

Readability Controllable Biomedical Document Summarization

Zheheng Luo and Qianqian Xie* and Sophia Ananiadou

NaCTeM, The University of Manchester

{zheheng.luo, qianqian.xie, sophia.ananiadou}@manchester.ac.uk

Abstract

Different from general documents, it is recognised that the ease with which people can understand a biomedical text is eminently varied, owing to the highly technical nature of biomedical documents and the variance of readers' domain knowledge. However, existing biomedical document summarization systems have paid little attention to readability control, leaving users with summaries that are incompatible with their levels of expertise. In recognition of this urgent demand, we introduce a new task of readability controllable summarization for biomedical documents, which aims to recognise users' readability demands and generate summaries that better suit their needs: technical summaries for experts and plain language summaries (PLS) for laypeople. To establish this task, we construct a corpus consisting of biomedical papers with technical summaries and PLSs written by the authors, and benchmark multiple advanced controllable abstractive and extractive summarization models based on pre-trained language models (PLMs) with prevalent controlling and generation techniques. Moreover, we propose a novel masked language model (MLM) based metric and its variant to effectively evaluate the readability discrepancy between lay and technical summaries. Experimental results from automated and human evaluations show that though current control techniques allow for a certain degree of readability adjustment during generation, the performance of existing controllable summarization methods is far from desirable in this task.

1 Introduction

Automatic summarization for biomedical documents (Guo et al., 2020; DeYoung et al., 2021) such as clinical literature (Wang et al., 2020b; DeYoung et al., 2021), provides an efficient way for readers to acquire desirable biomedical information quickly. Unlike general documents, biomedical

*Corresponding author

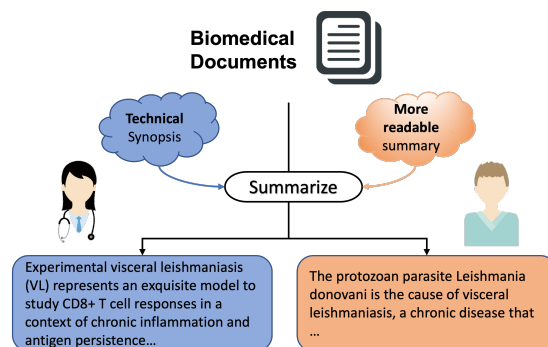


Figure 1: Example of our task. Summaries are generated according to users' demand for readability.

documents have characteristics of mounting scientific jargon (Plavén-Sigraý et al., 2017), and complex language structures (Friedman et al., 2002). Therefore, readers such as non-experts and professionals would seek textual information on different readability levels, since the variance of their biomedical knowledge affects their ease of understanding biomedical papers. For example, an in-domain expert might require accurate and clear technical summaries with medical jargon and professional language, to quickly grasp the main contributions of biomedical papers. In contrast, layperson readers usually require plain language summaries with less technical terms and more context of the research, which are easier to understand. Nevertheless, current biomedical summarization systems are only able to offer technical abstracts (Sotudeh et al., 2020; DeYoung et al., 2021; Xie et al., 2022b,a; Bishop et al., 2022) or lay language summaries (Guo et al., 2020; Chandrasekaran et al., 2020), fail to generate compatible summaries for various users according to their levels of expertise without considering the readability as an aspect to be controlled during summary generation (He et al., 2020). We argue that it is urgent to develop biomedical summarization approaches that can not only condense biomedical documents into

concise summaries but also adjust the readability level of summaries to improve the dissemination of scientific information.

Our research aims to tackle the problem, and thus propose a novel task of readability controllable biomedical document summarization, which is to automatically recognize users' readability demands and generate summaries that are compatible with their expertise level and needs, as shown in Figure 1. Specifically, in a binary readability level controlling setting, it is to produce technical summaries for experts, while plain language summaries (PLS) for laypeople. The task is challenging since: 1) it requires the model to accurately recognize different readability demands from limited guiding signals, 2) it requires a suitable selection of content from long biomedical documents for various readers guided by their readability demands, 3) it requires the model to learn not only lexical and syntactic adjustment but also paraphrasing according to users' needs. Since professionals pay more attention to clarity and accuracy while non-experts prefer summaries that are easier to understand.

To approach this task, we build the first corpus consisting of 28,124 biomedical literature with technical and plain language summaries written by the authors, then conduct a thorough analysis of the collected data including statistics, readability metrics, and textual features. Next, we examine several controlling techniques on prevalent pre-trained language models (PLMs) and evaluate their performance on our dataset. Apart from automatic assessment, we also bring in the human evaluation due to the inefficacy of current metrics for readability and text generation. To better characterize readability differences between technical summary and PLS, we further propose a novel masked noun phrase-based text complexity metric and its variant based on the masked language model (MLM). It is superior in modelling the semantic structure of biomedical texts compared to traditional metrics and existing MLM-based metrics.

Overall, our main contributions are summarised as follows: (1) We introduce a novel task of readability controllable biomedical document summarization. (2) We build a corpus¹ with 28,124 biomedical papers with their technical and plain language summaries, which will facilitate further exploration in this task. (3) We propose an MLM-

based text complexity metric, which surpasses existing readability evaluation metrics on our dataset. (4) We examined controlling techniques including prompts and multi-heads on both extractive and abstractive methods to adjust readability during summarization and found the performance is far from satisfying. To the best of our knowledge, this is the first effort to consider readability as a controllable attribute in scientific document summarization.

2 Related Work

2.1 Biomedical Text Summarization

Neural networks and PLMs have been explored for biomedical document summarization in recent years, due to their success in general text summarization (Cohan et al., 2018; Liu and Lapata, 2019a; Zhang et al., 2019; Wang et al., 2021). Sotudeh et al. (2020) improved radiology report summarization by incorporating medical ontology into a sequence-to-sequence summarizer. Wallace et al. (2020) investigated the BART model (Lewis et al., 2020) with domain specific pre-training strategies and input decorations for multi-document summarization of randomized controlled trials (RCTs). Progress in biomedical summarization has also been advanced due to the emergence of in-domain corpora. Cohan et al. (2018) and Wang et al. (2020b) compiled a large amount of biomedical literature with their abstracts as summaries. DeYoung et al. (2021) investigated if systematic reviews could be summarised from their cited clinical trials. Guo et al. (2020) mixed summarization and simplification by generating plain language summary conditioned on abstracts of systematic reviews.

