Evaluating Large Language Models for Predicting Protein Behavior under Radiation Exposure and Disease Conditions

Ryan Engel
Stony Brook University
ryengel@cs.stonybrook.edu

Gilchan Park
Brookhaven National Laboratory
gpark@bnl.gov

Abstract

The primary concern with exposure to ionizing radiation is the risk of developing diseases. While high doses of radiation can cause immediate damage leading to cancer, the effects of low-dose radiation (LDR) are less clear and more controversial. To further investigate this, it necessitates focusing on the underlying biological structures affected by radiation. Recent work has shown that Large Language Models (LLMs) can effectively predict protein structures and other biological properties. The aim of this research is to utilize open-source LLMs, such as Mistral, Llama 2, and Llama 3, to predict both radiation-induced alterations in proteins and the dynamics of protein-protein interactions (PPIs) within the presence of specific diseases. We show that fine-tuning these models yields state-of-the-art performance for predicting protein interactions in the context of neurodegenerative diseases, metabolic disorders, and cancer. Our findings contribute to the ongoing efforts to understand the complex relationships between radiation exposure and disease mechanisms, illustrating the nuanced capabilities and limitations of current computational models. The code and data are available at: https://github.com/Rengel2001/SURP_2024

1 Introduction

The exploration of the biological consequences of ionizing radiation on human health has long been a focal point of medical and environmental research. High doses of radiation are linked to immediate cellular damage and an increased risk of cancer (Wang et al., 2018). However, the implications of low-dose radiation (LDR) exposure remain a topic of significant debate. Emerging evidence suggests potential associations with various non-cancerous diseases, including neurodegenerative and cardiovascular diseases (Sharma et al., 2018; Kamiya et al., 2015). Additionally, others show that cancer

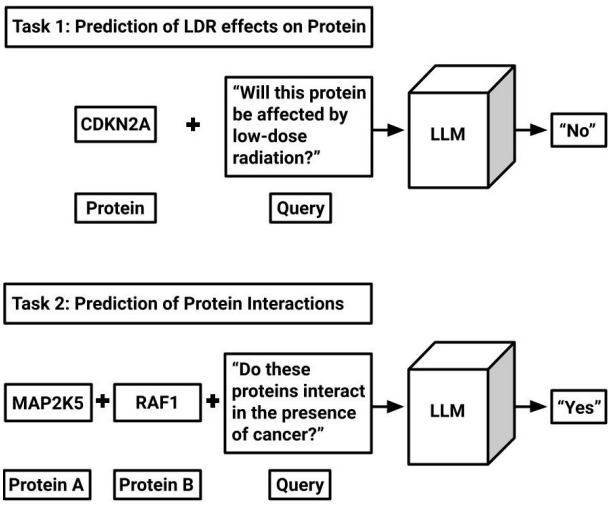

Figure 1: Tasks Utilizing LLMs for Protein Behavior Prediction.

is a result of low-dose radiation exposure (Shah et al., 2014; Hauptmann et al., 2020).

Understanding these effects at the molecular level, particularly in relation to protein structure and function, is crucial for developing protective measures. Similarly, protein-protein interactions (PPIs) are vital for various cellular processes and play a critical role in understanding disease mechanisms. Furthermore, there exists extensive PPI data, compiled into comprehensive public databases like BioGRID (Oughtred et al., 2021), STRING (Alanis-Lobato et al., 2016), HIPPIE (Szklarczyk et al., 2021), and Kegg (Kanehisa et al., 2017). Considerable research has been dedicated to understanding general protein interactions; however, there is a lack of studies examining protein interaction networks in the context of specific diseases.

The overarching goal of this research is to determine the efficacy of LLMs in accurately predicting complex biological processes related to protein function under various conditions. We employ three state-of-the-art LLMs, to analyze data from six diverse datasets. These datasets represent

two distinct categories. The first focuses on the effects of LDR on proteins, and the second highlights the PPI network present within specific diseases. We formalize this data into two binary classification tasks, which are illustrated in Figure 1. This approach not only demonstrates the versatility of LLMs in biological research but also paves the way for novel insights into the molecular dynamics influenced by radiation exposure and disease processes. Our contributions in this paper include:

1. Organizing 6 key datasets, which are then split into 13 subsets, each designed to emphasize different experimental conditions.

2. Conducting a comprehensive evaluation of three open-source LLMs, comparing the performance of pre-trained models with the fine-tuned models.

3. Investigating the level of knowledge that LLMs have regarding protein behaviors and reviewing their current limitations for these tasks.

4. Analyzing the proteins that occur in both the LDR datasets and the PPI datasets, to highlight which proteins in each network are significantly deregulated by radiation exposure.

2 Related Works

2.1 Low-Dose Radiation Research

There has been a great deal of research focused on the effects of radiation on biological systems. Many studies exploring the field use traditional methods and there has been significant progress (Khan and Wang, 2022; Tatjana Paunesku and Woloschak, 2021; Ji et al., 2019). However, the application of machine learning to these studies has been limited. Notably, one approach employed artificial neural networks (ANNs) within the Rosetta suite to predict protein post-translational modifications (PTMs) relevant to radiation-induced effects (Ertelt et al., 2024). Another study used machine learning to identify potential methionine oxidation sites, a modification also associated with oxidative stress from radiation (Aledo et al., 2017). These instances showcase the emerging intersection of computational power with radiation biology research.

2.2 PPI Prediction Methods

The abundance of PPI data has prompted significant advancements in molecular biology research. Recently, computational techniques employing machine learning and graph embeddings have been developed for PPI prediction. One approach employs Graph-BERT, ProtBERT, and SeqVec models within a PPI network graph, showcasing the efficacy of language models (Jha et al., 2023). Another emerging trend is the use of Convolutional Neural Networks (CNNs), with studies employing Bio2Vec coupled with CNNs to predict PPIs from sequences (Wang et al., 2019; Hashemifar et al., 2018). The PIPR method simplifies PPI prediction by using sequence data alone, surpassing many traditional models in both basic and complex PPI tasks (Chen et al., 2019). While many methods focus on general PPI prediction, the NECARE model (Qiu et al., 2021) excels at predicting cancer-associated PPIs using a deep learning framework with a Relational Graph Convolutional Network (R-GCN). Similarly, the symmetric logistic matrix factorization (symLMF) approach (Pei et al., 2021) accurately predicts PPIs, including those involved in neurodegenerative and metabolic disorders, outperforming most classifiers.

2.3 Language Models for Molecular Biology

Concurrently, advancements in computational biology have also leveraged language models and the transformer architecture (Vaswani et al., 2017) to achieve significant breakthroughs in biomolecular and proteomics research. At the forefront, AlphaFold (Jumper et al., 2021) has set a precedent by employing innovative deep learning techniques to predict protein structures with remarkable accuracy. Building upon these foundations, protein language models like ProGen2 (Nijkamp et al., 2022), ProGPT2 (Ferruz et al., 2022), and ProLlama (Lv et al., 2024), have further developed the applications of language modelling for proteomics. Additionally, this has led to advancements in general purpose biological language models like BioGPT (Luo et al., 2022), and BioMedLM (Bolton et al., 2024).

