Exploring Optimized Organic Fluorophore Search Through Simple Experimental Data-Driven VAE

Cheng-Wei Ju[1,*#], Yuzhi Xu[2,4*#], Yongrui Luo[5,#], Bo Li[6], Weikang Jiang[5], Jingyu Zhang[7], Hanzhi Bai[8], Zhiqiang Wang[9], Jiankai Ge[10], Ruiming Lin[1], Zehan Mi[1], Haozhe Zhang[1], Yifeng Tang[1], Michael S. Jones[1], and Xiaotian Li[3], John Z.H. Zhang[2,3,4]

1. Pritzker School of Molecular Engineering, The University of Chicago, Chicago, Illinois 60637, United States
2. Department of Chemistry, New York University, New York, New York 10003, United States
3. Faculty of Synthetic Biology and Institute of Synthetic Biology, Shenzhen Institute of Advanced Technology, Shenzhen 518055, P. R. China
4. Shanghai Frontiers Science Center of Artificial Intelligence and Deep Learning and NYU-ECNU Center for Computational Chemistry, NYU Shanghai, Shanghai 200062, P. R. China
5. Key Laboratory of Organofluorine Chemistry, Shanghai Institute of Organic Chemistry, Chinese Academy of Sciences, Shanghai 200032, P. R. China
6. QuanMol Tech, Inc., San Carlos, California 94070, United States
7. State Key Laboratory and Institute of Elemento-Organic Chemistry, College of Chemistry, Nankai University, Tianjin 300071, P. R. China
8. Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, P. R. China
9. Department of Electrical Engineering and Computer Science, Florida Atlantic University, Boca Raton, Florida, 33431, United States
10. Chemical and Biomolecular Engineering, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, United States

#These authors contributed equally to this work
E-mail: C.-W. Ju, cwju@uchicago.edu; Y. Xu, yuzhixu@nyu.edu

Abstract: Designing organic fluorescent molecules with tailored optical properties is challenging in decades, while the new avenue was opened by the statistical models. Inverse design has garnered considerable interest in organic materials science but concentrates on arbitrary design or theoretical properties. Here, we introduce a strategy that enables direct optimization of specific experimental properties in the inverse design process, utilizing a variational autoencoder (VAE) with a latent vector-based prediction model. Omitting the Kullback-Leibler divergence and separate training strategy successfully improved the generator's robustness and molecular diversity. We confirm the latent vectors obtained from VAE are powerful inputs for downstream prediction models of experimental properties, fluorescence energy and quantum yield. Our approach for the optimized search of organic fluorescent materials, substantiated by gradient space derived from latent vector and validated by newly synthesized and uncharacterized molecules, shows potential for broader applications in diverse organic material design.

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 optical properties of organic materials (**Fig 1A**).[11–17] For instance, the ChemFluor dataset served as the basis for Ju et al.'s 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 (**Fig S1**).[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 show promise in predictive modeling but their exploitation for reverse engineering has been limited.[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, generator has been explored in ML-assisted material design as well but concentrate either on arbitrary design or theoretical properties.[33,34]

Here, we questioned if the search for optimized compounds with specific experimental properties in materials science can also be achieved through an integrated generator-predictor framework (**Fig 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 datasets.[35,36] The limited size of experimental datasets will compromise the generator's efficacy. Additionally, this integration typically necessitates co-training of the decoder and predictor.[32] Lastly, predicting experimental properties—such as fluorescence wavelengths, quantum yields 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 RB-Boost VAE and a predictor to directly optimize organic fluorophores on a high dimensional space fitted from experimental energies (**Fig 1C**). We train the generator and predictor separately, and thus make the data fusion in the generator become possible. Utilizing the latent vectors from this RB-Boost VAE, we constructed a prediction model for the photophysical properties, including photoluminescence quantum yield (PLQY) and emission energy within error of quantum mechanical precision (~0.12 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.

