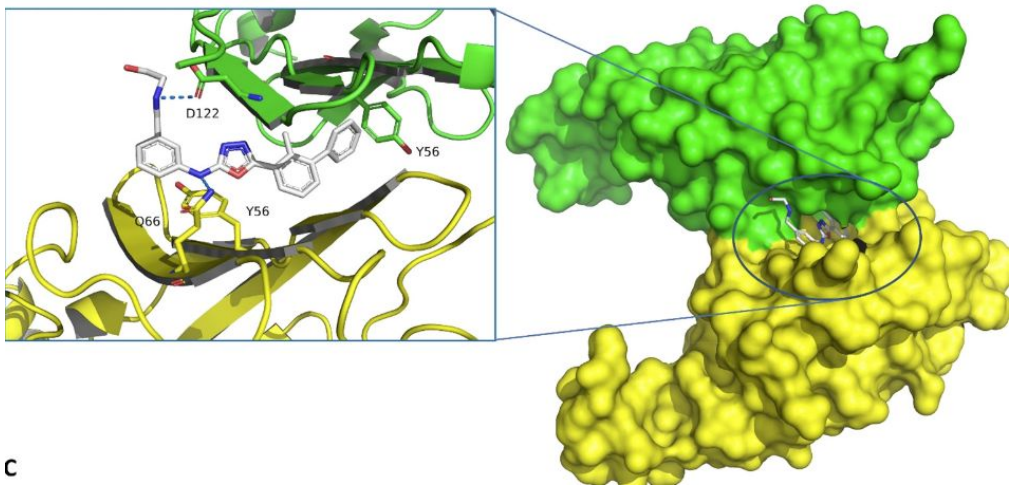
Latent Diffusion for Protein-Targeted Small Molecule Generation

Srikar K., Rushil Y., Arnav S., Tyler K.

Background

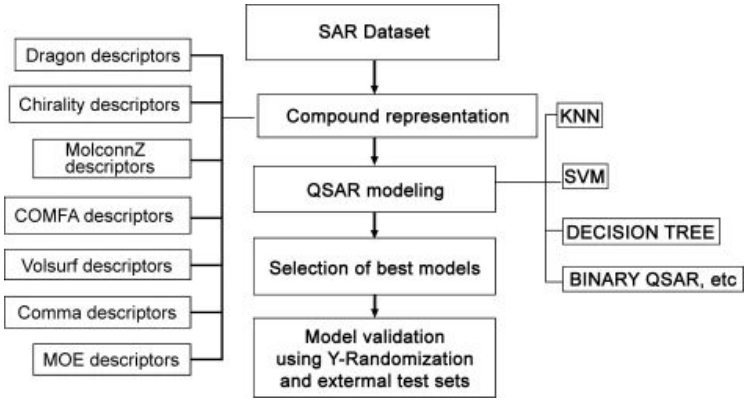
Protein and Small Molecule Interactions

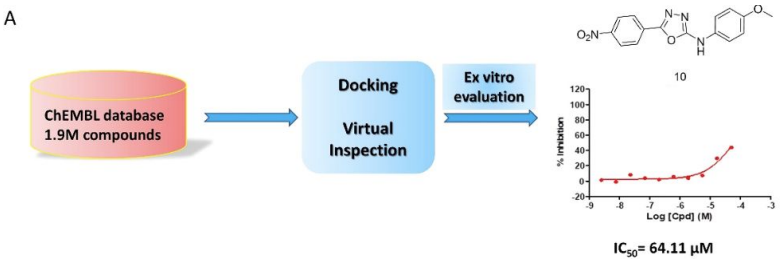
- modulate protein function via targeting active sites → **high specificity** and **selectivity** for therapeutics & **diversity of function**
- size (normally >900 Da) corresponds to **good diffusion properties** in vivo
- **high structural diversity** of small molecule candidates



Small Molecule Inhibitor Development

- **virtual screening** on existing databases (ChEMBL, DrugBank, etc.)
  - Ligand-based
  - Structure-based (docking & other simulations)
- **high-throughput screening** — high volume in vitro testing of candidates
- **QSAR** models — prediction of ligand characteristics from structural/computed features

