Generating Readily Synthesizable Dye Scaffolds with SyntheFluor

Ruhi Sayana [*1] **Kate Callon** [*1 2] **Jennifer Xu** [*1 3] **Jonathan Deutsch** [4] **James Zou** [1 5] **John Janetzko** [6 4] **Rabindra V. Shivnaraine** [4 7] **Kyle Swanson** [1]

Abstract

Developing new fluorophores needed for advanced bioimaging techniques requires the exploration of previously unexplored chemical space. Generative AI approaches for the creation of novel dye scaffolds are promising in that they explore diverse regions of chemical space, but previous attempts have yielded synthetically intractable dye candidates due to the absence of reaction constraints, thus impeding experimental validation. Here, we present SyntheFluor, a generative AI model that employs known reaction libraries and molecular building blocks to create readily synthesizable fluorescent molecule scaffolds. SyntheFluor designed 11,590 molecules, which were filtered to a set of 19 diverse candidate molecules predicted to have dye-like properties. These 19 candidates were further examined by time-dependent density functional theory calculations, and 14 were successfully synthesized and 13 were experimentally validated. The photophysical properties of the three most fluorescent molecules were characterized in depth, and the top scaffold in particular showed robust fluorescence properties comparable to a known dye, demonstrating the utility of SyntheFluor.

[*]Equal contribution [1]Department of Computer Science, Stanford University, Stanford, California, USA [2]Curai Health, San Francisco, California, USA [3]Insitro, San Francisco, California, USA [4]Department of Molecular and Cellular Physiology, Stanford University, Stanford, California, USA [5]Department of Biomedical Data Science, Stanford University, Stanford, California, USA [6]Department of Biochemistry and Molecular Genetics, University of Colorado Anschutz Medical Campus, Denver, Colorado, USA [7]Greenstone Biosciences, Palo Alto, California, USA. Correspondence to: John Janetzko <john.janetzko@cuanschutz.edu>, Kyle Swanson <swansonk@stanford.edu>, Rabindra V. Shivnaraine <rvshiv@stanford.edu>.

Proceedings of the Workshop on Generative AI for Biology at the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

1. Introduction

Fluorescent dyes with highly optimized photophysical properties are critical for advanced imaging methods to detect and interrogate biomolecules (Suzuki et al., 2007; Datta et al., 2021). A variety of microscopy techniques, such as STORM (Stochastic Optical Reconstruction Microscopy), PALM (Photoactivated Localization Microscopy) (Lelek et al., 2021; Henriques et al., 2011), single-molecule FRET (Fluorescence Resonance Energy Transfer) (Sasmal et al., 2016), single-particle tracking (SPT), light-sheet microscopy (Santi, 2011), multi-photon fluorescence microscopy (Gratton et al., 2001), and fluorescence recovery after photobleaching (FRAP), have significantly enhanced our ability to investigate cellular structures and processes with improved spatio-temporal resolution. As these methodologies evolve, the demand for a diverse array of fluorescent dyes—each characterized by unique photophysical properties tailored to each specific imaging technique—becomes increasingly critical. For example, techniques such as SPT would benefit from dyes that are longer lived, whereas FRAP would benefit from dyes that are rapidly bleached.

Recently, machine learning (ML) models have emerged as a potentially transformative tool for *de novo* molecular design (Sousa et al., 2021). However, they often face challenges due to the generation of synthetically intractable scaffolds (Gao & Coley, 2020).

One approach that addresses this limitation is SyntheMol-RL (Swanson et al., 2025), a generative AI model that uses reinforcement learning (RL) to design easy-to-synthesize molecules with desirable properties. SyntheMol-RL assembles molecules from a set of molecular building blocks and well-established reactions available in the Enamine REAL Space (Grygorenko et al., 2020), which contains over 30 billion molecules that are easy to synthesize. Its architecture utilizes graph neural networks to predict molecular properties, and it dynamically weights these predicted properties to generate molecules with optimal combinations of properties. In its original form, SyntheMol-RL was applied to generate antibiotic candidates targeting *Staphylococcus aureus* with *in vitro* and *in vivo* validation of the generated molecules.

In this study, we developed SyntheFluor, a new version of SyntheMol-RL that assembles readily-synthesizable flu-

orescent molecule scaffolds. Our main contributions are outlined below.

1. To guide SyntheFluor, we train both graph neural network (GNN) and multilayer perceptron (MLP) property predictors on experimental measurements of three critical fluorescent properties: photoluminescence quantum yield (PLQY), absorption wavelength, and emission wavelength.

2. To expand SyntheFluor's chemical space, we introduce 57 new reactions relevant for fluorescent molecule design to complement the 13 existing reactions in SyntheMol-RL.

3. We design SyntheFluor to simultaneously optimize for four properties that are essential for fluorescence—PLQY, absorption wavelength, emission wavelength, and π-conjugated network size—compared to only two properties for SyntheMol-RL.

4. We applied SyntheFluor to generate 11,590 candidate molecules, of which we experimentally validated 13 molecules for their fluorescent properties. This resulted in the discovery of multiple diverse fluorescent compounds, including one with brightness comparable to a known dye.

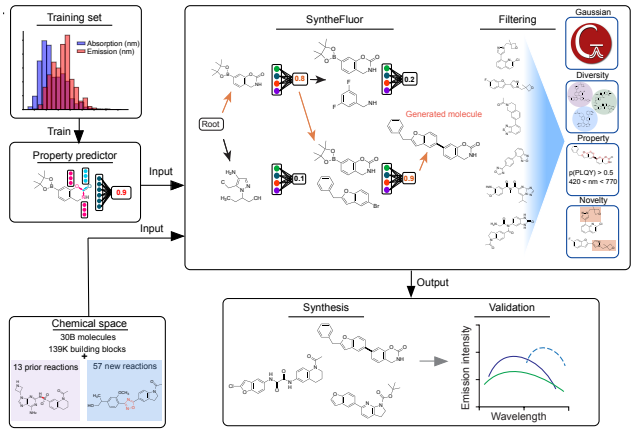


Figure 1. SyntheFluor pipeline overview. First, the training set is curated, followed by property predictor model development. Next, SyntheFluor generates molecules using the property predictors within a chemical space of synthesizable molecules. These molecules are then filtered based on TD-DFT calculations, structural diversity, predicted property thresholds, and novelty. Finally, selected molecules are synthesized and experimentally validated.

2. Related Work

Machine learning (ML) approaches have been used to identify promising candidates for fluorescent dyes. Prior work has successfully employed ML to accurately predict fluorescent properties, such as PLQY, absorption, and emission to identify molecules with potential fluorescence within databases of known compounds (Wang et al., 2021; Ye et al., 2020; Ju et al., 2021; Huang et al., 2024; Bu & Peng, 2023). Trained property predictors can then be used to identify promising fluorescent molecules from large databases (Wang et al., 2021). However, these methods are limited by the finite library sizes of existing chemical compounds, reducing the diversity of the resulting candidates.

