Leveraging Large Language Models for Molecular Generation in OLED Materials Discovery

Wei Xu, Han Chen*, Xinshi Lin*, Ruifeng He**, Jingyao Song**, Lan Ma**
*TCL AI Lab, Hong Kong, China
**Guangzhou China Ray Optoelectronic Materials Co. Ltd., Guangzhou, China

Abstract

Molecular generation is a crucial step in the discovery of OLED materials. Generative AI, particularly LLMs has shown remarkable capabilities in text-based generative applications. We fine-tuned an LLM to perform two molecular generation tasks. The generated molecules were verified to have the desired properties and closely align with chemical principles. This work demonstrates an effective language modeling for molecular generation in OLED materials discovery.

Author Keywords

AI; OLED; Molecular Generation; Large Language Model.

1. Introduction

OLED technology is revolutionizing the display industry with its advantages in color contrast, response time, energy efficiency, flexibility, etc. [1]. A typical OLED device has multiple layers of organic semiconductor materials with complex molecular structures. To identify high performance OLED materials in enormous organic chemical space is extremely challenging.

A practical approach is to perform virtual high-throughput screening based on quantitative structure-property relationship (QSPR) modeling. Recent studies show the great promise of AI-based QSPR modeling for virtual high-throughput screening and analysis [2, 3, 4, 5]. However, designing candidate molecules for screening highly relies on chemical expertise, which has its limitations. Generative AI models [6, 7] have shown remarkable capabilities in learning data distributions and generating new samples that align with desired distributions. Recent advancements in Transformer-based language models [8, 9, 10, 11] have sparked interest in their application to molecular generation for OLED materials [12]. Nonetheless, there is still room for improvement of the model performance in practical design scenarios.

In this work, we adopted a text-based representation of OLED materials and framed the molecular generation problem as a text-to-text generation task by Large Language Models (LLMs). To construct a large and diverse dataset for fine-tuning LLMs, we conducted a massive combinatorial enumeration based on a specific molecular core and commonly used functional groups for our target materials. Molecular properties were labeled using our molecular property prediction models. We fine-tuned an LLM to address two generative tasks: molecular generation with desired properties and molecular optimization towards desirable properties. The generated molecules exhibited a high level of novelty, with molecular properties that met the desired properties and aligned with chemical principles. This work showcases a practical solution by utilizing LLMs for molecular generation in the discovery of OLED materials.

2. Large Language Models for Molecular Generation

OLED materials are organic compounds with complex molecular structures. String representations, such as the Simplified Molecular-Input Line-Entry System (SMILES) [13] is widely used for describing molecular structures and facilitating molecular data interoperability. In this work, we treated the SMILES representation as a chemical language and utilized LLMs to model molecular generation tasks in the discovery of OLED materials.

Large Language Models

LLMs are language models trained on large datasets using deep learning techniques to capture complex contextual relationships. LLMs have demonstrated impressive capabilities in natural language generation and understanding. We expect that LLMs' general capabilities of natural language can be transferred to chemical language for understanding molecular structure and its underlying design principles. The main architectures of LLMs for text generation tasks include decoder-only Transformers such as GPTs [10, 11] and encoder-decoder Transformers such as T5 models [14, 15], as illustrated in Figure 1.

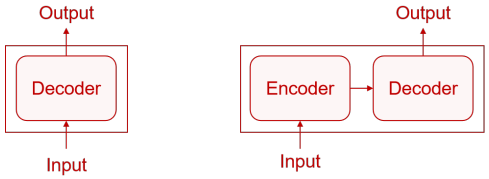

Figure 1. Sketches of decoder-only (left) and encoder-decoder (right) Transformer architectures.

The autoregressive nature of GPTs makes them highly effective at generating locally coherent and relevant text. T5 utilizes an encoder-decoder architecture, where the self-attention mechanism of the encoder part and the cross-attention mechanism of the encoder-decoder part can attend to all input tokens. In principle, T5 is better suited to capturing the characteristics of the SMILES representation, especially the long-range dependencies between SMILES characters. We believe it is advantageous to ensure the validity of lengthy SMILES strings and generate chemically valid molecules.

Molecular Generation as Text-to-Text Task

The greatest advantage of text-based language modeling is its universality. LLMs offer a general framework for addressing various single/multiple objective molecular generation problems using a consistent text-to-text generation model. Generating molecules with desired properties and optimizing molecules towards desirable properties are two common molecular generation tasks. Traditionally, chemists modify a reference

ISSN 0097-996X/25/5602-1201-$1.00 © 2025

molecule and perform experiments or simulations to verify the effects of these modifications. This trial-and-error approach is inefficient and costly. By adopting text-based molecular representations, these molecular generation tasks can be framed as text-to-text generation tasks, taking a text-represented reference molecule with task-related qualifiers as input and generating desirable molecules in text format as output, as illustrated in Figure 2.

