

Research Article

Open Access



Data-driven OLED candidate design: a generative model from independent-property domains to the comprehensive performance enhancement

Xinxin Niu¹, Zhiyao Su¹, Luyu Wang¹, Wenbin Shi¹, Hengyue Zhang¹, Yanfeng Dang^{1,*}, Yuan Yuan^{1,*}, Yajing Sun^{1,*}, Wenping Hu^{1,2}

¹Key Laboratory of Organic Integrated Circuits, Ministry of Education and Tianjin Key Laboratory of Molecular Optoelectronic Sciences, Department of Chemistry, School of Science, Tianjin University, Tianjin 300072, China.

²Joint School of National University of Singapore and Tianjin University, Fuzhou 350207, Fujian, China.

***Correspondence to:** Dr. Yanfeng Dang, Dr. Yuan Yuan, Dr. Yajing Sun, Key Laboratory of Organic Integrated Circuits, Ministry of Education and Tianjin Key Laboratory of Molecular Optoelectronic Sciences, Department of Chemistry, School of Science, Tianjin University, No. 92 Weijin Road, Nankai District, Tianjin 300072, China. E-mail: yanfeng.dang@tju.edu.cn; yyuan@tju.edu.cn; syj19@tju.edu.cn

How to cite this article: Niu, X.; Su, Z.; Wang, L.; Shi, W.; Zhang, H.; Dang, Y.; Yuan, Y.; Sun, Y.; Hu, W. Data-driven OLED candidate design: a generative model from independent-property domains to the comprehensive performance enhancement. *J. Mater. Inf.* **2025**, *5*, 45. <https://dx.doi.org/10.20517/jmi.2025.22>

Received: 2 Apr 2025 **First Decision:** 23 May 2025 **Revised:** 7 Jun 2025 **Accepted:** 12 Jun 2025 **Published:** 23 Jul 2025

Academic Editors: Ming Hu, Bohayra Mortazavi **Copy Editor:** Pei-Yun Wang **Production Editor:** Pei-Yun Wang

Abstract

The discovery of high-performance organic light-emitting diode (OLED) materials is hindered by conventional human-aware design methodologies and the scarcity of pure organic luminescent scaffolds. Although machine learning models have improved the efficiency of high-throughput screening for OLED candidates, their effectiveness is still limited by the small size and low quality of available experimental datasets. In this study, we introduced LumiGen, an integrated framework for the *de novo* design of high-quality OLED candidate molecules with targeted photophysical properties. A sampling-screening iterative process was designed to gradually refine the molecular selection, enabling the transition from independent-property optimization to all-rounded OLED candidates. Among the collected high-quality OLED candidate molecules, computational estimates indicate that the optical properties of most molecules (approximately 80.2%) meet the required criteria. During the iterative training process, the Sampling Augmentor enhances the proportion of OLED candidate molecules by over threefold (from 6.56% to 21.13%). Additionally, we successfully synthesized a new molecular scaffold from the OLED candidates, achieving a photoluminescence quantum yield of up to 88.6%. According to the statistics, only 0.33% of the molecules in the dataset outperform our synthesized molecules in terms of overall optical performance.



© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, sharing, adaptation, distribution and reproduction in any medium or format, for any purpose, even commercially, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.



LumiGen demonstrates the ability to learn molecular distribution patterns from disjoint labeled datasets, enabling the direct generation of all-round OLED candidates, thereby advancing OLED material discovery.

Keywords: Machine learning, luminescent molecules, *de novo* design, OLED

INTRODUCTION

Chromophores absorb light at specific wavelengths, triggering electronic transitions between molecular energy states, and subsequently emit light, making them suitable for applications such as organic light-emitting diodes (OLEDs)^[1-5]. In comparison with liquid crystal display (LCD) and traditional light-emitting diode (LED) technology, the self-emitting properties of OLED display technology provide outstanding image quality and energy efficiency, thereby making it the preferred choice for high-end display solutions^[6-10]. Nevertheless, intrinsic constraints associated with fluorescent (first-generation), phosphorescent (second-generation), and thermally activated delayed fluorescence (TADF, third-generation) materials restrict their deployment in OLEDs^[11]. For example, anthracene-based OLEDs suffer from relatively low external quantum efficiency (EQE), which generally remains below 10% in most studies due to forbidden triplet transitions and ineffective light output coupling^[12]. Moreover, the photoluminescence quantum yield (PLQY) of pure organic room-temperature phosphorescent materials (RTP) is notably low, often less than 5%, which is attributed to weak spin-orbit coupling^[11]. Due to electron and hole separation, TADF molecules typically exhibit a broad full width at half maximum (FWHM), which compromises color purity in display applications^[11]. The current scarcity of pure organic luminescent skeletons presents a significant challenge in meeting the diverse luminous demands of OLED devices^[13]. Furthermore, the ambiguity of structure-activity relationships represents a significant limitation to traditional molecular design methods, which rely on existing photophysical chemical knowledge. This approach is inherently inefficient, as it is subject to the constraints of experimental or other human-led molecular design^[14].

With advances in high-throughput screening, open material datasets, and machine learning (ML)-driven property predictors, it has become increasingly feasible to screen materials to identify promising candidates. For example, Joung *et al.* trained an ML model using experimental spectral data, successfully predicting three high-quality luminescent molecules designed by scientists^[15]. Similarly, Shi *et al.* developed a PLQY prediction model based on 230 experimental samples of TADF and performed high-throughput screening to computationally identify potential candidates for deep-blue OLED applications^[16]. Sun *et al.* further improved the performance of predictors for maximum absorption and emission wavelengths, FWHM, and PLQY^[17]. Although ML predictors have greatly enhanced screening efficiency, these approaches primarily rely on human-led chemical intuition and predefined molecular fragments, limiting their potential for discovering novel molecular scaffolds^[15]. Furthermore, these methods often fail to generalize beyond training data, as reflected in their unsatisfactory performance on external datasets^[17]. Generative ML models have emerged as a promising alternative, offering the ability to design novel molecular scaffolds from scratch, unrestricted by predefined fragment libraries. Weiss *et al.* utilized a molecular diffusion model trained on 475,000 computational molecular data points to generate molecules with specific frontier molecular orbital gaps^[18]. Similarly, Zeni *et al.* trained a generative model on over 600,000 materials to design novel materials with targeted physical and chemical properties, such as bulk modulus and magnetic density and further conducted experimental validation to demonstrate the feasibility of the generated materials^[19]. Popular generative models, including variational autoencoders (VAEs), generative adversarial networks (GANs), and diffusion models, have achieved remarkable breakthroughs in materials science^[20-22]. However, these models typically require hundreds of thousands of training samples, making them less practical for data-scarce scenarios. Especially, the experimental luminescent molecules dataset typically

encompasses only a few hundred samples, with the larger luminescent experimental datasets only containing 10^3 entries^[18]. Insufficient data hampers the ability to impose attribute-conditioned constraints in generative models without leveraging large computational datasets, which often contain approximately 10^5 entries^[16,18]. Long short-term memory (LSTM) models excel in distribution learning for small datasets^[23]. However, under conditions of low-fidelity and limited data, enforcing property constraints becomes challenging^[24]. The process of selecting potential OLED candidates remains constrained by the limited number of known materials, experimental optical measurement errors, and the reliance on human expertise^[25]. With the rapid emergence of diverse ML approaches, selecting targeted and efficient methods for OLED candidate screening and enabling an autonomous iterative learning framework are crucial steps toward bridging the gap between research and application, ultimately enabling the successful synthesis and validation of OLED candidates.