To construct a VAE with improved molecular diversity, we used SELF-referencing Embedded Strings (SELFIES), a much more robust molecular string representation compared to SMILES, which can be easily standardized and transferred into a one-hot format (**Fig 1D**).[37] We first validate performance on the QM9 dataset using a simple VAE model, which produces a reconstruction rate of 98.5% (**Fig 1E** and **Table S1**). However, when we move to a dataset with larger fluorescent molecules, ChemFluor30 (a sub-dataset of ChemFluor with the molecules smaller than 30 heavy atoms), the performance is largely reduced. Meanwhile, the small size of this dataset (~2,000 molecules) also motivated us to improve its broader molecular diversity. To overcome these challenges, we first modify the loss function of the VAE by excluding the Kullback-Leibler divergence inspired by traditional autoencoder, named as RB VAE (ReBuild VAE) (**Method S1.1.2**). Although KL divergence typically contributes to the regularization of the model, its exclusion allows the model to prioritize minimizing reconstruction loss and increase the reconstruction rate, which is critical for the direct optimization process. Additionally, we rationalized that integrating diverse molecular scaffolds during the training process can broaden the scope of information sampling within the latent layer. Therefore, we employed data fusion to compensate for the dataset's limitations by incorporating additional molecules that align with established protocols (**Method S1.1.3**). Complementarily, we performed targeted augmentation *via* PubChem, enriching our dataset with structurally analogous molecules (named as RB-Boost VAE). This dual strategy not only expanded our dataset but also introduced a wider array of molecular characteristics. The efficacy of these enhancements was confirmed by an improved reconstruction rate, which saw an increase from 59% to 64% when benchmarked against the original model.

We then evaluated between the enhanced RB-Boost VAE and the original model by perturbing a subset of latent vectors to generate molecules (**Fig 1F** and **Fig S2**). The RB-Boost VAE demonstrated superior performance, generating an average of 6.3 times more total distinct molecules with a broader chemical feature set, indicative of a more complex chemical space encapsulated during model training (**Fig S2A**). This contrasted with the original VAE, which tends to generate more similar structures. Moreover, the RB-Boost VAE facilitated the generation of transitional molecular structures through interpolation between two selected latent vectors (**Fig 1G** and **Fig S3**). Despite some resulting in non-viable 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.

With the establishment of the generator, we move to the prediction model. To adapt our VAE for chemical prediction, we train our predictor separately based on the learned latent space. 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 Gradient Boosting Regression Trees (GBRT), lauded for its predictive precision in our prior research (**Table S2**). The model results in a mean absolute error (MAE) of 0.128 eV using latent vectors as the input, surpassing TD-DFT accuracy (~0.15 to

0.20 eV) and are sufficient for utilizing in virtual screening (**Fig 1H** and **Table S3**).[38–42] A similar MAE of 0.124 eV was obtained from one-hot SELFIES as input indicated the high fidelity of the latent vector generated from SELFIES. Utilizing t-SNE visualizations, we observe the cluster of various structures such as Rhodamine and BODIPY derivatives (**Fig 1I**). Meanwhile, the analogous distributions between latent vectors and SELFIES proves that they are high fidelity predictors, while the distinct from ECFP4 suggest their uniqueness (**Fig S4**).

Furthermore, we assess PLQY predictions within latent space. Considering the distribution of PLQY and real-world situations, we apply 0.25 as a threshold to classify the bright and dark molecules (**Fig 1J**). Our classifier discerns between bright and dark materials with an accuracy of 0.81, rendering it suitable for practical predictive applications (**Fig 1K**).

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 (**Method S1.1.5**). We applied this approach with molecules shown in the center of **Fig 2A**, where the manipulation of latent vectors yielded diverse molecules with predicted emission energies ranging from 1.9 eV to 2.3 eV. 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 methods to estimate the HOMO-LUMO gap of molecules with similar skeleton generated in this high-dimensional space (**Fig S5)**.[43] Molecules with similar skeleton are selected here for the computational validation since we want to minimize the structure diversity that increase the complexity and difference between computational and experimental properties. The correlation further supports the validity of our approach (**Fig 2B**). Although it needs to be recognized that (1) semiempirical methods is not accurate; (2) calculated H-L gap only reflect the electronic structure in ground state while emission is highly related to excited state, we rational that molecules with similar skeleton should at least have similar trend between H-L gap and fluorescence wavelength. This localized optimization highlights our approach's potential in fine-tuning 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 subjected to the RB-Boost VAE with successful reproduction of **1** and **2**, whereas **3** underwent a transformation to 5-methyl-1H-pyrrolo[1,2-a]imidazole **4** (**Fig 2C**). Later, to evaluate the performance of the predictor, we characterized their fluorescence spectra in CH_2Cl_2 (**Fig 2D**). 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 fine-tuning. Due to the complexity introduced by the high-dimensional latent space, we centered our exploration on the nearby molecules of imidazopyridine derivative **5** (**Fig 2E** and **Fig S6-10**). We choose molecule **6** with an extended π-system, for its plausible structure and predicted red-shifted emission compare 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 (**Fig 2F**).[44]

