
FAST AND FLEXIBLE 3D MOLECULE DESIGN FRAMEWORK FOR NOVEL ORGANIC OPTOELECTRONIC MATERIALS

Kele Xu^{1,2,†,*}, Guojiang Zhao^{3,†}, Zheng Cheng^{4,†}, Tianjiao Wan^{1,2}, Ming Feng⁵, Shuqi Lu³,
Hongshuai Wang³, Zijian Gao^{1,2}, Qi Ou⁶, Liang Hu⁵,
Dawei Feng^{1,2}, Zifeng Zhao^{4,*}, Zhifeng Gao^{3,*}

¹College of Computer Science and Technology, National University of Defense Technology, Changsha, China

²National Key Laboratory of Parallel and Distributed Processing, Changsha, China

³DP Technology, Beijing, China

⁴AI for Science Institute, Beijing, China

⁵College of Electronic and Information Engineering, Tongji University, Shanghai, China

⁶SINOPEC Research Institute of Petroleum Processing Co., Ltd, Beijing, China

ABSTRACT

Organic optoelectronic materials (OOMs) are pivotal for advancing technologies such as organic photovoltaics and light-emitting diodes. Traditional methods for discovering new OOMs are inefficient and limited by chemical space exploration. We introduce *O*²-GEN, a novel framework leveraging a 3D pretraining backbone trained on a diverse dataset of over ten million molecules, enabling comprehensive exploration of chemical space. *O*²-GEN excels in generating fused-ring systems and conjugated fragment assemblies, achieving nearly 100% validity and novelty. It significantly outperforms existing models in speed and chemical structural validity, particularly for larger molecules. The framework supports both global and local generation modes, allowing for the creation of new molecules or modifications of existing structures. Additionally, *O*²-GEN integrates a property selector fine-tuned with density functional theory data, enabling precise multi-property screening. This framework offers a powerful tool for rational design and high-throughput screening of OOMs, with potential applications to drug discovery and energy materials.

1 Introduction

Over the past few decades, organic optoelectronic devices, as a cutting-edge technology, have drawn substantial research investments and have achieved remarkable commercial success, such as organic photovoltaics (OPVs) [1, 2], organic light-emitting diodes (OLEDs) [3, 4]. Organic optoelectronic molecules (OOMs) [5, 6, 7], serving as the fundamental constituents of organic optoelectronic devices, confer on these devices inherent attributes, including high efficiency, flexibility, lightweight, solution processability, and low cost. Thus, the exploration of novel OOM structures constitutes one of the primary driving forces propelling the progress within the organic optoelectronic device field.

OOMs typically consist of conjugated fused-ring system or conjugated fragments assemblies. This structural arrangement forms the basis for their optoelectronic activities. Evidently, the novelty of OOMs emerges from the diversity of the conjugated fused-rings and the variety of connection modalities between the conjugated fragments. Both factors exert an impact on the electronic structure of isolated molecules and packing mode in condensed state, further dictating the optoelectronic properties of OOMs and the performance of the associated devices [8, 9, 10].

The traditional approaches for searching for new OOMs have predominantly relied on the expertise of specialists, fortuitous discoveries, and inefficient trial-and-error procedures, which have imposed constraints on the exploration of chemical spaces [11, 12]. Recent advancements in deep learning [13, 14] have transformed such processes into in silico

[†]These authors contributed equally to this work.

^{*}Corresponding author.

rational design, accelerating the discovery of pharmaceutical compounds [15, 16, 17, 18], crystals [19, 20], energy materials [21] and small molecules [22].

However, existing deep generative models encounter significant difficulties when dealing with OOMs. Variational Autoencoders (VAE) [23, 24, 25, 26] and Generative Adversarial Networks (GAN) [27, 28, 29, 30], which utilize sequential or graph-based molecular representations, face challenges in generating accurate three-dimensional (3D) molecular conformations, which are important for the optoelectronic properties of OOMs. Recent endeavors integrating 3D-equivariant networks with diffusion [31, 32] have demonstrated promising outcomes in small molecule generation. Nevertheless, our experiments reveal that these methods exhibit slow convergence and low validity when applied to fused-ring molecules. Geometric Bayesian Flow models [33], despite harnessing 3D structural information and diminishing input noise to achieve faster convergence for smaller molecules, also yield low validity results for OOMs. As a pivotal development, Tomer Weiss et al. [11] proposed a method of mitigating complexity by using the graph of rings (GOR) representation. This approach significantly boosts the efficiency and validity of generating fused-ring molecules. However, this method restricts, to some extent, the chemical space to be searched and the novelty of the generated molecules [11].

In this work, we propose a universal framework, O^2 -GEN, for designing OOMs. O^2 -GEN adopts the virtual particle method [34], with a 3D pretraining backbone on 3D molecular structures facilitating its rational convergence. In the validation task on two fused-ring molecule datasets, new structures were generated rapidly with nearly 100% validity and novelty. For larger molecules composed of conjugated fragments, O^2 -GEN can generate complete molecules ~ 30 times faster than Diffusion and Flow models, with a tremendous improvement in the integrality of fragments assemblies. In particular, our framework has great flexibility, allowing designated local region modification of a known structure to empower researchers for fine-tuning molecular properties.

Furthermore, O^2 -GEN incorporates an efficient and accurate property selector by fine-tuning the pre-trained 3D molecular structure backbone with a limited amount of Density Functional Theory (DFT) [35] calculation data. This selector can be used to screen the generated molecules according to multiple downstream properties with near DFT accuracy. Through these means, the overarching objective of generating new-structured molecules with multi-objective characteristics is realized.

2 Results

2.1 Overview of O^2 -GEN

O^2 -GEN (Figure 1) encompasses a molecule generator specifically engineered for OOMs design, one (or multiple, depending on specific requirements) property selector, and a 3D pre-trained backbone network shared by both components. This architecture enables the efficient generation and property screening of OOMs.

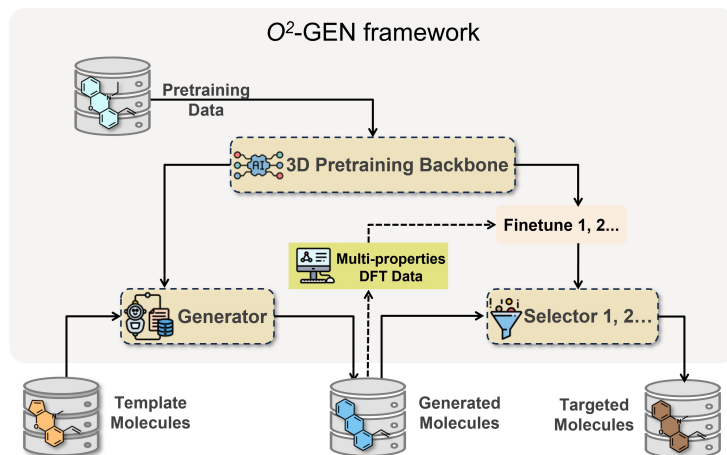


Figure 1: The workflow of O^2 -GEN

2.1.1 3D Pretraining

The pre-training data (Figure 2.a) consists of the open-source optoelectronic molecular datasets with pure condensed fused-ring molecules (COMPAS-1X [36], PAS [11]) and aromatic ring assemblies (CMR [37], Harvard Organic Photovoltaic Database [38], Ir complexes [7], TADF [39]). The detailed dataset names and their information are provided in Supplementary Table 1. In addition, considering that in many applications of OOMs require larger molecular weight, we developed a dataset named AOMAD (Aromatic Optoelectronic Molecular Assembly Database) through expert-guided fragment assembly (see Data part of Methods) to enhance the representativeness of molecular distributions in pre-training datasets. The pre-training dataset encompasses a total of over ten million molecules, meticulously curated to ensure the utmost diversity of OOMs.

We opted for the self-supervised pretraining representation learning approach of 3D molecular structures [40] (Figure 2.b) to better characterize the chemical space and the structural relationships of OOMs. This allows the 3D pretraining backbone to simultaneously learn the diversity of both the fused-ring structures and their connection modes.

