Generative models for Luminescence Molecules

Dias Kuatbekov
dept. of Computer Science
Nazarbayev University
Astana, Kazakhstan
dias.kuatbekov@nu.edu.kz

Aruzhan Amangeldina
dept. of Computer Science
Nazarbayev University
Astana, Kazakhstan
aruzhan.amangeldina@nu.edu.kz

Alexandr Alpatov
dept. of Computer Science
Nazarbayev University
Astana, Kazakhstan
alexandr.alpatov@nu.edu.kz

Aiana Yergaliyeva
dept. of Computer Science
Nazarbayev University
Astana, Kazakhstan
aiana.yergaliyeva@nu.edu.kz

Nazerke Yeraliyeva
dept. of Computer Science
Nazarbayev University
Astana, Kazakhstan
nazerke.yeraliyeva@nu.edu.kz

***Abstract*—luorescent molecules play critical roles in applications ranging from organic light-emitting diodes (OLEDs) to bioimaging and fluorescent dyes. Traditional synthesis of new fluorescent molecules is a resource-intensive process that demands extensive experimentation. Recent advances in cheminformatics powered by data-driven approaches offer promising alternatives to conventional methods. This project investigates the application of generative machine learning (ML) models, including Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs), to the synthesis of fluorescent molecules. Our models are trained to propose novel molecular structures, which are evaluated based on metrics such as validity, uniqueness, similarity to the training set, and predicted quantum yield. Preliminary results indicate that these generative approaches can effectively produce viable and unique fluorescent molecules, potentially reducing the need for empirical testing and accelerating discovery processes. luorescent molecules play critical roles in applications ranging from organic light-emitting diodes (OLEDs) to bioimaging and fluorescent dyes. Traditional synthesis of new fluorescent molecules is a resource-intensive process that demands extensive experimentation. Recent advances in cheminformatics powered by data-driven approaches offer promising alternatives to conventional methods. This project investigates the application of generative machine learning (ML) models, including Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs), to the synthesis of fluorescent molecules. Our models are trained to propose novel molecular structures, which are evaluated based on metrics such as validity, uniqueness, similarity to the training set, and predicted quantum yield. Preliminary results indicate that these generative approaches can effectively produce viable and unique fluorescent molecules, potentially reducing the need for empirical testing and accelerating discovery processes. F**

***Index Terms*—Fluorescent molecules, chromophores, generative models, Variational Autoencoders, Generative Adversarial Networks, cheminformatics, quantum yield.**

I. Introduction

The importance of fluorescent molecules in different field can hardly be understated. Fluorescent molecules have diverse applications prominently including the production of OLEDs used in modern displays and in bioimaging for non-invasive monitoring of biomolecular processes. [1]. However, discovery of new photoluminescent molecules remain a significant challenge due to the necessity of experimentation and expensive theoretical calculations. Concidering this challenges, there is a need for alternative approaches that could accelerate the discovery of new photoluminescent molecules. Advent of data driven cheminformatics pose as a significant step forward as a way to resolve the stated problem. Generative Machine Learning have showed success in similar domains, notably in data-driven drug design. We propose that most generative models for drug design can also be used to generate photoluminescent molecules. De-noveu generation of molecules for drug design usually leverage different revisions of generative models belonging to the family of Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs). Most of the available implementations can be tailored for de-noveu generation of molecules with photoluminescent properties. Nonetheless, up to this date, this remain a relatively unexplored domain. The objective of our project work is to explore the capabilites of generative models for the generation of molecules with photoluminescent properties. The project work report will cover how we leverage fine-tuning of pre-trained models to generate new molecules, access the generative models with well-established metrics such as validity, uniqueness and novelty, and also propose some custom metrics tailored for our task.

