Multiple sampling schemes and deep learning improve active learning performance in drug-drug interaction information retrieval analysis from the literature

Lang Li (lang.li@osumc.edu)
The Ohio State University
Weixin Xie
The Ohio State University
Kunjie Fan
The Ohio State University
Shijun Zhang
The Ohio State University

Research Article

Keywords: active learning, deep learning, drug-drug interaction, information retrieval, random negative sampling, positive sampling, similarity sampling, and uncertainty sampling

Posted Date: March 15th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1435945/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Journal of Biomedical Semantics on May 30th, 2023. See the published version at https://doi.org/10.1186/s13326-023-00287-7.

Abstract

Background: Drug-drug interaction (DDI) information retrieval (IR) is an important natural language process (NLP) task from the PubMed literature. For the first time, active learning (AL) is studied in DDI IR analysis. DDI IR analysis from PubMed abstracts faces the challenges of relatively small positive DDI samples among overwhelmingly large negative samples. Random negative sampling and positive sampling are purposely designed to improve the efficiency of AL analysis. The consistency of random negative sampling and positive sampling is shown in the paper.

Results: PubMed abstracts are divided into two pools. Screened pool contains all abstracts that pass the DDI keywords query in PubMed, while unscreened pool includes all the other abstracts. At a prespecified recall rate of 0.95, DDI IR analysis performance is evaluated and compared in precision. In screened pool IR analysis using supporting vector machine (SVM), similarity sampling plus uncertainty sampling improves the precision of AL over uncertainty sampling, from 0.89 to 0.92 respectively. In the unscreened pool IR analysis, the integrated random negative sampling, positive sampling, and similarity sampling improve the IR analysis performance over uncertainty sampling along, from 0.72 to 0.81 respectively. When we change the SVM to a deep learning method, all sampling schemes consistently benefit DDI AL analysis in both screened pool and unscreened pool. Deep learning has significant improvement of precision over SVM, 0.96 vs 0.91 in screened pool, and 0.90 vs 0.81 in the unscreened pool, respectively.

Conclusions: By integrating various sampling schemes and deep learning algorithms into AL, the DDI IR analysis from literature is significantly improved. The random negative sampling and positive sampling are highly effective methods in improving AL analysis where the positive and negative samples are extremely imbalanced.

Background

Drug-drug interaction (DDI) is one of the major risk factors that cause adverse drug events (ADEs). Nearly 22% and 9% of ED visits and hospitalizations, respectively, are caused by DDIs [1–4].DDIs are most prevalent among older adults because of the disproportionately high prevalence of polypharmacy [5–7]. DDI is a major research topic in pharmacokinetics (PK) and pharmaco-epidemiology (PE) studies. The DDI pharmacology mechanisms usually are investigated in PK studies, in which the change of one drug's metabolism and transportation are compared in the presence and absence of another drug [8]. PK DDI studies sometimes are performed *in vitro*, i.e. using either recombinant enzymes, or human liver microsome, or hepatocyte. Clinical PK DDI study is another important approach in assessing whether one drug exposure is altered by another co-committed drugs [8–9]. Using large scale claim or EHR databases, pharmaco-epidemiological studies, on the other hand, focus on whether DDIs change ADE risks in targeted patient populations [10]. DDI induced ADEs, sometimes, are also published in patient case reports [11–12]. If drug combinations are tested in controlled clinical trials, their efficacy and ADEs are always compared to those of single drugs [13]. An enormous amount of DDI information has been published in the biomedical literature. It has been a great interest in mining and curating these DDI

information for assisting physicians and patients in preventing DDIs and their associated ADEs. In this paper, we will focus on mining published DDI studies related to ADEs. They are either pharmacoepidemiology studies, or case reports, or controlled clinical trials.

There are two major DDI text mining tasks from PubMed: information retrieval (IR) [14] and information extraction (IE). The goal of DDI IR is to identify DDI relevant publications or abstracts, while DDI IE is to extract DDI pairs from the DDI relevant publications or abstracts. DDI IR is always the first step in identifying DDI relevant publications and abstracts. Then, DDI IE task relies on annotated DDI relationships in positively labeled DDI paper or abstracts generated from the DDI IR step. DDI IR and IE analyses were reviewed in our early paper in 2014[15]. In this paper, our literature review will focus on DDI IE and IR methods after 2014.

Deep learning (DL) techniques are clearly the major trend in recent DDI IE analysis. Zhao et al. [16] proposed a syntax convolutional neural network that combined a traditional convolutional neural network and external features (contexts, shortest path, part-of-speech) to extract DDIs. It obtained a F1-scores of 0.69 for DDI extraction. By integrating a recurrent neural network with multichannel word embedding, Zheng et al. [17] combined an attention mechanism and a recurrent neural network with long short-term memory (LSTM) units and obtained a system that performed well for DDI extraction (F1 = 0.77). Zhang et al. [18] integrated the shortest dependency paths and sentence sequence by a hierarchical recurrent neural networks-based method, which produced an F1-score of 0.73 for DDI extraction. Wang et al. [19] introduced the dependency-based technique to a bi-directional LSTM network, built a linear depth-first search and a breadth-first search, and it achieved an F1-score of 0.72 for DDI extraction. In a recent paper, Zhang et al. [20] shortest dependency path was integrated with both convolutional neuron network model and recurrent neuron network model in DDI IE analysis. It reported an F1-score of 0.75.

However, the research on DDI IR analysis has not been as advanced as DDI IE methodology development. Our DDI IR analysis in 2015 presented the most comprehensive comparisons among many machine learning (ML) methods. It demonstrated that linear discriminant analysis, logistic regression, and supporting vector machine all had similar performance, F1-score = 0.93, in identifying DDI related abstracts in PubMed [21]. However, if the recall rate was set as 0.95, DDI IR precision became as low as 0.67.

The under-developed DDI IR methodology is largely due to the lack of negatively labeled DDI PubMed in the existing DDI corpora [22] which contain only positively label DDI abstracts. While building up more negatively and positively labeled DDI abstracts shall certainly help in further developing DDI IR methodology, it is more interesting to explore the interactive process between DDI annotations and DDI IR optimization. This falls into one territory of artificial intelligent field, active learning (AL) [23]. AL attempts to maximize the performance of the ML algorithms while annotating as few samples as possible [24]. The application of AL to biomedical text mining is rather limited. As one example, its use to mine text in electronic medical record data to identify disease phenotype reduced the number of annotated samples required to achieve an AUC of 0.95 by 68% in predicting patients with rheumatoid arthritis.

Introduced by Lewis and Gale in 1994[25], AL optimizes ML algorithms sequentially based on user feedback. AL uses uncertainty sampling to guide ML training on new samples for which ML has demonstrated the lowest predictive performance. The primary AL research has focused on uncertainty sampling schemes, such as least confidence, margin sampling, entropy, query by committee, expected model change, expected error reduction and variance reduction[26]. Although AL can potentially improve the DDI IR analysis, there are several challenges that motivate the development of new AL methodology in this paper. Firstly, positively labeled DDI abstracts in the current DDI corpora were selected from a query of keywords, such as “drug interaction,” limits the identification of all relevant general abstracts in PubMed. If this is ignored in AL, it will lead to a biased DDI IR analysis. Secondly, more than 99% of PubMed abstracts are unrelated to DDIs, and the labeling of positive samples is more labor intensive, and therefore more expensive than labeling negative samples. Thus, to be more cost effective, AL should take greater advantage of the large-scale availability of negative data, but current uncertainty sampling schemes do not deal with them. DL approaches have been developed and implemented for the DDI IE analysis, but not yet for DDI IE. DL could significantly improve the performance of AL with respect to DDI IR.