2.2 Controllable Text Summarization

Recent efforts on controllable text summarization mostly focus on news articles. Fan et al. (2018) has leveraged PLMs with special tokens prepended to the input, to control the length, entities, and style of the generated summary. Zheng et al. (2020) and He et al. (2020) further extended prompts, keywords and entities as guiding markers. Chan et al. (2021) proposed the constrained markov decision process (CMDP) based method to control attributes of summarization. Other works have tried exerting control in decoding. HydraSum (Goyal et al., 2021) distributed different values of an attribute into multiple decoders and leveraged a gate mechanism to gain control over properties such as abstractness and length. Amplayo et al. (2021) and Amplayo

¹can be downloaded from <http://www.nactem.ac.uk/readability/>

and Lapata (2021) focused on the aspect control of opinion summarization on reviews. To the best of our knowledge, our work is the first effort to consider readability as a controllable attribute in scientific document summarization, which is important for specific-domain, especially biomedical science.

2.3 Readability Metrics

Readability is defined as the ease with which a reader can understand a piece of text. Many factors are involved in determining readability, such as lexical and syntactic sophistication, discourse cohesion, and background knowledge (Crossley et al., 2017). Prior work on lay summarization (Guo et al., 2020) evaluated their corpus by traditional readability formulas like Flesch-Kincaid Grade Level (Kincaid et al., 1975) which is inefficient in revealing the readability differences in scientific writings. Martinc et al. (2021b) has shown the potential of the PLM in estimating text readability. Devaraj et al. (2021) used an MLM-based metric to better classify technical abstracts and PLS of medical reviews. In this work, we propose an advanced MLM-based metric to manifest the readability differences among summaries in our corpus and evaluate the output of tested models.

3 Task Overview

Definition. The objective of this task is to generate summaries of biomedical documents on different readability levels based on users' demands. Let $D = \{d_1, d_2, \dots, d_k\}$ denotes the set of source documents, each document $d_i = \{x_{i,1}, x_{i,2}, \dots, x_{i,n}\}$ can be represented by the sequence of n tokens, S_i stands for the target summary of document d_i that is represented by the sequence of m tokens: $\{s_{i,1}, s_{i,2}, \dots, s_{i,m}\}$, $m \ll n$. r means the readability level the user might want. The task can be formulated as a conditional generative problem as follows:

$$P(S|D, r) = \prod_i^k P(S_i|d_i, r) \quad (1)$$

which maximizes the probability of generating S when given the document set D and the readability demand r . In this work, since the exploration of readability controlling summarization is still in an initial stage, we start with single document input with a binary readability control between "technical" and "plain language" and leave more fine-grained control to future work. We consider r_t

Dataset	docs	avg. doc length	avg. abs length	avg. PLS length
PubMed	133,000	3,016	203	-
CDSR	7,805	-	714	374
Ours	28,124	6,697	287	204

Table 1: Statistics of our PLOS datasets compared with existing biomedical summarization corpora PubMed (Cohan et al., 2018) and CDSR (Guo et al., 2020)

mean the demand for *technical summary* that is suitable for experts, while r_p means the demand for *plain language summary* (PLS) for laypeople readers. Thus, we have both technical target summary S_i^t and plain language target summary S_i^p for each input document d_i , to train the model. Additionally, a technical summary and a PLS generated from the same document by the same model will be referred to as a pair of summaries in this paper.

Evaluation. The most commonly used metric for evaluating summarization models is ROUGE (Lin, 2004), which has served as a standard in various text generation tasks. However, a recent study (Bhandari et al., 2020) has shown that ROUGE scores do not always agree with human evaluation when assessing generated summaries. Also, traditional readability metrics are found unable to show the significant readability difference between the technical summary and their simplified counterparts (Devaraj et al., 2021). Thus, we conducted both automatic and human evaluations to assess the readability and general qualities of generated summaries.

4 Dataset Description

4.1 Data Compilation

We constructed the corpus consisting of peer-reviewed biomedical research papers with the technical summaries and PLSs from journals including PLOS² Biology, PLOS Computational Biology, PLOS Genetics, PLOS Medicine, PLOS Neglected Tropical Diseases, and PLOS Pathogens, cover a broad range of biomedical research subjects. The PLSs are placed under the section *Author Summary* in the format of the PLOS articles and written by the authors following the requirement of PLOS submission guidelines which suggest highlighting where the work fits within a broader context, presenting the significance simply and avoiding the

²<https://journals.plos.org/plosone/>

use of acronyms and complex terminology³.

To build the corpus, we downloaded the whole PLOS article dataset (up to 4th April 2022)⁴. Then, after filtering out papers without plain language summaries, we extracted the full text, the abstract as technical summary, and the *Author Summary* as PLS from the remaining papers, resulting in 28,124 document-technical summary-PLS triplets. We randomly sample 1,000 triplets respectively for validation and test, leaving the rest 26,124 for training. Table 1 shows the main statistics of our dataset and other biomedical summarization corpora. Compared to previous work, our dataset is the first that contains both technical summaries and PLS, and our source documents are much longer, making the task more challenging.

4.2 Quantitative Analysis

To investigate the readability differences between technical summaries and PLSs, we examined various metrics and calculated Spearman's correlations with the readability levels of summaries in our dataset.

Traditional readability formula. Three established heuristics-based readability metrics are adopted, including the Flesch-Kincaid Grade (Kincaid et al., 1975), Coleman-Liau Index (Coleman and Liau, 1975), and automated readability index (ARI) (Senter and Smith, 1967), which are used to approximate the U.S. grade level to understand a written text. These metrics rely on shallow features like the length of a sentence or the number of characters in words and thus are unable to fully reveal the gap between the summaries of biomedical documents on different readability levels (Gao et al., 2019).