In contrast, generative approaches, such as SyntheFluor, enable the design of novel and diverse compounds (Tan et al., 2023). Generative models have been gaining traction for *de novo* molecular design (Sousa et al., 2021). With access to a vast chemical space, these models allow for the discovery of novel scaffolds that do not currently exist in fluorophore libraries. However, while some generative AI approaches have been employed to discover novel fluorescent molecular structures (Sumita et al., 2022; Tan et al., 2023), few consider the fragments and reactions needed to develop synthesizable molecules. As a result, current generative approaches to molecule design often produce synthetically intractable scaffolds (Gao & Coley, 2020).

One notable generative approach to fluorescent molecule design used the model ChemTS to generate fluorescent dye candidates, of which one novel fluorophore scaffold was experimentally validated (Sumita et al., 2022). This compute-intensive approach (utilizing 1024 cores over 5 days to generate candidates) used random forest models and time-dependent density functional theory (TD-DFT) quantum chemistry calculations in the generative process.

SyntheFluor, unlike ChemTS and most generative models for small molecules, efficiently generates candidate molecules (in $<$ 24 hours with 32 cores) that are readily synthesizable, enabling the synthesis of a larger proportion of generated compounds and easing the transition from *in silico* design to experimental validation.

3. Methods

SyntheFluor generates fluorescent molecule candidates by using fluorescent property predictors to guide the exploration of an RL algorithm within a chemical space of easily synthesizable molecules (Figure 1). Below, we introduce the property predictors (Section 3.1) and the chemical space (Section 3.2) followed by the SyntheFluor generative model (Section 3.3), which uses RL to learn how to generate molecules to optimize a dynamically weighted combination of fluorescent properties. After generation, additional fluorescent properties are calculated via the Gaussian software, which runs physics simulations, to help filter the generated molecules (Section 3.4).

3.1. Property Prediction

SyntheFluor designs fluorescent molecules by optimizing for four fluorescent properties: PLQY, absorption wavelength, emission wavelength, and sp^2 network size. The first three are predicted using machine learning models while the last is calculated directly, as detailed below.

3.1.1. Neural Property Predictors

Two machine learning model architectures were designed to predict PLQY, absorption wavelength, and emission wavelength: (1) Chemprop (Yang et al., 2019), a graph neural network that processes molecular graphs and computed features, and (2) an MLP, which uses only computed molecular features. Two types of features were tested for both models: Morgan fingerprints, which encode local chemical structures, and RDKit features, comprising 200 physicochemical properties computed with RDKit (Team, 2024). In both cases, molecular features were augmented with four experimentally derived solvent properties—polarizability (SP), dipolarity (SdP), acidity (SA), and basicity (SB)—corresponding to the solvent used during experimental measurements. These features are relevant for assessing the solvent effect (Catalán, 2009). Figure 2 depicts the resulting four different property prediction architectures: (1) Chemprop-Morgan, (2) Chemprop-RDKit, (3) MLP-Morgan, and (4) MLP-RDKit.

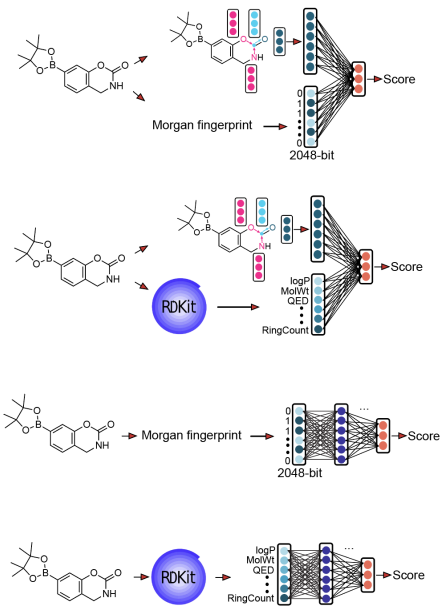


Figure 2. Visualizations of graph neural network architectures combined with either Morgan fingerprints or RDKit fingerprints (top two) and MLP architectures using Morgan fingerprints or RDKit fingerprints (bottom two).

The Chemprop-Morgan model consists of the Chemprop GNN model augmented with the Morgan fingerprint, which indexes the presence of specific substructures centered around each atom in the molecule. The model takes as input a molecular graph representation of each training molecule, with atoms as nodes and bonds as edges. The GNN aggregates features – such as atom and bond type for each atom and bond in the molecule – through three message passing steps, creating vector representations of local neighborhoods in the molecule within the neural network layers. The 300-dimensional vector representation is concatenated with Morgan fingerprints, which are 2,048 bits and were calculated with a radius of 2 using the cheminformatics package RDKit's `GetMorganFingerprintAsBitVect` function. Four numerical solvent features (SP, SdP, SA, SB) corresponding to polarizability, dipolarity, acidity, and basicity are also concatenated to the feature vector. The combined feature vector is passed through an MLP with one hidden layer, with a final activation that is a sigmoid for classification tasks (PLQY) or a linear layer for regression tasks (absorption, emission). The Chemprop-RDKit model consists of the same GNN, but instead of the Morgan fingerprint, 200 molecular features computed by RDKit and the four solvent features are appended to the 300-dimensional vector output from the GNN and input to the MLP layer. The MLP-Morgan and MLP-RDKit models have the same architecture as the MLP layer in the corresponding Chemprop models but do not include the GNN; thus, each model takes either a 2,052-dimensional feature vector (Morgan fingerprint with solvent features) or 204-dimensional feature vector (RDKit fingerprint with solvent features).

3.1.2. sp^2 Network Size Algorithm

The presence of a π-conjugated system is critical for a molecule to be fluorescent (Yamaguchi et al., 2008; Zhang et al., 2023), and larger π-conjugated systems reduce the HOMO-LUMO gap, shifting electronic transitions to the visible spectrum. To incorporate this information, we derive an sp^2 network size algorithm that utilizes a depth-first search (DFS) approach to calculate the size of the largest network of connected atoms with sp^2 hybridization in each molecule's molecular graph representation (see Appendix A).

3.2. Chemical Space

SyntheFluor is designed to generate molecules within the Enamine REAL Space, which contains over 30 billion molecules that can be readily synthesized in 3-4 weeks from combinations of around 139,000 molecular building blocks (Grygorenko et al., 2020). SyntheFluor combines these building blocks using a set of 70 chemical reactions, of which 13 were previously used by SyntheMol-RL and 57 are new to SyntheFluor. These 57 new reactions were specifically selected since many of them produce molecules with

extended aromatic systems – a critical feature for fluorescent dyes (Figure 3).

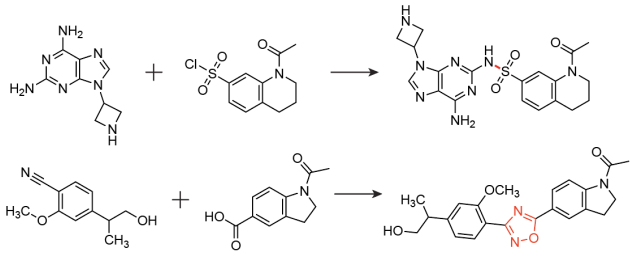


Figure 3. An example of a non-ring forming reaction in the original set of 13 reactions (top), and a ring-forming reaction in the extended set of 70 reactions (bottom).

