GenCompareSum: a hybrid unsupervised summarization method using salience

Jennifer A Bishop[1], Qianqian Xie[1], Sophia Ananiadou[1,2]
[1]National Centre for Text Mining, Department of Computer Science, The University of Manchester, Manchester, United Kingdom
[2]Alan Turing Institute, London, United Kingdom
{jennifer.bishop-2@ postgrad.,qianqian.xie@,sophia.ananiadou@}manchester.ac.uk

Abstract

Text summarization (TS) is an important NLP task. Pre-trained Language Models (PLMs) have been used to improve the performance of TS. However, PLMs are limited by their need of labelled training data and by their attention mechanism, which often makes them unsuitable for use on long documents. To this end, we propose a hybrid, unsupervised, abstractive-extractive approach, in which we walk through a document, generating salient textual fragments representing its key points. We then select the most important sentences of the document by choosing the most similar sentences to the generated texts, calculated using BERTScore. We evaluate the efficacy of generating and using salient textual fragments to guide extractive summarization on documents from the biomedical and general scientific domains. We compare the performance between long and short documents using different generative text models, which are finetuned to generate relevant queries or document titles. We show that our hybrid approach out-performs existing unsupervised methods, as well as state-of-the-art supervised methods, despite not needing a vast amount of labelled training data.

1 Introduction

Recent advancements in transformer-based architectures have enabled improvements in natural language processing (NLP) tasks. The use of encoder-decoder models, such as the T5 language model (Raffel and al., 2020) in generative linguistic tasks, such as abstractive summarization (Cachola et al., 2020) and query generation (Nogueira and Lin., 2019; Klein and Nabi, 2019), have been shown to significantly improve performance over existing methods. Bidirectional-encoder transformer architectures, namely BERT-based PLMs (Devlin et al., 2018) have also proven to be powerful for a broad range of NLP tasks, including text summarization (Liu and Lapata, 2019).

Whilst transformers have made great advancements in their ability at capturing semantic knowledge, they have also introduced new limitations. Firstly, they are restricted by the number of tokens that they can process at any one time. Another issue is the computational cost of finetuning the attention mechanisms embedded in transformers. These constraints are challenging for recent text summarization methods, often resulting in analysis being done on a truncated version of a document (Cachola et al., 2020; Liu and Lapata, 2019; Xu et al., 2020; Zhong et al., 2020; Dou et al. 2021; Zhang and Zhao, 2020). Since summarization should be able to succinctly capture the meaning of very long documents in a few sentences, the requirement to truncate a document before summarization is a major disadvantage. As a result, recent works have shifted their attention towards addressing the issue of long document summarization (Xiao and Carenini, 2020; Grail et al., 2021; Rohde et al., 2021; Xiao and Carenini, 2019). However, these are mostly supervised methods, requiring large amounts of labelled training data, which are often unavailable or time-consuming and costly to produce.

We address the challenges of supervised methods by adopting a hybrid unsupervised approach, where the PLMs are required only to act on short sections of the document at any time, meaning that our method can be extended to any document length. Furthermore, by nature of it being an unsupervised approach, it does not require manually labelled training data for the extractive summarization task. To-date, unsupervised methods for text summarization have generally used graph-based methods (Erkan and Radev, 2004; Mihalcea and Tarau, 2004; Liang et al 2021; Zheng and Lapata, 2019; Done et al., 2021), the more recent of these using

Proceedings of the BioNLP 2022 workshop, Dublin, Ireland, pages 220–240

transformer-based embeddings to calculate weights between the nodes in the graph (Zheng and Lapata, 2019; Done et al., 2021). We differ from these previous approaches as we do not use a graph-based model and instead evaluate the effectiveness of a novel approach – generating and using salient textual fragments to guide the extractive summarization. Moreover, earlier unsupervised, graph-based methods have been criticised in their ability to effectively represent documents which present multiple facts (Liang et al., 2021). Our method addresses this by generating multiple salient texts per document, thus enabling it to represent multiple facts per document.

Text summarization methods are divided into extractive and abstractive groupings. Extractive methods select the most relevant sentences from a document and abstractive methods consider the most relevant pieces of information to produce new textual fragments which convey the core message. Although abstractive summarization has the potential to be more succinct and readable, in its current state it cannot be trusted to be factually consistent (Wallace et al., 2021), making it unsuitable in many practical applications, such as summarization of biomedical articles for use by clinicians. Furthermore, Huang (2020) showed extractive techniques to outperform their abstractive counterparts in human evaluation. See (2017) recognises the advantage of hybrid extractive-abstractive summarization methods and uses a pointer-generator approach, where the model is mostly abstractive, but identifies and copies key facts directly from the source document to try to reduce factual inconsistency. We consider these factors and choose also to opt for a hybrid approach; however, we differ from See (2017) in our use of abstractive models. Specifically, we use transformer-based models for the generation of salient points, but ultimately, we generate an extractive summarization to ensure factual consistency.

Our method, GenCompareSum is a two-step hybrid summarization approach. GenCompareSum first splits a document into sections of several sentences and walks through them, generating salient textual fragments which represent each section. We experiment with different generative models, which are finetuned to predict either queries or document titles, that best represent a section of the document. Our method then uses these generated textual fragments to guide an unsupervised extractive summarization by calculating the BERTScore similarity between each of the generated texts and each of the sentences in the source document. We then select the sentences with the highest scores to form the extractive summarization. We evaluate our approach on short and long versions of data sets from the biomedical and scientific domains. Furthermore, we compare the use of different PLMs for generating salient textual fragments.

Our main contributions are as follows:

1. A novel two-step unsupervised hybrid abstractive-extractive summarization method, which generates salient textual fragments - queries and document titles - which represent sections of a document, and then uses them to guide the extractive summarization step.
2. The fusion of state-of-the-art PLMs with unsupervised approaches, to achieve a summary which harnesses the semantic knowledge of transformer-based models, whilst being extendable to any length document, without requiring a large corpus of training data.
3. Evaluation results demonstrate our hybrid method outperforms both existing unsupervised methods and state-of-the-art supervised methods, both on long and short documents.

2 Methods

We propose GenCompareSum, a hybrid abstractive-extractive model, which makes use of transformer-based architectures but is extendable to any document length, can represent multiple facts, and does not require vast amounts of training data. The method is comprised of two steps: first, using a generative model to produce salient textual fragments, i.e., queries or document titles, which represent key points from across a document, then a comparison between these salient fragments and each sentence, to select the most important sentences from across the document. A representation of our method can be seen in Figure 1. We make our code publicly available[1].