Following insights from previous works [41, 42, 43], we analyzed our 3D pretraining backbone's learned representations using t-SNE method [44]. The distribution of randomly selected 80,000 molecules from the validation set is colored by the molecular weights. As illustrated in Figure 2.c, the map displays the wide chemical space captured by this backbone. The highlighted molecules show that those with closer distances on the t-SNE map exhibit higher similarities in topological structures and functional groups, confirming the backbone's ability to identify complex molecular structures. It is critical for OOMs generation and property prediction.

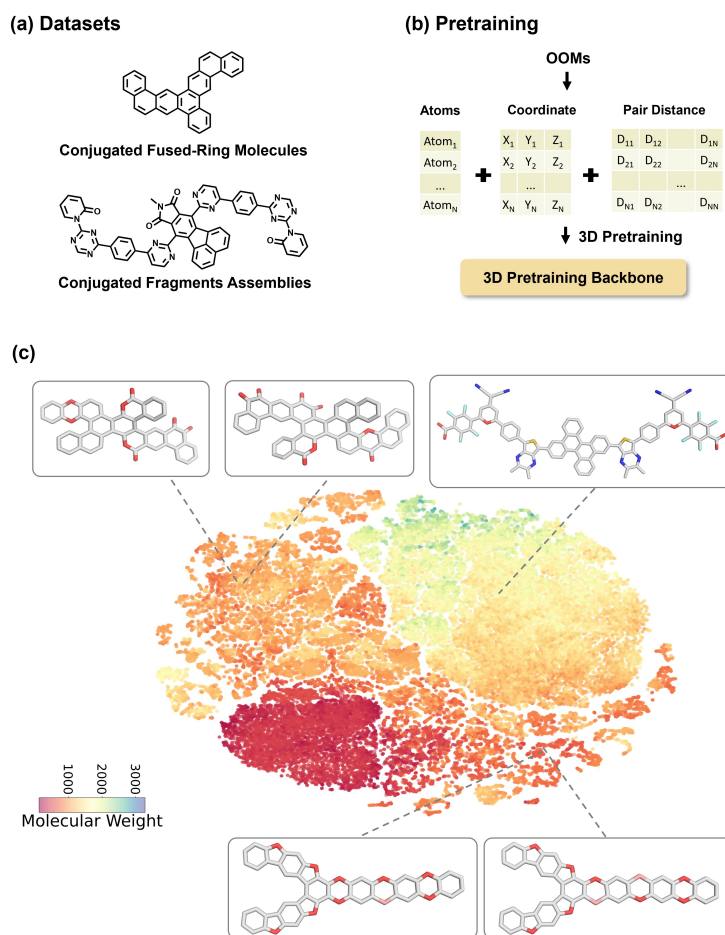


Figure 2: Details of 3D pretraining backbone. a) Data collection for OOMs; b) Pretraining process; c) t-SNE analysis.

2.1.2 Generator details and performances

The generator is inspired by the VD-GEN method [34]. Virtual Particles (VPs) are uniformly initialized within a predefined space and iteratively moved to regions of high-probability atomic distribution (Figure 3.a). This movement is guided by the pre-trained 3D backbone. Depending on different application scenarios, the generator can utilize specific molecular structures as initial templates to define the sampling space. This helps the virtual particles converge more rapidly during iteration, enhancing the generation speed, validity, and functionality of OOMs. Such features also enable us to modify and regenerate any part of a specific molecule, demonstrating the high flexibility of the generator. Next, we will evaluate the performance across different datasets. To ensure a fair comparison among different models, we fine-tuned the generator on the corresponding datasets prior to evaluation.

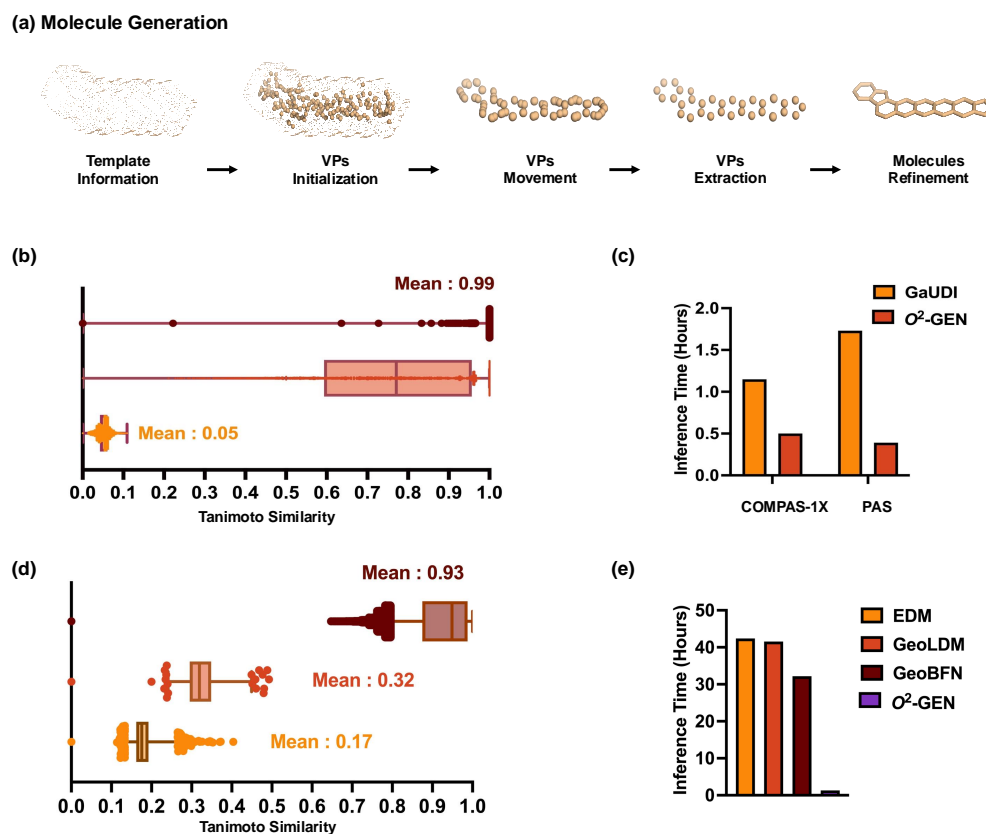


Figure 3: Details of Generator. a) The basic process of molecule generation by the Virtual Particle method; b) Molecular diversity in the COMPAS-1X dataset, comparing fingerprint similarity between each molecule and its closest match within the original dataset, among generated molecules following a 1:9 ratio, and between the generated and original datasets. The lines on the graph, from top to bottom, correspond respectively to these three categories. c) Comparison of inference times (hours) for processing 10,000 molecules across different methods and datasets. d) Molecular diversity in the AOMAD dataset, comparing fingerprint similarity between each molecule and its closest match within the original dataset, among generated molecules following a 1:9 ratio, and between the generated and original datasets. The lines on the graph, from top to bottom, correspond respectively to these three categories. e) Comparison of inference times (hours) for processing 10,000 molecules across different methods and datasets.

Generation task on fused-ring molecule databases The section demonstrates the framework's performance in generating novel and feasible fused-ring molecules. The framework was first tested on datasets like COMPAS-1X [36] and PAS [11], which contain fused-ring molecules with medium size. Analysis of atomic length distributions shows COMPAS-1X containing cata-condensed polybenzenoid hydrocarbons (cc-PBHs) within a range of 12 to 72, and PAS displaying a range from 10 to 66 in heterocyclic structures. The atomic length distributions are available in Supplementary Figure 1. We benchmarked against GaUDI [11] due to its state-of-the-art performance in generating

polycyclic molecules using EDM [31] and Graph of Rings (GOR) representations [45]. To validate our framework, we generated 10,000 molecules per dataset using the OOMs generator, assessing them based on the following metrics [11, 31, 32]: Validity (percentage of valid molecules as measured by RDKit [46]), Uniqueness (percentage of unique valid molecules), Novelty (percentage of valid molecules not in the training set), and V&U&N (the fraction of valid, unique and novel molecules that are not present in the training set) [47].