3.3. SyntheFluor

We developed SyntheFluor to generate fluorescent molecule candidates (Figure 4). SyntheFluor retains the core generative process of SyntheMol-RL (Swanson et al., 2025), which uses an RL value function – implemented as either a Chemprop or MLP model – to guide its selection of molecular building blocks to form a molecule. The value function learns to compute the expected property score of one or more building blocks, allowing SyntheMol-RL to select combinations of building blocks that lead to promising full molecules.

3.3.1. RL Algorithm

SyntheFluor uses the SyntheMol-RL reinforcement learning algorithm to generate molecules. This algorithm takes a chemical synthesis tree T as input. Each node $N \in T$ has N_{mol}, which is a set of one or more molecular building blocks from a chemical space (e.g., the Enamine REAL Space). SyntheMol-RL defines a value function $V(N)$ on all nodes, which is a model that takes in the building blocks in a node's N_{mol} as input and outputs a prediction of the property score. During each rollout, $V(N)$ is applied to all nodes created at a given step, and nodes are sampled proportional to $e^{V(N)/\tau}$, where τ is a temperature parameter that can be tuned to affect the RL policy's exploration or exploitation. After a molecule m is constructed at the end of a rollout, it is scored by a weighted combination of L property predictors, M_k for $k \in \{1, ..., L\}$ with weights w_k for $k \in \{1, ..., L\}$ to obtain the molecule's overall property score, $p(m) = \sum_{k=1}^{L} w_k * M_k(m)$.

Extending from the original application of SyntheMol-RL to antibiotic design, which only optimized for two properties, SyntheFluor optimizes for four properties. The reward function of SyntheFluor evaluates the quality of full, generated molecules by scoring their fluorescent properties, and

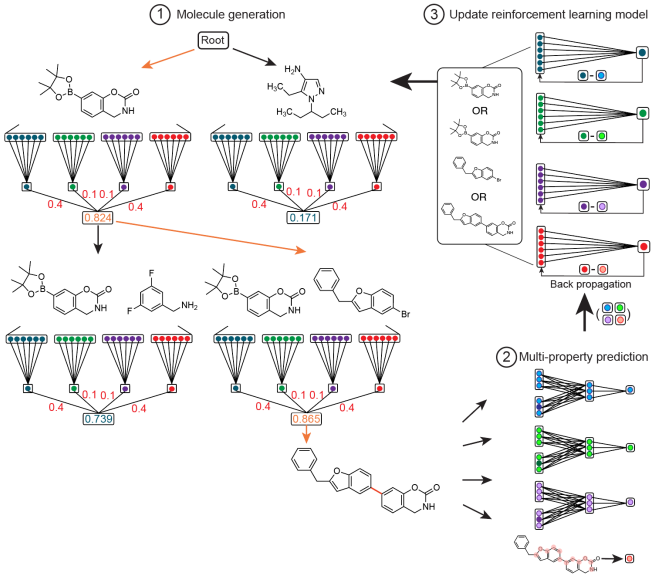


Figure 4. Schematic of the SyntheFluor reinforcement learning algorithm. Step 1 shows the selection of building blocks with intermediate RL value scoring conducted by four MLP-Morgan models (PLQY in blue, absorption in green, emission in purple, sp^2 in red), and the pairing of the final selected building blocks via Reaction 2718 to create the target molecule. Step 2 shows the evaluation of the candidate model via four reward models (Chemprop-Morgan for PLQY in light blue, absorption in light green, and emission in light purple, and the sp^2 algorithm for sp^2 network size in light red). Step 3 shows how the reward scores from Step 2 are used to update the corresponding RL value function MLP models in Step 1 and re-weight the building blocks for the next rollout.

these scores provide feedback to the RL value function to improve its ability to evaluate the quality of the molecule's component building blocks.

Following the RL-MLP version of SyntheMol-RL, we implement the value function using MLP models, and we implement the reward function using Chemprop models. We chose this version over the RL-Chemprop version of SyntheMol-RL, which uses Chemprop models for both value function and reward function, due to the increased speed of the MLP models with only a minor loss of accuracy.

Specifically, for the RL value function $V(N)$, we employed MLP-Morgan models for the four properties PLQY, absorption wavelength, emission wavelength, and sp^2 network size. $V(N)$ is a weighted combination of models where $V(N) = \sum_{k=1}^{L} w_k * Z_k(N_{mol})$, and $Z_1, ..., Z_L$ are the MLP-Morgan models ($L = 4$), and $w_1, ..., w_L$ are the same property weights used in the reward property score

$p(m)$. After each rollout, the algorithm stores tuples of $(N, M_1(m), ..., M_L(m))$ for every node N created along the path of nodes that ultimately led to molecule m, creating a training set of nodes and the property prediction scores of the final molecule created from them. The RL value models are trained at set intervals over the rollouts to predict property prediction scores for a generated molecule based on the building blocks of a node in its path using a mean squared error (MSE) loss, thus updating $V(N)$. For the reward function property predictors M_k, we utilized three Chemprop-Morgan models trained to predict PLQY, absorption wavelength, and emission wavelength, as well as the sp^2 network size algorithm (see Appendix A).

Notably, to ensure that fluorescence properties were assessed under relevant aqueous experimental conditions, all molecules evaluated by the MLP and Chemprop models within SyntheFluor were represented as a concatenation of their Morgan fingerprint with the four solvent features – SP, SdP, SA, and SB – corresponding to water.

3.3.2. Dynamic Weighting

SyntheFluor utilizes the SyntheMol-RL dynamic tuning mechanism to automatically adjust the RL temperature and property weights over time to optimize the generated molecules for both diversity and the four desired properties.

The RL temperature is important in defining the balance of exploration and exploitation and therefore the diversity of the generated molecules. The dynamic tuning method adjusts the RL temperature to obtain a molecular similarity among generated molecules of λ^* on average during generation. We set the RL temperature target similarity $\lambda^* = 0.6$, which means that on average the Tanimoto similarity of a newly generated molecule to the most similar previously generated molecule is 0.6.

Dynamic property weight tuning computes the average success rate on each rollout and adjusts the property weights based on the rolling average success rate. The success rate is determined by whether each generated molecule surpasses pre-determined success thresholds for each property.

3.4. Gaussian

To further estimate the fluorescent properties of some of the generated molecules, we used the software Gaussian to perform time-dependent density functional theory (TD-DFT) calculations (Jacquemin et al., 2011). We first used the B3LYP functional with the 3-21G* basis set to optimize the molecular geometry, facilitating convergence by bypassing eigenvalue checks, calculating the full force constant matrix, and setting the maximum number of optimization cycles to 1,000. Solvent effects were simulated using the Self-Consistent Reaction Field (SCRF) approach, with water modeled as the implicit solvent. After geometry optimization, TD-DFT computed the electronic excited states, specifically the first five singlet states.

These calculations provided excitation wavelengths, oscillator strengths, and dipole moments. The initial molecular coordinates were determined using a Merck molecular force field (MMFF) as calculated by RDKit, and the most stable conformations were selected. For each molecule, a Gaussian calculation was run on 12 CPUs per job, either until full optimization or for up to 48 hours. We reported the excited-state energy, oscillator strength, and dipole moment of the final optimization round.

