PhD thesis, Universität Würzburg, 2015.

[84] Ishii, H. *Charge Transport Simulations for Organic Semiconductors, in Molecular Technology*; John Wiley & Sons, Ltd, 2018; Chapter 1, pp 1–23.

[85] Dimitrakopoulos, C.; Malenfant, P. Organic Thin Film Transistors for Large Area Electronics. *Advanced Materials* **2002**, *14*, 99–117.

[86] Schweicher, G. et al. Chasing the "Killer" Phonon Mode for the Rational Design of Low-Disorder, High-Mobility Molecular Semiconductors. *Advanced Materials* **2019**, *31*, 1902407.

[87] Giannini, S.; Carof, A.; Ellis, M.; Yang, H.; Ziogos, O. G.; Ghosh, S.; Blumberger, J. Quantum localization and delocalization of charge carriers in organic semiconducting crystals. *Nature Communications* **2019**, *10*, 3843.

[88] Tsutsui, Y. et al. Unraveling Unprecedented Charge Carrier Mobility through Structure Property Relationship of Four Isomers of Didodecyl[1]benzothieno[3,2-b][1]benzothiophene. *Advanced Materials* **2016**, *28*, 7106–7114.

[89] Street, R. *Technology and Applications of Amorphous Silicon*; Springer Series in Materials Science; Springer Berlin Heidelberg, 1999.

[90] Giannini, S.; Ziogos, O. G.; Carof, A.; Ellis, M.; Blumberger, J. Flickering Polarons Extending over Ten Nanometres Mediate Charge Transport in High-Mobility Organic Crystals. *Advanced Theory and Simulations* **2020**, *3*, 2000093.

[91] Stehr, V.; Fink, R. F.; Tafipolski, M.; Deibel, C.; Engels, B. Comparison of different rate constant expressions for the prediction of charge and energy transport in oligoacenes. *WIREs Computational Molecular Science* **2016**, *6*, 694–720.

[92] Marcus, R. A. On the Theory of Oxidation-Reduction Reactions Involving Electron Transfer. I. *Journal of Chemical Physics* **1956**, *24*, 966–978.

[93] Marcus, R. A. Electron Transfer Reactions in Chemistry. Theory and Experiment. *Reviews of Modern Physics* **1993**, *65*, 599–610.

[94] Fetherolf, J. H.; Golež, D.; Berkelbach, T. C. A Unification of the Holstein Polaron and Dynamic Disorder Pictures of Charge Transport in Organic Crystals. *Phys. Rev. X* **2020**, *10*, 021062.

[95] Senthilkumar, K.; Grozema, F. C.; Bickelhaupt, F. M.; Siebbeles, L. D. A. Charge transport in columnar stacked triphenylenes: Effects of conformational fluctuations on charge transfer integrals and site energies. *The Journal of Chemical Physics* **2003**, *119*, 9809–9817.

[96] Valeev, E. F.; Coropceanu, V.; da Silva Filho, D. A.; Salman, S.; Brédas, J.-L. Effect of Electronic Polarization on Charge-Transport Parameters in Molecular Organic Semiconductors. *Journal of the American Chemical Society* **2006**, *128*, 9882–9886.

[97] Schober, C.; Reuter, K.; Oberhofer, H. Critical analysis of fragment-orbital DFT schemes for the calculation of electronic coupling values. *Journal of Chemical Physics* **2016**, *144*, 054103.

[98] Ghan, S.; Kunkel, C.; Reuter, K.; Oberhofer, H. Improved projection-operator diabatization schemes for the calculation of electronic coupling values. *The Journal of Chemical Physics* **2020**,

[99] Chen, X.-K.; Zou, L.-Y.; Guo, J.-F.; Ren, A.-M. An efficient strategy for designing n-type organic semiconductor materials—introducing a six-membered imide ring into aromatic diimides. *Journal of Materials Chemistry* **2012**, *22*, 6471–6484.

[100] Winkler, C.; Jeindl, A.; Mayer, F.; Hofmann, O. T.; Tonner, R.; Zojer, E. Understanding the Correlation between Electronic Coupling and Energetic Stability of Molecular Crystal Polymorphs: The Instructive Case of Quinacridone. *Chemistry of Materials* **2019**, *31*, 7054–7069.

