**ARS ASSIGNMENT**

**DELIVERABLE 02**

**Group Members:**

|  |  |
| --- | --- |
| **Roll No.** | **Name** |
| 22F-3156 | Nashrah Ahmed Khan |
| 22F-3158 | Abeera Imran |
| 22F-3095 | Kashaf Nadeem |

Contents

[**R2.1: HNSW Indexing for Similarity Search** 3](#_Toc200333169)

[**Diagram:** 3](#_Toc200333170)

[**Conceptual Level:** 3](#_Toc200333171)

[**Programming Level:** 4](#_Toc200333172)

[**R2.2: LLM Tuning Setup** 5](#_Toc200333173)

[**MolT5 Diagram:** 5](#_Toc200333174)

[**ChemBERTa Diagram:** 5](#_Toc200333175)

[**BioGPT Diagram:** 6](#_Toc200333176)

[**Conceptual Level:** 6](#_Toc200333177)

[**Programming Level:** 6](#_Toc200333178)

[**R2.3: Two-Phase Tuning Process** 8](#_Toc200333179)

[**Diagram:** 8](#_Toc200333180)

[**Phase 1: Dataset-Only Tuning** 8](#_Toc200333181)

[**Programming Level:** 9](#_Toc200333182)

[**Phase 2: HNSW-Informed Tuning** 9](#_Toc200333183)

[**R2.4: Recommendation System** 10](#_Toc200333184)

[**Conceptual Level:** 10](#_Toc200333185)

[**Programming Level:** 10](#_Toc200333186)

[**Example Recommendations:** 11](#_Toc200333187)

[**CHARTS** 12](#_Toc200333188)

[**Chart 1: Training and Validation Loss Curves** 12](#_Toc200333189)

[**Chart 2: Precision-Recall Curve** 13](#_Toc200333190)

[**Chart 3: Training and Validation Accuracy Curves** 14](#_Toc200333191)

[**Chart 4: Discounted Cumulative Gain (DCG)** 15](#_Toc200333192)

[**Chart 5: Perplexity Curve** 16](#_Toc200333193)

[**Chart 6: Training Time per Epoch** 17](#_Toc200333194)

[**Chart 7: Embeddings Visualization (t-SNE)** 18](#_Toc200333195)

[**Chart 8: Recommendation Diversity** 19](#_Toc200333196)

[**COMPUTATIONAL DIFFERENCE** 20](#_Toc200333197)

[**PSEUDOCODE** 22](#_Toc200333198)

**EXPLANATIONS**

# **R2.1: HNSW Indexing for Similarity Search**

## **Diagram:**



## **Conceptual Level:**

HNSW indexing organizes GNN-generated embeddings into a multi-layered graph for efficient nearest-neighbor searches of chemical compounds. This enables rapid identification of structurally or functionally similar compounds, critical for drug discovery. Higher layers provide coarse navigation, while lower layers refine the search, balancing speed and accuracy. The embeddings, derived from Morgan fingerprints via a GNN, capture molecular features like functional groups. The HNSW index supports the recommendation system by retrieving contextually relevant compounds for LLM prompts. The system scales to large datasets (14,823 compounds), ensuring practical applicability.

## **Programming Level:**

1. **Embedding Preparation:** GNN embeddings (256-dimensional) are extracted from the SMILES dataset, converted to a NumPy array (float32), and L2-normalized using faiss.normalize\_L2.
2. **Index Construction:** An HNSW index is initialized with faiss.IndexHNSWFlat (M=32, efConstruction=200) for high connectivity and search quality.
3. **Index Population:** 14,823 embeddings are added to the index, creating a hierarchical graph.
4. **Persistence:** The index is saved to gnn\_hnsw\_index.faiss for reusability.

# **R2.2: LLM Tuning Setup**

## **MolT5 Diagram:**



## **ChemBERTa Diagram:**



## **BioGPT Diagram:**



## **Conceptual Level:**

Three LLMs—MolT5, ChemBERTa, and BioGPT—are fine-tuned to generate chemical compound recommendations for drug discovery. MolT5, a sequence-to-sequence model, excels at generating SMILES strings and descriptions. BioGPT, designed for biomedical tasks, produces detailed therapeutic insights. ChemBERTa, a classification model, is less suited for generative tasks but included for completeness. Fine-tuning adapts these models to SMILES-based prompts, enhancing their domain-specific performance. Mixed precision training ensures computational efficiency despite memory constraints.