4. Results

We applied SyntheFluor to generate candidate fluorescent molecules. Below, we introduce the dataset of known fluorescent compounds (Section 4.1), which we used to train the property prediction models (Section 4.2) that guide SyntheFluor. Then, we detail how we applied SyntheFluor to generate molecules (Section 4.3), followed by a series of filtering steps to select the most promising molecules (Section 4.4). Finally, we describe the synthesis and experimental validation of our top molecules (Section 4.5).

4.1. Fluorescence Dataset

The property prediction models were trained using the ChemFluor dataset (Ju et al., 2021). This dataset contains 2,912 unique molecules dissolved in 63 different solvents, resulting in 4,336 unique molecule-solvent pairs. In addition to the SMILES strings that correspond to each molecule and the solvent the molecule is dissolved in, the dataset also contains experimentally derived PLQY values, absorption spectra, and emission spectra as well as the solvent constants (SP, SdP, SA, and SB).

Since PLQY, absorption, and emission measurements were not available for all entries, three separate training sets were curated, one for each prediction task. All molecule-solvent entries that contained the SMILES string; SP, SdP, SA, and SB entries; and the relevant measurement (either PLQY value, absorption wavelength, or emission wavelength) were included in the relevant dataset. For the 50 duplicate molecule-solvent pairs, the measurement of interest was averaged across identical entries. This resulted in 3,055 molecule-solvent pairs with PLQY measurements, 4,202 molecule-solvent pairs with absorption wavelengths, and 4,333 molecule-solvent pairs with emission wavelengths.

4.2. Developing Property Prediction Models

Using this dataset, we trained four different property prediction architectures to predict each fluorescence property: (1) Chemprop-Morgan, (2) Chemprop-RDKit, (3) MLP-

Morgan, and (4) MLP-RDKit (all using the Chemprop package v1.6.1). Chemprop and MLP models were trained as either binary classifiers or regressors, depending on the task. PLQY prediction was modeled as a binary classification task using a threshold of $PLQY > 0.5$, while absorption and emission wavelength predictions were modeled as regression tasks. All models were trained using 10-fold cross-validation with an 80% training, 10% validation, and 10% testing split, completing in under 60 minutes on an 8-CPU machine. PLQY models were trained on 3,055 molecule-solvent pairs, absorption models on 4,202 pairs, and emission models on 4,333 pairs.

Across all three tasks (PLQY, absorption, and emission) and both the Chemprop and MLP model architectures, the Morgan fingerprints outperformed the RDKit features (Figure 5). The Chemprop-Morgan and MLP-Morgan architectures showed comparable performance on the PLQY classification task (Chemprop-Morgan: ROC-AUC = 0.895 $\pm$ 0.019 ; MLP-Morgan: ROC-AUC = 0.896 $\pm$ 0.019). Chemprop-Morgan demonstrated a slight advantage over MLP-Morgan for both absorption (Chemprop-Morgan MAE = 13.118 $\pm$ 1.203; MLP-Morgan MAE = 13.657 $\pm$ 1.083) and emission (Chemprop-Morgan MAE = 18.951 $\pm$ 0.986; MLP-Morgan MAE = 19.829 $\pm$ 1.268) regression tasks. Based on its superior performance across most tasks, the Chemprop-Morgan architecture was selected for the reward scoring function in the SyntheFluor generation process, while the MLP-Morgan architecture was selected for the RL value function due to it's faster speed.

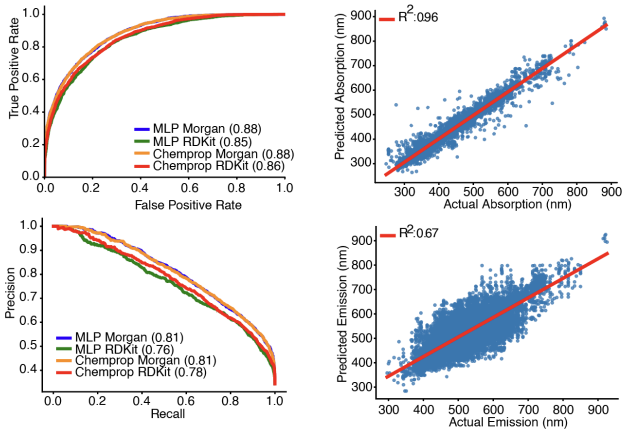


Figure 5. ROC curves and PRC curves for PLQY classification models (left). Predicted absorption versus actual absorption and predicted emission vs actual emission for the corresponding Chemprop-Morgan regression models (right).

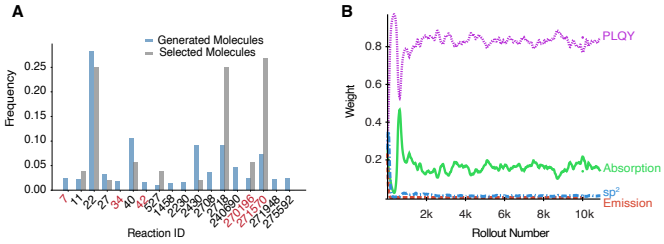


Figure 6. A) Histogram of reaction IDs used for generated molecules (blue) and selected molecules (gray). Reaction IDs in red are part of the extended set. B) Line plot showing the change in weight associated with each MLP model (PLQY, absorption, emission, and sp^2) in SyntheFluor's value function over the 10,000 rollouts.

4.3. Generating Molecules with SyntheFluor

We built SyntheFluor using the above property prediction models to guide its generation. For dynamic property weight tuning, we used the following success thresholds. PLQY: the probability of PLQY > 0.5 (the classification threshold) is at least 0.5 (i.e., $p(PLQY > 0.5) \geq 0.5$). Absorption: the predicted wavelength is within 420 nm to 750 nm (i.e., visible spectrum). Emission: the predicted wavelength is within 420 nm to 750 nm (i.e., visible spectrum). sp^2: the largest sp^2 network is ≥ 12 atoms.

We ran SyntheFluor for 10,000 rollouts, completing the generation process in 16 hours, 38 minutes, and 26 seconds and generating 11,590 candidate fluorescent molecules. These generated molecules used 18 unique reactions, five of which were from the new set of reactions (Figure 6A). Notably, SyntheFluor dynamically place the highest weight on the PLQY property throughout the generation process (Figure 6B).

To evaluate SyntheFluor's efficacy in generating molecules with optimized fluorescent properties, we compared the PLQY, absorption wavelengths, emission wavelengths, and sp^2 network sizes of the generated molecules against a random sample of 10,000 molecules from the Enamine REAL Space. Molecules generated by SyntheFluor had a higher probability of $PLQY > 0.5$ and larger sp^2 network sizes compared to the random sample while matching the overall absorption and emission wavelength distributions (Figure 7), thereby demonstrating that SyntheFluor successfully enriched for key fluorescent properties.