In summary, we've successfully demonstrated the feasibility of optimized organic materials search based on experimental properties through a novel combination of VAE and a predictor. 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. Furthermore, synthesized molecules support the feasibility of our generator and predictor. Using a fluorophore skeleton as an example, we designed and synthesized **7** and confirmed bright blue emission, further demonstrating the possibility of our strategy in materials discovery. This streamlined workflow not only enables fine-tuning of molecular properties for optimized compounds but also heralds a new era of material design, with promising applications in the development of OLEDs, and extending potential to OPVs and OFETs.

Figure1. RB-Boost VAE and Predictor.

(A) The schematic for direct design.

(B) The schematic for inverse design by searching for targeted molecules.

(C) Diagram for different molecular representation.

(D) Overview of the methodology for optimized materials search developed in this study.

(E) Evaluation of VAE model accuracy across various datasets. RB VAE refers to the model excluding the Kullback-Leibler divergence, while RB-Boost VAE incorporates data fusion into the RB VAE framework.

(F) RB-Boost VAE enhances diversity in molecular generation. Only typical molecules generated by disturbing a randomly selected and consistent subset of vectors are shown here.

(G) Capability of RB-Boost VAE in generating viable molecules from latent vector interpolations.

(H) The predictor based on GBRT with latent vector (left) or SELFIES (right) well reproduces the emission energy in the test set.

(I) T-distributed stochastic neighbor embedding (t-SNE) of latent vectors. Colors indicate the emission energies.

(J) The distribution of PLQY in the dataset. 0.25 is set as the threshold for bright and dark molecules.

(K) The prediction performance of PLQY with latent vector as input.

Figure2. 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 obtained in 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) Fine-tuning on the fluorophore skeleton (imidazopyridine) by exploring nearby molecules and controlling synthesis complexity.

(F) The fluorescence spectrum of molecule **7**.

Conflict:

B.L. is a founder and equity holder for QuanMol Tech, Inc.

Acknowledge:

We thank Mr. Qianzhen Shao (Vanderbilt University) for helpful discussion and constructive suggestions.

Supporting Information:

Detailed information for the method, supplementary tables and figures, and protocol for molecular synthesis (PDF).

Data availability:

We express our sincere gratitude to Joung et al. for generously sharing their open-access dataset[12]. The data and datasets utilized in this manuscript are derived from prior publications and the PubChem database, specifically references 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.

Reference:

(1) Wu, L.; Liu, J.; Li, P.; Tang, B.; D. James, T. Two-Photon Small-Molecule Fluorescence-Based Agents for Sensing, Imaging, and Therapy within Biological Systems. *Chem. Soc. Rev.* **2021**, *50* (2), 702–734. https://doi.org/10.1039/D0CS00861C.

(2) Dai, M.; Jae Yang, Y.; Sarkar, S.; Han Ahn, K. 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. https://doi.org/10.1039/D3CS00475A.

(3) Itoh, T. Fluorescence and Phosphorescence from Higher Excited States of Organic Molecules. *Chem. Rev.* **2012**, *112* (8), 4541–4568. https://doi.org/10.1021/cr200166m.

(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.* **2023**, e202315217. https://doi.org/10.1002/anie.202315217.

(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.* **2023**, *120* (46), e2310883120. https://doi.org/10.1073/pnas.2310883120.

(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. https://doi.org/10.1002/anie.201308826.

