Representation of biomedical knowledge in transformer-based deep neural networks

Alexandra Elbakyan

2025

Table of Contents

Problem Definition and Background

Biomedical knowledge is important for understanding of human diseases, as it enables precise and timely diagnostics and development of effective treatments for various conditions. However, even though some part of biomedical knowledge is available in the form of structured, machine-readable databases, the major part remains represented as unstructured text, i.e. as a publications in research journals and books. PubMed, the largest index of biomedical literature, includes more than 37 million citations to date. Recent advances in natural language processing made possible large-scale, automatic extraction, analysis and synthesis of that knowledge, further advancing our understanding of the inner workings of human body.

One of the earliest successfull experiments was data mining over 70,000 research abstracts performed on IBM Watson supercomputer in 2014 [1]. Using TF-IDF metrics, algorithm was able to discover previously unknown protein kinases for TP53 — a tumour suppressor gene that is often damaged or missing in cancer patients. The discovery was possible because embedding vectors of P53-phosphorylating protein kinases tended to cluster with already-known molecules with similar properties.

In the past few years, transformer deep neural networks have revolutionized the field of computational language processing, providing a more advanced approach to automated knowledge extraction and representation. Transformer networks are pre-trained on a large amounts of textual data and after that can be fine-tuned to perform specific task such as text classification, named entity recognition or relation extraction.

In this work, transformer neural networks that were pre-trained on large volumes of biomedical texts will be applied to the following tasks:

- cancer driver gene identification
- treatment outcome prediction.

Cancer driver gene identification. While tumour cells typically carry a lot of mutations, only alterations in some genes are important in developing cancer. These are primarily genes that are involved in cell division process, either encouraging cell to multiply (oncogenes e.g. EGFR or KRAS) or blocking further multiplication (tumour suppressor genes e.g. TP53).

Methods for identification of cancer-driver genetic mutations is an active area of research [2]. Identifying genes involved in cancer enables development of targeted and even personalized therapies, that have come to play an increasingly important role over the past two decades [3]. Unlike chemotherapy or radiotherapy, that affect both healthy and malignant cells, targeted therapies act in a more precise way, producing rapid tumour regression without inflicting toxic damage to healthy tissues.

In 2004, a list of 291 cancer genes was compiled based on manual analysis of published literature: only those mutations that are highly unlikely to be due to chance were included [4]. More recently, different computational tools have been developed that attempt to identify driver genes based on large-scale analysis of patient tumour samples, biological and chemical properties of amino acids, networks of interactions and etc.; many approaches rely on machine learning algorithms, such as SVM or random forests [5].

COSMIC Cancer Gene Census is an expert-curated dataset that currently lists 728 genes driving human cancer [6]. The list is not exhaustive, as it includes only those genes that have enough evidence support to date. Therefore, computational tools can identify a larger number of potential drivers: for example, automated analysis of 8000+ tumour samples performed by Moonlight revealed 1000+ cancer driver genes [7].

In the first task, transformer neural networks will be trained to predict whether a specific gene can be a cancer-driver gene or not, based on knowledge available in published biomedical literature.

Treatment outcome prediction. Drug response prediction is another active area of research where AI methods play an important role. Being able to predict response to drug is especially important in cancer therapy, because tumours associated with specific mutations will respond only to some treatments while remaining resistant to others.

Deep-learning systems that rely on features such as gene expression profile or molecular structure have shown to be quite effective [8,9]. In the second task I will test whether biomedical language models can also be used to predict treatment outcome for specific drug and genetic profile in cancer. That opens potential for the range of applications, including automatic generation of personalized treatment plans.

Datasets

Datasets used for training and evaluation:

- COSMIC, the Catalogue Of Somatic Mutations In Cancer [6]
- Diseases database of gene-disease connections, collected by text-mining of Pubmed abstracts and full-texts [10]
- Human Disease Ontology, a classification of human diseases [11]
- CIViC, Clinical Interpretation of Variants in Cancer, a community-curated knowledge base of associations between genetic mutations, drugs and treatment outcomes assembled from literature [12].

