JACS Au
AN OPEN ACCESS JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

Open Access
This article is licensed under CC-BY 4.0

pubs.acs.org/jacsau
Article

Exploring Optimized Organic Fluorophore Search through Experimental Data-Driven Adaptive β-VAE

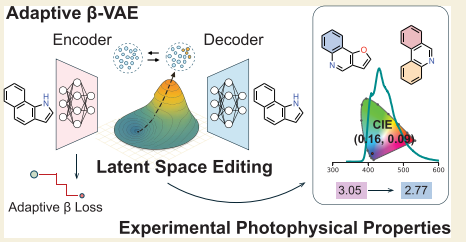
Yuzhi Xu, Yongrui Luo, Bo Li, Weikang Jiang, Jinyu Zhang, Jiangbo Wei, Hanzhi Bai, Zhiqiang Wang, Jiankai Ge, Ruiming Lin, Zehan Mi, Haozhe Zhang, Yifeng Tang, Michael S. Jones, Xiaotian Li, John Z.H. Zhang, and Cheng-Wei Ju*

Cite This: JACS Au 2025, 5, 3082−3091

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Designing organic fluorescent molecules with tailored optical properties has been a long-standing challenge. Recently, statistical models have opened new avenues for tackling this problem. Inverse design has attracted considerable attention in organic materials science; however, most existing approaches focus on arbitrary design or theoretical properties. Here, we introduce a strategy that enables the direct optimization of specific experimental properties during the inverse design process. Our method employs an adaptive β-variational autoencoder (adaptive β-VAE) combined with a latent vector-based prediction model. By dynamically tuning the Kullback−Leibler divergence scaling factor (β) and employing a separate training strategy, we enhance both the robustness of the generator and the diversity of the generated molecules. We demonstrate that latent vectors from the adaptive β-VAE serve as powerful inputs for downstream prediction models of experimental properties, such as fluorescence energy and quantum yield. Our optimized search framework for organic fluorescent materials—guided by gradients in latent space and validated by newly synthesized molecules sampled from optimal regions in the high-dimensional space—shows strong potential for broader applications in the design of diverse organic materials.



KEYWORDS: *molecular modeling, optimization, inverse molecular design, molecules, optical properties, fluorescence*

Downloaded via 99.7.2.48 on September 9, 2025 at 19:20:46 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

INTRODUCTION

The design of small-molecule organic fluorophores has become a central focus in biological research and material science due to the advent of fluorescence-based applications.[1−4] Despite this interest, the controlled synthesis of fluorophores remains challenging because of the intricate relationship between structure and properties.[5−7] Traditional first-principles calculations offer a partial solution; however, they often fail to balance computational speed with accuracy and can only work on limited properties.[8−10] Recent advances in machine learning (ML) have provided alternative pathways for predicting the optical properties of organic materials (Figure 1A).[11−17] For instance, the ChemFluor data set reported by us served as the basis for our reported ML model for photophysical property prediction.[11] Similarly, Joung et al. utilized a deep learning framework to predict a range of optical properties.[12]

The success of statistical models raises the possibility of inverse design and the targeted search for optimized compounds (Figure S4).[18−21] The challenge of inverse design with predictive models for organic materials comes from the reliance on molecular descriptors, which translate molecular structures into machine-readable formats.[22,23] This translation is unidirectional, preventing the reconstruction of molecular architectures from descriptors alone, thus limiting the scope for reverse engineering. Graph neural networks (GNNs) have shown promise in both predictive modeling and, more recently, inverse design. Nonetheless, due to their limited receptive fields and higher data requirements, fingerprint-based models remain advantageous for capturing global molecular features and enabling data-efficient training on experimentally derived data sets.[24,25] Additionally, the discrete nature of these variables (such as molecular fingerprints) complicates the computation of gradients during optimization, posing a barrier to the seamless application of conventional optimization techniques.[26,27] In response to these challenges, various generator architectures have garnered substantial interest.[28−30] Early work by Aspuru-Guzik et al. on a SMILES-based variational autoencoder (VAE) opened avenues for optimized compound searches, albeit limited to small molecules.[31,32] Moreover, the generator has been explored in ML-assisted material design as well but concentrates either on arbitrary design or theoretical properties.[33,34]

Here, we questioned whether the search for optimized compounds with specific experimental properties in materials

Received: January 15, 2025
Revised: May 19, 2025
Accepted: May 20, 2025
Published: June 30, 2025

© 2025 The Authors. Published by American Chemical Society

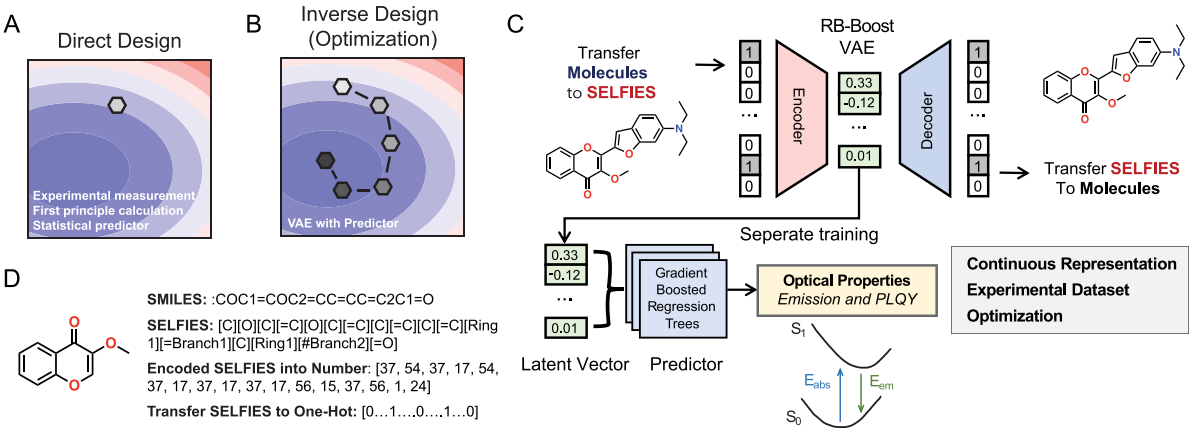


Figure 1. Schematic overview of organic material optimization. (A) The schematic for direct design. (B) The schematic for inverse design by searching for targeted molecules. (C) Overview of the methodology for optimized materials search developed in this study. (D) Diagram for various molecular representations.

