Deep Learning based molecule generation for developing novel therapeutics for neurodegenerative diseases

By
Piyush Joshi
MT22194

Under the Supervision of Dr. N. Arul Murugan
Indraprastha Institute of Information Technology, Delhi
May, 2024

Deep Learning based molecule generation for developing novel therapeutics for neurodegenerative diseases

By
Piyush Joshi
(MT22194)

Submitted
in partial fulfillment of the requirements for the degree of Master of Technology

To
Indraprastha Institute of Information Technology Delhi Month, 2024

CERTIFICATE

This is to certify that the thesis titled *"Deep Learning based molecule generation for developing novel therapeutics for neurodegenerative diseases"* being submitted by Piyush Joshi (MT22194) to the Indraprastha Institute of Information Technology, Delhi, for the award of the Master of Technology, is an original research work carried out by her under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

May, 2024, Dr. N. Arul Murugan

Department of Computational Biology

Indraprastha Institute of Information Technology, Delhi

New Delhi 110 020

ACKNOWLEDGEMENTS

I want to sincerely thank Dr. N Arul Murugan, who oversaw my thesis, for his invaluable guidance and unwavering support during the project. Without his support and competent understanding, this thesis could not have been completed. I want to express my gratitude to the faculty and staff at IIIT, Delhi, for their unwavering support during the whole undertaking. I am appreciative of my fellow classmates' encouragement and assistance. Lastly, I want to express my gratitude to my family, who have always supported and inspired me. Without them, my academic journey would not have been possible.

ABSTRACT

Deep learning models and generative AI have revolutionized drug discovery, especially with regard to neurodegenerative diseases like Alzheimer's. The developing of canonical SMILES (Simplified Molecular Input Line Entry System) for certain targets, such as BACE (Beta-secretase),a crucial enzyme linked to the disease progression of neurodegenerative disease, is a noteworthy application of these technologies. These models make use of large datasets to identify complex patterns in chemical structures, which allows them to synthesize new compounds with the necessary properties. The trick with BACE is to design compounds that can block its action only, without causing unwanted side effects.

Driven by advanced deep learning architectures like transformers or bidirectional recurrent neural networks (RNNs), generative AI generates and refines SMILES strings iteratively until they meet predetermined standards for drug-like qualities, selectivity, and efficacy. This method drastically shortens the time and cost required for experimental synthesis and assessment, which speeds up the drug discovery process. These AI-driven approaches also make it easier to explore over large chemical landscapes, which may reveal new treatment prospects that traditional approaches might miss.

As generative AI is iterative, researchers may gradually refine and enhance the quality of the compounds that are produced. Through the integration of feedback derived from computational assessments and experimental data, the model is capable of improving upon its errors and producing BACE inhibitors that are more potent and the compounds that are produced are docked against the BACE molecule to calculate the biding affinity,how effectively the molecule can bind with the BACE molecule. Iterative processes reduce the need for expensive and time-consuming laboratory studies while accelerating the identification of promising drugs candidates.

Apart from producing new compounds, generative AI can also be applied to refine lead compounds that already exist for improved BACE inhibition. Through iterative modifications of the molecules based on feedback from experiments and computer predictions, researchers can optimize the attributes of lead compounds to enhance their safety, selectivity, and efficacy.

All things considered, deep learning models and generative AI have great potential to advance drug discovery for neurodegenerative illnesses. These technologies enable researchers to find Alzheimer's and other debilitating disorders' cures more quickly by fusing complex algorithms with computational capacity. There is growing optimism over the development of transformational therapeutics for neurodegenerative disorders as ongoing research continues to hone and enhance the capabilities of generative AI and deep learning models.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

1.Introduction:

1.1 Drug designing

Within the dynamic landscape of drug discovery and molecular design, the integration of advanced computational methods has become indispensable [1]. Generative models have emerged as a transformative approach for generating novel molecules with targeted properties, rooted in machine learning algorithms [3]. This introduction aims to explore the practical application of generative models in molecular design, delving into their methodologies, real-world implementations, and implications across various industries [5]. By comprehensively understanding the principles and capabilities of generative models, researchers and professionals can harness their potential to drive innovation and address pressing challenges in drug discovery and materials science [44].

Generative models represent a subset of artificial intelligence techniques focused on creating new data instances from a given dataset [2]. In molecular design, these models analyze extensive repositories of chemical structures and associated properties to learn underlying patterns and relationships [25]. Through iterative training processes, generative models acquire the ability to generate novel molecular structures that adhere to specified constraints and objectives [27].

1.2 Generative ai for drug designing

One primary application of generative models in molecular design is the discovery of novel drug candidates [28]. Traditional drug discovery processes are often time-consuming and costly, necessitating extensive experimental validation of potential compounds. Generative models offer a complementary approach by generating virtual compound libraries, subsequently screened for desirable pharmacological properties using in silico methods [29]. This accelerates early-stage drug discovery, enabling researchers to focus resources on the most promising candidates.

Moreover, generative models play a crucial role in materials science by facilitating the design of new materials with tailored properties, from polymers to catalysts [30]. This ability to generate molecular structures with predefined characteristics holds immense potential for advancing various industrial applications [31]. Leveraging generative models, researchers can explore vast chemical spaces and identify novel materials with enhanced performance, durability, and sustainability.

However, the adoption of generative models in molecular design presents challenges. Ensuring the reliability and interpretability of generated molecular structures is paramount [32]. While generative models produce vast quantities of novel compounds, verifying their safety, efficacy, and adherence to regulatory standards remains a critical consideration. Additionally, integrating domain knowledge and

expert insights is essential for guiding the generation process and validating the feasibility of generated molecules [44].

Despite challenges, rapid progress in generative modeling techniques offers exciting opportunities for innovation in drug discovery and materials science [20]. Advanced algorithms, coupled with increasing computational power, enable researchers to explore complex chemical spaces and discover novel molecular structures with unprecedented efficiency [14]. Interdisciplinary collaboration between computer scientists, chemists, and materials engineers fosters synergy in developing robust generative models and translating their outputs into tangible solutions [9].

In conclusion, the integration of generative models in molecular design represents a paradigm shift in drug discovery and materials science [26]. By harnessing the power of artificial intelligence, researchers can expedite the process of discovering novel molecules with desired properties, paving the way for breakthroughs in healthcare, materials technology, and beyond [5]. As we advance our understanding of generative modeling techniques, the possibilities for innovation are limitless, promising a future driven by data-driven insights and computational ingenuity [7].The use of the generative ai models is to make the novel molecules against the bace inhibitor,There are several generative ai models that are used to make the models against the bace inhibitor [8].

1.3 Virtual screening of molecules after model deployement

In the realm of molecular design, artificial intelligence has revolutionized the discipline by providing researchers with strong tools to explore large chemical regions and expedite the drug development process [10]. AI is particularly good at creating new compounds with certain characteristics or target interactions, opening up possibilities that are not possible with conventional techniques. Researchers can input target profiles or desired structural properties using complex algorithms, and the system will suggest possible candidate compounds [11]. This strategy has found compounds that manual procedures alone could have missed, greatly improving molecular design [20.

The next crucial stage after creating compounds is assessing their safety and effectiveness profiles. During this stage, a computational method called virtual screening is essential because it mimics the chemical interactions that occur between the produced compounds and particular target proteins or biomolecules. Because beta-secretase 1 (BACE) contributes to the synthesis of amyloid-beta peptides, which are associated with the advancement of Alzheimer's disease, it is a major focus in research on the condition. Sophisticated algorithms and simulations are utilized in virtual screening to forecast the binding affinity and interaction patterns of potential compounds with the target protein [23].

The drug development process has been optimized through the integration of AI in virtual screening and molecular design, which combines rigorous experimental

validation with computational predictions [31]. By fine-tuning generative models using screening data, machine learning algorithms enhance the system's capacity to produce molecules with the required characteristics or interactions. therapeutic discovery has accelerated thanks in part to this iterative process, which may also lead to the discovery of innovative therapeutic candidates with significant potential to improve patient outcomes [32].

The drug development process has been optimized through the integration of AI in virtual screening and molecular design, which combines rigorous experimental validation with computational predictions [35]. By fine-tuning generative models using screening data, machine learning algorithms enhance the system's capacity to produce molecules with the required characteristics or interactions [36]. therapeutic discovery has accelerated thanks in part to this iterative process, which may also lead to the discovery of innovative drugs with significant potential to improve patient outcomes [37].

CHAPTER 2

2.Materials and Methods:

2.1 Variational autoencoder –

The model described in the paper titled facilitates the generation of new molecules by effectively exploring vast chemical compound spaces. This model comprises three main components: Encoder, Decoder, and Predictor [1]. The Encoder converts the discrete representation of a molecule into a continuous vector in the real-valued domain, while the Decoder performs the opposite operation, converting continuous vectors back into discrete molecule representations [2]. Additionally, the Predictor estimates chemical properties based on the latent continuous vector representation of the molecule. By employing continuous representations, the model enables efficient navigation through chemical compound spaces to discover optimized functional compounds [3].

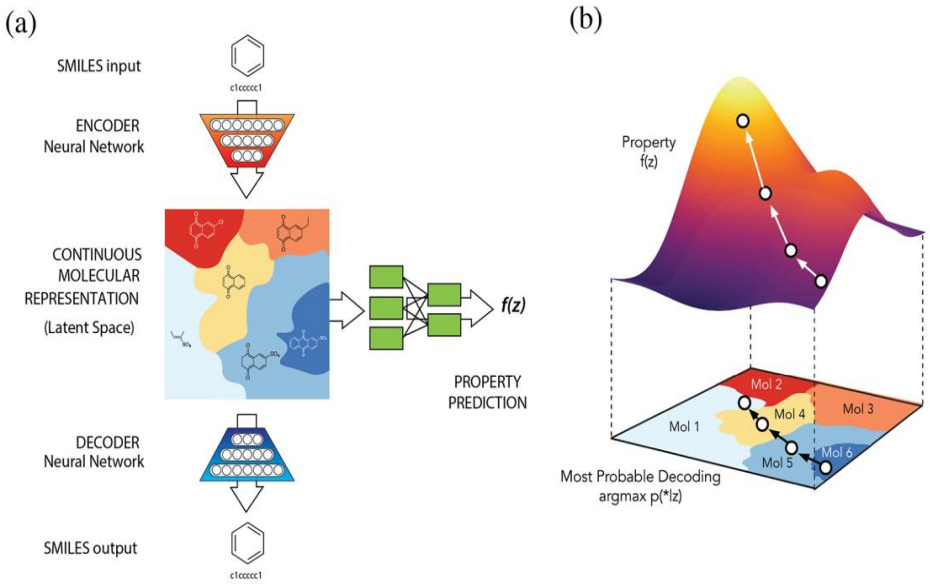

Figure – 1 variational autoencoders.

In Figure (a), we observe an autoencoder utilized for molecule design, along with a joint property prediction model. The process starts with a discrete representation of a molecule, like a SMILES string, which undergoes transformation by the encoder network. This network converts each molecule into a vector within the latent space, effectively creating a continuous representation [9]. Subsequently, given a point in the latent space, the decoder network generates a corresponding SMILES string. Additionally, a multilayer perceptron network estimates the value of target properties linked with each molecule [10].

Figure (b) demonstrates gradient-based optimization within the continuous latent space. After training a surrogate model to predict the properties of molecules based on their latent representation, the optimization procedure begins [11]. By optimizing the surrogate model concerning the latent representation, new representations are identified, expected to align with specific desired properties. These new representations can then be decoded into SMILES strings, allowing for empirical testing of their properties [12].

The molecule that is generated by the variational autoencoder by using the zinc dataset having 2 lakh smiles in the dataset.

Dataset – The dataset is comprises of the smiles and the corresponding property of the smile which is qed,logP,SAS.

Training Dataset – 249455 of the smiles molecules that are present in the zinc dataset

	smiles	logP	qed	SAS
0	CC(C)(C)c1ccc2occ(CC(=O)Nc3ccccc3F)c2c1	5.05060	0.702012	2.084095
1	C[C@@H]1CC(Nc2cncc(-c3nncn3C)c2)C[C@@H](C)C1	3.11370	0.928975	3.432004
2	N#Cc1ccc(-c2ccc(O[C@@H](C(=O)N3CCCC3)c3ccccc3)...	4.96778	0.599682	2.470633
3	CCOC(=O)[C@@H]1CCCN(C(=O)c2nc(-c3ccc(C)cc3)n3c...	4.00022	0.690944	2.822753
4	N#CC1=C(SCC(=O)Nc2cccc(Cl)c2)N=C([O-])[C@H](C#...	3.60956	0.789027	4.035182

Figure 2 – Zinc dataset.

The model is trained on the whole zinc dataset with the only taken property in the model which is qed property.
The molecules that are produced by the variational autoencoder is docked against the bace inhibitor and get the results from that.

2.2 GPT-2 –

Utilizing learned representations from a broad-domain dataset offers a solid starting point for subsequent tasks. This transfer of domain knowledge proves particularly valuable in scenarios with scant data, where models trained from the ground up struggle to discern patterns from limited samples [16]. The effectiveness of transformer models has cemented transfer learning as the go-to approach for learning downstream tasks in natural language processing [17]. Tools like HuggingFace Transformers, BertViz, and Adapter Hub have gained widespread adoption among practitioners in the NLP community.
Our transformer decoder is designed with similarities to GPT-2, but it employs a distinct approach to tokenization [18]. Instead of byte-level encoding, we utilize

character-level byte-pair encoding [18]. Initially, we establish a base vocabulary by reserving 72 characters from the SMILES alphabet. We then enhance this vocabulary by integrating up to 1000 of the most frequent merges [20]. The model incorporates parameterized token and position embeddings, with 8 attention heads and 4 attention blocks. It operates with an embedding/hidden dimension of 512, totalling 13.4 million parameters [34].

Training dataset – chEMBL dataset contains 2327927 molecules which are used for training the model.

