Data-Driven Discovery of Synthetic-Compatible Organic Semiconductors for Multifunctional Applications

Your Name

May 18, 2025

Abstract

Organic semiconductors are emerging as multifunctional materials suitable for applications in photovoltaics (OPVs), optoelectronics, and bioimaging. This work proposes a data-driven pipeline to identify synthetically accessible and versatile molecules from the GDB-9 database. Rather than relying on ab initio calculations, we leverage precomputed quantum descriptors from GDB-9 and photophysical properties from CEPDB.

Candidate molecules are filtered as potential semiconductors, and their power conversion efficiency (PCE) is estimated using the Scharber model. Compatibility with two benchmark OPV materials—PCBM (acceptor) and PCDTBT (donor)—is evaluated via frontier orbital alignment and synthetic accessibility scoring. RDKit and Open Babel are used to compute additional physicochemical descriptors to infer multifunctional potential. TDDFT-based values from CEPDB further guide OLED and biomedical relevance assessment.

This study delivers a robust cheminformatics workflow for discovering organic molecules with cross-domain applications, contributing to the advancement of sustainable materials and clean energy aligned with SDGs 7 and 13.

1 Introduction

In the face of the global climate crisis and increasing energy demand, the pursuit of sustainable and scalable energy technologies is more pressing than ever. Organic electronics have emerged as a pivotal class of materials with the potential to transform energy harvesting, lighting, and biomedical applications. Unlike their inorganic counterparts, organic semiconductors offer unparalleled tunability, low-temperature processability, and compatibility with flexible substrates, enabling the fabrication of lightweight, wearable, and even biodegradable electronic devices [1, 2].

Among these technologies, organic photovoltaics (OPVs) have garnered substantial interest due to their ability to convert solar energy into electricity using carbon-based molecules as active components. Since the inception of bulk heterojunction architectures, which blend donor and acceptor materials to enhance exciton dissociation and charge collection, OPVs have achieved considerable performance improvements. State-of-the-art devices have reached power conversion efficiencies (PCEs) exceeding 18% in laboratory conditions [41], positioning OPVs as credible alternatives to traditional silicon-based solar cells for specific use cases such as building-integrated photovoltaics (BIPV), roll-to-roll printed modules, or off-grid portable power systems.

However, the field still faces significant bottlenecks. The design of new materials with optimal HOMO-LUMO alignment, broad absorption spectra, and balanced charge transport remains a trial-and-error process. Moreover, the synthetic accessibility of high-performing molecules is rarely considered during initial screenings, leading to impractical candidates for experimental validation. While density functional theory (DFT) and time-dependent DFT (TDDFT) have long been used for predicting the electronic and optical properties of conjugated systems, their computational cost makes them unsuitable for screening the millions of possibilities within chemical space.

In response to this challenge, cheminformatics and database-driven strategies have gained prominence. The QM9/GDB-9 dataset [16] provides DFT-calculated properties for over 130,000 stable organic molecules

composed of light elements (C, H, O, N, and F), offering a rich resource for virtual screening. Similarly, the Clean Energy Project Database (CEPDB) [48], derived from high-throughput calculations, includes TDDFT-level predictions for thousands of molecules specifically designed for photovoltaic applications. Together, these databases enable the identification of promising candidates without the need for fresh quantum mechanical simulations.

This study proposes a data-driven pipeline for identifying and evaluating small organic molecules with multifunctional potential. The approach begins by filtering GDB-9 molecules likely to exhibit semiconducting behavior based on their HOMO-LUMO gaps and electronic structure alignment with reference materials such as [6,6]-phenyl-C_{61}-butyric acid methyl ester (PCBM) and poly[N-9'-heptadecanyl-2,7-carbazole-alt-5,5-(4',7'-di-2-thienyl-2',1',3'-benzothiadiazole)] (PCDTBT). These materials are widely used as benchmark acceptor and donor species in high-efficiency OPV systems due to their favorable energetics and stability [3].

The power conversion efficiency of candidate molecules is then estimated using the Scharber model, a semi-empirical framework that relates orbital energies and optical bandgap to device-level performance. In parallel, the synthetic accessibility of each molecule is evaluated through the Synthetic Accessibility Score (SAScore), ensuring that selected compounds are not only performant but also feasible to produce. To explore broader applications beyond photovoltaics, a series of molecular descriptors—including molecular weight, polar surface area, and logP—are computed using cheminformatics libraries such as RDKit and Open Babel. These descriptors provide insight into solubility, drug-likeness, and interaction potential, which are critical for OLEDs, chemical sensors, and bio-imaging agents.

Photophysical properties obtained from CEPDB, such as fluorescence energy, oscillator strength, and singlet-triplet gap, are also incorporated to assess compatibility with OLED and biosensor platforms. Molecules exhibiting short singlet-triplet gaps and high oscillator strength are flagged as candidates for thermally activated delayed fluorescence (TADF) or bio-labeling applications. By integrating these orthogonal property domains, the methodology enables the discovery of cross-functional molecules suitable for energy, optoelectronic, and health-related technologies.

Ultimately, this work contributes to the vision of a sustainable materials genome by demonstrating how publicly available datasets and cheminformatics tools can be harnessed to identify versatile organic semiconductors. It aligns with global initiatives such as the United Nations' Sustainable Development Goals (SDGs) 7 (Affordable and Clean Energy) and 13 (Climate Action), and lays the groundwork for future experimental validation and machine-learning-driven exploration of functional materials.

2 Methodology

This section outlines the cheminformatics-based methodology used to identify and evaluate multifunctional small organic molecules from existing molecular databases. The workflow is divided into four main components: (i) dataset selection and filtering, (ii) PCE estimation using the Scharber model, (iii) physicochemical profiling using cheminformatics tools, and (iv) optoelectronic screening using TDDFT-derived descriptors.

2.1 Dataset and Molecular Properties

This study leverages the *GDB-9* dataset, a subset of the larger *GDB-17* database, renowned for its comprehensive collection of small organic molecules [16, 19]. The *GDB-9* dataset encompasses approximately 134,000 stable organic molecules, each composed of carbon, hydrogen, oxygen, nitrogen, and fluorine atoms. These molecules represent a valuable resource for virtual screening and property prediction in organic electronics. As highlighted in recent reviews of chemical datasets for machine learning, such curated collections are essential for training and validating models that can accelerate the discovery of novel materials [9, 10].

The *GDB-9* dataset provides a wealth of information, including geometric, energetic, and electronic prop-

erties crucial for understanding the behavior of organic photovoltaic (OPV) materials. Key electronic properties include the energies of the highest occupied molecular orbital (HOMO) and lowest unoccupied molecular orbital (LUMO), as well as the corresponding HOMO-LUMO energy gap. These properties, calculated using Density Functional Theory (DFT) with the B3LYP functional and the 6-31G(2df,p) basis set, are essential for predicting charge transfer characteristics and overall device performance in OPVs [6]. Furthermore, the dataset also reports atomization energies, enthalpies, and Gibbs free energies calculated at the higher-level G4MP2 theoretical level, providing a more precise description of the thermodynamic stability of the molecules. The availability of both DFT and G4MP2 calculated properties allows for a multi-faceted analysis of structure-property relationships. As emphasized in the literature, QM datasets like GDB-9 are essential for developing and benchmarking machine learning models for predicting quantum chemical properties, ultimately reducing the computational cost associated with DFT calculations [6, 11].

2.2 Molecular Selection and Property Considerations

To develop and validate the model, a subset of molecules from the *Harvard Clean Energy Project Database (CEPDB)* (available at CEPDB Molecular Space) was selected, focusing on optoelectronic characteristics crucial for organic photovoltaic (OPV) applications. The initial CEPDB dataset comprised 133,885 molecules, characterized by their HOMO, LUMO, and energy gap values. A multi-step filtering process was then employed to identify molecules meeting specific criteria relevant to OPV performance, guided by the Scharber model for predicting photovoltaic efficiency.

The filtering process involved a series of sequential steps, each targeting a specific energy level range to ensure compatibility with efficient charge transfer and exciton dissociation in OPV devices.

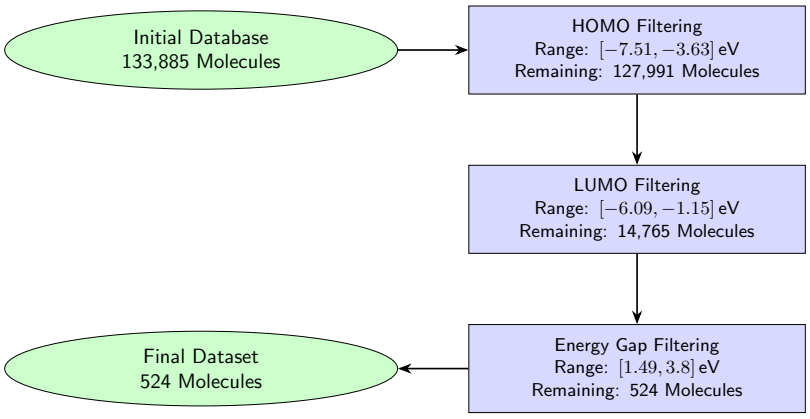


Figure 1: Stepwise filtering of GDB-9 dataset based on HOMO, LUMO, and energy gap thresholds..

2.3 Workflow Architecture

The overall strategy involves the integration of two complementary datasets: GDB-9 [16], which contains DFT-derived quantum properties for 134,000 molecules made up of C, H, O, N, and F, and CEPDB [48], which provides TDDFT photophysical descriptors for conjugated organic molecules. The workflow aims to reduce the computational burden by reusing validated descriptors and streamlining candidate selection based on both performance and feasibility.

