UNVEILING ZERO SHOT PREDICTION FOR GENE ATTRIBUTES THROUGH INTERPRETABLE AI

Ala Jararweh[*,1,2], **Oladimeji Macaulay**[*,2], **David Arredondo**[*,2], **Olufunmilola M Oyebamiji**[2],

Luis Tafoya [2], **Kushal Virupakshappa** [2] & **Avinash Sahu**[1,2]

[1] Department of Computer Science, The University of New Mexico

[2] Comprehensive Cancer Center, The University of New Mexico
{KVirupakshappa,ASahu}@salud.unm.edu

ABSTRACT

Representation learning has transformed the prediction of structures and functions of genes and proteins by employing sequence, expression, and network data. Yet, this approach taps into just a fraction of the knowledge accumulated over more than a century of genetic research. Here, we introduce GeneLLM, an interpretable transformer-based model that integrates textual information through contrastive learning to refine gene representations. While it has been posited that such knowledge representation could result in a bias towards well-characterized genes, GeneLLM surprisingly shows high accuracy across eight gene-related benchmarks, not only matching but often outperforming task-specific models, with a 50% increase in accuracy over its closest solubility-specific competitor. It demonstrates robust zero-shot learning capabilities for unseen gene annotations. The model's interpretability and our multimodal strategic approach to mitigating inherent data biases bolster its utility and reliability, particularly in biomedical applications where interpretability is paramount. Our findings affirm the complementary nature of unstructured text to structured databases in enhancing biomedical predictions, while conscientiously addressing interpretability and bias for AI deployment in healthcare. The code and datasets can be found at `https://www.avisahuai.com/tools` on request.

1 INTRODUCTION

Genes encode proteins, that drive biological processes, and are fundamental to the functions of living organisms (Alberts, 2017). Gene and protein functions help us explain their roles in individual cells and in human health and disease, yet our understanding of many genes remains incomplete. This is attributed to their complexity and variability across different cellular, individual, and environmental contexts (Virolainen et al., 2023). Traditional laboratory-based models, which capture only a very small subset of these contexts, are insufficient on their own for understanding this complexity. To complement laboratory approaches, task-specific machine learning models have been developed to further predict gene attributes (Novakovsky et al., 2023; Piya et al., 2023); however, they are constrained by the need for large task-specific training datasets, limiting their broader applicability.

Offering a versatile alternative to traditional task-specific models, the advent of foundation models has introduced a new paradigm in machine learning. These models, once pre-trained on large unlabeled datasets, can be fine-tuned for a wide array of predictive tasks and often outperform task-specific models (Bommasani et al., 2021). The advantage of foundation models lies in their ability to operate with minimal labeled data (few-shot learning) and sometimes with no labeled data (zero-shot learning) (Zhou et al., 2023).

Large text bodies are utilized by Large Language Models (LLMs) to identify statistical relationships between words, demonstrating their capability to encapsulate comprehensive knowledge in unstructured text (Naveed et al., 2023). Many AI models do not incorporate the extensive literature knowledge that is available and instead use only expression or sequence data to create embeddings of genes and cells for downstream prediction of structure and gene annotations (Pesaranghader et al., 2022). Structured information such as Gene Ontology (GO) (Carbon & Mungall, 2018), which categorizes genes and captures the relationships between them, can be injected into LLM knowledge derived from unstructured text to enhance the predictive power of a model. Such knowledge injection could be accomplished through methods like contrastive learning (Tian et al., 2019; Zhang et al., 2022) and Bootstrap Your Own Latent(BYOL) (Grill et al., 2020).

Here, we enable zero-shot learning for gene tasks by introducing an interpretable large language model to harness the vast unstructured textual data on genes. We first extract a summary of every gene and input it to an LLM pretrained on a biomedical text corpus, producing initial gene embeddings. These embeddings are further refined by incorporating Gene Ontology (GO) information through contrastive learning, enriching the embeddings with structured biological knowledge. Subsequent sections will detail the relevant literature, our methodology, and the efficacy of our approach in various predictive tasks, including zero-shot learning and cell- and gene-specific predictions.

2 Related Work

Gene and protein representation learnings have primarily focused on expression, sequence, or network data (Theodoris et al., 2023; Du et al., 2019b; John Jumper & Hassabis, 2021), driving advancements in gene-gene interaction and 3D structure predictions as well as cell property elucidation from single-cell RNA-Seq data (de Guia et al., 2020). Despite their efficacy in disease association and cancer classification, they primarily rely on quantitative data, potentially overlooking the contextual information embedded in textual sources. To improve protein representations, ProteinBERT (Brandes et al., 2022) and OntoProtein (Zhang et al., 2022) demonstrate the potential of integrating protein sequences with Gene Ontology (GO) using self-supervised and contrastive learning, respectively. Our work extends this multimodal approach by incorporating GO annotations and textual information.

Recent advances in NLP models, such as BERT (Devlin et al., 2019), LLaMA (Touvron et al., 2023), and GPT (Radford & Narasimhan, 2018), have revolutionized the utilization of unstructured biomedical texts from repositories like PubMed[1] and Europe PMC[2]. However, the application of such models for gene and cell-specific predictions remains understudied.

Recent studies have sounded the alarm on the issue of bias amplification in AI (Gatzemeier, 2021), particularly in healthcare where biases in training data, such as knowledge biases, can result in significant disparities like the under-diagnosis of underserved populations (Seyyed-Kalantari et al., 2021). These studies call for transparency and interpretability to ensure equitable healthcare (Vokinger et al., 2021). Our work focuses on knowledge bias in gene data by a multimodal strategy for bias mitigation.

3 GeneLLM

The GeneLLM framework incorporates a pre-trained LLM augmented with Gene Ontology (GO) knowledge through contrastive learning (CL) (Figure 1). This approach aims to infuse the language representations of gene summaries with the information from the Gene Ontology. Here, we detail the configuration of the pre-trained encoder, followed by a description of the CL, and finally discuss the methodologies implemented to interpret the model and the development of cell embeddings.

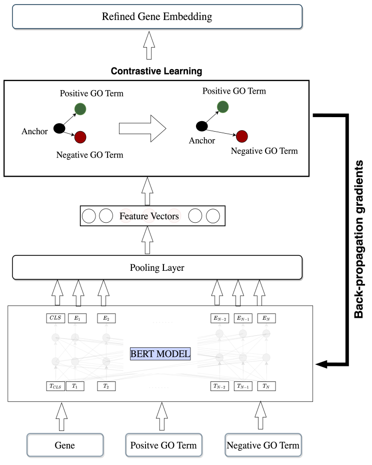


Figure 1: A high overview of **GeneLLM**. Contrastive learning enriches the gene representations by introducing GO term relationships between genes.

