SOORENA: Self-lOOp containing or autoREgulatory Nodes in biological network Analysis

Hala Arar[1], Jehad Aldahdooh[2], Payman Nickchi[1], Mohieddin Jafari[3,4,5]

1) Department of Statistics, University of British Columbia, Vancouver, BC, Canada
2) Research Programs Unit, University of Helsinki, Helsinki, Finland
3) Department of Biochemistry and Developmental Biology, University of Helsinki, Helsinki, Finland
4) Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland
5) Tampere Institute for Advanced Study, Tampere University, Tampere, Finland

* Corresponding author:

Mohieddin Jafari,

mohieddin.jafari@helsinki.fi

Abstract

Autoregulatory mechanisms, in which proteins modify their own activity or expression, are fundamental components of biological regulatory systems but remain challenging to identify systematically within the scientific literature. Manual curation is outpaced by publication growth, with self-regulation often described implicitly. To address the lack of automated tools for identifying protein autoregulatory mechanisms, we present SOORENA, a two-stage transformer-based model designed to predict and classify such mechanisms within PubMed abstracts. In Stage 1, the model determines whether a publication describes any form of protein autoregulation. In Stage 2, positive instances are further classified into one of seven mechanistic categories: autophosphorylation, autoubiquitination, autocatalytic activity, autoinhibition, autolysis, autoinducer production, and autoregulation. SOORENA was fine-tuned from PubMedBERT using a curated dataset of 1,332 experimentally validated abstracts sourced from UniProt-referenced publications. On a held-out test set, Stage 1 achieved an accuracy of 96.0% and a precision of 97.8%, effectively minimizing false positive propagation. Stage 2 demonstrated robust performance across all classes, with an overall accuracy of 95.5% and a macro-F1 score of 96.2%, including perfect classification for the two least-represented categories. Error analysis revealed that most misclassifications occurred between mechanistically related categories, suggesting that the model's learned representations reflect underlying biological relationships. We deployed SOORENA as a Shiny app enabling interactive search, metadata-based filtering, and ranking of predictions by model confidence alongside standardized ontology definitions to support scientific exploration. These results demonstrate that domain-specific language models can scale the discovery and curation of biologically critical self-regulatory mechanisms.

1. Introduction

Autoregulation refers to a biological process in which a protein modulates its own abundance or activity through direct molecular feedback, enabling precise control of cellular function [1]. These self-regulatory processes are widespread across signaling, transcriptional, and metabolic networks, where they contribute to homeostasis, robustness, and rapid adaptation to changing environments [2]. Autoregulatory motifs are among the most common topological features observed in biological regulatory systems, particularly in transcription factor networks and enzyme signaling pathways [3–5]. Notably, such motifs (often represented as self-loops in network models) cannot be eliminated through any known model reduction techniques in dynamic network modeling [6,7]. This underscores their fundamental importance in preserving the integrity and functionality of biological models.

Autoregulation operates through diverse molecular mechanisms. Enzymatic self-modification includes autophosphorylation by kinases, autoubiquitination by E3 ligases, autocatalytic reactions in metabolic enzymes, and autolysis in proteases [8–11]. At the gene expression level, transcription factors can activate or repress their own transcription, forming positive or negative feedback loops that modulate response dynamics and reduce expression noise [12–14]. These mechanisms play essential roles in human disease. Dysregulated autoregulatory signaling contributes to oncogenic kinase activation, neurodegenerative protein dysfunction, and bacterial quorum-sensing (QS) pathways that drive antimicrobial resistance, making autoregulatory proteins promising therapeutic targets [15,16].

Despite their importance, autoregulatory mechanisms remain under-characterized in public databases and are difficult to extract computationally from scientific literature. Mechanistic descriptions are often embedded implicitly in text, using heterogeneous phrasing rather than standardized vocabulary. For example, a statement such as "the kinase phosphorylates itself" implies autophosphorylation without naming the mechanism explicitly. Our keyword presence analysis supports this challenge. For example, fewer than half of autophosphorylation publications include the literal term in their abstracts, and similar gaps were observed across other mechanisms. As a result, conventional keyword-based search and rule-based text mining struggle to retrieve relevant literature when mechanistic descriptions are phrased implicitly.

Although manually curated resources such as UniProt provide high-quality mechanistic annotations [17], expert review cannot keep pace with the rapid expansion of biomedical literature. Publishing output now exceeds 1.5 million new scientific articles annually [18], meaning that reviewing even a targeted subset of 250,000 abstracts would require years of sustained curator effort. This creates a substantial bottleneck in keeping biological knowledge current and comprehensive.

Natural language processing (NLP) approaches using neural language models have transformed information extraction in biomedicine by enabling semantic interpretation beyond surface keywords [18,19][20]. Domain-specific transformer architectures (for example, PubMedBERT) capture biomedical terminology and contextual meaning more effectively than general-domain models. This facilitates the detection of specialized mechanistic relationships in text, such as autoregulation.

Here, we present Self-lOOp containing or autoREgulatory Nodes in biological network Analysis (SOORENA), a transformer-based model with two stages for automated identification and mechanistic classification of protein autoregulation in scientific literature. Stage 1 identifies abstracts describing any form of autoregulation, whereas Stage 2 classifies positive abstracts into seven mechanism categories: autophosphorylation, autoubiquitination, autocatalytic activity, autoinhibition, autolysis, autoinducer production, and autoregulation of gene expression. This architecture improves computational efficiency by restricting the fine-grained classification to relevant publications while preserving high precision in mechanism screening.

SOORENA was fine-tuned on 1,332 manually curated and experimentally validated mechanism annotations mapped from UniProt to PubMed abstracts. Because our dataset was highly unbalanced (with 719 examples of autophosphorylation but only 38 of autoinducer production), we used a weighted loss function and evaluation metrics that better captured performance on the underrepresented classes. These design choices enabled robust recognition of both common and rare autoregulatory mechanisms.

Finally, we deployed SOORENA across 252,880 PubMed abstracts to create the largest resource to date for predicted autoregulatory mechanisms in the literature. An interactive web application allows researchers to explore predictions, filter by metadata, and access ontology-linked standardized terms. SOORENA provides a practical framework to accelerate the discovery and curation of protein self-regulatory mechanisms by integrating expert-curated training data, domain-specific language models, and scalable inference tools.

2. Methods

2.1 Data Collection

We constructed a large-scale biomedical text corpus by integrating publication metadata from PubMed with manually curated annotations of autoregulatory mechanisms from UniProt. Specifically, we utilized the Swiss-Prot subset of UniProt, which contains reviewed protein entries with experimentally validated mechanism annotations. Although UniProt references a vast number of publications (estimated at over 200 million, many of which remain unverified), our dataset focuses on a high-confidence subset derived from curated entries. Abstracts were retrieved via the PubMed API and used for language modeling, while UniProt served as the source of validated autoregulatory labels, which are embedded within individual protein records.