Methods

DDI corpus and annotation guideline

There are two sample pools in this study. The first one is called screened sample pool, which are the abstracts in PubMed through keyword queries: [“drug interaction” AND (Type of Study)] and [“drug combination” AND (Type of Study)]. The “Type of Study” is defined in Table 1: clinical trial, pharmaco-epidemiology study, and case report. Based on the criteria for DDI abstract selection in Table 1, sample abstracts are reviewed and annotated. A corpus [27] is built, which has 933 positive DDI abstracts and 799 negative abstracts. They are the initial labeled samples in the screened sample pool. Table 1 presents inclusion and exclusion criteria for the screened sample pool abstract selection. 5,000 abstracts are randomly selected from screened samples as the screened sample pool in this study.

The other sample pool is called unscreened sample pool. It is made up of 10,000 abstracts that are randomly selected from PubMed and are not overlapped with screened sample pool. This unscreened sample pool, on the other hand, contains data are largely not DDI relevant. Data distribution for screened sample pool and unscreened sample pool is shown in Table 2.

Two annotators with complementary skills in biology and informatics develops this corpus. Mrs. Shijun Zhang, has a master’s degree in biology, and has worked in Dr. Li’s lab for 5 years with the primary research responsibility of corpus development for drug-interaction text mining[22]; Mrs. Weixin Xie, a PhD student in medical informatics, has conducted pharmacology and drug-interaction text-mining research under Dr. Li’s supervision. Training and education in labeling have an initial calibration step, in which two individuals label each abstract according to the inclusion and exclusion criteria outlined in Table 1, the

agreement between their labels is then evaluated for the first 30 positive abstracts (30 in each of the three DDI categories), and they receive further training based on that analysis.

Table 1
Inclusion and exclusion criteria for clinical DDI abstract selection

Inclusion (positive)	Clinical trial DDI study: Phase I/II/III clinical trials in which drug combination and/or single drug ADE are evaluated and reported.
	Pharmaco-epidemiological DDI study: Pharmaco-epidemiology studies in which ADEs from drug combinations are reported and compared to single drug ADEs.
	DDI and ADE case reports: DDI-induced ADE cases in which the time sequential drug and ADE are reported in clinical care settings.
Exclusion (negative)	Clinical PK DDI study: both single drug and drug combination exposures (i.e. pharmacokinetics) are evaluated either in patients or healthy volunteers.
	Clinical PK PG study: the single drug exposure (i.e. pharmacokinetics) is evaluated among patients that have different genotypes in CYP450 and UGT enzymes and drug transporters.
	in vitro PK study: substrate depletion and metabolite formation study is for the fm data collection; and inhibition study is for the Ki data collection.
	Drug interaction detection algorithms or software
	Compliance of avoiding DDI
	Concordance of DDI reporting among different drug interaction knowledge base.
	Comparison of the performance of DDI clinical decision systems
	Drug-alcohol/food interactions
	Drug/test interactions
	Case report studies
	Review papers
	Cell culture and animal studies
	Other studies that are not related to drug interactions.

Table 2
Statistics of DDI corpus

Data Source	Sample pool	Data set	Sample size	Initial training set	Initial validation set
PubMed	Screened sample pool	Labeled Positive	933	100 +*	50 +*
		Labeled Negative	799	100 -*	50 -*
		Unlabeled screened samples	3,169		50 R*
	Unscreened sample pool	Unlabeled unscreened samples	9,999	100 +*	50 +*
				100 R*	50 -*
					50 R*

Note: +* (labeled positive samples), -* (labeled negative samples), R* (random negative samples).

Sampling strategies in active learning

- Uncertainty sampling in AL refers to selecting the least confidence new samples, e.g. abstracts with predicted probability around 0.5 in a binary classification (i.e. DDI relevant or not), for the next round labeling and training in machine learning analysis.
- Positive sampling refers to selecting the most certain positive new samples, e.g. predicted probability close to 1 in a binary classification, for next round labeling and training in machine learning analysis.
- Random negative sampling Because more than 99% of unscreened pool abstracts are not DDI related, a random subset of unscreened pool is chosen as negative samples. These random negative samples may contain very small fraction of positive samples [28].
- Similarity sampling *aims to quick screen out samples that more like samples in corpus*, the cosine similarity (cosSIM) based on TF-IDF (Term Frequency-Inverse Document Frequency) [29] of each unlabeled sample and all the samples in corpus is used to evaluated. The TF(t) and IDF(t) of term t (word t) are formulated as

$$TF(t) = \frac{termt\#inanabstract}{totalterm\#inanabstract}; IDF(t) = ln\frac{Total\#ofabstract}{\#ofabstractwithtermt};$$

TF(t) measures how frequently a term *t* occurs in an abstract, and *IDF(t)* measures how important the term t is. In fact, certain terms that occur too frequently have little power in determining the relevance, therefore, we need to weigh up the effects of the less frequently occurring terms. And then, we got the TFIDF for term t by computing the following:

$$TFIDF(t) = TF(t) \times IDF(t);$$

Above multiplying TF(t) and IDF(t) results in the TFIDF score of a term t in an abstract. The higher the sore, the more relevant that term is in that particular abstract. For each abstract, we derived 30 key terms

with high TFIDF, and their frequency vector of each abstract was generated to calculate the cosine similarity (cosSIM). For example, abstracts A and B are two n-dimensional vectors, $A = (A_1, \ldots, A_n)$ and $B = (B_1, \ldots, B_n)$, using the formula below we can find out the cosine similarity between *A* and *B*:

$$cosSIM\,(A, B) = \frac{|A \bullet B|}{|A| \times |B|}$$

.

Here, the cosine similarity of abstract A and B ranges from 0 to 1. In this study, sample in pools has its similarity values with samples which are DDI related abstracts, the higher the similarity value, the more DDI related abstract likely. Similarity sampling is applied in conjunction with other sampling strategies in two sample pools, they will benefit to quick find more informative samples from unlabeled pools, so that is more cost effective for training models.

Existing AL analyses only uses uncertainty sampling. In this paper, we will study whether random negative sampling, positive sampling and similarity sampling will increase the performance of AL analysis.

Active learning with random negative sampling converges to the same optimal classifier as active learning

In our AL analysis, the absence of manual labeling reduces the expense involved with negative random negative sampling, but a small fraction of mislabeled negative samples requires correction to avoid classifier bias. Through the iterative AL process, we expect the asymptotic reduction of this bias to zero as the sample size grows. However, this random negative sampling scheme is beyond the scope of the current AL framework[30–31], which allows no mislabeled samples. Here is a heuristic proof to clarify the convergence of AL with negative random negative sampling to the same AL optimal classifier.

