

Inverse-Design on Molecules in Solar Cells

Dissertation

zur Erlangung des Grades

des Doktors der Naturwissenschaften

der Naturwissenschaftlich-Technischen Fakultät

der Universität des Saarlandes

von

Chencheng Fan

Saarbrücken

2022

Tag des Kolloquiums: 7. Dezember 2022

Dekan: Prof. Dr. Ludger Santen

Berichterstatter: Prof. Dr. Michael Springborg
Prof. Dr. Benoît Champagne

Akad. Mitglied: Dr. Diego M. Andrada

Vorsitz: Prof. Dr. Martin H. Müser

I would like to dedicate this thesis to my beloved parents . . .

Acknowledgements

I would like to acknowledge my supervisor, Prof. Springborg for guiding my Ph.D. project and leading me in the fantastic world of quantum and computational chemistry. I acknowledge my parents and family for their great love, which gave me the courage and strength to get up every time after falling. I acknowledge my dear friends for our friendship which warms me in the cold and lit the lights in the dark. I acknowledge my colleagues from whom I learned a lot in many aspects. Last but not the least, I acknowledge Philipp Thiel for translating the English abstract into German.

Abstract

Solar cells have attracted much attention due to the climate change and the energy crisis. But it is still a tremendous challenge to design break-through molecules with high cell efficiency. In contrast to the traditional “trial and error” approaches that design the molecules first and study their properties subsequently, the inverse design approach pursued in this Ph.D. project can be used in automatically identifying structures with the optimum property within a chemical space defined by a backbone and possible substituents. In the present work, this approach has been applied to optimize porphyrins in dye-sensitized solar cells and other systems for photovoltaics. The predicted optimal molecules give guidance for future molecular design. Our approach is also applied to study the role of the backbone when optimizing functional groups. The results show that for some properties, the modifications of the backbone have remarkable impacts on the properties and the optimal sets of functional groups. On the other hand, for other properties, these effects are less pronounced. Although solar cells are the motivation of our present inverse-design approaches and the current playground, in principle every conceivable property can be optimized using our inverse-design approaches. In addition, the studied systems can be almost every class of materials, even macro-molecules or monomers. Accordingly, our inverse-design approach is applicable for a wide range of systems and properties.

Abstract (German)

Solarzellen werden als eine Schlüsseltechnologie für die Energiewende betrachtet, aber das Design von Molekülen, die hohe Zelleffizienzen erzielen, ist immernoch eine enorme Herausforderung. Im Gegensatz zu klassischen “trial and error” Designansätzen in denen Moleküle erst synthetisiert und anschließend untersucht werden, verfolgt die vorliegende Arbeit einen inversen Designansatz. Hierbei werden, für ein gegebenes chemisches System aus Gerüst und Substituenten, die Systeme mit den optimalen Eigenschaften automatisch bestimmt. In der vorliegenden Arbeit wurden Porphyrine und andere konjugierte Systeme für die Anwendung in Farbstoffsolarzellen und ähnlichen Photovoltaiksystemen untersucht. Die vorhergesagten Moleküle können als Ausgangspunkt für weitere Studien dienen. Ferner wurde der Einfluss des Gerüsts auf die Eigenschaften des optimalen Moleküls und seiner Substituenten untersucht. Die Ergebnisse zeigen, dass für einige Eigenschaften Gerüstmodifikationen grundlegenden Einfluss auf die Eigenschaften und Substitutionsmuster haben, andere Eigenschaften hingegen wenig betroffen sind. Obwohl der hier eingesetzte inverse Designansatz nur für Solarzellen angewandt wurde, steht einer Anwendung für andere chemische Anwendungen wie z. B. heterogene Katalyse und Wasserstoffspeicherung nichts im Wege. Inverses Design kann auf fast alle Materialklassen von Monomeren bis zu Makromolekülen angewandt und für ein breites Feld von Eigenschaften verwendet werden.

Table of contents

Acknowledgements	iii
Abstract	v
Abstract (German)	vii
List of figures	xi
1 Introduction	1
1.1 Solar cell	1
1.2 DSSC (Dye-sensitized solar cell) and dye-sensitizer	3
1.3 The inverse-design approach	7
2 Methodology	11
2.1 Genetic algorithms (GAs)	11
2.1.1 Foundations	11
2.1.2 Search space	12
2.1.3 Fitness score	12
2.1.4 Operators	12
2.1.5 The flowchart	13
2.1.6 The advantages and disadvantages	14
2.2 The electronic-structure calculator: DFTB and TD-DFTB	15
2.2.1 DFTB model	16
2.2.2 TD-DFTB model	18
2.3 SMILES — A Simplified Chemical Language	20
2.4 UV-Vis spectroscopy	25
2.4.1 Important terminologies in UV-Vis spectroscopy	26
2.4.2 Factors affecting the position of UV-Visible bands	27
2.5 Theoretical methods to analyze output-molecules	30

2.5.1	DFT	30
2.5.2	Fukui function	36
2.5.3	Indices to evaluate charge-transfer excitations	38
3	My publications	41
3.1	Publication I	41
3.2	Publication II	71
3.3	Publication III	83
3.4	Publication IV	93
3.5	Publication V	105
4	Summary	119
	References	121

List of figures

1.1	Best research-solar-cell efficiencies. (This plot is reproduced with the courtesy of the National Renewable Energy Laboratory, Golden, CO.)	2
1.2	DSSC principle [1]	4
1.3	Some representative dye-sensitizers	6
2.1	Illustration of the search space of the genetic algorithm with the analogies of genes and chromosomes.	12
2.2	An example of the crossover operator to exchange the genes of two parents and generate two offsprings. Only one of the offsprings is shown.	13
2.3	An example of the mutation operator to mutate the gene of an individual.	13
2.4	Cyclohexane as an example of how to represent cyclic structures [2].	21
2.5	Digit 1 is used twice in SMILES: O1CCCCC1N1CCCCC1, [2], as an example to illustrate that the digit can be repeatedly used when denoting cycles in the SMILES representation.	21
2.6	Two valid SMILES notations for the same structure of 1-methyl-3-bromocyclohexene-1	22
2.7	Generating SMILES for cubane: C12C3C4C1C5C4C3C25.	22
2.8	SMILES allows partial specifications for cis-trans isomerism [2].	23
2.9	Specified and unspecified chirality representations in SMILES [2].	24
2.10	Three types of valence electrons in aldehyde group. [3]	25
2.11	The transition energies for different types of transitions. [3]	26
2.12	The transition energy of $\pi \rightarrow \pi^*$ decreases as the conjugation system prolong. [3]	27
2.13	UV-Vis spectroscopy of $\text{H}(\text{CH}=\text{CH})_n\text{H}$. [3]	28
2.14	Two isomers, trans-stilbene (left) and cis-stilbene (right).	28
2.15	UV-Vis spectroscopy of phenol. [3]	29
2.16	UV-Vis spectroscopy of aniline. [3]	29
2.17	Phenol and phenoxide anion in base	29

2.18 Aniline and protonated aniline in acid 29

Chapter 1

Introduction

1.1 Solar cell

Along with continuous growth of the global population, energy demands have been escalating. But the market of fossil energy may have dramatic variations due to the war or politics. In addition, greenhouse gas emissions cause climate change and threaten species' safety on the earth. So the researchers worldwide are motivated to search for alternative, clean and sustainable energy. Among the diverse renewable energy, solar energy has big advantages and has the potential to meet the enormous energy demands because solar light is abundant, inexhaustible, universal, and eco-friendly [4]. According to world energy outlook 2020 [5], solar photovoltaics will have continuous growth and the experts predicted that the annual increase can be as much as 162GW by 2022, which is 1.5 times the amount in 2019 before COVID-19.

The performance of a solar cell is assessed through power conversion efficiency (PCE, labeled as η), defined as eq.(1.1) at the standard conditions [1]. J_{sc} is the short-circuit photo-current density (where $V=0$). V_{oc} is the open-circuit photo-voltage (where $J=0$). FF is the fill factor.

$$\eta = \frac{J_{sc} \cdot FF \cdot V_{oc}}{P_{in}} \times 100\% \quad (1.1)$$

The general working principle for traditional solar cells includes the absorption of light which generates electron-hole pairs at the p-n junction, the separation of charge carriers while electrons move towards n-type electrode and holes move towards p-type electrode, and the injection of carriers into the external circuit. The theoretical upper limit for η value at AM1.5G (the global standard spectrum) is 33.7% [6] for a single-junction solar cell with a bandgap of 1.34 eV, whereas 68.7% [7] for an almost infinite stack of p-n junctions. This is called the Shockley-Queisser limit.

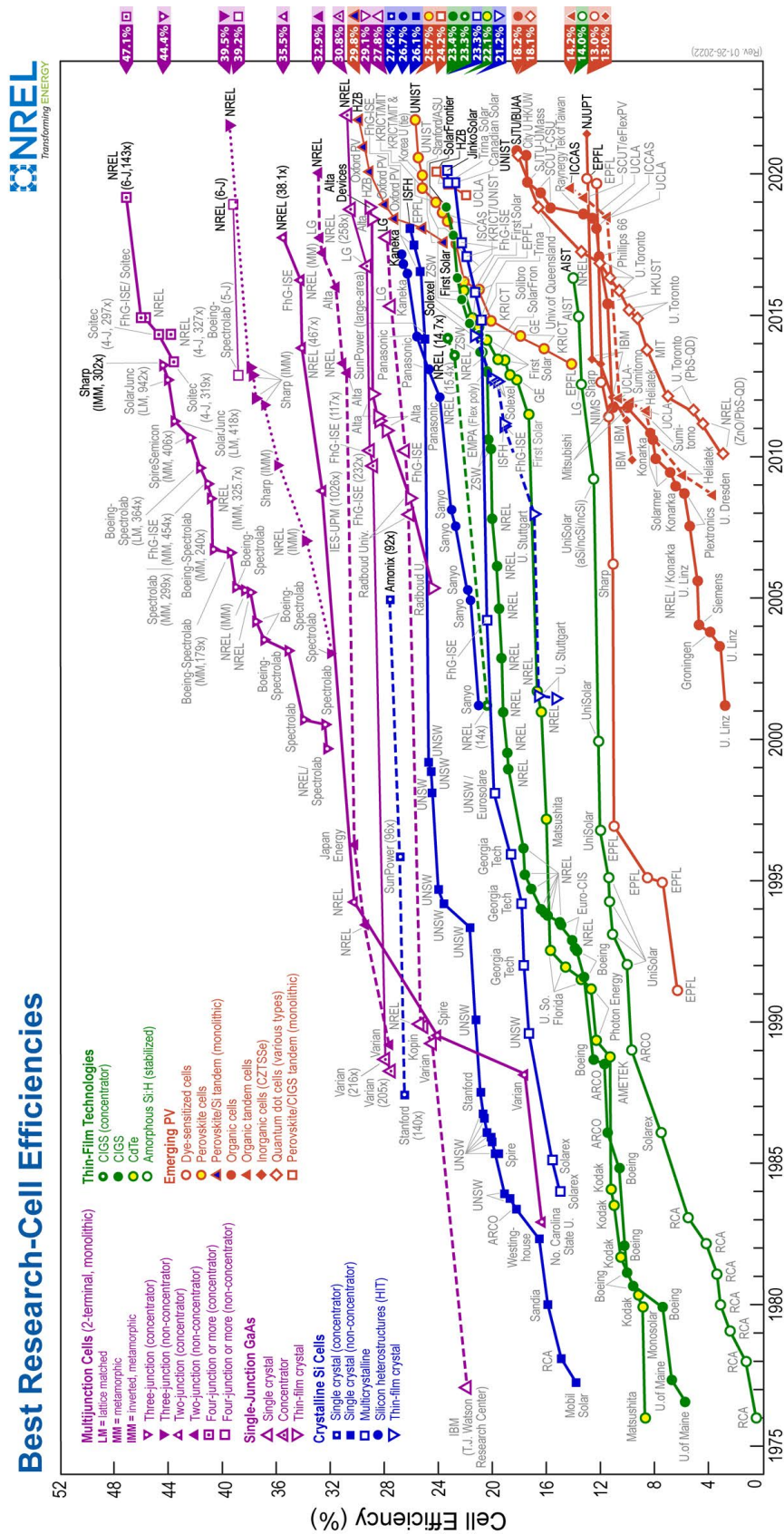


Fig. 1.1 Best research-solar-cell efficiencies. (This plot is reproduced with the courtesy of the National Renewable Energy Laboratory, Golden, CO.)

Fig. 1.1 depicts the records of best research-solar-cell efficiencies from 1970s till now. The 1st generation of solar cells is formed by crystalline silicon (c-Si), such as mono-crystalline (mono-Si, $\eta = 26.7\%$) and multi-crystalline (multi-Si, $\eta=23.3\%$). Crystalline silicon photovoltaics are the dominant product in the global market for the past fifty years, due to their beneficial characteristics [4] such as efficient electricity generation under full sunlight, stable photovoltaic performance in all climatic conditions, and the systematic mature research and development pipelines. However, the 1st generation of solar cells also has some drawbacks, including demanding intensive energies during production processes, not beautiful, bad photovoltaic performances when light intensity is low and low solar energy absorption. So they are not widespread use in building-integrated-photovoltaics, portable electronics, and indoor applications [4]. The 2nd generation of solar cells is thin-film solar cells, such as amorphous Si (a-Si), cadmium telluride (CdTe), and copper indium gallium selenide (CIGS), which are flexible and low cost but low efficiency in general ($\eta=15-20\%$).

The 3rd generation of solar cells has the advantage of using low-cost and ample materials with feasible (or easy) fabrication methods, such as dye-sensitized solar cells (DSSC, $\eta = 13\%$), organic solar cells (OSC, $\eta = 18.2\%$) and perovskite solar cells (PSC, $\eta = 25.5\%$). Their drawbacks are comparatively lower power conversion efficiency and low photovoltaic stability, which have made them difficult to compete with the existing commercial solar cells for bulk electricity generation outdoors [4]. However, they can be made as thin and light-weight flexible solar modules, meeting the requirements of portable electronics. In addition, the 3rd generation photovoltaics have higher efficiency than other existing technology under dim light (typical indoor conditions). So they are promising to be used for the wireless sensors which are commonly used in the devices in the “internet of things” (the name “internet of thing” originates from that devices are connected to the internet and be controlled through the internet).

1.2 DSSC (Dye-sensitized solar cell) and dye-sensitizer

The dye-sensitized solar cell (DSSC) was invented by Michael Grätzel in 1985. The state-of-the-art device was reported in 1991 ($\eta = 7.1\%$) and separates light absorption from charge carrier transport. DSSC is composed of the working electrode (the anode, including transparent conducting oxide glass, the semiconductor, and the dye-sensitizer), the electrolyte, and the counter electrode (the cathode). The working principle of DSSC is shown in Fig. 1.2.

It involves the following processes:

1. photo-excitation of the dye

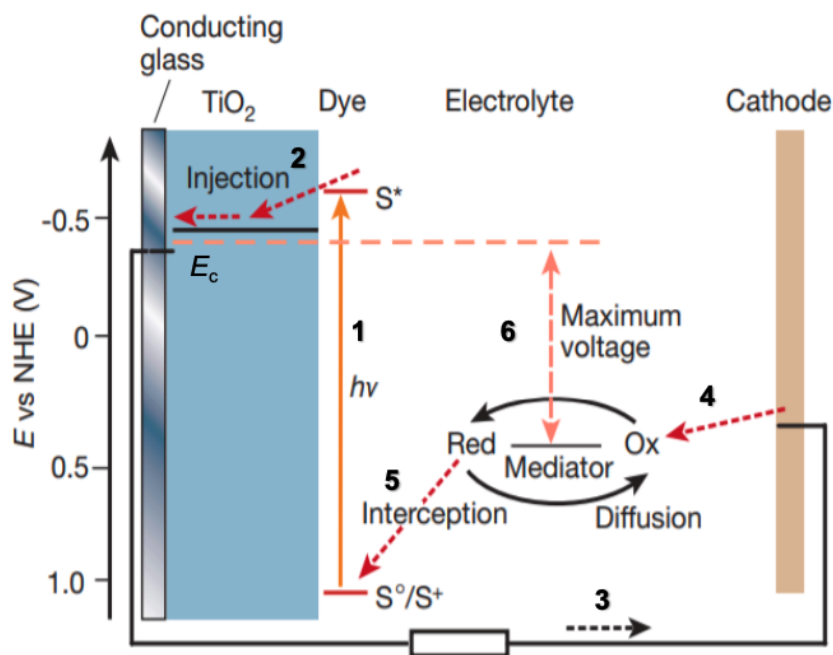


Fig. 1.2 DSSC principle [1]

2. injection of the excited electrons into the conduction band (CB) of the semiconductor (TiO_2 in most cases)
3. electron-transport through the conduction band of the semiconductor, transparent conducting oxide glass (TCO), and the external circuit
4. reduction of the electrolyte at the cathode
5. regeneration of the dye by electrolyte

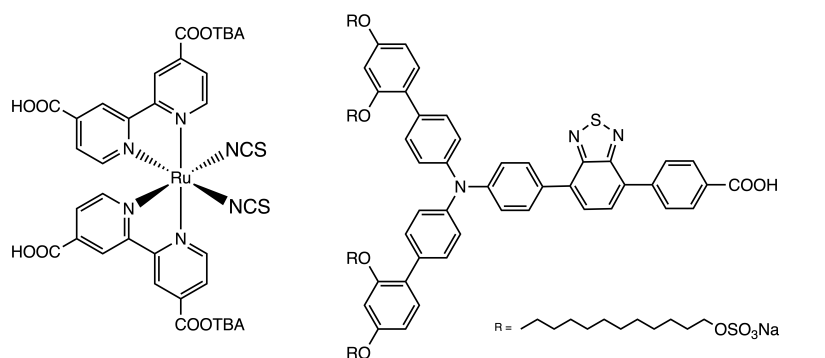
Certain processes can lead to a lowering of η , such as the relaxation of dye molecule in the excited state, the recombination of electrons in the conduction band of the semiconductor or transparent conducting oxide glass with electrolyte or oxidized dye, and trapping of electrons in the semiconductor.

As mentioned earlier, DSSC can be used as an efficient photovoltaic to power electronic devices (like wireless sensors) with indoor light. In [4], the authors highlight the recent progress in developing new materials for producing high-performance DSSC. For example, alternative electrolytes like cobalt (Co) or copper (Cu) redox systems have shown remarkable progress and have significantly increased the energy conversion efficiency of DSSC. The traditional iodide/triiodide redox system is known for its impressive conversion efficiency and robust long-term stability, with intrinsic bottlenecks such as lower redox potential and corrosive nature (which limit the photovoltaic performance of the fabricated DSSC). Co

redox electrolyte has tremendous potential for achieving high conversion efficiencies with high open-circuit voltages, while the major limitations include high recombination rates and mass transport limitations due to bulky ligands which limit the short current densities under full sunlight illumination. Moreover, the review [4] illustrates the progress in advanced TiO₂ photoelectrodes, sealing techniques, stability controls, etc., which all are important to increase the DSSC performance.

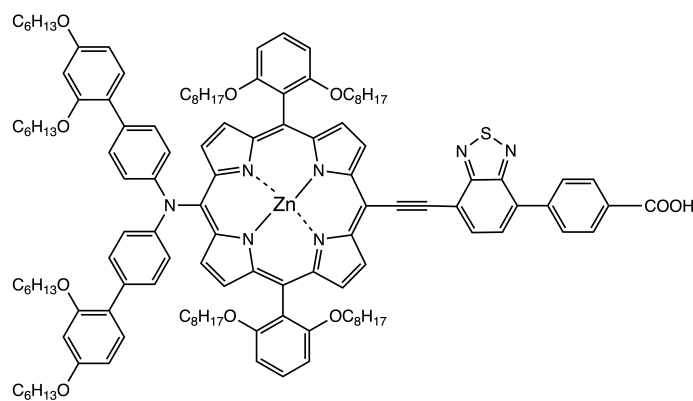
In spite of many components in DSSC, the dye-sensitizers play an important role in determining η and should fulfill various essential characteristics [1]. At first, the dye should have strong absorption at the visible region and near-IR. Then, the dye should have at least one anchoring group (such as -COOH, -H₂PO₃, -SO₃H, etc.) to strongly bind onto the semiconductor surface. Next, the LUMO (the lowest unoccupied molecule orbital) energy of the dye should be higher than the conduction band edge of the semiconductor, while the HOMO (the highest occupied molecule orbital) energy of the dye should be lower than the redox potential of the electrolyte. In addition, the dye should avoid unfavorable aggregation by structural optimization or addition of co-adsorbers since the dye aggregation can lead to the situation that electrons can't be injected into the semiconductors. In the end, the dye should be photo-, thermal and electrochemical stability. Based on these requirements, many dye-sensitizers have been designed and applied to DSSC in the past decades, including metal complexes (like Ruthenium complexes), porphyrins, phthalocyanines, and metal-free organic dyes (like coumarin, indoline, triarylamine, carbazole, etc.). Fig.1.3 shows some representative dye-sensitizers. Especially, porphyrins have attracted much interest because they exist in chlorophyll which has primary roles in photosynthesis, exhibit intense spectral response bands in the near-IR region, and possess good chemical, photo-, and thermal stability, and accordingly provide good potential candidates for photovoltaic applications [1].

However, the optimization of dye-sensitizers is quite difficult and a big challenge. Because adjusting the structures and properties often bring both positive and negative effects. For example, increasing the conduction band edge energy (E_c) can lead to an increase of V_{oc} and a decrease of J_{sc} . Because cell voltage corresponds to ΔV between conduction band edge energy and redox potential of the electrolyte, i.e. "Maximum voltage" in Fig.1.2, and the decrease of J_{sc} is related to the decrease of driving force for electron injection. As seen in eq. (1.1), η is proportional to the product of V_{oc} and J_{sc} . Therefore, the final effect of increasing E_c will be a compromise between these two effects. Currently, the highest η records of DSSC are 13% by the porphyrin named SM315 combined with cobalt (II/III) redox system [8] and 13.5% by dye MS5 with copper (II/I)-based electrolyte [9], which is far lower than Shockley-Queisser limit and implies a huge potential to be further improved.

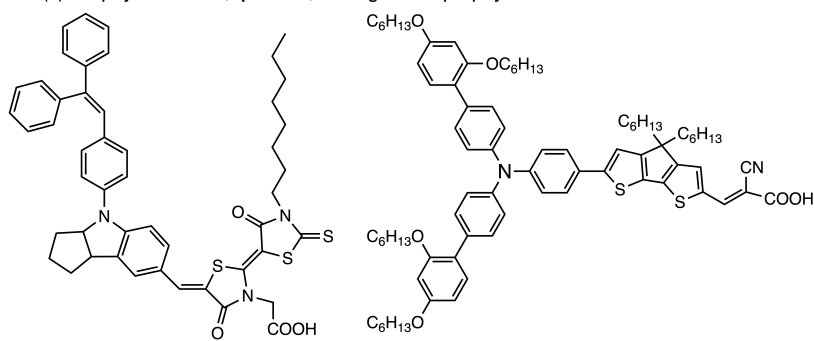


(a) Ru complexes N719; $\eta = 10.2\%$;
reference dye for studies

(b) Organic dye MS5; $\eta = 13.5\%$; the highest in dye-
sensitized solar cell



(c) Porphyrin SM315; $\eta = 13\%$; the highest in porphyrin-sensitized solar cell



(d) Indoline D205; $\eta = 9.5\%$

(e) Triarylamine Y123; $\eta = 10.3\%$

Fig. 1.3 Some representative dye-sensitizers

As introduced in my publication [10] when aiming at finding breakthrough dyes for DSSC, the common “trial and error” approach is very time-consuming, expensive, and inefficient. Normally, scientists design and synthesize novel dyes based on personal experience and ideas. Subsequently, “the cells are assembled, and the performance is checked. In parallel, theoretical calculations may be carried through with the aim of studying their electronic structures and electronic properties. There is an increasing trend towards using computational methods to rationalize the design of new dyes, in particular using quantitative structure-property relationship (QSPR) models. Some of those approaches are based on machine learning or data mining tools, which require powerful computer resources.” [10] An alternative approach, which is the one I pursued in my Ph.D., is the inverse-design approach.

1.3 The inverse-design approach

In contrast to the traditional design that starts with structures and then obtains properties, the inverse design approaches output the optimized structures for the property of interest. Although inverse-design methods in combination with electronic-structure calculations have been proposed almost 20 years ago [11, 12], they have not found wide applicability yet. One study was performed by Venkatraman *et al.* [13, 10], who used a genetic algorithm to optimize coumarin dyes for dye-sensitized solar cells. The overall approach of that work is related to my work that is presented in Sec. 3.1 but the details of the approach, the class of molecular systems, and properties are different from mine. Another relevant study is by O’Boyle [14, 15], who used a genetic algorithm to search the space of six or eight monomer units and examined over 90000 copolymers, finding the optimal structures with matching optical excitation energies and excited-state energies for highly efficient solar cells. In addition, Douguet [16, 15] developed an interesting approach for small organic drug molecules that can parse SMILES and then utilize a genetic algorithm, but the main problem is that the parsing operator is so complicated that the application is confined to small molecules.

Our inverse-design approaches can run on a normal desktop and require no expensive hardware, therefore we call it PooMa, the Poor Man’s Materials Optimization. Although we have three versions of PooMa, the framework is consistent which includes a feasible way of describing, designing, and constructing the different molecules, a genetic-algorithm approach for an efficient and fast search in the chemical space defined by the backbone and possible functional groups, a sufficiently accurate and fast electronic-structure calculator, and an intensive property that shall be optimized [15].

In 2017, my supervisor Prof. Springborg [17] presented the initial idea and frame of our inverse design approach and applied PooMa (1.0) to optimize given Si—Ge clusters for which the pre-chosen atomic positions are occupied by either Si or Ge atom for four electronic properties. “The results demonstrated that it is in principle possible to optimize the properties within a given class of systems, but many issues remained open.” [12] The approach and the test system didn’t involve a change in the number of atoms or valence electrons. When calculating the sunlight absorption property, the oscillator strength was simplified to be 1.

In a subsequent work (see Sec. 3.3, Publication III [12]) by Dr. Huwig and me, we presented a more general inverse-design approach for organic molecules and applied PooMa (2.0) on benzene derivatives as test systems. Considering that the organic molecules have rigid requirements on bond lengths and angles, it was a big leap for the approach to be able to automatically construct molecules with functional groups that consist of more atoms while earlier test systems were clusters, and each “functional group” consists of just one atom. PooMa (2.0) could optimize the structures for seven intensive electronic properties related to solar cells, especially the oscillator strength was calculated explicitly for sun-light absorption. The results predicted some functional groups are beneficial to the properties, while chemical and physical understandings are not always capable of explaining the outcome.

In 2019, I applied PooMa (2.1) to optimizing the real dye sensitizer, porphyrin in dye-sensitized solar cells, (see Sec. 3.1, Publication I [10]). Based on experience and literature, I picked α -substituted zinc-porphyrin as the backbone and 41 functional groups were possibly to be inserted at the four α sites. Furthermore, I optimized structures for the overall solar-cell efficiency η through a quantitative-structure-property-relationship (QSPR) model which was trained to predict the η from five electronic properties. The work provide suggestions and directions on porphyrin design, and it was the first time that we applied our inverse-design approach to a practical application.

In 2020, I introduced PooMa (2.1) to Dr. Khazaal and guided him to apply it to optimizing thiophene oligomers in bulk-heterojunction solar cells (see Sec. 3.4, Publication IV [18]) and cyanopyridone in organic solar cells (see Sec. 3.5, Publication V [19]).

“These studies focused on identifying functional groups that are attached to a given backbone (benzene, porphyrin, etc.) When optimizing organic (often, conjugated) molecules, it is often assumed that for similar backbones the same sets of functional groups are identified as providing the best materials.” [15] To study this assumption and the role of the backbone when optimizing functional groups, in 2021, I improved the inverse-design approach and applied PooMa (3.0) on five similar but related backbones (benzene, pyridine, pyridazine, pyrimidine, and pyrazine). (See Sec. 3.2, Publication II [15])

The PooMa (2.0, 2.1) offered one useful approach but did suffer from some problems. The biggest one is that it was tedious and difficult to prepare the structural information of each functional group and each substitute site in order to construct a reasonable molecule structure. “Frequent failures include that the functional groups are not inserted with the right angle or the neighboring functional groups overlap.” [15] In PooMa (3.0) I made use of SMILES (Simplified Molecular Input Line Entry System) together with the RDKit program, and developed with the purpose of describing the molecular structures efficiently. Using SMILES to describe molecular structures, the latest PooMa (3.0) can easily be used to study almost every class of materials — even macro-molecules or monomers. The genetic algorithm was adapted to SMILES while keeping its efficiency. In addition, the latest electronic-structure calculator SCC-DFTB (Self-consistent-charge density-functional-tight-binding method, also called DFTB+) is also more accurate in calculating excited states and charge transfer process but still efficient, compared to the previous electronic calculator DFTB (density-functional-tight-binding). More and very complicated properties are taken into account. Chapter 2 gives a detailed introduction to SMILES, the genetic algorithm, DFTB, DFTB+, and the theoretical methods to analyze output molecules by PooMa.

During my Ph.D., I improved our inverse-design approaches for optimizing molecules for solar cells or for studying the role of the backbone when optimizing functional groups for properties. The details of the work are presented in my publications in Chapter 3.

Chapter 2

Methodology

This chapter gives a detailed introduction to the methodology used in developing PooMa and analyzing molecules that result from the application of PooMa.

2.1 Genetic algorithms (GAs)

Genetic algorithms (GAs) are adaptive heuristic search algorithms and belong to evolutionary algorithms. GAs were originally inspired by the ideas of natural selection and genetics. Despite certain randomness, the genetic algorithm is not a completely random search, but directs the search intelligently into the region of better performance in solution space. GAs can be used for optimization problems and search problems, and usually get satisfactory solutions [20].

GAs is designed to simulate the process of natural selection, i.e. the individuals who can adapt at best to the environment can survive, reproduce, and go to the next generation. This “survival of the fittest” is simulated until the survived individuals don’t change for some consecutive generations, then the problem is seen as having identified an optimal solution. “Each generation is made up of a population of individuals, while each individual represents a possible solution in the search space. Each individual is represented as a string of character/integer/float/bits. This string is analogous to the chromosome in genetics.” [20]

2.1.1 Foundations

The foundations of GAs are described below, which use analogies of genetic structure and behavior of chromosomes of the population.

- Individuals of the population compete for resources and mates.

- The successful (fittest) individuals have the chance to mate and create offsprings so that the “good genes” of “fittest” parents propagate throughout the generations.
- Sometimes an offspring is better than both parents.
- Each subsequent generation becomes increasingly adaptive to the environment.

2.1.2 Search space

The population of individuals is kept within search space. Each individual represents a solution in search space for the given problem, and is usually a string (analogous to chromosome, see Fig. 2.1) which is composed of characters/integers/floats/bits (analogous to genes, see Fig. 2.1) [20].

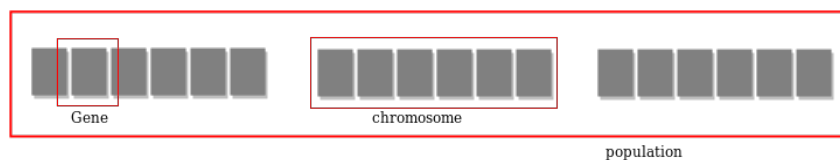


Fig. 2.1 Illustration of the search space of the genetic algorithm with the analogies of genes and chromosomes.

2.1.3 Fitness score

Each individual is attributed a “fitness score” as the ability of the individual to “compete”. Genetic algorithms seek the individuals with optimal (or near optimal) fitness scores which are able to survive among several consecutive generations. Each generation has a population of N individuals (chromosome/solutions). The individuals with better fitness scores are given chance to mate and produce offsprings by combining the chromosomes of parents. The process of mating and reproducing will be terminated when all of the old population have mated. The population size is static, therefore, by comparing the fitness scores, “unfit” individuals will always die and make room for new better fitter individuals to enter. So each new generation will never be worse than the previous generation. If no offspring is able to improve the overall performance for some successive generations, the genetic algorithm is said to be converged, and (near) optimal solutions have been found.

2.1.4 Operators

After creating the initial generation, the genetic algorithms use the following operators to generate new individuals.

- **Selection Operator** is used to select the individuals with good fitness scores and allow them to mate. [20]
- **Crossover Operator** is used for mating between individuals. After selecting two individuals by selection operator and choosing crossover sites randomly, the genes at these crossover sites will be exchanged [20] by the crossover operator, thus creating two completely new offspring. Fig. 2.2 gives an example.

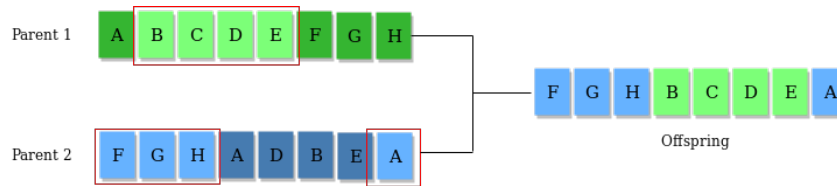


Fig. 2.2 An example of the crossover operator to exchange the genes of two parents and generate two offsprings. Only one of the offsprings is shown.

- **Mutation Operator** is used to insert random genes in an offspring to maintain the diversity in the population to avoid premature convergence. [20] Fig. 2.3 gives an example.



Fig. 2.3 An example of the mutation operator to mutate the gene of an individual.

2.1.5 The flowchart

The pseudo code of the genetic algorithm is as the followings:

1. Randomly initialize the population
2. Determine the fitness of the members of the initial population
3. Until population converge in consecutive generations, repeat the following process:
 - (a) Perform selection operator to select parents from the population
 - (b) Perform crossover operator to generate new offsprings. If the crossover is not able to generate new individuals in the late stage when many individuals have been evaluated, abandon the crossover and randomly generate new individuals as offsprings.

- (c) Perform mutation operator on the offsprings or skip the mutation randomly.
- (d) Calculate fitness for the offsprings
- (e) Compare the population with the offsprings. New fitter individuals will enter the population and replace the least fit individuals.

2.1.6 The advantages and disadvantages

“Many variants of genetic algorithms have been developed and applied to a wide range of optimization problems, from graph coloring to pattern recognition, from discrete systems (like the traveling salesman problem) to continuous systems (like the efficient design of airfoil in aerospace engineering), and from financial markets to multi-objective engineering optimization.” [21]

Compared to traditional optimization algorithms, GAs have many advantages. One is the ability to deal with various types of optimization, some are even complex, whether stationary or non-stationary (i.e. independent of time or not), linear or nonlinear, continuous or discontinuous, with or without random noise. Another advantage is parallelism. For example, the population can explore the search space in many directions simultaneously [21]. Different parameters, even different groups of individuals can be manipulated at the same time. Therefore, the parallelism of GAs can be easily implemented. Moreover, GAs are robust and can provide optimization over a large space. Unlike traditional artificial intelligence, genetic algorithms do not fail due to a slight change of input or in presence of noise. [20]

However, genetic algorithms also have some disadvantages [22]:

- For complicated problems, the repetitive evaluation of the fitness function is very time-consuming. “Then it may be necessary to forgo an exact evaluation and use an approximated fitness that is computationally efficient.” [22]
- As the complexity of the problem increases, the performance of genetic algorithms decreases. Therefore, it is necessary to break down the complicated problems into the simplest possible problems, for example, designing airfoils rather than the whole aircraft.
- The “better” solution by the genetic algorithm is only in comparison to other solutions. As a result, the stop criterion is not clear. [22]
- In many problems, genetic algorithms tend to converge towards local optima rather than the global optimum of the problem, which depends on the shape of the fitness landscape. To alleviate the problem, people may use a different fitness function, increase the rate

of mutation, or use selection techniques that maintain a diverse population of solutions, or simply replace part of the population with randomly generated individuals when most of the population is too similar to each other. [22]

- It is difficult to operate on dynamic data whose search space may be changed and the target solution of optimization may change after including new data.
- GAs are not suitable for classification problems or decision problems (whose fitness measure is right or wrong), because there is no way to find the converged solution, i.e. no hill in the fitness landscape.
- For some instances, other optimization algorithms may be more efficient than GAs in the speed of convergence. [22]

To sum up, when applying the genetic algorithm, many things need to be taken care of, such as the formulation of the fitness function, the size of the population, the choice of important parameters including the rate of mutation and crossover, as well as the selection criteria of the new population. [22] “Any inappropriate choice will make it difficult for the algorithm to converge, or it will simply produce meaningless results. Despite these drawbacks, genetic algorithms remain one of the most widely used optimization algorithms in modern nonlinear optimization.” [21]

2.2 The electronic-structure calculator: DFTB and TD-DFTB

Density Functional Theory (DFT) is the benchmark method to calculate electronic structure for large and complicated systems with good accuracy. When systems and time scales get larger, force-field models instead are the usual techniques to adopt. Between these methods are semi-empirical methods, which are derived from approximations to Hartree-Fock or DFT based methods. Especially, density functional based tight binding (DFTB) [23] is a simplification of Kohn-Sham DFT to a tight binding form (which only considers valence electrons while the core electrons are fixed), which reduces the complexity of the DFT method.

Since our inverse-design approaches are based on calculating tens of thousands of molecules on the fly, we use DFTB and TD-DFTB (time-dependent DFTB) as the electronic-structure calculator, which is a result of the compromise between accuracy and computational cost.

2.2.1 DFTB model

Expansion of the total energy

The total energy of the DFTB models can be seen as expanding the total energy of Kohn-Sham (KS) DFT into a Taylor series, see eq. (2.1). The Taylor series can be first order (leading to the DFTB1 model), or up to second order (giving the DFTB2 model) or up to third order (DFTB3 model). The chosen reference density ρ_0 is the superposition of neutral atomic densities, which is perturbed by the fluctuations for the ground state density, i.e. $\rho(r) = \rho_0(r) + \delta\rho(r)$. [23]

$$E^{DFTB3}[\rho_0 + \delta\rho] = E^0[\rho_0] + E^1[\rho_0, \delta\rho] + E^2[\rho_0, (\delta\rho)^2] + E^3[\rho_0, (\delta\rho)^3] \quad (2.1)$$

with

$$E^0[\rho_0] = \frac{1}{2} \sum_{AB} \frac{Z_A Z_B}{R_{AB}} - \frac{1}{2} \iint \frac{\rho_0(\mathbf{r})\rho_0(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}d\mathbf{r}' - \int V^{xc}[\rho_0]\rho_0(\mathbf{r})d\mathbf{r} + E^{xc}[\rho_0] \quad (2.2)$$

$$E^1[\rho_0, \delta\rho] = \sum_i n_i \langle \psi_i | \hat{H}[\rho_0] | \psi_i \rangle \quad (2.3)$$

$$E^2[\rho_0, (\delta\rho)^2] = \frac{1}{2} \iint \left(\frac{1}{|\mathbf{r} - \mathbf{r}'|} + \frac{\delta^2 E^{xc}[\rho]}{\delta\rho(\mathbf{r})\delta\rho(\mathbf{r}')} \Big|_{\rho_0} \right) \delta\rho(\mathbf{r})\delta\rho(\mathbf{r}') d\mathbf{r}d\mathbf{r}' \quad (2.4)$$

$$E^3[\rho_0, (\delta\rho)^3] = \frac{1}{6} \iiint \frac{\delta^3 E^{xc}[\rho]}{\delta\rho(\mathbf{r})\delta\rho(\mathbf{r}')\delta\rho(\mathbf{r}'')} \Big|_{\rho_0} \times \delta\rho(\mathbf{r})\delta\rho(\mathbf{r}')\delta\rho(\mathbf{r}'') d\mathbf{r}d\mathbf{r}'d\mathbf{r}'' \quad (2.5)$$

with XC being the exchange correlation energy and potential.

In total, three DFTB models have been developed. The DFTB1 model [24, 25] only considers the first order and non-self-consistent charge, and is implemented in PooMa (1.0, 2.0, 2.1). DFTB1 was originally called DFTB or non-SCC DFTB. DFTB2 model [26] considers the second order and self-consistent charge (SCC) for density fluctuations. DFTB2 model is implemented in PooMa (3.0), and was originally called SCC-DFTB. DFTB3 model [23] considers the third order and is developed to improve the calculations for hydrogen-bonded complexes and proton affinities. It was not adopted in PooMA, because it doesn't improve the accuracy much for the systems of our interest while requiring more computational costs. Here I will give a more detailed introduction on DFTB1 and DFTB2 which were used in PooMA, but only a brief introduction on DFTB3 for completeness.

DFTB1

DFTB1 method uses three approximations: only $E^0[\rho_0]$ (eq. (2.2)) and $E^1[\rho_0, \delta\rho]$ (eq. (2.3)) are taken into account; the orbitals Ψ_i are represented by a valence-only minimal basis set (ϕ_μ) (eq. (2.6)) using LCAO (linear combination of atomic orbitals) ansatz; the hamiltonian operator $\hat{H}[\rho_0]$ is limited to a two-center approximation. [23]

$$\Psi_i = \sum_{\mu} c_{\mu i} \phi_{\mu} \quad (2.6)$$

The minimal atomic orbital basis set ϕ_{μ} is computed as eq.(2.7), solving the atomic Kohn-Sham equations in DFT with an additional confining potential to confine the atomic orbitals. So the orbitals are slightly compressed atomic-like which is realistic when describing binding situations. r_0 depends on the specific atoms or bond types, which can be found in publications when doing parameterization. $\hat{H}[\rho_0]$ in eq. (2.3) is the Hamilton operator depending on the reference atomic densities ρ_A of neutral atoms $\{A\}$. To reduce the computational cost, it makes use of the frozen core approximation, which only considers the valence orbitals.

$$\left[-\frac{1}{2} \nabla^2 + V^{eff}[\rho^{atom}] + \left(\frac{r}{r_0} \right)^n \right] \phi_{\mu} = \varepsilon_{\mu} \phi_{\mu} \quad (2.7)$$

The Hamiltonian matrix elements can be represented using a linear combination of atomic orbitals (LCAO) basis. The diagonal elements $H_{\mu\mu}^0$ can be seen as the atomic eigenvalues. The nondiagonal elements $H_{\mu\nu}^0$ are calculated in a two-center approximation (neglecting the three center terms) as eq.(2.8). The Hamiltonian matrix elements only need to be computed once, normally using the dimers as the model to calculate Kohn-Sham equations and obtaining the H^0 values corresponding to different element pairs and inter-atomic distance.

$$H_{\mu\nu}^0 = \langle \phi_{\mu} | \hat{H}[\rho_0] | \phi_{\nu} \rangle \approx \left\langle \phi_{\mu} \left| -\frac{1}{2} \nabla^2 + V[\rho_A + \rho_B] \right| \phi_{\nu} \right\rangle, \mu \in A, \nu \in B \quad (2.8)$$

According to eq.(2.2), the total energy $E^0[\rho_0]$ is only related to the reference density and not related to the specific system where $\delta\rho$ gets involved. So it can be calculated in advance and then be applied to the specific system. In DFTB model, $E^0[\rho_0]$ is approximately defined as a sum of pair potentials, which are called repulsive energy, (see eq. (2.9)). The V_{AB} are determined by fitting to DFT or empirical results. Forces are calculated with the Hellmann–Feynman theorem and derivatives of the energy. [23]

$$E^0[\rho_0] \approx E_{rep} = \frac{1}{2} \sum_{AB} V_{AB}^{rep} \quad (2.9)$$

The DFTB1 model has good performance if the electron density of the system can be seen as a sum of atomic density. However, the inaccuracy of DFTB1 will increase when chemical bonding is influenced significantly by a delicate charge balance between the atomic components, particularly in molecules with hetero nuclear and polar semiconductors. So it is necessary to extend the DFTB with E^2 and E^3 terms in order to improve the description of total energies, forces, and transferability in the presence of considerable long-range Coulomb interactions [26].

DFTB2 and DFTB3

To approximate E^2 (see Eq. (2.4)), the density fluctuations are defined as the superposition of atomic components, which approximately decay exponentially and isotropically as shown in eq. (2.10).

$$\delta\rho(\mathbf{r}) = \sum_A \delta\rho_A(\mathbf{r} - \mathbf{R}_A) \approx \frac{1}{\sqrt{4\pi}} \sum_A \left(\frac{\tau_A^3}{8\pi} e^{-\tau_A|\mathbf{r}-\mathbf{R}_A|} \right) \Delta q_A \quad (2.10)$$

Then, the second order term E^2 can be written in eq.(2.11) if neglecting the exchange correlation (XC) contributions. E^2 needs to be solved self-consistently because E^2 depends on the Mulliken charges q_A (since $\Delta q_A = q_A - Z_A$, Z_A refers to the charge for the neutral atom) which in turn depend on the molecular orbital coefficients, $c_{\mu i}$. [23] Therefore, DFTB2 model is also called self-consistent-charge DFTB (SCC-DFTB).

$$E^2(\tau_A, \tau_B, R_{AB}) = \frac{1}{2} \sum_{AB(\neq A)} \gamma_{AB}(\tau_A, \tau_B, R_{AB}) \Delta q_A \Delta q_B \quad (2.11)$$

The third order term is defined as eq.(2.12), which is important when Δq_A is large, i.e. the charge is significantly different from that of the neutral atom. Besides, DFTB3 modifies Γ_{AB} to describe the interactions between hydrogen and first row elements better [23].

$$E^{DFTB3} = \sum_i \sum_{AB} \sum_{\mu \in A} \sum_{\nu \in B} n_i c_{\mu i} c_{\nu i} H_{\mu\nu}^0 + \frac{1}{2} \sum_{AB} \Delta q_A \Delta q_B \gamma_{AB}^h + \frac{1}{3} \sum_{AB} \Delta q_A^2 \Delta q_B \Gamma_{AB} + \frac{1}{2} \sum_{AB} V_{AB}^{rep} \quad (2.12)$$

2.2.2 TD-DFTB model

Runge and Gross laid the foundations for time-dependent density functional theory (TD-DFT) by generalizing the Hohenberg and Kohn's theorem to time-dependent external potentials. On that basis, Casida developed the formalism for calculating the linear response of the

electron density to a perturbation in the external potential, and derived an eigenvalue equation for excited states with single orbital transitions, which is the most popular method now to calculate excited states and their properties [27].

Nonetheless, TD-DFT calculations for excited states result in much more expensive computational cost, compared to DFT calculations on ground states. There are different strategies to reduce the cost and complexity, including dividing into subsystems, omitting some terms, cutting off the space for single orbital transitions, and approximating the integrals. As one of the methods to approximate integrals, Niehaus *et al.* adapted Casida's linear response approach to DFTB frame, and developed TD-DFTB (time-dependent density functional theory based tight binding). TD-DFTB is found to produce good results for $\pi \rightarrow \pi^*$ transitions and is suitable for calculating UV-Vis spectra with significantly reduced computational costs and applicable to large systems. TD-DFTB has been applied where TD-DFT is not feasible. On the other hand, TD-DFTB starts from the electronic structure calculated by DFTB, which leads to the inherited inaccuracy of TD-DFTB. In addition, TD-DFTB is restricted to systems which only involve elements whose DFTB parameters are available. [27]

To calculate excitation energies, first, a self-consistent field calculation is done to get the single-particle Kohn-Sham orbitals ψ_i and the orbital energies ε_i . Secondly, the coupling matrix is built, in eq.(2.13), which gives the response of the self-consistent field when experiencing a change in the electron density. Then, for closed-shell systems, the excitation energies (w_I) are calculated by solving the eigenvalue problem of eq.(2.14). In the end, the oscillator strengths can be calculated using eq.(2.15).

$$K_{ij\sigma,kl\tau} = \iint \psi_i(\mathbf{r})\psi_j(\mathbf{r}) \times \left(\frac{1}{|\mathbf{r}-\mathbf{r}'|} + \frac{\delta^2 E_{xc}}{\delta\rho_\sigma(\mathbf{r})\delta\rho_\tau(\mathbf{r}')} \right) \times \psi_k(\mathbf{r}')\psi_l(\mathbf{r}') d\mathbf{r}d\mathbf{r}' \quad (2.13)$$

where σ and τ are spin indices.

$$\sum_{ij\sigma} [w_{ij}^2 \delta_{ik} \delta_{jl} \delta_{\sigma\tau} + 2\sqrt{w_{ij}w_{kj}} K_{ij\sigma,kl\tau} \sqrt{w_{kj}}] F_{ij\sigma}^I = w_I^2 F_{kl\tau}^I \quad (2.14)$$

where $w_{ij} = \varepsilon_j - \varepsilon_i$ (i, k are occupied Kohn-Sham orbitals, whereas j, l are unoccupied ones).

$$f_I = \frac{2}{3} w_I \sum_{k=x,y,z} \left| \sum_{ij} \langle \psi_i | \vec{r}_k | \psi_j \rangle \sqrt{\frac{w_{ij}}{w_I}} (F_{ij\uparrow}^I + F_{ij\downarrow}^I) \right|^2 \quad (2.15)$$

where the transition-dipole matrix elements are given in eq.(2.16).

$$\langle \psi_i | \mathbf{r} | \psi_j \rangle = \sum_{\alpha} \mathbf{R}_{\alpha} q_{\alpha}^{ij} \quad (2.16)$$

2.3 SMILES — A Simplified Chemical Language

PooMa (3.0) is developed by making use of SMILES (Simplified Molecular Input Line Entry System), which is a very useful line notation to represent molecules, even reactions. SMILES is very useful, popular and requires very little data space compared to other methods of representing molecular structures. With SMILES, people don't need to draw the 2D picture to describe the molecule. SMILES is truly a language with vocabulary and grammar rules.

SMILES encoding rules

There are five general SMILES encoding rules, corresponding to specification of atoms, bonds, branches, ring closures, and disconnections. [2]

1. **Atoms** are required to be represented by their atomic symbols, while the second letter must be the lower case for elements with two letters, for example Cl or Br. The atoms don't need to have brackets if their valences are "normal valences" and elements are in the organic subset. Otherwise, the brackets are required. The organic subset includes B, C, N, O, P, S, F, Cl, Br, and I. The "normal valences" are B (3), C (4), N (3,5), O (2), P (3,5), S (2,4,6), and halogens(1). Within the brackets, the number of attached hydrogens must be clarified by the number following symbol H. Similarly, the charge needs to be specified by the (possibly repeated) occurrence of + or - (for example [Fe+++] is iron (III) cation, [OH-] is hydroxyl anion), or alternatively a number following the symbol + or - (for example [Fe+3] is also iron (III) cation, [Fe+2] is iron (II) cation). If unspecified, it is assumed that there is no attached hydrogen and charge. Atoms in aromatic rings are denoted with lower case letters, for example, capital letter C for aliphatic carbon while lower case c for the aromatic carbon. Examples are the following ones: C for methane (CH₄), P for phosphine (PH₃), N for ammonia (NH₃), S for hydrogen sulfide (H₂S), O for water (H₂O), Cl for hydrochloric acid (HCl), [S] for elemental sulfur, [Au] for elemental gold.
2. **Bonds** can take one out of four types, single, double, triple, and aromatic bonds, which are represented respectively by the symbols -, =, #, and : . The single and aromatic bonds are always omitted, while double and triple bonds have to be explicitly

stated. For example, CC for ethane (CH₃CH₃), C=O for formaldehyde (CH₂O), C=C for ethene (CH₂=CH₂), O=C=O for carbon dioxide (CO₂), COC for dimethyl ether (CH₃OCH₃), C#N for hydrogen cyanide (HCN), CCO for ethanol (CH₃CH₂OH), [H][H] for molecular hydrogen (H₂).

- Branches:** are denoted by enclosing them in parentheses and can be nested or stacked. The branches are always assumed to connect to the atom on the left. For example, CCN(CC)CC for triethylamine, CC(C)C(=O)O for isobutyric acid, C=CC(CCC)C(C(C)C)CCC for 3-propyl-4-isopropyl-1-heptene.
- Cyclic Structures** are denoted by breaking one bond in each ring. For example, cyclohexane in Fig. 2.4 [2] exemplifies how to represent cyclic structures. It needs to be noted that the “breaking bond” can be any bond in the ring, so different valid notations for the same system can be generated.

A digit is denoted following the two atomic symbols between which the bonds are broken. (The digits can be repeatedly used, for example in Fig. 2.5, the digit 1 is used twice in the specification.). Then, the SMILES are constructed based on the above rules. Even breaking the same bond in the ring, different ways to denote branches can generate different valid SMILES. For example in Fig. 2.6 for the same structure of 1-methyl-3-bromo-cyclohexene-1, (a) and (b) see Br or CCC1 respectively as branches and generate two valid SMILES. A possible case is that a single atom may have more than one ring closure. For example in Fig. 2.7 [2], cubane has two atoms related to two rings' closure if breaking the three bonds, generating the SMILES for cubane as C12C3C4C1C5C4C3C25.

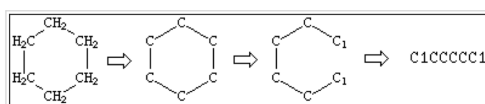


Fig. 2.4 Cyclohexane as an example of how to represent cyclic structures [2].

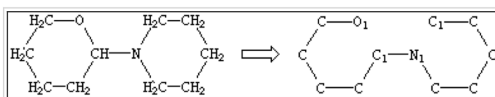


Fig. 2.5 Digit 1 is used twice in SMILES: O1CCCCC1N1CCCCC1, [2], as an example to illustrate that the digit can be repeatedly used when denoting cycles in the SMILES representation.

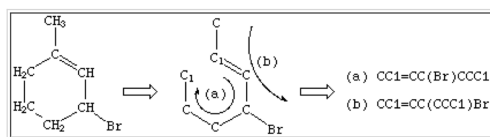


Fig. 2.6 Two valid SMILES notations for the same structure of 1-methyl-3-bromocyclohexene-1

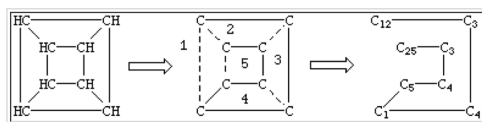


Fig. 2.7 Generating SMILES for cubane: C12C3C4C1C5C4C3C25.

