SuMe: A Dataset Towards Summarizing Biomedical Mechanisms

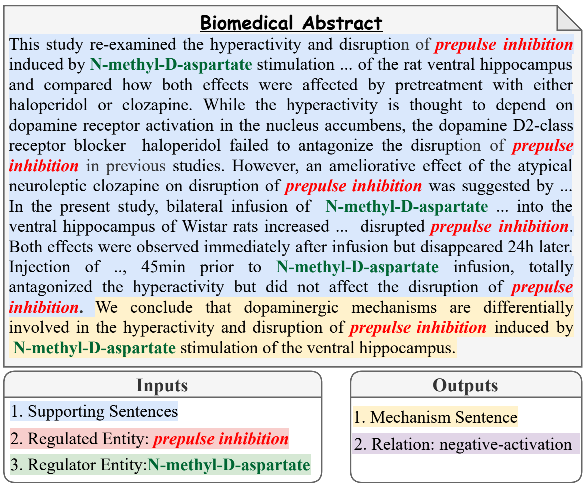
Anonymous ACL submission

Abstract

Can language models read biomedical texts and explain the biomedical mechanisms discussed? In this work we introduce a biomedical mechanism summarization task. Biomedical studies often investigate the mechanisms behind how one entity (e.g., a protein or a chemical) affects another in a biological context. The abstracts of these publications often include a focused set of sentences that present relevant supporting statements regarding such relationships, associated experimental evidence, and a concluding sentence that summarizes the mechanism underlying the relationship. We leverage this structure and create a summarization task, where the input is a collection of sentences in an abstract and the output includes the main relationships and a natural language sentence that summarizes the mechanism. Using a small amount of manually labeled mechanism sentences, we train a mechanism sentence classifier to filter a large biomedical abstract collection and create a summarization dataset with 22k instances [1]. We also introduce a pretraining conclusion generation task with 611k samples. Our benchmarking experiments with large language models show that the pretraining is helpful for the original task, but the model performance isn't still satisfactory and this task presents significant challenges in biomedical language understanding and summarization.

1 Introduction

Understanding biochemical mechanisms such as protein signaling pathways is one of the central pursuits of biomedical research (Arighi et al., 2011; Krallinger et al., 2017; Demner-Fushman et al., 2020). Biomedical research has advanced tremendously in the past few decades, to the point where we now suffer from "an embarrassment of riches:" publications are generated at such a rapid pace (PubMed[2] has indexed more than 1 million publications per year in the past 8 years!) that these mechanisms must be summarized, if humans are to keep up with the big picture behind this massive body of work. In this paper we introduce a novel dataset and task that couples elements of biochemical mechanisms with their textual summaries. In this initial effort, we focus on individual elements of these mechanisms, i.e., single interactions (positive or negative activations) between pairs of biochemical entities such as proteins. In particular, we introduce an instance of an explainable relation extraction problem, where interactions between two biochemical entities are mechanistically summarized in plain text. The proposed task is coupled with a novel dataset called SuMe, which should facilitate the development of methods that can extract and explain biomedical mechanisms. The contributions of this paper are the following:



Biomedical Abstract

This study re-examined the hyperactivity and disruption of ***prepulse inhibition*** induced by **N-methyl-D-aspartate** stimulation ... of the rat ventral hippocampus and compared how both effects were affected by pretreatment with either haloperidol or clozapine. While the hyperactivity is thought to depend on dopamine receptor activation in the nucleus accumbens, the dopamine D2-class receptor blocker haloperidol failed to antagonize the disruption of ***prepulse inhibition*** in previous studies. However, an ameliorative effect of the atypical neuroleptic clozapine on disruption of ***prepulse inhibition*** was suggested by ... In the present study, bilateral infusion of **N-methyl-D-aspartate** ... into the ventral hippocampus of Wistar rats increased ... disrupted ***prepulse inhibition***. Both effects were observed immediately after infusion but disappeared 24h later. Injection of .., 45min prior to **N-methyl-D-aspartate** infusion, totally antagonized the hyperactivity but did not affect the disruption of ***prepulse inhibition***. We conclude that dopaminergic mechanisms are differentially involved in the hyperactivity and disruption of ***prepulse inhibition*** induced by **N-methyl-D-aspartate** stimulation of the ventral hippocampus.

Inputs

1. Supporting Sentences
2. Regulated Entity: ***prepulse inhibition***
3. Regulator Entity:**N-methyl-D-aspartate**

Outputs

1. Mechanism Sentence
2. Relation: negative-activation

Figure 1: Example of an entry in the SuMe dataset. Some supporting text was removed to save space. The input is the supporting sentences with the main two entities. The output is the relation type and a sentence concluding the mechanism underlying the relationship.

