Dept of Physics, IIT Delhi

Course Code : PYD411
Academic Year : 2023 - 2024, Semester - I

Quantum Machine Learning in De Novo Drug Discovery

Vinamr Jain (2020PH10734)

Adviser: Prof. Abhishek Dixit
Prof. Kedar Bhalchandra Khare

Abstract: Recent years have witnessed groundbreaking advancements in drug discovery and the life sciences, driven by the integration of machine learning and artificial intelligence (AI) techniques. The next paradigm shift in technology is predicted to be quantum computing, with one of its early practical applications being quantum chemistry simulations. This research project report explores the near-term applications of quantum computing in generative chemistry and its advantages in revolutionizing drug discovery. Traditional drug discovery pipelines are time-consuming and expensive, involving the exploration of vast chemical space. Deep generative models, such as Generative Adversarial Networks (GANs), have shown promise in generating molecular structures with desired properties. However, classical GANs face limitations in exploring complex chemical spaces. This report proposes a qubit-efficient quantum GANs to efficiently navigate the expansive chemical space with few qubits. The combination of AI and quantum computing holds the potential to significantly accelerate drug development, reduce costs, and improve pharmaceutical research outcomes. The report discusses the challenges faced by hybrid quantum-classical GANs in generating unique and valid molecules but highlights their promising quantum advantages. The convergence of quantum machine learning and de novo drug discovery offers a compelling avenue for future research and development in the pharmaceutical industry.

Signature of student 1: ..... Signature of the adviser: .....

1 Motivation

In recent years, the transformative potential of quantum computing in fields like generative chemistry and drug discovery has become increasingly evident. Quantum computations offer the prospect of exponential speed-up in solving complex problems compared to their classical counterparts, making them invaluable for applications ranging from protein structure prediction to network analysis. The ability of quantum computers to process big data efficiently, run machine learning algorithms, simulate quantum systems, and optimize solutions can revolutionize the life sciences[1].

From a software perspective, existing quantum algorithms show great promise in drug discovery and biology. They enable faster ground-state calculations, linear systems processing, and more accurate predictions of drug-receptor interactions and protein folding. The fast progress in quantum machine learning (QML) methods complements classical AI approaches for generative chemistry, further advancing the potential of quantum computing in this domain.

Among classical generative models in drug discovery, Generative Adversarial Networks (GANs) have been highly successful. Quantum GANs (QuGANs) represent a significant leap forward, as they offer quantum speed-up, enabling the learning of more complex molecular representations due to the amplitude amplification property. QuGANs also provide more efficient exploration of the exponentially large chemical space and the ability to sample from distributions that are challenging for classical modeling[2].

The integration of quantum computational models into generative chemistry, with accuracy comparable to classical methods, is poised to impact various areas of life sciences positively. As quantum hardware advances, with the prospect of devices featuring thousands of physical qubits, early adopters stand to benefit greatly. Drug discovery and generative chemistry stand out as domains that can rapidly incorporate quantum computing into their pipelines. However, the development of new quantum software solutions and algorithms remains crucial for the successful application of quantum computing in generative chemistry. Key areas of focus include optimizing quantum GANs and quantum autoencoders[1].

The traditional drug discovery and development process is both time-consuming and costly, taking 10-18 years and involving expenditures in the billions. This process has recently shifted towards biologic drugs that selectively target disease-related cellular ligands. Identifying ligands with the right structural complementarity to disease-associated protein pockets necessitates navigating an immensely vast chemical space, a challenge known as de novo drug discovery. Classical generative models have limitations in exploring certain regions of this space due to exponential choices in molecule generation.

Quantum computing holds several advantages over classical computing, particularly in the fields of chemistry simulation, machine learning, and optimization. Quantum GANs, in particular, are poised to harness these advantages. They offer the potential for quantum speedup, enabling the learning of richer molecular representations and efficient exploration of the chemical space with relatively few qubits.

The drug discovery process is resource-intensive, with a low success rate, making it a prime candidate for quantum computing solutions. The complexity of molecular simulations, which includes lead identification and optimization, is a major challenge for classical computers but can be addressed effectively using quantum computing, given that molecules adhere to quantum mechanics.

Quantum computing represents a revolutionary approach to generative chemistry and drug discovery, with the potential to dramatically accelerate the development of

novel therapeutics while reducing costs and improving research outcomes. The synergy of quantum computing's computational power and the ever-evolving field of generative chemistry holds immense promise for the future of pharmaceutical research and development.