2.1.1 PubMed Corpus

We obtained a subset of 262,819 publications from PubMed, each containing a PubMed Identifier (PMID), Title, Abstract, Journal, and Authors. These publications were randomly selected from over two million abstracts corresponding to journals indexed in UniProt to ensure representative coverage of literature relevant to protein function. The dataset included 8,607 records lacking abstracts but contained no duplicate PMIDs.

2.1.2 UniProt Annotations

We extracted experimentally validated autoregulatory mechanism annotations from the Universal Protein Resource [17]. These annotations are generated by expert curators based on experimental evidence reported in the referenced publications. The UniProt

dataset consisted of 1,323,976 rows and 13 metadata fields describing protein-level experimental findings. Each entry is tied to a protein via a UniProt accession number (AC) and may reference supporting literature using PMIDs embedded in the RX field. Each unique AC may have more than one reference available. Up to three fields (RP, RT, and RC, which we later named them as Term_in_RP, Term_in_RT, Term_in_RC) contain mechanistic terms indicating evidence that a protein regulates its own activity or abundance. These term columns were aggregated into a single "Terms" column for model training. Of all entries, 1,823 were labeled with autoregulatory mechanisms (number of rows with a known autoregulatory mechanism), while 1,322,153 were unlabeled. A total of 181,779 rows had missing PMIDs, including 10 labeled entries, which were removed because the model required PubMed-linked text for training. Additionally, 89,620 PMIDs appeared multiple times, since each protein record occupied its own row; these were later aggregated so that each PMID corresponded to a single, combined entry. After aggregation, the dataset was reduced from 1,142,197 rows to 268,619 unique PMIDs, of which 1,374 were labeled with autoregulatory mechanisms. The reduction in labeled entries occurred because multiple protein records often referenced the same PMID, and these were merged into a single aggregated entry.

2.2 Label Construction and Data Integration

2.2.1 Merging PubMed and UniProt

UniProt annotations were mapped to PubMed metadata using PMIDs as a common identifier. A join between two tables was applied to ensure that all PubMed publications were retained, even if no corresponding UniProt record existed. Mechanism labels from

UniProt were added only when a PMID match was found. The resulting merged dataset contained 262,819 PubMed publications in total, of which 1,349 were labeled with autoregulatory mechanisms and 261,470 were unlabeled. The slight reduction in labeled entries from 1,374 to 1,349 occurred because a small number of labeled PMIDs from UniProt were not present in the PubMed corpus. A total of 8,607 entries from the PubMed dataset were dropped because they lacked abstracts. Only two of these were labeled, and abstracts were required for model training, leaving 1,347 labeled entries at this stage.

2.2.2 Term Analysis

To characterize the labeled data, all mechanistic terms associated with autoregulatory behavior were extracted from the UniProt annotations. We identified a total of 15 unique terms across the labeled subset, which shows substantial class imbalance. The most frequent mechanism was autophosphorylation (719 instances), followed by autocatalytic activity (133), autoubiquitination (121), autoregulation (119), and autoinhibition (84). Less common terms included autolysis (41) and autoinducer production (31), while several variants, such as autokinase, autocatalysis, and autophosphatase, appeared only rarely. The ratio between the most and least frequent terms (719:1) highlights the strong skewness toward phosphorylation-related mechanisms. This is not surprising, as this is a common trend in molecular biology due to extensive kinase research. These distributions informed later normalization and filtering steps to ensure that rare but biologically meaningful mechanisms were retained during modeling.

To assess how often mechanistic terms appeared explicitly in the literature, we examined whether each annotated term occurred literally within its corresponding abstract

text. The results showed wide variability across mechanism types. Terms such as autoactivation, autoinduction, and autolysis appeared in nearly all labeled publications, while others were frequently absent from the text despite being experimentally confirmed in UniProt - most notably autoubiquitination (14.0%), autophosphorylation (47.3%), and autocatalytic (41.4%). This finding highlights that many descriptions of autoregulatory mechanisms are phrased implicitly (for example, “the kinase phosphorylates itself”) rather than through standardized terminology. As a result, keyword searches or rule-based text mining miss a large portion of the relevant literature. This highlights the need for a context-aware language model that can understand implicit mechanistic descriptions.

A small subset of publications contained more than one autoregulatory mechanism within the same abstract. Specifically, 31 publications (approximately 2.3% of all labeled data) referenced multiple mechanisms, with an average of 2.1 mechanisms per publication and a maximum of 3.

2.2.3 Feature Engineering

To prepare the text for model training, the title and abstract of each publication were concatenated into a single text field to provide richer contextual information for mechanism detection. The resulting field was stored as text and served as the primary input for all subsequent modeling stages. The text was then cleaned using a series of preprocessing steps: HTML entities such as “&” were decoded, URLs and email addresses were removed, and excess whitespace was normalized. These operations ensured that the model processed standardized and noise-free text while preserving important biomedical terminology.

To evaluate data quality, text length was measured as the number of characters per abstract. Only two publications contained fewer than 100 characters, indicating minimal issues with incomplete text. Both of these publications were unlabeled. Across the dataset, the mean text length was 1,348 characters (SD = 432, median = 1,351), suggesting a consistent abstract structure across publications. Labeled publications were slightly longer (mean = 1,403, SD = 382; median = 1,402) than unlabeled ones (mean = 1,348; SD = 433, median = 1,351).

We further standardized the labeled data by normalizing spelling and variant forms of mechanistic terms to ensure consistent representation. Terms such as autoregulatory, autoinhibitory, autocatalysis, and autoinduction were replaced with their canonical forms (autoregulation, autoinhibition, autocatalytic, and autoinducer). After normalization, 11 unique mechanisms remained. To ensure balanced training, we retained only terms with at least 35 examples, resulting in seven mechanism classes summarized in **Table 1**. Despite this filtering, a 719:38 imbalance persisted between the most and least common mechanisms.

A small subset of publications (18 publications, 1.3 % of labeled data) referenced multiple mechanisms, with up to three mechanisms in a single abstract. However, this subset was insufficient to train a reliable multi-label classifier because of the sample size. Therefore, in the current version of SOORENA, each publication was assigned a single primary mechanism label (the first listed term) to maintain class balance and ensure stable model training. After all cleaning, normalization, and filtering steps, the final modeling dataset consisted of 254,197 total publications, of which 1,332 were labeled with one of the seven supported mechanisms.