[1] dataset will be published upon acceptance of the paper

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

(1) We introduce the SuMe dataset, which is constructed semi-automatically from publication abstracts. The dataset contains tuples of support sentences, mechanistic information such as the two biochemical entities in focus and the relation that holds between them, and a textual summary of this interaction (see Figure 1). The entities and relations are extracted using an existing biomedical information extraction system (Valenzuela-Escárcega et al., 2018). The mechanism summaries are extracted using a semi-automatic bootstrapping process. First, with the help of biomedical experts, we gathered a small set of mechanism sentences. We then train a mechanism sentence classifier by fine-tuning BioELECTRA (Kanakarajan et al., 2021), a biomedical domain language model (LM). We use this LM to collect a large set of approximately 22k mechanism summarization instances. The entire dataset construction is summarized in Figure 2. Five domain experts manually evaluated the quality of a dataset sample of 125 instances, and concluded that the generated dataset has reasonable quality.

(2) Using the above manually-curated sample, we evaluated the capacity of multiple neural LMs to generate the underlying biochemical relations, and the corresponding mechanism sentences. In particular, we analyzed GPT2 (Radford et al., 2019), scientific GPT2 (Papanikolaou and Pierleoni, 2020), T5 (Raffel et al., 2020a), SciFive (Phan et al., 2021), and BART (Lewis et al., 2019). The results indicate that the proposed task is quite challenging. We also defined a pretraining task with 611K instances to improve these LMs. In summary, this first empirical benchmark and analyses indicate that this is a meaningful and complex research problem that deserves further investigation.

2 Related Work

We address mechanism generation, which can be seen as a combination of explainable relation extraction and summarization. There is a huge body of work that addresses explainable methods (e.g., for relation extraction (Shahbazi et al., 2020) or explainable QA (Thayaparan et al., 2020)). Many prior works in relation and event extraction treat explanations as the task of selecting or ranking sentences that support a relation (e.g., (Shahbazi et al., 2020; Ghaeini et al., 2019; Lev et al., 2019; Çano and Bojar, 2020; Yasunaga et al., 2019)). Our work differs from these in that it focuses on *generating mechanisms* underlying a relation from supporting sentences, rather than identifying existing sentences.

Our work can also be viewed in the context of reading and generating information from scientific texts. Most work in this area focus on generating summaries using scientific publication and some times in combination with external information (Yasunaga et al., 2019; DeYoung et al., 2020; Collins et al., 2017; Wang et al., 2018a, 2019)
Some works even seek to generate part of the scientific papers. For example, TLDR (Cachola et al., 2020) introduces a task and a dataset to generate TLDRs for papers. They exploit titles and an auxiliary training signal in their model. ScisummNet (Yasunaga et al., 2019) introduces a large manually annotated dataset for generating paper summaries by utilizing their abstracts and citations. TalkSumm (Lev et al., 2019) generates summaries for scientific papers by utilizing videos of talks at scientific conferences. PaperRobot (Wang et al., 2019) generates a paper's abstract, title, and conclusion using a knowledge graph. FacetSum (Cohan et al., 2018) used Emerald journal articles to generate 4 different abstractive summaries, each targeted at specific sections of scientific documents. Nikolov et al. (2018) introduce two novel multi-sentence summarization datasets from scientific articles, and test the suitability of a wide range of existing extractive and abstractive neural network-based summarization approaches, e.g., generate abstracts from paper content, and generate titles from abstracts. Wang et al. (2018b) generate abstracts as a conditioned, iterative text generation problem, and design a new writing-editing network with an attentive revision gate to iteratively examine, improve, and edit the abstract. More recently, Meng et al. (2021) introduce a new dataset to generate 4 different summaries for long scientific documents.

In addition to the specifics of the output that we target, our work is different from all these other works because our proposed summarization task is grounded with the underlying biomedical event discussed, rather than focusing on generic summarization, which may lose the connection to the underlying biology that is the core material discussed in these papers.

3 Mechanism Summarization

Our goal is to develop a task and a dataset that pushes models towards understanding the mechanisms that underlie the relationships between enti-

