# LitGen: Genetic Literature Recommendation Guided by Human Explanations

Allen Nie[1,4], Arturo L. Pineda[1], Matt W. Wright[1,2], Hannah Wand[1,2,3], Bryan Wulf[1], Helio A. Costa[1], 2, Ronak Y. Patel[5], Carlos D. Bustamante[1,6], James Zou[1,4,6]

[1]*Department of Biomedical Data Science, Stanford University School of Medicine*
[2]*Department of Pathology, Stanford University School of Medicine*
[3]*Department of Cardiology, Stanford Healthcare*
[4]*Department of Computer Science, Stanford University*
[5]*Department of Molecular and Human Genetics, Baylor College of Medicine*
[6]*Chan-Zuckerberg Biohub*

As genetic sequencing costs decrease, the lack of clinical interpretation of variants has become the bottleneck in using genetics data. A major rate limiting step in clinical interpretation is the manual curation of evidence in the genetic literature by highly trained biocurators. What makes curation particularly time-consuming is that the curator needs to identify papers that study variant pathogenicity using different types of approaches and evidences—e.g. biochemical assays or case control analysis. In collaboration with the Clinical Genomic Resource (ClinGen)—the flagship NIH program for clinical curation—we propose the first machine learning system, LitGen, that can retrieve papers for a particular variant and filter them by specific evidence types used by curators to assess for pathogenicity. LitGen uses semi-supervised deep learning to predict the type of evi+dence provided by each paper. It is trained on papers annotated by ClinGen curators and systematically evaluated on new test data collected by ClinGen. LitGen further leverages rich human explanations and unlabeled data to gain 7.9%-12.6% relative performance improvement over models learned only on the annotated papers. It is a useful framework to improve clinical variant curation.

*Keywords*: Machine learning; Natural Language Processing; Clinical Genome; Variant pathogenicity curation

## 1. Introduction

The diversity of genetic variations that exist in the modern human population are slowly been recognized and discovered. Some of these variations are responsible for well-known physical differentiation across humans (e.g. hair color[1]), other variants can predict the development of inherited diseases like sickle-cell anemia or cystic fibrosis, and a few others are protective of disease, like some variations of PCSK9 which lowers the risk for coronary heart disease.[2] However, little is known overall about the more than 650 million variants known to date across the human genome.[3] In PubMed using the search term genetic variation returns over one million manuscripts, with almost half of them generated in the last 10 years.

Our understanding of previous published studies linking human genetic variants with medical syndromes and phenotypic traits is still limited. In 2013, the United States National Center for Biotechnology Information (NCBI) established the Clinical Genome program,[4] with the goal of defining the clinical relevance of key genes and variants through several gene and variant curation expert panels. These experts meet regularly to consider new evidence in the literature to curate and assess the pathogenicity of variants. The variant curation process combines clinical, genetic, population, and functional evidence with expert review to classify variants into 1 of 5 categories (Pathogenic, Likely Pathogenic, Variant of Unknown Significance, Likely Benign, Benign) according to the joint 2015 American College of Medical Genetics (ACMG), and Association for Medical Pathology (AMP) guidelines on clinical significance.[5]

The ACMG/AMP guidelines provide a set of criteria, and a curator searches for evidence and evaluates whether or not the evidence is sufficient to mark each criterion as met. A pathogenicity classification for each variant is calculated from the totality of the evidence evaluated using the ACMG/AMP criteria. Many of these criteria are mostly evaluated using pertinent information gleaned from publications, and finding the relevant publications that contain relevant evidence is a significant challenge to curators.

**The workflow of curating variants of clinical relevance.** The ClinGen procedure for biocurators[a] defines four steps to assess the pathogenicity of a variant: 1) select a variant of interest with and the suspected disease or mode of inheritance; 2) review available literature evidence about the disease; 3) curate evidence according to the ACMG/AMP criteria; 4) propose a level of pathogenicity. This process is assisted by ClinGens Variant Curation Interface[b]. Biocurators outside of the ClinGen environment follow a similar procedure. In the third step, when biocurators consider each of the ACMG/AMP criteria to systematically evaluate if the considered variant has some available literature. VCI further groups ACMG/AMP criteria into evidence types, many of which require evidence from published literature. Assessing which paper is relevant for each of the evidence types has a high burden of time and effort on the biocurator. To the best of our knowledge, there is currently no tool to automatically facilitate this task.

**Our contribution** We built a machine learning system LitGen that recommends papers to biocurators based on the evidence types presented in the paper. We believe this is the first system that analyzes papers for content on clinically relevant evidence types beyond variant name normalization or information matching.[6–8] We also contribute to the research area of semi-supervised learning with explanations. LitGen effectively uses explanations to guide semi-supervised learning. A thorough evaluation on new ClinGen data demonstrates that LitGen outperforms competitive baselines by a large margin.

