DEEP LEARNING-GUIDED DESIGN OF FLUORESCENT SMALL MOLECULES

Clare Phelps
Department of Computer Science
Stanford University
clphelps@stanford.edu

Heegwang Roh
Department of Chemistry
Stanford University
hgroh@stanford.edu

Finsam Samson
Department of Computer Science
Stanford University
finsam@stanford.edu

Srikant Sagireddy
Department of Chemical Engineering
Stanford University
srikant@stanford.edu

ABSTRACT

The design of novel fluorescent small molecules with improved properties is critical for advancements in biological imaging and diagnostics. Traditional approaches have predominantly relied on modifying existing scaffolds, limiting the exploration of the vast chemical space. In this study, we leverage a deep learning-guided generative model, SyntheMol, to design new fluorescent molecules with enhanced brightness and stability. We develop various deep learning-based property prediction models to predict the photoluminescence quantum yield (PLQY) given a molecule's chemical structure, which is then used in conjunction with SyntheMol to generate promising candidate molecules. We apply chemical intuition for filtering generated molecules based on conjugation levels and aqueous solubility, resulting in a set of promising candidates with unique scaffolds different from classical fluorescent molecules. Our approach demonstrates the potential of deep learning-guided generative models in the design of novel fluorescent molecules, paving the way for future experimental validation and application in biological research.

1 Introduction

The development of new chemical tools has been a key driver of many significant biological discoveries, and the use of small molecule fluorescent probes is a prime example. Fluorescent probes are essential in various aspects of biology, such as tracking individual biomolecules, labeling subcellular organelles, monitoring cellular activity, and diagnosing diseases [1]. Despite notable advancements in the field, there is still a critical need for fluorescent probes with better physical properties [2]. Improving dye brightness can enable the identification of biomolecules with lower limit-of-detection and enhance super-resolution imaging with lower spatial resolution. Similarly, dyes with higher stability (i.e., resistant to photobleaching) can enable longer-term real-time imaging and tracking, facilitating the study of biological phenomena over an extended timescale.

We apply a deep-learning-based approach to tackle this challenge. Specifically, we trained molecular property prediction models for a dye's brightness and stability, given its chemical structure, using a pre-existing database of fluorescent probes [3]. We subsequently leveraged SyntheMol, a recently-developed generative model for molecules of desired property and synthetic availability, to synthesize candidate fluorescent dyes.

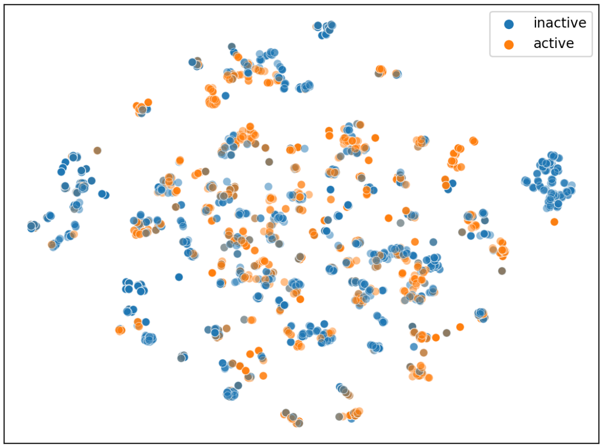


Figure 1: tSNE embedding of training set.

2 Methods

2.1 Data

Data were collected from the Chemfluor dataset released by Ju *et al.* [4]. The dataset contains 4386 samples, with photoluminescence quantum yields (PLQY) for approximately 3000 small molecules represented as SMILES strings. PLQY is defined as the ratio between the number of photons emitted and the number of photons absorbed by a small molecule. All PLQY values in the dataset are between 0 and 1. The dataset also includes features to describe the solvent the PLQY measurements were obtained in. This results in duplicate SMILES strings in the dataset. In this work, we do not include the solvent features, but show in the next section that it may help improve model performance. We also do not remove any duplicate SMILES strings from the dataset to ensure a sufficiently large sample size, but this may result in noisy activity labels. We binarize the PLQY values, such that molecules with a PLQY greater than or equal to 0.3 are considered active, and molecules with PLQY less than 0.3 are considered inactive. This resulted in 46% of the molecules in the dataset being considered active, and 54% of the molecules being considered inactive. We perform a tSNE visualization of our training dataset in Fig. 1.