[1] https://pubmed.ncbi.nlm.nih.gov/

[2] https://europepmc.org

3.1 GENELLM ENCODER FOR INITIAL EMBEDDING

For gene embeddings, we utilize BiomedBERT (Gu et al., 2020), a BERT (Devlin et al., 2019) based text encoder specifically trained on biomedical text corpora. BiomedBERT has outperformed general-domain language models in biomedical natural language processing applications. To generate initial (GeneLLM-Base) embeddings, each gene or GO term text summary is tokenized into a sequence of N tokens $(x_1, \ldots, x_N)$, encompassing special tokens $[CLS]$ and $[SEP]$. These tokens are fed into BiomedBERT, which in turn outputs a series of token embeddings $(T_1, \ldots, T_N)$. We combine these embeddings to obtain summary-level embedding E. After assessing various pooling methodologies, mean pooling was selected based on its performance, thereby defining the summary-level embedding as $E = (T_1 + T_2 + \ldots + T_N)/N$. The resulting embedding is a 768-dimensional vector representing the gene/GO summary.

3.2 CONTRASTIVE LEARNING

Contrastive learning (CL) is an approach that aims to learn representations by instructing a model on which data points are similar or dissimilar. CL is used to draw the representations of semantically similar genes closer together in the embedding space while pushing apart those of dissimilar genes. Semantically similar genes are identified as those sharing a Gene Ontology (GO) term. We investigated two methodologies to inject gene-GO relationships into the GeneLLM embeddings:

The **first** approach minimizes the distances between gene pairs sharing the same GO terms (g_a, g_p) and maximizes the distance between unrelated genes (g_a, g_n), aiming to minimize the objective function:

$$\mathcal{L} = \sum_{i=1}^{M} \Big[\max\Big(0,\, \delta + d(E_{g_a}^{(i)}, E_{g_p}^{(i)}) - d(E_{g_a}^{(i)}, E_{g_n}^{(i)})\Big)\Big] \tag{1}$$

where δ $(0 < \delta < 1)$ is the margin that provides a buffer between the distances of positive and negative pairs, $\mathcal{L}$ is the loss computed over a set of M triplets, $d(x, y)$ denotes the distance between two embeddings, and $E_g^{(i)}$ represents the i-th gene embedding. This CL method utilized 7.6 million gene-GO annotations (Carbon & Mungall, 2018). However, it did not explicitly embed the GO terms, so could not generalized to unseen GO terms.

The **second** approach involves co-embedding genes and GO terms into a shared embedding space. In this method, 18,000 GO terms and 15,000 genes were co-embedded to create gene-GO maps. This CL method utilizes 235,000 confident gene-GO relationships, S (details in the Appendix D). It aims to bring anchor genes (g_a) closer to their corresponding GO terms (t_p, the positives) in the shared embedding space E, while maximizing the distance between g_a and unrelated GO terms (t_n, the negatives). The loss function is defined as:

$$\mathcal{L} = \sum_{(g_a, t_p) \in S,\, (g_a, t_n) \notin S} \big[\max\big(0,\, \delta + d(E_{g_a}, E_{t_p}) - d(E_{g_a}, E_{t_n})\big)\big] \tag{2}$$

The margin δ $(0 < \delta < 1)$ ensures that the positive pairs are closer to the anchor than the negative pairs. The second CL approach, by embedding GO terms explicitly, enables the prediction of GO-gene pairs for unseen GO terms, facilitating zero-shot learning, as discussed in Section 4.

3.3 SHAPLEY ANALYSIS

SHAP (**SH**apley **A**dditive Ex**P**lanations) values quantify the impact of each feature on the difference between the actual model output and the expected baseline output (Lundberg & Lee, 2017). In LLMs, input features are often sub-word tokens, which may not be inherently interpretable; however, SHAP values are additive, meaning the sum of SHAP values for all features in a sample equals the model's output. In our analysis, we aggregate the contributions of all tokens within each word to determine the total SHAP of each word in a text. Contributions are calculated using the SHAP partition explainer (Lundberg, 2024), which calculates SHAP values for each token, or Owen values for groups of tokens in cases where text inputs are prohibitively large due to the exponential runtime of the exhaustive SHAP algorithm (Owen, 1977). Finally, the SHAP value of a word can vary contextually and may follow multimodal distributions; thus, we use the 90th percentile of SHAP values to denote word importance, ensuring importance is calculated in critical instances and resilience against outliers.

3.4 Cell embeddings

To obtain GeneLLM embeddings of cells, we utilize their gene expression data. GeneLLM cell embeddings C, representing M cells in D dimensions, are calculated as $C = GE$, where G is the gene expression matrix of size $M \times N$ and E is the gene embeddings matrix of size $N \times D$.