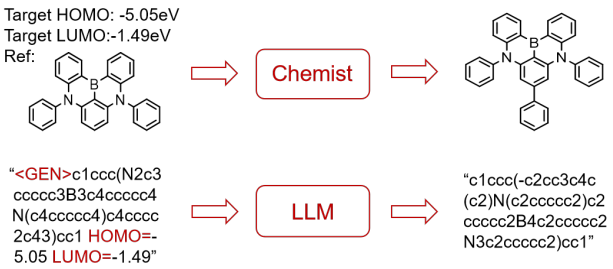

Figure 2. Language modeling for molecular generation.

In practical applications, the generated molecules should retain essential features of the original one, such as the backbone or key functional groups. This constraint can be quantified using molecular similarity scores. Tanimoto similarity is a widely used similarity measure for molecules based on molecular fingerprints [16, 17]. The Tanimoto similarity score ranges from 0 to 1. In our context, a higher Tanimoto similarity score indicates a higher degree of similarity between two molecules.

3. A Practical Application

Blue Dopants Optimization

The match of frontier molecular orbitals, i.e., the Highest Occupied Molecular Orbital (HOMO) and the Lowest Unoccupied Molecular Orbital (LUMO), is vital for balancing charge transport and achieving high operational efficiency.

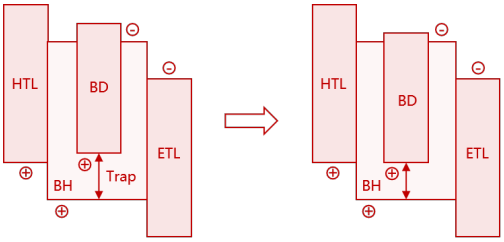

Figure 3. Optimizing energy level for charge transport.

We are using Multi-Resonance Thermally Activated Delayed Fluorescence (MR-TADF) materials as our blue dopants. A challenge we face is that, in comparison to host materials, the relatively high-lying HOMO level of our current dopants traps holes and leads to imbalanced charge transport during operation. One common strategy is to lower both the HOMO and LUMO levels of dopants through molecular optimization while ensuring that the first singlet state energy (S_1) remains unchanged, as illustrated in Figure 3. We reformulated this molecular optimization problem as a text-to-text generation task, where we input text-represented molecules with task-related qualifiers and generate text-represented molecules as output.

Data Preparation

We are focusing on a specific category of MR-TADF materials, as depicted in Figure 4, where Ar_1, Ar_2, Ar_3, Ar_4, and Ar_5 represent functional groups. We conducted combinatorial enumeration based on this structural formula together with commonly used functional groups based on our design preferences. Specifically, we selected 6 functional groups for Ar_1 and Ar_2 each, 23 functional groups for Ar_3 and Ar_4 each, and 29 functional groups for Ar_5. Each functional group offers various fusion/connection positions. In total, we generated more than 200,000 molecules and randomly sampled about 30,000 molecules to construct our fine-tuning dataset.

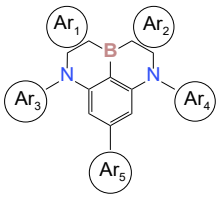

Figure 4. Structural formula of an MR-TADF material.

We labeled all the molecules with HOMO, LUMO, S_1, and T_1 energies using our molecular property prediction models [2, 5]. To empower LLM to capture the intrinsic QSPR, we augmented the dataset using randomly sampled 20% molecules with HOMO, LUMO, S_1, and T_1 energies from a subset of PubChem dataset [18], and 20% molecules with HOMO-LUMO energy gap from PCQM4Mv2 dataset [19]. We constructed the fine-tuning dataset in the format of "<PROP>INPUT OUTPUT", where "PROP" is the property tag, such as "HOMO", "LUMO", "S1", "T1", and "WEIGHT" (molecular weight). "INPUT" is the SMILES representation of the molecule, and "OUTPUT" is the property value. These property-labeled data formed the foundational component of the dataset for teaching LLMs to understand the underlying QSPR contextually.