Language model based metric. Predicted probabilities for masked tokens by language models have shown effectivity in measuring readability. Devaraj et al. (2021) assumed that language models trained on general domain text would align lower likelihoods to tokens in technical jargon and higher likelihoods to common tokens. They proposed the BERT (Devlin et al., 2019) and SciBERT (Beltagy et al., 2019) based metric, which randomly masks 15% of tokens in each sentence, and computes the average likelihoods of original masked tokens in the distributions from the model output. This method outperforms traditional metrics such

as Flesch-Kinard Grade in distinguishing plain language summaries against technical abstracts of systematic reviews. We will refer to their metric as masked random token-based text complexity (MRTTC).

Masked noun phrase-based metric. It is recognized there are many technical terms existing in biomedical texts in the form of noun phrase (NP), which should be considered as a complete semantic unit. Thus token-level masking has its limitation since random token-level masking may corrupt the semantic integrity of technical terms in biomedical texts. Rather than randomly masked tokens, the likelihood of whole noun phrases predicted by a language model could be a finer indicator to discriminate between plain language and technical texts. Therefore, we propose the novel masked noun phrase-based text complexity metric, denoted as NPTC. We first leverage Spacy (Honnibal and Montani, 2017) to extract all NPs in each document, then filter out NPs that only consist of stop-words to prevent disturbance of common tokens. Next, we mask tokens of each NP in the summaries of each document, to create a masked summary and use a BERT pre-trained on general text⁵ to predict the probability sequence of masked NPs. Lastly, the likelihoods of the target tokens in each masked NP are averaged as the likelihood of the NP, which is further averaged across the document to achieve its final score.

Moreover, we follow Martinc et al. (2021a) who argue that compared to directly sum word negative log-likelihood (WNLL), assigning weights to WNLL depending on the ranking of their values in a text is a more effective way to model readability given that several unreadable words might drastically increase the difficulty of the entire text. Based on these considerations, we propose the ranked NP-based text complexity (RNPTC). Specifically, after obtaining the probability of each NP, we sort them in descending order, and then use the following formula to average over the probabilities:

$$RNPTC = -\log\left(\frac{\sum_{i=1}^{|NPs|} P_{NP(i)}/\sqrt{i}}{|NPs|}\right) \quad (2)$$

where $|NPs|$ denotes the total amount of NPs, i stands for the rank of the current NP, and $P_{NP(i)}$ is the probability of the NP that is ranked i . By assigning the reciprocal of square rooted ranking as the weight to every NP, the complexities of the most

³<https://journals.plos.org/ploscompbiol/s/submit-guidelines>

⁴<https://plos.org/text-and-data-mining/>

⁵<https://huggingface.co/bert-base-uncased>

Metric	Avg. on TS	Avg. on PLS	Spearman's ρ
RNPTC	2.63	2.45	0.337
NPTC	1.84	1.72	0.183
MRTTC	0.81	0.78	0.148
C-L Index	16.50	15.93	0.140
F-K Grade	14.99	14.76	0.065
ARI	17.90	17.41	0.102

Table 2: Average values of different metrics in our dataset and their Spearman’s correlation with readability. TS means technical summaries.

Metric	Avg. on Abstract	Avg. on PLS	Spearman's ρ
RNPTC	2.78	2.46	0.584
NPTC	1.81	1.50	0.500
MRTTC	0.93	0.81	0.513
C-L Index	14.71	14.22	0.138
F-K Grade	11.80	12.67	-0.190
ARI	14.24	15.43	-0.217

Table 3: Average values of different metrics in the abstracts and PLS of CSDR and their Spearman’s correlation with readability.

difficult NPs are emphasised and the interference of common phrases is reduced. The detailed process of RNPTC is illustrated in Algorithm 1 at Appendix B. For a clearer comparison, we take the negative logarithm of Devaraj et al. (2021)’s metric in the experiment.

To illustrate the effectivity of our proposed metrics, we compare Spearman’s correlations of examined metrics with the readability levels on our dataset and the CDSR dataset(Cochrane Database of Systematic Reviews) of which the detailed collecting process is illustrated in Appendix C. The results are shown in Table 2 and 3. Compared to PLM-based metrics including RNPTC, NPTC and MRTTC, traditional formulas (C-L Index, F-K Grade, ARI) have lower Spearman’s correlation scores. This indicates that they are not good indicators of readability differences between technical texts and plain language ones due to their dependence on shallow statistical features. Our proposed metrics surpasses all other methods in correlation with the readability level in both datasets. This shows that the consideration for fine-grained semantic structure of biomedical texts in our metrics RNPTC and NPTC is helpful to discriminate readability differences between technical summaries and PLS.

4.3 Qualitative Analysis

To better demonstrate how the authors adjust the readability when writing summaries for different audiences, we manually read through some random samples from our dataset, and suggest the distinctions of the writing styles are rooted in their choice of content and language. Under these two major subjects, we further defined four specific aspects and present them in Table 4.

Content. When choosing content for writing summaries on different readability levels for their work, the authors usually keep the general compositions similar. But they are likely to include detailed experimental design and quantitative data which are helpful in showing the confidence of the experiments when drafting a technical summary while omitting these descriptions in PLS. Moreover, we have noticed that authors usually add a few sentences at the beginning of PLS to explain the subject of their research and the context around it, smoothing laypeople readers’ understanding of the research question.

Language. From the perspective of language use, there are two main distinctions. First, scientific jargon has been heavily used in technical summaries for accuracy and conciseness, but either removed or replaced with more common expressions in PLS. Second, the authors are also observed to use simpler syntactic structure in PLS to enhance readability.

5 Baselines for Readability Controllable Summarization

Existing text summarization methods can be classified into two main ways, extractive and abstractive approaches. We carry out experiments with both methods in this task.