4 Results and Discussion

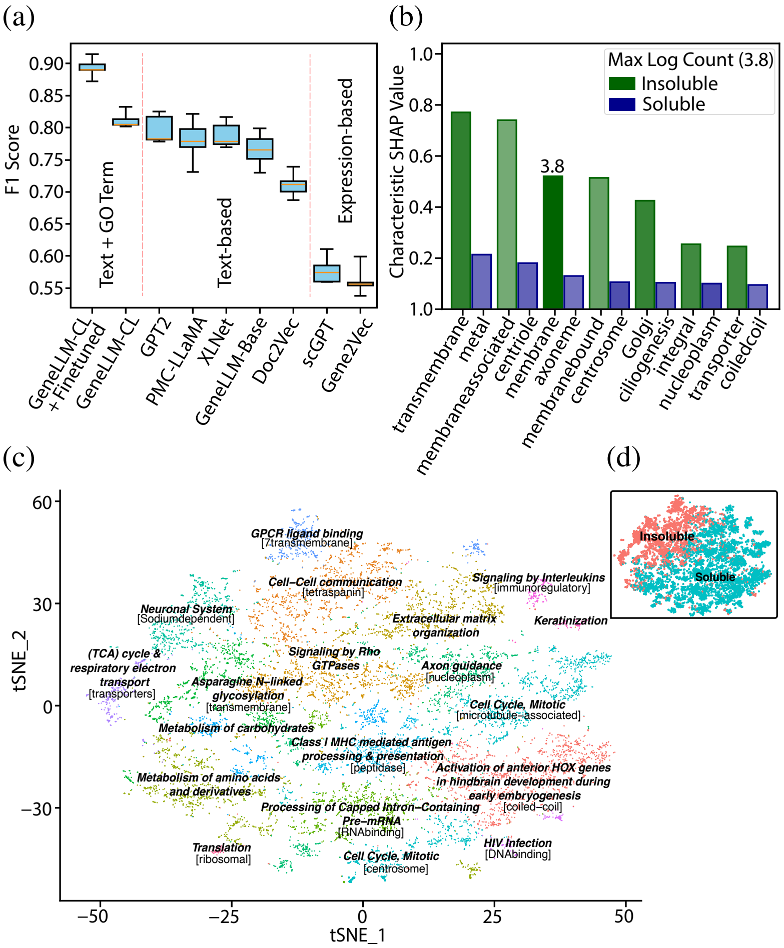


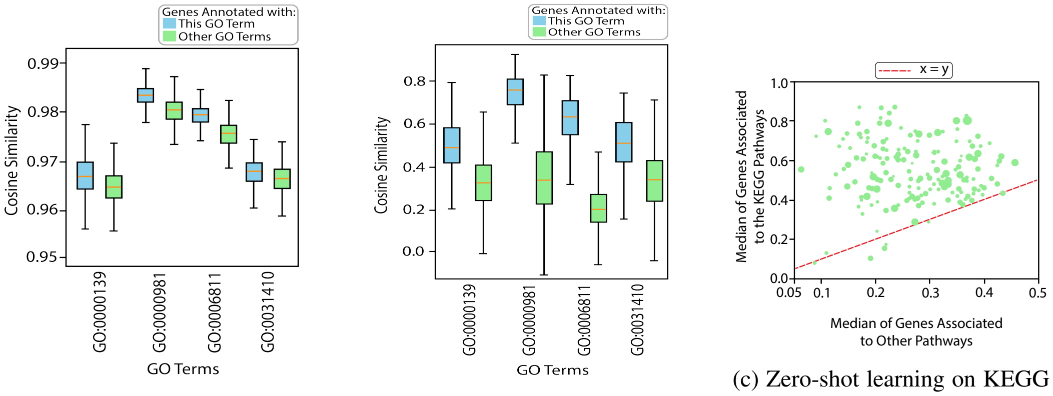
Figure 2: Performance and interpretability GeneLLM on solubility benchmark. (a) Performance of GeneLLM and baselines in predicting solubility of protein products of genes. (b) Top 7 words with the highest characteristic SHAP values across all gene summaries shown for two categories. (c) A t-SNE plot depicting gene clusters, with pathway enrichment analysis labels in bold, complemented by important SHAP-derived (frequent) terms in brackets. An example of interpretative insight: the ”GPCR ligand binding” clusters are ”insoluble” because they are ”transmembrane”. (d) Cluster showing predicted solubility of clusters in (c).

Evaluating GeneLLM for Solubility Prediction: Protein function is closely linked to its solubility in aqueous medium, impacting roles such as transporter, receptor, pharmacological, enzyme activity(Dyson et al., 2008). Experimentally determining solubility is laborious and expensive, requiring protein expression and purification followed by solubility tests under different conditions (Wang & Zou, 2023b). GeneLLM, in predicting protein solubility from gene products, outperformed baseline methods in Figure 2a and benefited from contrastive learning and fine-tuning. Text-based approaches outperformed expression-based approaches. GeneLLM significantly eclipsed the performance of 13 dedicated solubility-task models, with an impressive 50% increase in accuracy over its closest competitor. (Table 3) .

Model	Dosage Sensitivity	BivalentVs Lys4 Methylated	BivalentVs Non Methylated	Tf range	Tf target type	Solubility	Subcellular localization	Conservation (Pearson Corr.)
Majority Classifier	0.73 ± —	0.58 ± —	0.75 ± —	0.73 ± —	0.41 ± —	0.52 ± —	0.39 ± —	— ± —
GPT2	0.74 ± 0.04	**0.86** ± **0.04**	0.80 ± 0.11	0.71 ± 0.03	0.18 ± 0.02	0.80 ± 0.02	0.77 ± 0.01	0.31 ± 0.02
Doc2Vec	0.74 ± 0.04	0.84 ± 0.06	0.78 ± 0.05	0.66 ± 0.07	0.26 ± 0.01	0.71 ± 0.03	0.69 ± 0.02	0.34 ± 0.01
PMC-LLaMA	0.86 ± 0.05	0.77 ± 0.04	**0.84** ± **0.07**	0.64 ± 0.08	0.08 ± 0.01	0.78 ± 0.03	0.69 ± 0.01	**0.55** ± **0.01**
XLNet	0.74 ± 0.06	0.84 ± 0.06	0.83 ± 0.08	0.69 ± 0.05	0.12 ± 0.01	0.79 ± 0.02	0.76 ± 0.01	0.40 ± 0.01
Gene2Vec	0.84 ± 0.04	0.84 ± 0.06	0.75 ± 0.06	**0.75** ± **0.08**	0.21 ± 0.01	0.56 ± 0.02	0.54 ± 0.02	0.50 ± 0.02
BERT-Base	0.76 ± 0.09	0.83 ± 0.06	0.77 ± 0.10	0.68 ± 0.04	0.17 ± 0.01	0.77 ± 0.02	0.76 ± 0.01	0.43 ± 0.01
GeneLLM	**0.87** ± **0.06**	**0.86** ± **0.09**	0.82 ± 0.08	0.74 ± 0.07	**0.49** ± **0.04**	**0.89** ± **0.01**	**0.83** ± **0.01**	0.53 ± 0.01

Table 1: Comprehensive evaluation of GeneLLM: analyzing 5-fold Accuracy outcomes for contrastive learning-enhanced gene embeddings across a spectrum of gene prediction tasks.

Evaluating GeneLLM on Gene Tasks: *Baselines*: Refer to Appendix B.2. : Our evaluation includes eight gene-related tasks: Dosage Sensitivity, Chromatin State Predictions, Transcription Factor (TF) Range Prediction, TF Target Type Identification, Protein Localization, Solubility, and Gene Conservation (details in Appendix B.1.).*Results* (Table 1): GeneLLM exhibited superior performance in Dosage Sensitivity, a critical factor for interpreting copy number variants in genetic diagnostics. GeneLLM also surpassed the baselines in TF Target Type Identification, Protein Localization, and Solubility. These results suggest that textual data is particularly effective for tasks necessitating a comprehensive understanding of biological processes and molecular functions (details in Table 1). However, GeneLLM did not perform as strongly in Chromatin State and TF Range Predictions, possibly reflecting certain dimensions about gene regulation lacking in text.



(a) GeneLLM-Base

(b) GeneLLM-CL

(c) Zero-shot learning on KEGG pathways

Figure 3: GeneLLM's Zero Shot Prediction : (a,b) Impact of Contrastive Learning on Gene-GO Term Similarity Prediction. This figure presents cosine similarity scores between genes and GO terms for (a) GeneLLM-Base and (b) GeneLLM-CL models. For each of the four selected GO terms, the plot compares the scores of genes annotated with the respective GO term to those not annotated with it (details in Appendix D). Note the scale variation across the models (GeneLLM-Base similarities narrower). (c) GeneLLM-CL (trained solely on GO-terms) prediction of gene-KEGG relationships (dots in the scatter plot). The plot compares the median similarity of genes related to KEGG pathways (y-axis) against genes not related (x-axis). Pathways above the X=Y line represent a stronger association with related genes, showcasing the model's predictive accuracy.

SHAP enables interpretability: Our SHAP analysis elucidates the model's decision-making process in solubility classification, where it distinguishes between *membrane* and *soluble* classes, encoded as 0 and 1. Despite this encoding, the term *transmembrane* emerges as the most important word for the *membrane* class (see Figure 2b). Clustering of GeneLLM's embeddings identifies key terms for gene clusters (see Figure 2c). For example, the word *DNA-binding* indicates solubility in the *HIV Infection* cluster; non-membrane nuclear DNA-binding proteins such as TAF1 are active in nuclei, which are areas affected by HIV (Burley & Roeder, 1996) (Figure 2c,d). Similarly, the word *ribosomal* indicates solubility in the *Translation* cluster; ribosomes are non-membrane-bound organelles that play a primary role in translation. Additional examples are provided in the Appendix.

