

Data-Driven Tailoring of Molecular Dipole Polarizability and Frontier Orbital Energies in Chemical Compound Space

Szabolcs Góger, Leonardo Medrano Sandonas, Carolin Müller, and Alexandre Tkatchenko*

University of Luxembourg, Department of Physics and Materials Science, L-1511 Luxembourg City, Luxembourg

E-mail: alexandre.tkatchenko@uni.lu

Abstract

Understanding correlations – or lack thereof – between molecular properties is crucial for enabling fast and accurate molecular design strategies. In this contribution, we explore the relation between two key quantities describing the electronic structure and chemical properties of molecular systems: the energy gap between the frontier orbitals and the dipole polarizability. Based on the recently introduced QM7-X dataset, augmented with accurate molecular polarizability calculations as well as analysis of functional group compositions, we show that polarizability and HOMO–LUMO gap are uncorrelated when considering sufficiently extended subsets of the chemical compound space. The relation between these two properties is further analyzed on specific examples of molecules with similar composition as well as homooligomers. Remarkably, the freedom brought by the lack of correlation between molecular polarizability and HOMO–LUMO gap enables the design of novel materials, as we demonstrate on the example of organic photodetector candidates.

1 Introduction

Data-driven molecular design is an increasingly pursued strategy in chemical physics and computational chemistry. The search for novel molecules with tailored physicochemical properties for a given functionality is continuously motivating the development of a great variety of computer-aided molecular design approaches^{1–3}. The ultimate goal is to establish a feasible protocol that can be used for exploring the chemical compound space (CCS) through systematic targeting of physical properties. Physicochemical quantities, such as color, conductivity, excited state lifetime, electron affinity, ionization potential, and solubility, are commonly used in the design of molecular photosensitizers or op-

toelectronic devices, for instance^{4–7}. Given the complexity of a multi-property design task, it is essential to first have a solid grasp of the physical relationships between the various target properties⁸.

Within this context, two fundamental quantum-mechanical (QM) electronic properties are the optical gap and the molecular dipole polarizability (α). Optical gap is an experimental property that measures the energy corresponding to the lowest observed optical transition. Many computational studies use the HOMO–LUMO gap ΔE_{HL} (the difference between the energies of frontier molecular orbitals in the ground state) as a starting point in approximating experimental optical gaps. HOMO–LUMO gap is also related to chemical reactivity and excitation energies; and thus plays an important role in the prediction of optical absorption spectra, refractive indices and conductivity^{9–12}. For correctness of terminology, HOMO–LUMO gap obtained from density functional calculations should be referred to as Kohn-Sham (KS) gap. Although the relations between different gaps (Kohn-Sham, fundamental, and optical) are subtle and have been discussed in detail in the literature^{13,14}, in this manuscript we will use the KS gap as a proxy for observable experimental properties.

The molecular dipole polarizability α (referred to simply as polarizability in the manuscript), on the other hand, describes the dipolar response of a molecule to an external electric field, becoming a key quantity for understanding intra- and intermolecular interactions (*e.g.*, dispersion interactions, substituent and solvent effects as well as supramolecular structure formation) and for determining spectroscopic properties of molecules (Raman and sum frequency spectroscopy)^{15–21}. These features make both polarizability and HOMO–LUMO gap essential in the derivation of structure-property/property-property relationships and, consequently, in the development of design strategies for molecules with a targeted array of QM

properties for applications such as molecular dyes²², optoelectronic devices²³, molecular junctions^{24,25}, heterogeneous catalysts²⁶ and materials for non-linear optics^{27,28}.

Various computational methods and predictive models have been developed to estimate HOMO–LUMO gaps and polarizabilities for organic molecules with different levels of tradeoff between accuracy and computational cost^{29–32}. Lately, it has become feasible to access a plethora of highly-accurate QM properties – including ΔE_{HL} and α – for large swaths of the chemical compound space (CCS)^{33–37}. Comprehensive analyses of these extensive datasets may help to understand the deeper physical picture behind the inherent property-property relationships. With this motivation, we herein perform an exhaustive investigation of the two-dimensional space defined by HOMO–LUMO gap and polarizability (*i.e.*, $(\Delta E_{\text{HL}}, \alpha)$ -space) for small organic molecules with the aim to gain insights into the intrinsic relationship between these two properties. We find that while correlation might appear in homologous molecules, if a large enough subspace of CCS is considered, HOMO–LUMO gap and polarizability are essentially uncorrelated and their 2D space is represented as a structureless “blob”. Through the analysis of diverse molecular sets, it is shown that this lack of correlation can be related to the fact that the polarizability is primarily determined by the atomic composition, while the HOMO–LUMO gap heavily depends on the arrangement of the atoms into chemical functional groups. Hence, we expect that our findings will assist the development of novel design principles in which the control of multiple electronic properties is relevant, as we finally demonstrate on the case of molecular photodetectors.

The outline of the paper is as follows: in Sec. 2, we review accurate and approximate models for polarizability and HOMO–LUMO gap. In Sec. 3, we exhaustively examine the polarizabilities (α) and HOMO–LUMO gaps (ΔE_{HL}) of diverse molecular sets. In doing this, we have extended the QM7–X dataset³⁷ with functional group information as well as polarizabilities calculated with the hybrid PBE0 functional. In assessing our computational setting, we tested the predictive power of this functional against coupled cluster CCSD(T) calculations, and found an overall accuracy of 1.9%. As a first order approximation to predicting polarizabilities of small organic systems, we consider a linear combination of atomic contributions in Sec. 3.1. In Sec. 3.2, we then perform PBE0 calculations of polarizability (see Sec. 6.1 for computational details) for homologous molecules and explore the relationship with their HOMO–LUMO gaps. A statistical analysis of the $(\Delta E_{\text{HL}}, \alpha)$ -space using a subset of molecules contained in QM7–X dataset is carried out in Sec. 3.3. Our proposed design principle is further discussed and demonstrated on the case of organic photodetectors, see Sec. 4. The computational methods as well as the dataset used are presented in Sec. 6, follow-

ing the main conclusions of the manuscript in Sec. 5.

2 Models for Polarizability and Frontier Orbital Energy Gap

Since our main focus is to have a better understanding of the relationship between polarizability and HOMO–LUMO gap in organic molecules, we first revisit different qualitative and quantitative models employed to compute them. Typically, accurate electronic structure methods, such as coupled cluster or hybrid DFT are used to determine these properties³⁸. Practically, any mean-field electronic structure method allows to calculate a HOMO–LUMO gap, but the results are sensitive to the method chosen³⁹. Moreover, since the Kohn-Sham gap is not an experimental quantity, the ambiguity regarding the accuracy of different methods is not trivial to resolve. Polarizability (α) is typically obtained from finite field, coupled perturbed Hartree-Fock or density functional perturbation theory (DFPT) calculations^{29,30,40}. However, these electronic structure methods need considerable computational resources when dealing with larger molecules or significant swaths of the CCS. Accordingly, we will next discuss alternative physical models, empirical correlations as well as approximate methods to obtain these QM properties. We will start with examining the polarizability, for which analytical models (such as the quantum Drude oscillator, or QDO) as well as empirical correlations and predictive semiempirical methods are available. After this, the models for HOMO–LUMO gap will be mentioned, before concluding the section by analyzing what is known about the correlation between these two quantities.

A connection between HOMO–LUMO gap and polarizability can be anticipated starting from the perturbative expression for polarizability using the dipole moment operator $\hat{\mu}$ within second order perturbation theory as^{17,41}

$$\overleftrightarrow{\alpha} = 2 \sum_{n \neq 0}^{\infty} \frac{\langle \Psi_0 | \hat{\mu} | \Psi_n \rangle \otimes \langle \Psi_n | \hat{\mu} | \Psi_0 \rangle}{E_n - E_0}, \quad (1)$$

where Ψ_0 and E_0 are the ground state wavefunction and energy, respectively, and n is the index of the excited states. Indeed, since $\Delta E_{\text{HL}} = E_1 - E_0$ is commonly much smaller than the energy gap of higher excited states, the first term of the sum in Eq. 1 provides a first-order approximation to the infinite series and, hence, there could exist an inversely proportional relationship between ΔE_{HL} and α , *i.e.*, $\alpha \propto (\Delta E_{\text{HL}})^{-1}$.

