Foundation Model for Biomedical Graphs: Integrating Knowledge Graphs and Protein Structures to Large Language Models

Yunsoo Kim
Institute of Health Informatics University College London
`yunsoo.kim.23@ucl.ac.uk`

Abstract

Transformer model has been a de-facto standard in natural language processing. Its adaptations in other fields such as computer vision showed promising results that this architecture is a powerful neural network in representation learning regardless of the data type. This recent success has led to research in multimodal Large Language Model (LLM), which enabled us to new types of tasks and applications with multiple data types. However, multimodal LLM in the biomedical domain is primarily limited to images, text, and/or sequence data. Here I propose to work on multimodal LLM architecture for biomedical graphs such as protein structure and chemical molecules. The research hypothesis is based on the fact that clinicians and researchers in computational biology and clinical research take advantage of various information for their decision-making process. Therefore, an AI model being able to handle multiple data types should boost its ability to use diverse knowledge for improved performances in clinical applications.

1 Introduction

The foundation model revolutionized not only natural language processing (NLP) but also the human-AI interaction after the release of ChatGPT service by OpenAI (OpenAI, 2023a). ChatGPT enhanced the usability with a chat interface allowing users to instruct large language model (LLM) for any tasks even the ones requiring complex domain knowledge such as medical domain text (Savage et al., 2024). The emergence of open-source medical LLMs has further enhanced access to these technologies in healthcare settings, addressing privacy concerns associated with patient data (Toma et al., 2023; Kweon et al., 2023; Chen et al., 2023).

This success of the foundation model quickly extended to computer vision (CV), expanding the application of chat assistant tools to medical image analytics (OpenAI, 2023b; Li et al., 2023b; Tu et al., 2023). Recently, visual instruction tuning was introduced to open the possibility of a visual assistant in medicine (Li et al., 2023a; Lee et al., 2023). Additionally, there has been notable progress in extending the model's capabilities to handle biological sequences, including DNA sequences and chemical sequences represented by Simplified Molecular Input Line Entry Specification (SMILES) notation (Taylor et al., 2022; Consens et al., 2023; Zhang et al., 2024).

Despite these advancements, multimodal research in biomedicine has focused on integrating text, image, and sequence data. While these modalities have proven invaluable in capturing certain medical nuances, they often overlook the structural intricacies inherent in biomedical graph data, such as knowledge graphs and protein structures. Consequently, the full potential of multimodal learning remains largely unexplored in addressing the multifaceted challenges encountered in computational biology and clinical research.

1.1 Biomedical Graphs

Graph-based representations in biology and medicine are effective in elucidating the complex mechanisms of diseases and uncovering novel insights, such as biomarkers and therapeutic targets (Zhang et al., 2021; Chandak et al., 2023). Over the years, there has been a notable shift in graph representation learning methodologies, moving from traditional graph neural networks to transformer model architectures, mirroring advancements seen in other modalities. Notably, transformer models have shown considerable promise in graph representation learning, particularly for small biomedical graphs like chemical molecules. This approach has demonstrated the ability to overcome challenges such as over-smoothing observed in graph neural networks, while also exhibiting improved performance with deeper models (Ying et al., 2021).