II. Methodology

A. Data Description

The dataset was obtained from the "Experimental database of optical properties of organic compounds article, available at [1]. The dataset contains 20236 chromophore molecules under different solvents. The molecules are provided in Simplified Molecular Input Line Entry System (SMILES) notation. This string notation represents chemical structures in a format that can be processed by computers. The dataset describes various characteristics of chromophores. The quantum yield is of particular interest for this project as it refers to the efficiency with which the molecule emits light upon excitation. The

average quantum yield value across available chromophore molecules accounted for 0.343.

B. ChemVAE

The Chemical Variational Autoencoder (ChemVAE) is a generative model specifically designed for the generation and optimization of molecular structures [2]. Like other Variational Autoencoder (VAE) architectures, ChemVAE comprises an encoder and a decoder. The encoder features a combination of convolutional layers and Gated Recurrent Units (GRUs) to compress the SMILES representation of a molecule into a latent space. The decoder, primarily composed of GRUs, reconstructs the molecule from this latent representation. The primary training objective is to minimize the reconstruction loss, optimizing the accuracy of molecular generation.

During training, ChemVAE learns a continuous latent distribution, which means that post-training, sampling can be conducted from various parts of this distribution. The degree of similarity or diversity of the generated molecules compared to those in the training set can be adjusted by manipulating the sample mean. This flexibility allows for the exploration of both structurally similar and different molecular structures. Given the computational demands of training, the model was initially pre-trained on 50,000 molecules from the ZINC dataset over 30 epochs. Subsequently, we fine-tuned the model using our dataset of 21,000 unique chromophores for an additional 10 epochs. The rationale behind this approach is that pre-training helps to establish a latent space populated by generally valid molecules, while fine-tuning adjusts this space to better suit our specific task of generating luminescent molecules.

We sampled from the latent space with a z-score of 5, resulting in the generation of 188 distinct molecules.

C. Leveraging VAE to generate molecules

To improve and overcome limitations of the existing chemVAE, the aim of training the model on the bigger dataset was set. VAE model that is trained on a dataset of 250k molecules(ZINC) was chosen as the model that would be another example of VAE implementation. Unfortunately, the initial goal was not achieved due to the large size of the dataset. However, this model was different in terms of executing steps. It tokenizes the SMILES strings to identify unique characters (tokens) necessary for encoding sequences. Next, it initializes an AspuruGuzikAutoEncoder model from DeepChem open-source toolchain built on TensorFlow and Pytorch. Molecules meeting specific criteria (e.g., atom count between 10 and 50) are selected for further analysis. In the context of a VAE, leverage refers to how the model utilizes its learned latent space representation to generate new data samples.

D. MolGen

The model architecture proposed in [3] was employed to generate chromophore molecules in SMILES format. This architecture constitutes a generative adversarial network (GAN) comprising two primary components: a generator and a recurrent discriminator.

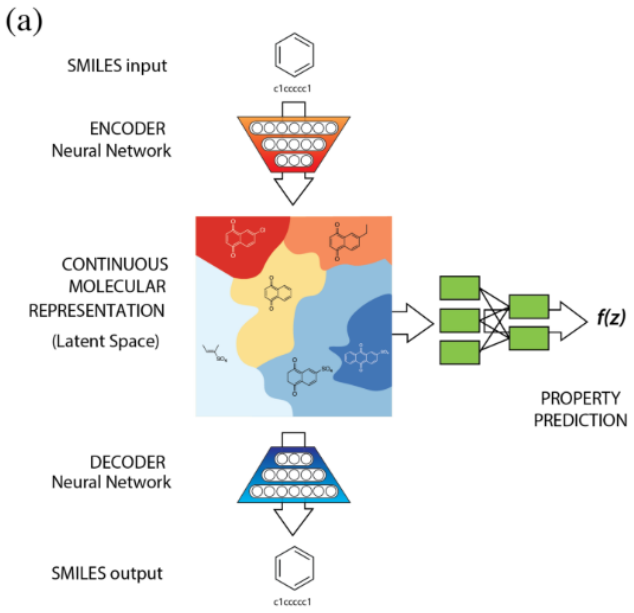

Fig. 1. VAE for molecular design

The generator learns to produce SMILES strings. Initially, a random noise vector z representing the latent space from which molecules are generated is fed into the generator. Subsequently, this vector is projected into a higher-dimensional space to capture its complex structural patterns. The projected vector is then inputted into an LSTM cell, which generates a sequential SMILES output.