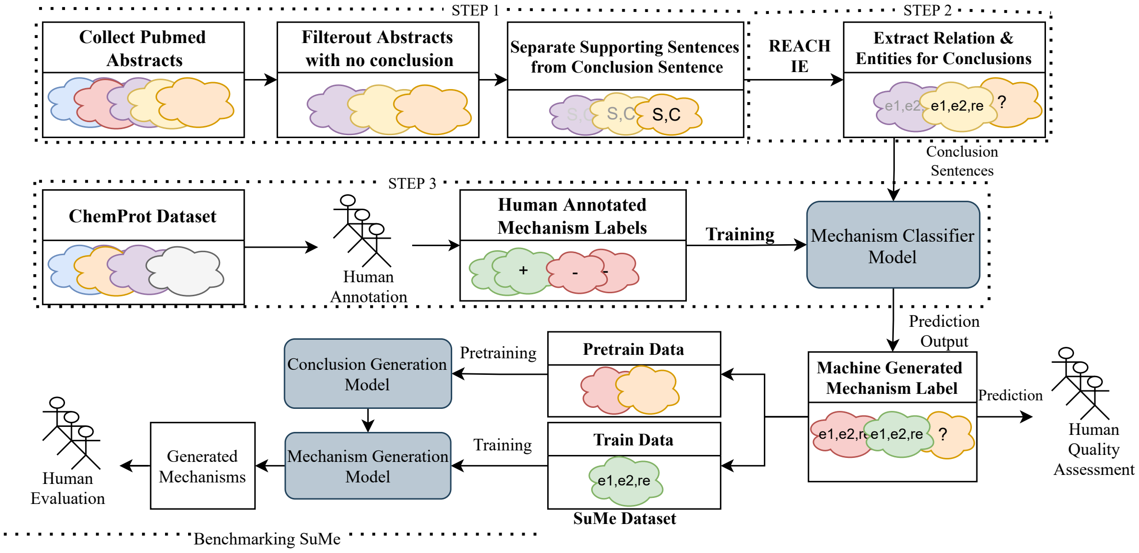


Figure 2: The overall bootstrapping pipeline for SuMe dataset collection and human evaluation. The main idea behind the pipeline is to collect relatively easy to acquire judgments from domain experts to then bootstrap and generate a weakly-labeled large training corpus. We further assess the quality of the resulting dataset through another round of human evaluation, which also yields a smaller curated evaluation dataset.

ties from biomedical literature. From a language processing perspective, we can view mechanisms as a form of explanation that justifies the relationship or connection between entities. From a biomedical science perspective, a mechanism provides two types of explanatory information, which we use to characterize mechanism sentences:

Why is the relation true? A sentence can be a mechanism, if it explains *why* the relation exists between the two main entities. For example, one protein (say A) might be up-regulate another (say B), which in turn inhibits yet another protein (say C). This provides the causal reasoning to conclude the relation that protein A inhibits protein C.

How does the relation come about? Another kind of explanatory information is the one that describes the process or manner in which the relation exists between the pair of entities. For example, one protein (say A) may activate another protein (say B) via a specific process.

These provide a way to specify what constitutes a mechanism sentence and help us to locate mechanism sentences in the literature. In particular, we consider abstracts which discuss studies that lead to conclusions about such mechanisms. Typically, these abstracts provide a short collection of sentences that describe the goals of the study, the methods used, the experimental observations, the findings, which can be used to substantiate the conclusions that establish the relation of interest, and the mechanism underlying the relation. This suggests a language processing task that tests for ability to understand biomedical mechanisms: given the preceding sentences in the abstract can a model accurately generate the underlying mechanism?

In this section, we first formally define this task, and then describe the auxiliary tasks we devised to help generating such explanations.

3.1 Task Definition

Given a set of sentences from a scientific abstract (referred to as *supporting sentences*) and a pair of entities (e_i, e_j) that are the focus of the abstract, generate the *conclusion sentence* that explains the mechanism behind the pair entities and output a relation that connects these entities (e.g., positive_activation(e_i, e_j)). Figure 1 shows an example of such a tuple of supporting sentences, focus entities, relation, and mechanism sentence. As the example illustrates, mechanism sentences describe some pathway often involving another entity or a process (e.g., *dopaminergic mechanism*), require identifying and combining information from multiple relevant sentences, and non-trivial inferences regarding the relationship between the entities (e.g., recognizing that the different effects on *prepulse inhibition* imply differential involvement).

Given an abstract of a scientific literature we need four pieces of information: 1. The two focus entities of the abstract. 2. The relation between enties. 3. Sentences from the abstract in support of this relation. 4. The conclusion sentence where the mechanism underlying the relation is summarized.

4 SuMe Dataset

We aim to create a large scale dataset for the mechanism summarization task defined above. However, identifying instances for this task requires domain expertise and cannot be easily done at scale. Instead, here we employ a bootstrapping process, where we first annotate a small amount of data to build a mechanism sentence classifier that can then helps us collect a large scale dataset for mechanism summarization. The key observation here is that identifying sentences that express a mechanism is a simpler task than targeted mechanism summarization task, and, thus, should be learnable from smaller amounts of data. We outline the process we use for creating our mechanism summarization dataset, SuMe, and an expert evaluation of its quality next.

4.1 SuMe Construction Process

We construct SuMe using biomedical abstracts from the PubMed open access subset[3]. Starting from 1.1M scientific papers, we followed the following sequence of bootstrapping steps to prepare the SuMe dataset. The following steps are also elaborated in Figure 2.