Evaluation Framework

Cancer gene identification. Two balanced datasets with equal number of positive and negative samples were created. Performance of different models was evaluated using 4-fold cross-validation. In the first round performance of models was estimated on a simple baseline task: classification between actual gene names and randomly generated sequences. Each model was trained over 8 epochs, and maximum, median and minium accuracy scores reached in every epoch were compared. In adition to that, training time required by each model to reach maximum accuracy was estimated. In the second round, models that performed best on baseline task were trained to do classification between cancer driver gene vs. gene not associated with cancer.

Treatment outcome prediction. Language model was prompted to answer «yes» «no» or «maybe» to the question asking if some treatment will help patient with a specific genetic mutation and diagnosis. The number of different answers was recorded along with percentage of correct answers.

Approach & Justification

The rationale for using large language models trained on biomedical texts to do cancer gene identification comes from the fact that new knowledge can often be derived from information that is already known and published. Literature-mining have been used in the past to discover previously unknown entities with specific properties, such as p53-phosphorylating protein kinases [1] or materials with thermoelectric properties identified by word2vec embeddings [13].

There are several biomedical language models currently available. These models differ in architecture, number of parameters and training dataset. BERT architecture is most numerous, including BioBERT [14], SciBERT [15], BiomedBERT [16] and TinyPubMedBERT, as well as BioMegatron [17]. Furthermore, each of the models listed has different versions: cased or uncased, basic or large. That yields 11 BERT-based models in total. LlaMA architecture has older PMC-LlaMA with 13B parameters and newer MedLLaMA with 8B parameters [18]. There are also two versions of BioGPT (basic and large) [19].

To select the best performing model, all models (19 in total) were initially evaluated on a basic baseline task: to differentiate between name of a cancer driver gene and randomly generated string of similar length (for BERT and LlaMA, I also included base model into evaluation) and after baseline performance were established, models with highest accuracy scores were trained to do cancer gene identification.

Dataset preparation. The list of known cancer driver genes was extracted from COSMIC dataset, providing 753 gene names in total. A gene name is a string consisting from 2 to 9 characters with average length of 4.8 and median 5. These were used as positive samples. Then two datasets were created, with two different kinds of negative samples:

(a) random alphanumeric sequences of the same length (2-9 characters with average length 4.77 and median 5)

(b) names of genes that are not associated with cancer: these were extracted from Diseases dataset. The dataset provides ~26K gene names mined from literature. Each gene-disease association in dataset has confidence score from 0 to 4, where 1 is equal to one standard deviation from random, and 4 represents well-known and established associations. For cancer diseases, number of genes at confidence level was:

≥ 1: 21299
≥ 2: 5767
≥ 3: 844

I excluded genes at level ≥ 2 from negative samples, as they can potentially be drivers, but do not yet have enough evidence. Gene names with non-alphanumeric symbols were also filtered, as they do not occur in positive subset of known driver genes. The list of samples was additionally calibrated for length distribution.

The number of negative samples was selected (generated) to be the same as positive (753) resulting in two balanced datasets for training and evaluation.

