Chinese Chemical Letters 37 (2026) 112633

Contents lists available at ScienceDirect

Chinese Chemical Letters

journal homepage: www.elsevier.com/locate/cclet

Data-driven generation and efficient screening of MR-TADF materials

Ya-Jun Yin[a,b,1], Li-Fang Yin[a,1], Yi Zhao[a,1], Xin Xu[a], Yu-Qi Xia[a], Jing-Jing Zhao[a], Jia-Qi Bai[a], Guang-Jun Nan[c,*], Ji-Fen Wang[b], Lu-Yi Zou[a,*]

[a] Institute of Theoretical Chemistry, College of Chemistry, Jilin University, Changchun 130023, China
[b] School of Resources and Environmental Engineering, Shanghai Polytechnic University, Shanghai Engineering Research Center of Advanced Thermal Functional Materials, Shanghai 201209, China
[c] Department of Physics, Zhejiang Normal University, Jinhua 321004, China

ARTICLE INFO

Article history:
Received 28 November 2025
Revised 5 March 2026
Accepted 16 March 2026
Available online 17 March 2026

Keywords:
MR-TADF
Machine learning
Transformer-VAE
Singlet-triplet energy gap
Full width at half maximum

ABSTRACT

The pursuit of high color purity in next-generation displays has spurred intense interest in multiple-resonance thermally activated delayed fluorescence (MR-TADF) emitters, which combine narrowband emission with high exciton utilization efficiency. However, the intricate relationship between molecular structure and photophysical properties renders rational design highly challenging. Herein, we present an integrated computational framework that unifies automated data extraction, *de novo* molecular generation and multi-property prediction for the targeted discovery of MR-TADF materials. Our approach leverages large-language-model-based literature mining and chemical structure recognition to construct a curated dataset (MR585), which guides a Transformer-based variational autoencoder for generative molecular design. Machine learning models trained on this dataset achieve test-set R^2 values of 0.692 for predicting the singlet-triplet energy gap (ΔE_{ST}) and 0.788 for full width at half maximum (FWHM). From the generated candidates, we identify 15 structurally novel emitters with ΔE_{ST} values of 0.13–0.33 eV and narrow FWHMs of 26.2–46.6 nm. Theoretical analysis further reveals that two sulphur-containing BCzBN derivatives exhibit enhanced spin-orbit coupling, accelerated reverse intersystem crossing and reduced efficiency roll-off compared to their parent structures. This work establishes a comprehensive AI-driven platform that bridges automated knowledge extraction with generative chemistry, offering an efficient and scalable pathway towards high-performance narrowband emitters.

© 2026 Published by Elsevier B.V. on behalf of Chinese Chemical Society and Institute of Materia Medica, Chinese Academy of Medical Sciences.

In recent years, thermally activated delayed fluorescence (TADF) materials have attracted considerable attention as next-generation emitters for organic light-emitting diodes (OLEDs), due to their ability to achieve high exciton utilization efficiency without relying on precious metals [1]. These materials operate by minimizing the energy gap (ΔE_{ST}) between the lowest singlet (S_1) and triplet (T_1) excited states, thereby facilitating reverse intersystem crossing (RISC), a process in which non-radiative triplet excitons are thermally upconverted into radiative singlet excitons. This mechanism theoretically enables an internal quantum efficiency approaching 100 % [2–4]. Conventional TADF emitters typically adopt a donor–acceptor (D-A) molecular architecture to spatially separate the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO), effectively reducing ΔE_{ST} [5–9]. However, the torsional distortion between the donor and acceptor units often induces strong vibronic coupling between the ground (S_0) and S_1 states, accompanied by significant structural relaxation in the excited state. As a result, the emission spectra become markedly broadened, with full width at half maximum (FWHM) typically ranging from 70 nm to 100 nm [10,11], which fails to meet the stringent color purity requirements of high-resolution display technologies.

To overcome the intrinsic limitation of conventional D–A type TADF materials, which struggle to simultaneously achieve a small ΔE_{ST} and narrow FWHM, Hatakeyama et al. introduced the concept of multiple-resonance TADF (MR-TADF). In this design, rigid aromatic backbones are functionalized with alternating electron-deficient (*e.g.*, boron) and electron-rich (*e.g.*, nitrogen) heteroatoms, creating opposing resonance effects that spatially alternate the HOMO and LUMO distributions at the atomic level. This electronic arrangement effectively reduces ΔE_{ST} while enhancing the RISC rate [12,13]. Furthermore, the structural rigidity suppresses excited-state relaxation and vibronic coupling, enabling narrowband emission without compromising exciton utilization efficiency

* Corresponding authors.
E-mail addresses: gjnan@zjnu.edu.cn (G.-J. Nan), zouly@jlu.edu.cn (L.-Y. Zou).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.cclet.2026.112633
1001-8417/© 2026 Published by Elsevier B.V. on behalf of Chinese Chemical Society and Institute of Materia Medica, Chinese Academy of Medical Sciences.

[14,15]. As a result, MR-TADF emitters strike an optimal balance between efficiency and color purity, demonstrating great promise for high-resolution OLED applications [16].

Nevertheless, the intricate coupling between molecular structure and photophysical processes remains a key bottleneck, limiting the photothermal stability of MR-TADF materials. In addition, device fabrication and performance validation are labor-intensive and highly process-dependent, posing practical barriers to further development in this field [17]. Current research predominantly focuses on local modifications of known B/N-based resonance frameworks, yielding incremental progress in property modulation. For instance, Jin *et al*. introduced various donor and acceptor substituents at para- or meta-positions on the BNCz backbone, achieving fine-tuned emission properties and color control. Their results verified the effectiveness of the "fixed backbone + local substitution" strategy in balancing device efficiency and spectral purity [18]. However, most existing studies remain confined to optimizing a limited number of known molecular scaffolds, lacking systematic exploration of entirely new structural frameworks.

With the rapid advancement of computational methods, theoretical modeling has become increasingly vital for guiding the rational design of MR-TADF materials. These approaches not only aid in performance optimization but also deepen understanding of structure-property relationships, thereby accelerating the discovery of new-material with improved efficiency and success rates. However, the evaluation of excited-state properties in MR-TADF systems still depends heavily on quantum-chemical methods such as time-dependent density functional theory (TD-DFT) and spin-component-scaled second-order coupled cluster (SCS-CC2) calculations [19,20]. Although these methods are highly reliable, their substantial computational cost renders them impractical for large-scale high-throughput screening of molecular candidates.

Recent advances in artificial intelligence (AI) offer promising pathways to overcome these limitations. Future research on MR-TADF materials should focus on two key directions: First, exploring molecular frameworks that go beyond traditional B/N resonance architectures to enhance structural diversity; and second, integrating generative AI with automated design methodologies to accelerate the intelligent evolution of materials discovery [18,21]. In line with this trend, an increasing number of studies have begun to employ machine learning (ML) models as reliable surrogates for or alternatives to conventional quantum-chemical calculations, thereby attaining prediction accuracy and enhancing efficiency of structural exploration [22–25]. For instance, Shi *et al*. combined electrotopological state (EState) and electronic-property descriptors with Extreme Gradient Boosting (XGBoost) and neural networks to construct a photoluminescence quantum yield (PLQY) regression model for 402 MR-TADF molecules, achieving a test accuracy of up to 88 % [26]. By coupling this model with a variational autoencoder (VAE), they further realized large-scale virtual molecule generation and high-throughput screening, effectively expanding the search space for high-PLQY blue emitters [26]. Similarly, Chen *et al*. developed a structure-aware ensemble learning model that achieves high-precision predictions of FWHM and PLQY even under small-sample conditions; their Shapley Additive Explanations (SHAP)–based analysis clarified the roles of molecular rigidity, resonance fragments, and photophysical behavior [27]. Notably, the quantitative structure-property relationships model developed by Adachi's group, which combines molecular fingerprints with kernel partial least squares, exhibits strong predictive performance for the peak emission wavelength ($R^2 = 0.87$, RMSE = 21.4) and FWHM of over 400 MR-TADF molecules [25]. This model successfully guided the design and synthesis of the deep-blue emitter ν-DABNA-O-xy, yielding OLED devices with an external quantum efficiency of 41.3 % [25], further demonstrating the practical utility of machine learning in real device-level MR-TADF material design.

Despite these advances, current approaches still exhibit notable constraints. Most models depend heavily on high-quality three-dimensional structural inputs and quantum-chemistry-derived prior data, which hinders rapid deployment during the early stages of molecular generation. Furthermore, existing studies tend to focus on optimizing a single property, such as PLQY, without systematically accounting for the trade-off between structural rationality and synthetic feasibility, thereby limiting the overall efficiency of multi-objective screening. Additionally, earlier efforts primarily relied on limited experimental datasets or statistical regression based on synthesized molecules, resulting in narrow structural coverage and insufficient generalization to unseen chemical spaces [27–29].

