Clarification of the Research Rationale and Objectives

Scientific Problem

Although extensive databases such as GDB-9 and CEPDB provide validated quantum chemical descriptors, their integration for cross-domain molecular discovery remains limited. In particular, while GDB-9 offers a vast repository of small molecules with computed HOMO-LUMO values and energy gaps, it lacks application-specific property ranges.

In contrast, CEPDB provides interval-based performance descriptors (HOMO, LUMO, gap, Voc, Jsc, and PCE) for a broad set of organic semiconductors, including optoelectronic materials. This makes it a useful reference database to define the physicochemical profile of high-performance molecules. Therefore, GDB-9 molecules were screened using property ranges derived from CEPDB, enabling the identification of small semiconductors with multifunctional potential.

Despite the atomic precision of ab initio DFT methods, their computational cost limits large-scale screening. Projects like the Clean Energy Project required considerable resources, often inaccessible to many labs. Hence, data-driven approaches remain essential for high-throughput discovery.

Research Question

How can we mine and refine GDB-9 using CEPDB-derived property intervals to identify molecules that are:
- Compatible with reference OPV materials (e.g., PCBM, PCDTBT),
- Synthetically feasible using cheminformatics scoring tools (e.g., SAScore),
- Applicable in multiple domains such as OLEDs and biomedical sciences?

Hypothesis

We hypothesize that a filtering approach based on:
1. Electronic descriptors from GDB-9,
2. Benchmark property ranges derived from CEPDB,
3. PCE predictions from the Scharber model,
4. Synthetic accessibility scoring (SAScore),
5. Physicochemical profiling (LogP, TPSA, QED),

will allow us to identify promising small molecules with synthetic compatibility and cross-domain functionality.

General Objective

Clarification of the Research Rationale and Objectives

To develop a computational strategy that combines quantum databases and cheminformatics to prioritize synthetically accessible small molecules for photovoltaic and biomedical use.

Specific Objectives

1. Use CEPDB to define HOMO, LUMO, and energy gap thresholds.
2. Filter GDB-9 molecules that fall within these electronic ranges.
3. Estimate PCE and Voc using the Scharber model.
4. Assess synthetic feasibility with SAScore.
5. Apply CEPDB photophysical descriptors (e.g., fluorescence energy, singlet-triplet gap) for secondary application potential.
6. Compute LogP, TPSA, QED to evaluate drug-likeness or sensory relevance.
7. Perform clustering to identify molecular families balancing performance and feasibility.