Molecule Generation of Drugs Using VAE

K.B.Anusha*[1] Modalavalasa Divya[2] K.Madhuri Pravallikha Rani[3],B.Satvika[4]

P.Tarun[5], G.Vaishnavi[6], S.Linga Raju[7]

[1,2]Assistant Professor, *Department of Computer Science & Engineering, Aditya Institute of Technology and Management, Tekkali-532201, India*.

[3,4,5,6,7] *B. Tech. Students, Department of Computer Science & Engineering, Aditya Institute of Technology and Management, Tekkali-532201, India*.

anushakb91@gmail.com

Abstract. The field of drug discovery and development has witnessed a transformative change during the previous few years with application in artificial intelligence and ML techniques. Among these, Variational Autoencoders (VAEs) have emerged promising instrument for the generative designing of drug molecules. In relation to drugs discovery, VAEs have been employed to encode and decode chemical structures, facilitating the generation of drug molecules. This is achieved through the encoding of chemical configurations into an uninterrupted latent space, where the generative capacity of the model can be harnessed to create diverse and potentially pharmacologically relevant compounds. Key components of this approach include the representation of molecules as graphs or SMILES (Simplified Molecular Input Line Entry System) strings, In development of specialized loss functions to optimize characteristics of molecules, and the investigating the latent space to produce molecules with desired characteristics. Hence, In this project, we build compounds for drug discovery using a variational autoencoder.

Keywords: VAE,SMILES,KL Divergence loss, Categorical cross entropy

1 Introduction

In previous cases, A notable in the application of deep learning techniques in cheminformatics. Despite the fact that generative models in de novo drug creation has had a significantly greater impact, the utilization of DL techniques to swap out conventional machine learning techniques has had a considerable influence. Drug discov-ery is a field that seeks innovative medicinal molecules to solve complicated health concerns by combining cutting-edge science and computational innovation. Tradi-tional drug discovery involves a lengthy and resource-intensive process of screening chemical compounds to identify potential therapeutic agents. In recent years, a break-through method to drug molecule generation has evolved, harnessing the capabilities of generative models, particularly Variational Autoencoders (VAEs).

K. R. Madhavi et al. (eds.), *Proceedings of the International Conference on Computational Innovations and Emerging Trends (ICCIET 2024)*, Advances in Computer Science Research 112,
https://doi.org/10.2991/978-94-6463-471-6_17

Even though the original purpose of this kind of model was unsupervised learning, [1] [2] For both supervised and semi-supervised learning, its efficacy has been demonstrated [3][4].

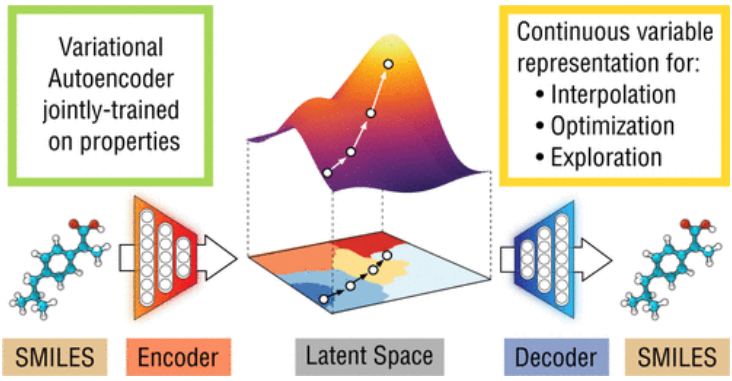

Fig. 1. General Structure of variational Autoencoder

A variational autoencoder is a model that produces a noise distribution similar to the prior one. These models are often trained using the expectation-maximization meta-algorithm. Such a technique is usually unmanageable, requiring the discovery of variational posteriors, or q-distributions, in order to maximize a lower constraint on the data probability. Usually, for each every data point, these q-distributions are parameterized in a separate optimization process. Conversely, variational autoencoders use an amortized NN approach to cooperatively optimize over data points. Using the real data points as input, the neural network creates parameters for the variational distribution.