The other component of the dataset was molecular pair data used to guide LLMs in capturing the structural differences between paired molecules and understanding their effects in a contextual manner. To focus on shifting HOMO and LUMO levels while maintaining S_1 energy unchanged, we only selected molecules with S_1 energies fluctuating within a small interval, such as ranging from 2.6eV to 2.8eV. We randomly sampled 5000 molecules to construct all the pairs and screened them using Tanimoto similarity of their Morgan fingerprint with radius of 3, with a similarity score greater than 0.8. The process yielded around 8,000 molecular pairs with high degree of similarity. These molecular pairs were converted into the format of "<EDIT-PROP-QUALIFIER>INPUT OUTPUT", where "PROP" denotes the property considered a constraint in generation, such as "HOMO" and "LUMO". The "QUALIFIER" indicates the target of generation, using terms such as "LARGE" and "SMALL". For example, <EDIT-HOMO-SMALL> indicates that the model should generate molecules with a smaller HOMO energy value, while <EDIT-HOMO-LARGE> indicates generating molecules with a large HOMO energy value. "INPUT" is the SMILES representation of the reference molecule, and "OUTPUT" is the paired molecule

with the desirable property. We repurposed the molecular pair data to enhance the capabilities of LLMs in generating molecules with desired properties. We restructured the molecular pairs in the format of "<GEN>REFERENCE HOMO=X1 LUMO=X2 S1=X3 T1=X4 WEIGHT=X5 OUTPUT", where "<GEN>" indicates the task, "REFERENCE" is the SMILES-represented reference molecule, "X1/2/3/4/5" is the property value, and "OUTPUT" represents the molecule with those properties. Overall, we constructed 392,176 texts for fine-tuning LLMs.

Model Fine-Tuning and Performance

Since SMILES representation is a sequence of characters, we utilized ByT5 [14], a byte-level T5 model pre-trained directly on raw characters from natural language datasets. We conducted full parameters fine-tuning of ByT5-Small (300 million parameters) for 10 epochs with a batch size of 192 and a learning rate of 0.001. The fine-tuning of 10 epochs took around 5 hours using 6 NVIDIA GeForce RTX 3090 GPUs.

In our model evaluation, we defined "validity" as the percentage of generated molecules that are chemically valid, "novelty" as the percentage of valid molecules not present in the fine-tuning dataset, and "uniqueness" as the percentage of unique molecules. We sampled 1000 molecules to assess these metrics. For different tasks, we assessed the accuracy of generation using different metrics. For the generating molecules with desired properties, we evaluated Mean Absolute Error (MSE) and Root Mean Squared Error (RMSE), while for the generating molecules towards desirable properties, we defined "Top-*K* accuracy", which is the percentage of samples for which the desired molecule is among the top *K* generated molecules with the highest confidence.

Table 1. Performance metrics of two generation tasks

Task	Validity	Unique-ness	Novelty	Accuracy
Targeting HOMO	0.708	0.998	0.946	MAE: 0.181eV RMSE: 0.224eV
Lowering HOMO	0.790	0.949	0.954	Top-10: 0.920
Raising HOMO	0.797	0.952	0.952	Top-10: 0.940

As shown in Table 1, ByT5's general language ability leads to a high validity score for the generated SMILES strings. The high uniqueness scores show its effectiveness for generating distinct molecules. More importantly, the high novelty scores indicate ByT5's capability to generate new molecules beyond our pre-defined molecular dataset. For the task of targeting HOMO, the low MAE and RMSE values indicate that the generated molecules closely match the desired properties. For the task of lowering and raising HOMO, there is almost certainly one molecule among the first 10 generated satisfies the set targets.

We employed the fine-tuned ByT5 as a molecular optimizer to optimized our MR-TADF dopants by lowering HOMO and LUMO levels while maintaining the S_1 energy unchanged. Figure 5 shows two generated molecules that represent two optimization strategies. In Figure 5, both strategies retain the core structure of the reference molecule while incorporating electron-deficient fragments. These modifications typically lead to lower HOMO and LUMO levels, which align with our chemical experience.

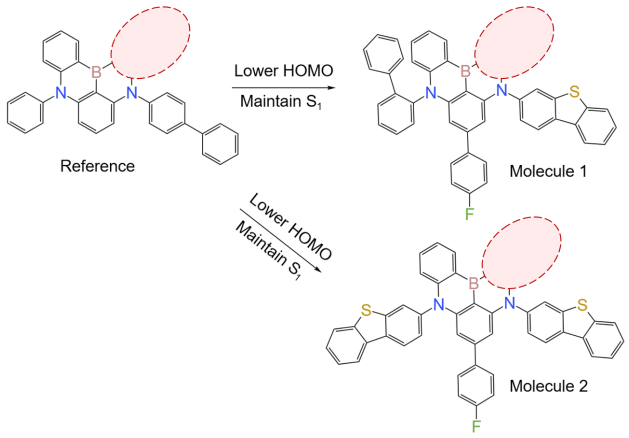

Figure 5. Molecular optimization strategies proposed by ByT5.

We verified the property shift through Density Functional Theory (DFT) calculation and AI prediction [2, 5], as listed in Table 2. Both HOMO and LUMO levels of the reference molecule are lowered effectively by over 0.1eV, while S_1 energy remains unchanged.