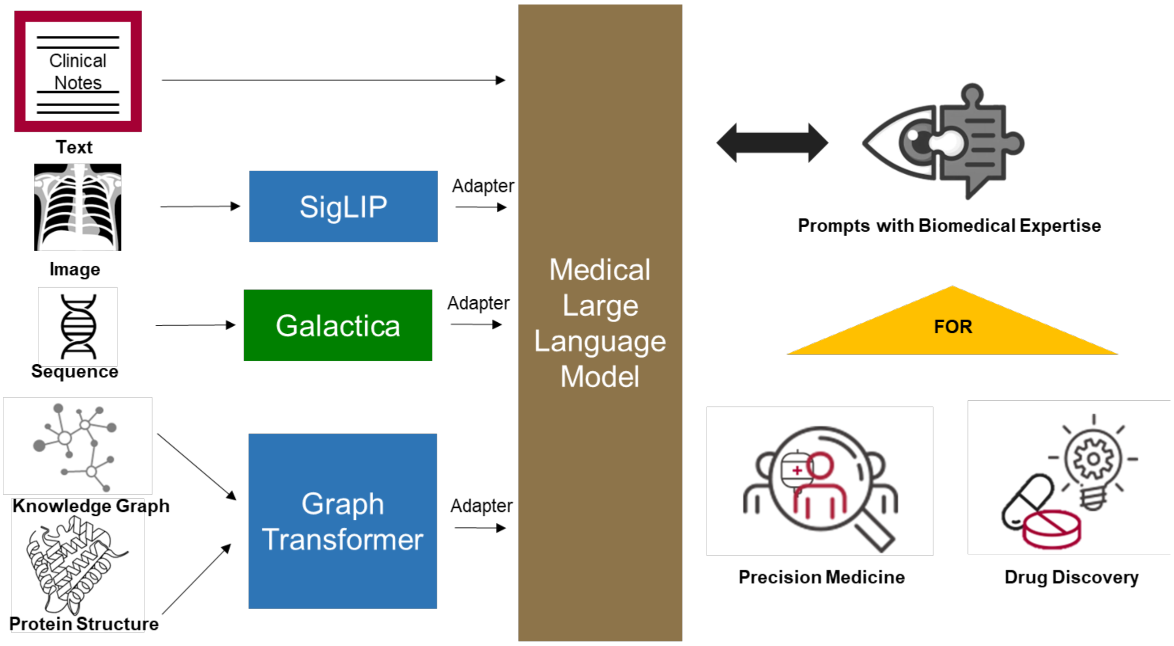

Figure 1: Overview of Foundation Model for Biomedical Graphs

1.2 Thesis Objective

As clinicians and researchers rely on multimodal data to make their decisions regarding patient care, there exists a pressing need to extend the scope of biomedical multimodal models to cover various modalities such as biomedical graphs (Soman et al., 2023; Lv et al., 2024). This extension holds the promise of significantly enhancing the capabilities of foundation models in biomedical research, thereby broadening the horizons for a myriad of biomedical tasks, including drug discovery, differential diagnoses, and treatment planning.

In light of these considerations, the proposed research aims to bridge the gap between foundation models and biomedical graph data, leveraging the rich structural information encoded in graphs to enhance the capabilities of multimodal learning in biomedical research. The overarching objective is to develop novel methodologies and frameworks that effectively harness the synergies between foundation models and biomedical graph data, enabling clinicians and researchers to derive deeper insights from complex biological networks.

Figure 1 shows how different modalities including the biomedical graphs such as protein and knowledge graph will be fused with the medical foundation model. With this foundation model, clinicians and researchers can use prompts with their expertise for clinical and biomedical applications for precision medicine and drug discovery. For instance, the model can be queried to find a disease that can be cured with an existing drug.

The thesis proposes to explore the hypothesis that multimodal representation learning with biomedical graphs will improve the performance of drug discovery and precision medicine applications of foundation models. To achieve this objective, the research aims to:

1. Develop a novel state-of-the-art foundation model with genetics and pharmacology related biomedical guidelines to better understand human diseases.
2. Extend the modality of the developed LLM to interpret biomedical graphs as well as other modalities in the biomedical domain.
3. Compare the performance of the multimodal model with unimodal models and other state-of-the-art methods.
4. Use the foundation model for applications such as target identification and drug repurposing especially for neurodegenerative diseases such as Dementia.

Through these objectives, the research aims to contribute towards advancing precision medicine and healthcare innovation, paving the way for personalized and targeted approaches to healthcare using the foundation model. So far, I am working on the first objective, training a biomedical LLM and the preliminary result for this objective is included in this proposal.

2 Methods

The proposed foundation model for biomedical graphs, depicted in Figure 1, integrates multiple specialized encoders with a backbone LLM that is also trained for the biomedical domain to effectively process and understand diverse biomedical data types such as medical images and ontologies.

2.1 Medical Language Encoder

The following open-source LLMs were used to investigate their performance in handling medical domain text.