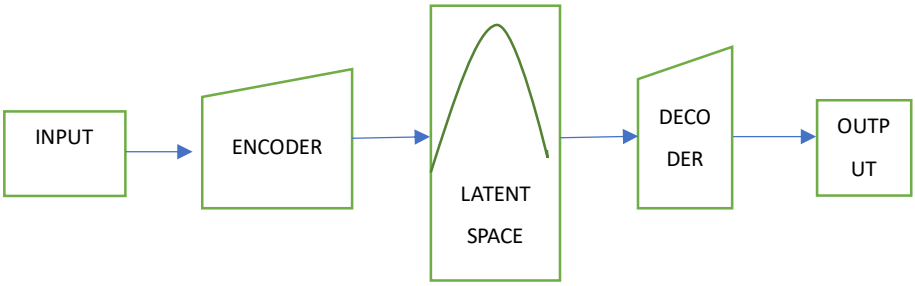

Fig. 2. Architecture of VAE

The architecture shown in Fig. 2 is designed to learn both the posterior of a probabilistic generative model, or encoder, and the model itself. We designate the continuous latent variable as z and the observation as x. The likelihood $p\theta(x|z)$, where θ is the learnable parameter, is used by the decoder to model the probabilistic generating process of x given z. To approximate the posterior, the encoder utilizes a model $q\varphi(z|x)$ characterized by φ.Both the encoder and the decoder are concurrently learned by raising the evidence lower bound (ELBO) of the marginal likelihood:

$$ELBO(\varphi,\theta) = Eq\varphi(z|x)\ logp\theta\ (x|z) - KL\ q\varphi(z|x)||p\ (z)$$

where the Kullback-Leibler (KL) divergence is denoted by KL, and the differential parameters are θ and φ. Gradient ascending can be used to enhance the evidence lower bound.
VAEs have a number of advantages. They can produce new data points from the learnt latent space distribution, which enables generative modeling. Additionally, they enable continuous representations of latent spaces, which enables us to create new data points by interpolating between various latent space points. Finally, because the probabilistic encoding motivates the model to provide a more reliable data representation,VAEs are less prone to overfitting than typical autoencoders.

2 Literature Survey

Based on DL, tactics have been created to address the drawbacks of earlier approaches as a result; the DL algorithms have become more popular in recent years. In order we express atomic information, the majority of existing techniques use deep networks to learn how to represent chemical properties and create the connection between atoms.
In 2020, ChengxiZang, Fei Wang , proposed MoFlow- a molecular graph-generating invertible flow model. Molecular graphs and their latent representations can be learned to map invertibly via flow, according to the graph generative model MoFlow. Our MoFlow uses a new graph flow to create atoms (nodes) with bonds,uses a posthoc validity correction to put them together into a molecular graph that is valid chemically,and first uses Glow-based model to construct bonds (edges).

In 2018, Rafael Gómez-Bombarelli, Jennifer N. Wei ,suggested the model outlined in the publication. He uses continuous, Data-driven molecular representation, automatic chemical design creates novel molecules by effectively exploring the open-ended areas within chemical compounds. Three parts make up the model: the encoder, decoder, and predictor. A molecule's discrete representation is transformed The encoder con-verts the input data into a continuous real-valued vector, which is subsequently con-verted returning to discrete models of molecules by decoder. Based on the molecule's latent continuous vector representation, the Predictor makes chemical property esti-mates. Gradient-based optimization can effectively direct the search for optimal func-tional molecules when continuous representations are used.

