# Retrieving Knowledge of Molecular Regulatory Mechanisms from PubMed Titles via an Event Extraction Approach

David A. Spellman[2], Jason Xiaotian Dou[1], *Student Member, IEEE*, Aaron Fangzheng Wu[2], Sumin Jo[1,4], Yu-Chiao Chiu[3,4], Yufei Huang[1,3,4]

***Abstract*— This work tackles three main issues in information extraction (IE) from biomedical literature: 1. How to design models to extract the structured and flexible representation of molecular regulatory mechanisms (MRM) from the literature that captures the biological features comprehensible to an expert. 2. How to overcome the limitation of existing abstract-based solutions for IE that are costly to annotate and insufficient to extract MRMs for answering regulatory questions in downstream tasks. 3. How to overcome the challenges of a lack of annotated training data for MRM such that different machine learning models for extracting MRM can be trained and evaluated. To address these issues, a novel event extraction (EE) driven approach to automatically extract MRM from paper titles is proposed. We have designed an EE template for MRM that homogenizes the characterization of MRM in titles, making the prediction of MRM possible. We have created the first training dataset for MRM with human-annotated event arguments that capture the essential components of the molecular mechanisms. Our comprehensive evaluations have demonstrated strong performance from the tested models. In addition, further strategies for fine-tuning pretrained biomedical language models are proposed. This work suggests a promising direction for an event-based solution for the extraction of molecular mechanisms from biomedical literatures.**

## I. Introduction

This study explores the feasibility of textual IE in the medical research domain, aiming to answer mechanistic regulatory questions by extracting MRMs from PubMed literature. Natural language-based IE can speed up medical research and development by providing timely knowledge. LMs can support question answering systems for MRMs by generating deep embeddings to train a classifier to mine biological concepts from text. Experts can use such systems to answer specialized questions. Recent advances in neural language modeling, such as the transformer, have enhanced natural language IE through large scale semi-supervised pre-training. Pre-trained Language Models (PLMs trained with tasks like MLM on huge unlabeled domain specific corpuses show excellent performance in their domain with minimal supervised fine-tuning. IE can automate the building of a knowledge graph, which can serve as a basis for an MRM question answering and inference system. Such automation can overcome the challenges and costs of keeping question answering systems updated. A system that can answer biological regulatory questions could help researchers discover new drugs and potential causes of disease.

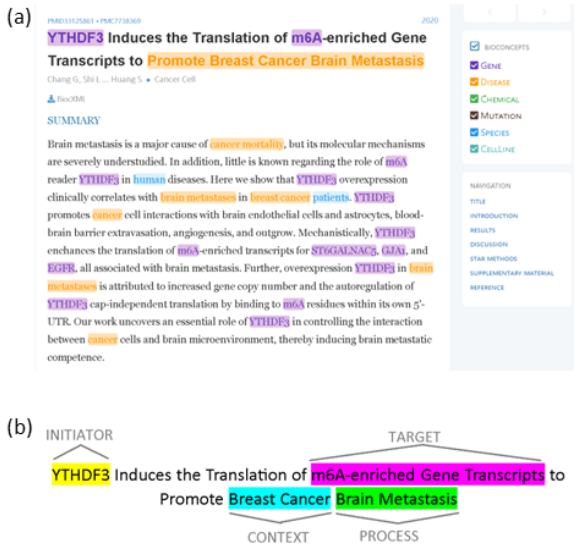


Figure 1. (a) Example PubTator named-entity-extraction (b) Event extraction from title

Current natural language IE techniques used in the field of biomedicine involve extracting information from the abstract or entire document utilizing named entity recognition, relation extraction, or sometimes event extraction (EE). However, abstract based techniques are not optimal, as they overlook the importance of the title in PubMed publications which capture the novelty of their MRM findings. Many titles contain the key details of these MRMs, and extracting information from the title in isolation as a first step greatly reduces the risk of the model failing to identify the key findings. Moreover, this approach reduces the annotation labor required for fine-tuning a PLM. Named entity recognition-based techniques such as PubTator (Fig. 1) are not suitable to extract MRMs because they cannot effectively organize these entities into a mechanistic summary that incorporates the biological context of the regulation. Fig. 1 shows an example of the named entity recognition performed by PubTator and how it compares with MRM extraction via EE. Relation extraction methods such as Peng et al, 2018 are not suitable for representing MRMs because they can only accommodate the relationships between two entities at a time. MRM extraction is contrasted with this type of relation extraction in Fig. 2, demonstrating that relation extraction fails to capture the additional context that EE can

[1]Department of Electrical and Computer Engineering (ECE), University of Pittsburgh, Pittsburgh, USA.
[2]Department of Computer Science (CS), University of Pittsburgh, Pittsburgh, USA.
[3]Department of Medicine, University of Pittsburgh, Pittsburgh, USA.
[4]UPMC Hillman Cancer Center, University of Pittsburgh Medical Center, Pittsburgh, USA.