1. Finding Conclusion Sentences: First, we use simple lexical patterns to find abstracts with clearly specified conclusion sentence. All abstracts which has any form of *conclude* word (*conclusion, concluded, concluding, concludes*, etc.) at the very end of the text are extracted here. We use this matching process to also split the abstracts into the set of supporting sentences (the ones that lead up to the conclusion) and the conclusion sentence.

2. Extracting Main Entities & Relation Starting with the abstracts which are now in the form of (supporting sentences, conclusion sentence), we then run a biomedical relation extractor, REACH (Valenzuela-Escárcega et al., 2018), which can identify protein-protein and chemical-protein relations between entities. In this work, we focus on the relations where one entity is the controller and another entity is the controlled entity and the relation between them is either *positive/negative activation* or *positive/negative regulation*. If an abstract doesn't return any such relation, we keep that for the pretraining step (as described in Section 5.3), otherwise we use it for the main task.

3. Filtering for Mechanism Sentences: We then filter out the instances to only retain those whose conclusion sentences are indeed a mechanism sentence. To this end, we devised a bootstrapping process where we first collect supervised data to train a classifier. To collect likely mechanism sentences we made use of the ChemProt (Peng et al., 2019) relation extraction dataset which contains sentences annotated with positive and negative regulation relations between entities. However, not all of these sentences necessarily explain the mechanism behind these relations. We asked 21 experts (grad students in a biomedical department) to inspect each sentence and rate whether it explains the mechanism behind the ChemProt annotated relation on a four-point Likert scale. For each sentence, an annotator can select between *Clearly a Mechanism, Plausibly a Mechanism, Clearly not a Mechanism,* and *Not Sure.* Each sentence is annotated by three experts and we find the inter-annotator agreement between users to be $\kappa = 73\%$ (Fleiss Kappa (Landis and Koch, 1977)). The final label for a sentence is selected based on the majority voting after combining *Clearly a Mechanism* and *Plausible a Mechanism* labels. Finally, each sentence is labeled as a *Mechanism*, or *Non-Mechanism*. The resulting dataset contained 439 *Mechanism* sentences and 447 *Non-Mechanism* sentences.

Using this small scale mechanism sentence dataset, we train classifiers to identify mechanism sentences, where the positive label indicates that the underlying sentence is a mechanism sentence. We compared the performance of finetuning BioBERT (Lee et al., 2020), SciBERT (Beltagy et al., 2019), BiomedNLP (Gu et al., 2020), and BioELECTRA (Kanakarajan et al., 2021) models. BioELECTRA performed the best with 74% macro F1 for mechanism sentence classification. We use the trained mechanism sentence classifier to label all conclusion sentences from the previous step and instances with the predicted mechanism sentences are used to create SuMe dataset.

We separate out the abstracts for which the conclusion sentences are predicted to have non-mechanism related conclusions as additional related data that can be use for pretraining the generation models we eventually train for the mechanism summarization task (as we describe in Section 5.3).

The above procedure results in a dataset that al-

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

Dataset	Train	Dev	Test
Abstracts	20765	1000	1000
Avg. #words in conc.	33.7	34.9	33.5
Avg. #words in supp.	187.5	187.9	186.7
Avg. #sent. in supp.	12.15	12.44	12.33
#Unique controller	8094	759	777
#Unique controlled	6684	717	687
#Unique pair entities	19229	988	989
#Unique entities	12685	1357	1364

Table 1: Dataset Statistics: Each dataset contains a number of unique abstracts, a supporting set (supp.), a mechanism sentence (conc.) a pair of entities. The first entity is called the regulator entity (regulator) and the second one is called the regulated entity (regulated)

lows us to define the following mechanism summarization task: Given a set of supporting sentences from an abstract and a pair of entities (e_i, e_j), generate a relation that connects these entities and a sentence that explains the mechanism that was the focus of the study. The statistics of the dataset are shown in Table 1.

4.2 SuMe Quality

Our goal was to create a large scale albeit bootstrapped dataset that can be used to train large language generation models. A key question to answer here is what is the quality of the resulting dataset. To assess this we asked three biomedical experts to evaluate a random sample from the dataset. The experts were given the set of input supporting sentences, the potential mechanism sentence, and the relation between main entities. They were asked the following three questions

1. Is the expected output relation associated with the instance valid?
2. Is the output sentence expected for this sentence an actual mechanism sentence?
3. Can the mechanism and relation be concluded given the input supporting sentences?

The first question checks for the quality of the automatically extracted relations. The second assesses the impact of the mechanism sentence classifier. Answers to these first two can help estimate the noise in the dataset. The final question helps quantify what fraction of times the information to generate the mechanism sentence is not entirely part of the input supporting sentences, which can make for harder instances requiring external knowledge.