4.4. Filtering Generated Molecules

To identify the most promising candidates, we applied a multi-step filtering process. First, molecules with an sp^2 network size smaller than 12 were removed, excluding 5,479 molecules. Next, only molecules with $p(PLQY > 0.5) \geq$

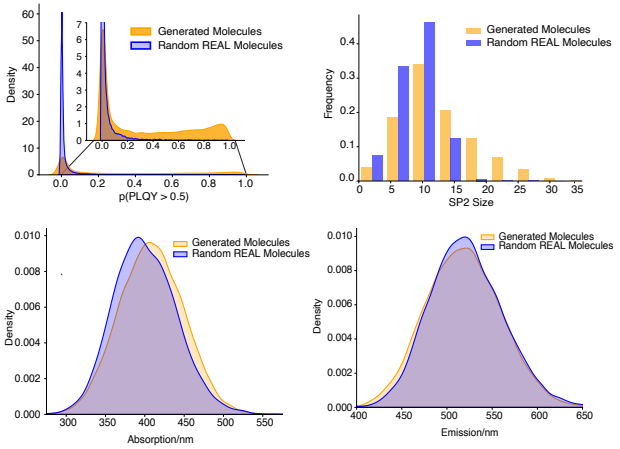


Figure 7. Distribution of PLQY probabilities (top left) and histogram of sp^2 network sizes (top right) on the generated molecules versus a random sample of 10,000 molecules in the REAL Space. Distribution of Chemprop-Morgan predicted absorption wavelengths (bottom left) and emission wavelengths (bottom right) on generated molecules versus the REAL Space random sample.

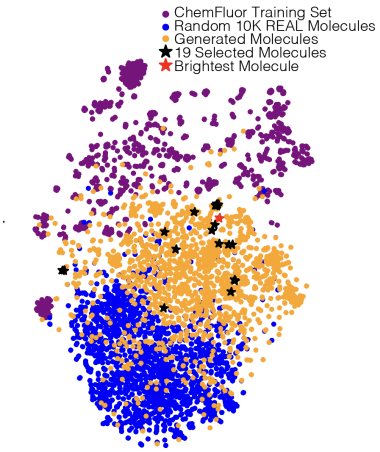


Figure 8. A t-SNE representation of ChemFluor training set molecules, randomly selected molecules in REAL Space, generated molecules, the final 19 selected molecules, and the brightest of validated compounds.

0.5 were retained, eliminating 4,256 molecules. Molecules with predicted absorption and emission wavelengths outside the visible range (420–750 nm) were also removed, excluding 21 molecules based on absorption and 1,203 based on emission. This left 631 molecules.

We evaluated the novelty of these molecules by calculating the Tanimoto similarity between each generated molecule and each ChemFluor molecule. Of the 631 generated molecules that passed the PLQY, absorption, emission, and sp^2 filtering steps, 630 had Tanimoto similarities < 0.5 with all ChemFluor molecules, indicating novel structures. Additionally, a qualitative comparison was performed using a t-SNE analysis on the Morgan fingerprints of these molecules with Tanimoto similarity as the distance metric and PCA for initialization, comparing 2,000 samples each from Enamine REAL molecules, the ChemFluor dataset, and SyntheFluor-generated molecules (Figure 8). The t-SNE plot revealed that the generated molecules occupy a novel chemical space. The fact that the generated molecules appear to occupy the subset of Enamine REAL space that is closest to the ChemFluor molecules indicates that they are more likely to possess fluorescent properties than random REAL molecules.

Next, to ensure that we test generated molecules with structural diversity, we grouped the 631 generated molecules into 100 clusters using K-means clustering on Morgan fingerprints using Tanimoto similarity. Then, we manually selected one molecule per cluster to maintain diversity, yielding 52 candidates. Of these, 34 (65%) were available for synthesis by Enamine.

To evaluate the potential fluorescence of the 34 candidate molecules, we used Gaussian to estimate the excitation wavelengths, oscillator strengths, and dipole moments. Gaussian-estimated excitation wavelengths correlated well with experimentally derived emission wavelengths from a subset of ChemFluor molecules (R^2= 0.63) (Supplementary Figure A1). We kept molecules with oscillator strengths above 0.1, where the oscillator strength quantifies the probability of absorption or emission in an electromagnetic transition. Of our 34 molecules, 19 passed the oscillator strength threshold and were thus selected for synthesis and subsequent experimental validation.

4.5. Experimental Characterization

Fourteen of the 19 candidates were successfully synthesized by Enamine (Kyiv, Ukraine), and their identity and purity were confirmed by LC-MS. However, one molecule was decomposed at the time of receipt, leaving 13 molecules for experimental testing. Excitation and emission scans were performed to determine the spectra, Ex_{max}, and Em_{max} for the compounds, and these values were compared to quinine sulfate as a reference standard (Drobnik & Yeargers, 1966) (Figure 9A; see Appendix B for detailed experimental methods).

Based on emission intensities, the three brightest compounds were identified as Compounds 13, 2, and 11, in descending order of brightness (Supplementary Figure A2). Compound 13 had the highest emission intensity, comparable to a known and regularly used fluorescent dye, quinine sulfate. Additionally, six additional compounds showed

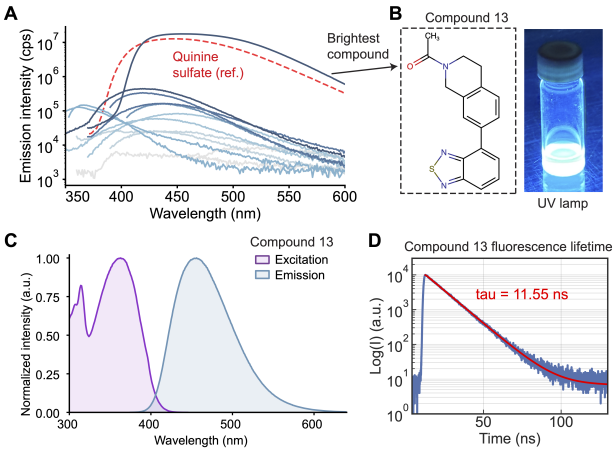


Figure 9. A) Emissions spectra of 13 synthesized molecules compared to the quinine sulfate standard at 10 mM. B) Structure of the brightest compound (13) at 10 mM in chloroform under UV lamp excitation. C) Normalized excitation and emissions spectrum of compound 13 (Ex_{max} = 363 nm, Em_{max} = 460 nm). D) Fluorescence lifetime of compound 13 obtained from time-correlated single photon counting. Red line is double exponential fit with mean lifetime of 11.55 ns.

fluorescence emission within two orders of magnitude of compound 13 (Figure 9A). Compound 13 was visibly the brightest upon UV excitation (Figure 9B); its excitation and emission spectra are shown in Figure 9C.

The top three fluorescent compounds were structurally diverse and each contained a distinct chromophore: a benzothiadiazole, a benzofuran, and an isoxazolopyridine. Fluorescence lifetime measurements were performed on these top three brightest compounds using a 10 MHz pulsed nanoLED at 405 nm, with emission recorded at the respective Em_{max} for each compound. Time-resolved fluorescence decay curves were tail-fitted to a bi-exponential model, yielding fluorescence lifetimes of 11.55 ns for Compound 13 (Figure 9D), 1.8 ns for Compound 2, and 1.5 ns for Compound 11 (Supplementary Figure A2, Supplementary Figure A3).

Our experimental validation identified three promising, structurally diverse scaffolds, one of which fluoresced in a range comparable to a known dye. Additionally, six of our 13 chemically stable dyes exhibited fluorescence within two orders of magnitude of our brightest compound, indicating SyntheFluor's utility in identifying promising, diverse candidate scaffolds for fluorescent molecules.