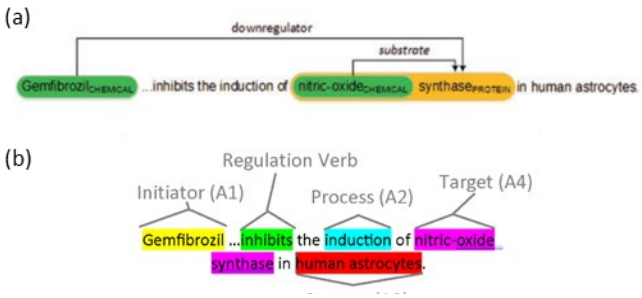


Figure 1. (a) Relation extraction from Peng et al, 2018 (b) Labelling based on EE template proposed in this work.

provide. EE is a better option for engineering the extraction of MRMs, as it allows for the design of a template with any number of arguments, where each argument corresponds to an entity playing a key role in an MRM. However, previous EE approaches for biomedical IE work such as Hope et al, 2020 have not treated the title as a sentence separate from the abstract with high importance in facilitating extraction as the first information seeking step. In addition, these prior biomedical EE methodologies have not proposed the use of an EE template designed for MRMs.

In order to capture the details required for a biomedical expert to answer questions about mechanistic interactions involving drugs, RNAs, proteins, genes, pathways, and diseases, this study explores the fine-tuning of biomedical PLMs utilized to generate richly contextualized embeddings that are crucial for training a high-quality EE classifier. The MRM EE template, which consists of four critical arguments along with the regulation verb, is developed to enable the fine-tuning of each PLM and the training of its corresponding classifier. Subsequently, the first MRM EE dataset is constructed, comprising approximately 400 titles from "N6-methyladenosine (m6A)" related PubMed articles, where each title is annotated with the regulation verb and the four arguments for the corresponding MRM. We chose the topic of m6A to create a training dataset, as it is an emerging area with a manageable number of published papers to annotate. DYGIE++ (Wadden et al, 2019) was then employed to perform both the fine-tuning of each PLM and the training of the EE classifier, leveraging the features provided in the embedding-vector output from the PLM in order to learn how to identify the components of the MRM event template. The results from this process are then presented to demonstrate the potential of this technique in producing a question answering and inference framework for MRMs.

## II. Training Data Generation

To define an EE template for MRMs, a set of regulation verbs were identified that signified some form of either positive or negative biomolecular regulation. These regulation verbs were then used to filter for titles with MRMs, along with a set of filter words that were used to identify titles that were likely to not contain an MRM. The MRM event template design includes arguments that can be filled by any of the following: genes, cellular and molecular processes, cell types, diseases, proteins, RNAs, chemicals, pathways, physiological conditions, specie names, and organ types. There are four of these arguments, and all arguments relate to both the regulation verb and other arguments in the MRM event

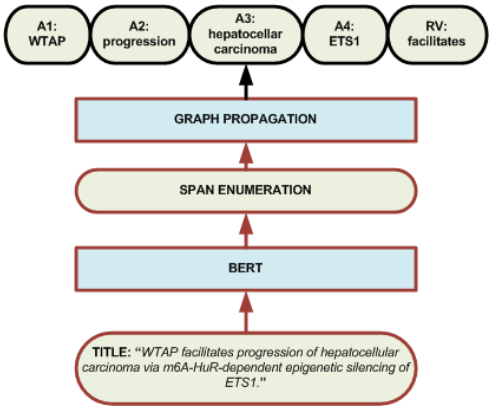


Figure 2. DYGIE++ architecture adapted to title event extraction.