2.4 General Purpose LLMs

Large-scale language models like Llama (Touvron et al., 2023a), and its subsequent iterations including Llama 2 (Touvron et al., 2023b), Llama 3 (AI@Meta, 2024) and Alpaca (Taori et al., 2023), have highlighted the importance of data design and task-specific training in improving model performance across a variety of tasks. Additionally, the creation of the Mistral (Jiang et al., 2023) model

helps to bring open-source LLMs to the forefront of scientific innovation. These strides in LLM research have brought significant advancements in other scientific disciplines (Zhang et al., 2024). We aim to utilize such LLMs to further advance research on LDR exposure and to analyze how this might affect protein networks and specific diseases.

3 LLMs and Datasets

In this study, we employ three open-source LLMs, Mistral (7B), Llama 2 (7B), and Llama 3 (8B), to investigate two primary areas of biological research: the effects of low-dose radiation (LDR) on proteins and the dynamics of PPIs in the context of specific diseases. These models were chosen because of their state-of-the-art performance in many natural language processing (NLP) tasks. Additionally, their open-source nature allows for broad accessibility and modification by researchers across disciplines, which promotes transparency and collaborative advancements in both NLP and other scientific domains.

To facilitate a comprehensive analysis, our methodology encompasses six core datasets, which are further subdivided into 13 distinct subsets based on specific experimental parameters and objectives. The first 3 datasets primarily explore the effects of LDR on protein deregulation. These 3 sets are further divided into 10 subsets, emphasizing different experimental conditions. The last 3 datasets focus on PPIs in the presence of specific diseases, namely neurodegenerative, metabolic, and cancer.

The subsets of the LDR data are much smaller than the PPI datasets, which is why these were combined into dataset 3c. Dataset 3c's larger size is shown in comparison with the other datasets in Figure 2. The details about each dataset is outlined in Appendix A.

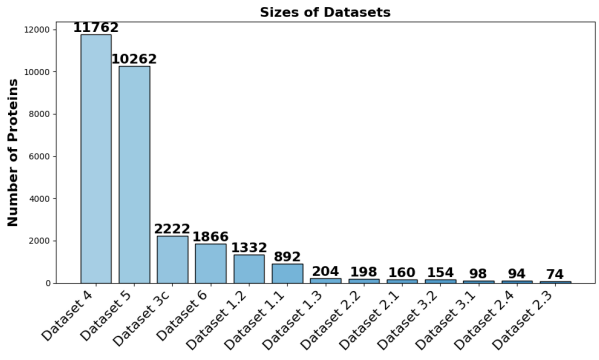

Figure 2: Comparison of Dataset Sizes

4 Experiments

The methodology for analyzing each dataset began with data pre-processing, executed through a Python script tailored to appropriately structure the raw data. Subsequently, this processed data was used to create prompts that fit the prompting strategies outlined in Appendix A. These prompts were then saved in a JSON file, and were subsequently used as input to the LLMs.

For deploying the models, a separate Python script using the Hugging Face Transformers library loaded the models onto 4×NVIDIA A100 80GB GPUs. These pre-trained models were then presented with the JSON file prompts and the performance of each model was recorded.

4.1 Experimental Setup

Our experimental setup across the datasets implemented a binary classification task, instructing the models to produce a "yes" or "no" answer in response to each prompt. The generated responses from each model necessitate the deployment of an algorithm to parse these outputs effectively. If the given string "yes" or "no" is not found in the model's response, this response is marked as the opposite of the true label. This is a result of using causal language models, which are designed for text generation. To optimize this task, the "Data Collator for Completion-Only Language Models" and the SFT (Supervised fine-tuning) Trainer from the Hugging Face library were utilized in training the models to give the correct response structure.

4.2 Data Split

We structured the training process differently for each of the two tasks. For the LDR task, we divided the prompts for each dataset into an 80/10/10 split for training, validation, and testing, respectively. The PPI datasets 4 and 5 utilized a 5-fold cross validation setup, where 4 sets were used for training and 1 set was used for testing in each fold. Similarly, the PPI dataset 6 used a 5-fold cross validation setup but instead 3 sets were used for training, 1 set for validation, and 1 set for testing. This was carried out to replicate the experimental conditions used in the benchmark models.

4.3 Fine-Tuning

During the training process, we employed Parameter Efficient Fine-Tuning (PEFT) (Mangrulkar et al., 2022), a method focused on selectively modi-

fying a subset of the model's parameters rather than the entire set. Low-Rank Adaptation (LoRA) (Hu et al., 2021) is a specialized PEFT technique that was utilized when fine-tuning the LLMs for these tasks. Additionally, we used QLoRA (Dettmers et al., 2023) to reduce the GPU memory required for training Llama 3 on datasets 4 and 5. This approach was essential because the combined size of these datasets and the 8 billion parameter model required more efficient memory usage than traditional LoRA.

5 Results

Every phase of the model training process was documented and analyzed. The evaluation metrics used include accuracy, Matthews Correlation Coefficient (MCC), specificity, macro precision, and macro F1 Score.

Model	Acc.	MCC	Spec.	Prec.	F1
Mistral (3-shot)	0.367	-0.343	0.068	0.290	0.304
Llama 2 (3-shot)	0.556	0.110	0.386	0.558	0.541
Llama 3 (3-shot)	0.489	0.0	1.0	0.244	0.328
Mistral (LoRA)	0.500	0.058	**0.977**	0.580	0.369
Llama 2 (LoRA)	0.522	0.061	0.750	0.534	0.500
Llama 3 (LoRA)	**0.567**	**0.155**	0.773	**0.585**	**0.551**

Table 1: Performance Comparison for Dataset 1.1

Model	Acc.	MCC	Spec.	Prec.	F1
Mistral (3-shot)	0.291	-0.461	0.027	0.210	0.239
Llama 2 (3-shot)	**0.567**	**0.153**	0.479	**0.578**	**0.566**
Llama 3 (3-shot)	0.545	0.0	**1.0**	0.272	0.353
Mistral (LoRA)	0.493	0.007	0.384	0.503	0.490
Llama 2 (LoRA)	0.537	-0.006	0.932	0.494	0.401
Llama 3 (LoRA)	0.552	0.054	0.945	0.554	0.420

Table 2: Performance Comparison for Dataset 1.2

Model	Acc.	MCC	Spec.	Prec.	F1
Mistral (3-shot)	0.286	-0.230	0.133	0.368	0.279
Llama 2 (3-shot)	0.381	-0.067	0.267	0.467	0.381
Llama 3 (3-shot)	**0.714**	0.0	**1.0**	0.357	0.417
Mistral (LoRA)	0.381	-0.241	0.400	0.391	0.358
Llama 2 (LoRA)	0.381	-0.447	**0.533**	0.286	0.276
Llama 3 (LoRA)	0.571	**0.279**	0.467	**0.630**	**0.568**