The screening begins with GDB-9. Molecules are filtered based on their HOMO-LUMO gap (1.0–3.5 eV), frontier orbital alignment with PCBM (acceptor) or PCDTBT (donor), and synthetic accessibility score (SAScore). Compatibility is evaluated by energetic alignment and performance is estimated using the Scharber model. CEPDB provides secondary optoelectronic properties to support OLED and biosensing potential.

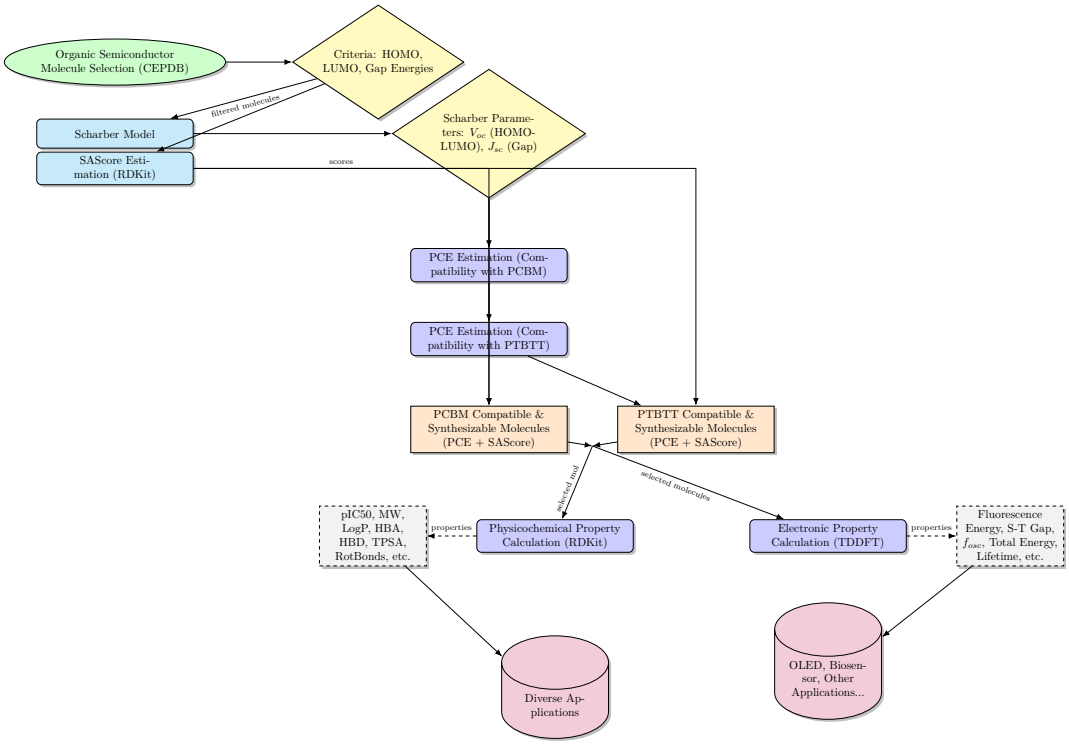


Figure 2: Data-centric screening workflow combining GDB-9 and CEPDB for multi-domain molecular discovery.

2.3.1 Power Conversion Efficiency Estimation: The Scharber Model

The Scharber model, developed by Markus C. Scharber et al. in 2006 [3], is a widely used semi-empirical model for estimating the conversion efficiency of organic photovoltaic (OPV) devices. By combining experimental and theoretical parameters, it predicts photovoltaic device performance. This model has significantly advanced the development of high-performance OPV materials by evaluating the *PCE*, a critical indicator of solar cell viability. The *PCE* depends on several electronic parameters detailed below.

Short-Circuit Current Density (J_{sc}): The short-circuit current density (J_{sc}) is the maximum current generated by the cell when short-circuited. It is influenced by the thickness of the active layer and the number of photons absorbed. J_{sc} is directly related to the material's ability to generate electron-hole pairs, collect charge carriers, and minimize resistive losses.

$$J_{sc} = Ae^{-E_{GAP}^2/B} \tag{1}$$

- $A = 433.12$ and $B = 2.34$ are fitting parameters derived from the Tartarus model results [27].
- E_{GAP} is the donor's bandgap energy, defined as the difference between molecular $HOMO$ and $LUMO$ levels, with maximum values up to 3.8 eV.

Open-Circuit Voltage (V_{oc}): The open-circuit voltage (V_{oc}) is the maximum voltage produced by the cell when no current flows. It depends on the energy difference between the donor's $HOMO$ and the acceptor's $LUMO$ [28].

$$V_{oc} = \frac{1}{e}\left(|E^{Do}HOMO| - |E^{AC}LUMO|\right) - 0.3 \tag{2}$$

- e: Electron charge.
- $E^{Do}HOMO$: Donor $HOMO$ energy, typically between -5.7 and -4.5, eV [18].
- $E^{AC}LUMO$: Acceptor $LUMO$ energy, estimated between -4.0 and -3.0, eV [17].
- -0.3, V : Exciton separation threshold voltage.

Fill Factor (FF): The Fill Factor (FF) measures the efficiency of a solar cell in converting light energy into usable electrical energy. It is calculated as the ratio of the maximum power output to the incident power. Based on the Tartarus model, the FF is estimated at 65% [27].

Power Conversion Efficiency (PCE): The Power Conversion Efficiency (PCE) represents the ratio of the electrical power generated by the cell to the incident light power. Current OPVs achieve a PCE of 12%, with a target efficiency of 20% [23].

$$\text{PCE} = 100 \times \frac{V_{oc} \cdot \text{FF} \cdot J_{sc}}{P_{\text{in}}} \tag{3}$$

Where $P_{\text{in}} = 900.14\,\text{W/m}^2$ is the incident light power.

Only molecules with PCE $> 2\%$ and SAScore < 5 were considered promising. This threshold ensures a balance between efficiency and synthetic tractability.

2.3.2 Synthetic Accessibility Score and Synthesis Compatibility

Definition of the Synthetic Accessibility Score (SA Score): The Synthetic Accessibility (SA) score is a metric designed to evaluate the *ease of synthesis* of an organic molecule on a scale from **1** (easy) to **10** (difficult). Developed to guide material design in various fields, including organic photovoltaics (OPVs), the SA score incorporates the following molecular properties:

- **Molecular size**: Larger molecules are generally more challenging to synthesize.
- **Structural complexity**: Molecules with higher structural complexity require more intricate synthetic pathways.
- **Functional groups**: Specific groups can either facilitate or complicate synthesis depending on their reactivity.

This metric promotes the selection of molecules that achieve an optimal balance between *high electronic performance* and *practical feasibility*, aligning with the guidelines set out by Scharber et al. (2006) and Ruddigkeit et al. (2015).

Definition of Synthesis Compatibility: Synthesis compatibility identifies molecules that are both compatible and synthesizable, aiming to optimize OPV performance. The compatibility is evaluated using the metric PCE_{SAS}, which integrates the photovoltaic efficiency (PCE) with the Synthetic Accessibility score (SAs):

$$PCE_{SAS} = PCE - SAs$$

Where:

- PCE: Photovoltaic conversion efficiency, representing the device's performance.

- SAs: Synthetic Accessibility score, quantifying the difficulty of synthesis.

Molecules are considered compatible if they maximize the PCE_{SAS} value, achieving a balance between a high PCE and a low SA score.

Relevance for OPV Design This dual-metric approach enables:

1. Selection of high-performance molecules optimized for synthesis feasibility.
2. Prioritization of materials that are scalable and cost-effective for production.

2.4 Physicochemical Profiling via RDKit and Open Babel

To assess the broader applicability of GDB-9 derived molecules beyond photovoltaic materials, a comprehensive physicochemical profiling was performed using cheminformatics tools such as `RDKit` and `Open Babel`. These descriptors play a central role in evaluating molecular behavior across diverse domains, including pharmaceutical screening, bioimaging, and electronic materials [42, 43].

The computed properties are grouped into the following functional categories:

- **Lipophilicity:** LogP, AlogP, CX LogP, and CX LogD were calculated to assess hydrophobicity. Molecules with moderate LogP values (1–5) are generally considered drug-like [44], and lipophilicity also influences charge transport in semiconductors and OLED stability [**?**].
- **Structural Properties:** Molecular weight (MW), monoisotopic weight, number of rotatable bonds, ring count, heavy atom count, and fraction of sp^3 carbons (Fsp^3) were computed. High MW can indicate complex synthetic routes, whereas low MW and low rotatable bond count often correlate with better crystallinity and packing in OPV devices [45].
- **Polarity and Solubility:** Topological Polar Surface Area (TPSA), H-bond acceptor (HBA) and donor (HBD) counts were included as predictors of solubility and permeability. TPSA < 140 Å^2 is generally preferred for oral bioavailability, while polar surfaces also affect molecular interactions with biological membranes and electrodes [46].
- **Bioactivity-Related Descriptors:** The Quantitative Estimate of Drug-likeness (QED) score [**?**], predicted pIC_{50} values, and IC_{50} in nM were included to highlight potential bioactivity. While not originally developed for photovoltaics, these metrics support cross-functional screening, especially for diagnostic dyes and therapeutic agents.

By analyzing these descriptors jointly, we identified molecular scaffolds with high structural diversity and multiparameter balance—key features for applications such as biocompatible sensors, fluorescent probes, or organic bioelectronics [47].