2.2 Model architectures

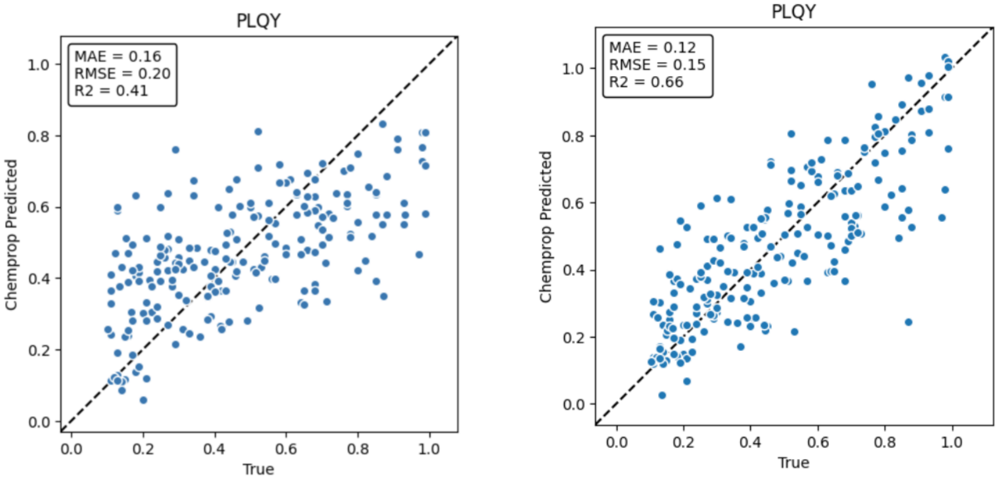
Our team has designed a total of six models to predict the PLQY (photoluminescence quantum yield) activity of molecules: two regression models and four classification models. These models are based on the message passing neural network called Chemprop [3], and a random forest model. The primary goal of these models is to accurately determine the photoluminescent behavior of molecules, enabling the efficient design and analysis of new materials with desired properties.

The regression models utilize Simplified Molecular Input Line Entry System (SMILES) strings as input data. SMILES strings are a compact and efficient way to represent molecular structures and their specific arrangements of atoms and bonds [5]. By analyzing these strings, the regression models are able to output the predicted PLQY value of a given molecule.

On the other hand, the classification models also use SMILES strings as input, but focus on predicting the probability that the input SMILES string will display active photoluminescent behavior ($PLQY > 0.3$). This probability gives researchers an idea of how likely it is that a molecule will demonstrate desirable photoluminescent properties.

The message passing neural network, Chemprop, serves as the foundation for these models. With the trained property predictors, we use SyntheMol [6], a generative model, to find molecules with high PLQY values. The models are described in detail below.

Chemprop. Chemprop works by iteratively passing messages between atoms in a molecule, effectively learning the local chemical environment and aggregating this information. This message-passing architecture allows Chemprop to learn meaningful molecular features essential for predicting the photoluminescent behavior of different molecules.



(A) SMILES only

(B) SMILES + solvent features

Figure 2: Chemprop regression PLQY predictor results.

Chemprop regression using only SMILES strings as input to predict PLQY performs reasonably well ($R^2 = 0.41$) (Fig. 2(A)). We also found that including solvent features (polarity, polarizability, dipolarity, acidity, and basicity) can improve the PLQY prediction accuracy ($R^2 = 0.66$) (Fig. 2(B)).

Chemprop + RDKit. The Chemprop + RDKit model is similar to the Chemprop model described above. After the message passing step in Chemprop obtains a feature vector of the molecule, we add 200 molecular features computed by RDKit. This new, expanded feature vector is used to predict the properties of the molecule. This gives the model information about global features that may be important to fluorescence.

Chemprop + Morgan. The Chemprop + Morgan model is similar to the Chemprop + RDKit model described above. However, instead of adding the 200 descriptors from RDKit, we add Morgan fingerprints, radius 2 and 2048 bits, to the feature vector obtained from the message passing steps. Since the Morgan fingerprints capture the local environment surrounding each atom, including ring structure and bond types, we hope this helps the model learn the importance of conjugated ring structures for fluorescence.

Random Forest. The random forest model uses the Morgan fingerprints, radius 2 and 2048 bits, as inputs to a random forest classifier with 100 decision trees. The prediction of the model is an average of the predictions of the 100 trees.

SyntheMol. SyntheMol is a generative model that outputs synthesizable molecules with desired molecular properties using a molecular property predictor along with a noisy Monte Carlo search of molecules that can be synthesized from a set of initial building blocks and reactions. The above models (Chemprop variants and Random Forest) can be used as molecular property predictors in SyntheMol.