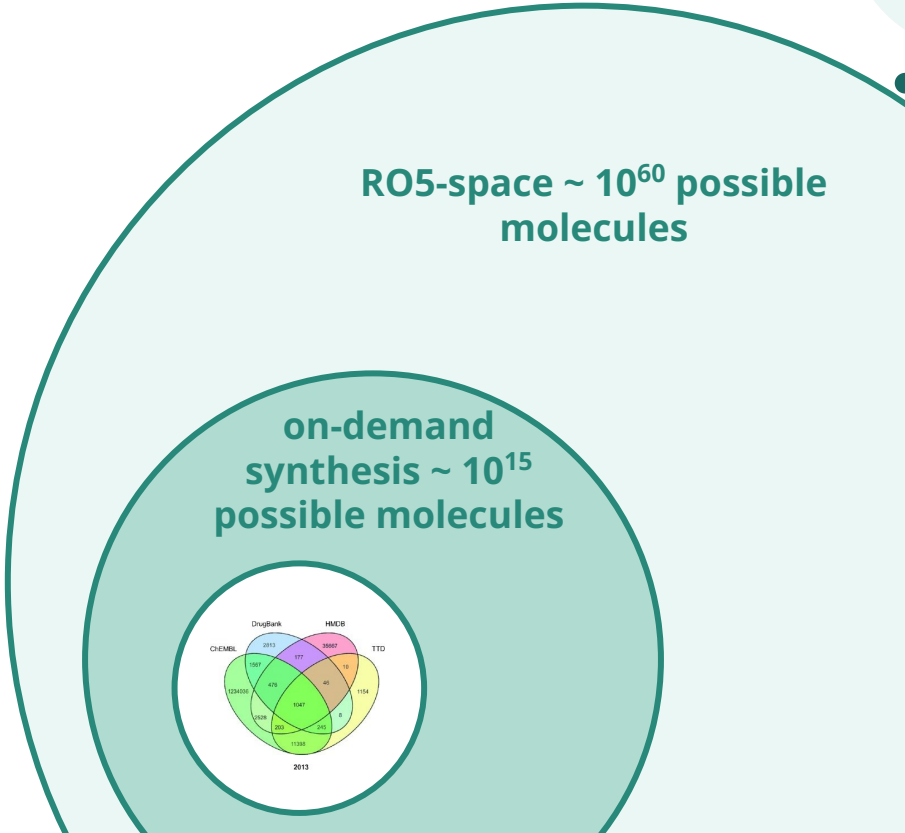




Current Limitations

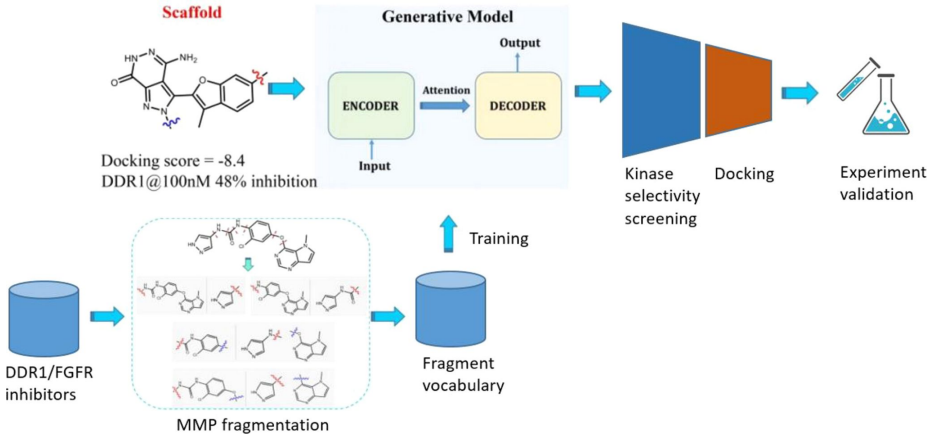
Challenges in Small Molecule Discovery

- current methods are **limited in scope** to already-discovered & highly-adjacent compounds
  - ChEMBL: ~2.4 million
  - DrugBank: ~16.5K
- simulations can be **expensive** and **inaccurate**
- expensive and **impractical** to do in vitro analysis of large # of chemicals

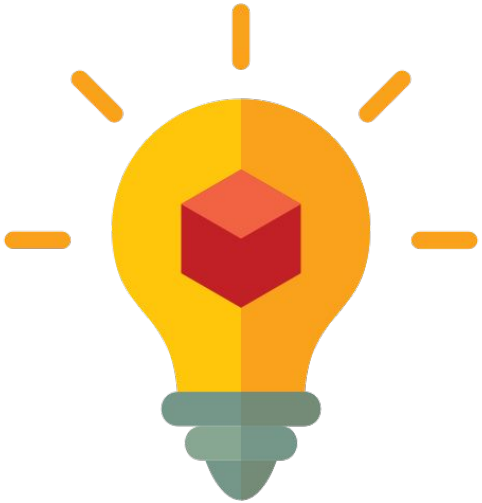


Generative Models for Discovery

- generative models **not limited** to already-explored drug space → new chemistry & increased diversity
- Example generative model types include **RNNs**, **GANs**, and **VAEs**
- turn toward **NLP approaches** (SMILES, sequence info) → transformer-based architecture & other approaches