In this work, we introduce LumiGen, an integrated framework designed for the *de novo* design of high-quality OLED candidate molecules. To learn molecular distribution patterns from single-property datasets, effectively capture independent property domains, and directly generate molecules with well-rounded performance across multiple properties, LumiGen integrates a Molecular Generator, a Spectral Discriminator, and a Sampling Augmentor. The Molecular Generator explores the chemical space of high-quality independent-property luminescent molecules, effectively mitigating issues related to over-reliance on human intervention and randomness in candidate selection. Given the errors in experimental measurements and the scarcity of experimental data, we have implemented a multi-expert voting strategy and elite selection approach within the Spectral Discriminator. This shift in focus moves from precisely predicting molecular luminescent properties to comprehensively identifying high-quality luminescent molecular sets (MolElite). Time-dependent density functional theory (TD-DFT) calculations of the generated candidate molecules validate that LumiGen achieves an accuracy of around 80.2%. During the operation of the Sampling Augmentor, the generation rate of high-quality (Elite) molecules is enhanced by more than threefold (from 6.56% to 21.13%), while the proportion of lower-quality molecules (Mediocrity) molecules decreases by over fivefold (from 0.775% to 0.131%). Furthermore, we have successfully synthesized a new molecular skeleton from MolElite, whose spectral characteristics fully meet the expectations set by LumiGen. Experimental measurements revealed an FWHM of 49.8 nm, an extinction coefficient of $5.25 \times 10^4 \text{ M}^{-1}\cdot\text{cm}^{-1}$, and a high quantum yield of 88.6% in dichloromethane solution. In the original dataset, only 0.33% of the molecules outperform our synthesized molecules in terms of overall optical performance. LumiGen fills the gap in direct generative models for organic optoelectronic functional molecules, addressing the challenge of imbalanced experimental data, and positioning itself as a powerful tool for the future development of organic optoelectronic materials.

MATERIALS AND METHODS

Data collection and analysis

To harness the capabilities of artificial intelligence (AI) for screening OLED candidate molecules, it is essential to evaluate the available data resources. Several open material databases of luminescent molecules exist, including the DB_{exp} dataset of chromophore experimental data, the ASBase dataset of aggregation-induced emission (AIE) molecules, the FORMAD dataset containing computed excited-state properties, and some smaller experimental TADF datasets^[9,25-29]. In this study, we focus primarily on DB_{exp} and ASBase, as they directly pertain to OLED performance, providing key luminescence properties such as PLQY, Φ_{QY} , and maximum emission wavelength, λ_{emi} , FWHM, σ_{emi} , and extinction coefficient, ϵ_{max} . While FORMAD offers detailed excited-state energy calculations, it may not be as intuitive for our objective compared to experimental datasets^[29]. Additionally, existing experimental TADF datasets are typically limited to a few hundred molecules, are not publicly available, or lack critical optical property data, making them unsuitable for direct use.

DB_{exp} was compiled by Joung *et al.* in 2020 and is the largest experimental luminescent molecular dataset with the most complete luminous properties by now^[25]. The samples in the DB_{exp} dataset are derived from 1,358 papers containing organic luminescent compounds. It includes 19,280 luminescent molecule-solvent pairs, comprising 6,690 unique luminescent molecules and 374 solvents. The samples from the DB_{exp} dataset display a Gaussian-like distribution of λ_{emi} , σ_{emi} , and ε_{max} , as depicted in Figure 1. Nevertheless, in terms of Φ_{QY} , the distribution skews towards the lower value region, which could be ascribed to the fact that before 2010, it was experimentally challenging to obtain organic luminescent molecules with a quantum yield greater than 0.75^[25]. This distribution phenomenon reflects the lengthy trial-and-error process in the experimental development of high-quality luminescent molecules. Moreover, the bias of quantum yield also makes it difficult to train the model for judging photophysical properties.

The limited data scale and low fidelity of experimental data hinder the seamless integration of the generation and identification. Instead of blindly sampling across the entire chemical space, a more effective strategy is to focus sampling within regions containing molecules with superior individual properties, followed by a targeted screening process. This optimization approach could significantly enhance the efficiency of OLED candidate molecule screening. Recognizing this potential, we selected three high-quality subsets of molecules from the DB_{exp} for the exploration of chemical space. Each set consists of 300 molecules, meticulously chosen based on possessing the narrowest σ_{emi} , the highest Φ_{QY} , and the highest ε_{max} .

Detailed construction of LumiGen

In this work, we designed a molecular generation algorithm framework, LumiGen, which could be used for *de novo* luminescent molecular generation with modified photophysical properties. LumiGen is composed of three main parts, which work at once and can form a loop [Figure 2]. Part I is the Molecular Generator for novel luminescent molecule generation. Part II is a Spectral Discriminator evaluated for comprehensive spectral properties. Part III is a Sampling Augmentor, which can enhance the sampling of new features and improve the excellence rate of the molecules generated in the next stage. Through the collaboration of these three components, LumiGen progressively learns molecular distribution patterns from disjoint labeled datasets, enabling the direct generation of all-round OLED candidates.

In the construction of the Molecular Generator, we adopted a pre-training and transfer learning strategy to efficiently generate chemically permissible molecules. We utilize a pre-trained LSTM model to capture the structural and synthetic preference of molecules on the ChEMBL24 dataset, which is enriched with a diverse array of both experimentally synthesized and naturally occurring molecular compounds^[23,30]. During the training process, each molecule was encoded as a one-hot vector. The LSTM model is trained to predict the conditional probability distribution of each token of the encoded simplified molecular input line entry system (SMILES) sequence^[23]. Based on the transfer learning strategy, the LSTM model is adeptly applied to high-quality independent-property luminescent molecular sets. The molecular generator is implemented as a four-layer LSTM network, trained on SMILES strings using categorical cross-entropy loss and the Adam optimizer. During transfer learning, the first layer's parameters are frozen, while the second layer is fine-tuned with a reduced learning rate to adapt to the downstream task in Supplementary Figure 1. By leveraging its learned representations, the LSTM model efficiently samples the target chemical space. This capability enables the LSTM to generate a variety of structurally diverse molecules, effectively bridging gaps within the target chemical space and enhancing our library of potential luminescent candidates.

A Spectral Discriminator is crafted based on a graph convolutional neural network (GCN)^[31]. In the feature extraction stage, the atomic and bonding information of luminescent molecules and solvent molecules are encoded separately, and connected to form a composite feature matrix [Supplementary Figure 2 and

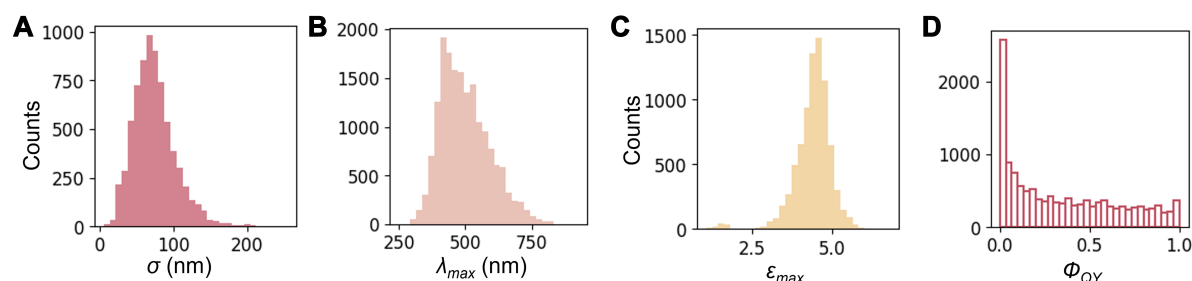


Figure 1. DB_{exp} dataset distributions for (A) λ_{emir} , (B) σ_{emir} , (C) $\lg(\epsilon_{max})$, and (D) Φ_{QY} .