Equation 1 can only be analytically evaluated for simple model systems (such as the hydrogen atom or a quantum Drude oscillator). For many-electron systems, the sum can only be evaluated numerically and requires including bound-bound and bound-continuum transi-

tion dipoles⁴². Modeling atoms or larger coarse-grained fragments with QDOs and solving the dipole-dipole screening equations is known to be an effective method to predict polarizability, and it is also the basis for the Many-Body Dispersion (MBD) method^{43,44}. Since the response properties of all atoms and molecules can be represented by QDOs by carefully setting the three parameters {charge q , frequency ω , mass μ } of the model, the analysis of the polarizability of the QDO Hamiltonian should generally be transferable to any systems. Therefore, we devote some attention to this model.

Due to selection rules of the dipole operator, only the first excited state contributes to the dipole polarizability of a QDO^{41,45}, making it effectively a two-state system

$$\alpha_{\text{QDO}} = 2q^2 \frac{\langle \Psi_0 | \hat{\mu} | \Psi_1 \rangle \langle \Psi_1 | \hat{\mu} | \Psi_0 \rangle}{E_1 - E_0} = \frac{q^2}{\mu\omega^2}, \quad (2)$$

where q is the magnitude of the charge bound by a harmonic potential with frequency ω , having a mass μ . The HOMO–LUMO gap of a QDO is $\Delta E_{\text{HL}} = \hbar\omega$, which indeed appears in the denominator. However, α can be separately controlled through the other two individual QDO parameters $\{q, \mu\}$, independently from ΔE_{HL} . This means that for the QDO model, the polarizability and the HOMO–LUMO gap are mutually related, yet they could be tuned separately from each other.

The idea of approximating the polarizability using an effective two-state system (so-called Unsöld approximation)^{41,46} is also useful for understanding qualitative trends. Within this approximation, polarizability is written using an average excitation ΔE as a fitting parameter

$$\overleftrightarrow{\alpha} = \frac{2}{\Delta E} \sum_{n \neq 0} \langle \Psi_0 | \hat{\mu} | \Psi_n \rangle \otimes \langle \Psi_n | \hat{\mu} | \Psi_0 \rangle. \quad (3)$$

Setting the average excitation to ΔE_{HL} is therefore exact for the QDO model, but the connection between these quantities for many-electron systems is not known in general⁴⁷.

Investigating correlations between polarizability and various molecular properties can lead to useful relationships such as the recent observation that polarizability scales with the fourth power of the characteristic size of the system⁴¹. The correlation between polarizability and orbital energies is relevant from a theoretical standpoint, as it forms the foundation of Pearson’s hard-soft acid-base (HSAB) theory^{48,49}. Based on recent theoretical works, we can postulate that polarizability can be expressed as a function of two factors accounting for *i*) ground state geometry (e.g., van der Waals radius or molecular volume) and *ii*) electronic structure (e.g., ionization energy or hardness)^{27,41,50–55}. While these correlations provide useful conceptual insights, they have not been put to use for constructing accurate numerical predictions.

There are two types of predictive models for polar-

izability with lower computational burden than electronic structure calculations. Firstly, approximations for polarizability can be constructed based on the group contribution principle, which divides polarizability into atomic or bond contributions^{26,56}. These models can offer somewhat accurate predictions with minimal molecular information and computational effort, and we will assess such models in this work. As a second approach, machine learning (ML) models have been proposed as a cost-effective solution with improved accuracy³¹. However, the training process and accuracy of the ML models are strongly dependent on the features of the dataset (e.g., chemical diversity, molecular size, number of samples) as well as on the ML method itself.

For the case of HOMO–LUMO gap, there is a well-established underlying physical principle in determining this property: it is known that the HOMO–LUMO gap of individual functional groups (called chromophores in this context) is transferable, with values documented in standard reference texts⁵⁷. These chromophores also form the foundation for both accurate ML models and earlier empirical rules for the prediction of HOMO–LUMO gaps^{58,59}. The HOMO–LUMO gap of a single functional group can be understood based on the molecular orbital theory, the most common version of which is the Hückel theory for conjugated systems. For instance, the inverse proportionality between the number of monomers and the HOMO–LUMO gap of polyenes is well-explained within this theory⁶⁰. In the case of non-interacting functional groups, their optical spectra are effectively independent and, consequently, the frontier energy gap of a molecule is determined by the lowest value for the constituent functional groups, making it an inherently size-independent (intensive) property.

In agreement with the analysis of the QDO model, recent studies relying on large datasets (7 k structures from the GDB-13 dataset as well as the tmQM dataset of 86 k transitional metal complexes) suggest that there’s no overall correlation between HOMO–LUMO gap and polarizability^{61,62}. Nevertheless, correlation has been observed both experimentally and computationally for different classes of structures (e.g., organic dyes and inorganic clusters^{22,63–65}), with notable exception of smaller systems where the HOMO–LUMO transition is symmetry forbidden²¹. In the following section, we explore the source of such seemingly contradictory results by showing that investigating a reduced subset of the chemical compound space can lead to correlations between quantities that are generally uncorrelated.

3 Results and Discussion

3.1 First Order Linear Atomic Additive Model for Polarizability

The simplest atomic additive method (motivated by that of Bosque³²) approximates the polarizability of a

Table 1: Revised linear regression parameters for the atomic additive polarizability model of Bosque *et al.* Note that the values in the original paper are presented in \AA^3 , whereas the values here are in bohr^3 . The relatively low influence of the intercept can be seen by comparing the last two rows; the rest of our manuscript uses the parameters presented under "This work"

	Intercept	C	Cl	H	N	O	S
Bosque	2.14	10.20	14.60	1.17	6.95	3.85	20.20
This work	1.71	10.10	12.70	0.87	7.88	4.00	19.10
No intercept	0.00	10.37	13.00	0.88	8.11	4.24	19.37

molecule via a linear combination of the number of each atom-types n weighted with a type-specific factor C_i , together with an intercept m

$$\alpha = m + \sum_i C_i n_i. \quad (4)$$

Bosque’s model was fit directly using experimental data of 426 compounds. The fitted m and C_i values for C, Cl, H, N, O, and S are listed in Table 1. Accordingly, we have here used the QM7-X dataset³⁷ (see Sec. 6) to validate the accuracy and reassess the model parameters on a significantly larger swath of the CCS. In doing so, we have considered the first conformer for each entry in the QM7-X dataset; a total of ≈ 13 k structures. The linear regression parameters optimized on QM7-X yield the results listed in Table 1. Bosque’s parameters hold up relatively well for QM7-X molecules, accounting for a correlation coefficient (R^2 value) of 0.65 with a mean absolute percentage error (MAPE) of 6.11 %. However, the re-fitted parameters improve the correlation coefficient to 0.72 and reduce MAPE value to 3.94 %, *i.e.*, the prediction accuracy is increased by a factor of 1.6. Note that the presence of the intercept m in Eq. 4 is just an artifact of the model, since the prediction should be zero when no atoms are present. Inclusion or omission of m , however, changes neither the goodness of the regression, nor the numerical value of the atomic contributions to a meaningful degree (the mean absolute error of the linear model goes from 3.078 a.u. to 3.079 a.u.; similarly to what had also been observed by the authors). Therefore, we decided to include the intercept in our further analysis, to be consistent with Bosque’s approach.

A shortcoming of atomic additive methods is that the same polarizability is predicted for all structural isomers, since only the total number of each atom-types is used in the prediction. This is manifested in Fig. 1 as having a systematic error within each possible α_{pred} value, and further demonstrated in the inset on the case of molecules with chemical formula $\text{C}_6\text{H}_8\text{O}$. Indeed, the reference polarizabilities for this given chemical formula span a range of 30 a.u., but the predicted value is 73.4 a.u. for all molecules, irrespective of the chemical arrangement of the atoms. From Fig. 1, it can also be inferred that such a simple additive model will only become worse for molecules of increasing size. Indeed, a trend appears where larger molecules exhibit stronger deviations towards higher polarizabilities – a trend that

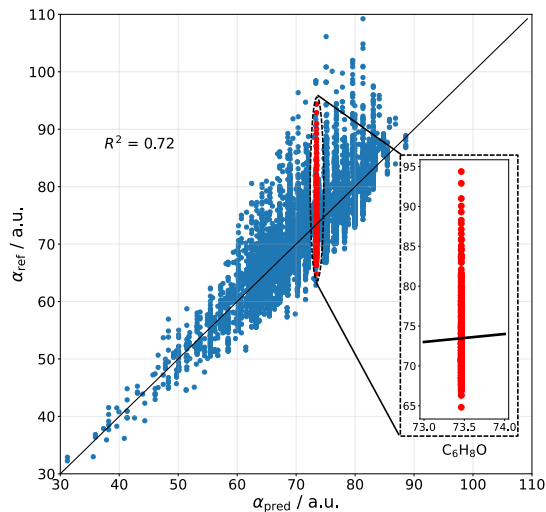


Figure 1: Performance of the atomic additive method. The linear regression parameters are fit utilizing our dataset of 13k molecules computed at the DFT-PBE0 level of theory (subset of the QM7-X dataset³⁷, see Fig. S 3 of ESI[†]). The inset shows the inherent shortcoming of the model, predicting the same polarizability for all $\text{C}_6\text{H}_8\text{O}$ isomers.

an additive model is unable to describe. This can be especially the case for polymeric molecules which form long chains, whose polarizability is highly anisotropic and it behaves increasingly non-additively with size.