Experiments & Results

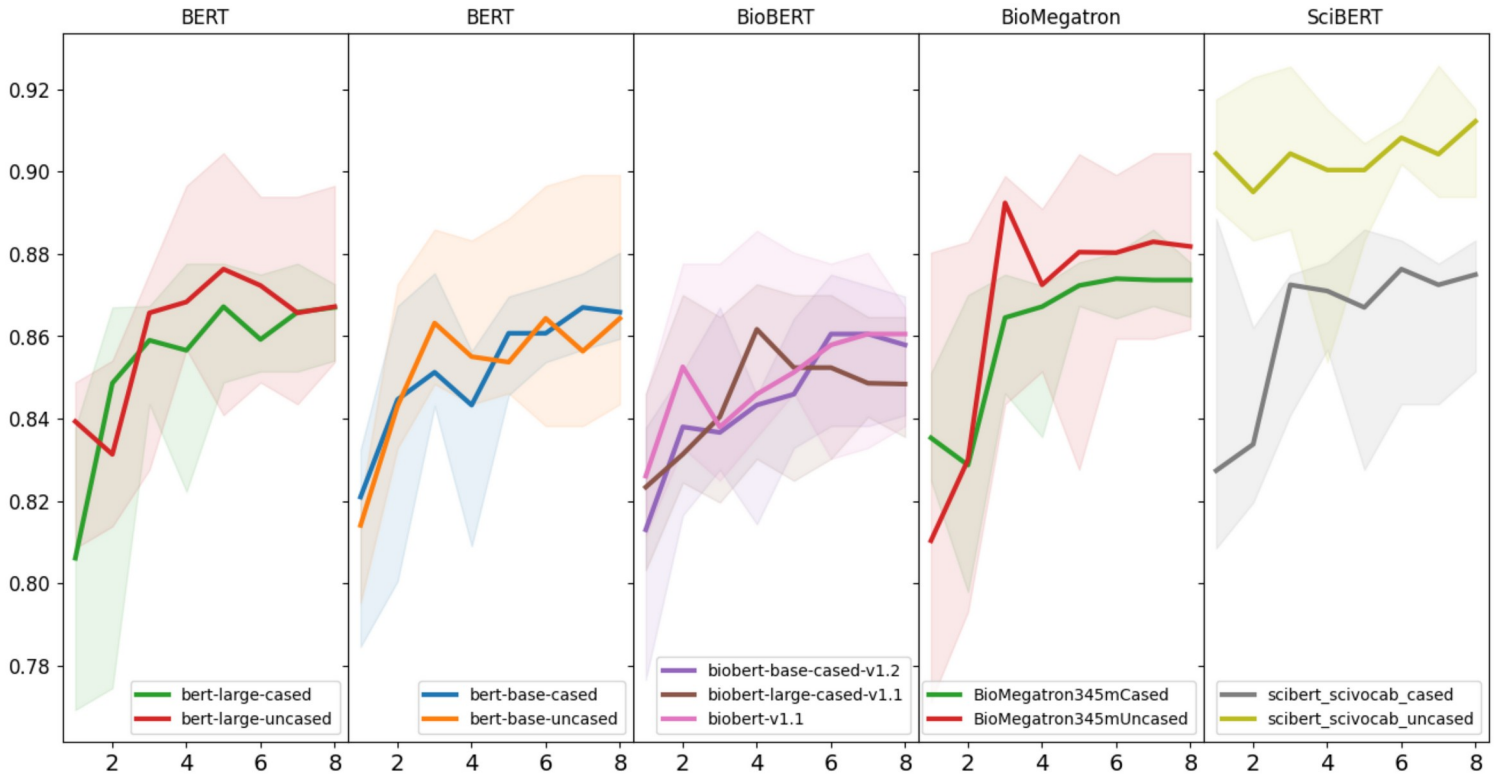

Figure 1. Accuracy scores on validation dataset achieved by BERT-mase models over N training epochs accross K folds, N = 8, K = 4. Line represents median accuracy score in a single epoch. Area between minimum and maximum accuracy is filled.

Gene vs. random sequence classification. Accuracy scores on validation dataset for BERT models is shown on Figures 1, 2 and Table B1 (Appendix B). Surprisingly, base BERT models without domain adaptation achieved decent accuracy scores around 0.86 on average; uncased variants performed slightly better than cased, even though difference was very small. Potential explanation for this might be, that distribution of characters in randomly generated sequences is statistically different from that of real gene names, or that BERT model received some biomedical texts in input data among others and was able to harness that knowledge.

BioMegatron achieved accuracy levels only a little higher than base model, while performance of BioBERT in all variants was somewhat lower. TinyPubMedBERT got the worst results, with maximum accuracy not reaching higher than 0.82.

SciBERT and BiomedBERT became the winners among BERT-based models, reaching scores of 0.90 and higher. In case of SciBERT the difference between performance of cased and uncased variant was very significant. Interestingly, BiomedBERT model with smaller number of parameters trained on full texts performed better than larger model trained on abstracts only, although the difference was not significant. It can also be noted that while other models achieved best results after 3-4 epochs, both SciBERT (uncased) and BiomedBERT performed good classification right from the beginning and did not learn much with training.

LlaMA and GPT models (Figure 3) performed significantly better than BERT, with top result around 0.94 shown by large version of BioGPT; the same flat training curve can be noticed here. LlaMA models display a much smoother learning curve; here PMC-LlaMA with 13B parameters became a winner (0.95) and overcome newer 8B models based on LlaMA 3. Surprisingly, biomedical LlaMA 3 performed worse than corresponding base model.