5.1 Skeleton Model

Prevalent PLMs such as BERT (Devlin et al., 2019) and BART (Lewis et al., 2019) using full attention mechanisms are faced with the out-of-memory problem when training with long sequences. Nevertheless, the average amount of words in a biomedical scientific document is usually several thousand, and shortening the input into hundreds of tokens would significantly reduce information that could be used by summarization models. Thus we take on Longformer-Encoder-Decoder (LED) (Beltagy et al., 2020) as the underlying architecture. LED is implemented with local attention based on a sliding

	Distinction	Technical Summary	Plain Language Summary
Content	Experiment Details	1300 individuals with minor TT were recruited and randomly assigned to ... At two-years all epilation arm participants were offered free surgery. At four-years 1151 (88.5%) were re-examined: 572 (88%) and 579 (89%) from epilation and surgery arms, respectively. At that time, 21.1% of the surgery arm participants had recurrent TT; 189/572 (33%) of the epilation arm had received surgery.	We conducted a randomized controlled trial of epilation versus eyelid surgery (the main treatment option) in 1300 people with mild trichiasis in Ethiopia.
	Background Information	West Nile virus (WNV) is transmitted to vertebrate hosts by mosquitoes as they take a blood meal.	Since it was first introduced into the United States in 1999, West Nile virus (WNV) has caused significant disease in humans, horses, and other animals. WNV is transmitted to humans and other ...
Language	Scientific Jargon	Experimental visceral leishmaniasis (VL) represents an exquisite model to study CD8+ T cell responses in a context of chronic inflammation and antigen persistence , since it is characterized by chronic infection in the spleen and CD8+ T cells are required for the development of protective immunity .	Preliminary studies suggested that CD8+ T cells can kill parasites and limit disease ; however, studying these important killer cells has been hindered , because we do not know what parasite molecules they recognize
	Syntactic structure	Hookworm-related cutaneous larva migrans (CLM) is a common but neglected tropical skin disease caused by the migration of animal hookworm larvae in the epidermis.	Hookworm-related cutaneous larva migrans (CLM) is a parasitic skin disease common in developing countries with hot climates. The disease is caused by animal hookworm larvae...

Table 4: Difference between the writing styles of technical summary and PLS. Red texts are contents which are included in technical summaries and omitted in PLS. Blue texts indicate complimentary information provided for PLS but not for technical summaries. Green texts represent more layperson-friendly expression while purple means scientific jargon. Teal texts stand for simplified syntactic structures of texts in pink.

window for all tokens in the sequence and global attention for certain tokens to learn task-specific representation. By the double attention design, LED reduces the computing requirements for long sequences and has shown better performance in summarizing long documents like Pubmed and Arxiv papers.

5.2 Control Methods

Control Prompts. Prior works (He et al., 2020) aforementioned in Section 2.2 have proved that combined with PLMs, prompts like special tokens are effective to guide the model to generate controlled text while training on corresponding data with certain attributes. Thus we design two types of markers $|PLAIN|$ and $|PRO|$ to respectively represent signals for producing PLS and technical summaries. During training, the special tokens are prepended to the source text to form the input source documents for their corresponding summaries. Also under the double attention setting in LED, to fully exert the guiding effect of these special tokens, we set them to conduct global attention with other tokens. In abstractive mode, we train the LED in a sequence-to-sequence style and used beam search during inference. For the extractive version, we followed the design in Liu and Lapata (2019b) by taking the encoder of LED and putting two transformer layers with a final classifier upon it. Then an end-of-sentence token is inserted after every sentence in the source documents and their hidden states output from the LED encoder are

used as representations of corresponding sentences into the following extractive layers. The model is optimized to pick sentences that appear in the summaries extracted by a greedy selection method (Liu and Lapata, 2019b) which will be referred to as Oracle extraction hereinafter. During inference, we select top- k sentences by their scores as the final summary.

Multi Heads. Goyal et al. (2021) has shown that the attributes of generated summaries like length can be distributed into multiple decoders in a sequence-to-sequence structure, then by decoding from different heads the user can obtain either long or short summaries. Inspired by their multi heads design, we tried adding one more decoder into the LED model for generating summaries of different readabilities. Specifically, in our abstractive framework, the encoder is kept shared but two decoders are set independent of learning either technical or plain language style generation. We trained this multi-decoder model with the gate mechanism in which the probabilities predicted by the heads of two decoders would be multiplied by $1 - g$ and g respectively. We adjust g equals 1 for technical summaries generation and 0 for PLS. During inference, we sample summaries from the two heads separately. In the extractive multi-heads, we keep the encoder shared and create two extractive heads to select sentences for different readabilities. The model is trained via the same gate mechanism as the abstractive one. In both models, we set the start

Method		R1	R2	RL
Oracle	tech	62.79	31.36	58.26
	plain	57.08	25.31	52.39
CP Abs	tech	48.51	16.89	44.53
	plain	45.69	15.40	41.89
CP Ext	tech	49.52	16.63	45.50
	plain	45.30	13.00	40.99
MH Abs	tech	49.87	17.85	46.02
	plain	47.31	14.73	43.08
MH Ext	tech	49.17	16.75	39.59
	plain	45.40	13.03	40.99

Table 5: ROUGE scores of generated summaries, CP stands for Control Prompts and MH is the abbreviation of Multi Heads

of sentence token $\langle s \rangle$ in LED for global attention. Due to length limitations, please see Appendix A for detailed experimental setups.

6 Results

6.1 Automatic Evaluation

In Table 5, we list the ROUGE scores of generated summaries from five tested methods. The upper bound of ROUGE score is established by the Oracle extraction in the first 2 rows, which displays a distinct lexical difference between technical summaries and plain language summaries since the lower ROUGE scores indicate that PLS are harder to approach by directly selecting sentences from the document. For the same reason, though competitive in summarizing technical summaries, extractive methods are largely outperformed by abstractive methods in generating PLS, manifesting the advantage of abstractive methods in adjusting the style of expression. Among all methods, multi heads abstractive LED (MH Abs) performs best on ROUGE of both kinds of readability. This can be attributed to the additional parameters provided by the multi heads technique for a model to adapt to generation for different readability demands. However, there is no evident difference between the two control techniques on extractive methods, indicating that under our training setting, more parameters are not helpful for controlling in an extractive way.