[1] https://github.com/jbshp/GenCompareSum

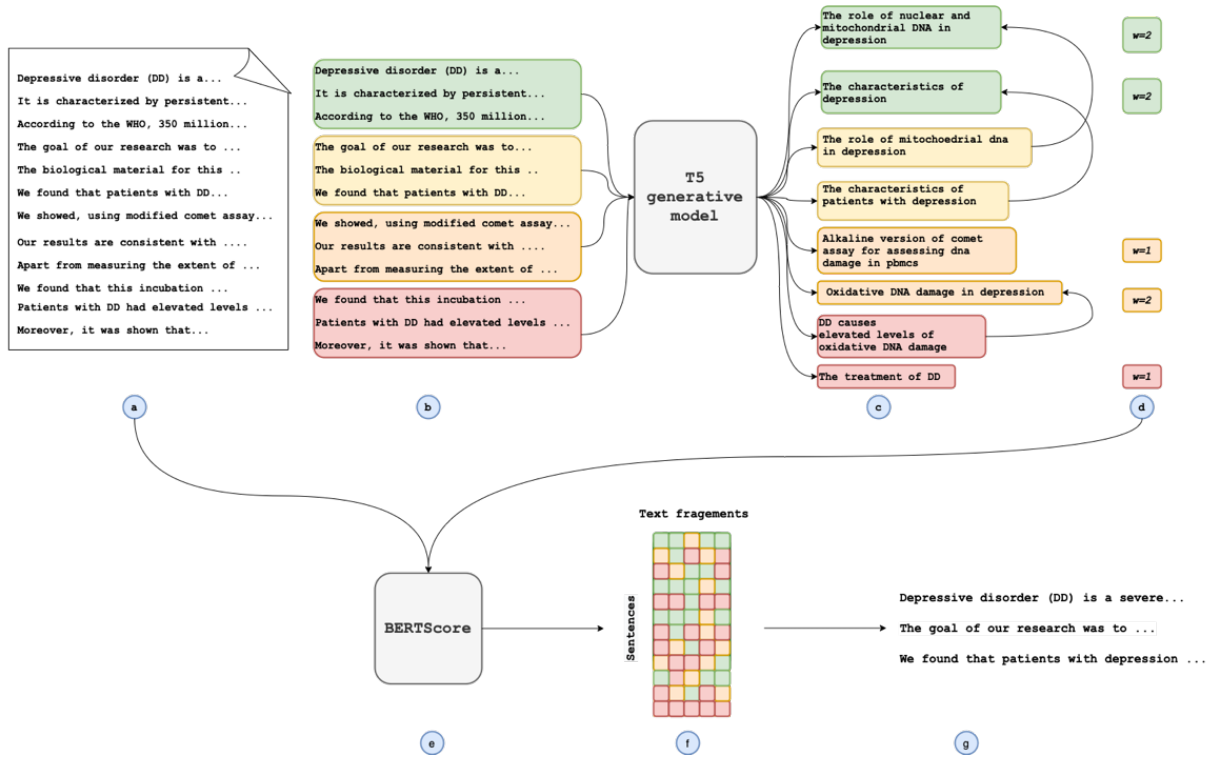

Figure 1: GenCompareSum pipeline. (a) We split the document into sentences. (b) We combine these sentences into sections of several sentences. (c) We feed each section into the generative text model and generate several text fragments per section. (d) We aggregate the questions, removing redundant questions by using n-gram blocking. Where aggregation occurs, we apply a count to represent the number of textual fragments which were combined and use this as a weighting going forwards. The highest weighted textual fragments are then selected to guide the summary. (e) The similarity between each sentence from the source document and each selected textual fragment is calculated using BERTScore. (f) We create a similarity matrix from the scores calculated in the previous step. These are then summed over the textual fragments, weighted by the values calculated in step (d), to give a score per sentence. (g) The highest scoring sentences are selected to form the summary.

2.1 Text splitting

Given a document D, we first split it into sentences s, such that $D = \{s_1, \ldots, s_n\}$, using the Stanford CoreNLP software package (Manning et al., 2014). We then combine sentences into document sections of x sentences, i.e., $D = \{p_1, \ldots, p_m\}$; $m = ceil\left(\frac{n}{x}\right)$. We chose not to use any pre-defined sections already existing within the documents as we found that the documents were not consistently extracted into their sections well across the different datasets. Splitting the document into a consistent number of sentences per section removes the requirement for high quality text extraction into document sections. The number of sentences x used to form the short text sections was decided via experimentation on the validation data sets.

2.2 Salient text generation

T5 (Raffel and al., 2020) is a sequence-to-sequence model, pre-trained on a cleaned and pre-processed version of the Common Crawl[2] data set – a data set consisting of textual content scraped from the internet. T5-based models have been shown to be high performing sequence-to-sequence models across a range of generative tasks, from question generation (Nogueira and Lin, 2019), to graph-to-text generation (Ribeiro et al., 2021), to generative common-sense reasoning (Yuchen Lin, et al., 2020), to abstractive text summarization (Zhang and Zhao, 2020; Goodwin, 2020). The T5 model uses an encoder-decoder architecture and is pre-trained via an unsupervised task in which 15% of tokens are masked; the masked words can be individual words or a span of words; the target of the training objective is to predict these

[2] https://commoncrawl.org

masked words, given the un-masked tokens and their respective positions. For downstream tasks, the pre-trained T5 model is finetuned using pairs of input and output sequences. A diagram of the T5 architecture and its pre-training and finetuning settings can be seen in Appendix A. We experiment with several T5-based models for the salient text generation task.

We use each section, p, as an input to a generative model to give k salient texts t, which aim to encapsulate the key facts of that section:

$$\{t_1, \dots, t_k\} = text\_gen(p). \quad (1)$$

In the case where p is longer than 512 tokens, it is truncated. We then aggregate the generated textual fragments from across the document sections to give $T = \{t_1, \dots, t_{mk}\}$.

For the generation of the textual fragments, we first experiment with a T5-based model finetuned with a query generation task in the general domain. This model, provided textual input, aims to generate queries which ask the most relevant questions of it. We use docTTTTTquery (Nogueira and Lin, 2019), a question generation model trained on the MS-MARCO data set (Bajaj et al., 2018), which is a question-answer data set generated from Bing's[3] search query logs. Surita et al. (2020), showed this pre-trained model to be effective at generating questions for long, biomedical texts.

Second, we follow the approach taken by Nogueira and Lin (2019) and finetune our own model on long-answer - query pairs from the biomedical domain, details of which can be found in Appendix B. We refer to this model as 't5-med-query'.

Last, we experiment using an open-source T5-based model, finetuned on abstract-title pairs from the scientific domain[4]. This approach has shown to be effective at proxying highly abstractive summaries (Cachola et al., 2020). We apply this model to our problem space, generating potential 'titles' for each document section. We refer to this model as 't5-s2orc-title'.

2.3 N-gram blocking

N-gram blocking is a technique which is applied to reduce redundancy and improve coverage in summarization models (Liu and Lapata, 2019). We apply n-gram blocking to the generated textual fragments, resulting in $T^* \subseteq T$, where $T^* = \{t_1, \dots t_{l,l \leq mk}\}$. Where we have removed generated texts by applying this technique, we keep a count of how many times a similar textual fragment was seen before n-gram blocking. We associate this count with the remaining generated text after n-gram blocking. We refer to these counts as weights, which can be described by $w = \{w_1, \dots, w_l\}$, such we have one weight associated with each generated textual fragment remaining after n-gram blocking. A visualization of this can be seen in steps c and d of Figure 1. We then take the top $q; q < l$ generated texts after ordering by the weight.

2.4 Text vector comparison

BERT-based comparisons have been shown to outperform traditional sentence comparative metrics like TF-IDF when used in unsupervised summarization tasks (Done et al., 2021). Furthermore, they have been demonstrated to align better with human judgement of text similarity than n-gram matching approaches during evaluation, likely due to their ability to match based on semantic meaning and their penalization of word re-ordering which changes a text's meaning (Zhang et al,. 2020). BERTScore (Zhang,et al., 2020) uses BERT-based token embeddings, calculates the cosine similarity between them and uses greedy matching to match each token in the first text to its most similar token in the second; these scores are averaged across the sentences to give precision, recall and F1 scores which quantify the similarity between two texts.

[3] https://www.bing.com

[4] https://huggingface.co/doc2query/S2ORC-t5-base-v1

Data set	Instances			Input length – Truncated document		Input length – Full document		Target length	
	Train	Val	Tokens	Tokens	Sentences	Tokens	Sentences	Tokens	Sentences
PubMed	117108	6631	3209	525	20	3209	124	208	9
S2ORC	47474	9490	4312	523	19	4312	154	250	9
CORD-19	31505	6299	5240	525	18	5240	206	232	8
ArXiv	202917	6436	6515	528	20	6515	249	279	11

Table 1: Description of the four data sets used in the extractive summarization experiments. For each data set, we give the number of articles in each of the train, validation and test splits, the mean number of tokens and sentences in the input research article, as well as the mean number of tokens and sentences in the gold summary (abstract) of the articles.