[101] Yu, M.; Wang, X.; Du, X.-F.; Kunkel, C.; Garcia, T. M.; Monaco, S.; Schatschneider, B.; Oberhofer, H.; Marom, N. Anomalous pressure dependence of the electronic properties of molecular crystals explained by changes in intermolecular electronic coupling. *Synthetic Metals* **2019**, *253*, 9 – 19.

[102] Tu, Z.; Yi, Y.; Coropceanu, V.; Brédas, J.-L. Impact of Phonon Dispersion on Nonlocal Electron–Phonon Couplings in Organic Semiconductors: The Naphthalene Crystal as a Case Study. *The Journal of Physical Chemistry C* **2018**, *122*, 44–49.

[103] Nematiaram, T.; Troisi, A. Strategies to reduce the dynamic disorder in molecular semiconductors. *Mater. Horiz.* **2020**, *7*, 2922–2928.

[104] Chung, H.; Diao, Y. Polymorphism as an emerging design strategy for high performance organic electronics. *Journal of Materials Chemistry C* **2016**, *4*, 3915–3933.

[105] Landi, A.; Troisi, A.; Peluso, A. Explaining different experimental hole mobilities: influence of polymorphism on dynamic disorder in pentacene. *Journal of Materials Chemistry C* **2019**, *7*, 9665–9670.

[106] Lederer, J.; Kaiser, W.; Mattoni, A.; Gagliardi, A. Machine Learning–Based Charge Transport Computation for Pentacene. *Advanced Theory and Simulations* **2019**, *2*, 1800136.

[107] Wang, C.-I.; Braza, M. K. E.; Claudio, G. C.; Nellas, R. B.; Hsu, C.-P. Machine Learning for Predicting Electron Transfer Coupling. *The Journal of Physical Chemistry A* **2019**, *123*, 7792–7802.

[108] Çaylak, O.; Yaman, A.; Baumeier, B. Evolutionary Approach to Constructing a Deep Feedforward Neural Network for Prediction of Electronic Coupling Elements in Molecular Materials. *Journal of Chemical Theory and Computation* **2019**, *15*, 1777–1784.

[109] Austin, I.; Mott, N. Polarons in crystalline and non-crystalline materials. *Advances in Physics* **1969**, *18*, 41–102.

[110] Norton, J. E.; Brédas, J.-L. Polarization Energies in Oligoacene Semiconductor Crystals. *Journal of the American Chemical Society* **2008**, *130*, 12377–12384.

[111] McMahon, D. P.; Troisi, A. Evaluation of the External Reorganization Energy of Polyacenes. *The Journal of Physical Chemistry Letters* **2010**, *1*, 941–946.

[112] Martinelli, N. G.; Idé, J.; Sánchez-Carrera, R. S.; Coropceanu, V.; Brédas, J.-L.; Ducasse, L.; Castet, F.; Cornil, J.; Beljonne, D. Influence of Structural Dynamics on Polarization Energies in Anthracene Single Crystals. *The Journal of Physical Chemistry C* **2010**, *114*, 20678–20685.

[113] Nelsen, S. F.; Blackstock, S. C.; Kim, Y. Estimation of inner shell Marcus terms for amino nitrogen compounds by molecular orbital calculations. *Journal of the American Chemical Society* **1987**, *109*, 677–682.

[114] Reimers, J. R. A practical method for the use of curvilinear coordinates in calculations of normal-mode-projected displacements and Duschinsky rotation matrices for large molecules. *The Journal of Chemical Physics* **2001**, *115*, 9103–9109.

[115] Kato, T.; Yamabe, T. Vibronic interactions and superconductivity in acene anions and cations. *The Journal of Chemical Physics* **2001**, *115*, 8592–8602.

[116] Brückner, C.; Engels, B. A theoretical description of charge reorganization energies in molecular organic P-type semiconductors. *Journal of Computational Chemistry* **2016**, *37*, 1335–1344.