2 Theory

2.1 Generative Adversarial Networks

In generative adversarial networks (GANs), the goal is to generate data resembling the training data by training a generator and a discriminator simultaneously. The generator creates fake data resembling the real dataset, and the discriminator distinguishes real from fake data. They engage in a competitive training process, with the generator aiming to produce data indistinguishable from the training set. This process is represented by a two-player minimax game, where the discriminator seeks to maximize its ability to discern real from fake data, while the generator aims to minimize this discriminative probability, summarized by the value function[3]-

$$\underset{G}{min}\ \underset{D}{max}\ V(D,G) = E_{x\sim p_{data}}[logD(x)] + E_{z\sim p_z}[log(1-D(G(z))]$$

- **x**: real data sample
- **z**: latent vector
- **D(x)**: probability of the discriminator classifying real data as real
- **G(z)**: fake data
- **D(G(z))**: probability of discriminator classifying fake data as real

2.2 Variational Quantum Circuits

VQCs comprise of controlled gates, namely the Controlled-X (CNOT) gate and Controlled-Z gate, along with single-qubit rotations (Rx, Ry, Rz, and R). The CNOT gate operates on two qubits, typically termed the control qubit (first qubit) and the target qubit (second qubit).

- **Initialization Layer:** This layer involves the application of Rx, Ry, Rz, and R gates to prepare an initial state. The rotation angles for these gates can be sampled from either a uniform or Gaussian distribution.
- **Parametrized Layers:** These layers, which can be iterated L times, incorporate CNOT gates, CZ gates, and parametrized rotational gates. The parameters, specifically the rotation angles, are adjustable and can be learned through back-propagation during training.
- **Measurement:** The measurement stage calculates the expected value for each qubit.

Figure 1(c) shows matrix representation of example quantum gates.

2.3 Evaluation Metrics

1. **Quality Metrics:**
    - **Validity:** Ratio of valid molecules to all generated molecules.
    - **Uniqueness:** Ratio of unique molecules to valid molecules.

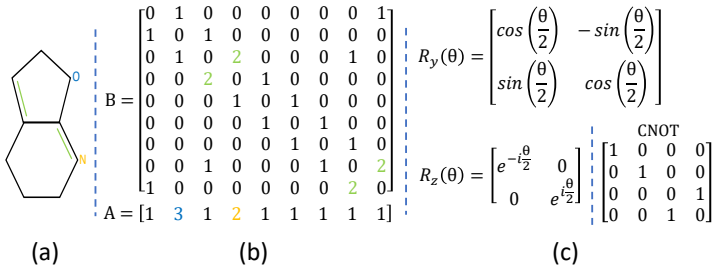

Figure 1: (a-b) A sample molecular graph from QM9 denoted by its corresponding atom vector A and bond matrix B; (c) Example Quantum Gates

- **Novelty:** Ratio of valid molecules not in the training dataset to all valid molecules.
- **Diversity:** Measures how diverse generated molecules are compared to the training dataset.

2. **Drug Properties:**
   - **Quantitative Estimation of Drug-Likeness (QED):** Measures the likelihood of a molecule being a drug based on desirability.
   - **Solubility:** Reports the n-octanol-water partition coefficient (logP), indicating hydrophilicity.
   - **Synthesizability (SA):** Quantifies ease of synthesis based on molecular complexity and fragment contributions.
3. **Dataset:** All experiments use the QM9 dataset, derived from the GDB-17 chemical database. QM9 contains 133,171 molecules with up to nine non-hydrogen atoms (C, N, O, F). Average QED, solubility, and SA of QM9 molecules are 0.461, 0.289, and 0.327, respectively.
4. **Frechet Distance:** Measures similarity between real and generated molecule distributions
5. **Kullback-Leibler (KL) divergences:** The KL-divergence score is a statistical metric used to measure how well a generative model's output aligns with the physicochemical properties of molecules in a training dataset. A higher KL-divergence score indicates a better match and suggests that the generative model accurately captures these properties, while a lower score implies room for improvement in replicating the properties of reference molecules.

3 Computational Setup

3.1 QGAN-HG

In our first setup, as described in the work by Li et al. (2021) [2], the Quantum GAN with a hybrid generator (QGAN-HG) is constructed with two key components. The first component is a parameterized quantum circuit responsible for producing a feature vector with a dimension defined by the number of qubits. The second component is a classical deep neural network, which generates an atom vector and a bond matrix to represent the molecular structure of drug compounds graphically. An alternative variation of the QGAN-HG is the patched Quantum GAN with a hybrid generator (P-QGAN-HG). In this variant, the quantum circuit is formed by concatenating several quantum sub-circuits, offering an adapted approach. In the case of using real data, as