Inhibitor dataset – 7234 molecules are present in the dataset.

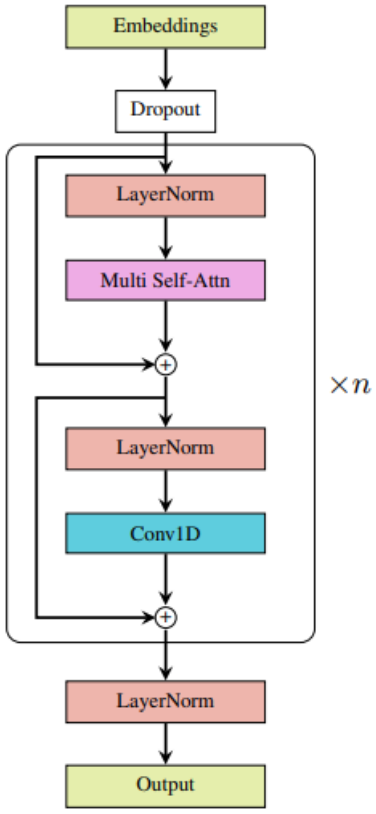

Figure – 3 GPT-2 architecture.

1. We have smiles gpt package that contains all the necessary information about the canonical smiles that we are using in this part [21].
2. The GPT-2 model is taken from the hugging face library,we have used all the tokenizers and the token file from the 10 million checkpoint because the model is trained on the 10 million pubchem dataset [22].
3. Vocabulary,merges checkpoints and tokenizers are taken from the already saved checkpoint that is trained on the 10 million molecules of the pubchem.We use the tokenzers and the checkpoint details of the 10 million pubchem dataset on which the gpt-2 model is run and the information is stored.
4. Train the gpt-2 model with this configuration on the chembl dataset that contains the canonical smile and generating the smiles [23].

5. Taking the vocabulary and the tokenizers from the 10 million dataset,we take this and just used in finetuning on the chembl dataset that contain the 2 million smiles strings [24].
6. The model is taken from the huggingface library usually used to taken the given model from the huggingface library.
7. The model is finetuned on the whole chembl dataset smiles and the molecules are generated [26].
8. The generated molecules are docked against the bace inhibitor to get the results in this [27].

2.3 FBRNN(Forward-backward rnn)

Within a recurrent neural network (RNN) framework, the Forward-Backward Recurrent Neural Network (FBRNN) model is an intricate architecture that integrates the advantages of both forward and backward information flow [31]. The FBRNN uses bidirectional processing to capture context from both past and future inputs, unlike traditional RNNs that only process sequential data in one direction [32]. This improves the network's understanding of temporal dependencies and performance in tasks like sentiment analysis, speech recognition, and sequence labeling.

The forward RNN, which processes input data in the natural order, and the backward RNN, which processes the input data in reverse order, are the two RNNs that operate in tandem to form the basis of the FBRNN [33].

The FBRNN can extract deeper contextual information and produce predictions that are more accurate by combining data from both directions [34]. This two-way method works especially well in situations when decision-making requires context from both past and future components.

The FBRNN model's superiority over conventional RNNs in managing long-range dependencies is one of its main features [35]. The model can take into account a wider context when making predictions since the forward pass gathers information about past inputs and the backward pass gets information about future inputs [36]. This ability is useful for activities like machine translation, where it's necessary to know the words that come before and after a sentence or phrase in order to understand its context [37].

Furthermore, as compared to unidirectional RNNs, the FBRNN model shows better generalization and robustness. The model may develop more resilient representations and handle input sequence fluctuations more effectively by utilizing bidirectional processing, which improves performance on unobserved data [39]. Because of this, the FBRNN is a useful tool for applications that need to be highly accurate and reliable in practical settings [40].

The FBRNN has been effectively used in a variety of disciplines in practical

applications [41]. It has been applied to natural language processing applications where knowledge of the sentence's word context is crucial, like named entity recognition, part-of-speech tagging, and sentiment analysis [42]. Because the FBRNN can capture temporal relationships, speech recognition accuracy has increased when transcribing spoken language [43].

In terms of sequence modeling and analysis, the Forward-Backward Recurrent Neural Network (FBRNN) model is a major development. Compared to conventional unidirectional RNNs, the FBRNN achieves improved levels of accuracy and robustness by utilizing bidirectional processing to capture extensive contextual information [44]. Because of this, it's a useful tool for many applications in natural language processing, machine learning, and artificial intelligence.

Training dataset – chEMBL dataset contains 2327927 molecules which are used for training the model.
Inhibitor dataset – 7234 molecules are present in the dataset.

1.In this the canonical smiles dataset is divided into the train-test split into the train dataset and to the test dataset.

2.The sequence that is present in the train dataset and to the test dataset is one hot-encoded labels [34]

3.Take the index of the maximum value of the characters in the sequence,we are one hot encoded each character of the sequence and then we are taking that index that shows the maximum index in the model.

4.In bidirectional lstm we have no knowledge of the future tokens we have the only knowledge of the tokens where we are proceeding [35].

5.In this we are using the TwoOutLSTM_v2 is not the bidirectional instead of this,it processes the the input sequence in a single direction,After processing the input sequence the models output is divided into the two parts,the forward prediction and to the backward prediction [36].

6.The forward prediction represents the model’s prediction based on processing the input sequence input sequence in the forward direction,similarly the backward prediction here is the prediction based on processing the input sequence in the reverse direction (where the model does not actually performing by breaking the output into the two we are assuming that we are processing the input sequence in both the direction).

7.The mean cross entropy loss is calculated for both the forward and the backward predictions separately,the losses that we are getting from the forward and the backward prediction are added together to get the total loss of the molecule [38].

8.The total loss for the molecule is used to compute gradients via the backpropagation method in this (molecule_loss.backward()),We are doing that

making the gradients are zeroed (self._optimizer.zero_grad()) before computing the gradients for the current molecule,The optimizer (torch.optim.Adam) takes the step based on the gradients to update the model parameters (self._optimizer.step()).

9.The loss computation per token are stored in the **statistic** array,we are normalizing these losses by dividing with the number of tokens that are present in the molecule [39].

10.The training of the model can be done by iterating over the epoch and the batches,updating the model parameters using the mini-batch gradient techniques.

11.The training loss is calculated for the model and compared with the validation loss,here the validation dataset is taken as the test dataset in this [40].

12.The trained model is saved and 100 unique canonical smile is generated using the **sampler.py** file.

13.The trained model is utilizing the dataset of the ChEmbl smiles dataset has **2327927** molecules in the dataset out of these,the pre-processed molecule getting after the preprocessing of the dataset is done is **1048576 [41].**

14.The epoch of the trained model are saved and taking the 9^{th} epoch of the trained model to fine tuned on the inhibitor dataset [42].

15.

```
[MODEL]
model = FBRNN
hidden_units = 256

[DATA]
data = SMILES_BIMODAL_FBRNN_fixed
encoding_size = 55
molecular_size = 77

[TRAINING]
epochs = 10
learning_rate = 0.001
n_folds = 1
batch_size = 512

[EVALUATION]
samples = 100
temp = 0.7
starting_token = G

[FINETUNING]
start_model = FBRNN_fixed_512
```

16. For fine-tuning of the model the below [FINETUNING] should be added above,

[FINETUNING]

start_model = FBRNN_fixed_512

the data should be preprocessed separately for the training dataset and to the finetuning dataset [43].

17.The fine-tuning dataset contains the 7235 molecules after the pre-processing of the dataset is done 5794 molecules are generated in the csv file taking the 9th epoch of the trained model indexing from 0 to 9 and having the 10 epochs in this [39].The model is finetuned on this dataset and the molecules are generated using the sampler.py files which are unique molecules [40].

18.The tanimoto similarity is calculated between the generated molecules and the training dataset if the generated molecule shows the tanimoto similarity less than 100% then we can save that canonical smiles in the csv file,similar to that one the molecule that is generated after the finetuning of the model on the inhibitor dataset for this one also the molecules that is generated after the fine-tuning of the model,tanimoto similarity is calculated between the molecules that is generated after the fine-tuning of the model and to the whole ChEMBL smiles dataset which contain **2327927** molecules [41].

19.The model take the Starting token as the G token,from the G token the forward prediction that we are taking is generating the sequence from the left to right direction from the token G,and the backward prediction is generating the token from the right to left direction in the molecule,the final molecule that is generated after this does not include the token G in the molecule [40].

20.In this the forward and the backward model generate the tokens from the token “G” two tokens are generated on each time step on each sides [42].

Figure 4 – Working of the FBRNN generating of the tokens at both the time-steps in each direction.

CHAPTER 3

3 Results and Discussion

3.1 Variational autoencoder –The result of the generated molecules from the zinc dataset by using the variational autoencoder which shows the binding affinity mentioned below.

1w50 is the apo structure of the bace

Tanimoto similarity is calculated between this generated smiles to the whole ChEMBL dataset which shows the tanimoto similarity less than 100%.

Training Dataset – 249455 of the smiles molecules that are present in the zinc dataset

Molecule -1

The molecule when docked with the bace shows the binding affinity of -7.5.

Ligand	Binding Affinity	rmsd/ub	rmsd/lb
1w50_output_0_uff_E=105.16	-7.5	0	0
1w50_output_0_uff_E=105.16	-7.4	22.47	20.532
1w50_output_0_uff_E=105.16	-7	5.834	1.708
1w50_output_0_uff_E=105.16	-6.9	5.974	1.661
1w50_output_0_uff_E=105.16	-6.9	6.346	2.329
1w50_output_0_uff_E=105.16	-6.8	6.76	1.814
1w50_output_0_uff_E=105.16	-6.7	6.072	2.939
1w50_output_0_uff_E=105.16	-6.7	23.898	21.914
1w50_output_0_uff_E=105.16	-6.4	34.348	31.661

Figure – 5 Molecule-1 affinity.

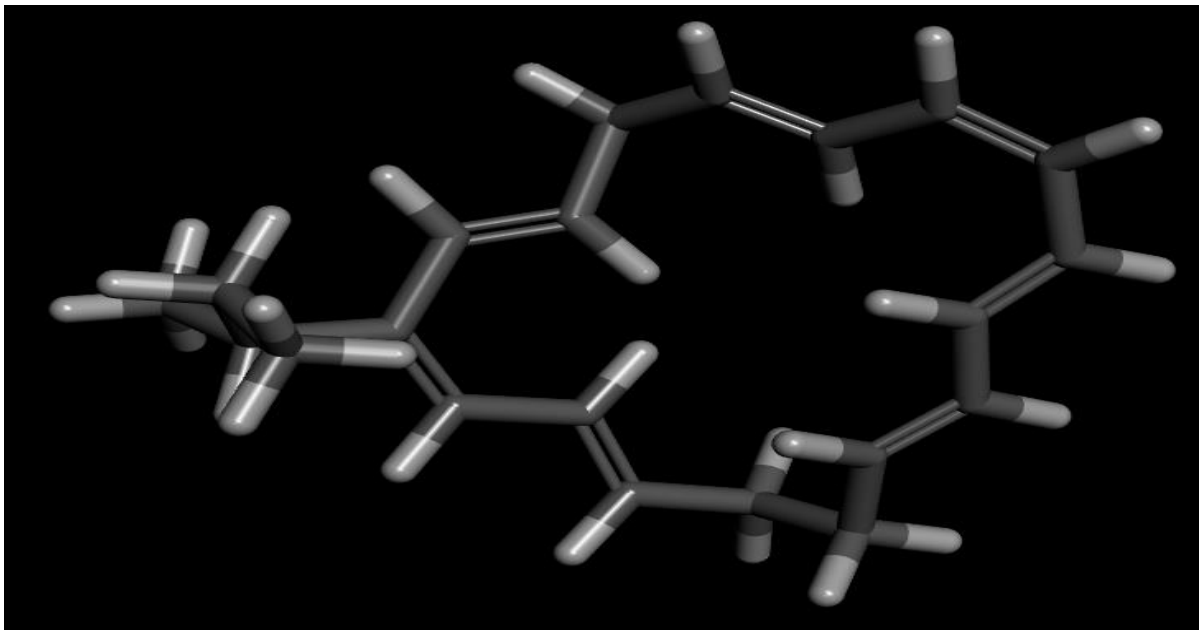

Figure – 6 Molecule-1.

Molecule – 2

The molecule when docked with the bace shows the binding affinity of -7.1.

Ligand	Binding Affinity	rmsd/ub	rmsd/lb
1w50_output_1_uff_E=226.59	-7.1	0	0
1w50_output_1_uff_E=226.59	-7	19.703	17.528
1w50_output_1_uff_E=226.59	-6.8	3.362	1.742
1w50_output_1_uff_E=226.59	-6.7	6.903	5.011
1w50_output_1_uff_E=226.59	-6.7	7.69	3.788
1w50_output_1_uff_E=226.59	-6.6	8.867	5.109
1w50_output_1_uff_E=226.59	-6.6	3.839	2.416
1w50_output_1_uff_E=226.59	-6.6	7.633	2.576
1w50_output_1_uff_E=226.59	-6.5	7.804	5.238

Figure – 7 Molecule-2 affinity.