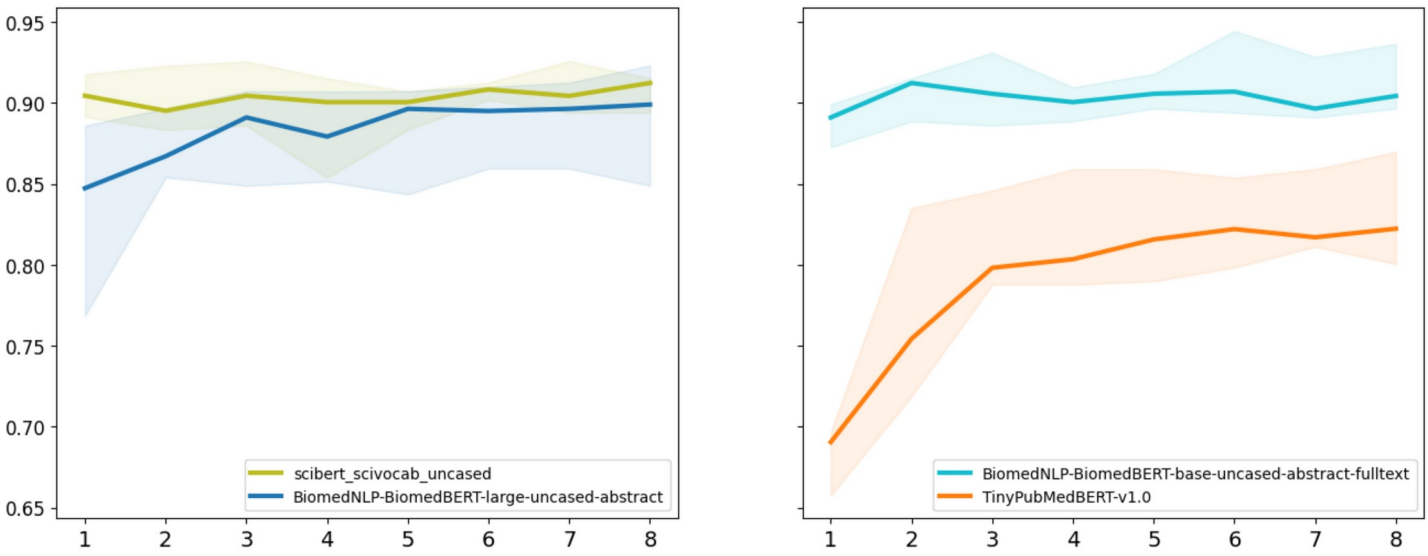

Figure 2. Accuracy scores on validation dataset achieved by BiomedBERT, SciBERT and TinyPubMedBERT over N training epochs accross K folds, N = 8, K = 4. Line represents median accuracy score in a single epoch. Area between minimum and maximum accuracy is filled.

