# BioNLI: Generating a Biomedical NLI Dataset Using Lexico-semantic Constraints for Adversarial Examples

**Mohaddeseh Bastan**
Stony Brook University
mbastan@cs.stonybrook.edu

**Mihai Surdeanu**
University of Arizona
msurdeanu@email.arizona.edu

**Niranjan Balasubramanian**
Stony Brook University
niranjan@cs.stonybrook.edu

## Abstract

Natural language inference (NLI) is critical for complex decision-making in biomedical domain. One key question, for example, is whether a given biomedical mechanism is supported by experimental evidence. This can be seen as an NLI problem but there are no directly usable datasets to address this. The main challenge is that manually creating informative negative examples for this task is difficult and expensive. We introduce a novel semi-supervised procedure that bootstraps an NLI dataset from existing biomedical dataset that pairs mechanisms with experimental evidence in abstracts. We generate a range of negative examples using nine strategies that manipulate the structure of the underlying mechanisms both with rules, e.g., flip the roles of the entities in the interaction, and, more importantly, as perturbations via logical constraints in a neuro-logical decoding system (Lu et al., 2021b).

We use this procedure to create a novel dataset for NLI in the biomedical domain, called BioNLI and benchmark two state-of-the-art biomedical classifiers. The best result we obtain is around mid 70s in F1, suggesting the difficulty of the task. Critically, the performance on the different classes of negative examples varies widely, from 97% F1 on the simple role change negative examples, to barely better than chance on the negative examples generated using neuro-logic decoding.[1]

| |
|---|
| **Premise:**The outflow of uracil from the yeast Saccharomyces cerevisiae is known to be relatively fast in certain circumstances, to be retarded by proton conductors and to occur in strains lacking a uracil proton symport. In the present work, it was shown that uracil exit from washed yeast cells is an active process, creating a uracil gradient of the order of -80 mV relative to the surrounding medium. Glucose accelerated uracil exit, while retarding its entry. DNP or sodium azide each lowered the gradient to about -30 mV, simultaneously increasing the rate of uracil entry. They also lowered cellular ATP content. Manipulation of the external ionic conditions governing delta mu H+ at the plasma membrane had no detectable effect on uracil transport in yeast preparations thoroughly depleted of ATP. |
| **Consistent Hypothesis:**It was concluded that <re> uracil <er> exit is probably not driven by the s <el> proton <le> gradient but may utilize ATP directly. |
| **Adversarial Hypothesis:**It is concluded that <el> uracil <le> exit from S. cerevisiae is an active process facilitated by a <re> proton <er> gradient and ATP. |

Table 1: Example of a premise/hypothesis pair in the BioNLI dataset, as well as of an adversarial hypothesis that was automatically generated by an encoder-decoder network that manipulated the lexico-semantic constraints in the original hypothesis. Here the regulator entity is marked as <re> entity <er>, and the regulated entity is marked as <el> entity <le>.

## 1 Introduction

Biomedical research has progressed at a tremendous pace, to the point where PubMed[2] has indexed well over 1M publications per year in the past eight years. Many of these publications include high-level mechanistic knowledge, e.g., protein-signaling pathways, which is critical for the understanding of many diseases (Valenzuela-Escarcega et al., 2018), but which must be supported by lower-level experimental evidence to be trustworthy. Developing models that can understand and reason about such mechanisms is crucial for supporting effective access to the rich biomedical knowledge (Bastan et al., 2022). In particular, the current information deluge motivates the need for developing tools that can answer the question: "Is a given mechanism supported by experimental evidence?". This can be seen as a biomedical natural language inference (NLI) problem. Despite the prevalence of many biomedical NLP datasets (Demner-Fushman et al., 2020; Bastan et al., 2022; Krallinger et al., 2017), there are no datasets that can be directly used to address this task.

However, manually creating a biomedical NLI

[1]Code and data is available at https://github.com/StonyBrookNLP/BioNLI

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



*Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5122 - 5133
December 7-11, 2022 

dataset that focuses on mechanistic information is challenging. Table 1, which contains an actual example from our proposed dataset, highlights several difficulties. First, understanding biomedical mechanisms and the necessary experimental evidence that supports (or does not support) them requires tremendous expertise and effort (Kaushik et al., 2019). For example, the premise shown is considerably larger than the average premise in other open-domain NLI datasets such as SNLI (Bowman et al., 2015), and is packed with domain-specific information. Second, negative examples are seldom explicit in publications. Creating them manually risks introducing biases, simplistic information, and systematic omissions (Wu et al., 2021).

In this work, we introduce a novel semi-supervised procedure for the creation of biomedical NLI datasets that include mechanistic information. Our key contribution is automating the creation of negative examples that are informative without being simplistic. Intuitively, we achieve this by defining lexico-semantic constraints based on the mechanism structures in the biomedical literature abstracts. Our dataset creation is as follows:

**(1)** We extract positive entailment examples consisting of a premise and hypothesis from abstracts of PubMed publications. We focus on abstracts that contain an explicit conclusion sentence, which describes a biomedical interaction between two entities (a regulator and a regulated protein or chemical). This yields premises that are considerably larger than premises in other open-domain NLI datasets: between 3 – 15 sentences.