To address these challenges, our group previously developed a SOGCN (Screening for OLED materials by Graph Convolutional neural Networks) model for molecular property prediction [30]. Using the MR220 dataset and a tailored data-augmentation strategy, the model achieved high-accuracy predictions of key photophysical parameters such as ΔE_{ST} and FWHM for MR-TADF materials, with predicted values showing strong agreement with experimental measurements. This significantly improved the efficiency of molecular screening. Building upon this foundation, the present study establishes a comprehensive intelligent framework encompassing automated data acquisition, de novo molecular generation, and multi-property prediction to enable systematic and targeted design of MR-TADF emitters.

In the data-acquisition process, we employed the Qwen Max large language model (LLM) in combination with prompt-engineering techniques to automate the extraction of chemical knowledge from scientific literature. Dedicated prompt templates were designed to identify key parameters such as ΔE_{ST} and FWHM, guiding the model to accurately locate and interpret dispersed textual descriptions and convert them into standardized data formats. Compared with the rule-based and template-matching ChemDataExtractor system [31], our approach provides a more adaptive pathway for chemical information extraction by emphasizing deep semantic understanding and cross-sentence reasoning. It enables the model to capture latent semantic relationships embedded in complex narrative contexts, thereby reducing manual effort while ensuring data consistency and quality.

To enhance model generalization, we adopt a SMILES randomization strategy for data augmentation. By systematically generating multiple equivalent SMILES sequences for each molecule, the model learns intrinsic compositional patterns of key structural motifs, such as conjugated backbones and B/N-containing heteroatomic frameworks, mitigating overfitting and sequence bias inherent to SMILES-based learning [27,28].

In the structure generation step, a Transformer-based VAE (Transformer-VAE) module was employed for latent-space modeling. A physics-guided regularization term was incorporated into the training objective, which ensures that, while learning the grammatical rules of SMILES representation, the model also captures essential physical characteristics of MR-TADF systems. This design enables the generation of candidate molecules that exhibit both structural diversity and physical plausibility.

In the property-prediction module, we developed a lightweight regression model that takes SMILES strings as input and rapidly predicts ΔE_{ST} and FWHM with high accuracy. This provides a computationally efficient surrogate for quantum-chemical methods when dealing with large molecular systems, offering a practical and efficient alternative for high-throughput screening.

Finally, to enable multi-objective optimization, our framework integrates multiple molecular descriptors, including topological scores, heteroatom density within the main skeleton, synthetic accessibility indices, and predicted excited-state properties. The integration of these multidimensional features provides a comprehensive and quantitative basis for rational molecular selection. As il-

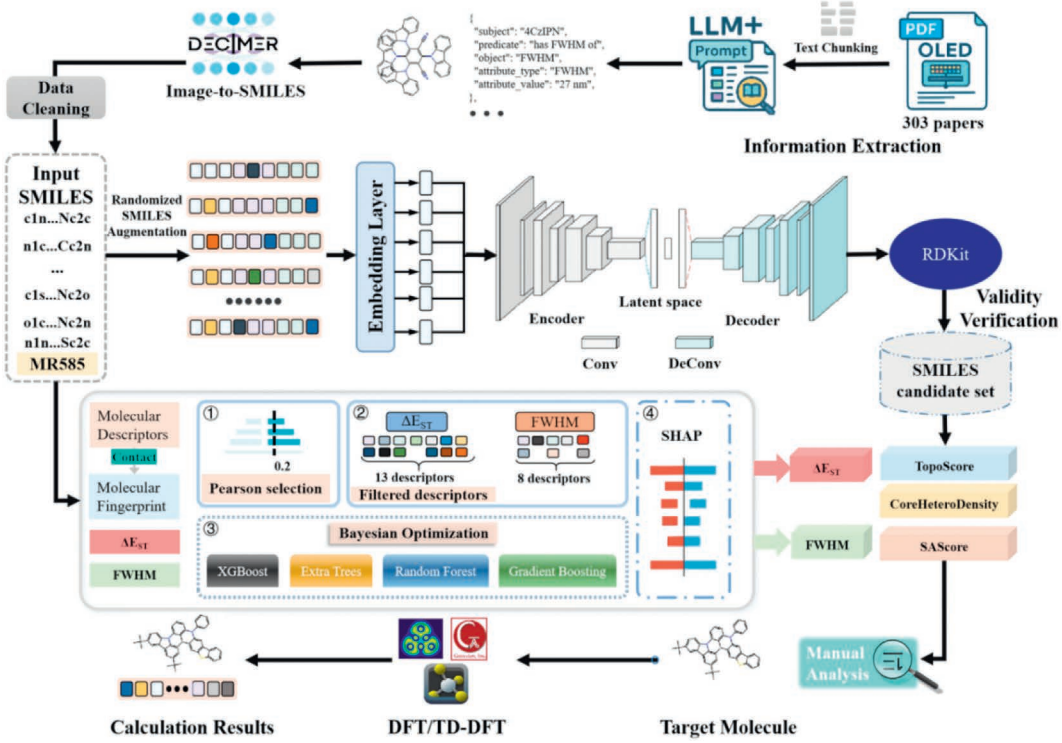


Fig. 1. Overall workflow of the data-driven MR-TADF design platform.

lustrated in Fig. 1, the resulting "generation–prediction–screening" integrated platform establishes a fully connected workflow, providing a robust support for systematic design and performance optimization of MR-TADF emitters.

To systematically explore the structural diversity of MR-TADF materials and construct a high-quality dataset, we established a comprehensive data curation and annotation pipeline. During the literature retrieval stage, we primarily searched major academic databases (*e.g.*, Web of Science Core Collection, Scopus, and Google Scholar), complemented by targeted searches on journal platforms; query terms were formulated by combining keywords such as "MR-TADF/multi-resonance TADF", "narrowband OLED", "BN-containing emitter/B,N-doped", "ΔE_{ST}", and "FWHM". We then employed the Qwen Max large language model, guided by structured prompts specifically designed for photophysical parameter extraction, to automatically identify and extract key data (*e.g.*, ΔE_{ST} and FWHM) from 265 relevant research articles. During extraction, the model leveraged contextual cues to impose semantic constraints between compound names and their corresponding numerical values, ensuring referential consistency and unambiguous interpretation (Fig. 2). This strategy significantly improved extraction accuracy under complex linguistic contexts.

For molecular structures presented as images in the literature, we used the DECIMER image recognition tool to automatically convert graphical depictions into standardized SMILES representations [32–34]. Since literature figures often contain non-structural elements such as atom numbering, reaction arrows, and condition annotations, which may interfere with recognition, we performed a lightweight manual preprocessing step before DECIMER parsing when necessary. Specifically, the original figure was cropped and cleaned using standard image editing software to retain only the molecular structure region, while non-structural overlays such as numbering labels, arrows, condition text, leader lines, and other annotations were masked or removed. The processed images were then resubmitted to DECIMER for recognition, and any remaining ambiguous cases were resolved by manual inspection and correction. All recognized structures were manually verified to ensure consistency between the molecular diagrams and the corresponding SMILES codes. For molecules appearing across multiple sources, we performed deduplication by comparing InChIKeys and retained the most complete records, merging duplicate entries into a unified data entry for each molecule. As a result, we constructed a dedicated MR-TADF dataset containing 585 molecules and their associated excited-state parameters (referred to as MR585), which serves as the core foundation for subsequent model training and performance prediction within the generative framework.

To evaluate the chemical diversity of the constructed dataset and prevent the generative model from being constrained by redundant or structurally homogeneous samples, two complementary quantitative metrics were employed. First, the structural Shannon information entropy was calculated based on InChIKey identifiers to assess the uniformity of structural-type distribution within the training set. Second, Tanimoto distances were computed using Morgan fingerprints (radius = 2, 1024 bits) to characterize the pairwise molecular similarity distribution [35]. These two metrics respectively reflect the evenness of structural-type representation and the degree of local configurational variation among molecules. The corresponding mathematical expressions are given below:

$$H = -\sum_{i=1}^{M} p_i \log_2 p_i$$
$$\bar{D} = \frac{2}{N(N-1)} \sum_{1 \le i < j \le N} \left(1 - \frac{A_i \cdot A_j}{\| A_i \|^2 + \| A_j \|^2 - A_i \cdot A_j} \right) \quad (1)$$

Here, H denotes the structural Shannon information entropy, and $\bar{D}$ represents the average Tanimoto distance. p_i is the occurrence frequency of the i-th structural class in the training set. A_i and A_j denote the fingerprint vectors of the i-th and j-th molecules, respectively, and N is the total number of samples. The term $A_i \cdot A_j$ corresponds to the number of shared structural features between two molecules. The constraint "$1 \le i < j \le N$" in

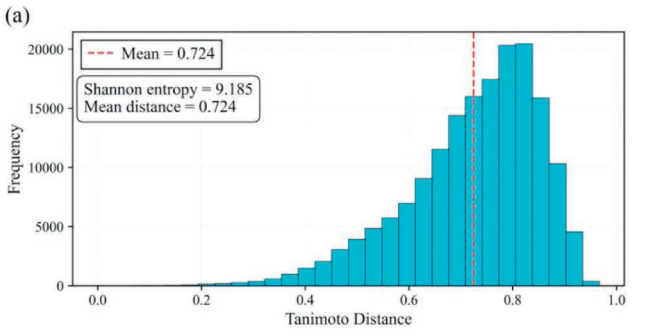


Fig. 2. Workflow for extracting material properties from the literature.

the equation is applied to avoid redundant computation and self-comparison.