5. Discussion

In this study, we developed SyntheFluor, a generative AI model capable of generating diverse, readily-synthesizable fluorescent dye scaffolds from a vast chemical space of over 30 billion molecules. SyntheFluor's synthesis-aware approach, in which scaffolds can be accessed from readily-available building blocks and established reactions, enables facile experimental validation. We showed that three SyntheFluor-generated molecules absorbed UV light and fluoresced in the visible range. These scaffolds had diverse chromophores, demonstrating the potential of SyntheFluor for the design of new fluorescent molecules.

To accomplish our goals, we needed SyntheFluor to robustly predict fluorescent properties. Because reinforcement learning value models require a model architecture that is differentiable, we exclusively experimented with neural property predictor architectures, including graph neural networks and MLPs.

Since PLQY, absorption, and emission properties depend on solvent, it was necessary to encode solvent features in our property predictor models. Doing so increased model expressivity and allowed us to train our model on more molecules, given that there exists no standardized experimental training data. Future improvement is likely to come from curating a larger fluorescence dataset, particularly one with a wider range of absorption and emission wavelengths.

Central to SyntheFluor's development from the core SyntheMol-RL algorithm is the adaptation of the multi-parameter objective from two antibiotic properties to four fluorescent properties: PLQY, absorption wavelength, emission wavelength, and sp^2 network size.

Optimizing for these four parameters ensured that the candidate molecules possess the structural and electronic properties necessary for fluorescence. Additionally, to increase the number of promising candidate scaffolds generated by SyntheFluor, we incorporated 57 new reactions, many of which have the potential to expand sp^2 networks. Eight of the fourteen molecules that were successfully synthesized used a reaction from this extended set, indicating the importance of including these reactions.

We selected 19 diverse molecules for experimental validation, 14 of which were successfully synthesized, and 13 of which were chemically stable. Of these 13, one compound (compound 13) was by far the brightest; however, 6 additional compounds showed fluorescence emission within two orders of magnitude of compound 13. Notably, the chromophores of the three most strongly fluorescent molecules were all different. From brightest to dimmest, these molecules contained a benzothiadiazole, a benzofuran, and an isoxazolopyridine, respectively. These three molecules all fluoresce in a similar spectral region, but their lifetimes span an order of magnitude range. While benzofuran- and benzothiadiazole-based fluorophores have been previously described, these two derivatives have not

previously been synthesized and evaluated for their fluorescent properties (Belmonte-Vázquez et al., 2019; Chen et al., 2023; Neto et al., 2022; Niu et al., 2015). Both scaffolds have been seen to be highly tunable and photostable, which has resulted in their use in various bioimaging modalities. Importantly, owing to the design of SyntheFluor, it is straightforward to obtain derivatives of these experimentally validated molecules for further optimization.

SyntheFluor is the first generative AI model to design diverse, readily-synthesizable fluorescent scaffolds, enabling an easy path from AI-driven molecular design to experimental validation. Future work will enhance both its property predictors via an extended training set and its building block and reaction scoring capabilities for better fluorescent molecule design and tunability.

Impact Statement

This paper presents work aimed at advancing the field of Machine Learning. The underlying architecture of SyntheFluor, SyntheMol-RL, carries potential societal implications. SyntheMol-RL can be dual-use: it may contribute positively by aiding the development of useful molecules such as fluorescent dyes or antibiotics, but it also carries the risk of misuse, such as the generation of toxic compounds. That said, we do not believe this specific paper raises any immediate societal concerns. The creation of fluorescent compounds is a well-studied area, with primary applications in industrial and academic research.

Acknolwedgement

J.J. acknowledges funding support from NIH (K99GM147609) and the Damon Runyon Cancer Research Foundation (DRG-2318-18). K.S. acknowledges the support of the Knight-Hennessy Scholarship and the Stanford Bio-X Fellowship.

References

Belmonte-Vázquez, J. L., Avellanal-Zaballa, E., Enríquez-Palacios, E., Cerdán, L., Esnal, I., Bañuelos, J., Villegas-Gómez, C., Lopez Arbeloa, I., and Peña-Cabrera, E. Synthetic approach to readily accessible benzofuran-fused borondipyrromethenes as red-emitting laser dyes. *The Journal of Organic Chemistry*, 84(5):2523–2541, 2019.

Bu, Y. and Peng, Q. Designing promising thermally activated delayed fluorescence emitters via machine learning-assisted high-throughput virtual screening. *The Journal of Physical Chemistry C*, 127(49):23845–23851, 2023.

Catalán, J. Toward a generalized treatment of the solvent effect based on four empirical scales: dipolarity (sdp, a new scale), polarizability (sp), acidity (sa), and basicity (sb) of the medium. *The Journal of Physical Chemistry B*, 113(17):5951–5960, 2009.

Chen, S.-H., Cao, X.-Y., Hu, P.-T., Jiang, K., Liang, Y.-T., Xu, B.-J., Li, Z.-H., and Wang, Z.-Y. Full-color emission of fluorinated benzothiadiazole-based d–a–d fluorophores and their bioimaging applications. *Materials Advances*, 4(24):6612–6620, 2023.

Datta, R., Gillette, A., Stefely, M., and Skala, M. C. Recent innovations in fluorescence lifetime imaging microscopy for biology and medicine. *Journal of Biomedical Optics*, 26(7):070603–070603, 2021.

Drobnik, J. and Yeargers, E. On the use of quinine sulfate as a fluorescence standard. *Journal of Molecular Spectroscopy*, 19(1):454–455, 1966. ISSN 0022-2852. doi: https://doi.org/10.1016/0022-2852(66)90267-0. URL https://www.sciencedirect.com/science/article/pii/0022285266902670.

Gao, W. and Coley, C. W. The synthesizability of molecules proposed by generative models. *Journal of chemical information and modeling*, 60(12):5714–5723, 2020.

Gratton, E., Barry, N. P., Beretta, S., and Celli, A. Multiphoton fluorescence microscopy. *Methods*, 25(1):103–110, 2001.

Grygorenko, O. O., Radchenko, D. S., Dziuba, I., Chuprina, A., Gubina, K. E., and Moroz, Y. S. Generating multibillion chemical space of readily accessible screening compounds. *Iscience*, 23(11), 2020.

Henriques, R., Griffiths, C., Hesper Rego, E., and Mhlanga, M. M. Palm and storm: unlocking live-cell super-resolution. *Biopolymers*, 95(5):322–331, 2011.

Huang, W., Huang, S., Fang, Y., Zhu, T., Chu, F., Liu, Q., Yu, K., Chen, F., Dong, J., and Zeng, W. Ai-powered mining of highly customized and superior esipt-based fluorescent probes. *Advanced Science*, 11(35):2405596, 2024.

Jacquemin, D., Mennucci, B., and Adamo, C. Excited-state calculations with td-dft: From benchmarks to simulations in complex environments. *Physical Chemistry Chemical Physics*, 13(38):16987, 2011. doi: 10.1039/c1cp22144b.

Ju, C.-W., Bai, H., Li, B., and Liu, R. Machine learning enables highly accurate predictions of photophysical properties of organic fluorescent materials: Emission wavelengths and quantum yields. *Journal of Chemical Information and Modeling*, 61(3):1053–1065, 2021. doi: 10.1021/acs.jcim.0c01203. URL https://doi.org/10.1021/acs.jcim.0c01203. PMID: 33620207.