Let us use a similar notion to that of Balcan and Long[32]. We assume that the data points (x, y) are drawn from an unknown underlying distribution D_{XY} over $X \times Y$. X is called the feature space (e.g. word frequencies in abstracts), and Y is the abstract label. Here, $Y = \{\pm 1\}$ and $X = \mathbb{R}^d$, and d is the dimension. Without loss of generality, we further assume that the feature space X is centralized in 0 after linear transformation. Let $\mathbb{C}$ be a class of linear classifiers through the origin, that is $\mathbb{C} = \left\{ sign\,(w \bullet x) : w \in \mathbb{R}^d, \|w\| = 1 \right\}$. In an AL, the goal is to identify a classifier $w \in \mathbb{C}$ of small misclassification error, where $err\,(w) = P_{(x,y)\tilde{D}_{XY}}\,[sign(\,w \bullet x) \neq y]$. Balcan and Long showed that with arbitrary small error ϵ and probability δ, an AL needs at most $O\left((d + \log(1/\delta) + loglog\,(1/\epsilon))\log(1/\epsilon)\right)$labeled samples to identify a classifier with misclassification error less thanϵ and probability higher than $1 - \delta$. This AL theory requires no misclassification error among sample labels.

In the unscreened sample pool, let us assume the mislabeling negative sample size,$n_{\mp}$, is much smaller than negative samples in random negative sampling N_{-}, i.e. $n_{\mp} \ll N_{-}$. The true positive samples in the training set,N_{+} is also smaller than N_{-}. Therefore, the error rate before the AL classifier is approximated in Eq. (1). Using the AL classifier, there will be $n_{\mp} \times N_{+}/(N_{-}+N_{+})$ mislabeled negative samples predicted to be positive, and their labels will be calibrated through the manual label in the AL. After AL calibration, the error rate of will be reduced to Eq. (2).

Error rate before AL

(1) ; *Error rate after AL*. (2)

Practically, considering, $n_{\mp} = N_{-}/1000$ and $N_{+} = N_{-}/4$. The error is $\epsilon + 0.001$ before AL, and $\epsilon + 0.001/5$ after one step AL calibration, and $\epsilon + 0.001/5^{m}$ after m steps. Therefore, the misclassification error due to the mislabeled data will go to zero extremely fast. This heuristic proof has not yet considered the complications such as nonlinear classified, general log-concave distributions, and inseparable positive and negative data in $\mathbb{R}^{d}$. Existing AL theories [32–33] have shown and supported that error ϵ holds with required $((d+\log(1/\delta)+loglog\,(1/\epsilon))\log(1/\epsilon))$labeled samples under these conditions. We, however, will use the similar argument to show that the mislabeled error $n_{\mp}/(N_{-}+N_{+})$ will become small after a number of AL steps.

Positive sampling increases the speed of AL optimization when sample population is overwhelming negative

Following the same annotation, ϵ is the prespecified misclassification error. Collecting positively labeled samples is not an easy task using uncertainty sampling alone when sample population is overwhelming negative. Here, population positive sample size N_{+} is significantly smaller than population negative sample size, N_{-}. The misclassification error rate of negative samples is $\epsilon \times \frac{N_{-}}{N_{-}+N_{+}}$, while the misclassification error of positive samples is $\epsilon \times \frac{N_{+}}{N_{-}+N_{+}}$. Given a positive sample in the sample pool available for selection, uncertainty sampling focuses on misclassified samples, and it has a $\epsilon \times \frac{N_{+}}{N_{-}+N_{+}}$ chance in selecting this positive sample. On the hand, in positive sampling, top α, a percentage, positively predicted samples will be selected. Hence, positive samples have $\left(1-\epsilon \times \frac{N_{+}}{N_{-}+N_{+}}\right) \times \alpha \cong \alpha$ chance to be selected, since$N_{-} \gg N_{+}$. Therefore, positive sampling will have $\alpha/(\epsilon \times \frac{N_{+}}{N_{-}+N_{+}}) = \frac{\alpha}{\epsilon} \times \frac{N_{-}+N_{+}}{N_{+}}$ times higher probability in selecting the positive samples than uncertainty sampling. Practically, considering N_{+} = 50,000 positive DDI or PG related abstracts, and N_{-} = 25,000,000 negative abstracts in PubMed, a misclassification rate $\epsilon = 0.20$, and top $\alpha = 20\%$ positively predicted samples are selected, positive sampling has $\frac{\alpha}{\epsilon} \times \frac{\mathrm{N_{-}+N_{+}}}{\mathrm{N_{+}}} = 501$ times higher chance in selecting this positive sample in AL than uncertainty sampling does.

Active learning implementation in multiple sampling schemes **in the unscreened sample pool (**Fig. 1**)**

- Random negative sampling and datasets: According to the random negative sampling scheme, the initial training set contains 100 random negative samples from unscreened sample pool and 100 labeled positive samples from screened sample pool. Machine learning model ML1 is trained out. While the initial external validation set is made of 50 labeled negative samples, 50 positive samples and 50 random negative samples.
- Uncertainty sampling, positive sampling, and similarity sampling: to predict the unlabeled samples in sample pool, a random subset (100 samples) with the low confidence samples (uncertainty sampling) and high confidence positively predicted samples (positive sampling) are collected from the unscreened sample pool. In the meantime, combined with the similarity values these extracted samples are similar with the samples in corpus, the top 20 samples with high similarity are extracted and manually reviewed (similarity sampling).
- Updating training and validation sets: the reviewed and labeled samples from previous multiple sampling processing are divided and distributed equally into the initial training and external validation data sets. The new training set and external validation set for next round are produced.
- Re-training: Using the updated training set, ML_1 is re-trained, and the multiple sampling scheme is applied again. Totally, four iterations are performed in active learning analysis.
- Performance evaluation: The performance of ML_1 from all rounds of AL analysis are evaluated using the updated external validation data set.

Active learning implementation with multiple sampling schemes **in the screened sample pool (**Fig. 1**)**

- Datasets: The initial training set contains 100 positive samples and 100 negative samples. Machine learning model ML_2 is trained out. While the initial external validation set is made of 50 labeled negative samples, 50 positive samples and 50 random negative samples.
- Uncertainty sampling and similarity sampling: Due to AL in screened sample pool uses labeled samples as training sets, only uncertainty sampling and similarity sampling are applied in screened sample pool. After predicting the unlabeled samples in screened sample pool, a random subset (100 samples) with the low confidence samples (uncertainty sampling) are collected. Then, combined with the similarity value that extracted samples are similar with the samples in corpus, the top 20 samples with high similarity are extracted and manually reviewed (similarity sampling).
- Updating training and validation sets: the reviewed and labeled samples from previous multiple sampling processing are divided and distributed equally into the initial training and external validation data sets. The new training set and external validation set for next round are produced.
- Re-training: Using the updated training set, ML_2 is re-trained, and the multiple sampling scheme is applied again. Totally, four iterations are performed in active learning analysis.
- Performance evaluation: The performance of ML_2 from four rounds are evaluated using the updated external validation data set.

Data preprocessing

All the abstracts are processed after downloading from PubMed. They are parsed with desired content (titles and abstracts), and are converted into GENIA format. Multiple abstract files are saved as text format in a folder. After going through Lowercase converting and StopwordsTokenizer, a Doc object for each file consisting of the text split on single space characters is transformed by basic whitespace tokenizer. This Doc is to produce to tokens that feed into models.