5. **Disconnected Compounds:** are written as individual structures separated by a “.” (period). For example, the SMILES for sodium phenoxide can be written as [Na+].[O-]c1ccccc1. The SMILES of an ion can be nested, for example sodium phenoxide can also be written as c1cc([O-].[Na+])ccc1. There is no default requirement to the charges and the net total charge can be non-zero. In short, the adjacent atoms separated by period “.” implies that the atoms are not bonded to each other.

Canonicalization

Although the same structure may have different valid SMILES as illustrated above, a unique SMILES for each structure exists. These SMILES are called “generic SMILES” if atoms and bonds are denoted without chiral or isotopic information. On the contrary to possibly many valid “generic SMILES” for a given structure, people can generate only one special generic SMILES through the canonicalization algorithm to represent the structure. This special generic SMILES is called “canonical SMILES” or “unique SMILES”. Different software (like openbabel and RDkit programs) may use different canonicalization algorithms, therefore the “canonical SMILES” may be different for the same structure in different software. However, within the specified software, the “canonical SMILES” is unique for the given structure. For example, the generic SMILES “OCC”, “[CH3][CH2][OH]”, “C-C-O” or “C(O)C” all correspond to the canonical SMILES “CCO”. Another example can be the generic SMILES “OC(=O)C(Br)(Cl)N”, “ClC(Br)(N)C(=O)O” or “O=C(O)C(N)(Br)Cl” all correspond to the canonical SMILES “NC(Cl)(Br)C(=O)O”. [2]

If SMILES include not only atoms and bonds but also chiral and isotopic information, the SMILES are called isomeric SMILES. A canonical isomeric SMILES is called absolute SMILES.

Isomeric SMILES

Isomeric SMILES use SMILES rules about isotopes, configuration of double bonds and chirality.

1. **Isotopes** are specified by adding the atomic mass inside brackets before the atomic symbol. For example, [12C] for carbon-12, [13C] for carbon-13, [C] for carbon (unspecified mass), [13CH4] for C-13 methane.
2. **Configuration Of Double Bonds** are written with “directional bonds” / and \ which can be seen as single bond but indicate relative directions of the bonds. For example, F/C=C/F or F\C=C\F can represent E-1,2-difluoroethene, while F/C=C\F or F\C=C/F can represent Z-1,2-difluoroethene.

Especially, SMILES makes it possible to partially specify cis-trans isomerism by using local representation (rather than absolute representation). An example is shown in Fig. 2.8.

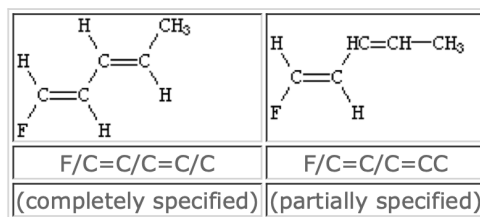


Fig. 2.8 SMILES allows partial specifications for cis-trans isomerism [2].

3. **Configuration Around Tetrahedral Centers:** Tetrahedral is the simplest and most common type of chirality, where four atoms are arranged around the central atom which is called the chiral center. If all the four neighbors are different, the two mirror structures will be different from each other. The two mirror structures are called enantiomers and are the only two forms that a tetrahedral center can have. If two (or more) of the four atoms around the central atom are identical to each other, the center atom will have no chirality because the mirror structures can be overlapped in space.

In SMILES, the chiral centers are indicated by the symbol @ or @@ following the atomic symbol of the chiral atom. The symbol @ indicates that the four neighboring atoms are listed anticlockwise. The symbol @@ indicates that the four neighboring atoms are listed clockwise. If there is no clarification about a possible chiral atom, it is also acceptable to not specify chirality.

For example in Fig. 2.9, if looking from the amino N to the chiral C (as the SMILES is written), the other three neighbors appear anticlockwise in the order that they are written in the top SMILES, N[C@](C)(F)C(=O)O (methyl-C, F, carboxy-C), and clockwise in the bottom one, N[C@@](F)(C)C(=O)O. [2]

Actually besides tetrahedral, SMILES can also handle other chirality, such as allene-like, square-planar, trigonal-bipyramidal, and octahedral, etc. The detailed rules can be found in [2].

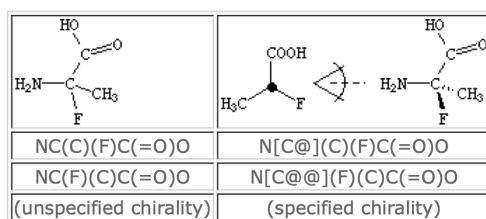


Fig. 2.9 Specified and unspecified chirality representations in SMILES [2].

SMILES Conventions

In addition, SMILES also have some commonly used conventions in the following aspects.

1. **Hydrogens** are normally not specified for organic structures. There are only four following exceptions which require explicit hydrogen specification. The first case is the charged hydrogen, like a proton [H+]. The second case is when the hydrogen is placed adjacent to other hydrogens, like a molecular hydrogen, [H][H]. The third case is that the hydrogens are connected to more than one atom, for example the bridging hydrogens. The fourth case is isotopic hydrogen, like heavy water [2H]O[2H].
2. **Aromaticity** can be deduced by aromaticity-detection algorithms in SMILES even the user input aliphatic (Kekulé-like) structure. However, people are used to entering aromatic structures directly by using lower case atomic symbols according to the rules. Only C, N, O, P, S, As, Se, and * (wildcard) can be considered aromatic. Exocyclic double bonds don't influence aromaticity.

To identify whether a molecule is aromatic or not, Hückel's rule is applied in the aromaticity-detection algorithm. The aromaticity is satisfied when all the atoms in the ring are sp^2 hybridized and the number of p-electrons satisfies Hückel's $4N+2$ criterion. The result is irrelevant for writing the SMILES in the aromatic way or the aliphatic way. For example, no matter whether writing benzene as c1ccccc1 or C1=CC=CC=C1 (the Kekulé form), the aromaticity-detection algorithm will detect the aromaticity

and convert the input SMILES to the aromatic way c1ccccc1. An opposite example could be cyclobutadiene and cyclooctatetraene which are anti-aromatic structures and whose output SMILES by the aromaticity-detection algorithm will be C1=CC=C1 and C1=CC=CC=CC=C1, although the user may input the wrong aromatic SMILES c1ccc1 and c1cccccc1.

- Others:** There are some other conventions in [2] about tautomer and bond conventions. The user rather than SMILES decide which tautomeric structure or valence conventions to be used to represent the structure. For example, nitromethane can be written as CN(=O)=O or the charge separated C[N+](=O)[O-].

2.4 UV-Vis spectroscopy

UV-Vis spectroscopy is treated by our inverse-design approaches and very important for solar cells. UV-Vis spectroscopy is based on the transitions of valence electrons. In molecules there are among others three different types of valence electrons, σ electrons for single bonds, π electrons for unsaturated bonds, and n electrons of oxygen, nitrogen, sulfur and halogen. Fig. 2.10 shows the three types of valence electrons in the aldehyde group. The energy level increases in the order $\sigma < \pi < n < \pi^* < \sigma^*$. There are four types of transitions and they decrease energetically in the order: $\sigma \rightarrow \sigma^* > n \rightarrow \sigma^* > \pi \rightarrow \pi^* > n \rightarrow \pi^*$, as shown in Fig. 2.11.

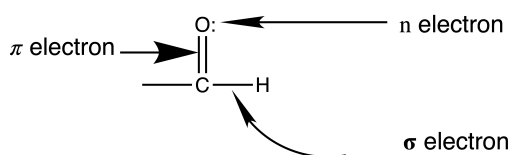


Fig. 2.10 Three types of valence electrons in aldehyde group. [3]

The transition from HOMO (highest occupied molecular orbital) to LUMO (lowest unoccupied molecular orbital) has the lowest transition energy. Neglecting relativistic effects (i.e. considering systems with solely lighter elements), the transitions are allowed when the spin doesn't change. On the contrary, the transitions are forbidden when the spin changes. This is called the spin selection about electron transitions. Transitions between states of the same parity are not allowed, which applies to chromophores that are centrosymmetric. So transitions $g \rightarrow u$ and $u \rightarrow g$ are allowed while $g \rightarrow g$ and $u \rightarrow u$ are forbidden. This is called Laporte selection. Violation of Laporte and spin selection rules will result in low intensity for the transition.

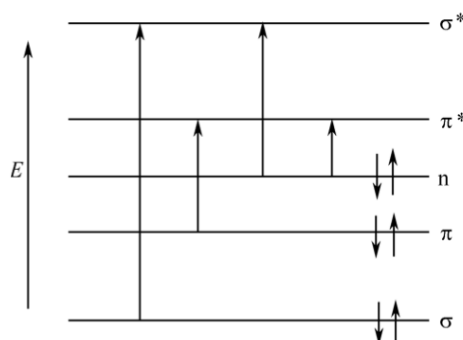


Fig. 2.11 The transition energies for different types of transitions. [3]

2.4.1 Important terminologies in UV-Vis spectroscopy

- **Chromophore** is defined as an isolated covalently bonded group that shows a characteristic absorption in UV/Visible region. The chromophores are normally functional groups with π electrons, such as carbonyl, double bond, triple bond, benzene ring, nitro, etc. The transitions are normally $\pi \rightarrow \pi^*$ and $n \rightarrow \pi^*$.
- **Auxochrome** is a functional group of atoms attached to the chromophore which modifies the ability of the chromophore to absorb light, altering the wavelength or intensity of the absorption [28]. But auxochromes themselves do not absorb ultraviolet or visible radiation. Auxochromes are normally atoms or functional groups with lone pairs of electrons which have p- π conjugation effects (with the π bond in chromophores), such as hydroxyl, alkoxy, amino, halogen, aldehyde, methyl mercaptan, etc.
- **Bathochromic Shift or Red shift** occurs when the absorption maximum is shifted towards a longer wavelength or lower energy, i.e. moving the band towards the direction of red color in the spectrum. This can occur when introducing auxochromes or changing the solvent polarity.
- **Hypsochromic Shift or Blue Shift** is the case when the absorption maximum is shifted towards the shorter wavelength or higher energy, in contrast to bathochromic shift (red shift), i.e. moving the band toward the direction of blue color in the spectrum. The reason for this can be the occurrence of auxochromes or changes in solvent polarity.
- **Hyperchromic Effect** is observed when absorption intensity is increased. Normally, auxochromes can cause hyperchromic shift. For example benzene shows B-band (benzenoid band, the secondary band in UV-Vis spectra of aromatic compounds) at 256 nm and that the maximum of molar absorptivity ϵ_{max} is 200, whereas aniline shows

B-band at 280 nm and that the maximum of molar absorptivity ϵ_{max} is 1430. The increase in the value of ϵ_{max} is due to the hyperchromic effect of auxochrome amino group.

- **Hypochromic Effect** is seen when absorption intensity is decreased. The reason can be the distortion of the geometry due to a functional group. For example, biphenyl shows the absorption maximum at 250 nm and that the maximum of molar absorptivity ϵ_{max} is 19,000, whereas 2-methyl biphenyl shows the absorption maximum at 237 nm, and that ϵ_{max} is 10250. The hypochromic effect originates from the effect that the methyl group can distort coplanarity of the chromophore and make it lose conjugation.

2.4.2 Factors affecting the position of UV-Visible bands

- **Effects of conjugation:** Two or more chromophores are conjugated, leading to increasing the energy of HOMO, decreasing the energy of LUMO, reducing the transition energy of $\pi \rightarrow \pi^*$ transition (for example in Fig. 2.12 illustrated the tendency), shifting the absorption maxima to a larger wavelength (red shift), increasing the intensity (as shown in Fig. 2.13)

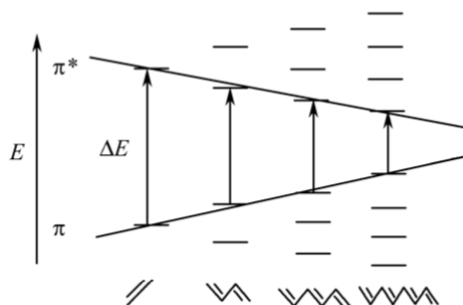


Fig. 2.12 The transition energy of $\pi \rightarrow \pi^*$ decreases as the conjugation system prolong. [3]

When the methyl group is connected to conjugation systems, it can also produce a weak bathochromic shift. Because the σ electron in C-H bond of the methyl group overlaps spatially with the π electron in the conjugation system, i.e. the hyperconjugation effect. It enlarges the conjugation space, reduces the transition energy of $\pi \rightarrow \pi^*$, and redshifts the absorption band.

- **Effect of Steric hindrance:** Steric hindrance can distort or destroy the co-planarity of the conjugated chromophores, leading to the conjugation effects being reduced or

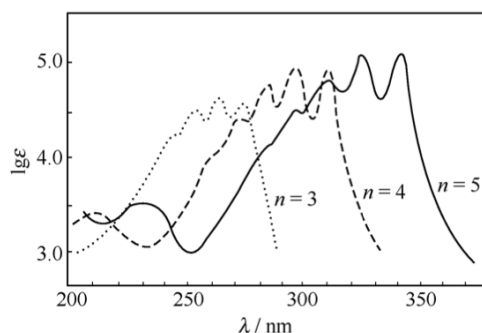


Fig. 2.13 UV-Vis spectroscopy of $\text{H}(\text{CH}=\text{CH})_n\text{H}$. [3]

disappearing. The absorption bands have blue shifts and the intensity of the bands are decreased, or the absorption bands disappeared.

For geometrical isomerism, trans-isomers often show absorption peaks at longer wavelength with more intensity, than cis-isomers due to the effects of steric hindrance. In cis-isomers, the planarity is often distorted more or less because bulky functional groups are on the same side of the double bond. For example, the trans-stilbene (see Fig. 2.14) and cis-stilbene obey this rule. Trans-stilbene has a more effective π - π overlap due to coplanarity, and the conjugation effect is stronger than cis-stilbene.

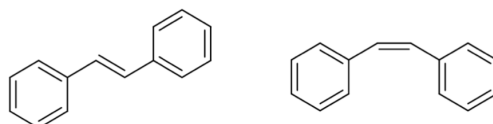


Fig. 2.14 Two isomers, trans-stilbene (left) and cis-stilbene (right).

- Effects of pH:** The changes in pH can make the conjugation system longer or shorter, leading to changes in the absorption peak, especially obvious for unsaturated acid, enol, phenol (see Fig. 2.15) and aniline (see Fig. 2.16). For example, in base, phenol exists in the form of phenoxide anion, which increases the conjugation through resonance (the lone pairs on the oxygen is delocalized over the π -system of the aromatic ring as shown in Fig. 2.17) and results in a bathochromic shift (red shift) and a hyperchromic effect. Another example, in acid, the NH_2 group in aniline exists in the form of NH_3^+ and the original p- π conjugation of NH_2 (in neutral or basic conditions) disappeared as shown in Fig. 2.18, which results in a hypsochromic shift (blue shift).

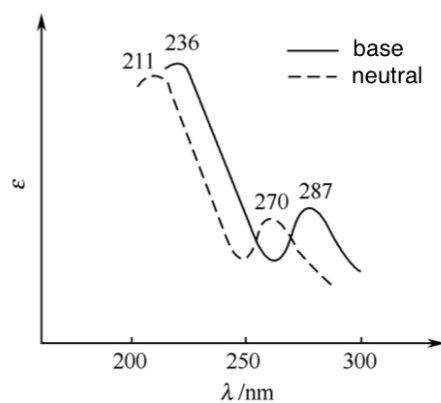


Fig. 2.15 UV-Vis spectroscopy of phenol. [3]

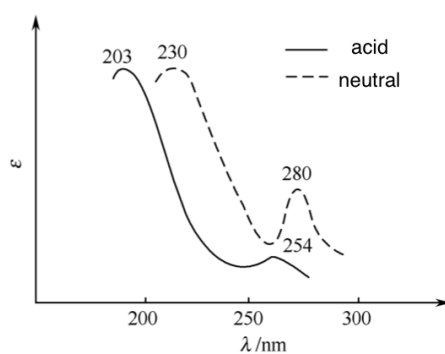


Fig. 2.16 UV-Vis spectroscopy of aniline. [3]

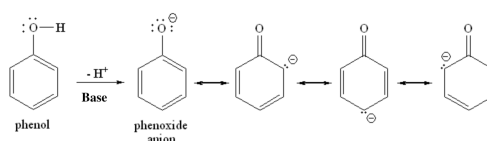


Fig. 2.17 Phenol and phenoxide anion in base

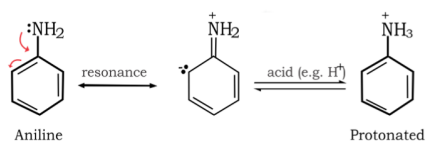


Fig. 2.18 Aniline and protonated aniline in acid

2.5 Theoretical methods to analyze output-molecules

The following theoretical methods are employed to have a deeper understanding of the molecules that result from our inverse design approaches.

2.5.1 DFT

Schrödinger equation

One of the foundation of quantum mechanics is the (time-independent) Schrödinger equation, see eq.(2.17).

$$\hat{H}\psi = E\psi \quad (2.17)$$

with \hat{H} is the Hamilton operator, ψ is the wave-function of the system (the eigenfunction of the equation), and E is the energy of the system (the eigenvalue of the equation). The Hamilton operator can be written as the sum of the potential energy and kinetic energy as eq.(2.18).

$$\hat{H} = -\frac{\hbar^2}{2m}\nabla^2 + V \quad (2.18)$$

where \hbar is $\frac{h}{2\pi}$, m is mass, V is potential energy, and ∇ is Laplace operator, eq.(2.19).

$$\nabla^2 = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \quad (2.19)$$

Here, x , y and z are coordinates of three-dimension Cartesian position.

For a more general molecular system with many particles (nuclei and electrons), the total Hamilton operator consists of five terms: the kinetic-energy operator for the nucleus, the kinetic-energy operator for the electrons, the potential-energy operator for nucleus-nucleus interactions, the potential-energy operator for electron-electron interactions, and the potential-energy operator for nucleus-electron interactions. [29] The Schrödinger equation becomes as eq.(2.20), where \vec{X} and \vec{x} are the coordinates for nuclei and electrons.

$$\hat{H}\psi = (\hat{H}_{k,n} + \hat{H}_{k,e} + \hat{H}_{p,n-n} + \hat{H}_{p,e-e} + \hat{H}_{p,n-e})\psi(\vec{X}, \vec{x}) = E \cdot \psi(\vec{X}, \vec{x}) \quad (2.20)$$

The Born–Oppenheimer approximation assumes that the nuclei can be seen as static. The physical idea is that electrons move much faster than nuclei. Based on the Born–Oppenheimer

approximation, the wavefunction ψ can be written in eq.(2.21) through a factor Ansatz.

$$\psi(\vec{X}, \vec{x}) = \psi_n(\vec{X}) \cdot \psi_e(\vec{X}; \vec{x}) \quad (2.21)$$

where $\psi_n(\vec{X})$ is the wavefunction describing the nuclei, and it only depends on the position of nuclei \vec{X} ; while $\psi_e(\vec{X}; \vec{x})$ is the wavefunction describing the electrons, and it depends not only on positions of nuclei \vec{X} but also on positions of electrons \vec{x} .

Subsequently the Schrödinger equation splits into two equations:

$$\hat{H}_e \psi_e(\vec{X}; \vec{x}) = (\hat{H}_{k,e} + \hat{H}_{p,e-e} + \hat{H}_{p,n-e}) \psi_e(\vec{X}; \vec{x}) = E_e \psi_e(\vec{X}; \vec{x}) \quad (2.22)$$

$$\hat{H}_n \psi_n(\vec{X}) = (\hat{H}_{k,n} + \hat{H}_{p,n-n}) \psi_n(\vec{X}) = E_n \psi_n(\vec{X}) \quad (2.23)$$

where $\hat{H}_{k,n}$ is zero due to the Born-Oppenheimer approximation that the nuclei are static. E_e is the electronic energy and E_n is the nuclei energy. So the total energy of the system is given in eq.(2.24).

$$E_{total} = E_e + E_n = E_e + \frac{1}{2} \sum_{k_1 \neq k_2=1}^M \frac{1}{4\pi\epsilon_0} \frac{Z_{k_1} Z_{k_2} e^2}{|\vec{R}_{k_1} - \vec{R}_{k_2}|} \quad (2.24)$$

Hartree-Fock approximation

To approximately solve the electronic Schrödinger equation (eq.(2.22)), one may make use of the Hartree-Fock method that uses a Slater determinant (eq.(2.25)) to approximately represent the wavefunction.

$$\psi_e \approx \phi(\vec{x}_1, \vec{x}_2, \dots, \vec{x}_N) = \frac{1}{\sqrt{N!}} \begin{vmatrix} \phi_1(\vec{x}_1) & \phi_2(\vec{x}_1) & \dots & \phi_N(\vec{x}_1) \\ \phi_1(\vec{x}_2) & \phi_2(\vec{x}_2) & \dots & \phi_N(\vec{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(\vec{x}_N) & \phi_2(\vec{x}_N) & \dots & \phi_N(\vec{x}_N) \end{vmatrix} \quad (2.25)$$

where ϕ_1, ϕ_2, \dots , and ϕ_N are single-electron wave-functions and the $\vec{x}_1, \vec{x}_2, \dots, \vec{x}_N$ are electron coordinates. According to linear algebra, if exchanging any two rows or columns, the sign of eq.(2.25) will change, which makes it satisfy the anti-symmetry principle. In addition, eq.(2.25) makes sure that the electrons all occupy different electron states, otherwise the rows will be equivalent and the determinant will be zero.

To sum up, the usually employed Hartree-Fock approach is based on the five following approximations [30]:

1. It is under the assumption of Born–Oppenheimer approximation described above.
2. It neglects relativistic effects. For example, “the momentum operator is assumed to be completely non-relativistic.” [30]
3. When applying the variational principle to determine an approximate solution, the solution is often written in terms of a linear combination of a finite number of basis functions. The finite basis set is assumed to be approximately complete. [30]
4. Each single-electron function in each Slater determinant describes an energy eigenfunction (an orbital).
5. The mean-field approximation is implied. So Hartree-Fock method doesn't consider electron correlation effects. Due to the mean-field approximation, sometimes the experimental results can differ from the computational results. The correlation effects are defined as:

$$E_{exact} - E_{HF} = E_{correlation} \quad (2.26)$$

Based on these five approximations, for closed-shell systems with doubly occupied orbital wave-functions, the Hartree-Fock equations for the individual orbital wavefunctions are as eq.(2.27).

$$\hat{F}(1)\phi_i(1) = \varepsilon_i\phi_i(1) \quad (2.27)$$

where (1) indicates that the operator is 1-electron in nature. [30] $\phi_i(1)$ are a set of one-electron wave functions, called the Hartree–Fock molecular orbitals. [30] $\hat{F}(1)$ is Fock operator, defined as eq.(2.28).

$$\hat{F}(1) = \hat{h}_1 + \sum_{j=1}^{N/2} [2\hat{J}_j(1) - \hat{K}_j(1)] \quad (2.28)$$

where \hat{h}_1 is that part of electronic Hamilton operator that is related to only one electron, i.e. $\hat{H}_{k,e}$ and $\hat{H}_{n,e}$ in eq.(2.22). $\hat{J}_j(1)$ is the Coulomb operator, which is the classical Coulomb potential from an electron in the j-th orbital. [30] $\hat{K}_j(1)$ is the exchange operator, which is quantum mechanical exchange operator originating from the fact that the total N-electron wave function is anti-symmetric when interchanging any two electrons. It is a result of this anti-symmetry and of the indistinguishability of the electrons. [30]

“Since the Fock operator depends on the orbitals used to construct the corresponding Fock matrix, the eigenfunctions of the Fock operator are in turn new orbitals, which can be

used to construct a new Fock operator. In this way, the Hartree–Fock orbitals are optimized iteratively until the change in total electronic energy falls below a predefined threshold. In this way, a set of self-consistent one-electron orbitals is calculated.” [30] Roothan suggested that instead of varying all orbitals in all points (i.e. vary infinitely many parameters), only a finite variation was considered by expanding the orbitals in a set of fixed basis functions as eq.(2.29).

$$\phi_l(\vec{x}) = \sum_{p=1}^{N_b} \chi_p(\vec{x})c_{pl} \quad (2.29)$$

where the basis functions χ as well as their number N_b have been chosen in advance and only the expansion coefficients c_{pl} are varied. Then Hartree-Fock-Roothan equation is obtained as eq.(2.30).

$$F \cdot c_l = \epsilon_l \cdot O \cdot c_l \quad (2.30)$$

where F is the Fock matrix, O is the overlap matrix, c_l is the vector containing (sought) coefficients of the l th orbital to the different basis functions. eq.(2.30) can be solved using standard matrix routines, which yields a new set of coefficients until the procedure is converged.

A set of basis functions is called a basis set. It can be composed of atomic orbitals (usual choice in quantum chemistry community, for example Gaussian-type orbitals, Slater-type orbitals, etc), plane waves (typically within the solid state community), or numerical atomic orbitals.

Due to the electronic wavefunction is much too complicated than what is needed for calculating experimental observables. Density functional theory (DFT) appeared, as an alternative to the Hartree-Fock method. In addition, DFT takes both exchange and correlation effects into consideration.

Denisty functional theory (DFT)

Density functional theory is used for studies of the properties of a many-electron system on the basis of electron densities in the system, instead of trying to get the approximate solution of the complicated wave-function and subsequently obtain the properties.

- **Hohenberg-Kohn theorems:** DFT was put on a firm theoretical foundation with the publication of the Hohenberg-Kohn theorems:

Theorem 1: The external potential $V_{ext}(r)$ (and hence the total energy E_{total}), is a unique functional of the electron density $\rho(r)$. [31] It states that once you know the ground-state electron density in position space, any ground-state property is uniquely

defined, i.e. any ground-state property is a functional of the electron density in position space.

Theorem 2: Only if the density is the ground state density, the system energy reaches the lowest energy, which means any other energies with random density (that, however, shall give the same number of particles) of the system is larger than energy with ground-state density. [31]

$$E_{GS}[\rho(r)] \geq E_{GS}[\rho_{GS}(r)] \quad (2.31)$$

where the subscript GS is the abbreviation of the ground state. $E_{GS}[\rho(r)]$ is the system energy calculated by random density, while $E_{GS}[\rho_{GS}(r)]$ is the minimum energy due to the density state is in the ground state. [32]

- **Kohn-Sham method**

However, the Hohenberg-Kohn theorems provide a formalistic proof for the validity of the correctness of density functional theory, but do not provide any practical scheme for calculating ground-state properties from the electron density. This was provided by Kohn-Sham method, based on the non-interacting Schrödinger equation (more clearly, Schrödinger-like equation) of a fictitious system (the “Kohn–Sham system”) of non-interacting electrons that generate the same density as the system of interest of interacting particles. [31] The Kohn–Sham equation contains a local effective (fictitious) external potential in which the non-interacting electrons move, typically denoted as $V_{eff}(r)$, called the Kohn–Sham potential [31]. The Kohn–Sham wavefunction is a single Slater determinant constructed from a set of orbitals that are the lowest-energy solutions to eq.(2.32)

$$\left(-\frac{\hbar^2}{2m} \nabla^2 + V_{eff}(r)\right)\varphi_i = \varepsilon_i \varphi_i \quad (2.32)$$

This eigenvalue equation is the typical representation of the Kohn–Sham equations. Here ε_i is the orbital energy of the corresponding Kohn–Sham orbital φ_i . [31]The density for an N -particle system is calculated from eq.(2.33).

$$\rho(r) = \sum_{i=1}^N |\phi_i(r)|^2 \quad (2.33)$$

where N is the number of electrons in the system.

Then, the total energy is expressed as a functional of the electron density as eq.(2.34).

$$E[\rho] = T[\rho] + \int V_{ext}\rho(\mathbf{r})d\mathbf{r} + E_c + E_{xc}[\rho] \quad (2.34)$$

where T is the Kohn-Sham kinetic energy, V_{ext} is the external potential acting on the interacting system (at minimum, for a molecular system, the electron–nuclei interaction), E_c is the classic Coulomb interaction energy (defined as $\frac{1}{2} \iint \frac{\rho(\vec{r}_1)\rho(\vec{r}_2)}{|\vec{r}_1-\vec{r}_2|} d\vec{r}_1 d\vec{r}_2$), E_{xc} is the exchange–correlation energy.

Then the Kohn–Sham potential V_{eff} is written as eq.(2.35).

$$V_{eff}(\mathbf{r}) = V_{ext}(\mathbf{r}) + V_c(\mathbf{r}) + V_{xc}(\mathbf{r}) \quad (2.35)$$

where V_c is the Coulomb potential, V_{xc} is the exchange–correlation potential (defined as $\frac{\delta E_{xc}[\rho]}{\delta \rho(\mathbf{r})}$). Local-density approximations (LDA) are a class of approximations to the exchange–correlation (XC) energy functional in density functional theory (DFT) that depend solely upon the value of the electronic density at each point in space (and not, for example, derivatives of the density or the Kohn–Sham orbitals), written in eq.(2.36). [33] A straightforward generalization of the LDA to include electron spin is local-spin-density approximation (LSDA) as eq.(2.37). Furthermore, if we let exchange–correlation potential V_{xc} depend also on $|\vec{\nabla}\rho|$ and $\nabla^2\rho, \dots$, the approximation is called generalized gradient approximations (GGA) or non-local-density approximations.

$$E_{xc}^{LDA}[\rho] = \int \rho(\mathbf{r})\epsilon_{xc}(\rho(\mathbf{r}))d\mathbf{r} \quad (2.36)$$

$$E_{xc}^{LSDA}[\rho_\alpha, \rho_\beta] = \int \rho(\mathbf{r})\epsilon_{xc}(\rho_\alpha, \rho_\beta)d\mathbf{r} \quad (2.37)$$

where the spin polarized system employs two spin-densities, ρ_α and ρ_β with $\rho = \rho_\alpha + \rho_\beta$.

Since the beginning of 1990s, hybrid methods appeared that led to more accurate energies and structures, by combining Hartree-Fock and density-functional treatments of exchange effects, whereas correlation effects are treated within the DFT scheme. For example, the popular B3LYP (Becke, 3-parameter, Lee–Yang–Parr) exchange–correlation functional is in eq.(2.38).

$$E_{xc}^{B3LYP} = (1-a)E_x^{LSDA} + aE_x^{HF} + b\Delta E_x^B + (1-c)E_c^{LSDA} + cE_c^{LYP} \quad (2.38)$$

where $a = 0.20$, $b = 0.72$, $c = 0.81$. ‘ e ’ denotes the exchange effect, while ‘ c ’ denotes the correlation effect. ‘ HF ’ is Hartree-Fock approximation; ‘ $LSDA$ ’ is local spin density approximation; ‘ B ’ and ‘ LYP ’ are functional by Becke and Lee, Yang and Parr respectively, which all belong to generalized-gradient approximation (GGA).

2.5.2 Fukui function

The Fukui function of Kenichi Fukui, is used to describe the change in the electron density of a molecular system due to a change of the total number of electrons. The condensed Fukui function is similar but for an atom rather than a point in space as in Fukui function. The Fukui function and condensed one are helpful to predict the most electrophilic and nucleophilic sites in a molecule system.

The Fukui function is defined as eq.(2.39) [34]:

$$f(r) = \left(\frac{\partial \rho(r)}{\partial N} \right)_v \quad (2.39)$$

where N is the number of electrons in the present system and the constant term v in the partial derivative is the external potential. “Generally the external potential only comes from nuclear charges, and thus v can be seen as a nuclear coordinate for an isolated chemical system. The partial derivative cannot be directly evaluated due to the discontinuity when N is an integer. To resolve this difficulty, the Fukui function is defined using one-sided derivatives.” [35, 36] ([35] is my publication in my master.)

- Nucleophilic attack:

$$f^+(r) = \left(\frac{\partial \rho(r)}{\partial N} \right)_v^+ = \lim_{\varepsilon \rightarrow 0^+} \frac{\rho_{N+\varepsilon}(r) - \rho_N(r)}{\varepsilon} = \rho_{N+1}(r) - \rho_N(r) \approx \rho^{LUMO}(r) \quad (2.40)$$

- Electrophilic attack:

$$f^-(r) = \left(\frac{\partial \rho(r)}{\partial N} \right)_v^- = \lim_{\varepsilon \rightarrow 0^+} \frac{\rho_N(r) - \rho_{N-\varepsilon}(r)}{\varepsilon} = \rho_N(r) - \rho_{N-1}(r) \approx \rho^{HOMO}(r) \quad (2.41)$$

- Radical attack:

$$f^0(r) = \frac{f^+(r) + f^-(r)}{2} = \frac{\rho_{N+1}(r) - \rho_{N-1}(r)}{2} \approx \frac{\rho^{HOMO}(r) + \rho^{LUMO}(r)}{2} \quad (2.42)$$

“When a molecule accepts electrons, the electrons tend to go to places where $f^+(r)$ is large because it is at these locations that the molecule is most able to stabilize additional electrons. Therefore a molecule is susceptible to nucleophilic attack at sites where $f^+(r)$ is large. Similarly, a molecule is susceptible to electrophilic attack at sites where $f^-(r)$ is large, this is because these are the regions where electron removal destabilizes the molecule the least.” [35]

“The Fukui function is a real space function, which is commonly studied by means of visualization of the iso-surface. In order to achieve quantitative comparison between different sites, one can calculate the condensed Fukui function based on atomic charges. In the condensed version, the atomic population number is used to represent the amount of electron density distribution around an atom. The definition of the condensed Fukui function for an atom, say A, can be written as:” [35]

- Nucleophilic attack:

$$f_A^+(r) = q_N^A - q_{N+1}^A \quad (2.43)$$

- Electrophilic attack:

$$f_A^-(r) = q_{N-1}^A - q_N^A \quad (2.44)$$

- Radical attack:

$$f_A^0 = (q_{N-1}^A - q_{N+1}^A)/2 \quad (2.45)$$

where q^A is the atomic charge.

Furthermore, the definition of condensed Fukui for an atom (say A) is extended to a functional group (say M), which is written as: [35]

- Nucleophilic attack:

$$f_M^+(r) = \sum_A f_A^+ = \sum_A (q_N^A - q_{N+1}^A) \quad (2.46)$$

- Electrophilic attack:

$$f_M^-(r) = \sum_A f_A^- = \sum_A (q_{N-1}^A - q_N^A) \quad (2.47)$$

- Radical attack:

$$f_M^0(r) = \sum_A f_A^0 = \sum_A (q_{N-1}^A - q_{N+1}^A)/2 \quad (2.48)$$

2.5.3 Indices to evaluate charge-transfer excitations

Tangui Le Bahers [37] proposed a method for analyzing charge-transfer (CT) during electron transition. On that basis, Tian Lu made some extensions and improvements, which were implemented in Software Multiwfn.

The electron density variation between the excited state (EX) and the ground state (GS) is defined as eq.(2.49).

$$\Delta\rho(r) = \rho_{EX}(r) - \rho_{GS}(r) \quad (2.49)$$

It should be noticed that the geometries used in calculating $\rho_{EX}(r)$ and $\rho_{GS}(r)$ must be identical, otherwise the resulting $\Delta\rho(r)$ is meaningless. Therefore, it can only be used to characterize “vertical” processes. [36]

“The $\Delta\rho$ value can be divided into positive and negative parts, namely ρ_+ and ρ_- . Of course, the integral of $|\rho_+|$ and $|\rho_-|$ should be equal, and if evident inequality is observed, that means the error in the numerical integral is unneglectable, and a denser grid is required. Even though what was analyzed is single-electron excitation, the magnitude of ρ_+ and ρ_- as well as their integrals can also be theoretically larger than 1.0 as the excitation of an electron leads to a distribution reorganization of the remaining electrons, which also contribute to $\Delta\rho$.”[35] ([35] is my publication in my master.)

The transferred charge q_{CT} is the magnitude of the integral of ρ_+ and ρ_- over the whole space, in a.u.. The barycenter of positive and negative parts of $\Delta\rho$ can be computed shown in eq.(2.50), eq.(2.51).

$$R_+ = \frac{\int r\rho_+(r)dr}{\int \rho_+(r)dr} \quad (2.50)$$

$$R_- = \frac{\int r\rho_-(r)dr}{\int \rho_-(r)dr} \quad (2.51)$$

The Cartesian components of R_+ will be referred to as X_+ , Y_+ , Z_+ below, while those of R_- will be referred to as X_- , Y_- , Z_- .

The distance between the two barycenters, the D_{CT} index, is defined as eq.(2.52), which characterizes a total charge-transfer length..

$$\sqrt{(D_{CT,x})^2 + (D_{CT,y})^2 + (D_{CT,z})^2} \equiv |\mathbf{R}_+ - \mathbf{R}_-| \quad (2.52)$$

$$D_{CT,x} = |X_+ - X_-|, D_{CT,y} = |Y_+ - Y_-|, D_{CT,z} = |Z_+ - Z_-| \quad (2.53)$$

Then, a dipole moment relating the ground and excited states (μ_{CT}) can be defined as eq.(2.54) in a.u. (Debye).

$$\mu_{CT} = D_{CT} \cdot q_{CT} \quad (2.54)$$

Chapter 3

My publications

3.1 Publication I



PCCP

PAPER

View Article Online
View Journal | View IssueCite this: *Phys. Chem. Chem. Phys.*,
2019, 21, 5834

Application of an inverse-design method to optimizing porphyrins in dye-sensitized solar cells†

Chencheng Fan,^a Michael Springborg^{ab} and Yaqing Feng^{cde}

Dye-sensitized solar cells (DSSCs) have attracted much interest during the past few decades. However, it is still a tremendous challenge to identify organic molecules that give an optimal power conversion efficiency (PCE). Here, we apply our recently developed, inverse-design method for this issue with the special aim of identifying porphyrins with promisingly high PCE. It turns out that the calculations lead to the prediction of 15 new molecules with optimal performances and for which none so far has been studied. These porphyrin derivatives will in the near future be synthesized and subsequently tested experimentally. Our inverse-design approach, PooMa, is based on the strategy of providing suggestions for molecular systems with optimal properties. PooMa has been developed as a tool that requires minimal resources and, therefore, builds on various approximate methods. It uses genetic algorithm to screen thousands (or often more) of molecules. For each molecule, the density-functional tight-binding (DFTB) method is used for calculating the electronic properties. In the present work, five different electronic properties are determined, all of which are related to optical performance. Subsequently, a quantitative structure–property relationship (QSPR) model is constructed that can predict the PCE through those five electronic properties. Finally, we benchmark our results through more accurate DFT calculations that give further information on the predicted optimal molecules.

Received 18th December 2018,
Accepted 16th February 2019

DOI: 10.1039/c8cp07722c

rsc.li/pccp

1 Introduction

The identification of materials for renewable-energy applications has become one of the most imperative issues within the global-energy strategy. Compared with silicon photovoltaic cells, dye-sensitized solar cells (DSSCs) exhibit advantages including ease of fabrication, low production costs, and compatibility with flexible substrates.

The performance of DSSCs can be quantified through various photovoltaic parameters such as the overall power conversion efficiency (PCE), the photo-current density measured at short-circuit (J_{sc}), the open circuit voltage (V_{oc}), the fill factor

(FF), and the intensity of the incident light (P_{in}). Specifically, the PCE gives the overall performance of the solar cell (eqn (1)).

$$\text{PCE (\%)} = \frac{J_{sc} \cdot \text{FF} \cdot V_{oc}}{P_{in}} \times 100\% \quad (1)$$

Considering the working principles of DSSC (Fig. 1), when illuminated with light, the electrons in the ground state of the dye are excited to higher-energy excited states. Subsequently, the excited electrons are injected into the conduction band of a

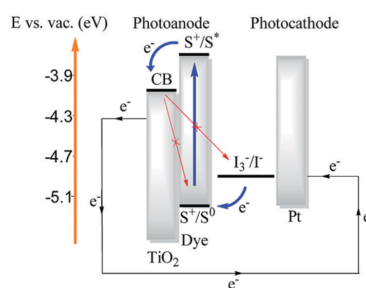


Fig. 1 The mechanism of a DSSC.²

^a Physical and Theoretical Chemistry, University of Saarland, 66123 Saarbruecken, Germany. E-mail: chencheng.fan@uni-saarland.de, m.springborg@mx.unisaarland.de

^b Materials Science, Tianjin University, 300350 Tianjin, China

^c School of Chemical Engineering and Technology, Tianjin University, 300350 Tianjin, China. E-mail: yafeng@tju.edu.cn

^d Collaborative Innovation Center of Chemical Science and Engineering (Tianjin), 300072 Tianjin, China

^e Tianjin Engineering Research Center of Functional Fine Chemicals, China

† Electronic supplementary information (ESI) available: Data set and substitutes pool. See DOI: 10.1039/c8cp07722c

semiconductor (in most cases, TiO_2) and the oxidized dye sensitizers are regenerated by taking electrons from the redox couple in the electrolyte. Therefore, as a key component in DSSCs, the dyes play an important role in determining the corresponding cell performance and should fulfill various essential characteristics:¹

1. The dye should have a good absorption of solar light.
2. The dye should have at least one anchoring group (COOH , PO_3H_2 , *etc.*) to bind strongly the dye to the semiconductor surface.
3. The energy of the LUMO of the dye should be higher than the conduction band edge of the semiconductor (which in the present work will be TiO_2) so that electrons can be transferred efficiently from the excited dye to the conduction band of TiO_2 . Moreover, the energy of the HOMO of the dye should be lower than the energy level of the electrolyte for the dye regeneration.
4. Unfavorable dye aggregation on the semiconductor surface should be avoided through optimization of the molecular structure of the dye or by addition of coadsorbents that prevent aggregation. However, controlled dye aggregates (J-aggregates) can lead to an improved performance compared with a monomer dye layer.
5. The dye should be photostable, thermally stable, and electrochemically stable.

However, there are many other factors that also influence the performance of a DSSC making it difficult to optimize the system. For example,¹ it may be beneficial for the performance of the DSSC to slow down the electron injection by increasing the energy at the conduction band edge (E_c). On the one hand, this should lead to an increase in V_{oc} , but, on the other hand, this will simultaneously lead to a decrease in the driving force for electron injection, which may result in a decrease in J_{sc} . Accordingly, the performance of a DSSC may be quantified through the product of current and voltage, which will represent a compromise between these two effects.

Another example¹ is that, generally, it seems to be favorable to have a larger spatial separation of HOMO and LUMO orbitals because this can imply an increased electron-density in those parts of the molecules that are close to the anchor group and to the TiO_2 substrate, while simultaneously decreasing the electron-density in those part of the molecules that are closest to the electrolyte. However, it should be noted that such a design will lead to a dye with a high reorganization energy, which will influence the performance negatively.

Therefore, considering the complexity of the DSSCs, it is still a big challenge to improve the solar cell efficiency so much that they can meet the standards of the photovoltaic cells market.

In 2014, Michael Grätzel and his co-workers studied the porphyrin named SM315 within the concept of DSSCs and obtained a PCE of 13%, which is the current record for DSSCs.³ Actually, among the organic-based dyes, porphyrins have always attracted much attention because they have a natural role in light harvesting, notably in the wavelength range of 400–700 nm.⁴ Moreover, different substituents can be easily attached to the core of porphyrin, which offers plenty of possibilities for molecular design. During the last decades, many novel porphyrins have been designed and applied in DSSCs. Most often, one considers porphyrins with a donor- π bridge-acceptor (D- π -A) framework, together with its variations, such as D-A- π -A, D- π -A- π -D,

A- π -D- π -A, *etc.*⁵ The D- π -A structure is beneficial for an efficient intra-molecular charge transfer (ICT) from donor to acceptor and is helpful for the electron injection into the conduction band of TiO_2 .⁵

In addition, many studies have focused on theoretical calculations with the aim of analysing the electronic structures and properties of porphyrins in DSSCs, and thereby providing a deeper understanding of the DSSCs that can lead to a more rational design of dyes. Thus, Giribabu and co-workers found that the zinc porphyrins give a better performance in comparison with the corresponding free-base porphyrins.⁶ Nazeeruddin and co-workers found that porphyrins with Zn are better than porphyrins containing Cu. Porphyrins with a phosphonate anchoring group showed lower efficiencies than those with a carboxylate anchoring group.⁷ Kim and co-workers found that charge-separation is one of the main factors that govern the solar cell efficiency.⁸ The role of *para*-alkyl substituents on *meso*-phenylporphyrin-sensitized solar cells was studied by Ballester and co-workers, who reported that the presence of hydrophobic alkyl chains in the molecular structures of porphyrins can decrease the charge recombination between the injected electrons and the electrolyte.⁹

However, it has been estimated that the maximum achievable theoretical PCE amounts to 32% for DSSCs,¹⁰ implying that the current record of 13% can be improved significantly. The common approach of arriving at an improvement is, based on personal experience and ideas, to design and synthesize novel dyes. Subsequently, these are assembled and their cell performance is checked. In parallel, theoretical calculations may be carried through with the aim of studying their electronic structures and electronic properties. This “trial and error” approach is very time-consuming, expensive, and inefficient when aiming at finding breakthrough dyes for DSSCs. There is an increasing trend towards using computational methods to rationalize the design of new dyes,^{11,12} in particular using quantitative structure–property relationship (QSPR) models. Some of those approaches are based on machine learning or data mining tools. Our study is related to that of Venkatraman *et al.*¹¹ who used genetic algorithm to optimize coumarin dyes for dye-sensitized solar cells. However, many details of our computational approach as well in the class of molecular systems and properties that we are considering differ from those of Venkatraman *et al.*

In the present work, we take known findings about DSSCs into consideration and adapt our inverse design method “PooMa”^{13,14} to the DSSCs with the purpose of identifying—eventually new—porphyrins with promisingly large values for the PCE. PooMa combines a stoichiometry-optimization approach based on genetic algorithms, an automatic approach for constructing molecules, an inexpensive but sufficiently accurate approach for calculating the relevant electronic properties of DSSCs, and a mathematical formulation of the performance function whose value shall be optimized. In the present study, the PooMa calculations lead to the prediction of 15 molecules with optimal values for the PCE, all of which has not been studied earlier. Subsequently, these modified porphyrins are

PCCP

View Article Online

Paper

studied through standard DFT and TD-DFT methods. Moreover, it is planned that they will be synthesized in the near future.

It is clearly a strong simplification to attempt to describe the performance of the DSSCs using solely properties of the individual, isolated dyes. A more accurate description could have been obtained by including, for instance, the interfacial properties of dyes adsorbed on a TiO₂ film (such as adsorption configurations, adsorption energy, charge transfer at the TiO₂-dye interface),^{15,16} the Fermi energy of TiO₂,¹⁷ charge recombination at TiO₂-electrolyte interface¹⁸ *etc.* However, the rationale behind PooMa is not to provide exact results and to identify uniquely the best possible systems, but rather to provide interesting suggestions for useful systems by exploring a larger part of chemical space than what usually is done. Moreover, PooMa is based on being computationally efficient which would be partly in conflict with attempts to try to include larger parts of the DSSCs in the modelling. With these limitations in mind, we are convinced that PooMa is an interesting alternative to the try-and-error approaches often used as well as to the computationally much more costly big-data approaches.

2 PooMa: the inverse design approach

2.1 The genetic algorithm and construction of molecules

The main idea of our approach is to identify functional groups and their positions at a given backbone. Since α substituted zinc-porphyrins are known to show good performance,¹⁹ we use the zinc-porphyrin as the backbone and the four α sites as the sites for functional groups (see Fig. 2). In view of the feasibility of synthesis, H atoms are kept at the eight β positions.

Currently in the pool of possible functional groups, we have collected 41 functional groups (see Fig. S1 in the ESI†). All of these are extracted from published dye sensitizers. Moreover, the groups 1–17 are “anchor types” containing the carboxyl group COOH, while groups 18–41 are “donor types”. Since the porphyrins should have at least one anchor group for the transfer of electrons from the dye to the electrode, we require that the α 1 site has to be occupied by an anchor group. For the other sites, any functional group can be attached.

The core of our inverse design approach is a genetic algorithm (GA) that is constructed to optimize the solar cell performance among thousands of molecules with different functional groups. More details can be found in our earlier work.^{13,14} Therefore, here we give just a short description, referring to Fig. 3. Our genetic algorithm consists of the following steps:

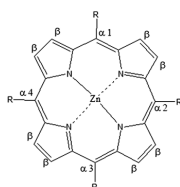


Fig. 2 Zinc-porphyrin as backbone and the four α sites as available sites for functional groups.

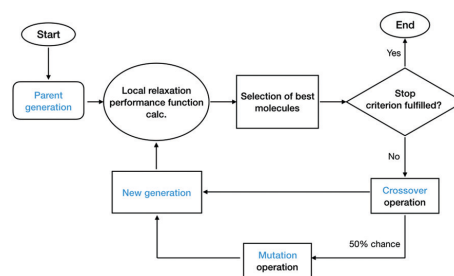


Fig. 3 Flowchart of the genetic algorithm.

1. Create randomly 20 porphyrins, which differ in the arrangement of the functional groups and which all are different also after applying symmetry operations.
2. Relax their structures locally and calculate their values for the performance function that shall be optimized.
3. Separate the molecules (“Parent molecules”) randomly into 10 pairs.
4. For each pair, interchange a (randomly generated) part of the functional groups through a crossover operation to create new “Children molecules”. The children may experience mutations whereby some functional group on one site is replaced randomly with another functional group, except on α 1 site at which the replacement can only be through an anchor type. It is verified that all the children and parent molecules are different also after applying symmetry operations.
5. Relax the structures of the children molecules and calculate their values for the performance function.
6. Out of the 20 “Parent molecules” and 20 “Children molecules”, choose those 20 molecules with the best performance function as parents for the next generation.
7. If the 20 best molecules are unchanged for 10 generations, stop the calculation, otherwise return to step 3.

2.2 DFTB method

In our recently developed PooMa (POOr MAn’s MAterials optimization) approach we have made a compromise between accuracy and computational speed. Thus, calculations based on density functional theory (DFT) can give comparatively accurate results at a relatively expensive cost but since our approach is based on studying the electronic structure of maybe even tens of thousands of molecules, DFT-based methods become too expensive for our aim.

Instead, we use the non-self-consistent-charge density-functional tight-binding (non-SCC DFTB) method by Seifert and co-workers.²⁰ Thereby, the computational efforts can be reduced drastically while simultaneously maintaining a reasonable accuracy. With this method, the total energy relative to that of the non-interacting atoms is given as

$$E_{\text{tot}} \simeq \sum_i^{\text{occ}} \varepsilon_i - \sum_j \sum_m^{\text{occ}} \varepsilon_{jm} + \frac{1}{2} \sum_{j \neq k} U_{jk} (\vec{R}_j - \vec{R}_k) \quad (2)$$

where ε_i is the energy of the i th occupied orbital for the system of interest and ε_{jm} is that of the m th occupied orbital for the

isolated j th atom. U_{jk} is a short-ranged pair potential between atoms j and k . Non-SCC implies that we do not take charge fluctuations and the resulting interactions into account. The U_{jk} is so adjusted that experimental data or results of *ab initio* calculations for certain test systems are accurately reproduced. Valence electrons are treated explicitly in the DFTB formalism, whereas the effects of the core electrons are included within a frozen-core approximation.

The orbital energies ε_i are obtained from the single-particle Kohn–Sham equations

$$\hat{h}_{\text{eff}}\psi_i(\vec{r}) = \left[-\frac{\hbar^2}{2m}\nabla^2 + V(\vec{r}) \right] \psi_i(\vec{r}) = \varepsilon_i\psi_i(\vec{r}) \quad (3)$$

whereby the potential $V(\vec{r})$ is approximated as the superposition of potentials of the isolated atoms,

$$V(\vec{r}) = \sum_m V_m(|\vec{r} - \vec{R}_m|) \quad (4)$$

with \vec{R}_m being the position of the m th nucleus. Moreover, $\psi_i(\vec{r})$ is approximated through a linear combination of atom-centered functions,

$$\psi_i(\vec{r}) = \sum_{\vec{R}} \sum_j c_{i,\vec{R},j} \chi_{\vec{R},j}(\vec{r}) \quad (5)$$

with $\chi_{\vec{R},j}(\vec{r})$ being the j th function centered at the atom at position \vec{R} . Within the DFTB approach it is furthermore assumed that all matrix elements $\langle \chi_{\vec{R}_1,j_1} | \hat{h}_{\text{eff}} | \chi_{\vec{R}_2,j_2} \rangle$ can be obtained from calculations on diatomic molecules.