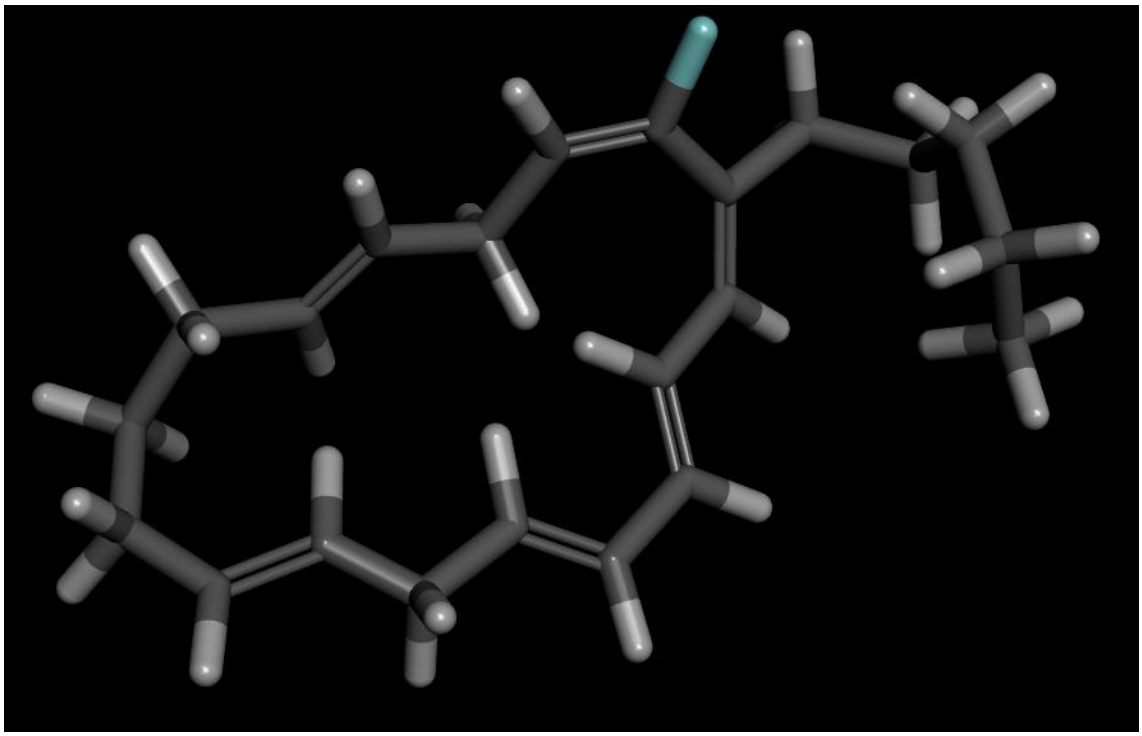

Figure – 8 Molecule-2.

Molecule – 3

The molecule when docked with the bace shows the binding affinity of -4.2.

Ligand	Binding Affinit	rmsd/ub	rmsd/lb
1w50_output_2_uff_E=53.52	-4.2	0	0
1w50_output_2_uff_E=53.52	-4.1	5.894	3.349
1w50_output_2_uff_E=53.52	-4.1	4.15	2.315
1w50_output_2_uff_E=53.52	-4	3.28	1.837
1w50_output_2_uff_E=53.52	-3.9	4.055	2.701
1w50_output_2_uff_E=53.52	-3.9	4.633	2.578
1w50_output_2_uff_E=53.52	-3.9	6.014	3.071
1w50_output_2_uff_E=53.52	-3.9	5.807	3.136
1w50_output_2_uff_E=53.52	-3.9	7.061	4.368

Figure – 9 Molecule -3 affinity.

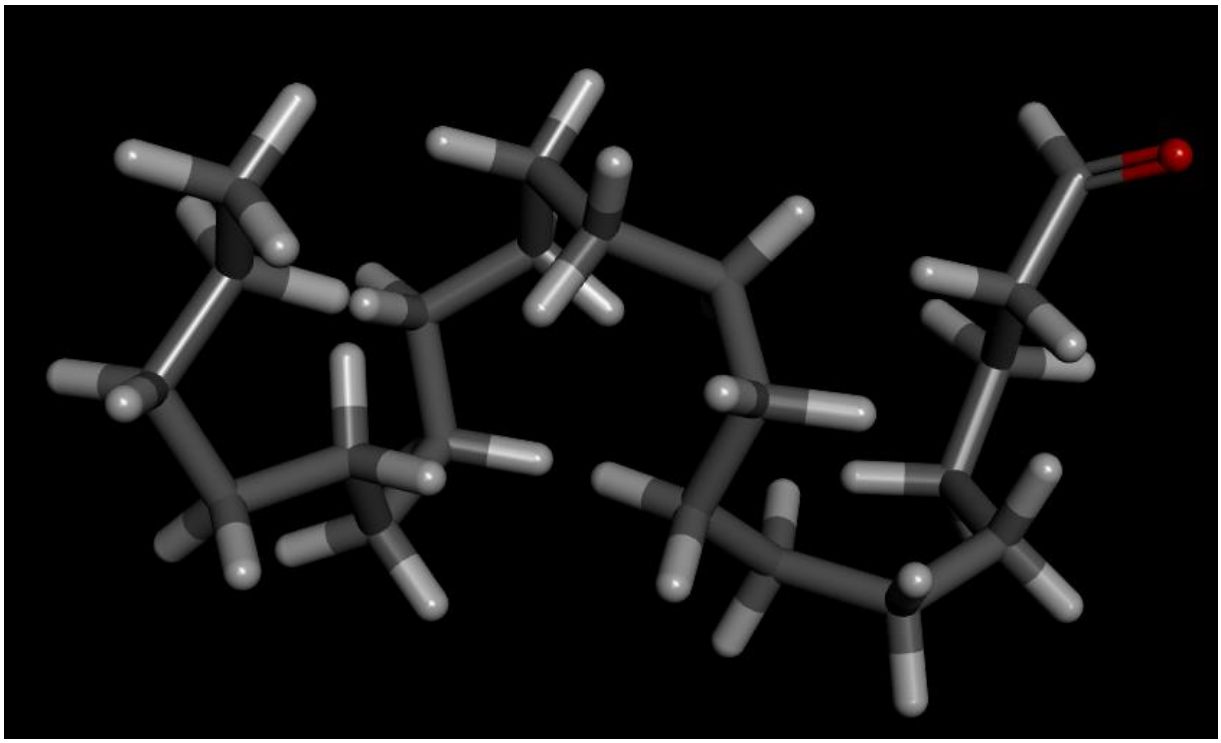

Figure – 10 Molecule-3.

Molecule – 4

The molecule when docked with the bace shows the binding affinity of -4.2.

Ligand	Binding Affinity	rmsd/ub	rmsd/lb
1w50_output_3_uff_E=39.83	-4.2	0	0
1w50_output_3_uff_E=39.83	-4	4.605	2.383
1w50_output_3_uff_E=39.83	-4	3.814	2.004
1w50_output_3_uff_E=39.83	-4	4.102	2.442
1w50_output_3_uff_E=39.83	-3.9	2.684	1.802
1w50_output_3_uff_E=39.83	-3.9	5.502	3.36
1w50_output_3_uff_E=39.83	-3.9	5.108	2.288
1w50_output_3_uff_E=39.83	-3.9	2.555	1.912
1w50_output_3_uff_E=39.83	-3.8	4.623	1.565

Figure – 11 Molecule -4 affinity.

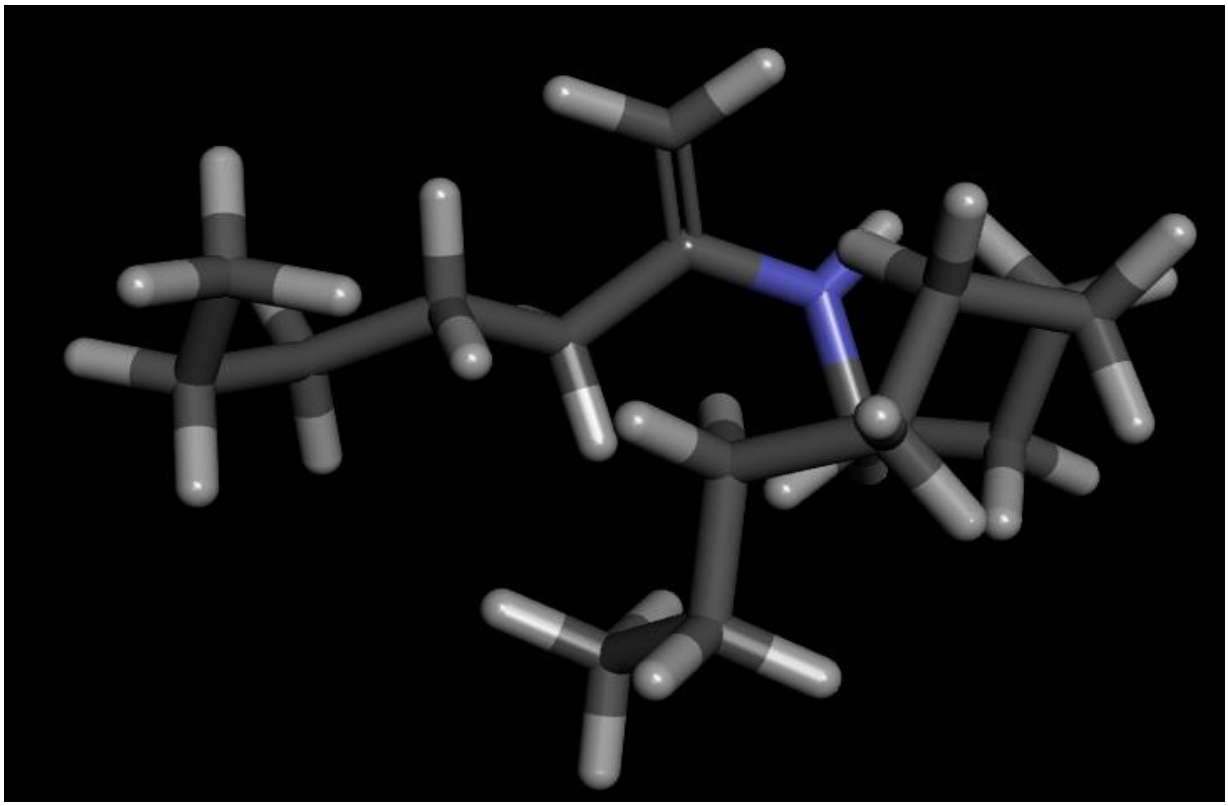

Figure – 12 Molecule-4.

3.2 GPT-2 - The result of the generated molecules from the **ChEMBL dataset** by using the Gpt-2 model finetuned on this dataset **21 molecules** are produced from this dataset out which only **18 molecules** are not showing 100% tanimoto similarity with respect to the chembl dataset,Tanimoto similarity is calculated between the each of the smile that is produced from the chembl dataset to all of the smiles that are present in the chembl dataset one by one [22].

Training dataset – ChEMBL dataset contains 2327927 molecules which are used for training the model.

The mentioned below are the 18 molecules -

1.Cn1cnnc1CN1CCNCC1

2.CC(C)[C@@H](N)C(=O)C(=O)Nc1ccccc1

3.CN(C)CCNC(=O)c1cc(N)no1

4.O=C(c1ncccn1)N1CCN(Cc2ccccc2)CC1

5.O=c1oc(=S)[nH]c2ccc(Br)cc12

6.Cc1cc(N=N)ccc1O

7.C[C@H]1C[SH]=C2N=NC(=CC3=NNC3=O)OCCN21

8.Cc1ccoc1C1=NS(=O)(=O)NC(=O)CN1

9.N=C(N)N1CCC[C@H](Nc2ccc(F)c(F)c2)C1

10.O=C(O)c1cccc(Oc2ccccc2)c1O

11.CCCc1noc(-c2c(Cl)cccc2Cl)n1

12.N[C@H]1CN(C(=O)C(F)(F)c2ccccc2)[C@@H]2[C@H]1C(=O)N2Cc1ccccc1

13.Cn1nc(C#N)cc1N1CCNC1

14.CCCn1nc(C(F)(F)F)c2ccccc21

15.NC[C@@H]1CCc2c(sc3ccccc23)C1

16.CCOC(=O)C=Cc1ccc(Cl)c(Br)c1

17.Cc1c[nH]c(CN2CCN(CC(N)=O)CC2)n1

18.CN1CCC[C@@H](C(=O)O)CC1

This 18 molecule are docked against the bace molecule and the following results that we get are –

The 18 molecules are docked against the bace having the pdb-id 1w50 which are mentioned below with binding affinity,rmsd/ub,rmsd/lb-

Ligand	Binding Affinity	rmsd/ub	rmsd/lb
1w50_molecule_0_uff_E=427.01	-5.1	0	0
1w50_molecule_0_uff_E=427.01	-4.7	30.948	29.256
1w50_molecule_0_uff_E=427.01	-4.6	2.085	1.483
1w50_molecule_0_uff_E=427.01	-4.5	30.92	29.094
1w50_molecule_0_uff_E=427.01	-4.5	2.713	1.627
1w50_molecule_0_uff_E=427.01	-4.4	24.652	23.423
1w50_molecule_0_uff_E=427.01	-4.3	31.424	29.647
1w50_molecule_0_uff_E=427.01	-4.3	28.543	27.63
1w50_molecule_0_uff_E=427.01	-4.2	30.07	29.248
1w50_molecule_1_uff_E=170.45	-5.9	0	0
1w50_molecule_1_uff_E=170.45	-5.9	6.889	2.912
1w50_molecule_1_uff_E=170.45	-5.6	20.334	17.767
1w50_molecule_1_uff_E=170.45	-5.6	3.122	2.481
1w50_molecule_1_uff_E=170.45	-5.4	6.467	3.068
1w50_molecule_1_uff_E=170.45	-5.4	18.61	16.923
1w50_molecule_1_uff_E=170.45	-5.3	18.816	17.063
1w50_molecule_1_uff_E=170.45	-5.1	18.439	15.909
1w50_molecule_1_uff_E=170.45	-5.1	19.265	17.387
1w50_molecule_2_uff_E=317.62	-5.1	0	0
1w50_molecule_2_uff_E=317.62	-5	5.661	3.253
1w50_molecule_2_uff_E=317.62	-4.9	16.455	15.305
1w50_molecule_2_uff_E=317.62	-4.8	6.719	1.878
1w50_molecule_2_uff_E=317.62	-4.8	7.542	2.664
1w50_molecule_2_uff_E=317.62	-4.8	15.536	13.88
1w50_molecule_2_uff_E=317.62	-4.8	7.054	1.928
1w50_molecule_2_uff_E=317.62	-4.7	30.88	30.048
1w50_molecule_2_uff_E=317.62	-4.7	31.4	29.214
1w50_molecule_3_uff_E=287.17	-6.6	0	0
1w50_molecule_3_uff_E=287.17	-6.2	8.038	2.633
1w50_molecule_3_uff_E=287.17	-6.2	7.503	2.587
1w50_molecule_3_uff_E=287.17	-5.9	8.106	2.449
1w50_molecule_3_uff_E=287.17	-5.9	7.706	2.522
1w50_molecule_3_uff_E=287.17	-5.9	8.165	3.498
1w50_molecule_3_uff_E=287.17	-5.8	8.479	3.367
1w50_molecule_3_uff_E=287.17	-5.8	2.352	1.526
1w50_molecule_3_uff_E=287.17	-5.8	7.804	2.914
1w50_molecule_4_uff_E=101.59	-5.6	0	0
1w50_molecule_4_uff_E=101.59	-5.3	21.15	20.611
1w50_molecule_4_uff_E=101.59	-5.1	18.133	17.322