Table 1. Final distribution of autoregulatory mechanisms retained for model training. *Percentages are relative to the total of 1,332 labeled publications.*

Mechanism	Publications (n)	Percentage (%)
Autophosphorylation	719	54.0
Autoregulation	163	12.2
Autocatalytic	147	11.0
Autoinhibition	122	9.2
Autoubiquitination	121	9.1
Autolysis	41	3.1
Autoinducer	38	2.9
Total	1,332	100

2.3 Train-Validation-Test Splits

Following dataset preparation, a total of 254,197 publications were retained, of which 1,332 contained experimentally validated autoregulatory mechanisms. These labeled examples were stratified by mechanism type to preserve class proportions across training, validation, and test sets. This stratification ensured that each mechanistic category was proportionally represented during model development, thereby minimizing sampling bias toward the mechanism class with the majority proportion (autophosphorylation).

Using this stratified procedure, 70% of labeled publications (n = 932) were assigned to the training set, while the remaining 30% (n = 400) were split evenly between validation and test sets (n = 200 each). This split was performed only on the labeled subset, meaning that the remaining 252,865 unlabeled publications were excluded from these initial splits

and later incorporated during binary classification (Stage 1). The overall composition of datasets across the original stratified split, Stage 1 (binary detection), and Stage 2 (multiclass classification) is summarized in **Table 2**.

Table 2. Dataset composition across original, Stage 1, and Stage 2 splits

Split	Original (labeled only)	Stage 1 (pos + neg)*	Stage 2 (multi-class)
Train	932	2,796 (932 + 1864)	932
Test	200	600 (200 + 400)	200
Validation	200	600 (200 + 400)	200
Total	1,332	3,996	1,332

* Stage 1 includes unlabeled negatives sampled at a 2:1 ratio relative to labeled positives for each split.

In the original split, the data were used to train a single-stage supervised model limited to the labeled corpus. For Stage 1, the task was reformulated as binary classification, where publications describing autoregulatory mechanisms were labeled 1, and randomly sampled unlabeled publications were treated as 0. Each split was expanded by adding unlabeled negatives at a 2:1 ratio, tripling the training volume while preserving the same class balance across validation and test sets. This strategy improved the model's ability to distinguish between mechanistic and non-mechanistic abstracts, reducing the likelihood of false positives before passing candidates to Stage 2.

For Stage 2, only the original labeled subsets (932 train, 200 validation, 200 test) were used. This stage addressed multi-class mechanism classification, predicting one of seven mechanism types. Because the dataset remained highly imbalanced (dominated by

autophosphorylation), class weights were applied during loss computation to equalize contributions from minority classes. The class-wise counts within the labeled corpus are shown in **Table 3**.

Table 3. Per-class distribution within labeled dataset (1,332 publications)

Mechanism	Train (n = 932)	Validation (n = 200)	Test (n = 200)	Total (n = 1,332)
Autophosphorylation	503	108	108	719
Autoregulation	114	24	25	163
Autocatalytic	103	22	22	147
Autoinhibition	85	18	19	122
Autoubiquitination	84	18	19	121
Autolysis	28	6	7	41
Autoinducer	15	4	4	38
Total	932	200	200	1,332

These counts confirm that stratified sampling preserved each mechanism's class proportions across splits. The training subset captured ~70% of each class, while validation and test sets maintained balanced representation for evaluation. Together, this framework produced a robust two-stage learning design: Stage 1 for binary detection of autoregulatory mechanisms across both labeled and unlabeled literature, and Stage 2 for fine-grained classification of mechanism type within the labeled subset.

2.4 Model Architecture

SOORENA is a two-stage trained transformer-based classification model built on PubMedBERT (microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext), a BERT-base architecture pretrained solely on PubMed abstracts and full-text articles to capture biomedical terminology and contextual semantics [20,21]. Stage 1 detects whether a publication describes any autoregulatory mechanism, serving as a screening layer to filter the broader biomedical literature. Stage 2 then assigns Stage 1-positive publications to one of seven distinct mechanism categories. PubMedBERT has demonstrated superior performance over general-domain BERT variants in biomedical entity recognition and relation extraction tasks [20].

Both stages use the same architecture: a PubMedBERT encoder followed by a dropout layer (p = 0.1) and a linear classification head. The stages are fine-tuned independently, allowing hyperparameter optimization and checkpoint selection specific to the prediction task (**Fig. 1**).

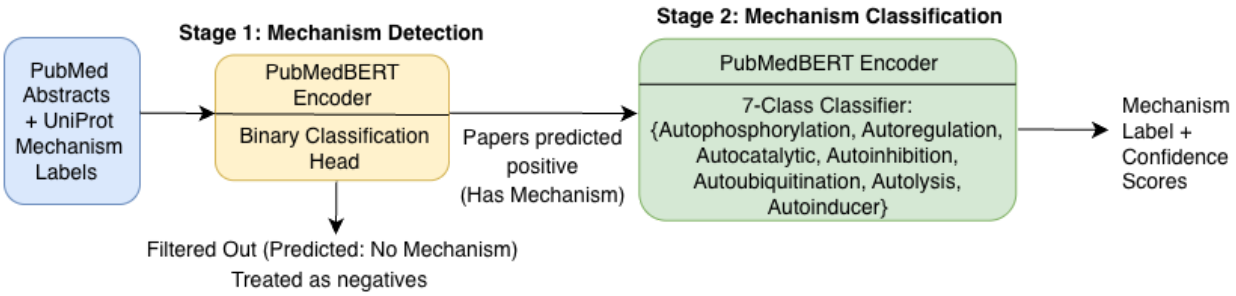

Figure 1. **Overview of the SOORENA two-stage autoregulatory mechanism detection architecture.**

2.5 Training Procedure

Both stages were fine-tuned using PyTorch 2.0 and the Hugging Face Transformers library (v4.30). Inputs were tokenized to a maximum length of 512 tokens and processed in mini-batches of 16 samples. The AdamW optimizer (learning rate of 2×10^{-5}) was used with a linear warmup schedule applied to the first 10% of training steps. All data splits used a fixed random seed (42) to ensure reproducibility. AdamW was chosen for its improved stability and regularization when fine-tuning transformers compared to standard Adam [22].

Stage 1 training employed binary cross-entropy loss and ran for three epochs (525 update steps). Model selection was based on the validation F1 score, which balances precision and recall to minimize false positives entering Stage 2. Stage 2 training used weighted cross-entropy loss to handle class imbalance across mechanism types and ran for four epochs (236 update steps). The class weights, inversely proportional to class frequency, are shown in **Table 4**. The final Stage 2 model checkpoint was chosen based on the highest validation macro-F1 score, ensuring balanced performance across both common and rare mechanisms.