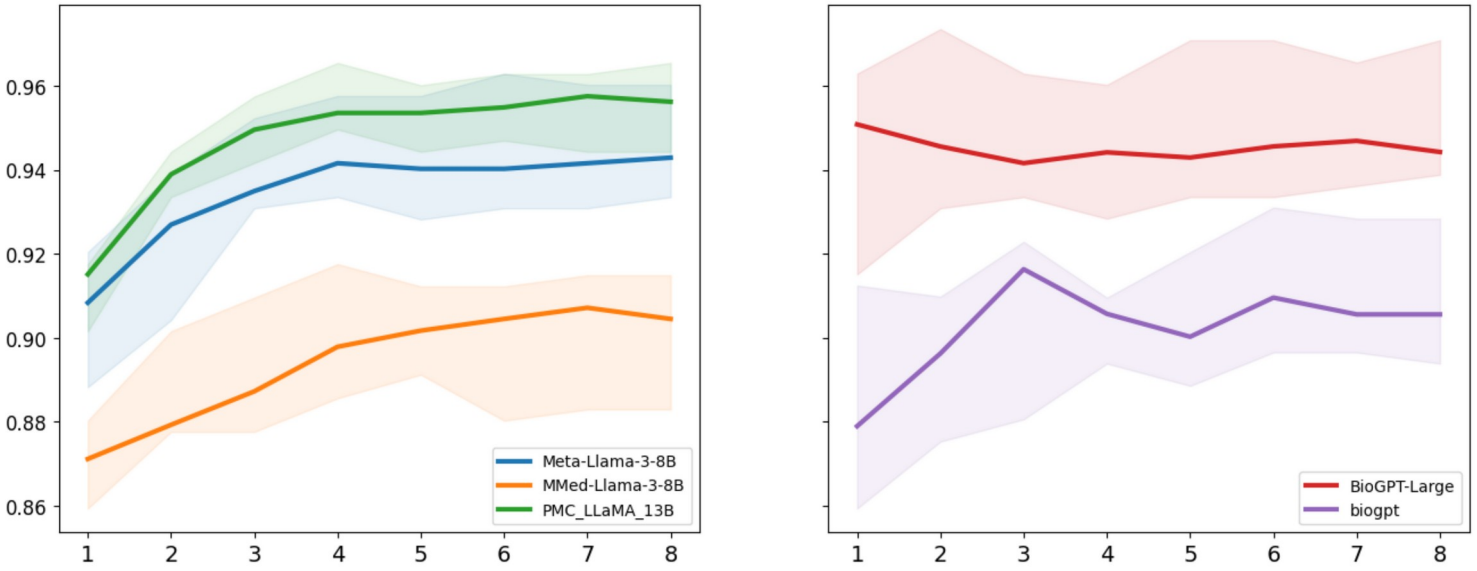

Figure 3. Accuracy scores on valiation dataset achieved by LlaMA and GPT models over N training epochs accross K folds, N = 8, K = 4.

Cancer driver vs. non-driver gene classification. Based on results of baseline task, following models were selected: SciBERT (cased), BioMedBERT (base, fulltext), PMC-LlaMA (13B) and BioGPT (large). Performance of all four models on both classification tasks are available in Appendix A and Table B2 (Appendix B). While both GPT and LlaMA performed better than BERT on both tasks, higher scores (0.80) were achieved by GPT on second task but not first. This, combined with the fact that BioGPT takes around 1min to train, which is almost 10x times faster than LlaMA, makes this model a clear winner.

Treatment outcome prediction. BioGPT-QA model was tested on CIVIC dataset using QA template, but the testing did not yield good result.

Ultimate Judgment, Analysis and Limitations

Performance on classification task has confirmed general understanding of LlaMA and GPT as more advanced models than BERT. That conclusion might seem trivial, given that these models have significantly larger number of parameters. However, the same performance metrics show that larger model is not always more efficient: results from base and large BERT were practically the same; base BiomedBERT trained on fulltext performed better than large one trained on abstracts only; and BioGPT outperformed PMC-LlaMa despite being 8 times smaller. Model architecture and training methods and dataset are equally important factors.

That specialized biomedical models were better in classification task than general-purpose LLMs shows that these models indeed learn some knowledge while training and do not rely simply on statistical features of input data. However, LLMs are not yet precise or accurate enough to enable, for example, automated medicine recommendations in complex cases. Still, these models can find applications in research, such as generating new hypotheses to be tested (i.e. suggest genes that might be potential targets for new therapies).

References

[1] S. Spangler et al., “Automated hypothesis generation based on mining scientific literature,” in Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, New York New York USA: ACM, Aug. 2014, pp. 1877–1886. doi: 10.1145/2623330.2623667.

[2] M. Nourbakhsh, K. Degn, A. Saksager, M. Tiberti, and E. Papaleo, “Prediction of cancer driver genes and mutations: the potential of integrative computational frameworks,” Briefings in Bioinformatics, vol. 25, no. 2, p. bbad519, Jan. 2024, doi: 10.1093/bib/bbad519.

[3] M. R. Waarts, A. J. Stonestrom, Y. C. Park, and R. L. Levine, “Targeting mutations in cancer,” Journal of Clinical Investigation, vol. 132, no. 8, p. e154943, Apr. 2022, doi: 10.1172/JCI154943

[4] P. A. Futreal et al., “A census of human cancer genes,” Nat Rev Cancer, vol. 4, no. 3, pp. 177–183, Mar. 2004, doi: 10.1038/nrc1299.

[5] D. Ostroverkhova, T. M. Przytycka, and A. R. Panchenko, “Cancer driver mutations: predictions and reality,” Trends in Molecular Medicine, vol. 29, no. 7, pp. 554–566, Jul. 2023, doi: 10.1016/j.molmed.2023.03.007.