(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. https://doi.org/10.1021/ar500370v.

(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. https://doi.org/10.1021/ar400306k.

(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. https://doi.org/10.1007/s11426-013-4916-7.

(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. https://doi.org/10.1038/nmat4717.

(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. https://doi.org/10.1021/acs.jcim.0c01203.

(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. https://doi.org/10.1021/jacsau.1c00035.

(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. https://doi.org/10.1039/D0RA05014H.

(14) P. Greenman, K.; H. Green, W.; Gómez-Bombarelli, R. Multi-Fidelity Prediction of Molecular Optical Peaks with Deep Learning. *Chem. Sci.* **2022**, *13* (4), 1152–1162. https://doi.org/10.1039/D1SC05677H.

(15) G. Terrones, G.; Duan, C.; Nandy, A.; J. Kulik, H. Low-Cost Machine Learning Prediction of Excited State Properties of Iridium-Centered Phosphors. *Chem. Sci.* **2023**, *14* (6), 1419–1433. https://doi.org/10.1039/D2SC06150C.

(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. https://doi.org/10.1021/accountsmr.1c00238.

(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), e263. https://doi.org/10.1002/agt2.263.

(18) Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse Molecular Design Using Machine Learning: Generative Models for Matter Engineering. *Science* **2018**, *361* (6400), 360–365. https://doi.org/10.1126/science.aat2663.

(19) Kim, B.; https://orcid.org/0000-0002-8283-4572; Kim, J.; _; _; -3844-8789. Inverse Design of Porous Materials Using Artificial Neural Networks. *Sci. Adv.* **2020**, *6* (1), eaax9324. https://doi.org/10.1126/sciadv.aax9324.

(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. https://doi.org/10.1002/advs.201902607.

(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. https://doi.org/10.1038/s41524-018-0128-1.

(22) Xu, Y.; Ge, J.; Ju, C.-W. Machine Learning in Energy Chemistry: Introduction, Challenges and Perspectives. *Energy Adv.* **2023**, *2* (7), 896–921. https://doi.org/10.1039/D3YA00057E.

(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. https://doi.org/10.1021/acsami.1c05536.

(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. https://doi.org/10.1021/acs.jcim.3c01250.

(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. https://doi.org/10.1038/s43246-022-00315-6.

(26) Lian, J.; Meng, F.; Wang, W.; Zhang, Z. Recent Trends in Fluorescent Organic Materials for Latent Fingerprint Imaging. *Front. Chem.* **2020**, *8*.

(27) Ilnicka, A.; Schneider, G. Compression of Molecular Fingerprints with Autoencoder Networks. *Mol. Inform.* **2023**, *42* (6), 2300059. https://doi.org/10.1002/minf.202300059.

(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), eabj3906. https://doi.org/10.1126/sciadv.abj3906.

(29) Tang, Y.; Y. Kim, J.; M. Ip, C. K.; Bahmani, A.; Chen, Q.; G. Rosenberger, M.; P. Esser-Kahn, A.; L. Ferguson, A. Data-Driven Discovery of Innate Immunomodulators via Machine Learning-Guided High Throughput Screening. *Chem. Sci.* **2023**, *14* (44), 12747–12766. https://doi.org/10.1039/D3SC03613H.

(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, Multiproperty-Driven Molecular Discovery: From Predictions to Measurements and Back. *Science* **2023**, *382* (6677), eadi1407. https://doi.org/10.1126/science.adi1407.

(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. https://doi.org/10.1021/acsami.9b01226.

(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. https://doi.org/10.1021/acscentsci.7b00572.

(33) Alverson, M.; G. Baird, S.; Murdock, R.; Sin-Hang Ho, (Enoch); Johnson, J.; D. Sparks, T. Generative Adversarial Networks and Diffusion Models in Material Discovery. *Digit. Discov.* **2024**, *3* (1), 62–80. https://doi.org/10.1039/D3DD00137G.

(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. https://doi.org/10.1088/2632-2153/abf5b7.

(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), eaay4275. https://doi.org/10.1126/sciadv.aay4275.