template. Argument one is best described as the initiator, which is most often a gene, but can also be a protein, RNA, small molecule, physiological condition, or in some cases a sub-event that must occur in order to trigger the mechanism. The initiator is the argument that causes the mechanism and can be viewed as the furthest upstream argument. Argument two is the process, or the entity that the mechanism is regulating. The third argument is the context, which adds additional information to the process, such as what disease the process occurs in, what organ is affected by the process, or more information about the identity of the process. The context argument is usually a disease, but can also be an organ, species name, RNA, or protein, depending on the scope of the biological mechanism and whether it is describing something on a micro or macro scale biologically. The fourth and final argument is the target. This is the entity that is acted upon by the initiator argument in order to cause the mechanism. The target argument is most commonly a gene, but it can also be an RNA, protein, pathway, or more specific site on some protein or RNA. Using these definitions, the full event template can be stated: the *initiator* acts upon the *target* to regulate the *process* in the provided *context*. For instance, a title from the dataset states, "METTL3 Promotes Esophageal Squamous Cell Carcinoma Metastasis Through Enhancing GLS2 Expression." In this example the initiator gene is "METTL3", the regulation verb is "Promotes", the process is "Metastasis", the context is "Esophageal Squamous Cell Carcinoma", and the target gene is "GLS2."

We extracted 2,656 papers published between 2013-2021 using the keyword "m6A" from PubMed using PubTator. A total of 400 titles that contained a regulatory mechanism were kept. For each title, arguments of the MRM EE template were annotated by two annotators and an additional annotator was involved to resolve the discrepancies. An example of a title that does not contain an MRM is, "FTO Plays an Oncogenic Role in Acute Myeloid Leukemia as a N6-Methyladenosine RNA Demethylase", as it contains no regulation verb and provides no mechanistic details of how "FTO" affects "Acute Myeloid Leukemia" through a form of biomolecular regulation. In contrast, "METTL3/YTHDF2 m6A axis accelerates colorectal carcinogenesis through epigenetically suppressing YPEL5", contains an MRM because the verb "accelerates" expresses what form of regulatory affect the "METTL3/YTHDF2 m6A axis" has on "colorectal carcinogenesis", and states how this mechanism functions

through targeting "YPEL5". Out of the 400 titles that could be annotated some were difficult to label because of multiple entities mapping to the same argument or because one of the arguments was missing. An example of a title made challenging due to the presence of an extra entity reads, "WTAP promotes myocardial ischemia/reperfusion injury by increasing endoplasmic reticulum stress via regulating m6A modification of ATF4 mRNA." This title has two potential target arguments, "endoplasmic reticulum stress" and "ATF4 mRNA". In the title, "endoplasmic reticulum stress" is increased. In order to increase it "ATF4 mRNA" must be targeted first through regulating its m6A modification. This case is addressed by labeling the final gene in the title as the target, or in this case "ATF4 mRNA." An example of a title missing an argument is, "METTL3-mediated m6A mRNA modification of FBXW7 suppresses lung adenocarcinoma", where the process argument is missing and the regulation verb acts upon the entity that would usually be the context. The solution to this case is to have no labeled context, and to label "lung adenocarcinoma" as the process being regulated.

## III. Methods

Training was performed with DYGIE++, (Wadden et al, 2019), which fine-tunes a Bidirectional Encoder Representations from Transformers (BERT) PLM for producing the embedding vectors necessary for using an EE classifier. The model uses a dynamic span graph algorithm to greedily produce all potential MRM argument spans. A feed-forward network is trained to prune the graph by eliminating all span-pairs that fall below the probability threshold for event argument classification.

$$\mathbf{u}_x^t(i) = \sum_{j \in B_x(i)} V_x^t(i,j) \odot \mathbf{g}_j^t \qquad (1)$$

Equation (1) shows how the dynamic span graph algorithm is applied to training in order to update what trigger-argument span pairs are most probable to belong in the event template being learned. This involves a step where the event classifier produces a score for all trigger-argument span pairs, followed by a final pruning step where all triggers and arguments that are unlikely to belong in the event are removed. In this equation the vector **g** is the embedded span representation in the graph, vector **u** is the updated span embedding after DYGIE++ has performed an iteration of fine-tuning, the matrix V represents a feed forward neural network, and the dot within a circle represents an element-wise multiplication. Fig. 3 provides a schematic for DYGIE++, and the different steps the model takes in order to extract an event from a single title.