[6] Z. Sondka, S. Bamford, C. G. Cole, S. A. Ward, I. Dunham, and S. A. Forbes, “The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers,” Nat Rev Cancer, vol. 18, no. 11, pp. 696–705, Nov. 2018, doi: 10.1038/s41568-018-0060-1.

[7] A. Colaprico et al., “Interpreting pathways to discover cancer driver genes with Moonlight,” Nat Commun, vol. 11, no. 1, p. 69, Jan. 2020, doi: 10.1038/s41467-019-13803-0.

[8] S. Chawla et al., “Gene expression based inference of cancer drug sensitivity,” Nat Commun, vol. 13, no. 1, p. 5680, Sep. 2022, doi: 10.1038/s41467-022-33291-z.

[9] P. Li, Z. Jiang, T. Liu, X. Liu, H. Qiao, and X. Yao, “Improving drug response prediction via integrating gene relationships with deep learning,” Briefings in Bioinformatics, vol. 25, no. 3, p. bbae153, Mar. 2024, doi: 10.1093/bib/bbae153.

[10] D. Grissa, A. Junge, T. I. Oprea, and L. J. Jensen, “Diseases 2.0: a weekly updated database of disease–gene associations from text mining and data integration,” Database, vol. 2022, p. baac019, Mar. 2022, doi: 10.1093/database/baac019.

[11] L. M. Schriml et al., “The Human Disease Ontology 2022 update,” Nucleic Acids Research, vol. 50, no. D1, pp. D1255–D1261, Jan. 2022, doi: 10.1093/nar/gkab1063.

[12] M. Griffith et al., “CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer,” Nat Genet, vol. 49, no. 2, pp. 170–174, Feb. 2017, doi: 10.1038/ng.3774.

[13] V. Tshitoyan et al., “Unsupervised word embeddings capture latent knowledge from materials science literature,” Nature, vol. 571, no. 7763, pp. 95–98, Jul. 2019, doi: 10.1038/s41586-019-1335-8.

[14] J. Lee et al., “BioBERT: a pre-trained biomedical language representation model for biomedical text mining,” Bioinformatics, vol. 36, no. 4, pp. 1234–1240, Feb. 2020, doi: 10.1093/bioinformatics/btz682.

[15] I. Beltagy, K. Lo, and A. Cohan, “SciBERT: A Pretrained Language Model for Scientific Text,” 2019, doi: 10.48550/ARXIV.1903.10676.

[16] S. Chakraborty, E. Bisong, S. Bhatt, T. Wagner, R. Elliott, and F. Mosconi, “BioMedBERT: A Pre-trained Biomedical Language Model for QA and IR,” in Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain (Online): International Committee on Computational Linguistics, 2020, pp. 669–679. doi: 10.18653/v1/2020.coling-main.59.

[17] H.-C. Shin et al., “BioMegatron: Larger Biomedical Domain Language Model,” in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online: Association for Computational Linguistics, 2020, pp. 4700–4706. doi: 10.18653/v1/2020.emnlp-main.379.

[18] C. Wu, W. Lin, X. Zhang, Y. Zhang, W. Xie, and Y. Wang, “PMC-LLaMA: toward building open-source language models for medicine,” Journal of the American Medical Informatics Association, vol. 31, no. 9, pp. 1833–1843, Sep. 2024, doi: 10.1093/jamia/ocae045.

[19] R. Luo et al., “BioGPT: generative pre-trained transformer for biomedical text generation and mining,” Briefings in Bioinformatics, vol. 23, no. 6, p. bbac409, Nov. 2022, doi: 10.1093/bib/bbac409.