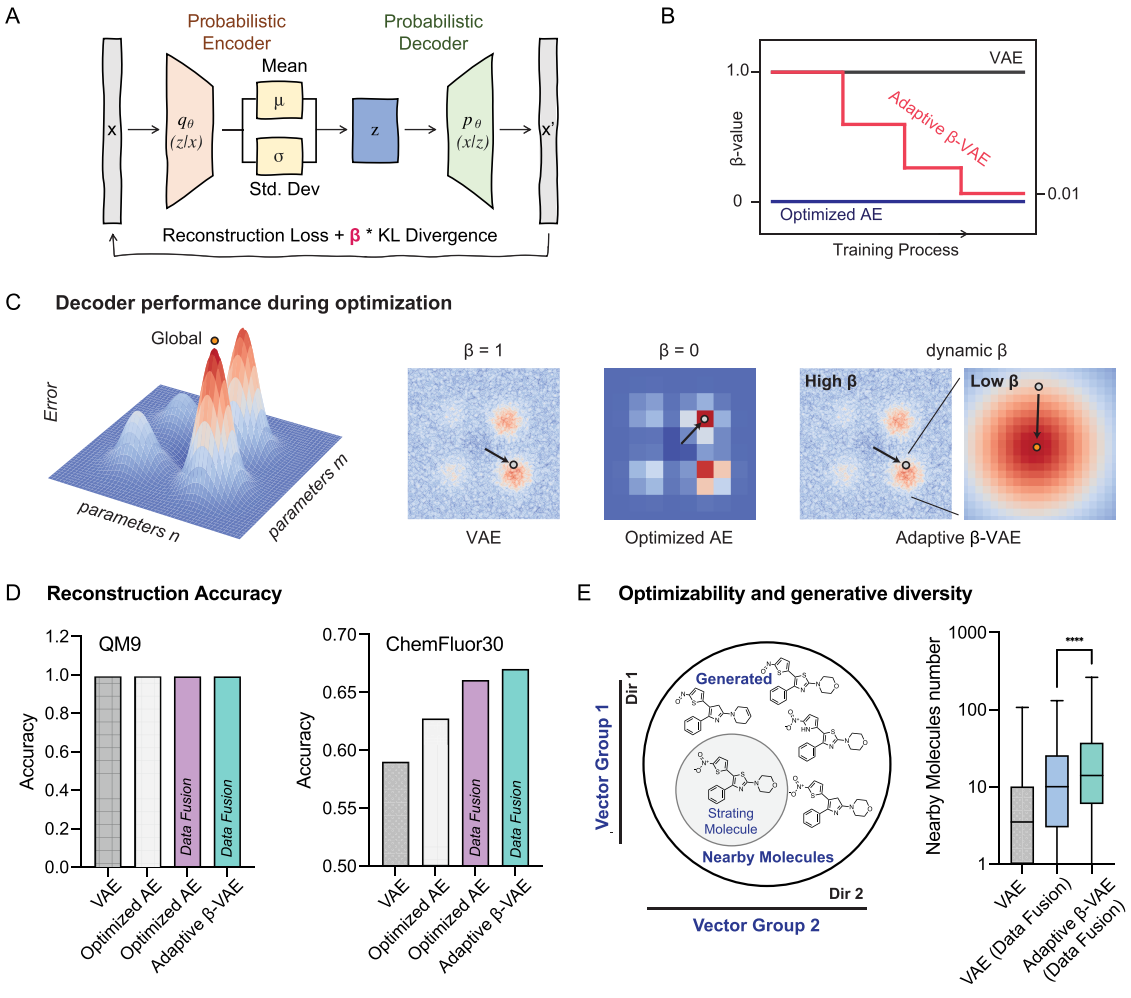


Figure 2. Enhancing molecular diversity and accuracy through adaptive β-VAE. (A) Schematic for the VAE and β-VAE. (B) The change of β value during the training process in VAE, optimized AE, and adaptive β-VAE. (C) Conceptual illustration of the optimization process under different β settings. (Left) Error landscape with respect to hyperparameters n and m. The goal of tuning the β value is to guide the model toward the global optimum (marked in orange). (Middle) Latent space representations under different KL divergence weights: standard VAE (β = 1) encourages structured diversity but may lead to noisy representations and convergence to a local minimum; AE (β = 0) drives the model toward a global minimum in terms of reconstruction error but sacrifices diversity and may ignore the true global optimum. Adaptive β-VAE transitions from global exploration (high β) to localized accuracy (low β), effectively balancing both objectives. Heatmaps represent reconstruction error. Arrows indicate training trajectories. (D) Evaluation of the reconstruction accuracy of different AE variants across various data sets. (E) Adaptive β-VAE enhances diversity in molecular generation. The box plot shows the capability of adaptive β-VAE in generating viable molecules from latent vector interpolations, while only molecules generated by disturbing a randomly selected and consistent subset of vectors are used for the counts.

https://doi.org/10.1021/jacsau.5c00052
JACS Au 2025, 5, 3082−3091

science can also be achieved through an integrated generator-predictor framework (Figure 1B). This approach, however, presents several challenges that impede large-scale exploration. Primarily, the combination of generation and prediction tools has predominantly focused on properties derived from quantum chemical computations due to the limited scarcity of experimental data sets.[35−37] The limited size of experimental data sets will compromise the generator's efficacy. Additionally, this integration typically necessitates cotraining of the decoder and predictor.[32] Lastly, predicting experimental properties—such as fluorescence wavelengths, photoluminescence quantum yield (PLQY) in organic fluorophores, power conversion efficiencies (PCEs) in organic photovoltaics (OPVs), and charge carrier mobility in organic field-effect transistors (OFETs)—proves substantially more difficult than computational attributes due to the multifaceted influences in real-world experimental conditions.

To answer these questions, we developed a workflow leveraging an adaptive β-VAE and a predictor to directly optimize organic fluorophores on a high dimensional space fitted from experimental energies (Figure 1C). SELFIES, a robust and standardizable molecular string representation, was utilized for reliable encoding in a one-hot format (Figure 1D and Method S1.1.1). The application of this encoding is to reduce the model's dependency on learning syntax alongside molecular structure, thereby minimizing syntax-related errors during generation. We train the generator and predictor separately and thus make the data fusion in the generator become possible. Dynamic tuning of the scaling factor of the Kullback−Leibler divergence (KL divergence), β, which regulated the strength of the regularization, can generate a more flexible latent space representation and improve the decoder's reconstruction ability. Utilizing the latent vectors from this adaptive β-VAE, we constructed a prediction model for the photophysical properties, including PLQY and emission energy within the error of quantum mechanical precision (~0.13 eV). Then, we visualize the high-dimensional space to confirm the possibility of target molecular optimization. Experimental validation with newly synthesized molecules sampled from optimal regions of high-dimensional space successfully confirms the feasibility of our generator and predictor. Applying our method in a fluorophore skeleton, we synthesized a new compound with bright blue emission, showcasing our strategy's potential for material discovery. Our workflow proves the feasibility of inverse design achieved through target optimization and signals a transformative approach to diverse organic material design.

RESULTS AND DISCUSSION

Adaptive β-VAE for Molecular Reconstruction

Traditional autoencoders (AE) focus on compressing and reconstructing data but lack control over the latent space, limiting their usefulness for generating diverse molecular structures. VAE, on the other hand, provides structured latent spaces that are ideal for molecular generation by introducing a KL divergence term. However, this structure can sometimes overconstrain the model, reducing reconstruction efficiency. To address this, β-VAE was introduced, adding a scaling factor, β, before the KL divergence term (Figure 2A). Adjusting β provides more flexibility: lower β values reduce the influence of KL divergence, allowing for higher reconstruction accuracy, while higher β values increase the regularization effect, encouraging diverse generation. β-VAE allows for control over the balance between reconstruction accuracy and diversity. However, a fixed β may still be suboptimal, as different stages of training demand varying levels of regularization.

To tackle this challenge, we propose a β-VAE variant, named adaptive β-VAE. Our approach uses a dynamic β that changes over the course of training: we start with a high β to build a globally diverse latent space and then gradually decrease β, allowing the model to emphasize local reconstruction accuracy (Figure 2B, see Methods S1.1.2 and S1.1.3 for details). This adaptive strategy enables the model to explore a broader range of possibilities at early stages and converge to accurate solutions later, striking a balance between AE's reconstruction focus and VAE's generative flexibility (Figure 2C). This approach is particularly effective for specific data types, such as chemical small molecules, where both local detail and global diversity are essential.

Cotraining prediction and generative models have been common in molecular generation; however, this approach is limited by experimental data sets, which are often small in size, restricting the model's generative diversity. To overcome this, we opted for separate training of the generator and predictor, allowing each model to specialize without data set limitations. Additionally, we incorporated diverse molecular scaffolds through data fusion, enriching our data set with compatible molecules per established protocols (Method S1.1.4). This strategy broadens latent space sampling and enhances generative diversity, addressing real-world challenges in molecular generation.