## 2. Clinical Variant Curation Data

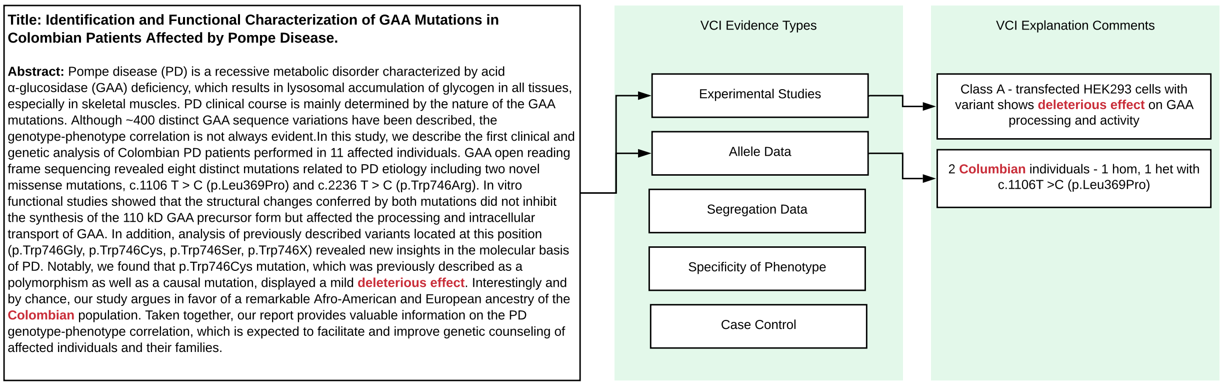
### 2.1. *ClinGen's Variant Curation Interface (VCI)*

The data that we use to develop LitGen are collected through ClinGens Variant Curation Interface (VCI). VCI is a curation web tool that was designed to support variant curation based on the ACMG/AMP Guidelines and serves as a platform for the standardized curation

[a] https://clinicalgenome.org/site/assets/files/3677/clingen_variant-curation_sopv1.pdf

[b] https://clinicalgenome.org/curation-activities/variant-pathogenicity/

of clinical variants by ClinGens Variant Curation Expert Panels. This pool of evidence can then be utilized by all VCI users when evaluating each of the ACMG/AMP criteria in turn within the interface. The VCI allows a user to provide an explanation comment describing the rationale for their evaluation in a text field, and to provide a PubMed ID linking to the relevant published literature that contains the data that supports their evaluation. The VCI allows the curator to assert whether the paper is relevant for a subset of evidence types. Here we focus on the five most common evidence types (Fig. 1).



**Title: Identification and Functional Characterization of GAA Mutations in Colombian Patients Affected by Pompe Disease.**

**Abstract:** Pompe disease (PD) is a recessive metabolic disorder characterized by acid α-glucosidase (GAA) deficiency, which results in lysosomal accumulation of glycogen in all tissues, especially in skeletal muscles. PD clinical course is mainly determined by the nature of the GAA mutations. Although ~400 distinct GAA sequence variations have been described, the genotype-phenotype correlation is not always evident.In this study, we describe the first clinical and genetic analysis of Colombian PD patients performed in 11 affected individuals. GAA open reading frame sequencing revealed eight distinct mutations related to PD etiology including two novel missense mutations, c.1106 T > C (p.Leu369Pro) and c.2236 T > C (p.Trp746Arg). In vitro functional studies showed that the structural changes conferred by both mutations did not inhibit the synthesis of the 110 kD GAA precursor form but affected the processing and intracellular transport of GAA. In addition, analysis of previously described variants located at this position (p.Trp746Gly, p.Trp746Cys, p.Trp746Ser, p.Trp746X) revealed new insights in the molecular basis of PD. Notably, we found that p.Trp746Cys mutation, which was previously described as a polymorphism as well as a causal mutation, displayed a mild **deleterious effect**. Interestingly and by chance, our study argues in favor of a remarkable Afro-American and European ancestry of the **Colombian** population. Taken together, our report provides valuable information on the PD genotype-phenotype correlation, which is expected to facilitate and improve genetic counseling of affected individuals and their families.

Class A - transfected HEK293 cells with variant shows **deleterious effect** on GAA processing and activity

2 **Columbian** individuals - 1 hom, 1 het with c.1106T >C (p.Leu369Pro)

Fig. 1. Paper annotation workflow. From a paper on PubMed (left), the curator selects which subset of the five variant curation (VCI) evidence types that the paper is relevant for (middle), and provide explanations for the selection (right). We highlight some keywords for emphasis. LitGen's goal is to predict which evidence types are relevant given a paper.

## 2.2. *Labeled papers*

We extracted all papers entered by VCI users between October 2016 and March 2019. The collected data include 1543 unique papers which contained clinical information on 932 unique variants. We randomly split this set of papers into train, dev, and test set by 0.9/0.05/0.05. Additionally, we collected a new set of 358 papers entered from April 2019 to May 2019 as a holdout evaluation dataset. Papers in this holdout evaluation dataset are entirely new. Table 1 shows the distribution of these two datasets. Each paper contains information that can be categorized into different evidence types that curators used to assert clinical pathogenicity. Curators can optionally provide an explanation comment for each type of evidence. In this manuscript, we focused on the top 5 VCI evidence types by the number of unique papers—these are Case Control, Specificity of Phenotype, Allele Data, Experimental Studies, and Segregation Data. These 5 evidence types covers 84% of all papers annotated in the VCI.