Machine learning and deep learning analyses

Supporting vector machine (SVM) is used as the traditional machine learning method in AL. The appearance frequency of terms from the Doc followed Poisson distribution and was represented as a categorical term-document occurrence matrix based on the word count. The terms with low frequency SDs were considered to lack useful information and specificity. Therefore, the terms with frequency SD > 0.03 were selected as features and used to train models.

FastText [34–35] is used as a relatively simple deep learning (DL) algorithm in AL analysis. We utilize the “torch” module for text mining package in python. FastText is a multi-step approach for text classification (Fig. 2).

- Input layer: It is a document consisting of words, for example, “loratadine", " increases", " the", " myopathy", " risk", " of ", "simvastatin”.
- Embedding layer: It maps words and the character N-grams (N = 2) into embedding vectors by looking up the hashed dictionary according to the global vectors (GloVe). The input words and N-grams are represented as an array that would be taken as input and extract the features.
- Pooling layer: A fixed-length vector by performing feature selection, the pooling layer performs element-wise averaging over all the word embeddings, followed by the output layer.
- Softmax regression: The sigmoid function $\varphi(z) = \frac{1}{1+e^{-z}}$ is used to formulate the prediction probability for an abstract: DDI positive and DDI negative.

Performance evaluation

DDI IR AL analysis is evaluated using the following evaluation matrices: Precision (P) = TP/(TP + FP), Recall (R) = TP/(TP + FN), and the F1-score = (2*P*R)/(P + R). P is reported when R is set as 0.95. This pre-specified high recall rate serves the purpose that we will miss only a small fraction of DDI relevant paper, i.e. 0.05, in our DDI IR analysis.

Results

Random negative sampling plays an effective role in unscreened sample pool

Random negative sampling expects that DDI-related abstracts are only a very small fraction of unscreened sample pool. According to the distribution of positive and negative samples in DDI corpus, 1,000 samples are randomly selected from unscreened sample pool and on displayed. As Fig. 3 shown, PCA analysis tells the distribution of most random samples from unscreened pools are the same as the negative samples' distribution. It further illustrates that most samples in the unscreened sample pool are non-DDI related abstracts from the perspective of clustering, which makes random negative sampling more reasonable and efficient in unscreened sample pool.

Negatively labeled abstracts are different between screened sample pool and unscreened sample pool

Between screened and unscreened samples pools, we found the distribution of the negatively labeled samples are different. Using t-SNE (t-Distributed Stochastic Neighbor Embedding) visualization, it maps the high-dimensional data of abstracts to a lower dimensional space. We randomly select 300 samples in each of the two pools. Based on the data preprocessing and word embedding of the high dimensional characteristics they have, t-SNE reduced to 2 dimensions. Using the top two dimensions of t-SNE analysis, Fig. 4 shows two distributions of negatively labeled samples between the screened sample pool and the unscreened sample pool. The color in the contour plots represent the local density of the samples, and darker colored areas indicates higher density. Apparently, screened samples and unscreened samples have differently distributed negative samples. Therefore, this observation suggests that different classifiers are needed between screened sample pool and unscreened sample pool.

Stratified AL analysis identifies more DDI relevant abstract than DDI IR analysis using only screened sample pool

PubMed comprises more than 32 million citations for biomedical literature, through the query [“drug interaction” AND (Type of Study)] and [“drug combination” AND (Type of Study)], totally 142,520 relevant literature was obtained from the year 1956 to 2021. To further verify that DDI relevant abstracts belong to very few of them, we random selected 1,000 samples from screened sample pool (pass the query) and unscreened sample pool. After manual reviewing and labeling, 25 out of 1,000 samples in screened sample pool are positive, and only 1 out of 1,220 are positive for the unscreened sample pool. It preliminary estimated that DDI relevant abstracts (positive samples) are 2.5% and 0.1% in screened and unscreened pool, respectively. Therefore, the estimated fraction of DDI relevant abstracts in two pools is about 3,563 and 26,230. Therefore, if we just use the samples in screened sample pool, we will miss potentially a large number of positive abstracts in the unscreened pool.

Multiple sampling schemes improve the performance of AL analysis

SVM is used as the machine learning method for AL with multiple sampling schemes in screened sample pool and unscreened sample pool. In AL analysis, recall rate is all pre-specified at 0.95.

- **Screened sample pool Fig. 5A** compares the performance of traditional uncertainty sampling AL (Un) and AL with uncertainty sampling and similarity sampling (UnS). When recall is set as 0.95, Un keeps increasing precision from 0.75 to 0.90 from round 1 to 3 in AL analysis, until the precision performance drops in round 4. UnS, on the other hand, consistently improves the precision from round 1 to 4, from 0.86 to 0.92. This analysis demonstrates that the UnS has more steady and significant improvement of AL performance than the traditional Un method.
- **Unscreened sample pool Fig. 5B** presents the performance of three sampling schemes AL: uncertainty sampling + random negative sampling (UnR); uncertainty sampling + random negative sampling + positive sampling (UnRP); and uncertainty sampling + random negative sampling + positive sampling + similarity sampling (UnRPS). In round 1, when the recall is set at 0.95, UnRPS, UnRP, UnR have precisions of (0.75, 0.74, 0.63), respectively. Both UnRPS and UnRP out-perform UnR. UnR's precision increases to 0.73 from round 1 to 3, but drops in round 4. However, both UnRPS and UnRP keeps stable increases in precision from round 1 to 4, and finally UnRPS has the best precision at 0.81. This analysis suggests that combined uncertainty sampling, random negative sampling, and similarity sampling leads to the best performance.

Deep learning method out-performs the machine learning method in AL analysis

The performance of the embedding-based deep learning algorithm (FastText) is compared to SVM in AL analysis. Similar to the previous analysis, the recall rate is set as 0.95. The precisions are analyzed and reported in separate AL analyses from screened samples and unscreened samples.

- **Screened sample pool** Using FastText, at the beginning, i.e. round 1, FastText with UnS reaches a precision 0.90 already. It out performs FastText with Un (precision = 0.86), SVM with UnS (precision = 0.86) and SVM with Un (precision = 0.75). During AL process, FastText with either Un and UnS sampling scheme improve the precision from round 1 to 4, though Un shows a larger variation than UnS. At the end, FastText with UnS has the best precision = 0.96. These trends are shown in **Fig. 5A**. These data suggests that FastText, a DL method, has improved AL performance than SVM.
- **Unscreened sample pool** The performance of FastText in AL with multiple sampling schemes are compared to SVM in unscreened sample pool (Fig. 5B). At the baseline, i.e. round 1, FastText with UnRPS or UnRP have the comparable best performance, precision = 0.80 and 0.81, respectively. Their precision steadily improve from round 1 to 4, and reach to 0.90 and 0.88, respectively. These numbers are noticeably higher than those from SVM method with multiple sampling schemes.