As shown in Table 1, both O^2 -GEN and GaUDI have V&U&N of over 97% on the PAS dataset. However, on the COMPAS-1X dataset, GaUDI's unique and novelty performance drops significantly, while O^2 -GEN maintains a V&U&N close to 100%. To account for this difference, we characterized the two datasets using the molecular fingerprint method [46]. The average fingerprint similarity of COMPAS-1X and PAS dataset (Figure 3.b and Supplementary Figure 2)) is 0.99 and 0.67 among their original molecules, respectively. This result indicates the molecular structure diversity of COMPAS-1X dataset is extremely limited, while the PAS dataset covers a much larger chemical space. GaUDI is strongly restricted by the diversity of the training data, while the 3D pretraining backbone of O^2 -GEN can provide prior information on pre-trained molecules. As a result (Figure 3.b), generated molecules showing an average molecular fingerprint with the similarity of 0.75. Further analysis revealed a significantly lower average fingerprint similarity of 0.05 between the generated molecules and those in the original dataset. A similar trend of change in molecular fingerprint similarity can also be observed on the PAS dataset (Supplementary Figure 2), indicating the successful exploration of larger chemical space by O^2 -GEN. On the other hand, the GOR method of GaUDI treats rings as single tokens, which accelerates the convergence of the model, but it also limits the diversity of explored molecular structures, diminishing the novelty of the generated molecules.

Furthermore, the O^2 -GEN framework significantly accelerates the speed of molecular generation. As shown in Figure 3.c, when generating 10,000 molecules, on the COMPAS-1X and PAS datasets, the molecular generation speed of O^2 -GEN increases by 56.52% and 77.46% respectively compared with GaUDI.

Table 1: The performance of GaUDI and O^2 -GEN on COMPAS-1X and PAS datasets

Datasets	COMPAS-1X				PAS			
Generative Models	Validity (%)	Unique (%)	Novelty (%)	V&U&N (%)	Validity (%)	Unique (%)	Novelty (%)	V&U&N (%)
GaUDI	99.40	86.57	20.66	17.78	99.64	99.93	97.51	97.09
O^2 -GEN	100.00	99.92	100.00	99.92	98.48	100.00	100.00	98.48

Generation task on conjugated fragments assemblies databases The generation of conjugated fragments assemblies offers another perspective for examining the generation model. To this end, we selected a recently published organic optoelectronic molecular dataset, FORMED [48]. It contains fragments assembled OOMs assembled by conjugated fragments. The atomic length distributions of FORMED range from 2 to 132. Considering that in many application scenarios of optoelectronic materials require larger molecular weight, we developed a larger dataset named AOMAD (Aromatic Optoelectronic Molecular Assembly Database) through expert - guided fragment assembly. The atomic length distributions of this dataset range from 66 to 338. Details are shown in Supplementary Figure 1. GaUDI [11] was excluded because its generator design is not suitable for tasks involving the conjugated fragments assemblies generation task. We compared EDM [31], GeoLDM [32], and GeoBFN [33], known for their proficiency in smaller molecular tasks, as new baseline models.

First, the performance of uniqueness and novelty for all models is close to 100% (except that the uniqueness ratio of molecules generated by EDM on AOMAD database is 88.56%). Therefore, the rows of uniqueness, novelty, and V&U&N are omitted in Table 2.

Table 2: The performance of baseline models and O^2 -GEN on FORMED and AOMAD datasets.

Datasets	FORMED			AOMAD		
Generative Models	Validity (%)	Connected Ratio (%)	Avg. Rings/Mol	Validity (%)	Connected Ratio (%)	Avg. Rings/Mol
EDM	87.39	39.01	2.89	75.79	0.00	4.38
GeoLDM	84.90	30.43	2.93	78.49	0.00	2.99
GeoBFN	70.78	51.90	2.85	49.32	1.69	9.59
O^2 -GEN	97.34	86.11	2.94	96.86	30.93	14.78

Nevertheless, the task of learning the distribution of fused-rings and the fragment connection modalities of molecules in FORMED and AOMAD poses a great challenge to the baseline models. The molecules generated by these methods almost all have a large number of ring breakages, overlaps, and other situations that violate the basic chemical rules as shown in Supplementary Figure 3, resulting in a low validity. As shown in Table 2, the validity of the baseline models ranges from 70% to 88% (FORMED datasets), and with the increase of molecular weight and the number of rings on AOMAD datasets, the validity of all baseline models significantly drops to 49%-79%. In contrast, O^2 -GEN still maintains the validity of nearly 97% on both datasets.

To further clearly evaluate the models' ability to generate integral fragments assembled OOMs, we define two new metrics: the connected ratio, which is the proportion of valid molecules where all fragments are connected to each other; and the Avg. Rings / Mol, which is the average number of rings in each valid molecule. On the FORMED dataset, 86.11% of the molecules generated by O^2 -GEN are intact, while this ratio for other baseline models ranges from 39% to 51.9%. On the AOMAD datasets, due to the increase in molecular size and the number of rings, the connection ratio of molecules generated by the three baseline models is close to 0, while O^2 -GEN still remains above 30%. The average number of rings in the obtained molecules is 14.78, which is also much higher than that of the baseline models. The introduction of pre-training 3D information enables our generator to more effectively construct connected fragments assembled OOMs.

Furthermore, we used the molecular fingerprint method to depict the specific characteristics of the novelty of the molecules generated by O^2 -GEN. As shown in Figure 3.d, the average fingerprint similarity between the most similar molecules in the original dataset is 0.93, while the similarity between the molecules we generated and the most similar molecule pairs in the original dataset is only 0.17. This indicates that our framework has the ability to generate novel large-scale OOMs in a broader chemical space.

It is worth noting that when dealing with complex large fragments assembled OOMs, the baseline models have difficulty converging within a limited time. For generating 10,000 molecules, the inference times of EDM, GeoLDM, and GeoBFN for generating molecules are as high as 42.40 h, 41.57 h, and 32.17 h, respectively; while the inference time of O^2 -GEN is only 1.3 h. This demonstrates the rapidity and efficiency of O^2 -GEN in handling such systems.

2.1.3 Selector details and performances

Accurate and fast property screening is crucial for identifying OOMs that simultaneously meet multiple property targets. Our selector is based on the reported molecular pre-training framework, Uni-Mol [40]. It shares the same 3D pre-trained backbone with the generator (Figure 1.a). By fine-tuning with specific DFT data, it can meet the prediction requirements for specific properties.

On the basis of the demand for predicting the properties of real-world OOMs, we selected key optical and redox properties such as the Highest Occupied Molecular Orbital Energy (HOMO), Lowest Unoccupied Molecular Orbital Energy (LUMO), Ionization Potential Energy (IP), Electron Affinity Energy (EA), and Energy of Absorption Maximum (Abs Max) as the prediction targets. Through finetuning with the thousands of DFT-level data points generated with O^2 -GEN, the correlation coefficient R^2 of all five properties between the predicted values of our selector and the DFT results all exceeded 0.88, with the mean absolute error (MAE) ranging from 0.04 to 0.13 eV. (For detailed calculation procedures, refer to the Method section, and see Supplementary Figure 4) All of these results demonstrate that our selectors are capable of predicting related properties with QM accuracy.

2.1.4 Case study: The non-fullerene acceptor generation of OPV

Global generation To further validate the ability of O^2 -GEN to create novel OOMs in real-world scenarios, we selected the fused-ring core of the star non-fullerene acceptor Y6 [49, 50] in the OPV field as the template molecule, and generated 600 brand-new molecules. Figure 4.a shows representative molecules in the generated dataset. Details of the statistical distribution of the structural data (number of atoms in the molecule, types of heavy atoms, number of rings...) are presented in Supplementary Figure 5.

Compared with the original template molecule, the generated molecules successfully retained a similar molecular size and contour, avoiding excessive divergence in the chemical space. At the same time, the generation model created a large number of novel and reasonable structures, such as new heteroatoms, peri-condensed fused-ring structures (i.e., an atom is shared by three rings simultaneously), and new fragment connection patterns. The generated molecules still retain a large conjugated system, which is the structural basis for the optoelectronic function of the molecule, providing a structural guarantee for subsequent screening.