Table 4. Class weights used for Stage 2 multi-class training

Mechanism	Class weight
Autophosphorylation	0.27
Autoregulation	1.21
Autocatalytic	1.29

Autoinhibition	1.57
Autoubiquitination	1.60
Autolysis	4.93
Autoinducer	4.93

2.6 Evaluation

2.6.1 Performance Metrics

We assessed model performance using multiple complementary metrics: Accuracy, Precision, Recall, F1 Score, Macro-F1, and Weighted-F1. For Stage 2, macro-F1 was prioritized as the primary metric because it prevents strong performance on common classes (autophosphorylation) from masking poor performance on rare ones (autolysis, autoinducer). This is critical in biomedical classification, where underrepresented mechanisms may be scientifically important despite low frequency.

2.6.2 Error and Confidence Analysis

Confusion matrices were generated to inspect class-specific error patterns. Prediction confidence was used to identify uncertain classifications, especially in Stage 1, where the negative class contains label noise.

2.7 Computational Resources

All experiments were conducted on an Apple M1 Max chip (32 GB unified memory) without GPU acceleration. The code supports GPU training when available. Stage 1 required

approximately 33 minutes per epoch, and Stage 2 required approximately 12 minutes per epoch.

3. Results

3.1 Stage 1: Binary Classification Performance

Stage 1 classified publications as either containing or not containing evidence of an autoregulatory mechanism. On the held-out test set (n = 600), the model achieved strong performance, accurately identifying publications that should progress to mechanistic classification. Overall accuracy was 96.0%, with a precision of 97.8% and a recall of 90.0% for the positive class, yielding an F1 score of 93.8% (**Table 5**). The model's high precision ensured that very few false positives were propagated into Stage 2.

Errors were dominated by false negatives, with 20 mechanism-containing publications missed and 4 publications incorrectly flagged as positive. This conservative behavior aligns with the design objective of minimizing uncertain cases passed to Stage 2. The confusion matrix (**Fig. 2**) illustrates that misclassifications did not arise from systematic bias toward specific negative subsets but rather from ambiguous descriptions in the original text.

Validation F1 improved from 90.2% at epoch 1 to a peak of 91.8% at epoch 2, followed by a slight decrease at epoch 3, indicating minor overfitting. The best epoch-2 checkpoint was therefore selected for final evaluation. Collectively, these results demonstrate that Stage 1 acts effectively as a high-precision filtering stage for biomedical corpora for fine-grained mechanistic classification.

Table 5. Stage 1 Test Performance (n = 600)

Metric	Score (%)
Accuracy	96.0
Precision	97.8
Recall	90.0
F1	93.8

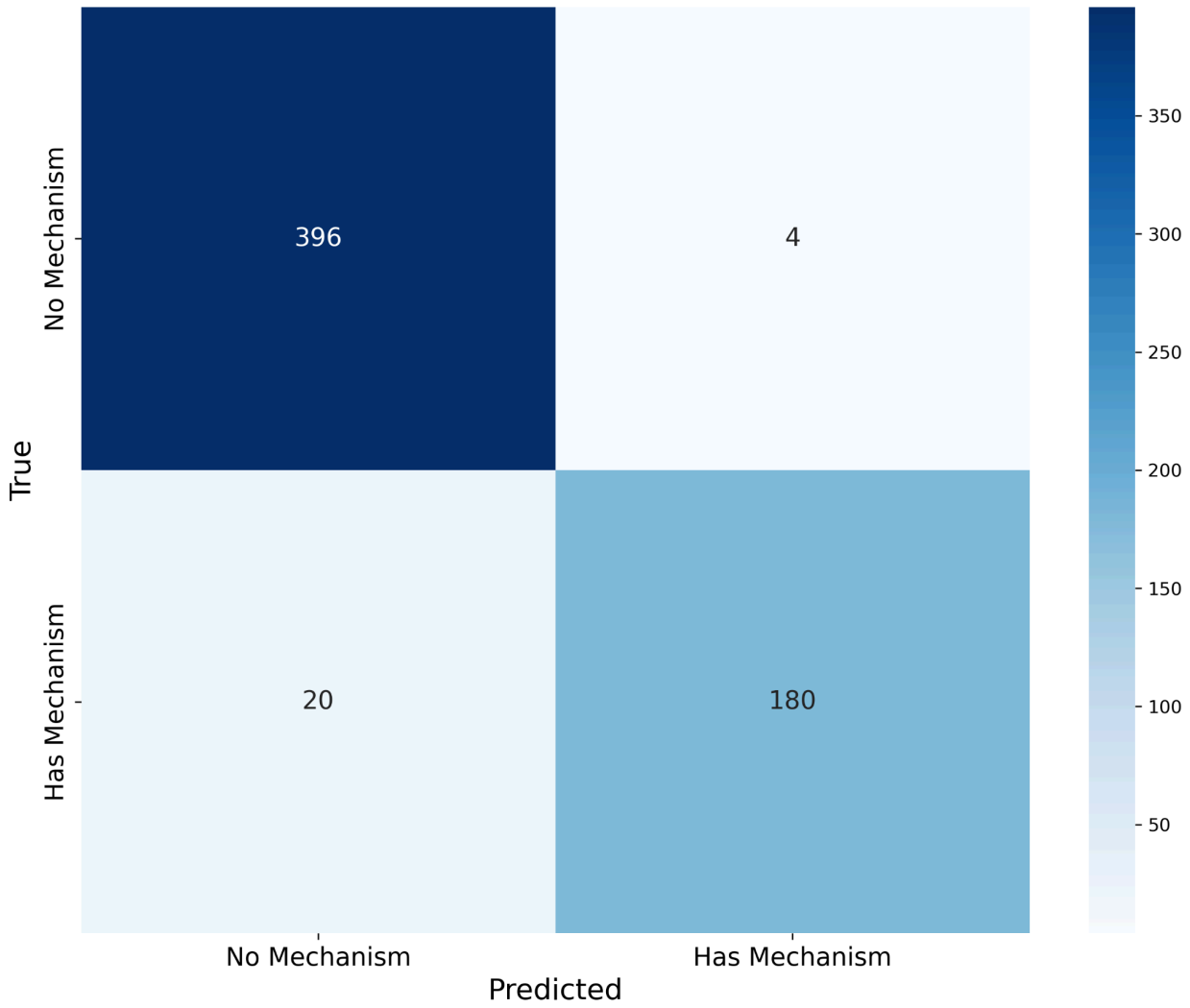

Figure 2. **Confusion matrix for Stage 1 binary classification showing true versus predicted labels for 600 held-out test abstracts.**

3.2 Stage 2: Multi-Class Mechanism Classification

Stage 2 assigned each Stage 1-positive publication to one of seven autoregulatory mechanism categories. On the held-out test set (n = 200), the model achieved 95.5% accuracy, 96.2% macro-F1, and 95.5% weighted-F1 (**Table 6**). These results demonstrate robust generalization despite substantial class imbalance in the training data.