Contrastive learning enables zero-shot predictions: Next, we examined GeneLLM's ability to discern relationships between genes and Gene Ontology (GO) terms. By embedding genes and GO terms in the same space, we observed that without contrastive learning, GeneLLM's contextual similarity measures were no better than random (Figure 3a). However, contrastive learning enabled GeneLLM to capture gene-GO term relationships and generalize to new, unseen GO categories, demonstrating zero-shot learning (Figure 3b). Furthermore, even without training on KEGG pathways, GeneLLM with GO-based contrastive learning accurately identified gene-KEGG pathway links, further demonstrating its zero-shot learning ability (see Figure 3c). Zero-shot learning failed for some GO and KEGG terms, likely due to limited knowledge available, as detailed in subsequent sections.

Evaluating GeneLLM cell embeddings: We assessed GeneLLM's cell embeddings for their ability to differentiate cell types in human peripheral blood mononuclear cells (10x Genomics, 2019). We chose GPT-2 for comparison as it represents the highest-performing baseline in the solubility task. GeneLLM-CL embeddings outperformed both GeneLLM-Base and GPT-2 (Table 2). For a comprehensive comparison, refer to Appendix C.

Mitigating bias in gene representation with Multi-Modal Learning: Human knowledge of genes is biased; some genes are well-studied due to various associations, while many remain under-researched. Furthermore, gene knowledge is often shaped by genetic research subject to human biases (Stoeger et al., 2018), and exacerbated by publication biases. AI models trained on such data risk perpetuating these biases. Given the challenge of quantifying human biases (Viswanathan et al., 2017), we focus on the potential bias from limited knowledge. We demonstrated this bias's impact on AI through two analyses. Firstly, we used the summary lengths of KEGG pathways as proxies for their knowledge levels, assessing our model's accuracy on well-studied versus less-studied KEGG pathways. GeneLLM's contextual similarity, as determined by zero-shot learning, increased with information availability (see Figure 4, full figure in Appendix E), indicating prediction bias.

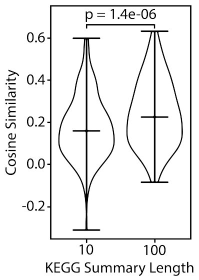


Figure 4: The effect of summary length on KEGG pathway predictions.

Secondly, for solubility prediction, genes with scarce online information showed worse model performance (Figure 5). This bias was mitigated by integrating GeneLLM's knowledge representation with data beyond text, as shown when GeneLLM embeddings combined with Gene2Vec embeddings narrowed the performance gap for under-researched genes (Figure 5). This suggests that representations from other modalities can complement text-derived information, and demonstrate a strategy to mitigate bias.

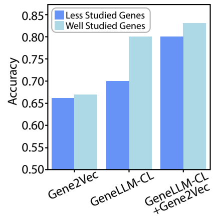


Figure 5: Multimodal fusion mitigate bias: GeneLLM's lower accuracy for solubility task of lesser-known genes, indicating knowledge bias, improves with Gene2Vec fusion.

5 Conclusion

We introduced GeneLLM, a model designed to predict gene and cell characteristics and to understand biological processes. GeneLLM was benchmarked against a diverse set of eight gene-related tasks, often outperforming task-specific models, as seen from the performance of foundation models in other domains. The use of contrastive learning has enhanced the ability of LLMs to predict a wide range of gene- and cell-related tasks and has enabled zero-shot predictions. By showcasing the interpretability of GeneLLM's predictions, we underscore the expanded utility and reliability of AI models, demonstrating that interpretability enhances the overall value of such technologies in biomedical applications and their potential for clinical adoption. This approach also provides evidence supporting the hypothesis that information contained in text is complementary to that found in structured databases. Overall, GeneLLM offers a pathway to enhance biomedical prediction capabilities while mitigating challenges related to AI model interpretability, and bias.

REFERENCES

10x Genomics. Single Cell Immune Profiling Dataset by Cell Ranger 3.1.0. https://support.10xgenomics.com/, July 2019. PBMCs from C57BL/6 mice (v1, 150x150).

Bruce Alberts. *Molecular biology of the cell*. Garland science, 2017.

José Juan Almagro Armenteros, Casper Kaae Sønderby, Søren Kaae Sønderby, Henrik Nielsen, and Ole Winther. Deeploc: prediction of protein subcellular localization using deep learning. *Bioinformatics*, 33(21):3387–3395, 2017.

Shokufeh Bagheri, Rasool Haddadi, Sahar Saki, Masoumeh Kourosh-Arami, and Alireza Komaki. The effect of sodium channels on neurological/neuronal disorders: A systematic review. *International Journal of Developmental Neuroscience*, 81(8):669–685, 2021.

Bradley E Bernstein, Tarjei S Mikkelsen, Xiaohui Xie, Michael Kamal, Dana J Huebert, James Cuff, Ben Fry, Alex Meissner, Marius Wernig, Kathrin Plath, et al. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, 125(2):315–326, 2006.

Coilin Boland, Samir Olatunji, Jonathan Bailey, Nicole Howe, Dietmar Weichert, Susan Kathleen Fetics, Xiaoxiao Yu, Javier Merino-Gracia, Clement Delsaut, and Martin Caffrey. Membrane (and soluble) protein stability and binding measurements in the lipid cubic phase using label-free differential scanning fluorimetry. *Analytical chemistry*, 90(20):12152–12160, 2018.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. Proteinbert: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38(8):2102–2110, 2022.

S K Burley and R G Roeder. Biochemistry and structural biology of transcription factor iid (tfiid). *Annual Review of Biochemistry*, 65:769–799, 1996.

M Cecilia Caino and Dario C Altieri. Molecular pathways: mitochondrial reprogramming in tumor progression and therapy. *Clinical Cancer Research*, 22(3):540–545, 2016.

Seth Carbon and C Mungall. Gene ontology data archive. *Dataset on Zenodo*, 2018.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, and Bo Wang. scgpt: Towards building a foundation model for single-cell multi-omics using generative ai. *bioRxiv*, 2023.

Christian Dallago, Konstantin Schütze, Michael Heinzinger, Tobias Olenyi, Maria Littmann, Amy X Lu, Kevin K Yang, Seonwoo Min, Sungroh Yoon, James T Morton, et al. Learned embeddings from deep learning to visualize and predict protein sets. *Current Protocols*, 1(5):e113, 2021.