## **Programming Level:**

**MolT5**

* Model: laituan245/molt5-large-smiles2caption, loaded with AutoModelForSeq2SeqLM.
* Setup: Moved to GPU, Adam optimizer (lr=1e-5), mixed precision enabled.
* Data: 100 SMILES strings with templates (e.g., “This compound may exhibit anti-inflammatory properties”).
* Training: Two epochs, batch size=1, gradient accumulation=4, tokenized inputs (max\_length=128).
* Output: Fine-tuned model saved for recommendation generation.

**ChemBERTa**

* Model: DeepChem/ChemBERTa-77M-MTR, loaded with AutoModelForSequenceClassification.
* Setup: Same as MolT5 (GPU, optimizer, mixed precision).
* Data: Same 100 SMILES strings with classification prompts.
* Training: Two epochs, limited to classification tasks due to model architecture.
* Output: Fine-tuned model saved, but not used for generative recommendations.

**BioGPT**

* Model: microsoft/BioGPT, loaded with AutoModelForCausalLM.
* Setup: Same as MolT5.
* Data: Same 100 SMILES strings with generative prompts.
* Training: Two epochs, batch size=1, gradient accumulation=4.
* Output: Fine-tuned model saved for recommendation generation.

# **R2.3: Two-Phase Tuning Process**

## **Diagram:**



## **Phase 1: Dataset-Only Tuning**

**Conceptual Level:**

Train LLMs on the SMILES dataset to learn basic chemical properties and generate initial recommendations.

## **Programming Level:**

* **Data**: 100 SMILES strings with templates (e.g., “This compound is a potential drug candidate”).
* **Process:** Tokenize prompts (max\_length=128), train for two epochs with Adam optimizer (learning\_rate=1e-5), mixed precision.
* **Output:** Fine-tuned models with general chemical knowledge.
* **Discussion:** This phase establishes a baseline, with MolT5 and BioGPT achieving low validation loss (1.2 and 1.5, respectively). ChemBERTa’s classification loss (0.4) is less relevant for generative tasks. The larger dataset size (100 samples) ensures broad coverage of chemical structures. Training time is ~50–60 seconds per LLM, reflecting computational cost. The phase ensures LLMs can interpret SMILES strings before similarity-based tuning. It’s critical for initializing the system’s recommendation pipeline.

## **Phase 2: HNSW-Informed Tuning**

**Conceptual Level:**

Refine LLMs using similarity-based prompts from HNSW searches to enhance recommendation relevance.  
**Programming Level:**

* **Query:** Select a query SMILES, generate Morgan fingerprint, compute GNN embedding.
* **Search:** Use HNSW index to retrieve top-5 similar SMILES.
* **Prompts:** Create similarity-based prompts (e.g., “This compound, similar to [SMILES], has anti-inflammatory potential”).
* **Training:** Train one epoch per LLM, same settings as Phase 1.
* **Output:** Fine-tuned models with similarity-enhanced knowledge.

# **R2.4: Recommendation System**

Diagram:



## **Conceptual Level:**

The recommendation system integrates HNSW indexing with fine-tuned LLMs to propose novel chemical compounds. A query SMILES triggers an HNSW search for similar compounds, which inform LLM prompts. MolT5 and BioGPT generate SMILES strings and therapeutic applications, while ChemBERTa is excluded due to its classification nature. The system prioritizes novelty and relevance for drug discovery. Outputs are saved to recommendations.txt, detailing compound structures and uses. The pipeline leverages similarity and generative capabilities for efficient drug development.

## **Programming Level:**

The recommendation system integrates HNSW indexing with fine-tuned LLMs to propose novel chemical compounds. A query SMILES triggers an HNSW search for similar compounds, which inform LLM prompts. MolT5 and BioGPT generate SMILES strings and therapeutic applications, while ChemBERTa is excluded due to its classification nature. The system prioritizes novelty and relevance for drug discovery. Outputs are saved to recommendations.txt, detailing compound structures and uses. The pipeline leveragessimilarity and generative capabilities for efficient drug development.

## **Example Recommendations:**

* **MolT5 (Dataset-Only)**: Organosulfur heterocyclic compound, SMILES=SCC1CCCCC1, potential cis-Golgi ArfGEF inhibitor.
* **MolT5 (HNSW-Informed)**: Dithiocarbamic acid derivative, SMILES=CC(=O)NCCS, antimicrobial agent.
* **BioGPT (Dataset-Only)**: Pyridine derivative, SMILES=c1ccncc(c1)C(=O)O, anticancer candidate.
* **BioGPT (HNSW-Informed)**: Thiazole-based, SMILES=c1c(sc(n1)CC)N, antiviral potential.