Performance across individual mechanism types was similarly strong (**Table 7**). Autophosphorylation, the most common mechanism, achieved 99.0% precision and 92.5% recall (F1 = 95.6%), indicating effective representation learning even when class prevalence might bias toward overprediction. Mechanisms with intermediate representation, including autoregulation, autocatalytic activity, autoinhibition, and autoubiquitination, achieved F1 scores above 91%.

Strikingly, the rarest mechanisms, autolysis and autoinducer production (n = 6 each), were classified perfectly (100% precision, recall, and F1). This highlights the effectiveness of weighted loss in preserving minority-class performance and preventing collapse toward majority categories. Together, these findings show that Stage 2 reliably captures mechanistic diversity across the full spectrum of autoregulatory behaviors.

Table 6. Overall Stage 2 test-set classification performance

Metric	Value (%)
Accuracy	95.5
Macro precision	94.6
Macro recall	98.1
Macro F1	96.2
Weighted precision	95.9

Weighted recall	95.5
Weighted F1	95.5

Table 7. Per-class Stage 2 classification performance on the test set

Mechanism	Precision (%)	Recall (%)	F1 score (%)	Support (n)
Autophosphorylation	99.0	92.5	95.6	107
Autoregulation	92.3	100.0	96.0	24
Autocatalytic	91.7	100.0	95.6	22
Autoinhibition	94.4	94.4	94.4	18
Autoubiquitination	85.9	100.0	91.9	17
Autolysis	100.0	100.0	100.0	6
Autoinducer	100.0	100.0	100.0	6

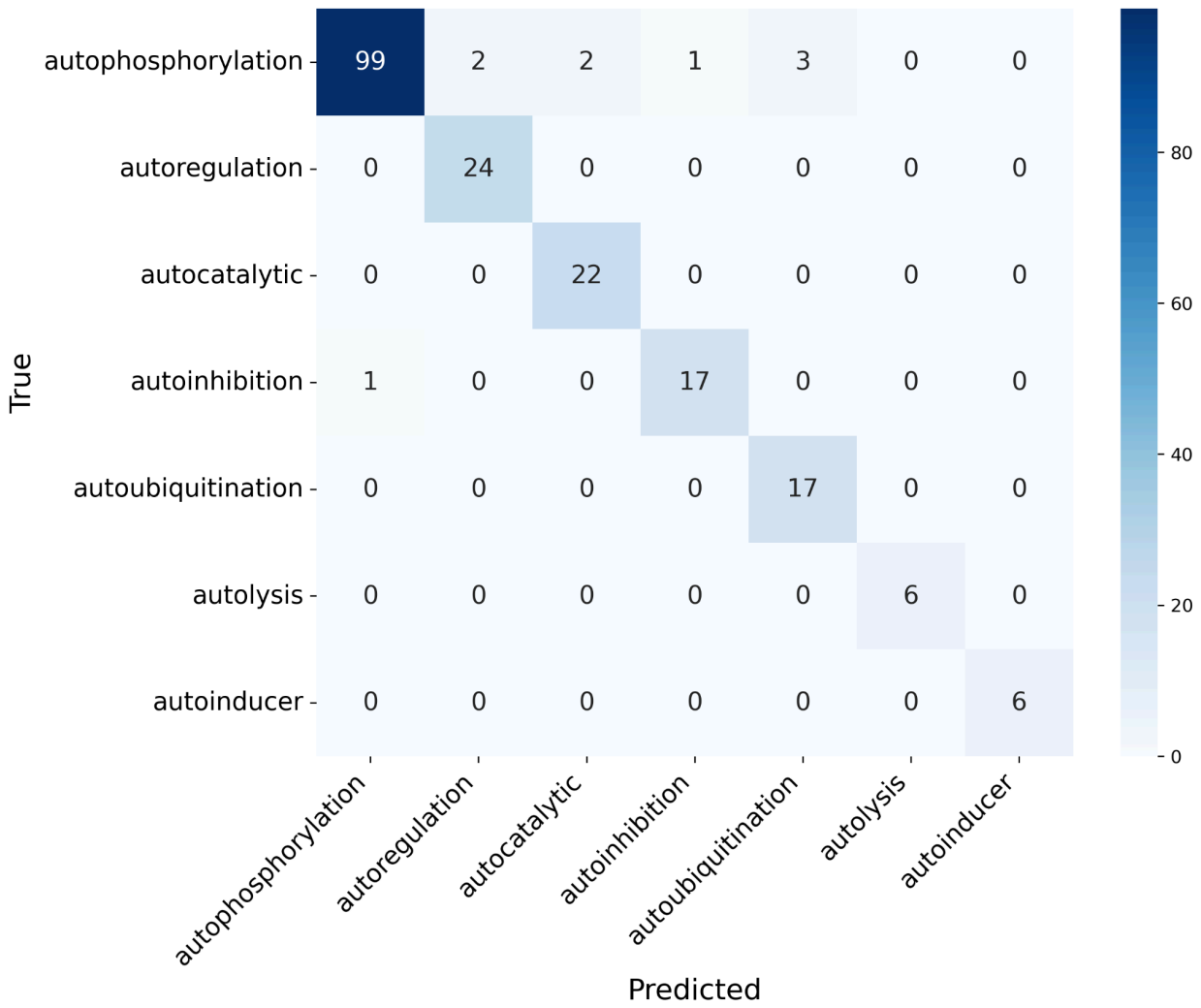

Figure 3. **Confusion matrix for Stage 2 multi-class classification showing true versus predicted mechanism labels for 200 test abstracts.**

3.3 Error & Uncertainty Analysis

Across the Stage 2 test set, the model misclassified 9 of 200 publications (4.5% error rate). Most errors occurred between mechanistically related categories, particularly autophosphorylation and autocatalytic activity, which share overlapping biochemical terminology describing self-activation mediated by the enzyme. Importantly, no

misclassifications occurred between mechanistically distant categories (e.g., autolysis vs. autoinducer production), indicating that the model captures biologically meaningful distinctions rather than relying on superficial keyword patterns (**Fig. 3**).

Prediction confidence scores, representing the model's estimated probability for each predicted class, provided additional insight into model reliability. Correctly classified Stage 2 examples had a mean confidence of 92.7%, whereas misclassified examples showed substantially lower confidence (mean: 67.3%). This separation suggests that confidence estimates can function as a practical uncertainty measure for downstream expert review.