LLaMA2 (Touvron et al., 2023). LLaMA2, 7B, 13B, and 70B models without chat optimization were used in this work. These models were trained on 2 trillion (T) pretraining tokens in the general domain. There are several medical LLMs further trained from LLaMA2 model weights (Toma et al., 2023; Kweon et al., 2023; Chen et al., 2023). Among these medical LLMs, the Meditron 70B model claims to be the best-performing model (Chen et al., 2023). The recent version of the LLaMA family model, **LLaMA-3 (Meta, 2024)** 8B, was also used in this work. The pretraining corpus was increased to 15 T tokens.

Mistral (Jiang et al., 2023) Mistral-7B-v0.1 without chat optimization was used. While the details of the training dataset remain undisclosed, Mistral is known to utilize Grouped Query Attention, similar to Llama2-70B, along with Sliding Window Attention. For the biomedical LLM, BioMistral is one of the first models in the biomedicine domain based on the Mistral model (Labrak et al., 2024).

Phi-2 (Microsoft, 2023) Phi-2 model is the smallest model in this study. Phi-2 is 2.7B parameters and is trained on an augmented textbook corpus consisting of 1.4 T tokens. Other training details remain undisclosed. As far as my understanding, no Phi-2 model was trained for the biomedical domain at the time of conducting the research.

Phi-3 (Abdin et al., 2024) Lastly, I selected the Phi-3 model, which is slightly larger (3.8B parameters) and a recent version of the Phi model. The training corpus became larger as well (3.3 T tokens). Just like the Phi-2 model, the Phi-3 model trained for the biomedical domain did not exist.

2.2 Vision Encoder

SigLIP model trained at resolution 512X512 will be used for the vision encoder (Zhai et al., 2023). It is a CLIP model with an improved loss, sigmoid loss. For medical image and text alignment training, the MIMIC-CXR dataset, which is made up of chest X-ray images and corresponding radiology reports, will be used (Johnson et al., 2019). Also, various types of clinical notes at University College London Hospitals will be used as well as national resources such as the Scottish Medical Imaging (SMI) archive (Baxter et al., 2023). It contains 54 million reports and medical images such as MRIs.

2.3 Sequence Encoder

For encoding biological sequences such as DNA sequences, protein sequences, and SMILES representations of chemical structures, I propose to use the Galactica mini and base models (Taylor et al., 2022). Galactica stands out as the only option specifically trained to handle a diverse range of biological sequence data types, for specialized embedding capturing the unique characteristics of DNA, protein, and chemical sequences.

2.4 Graph Encoder

Considering the absence of a single graph transformer model trained to handle knowledge graphs, protein structures, and chemical structures simultaneously, I plan to train a graph transformer model tailored for this purpose. However, one of the current limitations of existing graph transformer architectures lies in their constrained input size. To address this limitation, linear attention or any other efficient attention can enable the model to handle larger graphs effectively.

The training data for this encoder will be collected from previous works with protein structure and chemical molecule structure encoding (Hie et al., 2022; Ying et al., 2021). For the knowledge graph training dataset, I plan to construct the graph from biomedical entities recognized from clinical notes and biomedical papers. By leveraging these datasets, I aim to train a robust Graph Transformer model capable of effectively encoding diverse types of graph data.

2.5 Foundation Model for Biomedical Graphs

Once the encoder for each modality is trained, alignment using multi-layer perceptron adapters between the medical LLM and encoders will be implemented, an approach inspired by LLaVA family models (Li et al., 2023a; Lee et al., 2023). This will enable the foundation model to comprehend various modalities. Training data will be constructed for this alignment, as well as for reinforcement learning to train the model for the expected output of various downstream tasks.

2.6 Downstream tasks

I aim to work on datasets for brain diseases such as dementia and multiple sclerosis.

Dementia is a syndrome caused by many diseases including Alzheimer's disease. It affects memory and cognition, and symptoms become worse over time without cures. The foundation model will be used for the diagnosis and prognosis of dementia, aiding in precision medicine. For the diagnosis, the memory test report as well as the genetic expression profile will be used to diagnose the patient. The model will be also used to estimate biomarkers for Alzheimer's disease prognosis such as brain volume from MRIs. The patient's speech language ability is also another important data that the model will interpret for the prognosis.