The recurrent discriminator is a Recurrent Neural Network which differentiates between true and fake (generated) SMILES strings. It computes a probabilistic likelihood that indicates the extent to which the input is true. This likelihood value is further leveraged to compute a reward for the generator to stimulate its learning, mimicking Reinforcement Learning. The reward is calculated following the equation:

$$R = (2 * y\_pred - 1) \tag{1}$$

The model architecture is demonstrated in fig 2. The model was trained for 70000 steps with a batch size of 128, 256 hidden layers, and a learning rate of 1e-4. For every 2000 steps, the model performance was evaluated by generating 100 molecules and assessing their validity using the RDKit library. The progression of the validity score is demonstrated in fig 3.

E. MolGAN

One prominent alternative to the above approaches was to try and predict molecules not based on their SMILES, but based on their Graph representation. In this method, each molecule is represented by an undirected graph with a set of edges E and a set of nodes V. Each node represents an atom

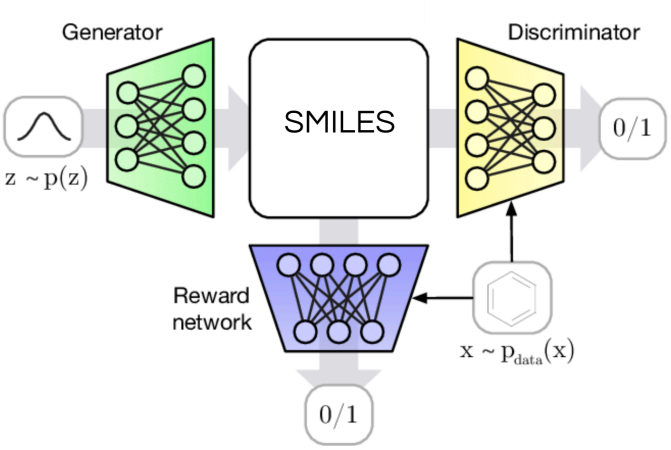

Fig. 2. MolGen architecture

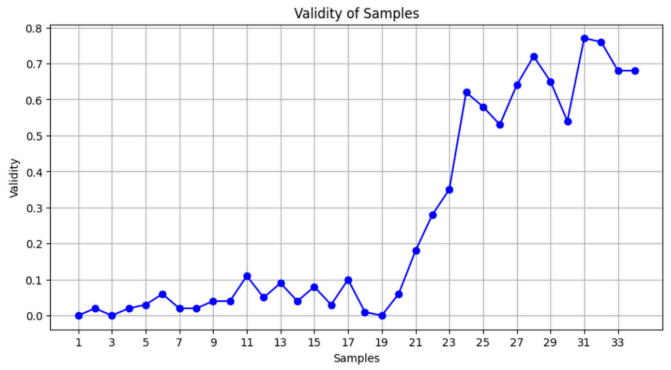

Fig. 3. MolGen training process

and each edge represents an atomic bond. Thus, each atom is encoded into a one-hot vector x_i, indicating its type and each bond between atoms x_i and x_j is represented as an entry in an adjacency matrix $\mathbf{A}_{ij}$.

The most promising Graph-based generative model we've found was MolGAN - an implicit, likelihood-free generative model for small molecular graphs that circumvents the need for expensive graph matching procedures or node ordering heuristics of previous likelihood-based methods [4]. The MolGAN architecture consists of three main components: a generator G, a discriminator D and a reward network R. The Generator and Discriminator are playing a minimax game, while the Reward network is used to add an element of Reinforcement Learning (RL) to the model.