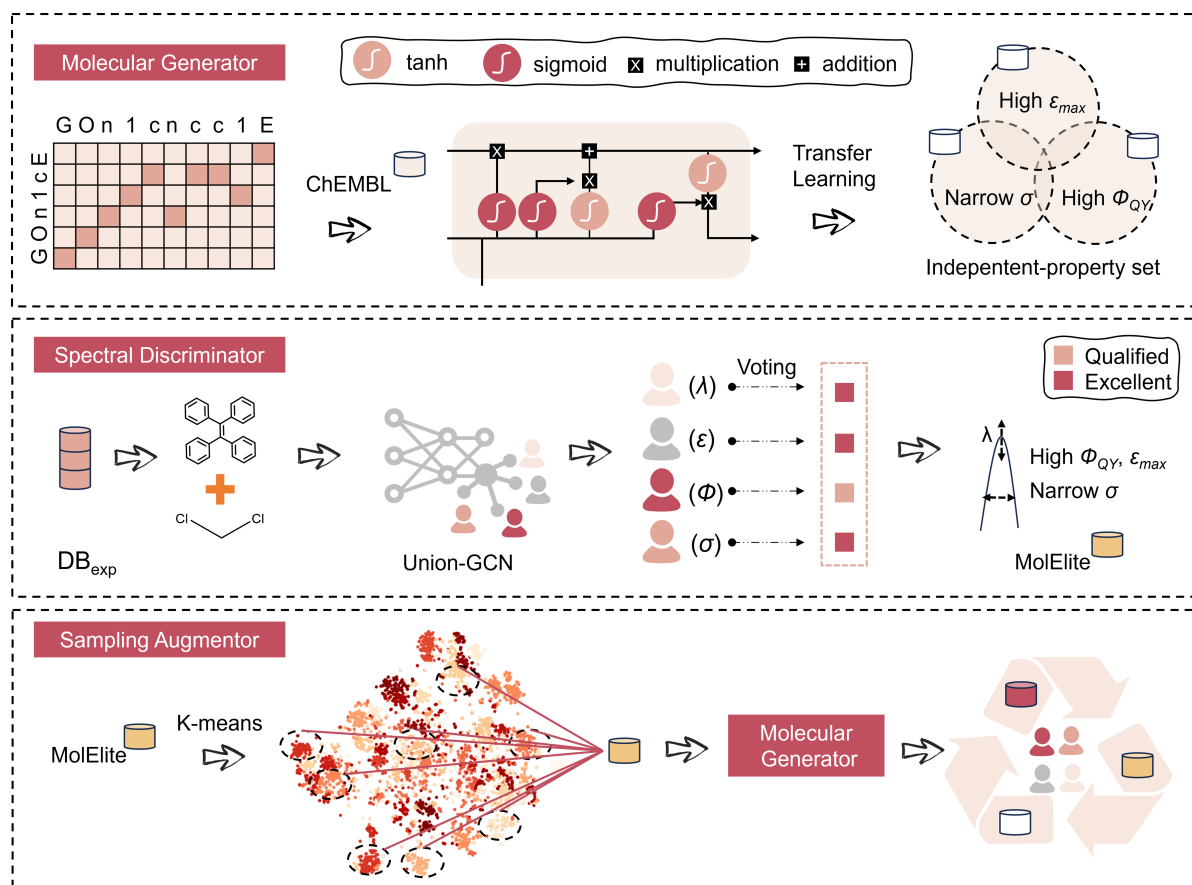


Figure 2. LumiGen framework, comprising a Molecular Generator, a Spectral Discriminator, and a Sampling Augmentor. The Molecular Generator leverages a ChEMBL24 pre-trained LSTM model to produce three types of high-quality luminescent molecules. The Spectral Discriminator is trained on the DB_{exp} dataset. The Sampling Augmentor generates new high-quality subsets through clustering, completing the loop from lab to ML. LSTM: Long short-term memory; ML: machine learning.

Supplementary Table 1]. This feature construction strategy is superior for differentiating luminescent molecules from solvent molecules, thus accurately simulating changes in luminescent properties when molecules are in different solvents. Due to the unevenness of the experimental dataset, where different photophysical properties of molecules may not be reported simultaneously, we have implemented a multi-expert voting strategy to comprehensively evaluate a luminous molecule, known as Union-GCN. Certain experts are dedicated to predicting quantum yield, while some focus on predicting FWHM or maximum emission wavelength. Experts who specialize in the same property are identified as experts within the same

domain. In contrast, those experts tasked with predicting different photophysical properties are recognized as experts from distinct domains. This specialized approach not only tailors the predictive capabilities of each model to specific aspects of luminescent properties but also ensures that insights from diverse domains can be integrated. To mitigate the impact of experimental errors and data insufficiencies, our elite selection strategy integrates the collective decisions of the Union-GCN. To ensure model diversity within the Union-GCN ensemble, we use identical network architectures and hyperparameters across all models, while varying the random seeds for training. This approach introduces diversity through different data permutations and initialization paths, leading to complementary predictions. This strategy shifts the focus from precisely predicting the luminescent properties of molecules to comprehensively identifying high-quality luminescent molecular sets. Specifically, we randomly select 80% of the luminescent molecule samples as a training set and use different random seeds to train Union-GCN. As a result, we obtained multiple experts from the same and different domains. Then, Union-GCN predicts the luminescent properties of sampled molecules in a hypothetical solvent and evaluates the consistency of experts from the same domains. If there is significant divergence among experts within the same domain [e.g., different experts in (λ)], it indicates the difficulty of accurately evaluating the luminescent properties based on the current dataset. A molecule can only advance to the next round of classification if all experts within the same domain agree that it belongs to the same grade in Table 1. Thirdly, Spectral Discriminator channels molecules with three “Excellent” luminescent properties, as determined by experts from three different domains [e.g., (ϵ), (σ), and (λ)], into the high-quality MolElite and those with three “Qualified” properties into the MolMediocrity.

In the Sampling Augmentor, we utilize Morgan fingerprints, a type of circular fingerprint for molecules, as a descriptor to facilitate the clustering of molecules within the MolElite. These fingerprints capture the molecular structure by encoding the presence of specific chemical substructures within a fixed radius around each atom, providing a robust basis for comparing molecular similarities. The entire dataset is divided into 50 distinct groups using the k-means algorithm depending on the distribution and diversity of the molecular structures within the dataset. By dividing the molecules into well-defined groups, the generator is exposed to a wide array of structural motifs. By selecting 300 molecules from different clusters for the next generation, we maintain a diverse and representative high-quality light-emitting molecule subset. This approach fosters a broad exploration of the chemical space, enhancing the likelihood of discovering novel, efficient luminescent materials in each iteration of the cycle, from molecular generation through to spectral identification. The module selection and hyperparameter optimization procedures are illustrated in Supplementary Figures 3 and 4.

RESULTS AND DISCUSSION

Baseline performance evaluation of LumiGen

To comprehensively evaluate the baseline performance of the LumiGen framework, we conducted a systematic analysis of its three key components: Molecular Generator, Spectral Discriminator, and Sampling Augmentor. The Molecular Generator was assessed in terms of chemical space exploration, molecular novelty, and diversity, verifying its capability to reduce human intervention and optimize sampling quality. The performance of the Spectral Discriminator was evaluated by examining its prediction accuracy for photophysical properties in different solvent environments, assessing its ability to capture the interactions between emissive molecules and solvents. Additionally, the iterative optimization process of the Sampling Augmentor was monitored, quantifying changes in the proportion of high-quality molecules to determine its effectiveness in enhancing screening efficiency and refining molecular generation. The following sections provide a detailed performance assessment of each component within LumiGen.