2.3 Model training

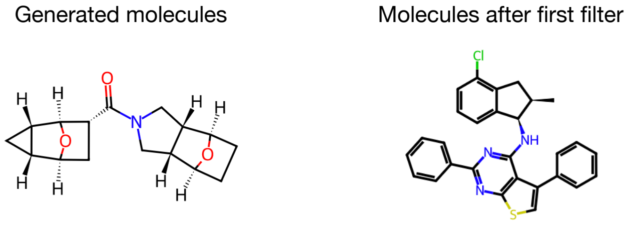
All models we trained used a random 80/10/10 train/validation/test split for each of five cross-validation folds. The models were trained on the train set and evaluated on the test set. The Chemprop models were trained for 30 epochs using the Adam optimizer. For the classification models, a binary cross-entropy loss was used for training. The classification models were evaluated using both the area under the receiver operating characteristic curve (ROC-AUC) and area under the precision recall curve (PRC-AUC), as seen in Table 1. For the regression models, a mean-squared error loss was used for training. The regression models were evaluated using the coeffecient of determination (R^2). When using the models for generating molecules with SyntheMol, we compute the average score of the ensemble of five models, from the five cross-validation folds.

2.4 Generating molecules

We applied SyntheMol with the PLQY property predictors described above to discover potentially fluorescent small molecules. The models took 9 to 10 hours to generate molecules with 20,000 rollouts. Table 2 contains additional information about parameters used when generating molecules and the number of molecules generated.

Table 1: Summary of test performance metrics for each of the four classification models.

Model	Chemprop	Chemprop + RDKit	Chemprop + Morgan	Random Forest
ROC-AUC	0.822 ± 0.005	0.835 ± 0.042	0.861 ± 0.001	0.868 ± 0.022
PRC-AUC	0.700 ± 0.009	0.739 ± 0.012	0.793 ± 0.003	0.816 ± 0.018


Generated molecules

Molecules after first filter

Figure 3: Representative structure of generated molecules before (left) and after (right) filtering for conjugation.

After generating molecules, our goal was to select three molecules that could be experimentally validated. We had two criteria for selection: (1) the molecule had to contain a certain degree of conjugation, and (2) the molecule had to be water-soluble. The first criterion stems from the fact that the majority of known fluorescent small molecules contain conjugated fragments in their structure. The second criterion stems from the application of fluorescent small molecules in biological imaging, which are mainly aqueous environments.

2.5 Filtering for conjugation

To ensure that the generated molecules had a sufficient level of conjugation, we only select molecules with greater than or equal to 24 aromatic atoms in their structure. This corresponds to roughly 4 conjugated ring structures in each generated molecule. After applying this filter, we find that very few of the generated molecules actually meet this requirement (Table 2).

Table 2: Summary of parameters and results from generating molecules with SyntheMol.

Model	Chemprop	Chemprop + RDKit	Chemprop + Morgan	Random Forest	Regression
Rollouts	20,000	20,000	20,000	20,000	20,000
max_reactions	1	1	1	1	5
Molecules	24,340	23,765	23,949	24,125	57,724
Conjugated molecules	13	29	155	4	7,585

To understand why, we visualize the structures of the generated molecules in Fig. 3. We see that before filtering, many of the molecules contain three- and five-membered ring structures without conjugation. After filtering for conjugation, we select molecules with more traditional five- and six-membered conjugated ring structures. This suggests that the model is able to capture the importance of ring structures in fluorescent molecules, but it does not learn the importance of conjugation in fluorescent molecules. This is an important area of improvement since our chemical intuition tells us that conjugation is highly influential on the fluorescent properties of molecules.

2.6 Aqueous solubility classifier

To find molecules that would likely be more suitable for biological environments, we implemented a Chemprop solubility classifier using AqSolDB, a dataset containing solubility information for approximately 9000 unique compounds, as a training set [7]. Solubility is defined as LogS, which is the base-10 log of the solubility (mol/L) of a compound in a solvent. For our classifier, LogS ≥ -2 was defined as water soluble (1), and LogS < -2 was defined as water insoluble (0). The solubility classifier uses SMILES strings and RDKit descriptors to make its predictions (Table 3).

Table 3: Solubility classifier performance (5-fold cross-validation).