Key advantages of MolGAN include:

- Likelihood-free estimation.
- Graph-based representation of the molecules.
- Improved Loss function, based on the loss function of Wasserstein GAN (WGAN).
- Deterministic Policy Gradient, that allows us to customize the Reward network.

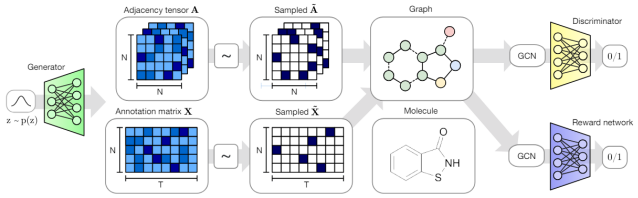

Fig. 4. MolGAN Architecture

F. TransformerVAE

In the Transformer VAE model the encoder and decoder use a Transformer architecture to map input molecules into the latent space and decode latent representations back into molecular structures respectively [5]. Combining Transformer into VAE allows for capturing more complex dependencies in molecular structures.

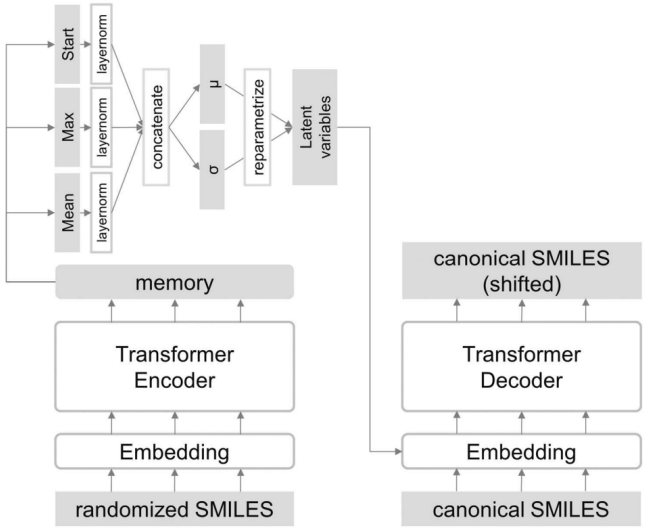

Fig. 5. Structure of Transformer VAE model

The architecture of Transformer VAE is shown in fig 5. The sinusoidal positional encoding is added with embedded randomized SMILES and together they serve as input into the encoder. The mean and maximum of the memory are pooled from the output of the encoder and then are concatenated with the initial token memory. Using the pooled memory, the mean and variance of the posterior distribution of latent variables are calculated. Lastly, the latent variables are reparameterized and are added to the embedded input of the decoder [5].

As seen in fig 6 the TransformerVAE shows better or comparable performance than existing models in generating novel molecules across different metrics.

III. Results

A. ChemVAE

ChemVAE was able to generate molecules that scored high on validity, uniqueness and novelty metrics. We further compare the generated molecules on similarity with the training set