[117] Kera, S.; Hosoumi, S.; Sato, K.; Fukagawa, H.; Nagamatsu, S.-i.; Sakamoto, Y.; Suzuki, T.; Huang, H.; Chen, W.; Wee, A. T. S.; Coropceanu, V.; Ueno, N. Experimental Reorganization Energies of Pentacene and Perfluoropentacene: Effects of Perfluorination. *The Journal of Physical Chemistry C* **2013**, *117*, 22428–22437.

[118] Atahan-Evrenk, S. Computational investigation of intramolecular reorganization energy in diketopyrrolopyrrole (DPP) derivatives. *Turkish Journal of Chemistry* **2018**, *42*, 869–882.

[119] Lin, K.-H.; Corminboeuf, C. FB-REDA: fragment-based decomposition analysis of the reorganization energy for organic semiconductors. *Physical Chemistry Chemical Physics* **2020**, *22*, 11881–11890.

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

[121] Antono, E.; Matsuzawa, N. N.; Ling, J.; Saal, J. E.; Arai, H.; Sasago, M.; Fujii, E. Machine-Learning Guided Quantum Chemical and Molecular Dynamics Calculations to Design Novel Hole-Conducting Organic Materials. *The Journal of Physical Chemistry A* **2020**, *124*, 8330–8340.

[122] Chang, Y.-C.; Chao, I. An Important Key to Design Molecules with Small Internal Reorganization Energy: Strong Nonbonding Character in Frontier Orbitals. *The Journal of Physical Chemistry Letters* **2010**, *1*, 116–121.

[123] Chen, W.-C.; Chao, I. Molecular Orbital-Based Design of -Conjugated Organic Materials with Small Internal Reorganization Energy: Generation of Nonbonding Character in Frontier Orbitals. *The Journal of Physical Chemistry C* **2014**, *118*, 20176–20183.

[124] Kuo, M.-Y.; Liu, C.-C. Molecular Design toward High Hole Mobility Organic Semiconductors: Tetraceno[2,3-c]thiophene Derivatives of Ultrasmall Reorganization Energies. *The Journal of Physical Chemistry C* **2009**, *113*, 16303–16306.

[125] Zhu, R.; Duan, Y.-A.; Geng, Y.; Wei, C.-Y.; Chen, X.-Y.; Liao, Y. Theoretical evaluation on the reorganization energy of five-ring-fused benzothiophene derivatives. *Computational and Theoretical Chemistry* **2016**, *1078*, 16–22.

[126] Oshi, R.; Abdalla, S.; Springborg, M. Theoretical study on functionalized anthracene and tetraceneas starting species to produce promising semiconductor materials. *Computational and Theoretical Chemistry* **2018**, *1128*, 60–69.

[127] Hutchison, G. R.; Ratner, M. A.; Marks, T. J. Hopping Transport in Conductive Heterocyclic Oligomers: Reorganization Energies and Substituent Effects. *Journal of the American Chemical Society* **2005**, *127*, 2339–2350.

[128] Misra, M.; Andrienko, D.; Baumeier, B.; Faulon, J.-L.; von Lilienfeld, O. A. Toward Quantitative Structure–Property Relationships for Charge Transfer Rates of Polycyclic Aromatic Hydrocarbons. *Journal of Chemical Theory and Computation* **2011**, *7*, 2549–2555.

[129] Bronstein, H.; Nielsen, C. B.; Schroeder, B. C.; McCulloch, I. The role of chemical design in the performance of organic semiconductors. *Nature Reviews Chemistry* **2020**, *4*, 66–77.

[130] Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **2004**, *432*, 862–865.

[131] Sotriffer, C.; Mannhold, R.; Kubinyi, H.; Folkers, G. *Virtual Screening: Principles, Challenges, and Practical Guidelines*; Methods and Principles in Medicinal Chemistry; Wiley, 2011.

[132] Curtarolo, S.; Hart, G. L. W.; Nardelli, M. B.; Mingo, N.; Sanvito, S.; Levy, O. The high-throughput highway to computational materials design. *Nature Materials* **2013**, *12*, 191–201.

[133] 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.* **2015**, *45*, 195–216.