Model	**Chemprop + RDKit**
ROC-AUC	0.958 ± 0.002
PRC-AUC	0.934 ± 0.006

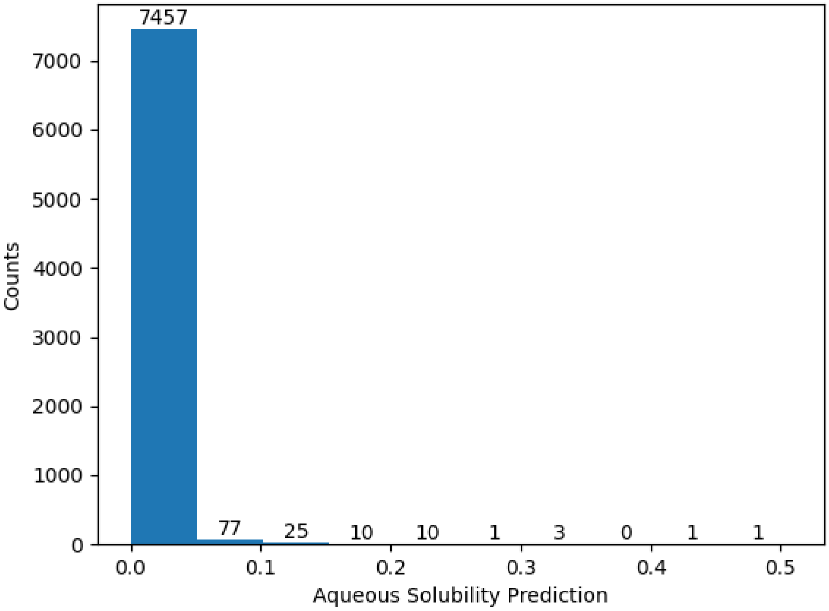


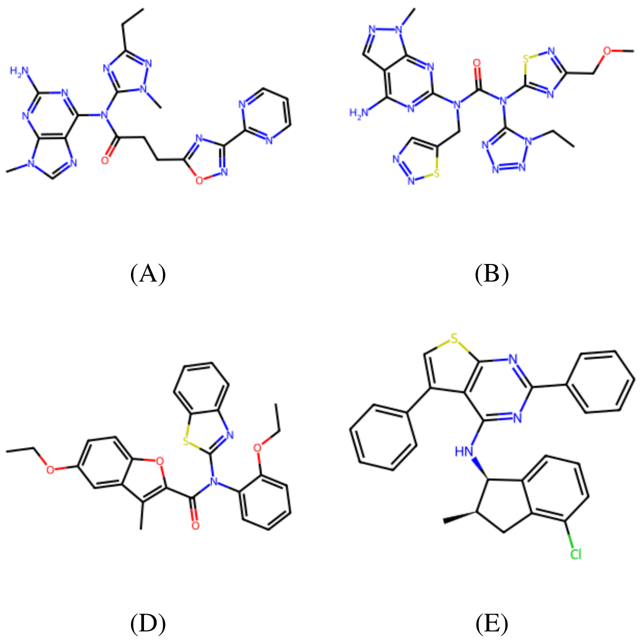
Figure 4: Distribution of Solubility Predictions of Generated Molecules.

Most of the generated molecules are predicted to have very low water solubilities. For example, after filtering the molecules generated using the Chemprop regression model (yielding 7585 sufficiently conjugated molecules) and then predicting their solubilities, more than 99% have predicted solubilities less than 0.1 (Fig. 4).

Of these generated molecules, two received particularly high scores ((A) 0.508 and (B) 0.417) (Fig. 5). This may be due to the high number of nitrogen atoms in these molecules, as there are many potential sites for hydrogen bonding.

3 Results

After generating molecules using SyntheMol and our various PLQY property predictors, we filtered them for sufficient conjugation (Section 2.5) and predicted their water solubilities (Section 2.6). Our top candidates for fluorescent dyes are shown in Fig 5, and their predicted properties are shown in Table 4. We used several PLQY predictors (Chemprop regression, random forest, SVM) to validate the PLQY predicted by SyntheMol during the generation process. The molecules are all predicted to be sufficiently fluorescent (PLQY > 0.3). Although the PLQY predictions for a given molecule are not entirely consistent among predictors, the predictors generally agree on the relative fluorescence of the molecules. For example, all predictions indicate molecule B as having the greatest PLQY. Molecule B also has a particularly high solubility score (0.417), so it would be considered our top candidate. Molecule A has the highest solubility score (and the only one greater than 0.5) in addition to high PLQY predictions, so it would also be a top candidate. Molecules A and B were generated using the Chemprop regression PLQY predictor, which was also the only SyntheMol run where the maximum number of reactions was greater than one (five reactions). This is likely why there was a higher proportion of sufficiently conjugated molecules generated during this run and also why there were a few molecules with particularly high solubility scores.