Table 3: Performance Comparison for Dataset 1.3

Model	Acc.	MCC	Spec.	Prec.	F1
Mistral (3-shot)	0.125	-0.745	0.0	0.083	0.111
Llama 2 (3-shot)	0.438	-0.035	0.3	0.482	0.435
Llama 3 (3-shot)	0.625	0.0	**1.0**	0.313	0.385
Mistral (LoRA)	0.688	0.313	0.8	0.664	0.654
Llama 2 (LoRA)	0.688	0.423	0.6	0.706	0.686
Llama 3 (LoRA)	**0.813**	**0.592**	**0.9**	**0.809**	**0.792**

Table 4: Performance Comparison for Dataset 2.1

Model	Acc.	MCC	Spec.	Prec.	F1
Mistral (3-shot)	0.095	-0.767	0.0	0.059	0.087
Llama 2 (3-shot)	0.524	**0.224**	0.4	**0.607**	**0.523**
Llama 3 (3-shot)	**0.714**	0.0	**1.0**	0.357	0.417
Mistral (LoRA)	**0.714**	0.0	**1.0**	0.357	0.417
Llama 2 (LoRA)	0.286	0.0	0.0	0.143	0.222
Llama 3 (LoRA)	0.524	-0.167	0.667	0.417	0.417

Table 5: Performance Comparison for Dataset 2.2

Model	Acc.	MCC	Spec.	Prec.	F1
Mistral (3-shot)	0.25	-0.5	0.25	0.25	0.25
Llama 2 (3-shot)	0.375	-0.378	0.0	0.214	0.273
Llama 3 (3-shot)	0.5	0	**1.0**	0.25	0.333
Mistral (LoRA)	**0.625**	**0.258**	0.5	**0.633**	**0.619**
Llama 2 (LoRA)	**0.625**	**0.258**	0.75	**0.633**	**0.619**
Llama 3 (LoRA)	0.5	0	**1.0**	0.25	0.333

Table 6: Performance Comparison for Dataset 2.3

Model	Acc.	MCC	Spec.	Prec.	F1
Mistral (3-shot)	0.1	-0.816	0.167	0.1	0.091
Llama 2 (3-shot)	0.4	-0.102	0.167	**0.438**	**0.375**
Llama 3 (3-shot)	**0.6**	**0.0**	**1.0**	0.3	**0.375**
Mistral (LoRA)	**0.6**	**0.0**	**1.0**	0.3	**0.375**
Llama 2 (LoRA)	0.4	**0.0**	0.0	0.2	0.286
Llama 3 (LoRA)	0.4	**0.0**	0.0	0.2	0.286

Table 7: Performance Comparison for Dataset 2.4

Model	Acc.	MCC	Spec.	Prec.	F1
Mistral (3-shot)	0.364	**0.0**	0.0	0.182	0.267
Llama 2 (3-shot)	0.364	-0.463	0.571	0.25	0.267
Llama 3 (3-shot)	**0.636**	**0.0**	**1.0**	**0.318**	**0.389**
Mistral (LoRA)	0.364	**0.0**	0.0	0.182	0.267
Llama 2 (LoRA)	0.364	**0.0**	0.0	0.182	0.267
Llama 3 (LoRA)	0.273	-0.418	0.0	0.15	0.214

Table 8: Performance Comparison for Dataset 3.1

Model	Acc.	MCC	Spec.	Prec.	F1
Mistral (3-shot)	0.438	0.0	0.0	0.219	0.304
Llama 2 (3-shot)	0.438	-0.098	0.333	0.450	0.435
Llama 3 (3-shot)	**0.625**	**0.293**	**1.0**	**0.8**	**0.5**
Mistral (LoRA)	0.563	0.0	**1.0**	0.281	0.360
Llama 2 (LoRA)	0.438	0.0	0.0	0.219	0.304
Llama 3 (LoRA)	0.563	0.0	**1.0**	0.281	0.360

Table 9: Performance Comparison for Dataset 3.2

Model	Acc.	MCC	Spec.	Prec.	F1
Mistral (3-shot)	0.247	-0.593	0.0	0.173	0.198
Llama 2 (3-shot)	0.475	-0.015	**0.769**	0.491	0.443
Llama 3 (3-shot)	0.493	-0.038	0.317	0.480	0.473
Mistral (LoRA)	**0.552**	**0.090**	0.423	0.546	**0.540**
Llama 2 (LoRA)	0.547	0.066	0.154	**0.549**	0.459
Llama 3 (LoRA)	0.516	0.014	0.375	0.507	0.502

Table 10: Performance Comparison for Dataset 3c

Tables 1-13 indicate the performance of both the pre-trained models, and their fine-tuned counterparts on each of the 13 datasets. We evaluated the pre-trained models using the same procedure as the fine-tuned models, the only difference is that the

Model	Acc. (%)	MCC (%)	Spec. (%)	Prec.(%)	F1 (%)
Mistral (3-shot)	38.44±0.46	-31.79±0.54	4.15±0.16	28.16±0.28	30.23±0.25
Llama 2 (3-shot)	55.14±0.16	13.18±0.29	23.79±0.57	58.47±0.23	50.23±0.17
Llama 3 (3-shot)	50.38±0.36	6.17±0.27	**1.0±0.0**	75.10±0.18	34.18±0.17
Mistral (LoRA)	62.34±6.88	25.53±14.37	97.89±1.17	48.49±12.84	51.97±10.36
Llama 2 (LoRA)	87.28±0.41	76.63±1.03	88.59±0.95	87.33±0.22	87.28±0.31
Llama 3 (QLoRA)	**88.27±1.08**	**76.92±2.12**	92.81±1.06	**88.58±1.08**	**88.26±1.07**
SymLMF (Reported)	86.11±1.05	74.29±2.07	N/A	83.24±1.28	N/A

Table 11: Performance Comparison for Dataset 4

Model	Acc.	MCC	Spec.	Prec.	F1
Mistral (3-shot)	35.52±0.76	-39.08±0.82	1.98±0.14	23.65±0.56	27.33±0.48
Llama 2 (3-shot)	51.45±0.81	6.66±1.16	6.42±0.37	57.66±1.31	39.09±0.57
Llama 3 (3-shot)	56.03±0.21	12.12±0.46	53.69±0.34	56.06±0.23	56.00±0.21
Mistral (LoRA)	62.18±6.61	25.08±13.33	**93.80±3.89**	57.70±11.91	51.93±10.16
Llama 2 (LoRA)	84.63±0.24	69.20±0.66	84.06±0.92	84.51±0.34	84.43±0.34
Llama 3 (QLoRA)	**91.28±0.87**	**82.57±1.73**	90.41±1.08	**91.29±.86**	**91.28±0.87**
SymLMF (Reported)	81.37±1.04	63.31±2.07	N/A	77.70±1.07	N/A

Table 12: Performance Comparison for Dataset 5

Model	Acc.	MCC	Spec.	Prec.	F1
Mistral (3-shot)	41.91±1.19	-23.81±1.55	5.15±0.43	32.45±0.95	32.79±0.69
Llama 2 (3-shot)	57.61±0.91	16.69±2.40	39.18±1.42	59.0±1.32	56.09±1.0
Llama 3 (3-shot)	53.96±1.62	16.40±2.75	**97.83±0.55**	67.25±2.83	42.91±1.68
Mistral (LoRA)	83.76±8.12	68.81±15.40	93.30±2.15	79.20±12.41	80.69±10.85
Llama 2 (LoRA)	93.25±0.84	86.84±1.49	93.39±1.38	93.55±0.74	93.22±0.84
Llama 3 (LoRA)	**93.94±0.25**	**88.04±0.47**	91.83±1.12	**94.09±0.22**	**93.92±0.25**
NECARE (Reported)	N/A	84.0±3.0	92.0±2.0	90.0±2.0	90.0±2.0

Table 13: Performance Comparison for Dataset 6

models were prompted with example questions or “shots“ before the dataset prompt was given. The term "3-shot" refers to the 3 example questions prompted before the dataset's prompt. The results of these experiments demonstrated that LLMs were particularly effective when fine-tuned on larger, well-structured datasets, as evidenced by their success in the PPI prediction task.