APPENDIX A

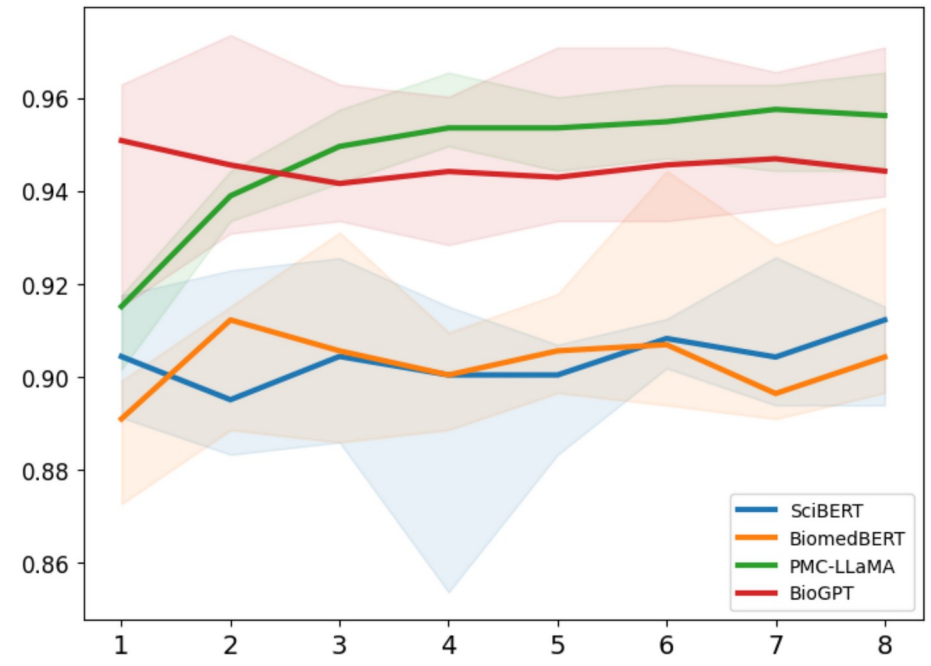

Figure A1.

Accuracy of top-performing models over 8 training epochs accross 4 folds on gene vs. random sequence classification task.

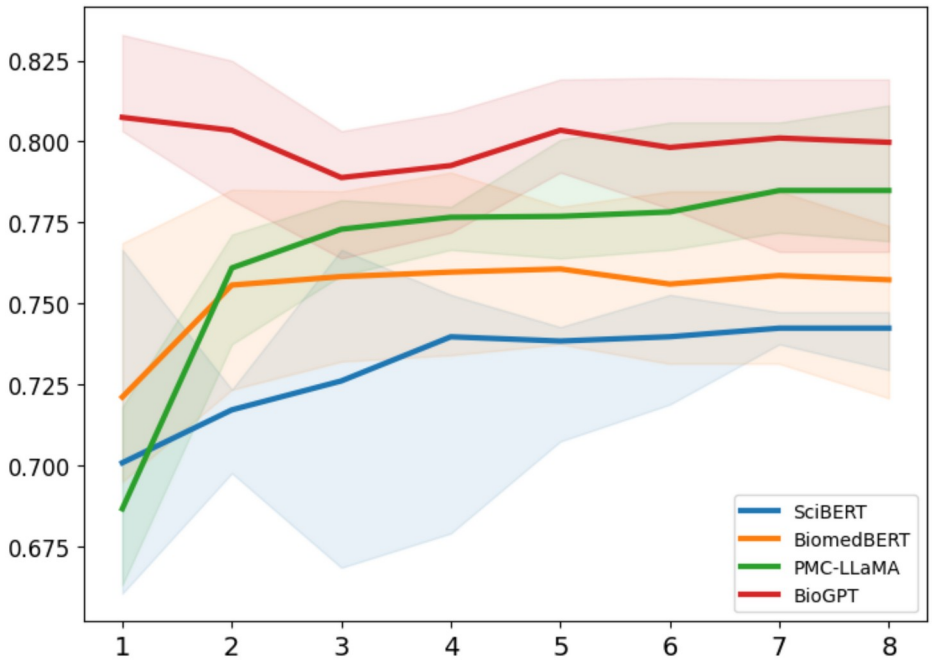

Figure A2.

Accuracy of top-performing models over 8 training epochs accross 4 folds on cancer driver gene vs. gene not associated with cancer.