Table 1. Researchers Evolutions on different models Descriptions

S.No.	Paper Source	Dataset	Model	Result
1.	[6]	Zinc dataset	MOFLOW	Moflow find the similar molecules and property improvement.
2.	[7]	Zinc dataset	LSTM,Cycle GAN	Our LA-CycleGAN model achieved 0.62 in replicating the optimal LogP distribution for the collection of Aliphatic Rings when compared to

				Mol-CycleGAN. Both the overall molecular weight and the overall Mol-Weight have increased.
3.	**[8]**	KEGG, DrugBank, SuperTarget dataset	CCGVAE	The model enhances CGVAE in both datasets; in particular, the reconstruction job performs better when the histogram of valences is used.
4.	[9]	QM9 dataset,PDB bind dataset	Scalable Quantum Generative Autoencoder	Following learning, including classical and quantum AEs.More precise reconstruction is achieved by avoiding a longer learning period through the inefficiency of quantum circuit simulation.
5.	[10]	CHEMBL,ZINC, MOSES dataset	RNN	MolecularRNN learns diverse distributions through unsupervised pretraining, generating 98% valid molecules in inference.
6.	**[11]**	Qm9 chemical dataset	MolGAN	Despite having a 90% novelty score, CharacterVAE performs poorly in terms of validity. Conversely, MolGAN has good marks for both originality and validity.
7.	[12]	QM9,CHEMBL dataset	LSTM,GuacaMol Benchmark baselines	Overall, this model performs better than standard graph-based techniques. Utilize GuacaMol's distribution-learning metrics to assess the model, including validity, novelty, uniqueness, KL-divergence47, and Fréchet ChemNet Distance48 scores.
8.	[13]	CHEMBL dataset	Latent GAN	The percentage of LatentGAN-generated targets that were anticipated to be active for EGFR, HTR1A, and S1PR1 was 71%, 71%, and 44%, respectively.

9.	[14]	ZINC dataset	AE	When training alongside a property prediction task, the autoencoder demonstrated strong predictive power and the capacity to optimize molecules gradient-wise in the resulting smoothed latent space.
10.	[15]	MNIST and Frey Face datasets3	Variational Bayes,SGVB	A novel estimator of the variational lower bound, Stochastic Gradient VB (SGVB), for efficient approximate inference with continuous latent variables

3 Methodology

We build the pharmacological compounds in this study using variational autoencoder.To ensure that the latent space has favourable qualities that support the generative process and to avoid over fitting, an autoencoder whose training is regularized is called a variational autoencoder. The goal of a variational autoencoder, like a standard autoencoder, lowers the amount of reconstruction error that exists in between encoded-decoded data, original data.

It has an encoder and a decoder built into its architecture. To integrate this, however, we do tweak the encoding-decoding process somewhat after the latent space has been regularized. In particular, we represent an input not as a single point but as a distribution throughout the la-tent space. The molecule's feature matrix and graph adjacency matrix are fed into the encoder. The attributes z_mean and log_var, which describe the molecule's latent-space representation, are first processed by a convolution layer before being compressed and managed by several dense layers.

3.1 Graph Convolution layer:

Neighborhood aggregations with non-linear modifications are implemented using the relational graph convolution layer. The following are these layers:

A_hat * H_hat**(l+1) * W**(l) * D_hat**(-1) = σ(H_hat**(l+1))

In the case of the inverse di-agonal of A, the degree tensor is D_hat** (-1). A_hat, W_hat**(l) represents the trainable weight tensor at the l-th layer, and σ indicates the non-linear transformation (typically a ReLU activation,For any bond type (relation), the degree tensor expresses precisely how many bonds are connected to each atom in the diagonal.

Upon receiving a latent-space representation as input, the Decoder forecasts the feature matrix and graph adjacency matrix of the pertinent molecules.

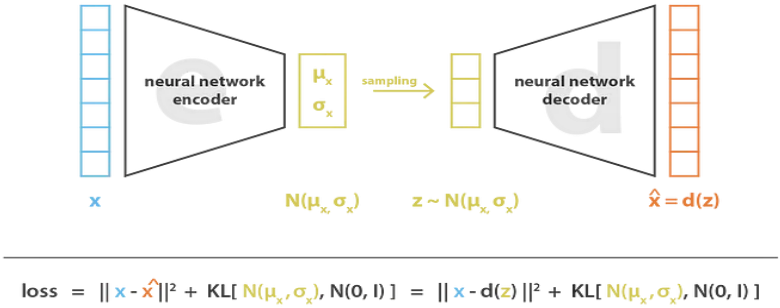

Fig. 3. Illustration of an variational autoencoder with its loss function.Data Security Measures

This model optimizes 4 losses during training: graph loss, KL divergence loss, loss in property prediction, and categorical crossentropy.