In Table 6, we further show the difference in readability scores between pairs of summaries evaluated by RNPTC and three traditional functions. The values of target summaries in the test set (PLOS Test) are added in the last row for comparison. For technical summaries, the generated summaries should have higher readability scores which indicate more complicated tokens such as terms that could embed key information, while out-

Method		FKG	CLI	ARI	RNPTC
Oracle	tech	14.37	14.79	17.02	2.57
	plain	14.46	15.01	17.00	2.50
CP Abs	tech	14.85	15.59	17.40	2.46
	plain	15.13	15.83	17.73	2.40
CP Ext	tech	14.91	15.60	17.43	2.60
	plain	15.04	15.80	17.58	2.53
MH Abs	tech	14.25	15.07	16.63	2.41
	plain	14.75	15.12	17.18	2.35
MH Ext	tech	15.02	15.65	17.57	2.60
	plain	15.13	15.91	17.73	2.51
PLOS Test	tech	14.90	16.45	17.75	2.57
	plain	14.80	15.98	17.43	2.41

Table 6: Readability scores of generated summaries

put for PLSs is expected to get lower scores by using less jargon but more common words. From traditional readability scores, most generated PLSs are of no difference or even harder to read than their technical counterparts. As mentioned before in section 4.2, these traditional readability formulas can largely be influenced by shallow statistic features such as the number of tokens and sentences. Thus, they may be deficient to reflect the readability differences.

Unlike these metrics, our proposed RNPTC demonstrates that the text complexity is generally lower in PLSs. However, the variance is slight compared to that between pairs of target summaries, which indicates these controlling models are not ideal to adjust their output under different readability demands. Moreover, from RNPTC, PLSs and technical summaries generated via extractive methods are generally more complicated than those from abstractive methods and target summaries, suggesting the high lexical complexity of sentences from original documents. With regards to generated summaries of abstractive methods, their RNPTCs are lower than those of target summaries especially for technical ones, implying they might fail to provide enough key information for expert readers. This could be caused by the lower contextual tendency of PLM-based text generation (Gehrmann et al., 2019). Thus when generating summaries, these PLM-based models tend to select tokens that will maximize the global probability while avoiding technical terms since their predicting probabilities can be lower than common tokens.

6.2 Textual Variance

Observing the slight difference in numerical readability metrics between pairs of generated summaries, we wonder how much textual variance can controlling techniques actually lead to when sum-

marizing for different readability demands. Hence, we take the ratio of the number of n-grams in a PLS that appear in the corresponding technical summary to the total number of n-grams in the PLS as the indicator for similarity. It is assumed that fewer overlapping suggests stronger control of a summarization model. In Table 7, we compared the n-gram overlapping fractions among pairs of target summaries in our whole PLOS dataset (PLOS Whole) and the test subset (PLOS Test) as well as pairs of generated summaries from the five examined methods.

Between pairs of target summaries, the similitude is evidently smaller compared to all generated results, showing the limited ability of the controlling techniques to adjust output according to readability demands. Two extractive methods suffer the most from the overlapping problem, approximately 80 per cent of 4-grams in their PLSs are also in the technical summaries, which could be due to the already large n-grams overlapping fraction in the Oracle extraction. Meanwhile, training with more discernable pairs of target summaries, abstractive methods outperform extractive ones significantly in generating more distinctive contents when given different readability demands. However, the overlapping ratio is still high, matching the little difference in readability evaluation.

Ngrams	N=1	N=2	N=3	N=4
PLOS Whole	0.67	0.25	0.12	0.08
PLOS Test	0.67	0.26	0.13	0.08
Oracle	0.73	0.42	0.33	0.30
CP Abs	0.80	0.59	0.51	0.47
CP Ext	0.91	0.84	0.82	0.80
MH Ext	0.87	0.76	0.73	0.72
MH Abs	0.76	0.47	0.35	0.29

Table 7: N-gram overlap fraction between PLSs and technical summaries

6.3 Human Evaluation

We further conduct a human evaluation to assess the two main aspects to be tested in our task: controlling performance and general quality. More details about the setup can be found in Appendix D. Examples of evaluated target and generated summaries can be found in Appendix E, Table 9.

When assessing controlling performance, we split evaluators into *experts* and *laypeople* according to their education background and ask them to rate the readability on a scale of 1-5 to see how biomedical expertise affects people's judgement on readability. In respect of general quality, we inform

Aspect		Target		Generated	
		Tech	Plain	Tech	Plain
Readability	<i>experts</i>	4.90	4.83	4.50	4.45
	<i>laypeople</i>	3.37	4.43	3.83	3.90
Relevance		-	-	3.39	3.54
Coherence		4.62	4.72	4.01	4.05
Grammar		4.83	4.90	4.39	4.35

Table 8: Human evaluation of target and generated summaries

the evaluators to rate each generated summary from 1 to 5 in three aspects: 1) relevance: to what extent does the summary cover the important information⁶. 2) grammar: how good the summary is grammatically. 3) coherence: is the content well-structured in the summary. The results are shown in Table 8.

In the judgment of the *experts*, the readability differences between pairs of target summaries are small while *laypeople* evaluators evidently discern the gap between target technical summaries and PLS. In both groups, the scores of pairs of generated summaries are quite close. And for *laypeople* group the readability difference between generated TS and PLS is modest compared to that between target technical summaries and PLSs, matching the slight effect of control shown in previous analyses. From the gap between the two groups' readability scores of technical summaries, we can see the influence of domain expertise on the ease of understanding biomedical texts. In respect of general quality, we can find that grammatically generated texts are only slightly worse than human writings. However, when it comes to relevance, generated summaries achieve only mediocre performance, suggesting the difficulty of capturing key information in long biomedical documents. Also, the lower coherence of generated summaries reveals the inability of PLMs to keep the output sentences well-connected when generating paragraph-level texts.

7 Conclusion

In this work, we introduce a novel controlling summarization task that aims at encapsulating biomedical scientific literature on different readability levels. We draw the full text, abstracts, and plain language summaries from papers published in peer-reviewed biomedical journals to build the first corpus for the task. We propose an effective MLM-

⁶we here ask annotators to compare generated summaries to target summaries so the relevance of target summaries is omitted

based readability metric and its variant, which outperform traditional and previous MLM-based readability measures in distinguishing technical summaries against PLSs. We leverage the Longformer-Encoder-Decoder (LED) as our skeleton model and examine prevalent generation controlling techniques on both extractive and abstractive methods. We found that though the abstractive approach combined with multi decoders can lead to higher-quality summaries and larger textual differences under binary readability demands compared to extractive methods, all tested models fail to show satisfying readability controlling ability. In the future, we would like to investigate how to introduce a more powerful teaching force such as language models trained with corpora of different readability levels to guide the controlling of summarization models.