1w50_molecule_4_uff_E=101.59	-4.6	33.268	32.751
1w50_molecule_4_uff_E=101.59	-4.5	31.331	30.261
1w50_molecule_4_uff_E=101.59	-4.5	30.047	29.066
1w50_molecule_4_uff_E=101.59	-4.5	34.258	33.58
1w50_molecule_4_uff_E=101.59	-4.4	33.064	31.975
1w50_molecule_4_uff_E=101.59	-4.4	3.152	2.699
1w50_molecule_5_uff_E=150.62	-5.4	0	0
1w50_molecule_5_uff_E=150.62	-5.2	4.068	2.955
1w50_molecule_5_uff_E=150.62	-5.1	3.88	3.091
1w50_molecule_5_uff_E=150.62	-5.1	4.335	3.227
1w50_molecule_5_uff_E=150.62	-5	3.029	2.343
1w50_molecule_5_uff_E=150.62	-4.8	33.601	32.552
1w50_molecule_5_uff_E=150.62	-4.7	17.037	16.455
1w50_molecule_5_uff_E=150.62	-4.6	15.896	15.05
1w50_molecule_5_uff_E=150.62	-4.6	28.64	28.058
1w50_molecule_6_uff_E=1980.22	-6.5	0	0
1w50_molecule_6_uff_E=1980.22	-6	4.814	3.051
1w50_molecule_6_uff_E=1980.22	-6	6.522	2.957
1w50_molecule_6_uff_E=1980.22	-5.9	30.131	26.801
1w50_molecule_6_uff_E=1980.22	-5.9	28.393	27.621
1w50_molecule_6_uff_E=1980.22	-5.9	3.694	3.216
1w50_molecule_6_uff_E=1980.22	-5.7	7.201	4.453
1w50_molecule_6_uff_E=1980.22	-5.7	23.161	21.55
1w50_molecule_6_uff_E=1980.22	-5.6	29.416	27.963
1w50_molecule_7_uff_E=819.49	-7.5	0	0
1w50_molecule_7_uff_E=819.49	-7.3	2.09	1.518
1w50_molecule_7_uff_E=819.49	-6.4	5.195	2.589
1w50_molecule_7_uff_E=819.49	-6.3	5.544	2.623
1w50_molecule_7_uff_E=819.49	-6.3	30.77	28.125
1w50_molecule_7_uff_E=819.49	-6.2	5.73	2.678
1w50_molecule_7_uff_E=819.49	-5.9	14.515	11.739
1w50_molecule_7_uff_E=819.49	-5.9	14.399	12.051
1w50_molecule_7_uff_E=819.49	-5.9	3.731	2.501
1w50_molecule_8_uff_E=293.37	-6.7	0	0
1w50_molecule_8_uff_E=293.37	-6.6	4.221	2.63
1w50_molecule_8_uff_E=293.37	-6.3	5.601	3.705
1w50_molecule_8_uff_E=293.37	-6.3	4.069	2.932
1w50_molecule_8_uff_E=293.37	-6.3	2.392	1.44
1w50_molecule_8_uff_E=293.37	-6.1	4.196	2.608
1w50_molecule_8_uff_E=293.37	-6.1	5.207	3.815
1w50_molecule_8_uff_E=293.37	-6	5.357	4.312
1w50_molecule_8_uff_E=293.37	-5.9	6.298	4.266
1w50_molecule_9_uff_E=246.38	-5.9	0	0
1w50_molecule_9_uff_E=246.38	-5.9	5.858	4.016
1w50_molecule_9_uff_E=246.38	-5.8	5.704	3.887
1w50_molecule_9_uff_E=246.38	-5.8	2.246	1.51

1w50_molecule_9_uff_E=246.38	-5.8	33.851	32.445
1w50_molecule_9_uff_E=246.38	-5.7	5.743	3.889
1w50_molecule_9_uff_E=246.38	-5.7	24.501	22.325
1w50_molecule_9_uff_E=246.38	-5.5	6.185	1.874
1w50_molecule_9_uff_E=246.38	-5.5	25.009	23.051
1w50_molecule_10_uff_E=354.50	-6.4	0	0
1w50_molecule_10_uff_E=354.50	-6.3	3.288	2.126
1w50_molecule_10_uff_E=354.50	-5.9	2.083	1.706
1w50_molecule_10_uff_E=354.50	-5.6	17.55	15.42
1w50_molecule_10_uff_E=354.50	-5.6	17.425	15.414
1w50_molecule_10_uff_E=354.50	-5.5	18.991	16.146
1w50_molecule_10_uff_E=354.50	-5.4	18.648	16.2
1w50_molecule_10_uff_E=354.50	-5.3	15.936	14.446
1w50_molecule_10_uff_E=354.50	-5.3	17.81	16.521
1w50_molecule_11_uff_E=1083.00	-7.6	0	0
1w50_molecule_11_uff_E=1083.00	-7.3	4.046	2.711
1w50_molecule_11_uff_E=1083.00	-7.1	7.253	4.094
1w50_molecule_11_uff_E=1083.00	-7.1	2.552	1.796
1w50_molecule_11_uff_E=1083.00	-7.1	5.243	2.749
1w50_molecule_11_uff_E=1083.00	-7.1	2.113	1.609
1w50_molecule_11_uff_E=1083.00	-7	5.954	2.66
1w50_molecule_11_uff_E=1083.00	-6.9	5.99	3.132
1w50_molecule_11_uff_E=1083.00	-6.8	6.051	2.957
1w50_molecule_12_uff_E=479.15	-5.4	0	0
1w50_molecule_12_uff_E=479.15	-5.1	1.891	1.581
1w50_molecule_12_uff_E=479.15	-5.1	30.699	29.561
1w50_molecule_12_uff_E=479.15	-4.9	30.165	27.498
1w50_molecule_12_uff_E=479.15	-4.9	30.586	29.23
1w50_molecule_12_uff_E=479.15	-4.6	33.166	31.498
1w50_molecule_12_uff_E=479.15	-4.5	52.658	51.638
1w50_molecule_12_uff_E=479.15	-4.4	44.131	42.043
1w50_molecule_12_uff_E=479.15	-4.4	37.105	35.981
1w50_molecule_13_uff_E=336.32	-6.7	0	0
1w50_molecule_13_uff_E=336.32	-6.4	4.756	2.659
1w50_molecule_13_uff_E=336.32	-5.9	3.611	2.254
1w50_molecule_13_uff_E=336.32	-5.9	6.142	3.669
1w50_molecule_13_uff_E=336.32	-5.6	13.531	11.145
1w50_molecule_13_uff_E=336.32	-5.5	16.119	15.088
1w50_molecule_13_uff_E=336.32	-5.4	17.553	16.473
1w50_molecule_13_uff_E=336.32	-5.4	13.068	11.107
1w50_molecule_13_uff_E=336.32	-5.4	13.772	11.819
1w50_molecule_14_uff_E=381.00	-5.9	0	0
1w50_molecule_14_uff_E=381.00	-5.6	3.461	2.57
1w50_molecule_14_uff_E=381.00	-5.5	5.56	2.581
1w50_molecule_14_uff_E=381.00	-5.5	5.624	4.283
1w50_molecule_14_uff_E=381.00	-5.4	5.843	4.153

1w50_molecule_14_uff_E=381.00	-5.4	2.547	1.624
1w50_molecule_14_uff_E=381.00	-5.3	5.668	2.765
1w50_molecule_14_uff_E=381.00	-5.3	33.897	33.065
1w50_molecule_14_uff_E=381.00	-5.3	5.91	2.793
1w50_molecule_15_uff_E=110.10	-5.5	0	0
1w50_molecule_15_uff_E=110.10	-5.5	25.953	24.427
1w50_molecule_15_uff_E=110.10	-5.5	20.49	19.227
1w50_molecule_15_uff_E=110.10	-5.4	23.83	22.776
1w50_molecule_15_uff_E=110.10	-5.3	25.565	24.046
1w50_molecule_15_uff_E=110.10	-5.2	26.558	25.005
1w50_molecule_15_uff_E=110.10	-5	26.736	25.186
1w50_molecule_15_uff_E=110.10	-5	21.152	19.784
1w50_molecule_15_uff_E=110.10	-4.9	36.071	35.225
1w50_molecule_16_uff_E=353.49	-5.4	0	0
1w50_molecule_16_uff_E=353.49	-5.4	21.819	20.812
1w50_molecule_16_uff_E=353.49	-5.3	23.481	21.358
1w50_molecule_16_uff_E=353.49	-5.2	30.491	29.058
1w50_molecule_16_uff_E=353.49	-5.1	24.545	21.795
1w50_molecule_16_uff_E=353.49	-5	13.677	11.125
1w50_molecule_16_uff_E=353.49	-5	3.41	2.443
1w50_molecule_16_uff_E=353.49	-5	22.79	21.568
1w50_molecule_16_uff_E=353.49	-4.9	24.038	21.149
1w50_molecule_17_uff_E=223.59	-4.7	0	0
1w50_molecule_17_uff_E=223.59	-4.5	3.571	2.587
1w50_molecule_17_uff_E=223.59	-4.4	2.176	1.111
1w50_molecule_17_uff_E=223.59	-4.3	22.035	21.023
1w50_molecule_17_uff_E=223.59	-4.3	33.715	32.533
1w50_molecule_17_uff_E=223.59	-4.2	12.011	10.698
1w50_molecule_17_uff_E=223.59	-4.2	32.885	31.627
1w50_molecule_17_uff_E=223.59	-4	31.241	29.412
1w50_molecule_17_uff_E=223.59	-3.9	33.362	32.274

Figure 13 – Showing the docking of the 18 molecules (ChEMBL dataset) with the respect to bace(1W50).

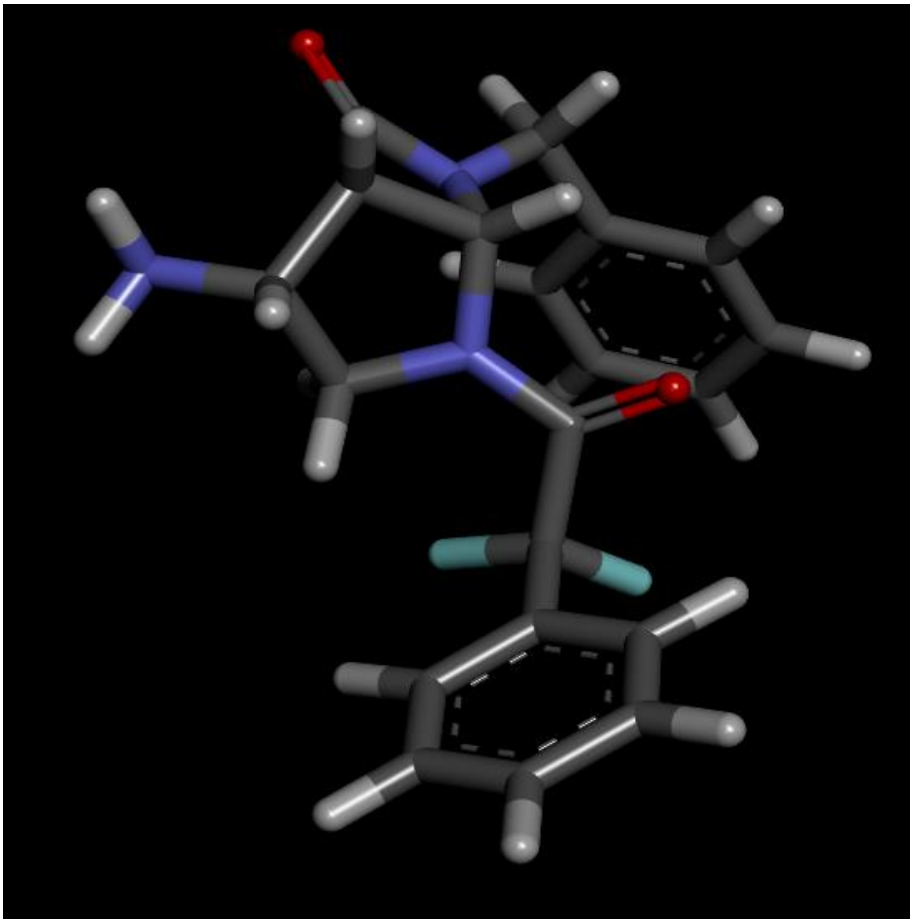

Figure – 14 Molecule_11.

The above mentioned molecule_11 shows the binding affinity of -7.6.

When the Gpt-2 model is trained on the **inhibitor dataset 52 molecules were generated from the model**,out of this each smile of the molecule is taken and the tanimoto similarity is calculated between the each smile that generated by the gpt-2 model with that to all of the smiles that are present in the chembl dataset [23].**Out of this 32 molecules are selected from the 52 molecules because they are not showing the 100% tanimoto similarity or less than 100% tanimoto similarity [38].**

Dataset – Inhibitor dataset contains 7234 molecules.