## 2.3. *Unlabeled papers*

In order to investigate whether semi-supervised learning can improve our model's performance, we collect a larger set of unlabeled papers through the following pipeline. We use ClinGen Allele Registry[3] to find the rsid of the variant if a clinical variant ID is provided. We use

Table 1. Labeled data summary: number of papers and explanations by VCI evidence type.

| | | # unique papers | | # explanations | |
|---|---|---|---|---|---|
| Evidence types in the VCI | ACMG criteria | Train | Holdout | Train | Holdout |
| Experimental Studies | BS3, PS3 | 385 | 74 | 732 | 80 |
| Allele Data | BP2, PM3 | 441 | 86 | 971 | 103 |
| Segregation Data | BS4, PP1 | 232 | 40 | 271 | 40 |
| Specificity of Phenotype | PP4 | 482 | 26 | 993 | 28 |
| Case Control | PS4 | 656 | 264 | 952 | 331 |
| **Total** | | **1543** | **358** | **3919** | **582** |

Training data collected during Oct 2016 to Mar 2019. Holdout evaluation data collected during April 2019 to May 2019. Note that we do not allow the algorithm to use explanations during test time. We have 1543 labeled data points for training.

LitVar API, a new service provided by NCBI,[9] to retrieve relevant literature of a given variant. LitVar scanned and indexed all of PubMed abstracts and PubMed Central full papers. We use this pipeline to retrieve all relevant papers to all variants curated through the ClinGen VCI. ClinGen Allele Registry found rsid for 877 of 932 variants (94.1%). We further found 742 (79.6%) variants that have been mentioned in the literature indexed in LitVar. We queried 4477 papers in total from LitVar, and 650 of these papers overlap with papers that have already entered into ClinGen by curators. Excluding these papers, we have 3827 new papers. We release all of our code and data at `https://github.com/windweller/ClinGenML/`.

## 3. Method

We use the following notations to describe our data. Each paper in our dataset is annotated with at least one VCI evidence type and the associated explanation comments on the rationale of selection. For the labeled papers dataset, we have $(x, \boldsymbol{y}) \in (\mathcal{X}, \boldsymbol{\mathcal{Y}})$ where $\boldsymbol{y} \in [0,1]^m$ for m labels and $m = 5$ in our case. Here $x$ represents the paper title and abstract. This is a multi-label setting because each paper can contain multiple evidence types. Each explanation comment is associated with exactly one evidence type. We can regard it as $(e, y) \in (E, \mathcal{Y})$, where $e$ is the explanation text and $y \in \{1, ..., m\}$ describes the evidence type.

### 3.1. *BiLSTM baseline*

We aim to train a competitive supervised learning algorithm on the labeled data. We use the state-of-the-art text processing algorithm for our model: long-short-term memory networks (LSTMs). It has been used in many natural language processing applications,[10] generating complex human responses,[11] and well-adopted in clinical text processing.[12,13] We use the bidirectional variant of this algorithm proposed by Graves et al..[14]

For each paper abstract $x = w_1, ..., w_T$, we compute the hidden state vectors $H = [h_1, ..., h_T]$. We compute the vector representation of the abstract $c(x)$ using the global max-pooling over the temporal dimension suggested by Collobert & Weston.[15] At last, we predict whether an evidence type $y_i$ exist through a sigmoid binary classifier with parameter $\theta_i$. We compute the binary cross-entropy loss through the predicted labels $\hat{\boldsymbol{y}} = [\hat{y}_1, ..., \hat{y}_m]$ and true labels $\boldsymbol{y}$.

$$H = [h_1, ..., h_T] = \text{BiLSTM}(w_1, ..., w_T),\ H \in \mathbb{R}^{T \times d} \tag{1}$$

$$\boldsymbol{c}(x) = [\max(H_{\cdot,1}), \max(H_{\cdot,2}), ..., \max(H_{\cdot,d})], \boldsymbol{c}(x) \in \mathbb{R}^d \tag{2}$$

$$P(y_i) = \hat{y}_i = \sigma(\theta_i^\intercal \boldsymbol{c}(x)),\ \text{for } i = 1, ..., m \tag{3}$$

$$\mathcal{L}_{\text{BCE}}(x, \hat{\boldsymbol{y}}, \boldsymbol{y}) = -\tfrac{1}{m} \textstyle\sum_{i=1}^{m} y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \tag{4}$$