We first validated our model on the QM9 data set with three VAE variants (β = 1 VAE; β = 0, optimized AE; and adaptive β-VAE), achieving a high reconstruction rate (>98%) across all variants (Figure 2D and Table S1). When our strategy was applied to a more challenging data set with larger fluorescent molecules, ChemFluor30 (a subdata set of ChemFluor with molecules smaller than 30 heavy atoms), the adaptive β-VAE with data fusion showed a clear performance improvement, raising reconstruction rates from 59% to 67%, a relative increase of ~13% (Figure 2D). Ablation experiment confirmed the impotence of scheduling strategy for the β value in adaptive β-VAE (Table S2). This dual strategy not only increase the reconstruction accuracy but also enhanced the representation of diverse molecular characteristics.

We then evaluated the enhanced adaptive β-VAE and the VAE by perturbing a subset of latent vectors to generate molecules (Figures 2E and S5). The adaptive β-VAE demonstrated superior performance, generating an average of 8.2 times more total distinct molecules with a broader chemical feature set, indicative of a more complex chemical space encapsulated during model training (Figure 2E). This contrasted with the original VAE, which tends to generate more similar structures. Moreover, the adaptive β-VAE facilitated the generation of transitional molecular structures through interpolation between two selected latent vectors (Figures 2E and S6). Despite some resulting nonviable molecules, the majority of these intermediate structures were coherent and synthesizable, emphasizing the strength of our strategy in refining the VAE architecture to generate a wide range of diverse molecules.

https://doi.org/10.1021/jacsau.5c00052
JACS Au 2025, 5, 3082−3091

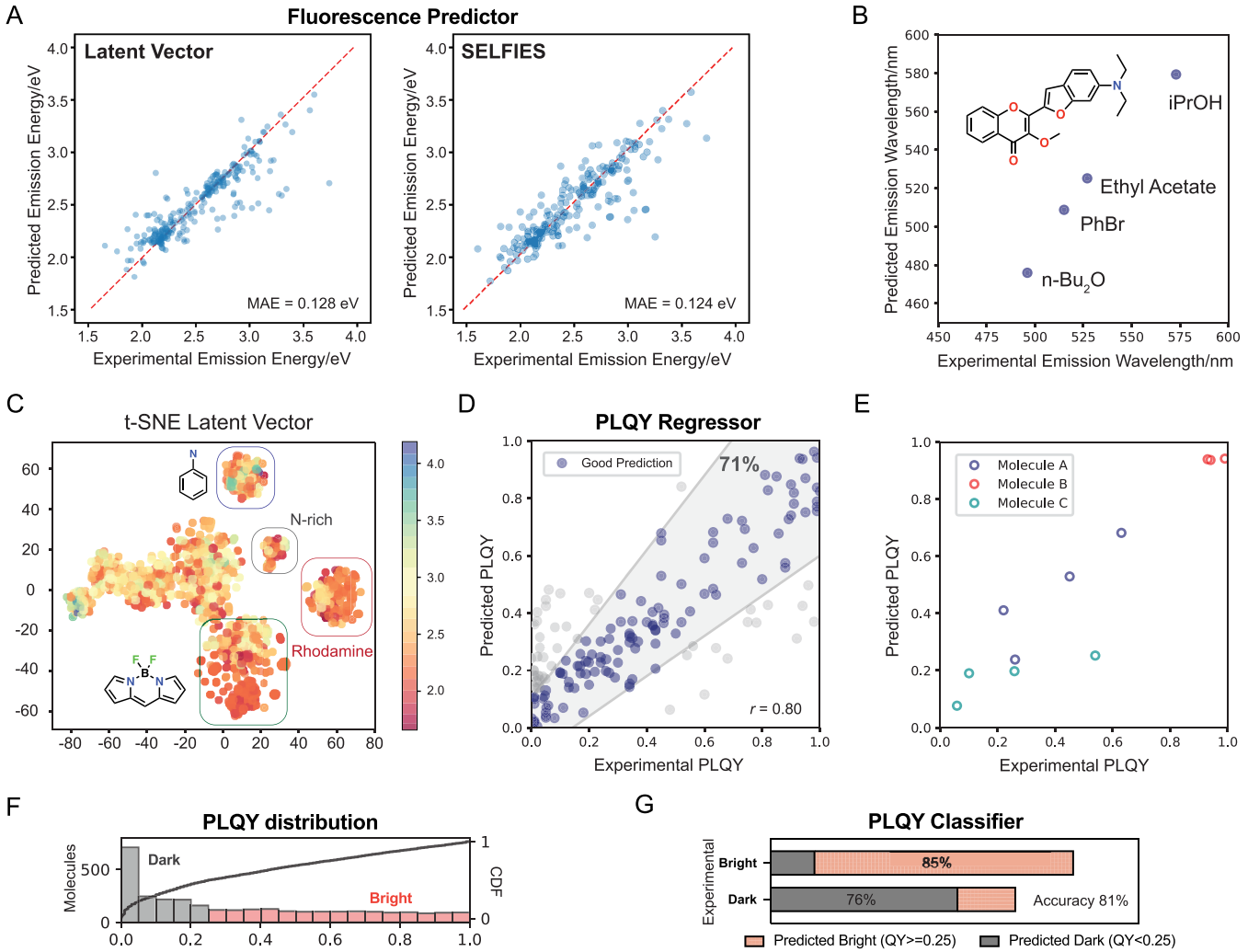


Figure 3. Latent vector-based predictor for optical properties. (A) The predictor for the emission energy based on GBRT with latent vector (left) or SELFIES (right) well reproduces the fluorescence in the test set for unseen molecules with different solvents. (B) The fluorescence wavelength of one typical molecule in the test set in various solvents. (C) t-SNE of latent vectors obtained from adaptive β-VAE. Colors indicate the emission energies (eV). (D) The regressor for the PLQY based on GBRT with the latent vector as input can reproduce the brightness for unseen molecules. (E) PLQY for selected molecules in the test set under various solvents. (F) The distribution of PLQY in the data set. 0.25 is set as the threshold for bright and dark molecules. (G) The prediction performance of PLQY classifier with the latent vector as input.

Predictor Based on GBRT with the Latent Vector for Experimental Optical Properties

With the establishment of the generator, we move to the prediction model. To adapt our VAE for chemical property prediction, we train the predictor separately using the latent space learned from the *ChemFluor30* data set, which contains experimentally measured photophysical properties. This approach diverges from the conventional joint training approach, which often restricts chemical diversity.[32] Our investigation prioritizes emission energies—key optical properties for organic emitters. We adopt the Gradient Boosting Regression Tree (GBRT), lauded for its predictive precision in our prior research (Method S1.1.5 and Tables S3 and S4). The model results in a mean absolute error (MAE) of 0.128 eV for unseen molecules in different solvents using latent vectors as the input, surpassing TD-DFT accuracy (~0.20 eV), and is sufficient for utilizing in virtual screening (Figure 3A and Table S5)[38−42] A similar MAE of 0.124 eV was obtained from one-hot SELFIES as input indicating the high fidelity of the latent vector generated from SELFIES. Furthermore, the model successfully reproduces the trends of specific molecules in different solvents (Figures 3B and S7). To externally validate the model, we test it on a data set of NDI, Rhodamine, and Coumarin molecules previously used as benchmarks. The model maintains a strong performance with an MAE of 0.20 eV, comparable to TD-DFT accuracy (Figure S8). Utilizing T-distributed stochastic neighbor embedding (t-SNE) visualizations, we observe the cluster of various structures such as Rhodamine and BODIPY derivatives (Figure 3C). Meanwhile, the analogous distributions between latent vectors and SELFIES prove that they are high-fidelity predictors, while the distinct from ECFP4 suggests their uniqueness (Figure S9). Furthermore, based on the predictor, we confirm that the molecules generated by adaptive β-VAE exhibit a greater diversity in their predicted emission energies (Figure S10).