Discussion

This study performed a comprehensive investigation on how various sampling schemes and machine learning algorithms improve AL for DDI IR analysis from literature. This is also the first time that AL is studied for its performance in DDI IR analysis. DDI IR analysis from PubMed abstracts faces the challenges of relatively small positive DDI samples and overwhelmingly large negative samples. New sampling schemes, including random negative sampling and positive sampling, are purposely designed to address these challenges. They reduce annotation labor and improve the efficiency of AL analysis. The theoretical consistency of random negative sampling and positive sampling is also shown in the paper.

Practically, PubMed abstracts are divided into two pools. Screened pool contains all abstracts that pass the DDI keywords query in PubMed, while unscreened pool includes all the other abstracts. Our preliminary analysis reveals that the unscreened pool contains seven times more DDI related abstracts, 26,230, than the screened pool, 3563. This shows that we cannot only rely on PubMed query in retrieve DDI related abstracts.

At a prespecified recall rate of 0.95, DDI IR analysis performance is evaluated and compared in precision. In screened pool IR analysis using supporting vector machine (SVM), similarity sampling plus uncertainty sampling improves the precision of AL over uncertainty sampling, from 0.89 to 0.92 respectively. In the unscreened pool IR analysis, the integrated random negative sampling, positive sampling, and similarity sampling improve the IR analysis performance over uncertainty sampling along, from 0.72 to 0.81 respectively. When we change the SVM to a deep learning method, all sampling schemes consistently benefit DDI AL analysis in both screened pool and unscreened pool. Deep learning also has significant improvement of precision over SVM, 0.96 vs 0.91 in screened pool, and 0.90 vs 0.81 in the unscreened pool, respectively. Please note that the recall is all set 0.95 for all occasions in our IR analysis. The 0.96 and 0.90 precision performance are extraordinary.

Random negative sampling and positive sampling are effective methods in improving AL analysis when a sample pool is dominated with negative samples. In our DDI IR analysis, they effectively reduce the annotation workload, and improve the IR analysis performance. We believe these two sampling schemes are equally effective to other NLP applications where the positive and negative samples are imbalanced.

Conclusion

This paper developed multiple sampling schemes and deep learning algorithms, and implemented them in the active learning (AL). This is the first time that AL is developed to preform drug-drug interaction information retrieval (DDI IR) analysis. The superior performance of deep learning to the conventional machine learning approaches is a major conclusion in AL DDI IR analysis. We further demonstrate that both positive sampling and random negative sampling schemes are highly effective sampling scheme in AL analysis, when positive samples are extremely small and negative samples are overwhelmingly large.

Abbreviations

DDI: drug-drug interaction; AL: active learning; ADEs: adverse drug events; ML: machine learning; DL: deep learning; IR: information retrieval; IE: information extraction; PK: pharmacokinetics; PE: pharmaco-epidemiology; SVM: supporting vector machine.

Declarations

Acknowledgements

Not applicable.

Authors' contributions

LL and WX provide paper ideas and designed the study planning, as well as writing the manuscript. LL and SJ were involved in data curation and statistics. KJ and WX were major contributors in programming. All authors read and approved the final manuscript.

Availability of data and materials

The datasets generated and analyzed during the current study are available in GitHub at https://github.com/zha204/Deep-AL-for-DDI-IR.

Funding

This work was supported by the National Institutes of Health grants (U01 CA248240, P30 HD106451 and R01 LM011945).

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

All authors declare no conflict of interest.

References

1. Patel PS, Rana DA, Suthar JV, Malhotra SD, Patel VJ. A study of potential adverse drug-drug interactions among prescribed drugs in the medicine outpatient department of a tertiary care teaching hospital. J Basic Clin Pharm. 2014; 5:44-8. doi: 10.4103/0976-0105.134983
2. Percha B, Altman RB. Informatics confronts drug-drug interactions. Trends Pharmacol Sci. 2013;34:178-84. doi: 10.1016/j.tips.2013.01.006
3. Budnitz DS, Pollock DA, Weidenbach KN, Mendelsohn AB, Schroeder TJ, Annest JL. National surveillance of emergency department visits for outpatient adverse drug events. JAMA. 2006;296:1858-66. doi: 10.1001/jama. 296.15.1858
4. Dechanont S, Maphanta S, Butthum B, Kongkaew C.Hospital admissions/visits associated with drug-drug interactions: a systematic review and meta-analysis. Pharmacoepidem Dr S. 2014;23(5):489-497.
5. Magro L, Moretti U, Leone R.Epidemiology and characteristics of adverse drug reactions caused by drug-drug interactions. Expert Opin Drug Saf. 2012;11(1):83-94.
6. Maher RL, Hanlon J, Hajjar ER. Clinical consequences of polypharmacy in elderly. Expert Opin Drug Saf. 2014; 13(1):57-65.
7. Bourgeois FT, Shannon MW, Valim C, Mandl KD. Adverse drug events in the outpatient setting: an 11-year national analysis. Pharmacoepidemiol Drug Saf. 2010;19:901-10. doi: 10.1002/pds.1984
8. Bjornsson TD, Callaghan JT, Einolf HJ, Fischer V, Gan L, Grimm S, et al. The conduct of in vitro and in vivo drug-drug interaction studies: a Pharmaceutical Research and Manufacturers of America (PhRMA) perspective. Drug Metab Dispos. 2003;31(7):815-32. doi: 10.1124/dmd.31.7.815.
9. Bjornsson TD, Callaghan JT, Einolf HJ, Fischer V, Gan L, Grimm S, et al. The conduct of in vitro and in vivo drug-drug interaction studies: a PhRMA perspective. J Clin Pharmacol. 2003;43(5):443-69.
10. Hennessy S, Leonard CE, Gagne JJ, Flory JH, Han X, Brensinger CM, et al. Pharmacoepidemiologic Methods for Studying the Health Effects of Drug-Drug Interactions. Clin Pharmacol Ther. 2016;99(1):92-100. doi: 10.1002/cpt.277.
11. Burns H, Russell L, Cox ZL. Statin-induced rhabdomyolysis from azithromycin interaction in a patient with heterozygous SLCO1B1 polymorphism. J Clin Pharm Ther. 2021;46(3):853-855. doi: 10.1111/jcpt.13327.
12. De Luca M, Iacono O, Lucci R, Guardasole V, Bosso G, Cittadini A, et al. Atorvastatin-linked rhabdomyolysis caused by the simultaneous intake of amoxicillin clavulanic acid. J Basic Clin Physiol Pharmacol. 2021;32(1): 2020-0108.
13. Humphrey RW, Brockway-Lunardi LM, Bonk DT, Dohoney KM, Doroshow JH, Meech SJ, et al. Opportunities and challenges in the development of experimental drug combinations for cancer. J Natl Cancer Inst. 2011;103(16):1222-6. doi: 10.1093/jnci/djr246.
14. Martin Krallinger, Obdulia Rabal, Analia Lourenc´ o, Julen Oyarzabal, Alfonso Valencia. Information Retrieval and Text Mining Technologies for Chemistry. Chem. Rev. 2017;117(12):7673-