The 32 molecules that do not show 100% tanimoto similarity when compared to the chembl dataset are mentioned below –

1.CCCN(CCC)C(=O)C1=CC(C(=O)N[C@@H](Cc2cc(F)cc(F)c2)[C@@H](O)[C@H](Cc2ccccc2)N2CCCCC2=O)=C1CC#CCOc1cnc(C(=O)Nc2ccc(F)c([C@]3(C(F)F)N=C(N)O[C@@H]4C[C@@H]43)c2)c(Cl)c1

2.CO[C@H]1CC[C@]2(CC1)Cc1ccc(-c3cccc(Cl)c3)cc1[C@@]21N=C(N)N(CC(C)C)C1=O

3.CC(C)c1cnc(C(=O)Nc2ccc(F)c([C@@]3(C(F)F)N=C(N)O[C@@H]4C[C@@H]43)c2)cn1

4.CCOC1=CC(C(=O)N[C@@H](Cc2ccccc2)[C@@H](O)CN[C@H](Cc2ccccc2)C[C@@H]2CCCC2=O)=C1

5.C#Cc1cnc(C(=O)Nc2ccc(F)c([C@@]3(C(F)F)N=C(N)O[C@@H]4C[C@@H]43)c2)c(Cl)c1

6.CO[C@H]1CC[C@]2(CC1)Cc1ccc(OF)cc1[C@@]21N=C(N)N(C)C1=O

7.C[C@@]1(c2cc(Nc3ncc(Br)cn3)ccc2F)CCSC(N)=N1

8.CO[C@H]1CC[C@]2(CC1)Cc1ccc(OCC#N)cc1[C@@]21N=C(N)N(C)C1=O

9.COc1cccc(CNC[C@@H](O)[C@@H](Cc2ccccc2)NC(=O)[C@H](CC(C)C)NC(=O)c2cccc(C(C)(C)C)c2)c1

10.COc1cnc(C(=O)Nc2ccc(F)c([C@@]3(C)C[C@@H](C(F)(F)F)OC(N)=N3)c2)cn1

11.CC(=O)[C@@H](N)C(=O)N[C@H](C(=O)N[C@H](C(=O)N[C@@H](CCC(=O)O)C(C)C)C(=O)N[C@@H](CC(C)C)C(C)C)C(C)C

12.CC(=O)c1cc2cc(c1)C(=O)N[C@@H](Cc1ccccc1)[C@H](O)CNC[C@H](C[C@H](Cc1ccccc1)S(C)(=O)=O)NC2=O

13.CC1=N[C@@H](c2cccc(NC(=O)c3ccc(OC(F)F)cc3)c2)N=C1N

14.C[C@H](NC(=O)C[C@@H](O)C(=O)NC(=O)N[C@@H](Cc1cc(F)cc(F)c1)[C@@H](O)[C@H](Cc1ccccc1)NC(=O)c1ccccc1CCCCN)C(=O)NC1CCCCC1

15.N#Cc1ccc(C(=O)Nc2ccc(F)c([C@@]3(C(F)F)N=C(N)O[C@@H]4C[C@@H]43)c2)cn1

16.COc1ccc(C(=O)Nc2cccc([C@@]3(C)COCC(N)=N3)c2)cn1

17.CC#Cc1cncc(-c2ccc3c(c2)[C@@]2(COC(N)=N2)C2(COC2)N=C(C)O3)c1

18.Cc1cnc(C(=O)Nc2ccc(F)c([C@@]3(CF)C[C@@H](C(F)(F)F)C(N)=N3)c2)cn1

19.COc1cccc(-c2cccc([C@]3(C)N=C3N)c2)c1

20.C[C@@]1(c2nc(NC(=O)c3ccc(F)cn3)ccc2F)CO[C@H](C(F)(F)F)C(N)=N1

21.CO[C@H]1CC[C@]2(CC1)Cc1ccc(OCC(F)(F)F)cc1[C@]21N=C(N)N(C)C1=O

22.N#Cc1ccc(C(=O)Nc2ccc(F)c([C@@]3(C(F)F)N=C(N)O[C@@H]4C[C@@H]43)c2)cn1

23.C[C@@]1(c2cc(NC(=O)c3ccc(C#N)cn3)ccc2F)C[C@@H](C(F)(F)F)SC(N)=N1

24.COc1nc(N2C[C@H]3C(=O)N(C)C(N)=N[C@@]3(c3ccc(F)cc3F)C2)nc(C)c1F

25.COc1cnc(C(=O)Nc2ccc(F)c([C@]3(C(F)F)N=C(N)O[C@@H]4C[C@@H]43)n2)cn1

26.COc1ccc2c(c1)[C@@]1(COC(N)=N1)c1cc(-c3cncc(Cl)c3)ccc1O2

27.CN1C(=O)[C@@](C)(c2nc(NC(=O)c3ncc(F)cc3F)ccc2F)N=C1N

28.C[C@@]1(c2cc(Nc3ccccn3)ccc2F)COCC(N)=N1

29.C[C@@]1(c2nc(NC(=O)c3ccc(F)cn3)ccc2F)CO[C@H](C(F)(F)F)C(N)=N1

30.N#Cc1cnc(C(=O)Nc2ccc(F)c([C@]3(C(F)F)COCC(N)=N3)c2)c(Cl)c1

31.Cc1cncc(-c2ccc3c(c2)[C@]2(CCSC(N)=N2)[C@]2(COC2)[C@]2(CCC2)O3)c1

32.CCCN(CCC)C(=O)C1=CC(C(=O)N[C@@H](Cc2cc(F)cc(F)c2)[C@@H](O)[C@H](Cc2ccccc2)N2CCCCC2=O)=C1

The 32 molecules are docked against the bace having the pdb-id 1w50 which are mentioned below with binding affinity,rmsd/ub,rmsd/lb-

Ligand	Binding Affinity	rmsd/ub	rmsd/lb
1w50_molecule_0_uff_E=366.74	-7.7	0	0
1w50_molecule_0_uff_E=366.74	-6.8	7.423	4.16
1w50_molecule_0_uff_E=366.74	-6.7	3.235	2.327
1w50_molecule_0_uff_E=366.74	-6.6	7.503	4.216
1w50_molecule_0_uff_E=366.74	-6.5	5.982	4.579
1w50_molecule_0_uff_E=366.74	-6.5	7.974	3.875
1w50_molecule_0_uff_E=366.74	-6.4	5.274	3.856
1w50_molecule_0_uff_E=366.74	-6.4	3.58	2.844
1w50_molecule_0_uff_E=366.74	-6	24.752	23.559
1w50_molecule_1_uff_E=810.48	-7.6	0	0
1w50_molecule_1_uff_E=810.48	-7.3	7.633	3.605
1w50_molecule_1_uff_E=810.48	-7.2	6.118	3.834
1w50_molecule_1_uff_E=810.48	-7.2	8.651	4.466
1w50_molecule_1_uff_E=810.48	-7.2	8.882	5.644
1w50_molecule_1_uff_E=810.48	-7.1	7.546	3.053
1w50_molecule_1_uff_E=810.48	-7.1	25.986	21.974
1w50_molecule_1_uff_E=810.48	-7.1	7.29	3.386
1w50_molecule_1_uff_E=810.48	-7	25.98	22.276

1w50_molecule_2_uff_E=400.91	-7.7	0	0
1w50_molecule_2_uff_E=400.91	-7.6	20.931	18.11
1w50_molecule_2_uff_E=400.91	-7.6	4.675	2.935
1w50_molecule_2_uff_E=400.91	-7.5	4.118	2.082
1w50_molecule_2_uff_E=400.91	-7.4	4.891	3.653
1w50_molecule_2_uff_E=400.91	-7.4	9.213	4.695
1w50_molecule_2_uff_E=400.91	-7.3	4.511	2.71
1w50_molecule_2_uff_E=400.91	-7.3	4.382	2.004
1w50_molecule_2_uff_E=400.91	-7.2	33.444	30.675
1w50_molecule_3_uff_E=403.44	-8	0	0
1w50_molecule_3_uff_E=403.44	-7.6	6.304	3.588
1w50_molecule_3_uff_E=403.44	-7.4	8.006	4.762
1w50_molecule_3_uff_E=403.44	-7.3	6.254	3.85
1w50_molecule_3_uff_E=403.44	-7.3	6.453	4.474
1w50_molecule_3_uff_E=403.44	-7.1	4.767	3.082
1w50_molecule_3_uff_E=403.44	-7.1	6.74	4.108
1w50_molecule_3_uff_E=403.44	-6.9	7.4	4.286
1w50_molecule_3_uff_E=403.44	-6.8	8.021	4.431
1w50_molecule_4_uff_E=298.68	-8.2	0	0
1w50_molecule_4_uff_E=298.68	-7.9	22.098	18.958
1w50_molecule_4_uff_E=298.68	-7.8	21.261	16.952
1w50_molecule_4_uff_E=298.68	-7.7	20.57	17.483
1w50_molecule_4_uff_E=298.68	-7.4	3.028	2.384
1w50_molecule_4_uff_E=298.68	-7.4	2.227	1.794
1w50_molecule_4_uff_E=298.68	-7.4	33.177	29.418
1w50_molecule_4_uff_E=298.68	-7.2	21.36	17.629
1w50_molecule_4_uff_E=298.68	-7.2	21.173	18.481
1w50_molecule_5_uff_E=341.79	-8.4	0	0
1w50_molecule_5_uff_E=341.79	-7.7	20.785	18.191
1w50_molecule_5_uff_E=341.79	-7.7	20.467	18.007
1w50_molecule_5_uff_E=341.79	-7.6	33.103	30.732
1w50_molecule_5_uff_E=341.79	-7.5	20.965	19.197
1w50_molecule_5_uff_E=341.79	-7.5	21.56	17.094
1w50_molecule_5_uff_E=341.79	-7.4	20.357	17.067
1w50_molecule_5_uff_E=341.79	-7.4	4.739	3.229
1w50_molecule_5_uff_E=341.79	-7.3	20.827	17.917
1w50_molecule_6_uff_E=390.09	-7.7	0	0
1w50_molecule_6_uff_E=390.09	-7.7	7.174	3.624
1w50_molecule_6_uff_E=390.09	-7.4	2.212	1.691
1w50_molecule_6_uff_E=390.09	-7.3	6.259	3.807
1w50_molecule_6_uff_E=390.09	-7.1	2.098	1.698
1w50_molecule_6_uff_E=390.09	-7.1	4.402	2.647
1w50_molecule_6_uff_E=390.09	-7	2.895	2.033
1w50_molecule_6_uff_E=390.09	-7	3.734	2.532
1w50_molecule_6_uff_E=390.09	-7	5.788	3.266
1w50_molecule_7_uff_E=1396.87	-8.1	0	0

1w50_molecule_7_uff_E=1396.87	-8	7.436	3.99
1w50_molecule_7_uff_E=1396.87	-7.9	6.532	3.998
1w50_molecule_7_uff_E=1396.87	-7.9	2.259	1.9
1w50_molecule_7_uff_E=1396.87	-7.8	3.842	2.321
1w50_molecule_7_uff_E=1396.87	-7.8	7.206	3.208
1w50_molecule_7_uff_E=1396.87	-7.6	25.277	22.609
1w50_molecule_7_uff_E=1396.87	-7.6	7.109	3.077
1w50_molecule_7_uff_E=1396.87	-7.5	7.705	3.814
1w50_molecule_8_uff_E=375.13	-6.5	0	0
1w50_molecule_8_uff_E=375.13	-6.3	2.956	2.021
1w50_molecule_8_uff_E=375.13	-6.3	3.217	1.983
1w50_molecule_8_uff_E=375.13	-6.3	4.005	2.681
1w50_molecule_8_uff_E=375.13	-6.2	4.894	2.614
1w50_molecule_8_uff_E=375.13	-6.2	8.47	2.548
1w50_molecule_8_uff_E=375.13	-6.1	4.296	2.864
1w50_molecule_8_uff_E=375.13	-6.1	6.964	2.182
1w50_molecule_8_uff_E=375.13	-6.1	4.118	2.842
1w50_molecule_9_uff_E=299.84	-7.3	0	0
1w50_molecule_9_uff_E=299.84	-7	20.9	19.743
1w50_molecule_9_uff_E=299.84	-7	20.704	19.526
1w50_molecule_9_uff_E=299.84	-6.8	2.791	2.303
1w50_molecule_9_uff_E=299.84	-6.8	2.556	1.884
1w50_molecule_9_uff_E=299.84	-6.7	6.483	3.447
1w50_molecule_9_uff_E=299.84	-6.7	7.245	3.508
1w50_molecule_9_uff_E=299.84	-6.6	6.252	2.16
1w50_molecule_9_uff_E=299.84	-6.6	6.054	2.628
1w50_molecule_10_uff_E=397.52	-7.2	0	0
1w50_molecule_10_uff_E=397.52	-7.1	22.36	20.481
1w50_molecule_10_uff_E=397.52	-7.1	23.16	21.262
1w50_molecule_10_uff_E=397.52	-7.1	34.804	32.815
1w50_molecule_10_uff_E=397.52	-6.7	22.846	20.426
1w50_molecule_10_uff_E=397.52	-6.7	21.926	19.786
1w50_molecule_10_uff_E=397.52	-6.6	21.268	19.842
1w50_molecule_10_uff_E=397.52	-6.5	38.365	36.942
1w50_molecule_10_uff_E=397.52	-6.5	33.586	31.72
1w50_molecule_11_uff_E=954.80	-7.8	0	0
1w50_molecule_11_uff_E=954.80	-7.8	4.417	2.885
1w50_molecule_11_uff_E=954.80	-7.6	8.251	4.17
1w50_molecule_11_uff_E=954.80	-7.5	8.056	4.606
1w50_molecule_11_uff_E=954.80	-7.4	8.393	4.157
1w50_molecule_11_uff_E=954.80	-7.4	7.151	3.792
1w50_molecule_11_uff_E=954.80	-7.2	5.126	2.979
1w50_molecule_11_uff_E=954.80	-7.1	4.797	2.905
1w50_molecule_11_uff_E=954.80	-7.1	33.06	30.76
1w50_molecule_12_uff_E=553.13	-9.3	0	0
1w50_molecule_12_uff_E=553.13	-8.5	3.478	2.331

