Neural networks for open and closed Literature-based Discovery

Gamal Crichton *, **Simon Baker, Yufan Guo, Anna Korhonen**

Language Technology Laboratory, TAL, University of Cambridge, Cambridge, United Kingdom

* gkoc2@cam.ac.uk

Abstract

Literature-based Discovery (LBD) aims to discover new knowledge automatically from large collections of literature. Scientific literature is growing at an exponential rate, making it difficult for researchers to stay current in their discipline and easy to miss knowledge necessary to advance their research. LBD can facilitate hypothesis testing and generation and thus accelerate scientific progress. Neural networks have demonstrated improved performance on LBD-related tasks but are yet to be applied to it. We propose four graph-based, neural network methods to perform open and closed LBD. We compared our methods with those used by the state-of-the-art LION LBD system on the same evaluations to replicate recently published findings in cancer biology. We also applied them to a time-sliced dataset of human-curated peer-reviewed biological interactions. These evaluations and the metrics they employ represent performance on real-world knowledge advances and are thus robust indicators of approach efficacy. In the first experiments, our best methods performed 2-4 times better than the baselines in closed discovery and 2-3 times better in open discovery. In the second, our best methods performed almost 2 times better than the baselines in open discovery. These results are strong indications that neural LBD is potentially a very effective approach for generating new scientific discoveries from existing literature. The code for our models and other information can be found at: https://github.com/cambridgeltl/nn_for_LBD.

Introduction

Literature-based Discovery (LBD) aims to discover new knowledge by connecting information which have been explicitly stated in literature to deduce connections which have not been explicitly stated. Its pioneer is Don Swanson who hypothesised that the combination of two separately published results indicating an A-B relationship and a B-C relationship are evidence of an A-C relationship which is unknown or unexplored. He used this to propose fish oil as a treatment for Raynaud syndrome due to their shared relationship with blood viscosity [1]. This hypothesis was later shown to have merit in a prospective study [2] and he continually proposed other discoveries using similar methods [3–5]. LBD comes in two flavours: open and closed discovery. In open discovery, only the A is given and Cs are deduced using the various A-B-C relationships in existence while in closed discovery the A and C are given and the goal is to quantify the existence of relevant Bs.

LBD has evolved to involve using computers to discover many such connections automatically from large collections of literature. Thus, it can facilitate both hypothesis testing and generation to give tangible support to scientific research [6, 7]. Scientific literature is growing at an exponential rate [8], making it difficult for researchers to stay current in their discipline. This, along with the increasing necessity of researchers to specialize has led to an environment where discoveries in one area are not known outside of it [9] and valuable logical connections between disparate bodies of knowledge remain unnoticed [10]. This means there is a very real chance that knowledge which can be combined to form or crystallise breakthrough-inducing hypotheses are dispersed throughout the literature. LBD can help researchers to quickly discover and explore hypotheses as well as gain information on relevant advances inside and outside of their niches and increase interdisciplinary information sharing. Thus as the scientific literature grows, the necessity for LBD as a research tool increases.

LBD has already been used to identify new connections between biomedical entities and new candidate genes and treatments for illnesses [6] and to propose treatments for Parkinson's Disease and Multiple Sclerosis [11, 12]. It has seen use in drug development and repurposing [13, 14] as well as predicting adverse drug reactions [15, 16]. It has also been used to propose new potential cancer treatments [17] and identify promising research collaborations [18].

The recently-released LION LBD system [19] reports state-of-the-art results in LBD. It uses PubTator [20] for annotating PubMed scientific articles with concepts such as chemicals, genes/proteins, mutations, diseases and species; as well as sentence-level annotation of cancer hallmarks that describe fundamental cancer processes and behaviour [21]. It uses co-occurrence metrics to rank relations between concepts and performs both open and closed discovery.

Neural networks have been successful in related tasks such as Knowledge Discovery and Natural Language Process (NLP) in recent years. Whether they can be used to give improved results in LBD is unexplored (except for recent exploratory work by [22]). In this paper we make two main contributions: four graph-based neural approaches to LBD; and evaluations of them on two real-world biomedical datasets using informative metrics. These datasets tested their ability to rank future published biomedical discoveries: one is the Cancer Discovery dataset used by [19] and the other consisting of human-verified, peer-reviewed biomedical interactions.