5.1 Performance

When fine-tuned with QLoRA, Llama 3 shows superior performance on the PPI prediction task for each of the three datasets. On the neurodegenerative (Table 11) and metabolic disorder (Table 12) PPI prediction tasks, it scores an accuracy of 88.27% and 91.28% respectively. These values outperform the current best model SymLMF (Pei et al., 2021), which achieves only 86.11% and 81.37%.

Furthermore, this model fine-tuned with LoRA achieves a precision of 96.9%, which outperforms the 94% precision achieved with NECARE (Qiu et al., 2021) as shown in table 13. It is clear that the fine-tuned Llama 3 model is currently the best prediction method for identifying PPIs in the presence neurodegenerative diseases, metabolic disorders, and cancer.

5.2 Discussion

We show that fine-tuning the LLMs can increase performance by a substantial margin. However, this depends heavily on the size of the dataset used to train the model, and the specific prompting techniques used. While fine-tuning significantly boosted accuracy in datasets 4, 5, and 6 by up to 50%, model performance on datasets 1, 2, and 3 exhibited less pronounced improvements after fine-tuning (Tables 1-10).

In analyzing the discrepancies in model performance between the PPI and LDR tasks, one notable difference lies in the composition of the prompts used for each task. For the PPI task, each prompt includes two variable protein names. This dual-protein structure of the prompts likely provides the model with a relational context that aids in discerning interaction patterns between proteins, facilitating more effective learning and prediction.

In contrast, the LDR task prompts feature only one variable protein name, potentially limiting the model's learning and predictive capabilities due to insufficient relational or comparative data. The single protein name reduces the available contextual cues for predicting deregulation. This prompt de-

sign likely contributes to the lower accuracy in the LDR task, as the model may struggle to infer the broader biological impacts of LDR exposure from a solitary protein reference.

These results not only illustrate the current constraints of these models but also suggests potential avenues for improvements, such as the development of more domain-specific datasets related to LDR, or the application of prompt engineering techniques.

6 Evaluation of Model Predictions

In this section, we analyze the predictions made by the LLMs and highlight some of the proteins that were correctly and incorrectly identified. We focus on interpreting the results obtained from our experiments in tables 1-13, examining the predictions made and identifying patterns in each model's output to understand their current limitations.

6.1 Correctly Identified Proteins

After analyzing the model output for each LLM, there are a few commonalities between the correctly identified proteins. Many of the names follow standard naming conventions in molecular biology, such as using abbreviations or acronyms that represent the function or family of the protein. Some examples include: SLC (Solute Carrier) proteins slc9a6, slc3a2, slc27a4, slc1a1, slc38a3, and RP (Ribosomal Protein) proteins rpl24, rpl22, rpl9, rpl15, rps11, rps25, rps13, rps27rt.

Additionally, the proteins correctly identified seem to belong to various functional categories, such as cytoskeletal proteins: tubb4a, tubb, actb, signaling proteins: hras, gsk3b, camk2a, camk4, rab3b, and metabolic enzymes: aldh3b1, aldh1l1, psat1, cpt2, pnpo, ak5, pgm3.

Overall, the correctly identified proteins cover a diverse range of cellular functions, including signaling, metabolism, transport, cytoskeletal organization, and many others. The naming conventions and functional hints within the protein names suggest that these proteins are well-studied and recognized by the models, potentially due to their importance in various biological processes and their prevalence in scientific literature.

6.2 Incorrectly Identified Proteins

When contrasting the incorrectly identified proteins with the correctly identified ones, a few key differences can be observed. Specifically, the incorrectly identified protein names seem to follow less standardized naming conventions compared to the correctly identified ones. They lack common abbreviations or acronyms that indicate their functional categories or protein families.

Furthermore, it is more challenging to infer the functional categories or processes that the incorrectly identified proteins are involved in based solely on their names. These proteins could be less well-known or less extensively researched, making it more challenging for the models to accurately identify them.

Additionally, LLMs might have biases or limitations in their training data or algorithms, which could contribute to the discrepancies in identification accuracy. Ultimately, the correctly identified proteins seem to follow more recognizable naming conventions, belong to well-characterized functional categories, and potentially have a more substantial presence in scientific literature, which could explain why they were more accurately identified by the models compared to the incorrectly identified ones.

7 Dataset Cross-Reference Analysis

Independent of the LLM experiments, we conduct a dataset cross-reference analysis to identify the common proteins between the LDR and PPI datasets, highlighting those that may be involved in both processes. Through this extensive analysis, we gained a deeper understanding of the data utilized for training these LLMs and enhanced our understanding of the protein dynamics involved in both radiation response and disease mechanisms.

We identified overlaps between the PPI datasets 4, 5, and 6, and the combined LDR dataset 3c. The positive interaction pairs were identified for each of datasets 4 (11,762 proteins), 5 (10,262 proteins), and 6 (1,866 proteins). Subsequently, the significantly affected proteins in the combined dataset 3c were identified (1,111 proteins). Our findings show that the highest percentage of overlap with the LDR data was with dataset 4, the neurodegenerative PPI dataset.

7.1 Dataset Analysis Metrics

The metrics used for these experiments include the percentage of overlap, the multiset coverage, the Jaccard index, and the weighted Jaccard index. The difference between the percentage overlap and the multiset coverage is that the multiset coverage takes into account the frequency of reoccurring proteins

between all interactions. In other words, multiset coverage includes duplicate protein names, where the percentage overlap uses only unique proteins names.

The reasoning for calculating both multiset coverage and percent overlap is because if a specific protein occurs frequently in the protein interaction network, it likely contributes more to the overall biological structure. Thus, including the duplicate proteins in the calculation of multiset coverage illustrates the extent to which these proteins affect the network.

Additionally, the Jaccard index was used for calculating the set similarity for the unique proteins, and the weighted Jaccard index was used when accounting for duplicate proteins. These values measure the similarity between the two sets, while accounting for their sizes through normalization.