Table 1. Rules for grade assignment

Grade	$\lg(\epsilon)$ ($\text{M}^{-1}\cdot\text{cm}^{-1}$)	Φ_{QY}	σ_{emi} (nm)	λ_{emi} (nm) ^a
Excellent	> 4.5	> 0.7	< 60	$\Delta\lambda < 20$
Qualified	3.5-4.5	0.4-0.7	60-100	20-40
Bad	< 3.5	< 0.4	> 100	$\Delta\lambda > 40$

^a $\Delta\lambda = \lambda_{\text{max}} - \lambda_{\text{min}}$, λ_{max} and λ_{min} represent the maximum and minimum emission wavelengths predicted by Union-GCN, respectively. GCN: Graph convolutional neural network.

The pre-trained LSTM model achieved high-performance metrics with a validity index of 97.4%, a uniqueness index of 97.3%, and a novelty index of 93.8% for the generated SMILES. The high validity indicates that the model can efficiently translate generated molecules back into molecular structure, showcasing its capability to generate chemically meaningful SMILES strings. The considerable proportion of novelty across the dataset highlights the model's ability to explore a broad chemical space, substantiating its efficacy in the *de novo* generation of novel molecules. When the learning chemical space from ChEMBL24 to high-quality luminescent molecule subsets, the validity and novelty of generated data indices slightly decreased to 81.5% and 81.1%, respectively, while the uniqueness increased to 99.8%. These changes can be attributed to the increased complexity of the luminescent molecular structures. Subsequently, the Molecular Generator samples every ten training sessions to monitor the learning process and generate new molecules. A redundancy rate of 63% by the 20th sampling indicates sufficient exploration within the target space of luminescent molecules.

Then, uniform manifold approximation and projection (UMAP) technology and the fraction of sp^3 -hybridized carbon atoms (F_{sp^3}) are employed to visualize the movement of the chemical space center. In Figure 3A, the chemical space of ChEMBL is not only in close proximity to but also partially overlaps with that of the DB_{exp} . Nevertheless, a clear distinction is evident when comparing the target spaces occupied by high-quality luminescent molecule subsets. A significant proportion of the chemical space associated with molecules exhibiting narrow FWHM, high extinction coefficient, and high quantum yield displays discrepancies from the established DB_{exp} , indicating the existence of hitherto unexplored regions. These unknown chemical regions are of critical importance, as they are likely to contain molecules with enhanced luminescent properties that have not yet been realized experimentally.

Otherwise, molecules from ChEMBL exhibit significantly higher F_{sp^3} compared to those in the DB_{exp} in Figure 3B. A reduction in F_{sp^3} typically results in increased molecular rigidity, which can enhance luminescence efficiency by limiting intermolecular rotations and reducing non-radiative transitions. The stability and alignment of F_{sp^3} with target space characteristics during the transfer learning process illustrate the Molecular Generator's adaptability in learning structural features. Furthermore, Shannon entropy (SSE) consistently averaged around 0.9 during the transfer learning process, indicating substantial diversity within sampled molecules^[32]. These findings underline the effectiveness of our transfer learning approach in generating novel and diverse high-quality molecular structures from high-quality subsets, which is also shown in Figure 3C and D.

Moreover, we evaluated the performance of the Spectral Discriminator. Firstly, as Union-GCN constructs and learns from luminescent molecules and solvents as discrete molecular graphs, the model's performance exceeds that of GCNs with a single graph as input by approximately 3.15 nm in predicting maximum emission wavelength in Supplementary Tables 2 and 3. This illustrates the benefit of Union-GCN in simulating interactions between luminescent molecules and solvents, emphasizing its capacity to capture accurately the intricate interaction that influences luminescent properties in diverse solvent environments.

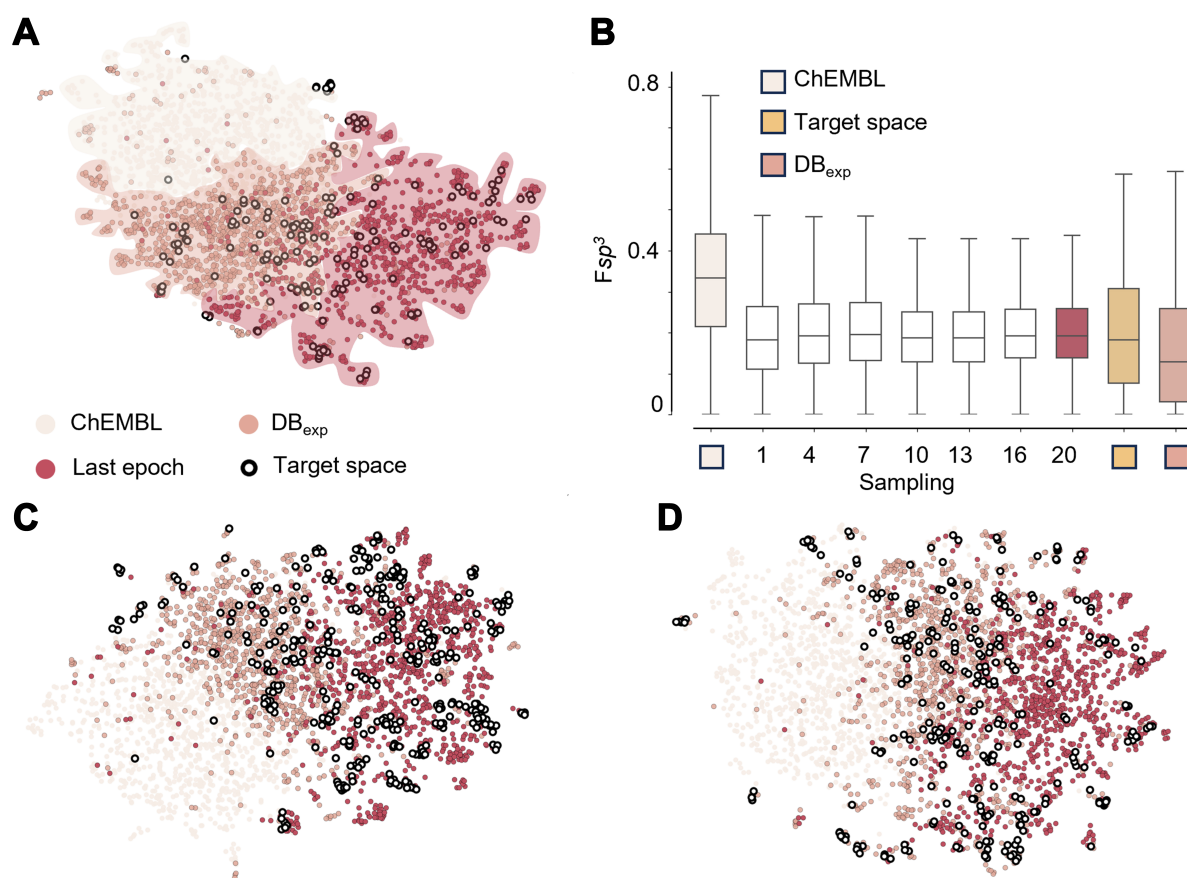


Figure 3. (A) UMAP for target space with narrow σ_{emi} ; (B) The box plots of Fsp^3 variations corresponding to different chemical spaces are shown for narrow σ_{emi} ; (C) UMAP for target space with high Φ_{QY} ; (D) UMAP for target space with high $lg(\epsilon_{max})$. UMAP: Uniform manifold approximation and projection.

Additionally, Union-GCN exhibits enhanced accuracy, with a mean absolute error (MAE) of 21.97 nm and a root mean square error (RMSE) of 30.85 nm in predicting maximum emission wavelength in Figure 4A. These normalized MAE represent only 2.7% of the total maximum emission wavelength range from 247 to 1,050 nm, demonstrating remarkable precision. With regard to other photophysical properties, the relatively higher error observed in quantum yield prediction can primarily be attributed to the susceptibility of quantum yield measurements to significant experimental errors and a weaker correlation with molecular structure, which is evidenced by previous statistical analysis^[26].