The accuracy of the model and reconstruction is evaluated by the loss function for categorical cross-entropy. Root MSE discrepancy between the real and anticipated features is computed by the feature prediction loss after the latent representation has been run through the feature prediction model. Model feature prediction is optimized with the usage of binary cross entropy. The model and characteristic (QED) prediction also have an impact on the gradient penalty. The gradient cross model of the original neural network is improved by the gradient penalty, which is a soft 1-Lipschitz continuity this gives the loss function more regularity. Our model would be inferred to predict across random latent space, and we would attempt to produce 100 new legitimate molecules.

4 Results and Discussions

SMILES	logP	Qed	SAS
COc1ccc(C(=O)N(C)[C@@H](C)C/C(N)=N/O)cc1O	0.9978	0.327297	2.852316
C[C@@H]1CC(Nc2cncc(-c3nncn3C)c2)C[C@@H](C)C1	3.1137	0.928975	3.432004
N#CC1=C(SCC(=O)Nc2cccc(Cl)c2)N=C([O-])[C@H](C#N)C12CCCCC2	3.60956	0.789027	4.035182
CCOC(=O)[C@@H]1CCCN(C(=O)c2nc(-c3ccc(C)cc3)n3c2CCCCC3)C1	4.00022	0.690944	0.690944
CC[NH+](CC)[C@](C)(CC)[C@H](O)c1cscc1Br	2.6374	0.824369	5.091438
CC(C)(C)c1ccc2occ(CC(=O)Nc3ccccc3F)c2c1	5.0506	0.702012	2.084095

For virtual screening, we utilize the ZINC dataset, a publicly available database of pharmaceuticals.. Data set contains chemical formulas in SMILES representation in conjunction with pertinent chemical measures, including QED (Qualitative Estimate of Drug-likeness), SAS (Synthetic Accessibility Score), and logP (Water–Octanal Partition Coefficient).training on the Zinc 250K dataset, which contains 250,000 molecules similar to drugs, in that order. A random subset of 1000 ZINC 250K was used to train a VAE.

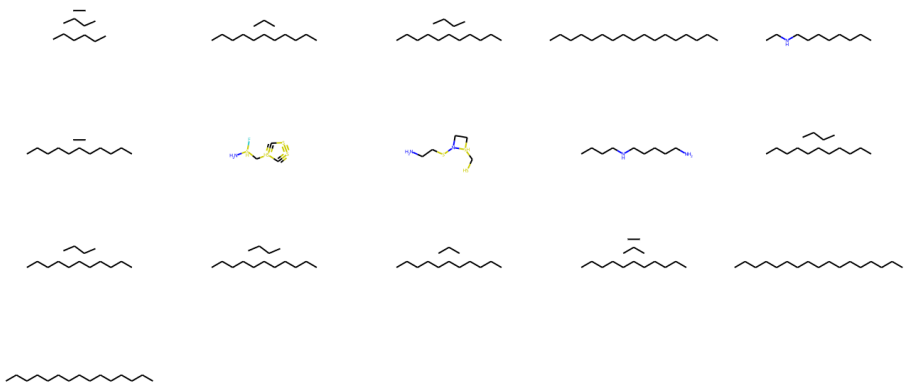

Fig. 4. Legitimate molecules from various locations in the latent space

The above Fig. 4 creates new, legitimate molecules from various locations in the latent space. The model's parameters are changed during training in order to minimize a loss function. The difference between the target values in the training data and expected output can be computed by loss function. Reducing this discrepancy shows that the model is picking up the ability to make precise predictions from the training set.