7.2 Neurodegenerative Diseases PPI

The neurodegenerative diseases PPI dataset exhibited the highest percentage of unique protein overlap (14.02%) and multiset coverage (22.21%). The Jaccard Index for unique proteins was 0.0633 and the Weighted Jaccard Index was 0.2546, indicating a significant shared profile. This neurodegenerative PPI dataset contains 820 unique proteins in the PPI network. There were 115 unique proteins identified to overlap between the LDR data and PPI data. Some of these proteins include MAPT (Microtubule-Associated Protein Tau) (Medeiros et al., 2011), HTT (Huntingtin) (Tabrizi et al., 2019; Jimenez-Sanchez et al., 2017), APP (Amyloid Precursor Protein) (de la Vega et al., 2021; X et al., 2021), and GFAP (Glial Fibrillary Acidic Protein) (Yang and Wang, 2015; Kunchok et al., 2019), each of which have been shown to be linked with neurodegenerative diseases.

7.3 Metabolic Disorders PPI

The metabolic disorders PPI dataset showed a 7.14% overlap with unique proteins, and a multiset coverage of 13.78%. Both the Jaccard Index (0.0357) and Weighted Jaccard Index (0.1420) were lower compared to the neurodegenerative dataset, indicating less similarity with the LDR dataset but still notable overlap. This metabolic diseases PPI dataset contains 1036 unique proteins in the network. There were 74 unique proteins identified between both sets. Some of these proteins include ALDH2 (Aldehyde Dehydrogenase 2) (Wang et al., 2021; Chen et al., 2022), ACE (Angiotensin-Converting Enzyme) (Fountain et al., 2024), and ACAD8 (Acyl-CoA Dehydrogenase 8) (Zhuang et al., 2022), which have been shown to link to metabolic disorders.

7.4 Cancer PPI

The overlap in the cancer PPI dataset was more modest, with an 8.84% overlap and 4.72% multiset coverage, highlighting 19 unique overlapping proteins. The Jaccard Index was notably low at 0.0145, and the Weighted Jaccard Index at 0.0305. Some notable proteins identified include PAK1 (P21-Activated Kinase 1) (Belli et al., 2023), GRM1 (Glutamate Metabotropic Receptor 1) (Mehta et al., 2013; Nord et al., 2014), ANK1 (Ankyrin 1) (Tessema et al., 2017), and PTEN (Phosphatase and Tensin Homolog) (Liu et al., 2015). These proteins are illustrated in Figure 3. The highlighted proteins are also found in the combined LDR dataset, indicating that these proteins are significantly deregulated after exposure to LDR.

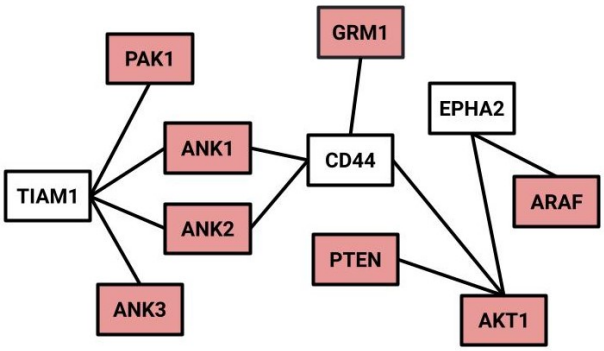

Figure 3: Cancer Protein Interaction Network. Highlighted Proteins are Significantly Affected by LDR.

7.5 Comparison

The higher overlap and Jaccard indices for dataset 4 show that there are more proteins in this network that are affected by LDR compared to those in the metabolic and cancer datasets. Similarly, the overlap of unique proteins between dataset 3c and dataset 6 is more than the overlap between datasets 3c and 5 despite its significantly larger size. This data suggests a higher probability that LDR affects cancer when compared to metabolic disorders. By highlighting the specific proteins overlapping between these datasets, we have identified key points for future research that can help bridge the gap between LDR exposure and disease mechanisms.

8 Conclusion

This study presents an exploration of the capabilities of LLMs in predicting the molecular dynamics of proteins under various conditions. By employing three state-of-the-art LLMs across multiple datasets, our research offers valuable insights into the potential utility and limitations of computational models for these tasks.

The fine-tuning process using LoRA proved to be a pivotal factor in enhancing model performance, demonstrating notable improvements in accuracy and predictive capabilities. Improving the accuracy of these models is key, because a major limitation of LLMs is their tendency to hallucinate, or give false information. Utilizing parameter efficient fine-tuning strategies helps to alleviate this problem while also maintaining an efficient computational complexity. Through the use of PEFT, the Llama 3 model outperforms the current best models for the PPI prediction tasks, indicating its potential for future advancements in biomolecular research.

Our analysis of protein identification by these models revealed intriguing patterns. Correctly identified proteins often belonged to well-characterized functional categories and were represented by standard naming conventions, suggesting that the pre-training on extensive biomedical literature may have equipped the models with a robust foundation of biological knowledge. Conversely, proteins that were incorrectly identified typically lacked these characteristics, possibly indicating areas where LLMs could benefit from further training or more focused dataset enrichment.

The cross-referencing of proteins affected by LDR with those involved in PPIs of neurodegenerative, metabolic, and cancer-related processes brought forth specific proteins that could be further explored in future studies. Notably, the neurodegenerative PPI dataset showed the highest overlap, where 115 unique proteins were identified in both datasets. These results highlight exactly which proteins in the PPI networks are significantly deregulated after LDR exposure, which could help to advance our understanding of how LDR affects disease mechanisms.

In conclusion, the integration of LLMs into biological research, particularly using fine-tuning techniques like LoRA, holds promising potential for advancing our understanding of the molecular mechanisms underpinning disease and radiation exposure. The versatility and scalability of these models make them instrumental tools in the ongoing quest to decode complex biological data. Their capacity to learn patterns and generate insights from extensive datasets holds immense promise for future research endeavors. Future work should focus on expanding the datasets, specifically the LDR data, and refining model architectures to further enhance the precision and applicability of LLMs in scientific discovery.

Acknowledgments

This material is based upon work for the LUCID: Low-dose Understanding, Cellular Insights, and Molecular Discoveries program supported by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, under Contract DE-AC02-06CH11357.

References

AI@Meta. 2024. Llama 3 model card.

Gregorio Alanis-Lobato, Miguel A. Andrade-Navarro, and Martin H. Schaefer. 2016. HIPPIE v2.0: enhancing meaningfulness and reliability of protein–protein interaction networks. *Nucleic Acids Research*, 45(D1):D408–D414.

Juan C. Aledo, Francisco R. Cantón, and Francisco J. Veredas. 2017. A machine learning approach for predicting methionine oxidation sites. *BMC Bioinformatics*, 18(1):430.

Z Barjaktarovic, J Merl-Pham, O Azimzadeh, S J. Kempf, K Raj, M J. Atkinson, and S Tapio. 2017. Low-dose radiation differentially regulates protein acetylation and histone deacetylase expression in human coronary artery endothelial cells. *International Journal of Radiation Biology*, 93(2):156–164. PMID: 27653672.

Stefania Belli, Daniela Esposito, Alessandra Allotta, Alberto Servetto, Paola Ciciola, Ada Pesapane, Claudia M. Ascione, Fabiana Napolitano, Concetta Di Mauro, Elena Vigliar, Antonino Iaccarino, Carmine De Angelis, Roberto Bianco, and Luigi Formisano. 2023. Pak1 pathway hyper-activation mediates resistance to endocrine therapy and cdk4/6 inhibitors in er+ breast cancer. *npj Breast Cancer*, 9:48.