**(2)** We generate a wide range of negative examples by manipulating the structure of the underlying mechanisms both with rules, e.g., flip the roles of the entities in the interaction, and, more importantly, by imposing the perturbed conditions as logical constraints in a neuro-logical decoding system (Lu et al., 2021b). This battery of strategies produces a variety of negative examples, which range in difficulty, and, thus, provide an important framework for the evaluation of NLI methods.

We employ this procedure to create a new dataset for natural language inference (NLI) in the biomedical domain, called BioNLI. Table 1 shows an actual example from BioNLI. The dataset contains 13489 positive entailment examples, and 26907 adversarial negative examples generated using nine different strategies. An evaluation of a sample of these negative examples by human biomedical experts indicated that $86\%$ of these examples are indeed true negatives. We trained two state-of-the-art neural NLI classifiers on this dataset, and show that the overall F1 score remains relatively low, in the mid 70s, which indicates that this NLI task remains to be solved. Critically, we observe that the performance on the different classes of negative examples varies widely, from $97\%$ accuracy on the simple negative examples that change the role of the entities in the hypothesis, to $55\%$ (i.e., barely better than chance) on the negative examples generated using neuro-logic decoding. Further, given how the dataset is constructed we can also test if models produce consistent decisions on all adversarial negatives associated with a mechanism, giving deeper insight into model behavior. Thus, in addition of its importance in the biomedical field, we hope that this dataset will serve as a benchmark to test models' language understanding abilities.

## 2 Related Work

Previous work on NLI in scientific domains include: medical question answering (Abacha and Demner-Fushman, 2016), entailment based text exploration in health care (Adler et al., 2012), entailment recognition in medical texts (Abacha et al., 2015), textual inference in clinical trials (Shivade et al., 2015), NLI on medical history (Romanov and Shivade, 2018), and SciTail (Khot et al., 2018) which is created from multiple-choice science exams and web sentences. These datasets either have modest sizes (Abacha et al., 2015), target specific NLP problems such as coreference resolution or named entity extraction (Shivade et al., 2015), and make use of experts in the domain to generate inconsistent data which is costly and labor-intensive. Additionally, they also focus on sentence-to-sentence entailment tasks, where both the premise and the hypothesis are no longer than one sentence. Most importantly, none of these are directly aimed at inference on mechanisms in biomedical literature.

Our work is also related to NLI tasks that go beyond sentence-level entailments. For example, (Yin et al., 2021) include premises longer than a sentence, but only use three simple rule-based methods to create negative samples. (Yan et al., 2021; Nie et al., 2019) use larger contexts as premises for the NLI task but only on general purpose domains like news, fiction, and Wiki. On the other hand, the BioNLI dataset is an inference problem with large contexts as premises but in the biomedical domain

which often requires handling more complex texts and domain knowledge.

There is also a growing body of research into exploring factual inconsistency in text generation models (Maynez et al., 2020; Zhu et al., 2021; Utama et al., 2022). We take advantage of the known weakness of generation models for hallucination and also employ a constraint based neurological decoding from recently introduced decoding methods (Lu et al., 2021b,a; Kumar et al., 2021) to generate adversarial examples for BioNLI dataset.

## 3 BioNLI Creation

We model the task of understanding if a high-level mechanistic statement is supported by lower-level experimental evidence as natural language inference (NLI). The goal of NLI is to understand whether the given *hypothesis* can be entailed from the *premise* or not (Dagan et al., 2005). This is typically modeled with three labels (entailed or not, plus a neutral class if the two texts are unrelated). In our case, the premise contains the experimental evidence, while the hypothesis summarizes the higher-level mechanistic information. Both of these texts are extracted from abstracts of biomedical publications, where the beginning sentences (the *supporting set*) describe experimental evidence, and a *conclusion* sentence summarizes the mechanistic information that is entailed by these experiments.

In this work, we introduce the BioNLI dataset, an NLI dataset automatically created from a set of abstracts of PubMed open-access publications. We collected all the abstracts which contain a conclusion sentence with mechanistic information at the end of the abstract, and filter out the rest. Following previous work in mechanism generation (Bastan et al., 2022), we focus on conclusion sentences that discuss binary biochemical interactions between a regulator and a regulated entity (both of which are proteins or chemicals). We then generate negative examples by manipulating the structure of the conclusion sentences.

In the following subsections we describe in detail the generation of both positive and negative examples in BioNLI.

### 3.1 Identifying Abstracts with Mechanistic Information

To identify abstracts that contain conclusion sentences with such binary biochemical interactions, we followed the same procedure and dataset [3] as (Bastan et al., 2022) . That is, we used a series of patterns (e.g., finding words that start with *conclud* all patterns are described in Appendix A) to identify conclusion sentences at the end of abstracts, and consider the previous ones as the supporting set. We analyzed the SuMe dataset and found that 91% of the abstracts end with conclusion sentences, which indicates that the filtering heuristic is robust.

Further, we take advantage of the structured text in the biomedical domain, by focusing on abstracts that describe some mechanism between two biochemical entities. One of the main entities is called *regulator entity* and is marked with <*re*> *entity* <*re*> inside the text; the other main entity is called *regulated entity* and is marked with <*el*> *entity* <*le*> inside the text. We will use this structure to generate negative examples by modifying it.