2.3 Performance function and property descriptors

Since the PCE quantifies the overall performance of the solar cell, we shall let PCE be the performance function whose value shall be optimized when identifying the best molecules. Even though the factors that affect the PCE are quite complicated and difficult to access with electronic-properties calculations, the electronic properties of a sensitizer are very important for the overall performance of a DSSC. Thus, we choose the followings five electronic property descriptors whose values subsequently shall be used to predict the PCE:

1. The energy gap G between the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO).

$$G = \varepsilon_{\text{LUMO}} - \varepsilon_{\text{HOMO}} \quad (6)$$

2. We define the sun absorption per valence electron A through eqn (7), whereby N_e is the number of valence electrons of the dye, i denotes an occupied orbital and a denotes an unoccupied orbital. f_{ia} is the oscillator strength for single electron excitation from the i th orbital to the a th orbital. The spectral function $P(\varepsilon_a - \varepsilon_i)$ for the solar spectrum is calculated using a black-body approximation as given in eqn (8).

$$A = \frac{\sum_{i,a} f_{ia} \cdot P(\varepsilon_a - \varepsilon_i)}{N_e} \quad (7)$$

$$P(\varepsilon_a - \varepsilon_i) = \frac{8\pi(\varepsilon_a - \varepsilon_i)^3}{(2\pi\hbar c)^3} \left[\exp\left(\frac{\varepsilon_a - \varepsilon_i}{k_B T}\right) - 1 \right]^{-1} \quad (8)$$

The oscillator strength f_{ia} is calculated using the approximated TD-DFTB (time-dependent density-functional tight-binding) method.²¹ Since we only consider one-electron excitations and do not take charge fluctuations into account, the expression for f_{ia} can be simplified as given below in eqn (9), whereby the transition dipole moment \vec{d}_{ia} is approximated in terms of Mulliken transition charges $q_{n,ia}$ and the corresponding coordinates \vec{R}_n of the atom n for transition $i \rightarrow a$.

$$f_{ia} = \frac{4}{3}(\varepsilon_a - \varepsilon_i) \left| \vec{d}_{ia} \right|^2 = \frac{4}{3}(\varepsilon_a - \varepsilon_i) \left| \sum_n (\vec{R}_n \cdot q_{n,ia}) \right|^2 \quad (9)$$

3. Light-harvesting efficiency per valence electron, LHE. All electronic transitions are taken into account for excitations in the UV-Vis range.

$$\text{LHE} = \frac{\sum_{i,a} (1 - 10^{-f_{ia}})}{N_e} \quad (10)$$

4. Average orbital overlap per valence electron O_{av} (eqn (11)), that takes into account all single overlaps for each pair of occupied and unoccupied molecular orbitals with a weight factor according to its transition probability $f_{ia}P(\varepsilon_a - \varepsilon_i)$. The individual overlap O_{ia} (eqn (12)) is calculated using Mulliken gross populations $q_{n,i}$ and $q_{n,a}$.

$$O_{\text{av}} = \frac{\sum_{i,a} f_{ia} P(\varepsilon_a - \varepsilon_i) O_{ia}}{N_e} \quad (11)$$

$$O_{ia} = \sum_n |q_{n,i} \cdot q_{n,a}| \quad (12)$$

5. Average orbital distance per valence electron D_{av} (eqn (13)), similar to O_{av} , that takes into account all the single-orbital distances with a weight $f_{ia}P(\varepsilon_a - \varepsilon_i)$. Also here, the single orbital distance D_{ia} (eqn (14)) is calculated using Mulliken gross populations.

$$D_{\text{av}} = \frac{\sum_{i,a} f_{ia} P(\varepsilon_a - \varepsilon_i) D_{ia}}{N_e} \quad (13)$$

$$D_{ia} = \left| \sum_n q_{n,i} \vec{R}_n - \sum_n q_{n,a} \vec{R}_n \right| \quad (14)$$

In several cases we are studying properties per valence electron. The reason is that our goal is to optimize intensive properties and not extensive properties in order to avoid that the largest molecules are found to be the best ones.

2.4 DFT calculations

Additional DFT and TD-DFT calculations are used to have a further theoretical understanding of the porphyrin molecules that result from the PooMa calculations.

The ground state geometry optimization of all molecules are carried through using the B3LYP functional with a 6-311G(d,p)

basis set. Keeping the geometries that are obtained through those calculations, we studied the excited states using the CAM-B3LYP functional with a 6-311G(d,p) basis set. The conductor-like polarizable continuum model (C-PCM) is used to simulate the environment of tetrahydrofuran (THF). All DFT and TD-DFT calculations are performed using the Gaussian09 software package.²² Subsequently, the Fukui functions, condensed Fukui functions, and charge transfer indices are calculated using the Multiwfn software package.²³

2.5 Fukui function

The Fukui function is important within conceptual density functional theory and is defined as²³

$$f(\vec{r}) = \left(\frac{\partial \rho(\vec{r})}{\partial N} \right)_V \quad (15)$$

where N is the number of electrons in the system of interest and V is the external potential. Using a finite-difference approximation, the Fukui function can be calculated unambiguously for three situations:

Nucleophilic attack:

$$f^+(\vec{r}) = \left(\frac{\partial \rho(\vec{r})}{\partial N} \right)_V^+ = \lim_{\varepsilon \rightarrow 0^+} \frac{\rho_{N+\varepsilon}(\vec{r}) - \rho_N(\vec{r})}{\varepsilon} \simeq \rho_{N+1}(\vec{r}) - \rho_N(\vec{r}) \quad (16)$$

Electrophilic attack:

$$f^-(\vec{r}) = \left(\frac{\partial \rho(\vec{r})}{\partial N} \right)_V^- = \lim_{\varepsilon \rightarrow 0^+} \frac{\rho_N(\vec{r}) - \rho_{N-\varepsilon}(\vec{r})}{\varepsilon} \simeq \rho_N(\vec{r}) - \rho_{N-1}(\vec{r}) \quad (17)$$

Radical attack:

$$f^0(\vec{r}) = \frac{f^+(\vec{r}) + f^-(\vec{r})}{2} \simeq \frac{\rho_{N+1}(\vec{r}) - \rho_{N-1}(\vec{r})}{2} \quad (18)$$

When a molecule accepts electrons, the additional electrons tend to go to those parts of space where $f^+(\vec{r})$ is large because this is where the molecule is most stable against the addition of electrons. Therefore, a molecule is susceptible to nucleophilic attack at places where $f^+(\vec{r})$ is large. Similarly, a molecule is susceptible to electrophilic attack at sites where $f^-(\vec{r})$ is large, because these are the regions where an electron removal destabilizes the molecule the least.

Furthermore, the “condensed” version of the Fukui function based on atomic charges can be helpful in providing a more quantitative description. For an atom A, the condensed Fukui function is defined as²³

Nucleophilic attack:

$$f_A^+ = q_N^A - q_{N+1}^A \quad (19)$$

Electrophilic attack:

$$f_A^- = q_{N-1}^A - q_N^A \quad (20)$$

Radical attack:

$$f_A^0 = (q_{N-1}^A - q_{N+1}^A)/2 \quad (21)$$

Here, q^A is the atomic charge according to a Hirshfeld population analysis.

2.6 Charge transfer indices

Le Bahers²⁴ proposed a method for analyzing the charge-transfer (CT) during an electronic transition. The electron density variation between the excited state (EX) and the ground state (GS) is defined as

$$\Delta\rho(r) = \rho_{\text{EX}}(r) - \rho_{\text{GS}}(r) \quad (22)$$

The geometry of the excited state and the ground state must be identical. $\Delta\rho(r)$ can be separated into positive and negative parts, namely $\rho_+(r)$ and $\rho_-(r)$. The transferred charge q_{CT} is the magnitude of the integral of $\rho_+(r)$ and $\rho_-(r)$ over the whole space. The barycenters of the positive and negative parts can be defined as

$$R_+ = \int r \rho_+(r) dr / \int \rho_+(r) dr \quad (23)$$

$$R_- = \int r \rho_-(r) dr / \int \rho_-(r) dr \quad (24)$$

The spatial distance between the two barycenters of density distributions can subsequently be used to quantify the CT excitation length D_{CT} ,

$$D_{\text{CT}} = |R_+ - R_-|. \quad (25)$$

From this, a dipole moment relating the ground and the excited states (μ_{CT}) can be defined as

$$\mu_{\text{CT}} = D_{\text{CT}} q_{\text{CT}} \quad (26)$$

3 The QSPR model

Quantitative structure–property relationship (QSPR) models aim to establish a mathematical relation between one set of descriptors characterizing the system and a property of interest. Here, we will construct a QSPR model that uses the five electronic properties as descriptors to predict the PCE.

3.1 Data set

In total, 206 dye-sensitizers including porphyrins and some small molecules, compiled from the literature, were chosen as forming the data set (see Table S1 in ESI†) from which the QSPR model should be derived. Ideally all values should result from a single source, for example from the same laboratory, but such a data set would be very limited in size. To obtain data that are reasonably comparable despite being from different sources, we have chosen data that satisfy:

1. The electrolyte is iodine based.
2. The semiconductor is TiO₂.
3. No polymer sensitizer.
4. Only one dye sensitizer is used without co-adsorbent or co-sensitizer.
5. The global standard AM 1.5G light source with a light intensity of 100 mW cm⁻² is used.

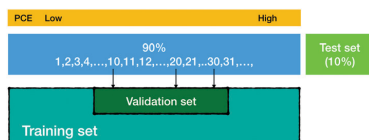


Fig. 4 The whole data set is split into training set, validation set, and test set.

Although other experimental parameters (like semiconductor thickness, cell fabrication, the amount of dyes absorbed on TiO_2 , exposure time) can influence the performance, further constraints based on these factors would reduce the size of the data set too much and were therefore ignored.

In order to obtain and validate a reliable QSPR model, we split the whole data set into a training set, a validation set, and a test set (Fig. 4). Specifically, at first we randomly select 10% of the data as constituting the test set. The remaining data is sorted according to ascending order of the PCE. Every 10th data (such as the 10th, 20th, 30th, *etc.*) are selected as forming the validation set while the other ones constitute the training set. We use the training set in constructing the QSPR model. Subsequently, we examine its performance using the validation set. Finally, we examine the model with further randomly constructed test sets.

3.2 Correlation coefficients

The correlation coefficient between two variables X and Y is defined as given in eqn (27). Here, cov is the covariance, σ_X and σ_Y are the standard deviations, E is the expectation value, and μ_X and μ_Y are the mean values of X and Y .

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} \quad (27)$$

With the help of the correlation coefficient only linear relationships between the two variables X and Y can be identified. Moreover, $-1 \leq \text{corr}(X, Y) \leq 1$. When $\text{corr}(X, Y)$ approaches 0, X and Y possess no linear relationship. As $\text{corr}(X, Y)$ approaches ± 1 , the variables possess an increased linear relationship. The correlation coefficient is symmetric, $\text{corr}(X, Y) = \text{corr}(Y, X)$.

According to Cohen's standard we have that, if the correlation coefficient lies between the 0.1 and 0.29, the two variables have a small linear dependence; if the correlation coefficient is between 0.3 and 0.49, the two variables have a medium linear dependence; and if the correlation coefficient is larger than 0.5, the two variables have a large linear dependence. We shall use these criteria in evaluating the interdependence of our variables.

3.3 Model evaluation

The quality of a model can be quantified through the coefficient R^2 given below in eqn (28). Here, N is the number of data in the training set, $\text{PCE}_i^{\text{obs}}$ is the observed/experimental PCE value for the i th sample, and $\text{PCE}_i^{\text{cal}}$ is the calculated PCE value for

this system. $\text{PCE}_{\text{mean}}^{\text{obs}}$ is the arithmetic mean of the observed PCE values in the training set.

$$R^2 = 1 - \frac{\sum_{i=1}^N (\text{PCE}_i^{\text{obs}} - \text{PCE}_i^{\text{cal}})^2}{\sum_{i=1}^N (\text{PCE}_i^{\text{obs}} - \text{PCE}_{\text{mean}}^{\text{obs}})^2} \quad (28)$$

For the ideal model, R^2 equals 1. As the value of R^2 deviates from 1, the quality of the model is reduced. Then, one possible approach for improving the model could be to increase the number of descriptors with, however, the accompanying risk that the QSPR model becomes overdetermined.

The predictive power of the model is evaluated by the predictive squared correlation coefficient (Q^2), the mean absolute error (MAE), and the root of mean square error (RMSE) of the validation set and test set. Q^2 is given in eqn (29), where N is the number of samples in the validation set/test set, $\text{PCE}_i^{\text{pred}}$ is the predicted PCE value of the i th sample, and $\text{PCE}_i^{\text{obs}}$ is the observed/experimental value. $\text{PCE}_{\text{mean}}^{\text{train}}$ is the arithmetic mean of the observed PCE values in the training set.

$$Q^2 = 1 - \frac{\sum_{i=1}^N (\text{PCE}_i^{\text{pred}} - \text{PCE}_i^{\text{obs}})^2}{\sum_{i=1}^N (\text{PCE}_i^{\text{obs}} - \text{PCE}_{\text{mean}}^{\text{train}})^2} \quad (29)$$

The MAE and RMSE are given in eqn (30) and (31), where N is the number of samples in the training/validation/test set.

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |\text{PCE}_i^{\text{pred}} - \text{PCE}_i^{\text{obs}}| \quad (30)$$

$$\text{RMSE} = \sqrt{\frac{1}{N} \left[\sum_{i=1}^N (\text{PCE}_i^{\text{pred}} - \text{PCE}_i^{\text{obs}})^2 \right]} \quad (31)$$

4 Results and discussion

4.1 Analysis of the descriptors

At first we verified that the descriptors take different values for different systems in order to assure that they indeed are able to distinguish between different samples. Table 1 shows that all the five descriptors have standard deviations that are comparable with the mean values, *i.e.*, for each descriptor the data are sufficiently well scattered.

Then we examined whether the descriptors are linearly dependent through their correlation coefficient. Table 2 shows that all the five descriptors are independent since the largest correlation coefficient equals 0.74 (between O_{av} and A).

Table 1 Mean values and standard deviations of the electronic properties G , O_{av} , D_{av} , A , and LHE

Variables	G	O_{av}	D_{av}	A	LHE
μ	0.054	0.085	16.7	2.84	0.089
σ	0.013	0.031	10.4	0.75	0.020

View Article Online

Paper

PCCP

Table 2 The correlation coefficients between G , O_{av} , D_{av} , A , LHE, and PCE

Variables	G	O_{av}	D_{av}	A	LHE
O_{av}	0.47				
D_{av}	-0.43	-0.41			
A	0.23	0.74	0.10		
LHE	-0.11	0.34	0.37	0.7	
PCE	-0.21	-0.24	-0.08	-0.37	-0.29

Therefore, the five electronic properties G , O_{av} , D_{av} , A , LHE can be used as independent variables in the QSPR model.

In addition, from the results for the correlation coefficients between any of the electronic properties and PCE (see Table 2) we conclude that a QSPR model can only contain non-linear relationships: according to Cohen's standard, only A shows a medium linear relationship with PCE while the rest electronic properties show only small or no linear relationship.

4.2 The QSPR model

We adopted a non-linear mathematical function for our QSPR model (eqn (32)), where the PCE for the i th system is denoted Z_i and the value for the j th of the five descriptors is denoted X_{ij} . As can be seen, the expression consists of a constant plus terms that depend on one, two, three, four, or five descriptors.

$$\begin{aligned}
 Z_i \approx & c_0 + \sum_j c_j X_{ij}^{p_j} + \sum_{jk} c_{jk} X_{ij}^{p_j} X_{ik}^{p_k} + \sum_{jkl} c_{jkl} X_{ij}^{p_j} X_{ik}^{p_k} X_{il}^{p_l} \\
 & + \sum_{jklm} c_{jklm} X_{ij}^{p_j} X_{ik}^{p_k} X_{il}^{p_l} X_{im}^{p_m} \\
 & + c_{jklmn} X_{ij}^{p_j} X_{ik}^{p_k} X_{il}^{p_l} X_{im}^{p_m} X_{in}^{p_n}
 \end{aligned}
 \quad (32)$$

whereby in each summation we include only terms involving different descriptors.

The fitting will result in the coefficients c and the powers p . In order to reduce the effects of numerical noise and to make the QSPR model more stable, terms in eqn (32) involving powers p smaller than 0.3 were removed. In addition, if it subsequently was found that two terms were strongly dependent (a value for corr larger than 0.9) one of the terms was deleted, too. Finally, if any term was found to be irrelevant for PCE (*i.e.*, corr is smaller than 0.1) this term was also deleted.

At the end, the final expression of our QSPR model becomes (in Hartree atomic units):

$$\begin{aligned}
 \eta = & 6.01 + 6.54G^{-0.44} - 1.81O^{-0.51} + 0.05D^{0.46} - 1.47A^{1.15} \\
 & - 4.14LHE^{-0.54} - 7.24G^{-0.63}A^{-0.57} \\
 & + 7.95G^{-0.36}A^{-0.55}LHE^{-0.32} \\
 & - 0.43G^{-0.65}O^{0.41}A^{-0.79}LHE^{-0.5}.
 \end{aligned}
 \quad (33)$$

As shown in Fig. 5, the resulting predicted PCE values show the same trends as the experimental PCE values among the

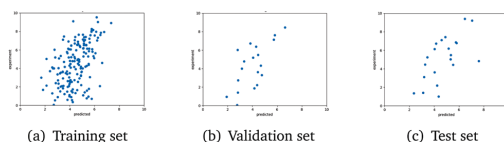


Fig. 5 Predicted and experimental PCE values for training set, validation set, and test set.

Table 3 Model validation: R^2 , Q^2 , MAE, and RMSE for training set, validation set and test set

	R^2	Q^2	MAE	RMSE
Training set	0.23	—	1.63	1.98
Validation set	—	0.42	1.58	1.78
Test set	—	0.42	1.61	1.87

training set, the validation set, and the test set. Table 3 shows however also that the model is not perfect. We emphasize that a perfect model should not be expected because of various reasons. At first, all the comments at the end of Section 1 apply, *i.e.*, reducing the performance of a DSSC to the properties of the single dye is a tremendous simplification when taking into account the complexity of the DSSCs. Moreover, there are significant experimental uncertainties (occasionally, even the same system studied experimentally by the same group at different times has resulted in different values for the PCE). Thus, for our purpose, we are convinced that the quality of our QSPR model is reasonable. Thus, when our QSPR is used in PooMa as performance function to identify good porphyrins, the absolute error between predicted PCE and experimental PCE is less important.

We add that when repeating the procedure of constructing the QSPR model through a different, randomly constructed training set, the mathematical expression changes only little, showing that our approach is numerically robust. Furthermore, when subsequently applying these slightly different mathematical functions in PooMa essentially the same porphyrins are found as being those performing the best.

4.3 DFT results

The PooMa calculations resulted in the 15 porphyrins that are shown in Fig. 6. These porphyrins are novel and have not been studied earlier. All the porphyrins share the same strong donor (iminostilbene). Moreover, some functional groups show up frequently. In addition, at least one of the COOH groups is used as anchor group, while in some cases additional groups that could also act as anchor groups occur although not necessarily acting as anchor groups.

Fig. 7 shows that the energy level alignments of all resulting porphyrins match the requirements for DSSCs, *i.e.*, the energy of HOMO is lower than the potential of the electrolyte and the energy of LUMO is higher than the conduction band of TiO_2 .

Fig. 8 shows the Fukui function. We remind that the electrons tend to flow from those regions where f^- is large to those regions where f^+ is large. For a more quantitative analysis

Paper

View Article Online

PCCP

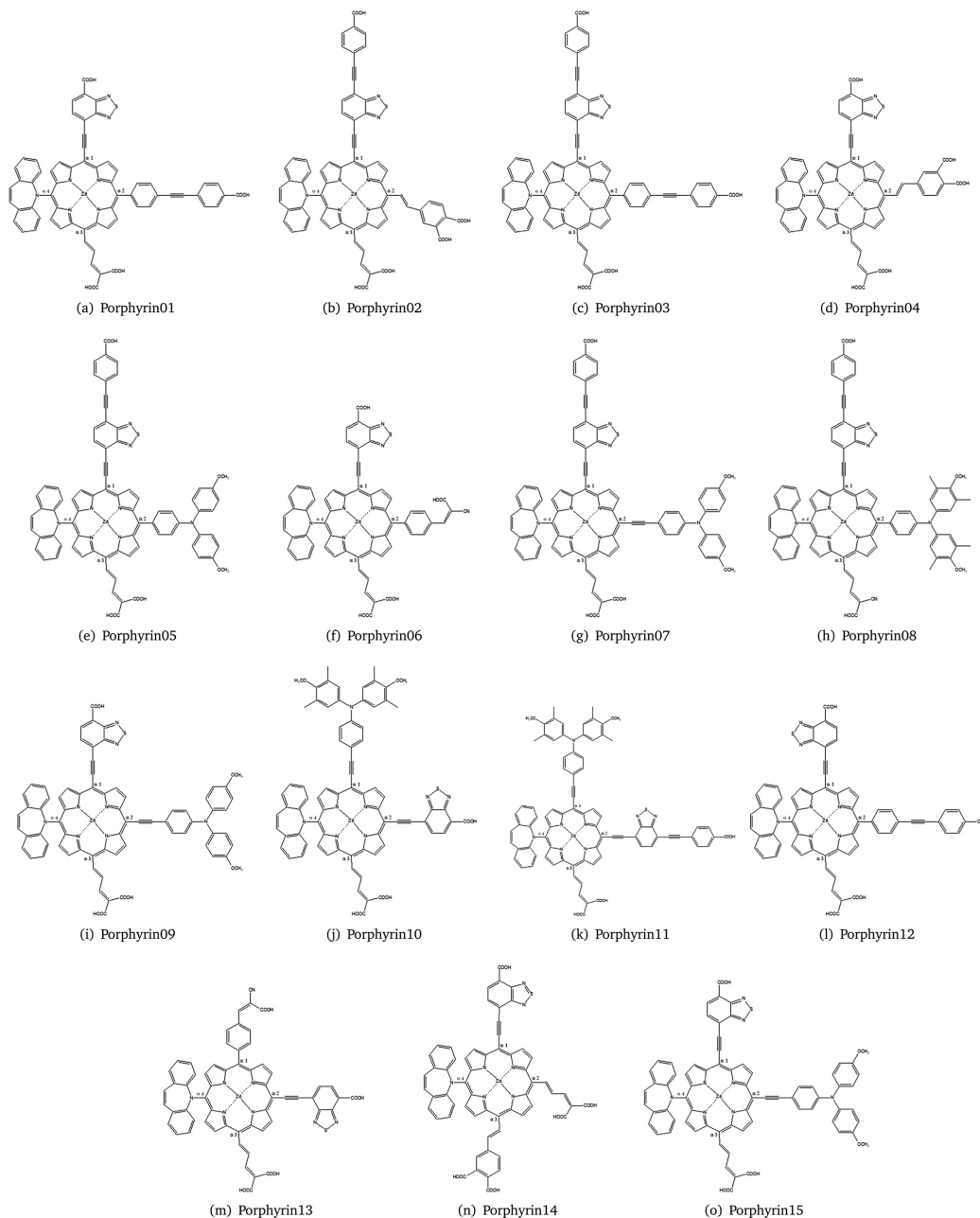


Fig. 6 The molecular structures identified by the PooMa calculations.

we show the condensed Fukui function in Table 4. Thereby, we recognize two charge-flow patterns, *i.e.*, a flow from the porphyrin ring to a functional group, a pattern that can be recognized for

Porphyrin01–04, 06, 12–14, whereas the other porphyrins show a pattern with a flow from one functional group to the porphyrin ring and to another functional group.

View Article Online

Paper

PCCP

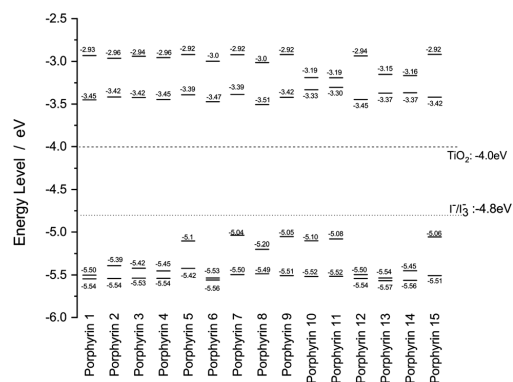


Fig. 7 The energies of the four frontier molecular orbitals (HOMO–1, HOMO, LUMO, LUMO+1) of the 15 substituted porphyrins as obtained in the DFT calculations.

Table 4 Condensed Fukui function of Porphyrin01–15

	Core		$\alpha 1$		$\alpha 2$		$\alpha 3$		$\alpha 4$	
	f^-	f^+	f^-	f^+	f^-	f^+	f^-	f^+	f^-	f^+
Porphyrin01	0.52	0.41	0.20	0.36	0.07	0.03	0.14	0.17	0.06	0.03
Porphyrin02	0.47	0.38	0.22	0.40	0.10	0.03	0.13	0.15	0.06	0.03
Porphyrin03	0.49	0.40	0.25	0.39	0.06	0.03	0.14	0.17	0.06	0.03
Porphyrin04	0.49	0.41	0.18	0.36	0.12	0.03	0.13	0.17	0.07	0.03
Porphyrin05	0.15	0.38	0.06	0.40	0.71	0.03	0.04	0.17	0.02	0.03
Porphyrin06	0.52	0.41	0.20	0.35	0.06	0.03	0.15	0.17	0.07	0.03
Porphyrin07	0.25	0.37	0.09	0.40	0.55	0.04	0.06	0.16	0.03	0.02
Porphyrin08	0.16	0.39	0.07	0.32	0.69	0.03	0.06	0.23	0.02	0.03
Porphyrin09	0.23	0.39	0.07	0.36	0.61	0.04	0.05	0.17	0.03	0.03
Porphyrin10	0.25	0.36	0.58	0.04	0.08	0.49	0.07	0.08	0.03	0.04
Porphyrin11	0.27	0.32	0.52	0.04	0.10	0.53	0.08	0.08	0.03	0.03
Porphyrin12	0.50	0.41	0.19	0.36	0.06	0.03	0.14	0.17	0.11	0.03
Porphyrin13	0.52	0.38	0.05	0.02	0.20	0.47	0.14	0.08	0.09	0.04
Porphyrin14	0.49	0.37	0.19	0.45	0.13	0.08	0.11	0.08	0.08	0.03
Porphyrin15	0.24	0.39	0.07	0.36	0.60	0.04	0.06	0.17	0.03	0.03

Table 6 shows the results of the TD-DFT calculations. For all 15 porphyrins, the maximum absorption is in the 380–500 nm range, which corresponds to the so-called Soret bands. All the

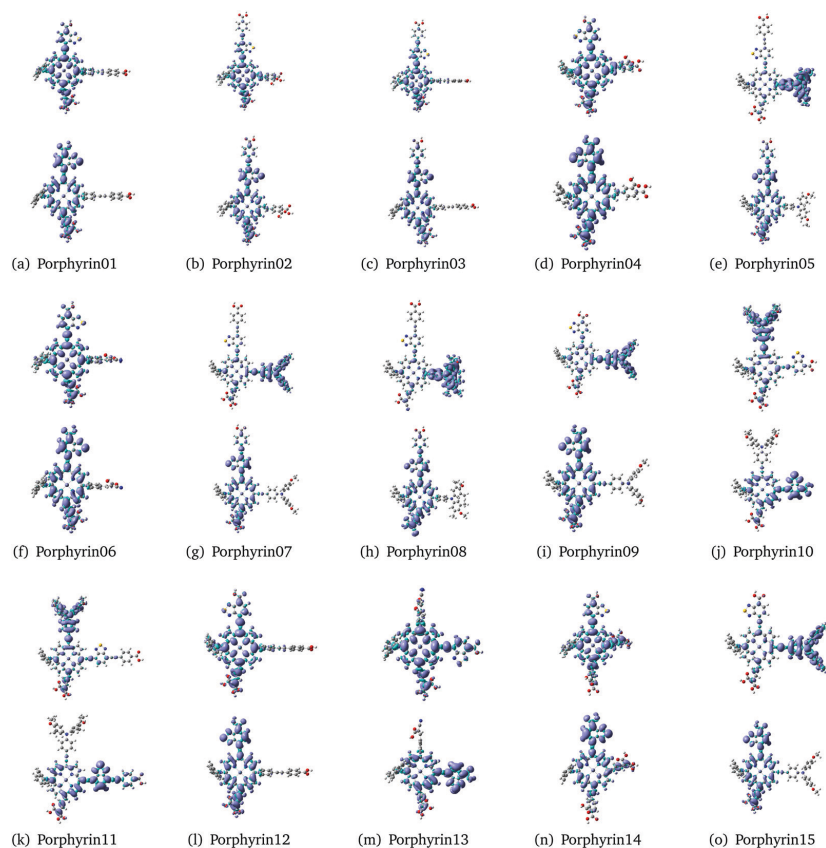


Fig. 8 The Fukui functions for the 15 substituted porphyrins. For each molecule, the upper part shows f^- and the lower part shows f^+ .

Paper

Table 5 The charge transfer indices^a q_{CT} , D_{CT} , μ_{CT} of Porphyrin01–15

	$q_{CT}/a.u.$	$D_{CT}/\text{Angstrom}$	μ_{CT}/Debye
Porphyrin01	0.61	0.05	0.16
Porphyrin02	0.64	0.06	0.18
Porphyrin03	0.67	0.05	0.16
Porphyrin04	0.59	0.06	0.17
Porphyrin05	0.72	0.18	0.61
Porphyrin06	0.58	0.09	0.25
Porphyrin07	0.75	0.45	1.59
Porphyrin08	0.76	0.08	0.30
Porphyrin09	0.68	0.55	1.82
Porphyrin10	0.72	0.61	2.11
Porphyrin11	0.79	0.58	2.21
Porphyrin12	0.61	0.09	0.27
Porphyrin13	0.58	0.19	0.54
Porphyrin14	0.59	0.24	0.68
Porphyrin15	0.61	0.05	0.16

^a The excited state is the first singlet excited state.Table 6 Oscillator strength, LHE^a and excitation contribution (only above 10%) at the calculated absorption maximum using the TD-DFT method

Molecules	$\lambda_{\text{max}}^{\text{abs}}$ (nm)	Oscillator strength	LHE	Main configuration
Porphyrin01	426	1.65	0.977	H → L+1 (57%) H-1 → L (36%)
Porphyrin02	475	1.84	0.986	H-1 → L+1 (34%) H → L+2 (25%) H-3 → L (13%) H → L (11%)
Porphyrin03	470	1.70	0.980	H-1 → L+1 (32%) H → L+2 (33%) H-3 → L (13%)
Porphyrin04	443	1.66	0.978	H → L+1 (47%) H-1 → L (40%)
Porphyrin05	472	1.53	0.970	H → L+1 (23%) H-2 → L+2 (20%) H-4 → L (14%)
Porphyrin06	452	1.69	0.979	H-2 → L+1 (10%) H-1 → L+1 (47%) H → L+2 (21%) H → L (15%)
Porphyrin07	469	2.31	0.995	H-2 → L (40%) H → L+1 (35%)
Porphyrin08	469	1.52	0.970	H → L+1 (30%) H-2 → L+2 (20%) H-4 → L (16%)
Porphyrin09	474	2.07	0.992	H-2 → L (36%) H → L+1 (34%)
Porphyrin10	471	2.21	0.994	H-2 → L (48%) H → L+1 (22%) H-2 → L+1 (10%)
Porphyrin11	470	2.14	0.993	H-2 → L (41%) H → L+1 (17%) H-2 → L+1 (14%) H-2 → L+2 (12%)
Porphyrin12	451	1.58	0.973	H-1 → L+1 (41%) H → L+2 (19%) H → L (14%)
Porphyrin13	444	2.16	0.993	H-1 → L (48%) H → L+1 (41%)
Porphyrin14	475	1.59	0.974	H-1 → L+1 (52%) H → L (25%) H → L+2 (12%)
Porphyrin15	474	2.10	0.992	H-2 → L (36%) H → L+1 (34%)

^a LHE corresponding to maximum oscillator strength. H represents HOMO, while L represents LUMO.

15 porphyrins show comparatively high oscillator strengths and light-harvest efficiencies (LHE). Moreover, the main contribution to the absorption peak is electronic transitions from the five energetically highest, occupied orbitals to the three energetically lowest, unoccupied ones.

Following Le Bahers' approach, we show in Table 5 the calculated charge transfer indices which only considered the first excited singlet state. In particular, Porphyrin11, Porphyrin10, Porphyrin9, and Porphyrin07 have large values for the intramolecular charge transfer for the first excited singlet state. However, other excited states can be critical too in intramolecular charge transfer, which will not be discussed further here.

5 Conclusions

The main purpose of the present study was to apply our recently developed inverse-design method (PooMa) to a practical application, *i.e.* identifying efficient α substituted porphyrin-dyes to be used in dye-sensitized solar cells. A second, equally important, purpose was to provide suggestions for new molecules with hopefully better performance than the so-far synthesized ones.

PooMa is based on genetic algorithms that make it possible to identify efficiently the optimal molecules out of eventually thousands or more of molecules. In order to reduce the computational demands we used the DFTB method for the electronic-structure calculations, which provides a good compromise between efficiency and accuracy. For each molecule, five property descriptors (G , O_{av} , D_{av} , A , and LHE) were calculated. We developed a QSPR model that could establish a mathematical relationship between these five property descriptors and the PCE.

It is interesting that the PooMa calculations led to 15 porphyrins of which none has been studied so far. Thus, this demonstrates that an approach like PooMa may make it possible to extend the chemical space and provide suggestions for interesting systems beyond what the more traditional use of chemical intuition provides. Moreover, we used DFT methods to provide further information on the porphyrins that PooMa identified. Finally, for the sake of completeness we add that the suggested porphyrins will be synthesized and tested in experiment in the near future.

The main ideas behind genetic algorithms are not new. However, the combination with a simple, fast, but still accurate method (DFTB) for the calculation of electronic properties is a unique feature of our method that makes it generally applicable also for other properties and/or classes of systems. Thus, our inverse design method is not limited to solar cells or to optical properties. New electronic properties can be added in PooMa and new performance functions can be optimized. Other areas, such as heterogeneous catalysis, hydrogen storage, *etc.*, provide other possible applications of PooMa.

Conflicts of interest

The authors declare no conflict of interest.

View Article Online

PCCP

Paper

Acknowledgements

This work was supported by the German Research Council through Project Sp439/41-1, National Natural Science Foundation of China (No. 21761132007).

References

- 1 A. Hagfeldt, G. Boschloo, L. Sun, L. Kloo and H. Pettersson, *Chem. Rev.*, 2010, **110**, 6595–6663.
- 2 L. Zhang and J. M. Cole, *J. Mater. Chem. A*, 2017, **5**, 19541–19559.
- 3 S. Mathew, A. Yella, P. Gao, R. Humphry-Baker, B. F. E. Curchod, N. Ashari-Astani, I. Tavernelli, U. Rothlisberger, M. K. Nazeeruddin and M. Grätzel, *Nat. Chem.*, 2014, **6**, 242–247.
- 4 M. P. Balanay and D. H. Kim, *Phys. Chem. Chem. Phys.*, 2008, **10**, 5121–5127.
- 5 X. Liao, H. Zhang, J. Huang, G. Wu, X. Yin and Y. Hong, *Dyes Pigm.*, 2018, **158**, 240–248.
- 6 C. V. K. Lingamallu Giribabu and P. Y. Reddy, *J. Porphyrins phthalocyanines*, 2006, **10**, 1007–1016.
- 7 M. K. Nazeeruddin, R. Humphry-Baker, D. L. Officer, W. M. Campbell, A. K. Burrell and M. Grätzel, *Langmuir*, 2004, **20**, 6514–6517.
- 8 M. P. Balanay, C. V. P. Dipaling, S. H. Lee, D. H. Kim and K. H. Lee, *Sol. Energy Mater. Sol. Cells*, 2007, **91**, 1775–1781.
- 9 A. Forneli, M. Planells, M. A. Sarmentero, E. Martinez-Ferrero, B. C. O'Regan, P. Ballester and E. Palomares, *J. Mater. Chem.*, 2008, **18**, 1652–1658.
- 10 H. J. Snaith, *Adv. Funct. Mater.*, 2010, **20**, 13–19.
- 11 V. Venkataraman, S. Abburu and B. K. Alsberg, *Phys. Chem. Chem. Phys.*, 2015, **17**, 27672–27682.
- 12 H. Li, Z. Zhong, L. Li, R. Gao, J. Cui, T. Gao, L. H. Hu, Y. Lu, Z. M. Su and H. Li, *J. Comput. Chem.*, 2015, **36**, 1036–1046.
- 13 K. Huwig, C. Fan and M. Springborg, *J. Chem. Phys.*, 2017, **147**, 234105.
- 14 M. Springborg, S. Kohaut, Y. Dong and K. Huwig, *Comput. Theor. Chem.*, 2017, **1107**, 14–22.
- 15 H. G. Tsai, J. Hu, C. Tan, Y. Sheng and C. Chiu, *J. Phys. Chem. A*, 2016, **120**, 8813–8822.
- 16 W. Fan, D. Tan and W. Deng, *Phys. Chem. Chem. Phys.*, 2011, **13**, 16159–16167.
- 17 M. Karami, A. R. S. Beni and B. Hosseinzadeh, *Surf. Sci.*, 2017, **664**, 110–119.
- 18 E. Maggio, N. Martsinovich and A. Troisi, *J. Chem. Phys.*, 2012, **137**, 22A508.
- 19 M. Urbani, M. Grätzel, M. K. Nazeeruddin and T. Torres, *Chem. Rev.*, 2014, **114**, 12330–12396.
- 20 G. Seifert, D. Porezag and T. Frauenheim, *Int. J. Quantum Chem.*, 1996, **58**, 185–192.
- 21 T. A. Niehaus, S. Suhai, F. Della Sala, P. Lugli, M. Elstner, G. Seifert and T. Frauenheim, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2001, **63**, 085108.
- 22 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman and D. J. Fox, *Gaussian 09 Revision D.01*, Gaussian Inc., Wallingford, CT, 2013.
- 23 T. Lu and F. Chen, *J. Comput. Chem.*, 2012, **33**, 580–592.
- 24 T. Le Bahers, C. Adamo and I. Ciofini, *J. Chem. Theory Comput.*, 2011, **7**, 2498–2506.

Electronic Supplementary Material (ESI) for Physical Chemistry Chemical Physics.
This journal is © the Owner Societies 2019

Supporting information for:

**Application of an inverse-design method to optimizing
porphyrins in dye-sensitized solar cells**

Chencheng Fan^{*a}, Michael Springborg^{a,b}, Yaqing Feng^c

^a *Physical and Theoretical Chemistry, University of Saarland, 66123
Saarbruecken, Germany.*

^b *Materials Science, Tianjin University, 300350 Tianjin, China.*

^c *School of Chemical Engineering, Tianjin University, 300350 Tianjin,
China.*

E-mail addresses:

chencheng.fan@uni-saarland.de (C. Fan)

m.springborg@m.x.uni-saarland.de (M. Springborg)

yqfeng@tju.edu.cn (Y. Feng)

Fig. 1 shows the 41 functional groups that were considered in the PooMa calculations. Moreover, Table 1 lists the 206 dyes that were identified in earlier, experimental studies plus the values for their five descriptors as obtained in the DFTB calculations.

Table 1: The set of 206 dyes whose PCE was determined experimentally in other studies. Also shown are the calculated values for their electronic properties as calculated using the DFTB method.

Molecule	G	O	D	A	LHE	PCE (exp.)	Molecule name in Ref.	Ref.
1	0.0667663	0.1098637	5.7343099	3.4535395	0.1145571	3.65	Dye-1	1
2	0.0663543	0.1096820	6.0609838	3.4544885	0.1138443	2.92	Dye-2	1
3	0.0664988	0.1103366	4.8131561	3.4586567	0.1141258	1.85	Dye-3	1
4	0.0667569	0.1106514	4.6365321	3.4549942	0.1144396	3.60	Dye-4	1
5	0.0585169	0.1340320	10.2152042	4.0362108	0.1120592	3.12	Ia	2
6	0.0581445	0.1203809	9.4677291	3.7705729	0.1056630	2.12	Ib	2
7	0.0582487	0.1049978	8.3026409	3.2347226	0.0894526	1.56	Ic	2
8	0.0585103	0.0972312	7.3989790	3.0272590	0.0844352	1.01	Id	2
9	0.0651194	0.1489580	5.0898235	3.8586935	0.1140304	3.55	IIa	2
10	0.0646801	0.1383724	4.5471286	3.6183750	0.1074025	4.18	IIb	2
11	0.0649658	0.1212406	4.3987205	3.1155014	0.0917486	4.79	IIc	2
12	0.0646400	0.1126151	3.8027126	2.9405910	0.0863881	5.08	IId	2
13	0.0522444	0.0926442	25.2075402	3.9906128	0.1171719	4.05	4a	3
14	0.0538638	0.1097202	20.7111210	3.8794755	0.1079478	5.26	4b	3
15	0.0451378	0.1281958	16.3328951	4.1582659	0.1109642	2.62	4c	3
16	0.0609398	0.1451275	6.5802371	4.5240353	0.1094510	5.05	LP-1	4
17	0.0610683	0.1323146	6.6710082	4.0286502	0.0959868	6.04	LP-2	4
18	0.0609842	0.1204137	5.9517260	3.6794902	0.0880607	5.35	LP-3	4
19	0.0616440	0.1304654	5.6258422	3.9561696	0.1209219	4.02	LP-4	5
20	0.0577100	0.0921068	19.6625025	3.6961125	0.1280541	4.47	LP-5	5

Table 1 continued from previous page

Molecule	G	O	D	A	LHE	PCE (exp.)	Molecule name in Ref.	Ref.
21	0.0421651	0.1225426	20.2310204	4.2740513	0.1346527	6.14	LP-6	5
22	0.0431682	0.0832682	31.9452022	3.6657477	0.1031059	6.46	LP-11	6
23	0.0300687	0.1006932	37.1778862	4.0805418	0.1069056	7.37	LP-12	6
24	0.0651594	0.0924354	2.5485089	2.6158141	0.0594725	3.94	N-1	7
25	0.0636379	0.0849677	0.9614710	2.4923765	0.0635152	2.93	N-2	7
26	0.0614332	0.0902923	3.9628319	2.7286297	0.0559813	4.16	N-3	7
27	0.0623167	0.1319958	16.7910857	4.4047664	0.0927662	1.69	cz-4	8
28	0.0624102	0.1261617	15.8280472	4.1983894	0.0886179	2.13	cz-6	8
29	0.0623065	0.1152013	14.7572428	3.8539791	0.0813854	1.30	cz-10	8
30	0.0657199	0.1581280	4.5349031	4.5953904	0.1068553	3.01	ZnP	9
31	0.0652765	0.1240605	9.0586813	3.7328133	0.1144729	3.47	CZ-ZnP	9
32	0.0540346	0.1245887	8.9207224	3.6745778	0.1086988	3.62	IDB-ZnP	9
33	0.0542926	0.1312902	4.9425365	3.6583189	0.1141326	2.55	ISB-ZnP	9
34	0.0641569	0.1674612	4.4436781	4.5929910	0.1072287	0.95	4a	10
35	0.0643061	0.1241872	3.5871889	3.3790816	0.0777495	1.23	4b	10
36	0.0613149	0.1523029	4.0288707	4.2330484	0.1136030	0.55	5a	10
37	0.0614268	0.1202159	2.5602694	3.2368591	0.0848363	0.90	5b	10
38	0.0305762	0.0850418	7.1037800	2.1842021	0.0630336	4.84	GY21	11
39	0.0381284	0.0913519	5.7311743	2.1392080	0.0654496	8.90	GY50	11
40	0.0698013	0.1340860	12.2788772	3.9427723	0.1108234	4.11	Zn-1	12
41	0.0694326	0.1195924	11.2653050	3.4081137	0.0944323	4.80	Zn-2	12
42	0.0714856	0.1359144	5.2975855	3.6566619	0.1040171	0.89	Zn-4	12
43	0.0681198	0.1270787	9.9687732	3.8465629	0.1137559	5.20	Zn-3	13

57

Table 1 continued from previous page

Molecule	G	O	D	A	LHE	PCE (exp.)	Molecule name in Ref.	Ref.
44	0.0701037	0.1369030	8.4470794	3.9188586	0.1102227	4.00	Zn-5	13
45	0.0606876	0.1390918	17.4673304	4.2028109	0.1092884	4.00	Zn-8	13
46	0.0699891	0.1261839	12.3239482	3.7383974	0.1082790	2.40	Zn-11	13
47	0.0647567	0.1188279	23.4547244	4.1144920	0.1151459	3.70	Zn-13	13
48	0.0574624	0.1354942	16.3196090	4.0872576	0.1061092	5.10	GD-1	14
49	0.0576122	0.1242851	15.3547937	3.8095844	0.0999381	7.10	GD-2	14
50	0.0578275	0.1148293	14.2016937	3.5310991	0.0931519	5.80	GD-3	14
51	0.0579618	0.1003757	12.4794999	3.0905908	0.0813659	6.40	GD-4	14
52	0.0580412	0.0798598	10.0239670	2.4717253	0.0653706	5.30	GD-5	14
53	0.0574176	0.1220547	14.2477433	3.5191403	0.0919278	6.10	GD-6	14
54	0.0633787	0.0996594	11.8911212	2.9677537	0.0637155	2.08	1b-d-Zn	15
55	0.0583384	0.0724674	16.5569083	3.2440629	0.0688259	2.37	2b-bd-Zn	15
56	0.0518869	0.0723033	17.3013189	3.0943628	0.0677597	3.03	2b-bdta-Zn	15
57	0.0711670	0.1112677	2.3188779	2.5824263	0.0704301	4.34	YD-0	16
58	0.0571101	0.1044525	3.6465939	2.5374905	0.0726405	6.15	YD-1	16
59	0.0707163	0.1077785	2.4395716	2.5352951	0.0589497	6.56	YD-2	16
60	0.0477547	0.1072710	4.8059315	2.6988464	0.0774612	5.34	YD-3	16
61	0.0529775	0.0992576	3.3021361	2.3825721	0.0687246	5.65	YD-4	16
62	0.0395915	0.0795812	4.2701956	2.1760608	0.0820473	2.10	YD-5	16
63	0.0458267	0.1040151	5.9300487	2.6847999	0.0822526	5.13	YD-6	16
64	0.0434028	0.0969086	14.8000218	2.9669484	0.0867171	4.38	YD-7	16
65	0.0513111	0.0706299	14.4246611	2.4099288	0.0734713	4.27	YD-8	16
66	0.0711670	0.1112677	2.3188779	2.5824263	0.0704301	5.00	YD0	17

Table 1 continued from previous page

Molecule	G	O	D	A	LHE	PCE (exp.)	Molecule name in Ref.	Ref.
67	0.0707163	0.1077785	2.4395716	2.5352951	0.0589497	7.10	YD2	17
68	0.0542112	0.0772139	4.0564086	2.3031583	0.0732740	6.80	YD14	17
69	0.0439099	0.0694964	15.4810452	2.6246536	0.0880480	4.20	YD15	17
70	0.0519541	0.0689608	8.8103707	2.2925680	0.0848171	5.50	YD16	17
71	0.0561651	0.0926398	3.2137225	2.1946879	0.0616570	7.00	YD17	17
72	0.0317356	0.0430464	28.3898544	2.2488689	0.1006172	6.69	XY1	18
73	0.0307369	0.0464320	27.0144801	2.3217648	0.0995922	6.89	XY2	18
74	0.0246106	0.0435574	30.4748570	2.2702888	0.0990199	5.50	XY3	18
75	0.0597190	0.0587322	22.1780946	2.3657366	0.0860160	3.64	CCT1A	19
76	0.0533135	0.0637261	27.7841165	2.6894384	0.0914474	4.80	CCT2A	19
77	0.0489295	0.0682620	35.6102724	2.9571414	0.0925016	5.25	CCT3A	19
78	0.0555831	0.0730346	4.2128616	1.9403036	0.0561517	7.60	YD-2-OCS	20
79	0.0555831	0.0730346	4.2128616	1.9403036	0.0561517	9.40	YD-2-OCS	20
80	0.0741501	0.1123608	1.0770695	2.3705037	0.0688793	4.01	1PEP	21
81	0.0740513	0.1280516	3.6765154	2.6671958	0.0797870	2.55	2PEP	21
82	0.0740785	0.1327977	10.0882432	2.9577075	0.0915939	0.58	3PEP	21
83	0.0612046	0.0808118	17.5427761	2.7310944	0.0796227	2.22	Q1	22
84	0.0643443	0.0793069	13.3271717	2.5787145	0.0770773	5.51	Q2	22
85	0.0651000	0.0628752	8.7421524	1.8751841	0.0563655	7.13	XW1	22
86	0.0633436	0.0679377	9.2710832	1.9548001	0.0590251	6.84	XW2	22
87	0.0601367	0.0710189	6.1817574	1.8861611	0.0571289	7.32	XW3	22
88	0.0615689	0.0691939	3.7172860	1.7146918	0.0516344	7.94	XW4	22
89	0.0552887	0.0758360	32.3656110	2.8962656	0.1117750	9.10	TA-ST-CA	23

Table 1 continued from previous page

Molecule	G	O	D	A	LHE	PCE (exp.)	Molecule name in Ref.	Ref.
90	0.0524638	0.0783148	29.8457557	3.3269538	0.1064191	6.72	DPTP	24
91	0.0488015	0.0745953	28.6722651	3.0716907	0.0952905	7.05	OMeDPTP	24
92	0.0503438	0.0542790	20.7832498	2.2597189	0.0717968	7.64	OHxDPTP	24
93	0.0481101	0.0767060	31.6898283	3.3072438	0.1135833	5.22	M-TP	24
94	0.0520232	0.0448104	12.8216249	1.4405117	0.0593717	4.60	PFT1	25
95	0.0527270	0.0580806	16.1220258	1.8478817	0.0745614	5.81	PFT2	25
96	0.0498505	0.0463602	18.2618960	1.7550009	0.0653839	6.10	PFT4	25
97	0.0490678	0.0578566	25.0361823	2.2980821	0.0810550	6.19	PFT5	25
98	0.0500541	0.0575150	23.8612992	2.2342677	0.0812448	5.76	PFT6	25
99	0.0549575	0.0884838	34.8488172	3.4904077	0.1125149	3.50	DP1	26
100	0.0461346	0.0996140	24.9309775	4.0957710	0.1132304	4.70	DP2	26
101	0.0493274	0.0817432	37.2108174	3.5345121	0.1058133	1.50	DP3	26
102	0.0407387	0.0892162	27.3739533	3.8124314	0.1063739	2.20	DP4	26
103	0.0841135	0.2057307	7.1010683	2.7208610	0.0664321	1.32	PR1	27
104	0.0517700	0.1100617	28.0466053	3.7487154	0.1069148	2.01	PR2	27
105	0.0552221	0.0880710	37.5865174	3.2462465	0.1093803	4.51	1a	28
106	0.0515456	0.0737763	38.6887253	3.1875722	0.1159729	4.30	2a	28
107	0.0448888	0.0806014	43.4726481	3.4968618	0.1139284	1.87	3a	28
108	0.0617975	0.0825254	32.6481218	3.0964956	0.1122582	3.40	4a	28
109	0.0508722	0.0699902	37.6127881	3.1095210	0.1164795	4.81	5a	28
110	0.0563576	0.0601751	26.4729997	2.2710818	0.0797090	4.01	1b	28
111	0.0520953	0.0519731	27.7355773	2.2763563	0.0844611	4.53	2b	28
112	0.0453090	0.0571239	30.8566665	2.5021194	0.0831224	3.60	3b	28

∞

Table 1 continued from previous page

Molecule	G	O	D	A	LHE	PCE (exp.)	Molecule name in Ref.	Ref.
113	0.0641378	0.0582832	21.5817121	2.0887966	0.0829236	3.31	4b	28
114	0.0519093	0.0478448	25.8112761	2.1736651	0.0837734	4.89	5b	28
115	0.0345478	0.1158408	22.4026301	3.6692020	0.1043632	0.95	4a	29
116	0.0313985	0.0958721	24.1906637	3.2062762	0.1334664	4.51	4b	29
117	0.0500921	0.0891723	37.2959570	3.6383843	0.1148349	3.68	3a	30
118	0.0439812	0.0787065	37.1605931	3.4551607	0.1125992	0.35	3b	30
119	0.0553125	0.0651710	8.2952616	2.0958641	0.0608170	7.95	JY47	31
120	0.0360465	0.0753090	11.4781473	2.3113859	0.0710253	8.44	JY48	31
121	0.0361144	0.0742704	12.0550747	2.3025053	0.0739827	7.84	JY49	31
122	0.0726905	0.1329466	2.1572703	3.2870702	0.0677228	1.87	PD1	32
123	0.0542564	0.0879329	9.4990120	3.2851808	0.1168215	2.74	PD2	32
124	0.0411915	0.0919867	23.8557771	3.0991290	0.0867796	1.61	CVHTP	33
125	0.0618008	0.0858590	22.5883618	2.9804705	0.0835124	1.12	CVHTC	33
126	0.0413066	0.0959144	28.5044492	3.4460743	0.0980040	3.20	CVTC-H-CVTP	33
127	0.0394775	0.0372875	12.1225083	1.3838718	0.0763475	6.20	C1	34
128	0.0456119	0.0312782	11.7898422	1.3341227	0.0758650	4.74	C2	34
129	0.0344904	0.0305524	13.4128932	1.3582548	0.0765701	1.42	C3	34
130	0.0291520	0.0716024	10.9348966	1.6340496	0.0940685	6.14	ND	35
131	0.0460617	0.0628970	20.4731641	2.6530180	0.0837996	7.42	SC1	36
132	0.0455990	0.0645827	19.3184989	2.7756639	0.0840953	6.45	SC2	36
133	0.0456559	0.0517875	16.1174775	2.2321817	0.0688872	6.53	SC3	36
134	0.0541271	0.0490485	31.4490748	2.5346917	0.0930642	6.90	M81	37
135	0.0534775	0.0699740	38.1812534	3.1046537	0.0996195	6.15	M82	37