Joseph M. de Guia, Madhavi Devaraj, and Carson K. Leung. Deepgx: deep learning using gene expression for cancer classification. In *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '19, pp. 913–920, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450368681. doi: 10.1145/3341161.3343516. URL https://doi.org/10.1145/3341161.3343516.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

Jingcheng Du, Peilin Jia, YuLin Dai, Cui Tao, Zhongming Zhao, and Degui Zhi. Gene2vec: distributed representation of genes based on co-expression. *BMC Genomics*, 20(1):82, Feb 2019a. ISSN 1471-2164. doi: 10.1186/s12864-018-5370-x.

Jingcheng Du, Peilin Jia, Yulin Dai, Cui Tao, Zhongming Zhao, and Degui Zhi. Gene2vec: distributed representation of genes based on co-expression. *BMC Genomics*, 20(1):82–, 2019b. ISSN 14712164. doi: 10.1186/s12864-018-5370-x. URL https://doi.org/10.1186/s12864-018-5370-x.

Michael R Dyson, Rajika L Perera, S Paul Shadbolt, Lynn Biderman, Krystyna Bromek, Natalia V Murzina, and John McCafferty. Identification of soluble protein fragments by gene fragmentation and genetic selection. *Nucleic acids research*, 36(9):e51–e51, 2008.

S Gatzemeier. Ai bias: Where does it come from and what can we do about it. *Data Science W231-Behind the Data: Humans and Values*, 2021.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. *CoRR*, abs/2006.07733, 2020. URL https://arxiv.org/abs/2006.07733.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *CoRR*, abs/2007.15779, 2020. URL https://arxiv.org/abs/2007.15779.

A James Hudspeth, Thomas M Jessell, Eric R Kandel, James Harris Schwartz, and Steven A Siegelbaum. *Principles of neural science*. McGraw-Hill, Health Professions Division, 2013.

Ting Jiang, Shaohan Huang, Zhongzhi Luan, Deqing Wang, and Fuzhen Zhuang. Scaling sentence embeddings with large language models, 2023.

Alexander Pritzel Tim Green Michael Figurnov Olaf Ronneberger Kathryn Tunyasuvunakool Russ Bates Augustin Žídek Anna Potapenko1 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 John Jumper, Richard Evans and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *nature*, 2021.

Quoc V. Le and Tomas Mikolov. Distributed representations of sentences and documents, 2014.

Monkol Lek, Konrad J Karczewski, Eric V Minikel, Kaitlin E Samocha, Eric Banks, Timothy Fennell, Anne H O'Donnell-Luria, James S Ware, Andrew J Hill, Beryl B Cummings, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285–291, 2016.

Zhijun Liao, Gaofeng Pan, Chao Sun, and Jijun Tang. Predicting subcellular location of protein with evolution information and sequence-based deep learning. *BMC bioinformatics*, 22:1–23, 2021.

Scott Lundberg. SHAP partition explainer. https://github.com/shap/shap/blob/master/shap/explainers/_partition.py, 2024.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Nick Barnes, and Ajmal Mian. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*, 2023.

Zhihua Ni, Xiao-Yu Zhou, Sidra Aslam, and Deng-Ke Niu. Characterization of human dosage-sensitive transcription factor genes. *Frontiers in genetics*, 10:1208, 2019.

Gherman Novakovsky, Nick Dexter, Maxwell W Libbrecht, Wyeth W Wasserman, and Sara Mostafavi. Obtaining genetics insights from deep learning via explainable artificial intelligence. *Nature Reviews Genetics*, 24(2):125–137, 2023.

Guilliermo Owen. Values of games with a priori unions. In Rudolf Henn and Otto Moeschlin (eds.), *Mathematical Economics and Game Theory*, pp. 76–88, Berlin, Heidelberg, 1977. Springer Berlin Heidelberg. ISBN 978-3-642-45494-3.

Ahmad Pesaranghader, Stan Matwin, Marina Sokolova, Jean-Christophe Grenier, Robert G Beiko, and Julie Hussin. deepSimDEF: deep neural embeddings of gene products and gene ontology terms for functional analysis of genes. *Bioinformatics*, 38(11):3051–3061, 05 2022. ISSN 1367-4803. doi: 10.1093/bioinformatics/btac304. URL https://doi.org/10.1093/bioinformatics/btac304.

Antara Anika Piya, Michael DeGiorgio, and Raquel Assis. Predicting gene expression divergence between single-copy orthologs in two species. *Genome Biology and Evolution*, 15(5):evad078, 2023.

Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018. URL https://api.semanticscholar.org/CorpusID:49313245.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

Ritika Ramani, Katie Krumholz, Yi-Fei Huang, and Adam Siepel. Phastweb: a web interface for evolutionary conservation scoring of multiple sequence alignments using phastcons and phylop. *Bioinformatics*, 35(13):2320–2322, 2019.

Laleh Seyyed-Kalantari, Haoran Zhang, Matthew BA McDermott, Irene Y Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature medicine*, 27(12):2176–2182, 2021.

Hashem A Shihab, Mark F Rogers, Colin Campbell, and Tom R Gaunt. Hipred: an integrative approach to predicting haploinsufficient genes. *Bioinformatics*, 33(12):1751–1757, 2017.

Thomas Stoeger, Martin Gerlach, Richard I. Morimoto, and Luís A. Nunes Amaral. Large-scale investigation of the reasons why potentially important genes are ignored. *PLOS Biology*, 16(9): 1–25, 09 2018. doi: 10.1371/journal.pbio.2006643. URL https://doi.org/10.1371/journal.pbio.2006643.

Christina V Theodoris, Ling Xiao, Anant Chopra, Mark D Chaffin, Zeina R Al Sayed, Matthew C Hill, Helene Mantineo, Elizabeth M Brydon, Zexian Zeng, X Shirley Liu, et al. Transfer learning enables predictions in network biology. *Nature*, pp. 1–9, 2023.

Vineet Thumuluri, José Juan Almagro Armenteros, Alexander Rosenberg Johansen, Henrik Nielsen, and Ole Winther. Deeploc 2.0: multi-label subcellular localization prediction using protein language models. *Nucleic Acids Research*, 50(W1):W228–W234, 2022.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. *CoRR*, abs/1910.10699, 2019. URL http://arxiv.org/abs/1910.10699.

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. Llama: Open and efficient foundation language models, 2023. URL http://arxiv.org/abs/2302.13971. cite arxiv:2302.13971.

Samuel J Virolainen, Andrew VonHandorf, Kenyatta CMF Viel, Matthew T Weirauch, and Leah C Kottyan. Gene–environment interactions and their impact on human health. *Genes & Immunity*, 24(1):1–11, 2023.

M. Viswanathan, C.D. Patnode, N.D. Berkman, et al. Assessing the risk of bias in systematic reviews of health care interventions. https://www.ncbi.nlm.nih.gov/books/NBK519366/, Dec 13 2017.

Kerstin N Vokinger, Stefan Feuerriegel, and Aaron S Kesselheim. Mitigating bias in machine learning for medicine. *Communications medicine*, 1(1):25, 2021.

Chao Wang and Quan Zou. Prediction of protein solubility based on sequence physicochemical patterns and distributed representation information with deepsolue. *BMC Biology*, 21, 01 2023a. doi: 10.1186/s12915-023-01510-8.