Statistical analysis of the MR-TADF dataset (Fig. 3) reveals that its molecular structures exhibit high diversity within the chemical space. As shown in Figs. 3a and c, the MR585 dataset exhibits significant structural diversity in molecular space. Pairwise similarity analysis using Morgan fingerprints yields a mean Tanimoto distance of 0.724 and a Shannon entropy of 9.185, indicating low overall structural similarity and broad coverage of chemical space. Currently, publicly available and standardized datasets specifically targeting MR-TADF materials remain limited, and many of the reported datasets are constructed primarily for task-specific modeling purposes without releasing complete molecular structures and annotations in a reusable form [36]. To ensure a fair and reproducible dataset-level comparison, we compared MR585 with two representative and publicly available MR-TADF datasets previously published by our group, namely MR220 [30] and MR262 [28]. As shown in Fig. 3c, MR585 achieves the highest values in both Shannon entropy (9.185 *vs.* 7.687 and 7.890) and mean Tanimoto distance (0.724 *vs.* 0.698 and 0.704), suggesting a more balanced scaffold distribution and lower structural redundancy. These characteristics are favorable for learning more transferable structure-property relationships. This characteristic enables ML models to learn broader structure–property mapping relationships, thereby enhancing their extrapolation predictive capability. Fig. 3b displays the two-dimensional projection obtained by applying t-distributed stochastic neighbor embedding (t-SNE) dimensionality reduction to the molecular fingerprints. The t-SNE method effectively preserves local structural relationships in high-dimensional space, clearly revealing the distribution patterns of molecules in the chemical space. The results show that the molecules are widely distributed in the projection without forming large-scale clusters, further confirming the global diversity of the dataset. At the same time, several local clustering regions can be observed, indicating high local similarity within certain structural scaffold families. The presence of such local similarity helps the model capture the influence of subtle structural changes on properties, thereby improving prediction accuracy. Coloration results indicate a significant correlation between the number of heteroatoms (including B, N, O, S, Se, P, *etc.*) and the distribution of molecules in the chemical space: molecules with a higher number of heteroatoms (*e.g.*, 18) tend to cluster in specific regions, while those with fewer heteroatoms (*e.g.*, only 1) are more dispersed. Representative structures with the highest and lowest numbers of heteroatoms are further annotated in the figure, intuitively illustrating their structural differences. This demonstrates that the number of heteroatoms, as a key structural descriptor, is closely related to the positioning of molecules in the chemical space, providing important clues for understanding how structural features influence photophysical properties (such as ΔE_{ST} and FWHM).

From the perspective of SMILES encoding and generative modeling, the presence of heteroatoms affects not only the fingerprint representation but also the local pattern formation in sequence modeling. For sequence-based generative architectures such as the Transformer-VAE, inserting heteroatoms into a carbon-dominated backbone introduces new syntactic nodes and turning points in the molecular sequence, thereby altering the conditional probability distribution for subsequent token prediction. This implies that



(c)

Dataset	Shannon entropy	Mean distance
MR220	7.687	0.698
MR262	7.890	0.704
MR585	9.185	0.724

Fig. 3. Structural diversity analysis and chemical space distribution of the MR-TADF datasets. (a) MR585: Histogram of pairwise Tanimoto distances (Morgan fingerprints, $r = 2$, 1024 bits). (b) MR585: 2D t-SNE projection colored by heteroatom count (perplexity = 30, random_state = 42). (c) Comparison of Shannon entropy and mean Tanimoto distance for MR220, MR262, and MR585.

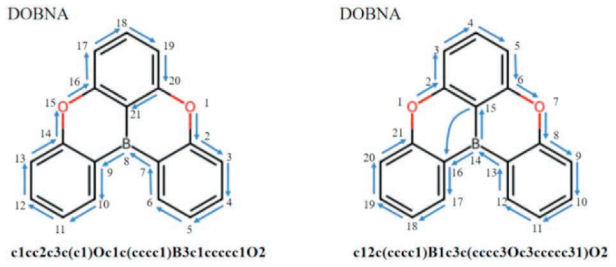


Fig. 4. Different SMILES representations for DOBNA [37].

heteroatoms serve not only a chemical function but also act as "structural branching points" in sequence space, significantly shaping the model's ability to learn local structural motifs and expanding the diversity of representations within the latent space.

To mitigate the model's structural bias toward specific SMILES sequences and enhance its ability to generalize across diverse molecular topologies, a SMILES randomization strategy was employed for data augmentation. This method systematically alters the traversal order of atoms and bonding paths while strictly preserving the underlying molecular topology, thereby generating multiple semantically equivalent SMILES representations for each molecule (Fig. 4). Such a design encourages the model to focus on learning the intrinsic structural features of molecules rather than memorizing fixed string patterns, effectively strengthening its capacity for deeper chemical structure understanding during training.

To achieve systematic data augmentation, the original molecular set is defined as $M = m_1, m_2, \ldots, m_N$. For each molecule m_i, K semantically equivalent SMILES representations are generated through a randomization algorithm. The augmented training set can thus be expressed as:

$$\mathcal{D}_{aug} = \bigcup_{m_i \in M} \{\text{SMILES}(m_i) \cup \{\text{SMILES}^{(1)}_{\text{rand}}(m_i), \ldots, \text{SMILES}^{(K)}_{\text{rand}}(m_i)\}\} \quad (2)$$

where $\text{SMILES}^{(\text{k})}_{\text{rand}}(\text{m}_\text{i})$ denotes the K-th randomized SMILES variant generated through stochastic traversal of the molecular graph.

The generative model adopts a Transformer-VAE comprising an encoder, a latent-space module, and a decoder [38]. The encoder stacks eight layers of multi-head self-attention and position-wise feed-forward networks and employs a pre-layer normalization scheme to improve optimization stability in deep architectures. Input SMILES strings are mapped to dense embeddings and augmented with sinusoidal positional encodings to capture the ordering of atomic connectivities and local structural context, enabling the model to learn latent regularities such as aromatic ring topology and heteroatom distribution.

In contrast, VAE variants built on recurrent neural networks or fully connected layers often struggle to preserve chemically valid atom connectivity for structurally complex, higher-molecular-weight compounds, leading to bond inconsistencies or repetitive token outputs (*e.g.*, runs of C atoms) [39,40]. By leveraging global self-attention to model long-range interatomic dependencies, the Transformer-VAE produces sequences that better adhere to molecular structural rules and thereby improves the structural validity of generated MR-TADF candidates.

After the encoder output is globally pooled and projected into the latent space, the mean vector μ and the logarithmic variance $\log(\sigma^2)$ are obtained. The latent variable is then generated through the reparameterization trick, expressed as:

$$z = \mu + \sigma \cdot \varepsilon,\ \varepsilon \sim \mathcal{N}(0, I) \quad (3)$$

here, z denotes the latent variable vector representing the compressed structural features of a molecule in the latent space. μ is the mean vector, predicted by the encoder, indicating the central position of the latent distribution; σ is the standard deviation vector, controlling the spread of the distribution in latent space; ε is a random noise term sampled from a standard normal distribution $\mathcal{N}(0, \text{I})$, enabling differentiable stochastic sampling; and I is the identity matrix that ensures independence across sampling dimensions.

The decoder adopts a Transformer-based architecture that conditions on the latent vector z to autoregressively generate SMILES sequences. Structural validity is verified using RDKit (version 2024.9.4), and chemically invalid outputs are filtered out prior to downstream analysis.

Training is performed with the Adam optimizer [41] using an initial learning rate of 1.0×10^{-4}, a batch size of 256, and a fixed training horizon of 10,000 epochs to ensure comparability across different augmentation-ratio settings. The learning-rate schedule includes a warmup phase (4000 steps) followed by a square-root decay, and the KL term is applied with a fixed weight of $\beta = 0.001$ to stabilize optimization and mitigate posterior collapse. Recognizing that no single metric can comprehensively capture generative performance, we systematically evaluate the model along four complementary dimensions (validity, diversity, novelty and duplication rate), thereby providing a more complete and interpretable assessment protocol. Formal definitions and symbol conventions for all metrics are provided in Table S1.

To systematically evaluate the effect of data augmentation magnitude on generative model performance, comparative experiments were conducted across augmentation factors ranging from 2× to

10×. In each trial, 10,000 SMILES sequences were generated to examine how structural validity and uniqueness varied with the level of augmentation. As summarized in Table S2 (Supporting information), increasing the augmentation factor markedly enhanced molecular uniqueness. A 6× augmentation achieved a relatively balanced trade-off between uniqueness (3.14 %) and structural validity (66.8 %). However, this "balance point" does not represent the optimal configuration, as the primary goal of this study is to maximize chemical-space exploration and yield the largest number of structurally distinct molecules. From this perspective, 10× augmentation, despite a lower validity of 59.8 %, attained the highest uniqueness (4.01 %) and the best overall success rate (2.4 %), indicating its superior effectiveness in generating novel structures.