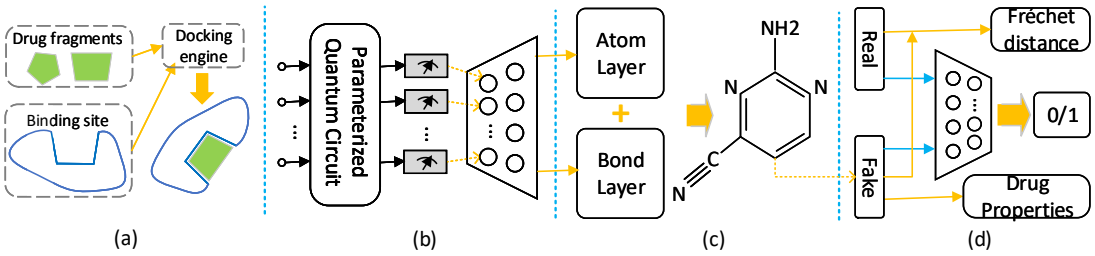

Figure 2: (a) Only generated molecules that have high affinity towards the receptor binding sites are considered as valid; (b) quantum stage (which is a parameterized quantum circuit with last-layer N measuring the expectation values) and classical stage (neural network with last-layer out-feature dimension of 512) separated by blue dotted line; (c) application of atom layer and bond layer for generating synthetic molecular graphs (one example synthetic molecule is given); (d) a batch of real molecules from training dataset (QM9 in this case) and a batch of synthetic molecules generated from (c) are fed into classical discriminator for real/synthetic prediction and FD score calculation, and drug properties for synthetic molecules are evaluated using RDKit package. The prediction losses from discriminator are back-forwarded to two neural networks as well as quantum circuit for updating all parameters simultaneously in each training epoch [2].

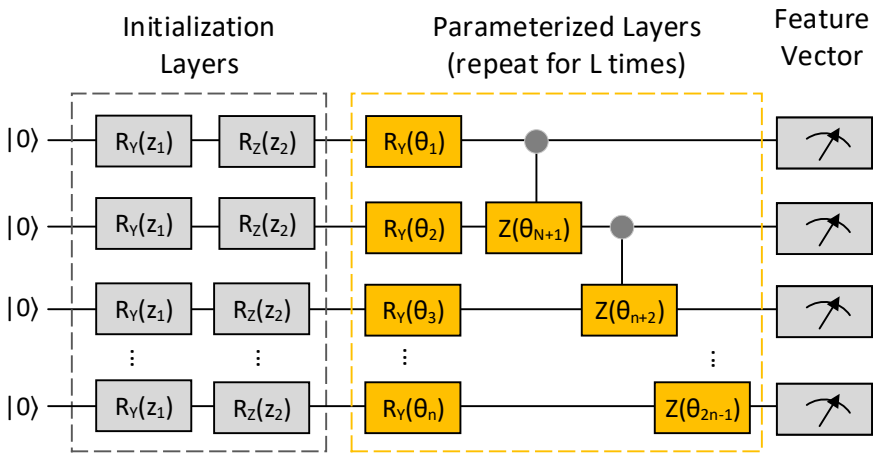

Figure 3: Parameterized quantum circuit to obtain feature vector of N dimensions. The circuit is composed of initialization layers, repeatable parameterized layers and measurement layer. Two CNOT gates for each ZZ interaction for creating entanglement are not shown here[2]

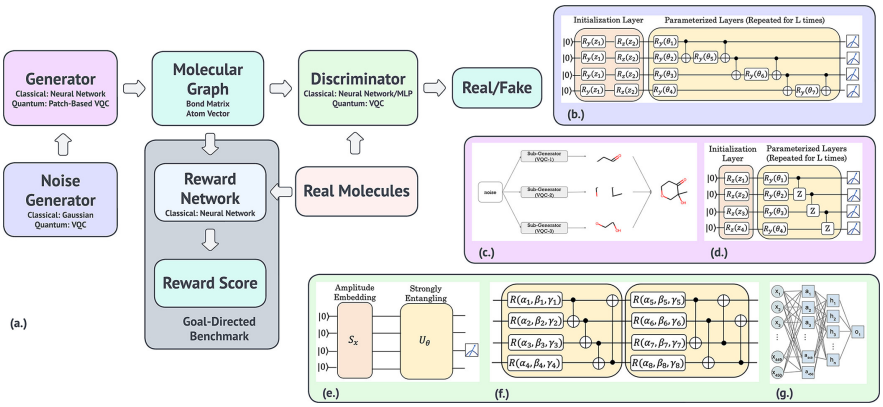

Figure 4: Overall pipeline[4]. (a) The overall pipeline of MolGAN with different combinations of classical/quantum components. The reward neural network branch is enabled in the goal-directed benchmark. The classical noise generator samples from the Gaussian distribution, and the quantum one uses the variational quantum circuit (VQC). The classical generator is built by neural networks, and the quantum one uses the patch-based VQC to generate the molecular graph. The molecular graph is represented by a bond matrix and atom vector. The classical discriminator is built by a graph-based neural network or multilayer perceptron (MLP), and the VQC is used in the quantum one. (b) The example of VQC in the noise generator. (c) The patch method uses multiple VQCs as subgenerators. Each subgenerator takes noise as input and outputs a partial part of the final molecular graph. The final molecular graph is constructed by concatenating all the partial patches together. (d) The example of VQC in the quantum generator. (e) The VQC of the quantum discriminator consists of the amplitude embedding circuit (Sx), the strong entanglement layers (U_θ), and the measurement. (f) The VQC of strongly entanglement layers contains multiple CNOT gates and parametrized rotational gates (R). (g) MLP-based discriminator architecture in MolGAN-CC.