Local generation From the perspective of experimental field experts, global generation could produce a large number of molecules with reasonable chemical structures but lacking in functionality. These molecules cannot directly assist researchers in quickly focusing on specific needs. Considering that the classic paradigm of actual scientific research is to make improvements based on successful cases, O^2 -GEN's local generation mode supports an arbitrary part of the template molecule as a variable region for generation. Compared with global generation mode, local generation further restricts the chemical space, thus improving the functionality of the generated structures. On the other hand, only synthesizable molecules have application value. The global generation mode of O^2 -GEN model does not consider the feasibility of synthetic methodology or routes. However, molecules generated through local generation can be derived from successfully synthesized molecules or building blocks, and ideally, they also have better synthesizability.

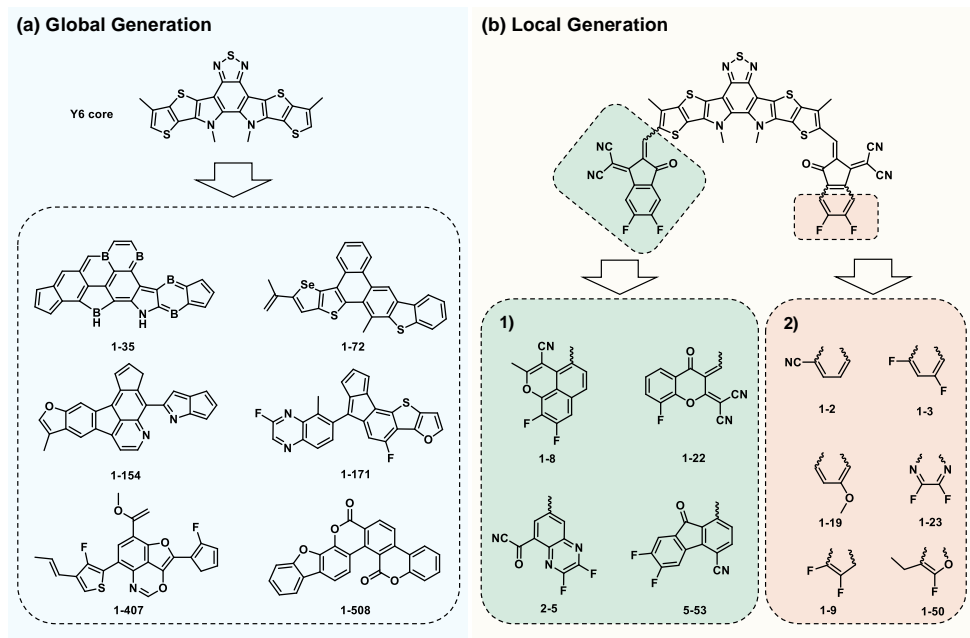


Figure 4: Global and local modes of O^2 -GEN to generate OOMs.

We take the terminal group of Y6 as an example to demonstrate the partial generation ability of O^2 -GEN and examine the obtained 521 molecules. The newly generated terminal groups exhibited rich and novel structures as shown in Figure 4.b1. Compared with the template molecule, the number and type of rings, the types of substituents, and heteroatoms also showed greater richness. Furthermore, we selected 6 atoms on half of the ring (Figure 4.b2) to further test the local generation mode for fused-ring fragments. Under this condition, the chemical space explored was further compressed. Nevertheless, O^2 -GEN still presented numerous novel and valuable structures in 150 generated samples.

Considering the requirement of wet experiment, non-fullerene acceptors should have a precise energy level matching with the donors and the adjacent charge-extraction layer, thus the ideal parameters depend on the researcher's system and objectives. Therefore, we present the distributions of the IP, EA, and Abs Max of the molecules generated under the two modes in Supplementary Figure 6. The diverse structures ensure the novelty of the molecules and the variety of downstream properties, creating sufficient scope for multi-target property screening.

3 Discussion

The performance of organic optoelectronic devices relies on OOMs. Although deep learning has sped up material discovery, current deep generative models face difficulties with fused-ring molecules and conjugated fragment assemblies. Our proposed framework, O^2 -GEN, provides a solution to address these challenges and offers a powerful tool for the needs of experimental researchers.

While O^2 -GEN excels in generating novel structures, the synthesizability of these molecules remains a challenge. Although local generation mode improves the likelihood of synthesizable structures, further integration with retrosynthetic analysis tools can enhance the practical applicability of the generated molecules.

The O^2 -GEN framework offers several key advantages over existing methods. The incorporation of a 3D pre-trained backbone allows the model to capture the complex structural relationships and electronic properties of OOMs, leading to more novel generation and accurate predictions. VPs method provides a fast and maneuverable way to generate molecules, endows the flexibility of O^2 -GEN in both global and local generation modes, and provides researchers with a powerful tool for fine-tuning molecular properties. This is particularly valuable in the context of organic optoelectronics, where small structural modifications can have a profound impact on device performance. The ability to generate molecules with specific properties, such as precise energy levels or absorption maxima, opens up new possibilities for the design of tailored materials for specific applications.

Besides, the potential applications of O^2 -GEN extend beyond organic optoelectronics. The framework's ability to generate and screen molecules with high accuracy and efficiency can be applied to other fields, such as drug discovery, catalysis, and energy storage. By leveraging the principles of rational design and high-throughput screening, O^2 -GEN could accelerate the discovery of novel materials across a wide range of disciplines.

4 Methods

4.1 Data

Details of AOMAD (Aromatic Optoelectronic Molecular Assembly Database): In the high-throughput generation of a dataset of candidate optoelectronic molecules with higher molecular weight, we employed commonly used optoelectronic molecular fragments from the literature to construct a database of candidate molecules based on predefined rules. Specifically, the conjugated molecules were composed of various types of building blocks, including donor (D), spacer (S), acceptor (A), and capping (C) units. The combination of fragments was achieved through an in-house algorithm that systematically enumerated all possible valid combinations of the selected fragments, adhering to the predefined architectural rules. Specifically, the binding sites of each fragment are marked with X in Supplementary Figure 7, and we connected the combined sites of different fragments to each other to form new molecules. The molecular geometry after splicing was finally optimized using the MMFF94 [51] force field. Two primary molecular architectures were generated: "CADAC" and "CSADASC" types. The CADAC molecules consist of a linear arrangement of capping, acceptor, donor, acceptor, and capping units, resulting in 18,480 unique structures. The CSADASC molecules, on the other hand, feature a more complex architecture with additional spacer units, yielding 1,677,622 distinct configurations. This approach ensures a diverse and comprehensive library of potential optoelectronic materials, facilitating subsequent computational screening and analysis for optimal molecular candidates.

4.2 3D Pretraining Backbone

To systematically investigate the structural properties of OOMs and efficiently explore their extensive chemical space, we have utilized Uni-Mol [40] for 3D pretraining. Uni-Mol is a universal 3D framework for molecular representation learning that effectively learns and outputs 3D positional information through pretraining with a large volume of unlabeled data. The model takes atom types and atomic coordinates as inputs, where atomic representations are initialized via an embedding layer. Pairwise representations are derived from invariant spatial positional encoding based on pairwise Euclidean distances, ensuring global rotational and translational invariance (SE(3)-invariance). Within the model, these representations interact through the self-attention mechanism of the Transformer [52], allowing for the capture of both local and global molecular features.

During pretraining, Uni-Mol focuses on two core tasks: 3D position reconstruction and masked atom prediction, both aimed at capturing the deep structural characteristics of OOMs. In the 3D position reconstruction task, the model learns molecular geometry by correcting perturbed atomic coordinates. Specifically, it recovers the original unperturbed 3D coordinates from randomly distorted inputs. To encode 3D positional information, Uni-Mol constructs spatially invariant position encodings based on the Euclidean distances between atomic pairs. These encodings facilitate the interaction between atomic and pairwise representations within the multi-head self-attention mechanism. Following the formulation in Uni-Mol, for any given atom pair (i, j) , the update at layer $(l + 1)$ is defined as:

$$q_{ij}^{l+1} = q_{ij}^l + \sum_{h=1}^H \frac{Q_i^{l,h} (K_j^{l,h})^T}{\sqrt{d}} \quad (1)$$

Here, q_{ij}^l denotes the pairwise representation at layer l , while $Q_i^{l,h}$ and $K_j^{l,h}$ represent the query and key embeddings of atoms i and j in the multi-head attention mechanism, respectively. The term d corresponds to the hidden representation dimension, and H denotes the number of attention heads.