Furthermore, we also assessed PLQY predictions within the latent space. PLQY is one of the most critical factors affecting the fluorescence intensity of organic fluorescent materials, yet attempts at its prediction remain limited. Our regressor achieves reasonable accuracy for unseen molecules across

https://doi.org/10.1021/jacsau.5c00052
JACS Au 2025, 5, 3082−3091

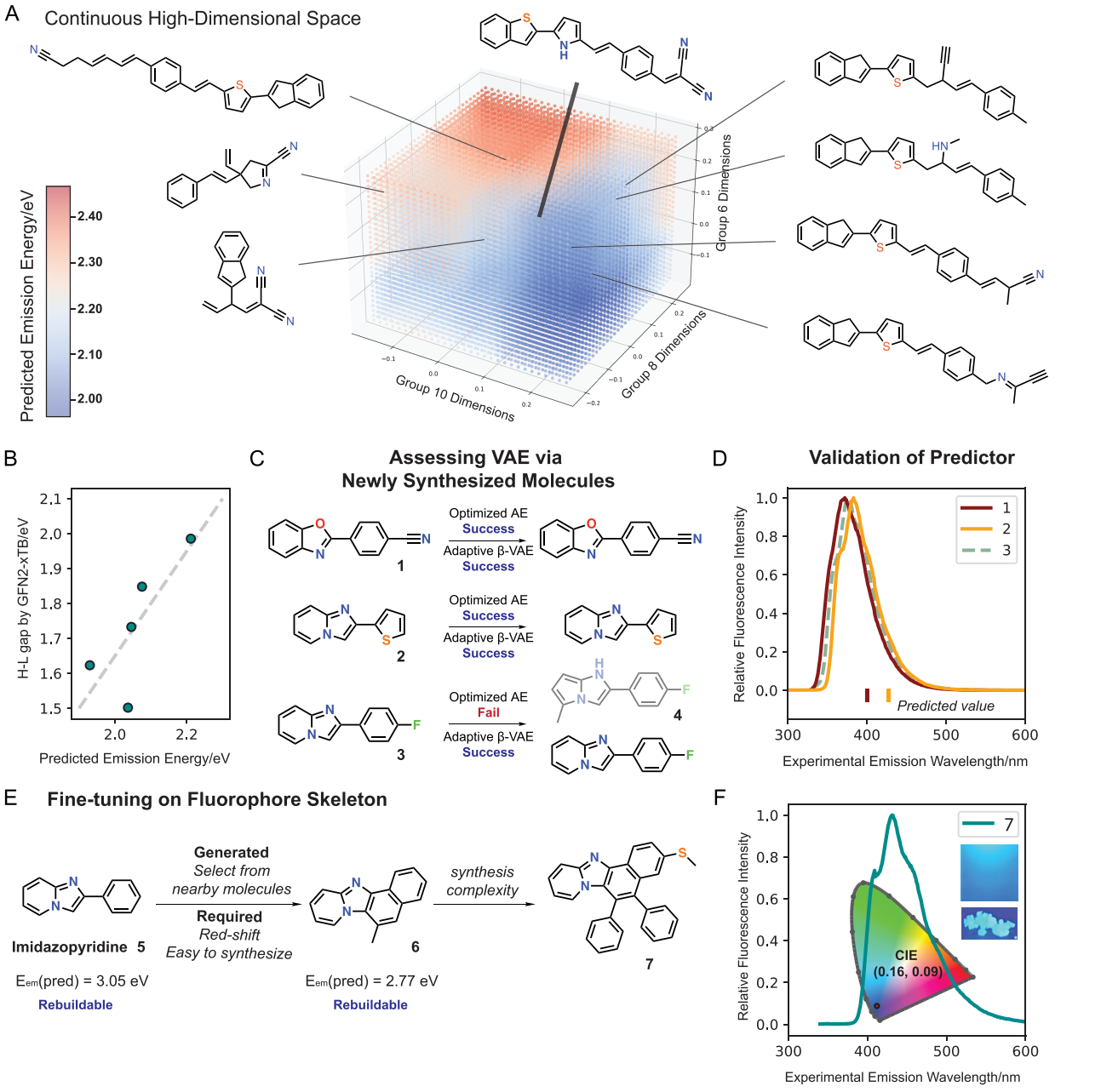


Figure 4. High-dimensional latent space analysis and synthesis validation. (A) Visualization and analysis of the continuous high-dimensional space, indicating potential for optimization. (B) Correlation between the HOMO−LUMO gap calculated by GFN2-xTB and the predicted emission energy of molecules with similar backbone sampled from high-dimensional space. (C) External validation of RB-Boost VAE using uncharacterized synthesized molecules. (D) Comparison of experimental fluorescence spectra with predicted emission energies for uncharacterized molecules, illustrating prediction accuracy. (E) Editing on the fluorophore skeleton (imidazopyridine) by exploring nearby molecules and controlling synthesis complexity. (F) The fluorescence spectrum of molecule **7**.

various solvents (r = 0.80, Figure 3D), making it suitable for prescreening fluorophore candidates. To study the prediction accuracy for real-world problems, we define an accurate PLQY prediction if the absolute error is less than 30% of the true value plus 0.1, an empirically chosen threshold that reflects typical practical tolerance and experimental uncertainty.[ref] Over 70% of unseen molecules can be accurately predicted, outperforming TD-DFT-based estimations.[43,44] It is also important to note that TD-DFT cannot easily or broadly estimate the PLQY. Only a few studies have attempted such calculations under specific physical assumptions, and these approaches are limited to a range of molecular systems. Additionally, our model can also well reproduce solvent effects (Figure 3E). Considering the distribution of PLQY and real-world situations, we apply 0.25 as a threshold to classify the bright and dark molecules (Figure 3F).[45−47] Our classifier discerns between bright and dark materials with an accuracy of 0.81, rendering it suitable for practical predictive applications (Figure 3G).

Synthesis Validation of the Framework

Based on the demonstrated performance of our generator and predictor, we have utilized vector group tuning to visualize the high-dimensional space in a 3D plot, facilitating precise structural adjustment and exploration (Methods S1.1.6 and

https://doi.org/10.1021/jacsau.5c00052
JACS Au 2025, 5, 3082−3091

S1.1.7). We applied our approach with molecules shown in the center of Figure 4A, where the manipulation of latent vectors yielded diverse molecules with predicted emission energies ranging from 1.95 to 2.45 eV (Figure 4A for model based on adaptive β-VAE and Figure S11 for optimized AE). To validate the reliability of the predicted fluorescence energy in the generated high-dimensional space, we employed Semiempirical Tight Binding, GFN2-xTB, a semiempirical quantum mechanical method to estimate the HOMO−LUMO gap of several molecules with a similar skeleton generated in this high-dimensional space (Figure S12).[48] Molecules with similar skeletons are selected here for the computational validation since we want to minimize the structure diversity that increases the complexity and difference between computational and experimental properties. The correlation further supports the validity of our approach (Figure 4B). Although it needs to be recognized that (1) semiempirical methods are not accurate and (2) calculated H-L gap only reflects the electronic structure in the ground state while emission is highly related to the excited state, we rationalize that molecules with a similar skeleton should at least have similar trend between H-L gap and fluorescence wavelength. This localized optimization highlights our approach's potential in editing molecular structures and properties, confirming its utility in precision design.