We weight the score by w, the count representing the number of textual fragments which were aggregated during n-gram blocking to give:

$$score_i = \sum_1^{Z=q} w_z * BERTScore(s_i, t_z) \quad (2)$$

We then select the sentences with the highest score to form our summary and reorder them back into the sequence that they appear within the original document.

3 Experiments

3.1 Data sets

We evaluate the efficacy of our hybrid summarization model with four publicly available data sets from the biomedical and scientific domains. All four data sets consist of full-article research papers and their corresponding abstracts. In line with previous literature, we use their abstracts as the target summaries. The data sets included in our experiments are CORD-19 (Wang et al., 2020), PubMed and ArXiv (Cohan et al. 2018), and S2ORC (Lo et al., 2020). The CORD-19 data set used is the version released on 2020-06-28, containing 57,037 articles relating to COVID-19. The S2ORC data set is a large corpus of scientific literature across several domains; we select a random subset of 63,709 articles tagged as being from the biological and biomedical domains. The PubMed and ArXiv data sets are from the biomedical and scientific domains respectively.

For the S2ORC and CORD-19 data sets, we split the data set by sampling randomly to create training/validation/test sets using the ratio 75/15/10. For the PubMed and ArXiv data sets, we use the train/validation/test sets given in the resources associated with the original paper.

Since most previous literature using transformer-based models in their methods either evaluates them on short or truncated texts (Cachola et al., 2020; Liu and Lapata, 2019; Xu et al., 2020; Zhong et al., 2020; Dou et al., 2021; Zhang and Zhao, 2020), we also create short data sets for evaluating our models. We create these data sets by truncating documents to the end of the sentence which contains their 512th token. We evaluate our models both on the short and full-text versions of the four data sets described above. Table 1 gives, for each data set, the mean number of tokens and sentences for the documents and their target summaries.

As training is not required for unsupervised models, for these methods only the test data sets are used. To train the supervised method, BERTExtSum (Liu and Lapata, 2019), which we implement for comparison, we use the training data set to train the model and the validation data set to select the best performing epoch for evaluation on the test set.

3.2 Parameter selection

To select the optimal parameters for our models, we take a constant but random sample of 1000 articles from the PubMed validation data set and

experiment with different combinations of the parameters, details of which can be found in Appendix C. Different methods for calculating text similarity were also compared, namely, BERTScore, SimCSE (Gao et al., 2021) and Sentence Transformers (Reimers and Gurevych, 2019), with BERTScore shown to be the highest performing against a ROUGE metric for the extractive summarization task, details of this analysis can be found in Appendix D.

3.3 Implementation details

We run all experiments requiring GPUs on NVIDIA Quadro RTX 6000 hardware. We report all our results in terms of ROUGE-1, ROUGE-2 and ROUGE-L scores (Lin, 2004), calculated using pyrouge[5] python package.

Several extractive text summarization methods are compared across the short and full-text versions of the four scientific data sets. For the short-text data sets, we take 6 sentences to generate our predictive summary. We choose to give results on a short-text summary for a fair comparison against supervised methods, which are restricted by the length of document that they can easily summarize. For the full-text articles, the number of sentences that we select for the predictive summary is the same as the average number of sentences in the target summaries for a given data set, shown in Table 1. E.g., for the PubMed data set, we select 9 sentences to summarize the full text article.

3.4 Related work

ORACLE summaries indicate the upper bound for extractive text summarization. We calculate ORACLE summaries by adapting code from Liu and Lapata (2019), which applies greedy sentence selection to maximise ROUGE scores.

As baseline methods for comparison, we implement the LEAD method, taking the first n sentences to form the summary, and the RANDOM method, taking a random sample of n sentences to form the summary.

We also compare our method to unsupervised extractive methods, LexRank (Erkan and Radev, 2004), TextRank (Mihalcea and Tarau, 2004) and SumBasic (Nenkova and Vanderwende, 2005), all of which were implemented using the sumy[6] package. LexRank (Erkan and Radev, 2004) and TextRank (Mihalcea and Tarau, 2004) are both graph-based models, based on Google's PageRank algorithm (Brin and Page, 1998), which assume that the sentences with the highest centrality are the most important and use these to form a summary. SumBasic simply assumes that sentences containing the words which are used with the highest frequency across the whole document will be the most important.

Additionally, we compare our method to BERTExtSum (Liu and Lapata, 2019), a state-of-the-art supervised method using BERT-based transformer models. For evaluation on the short data sets, where the documents are truncated at the end of the sentence containing the 512th token, we use their implementation without modification to train and evaluate the models. For the full-text article, we adapt their code, denoted BERTExtSum*, to cycle through the article in 512 token-length blocks and predict the best sentences to select from across this cycle. However, due to hardware limitations and the computational intensity of the attention calculation, we were still required to truncate the document at 1024 tokens to evaluate this method.

Lastly, we implement GenCompareSum and compare the performance between using different generative text models: docTTTTTquery, t5-med-query, and t5-s2orc-title.

4 Experimental Results

4.1 Automatic evaluation

We report the results of our unsupervised hybrid abstractive-extractive method on the extractive summarization task in Table 2.

For the short documents, our method GenCompareSum (t5-s2orc-title) performs best across three out of four of the data sets, and second-best for the fourth data set. There is no clear 'second-best' model out of the methods compared for the short data sets.

[5] https://github.com/bheinzerling/pyrouge

[6] https://github.com/miso-belica/sumy

Model	PubMed			S2ORC			CORD-19			ArXiv		
	R1	R2	RL	R1	R2	RL	R1	R2	RL	R1	R2	RL
Short Document												
ORACLE	47.27	22.85	43.20	49.29	25.42	45.52	43.47	17.75	39.28	47.29	18.49	41.90
RANDOM	34.98	10.82	31.37	34.69	11.06	31.30	31.64	7.91	28.10	34.53	8.88	30.26
LEAD	35.39	12.07	32.28	**40.50**	**16.72**	**37.68**	34.80	10.17	31.67	34.35	8.75	30.61
LexRank	38.48	13.05	34.92	39.44	14.57	36.13	35.65	10.17	32.11	38.98	11.44	34.64
TextRank	38.15	12.99	34.77	40.17	14.84	36.63	36.25	10.61	32.53	37.97	11.58	33.53
SumBasic	36.11	11.06	32.67	35.99	11.99	32.87	33.63	8.82	30.22	37.14	9.83	33.06
BERTExtSum	38.78	14.47	35.43	39.41	16.14	36.38	34.68	10.34	31.42	39.36	11.74	35.09
GenCompareSum (docTTTTTquery)	37.82	13.12	32.41	38.31	14.27	35.17	33.77	9.73	30.66	38.59	11.49	34.50
GenCompareSum (t5-med-query)	38.54	13.67	35.06	38.96	14.78	35.80	36.77	11.24	33.29	38.92	11.59	34.76
GenCompareSum (t5-s2orc-title)	**39.19**	14.35	**35.65**	40.16	15.84	36.91	**36.84**	**11.35**	**33.35**	**39.66**	**12.30**	**35.38**
Long Document												
ORACLE	61.76	36.78	57.61	64.11	39.21	60.16	59.10	32.09	54.63	60.16	32.17	54.97
RANDOM	37.26	11.19	33.66	37.12	10.23	33.73	33.37	7.70	29.98	34.20	8.70	30.64
LEAD	37.23	11.11	33.67	40.50	16.72	37.68	34.61	10.17	31.68	34.70	10.27	31.37
LexRank	41.02	15.83	37.18	42.60	15.84	38.97	39.50	12.65	35.68	33.94	12.09	30.62
TextRank	34.53	12.98	30.99	36.58	13.23	33.10	32.99	10.39	24.47	26.57	9.20	23.74
SumBasic	40.61	12.42	36.54	36.63	10.43	33.68	33.88	8.24	30.86	33.18	7.75	30.29
BERTExtSum*	41.87	16.01	**38.51**	**43.56**	**17.85**	**40.40**	38.95	12.17	35.48	**40.65**	14.01	**36.89**
GenCompareSum (docTTTTTquery)	40.54	14.77	36.83	40.78	14.24	37.43	36.84	11.19	33.51	38.19	12.76	34.55
GenCompareSum (t5-med-query)	41.60	15.67	37.79	41.84	15.10	38.35	39.33	12.31	35.74	37.17	11.97	33.95
GenCompareSum (t5-s2orc-title)	**42.10**	**16.51**	38.25	43.39	16.84	39.82	**41.02**	**13.79**	**37.25**	39.96	**15.15**	36.19