Overall, increasing the augmentation factor does not lead to a linear improvement in model performance. Instead, moderate to high augmentation levels (6× −10×) are more effective at broadening the model's structural exploration space and reducing overlap with structures already present in the training set. Accordingly, we selected 10× augmentation based on the comparative results in Table S2. Under a fixed generation budget (10,000 samples per setting), 10× achieved the highest novelty/uniqueness (4.01 %) and the best overall success rate (2.4 %), which matches our primary objective of maximizing chemical-space exploration. Lower ratios (*e.g.*, 2× −6×) provide higher validity, but yield fewer structurally distinct molecules. Augmentation levels beyond 10× were not pursued because further increases offer limited gains in uniqueness while the reduced validity would substantially increase the burden of downstream screening and filtering. With 10× augmentation, combined with structural screening and similarity filtering (Tanimoto threshold = 0.55), 15 structurally unique and chemically plausible molecules were retained (Fig. S1 in Supporting information), highlighting the advantage of higher augmentation in enhancing structural diversity. The choice of the 0.55 cutoff is also supported by systematic analyses: For commonly used fingerprints such as ECFP4, compounds with substantial structural differences typically show Tanimoto coefficients in the range of 0.4–0.6 [42]. Therefore, we adopt 0.55 as the filtering threshold to maximize structural diversity while maintaining chemical reasonableness.

To enable rapid performance evaluation of generated molecules, a machine learning-based screening mechanism was developed specifically for predicting the excited-state properties of MR-TADF systems. The framework aims to provide fast and accurate predictions of key parameters, ΔE_{ST} and FWHM, serving as a reliable foundation for subsequent molecular prioritization. Considering that MR-TADF compounds often exhibit intricate resonance structures and that available datasets remain limited, the proposed approach avoids the time-consuming structural optimization typically required in conventional quantum-chemical methods. Instead, a lightweight regression model was constructed by integrating two-dimensional molecular descriptors with fingerprint-based structural features. This hybrid strategy maintains high predictive accuracy while substantially reducing computational cost, thereby offering a practical and scalable pathway for high-throughput molecular screening.

To support the construction of the predictive model, a structural space analysis was first performed on the 585 MR-TADF molecules. Each molecule was processed using the RDKit toolkit to generate its binary Morgan fingerprint vector $\mathrm{f}_i \in \{0,1\}^{1024}$ with a radius of 2. Subsequently, the Tanimoto distance d_{ij} between any two molecules (i, j) was calculated as:

$$d_{ij} = 1 - \frac{\left|\mathrm{f}_i \cap \mathrm{f}_j\right|}{\left|\mathrm{f}_i \cup \mathrm{f}_j\right|} = 1 - \frac{\sum_{k=1}^{1024} \mathrm{f}_{i,k} \cdot \mathrm{f}_{j,k}}{\sum_{k=1}^{1024} \max\left(\mathrm{f}_{i,k}, \mathrm{f}_{j,k}\right)} \quad (4)$$

where d_{ij} represents the structural dissimilarity between molecules i and j; $\mathrm{f}_i \cap \mathrm{f}_j$ and $\mathrm{f}_i \cup \mathrm{f}_j$ denote the intersection and union of structural features encoded in their fingerprint vectors, respectively. A smaller d_{ij} value indicates a higher degree of structural similarity between the two molecules.

Accordingly, the distance matrix $D = [d_{ij}] \in \mathbb{R}^{585 \times 585}$ was obtained, representing the pairwise topological dissimilarities among all MR-TADF molecules. To visualize the overall structural distribution and capture how molecular features vary with ΔE_{ST} and FWHM, classical multidimensional scaling (MDS) was employed to project the high–dimensional distance matrix onto a two-dimensional plane [43]. The transformation is defined as:

$$X = \mathrm{MDS}(\mathrm{D}) \in \mathbb{R}^{585 \times 2} \quad (5)$$

where $\mathbb{R}^{585 \times 2}$ denotes the two-dimensional real space containing 585 sample points, and $\mathrm{X}_i = (\mathrm{x}_i, \mathrm{y}_i)$ represents the coordinates of molecule i in the 2D projection space. This allows direct visualization of the relative distribution and structural relationships among different MR-TADF molecules.

We visualized and analyzed two key photophysical properties (ΔE_{ST} and FWHM), within a unified structural feature space (Fig. 5). Statistical analysis of the dataset shows that ΔE_{ST} ranges from 0.004 eV to 0.680 eV, while FWHM spans 12–116 nm. This indicates that the studied molecules exhibit a broad variation range in both critical photophysical properties, laying a solid data foundation for in-depth exploration of their intrinsic correlation. To screen high-quality candidate molecules with both small ΔE_{ST} and narrow FWHM, this study adopted a multi-objective collaborative screening strategy. First, molecules were independently ranked based on ΔE_{ST} and FWHM respectively. Then, molecules that ranked high in both properties and had consistent structures were selected to form an intersection subset. Finally, molecules that ranked among the top 10 in both properties were further filtered from this subset (Fig. 5c). This screening method ensures that the selected molecules perform excellently in key performance indicators, thereby guaranteeing their overall performance reaches an optimal level. Structural analysis of the screened intersection subset revealed that local structures such as rigid B/N-fused polycyclic fragments appear frequently. The repeated occurrence of these structural units indicates that they play a crucial role in the synergistic regulation of ΔE_{ST} and FWHM.

In constructing input features, this study comprehensively characterized molecular structures while avoiding human-imposed biases by automatically computing all 217 two-dimensional molecular descriptors from SMILES representations using RDKit. These descriptors encompass a broad range of structural and physicochemical properties, such as topological complexity, polar surface area, charge distribution, EState energy indices, and atomic connectivity patterns, thus providing a systematic description of both 6the electronic and geometric features of MR-TADF materials.

Given that the excited-state properties of MR-TADF systems are jointly governed by orbital distribution, polarization effects, and molecular rigidity, the parameters ΔE_{ST} and FWHM exhibit an approximately linear response to local structural perturbations (Fig. 6). Statistical analysis of the molecular dataset revealed a Pearson correlation coefficient of $r = 0.175$ ($p = 8.67 \times 10^{-5}$) between ΔE_{ST} and structural difference (Tanimoto distance), and $r = 0.127$ ($p = 4.61 \times 10^{-3}$) between FWHM and structural difference. Although the correlations are moderate, their statistical significance indicates that molecular topological perturbations can linearly reflect variations in excited-state properties under small-sample conditions, supporting the use of the Pearson coefficient as a valid metric for feature–property association.

Accordingly, the Pearson correlation coefficient (PCC) was employed to quantify the linear relationship between each molecular

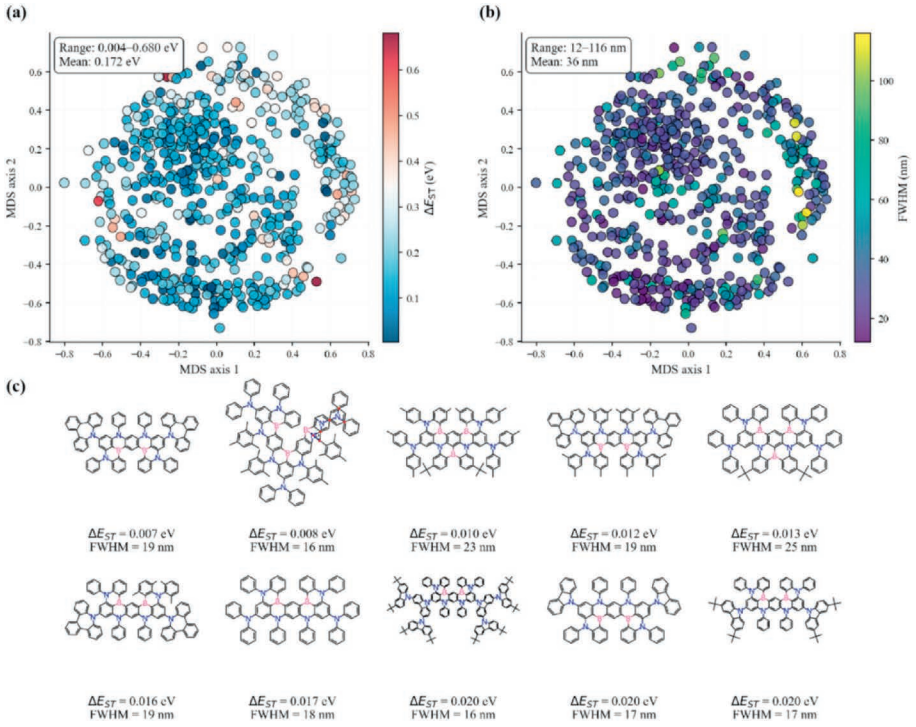


Fig. 5. Distribution of ΔE_{ST} and FWHM within the same structural feature space. (a) Visualized distribution of ΔE_{ST} and (b) the corresponding distribution of FWHM. Comparing both properties under identical embedding coordinates enables the identification of molecules exhibiting simultaneously small ΔE_{ST} and narrow FWHM. The dataset was independently ranked by ΔE_{ST} and FWHM, and (c) the top 10 molecules from each ranking were intersected.