1w50_molecule_12_uff_E=553.13	-8.2	19.67	16.676
1w50_molecule_12_uff_E=553.13	-8.1	5.333	3.519
1w50_molecule_12_uff_E=553.13	-7.8	19.445	16.631
1w50_molecule_12_uff_E=553.13	-7.8	32.938	30.373
1w50_molecule_12_uff_E=553.13	-7.8	22.592	18.051
1w50_molecule_12_uff_E=553.13	-7.6	20.383	18.239
1w50_molecule_12_uff_E=553.13	-7.6	19.009	15.131
1w50_molecule_13_uff_E=1379.66	-7.5	0	0
1w50_molecule_13_uff_E=1379.66	-7	2.844	1.72
1w50_molecule_13_uff_E=1379.66	-6.9	7.958	2.953
1w50_molecule_13_uff_E=1379.66	-6.9	19.991	16.574
1w50_molecule_13_uff_E=1379.66	-6.7	6.874	2.419
1w50_molecule_13_uff_E=1379.66	-6.7	6.372	2.447
1w50_molecule_13_uff_E=1379.66	-6.6	8.38	3.18
1w50_molecule_13_uff_E=1379.66	-6.5	7.905	2.809
1w50_molecule_13_uff_E=1379.66	-6.3	2.702	1.95
1w50_molecule_14_uff_E=576.82	-7.3	0	0
1w50_molecule_14_uff_E=576.82	-7	8.151	2.322
1w50_molecule_14_uff_E=576.82	-6.8	5.744	2.234
1w50_molecule_14_uff_E=576.82	-6.7	7.972	4.39
1w50_molecule_14_uff_E=576.82	-6.7	5.508	2.427
1w50_molecule_14_uff_E=576.82	-6.6	7.695	2.971
1w50_molecule_14_uff_E=576.82	-6.6	8.187	3.321
1w50_molecule_14_uff_E=576.82	-6.6	6.564	2.868
1w50_molecule_14_uff_E=576.82	-6.6	5.755	2.118
1w50_molecule_15_uff_E=1584.83	-8	0	0
1w50_molecule_15_uff_E=1584.83	-7.9	2.533	1.765
1w50_molecule_15_uff_E=1584.83	-7.5	23.393	21.132
1w50_molecule_15_uff_E=1584.83	-7	9.151	4.036
1w50_molecule_15_uff_E=1584.83	-6.9	5.006	3.471
1w50_molecule_15_uff_E=1584.83	-6.8	23.979	21.696
1w50_molecule_15_uff_E=1584.83	-6.8	24.209	21.694
1w50_molecule_15_uff_E=1584.83	-6.6	24.942	22.597
1w50_molecule_15_uff_E=1584.83	-6.6	4.116	3.436
1w50_molecule_16_uff_E=754.38	-7.1	0	0
1w50_molecule_16_uff_E=754.38	-7	6.754	3.963
1w50_molecule_16_uff_E=754.38	-6.8	7.288	4.019
1w50_molecule_16_uff_E=754.38	-6.5	6.319	2.663
1w50_molecule_16_uff_E=754.38	-6.5	25.948	23.335
1w50_molecule_16_uff_E=754.38	-6.5	6.837	3.407
1w50_molecule_16_uff_E=754.38	-6.5	5.506	3.543
1w50_molecule_16_uff_E=754.38	-6.5	6.448	2.628
1w50_molecule_16_uff_E=754.38	-6.5	6.961	3.299
1w50_molecule_17_uff_E=1568.26	-7.6	0	0
1w50_molecule_17_uff_E=1568.26	-7.2	5.252	3.954
1w50_molecule_17_uff_E=1568.26	-7.1	5.423	3.693

1w50_molecule_17_uff_E=1568.26	-7	18.437	15.606
1w50_molecule_17_uff_E=1568.26	-7	8.917	4.117
1w50_molecule_17_uff_E=1568.26	-7	3.484	2.642
1w50_molecule_17_uff_E=1568.26	-6.8	9.925	5.061
1w50_molecule_17_uff_E=1568.26	-6.8	2.565	2.229
1w50_molecule_17_uff_E=1568.26	-6.8	3.643	2.675
1w50_molecule_18_uff_E=1585.74	-7.8	0	0
1w50_molecule_18_uff_E=1585.74	-7.6	2.487	1.762
1w50_molecule_18_uff_E=1585.74	-7.4	2.443	1.635
1w50_molecule_18_uff_E=1585.74	-7.4	8.378	3.503
1w50_molecule_18_uff_E=1585.74	-7.2	9.748	3.841
1w50_molecule_18_uff_E=1585.74	-6.9	33.262	30.09
1w50_molecule_18_uff_E=1585.74	-6.8	6.629	4.249
1w50_molecule_18_uff_E=1585.74	-6.8	6.354	4.608
1w50_molecule_18_uff_E=1585.74	-6.7	2.798	1.87
1w50_molecule_19_uff_E=576.14	-7.6	0	0
1w50_molecule_19_uff_E=576.14	-7.4	8.19	5.022
1w50_molecule_19_uff_E=576.14	-7.4	32.853	29.508
1w50_molecule_19_uff_E=576.14	-7.3	7.299	4.203
1w50_molecule_19_uff_E=576.14	-7.2	5.256	3.516
1w50_molecule_19_uff_E=576.14	-7.2	28.01	25.671
1w50_molecule_19_uff_E=576.14	-7.2	7.001	5.449
1w50_molecule_19_uff_E=576.14	-7.1	6.665	3.854
1w50_molecule_19_uff_E=576.14	-6.8	32.724	30.99
1w50_molecule_20_uff_E=1167.93	-8.6	0	0
1w50_molecule_20_uff_E=1167.93	-8.1	4.216	2.738
1w50_molecule_20_uff_E=1167.93	-7.9	7.931	3.934
1w50_molecule_20_uff_E=1167.93	-7.9	6.135	3.141
1w50_molecule_20_uff_E=1167.93	-7.8	3.55	2.281
1w50_molecule_20_uff_E=1167.93	-7.7	3.03	1.745
1w50_molecule_20_uff_E=1167.93	-7.6	7.364	2.647
1w50_molecule_20_uff_E=1167.93	-7.5	7.313	3.09
1w50_molecule_20_uff_E=1167.93	-7.4	3.13	2.087
1w50_molecule_21_uff_E=2952.74	-7.3	0	0
1w50_molecule_21_uff_E=2952.74	-6.8	18.846	17.683
1w50_molecule_21_uff_E=2952.74	-6.6	19.182	17.345
1w50_molecule_21_uff_E=2952.74	-6.5	16.669	13.713
1w50_molecule_21_uff_E=2952.74	-6.5	19.227	17.27
1w50_molecule_21_uff_E=2952.74	-6.4	7.641	4.092
1w50_molecule_21_uff_E=2952.74	-6.4	18.012	14.328
1w50_molecule_21_uff_E=2952.74	-6.3	19.07	17.171
1w50_molecule_21_uff_E=2952.74	-6.3	18.86	17.36
1w50_molecule_22_uff_E=836.19	-7.4	0	0
1w50_molecule_22_uff_E=836.19	-7.4	6.964	3.14
1w50_molecule_22_uff_E=836.19	-7.2	2.142	1.528
1w50_molecule_22_uff_E=836.19	-7.2	8.096	4.559

1w50_molecule_22_uff_E=836.19	-7.1	2.874	1.344
1w50_molecule_22_uff_E=836.19	-7.1	8.513	4.923
1w50_molecule_22_uff_E=836.19	-7	3.805	2.196
1w50_molecule_22_uff_E=836.19	-6.8	6.218	3.028
1w50_molecule_22_uff_E=836.19	-6.8	9.823	4.034
1w50_molecule_23_uff_E=341.13	-7.2	0	0
1w50_molecule_23_uff_E=341.13	-7.1	22.138	20.094
1w50_molecule_23_uff_E=341.13	-7	31.938	29.066
1w50_molecule_23_uff_E=341.13	-6.9	21.822	18.537
1w50_molecule_23_uff_E=341.13	-6.9	20.842	17.545
1w50_molecule_23_uff_E=341.13	-6.9	21.355	17.714
1w50_molecule_23_uff_E=341.13	-6.9	21.404	17.088
1w50_molecule_23_uff_E=341.13	-6.7	33.191	30.917
1w50_molecule_23_uff_E=341.13	-6.6	23.772	17.633
1w50_molecule_24_uff_E=379.20	-7.6	0	0
1w50_molecule_24_uff_E=379.20	-7.5	32.754	30.083
1w50_molecule_24_uff_E=379.20	-7.5	7.155	3.556
1w50_molecule_24_uff_E=379.20	-7.4	32.763	30.293
1w50_molecule_24_uff_E=379.20	-7.3	6.781	4.091
1w50_molecule_24_uff_E=379.20	-7.3	7.563	2.888
1w50_molecule_24_uff_E=379.20	-7.3	7.532	2.901
1w50_molecule_24_uff_E=379.20	-7.2	24.089	21.185
1w50_molecule_24_uff_E=379.20	-7.1	3.76	2.064
1w50_molecule_25_uff_E=1642.22	-7.5	0	0
1w50_molecule_25_uff_E=1642.22	-7.3	23.082	19.131
1w50_molecule_25_uff_E=1642.22	-6.8	8.609	4.838
1w50_molecule_25_uff_E=1642.22	-6.7	23.291	20.585
1w50_molecule_25_uff_E=1642.22	-6.6	26.441	21.784
1w50_molecule_25_uff_E=1642.22	-6.5	8.728	5.086
1w50_molecule_25_uff_E=1642.22	-6.3	24.344	22.382
1w50_molecule_25_uff_E=1642.22	-6.2	6.472	5.08
1w50_molecule_25_uff_E=1642.22	-6.1	23.511	19.578
1w50_molecule_26_uff_E=292.48	-7.7	0	0
1w50_molecule_26_uff_E=292.48	-7.2	3.036	2.287
1w50_molecule_26_uff_E=292.48	-7.2	6.535	3.625
1w50_molecule_26_uff_E=292.48	-7.1	6.98	4.587
1w50_molecule_26_uff_E=292.48	-7.1	23.954	22.412
1w50_molecule_26_uff_E=292.48	-6.9	5.122	4.485
1w50_molecule_26_uff_E=292.48	-6.9	6.519	4.38
1w50_molecule_26_uff_E=292.48	-6.8	2.689	2.09
1w50_molecule_26_uff_E=292.48	-6.8	6.577	3.718
1w50_molecule_27_uff_E=443.29	-7.8	0	0
1w50_molecule_27_uff_E=443.29	-7.8	7.067	4.433
1w50_molecule_27_uff_E=443.29	-7.5	25.934	23.648
1w50_molecule_27_uff_E=443.29	-7.4	7.071	4.906
1w50_molecule_27_uff_E=443.29	-7.3	7.613	3.393

1w50_molecule_27_uff_E=443.29	-7.1	8.033	5.155
1w50_molecule_27_uff_E=443.29	-7	29.284	27.28
1w50_molecule_27_uff_E=443.29	-6.9	8.085	5.287
1w50_molecule_27_uff_E=443.29	-6.9	9.473	5.301
1w50_molecule_28_uff_E=699.68	-7.3	0	0
1w50_molecule_28_uff_E=699.68	-7	5.322	3.364
1w50_molecule_28_uff_E=699.68	-6.7	4.587	2.673
1w50_molecule_28_uff_E=699.68	-6.7	4.995	3.571
1w50_molecule_28_uff_E=699.68	-6.7	5.237	4.164
1w50_molecule_28_uff_E=699.68	-6.7	23.947	22.246
1w50_molecule_28_uff_E=699.68	-6.7	5.096	3.547
1w50_molecule_28_uff_E=699.68	-6.6	5.835	3.285
1w50_molecule_28_uff_E=699.68	-6.3	5.617	3.444
1w50_molecule_29_uff_E=1598.09	-7.9	0	0
1w50_molecule_29_uff_E=1598.09	-7.7	1.648	1.07
1w50_molecule_29_uff_E=1598.09	-7.6	5.839	3.14
1w50_molecule_29_uff_E=1598.09	-7.5	8.66	4.023
1w50_molecule_29_uff_E=1598.09	-7.4	7.727	4.727
1w50_molecule_29_uff_E=1598.09	-7.4	5.338	3.502
1w50_molecule_29_uff_E=1598.09	-7.3	4.555	3.265
1w50_molecule_29_uff_E=1598.09	-7.3	4.642	3.617
1w50_molecule_29_uff_E=1598.09	-7.2	2.574	1.353
1w50_molecule_30_uff_E=1643.99	-7.5	0	0
1w50_molecule_30_uff_E=1643.99	-7.4	17.362	14.404
1w50_molecule_30_uff_E=1643.99	-7.2	38.923	34.569
1w50_molecule_30_uff_E=1643.99	-7	34.804	32.661
1w50_molecule_30_uff_E=1643.99	-6.9	37.9	33.773
1w50_molecule_30_uff_E=1643.99	-6.9	39.464	36.957
1w50_molecule_30_uff_E=1643.99	-6.7	34.788	31.859
1w50_molecule_30_uff_E=1643.99	-6.7	34.953	32.757
1w50_molecule_30_uff_E=1643.99	-6.7	36.817	32.643
1w50_molecule_31_uff_E=1664.81	-8.1	0	0
1w50_molecule_31_uff_E=1664.81	-8	4.499	2.449
1w50_molecule_31_uff_E=1664.81	-8	2.239	1.862
1w50_molecule_31_uff_E=1664.81	-7.8	4.057	2.321
1w50_molecule_31_uff_E=1664.81	-7.5	4.076	2.891
1w50_molecule_31_uff_E=1664.81	-7.5	3.616	2.78
1w50_molecule_31_uff_E=1664.81	-7.5	4.613	2.348
1w50_molecule_31_uff_E=1664.81	-7.4	3.136	2.337
1w50_molecule_31_uff_E=1664.81	-7.4	3.379	2.743

Figure – 15 The 32 molecules (inhibitor dataset) are docked against the bace.