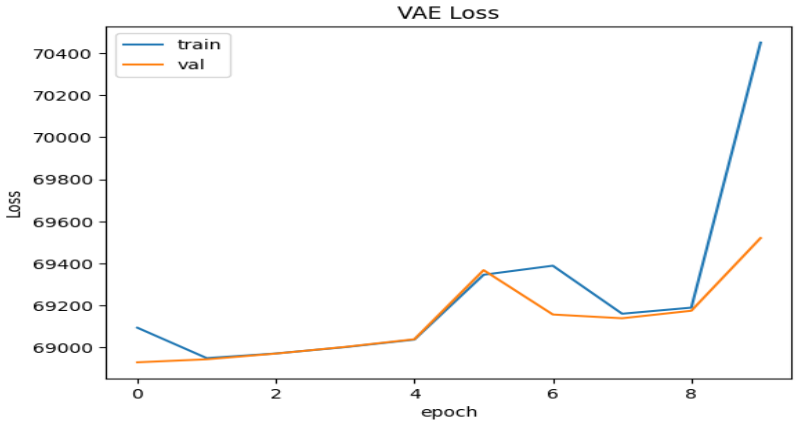

Fig. 5. The training and validation loss

The above Fig. 5 display the training and validation loss. Goal of our model is to generalize successfully to new, previously unknown data. To ensure this, the model's performance is assessed against a validation set that it has never seen before. The validation loss is derived using the same loss function as the prediction loss, but it is applied to the validation set's predictions. A lower validation loss means that the model is not overfitting to the training data, implying higher generalization performance.

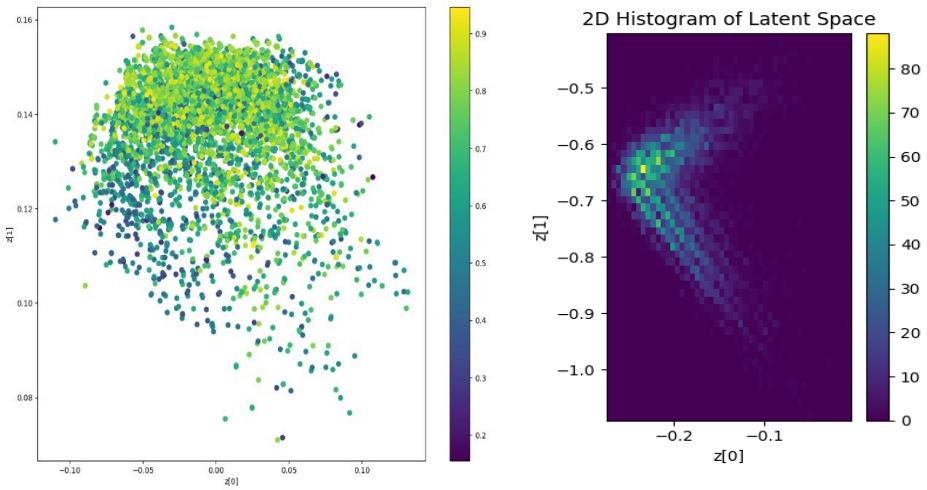

Fig. 6. Molecular property-related latent space clusters

The above Fig. 6 show molecular property-related latent space clusters (SAS).The latent space is a compressed representation of the input features that is anticipated to capture crucial SAS-related qualities. Each point in the latent space reflects a distinct encoding of the synthetic accessibility of a chemical.

5 Conclusion

Over several decades, de novo drug design has evolved. Techniques based on both structure and ligand have made substantial progress. In this paper, we propose a Molecule Generation of Drugs using VAE. Due to the characteristics of SMILES strings, VAE-based architecture frequently are very good Potential SMILES candidates creation utilizing char-by-char. The VAE alone generative models are capable of producing chemically valid SMILES strings once they have learned the SMILES syntax. Designing compounds with desired molecular characteristics is critical in drug discovery. To produce this biased molecular design, four optimization methodologies are used. A tiny dataset with specific desirable qualities is utilized to fine-tune . This overarching structure for producing molecules that are prejudiced clearly relies on a limited amount of well-known labeled data. Four tasks are used to validate this model: property optimization, constrained property optimization, display of the

continuous latent space, and production and reconstruction of molecular graphs. This technique enables a novel type of guided gradient-based chemical space search.