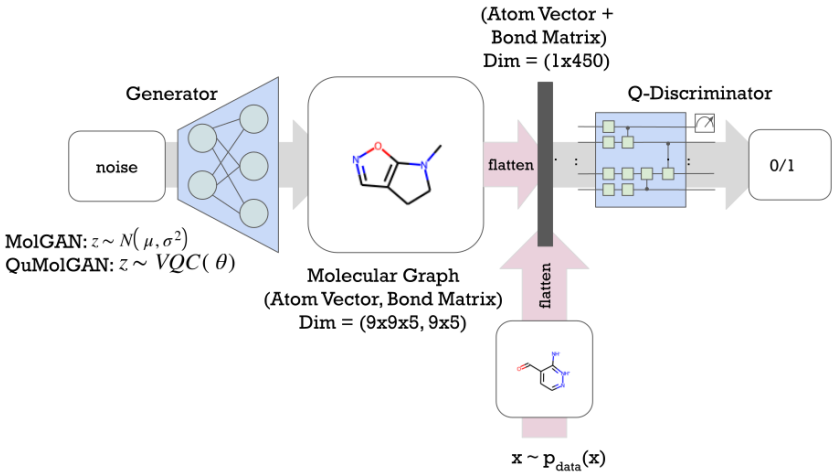

Figure 5: MolGAN Quantum Discriminator Schema[4]

illustrated in Figure 2 (d), an essential step involves the preparation of quantum states. Typically, this is accomplished through amplitude encoding, which translates classical data into a quantum state. This stage consumes N log(M) qubits, where N signifies the size of the training dataset, and M denotes the feature dimension. A quantum layer is introduced to harness the remarkable expressive capabilities of variational quantum circuits, which efficiently perform low-rank matrix computations in $O(poly(\log(M)))$ time, providing exponential speedup. The variational quantum circuit, as depicted in Figure 3, comprises three distinct stages: initialization, parameterized (repeated for L layers with $L(2N-1)$ parameter count), and measurement stages. Within this quantum circuit, two parameters, z1 and z2, are uniformly sampled from the interval $[-\pi, \pi]$, effectively serving as replacements for the random Gaussian noise input commonly found in classical GANs. Following the application of the initialization layers, the quantum state is prepared mathematically as: $|z\rangle = (\bigotimes_{i=1}^{N} R_Z(z_2)R_Y(z_1)|0\rangle)^{\bigotimes^N}$. Let's denote the parameterized layers, repeated L times, as the unitary matrix U(θ). Consequently, the final quantum state is expressed as: $|\Psi(z)\rangle = U(\theta)|z\rangle$. A sequence of measurement operators is applied to derive the expectation value for each qubit, forming the feature vector that will be subsequently input to the classical neural network.

In the classical component of the hybrid generator, a conventional neural network is employed. Its input layer receives the feature vector, composed of the expectation values. The concluding layer of this network is divided into two parts: one for generating atom vectors and the other for creating bond matrices. Similar to the approach used in MolGAN, a categorical re-parameterization step with Gumbel-Softmax is applied. This step facilitates the calculation of gradients during the backward pass and is instrumental in producing discrete synthetic molecular graphs. The patched Quantum GAN with a hybrid generator shares the same core components as previously described. However, there's a notable difference in the architecture of the parameterized quantum circuit. In this case, the quantum circuit comprises multiple quantum sub-circuits.

P-QGAN-HG offers the advantage of requiring fewer quantum resources, as these multiple sub-circuits can be executed either sequentially or in parallel. This approach brings increased efficiency to the learning process, with each sub-circuit being simulated more effectively. However, it's important to note one of the evident trade-offs: a reduction in expressive power. This reduction is due to the quantum state dimension diminishing from 2^N to $2^{(N/2)}$ when multiple circuits with half the size are employed in the Hilbert space.

The discriminator is kept the same as MolGAN in this setup.

3.2 Quantum Noise, Generator and Discriminator

Following the works of Kao et al. (2023)[4], we incorporate VQCs for noise generation and for replacing the generator and the discriminator networks. In QuMolGANs (Quantum Noise, classical Generator and Discriminator), MolGAN is used as the base model. In classical GANs, the inputs of the generator are sampled from a uniform or a Gaussian distribution, while here in contrast, we use a variational quantum circuit for the generator input generation (see Figure 4 (b)) to demonstrate the quantum advantage.