### 3.2. *Leveraging unlabeled data*

After training a competitive baseline model only on limited labeled data, we explore the possibility of leveraging unlabeled paper by using a proxy labeling model. Proxy-label approach to semi-supervised learning has been generally shown to improve the performance of the final model. This approach aims to produce proxy labels on unlabeled data, which later are used as targets together with labeled data to train the final model. These proxy labels do not reflect the ground truth labels, but they might provide some signals for learning.[16,17]

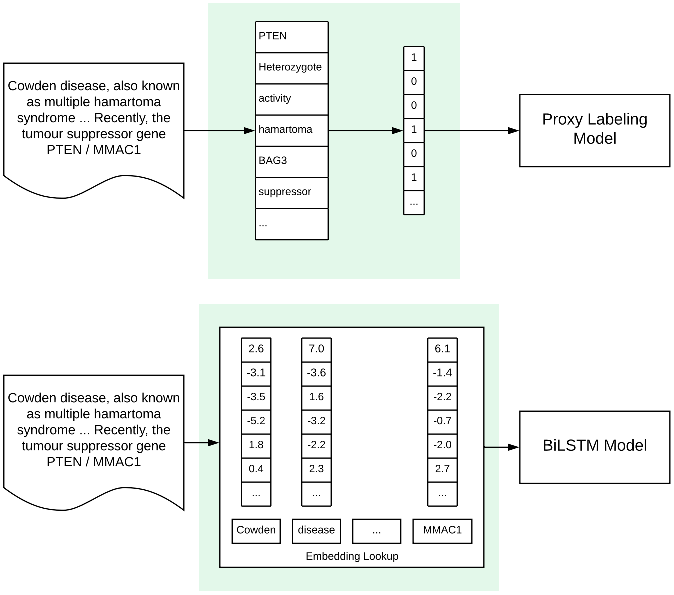

Cowden disease, also known as multiple hamartoma syndrome ... Recently, the tumour suppressor gene PTEN / MMAC1
PTEN
Heterozygote
activity
hamartoma
BAG3
suppressor
...
1
0
0
1
0
1
...
Proxy Labeling Model
Cowden disease, also known as multiple hamartoma syndrome ... Recently, the tumour suppressor gene PTEN / MMAC1
2.6
-3.1
-3.5
-5.2
1.8
0.4
...
7.0
-3.6
1.6
-3.2
-2.2
2.3
...
6.1
-1.4
-2.2
-0.7
-2.0
2.7
...
Cowden
disease
...
MMAC1
Embedding Lookup
BiLSTM Model

Fig. 2. **Naive Unlabeled:** The two views of inputs for the proxy labeling model and the BiLSTM model.

We train a random forest model to predict evidence types on our labeled dataset (represented as bag-of-words). We then apply this random forest to predict labels for each unlabeled paper; we call these the proxy labels. Finally we train our BiLSTM model on proxy-labeled unlabeled data and labeled data together. We refer to this strategy as **Naive Unlabeled**, because it is a simple and direct approach to use the unlabeled papers. The point of using the random forest to generate the proxy labels is that it contains different inductive bias compared to the original BiLSTM. Zhou & Goldman[17] showed that when the proxy labeling models have different bias compared to the final classifier, the generated proxy labels can often improve the model's performance.

### 3.3. *Explanations in multitask learning*

Beyond building a strong BiLSTM baseline and incorporate proxy labeling methods on unlabeled data, another important feature of our curation dataset is that we have human-provided explanations associated with each paper. Each explanation is a concise summary of *why* the curator asserted that a paper provides a particular type of pathogenicity evidence. We hypothesized that these explanations could help us to generating features that are salient for evidence predictions. Contrary to using humans to label each training example, which is very costly both in terms of time and resource, recent works have explored whether human-provided explanations will allow models to learn beyond instance-level labels. Early works focus on using semantic parsing over human explanations to obtain labeling functions.[18,19] However, such approaches are limited to explanations that have fixed format such as "X because of Y and Z". The explanations provided by our curators are free text and do not conform to predefined templates. An innovation of our work is on how to leverage these explanations.

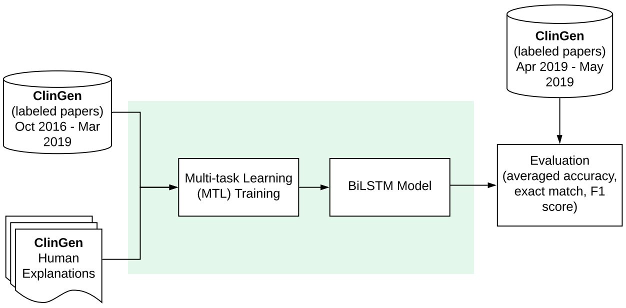


Fig. 3. Multi-task learning pipeline that leverages labeled data and explanations

A simple way to use the explanations is to treat them simply as additional labeled examples where the label is the associated VCI evidence type. We build a multi-task learning objective, where the BiLSTM model is asked to optimize for two tasks: predicting whether a paper contains information relevant to a VCI evidence type (original task, loss marked as $\mathcal{L}_1$), as well as whether an explanation is provided as rationale for a VCI evidence type (explanation prediction task, loss marked as $\mathcal{L}_2$). For each epoch, we train on two tasks separately: first on the explanation prediction task, and after iterating through all batches of explanations, we train on the original paper abstract prediction task. We use a scalar hyperparameter $\lambda \in [0, 1]$

to scale the loss of the explanation prediction task. We call this approach **Naive exp**.