### 3.2 Positive Instances

For positive examples, we simply use the original conclusion sentence from the abstract as the hypothesis and the supporting set as the premise. These sentences are likely to be accurate as they are written by domain experts, and also peer-reviewed by other scientists.

### 3.3 Adversarial Instances

The key contribution of this paper is on the automatic creation of meaningful, yet difficult negative examples without the use of experts. We introduce multiple strategies for creating negative examples. We group these strategies into two groups: *rule-based* and *neural-based counterfactuals*, both of which are detailed below. We show examples of these strategies in Table 4.

#### 3.3.1 Rule-Based Counterfactuals

This category consists of rule-based methods that convert a correct conclusion sentence (i.e., the hypothesis) into an instance that is not entailed by the given supporting set by perturbing parts of its semantic structure. Most of them are used in general-domain factual consistency evaluating systems (Kryściński et al., 2019; Zhu et al., 2020):

**Swap Entity Names (SEN)**: Swapping the entity names. This flips the roles of the entities in the interaction, i.e., the regulator becomes the regu-

[3] This paper works at a higher level of abstraction, which contains causal semantic relations (or "activations"), which are always directed and not necessarily asymmetric.

lated and vice versa, which contradicts the original evidence.

**Swap Entity Positions (SEP)**: In this perturbation we swap the positions of the two entities in text.

**Swap Random Entity (SRE)**: In this perturbation one of the main entities is randomly swapped with a different entity that occurs in the supporting set and has the same entity type. We use SciSpaCy (Neumann et al., 2019) and the built in *en_ner_bionlp13cg_md* model to detect entity types.

**Swap Random Entity with Out of text entity (SREO)**: In this perturbation we swap one of the two entities in the interaction with a random entity from out of the context which was not available in the supporting set but has the same type as the main entity. Similarly, we use SciSpaCy with the same model to detect entity types.

**Verb Negation (VNeg)**: We randomly select one of the predicates in the original conclusion and change its polarity, e.g., from positive to negative or vice versa.

**Swap Numbers (SN)**: If the conclusion contains a number, it is swapped with a different number, randomly chosen from the supporting set.

**Lexical Polarity Reversal (LPR)**: We collected a list of terms describing mechanistic interactions (e.g., *inhibition* and *promotion*), and swapped them with their antonyms when encountered in the hypothesis.

#### 3.3.2 Neural-based Counterfactuals

The above methods are relatively simple perturbations, which might be easily detected by transformer-based classifiers. To counteract this potential limitation, we take advantage of transformer-based generation methods to create more complex and diverse set of negative examples. In this category we have two main approaches:

**Mechanism Generation (GEN)**: We use a model pretrained on a mechanism generation task in the same context (Bastan et al., 2022) to generate mechanism sentences (and the relation between main entities) for each abstract in our dataset. As our dataset overlaps with the one from (Bastan et al., 2022), we implemented a 5-fold cross validation, and retrained the model with the corresponding training set in each fold. That is, for each split, we train with 4 folds and generate the output for the other fold. The generated texts which get a BLEURT score lower than $\lambda$, and predict the relation between two main entities incorrectly are selected as counterfactuals. Here, we set $\lambda = 0.45$.

**Neurologic Decoding (GEN-ND)**: Neurologic decoding is a decoding algorithm that enables neural language models to generate text while satisfying complex lexical constraints (Lu et al., 2021b). We take advantage of this decoding method to impose different structure-aware constraints. For generation, we use the same model as the one described in GEN approach. For decoding this model, we define the following constraints which result in generating negative examples. (we combine all three categories in our results table, naming the entire group *GEN-ND*):

**(1)** Neurologic Decoding with SEN Constraints (GEN-ND-SEN): We imposed as positive constraints (i.e., constraints that should be satisfied during decoding) that the two entities be present in the output, but we swapped their names. That is, the regulator and regulated entities are swapped; if both of them are satisfied in the generated text, the instance is used as a negative example. For example if the original conclusion has the following pattern:

$$... < re > entity1 < er > ... < el > entity2 < le >$$

We add the following constraints to the neurological decoding:

$$[[< re > entity2 < er >], [< el > entity1 < le >]]$$

Compared with the general SEN introduced in section 3.3.1, by using these constraints we force the generation model to generate a natural yet negative and completely new sentence.

**(2)** Neurologic Decoding with SRE Constraints (GEN-ND-SRE): Similarly, we swapped one of the main entities in the positive constraints with a random entity from the supporting text. To make sure the generated sentence is not too similar to the original conclusion sentence, we used the generated sentence only if both constraints are satisfied and the semantic similarity between the text and the original text is smaller than $\delta$. To compute the semantic similarity we use BioLinkBERT (Yasunaga et al., 2022). We set $\delta$ to 0.9.