To differentiate between structural isomers, a descriptor that accounts for different geometric properties (for example, radius of gyration) might be constructed, since polarizability is an extensive property^{52,55}. This extensivity is only partially captured by atomic additive methods, insofar as increasing the number of atoms in a molecule is inherently increasing the size as well. More accurate models should also differentiate between similar atoms based on their surrounding chemical environments, as it is done for example in the self-consistent screening approach used in the Many-Body Dispersion (MBD) method^{43,44}. Therefore, while the shown first-order linear model is limited by its accuracy, it can serve as a baseline for more accurate methods involving coupling between atoms in a molecule.

To summarize, a first-order approximation to polarizability can be constructed just by using an atomic additive model without explicit knowledge of the molecular spatial arrangement or the local chemical environments. While the predictive power of such a model

is rather restricted, its rough correlation with reference electronic-structure calculations (see Fig. 1) gives a clear evidence that a significant fraction of the polarizability is determined by just the atomic composition.

3.2 Case Studies for the Relation between the HOMO-LUMO Gap and Polarizability

Experimental studies often focus on examining molecules with similar electronic structures, leading to hidden correlations between optical gap and polarizability. In our path towards the general understanding of the relationship between these QM properties, we now examine two different cases: (i) molecules having the same atom-type composition but slightly different chemical compositions and (ii) molecules with the chemical properties fixed while increasing the system size (*e.g.*, oligomers). In Sec. 3.3, we discuss a broader case in which a large subset of CCS is analyzed.

Constitutional isomers. In general, the functional groups in a molecule govern the nature and order of the molecular orbitals, determining the HOMO–LUMO gap and the orbitals involved in the electronic transitions. To explore the relationship between HOMO–LUMO gap and polarizability as a function of chemical functionality, we present select examples of constitutional isomers, *i.e.*, molecules with the same atomic composition that belong to different substance classes due to the presence of different functional groups.

As a first example, two constitutional isomers with the formula C_5H_8O , namely an α,β - (3-penten-2-one) and a β,γ -unsaturated enone (4-penten-2-one) is considered (see Fig. 2(a)). It is noticeable that 3-penten-2-one has a smaller gap compared to 4-penten-2-one (by 1.12 eV) because delocalization results in a greater mobility of π -electrons throughout the molecular structure. However, both molecules present a similar polarizability coming from the identical atomic composition as well as similar total size.

A second set of constitutional isomers with the formula $C_8H_{14}O$ was constructed for molecules bearing a C=O (oxo-group) and C=C (alkene-group) on an octane backbone. These isomers are thus formed by the following substance classes: one ketene, one conjugated aldehyde, four conjugated ketones, five non-conjugated aldehydes and eight non-conjugated ketones (see Fig. 2(b)). While these structures are chemically quite different, their orbital symmetries are largely similar, leading to a correlation between their polarizability and HOMO–LUMO gap. Notice, however, that the polarizabilities of the structures are all within 4% of each other, whereas the variation of HOMO–LUMO gap is about five times larger. As such, the statement that polarizability is mainly determined by the atomic composition and HOMO–LUMO gap by the chemical composition seems to hold, even though some correlation between these two quantities is observed due to

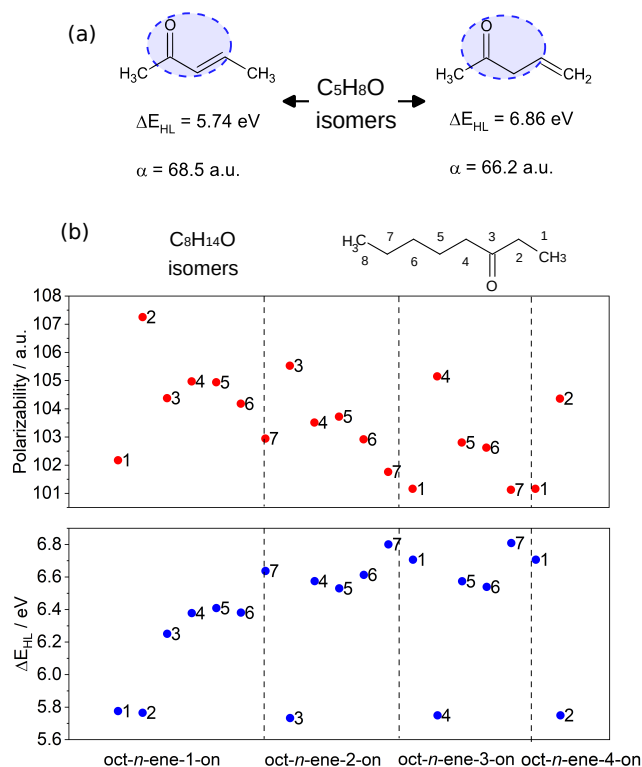


Figure 2: Two molecules, having the same atomic composition but different chemical properties show similar polarizabilities but different HOMO–LUMO gaps (a). HOMO–LUMO gap and polarizability of all possible linear structures having eight carbon atoms, an oxo group and a double bond between two of the carbons (the values shown are results of PBE0 calculations as described in Sec. 6.1). The numbering of carbon atoms is shown in the case of octane-3-on (b).

the similarity of the structures.

Homologous series of molecules. As previously elaborated, HOMO–LUMO gap and polarizability can seemingly correlate for molecules that belong to a homologous series. This can be explained by the fact that the electronic nature and order of the frontier orbitals is often identical for structurally and electronically similar molecules. Consequently, the decrease in the HOMO–LUMO gap can correlate with the increase in polarizability when considering molecules of a homologous series with an increasing number of repeating units. To support this assumption, we consider in the following a series of oligomers, namely alkanes (C_mH_{2m+2}) and alkenes (C_mH_{2m} ; see Fig. 3). The example is taken from Afzal *et al.*⁶⁶, with polarizability and HOMO–LUMO gap recalculated within our computational setup (*cf* Sec. 6).

Fig. 3 shows a decreasing behaviour of HOMO–LUMO gap for oligoethylene and oligoacetylene as a function of the number of monomers n , in agreement with previous works as well as qualitative predictions from the Hückel model^{60,67}. Indeed, we have found that the absence of a qualitative change to the electronic structure within the ethylene oligomers leads

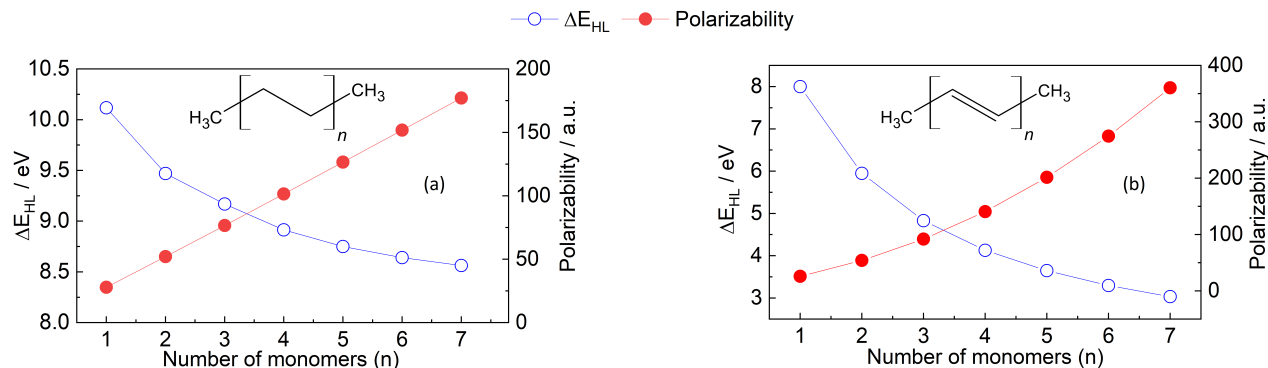


Figure 3: HOMO–LUMO gap (blue dots) and polarizability (red dots) of the oligomers of (a) ethylene and (b) acetylene. The calculations of both properties were carried out as described in Sec. 6.1.