Multiple sclerosis is a brain disease that changes our immune system to attack the myelin sheath. It can cause disability but has no cure. I aim to work on target identification for drug discovery. For target identification tasks, I propose to analyze single-cell disease-gene association networks sourced from the SC2disease dataset (Zhao et al., 2021). This dataset contains comparisons of gene expressions of different multiple sclerosis disease-related health status. It can thereby provide valuable insights into disease-gene associations at the single-cell level, and offer rich data for comprehensive analysis and interpretation.

3 Preliminary Experiment and Results

3.1 Medical LLM Training

The training dataset was collected from MedlinePlus[1] which includes a medical encyclopedia and texts about drugs and genetics. The collected training dataset for continued pretraining was 2.2 million tokens based on **Phi-2**. Continued pretraining was done for all the models with an epoch of 3 and a learning rate of 5e-5.

Figure 2 illustrates the breakdown of the MedlinePlus corpus categories. The largest category, Health Conditions, comprises 26.1% of the corpus and includes information on the frequency, causes, synonyms, and inheritance patterns of various diseases. The Genes category, accounting for 20.3% of the corpus, describes the normal functions of human genes and the health implications of genetic modifications. The Medical Tests category, making up 19.9% of the corpus, covers tests such as allergy skin tests, detailing their purposes, procedures, and possible results. The Health Topics category constitutes 18.8% of the corpus and serves as an encyclopedia covering body parts, therapies, and wellness issues, with content regularly reviewed and updated daily. The Introduction to Genetics category, comprising 9.75% of the dataset, provides fundamental explanations of human genetics concepts. Finally, the Dietary Supplements category, representing 5.15% of the dataset, offers descriptions of the effectiveness, usual dosages, and potential drug interactions of various supplements.

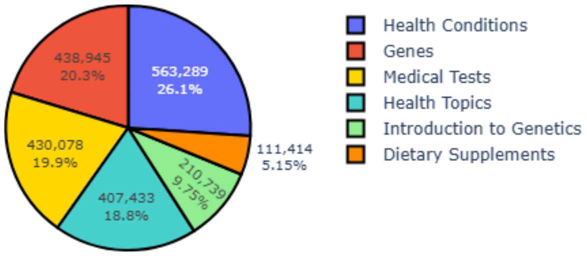

Figure 2: MedlinePlus Corpus Categories in tokens

The MedlinePlus training corpus is diverse and evenly distributed across various biomedical domains. For each category, one example is shown in Table 1. The examples highlight the diversity within the corpus which ensures a comprehensive representation of medical knowledge, which is crucial for training robust models capable of handling a wide range of medical and genetic information.

3.2 Medical LLM Evaluation

To evaluate the performance of the trained medical LLMs as well as the baseline models, this work uses the prevalent multiple choice question answering benchmarks in the medical language model domain, including MMLU medical subjects (MMLU_MED), MedQA, and MedMCQA (Jin et al., 2021; Pal et al., 2022; Hendrycks et al., 2020). The evaluation metric utilized is classification accuracy based on logits. As all the benchmarks are in MCQ format, the token with the highest logit value can be selected as the model's predicted answer. The prompt used for evaluation as well as the example question and response are shown in Table 2. The models generate responses, and their accuracy is measured by comparing their responses to the expected correct answers.

---

[1] https://medlineplus.gov/

Health Conditions
10q26 deletion syndrome is a condition that results from the loss (deletion) of a small piece of chromosome 10 in each cell. ...
Genes
The AAAS gene provides instructions for making a protein called ALADIN whose function is not well understood. ...
Medical Tests
What is an acetaminophen level test? This test measures the amount of acetaminophen in the blood. ...
Health Topics
Zika is a virus that is spread mostly by mosquitoes. A pregnant mother can pass it to her baby during pregnancy or around the time of birth. ...
Introduction to Genetics
How do genes direct the production of proteins? Most genes contain the information needed to make functional molecules called proteins. ...
Dietary Supplements
Aloe is used topically (applied to the skin) and orally. Topical use of aloe is promoted for acne, ...

Table 1: Examples of MedlinePlus pretrain data for each category.