There are inherent problems to this approach. First of all, when we train on $(e, y)$, the explanations have a different length distribution compares to $x$, the paper abstracts. Explanations tend to be shorter and more succinct. Since we are using the same BiLSTM model to process both texts, we are learning from two data distributions. Second, even though both explanations and papers are associated with a VCI evidence type, one explanation can only exclusively be used to justify for one VCI evidence type, while a paper can be associated with multiple VCI evidence types. Therefore, the nature of data-to-label mapping is different for the two tasks. The last problem is that explanations are noisy. Curators submit these explanations often as a comment or additional information to support their choice of paper. Not all words in explanations are useful for the original task. We address all three problems by proposing our new approach: use explanations to perform feature selection, and then use the selected features to proxy label the unlabeled papers.

### 3.4. *Explanations as feature selection for proxy labeling*

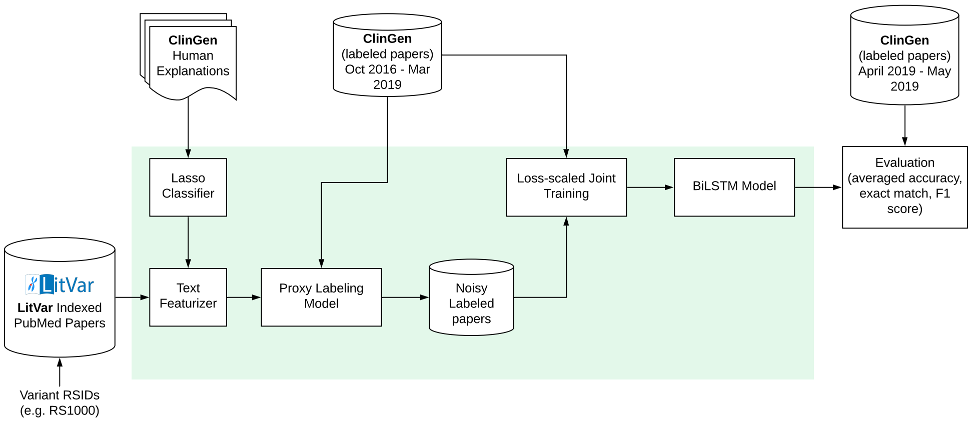
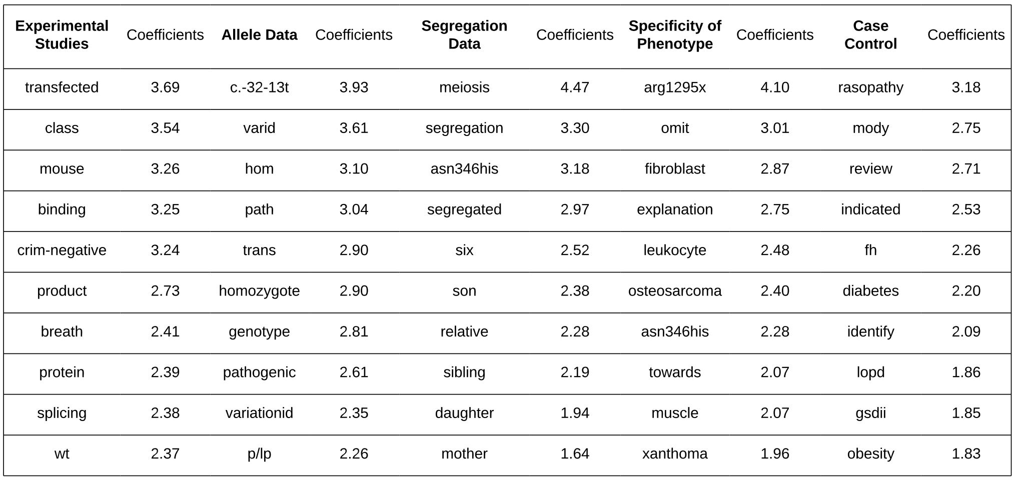


Fig. 4. General training pipeline that leverages unlabeled data guided by explanations.

We first train a Lasso classifier (a logistic regression classifier with $L_1$ regularization) on the frequency-encoded unigram feature representation of the explanations. Our Lasso classifier obtains a coefficient on each word that determines whether the word is important for the prediction of which VCI evidence type an explanation is associated with. Our Lasso classifier obtains 89.0% accuracy on this classification task. This shows that explanations are easier to classify compared to paper abstracts and they contain useful signals that can be leveraged. We extract words that have non-zero coefficients. We display some of these words in Figure 5. In total, we are able to find 799 words that have non-zero coefficient out of 7550 words contained in explanations. We use these 799 words as the selected features and then follow the same proxy labeling strategy as the Naive Unlabeled algorithm.