Lelek, M., Gyparaki, M. T., Beliu, G., Schueder, F., Griffié, J., Manley, S., Jungmann, R., Sauer, M., Lakadamyali, M., and Zimmer, C. Single-molecule localization microscopy. *Nature reviews methods primers*, 1(1):39, 2021.

Neto, B. A., Correa, J. R., and Spencer, J. Fluorescent benzothiadiazole derivatives as fluorescence imaging dyes: a decade of new generation probes. *Chemistry–A European Journal*, 28(4):e202103262, 2022.

Niu, G., Liu, W., Wu, J., Zhou, B., Chen, J., Zhang, H., Ge, J., Wang, Y., Xu, H., and Wang, P. Aminobenzofuran-fused rhodamine dyes with deep-red to near-infrared emission for biological applications. *The Journal of Organic Chemistry*, 80(6):3170–3175, 2015.

Santi, P. A. Light sheet fluorescence microscopy: a review. *Journal of Histochemistry & Cytochemistry*, 59(2):129–138, 2011.

Sasmal, D. K., Pulido, L. E., Kasal, S., and Huang, J. Single-molecule fluorescence resonance energy transfer in molecular biology. *Nanoscale*, 8(48):19928–19944, 2016.

Sousa, T., Correia, J., Pereira, V., and Rocha, M. Generative deep learning for targeted compound design. *Journal of Chemical Information and Modeling*, 61(11):5343–5361, 2021.

Sumita, M., Terayama, K., Suzuki, N., Ishihara, S., Tamura, R., Chahal, M. K., Payne, D. T., Yoshizoe, K., and Tsuda, K. De novo creation of a naked eye–detectable fluorescent molecule based on quantum chemical computation and machine learning. *Science Advances*, 8(10):eabj3906, 2022.

Suzuki, T., Matsuzaki, T., Hagiwara, H., Aoki, T., and Takata, K. Recent advances in fluorescent labeling techniques for fluorescence microscopy. *Acta histochemica et cytochemica*, 40(5):131–137, 2007.

Swanson, K., Liu, G., Catacutan, D. B., McLellan, S., Arnold, A., Tu, M. M., Brown, E. D., Zou, J., and Stokes, J. Synthemol-rl: a flexible reinforcement learning framework for designing novel and synthesizable antibiotics. *bioRxiv*, 2025. doi: 10.1101/2025.05.17.654017. URL https://www.biorxiv.org/content/early/2025/05/17/2025.05.17.654017.

Tan, Z., Li, Y., Wu, X., Zhang, Z., Shi, W., Yang, S., and Zhang, W. De novo creation of fluorescent molecules via adversarial generative modeling. *RSC Advances*, 13(2): 1031–1040, 2023.

Team, T. R. D. Rdkit: Open-source cheminformatics, 2024. URL https://www.rdkit.org/. Accessed: 2024-12-15.

Wang, Y., Cai, L., Chen, W., Wang, D., Xu, S., Wang, L., Kononov, M. A., Ji, S., and Xian, M. Development of xanthene-based fluorescent dyes: Machine learning-assisted prediction vs. td-dft prediction and experimental validation. *Chemistry-Methods*, 1(8):389–396, 2021.

Yamaguchi, Y., Matsubara, Y., Ochi, T., Wakamiya, T., and Yoshida, Z.-i. How the π conjugation length affects the fluorescence emission efficiency. *Journal of the American Chemical Society*, 130(42):13867–13869, 2008.

Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., et al. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, 59(8):3370–3388, 2019.

Ye, Z.-R., Huang, I.-S., Chan, Y.-T., Li, Z.-J., Liao, C.-C., Tsai, H.-R., Hsieh, M.-C., Chang, C.-C., and Tsai, M.-K. Predicting the emission wavelength of organic molecules using a combinatorial qsar and machine learning approach. *RSC advances*, 10(40):23834–23841, 2020.

Zhang, J., Chen, M., Ren, X., Shi, W., Yin, T., Luo, T., Lan, Y., Li, X., and Guan, L. Effect of conjugation length on fluorescence characteristics of carbon dots. *RSC advances*, 13(40):27714–27721, 2023.

Appendix

A. sp2 Algorithm

Algorithm 1 sp^2 Network Size

Input: Molecule, represented as an adjacency matrix
Output: Size of the largest connected component of sp^2 atoms

Compute $sp2\_atom\_idxs \leftarrow$ indices of all sp^2 atoms in the molecule
Compute $sp2\_neighbors[atom\_idx] \leftarrow$ sp^2 neighbors of each sp^2 atom
$visited\_global \leftarrow \emptyset$
$max\_count \leftarrow 0$

function DFS($atom\_idx, visited\_local$)
  Append $atom\_idx$ to $visited\_local$
  Append $atom\_idx$ to $visited\_global$
  for each $neighbor\_idx$ in $sp2\_neighbors[atom\_idx]$ **do**
    if $neighbor\_idx \notin visited\_local$ **then**
      DFS($neighbor\_idx, visited\_local$)
    end if
  end for
  return len($visited\_local$)
end function

for each $atom\_idx$ in $sp2\_atom\_idxs$ **do**
  if $atom\_idx \notin visited\_global$ **then**
    $component\_size \leftarrow$ DFS($atom\_idx, \emptyset$)
    $max\_count \leftarrow \max(max\_count, component\_size)$
  end if
end for

Output: max_count

B. Experimental Validation Measurements

B.1. Measurements of excitation, emission, and fluorescence lifetimes

Excitation and emission spectra were collected using a Fluorolog 3 spectrofluorometer (Horiba Jobin Yvon). The spectra of 10 mM of each molecule in chloroform were acquired through 1 nm increment wavelength scans with excitation and emission slit widths set to 4 nm and a 0.1 s integration time. For excitation, a tungsten lamp served as the source, while photon collection was obtained using a photomultiplier tube (PMT) and recorded as counts per second (CPS).

Fluorescence lifetimes were measured using a Horiba time-correlated single-photon counting (TCSPC) unit, equipped with a nanoLED405LH (pulse duration: 705 ps) operated at a repetition rate of 1 MHz. Emission signals were collected using a PMT across 4,096 channels with a total time span of 200 ns (0.055 ns/channel). The instrument response function (IRF) was recorded using a 1000-fold dilution of the Ludox-40 scattering solution obtained from Sigma, generating a FWHM of 660 ps. Fluorescence decay profiles were tail fit to a double exponential model in Python to determine the amplitude-weighted mean fluorescence lifetime (τ).

B.2. Quantum yield and molar extinction coefficient

The fluorescent quantum yield (Φ_f), defined as the ratio of photons emitted to photons absorbed by a fluorescent molecule, was determined using the relative method with quinine sulfate as the fluorescence standard ($\Phi_{std} = 0.62$). Emission spectra were collected for 20 μM quinine sulfate in 0.1 M sulfuric acid and 20 μM compound X in chloroform, using a 330 nm

excitation wavelength. The absorbance of both solutions at 330 nm was approximately identical (∼0.077). The relative quantum yield was calculated from the integrated emission spectra using the formula: $\Phi_f = \Phi_{std} \times \frac{I}{I_{std}} \times \frac{n^2}{n_{std}^2}$, where I and I_{std} are the integrated fluorescence intensities of the sample and standard, respectively, and n and n_{std} are the refractive indices of the solvents. Measurements were performed under identical conditions to ensure an accurate comparison with the reference standard.