Table 1 continued from previous page

Molecule	G	O	D	A	LHE	PCE (exp.)	Molecule name in Ref.	Ref.
136	0.0479043	0.0497461	15.8997758	1.9558896	0.0709909	4.43	M83	37
137	0.0500849	0.0529846	16.4925284	1.9671548	0.0680051	7.59	M84	37
138	0.0582135	0.0878282	14.5325216	2.8137256	0.0696394	5.99	M85	37
139	0.0841303	0.1048041	8.8130411	1.8789629	0.0704406	0.72	DEA-Q	38
140	0.0506190	0.0810102	19.6715293	2.4798638	0.0966150	0.75	CBZ-Q	38
141	0.0427224	0.0587965	19.0172584	2.1543939	0.0815758	3.07	BPA-Q	38
142	0.0604350	0.0693764	19.8577933	2.8935371	0.1174360	5.65	FWD1	39
143	0.0580204	0.0904394	25.7566080	3.8867700	0.1120075	6.04	FWD2	39
144	0.0603457	0.0841191	24.1186544	3.0040282	0.1209318	5.19	FWD3	39
145	0.0447695	0.0547394	14.2904317	1.9386081	0.0923815	5.64	ZHG5	40
146	0.0451923	0.0616264	15.9976067	2.0818389	0.0926777	5.32	ZHG6	40
147	0.0458692	0.0373386	23.6356043	1.8225969	0.0889805	2.74	ZHG7	40
148	0.0494562	0.0880206	13.2479744	2.3711061	0.0752384	7.55	TBCPCA-1	41
149	0.0492978	0.0832383	12.9728251	2.2690494	0.0724452	7.97	TBCPCA-2	41
150	0.0485320	0.0512228	14.9746384	2.0241526	0.0883079	7.11	TBTCPCA-1	41
151	0.0483990	0.0496290	14.8755703	1.9700435	0.0861409	6.17	TBTCPCA-2	41
152	0.0492426	0.0283154	12.9889047	2.1688351	0.0802441	1.40	FBA1	42
153	0.0602150	0.0240906	13.0003237	1.6102805	0.0687187	2.10	FBA2	42
154	0.0600001	0.0184357	9.9806662	1.2505838	0.0533310	2.50	FBA3	42
155	0.0644035	0.0660986	26.6569484	2.6446677	0.0830919	5.20	CD-1	43
156	0.0663843	0.0592458	24.1027450	2.3756802	0.0912346	4.10	CD-2	43
157	0.0687606	0.0537808	22.3137242	2.2149987	0.0990528	3.50	CD-3	43
158	0.0156502	0.0539460	18.7662736	1.9177332	0.0854602	3.37	SC32	44

Table 1 continued from previous page

Molecule	G	O	D	A	LHE	PCE (exp.)	Molecule name in Ref.	Ref.
159	0.0442793	0.0535942	23.5056563	2.2802517	0.0823760	6.52	SC33	44
160	0.0437057	0.0534638	23.9861220	2.2521598	0.0812267	7.07	SC35	44
161	0.0430359	0.0390825	20.6383475	1.8277278	0.0692110	7.38	SC36	44
162	0.0442080	0.0424417	21.7986307	1.9608395	0.0737957	6.70	SC36N	44
163	0.0527280	0.0575476	17.7988149	3.0560310	0.0944754	7.44	D1	45
164	0.0522667	0.0940092	22.8766610	3.0570864	0.0904328	6.64	D2	45
165	0.0521401	0.0628896	31.3356719	2.8531441	0.0880253	6.72	D3	45
166	0.0527717	0.0704874	37.1575169	3.0258481	0.0983773	6.06	D4	45
167	0.0659430	0.0741071	21.0042222	2.4773849	0.0765677	7.39	QX11	46
168	0.0695141	0.0792216	15.6467279	2.3310899	0.0709776	7.64	QX12	46
169	0.0657324	0.0589542	16.9540920	1.9888067	0.0620891	6.95	QX13	46
170	0.0692903	0.0572709	13.9197013	1.8434155	0.0581264	6.52	QX14	46
171	0.0317029	0.0510672	40.7211582	2.8327700	0.0826220	1.10	DPP1	47
172	0.0322922	0.0353959	25.9853729	1.9561253	0.0669555	1.40	DPP2	47
173	0.0315614	0.0362411	30.0719430	2.0737607	0.0601351	1.80	DPP3	47
174	0.0345150	0.0332382	26.2949546	1.8979877	0.0684235	2.20	DPP4	47
175	0.0417351	0.0745864	50.8924433	3.7784633	0.1065326	0.60	PE4T	47
176	0.0468692	0.0581254	28.2242522	2.7360594	0.0668775	3.40	MK-2	47
177	0.0392516	0.0690701	15.5298974	2.4799794	0.0930063	7.19	QBT-3	48
178	0.0388931	0.0701548	13.8409102	2.4168808	0.0868444	6.78	CTY-1	48
179	0.0292251	0.0640481	17.4939690	2.5255762	0.0904998	8.27	CTY-2	48
180	0.0388931	0.0701548	13.8409102	2.4168808	0.0868444	7.60	CTY-3	48
181	0.0292251	0.0640481	17.4939690	2.5255762	0.0904998	8.10	CTY-4	48

Table 1 continued from previous page

Molecule	G	O	D	A	LHE	PCE (exp.)	Molecule name in Ref.	Ref.
182	0.0712902	0.0833525	1.2577618	2.0866102	0.0536498	5.03	ZnP3C	49
183	0.0504375	0.0716576	26.5928157	2.8331785	0.0632194	1.65	X66	50
184	0.0508574	0.0682930	23.8054083	2.6036665	0.0662447	3.67	X67	50
185	0.0505090	0.0706008	24.3234921	2.6846997	0.0675389	3.24	X68	50
186	0.0509958	0.0609839	22.9207043	2.4686164	0.0739502	4.21	X69	50
187	0.0671807	0.0805556	1.0923534	1.8670982	0.0481617	9.01	LD14	51
188	0.0669543	0.0880626	1.2356082	1.9481770	0.0490203	9.53	LW4	51
189	0.0669449	0.0900709	2.4224591	2.0487544	0.0517304	8.16	LW5	51
190	0.0489638	0.1000547	2.5089851	2.1347794	0.0576524	9.21	LW24	51
191	0.0525938	0.1101084	17.2953206	3.1584463	0.0835753	4.39	T2-1	52
192	0.0225351	0.0843428	22.3894841	2.8046683	0.1199625	5.89	PZ-1	52
193	0.0180974	0.0672775	22.0673579	2.4616481	0.1245091	2.62	PZ-2	52
194	0.0209844	0.0739312	22.0094535	2.5514273	0.1238206	4.97	PZ-3	52
195	0.0208503	0.0634133	21.2871276	2.3910134	0.1047528	6.35	PZ-4	52
196	0.0689817	0.0648257	12.9473879	3.2856018	0.0957432	1.37	OF-Car-Car	53
197	0.0819577	0.0826697	4.6338681	2.7784854	0.0916952	2.04	OF-Py-Py	53
198	0.0705968	0.0729861	32.3979744	3.2201559	0.0995803	0.07	OF-Cat-Cat	53
199	0.0749618	0.0810413	27.3428612	2.9627925	0.0992052	1.51	OF-Py-Car	53
200	0.0753765	0.0785730	10.4911300	3.1440620	0.0995364	0.06	OF-Py-Cat	53
201	0.0632342	0.0887058	20.0629795	2.8690664	0.1030617	1.75	MH	54
202	0.0601685	0.0807677	21.7355627	3.0067993	0.1096920	0.99	MT	54
203	0.0569366	0.0858796	10.8433973	3.6573353	0.0965582	2.70	DH	54
204	0.0551298	0.0692840	11.0962548	3.6412735	0.1094707	1.42	DT	54

Table 1 continued from previous page

Molecule	G	O	D	A	LHE	PCE (exp.)	Molecule name in Ref.	Ref.
205	0.0642779	0.0940782	18.5053841	3.0665354	0.1097282	2.31	TPAC1	54
206	0.0578251	0.0767625	11.2128556	3.6531465	0.1069505	2.73	TPAC2	54

References

- [1] X. Liu, C. Li, X. Peng, Y. Zhou, Z. Zeng, Y. Li, T. Zhang, B. Zhang, Y. Dong, D. Sun, P. Cheng and Y. Feng, *Dyes and Pigments*, 2013, **98**, 181–189.
- [2] Z. Zeng, B. Zhang, C. Li, X. Peng, X. Liu and S. Meng, *Dyes and Pigments*, 2014, **100**, 278–285.
- [3] Y. Liang, X. Xue, W. Zhang, C. Fan, Y. Li, B. Zhang and Y. Feng, *Dyes and Pigments*, 2015, **115**, 7–16.
- [4] F. Lu, J. Zhang, Y. Zhou, Y. Zhao, B. Zhang and Y. Feng, *Dyes and Pigments*, 2016, **125**, 116–123.
- [5] F. Lu, Y. Feng, X. Wang, Y. Zhao and G. Yang, *Dyes and Pigments*, 2017, **139**, 255–263.
- [6] F. Lu, X. Wang, Y. Zhao, G. Yang, J. Zhang, B. Zhang and Y. Feng, *Journal of Power Sources*, 2016, **333**, 1–9.
- [7] N. Zhang, B. Zhang, J. Yan, X. Xue and X. Peng, *Renewable Energy*, 2015, **77**, 579–585.
- [8] X. Xue, W. Zhang, N. Zhang, C. Ju, X. Peng, Y. Yang, Y. Liang, Y. Feng and B. Zhang, *RSC Advances*, 2014, **4**, 8894–8900.
- [9] Y. Zhou, N. A. Lee, K. T. Ngo and X. Peng, *RSC Advances*, 2015, **5**, 41193–41202.
- [10] N. Zhang, B. Zhang, L. Sun and Y. Li, *Research on Chemical Intermediates*, 2015, **41**, 8713–8724.
- [11] A. Yella, C.-L. Mai, S. M. Zakeeruddin, S.-N. Chang, C.-H. Hsieh, C.-Y. Yeh and M. Grätzel, *Angewandte Chemie International Edition*, 2014, **53**, 2973–2977.
- [12] M. K. Nazeeruddin, R. Humphry-Baker, D. L. Officer, W. M. Campbell, A. K. Burrell and M. Grätzel, *Langmuir*, 2004, **20**, 6514–6517.
- [13] Q. Wang, W. M. Campbell, E. E. Bonfantani, K. W. Jolley, D. L. Officer, P. J. Walsh, K. Gordon, R. Humphry-Baker, M. K. Nazeeruddin and M. Grätzel, *Journal of Physical Chemistry B*, 2005, **109**, 15397–15409.

- [14] W. M. Campbell, K. W. Jolley, P. Wagner, K. Wagner, P. J. Walsh, K. C. Gordon, L. Schmidt-Mende, M. K. Nazeeruddin, Q. Wang, M. Grätzel and D. L. Officer, *Journal of Physical Chemistry C*, 2007, **111**, 11760–11762.
- [15] J. K. Park, H. R. Lee, J. P. Chen, H. Shinokubo, A. Osuka and D. Kim, *Journal of Physical Chemistry C*, 2008, **112**, 16691–16699.
- [16] C.-P. Hsieh, H.-P. Lu, C.-L. Chiu, C.-W. Lee, S.-H. Chuang, C.-L. Mai, W.-N. Yen, S.-J. Hsu, E. W.-G. Diau and C.-Y. Yeh, *Journal of Materials Chemistry*, 2010, **20**, 1127–1134.
- [17] S.-L. Wu, H.-P. Lu, H.-T. Yu, S.-H. Chuang, C.-L. Chiu, C.-W. Lee, E. W.-G. Diau and C.-Y. Yeh, *Energy & Environmental Science*, 2010, **3**, 949–955.
- [18] Y. Li, B. Xu, P. Song, F. Ma and M. Sun, *The Journal of Physical Chemistry C*, 2017, **121**, 12546–12561.
- [19] S. Jungstittiwong, R. Tarsang, T. Sudyoadsuk, V. Promarak, P. Khongpracha and S. Namuangruk, *Organic Electronics: physics, materials, applications*, 2013, **14**, 711–722.
- [20] A. Yella, H.-W. Lee, H. N. Tsao, C. Yi, A. K. Chandiran, M. Nazeeruddin, E. W.-G. Diau, C.-Y. Yeh, S. M. Zakeeruddin and M. Grätzel, *Science*, 2011, **334**, 629 LP – 634.
- [21] T. Keawin, R. Tarsang, K. Sirithip, N. Prachumrak, T. Sudyoadsuk, S. Namuangruk, J. Roncali, N. Kungwan, V. Promarak and S. Jungstittiwong, *Dyes and Pigments*, 2017, **136**, 697–706.
- [22] Y. Wang, B. Chen, W. Wu, X. Li, W. Zhu, H. Tian and Y. Xie, *Angewandte Chemie - International Edition*, 2014, **53**, 10779–10783.
- [23] S. Hwang, J. H. Lee, C. Park, H. Lee, C. Kim, C. Park, M.-H. Lee, W. Lee, J. Park, K. Kim *et al.*, *Chemical Communications*, 2007, 4887–4889.
- [24] L.-Y. Lin, C.-H. Tsai, K.-T. Wong, T.-W. Huang, C.-C. Wu, S.-H. Chou, F. Lin, S.-H. Chen and A.-I. Tsai, *Journal of Materials Chemistry*, 2011, **21**, 5950–5958.
- [25] C.-J. Liang, C. P. Kumar, C.-T. Li and J. T. Lin, *Asian Journal of Organic Chemistry*, 2018, 1–11.

- [26] D. S. Patil, K. K. Sonigara, M. M. Jadhav, K. C. Avhad, S. Sharma, S. S. Soni and N. Sekar, *New Journal of Chemistry*, 2018, **42**, 4361–4371.
- [27] S. Kotteswaran, M. S. Pandian and P. Ramasamy, *Journal of Materials Science: Materials in Electronics*, 2018, **29**, 6672–6678.
- [28] H. Masui, M. M. Maitani, S. Fuse, A. Yamamura, Y. Ogomi, S. Hayase, T. Kaiho, H. Tanaka, Y. Wada and T. Takahashi, *Asian Journal of Organic Chemistry*, 2018, **7**, 458–464.
- [29] S. S. Fernandes, A. Pereira, D. Ivanou, A. Mendes and M. M. M. Raposo, *Dyes and Pigments*, 2018, **151**, 89–94.
- [30] S. S. M. Fernandes, M. C. R. Castro, A. I. Pereira, A. Mendes, C. Serpa, J. Pina, L. L. G. Justino, H. D. Burrows and M. M. M. Raposo, *ACS Omega*, 2017, **2**, 9268–9279.
- [31] B. Pan, Y. Z. Zhu, D. Ye and J. Y. Zheng, *Dyes and Pigments*, 2018, **150**, 223–230.
- [32] L. Bao, R. Cheruku, S. Thogiti, P. Ho, W. Yang and J. Kim, *Nanoscience and Nanotechnology Letters*, 2017, **9**, 1–7.
- [33] H. Wei, J. Shen, Y. Liu, T. Huang, Q. Zhang, J. Zhao and X. Zhao, *Dyes and Pigments*, 2018, **149**, 789–795.
- [34] L. Zheng, Q. Cao, J. Wang, Z. Chai, G. Cai, Z. Ma, H. Han, Q. Li, Z. Li and H. Chen, *ACS Omega*, 2017, **2**, 7048–7056.
- [35] P. Ferdowsi, Y. Saygili, W. Zhang, T. Edvinson, L. Kavan, J. Mokhtari, S. M. Zakeeruddin, M. Grätzel and A. Hagfeldt, *ChemSusChem*, 2018, **11**, 494–502.
- [36] C. Shen, Y. Wu, W. Zhang, H. Jiang, H. Zhang, E. Li, B. Chen, X. Duan and W. H. Zhu, *Dyes and Pigments*, 2018, **149**, 65–72.
- [37] H. Cheng, Y. Wu, J. Su, Z. Wang, R. P. Ghimire, M. Liang, Z. Sun and S. Xue, *Dyes and Pigments*, 2018, **149**, 16–24.
- [38] M. Mao, J. B. Wang, X. L. Liu, G. H. Wu, X. Q. Fang and Q. H. Song, *Spectrochimica Acta - Part A: Molecular and Biomolecular Spectroscopy*, 2018, **190**, 23–32.

- [39] H. Wang, B. Bao, X. Hu and J. K. Fang, *Electrochimica Acta*, 2017, **250**, 278–284.
- [40] S. G. Chen, H. L. Jia, X. H. Ju and H. G. Zheng, *Dyes and Pigments*, 2017, **146**, 127–135.
- [41] K. Stalindurai, A. Karuppasamy, J. D. Peng, K. C. Ho and C. Ramalingan, *Electrochimica Acta*, 2017, **246**, 1052–1064.
- [42] I. Pecnikaj, D. Minudri, L. Otero, F. Fungo, M. Cavazzini, S. Orlandi and G. Pozzi, *New Journal of Chemistry*, 2017, **41**, 7729–7738.
- [43] R. L. Vekariya, J. V. Vaghasiya and A. Dhar, *Organic Electronics: physics, materials, applications*, 2017, **48**, 291–297.
- [44] S. Chaurasia, C. T. Li, M. B. Desta, J. S. Ni and J. T. Lin, *Chemistry - An Asian Journal*, 2017, **12**, 996–1004.
- [45] C. H. Siu, L. T. L. Lee, P. Y. Ho, C. L. Ho, T. Chen, S. Suramitr, S. Hannongbua, Z. Xie, M. Wei and W. Y. Wong, *Chemistry - An Asian Journal*, 2017, **12**, 332–340.
- [46] X. Qian, R. Yan, C. Xu, L. Shao, H. Li and L. Hou, *Journal of Power Sources*, 2016, **332**, 103–110.
- [47] I. Imae, Y. Ito, S. Matsuura and Y. Harima, *Organic Electronics: physics, materials, applications*, 2016, **37**, 465–473.
- [48] J. S. Ni, T. Y. Chiu, W. S. Kao, H. J. Chou, C. C. Su and J. T. Lin, *ACS Applied Materials and Interfaces*, 2016, **8**, 23066–23073.
- [49] A. Charisiadis, V. Nikolaou, K. Karikis, C. Giatagana, K. Chalepli, K. Ladomenou, S. Biswas, G. D. Sharma and A. G. Coutsolelos, *New Journal of Chemistry*, 2016, **40**, 5930–5941.
- [50] Z. Lu, P. Dai, C. Wang, M. Liang, X. Zong, Z. Sun and S. Xue, *Tetrahedron*, 2016, **72**, 3204–3212.
- [51] J. Lu, H. Li, S. Liu, Y.-c. Chang, H.-p. Wu, Y. Cheng, E. W.-g. Diao and M. Wang, *Physical Chemistry Chemical Physics*, 2016, **18**, 6885–6892.
- [52] X. Zhang, F. Gou, J. Shi, H. Gao, C. Xu, Z. Zhu and H. Jing, *RSC Advances*, 2016, **6**, 106380–106386.

- [53] Y. Ooyama, K. Furue, T. Enoki, M. Kanda, Y. Adachi and J. Ohshita, *Physical Chemistry Chemical Physics*, 2016, **18**, 30662–30676.
- [54] B. Hosseinzadeh, A. S. Beni, M. Azari, M. Zarandi and M. Karami, *New Journal of Chemistry*, 2016, **40**, 8371–8381.

3.2 Publication II

Role of the Backbone when Optimizing Functional Groups—A Theoretical Study Based on an Improved Inverse-Design Approach

Chencheng Fan,* Mohammad Molayem, Michael Springborg, Moritz Kick, and Yaqing Feng

Cite This: <https://doi.org/10.1021/acs.jpca.1c10437>

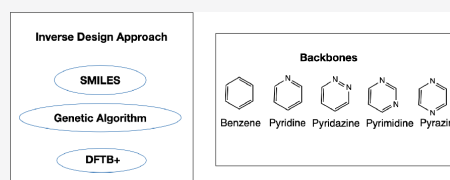
Read Online

ACCESS |

Metrics & More

Article Recommendations

ABSTRACT: We present an improved inverse-design approach for automatically identifying molecular (or other) systems with optimal values for prechosen properties. The new approach uses SMILES (simplified molecular input line entry system) to describe molecular structures efficiently, a genetic algorithm to optimize the molecules automatically, and the DFTB+ (self-consistent charge density functional tight-binding) method to calculate electronic properties. Thereby, almost every class of materials—even macromolecules or monomers—can be studied easily. Without crossover operators but with only mutation operators, the genetic algorithm is more adaptive to SMILES while keeping its efficiency. DFTB+ is more accurate than the DFTB method used in our previous inverse-design approach for the study of excited states and charge transfer processes. The improved approach is applied to optimize benzene, pyridine, pyridazine, pyrimidine, and pyrazine derivatives for seven electronic properties, which all are highly relevant and important for the performance of molecules in solar cells. We found that for some electronic properties, the precise composition and structure of the backbone have remarkable impacts on the value of the electronic properties and/or on the set of functional groups that leads to the best performance. On the contrary, for other properties, these effects are less pronounced. The reasonable optimal functional groups and/or substitution patterns are reported.



1. INTRODUCTION

The identification of optimal materials for specific technical and/or scientific purposes remains a challenging task. The main problem is the almost inexhaustible chemical space.¹ Most often, simple trial-and-error approaches are used for the design of new materials. Thereby, based on experience, instinct, and information from other sources, new materials are synthesized. Subsequently, the properties of interest are studied, and, if possible, relations between property on the one side and structure and composition on the other side are identified that ultimately can lead to the prediction of better materials. This approach is very time-consuming and has low efficiency. It can easily happen that intensive work has to be discarded for the simple reason that the new materials do not possess the desired properties.

Computational screening can help to reduce the search space. A currently very active approach is that of "big data", which is based on collecting vast and accurate information for as many systems and properties as possible.² Subsequently, this huge amount of information is analyzed with the help of high-performance IT technology, and ultimately the thereby gained knowledge is used to predict new materials with better properties. This approach clearly requires powerful computer resources.

An alternative approach, which is the one we shall pursue here, automatically determines the chemical composition of systems with optimal properties within a larger class of

systems. This is the idea behind the inverse-design approaches. Although inverse-design approaches in combination with electronic-structure calculations were proposed more than 20 years ago,³ they have not yet found wider applicability. In 2017, we presented our original idea for an inverse-design approach and applied it initially to Si–Ge clusters as test systems.⁴ The results demonstrated that, in principle, it is possible to identify a stoichiometry that leads to optimized properties within a given class of systems. In a subsequent work, we presented a more general inverse-design approach together with test results on benzene derivatives.⁵ Different from the big data approach that requires expensive computational resources, our inverse-design approaches can run on a normal laptop without expensive hardware or software. In 2019, we applied this approach to identify porphyrin derivatives with optimal performance in dye-sensitized solar cells, which has been a tremendous challenge over the past decades.⁶

Received: December 9, 2021

Revised: January 25, 2022

However, these studies focused on identifying functional groups that are attached to a given backbone (benzene, porphyrin, etc.). When optimizing organic (often, conjugated) molecules, it is often assumed that for similar backbones the same sets of functional groups are identified as providing the best materials. In the present work, we study the validity of this assumption and the role of the backbone when optimizing functional groups. To this end, we apply an improved version of our inverse-design approach to identify the “best” sets of functional groups for different but related backbones. We use the same pool of functional groups for all calculations and consider seven electronic properties. Thus, the outcome is the best set of functional groups for different properties and backbones.

It needs to be noted that our current goal is to study whether the same functional groups are found to be optimum when optimizing some electronic properties for related backbones rather than suggest new molecules for solar cells. In order to reduce the computational costs, we have accordingly considered simpler model systems (molecules related to benzene) and a set of properties that all are relevant for solar-energy harvesting without being exhaustive. Moreover, we are also using a simplified electronic-structure calculator. All these assumptions can be relaxed but sufficient for the question at hand.

In the next section, we describe our improved inverse-design approach as well as our motivation for introducing these improvements. Subsequently, the results are described and discussed in Section 3. We summarize our findings in Section 4.

2. METHODS

Initially, in our inverse-design approaches we define a backbone (to which functional groups will be attached) and a pool of K functional groups. For a given backbone, we determine N substitutional sites, to each of which a functional group can be attached. If possible symmetries of the system are ignored, the total number of possible molecules equals K^N , which easily becomes too large to allow for studying all of them. Instead, we use genetic algorithms to identify candidate systems for the best ones for the property of interest. Nevertheless, still many systems (typically 3 000–12 000 systems depending on the backbone and the property) will have to be studied, making it important to apply efficient computational methods for the calculations of electronic structures and properties.

Therefore, our inverse-design approach involves different ingredients, including an efficient way of describing, designing, and constructing the different molecules, a genetic-algorithm approach for an efficient and fast search in the chemical space defined by the complete set of K^N possible molecules, a sufficiently accurate and fast electronic-structure calculator, and an intensive property that shall be optimized. Our previous inverse-design approach⁵ offered one solution but did suffer from some problems that we have removed in the current work. Here, we shall describe our new and improved approach in comparison with our earlier one.⁵

2.1. Describing, Designing, and Constructing the Molecules. In our earlier work,⁵ each molecule was characterized by a string of N integers with each integer representing a site at which a functional group can be attached. Each integer takes a value between 1 and K describing which of the functional groups is placed at that site. Thus, each system

with a given set of functional groups is uniquely defined through the string of N integers. In addition, the structure of each functional group was described in a predefined coordinate system in real space. For each of the N substitutional sites, we defined a local coordinate system characterized by both its origin and its orientation in the global coordinate system.

It turned out that this way of constructing the molecules was not very efficient. The initial structural information on each functional group and each substitutional site has to be very realistic so that a reasonable initial molecular structure can be generated for further structural optimization. However, when there are many substitutional sites and functional groups, it becomes tedious and difficult. Frequent failures include that the functional groups are not inserted at the right angle or that neighboring functional groups overlap.

Instead, in the new approach, we use an easier and more efficient way to describe and construct molecules through SMILES (simplified molecular input line entry system). In principle, almost every class of materials then can be constructed easily, even for macromolecules or monomers.

SMILES is a line notation for entering and representing molecules. There are six generic SMILES encoding rules, corresponding to the specification of atoms, bonds, branches, ring closures, disconnections, and canonical SMILES.⁷

Also, earlier methods have been presented that are designed for combining the use of SMILES and evolutionary algorithms, but they suffered from some limitations. Thus, Mark Shackelford⁸ developed an approach for drug discovery. At first, it generates an initial population of molecules through a molecular creation process; i.e., a SMILES string is constructed by selecting molecular fragments and atoms from a user-specified list. Then, standard chemical rules and other restrictions (e.g., an upper limit for the molecular weight) are used to determine whether a so-generated molecule can be considered as being realistic. Afterward, genetic algorithms are used to select molecules through three selection operators and subsequently to generate new molecules through 10 mutation operators, including inserting/removing/switching an atom, replacing a single atom with a fragment, removing a bond, cutting/adding a ring, etc. The operators are very flexible and complicated, but the limitation is that the fraction of chemically unrealistic SMILES is high.

Dominique Douguet⁹ developed another approach for small organic drug molecules. The approach can parse each molecule's SMILES in the initial population by scanning chains and detecting rings and parentheses so that they are never cut during substructure extraction. Then, the approach uses a crossover operator to combine substructures from two parents and a mutation operator to modify the substructure from one parent. These operators incorporate chemical knowledge like valency rules. However, its application is confined to small molecules because for bigger molecules parsing SMILES is complicated and goes wrong sometimes, and when generating new molecules, the ratio of valid SMILES is notably lower than that for small molecules.

Noel M. O'Boyle¹⁰ developed an approach which uses genetic algorithms to search the space of six or eight monomer units and examined over 90 000 copolymers, finding the optimal structures with matching optical excitation energies and excited-state energies for highly efficient solar cells. The author used Open Babel 2.2.3 to generate the monomers' 3D structures and the Gaussian 09 software to calculate the electronic properties of the monomers. However, through

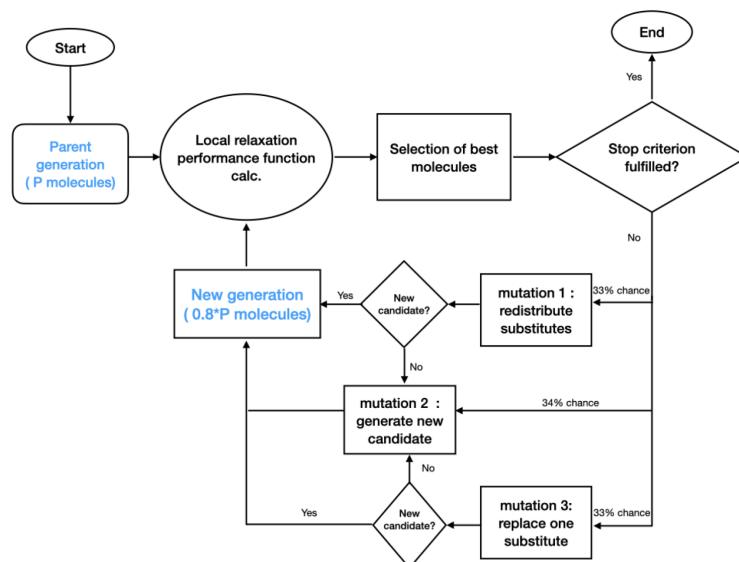


Figure 1. Flowchart of the genetic algorithm.

hundreds of tests we found that for organic molecules Open Babel 2.4.1 sometimes cannot generate reasonable structures, whereas the RDkit program (2018.09.2.0)¹¹ in almost 100% of the cases can. In addition, the Gaussian 09 software is accurate but not fast enough when, as in our case, some thousands of molecules are to be studied. The optical excitation energies and excited-state energies are two important properties but are not sufficient to fully characterize the optical properties of a given material. Therefore, our improved inverse-design approach adopts the RDkit program (2020.09.10) and DFTB + (21.1), which is a faster but accurate enough method. Moreover, here we are studying a larger number of electronic properties.

Different from those approaches, our inverse-design approach adopts an explicit but efficient way of implementation when using SMILES with genetic algorithms. Instead of complicated operators to parse SMILES or to check the validity of SMILES, a “Template SMILES” is defined, which subsequently is modified to get chemically realistic systems. The template SMILES is the SMILES for the core of our molecules and with a special notation at the places where subsequently the SMILES notation for the functional groups will be inserted. We can easily generate SMILES for almost every class of materials, even for macromolecules or monomers, and the percentage of chemically valid SMILES is close to 100%.

In detail, our approach consists of the following steps:

1. Define a pool of (in our case) 18 functional groups using SMILES, including the following: $-\text{CH}_3$, $-\text{CN}$, $-\text{F}$, $-\text{Cl}$, $-\text{Br}$, I , $-\text{OH}$, OCH_3 , $-\text{NH}_2$, $-\text{N}(\text{CH}_3)_2$, $-\text{CHO}$, $-\text{COOH}$, $-\text{C}=\text{CH}_2$, $-\text{C}\equiv\text{CH}$, $-\text{NO}_2$, $-\text{SH}$, $-\text{SCH}_3$, and $-\text{H}$.
2. Define the Template SMILES that consists of the backbone and the substitutional sites.
3. Place the functional groups at the substitutional sites.

4. Create the 3D molecular structure using the RDkit program.

Since each molecule is attributed a “Canonical SMILES” which is the character of the molecule, we can easily recognize and discard repeated systems including symmetrically identical cases.

2.2. The Genetic-Algorithm Approach. Our earlier approach^{4–6} used a genetic algorithm with crossover and mutation operators to generate new molecules after the initial population. Specifically, the crossover operator split the “parent population” into pairs and cut equivalently each member of each pair into two parts that are subsequently interchanged. The mutation operator selects a random substitutional site of a molecule to mutate with another functional group. There, the crossover operator was easy to implement because each molecule was described uniquely through a string of N integers (as discussed in Section 2.1), and strings for different molecules all have the same length.

However, when using SMILES to describe the molecules, crossover operators are difficult to implement without the complicated SMILES-parse (which is used for deciding the cutting places so that a valid SMILES results when interchanging the two parts). So, in the current approach we replace the crossover operator with a new mutation operator to redistribute the functional groups of a molecule. The “traditional” mutation operators are kept that consist of a single replacement of a functional group at a random site or the generation of a completely new set of molecules. The genetic algorithm consists of the following steps (see also the flowchart in Figure 1):

1. Create a pool of P (we chose the value 20) molecules (parents). Each molecule is constructed by placing functional groups randomly. It is guaranteed that these P molecules (as well as the children molecules in the later steps) all are different even after taking symmetry into

account, by checking that the canonical SMILES are not duplicated.

- For each molecule, the structure is relaxed locally, and the performance function is calculated.
- Generate a new generation (children) containing c^*P molecules as follows. (For the constant c , we chose the value 0.8. According to our experience, $c = 0.8$ constitutes a good compromise between diversity (a too small value would result in too few children that are different from the parents) and efficiency (a too large value would produce too many children which requires a longer time to calculate).) At first, select a random member from the P parent molecules. Then, generate a child that has not been considered previously by applying one out of the following three mutation operators: the functional groups are redistributed (33% probability), or one functional group is replaced by any other in the pool of functional groups (33% probability), or the parent is abandoned and a totally new candidate is generated (34% probability). In the case that the former two operators cannot provide a new child within 100 attempts, it turns to the last operator. At the end, the child will enter the new generation pool. This process is repeated until all the c^*P children molecules have been created.
- Relax the structures of the children molecules and calculate their performances.
- Out of the P parents and c^*P children, keep those P molecules with the best performance function as parents for the next generation.
- If the P best molecules are unchanged for Q (we chose the value 40) generations, the calculation is terminated. Otherwise, return to step 3. Besides this optimal termination, there are two more conditions that could lead to a termination of the calculation. This is the case if the calculation has treated X (we chose the value 1000) generations without convergence, or if more than Y (we chose the value 2000) attempts for generating new molecules have failed.

2.3. The Electronic-Structure Calculator. When calculating electronic properties, we use in the present work the DFTB+ code (version 21.1)¹² as implemented in the Atomic Simulation Environment (ASE 3.18.0).¹³ As an improvement over the DFTB method (which we adopted in our earlier approach), the DFTB+ method^{14–16} has the main advantage that self-consistent charge (SCC) calculations can be carried out. Therefore, it is more accurate when calculating excited states and charge transfer processes, but it is still computationally inexpensive.

Within the DFTB+ method, the total energy of a system relative to that of the noninteracting atoms is given as^{14–16}

$$E_{\text{tot}} = \sum_i^{\text{occ}} \epsilon_i - \sum_j \sum_m^{\text{occ}} \epsilon_{jm} + E_{\text{rep}} + \frac{1}{2} \sum_{j,k}^N \Delta q_j \Delta q_k \gamma_{jk} \quad (1)$$

where the first two terms represent the single-electron energy. The third term describes short-range repulsions. The last term is the second-order energy term in density fluctuations. ϵ_i is the energy of the i th occupied orbital for the system of interest, and ϵ_{jm} is that of the m th occupied orbital for the isolated j th atom. Δq_j is the Mulliken net charge on atom j , γ_{jk} depends on the distance between the atoms j and k and on the Hubbard parameters U_j and U_k . Valence electrons are treated explicitly

in the DFTB+ formalism, whereas the effects of the core electrons are included within a frozen-core approximation.

The orbital energies ϵ_i are obtained from the single-particle Kohn–Sham equations

$$\hat{h}_{\text{eff}} \psi_i(\vec{r}) = \left[-\frac{\hbar^2}{2m} \nabla^2 + V(\vec{r}) \right] \psi_i(\vec{r}) = \epsilon_i \psi_i(\vec{r}) \quad (2)$$

whereby the potential $V(\vec{r})$ is approximated as the superposition of potentials of the isolated atoms

$$V(\vec{r}) = \sum_m V_m(|\vec{r} - \vec{R}_m|) \quad (3)$$

with \vec{R}_m being the position of the m th nucleus. Moreover, $\psi_i(\vec{r})$ is approximated through a linear combination of atom-centered functions

$$\psi_i(\vec{r}) = \sum_{\vec{R}} \sum_j c_{i,\vec{R},j} \chi_{\vec{R},j}(\vec{r}) \quad (4)$$

with $\chi_{\vec{R},j}(\vec{r})$ being the j th function centered at the atom at position \vec{R} . Within the DFTB+ approach, it is furthermore assumed that all matrix elements $\langle \chi_{\vec{R}_1, j_1} | \hat{h}_{\text{eff}} | \chi_{\vec{R}_2, j_2} \rangle$ can be obtained from calculations on diatomic molecules.

We study excited states within the framework of the time-dependent density functional tight-binding (TD-DFTB) formalism.¹⁷ The excitation energy w_I for the excited state I is determined from the eigenvalue problem

$$\sum_{ij\sigma} [w_{ij}^2 \delta_{ik} \delta_j \delta_{\sigma\tau} + 2\sqrt{w_{ij} K_{ij\sigma, k\tau}} \sqrt{w_{kl}} F_{ij\sigma}^I] = w_I^2 F_{k\tau}^I \quad (5)$$

with

$$K_{ij\sigma, k\tau} = \iint \psi_i(\vec{r}) \psi_j(\vec{r}') \times \left(\frac{1}{|\vec{r} - \vec{r}'|} + \frac{\delta^2 E_{xc}}{\delta \rho_\sigma(\vec{r}) \delta \rho_\tau(\vec{r}')} \right) \times \psi_k(\vec{r}) \psi_l(\vec{r}') d\vec{r} d\vec{r}' \quad (6)$$

where $w_{ij} = \epsilon_j - \epsilon_i$. i and k mark occupied orbitals, whereas j and l mark unoccupied ones. σ and τ are spin indices.

Subsequently, the oscillator strengths can be calculated from

$$f^I = \frac{2}{3} w_I \sum_{k=x,y,z} \left| \sum_{ij} \langle \psi_i | \vec{r}_k | \psi_j \rangle \sqrt{\frac{w_{ij}}{w_I}} (F_{ij\tau}^I + F_{ij\sigma}^I) \right|^2 \quad (7)$$

2.4. The Properties. For solar-energy harvesting, the power conversion efficiency (PCE) is the property that is optimized. It is defined as

$$\text{PCE} (\%) = \frac{J_{\text{sc}} \cdot \text{FF} \cdot V_{\text{oc}}}{P_{\text{in}}} \times 100\% \quad (8)$$

P_{in} is the intensity of the incident light, and FF is the fill factor, which can be influenced by external factors such as electrode thickness and composition of the electrolyte.¹⁸ J_{sc} (the short-circuit current) and V_{oc} (the open-circuit voltage) are influenced by the dye and very critical for PCE.

Here we shall study seven important molecular properties that all have impacts on the PCE. They all are defined as intensive properties (see below) to avoid that the best performance is found for the smallest or largest system. Moreover, the light harvest efficiency LHE and the thermodynamic driving force for electron injection ΔG_{inj} are

D

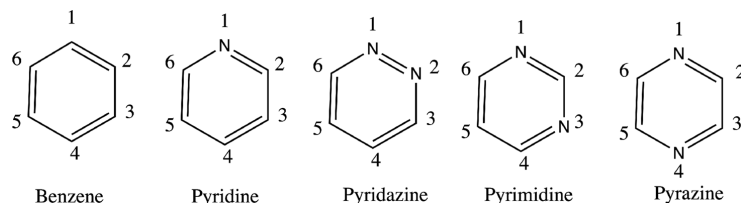


Figure 2. Structures of benzene, pyridine, pyridazine, pyrimidine, and pyrazine. The numbers next to the substitutional sites refer to the positions listed in Tables 1–7 but do not influence the optimization. Symmetries make the sequence of functional groups in substitutional sites for a given molecule nonunique. For benzene, the sets of functional groups in sequences (1, 2, 3, 4, 5, 6), (2, 3, 4, 5, 6, 1), (3, 4, 5, 6, 1, 2), (4, 5, 6, 1, 2, 3), (5, 6, 1, 2, 3, 4), (6, 1, 2, 3, 4, 5), (6, 5, 4, 3, 2, 1), (5, 4, 3, 2, 1, 6), (4, 3, 2, 1, 6, 5), (3, 2, 1, 6, 5, 4), (2, 1, 6, 5, 4, 3), and (1, 6, 5, 4, 3, 2) represent equivalent molecules. Similarly, for pyridine the sets (2, 3, 4, 5, 6) and (6, 5, 4, 3, 2), for pyridazine the sets (3, 4, 5, 6) and (6, 5, 4, 3), for pyrimidine the sets (2, 4, 5, 6) and (2, 6, 5, 4), and for pyrazine the sets (2, 3, 5, 6), (3, 2, 6, 5), (5, 6, 2, 3), and (6, 5, 3, 2) are equivalent. Therefore, when comparing different entries in Tables 1–7 one has to take these symmetries into account.

related to J_{sc} .¹⁹ The dipole moment of the ground state (μ_{norm}) is related to V_{oc} .¹⁹ The rest of properties are related to sun absorption. We did not include the LUMO energy because it is largely correlated to ΔG_{inj} and the HOMO–LUMO energy gap G . We tried to study relatively independent properties.

1. The light harvest efficiency, LHE

$$LHE = \sum_i (1 - 10^{-f_i/N_e}) \quad (9)$$

where f_i is the oscillator strength of the i th singlet excitation. In our calculations we consider only excitations whose excitation energy is below 12.4 eV, corresponding to UV–vis light. N_e is the number of valence electrons.

2. The absorption energy corresponding to the wavelength for which the absorption maximum is found, λ_{max} .
3. The thermodynamic driving force for electron injection, ΔG_{inj} , which is the free energy change defined as²⁰

$$-\Delta G_{inj} = -e(E_{OX}^{dye*} - E_{CB}) \quad (10)$$

where E_{CB} is the reduction potential of the conduction band of TiO_2 , i.e., -0.5 V (vs NHE, pH = 7). E_{OX}^{dye*} is the oxidation potential of the dye (vs NHE) in the excited state

$$eE_{OX}^{dye*} = eE_{OX}^{dye} - E_{0-0} \quad (11)$$

E_{0-0} is the energy difference between the ground and excited states. Accordingly

$$\Delta G_{inj} = e(E_{OX}^{dye} - E_{CB}) - E_{0-0} \quad (12)$$

$\Delta G_{inj} < 0$ implies that the reaction can happen spontaneously.

According to Koopmans' theorem, the first ionization energy of a molecule is equal to the negative of the HOMO energy, and the lowest electron affinity is equal to the negative of the LUMO energy when choosing the potential energy zero infinitely far away. Then

$$\Delta G_{inj} = -E_{HOMO}^{dye} + E_{LUMO}^{TiO_2} - E_{0-0} \quad (13)$$

Here, we simply set $E_{LUMO}^{TiO_2} = -4.0$ eV as the reference²¹ and use the absorption energy λ_{max} as the value for E_{0-0} .

4. Length of the dipole moment of the ground state, μ_{norm} , which has been found to be important for the PCE²²

$$\mu_{norm} = \frac{\sqrt{(\mu_x)^2 + (\mu_y)^2 + (\mu_z)^2}}{N_e} \quad (14)$$

5. The HOMO–LUMO energy gap, G

$$G = \epsilon_{LUMO} - \epsilon_{HOMO} \quad (15)$$

6. The sun light absorption, A

$$A = \sum_i \frac{f_i}{N_e} \cdot P(\epsilon_i) \quad (16)$$

where f_i is the oscillator strength and $P(\epsilon_i)$ is the spectral function for the solar spectrum described with a blackbody approximation

$$P(\epsilon_i) = \frac{8\pi\epsilon_i^3}{(2\pi\hbar c)^3} \left[\exp\left(\frac{\epsilon_i}{k_B T}\right) - 1 \right]^{-1} \quad (17)$$

Again, all the singlet excitations whose energy is below 12.4 eV are taken into account.

7. Degree of unsaturation, DU

$$DU = \frac{2 \cdot N_{triple\ bonds} + N_{double\ bonds} + 0.5 \cdot N_{aromatic\ atoms}}{N_e} \quad (18)$$

where $N_{triple\ bonds}$, $N_{double\ bonds}$, and $N_{aromatic\ atoms}$ are the number of triple bonds, double bonds, and atoms in aromatic rings, respectively.

3. RESULTS AND DISCUSSION

The purpose of this work is to study the role of the backbones when optimizing functional groups and to check the assumption that for similar backbones the same sets of functional groups are optimal for a given property. We optimized the functional groups that are attached to different backbones for the seven electronic properties described in Section 2.4, i.e., the sun light absorption A , the absorption energy corresponding to the wavelength for which the absorption maximum is found λ_{max} , the light harvest efficiency LHE, the degree of unsaturation DU, the dipole moment of the ground state μ_{norm} , the HOMO–LUMO energy gap G , and the thermodynamic driving force for electron injection ΔG_{inj} . The different backbones we chose are benzene, pyridine, pyridazine, pyrimidine, and pyrazine (Figure 2). The four heterocyclic compounds are closely related to benzene including being aromatic while the nitrogen atoms in the backbones may influence the optimization through an electron-withdrawing induction effect ($-I$) or the lone electron pair in sp^2 hybrid orbitals.

Table 1. Optimized Functional Groups That Are Attached to the Five Backbones in Figure 2 for the Absorption A^a

	substitutional sites						A (au)
	1	2	3	4	5	6	
benzene	NH ₂	C≡CH	NO ₂	OH	CH=CH ₂	C≡CH	4.94
	NH ₂	C≡CH	CH=CH ₂	OH	C≡CH	CHO	4.87
	OH	CH=CH ₂	C≡CH	OH	C≡CH	NO ₂	4.85
	NH ₂	C≡CH	CHO	NH ₂	C≡CH	CHO	4.78
	NH ₂	C≡CH	C≡CH	NH ₂	CHO	C≡CH	4.73
pyridine		OH	CH=CH ₂	CHO	OH	CH=CH ₂	5.01
		OH	C≡CH	CHO	OH	CH=CH ₂	4.69
		F	CH=CH ₂	CHO	OH	CH=CH ₂	4.58
		NH ₂	CH=CH ₂	CHO	NH ₂	CH=CH ₂	4.53
		F	C≡CH	CHO	NH ₂	CH=CH ₂	4.51
pyridazine			CH=CH ₂	NH ₂	CHO	CH=CH ₂	3.32
			CH=CH ₂	OH	CHO	CH=CH ₂	3.24
			CH=CH ₂	NH ₂	C≡CH	CH=CH ₂	3.09
			N(CH ₃) ₂	OH	CHO	CH=CH ₂	3.02
			CH=CH ₂	CHO	C≡CH	N(CH ₃) ₂	2.96
pyrimidine		CH=CH ₂		C≡CH	N(CH ₃) ₂	C≡CH	3.80
		CHO		OH	N(CH ₃) ₂	OH	3.80
		NO ₂		OH	N(CH ₃) ₂	F	3.77
		NO ₂		CH=CH ₂	N(CH ₃) ₂	C≡CH	3.69
		CHO		CH=CH ₂	N(CH ₃) ₂	CH=CH ₂	3.62
pyrazine		NH ₂	CH=CH ₂		NH ₂	CH=CH ₂	5.16
		CH=CH ₂	CH=CH ₂		CH=CH ₂	CH=CH ₂	4.93
		CH=CH ₂	OH		CH=CH ₂	OH	4.92
		CH=CH ₂	CH=CH ₂		CH=CH ₂	NH ₂	4.68
		F	CH=CH ₂		OH	CH=CH ₂	4.54

^aFor each system, we present those five best molecules in the results of our inverse-design calculations. The substitutional sites are noted as in Figure 2. The right column gives the calculated values for A in au.

We carried out three sets of calculations for each property and backbone with different random seeds in order to test the stability of the approach. We found that sometimes the optimized molecules are identical (sometimes not!) although the values of the property differ only a little. However, we also found that the functional groups and substitution patterns are always similar. Therefore, the results of our inverse-design approach should be considered as providing insights into molecule design rather than the absolutely optimized systems. It can be important to adopt functional groups/substitution patterns that are common for a given backbone and property, whereas other functional groups are less important and can be chosen according to practical synthesis. For simplicity, we list only one set of results for each property and backbone in Tables 1–7.

We shall start our discussion with the results of Table 1 for the sun light absorption A . The number and positions of nitrogen atoms in the backbone have a strong effect on the value of A , giving high values for benzene, pyridine, and pyrazine but low values for pyridazine and pyrimidine. Thus, there is no simple scaling with the number of nitrogen atoms in the backbone, but also their positions are crucial.

The optimized functional groups are either chromophores or auxochromes. With the six- π -electron delocalization system of the backbone, the chromophores have π -conjugation effects, and the auxochromes have p - π -conjugation effects. For benzene, the best systems possess two auxochromes at opposite sites together with four chromophores at the remaining sites. For pyridine, one nitrogen atom replaces one carbon atom in the benzene ring, so the backbone possesses one less site for substitutions and often a

chromophore is removed rather than an auxochrome. In addition, the auxochromes are attached to one α site and its opposite site, whereas chromophores are attached to the other α site and also at its opposite site. For pyridazine, pyrimidine, and pyrazine, compared with benzene, two nitrogen atoms replace two carbon atoms. For pyrimidine and pyrazine, we do not recognize a special trend toward removing chromophores or auxochromes. For pyridazine, we observe a tendency to keep one auxochrome at site 4 or no auxochrome. For pyrimidine, a chromophore is often at site 2 and an auxochrome at site 5.

When the absorption energy corresponding to the wavelength for which the absorption maximum is found, λ_{max} is optimized (Table 2), the value increases with the number of nitrogen atoms in the backbone; i.e., pyridazine, pyrimidine, and pyrazine possess larger values of λ_{max} than pyridine while benzene has the smallest value.

For all five compounds, the optimized pattern is “push–pull” structures with electron donors (which have conjugation effects, like $-\text{N}(\text{CH}_3)_2$, $-\text{NH}_2$, $-\text{OH}$, $-\text{OCH}_3$, etc.) and electron acceptors (which have conjugation and induction effects, like $-\text{NO}_2$, $-\text{CHO}$, $-\text{COOH}$, etc.). Specially, the $-\text{NO}_2$ group is outstandingly repeatedly found except for the case of pyrazine. For example, $-\text{NO}_2$ is always located at site 3 for pyridazine and at site 2 for pyrimidine.

Next, for the light harvest efficiency LHE (Table 3), the value decreases with the number of the nitrogen atoms in the backbone; i.e., benzene shows a larger value than pyridine while pyridazine, pyrimidine, and pyrazine have the smallest value. It is common for all five compounds that the best candidates possess structures with all the substitute sites

Table 2. As in Table 1 but for λ_{max} , i.e., the Absorption Energy Corresponding to the Wavelength for Which the Absorption Maximum Is Found^a

	substitutional sites						λ_{max} (eV)
	1	2	3	4	5	6	
benzene	Br	NO ₂	CHO	OH	N(CH ₃) ₂	OH	2.34
	Br	NO ₂	Br	SH	N(CH ₃) ₂	OH	2.50
	Br	NO ₂	COOH	OH	N(CH ₃) ₂	OH	2.5
	Br	NO ₂	CN	OH	N(CH ₃) ₂	SH	2.55
	Cl	NO ₂	COOH	OH	N(CH ₃) ₂	OH	2.56
pyridine		OH	CHO	NO ₂	NH ₂	CH=CH ₂	2.61
		OH	NO ₂	Br	NH ₂	CH=CH ₂	2.66
		OH	NO ₂	CH=CH ₂	NH ₂	CH=CH ₂	2.67
		OH	CH=CH ₂	NO ₂	NH ₂	CH=CH ₂	2.71
		OCH ₃	CH=CH ₂	NO ₂	NH ₂	CH=CH ₂	2.73
pyridazine			NO ₂	COOH	Br	N(CH ₃) ₂	3.08
			NO ₂	Br	OH	N(CH ₃) ₂	3.08
			NO ₂	Cl	OH	N(CH ₃) ₂	3.16
			NO ₂	CHO	OCH ₃	N(CH ₃) ₂	3.18
			NO ₂	COOH	Cl	N(CH ₃) ₂	3.19
pyrimidine		NO ₂		OCH ₃	N(CH ₃) ₂	OH	2.75
		NO ₂		SCH ₃	N(CH ₃) ₂	OCH ₃	2.80
		NO ₂		SH	N(CH ₃) ₂	CH ₃	2.91
		NO ₂		OCH ₃	N(CH ₃) ₂	COOH	3.05
		NO ₂		SCH ₃	N(CH ₃) ₂	C≡CH	3.06
pyrazine		NH ₂	CHO		OH	CH=CH ₂	2.92
		OH	CH=CH ₂		OCH ₃	CHO	3.03
		OH	CH=CH ₂		OCH ₃	CH=CH ₂	3.06
		OH	CHO		OCH ₃	CH=CH ₂	3.07
		Cl	CH=CH ₂		NH ₂	NO ₂	3.10

^aThe right column gives the calculated values for λ_{max} in eV.

replaced with -I. Therefore, the iodide atom is believed to be extremely beneficial for LHE. The other possible good substitutes could be -Cl, -Br, -SH, and -C≡CH.

For the degree of unsaturation DU (Table 4), -C≡CH and -C≡N are repeatedly found for all five compounds regardless of the backbones. This finding can be understood from the fact that DU is related only to the geometric structure, and -C≡CH and -C≡N have the largest values of unsaturation among the functional groups.

For the dipole moment of the ground state μ_{norm} (Table 5), all the five backbones have similar values and commonly have push-pull structures with electron donors (like -NH₂ and -OH) and electron acceptors (like -NO₂ and -C≡N). Furthermore, -NO₂ and -NH₂ are always located on the opposite substitute sites of the five backbones. The sites for -NO₂ are very special, such as site 6 for pyridazine and site 2 for pyridine, pyrimidine, and pyrazine, i.e., always next to the nitrogen atom which can play a "pulling" role by an electron-withdrawing induction effect.

For the HOMO-LUMO energy gap G (Table 6), all five cyclic backbones contain electron acceptors -CHO and -NO₂ in the optimized candidates. However, there are also other good functional groups for the five backbones, such as electron donors -N(CH₃)₂ and -OH for benzene and -N(CH₃)₂ for pyrazine and electron acceptors -I and -COOH for pyridine, pyrimidine, and pyridazine. The values of G increase with the number of nitrogen atoms in the backbone; i.e., benzene has smaller value of G than pyridine while pyridazine, pyrimidine, and pyrazine have the largest value. The positions of nitrogen atoms play also a role since pyrazine has a larger value of G than pyrimidine and pyridazine.

In Table 7, all five classes of molecules have similar negative values of the thermodynamic driving force ΔG_{inj} and similar optimized functional groups. The electron-withdrawing groups rarely occur in the optimized molecules, but electron-donating groups (like -N(CH₃)₂, -OCH₃, and -SCH₃, etc.) are always found since these are useful when injecting electrons from the molecule into the TiO₂ conduction band.

In total, for some electronic properties like A , λ_{max} , LHE, and G , the backbone is very crucial and has a strong effect on the value of the property and/or optimized functional groups. On the other hand, for other electronic properties, the effect of the backbone is less pronounced, and different backbones result in very similar functional groups as being the best ones.