For each paper abstract, we record the frequency of these 799 unique words in the abstract and ignore all other words. Originally, in section 3.2, we naively encode the paper abstract



| **Experimental Studies** | Coefficients | **Allele Data** | Coefficients | **Segregation Data** | Coefficients | **Specificity of Phenotype** | Coefficients | **Case Control** | Coefficients |
|---|---|---|---|---|---|---|---|---|---|
| transfected | 3.69 | c.-32-13t | 3.93 | meiosis | 4.47 | arg1295x | 4.10 | rasopathy | 3.18 |
| class | 3.54 | varid | 3.61 | segregation | 3.30 | omit | 3.01 | mody | 2.75 |
| mouse | 3.26 | hom | 3.10 | asn346his | 3.18 | fibroblast | 2.87 | review | 2.71 |
| binding | 3.25 | path | 3.04 | segregated | 2.97 | explanation | 2.75 | indicated | 2.53 |
| crim-negative | 3.24 | trans | 2.90 | six | 2.52 | leukocyte | 2.48 | fh | 2.26 |
| product | 2.73 | homozygote | 2.90 | son | 2.38 | osteosarcoma | 2.40 | diabetes | 2.20 |
| breath | 2.41 | genotype | 2.81 | relative | 2.28 | asn346his | 2.28 | identify | 2.09 |
| protein | 2.39 | pathogenic | 2.61 | sibling | 2.19 | towards | 2.07 | lopd | 1.86 |
| splicing | 2.38 | variationid | 2.35 | daughter | 1.94 | muscle | 2.07 | gsdii | 1.85 |
| wt | 2.37 | p/lp | 2.26 | mother | 1.64 | xanthoma | 1.96 | obesity | 1.83 |

Fig. 5. We display a set of keywords that are the most positively associated with each VCI evidence types from human explanations by training a lasso model on unigram features. Coefficients refer to Lasso coefficients.

obtaining a vector equivalent to the size of the entire vocabulary space (after removing stop words and punctuation), which is 16860. We have now restricted the dimension of the vector representing the paper abstract from 16860 to 799. We refer to this feature selection process as **Exp-guided**. We then train our random forest proxy labeling model on the paper abstracts and use it to generate proxy labels for unlabeled data. At last, we train the BiLSTM model on both proxy labeled unlabeled data and ground-truth data. We refer to this setting as the **LitGen** model.

Another advantage to our explanation-guided feature selection process is that we can now automatically generate labeling functions without semantic parsing. We use a simple heuristic to binarize the coefficients in our Lasso classifier: each of the 799 words is a labeling function. If the word has a positive coefficient, we output +1 when we encounter this word. If the word has a negative coefficient, we output -1. When the word is missing, we output 0. This allows us to leverage labeling function aggregation algorithm such as Snorkel-MeTal.[20] We include this result to show that by selecting features from explanations, we are able to leverage multiple approaches in semi/weakly supervised learning. We refer to this setting as **Exp-guided Snorkel**.

## 4. Experimental results

### 4.1. *Evaluation metrics*

We use the following metrics to evaluate model performance in predicting the evidence types given a paper. We compute the average accuracy (Avg Accu) across VCI evidence types. Accuracy reflects how correctly the model can determine whether the paper contains a type of evidence or not. We compute the exact match ratio (EM) as well, which is a more strict metric that requires the model's predictions to exactly match every ground truth label. Finally, we also compute the average $F_1$ score weighted by the number of examples in each evidence type (Wgt F1). All the models are trained on data up to March 2019 and are evaluated on new ClinGen paper annotated from April to May 2019.

### 4.2. *Performance comparison*

We train the **LitGen model** based on the strategies described in the method section. We also consider a baseline classifier that randomly predicts the value of each label based on the class balance of the training data. We evaluate all trained models on a final holdout set of 347 disjoint papers. In Table 2, we show the performance of our proposed methods to incorporate explanations into the supervised learning and proxy-label semi-supervised learning pipelines.

Table 2. Performance of different training strategies for LitGen model.

| | Apr 2019 to May 2019 | | |
|---|---|---|---|
| Strategy | Avg Accu | EM | Wgt $F_1$ |
| Baseline (Majority) | 62.9 | 8.7 | 36.0 |
| BiLSTM | 82.6 | 45.2 | 62.7 |
| BiLSTM + Naive Exp | 83.8 | 48.7 | 66.5 |
| BiLSTM + Naive Unlabled | 83.9 | 50.1 | 65.7 |
| BiLSTM + Naive Exp + Naive Unlabeled | *82.9* | *48.4* | *66.4* |
| BiLSTM + Exp-guided Snorkel | 84.0 | 50.1 | 66.8 |
| **LitGen**: BiLSTM + Exp-guided Unlabeled | **85.0** | **51.6** | **68.1** |

**Unlabeled data and explanations both help** We observe the improvement over BiLSTM model when training on proxy-labeled paper abstracts and leveraging explanations: both *BiLSTM + Naive Unlabled* and *BiLSTM + Naive Exp* outperform *BiLSTM* on all the evaluation metrics. That naive training on explanations leads to improvement shows that explanations do provide learning signals for the model.