Limitations

In this paper, we examined exerting control on the readability of the output of PLM-based summarization models.

Firstly, there are only texts of binary readability levels in the experimented corpus, limiting more fine-grained readability control under our supervised training methods.

Secondly, though distinctions of content and language are observed in pairs of generated summaries, the degree of control is still far from satisfying. We assume that introducing pre-trained discriminators (Dathathri et al., 2019) into the summarization progress might help models push the output further towards a more technical or plain language direction.

Last but not least, due to the length and complexity of biomedical documents, we solely evaluate the relevance of generated summaries while leaving their faithfulness to source documents unchecked. Faithfulness or factuality of scientific summaries are of critical significance but receive little attention owing to the difficulty, we encourage future work to combine question generation and reading comprehension (Wang et al., 2020a) into the assessment of faithfulness in scientific document summarization.

Ethics Statement

This paper presents a corpus built upon part of the whole article dataset from the PLOS corporation which is freely open to the public. The advancement of large pre-trained language models

has greatly boosted the improvement of summarization models in various domains including the biomedical area. High quality, especially factual summaries would facilitate practice and research in both clinical and scientific communities. Yet, current state-of-the-art PLM-based models are unable to guarantee the faithfulness or factuality of generated summaries (Maynez et al., 2020). Thus we suggest any output of our proposed model should be manually examined by domain experts before using for any purpose.

Acknowledgement

We would like to thank the evaluators (Jimin Huang, Feng Zhou, Shenghan Xie, Tairan Yang, Wang Yang, Yuandong Tao) for their valuable time and the Computational Shared Facility at The University of Manchester for offering computing devices. We also thank Jimin Huang and Anonymous Reviewers for their suggestions and feedback on our paper.

References

- Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2021. [Aspect-controllable opinion summarization](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6578–6593. Association for Computational Linguistics.
- Reinald Kim Amplayo and Mirella Lapata. 2021. [Informative and controllable opinion summarization](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 2662–2672. Association for Computational Linguistics.
- Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. [Scibert: A pretrained language model for scientific text](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3613–3618. Association for Computational Linguistics.
- Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.
- Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9347–9359.

- Jennifer Bishop, Qianqian Xie, and Sophia Ananiadou. 2022. Gencomparesum: a hybrid unsupervised summarization method using salience. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 220–240.
- Hou Pong Chan, Lu Wang, and Irwin King. 2021. Controllable summarization with constrained markov decision process. *Transactions of the Association for Computational Linguistics*, 9:1213–1232.
- Muthu Kumar Chandrasekaran, Guy Feigenblat, Eduard Hovy, Abhilasha Ravichander, Michal Shmueli-Scheuer, and Anita de Waard. 2020. [Overview and insights from the shared tasks at scholarly document processing 2020: CL-SciSumm, LaySumm and LongSumm](#). In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 214–224, Online. Association for Computational Linguistics.
- 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. *arXiv preprint arXiv:1804.05685*.
- Meri Coleman and Ta Lin Liao. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.
- Scott A Crossley, Stephen Skalicky, Mihai Dascalu, Danielle S McNamara, and Kristopher Kyle. 2017. Predicting text comprehension, processing, and familiarity in adult readers: New approaches to readability formulas. *Discourse Processes*, 54(5-6):340–359.
- Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*.
- Ashwin Devaraj, Iain J Marshall, Byron C Wallace, and Junyi Jessy Li. 2021. Paragraph-level simplification of medical texts. *arXiv preprint arXiv:2104.05767*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.
- Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu Wang. 2021. Ms2: Multi-document summarization of medical studies. *arXiv preprint arXiv:2104.06486*.
- Angela Fan, David Grangier, and Michael Auli. 2018. Controllable abstractive summarization. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 45–54.
- Carol Friedman, Pauline Kra, and Andrey Rzhetsky. 2002. Two biomedical sublanguages: a description based on the theories of zellig harris. *Journal of biomedical informatics*, 35(4):222–235.
- Yanjun Gao, Chen Sun, and Rebecca J. Passonneau. 2019. [Automated pyramid summarization evaluation](#). In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 404–418, Hong Kong, China. Association for Computational Linguistics.
- Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. 2019. Gltr: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116.
- Tanya Goyal, Nazneen Fatema Rajani, Wenhao Liu, and Wojciech Kryściński. 2021. Hydrasum: Disentangling stylistic features in text summarization using multi-decoder models. *arXiv preprint arXiv:2110.04400*.
- Yue Guo, Wei Qiu, Yizhong Wang, and Trevor Cohen. 2020. Automated lay language summarization of biomedical scientific reviews. *arXiv preprint arXiv:2012.12573*.
- Junxian He, Wojciech Kryściński, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2020. Ctrlsum: Towards generic controllable text summarization. *arXiv preprint arXiv:2012.04281*.
- Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.
- J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.
- 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*.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. [BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