Model Goals & Novelty



Model Goals

Simple Terms

Input Protein Sequence, output SMILES of protein-binding small molecule

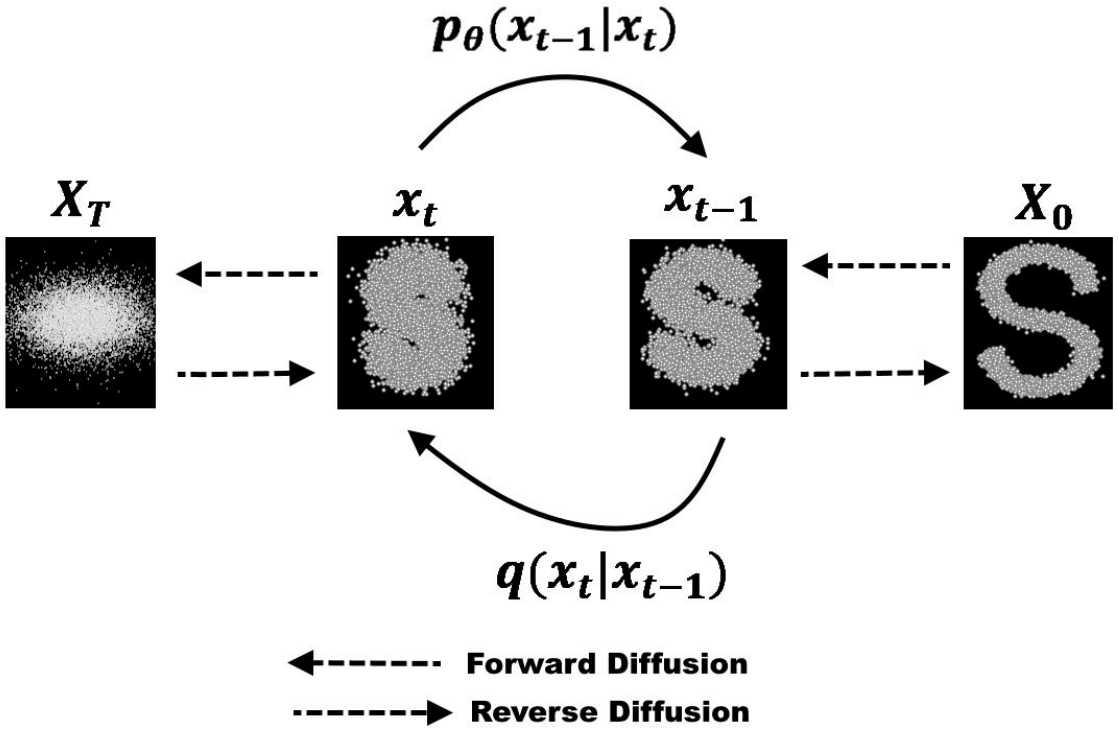
Academic Speak

Generation molecular embeddings directly from protein embeddings via **latent sequence diffusion**, thereby creating a novel link between protein sequences and potential interacting molecules.

Our Novelty

- **Innovative Approach to Molecule Generation:**
  - Unlike traditional methods that focus on molecular structures, our model explores latent sequence diffusion, offering a new perspective in molecule generation.
  - To our knowledge, the *first* of its kind in protein-targeted small molecule research.

What is Diffusion?



Data Curation

BindingDB

- **Components**
  - High-Affinity Protein Interactions: Proteins with ligand interactions stronger than 100 nM IC50/Kd/EC50.
  - Ligand Criteria: Compounds with PubChem CIDs, representable by SMILES, <1000 Da, linked to Uniprot IDs.
  - Protein Characteristics: Sequences ranging from 80 to 1000 amino acids, functioning as monomers
- **Selection**
  - A dataset of 5000 protein, molecule pairs were chosen at random
  - Train: Test: Validation Split → 70:15:15

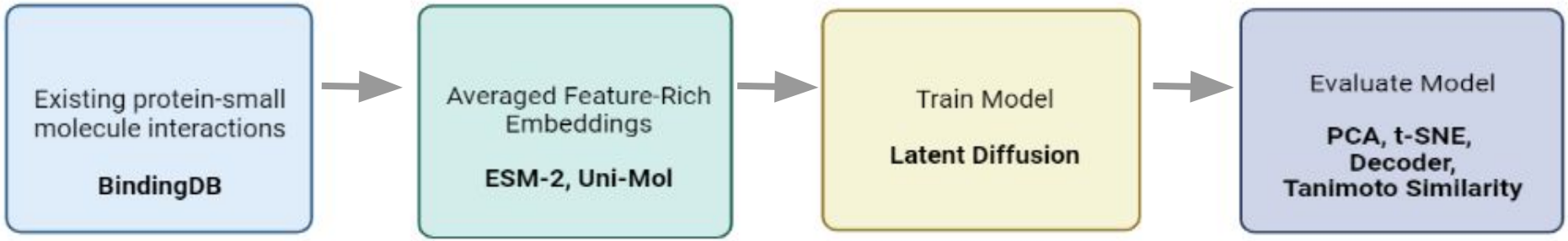
Improvements to Data Curation

- **Refine via clustering**
  - Refine the dataset by sequence clustering at the 30% identity level using the tool Mmseqs2, aiming to cover a diverse range of sequences for training, validation, and testing.
- **Validate via PDB**
  - The validation datasets should be carefully assembled by pairing PDB protein-ligand co-crystal structures with BindingDB entries.

ESM-2 + Uni-Mol Embeddings

- **ESM-2:**
  - Condenses High-Dimensional Data: It effectively reduces the complexity of protein sequences into a single embedding dimension shape vector, preserving essential information in a more manageable form.
  - Captures Sequential Information: ESM-2's embeddings are adept at encapsulating the sequential nature of protein sequences, which is crucial for understanding the structural and functional aspects of proteins. However, we used *average* ESM-2 embeddings.
- **Uni-Mol:**
  - Averaging of Tokens: Uni-Mol uses an approach that averages the embeddings of SMILES tokens, which are representations of molecular structures, effectively capturing the chemical characteristics of molecules.
  - Reduction to Single Vector: The model condenses the entire sequence and SMILES tokenized sequence length into a single embedding dimension shape vector, efficiently representing complex molecular information.