		Our model	HMM	NGram	Combinatorial	CharRNN	AAE	VAE	JTN-VAE	LatentGAN
Valid(↑)		0.8761±0.0035	0.0760±0.0322	0.2376±0.0025	**1.0000±0.0000**	0.9748±0.0264	0.9368±0.0341	0.9767±0.0012	**1.0000±0.0000**	**0.8966±0.0029**
unique@1000(↑)		**1.0000±0.0000**	0.6230±0.1224	0.9740±0.0108	0.9983±0.0015	**1.0000±0.0000**	**1.0000±0.0000**	**1.0000±0.0000**	**1.0000±0.0000**	**1.0000±0.0000**
unique@10000(↑)		**1.0000±0.0001**	0.5671±0.1424	0.9217±0.0019	0.9909±0.0009	0.9994±0.0003	0.9973±0.0020	0.9984±0.0005	0.9996±0.0003	
FCD(↓)	Test	1.3380±0.0735	24.4661±2.5251	5.5069±0.1027	4.2375±0.0370	**0.0732±0.0247**	0.5555±0.2033	0.0930±0.0125		0.2968±0.0087
	TestSF	1.7098±0.0740	25.4312±2.5599	6.2306±0.0966	4.5113±0.0274	**0.5204±0.0379**	1.0572±0.2375	0.5670±0.0338		0.8281±0.0117
SNN(↑)	Test	0.4803±0.0025	0.3876±0.0107	0.5209±0.0010	0.4514±0.0003	0.6015±0.0206	0.6081±0.0043	**0.6257±0.0005**	0.5477±0.0076	0.5371±0.0004
	TestSF	0.4658±0.0026	0.3795±0.0107	0.4997±0.0005	0.4388±0.0002	0.5649±0.0142	0.5677±0.0045	**0.5783±0.0008**	0.5194±0.0070	0.5132±0.0002
Frag(↑)	Test	0.9927±0.0016	0.5754±0.1224	0.9846±0.0012	0.9912±0.0004	**0.9998±0.0002**	0.9910±0.0051	0.9994±0.0001		0.9986±0.0004
	TestSF	0.9898±0.0018	0.5681±0.1218	0.9815±0.0012	0.9904±0.0003	0.9983±0.0003		**0.9985±0.0003**	0.9947±0.0002	0.9972±0.0007
Scaf(↑)	Test	0.7159±0.0192	0.2065±0.0481	0.5302±0.0163	0.4445±0.0056	0.9242±0.0058		**0.9386±0.0021**	0.8964±0.0039	0.8867±0.0009
	TestSF	**0.1893±0.0081**	0.0490±0.0180	0.0977±0.0142	0.0865±0.0027	0.1101±0.0081		0.0588±0.0005	0.1009±0.0105	0.1072±0.0098
IntDiv(↑)		0.8531±0.0013	0.8466±0.0403	**0.8738±0.0002**	0.8732±0.0002		0.8557±0.0051	0.8558±0.0004	0.8551±0.0034	0.8565±0.0007
Filters(↑)		0.9706±0.0005	0.9024±0.0489	0.9582±0.0010	0.9557±0.0018	0.9943±0.0034	0.9960±0.0006	**0.9970±0.0002**	0.9760±0.0016	0.9735±0.0006
Novelty(↑)		0.9911±0.0005	**0.9994±0.0010**	0.9694±0.0010	0.9876±0.0008	0.8419±0.0309	0.7931±0.0285	0.6949±0.0069	0.9143±0.0058	0.9408±0.0006

Fig. 6. Comparison of the generative ability of TransformerVAE with other models

via Tanimoto score. The analysis shows that generated moelcules are very dissimilar to molecules used to train chemVAE. chemVAE shows the best results for average quantum yield results.

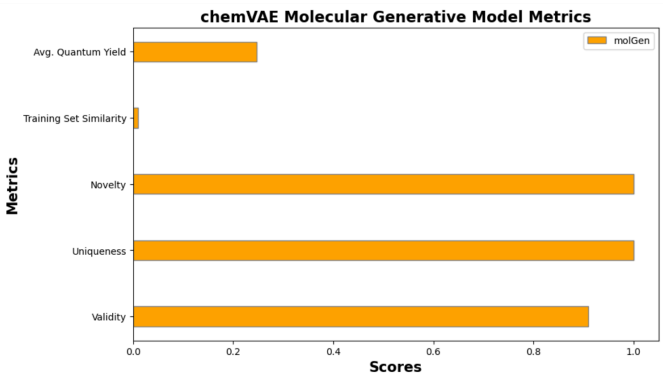

Fig. 7. Evaluation metrics results for Chemical VAE

B. *levVAE*

When comparing the second VAE model to other models using metrics, it demonstrated strong performance in terms of both validity and novelty. It is self-explanatory results as the VAE is the generative model that is focused producing the data that is different from the training data. Notably, this model showed even higher validity than ChemVAE. However, its quantum scores were relatively low, consistent with the performance of the other two models.