In the pair-to-atom communication process, the model incorporates pairwise representations as bias terms in the self-attention mechanism, further enhancing its ability to model 3D structural information. The computation is formulated as:

$$\text{Attention}(Q_i^{l,h}, K_j^{l,h}, V_j^{l,h}) = \text{softmax} \left(\frac{Q_i^{l,h} (K_j^{l,h})^T}{\sqrt{d}} + q_{ij}^{l-1,h} \right) V_j^{l,h} \quad (2)$$

where $V_j^{l,h}$ represents the value embedding, and $q_{ij}^{l-1,h}$ denotes the pairwise representation from the previous layer.

To further refine molecular geometry, Uni-Mol learns to recover atomic positions from perturbed coordinates. The objective of this task is to compensate for relative atomic displacements, thereby predicting accurate 3D coordinates,

formulated as:

$$\hat{x}_i = x_i + \sum_{j=1}^n (x_i - x_j) \text{ReLU}((q_{ij}^L - q_{ij}^0)U) W \quad (3)$$

Here, x_i and \hat{x}_i denote the original and predicted coordinates, respectively. The pairwise representations q_{ij}^0 and q_{ij}^L correspond to the initial and final states, while U and W are trainable projection matrices. In the masked atom prediction task, the model needs to identify randomly masked atomic types within OOMs. This process helps the model learn both local atomic environments and global molecular interactions. Through these pretraining tasks, Uni-Mol serves as the 3D pretraining backbone of O^2 -GEN, effectively encoding key structural information for downstream tasks such as molecular generation and property screening.

4.3 O^2 -GEN's Generator

To perform OOMs generation, two 3D pretraining backbones are utilized, providing efficient and accurate atomic-level representations by leveraging extensive 3D structural information. This enhances the structural validity and convergence speed of generated molecules. The generator in O^2 -GEN is inspired by the VD-GEN [34], which initializes VPs uniformly within a predefined space. Under the guidance of the 3D pretraining backbone, these VPs are iteratively moved toward regions with high-probability atomic distributions. During generation, the initialization of VPs can be tailored to specific applications by incorporating predefined molecular structures as templates, thereby defining the sampling space.

Firstly, given the prediction space \mathbf{S} , n VPs are randomly and uniformly distributed in the space (where n represents the number of atoms in the reference molecule). Let \mathbf{V}_0 denote the initialized VPs. Given \mathbf{V}_0 , their distribution is updated by moving them in 3D space to approximate the real atomic distribution. The movement of VPs follows an iterative process, where in each iteration, the model takes the positions and types of VPs from the previous step as input and predicts their updated positions and types. This process, referred to as Particle Movement (PM), can be formulated as:

$$\mathbf{V}_{r+1} = f_{pm}(\mathbf{V}_r, \mathbf{S}; \theta_{pm}), \quad \mathbf{V}_r = \{(\mathbf{x}_i^r, \mathbf{y}_i^r)\}_{i=1}^n \quad (4)$$

where \mathbf{V}_r represents the set of n virtual particles at iteration r , and f_{pm} is an SE(3)-equivariant model that takes 3D coordinates as input, corresponding to our 3D pretraining backbone. The term θ_{pm} denotes the learnable parameters. The variables $\mathbf{x}_i^r \in \mathbb{R}^t$ and $\mathbf{y}_i^r \in \mathbb{R}^3$ represent the atomic type and coordinates, respectively, where t is the number of atomic types. The number of iterations in the particle movement stage is denoted as R_1 . Since \mathbf{V}_{R_1} generated in the PM stage is an approximation of atomic distributions rather than molecular distribution, a Molecule Extraction (ME) step is required to refine VPs into meaningful molecular representations. It employs the pretraining backbone h_{me} to filter and aggregate n virtual particles into m atoms, formulated as:

$$\mathbf{W}_0 = h_{me}(\mathbf{V}_{R_1}, \mathbf{S}; \theta_{me}) = \{(\hat{\mathbf{x}}_i^0, \hat{\mathbf{y}}_i^0)\}_{i=1}^m \quad (5)$$

where θ_{me} represents learnable parameters, and \mathbf{W}_0 denotes the extracted molecular structure composed of m atoms.

Since some VPs may deviate significantly from their target positions, those exceeding a predefined threshold z are filtered out. Next, the model predicts the merging probability between each pair of VPs and constructs a probability matrix to cluster VPs accordingly. Once the initial molecular structure \mathbf{W}_0 is obtained, further optimization of its 3D coordinates is required to refine the molecular geometry. Similar to Particle Movement, this step, referred to as Molecule Refinement (MR), follows an iterative process in which atomic positions are continuously adjusted using a different set of model parameters. The refinement process is formulated as:

$$\mathbf{W}_{r+1} = f_{mr}(\mathbf{W}_r, \mathbf{S}; \theta_{mr}), \quad \mathbf{W}_r = \{(\hat{\mathbf{x}}_i^r, \hat{\mathbf{y}}_i^r)\}_{i=1}^m \quad (6)$$

where f_{mr} is an SE(3)-equivariant model with learnable parameters θ_{mr} , corresponding to our second 3D pretraining backbone. The term \mathbf{W}_r represents the atomic set at iteration r , and R_2 denotes the total number of refinement iterations, ultimately yielding the final generated molecule.

Each stage has its corresponding loss function. In the Particle Movement stage, the goal of the f_{pm} model is to move VPs to the positions of the template molecule atoms so that the distribution of VPs approximates the atomic distribution of the molecule. Each VP is directly assigned a real atom as a training target. Following the principle of least action [53], assigning each VP to the nearest real atom as the training target is the optimal choice. Given the real atomic distribution:

$$\mathbf{G} = \{(\mathbf{x}_i^g, \mathbf{y}_i^g)\}_{i=1}^m : a_i = \arg \min_{j=1}^m \|\mathbf{y}_i^0 - \mathbf{y}_j^g\|_2 \quad (7)$$

where \mathbf{y}_i^0 represents the initial position of the i -th VP. Given the target a_i , the loss function in this stage consists of four components: the negative log-likelihood (NLL) loss for VP types, a truncated L2 loss for VP 3D coordinates, an L1 distance loss for VP-VP pairs and VP-space pairs, and a regularization loss to constrain movement between consecutive iterations. Therefore, at iteration r , the training loss function for this stage is expressed as:

$$\begin{aligned}\mathcal{L}_{pm} = & \frac{1}{n} \sum_{i=1}^n (\text{NLL}(\bar{\mathbf{x}}_i^r, \mathbf{x}_{a_i}^g) + \text{clip}(\|\mathbf{y}_i^r - \mathbf{y}_{a_i}^g\|_2, \tau)) \\ & + \frac{1}{n} \sum_{j=1}^n \|\mathbf{d}_{ij}^r - \mathbf{d}_{a_i, a_j}^g\|_1 + \frac{1}{u} \sum_{j=1}^u \|\mathbf{c}_{ij}^r - \mathbf{c}_{a_i, j}^g\|_1 \\ & + \max(\|\mathbf{y}_i^r - \mathbf{y}_i^{r-1}\|_2 - \delta, 0),\end{aligned}\quad (8)$$

where $\bar{\mathbf{x}}_i^r$ ($\mathbf{x}_{a_i}^g$) represents the predicted (true) atomic type of the i -th VP; \mathbf{y}_i^r ($\mathbf{y}_{a_i}^g$) represents the predicted (true) coordinates of the i -th VP; τ is the truncation threshold for the coordinate loss; \mathbf{d}_{ij}^r (\mathbf{d}_{a_i, a_j}^g) is the predicted (true) distance between VP pairs; \mathbf{c}_{ij}^r ($\mathbf{c}_{a_i, j}^g$) is the predicted (true) distance between VPs and spatial atoms; and δ is the threshold for movement constraint.

It is worth noting that during training, the iteration step r is uniformly sampled between 1 and R , where R is the maximum number of iterations (for this stage, $R = R_1$). Before iteration r , the model operates only in forward mode, without computing the loss or performing backpropagation. Finally, at iteration r , gradient computation and backpropagation are enabled. During inference, iterative sampling is not used.