(A) (B) (C)

(D) (E)

Figure 5: Top molecules.

Table 4: Predicted properties of top molecules.

Molecule	Generation PLQY predictor	PLQY prediction (Chemprop regression)	PLQY prediction (Random forest)	PLQY prediction (SVM)	Solubility prediction
A	Chemprop regression	0.789	0.519	0.485	0.508
B	Chemprop regression	0.922	0.631	0.523	0.417
C	Chemprop + Morgan regression	0.558	0.483	0.489	0.032
D	Chemprop + RDKit classification	0.334	0.441	0.420	0.0004
E	Chemprop + Morgan classification	0.488	0.475	0.462	0.0002

4 Discussion

Despite the wide range of applications for fluorescent molecules, the field of designing novel fluorescent molecules has predominantly relied on modifying existing scaffolds such as Rhodamine, Coumarine, and Xanthene [8] [9]. However, in this study, we took a different approach by harnessing the potential of SyntheMol, a generative algorithm capable of designing molecules with specific properties.

Using SyntheMol, we designed a set of molecules and subjected them to further screening based on conjugation length and predicted solubility in water. Through this process, we identified five final candidates. Notably, these generated molecules possess unique scaffolds that differ from classical fluorescent molecules. This demonstrates the power of training a generative algorithm using a comprehensive dataset of chemical properties from various molecules. It holds great promise for generating novel scaffolds with desired chemical properties.

To optimize the algorithm for generating high PLQY molecules, several modifications can be considered. Firstly, improvements can be made to the set of building blocks and chemical reactions employed. Conjugation, characterized by alternating single and double bonds, is a well-known critical feature of fluorescent molecules [10]. However, an issue arises when a building block contains a reactive functional group attached to an sp^3-hybridized carbon, as conjugation is disrupted irrespective of the building block's score (see Fig. 6). Addressing this challenge, two strategies can be pursued: incorporating a set of building blocks with a higher representation of conjugated reactive functional groups or integrating an additional scoring function that evaluates the degree of conjugation after each chemical reaction. By ensuring the preservation of conjugation, these modifications hold the potential to significantly improve the algorithm's performance in designing high PLQY molecules.

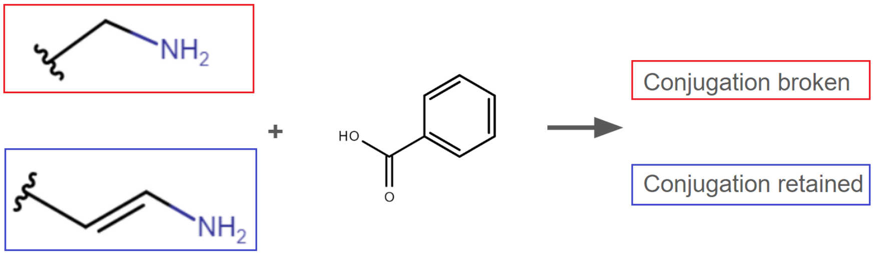


Figure 6: An example schematic of amine-carboxylate coupling showing how conjugation can be broken. A reaction of an aliphatic amine building block (red) will lead to a disconnected conjugation. Only when amino group is already attached to an sp^2-hybridized carbon, the conjugation can be retained (blue).

In addition, an additional filtering step could be introduced after each iteration to enhance the diversity of generated molecules. It was observed that many of the identified hits exhibited remarkably similar core structures (Fig. 7). This outcome is expected, as these core structures tend to possess high predicted PLQY values and are likely to persist throughout the calculation. To overcome this limitation, a filtering process can be implemented whereby only a limited number of molecules with identical core structures are selected for the subsequent iteration. By prioritizing molecular diversity at each stage, this filtering step would encourage the generation of a broader range of novel molecules, leading to increased exploration of the chemical space and potentially uncovering molecules with even higher PLQY values.