**Naive joint training hurts performance** However, even though training on explanation prediction task or training on proxy-labeled paper abstracts each improves the final model's performance, such effect is not additive when we train on both. *BiLSTM + Naive Exp + Naive Unlabeled* performs relatively poorly. We have discussed potential drawback of training naively on explanations such as text length distribution mismatch and noisy explanation text.

**Using explanations for feature selection outperforms all** Explanations contain valuable learning signals but are noisy in its writing. When we use them for feature selection, choosing words that are determined important by a Lasso classifier, we accomplish two goals at once: 1) reducing the overall document feature vector dimension for the random forest proxy labeling model; 2) provide a set of labeling functions that can be leveraged by algorithms like Snorkel-MeTal. We can see in Table 2, this approach produces two best performing final models.

### 4.3. *Performance of Proxy Labeling Model*

Table 3. Evaluation of the quality of generated proxy labels on the holdout test set.

| | Apr 2019 to May 2019 | | |
|---|---|---|---|
| Labeling model | Avg Accu | EM | Wgt $F_1$ |
| Naive Unlabeled | 81.2 | 40.3 | 53.2 |
| Exp-guided Unlabeled | 82.8 | 46.1 | 60.0 |
| Exp-guided Snorkel | 11.5 | 2.6 | 42.3 |

We performed additional analysis to gain more insights into the improved performance of LitGen due to proxy labels on the unlabeled papers. Since we do not have access to ground truth labels for the unlabeled papers, we evaluate the performance of the proxy labeling models on the holdout evaluation dataset that we used to evaluate our BiLSTM model. It is notable that the random forest with explanation-guided feature selection (Exp-guided Unlabeled) gives reasonably accurate proxy labels, and is indeed more accurate than the Naive Unlabeled which does not have this feature selection. Moreover because this random forest derived proxy label provides complementary signal, training the original BiLSTM on this additional data leads to additional improvements and give rise to our final LitGen algorithm. We note that popular weak supervision algorithm, Snorkel, performs poorly with our automatic labeling functions.

### 4.4. *Performance by Evidence Types*

Table 4. Accuracy of baseline (always guess the majority class), BiLSTM and LitGen model for each evidence type.

| Evidence type | Baseline (Majority) | BiLSTM | LitGen |
|---|---|---|---|
| Experimental Studies | 63.1 | 85.6 | **86.7** |
| Allele Data | 65.7 | 80.4 | **83.0** |
| Segregation Data | 73.8 | 88.8 | 88.8 |
| Specificity of Phenotype | 66.0 | 87.0 | **90.2** |
| Case Control | 45.8 | 71.2 | **76.4** |

We show the performance of our model on each of the evidence types in Table 4. We can see that one of the most difficult class to predict is the evidence type "segregation data".

We conjecture that this is because we only used paper abstracts. Most segregation data are mentioned in the actual content of the paper. However, it remains a major challenge for a deep learning system to consume input as long as a full scientific paper. One of the easiest evidence types to learn is “experimental studies” because curators mostly look for experimental procedure keywords and most of them are present in the abstracts.

## 5. Discussion

**Automatic literature recommendation for variant curation** We propose a new goal for the field of literature recommendation: automatically generate semantic tags according to VCI evidence types to aid biocurators in filtering papers. We are operating under a low-resource setting where few papers have currently been annotated by experts. However, such annotations are very rich and often contain explanations to justify curator’s decisions to submit a paper as evidence. We propose a pipeline that leverages explanations beyond semantic parsing and can be automatically learned by training a Lasso classifier.

**Implication for the curation pipelines** In the era of implementing genomic medicine, machine assistance is needed for scalability. Human time should be reserved for steps that need true domain expertise and critical interpretation. A feasible model for systematic curation at scale would automate the generation and delivery of gene or variant level information to expert biocurators that can then critique the quality and relevance of the evidence in the context of a specific disease. This reduces the time it takes to identify evidence of interest that need more in depth human review.

Our machine learning model for predicting relevant literature by variant and evidence type is well suited for a semi-automated model of curation at scale. Early efforts in automated literature curation have been able to recommend papers by matching for the variant of interest. The added functionality suggests what type of evidence helps to further streamline curation workflow and efficiency by pre-mapping evidence onto predicted ACMG/AMP criteria. Displaying papers by evidence type also matches the natural organization of curation interfaces such as the VCI, making this an even more feasible tool to implement and have true clinical impact. LitGen is not meant to replace biocurators, but rather to facilitate the curation process by prioritizing papers that are more likely to contain particular types of evidence.

**Acknowledgement** This project is supported by NIH ClinGen. J.Z. is also supported by the Chan-Zuckerberg Biohub. We would like to thank S. Plon and A. Milosavljevic for feedback.