to a relatively small HOMO–LUMO gap change going from $n = 1 \rightarrow 7$ (≈ 1.5 eV) compared to the acetylene oligomers, where every monomer modifies the conjugation, producing a more significant change of ≈ 5.0 eV. Unlike HOMO–LUMO gap, the behavior of polarizability in the molecular chains can not be simply explained. The observation that polarizability monotonously increases with n is in line with both the principles of atomic additive models and the correlation with molecular size. However, the absolute magnitude of the polarizability values is significantly different for the two sets of oligomers, and this difference increments with increasing the number of monomers. This quantitative difference can neither be explained by atomic additive models nor correlations using molecular size, but it correlates with the reductions in HOMO–LUMO gap. Notably, not even AlphaML³¹ can predict this behavior: the model predicts 181 a.u. for oligoethylene and 227 a.u. for oligoacetylene in the $n = 7$ case, with the DFPT results being 177 a.u. and 360 a.u., respectively. The difference due to conjugation is therefore underestimated by a factor of four even when using ML methods, and this error is expected to increase with increasing the chain length. Up to now, we are not aware of any other simple polarizability estimation method that can accurately predict the values in Fig. 3. These findings provide clear evidence that further work is necessary to enhance our understanding and improve and accuracy of computational methods used for calculating polarizability, even for relatively simple molecules such as hydrocarbon oligomers.

3.3 Clustering of Structures in the $(\Delta E_{\text{HL}}, \alpha)$ -Space

All previously presented examples might suggest that there is a correlation between HOMO–LUMO gap and polarizability. However, these examples considered similar molecules with respect to their functionality or chemical composition — factors that essentially determine both the HOMO–LUMO gap and polarizability. From optical spectroscopy, it is known that the optical

gap is primarily determined by the functional groups in a molecule. This is reflected in characteristic optical gaps (vertical excitation energies of the lowest electronic transitions) per functional groups, *e.g.*, the $\pi\pi^*$ absorption of an isolated alkene-group as chromophore is between 7.51 and 6.70 eV. Since we are assuming that HOMO–LUMO gap is a good starting point in determining the optical gap of a molecule, it would be expected to find that ΔE_{HL} values are also clustered by certain functional groups. In contrast, our analysis has shown that polarizability (α) is primarily determined by the atomic composition of a molecule. To draw more general conclusions about these properties, we analyse the two-dimensional (2D) property space defined by HOMO–LUMO gap and polarizability for a selected subset of QM7–X molecules⁸ (see Sec. 6.2). This dataset enables us to study the $(\Delta E_{\text{HL}}, \alpha)$ relationship more broadly because it covers a considerable number and variety of chemical compounds.

Fig. 4(a) shows the $(\Delta E_{\text{HL}}, \alpha)$ -space for the QM7–X molecules – indicating no direct relationship between the two quantities across the chemical compound space spanned by this dataset ($R^2 = 0.13$). Furthermore, the role of the two main factors that determine ΔE_{HL} (functionality) and α (atomic composition) are highlighted in Fig. 4. The panels (b) and (c) exemplarily display the distributions of ΔE_{HL} and α for aldehydes and primary alcohols, *i.e.*, molecules that bear one of the respective functional groups. The subplots (d) and (e) show the respective distributions for molecules with equal atomic positions, namely with the molecular formulas $\text{C}_4\text{H}_8\text{O}$ and $\text{C}_4\text{H}_9\text{N}$, respectively.

Functional groups & HOMO-LUMO gap.

In Figs. 4(b,c), we highlight the frequency plots of ΔE_{HL} and α values for all non-conjugated aldehydes (blue) and primary alcohols (pink) of our select dataset. Fig. 4(c) clearly reflects the common notion of chromophores, namely that HOMO–LUMO gap is mainly determined by the type of chromophore (*e.g.*, aldehyde or primary alcohol group) and the character of the lowest energy electronic transition (*e.g.*, $n\pi^*$ - or $n\sigma^*$ -transition). Thus, ΔE_{HL} values for aldehydes only

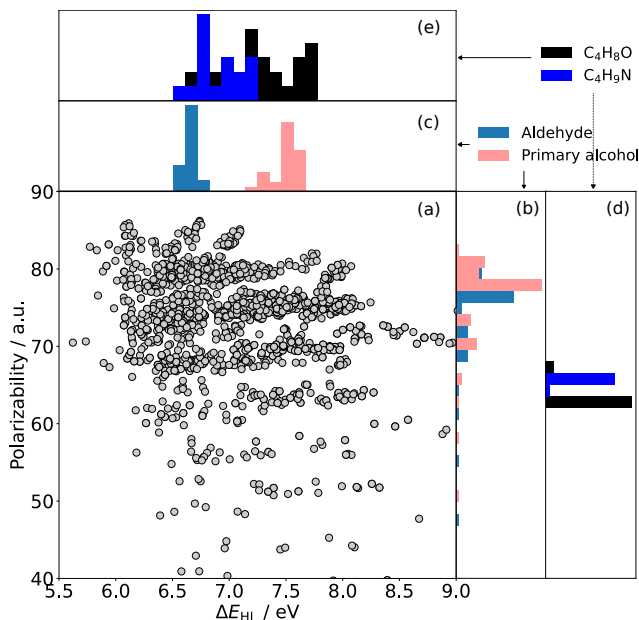


Figure 4: (a) Polarizability (α) vs. HOMO-LUMO gap (ΔE_{HL}) for molecules of the subset of QM7-X under study (see text). Histograms of the HOMO-LUMO gaps (b) for all non-conjugated aldehydes (blue) and primary alcohols (pink) and (d) for structures having the atomic composition $\text{C}_4\text{H}_8\text{O}$ (black) and $\text{C}_4\text{H}_9\text{N}$ (blue). Histograms of polarizabilities (c) for all non-conjugated aldehydes (blue) and primary alcohols (pink) and (e) for structures having the atomic composition $\text{C}_4\text{H}_8\text{O}$ (black) and $\text{C}_4\text{H}_9\text{N}$ (blue). The difference between the clustering in the two quantities is reflected in the degree of separation between the histograms.

show a value of *circa* 6.5 eV while, for primary alcohol group, they extend from 7.2 eV to 7.8 eV.

To fully explore the role of functional groups on the ($\Delta E_{\text{HL}}, \alpha$) relationship, the QM7-X molecules were categorized into twelve major classes based on the functional groups they are bearing (see Sec. 6.2). Fig. S 4[†] shows that the distribution of all detected functional groups in the dataset, confirming that ΔE_{HL} is clustered along the chemical properties of the molecules.

Unlike ΔE_{HL} , molecules containing aldehydes (blue) and primary alcohols (pink) exhibit polarizabilities that extend throughout the entire range of the dataset (*cf.* Fig. 4(b)). This finding is further reflected in the average Kolmogorov-Smirnov-metric (measuring the statistical distance between two general distributions, see Sec. II of ESI[†]) of the individual molecular classes in the ($\Delta E_{\text{HL}}, \alpha$)-space, which is 0.81 and 0.40 for ΔE_{HL} and α , respectively. Our analysis then demonstrates that functional groups primarily affect HOMO-LUMO gap rather than the polarizability, resulting in well-defined molecular clusters on the ΔE_{HL} -axis.

Atomic composition & polarizability.

According to the Kolmogorov-Smirnov analysis in Sec. II of the ESI[†], the functional groups only indirectly influence the magnitude of the polarizability in a given

molecule, whereas the atomic composition is a crucial factor for the determination of the polarizability. This finding is also in line with the fact that a good correlation is achieved between the first-order atomic additive model and the reference DFT data shown in Fig. 1. The ΔE_{HL} and α values for a set of two constitution isomers, namely with the chemical formula $\text{C}_4\text{H}_8\text{O}$ (including aldehydes, dialkyl ethers, enol ethers, as well as primary and secondary alcohols) and $\text{C}_4\text{H}_9\text{N}$ (including carbonitriles and primary/secondary aliphatic amines) is also presented in Figs. 4(d), showing a narrow polarizability distribution. These results are another clear evidence that α , to a reasonable approximation, is independent of the actual chemical arrangement of the atoms in the molecule but it mainly depends on the total number of atom-types.

In summary, we can conclude that the lack of overall correlation observed in ($\Delta E_{\text{HL}}, \alpha$)-space is a consequence of two main facts: (i) the HOMO-LUMO gap is determined by the nature of the chemical composition (*cf.* Fig. 4c vs. 4e) and (ii) the polarizability is largely determined by the atomic composition (*cf.* Fig. 4b vs. 4d).