4. CONCLUSIONS

The purpose of the present work is to study the role of the backbone in optimizing functional groups and to examine the assumption that for similar backbones the same set of functional groups results in optimal values for a given property. To address this issue, we at first improved our previous inverse-design approach.^{5,6}

Our new approach includes an easier and more efficient way to construct the molecules through SMILES, so that almost every class of materials can be studied, even macromolecules or monomers. Specifically, a "Template SMILES" was defined and subsequently used in generating chemically realistic SMILES in close to 100% of the cases, without having to invoke the complicated operators of parsing SMILES or checking the validity of SMILES. Besides, each candidate is labeled through "Canonical SMILES" (which is the characteristic SMILES of the molecule), so no molecule will be studied more than once.

Table 3. As in Table 1 but for the Light Harvest Efficiency LHE^a

	substitutional sites						LHE
	1	2	3	4	5	6	
benzene	I	I	I	I	I	I	0.33
	I	I	I	I	I	Br	0.32
	I	I	I	I	I	Cl	0.30
	I	I	I	I	I	SH	0.30
	I	I	I	I	I	C≡CH	0.30
pyridine	I	I	I	I	I	I	0.28
	I	I	I	I	I	Br	0.27
	I	I	I	I	Br	I	0.27
	I	I	Br	I	I	I	0.27
pyridazine	I	I	I	I	I	C≡CH	0.27
	I	I	I	I	I	I	0.25
	I	I	I	I	I	Br	0.23
	I	I	I	I	Br	I	0.23
	I	I	I	I	I	C≡CH	0.23
pyrimidine	I	I	I	I	I	I	0.24
	Br	I	I	I	I	I	0.23
	C≡CH	I	I	I	I	I	0.23
	I	I	I	Br	I	I	0.23
	I	I	I	I	Br	I	0.23
pyrazine	I	I	I	I	I	I	0.25
	I	I	I	I	Br	I	0.23
	I	I	C≡CH	I	I	I	0.23
	I	I	I	I	Cl	I	0.22
	I	I	Br	Br	Br	0.22	

^aThe right column gives the calculated values for LHE.

In addition, our new approach is based on an alternative genetic algorithm for the molecular optimization, which has no crossover operator but only mutation operators. Thereby, the genetic algorithm becomes more adaptive to SMILES while simultaneously keeping the efficiency of the genetic algorithm.

Furthermore, our new approach uses a more accurate but still computationally inexpensive electronic-structure method (DFTB+, i.e., self-consistent-charge DFTB) which in particular provides more accurate results on excited states and charge transfer processes.

Although the solar cell is at the focus in the present work and seven electronic properties that are relevant for solar-energy harvesting are considered here, in principle every conceivable property can be optimized. Therefore, our inverse-design approach can have a wider range of applications beyond the solar cell. Besides, the DFTB+ method in our inverse-design approach can easily be replaced by other electronic-structure methods when this should be desirable or necessary.

When studying our theoretical question, we considered molecular backbones related to benzene but with zero, one, or two CH groups replaced by N atoms, whereby very similar backbones are obtained. The optimizations resulted in that for some properties (the sunlight absorption A , the absorption energy corresponding to the wavelength of the absorption maximum λ_{max} , light harvest efficiency LHE, and HOMO–LUMO energy gap G) the number and/or positions of the nitrogen atoms in the backbones have remarkable impacts on the values of the properties and/or on the optimal sets of functional groups. However, for other properties, these effects are less pronounced.

Table 4. As in Table 1 but for the Degree of Unsaturation DU^a

	substitutional sites						DU
	1	2	3	4	5	6	
benzene	CN	CN	CN	C≡CH	CN	H	0.19
	CN	CN	CN	CN	C≡CH	H	0.19
	CN	CN	C≡CH	CN	CN	H	0.19
	CN	CN	C≡CH	CN	C≡CH	H	0.19
	CN	C≡CH	CN	CN	C≡CH	H	0.19
pyridine	CN	C≡CH	C≡CH	C≡CH	C≡CH	H	0.18
	CN	C≡CH	C≡CH	CN	CN	H	0.18
	CN	C≡CH	CN	CN	C≡CH	H	0.18
	CN	CN	C≡CH	C≡CH	C≡CH	H	0.18
	CN	CN	C≡CH	C≡CH	C≡CH	H	0.18
pyridazine	CN	C≡CH	C≡CH	C≡CH	C≡CH	C≡CH	0.18
	CN	C≡CH	C≡CH	CN	C≡CH	C≡CH	0.18
	CN	C≡CH	C≡CH	CN	CN	C≡CH	0.18
	CN	C≡CH	C≡CH	CN	C≡CH	C≡CH	0.18
	CN	C≡CH	CN	CN	C≡CH	C≡CH	0.18
pyrimidine	CN	CN	C≡CH	C≡CH	CN	C≡CH	0.18
	CN	CN	C≡CH	CN	C≡CH	C≡CH	0.18
	C≡CH	CN	C≡CH	CN	C≡CH	CN	0.18
	C≡CH	C≡CH	C≡CH	C≡CH	CN	CN	0.18
	CN	CN	CN	CN	H	C≡CH	0.17
pyrazine	C≡CH	C≡CH	C≡CH	CN	CN	C≡CH	0.18
	CN	CN	CN	C≡CH	C≡CH	C≡CH	0.18
	CN	C≡CH	C≡CH	CN	CN	C≡CH	0.18
	C≡CH	CN	CN	CN	CN	C≡CH	0.18
	CN	CN	CN	C≡CH	H	C≡CH	0.17

^aThe right column gives the calculated values for DU.

Table 5. As in Table 1 but for the Dipole Moment of the Ground State μ_{norm} ^a

	substitutional sites						μ_{norm} (D)
	1	2	3	4	5	6	
benzene	OH	H	NH ₂	H	CN	NO ₂	0.16
	H	H	NH ₂	H	H	NO ₂	0.15
	H	H	NH ₂	H	CN	NO ₂	0.15
	H	H	NH ₂	H	OH	NO ₂	0.15
	H	NH ₂	NH ₂	H	CN	NO ₂	0.15
pyridine		NO ₂	H	H	NH ₂	H	0.16
		NO ₂	OH	H	NH ₂	H	0.16
		NO ₂	CN	H	NH ₂	NH ₂	0.16
		NO ₂	CN	H	NH ₂	H	0.16
		NO ₂	F	H	NH ₂	H	0.16
pyridazine			NH ₂	H	H	NO ₂	0.15
			NH ₂	H	OH	NO ₂	0.15
			NH ₂	NH ₂	H	NO ₂	0.15
			NH ₂	H	F	NO ₂	0.15
			NH ₂	H	CN	NO ₂	0.15
pyrimidine		NO ₂		H	NH ₂	H	0.17
		NO ₂		H	NH ₂	OH	0.16
		NO ₂		H	NH ₂	NH ₂	0.16
		NO ₂		H	N(CH ₃) ₂	H	0.16
		NO ₂		OH	NH ₂	NH ₂	0.16
pyrazine		NO ₂	CN		NH ₂	NH ₂	0.16
		NO ₂	NO ₂		NH ₂	NH ₂	0.16
		NO ₂	F		NH ₂	NH ₂	0.16
		NO ₂	OH		NH ₂	OH	0.16
		NO ₂	H		NH ₂	H	0.16

^aThe right column gives the calculated values for μ_{norm} in Debye.Table 6. As in Table 1 but for the HOMO–LUMO Energy Gap G^a

	substitutional sites						G (eV)
	1	2	3	4	5	6	
benzene	N(CH ₃) ₂	OH	NO ₂	OH	N(CH ₃) ₂	NO ₂	1.15
	N(CH ₃) ₂	OH	NO ₂	OH	OCH ₃	CHO	1.17
	N(CH ₃) ₂	OH	NO ₂	OH	SH	CHO	1.25
	N(CH ₃) ₂	OH	NO ₂	OH	OCH ₃	NO ₂	1.28
	N(CH ₃) ₂	OH	NO ₂	OH	C≡CH	CHO	1.31
pyridine		OH	NO ₂	CHO	I	CHO	1.44
		Cl	NO ₂	CHO	COOH	CHO	1.45
		I	NO ₂	CHO	I	COOH	1.47
		OCH ₃	NO ₂	CHO	COOH	CHO	1.48
		COOH	NO ₂	CHO	COOH	CHO	1.49
pyridazine			CHO	COOH	CHO	NO ₂	1.45
			CHO	I	COOH	CHO	1.54
			CHO	COOH	I	COOH	1.55
			CHO	CHO	I	NO ₂	1.61
			NO ₂	CHO	I	NO ₂	1.62
pyrimidine		Cl		CHO	COOH	CHO	1.55
		NO ₂		CHO	COOH	NO ₂	1.59
		NO ₂		CHO	I	CHO	1.59
		NO ₂		COOH	I	COOH	1.63
		N(CH ₃) ₂		NO ₂	NH ₂	NO ₂	1.64
pyrazine		NO ₂	N(CH ₃) ₂		NO ₂	N(CH ₃) ₂	1.67
		NO ₂	N(CH ₃) ₂		CHO	N(CH ₃) ₂	1.69
		CHO	N(CH ₃) ₂		CHO	N(CH ₃) ₂	1.71
		NO ₂	N(CH ₃) ₂		NO ₂	NH ₂	1.74
		NO ₂	NH ₂		CHO	N(CH ₃) ₂	1.75

^aThe right column gives the calculated values for G in eV.

Table 7. As in Table 1 but for the Thermodynamic Driving Force for Electron Injection ΔG_{inj}^a

	substitutional sites						ΔG_{inj} (eV)
	1	2	3	4	5	6	
benzene	N(CH ₃) ₂	H	OCH ₃	OH	SCH ₃	H	-12.27
	N(CH ₃) ₂	H	N(CH ₃) ₂	OH	SCH ₃	H	-12.07
	N(CH ₃) ₂	H	SCH ₃	SH	OCH ₃	H	-11.86
	N(CH ₃) ₂	H	CH ₃	H	SCH ₃	OH	-11.79
	SCH ₃	H	OH	SCH ₃	H	OCH ₃	-11.69
pyridine		N(CH ₃) ₂	NH ₂	H	SCH ₃	N(CH ₃) ₂	-12.38
		NH ₂	SCH ₃	H	N(CH ₃) ₂	OCH ₃	-12.22
		N(CH ₃) ₂	NH ₂	SCH ₃	H	N(CH ₃) ₂	-12.14
		NH ₂	H	SCH ₃	NH ₂	N(CH ₃) ₂	-12.11
		OH	N(CH ₃) ₂	H	N(CH ₃) ₂	SCH ₃	-12.10
pyridazine			OCH ₃	N(CH ₃) ₂	NH ₂	SCH ₃	-11.64
			N(CH ₃) ₂	OH	CH ₃	SCH ₃	-11.58
			N(CH ₃) ₂	N(CH ₃) ₂	OH	SCH ₃	-11.58
			N(CH ₃) ₂	NH ₂	N(CH ₃) ₂	SCH ₃	-11.56
			N(CH ₃) ₂	C≡CH	OCH ₃	SCH ₃	-11.52
pyrimidine		N(CH ₃) ₂		NH ₂	OH	SCH ₃	-12.0
		N(CH ₃) ₂		OCH ₃	OH	SCH ₃	-11.87
		OH		SCH ₃	OH	N(CH ₃) ₂	-11.78
		OCH ₃		N(CH ₃) ₂	OH	SCH ₃	-11.77
		NH ₂		NH ₂	OH	SCH ₃	-11.74
pyrazine		N(CH ₃) ₂	OH		SCH ₃	OH	-12.14
		N(CH ₃) ₂	OCH ₃		OH	SCH ₃	-11.99
		N(CH ₃) ₂	OCH ₃		H	SCH ₃	-11.79
		N(CH ₃) ₂	OH		H	SCH ₃	-11.76
		N(CH ₃) ₂	NH ₂		SCH ₃	H	-11.71

^aThe right column gives the calculated values for ΔG_{inj} in eV.

In addition, the “push–pull” structure is found to be optimal for λ_{max} and the dipole moment of the ground state μ_{norm} , a finding that is in agreement with experimental molecular design strategies^{2,3} for dye-sensitized solar cells. The pattern with strong electron donors and without electron acceptors is found to be optimal for the thermodynamic driving force ΔG_{inj} . The iodine is strongly beneficial for the light harvest efficiency LHE.

AUTHOR INFORMATION

Corresponding Author

Chencheng Fan – Department of Physical and Theoretical Chemistry, University of Saarland, 66123 Saarbrücken, Germany; orcid.org/0000-0003-0170-5777; Email: s8chfann@stud.uni-saarland.de, 1094440772@qq.com

Authors

Mohammad Molayem – Department of Physical and Theoretical Chemistry, University of Saarland, 66123 Saarbrücken, Germany

Michael Springborg – Department of Physical and Theoretical Chemistry, University of Saarland, 66123 Saarbrücken, Germany; orcid.org/0000-0002-5036-8239

Moritz Kick – Fraunhofer Institute for Solar Energy Systems ISE, 79110 Freiburg, Germany

Yaqing Feng – School of Chemical Engineering and Technology, Tianjin University, 300350 Tianjin, P. R. China

Complete contact information is available at: <https://pubs.acs.org/10.1021/acs.jpca.1c10437>

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the German Research Council through Project (Sp439/41-1) and the National Natural Science Foundation of China (No. 21761132007).

REFERENCES

- Reymond, J.-L. The chemical space project. *Acc. Chem. Res.* **2015**, *48*, 722–730.
- Agrawal, A.; Choudhary, A. Perspective: Materials informatics and big data: Realization of the “fourth paradigm” of science in materials science. *APL Mater.* **2016**, *4*, 053208.
- Franceschetti, A.; Zunger, A. The inverse band-structure problem of finding an atomic configuration with given electronic properties. *Nature* **1999**, *402*, 60–63.
- Springborg, M.; Kohaut, S.; Dong, Y.; Huwig, K. Mixed Si-Ge clusters, solar-energy harvesting, and inverse-design methods. *Comput. Theor. Chem.* **2017**, *1107*, 14–22.
- Huwig, K.; Fan, C.; Springborg, M. From properties to materials: An efficient and simple approach. *J. Chem. Phys.* **2017**, *147*, 234105.
- Fan, C.; Springborg, M.; Feng, Y. Application of an inverse-design method to optimizing porphyrins in dye-sensitized solar cells. *Phys. Chem. Chem. Phys.* **2019**, *21*, 5834–5844.
- SMILES—A Simplified Chemical Language; Daylight Chemical Information Systems, Inc., Laguna Niguel, CA (accessed May 20, 2020); <https://www.daylight.com/dayhtml/doc/theory/theory.smiles.html>.
- Shackelford, M. Evolutionary Algorithm for Drug Discovery-Interim Design Report. arXiv:1403.4871, March 19, 2014.
- Douguet, D.; Thoreau, E.; Grassy, G. A genetic algorithm for the automated generation of small organic molecules: Drug design using

an evolutionary algorithm. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 449–466.

(10) O'Boyle, N. M.; Campbell, C. M.; Hutchison, G. R. Computational Design and Selection of Optimal Organic Photovoltaic Materials. *J. Phys. Chem. C* **2011**, *115*, 16200–16210.

(11) RDKit, version 2018.09.2.0; Open-source cheminformatics, 2018; <http://www.rdkit.org>.

(12) Hourahine, B.; Aradi, B.; Blum, V.; Bonafé, F.; Buccheri, A.; Camacho, C.; Cevallos, C.; Deshayé, M.; Dumitrică, T.; Dominguez, A.; et al. DFTB+, a software package for efficient approximate density functional theory based atomistic simulations. *J. Chem. Phys.* **2020**, *152*, 124101.

(13) Larsen, A. H.; Mortensen, J. J.; Blomqvist, J.; Castelli, I. E.; Christensen, R.; Dulak, M.; Friis, J.; Groves, M. N.; Hammer, B.; Hargus, C.; et al. The atomic simulation environment—a Python library for working with atoms. *J. Phys.: Condens. Matter* **2017**, *29*, 273002.

(14) Elstner, M.; Porezag, D.; Jungnickel, G.; Elsner, J.; Haugk, M.; Frauenheim, T.; Suhai, S.; Seifert, G. Self-consistent-charge density-functional tight-binding method for simulations of complex materials properties. *Phys. Rev. B* **1998**, *58*, 7260.

(15) Porezag, D.; Frauenheim, T.; Köhler, T.; Seifert, G.; Kaschner, R. Construction of tight-binding-like potentials on the basis of density-functional theory: Application to carbon. *Phys. Rev. B* **1995**, *51*, 12947.

(16) Seifert, G.; Porezag, D.; Frauenheim, T. Calculations of molecules, clusters, and solids with a simplified LCAO-DFT-LDA scheme. *Int. J. Quantum Chem.* **1996**, *58*, 185–192.

(17) Niehaus, T. A.; Suhai, S.; Della Sala, F.; Lugli, P.; Elstner, M.; Seifert, G.; Frauenheim, T. Tight-binding approach to time-dependent density-functional response theory. *Phys. Rev. B* **2001**, *63*, 085108.

(18) López-López, C.; Colodrero, S.; Calvo, M. E.; Míguez, H. Angular response of photonic crystal based dye sensitized solar cells. *J. Phys. Chem. C* **2013**, *6*, 1260–1266.

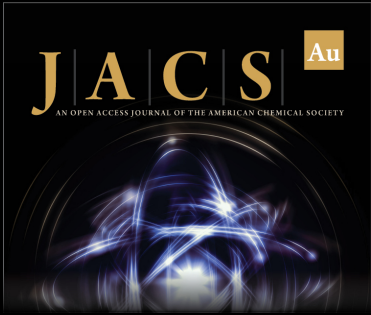
(19) Zhang, J.; Li, H.-B.; Sun, S.-L.; Geng, Y.; Wu, Y.; Su, Z.-M. Density functional theory characterization and design of high-performance diarylamine-fluorene dyes with different π spacers for dye-sensitized solar cells. *J. Mater. Chem.* **2012**, *22*, 568–576.

(20) Katoh, R.; Furube, A.; Yoshihara, T.; Hara, K.; Fujihashi, G.; Takano, S.; Murata, S.; Arakawa, H.; Tachiya, M. Efficiencies of electron injection from excited N3 dye into nanocrystalline semiconductor (ZrO₂, TiO₂, ZnO, Nb₂O₅, SnO₂, In₂O₃) films. *J. Phys. Chem. B* **2004**, *108*, 4818–4822.

(21) Preat, J.; Jacquemin, D.; Michaux, C.; Perpète, E. A. Improvement of the efficiency of thiophene-bridged compounds for dye-sensitized solar cells. *Chem. Phys.* **2010**, *376*, 56–68.

(22) De Angelis, F.; Fantacci, S.; Selloni, A.; Grätzel, M.; Nazeeruddin, M. K. Influence of the sensitizer adsorption mode on the open-circuit potential of dye-sensitized solar cells. *Nano Lett.* **2007**, *7*, 3189–3195.

(23) Liu, B.; Zhu, W.; Wang, Y.; Wu, W.; Li, X.; Chen, B.; Long, Y.-T.; Xie, Y. Modulation of energy levels by donor groups: an effective approach for optimizing the efficiency of zinc-porphyrin based solar cells. *J. Mater. Chem.* **2012**, *22*, 7434–7444.



JACS Au
AN OPEN ACCESS JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

Editor-in-Chief
Prof. Christopher W. Jones
Georgia Institute of Technology, USA

Open for Submissions

pubs.acs.org/jacsau ACS Publications
Most Trusted Most Cited Most Read

3.3 Publication III

From properties to materials: An efficient and simple approach

Cite as: J. Chem. Phys. **147**, 234105 (2017); <https://doi.org/10.1063/1.5009548>

Submitted: 17 October 2017 . Accepted: 20 November 2017 . Published Online: 18 December 2017

Kai Huwig, Chencheng Fan, and Michael Springborg



View Online



Export Citation



CrossMark

ARTICLES YOU MAY BE INTERESTED IN

[Reliability assessment for large-scale molecular dynamics approximations](#)

The Journal of Chemical Physics **147**, 234106 (2017); <https://doi.org/10.1063/1.5009431>

[An atomistic fingerprint algorithm for learning ab initio molecular force fields](#)

The Journal of Chemical Physics **148**, 034101 (2018); <https://doi.org/10.1063/1.5008630>

[Molecular dynamics based enhanced sampling of collective variables with very large time steps](#)

The Journal of Chemical Physics **148**, 024106 (2018); <https://doi.org/10.1063/1.4999447>

PHYSICS TODAY
WHITEPAPERS

ADVANCED LIGHT CURE ADHESIVES

Take a closer look at what these environmentally friendly adhesive systems can do

READ NOW

PRESENTED BY
MASTERBOND
ADVANCED POLYMER SYSTEMS

J. Chem. Phys. **147**, 234105 (2017); <https://doi.org/10.1063/1.5009548>

147, 234105

© 2017 Author(s).



From properties to materials: An efficient and simple approach

Kai Huwig,^{1,a)} Chencheng Fan,^{1,b)} and Michael Springborg^{1,2,c)}

¹Physical and Theoretical Chemistry, University of Saarland, 66123 Saarbrücken, Germany

²School of Materials Science and Engineering, Tianjin University, Tianjin 300072, China

(Received 17 October 2017; accepted 20 November 2017; published online 18 December 2017)

We present an inverse-design method, the poor man's materials optimization, that is designed to identify materials within a very large class with optimized values for a pre-chosen property. The method combines an efficient genetic-algorithm-based optimization, an automatic approach for generating modified molecules, a simple approach for calculating the property of interest, and a mathematical formulation of the quantity whose value shall be optimized. In order to illustrate the performance of our approach, we study the properties of organic molecules related to those used in dye-sensitized solar cells, whereby we, for the sake of proof of principle, consider benzene as a simple test system. Using a genetic algorithm, the substituents attached to the organic backbone are varied and the best performing molecules are identified. We consider several properties to describe the performance of organic molecules, including the HOMO-LUMO gap, the sunlight absorption, the spatial distance of the orbitals, and the reorganisation energy. The results show that our method is able to identify a large number of good candidate structures within a short time. In some cases, chemical/physical intuition can be used to rationalize the substitution pattern of the best structures, although this is not always possible. The present investigations provide a solid foundation for dealing with more complex and technically relevant systems such as porphyrins. Furthermore, our "properties first, materials second" approach is not limited to solar-energy harvesting but can be applied to many other fields, as briefly is discussed in the paper. *Published by AIP Publishing.* <https://doi.org/10.1063/1.5009548>

I. INTRODUCTION

The identification of optimal materials for specific technical and/or scientific purposes remains being a challenging task. The main problem is the almost inexhaustible chemical space.^{1,2} Most often, simple trial-and-error approaches are used for the design of new materials with desired properties. Experimental high-throughput methods or computational screening procedures can help to reduce the search space.³⁻⁷ A currently very active approach is to collect much, often very accurate, information on as many systems and properties as possible, resulting in what has been called "big data," where this huge amount of information is analyzed with the help of high-performance IT systems,^{8,9} an approach that clearly requires powerful computer resources.

An alternative, which is the one we shall pursue here, is to apply simple, approximate methods to calculate the materials properties "on the fly" combined with an efficient optimization approach. Although such inverse-design methods in combination with electronic-structure calculations have been proposed almost 20 years ago,¹⁰ they have not found wide applicability.¹¹⁻¹³ Recently, we presented the initial ideas for an inverse-design approach and applied it to some mixed Si—Ge clusters as a test system.¹⁴ The results demonstrated that it is, in

principle, possible to optimize the properties within a given class of systems, but many issues remained open. In the present work, we shall address most of those open issues, in particular, discussing how to construct molecules automatically and how to identify useful quantities whose values shall be optimized.

As will be discussed further below, our approach can, in principle, be used for many classes of materials and for very many properties. However, we shall here optimize organic molecules and, similar to our earlier work, regarding their properties for solar-cell applications, using this as a test case and illustration. Instead of following the big data trend with expensive computational resources, our approach is based on moving in the opposite direction, i.e., to provide a simple and cost-saving alternative. In this spirit, we use a simple electronic-structure method, which, however, could be replaced with other electronic-structure methods if desirable without having to modify other parts of our approach. This method is combined with an efficient and fast optimization procedure. Our aim is not to be as accurate as possible but instead to identify trends and provide structural motifs for systems that are worthwhile to study experimentally. Because of the approximate nature of our approach, we are able to carry through a detailed study and obtain results in a short time. Actually, our computer program can run on a normal laptop or desktop without expensive hardware or software. In this spirit, our approach will be called PooMa, the Poor Man's Materials Optimization.

Even though we shall use organic molecules within the concept of solar-energy harvesting mainly as a test for our

^{a)}k.huwig@mx.uni-saarland.de

^{b)}s8chfann@stud.uni-saarland.de

^{c)}m.springborg@mx.uni-saarland.de

approach, we emphasize that these systems and properties are currently highly relevant. Within the area of eco-friendly and sustainable renewable energy sources, solar cells are among the most promising systems that also are widely used.^{15–17} To a great extent, the efficiency of solar cells has increased since the 1970s while simultaneously the costs have decreased,¹⁸ and still many efforts are undertaken to improve further the efficiency for solar-energy harvesting and to lower the costs. Over time, also a wide range of different types of cells has been invented, including silicon-based cells,^{19,20} perovskites,^{21,22} quantum dots,²³ dye-based cells,^{24,25} and many others.^{26–28}

Most relevant for our purpose is the work of Reagan and Grätzel that formed the basis for the large interest in dye-sensitized solar cells.²⁹ Besides the low costs of the materials compared to silicon-based solar cells, the good efficiency and stability of these systems are convincing. In dye-sensitized solar cells, organic molecules are the light-absorbing components. The operating principle of such a solar cell is as follows. Via sun-light absorption, an electron in an occupied orbital is excited to a virtual orbital of the dye. Subsequently, this electron is transferred to a semiconductor electrode (e.g., TiO₂) and transported to an external circuit. The mediator, a redox couple (e.g., Co²⁺/Co³⁺), is reduced by this electron, and finally the organic dye is regenerated by the redox system.

The molecules used in such cells consist of an organic core to which different substituents/functional groups are attached. Due to the wide variety of possible functional groups, the optical properties of these molecules can be tailored in a wide range. This makes the dye-sensitized solar cells fascinating and complex at the same time. To investigate these systems by pure experimental research is very cost- and time-consuming. Efficient theoretical techniques can be used to pre-screen the large pool of possible substances and to provide a smaller set of good candidate structures that can be examined further experimentally.

The main idea behind our theoretical approach that will be presented here is to identify that composition of a large class of systems that gives the best value for a predefined property, without the need of investigating all possible systems. Related inverse-design methods have already been applied to other problems in materials science, such as superconducting alloys and photovoltaic materials, crystals, or even molecules for dye-sensitized solar cells.^{10,11,30} Although all methods share the main ideas, there are important details that make it relevant to present our approach in more detail. This will be done in Sec. II, whereas some representative results are presented in Sec. III. Our work concludes with a summary in Sec. IV.

II. THE THEORETICAL APPROACH: PooMa

The main assumption behind our approach is that we have a given backbone to which single atoms or groups of atoms are attached at a given set of sites. The question we are addressing is where to put which group of atoms in order to optimize some property. In the case of the dye-sensitized solar cells, the backbone will be the core (e.g., the porphyrin core) of the dye and the set of sites will be those at which functional groups can

be attached. Alternatively, the core may be a surface of some crystalline compound, and the sites could be the positions of the next (virtual) layer. By placing individual atoms or groups of atoms at those sites, it may be hoped that the catalytic performance of this surface can be optimized. Yet another example is that of an empty crystal structure for which the atom type (including the possibility of having a vacancy) at each site is optimized for whatever purpose.

In the general case, you will have N sites and to each of those attach one out of K groups of atoms (including vacancies, eventually). Thus, each system can be characterized by a string of N integers, each integer being a number between 1 and K . This gives a total of K^N different systems, although symmetry may reduce this.

Our approach consists of the following four parts:

1. An approach that allows for an efficient optimization of the systems. To this, we will apply genetic algorithms as explained in Subsection II A. Moreover, in our preliminary work¹⁴ we demonstrated that this could work well when optimizing clusters consisting of Si and Ge atoms with a pre-chosen structure.
2. An efficient approach for attaching functional groups at given sites.
3. An approach for calculating the property that shall be optimized. Here, we shall use a parameterized density-functional tight-binding method but emphasize that it is possible to replace this method through other methods if so desired.
4. The identification of the property that shall be optimized. Thereby, it shall be avoided that the calculations preferably give the largest possible systems so that the property shall not automatically increase with system size.

In the present work, we shall not only focus on issues 2 and 4 but also describe the other two approaches.

A. The genetic algorithm

For the automatic optimization of properties, we use a genetic algorithm. We consider molecules that consist of the same core but differ in the functional groups that are attached to this core. Any molecule of our interest can be characterized by a string of N integers. Each integer, lying between 1 and K , represents a functional group. The goal is to identify that string of integers, i.e., that substitution pattern, that gives the best performance regarding a predefined property. To this purpose, we apply genetic algorithms on the strings of integers.

In short, our genetic algorithm consists of the following steps:

1. Create randomly $2P$ such molecules (parents). Make sure that these molecules are all different, including that no pairs are equivalent due to symmetry.
2. Relax their structures locally.
3. For each, calculate the function value that describes its performance and that shall be optimized.
4. Separate the molecules randomly into P pairs.
5. From each pair of the 2 molecules, create 2 children through a cutting and mating process. This implies that the two strings of N integers for the parents are cut into parts containing M and $N - M$ integers with M

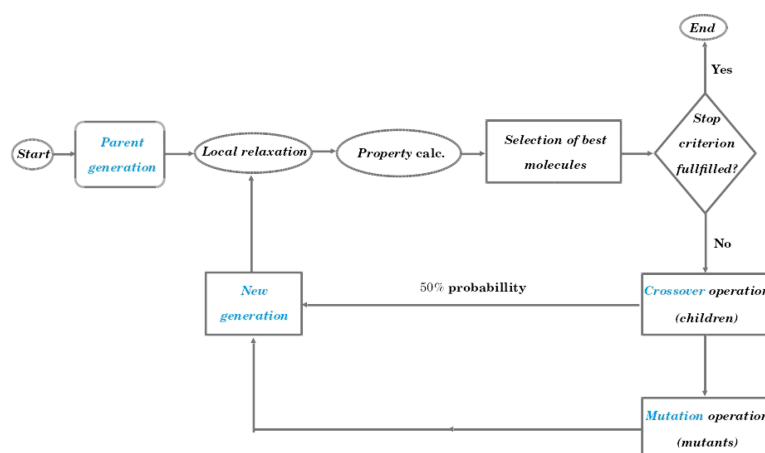


FIG. 1. Flowchart of the genetic algorithm.

($0 < M < N$) being a random number. The two equivalent parts of the parents are interchanged.

6. Include mutations. This means that a random number decides whether a randomly chosen integer in one of the children shall be changed randomly into another integer.
7. Make sure that all the children and parents are different, including that they are not equivalent due to symmetry. If not, repeat 4, 5, and 6.
8. Treat the children as under 2 and 3.
9. Out of the resulting $4P$ molecules, choose those $2P$ with the best performances as the parents for the next generation.
10. If these $2P$ molecules are unchanged for several generations, stop, otherwise return to 4.

A flowchart showing our genetic algorithm can be found in Fig. 1.

The advantage of the genetic algorithm is that it ultimately can lead to up to $2P$ different systems with optimal performances. Thus, as we shall see below in our test case, we will be able to identify building elements that repeat themselves for those systems with good values for the property at hand. This means that even though our approach may suggest molecules that hardly can be synthesized (notice that our approach does not include a test whether the resulting systems can be realized), the identification of common building blocks can be useful in synthesizing other molecular systems containing those building blocks and having good performances.

B. Construction of molecules

In our earlier work, we demonstrated that the genetic-algorithm-based optimization could work for a particularly simple case, i.e., one where a set of positions should be occupied by either Si or Ge atoms. This test case was simple since it did not involve a changing number of atoms or valence electrons and since each “functional group” consists of just one atom. It is much more difficult when a functional group contains more atoms so that also its orientation in space may change depending on at which site it is being attached. In this subsection, we shall describe our approach for achieving this in an efficient way.

Each of the systems of our interest consists of two parts: a “naked” core system (eventually including an anchor group in the case of solar-energy harvesting) and the functional groups. The anchor group is that part of the dye molecule that is attached to the semiconductor. It ensures the transfer of the electrons from the dye to the electrode.

All necessary structural data are written in one input file, and the information is transferred to a small subroutine in our Fortran source code, where the molecules are built. Figure 2 shows the construction procedure exemplarily for a benzene derivative. First the “naked” core system is constructed (1). Then the “empty” substituent sites at which the groups will be attached are defined. For each such site, we define both a typical position (2) and a typical orientation so that a functional group will be placed relative to this position and according to the typical orientation. In order to make the calculations

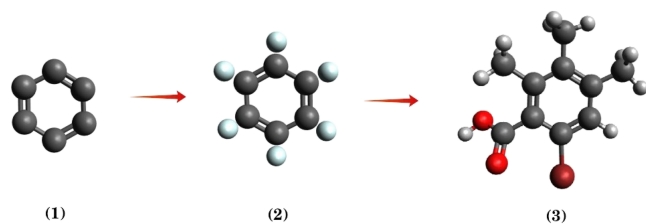


FIG. 2. Steps for molecule construction. “Naked” core system (1), “naked” core system with “empty” substituent sites (2), final molecule (3).

as efficient as possible, it is important that the information on each functional group includes a typical distance relative to the reference position of the empty sites and a realistic orientation. This also implies that it is important that the typical positions and orientations of different empty sites are comparably realistic. In the subsequent calculations, the functional groups will be attached to these empty sites (3).

Functional groups and anchor groups can be attached to the core in the same way, whereby it can be included that the anchor groups stay fixed.

For the calculations to be reported below, we used a pool of 18 functional groups: $-\text{CH}_3$, $-\text{H}$, $-\text{CN}$, $-\text{F}$, $-\text{Cl}$, $-\text{Br}$, $-\text{I}$, $-\text{OH}$, $-\text{OCH}_3$, $-\text{NH}_2$, $-\text{N}(\text{CH}_3)_2$, $-\text{CHO}$, $-\text{COOH}$, $-\text{HC}=\text{CH}_2$, $-\text{C}\equiv\text{CH}$, $-\text{NO}_2$, $-\text{SH}$, and $-\text{SCH}_3$. As an anchor group, we chose the most simple case, the carboxyl group $-\text{COOH}$.

C. Electronic-structure calculation

Whereas we consider the approaches presented in Subsections II A and II B as being crucial, the choice of the electronic-structure method that shall be used to determine both the relaxed structures of the molecules and, subsequently, the numerical values of the property that shall be optimized is not fixed. The method shall be chosen according to the desired accuracy compared to the available computer resources, whereby it should be taken into account that a typical calculation may easily require that some few thousands of systems are studied.

Here, we have chosen to use the non-self-consistent charge density-functional tight-binding (DFTB) method of Seifert and co-workers,^{31,32} which is a choice that is based less on accuracy and more on speed and simplicity. Within the DFTB method, the total energy of a system relative to that of the non-interacting atoms can be expressed as a sum of three terms,

$$E_{\text{tot}} = \sum_i^{\text{occ}} \epsilon_i - \sum_j \sum_m^{\text{occ}} \epsilon_{jm} + \frac{1}{2} \sum_{j \neq k} U_{jk} (|\vec{R}_j - \vec{R}_k|). \quad (1)$$

The first two terms represent the energies of the occupied orbitals of the system minus those of the neutral, isolated atoms, respectively. U_{jk} is a short-ranged pairwise repulsive potential between two atoms j and k . Non-SCC implies that we do not take charge fluctuations and the resulting interactions into account.

The orbital energies are calculated by solving the Kohn-Sham equations,

$$\hat{h}\psi_i = \epsilon_i\psi_i. \quad (2)$$

Thereby, the molecular orbital ψ_i is expanded in a minimal basis set of atom-centered non-orthogonal basis functions,

$$\psi_i(\vec{r}) = \sum_{jp} c_{jpi} \phi_{jp}(\vec{r}), \quad (3)$$

where ϕ_{jp} is the j th function centered at the p th atom.

The single-particle, Kohn-Sham operator is written as a kinetic-energy term and a superposition of atom-centered potentials,

$$\hat{h} = -\frac{\hbar^2}{2m} \nabla^2 + \sum_n V(|\vec{r} - \vec{R}_n|), \quad (4)$$

with \vec{R}_n being the position of the n th atom. By assuming that $\langle \phi_{j_1 p_1}(\vec{r}) | V(|\vec{r} - \vec{R}_n|) | \phi_{j_2 p_2}(\vec{r}) \rangle$ vanishes unless at least one of the atoms p_1 and p_2 equals n , all relevant matrix elements $\langle \phi_{j_1 p_1} | \phi_{j_2 p_2} \rangle$ and $\langle \phi_{j_1 p_1} | \hat{h} | \phi_{j_2 p_2} \rangle$ can be extracted from accurate calculations on diatomic molecules.

Finally, the potential U_{jk} is so adjusted that experimental data or results of *ab initio* calculations for certain test systems are accurately reproduced. Valence electrons are treated explicitly in the DFTB formalism, whereas the effect of the core electrons is included within a frozen-core approximation.

D. Performance function

A crucial issue is to define the function whose value shall be optimized in order to arrive at the best possible system(s). This function should not grow with the system size since the calculations then would predict that the largest system is the best one, i.e., it should be an intensive property. However, it is not obvious how to distinguish between the performances of different molecules that differ in more than just the size. Ultimately, there is no final answer to the question of how to define the performance function that is applicable in any situation. Instead, we shall here discuss various properties that we consider useful in the context of solar-energy harvesting. We add that in our earlier work,¹⁴ where we studied a set of functions for mixed Si-Ge clusters, the issue of having an intensive property was not relevant: all systems had the same number of atoms and number of valence electrons.

In a first attempt to quantify the performance of molecular systems within the concept of dye-sensitized solar cells, we shall here consider several simple properties for which the molecules subsequently will be optimized. For each computation, one property will be considered. It is relevant to add that whether and how the various properties quantify the performance of the molecular systems in solar-energy harvesting is an open question. Moreover, as we found in our earlier study,¹⁴ we cannot exclude that some of the properties are mutually dependent. Since we, in the present study, are varying whole functional groups with different numbers and types of atoms, some molecules become larger than the other ones. In many cases, the properties we consider depend on the number of orbitals of a system and, therefore, on the size. To remove this size dependency, the corresponding properties are divided by the number of valence electrons, N_e .

The properties we shall study are the following:

- The energy gap G between the highest molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO). A good performance is expected when G is as small as possible.

$$G = \epsilon_{\text{LUMO}} - \epsilon_{\text{HOMO}}. \quad (5)$$

- The absorption A of a molecule calculated approximately as the sum of the probabilities $P(\epsilon_a - \epsilon_i)$ for single-electron excitations, weighted with the oscillator strength f_{ia} ,

$$A = \frac{\sum_{i,a} f_{ia} P(\epsilon_a - \epsilon_i)}{N_e}. \quad (6)$$

Here, and below, i denotes an occupied orbital and a denotes an unoccupied one.

The spectral function $P(\epsilon_a - \epsilon_i)$ for the solar spectrum is calculated using a black-body approximation. Then

$$P(\epsilon_a - \epsilon_i) = \frac{8\pi(\epsilon_a - \epsilon_i)^3}{(2\pi\hbar c)^3} \left[\exp\left(\frac{\epsilon_a - \epsilon_i}{k_B T}\right) - 1 \right]^{-1}. \quad (7)$$

The oscillator strength f_{ia} is calculated within the framework of the approximated TD-DFTB (time-dependent density-functional tight-binding) method.^{33–35} Since we only consider one-electron excitations and we do not take into account charge fluctuations, the expression can be simplified compared to the original expression. Thereby, the transition dipole moments \vec{d}_{ia} are approximated in terms of Mulliken transition charges $q_{n,ia}$ and the corresponding coordinates \vec{R}_n of the atoms n for transition $i \rightarrow a$,

$$f_{ia} = \frac{4}{3}(\epsilon_a - \epsilon_i)|\vec{d}_{ia}|^2 = \frac{4}{3}(\epsilon_a - \epsilon_i) \left| \sum_n \vec{R}_n q_{n,ia} \right|^2. \quad (8)$$

- Light-harvesting efficiency, LHE. Using the oscillator strength f_{ia} , the light-harvesting efficiency can be calculated. The larger the LHE, the more light can be harvested by a molecule. All transitions with energies below a pre-defined threshold are taken into account,

$$\text{LHE} = \frac{\sum_{i,a} (1 - 10^{-f_{ia}})}{N_e}. \quad (9)$$

- Orbital overlap O_{av} . To obtain high solar cell efficiency, it is important that the recombination of electron-hole pairs is prevented. Therefore we consider properties that are related to the spatial distribution of the molecular orbitals (MOs). The overlap O_{ia} between an occupied and an unoccupied orbital should be as small as possible and is calculated using atomic gross Mulliken populations $q_{n,i}$ and $q_{n,a}$,

$$O_{ia} = \sum_n |q_{n,i} \cdot q_{n,a}|. \quad (10)$$

This overlap is calculated for each pair of occupied/unoccupied MOs and summed up over all transitions $i \rightarrow a$ to obtain an average value O_{av} for a single molecule. To weight a single overlap according to its transition probability, we multiply with $f_{ia}P(\epsilon_a - \epsilon_i)$,

$$O_{av} = \frac{\sum_{i,a} f_{ia} P(\epsilon_a - \epsilon_i) O_{ia}}{N_e}. \quad (11)$$

- Orbital distance D_{av} . Since we only consider single-electron excitations, we expect that a reduced recombination rate occurs when the spatial distance between an occupied and an unoccupied orbital is large. We quantify this distance using atomic Mulliken gross populations, i.e.,

$$D_{ai} = \left| \sum_n q_{n,a} \vec{R}_n - \sum_n q_{n,i} \vec{R}_n \right|. \quad (12)$$

Just like for the orbital overlap, these distances are weighted and an average value is determined,

$$D_{av} = \frac{\sum_{i,a} f_{ia} P(\epsilon_a - \epsilon_i) D_{ai}}{N_e}. \quad (13)$$

- LUMO distribution, DL. This is another property that quantifies the spatial distribution of the lowest unoccupied molecular orbital but this time by considering its distribution on the anchor group (ag). To obtain a sufficiently good transfer between the dye molecules and the semiconductor electrode, the LUMO should have a noticeable contribution from the atoms of the anchor group. Also this property will be quantified in terms of the atomic Mulliken gross populations,

$$\text{DL} = \sum_{n \in \text{ag}} q_{n,\text{LUMO}}. \quad (14)$$

- The reorganisation energies λ_+ and λ_- . In a dye-sensitized solar cell, electrons are transferred between the molecules. The charge transfer rate can be described by the Marcus theory,³⁶

$$k_{\text{tr}} \propto \frac{1}{\sqrt{\lambda}} t^2 \exp\left(-\frac{\lambda}{4k_B T}\right), \quad (15)$$

where k_{tr} is the charge transfer rate, λ is the reorganisation energy, t is the transfer integral, T is the temperature, and k_B is the Boltzmann constant. Unfortunately, the transfer integral t depends on the relative orientation and distance between adjacent molecules. Calculating this parameter is therefore very demanding and conflicts with our idea of a poor man's materials optimization. In contrast, the reorganisation energy λ of a single molecule is much easier to calculate. Since the charge transfer process should be fast and efficient, this energy of a molecule, while forming cations and anions, should be small. In order to quantify this property using a single number, we average the energies for holes λ_+ and electrons λ_- . The latter are defined by

$$\begin{aligned} \lambda_+ &= (E_0^+ - E_+^+) + (E_+^0 - E_0^0), \\ \lambda_- &= (E_0^- - E_-^-) + (E_-^0 - E_0^0). \end{aligned} \quad (16)$$

E_0^+ (E_0^-) is the total energy of the cation (anion) in the neutral geometry and E_+^+ (E_-^-) is the energy of the neutral molecule in the optimized cationic (anionic) geometry. E_0^0 , E_+^0 , and E_-^0 are the energies for the neutral, cationic, and anionic molecule in their equilibrium structures, respectively.

III. RESULTS

Using benzene as a test system, we shall here discuss the outcome of calculations, whereby functional groups have been attached to the benzene core in order to optimize the properties discussed in Sec. II. As described in Sec. II B, we replace one of the six H atoms with the anchor group $-\text{COOH}$, whereas each of the five remaining H atoms was replaced by one out of the 18 different substituents. Taking symmetry into account, this gives a total of $\frac{1}{2}18^5 \approx 10^6$ different molecules, which hardly can be studied. We shall here demonstrate how our PooMa approach can provide interesting and relevant information that, in principle, can be explored further through synthetic/experimental work. We emphasize that our present results are not the main focus of this work since it only deals with benzene



FIG. 3. The five benzene derivatives with the smallest values for G . Functional groups that occur more frequently are explicitly marked. The performance decreases from left to right.

derivatives but that they may give an insight into the capabilities of our approach and what information these results can provide.

Figure 3 shows the five benzene derivatives with the best (lowest) values for the HOMO-LUMO gap, G . As can be seen, the best candidates share certain mutual functional groups (circled functional groups). Within the framework of our evolutionary algorithm, these substituents may be considered as constituting good fragments. Moreover, all the obtained structures possess a substitution pattern with electron-withdrawing (e.g., $-\text{NO}_2$ and $-\text{CHO}$) and electron-donating groups (e.g., $-\text{NMe}_2$ and $-\text{NH}_2$). This finding is supported by the general strategy for the synthesis of molecules with small HOMO-LUMO gaps: the incorporation of electron-withdrawing and electron-donating groups or fragments in the substitution pattern and the extension of the conjugated system.³⁷

Optimizing the solar-energy absorption and the light-harvesting efficiency results in the systems shown in Fig. 4. The top row represents the molecules with the best (highest) values for absorption A , and the bottom row represents those for the harvesting efficiency LHE. According to chemical intuition, the conjugating effect of $-\text{CHO}$, $-\text{C}\equiv\text{CH}$, and $-\text{NMe}_2$ is beneficial for sunlight absorption and sunlight harvesting efficiency. The repeated occurrence of iodine in the substitution pattern of the molecules for the latter property is, on the other hand, less easy to explain through chemical or physical understanding.

If the spatial separation of the occupied and unoccupied molecular orbitals is large, it may be hoped that electron-hole recombination is prevented. Therefore, we also optimized this quantity. Figure 5 shows those molecules that have a large average orbital distance (upper row) and a small average orbital overlap (bottom row). For the first case, all obtained molecules possess an electron-withdrawing group ($-\text{CHO}$) and an electron-donating group ($-\text{NMe}_2$), which are located opposite to each other. Such substituents and arrangements are obviously beneficial for a large orbital distance. Moreover the two best molecules show such an arrangement even twice. On the other hand, the structures with a small orbital overlap predominantly include electron-donating groups such as $-\text{NMe}_2$ and $-\text{SMe}/-\text{OMe}$. Electron-withdrawing groups are completely absent, except for the presence of a fluorine atom in a single molecule. As above, these substitution patterns also are not easily understood in terms of chemical intuition.

Optimizing DL that describes the distribution of the LUMO on the carboxylic acid anchor group leads to the molecules shown in Fig. 6. Again, certain good fragments occur, namely, two opposing amino groups (circled orange). The fixed anchor group possesses strong electron-withdrawing properties and mainly electron-donating groups ($-\text{OH}$, $-\text{OMe}$) are favoured and prevail within the evolutionary algorithm, except for the fluorine atoms with a strong inductive electron-withdrawing effect. No NO_2 or CN groups occur. So, the COOH group does not have to compete with too

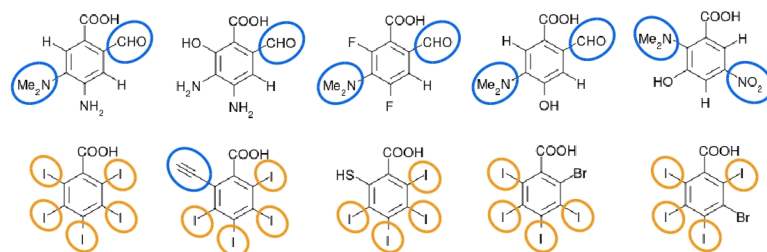


FIG. 4. The top row shows the molecules with the best (largest) value for A , and the bottom row shows those with the best (largest) value for LHE. Functional groups with conjugating effects are circled blue, whereas the iodine substituents are circled orange. The performance decreases from left to right.

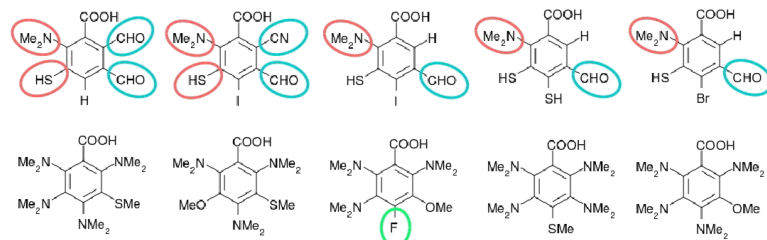


FIG. 5. The top row shows the molecules with the best (largest) value for D_{av} , and the bottom row shows those with the best (smallest) value for O_{av} . Functional groups that occur several times and seem to be beneficial for a good distance are circled. For the overlap, all favoured groups, except the one fluorine atom (circled), belong to the electron-donating ones. The performance decreases from left to right.

234105-7 Huwig, Fan, and Springborg

J. Chem. Phys. 147, 234105 (2017)

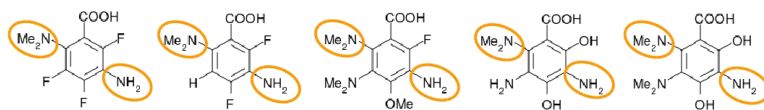


FIG. 6. The five molecules with the best (largest) DL. Multiply occurring substituents are circled. The performance decreases from left to right.

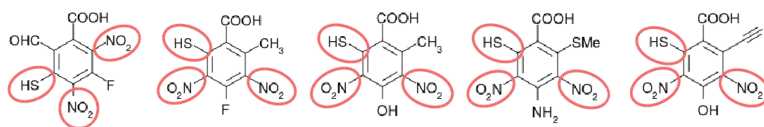


FIG. 7. The five molecules with the best (smallest) λ . Multiply occurring substituents are circled. The performance decreases from left to right.

many other electron-withdrawing groups, which is particularly advantageous when aiming for a high proportion of the LUMO on the anchor group. But chemical intuition fails for the explanation for the order of the structures. It is exactly those two best structures that have the fluorine atoms, whereas for the fourth and fifth structures, only electron-donating substituents appear.

The last property we considered is the reorganisation energy λ . The molecules obtained, when optimizing this performance function, are shown in Fig. 7. In this case, the good fragments are two nitro groups in the meta position to each other and a SH substituent. Due to the NO_2 groups with a $-M$ effect and the SH group with a $+M$ effect, one could construct several mesomeric structures for these molecules indicating that they are highly delocalised systems. This electron delocalisation ensures that the total energy of the system and hence the structure does not change much when forming a cation or anion.

IV. CONCLUSIONS

The main goal of the present work was to present a simple and efficient theoretical scheme with which molecular (or other) systems with optimal values for certain properties automatically are identified. The method is approximate, and the resulting molecules are not studied within the context of being realistic. Nevertheless, as we demonstrated through our examples, the resulting molecules may contain a certain structural pattern that then must be considered as being “good” so that also other related molecules that may easily be synthesized can be expected to have good performances.

Our approach is based on identifying the “best” molecules out of a large class of related ones that differ mainly in substitutions. As an alternative to having to rely on results of expensive, experimental or theoretical studies on such systems, our method is based on searching the molecular space in an intelligent way on the fly. Thereby, it can be implemented on a simple laptop or desktop and requires, accordingly, no expensive resources. In this spirit, we have called our approach PooMa, the Poor Man’s Materials Optimization.

The method combines a genetic-algorithm approach for the optimization with an automatic and efficient way of generating the molecular systems. For the calculation of the properties, we use at the moment, the simple DFTB method that, however, can be replaced by other more accurate methods if this is considered desirable. Finally, the method requires

a mathematical definition of the property whose value shall be optimized.

The outcome of the calculations is a smaller set of optimal systems for a given property. As we showed in our previous work,¹⁴ it may happen that the calculations do not identify the absolutely best system although still very good ones. However, this information is also very useful for experimental studies since our approach can give not only one but several molecules that, in addition, may share certain structural features that then must be considered as being good for the property and system class at hand.

In the present work, we demonstrated the applicability of our approach by studying benzene derivatives within the concept of solar cell applications. However, many other systems and/or properties can be treated with this type of approach too. This includes conjugated organic molecules for non-linear optics applications where functional groups attached to the backbone can be optimized. Another possibility is to identify optimal materials for heterogeneous catalysis, whereby the composition and arrangement of the surface atoms (including, eventually, vacancies) of the catalyst can be optimized. Yet another possibility is to determine the structure and composition of porous solids for the purpose of storage of smaller molecules such as hydrogen. The economical nature of our method leads to limited accuracy, but since we only want to identify trends and pre-screen huge structure pools, we think this approach is justified. Our goal is to give information to experimentalists and make suggestions.

For each property for our test system, our results predicted that the occurrence of certain functional groups will be beneficial to the property at hand. For different properties, the good fragments are different. A general assumption that functional groups are the most promising is not easily identified. Moreover in some cases, chemical/physical understanding is not always capable of explaining the outcome of our calculations. This was, for example, the case of the light harvesting efficiency. In other cases, chemical intuition can very well be of help to understand the results as we found for the solar-light absorption or for the HOMO-LUMO gap.

As a continuation of the present work, the importance of the different properties that we have calculated in actually quantifying the performance of the molecular systems in solar-energy harvesting will be studied. Then, a single quantity that quantifies the performance and that is constructed from the properties we are calculating shall be identified and ultimately optimized. Furthermore, the application to technically more

234105-8 Huwig, Fan, and Springborg

J. Chem. Phys. **147**, 234105 (2017)

relevant systems such as porphyrins or BODIPY will be pushed forward.