[134] Hachmann, J.; Olivares-Amaya, R.; Atahan-Evrenk, S.; Amador-Bedolla, C.; Sánchez-Carrera, R. S.; Gold-Parker, A.; Vogt, L.; Brockway, A. M.; Aspuru-Guzik, A. The Harvard Clean Energy Project: Large-Scale Computational Screening and Design of Organic Photovoltaics on the World Community Grid. *The Journal of Physical Chemistry Letters* **2011**, *2*, 2241–2251.

[135] Hachmann, J.; Olivares-Amaya, R.; Jinich, A.; Appleton, A. L.; Blood-Forsythe, M. A.; Seress, L. R.; Román-Salgado, C.; Trepte, K.; Atahan-Evrenk, S.; Er, S.; Shrestha, S.; Mondal, R.; Sokolov, A.; Bao, Z.; Aspuru-Guzik, A. Lead candidates for high-performance organic photovoltaics from high-throughput quantum chemistry – the Harvard Clean Energy Project. *Energy and Environmental Science* **2014**, *7*, 698–704.

[136] Taylor, R.; Wood, P. A. A Million Crystal Structures: The Whole Is Greater than the Sum of Its Parts. *Chemical Reviews* **2019**, *119*, 9427–9477.

[137] The RDKit: Open-Source Cheminformatics Software. `http://www.rdkit.org`.

[138] Saldívar-González, F. I.; Medina-Franco, J. L. In *Small Molecule Drug Discovery*; Trabocchi, A., Lenci, E., Eds.; Elsevier, 2020; pp 83–102.

[139] Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **2014**, *1*.

[140] Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *Journal of Chemical Information and Modeling* **2012**, *52*, 2864–2875.

[141] von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A. Exploring chemical compound space with quantum-based machine learning. *Nature Reviews Chemistry* **2020**, *4*, 347–358.

[142] Glavatskikh, M.; Leguy, J.; Hunault, G.; Cauchy, T.; Da Mota, B. Dataset's chemical diversity limits the generalizability of machine learning predictions. *Journal of Cheminformatics* **2019**, *11*, 69.

[143] Stuke, A. Machine learning for spectroscopic properties of organic molecules. 2020.

[144] Russell, S.; Norvig, P. *Artificial Intelligence: A Modern Approach*; Always learning; Pearson, 2016.

[145] Rajan, K. Materials informatics. *Materials Today* **2005**, *8*, 38–45.

[146] Leach, A.; Gillet, V. *An Introduction to Chemoinformatics*; Springer Netherlands, 2007.

[147] Zunger, A. Inverse design in search of materials with target functionalities. *Nature Reviews Chemistry* **2018**, *2*, 0121, Perspective.

[148] Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The rise of deep learning in drug discovery. *Drug Discovery Today* **2018**, *23*, 1241–1250.

[149] Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **2018**, *361*, 360–365.

[150] Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547–555.

[151] Himanen, L.; Geurts, A.; Foster, A. S.; Rinke, P. Data-Driven Materials Science: Status, Challenges, and Perspectives. *Advanced Science* **2019**, *6*, 1900808.

[152] Schmidt, J.; Marques, M. R. G.; Botti, S.; Marques, M. A. L. Recent advances and applications of machine learning in solid-state materials science. *npj Computational Materials* **2019**, *5*, 83.

[153] Carleo, G.; Cirac, I.; Cranmer, K.; Daudet, L.; Schuld, M.; Tishby, N.; Vogt-Maranto, L.; Zdeborová, L. Machine learning and the physical sciences. *Reviews of Modern Physics* **2019**, *91*, 045002.

[154] Manzhos, S.; Carrington, T. Neural Network Potential Energy Surfaces for Small Molecules and Reactions. *Chemical Reviews* **0**, *0*, null, PMID: 33021368.

[155] Stokes, J. M. et al. A Deep Learning Approach to Antibiotic Discovery. *Cell* **2020**, *180*, 688 – 702.e13.

[156] Cheng, B.; Mazzola, G.; Pickard, C. J.; Ceriotti, M. Evidence for supercritical behaviour of high-pressure liquid hydrogen. *Nature* **2020**, *585*, 217–220.