MolGAN-QC (classical noise, quantum generator, classical discriminator) follows the same methodology as QGAN-HG as proposed above. Here we employ multiple Variational Quantum Circuits (VQCs) as sub-generators, where each sub-generator is tasked with constructing a specific portion of the final output—such as the molecular graph in our study. The ultimate molecular graph, comprising the atom vector and bond matrix, is formed by concatenating all the partial patches together, as illustrated in Figure 4c. The sub-generators adhere to the same ansatz architecture, as depicted

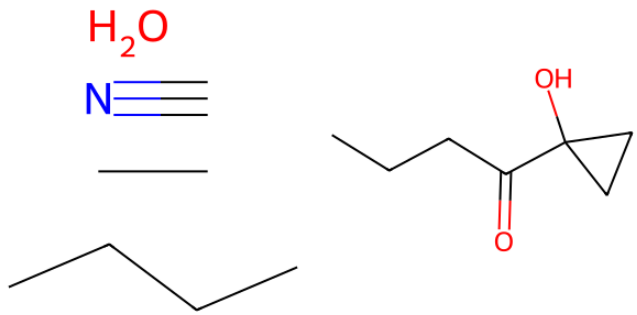

Figure 6: Sample generated molecules in QuMolGAN-HR

in Figure 4d. Each ansatz circuit within the sub-generator undergoes three stages: the preparation of the initialization state, a single layer of a 4-qubit circuit, and the measurement. The initialization layer involves Ry gates, with the rotation angles (z_i) being sampled from a uniform distribution. The parametrized layers, which can be repeated for L times, include Controlled-Z (CZ) gates and a specific type of parametrized rotational gates, Ry, whose parameters (θ_k) are learnable through back-propagation.

The quantum discriminator (MolGAN-CQ) operates by taking the molecular graph as input and discerning whether it originates from the generator (fake) or the authentic data distribution (real). The Variational Quantum Circuit (VQC) in the quantum discriminator comprises three key components: the amplitude encoding layer (S_x), the strongly entangling layers (U_θ), and the measurement, as visually represented in Figure 4e. Amplitude encoding is employed to encode the atom vector and bond matrix, utilizing the operator S_x. The strongly entangling layers, denoted by U_θ, incorporate multiple Controlled-NOT (CNOT) gates and parametrized rotational gates $R(\alpha, \beta, \gamma)$ (Figure 4f). Within each layer, each qubit initiates with parametrized rotational gates $R(\alpha_i, \beta_i, \gamma_i)$, followed by a CNOT gate. The parametrized angles $\alpha_i, \beta_i, \gamma_i$ are subject to learning through back-propagation. The measurement stage calculates the expected value of one qubit, and this outcome is utilized to determine the authenticity of the input molecular graph—whether it is real or generated (See 5).

To assess quantum advantage in the goal-directed benchmark, we activate the reward network branch within the original schema, depicted by the gray block in Figure 4a. In this configuration, the generator undergoes training through a linear combination of the Wasserstein GAN (WGAN) loss and the reinforcement learning (RL) loss:

$$L(\omega) = \alpha \cdot L_{\text{WGAN}}(\omega) + (1 - \alpha) \cdot L_{\text{GAN RL}}(\omega)$$

where $\alpha \in [0, 1]$ serves as a hyperparameter governing the tradeoff between the WGAN loss and RL loss. The variable ω represents the inputs to the networks.

4 Results and discussion

Figures 6 & 7 illustrate sample molecules produced using different architectures described in the report. Firstly, the QGAN-HG model was trained for a VQC with 8 qubits, 1 layer and without patches. Medium reduced (MR) configuration was used with $\sim 59k$ parameters in the Generator, mini-batch size of 16 trained for 5000 iterations, with learning rate of 0.001 for the generator and discriminator each. Figure 8 and 9 show the Frechet Distance and the drug properties obtained in the paper [2]. Figure 10, 11 and 7 (b) Show the reproduced Freched Distance, drug property results and valid generated molecule respectively. As can be seen, QGAN-HG outperforms classical MolGAN in both drug properties and Frechet distance, even with a significant