We asked 5 biomedical experts to evaluate 125 randomly selected samples. The purpose of having this set is two fold, first to evaluate the quality of the data collection process, second to collect a clean human evaluated dataset which can be used as an extra test set. The results of the dataset evaluation are shown in Table 2. This evaluation shows that the generated dataset is of reasonable quality, and can serve as a meaningful resource for training models for biomedical summarization.

Quality	Correct
Entities & Relation Extraction	90%
Mechanism Sentence Classifier	85%
Concludable	86%
All Acceptable	81%

Table 2: Dataset Quality: We asked three main questions. This table shows what percentage of each category is acceptable. The last question shows what percentage of the sentences are approved in all questions.

5 Evaluation

Our evaluation focuses on the following questions:

1. Benchmarking: What is the performance of generic and domain-adapted large scale language generation models on SuMe?
2. Effect of pretraining: What is the impact of using the additional data via pretraining?
3. Effect of modeling supporting sentences: What is the impact of selecting a subset of supporting sentences?
4. Error analysis: What are the main failure modes of language generation models?

5.1 Experimental Setup

We use SuMe to benchmark language generation models and measure their ability to correctly identify the relation between the focus entities and to summarize the mechanism behind the relation based on the input sentences from the abstract.

Models: We compare pretrained GPT-2 (Radford et al., 2019), T5 (Raffel et al., 2020b), BART (Lewis et al., 2019) models and two domain-adapted models, GPT2-Pubmed (Papanikolaou and Pierleoni, 2020), and SciFive (Phan et al., 2021), which were trained on scientific literature.

Evaluation Metrics: We conduct both automatic and manual evaluation of the model outputs.

Relation Generation (RG): The models are supposed to generate the relation type with a marker in and then generate the mechanism that underlies this relation. There are two types of relations in the dataset: positive and negative. We evaluate the model's output as we would for a corresponding

Model	RG (F1)	BLEURT	Rouge-1	Rouge-2	Rouge-L
BART	76	42.49	46.54	25.92	35.34
GPT2	74	44.19	46.54	28.32	38.78
T5	72	44.41	48.26	27.63	38.77
GPT2-Pubmed	78	46.33	48.37	29.55	40.19
SciFive	**79**	**47.81**	**52.10**	**32.62**	**43.31**

Table 3: Benchmarking performance of strong language generation models and some domain-adapted models. We present standard automatic evaluations measures for the mechanism sentence generation task along with F1 for the generated relations. The science domain versions of both GPT2 and T5 work better than the original versions.

classification task, i.e., the generated relation is deemed correct if it exactly matches the correct relation name. We report F1 numbers for this binary classification task.

Mechanism Generation: We evaluate the quality of the generated explanations using two language generation metrics: the widely-used ROUGE (Lin, 2004) scores, and to address the recent concerns on the usage of these scores in capturing conceptual information (Novikova et al., 2017) we additionally report BLEURT score (Sellam et al., 2020) which is able to better account for more complicated semantic mismatches between the generated sentence and the gold reference. We use a recent version, the BLEURT-20 model (Pu et al., 2021) that has been shown to be more effective. We compare the generated text as the hypothesis with the actual text as the reference.

Fine-tuning and Training Details: All models were fine-tuned on the training portion of SuMe for 20 epochs. For each model, we evaluate the average of BLEURT and Rouge-L score on the validation set and the one with the highest score is chosen for prediction. The learning rate is set to 6e-5, we use AdamW (Loshchilov and Hutter, 2017) optimizer with $\epsilon = 1e-8$. The input token is limited to 512 tokens, and the generated token is maxed out at 128 tokens. We select batch size of 8 with gradient accumulation steps of two.

5.2 Automatic Evaluation Results

Table 3 compares the performance of the five language generation models on both the relation generation (RG) and mechanism generation tasks.

Fine-tuning the domain-adapted models, GPT2-Pubmed and SciFive, is better than fine-tuning the standard pre-trained models for both relation and mechanism generation tasks. SciFive achieves the best performance with more than a 7.5% increase in BLEURT score and more than 9.7% increase in RG F1 over the standard T5 model, highlighting the importance of domain adaptation for the SuMe tasks defined over scientific literature.

The overall numbers (coupled with the human evaluation in Section 5.5) suggest that mechanism generation is a difficult and challenging task.

The models achieve better performance on the relation generation task but there is still a substantial room for improvement here with the best model achieving an F1 of 79. If the model is unable to generate the relation correctly, then the mechanism it generates is not useful. Ideally we want models to correctly generate both the relation and the mechanism that underlies it. We also evaluated the correlation between BLEURT score and relation generation classification score. Our analysis shows that when the model generates an accurate relation, it get's higher BLEURT score while when it generates an incorrect relation, the BLEURT score is lower by 10%. (50.02 vs 45.08)

5.3 Pretraining with Conclusion Generation

Next we analyze the impact of pre-training the models on a related task of generating conclusion (instead of mechanism) sentences, for which we can obtain data at scale without any labeling effort. SuMe includes 611K instances of this kind which is an order of magnitude larger than the mechanism summarization instances.

