Development of a Machine Learning Tool to Identify Biochemical Features of Proteins

Hala Arar Zheng He Yining Zhou Mingyang Zhang

2025-05-09

Executive Summary

This proposal outlines the development of a machine learning tool aimed at identifying and cataloging biochemical features of proteins, particularly focusing on autoregulatory mechanisms. Collaborating with the University of Helsinki, we will use a combination of BERT and domain-specific variants like BioBERT to create a tool capable of analyzing biomedical text and identifying autoregulatory features. The tool will help in drug development and disease modeling by providing researchers with an efficient method to detect self-regulating proteins, thus enabling the identification of new drug targets and expanding knowledge on disease mechanisms. The final deliverables will include a fine-tuned machine learning model and an interactive Shiny app for querying model predictions.

Introduction

We are excited to collaborate with the University of Helsinki, one of the oldest and most prestigious universities in Finland that's renowned for its contributions to research and consistently ranking in the top 1% of global university rankings (Helsinki (n.d.)). More specifically, we are working with Dr. Mohieddin Jafari, a principal investigator with Jafari Labs research group. This interdisciplinary project brings together research from the Department of Biochemistry and Developmental Biology and data science techniques. The department specializes in systems pharmacology, drug development, and disease modeling.

Cells have the remarkable ability to regulate themselves through various mechanisms. Self-loops are simple feedback mechanisms within biological systems where a molecule or gene influences itself. Autoregulation is a key aspect of cellular regulation, where certain proteins control their own expression by binding to their own promoters (Bateman (2008)). This process is vital in regulating a range of transcription factors, which are essential for various cellular functions, including cell cycle control, inducible responses, and cell type-specific activities (Bateman (2008)).

Despite the crucial role that self-loops play in regulating cellular processes, we currently lack a clear, organized method for tracking these self-loops in biological networks. Without a comprehensive database, it is difficult to understand how proteins influence vital functions like cell growth, metabolism, and disease development. To solve this problem, we are developing a machine learning tool that will identify and catalog these self-regulatory features in proteins. This tool will not only help us identify self-regulating proteins but also discover new drug targets. Understanding self-loops is crucial for understanding disease mechanisms allowing researchers to better grasp how these loops affect cellular behavior and contribute to diseases.

Data Science Techniques

Data Description

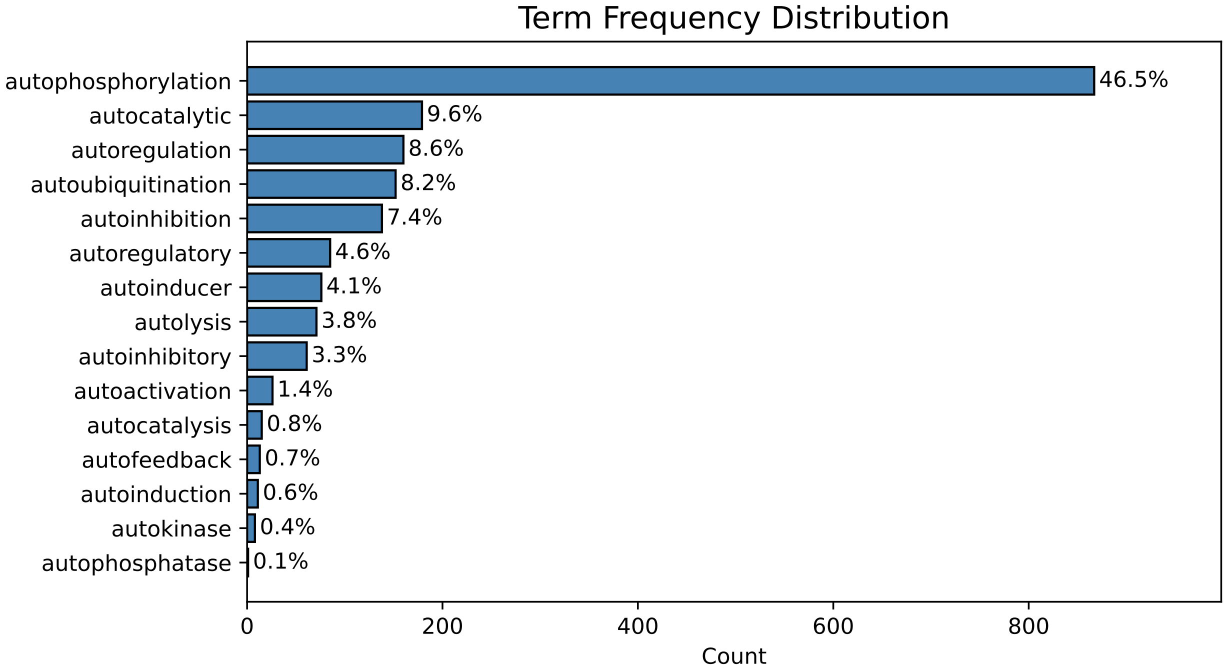
To build a machine learning model that identifies biochemical regulatory features of proteins, our project relies on two primary datasets provided by our partner. The first data frame contains metadata from over 260,000 scientific articles, including PubMed identification number (PMIDs), titles, and abstracts. The second data frame consists of approximately 1.32 million protein annotations, where 1,823 entries (0.14%) contain regulatory labels indicating potential self-regulation mechanisms such as *autophosphorylation* or *autoinhibition*.