Chao Wang and Quan Zou. Prediction of protein solubility based on sequence physicochemical patterns and distributed representation information with deepsolue. *BMC biology*, 21(1):1–11, 2023b.

Guohua Wang, Fang Wang, Qian Huang, Yu Li, Yunlong Liu, Yadong Wang, et al. Understanding transcription factor regulation by integrating gene expression and dnase i hypersensitive sites. *BioMed research international*, 2015, 2015.

Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-llama: Towards building open-source language models for medicine, 2023.

Chunlei Wu, Ian MacLeod, and Andrew I Su. Biogps and mygene. info: organizing online, gene-centric information. *Nucleic acids research*, 41(D1):D561–D565, 2013.

Dehua Yang, Qingtong Zhou, Viktorija Labroska, Shanshan Qin, Sanaz Darbalaei, Yiran Wu, Elita Yuliantie, Linshan Xie, Houchao Tao, Jianjun Cheng, et al. G protein-coupled receptors: structure-and function-based drug discovery. *Signal transduction and targeted therapy*, 6(1):7, 2021.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding, 2020.

Ningyu Zhang, Zhen Bi, Xiaozhuan Liang, Siyuan Cheng, Haosen Hong, Shumin Deng, Jiazhang Lian, Qiang Zhang, and Huajun Chen. Ontoprotein: Protein pretraining with gene ontology embedding, 2022.

C Zhou, Q Li, C Li, J Yu, Y Liu, G Wang, K Zhang, C Ji, Q Yan, L He, et al. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. arxiv. *arXiv preprint arXiv:2302.09419*, 2023.

A Datasets Collection

Gene Summaries. The gene summary dataset was curated from two publicly accessible databases. We constructed a gene summary by concatenating the description of the gene obtained from [3] Wu et al. (2013), and the gene function obtained from UniProt into one unified summary. We then preprocess the dataset by removing certain keywords such as PubMed ID, author's name, isoform ID, and other identifying content. After preprocessing and removing duplicate summaries, there are a total of 14,450 remaining gene summaries.

Gene Ontology Data. We further enhance the gene representations by minimizing the distance to related Gene Ontology (GO) representations. We collected a total of 235,000 gene-to-GO annotations that span 18479 different GO terms. We hide 3000 GO terms and 68,000 annotations for testing the model performance. The Gene Ontology (GO) annotations were downloaded from AmiGO 2 website [4] Carbon & Mungall (2018). The preprocessing of GO term summaries was carried out in the same manner as that for gene summaries.

Single cell transcriptome Data. Single-cell transcriptome of Peripheral Blood Mononuclear Cells (PBMC) was downloaded from [5]. A total of 2700 cells with 24447 gene expression level each was downloaded, and only genes we have summary for were used from the dataset. After preprocessing, a total of 2,700 cells and 14,450 genes were used for the classification task.

[3] `https://mygene.info`

[4] `https://amigo.geneontology.org/amigo/dd_browse`

[5] `https://cf.10xgenomics.com/samples/cell/pbmc3k/pbmc3k_filtered_gene_bc_matrices.tar.gz`

Solubility Data. The solubility annotations dataset was obtained from a publication by Dallago et al. (2021), it contains 1499 protein sets labelled as either *Soluble* or *Membrane*. To obtain gene annotations, we use protein products where a gene is annotated *soluble* if one of its protein products is soluble, otherwise, it is labeled *membrane*.

KEGG Pathways. Similar to GO terms, we also test the model performance on KEGG pathway annotation. We obtained 1522 gene-to-KEGG annotations that span 263 KEGG pathways for training the model. We also held out a small set of 157 gene-to-KEGG relationships that span 30 KEGG pathways. The KEGG pathway annotations and summaries were collected from[6]. The preprocessing of KEGG summaries was performed in the same manner as that for the gene summaries.

B Gene-Related Tasks

B.1 Tasks

We evaluate the model performance on a variety of gene-related tasks, either classification or regression tasks. We now describe each task and the dataset collection sources.

Solubility. Distinguishing between membrane protein and soluble protein is important in Proteomics. Soluble proteins are a part of the cytosol, they can travel across membranes, and serve a wide range of functions both inside and outside the cells Dyson et al. (2008), while membrane proteins serve structural as well as functional roles in the cell as transporters, receptors, and enzymes. Over 50% of medications available on the market target membrane proteins, many of which have significant pharmacological implications Boland et al. (2018). We use GeneLLM to distinguish these classes for a given gene based on the textual description of its function. We curate `soluble` and `membrane` annotations of genes from previously reported datasets(Wang & Zou, 2023a).

Chromatin State Prediction. In epigenetic studies, Chromatin states are identified by modification of histones and methylations, which not only provide a basis for genes to be segmented into biologically meaningful units but also help determine their role in regulating gene expression. Bivalent Chromatin states, a hallmark of Embryonic Stem Cells (ESCs), are characterized by the presence of both repressive histone methylation-H3K27me3, the larger region, and activating histone methylation-H3K4me3, the smaller region. This Bivalent Chromatin structure marks developmental genes and maintains their promoters in a poised state, ready for activation during differentiation processes (Bernstein et al., 2006). We finetuned GeneLLM to classify between bivalent genes and genes that had either unmethylated promoters or were solely marked by H3K4me3, as reported by Theodoris et al. (2023). For our work, we utilize knowledge representations from the gene summaries instead of single-cell transcriptomic data. We used 184 selected annotations available in the datasets for the `bivalent` and `H3K4me3` classification task and 147 selected annotations available in the labeled datasets for the `bivalent` and `unmethylated` classification task from Theodoris et al. (2023) to finetune GeneLLM.

Dosage Sensitivity. Specific genes are said to be dosage sensitive when the variations in gene dosage (copy number variations) can cause phenotypic changes. For this task, we evaluate our model on a curated dataset of two annotations, namely dosage-sensitive and dosage-insensitive genes, from previously reported studies (Theodoris et al., 2023; Lek et al., 2016; Shihab et al., 2017; Ni et al., 2019).

Subcellular Localization. Understanding protein subcellular locations is essential for understanding its function and physiochemical properties. Computational methods are required in protein analysis research because traditional protein subcellular localization methods are laborious and time-consuming (Liao et al., 2021). Also, it can help identify possible targets for therapy and understand illnesses associated with abnormal subcellular localization (Thumuluri et al., 2022). We fine-tuned GeneLLM to distinguish between the subcellular localization of each gene. We utilize a set of gene annotations that spans the following 3 annotations: `Cytoplasm`, `Cell membrane`, and `Nucleus`, from the UniProt database, as described in Almagro Armenteros et al. (2017).