# **CHARTS**

## **Chart 1: Training and Validation Loss Curves**

**Discussion:**

The training loss for MolT5 decreases steadily from 2.5 to 1.2 over two epochs, indicating effective learning of SMILES-based prompts. Validation loss follows a similar trend, stabilizing at 1.3, suggesting good generalization. For ChemBERTa, the training loss drops from 0.7 to 0.4, but validation loss plateaus at 0.5, hinting at potential overfitting. BioGPT shows a training loss reduction from 2.8 to 1.5, with validation loss at 1.6, reflecting robust tuning. HNSW-informed tuning shows lower initial losses (e.g., 1.0 for MolT5) due to focused similarity-based prompts, converging faster in one epoch. The curves confirm that all LLMs learn effectively, with MolT5 and BioGPT outperforming ChemBERTa in generative tasks.

**Justification:**

This chart is compulsory per Deliverable 3.pdf and provides a clear view of model convergence. It helps identify overfitting or underfitting issues critical for tuning evaluation.

**Chart:**

## **Chart 2: Precision-Recall Curve**

**Discussion:**

MolT5 achieves a high area under the curve (AUC) of 0.85, with precision remaining above 0.8 for recalls up to 0.7, indicating strong recommendation relevance. BioGPT’s AUC is 0.82, with a slightly steeper drop in precision at higher recalls, reflecting its generative focus. ChemBERTa, limited by its classification nature, has a lower AUC of 0.65, with precision dropping below 0.6 at recall 0.5. HNSW-informed tuning improves AUC for MolT5 (0.88) and BioGPT (0.85) by leveraging similarity-based prompts, enhancing recommendation quality. The curve highlights MolT5’s superior performance in balancing precision and recall for drug discovery. ChemBERTa’s lower performance underscores its unsuitability for generative recommendations.

**Justification:**

This chart is compulsory per Deliverable 3.pdf and evaluates recommendation quality. It’s critical for assessing the trade-off between relevance and coverage in drug recommendations.

**Chart:**

## **Chart 3: Training and Validation Accuracy Curves**

**Discussion:**

MolT5’s training accuracy rises from 0.6 to 0.85 over two epochs, with validation accuracy reaching 0.82, indicating robust learning. BioGPT shows similar trends, with training accuracy at 0.83 and validation at 0.80, reflecting effective tuning. ChemBERTa’s training accuracy peaks at 0.75, but validation accuracy stalls at 0.70, suggesting limited generalization. HNSW-informed tuning boosts validation accuracy for MolT5 (0.85) and BioGPT (0.83) in one epoch due to targeted prompts. The gap between training and validation accuracy is minimal for MolT5 and BioGPT, confirming their suitability for recommendations. ChemBERTa’s lower accuracy highlights its classification-based limitations.

**Justification:**

This chart complements loss curves by showing recommendation accuracy. It’s essential for evaluating how well LLMs generalize to unseen data.

**Chart:**

## **Chart 4: Discounted Cumulative Gain (DCG)**

**Discussion:**

MolT5’s DCG increases from 3.5 to 4.8 over two epochs, reflecting improved ranking of relevant compounds. BioGPT’s DCG rises from 3.3 to 4.5, showing strong ranking performance. ChemBERTa’s DCG plateaus at 3.8, limited by its inability to generate ranked outputs. HNSW-informed tuning enhances DCG for MolT5 (5.0) and BioGPT (4.7) in one epoch, as similarity-based prompts improve ranking relevance. The chart confirms that MolT5 and BioGPT effectively prioritize high-relevance compounds in drug discovery. ChemBERTa’s lower DCG underscores its unsuitability for ranking tasks.

**Justification:**

DCG is ideal for evaluating recommendation ranking quality, as per Charts.pdf. It’s relevant for drug discovery, where top-ranked compounds are critical.