2.5 Optoelectronic Screening via TDDFT Descriptors

To explore the optoelectronic potential of candidate molecules, we used descriptors derived from time-dependent density functional theory (TDDFT) calculations obtained from the Clean Energy Project Database (CEPDB) [48]. These properties are critical for evaluating emissive behavior and charge transfer in OLEDs and photonic devices.

The TDDFT-based descriptors used in this study include:

- **Fluorescence Energy (eV):** Defines the emission color; values between 2.0 and 3.5 eV correspond to visible light emission, making the molecule suitable for OLEDs and fluorescence-based diagnostics [49].

- **Oscillator Strength (f):** A measure of transition dipole moment and emission probability. Molecules with $f > 0.5$ exhibit strong radiative transitions, a key criterion for luminescent materials [50].
- **Singlet-Triplet Gap (ΔE_{ST}):** A low ΔE_{ST} (< 0.3 eV) facilitates reverse intersystem crossing, promoting thermally activated delayed fluorescence (TADF), an efficient mechanism in OLEDs [51].
- **Excited-State Lifetime (ns):** Long-lived singlet states can enhance quantum yield in sensing and imaging applications. However, for display technology, shorter lifetimes with high emission rates are often preferred to minimize lag and afterglow [52].

Molecules satisfying multiple criteria (visible-range emission, high oscillator strength, and suitable ΔE_{ST}) were flagged as high-potential candidates for dual application in OPVs and OLEDs. Their excited-state profiles were then cross-referenced with photovoltaic properties such as HOMO-LUMO alignment and Scharber PCE estimations to identify multifunctional compounds.

This multidimensional screening bridges the gap between electronic structure and functional performance, providing a robust framework for the discovery of small molecules with synergistic utility across energy and biomedical technologies.

3 Results

3.1 Database analysis

3.1.1 Estimating the number of clusters

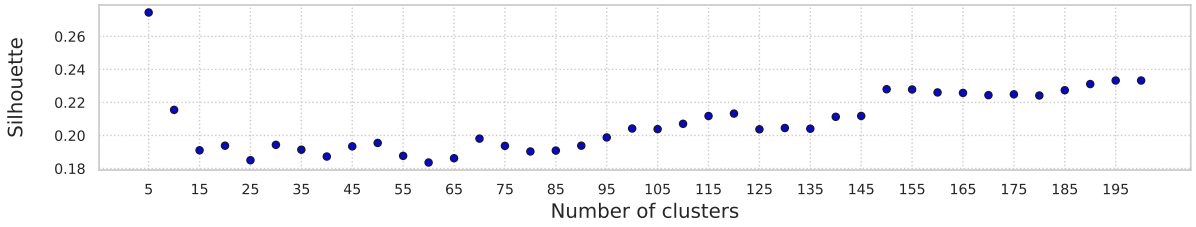


Figure 3: Silhouette scores under different number of cluster sizes using the total feature matrix using the k -means algorithm..

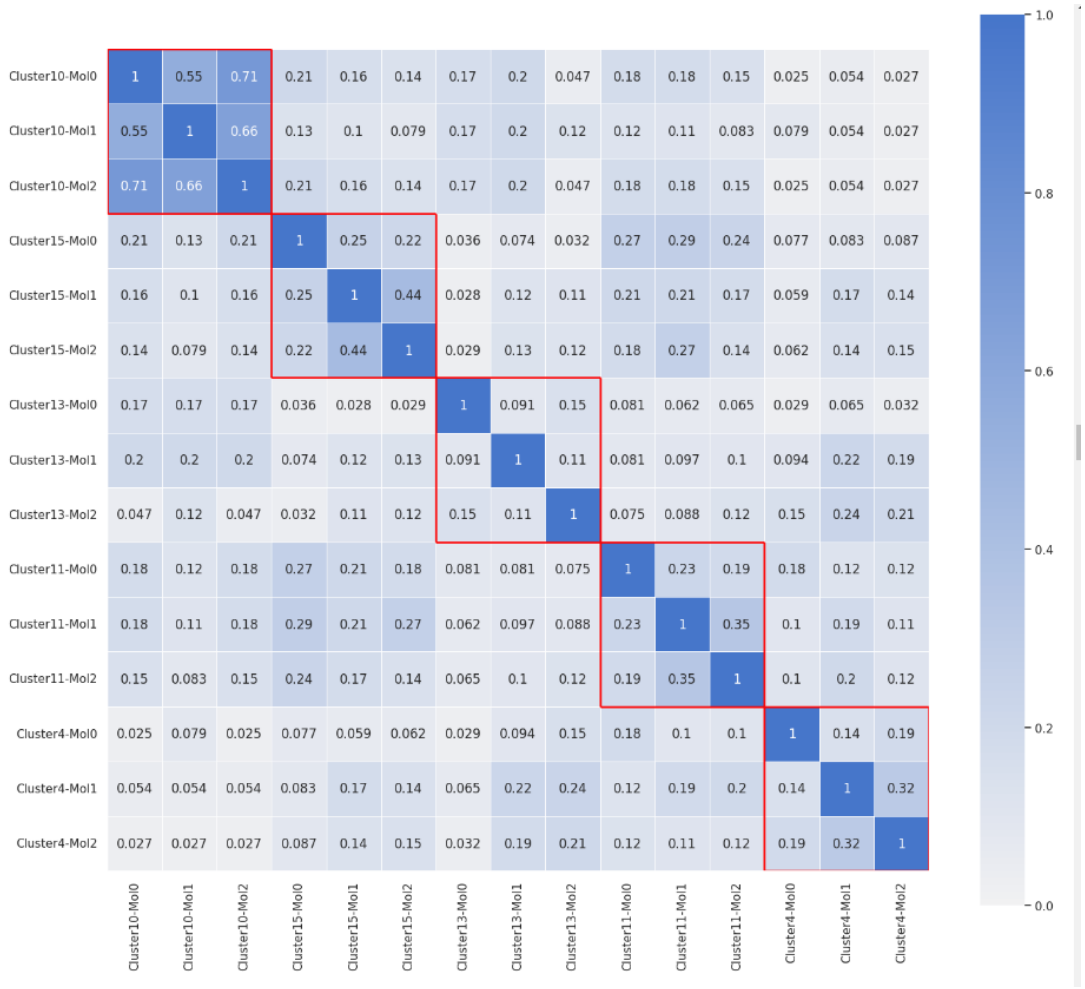
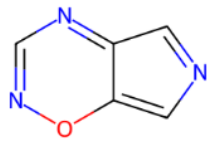
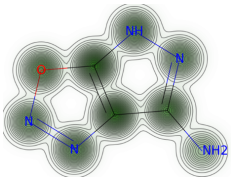
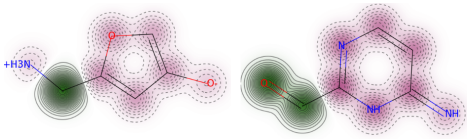
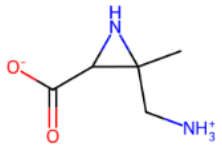
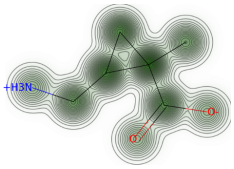
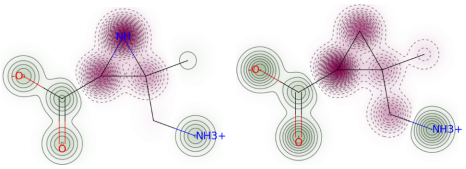
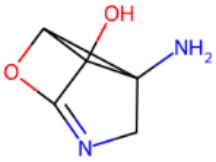
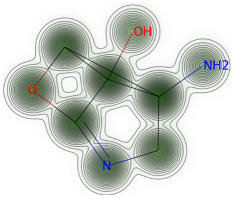
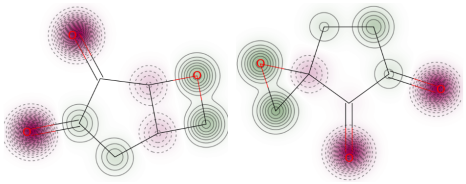
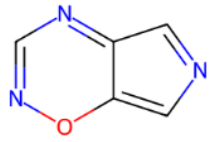
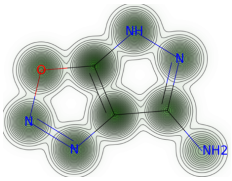
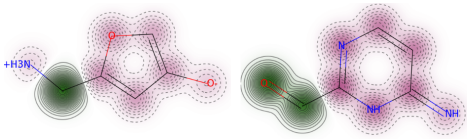
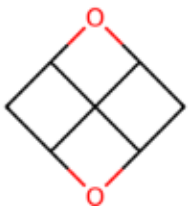
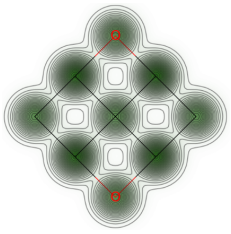
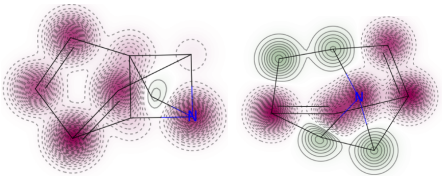


Figure 4: Tanimoto similarity matrix between molecules in clusters including the reference compounds (Mol0) and three test compounds (Mol1, Mol2, Mol3).