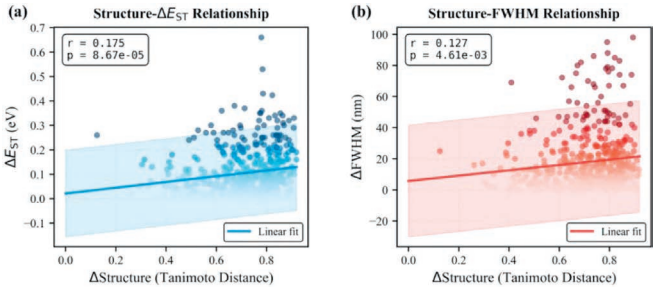


Fig. 6. (a) Variation of ΔE_{ST} with structural perturbation. (b) Variation of FWHM with structural perturbation. The x-axis (ΔStructure, Tanimoto distance) indicates the extent of structural difference, with larger values representing greater dissimilarity.

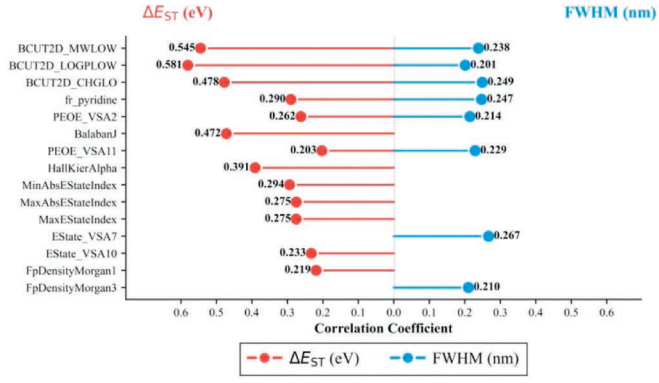


Fig. 7. The figure illustrates the distribution of Pearson correlation coefficients for ΔE_{ST} (in red) and FWHM (in blue) with various categories of molecular descriptors.

descriptor and the target property, expressed as:

$$\rho_{X,Y} = \frac{\sum_{i=1}^{n}\left(X_i - \overleftarrow{X}\right)\left(Y_i - \overleftarrow{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \overleftarrow{X}\right)^2}\cdot\sqrt{\sum_{i=1}^{n}\left(Y_i - \overleftarrow{Y}\right)^2}} \tag{6}$$

where X_i and Y_i denote the descriptor value and the corresponding measured property value of the i th sample, while $\bar{X}$ and $\bar{Y}$ represent the sample means. The coefficient ρ ranges from $[-1, 1]$; the closer the absolute value of ρ is to 1, the stronger the linear correlation, with the sign indicating the direction of the relationship.

To mitigate noise and redundancy in small-sample modeling, PCC between each molecular descriptor and the target properties (ΔE_{ST} and FWHM) were computed. Descriptors with absolute correlation values below 0.2 were excluded, retaining only those that exhibited significant contributions to property variation. The filtered results revealed that ΔE_{ST} showed strong correlations with 13 descriptors, whereas FWHM was closely associated with 8 descriptors (Fig. 7).

Analysis of the correlation distribution and the physical meanings of the descriptors (Table S3 in Supporting information) highlights a clear distinction in feature sensitivity between ΔE_{ST} and FWHM. ΔE_{ST} correlates significantly with electronic state indices (*e.g.*, EState and EState_VSA series) and topological complexity descriptors (*e.g.*, BalabanJ, HallKierAlpha), indicating that its variation is primarily governed by the nonuniformity of intramolecular electron distribution and the connectivity of molecular skeletons. In contrast, FWHM shows positive correlations with descriptors related to polar surface area and local charge distribution (*e.g.*, PEOE_VSA series, BCUT2D_CHGLO), suggesting that FWHM increases with greater molecular polarization and charge delocalization—reflecting property changes driven mainly by charge redistribution and geometric relaxation in the excited state.

Building upon the preceding analysis and the structural commonalities revealed in Fig. 7c, it becomes evident that molecular

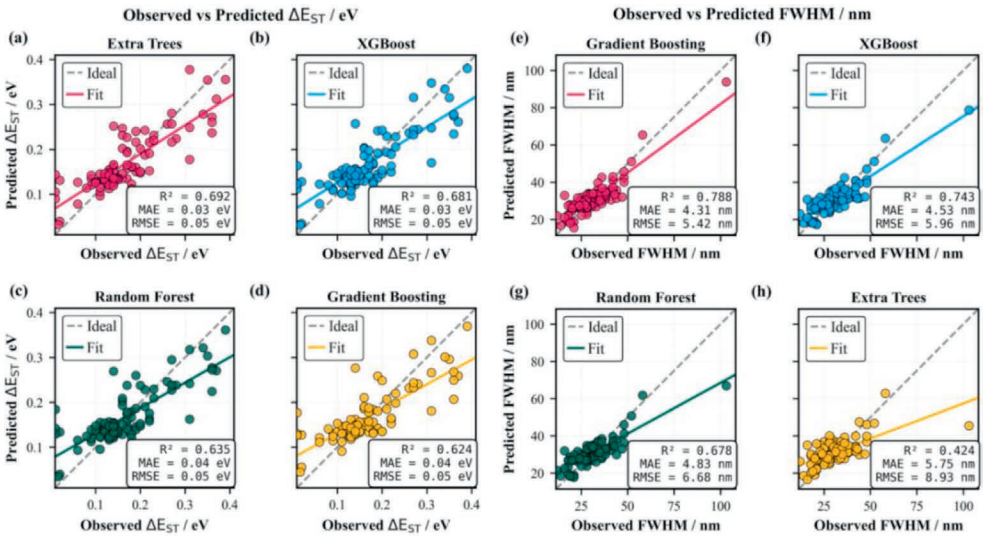


Fig. 8. Comparison of ML models for predicting ΔE_{ST} and FWHM. (a-d) Parity plots for ΔE_{ST} predictions. (e-h) Parity plots for FWHM predictions.

properties are jointly governed by both local structural motifs and global topological patterns. To more effectively capture these two levels of information, this study incorporated Morgan fingerprints (radius = 2, 2048 bits) as supplementary features to the previously selected core descriptors. This integration aimed to combine molecular topology with chemical descriptors, thereby constructing a more comprehensive and high-dimensional feature representation. Consequently, the ΔE_{ST}-related feature set comprised 13 selected descriptors together with the 2048-bit fingerprint encoding, while the FWHM-related set consisted of 8 descriptors combined with an equally sized fingerprint representation-forming a feature space that balances structural specificity and high-dimensional correlation under limited data conditions.

Based on this composite feature space, four types of tree-based regression models were constructed, namely Random Forest, Extra Trees, Gradient Boosting, and XGBoost, covering both parallel ensemble and sequential gradient-boosting paradigms [44–47]. During model training, BayesSearchCV was employed for automated hyperparameter optimization, using the tree-structured parzen estimator to build a surrogate performance model [48]. The search process was dynamically guided by the expected improvement criterion [49], enabling efficient exploration of the parameter space and global model performance optimization.

The predictive performance of the four regression models for both ΔE_{ST} and FWHM was systematically evaluated (Fig. 8). In the ΔE_{ST} prediction task, the Extra Trees model achieved the best performance, with an $R^2 = 0.692$, RMSE = 0.05 eV, and MAE = 0.03 eV (Fig. 8a), demonstrating excellent overall fitting accuracy. The Random Forest and XGBoost models also exhibited strong predictive capability, reaching $R^2 = 0.635$ and $R^2 = 0.681$, respectively (Figs. 8b and c). Although the Gradient Boosting model showed minor fitting deviations in certain sample ranges, it still achieved $R^2 = 0.624$ (Fig. 8d), confirming its robustness under limited-sample conditions.

For the FWHM prediction, the Gradient Boosting model outperformed all others, yielding $R^2 = 0.788$, RMSE = 5.42 nm, and MAE = 4.31 nm (Fig. 8e). The XGBoost model followed closely with $R^2 = 0.743$ (Fig. 8f), while the Random Forest and Extra Trees achieved $R^2 = 0.678$ and $R^2 = 0.424$, respectively (Figs. 8g and h). Overall, all models exhibited a strong positive linear correlation between predicted and experimental values. Notably, the gradient-based ensemble approaches (XGBoost and Gradient Boosting) produced regression slopes closer to the ideal prediction line, suggesting their superior capability in capturing nonlinear structure-property relationships and complex feature interactions.

To further elucidate the structural response mechanisms underlying the differences between ΔE_{ST} and FWHM (Fig. 9), SHAP analysis was conducted on the predictive models.