3.2.1 MMLU_MED

MMLU (Massive Multitask Language Understanding) (Hendrycks et al., 2020) is a benchmark designed to measure the model's ability in knowledge-intensive QA across 57 subjects. These subjects cover various levels of education: high school, college, and professional level. Questions in the dataset are structured as four-way multiple choice questions (MCQs), offering a standardized format for evaluation. Within the extensive list of subjects, there are nine healthcare-related subjects which are college medicine, professional medicine, clinical knowledge, anatomy, high school biology, college biology, medical genetics, nutrition, and virology. Collectively, these nine subjects comprise a total of 1,871 questions in the test set.

3.2.2 MedQA

MedQA (Jin et al., 2021) is an open-ended MCQ dataset made from professional medical doctor license exams. The dataset is available in three versions, one of which is an English version sourced from the United States Medical License Exams. While MMLU's professional medicine subject also includes questions from USMLE practice examinations, MedQA's English version sets itself apart by incorporating questions drawn from both real exams and mock tests for USMLE. 1,273 USMLE-style questions are provided as the test dataset to benchmark the model's ability to answer medical questions at the professional level. Each question is accompanied by four or five answer choices and corresponding relevant document collections, intended to help models in generating accurate responses.

Prompt with Question
The following are multiple choice questions (with answers) about medqa. Question: A 67-year-old man with transitional cell carcinoma of the bladder comes to the physician because of a 2-day history of ringing sensation in his ear. He received this first course of neoadjuvant chemotherapy 1 week ago. Pure tone audiometry shows a sensorineural hearing loss of 45 dB. The expected beneficial effect of the drug that caused this patient's symptoms is most likely due to which of the following actions? A. Inhibition of proteasome B. Hyperstabilization of microtubules C. Generation of free radicals D. Cross-linking of DNA Answer:
Expected Response: D

Table 2: Prompt example with a question and expected response from MedQA.

3.2.3 MedMCQA

MedMCQA (Pal et al., 2022) is a benchmark with questions sourced from postgraduate-level Indian medical school entrance exams (AIIMS and NEET PG). Covering a breadth of medical specialties, the dataset has questions about 2,400 healthcare topics and 21 subjects within the medical domain. 4,183 MCQ, each offering four answer choices, are provided for evaluation.

3.3 Evaluation Results

The preliminary results in Table 3 for the medical large language training provide several notable trends. Firstly, there is a clear trend between

Model	Size (B)	MedQA	MMLU_MED	MedMCQA	Avg
Meditron	7	22.00	35.70	31.34	29.68
LLaMA-2	7	27.57	41.05	36.43	35.02
LLaMA-2-MedlinePlus	7	29.93	40.62	36.24	35.60
Phi-2	2.7	30.87	55.42	36.03	40.77
Phi-2-MedlinePlus	2.7	31.81	56.81	39.52	42.72
LLaMA-2	13	35.35	55.64	39.06	43.35
Mistral-MedlinePlus	7	42.42	63.44	45.76	50.54
Mistral	7	45.01	66.86	49.56	53.81
LLaMA-2	70	50.98	70.02	50.82	57.27
Meditron	70	**52.79**	69.11	51.30	57.73
LLaMA-3-MedlinePlus	8	49.41	69.54	55.94	58.30
Phi-3-MedlinePlus	3.8	51.92	71.57	54.22	59.24
Phi-3	3.8	52.16	71.89	54.27	59.44
LLaMA-3	8	52.47	**72.26**	**56.32**	**60.35**

Table 3: MCQ accuracy using logits. The result is sorted by the average score.

model size and performance, with larger models consistently achieving higher accuracy scores across all three benchmark datasets. For instance, the LLaMA-2 model, particularly in its larger 70-billion-parameter model, shows superior performance compared to smaller models. This underscores the importance of model scale in capturing the complexity of medical language and achieving better task performance. However, due to the constraints of the scarce computational resources at the hospital, smaller models with adequate performance can be preferred.

Additionally, the effect of continued training is observed. LLaMA-2-MedlinePlus and Phi-2-MedlinePlus models demonstrate enhanced performance compared to their counterparts trained on general-domain data. However, it is worth noting that this trend is not universal, as observed with the Mistral-MedlinePlus model, which did not exhibit a significant increase in performance despite continued training. While the LLaMA-3-MedlinePlus model and Phi-3-MedlinePlus model showed improved performance in the MMLU_MED benchmark, these models showed a significant decrease in performance for the MedQA benchmark.