7761. https://doi.org/10.1021/acs.chemrev. 6b00851

15. Wu HY, Chiang CW, Li L. Text mining for drug-drug interaction. Methods Mol Biol. 2014; 1159:47-75. doi: 10.1007/978-1-4939-0709-0_4.
16. Zhao Z, Yang Z, Luo L, Lin H, Wang J. Drug drug interaction extraction from biomedical literature using syntax convolutional neural network. Bioinformatics. 2016;32:3444-53. doi: 10.1093/bioinformatics/btw486
17. Zheng W, Lin H, Luo L, Zhao Z, Li Z, Zhang Y, et al. An attention-based effective neural model for drug-drug interactions extraction. BMC Bioinformatics. 2017;18:445. doi: 10.1186/s12859-017-1855-x
18. Zhang Y, Zheng W, Lin H, Wang J, Yang Z, Dumontier M. Drug-drug interaction extraction via hierarchical RNNs on sequence and shortest dependency paths. Bioinformatics. 2018;34:828-835. doi: 10.1093/bioinformatics/btx659
19. Wang W, Yang X, Yang C, Guo X, Zhang X, Wu C. Dependency-based long short term memory network for drug-drug interaction extraction. BMC Bioinformatics. 2017;18:578. doi: 10.1186/s12859-017-1962-8
20. Zhang Y, Lin H, Yang Z, Wang J, Zhang S, Sun Y, et al. A hybrid model based on neural networks for biomedical relation extraction. J Biomed Inform. 2018;81:83-92. doi: 10.1016/j.jbi.2018.03.011.
21. Kolchinsky A, Lourenço A, Wu HY, Li L, Rocha LM. Extraction of pharmacokinetic evidence of drug-drug interactions from the literature. PLoS One. 2015;10(5):e0122199. doi: 10.1371/journal.pone.0122199.
22. Wu HY, Karnik S, Subhadarshini A, Wang Z, Philips S, Han X, et al. An integrated pharmacokinetics ontology and corpus for text mining. BMC Bioinformatics. 2013;14:35. doi: 10.1186/1471-2105-14-35.
23. Burr Settles, Mark Craven, Lewis Friedland. Active learning with real annotation costs. Appears in Proceedings of the NIP Workshop on Cost-senstitive learning. 2008
24. Mahnoosh Kholghi, Laurianne Sitbon, Guido Zuccon, Anthony Nguyen. Active learning reduces annotation time for clinical concept extraction. Int J Med Inform. 2017; 106:25-31. doi: 10.1016/j.ijmedinf.2017.08.001.
25. Lewis, DD, Gale, WA. A sequential algorithm for training text classifier. SIGIR. 1994;3-12. doi:10.1007/978-1-4471-2099-5_1
26. Culotta A, McCallum A. Reducing labeling effort for structured prediction tasks. In: AAAI: 2005;2:746-751.
27. Xie WX, ZHANG SJ, Li L. An Integrated Repository of Drug Interaction Corpora. Manuscript.
28. Xie WX, Wang LM, Cheng Q, Wang XY, Wang Y, Bi HY, et al. Integrated random negative sampling and uncertainty sampling in active learning improve clinical drug safety drug-drug interaction information retrieval. Front. Pharmacol.2021;11:582470. https://doi.org/10.3389/ fphar. 2020. 582470

29. Salton, G. and Buckley, C. Term-weighting approaches in automatic text retrieval. Information Processing & Management.1988;24 (5): 513-523.
30. Yang L, Zhang Y, Chen J, Zhang S, Chen DZ. Suggestive Annotation: A Deep Active Learning Framework for Biomedical Image Segmentation. MICCAI 2017. 2017;10435. https://doi.org/10.1007/978-3-319-66179-7_46
31. Hanneke S. Rates of convergence in active learning. The Annals of Statistics. 2011;39(1):333-361.
32. Balcan MF, Long P. Active and passive learning of linear separators under log-concave distributions. In: Conference on Learning Theory. 2013;288-316.
33. Balcan MF, Broder A, Zhang T. Margin based active learning. In: International Conference on Computational Learning Theory: 2007; 35-50.
34. Piotr Bojanowski, Edouard Grave, Armand Joulin, Tomas Mikolov. Enriching Word Vectors with Subword Information. Transactions of Association for computational linguistics. 2017; 5:135-146.
35. Armand Joulin, Edouard Grave, Piotr Bojanowski, Tomas Mikolov. Bag of Tricks for Efficient Text Classification. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. 2017;427-431.