To further corroborate our strategy's efficacy, we synthesized and analyzed novel molecules. Initially, derivatives of benzoxazole and imidazopyridine (**1**−**3**) were analyzed by using the optimized AE and adaptive β-VAE. Compounds **1** and **2** were successfully reproduced by both methods. However, compound **3** underwent a transformation to 5-methyl-1*H*-pyrrolo[1,2-*a*]imidazole **4** with the optimized AE but was accurately reproduced by the adaptive β-VAE (Figure 4C). Later, to evaluate the performance of the predictor, we characterized their fluorescence spectra in CH_2Cl_2 (Figure 4D). Although the absolute error is around 0.20 eV, the model accurately reflected the emission trend for **1** and **2**, which possess a similar biaryl backbone. Following this initial validation of the generator and predictor, we investigated the utility of our strategy in optimized compound searches and molecular editing. Due to the complexity introduced by the high-dimensional latent space, we centered our exploration on the nearby molecules of imidazopyridine derivative **5** (Figure 4E). We choose molecule **6** with an extended π-system, for its plausible structure and predicted red-shifted emission compared with **5** (3.05 eV to 2.77 eV). Considering synthetic feasibility and our laboratory's compound library, we synthesized **7** based on the backbone of **6**. The photophysical characterization of **7** revealed its bright blue emission with a CIE coordinate (0.16, 0.09), indicating its potential as a blue OLED emitter (Figure 4F).[49]

CONCLUSIONS

In summary, we successfully leveraged the latent vector space to enable optimized molecule generation with experimentally relevant properties through a combination of adaptive β-VAE and a predictor. Specifically, we applied adaptive β-VAE, which employs a dynamically tuned scaling factor, β, for KL divergence to regulate the strength of regularization. This tuning of β enabled a flexible latent space representation, enhancing both the reconstruction accuracy and molecular diversity. Unlike traditional workflows, our predictor actively informed the selection of latent vectors, optimizing the search for "dream molecules" with tailored properties. We confirmed the practicality of our method in searching for optimized compounds by (1) the evaluation of the predictor performance, (2) visualization of the latent space with predicted emission energy validated by semiempirical quantum mechanical methods, and (3) experimental validation of synthesized molecules. Using a fluorophore skeleton as an example, we designed and synthesized compound **7**, which exhibited bright blue emission, demonstrating the feasibility and potential of our strategy in materials discovery.

This streamlined workflow not only enables editing of molecular properties for optimized compounds but also heralds a new era of material design with promising applications in the development of OLEDs, OPVs, and OFETs. Despite its success, the current approach has limitations, including the reliance on relatively small experimental data sets and the need for improved predictors for complex experimental properties. Furthermore, systematically benchmarking AI-driven reverse design against heuristic-driven expert strategies would be valuable for understanding the full potential of these data-driven approaches. Future work will address these challenges by expanding experimental data sets, integrating diffusion model with advanced neural network predictors, and exploring multitask learning frameworks.[50−55] These efforts will further enhance the robustness, accuracy, and versatility of AI-driven molecular design, paving the way for transformative applications across materials science and biotechnology.

METHODS

Variational Autoencoder

The VAE, developed by Diederik P. Kingma and Max Welling, reframes statistical inference issues as optimization problems.[56] In a VAE, the input data is sampled from a parametrized distribution, and the encoder and decoder are trained together to minimize the reconstruction error between the parametric and true posterior distributions.

When the model receives input x, the encoder compresses it into the latent space. The decoder then takes information sampled from this space to produce an output x as similar as possible to x. However, rather than encoding an input as a single point in the latent space, the VAE represents it as a distribution over this space. Thus, the encoder returns a distribution over the latent space instead of a single point. A regularization term is added to the loss function over this distribution to ensure a well-organized latent space conducive to the generative process.

The VAE's primary mechanism involves maximizing the evidence lower bound (ELBO). The ELBO is formulated as follows

$$\mathrm{ELBO} = E_{q_\phi(z|x)}[\log p_\theta(x|z)] - \mathrm{KL}[q_\phi(z|x)|p(z)]$$

Here, $q_\phi(z|x)$ represents the approximate posterior distribution of the latent space learned by the encoder, $p_\theta(x|z)$ is the conditional probability distribution of the data generated by the decoder, $p(z)$ is the prior distribution of the latent space, and KL denotes the Kullback−Leibler divergence, measuring the divergence between two distributions. By maximizing the ELBO, the VAE aims to improve the quality of data reconstruction while maintaining an effective organizational structure in the latent space. We use the framework of VAE, as shown in Figure S1.

Data Sets

Two data sets have been applied in this work to construct a generative model: (1) *QM 9_sub*, which contains 25000 small organic molecules obtained randomly from QM9 data set.[57] Molecules in the data set consist of H, C, O, N, and F and contain up to 9 heavy atoms. As shown by the distribution in Figure S2A, about 80% of the molecules

https://doi.org/10.1021/jacsau.5c00052
JACS Au 2025, 5, 3082−3091

contain 9 atoms. Compounds in this data set is considered small organic molecules; (2) ChemFluor is composed of more than 4300 experimental solvated organic fluorescent materials (around 3000 distinct compounds) and 11,000 data (λ_{abs}, λ_{em}, and PLQY).[11] Most of the molecules contain more than 20 atoms (Figure S2b). A subdata set of ChemFluor named *ChemFluor30*, which contains 2280 molecules with atomic number less than 31, has been used in this work. 80% of the data set is randomly selected as used as the training set. 10% of the data set is used as the validation set and 10% of the data set is used as the test set. The percentage of the molecules that successfully reproduced by VAE is used to evaluate the performance of various models. To comprehensively analyze the rebuild rate of different decoders and encoders, we evaluate VAE, optimized AE, and adaptive β-VAE in both QM9_sub and ChemFluor30 data set.

Optimized AE and Adaptive β-VAE

In our research, we present a variant AE, termed optimized AE in our work, which adapts the traditional ELBO by excluding the KL divergence. This alteration allows the model to primarily focus on learning the latent distribution without diverging toward generating novel molecular structures.

We also developed a variant of the β-VAE, termed adaptive β-VAE, which adapts the traditional ELBO by modifying the KL divergence. This alteration allows the model to primarily focus on learning the latent distribution without diverging toward generating novel molecular structures. To be more specific, in the basic β-VAE, the parameter modifies the objective by introducing a β-term to balance the reconstruction and regularization

$$\mathcal{L}_\beta(\theta, \Phi; \mathbf{X}) = \mathbb{E}_{q_\phi(\hat{\mathbf{z}}\mathbf{X})}[\log p_\theta(\hat{\mathbf{X}}\mathbf{Z})] - \beta D_{KL}(q_\phi(\hat{\mathbf{Z}}\mathbf{X})\widehat{P(\mathbf{Z})})$$

The choice of a fixed β value is known to influence the learned representation. A larger β emphasizes disentanglement and a smoother latent space at the expense of reconstruction fidelity, while a smaller β prioritizes data fidelity over latent regularization. Although a single β-value is conceptually simple, it cannot adapt to the evolving needs of the training process. Early in training, encouraging a well-structured latent space can prevent representations from collapsing into narrow regions. Later in training, allowing more focus on reconstruction can refine the learned distributions and ensure high-quality decoding.

Therefore, we use an adaptive strategy to modify this process: An exponential decay schedule is one of the simplest adaptive strategies. Suppose β_{start} is the initial β-value, β_{end} is a lower bound, and $\rho \in (0,1)$ is a decay rate. At epoch t, we define $\beta^{(t)} = \max(\beta_{end}, \beta_{start} \cdot \rho^t)$. In early epochs, $\beta^{(t)} \approx \beta_{start}$, which is typically chosen to be ≥ 1 to ensure a well-regularized latent space. As the training progresses, $\beta^{(t)}$ smoothly decreases, shifting the balance toward more accurate reconstructions.