We merged these two datasets using the PMID as a shared key, resulting in a single dataset with 1,323,976 entries. From the merged data, we retained six essential columns: a protein accession identifier (AC), organism name (OS), PMID, title, abstract, and regulatory terms. The regulatory terms were originally spread across three separate columns: *Term_in_RP*, *Term_in_RT*, and *Term_in_RC*. For clarity and consistency, we combined these three columns into a single unified column named *Terms*, which contains the cleaned and deduplicated annotations for each entry.

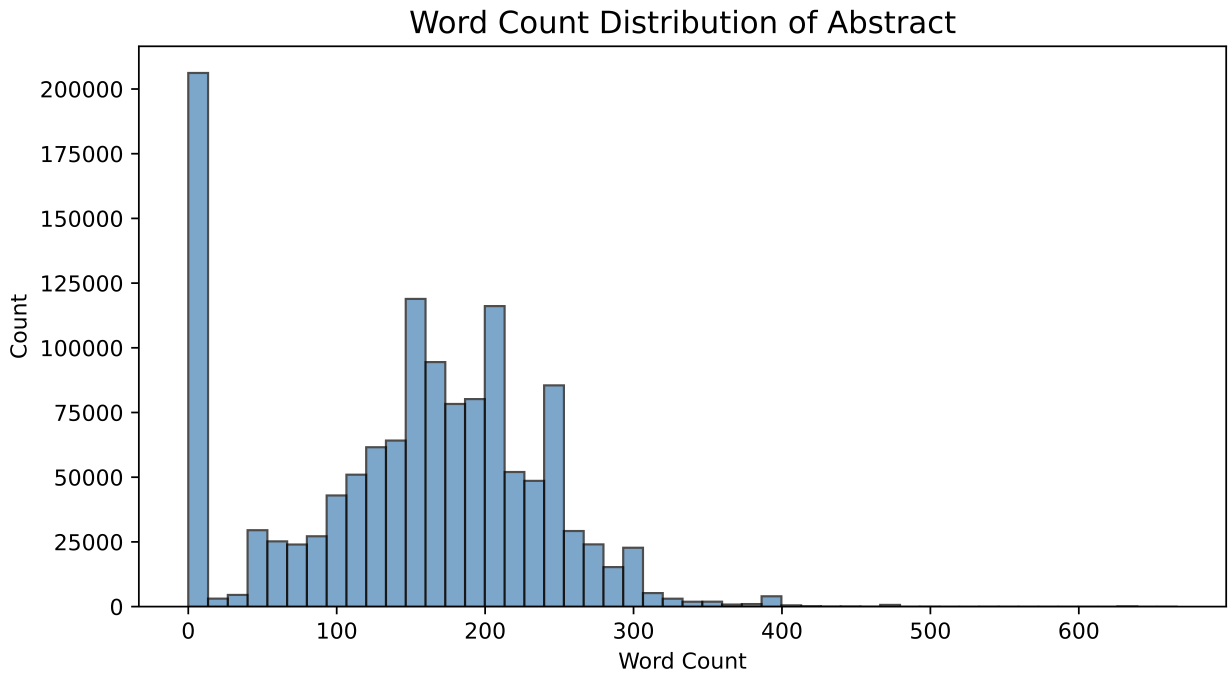
Table 1: Final dataset structure

	AC	OS	PMID	Title	Abstract	Terms
1085	P63104	Homo sapien...	29357390	Herpesvirus...	The N-termi...	autoubiquit...
5416	Q64264	Mus musculu...	18599790	Sporadic au...	Sudden infa...	autoinhibition
5891	Q9Y6E2	Homo sapien...	29470543	Translation...	The efficie...	autoregulation
5926	Q7L1Q6	Homo sapien...	29470543	Translation...	The efficie...	autoregulation
9655	Q13131	Homo sapien...	17088252	Conserved a...	AMP-activat...	autoinhibitory

Despite the large volume of data, only about 0.14% of the entries (1,823 rows) contain any regulatory term. Among these labeled entries, the majority (98%) are annotated with a single term, and just a small fraction (2%) are labeled with multiple regulatory types. The distribution of these terms is highly imbalanced: nearly half of the labeled entries mention the term *autophosphorylation*, while others like *autoubiquitination* and *autoinhibition* occur less frequently. We also found that many of these terms are explicitly mentioned in the original article metadata. Specifically, about 48% of labeled entries contain the term in the article title, and 47% include it in the abstract.