Figures

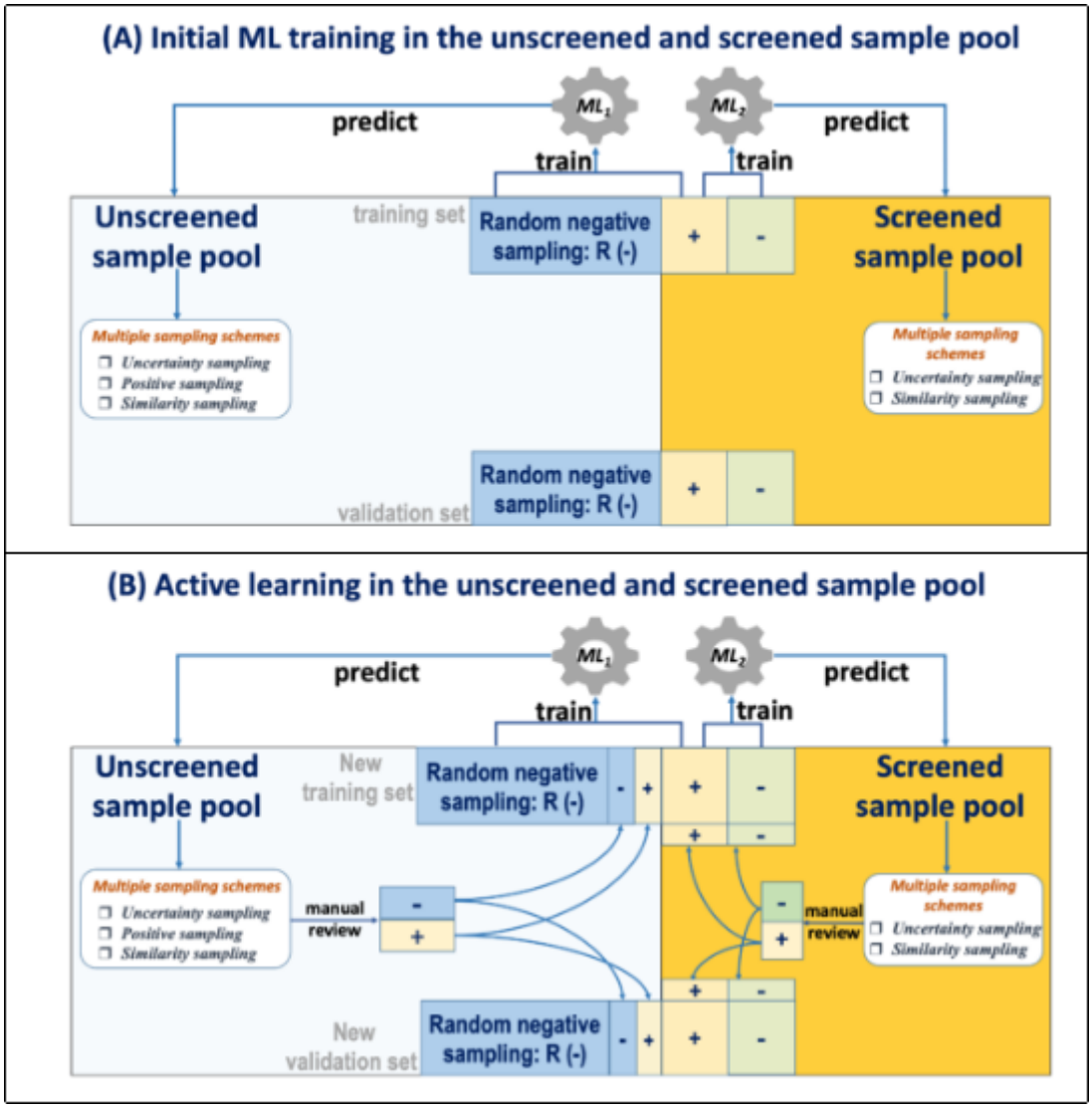


Figure 1

Stratified active learning with multiple sampling schemes

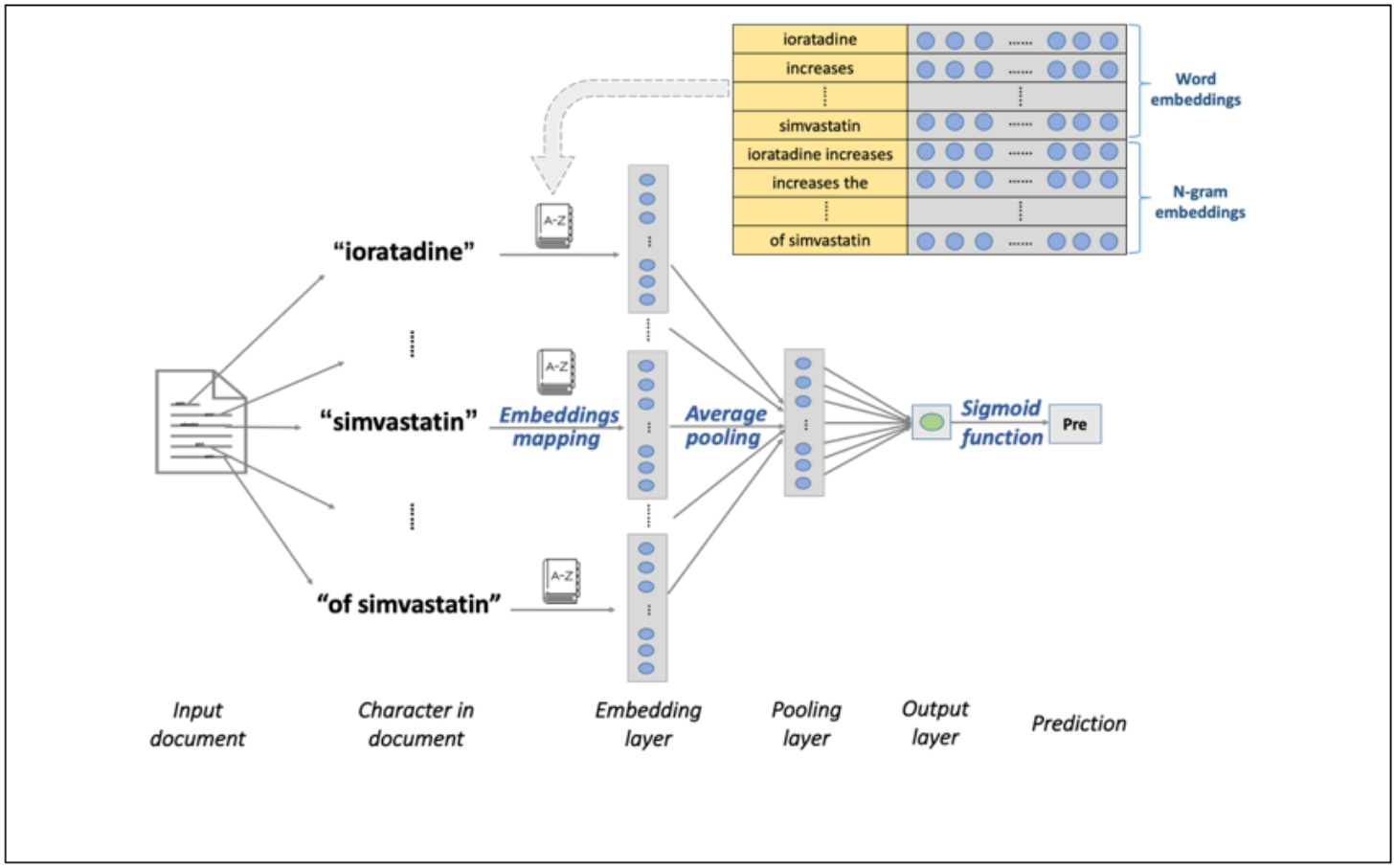


Figure 2

FastText scheme

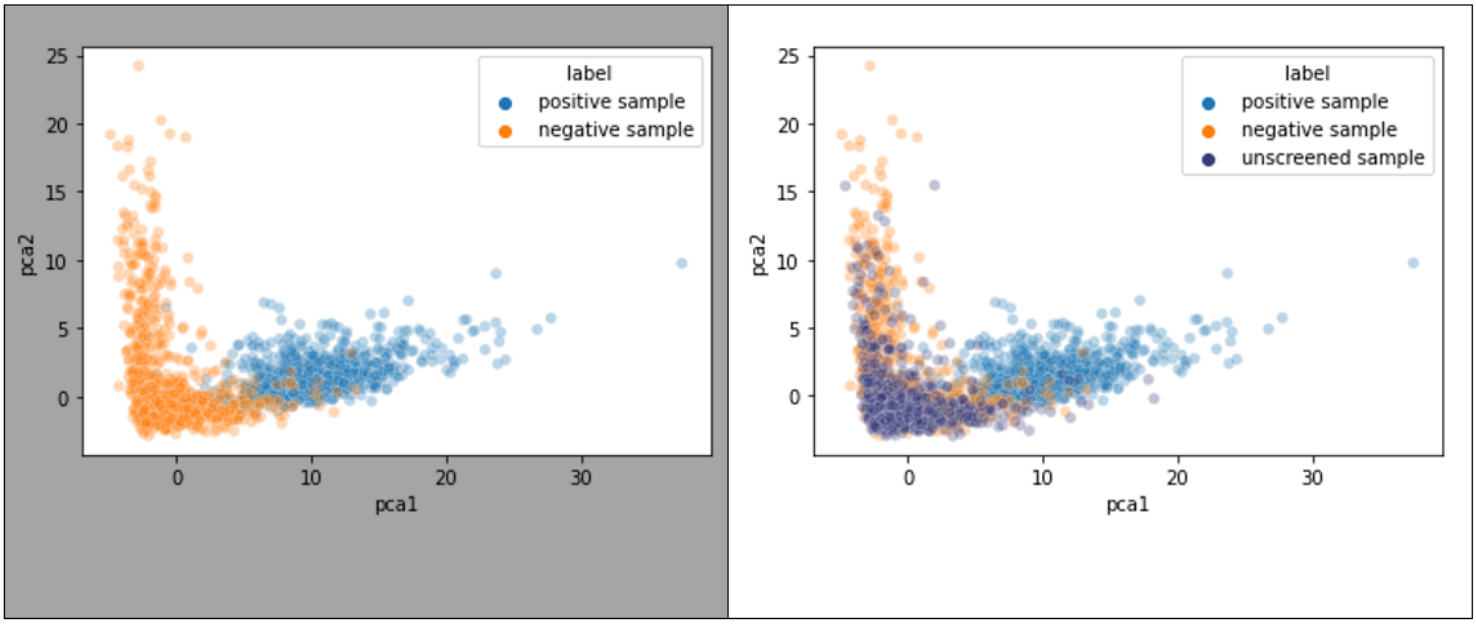


Figure 3

Positive and negative samples distribution in screened and unscreened sample pool