4 Case Study: Design of Photodetectors

A common challenge in materials science is the effective design of photodetectors. These optoelectronic devices capture light and convert it to electric signal, therefore playing an important role in sensing, monitoring and optical communication. The wide range of physicochemical properties spanned by organic molecules enables various design strategies, which ultimately led to the emerging field of organic photodetectors^{68,69}. HOMO-LUMO gap is one of the key quantities that can be used to approximate the coupling strength of molecules with light, thereby any design strategy motivated by optics will be initially based on this property^{70,71}. Since the fundamental function of photodetectors is to convert light into electrical current, controlling the electrochemical behavior is also crucial. Specifically, the electrochemical work function plays a critical role in the description of organic photodetectors^{72,73}, as opposed to organic semiconductors, where the focus is usually on the charge carrier mobility⁷⁴. The work function ϕ of an electrode is known to change with the polarizability of the absorbed molecules as well as the surface coverage, as described by the Topping equation⁷⁵ (written for a square lattice)

$$e\Delta\phi = \pm \frac{e\mu\theta}{\varepsilon_0 d^2} \left(1 + 9\alpha' \left(\frac{\theta}{d^2} \right)^{3/2} \right)^{-1}. \quad (5)$$

This expression highlights that the work function ϕ also depends on the the dipole moment μ_0 and polarizability α of the molecules, besides the surface coverage θ and

the lattice constant of the absorbate d . Notice that an effective polarizability α' is used to represent the properties of the absorbed molecules in Eq. 5, which is usually an order of magnitude larger than free molecular polarizability^{76,77}. Despite being acknowledged to fluctuate with the coverage rate, this equation can serve as a useful initial reference to screen potential molecules for photosensitizers according to the intended work function^{72,73}. Indeed, this relationship between both properties makes it important to regulate the polarizability of molecules for achieving a desired electrochemical behavior. Through this connection, it can be seen that molecules with higher polarizability tend to facilitate electron injection while those with lower polarizability tend to facilitate hole injection⁷³.

In the preceding sections, we have postulated that polarizability and HOMO–LUMO gap are uncorrelated if a large enough subset of the CCS is considered. This law can now be translated to the domain of organic photodetectors: since HOMO–LUMO gap and polarizability are generally independent, it should be possible to design a photodetector with a given detection peak having an arbitrary work function. Alternatively, if matching of electrochemical properties of different systems is the goal, it should be possible to design organic photodetectors with each having arbitrary optical detection windows, yet having the same effect on the work function of electrodes. To demonstrate this statement, we use a dataset generated by Xu *et al.*⁷¹, who employed a self-improving Bayesian search to predict possible photodetector molecules in a large subset of CCS. The selection criterion for possible photodetectors was based on both HOMO–LUMO gap and singlet-triplet energy gap, which were evaluated from ground state DFT and TD-DFT calculations, respectively. From all predicted molecules having a donor and an acceptor site (DA structures), we have only selected those cases which have the same atom types as QM7–X molecules (see also in Table 1), leading to a total of 5,311 structures. Using the atomic additive model described in Sec. 3.1, we have estimated the polarizabilities of the selected structures; the plot of the polarizability versus HOMO–LUMO gap is shown in the top panel of Fig. 5. Here, one can see that most structures are found in a relatively extended region having ΔE_{HL} between 3 and 5 eV and α between 200 and 400 a.u., with the possibility to find outliers in all directions around this cluster. In particular, if a high shift in the work function is desired, there appear to be several good candidates with varying optical absorption ranges (see bottom of the graph).

Moreover, our calculations show that the polarizability of these structures can vary by a factor of up to six, depending on the specific values of the HOMO–LUMO gap. To demonstrate this flexibility in α , Fig. 5 also shows the four molecules corresponding to the four quartiles having a HOMO–LUMO gap of 4 ± 0.1 eV, selected to correspond to the maximum density of data. For this specific HOMO–LUMO gap, polarizability changes between 149.9 a.u. and 451.1 a.u. Tak-

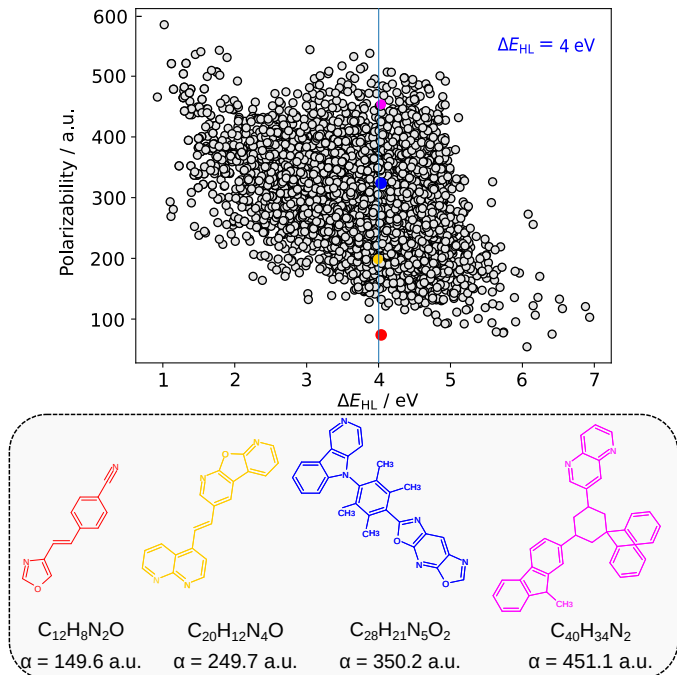


Figure 5: HOMO–LUMO gap and polarizability of the structures in the donor-acceptor (DA) dataset of Xu *et al.*⁷¹. The maximum of the HOMO–LUMO gap density (4 eV) is marked with a blue line. The four structures corresponding to the four quartiles in the predicted polarizability (within ± 0.1 eV) are also shown, together with the predicted values (in a.u.).

ing the tenfold enhancement between the polarizability of the free molecule and the absorbed α' into account and using approximate values of $\mu = 4$ D and $d = 1.5$ nm with a full surface coverage, this would mean that changes in work function could range from 0.9 eV to 1.5 eV. These variations are larger than usually achievable by modifications of a semiconductor structure or controlling the surface coverage^{77,78}. Therefore, our analysis shows that work function can be, for practical purposes, freely tailored, even with a very specific design requirement on HOMO–LUMO gap.

This flexibility is also relevant in the task of designing wavelength-selective detectors, which would imply hard constraint on HOMO–LUMO gap. If ΔE_{HL} and α could not be controlled independently, then optical design restrictions would directly influence the electrochemical behavior. The decoupling of ΔE_{HL} and α means that the wavelength of detection and the work function can be controlled independently. Fine-tuning the work functions to achieve matching on the metal-organic interface at the electrode is crucial for efficiency. Thus, with the existent “freedom of design” in $(\Delta E_{\text{HL}}, \alpha)$ -space, we have demonstrated that an efficient detection can be theoretically achieved for any detection wavelength. Alternatively, since the work function can be tailored to match any detection wavelength, it is also possible to design detectors for different detection ranges having equivalent electrochemical proper-

ties such as sensitivity, dark current, adhesion behavior as well as any other properties determined by the work function.

5 Conclusions

Predictive molecular design is an emerging tool in modern molecular physics and chemistry which heavily relies on the understanding of relationships between key structural and electronic properties. Identifying and explaining correlations between properties necessitates either deep physical understanding or exhaustive data analysis. Herein, we present a comprehensive investigation of the intricate interplay between the HOMO–LUMO gap and dipole polarizability – two central properties in designing molecules with tailored optical properties and intermolecular interactions.

Despite the fact that both quantities have their root in the molecular electronic spectrum, understanding their correlation is quite complex. On one hand, the properties are essentially uncorrelated when accounting for a vast chemical space. On the other hand, when examining a small subset of the chemical compound space with similar functionalities, such as homologous series of molecules like oligomeric hydrocarbons, we show that the properties can be observed as being correlated.

To perform a data-driven analysis, we extended the QM7–X database with functional group labels and accurate polarizabilities to explain the physical cause of this phenomenon. Our results demonstrate that the atomic composition has a major role in determining polarizability, while the arrangement of these atoms into chemical functional groups dictates the HOMO–LUMO gap. The physical origin of molecular polarizability was elaborated by studying conceptual models as well as interpreted with the help of a first order linear atomic additive model. Finally, the “freedom of desing” arising from the interaction of HOMO–LUMO gap and polarizability was used on the example of organic photodetectors, demonstrating that the electrochemical properties of such molecules can be freely tailored even with specific requirement on the optical properties. The theoretical insights gained from this work can give the basis for expanding the understanding of the relationship between HOMO–LUMO gap and polarizability by incorporating additional descriptors such as molecular size and electronic mobility. Additionally, the unraveled “freedom of desing” could be applied to the development of new compounds with tailored optical and electronic properties for use in applications such as organic electronics, sensing or energy harvesting.