**(3)** Neurologic Decoding with Negative Constraints (GEN-ND-NG): The third and last method we tried uses negative constraints, i.e., its lexical artifacts should *not* be present in the generated text. In

| | Dataset | Train | Dev | Test | Sum |
|---|---|---|---|---|---|
| | + | 8489 | 3000 | 2000 | 13489 |
| - | SEN | 2064 | 3000 | 2000 | 7064 |
| | SEP | 2022 | 3000 | 2000 | 7022 |
| | SRE | 81 | 20 | 15 | 116 |
| | SREO | 1466 | 2584 | 1524 | 5574 |
| | VNeg | 837 | 1395 | 810 | 3042 |
| | SN | 615 | 543 | 314 | 1472 |
| | LPR | 711 | 623 | 340 | 1674 |
| | GEN-ND | 547 | 141 | 102 | 790 |
| | GEN | 165 | 30 | 21 | 216 |
| | Total | 8508 | 11336 | 7126 | 26970 |
| | Unique | 8508 | 3000 | 2000 | 13508 |
| | Total | 16997 | 14336 | 9126 | 40459 |
| | Unique | 16997 | 3000 | 2000 | 21997 |

Table 2: Dataset statistics of the larger distribution. Each instance is perturbed as many times as possible for the dev and test sets and once for the training set.

| | Dataset | Train | Dev |
|---|---|---|---|
| | + | 2790 | 2453 |
| - | SEN | 413 | 2453 |
| | SEP | 452 | 2453 |
| | SRE | 80 | 22 |
| | SREO | 335 | 2058 |
| | VNeg | 159 | 1214 |
| | SN | 256 | 625 |
| | LPR | 311 | 728 |
| | GEN-ND | 586 | 84 |
| | GEN | 162 | 33 |
| | Total | 2754 | 9670 |
| | Total Unique | 2754 | 2453 |
| | Total | 5544 | 12123 |
| | Total Unique | 5544 | 2453 |

Table 3: Dataset statistics of the balanced distribution. We sampled over perturbed classes to create a balance dataset so that no rule-based category have more than 500 instances in the train set. Test set is same as Table 2

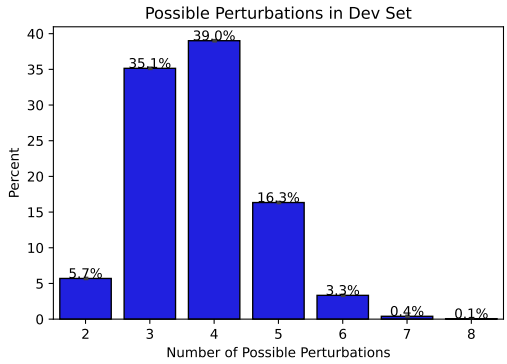


Figure 1: Distribution of possible perturbations over the dev set

particular, the negative constraints we defined contained the original entities. By using the original entities as negative constraints the generated text receives a higher score if the main entities are not shown in their own roles (i.e., neither regulator nor regulated entities are not enclosed with specific markers).

### 3.4 Dataset Statistics

The resulting dataset is summarized in two tables. Table 2 shows the maximum number of possible perturbations on each instance. For example, all instances can be perturbed with SEP and SEN approaches, while only the ones that have a number in both conclusion and supporting set can be perturbed with the SN approach.

The distribution of the possible perturbations over the dev set is shown in Figure 1. As we see in this figure, all sentences can get at least two perturbation SEN and SEP. $5.7\%$ of the instances can not get any other perturbations based on their structure. $35.1\%$ of the data get 3 different adversarial examples. Most of the data, which is about $39\%$ of them, can be perturbed with 4 different approaches. The lowest category is 8 perturbations which are only about $0.1\%$ of the data. We don't have any instance which can be perturbed with all 9 possible methods explained in section 3.3, which shows the diversity and variety on the adversarial instance generation approaches.

Note that while our goal is to produce a dataset with as many high quality examples as possible for each category, downstream applications can adjust the distribution of the training categories to be uniform or biased towards specific categories as needed based on their requirements. To study the impact of a balanced distribution of adversarial examples, we sampled the positive and negative classes to produce a balanced dataset. Table 3 shows the distribution of the adversarial categories in this balanced dataset. We evaluate with the original collection as well as with this balanced dataset (see Table 5).

Table 4 shows a set of rule-based and neural-based adversarial examples. The main entities are enclosed with specific markers and are swapped with different methods. We also see a completely new

and negative generated text based on the supporting set with the generation approach (GEN-SEN).

### 3.5 Quality Control

To ensure the quality of the collected dataset, we asked two experts (graduate students in the biomedical domain) to inspect 50 randomly selected adversarial examples generated by the neural-based counterfactual methods (section 3.3.2). We sampled 25 examples each from GEN and GEN-ND methods. Given the abstract and the generated sentence, the experts assessed whether the generated sentence is indeed inconsistent with the given supporting set, meaning that the sentence cannot be concluded given the supporting set. The expert analysis shows that the neural-based counterfactuals are of high quality. They find that $88\%$ of adversarial examples from the GEN method and $84\%$ from the GEN-ND method are correct negative examples, averaging to $86\%$ overall.

## 4 Evaluation

In this section we benchmark the performance of state-of-the-art (SOTA) biomedical language models on the BioNLI dataset. Our evaluation is aimed at assessing the following aspects of the BioNLI:

1. How difficult is the inference task captured by the dataset?
2. What kinds of perturbations are difficult for the models?
3. How consistent are the models on adversarial instances?