In the ΔE_{ST} model (Fig. 9a), the key influential features were mainly associated with the internal electronic distribution and molecular topology, with descriptors such as FP_1159, FP_106, and HallKierAlpha showing significant contributions. Among them, FP_1159 had the highest impact (Fig. 9c), corresponding to structural fragments primarily located around the B atom. As a relatively electropositive center, the B atom modulates the local electron density of neighboring atoms, thereby influencing the spatial separation of HOMO and LUMO orbitals and consequently regulating ΔE_{ST}. The HallKierAlpha descriptor reflects the rigidity and connectivity of the molecular skeleton-higher values indicate a more conjugated and delocalized structure, which helps fine-tune the energy gap magnitude.

In contrast, the FWHM model (Fig. 9b) exhibits a clear geometry-charge co-regulation pattern. Key descriptors include FpDensityMorgan3, BCUT2D_MWLOW, and FP_574. FpDensityMorgan3 captures the connection strength and bonding type among atoms, representing the coupling between electronic states and molecular geometry; its importance highlights that the interplay between vibrational modes and electronic transitions is a dominant factor influencing spectral broadening. BCUT2D_MWLOW, which links molecular weight with topological shape, explains how rigidity and polarity modulate emission bandwidth. FP_574, meanwhile, identifies the influence of local charge distribution on FWHM; the corresponding structural fragment (marked in blue as Bit #574 in Fig. 9c) is mainly located near N atoms, suggesting that the local polarity and charge environment in these regions play a key role in determining the extent of spectral broadening.

To evaluate the potential applicability of the 15 newly generated molecules in the MR-TADF domain, a comprehensive set of key parameters was systematically calculated. The evaluation framework encompassed both photophysical properties, such as ΔE_{ST} and FWHM, and structural descriptors, including the topological score (TopoScore), heteroatom density of the molecular backbone, and synthetic accessibility score (SAScore). Among these, ΔE_{ST} and FWHM are directly associated with emission efficiency and color purity; TopoScore and heteroatom density jointly reflect the complexity, stability, and orbital-tuning capacity of the conju-

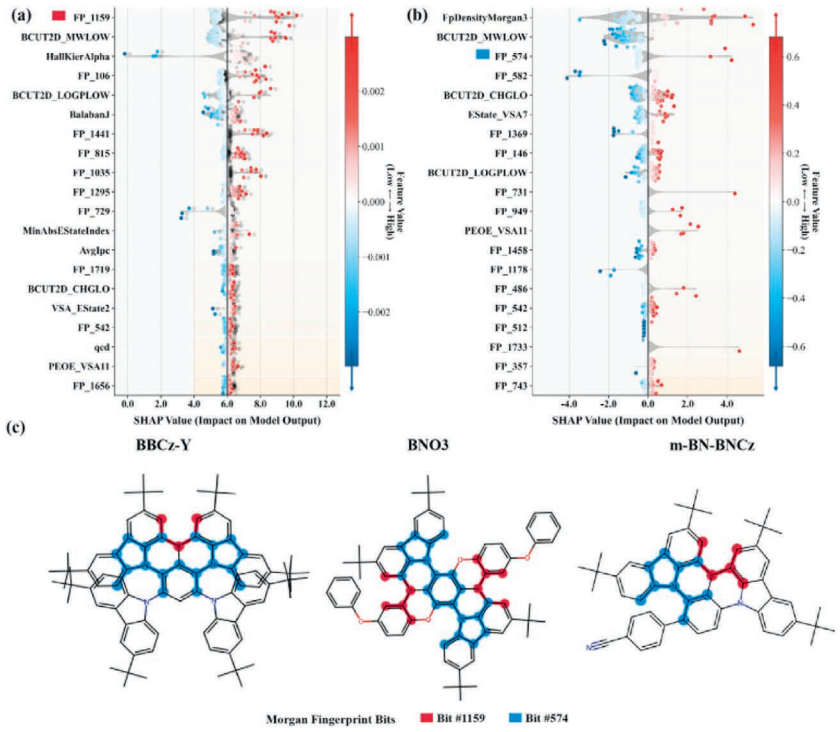


Fig. 9. SHAP analysis and visualization of key structural features in the ΔE_{ST} and FWHM predictive models. (a) SHAP feature contributions for the ΔE_{ST} model. (b) SHAP feature contributions for the FWHM model. (c) Spatial distribution of the most influential fingerprint bits within representative molecular structures (red: Bit #1159; blue: Bit #574).

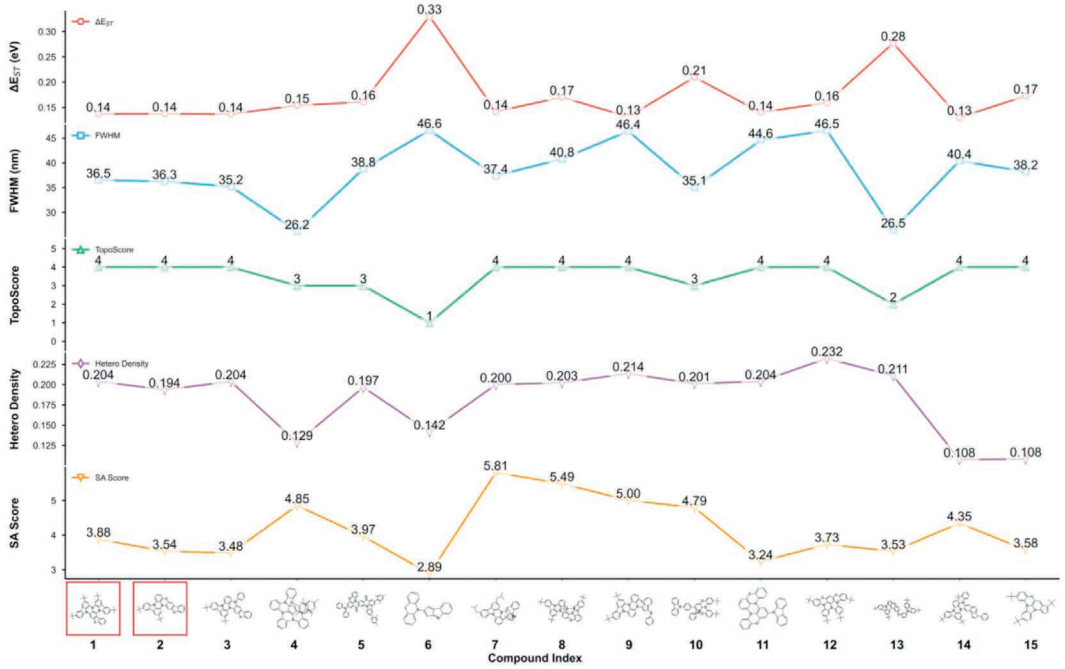


Fig. 10. Multi-attribute visualization of generated MR-TADF candidates. The five parallel subplots display the predicted ΔE_{ST}, FWHM, TopoScore, heteroatom density, and SAScore of the de-duplicated candidate set. Two lead candidates selected for theoretical validation are highlighted in red.

gated skeleton; while SAScore serves as an indicator of the feasibility of synthetic routes. This multidimensional evaluation system thus provides a quantitative basis for screening high-performance yet synthetically accessible MR-TADF candidates.

The multidimensional assessment results (Fig. 10) indicate that the 15 generated candidate molecules exhibit strong potential in both photophysical performance and structural characteristics. From a performance perspective, their ΔE_{ST} values fall within the range of 0.13–0.33 eV and the FWHM values span 26.2–46.6 nm, suggesting potentially high exciton utilization efficiency and good color purity. Structurally, the relatively high TopoScore values (typically > 3) reflect the rigid, polycyclic conjugated backbones characteristic of MR-TADF emitters, while the heteroatom density of the main skeleton (0.10–0.24) indicates a reasonable B/N doping



Fig. 11. Molecular structures of the investigated compounds.

ratio that is favorable for constructing balanced donor-acceptor electronic structures. In addition, the SAScore values range from 2.89 to 5.81, implying that most candidates retain acceptable synthetic feasibility while maintaining a certain degree of structural complexity.

To further strengthen the "generation-screening-application" closed-loop logic, we explicitly incorporated SAScore as a synthetic feasibility constraint in the screening process and stratified the 15 candidate molecules accordingly. Based on the distribution of SAScore, nine molecules with SAScore $\leq$ 4.0 (compounds **1**–**3**, **5**, **6**, **11**–**13**, and **15**) were prioritized for follow-up validation; four molecules with SAScore in the range of 4.0–5.0 (compounds **4**, **9**, **10**, and **14**) were considered as intermediate-priority candidates; and two molecules with SAScore $>$ 5.0 (compounds **7**, **8**), due to higher synthetic complexity, were assigned lower priority for short-term validation.