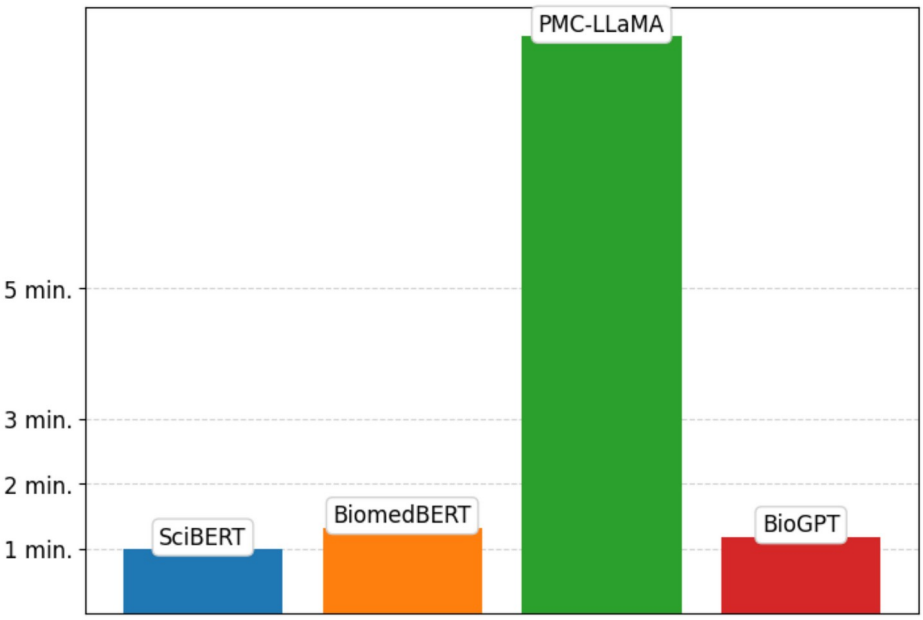

Figure A3.

Average training time, accross 4 folds, to reach maximum accuracy score, for top-performing models.

APPENDIX B

	model	params	time	acc med	acc std	acc min	acc max
1	PMC-LLaMA 2 13B	13.1B	7 min	0.96	0.01	0.95	0.97
2	BioGPT-large	1.6B	1 min	0.95	0.01	0.95	0.97
3	LLaMA 3 8B	8.1B	4 min	0.94	0.01	0.93	0.96
4	SciBERT uncased	109.9M	2 min	0.92	0.01	0.9	0.93
5	BioGPT	346.8M	48 sec	0.92	0.01	0.9	0.93
6	BiomedBERT fulltext	109.5M	2 min	0.91	0.01	0.9	0.94
7	MMed-LLaMA 3 8B	8.1B	4 min	0.9	0.01	0.9	0.92
8	BiomedBERT abstract	335.1M	7 min	0.9	0.02	0.86	0.92
9	BioMegatron uncased	335.2M	8 min	0.9	0.02	0.86	0.9
10	SciBERT cased	109.9M	2 min	0.88	0.01	0.86	0.89
11	BERT-large cased	333.6M	5 min	0.88	0.01	0.87	0.88
12	BERT-large uncased	335.1M	5 min	0.88	0.02	0.86	0.9
13	BioMegatron cased	333.6M	9 min	0.88	0.01	0.87	0.89
14	BERT cased	108.3M	2 min	0.87	0.01	0.86	0.88
15	BERT uncased	109.5M	2 min	0.87	0.02	0.86	0.9
16	BioBERT cased	108.3M	2 min	0.86	0.01	0.84	0.88
17	BioBERT-large cased	108.3M	2 min	0.86	0.02	0.85	0.89
18	BioBERT uncased	364.3M	5 min	0.86	0.01	0.85	0.87
19	TinyPubMedBERT	14.4M	30 sec	0.82	0.02	0.82	0.87

Table B1. Accuracy scores on valiation dataset for all models on gene vs. random sequence classification task over N training epochs accross K folds, N = 8, K = 4. Models are sorted by median max. accuracy in all folds descending, training time ascending.

	model	params	time	acc med	acc std	acc min	acc max
1	BioGPT-large	1.6B	2 min	0.82	0.01	0.8	0.83
2	PMC-LLaMA 2 13B	13.1B	10 min	0.79	0.01	0.77	0.81
3	BiomedBERT fulltext	109.5M	2 min	0.76	0.02	0.74	0.79
4	SciBERT uncased	109.9M	1 min	0.74	0.01	0.74	0.77

Table B2. Accuracy scores on valiation dataset for selected models on cancer driver vs. other gene classification task.