Related work

Literature-based Discovery (LBD)

LBD seeks to discover previously unknown associations or hidden links between pieces of existing knowledge by analysing literature in an automated or semi-automated way using various computational approaches and algorithms [23, 24]. It has mostly been deployed in the biomedical domain, but it has also been used outside of it as it has been applied to research into developing water purification systems, accelerating development of developing countries and identifying promising research collaborations [18, 25, 26].

[1] defined the most basic and widespread type of LBD, called the ABC paradigm because it centres around three concepts called A, B and C (e.g. [27–29]). It states that if there is a connection between A and B and one between B and C, then there is one between A and C which, if not explicitly stated, is yet to be explored. The ABC paradigm has two types: *open* and *closed* discovery.

In open discovery, only A is given. The approach finds Bs and uses them to return possibly interesting Cs to the user, thus *generating hypotheses* from A. With closed discovery, the A and

C are given to the approach which seeks to find the Bs which can link the two, thus *testing a hypothesis* about A and C.

[30] distinguishes between traditional approaches and 'emergent' paradigms which will define the field in the future (e.g. [31, 32]). One of the characteristics of these is their use of techniques borrowed from other research fields including link prediction on graphs and machine learning which offer different approaches to LBD and address its problems. This work provides a blend by using the ABC paradigm but harnessing machine learning models inspired by link prediction on graphs.

Evaluating LBD systems

Evaluation is difficult in LBD for several reasons: disagreement about the role of LBD systems in research and thus what makes a successful one; difficulty in determining how useful, interesting or actionable a discovery is; and difficulty in objectively defining a 'discovery', which hinders the creation of a standard evaluation set which quantifies when a discovery has been replicated or found. Nonetheless, several methods have been employed in previous work.

A popular methods used in LBD is to *replicate previous discoveries* [28, 33, 34]. These are usually LBD-based discoveries as they are relatively easy to quantify compared to other discoveries. This means that there are only a handful of such discoveries and there is a danger of designing approaches which are tuned to perform well on these discoveries but do not generalise. In this evaluation, the literature before the discovery to be replicated is used to generate a ranked list of discovery candidates as target or linking terms. Success is measured by reporting the rank of the term(s) of interest; the higher the rank, the better the approach.

Literature- or time-slicing involves splitting the existing literature at a point in time. The approach is then exposed to the literature before the split and is evaluated by how many of the discoveries in the later period it can discover. Unclear definition of a discovery and an inability to determine if a discovery is wrong or simply new are critiques of this approach. In the absence of a perfect gold standard, this approach estimates it by finding instances of the defined relationships in the test set which are not in the training set and can be reasonably inferred from it. This means that evaluation depends on what constitutes a relationship for the given dataset. If a noisy relationship is used, the evaluation will be easy to perform well on. Previous systems have used term co-occurrences [35], relationships from external biomedical resources (e.g SemMedDB) [32] and semantic relationships [36]. A high precision approach would be to get expert opinion to generate the gold standard [37], but this is time-consuming, expensive and tends to produce low recall rates.

The advantage of this evaluation is that it produces an indicator of an approach's performance on a large number of test instances. This raises the need for evaluation metrics which can quantify performance on large, ranked lists. LBD works have used metrics popular in Information Retrieval [38] which include Precision, Recall, Area Under the Curve (AUC), Precision at k, Mean Average Precision (MAP) etc.

Proposing new discoveries or treatments goes beyond replicating past discoveries or predicting time-sliced instances of a particular relationship and shows that a system is capable of being used in realistic situations [13, 33, 39, 40]. This is usually accompanied by peer-reviewed publication in the domain or vetting by a domain expert.

Neural networks in the biomedical domain

While they are yet to be applied to LBD, the versatility of neural networks have been shown in their application to a broad range of biomedical tasks. They have been used to predict mental health conditions from tweets [41], recognise biomedical entities in text [42, 43], classify

hallmarks of cancer in text [44] and predict links representing Drug-Target Interactions (DTIs) and Protein-Protein Interactions (PPIs) in biomedical graphs [45]. On the biomedical image front, they have been used for classifying biomedical images [46, 47], segmenting 3D biomedical images [48] and segmenting and enhancing cardiac images [49].

An excellent recent overview of the use of neural networks in the biomedical domain is [50]. They point out that beyond the well-known applications to diagnosis, neural networks are increasingly being used to inform healthcare management decisions.