### 4.1 Implementation Details

We fine-tune two state-of-the-art models in biomedical domain:
(i) PubMedBERT (Gu et al., 2020) was pretrained from scratch with texts from the biomedical domain and shown to be effective for a wide variety of biomedical NLP tasks including NER, QA, and sentence similarity.
(ii) BioLinkBERT (Yasunaga et al., 2022) augments PubMedBERT by pre-training jointly on linked biomedical articles.
We fine-tune the top 3 layers of the base-sized pretrained models from Hugging Face (Wolf et al., 2019) using PyTorch (Paszke et al., 2019). We use AdamW (Loshchilov and Hutter, 2017) with a learning rate of $1e-4$ by manually tuning 5 different values. We use the original hyper-parameters of the models and the NLI label prediction is done via binary classification using `CLS` token. The sequence length we use here is 512 and the beam size is 16. We train each model for 20 epochs and choose the best one based on the performance (macro F1) over the dev set.

The performance of both these models is listed in Table 5. The table reports positive, negative, and overall F1 scores, as well performance for the various types of negative examples (recall). At the time of prediction, we use a binary classifier. Hence, we don't have a fine-grained negative category predictions, therefore, we can't calculate precision. Instead, we report recall for fine-grained negative categories and F1 score for positive and overall negative prediction. We also report macro-F1 for positive and negative classes. In addition, the table includes an ablation experiment, where the performance of the classifiers trained only on the hypothesis ("hypo-only") is contrasted with the classifier trained on the entire data ("premise+hyp"). We also include the performance of the balanced distribution in parentheses, to compare with the overall distribution.

### 4.2 Overall Difficulty

Table 5 indicates that the overall performance of the best model on the positive class is 77%, and 79% on all negative examples (macro-average). If we only consider the difficult negative classes (SRE, LPR, GEN-ND, GEN), the best model's performance on the negative categories drop considerably to 55.4%, i.e., only slightly better than a random classifier.

This table also highlights the difficulty of the generated categories. While traditional approaches of the adversarial example creation, (i.e., SEN and SEP), are solvable with large transformer-based models, the more complex negative examples produced using generation are considerably more difficult to be classified correctly.

Table 5 calls attention to another feature of the BioNLI dataset: classifiers trained with the balanced training data (shown in parentheticals) perform better on minority categories, while the models produced after training on the larger distribution perform better on other categories. This highlights the versatility of the dataset, as well as the importance of customizing the data distribution (including that of negative examples!) for each use case.

### 4.3 Difficulty of Adversarial Instances

Table 5 indicates that some categories of the adversarial examples are more difficult than the rest.

| | |
|---|---|
| **Abstract:** | We investigated whether intracellular pH (pH(i)) is a causal mediator in abscisic acid (ABA)-induced gene expression. We measured the change in pH(i) by a "null-point" method during stimulation of barley (Hordeum vulgare cv Himalaya) aleurone protoplasts with ABA and found that ABA induces an increase in pH(i) from 7.11 to 7.30 within 45 min after stimulation. This increase is inhibited by plasma membrane H(+)-ATPase inhibitors, which induce a decrease in pH(i), both in the presence and absence of ABA. This ABA-induced pH(i) increase precedes the expression of RAB-16 mRNA, as measured by northern analysis. ABA-induced pH(i) changes can be bypassed or clamped by addition of either the weak acids 5,5-dimethyl-2,4-oxazolidinedione and propionic acid, which decrease the pH(i), or the weak bases methylamine and ammonia, which increase the pH(i). Artificial pH(i) increases or decreases induced by weak bases or weak acids, respectively, do not induce RAB-16 mRNA expression. Clamping of the pH(i) at a high value with methylamine or ammonia treatment affected the ABA-induced increase of RAB-16 mRNA only slightly. However, inhibition of the ABA-induced pH(i) increase with weak acid or proton pump inhibitor treatments strongly inhibited the ABA-induced RAB-16 mRNA expression. |
| **Conclusion:** | We conclude that, although the <el> ABA <le>-induced the <re> pH <er>(i) increase is correlated with and even precedes the induction of RAB-16 mRNA expression and is an essential component of the transduction pathway leading from the hormone to gene expression, it is not sufficient to cause such expression. |
| SEN | We conclude that, although the <el> pH <le>-induced the <re> ABA <er>(i) increase is correlated with and even precedes the induction of RAB-16 mRNA expression and is an essential component of the transduction pathway leading from the hormone to gene expression, it is not sufficient to cause such expression. |
| SEP | We conclude that, although the <re> pH <er>-induced the <el> ABA <le>(i) increase is correlated with and even precedes the induction of RAB-16 mRNA expression and is an essential component of the transduction pathway leading from the hormone to gene expression, it is not sufficient to cause such expression. |
| SREO | We conclude that, although the <el> integrin <le>-induced the <re> pH <er>(i) increase is correlated with and even precedes the induction of RAB-16 mRNA expression and is an essential component of the transduction pathway leading from the hormone to gene expression, it is not sufficient to cause such expression. |
| VNeg | We conclude that, although the <el> ABA <le>-induced the <re> pH <er>(i) increase is not correlated with and even precedes the induction of RAB-16 mRNA expression and is an essential component of the transduction pathway leading from the hormone to gene expression, it is not sufficient to cause such expression. |
| LPR | We conclude that, although the <el> ABA <le>-induced the <re> pH <er>(i) decrease is correlated with and even precedes the induction of RAB-16 mRNA expression and is an essential component of the transduction pathway leading from the hormone to gene expression, it is not sufficient to cause such expression. |
| Generation | We conclude that the <re> ABA <er> -induced increase in <el> pH <le> (i) precedes the expression of RAB-16 mRNA. |

Table 4: Example of the generated adversarial instance for the BioNLI dataset using lexico semantic constraints. Regulator entities are enclosed in *<re> <er>* tags and regulated entities are enclosed in *<el> <le>* tags. The red texts show the negated phrases.