We study the effect of this pretraining task by varying the amount of pretraining data. We analyze the impact in terms of the overall effectiveness and the amount of fine-tuning (number of epochs) needed to converge when finetuning.

Pretraining Data Size: We pretrain the SciFive model on the conclusion generation task with increasing amount of data (100K increments), and measure the performance of finetuning the pretrained models on the mechanism summarization task. Figure 3 shows that there is a trend of improved performance suggesting that pretraining is beneficial for learning to generate mechanisms.

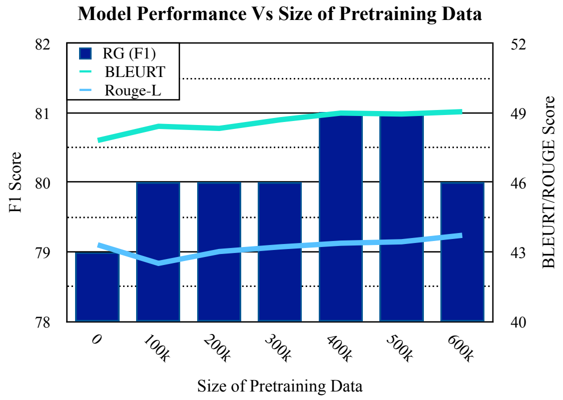


Figure 3: Comparison of relation generation F1 (left y-axis/blue bars) and the mechanism generation measures (right y-axis/teal+Blue curves) against the amount of pretraining. As we increase the size of the pretraining data, the model performance improves in both aspects.

Number of Epochs: We also compare the impact of the amount of pretraining on the number of epochs needed for convergence in fine-tuning. Figure 4 compares pretrained models with different number of pretraining epochs (x-axis) in terms of their overall effectiveness (BLEURT score bars) and the number of epochs to convergence (Fine-tuning epochs curve). The figure shows that when we continue pretraining, not only does the resulting model performs better, but it also converges sooner taking fewer number of epochs to reach higher effectiveness. Together these results suggest potential for the auxiliary data available in the SuMe dataset.

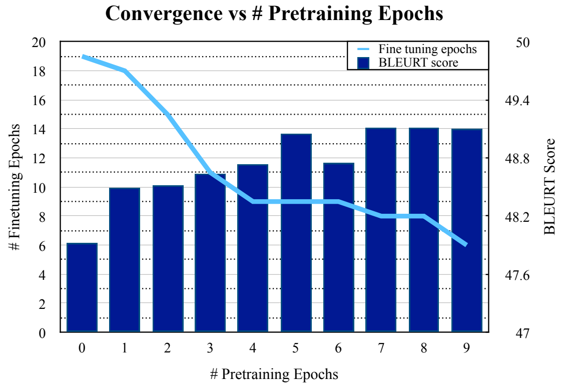


Figure 4: Number of pretraining epochs vs. number of fine-tuning epochs for each pretrained model until convergence.

5.4 Modeling Supporting Sentences

Will it help to model the subset of sentences within the inputs sentences that provide the best support for generating the mechanism sentence? This kind of an extractive step has been used previously in summarization tasks to reduce the amount of irrelevant information in the input (Narayan et al., 2018; Liu and Lapata, 2019). To understand the utility of this, we built a pseudo-oracle that finds the sentences that have the best overlap (measured via BLEURT score (Sellam et al., 2020)) with the output mechanism sentence. Then, we trained the SciFive model and pretrained version to only use the top few sentences according to BLEURT score such that input size is now half of the original input size. We find that this only provides improvements in BLEURT score over using the entire set of input sentences for the basic SciFive model and the gains reduce with the pretrained model. Unlike standard summarization tasks there are fewer completely unrelated sentences in the abstracts and generating the mechanism sentences remains challenging even when we are able to identify the most relevant sentences within this set suggesting that the task remains hard even when the most important sentences are somehow known to the model.

Supporting Set	BLEURT	Rouge-L
SciFive	47.81	43.31
+Oracle	49	43.07
+Pretraining	49.05	43.72
+Pretraining+Oracle	49.64	43.81

Table 4: The effect of selecting supporting sentences with highest BLEURT score.

5.5 Manual Evaluation

We also conduct a manual evaluation of the outputs from the best model — the SciFive model that was pretrained with the conclusion generation task. We asked 3 biomedical experts to evaluate output sentences for 100 instances and answer the following three questions:

1. Does the generated sentence contain a mechanism that explains the relation between the two main entities?
2. Is the information in the generated mechanism supported by the information contained in the supporting set of sentences?
3. Is the generated mechanism factually correct?

The results show that the model learns to generate mechanism sentences that connect the main entities for most of the instances (79%). It also produces output for which there is support in the input sentences in a majority of the cases (53%).