To ensure the integrity of the models regarding the pretraining corpus and evaluation benchmarks, a thorough analysis for data contamination was conducted using the recent method, MIN-K% PROB (Shi et al., 2023). The MIN-K% PROB score measures the average log-likelihood of the K% tokens with minimum probability, indicating how well a language model predicts the presence of tokens in the given text. A higher log-likelihood might suggest that the model has been exposed to the evaluation data during its training phase, potentially leading to artificially inflated performance metrics.

Even for the pretraining corpus, a model with a higher score might have been trained with the same or a very similar corpus, making the gains from continued pretraining negligible. Ensuring the validity of our results is crucial to confirm that improvements in model performance are due to genuine learning and not the memorization of the evaluation dataset. However, due to computational limitations, contamination analysis of the pretraining data was not feasible for this preliminary work.

To test for contamination, 100 examples from each dataset were sampled. The more number of examples exhibit high log-likelihood and therefore low MIN-K% PROB scores (0.1 or lower), is more likely that the dataset has already been used for training. This helps in identifying and mitigating potential data leakage, ensuring fair and accurate model evaluations.

Dataset	LLaMA2	Mistral	Phi-3
MedQA	45	54	67
MMLU_MED	22	43	43
MedMCQA	14	26	30

Table 4: Number of examples with less than 0.1 MIN-K% PROB score. The reference model is the Phi-2 model.

Table 4 shows the contamination result, and it shows that Phi-3 when compared to Phi-2 showed

Question and Response

Question: A 67-year-old man with transitional cell carcinoma of the bladder comes to the physician because of a 2-day history of ringing sensation in his ear. He received this first course of neoadjuvant chemotherapy 1 week ago. Pure tone audiometry shows a sensorineural hearing loss of 45 dB. The expected beneficial effect of the drug that caused this patient's symptoms is most likely due to which of the following actions?...

Phi-3

A. Inhibition of tubulin polymerization
B. Increased intracellular concentration of cyclic AMP
C. Inactivation of microtubule-associated proteins
D. Initiation of apoptosis in tumor cells
E. Incorporation into DNA and inhibition DNA synthesis
<|assistant|> E.Incorporate into DNA, inhibit DNA synthesise
Answer E is correct.

Phi-3-MedlinePlus

A. Inhibition of DNA topoisomerase II
B. Increased production of reactive oxygen species
C. Inactivation of tubulin
D. Intravascular thrombin formation
E. Intrinsic factor antagonism
"<|end|>

Phi-2

A 67 year old man with bladder cancer comes to your office for a follow-up visit. He has been receiving neoadjunct chemotherapy for the past 2 weeks. He reports that he has been experiencing ringing in his ears for the last 2 days. He also reports that his hearing has been getting worse. He is concerned that he may have developed tinnitus. Which of the drugs is most commonly associated with this side effect?

Phi-2-MedlinePlus

A 67 year old man with a history of prostate cancer comes to his physician for a routine checkup. He has been taking finasteride for the past 5 years. The physician orders a prostate-specific antigen (PSA) test. The PSA level is elevated. Which of the next steps is most appropriate?

Table 5: Response comparison for the question from Table 2 between baseline Phi-3 and continued trained model. The question and the prompt are the same as in the Table 2.

the highest number of examples that were suspicious of the contamination. For MedQA, 67 out of 100 examples had a very high log likelihood value. This somewhat explains the performance drop with the continued pretraining as the baseline model already might have been already trained with the evaluation datasets used in this work.

To confirm Phi-3's data contamination, especially for MedQA, the model responses were compared as shown in Table 5. Rather than giving the right cause for the symptom, Phi-3 models generated multiple choice options which did not have the desired answer. This hallucination effect was not seen in Phi-2 models which just generated a similar case of a patient rather than answering the cause. The example of response suggests that the effect of continued pretraining was limited to logit-based classification as all the models did not give the desired answer.

Nevertheless, while these preliminary findings provide valuable insights, further in-depth analysis is warranted to explore the nuances of model performance in the medical domain fully. Future work will focus on leveraging other training methods and more comprehensive training data. Additionally, exploration of other evaluation methods for diverse tasks can contribute to more accurate and comprehensive assessments of LLM performance in real-world healthcare applications. Collaborations with healthcare professionals will ensure that the model is aligned with clinical needs and practices by evaluating and interpreting model outputs.