Based on this multi-objective evaluation, molecules **1** and **2** were selected as priority targets for in-depth quantum-chemical validation. On the one hand, both molecules rank highest in the overall multi-objective screening and thus represent high-potential candidates under the proposed strategy. More importantly, both have clear structural precedents in the existing MR-TADF literature, which reduces the structural uncertainty when progressing from generated candidates to verifiable molecules. Specifically, Molecule **1** shares a highly similar multi-resonance scaffold with reported emitters such as NBO, with the primary difference being an O $\rightarrow$ S substitution at a key site [50]. Molecule **2** retains the core structural motifs of representative systems such as tCzBT2B and Cz-BSN, while introducing targeted refinements in local fragments [51,52]. This generation pathway, element substitution and localized structural tuning around experimentally validated scaffolds, demonstrates that the model does not assemble structures arbitrarily, but instead proposes differentiated candidates in the vicinity of chemically feasible frameworks, providing a more reliable starting point for high-level theoretical validation. At the same time, given that high-accuracy methods such as SCS-CC2 are highly sensitive to molecular size and incur rapidly increasing computational costs, extending in-depth validation to all 15 candidates would be prohibitive; moreover, candidates with smaller molecular sizes did not exhibit superior predicted performance or empirical structural advantages over molecules **1** and **2**. Consequently, prioritizing molecules **1** and **2** represents a balanced compromise between scientific rigor and practical feasibility.

To verify the predicted performance of the generated and screened candidates, molecules **1** and **2** were subjected to comprehensive theoretical calculations. Both molecules are based on the classical MR-TADF skeleton **BCzBN** (Fig. 11) [53]. Compared with **BCzBN**, Molecule **1** replaces one nitrogen atom with sulfur and introduces a benzothiophene group fused to the upper phenyl ring, whereas Molecule **2** retains the **BCzBN** backbone but substitutes the central phenyl ring with a benzothiophene unit.

Frontier molecular orbitals analysis reveals that all investigated molecules exhibit a typical MR feature characterized by the spatially alternating distribution of HOMO and LUMO (Fig. 12), which effectively suppresses vibrational coupling in the excited state and enables narrowband emission [54,55]. Compared with **BCzBN**,

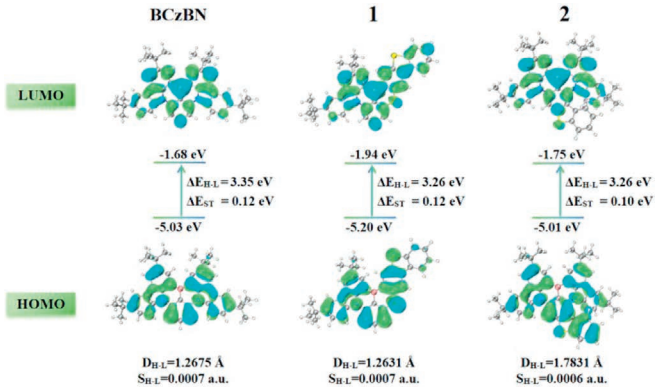


Fig. 12. HOMO and LUMO energy levels, $\Delta E_{H\text{-}L}$, ΔE_{ST}, and orbital distributions of the investigated molecules.

molecules **1** and **2** show a more delocalized FMOs distribution due to the expanded π-conjugated framework, leading to a reduced ΔE_{ST} and a corresponding red-shift in emission wavelength (Table S4 in Supporting information). Notably, the HOMO of molecule **2** extends entirely over the benzothiophene fragment, while its LUMO is localized near the sulfur atom. This asymmetric electron distribution enhances the intramolecular charge-transfer character, increasing the HOMO-LUMO separation distance ($D_{H\text{-}L}$) and decreasing their overlap ($S_{H\text{-}L}$), thereby facilitating a reduction in ΔE_{ST}.

Although MR-TADF materials exhibit great potential for high-resolution display applications owing to their intrinsically narrowband emission, their practical device performance is often constrained by efficiency roll-off. This limitation is mainly attributed to the relatively low reverse intersystem crossing rate (k_{RISC}): Under high current densities, triplet excitons tend to accumulate, thereby activating loss channels such as triplet–triplet annihilation. Fig. S2 (Supporting information) summarizes the key excited-state decay rates of all investigated molecules, while the corresponding detailed calculated results and parameters are compiled in Table 1 and Tables S5 and S6 (Supporting information). Compared with the parent scaffold **BCzBN**, the two molecules generated in this study (molecules **1** and **2**) exhibit markedly enhanced spin–orbit coupling (SOC) upon sulfur incorporation, benefiting from the heavy-atom effect and mechanistically favoring an increased k_{RISC}. When assessed using the figure of merit (FOM) proposed by Diesing and co-workers to quantify efficiency roll-off, molecules **1** and **2** display a substantially reduced roll-off tendency owing to their accelerated RISC process (Table 1).

Vibronically resolved emission spectra calculations (Figs. S3-S5 in Supporting information) show that the calculated FWHM of **BCzBN** (26 nm) agrees well with the experimental value (29 nm), confirming the reliability of the computational approach. Although the introduction of sulfur atoms induces slight conformational distortion, leading to moderately broader FWHMs for molecules **1** and **2** (30 and 37 nm, respectively), their emissions still fall within the narrowband regime, indicating strong potential for high-resolution display applications. Collectively, these findings demonstrate that the proposed molecular generation framework not only preserves high color purity but also provides a feasible strategy for producing high-performance MR-TADF materials capable of mitigating efficiency roll-off.

In conclusion, this study establishes an integrated framework encompassing data acquisition, molecular generation, and performance prediction for MR-TADF emitters, aiming to overcome the limitations of conventional quantum-chemical workflows, namely high computational cost, limited data availability, and restricted structural innovation. By leveraging the Qwen Max large lan-

Table 1
Calculated ΔE_{ST} (eV), SOC, k_r^S, k_{ISC}, k_{RISC} (s^{-1}), and FOM of the investigated molecules.

Sample	ΔE_{ST}	$<S_1\|HSOC\|T_1>$	k_{ISC}	k_{RISC}	k_r^S	FOM
BCzBN	0.12/0.13 [53]	0.07	7.61×10^5	2.16×10^3	1.37×10^8	2.86×10^3
1	0.12	0.15	3.51×10^6	1.04×10^4	8.14×10^7	1.32×10^4
2	0.10	0.20	7.56×10^6	4.41×10^4	7.46×10^7	5.18×10^4

guage model and prompt engineering, key photophysical parameters (ΔE_{ST} and FWHM) were automatically extracted and structured from the literature, while DECIMER enabled the conversion of molecular diagrams into SMILES representations.

In the molecular generation module, a Transformer-VAE-based latent space modeling approach was introduced. Combined with SMILES randomization for data augmentation, the model achieved a balance between structural validity and diversity. Under 10× augmentation, the model reached a validity rate of 59.8 % and the highest uniqueness rate of 4.01 %, generating a broader range of novel structures, consistent with the goal of maximizing molecular diversity.

For excited-state property prediction, a lightweight machine-learning screening mechanism based on 2D molecular descriptors was developed to independently predict ΔE_{ST} and FWHM. Among the compared models, Extra Trees achieved the best performance for ΔE_{ST} prediction ($R^2 = 0.692$), while XGBoost delivered the highest accuracy for FWHM ($R^2 = 0.788$). SHAP analysis further elucidated the structure–property relationships: ΔE_{ST} was mainly influenced by electronic state indices and topological complexity descriptors, whereas FWHM was more dependent on polar surface area, charge distribution, and local geometric factors, consistent with literature reports on FWHM's multifactorial dependence.

Among the 15 candidate molecules generated through this workflow, ΔE_{ST} values ranged from 0.13 eV to 0.33 eV and FWHM values from 26.2 nm to 46.6 nm, indicating promising exciton utilization and color purity, though further improvement is required to achieve ultra-low ΔE_{ST} ($\leq$0.10 eV) and narrower FWHM ($\leq$30–35 nm). Multi-parameter evaluation incorporating TopoScore, heteroatom density, and SAScore identified molecules **1** and **2** as the most promising candidates. Compared with the emitter **BCzBN**, both molecules-derived from the **BCzBN** backbone-exhibited significantly enhanced SOC and an order-of-magnitude increase in k_{RISC} due to the sulfur heavy-atom effect, leading to reduced efficiency roll-off and confirming the consistency between model generation and predicted performance.

Overall, this study develops an integrated data mining-generation-prediction-screening workflow that unifies literature curation, candidate discovery, and performance evaluation within a single framework, substantially improving the efficiency of MR-TADF candidate generation and screening while providing a data-driven basis for systematically summarizing structure-property relationships; with MR585 as the training foundation, the successful validation of molecules **1** and **2** further indicates that the model is not arbitrarily assembling molecular structures, but can capture physically meaningful patterns under the design paradigm of element substitution and local motif tuning and generate new structures beyond those present in the dataset with improved predicted performance. Nevertheless, the current framework remains constrained by the coverage of MR585: Some heteroatom–containing structures are still underrepresented. In particular, systems containing Se or P are relatively scarce and were not the primary focus of validation in this work, so generalization in these chemical spaces may be limited, particularly because such systems often introduce synthetic and stability considerations beyond photophysical metrics (*e.g.*, Se-containing molecules can be prone to oxidation, may involve weaker C-Se bonds, and can be highly sensitive to air, temperature, and reaction conditions), implying that candidates in these regions should be judged jointly with synthetic-feasibility and operational-stability constraints. In addition, ultra-narrowband targets with $\Delta E_{ST} \leq 0.10$ eV and FWHM $\leq$ 30 nm constitute only a small fraction of the present dataset, restricting both learning and rigorous evaluation in this extreme regime; moreover, these high-performance regions often correspond to highly optimized electronic structures that are extremely sensitive to subtle structural perturbations, placing higher demands on descriptor completeness and model resolution. Accordingly, our discussion is centered on the property ranges covered by the generated candidates ($\Delta E_{ST} = 0.13$–0.33 eV; FWHM = 26.2–46.6 nm), and future work will expand data coverage for underrepresented chemistries and extreme-property regimes, refine structural and excited-state descriptors, and further incorporate feasibility and reliability constraints into the generation and screening process to improve applicability across broader chemical spaces and enhance consistency between predicted properties and device-level performance.