Table 2: Results of the extractive summarization task on the PubMed, ArXiv, s2orc and CORD-19 data sets. The short text version of the data set consists of the articles truncated at the end of the sentence containing the 512th token. We select 6 sentences for the short text summary. For the full-text document prediction, we use select the average number of sentences in the gold summaries of the respective data sets, which are given in Table 1, PubMed: 9, S2ORC: 9, CORD-19: 8, ArXiv: 11. Bold font indicates the top result within a data set, underlined font indicates the second-best result.

Interestingly, for the S2ORC data set, the method outperforming all others is LEAD, i.e., taking the first sentences from the document as the predictive summary. However, in evaluation of the full-text version of the S2ORC data set, it does not hold that LEAD is the best method, and it is seen to be outperformed by several other methods.

For the long document data sets, GenCompareSum (t5-s2orc-title) outperforms all other unsupervised models. A strong unsupervised baseline, LexRank has been shown in prior literature to give competitive performance when compared to supervised approaches (Cohan et al., 2018; Subramanian, Li and Pilault, 2020). In-line with these works, we show LexRank to be the best-performing unsupervised method after our own.

Our method, GenCompareSum (t5-s2orc-title), outperforms LexRank by a large margin – an average $\Delta\text{R1}, \Delta\text{R2}, \Delta\text{R1}$ of 2.35, 1.47, 2.27 across the four data sets. We also demonstrate a slight performance increase over our implementation of the supervised method BERTExtSum*, which we adapted to run over longer documents. The same calculation across the data sets with BERTExtSum* shows us outperforming $\Delta R1, \Delta R2, \Delta R1$ by 0.36, 0.56, 0.06 across the four data sets. Given that our method is unsupervised, and therefore does not require labelled training data and can be

extended to any document length, we believe this is a significant improvement.

Considering the different implementations of GenCompareSum, we can see that, as expected, our results show that finetuning on in-domain data gives notable performance increases. Table 2 shows that the $\Delta R1$ between an out-of-domain query generation model (docTTTTTquery) and a query generation model trained on biomedical data (t5-med-query) were as high as 3 and 2.49 for the short and long articles respectively, for the CORD-19 biomedical data set. However, for the ArXiv data set, which consists of predominantly physical and computer science related research articles, the performance decreased when using the t5-med-query generative model instead of the general domain docTTTTTquery model.

Our best-performing GenCompareSum model, t5-s2orc-title, uses a generative PLM finetuned on document-title pairs from the S2ORC data set to guide the extractive summarization. In many ways, a title can be considered as a highly abstractive summarization (Cachola et al., 2020). A major advantage of this finding is that, although it does require training data to finetune this generative model, document-title pairs are readily available across many domains, thus a model can easily be trained for a specific task without needing extensive manual labelling effort. Furthermore, this model, although finetuned on biomedical and scientific data, is finetuned on a very broad range of documents within these fields. We demonstrate that, despite the broad coverage of fields in its training data, it performs very well when applied to data from a more specific domain, e.g., biomedicine in the PubMed and CORD-19 data sets.

Lastly, we observe that there is big difference in ORACLE scores between the short and full text data sets. Although our models out-perform all other methods evaluated for both short and full text documents, the gap between the best predictive scores in our experiments and the ORACLE upper bound is large for long documents, suggesting that much more research could be done in this space. Furthermore, based on this observation, we also hypothesise that predicting summaries from short documents is a significantly easier task than doing the same for long documents. This is supported by TextRank performing worse on the long documents than on the truncated versions. We believe this is explainable both by the fact that there are much fewer sentences to choose from within a shorter document (we select approximately 32% of all sentences across the data sets for short document predictions and 5% for the full documents), thus less room for error. Furthermore, previous work has shown that often the most important parts of the document are towards the beginning of it (Zheng et al., 2019), implying that there is less 'noise' (i.e., unimportant sentences) to select from a truncated document.

4.2 Qualitative analysis

In Appendix E we provide a randomly sampled PubMed document, the associated generated salient fragments, and the predicted extractive summary given by each of the three GenCompareSum methods. We also provide the gold summary (document abstract) for comparison. In Appendix F, we give the same for a randomly sampled document from the ArXiv data set. In this section, we comment on the difference between the texts generated by the difference T5-based models and hypothesise on how this influences the extractive summary.

The docTTTTTquery model produces questions which are relatively general and imply little biomedical knowledge when provided the PubMed document as input. In this setting, it produces textual fragments such "what is nlrp3". Interestingly, it does manage to produce more complex texts from sections of the ArXiv data set, such as: "what is the contribution of the spiral arm to the resonant structure in the solar neighborhood?".

In comparison, the t5-med-query model, whilst also generating questions, better encapsulates biomedical concepts when given a document from the PubMed dataset, e.g., "what is the role of nuclear and mitochondrial dna damage and repair in people with depression?". However, in line with the ROUGE results given in Section 4.1, it seems to perform less well on out-of-domain (i.e., scientific rather than biomedical) literature, and appears to default to a more general question generation model, generating texts for the ArXiv document such as "what is the effect of a spiral arm?".

The t5-s2orc-title model generates texts which read much more like very short, highly

abstractive summaries. E.g., for the PubMed article, it generated the textual fragment: "the role of the nuclear and mitochondrial dna in depression" and for the ArXiv article it generated: "the spiral arm contribution to the resonant structure of the solar neighborhood". Although outperformed by the title-generation model t5-s2orc-title in the automatic evaluation, on analysis of the generated textual fragments, the query generation models do seem to effectively represent the important facts from an article, especially in the biomedical domain. We hypothesise that our use of BERTScore, to calculate the similarity between salient texts and document sentences, favours the title generation model due to it calculating the similarity between words in different texts and not being designed to answer questions. In future work, we would like to experiment further with the combination of the query generation models and extractive question answering approaches for the extractive summarization task.

5 Future work

In this section we suggest future directions for our research. Firstly, we highlight that our method is generalizable and not restricted to T5-based architectures for the generation of salient text fragments. Therefore, we would like to experiment with different models for this step, e.g., BART-based (Lewis, 2019) models, or models trained with different data.

Another interesting direction would be the inclusion of zoning into the method. As mentioned previously, we chose not to use an article's pre-defined sections as they are often not available. However, it would be interesting to predict a classification for a sentence within the text (e.g., 'Results' for a scientific article), and to incorporate this into the model.

We would also like to evaluate our models on other data sets and domains in future research, e.g., clinical notes, and would like to carry out a human evaluation to validate the results, ideally with experts from the same domain as the data being summarized. Human evaluation would allow for aspects such as fluency, factual consistency, and coherence to be assessed, which have been shown not to necessarily align well with ROUGE evaluation in previous works (Kryscinski et al., 2019; Huang et al., 2020).