## IV. Results

All models were trained with 396 of the total 500 titles in the m6A dataset, including any title that contained at least one regulation verb along with two or more of the four arguments in the MRM event template. Each model was trained using 5-fold cross-validation, where three of the five splits were combined for the training set, one split was used as the development set, and the final split was held out as the test set. Three different baseline model encoder types were trained: PubMed BERT Base Uncased, Clinical PubMed BERT Base Uncased, and PubMed BERT Large Uncased. The predictions

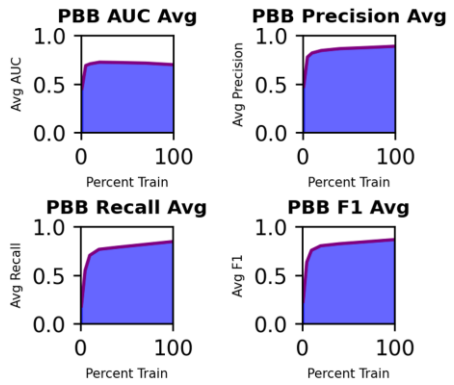


Figure 3. PubMed BERT-base performance metrics vs percent train.

across all five folds were scored as true positives, false negatives, and false positives using the standards outlined by the state-of-the-art ACE event extraction literature (ACE 2005). Precision, recall, and F1 metrics were calculated for all arguments and the regulation verb across all three models. ROC curves were produced for the regulation verb and each argument using the ROC scikit-learn implementation, and the AUROC was calculated using the same package.

In order to show how the m6A MRM dataset was affecting the performance of the models, experiments were run using varied percentages of the available training data. These experiments were structured so that all three model types saw the same sets of training examples for each fold one through five, so that data instance variation could not give one model a random advantage over the others. All three models were trained with one, five, ten, twenty, forty, seventy, and one hundred percent of available training data. A sample of the results are shown in Table 1.

The baseline model used SciBERT as the PLM and saw F1 metric scores above 0.8 for all arguments save the target. The regulation verb achieved the highest extraction accuracy with an F1 of 0.97. The initiator performed the best of the arguments, with an F1 of 0.86. The other three PLMs pretrained on PubMed data all outperformed the baseline, except for PubMed Bert Large, when fine-tuned with maximum training examples. The best of these models, PubMed BERT Base, improved the F1 across all four arguments by approximately 0.03 for each. Only the regulation verb failed to improve when compared to the baseline. A more detailed summary of the F1 metrics can be viewed in Table 1. All four of these models perform flawless extraction on the majority of titles that easily fit the template such as, "DDX3 modulates cisplatin resistance in OSCC through ALKBH5-mediated m6A-demethylation of FOXM1 and NANOG." In this example the baseline model correctly extracts "DDX3" as the initiator, "cisplatin resistance" as the process, "OSCC" as the context, "FOXM1 and NANOG" as the target, and "modulates" as the regulation verb. These models make a mistake in one of every six titles. The most frequent forms of mistake are either missing an argument, missing a token, or extracting an extra token. An example where the model struggles and misses the target is, "Hypoxia Promotes Vascular Smooth Muscle Cell (VSMC) Differentiation of Adipose-Derived Stem Cell (ADSC) by Regulating Mettl3 and Paracrine Factors." Here the baseline model correctly extracted the initiator "Hypoxia", the process "Vascular

TABLE I. F1 Comparison of Models Trained with 1, 10 and 100% of Dataset.

| Model | Argument 1 F1 | | | Argument 2 F1 | | | Argument 3 F1 | | | Argument 4 F1 | | | Regulation Verb F1 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1% | 10% | 100% | 1% | 10% | 100% | 1% | 10% | 100% | 1% | 10% | 100% | 1% | 10% | 100% |
| PubMed BERT Base | 0.13 | 0.81 | 0.90 | 0.21 | 0.65 | 0.82 | 0 | 0.77 | 0.85 | 0 | 0.64 | 0.78 | 0.81 | 0.91 | 0.97 |
| PubMed BERT Large | 0 | 0.79 | 0.86 | 0 | 0.52 | 0.77 | 0.02 | 0.56 | 0.78 | 0.01 | 0.59 | 0.72 | 0.23 | 0.91 | 0.96 |
| Clinical PubMed BERT | 0 | 0.78 | 0.88 | 0.14 | 0.68 | 0.80 | 0.19 | 0.79 | 0.83 | 0 | 0.54 | 0.75 | 0.91 | 0.92 | 0.96 |
| SciBERT | | | 0.86 | | | 0.80 | | | 0.81 | | | 0.75 | | | 0.97 |

Smooth Muscle Cell (VSMC) Differentiation", the context "Adipose-Derived Stem Cell", and the regulation verb "Promotes", but incorrectly predicts no target, when "Mettl3 and Paracrine Factors" is the span that should be extracted for the target.