The molar extinction coefficient (ε) was determined using the Beer-Lambert law: $A = \varepsilon cl$, where A is the maximum absorbance, c is the molar concentration of the solution and l is the path length of the cuvette (1 cm). Absorbance was measured using a Beckman DU 640 spectrophotometer with a 1 cm quartz cuvette and compound titrations ranging from 20 to 500 μM. ε was obtained from the slope of the concentration versus absorbance data using linear regression (y-intercept = 0).

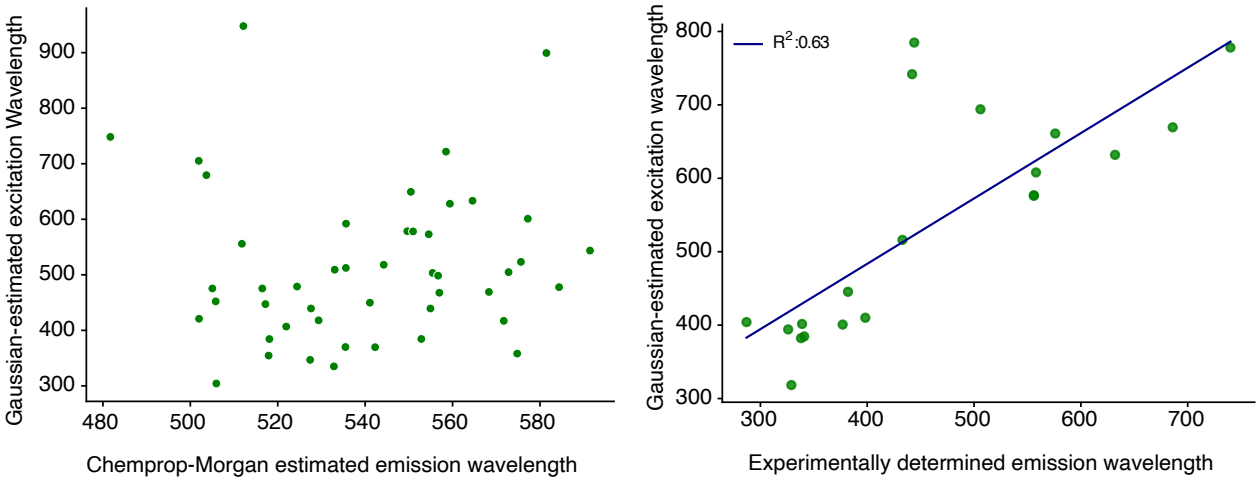
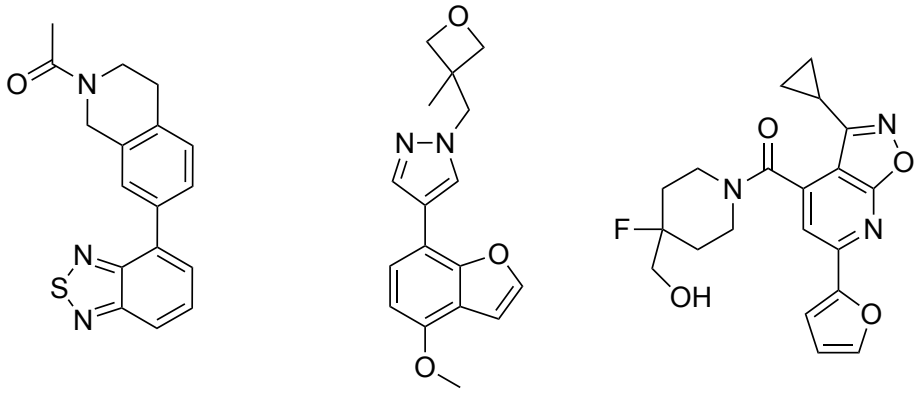


Figure A1. Plots of Gaussian-estimated excitation wavelength versus Chemprop-Morgan estimated emission wavelength among a subset of generated molecules (left) and of Gaussian-estimated excitation wavelength versus experimentally-determined emission wavelength among a subset of ChemFluor molecules (right).



	Cmpd 13	Cmpd 2	Cmpd 11
Brightness (cps):	17695410	445020	338560
Lifetime (ns):	11.5	1.8	1.5
Quantum yield:	0.62	*N.D.*	*N.D.*
Molar extinction coefficient (ε) ($L \cdot mol^{-1} \cdot cm^{-1}$):	6000	*N.D.*	*N.D.*

Figure A2. Fluorescent properties of the three brightest compounds generated by SyntheFluor.

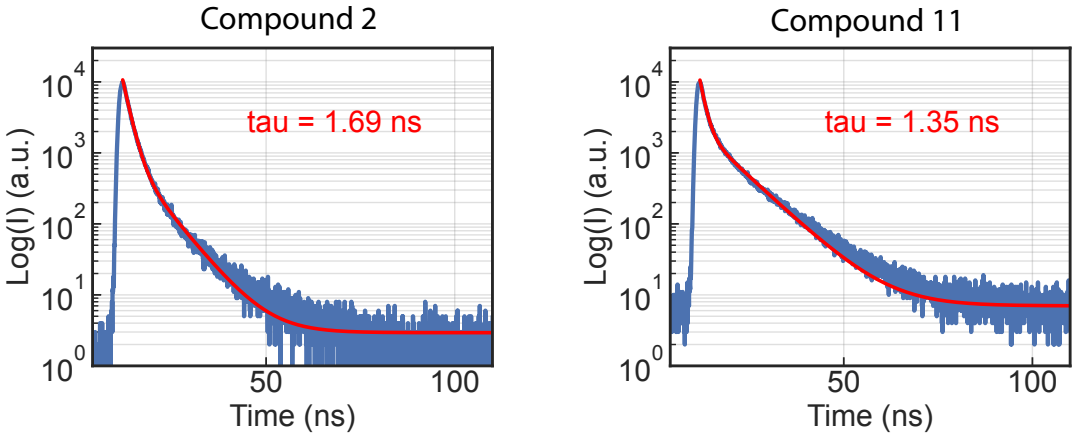


Figure A3. Fluorescence lifetime decay profiles of SyntheFluor compounds. Decay profiles are fit to a double exponential decay function (red curve) with amplituded-weighted mean lifetime tau.

Data and Code Availability

All the data and code are available here: https://drive.google.com/file/d/1C99rnq0_PftLi4l8u3GH_CjdplHegCOe

Supporting Information Available

Remaining supplementary figures and tables available here: https://drive.google.com/drive/folders/1ROjS31ED06oJ4em30prLLiHWvnff78in

Figure S6: UV/HPLC and mass spectrometry analysis of compound 13

Figure S7: UV/HPLC and mass spectrometry analysis of compound 2

Figure S8: UV/HPLC and mass spectrometry analysis of compound 11

Table S1: Metrics for model performance on PLQY classification (Excel spreadsheet)

Table S2: 13 synthesized compounds ordered by brightness (CPS) (Excel spreadsheet)