Although we have used the dye-sensitized solar cells as motivation and playground, our approach is not limited to solar cells or to optical properties. In principle, every conceivable property can be optimized and other areas such as heterogeneous catalysis or hydrogen storage can be explored. The main idea of first having a property that gives a system with a certain composition as a result is not new. But the combination with a simple, fast approach, that requires no expensive computational resources, is a unique feature of our method. In the age of big data analysis, we have hopefully shown that there are also other promising ways of investigating huge search spaces.

ACKNOWLEDGMENTS

This work was supported by the German Research Council (DFG) through Project No. SP 439/41.

- ¹J.-L. Reymond and M. Awale, "Exploring chemical space for drug discovery using the chemical universe database," *ACS Chem. Neurosci.* **3**, 649–657 (2012).
- ²J.-L. Reymond, "The chemical space project," *Acc. Chem. Res.* **48**, 722–730 (2015).
- ³X.-D. Xiang, X. Sun, G. Briceño, Y. Lou, K.-A. Wang, H. Chang, W. G. Wallace-Freedman, S.-W. Chen, and P. G. Schultz, "A combinatorial approach to materials discovery," *Science* **268**, 1738–1740 (1995).
- ⁴B. Jandeleit, D. J. Schaefer, T. S. Powers, H. W. Turner, and W. H. Weinberg, "Kombinatorische Materialforschung und Katalyse," *Angew. Chem.* **111**, 2648–2689 (1999); Combinatorial materials science and catalysis, *Angew. Chem., Int. Ed.* **38**, 2494–2532 (1999).
- ⁵W. F. Maier, K. Stöwe, and S. Sieg, "Combinatorial and high-throughput materials science," *Angew. Chem., Int. Ed.* **46**, 6016–6067 (2007).
- ⁶J. Greeley, T. F. Jaramillo, J. Bonde, I. Chorkendorff, and J. K. Nørskov, "Computational high-throughput screening of electrocatalytic materials for hydrogen evolution," *Nat. Mater.* **5**, 909–913 (2006).
- ⁷S. Curtarolo, G. L. W. Hart, M. Buongiorno Nardelli, N. Mingo, S. Sanvito, and O. Levy, "The high-throughput highway to computational materials design," *Nat. Mater.* **12**, 191–201 (2013).
- ⁸L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, and M. Scheffler, "Big data of materials science: Critical role of the descriptor," *Phys. Rev. Lett.* **114**, 105503 (2015).
- ⁹A. Agrawal and A. Choudhary, "Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science," *APL Mater.* **4**, 053208 (2016).
- ¹⁰A. Franceschetti and A. Zunger, "The inverse band-structure problem of finding an atomic configuration with given electronic properties," *Nature* **402**, 60–63 (1999).
- ¹¹D. Xiao, L. A. Martini, R. C. Snoberger III, R. H. Crabtree, and V. S. Batista, "Inverse design and synthesis of acac-coumarin anchors for robust TiO₂ sensitization," *J. Am. Chem. Soc.* **133**, 9014–9022 (2011).
- ¹²X. Zhang, Y. Wang, J. Lv, C. Zhu, Q. Li, M. Zhang, Q. Li, and Y. Ma, "First-principles structural design of superhard materials," *J. Chem. Phys.* **138**, 114101 (2013).
- ¹³T. Weymuth and M. Reiher, "Inverse quantum chemistry: Concepts and strategies for rational compound design," *Int. J. Quantum Chem.* **114**, 823–837 (2014).
- ¹⁴M. Springborg, S. Kohaut, Y. Dong, and K. Huwig, "Mixed Si–Ge clusters, solar-energy harvesting, and inverse-design methods," *Comput. Theor. Chem.* **1107**, 14–22 (2017).
- ¹⁵M. A. Green, *Solar Cells: Operating Principles, Technology, and System Applications* (Prentice-Hall, Inc., Englewood Cliffs, NJ, USA, 1982).
- ¹⁶L. M. Fraas and L. D. Partain, *Solar Cells and Their Applications* (John Wiley & Sons, Hoboken, US, 2010).
- ¹⁷S. Sharma, K. K. Jain, and A. Sharma, "Solar cells: In research and applications—A review," *Mater. Sci. Appl.* **6**, 1145–1155 (2015).
- ¹⁸N. Asim, K. Sopian, S. Ahmadi, K. Saeedfar, M. A. Alghoul, O. Saadatian, and S. H. Zaidi, "A review on the role of materials science in solar cells," *Renewable Sustainable Energy Rev.* **16**, 5834–5847 (2012).
- ¹⁹J. Zhao, A. Wang, P. Altermatt, and M. A. Green, "Twenty-four percent efficient silicon solar cells with double layer antireflection coatings and reduced resistance loss," *Appl. Phys. Lett.* **66**, 3636–3638 (1995).
- ²⁰C. Battaglia, A. Cuevas, and S. De Wolf, "High-efficiency crystalline silicon solar cells: Status and perspectives," *Energy Environ. Sci.* **9**, 1552–1576 (2016).
- ²¹G. Hodes, "Perovskite-based solar cells," *Science* **342**, 317–318 (2013).
- ²²M. A. Green, A. Ho-Baillie, and H. J. Snaith, "The emergence of perovskite solar cells," *Nat. Photonics* **8**, 506–514 (2014).
- ²³R. P. Raffaele, S. L. Castro, A. F. Hepp, and S. G. Bailey, "Quantum dot solar cells," *Prog. Photovoltaics: Res. Appl.* **10**, 433–439 (2002).
- ²⁴A. Hagfeldt, G. Boschloo, L. Sun, L. Klöö, and H. Pettersson, "Dye-sensitized solar cells," *Chem. Rev.* **110**, 6595–6663 (2010).
- ²⁵M. Grätzel, "Dye-sensitized solar cells," *J. Photochem. Photobiol. C* **4**, 145–153 (2003).
- ²⁶G. Li, R. Zhu, and Y. Yang, "Polymer solar cells," *Nat. Photonics* **6**, 153–161 (2012).
- ²⁷W. U. Huynh, J. J. Dittmer, and A. P. Alivisatos, "Hybrid nanorod-polymer solar cells," *Science* **295**, 2425–2427 (2002).
- ²⁸M. Wiesenfarth, S. P. Philipps, A. W. Bett, K. Horowitz, and S. Kurtz, "Current status of concentrator photovoltaic (CPV) technology," Report Version 1.3, Fraunhofer Institute for Solar Energy systems (NREL), 2017.
- ²⁹B. O'Regan and M. Grätzel, "A low-cost, high-efficiency solar cell based on dye-sensitized colloidal TiO₂ films," *Nature* **353**, 737–740 (1991).
- ³⁰L. Yu, R. S. Kokenyesi, D. A. Keszler, and A. Zunger, "Inverse design of high absorption thin-film photovoltaic materials," *Adv. Energy Mater.* **3**, 43–48 (2013).
- ³¹D. Porezag, Th. Frauenheim, Th. Köhler, G. Seifert, and R. Kaschner, "Construction of tight-binding-like potentials on the basis of density-functional theory," *Phys. Rev. B* **51**, 12947–12957 (1995).
- ³²G. Seifert, D. Porezag, and Th. Frauenheim, "Calculations of molecules, clusters, and solids with a simplified LCAO-DFT-LDA scheme," *Int. J. Quantum Chem.* **58**, 185–192 (1996).
- ³³T. A. Niehaus, S. Suhai, F. Della Sala, P. Lugli, M. Elstner, G. Seifert, and Th. Frauenheim, "Tight-binding approach to time-dependent density-functional response theory," *Phys. Rev. B* **63**, 085108 (2001).
- ³⁴T. A. Niehaus, "Approximate time-dependent density functional theory," *J. Mol. Struct.: THEOCHEM* **914**, 38–49 (2009).
- ³⁵R. Rüger, E. van Lenthe, T. Heine, and L. Visscher, "Tight-binding approximations to time-dependent density functional theory—A fast approach for the calculation of electronically excited states," *J. Chem. Phys.* **144**, 184103 (2016).
- ³⁶R. A. Marcus, "Electron transfer reactions in chemistry. Theory and experiment," *Rev. Mod. Phys.* **65**, 599–610 (1993).
- ³⁷D. F. Perepichka and M. R. Bryce, "Molecules with exceptionally small HOMO-LUMO gaps," *Angew. Chem., Int. Ed.* **44**, 5370–5373 (2005).

3.4 Publication IV

Computational Condensed Matter 25 (2020) e00503



Contents lists available at ScienceDirect

Computational Condensed Matter

journal homepage: <http://ees.elsevier.com/cocom/default.asp>

Regular article

Application of an inverse-design method for designing new branched thiophene oligomers for bulk-heterojunction solar cells

Abdullah S. Khazaal^{a,*}, Michael Sprinborg^{b,c}, Chencheng Fan^b, Kai Huwig^b^a Chemistry Department, College of Science, Tikrit University, 34001, Salahuddin, Iraq^b Physical and Theoretical Chemistry, University of Saarland, 66123, Saarbrücken, Germany^c Materials Science, Tianjin University, 300350, Tianjin, China

ARTICLE INFO

Article history:

Received 30 May 2020

Received in revised form

24 August 2020

Accepted 25 August 2020

Keywords:

Bulk heterojunction solar cells

Density functional theory

Inverse-design method

Power conversion

Thiophene oligomers

ABSTRACT

Using our recently developed theoretical inverse-design method (PooMa) a new series of branched oligothiophene molecules for photovoltaic donors are designed. PooMa uses a genetic algorithm to screen a huge pool of compounds combined with a fast electronic-structure method (Density-Functional Tight-Binding, DFTB) with reasonable accuracy. Here, we apply this inverse-design method to identify a set of 20 branched oligothiophene systems with promisingly high efficiencies by using a Quantitative Structure Property Relation (QSPR) model based on five electronic descriptors that describe the performance of organic solar cells. We consider a pool of oligomers that are modified by attaching to each of 7 different sites one out of 22 functional groups, i.e., a pool of $22^7 \approx 2.5 \times 10^9$ molecules. Subsequently, density-functional-theory (DFT) and Time-Dependent-DFT (TD-DFT) calculations in the gas phase with a 6-31G(d,p) basis set have been carried through to give further information on the suggested oligomers. Bulk-heterojunction photovoltaic cells were designed with the suggested oligothiophenes as donors and Phenyl-C61-butyric acid methyl (PCBM) derivatives as acceptors. Conversion efficiencies of the designed photovoltaic cells were examined with the Scharber diagram model.

© 2020 Elsevier B.V. All rights reserved.

1. Introduction

The production of cheap, clean, and renewable energy through solar energy harvesting has become of enormous importance in the fight against man-made environmental damage. Thereby, organic solar cells represent a future generation for low-cost power production as they offer low-temperature and flexible substrate processing. The combination of the plastic and semiconducting properties of the organic conducting polymers (COPs) makes them attractive for the development of devices [1–10]. The π -electrons in COPs are delocalized and give an energy bandgap whose value is comparable to that of more conventional semiconductors. This property is combined with high electric conductivity when charge carriers are created through oxidation or reduction or via electronic excitation process. The small bandgap for the π -orbitals makes it possible to obtain emission and absorption of light in the visible region. This unequalled combination of optical and electrical characteristics together with the mechanical properties of plastics has

led to the manufacturing of plastic devices like organic solar cells, organic photovoltaic devices (OPVs), and light-emitting diodes (OLEDs) [11,12].

With the purpose of identifying optimal COPs in organic devices, theoretical and experimental methods have repeatedly been applied (see, e.g. Ref. [13–16]). However, the very large number of possible materials makes this a very big challenge, and the commonly used try-and-error approach may not be optimal. Computational studies methods can help in reducing the required efforts [17–19]. A further improvement in the theoretical studies is obtained when these are used in an efficient screening of the possible materials, although this is an approach that still is at its infancy. In this paper, we shall present the results of such an approach.

Of the important and promising materials are thiophene-based oligomers that have found use in many applications as diverse as OPVs [19,20], OLEDs [21,22], and organic field-effect transistors (OFETs) [23–25]. The efforts have concentrated on donor materials

* Corresponding author.

E-mail address: abdslem77@tu.edu.iq (A.S. Khazaal).

of polymers and p-type conjugated small-molecule systems that have a large absorption of sunlight, ease of film formation, high mobility of charge carriers, and appropriate energy levels of frontier orbitals. In OPVs, conjugated oligomers have some critical advantages over their conjugate polymer systems for several reasons. At first, they have a well-defined structure, which ensures constant performances of the devices because of the synthetic reproducibility. Second, they are thermally stable and can be processed both from solution and in the gas phase. Third, their high crystallinity is useful for the transport of charge carriers [26]. Oligomers of π -conjugated thiophene-based systems represent favorable materials for OPVs and have received more attention when attempting to improve the performance of organic solar cells [27,28].

The power conversion efficiency of OPVs can be increased through bulk heterojunction (BHJ) fabrication by introducing fullerene derivatives as acceptors with conjugated polymer or small conjugated molecules as donors [3]. The corresponding energy level diagrams and architectures of the BHJ device are shown in Fig. 1. In the general case, a BHJ device is constructed in one out of two architectures: the conventional and the inverted one. Through the absorption of sunlight, an electron jumps from the HOMO of the donor to the LUMO of the acceptor leading to the generation of a hole that is delivered at the anode. In parallel, the electron is transported to the cathode [29].

To explore new derivatives of thiophene oligomers with experimental methods is expensive and time-consuming. Instead, we follow here an alternative theoretical approach (i.e., an inverse-design method) to identify optimal oligomers out of a large pool by using various molecular properties in estimating their performance

in solar cell applications. The application of inverse-design methods in optimizing materials has been suggested by others [30–34] even though our implementation differs in several aspects from earlier assumptions. Thus, the performance of the combination of the simple inverse-design approach with the Density-Functional Tight-Binding (DFTB) method for calculating the electronic properties of a given system was initially demonstrated for some Si–Ge mixture clusters and benzene derivatives for solar energy harvesting. It was demonstrated that this method can provide good candidate structures in a short time, which represents a significant improvement over a blind search and can be used on simple computer architecture (desktop or laptop) and does not require large resources. For this reason, it is called PooMa, the Poor Man's Materials Optimization [35,36]. The main purpose of these first studies was to demonstrate the ability of our inverse-design method in optimizing properties.

In the present work, PooMa is used to propose 20 substituted molecular systems based on the thiophene tetramer (Tn) with optimal values for the power conversion efficiency (PCE). A drawback of the theoretical approach is that it does not take the stability or synthesizability into account, but from the larger number of proposed systems it is possible to identify more general features that should be “good” and, hence, be parts of systems that ultimately can be synthesized. In addition, by studying these Tn oligomers through standard density functional theory (DFT) and its derivative time-dependent DFT (TD-DFT) calculations, a further understanding of the geometrical and electronic properties of these Tn oligomers in their ground and excited states can be obtained.

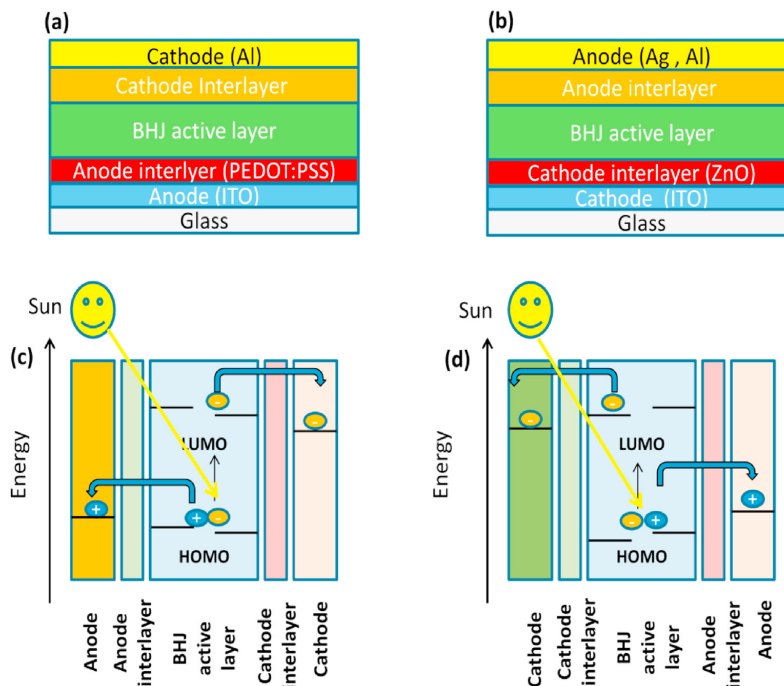


Fig. 1. Schematic description of BHJ device according to (a) conventional architecture and (b) inverted architecture, together with the energy diagrams for (c) conventional and (d) inverted devices.

2. Theoretical details

2.1. Molecules and genetic algorithm

Our home-made PooMa program is designed to search in the space of all possible substitutions of functional groups at fixed sites of a given backbone core. In this work, the tetrathiophene oligomer represents the backbone core with eight sites (one site is fixed for the anchor group) for the substitution of electron pushing and pulling groups. 22 functional groups with different electron-donating and -withdrawing properties were studied with the purpose of optimizing sun-light harvesting. The groups can be separated into the -H group, nine electron-donating groups (-NH₂, -NHCH₃, -OH, -OCH₃, -NHCOCH₃, -OCOCH₃, -C₆H₅, -CH=CH₂, and -CH₃), eight electron-withdrawing groups (-NO₂, -CN, -COCl, -CHO, -COCH₃, -COOH, -COOCH₃, -CONH₂), and four halogen atoms (-F, -Cl, -Br and -I). These groups were chosen because of their small size, which will facilitate the synthesis procedure and lead to more planar geometries in comparison with larger groups.

In addition, as an anchor group [37], -COOH was placed at position 1 (R₁), whereas the other seven sites were made available for all groups (see Fig. 2).

In our case, each molecule can be characterized through a string with $N = 8$ integers, whereby seven of these integers can take a value up to $K = 22$. N is the number of positions at which functional groups can be attached, and K is the number of functional groups. In the calculations, we keep a constant number, P , of molecules. We add that in order to obtain realistic descriptions of the polythiophene polymer, systems with more than 4 monomers are considered necessary [38–43]. However, because the size of the pool of possible systems scales exponentially with the number of monomers, we limited ourselves to studying the tetrathiophene, being convinced that already for this system relevant and interesting information can be obtained.

An important part of our inverse-design method is a genetic algorithm [44,45] that is used in reducing drastically the number of systems to be treated out of the 22⁷ possible ones (the R₁ site was used for the -COOH anchor group) in the optimization process.

In brief, the genetic algorithm consists of the following steps:

1. Initially, $P = 20$ substituted tetrathiophenes systems are created with different sequences of functional groups also when considering symmetry operations. These first 20 substituted systems form the so-called parent molecules (first generation).
2. The structures of the first generation are relaxed locally after which their performance function that shall be optimized is calculated.
3. The $P = 20$ parent molecules are separated randomly into 10 pairs. For each pair, two new children are created via a cutting and mating process whereby each string of N integers is cutting randomly into two parts that interchanged as described in detail in Ref. [36]. Mutations may be considered to replace some

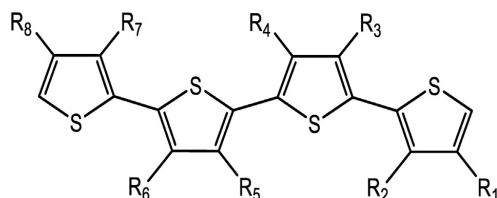


Fig. 2. The backbone of tetrathiophene with the substitution sites (R1-R8).

functional group at a certain site randomly with another functional group, i.e., a randomly chosen integer for one of the children may be changed randomly into another integer except at the site R₁ at which the anchor group -COOH is fixed. All parent and children molecules must be different also after applying symmetry; otherwise, step 3 will be repeated.

4. The structures of the children's molecules are relaxed and their performance function is calculated. Of the 20 parents and 20 children molecules, those 20 with the best performance function are chosen as parents for the next generation.
5. The calculation will be stopped if the new parent molecules are unchanged for ten generations, otherwise go to 3.

2.2. DFTB approach and performance function

During the PooMa calculations, a larger number of systems will be treated so that a computationally efficient approach is needed for studying the individual system, i.e., a method with an acceptable compromise between computational speed and accuracy. To this end we use the non-self-consistent-charge density-functional tight-binding (non-SCC-DFTB) approach of Seifert and coworkers [46,47]. The DFTB method does not consider charge fluctuations and the resulting interactions. Therefore, the computational cost can be reduced while keeping a reasonable accuracy in particular for systems with smaller charge transfers. Also in our earlier studies, we have applied the DFTB method [35,36] and the reader is referred to those for a more detailed description.

The overall power conversion efficiency (PCE) of the OPVs will be quantified through various photovoltaic parameters including the short-circuit current density (J_{sc}), the fill factor (FF), the open-circuit voltage (V_{oc}) and the intensity of the incident light (P_{in}) [48],

$$PCE(\%) = \frac{J_{sc} FF V_{oc}}{P_{in}} \times 100\% \quad (1)$$

These parameters depend on all parts of the cell, i.e., the organic molecules, the nature of the electrodes, the electrolyte, and the working conditions. It is hardly possible to deal with such complex systems efficiently with electronic-structures calculations. Thus, we shall use a recently developed Quantitative-Structure-Property-Relation (QSPR) model for describing the efficiency of the solar cells through properties of individual, isolated molecular systems [49]. According to that model, we calculate the performance function through five parameters as formulated in the QSPR model [49] and this represents the quantity that shall be optimized in the calculations of PooMa.

The five parameters that we use to model the PCE are described below:

The first parameter that we consider is the energy gap G between the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO). A good value of PCE is expected for a small value of the energy gap.

$$G = \epsilon_{LUMO} - \epsilon_{HOMO} \quad (2)$$

The second parameter is the sun-light absorption A which is approximated as the sum of the probabilities $P(\epsilon_a - \epsilon_i)$ for single-electron excitations, weighted with the oscillation strength f_{ia} .

$$A = \frac{\sum_{ia} f_{ia} P(\epsilon_a - \epsilon_i)}{N_e} \quad (3)$$

Here, and below, i and a represent an occupied orbital and an unoccupied orbital, respectively, and N_e is the number of valence electrons of the Tn systems. We divide the properties by the

number of valence electrons to optimize intensive properties and not extensive properties in order to avoid that the largest or smallest molecular systems are predicted to be the best ones. The spectral function $P(\epsilon_a - \epsilon_i)$ is modeled using a black-body approximation:

$$P(\epsilon_a - \epsilon_i) = \frac{8\pi(\epsilon_a - \epsilon_i)^3}{(2\pi\hbar c)^3} \left[\exp\left(\frac{\epsilon_a - \epsilon_i}{k_B T}\right) - 1 \right]^{-1} \quad (4)$$

The oscillation strength f_{ia} is calculated using an approximated TD-DFT (time-dependent density-functional theory) method [50]. Since we only consider one-electron excitations and ignore charge fluctuations, f_{ia} can be simplified as

$$f_{ia} = \frac{4}{3}(\epsilon_a - \epsilon_i) \left| \sum_n \vec{R}_{n,a} \cdot \vec{q}_{n,ia} \right|^2 \quad (5)$$

$\sum_n \vec{R}_{n,a} \cdot \vec{q}_{n,ia}$ is the transition dipole moment in terms of Mulliken transition charges $q_{n,ia}$, and the corresponding coordinates \vec{R}_n of the atom n for the $i \rightarrow a$ transition.

The third parameter is the light-harvesting efficiency (LHE). The larger the LHE, the better is the efficiency,

$$\text{LHE} = \frac{\sum_{i,a} (1 - 10^{-f_{ia}})}{N_e} \quad (6)$$

The fourth parameter that we consider is the orbital overlap average O_{av} which is calculated as the sum of the overlap between each pair of occupied and unoccupied molecular orbitals O_{ia} over all transitions i to a and weighted by the transition probability $f_{ia}P(\epsilon_a - \epsilon_i)$. Thus, from

$$O_{ia} = \sum_n |q_{n,i} \cdot q_{n,a}| \quad (7)$$

with $q_{n,i}$ and $q_{n,a}$ being the atomic gross Mulliken populations of the atom n for the orbitals i and a , we define

$$O_{av} = \frac{\sum_{i,a} f_{ia} P(\epsilon_a - \epsilon_i) O_{ia}}{N_e} \quad (8)$$

Finally, the orbital distance average D_{av} is the fifth parameter that we use. This is calculated as in Eq. (8) but by replacing O_{ia} by single-orbital distances D_{ia} .

$$D_{av} = \frac{\sum_{i,a} f_{ia} P(\epsilon_a - \epsilon_i) D_{ia}}{N_e} \quad (9)$$

with

$$D_{ia} = \left| \sum_n q_{n,i} \vec{R}_n - \sum_n q_{n,a} \vec{R}_n \right| \quad (10)$$

In experimental studies, the systems involve other factors (electrolyte, electrodes) that influence the performances of these systems, too. It is, therefore, an approximation to model the performance through properties of only the isolated dyes. However, for different systems with similar electrolytes and electrodes, we believe that we will be able to describe the performance variation when changing dyes. Therefore, our QSPR model is based on modeling measured performances of 206 molecular systems that mainly differ in the dye but otherwise was very similar. The resulting QSPR model is described in detail in Ref. [49].

2.3. DFT and TD-DFT calculations

The PooMa calculations result in a smaller set of systems with optimal properties. The strategy behind PooMa is to reduce the number of required computations by using the genetic algorithms which allows for a significant reduction in the number of molecules to be studied. However, this number may still be of the order of some few 1000s (in our present work, we considered 1520 molecules). Consequently, only through the combination of the genetic algorithm with the DFTB method, the PooMa approach becomes efficient.

The number of systems considered in the PooMa calculations, i.e., 1520, is clearly much smaller than the total number of possible systems so that it cannot be guaranteed that indeed the "best" ones have been identified. Nevertheless, we are convinced that PooMa provides a set of good systems that it is worthwhile to study in more detail compared to approaches where systems are proposed based on more or less subjective suggestions.

To get more insight into the properties of the obtained Tn systems, additional DFT and TD-DFT calculations are used for these systems and for studying their interactions with fullerene derivatives. The ground state geometry optimization is performed using the B3LYP (Becke 3-parameter Lee-Yang-Parr) exchange-correlation functional [51–53] together with a 6-31G(d,p) basis set. All calculations related to the excited states were done using TD-DFT theory at the CAM-3LYB/6-31G(d,p) level [54]. All calculations were performed in the gas phase with the Gaussian09 software package [55]. The resulting output was analyzed further using the Multiwfn software package [56].

2.4. Charge transfer based on the electron-density difference

To analyze charge-transfer (CT) during the vertical electron transitions we shall use a method proposed earlier [57] for the one-dimensional case and in the Multiwfn package generalized to the case of three dimensions. The change in the electron density between the excited state (Ex) and the ground state (Gs) for fixed geometry is given as

$$\Delta\rho(r) = \rho_{\text{Ex}}(r) - \rho_{\text{Gs}}(r) \quad (11)$$

$\rho_{\text{Gs}}(r)$ and $\rho_{\text{Ex}}(r)$ are the electronic densities associated with the ground and the excited states. $\Delta\rho(r)$ can be written as a sum of positive and negative parts $\rho_+(r)$ and $\rho_-(r)$. The barycenters of these two are given as

$$R_+ = \frac{\int r \rho_+(r) dr}{\int \rho_+(r) dr} \quad (12)$$

$$R_- = \frac{\int r \rho_-(r) dr}{\int \rho_-(r) dr} \quad (13)$$

The charge-transfer excitation length D_{CT} is then defined as the spatial distance between the two barycenters R_+ and R_- :

$$D_{\text{CT}} = |R_+ - R_-| \quad (14)$$

The transferred charge q_{CT} is defined as the integral over all space of $\rho_+(r)$ or $\rho_-(r)$. Analogously, a charge-transfer dipole moment (μ_{CT}) can be defined as:

$$\mu_{\text{CT}} = D_{\text{CT}} \cdot q_{\text{CT}} \quad (15)$$

The value of μ_{CT} is qualitatively related to the difference between the computed dipole moments for the ground and the excited states ($\Delta\mu_{ES-gs}$).

The electron and hole mobilities depend strongly on the internal reorganization energies. The reorganization energy is defined as the sum of geometrical relaxation energies when the species go from the ground state geometry to the charged state geometry, and vice versa. The latter can be defined as [58]:

$$\lambda_e = \lambda_- + \lambda_0 = (E_0^- - E_-) + (E_0^0 - E_0^0) \quad (16)$$

$$\lambda_h = \lambda_+ + \lambda_0 = (E_0^+ - E_+^+) + (E_0^0 - E_0^0) \quad (17)$$

Here, λ_e and λ_h are reorganization energies for electron and hole, respectively; E_0^- (E_0^+) is the energy of anion (cation) calculated for the optimized structure of the neutral molecules, E_0^0 (E_+^0) the energy of the neutral state in the optimized geometry of the anion (cation); and E_0^0 , E_- and E_+^+ the total energies of the neutral, anionic, and cationic species in their equilibrium structures, respectively.

2.5. Molecular frontier orbitals and chemical reactivity parameters

The molecular frontier orbitals are important for the power conversion efficiency of bulk heterojunction (BHJ) solar cells. Generally, a higher value of V_{OC} gives better power efficiency. This voltage in BHJ solar cells can be determined theoretically as the difference between the absolute values of the energies of the HOMO of the donor molecule (Tn) and the LUMO of the electron acceptor (PCBMs) using [59]:

$$V_{OC} = \frac{1}{e} (|E_{HOMO}^{Donor}| - |E_{LUMO}^{Acceptor}|) - 0.3 \text{ V} \quad (18)$$

where e represents the elementary charge and the value of 0.3 V is an empirical term [60].

Another important parameter is the driving force (ΔE_{DV}) required for exciton dissociation. This is determined as the difference between LUMO energy levels of the donor (Tn) and the acceptor (Phenyl-C61-butyric acid methyl derivatives, PCBMs), whereby a larger value (larger than 0.3 eV) suggests a better charge separation [61],

$$E_{DV} = E_{LUMO}^{Donor} - E_{LUMO}^{Acceptor} \quad (19)$$

To evaluate the ability of our systems (Tn) and PCBMs molecules to donate or accept electrons, we calculated various parameters, i.e., the chemical hardness (η), the electronegativity (χ), and the electrophilicity power (ω) according to Refs. [62,63]:

$$\eta = \frac{IP - EA}{2} \quad (20)$$

$$\chi = \frac{IP + EA}{2} \quad (21)$$

$$\omega = \frac{\chi^2}{2\eta} \quad (22)$$

where IP and EA represent the adiabatic ionization potential and the adiabatic electron affinity, respectively,

$$IP = E_+^+ - E_0^0 \quad (23)$$

$$EA = E_0^0 - E_- \quad (24)$$

The quantities in Eqs. (23) and (24) were defined above.

2.6. Optical properties

To understand the electronic transitions and emissions, TD-DFT/CAM-B3LYP was applied for the optimized structures to obtain the vertical excitation energies and their corresponding oscillation strengths. The combination of B3LYP for the structure optimization and CAM-B3LYP for the excitation properties has been found in other studies to provide a good choice [64–66]. To study the fluorescence properties of the systems, we performed also TD-DFT/CAM-B3LYP calculations for the geometry optimization of the lowest singlet excited state. The Stokes shift of the systems is calculated as the difference between the maximum wavelengths of emission and excitation.

Furthermore, the radiative lifetimes (τ) were calculated for the spontaneous emission by using the Einstein transition probabilities according to Ref. [67]:

$$\tau = \frac{c^3}{2(E_{Flu})^3 f} \quad (25)$$

where c is the speed of light, E_{Flu} is the fluorescence energy, and f the oscillation strength.

Using Koopmans' theorem, the oxidation potential energies of the systems in the ground state E^{OX} was estimated as the negative of the HOMO energy [68], while the oxidation potential of the excited state can be estimated according to Ref. [69]:

$$E^{OX*} = E^{OX} - E^{00} \quad (26)$$

E^{00} represents the vertical transition energy corresponding to the maximum wavelength in the absorption spectrum.

3. Results and discussion

3.1. Structures of the identified thiophene oligomers

For the thiophene tetramer of Fig. 2, the PooMa calculations resulted in 20 structures with optimal properties and differing in the functional groups attached to the backbone. These are depicted in Fig. 3. The different systems are ordered according to the efficiency as given by the QSPR model. In all cases, the HOMO is localized to the three rings with the sites R1–R6, whereas the LUMO is localized around the two rings with the sites R5–R8. Most of the molecules contain $-\text{OCH}_3$, $-\text{H}$, $-\text{OH}$, or $-\text{NH}_2$ in addition to the $-\text{COOH}$ anchor group as electron donor and $-\text{COOH}$, $-\text{OCH}_3$ as an electron acceptor.

3.2. Geometric properties and frontier orbitals

To have a further theoretical understanding of the molecules obtained in the PooMa calculations, more accurate methods (DFT and TD-DFT) were used. The optimized geometries of the molecules are shown in Appendix A. Fig. A3. Table A1 (Appendix A.) lists some geometrical parameters of the proposed molecules (i.e., the bridge bond lengths and the dihedral angles between thiophene units) in the ground and the excited states. The lengths of the bridge bonds B2 and B3 are almost unchanged for all molecules and lie in the range 1.44–1.45 Å for the ground state (S_0) and are slightly smaller for the excited state (S_1). The length of the bridge bond B1 for the S_0 and S_1 states are in the range 1.46–1.47 and 1.41–1.42 Å,

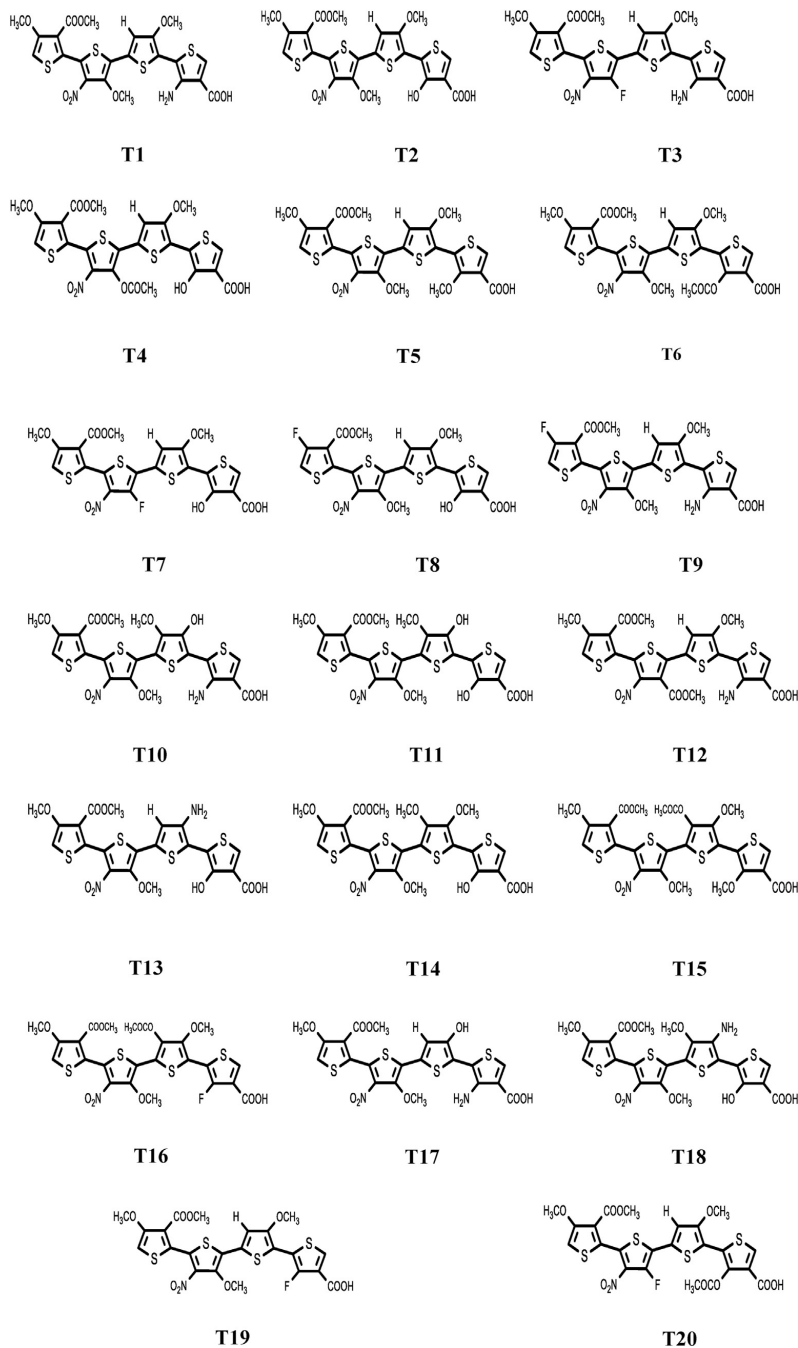


Fig. 3. The molecular structures of the best performing Tn systems obtained from the PooMa calculations. The performance decreases from T1 to T20.

respectively. Accordingly, the bridge bonds possess some C=C character in the excited state.

The dihedral angles TA1, TA2, and TA3 are listed in Appendix A. Table A2 give information on the coplanarity of the studied molecules. It is seen that for the ground state, the molecules are very close to being planar for most of the units except those containing the R7 and R8 groups. Upon excitation, a smaller increase in the coplanarity is observed, which may result in a faster transfer of photo-induced electrons from S_0 to S_1 .

As can be observed in Fig. 4 and Figs. A4 and A5 (Appendix A), the HOMO has some anti-bonding character between the oligomer units. Further, an analysis of the frontier orbitals reveals a strong density localization on the terminal motifs including the thiophene units with the R5-R8 groups for the LUMO, while the HOMO level is spread over the thiophene units with R1-R6 groups. The HOMO and LUMO energies and the bandgap E_g of the proposed compounds are listed in Appendix A. Table A2. It can be seen that the T20, T3, T4, and T7 possess the highest energy gaps, while the lowest value is found for T5, which may lead to larger values for V_{oc} and J_{sc} .

The theoretically obtained maximum open-circuit (V_{oc}) and driving force (E_{DV}) of the complexes of the donor molecules and the PCBM: PCBM-71 ([6,6]-Phenyl-C₇₁-butyric acid methyl ester), PCBM ([6,6]-Phenyl-C₆₁-butyric acid methyl ester) and C60 ((C60-Ih) [5,6]fullerene) are determined using Eqs. (18) and (19), respectively. The best alignment of the HOMO and LUMO levels of the acceptor with the levels of the proposed donors was found for C60 (See Appendix A. Table A2), whereas all LUMO energy levels of Tn are higher than that of C60. The latter gave the largest value for the driving forces (higher than 0.3 eV) out of the other two acceptors, which ensures an efficient electron transfer from the donor to the acceptor. Some Tn complexes with PCBM-71 show some negative values of E_{DV} , which makes the transfer of electrons from these compounds to the LUMO of PCBM-71 very difficult. The theoretical values for V_{oc} of the Tn systems with all acceptors suggest a possible efficient electron injection from donor to acceptor. Therefore, devices constructed from Tn and PCBM with the values of E_{DV} higher than 0.3 eV can exhibit a high efficient injection of electrons and improve the overall performance of bulk-heterojunction cells.(see in Table 1)

3.3. Charge transport properties and quantum chemical parameters

The charge transport properties for the $S_0 \rightarrow S_1$ excitation (q_{CT} , μ_{CT} , and D_{CT}) are presented in Table 2. The values of the transferred charges (q_{CT}) fall in a narrow range of 0.290–0.343 a.u. Generally, a higher value of q_{CT} suggests a larger exciton separation and thus an increased J_{sc} . The small D_{CT} values show that the charge transfer is over a shorter distance in the Tn systems. Together with q_{CT} , this leads to no significant variation in the dipole moment μ_{CT} due to the excitation.

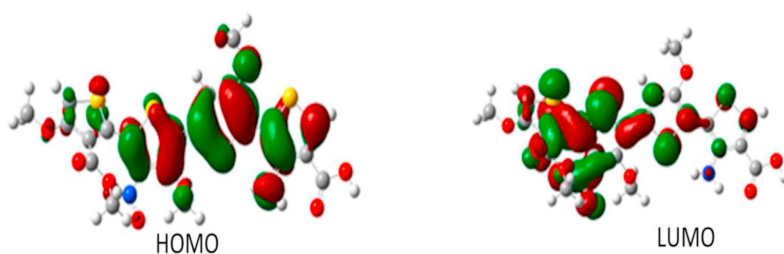


Fig. 4. Plots of the HOMO and LUMO orbitals of the suggested T1 system.

Table 1
Theoretical open-circuit voltage (V_{oc}) in V and driving force (E_{DV}) in eV of Tn systems T1-T20 with PCBM.

Tn Systems	PCBM-71		PCBM		C60	
	V_{oc}	E_{DV}	V_{oc}	E_{DV}	V_{oc}	E_{DV}
T1	1.647	-0.117	1.400	0.130	1.110	0.420
T2	1.641	-0.085	1.394	0.162	1.105	0.452
T3	2.393	0.143	2.146	0.390	1.857	0.679
T4	2.266	0.270	2.019	0.517	1.729	0.807
T5	1.276	-0.030	1.029	0.217	0.739	0.506
T6	1.780	-0.132	1.533	0.114	1.243	0.404
T7	2.384	0.130	2.137	0.377	1.848	0.667
T8	1.834	-0.204	1.587	0.044	1.297	0.333
T9	1.818	-0.186	1.570	0.061	1.281	0.351
T10	1.534	0.135	1.287	0.382	0.997	0.671
T11	1.529	0.124	1.282	0.371	0.993	0.660
T12	1.912	0.231	1.665	0.478	1.376	0.768
T13	1.673	-0.046	1.426	0.201	1.136	0.491
T14	1.645	-0.001	1.398	0.246	1.108	0.536
T15	1.279	0.067	1.031	0.314	0.741	0.604
T16	1.704	-0.045	1.457	0.202	1.168	0.492
T17	1.677	-0.045	1.430	0.202	1.141	0.491
T18	1.500	0.133	1.253	0.380	0.963	0.670
T19	1.772	-0.126	1.526	0.121	1.236	0.411
T20	2.453	0.085	2.206	0.332	1.917	0.622

The calculated reorganization energies for the hole and electron are tabulated in Table 2. It is well-known that lower values of λ_h and λ_e lead to higher charge transfer rates [70]. The smallest values of λ_h and λ_e are found for T8, which we, therefore, consider as the best hole and electron transport molecule of our Tn systems. Besides, the difference between the λ_h and λ_e for T16 is only 0.02 eV, implying that T16 has better equilibrium properties for hole and electron properties. Thus, T16 may be utilized as a good candidate molecule for ambipolar charge transport materials.

Generally, it may be suggested that molecules with a large dipole moment μ can easily interact with other molecules. From Table 2 we recognize that the highest values of μ are found for T8 and T7, suggesting that these compounds easier can interact with the PCBM. On the other hand, the Tn systems have the lowest values of the hardness, η compared to PCBM with some exceptions for PCBM and PCBM-71 (see Table 2). Since η quantifies the resistance to changes in the number of electrons, the Tn systems act rather as electron donors and PCBM derivatives as electron acceptors. The electronegativity and electrophilicity values of the PCBM are higher than those of the Tn systems, implying that the PCBM will attract electrons from the Tn systems (see Table 2).

3.4. Absorption and fluorescence properties

With the help of TD-DFT/CAM calculations, the electronic transition energies of our π -conjugated molecules can be obtained

Table 2

The charge transport properties (q_{CT} /a.u., μ_{CT} /Debye, D_{CT} /Å, λ_h , and λ_e /eV) of the Tn systems and the quantum chemical parameters (χ /eV, x /eV, and x /eV and μ /Debye) for the Tn systems and the PCBM.

Tn systems	q_{CT}	μ_{CT}	D_{CT}	λ_h	λ_e	η	x	ω	μ
T1	0.309	0.369	0.248	0.53	0.73	2.15	3.65	3.09	5.09
T2	0.303	0.134	0.092	0.46	0.78	2.31	3.73	3.00	5.69
T3	0.296	0.408	0.287	0.45	0.62	2.18	3.78	3.29	7.77
T4	0.319	0.127	0.083	0.44	0.64	2.24	3.80	3.21	7.83
T5	0.311	0.205	0.137	0.59	0.73	2.26	3.81	3.21	5.69
T6	0.324	0.196	0.126	0.51	0.73	2.29	3.92	3.36	5.19
T7	0.290	0.193	0.138	0.41	0.60	2.24	3.82	3.26	8.23
T8	0.291	0.290	0.207	0.39	0.57	2.24	3.88	3.35	8.74
T9	0.297	0.432	0.303	0.43	0.59	2.15	3.84	3.42	7.50
T10	0.314	0.084	0.056	0.72	0.61	2.39	3.65	2.78	6.08
T11	0.311	0.141	0.095	0.59	0.60	2.31	3.70	2.96	7.56
T12	0.320	0.235	0.153	0.58	0.77	2.16	3.71	3.19	5.46
T13	0.297	0.085	0.059	0.50	0.60	2.26	3.70	3.03	6.30
T14	0.322	0.106	0.069	0.61	0.76	2.27	3.58	2.82	5.48
T15	0.343	0.042	0.025	0.71	0.59	2.34	3.86	3.19	7.48
T16	0.327	0.034	0.022	0.66	0.64	2.35	3.99	3.38	7.77
T17	0.298	0.118	0.083	0.70	0.67	2.27	3.82	3.22	4.74
T18	0.313	0.111	0.074	0.63	0.75	2.31	3.62	2.83	6.24
T19	0.297	0.132	0.093	0.49	0.65	2.27	3.93	3.40	7.74
T20	0.311	0.229	0.154	0.44	0.58	2.30	4.01	3.50	7.80
PCBM-71	***	***	***	***	***	2.28	4.34	4.13	4.04
PCBM	***	***	***	***	***	2.35	4.38	4.07	3.01
C60	***	***	***	***	***	2.52	5.05	5.06	0.00

accurately [71]. Therefore, we used it for our systems and present the results in Table 3, i.e., the maximum absorption wavelengths (λ_{max}), oscillation strengths (f), vertical excitation energies (E_{ex}), and the major contribution of the molecular orbitals (MOs).

According to Table 3, the vertical excitations of the Tn systems are in the range 2.74–3.45 eV with the lowest (best) values for T9 and T8. It is clear that the lowest-lying transition can be varied by chosen appropriate functional groups. We add that the lowest excitation of the Tn systems has a higher oscillation strength compared to other energetically low excitations. Moreover, the $S_0 \rightarrow S_1$ excitation has its dominating contribution from the HOMO \rightarrow LUMO excitation.

Next, we studied the fluorescence properties of Tn systems by using the TDDFT/CAM-B3LYP method. We present in Table 4 the maximum wavelengths (λ_{em}), the fluorescence energies of the lowest excitation (E_{Flu}), the oscillation strengths (f), the involved

molecular orbitals (MOs), the radiative lifetimes (τ), and the Stokes shifts (SS) for the Tn systems. As for the absorption, also the $S_1 \rightarrow S_0$ emission of the Tn systems can be assigned to $\pi^* \rightarrow \pi$ and LUMO \rightarrow HOMO transitions. The Stokes shifts of T2, T18, and T14 are larger than those of the other Tn systems, indicating that these compounds have a larger structural change between the ground and the excited states. The compounds with smaller Stokes shifts delay the injection of electrons from the HOMO of the donor to the LUMO of PCBM. We also see that the radiative lifetimes lie in the range 7–21 ns whereby a large value delays charge recombination and enhance the efficiency of solar cells.

As shown in Fig. 5, the oxidation potentials of the excited states of the Tn systems fall in the range 1.49–2.87 eV. The most oxidizing molecules (i.e., the lowest oxidation potentials in the excited states) are T15, T5, and T10, which implies that these systems have the best electron-donation properties.

Table 3

The calculated maximum absorption wavelengths (λ_{max} /nm), oscillation strengths (f), $S_0 \rightarrow S_1$ vertical excitation energies (E_{ex} /eV), and the contribution of the molecular orbitals (MOs) involved in the excitation of the Tn systems (T1–T20).

Tn systems	λ_{max}	f	E_{ex}	The major contribution of MOs
T1	425	0.76	2.92	HOMO \rightarrow LUMO (78%)
T2	405	0.76	3.10	HOMO \rightarrow LUMO (78%)
T3	430	0.77	2.90	HOMO \rightarrow LUMO (79%)
T4	408	0.78	3.04	HOMO \rightarrow LUMO (75%)
T5	399	0.83	3.11	HOMO \rightarrow LUMO (83%)
T6	392	0.85	3.20	HOMO \rightarrow LUMO (85%)
T7	414	0.80	2.99	HOMO \rightarrow LUMO (81%)
T8	436	0.71	2.84	HOMO \rightarrow LUMO (85%)
T9	452	0.68	2.74	HOMO \rightarrow LUMO (84%)
T10	370	0.71	3.45	HOMO \rightarrow LUMO (76%)
T11	385	0.85	3.22	HOMO \rightarrow LUMO (88%)
T12	400	0.73	3.10	HOMO \rightarrow LUMO (66%)
T13	405	0.79	3.06	HOMO \rightarrow LUMO (78%)
T14	389	0.67	3.18	HOMO \rightarrow LUMO (74%)
T15	395	0.71	3.14	HOMO \rightarrow LUMO (84%)
T16	380	0.66	3.26	HOMO \rightarrow LUMO (82%)
T17	385	0.58	3.22	HOMO \rightarrow LUMO (69%)
T18	390	0.67	3.18	HOMO \rightarrow LUMO (71%)
T19	401	0.71	3.09	HOMO \rightarrow LUMO (81%)
T20	399	0.83	3.11	HOMO \rightarrow LUMO (81%)

Table 4

Emission spectral data for Tn systems. E_{Flu} /eV of $S_1 \rightarrow S_0$, λ_{em} /nm, f , MOs contribution, τ /ns, and SS/nm.

Tn systems	λ_{em}	E_{Flu}	f	MOs major contribution	τ	SS
T1	634	1.96	0.32	LUMO \rightarrow HOMO (90.1%)	19	209
T2	656	1.89	0.41	LUMO \rightarrow HOMO (96.5%)	16	251
T3	584	2.12	0.55	LUMO \rightarrow HOMO (90.5%)	9	154
T4	582	2.13	0.42	LUMO \rightarrow HOMO (89.5%)	12	174
T5	578	2.14	0.47	LUMO \rightarrow HOMO (89.9%)	11	179
T6	569	2.18	0.48	LUMO \rightarrow HOMO (89.6%)	10	177
T7	561	2.21	0.56	LUMO \rightarrow HOMO (90.4%)	8	147
T8	602	2.06	0.43	LUMO \rightarrow HOMO (90.8%)	13	166
T9	628	1.98	0.42	LUMO \rightarrow HOMO (91.0%)	14	176
T10	575	2.16	0.39	LUMO \rightarrow HOMO (89.4%)	13	205
T11	575	2.16	0.43	LUMO \rightarrow HOMO (89.1%)	12	190
T12	568	2.18	0.45	LUMO \rightarrow HOMO (88.4%)	11	168
T13	585	2.12	0.45	LUMO \rightarrow HOMO (89.1%)	12	180
T14	635	1.95	0.29	LUMO \rightarrow HOMO (90.8%)	21	246
T15	558	2.22	0.43	LUMO \rightarrow HOMO (90.4%)	11	163
T16	550	2.25	0.44	LUMO \rightarrow HOMO (91.2%)	10	170
T17	577	2.15	0.46	LUMO \rightarrow HOMO (87.2%)	11	192
T18	640	1.94	0.30	LUMO \rightarrow HOMO (88.7%)	20	250
T19	567	2.19	0.48	LUMO \rightarrow HOMO (90.0%)	10	166
T20	542	2.29	0.60	LUMO \rightarrow HOMO (89.8%)	7	143

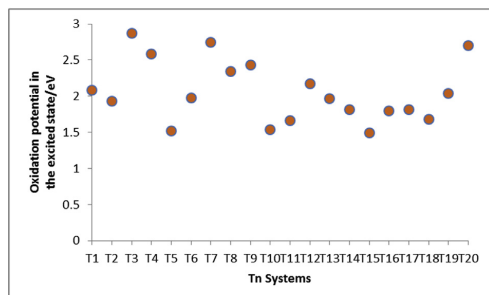


Fig. 5. Oxidation potential in the excited states of the 20 Tn systems.

3.5. Tn/PCBM power conversion efficiency

The photovoltaic performances of the Tn donor systems can be estimated theoretically from the relationship between their LUMO energy levels and bandgaps using the Scharber diagram [59]. The Scharber model has been used to predict the efficiency of organic solar cells combining PCBM acceptor with polymers or small molecular donor [72–74]. Applying this to our Tn donor systems together with the PCBM acceptor gives the results shown in Fig. 6. Then, according to the Scharber model, the suggested molecules T8 and T9 are expected to have the highest conversion efficiencies of 4% and 3% respectively.

4. Conclusions

The main result of this work is the proposal of twenty branched oligothiophene molecules (Tn) as candidates for optimal performance in solar-energy harvesting. These were identified using a recently developed inverse-design approach (PooMa). Each molecule is based on four thiophene units with seven sites to which various functional groups could be attached beside one site fixed for the anchor group (-COOH). Twenty-two substituents were considered. For identifying the best solar cell performances, we used a QSPR model based on five molecular properties (G , A , LHE, O_{av} , and D_{av}). PooMa is combined with a fast and efficient electronic-structure method (DFTB). Although our calculations did not take into account whether the proposed systems could be synthesized, we believe that our approach provides a qualified

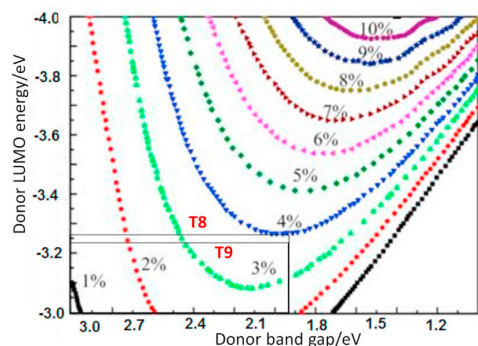


Fig. 6. Scharber diagram to estimate conversion efficiencies of the Tn/PCBM solar cells.

suggestion for optimal systems and that if they themselves cannot be produced, it would be possible to synthesize molecules related to our proposed ones.