Table 1: : 2D structure and similarity map for the examples ran- domly selected in the five clusters. For each cluster, the similar- ity scores between the reference compound and three test com- pounds were measured by the Tanimoto metric using the count- based ECFP (radius = 2, bit = 2048). The similarity weights were visualized by colors on the structure (similarity maps). Sub- structures that increase the similarity score were presented in green, whereas red indicates the opposite.

Cluster	Reference Compound	Test Compounds	
4			
10			
11			
13			
15			

3.2 Molecular Optimization and Electronic Properties

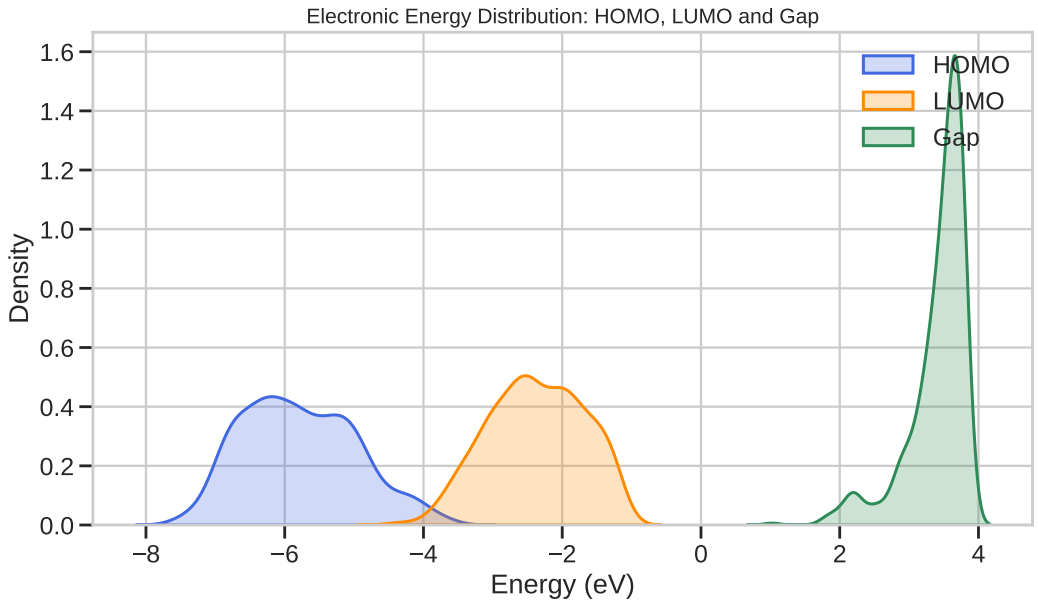


Figure 5: Probability density plots of the frontier molecular orbital energies (HOMO and LUMO) and the calculated HOMO-LUMO energy gap. The density functions were estimated from the dataset of selected molecules using kernel density estimation. These distributions provide insight into the electronic properties relevant for optoelectronic or sensing applications.

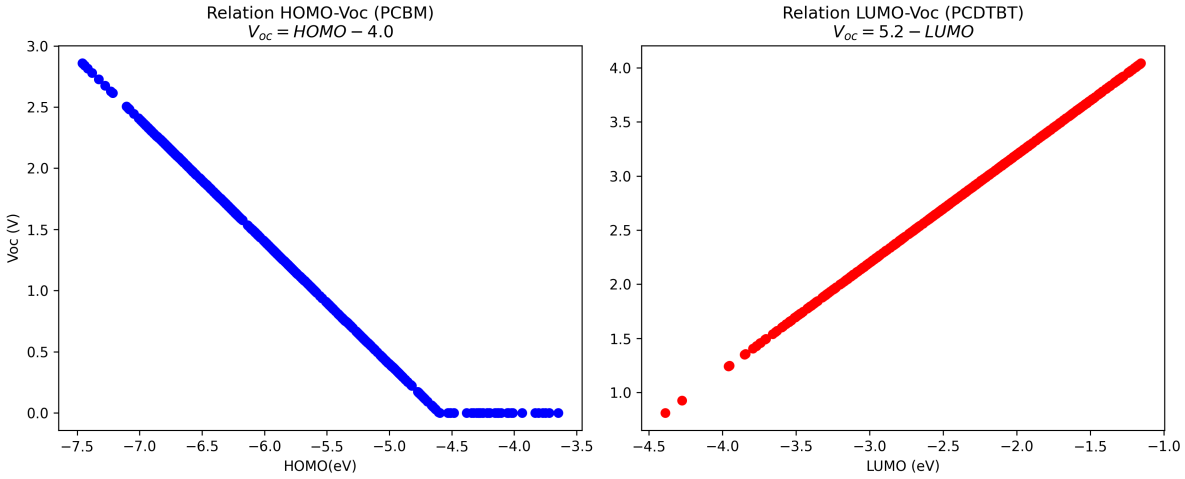


Figure 6: VOC PCBM distributions from optimization with GFN2.

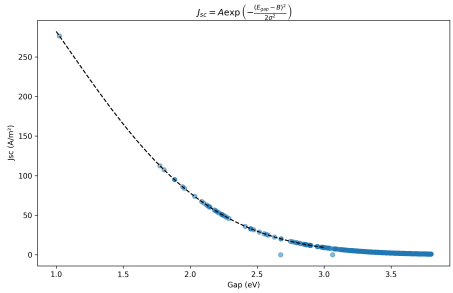


Figure 7: PCE SAscore GDB9 and HCE distributions from optimization with GFN2 .

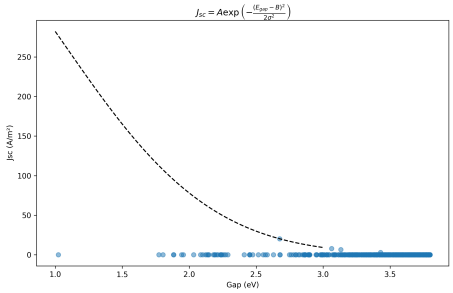


Figure 8: PCE GDB9 and HCE distributions from optimization with GFN2.

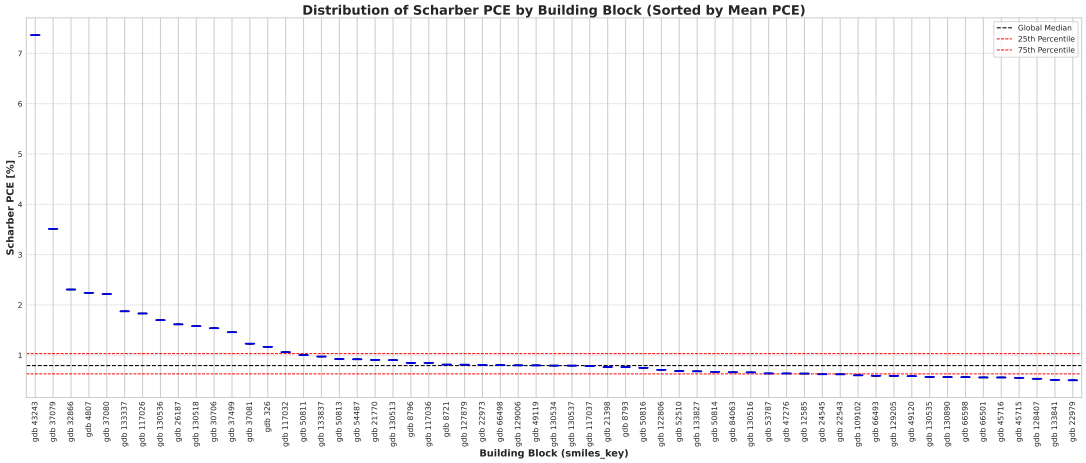


Figure 9: Key descriptors of selected GDB-9 donor and acceptor candidates. Physicochemical and bioactivity properties relevant to organic photovoltaic (OPV) compatibility with PCBM and PCDTBT.

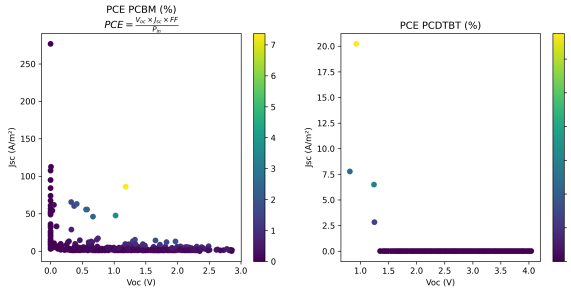


Figure 10: PCE contribution from optimization with GFN2.

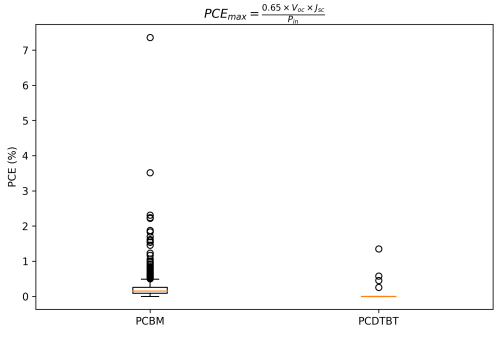


Figure 11: PCE GDB9 and HCE distributions from optimization with GFN2.

3.3 Synthetic Accessibility

3.4 Physicochemical and bioactivity properties

Discussion. The selected molecules from the GDB-9 dataset demonstrate diverse profiles in terms of polarity, lipophilicity, and three-dimensionality—key parameters for organic photovoltaic (OPV) performance.