Gold Mechanism	**Generated Mechanism**
On the basis of these observations, we conclude that *IL-18* induces *MCP-1* production through the PI3K/Akt and MEK/ERK1/2 pathways in macrophages.	In conclusion, *IL-18* induces *MCP-1* production in macrophages through the PI3K/Akt and MEK/ERK1/2 pathways.
Taken together, we conclude that *DeltaNp73* negatively regulates *NGF-mediated neuronal differentiation* by transrepressing TrkA.	In conclusion, our results indicate that *DeltaNp73* negatively regulates *NGF-mediated neuronal differentiation* by transcriptionally repressing the expression of TrkA.
We conclude that, without modulatory factors which play a role in vivo, *NGF* can enhance the synthesis of *tyrosine* hydroxylase n sympathetic ganglia in vitro, provided organ culture conditions which permit optimal survival of adrenergic neurons are selected.	This conclusion was further supported by pulse-labeling of *tyrosine* hydroxylase with [3H]leucine, which showed that *NGF* increased synthesis of *tyrosine* in sympathetic ganglia by approximately 50%.
We conclude that 20 mm alcohol/submaximal CCK as well supramaximal CCK stimulation can trigger pathologic basolateral exocytosis in pancreatic acinar cells via PKC alpha-mediated activation of *Munc18c* , which enables *Syntaxin-4* to become receptive in forming a SNARE complex in the BPM	We conclude that alcohol can induce a clinically relevant form of pancreatitis by blocking apical exocytosis and redirecting exocytosis to less efficient BPM, mimicking supramaximal CCK stimulation.

Table 5: Examples of the generated outputs by the model. The first two are good outputs where the mechanism is a simple paraphrase of the expected gold mechanism, while the next two illustrate the types of semantic errors we observe. The main entities are makred in *Italics*. The phrase explaining the mechanism in gold data is in blue, in good generation is in green, and in bad generation is in red.

The experts found that the output statements to be scientifically correct in many cases(58%). In summary, however, only 32% of the outputs were acceptable in all questions and were deemed to be good mechanism sentences. This again highlights the significant challenge posed by this task.

5.6 Error Analysis

To understand the frequent failure modes of the model, we manually categorized the errors in a hundred outputs that had the worst BLEURT scores with the reference mechanism sentences. We find the following main categories of errors:

Missing Entities (35%) – The most prevalent issue is the absence of one of the main entities in the generated sentence. Despite this being a necessary feature in all of the mechanism sentences in the training data, the prevalence of this error shows that models find it difficult to track the main entities during generation.

Incorrect Mechanism (24%) – The model is unable to generate the correct mechanism even though it is able to identify the correct relation and fills in some information that is either unrelated to or unsupported by the input sentences.

Flipped Relation (19%) – The model predicts the incorrect relation and generates a mechanism that is faithful to this incorrect relation. Improving relation generation is thus an important step for improving mechanism generation.

Non Mechanisms (11%) – While the model learns to generate mechanism like sentences for the most part, it sometimes still fails to produce sentences that contain any mechanism at all.

Multiple pieces of information (11%) – Some mechanisms are complex in that they require combining multiple bits of information from different input sentences and manages to only generate part of this complex mechanism.

Table 5 shows example generated mechanisms. The first example shows a generated mechanism that is almost the same as the gold mechanism with only a slight syntactic change. The second example shows a generated mechanism which also conveys the gold mechanism accurately with a paraphrasing that expands the technical term TRANSPRESSING. The third shows a bad output which contains a mechanism but not of the relation connecting the main entities. The fourth example presents a case where the information is correct but it does not even mention the main entities.

6 Conclusions

We introduced SuMe, a dataset for biomedical mechanism summarization. This dataset is coupled with a challenging summarization task, which requires the generation of mechanism participants as well as a textual summary of the mechanism, using as input multiple sentences from actual publication abstracts. We evaluated the complexity of the task using multiple neural language models. Our evaluation suggests that the proposed task is learnable, but we are far from solving it. We also introduce a pretraining task which is generally easier, and broadly scalable to improve the baselines.

All in all, we believe that the proposed dataset and associated task are an useful step towards building true information-access applications for the biomedical literature.

7 Ethical Considerations

The dataset is constructed from publicly available scientific literature. The domain experts were compensated for their time at the rate of $20/hour which is above the minimum hourly wage in the state of New York. The task and dataset are aimed at developing models that are able to better understand and reason about mechanisms underlying biomedical relations. Our results suggest current models are far from producing consistently reliable outputs and are not ready for practical use at this stage.

References

Cecilia N Arighi, Zhiyong Lu, Martin Krallinger, Kevin B Cohen, W John Wilbur, Alfonso Valencia, Lynette Hirschman, and Cathy H Wu. 2011. Overview of the biocreative iii workshop. *BMC bioinformatics*, 12(8):1–9.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: Pretrained language model for scientific text. In *EMNLP*.