(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. https://doi.org/10.1021/acs.jpclett.8b00635.

(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. https://doi.org/10.1088/2632-2153/aba947.

(38) Laurent, A. D.; Jacquemin, D. TD-DFT Benchmarks: A Review. *Int. J. Quantum Chem.* **2013**, *113* (17), 2019–2039. https://doi.org/10.1002/qua.24438.

(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. https://doi.org/10.1021/acs.jpclett.1c02506.

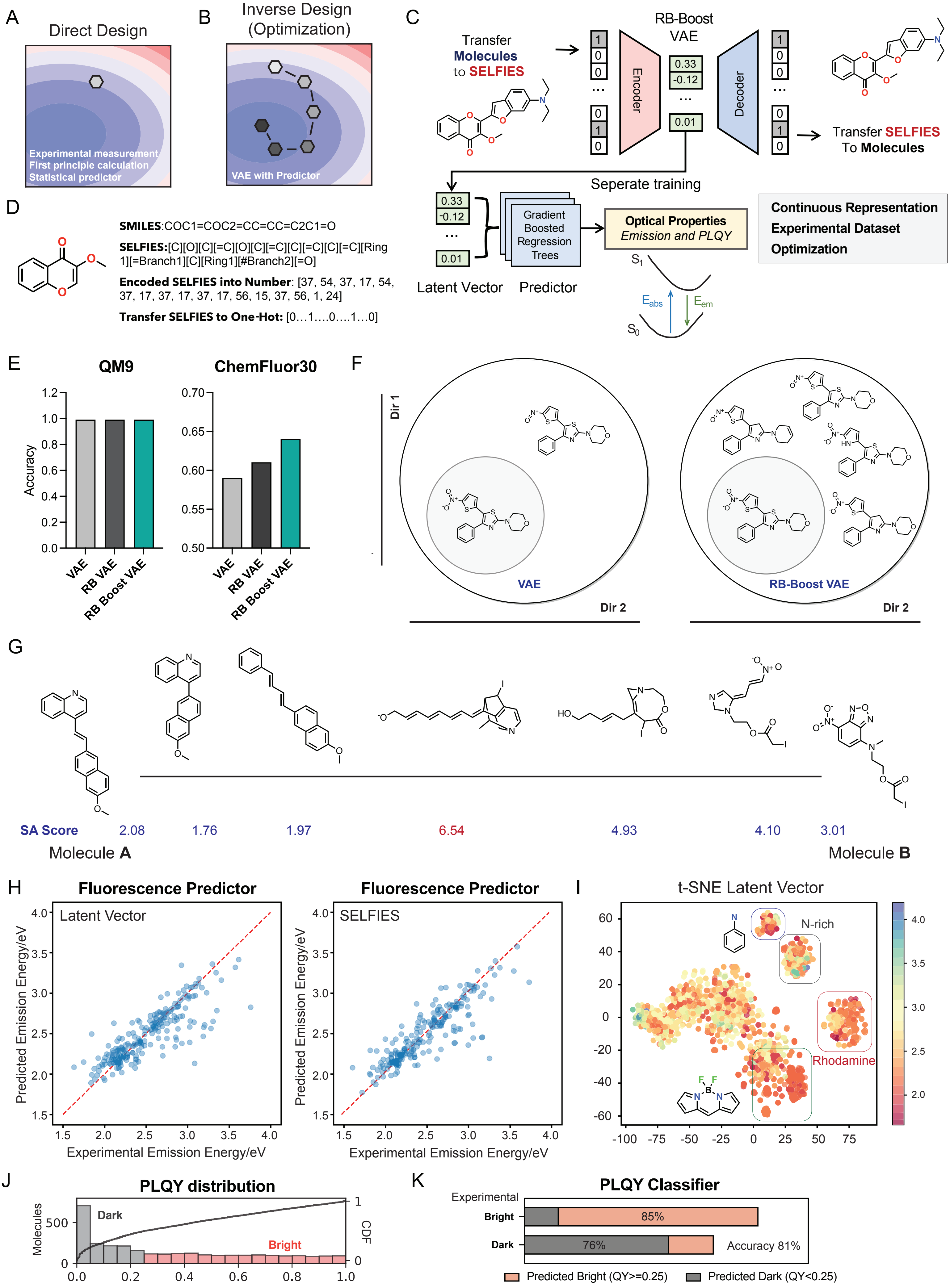
(40) Chantzis, A.; Cerezo, J.; Perrier, A.; Santoro, F.; Jacquemin, D. Optical Properties of Diarylethenes with TD-DFT: 0–0 Energies, Fluorescence, Stokes Shifts, and Vibronic Shapes. *J. Chem. Theory Comput.* **2014**, *10* (9), 3944–3957. https://doi.org/10.1021/ct500371u.