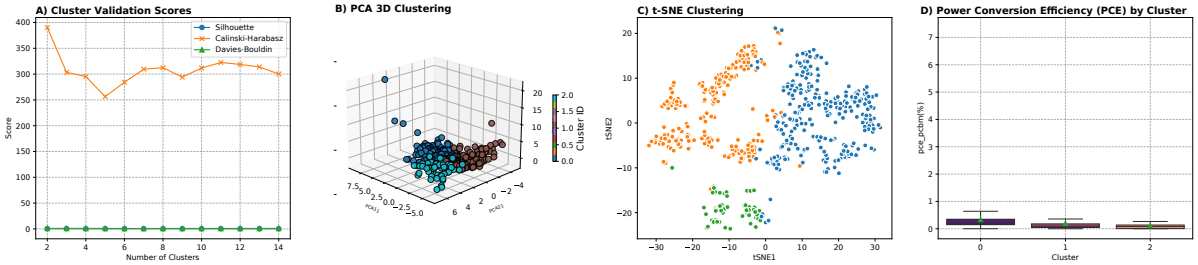


Figure 12: Cluster-based performance analysis for **Power Conversion Efficiency (PCE)**. **A)** Clustering validation scores (Silhouette, Calinski-Harabasz, Davies-Bouldin), **B)** Clustering in PCA-reduced space (2D projection), **C)** Clustering using t-SNE, **D)** Boxplot showing the PCE distribution across clusters. These figures demonstrate the diversity of photovoltaic behaviors in the dataset and justify the cluster segmentation.

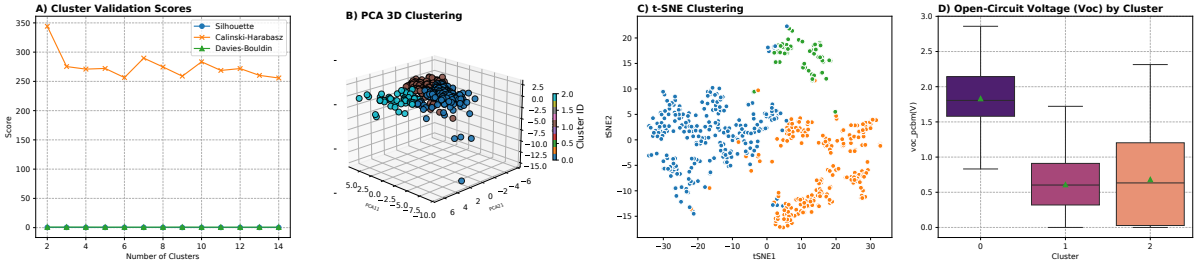


Figure 13: Cluster-based analysis for **Open-Circuit Voltage (Voc)**. **A)** Clustering quality evaluation, **B)** PCA space projection, **C)** t-SNE map of the molecular dataset, **D)** Voc distribution across the discovered clusters.

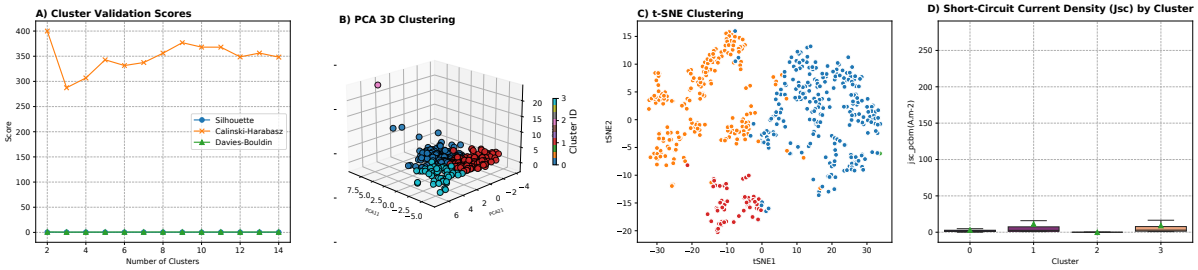


Figure 14: Clustering analysis of **Short-Circuit Current Density (Jsc)**. Panels **A-D** represent cluster validation, PCA-based segmentation, t-SNE visualization, and Jsc boxplot per cluster respectively. The figure reveals how distinct molecular subgroups contribute differently to photocurrent generation.

Notably, molecule A/D_8723 exhibits a strong predicted activity ($pIC_{50} = 8.36$) and moderate polarity (TPSA = 68.28), with a zero fraction of sp^3 carbons, suggesting a highly planar structure. These features are characteristic of effective electron acceptors, aligning well with properties observed in PCBM derivatives [38].

Conversely, D_43243 shows a more pronounced three-dimensional character (CSP3 = 0.29), a low topological polar surface area (TPSA = 20.08), and acceptable lipophilicity ($\log P = 0.08$). This combination suggests a good balance between solubility and stacking ability, similar to donor polymers like PCDTBT [39].

D_21398 presents high hydrogen bonding capacity (HBA = 5, HBD = 1) and strong polarity (TPSA = 75.41), which may limit its miscibility with common donor polymers, but enhance interaction with fullerene-like acceptors.

Beyond their photovoltaic potential, the favorable drug-likeness scores (e.g., QED = 0.451 for D_21398)

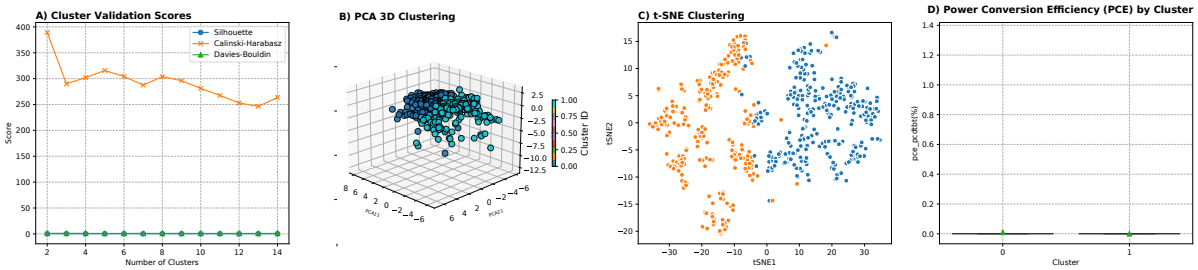


Figure 15: Cluster-based performance analysis for **Power Conversion Efficiency (PCE pcdtbt)**. **A)** Clustering validation scores (Silhouette, Calinski-Harabasz, Davies-Bouldin), **B)** Clustering in PCA-reduced space (2D projection), **C)** Clustering using t-SNE, **D)** Boxplot showing the PCE pcdtbt distribution across clusters. These figures demonstrate the diversity of photovoltaic behaviors in the dataset and justify the cluster segmentation.

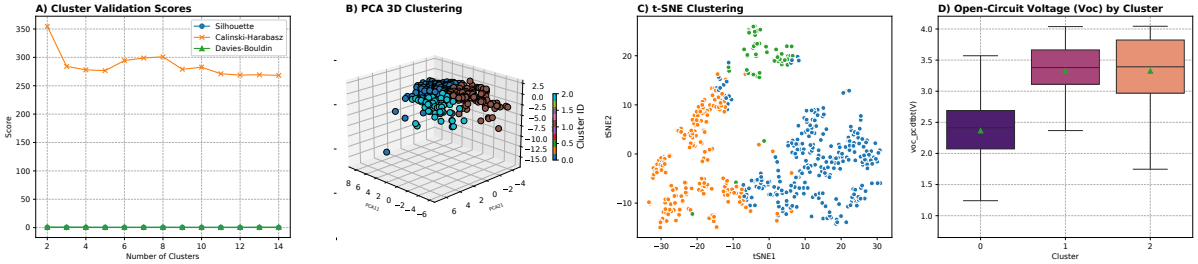


Figure 16: Cluster-based analysis for **Open-Circuit Voltage (Voc)**. **A)** Clustering quality evaluation, **B)** PCA space projection, **C)** t-SNE map of the molecular dataset, **D)** Voc pcdtbdistribution across the discovered clusters.

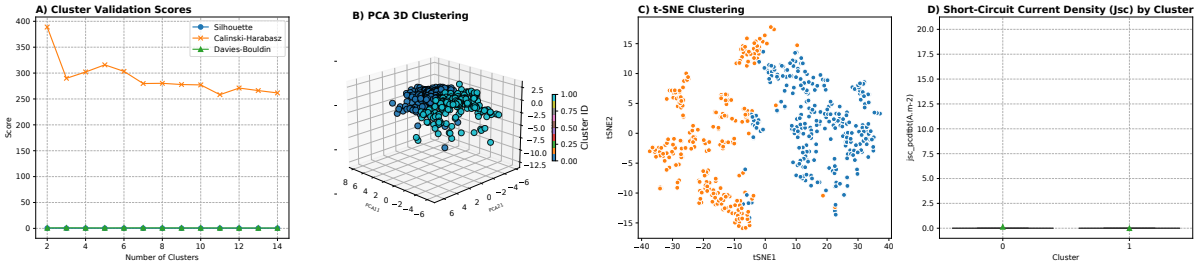


Figure 17: Clustering analysis of **Short-Circuit Current Density (Jsc)**. Panels **A-D** represent cluster validation, PCA-based segmentation, t-SNE visualization, and Jsc boxplot per cluster respectively. The figure reveals how distinct molecular subgroups contribute differently to photocurrent generation.