The Molecule Extraction stage involves two tasks. The first task is to predict the deviation between VPs and their target positions, where the error is binned into one-hot encoding, converting it into a classification task to ensure stable training. The second task is to predict which VPs should be merged. Ideally, VPs assigned to the same target atom should be merged. Thus, for VP pairs with the same target atom, the training label is set to True. However, in practice, the distribution of True and False labels in VP pairs is highly imbalanced, so Focal Loss (FL) is adopted to handle this issue. The loss function for the molecule extraction stage is defined as:

$$\mathcal{L}_{fe} = \mathcal{L}_{\text{error_pred}} + \mathcal{L}_{\text{merge}} = \frac{1}{n} \sum_{i=1}^n \text{NLL}(\bar{\mathbf{s}}_i, \mathbf{s}_i) + \frac{1}{l^2} \sum_{i=1}^l \sum_{j=1}^l \text{FL}(\bar{\mathbf{r}}_{ij}, \mathbf{r}_{ij}) \quad (9)$$

where \mathbf{s}_i represents the one-hot encoded target error vector; $\bar{\mathbf{s}}_i$ represents the predicted probability vector; l is the number of filtered virtual particles; \mathbf{r}_{ij} is the target merging type for VP pairs; and $\bar{\mathbf{r}}_{ij}$ is the predicted merging type probability.

The Molecule Refinement stage is very similar to the Particle Movement stage, but its training target is slightly different. Specifically, the training target of the i -th atom is the most frequently occurring target atom in its cluster w_i , rather than the nearest atom. Formally, the target atom of the i -th atom is defined as:

$$b_i = \text{most_frequent}(\{a_j \mid j \in w_i\}) \quad (10)$$

4.4 VPs Initialization

Given the near-planar nature of most OOMs, our approach begins with an operation that rotates the molecular coordinates closer to the XY plane, utilizing an equivariant rotation matrix transformation. This process aims to improve the training process by aligning with OOMs' intrinsic geometrical properties and facilitating improved comprehension of atomic distributions. We utilize a method based on rotation matrices and Singular Value Decomposition (SVD) [54] for rotating the three-dimensional coordinates of a given molecule to be as near parallel to the x-y plane as possible:

First, we calculate the centroid \vec{C} of the molecular coordinates to identify the geometric center of the molecule. The centroid is derived from the arithmetic mean of all coordinate points in the matrix $Y \in \mathbb{R}^{n \times 3}$:

$$\vec{C} = \frac{1}{N} \sum_{i=1}^N \mathbf{y}_i \quad (11)$$

where N is the number of atoms. This step is essential for relocating the coordinate points to a coordinate system centered at the centroid.

To achieve the goal of aligning the molecule as closely as possible with the x-y plane, we identify a normal vector \vec{N} that indicates the molecule's present orientation. This vector is derived by performing SVD on the coordinates once they have been adjusted by subtracting the centroid. Within the matrix \mathbf{V} , which contains the right singular vectors from this decomposition, the vector \vec{N} is identified as the column corresponding to the smallest singular value. This column, usually the final one in \mathbf{V} , serves to represent \vec{N} , defining the molecular orientation in relation to the x-y plane.

Given \vec{N} and its relation to the z-axis, we employ Rodrigues' rotation formula [55] to devise a rotation matrix \mathbf{R} that orients \vec{N} with the z-axis. This formula is represented as:

$$\mathbf{R} = \mathbf{I} + \mathbf{K} + \mathbf{K}^2 \frac{1 - c}{s^2} \quad (12)$$

Here, \mathbf{I} denotes the identity matrix, and \mathbf{K} is the skew-symmetric matrix based on the cross product $\vec{v} = \vec{N} \times \vec{z}$, where $\vec{z} = [0, 0, 1]$. The dot product $c = \vec{N} \cdot \vec{z}$ measures the cosine of the angle between \vec{N} and \vec{z} , and $s = \|\vec{v}\|$ is the magnitude of \vec{v} , indicating the sine of the angle between \vec{N} and \vec{z} . The skew-symmetric matrix \mathbf{K} is constructed from \vec{v} 's components v_x, v_y, v_z as follows:

$$\mathbf{K} = \begin{bmatrix} 0 & -v_z & v_y \\ v_z & 0 & -v_x \\ -v_y & v_x & 0 \end{bmatrix} \quad (13)$$

Then, by multiplying the matrix Y with the rotation matrix, we obtain the rotated coordinates Y' , thereby bringing the molecular representation closer to the x-y plane. This is mathematically represented as:

$$Y' = Y \cdot \mathbf{R} \quad (14)$$

Next, we construct the initial VPs space for molecules oriented to closely align with the x-y plane. We present two efficient sampling strategies. For the coarse-grained sampling strategy, this approach involves sampling VPs within a broadly predefined space, without any prior molecular structure information. During training, the number of VPs is dynamically adjusted according to the total number of atoms in the molecule, falling within specified lower and upper ratio bounds. This adjustment allows the VPs to adapt to the size of the molecule. A cubic space is constructed using atomic coordinates, within which VPs are uniformly generated. The VPs are then continuously shifted by the generator to model the actual atomic distribution. Furthermore, to reduce resource consumption, the initial space for VPs can be defined based on the smallest volume constructed by projecting the molecular atomic coordinates on the x, y, and z axes (this can be implemented via PCA [56]). During inference, this predefined space distribution from the training set is used to initialize space for generation. For the fine-grained sampling strategy, considering the need for efficient generation of large, complex OOMs that fit real-world scenarios, we employ prior spatial structural information for effective generation and structural refinement. Specifically, we construct resource-optimization spatial grids based on the spatial structure information of template molecules and initialize VPs within these grids around each atom to better simulate the complex atomic distribution of such OOMs. This approach enables faster iterative generation of the desired structural OOMs.

4.5 Screening Selector Finetuning

We utilize a specialized 3D pretraining backbone tailored for OOMs to build a regression model. This model learns generated molecular conformations and their properties through DFT calculations. Due to the high cost of DFT calculations, only a subset of samples undergo DFT calculations to obtain their properties. Uni-Mol's generalization capabilities enable broader applicability with minimal data. The model is then trained using these conformations as inputs and their DFT properties as labels, optimizing with a smooth L1 loss function:

$$\text{Smooth}_{L1}(\hat{z}, z) = \begin{cases} 0.5(\hat{z} - z)^2 & \text{if } |\hat{z} - z| < 1 \\ |\hat{z} - z| - 0.5 & \text{otherwise} \end{cases} \quad (15)$$

The overall loss is then calculated as the sum of the smooth L1 loss for all predictions:

$$\text{Loss} = \sum_{i=1}^N \text{Smooth}_{L1}(\hat{z}_i, z_i) \quad (16)$$

where \hat{z} (or \hat{z}_i) represents the value predicted by the model, and z (or z_i) represents the actual target value.

After experimentation, we discovered that combining the original conformations with force-field optimizations from RDKit improves model performance. Moreover, compared to traditional inverse design prediction models, which are time-intensive, the proposed screening selector achieves rapid convergence, potentially increasing efficiency.

Finetuning data production:

All Density Functional Theory (DFT) and Time-Dependent Density Functional Theory (TD-DFT) [57] calculations were performed using Gaussian 16 [58]. The functional employed was B3LYP [59], and the basis set chosen was def2-SVP [60]. Dispersion corrections were considered using the D3BJ [61] method during the calculations. For different downstream tasks, the datasets are organized as follows:

- 1) HOMO/LUMO: total of 7,130 data points, with 5,704 for training and 1,426 for validation.
- 2) IP: 5,249 data points, including 4,200 for training and 1,049 for validation.
- 3) EA: 4,648 data points, with 3,683 used for training and 965 for validation.
- 4) Abs Max: A total of 5,356 data points, comprising 4,285 for training and 1,071 for validation.

These datasets were split using an 8:2 cross-validation ratio, where 80% of the data was used for training and 20% for validation.