## References

1. W. Branicki, F. Liu, K. van Duijn, J. Draus-Barini, E. Pośpiech, S. Walsh, T. Kupiec, A. Wojas-Pelc and M. Kayser, Model-based prediction of human hair color using dna variants, *Human genetics* **129**, 443 (2011).
2. J. C. Cohen, E. Boerwinkle, T. H. Mosley Jr and H. H. Hobbs, Sequence variations in pcsk9, low ldl, and protection against coronary heart disease, *New England Journal of Medicine* **354**, 1264 (2006).
3. P. Pawliczek, R. Y. Patel, L. R. Ashmore, A. R. Jackson, C. Bizon, T. Nelson, B. Powell, R. R.

Freimuth, N. Strande, N. Shah *et al.*, Clingen allele registry links information about genetic variants, *Human mutation* **39**, 1690 (2018).

4. H. L. Rehm, J. S. Berg, L. D. Brooks, C. D. Bustamante, J. P. Evans, M. J. Landrum, D. H. Ledbetter, D. R. Maglott, C. L. Martin, R. L. Nussbaum *et al.*, Clingenthe clinical genome resource, *New England Journal of Medicine* **372**, 2235 (2015).
5. S. Richards, N. Aziz, S. Bale, D. Bick, S. Das, J. Gastier-Foster, W. W. Grody, M. Hegde, E. Lyon, E. Spector *et al.*, Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the american college of medical genetics and genomics and the association for molecular pathology, *Genetics in medicine* **17**, p. 405 (2015).
6. C.-H. Wei, L. Phan, J. Feltz, R. Maiti, T. Hefferon and Z. Lu, tmvar 2.0: integrating genomic variant information from literature with dbsnp and clinvar for precision medicine, *Bioinformatics* **34**, 80 (2017).
7. J. Birgmeier, M. Haeussler, C. A. Deisseroth, K. A. Jagadeesh, A. J. Ratner, H. Guturu, A. M. Wenger, P. D. Stenson, D. N. Cooper, C. Ré, J. A. Bernstein and G. Bejerano, Amelie accelerates mendelian patient diagnosis directly from the primary literature, *bioRxiv* (2017).
8. V. Kuleshov, J. Ding, C. Vo, B. Hancock, A. Ratner, Y. Li, C. Ré, S. Batzoglou and M. Snyder, A machine-compiled database of genome-wide association studies, *Nature Communications* **10**, p. 3341 (2019).
9. A. Allot, Y. Peng, C.-H. Wei, K. Lee, L. Phan and Z. Lu, Litvar: a semantic search engine for linking genomic variant data in pubmed and pmc, *Nucleic acids research* **46**, W530 (2018).
10. T. Mikolov, Statistical language models based on neural networks, *Presentation at Google, Mountain View, 2nd April* (2012).
11. A. Nie, E. Bennett and N. Goodman, Learning to explain: Answering why-questions via rephrasing, in *Proceedings of the First Workshop on NLP for Conversational AI*, (Association for Computational Linguistics, Florence, Italy, August 2019).
12. A. Nie, A. Zehnder, R. L. Page, Y. Zhang, A. L. Pineda, M. A. Rivas, C. D. Bustamante and J. Zou, Deeptag: inferring diagnoses from veterinary clinical notes, *npj Digital Medicine* **1**, p. 60 (2018).
13. Y. Zhang, A. Nie, A. Zehnder, R. L. Page and J. Zou, Vettag: improving automated veterinary diagnosis coding via large-scale language modeling, *npj Digital Medicine* **2**, p. 35 (2019).
14. A. Graves, S. Fernández and J. Schmidhuber, Bidirectional lstm networks for improved phoneme classification and recognition, in *International Conference on Artificial Neural Networks*, 2005.
15. R. Collobert and J. Weston, A unified architecture for natural language processing: Deep neural networks with multitask learning, in *Proceedings of the 25th international conference on Machine learning*, 2008.
16. A. Blum and T. Mitchell, Combining labeled and unlabeled data with co-training, in *Proceedings of the eleventh annual conference on Computational learning theory*, 1998.
17. Y. Zhou and S. Goldman, Democratic co-learning, in *16th IEEE International Conference on Tools with Artificial Intelligence*, 2004.
18. S. Srivastava, I. Labutov and T. Mitchell, Joint concept learning and semantic parsing from natural language explanations, in *Proceedings of the 2017 conference on empirical methods in natural language processing*, 2017.
19. B. Hancock, M. Bringmann, P. Varma, P. Liang, S. Wang and C. Ré, Training classifiers with natural language explanations, in *Proceedings of the conference. Association for Computational Linguistics. Meeting*, 2018.
20. A. Ratner, B. Hancock, J. Dunnmon, R. Goldman and C. Ré, Snorkel metal: Weak supervision for multi-task learning, in *Proceedings of the Second Workshop on Data Management for End-To-End Machine Learning*, 2018.