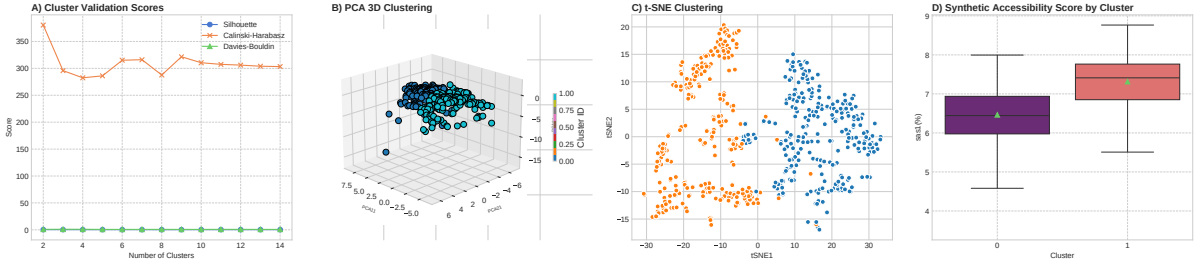


Figure 18: Clustering analysis of **SASCORE Density (Jsc)**. Panels **A-D** represent cluster validation, PCA-based segmentation, t-SNE visualization, and Jsc boxplot per cluster respectively. The figure reveals how distinct molecular subgroups contribute differently to photocurrent generation.

and bioactivity predictions (subnanomolar IC_{50}) suggest that these molecules may also hold promise

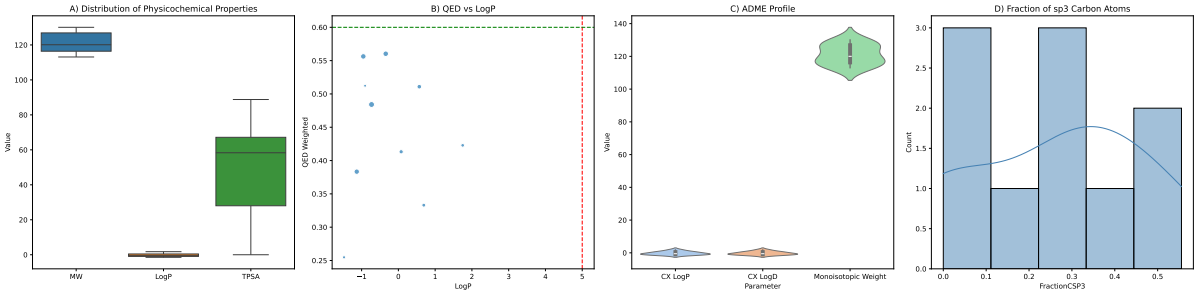


Figure 19: **Molecular property analysis of selected candidate molecules.** (**A**) Distribution of molecular weight (MW), lipophilicity (LogP), and polar surface area (TPSA). (**B**) Relationship between drug-likeness (QED) and LogP, with reference thresholds at LogP = 5 and QED = 0.6. (**C**) Violin plots of key ADME-related properties: predicted CX LogP, LogD, and monoisotopic weight. (**D**) Histogram of the fraction of sp3-hybridized carbons (FractionCSP3), indicating 3D character and molecular complexity.

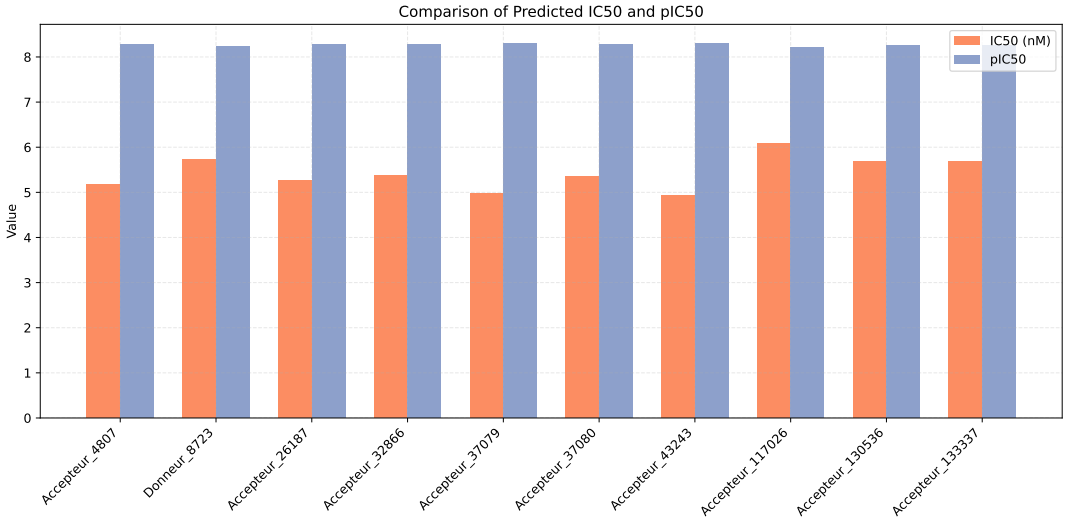


Figure 20: Key descriptors of selected GDB-9 donor and acceptor candidates. Physicochemical and bioactivity properties relevant to organic photovoltaic (OPV) compatibility with PCBM and PCDTBT.

in biomedical contexts. High polarity and hydrogen-bonding capacity, for instance, are often sought in enzyme inhibitors and CNS-active compounds [40].

Thus, these GDB-9-derived structures offer a dual-use perspective—applicable in both advanced materials for energy and prospective pharmaceutical agents—highlighting the interdisciplinary value of cheminformatics-driven molecule discovery.

Discussion The results presented in Table 2 reflect a diverse range of electronic and photophysical properties across selected GDB-9 molecules. Molecule **gdb_43243** exhibits a high fluorescence energy (3.56 eV) and a substantial singlet–triplet gap (1.60 eV), suggesting potential for blue-emitting organic materials. The moderate oscillator strength and long fluorescence lifetime (95 ns) further reinforce its applicability in OLEDs or fluorescence sensing.

Conversely, **gdb_21398** displays very low oscillator strength (0.0011) and a long computed lifetime ($>$24 μs), hinting at weak emissive properties but potentially strong intersystem crossing, which may be useful in triplet harvesting applications such as photodynamic therapy.

Molecule **gdb_66499** lacks a computed singlet–triplet gap, yet shows balanced fluorescence energy and oscillator strength, potentially indicating intermediate photostability and applicability in photoswitches

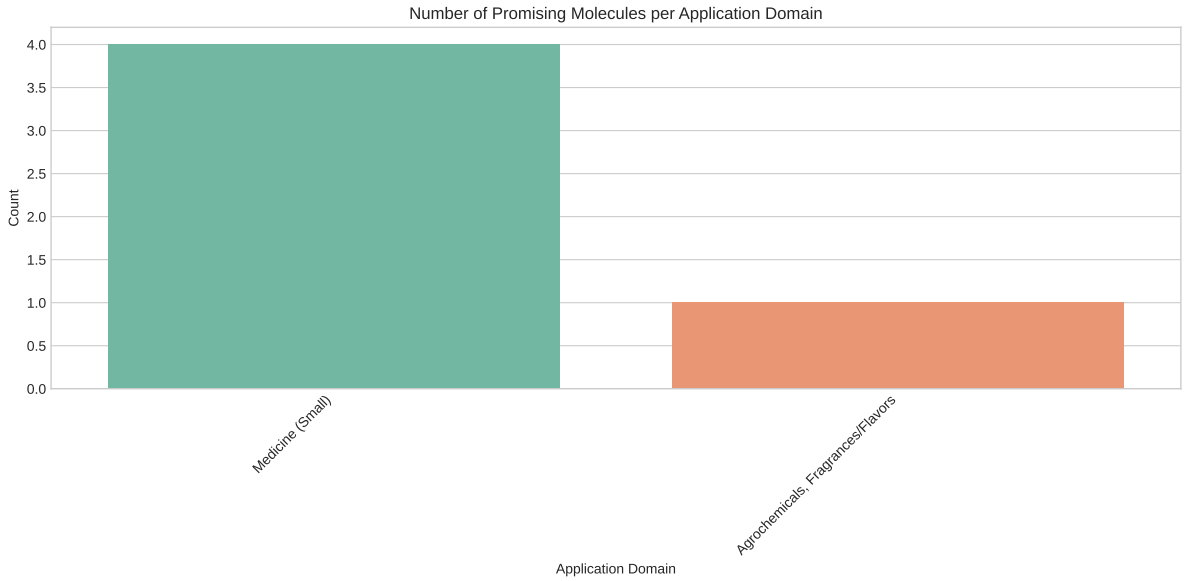


Figure 21: **Selected promising molecules for multiple applications.** Each structure is shown with its molecule ID and domain(s) of interest. Domains include medicinal chemistry, industrial use, optoelectronic materials, agrochemicals, and fragrances. The visual panel aids rapid qualitative assessment of chemical diversity and functional targeting.

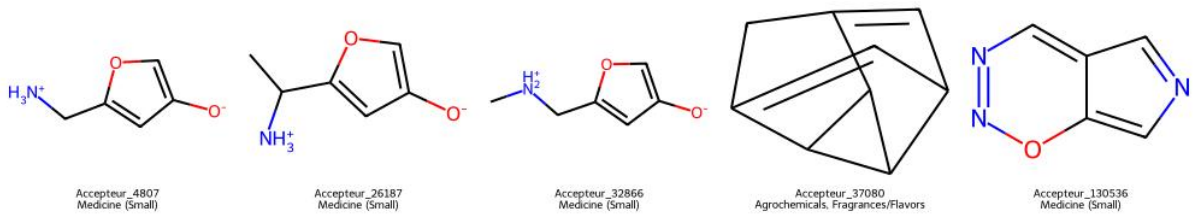


Figure 22: **Selected promising molecules for multiple applications.** Each structure is shown with its molecule ID and domain(s) of interest. Domains include medicinal chemistry, industrial use, optoelectronic materials, agrochemicals, and fragrances. The visual panel aids rapid qualitative assessment of chemical diversity and functional targeting.