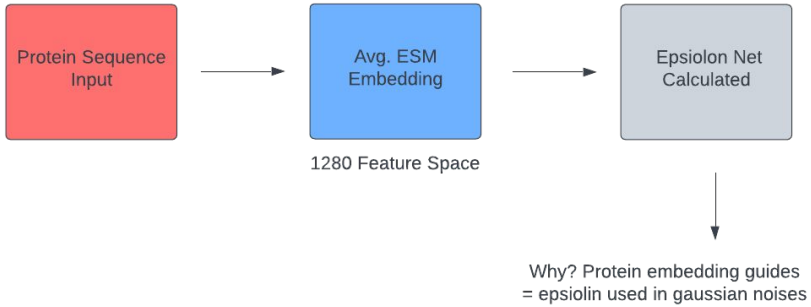
Workflow Diagram



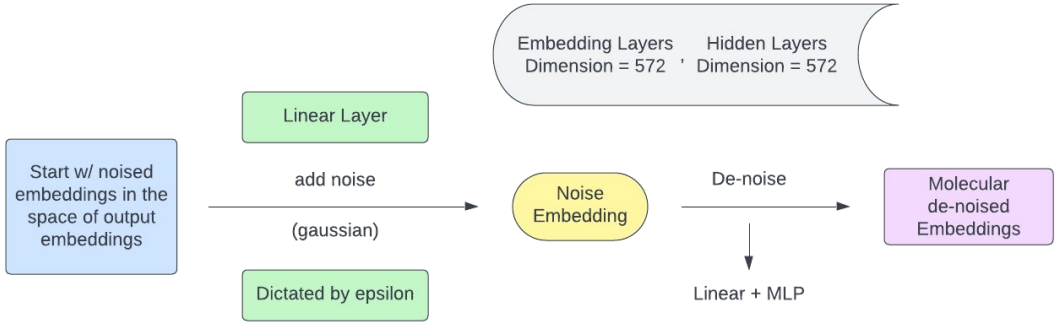
Model Architecture

Model Diagram

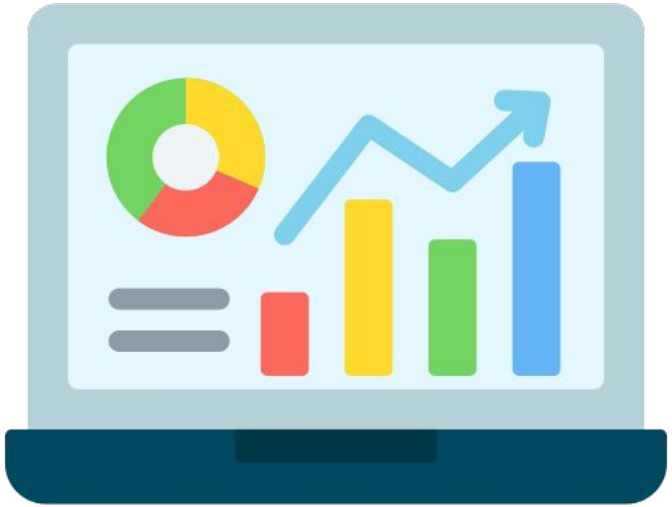
1



2

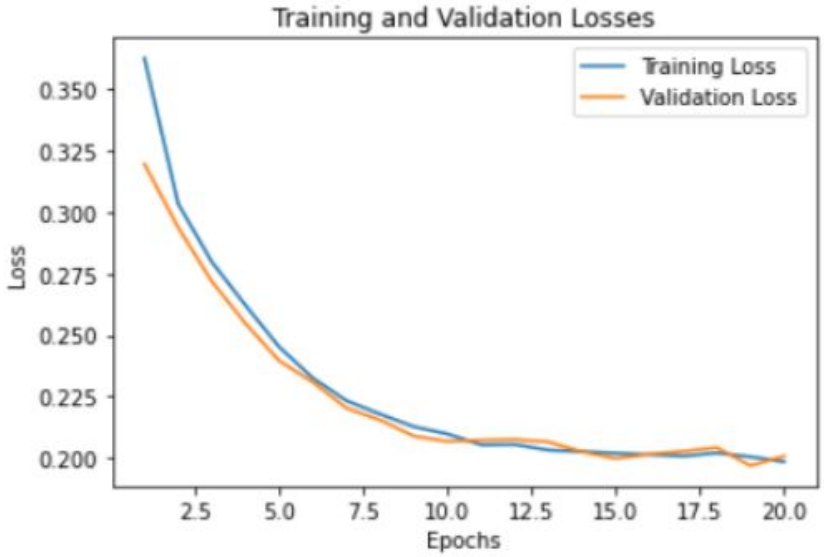


Results



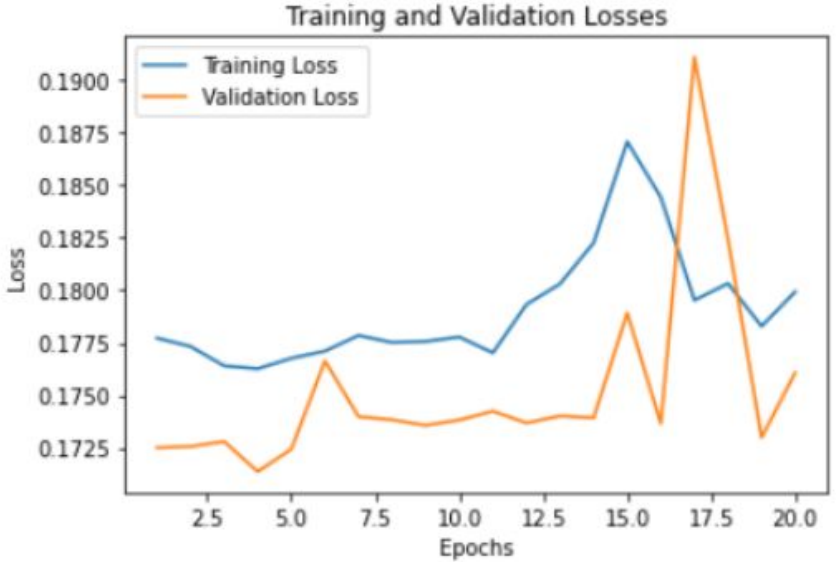
Model Loss (L1 Loss)

Finalized Model (Linear Epsilon Net)