While exponential decay is effective and simple, other heuristics can be employed:

Linear decay: β decreased linearly over epochs until reaching β_{end}

Piecewise scheduling: using a high β during the initial T_{switch} epoch and then abruptly lowering it thereafter.

$$\beta^{(t)} = \beta_{high} \text{ if } t < T_{switch},\ \beta^{(t)} = \beta_{low} \text{ if } t \geq T_{switch}$$

Performance-based adjustment: monitoring KL divergence and reconstruction loss during training and adjusting β accordingly. For instance, if the KL term becomes too small, β is temporarily increased; if reconstruction lags, β is decreased. These alternatives provide flexibility and can be tailored to specific data sets or training objectives. In our work, $\rho \in (0,1)$, we select the 0.95, and therefore, the final loss can be expressed as

$$\mathcal{L}_\beta(\theta, \phi; x) = E_{q_\phi(z|x)}[\log p_\theta(x|z)] - \beta D_{KL}(q_\phi(z|x)|p(z))$$

Data Fusion

We also have enhanced the diversity and recognition capabilities of our molecular data set by incorporating a subset of molecules from the *ChemFluor30* data set and expanding it through similarity-based augmentation using the PubChem database. Furthermore, to enrich our data set with high-quality chemical structures, we have integrated data from the Joung et al.[12] Following a stringent selection process that filters molecules based on a maximum atom count criterion of 31 atoms, we combined the data sets. The resultant augmented database contains a total of 5310 molecules, significantly broadening the chemical space for the training of our models.[4] This methodological enhancement facilitates the learning of a generalized molecular representation.

■ ASSOCIATED CONTENT

Data Availability Statement

We express our sincere gratitude to Joung et al. for generously sharing their open access data set.[12] The data and data sets utilized in this manuscript are derived from prior publications and the PubChem database, specifically refs 11 and 12. All the data and code in our work can be found on the CodeOcean platform at https://codeocean.com/capsule/7686798/tree/v1 and are publicly available.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/jacsau.5c00052.

General experimental procedures, materials, and instruments; detailed computational methods; predictor benchmarking and discussion; ^{1}H and ^{13}C NMR spectra for all compounds; metrics for VAE reconstruction; performance of the adaptive β-schedule; benchmarking results for latent-space predictors; 10 × 5-fold cross-validation statistics; comparisons across emission energy prediction methods; molecular encoding workflow; molecular design pathways; diversity analysis; latent space visualizations; solvent-dependent fluorescence data; and external set validation (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Cheng-Wei Ju − *Pritzker School of Molecular Engineering, The University of Chicago, Chicago, Illinois 60637, United States;* orcid.org/0000-0002-2250-8548; Email: chengwei.ju99@gmail.com

Authors

Yuzhi Xu − *Department of Chemistry, New York University, New York, New York 10003, United States; Shanghai Frontiers Science Center of Artificial Intelligence and Deep Learning and NYU-ECNU Center for Computational Chemistry, Shanghai 200062, P. R. China;* orcid.org/0000-0002-3325-5427

Yongrui Luo − *Key Laboratory of Organofluorine Chemistry, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, Shanghai 200032, P. R. China;* orcid.org/0000-0001-9137-2595

Bo Li − *QuanMol Tech, Inc., San Carlos, California 94070, United States*

Weikang Jiang − *Key Laboratory of Organofluorine Chemistry, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, Shanghai 200032, P. R. China*

Jinyu Zhang − *State Key Laboratory and Institute of Elemento-Organic Chemistry, College of Chemistry, Nankai University, Tianjin 300071, P. R. China*

Jiangbo Wei − *Department of Chemistry and Department of Biological Sciences, National University of Singapore, Singapore 117544, Singapore*

https://doi.org/10.1021/jacsau.5c00052
JACS Au 2025, 5, 3082−3091

Hanzhi Bai − *Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, P. R. China*
Zhiqiang Wang − *Department of Electrical Engineering and Computer Science, Florida Atlantic University, Boca Raton, Florida 33431, United States*
Jiankai Ge − *Chemical and Biomolecular Engineering, University of Illinois at Urbana−Champaign, Urbana, Illinois 61801, United States;* orcid.org/0000-0002-7370-2797
Ruiming Lin − *Pritzker School of Molecular Engineering, The University of Chicago, Chicago, Illinois 60637, United States;* orcid.org/0000-0002-6750-0566
Zehan Mi − *Pritzker School of Molecular Engineering, The University of Chicago, Chicago, Illinois 60637, United States*
Haozhe Zhang − *Pritzker School of Molecular Engineering, The University of Chicago, Chicago, Illinois 60637, United States;* orcid.org/0000-0001-6363-5271
Yifeng Tang − *Pritzker School of Molecular Engineering, The University of Chicago, Chicago, Illinois 60637, United States;* orcid.org/0000-0003-4247-6712
Michael S. Jones − *Pritzker School of Molecular Engineering, The University of Chicago, Chicago, Illinois 60637, United States*
Xiaotian Li − *Faculty of Synthetic Biology, Shenzhen University of Advanced Technology, Shenzhen 518055, P. R. China;* orcid.org/0009-0003-2526-0722
John Z.H. Zhang − *Department of Chemistry, New York University, New York, New York 10003, United States; Shanghai Frontiers Science Center of Artificial Intelligence and Deep Learning and NYU-ECNU Center for Computational Chemistry, Shanghai 200062, P. R. China; Faculty of Synthetic Biology, Shenzhen University of Advanced Technology, Shenzhen 518055, P. R. China;* orcid.org/0000-0003-4612-1863

Complete contact information is available at: https://pubs.acs.org/10.1021/jacsau.5c00052

Author Contributions

All authors have given approval to the final version of the manuscript.

Notes

The authors declare the following competing financial interest(s): B.L. is a founder and equity holder for QuanMol Tech, Inc.

■ ACKNOWLEDGMENTS

We thank Qianzhen Shao (Vanderbilt University) for helpful discussion and constructive suggestions. We also sincerely acknowledge the HPC support provided by Greene at NYU and Jubail at NYUAD.