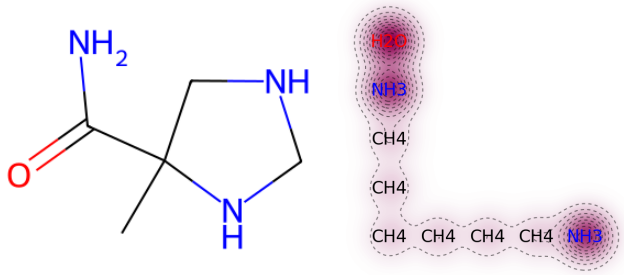

Figure 7: Sample generated molecule in QuMolGAN-NR and QGAN-HG respectively

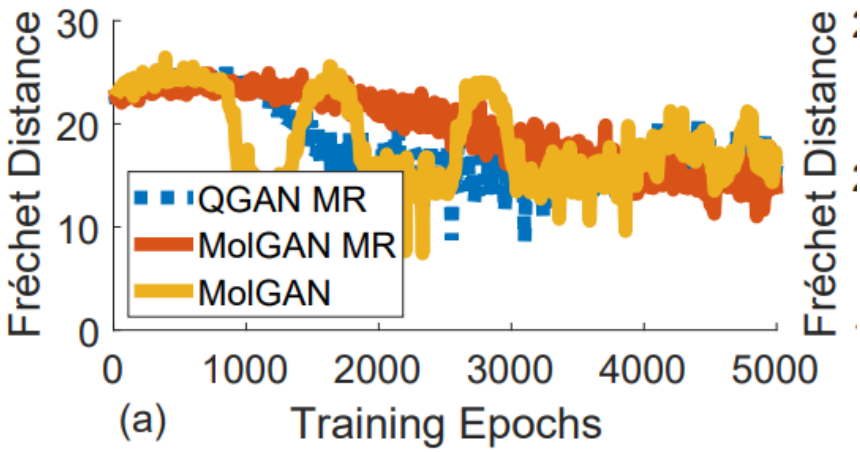

Figure 8: Metrics of Frechet distance for QGAN-HG MR as observed in the paper[2]

TABLE I

DRUG PROPERTIES OF 1000 GENERATED MOLECULES FROM ALL GAN VARIATIONS IN THIS PAPER. BEST RESULTS ARE SHOWN IN BOLD. SIGN '-' INDICATES THE CORRESPONDING METRIC FOR SAMPLED MOLECULES ARE NOT SUCCESSFULLY EVALUATED BY RDKIT.

Method	Druglikeliness	Solubility	Synthesizability	Diversity	Valid	Unique	Novel
MolGAN* [10]	0.50	**0.70**	0.11	**1.0**	**0.82**	0.21	**1.0**
MolGAN MR	0.47	0.60	0.14	**1.0**	0.31	0.70	**1.0**
MolGAN HR	-	-	-	**1.0**	0.10	**1.0**	**1.0**
QGAN-HG MR (proposed)	**0.51**	0.49	0.07	**1.0**	0.63	0.35	**1.0**
QGAN-HG HR (proposed)	-	-	-	**1.0**	0.03	**1.0**	**1.0**
QGAN-HG HR L2 (proposed)	-	-	-	**1.0**	0.02	**1.0**	**1.0**
QGAN-HG HR Q10 (proposed)	0.49	0.43	0.15	**1.0**	0.04	**1.0**	**1.0**
P2-QGAN-HG MR (proposed)	0.49	0.62	0.11	**1.0**	0.53	0.40	**1.0**
P4-QGAN-HG MR (proposed)	0.49	0.51	0.13	**1.0**	0.59	0.45	**1.0**
QGAN-HG MR (on IBM quantum computer)	0.48	0.50	**0.17**	**1.0**	0.38	0.92	**1.0**

Figure 9: Drug property results for QGAN-HG MR in the paper[2]

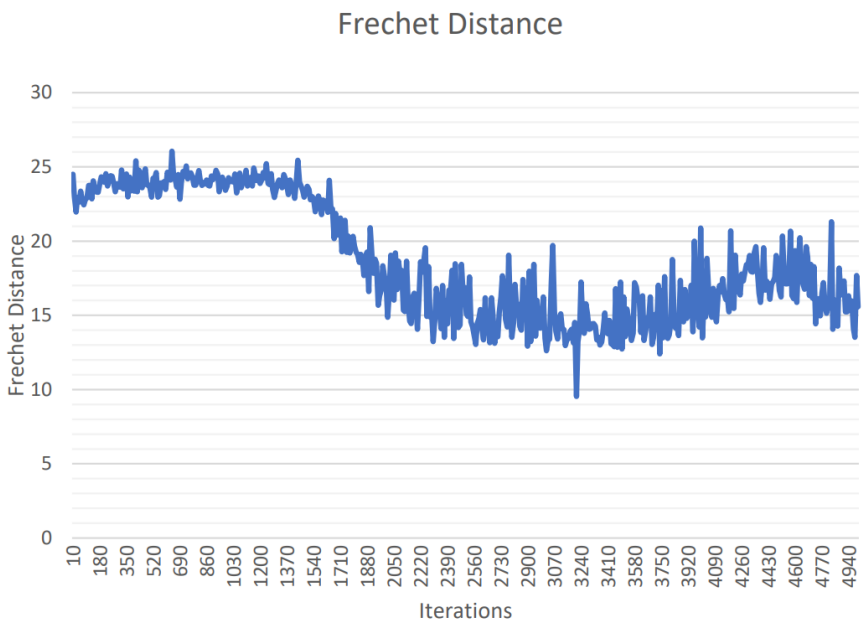

Figure 10: Frechet Distance Bonds and Atoms (Reproduced) for QGAN-HG

REPRODUCED DRUG PROPERTIES-

Diversity	Drugcandidate	logP	Novel	NP	QED	SA	Unique	Valid
1.00	0.4593	0.7101	1.00	0.7538	0.5081	0.2263	0.7857	0.875