Lastly, our analysis (details in Appendix D) showed that BERTScore was the best performing method for calculating text similarity for the extractive summarization task, outperforming methods using sentence embeddings. We hypothesise that this could be due to our evaluation metric being ROUGE scoring, which favours methods that produce summaries containing exactly the same words as the gold summary, rather than semantically similar sentences. Experimentation into different evaluation metrics for extractive summarization, including human evaluation, and how they correlate to the performance of our methods when using different models for calculating text similarity, is also an interesting direction for future work.

6 Conclusion

In this work we propose GenCompareSum, a novel two-step unsupervised hybrid abstractive-extractive method for text summarization. We evaluate the efficacy of using PLMs to generate salient textual fragments which represent the key points of a document – experimenting with generation of both queries and document titles - and using them to guide the second step, extractive summarization. We show that that our unsupervised method, which can be extended to any length of document and does not require a corpus of annotated training data, outperforms over both strong supervised and unsupervised baselines on long and short documents. Furthermore, we show that our best-performing model uses title-document pairs for the generative task, which are readily available across many domains without the need for manual labelling effort.

Author contributions

JA Bishop proposed the research themes, developed the code, conducted the experiments, and drafted the manuscript. Q Xie and S Ananiadou supervised all steps of the work and revised the manuscript. All authors approved the final version of the manuscript.

References

Bajaj, Payal, et al. 2018. "MS MARCO: A Human Generated MAchine Reading

COmprehension Dataset.” *arXiv:1611.09268.*

Beltagy, Iz, Kyle Lo, and Arman Cohan. 2019. "SciBERT: A pretrained language model for scientific text." *arXiv preprint arXiv:1903.10676.*

Brin, Sergey, and Lawrence Page. 1998. “The anatomy of a large-scale hypertextual web search engine.” *Computer networks and ISDN systems* 30 (1-7): 107-117.

Cachola, Isabel, Kyle Lo, Arman Cohan, and Daniel S. Weld. 2020. “TLDR: Extreme Summarization of Scientific Documents.” *Findings of the Association for Computational Linguistics: EMNLP 2020.* 4766-4777.

Cohan, Arman, et al. 2018. “A discourse-aware attention model for abstractive summarization of long documents.” *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference.* 615-621.

Cohan, Arman, et al. 2020. "Specter: Document-level representation learning using citation-informed transformers.*" arXiv preprint arXiv:2004.07180*.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. “Bert: Pre-training of deep bidirectional transformers for language understanding.” *arXiv preprint arXiv:1810.04805* .

Dou, Zi-Yi, Pengfei Liu, Hiroaki Hayashi, and Zhengbao and Neubig, Graham Jiang. 2021. “GSum: A General Framework for Guided Neural Abstractive Summarization.” *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* 4830-4842.

Erkan, Gunes, and Dragomir R. Radev. 2004. “Lexrank: Graph-based lexical centrality as salience in text summarization.” *Journal of artificial intelligence research* 22: 457-479

Gao, Tianyu, Xingcheng Yao, and Danqi Chen. 2021. "Simcse: Simple contrastive learning of sentence embeddings." *arXiv preprint arXiv:2104.08821.*

Travis Goodwin, Max Savery, and Dina Demner-Fushman. 2020. Towards Zero-Shot Conditional Summarization with Adaptive Multi-Task Fine-Tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2020,* pages 3215–3226, Online. Association for Computational Linguistics.

Grail, Quentin, Julien Perez, and Eric Gaussier. 2021. “Globalizing BERT-based Transformer Architectures for Long Document Summarization.” *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume.*

Huang, Dandan, et al. 2020. “What have we achieved on text summarization?” In *Proceedings of the 2020 Conference on Emperical Methods in Natural Language Processing (EMNLP)*, pages 446-469. Association for Computational Linguistics.

Jin, Qiao, et al. 2019. “PubMedQA: A Dataset for Biomedical Research Question Answering.” *arXiv preprint arXiv:1909.06146.*

Klein, Tassilo, and Moin Nabi. 2019. “Learning to answer by learning to ask: Getting the best of gpt-2 and BERT worlds.” *arXiv preprint arXiv:1911.02365.*

Kryscinski, Wojciech, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural Text Summarization: A Critical Evaluation. 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 540–551, Hong Kong, China. Association for Computational Linguistics.

Lewis, Mike, 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.*

Liang, Xinnian, et al. 2021. “Improving unsupervised extractive summarization with facet-aware modeling.” *Findings of the Association for Computational Linguistics: ACL-IJCNLP* .

Lin, Chin-Yew. 2004. “ROUGE: A package for automatic evaluation of summaries.” *Text Summarization Branches Out.* pages 74-81. Association for Computational Linguistics.

Liu, Yang, and Mirella Lapata. 2019. “Text summarization with pretrained encoders.” In *Preceedings of the 2019 Conference onon Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pges 3730-3740, Hong Kong, China. Association for Computational Linguistics.

Lo, Kyle, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel. Weld. 2020. “S2ORC: The Semantic Scholar Open Research Corpus.” *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* pages 4969-4983, Online. Association for Computational Linguistics.

Möller, Timo, Anthony Reina, Raghavan Jayakumar, and Malte Pietsch. 2020. “COVID-QA: A Question Answering Dataset for COVID-19” *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020,* Online. Association for Computational Linguistics.

Manning, Christopher D, and et al. 2014. “The Stanford CoreNLP natural language processing toolkit.” *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations.*

Mihalcea, Rada, and Paul Tarau. 2004. “Textrank: Bringing order into text.” *Proceedings of the 2004 conference on empirical methods in natural language processing.* Pages 404-411. Barcelona, Spain. Association for Computational Linguistics.

Nenkova, Ani, and Lucy Vanderwende. 2005. “The impact of frequency on summarization.” *Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005* 101.

Nentidis, Anastasios et al. 2021. “Overview of BioASQ 2021: The ninth BioASQ challenge on Large-Scale Biomedical Semantic Indexing and Question Answering.” *Experimental IR Meets Multilinguality, Multimodality, and Interaction* 239–263.

Nogueira, Rodrigo, and Jimmy Lin. 2019. “From doc2query to docTTTTTquery.” *Online preprint.*

Raffel, Colin, et al. 2020. “Exploring the limits of transfer learning with a unified text-to-text transformer.” *Journal of Machine Learning Research* 21: 1-67.

Reimers, Nils, and Iryna Gurevych. 2019. "Sentence-bert: Sentence embeddings using siamese bert-networks." *arXiv preprint arXiv:1908.10084.*

Leonardo F. R. Ribeiro, Martin Schmitt, Hinrich Schütze, and Iryna Gurevych. 2021. Investigating Pretrained Language Models for Graph-to-Text Generation. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 211–227, Online. Association for Computational Linguistics.

Rohde, Tobias, Xiaoxia Wu, and Yinhan Liu. 2021. “Hierarchical learning for generation with long source sequences.” *arXiv preprint arXiv:2104.07545.*

See, Abigail, Peter J Liu, and Christopher D Manning. 2017. “Get to the point: Summarization with pointer-generator networks.” *ACL 2017 - 55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers).* pages 1073-1083, Ancouver, Canada, Association for Computational Linguistics.

Subramanian, Sandeep, Raymond Li, Jonathan Pilault and Chris Pal. 2020. “On Extractive and Abstractive Neural Document Summarization with Transformer Language Models.” In *Proceedings ofthe 2020 Conference on Empirical Methods in Natural Language Processing.* pages 9308–9319, Online. Association for Computational Linguistics.