6 Computational Methods

6.1 Target Molecular Property Space

Generally, molecular design is a multi-property optimization problem and requires an exhaustive analysis of diverse structure-property and property-property relationships^{1,6}. In this contribution, we have opted to focus on the examination of the two-dimensional property space defined by ΔE_{HL} and α (*i.e.*, $(\Delta E_{\text{HL}}, \alpha)$ -space), as motivated in the introduction (see Sec. 1). To perform a purely data-driven study, we utilize the QM7–X dataset³⁷ containing 42 physicochemical properties of ≈ 4.2 M (equilibrium and non-equilibrium) organic molecules with up to seven heavy (non-hydrogen) atoms (including C, O, N, S and Cl), spanning a practically important subset of CCS. Accordingly, a subset of QM7–X considering only one equilibrium constitutional isomers and stereoisomers per unique molecular graph is selected for further analysis (≈ 13 k molecules). In QM7–X, the molecular structures were optimized using the third-order self-consistent charge density-functional tight binding method (DFTB3)⁷⁹ supplemented with a treatment of many-body dispersion/van der Waals interactions *via* the MBD approach^{44,80}. However, for our studies concerning polarizability, α was computed directly, employing density functional perturbation theory (DFPT)²⁹ by means of the PBE0⁸¹ functional as implemented in the FHI-aims code⁸² (version 190205). The *tight* basis set settings were also applied for all atoms (C, H, N, O, S and Cl). To ensure the transferability of the values, we store the molecular (mean) polarizability (denoted as α and simply referred to as *polarizability* in other parts of the manuscript)

$$\alpha = \frac{1}{3} (\langle \vec{\alpha} \rangle_{xx} + \langle \vec{\alpha} \rangle_{yy} + \langle \vec{\alpha} \rangle_{zz}), \quad (6)$$

which is independent of the molecular orientation. A second orientation-independent observable, the polarizability anisotropy ($\Delta\alpha$) is also often reported, defined as

$$(\Delta\alpha)^2 = 3(\langle \vec{\alpha} \rangle_{xy}^2 + \langle \vec{\alpha} \rangle_{xz}^2 + \langle \vec{\alpha} \rangle_{yz}^2) + \frac{1}{2}((\langle \vec{\alpha} \rangle_{xx} - \langle \vec{\alpha} \rangle_{yy})^2 + (\langle \vec{\alpha} \rangle_{xx} - \langle \vec{\alpha} \rangle_{zz})^2 + (\langle \vec{\alpha} \rangle_{yy} - \langle \vec{\alpha} \rangle_{zz})^2). \quad (7)$$

This quantity is mainly used in the description of macromolecules and supramolecular systems, and since our focus is small organic molecules, we don’t analyze the anisotropy in this manuscript.

To estimate the accuracy of the DFPT calculations, the respective mean polarizabilities were compared with the values documented in the QM7b database^{83,84}. This comparison ensures an accurate assessment of the prediction error due to the following two reasons: *i*) there is a large overlap between the structures in the QM7b and QM7–X databases, and *ii*) QM7b provides

highly accurate α values obtained at the linear-response coupled cluster singles and doubles (LR-CCSD) level of theory^{31,54}. Then, we computed the polarizabilities of 300 randomly selected structures of QM7b employing the same DFPT computational setup described above. The selected level predicts α with a mean average error of 1.9% with a standard deviation of 1.1% (cf. Fig. S 1 of the ESI[†]). This accuracy is higher with respect to average DFT methods which produce mean average errors between 2.5% and 3.8%⁸⁵, underlining the benefits of calculating polarizabilities using the PBE0 functional. Polarizability anisotropy is predicted with a mean average error of 10.2% with a standard deviation of 5.1%, which is in line with previously reported values^{86,87}. In general, the mean polarizability is slightly underestimated, whereas the anisotropy is almost always overestimated by PBE0.

6.2 Molecular Classification: Functional Groups

A workflow has been implemented to identify chemical functional groups from the molecular structure in two steps: firstly, we save the Cartesian coordinates of molecules in a MDL Molfiles format using the standard implementation in `Open Babel`⁸⁸. Secondly, `Checkmol`⁸⁹ is employed to detect the functional groups (204 tags) based on the connectivity tree. In total, 61 unique functional groups were detected for the subset of the ≈ 13 k QM7-X-molecules³⁷, demonstrating that the dataset covers a considerable sector of CCS (cf. Fig. S 2 of the ESI[†]). Since `Open Babel` predicts valencies only based on the distance between pairs of atoms, the functional group detection scheme is prone to errors for molecules with rare functional groups. Moreover, the functional group definitions of `Checkmol` have significant overlaps, e.g., the molecules detected as alkylamines are also detected as primary amines. To ensure that these shortcomings do not influence our conclusions, we base our analyses only on the subset of the 14 k molecules that have certain functional groups. These groups are chosen to be chemically important, non-overlapping, and each of these categories contain at least 500 entries. The number of structures containing one of these functional groups is 9604. For the analysis in section 3.3, only molecules containing a single functional group are considered, i.e., 1626 entries of our dataset (see Fig. S 3). Based on these constraints, the following eleven classes of molecules are identified: aldehydes, carbonitriles, dialkyl ether, enol ether, hydrazones, ketones, oximes, primary alcohols and amines, as well as secondary alcohols and amines (see labels in Fig. 4(a)).

Acknowledgement The authors acknowledge financial support from the Luxembourg National Research Fund: “DRIVEN”(PRIDE17/12252781) under the PRIDE program. Furthermore, the financial support from European Research Council, ERC Consol-

idator Grant “BeStMo (GA n725291)” and INTER-FWO project “MONODISP”, is also gratefully acknowledged. C.M. gratefully acknowledges support by the Alexander von Humboldt Foundation through a Feodor Lynen Research Fellowship.

Notes and references

- (1) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **2018**, *361*, 360–365.
- (2) von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A. Exploring chemical compound space with quantum-based machine learning. *Nat. Rev. Chem.* **2020**, *4*, 347–358.
- (3) Bilodeau, C.; Jin, W.; Jaakkola, T.; Barzilay, R.; Jensen, K. F. Generative models for molecular discovery: Recent advances and challenges. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2022**, *12*, e1608.
- (4) Müller, C.; Bold, S.; Chavarot-Kerlidou, M.; Dietzek-Ivanšić, B. Photoinduced electron transfer in triazole-bridged donor-acceptor dyads – A critical perspective. *Coord. Chem. Rev.* **2022**, *472*, 214764.
- (5) Kranz, C.; Wächtler, M. Characterizing photocatalysts for water splitting: from atoms to bulk and from slow to ultrafast processes. *Chem. Soc. Rev.* **2021**, *50*, 1407–1437, DOI: 10.1039/D0CS00526F.
- (6) Zhao, Z.-J.; Liu, S.; Zha, S.; Cheng, D.; Studt, F.; Henkelman, G.; Gong, J. Theory-guided design of catalytic materials using scaling relationships and reactivity descriptors. *Nat. Rev. Mater.* **2019**, *4*, 792–804.
- (7) Wang, X.; Zhang, G.; Yang, L.; Sharman, E.; Jiang, J. Material descriptors for photocatalyst/catalyst design. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2018**, *8*, e1369.
- (8) Sandonas, L. M.; Hoja, J.; Ernst, B. G.; Vazquez-Mayagoitia, A.; DiStasio Jr, R. A.; Tkatchenko, A. Quantum Mechanics Enables "Freedom of Design" in Molecular Property Space. *ChemRxiv* **2021**, DOI: 10.26434/chemrxiv-2021-q9rc2.
- (9) Katritzky, A. R.; Sild, S.; Karelson, M. Correlation and Prediction of the Refractive Indices of Polymers by QSPR. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1171–1176.
- (10) Takimiya, K.; Yamamoto, T.; Ebata, H.; Izawa, T. Design strategy for air-stable organic semiconductors applicable to high-performance field-effect