[6] `https://www.genome.jp/kegg/pathway.html`

Conservation. The PhastCon score is a measure derived from the PHAST (Phylogenetic Analysis with Space/Time models) package. It quantifies the evolutionary conservation of genomic sequences, offering insights into the functional significance and evolutionary pressures shaping these regions. Predicting PhastCon scores is essential for identifying functionally important genomic elements, such as coding sequences and regulatory regions, which are critical in evolutionary biology, functional genomics, and disease studies (Ramani et al., 2019). GeneLLM performance to predict Gene Conservation is measured by predicting the PhastonCon score with the help of knowledge representations of gene summaries obtained from GeneLLM as input. The Spearman and Pearson correlation coefficients are used to evaluate this regression task.

Transcription Factors (TFs) Range. TFs are the proteins that bind to certain sequences of DNA and control the transcription of genetic information from DNA to messenger RNA (Wang et al., 2015), with their influence ranging from short-range effects on nearby genes to long-range effects influencing distant genes. The dataset spans two different annotations, namely `short-range` and `long-range` genes, from Theodoris et al. (2023).

Transcription Factors (TFs) Target Type. We also study two specific transcription factors, GATA4 and TBX5 since they have crucial roles including heart development and function (Theodoris et al., 2023). We utilize annotations of genes that identify whether a gene is directly/indirectly regulated by one or both of these transcription factors. More specifically, we collect dataset annotations that span the following 5 labels: `gata4_indirect`, `gata4_direct`, `tbx5_indirect`, `tbx5_direct`, and `combo_targets`, from Theodoris et al. (2023).

B.2 Baselines

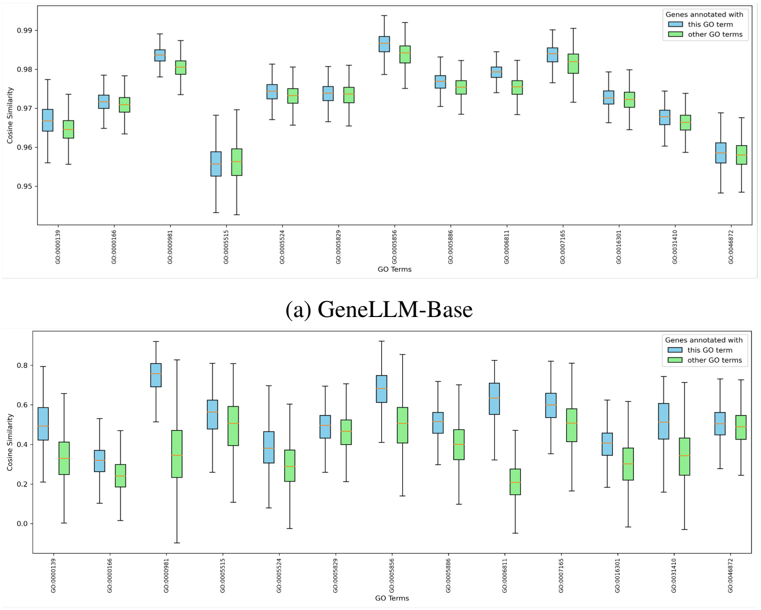
We conduct a comprehensive evaluation of our model against different baseline models. The baselines cover a wide selection of representation learning methods that are either trained on gene co-expression transcriptome, or text data. Below is a description of the proposed baselines:

- Majority Classifier: The most frequent class in the dataset is predicted for all genes. This classifier is chosen to show the distribution of the dataset and the hardness of the problem at hand.
- scGPT (Cui et al., 2023): is a transformer-based language model equipped with multi-head attention mechanisms, designed for gene and cell embedding tasks. This single-cell foundation model is pre-trained on 33 million normal human cells. We utilize the embedding from their largest pre-trained model (i.e. whole-human scGPT) to get the gene embeddings.
- Gene2Vec (Du et al., 2019a): we utilize the Gene2Vec embeddings that are trained on a wide range of gene co-expression datasets, thus absorbing rich and nuanced gene interactions and functions.
- Doc2Vec (Le & Mikolov, 2014) is a text-based embedding model that utilizes fixed-length feature representation to generate embeddings for variable-length text such as sentences, paragraphs, and documents. We get the embeddings by passing the gene summaries where we use an embedding size of 50, the maximum distance between the current and predicted word within a sentence of 2, all words with a total frequency of 1, and 40 training epochs.
- XLNet (Yang et al., 2020) is an autoregressive pretraining transformer-based model. We use the 12-layer xlnet-base-cased and CLS pooling to get the gene embeddings.
- PMC-LLaMA (Wu et al., 2023) is a LLaMA-based (Touvron et al., 2023) foundation language model that is pre-trained on the biomedical text and calibrated for medical domain applications. We get the embeddings from PMC-LLaMA by using Prompt-based last token pooling where we use the following prompt *"This sentence: {text} means in one word:[CLS]"* and utilize the contextualized embedding of the last token which would be the CLS token that is added by the tokenizer after the colon (Jiang et al., 2023).
- GPT-2 (Radford et al., 2019) is another open-source foundation language model that is trained on large text data and calibrated for downstream applications. We get the gene embeddings from the encoder of GPT2 by performing CLS pooling on the gene summaries.

The proposed baselines generate summary embeddings that are not task-specific, we therefore tailor our analysis to downstream tasks by augmenting a Linear/Logistic Regression (LR) on top of the embeddings.

Table 2: 5 fold CV Model performance metrics with accuracy and standard deviation on Cell type classification.

Model	GeneLLM-CL	GeneLLM-Base	GPT-2
SGDClassifier	**0.85 ± 0.02**	0.82 ± 0.02	0.79 ± 0.03
PassiveAggressive	**0.84 ± 0.01**	0.83 ± 0.02	0.79 ± 0.03
Perceptron	**0.83 ± 0.01**	0.83 ± 0.04	0.76 ± 0.04
LGBM	**0.82 ± 0.02**	0.81 ± 0.01	0.79 ± 0.02
XGB	0.81 ± 0.02	**0.81 ± 0.01**	0.79 ± 0.02
ExtraTrees	**0.77 ± 0.01**	0.72 ± 0.02	0.72 ± 0.02
RandomForest	**0.77 ± 0.01**	0.74 ± 0.01	0.73 ± 0.02
Bagging	**0.73 ± 0.02**	0.71 ± 0.02	0.69 ± 0.03
KNeighbors	**0.73 ± 0.01**	0.70 ± 0.01	0.69 ± 0.01
DecisionTree	**0.63 ± 0.03**	0.59 ± 0.04	0.57 ± 0.04
NearestCentroid	**0.63 ± 0.02**	0.51 ± 0.01	0.43 ± 0.02
GaussianNB	**0.62 ± 0.01**	0.48 ± 0.03	0.42 ± 0.04
BernoulliNB	**0.57 ± 0.02**	0.41 ± 0.03	0.35 ± 0.02
ExtraTree	**0.54 ± 0.03**	0.51 ± 0.02	0.48 ± 0.03
AdaBoost	**0.43 ± 0.07**	0.41 ± 0.04	0.39 ± 0.01



(a) GeneLLM-Base

(b) GeneLLM-CL

Figure 6: Discovering novel gene-to-GO term relationships by employing contrastive learning. Be aware that the scales of the two figures are not the same. GeneLLM-Base restricts the similarities within a narrower range, making the results less comparable, whereas GeneLLM-CL spreads the gene similarities throughout the entire y-axis, enhancing the distinction between GO terms and genes.