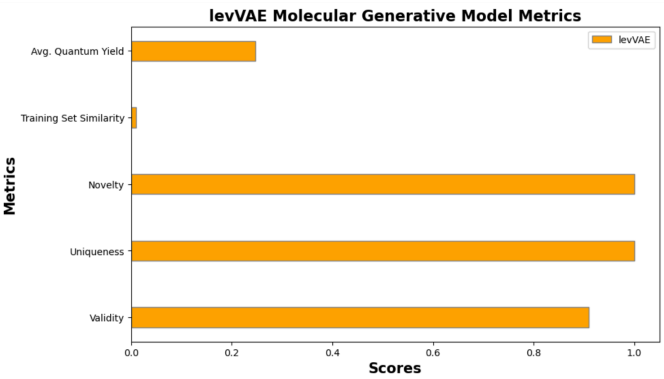

Fig. 8. Evaluation metrics results for LevVae

C. *MolGen*

As a result of training the MolGen model, slightly more than 7600 chromophore molecules were generated. The analysis revealed high scores on validity and uniqueness. The model scored relatively lower on novelty in contrast to other models. Comparatively higher resemblance to the training set according to the Tanimoto score was also observed. The model exhibited the second best results in a quantum yield of 0.23.

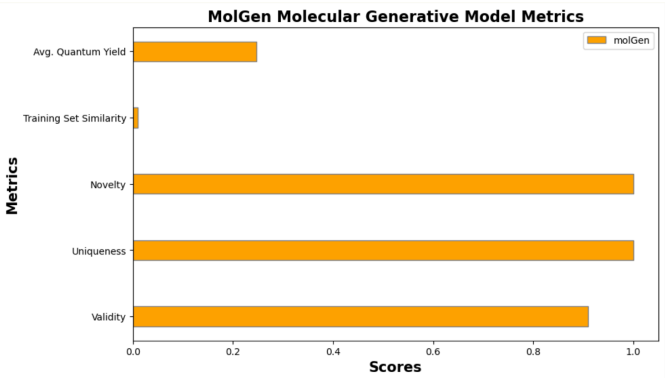

Fig. 9. Evaluation metrics results for MolGen

D. *MolGAN*

Unfortunately, we were unable to obtain any meaningful results from MolGAN due to various reasons. We tested 4 different implementations of MolGAN using differing backend libraries (JAX, TensorFlow, Torch). Most of them failed because of their lack of backward-compatibility and outdated code. Torch implementation compiled and started its training phase, but ultimately failed because of the lack of computation resources, namely RAM and Video Memory. Possible improvements include: refactoring the legacy code of different implementations, repeating the experiment with more computational power, and reimplementing the model using different/more optimized backend technology.

E. *TransformerVAE*

Unfortunately, with TransformerVAE we could not produce any tangible results due to errors occurring during the model training process. The possible reasons for that are complexity of the model, compatibility issues with our dataset format, and our gap in knowledge and skills.

TABLE I
COMPARISON OF GENERATIVE MODELS ON VARIOUS METRICS

Model	Validity	Uniqueness	Novelty	Training Set Similarity	Avg. Quantum Yield
ChemVAE	0.91	1.00	1.00	0.01	0.247
LevVAE	1.00	1.00	1.00	0.01	0.226
MolGen	1.00	1.00	0.83	0.08	0.23