Elliot Bolton, Abhinav Venigalla, Michihiro Yasunaga, David Hall, Betty Xiong, Tony Lee, Roxana Daneshjou, Jonathan Frankle, Percy Liang, Michael Carbin, and Christopher D. Manning. 2024. Biomedlm: A 2.7b parameter language model trained on biomedical text. *Preprint*, arXiv:2403.18421.

C. H. Chen, B. R. Kraemer, and D. Mochly-Rosen. 2022. Aldh2 variance in disease and populations. *Dis Model Mech*, 15(6):dmm049601.

Muhao Chen, Chelsea J T Ju, Guangyu Zhou, Xuelu Chen, Tianran Zhang, Kai-Wei Chang, Carlo Zaniolo, and Wei Wang. 2019. Multifaceted protein–protein interaction prediction based on Siamese residual RCNN. *Bioinformatics*, 35(14):i305–i314.

M. Pagnon de la Vega, V. Giedraitis, W. Michno, L. Kilander, G. Güner, M. Zielinski, M. Löwenmark, R. Brundin, T. Danfors, L. Söderberg, I. Alafuzoff, L. N. G. Nilsson, A. Erlandsson, D. Willbold, S. A. Müller, G. F. Schröder, J. Hanrieder, S. F. Lichtenthaler, L. Lannfelt, D. Sehlin, and M. Ingelsson. 2021. The uppsala app deletion causes early onset autosomal dominant alzheimer's disease by altering app processing and increasing amyloid β fibril formation. *Cold Spring Harbor Perspectives in Medicine*, 7(7):a024240.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

Moritz Ertelt, Vikram Khipple Mulligan, Jack B. Maguire, Sergey Lyskov, Rocco Moretti, Torben Schiffner, Jens Meiler, and Clara T. Schoeder. 2024. Combining machine learning with structure-based protein design to predict and engineer post-translational modifications of proteins. *PLOS Computational Biology*, 20(3):1–20.

Noelia Ferruz, Steffen Schmidt, and Birte Höcker. 2022. Protgpt2 is a deep unsupervised language model for protein design. *Nature Communications*, 13(1):4348.

J. H. Fountain, J. Kaur, and S. L. Lappin. 2024. Physiology, renin angiotensin system. *StatPearls*.

Somaye Hashemifar, Behnam Neyshabur, Aly A Khan, and Jinbo Xu. 2018. Predicting protein–protein interactions through sequence-based deep learning. *Bioinformatics*, 34(17):i802–i810.

M. Hauptmann, R. D. Daniels, E. Cardis, H. M. Cullings, G. Kendall, D. Laurier, M. S. Linet, M. P. Little, J. H. Lubin, D. L. Preston, D. B. Richardson, D. O. Stram, I. Thierry-Chef, M. K. Schubauer-Berigan, E. S. Gilbert, and A. Berrington de Gonzalez. 2020. Epidemiological studies of low-dose ionizing radiation and cancer: Summary bias assessment and meta-analysis. *J Natl Cancer Inst Monogr*, 2020(56):188–200. Erratum in: J Natl Cancer Inst Monogr. 2023 May 4;2023(61):e1. PMID: 32657347; PMCID: PMC8454205.

Daniela Hladik, Claudia Dalke, Christine von Toerne, Stefanie M. Hauck, Omid Azimzadeh, Jos Philipp, Marie-Claire Ung, Helmut Schlattl, Ute Rößler, Jochen Graw, Michael J. Atkinson, and Soile Tapio. 2020. Creb signaling mediates dose-dependent radiation response in the murine hippocampus two years after total body exposure. *Journal of Proteome Research*, 19(1):337–345. PMID: 31657930.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Kanchan Jha, Sourav Karmakar, and Sriparna Saha. 2023. Graph-bert and language model-based framework for protein–protein interaction identification. *Scientific Reports*, 13(1):5663.

K. Ji, Y. Wang, L. Du, et al. 2019. Research progress on the biological effects of low-dose radiation in china. *Dose-Response*, 17(1).

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

M. Jimenez-Sanchez, F. Licitra, B. R. Underwood, and D. C. Rubinsztein. 2017. Huntington's disease: Mechanisms of pathogenesis and therapeutic strategies. *Cold Spring Harbor Perspectives in Medicine*, 7(7):a024240.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Rhea Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. 2021. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589. Epub 2021 Jul 15. PMID: 34265844; PMCID: PMC8371605.

Kenji Kamiya, Kotaro Ozasa, Suminori Akiba, Ohstura Niwa, Kazunori Kodama, Noboru Takamura, Elena K Zaharieva, Yuko Kimura, and Richard Wakeford. 2015. Long-term effects of radiation exposure on health. *The Lancet*, 386:469–478. From Hiroshima and Nagasaki to Fukushima.

Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. 2017. Kegg: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res*, 45(D1):D353–D361. Epub 2016 Nov 28. PMID: 27899662; PMCID: PMC5210567.

Md Gulam Musawwir Khan and Yi Wang. 2022. Advances in the current understanding of how low-dose radiation affects the cell cycle. *Cells*, 11(3).

A. Kunchok, A. Zekeridou, and A. McKeon. 2019. Autoimmune glial fibrillary acidic protein astrocytopathy. *Current Opinion in Neurology*, 32(3):452–458.

Y Liu, X Hu, C Han, L Wang, X Zhang, X He, and X Lu. 2015. Targeting tumor suppressor genes for cancer therapy. *Bioessays*, 37:1277–1286.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6).

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. *Preprint*, arXiv:2402.16445.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and B Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. *Younes Belkada and Sayak Paul," PEFT: State-of-the-art Parameter-Efficient Fine-Tuning methods*.

R. Medeiros, D. Baglietto-Vargas, and F. M. LaFerla. 2011. The role of tau in alzheimer's disease and related disorders. *CNS Neurosci Ther*, 17(5):514–524.

MS Mehta, SC Dolfi, R Bronfenbrener, E Bilal, C Chen, D Moore, Y Lin, H Rahim, S Aisner, RD Kersellius, J Teh, S Chen, DL Toppmeyer, DJ Medina, S Ganesan, A Vazquez, and KM Hirshfield. 2013. Metabotropic glutamate receptor 1 expression and its polymorphic variants associate with breast cancer phenotypes. *PLoS One*, 8:e69851.

Erik Nijkamp, Jeffrey Ruffolo, Eli N. Weinstein, Nikhil Naik, and Ali Madani. 2022. Progen2: Exploring the boundaries of protein language models. *Preprint*, arXiv:2206.13517.

Karolin H. Nord, Henrik Lilljebjörn, Francesco Vezzi, Jenny Nilsson, Linda Magnusson, Johnbosco Tayebwa, Danielle de Jong, Judith V. M. G. Bovée, Pancras C. W. Hogendoorn, and Karoly Szuhai. 2014. Grm1 is upregulated through gene fusion and promoter swapping in chondromyxoid fibroma. *Nature Genetics*, 46:474–477.