Similar trends were observed in Stage 1. False negatives were more frequent than false positives, reflecting the model's intentional bias toward high precision. Negative predictions also exhibited a higher mean confidence (93.3%) than positive predictions (89.2%). Together, these findings indicate that SOORENA not only achieves strong classification accuracy but also provides interpretable confidence signals that support high-confidence curation while safely flagging ambiguous literature for manual evaluation.

3.4 Interactive Database: SOORENA Web Application

We developed SOORENA, an R Shiny app that integrates UniProt annotations, model predictions on 252,880 unlabeled PubMed abstracts, and the 1,332 experimentally labeled examples used for model training (**Fig. 4**). The app supports search, filtering, inspection, and export, and includes an embedded ontology to standardize terminology and aid interpretation.

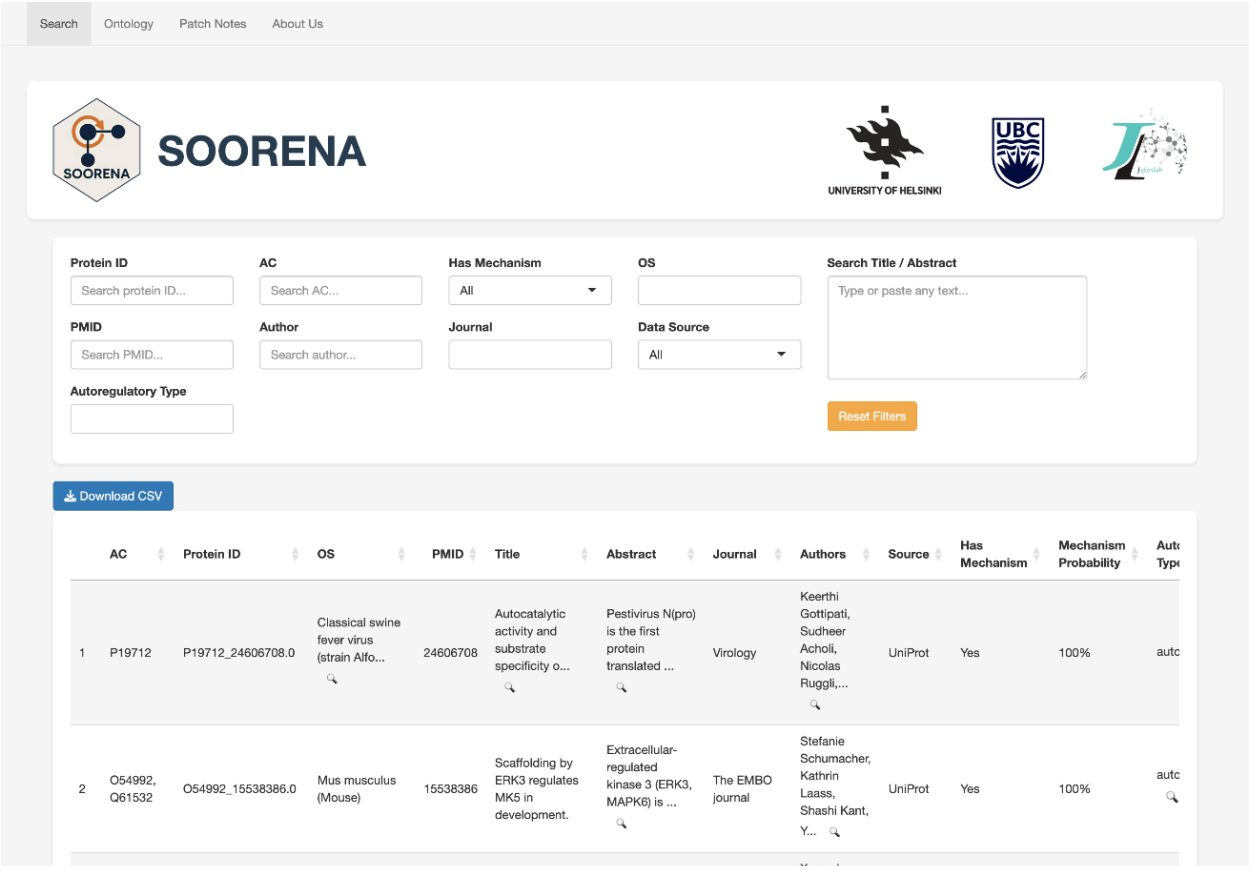

Figure 4. SOORENA web interface showing the Search tab, which supports field-specific filtering (e.g., Protein ID, PMID, organism, and data source) , and displays interactive, sortable table columns, and CSV export functionality.

Across the 252,880 unlabeled abstracts and titles, the model identified 14,374 publications (5.7%) as containing autoregulatory mechanisms and 238,506 (94.3%) as not containing one. Because the Stage 1 binary model used a subset of these unlabeled examples as pseudo-negatives during fine-tuning, some overlap exists between the Stage 1 training pool and the final inference set. However, these examples were not associated with any explicit labels, so their inclusion does not affect reported evaluation metrics and only serves to enrich the searchable database. The distribution of predicted mechanism types is

summarized in **Table 8**. The table shows dominance of autophosphorylation followed by autoubiquitination and autocatalytic activity.

Table 8. Predicted mechanism type distribution across 252,880 unseen PubMed abstracts

Mechanism	Count (n)	Percentage (%)
Autophosphorylation	8,256	57.4
Autoubiquitination	2,226	15.5
Autocatalytic	1,820	12.7
Autoregulation	1,188	8.3
Autoinhibition	516	3.6
Autolysis	192	1.3
Autoinducer	176	1.2
Total predicted mechanisms	14,374	100.0

On the Search tab, users can filter by protein identifiers (Protein ID, UniProt AC), organism (OS), PMID, author, journal, data source (UniProt-curated vs. model-predicted), and predicted mechanism. A free-text box enables case-insensitive keyword search across titles and abstracts. Results are rendered as an interactive table (DT) with sortable columns: AC, Protein ID, OS, PMID, Title, Abstract, Journal, Authors, Source, Has Mechanism, Mechanism Probability, Autoregulatory Type, and Type Confidence. Probabilities are shown as percentages, and long fields are truncated with a magnifier button for full inspection in a modal. Users can apply confidence cut-offs via the Has Mechanism and Autoregulatory Type controls, then export the filtered results as CSV for downstream review.

The Ontology tab provides the curated vocabulary used by SOORENA. A hierarchical tree presents high-level groupings followed by mechanism pages for the seven classes, each with a prose definition, core relations (is-a, part-of, regulates, has-input, has-output, occurs-in), and key references. From the Search table, each classified record includes a pop-up that shows the ontology path and the same definition block, keeping model outputs tied to consistent semantics.

The app is versioned through a Patch Notes tab that lists dated changes to data columns, evaluation summaries, UI components, ontology content, and About tab credits contributors and partners. Together, these elements make the system auditable and reproducible: users can trace any row back to its PMID, see the model's probability for mechanism presence and type, consult standardized definitions, and export subsets for manual validation.