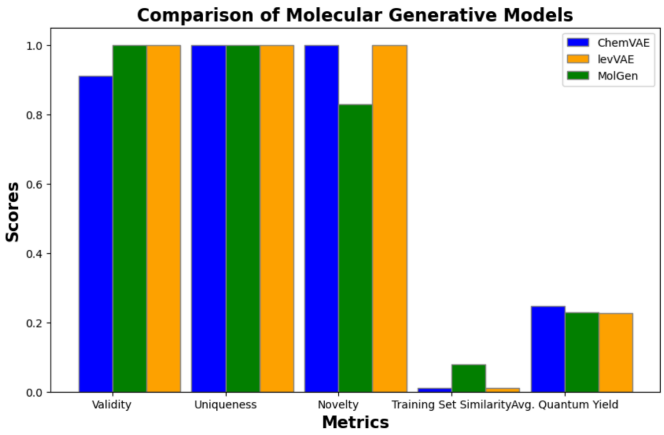

Fig. 10. Evaluation metrics across all models

IV. Discussion

Three models were trained on the available dataset and evaluated across several metrics. All models exhibited strong performance in validity, uniqueness, and novelty properties. According to the Tanimoto score, VAE-based models performed better at generating novel molecules in contrast to the GANs-based model. This can be attributed to the varying sampling strategies of the two models. VAEs sample from a learned continuous latent space, allowing for smooth interpolation between molecules and the generation of novel structures by exploring different regions of this space. GANs may struggle to maintain diversity and novelty, especially with a smaller dataset, as they rely on random noise vectors that may not effectively capture the underlying distribution of the data. GANs are also susceptible to mode collapse, where the generator produces limited diversity in its outputs, focusing on a few modes of data distribution. This can hinder the generation of novel molecules, especially if the dataset is small and does not cover the full diversity of molecular structures.
All models produced close quantum yield scores of over 0.20. It is noteworthy that while these quantum yield values might seem relatively low, it's consistent with the average quantum yield of the training set of 0.343. This in turn proves the models' capability to generate molecules with similar characteristics to those present in the training dataset.
During the exploration of several available models, various obstacles were faced. One of the limitations is the relatively small size of the dataset (20k) in comparison to benchmark datasets, such as the Zinc dataset of 250k and the QM9 dataset of 134k stable molecules. This in turn limits the capability to train large complex models as it may result in overfitting, limited generalization, and failure to learn meaningful representations. Additionally, outdated code and limited computational power of our devices hindered getting meaningful results from models, such as MolGAN. Another limitation is that the current models lack the consideration of synthesis difficulty and vivo stability, that is the models focus on generating molecules with desired properties and chemical structure without considering synthesis feasibility and stability in biological environments.

V. Conclusion

To summarize, the main aim of our work was evaluating the performance of various generative machine learning models in producing valid fluorescent molecules. The obtained results of using Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) show high potential in accelerating the discovery of new fluorescent compounds. By applying different models on the datasets with the various sizes, we aimed to examine them using several criteria and compare the effectiveness of each model. The metrics of the quantum yield of the molecule was added to the well-established metrics. To be precise, the models ChemVAE, MolGen, MolGAN, TransformerVAE were covered by our work. Although we couldnt obtain the results from the last two models, the detailed overview was conducted, including the performance metrics based on the theorethical data. The main criterias used to access the data like validity, uniqueness, novelty were anticipated accurately, with the average of the quantum yield being over 0.20. The limitations of the project included the available computational power, small dataset size and outdated code, which could not be implemented. Moreover, listed models lack of consideration of synthesis difficulty and in vivo stability of the generated molecules.

References

[1] J. F. Joung, M. Han, M. Jeong, and S. Park, "Experimental database of optical properties of organic compounds," Sep 2020.
[2] R. Gmez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernndez-Lobato, B. Snchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, and A. Aspuru-Guzik, "Automatic chemical design using a data-driven continuous representation of molecules," 2018. PMID: 29532027.
[3] Urchade, "Urchade/molgen: Molecule smiles generation with gan and reinforcement learning (training language gan from scratch)."
[4] N. D. Cao and T. Kipf, "Molgan: An implicit generative model for small molecular graphs," 2022.
[5] Y. Yoshikai, T. Mizuno, S. Nemoto, and H. Kusuhara, "A novel molecule generative model of vae combined with transformer for unseen structure generation," 2024.