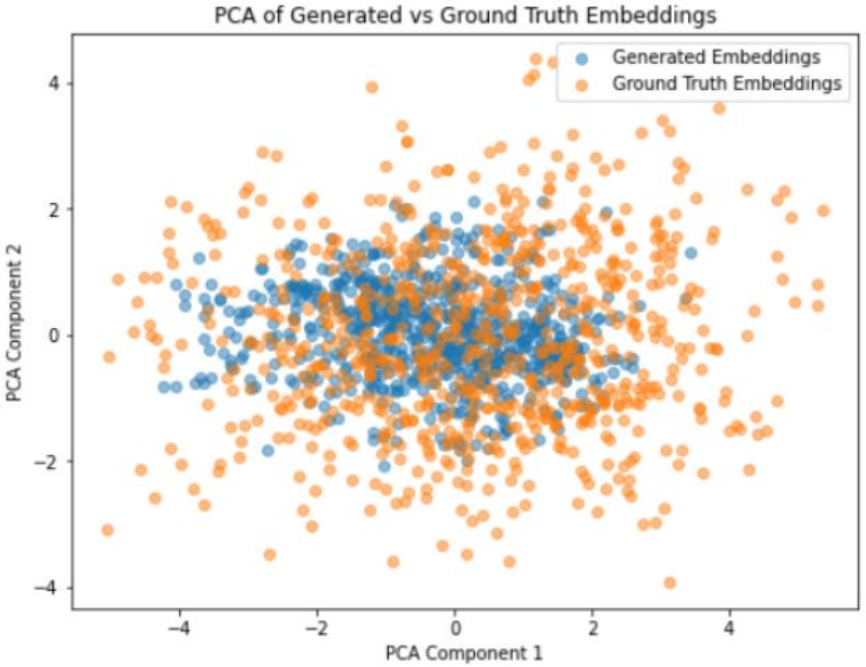


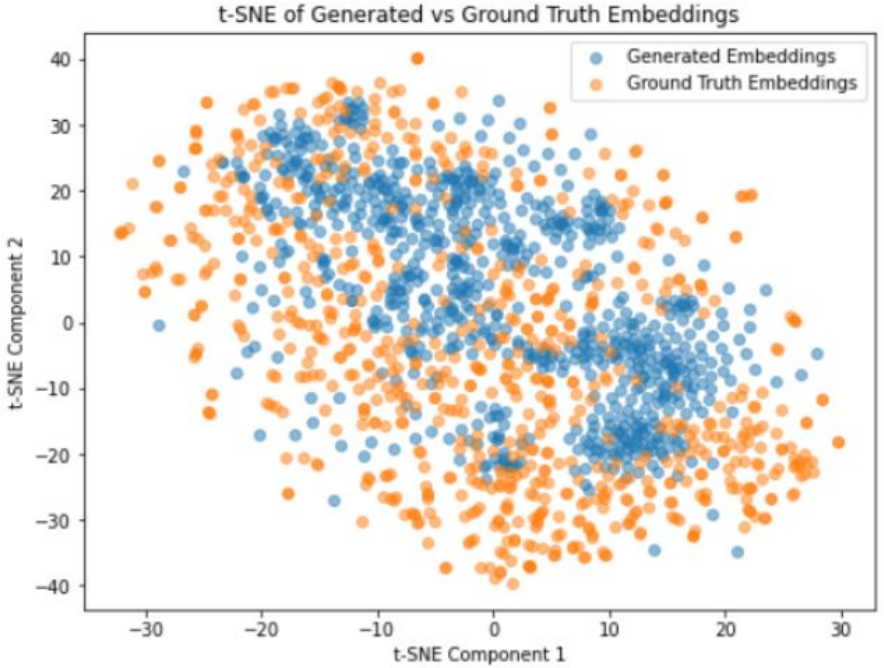
Important Note: Denoising architecture did not change model loss performance significantly.

One of tested variations (RNN Epsilon Net)