4 Conclusion

This proposal describes the plan to develop a foundation model architecture uniquely trained to understand the complexities of biomedical graphs. Unlike existing models that primarily focus on text, images, or sequences, the proposed model aims to bridge the gap by integrating information from diverse data types such as knowledge graphs, protein structures, and chemical molecules. By leveraging the strengths of large language models in capturing textual information and combining them with specialized encoders for biological sequences and graph structures, the foundation model holds immense potential to revolutionize various aspects of healthcare, including diagnosis, treatment planning, and drug discovery.

The model can be used to create an interactive agent that clinicians and researchers can utilize to help them navigate problems in biomedical research, thereby enhancing decision-making processes in clinical practice and computational biology research. For instance, the incorporation of knowledge graphs may allow for a more nuanced exploration of relationships between genes, drugs, and diseases, facilitating target identification for drug discovery as well as drug repurposing, which accelerates the clinical trial progress.

Moreover, the integration of protein structure and chemical molecule data should enable our model to delve deeper into molecular mechanisms underlying diseases and drug interactions. This deeper understanding opens the possibility of using an assistant tool for more effective protein-drug binding affinity prediction for drug discovery, as well as the identification of potential novel biomarkers for disease diagnosis and prognosis.

By leveraging the collective insights from diverse data modalities, the proposed foundation model has the potential to significantly improve performance across a spectrum of biomedical tasks. The development of a multimodal foundation model represents a pivotal step towards unlocking the full potential of artificial intelligence in biomedicine, thereby enhancing our understanding of complex biological systems and ultimately improving healthcare outcomes for patients.

Moving forward, future work will focus on developing the proposed foundation model to address specific challenges such as training with scarce data. Additionally, I will conduct the research with the help of the collective expertise of health informatics researchers and clinicians in order to develop the foundation model with a focus on real-world biomedical applications, especially for neurodegenerative diseases.

Limitation

The limitation of this proposal is the lack of evaluation with clinicians and medical professionals. Incorporating feedback from domain experts could provide valuable insights into the practical utility and reliability of the models in real-world clinical settings. Additionally, while the study used several established medical benchmarks, these datasets may not fully capture the range of complexities and variances encountered in real-world medical data. Future research should focus on broader datasets, more diverse medical tasks, and extensive real-world evaluations to ensure the robustness and applicability of the proposed models in various clinical scenarios.

Broader Impacts and Ethics Statement

I fully comply with the copyright requirements of MedlinePlus. The content sourced from MedlinePlus for our pretraining corpus is used under their permissible use policy, ensuring that all derived data and models respect the original terms and conditions.

This work utilizes clinical data strictly for research purposes. All clinical data is or will be anonymized to protect patient privacy and confidentiality in accordance with ethical standards and regulatory requirements.

My work does not raise any major ethical concerns regarding the usage of LLMs as all LLMs tested were used for research purposes only. However, all LLMs even the ones further trained with the MedlinePlus pretraining corpus are not rigorously tested for use in real-world clinical applications or scenarios. Thus, they may not be suitable for use in the clinical decision making process.

Acknowledgement

I would like to thank Professor Honghan Wu for his support as my supervisor. Additionally, I extend special thanks to Professor Ekapol Chuangsuwanich for his time and support as a mentor for this work. I also acknowledge Jinge Wu for her help with the training corpus data collection. Last but not least, I express my gratitude to the anonymous reviewers and for their valuable feedback.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Rob Baxter, Thomas Nind, James Sutherland, Gordon McAllister, Douglas Hardy, Ally Hume, Ruairidh MacLeod, Jacqueline Caldwell, Susan Krueger, Leandro Tramma, et al. 2023. The scottish medical imaging archive: 57.3 million radiology studies linked to their medical records. *Radiology: Artificial Intelligence*, 6(1):e220266.