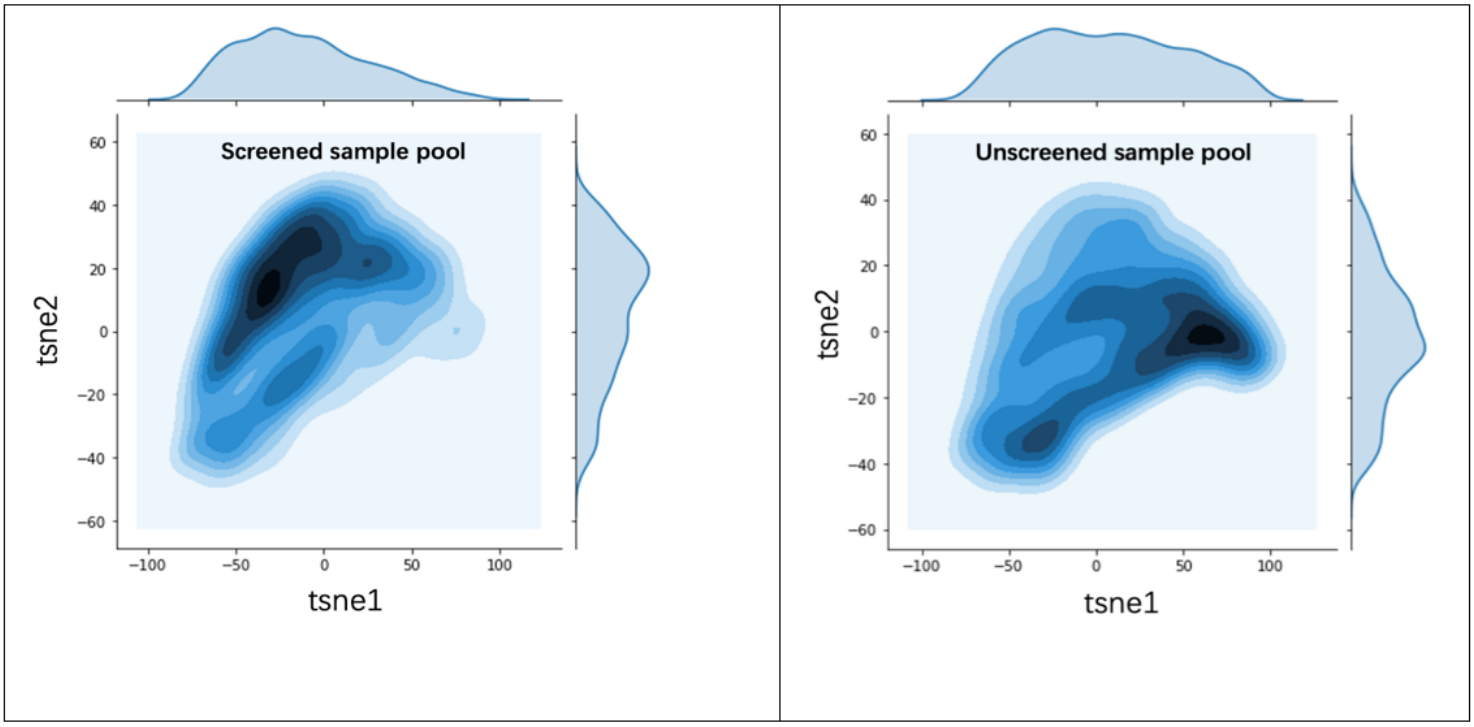
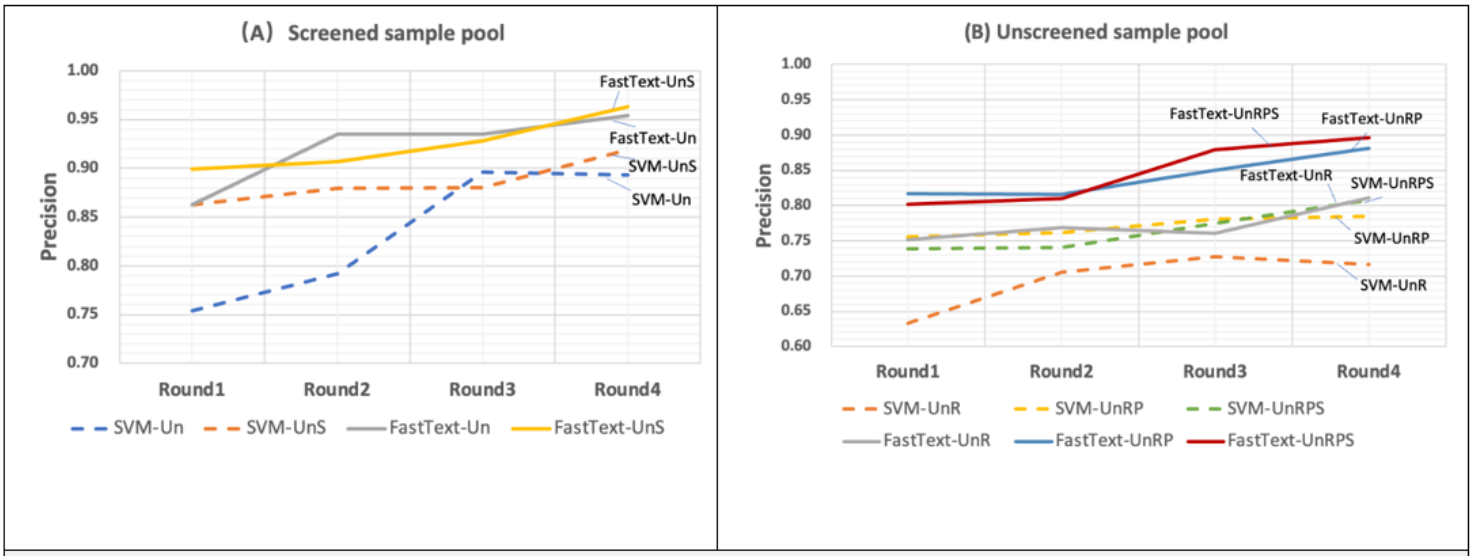


Figure 4

t-SNE analysis for two sample pools



Notes: *Precision in the figure presents the precision value when recall = 0.95.*

Un: uncertainty sampling; UnS: uncertainty sampling + similarity sampling; UnR: uncertainty sampling + random negative sampling; UnRP: uncertainty sampling + random negative sampling + positive sampling; UnRPS: uncertainty sampling + random negative sampling + positive sampling + similarity sampling.

Figure 5

Performance of Multiple sampling in screened and unscreened sample pools