In the initial sampling space, the MolElite set comprised 4,766 molecules, representing 6.56% of the total, while the MolMediocrity set accounted for only 0.775%. In the first iterative cycle, the proportion of molecules in MolElite rose to 21.13%, whereas that of the MolMediocrity fell to 0.254%. The increased elite proportion indicates that the enhanced sampling can continuously strengthen the learning process and optimize the generation step of molecular generation. Throughout this process, the novelty, validity, and uniqueness of the sampled molecules remained consistently high as shown in Figure 4B. Although the proportion of high-quality molecules increased, the maintenance of novelty, efficacy, and uniqueness demonstrates that the Molecular Generator can still generate a diverse range of molecular structures while pursuing optimization. As the number of iterations increased, the proportion of molecules in MolElite remained high, while the proportion of molecules classified as MolMediocrity showed a continuous decline.

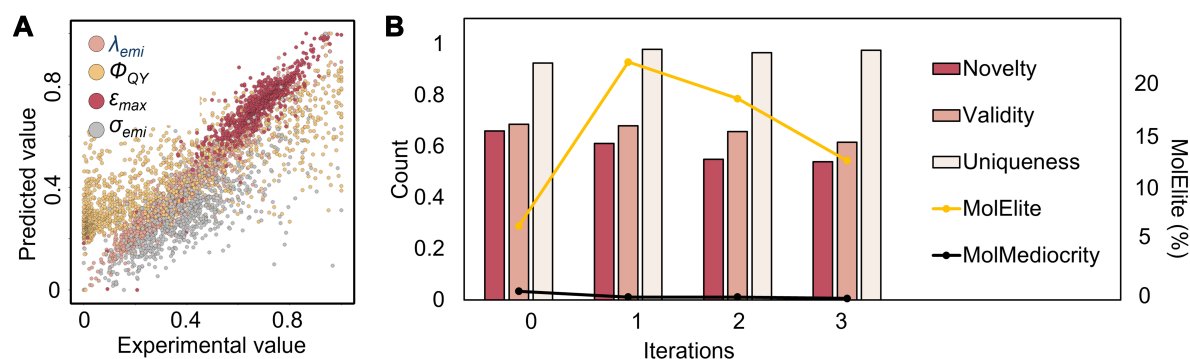


Figure 4. (A) Performance of Union-GCN in λ_{emi} , σ_{emi} , Φ_{QY} , and $\lg(\epsilon_{max})$; (B) The variation trend of novelty, availability, uniqueness, MolElite, and MolMediocrity of the sampled molecules during Sampling Augmentor training. Different properties are normalized to be displayed on the same image. GCN: Graph convolutional neural network.

In the whole process, the proportion of Elite molecules increases more than threefold, rising from 6.56% to 21.13%, while the fraction of Mediocrity molecules declines over fivefold, dropping from 0.775% to 0.131%. This indicates that the Molecular Generator gained a deeper understanding and capturing ability of elite molecule features over time, with the evolving model increasingly favoring the generation of higher-performance molecules.

Analysis of iteratively generated results

The Spectral Discriminator employs a comprehensive evaluation and categorization process to distinguish between high-quality and mediocre molecules. This process is applied to molecules sampled from repeatedly generative processes, resulting in two collections: the high-quality MolElite and the control group, the MolMediocrity. The maximum structural similarity indices between molecules in the MolElite and those in the DB_{exp} primarily range from 0.4 to 0.6, as illustrated in Figure 5A. Generally, a similarity index of less than 0.7 is indicative of significant structural novelty^[28]. That is to say, the molecules in MolElite exhibit a significant degree of structural novelty. As shown in Figure 5B and C, molecules in the MolElite typically exhibit consistent calculated emission wavelengths with the prediction of Spectral Discriminator. In the ultraviolet (UV) absorption spectra, extinction coefficients are all above 4.5. Moreover, the SA score for the target molecules used in transfer learning was approximately 3.5, slightly higher than the average score of 2.5 for the entire DB_{exp} in Figure 5D. Notably, the SA score peak for the MolElite aligns with that of the target molecules, whereas the MolMediocrity aligns more closely with the DB_{exp} . This disparity underscores a correlation between the complexity of molecular structures and their luminescent properties, successfully captured by the Spectral Discriminator. It highlights the complementary strengths of both the Molecular Generator and the Spectral Discriminator. The fluorescence absorption and emission spectra of molecules were plotted in both MolElite and MolMediocrity^[33]. Some molecules in the MolMediocrity suffer from misaligned maximum emission wavelength from TD-DFT calculation. Their extinction coefficient is less than 4.5 in Figure 5E and F. Simultaneously, the emission spectra of some molecules in MolElite exhibit a notably narrow FWHM under the same spectral plotting parameters, and detailed information is in Supplementary Figures 5-8. Among a randomly selected set of 80 Elite molecules, 69 meet the predefined criteria in terms of λ_{emi} and $\lg(\epsilon)$, accounting for 80.2% of the total. Furthermore, in accordance with the Fermi Golden rule, the radiative transition rate (k_r) and internal conversion rate (k_{ic}) of these molecules were calculated^[33]. When k_{ic} is of the same order of magnitude, the k_r of molecules in the MolElite is one to two orders of magnitude higher than those of the control group in Supplementary Table 4. These data illustrate the potential advantages of the MolElite in terms of luminous purity and efficiency. Despite the fluorene derivatives have been proven to have luminescent properties, the

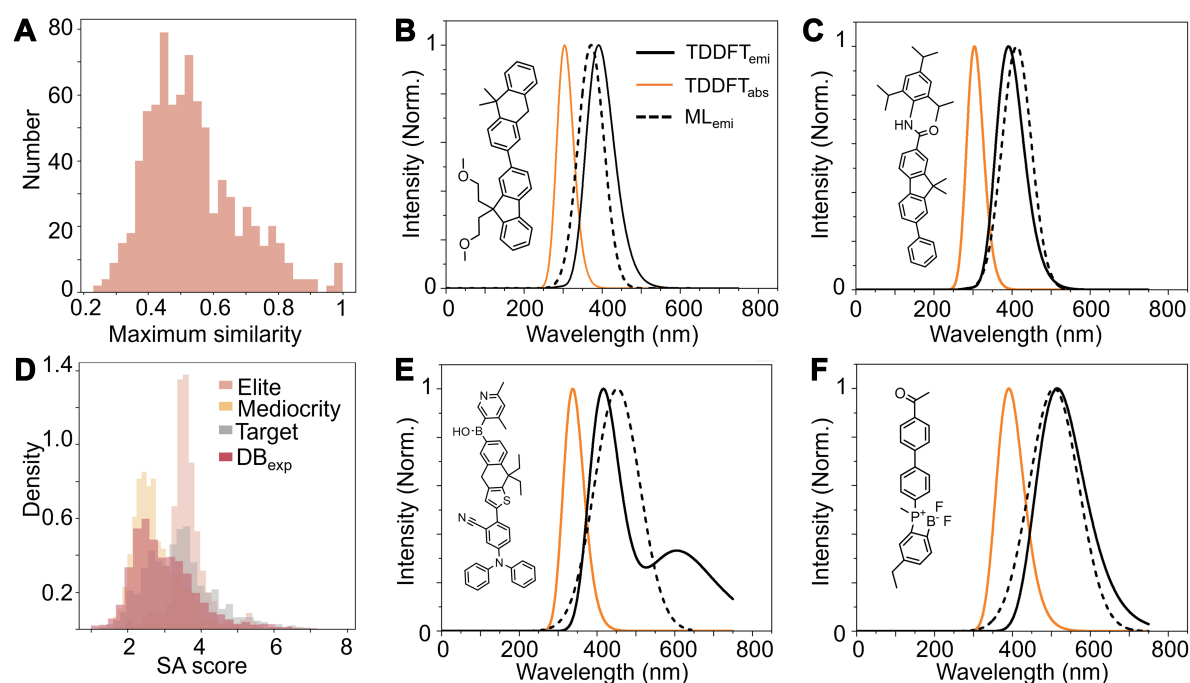


Figure 5. (A) The maximum similarity between molecules in the MolElite and molecules in the DB_{exp}; (B and C) Examples of molecules from the MolElite, along with their fluorescence absorption and emission spectra; (D) SA score distribution histogram of MolElite, MolMediocrity, DB_{exp}, and target molecules; (E and F) Examples of molecules from the MolMediocrity, along with their fluorescence absorption and emission spectra. SA: Synthetic accessibility.