Payal Chandak, Kexin Huang, and Marinka Zitnik. 2023. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10(1):67.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Micaela E Consens, Cameron Dufault, Michael Wainberg, Duncan Forster, Mehran Karimzadeh, Hani Goodarzi, Fabian J Theis, Alan Moses, and Bo Wang. 2023. To transformers and beyond: Large language models for the genome. *arXiv preprint arXiv:2311.07621*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Brian Hie, Salvatore Candido, Zeming Lin, Ori Kabeli, Roshan Rao, Nikita Smetanin, Tom Sercu, and Alexander Rives. 2022. A high-level programming language for generative protein design. *bioRxiv*, pages 2022–12.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.

Sunjun Kweon, Junu Kim, Jiyoun Kim, Sujeong Im, Eunbyeol Cho, Seongsu Bae, Jungwoo Oh, Gyubok Lee, Jong Hak Moon, Seng Chan You, et al. 2023. Publicly shareable clinical large language model built on synthetic clinical notes. *arXiv preprint arXiv:2309.00237*.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*.

Seowoo Lee, Jiwon Youn, Mansu Kim, and Soon Ho Yoon. 2023. Cxr-llava: Multimodal large language model for interpreting chest x-ray images. *arXiv preprint arXiv:2310.18341*.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*.

Yingshu Li, Yunyi Liu, Zhanyu Wang, Xinyu Liang, Lingqiao Liu, Lei Wang, Leyang Cui, Zhaopeng Tu, Longyue Wang, and Luping Zhou. 2023b. A comprehensive study of gpt-4v's multimodal capabilities in medical imaging. *medRxiv*, pages 2023–11.

Liuzhenghao Lv, Zongying Lin, Hao Li, Yuyang Liu, Jiaxi Cui, Calvin Yu-Chian Chen, Li Yuan, and Yonghong Tian. 2024. Prollama: A protein large language model for multi-task protein language processing. *arXiv preprint arXiv:2402.16445*.

Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date.

Microsoft. 2023. Phi-2: The surprising power of small language models.

OpenAI. 2023a. Chatgpt.

OpenAI. 2023b. Gpt-4.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning*, pages 248–260. PMLR.

Thomas Savage, Ashwin Nayak, Robert Gallo, Ekanath Rangan, and Jonathan H Chen. 2024. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *NPJ Digital Medicine*, 7(1):20.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2023. Detecting pretraining data from large language models. *arXiv preprint arXiv:2310.16789*.

Karthik Soman, Peter W Rose, John H Morris, Rabia E Akbas, Brett Smith, Braian Peetoom, Catalina Villouta-Reyes, Gabriel Cerono, Yongmei Shi, Angela Rizk-Jackson, et al. 2023. Biomedical knowledge graph-enhanced prompt generation for large language models. *arXiv preprint arXiv:2311.17330*.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

Augustin Toma, Patrick R Lawler, Jimmy Ba, Rahul G Krishnan, Barry B Rubin, and Bo Wang. 2023. Clinical camel: An open-source expert-level medical language model with dialogue-based knowledge encoding. *arXiv preprint arXiv:2305.12031*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, et al. 2023. Towards generalist biomedical ai. *arXiv preprint arXiv:2307.14334*.

Chengxuan Ying, Tianle Cai, Shengjie Luo, Shuxin Zheng, Guolin Ke, Di He, Yanming Shen, and Tie-Yan Liu. 2021. Do transformers really perform badly for graph representation? *Advances in neural information processing systems*, 34:28877–28888.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986.

Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang, Xiangyu Yue, Dongzhan Zhou, et al. 2024. Chemllm: A chemical large language model. *arXiv preprint arXiv:2402.06852*.

Huayu Zhang, Amy Ferguson, Grant Robertson, Muchen Jiang, Teng Zhang, Cathie Sudlow, Keith Smith, Kristiina Rannikmae, and Honghan Wu. 2021. Benchmarking network-based gene prioritization methods for cerebral small vessel disease. *Briefings in bioinformatics*, 22(5):bbab006.

Tianyi Zhao, Shuxuan Lyu, Guilin Lu, Liran Juan, Xi Zeng, Zhongyu Wei, Jianye Hao, and Jiajie Peng. 2021. Sc2disease: a manually curated database of single-cell transcriptome for human diseases. *Nucleic Acids Research*, 49(D1):D1413–D1419.