In the above mentioned molecules with the binding affinities the molecule_12 shows the binding affinity of -9.3 which is the highest binding affinity.

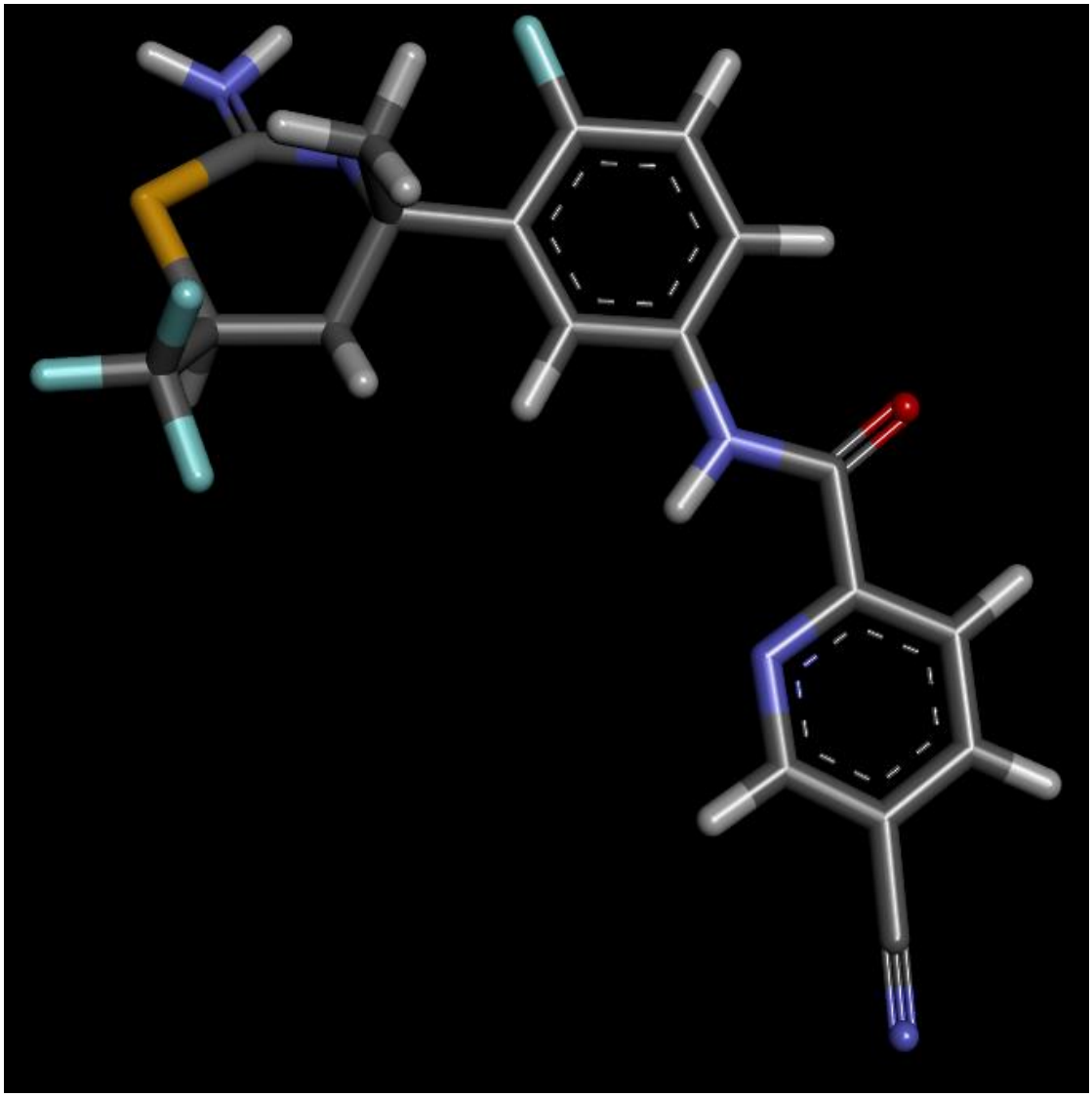

Figure – 16 Molecule-12.

The above mentioned molecule_12 shows the binding affinity of -9.3.

3.3 FBRNN-Forward backward RNN –

On the training on the ChEmbl smiles

Training dataset – chEMBL dataset contains 2327927 molecules

The training of the model generates the graph which shows the training loss as well as the validation loss with respect to the epoch.

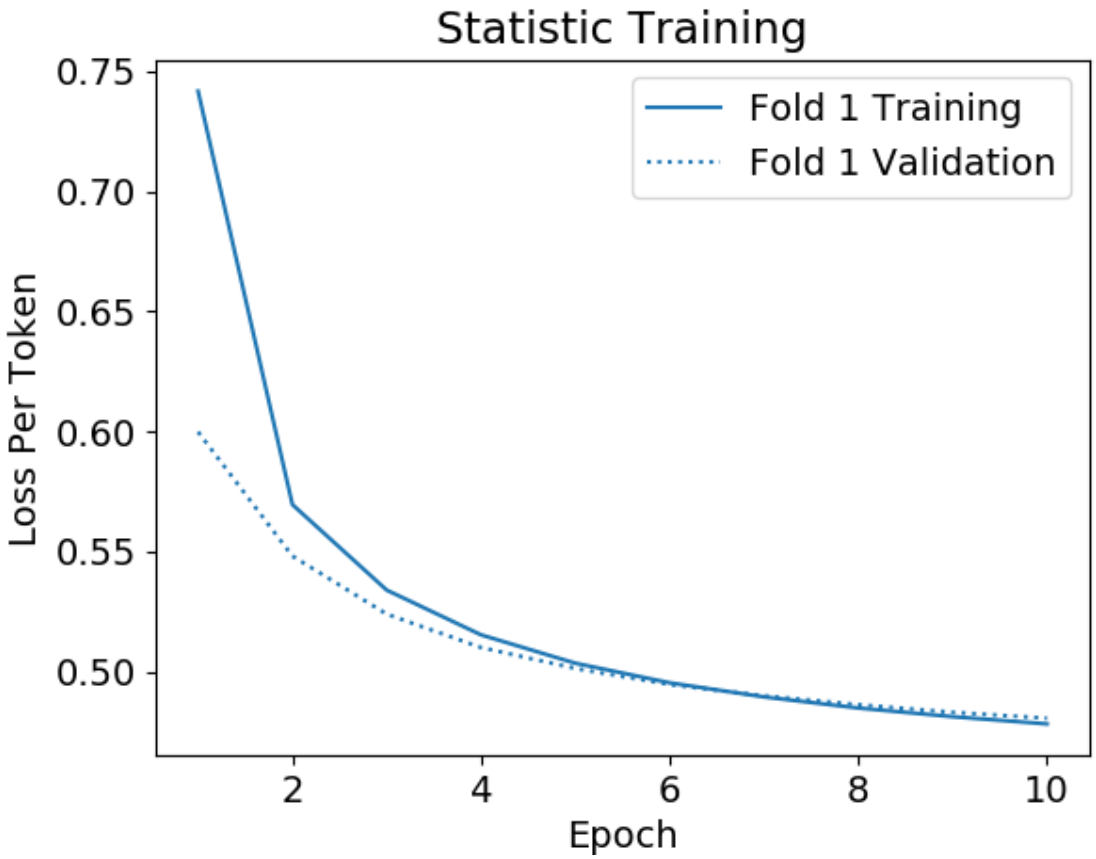

Figure 17 -The training and the validation loss is plotted against the number of epoch.

Here the validation dataset we are using is equivalent to the test dataset,no separate validation dataset is used in this.

Inhibitor dataset – 7234 molecules are present after the pretraining of the dataset only 5794 molecules are present.

The final molecules that are produced after the finetuning on the inhibitor dataset,approximately 100 molecules are produced in this and the tanimoto similarity is calculated between the generated molecules and the chEmbl dataset in this.

The highest binding affinity that between the molecule_10 and the bace is -9.7 this is the highest binding affinity that is calculated after the 100 molecules that are docked against the bace inhibitor in this.

Below mentioned is the canonical smile that shows the binding affinity of -9.7 with the bace molecule -

CC1=C[NH+]=C(C=C1)C2=CC3=C(C=C2)C4CCN(CCC5CC(N)C=C5C6CC36NC4)CC7=CC(=CC=C7)Cl

838	1w50_molecule_9_uff_E=874.79	-7.3	7.826	5.252
839	1w50_molecule_10_uff_E=2109.67	-9.7	0	0
840	1w50_molecule_10_uff_E=2109.67	-9.2	7.453	3.09
841	1w50_molecule_10_uff_E=2109.67	-8.8	2.551	2.18
842	1w50_molecule_10_uff_E=2109.67	-8.7	6.138	3.398
843	1w50_molecule_10_uff_E=2109.67	-8.7	4.671	3.7
844	1w50_molecule_10_uff_E=2109.67	-8.6	9.122	4.836
845	1w50_molecule_10_uff_E=2109.67	-8.4	9.735	4.785
846	1w50_molecule_10_uff_E=2109.67	-8.4	7.805	3.705
847	1w50_molecule_10_uff_E=2109.67	-8.3	24.637	19.956
848	1w50_molecule_2_uff_E=558.23	-7.6	0	0
849	1w50_molecule_2_uff_E=558.23	-7.5	2.266	1.67
850	1w50_molecule_2_uff_E=558.23	-7.5	25.07	22.054
851	1w50_molecule_2_uff_E=558.23	-7.1	5.957	3.157
852	1w50_molecule_2_uff_E=558.23	-7	6.753	3.896
853	1w50_molecule_2_uff_E=558.23	-7	2.241	1.574
854	1w50_molecule_2_uff_E=558.23	-7	5.463	3.57
855	1w50_molecule_2_uff_E=558.23	-6.9	7.183	3.071
856	1w50_molecule_2_uff_E=558.23	-6.8	24.825	21.835

Figure – 18 Highest binding affinity of Molecule-10.

The above figure shows the binding affinity of thee different molecules with the bace and the highest binding affinity that we are getting in this is -9.7 that is shown in this csv file.

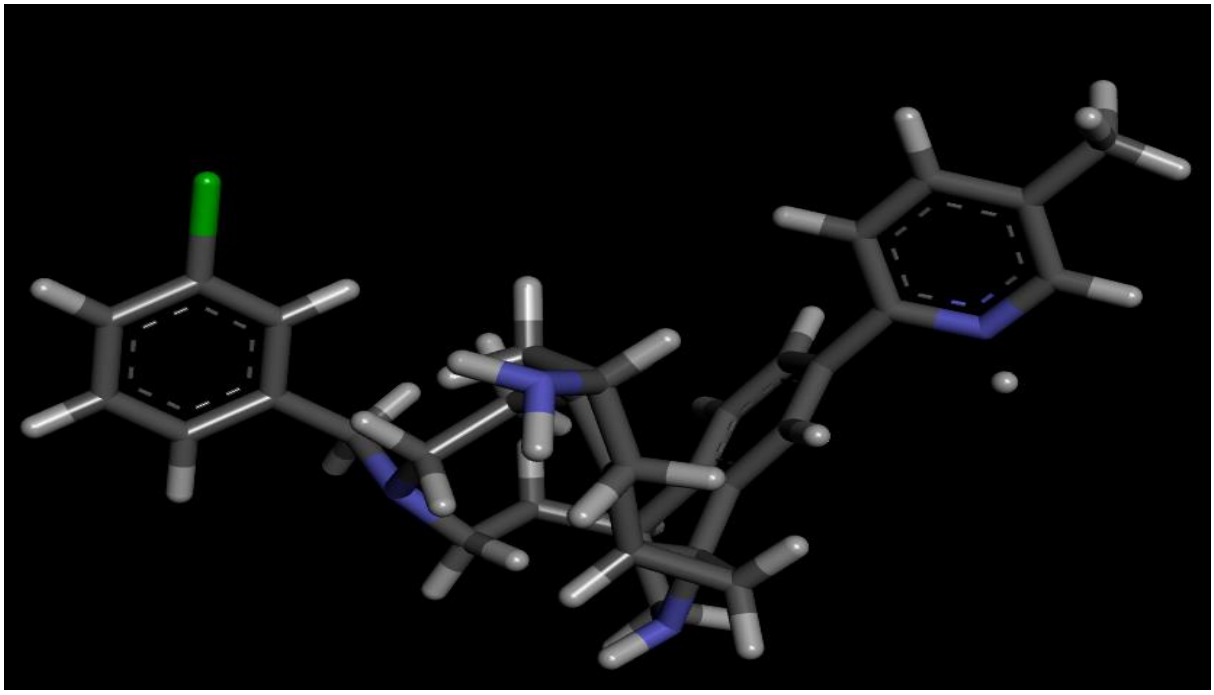

Figure 19 – Molecule_10 showing binding affinity of -9.7.

The above figure shows the molecule_10 which shows the binding affinity of -9.7.

In certain neurodegenerative diseases such as the alzheimers disease,beta-secretase-1(BACE) can indeed be overproduced or overactivated but the use of the generative ai technologies design the molecules which are designed against this target and the overproduction of the enzyme beta-secretase-1 is lead to the formation of the beta amyloid peptides that can lead to the formation of the plaques which are believed to contribute in this neurodegenerative process in the alzheimer's disease.