dihydroanthracene-fluorene linkage, and fluorenamide molecular skeletons shown in Figure 5B and C have not yet been synthesized nor has their luminescent performance been explored. TD-DFT calculations required an average of 235 min per molecule to assess both ground and excited state geometries. In contrast, Union-GCN predicted four essential optical properties within just 0.5 min on the same computational setup, showcasing its ability for rapid, large-scale predictive analysis.

Experimental validation and model expansion

To establish a bridge between model generation and experimental characterization, we used the 69 computationally validated molecules from the MolElite series as the initial proof-of-concept set. Subsequently, based on spectral calculation results and synthetic feasibility, human chemists ultimately selected 9,9-dimethyl-2-(6-phenylnaphthalen-2-yl)-9H-fluorene (Mol1) as the molecular scaffold for further study in Supplementary Figures 9–11. With an SA score of 2.0, Mol1 signifies a reduced complexity in synthesis, as shown in Figure 6A. We experimentally obtained Mol1 in a streamlined one-step synthesis process. In the experimental test, the fluorescence absorption and emission characteristics of Mol1 perfectly aligned with our design specifications, as demonstrated in Figure 6B. Notably, the fluorescence emission spectrum revealed an FWHM of 49.8 nm, and an $\lg(\epsilon)$ of 4.72. In the dichloromethane solution, Mol1 exhibited a high quantum yield of 88.6%. In comparison, the LumiGen predicted an FWHM of 53.0 nm, a maximum emission wavelength of 400.3 nm, and an $\lg(\epsilon)$ of 4.61 for Mol1. These luminescent properties are in good agreement with the experimental measurements. The quantum yield evaluated by MolElite was 0.592, slightly lower than the experimental value. By contrast, the TD-DFT calculated maximum emission wavelength is more red-shifted, and the computed extinction coefficient is greater than 5. These exceptional results not only fulfill our rigorous criteria but also validate the practical effectiveness and efficiency of the LumiGen framework.

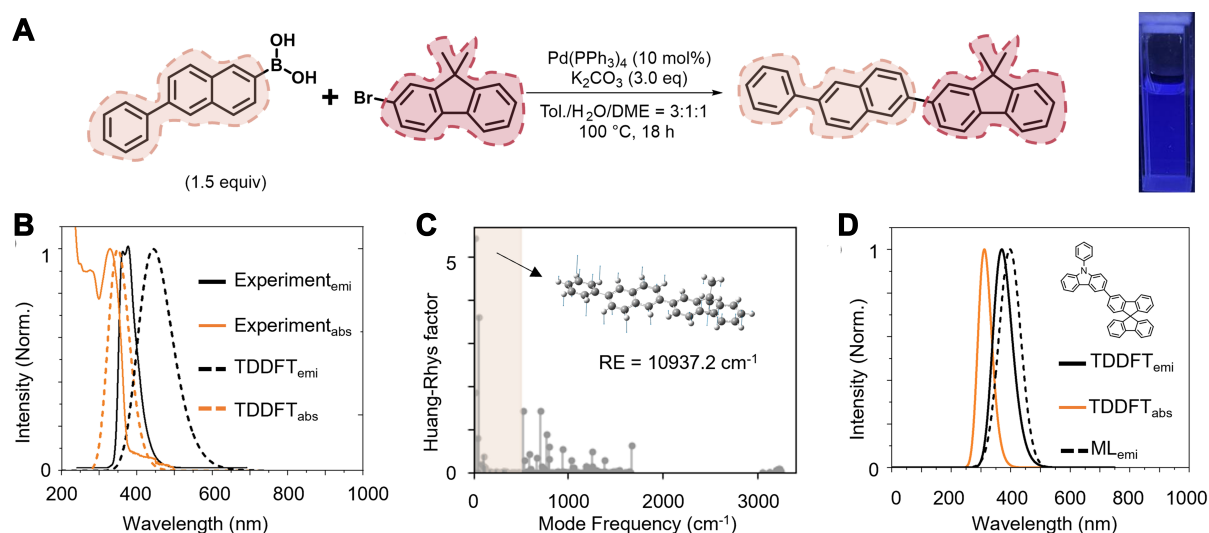
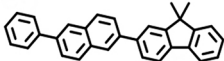
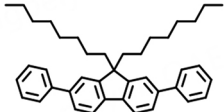
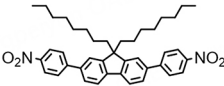


Figure 6. (A) Mol1 is synthesized from MolElite; (B) Fluorescence absorption and emission spectra of Mol1; (C) Huang-Rhys factors of Mol1. RE represents the RE from the S1 state to the S0 state; (D) Examples of molecules from the MolElite, along with their fluorescence absorption and emission spectra.

Furthermore, we conducted a statistical analysis of the number of molecules in the DB_{exp} dataset that outperformed Mol1 in terms of PLQY, $\lg(\epsilon)$, and FWHM. Specifically, we identified molecules that simultaneously satisfy the conditions: PLQY greater than 0.886, $\lg(\epsilon)$ greater than 4.72, and FWHM less than 49.8 nm. The analysis reveals that only 23 molecules (0.33%) in the DB_{exp} dataset outperform Mol1 in terms of the combined performance across PLQY, $\lg(\epsilon)$, and FWHM. In addition, we compared several previously synthesized molecules with Mol1 in terms of their luminescent properties, as summarized in Table 2^[34,35]. The structural similarities between these molecules and Mol1 are all above 0.7, indicating that they belong to the fluorene family of molecular scaffolds. Interestingly, the $\lg(\epsilon)$ of the three comparison molecules are in close proximity to that of Mol1, all greater than 4.5, demonstrating that these molecules exhibit strong absorption characteristics similar to Mol1. In terms of PLQY and FWHM, Mol2 exhibits properties that are remarkably close to those of Mol1. The PLQY of Mol2 is slightly lower than Mol1 (85.0% vs. 88.6%), while the FWHM is similar, reinforcing its potential for efficient luminescent applications. In contrast, the performance of Mol3 highlights the limitations of molecules that do not meet the ideal structure-property relationships. Mol3, with a similarity score of 0.74, shows promising structural alignment with Mol1. However, its PLQY is significantly reduced to just 11.0%, and its FWHM is extremely broad at 119.9 nm, which deviates substantially from our design criteria.