Rose Oughtred, Jennifer Rust, Christie Chang, Bobby-Joe Breitkreutz, Chris Stark, Andrew Willems, Lorrie Boucher, Genie Leung, Nadine Kolas, Frederick Zhang, Sonam Dolma, Jasmin Coulombe-Huntington, Andrew Chatr-aryamontri, Kara Dolinski, and Mike Tyers. 2021. The biogrid database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Science*, 30(1):187–200.

Fen Pei, Qingya Shi, Haotian Zhang, and Ivet Bahar. 2021. Predicting protein–protein interactions using symmetric logistic matrix factorization. *Journal of Chemical Information and Modeling*, 61(4):1670–1682. PMID: 33831302.

Jiajun Qiu, Kui Chen, Chunlong Zhong, Sihao Zhu, and Xiao Ma. 2021. Network-based protein-protein interaction prediction method maps perturbations of cancer interactome. *PLOS Genetics*, 17(11):1–19.

Z Schmal, A Isermann, D Hladik, C von Toerne, S Tapio, and CE Rübe. 2019. Dna damage accumulation during fractionated low-dose radiation compromises hippocampal neurogenesis. *Radiotherapy and Oncology*, 137:45–54.

D J Shah, R K Sachs, and D J Wilson. 2014. Radiation-induced cancer: a modern view. *British Journal of Radiology*, 85(1020):e1166–e1173.

Neel K. Sharma, Rupali Sharma, Deepali Mathur, Shashwat Sharad, Gillipsie Minhas, Kulsajan Bhatia, Akshay Anand, and Sanchita P. Ghosh. 2018. Role of ionizing radiation in neurodegenerative diseases. *Frontiers in Aging Neuroscience*, 10.

D. Szklarczyk, A. L. Gable, K. C. Nastou, D. Lyon, R. Kirsch, S. Pyysalo, N. T. Doncheva, M. Legeay, T. Fang, P. Bork, L. J. Jensen, and C. von Mering. 2021. The string database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res*, 49(D1):D605–D612. Erratum in: Nucleic Acids Res. 2021 Oct 11;49(18):10800. PMID: 33237311; PMCID: PMC7779004.

Sarah J. Tabrizi, Blair R. Leavitt, G. Bernhard Landwehrmeyer, Edward J. Wild, Carsten Saft, Roger A. Barker, Nick F. Blair, David Craufurd, Josef Priller, Hugh Rickards, Anne Rosser, Holly B. Kordasiewicz, Christian Czech, Eric E. Swayze, Daniel A. Norris, Tiffany Baumann, Irene Gerlach, Scott A. Schobel, Erika Paz, Anne V. Smith, C. Frank Bennett, and Roger M. Lane. 2019. Targeting huntingtin expression in patients with huntington's disease. *New England Journal of Medicine*, 380(24):2307–2316.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. GitHub. Available: `https://crfm.stanford.edu/2023/03/13/alpaca.html`.

Jelena Popović Tatjana Paunesku, Aleksandra Stevanović and Gayle E. Woloschak. 2021. Effects of low dose and low dose rate low linear energy transfer radiation on animals – review of recent studies relevant for carcinogenesis. *International Journal of Radiation Biology*, 97(6):757–768. PMID: 33289582.

M Tessema, CM Yingling, MA Picchi, G Wu, T Ryba, Y Lin, AO Bungum, ES Edell, A Spira, and SA Belinsky. 2017. Ank1 methylation regulates expression of microrna-486-5p and discriminates lung tumors by histology and smoking status. *Cancer Letters*, 410:191–200.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Jin-song Wang, Hai-juan Wang, and Hai-li Qian. 2018. Biological effects of radiation on cancer cells. *Military medical research*, 5:1–10.

Q. Wang, B. Chang, X. Li, and Z. Zou. 2021. Role of aldh2 in hepatic disorders: Gene polymorphism and disease pathogenesis. *J Clin Transl Hepatol*, 9(1):90–98.

Yanbin Wang, Zhu-Hong You, Shan Yang, Xiao Li, Tong-Hai Jiang, and Xi Zhou. 2019. A high efficient biological language model for predicting protein–protein interactions. *Cells*, 8(2).

Xiao X, Liu H, Liu X, Zhang W, Zhang S, and Jiao B. 2021. App, psen1, and psen2 variants in alzheimer's disease: Systematic re-evaluation according to acmg guidelines. *Frontiers in Aging Neuroscience*, 13:695808.

Z. Yang and K. K. Wang. 2015. Glial fibrillary acidic protein: from intermediate filament assembly and gliosis to neurobiomarker. *Trends Neurosci*, 38(6):364–374.

Qiang Zhang, Keyang Ding, Tianwen Lyv, Xinda Wang, Qingyu Yin, Yiwen Zhang, Jing Yu, Yuhao Wang, Xiaotong Li, Zhuoyi Xiang, Xiang Zhuang, Zeyuan Wang, Ming Qin, Mengyao Zhang, Jinlu Zhang, Jiyu Cui, Renjun Xu, Hongyang Chen, Xiaohui Fan, Huabin Xing, and Huajun Chen. 2024. Scientific large language models: A survey on biological & chemical domains. *Preprint*, arXiv:2401.14656.

D. Y. Zhuang, S. X. Ding, F. Wang, X. C. Yang, X. L. Pan, Y. W. Bao, L. M. Zhou, and H. B. Li. 2022. Identification of six novel variants of acad8 in isobutyryl-coa dehydrogenase deficiency with increased c4 carnitine using tandem mass spectrometry and ngs sequencing. *Front Genet*, 12:791869.

A Dataset Information

Dataset 1

The first dataset was provided from this study: "CREB Signaling Mediates Dose-Dependent Radiation Response in the Murine Hippocampus Two Years after Total Body Exposure" (Hladik et al., 2020). This study records the modulation of protein expressions in response to varied radiation exposures, categorizing proteins based on their upregulation or downregulation across three distinct radiation groups. A graphical representation of the significantly deregulated proteins can be seen in Chart A, this chart was presented in the original study and helps to visualize the structure and size of the dataset.

To construct a balanced representation of the data, we combine the identified upregulated and downregulated proteins, for each of the three radiation groups. Subsequently, we employ a randomized selection process, drawing an equitable count of proteins from the list of proteins deemed unaffected by LDR, as shown by the original study.

After cleaning the data, the number of proteins in each of the three subsets (1.1, 1.2, and 1.3) are 892, 1332, and 204 proteins respectively. Each subset maintains an equal distribution between proteins influenced by LDR and those unaffected, thus ensuring analytical balance. The LLMs are then tasked to evaluate the following query for each protein: "Given the options yes or no, will there be significant deregulation of the protein {protein x} 24 months post low-dose radiation exposure at {dosage level} Gy?".