[157] Timmermann, J.; Kraushofer, F.; Resch, N.; Li, P.; Wang, Y.; Mao, Z.; Riva, M.; Lee, Y.; Staacke, C.; Schmid, M.; Scheurer, C.; Parkinson, G. S.; Diebold, U.; Reuter, K. IrO2 Surface Complexions Identified Through Machine Learning and Surface Investigations. 2020.

[158] Padula, D.; Simpson, J. D.; Troisi, A. Combining electronic and structural features in machine learning models to predict organic solar cells properties. *Mater. Horiz.* **2019**, *6*, 343–349.

[159] Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAPRetrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. *Journal of Chemical Information and Computer Sciences* **1998**, *38*, 511–522.

[160] Degen, J.; Wegscheid-Gerlach, C.; Zaliani, A.; Rarey, M. On the Art of Compiling and Using 'Drug-Like' Chemical Fragment Spaces. *ChemMedChem* **2008**, *3*.

[161] Yoshikawa, N.; Hutchison, G. R. Fast, efficient fragment-based coordinate generation for Open Babel. *Journal of Cheminformatics* **2019**, *11*, 49.

[162] Walters, W. P. Virtual Chemical Libraries. *Journal of Medicinal Chemistry* **2019**, *62*, 1116–1124.

[163] Saldívar-González, F. I.; Huerta-García, C. S.; Medina-Franco, J. L. Chemoinformatics-based enumeration of chemical libraries: a tutorial. *Journal of Cheminformatics* **2020**, *12*, 64.

[164] Schneider, G. *De novo Molecular Design*; Wiley, 2013.

[165] Kawai, K.; Nagata, N.; Takahashi, Y. De Novo Design of Drug-Like Molecules by a Fragment-Based Molecular Evolutionary Approach. *Journal of Chemical Information and Modeling* **2014**, *54*, 49–56.

[166] Bian, Y.; Xie, X.-Q. S. Computational Fragment-Based Drug Design: Current Trends, Strategies, and Applications. *The AAPS Journal* **2018**, *20*, 59.

[167] Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *Journal of Medicinal Chemistry* **1996**, *39*, 2887–2893.

[168] Kruger, F.; Stiefl, N.; Landrum, G. A. rdScaffoldNetwork: The Scaffold Network Implementation in RDKit. *Journal of Chemical Information and Modeling* **2020**, *60*, 3331–3335.

[169] Mann, H. B.; Whitney, D. R. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Annals of Mathematical Statistics* **1947**, *18*, 50–60.

[170] Subramanian, A. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. *Proceedings of the National Academy of Sciences* **2005**, *102*, 15545–15550.

[171] Yang, H.; Li, J.; Wu, Z.; Li, W.; Liu, G.; Tang, Y. Evaluation of Different Methods for Identification of Structural Alerts Using Chemical Ames Mutagenicity Data Set as a Benchmark. *Chemical Research in Toxicology* **2017**, *30*, 1355–1364.

[172] Varin, T.; Gubler, H.; Parker, C. N.; Zhang, J.-H.; Raman, P.; Ertl, P.; Schuffenhauer, A. Compound Set Enrichment: A Novel Approach to Analysis of Primary HTS Data. *Journal of Chemical Information and Modeling* **2010**, *50*, 2067–2078.

[173] Settles, B. *Active Learning*; Synthesis Lectures on Artificial Intelligence and Machine Learning Series; Morgan & Claypool, 2012.

[174] Reker, D.; Schneider, G. Active-learning strategies in computer-assisted drug discovery. *Drug Discovery Today* **2015**, *20*, 458 – 465.

[175] 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 Computational Materials* **2019**, *5*, 21.

[176] Todeschini, R.; Consonni, V.; Mannhold, R.; Kubinyi, H.; Timmerman, H. *Handbook of Molecular Descriptors*; Methods and Principles in Medicinal Chemistry; Wiley, 2008.

[177] Wu, Z.; Ramsundar, B.; Feinberg, E.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chemical Science* **2018**, *9*, 513–530.