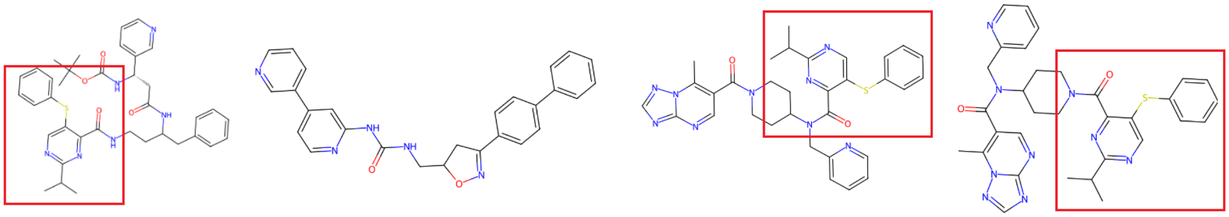


Figure 7: Best 4 molecules generated from SyntheMol, using Regression model with SMILES string and Morgan features as inputs. Hits 1, 3, and 4 share a similar core scaffold, boxed in red.

Lastly, future research could incorporate the generated molecules into an emission/excitation wavelength prediction model. Initial investigations revealed that the Chemprop algorithm exhibited superior performance in predicting these properties compared to PLQY (data not shown). By integrating this approach, an additional filtering step could be implemented to selectively retain dyes with spectra falling within the visible range or to identify pairs of multiple dyes with non-overlapping spectra, which are particularly valuable in biological research[11]. This integration would not only expand the scope of property prediction but also enable the generation of molecules that align with specific wavelength requirements, enhancing their utility in various applications.

5 Conclusion

SyntheMol, a generative algorithm for designing molecules with desired properties, successfully generated promising candidates exhibiting high photoluminescence quantum yields (PLQYs). The PLQYs of these candidate molecules were validated against other reliable PLQY prediction models, further confirming their potential. Incorporating chemical intuition into the algorithm through iterative modifications holds great promise for enhancing its overall performance. To assess and validate the algorithm's effectiveness, future experiments involving the synthesis of these molecules are necessary.

Acknowledgments

We thank our TA Kyle Swanson for his discussions and help on running SyntheMol. We thank Dr. Rabindra Vishwadev Shivnaraine for discussions on fluorescent molecules.

Code and Data

Our code and data are available on GitHub at https://github.com/sbsagireddy/fluorID.

References

[1] Roger Y. Tsien. Fluorescent probes of cell signaling. *Annual Review of Neuroscience*, 12(1):227–253, March 1989.

[2] Hisataka Kobayashi, Mikako Ogawa, Raphael Alford, Peter L. Choyke, and Yasuteru Urano. New strategies for fluorescent probe design in medical diagnostic imaging. *Chemical Reviews*, 110(5):2620–2640, December 2009.

[3] Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, Andrew Palmer, Volker Settels, Tommi Jaakkola, Klavs Jensen, and Regina Barzilay. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(8):3370–3388, July 2019.

[4] Cheng-Wei Ju, Hanzhi Bai, Bo Li, and Rizhang Liu. 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. PMID: 33620207.

[5] David Weininger. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, February 1988.

[6] Kyle Swanson, Gary Liu, Denise B. Catacutan, James Zou, and Jonathan M. Stokes. Generative ai for designing and validating easily synthesizable and structurally novel antibiotics. *Manuscript in preparation*, 2023.

[7] Murat Cihan Sorkun, Abhishek Khetan, and Süleyman Er. AqSolDB, a curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds. *Scientific Data*, 6(1):143, August 2019.

[8] Jonathan B Grimm, Anand K Muthusamy, Yajie Liang, Timothy A Brown, William C Lemon, Ronak Patel, Rongwen Lu, John J Macklin, Philipp J Keller, Na Ji, and Luke D Lavis. A general method to fine-tune fluorophores for live-cell and in vivo imaging. *Nature Methods*, 14(10):987–994, September 2017.

[9] Juan Tang, Michael A. Robichaux, Kuan-Lin Wu, Jingqi Pei, Nhung T. Nguyen, Yubin Zhou, Theodore G. Wensel, and Han Xiao. Single-atom fluorescence switch: A general approach toward visible-light-activated dyes for biological imaging. *Journal of the American Chemical Society*, 141(37):14699–14706, August 2019.

[10] Yoshihiro Yamaguchi, Yoshio Matsubara, Takanori Ochi, Tateaki Wakamiya, and Zen ichi Yoshida. How the conjugation length affects the fluorescence emission efficiency. *Journal of the American Chemical Society*, 130(42):13867–13869, September 2008.

[11] Jeff W Lichtman and José-Angel Conchello. Fluorescence microscopy. *Nature Methods*, 2(12):910–919, November 2005.