This is mostly seen in rule-based categories. In average, the neural-based methods generate more difficult sentences than the rule-based methods.

For instance, SEP and SEN instances are easier due to the structure (markers) in the dataset. Even without inspecting the supporting set, the model learns that the entity with *<re><er>* markers should be the subject of the text while the entity marked with *<el><le>* should be the object of the mechanism.

Some categories are easier to recognize with context. For instance SREO and SN approaches are not easily detectable with hypothesis only baselines. But, when the model is trained with both premise and hypothesis, these become easier because the contradiction can be recognized using information in the premise (i.e. abstract).

We have four difficult categories of adversarial examples (SRE, LPR, GEN-ND, and GEN) where the models perform only slightly better than a random classifier. These are the least-frequent classes; when we train the model under the balanced distribution the performance improves, somewhat. However, performance remains low, which underscores the need for further research on handling these difficult adversarial examples.

### 4.4 Model Consistency on Adversarial Instances

In addition to the per-perturbation evaluation, we also merged all available positive and negative instance for each entry of the dataset, and computed what percentage of them are classified correctly. The cumulative results are shown in Figure 2. While models are able to get reasonable accuracy overall, they are not consistent in their decisions. There are no cases, where both the positive instance and all of its associated negative instances are all classified correctly. There are only 30% of the cases where at least 70% of the perturbations

| | | | Full Distribution | | | | Balanced Distribution | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Model | | | PubMedBERT | | BioLinkBERT | | PubMedBERT | | BioLinkBERT | |
| Class | | | h-only | p+h | h-only | p+h | h-only | p+h | h-only | p+h |
| Positive | | | 0.65 | 0.76 | 0.68 | 0.77 | 0.57 | 0.69 | 0.60 | 0.69 |
| Negative | Rule-based | SEN | 0.90 | 0.96 | 0.91 | 0.97 | 0.88 | 0.92 | 0.93 | 0.96 |
| | | SEP | 0.92 | 0.97 | 0.93 | 0.98 | 0.89 | 0.92 | 0.93 | 0.96 |
| | | SRE | 0.33 | 0.50 | 0.33 | 0.50 | 0.50 | 0.67 | 0.33 | 0.67 |
| | | SREO | 0.69 | 0.98 | 0.64 | 0.99 | 0.60 | 0.95 | 0.69 | 0.97 |
| | | VNeg | 0.81 | 0.91 | 0.82 | 0.86 | 0.79 | 0.84 | 0.78 | 0.83 |
| | | SN | 0.64 | 0.82 | 0.54 | 0.81 | 0.52 | 0.78 | 0.59 | 0.82 |
| | | LPR | 0.56 | 0.56 | 0.48 | 0.59 | 0.50 | 0.50 | 0.48 | 0.49 |
| | | Macro-avg | 0.69 | 0.81 | 0.67 | 0.82 | 0.67 | 0.80 | 0.68 | 0.81 |
| | NN-based | GEN-ND | 0.47 | 0.6 | 0.53 | 0.56 | 0.57 | 0.66 | 0.64 | 0.68 |
| | | GEN | 0.33 | 0.57 | 0.33 | 0.57 | 0.57 | 0.68 | 0.33 | 0.68 |
| | | Macro-avg | 0.40 | 0.58 | 0.43 | 0.56 | 0.57 | 0.67 | 0.49 | 0.68 |
| | All negatives | | 0.63 | 0.76 | 0.61 | 0.76 | 0.65 | 0.77 | 0.63 | 0.79 |
| Macro-avg | | | 0.64 | 0.76 | 0.65 | 0.77 | 0.61 | 0.73 | 0.62 | 0.74 |

Table 5: Overall performance of two state-of-the-art models in the biomedical domain (PubMedBERT, BioLinkBERT) on both distributions. The models are fine-tuned using the data with premise (p+h) and without premise (h-only) on the BioNLI dataset. The metric used here is recall for fine-grained negative classes and F1 for positive and all negative categories. The different rows indicate the performance for the various kinds of positive and negative examples.

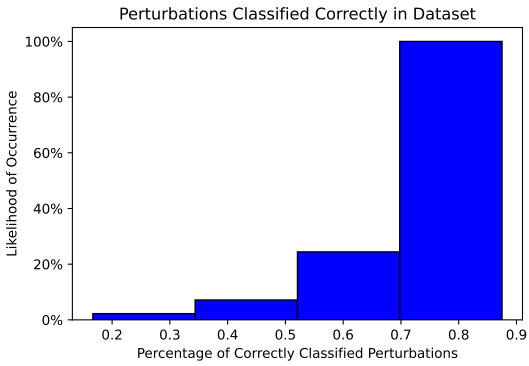


Figure 2: Percentage of correctly classified perturbations in the dev set.

| Error Category | Frequency % |
|---|---|
| Multiple pieces of information | 10 |
| Abbreviation | 10 |
| Unrelated information | 16 |
| Mechanism mix up | 18 |
| Noun phrase mix up | 20 |
| Entity Similarity | 26 |

Table 6: Distribution of different error categories in 50 incorrect classified samples

derived from the same positive instance are classified correctly. This further indicates the brittleness (or lack of robust reasoning) in current models suggesting avenues for further research.