In order to obtain further insight into the proposed systems, we applied ab initio DFT and its derivative TD-DFT. The predicted band gaps of the Tn systems lie in the range of 1.55–2.84 eV. The bridge bonds between neighboring thiophene rings undergo a slight decrease when passing from the S_0 to the S_1 state, which increases the electronic interactions between the thiophene units. The theoretical values of V_{oc} of the Tn systems combined with PCBM-7 are higher than the values found for PCBM and C60. A higher LUMO energy of Tn gives a better driving force (E_{dv}) when combined with PCBM derivatives. The values of the quantum chemical parameters show that the Tn systems are more prone to donate electrons to PCBMs. The values of the reorganization energies show that T8 is a good hole and electron transporting material. The calculated absorption wavelengths fall in the range 370–452 nm, whereas the emission wavelengths are in the range 542–656 nm. Finally, with the help of the Scharber diagram, we could identify higher power conversion efficiencies for T8 and T9, i.e., 4% and 3% respectively.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The corresponding author (Dr. Abdullah S. Khazaal) would like to thank the German Academic Exchange Service (DAAD), Bonn, Germany, for funding his research stay for university academics and scientists, 2018 (funding ID: 57378441) at the theoretical and physical chemistry department, University of Saarland, Germany.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.cocom.2020.e00503>.

References

- [1] S. Yanagida, G.K.R. Senadeera, K. Nakamura, T. Kitamura, Y. Wada, Polythiophene-sensitized TiO₂ solar cells, *J. Photochem. Photobiol. Chem.* 166 (2004) 75–80, <https://doi.org/10.1016/j.jphotochem.2004.04.039>.
- [2] A. Hayakawa, O. Yoshikawa, T. Fujieda, K. Uehara, S. Yoshikawa, High performance polythiophene/fullerene bulk-heterojunction solar cell with a TiO_x hole blocking layer, *Appl. Phys. Lett.* 90 (2007), 163517, <https://doi.org/10.1063/1.2730746>.
- [3] G. Denler, M.C. Scharber, C.J. Brabec, Polymer-fullerene bulk-heterojunction solar cells, *Adv. Mater.* 21 (2009) 1323–1338, <https://doi.org/10.1002/adma.200801283>.
- [4] S. Petzold, C. Wang, A. Khazaal, T. Osswald, Conjugated polymer photovoltaic solar cells: manufacturing and increasing performance, *Plast. Eng.* 66 (2010) 26–32, <https://doi.org/10.1002/j.1941-9635.2010.tb00589.x>.
- [5] T. Xu, L. Yu, How to design low bandgap polymers for highly efficient organic solar cells, *Mater. Today* 17 (2014) 11–15, <https://doi.org/10.1016/j.mattod.2013.12.005>.
- [6] Q. Fan, W. Su, X. Guo, B. Guo, W. Li, Y. Zhang, K. Wang, M. Zhang, Y. Li, A new polythiophene derivative for high efficiency polymer solar cells with PCE over 9%, *Adv. Energy Mater.* 6 (2016), 1600430 <https://doi.org/10.1002/aenm.201600430>.
- [7] Y. Qin, M.A. Uddin, Y. Chen, B. Jang, K. Zhao, Z. Zheng, R. Yu, T.J. Shin, H.Y. Woo, J. Hou, Highly efficient fullerene-free polymer solar cells fabricated with polythiophene derivative, *Adv. Mater.* 28 (2016) 9416–9422, <https://doi.org/10.1002/adma.201601803>.
- [8] M.N. Siddiqui, M. Mansha, U. Mehmood, N. Ullah, A.F. Al-Betar, A.A. Al-Saadi, Synthesis and characterization of functionalized polythiophene for polymer-sensitized solar cell, *Dyes Pigments* 141 (2017) 406–412, <https://doi.org/10.1016/j.dyepig.2017.02.041>.
- [9] A.S. Khazaal, Polymer nanocomposites based on graphene and titanium oxide for supercapacitors, *Asian J. Mater. Chem.* 3 (2018) 1–7, <https://doi.org/>

- 10.14233/ajmc.2018.ajmc-p56.
- [10] S.T. Mahmood, A.S. Khazaal, L.H. Alwaan, Preparation of polymer nano-composites and their application as supercapacitors, *Int. J. Nanoelectron. Mater.* 13 (2020) 169–180.
- [11] A.G. MacDiarmid, “Synthetic metals”: a novel role for organic polymers (Nobel lecture), *Angew. Chem. Int. Ed.* 40 (2001) 2581–2590, [https://doi.org/10.1002/1521-3773\(20010716\)40:14<2581::aid-anie2581>3.0.co;2-2](https://doi.org/10.1002/1521-3773(20010716)40:14<2581::aid-anie2581>3.0.co;2-2).
- [12] H. Shirakawa, Nobel lecture: the discovery of polyacetylene film—the dawning of an era of conducting polymers, *Rev. Mod. Phys.* 73 (2001) 713–718, <https://doi.org/10.1103/RevModPhys.73.713>.
- [13] A.J. Heeger, Semiconducting and Metallic polymers: the fourth generation of polymeric materials (Nobel lecture), *Angew. Chem. Int. Ed.* 40 (2001) 2591–2611, [https://doi.org/10.1002/1521-3773\(20010716\)40:14<2591::AID-ANIE2591>3.0.CO;2-0](https://doi.org/10.1002/1521-3773(20010716)40:14<2591::AID-ANIE2591>3.0.CO;2-0).
- [14] A. Dkhissi, F. Ouhib, A. Chaalane, R.C. Hiorns, C. Dagron-Lartigau, P. Iratçabal, J. Desbrières, C. Pouchan, Theoretical and experimental study of low band gap polymers for organic solar cells, *Phys. Chem. Chem. Phys.* 14 (2012) 5613–5619, <https://doi.org/10.1039/C2CP40170C>.
- [15] Y.Y. Smolin, S. Nejati, M. Bavarian, D. Lee, K.K.S. Lau, M. Soroush, Effects of polymer chemistry on polymer-electrolyte dye sensitized solar cell performance: a theoretical and experimental investigation, *J. Power Sources* 274 (2015) 156–164, <https://doi.org/10.1016/j.jpowsour.2014.10.028>.
- [16] Y.Y. Smolin, S. Janakiramam, M. Soroush, K.K.S. Lau, Experimental and theoretical investigation of dye sensitized solar cells integrated with crosslinked poly (vinylpyrrolidone) polymer electrolyte using initiated chemical vapor deposition, *Thin Solid Films* 635 (2017) 9–16, <https://doi.org/10.1016/j.tsf.2016.12.034>.
- [17] J.-L. Reymond, M. Awale, Exploring chemical space for drug discovery using the chemical universe database, *ACS Chem. Neurosci.* 3 (2012) 649–657, <https://doi.org/10.1021/cn300042z>.
- [18] J. Greeley, T.F. Jaramillo, J. Bonde, I.B. Chorkendorff, J.K. Nørskov, Computational high-throughput screening of electrocatalytic materials for hydrogen evolution, *Nat. Mater.* 5 (2006) 909–913, <https://doi.org/10.1038/nmat1752>.
- [19] A.L. Brisenot, T.W. Holcombe, A.I. Boukai, E.C. Garnett, S.W. Shelton, J.M. Fréchet, P. Yang, Oligo- and polythiophene/ZnO hybrid nanowire solar cells, *Nano Lett.* 10 (2010) 334–340, <https://doi.org/10.1021/nl9036752>.
- [20] T.L. Nguyen, T.H. Lee, B. Gautam, S.Y. Park, K. Gundogdu, J.Y. Kim, H.Y. Woo, Single component organic solar cells based on oligothiophene-fullerene conjugate, *Adv. Funct. Mater.* 27 (2017), 1702474, <https://doi.org/10.1002/adfm.201702474>.
- [21] G. Gigli, G. Barbarella, L. Favaretto, F. Cacialli, R. Cingolani, High-efficiency oligothiophene-based light-emitting diodes, *Appl. Phys. Lett.* 75 (1999) 439–441, <https://doi.org/10.1063/1.124403>.
- [22] F. Mariano, M. Mazzeo, Y. Duan, G. Barbarella, L. Favaretto, S. Carallo, R. Cingolani, G. Gigli, Very low voltage and stable p-i-n organic light-emitting diodes using a linear S, S-dioxide oligothiophene as emitting layer, *Appl. Phys. Lett.* 94 (2009), 63510, <https://doi.org/10.1063/1.3072798>.
- [23] G. Horowitz, F. Garnier, A. Yassar, R. Hajlaoui, F. Kouki, Field-effect transistor made with a sexithiophene single crystal, *Adv. Mater.* 8 (1996) 52–54, <https://doi.org/10.1002/adma.19960080109>.
- [24] F. Garnier, G. Horowitz, D. Fichou, A. Yassar, Molecular order in organic-based field-effect transistors, *Synth. Met.* 81 (1996) 163–171, [https://doi.org/10.1016/s0379-6779\(96\)03761-7](https://doi.org/10.1016/s0379-6779(96)03761-7).
- [25] R.P. Ortiz, J. Casado, V. Hernández, J.T.L. Navarrete, J.A. Letizia, M.A. Ratner, A. Facchetti, T.J. Marks, Thiophene-diazine molecular semiconductors: synthesis, structural, electrochemical, optical, and electronic structural properties; implementation in organic field-effect transistors, *Chem. Eur. J.* 15 (2009) 5023–5039, <https://doi.org/10.1002/chem.200802424>.
- [26] F. Zhang, D. Wu, Y. Xu, X. Feng, Thiophene-based conjugated oligomers for organic solar cells, *J. Mater. Chem.* 21 (2011) 17590–17600, <https://doi.org/10.1039/C1JM12801A>.
- [27] N. Noma, T. Tsuzuki, Y. Shirota, *z*-Thiophene octamer as a new class of photoactive material for photoelectrical conversion, *Adv. Mater.* 7 (1995) 647–648, <https://doi.org/10.1002/adma.19950070709>.
- [28] A. Mishra, C.-Q. Ma, P. Bäuerle, Functional oligothiophenes: molecular design for multidimensional nanoarchitectures and their applications, *Chem. Rev.* 109 (2009) 1141–1276, <https://doi.org/10.1021/cr800422z>.
- [29] S.M. McAfee, J.M. Topple, I.G. Hill, G.C. Welch, Key components to the recent performance increases of solution processed non-fullerene small molecule acceptors, *J. Mater. Chem. A* 3 (2015) 16393–16408, <https://doi.org/10.1039/C5TA04310G>.
- [30] T. Weymuth, M. Reiher, Inverse quantum chemistry: concepts and strategies for rational compound design, *Int. J. Quant. Chem.* 114 (2014) 823–837, <https://doi.org/10.1002/qua.24687>.
- [31] D. Xiao, I. Warnke, J. Bedford, V.S. Batista, Inverse molecular design for materials discovery, in: M. Springborg, J.-O. Joswig (Eds.), *Chemical Modelling*, vol. 10, Royal Society of Chemistry Cambridge, UK, 2013, pp. 1–31, <https://doi.org/10.1039/9781849737241-00001>.
- [32] A. Franceschetti, A. Zunger, The inverse band-structure problem of finding an atomic configuration with given electronic properties, *Nature* 402 (1999) 60–63, <https://doi.org/10.1038/46995>.
- [33] M. D’Avezac, J.-W. Luo, T. Chanier, A. Zunger, Genetic-algorithm discovery of a direct-gap and optically allowed superstructure from indirect-gap Si and Ge semiconductors, *Phys. Rev. Lett.* 108 (2012), 27401, <https://doi.org/10.1103/PhysRevLett.108.027401>.
- [34] A. Zakutayev, X. Zhang, A. Nagaraja, L. Yu, S. Lany, T.O. Mason, D.S. Ginley, A. Zunger, Theoretical prediction and experimental realization of new stable inorganic materials using the inverse design approach, *J. Am. Chem. Soc.* 135 (2013) 10048–10054, <https://doi.org/10.1021/ja311599g>.
- [35] M. Springborg, S. Kohaut, Y. Dong, K. Huwig, Mixed Si-Ge clusters, solar-energy harvesting, and inverse-design methods, *Comput. Theor. Chem.* 1107 (2017) 14–22, <https://doi.org/10.1016/j.comptc.2016.11.020>.
- [36] K. Huwig, C. Fan, M. Springborg, From properties to materials: an efficient and simple approach, *J. Chem. Phys.* 147 (2017), 234105, <https://doi.org/10.1063/1.5009548>.
- [37] J. Liu, T. Tanaka, K. Sivula, A.P. Alivisatos, J.M.J. Fréchet, Employing end-functional polythiophene to control the morphology of nanocrystal–polymer composites in hybrid solar cells, *J. Am. Chem. Soc.* 126 (2004) 6550–6551, <https://doi.org/10.1021/ja0489184>.
- [38] S. Bibi, H. Ullah, S.M. Ahmad, A.A. Shah, S. Bilal, A.A. Tahir, K. Ayub, Molecular and electronic structure elucidation of polypyrrole gas sensors, *J. Phys. Chem. C* 119 (2015) 15994–16003, <https://doi.org/10.1021/acs.jpcc.5b03242>.
- [39] H. Ullah, A.A. Shah, S. Bilal, K. Ayub, Doping and dedoping processes of polypyrrole: DFT study with hybrid functionals, *J. Phys. Chem. C* 118 (2014) 17819–17830, <https://doi.org/10.1021/jp505626d>.
- [40] H. Ullah, A.A. Shah, S. Bilal, K. Ayub, DFT study of polyaniline NH₃, CO₂, and CO gas sensors: comparison with recent experimental data, *J. Phys. Chem. C* 117 (2013) 23701–23711, <https://doi.org/10.1021/jp407132c>.
- [41] H. Ullah, K. Ayub, Z. Ullah, M. Hanif, R. Nawaz, A.A. Shah, S. Bilal, Theoretical insight of polypyrrole ammonia gas sensor, *Synth. Met.* 172 (2013) 14–20, <https://doi.org/10.1016/j.synthmet.2013.03.021>.
- [42] S.M. Ahmad, S. Bibi, S. Bilal, A.A. Shah, K. Ayub, Spectral and electronic properties of π -conjugated oligomers and polymers of poly(*o*-chloroaniline-*co*-*o*-toluidine) calculated with density functional theory, *Synth. Met.* 205 (2015) 153–163, <https://doi.org/10.1016/j.synthmet.2015.04.005>.
- [43] H. Ullah, A.A. Shah, K. Ayub, S. Bilal, Density functional theory study of poly(*o*-phenylenediamine) oligomers, *J. Phys. Chem. C* 117 (2013) 4069–4078, <https://doi.org/10.1021/jp311526u>.
- [44] J.H. Holland, *Adaptation in Natural and Artificial Systems: an Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*, MIT press, 1992.
- [45] J. Devillers, *Genetic Algorithms in Molecular Modeling*, first ed., Academic Press, London, 1996.
- [46] D. Porezag, T. Frauenheim, T. Köhler, G. Seifert, R. Kaschner, Construction of tight-binding-like potentials on the basis of density-functional theory: application to carbon, *Phys. Rev. B* 51 (1995) 12947–12957, <https://doi.org/10.1103/physrevb.51.12947>.
- [47] G. Seifert, D. Porezag, T. Frauenheim, Calculations of molecules, clusters, and solids with a simplified LCAO-DFT-LDA scheme, *Int. J. Quant. Chem.* 58 (1996) 185–192, [https://doi.org/10.1002/\(sici\)1097-461x\(1996\)58:2<185::aid-qua7>3.0.co;2-u](https://doi.org/10.1002/(sici)1097-461x(1996)58:2<185::aid-qua7>3.0.co;2-u).
- [48] M. Grätzel, Recent advances in sensitized mesoscopic solar cells, *Acc. Chem. Res.* 42 (2009) 1788–1798, <https://doi.org/10.1021/ar900141y>.
- [49] C. Fan, M. Springborg, Y. Feng, Application of an inverse-design method to optimizing porphyrins in dye-sensitized solar cells, *Phys. Chem. Chem. Phys.* 21 (2019) 5834–5844, <https://doi.org/10.1039/C8CP07722C>.
- [50] T.A. Niehaus, S. Suhai, F. Della Sala, P. Lugli, M. Elstner, G. Seifert, T. Frauenheim, Tight-binding approach to time-dependent density-functional response theory, *Phys. Rev. B* 63 (2001), 85108, <https://doi.org/10.1103/physrevb.63.085108>.
- [51] C. Lee, W. Yang, R.G. Parr, Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density, *Phys. Rev. B* 37 (1988) 785–789, <https://doi.org/10.1103/physrevb.37.785>.
- [52] A.D. Becke, Density-functional thermochemistry. III. The role of exact exchange, *J. Chem. Phys.* 98 (1993) 5648–5652, <https://doi.org/10.1063/1.464913>.
- [53] R. Behjatmanesh-Ardakani, M.D. Arab, A. Saleem, Z.M. Kotena, S. Bin Mohammad, NBO - NEDA and AIM studies on the interactions between benzocryptand [222B] and Li⁺, Na⁺, K⁺ and Ca²⁺, *Int. J. Pharmaceut. Sci. Rev.* 39 (2016) 45–53.
- [54] T. Yanai, D.P. Tew, N.C. Handy, A new hybrid exchange–correlation functional using the Coulomb-attenuating method (CAM-B3LYP), *Chem. Phys. Lett.* 393 (2004) 51–57, <https://doi.org/10.1016/j.cplett.2004.06.011>.
- [55] M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, G. Scalmani, V. Barone, et al., Gaussian 09 Revision D.01, 2013, Gaussian Inc. Wallingford CT.
- [56] T. Lu, F. Chen, Multiwfn: a multifunctional wavefunction analyzer, *J. Comput. Chem.* 33 (2012) 580–592, <https://doi.org/10.1002/jcc.22885>.
- [57] T. Le Bahers, C. Adamo, I. Ciofini, A qualitative index of spatial extent in charge-transfer excitations, *J. Chem. Theor. Comput.* 7 (2011) 2498–2506, <https://doi.org/10.1021/ct200308m>.
- [58] V.T.T. Huong, H.T. Nguyen, T.B. Tai, M.T. Nguyen, Π -conjugated molecules containing naphtho [2, 3-*b*] thiophene and their derivatives: theoretical design for organic semiconductors, *J. Phys. Chem. C* 117 (2013) 10175–10184, <https://doi.org/10.1021/jp401191a>.
- [59] M.C. Scharber, D. Mühlbacher, M. Koppe, P. Denk, C. Waldauf, A.J. Heeger, C.J. Brabec, Design rules for donors in bulk-heterojunction solar cells—towards 10% energy-conversion efficiency, *Adv. Mater.* 18 (2006) 789–794, <https://doi.org/10.1002/adma.200501717>.
- [60] R.A. Street, S.A. Hawks, P.P. Khlyabich, G. Li, B.J. Schwartz, B.C. Thompson,

A.S. Khazaal, M. Sprinborg, C. Fan et al.

Computational Condensed Matter 25 (2020) e00503

- Y. Yang, Electronic structure and transition energies in polymer–fullerene bulk heterojunctions, *J. Phys. Chem. C* 118 (2014) 21873–21883, <https://doi.org/10.1021/jp507097h>.
- [61] W. Taouali, M.E. Casida, A.M.H.M. Darghouth, K. Alimi, Theoretical design of new small molecules with a low band-gap for organic solar cell applications: DFT and TD-DFT study, *Comput. Mater. Sci.* 150 (2018) 54–61, <https://doi.org/10.1016/j.commatsci.2018.03.038>.
- [62] R.G. Parr, L. v Szentpály, S. Liu, Electrophilicity index, *J. Am. Chem. Soc.* 121 (1999) 1922–1924, <https://doi.org/10.1021/ja983494x>.
- [63] R. Shankar, K. Senthilkumar, P. Kolaidevel, Calculation of ionization potential and chemical hardness: a comparative study of different methods, *Int. J. Quant. Chem.* 109 (2009) 764–771, <https://doi.org/10.1002/qua.21883>.
- [64] M. Ans, M. Paramasivam, K. Ayub, R. Ludwig, M. Zahid, X. Xiao, J. Iqbal, Designing alkoxy-induced based high performance near infrared sensitive small molecule acceptors for organic solar cells, *J. Mol. Liq.* 305 (2020), 112829, <https://doi.org/10.1016/j.molliq.2020.112829>.
- [65] M. Ans, K. Ayub, X. Xiao, J. Iqbal, Tuning opto-electronic properties of alkoxy-induced based electron acceptors in infrared region for high performance organic solar cells, *J. Mol. Liq.* 298 (2020), 111963, <https://doi.org/10.1016/j.molliq.2019.111963>.
- [66] M. Ans, J. Iqbal, I.A. Bhatti, K. Ayub, Designing dithienonaphthalene based acceptor materials with promising photovoltaic parameters for organic solar cells, *RSC Adv.* 9 (2019) 34496–34505, <https://doi.org/10.1039/C9RA06345E>.
- [67] V. Lukeš, A. Aquino, H. Lischka, Theoretical study of vibrational and optical spectra of methylene-bridged oligofluorenes, *J. Phys. Chem.* 109 (2005) 10232–10238, <https://doi.org/10.1021/jp054248s>.
- [68] R.G. Pearson, Absolute electronegativity and hardness correlated with molecular orbital theory, *Proc. Natl. Acad. Sci. Unit. States Am.* 83 (1986) 8440–8441, <https://doi.org/10.1073/pnas.83.22.8440>.
- [69] W.-L. Ding, D.-M. Wang, Z.-Y. Geng, X.-L. Zhao, W.-B. Xu, Density functional theory characterization and verification of high-performance indoline dyes with D–A– π –A architecture for dye-sensitized solar cells, *Dyes Pigments* 98 (2013) 125–135, <https://doi.org/10.1016/j.dyepig.2013.02.008>.
- [70] X.-Q. Ran, J.-K. Feng, A.-M. Ren, W.-C. Li, L.-Y. Zou, C.-C. Sun, Theoretical study on photophysical properties of ambipolar spirofluorene derivatives as efficient blue-light-emitting materials, *J. Phys. Chem.* 113 (2009) 7933–7939, <https://doi.org/10.1021/jp903511r>.
- [71] Y. Geng, H. Li, S. Wu, Y. Duan, Z. Su, Y. Liao, The influence of thienyl-S, S-dioxidation on the photoluminescence and charge transport properties of dithienothiophenes: a theoretical study, *Theor. Chem. Acc.* 129 (2011) 247–255, <https://doi.org/10.1007/s00214-011-0928-6>.
- [72] X. Liu, M. Li, R. He, W. Shen, Theoretical investigations on fluorinated and cyano copolymers for improvements of photovoltaic performances, *Phys. Chem. Chem. Phys.* 16 (2014) 311–323, <https://doi.org/10.1039/c3cp53268b>.
- [73] J. Ku, Y. Gim, Y. Lansac, Y.H. Jang, N-Alkylthienopyrroledione versus benzothiadiazole pulling units in push–pull copolymers used for photovoltaic applications: density functional theory study, *Phys. Chem. Chem. Phys.* 18 (2016) 1017–1024, <https://doi.org/10.1039/C5CP06075C>.
- [74] Y. Gim, D. Kim, M. Kyeong, S. Byun, Y. Park, S. Kwon, H. Kim, S. Hong, Y. Lansac, Y.H. Jang, D–A–D-type narrow-bandgap small-molecule photovoltaic donors: pre-synthesis virtual screening using density functional theory, *Phys. Chem. Chem. Phys.* 18 (2016) 15054–15059, <https://doi.org/10.1039/c5cp07536j>.

3.5 Publication V



Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM

Optimizing small conjugated molecules for solar-cell applications using an inverse-design method

Abdullah S. Khazaal^{a,*}, Michael Springborg^{b,c}, Chencheng Fan^b, Kai Huwig^b

^a Chemistry Department, College of Science, Tikrit University, 34001, Salahuddin, Iraq

^b Physical and Theoretical Chemistry, University of Saarland, 66123, Saarbrücken, Germany

^c Materials Science, Tianjin University, 300350, Tianjin, China



ARTICLE INFO

Article history:
Received 2 October 2019
Received in revised form
17 May 2020
Accepted 18 May 2020
Available online 8 July 2020

Keywords:
Conjugated molecules
Density functional theory
Inverse-design method
Organic solar cells

ABSTRACT

Small organic conjugated molecules are key elements for low-cost photovoltaic devices. One example is cyanopyridone molecules. By modifying these molecules, for instance through optimally chosen functional groups attached to the backbone, their properties can be improved. However, the very large number of possible modifications makes it difficult to identify the best performing molecules. In the present work, we have used a computational inverse-design approach (PooMa) to identify the positions and types of functional groups attached to a modified cyanopyridone that lead to the best performance in solar-energy harvesting. A QSPR model based on five electronic descriptors has been used to determine the properties of solar cells. Our approach uses a genetic algorithm to search the chemical space containing 18^4 (104,976) substituted cyanopyridone systems and predicts out of those the best 20 molecules with optimal performance efficiencies (PCE). PooMa uses the Density-Functional Tight-Binding (DFTB) method for calculating the electronic properties. DFTB is a fast method with acceptable accuracy and, therefore, can be used on a normal desktop without expensive hard- or software. In order to get further information about our suggested systems, a DFT method and its derivative TD-DFT are applied.

© 2020 Elsevier Inc. All rights reserved.

1. Introduction

Organic photovoltaics (OPVs) is a promising technology for converting sunlight into electricity utilizing thin semiconductor films of organic materials. Accordingly, this field has received much attention from the industry [1]. Intensive studies have been devoted to developing such photovoltaic devices during the last decade [2–9]. An important improvement of OPVs is the introduction of the bulk heterojunction (BHJ) architecture in which the thin film contains a mixture of electron donors and acceptors [10,11]. The donor and acceptor molecules should possess a low energy gap between occupied and unoccupied orbitals and a broad absorption band. Moreover, in order to guarantee an efficient charge transport, the HOMO and LUMO orbitals should have energies within certain ranges, which will guarantee an efficient separation of charges and maximize the open-circuit voltage (V_{OC}) [12].

Instead of using polymeric materials in OPVs one may use small molecules (SM) which can have the advantage of easy preparation

and improved film processing. This includes a certain degree of self-organization, which can result in high mobility of charge carriers. The use of SMs can also overcome some problems that result when using polymeric materials, including a wide range of molecular weights, difficulties with purification, and the presence of impurities due to the end groups. These properties of SMs make them good candidate materials for solar-cell applications [13–15]. Experimental and theoretical efforts have shown that SMs of the donor-acceptor (D-A) type are favorable for OPVs which have led to a strong interest in such SMs for organic photovoltaic devices [16,17].

In particular, much interest has been devoted to conjugated SMs during the last two decades, systems that have wide usage in solar cells, light-emitting diodes, and sensors [18–29]. Aromatic systems containing hetero atoms of donor-acceptor (D-A) type show a variety of electronic characteristics which make them interesting for optoelectronic applications [30,31].

For the purpose of identifying new systems with even better properties, theoretical studies can be very helpful [32–34] since experimental studies of new molecules for solar-cell applications can be expensive and take a long time. In the present work, we

* Corresponding author.
E-mail address: abdslem77@tu.edu.iq (A.S. Khazaal).

follow an alternative, cheaper, and approximate theoretical approach based on an inverse-design method with the purpose of identifying the best-performing systems out of a large pool of candidate materials. We emphasize that the use of inverse-design methods in optimizing materials has been suggested and used by others [35–39] although our implementation differs in several aspects from the earlier proposals.

Here, we shall focus on molecules based on the cyanopyridone core, cf. Fig. 1. The cyanopyridone systems exhibit excellent optoelectronic properties [40,41] which make them good candidates for solar-cell applications. The inverse-design method combined with an efficient method for calculating electronic properties will here be used to identify a smaller set of optimally performing systems out of a large number of systems. The approach we shall use has earlier been applied to some Si–Ge clusters and to benzene derivatives for solar energy harvesting [42,43] whereby the main purpose of these studies was to demonstrate the capability of our inverse-design method in optimizing properties. These studies demonstrated that this method can provide a pool of good candidate structures in a short time and can be used on a smaller computer (desktop or laptop) and, accordingly, does not require expensive resources which is the reason that it is called PooMa, the Poor Man's Materials Optimization.

2. Computational details

2.1. PooMa method

Our home-made PooMa program is designed for treating the case that substitutions of certain atoms or groups on a given core backbone within a set of sites shall be optimized for some property. The core may be the surface of a crystalline compound to which an additional layer of atoms or groups of atoms shall be added. One may, for instance, identify the stoichiometry of this extra layer that leads to optimal catalytic performance.

However, in the present work, PooMa will be used in identifying the best substitutions out of 18 different functional groups that include H, nine electron-donating groups (N(CH₃)₂, NH₂, OCH₃, OH, SCH₃, SH, CH₃, CH=CH₂ and C≡CH) and eight electron-withdrawing groups (CN, NO₂, CHO, COOH, F, Cl, Br and I). These groups can be placed at each of the four sites, R₁, R₂, R₃, R₄, of the modified cyanopyridone that is depicted in Fig. 1.

The substitution of the eighteen groups at the four different sites gives a total of 18⁴ different systems. The inverse-design approach provides an efficient optimization of these systems without having to study them all. It is based on genetic algorithms and in the present work, it shall be used to optimize the systems for solar-energy harvesting.

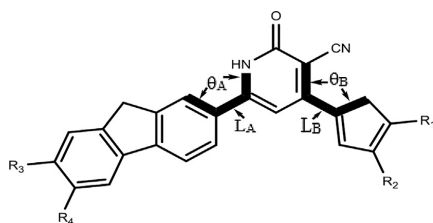


Fig. 1. The backbone of the cyanopyridone with the four sites available for the substitutions. θ_A and θ_B are dihedral angles quantifying the coplanarity of the compound, and L_A and L_B are bond lengths, whose values quantify the charge transfer following an electronic excitation.

In our case, each molecule can be characterized through a string with $N = 4$ integers, whereby each of these integers can take a value up to $K = 18$. N is the number of positions at which a functional group can be attached, and K is the number of functional groups. In the calculations, we keep a constant number, P , of molecules. We apply then a genetic algorithm [44,45] to this set of molecules. The genetic algorithm consists of the following steps (see also Fig. 2):

- 1 $P = 20$ substituted cyanopyridones are created randomly that then form the first so-called generation (parent molecules). It is checked that these molecules are different also when applying symmetry operations.
- 2 The parent members are relaxed locally, after which the values of the performance function that shall be optimized are calculated.
- 3 The P parent molecules are separated into pairs. For each pair, two new molecules (children) are constructed through a cutting and mating procedure whereby each string of N integers is cut randomly into two parts that are interchanged as described in detail in Ref. [43]. Mutations may be introduced to replace in a random way some substituted group through another group, i.e., a randomly chosen integer for one of the children may be changed randomly into another integer.
- 4 It is assured that all the children and parent molecules are different, also when taking symmetry into account. Eventually, step 3 may have to be repeated before this is achieved.
- 5 The children's molecules are treated as under 2. Subsequently, out of the 20 parents and 20 children molecules, those 20 molecules that have the best values of the performance function are kept as the parents for the next generation.
- 6 If the twenty new parent molecules in 5 are unchanged for 10 generations the calculation will be terminated, otherwise, return to 3.

2.2. The performance function

The efficiency of an OPV can be expressed as [46]:

$$\eta = \frac{J_{sc} FF V_{oc}}{P_{in}} \quad (1)$$

The photovoltaic parameters quantifying the power conversion efficiency of solar cells are the short-circuit current density (J_{sc}), the open-circuit photovoltage (V_{oc}), the fill factor (FF), and the intensity of the incident light (P_{in}). All these parameters depend on all parts of the photovoltaic cell, i.e., the nature of the electrodes, the electrolyte, and the organic molecules. With electronic-structure methods, it is not possible to study such a complex system efficiently. Therefore, we shall use an approach similar to what we presented recently for substituted porphyrins [47], i.e., use a QSPR model for describing the cell efficiency through properties of the individual, isolated molecules.

Accordingly, we shall calculate the performance function from the values of five molecular parameters that will be described below. Subsequently, η will be approximated through a function of those five parameters as given through a QSPR model [47] and this is the quantity that shall be optimized in the PooMa calculations.

The five parameters that we consider as being relevant to η are:

As the first parameter, we consider the energy gap G between the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO). In general, the lower value the better is the efficiency,

$$G = \epsilon_{LUMO} - \epsilon_{HOMO} \quad (2)$$

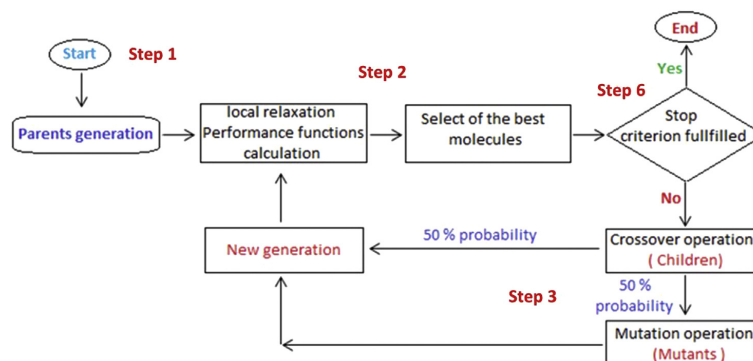


Fig. 2. Flowchart of the genetic algorithm.

Second, we use the absorption of sunlight per valence electron,

$$A = \frac{\sum_{i,a} f_{ia} \cdot P(\epsilon_a - \epsilon_i)}{N_e} \quad (3)$$

Here, we divide the property by the number of valence electrons N_e in order to arrive at an intensive property to avoid that the largest molecules are predicted to be the best candidates. Moreover, i and a denote occupied and unoccupied orbitals, respectively. f_{ia} is a weight factor that may be the oscillation strength. In that case, it is calculated using the framework of the approximated TD-DFT (time-dependent density-functional theory) method [48]. Since we consider only single-electron excitations, f_{ia} can be expressed as

$$f_{ia} = \frac{4}{3} (\epsilon_a - \epsilon_i) |\vec{d}_{ia}|^2 \quad (4)$$

\vec{d}_{ia} is the transition dipole moment that is approximated as $\sum_n \vec{R}_n \cdot q_{n,ia}$ in terms of Mulliken transition charges $q_{n,ia}$ and the corresponding coordinates \vec{R}_n of the atom n for the $i \rightarrow a$ transition.

$P(\epsilon_a - \epsilon_i)$ represents the solar spectrum and is modelled through a black-body spectrum:

$$P(\epsilon_a - \epsilon_i) = \frac{8\pi (\epsilon_a - \epsilon_i)^3}{(2\pi\hbar C)^3} \left[\exp\left(\frac{\epsilon_a - \epsilon_i}{k_B T}\right) - 1 \right]^{-1} \quad (5)$$

where k_B is the Boltzmann constant, C is the speed of light, and T is the temperature.

A third parameter is a light-harvesting efficiency (LHE) given as

$$\text{LHE} = \frac{\sum_{i,a} (1 - 10^{-f_{ia}})}{N_e} \quad (6)$$

We also consider the average of the orbital overlap O_{av} that is calculated as a weighted sum over O_{ia} , i.e., the overlap of each pair of occupied and unoccupied molecular orbitals over all transitions from i to a weighted by the transition probability $f_{ia} P(\epsilon_a - \epsilon_i)$ and divided by the number of valence electrons

$$O_{av} = \frac{\sum_{i,a} f_{ia} P(\epsilon_a - \epsilon_i) O_{ia}}{N_e} \quad (7)$$

here, $O_{ia} = \left| \sum_n q_{n,i} \cdot q_{n,a} \right|$, where $q_{n,i}$ and $q_{n,a}$ are the atomic

gross Mulliken populations of the atom n for the transition $i \rightarrow a$.

Finally, as the fifth parameter, we use the average value of the orbital distance D_{av} . We expect that the recombination rate will be reduced when the spatial distance D_{ia} between occupied and unoccupied orbitals becomes large. D_{av} is calculated as in Eq. (7) but with O_{ia} replaced by

$$D_{ia} = \left| \sum_n q_{n,a} \vec{R}_n - \sum_n q_{n,i} \vec{R}_n \right| \quad (8)$$

giving

$$D_{av} = \frac{\sum_{i,a} f_{ia} P(\epsilon_a - \epsilon_i) D_{ia}}{N_e} \quad (9)$$

In experimental studies, the systems contain other parts (electrolyte, electrodes) that also contribute to the performance of the systems. It is thus an approximation to model the performance through properties of just the isolated dyes. However, for different systems based on similar electrolytes and electrodes, we believe that our approach will be able to describe the variation in the performance when varying the dyes. Therefore, our QSPR model, described in detail in Fan et al. [47], is based on modeling the measured performances of 206 systems that mainly differed in the dye but otherwise were very similar. The resulting non-linear QSPR contained five electronic properties (independent variables) with an $R^2 = 0.23$ for the training set (MAE = 1.63; RMSE = 1.98), and $Q^2 = 0.42$ for the validation set (MAE = 1.58; RMSE = 1.78) and $Q^2 = 0.42$ for the test set (MAE = 1.61; RMSE = 1.87). While the model developed by Fan et al. is not perfect, the model aided the design of 15 novel porphyrins for dye-sensitized solar cells with optimal power conversion efficiency (PCE) [47].

2.3. DFTB approach

In order to determine the parameters of the last section, we shall use an electronic-structure method that is not too expensive, which is important because of the larger number of systems that will be treated within the PooMa approach. The method is chosen according to its compromise between accuracy and computational speed. Accordingly, we use the non-self-consistent density-functional tight-binding (DFTB) method of Seifert and co-workers. This method does not take into account the effects of fluctuations of charges and the resulting interactions [49,50]. The DFTB method will be used to determine the equilibrium structures of the

4

A.S. Khazaal et al. / Journal of Molecular Graphics and Modelling 100 (2020) 107654

molecular systems and subsequently to calculate the properties above. With this method, the total energy E_{tot} of a system relative to the energies of the non-interacting atoms is expressed as

$$E_{\text{tot}} \approx \sum_i^{\text{occ}} \epsilon_i - \sum_j \sum_m^{\text{occ}} \epsilon_{jm} + \frac{1}{2} \sum_{j \neq k} U_{jk} (\vec{R}_j - \vec{R}_k). \quad (10)$$

The first term represents the energies of the occupied orbitals of the system of interest and the second one constitutes the energies of the isolated neutral atoms. The third term U_{jk} represents short-ranged repulsive potentials between atoms k and j . The effects of the core electrons are included within a frozen-core approximation. The orbital energies are calculated using the Kohn-Sham equations for an effective one-particle potential $V_{\text{eff}}(\vec{r})$,

$$\hat{H}\psi_i(\vec{r}) = \epsilon_i \psi_i(\vec{r}) \quad (11)$$

here, the Hamilton operator \hat{H} involves the kinetic-energy operator

$$\hat{T} = -\frac{\hbar^2}{2m} \nabla^2 \quad (12)$$

and the potential-energy operator that is written as the superposition of the potentials of the isolated, neutral atoms

$$V_{\text{eff}}(\vec{r}) = \sum_m V_m(|\vec{r} - \vec{R}_m|) \quad (13)$$

\vec{R}_m is the position of the m th nucleus. The Kohn-Sham orbitals $\psi_i(\vec{r})$ are expanded in a basis set of atom-centered functions.

2.4. DFT method

The PooMa calculations combine the genetic algorithm for optimization with the DFTB method for electronic-structure calculations for the different molecules. The output of the PooMa calculations is a smaller set of molecules with optimal properties. The philosophy behind PooMa is to avoid relying on heavy computations since although the genetic algorithm allows for reducing drastically the number of molecules that shall be treated, this number is still typically some 1000s (in the present study, we considered 860 molecules). Therefore, combining the genetic algorithm with the DFTB method makes PooMa efficient.

In order to get more insight into the properties of the systems that result from the PooMa calculations, we have studied those using density-functional-theory (DFT) and time-dependent density-functional-theory (TD-DFT). The structural optimization of the ground states was carried through with the B3LYP (Becke 3-parameter Lee-Yang-Parr) exchange-correlation functional [51] and the 6-311G(d,p) basis set. Some recent studies have demonstrated the ability of the CAM-B3LYP method to predict excitation energies and absorption spectra of conjugated molecules [52–54]. Therefore, excited state calculations were carried out for the optimized structures of the ground states using the CAM-B3LYP method with a 6-311G(d,p) basis set in order to predict the vertical excitation energies and the absorption spectra. Solvent effects were taken into account for methanol as a solvent [36] using the conductor-like polarizable continuum model CPCM that represents an efficient tool to treat solvent effects for both excited and ground states [55].

All DFT and TD-DFT calculations were performed using the Gaussian09 software package [56]. The resulting output was analyzed further with the help of the Multiwfn software package [57].

2.5. Excitation and emission processes

We shall use an earlier proposed model to analyze excitation properties [58]. At first, we define the change in the electron density due to a vertical excitation,

$$\Delta\rho(\mathbf{r}) = \rho_{\text{ex}}(\mathbf{r}) - \rho_{\text{gs}}(\mathbf{r}) \quad (14)$$

$\rho_{\text{ex}}(\mathbf{r})$ and $\rho_{\text{gs}}(\mathbf{r})$ are the densities of the excited state and the ground state, respectively. Of obvious reasons, the spatial integral of $\Delta\rho(\mathbf{r})$ will vanish, implying that $\Delta\rho(\mathbf{r})$ will have regions where it is positive and regions where it is negative. Therefore, $\Delta\rho(\mathbf{r})$ can be separated into one part that equals $\Delta\rho(\mathbf{r})$ where this is positive and otherwise being zero and another part that equals $\Delta\rho(\mathbf{r})$ where this is negative and otherwise being zero. Calling those two parts $\rho_+(\mathbf{r})$ and $\rho_-(\mathbf{r})$, respectively, their barycenters are

$$R_+ = \frac{\int r \rho_+(\mathbf{r}) d\mathbf{r}}{\int \rho_+(\mathbf{r}) d\mathbf{r}} \quad (15)$$

$$R_- = \frac{\int r \rho_-(\mathbf{r}) d\mathbf{r}}{\int \rho_-(\mathbf{r}) d\mathbf{r}} \quad (16)$$

The length of the charge transfer excitation D_{CT} can be defined as

$$D_{\text{CT}} = |R_+ - R_-| \quad (17)$$

As a measure for the transferred charge upon excitation we use

$$q_{\text{CT}} = \int \rho_+(\mathbf{r}) d\mathbf{r} = - \int \rho_-(\mathbf{r}) d\mathbf{r}. \quad (18)$$

For one-electron excitations, q_{CT} lies between 0 and 1.

The change in the dipole moment due to the excitation can be quantified through

$$\mu_{\text{CT}} = D_{\text{CT}} q_{\text{CT}}. \quad (19)$$

Another important quantity is the reorganization energy λ . Here, we will define these for hole and electron transfer processes that are relevant when the charge is moving from one molecule to the next. These energies can be calculated from (see Fig. 3) [59]:

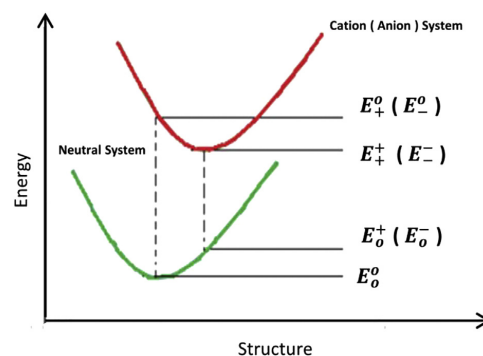


Fig. 3. Schematic representation of the energies of the neutral, and the cationic (anionic) systems used in the calculation of the reorganization energies.

$$\lambda_+ = (E_+^0 - E_0^0) + (E_0^+ - E_+^+) \quad (20)$$

$$\lambda_- = (E_-^0 - E_0^0) + (E_0^- - E_-^-) \quad (21)$$

here, E_0^0 , E_+^+ , and E_-^- represent the energies of the neutral, the cationic, and the anionic species in their equilibrium structures, respectively. E_+^0 (E_-^0) is the energy of the neutral molecule in the optimized cationic (anionic) geometry. Similarly, E_0^+ (E_0^-) is the total energy of the cation (anion) in the geometry of the neutral molecule. For a fast and efficient charge transfer, the reorganization energies should be small.

Vertical excitation energies describe UV/Vis absorption spectra as indicated in Fig. 4. These energies are calculated using TD-DFT for the ground-state structures.

The oxidation potential E^{OX} of the ground state can be estimated as the negative of E_{HOMO} by using Koopmans' theorem, while the oxidation potential of a molecule in an excited state, E^{OX*} can be estimated from

$$E^{OX*} = E^{OX} - E^0 \quad (22)$$

E^0 is the energy corresponding to the maximum in the absorption spectra [60].

The emission energy from the excited to the ground state can be calculated through the vertical S_1 to S_0 transition using the relaxed structure of the excited state [61]. The spontaneous emission decay rate from the excited state to the ground state can then be determined from

$$\kappa_r = \frac{4}{3} \frac{E_{Flu}^3}{C^3 \mu_{10}^2} \quad (23)$$

here, E_{Flu} is the energy related to the fluorescence process, i.e., the vertical energy of the emission from the excited to the ground state. Moreover, C and μ_{10} represent the speed of light and transition dipole, respectively. The radiative lifetime τ_r of the emission process can then be obtained from Ref. [62,63]

$$\tau_r = \frac{1}{\kappa_r} \quad (24)$$

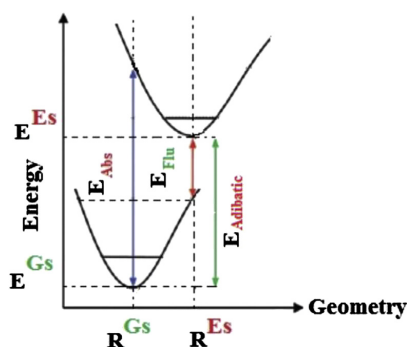


Fig. 4. Schematic representation of absorption and emission processes. R^{Gs} and R^{Es} are the geometries of the ground and excited states, respectively. E^{Gs} and E^{Es} are the corresponding energies of the excited and ground state. E^{Abs} , E^{Flu} , and $E^{Adiabatic}$ are the energies of absorption and fluorescence emission, and adiabatic transition, respectively.

2.6. Condensed Fukui functions

Parr and Yang [64,65] introduced the Fukui function $f(r)$ within the framework of density functional theory. $f(r)$ of a molecular system gives information on the reactivity of different sites in the system of interest. It is defined as

$$f(r) = \left[\frac{\partial \rho(r)}{\partial N} \right]_{v(r)} \quad (25)$$

$\rho(r)$ equals the electron density at r , N is the total number of electrons of the system, and $v(r)$ represents the external potential. Thus, the Fukui function describes where electrons preferably will be added or removed when their number changes. Parr and Yang used a finite difference approximation to calculate the Fukui functions. Depending on the charge-change process, different Fukui functions may be defined. Thus, $f^+(r)$ describes a nucleophilic attack where electrons are added to the system, and $f^-(r)$ describes an electrophilic attack where electrons are removed from the system. To a first approximation, we have

$$f^+(r) = \rho_{N+1}(r) - \rho_N(r) \approx \rho^{LUMO}(r) \quad (26)$$

$$f^-(r) = \rho_N(r) - \rho_{N-1}(r) \approx \rho^{HOMO}(r) \quad (27)$$

More recently, Yang and Mortier [66] presented a procedure for calculating the so-called condensed Fukui functions for individual atoms in the molecule based on separating the integral of the Fukui function over the complete space into atomic components. Denoting the atomic charges for the atom a according to the charge decomposition of Hirshfeld $q_a(N+1)$, $q_a(N-1)$, and $q_a(N)$ for the anion, the cation, and the neutral system, respectively, the condensed Fukui functions of each atom are defined as

$$f_a^- = q_a(N-1) - q_a(N) \quad (28)$$

$$f_a^+ = q_a(N) - q_a(N+1) \quad (29)$$

Equations (28) and (29) are used for electrophilic and nucleophilic attacks, respectively.

3. Results and discussion

3.1. The results of the PooMa optimization

At first, we shall discuss the systems that were identified in the PooMa calculations as being the ones with the best performances. They are shown in Fig. 5. The 20 different systems have different values for the properties that are used in the QSPR model for the PCE.

All the obtained molecules have a D - π - A structure, i.e., consist of an electron donor (D) and acceptor (A) linked through a conjugated bridge (π). Most of the optimized molecules have $N(CH_3)_2$, OCH_3 , or $COOH$ as electron donor and NO_2 , CHO , or $COOH$ as an electron acceptor. This is different from our findings for substituted porphyrins [47] implying that also the conjugated bridge is important.

3.2. Geometrical and electronic properties

The cyanopyridone derivatives that were obtained in the PooMa calculations were subsequently studied using the more accurate theoretical methods. For the charge separation following the electronic excitation due to the absorption of energy, some of the geometrical parameters are of crucial importance, including the

6

A.S. Khazaal et al. / Journal of Molecular Graphics and Modelling 100 (2020) 107654

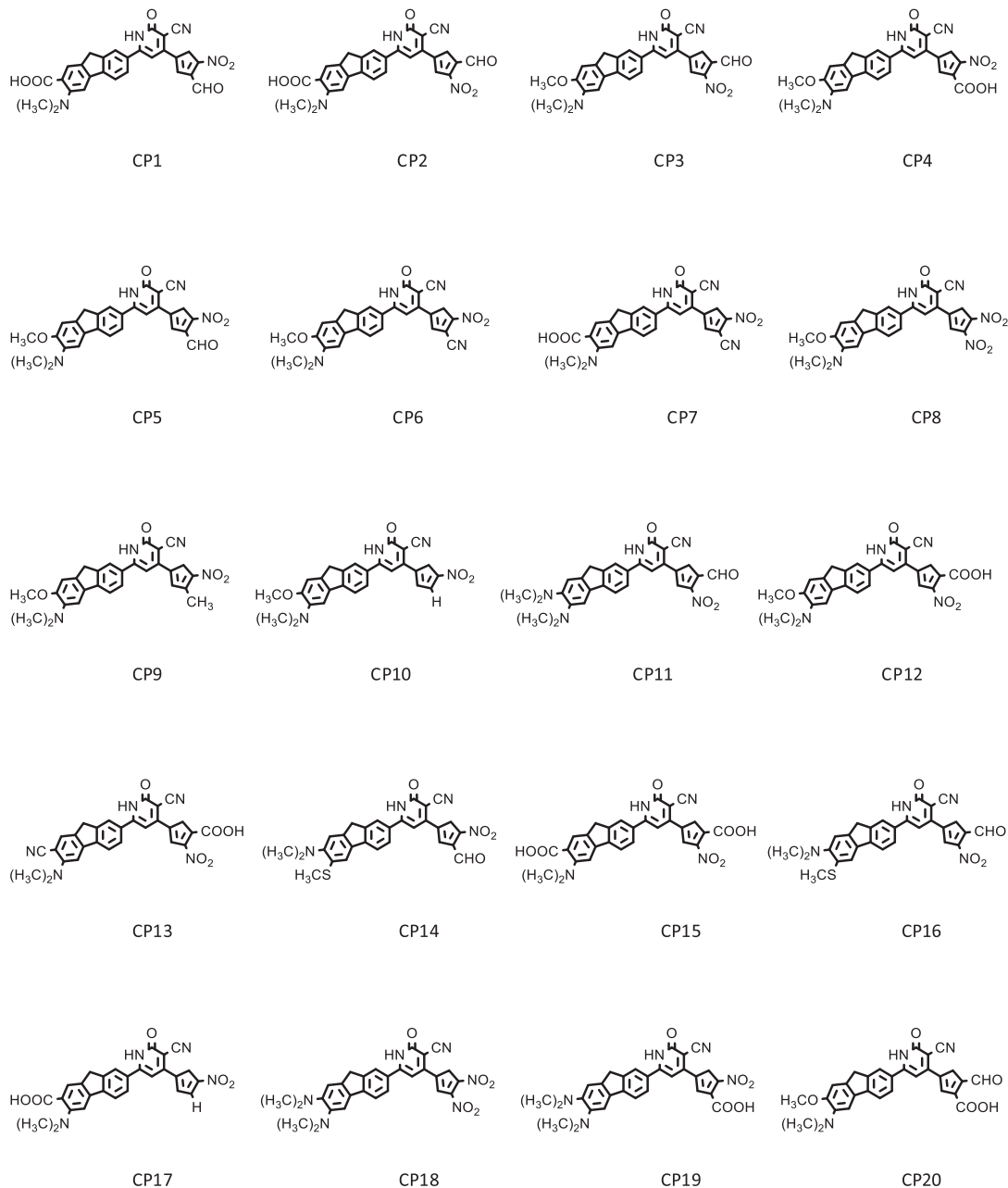


Fig. 5. The best performing molecules according to the PooMa calculations.

two bond lengths L_A and L_B and the two dihedral angles Θ_A and Θ_B that were shown in Fig. 1. The values of those, both in the ground state (S_0) and in the first excited (S_1) state, are tabulated in Table 1.

The values of L_A and L_B for the excited state are slightly smaller

than those of the ground state, which suggests that these bonds change towards C=C bonds. This implies an increased π -electron delocalization and a stronger interaction between the two parts separated by the cyanopyridone ring. This will favor the

Table 1The bond lengths L_A and L_B in Å and dihedral angles θ_A and θ_B in deg. for the molecules of Fig. 5 in both the ground state (S_0) and the first excited state (S_1).