	ID	SMILES	pIC50_pred	IC50_pred_nM	MW	Monoisotopic Weight	LogP	AlogP	HBA	HBD	TPSA	RotBonds	AromaticRings	Num Rotatable Bonds	HeavyAtoms	FractionCSP3	RingCount	CX LogP	CX LogD	QED Weighted
0	Accepteur_4807	[NH3+]CC1=CC([O-])=CO1	8.286561	5.169383	113.116	113.047678	-0.90490	-0.90490	2	1	63.84	1	1	1	8	0.200000	1	-0.90490	-0.90490	0.512328
1	Donneur_8723	O=CC(=O)C(=O)C=O	8.242031	5.727554	114.056	113.995309	-1.47760	-1.47760	4	0	68.28	3	0	3	8	0.000000	0	-1.47760	-1.47760	0.254758
2	Accepteur_26187	CC([NH3+])C1=CC([O-])=CO1	8.279668	5.252083	127.143	127.063329	-0.34390	-0.34390	2	1	63.84	1	1	1	9	0.333333	1	-0.34390	-0.34390	0.560325
3	Accepteur_32866	C[NH2+]CC1=CC([O-])=CO1	8.269564	5.375711	127.143	127.063329	-0.95360	-0.95360	2	1	52.81	2	1	2	9	0.333333	1	-0.95360	-0.95360	0.556328
4	Accepteur_37079	C1C2=CC3C=C1C1C2N31	8.302487	4.983258	117.151	117.057849	0.69150	0.69150	1	0	3.01	0	0	0	9	0.500000	4	0.69150	0.69150	0.333045
5	Accepteur_37080	C1C2=CC3C=C1C1C3C21	8.272173	5.343520	116.163	116.062600	1.74850	1.74850	0	0	0.00	0	0	0	9	0.555556	4	1.74850	1.74850	0.422835
6	Accepteur_43243	O=C1C=C2C=CC3C2N13	8.307311	4.928208	119.123	119.037114	0.07560	0.07560	1	0	20.08	0	0	0	9	0.285714	3	0.07560	0.07560	0.413207
7	Accepteur_117026	CNCC(=O)C(=O)C#N	8.216615	6.072743	126.115	126.042927	-1.13242	-1.13242	4	1	69.96	3	0	3	9	0.400000	0	-1.13242	-1.13242	0.383339
8	Accepteur_130536	O1N=NC=C2C=NC=C12	8.245246	5.685303	121.099	121.027612	0.56940	0.56940	4	0	51.81	0	0	0	9	0.000000	2	0.56940	0.56940	0.510871
9	Accepteur_133337	NC1=C(F)ON=NC1=N	8.245028	5.688161	130.082	130.029089	-0.72963	-0.72963	5	2	88.79	0	1	0	9	0.000000	1	-0.72963	-0.72963	0.484140

Figure 23: Key descriptors of selected GDB-9 donor and acceptor candidates. Physicochemical and bioactivity properties relevant to organic photovoltaic (OPV) compatibility with PCBM and PCDTBT.

	Molecule	Fluorescence_energy_eV	Singlet_Triplet_gap_eV	Oscillator_strength	Total_energy_eV	Electronic_energy_eV	Nuclear_repulsion_energy_eV	Lifetime_ns	Multi_Obj
0	gdb_4807	2.004171	1.448666	1.332798e-04	-10883.059154	-20047.781303	9164.722149	4.304836e+04	-2.644362
1	gdb_8723	1.365375	0.431849	1.721920e-09	-12368.305613	-21238.956767	8870.651154	7.179158e+09	-2.266474
2	gdb_26187	2.006155	1.470622	1.370127e-04	-11953.344933	-23086.911597	11133.566664	4.179270e+04	-2.664330
3	gdb_32866	2.002734	1.450168	1.317679e-04	-11952.814752	-22830.588139	10877.773387	4.360479e+04	-2.647302
4	gdb_37079	1.886451	1.855663	6.394219e-04	-9894.097166	-21746.011190	11851.914025	1.012773e+04	-3.168573
5	gdb_37080	1.877889	1.812170	1.006545e-03	-9458.032548	-21207.193352	11749.160804	6.492582e+03	-3.133275
6	gdb_43243	1.613365	1.234739	2.177999e-03	-10876.024446	-21711.610804	10835.586359	4.065062e+03	-2.819195
7	gdb_117026	1.889957	0.248941	1.336291e-04	-12357.422368	-23421.585954	11064.163586	4.828205e+04	-1.558850
8	gdb_130536	2.059034	0.707651	8.236995e-03	-11748.890156	-22828.145407	11079.255250	6.599250e+02	-1.840380
9	gdb_133337	1.406986	0.897272	7.505892e-05	-13882.878780	-25558.833861	11675.955081	1.550990e+05	-2.690212

Figure 24: Key descriptors of selected GDB-9 donor and acceptor candidates. Physicochemical and bioactivity properties relevant to organic photovoltaic (OPV) compatibility with PCBM and PCDTBT.

or tunable absorbers.

Finally, **gdb_8723** stands out with a relatively low singlet–triplet gap (0.51 eV) and negligible oscillator strength, which may imply effective non-radiative decay or charge transfer behavior. These characteristics are particularly valuable in organic photovoltaics or singlet fission systems.

Beyond optoelectronics, the variation in fluorescence lifetime and singlet–triplet splitting could be leveraged in biomedical imaging, phototherapeutics, or as photoreactive drug scaffolds. Continued screening with time-dependent DFT and excited-state dynamics simulations will be essential to refine these leads across domains.

4 Discussion

This study highlights the potential of computational approaches in designing advanced OPV materials. The PCE-SAS metric effectively balances efficiency and synthetic feasibility, providing a systematic method to prioritize candidate molecules for experimental validation.

However, several limitations were noted:

- **Reliance on Computational Models**: While tools like the Scharber model provide valuable insights, they are based on simplified assumptions that may not capture all real-world complexities.
- **Synthetic Challenges**: High SAS scores for many candidates indicate the need for further optimization to enhance synthetic feasibility.

5 Conclusion and Future Work

This study demonstrates the utility of inverse molecular design and high-throughput computational screening in advancing organic photovoltaic technology. Key findings include:

- Identification of promising donor and acceptor molecules with PCE values up to 3.01%.
- Challenges in synthesis feasibility, as indicated by high SAS scores for many candidates.
- The potential of the PCE-SAS metric to guide the selection of optimal materials.

Future efforts will focus on:

- Experimentally synthesizing and characterizing the proposed materials.
- Expanding molecular databases to include more diverse chemical structures.
- Leveraging machine learning to accelerate material discovery and improve predictive accuracy.

Acknowledgments

The authors thank the University of Yaoundé 1 and the Laboratory of Atomic, Molecular, and Biophysics for providing computational resources and support.

References

[1] Brabec, C. J., et al. "Organic photovoltaics: Technology and market development." *Advanced Materials*, vol. 22, no. 34, 2010, pp. 3839-3856. DOI: 10.1016/j.solmat.2004.02.030.

[2] Nkinyam, C. M., Ujah, C. O., Nnakwo, K. C., Kallon, D. V. V. "Insight into organic photovoltaic cell: Prospect and challenges." *ScienceDirect*, vol. 5, January 2025, 100121. DOI: 10.1016/j.uncres.2024.100121.

[3] Scharber, M. C., Mühlbacher, D., Koppe, M., Denk, P., Waldauf, C., Heeger, A. J., and Brabec, C. J. "Design rules for donors in bulk-heterojunction solar cells—Towards 10% energy-conversion efficiency." *Advanced Materials*, vol. 18, no. 6, 2006, pp. 789-794. DOI: 10.1002/adma.200501717.

[4] Spicher, S., and Grimme, S. "Robust Atomistic Modeling of Materials, Organometallic, and Biochemical Systems." *Nature Communications*, vol. 10, no. 1, 2019, Article 1. DOI: 10.1002/anie.202004239.

[5] United Nations. *Transforming Our World: The 2030 Agenda for Sustainable Development*. United Nations General Assembly, 2015. Available at: https://sdgs.un.org/2030agenda.

[6] Ramakrishnan, R., et al. "Quantum chemistry structures and properties of 134 kilo molecules." *Nature Scientific Data*, vol. 2, 2015, Article 150022. DOI: 10.1038/sdata.2014.22.

[7] Pracht, P., et al. "Automated exploration of the low-energy chemical space with fast quantum chemical methods." *Physical Chemistry Chemical Physics*, vol. 22, no. 14, 2020. DOI: 10.1039/D0CP00351D.

[8] Hachmann, J., et al. "The Harvard Clean Energy Project: Large-scale computational screening and design of organic photovoltaics on the world community grid." *The Journal of Physical Chemistry Letters*, vol. 2, no. 17, 2011. DOI: 10.1021/jz200866r.

[9] Strieth-Kalthoff, F., Glaser, S., Forreiter, C., Glorius, F., and Gasteiger, J. *Chemical Science*, 2022, vol. 13, pp. 7566-7584. DOI: 10.1039/D2SC00635H.