4. Discussion

4.1 Overview of Contributions

This study introduces SOORENA, a transformer-based system for automated discovery of protein autoregulatory mechanisms in biomedical literature. Using PubMed abstracts and experimentally validated proteins from UniProt, the model accurately identifies whether a publication describes autoregulation (96.0% accuracy) and assigns one of seven mechanistic categories with strong class-balanced performance (95.5% accuracy, 96.2% macro-F1). Importantly, SOORENA generalizes well to biologically rare mechanisms, achieving perfect test set performance for autolysis and autoinducer production despite very limited training instances. Deployment across 252,880 additional PubMed abstracts produced the first large-scale, searchable database of predicted autoregulatory mechanisms.

4.2 Effectiveness of the Two-Stage Architecture

The staged architecture reflects the structure of the biological question and improves reliability. Stage 1 prioritizes high precision to prevent irrelevant publications from propagating errors downstream, a well-established requirement in hierarchical classification pipelines [23]. Stage 2 uses weighted loss to counter class imbalance, a well-established requirement for biomedical ML tasks where minority classes are often scientifically most important [24].

This formulation addresses the limitations of single-stage classification, where class imbalance and task heterogeneity commonly degrade performance. Additionally, routing

only ~6% of publications into Stage 2 reduces computational cost by approximately 94%, enabling scalable whole-corpus inference on standard hardware.

4.3 Interpretation of Error Patterns

Misclassifications were concentrated among mechanistically adjacent categories. For example, autophosphorylation was occasionally misclassified as autocatalytic activity. Biochemically, these mechanisms are closely related: autophosphorylating kinases are executing a specific form of enzymatic autocatalysis [25,26]. Likewise, confusion among ubiquitin-dependent processes suggests contextual ambiguity in abstracts, where enzymatic roles (E2 vs. E3 ligases) are often unstated or implicitly described.

Crucially, no misclassifications occurred between mechanistically distant biological categories (e.g., autolysis vs. autoinduction), indicating that the model's representations capture mechanistic semantics rather than relying on lexical co-occurrence patterns. Probability calibration analysis further supports its practical usability: correctly classified predictions displayed higher confidence scores than misclassified ones, allowing confidence-threshold filtering to assist expert curation and quality control review [27].

4.4 Limitations

Reliance on UniProt introduces several known curation biases. Positive examples are skewed toward well-studied proteins and pathways, particularly kinase signaling, thereby inflating the autophosphorylation frequency relative to biological reality. A second key limitation is the use of abstracts rather than full-text articles. Mechanistic assertions often appear in results sections, figures, or supplemental materials [28], meaning that both

training labels and predictions are constrained by incomplete contextual information. This likely underestimates true model capability. Handling of multi-mechanism publications remains simplified. Although biologically meaningful co-regulation exists (e.g., autophosphorylation-dependent autoinhibition), only 18 such cases were identified, insufficient for a reliable multi-label model. The model's reliance on transformer-based semantics inference introduces a potential for latent confounders. For instance, detecting phosphorylation-related language does not guarantee that the event is auto-directed; distinguishing cis- from trans-regulatory contexts remains challenging at the abstract level.

4.5 Comparison to Alternative Approaches

Manual curation remains the gold standard for accuracy, but scales poorly and cannot keep pace with the rapid growth of publications [29]. Rule-based text mining can identify explicit mechanistic terms but often fails when descriptions are implicit or context-dependent, which is common for post-translational and feedback processes. SOORENA bridges this gap by combining transformer-based language understanding with domain-specific fine-tuning, achieving expert-comparable interpretation while enabling continuous, automated updates to regulatory knowledge bases

Transformer models pretrained on biomedical corpora, particularly PubMedBERT, show major advantages in biomedical reasoning over general-domain models such as BERT-base [19,20,30,31]. Our findings reinforce the importance of domain alignment during pretraining, especially when training labeled datasets are small, which is a frequent bottleneck in biomedical NLP.

4.6 Biological and Scientific Impact

SOORENA provides a scalable framework for expanding biochemical knowledge by identifying regulatory findings that remain uncurated into structured databases. High-confidence predictions indicate literature that may contain overlooked experimental evidence, and this helps accelerate the update cycle for resources such as UniProt. The comprehensive mapping of mechanism types across thousands of proteins also enables comparative biological analysis, offering potential insights into how self-regulatory processes have diversified evolutionarily across kinases, proteases, and quorum-sensing systems. Moreover, linking predicted mechanisms with protein domains, structural motifs, and interaction partners may support new hypotheses regarding the molecular features that predispose certain proteins to self-regulate, consistent with prior studies suggesting structural correlates of autoregulation [32,33].

Temporal patterns in autoregulatory discoveries may further reveal research trends and highlight neglected regulatory processes that warrant renewed experimental focus. Finally, integrating SOORENA predictions with structural and functional annotation pipelines could directly support targeted experimental design by identifying proteins most likely to exhibit self-regulatory control and thus prioritizing them for validation..

4.7 Future Directions

Future extensions of this work will focus on increasing the granularity, completeness, and biological context of autoregulatory mechanism detection. Incorporating full-text articles instead of abstracts would substantially improve mechanistic recall, as

essential experimental details often appear in figure legends, methods, and supplementary materials. Addressing this challenge will likely require long-sequence or hierarchical transformers capable of processing multi-kilobyte documents efficiently [34]. Expanding beyond the current single-label formulation to support multi-mechanism prediction is also important since many proteins employ layered regulatory strategies such as autophosphorylation-dependent autoinhibition. This will likely require modified classification architectures and additional annotated examples of co-occurring mechanisms. In parallel, advancing toward knowledge-augmented inference by leveraging protein interaction networks, domain annotations, and curated mechanistic pathways could provide a biologically grounded framework for disambiguating closely related mechanisms. Finally, integrating active learning driven by the model's well-calibrated uncertainty estimates could make expert validation more efficient, allowing iterative refinement of both the system and the underlying curated knowledge.

5. Conclusion

We developed SOORENA, a domain-adapted, two-stage trained transformer model capable of identifying and classifying autoregulatory mechanisms directly from biomedical literature with high accuracy. By integrating curated evidence from UniProt with large-scale PubMed text, the system detects regulatory signals even when explicit mechanistic terminology is absent, addressing an unmet need in computational curation. Strong performance across both common and rare mechanisms demonstrates that transformer-based models can reliably characterize nuanced biochemical processes in highly imbalanced settings. Deployment of SOORENA across the complete PubMed corpus produces the most extensive resource to date for exploring protein self-regulation, now accessible through an interactive, confidence-aware web interface that supports hypothesis generation and targeted expert review. Together, these contributions bridge the gap between experimental discovery and curated biological knowledge, establishing an automated foundation for accelerating the study of autoregulatory mechanisms at scale.