[178] Ward, L. et al. Matminer: An open source toolkit for materials data mining. *Computational Materials Science* **2018**, *152*, 60–69.

[179] Himanen, L.; Jäger, M. O. J.; Morooka, E. V.; Federici Canova, F.; Ranawat, Y. S.; Gao, D. Z.; Rinke, P.; Foster, A. S. DScribe: Library of descriptors for machine learning in materials science. *Computer Physics Communications* **2020**, *247*, 106949.

[180] Cereto-Massagué, A.; Ojeda, M. J.; Valls, C.; Mulero, M.; Garcia-Vallvé, S.; Pujadas, G. Molecular fingerprint similarity search in virtual screening. *Methods* **2015**, *71*, 58 – 63.

[181] Capecchi, A.; Probst, D.; Reymond, J.-L. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *Journal of Cheminformatics* **2020**, *12*, 43.

[182] Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* **2010**, *50*, 742–754.

[183] Duvenaud, D.; Maclaurin, D.; Aguilera-Iparraguirre, J.; Gómez-Bombarelli, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. 2015.

[184] Zhou, J.; Cui, G.; Zhang, Z.; Yang, C.; Liu, Z.; Wang, L.; Li, C.; Sun, M. Graph Neural Networks: A Review of Methods and Applications. 2019.

[185] Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Physical Review B* **2013**, *87*, 184115.

[186] Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Physical Review Letters* **2012**, *108*, 058301.

[187] Huo, H.; Rupp, M. Unified Representation of Molecules and Crystals for Machine Learning. 2018.

[188] Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons. *Physical Review Letters* **2010**, *104*, 136403.

[189] De, S.; Bartók, A. P.; Csányi, G.; Ceriotti, M. Comparing molecules and solids across structural and alchemical space. *Physical Chemistry Chemical Physics* **2016**, *18*, 13754–13769.

[190] Mitchell, T.; Carbonell, J.; Michalski, R. *Machine Learning: A Guide to Current Research*; The Springer International Series in Engineering and Computer Science; Springer US, 1986.

[191] Bishop, C. *Pattern Recognition and Machine Learning*; Information Science and Statistics; Springer, 2006.

[192] Murphy, K. *Machine Learning: A Probabilistic Perspective*; Adaptive Computation and Machine Learning series; MIT Press, 2012.

[193] Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

[194] Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; Adaptive Computation and Machine Learning series; MIT Press, 2016.

[195] Hofmann, T.; Schölkopf, B.; Smola, A. J. Kernel methods in machine learning. *Annals of Statistics* **2008**, *36*, 1171–1220.

[196] Rupp, M. Machine learning for quantum mechanics in a nutshell. *International Journal of Quantum Chemistry* **2015**, *115*, 1058–1073.

[197] Rasmussen, C.; Williams, C. *Gaussian Processes for Machine Learning*; Adaptative computation and machine learning series; University Press Group Limited, 2006.

[198] Williams, C.; Rasmussen, C. Gaussian Processes for Regression. Advances in neural information processing systems. Cambridge, MA, USA, 1996; pp 514–520.

[199] Ralaivola, L.; Swamidass, S. J.; Saigo, H.; Baldi, P. Graph kernels for chemical informatics. *Neural Networks* **2005**, *18*, 1093 – 1110.

[200] Srinivas, N.; Krause, A.; Kakade, S.; Seeger, M. Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design. Proceedings of the 27th International Conference on International Conference on Machine Learning. Madison, WI, USA, 2010; p 1015–1022.

[201] van der Maaten, L.; Postma, E.; Herik, H. Dimensionality Reduction: A Comparative Review. *Journal of Machine Learning Research - JMLR* **2007**, *10*.

[202] Chen, C.; Härdle, W.; Unwin, A. *Handbook of Data Visualization*; Springer Handbooks of Computational Statistics; Springer Berlin Heidelberg, 2007.

[203] Ceriotti, M.; Tribello, G. A.; Parrinello, M. Simplifying the representation of complex free-energy landscapes using sketch-map. *Proceedings of the National Academy of Sciences* **2011**, *108*, 13023–13028.