### 4.5 Error Analysis

We analyzed a set of 50 instances within the BioNLI dataset which are classified incorrectly. Table 6 outlines distribution of the errors made by BioLinkBERT model in different categories. In 10% of the misclassified instances, the mechanism behind the entities which is explained in the conclusion sentence needs multi hop reasoning which makes it difficult to classify correctly. Abbreviations can cause difficulties, when one of the main entities is mentioned in its full form in one sentence but abbreviated in the rest of the sentences. Distant entities also make inference harder. The model fails when the two main entities are in different sen-

tences with many unrelated fragments of text occur between them. Sometimes the abstracts can talk about two related experiments on the same entities, but the mechanism is only related to one of the experiment making it harder to assess entailment. A '-' (dash) between two words, can change the subject and object of a sentence. The entailment is very difficult when there is a subtle difference here. Finally, the similarity between entity names, or name overlaps, specifically in case of SRE and GEN-ND-SRE confuses the model about the correct entailment class. There are cases where the entity is swapped with another entity with similar name or with partially overlapped name, these cases seem to be difficult for the model to classify correctly.

## 5 Conclusion

In this paper, we introduced a novel semi-supervised procedure for the creation of biomedical NLI datasets that include mechanistic information. Our key contribution is automating the creation of negative examples that are informative without being simplistic. We achieve this by manipulating the lexico-semantic constraints in the mechanism structures captured in the hypotheses, which we implement both with rules and with neuro-logic decoding. To our knowledge, this is the first paper that employs neuro-logic decoding for the generation of adversarial examples. All in all, we implemented nine different strategies for the creation of adversarial examples.

We used this procedure to create the BioNLI dataset, which addresses NLI for mechanistic texts in the biomedical domain. An evaluation of a sample of these negative examples by human biomedical experts indicated that $86\%$ of these examples are indeed true negatives. We trained two state-of-the-art neural NLI classifiers on this dataset, and showed that the overall performance remains relatively low, which indicates that this NLI task is not solved. Critically, we observe that the performance on the different classes of negative examples varies widely, from $97\%$ accuracy on the simple negative examples that change the role of the entities in the hypothesis, to $55\%$ (i.e., barely better than chance) on the negative examples generated using neuro-logic decoding. We hope that this open-access dataset[4] will enable further research both on biomedical NLI, and on language understanding in general.

## 6 Acknowledgement

This material is based on research that is supported in part by the National Science Foundation under the award IIS #2007290. The authors would like to thank the anonymous reviewers and the area chair for their feedback on this work. We would also like to thank the biomedical experts who assessed the quality of the adversarial examples.

## 7 Limitations

Unlike many scientific NLI datasets (Romanov and Shivade, 2018; Shivade et al., 2015) no instance in the BioNLI dataset was directly annotated by human domain experts. Instead, following the trend of machine-generated datasets (Hartvigsen et al., 2022), we build upon recent developments in text generation and generate BioNLI automatically.

The only human annotation in this effort was performed by one expert on a sample of 50 sentences, to check the quality of automatically created negative examples. This minimal effort was justified by previous work in the biomedical space (citation hidden for review), in which we observed that experts had high inter-annotator agreement on the interpretations of scientific information in abstracts.

The premise-free experiments show the presence of artifacts in some categories of the BioNLI dataset, similar to several other NLI datasets (Romanov and Shivade, 2018; Bowman et al., 2015; Nangia et al., 2017). Addressing these artifacts remains an open research issue.

## 8 Ethical Considerations

Our data is collected solely from open-access publications in PubMed. We do not include any meta data (authors, publication venue, etc.) in the dataset. The created dataset is also open-access.

We believe our released dataset and software will contribute to society by promoting further NLI research and applications in the biomedical domain. Long term, we envision that this research will enable novel machine reading applications that automatically discover potential explanations and treatments for diseases that are still misunderstood today.

[4]Code and data is available at https://github.com/StonyBrookNLP/BioNLI

## References


Asma Ben Abacha and Dina Demner-Fushman. 2016. Recognizing question entailment for medical question answering. In *AMIA Annual Symposium Proceedings*, volume 2016, page 310. American Medical Informatics Association.

Asma Ben Abacha, Duy Dinh, and Yassine Mrabet. 2015. Semantic analysis and automatic corpus construction for entailment recognition in medical texts. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 238–242. Springer.

Meni Adler, Jonathan Berant, and Ido Dagan. 2012. Entailment-based text exploration with application to the health-care domain. In *Proceedings of the ACL 2012 System Demonstrations*, pages 79–84.

Mohaddeseh Bastan, Nishant Shankar, Mihai Surdeanu, and Niranjan Balasubramanian. 2022. Sume: A dataset towards summarizing biomedical mechanisms. *arXiv preprint arXiv:2205.04652*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*, pages 177–190. Springer.

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

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

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509*.

Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. 2019. Learning the difference that makes a difference with counterfactually-augmented data. *arXiv preprint arXiv:1909.12434*.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. Scitail: A textual entailment dataset from science question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

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

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Evaluating the factual consistency of abstractive text summarization. *arXiv preprint arXiv:1910.12840*.