Lastly, we examined the length of the abstracts to guide future modeling decisions. As illustrated in figure below, the majority of abstracts are within this word count range, which provides insight into the suitable input length for the model.


Word Count Distribution of Abstract
Count
200000
175000
150000
125000
100000
75000
50000
25000
0
Word Count
0
100
200
300
400
500
600

Overview of Techniques

The dataset will be trained in a deep learning pipeline, with BERT recommended as a strong starting point, specifically designed to detect and classify autoregulatory features in biomedical text. This model will be the core of our Shiny app, which allows users to search and explore the model's predicted outcomes. Instead of performing live prediction, the Shiny app will provide a searchable interface over pre-analyzed texts, enabling researchers and biologists to explore predicted autoregulatory terms and patterns interactively.

The BERT model is chosen because it is a transformer-based model that excels in multi-label and multi-class classification tasks, making it well to recognize and categorize multiple types of autoregulatory features in protein-related texts, rather than just a simple binary yes/no response (Le et al. (2021)). The model can capture the contextual relationships in complex biomedical text and it might perform even better when we adapt to domain-specific variants such as BioBERT and PubMedBERT (Kang et al. (2022)). We will also evaluate variants like DistilBERT for faster processing and BioBERT or PubMedBERT for better domain alignment, ensuring the best balance between accuracy, efficiency, and scalability.

Following the training process, the model generates a prediction for one or more autoregulatory types, such as autophosphorylation or autoinhibition, for protein-related text. The predictions are indexed and presented in the Shiny interface. This enables interactive examination between biomedical articles and the outputs.

Challenges

Despite the advantages of using BERT listed above, the dataset consists of a variety of challenges that require careful handling. First, the labeling is highly sparse, with only 0.14% of the entries labeled by terms indicating autoregulation, which limits the number of supervised signals available for the training process. Second, the label distribution is highly imbalanced with half of the labeled entries corresponding to a single term, autophosphorylation, while the other terms are each found only in a small number of examples. To address the severe class imbalance in our dataset, we plan to apply oversampling or undersampling strategies alongside cost-sensitive learning techniques such as class-weighted loss and focal loss, which focus training on underrepresented labels (Zhang et al., 2020; Lin et al., 2017). While the majority of labeled entries are dominated by a single term, supporting multi-label classification remains essential to capture samples with multiple regulatory features. Additionally, we are exploring advanced architectures such as Graph Convolutional Networks (GCNs) to model potential dependencies between co-occurring regulatory terms, which could further enhance representation learning (Sharma et al., 2023). Despite the majority of labeled entries corresponding to a single term, it is still important for the model to support multi-label classification to handle cases with multiple regulation types. Lastly, while the lengths of the input texts are mostly appropriate for BERT (with most falling between 256 and 384 tokens), proper preprocessing and tokenization must be ensured to maintain relevant context throughout both titles and abstracts.

Success Criteria

To ensure that this tool is both scientifically valid and practically useful, we evaluate its success based on quality, stakeholder usability, and future extensibility. When assessing our model, it is essential to use metrics that go beyond overall accuracy. This project aims to develop a tool that will help researchers identify autoregulation mechanisms in proteins from biomedical text. Given the highly imbalanced nature of our dataset, with terms such as autophosphorylation making up nearly half of all labeled entries, we focus on three key evaluation metrics: precision, recall, and micro F1-score. Precision captures how many of the predicted terms are actually correct. This is crucial because our partner's team will review model predictions directly. A high precision model minimizes false positives, ensuring researchers don't waste time filtering out incorrect terms or risk misleading downstream analysis (Dowd (2023)).

Recall, on the other hand, ensures that the model does not miss important terms. If the model only predicts terms it is confident about and misses any type of autoregulation when they are present, it limits the scientific value of the tool (Torgo & Ribeiro (2009)). This would hinder the growth of the database and the discovery of new regulatory pathways. High recall ensures that the model captures a wide range of relevant annotations, aligning with the partner's goal of expanding the self-loop catalog through automated analysis (Torgo & Ribeiro (2009)).