- transistors. *Sci. Technol. Adv. Mater.* **2007**, *8*, 273–276.
- (11) Mazinani, S. K. S.; Meidanshahi, R. V.; Palma, J. L.; Tarakeshwar, P.; Hansen, T.; Ratner, M. A.; Mujica, V. Polarizability as a Molecular Descriptor for Conductance in Organic Molecular Circuits. *J. Phys. Chem. C* **2016**, *120*, 26054–26060.
- (12) Gryn'ova, G.; Lin, K.-H.; Corminboeuf, C. Read between the Molecules: Computational Insights into Organic Semiconductors. *J. Am. Chem. Soc.* **2018**, *140*, 16370–16386.
- (13) Bredas, J.-L. Mind the gap! *Mater. Horiz.* **2014**, *1*, 17–19.
- (14) Stein, T.; Eisenberg, H.; Kronik, L.; Baer, R. Fundamental Gaps in Finite Systems from Eigenvalues of a Generalized Kohn-Sham Method. *Phys. Rev. Lett.* **2010**, *105*, 266802.
- (15) Karplus, M.; Kolker, H. J. Van der Waals Forces in Atoms and Molecules. *J. Chem. Phys.* **1964**, *41*, 3955–3961.
- (16) Hohm, U. Is There a Minimum Polarizability Principle in Chemical Reactions? *J. Phys. Chem. A* **2000**, *104*, 8418–8423.
- (17) Stone, A. J. *The Theory of Intermolecular Forces*; Oxford University Press: New York, 2016.
- (18) Xie, C.; Oganov, A. R.; Dong, D.; Liu, N.; Li, D.; Debela, T. T. Rational design of inorganic dielectric materials with expected permittivity. *Sci. Rep.* **2015**, *5*, 16769.
- (19) Sabirov, D. S. Polarizability as a landmark property for fullerene chemistry and materials science. *RSC Adv.* **2014**, *4*, 44996–45028.
- (20) Kamada, K.; Ueda, M.; Nagao, H.; Tawa, K.; Sugino, T.; Shmizu, Y.; Ohta, K. Molecular Design for Organic Nonlinear Optics: Polarizability and Hyperpolarizabilities of Furan Homologues Investigated by Ab Initio Molecular Orbital Method. *J. Phys. Chem. A* **2000**, *104*, 4723–4734.
- (21) Pouchan, C.; Bégué, D.; Zhang, D. Y. Between geometry, stability, and polarizability: Density functional theory studies of silicon clusters Si_n (n=3–10). *J. Chem. Phys.* **2004**, *121*, 4628–4634.
- (22) Targema, M.; Obi-Egbedi, N. O.; Adeoye, M. D. Molecular structure and solvent effects on the dipole moments and polarizabilities of some aniline derivatives. *Comput. Theor. Chem.* **2013**, *1012*, 47–53.
- (23) Kim, T.-D.; Lee, K.-S. D- π -A Conjugated Molecules for Optoelectronic Applications. *Macromol. Rapid Commun.* **2015**, *36*, 943–958.
- (24) Wang, D.; Fracasso, D.; Nurbawono, A.; Annadata, H. V.; Sangeeth, C. S. S.; Yuan, L.; Nijhuis, C. A. Tuning the Tunneling Rate and Dielectric Response of SAM-Based Junctions via a Single Polarizable Atom. *Adv. Mater.* **2015**, *27*, 6689–6695.
- (25) Kaasbjerg, K.; Flensberg, K. Strong Polarization-Induced Reduction of Addition Energies in Single-Molecule Nanojunctions. *Nano Lett.* **2008**, *8*, 3809–3814.
- (26) Jaque, P.; Toro-Labbé, A. Characterization of copper clusters through the use of density functional theory reactivity descriptors. *J. Chem. Phys.* **2002**, *117*, 3208–3218.
- (27) Meyers, F.; Marder, S. R.; Pierce, B. M.; Bredas, J. L. Electric Field Modulated Nonlinear Optical Properties of Donor-Acceptor Polyenes: Sum-Over-States Investigation of the Relationship between Molecular Polarizabilities (.alpha., .beta., and .gamma.) and Bond Length Alternation. *J. Am. Chem. Soc.* **1994**, *116*, 10703–10714.
- (28) Otto, P. Recent developments in the theoretical design of low-gap polymers and their nonlinear optical properties. *Int. J. Quantum Chem.* **1994**, *52*, 353–364.
- (29) Shang, H.; Raimbault, N.; Rinke, P.; Scheffler, M.; Rossi, M.; Carbogno, C. All-electron, real-space perturbation theory for homogeneous electric fields: theory, implementation, and application within DFT. *New J. Phys.* **2018**, *20*, 073040.
- (30) Karna, S. P. A “direct” time-dependent coupled perturbed Hartree–Fock–Roothaan approach to calculate molecular (hyper) polarizabilities. *Chem. Phys. Lett.* **1993**, *214*, 186–192.
- (31) Wilkins, D. M.; Grisafi, A.; Yang, Y.; Lao, K. U.; DiStasio, R. A.; Ceriotti, M. Accurate molecular polarizabilities with coupled cluster theory and machine learning. *Proc. Natl. Acad. Sci.* **2019**, *116*, 3401–3406.
- (32) Bosque, R.; Sales, J. Polarizabilities of Solvents from the Chemical Composition. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1154–1163.
- (33) Curtarolo, S.; Hart, G. L. W.; Nardelli, M. B.; Mingo, N.; Sanvito, S.; Levy, O. The high-throughput highway to computational materials design. *Nat. Mater.* **2013**, *12*, 191–201.

- (34) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **2014**, *1*, 140022.
- (35) Gromski, P. S.; Henson, A. B.; Granda, J. M.; Cronin, L. How to explore chemical space using algorithms and automation. *Nat. Rev. Chem.* **2019**, *3*, 119–128.
- (36) Gadaleta, D.; Benfenati, E. A descriptor-based analysis to highlight the mechanistic rationale of mutagenicity. *J. Environ. Sci. Health, Part C* **2021**, *39*, 269–292.
- (37) Hoja, J.; Sandonas, L. M.; Ernst, B. G.; Vazquez-Mayagoitia, A.; DiStasio Jr, R. A.; Tkatchenko, A. QM7-X, a comprehensive dataset of quantum-mechanical properties spanning the chemical space of small organic molecules. *Sci. Data* **2021**, *8*, 1–11.
- (38) Cramer, C. J. *Essentials of computational chemistry: theories and models*; John Wiley & Sons: West Sussex, 2013.
- (39) Ramakrishnan, R.; Hartmann, M.; Tapavicza, E.; von Lilienfeld, O. A. Electronic spectra from TDDFT and machine learning in chemical space. *J. Chem. Phys.* **2015**, *143*, 084111.
- (40) Medved', M.; Stachová, M.; Jacquemin, D.; André, J.-M.; Perpète, E. A. A generalized Romberg differentiation procedure for calculation of hyperpolarizabilities. *J. Mol. Struct.: THEOCHEM* **2007**, *847*, 39–46.
- (41) Szabó, P.; Góger, S.; Charry, J.; Karimpour, M. R.; Fedorov, D. V.; Tkatchenko, A. Four-Dimensional Scaling of Dipole Polarizability in Quantum Systems. *Phys. Rev. Lett.* **2022**, *128*, 070602.
- (42) Buckingham, R. The quantum theory of atomic polarization I—Polarization by a uniform field. *Proc. R. Soc. London, Ser. A* **1937**, *160*, 94–113.
- (43) Mayer, A.; Åstrand, P.-O. A Charge-Dipole Model for the Static Polarizability of Nanostructures Including Aliphatic, Olefinic, and Aromatic Systems. *J. Phys. Chem. A* **2008**, *112*, 1277–1285.
- (44) Tkatchenko, A.; DiStasio, R. A.; Car, R.; Scheffler, M. Accurate and Efficient Method for Many-Body van der Waals Interactions. *Phys. Rev. Lett.* **2012**, *108*, 236402.
- (45) Jones, A. P.; Crain, J.; Sokhan, V. P.; Whitfield, T. W.; Martyna, G. J. Quantum Drude oscillator model of atoms and molecules: Many-body polarization and dispersion interactions for atomistic simulation. *Phys. Rev. B* **2013**, *87*, 144103.
- (46) Sylvain, M. G.; Csizmadia, I. Average dipole polarizabilities from the unsold approximation and ab initio data. *Chem. Phys. Lett.* **1987**, *136*, 575–582.
- (47) Roberts, R. E. Scaled Unsöld Approximation for Atoms and Molecules. *J. Chem. Phys.* **1967**, *47*, 1873–1873.
- (48) Pearson, R. G. Hard and soft acids and bases, HSAB, part 1: Fundamental principles. *J. Chem. Educ.* **1968**, *45*, 581.
- (49) Pearson, R. G. Hard and soft acids and bases, HSAB, part II: Underlying theories. *J. Chem. Educ.* **1968**, *45*, 643.
- (50) Brinck, T.; Murray, J. S.; Politzer, P. Polarizability and volume. *J. Chem. Phys.* **1993**, *98*, 4305–4306.
- (51) Politzer, P.; Jin, P.; Murray, J. S. Atomic polarizability, volume and ionization energy. *J. Chem. Phys.* **2002**, *117*, 8197–8202.
- (52) Blair, S. A.; Thakkar, A. J. Relating polarizability to volume, ionization energy, electronegativity, hardness, moments of momentum, and other molecular properties. *J. Chem. Phys.* **2014**, *141*, 074306.
- (53) Choudhary, S.; Ranjan, P.; Chakraborty, T. Atomic polarizability: A periodic descriptor. *J. Chem. Res.* **2020**, *44*, 227–234.
- (54) Yang, Y.; Lao, K. U.; DiStasio, R. A. Influence of Pore Size on the van der Waals Interaction in Two-Dimensional Molecules and Materials. *Phys. Rev. Lett.* **2019**, *122*, 026001.
- (55) Fedorov, D. V.; Sadhukhan, M.; Stöhr, M.; Tkatchenko, A. Quantum-mechanical relation between atomic dipole polarizability and the van der Waals radius. *Phys. Rev. Lett.* **2018**, *121*, 183401.
- (56) Thakkar, A. J. A hierarchy for additive models of polarizability. *AIP Conf. Proc.* **2012**, *1504*, 586–589.
- (57) Gilbert, A.; Baggott, J. E. *Essentials of molecular photochemistry*; Wiley-Blackwell: Oxford, 1991.
- (58) Joung, J. F.; Han, M.; Hwang, J.; Jeong, M.; Choi, D. H.; Park, S. Deep Learning Optical Spectroscopy Based on Experimental Database: Potential Applications to Molecular Design. *JACS Au* **2021**, *1*, 427–438.
- (59) Woodward, R. B. Structure and the Absorption Spectra of α,β -Unsaturated Ketones. *J. Am. Chem. Soc.* **1941**, *63*, 1123–1126.