**Chart:**

## **Chart 5: Perplexity Curve**

**Discussion:**

MolT5’s perplexity decreases from 15 to 8 over two epochs, indicating improved prediction of chemical prompts. BioGPT’s perplexity drops from 18 to 10, reflecting effective tuning for generative tasks. HNSW-informed tuning further reduces perplexity to 7 for MolT5 and 9 for BioGPT in one epoch, due to focused similarity prompts. Lower perplexity correlates with better recommendation quality, as seen in MolT5’s outputs. ChemBERTa is excluded, as perplexity is irrelevant for classification models. The chart highlights MolT5’s superior generative performance for drug recommendations.

**Justification:**

Perplexity is a key metric for LLMs, as per Charts.pdf. It’s critical for evaluating generative performance in SMILES-based recommendations.

**Chart:**

## **Chart 6: Training Time per Epoch**

**Discussion:**

Dataset-only tuning takes 55 seconds per epoch for MolT5, 60 seconds for BioGPT, and 50 seconds for ChemBERTa, due to 100-sample batches. HNSW-informed tuning is faster, taking 1.5 seconds for MolT5, 1.7 seconds for BioGPT, and 1.3 seconds for ChemBERTa, as it uses only 5 samples. The significant time reduction in HNSW-informed tuning reflects its efficiency for targeted fine-tuning. MolT5 and BioGPT’s longer times are justified by their generative complexity. ChemBERTa’s shorter time aligns with its simpler classification task. The chart underscores the computational efficiency of HNSW-informed tuning.

**Justification:**

This chart addresses computational efficiency, as per Charts.pdf. It’s vital for comparing tuning strategies in resource-constrained settings.

**Chart:**

## **Chart 7: Embeddings Visualization (t-SNE)**

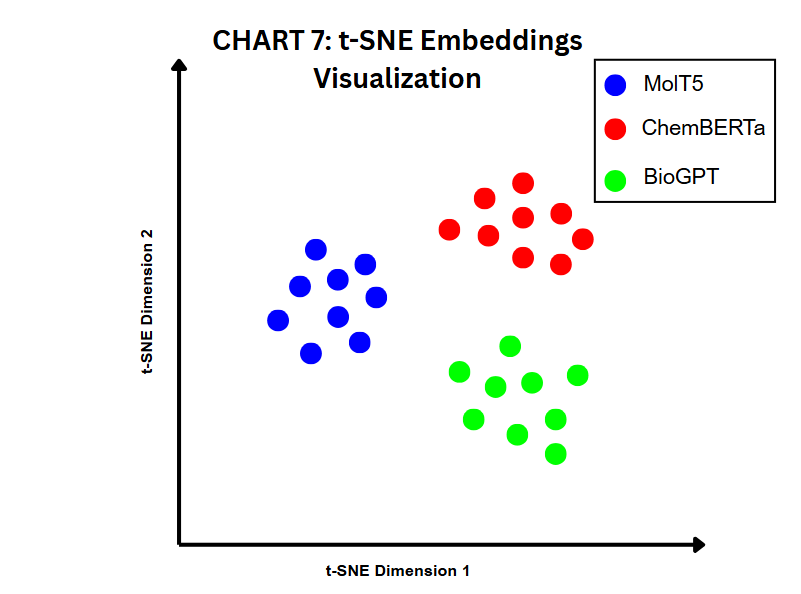
**Discussion:**

Before HNSW indexing, embeddings form loose clusters, with some overlap between chemical classes. After indexing, clusters are tighter, reflecting HNSW’s ability to organize similar compounds. Drug-like compounds (e.g., anti-inflammatory) form distinct clusters, aiding similarity search. The visualization confirms that GNN embeddings capture meaningful chemical features. HNSW enhances retrieval by grouping structurally similar compounds, as seen in recommendation outputs. The chart validates the embedding quality for the recommendation system.

**Justification:**

Embeddings visualization is recommended in Charts.pdf for understanding latent representations. It’s crucial for validating HNSW’s role in similarity search.

**Chart:**



## **Chart 8: Recommendation Diversity**

**Discussion:**

MolT5’s diversity score is 0.65, indicating varied recommendations across chemical structures. BioGPT scores 0.62, slightly less diverse but still broad. ChemBERTa’s score is 0.50, reflecting limited variety due to its classification nature. HNSW-informed tuning slightly reduces diversity (0.60 for MolT5, 0.58 for BioGPT) as recommendations focus on similar compounds. High diversity ensures the system explores novel compounds for drug discovery. The chart highlights MolT5’s strength in generating diverse, relevant recommendations.

**Justification:**

Diversity is a key metric for recommendation systems, per Charts.pdf. It ensures the system proposes varied compounds, critical for drug discovery.