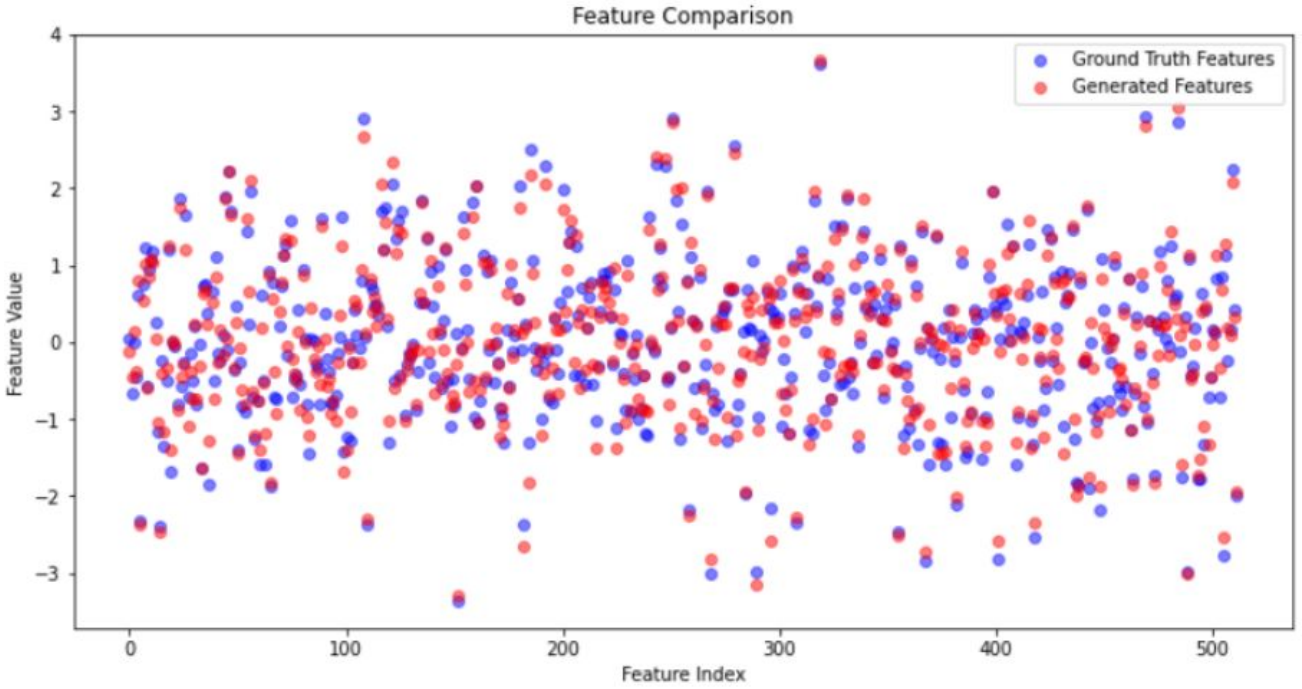


Generated vs. Ground Truth Latent Visualization





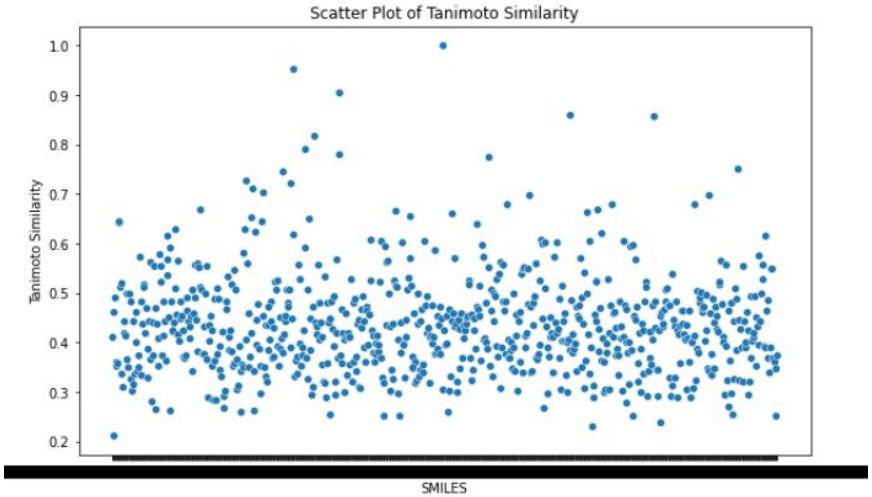
One Sample Generation (Expanded Chemical Space)

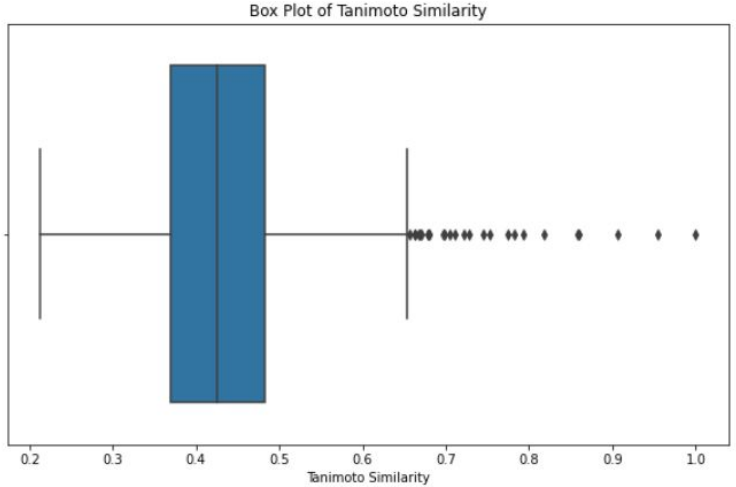


Decoded SMILES Output Comparisons (Look-up Table)