Data & code availability

All data and code used in this study are publicly available. The curated and standardized MR-TADF molecular dataset (MR585), including photophysical property labels, molecular fingerprint features, and graph-structure features, is available at: https://github.com/zouly-group/MR585_Dataset. The literature sources underpinning the MR585 dataset compilation are provided in Table S7. The literature-mining and data extraction pipeline used to build MR585 is available at: https://github.com/zouly-group/TADF-Literature-Mining-Molecular-Property-Workbench.

CRediT authorship contribution statement

Ya-Jun Yin: Writing – review & editing, Writing – original draft, Visualization, Project administration, Data curation, Conceptualization. **Li-Fang Yin:** Writing – review & editing, Resources, Data curation. **Yi Zhao:** Writing – review & editing, Conceptualization. **Xin Xu:** Writing – review & editing. **Yu-Qi Xia:** Data curation. **Jing-Jing Zhao:** Data curation. **Jia-Qi Bai:** Data curation. **Guang-Jun Nan:** Supervision, Project administration, Methodology. **Ji-Fen Wang:** Conceptualization. **Lu-Yi Zou:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 22573039) and the International Science and Technology Cooperation Project of Jilin Provincial Department of Science and Technology (No. 20240402047GH).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cclet.2026.112633.

References

[1] A. Farokhi, S. Lipinski, L.M. Cavinato, et al., Chem. Soc. Rev. 54 (2025) 266–340.
[2] S. Diesing, L. Zhang, E. Zysman-Colman, I.D.W. Samuel, Nature 627 (2024) 747–753.
[3] T. He, Z. Zhang, L. Yin, et al., Chem. Res. Chin. Univ. 41 (2025) 1133–1143.
[4] J.G. Yang, X. Feng, N. Li, et al., Sci. Adv. 9 (2023) eadh0198.
[5] T.F. He, A.M. Ren, Y.N. Chen, et al., Inorg. Chem. 59 (2020) 12039–12053.
[6] M. Liu, B. Hou, Y. Li, Y. Pan, B. Yang, Comput. Theor. Chem. 1231 (2024) 114414.
[7] Y. Takeda, Acc. Chem. Res. 57 (2024) 2219–2232.
[8] T.F. He, A.M. Ren, G.H. Li, et al., J. Phys. Chem. Lett. 12 (2021) 2232–2244.
[9] Y.L. Zhang, T.F. He, Z.K. Zhao, et al., Inorg. Chem. 62 (2023) 7753–7763.
[10] J.M. Ha, S.H. Hur, A. Pathak, J.E. Jeong, H.Y. Woo, NPG. Asia Mater. 13 (2021) 53.
[11] Z.K. Zhao, T.F. He, Q. Gao, et al., Inorg. Chem. 63 (2024) 17435–17448.
[12] T. Hatakeyama, K. Shiren, K. Nakajima, et al., Adv. Mater. 28 (2016) 2777–2781.
[13] Z. Yang, S. Li, L. Hua, et al., Chem. Sci. 16 (2025) 3904–3915.
[14] M.X. Du, J.O. Zhou, X.F. Luo, L. Duan, Moore More 1 (2025) 79–98.
[15] K.R. Naveen, H.I. Yang, J.H. Kwon, Commun. Chem. 5 (2022) 149.
[16] G. Ricci, A. Landi, J.C. Sancho-García, Y. Olivier, Adv. Opt. Mater. 13 (2025) 2402765.
[17] B.H. Jhun, Y. Park, H.S. Kim, et al., Nat. Commun. 16 (2025) 392.
[18] J.M. Jin, C. Shi, W.C. Chen, Y. Huo, Chem. Commun. 61 (2025) 10731–10746.
[19] K. Shizu, H. Kaji, Commun. Chem. 5 (2022) 53.
[20] D. Hall, J.C. Sancho-García, A. Pershin, et al., J. Phys. Chem. A 127 (2023) 4743–4757.
[21] J.M. Dos Santos, D. Hall, B. Basumatary, et al., Chem. Rev. 124 (2024) 13736–14110.
[22] A. Wu, Q. Ye, X. Zhuang, et al., Precis. Chem. 1 (2023) 57–68.
[23] Y. Chen, W. Yan, Z. Wang, J. Wu, X. Xu, J. Chem. Theory. Comput. 20 (2024) 9500–9511.
[24] Z. Wang, W. Zhang, M. Jiang, et al., J. Phys. Chem. Lett. 15 (2024) 12501–12512.
[25] H.S. Kim, H.J. Cheon, S.H. Lee, et al., Sci. Adv. 11 (2025) eadr1326.
[26] H. Shi, Y. Shi, Z. Liang, et al., Chem. Engin. J. 494 (2024) 153150.
[27] Z. Chen, J. Song, L. Hu, et al., Chin. Chem. Lett. 37 (2026) 111967.
[28] Y. Yin, L. Yin, Y. Zhao, et al., Chem. Res. Chin. Univ. 41 (2025) 1173–1185.
[29] S.Y. Zhang, Y. Zhao, L. Wang, Z.P. Liu, Y. Tang, CCS Chem. 7 (2025) 3493–3506.
[30] Y. Li, B. Zhang, A. Ren, et al., Chem. Engin. J. 501 (2024) 157676.
[31] D. Huang, J.M. Cole, Sci. Data 11 (2024) 80.
[32] J. Arús-Pous, S.V. Johansson, O. Prykhodko, et al., J. Cheminform. 11 (2019) 71.
[33] D. Weininger, J. Chem. Inf. Comput. Sci. 28 (1988) 31–36.
[34] K. Rajan, A. Zielesny, C. Steinbeck, J. Cheminformatics 12 (2020) 65.
[35] D. Rogers, M. Hahn, J. Chem. Inf. Model. 50 (2010) 742–754.
[36] H. Shi, Y. Shi, G. Zeng, et al., Chem. Engin. J. 525 (2025) 170146.
[37] F. Chen, L. Zhao, X. Wang, et al., Sci. China Chem. 64 (2021) 547–551.
[38] A. Vaswani, N. Shazeer, N. Parmar, et al., in: Proceedings of the 31st international conference on neural information processing systems, Curran Associates Inc., New York, USA, 2017, pp. 6000–6010.
[39] R. Gómez-Bombarelli, J.N. Wei, D. Duvenaud, et al., ACS Cent. Sci. 4 (2018) 268–276.
[40] M.H.S. Segler, T. Kogej, C. Tyrchan, M.P. Waller, ACS Cent. Sci. 4 (2018) 120–131.
[41] D.P. Kingma, J. Ba, in: Proceedings of 3rd international conference on learning representations (ICLR 2015), San Diego, CA, USA, 2015.
[42] S. Jasial, Y. Hu, M. Vogt, J. Bajorath, F1000Res. 5 (2016) 591.
[43] W.S. Torgerson, Psychometrika 17 (1952) 401–419.
[44] L. Breiman, Mach. Learn. 45 (2001) 5–32.
[45] T. Chen, C. Guestrin, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, Association for Computing Machinery, New York, USA, 2016, pp. 785–794.
[46] J.H. Friedman, Ann. Stat. 29 (2001) 1189–1232.
[47] P. Geurts, D. Ernst, L. Wehenkel, Mach. Learn. 63 (2006) 3–42.
[48] J. Bergstra, R. Bardenet, Y. Bengio, B. Kégl, Advances in Neural Information Processing Systems, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, K.Q. Weinberger (Eds.) Eds, Curran Associates, Inc., 2011.
[49] D.R. Jones, M. Schonlau, W.J. Welch, J. Glob. Optim. 13 (1998) 455–492.
[50] X.F. Luo, H.X. Ni, H.L. Ma, et al., Adv. Opt. Mater. 10 (2022) 2102513.
[51] Q. Li, Y. Wu, X. Wang, et al., Chem. Eur. J. 28 (2022) e202104214.
[52] C. Jiang, Y. Nie, C. Cao, et al., Adv. Sci. 13 (2026) e12796.
[53] Z. Huang, H. Xie, J. Miao, et al., J. Am. Chem. Soc. 145 (2023) 12550–12560.
[54] T.F. He, Y.N. Chen, X.L. Hao, et al., J. Phys. Chem. Lett. 17 (2026) 205–213.
[55] L. Yin, Y. Zhao, Q. Gao, et al., Chem. Sci. 17 (2026) 2599–2615.