First-principles calculations can partly explain the differences in photoluminescent properties observed within the same molecular family. Mol1 exhibits a high highest-occupied molecular orbital-lowest-unoccupied molecular orbital (HOMO-LUMO) overlap and strong oscillator strength ($f = 2.0051$) in its lowest excited state, both of which could contribute to an enhanced radiative transition rate, thereby potentially achieving a high PLQY. Furthermore, as shown in Figure 6C, Mol1 displays a series of normal modes with high Huang-Rhys factors in the 0-500 cm^{-1} vibrational frequency range, with normal mode 8 having a Huang-Rhys factor of 5.07, primarily associated with the overall antisymmetric torsional vibration of the molecule. In contrast, Mol2 and Mol3 exhibit lower Huang-Rhys factors in the same frequency range, and Mol1 also has a slightly smaller reorganization energy (RE) than Mol2. These factors collectively indicate that Mol1 has a potential advantage in PLQY. On the other hand, Mol1's vibrational modes are mainly concentrated in the low-frequency region near 0 cm^{-1} , where they have little impact on spectral

Table 2. Comparison of luminescent properties of Mol1 with other synthesized similar molecules

Molecule	λ_{emi} (nm) ^a	Φ_{qy}	σ_{emi} (nm)	$f(\text{S1} \rightarrow \text{S0})$	SA score
 (Mol1)	400	88.6%	49.8	2.00	2.0
 (Mol2)	361	85.0%	41.1	1.87	2.3
 (Mol3)	538	11.0%	119.9	1.66	2.7

SA: Synthetic accessibility; GCN: graph convolutional neural network.

broadening. Additionally, the vibrational coupling in its high-frequency region is significantly suppressed, resulting in a relatively narrow spectral bandwidth, which further contributes to an advantage in FWHM. In conclusion, Mol1 demonstrates computational advantages in both PLQY and FWHM, providing theoretical support for its exceptional photoluminescent performance within the same molecular family.

These properties suggest that Mol3, despite its structural similarity, does not meet the performance standards expected of high-efficiency luminescent molecules. The large FWHM and low quantum yield indicate substantial non-radiative losses, which is a critical disadvantage for any potential application in optoelectronic devices. In contrast, Mol1 demonstrates versatility in the optical performance comparison among these similar molecules, excelling in multiple luminescent indicators. Mol1 not only exhibits a lower SA score of 2.0, indicating better synthetic accessibility, but also demonstrates superior luminescent performance in terms of PLQY and FWHM. Mol1, as a proof of concept, confirms that LumiGen successfully achieves its goal of transitioning from independent-property domains to comprehensive performance enhancement. Its ability to combine superior photophysical properties illustrates the effectiveness of the approach and further establishes LumiGen as a powerful framework for future luminescent molecule design.

To verify the feasibility in the case of insufficient material data, we applied LumiGen to train a generation model in a smaller molecular dataset of AIE materials. ASBase currently contains over 1,000 AIE functional molecular materials, encompassing the photophysical and physicochemical properties in [Supplementary Table 5](#)^[26]. The Molecular Generator adeptly filled the gaps between the ASbase and the target spaces in [Supplementary Figures 12-14](#). Compared to the DB_{exp}, the UMAP dimensionality reduction plot using high-quality AIE molecules for transfer learning shows more clustering of the sampled molecules, probably due to the increased rigidity of AIE molecules. A notable finding was the particularly low Fsp^3 on templates with high quantum yield, suggesting targeted enhancements in molecular design during the transfer learning process. Our comparative analysis, supported by TD-DFT calculations, revealed superior luminescent properties in the MolElite, characterized by narrower FWHM, higher k_r , and higher extinction coefficient, while the MolMediocrity showed less favorable characteristics in [Supplementary Figures 7 and 8](#). Importantly, for the molecules shown in [Figure 6D](#) and their molecular scaffold derivatives, an EQE of nearly 25% was achieved for the sky-blue phosphorescent OLED^[36]. The adaptability for smaller datasets and accommodating experimental variability of LumiGen highlights its practicality for real-world applications. The emergence of sophisticated data extraction tools and advanced language models (such as ChatGPT)

promises to enrich the training datasets available for LumiGen, potentially promoting the development and testing of luminescent materials.

CONCLUSIONS

In this work, we introduced LumiGen, a novel framework combining a Molecular Generator, a Spectral Discriminator, and a Sampling Augmentor to achieve the *de novo* design of luminescent molecular generation. This tool is particularly adept at populating and filtering the chemical space of high-quality luminescent molecules, and the Sampling Augmentor is capable of optimizing the model to generate higher-quality luminescent molecules. In particular, the multi-expert voting and elite selection strategy effectively address substantial statistical errors in experimental data, thereby ensuring that only high-quality luminescent molecules are retained. The generated molecules display minimal structural resemblance to the original dataset and exhibit high SSE, demonstrating their novelty and diversity. The validity of LumiGen in distinguishing between MolElite and MolMediocrity is also confirmed by TD-DFT calculations, which are conducted with the aim of tailoring targeted luminescent molecules for advanced optoelectronic applications. By synthesizing and characterizing high-quality molecular scaffolds, LumiGen can seamlessly integrate theoretical predictions with experimental validations. What is more, LumiGen is effective in limited datasets ASBase, which makes it advantageous in resource-constrained situations. With the emergence of batch literature data extraction tools and large language models, we are set to enhance our training datasets, boosting the learning potential of our framework. As the pioneering ML framework for *de novo* luminescent molecular design, LumiGen strategically bridges gaps in high-quality experimental data to discover versatile luminescent molecules, poised to transform the search for next-generation display technologies.

DECLARATIONS

Acknowledgments

The authors gratefully acknowledge the National Supercomputer Center in Tianjin (Tianhe 3F) and the Scientific Computing Center of CIC, Tianjin University for providing computation facilities.

Authors' contributions

Data curation, methodology, writing - original draft: Niu, X.

Data curation, formal analysis: Su, Z.; Zhang, H.

Data curation: Wang, L.; Shi, W.

Writing - review and editing: Dang, Y.

Data curation, writing - original draft: Yuan, Y.

Funding acquisition, project administration, supervision, writing - review and editing: Sun, Y.

Funding acquisition, writing - review and editing: Hu, W.

Availability of data and materials

The code is open source at <https://github.com/YajingSun-Group/LumiGen>. The relevant datasets, model architecture, and generated data can be found in the links.

Financial support and sponsorship

This work was financially supported by the National Natural Science Foundation of China (22473085, 22003046 and 52121002), the Ministry of Science and Technology of China (2022YFA1204401) and Xiaomi Young Talents Program.

Conflicts of interest

All authors declared that there are no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Copyright

© The Author(s) 2025.