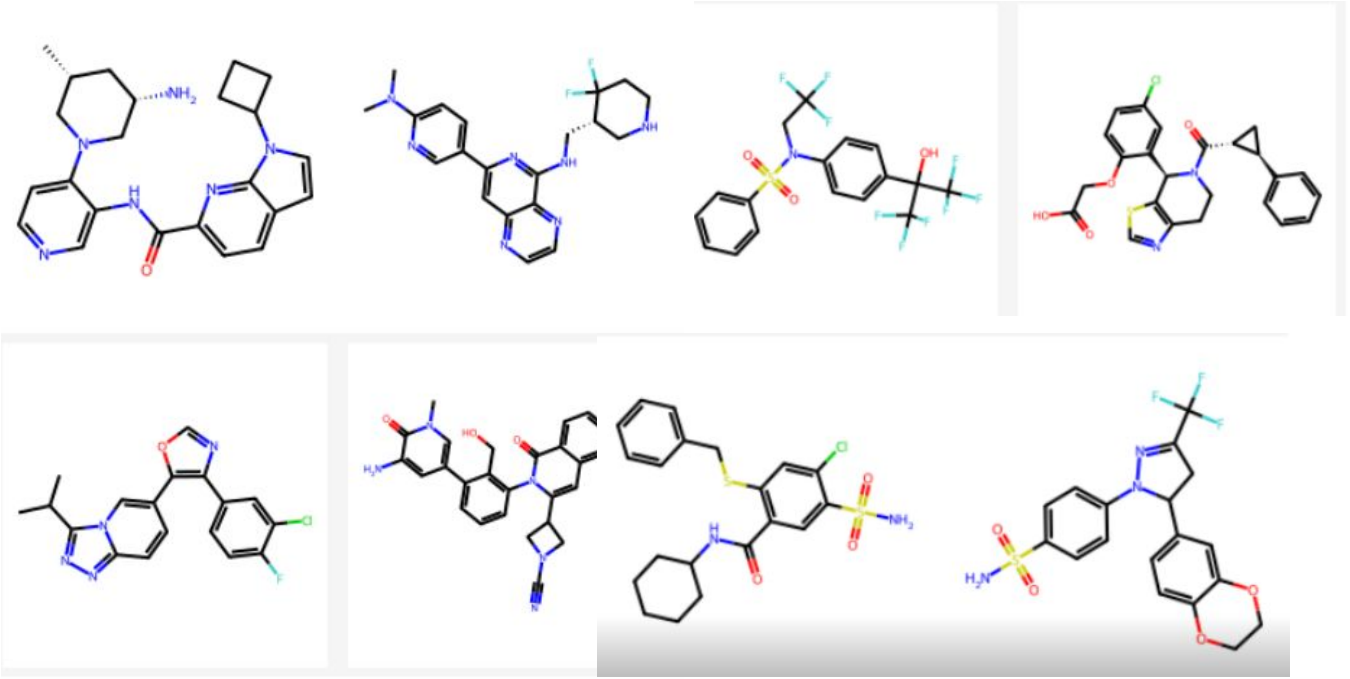
smiles	Decoded_SMILES	Min_Distance
C[C@@H]1C[C@H](N)CN(C1)c1ccncc1NC(=O)c1ccc2ccn...	CN(C)c1ccc(cn1)-c1cc2nccnc2c(NC[C@H]2CNCCC2(F)...	4.233210
CC(C)c1nnc2ccc(cn12)-c1ocnc1-c1ccc(F)c(Cl)c1	Cn1cc(cc(N)c1=O)-c1cccc(c1CO)-n1c(cc2cc(ccc2c1...	4.535190
CSc1nc(c([nH]1)-c1ccc(F)cc1)-c1ccnc(NC2CCCCC2C)c1	CC(C)c1cc(C(=O)N2Cc3ccc(CN4CCN(C)CC4)cc3C2)c(O...	4.616539
CCN(CC)CCN1C(=O)[C@](O)(c2c1cc(cc2C(F)(F)F)C(N...	COc1ccc(Cn2c(nnc2[C@H](Cc2ccccc2)NC(C)=O)[C@@H...	4.581069
NS(=O)(=O)c1cc(C(=O)NC2CCCCC2)c(SCc2ccccc2)cc1Cl	NS(=O)(=O)c1ccc(cc1)N1N=C(CC1c1ccc2OCCOc2c1)C(...	3.765682
...	...	...
NC(=O)c1ccc2cc(ccc2c1)C1(O)CCn2cncc12	Cc1c(-c2ccnc3c(F)cccc23)c2cc(C)ccc2n1CC(O)=O	3.492570
OC(c1ccc(cc1)N(CC(F)(F)F)S(=O)(=O)c1ccccc1)(C(...	OC(=O)COc1ccc(Cl)cc1[C@H]1N(CCc2ncsc12)C(=O)[C...	3.526263
CCc1cc2c(cc1C(=C)c1ccc(cc1)C(O)=O)C(C)(C)CCC2(C)C	Cc1c(-c2ccnc3c(F)cccc23)c2cc(C)ccc2n1CC(O)=O	3.844205
O=C1NCc2cc(ccc12)S(=O)(=O)NCCNC\C=C\c1ccc(cc1)...	Cn1ccc(Nc2nc(N)cc(n2)-c2cccc(c2CO)-n2ccc3cc(cc...	4.524452
CN1CC[C@H](CC1=O)c1cccnc1Oc1ccc(Nc2nc3ccccc3s2...	COc1ccccc1-c1nccnc1C1CN(C1)C(=O)c1nc2ccccc2[nH]1	4.654112

Tanimoto Similarity (Generated vs. Ground Truth)

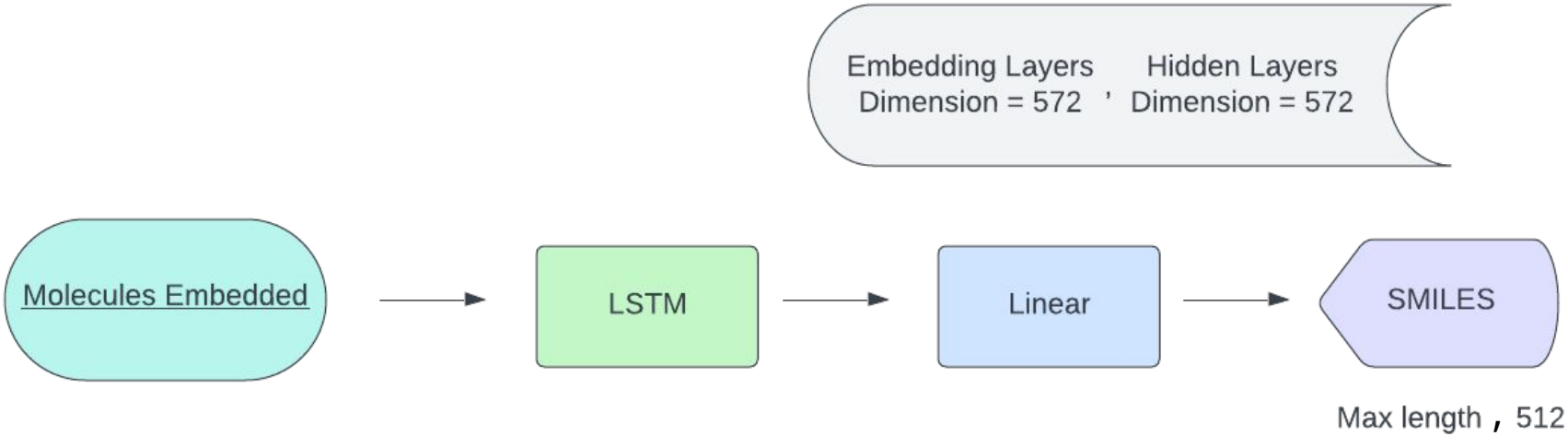




Decoded Structure Output Comparisons (Look-up Table)



