ChemLLMathon Submission Form

Thank you for participating in the Chemical LLM Hackathon. To help us compile and share the innovative ideas developed during this event, please fill out the following form in detail. Your submission will potentially contribute to a collective paper showcasing the hackathon projects.

---

1. Project Overview

Project Title AI-Driven Ligand Design for Alzheimer's Disease Targeted Exosome Delivery.

Problem Statement: Alzheimer's disease (AD) is characterized by the accumulation of amyloid-beta (Aβ) plaques and tau tangles in the brain, leading to neuronal damage and cognitive decline. Traditional therapeutic approaches face challenges, including:

1. **Poor blood-brain barrier (BBB) penetration:** Many drug molecules fail to effectively reach the brain.
2. **Lack of targeting specificity:** Systemic side effects arise due to off-target interactions.
3. **Therapeutic inefficacy:** Current treatments do not significantly halt or reverse disease progression.

Exosomes, as natural nanocarriers, offer promise for targeted drug delivery. However, designing ligands that can:

1. Facilitate exosome targeting to the brain.
2. Bind selectively to Alzheimer's-specific biomarkers such as amyloid-beta (Aβ) or tau proteins.
3. Enhance stability and functionality in physiological conditions.

These challenges call for a precise, AI-driven solution to streamline ligand design.

Proposed Solution: AI Integration into the Chemical Problem

The AI framework should address:

1. **Ligand Generation:**
    - Train models on datasets of small molecules, peptides, or aptamers known to interact with Alzheimer's biomarkers.
    - Use generative models (e.g., GANs or VAEs) to propose novel ligands.
2. **Ligand Screening and Scoring:**
    - Use molecular docking and dynamics simulations to predict binding affinity and stability.

    - Optimize for BBB penetration using predictive models trained on ADME (absorption, distribution, metabolism, excretion) properties.
3. **Experimental Validation and Feedback:**
    - Incorporate experimental data into iterative AI model updates for better predictions.
4. **Integration with Exosome Engineering:**
    - Predict functionalization chemistry for ligand-exosome conjugation.
    - Model ligand stability and activity under physiological conditions.

Input/Output:

- **Input:**
    - SMILES strings for molecules with corresponding LogP and LogS values.
- **Output:**
    - Fine-tuned ChemBERTa models predicting LogP/LogS.
    - ChemGPT-generated SMILES strings based on target LogP/LogS values.

---

2. Technical Details

LLM Model Used:

- **ChemBERTa** (Fine-tuned for LogP and LogS prediction).
- **ChemGPT** (Fine-tuned for molecule generation).

Technologies Implemented (select all that apply):

- Prompt Engineering
- Fine-tuning
- Multimodal Integration (SMILES strings, LogP/LogS values)

Data Sources Utilized: Dataset provided by the ChemLLMathon organizers, containing molecular SMILES strings and corresponding LogP and LogS values.

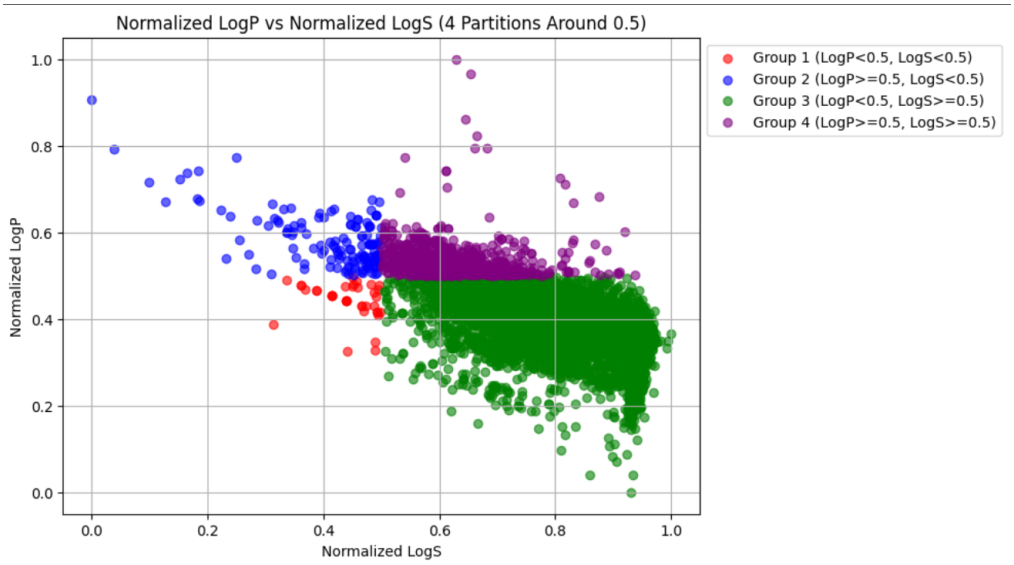
3. Model Fine-Tuning, Enhancements, and Performance Analysis