Dataset 2

The second dataset was provided from the study titled "DNA damage accumulation during fractionated low-dose radiation compromises hippocampal Neurogenesis" (Schmal et al., 2019). This research provides an evaluation of protein expression changes due to low-dose radiation (LDR), and gives information regarding the temporal aspects of radiation exposure on cellular processes. Similar to Dataset 1, we have provided

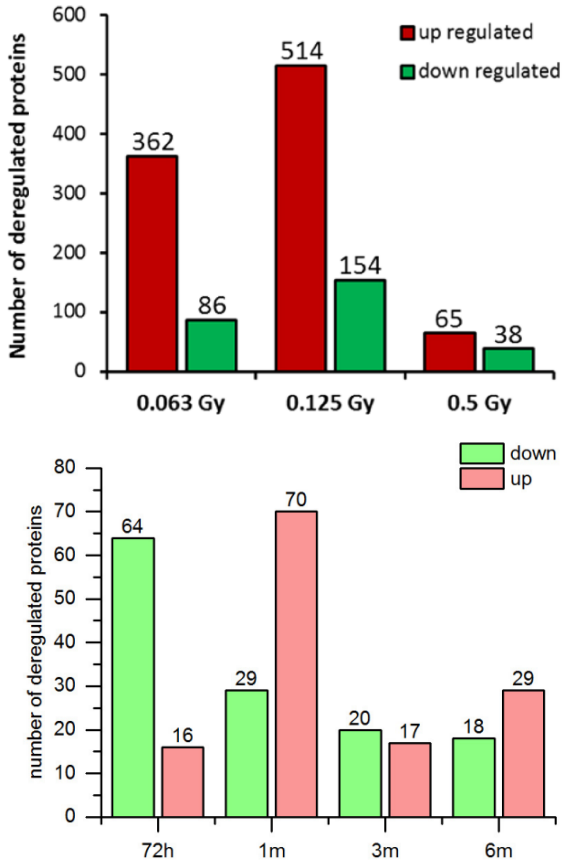

Figure 4: Chart A (Top) and Chart B (Bottom)

a graphical representation of the significantly deregulated proteins in Chart B, this chart was also presented in the original study.

This dataset encapsulates the regulatory status of proteins, upregulated or downregulated, across four distinct cohorts. Each cohort underwent an identical radiation dosage of 2.0 Gy, but the resultant protein expression was analyzed at different post-exposure intervals. Mirroring the methodology applied to the first dataset, we combined the upregulated and downregulated protein expressions, as indicated by the red and green columns for each group, and randomly sample the unaffected proteins. This approach ensures a balanced representation of the data for each group.

After the data cleaning process the number of proteins in each of the four subsets (2.1, 2.2, 2.3, and 2.4) are 160, 198, 74, and 94 proteins respectively. Similar to the first dataset, the LLMs are then tasked to evaluate the following query for each protein: "Given the options yes or no, will there be significant deregulation of the protein {protein x} {time} after exposure to low dose radiation at 2.0 Gy?".

Dataset 3

Dataset 3 was provided from the study titled "Low-dose radiation differentially regulates protein acetylation and histone deacetylase expression in human coronary artery endothelial cells" (Barjaktarovic et al., 2017). This work delves into the post-translational modifications, specifically acetylation, that occur in proteins of human coronary artery endothelial cells as a result of low-dose radiation (LDR) exposure. The administered radiation dose of 0.5 Gy and the subsequent temporal protein measurements offer valuable insights into the cellular responses.

In this study, the protein deregulation via acetylation was monitored at two time intervals: at 4 hours and then at 24 hours post-radiation exposure. The resulting subsets for analysis, capturing the 4 hour period and the 24 hour period, comprised 98 and 154 proteins, respectively. These two groups represents datasets 3.1 and 3.2.

To maintain a consistent evaluation strategy, the LLMs are given a prompt for each protein in the dataset: "Given the options yes or no, will there be an altered acetylation status of protein {protein x} 24 hours after exposure to low dose radiation at 0.5 Gy?".

Dataset 3c

Dataset 3c represents a strategic combination of datasets 1, 2, and 3. This integration was motivated by insights derived from the review of experiments 1 through 3, which suggested limitations in the approach's efficacy. Specifically, the chosen prompts for these experiments were potentially too narrowly defined, and the datasets themselves were not sufficiently sized to enable the LLMs to recognize the patterns within the data.

To address these challenges, we synthesized a comprehensive dataset combining the protein deregulation data from the first 3 datasets. The objective was to refine the training process for the LLMs using a larger dataset and significantly broader prompting strategy. The reformulated prompt used to train the LLMs is: "Given the options yes or no, will there be deregulation of

the protein {protein x} after low-dose radiation exposure?".

Dataset 3c includes an amalgamation of datasets 1.1, 1.2, 1.3, 2.1, 2.2, 2.3, 2.4, 3.1, 3.2. It is a combination of the proteins in each of the columns from charts A and B from Figure 17, along with the proteins from datasets 3.1 and 3.2. The repeated proteins between all datasets were removed. These proteins become deregulated across different time intervals and radiation dosage levels, resulting in a comprehensive dataset of 1,111 proteins.

A randomized sampling methodology was employed to select proteins that do not exhibit deregulation across these varied experimental conditions, which resulted in a dataset featuring 2,222 proteins. This augmented dataset size significantly enhances the LLMs' training ability, facilitating a more nuanced understanding of protein behavior in response to low-dose radiation exposure.

Dataset 4

Dataset 4 was provided by the study "Predicting Protein-Protein Interactions Using Symmetric Logistic Matrix Factorization" (Pei et al., 2021). In an effort to understand the Protein-Protein Interactions (PPIs) specific to disease mechanisms, this dataset focuses on the protein interactions associated with neurodegenerative diseases.

From the data provided, this study narrows its focus to a subset encompassing 820 proteins, which form a network of 5,881 positive (interaction present) and 5,881 negative (interaction absent) protein pairs. This gives us a total of 11,762 protein interactions. The LLMs are prompted with the following query for each protein pair: "Given the options yes or no, do proteins {protein x} and {protein y} interact in the presence of neurodegenerative disease?".

Dataset 5

Concurrent with the exploration of neurodegenerative diseases in Dataset 4, Dataset 5 focuses on metabolic disorders. Provided by the same study "Predicting Protein-Protein Interactions Using Symmetric Logistic Matrix Factorization" (Pei et al., 2021), this dataset shines a light on the protein interactions that might contribute to metabolic dysfunction.

This data is made up of 1,063 proteins, from which a balanced collection of 5,131 positive and 5,131 negative protein pairs is drawn. This leads to a total dataset size of 10,262 protein interactions. The LLMs will use a similar prompt to that used for dataset 4: "Given the options yes or no, do proteins {protein x} and {protein y} interact in the presence of a metabolic disorder?".

Dataset 6

Dataset 6 was provided from the study "Network-based protein-protein interaction prediction method maps perturbations of cancer interactome" (Qiu et al., 2021), which offers a focused lens on the protein interaction network within the context of cancer.

This data presents a network of protein interactions consisting of 933 positive instances—indicative of an interaction's presence—and 1,308 negative instances, signifying the absence of interaction. To achieve an even representation akin to previous datasets, we conduct a randomized selection, reducing the negative instances to 933, thereby equalizing the number of positive and negative samples and giving a total of 1,866 protein interactions. The prompt used for this dataset is similar to datasets 4 and 5: "Given the options yes or no, do proteins {protein x} and {protein y} interact in the presence of cancer?".