Surita, Gabriela, Rodrigo Nogueira, and Roberto Lotufo. 2020. “Can questions summarize a corpus? Using question generation for characterizing COVID-19 research.” *arXiv:2009.092900.*

Wallace, Byron C, Sayantan Saha, Frank Soboczenski, and Iain J Marshall. 2021. “Generating (Factual?) narrative summaries of RCTs: Experiments with neural multi-document summarization.” *AMIA Jt Summits Transl Sci Proc.* 605–614.

Wang, Lucy Lu, and et al. 2020. “CORD-19: The covid-19 open research dataset.” *arXiv preprint arXiv:2004.10706v4.*

Xiao, Wen, and Giuseppe Carenini. 2019. “Extractive summarization of long documents by combining global and local context.” *arXiv preprint arXiv:1909.08089.*

Xiao, Wen and Giuseppe Carenini. 2020. “Systematically Exploring Redundancy Reduction in Summarizing Long Documents.” *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing.* pages 516-528. Association for Computational Linguistics.

Xu, Shusheng, Xingxing Zhang, Yi Wu, Furu Wei, and Ming Zhou. 2020. “Unsupervised Extractive Summarization by Pre-training Hierarchical Transformers” *Findings of the Association for Computational Linguistics: EMNLP 2020.* pages 1784-1795, Online. Association for Computational Linguistics.

Yuchen Lin, Bill, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020. CommonGen: A Constrained Text Generation Challenge for Generative Commonsense Reasoning. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1823–1840, Online. Association for Computational Linguistics.

Zhang, Jingqing, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. “PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization” *Proceedings of the 37th International Conference on Machine Learning.* 1328-11339.

Zhang, Tianyi, and et al. 2020. “BERTScore: Evaluating Text Generation with BERT” *International Conference on Learning Representations.*

Zheng, Hoa, and Mirella Lapata. 2019. “Sentence Centrality Revisited for Unsupervised Summarization.” In *Procedings of the 57th Annual Meeting of the Association for Computational Liguistcs*. pages 6236-6247, Florence, Italy. Association for Computational Linguistics.

Zhong, M, P Liu, Y Chen, D Wang, X Qiu, and X. Huang. 2020. “Extractive summarization as text matching” *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* Pages 6197-6208, Online. Association for Computational Linguistics.

Zhu, Ming, Aman Ahuja, Da-Cheng Juan, Wei Wei, and Chandan K. Reddy. 2020. “Question answering with long multiple-span answers” *Findings of the Association for Computational Linguistics: EMNLP 2020.* Pages 3840-3849 , Online. Association for Computational Linguistics.

Appendix A. T5 model architecture.

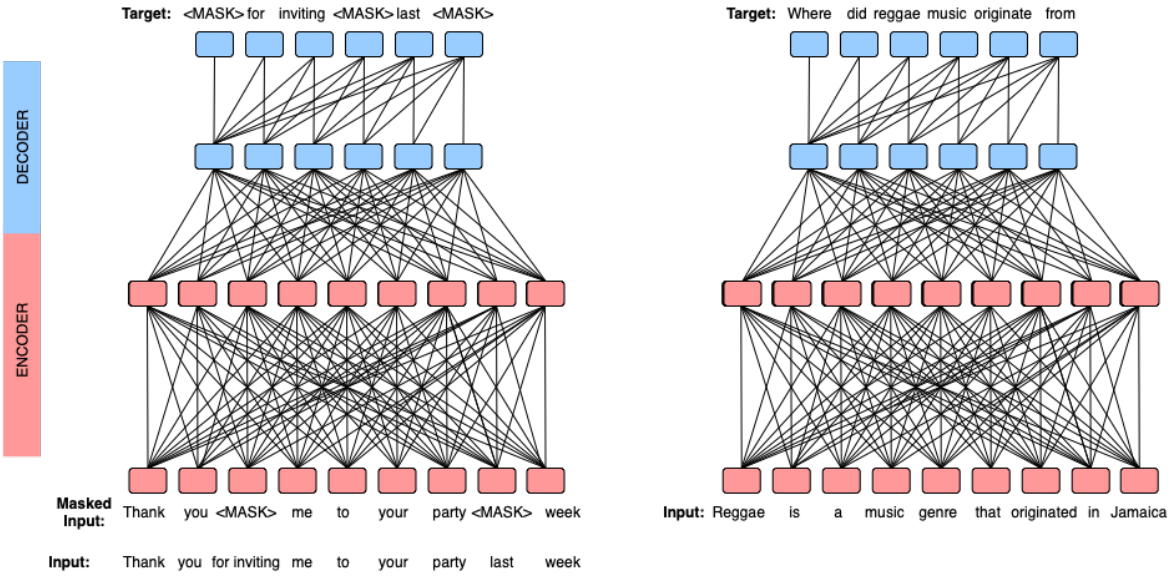

Figure 2: Representations of the two training settings of the T5 encoder-decoder model. The left diagram shows the unsupervised pretraining task, in which a tokenized text containing masked spans is passed to the encoder and the output target of the decoder is the prediction of the masked spans. The right diagram shows the supervised downstream task, where the pre-trained model is finetuned on pairs of tokenized sequences.

Appendix B. t5-med-query training.

To finetune our GenCompareSum (t5-med-query) model, we combine four biomedical data sets to make a large corpus of text-question pairs, where the questions can be answered by the long textual input. From the BioAsq data set (Nentidis et al., 2021), 3,433 'ideal answer'-question pairs were used, 2,720 text-question pairs from COVID-QA (Möller et al., 2020), where the paragraph containing the answer is used as the text input, 61,244 context-question pairs from PubMedQA (Jin et al. 2019), where the 'context' refers to the abstract without its 'conclusion' section, and 27,722 long answer-question pairs from the MASH-QA (Zhu et al. 2020) data set. The t5-base model is loaded and finetuned on this data set for 5 epochs, with a batch size of 8.

Appendix C. Parameter selection.

Parameter Name	Parameter Definition	Parameter range experimented with	Optimal parameter selected
T5 model temperature	Controls randomness of generative text model predictions	0.2-1	0.5
T5 input size (x)	Number of sentences used to form sections to input to T5 text generation model	2-12	4
T5 predictions per input (k)	Number of salient texts generated per section passed to the model	2-6	3
T5 prediction n-gram blocking	Number of consecutive word matches used to determine whether a generated text should be removed due to redundancy when compared to another generated text	No n-gram blocking, n=3, n=4	4
T5 generated texts used for comparison (q)	Number of generated texts used for comparison to the original document sentences	4-12	10
BERTScore embedding model	Base model used in BERTScore package for word-embedding comparison	bert-base-uncased[7], facebook/bart-large-mnli[8], allenai/longformer-large-4096[9], allenai/scibert_scivocab_uncased[10]	bert-base-uncased
BERTScore batch size	Batch size in BERTScore package	64 (used default)	64
Score weighting	Optional multiplication of scores by frequency of question occurrence	True/False	True
Sentence selecton n-gram blocking	Number of consecutive word matches used to determine whether a selected sentence should be removed due to redundancy when compared to another selected sentence	No n-gram blocking, n=3, n=4	4

Table 3: Parameters experimented with, and selected for use, in the GenCompareSum models.

[7] https://huggingface.co/bert-base-uncased

[8] https://huggingface.co/facebook/bart-large-mnli

[9] https://huggingface.co/allenai/longformer-base-4096

[10] https://huggingface.co/allenai/scibert_scivocab_uncased

Appendix D. Analysis of methods for calculating text similarity

In this section we compare different methods for calculating the similarity between the generated salient text fragments and the document sentences. We use our best performing model, GenCompareSum (s2orc-title), and implement different models for the text comparison step. We present results for the extractive summarization task on the PubMed 'Short Document' data set.