Figure 11: Reprodued Drug property results for QGAN-HG

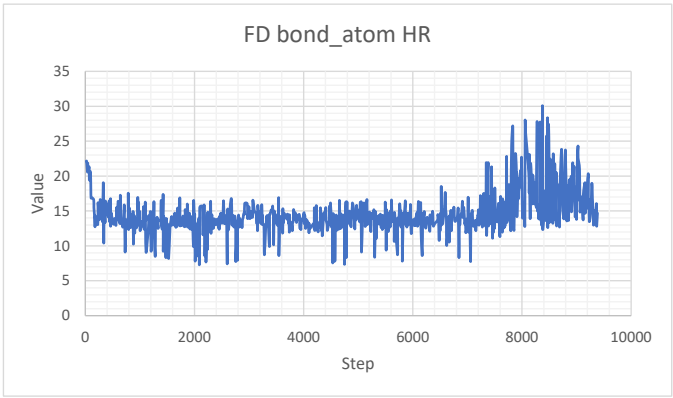

Figure 12: Frechet Distance Bonds and Atoms (Reproduced) for QuMolGAN-HR

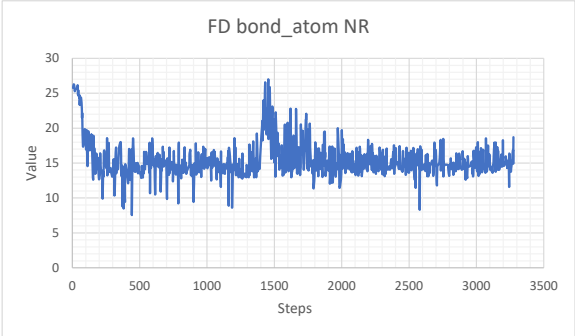

Figure 13: Frechet Distance Bonds and Atoms (Reproduced) for QuMolGAN-NR

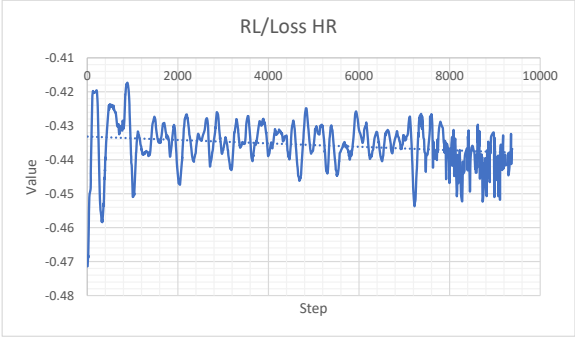

Figure 14: Reinforcement Loss (Reproduced) for QuMolGAN-HR

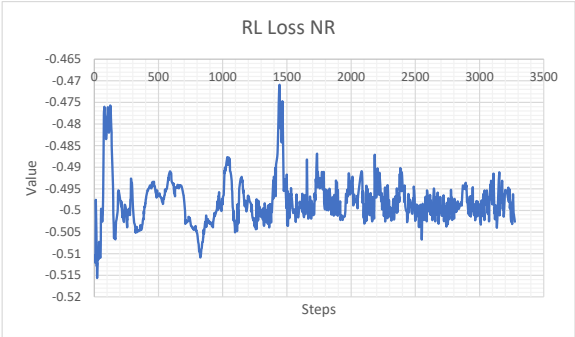

Figure 15: Reinforcement Loss (Reproduced) for QuMolGAN-NR

REPRODUCED DRUG SCORES HR-

Valid	Unique	Novel	NP	QED	Solute	SA	Diverse	Drug candidate
0.8472	0.1861	1.00	0.95	0.54	0.36	0.42	0.65	0.53

Figure 16: Reprodued Drug property results for QuMolGAN-HR

REPRODUCED DRUG SCORES NR (0 epochs)-

Valid	Unique	Novel	NP	QED	Solute	SA	Diverse	Drug candidate
0.5842	0.2319	nan	nan	nan	nan	nan	nan	0.40