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel S Weld. 2020. Tldr: Extreme summarization of scientific documents. *arXiv preprint arXiv:2004.15011*.

Erion Çano and Ondřej Bojar. 2020. Two huge title and keyword generation corpora of research articles. *arXiv preprint arXiv:2002.04689*.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. *CoRR*, abs/1804.05685.

Ed Collins, Isabelle Augenstein, and Sebastian Riedel. 2017. A supervised approach to extractive summarisation of scientific papers. *arXiv preprint arXiv:1706.03946*.

Dina Demner-Fushman, Kevin Bretonnel Cohen, Sophia Ananiadou, and Junichi Tsujii, editors. 2020. *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*. Association for Computational Linguistics, Online.

Jay DeYoung, Eric Lehman, Ben Nye, Iain J Marshall, and Byron C Wallace. 2020. Evidence inference 2.0: More data, better models. *arXiv preprint arXiv:2005.04177*.

Reza Ghaeini, Xiaoli Fern, Hamed Shahbazi, and Prasad Tadepalli. 2019. Saliency learning: Teaching the model where to pay attention. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4016–4025.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing.

Kamal raj Kanakarajan, Bhuvana Kundumani, and Malaikannan Sankarasubbu. 2021. BioELECTRA:pretrained biomedical text encoder using discriminators. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 143–154, Online. Association for Computational Linguistics.

Martin Krallinger, Martin Pérez-Pérez, Gael Pérez-Rodríguez, Aitor Blanco-Míguez, Florentino Fdez-Riverola, Salvador Capella-Gutierrez, Anália Lourenço, and Alfonso Valencia. 2017. The biocreative v. 5 evaluation workshop: tasks, organization, sessions and topics.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Guy Lev, Michal Shmueli-Scheuer, Jonathan Herzig, Achiya Jerbi, and David Konopnicki. 2019. Talksumm: A dataset and scalable annotation method for scientific paper summarization based on conference talks. *arXiv preprint arXiv:1906.01351*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Rui Meng, Khushboo Thaker, Lei Zhang, Yue Dong, Xingdi Yuan, Tong Wang, and Daqing He. 2021. Bringing structure into summaries: a faceted summarization dataset for long scientific documents. *arXiv preprint arXiv:2106.00130*.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint arXiv:1808.08745*.

Nikola I Nikolov, Michael Pfeiffer, and Richard HR Hahnloser. 2018. Data-driven summarization of scientific articles. *arXiv preprint arXiv:1804.08875*.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Yannis Papanikolaou and Andrea Pierleoni. 2020. Dare: Data augmented relation extraction with gpt-2. *arXiv preprint arXiv:2004.13845*.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets. *arXiv preprint arXiv:1906.05474*.

Long N Phan, James T Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. 2021. Scifive: a text-to-text transformer model for biomedical literature. *arXiv preprint arXiv:2106.03598*.

Amy Pu, Hyung Won Chung, Ankur P. Parikh, Sebastian Gehrmann, and Thibault Sellam. 2021. Learning compact metrics for mt. In *EMNLP*.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020a. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020b. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Hamed Shahbazi, Xiaoli Fern, Reza Ghaeini, and Prasad Tadepalli. 2020. Relation extraction with explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6488–6494.

Mokanarangan Thayaparan, Marco Valentino, and André Freitas. 2020. A survey on explainability in machine reading comprehension. *arXiv preprint arXiv:2010.00389*.

Marco A Valenzuela-Escárcega, Özgün Babur, Gus Hahn-Powell, Dane Bell, Thomas Hicks, Enrique Noriega-Atala, Xia Wang, Mihai Surdeanu, Emek Demir, and Clayton T Morrison. 2018. Large-scale automated machine reading discovers new cancer driving mechanisms. *Database: The Journal of Biological Databases and Curation*.

Qingyun Wang, Lifu Huang, Zhiying Jiang, Kevin Knight, Heng Ji, Mohit Bansal, and Yi Luan. 2019. Paperrobot: Incremental draft generation of scientific ideas. *arXiv preprint arXiv:1905.07870*.

Qingyun Wang, Zhihao Zhou, Lifu Huang, Spencer Whitehead, Boliang Zhang, Heng Ji, and Kevin Knight. 2018a. Paper abstract writing through editing mechanism. *arXiv preprint arXiv:1805.06064*.

Qingyun Wang, Zhihao Zhou, Lifu Huang, Spencer Whitehead, Boliang Zhang, Heng Ji, and Kevin Knight. 2018b. Paper abstract writing through editing mechanism. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 260–265. Association for Computational Linguistics.

Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander R Fabbri, Irene Li, Dan Friedman, and Dragomir R Radev. 2019. Scisummnet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7386–7393.