We compare BERTScore, a method which uses word embeddings to calculate the similarity between texts, with two other methods to calculate the similarity between texts using sentence embeddings. Sentence Transformers (Reimers and Gurevych, 2019) is trained with a triplet / siamese bert-based architecture and a training objective designed to minimize distances between similar sentences. We implement this method with their python package[11]. We compare both their suggested base model for the general domain 'all-mpnet-base-v2' and a model trained to calculate document-level similarity for scientific documents 'allenai-specter' (Cohan et al., 2020). We also implement SimCSE (Gao et al., 2021), which generates sentence embeddings with a model trained using contrastive learning. For this method, we use the general-domain base model which is suggested to be the best performing in SimCSE's documentation[12]. For the BERTScore method, we experiment with base models from the general domain, namely 'bert-base-uncased'[13], which was used in our implementations to give the results in Table 2 of the main manuscript, and a base model pretrained on data from the scientific domain (Beltagy et al., 2019), 'allenai/scibert_scivocab_cased'.

Table 4 gives the results. We can observe that BERTScore, implemented with a base model from the general domain, outperforms all other methods compared for calculating text similarity on the extractive summarization task when evaluated using ROUGE metrics.

Text similarity method	R1	R2	RL
BERTScore (bert-base-uncased)	39.19	14.35	35.65
BERTScore (allenai/scibert_scivocab_cased)	37.78	13.40	34.45
SentenceTransformer (all-mpnet-base-v2)	39.03	14.20	35.45
SentenceTransformer(allenai-specter)	38.20	13.41	34.67
SimCSE (princeton-nlp/sup-simcse-roberta-large)	38.62	13.73	35.07

Table 4: A comparison of ROUGE-1,-2 and -L results for the PubMed Short Document data set on the extractive summarization task, using different methods for calculating text similarity between generated salient texts and the document's sentences. The method is given in the first column, with the base model used in its implementation given in brackets.

[11] https://github.com/UKPLab/sentence-transformers
[12] https://github.com/princeton-nlp/SimCSE
[13] https://huggingface.co/bert-base-uncased

Appendix E. Example output of our method on a PubMed article.

PubMed Sample Document and Predictions	
PubMed Sample Document	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4329942/
PubMed Sample Abstract (Target Summary)	depressive disorder (dd) , including recurrent dd (rdd) , is a severe psychological di sease , which affects a large percentage of the world population . although pathogenes is of the disease is not known , a growing body of evidence shows that inflammation t ogether with oxidative stress may contribute to development of dd . since reactive oxy gen species produced during stress may damage dna , we wanted to evaluate the exten t of dna damage and efficiency of dna repair in patients with depression. material / we measured and compared the extent of endogenous dna damage single - and double - st rand breaks , alkali - labile sites , and oxidative damage of the pyrimidines and purine s in peripheral blood mononuclear cells isolated from rdd patients (n = 40) and healt hy controls (n = 46) using comet assay . we also measured dna damage evoked by h ydrogen peroxide and monitored changes in dna damage during repair incubation. we found an increased number dna breaks , alkali - labile sites , and oxidative modificatio n of dna bases in the patients compared to the controls . exposure to hydrogen peroxid e evoked the same increased damage in both groups . examination of the repair kineti cs of both groups revealed that the lesions were more efficiently repaired in the contro ls than in the patients. the first time we showed that patients with depression , compar ed with non - depresses individuals , had more dna breaks , alkali - labile sites , and o xidative dna damage , and that those lesions may be accumulated by impairments of t he dna repair systems . more studies must be conducted to elucidate the role of dna da mage and repair in depression .
Salient Texts - GenCompareSum (docTTTTTquery)	• what is nlrp3 • how long does it take for dna damage to be repaired • what is the oxidative modification of purines • what is the main activator of nlrp3 • what is the damage caused by dna repair • what is the role of mitochondrial dna in depression • what is oxidative damage in dna • does oxog cause depression • what is the dna damage response • what is the oxidative damage of pyrimidines and purines?
Predicted Summary - GenCompareSum (docTTTTTquery)	• since the findings described above are inconsistent , we wanted to determine if the oxidative modification of purines , like 8 - oxog , and pyrimidines are present in a higher degree in patients with depression than in controls . • to achieve these objectives , we measured and compared the extent of endog enous dna damage single - and double - strand breaks , alkali - labile sites , a nd oxidative damage of the pyrimidines and purines in pbmcs isolated from dd patients and healthy controls . • we evaluated the level of basal endogenous dna damage by subjecting pbmcs to comet assay procedure immediately after isolation from blood . • moreover , we estimated the extent of oxidative dna damage by employing m odified comet assay with 2 glycosylases : nth removing oxidized pyrimidines and hogg1 excising oxidized purines . • figure 3 shows mean dna damage changes in pbmcs of the patients with depr ession and the controls without psychiatric disturbances during the repair inc ubation . • the goal of our research was to examine the susceptibility of rdd patients to d na damage induced by oxidative stress by measuring the level of endogenous dna damage , including oxidative dna damage , the amount of dna damage in duced by h2o2 , and efficiency of dna damage repair in the patients as compa red to the controls without psychological disturbances . • apart from measuring the extent of endogenous dna damage , we also estimat ed the amount of dna damage induced by the incubation of pbmcs with h2o2 and efficiency of its repair .

	• additionally , we monitored the repair efficiency of the induced dna damage . • moreover , nlrp3 inflammasome , activation of which was detected in the pat ients pbmcs , was also found to inhibit dna repair after induction of oxidative stress .
Salient Texts - GenCompareSum (t5-med-query)	• what was the purpose of the study? • what is the alkaline version of the comet assay? • what is the effect of pbmcs on basal endogenous dna damage? • what is the incubation time for dna repair? • what is the role of nuclear and mitochondrial dna damage and repair in peopl e with depression? • is it possible to study the susceptibility of rdd patients to dna damage induce d by oxidative stress? • what is recurrent depressive disorder? • what is the association between 8 - oxog and depression in japanese office w orkers? • which is the most versatile nlr? • what enzymes are bifunctional glycosylases?
Predicted Summary - GenCompareSum (t5-med-query)	• moreover , we also wanted to know if the patients have elevated levels of oth er kinds of dna damage , such as strand breaks . • we evaluated the level of basal endogenous dna damage by subjecting pbmcs to comet assay procedure immediately after isolation from blood . • figure 2 shows basal endogenous dna damage and the damage induced after 10 - min incubation with 20 m h2o2 in pbmcs isolated from the patients and controls without psychiatric disturbances . • figure 3 shows mean dna damage changes in pbmcs of the patients with depr ession and the controls without psychiatric disturbances during the repair inc ubation . • figure 5 compares basal endogenous dna damage and the level of this param eter at the end of the repair incubation in pbmcs of the patients and the contr ols measured by the alkaline version of comet assay . • the goal of our research was to examine the susceptibility of rdd patients to d na damage induced by oxidative stress by measuring the level of endogenous dna damage , including oxidative dna damage , the amount of dna damage in duced by h2o2 , and efficiency of dna damage repair in the patients as compa red to the controls without psychological disturbances . • apart from measuring the extent of endogenous dna damage , we also estimat ed the amount of dna damage induced by the incubation of pbmcs with h2o2 and efficiency of its repair . • additionally , we monitored the repair efficiency of the induced dna damage . • there is a need for further studies to define the role of nuclear and mitochond rial dna damage and repair in people with depression , and their implications for clinical outcome .
Salient Texts - GenCompareSum (t5-s2orc-title)	• dna damage in patients with depression. • oxidative dna damage in depression • the oxidative dna damage in patients with renal failure • activation of nlrp3 by oxygen species in pbmc patients. • activation of mitochondrial nlrp3 in patients with pbmcs. • urinary 8-oxog in japanese office workers • the use of the alkaline version of comet assay for assessing dna damage in pb mcs • the role of the nuclear and mitochondrial dna in depression. • the role of the dna repair rate in the repair of pbmcs in patients with squamou s cell carcinoma.
Predicted Summary - GenCompareSum (t5 -s2orc-title)	• in agreement with this , activation of nlrp3 in pbmcs of the patients was acco mpanied by increased lipid peroxidation , which can be attributed to increase d oxidative stress and elevated mitochondrial ros (mtros) production .