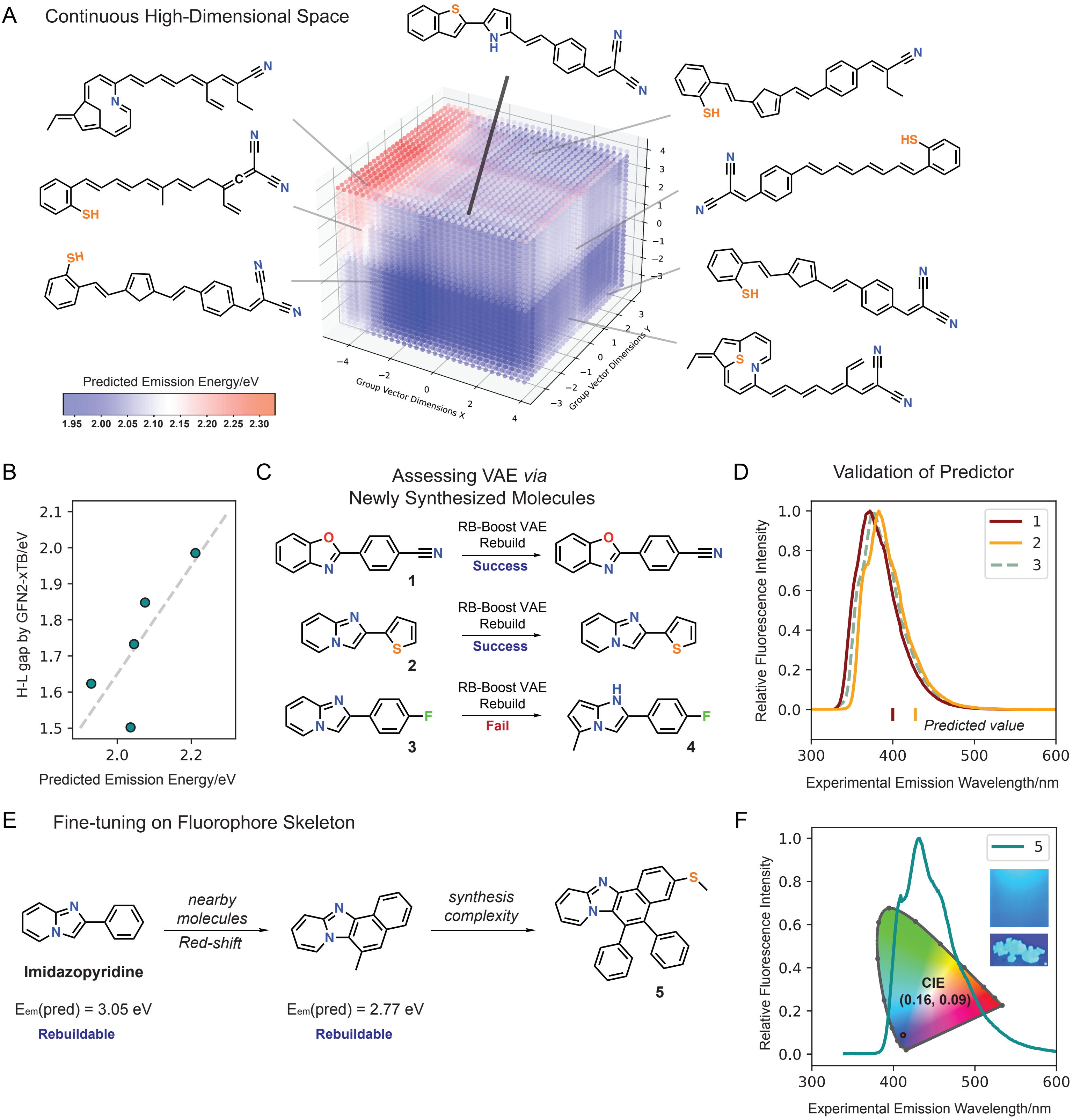
(41) Charaf-Eddin, A.; Planchat, A.; Mennucci, B.; Adamo, C.; Jacquemin, D. Choosing a Functional for Computing Absorption and Fluorescence Band Shapes with TD-DFT. *J. Chem. Theory Comput.* **2013**, *9* (6), 2749–2760. https://doi.org/10.1021/ct4000795.