Cyanopyridone Derivatives	S_0				S_1			
	L_A	L_B	θ_A	θ_B	L_A	L_B	θ_A	θ_B
CP1	1.472	1.473	-32.779	42.883	1.432	1.440	-15.451	19.628
CP2	1.472	1.473	-32.844	42.322	1.426	1.448	-12.446	25.824
CP3	1.469	1.473	-31.014	42.693	1.427	1.448	-11.901	24.365
CP4	1.468	1.473	-31.039	42.768	1.432	1.440	-15.108	19.256
CP5	1.468	1.474	-30.854	43.400	1.458	1.437	-25.361	10.493
CP6	1.468	1.475	-30.684	47.350	1.457	1.435	-24.887	12.744
CP7	1.472	1.474	-32.644	46.429	1.433	1.438	-16.078	20.340
CP8	1.468	1.474	-30.773	44.584	1.461	1.437	-27.731	-14.969
CP9	1.469	1.472	-30.966	43.248	1.426	1.447	-11.423	24.590
CP10	1.469	1.472	-30.690	42.258	1.427	1.445	-12.205	23.768
CP11	1.468	1.473	-30.806	42.635	1.436	1.445	-15.262	20.798
CP12	1.469	1.473	-31.222	41.693	1.442	1.454	-9.940	29.434
CP13	1.473	1.472	-33.476	41.774	1.442	1.453	-10.643	29.544
CP14	1.470	1.473	-31.734	43.281	1.444	1.438	-18.816	13.946
CP15	1.470	1.473	-31.811	42.595	1.427	1.450	-11.569	24.759
CP16	1.472	1.472	-32.781	41.435	1.422	1.454	-10.426	30.196
CP17	1.473	1.472	-32.511	42.067	1.426	1.446	-12.391	25.221
CP18	1.467	1.474	-30.532	45.151	1.461	1.438	-28.382	-15.171
CP19	1.468	1.473	-30.792	43.151	1.449	1.434	-20.295	10.798
CP20	1.469	1.472	-31.296	41.242	1.422	1.453	-9.377	28.595

Table 2

The values of the condensed Fukui function for the compounds CP1-CP20.

Compound	Core		R_1		R_2		R_3		R_4	
	f^+	f^-	f^+	f^-	f^+	f^-	f^+	f^-	f^+	f^-
CP1	0.53	0.51	0.31	0.00	0.14	0.00	0.00	0.08	0.00	0.40
CP2	0.54	0.51	0.21	0.00	0.23	0.00	0.00	0.08	0.00	0.40
CP3	0.52	0.59	0.21	0.00	0.24	0.00	0.00	0.09	0.00	0.30
CP4	0.62	0.59	0.30	0.00	0.06	0.00	0.00	0.09	0.00	0.30
CP5	0.52	0.59	0.32	0.00	0.14	0.00	0.00	0.04	0.00	0.30
CP6	0.55	0.58	0.32	0.00	0.10	0.00	0.00	0.04	0.00	0.30
CP7	0.56	0.51	0.31	0.00	0.10	0.00	0.00	0.08	0.00	0.40
CP8	0.52	0.58	0.30	0.00	0.15	0.00	0.00	0.09	0.00	0.30
CP9	0.68	0.59	0.25	0.00	0.05	0.00	0.00	0.09	0.00	0.30
CP10	0.68	0.59	0.28	0.00	0.02	0.00	0.00	0.09	0.00	0.30
CP11	0.52	0.53	0.21	0.00	0.24	0.00	0.00	0.24	0.00	0.20
CP12	0.67	0.59	0.14	0.00	0.17	0.00	0.00	0.09	0.00	0.30
CP13	0.70	0.51	0.13	0.00	0.14	0.00	0.00	0.08	0.00	0.39
CP14	0.52	0.52	0.32	0.00	0.14	0.00	0.00	0.23	0.00	0.22
CP15	0.53	0.52	0.21	0.00	0.24	0.00	0.00	0.23	0.00	0.16
CP16	0.69	0.51	0.13	0.00	0.15	0.00	0.00	0.08	0.00	0.40
CP17	0.69	0.51	0.26	0.00	0.02	0.00	0.00	0.08	0.00	0.40
CP18	0.52	0.53	0.30	0.00	0.15	0.00	0.00	0.24	0.00	0.20
CP19	0.62	0.53	0.30	0.00	0.06	0.00	0.00	0.24	0.00	0.20
CP20	0.70	0.59	0.19	0.00	0.08	0.00	0.00	0.09	0.00	0.30

intramolecular charge transfer. The coplanarity of the cyanopyridone derivatives is quantified through the values of the dihedral angles θ_A and θ_B . For the ground state, no molecule is coplanar but have dihedral angles in the range of 30–47 deg. Upon excitation from S_0 to S_1 , we see a decrease in these values that then lie in the range 9–30 deg. Also, this favors the charge transfer processes.

The values of the condensed Fukui functions are listed in Table 2. We notice that those parts of the molecules with the R_3 and R_4 groups are the electron-donating parts, whereas those parts with the R_1 and R_2 groups are the electron-accepting parts. However, both for electron-donating and -accepting processes, the core plays a major role. The Fukui functions themselves are shown in Figs. 6 and 7. These figures confirm the interpretation above but give also a more detailed picture of the charge-transfer processes.

Since the absolute values of the HOMO and LUMO energies, as well as the energy gap separating them, are important for organic solar cells we show those in Fig. 8. The calculated band gaps of our

systems lie in the range 1.88–2.82 eV and are smaller than that of the unsubstituted cyanopyridone which has a 3.24 eV energy gap. The highest band gap among the cyano derivatives we find for CP20 and the lowest band gaps for CP1, CP2, CP3, and CP14. We stress that a decrease in the band gap leads to an improvement in V_{oc} and J_{sc} .

3.3. Optical and charge transport properties

Next, using TD-DFT/CAM we studied the electronic transitions of our systems when being dissolved in methanol. The calculated values of the absorption wavelengths, vertical excitation energies (E_{ex}), and oscillator strengths (O_s) for the main absorption peak as well as the light harvesting efficiencies (LHE) are tabulated in Table 3. All the vertical excitation energies are lower than that of the unsubstituted system (3.486 eV) being in the range 2.894–3.268 eV with the lowest values for CP8, CP18, and CP19. These excitations can be attributed to $\pi - \pi^*$ transitions. The bathochromic shift in comparison with the unsubstituted cyanopyridone can be explained through an increase in the conjugation length of the systems. In addition, systems with two electron-donating groups in part A and two electron-withdrawing groups in part B show the highest bathochromic shift. The corresponding longer wavelength will facilitate the conversion of photo energy to electric current in the solar cells. For a system to have high absorption and J_{sc} , it should have large LHE, but as can be seen in Table 4, the LHE values of our systems are fairly close (0.778–0.898).

Fig. 9 shows the values of the oxidation potentials in the excited states for the substituted cyanopyridones. The most oxidizing compounds are CP1, CP2, CP13, and CP16, which implies that these compounds have the best electron-donation properties among the suggested systems. The optical properties suggest that all systems can serve for the conversion of sunlight into electrical energy.

The emission spectra for the $S_1 \rightarrow S_0$ transition is for all systems assigned to $\pi^* - \pi$ transitions. Table 4 shows that the luminescence transitions range from blue to green bands (418–559 nm). The Stokes shift of all systems falls in the range of 38–131 nm. The relatively large values of the Stokes shift may be attributed to significant structural rearrangements in the S_1 state [67]. The computed radiative lifetimes (τ_r) are listed in Table 5 with the

8

A.S. Khazaal et al. / Journal of Molecular Graphics and Modelling 100 (2020) 107654

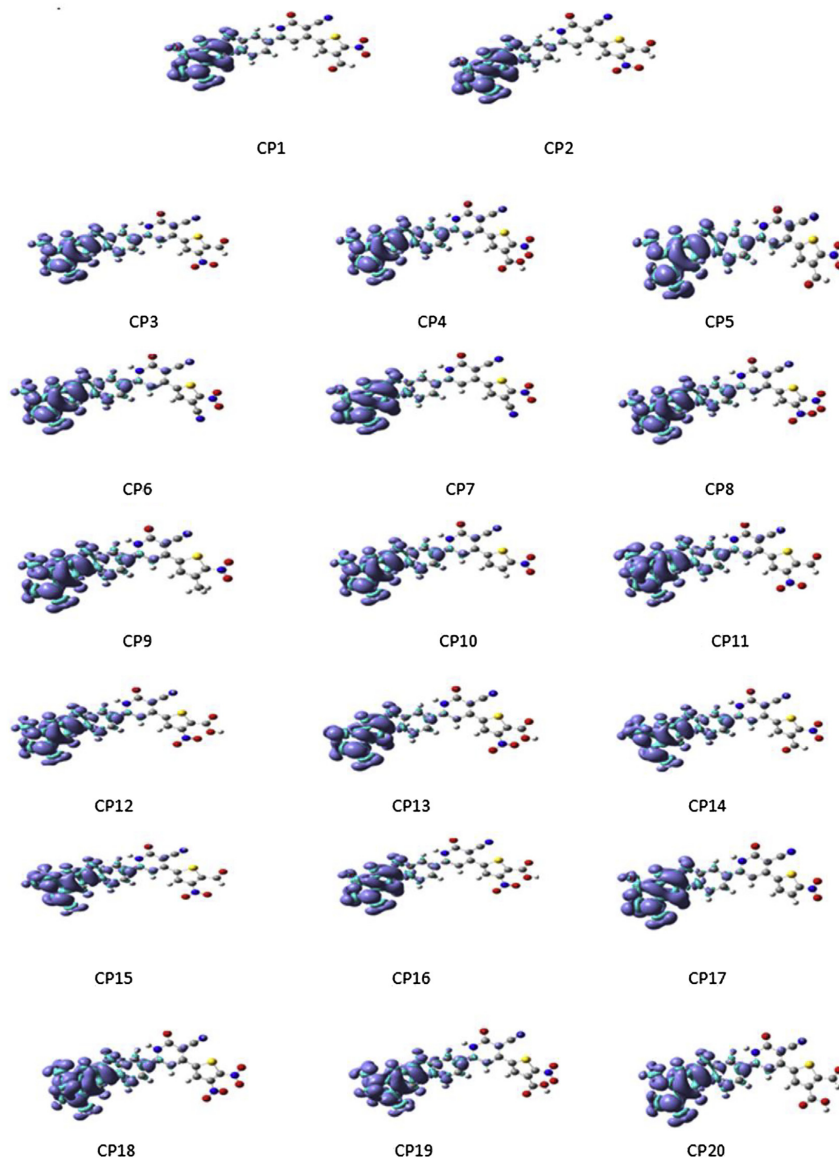


Fig. 6. The Fukui function f for the various cyanopyridone derivatives.

highest values found for CP18, CP8, CP6, and CP5. The high values of the radiative lifetimes correspond to low values of the LHE.

Table 5 shows the values of the transferred charge q_{CT} upon the $S_0 \rightarrow S_1$ excitation. In all cases, it lies in the range 0.279–0.317 a.u., i.e., it shows only small scatter. The largest value for the charge transfer length is found for CP18 (0.919 Å) and the lowest for CP16 (0.314 Å). The values for this property correlate strongly with the variation in the transition dipole moment μ_{CT} due to excitation as well as with the values of the reorganization energies of holes as

can be seen from Fig. 10.

The values of the reorganization energies for electrons and holes are tabulated in Table 5. We recognize that the compounds CP1, CP16, CP2, CP7, and CP17 have the lowest hole reorganization energies. Malgoli and Brédas reported that a λ_+ value of 0.29 eV for TPD implies a good hole-transport material [68]. When we compare that value with our results for the five compounds mentioned above, we conclude that these compounds represent good hole-transport materials, too. The lowest reorganization energies for

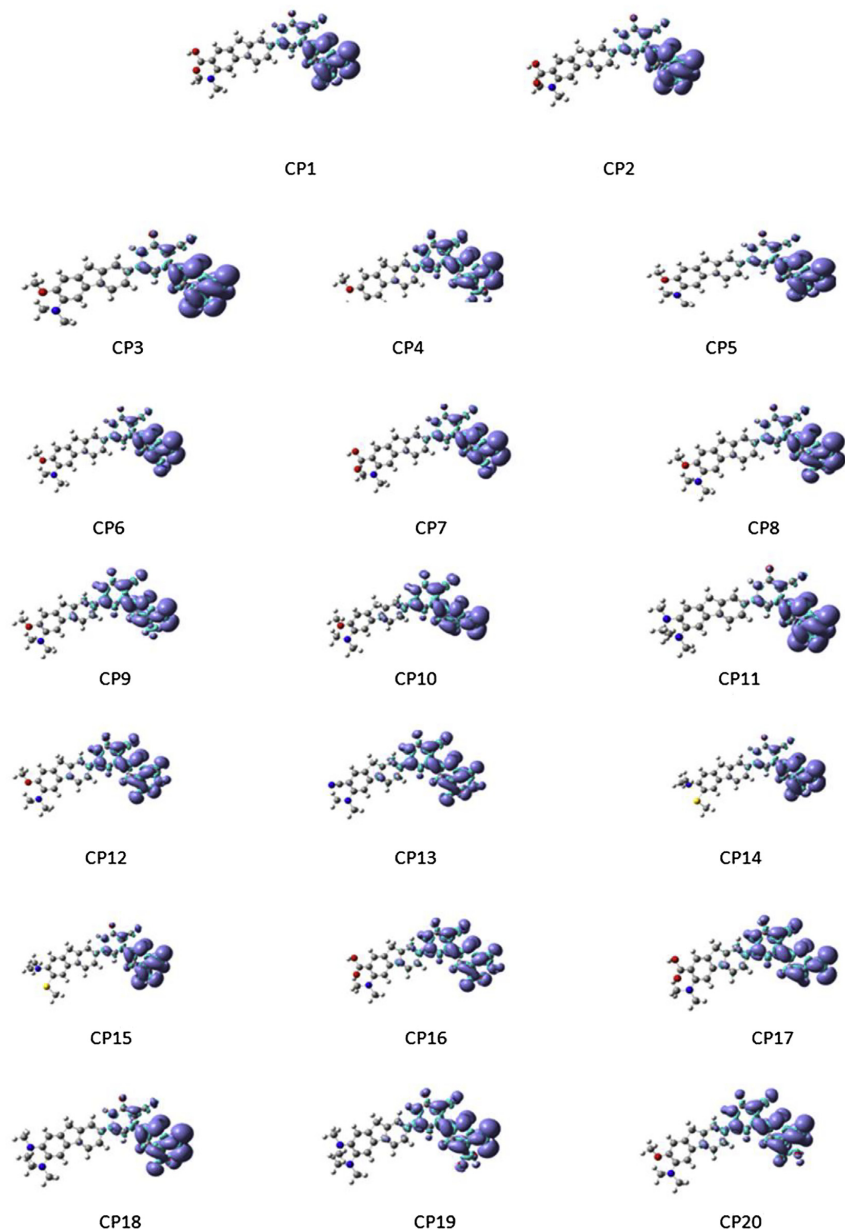


Fig. 7. The Fukui function f^+ for the various cyanopyridone derivatives.

electrons were found for CP2, CP15, CP11, CP3, and CP20. Of all systems, CP2 possesses the best hole and electron transport properties. The systems CP4, CP8, and CP12 have a small difference between λ_- and λ_+ values, which suggests that these systems possess good properties for ambipolar charge-transport processes.

4. Conclusions

In summary, we have used a recently developed inverse-design approach (PooMa) for designing a novel series of small conjugated molecules based on the cyanopyridone scaffold and including 18 different substituents at four sites (see Fig. 1). The aim was to

10

A.S. Khazaal et al. / Journal of Molecular Graphics and Modelling 100 (2020) 107654

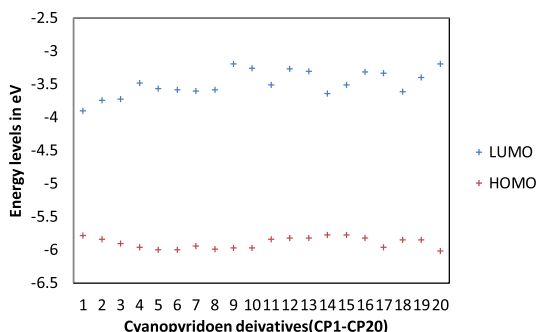


Fig. 8. The energies of the frontier molecular orbitals HOMO and LUMO for the 20 cyanopyridone derivatives.

Table 3

Absorption spectra data as obtained with the TD-DFT method for the cyanopyridone derivatives (CP1-CP20) together with results for the unmodified cyanopyridone compound (CP0). λ_{max} (nm) is the wavelength of maximum absorption, E_{ex} the electronic transition energy (eV) for the $S_0 \rightarrow S_1$ transition, Os the corresponding oscillator strength, and LHE is the light harvesting efficiency.

Compound	λ_{max}	E_{ex}	Os	LHE
CP0	356	3.486	0.928	0.822
CP1	386	3.215	0.675	0.789
CP2	380	3.267	0.788	0.837
CP3	400	3.100	0.847	0.858
CP4	405	3.064	0.819	0.848
CP5	410	3.024	0.663	0.783
CP6	412	3.008	0.670	0.786
CP7	388	3.200	0.673	0.788
CP8	415	2.989	0.654	0.778
CP9	397	3.124	0.946	0.887
CP10	399	3.105	0.914	0.878
CP11	412	3.001	0.876	0.867
CP12	393	3.156	1.017	0.904
CP13	371	3.344	0.955	0.889
CP14	397	3.126	0.723	0.811
CP15	389	3.191	0.892	0.872
CP16	375	3.302	0.853	0.860
CP17	379	3.268	0.819	0.848
CP18	428	2.894	0.657	0.780
CP19	417	2.976	0.859	0.862
CP20	392	3.166	0.994	0.898

identify those substituted molecules that perform best in solar-energy harvesting. To this end, we applied a QSPR model that is constructed for predicting the performance of a solar cell from five properties of the isolated molecules that in turn were calculated using the computational efficient DFTB method. The calculations predicted twenty systems with the best efficiencies. In order to get more detailed information about the suggested systems, DFT and TD-DFT calculations were carried through for the 20 molecules that gave more information on structure, molecular orbitals, as well as on electronic, optical, and charge transfer processes. The results show that the systems have the advantage of having high values of J_{sc} due to their small HOMO-LUMO gaps, and the low HOMO energies lead to a high value of V_{OC} . All the suggested systems showed a red shift in λ_{max} in comparison with the unsubstituted compound. The values of the radiative lifetime through the emission process and of the Stokes shift show a large spread. The reorganization energies show that some systems are good in hole transport and some can be used in ambipolar charge transport. The transport charge is largely constant for all the systems, while the charge

Table 4

Emission spectra data obtained with the TD-DFT method for the substituted cyanopyridones. λ_{max} (nm) is the wavelength at the maximum emission, ΔE (E_{flu}) the emission energy in eV of the $S_1 \rightarrow S_0$ transition, Os the corresponding oscillator strength, τ_r the radiative lifetime (ns), and SS is the Stokes shift (nm).

Compound	λ_{max}	ΔE	Os	τ_r	SS
CP1	427	2.903	0.862	3.42	41
CP2	418	2.956	0.980	2.86	38
CP3	450	2.755	1.064	3.11	50
CP4	460	2.690	1.017	3.45	55
CP5	503	2.463	0.886	5.13	93
CP6	505	2.456	0.892	5.11	93
CP7	477	2.598	0.860	4.76	89
CP8	516	2.400	0.873	5.56	101
CP9	448	2.767	1.133	2.91	51
CP10	504	2.462	1.102	4.22	105
CP11	489	2.537	1.117	3.69	77
CP12	443	2.799	1.203	2.67	50
CP13	411	3.020	1.124	2.41	40
CP14	486	2.549	0.925	4.58	89
CP15	461	2.688	1.085	3.39	72
CP16	415	2.990	1.057	2.61	40
CP17	419	2.958	0.996	2.84	40
CP18	559	2.219	0.897	6.64	131
CP19	516	2.401	1.076	2.48	99
CP20	441	2.809	1.181	2.70	49

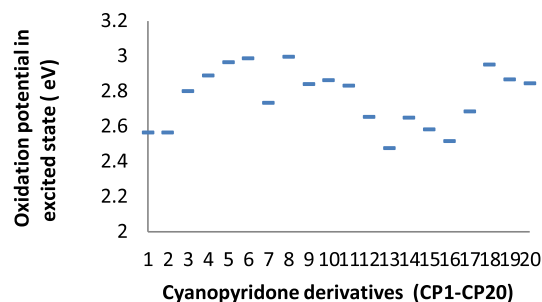


Fig. 9. The oxidation potential in the excited state of the substituted cyanopyridones.

Table 5

Charge-transfer parameters of the substituted cyanopyridones. The properties are the charge transfer excitation length (D_{CT}) in Å, the charge-transfer dipole moment (μ_{CT}) in Debye, the transferred charge (q_{CT}) in a.u., the reorganization energy for electrons λ_- and holes λ_+ in eV.

Compound	D_{CT}	μ_{CT}	q_{CT}	λ_-	λ_+
CP1	0.412	0.594	0.300	0.542	0.253
CP2	0.317	0.460	0.303	0.484	0.251
CP3	0.809	1.147	0.295	0.490	0.647
CP4	0.754	1.087	0.300	0.625	0.646
CP5	0.844	1.200	0.296	0.542	0.646
CP6	0.899	1.267	0.294	0.544	0.645
CP7	0.411	0.589	0.298	0.542	0.251
CP8	0.770	1.117	0.302	0.663	0.648
CP9	0.812	1.143	0.293	0.547	0.641
CP10	0.796	1.065	0.279	0.555	0.642
CP11	0.919	1.363	0.309	0.488	0.742
CP12	0.621	0.897	0.301	0.622	0.645
CP13	0.262	0.368	0.293	0.582	0.422
CP14	0.719	1.024	0.297	0.542	0.683
CP15	0.673	0.967	0.299	0.487	0.680
CP16	0.314	0.457	0.303	0.598	0.253
CP17	0.346	0.475	0.286	0.544	0.244
CP18	0.924	1.405	0.316	0.664	0.739
CP19	0.843	1.282	0.317	0.624	0.743
CP20	0.761	1.079	0.295	0.490	0.647

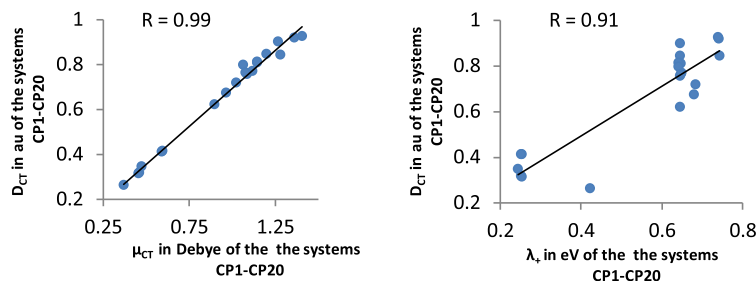


Fig. 10. The dependence of D_{CT} on μ_{CT} (left part) and on λ_+ (right part) for the systems CP1-CP20.

transfer length shows some scatter and varies in the same way as the dipole moment through the excitation process.

We hope that the proposed systems will be further studied experimentally to help upgrade the efficiency in solar-energy harvesting.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The corresponding author would like to thank the German Academic Exchange Service (DAAD), Bonn, Germany, for funding his research stay for university academics and scientists at the university of Saarland, 2018, funding-ID (57378441).

References

- [1] B. Kippelen, J.-L. Brédas, Organic photovoltaics, *Energy Environ. Sci.* 2 (2009) 251–261, <https://doi.org/10.1039/b812502n>.
- [2] B.C. Thompson, J.M.J. Fréchet, Polymer-fullerene composite solar cells, *Angew. Chem. Int. Ed.* 47 (2007) 58–77, <https://doi.org/10.1002/anie.200702506>.
- [3] P. Heremans, D. Cheyns, B.P. Rand, Strategies for increasing the efficiency of heterojunction organic solar cells: material selection and device architecture, *Acc. Chem. Res.* 42 (2009) 1740–1747, <https://doi.org/10.1021/ar9000923>.
- [4] M. Helgesen, R. Søndergaard, F.C. Krebs, Advanced materials and processes for polymer solar cell devices, *J. Mater. Chem.* 20 (2010) 36–60, <https://doi.org/10.1039/b913168j>.
- [5] J.A. Mikroyannidis, D.V. Tsakourinos, S.S. Sharma, Y.K. Vijay, G.D. Sharma, Conjugated small molecules with broad absorption containing pyridine and pyran units: synthesis and application for bulk heterojunction solar cells, *Org. Electron.* 11 (2010) 2045–2054, <https://doi.org/10.1016/j.orgel.2010.10.002>.
- [6] J.C. Bernède, L. Cattin, M. Morsli, S.R.B. Kanth, S. Patil, N. Stephant, Improvement of the efficiency of organic solar cells using the terthiophene-pyran-malononitrile (T3PM) as electron donor, through the use of a MoO_3/CuI anode buffer layer, *Layer Energy Procedia* 31 (2012) 81–88, <https://doi.org/10.1016/j.egypro.2012.11.168>.
- [7] M.C. Scharber, N.S. Sariciftci, Efficiency of bulk-heterojunction organic solar cells, *Prog. Polym. Sci.* 38 (2013) 1929–1940, <https://doi.org/10.1016/j.progpolymsci.2013.05.001>.
- [8] L. Lv, X. Wang, T. Dong, X. Wang, X. Wu, L. Yang, H. Huang, Significantly improving the efficiency of polymer solar cells through incorporating non-covalent conformational locks, *Mater. Chem. Front.* 1 (2017) 1317–1323, <https://doi.org/10.1039/C6QM00296j>.
- [9] W. Zhao, S. Li, H. Yao, S. Zhang, Y. Zhang, B. Yang, J. Hou, Molecular optimization enables over 13% efficiency in organic solar cells, *J. Am. Chem. Soc.* 139 (2017) 7148–7151, <https://doi.org/10.1021/jacs.7b02677>.
- [10] G. Yu, A.J. Heeger, Charge separation and photovoltaic conversion in polymer composites with internal donor/acceptor heterojunctions, *J. Appl. Phys.* 78 (1995) 4510–4515, <https://doi.org/10.1063/1.359792>.
- [11] G. Yu, J. Gao, J.C. Hummelen, F. Wudl, A.J. Heeger, Polymer photovoltaic cells: enhanced efficiencies via a network of internal donor-acceptor heterojunctions, *Science* 270 (1995) 1789–1791, <https://doi.org/10.1126/science.270.5243.1789>.
- [12] G. Dennler, M.C. Scharber, C.J. Brabec, Polymer-fullerene bulk-heterojunction solar cells, *Adv. Mater.* 21 (2009) 1323–1338, <https://doi.org/10.1002/adma.200801283>.
- [13] M.T. Lloyd, J.E. Anthony, G.G. Malliaras, Photovoltaics from soluble small molecules, *Mater. Today* 10 (2007) 34–41, [https://doi.org/10.1016/S1369-7021\(07\)70277-8](https://doi.org/10.1016/S1369-7021(07)70277-8).
- [14] B.P. Rand, J. Genoe, P. Heremans, J. Poortmans, Solar cells utilizing small molecular weight organic semiconductors, *Prog. Photovoltaics Res. Appl.* 15 (2007) 659–676, <https://doi.org/10.1002/ppp.788>.
- [15] J. Roncali, P. Frère, P. Blanchard, R. de Bettignies, M. Turbiez, S. Roquet, P. Leriche, Y. Nicolas, Molecular and supramolecular engineering of π -conjugated systems for photovoltaic conversion, *Thin Solid Films* 511–512 (2006) 567–575, <https://doi.org/10.1016/j.tsf.2005.12.014>.
- [16] N. Mataga, H. Chosrowjan, S. Taniguchi, Ultrafast charge transfer in excited electronic states and investigation into fundamental problems of exciplex chemistry: our early studies and recent developments, *J. Photochem. Photobiol., A* 6 (2005) 37–79, <https://doi.org/10.1016/j.jphotochem.2005.02.003>.
- [17] J. Hou, M.-H. Park, S. Zhang, Y. Yao, L.-M. Chen, J.-H. Li, Y. Yang, Band gap and molecular energy level control of conjugated polymer photovoltaic materials based on benzo[1,2-b:4,5-b']dithiophene, *Macromolecules* 41 (2008) 6012–6018, <https://doi.org/10.1021/ma800820r>.
- [18] L.M. Leung, W.Y. Lo, S.K. So, K.M. Lee, W.K. Choi, A high-efficiency blue emitter for small molecule-based organic light-emitting diode, *J. Am. Chem. Soc.* 122 (2000) 5640–5641, <https://doi.org/10.1021/ja000927z>.
- [19] B.R. Baker, R.Y. Lai, M.S. Wood, E.H. Doctor, A.J. Heeger, K.W. Plaxco, An electronic, aptamer-based small-molecule sensor for the rapid, label-free detection of cocaine in adulterated samples and biological fluids, *J. Am. Chem. Soc.* 128 (2006) 3138–3139, <https://doi.org/10.1021/ja056957p>.
- [20] B. Walker, A. Tamayo, J. Yang, J.Z. Brzezinski, T.-Q. Nguyen, Solution-processed small molecule-based blue light-emitting diodes using conjugated polyelectrolytes as electron injection layers, *Appl. Phys. Lett.* 93 (2008), 063302, <https://doi.org/10.1063/1.2968189>.
- [21] L. Duan, L. Hou, T.-W. Lee, J. Qiao, D. Zhang, G. Dong, L. Wang, Y. Qiu, Solution processable small molecules for organic light-emitting diodes, *J. Mater. Chem.* 20 (2010) 6392–6407, <https://doi.org/10.1039/b926348a>.
- [22] W. Ni, X. Wan, M. Li, Y. Wang, Y. Chen, A-D-A small molecules for solution-processed organic photovoltaic cells, *Chem. Commun. (J. Chem. Soc. Sect. D)* 51 (2015) 4936–4950, <https://doi.org/10.1039/C4CC09758K>.
- [23] X. Xu, R. Liu, L. Li, Nanoparticles made of π -conjugated compounds targeted for chemical and biological applications, *Chem. Commun. (J. Chem. Soc. Sect. D)* 51 (2015) 16733–16749, <https://doi.org/10.1039/C5CC06439B>.
- [24] J. Schill, A.P.H.J. Schenning, L. Brunsveld, Self-assembled fluorescent nanoparticles from π -conjugated small molecules: en route to biological applications, *Macromol. Rapid Commun.* 36 (2015) 1306–1321, <https://doi.org/10.1002/marc.201500117>.
- [25] F. Pfeiffer, G. Mayer, Selection and biosensor application of aptamers for small molecules, *Front. Chem.* 4 (2016) 1–21, <https://doi.org/10.3389/fchem.2016.00025>.
- [26] L. Zhou, L. Yang, M. Yu, Y. Jiang, C.-F. Liu, W.-Y. Lai, W. Huang, Inkjet-printed small-molecule organic light-emitting diodes: halogen-free inks, printing optimization, and large-area patterning, *ACS Appl. Mater. Interfaces* 9 (2017) 40533–40540, <https://doi.org/10.1021/acsami.7b13355>.
- [27] H. Bin, J. Yao, Y. Yang, I. Angunawela, C. Sun, L. Gao, L. Ye, B. Qiu, L. Xue, C. Zhu, C. Yang, Z.-G. Zhang, H. Ade, Y. Li, High-efficiency all-small-molecule organic solar cells based on an organic molecule donor with alkylsilyl-thienyl conjugated side chains, *Adv. Mater.* 30 (2018) 1–8, <https://doi.org/10.1002/adma.201706361>.
- [28] R. Ilmi, A. Haque, M.S. Khan, High efficiency small molecule-based donor materials for organic solar cells, *Org. Electron.* 58 (2018) 53–62, <https://doi.org/10.1016/j.orgel.2018.03.048>.
- [29] F. Guo, A. Karl, Q.-F. Xue, K.C. Tam, K. Forberich, C.J. Brabec, The fabrication of color-tunable organic light-emitting diode displays via solution processing, *Light Sci. Appl.* 6 (2017), e17094, <https://doi.org/10.1038/lsa.2017.94>.
- [30] V.P. Rao, A.K.-Y. Jen, K.Y. Wong, K.J. Drost, Novel push-pull thiophenes for

12

A.S. Khazaal et al. / Journal of Molecular Graphics and Modelling 100 (2020) 107654

- second order nonlinear optical applications, *Tetrahedron Lett.* 34 (1993) 1747–1750, [https://doi.org/10.1016/S0040-4039\(00\)60768-2](https://doi.org/10.1016/S0040-4039(00)60768-2).
- [31] P.R. Varanasi, A.K.-Y. Jen, J. Chandrasekhar, I.N.N. Namboothiri, A. Rathna, The important role of heteroaromatics in the design of efficient second-order nonlinear optical molecules: theoretical investigation on push-pull heteroaromatic stilbenes, *J. Am. Chem. Soc.* 118 (1996) 12443–12448, <https://doi.org/10.1021/ja960136q>.
- [32] J.-L. Reymond, M. Awale, Exploring chemical space for drug discovery using the chemical universe database, *ACS Chem. Neurosci.* 3 (2012) 649–657, <https://doi.org/10.1021/cn3000422>.
- [33] J. Greeley, T.F. Jaramillo, J. Bonde, I. Chorkendorff, J.K. Nørskov, Computational high-throughput screening of electrolytic materials for hydrogen evolution, *Nat. Mater.* 5 (2006) 909–913, <https://doi.org/10.1038/nmat1752>.
- [34] A.L. Briseno, T.W. Holcombe, A.I. Boukai, E.C. Garnett, S.W. Shelton, J.J.M. Fréchet, P. Yang, Oligo- and polythiophene/ZnO hybrid nanowire solar cells, *Nano Lett.* 10 (2010) 334–340, <https://doi.org/10.1021/nl9036752>.
- [35] T. Weymuth, M. Reiher, Inverse quantum chemistry: concepts and strategies for rational compound design, *Int. J. Quant. Chem.* 114 (2014) 823–837, <https://doi.org/10.1002/qua.24687>.
- [36] D. Xiao, I. Warnke, J. Bedford, V.S. Batista, Inverse molecular design for materials discovery, in: M. Springborg, J.-O. Joswig (Eds.), *Chemical Modelling*, vol. 10, Royal Society of Chemistry, Cambridge, UK, 2013, pp. 1–31, <https://doi.org/10.1039/9781849737241-00001>.
- [37] A. Franceschetti, A. Zunger, The inverse band-structure problem of finding an atomic configuration with given electronic properties, *Nature* 402 (1999) 60–63, <https://doi.org/10.1038/46995>.
- [38] M. d’Avezac, J.-W. Luo, T. Chanier, A. Zunger, Genetic-algorithm discovery of a direct-gap and optically allowed superstructure from indirect-gap Si and Ge semiconductors, *Phys. Rev. Lett.* 108 (2012), 027401, <https://doi.org/10.1103/PhysRevLett.108.027401>.
- [39] A. Zakutayev, Z. Zhang, A. Nagaraja, L. Yu, S. Lany, T.O. Mason, D.S. Ginley, A. Zunger, Theoretical prediction and experimental realization of new stable inorganic materials using the inverse design approach, *J. Am. Chem. Soc.* 135 (2013) 10048–10054, <https://doi.org/10.1021/ja311599g>.
- [40] X.-L. Zhao, T.C.W. Mak, Silver cages with encapsulated acetylenediide as building blocks for hydrothermal synthesis of supramolecular complexes with *n*-cyanopyridine and pyridine-*n*-carboxamide (*n* = 3,4), *Dalton Trans.* Issue 21 (2004) 3212–3217, <https://doi.org/10.1039/B406711H>.
- [41] C.T. Devaiah, B. Hemavathi, T.N. Ahipa, New blue emissive conjugated small molecules with low lying HOMO energy levels for optoelectronic applications, *Spectrochim. Acta A* 175 (2017) 222–228, <https://doi.org/10.1016/j.saa.2016.12.035>.
- [42] M. Springborg, S. Kohaut, Y. Dong, K. Huwig, Mixed Si-Ge clusters, solar-energy harvesting and inverse-design methods, *Comput. Theor. Chem.* 1107 (2017) 14–22, <https://doi.org/10.1016/j.comptc.2016.11.020>.
- [43] K. Huwig, C. Fan, M. Springborg, From properties to materials: an efficient and simple approach, *J. Chem. Phys.* 147 (2017), <https://doi.org/10.1063/1.5009548>, 234105.
- [44] J.H. Holland, *Adaptation in Natural and Artificial Systems: an Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*, Ph.D. Thesis, University of Michigan, Ann Arbor, MI, 1975.
- [45] J. Devillers, *J. Genetic Algorithms in Molecular Modeling*, Academic Press, London, 1996.
- [46] M. Grätzel, Recent advances in sensitized mesoscopic solar cells, *Acc. Chem. Res.* 42 (2009) 1788–1798, <https://doi.org/10.1021/ar900141y>.
- [47] C. Fan, M. Springborg, Y. Feng, Application of an inverse-design method to optimizing porphyrins in dye-sensitized solar cells, *Phys. Chem. Chem. Phys.* 21 (2019) 5834–5844, <https://doi.org/10.1039/C8CP07722C>.
- [48] R. Rüger, E. van Lenthe, T. Heine, L. Visscher, Tight-binding approximations to time-dependent density functional theory—A fast approach for the calculation of electronically excited states, *J. Chem. Phys.* 144 (2016), <https://doi.org/10.1063/1.4948647>, 184103.
- [49] D. Porezag, T. Frauenheim, T. Köhler, G. Seifert, R. Kaschner, Construction of tight-binding-like potentials on the basis of density-functional theory: application to carbon, *Phys. Rev. B* 51 (1995) 12947–12957, <https://doi.org/10.1103/physrevb.51.12947>.
- [50] G. Seifert, D. Porezag, T. Frauenheim, Calculations of molecules, clusters and solids with simplified LCAO-DFT-LDA scheme, *Int. J. Quant. Chem.* 58 (1996) 185–192, [https://doi.org/10.1002/\(sici\)1097-461x\(1996\)58:2<185::aid-qua7>3.0.co;2-u](https://doi.org/10.1002/(sici)1097-461x(1996)58:2<185::aid-qua7>3.0.co;2-u).
- [51] A.D. Becke, Density-functional thermochemistry. III. The role of exact exchange, *J. Chem. Phys.* 98 (1993) 5648–5652, <https://doi.org/10.1063/1.464913>.
- [52] B. Camino, M. de la Pierre, A.M. Ferrari, Photoelectrochemical properties of the CT1 dye : a DFT study, *J. Mol. Struct.* 1046 (2013) 116–123, <https://doi.org/10.1016/j.molstruc.2013.04.010>.
- [53] A. Irfan, R. Jin, A.G. Al-Sehemi, A.M. Asiri, Quantum chemical study of the donor-bridge-acceptor triphenylamine based sensitizers, *Spectrochim. Acta* 110 (2013) 60–66, <https://doi.org/10.1016/j.saa.2013.02.045>.
- [54] S. Jungsuttiwong, R. Tarsang, T. Sudyoasuk, V. Promarak, P. Khongpracha, S. Namuangruk, Theoretical study on novel double donor-based dyes used in high efficient dye-sensitized solar cells: the application of TDFT study to the electron injection process, *Org. Electron.* 14 (2013) 711–722, <https://doi.org/10.1016/j.orgel.2012.12.018>.
- [55] J. Tomasi, B. Mennucci, R. Cammi, Quantum mechanical continuum solvation models, *Chem. Rev.* 105 (2005) 2999–3093, <https://doi.org/10.1021/cr9904009>.
- [56] M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, G. Scalmani, V. Barone, G.A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A.V. Marenich, J. Bloino, B.G. Janesko, R. Gomperts, B. Mennucci, H.P. Hratchian, J.V. Ortiz, A.F. Izmaylov, J.L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V.G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J.A. Montgomery Jr., J.E. Peralta, F. Ogliaro, M.J. Bearpark, J.J. Heyd, E.N. Brothers, K.N. Kudin, V.N. Staroverov, T.A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A.P. Rendell, J.C. Burant, S.S. Iyengar, J. Tomasi, M. Cossi, J.M. Millam, M. Klene, C. Adamo, R. Cammi, J.W. Ochterski, R.L. Martin, K. Morokuma, O. Farkas, J.B. Foresman, D.J. Fox, *Gaussian 09 Revision D.01*, Gaussian Inc, Wallingford CT, 2013.
- [57] T. Lu, F. Chen, Multiwfn: a multifunctional wavefunction analyzer, *J. Comput. Chem.* 33 (2012) 580–592, <https://doi.org/10.1002/jcc.22885>.
- [58] T.L. Bahers, C. Adamo, I. Ciofini, A qualitative index of special extent in charge-transfer excitations, *J. Chem. Theor. Comput.* 7 (2011) 2498–2506, <https://doi.org/10.1021/ct200308m>.
- [59] J.-L. Brédas, D. Beljonne, V. Coropceanu, J. Cornil, Charge-transfer and energy-transfer processes in π -conjugated oligomers and polymers: a molecular picture, *Chem. Rev.* 104 (2004) 4971–5004, <https://doi.org/10.1021/cr040084k>.
- [60] H. Sadki, M. Bourass, M.N. Bennani, M. Bouachrine, New organic materials based on D- π -A structure for application in dye-sensitized solar cells, *Res. Chem. Intermed.* 44 (2018) 6071–6085, <https://doi.org/10.1007/s11164-018-3476-z>.
- [61] D. Jacquemin, E.A. Perpète, G. Scalmani, M.J. Frisch, I. Ciofini, C. Adamo, Absorption and emission spectra in gas-phase and solution using TD-DFT: formaldehyde and benzene as case studies, *Chem. Phys. Lett.* 421 (2006) 272–276, <https://doi.org/10.1016/j.cplett.2006.01.068>.
- [62] B. Lounis, M. Orrit, Single-photon sources, *Rep. Prog. Phys.* 68 (2005) 1129–1179, <https://doi.org/10.1088/0034-4885/68/5/r04>.
- [63] G. Nienhuis, C.T.J. Alkemade, Atomic radiative transition probabilities in a continuous medium, *Phys. B+C* 81 (1976) 181–188, [https://doi.org/10.1016/0378-4363\(76\)90256-4](https://doi.org/10.1016/0378-4363(76)90256-4).
- [64] R.G. Parr, W. Yang, Density functional approach to the frontier-electron theory of chemical reactivity, *J. Am. Chem. Soc.* 106 (1984) 4049–4050, <https://doi.org/10.1021/ja00326a036>.
- [65] W. Yang, R.G. Parr, Hardness, softness, and the Fukui function in the electronic theory of metals and catalysis, *Proc. Natl. Acad. Sci. Unit. States Am.* 82 (1985) 6723–6726, <https://doi.org/10.1073/pnas.82.20.6723>.
- [66] W. Yang, W.J. Mortier, The use of global and local molecular parameters for the analysis of the gas-phase basicity of amines, *J. Am. Chem. Soc.* 108 (1986) 5708–5711, <https://doi.org/10.1021/ja00279a008>.
- [67] J. Jayabarathi, V. Thanikachalam, N. Srinivasan, M.V. Perumal, Fluorescence spectral studies of some imidazole derivatives, *Spectrochim. Acta* 90 (2012) 125–130, <https://doi.org/10.1016/j.saa.2012.01.030>.
- [68] M. Malagoli, J.-L. Brédas, Density functional theory study of the geometric structure and energetics of triphenylamine-based hole-transporting molecules, *Chem. Phys. Lett.* 327 (2000) 13–17, [https://doi.org/10.1016/S0009-2614\(00\)00757-0](https://doi.org/10.1016/S0009-2614(00)00757-0).

Chapter 4

Summary

My research focused on designing molecules for properties by applying our inverse-design approaches to solar cells.

In the first stage, Dr. Huwig and I together developed PooMa (2.0) that can optimize organic molecules for a specific electronic property. PooMa (2.0) includes the automatic approach for generating molecules with the same backbone but having different substitutes, the genetic algorithm, the electronic-structure calculator, and an intensive property that shall be optimized. It can identify the optimum sets of functional groups for electronic properties related to dye-sensitized solar cells, including HOMO-LUMO gap, sunlight absorption, light harvest efficiency, orbital overlap, orbital distance, LUMO distribution, and the reorganization energy. Although the test system benzene was simple and far from practical applications, it is a benchmark that PooMa (2.0) offered one general solution to inverse design of organic molecules, compared with clusters as playgrounds in PooMa (1.0).

In the second stage, I improved our inverse-design approach to PooMa (2.1) that added a QSPR model (a mathematical relation) trained to predict the cell efficiency η from five electronic properties. Thus, our inverse-design approach could be applied to tackle the tremendous challenge to find optimal dye sensitizers in DSSC. The results led to the prediction of 15 porphyrins, none of which has been designed or studied before. DFT and TD-DFT methods were used to analyze the 15 porphyrins and it is found that their electronic properties were very promising. A pity is that they were too complicated to be synthesized successfully, so we didn't get experimental validation at the end. However, the purpose was achieved to identify optimal molecules through our inverse design in the large chemical space and provide suggestions for molecular design beyond chemical or physical intuition.

Later, I introduced PooMa (2.1) to Dr. Khazaal and guided him to apply it to his systems. Our inverse-design approach helped him to identify 20 thiophene oligomers for bulk-heterojunction solar cells and 20 cyanopyridone for organic solar cells. "It is clearly a strong

simplification to attempt to describe the performance of solar cells only using the properties of an individual molecule.” [10] Actually many factors can affect the cell performance, such as electrolyte, semi-conductor, the aggregation of molecules, etc. However, including more parts of solar cells in the modeling will increase the computational cost dramatically. Moreover, the rationale behind PooMa is not to identify exact molecules for solar cells, but to provide interesting suggestions for molecular design after exploring a larger part of chemical space than what usually is done [10].

In the third stage, I improved our inverse-design approach to PooMa (3.0) which used an easier and more efficient way to describe and construct molecules through SMILES (Simplified Molecular Input Line Entry System). Thereby, almost every class of materials — even macro-molecules or monomers — can be studied easily. PooMa (3.0) adopted the genetic algorithm that is more adaptive to SMILES while keeping its efficiency, and calculated seven important molecular properties that all have impacts on the η , including sunlight absorption A, the absorption energy corresponding to the wavelength for which the absorption maximum is found λ_{max} , the light harvest efficiency LHE, the degree of unsaturation DU, the dipole moment of ground state μ , the HOMO-LUMO energy gap G and the thermo-dynamic driving force for electron injection ΔG_{inj} . [15] Since this approach also take into account the self-consistent charge, the new electronic-structure calculator is more accurate for the study of excited states and charge transfer processes with acceptable computational cost.

By making use of the PooMa (3.0), I was able to easily study the assumption that in organic (often conjugated) molecular design the same sets of functional groups for similar backbones are identified as providing the best materials. The results show that for A, λ_{max} , LHE and G, it has remarkable impacts on the values of the properties and/or the optimal sets of functional groups by replacing 0, 1, or 2 CH groups of benzene backbone with N atoms. But for other properties, these effects are less pronounced. In the future when attempting to apply PooMa (3.0) on identifying breakthrough dye sensitizers in DSSC or other solar cells, it needs a QSPR model as in PooMa (2.1) to predict the η from the currently determined seven electronic properties.

Although we have used DSSC as the motivation and playground, our inverse-design approach is not limited to solar cells or optical properties. In principle, every conceivable property can be optimized and other areas can be explored, such as heterogeneous catalysis, hydrogen storage, etc. [12]

References

- [1] A. Hagfeldt, G. Boschloo, L. Sun, L. Kloo and H. Pettersson, *Chemical Reviews*, 2010, **110**, 6595–6663.
- [2] DAYLIGHT CIS Inc., 3. *SMILES - A Simplified Chemical Language*, <https://www.daylight.com/dayhtml/doc/theory/theory.smiles.html>, Online; accessed 20-May-2020.
- [3] Q. Deng, L. Liu and H. Deng, *Spectroscopy Analysis*, Science Press, Beijing, 2003.
- [4] M. Kokkonen, P. Talebi, J. Zhou, S. Asgari, S. A. Soomro, F. Elsehrawy, J. Halme, S. Ahmad, A. Hagfeldt and S. G. Hashmi, *Journal of Materials Chemistry A*, 2021, **9**, 10527–10545.
- [5] IEA, *World Energy Outlook 2020*, <https://www.iea.org/reports/world-energy-outlook-2020>.
- [6] S. Rühle, *Solar Energy*, 2016, **130**, 139–147.
- [7] A. D. Vos and H. J. Pauwels, *Applied physics*, 1981, **25**, 119–125.
- [8] S. Mathew, A. Yella, P. Gao, R. Humphry-Baker, B. F. Curchod, N. Ashari-Astani, I. Tavernelli, U. Rothlisberger, M. Nazeeruddin, M. Grätzel *et al.*, *Nature chemistry*, 2014, **6**, 242–247.
- [9] D. Zhang, M. Stojanovic, Y. Ren, Y. Cao, F. T. Eickemeyer, E. Socie, N. Vlachopoulos, J.-E. Moser, S. M. Zakeeruddin, A. Hagfeldt *et al.*, *Nature communications*, 2021, **12**, 1–10.
- [10] C. Fan, M. Springborg and Y. Feng, *Physical Chemistry Chemical Physics*, 2019, **21**, 5834–5844.
- [11] A. Franceschetti and A. Zunger, *Nature*, 1999, **402**, 60–63.
- [12] K. Huwig, C. Fan and M. Springborg, *The Journal of Chemical Physics*, 2017, **147**, 234105.
- [13] V. Venkatraman, S. Abburu and B. K. Alsberg, *Physical Chemistry Chemical Physics*, 2015, **17**, 27672–27682.
- [14] N. M. O’Boyle, C. M. Campbell and G. R. Hutchison, *The Journal of Physical Chemistry C*, 2011, **115**, 16200–16210.

- [15] C. Fan, M. Molayem, M. Springborg, M. Kick and Y. Feng, *The Journal of Physical Chemistry A*, 2022, **126**, 1289–1299.
- [16] D. Douguet, E. Thoreau and G. Grassy, *Journal of Computer-Aided Molecular Design*, 2000, **14**, 449–466.
- [17] M. Springborg, S. Kohaut, Y. Dong and K. Huwig, *Computational and Theoretical Chemistry*, 2017, **1107**, 14–22.
- [18] A. S. Khazaal, M. Springborg, C. Fan and K. Huwig, *Computational Condensed Matter*, 2020, **25**, e00503.
- [19] A. S. Khazaal, M. Springborg, C. Fan and K. Huwig, *Journal of Molecular Graphics and Modelling*, 2020, **100**, 107654.
- [20] A. Kumar, *Genetic Algorithms*, <https://www.geekforgeeks.org/genetic-algorithms/>, Online; accessed 20-May-2022.
- [21] X. Yang, *Nature-Inspired Optimization Algorithms (Second Edition)*, Academic Press, London, 2021.
- [22] Wikipedia contributors, *Genetic algorithm* — *Wikipedia, The Free Encyclopedia*, 2022, https://en.wikipedia.org/w/index.php?title=Genetic_algorithm&oldid=1081816519, [Online; accessed 13-May-2022].
- [23] B. Hourahine, B. Aradi, V. Blum, F. Bonafe, A. Buccheri, C. Camacho, C. Cevallos, M. Deshayé, T. Dumitrică, A. Dominguez *et al.*, *The Journal of Chemical Physics*, 2020, **152**, 124101.
- [24] G. Seifert, D. Porezag and T. Frauenheim, *International journal of quantum chemistry*, 1996, **58**, 185–192.
- [25] D. Porezag, T. Frauenheim, T. Köhler, G. Seifert and R. Kaschner, *Physical Review B*, 1995, **51**, 12947.
- [26] M. Elstner, D. Porezag, G. Jungnickel, J. Elsner, M. Haugk, T. Frauenheim, S. Suhai and G. Seifert, *Physical Review B*, 1998, **58**, 7260.
- [27] R. Rüger, E. Van Lenthe, T. Heine and L. Visscher, *The Journal of chemical physics*, 2016, **144**, 184103.
- [28] Wikipedia contributors, *Chromophore* — *Wikipedia, The Free Encyclopedia*, 2022, <https://en.wikipedia.org/w/index.php?title=Chromophore&oldid=1089015958>, [Online; accessed 12-June-2022].
- [29] M. Springborg and M. Zhou, *Quantum Chemistry*, De Gruyter, 2021.
- [30] Wikipedia contributors, *Hartree–Fock method* — *Wikipedia, The Free Encyclopedia*, 2022, https://en.wikipedia.org/w/index.php?title=Hartree%E2%80%93Fock_method&oldid=1073176129, [Online; accessed 6-June-2022].

- [31] Wikipedia contributors, *Kohn–Sham equations* — *Wikipedia, The Free Encyclopedia*, 2022, https://en.wikipedia.org/w/index.php?title=Kohn%E2%80%93Sham_equations&oldid=1070831947, [Online; accessed 12-June-2022].
- [32] H. Huang, *Ph.D. thesis*, Queen’s University Belfast, School of Chemistry and Chemical Engineering, 2020.
- [33] Wikipedia contributors, *Local-density approximation* — *Wikipedia, The Free Encyclopedia*, 2022, https://en.wikipedia.org/w/index.php?title=Local-density_approximation&oldid=1090094128, [Online; accessed 12-June-2022].
- [34] R. G. Parr and W. Yang, *Journal of the American Chemical Society*, 1984, **106**, 4049–4050.
- [35] C. Fan, B. Zhang, Y. Li, Y. Liang, X. Xue and Y. Feng, *Physical Chemistry Chemical Physics*, 2015, **17**, 30624–30631.
- [36] T. Lu, *Software Manual of Multiwfn (A Multifunctional Wavefunction Analyzer) version 3.7*, <https://sobereva.com/multiwfn>, Online; accessed 14-Aug-2020.
- [37] T. Le Bahers, C. Adamo and I. Ciofini, *Journal of Chemical Theory and Computation*, 2011, **7**, 2498–2506.