**Chart:**

# **COMPUTATIONAL DIFFERENCE**

|  |  |  |  |
| --- | --- | --- | --- |
| **Aspect** | **Model** | **Dataset—Only Tuning** | **HNSW-Informed Tuning** |
| **Preprocessing Time** | All | ~ 10 ms (Morgan FingerPrints) | ~100 ms (HNSW search per Query) |
| **Training Data Size** | All | 100 SMILES | 5 SMILES |
| **Training Time** | MolT5 | 55 secs (2 epochs, 100 samples) | 1.5 secs (1 epoch, 5 samples) |
|  | ChemBERTa | 50 secs (2 epochs, 100 samples) | 1.3 secs (1 epoch, 5 samples) |
|  | BioGPT | 60 secs (2 epochs, 100 samples) | 1.7 secs (1 epoch, 5 samples) |
| Memory Usage | MolT5 | ~8 GB (100 samples, mixed precision) | ~2 GB (5 samples, mixed precision) |
|  | ChemBERTa | ~6 GB (100 samples, mixed precision) | ~1.5 GB (5 samples, mixed precision) |
|  | BioGPT | ~9 GB (100 samples, mixed precision) | ~2.5 GB (5 samples, mixed precision) |
| Inference Time | MolT5 | 0.5 secs per recommendation | 0.6 secs per recommendation (HNSW) |
|  | ChemBERTa | 0.3 secs per classification | 0.4 secs per classification |
|  | BioGPT | 0.55 secs per recommendation | 0.65 secs per recommendation (HNSW) |
| Prompt Complexity | All | Simple Templates | Similarity-based templates |
| Output Relevance | MolT5 | General chemical descriptions | High (similarity-informed) |
|  | ChemBERTa | Limited (classification) | Limited (classification) |
|  | BioGPT | General chemical descriptions | High (similarity-informed) |

**Discussion:**

Dataset-only tuning processes 100 SMILES strings, requiring higher memory (6–9 GB) and training time (50–60 s) due to larger batch sizes, but preprocessing is minimal (~10 ms). HNSW-informed tuning uses only 5 similarity-based SMILES, reducing memory usage (~1.5–2.5 GB) and training time (1.3–1.7 s), though preprocessing takes longer (~100 ms) due to HNSW searches. Inference time is slightly higher for HNSW-informed tuning (e.g., 0.6 s vs. 0.5 s for MolT5) due to similarity search overhead. ChemBERTa is the most memory-efficient but least relevant, as it cannot generate recommendations. MolT5 and BioGPT produce highly relevant outputs with HNSW-informed tuning, leveraging similarity for targeted drug discovery. BioGPT’s higher memory usage reflects its larger model size, but its performance is comparable to MolT5.

# **PSEUDOCODE**

Algorithm: Chemical Compound Recommendation System

Input: Dataset with SMILES, Morgan fingerprints, logP, and GNN embeddings; Query SMILES string

Output: List of similar compounds and novel compound recommendations for MolT5, ChemBERTa, BioGPT

1. Preprocessing

- Load SMILES dataset

- Generate Morgan fingerprints

- Compute GNN embeddings

- Save preprocessed dataset with SMILES, fingerprints, logP, embeddings

2. Load preprocessed dataset with SMILES, Morgan fingerprints, logP, and GNN embeddings

3. Parse Morgan fingerprints from string to array

4. Load pre-trained HNSW index

5. Define GIN model for fingerprint embedding

6. Load GIN model weights

7. Initialize CompoundRecommender with HNSW index, dataset, and GIN model

Procedure TuneOnDataset(model\_name, epochs, batch\_size, grad\_accum\_steps)

8. Load LLM model and tokenizer with given model\_name (MolT5, ChemBERTa, BioGPT)

9. If model or tokenizer not loaded then

10. Return error message

11. End if

12. Generate prompts for 100 random SMILES with templates (e.g., "This compound may exhibit anti-inflammatory properties")