	• moreover , we induced oxidative dna damage in those pbmcs by incubating t hem with hydrogen peroxide , measured the kinetics of removing of such da mage , and compared the results between the patients and the controls . • we evaluated the level of basal endogenous dna damage by subjecting pbmcs to comet assay procedure immediately after isolation from blood . • figure 2 shows basal endogenous dna damage and the damage induced after 10 - min incubation with 20 m h2o2 in pbmcs isolated from the patients and controls without psychiatric disturbances . • figure 3 shows mean dna damage changes in pbmcs of the patients with depr ession and the controls without psychiatric disturbances during the repair inc ubation . • it is possible that increased oxidative dna damage occurs only in patients wit h more severe forms of depression , or in later stages of the disease develop ment . • these results indicate that in the patients , oxidative dna damage is less effici ently removed than in the controls . • moreover , nlrp3 inflammasome , activation of which was detected in the pat ients pbmcs , was also found to inhibit dna repair after induction of oxidative stress . • for the first time , we showed that patients with depression had elevated level s of dna breaks , alkali - labile sites , and oxidative dna damage , and that the se lesions may be accumulated by impairments of dna repair pathways .

Appendix F. Example output of our method on an ArXiv article.

ArXiv Sample Document and Predictions	
ArXiv Sample Document	https://arxiv.org/abs/0906.4682
ArXiv Sample Abstract (Target Summary)	we study the phase space available to the local stellar distribution using a galactic pot ential consistent with several recent observational constraints . we find that the induced phase space structure has several observable consequences . the spiral arm contribution to the kinematic structure in the solar neighborhood may be as important as the one produced by the galactic bar . we suggest that some of the stellar kinematic groups in the solar neighborhood , like t he hercules structure and the kinematic branches , can be created by the dynamical res onances of self - gravitating spiral arms and not exclusively by the galactic bar . a structure coincident with the arcturus kinematic group is developed when a hot stell ar disk population is considered , which introduces a new perspective on the interpret ation of its extragalactic origin . a bar - related resonant mechanism can modify this kinematic structure . we show that particles in the dark matter disk - like structure predicted by recent lcd m galaxy formation experiments , with similar kinematics to the thick disk , are affect ed by the same resonances , developing phase space structures or dark kinematic grou ps that are independent of the galaxy assembly history and substructure abundance . we discuss the possibility of using the stellar phase space groups as constraints to no n - axisymmetric models of the milky way structure .
Salient Texts - GenCompareSum (docTTTTTquery)	• what is the role of the bar in the local kinematic structure • what is the effect of the non axisymmetric galactic structure on the solar neig hborhood kinematic distribution? • what is the shape of the solar structure at @xmath27 • what is the structure of the hercules branch • what is the effect of a spiral arm • what is the hercules structure • how does the kinematics of the disk affect the galaxy? • which of the following structure is a contribution to the solar neighborhood k inematics? • what type of spiral arm is used to measure observations made in the solar nei ghborhood • what is the contribution of the spiral arm to the resonant structure in the solar neighborhood?
Predicted Summary - GenCompareSum (docTTTTTquery)	• however , it is unclear whether there is any dependence of the induced local solar neighborhood kinematics on the detailed galactic structure . • in order to study the effect of the non - axisymmetric galactic structure on th e solar neighborhood kinematic distribution , we have performed numerical i ntegrations of test particle orbits on the galactic plane , adopting the initial c onditions discussed in sect . • the induced kinematic distribution at the end of the simulation is studied by c onsidering the particles inside a circle of radius @xmath7 centered at the sol ar position . • therefore we focused on the recently induced kinematic structure in the solar neighborhood . • with these initial conditions , we can study the relatively rapid induced effect s of the non - axisymmetric component on the local kinematics . • we conclude that the contribution of the spiral arms to the solar neighborhoo d kinematics may be comparable to that of the bar . • in our simulations the positions of these kinematic arches are modified when the bar is added to the model . • furthermore , these simulations show the important role of the bar in the dev elopment of the local kinematic structure .

	• the spiral arm contribution to the resonant structure in the solar neighborhoo d may be comparable to that of the galactic bar . • in summary , the imprints of the non - axisymmetric galactic structure on the local stellar kinematics are strong .
Salient Texts - GenCompareSum (t5-med-query)	• what is the effect of dark matter kinematics on the bar - and spiral arm - indu ced phase space structure? • what is the main argument of @xcite? • what is the structure of the hercules? • what is the solar neighborhood? • what is the kinematic distribution of the particles? • what is the relationship between spiral arms and stellar behavior? • what is the galactic potential? • what is the required condition for a thick disk? • what is the difference between ic3 and ic2? • why is the observed velocity field a useful parameter for predicting the beha vior of galaxies?
Predicted Summary - GenCompareSum (t5-med-query)	• however , it is unclear whether there is any dependence of the induced local solar neighborhood kinematics on the detailed galactic structure . • moreover , the initial conditions hardly consider the evolution of the mw . • the induced kinematic distribution at the end of the simulation is studied by c onsidering the particles inside a circle of radius @xmath7 centered at the sol ar position . • therefore we focused on the recently induced kinematic structure in the solar neighborhood . • we conclude that the contribution of the spiral arms to the solar neighborhoo d kinematics may be comparable to that of the bar . • in our simulations the positions of these kinematic arches are modified when the bar is added to the model . • another unexpected aspect of the bar - and spiral arm - induced phase space s tructure is the effect on the local dark matter kinematics . • the spiral arm contribution to the resonant structure in the solar neighborhoo d may be comparable to that of the galactic bar . • the main differences to previous studies are the arm force contrast and force field shape? • in summary , the imprints of the non - axisymmetric galactic structure on the local stellar kinematics are strong .
Salient Texts - GenCompareSum (t5-s2orc-title)	• dark matter kinematics in the solar neighborhood • a note on the arcturus structure in a $xmath26$ plane • dark kinematic groups in the dark disk • the spiral arm contribution to the resonant structure of the solar neighborhoo d • the birth of stars in the disk with small velocity dispersion • the solar neighborhood kinematics and the spiral arms • spiral arms in the mw-type galaxies • the hercules branch of a galactic model using only a bar • theoretical study of the bar and spiral arm perturbations in the xci model • dark matter currents in the galactic dark disk
Predicted Summary - GenCompareSum (t5 -s2orc-title)	• in @xcite we presented a study of the solar neighborhood kinematic groups using a sample of 24,190 stars . • lastly , we investigate effects on the local dark matter kinematics , in particul ar in the disk - like dark matter structure recently predicted by lcdm models . • the induced kinematic distribution at the end of the simulation is studied by c onsidering the particles inside a circle of radius @xmath7 centered at the sol ar position . • therefore we focused on the recently induced kinematic structure in the solar neighborhood .

	• we conclude that the contribution of the spiral arms to the solar neighborhoo d kinematics may be comparable to that of the bar . • another unexpected aspect of the bar - and spiral arm - induced phase space s tructure is the effect on the local dark matter kinematics . • our results show that these models generate dark matter currents inside the g alactic dark disk . • the spiral arm contribution to the resonant structure in the solar neighborhoo d may be comparable to that of the galactic bar . • we show that the galactic non - axisymmetric potential develops dark kinema tic groups in the dark disk predicted in cosmological simulations of galaxy f ormation . • in summary , the imprints of the non - axisymmetric galactic structure on the local stellar kinematics are strong .