Sachin Kumar, Eric Malmi, Aliaksei Severyn, and Yulia Tsvetkov. 2021. Controlled text generation as continuous optimization with multiple constraints. *Advances in Neural Information Processing Systems*, 34:14542–14554.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Ximing Lu, Sean Welleck, Peter West, Liwei Jiang, Jungo Kasai, Daniel Khashabi, Ronan Le Bras, Lianhui Qin, Youngjae Yu, Rowan Zellers, et al. 2021a. Neurologic a* esque decoding: Constrained text generation with lookahead heuristics. *arXiv preprint arXiv:2112.08726*.

Ximing Lu, Peter West, Rowan Zellers, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021b. NeuroLogic decoding: (un)supervised neural text generation with predicate logic constraints. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4288–4299, Online. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Nikita Nangia, Adina Williams, Angeliki Lazaridou, and Samuel R Bowman. 2017. The repeval 2017 shared task: Multi-genre natural language inference with sentence representations. *arXiv preprint arXiv:1707.08172*.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019. Adversarial nli: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style,

high-performance deep learning library. *Advances in neural information processing systems*, 32.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. *arXiv preprint arXiv:1808.06752*.

Chaitanya Shivade, Courtney Hebert, Marcelo Lopetegui, Marie-Catherine De Marneffe, Eric Fosler-Lussier, and Albert M Lai. 2015. Textual inference for eligibility criteria resolution in clinical trials. *Journal of biomedical informatics*, 58:S211–S218.

Prasetya Ajie Utama, Joshua Bambrick, Nafise Sadat Moosavi, and Iryna Gurevych. 2022. Falsesum: Generating document-level nli examples for recognizing factual inconsistency in summarization. *arXiv preprint arXiv:2205.06009*.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel S Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. *arXiv preprint arXiv:2101.00288*.

Kun Yan, Lei Ji, Huaishao Luo, Ming Zhou, Nan Duan, and Shuai Ma. 2021. Control image captioning spatially and temporally. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2014–2025, Online. Association for Computational Linguistics.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. Linkbert: Pretraining language models with document links. *arXiv preprint arXiv:2203.15827*.

Wenpeng Yin, Dragomir Radev, and Caiming Xiong. 2021. DocNLI: A large-scale dataset for document-level natural language inference. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4913–4922, Online. Association for Computational Linguistics.

Chenguang Zhu, William Hinthorn, Ruochen Xu, Qingkai Zeng, Michael Zeng, Xuedong Huang, and Meng Jiang. 2020. Enhancing factual consistency of abstractive summarization. *arXiv preprint arXiv:2003.08612*.

Yaxin Zhu, Yikun Xian, Zuohui Fu, Gerard de Melo, and Yongfeng Zhang. 2021. Faithfully explainable recommendation via neural logic reasoning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3083–3090, Online. Association for Computational Linguistics.

# A Extraction Patterns for Identifying Conclusion Sentences

To extract the conclusion sentences from the abstract, we follow the recipe from Bastan et al. (2022). We filtered the abstracts from PubMed dataset which have a form of conclusion sentence at the end. In particular we filtered out all the abstracts that do not have any of the phrases in Table 7.

| Used Phrase |
|---|
| *we conclude that* |
| *it is concluded that* |
| *it was concluded that* |
| *we concluded that* |
| *we have concluded that* |
| *it has been concluded that* |
| *it may be concluded that* |
| *it was therefore concluded that* |
| *we therefore conclude that* |
| *we conclude* |
| *we thus conclude that* |
| *it is therefore concluded that* |
| *we further conclude that* |

Table 7: Used phrases to filter the abstracts

# B Hyper-parameter Selection

## B.1 Generation Lambda

One of the strategies to generate negative examples is the generation method (GEN). We trained a t5-large model on the SuMe dataset using 5-fold cross validation. Each time we trained on 4 folds and generated the output for the 5th fold. From the generated texts, we selected the ones which have lowest quality. That is, we selected generated sentences that contain both entities, the predicted relation labels are incorrect, and the Bleurt score of the generated sentence against the true mechanism sentence is lower than a threshold $\lambda$. In our preliminary experiments we found that $\lambda = 0.45$ yields a

| Parameter | Value | Details |
|---|---|---|
| min_tgt_length | 15 | min target length |
| max_tgt_length | 256 | max target length |
| bs | 4 | batch size |
| beam_size | 50 | beam size |
| length_penalty | 0.1 | length penalty for beam |
| ngram_size | 10 | ngrams occur once |
| prune_factor | 50 | candidates to keep |
| sat_tolerance | 2 | min satisfied constraints |
| beta | 2 | reward factor |

Table 8: Neurological decoding hyper parameters

good compromise between quality and yield. Analyzing the output of this hyper-parameter showed that $90\%$ of the sentences selected with this method are indeed true negative samples.

### B.2 Neurological Decoding Hyper-parameters

One of the strategies to generate negative examples is the generation method with the neurological decoding (GEN-ND) (Lu et al., 2021b). We used the source code introduced in their GitHub page[5]. The hyper-parameter details are shown in Table 8.

We also allowed for the use of negative or positive constraints, another choice that we use as a hyper-parameter.

[5] http://github.com/GXimingLu/neurologic_decoding