LSTM Decoder Architecture



Limitation of Architecture

- **Limitations:**
  - Static Context: Lacks temporal dynamics, providing the same context at each timestep.
  - Over Reliance on Initial Embeddings: Limits sequence diversity and adaptability.
  - Risk of Redundancy: Repetitive, unvaried outputs in longer sequences.
- **Alternatives**:
  - Positional Encodings: To add temporal information.
  - Direct Sequence Generation: Fastest using seq2seq internal model.
  - Autoregressive Generation: One sequence at a time for dynamic context.
  - Learned Transformations: Simplest method from vector to sequence.
  - Recommended for Our Application: Combine learned transformation with positional encoding for optimal results.

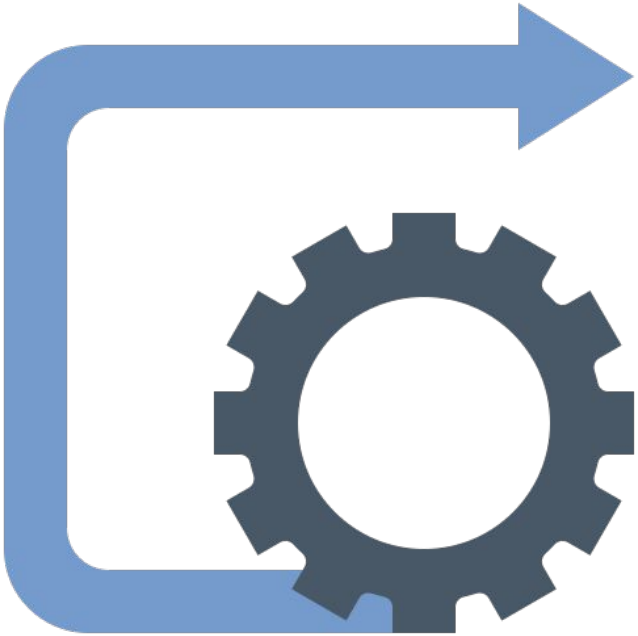
```
{'#': 1,
 '(': 2,
 ')': 3,
 '+': 4,
 '-': 5,
 '/': 6,
 '1': 7,
 '2': 8,
 '3': 9,
 '4': 10,
 '5': 11,
 '6': 12,
 '7': 13,
 '8': 14,
 '=': 15,
 '@': 16,
 'B': 17,
 'C': 18,
 'F': 19,
 'H': 20,
 'I': 21,
 'L': 22,
 'N': 23,
 'O': 24,
 'P': 25,
 'S': 26,
 '[': 27,
 '\\': 28,
 ']': 29,
 'c': 30,
 'e': 31,
 'i': 32,
 'l': 33,
 'n': 34,
 'o': 35,
 'r': 36,
 's': 37,
 '<PAD>': 0}
```

Decoded SMILES (LSTM-Decoder)

	Decoded_SMILES
0	CCCCCCCCCccccccccccccccccccccccccccccccccccccccc...
1	Cc1ccccccccccccccccccccccccccccccccccccccccc<P...
2	CC(Ccccccccccccccccccccccccccccccccccccccccccc...
3	CC(Ccc(ccccccccccccccccccccccccccccccccccccccc...
4	CC(==((ccccccccccccccccccccccccccccccccccccc<...
...	...
745	CC(ccccccccccccccccccccccccccccccc<PAD><PAD><...
746	FC(Fccccccccccccccccccc(((((((((((((((()))))))...
747	CC(=ccccccccccccccccccccccccccccc))))))))<P...
748	Oc1ccccccccccccccccccccccccccccccccccccccccc...
749	Cc1ccc(Ccccccccccccccccccccccccccccccccccccc...

750 rows × 1 columns

Limitations & Future Work



Limitations

- **Featurized Embeddings are Averaged**
  - The averaging of the embeddings takes away the sequential nature of the proteins and molecules.
  - Suggestion: Experiment with tokenized embeddings, not average embeddings.
- **Transition Model Complexity:**
  - Current linear architecture might be too simplistic for modeling noise addition and removal in molecule generation.
  - Suggestion: Incorporate convolutional layers or attention mechanisms for enhanced spatial and sequential understanding.

Limitations

- **Variance Schedule Rigidity:**
  - Predetermined variance schedule may not fit all protein-molecule interactions, potentially limiting diffusion effectiveness.
  - Suggestion: Implement an adaptive variance schedule for more effective diffusion processes
- **Output Interpretability:**
  - Generated molecular embeddings are abstract and challenging to interpret in chemical structure terms.
  - Suggestion: Add a decoding mechanism to convert embeddings into interpretable structures like SMILES.
- **Epsilon Network Design:**
  - Linear layers may not capture complex, non-linear relationships between protein and molecule embeddings.
  - Suggestion: Use transformer layers or gated recurrent units for better pattern recognition.

Future Avenues

- **Enhancing Diversity and Generalization:**
  - Explore adversarial training, domain adaptation, and few-shot learning to improve generalization and molecular diversity.
- **Efficiency Improvements:**
  - Optimize model architecture and use efficient algorithms to reduce computational demands.
- **Integration with Other Biological Data:**
  - Integrate additional biological data like gene expression profiles to generate more targeted molecules and specific diseases.

Questions?