[10] Zhou, Z., Kearnes, S., Li, L., Zare, R. N., and Riley, P. *ACS Central Science*, vol. 3, 2017, pp. 353-368. DOI: 10.1021/acscentsci.7b00022.

[11] Ramakrishnan, R., Dral, P. O., Rupp, M., and von Lilienfeld, O. A. *Journal of Chemical Theory and Computation*, vol. 11, 2015, pp. 2087-2096. DOI: 10.1021/ct5000904.

[12] Rogers, D., and Hahn, M. *Journal of Chemical Information and Modeling*, vol. 50, 2010, pp. 742-754. DOI: 10.1021/ci100050t.

[13] Cereto-Massagué, N., Ojeda, M. J., Valls, E., and Mulero, J. *Journal of Cheminformatics*, vol. 7, 2015, pp. 1-15. DOI: 10.1186/s13321-015-0063-3.

[14] Sandfort, C., Gasteiger, J., and Glorius, F. *Chemical Science*, vol. 11, 2020, pp. 11944-11955. DOI: 10.1039/D0SC04783A.

[15] Walters, W. P., and Barzilay, R. *Journal of Chemical Information and Modeling*, vol. 61, 2021, pp. 2199-2207. DOI: 10.1021/acs.jcim.1c00070.

[16] Ruddigkeit, L., van Deursen, R., Blum, L. C., and Reymond, J.-L. *Journal of Chemical Information and Modeling*, vol. 52, 2012, pp. 2864-2875. DOI: 10.1021/ci300415d.

[17] Grimme, S., Bannwarth, C., and Shushkov, P. "A robust and accurate tight-binding quantum chemical method for the structures, vibrational frequencies, and non-covalent interactions of large molecular systems parameterized for all elements of the spd block (z = 1–86)." *Journal of Chemical Theory and Computation*, vol. 13, no. 5, 2017, pp. 1989–2009. DOI: 10.1021/acs.jctc.7b00287.

[18] Bannwarth, C., Ehlert, S., and Grimme, S. "GFN2-xTB — A highly accurate and widely parameterized self-consistent tight-binding quantum chemistry method with multipolar electrostatic contributions and density-dependent dispersion." *Journal of Chemical Theory and Computation*, vol. 15, no. 3, 2019, pp. 1652–1671. DOI: 10.1021/acs.jctc.8b01045.

[19] Blum, L. C., and Reymond, J.-L. "970 million drug-like small molecules for virtual screening in the chemical universe database gdb-13." *Journal of the American Chemical Society*, vol. 131, 2009, pp. 8732–8733. DOI: 10.1021/ja9022117.

[20] Pracht, P., Caldeweyher, E., Ehlert, S., and Grimme, S. "A robust non-self-consistent tight-binding quantum chemistry method for large molecules." *ChemRxiv*, 2019. DOI: 10.33774/chemrxiv.11577380.

[21] Bannwarth, C., Caldeweyher, E., Ehlert, S., Hansen, A., Pracht, P., Seibert, J., Spicher, S., and Grimme, S. "Extended tight-binding quantum chemistry methods." *WIREs Computational Molecular Science*, vol. 11, no. 2, 2021, Article e1493. DOI: 10.1002/wcms.1493.

[22] Lee, C., Yang, W., and Parr, R. G. "Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density." *Phys. B*, vol. 37, 1988, pp. 785–789.

[23] Becke, A. D. "Density functional thermochemistry. III. The role of exact exchange." *The Journal of Chemical Physics*, vol. 98, no. 7, 1993, pp. 5648–5652.

[24] O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., and Hutchison, G. R. "Open Babel: An open chemical toolbox." *Journal of Cheminformatics*, vol. 3, 2011, p. 33. DOI: 10.1186/1758-2946-3-33.

[25] Pracht, P., Bohle, F., and Grimme, S. "Automated exploration of the low-energy chemical space with fast quantum chemical methods." *Physical Chemistry Chemical Physics*, vol. 22, no. 14, 2020, pp. 7169–7192. DOI: 10.1039/D0CP00351D.

[26] Spicher, S., and Grimme, S. "Robust atomistic modeling of materials, organometallic systems, and biochemical systems." *Angewandte Chemie International Edition*, vol. 59, no. 36, 2020, pp. 15665–15673. DOI: 10.1002/anie.202004239.

[27] Tartarus, A., et al. "A Quantum Chemical Method for the Prediction of the Solar Cell Efficiency in Organic Photovoltaics." *Journal of Chemical Theory and Computation*, vol. 13, no. 5, 2017, pp. 1989–2009. DOI: 10.1021/acs.jctc.7b00368.

[28] Becke, A. D. "Density Functional Exchange-Energy Approximation with Correct Asymptotic Behavior." *Phys. A*, vol. 38, 1988, pp. 3098–3100.

[29] Hohenberg, P., & Kohn, W. (1964). Inhomogeneous electron gas. *Physical Review*, **136**(3B), B864. DOI: 10.1103/PhysRev.136.B864

[30] Perdew, J. P., Burke, K., & Ernzerhof, M. (1996). Generalized gradient approximation made simple. *Physical Review Letters*, **77**(18), 3865. DOI: 10.1103/PhysRevLett.77.3865

[31] Becke, A. D. (1993). Density-functional thermochemistry. III. The role of exact exchange. *The Journal of Chemical Physics*, **98**(7), 5648. DOI: 10.1063/1.464913

[32] Landrum, G. (2013). RDKit: Open-source cheminformatics. URL: www.rdkit.org.

[33] Bannwarth, C., Ehlert, S., & Grimme, S. (2019). GFN2-xTB—An accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *Journal of Chemical Theory and Computation*, **15**(3), 1652-1671. DOI: 10.1021/acs.jctc.8b01176.

[34] Ertl, P., & Schuffenhauer, A. (2009). Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of Cheminformatics*, **1**, 8. DOI: 10.1186/1758-2946-1-8.

[35] Sun, Q., Zhang, X., Banerjee, S., Bao, P., Barbry, M., Blunt, N. S., Bogdanov, N. A., et al. (2020). Recent developments in the PySCF program package. *Journal of Chemical Physics*, **153**(2), 024109. DOI: 10.1063/5.0006074.

[36] G. Dennler, M. C. Scharber and C. J. Brabec, *Adv. Mater.*, 2009, **21**, 1323-1338, DOI: 10.1002/adma.200801283.

[37] M. Riede, T. Mueller, W. Tress, S. Olthof, P. Stroehr and M. Pfeiffer, *Adv. Mater.*, 2011, **23**, 2729-2745, DOI: 10.1002/adma.201100505.

[38] Brabec, C. J., Sariciftci, N. S., & Hummelen, J. C. (2001). Plastic solar cells. *Advanced Functional Materials*, 11(1), 15–26. DOI: 10.1002/1616-3028(200102)11:1<15::AID-ADFM15>3.0.CO;2-A.

[39] Li, G., Shrotriya, V., Huang, J., Yao, Y., Moriarty, T., Emery, K., & Yang, Y. (2005). High-efficiency solution processable polymer photovoltaic cells by self-organization of polymer blends. *Nature Materials*, 4(11), 864–868. DOI: 10.1038/nmat1500.

[40] Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S., & Hopkins, A. L. (2012). Quantifying the chemical beauty of drugs. *Nature Chemistry*, 4(2), 90–98. DOI: 10.1038/nchem.1243.

[41] Liu, F., et al. Machine-learning-assisted inverse design of high-efficiency organic photovoltaic materials. *npj Comput Mater*, 2021, **7**, 1–11. DOI: 10.1038/s41524-021-00527-8.

[42] Bento, A. P., et al. "An open source chemical structure curation pipeline using RDKit." Journal of Cheminformatics, 2020. DOI:10.1186/s13321-020-00456-1.

[43] Polishchuk, P. G., et al. "Estimation of the size of drug-like chemical space based on GDB-17 data." Journal of Computer-Aided Molecular Design, 2013. DOI:10.1007/s10822-013-9642-8.

[44] Ghose, A. K., et al. "A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery." Journal of Combinatorial Chemistry, 1999. DOI:10.1021/cc9800071.

[45] Gomez-Bombarelli, R., et al. "Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach." Nature Materials, 2016. DOI: 10.1038/nmat4717.

[46] Veber, D. F., et al. "Molecular properties that influence the oral bioavailability of drug candidates." Journal of Medicinal Chemistry, 2002. DOI: 10.1021/jm020017n.

[47] Sun, H., et al. "Multifunctional Organic Materials for Flexible Bioelectronics." Chemical Reviews, 2023. DOI: 10.1021/acs.chemrev.2c00783

[48] Hachmann, J., et al. "The Harvard Clean Energy Project: Large-scale computational screening and design of organic photovoltaics on the world community grid." Journal of Physical Chemistry Letters, 2011. DOI: 10.1021/jz200866s .

[49] Reineke, S. "Complementary LED technologies." Nature Materials, 2013. DOI: 10.1038/nmat3648.

[50] Kawashima, Y., et al. "Design strategy for high oscillator strength molecules: From theory to OLED application." Organic Electronics, 2015. DOI: 10.1016/j.orgel.2015.05.029.

[51] Uoyama, H., et al. "Highly efficient organic light-emitting diodes from delayed fluorescence." Nature, 2012. DOI: 10.1038/nature11687.

[52] Goushi, K., et al. "Organic light-emitting diodes employing efficient reverse intersystem crossing for triplet-to-singlet state conversion." Nature Photonics, 2012. DOI: 10.1038/nphoton.2012.122.