■ REFERENCES

(1) Wu, L.; Liu, J.; Li, P.; Tang, B.; James, T. D. Two-Photon Small-Molecule Fluorescence-Based Agents for Sensing, Imaging, and Therapy within Biological Systems. *Chem. Soc. Rev.* **2021**, *50* (2), 702−734.
(2) Dai, M.; Yang, Y. J.; Sarkar, S.; Ahn, K. H. Strategies to Convert Organic Fluorophores into Red/near-Infrared Emitting Analogues and Their Utilization in Bioimaging Probes. *Chem. Soc. Rev.* **2023**, *52* (18), 6344−6358.
(3) Itoh, T. Fluorescence and Phosphorescence from Higher Excited States of Organic Molecules. *Chem. Rev.* **2012**, *112* (8), 4541−4568.
(4) Jiang, G.; Liu, H.; Liu, H.; Ke, G.; Ren, T.-B.; Xiong, B.; Zhang, X.-B.; Yuan, L. Chemical Approaches to Optimize the Properties of Organic Fluorophores for Imaging and Sensing. *Angew. Chem., Int. Ed. Engl.* **2024**, *63*, No. e202315217.
(5) Ju, C.-W.; Wang, X.-C.; Li, B.; Ma, Q.; Shi, Y.; Zhang, J.; Xu, Y.; Peng, Q.; Zhao, D. Evolution of Organic Phosphor through Precision Regulation of Nonradiative Decay. *Proc. Natl. Acad. Sci. U.S.A.* **2023**, *120* (46), No. e2310883120.
(6) Choi, E. J.; Kim, E.; Lee, Y.; Jo, A.; Park, S. B. Rational Perturbation of the Fluorescence Quantum Yield in Emission-Tunable and Predictable Fluorophores (Seoul-Fluors) by a Facile Synthetic Method Involving C-H Activation. *Angew. Chem., Int. Ed.* **2014**, *53* (5), 1346−1350.
(7) Kim, E.; Lee, Y.; Lee, S.; Park, S. B. Discovery, Understanding, and Bioapplication of Organic Fluorophore: A Case Study with an Indolizine-Based Novel Fluorophore, Seoul-Fluor. *Acc. Chem. Res.* **2015**, *48* (3), 538−547.
(8) Shuai, Z.; Wang, D.; Peng, Q.; Geng, H. Computational Evaluation of Optoelectronic Properties for Organic/Carbon Materials. *Acc. Chem. Res.* **2014**, *47* (11), 3301−3309.
(9) Shuai, Z.; Xu, W.; Peng, Q.; Geng, H. From Electronic Excited State Theory to the Property Predictions of Organic Optoelectronic Materials. *Sci. China:Chem.* **2013**, *56* (9), 1277−1284.
(10) Gómez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Duvenaud, D.; Maclaurin, D.; Blood-Forsythe, M. A.; Chae, H. S.; Einzinger, M.; Ha, D.-G.; Wu, T.; Markopoulos, G.; Jeon, S.; Kang, H.; Miyazaki, H.; Numata, M.; Kim, S.; Huang, W.; Hong, S. I.; Baldo, M.; Adams, R. P.; Aspuru-Guzik, A. Design of Efficient Molecular Organic Light-Emitting Diodes by a High-Throughput Virtual Screening and Experimental Approach. *Nat. Mater.* **2016**, *15* (10), 1120−1127.
(11) Ju, C.-W.; Bai, H.; Li, B.; Liu, R. Machine Learning Enables Highly Accurate Predictions of Photophysical Properties of Organic Fluorescent Materials: Emission Wavelengths and Quantum Yields. *J. Chem. Inf. Model.* **2021**, *61* (3), 1053−1065.
(12) 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* (4), 427−438.
(13) Ye, Z.-R.; Huang, I.-S.; Chan, Y.-T.; Li, Z.-J.; Liao, C.-C.; Tsai, H.-R.; Hsieh, M.-C.; Chang, C.-C.; Tsai, M.-K. Predicting the Emission Wavelength of Organic Molecules Using a Combinatorial QSAR and Machine Learning Approach. *RSC Adv.* **2020**, *10* (40), 23834−23841.
(14) Greenman, P.; Green, W. H.; Gómez-Bombarelli, R. Multi-Fidelity Prediction of Molecular Optical Peaks with Deep Learning. *Chem. Sci.* **2022**, *13* (4), 1152−1162.
(15) Terrones, G.; Duan, C.; Nandy, A.; Kulik, J. H. Low-Cost Machine Learning Prediction of Excited State Properties of Iridium-Centered Phosphors. *Chem. Sci.* **2023**, *14* (6), 1419−1433.
(16) Axelrod, S.; Schwalbe-Koda, D.; Mohapatra, S.; Damewood, J.; Greenman, K. P.; Gómez-Bombarelli, R. Learning Matter: Materials Design with Machine Learning and Atomistic Simulations. *Acc. Mater. Res.* **2022**, *3* (3), 343−357.
(17) Gong, J.; Gong, W.; Wu, B.; Wang, H.; He, W.; Dai, Z.; Li, Y.; Liu, Y.; Wang, Z.; Tuo, X.; Lam, J. W. Y.; Qiu, Z.; Zhao, Z.; Tang, B. Z. ASBase: The Universal Database for Aggregate Science. *Aggregate* **2023**, *4* (1), No. e263.
(18) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse Molecular Design Using Machine Learning: Generative Models for Matter Engineering. *Science* **2018**, *361* (6400), 360−365.
(19) Kim, B.; Lee, S.; Kim, J. Inverse Design of Porous Materials Using Artificial Neural Networks. *Sci. Adv.* **2020**, *6* (1), No. eaax9324.
(20) Chen, C.-T.; Gu, G. X. Generative Deep Neural Networks for Inverse Materials Design Using Backpropagation and Active Learning. *Adv. Sci.* **2020**, *7* (5), 1902607.
(21) Kim, K.; Kang, S.; Yoo, J.; Kwon, Y.; Nam, Y.; Lee, D.; Kim, I.; Choi, Y.-S.; Jung, Y.; Kim, S.; Son, W.-J.; Son, J.; Lee, H. S.; Kim, S.; Shin, J.; Hwang, S. Deep-Learning-Based Inverse Design Model for