References

1. Dilokthanakul, Nat; Mediano, Pedro A. M.; Garnelo, Marta; Lee, Matthew C. H.; Salimbeni, Hugh; Arulkumaran, Kai; Shanahan, Murray (2017-01-13). "Deep Unsupervised Clustering with Gaussian Mixture Variational Autoencoders". arXiv:1611.02648 [cs.LG].
2. Hsu, Wei-Ning; Zhang, Yu; Glass, James (December 2017). "Unsupervised domain adaptation for robust speech recognition via variational autoencoder-based data augmentation". 2017 IEEE Automatic Speech Recognition and Understanding Workshop(ASRU).pp. 1623. arXiv:1707.06265.
3. Ehsan Abbasnejad, M.; Dick, Anthony; van den Hengel, Anton (2017). Infinite Variational Autoencoder for Semi-Supervised Learning
4. Xu, Weidi; Sun, Haoze; Deng, Chao; Tan, Ying (2017-02-12). "Variational Autoencoder for Semi-Supervised Text Classification". Proceedings of the AAAI Conference on Artificial Intelligence. 31 (1).
5. Kumar, DNS Ravi, N. Praveen, Hari Hara P. Kumar, Ganganagunta Srinivas, and M. V. Raju. "Acoustic Feedback Noise Cancellation in Hearing Aids Using Adaptive Filter." International Journal of Integrated Engineering 14, no. 7 (2022): 45-55.
6. Avanija, J., K. E. Kumar, Ch Usha Kumari, G. Naga Jyothi, K. Srujan Raju, and K. Reddy Madhavi. "Enhancing Network Forensic and Deep Learning Mechanism for Internet of Things Networks." (2023).
7. Wang F, Feng X, Guo X, Xu L, Xie L and Chang S (2021) Improving de novo Molecule Generation by Embedding LSTM and Attention Mechanism in CycleGAN. Front. Genet. 12:709500.
8. Constrained Graph Variational Autoencoders for Molecule Design Qi Liu, Miltiadis Allamanis, Marc Brockschmidt, Alexander L. Gaunt.
9. J. Li and S. Ghosh, "Scalable Variational Quantum Circuits for Autoencoder-based Drug Discovery," 2022 Design, Automation & Test in Europe Conference & Exhibition (DATE), Antwerp, Belgium, 2022, pp. 340-345.
10. Raju, S. Viswanadha, A. Vinaya Babu, G. V. S. Raju, and K. R. Madhavi. "W-Period Technique for Parallel String Matching." IJCSNS 7, no. 9 (2007): 162.
11. Kumar, Voruganti Naresh, U. Sivaji, Gunipati Kanishka, B. Rupa Devi, A. Suresh, K. Reddy Madhavi, and Syed Thouheed Ahmed. "A Framework For Tweet Classification And Analysis On Social Media Platform Using Federated Learning." Malaysian Journal of Computer Science (2023): 90-98..
12. Mahmood, O., Mansimov, E., Bonneau, R. et al. Masked graph modeling for molecule generation. Nat Commun 12, 3156 (2021).
13. Prykhodko, O., Johansson, S.V., Kotsias, PC. et al. A de novo molecular generation method using latent vector based generative adversarial network. J Cheminform 11, 74 (2019).
14. Rafael Gómez-Bombarelli,Jennifer N. Wei,David Duvenaud,José Miguel Hernández Lobato, Benjamín Sánchez-Lengeling,Dennis Sheberla Jorge Aguilera-Iparraguirre,Timothy D. Hirzel,Ryan P. Adams, and Alán Aspuru-Guzik
15. Kingma, Diederik P.; Welling, Max (2013-12-20). "Auto-Encoding Variational Bayes". arXiv:1312.6114 .