Results from the training variation experiment show that across the three pretrained BERT models, precision, recall, and F1 rapidly increase with more training examples up until about ten to twenty percent of the training splits, translating into about twenty to fifty training titles. Most arguments have an inflection point in the range of twenty to fifty titles, where performance gains from additional titles is lessened, perhaps only making a difference if hundreds or even thousands more training titles were annotated. Fig. 4 shows this trend plotted across all four metrics for the best model, PubMed BERT Base. Experiments were run using as little as two training titles per fold, and when this was performed all three models could not achieve an F1 score greater than 0.25 during argument classification. According to the AUROC plots, adding additional titles had no clear positive or negative trend after using about fifty titles. When looking at specific arguments or the regulation verb in isolation, as displayed in Table 1, there are some more interesting trends of greater magnitude. For example, some arguments gained a lot in precision or recall by adding more examples for a particular BERT model, but then the same argument when using a different BERT model did not see the same boost in performance with more training titles. For some models the regulation verb saw a significant drop in classification performance when more training examples were used. In some cases, arguments or the regulation verb saw a drop in performance around fifty percent, and then the performance climbed back up to a maximum performance at one-hundred percent training titles. Reasons for these differences include the fact that these BERT models have seen different pretraining data, and one BERT model has a better latent understanding of a given argument over another PLM. Another reason could be that some training examples in the dataset could be more confusing or less well annotated than others. Such instances could slightly damage the latent space representation until BERT sees more examples that allow it to climb out of a local optimum with respect to error. Overall, the additional training examples have no negative effect on AUROC, and serve to boost precision, recall, and F1 across the board. This shows that well labeled EE datasets such as the PubMed title m6A MRM dataset are necessary, and that current PLMs cannot perform EE as well in a near zero-shot setting.

## V. CONCLUSION AND FUTURE WORK

Title-based extraction of MRMs shows potential for future biomedical information retrieval and inference systems, but there is still room to explore and refine. Although many quality labels were produced for training and evaluation, the labels are still far from perfect. This is where the model can sometimes help identify inconsistencies in the annotations when reviewing the summary of the differences in the actual and predicted events. For example, if the labeler missed labeling a process or outcome such as "apoptosis" in the actual outcome argument value, the model might predict "apoptosis" as part of the outcome argument anyway. Further simplification of the gold labels in the m6A dataset is needed, so that the labels only include tokens that are crucial to the informational completeness of the EE. On top of this, having wordy annotations will make it more likely the model is penalized during evaluation for not including tokens that have no real bearing on whether extraction is correct from the perspective of MRMs. There may also be room for improvement by designing sub-templates of regulation instead of limiting extraction to a single template. This could help if using multiple templates offers a clearer dividing line between arguments. Moving forward, a second text classifier is being developed in order to recognize titles that contain MRMs, so that the EE classifier can be properly deployed on a wide range of PubMed articles.

Extraction from the title works well as the ignition step for biomedical IE. The limited variation in title structure and the concise nature of the title can function as an IE guide if it does not contain the critical MRM knowledge of a Bio Medical document outright. If the important information to be identified about a molecular mechanism is not in the title, then the abstract can be harnessed to help fill in gaps. Once the pipeline can reliably extract knowledge of MRMs, then investigating ways to construct a knowledge graph is the next step. With a knowledge graph that correctly represents knowledge of molecular mechanisms we can experiment with search algorithms. Our final goal is to assess what level of information retrieval and inference can be performed with our strategy. Furthermore, we plan to assess whether there exists a transformer-based decoder, such as a PubMed GPT, that could be fine-tuned to directly answer research questions pertaining to MRMs.

## VI. REFERENCES


[1] ACE (Automatic Content Extraction) English Annotation Guidelines for Events, Version 5.4.3 2005.07.01, Linguistic Data Consortium. http://www.ldc.upenn.edu/Projects/ACE/.

[2] Tom Hope, Aida Amini, David Wadden, Madeleine van Zuylen, Sravanthi Parasa, Eric Horvitz, Daniel Weld, Roy Schwartz, Hannaneh Hajishirzi. Extracting a Knowledge Base of Mechanisms from COVID-19 Papers. arXiv (https://arxiv.org/abs/2010.03824), 2020.

[3] Yifan Peng, Anthony Rios, Ramakanth Kavuluru, Zhiyong Lu, Extracting chemical–protein relations with ensembles of SVM and deep learning models, *Database*, Volume 2018, 2018, bay073, https://doi.org/10.1093/database/bay073.

[4] David Wadden, Ulme Wennberg, Yi Luan, Hannaneh Hajishirzi. Entity, Relation, and Event Extraction with Contextualized Span Representations. arXiv (https://arxiv.org/abs/1909.03546), 2019.

[5] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Édouard Duchesnay. Scikit-learn: Machine Learning in Python. JMLR 12, pp. 2825−2830, 2011.