- 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. 2019a. [Text summarization with pretrained encoders](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3728–3738. Association for Computational Linguistics.
- Yang Liu and Mirella Lapata. 2019b. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740.
- Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021a. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1):141–179.
- Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021b. [Supervised and unsupervised neural approaches to text readability](#). *Computational Linguistics*, 47:141–179.
- Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.
- Pontus Plavén-Sigray, Granville James Matheson, Björn Christian Schiffler, and William Hedley Thompson. 2017. [Research: The readability of scientific texts is decreasing over time](#). *eLife*, 6:e27725.
- RJ Senter and Edgar A Smith. 1967. Automated readability index. Technical report, Cincinnati Univ OH.
- Sajad Sotudeh, Nazli Goharian, and Ross W Filice. 2020. Attend to medical ontologies: Content selection for clinical abstractive summarization. *arXiv preprint arXiv:2005.00163*.
- Byron C. Wallace, Sayantan Saha, Frank Soboczenski, and Iain James Marshall. 2020. [Generating \(factual?\) narrative summaries of rcts: Experiments with neural multi-document summarization](#). *CoRR*, abs/2008.11293.
- Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020a. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020.
- Benyou Wang, Qianqian Xie, Jiahuan Pei, Prayag Tiwari, Zhao Li, et al. 2021. Pre-trained language models in biomedical domain: A systematic survey. *arXiv preprint arXiv:2110.05006*.
- Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, et al. 2020b. Cord-19: The covid-19 open research dataset. *ArXiv*.
- Qianqian Xie, Jennifer Amy Bishop, Prayag Tiwari, and Sophia Ananiadou. 2022a. Pre-trained language models with domain knowledge for biomedical extractive summarization. *Knowledge-Based Systems*, 252:109460.
- Qianqian Xie, Jimin Huang, Tulika Saha, and Sophia Ananiadou. 2022b. Gretel: Graph contrastive topic enhanced language model for long document extractive summarization. *arXiv preprint arXiv:2208.09982*.
- Haoyu Zhang, Yeyun Gong, Yu Yan, Nan Duan, Jianjun Xu, Ji Wang, Ming Gong, and Ming Zhou. 2019. [Pretraining-based natural language generation for text summarization](#). *CoRR*, abs/1902.09243.
- Changmeng Zheng, Yi Cai, Guanjie Zhang, and Qing Li. 2020. [Controllable abstractive sentence summarization with guiding entities](#). In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 5668–5678. International Committee on Computational Linguistics.

A Experiment Setup

All our experiments were run at a single NVIDIA Tesla A-100 GPU. We set input length as 8,192 for covering the whole text of our source document in LED and chose a learning rate of $3e - 5$ with warm-up for 3 epochs to finetune all the models from the checkpoint in hugging face⁷. With regard to the optimizer, we used AdamW. For generation, we use a beam size of 4, with no repeat n-gram equal to 3. In extractive methods, we select topk sentences when the candidate summary reaches the average token number of technical summaries and PLS. All our models are built with PyTorch and HuggingFace.

B RNPTC Algorithm

Algorithm 1 To compute a text complexity score given a document d and a PLM lm . The FORWARD function returns the output matrix where each row maps to a distribution over all the tokens in the vocabulary. The APPEND function adds a value in the list. The RANKMEAN function is calculated as in Equation 2.

```
1: procedure RNPTC( $d, lm$ )
2:   NPs  $\leftarrow$  Noun phrases list extracted from  $d$ 
3:   P  $\leftarrow$  Create empty NP probability list
4:   for  $i = \{1, \dots, |NPs|\}$  do
5:      $T \leftarrow$  Token sequence of NPs[i]
6:      $d' \leftarrow d$ 
7:      $p \leftarrow$  Create empty token probability list
8:     for all  $t \in T$  do
9:        $d'[t] \leftarrow$  [MASK]
10:    end for
11:     $output \leftarrow$  FORWARD( $lm, d'$ )
12:    for all  $t \in T$  do
13:       $prob \leftarrow output[t][d[t]]$ 
14:      APPEND( $p, prob$ )
15:    end for
16:    APPEND( $P, mean(p)$ )
17:  end for
18:   $\hat{P} \leftarrow$  descending sort( $P$ )
19:  return -Log(RANKMEAN( $\hat{P}$ ))
20: end procedure
```

C CDSR Collecting

The Cochrane Database of Systematic Reviews (CDSR) includes peer-reviewed systematic reviews covering various topics in the healthcare domain. In particular, each review in CDSR contains an abstract and a plain language summary which has been required from authors submitting a review

⁷<https://huggingface.co/allenai/led-base-16384>

since 2015. Prior to this, they were written by Cochrane staff with specialized training. Following the collecting process introduced by Guo et al. (2020)⁸ we extracted 8442 abstracts paired with their PLS from CDSR reviews from up to 14th Sept 2022. The data can be downloaded from the official API⁹ but it may differ from our experimented reviews due to the change from Cochrane. The average length of our collected CDSR abstracts by word is around 721 while the average length of PLS is about 395.

D Human Evaluation Setup

The evaluation samples are 10 documents drawn from the test set with their pairs of target summaries and the corresponding pairs of generated summaries by the multi-heads abstractive model (MH Abs) due to its highest ROUGE scores and lowest n-grams overlapping.

Our standard for eligible evaluators is being able to read and write in academic English and having an undergraduate degree. Specifically, we recruit four *experts* who have a medical undergraduate degree and either have or pursue a doctorate in the biomedical area and another four *laypeople* with no biomedical related degree or experience in medical institutions. The final score of each aspect is averaged across tested documents and marking evaluators.

⁸<https://github.com/qiuweipku/Plain language summarisation>

⁹<https://www.cochranelibrary.com/cdsr/reviews>

E Example Output

Target Technical Summary Sex differences in schizophrenia are well known, but their genetic basis has not been identified. We performed a genome-wide association scan for schizophrenia in an Ashkenazi Jewish population using DNA pooling. We found a female-specific association with rs7341475, a SNP in the fourth intron of the reelin (RELN) gene ($p = 2.9 \times 10^{-5}$ in women), with a significant gene-sex effect ($p = 1.8 \times 10^{-4}$). We studied rs7341475 in four additional populations, totaling 2,274 cases and 4,401 controls. A significant effect was observed only in women, replicating the initial result ($p = 2.1 \times 10^{-3}$ in women; $p = 4.2 \times 10^{-3}$ for gene-sex interaction). Based on all populations the estimated relative risk of women carrying the common genotype is 1.58 ($p = 8.8 \times 10^{-7}$; $p = 1.6 \times 10^{-5}$ for gene-sex interaction). The female-specific association between RELN and schizophrenia is one of the few examples of a replicated sex-specific genetic association in any disease.