[204] Reutlinger, M.; Schneider, G. Nonlinear dimensionality reduction and mapping of compound libraries for drug discovery. *Journal of Molecular Graphics and Modelling* **2012**, *34*, 108–117.

[205] Ceriotti, M. Unsupervised machine learning in atomistic simulations, between predictions and understanding. *The Journal of Chemical Physics* **2019**, *150*, 150901.

[206] Molnar, C. *Interpretable Machine Learning*; Lulu.com, 2020.

[207] van der Maaten, L.; Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **2008**, *9*, 2579–2605.

[208] McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. 2020.

[209] Isayev, O.; Fourches, D.; Muratov, E. N.; Oses, C.; Rasch, K.; Tropsha, A.; Curtarolo, S. Materials Cartography: Representing and Mining Materials Space Using Structural and Electronic Fingerprints. *Chemistry of Materials* **2015**, *27*, 735–743.

[210] Automatic Selection And Prediction tools for materials and molecules. `https://github.com/BingqingCheng/ASAP`.

[211] Osolodkin, D. I.; Radchenko, E. V.; Orlov, A. A.; Voronkov, A. E.; Palyulin, V. A.; Zefirov, N. S. Progress in visual representations of chemical space. *Expert Opinion on Drug Discovery* **2015**, *10*, 959–973.

[212] 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. *Journal of Chemical Information and Modeling* **2015**, *55*, 84–94.

[213] Vogt, M.; Stumpfe, D.; Maggiora, G. M.; Bajorath, J. Lessons learned from the design of chemical space networks and opportunities for new applications. *Journal of Computer-Aided Molecular Design* **2016**, *30*, 191–208.

[214] Bokeh Development Team, Bokeh: Python library for interactive visualization. 2020.

[215] Inc., P. T. Collaborative data science. 2015; `https://plot.ly`.

[216] Projection-Viewer. `https://github.com/chkunkel/projection_viewer`.

[217] Probst, D.; Reymond, J.-L. FUn: a framework for interactive visualizations of large, high-dimensional datasets on the web. *Bioinformatics* **2017**, *34*, 1433–1435.

[218] Fraux, G.; Cersonsky, R. K.; Ceriotti, M. Chemiscope: interactive structure-property explorer for materials and molecules. *Journal of Open Source Software* **2020**, *5*, 2117.

[219] Gütlein, M.; Karwath, A.; Kramer, S. CheS-Mapper 2.0 for visual validation of (Q)SAR models. *Journal of Cheminformatics* **2014**, *6*, 41.

[220] Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. KNIME: The Konstanz Information Miner. Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007). 2007.

[221] Sander, T.; Freyss, J.; von Korff, M.; Rufener, C. DataWarrior: An Open-Source Program For Chemistry Aware Data Visualization And Analysis. *Journal of Chemical Information and Modeling* **2015**, *55*, 460–473.

[222] Ishii, H.; Obata, S.; Niitsu, N.; Watanabe, S.; Goto, H.; Hirose, K.; Kobayashi, N.; Okamoto, T.; Takeya, J. Charge mobility calculation of organic semiconductors without use of experimental single-crystal data. *Scientific Reports* **2020**, *10*, 2524.

[223] Friederich, P.; Meded, V.; Poschlad, A.; Neumann, T.; Rodin, V.; Stehr, V.; Symalla, F.; Danilov, D.; Lüdemann, G.; Fink, R. F.; Kondov, I.; von Wrochem, F.; Wenzel, W. Molecular Origin of the Charge Carrier Mobility in Small Molecule Organic Semiconductors. *Advanced Functional Materials* **2016**, *26*, 5757–5763.

[224] Egger, A. T.; Hörmann, L.; Jeindl, A.; Scherbela, M.; Obersteiner, V.; Todorović, M.; Rinke, P.; Hofmann, O. T. Charge Transfer into Organic Thin Films: A Deeper Insight through Machine-Learning-Assisted Structure Search. *Advanced Science* **2020**, *n/a*, 2000992.

[225] MacLeod, B. P. et al. Self-driving laboratory for accelerated discovery of thin-film materials. *Science Advances* **2020**, *6*.