Figure 17: Reprodued Drug property results for QuMolGAN-NR

reduction in the number of parameters used. We then train QuMolGAN (Quantum Noise) with different parameter complexities, in the first case with No reduction (NR), we train a quantum circuit with 8 qubits, 3 layers with a learning rate of 0.04. The corresponding bond and atom frechet distance, RL loss and drug property scores are shown in Figures 13, 15 & 17 respectively. Figure 7 shows an exemplary molecule generated. Note that the code was run on hpc, 4 haswell cpu nodes were allocated with a walltime of 2:00:00 hrs, however the model did't converge and the job terminated before it completed even a single iteration. Hence, a few of the drug properties show a nan value. In the second case we train a Highly reduced (HR) parameter architecture (HR,MR,NR correspond to different generator convolution dimensions respectively), with 4 qubits, 3 layers and learning rate of 0.04. The model was trained for 1 epoch before the job was forcefully terminated. The QuMolGAN-HR bond and atom frechet distance, RL loss and drug property scores are shown in Figures 12, 14 & 16 respectively. Figure 6 show exemplary molecules generated for this model. Notice how this outperforms even QGAN-HG in drug property scores like Drug candidate, QED, SA and NP scores, while it underperforms on quality metrics like diversity, validity and uniqueness. This is possibly due to the model being undertrained and not reaching convergence.

The potential of quantum computation in producing novel drug molecules is apparent and evidenced in this study. However, the method is still far from being perfect. Simulating quantum computations is computationally extremely expensive and scales exponentially with the number of qubits. Current hardware is still noisy and intermediate scale (NISQ) and simulation is not possible to perform on the large scale limiting the usecase.

5 Conclusions

Quantum Generative models clearly have the potential of outperforming classical models in terms of drug properties, while significantly reducing the number of parameters used and simplifying model architecture. We demonstrated the potential of generating training-set-like small molecules using different hybrid quantum classical models. The quantum advantage in expressive power is evident. However, the training processing is resource-consuming and time-consuming, even on state of the art classical computers.

6 Future scope

The quantum advantage of the quantum discriminator model (MolGAN-CQ) is yet to be demonstrated and couldn't be done during the course of this project. We made an effort to see if we can implement a problem specific ansatz and integrate it with these hybrid quantum classical generative models, exploiting a the presence of a group

theoretic structure in our dataset of drug molecules if any, taking inspiration from Glick et al. (2021) [5], However we concluded that finding such a symmetry in drug molecules is a research problem of it's own.

Finally, there is scope of diving into other classical generative methods like autoencoders and explore the quantum advantage there, as has been done by Li & Ghosh (2022) [6].

Acknowledgements

We would like to thank Prof. Abhishek Dixit and Prof. Kedar Khare for their continuous support throughout the project. We also show appreciation for the quantum machine learning modules developed by pennylane and IBM. We thank IIT Delhi HPC facility for computational resources. This project wouldn't have been possible without the people that supported us both directly and indirectly

References

[1] Pyrkov, A., Aliper, A., Bezrukov, D., Lin, Y., Polykovskiy, D., Kamya, P., Ren, F., Zhavoronkov, A. (2023). Quantum computing for near-term applications in generative chemistry and drug discovery. Drug Discovery Today, 28(8), 103675. https://doi.org/10.1016/j.drudis.2023.103675

[2] J. Li, R. O. Topaloglu and S. Ghosh, "Quantum Generative Models for Small Molecule Drug Discovery," in IEEE Transactions on Quantum Engineering, vol. 2, pp. 1-8, 2021, Art no. 3103308, doi: 10.1109/TQE.2021.3104804.

[3] https://pennylane.ai/qml/demos/tutorial_quantum_gans

[4] Exploring the Advantages of Quantum Generative Adversarial Networks in Generative Chemistry Po-Yu Kao, Ya-Chu Yang, Wei-Yin Chiang, Jen-Yueh Hsiao, Yudong Cao, Alex Aliper, Feng Ren, Alán Aspuru-Guzik, Alex Zhavoronkov, Min-Hsiu Hsieh, and Yen-Chu Lin Journal of Chemical Information and Modeling 2023 63 (11), 3307-3318 DOI: 10.1021/acs.jcim.3c00562

[5] Glick, J. R., Gujarati, T. P., Corcoles, A. D., Kim, Y., Kandala, A., Gambetta, J. M., & Temme, K. (2021). Covariant quantum kernels for data with group structure. arXiv preprint arXiv:2105.03406.

[6] Li, J. and Ghosh, S., 2022, March. Scalable variational quantum circuits for autoencoder-based drug discovery. In 2022 Design, Automation & Test in Europe Conference & Exhibition (DATE) (pp. 340-345). IEEE. https://ieeexplore.ieee.org/document/9774564

[7] De Cao, N., & Kipf, T. (2018). MolGAN: An implicit generative model for small molecular graphs. arXiv preprint arXiv:1805.11973.