Generated Technical Summary Schizophrenia is a common psychiatric disorder of unknown aetiology. Sex differences in the risk of developing schizophrenia have been reported for a number of loci, but the robustness of these claims is open to doubt. We have carried out a genome-wide association study of schizophrenia using DNA pools of cases and controls from the Ashkenazi Jewish population. We identified a single SNP in Chromosome 7 (rs7341475) in the fourth intron of the reelin gene (RELN) that was replicated in the UK case-control sample with an effect in the same direction ($p = 1.3 \times 10^{-4}$). The odds ratio (ORG) for women in the combined replication set was 1.41 (95% confidence interval [CI] 1.3–1.6) for men and ORG of 0.97 for women. We tested the association in four other sample sets, three of European ancestry and one of Chinese descent. We found no evidence for stratification in the US sample, indicating that the risk for schizophrenia cannot be caused by population structure. Our result of a female-specific association of RELN with schizophrenia may suggest a possible pathway where sex hormones modulate gene expression, which by altering cortical structure, increases susceptibility to psychosis.

Target Plain Language Summary Schizophrenia is a complex mental disease, which includes symptoms of delusions, hallucinations, disorganized speech, aberrant behavior, lack of emotional expression, diminished motivation, and social withdrawal. The cause of schizophrenia is unknown, but there is extensive evidence that genetics play a significant role in its aetiology. We studied the genetic basis of schizophrenia by analyzing around 500,000 genetic variants distributed across the whole human genome in DNA from schizophrenic patients and controls. We analyzed separately the DNA from men and women, and identified a genetic variant that increases the risk of developing schizophrenia in women only. The genetic variant is estimated to increase the risk of schizophrenia for women carrying the risk variant by 1.4-fold. The genetic variant is in a gene called reelin, which is known to play a part in brain development. However, it is still unclear how this genetic variant predisposes to schizophrenia nor why it is specific to women only.

Generated Plain Language Summary Schizophrenia is a common psychiatric disorder of unknown aetiology. The heritability of schizophrenia is approximately 80%. However, sex differences in the risk of developing schizophrenia have so far been reported for a number of loci, but the robustness of these claims is open to doubt; results have yet to be corroborated. In this study, we carried out a genome-wide association study using DNA pools of cases and controls from the Ashkenazi Jewish population in Northern Ireland and the Republic of Ireland. We identified one SNP in the reelin gene, rs7341475, in the fourth intron of the gene. The SNP is located in Chromosome 7 (RELN), a gene previously studied for association with schizophrenia. We found that the frequency of this SNP is increased in women with schizophrenia, suggesting a possible pathway where sex hormones modulate gene expression, which by altering cortical structure, increases susceptibility to psychosis. Our result of a female-specific association of RELN with schizophrenia may suggest that sex hormones may mediate changes in the function of the RELN gene.

Target Technical Summary Pathogen perception by the plant innate immune system is of central importance to plant survival and productivity. The Arabidopsis protein RIN4 is a negative regulator of plant immunity. In order to identify additional proteins involved in RIN4-mediated immune signal transduction, we purified components of the RIN4 protein complex. We identified six novel proteins that had not previously been implicated in RIN4 signaling, including the plasma membrane (PM) H⁺-ATPases AHA1 and/or AHA2. RIN4 interacts with AHA1 and AHA2 both in vitro and in vivo. RIN4 overexpression and knockout lines exhibit differential PM H⁺-ATPase activity. PM H⁺-ATPase activation induces stomatal opening, enabling bacteria to gain entry into the plant leaf; inactivation induces stomatal closure thus restricting bacterial invasion. The *rin4* knockout line exhibited reduced PM H⁺-ATPase activity and, importantly, its stomata could not be re-opened by virulent *Pseudomonas syringae*. We also demonstrate that RIN4 is expressed in guard cells, highlighting the importance of this cell type in innate immunity. These results indicate that the Arabidopsis protein RIN4 functions with the PM H⁺-ATPase to regulate stomatal apertures, inhibiting the entry of bacterial pathogens into the plant leaf during infection.

Generated Technical Summary Plant innate immunity is essential for the survival of many pathogens. However, how pathogen-associated molecular patterns (PAMPs) and effector-triggered immunity (ETI) interact remains elusive. Here, we report the identification and characterization of the Arabidopsis RIN4 protein complex in planta. We demonstrate that the *rin4* knockout line exhibits decreased PM H⁺-ATPase activity, whereas the *rins2/rin4* double mutant line exhibits increased AHA1 and AHA2 activity. We also show that RIN1/AHA2 is expressed in guard cells along with other PTI and ETI signaling components. Our results support a model in which RIN2 associates with the C-terminal autoinhibitory domain of the PMH⁺-atPase to regulate leaf stomatal apertures in response to pathogen effectors.

Target Plain Language Summary Plants are continuously exposed to microorganisms. In order to resist infection, plants rely on their innate immune system to inhibit both pathogen entry and multiplication. We investigated the function of the Arabidopsis protein RIN4, which acts as a negative regulator of plant innate immunity. We biochemically identified six novel RIN4-associated proteins and characterized the association between RIN4 and the plasma membrane H⁺-ATPase pump. Our results indicate that RIN4 functions in concert with this pump to regulate leaf stomata during the innate immune response, when stomata close to block the entry of bacterial pathogens into the leaf interior.

Generated Plain Language Summary Plants are continuously exposed to a variety of microorganisms. In order to successfully avoid infection, they have evolved a series of defense mechanisms that work in concert to limit pathogen invasion and multiplication. The most well-studied bacterial effectors come from *P. syringae* pv. *tomato* (Pst), the causal agent of bacterial speck on Arabidopsis thaliana. Pst utilizes the type III secretion system (PTI) to deliver effector proteins into the plant cell during infection, resulting in effector-triggered immunity (ETI). However, how pathogen perception activates immune responses and signaling overlap between PTI and ETI remains elusive. In this study, we report the identification and characterization of the RIN4 protein complex in planta. We identified the PM H⁺-ATPases AHA1 and AHA2, whose interaction we characterized in greater detail. We also demonstrate that the *rin4* knockout line cannot re-open its stomata in response to virulent Pst. Importantly, we also show that Rin4 is expressed in guard cells along with other PTI signaling components. Our findings are consistent with a model in which RIN 4 associates with the C-terminal autoinhibitory domain (AHA1/AHA2) to regulate leaf stomatal apertures in

Table 9: Examples of target summary and generated output from Mutli Heads Abstractive Model