Intelligent Discovery of Organic Molecules. *npj Comput. Mater.* **2018**, *4* (1), 1−7.
(22) Xu, Y.; Ge, J.; Ju, C.-W. Machine Learning in Energy Chemistry: Introduction, Challenges and Perspectives. *Energy Adv.* **2023**, *2* (7), 896−921.
(23) Xu, Y.; Ju, C.-W.; Li, B.; Ma, Q.-S.; Chen, Z.; Zhang, L.; Chen, J. Hydrogen Evolution Prediction for Alternating Conjugated Copolymers Enabled by Machine Learning with Multidimension Fragmentation Descriptors. *ACS Appl. Mater. Interfaces* **2021**, *13* (29), 34033−34042.
(24) Heid, E.; Greenman, K. P.; Chung, Y.; Li, S.-C.; Graff, D. E.; Vermeire, F. H.; Wu, H.; Green, W. H.; McGill, C. J. Chemprop: A Machine Learning Package for Chemical Property Prediction. *J. Chem. Inf. Model.* **2024**, *64* (1), 9−17.
(25) Reiser, P.; Neubert, M.; Eberhard, A.; Torresi, L.; Zhou, C.; Shao, C.; Metni, H.; van Hoesel, C.; Schopmans, H.; Sommer, T.; Friederich, P. Graph Neural Networks for Materials Science and Chemistry. *Commun. Mater.* **2022**, *3* (1), 1−18.
(26) Le, T.; Winter, R.; Noé, F.; Clevert, D.-A. Neuraldecipher − Reverse-Engineering Extended-Connectivity Fingerprints (ECFPs) to Their Molecular Structures. *Chem. Sci.* **2020**, *11* (38), 10378−10389.
(27) Ilnicka, A.; Schneider, G. Compression of Molecular Fingerprints with Autoencoder Networks. *Mol. Inf.* **2023**, *42* (6), 2300059.
(28) Sumita, M.; Terayama, K.; Suzuki, N.; Ishihara, S.; Tamura, R.; Chahal, M. K.; Payne, D. T.; Yoshizoe, K.; Tsuda, K. De Novo Creation of a Naked Eye−Detectable Fluorescent Molecule Based on Quantum Chemical Computation and Machine Learning. *Sci. Adv.* **2022**, *8* (10), No. eabj3906.
(29) Tang, Y.; Kim, Y.; Ip, C. K. M.; Bahmani, A.; Chen, Q.; Rosenberger, G.; Esser-Kahn, P.; Ferguson, L. A. Data-Driven Discovery of Innate Immunomodulators via Machine Learning-Guided High Throughput Screening. *Chem. Sci.* **2023**, *14* (44), 12747−12766.
(30) Koscher, B. A.; Canty, R. B.; McDonald, M. A.; Greenman, K. P.; McGill, C. J.; Bilodeau, C. L.; Jin, W.; Wu, H.; Vermeire, F. H.; Jin, B.; Hart, T.; Kulesza, T.; Li, S.-C.; Jaakkola, T. S.; Barzilay, R.; Gómez-Bombarelli, R.; Green, W. H.; Jensen, K. F.; Autonomous, K. F. Autonomous, multiproperty-driven molecular discovery: From predictions to measurements and back. *Science* **2023**, *382* (6677), No. eadi1407.
(31) Dimitrov, T.; Kreisbeck, C.; Becker, J. S.; Aspuru-Guzik, A.; Saikin, S. K. Autonomous Molecular Design: Then and Now. *ACS Appl. Mater. Interfaces* **2019**, *11* (28), 24825−24836.
(32) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4* (2), 268−276.
(33) Alverson, M.; Baird, S. G.; Murdock, R.; Johnson, J.; Sparks, T. D.; et al. Generative Adversarial Networks and Diffusion Models in Material Discovery. *Digital Discovery* **2024**, *3* (1), 62−80.
(34) Flam-Shepherd, D.; Wu, T. C.; Aspuru-Guzik, A. MPGVAE: improved generation of small organic molecules using message passing neural nets. *Mach. Learn. Sci. Technol.* **2021**, *2* (4), 045010.
(35) Sun, W.; Zheng, Y.; Yang, K.; Zhang, Q.; Shah, A. A.; Wu, Z.; Sun, Y.; Feng, L.; Chen, D.; Xiao, Z.; Lu, S.; Li, Y.; Sun, K. Machine Learning−Assisted Molecular Design and Efficiency Prediction for High-Performance Organic Photovoltaic Materials. *Sci. Adv.* **2019**, *5* (11), No. eaay4275.
(36) Nagasawa, S.; Al-Naamani, E.; Saeki, A. Computer-Aided Screening of Conjugated Polymers for Organic Solar Cell: Classification by Random Forest. *J. Phys. Chem. Lett.* **2018**, *9* (10), 2639−2646.
(37) Krenn, M.; Häse, F.; Nigam, A.; Friederich, P.; Aspuru-Guzik, A. Self-Referencing Embedded Strings (SELFIES): A 100% Robust Molecular String Representation. *Mach. Learn. Sci. Technol.* **2020**, *1* (4), 045024.
(38) Laurent, A. D.; Jacquemin, D. TD-DFT Benchmarks: A Review. *Int. J. Quantum Chem.* **2013**, *113* (17), 2019−2039.
(39) Ju, C.-W.; French, E. J.; Geva, N.; Kohn, A. W.; Lin, Z. Stacked Ensemble Machine Learning for Range-Separation Parameters. *J. Phys. Chem. Lett.* **2021**, *12* (39), 9516−9524.
(40) Chantzis, A.; Cerezo, J.; Perrier, A.; Santoro, F.; Jacquemin, D. Optical Properties of Diarylethenes with TD-DFT: 0−0 Energies, Fluorescence, Stokes Shifts, and Vibronic Shapes. *J. Chem. Theory Comput.* **2014**, *10* (9), 3944−3957.
(41) Charaf-Eddin, A.; Planchat, A.; Mennucci, B.; Adamo, C.; Jacquemin, D. Choosing a Functional for Computing Absorption and Fluorescence Band Shapes with TD-DFT. *J. Chem. Theory Comput.* **2013**, *9* (6), 2749−2760.
(42) Hall, D.; Sancho-García, J. C.; Pershin, A.; Beljonne, D.; Zysman-Colman, E.; Olivier, Y. Benchmarking DFT Functionals for Excited-State Calculations of Donor−Acceptor TADF Emitters: Insights on the Key Parameters Determining Reverse Inter-System Crossing. *J. Phys. Chem. A* **2023**, *127* (21), 4743−4757.
(43) Lin, Z.; Kohn, A. W.; Van Voorhis, T. Toward Prediction of Nonradiative Decay Pathways in Organic Compounds II: Two Internal Conversion Channels in BODIPYs. *J. Phys. Chem. C* **2020**, *124* (7), 3925−3938.
(44) Kohn, A. W.; Lin, Z.; Van Voorhis, T. Toward Prediction of Nonradiative Decay Pathways in Organic Compounds I: The Case of Naphthalene Quantum Yields. *J. Phys. Chem. C* **2019**, *123* (25), 15394−15402.
(45) Zhao, L.; Li, J.; Li, L.; Hu, W. Recent Advances in Small-Molecule Organic Fluorescent Semiconductors. *J. Mater. Chem. C* **2024**, *12* (13745), 13745−13761.
(46) Liu, W.; Deng, S.; Zhang, L.; Ju, C.-W.; Xie, Y.; Deng, W.; Chen, J.; Wu, H.; Cao, Y. Short-Wavelength Infrared Organic Light-Emitting Diodes from A−D−A′−D−A Type Small Molecules with Emission Beyond 1100 nm (Adv. Mater. 39/2023). *Adv. Mater.* **2023**, *35* (39), 2370279.
(47) Yang, Q.-Y.; Lehn, J.-M. Bright White-Light Emission from a Single Organic Compound in the Solid State. *Angew. Chem., Int. Ed.* **2014**, *53* (18), 4572−4577.
(48) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *J. Chem. Theory Comput.* **2019**, *15* (3), 1652−1671.
(49) Lee, J.-H.; Chen, C.-H.; Lee, P.-H.; Lin, H.-Y.; Leung, M.; Chiu, T.-L.; Lin, C.-F. Blue Organic Light-Emitting Diodes: Current Status, Challenges, and Future Outlook. *J. Mater. Chem. C* **2019**, *7* (20), 5874−5888.
(50) Hung, S.-H.; Ye, Z.-R.; Cheng, C.-F.; Chen, B.; Tsai, M.-K. Enhanced Predictions for the Experimental Photophysical Data Using the Featurized Schnet-Bondstep Approach. *J. Chem. Theory Comput.* **2023**, *19* (14), 4559−4567.
(51) Yang, S.; Cho, K.; Merchant, A.; Abbeel, P.; Schuurmans, D.; Mordatch, I.; Cubuk, E. D. Scalable Diffusion for Materials Generation. *arXiv* **2023**, arXiv:2311.09235v2.
(52) Xu, Y.; Liu, X.; Xia, W.; Ge, J.; Ju, C.-W.; Zhang, H.; Zhang, J. Z. H. ChemXTree: A Feature-Enhanced Graph Neural Network-Neural Decision Tree Framework for ADMET Prediction. *J. Chem. Inf. Model.* **2024**, *64* (22), 8440−8452.
(53) Xu, M.; Powers, A. S.; Dror, R. O.; Ermon, S.; Leskovec, J. Geometric Latent Diffusion Models for 3D Molecule Generation. In *Proceedings of the International Conference on Machine Learning*; PMLR, 2023; pp 38592−38610.
(54) Queen, O.; McCarver, G. A.; Thatigotla, S.; Abolins, B. P.; Brown, C. L.; Maroulas, V.; Vogiatzis, K. D. Polymer Graph Neural Networks for Multitask Property Learning. *npj Comput. Mater.* **2023**, *9* (1), 90.
(55) Nigam, A.; Pollice, R.; Tom, G.; Jorner, K.; Willes, J.; Thiede, L.; Kundaje, A.; Aspuru-Guzik, A. Tartarus: A Benchmarking Platform for Realistic and Practical Inverse Molecular Design. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 3263−3306.
(56) Kingma, D. P.; Welling, M. Auto-Encoding Variational Bayes. *arXiv* **2022**, arXiv:1312.6114v11.

https://doi.org/10.1021/jacsau.5c00052
JACS Au 2025, 5, 3082−3091

(57) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Quantum Chemistry Structures and Properties of 134 Kilo Molecules. *Sci. Data* **2014**, *1* (1), 140022.

https://doi.org/10.1021/jacsau.5c00052
JACS Au 2025, 5, 3082−3091