(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. https://doi.org/10.1021/acs.jpca.2c08201.

(43) 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. https://doi.org/10.1021/acs.jctc.8b01176.

(44) 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. https://doi.org/10.1039/C9TC00204A.


A
Direct Design
Experimental measurement
First principle calculation
Statistical predictor
B
Inverse Design
(Optimization)
VAE with Predictor
C
Transfer
Molecules
to SELFIES
RB-Boost
VAE
Encoder
0.33
-0.12
0.01
Decoder
Transfer SELFIES
To Molecules
Seperate training
Latent Vector
Gradient Boosted Regression Trees
Predictor
Optical Properties
Emission and PLQY
S_1
E_{abs}
E_{em}
S_0
Continuous Representation
Experimental Dataset
Optimization
D
SMILES:COC1=COC2=CC=CC=C2C1=O
SELFIES:[C][O][C][=C][O][C][=C][C][=C][C][=C][Ring1][=Branch1][C][Ring1][#Branch2][=O]
Encoded SELFIES into Number: [37, 54, 37, 17, 54, 37, 17, 37, 17, 37, 17, 56, 15, 37, 56, 1, 24]
Transfer SELFIES to One-Hot: [0…1….0….1…0]
E
QM9
ChemFluor30
Accuracy
VAE
RB VAE
RB Boost VAE
F
Dir 1
Dir 2
VAE
RB-Boost VAE
G
SA Score
2.08
1.76
1.97
6.54
4.93
4.10
3.01
Molecule A
Molecule B
H
Fluorescence Predictor
Latent Vector
SELFIES
Predicted Emission Energy/eV
Experimental Emission Energy/eV
I
t-SNE Latent Vector
N-rich
Rhodamine
J
PLQY distribution
Molecules
Dark
Bright
CDF
K
PLQY Classifier
Experimental
Bright
Dark
85%
76%
Accuracy 81%
Predicted Bright (QY>=0.25)
Predicted Dark (QY<0.25)


A
Continuous High-Dimensional Space
Predicted Emission Energy/eV
1.95 2.00 2.05 2.10 2.15 2.20 2.25 2.30
Group Vector Dimensions X
Group Vector Dimensions Y
−4 −2 0 2 4
−3 −2 −1 0 1 2 3
4 3 2 1 0 −1 −2 −3
B
H-L gap by GFN2-xTB/eV
Predicted Emission Energy/eV
2.1 2.0 1.9 1.8 1.7 1.6 1.5
2.0 2.2
C
Assessing VAE via
Newly Synthesized Molecules
RB-Boost VAE
Rebuild
Success
Success
Fail
1
2
3
4
D
Validation of Predictor
Relative Fluorescence Intensity
Experimental Emission Wavelength/nm
1
2
3
Predicted value
1.0 0.8 0.6 0.4 0.2 0.0
300 400 500 600
E
Fine-tuning on Fluorophore Skeleton
Imidazopyridine
nearby
molecules
Red-shift
synthesis
complexity
E_{em}(pred) = 3.05 eV
Rebuildable
E_{em}(pred) = 2.77 eV
Rebuildable
5
F
Relative Fluorescence Intensity
Experimental Emission Wavelength/nm
5
CIE
(0.16, 0.09)
1.0 0.8 0.6 0.4 0.2 0.0
300 400 500 600