Acknowledgements

We gratefully acknowledge the contributions of Zheng He, Yining Zhou, and Mingyang Zhang in the early prototyping and application development phases of this work.

Data and code availability

The script and data are publicly available at (https://github.com/jafarilab/SOORENA).

Author Contributions

M.J. conceived the research project, and M.J. and P.N. jointly supervised and coordinated its execution. H.A. contributed to the modeling and prepared the initial draft of the manuscript. J.A. contributed to data collection for the training and test sets. All authors participated in writing, editing, and revising the manuscript.

Funding

This study was financially supported by the Jane and Aatos Erkko Foundation [Grant 220031 to M.J.] and Tampere University internal funding [Grant 31003463 to M.J.].

Conflict of interest

The authors declare no conflicts of interest.

References

1. Alon U. Network motifs: theory and experimental approaches. Nat Rev Genet. 2007;8: 450–461.

2. Tyson JJ, Novák B. Functional motifs in biochemical reaction networks. Annu Rev Phys Chem. 2010;61: 219–240.

3. Veliz-Cuba A, Stigler B. Boolean models can explain bistability in the lac operon. J Comput Biol. 2011;18: 783–794.

4. Ma'ayan A, Jenkins SL, Neves S, Hasseldine A, Grace E, Dubin-Thaler B, et al. Formation of regulatory patterns during signal propagation in a Mammalian cellular network. Science. 2005;309: 1078–1083.

5. Jafari M, Sadeghi M, Mirzaie M, Marashi S-A, Rezaei-Tavirani M. Evolutionarily conserved motifs and modules in mitochondrial protein-protein interaction networks. Mitochondrion. 2013;13: 668–675.

6. Saadatpour A, Albert R, Reluga TC. A REDUCTION METHOD FOR BOOLEAN NETWORK MODELS PROVEN TO CONSERVE ATTRACTORS. SIAM J Appl Dyn Syst. 2013;12: 1997–2011.

7. Veliz-Cuba A. Reduction of Boolean network models. J Theor Biol. 2011;289: 167–172.

8. Hubbard SR, Till JH. Protein tyrosine kinase structure and function. Annu Rev Biochem. 2000;69: 373–398.

9. Komander D, Rape M. The ubiquitin code. Annu Rev Biochem. 2012;81: 203–229.

10. Kapust RB, Waugh DS. Controlled intracellular processing of fusion proteins by TEV protease. Protein Expr Purif. 2000;19: 312–318.

11. Blackmond DG. "If pigs could fly" chemistry: a tutorial on the principle of microscopic reversibility. Angew Chem Int Ed Engl. 2009;48: 2648–2654.

12. Becskei A, Serrano L. Engineering stability in gene networks by autoregulation. Nature. 2000;405: 590–593.

13. Rosenfeld N, Elowitz MB, Alon U. Negative autoregulation speeds the response times of transcription networks. J Mol Biol. 2002;323: 785–793.

14. Nevozhay D, Adams RM, Murphy KF, Josic K, Balázsi G. Negative autoregulation linearizes the dose-response and suppresses the heterogeneity of gene expression. Proc Natl Acad Sci U S A. 2009;106: 5123–5128.

15. Soto C. Transmissible proteins: expanding the prion heresy. Cell. 2012;149: 968–977.

16. Papenfort K, Bassler BL. Quorum sensing signal–response systems in Gram-negative bacteria. Nature Reviews Microbiology. 2016;14: 576–588.

17. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 2021;49: D480–D489.

18. Brainard J. Scientists are drowning in COVID-19 papers. Can new tools keep them afloat? [cited 3 Nov 2025]. Available: https://www.science.org/content/article/scientists-are-drowning-covid-19-papers-can-new-tools-keep-them-afloat

19. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36: 1234–1240.

20. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Trans Comput Healthc. 2022;3: 1–23.

21. Devlin J, Chang M-W, Lee K, Toutanova K. Proceedings of the 2019 Conference of the North. Stroudsburg, PA, USA: Association for Computational Linguistics; 2019. doi:10.18653/v1/n19-1423

22. Loshchilov I, Hutter F. Decoupled weight decay regularization. arXiv [cs.LG]. 2017. doi:10.48550/ARXIV.1711.05101

23. Silla CN Jr, Freitas AA. A survey of hierarchical classification across different application domains. Data Min Knowl Discov. 2011;22: 31–72.

24. Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. J Big Data. 2019;6. doi:10.1186/s40537-019-0192-5

25. Reinhardt R, Leonard TA. A critical evaluation of protein kinase regulation by activation loop autophosphorylation. Elife. 2023;12. doi:10.7554/eLife.88210

26. Endicott JA, Noble MEM, Johnson LN. The structural basis for control of eukaryotic protein kinases. Annu Rev Biochem. 2012;81: 587–613.

27. Desai S, Durrett G. Calibration of pre-trained transformers. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, PA, USA: Association for Computational Linguistics; 2020. doi:10.18653/v1/2020.emnlp-main.21

28. Fang L, Chen Q, Wei C-H, Lu Z, Wang K. Bioformer: an efficient transformer language model for biomedical text mining. ArXiv. 2023. Available: https://www.ncbi.nlm.nih.gov/pubmed/36945685

29. Amberger JS, Bocchini CA, Scott AF, Hamosh A. OMIM.org: leveraging knowledge across phenotype-gene relationships. Nucleic Acids Res. 2019;47: D1038–D1043.

30. Aldahdooh J, Tanoli Z, Tang J. Mining drug-target interactions from biomedical literature using chemical and gene descriptions-based ensemble transformer model. Bioinform Adv. 2024;4: vbae106.

31. Aldahdooh J, Vähä-Koskela M, Tang J, Tanoli Z. Using BERT to identify drug-target interactions from whole PubMed. BMC Bioinformatics. 2022;23: 245.

32. Kobe B, Heierhorst J, Feil SC, Parker MW, Benian GM, Weiss KR, et al. Giant protein kinases: domain interactions and structural basis of autoregulation. EMBO J. 1996;15: 6810–6821.

33. Jafari M, Ansari-Pour N, Azimzadeh S, Mirzaie M. A logic-based dynamic modeling approach to explicate the evolution of the central dogma of molecular biology. PLoS One. 2017;12: e0189922.

34. Zaheer M, Guruganesh G, Dubey A, Ainslie J, Alberti C, Ontanon S, et al. Big bird: Transformers for longer sequences. arXiv [cs.LG]. 2020. doi:10.48550/ARXIV.2007.14062