- (60) Bahnick, D. A. Use of Huckel Molecular Orbital Theory in Interpreting the Visible Spectra of Polymethine Dyes: An Undergraduate Physical Chemistry Experiment. *Journal of Chemical Education* **1994**, *71*, 171.
- (61) Balcells, D.; Skjelstad, B. B. tmQM Dataset-Quantum Geometries and Properties of 86k Transition Metal Complexes. *J. Chem. Inf. Model.* **2020**, *60*, 6135–6146.
- (62) Montavon, G.; Rupp, M.; Gobre, V.; Vazquez-Mayagoitia, A.; Hansen, K.; Tkatchenko, A.; Müller, K.-R.; Von Lilienfeld, O. A. Machine learning of molecular electronic properties in chemical compound space. *New J. Phys.* **2013**, *15*, 095003.
- (63) Wang, J.; Yang, M.; Wang, G.; Zhao, J. Dipole polarizabilities of germanium clusters. *Chem. Phys. Lett.* **2003**, *367*, 448–454.
- (64) Alyar, H.; Kantarci, Z.; Bahat, M.; Kasap, E. Investigation of torsional barriers and nonlinear optical (NLO) properties of phenyltriazines. *J. Mol. Struct.* **2007**, *834-836*, 516–520.
- (65) De Proft, F.; Ayers, P. W.; Fias, S.; Geerlings, P. Woodward-Hoffmann rules in density functional theory: Initial hardness response. *J. Chem. Phys.* **2006**, *125*, 214101.
- (66) Afzal, M. A. F.; Cheng, C.; Hachmann, J. Combining first-principles and data modeling for the accurate prediction of the refractive index of organic polymers. *J. Chem. Phys.* **2018**, *148*, 241712.
- (67) Zade, S. S.; Bendikov, M. From Oligomers to Polymer: Convergence in the HOMO-LUMO Gaps of Conjugated Oligomers. *Org. Lett.* **2006**, *8*, 5243–5246.
- (68) Brédas, J. L.; Calbert, J. P.; da Silva Filho, D. A.; Cornil, J. Organic semiconductors: A theoretical characterization of the basic parameters governing charge transport. *Proc. Natl. Acad. Sci.* **2002**, *99*, 5804–5809.
- (69) Yang, D.; Ma, D. Development of Organic Semiconductor Photodetectors: From Mechanism to Applications. *Adv. Opt. Mater.* **2019**, *7*, 1800522.
- (70) Narsaria, A. K.; Poater, J.; Fonseca Guerra, C.; Ehlers, A. W.; Lammertsma, K.; Bickelhaupt, F. M. Rational design of near-infrared absorbing organic dyes: Controlling the HOMO–LUMO gap using quantitative molecular orbital theory. *J. Comput. Chem.* **2018**, *39*, 2690–2696.
- (71) Xu, S.; Li, J.; Cai, P.; Liu, X.; Liu, B.; Wang, X. Self-Improving Photosensitizer Discovery System via Bayesian Search with First-Principle Simulations. *J. Am. Chem. Soc.* **2021**, *143*, 19769–19777.
- (72) Castellani, M.; Winkler, S.; Bröker, B.; Baumgarten, M.; Müllen, K.; Koch, N. Work function increase of transparent conductive electrodes by solution processed electron acceptor molecular monolayers. *Appl. Phys. A* **2014**, *114*, 291–295.
- (73) Peng, X.; Hu, L.; Qin, F.; Zhou, Y.; Chu, P. K. Low Work Function Surface Modifiers for Solution-Processed Electronics: A Review. *Adv. Mater. Interfaces* **2018**, *5*, 1701404.
- (74) Kunkel, C.; Schober, C.; Margraf, J. T.; Reuter, K.; Oberhofer, H. Finding the Right Bricks for Molecular Legos: A Data Mining Approach to Organic Semiconductor Design. *Chem. Mater.* **2019**, *31*, 969–978.
- (75) Topping, J. On the mutual potential energy of a plane network of doublets. *Proc. R. Soc. London, Ser. A* **1927**, *114*, 67–72.
- (76) Maurer, R. J.; Ruiz, V. G.; Tkatchenko, A. Many-body dispersion effects in the binding of adsorbates on metal surfaces. *J. Chem. Phys.* **2015**, *143*, 102808.
- (77) Widdascheck, F.; Hauke, A. A.; Witte, G. A Solvent-Free Solution: Vacuum-Deposited Organic Monolayers Modify Work Functions of Noble Metal Electrodes. *Adv. Funct. Mater.* **2019**, *29*, 1808385.
- (78) Naghdi, S.; Sanchez-Arriaga, G.; Rhee, K. Y. Tuning the work function of graphene toward application as anode and cathode. *J. Alloys Compd.* **2019**, *805*, 1117–1134.
- (79) 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.
- (80) Ambrosetti, A.; Reilly, A. M.; DiStasio, R. A.; Tkatchenko, A. Long-range correlation energy calculated from coupled atomic response functions. *J. Chem. Phys.* **2014**, *140*, 18A508.
- (81) Adamo, C.; Barone, V. Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.* **1999**, *110*, 6158–6170.
- (82) Blum, V.; Gehrke, R.; Hanke, F.; Havu, P.; Havu, V.; Ren, X.; Reuter, K.; Scheffler, M. Ab initio molecular simulations with numeric atom-centered orbitals. *Comput. Phys. Commun.* **2009**, *180*, 2175–2196.
- (83) Blum, L. C.; Raymond, J.-L. 970 Million Drug-like Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. *J. Am. Chem. Soc.* **2009**, *131*, 8732.

- (84) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.
- (85) Hait, D.; Head-Gordon, M. How accurate are static polarizability predictions from density functional theory? An assessment over 132 species at equilibrium geometry. *Phys. Chem. Chem. Phys.* **2018**, *20*, 19800–19810.
- (86) Adamo, C.; Cossi, M.; Scalmani, G.; Barone, V. Accurate static polarizabilities by density functional theory: assessment of the PBE0 model. *Chem. Phys. Lett.* **1999**, *307*, 265–271.
- (87) Hickey, A. L.; Rowley, C. N. Benchmarking quantum chemical methods for the calculation of molecular dipole moments and polarizabilities. *J. Phys. Chem. A* **2014**, *118*, 3678–3687.
- (88) O’Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminf.* **2011**, *3*, 1–14.
- (89) Haider, N. Functionality Pattern Matching as an Efficient Complementary Structure/Reaction Search Tool: an Open-Source Approach. *Molecules* **2010**, *15*, 5079–5092.

TOC Graphic