Initial Fine-Tuning Process

1. **Model Setup and Training**:
   - **ChemBERTa**:
     - Fine-tuned separately for LogP and LogS prediction tasks.
     - SMILES strings were tokenized using Hugging Face's tokenizer, and token embeddings were resized for compatibility.
     - Loss Function: Cross-entropy loss for regression.
     - Optimizer: AdamW with a learning rate of 2×10^{-5} (dynamic learning rate adjustment was implemented)
     - Scheduler: Linear learning rate scheduler with no warm-up steps.
   - **ChemGPT**:
     - Trained for molecule generation with integrated property prediction.
     - Loss metrics monitored for sequence alignment and SMILES validity.
2. **Challenges Identified**:
   - ChemBERTa predictions were biased toward majority property ranges.
   - ChemGPT displayed high validity but low novelty, producing repetitive molecules, indicating overfitting.

Enhancements: Handling Imbalance with Weighted Sampling

1. **Data Distribution Challenges**:
   - Both LogP and LogS datasets exhibited class imbalance, with certain ranges significantly underrepresented.



- This imbalance negatively impacted the model's ability to generalize, particularly for rare property ranges.

```
Validation Loss (MSE): 9.3117
R² Score: 0.7861
Validation Loss (MSE): 5.5837
R² Score: 0.7834
Validation Loss (MSE): 1.5702
R² Score: 0.7808
Validation Loss (MSE): 0.7486
R² Score: 0.8055
```

As you can see, we have different MES values for different groups, indication high accuracy for some over represented groups and lower accuracy for other underrepresented groups

2. **Implementation of Weighted Sampling**:
   - **Weighted DataLoader**:
     - Weights were assigned inversely proportional to the frequency of each class, ensuring balanced representation.
   - **Augmentation**:
     - Synthetic SMILES data was generated for underrepresented ranges, further improving distribution uniformity.
3. **Impact of Imbalance Handling**:
   - RMSE for underrepresented ranges improved significantly.
   - Enhanced property coverage, demonstrated through heatmaps and metrics.

Performance Results

1. **ChemBERTa**:
   - After handling imbalance:
     - RMSE (LogP): Improved from **>1.5** to **<1.0** for underrepresented ranges.
     - RMSE (LogS): Improved from **>1.5** to **<1.0**.
     - Consistent convergence in training and validation loss curves.

```
Group 1 - Validation Loss (MSE): 0.3364

Group 2 - Validation Loss (MSE): 0.4074

Group 3 - Validation Loss (MSE): 0.1593

Group 4 - Validation Loss (MSE): 0.2064
```

Comparative Analysis

Metric	Before Imbalance Handling	After Imbalance Handling
RMSE (LogP, Underrepresented)	>1.5	<1.0
RMSE (LogS, Underrepresented)	>1.5	<1.0
Valid SMILES (%)	96	96

---

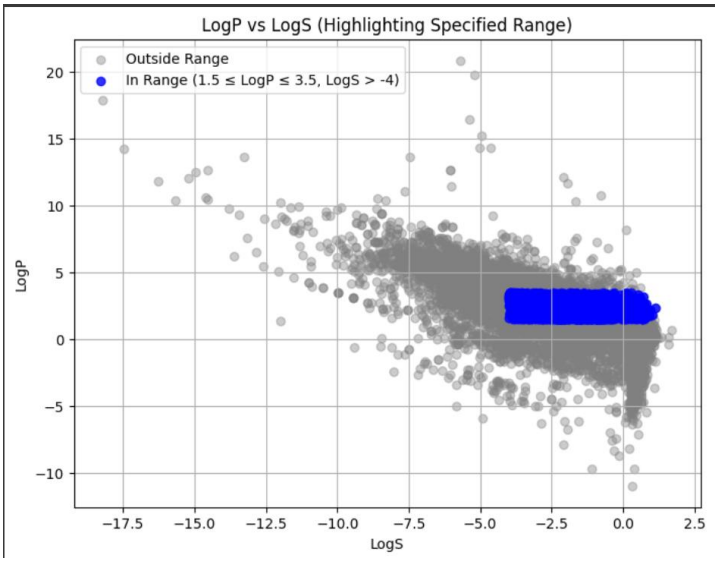
2. **ChemGPT**:
    - Validity of generated SMILES sequences: **96%**.