We selected the micro F1-score because it provides a better understanding of the model's overall performance across all terms, rather than focusing solely on the most common ones (Takahashi et al. (2022)). For example, a model that always predicts autophosphorylation for every abstract might achieve a high accuracy score but would fail to identify rare yet biologically significant terms like autoinduction or autophosphatase. The micro F1-score averages precision and recall across all predictions, weighting each prediction equally. This penalizes models that ignore minority terms, which is particularly important since researchers value rare self-loops just as much as the more common ones (Takahashi et al. (2022)).

Stakeholder

These evaluation metrics are closely aligned with the priorities of our stakeholders. The partner's lab requires high recall and micro F1 to ensure that the tool can cover a broad range of biological pathways. This is essential for advancing research in systems pharmacology and drug development, as it allows the model to capture a wide variety of autoregulatory features. Researchers using the Shiny app will prioritize precision, as they will directly interact with the model's predictions. A high precision ensures that they can trust the tool's outputs, minimizing the need for manual verification and maximizing its utility in research applications.

For our team, these metrics guide the evaluation and comparison of different BERT model variants, helping us select the best approach for deployment. By focusing on micro F1, precision, and recall, we ensure that our model performs well not only on training data but also in real-world applications. The goal is for the model to reliably predict meaningful regulatory features, supporting the advancement of biomedical research and providing valuable insights into self-regulatory mechanisms within proteins.

Timeline

Our project is structured into four key phases over eight weeks:

Week 1–2: Setup and Scoping

We began with setting up our team workflow, understanding the partner's needs, and finalizing our project scope. During this phase, we prepared and delivered both the proposal presentation and written report.

Week 3–Mid-Week 6: Modeling and Iteration

The core of our work will happen here. We are fine-tuning baseline BERT models and comparing domain-specific variants like BioBERT and PubMedBERT. Model iterations will be guided by performance on the evaluation metrics and feedback from our mentor and partner. Final model selection will occur at the end of this phase.

Late Week 6: Output Structuring and Prototyping

Once the best-performing model is selected, we will structure the predictions into a format that enables easy querying. We will then develop a Shiny app which allows researchers to explore predictions interactively.

Week 7–8: Deliverables and Wrap-Up

We will finalize and submit our runnable data product by June 9. This will be followed by our final presentation on June 12–13 and the final report by June 25. The project concludes with a poster presentation on June 26. We have allowed buffer time throughout to accommodate stakeholder feedback and iteration.

This timeline ensures steady progress toward a usable, reliable, and extensible tool aligned with both scientific and practical goals.

Conclusion

In summary, this proposal outlines the development of a machine learning tool to identify self-regulatory features in proteins, using advanced deep learning techniques like BERT. The tool will support drug development and disease modeling by providing an efficient method for researchers to identify autoregulation mechanisms, aiding in the discovery of new drug targets and expanding our understanding of disease processes. The timeline and success criteria ensure that this tool will be both scientifically rigorous and practically useful.

References

“‘{=latex}

Bateman, E. (2008). Autoregulation of eukaryotic transcription factors. *Progress in Nucleic Acid Research and Molecular Biology*, *60*, 133–168. https://doi.org/10.1016/S0079-6603(08)60892-2

Dowd, P. A. (2023). *Accuracy and precision* (B. S. Daya Sagar, Q. Cheng, J. McKinley, & F. Agterberg, Eds.). Springer. https://doi.org/10.1007/978-3-030-85040-1_432

Helsinki, U. of. (n.d.). *University of helsinki*. https://www.helsinki.fi/en.

Kang, H., Goo, S., Lee, H., Chae, J., Yun, H., & Jung, S. (2022). Fine-tuning of bert model to accurately predict drug–target interactions. *Pharmaceutics*, *14*(8), 1710. https://doi.org/10.3390/pharmaceutics14081710

Le, N. Q., Ho, Q.-T., Nguyen, T.-T.-D., & Ou, Y.-Y. (2021). A transformer architecture based on bert and 2D convolutional neural network to identify DNA enhancers from sequence information. *Briefings in Bioinformatics*, *22*(5). https://doi.org/10.1093/bib/bbab005

Takahashi, K., Yamamoto, K., Kuchiba, A., & al., et. (2022). Confidence interval for micro-averaged F1 and macro-averaged F1 scores. *Applied Intelligence*, *52*, 4961–4972. https://doi.org/10.1007/s10489-021-02635-5

Torgo, L., & Ribeiro, R. (2009). Precision and recall for regression. In J. Gama, V. S. Costa, A. M. Jorge, & P. B. Brazdil (Eds.), *Discovery science* (Vol. 5808, pp. 261–275). Springer. https://doi.org/10.1007/978-3-642-04747-3_26