REFERENCES

1. Xiong, Z.; Xiang, S.; Lv, Y.; Chen, B.; Zhang, Z. Hydrogen-bonded organic frameworks as an appealing platform for luminescent sensing. *Adv. Funct. Mater.* **2024**, *34*, 2403635. [DOI](#)
2. Yang, Q.; Hu, Z.; Zhu, S.; et al. Donor engineering for NIR-II molecular fluorophores with enhanced fluorescent performance. *J. Am. Chem. Soc.* **2018**, *140*, 1715-24. [DOI](#)
3. Hayashi, S.; Koizumi, T. Elastic organic crystals of a fluorescent π -conjugated molecule. *Angew. Chem. Int. Ed. Engl.* **2016**, *55*, 2701-4. [DOI](#) [PubMed](#)
4. Wang, C.; Adams, S. R.; Ahrens, E. T. Emergent fluorophores and their uses in molecular imaging. *Acc. Chem. Res.* **2021**, *54*, 3060-70. [DOI](#) [PubMed](#)
5. Tu, L.; Xie, Y.; Li, Z.; Tang, B. Aggregation-induced emission: red and near-infrared organic light-emitting diodes. *SmartMat* **2021**, *2*, 326-46. [DOI](#)
6. Yoshida, K.; Gong, J.; Kanibolotsky, A. L.; Skabara, P. J.; Turnbull, G. A.; Samuel, I. D. W. Electrically driven organic laser using integrated OLED pumping. *Nature* **2023**, *621*, 746-52. [DOI](#) [PubMed](#) [PMC](#)
7. Hong, G.; Gan, X.; Leonhardt, C.; et al. A brief history of OLEDs-emitter development and industry milestones. *Adv. Mater.* **2021**, *33*, e2005630. [DOI](#) [PubMed](#)
8. Thakur, K.; van der Zee, B.; Sachnik, O.; et al. Effect of *tert*-butylation on the photophysics of thermally activated delayed fluorescence emitters. *Adv. Photonics. Res.* **2024**, *5*, 2400022. [DOI](#)
9. Stavrou, K.; Franca, L. G.; Danos, A.; Monkman, A. P. Key requirements for ultraefficient sensitization in hyperfluorescence organic light-emitting diodes. *Nat. Photon.* **2024**, *18*, 554-61. [DOI](#)
10. Eggeman, A. S.; Illig, S.; Troisi, A.; Sirringhaus, H.; Midgley, P. A. Measurement of molecular motion in organic semiconductors by thermal diffuse electron scattering. *Nat. Mater.* **2013**, *12*, 1045-9. [DOI](#) [PubMed](#)
11. Wei, Q.; Fei, N.; Islam, A.; et al. Small-molecule emitters with high quantum efficiency: mechanisms, structures, and applications in OLED devices. *Adv. Opt. Mater.* **2018**, *6*, 1800512. [DOI](#)
12. Kuang, C.; Li, S.; Murtaza, I.; et al. Enhanced horizontal dipole orientation by novel penta-helicene anthracene-based host for efficient blue fluorescent OLEDs. *Small* **2024**, *20*, e2311114. [DOI](#) [PubMed](#)
13. Cho, H. H.; Congrave, D. G.; Gillett, A. J.; et al. Suppression of Dexter transfer by covalent encapsulation for efficient matrix-free narrowband deep blue hyperfluorescent OLEDs. *Nat. Mater.* **2024**, *23*, 519-26. [DOI](#) [PubMed](#) [PMC](#)
14. Gómez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; et al. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nat. Mater.* **2016**, *15*, 1120-7. [DOI](#)
15. Joung, J. F.; Han, M.; Hwang, J.; Jeong, M.; Choi, D. H.; Park, S. Deep learning optical spectroscopy based on experimental database: potential applications to molecular design. *JACS. Au.* **2021**, *1*, 427-38. [DOI](#) [PubMed](#) [PMC](#)
16. Shi, H.; Shi, Y.; Liang, Z.; et al. Machine learning-enabled discovery of multi-resonance TADF molecules: unraveling PLQY predictions from molecular structures. *Chem. Eng. J.* **2024**, *494*, 153150. [DOI](#)
17. Sun, M.; Fu, C.; Su, H.; et al. Enhancing chemistry-intuitive feature learning to improve prediction performance of optical properties. *Chem. Sci.* **2024**, *15*, 17533-46. [DOI](#) [PubMed](#) [PMC](#)
18. Weiss, T.; Mayo Yanes, E.; Chakraborty, S.; Cosmo, L.; Bronstein, A. M.; Gershoni-Poranne, R. Guided diffusion for inverse molecular design. *Nat. Comput. Sci.* **2023**, *3*, 873-82. [DOI](#) [PubMed](#)
19. Zeni, C.; Pinsler, R.; Zügner, D.; et al. A generative model for inorganic materials design. *Nature* **2025**, *639*, 624-32. [DOI](#) [PubMed](#) [PMC](#)
20. Alakhdar, A.; Poczos, B.; Washburn, N. Diffusion models in de novo drug design. *J. Chem. Inf. Model.* **2024**, *64*, 7238-56. [DOI](#) [PubMed](#) [PMC](#)
21. Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; et al. Automatic chemical design using a data-driven continuous representation of molecules. *ACS. Cent. Sci.* **2018**, *4*, 268-76. [DOI](#) [PubMed](#) [PMC](#)
22. Zhang, K.; Yang, X.; Wang, Y.; et al. Artificial intelligence in drug development. *Nat. Med.* **2025**, *31*, 45-59. [DOI](#)

23. Moret, M.; Friedrich, L.; Grisoni, F.; Merk, D.; Schneider, G. Generative molecular design in low data regimes. *Nat. Mach. Intell.* **2020**, *2*, 171-80. [DOI](#)
24. Sumita, M.; Terayama, K.; Suzuki, N.; et al. De novo creation of a naked eye-detectable fluorescent molecule based on quantum chemical computation and machine learning. *Sci. Adv.* **2022**, *8*, eabj3906. [DOI](#) [PubMed](#) [PMC](#)
25. Joung, J. F.; Han, M.; Jeong, M.; Park, S. Experimental database of optical properties of organic compounds. *Sci. Data.* **2020**, *7*, 295. [DOI](#) [PubMed](#) [PMC](#)
26. Gong, J.; Gong, W.; Wu, B.; et al. ASBase: the universal database for aggregate science. *Aggregate* **2023**, *4*, e263. [DOI](#)
27. Li, P.; Wang, Z.; Li, W.; Yuan, J.; Chen, R. Design of thermally activated delayed fluorescence materials with high intersystem crossing efficiencies by machine learning-assisted virtual screening. *J. Phys. Chem. Lett.* **2022**, *13*, 9910-8. [DOI](#)
28. Kim, H.; Lee, K.; Kim, J. H.; Kim, W. Y. Deep learning-based chemical similarity for accelerated organic light-emitting diode materials discovery. *J. Chem. Inf. Model.* **2024**, *64*, 677-89. [DOI](#)
29. Blaskovits, J. T.; Laplaza, R.; Vela, S.; Corminboeuf, C. Data-driven discovery of organic electronic materials enabled by hybrid top-down/bottom-up design. *Adv. Mater.* **2024**, *36*, e2305602. [DOI](#) [PubMed](#)
30. Zdrazil, B.; Felix, E.; Hunter, F.; et al. The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic. Acids. Res.* **2024**, *52*, D1180-92. [DOI](#) [PubMed](#) [PMC](#)
31. Guo, J.; Sun, M.; Zhao, X.; et al. General graph neural network-based model to accurately predict cocrystal density and insight from data quality and feature representation. *J. Chem. Inf. Model.* **2023**, *63*, 1143-56. [DOI](#)
32. Medina-Franco, J. L.; Martínez-Mayorga, K.; Bender, A.; Scior, T. Scaffold diversity analysis of compound data sets using an entropy-based measure. *QSAR. Comb. Sci.* **2009**, *28*, 1551-60. [DOI](#)
33. Niu, Y.; Li, W.; Peng, Q.; et al. MOlecular MATerials Property Prediction Package (MOMAP) 1.0: a software package for predicting the luminescent properties and mobility of organic functional materials. *Mol. Phys.* **2018**, *116*, 1078-90. [DOI](#)
34. Park, Y.; Lee, J.; Jung, D. H.; et al. An aromatic imine group enhances the EL efficiency and carrier transport properties of highly efficient blue emitter for OLEDs. *J. Mater. Chem.* **2010**, *20*, 5930. [DOI](#)
35. Kotaka, H.; Konishi, G.; Mizuno, K. Synthesis and photoluminescence properties of π -extended fluorene derivatives: the first example of a fluorescent solvatochromic nitro-group-containing dye with a high fluorescence quantum yield. *Tetrahedron. Lett.* **2010**, *51*, 181-4. [DOI](#)
36. Liu, X.; Liang, F.; Ding, L.; et al. The study on two kinds of spiro systems for improving the performance of host materials in blue phosphorescent organic light-emitting diodes. *J. Mater. Chem. C.* **2015**, *3*, 9053-6. [DOI](#)