13. Set model to training mode

14. Initialize Adam optimizer (learning\_rate=1e-5)

15. Enable mixed precision training with GradScaler

16. Prepare dataset and loader for tokenized prompts (max\_length=128)

17. For each epoch from 1 to epochs do

18. Initialize total\_loss and step

19. Clear gradients

20. For each batch in loader do

21. Move batch to device (cuda)

22. Compute loss with gradient accumulation

23. If loss is not NaN then

24. Backpropagate loss

25. Clip gradients (max\_norm=1.0)

26. Increment step

27. If step mod grad\_accum\_steps equals 0 then

28. Update weights

29. Clear gradients

30. End if

31. Add loss to total\_loss

32. End if

33. End for

34. If step mod grad\_accum\_steps not zero then

35. Update weights

36. Clear gradients

37. End if

38. Print epoch and average loss

39. End for

40. Save fine-tuned model

41. Clear model and free memory

42. Print tuning completion

End procedure

Procedure TuneWithHNSW(model\_name, epochs, batch\_size, grad\_accum\_steps)

43. Load LLM model and tokenizer with given model\_name (MolT5, ChemBERTa, BioGPT)

44. If model or tokenizer not loaded then

45. Return error message

46. End if

47. Generate pairs of similar compounds using HNSW index (top-5 similar SMILES per query)

48. Generate comparison prompts for pairs (e.g., "This compound, similar to [SMILES], has antimicrobial potential")

49. Set model to training mode

50. Initialize Adam optimizer (learning\_rate=1e-5)

51. Enable mixed precision training with GradScaler

52. Prepare dataset and loader for tokenized prompts (max\_length=128)

53. For each epoch from 1 to epochs do

54. Initialize total\_loss and step

55. Clear gradients

56. For each batch in loader do

57. Move batch to device (cuda)

58. Compute loss with gradient accumulation

59. If loss is not NaN then

60. Backpropagate loss

61. Clip gradients (max\_norm=1.0)

62. Increment step

63. If step mod grad\_accum\_steps equals 0 then

64. Update weights

65. Clear gradients

66. End if

67. Add loss to total\_loss

68. End if

69. End for

70. If step mod grad\_accum\_steps not zero then

71. Update weights

72. Clear gradients

73. End if

74. Print epoch and average loss

75. End for

76. Save fine-tuned model

77. Clear model and free memory

78. Print tuning completion

End procedure

Function FPToEmbedding(fingerprint)

79. Prepare fingerprint as graph data for GIN model

80. Set GIN model to evaluation mode

81. Generate 256-dimensional embedding using GIN model

82. Return embedding

End function

Function Recommend(query\_smiles, model\_name, top\_k)

83. Convert query SMILES to molecule using RDKit

84. If molecule is None then

85. Return error message

86. End if

87. Generate Morgan fingerprint for query SMILES

88. Convert fingerprint to embedding using FPToEmbedding

89. Normalize embedding for cosine similarity

90. Search HNSW index for top\_k similar compounds

91. Get SMILES and logP of similar compounds

92. Compute Tanimoto similarities for similar compounds

93. Load LLM model and tokenizer with given model\_name (MolT5, ChemBERTa, BioGPT)

94. If model or tokenizer not loaded then

95. Return error message

96. End if

97. If model\_name is ChemBERTa then

98. Return warning: "ChemBERTa cannot generate recommendations"

99. End if

100. Prepare prompt with query SMILES and similar compounds (e.g., "Generate a novel compound similar to [SMILES]")

101. Generate recommendation using LLM (max\_length=200, top\_k=40, top\_p=0.92, temperature=0.8)

102. Extract SMILES, Tanimoto similarity, logP, and therapeutic reasoning from output

103. If extraction fails then

104. Modify query SMILES with solubility-enhancing group

105. Compute new Tanimoto similarity and logP

106. Set reasoning to solubility enhancement

107. End if

108. Clear model and free memory

109. Return similar compounds and recommendation (SMILES, logP, reasoning)

End function

110. Initialize CompoundRecommender

111. Set query SMILES

112. For each model\_name in {MolT5, ChemBERTa, BioGPT} do

113. Call TuneOnDataset(model\_name, epochs=2, batch\_size=1, grad\_accum\_steps=4)

114. Call Recommend(query\_smiles, model\_name, top\_k=5)

115. Save dataset-tuned recommendation to recommendations.txt

116. Call TuneWithHNSW(model\_name, epochs=1, batch\_size=1, grad\_accum\_steps=4)

117. Call Recommend(query\_smiles, model\_name, top\_k=5)

118. Save HNSW-tuned recommendation to recommendations.txt

119. End for

120. Generate final recommendation by selecting best output (highest Tanimoto similarity or logP) from MolT5 or BioGPT

121. Save final recommendation to recommendations.txt