4.6 Models Training Hyperparameters

3D Pretraining Backbone:

We trained the Uni-Mol [40] using the Adam[62] optimizer with a learning rate set to 0.0001. The training process lasted for approximately 40 epochs, during which 15% of the atoms were masked. The batch size was set to 128, and the model consisted of 15 layers. Training was conducted on 8 Tesla-A100-80G GPUs.

O²-GEN's Generator Pretraining and Inference:

For both global and local generation modes, we adopted the 3D pretraining backbone designed for OOMs (as detailed in the previous section) and employed the VP initialization strategy. The generator was trained using the Adam optimizer with a learning rate of 0.0001. The training process lasted for 1,000,000 steps, equivalent to approximately 17 epochs. The batch size was set to 4, and the model consisted of 9 layers. Training was conducted on 4 Tesla-A100-80G GPUs. For inference, our generator and other baselines were all generated using a single V100 GPU. For detailed information, please refer to Supplementary Table 2 and Supplementary Table 3.

Screening Selector:

Downstream Selector Task - Highest Occupied Molecular Orbital/Lowest Unoccupied Molecular Orbital energy

We constructed a regression model based on the aforementioned 3D pretraining backbone to train the selector. The model was trained using the Adam optimizer with a learning rate of 0.00005, a batch size of 64, and consisted of 15 layers. The training process lasted for 1,000 epochs.

Downstream Selector Task - Energy of Maximum Absorption Peak

We constructed a regression model based on the aforementioned 3D pretraining backbone to train the selector. The model was trained using the Adam optimizer with a learning rate of 0.0003, a batch size of 32, and consisted of 13 layers. The training process lasted for 1,000 epochs.

Downstream Selector Task - Energy of Ionization Potential

We constructed a regression model based on the aforementioned 3D pretraining backbone to train the selector. The model was trained using the Adam optimizer with a learning rate of 0.0003, a batch size of 128, and consisted of 14 layers. The training process lasted for 1,000 epochs.

Downstream Selector Task - Energy of Electron Affinity

We constructed a regression model based on the aforementioned 3D pretraining backbone to train the selector. The model was trained using the Adam optimizer with a learning rate of 0.0003, a batch size of 16, and consisted of 15 layers. The training process lasted for 1,000 epochs.

5 Acknowledgments

This work is supported by National Science and Technology Major Project (2023ZD0121101).

References

- [1] Albert Polman, Mark Knight, Erik C Garnett, Bruno Ehrler, and Wim C Sinke. Photovoltaic materials: Present efficiencies and future challenges. *Science*, 352(6283):aad4424, 2016.
- [2] Soyeon Kim, Adi Prasetyo, Joo Won Han, Yongki Kim, Myunghun Shin, Jinhee Heo, Jung Ha Kim, Shinuk Cho, Yong Hyun Kim, Muhammad Jahandar, et al. Enhanced flexible optoelectronic devices by controlling the wettability of an organic bifacial interlayer. *Communications Materials*, 2(1):26, 2021.
- [3] Hiroki Uoyama, Kenichi Goushi, Katsuyuki Shizu, Hiroko Nomura, and Chihaya Adachi. Highly efficient organic light-emitting diodes from delayed fluorescence. *Nature*, 492(7428):234–238, 2012.
- [4] Minseok Jeong, Joonyoung F Joong, Jinhyo Hwang, Minhi Han, Chang Woo Koh, Dong Hoon Choi, and Sungnam Park. Deep learning for development of organic optoelectronic devices: efficient prescreening of hosts and emitters in deep-blue fluorescent oleds. *npj Computational Materials*, 8(1):147, 2022.
- [5] Oksana Ostroverkhova. Organic optoelectronic materials: mechanisms and applications. *Chemical reviews*, 116(22):13279–13412, 2016.
- [6] Panpan Yu, Yonggang Zhen, Huanli Dong, and Wenping Hu. Crystal engineering of organic optoelectronic materials. *Chem*, 5(11):2814–2853, 2019.
- [7] Zheng Cheng, Jiapeng Liu, Tong Jiang, Mohan Chen, Fuzhi Dai, Zhifeng Gao, Guolin Ke, Zifeng Zhao, and Qi Ou. Automatic screen-out of ir (iii) complex emitters by combined machine learning and computational analysis. *Adv Opt Mater*, 11(18):2301093, 2023.
- [8] Vinayak Bhat, Connor P Callaway, and Chad Risko. Computational approaches for organic semiconductors: from chemical and physical understanding to predicting new materials. *Chemical Reviews*, 123(12):7498–7547, 2023.
- [9] Yasuhiko Shirota and Hiroshi Kageyama. Organic materials for optoelectronic applications: Overview. *Handbook of Organic Materials for Electronic and Photonic Devices*, pages 3–42, 2019.
- [10] Hongmei Zhuo, Xiaojun Li, Jinyuan Zhang, Can Zhu, Haozhe He, Kan Ding, Jing Li, Lei Meng, Harald Ade, and Yongfang Li. Precise synthesis and photovoltaic properties of giant molecule acceptors. *Nature Communications*, 14(1):7996, 2023.
- [11] Tomer Weiss, Eduardo Mayo Yanes, Sabyasachi Chakraborty, Luca Cosmo, Alex M Bronstein, and Renana Gershoni-Poranne. Guided diffusion for inverse molecular design. *Nature Computational Science*, 3(10):873–882, 2023.
- [12] Cheng-Han Li and Daniel P Tabor. Generative organic electronic molecular design informed by quantum chemistry. *Chemical Science*, 14(40):11045–11055, 2023.
- [13] Addis S Fuhr and Bobby G Sumpter. Deep generative models for materials discovery and machine learning-accelerated innovation. *Frontiers in Materials*, 9:865270, 2022.
- [14] Edward O Pyzer-Knapp, Jed W Pitera, Peter WJ Staar, Seiji Takeda, Teodoro Laino, Daniel P Sanders, James Sexton, John R Smith, and Alessandro Curioni. Accelerating materials discovery using artificial intelligence, high performance computing and robotics. *npj Computational Materials*, 8(1):84, 2022.
- [15] Jaechang Lim, Seongok Ryu, Jin Woo Kim, and Woo Youn Kim. Molecular generative model based on conditional variational autoencoder for de novo molecular design. *Journal of Cheminformatics*, 10(1):1–9, 2018.
- [16] Miha Skalic, Davide Sabbadin, Boris Sattarov, Simone Sciabola, and Gianni De Fabritiis. From target to drug: generative modeling for the multimodal structure-based ligand design. *Molecular Pharmaceutics*, 16(10):4282–4291, 2019.
- [17] Mingyuan Xu, Ting Ran, and Hongming Chen. De novo molecule design through the molecular generative model conditioned by 3d information of protein binding sites. *Journal of Chemical Information and Modeling*, 61(7):3240–3254, 2021.
- [18] Matthew Ragoza, Tomohide Masuda, and David Ryan Koes. Generating 3d molecules conditional on receptor binding sites with deep generative models. *Chemical Science*, 13(9):2701–2713, 2022.
- [19] Tian Xie, Xiang Fu, Octavian-Eugen Ganea, Regina Barzilay, and Tommi Jaakkola. Crystal diffusion variational autoencoder for periodic material generation. *Preprint at arXiv:2110.06197*, 2021.
- [20] Luis M Antunes, Keith T Butler, and Ricardo Grau-Crespo. Crystal structure generation with autoregressive large language modeling. *Preprint at arXiv:2307.04340*, 2023.
- [21] Hongshuai Wang, Yujin Ji, and Youyong Li. Simulation and design of energy materials accelerated by machine learning. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 10(1):e1421, 2020.
- [22] Seiji Takeda, Toshiyuki Hama, Hsiang-Han Hsu, Victoria A Pionova, Dmitry Zubarev, Daniel P Sanders, Jed W Pitera, Makoto Kogoh, Takumi Hongo, Yenwei Cheng, et al. Molecular inverse-design platform for material industries. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2961–2969, 2020.