C Cell-Type Classification via Cell Embeddings

Table 2 presents a comprehensive comparison between the two calibrated cell embeddings generated by GeneLLM and those generated by GPT-2. GeneLLM-CL outperforms both baselines across the classifiers mentioned in the table. In this detailed analysis, GeneLLM-CL demonstrates superior performance in cell-type predictions by utilizing contrastive learning.

D Zero-Shot Learning

To enable zero-shot ability, we encapsulate gene embeddings by co-embedding GO-gene relationships. In our initial analysis, we utilize 457,000 relationships, however, after removing entries that involve missing values, the total number of relationships is reduced to 392,000. We further consider relevant relationships that strictly include the genes in our datasets which resulted in a final count of 235,000 significant relationships. Figure 6 illustrates the expanded list of GO terms and their related genes (blue) and non-related genes (green). The list spans the GO terms that strictly have more than 500 related genes. The relationship between GO terms and related genes is better emphasized when the contrastive learning objective is applied. The model also shows inconsistent prediction capacity for some GO terms (e.g. GO:0000981 and GO:0006811 versus GO0005829). We hypothesize that this is related to the length of summaries since these GO terms have smaller summaries compared to the overall distribution.

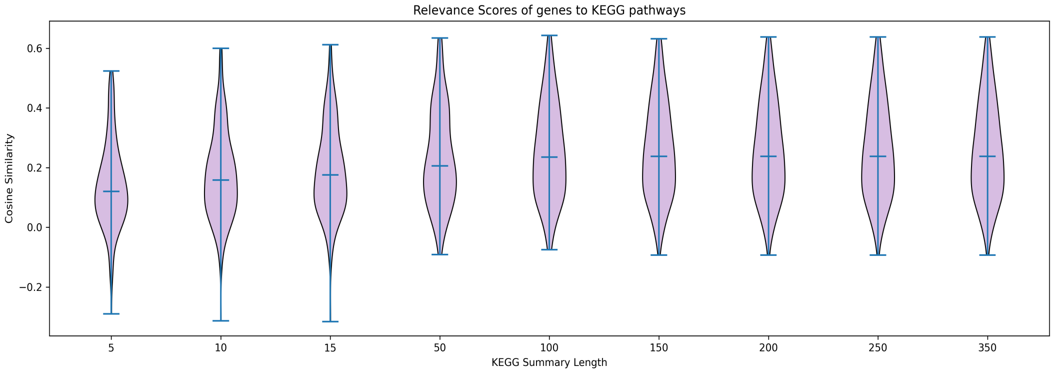


Figure 7: Performance of zero-shot learning on KEGG pathways reveals the significance of available knowledge on performance. This detailed study showcases how variations in informational input affect learning outcomes without direct training on KEGG pathways.

E Bias Mitigation via Multimodal Fusion

Figure 7 presents the full KEGG pathway predictions from GeneLLM-CL. The figure illustrates the relevance scores of our model on different summary lengths of the KEGG pathways. The average similarity for each KEGG pathway and relevant genes (D_r) and non-relevant genes (D_{nr}) are calculated using cosine similarity. We then calculate a relevance score for a KEGG pathway as the difference between D_r and D_{nr} (i.e. $D_r - D_{nr}$). We train the model with different summary lengths of 10 and 100 words per KEGG summary. The figure shows that when the amount of available information increases, the model performance is increased. The model can significantly diminish the level of bias in its predictions by learning from longer text summaries thus leading to more reliable outcomes.

F Solubility Clustering and Prediction

Solubility Prediction Apart from the baseline comparison for predicting solubility in Section 4, We compare the finetuned GeneLLM-CL against solubility-specific models. Table 3 illustrates our model performance compared to the solubility models mentioned in Wang & Zou (2023b).

Table 3: A comprehensive evaluation of gene products solubility of Methods from DeepSoluE (Wang & Zou, 2023b)

Model	**F1 Score**	**Accuracy**
RPSP	0.392	0.498
ccSOL omics	0.537	0.508
SKADE	0.168	0.492
SOLpro	0.468	0.52
Protein-Sol	0.585	0.516
DeepSol	0.239	0.529
rWH	0.485	0.54
ESPRESSO	0.583	0.538
CamSol	0.487	0.541
SWI	0.638	0.559
PROSO II	0.491	0.58
SoluProt	0.593	0.585
DeepSoluE	0.600	0.5952
GeneLLM-CL	0.81	0.81
GeneLLM-CL + Finetuned	**0.89**	**0.887**

GeneLLM-CL significantly outperformed the proposed baselines on the same set of annotations but features extracted from text summaries and GO term relationships, showcasing the effectiveness of contrastive learning in forecasting gene solubility.

Clustering Analysis. Clustering and enrichment analysis of gene embeddings from GeneLLM fine-tuned from solubility produced interpretable results. Gene clustering is based on embedding done using the shared nearest neighbor algorithm. Pathway enrichments were done on each cluster and labeled as the most enriched pathway. The cluster labeled as ”Cell-Cell communication” ”Neuronal System” and ”GPCR ligand binding” along with other genes related to membrane function were predicted by the model as membranes, as shown in Figure 2c. This is even though the model was trained on a limited dataset and was not directly exposed to these annotations (*membrane* vs *soluble*) during training, instead they were encoded as 0’s and 1’s. For example, a majority of the genes in the Neuronal System are involved in signal transduction. These genes encode a variety of proteins such as neurotransmitter receptors, ion channels, signaling enzymes, and other molecules involved in signal transmission and processing in the nervous system, which are majorly membrane protein functions (Hudspeth et al., 2013).

Interpretability. Our model interpretability analysis, utilizing SHAP, unveiled significant terms associated with each cluster, as illustrated in Figure 2c. Through SHAP analysis, we were able to directly link various terms to the function of each cluster in the membrane and soluble genes clusters. For instance, the term *7Transmembrane* is one of the most important words within the *GPCR Ligand Binding* cluster, underscoring the characteristic feature of GPCRs traversing the membrane seven times, a structural hallmark facilitating their critical signaling roles (Yang et al., 2021). The importance of the word *Sodiumdependent* aligns with the known dependency of the neuronal system on sodium ions (Na^+) for critical functions such as action potential propagation and neurotransmitter regulation, and their role in the cell membrane (Bagheri et al., 2021). Similarly, *transporters* was identified as one of the most important words in *(TCA) cycle & respiratory electron transport* cluster, this reflects the vital role of these membrane proteins in facilitating the biochemical pathways essential for the TCA cycle and electron transport chain activities (Caino & Altieri, 2016). The word *DNAbinding* in the context of *HIV Infection* relates to non-membrane proteins with DNA-binding properties such as TAF1 that function within the nucleus where HIV is active (Burley & Roeder, 1996).