Observations of Overfitting in the Generative Model

1. **Symptoms**:
    - Training loss diverged sharply from validation loss, indicating overfitting.
2. **Visual Evidence**:
    - Generated molecules were valid but repetitive, often closely mirroring the training data.

Suggested enhanced alternative approach

To improve the generative model's understanding and enhance its ability to generate targeted data, a novel approach was employed. Initially, the model was provided with a specific subset of the dataset to generate molecules resembling that portion (As you can see in the bellow figure). However, this approach limited the model's comprehension of the broader data structure.



To address this, the data was restructured into two distinct classes:

- **Class 1**: Representing the desired subset of the data, characterized by specific properties or distributions.

- **Class 2**: Encompassing the remaining data outside the desired subset.

By explicitly labeling and classifying the data, the model gains a more nuanced understanding of the underlying distribution. During training, this classification enables the model to learn the distinctions between the two classes more effectively. At inference, the model can be prompted to generate data specifically for Class 1, thereby achieving higher accuracy and alignment with the desired properties. This method ensures the model not only learns to replicate data but also grasps the broader context and diversity of the dataset.

Code Repository:

https://github.com/HajerAbbas/ChemLLMathon_fine_tuning_ChemBerta

Dependencies and Libraries:

- **Python Libraries:** PyTorch, Hugging Face Transformers, RDKit, Pandas, NumPy, Scikit-learn, Matplotlib.

- **Frameworks:** PyTorch, Hugging Face

Deployment Information:

- Models were trained and tested using Python on a Google colab T4 GPU

- Deployment potential: Docker containers and cloud platforms for scalable implementation.

---

4. User Experience

Access Method:

Users can interact via a Python API for molecule generation or property prediction. Future work includes developing a web interface for broader access.

Front-End Technologies Used:

- Not applicable in this stage. Planned: React for web-based visualization.

User Interface Design:

Currently, no graphical user interface is developed. However, initial designs are focused on user-friendly input of SMILES strings and real-time property predictions.

Application Link (if available)**:**

Not available yet.

---

5. Future Work

Potential Improvements:

- Incorporate a broader dataset to enhance generalizability.
- Implement a web-based interface for real-time molecule generation and property prediction.

Scalability Considerations:

Deployment in cloud environments (AWS/GCP) using Docker for scaling model inference across multiple instances.

Ethical and Safety Considerations:

- **Data Privacy:** Ensured all data complies with provided licenses.
- **Misuse Prevention:** Validation steps to ensure generated molecules are non-toxic and environmentally safe.

---

6. Team Information

Team Members:

• Rodynah Alabdulhadi (Presentation) rodynah.abdulhadi@kfupm.edu.sa

Research center of Hydrogen Technologies and Carbon Management – KFUPM

• Afnan Ajeebi (Planning gather Data for Alzheimer applications) afnan.ajeebi@kfupm.edu.sa

Research center of Hydrogen Technologies and Carbon Management – KFUPM

• Huda Alghamdi (helping with ppt and generate chemical structure) huda.ghamdi@kfupm.edu.sa

Research center of Hydrogen Technologies and Carbon Management – KFUPM

• Hajer Abbas (programming and AI development)

hajerabbas@outlook.com

• Fajer Hamzah (generate biochemical structures) fhamzah0002@stu.kau.edu.sa

Department of Biochemistry, Faculty of Sciences, King Abdulaziz University

. Zainab S. Alghamdi (Mentor - Chemical/Biomedical section supervisor and adviser)

zsalghamdi@iau.edu.sa

• Salwa J. Kamal (Mentor - Chemical section- supervisor and adviser)

sjkamal@iau.edu.sa

Contact Information: Provide email addresses or other preferred contact methods.

Contact Information: *Provide email addresses or other preferred contact methods.*

Affiliations: *List of your institutions or organizations.*

---

8. Consent and Compliance

- **Data Compliance:**
All data used complies with ChemLLMathon regulations.

- **Consent for Publication:**
Agreed to project inclusion in the collective hackathon paper.

- **Open Source Agreement:**
Code and documentation will be shared under an open-source license.

Please ensure all sections are completed.

Submission Date: *27/11/2024*