- [23] Changsheng Ma and Xiangliang Zhang. Gf-vae: a flow-based variational autoencoder for molecule generation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge management*, pages 1181–1190, 2021.
- [24] Dakuo He, Qing Liu, Yan Mi, Qingqi Meng, Libin Xu, Chunyu Hou, Jinpeng Wang, Ning Li, Yang Liu, Huifang Chai, et al. De novo generation and identification of novel compounds with drug efficacy based on machine learning. *Advanced Science*, page 2307245, 2024.
- [25] Orion Dollar, Nisarg Joshi, Jim Pfandtnr, and David AC Beck. Efficient 3d molecular design with an e (3) invariant transformer vae. *The Journal of Physical Chemistry A*, 127(37):7844–7852, 2023.
- [26] Nutaya Pravalphruekul, Maytus Piriyaajakonkij, Phond Phunchongharn, and Supanida Piyayotai. De novo design of molecules with multi-action potential from differential gene expression using variational autoencoder. *Journal of Chemical Information and Modeling*, 2023.
- [27] Nicola De Cao and T Kipf. Molgan: An implicit generative model for small molecular graphs. arxiv 2018. *Preprint at arXiv:1805.11973*, 2019.
- [28] Xiaohong Liu, Wei Zhang, Xiaochu Tong, Feisheng Zhong, Zhaojun Li, Zhaoping Xiong, Jiacheng Xiong, Xiaolong Wu, Zunyun Fu, Xiaoqin Tan, et al. Molfiltergan: a progressively augmented generative adversarial network for triaging ai-designed molecules. *Journal of Cheminformatics*, 15(1):1–14, 2023.
- [29] Paula A Marin Zapata, Oscar Méndez-Lucio, Tuan Le, Carsten Jörn Beese, Jörg Wichard, David Rouquié, and Djork-Arné Clevert. Cell morphology-guided de novo hit design by conditioning gans on phenotypic image features. *Digital Discovery*, 2(1):91–102, 2023.
- [30] Young Jae Lee, Hyungu Kahng, and Seoung Bum Kim. Generative adversarial networks for de novo molecular design. *Molecular Informatics*, 40(10):2100045, 2021.
- [31] Emiel Hoogetboom, Víctor García Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *International Conference on Machine Learning*, pages 8867–8887. PMLR, 2022.
- [32] Minkai Xu, Alexander Powers, Ron Dror, Stefano Ermon, and Jure Leskovec. Geometric latent diffusion models for 3d molecule generation. In *International Conference on Machine Learning*. PMLR, 2023.
- [33] Yuxuan Song, Jingjing Gong, Hao Zhou, Mingyue Zheng, Jingjing Liu, and Wei-Ying Ma. Unified generative modeling of 3d molecules with bayesian flow networks. In *The Twelfth International Conference on Learning Representations*, 2024.
- [34] Shuqi Lu, Lin Yao, Xi Chen, Hang Zheng, Di He, and Guolin Ke. 3d molecular generation via virtual dynamics. *Transactions on Machine Learning Research*, 2024.
- [35] Robert G Parr and Weitao Yang. Density-functional theory of the electronic structure of molecules. *Annual Review of Physical Chemistry*, 46(1):701–728, 1995.
- [36] Alexandra Wahab, Lara Pfuderer, Eno Paenurk, and Renana Gershoni-Poranne. The compas project: A computational database of polycyclic aromatic systems. phase 1: cata-condensed polybenzenoid hydrocarbons. *Journal of Chemical Information and Modeling*, 62(16):3704–3713, 2022.
- [37] David D Landis, Jens S Hummelshøj, Svetlozar Nestorov, Jeff Greeley, Marcin Dułak, Thomas Bligaard, Jens K Nørskov, and Karsten W Jacobsen. The computational materials repository. *Computing in Science & Engineering*, 14(6):51–57, 2012.
- [38] Steven A Lopez, Edward O Pyzer-Knapp, Gregor N Simm, Trevor Lutzow, Kewei Li, Laszlo R Seress, Johannes Hachmann, and Alán Aspuru-Guzik. The harvard organic photovoltaic dataset. *Scientific data*, 3(1):1–7, 2016.
- [39] Dingyun Huang and Jacqueline M Cole. A database of thermally activated delayed fluorescent molecules auto-generated from scientific literature with chemdataextractor. *Scientific Data*, 11(1):80, 2024.
- [40] Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: A universal 3d molecular representation learning framework. In *The Eleventh International Conference on Learning Representations*, 2023.
- [41] Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287, 2022.
- [42] Xiangxiang Zeng, Hongxin Xiang, Linhui Yu, Jianmin Wang, Kenli Li, Ruth Nussinov, and Feixiong Cheng. Accurate prediction of molecular properties and drug targets using a self-supervised image representation learning framework. *Nature Machine Intelligence*, 4(11):1004–1016, 2022.
- [43] Han Li, Ruotian Zhang, Yaosen Min, Dacheng Ma, Dan Zhao, and Jianyang Zeng. A knowledge-guided pre-training framework for improving molecular representation learning. *Nature Communications*, 14(1):7568, 2023.
- [44] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11), 2008.
- [45] Tomer Weiss, Alexandra Wahab, Alex M Bronstein, and Renana Gershoni-Poranne. Interpretable deep-learning unveils structure–property relationships in polybenzenoid hydrocarbons. *The Journal of Organic Chemistry*, 2023.
- [46] Greg Landrum et al. Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum*, 8:31, 2013.

- [47] Han Huang, Leilei Sun, Bowen Du, and Weifeng Lv. Learning joint 2-d and 3-d graph diffusion models for complete molecule generation. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.
- [48] J Terence Blaskovits, Ruben Laplaza, Sergi Vela, and Cl  mence Corminboeuf. Data-driven discovery of organic electronic materials enabled by hybrid top-down/bottom-up design. *Advanced Materials*, 36(2):2305602, 2024.
- [49] Jun Yuan, Yunqiang Zhang, Liuyang Zhou, Guichuan Zhang, Hin-Lap Yip, Tsz-Ki Lau, Xinhui Lu, Can Zhu, Hongjian Peng, Paul A. Johnson, Mario Leclerc, Yong Cao, Jacek Ulanski, Yongfang Li, and Yingping Zou. Single-junction organic solar cell with over 15 *Joule*, 3(4):1140–1151, 2019.
- [50] Safa Shoaee, Hoang M. Luong, Jiage Song, Yingping Zou, Thuc-Quyen Nguyen, and Dieter Neher. What we have learnt from pm6:y6. *Advanced Materials*, 36(20):2302005, 2024.
- [51] Thomas A Halgren. Merck molecular force field. i. basis, form, scope, parameterization, and performance of mmff94. *Journal of computational chemistry*, 17(5-6):490–519, 1996.
- [52] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.
- [53] Richard Feynman. *The Character of Physical Law, with new foreword*. MIT press, 2017.
- [54] V. Klema and A. Laub. The singular value decomposition: Its computation and some applications. *IEEE Transactions on Automatic Control*, 25(2):164–176, 1980.
- [55] Olinde Rodrigues. Des lois g  om  triques qui r  gissent les d  placements d’un syst  me solide dans l’espace, et de la variation des coordonn  es provenant de ces d  placements consid  r  s ind  pendamment des causes qui peuvent les produire. *Journal de Math  matiques Pures et Appliqu  es*, 5:380–440, 1840.
- [56] Andrzej Ma  kiewicz and Waldemar Ratajczak. Principal components analysis (pca). *Computers & Geosciences*, 19(3):303–342, 1993.
- [57] Miguel AL Marques and Eberhard KU Gross. Time-dependent density functional theory. In *A Primer in Density Functional Theory*, pages 144–184. Springer, 2003.
- [58] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox. Gaussian~16 Revision C.01, 2016. Gaussian Inc. Wallingford CT.
- [59] A Becke. Density-functional thermochemistry. iii. the role of exact exchange (1993) j. *Chem. Phys*, 98:5648.
- [60] Florian Weigend and Reinhart Ahlrichs. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for h to rn: Design and assessment of accuracy. *Physical Chemistry Chemical Physics*, 7(18):3297–3305, 2005.
- [61] Stefan Grimme, Stephan Ehrlich, and Lars Goerigk. Effect of the damping function in dispersion corrected density functional theory. *Journal of computational chemistry*, 32(7):1456–1465, 2011.
- [62] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Preprint at arXiv:1412.6980*, 2014.