Table 2. Electronic properties by DFT calculation and AI prediction

Mol.	HOMO/eV		LUMO/eV		S_1/eV	
	DFT	AI	DFT	AI	DFT	AI
Ref.	-4.972	-4.937	-1.501	-1.540	3.416	3.382
Mol.1	-5.084	-5.030	-1.745	-1.734	3.388	3.391
Mol.2	-5.088	-5.047	-1.739	-1.744	3.381	3.388

Overall, most of the proposed strategies align with chemical principles. More importantly, we noted some optimization strategies which are beyond our chemical experience.

4. Conclusion

In this work, we adopted SMILES representation for OLED materials and formulated two generative tasks: molecular generation with desired properties and molecular optimization towards desirable properties using ByT5, a byte-level text-to-text generation LLM. To construct large and diverse dataset for fine-tuning ByT5, we performed a massive combinatorial enumeration based on a specific molecular backbone and commonly used functional groups for our MR-TADF dopants. The molecular structures were labeled using our property prediction AI models.

The fine-tuning dataset comprised two types of data. One was molecular structures with labeled properties, which was used to teach ByT5 to understand intrinsic QSPR. The other one was molecular pairs consisted of similar molecules with difference in

HOMO and LUMO values, which was used to guide ByT5 to understand structural differences and their effects between paired molecules. The fine-tuned ByT5 model demonstrated excellent ability in generating chemically novel and unique molecules. The top generated molecules were verified to satisfy design requirements and align with chemical principles. Notably, some novel optimization strategies emerged in the generated molecules. This work highlights the remarkable capabilities of LLMs for molecular generation in the discovery of OLED materials.

5. References

[1] Hong G, Gan X, Leonhardt C, Zhang Z, Seibert J, Busch JM, Bräse S. A brief history of OLEDs-emitter development and industry milestones. Advanced Materials. 2021 Mar;33(9):2005630.

[2] Xu W, Shen J, Chen H, He R, Song X, Xia Z, Ma L, Song J. Graph-based AI workflow for OLED materials discovery. SID Symposium Digest of Technical Papers. 2023 Jun (Vol. 54, No. 1, pp. 1571-1574).

[3] Kim H, Kim S, Yoo D, Kim G, Koh E, Kim J, Park S, Kim S, Shin H, Cho H, Baek S. A novel OLED material discovery based on AI technology. SID Symposium Digest of Technical Papers 2024 Jun (Vol. 55, No. 1, pp. 1176-1178).

[4] Kim HJ, Lee J, Choi YK, Lee T, Yang JH, Ko SM, Jeong DW, Han S, Min J, Baek JH, Lee SW. Machine learning strategy towards inverse design of blue TADF emitter: training excited state properties based on density functional theory calculations. SID Symposium Digest of Technical Papers 2024 Jun (Vol. 55, No. 1, pp. 1183-1186).

[5] Xu W, Chen H, Xia Z, He R, Ma L, Song J. AI-enabled high-throughput analysis for OLED materials optimization. Proceedings of the International Display Workshops. 2024 Dec (Vol. 31, pp. 635-638).

[6] Kingma DP, Welling M. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114. 2013 Dec 20.

[7] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. Advances in Neural Information Processing Systems. 2014;27.

[8] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. Advances in Neural Information Processing Systems. 2017;30.

[9] Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. 2018 Oct 11.

[10] Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training.

[11] Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. OpenAI blog. 2019 Feb 24;1(8):9.

[12] Xu W, Chen H, He R, Song X, Ma L, Song J. Exploring potential of language models in OLED materials discovery. SID Symposium Digest of Technical Papers 2024 Jun (Vol. 55, No. 1, pp. 2163-2166).

[13] Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. Journal of Chemical Information and Computer Sciences. 1988 Feb 1;28(1):31-6.

[14] Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou Y, Li W, Liu PJ. Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research. 2020;21(140):1-67.

[15] Xue L, Barua A, Constant N, Al-Rfou R, Narang S, Kale M, Roberts A, Raffel C. ByT5: Towards a token-free future with pre-trained byte-to-byte models. Transactions of the Association for Computational Linguistics. 2022 Mar 25;10:291-306.

[16] Morgan HL. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. Journal of chemical documentation. 1965 May 1;5(2):107-13.

[17] Bajusz D, Rácz A, Héberger K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?. Journal of cheminformatics. 2015 Dec;7:1-3.

[18] Kim JH, Kim H, Kim WY. Effect of molecular representation on deep learning performance for prediction of molecular electronic properties. Bulletin of the Korean Chemical Society. 2022 May;43(5):645-9.

[19] Nakata M, Shimazaki T. PubChemQC project: a large-scale first-principles electronic structure database for data-driven chemistry. Journal of Chemical Information and Modeling. 2017 Jun 26;57(6):1300-8.