CHAPTER - 4

4.Conclusion and the Future scope

4.1 Conclusion

Creating compounds that target BACE by utilizing generative AI technologies is a novel strategy with a lot of promise [1]. With the use of AI-driven techniques, large chemical regions can be quickly explored, leading to the discovery of novel compounds with ideal characteristics for BACE inhibition. Researchers can predict the safety, selectivity, and efficacy profiles of prospective compounds and investigate complex chemical interactions with greater efficiency by employing artificial intelligence algorithms [25].
Combining generative AI with drug design against BACE not only speeds up the process of finding new treatments for neurological disorders like Alzheimer's, but also increases the probability of success. Moreover, AI-driven methods enable the iterative optimization and modification of potential compounds based on real-time data, resulting in more accurate and focused treatments [26].
Future developments in generative AI technologies, along with cooperative efforts from biologists, chemists, and computational scientists, will accelerate the creation of compounds that target BACE [27]. Drug development efforts could be revolutionized by this cooperative synergy, giving patients and caregivers dealing with neurodegenerative illnesses new hope [28].

To sum up, the application of generative AI technologies in the development of compounds that oppose BACE has significant potential for the treatment of neurodegenerative illnesses [29]. This novel approach to drug discovery not only expedites the process but also raises the possibility of discovering safe and highly effective treatments. As long as there is ongoing development and cooperative efforts, there is hope for applying AI-driven approaches to address challenging problems in healthcare and improve patient outcomes [30].

Furthermore, new opportunities for innovation in pharmaceutical research are created by the combination of generative AI and conventional drug design methodologies. Through the integration of biochemical knowledge and computational capabilities, researchers are able to investigate a wide range of molecular configurations and make previously unheard-of biological activity predictions [31]. By using a synergistic strategy, the time and resources needed for preclinical and clinical development are greatly reduced, and the identification of possible drug candidates is accelerated [32].

Furthermore, the development of virtual compound libraries is made possible by generative AI, broadening the extent of chemical space investigation beyond what can be accomplished with only conventional laboratory techniques [33]. The probability of discovering new lead compounds with the appropriate pharmacological characteristics -such as potency, selectivity, and ADMET (absorption, distribution, metabolism, excretion, and toxicity) profiles-is increased by using this virtual screening method [34].

In addition, by producing analogs with better drug-like qualities, such increased bioavailability and decreased toxicity, generative AI helps optimize lead compounds. By targeting BACE and other important molecular pathways linked to the evolution of neurodegenerative illnesses, safer and more effective therapies may be developed as a result of this iterative process of design, synthesis, and assessment [35].

In conclusion, generative AI technologies' incorporation into drug design signifies a paradigm shift in the field of pharmaceutical research and gives hitherto unseen chances for creativity and discovery [36]. Researchers can expedite the development of next-generation treatments and solve unmet medical needs in neurodegenerative disorders and beyond by leveraging the capabilities of AI-driven techniques [44].

4.2 Future scope

Large language models (LLMs) have great promise for the development of molecules that target BACE in neurodegenerative diseases [45]. These models have the potential to completely transform this field of drug design as they develop and become more predictive. Improving the produced molecules' selectivity and efficacy is crucial for progress since it guarantees that the compounds have the best possible characteristics to inhibit BACE with the least amount of off-target effects [44]. Furthermore, by evaluating individual genetic data and customizing molecular designs to particular genetic profiles, LLMs can significantly contribute to personalized medicine and improve treatment outcomes.

Furthermore, the multi-targeted strategy made possible by LLMs has enormous promise for the creation of treatments that not only target BACE but also other crucial pathways connected to neurodegeneration [23]. Improved treatment efficacy and synergistic benefits could result from this all-encompassing approach [22].

The ability of LLMs to support medication repurposing initiatives by locating compounds that already have BACE inhibitory activity and established safety profiles for accelerated development is a noteworthy additional feature. To thoroughly confirm LLM-generated compounds' safety, effectiveness, and pharmacological characteristics, it is still necessary to incorporate them into experimental validation procedures. Establishing criteria and protocols for evaluating LLM-generated compounds as possible medicines, taking safety, efficacy, and ethical issues by account, would also require cooperation with regulatory bodies [22].

Additionally, in order to stay current with new scientific findings and technical developments in the field of neurodegenerative disease research, LLMs must constantly study and develop [34]. This includes investigating combination treatments that address several facets of neurodegeneration at once in the hopes of enhancing neuroprotective and disease-modifying effects. All things considered, personalized approaches, multi-targeted strategies, advanced predictive capabilities, and cross-disciplinary collaborations characterize the future of molecule generation with LLMs against the BACE target in neurodegenerative processes, all of which point to

significant advancements in therapeutic interventions for neurodegenerative diseases [33].

Bibliography

[1] Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.;Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu,B.; Zhang, J.; Bryant, S. H. PubChem Substance and Compounddatabases.Nucleic Acids Res.2016,44, D1202−D1213.
[2] Polishchuk, P. G.; Madzhidov, T. I.; Varnek, A. Estimation of thesize of drug-like chemical space based on GDB-17 data.J. Comput.-Aided Mol. Des.2013,27, 675−679.
[3] Shoichet, B. K. Virtual screening of chemical libraries.Nature2004,432, 862−5.
[4] Scior, T.; Bender, A.; Tresadern, G.; Medina-Franco, J. L.;Martinez-Mayorga, K.; Langer, T.; Cuanalo-Contreras, K.; Agrafiotis,D. K. Recognizing Pitfalls in Virtual Screening: A Critical Review.J.Chem. Inf. Model.2012,52, 867−881.
[5] Cheng, T.; Li, Q.; Zhou, Z.; Wang, Y.; Bryant, S. H. Structure-Based Virtual Screening for Drug Discovery: a Problem-CentricReview.AAPS J.2012,14, 133−141.
[6] Pyzer-Knapp, E. O.; Suh, C.; Gómez-Bombarelli, R.; Aguilera-Iparraguirre, J.; Aspuru-Guzik, A. What Is High-Throughput VirtualScreening? A Perspective from Organic Materials Discovery.Annu.Rev. Mater. Res.2015,45, 195−216.
[7] Schneider, G. Virtual screening: an endless staircase?Nat. Rev.Drug Discovery2010,9, 273−276.
[8] Hachmann, J.; Olivares-Amaya, R.; Atahan-Evrenk, S.; Amador-Bedolla, C.; Sánchez-Carrera, R. S.; Gold-Parker, A.; Vogt, L.;Brockway, A. M.; Aspuru-Guzik, A. The Harvard clean energy project:large-scale computational screening and design of organic photo-voltaics on the world community grid.J. Phys. Chem. Lett.2011,2,2241−2251.
[9] Gómez-Bombarelli, R.; et al.Nat. Mater.2016,15, 1120−1127.
[10] Virshup, A. M.; Contreras-García, J.; Wipf, P.; Yang, W.;Beratan, D. N. Stochastic Voyages into Uncharted Chemical SpaceProduce a Representative Library of All Possible Drug-LikeCompounds.J. Am. Chem. Soc.2013,135, 7296−7303.
[11] Rupakheti, C.; Virshup, A.; Yang, W.; Beratan, D. N. StrategyTo Discover Diverse Optimal Molecules in the Small MoleculeUniverse.J. Chem. Inf. Model.2015,55, 529−537.
[12] Reymond, J.-L. The Chemical Space Project.Acc. Chem. Res.2015,48, 722−730.
[13] Reymond, J.-L.; van Deursen, R.; Blum, L. C.; Ruddigkeit, L.Chemical space as a source for new drugs.MedChemComm2010,1,30−38.
[14] Kanal, I. Y.; Owens, S. G.; Bechtel, J. S.; Hutchison, G. R.Efficient Computational Screening of Organic Polymer Photovoltaics.J. Phys. Chem. Lett.2013,4, 1613−1623.
[15] O'Boyle,N.M.;Campbell,C.M.;Hutchison,G.R.Computational Design and Selection of Optimal Organic PhotovoltaicMaterials.J. Phys. Chem. C2011,115, 16200−16210.
[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems, pages 5998–6008, 2017.
[17] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving language understanding by generative pre-training. 2018.
[18] J. Devlin, M. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational

Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), page 4171–4186. Association for Computational Linguistics, 2019.
[19] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T.L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A.M. Rush. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online, oct 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.
[20] J. Vig. A multiscale visualization of attention in the transformer model. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 37–42, Florence, Italy, July 2019. Association for Computational Linguistics. doi:10.18653/v1/P19-3007. URL https://www.aclweb.org/ anthology/P19-3007.
[21] J. Pfeiffer, A. Rücklé, C. Poth, A. Kamath, I. Vulic, S. Ruder, K. Cho, and I. Gurevych. Adapterhub: A framework ´ for adapting transformers. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 46–54, 2020.
[22] Z. Wu, B. Ramsundar, E.N. Feinberg, J. Gomes, C. Geniesse, A.S. Pappu, K. Leswing, and V. Pande. Moleculenet: a benchmark for molecular machine learning. In Chemical science, 9(2), pages 513—530, 2018.
[23] N. Brown, M. Fiscato, M.H.S. Segler, and A.C. Vaucher. Guacamol: Benchmarking models for de novo molecular design. Journal of Chemical Information and Modeling, 59(3):1096–1108, 2019. doi:10.1021/acs.jcim.8b00839. URL https://doi.org/10.1021/acs.jcim.8b00839.
[24] D. Weininger. Smiles, a chemical language and information system. introduction to methodology and encoding rules. In J.Chem. Inf. Comput. Sci, pages 28–31, 1988.
[25] M.H.S. Segler, T. Kogej, C. Tyrchan, and M.P. Waller. Generating focused molecule libraries for drug discovery with recurrent neural networks. ACS Central Science, 4(1):120–131, 2018. doi:10.1021/acscentsci.7b00512. URL https://doi.org/10.1021/acscentsci.7b00512. PMID: 29392184.
[26] S. Chithrananda, G. Grand, and B. Ramsundar. Chemberta: Large-scale self-supervised pretraining for molecular property prediction. arXiv preprint arXiv:2010.09885, 2020.
[27] P. Karpov, G. Godin, and I.V. Tetko. Transformer-cnn: Swiss knife for qsar modeling and interpretation. In Journal of Cheminformatics 12, 2020. doi:10.1186/s13321-020-00423-w.
[28] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. 2019.
[29] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly. Parameter-efficient transfer learning for nlp. arXiv preprint arXiv:1902.00751, 2019.
[30] A. Gupta, A. Müller, B. Huisman, J. Fuchs, P. Schneider, and G. Schneider. Generative recurrent networks for de novo drug design. Molecular Informatics, 37, 11 2017. doi:10.1002/minf.201700111.
[31] Dobson, C. M. Chemical Space and Biology. Nature 2004, 432, 824−828.

[32] Walters, W. P.; Stahl, M. T.; Murcko, M. A. Virtual screening-an overview. Drug Discov. Today 1998, 3, 160−178.
[33] Munk, M. E. Computer-Based Structure Determination: Then and Now. J. Chem. Inf. Comput. Sci. 1998, 38, 997−1009.
[34] Fink, T.; Reymond, J.-L. Virtual Exploration of the Chemical Universe up to 11 Atoms of C, N, O, F: Assembly of 26.4 Million Structures (110.9 Million Stereoisomers) and Analysis for New Ring Systems, Stereochemistry, Physicochemical Properties, Compound Classes, and Drug Discovery. J. Chem. Inf. Model. 2007, 47, 342−353.
[35] Wieland, T.; Kerber, A.; Laue, R. Principles of the Generation of Constitutional and Configurational Isomers. J. Chem. Inf. Comput. Sci. 1996, 36, 413−419.
[36] Miyao, T.; Arakawa, M.; Funatsu, K. Exhaustive Structure Generation for Inverse-QSPR/QSAR. Mol. Inf. 2010, 29, 111−125.
[37] Miyao, T.; Kaneko, H.; Funatsu, K. Inverse QSPR/QSAR Analysis for Chemical Structure Generation (from y to x). J. Chem. Inf. Model. 2016, 56, 286−299.
[38] Miyao, T.; Kaneko, H.; Funatsu, K. Ring-System-Based Exhaustive Structure Generation for Inverse-QSPR/QSAR. Mol. Inf. 2014, 33, 764−778.
[39] Fechner, U.; Schneider, G. Flux (2): Comparison of Molecular Mutation and Crossover Operators for Ligand-Based de Novo Design. J. Chem. Inf. Model. 2007, 47, 656−667.
[40] Devi, R. V.; Sathya, S. S.; Coumar, M. S. Evolutionary algorithms for de novo drug design□A survey. Appl. Soft Comput. 2015, 27, 543− 552.
[41] Hartenfeller, M.; Zettl, H.; Walter, M.; Rupp, M.; Reisen, F.; Proschak, E.; Weggen, S.; Stark, H.; Schneider, G. DOGS: ReactionDriven de Novo Design of Bioactive Compounds. PLoS Comput. Biol. 2012, 8, No. e1002380.
[42] Button, A.; Merk, D.; Hiss, J. A.; Schneider, G. Automated de Novo Molecular Design by Hybrid Machine Intelligence and RuleDriven Chemical Synthesis. Nat. Mach. Intell. 2019, 1, 307−315.
[43] Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. Learning Internal Representations by Error Propagation, ICS-8506; California University San Diego La Jolla, Institute for Cognitive Science, 1985. (14) Hopfield, J. J. Neural Networks and Physical Systems with Emergent Collective Computational Abilities. Proc. Natl. Acad. Sci. U.S.A. 1982, 79, 2554−2558.
[44] Makhzani, A.; Shlens, J.; Jaitly, N.; Goodfellow, I.; Frey, B. Adversarial Autoencoders. 2015, arXiv:1511.05644. arXiv preprint.