Node representations as embeddings

Graphs encode knowledge and can be processed to extract information which may not be easily seen. For machines to process them, graphs must be represented in a useable format, usually representing nodes as vectors of real numbers. Research on node representation devises methods which can create representations which preserve the original information in the graph. This information relates to the nature of the links and are classified as first or second (or higher) order proximity [51, 52]. Given two nodes, first order proximity is concerned with the strength of the direct link between them. Second order proximity compares their neighbourhoods and classes them as similar if their neighbourhoods are similar.

The quality of a method depends on its ability to preserve the proximities of a graph when creating representations. The node representations created by recent research represents each node as a vector in a space where similar nodes are located close to each other (node *embeddings*). There has been a proliferation of methods to create these node embeddings from graphs and it would be unwieldy to include all of them in this work. Comparisons between some of these can be found in [52]. We utilised a popular method whose implementation is freely available online, supported weighted edges and scaled to our large graphs: Large-scale Information Network Embedding (LINE) [51].

LINE explicitly defines two optimization functions to capture the structure of the graph. One captures first order proximity and the other captures second order proximity. [51] report that training their model with each setting then concatenating the outputs gives the best performance.

Materials and methods

Evaluation

Here we discuss the method of preparation of the datasets used for LBD and the metrics used for evaluation. The datasets contain information on the year that each link in the graphs was formed and the graphs were split by year of link formation for training and evaluation. The methods were given the earlier links and asked to predict later links.

Cancer case discoveries.   To facilitate direct comparison, we evaluate on the cases used in [19], which describes them at length. For completeness, we provide a summary. They are a set of five triples that represent specific recently-published discoveries (2011-2016) on the molecular biology of cancer that could have potentially been suggested by an LBD system in the past. They were selected and curated by cancer biologists. There are an additional five pairs of discoveries proposed by Swanson. The B connection was not simple in the Swanson cases so it was not possible to create triples to facilitate performing closed discovery on those cases. The details of these can be found in the Supplementary Document (S1 File) which accompanies this paper.

Each LBD approach is given a graph constructed only from literature up to five years before the publication date of the discovery and the model is then given the A and C nodes (in closed discovery) and asked to rank the B nodes. In open discovery it is given only the A node and

asked to rank all nodes within two hops (the C nodes). The approach's performance is quantified using the rank of the gold response in the returned list.

Time-slicing. The Cancer Discovery cases described above are strong evaluations for biomedical LBD systems because showing how a system would have ranked a discovery later published in a top-tier, peer-reviewed journal is a potent argument for its usefulness for LBD. However, the dataset is unsuitable for machine learning because it does not provide a development set to tune hyperparameters on; neither is it obvious how to create one. This, in addition to its limited size prompt the need for additional evaluation methods to gain a more generalised picture of performance of our approaches and models.

For this we choose a dataset which contained human-curated biomedical interactions which were published in peer-reviewed journals (details in "Datasets" Section). A graph created from the interactions in this dataset is time-sliced. From the post cut-off publication year, development and test sets are constructed. In some senses, this is not as stringent an evaluation and it is not possible to do closed discovery with it, but this provides robust additional evaluation of our open discovery approaches on a larger test set which is more indicative of approach generalizability.

Metrics. The evaluation metrics are important when analysing the performance of ranking systems. [19] reported median ranks over the groups of cases for the case discoveries. For comparability, we also report this along with the mean over the cancer and Swanson cases separately and combined.

For the time-sliced experiments, we additionally report MAP, Mean Reciprocal Rank (MRR) and Mean R-precision. There are 2 reasons for this: there is great variance between the amount of Cs which are ranked for each A so the mean rank can vary widely, distorting the results; and these metrics, especially the latter 2, give higher priority to correct scores ranked highly in the list, which is of importance in any ranking problem but especially so for LBD where investigating each proposal is a costly endeavour. Formal definitions of these evaluations are in Section 2 of the Supplementary Document (S1 File) which accompanies this paper.

Baselines

The baseline approaches are those used by [19]. They are 8 co-occurrence metrics accompanied by three aggregator functions and two accumulator functions (explained later in this section). We present a condensed version here for completeness (names in brackets are the shorthand they will be referred to going forward). More details can be found in the referred paper: Section 3.3 and full details in its Supplementary Information document.

- Co-occurrence count (Count): the number of sentences in which mentions of the entities connected by the edge co-occur.

- Document count (Doc-count): the number of documents in which mentions of the entities connected by the edge co-occur.

- Jaccard Index (Jaccard): the ratio of the size of the intersection over the size of the union of the sets of sentences in which the entities occur.

- Symmetric conditional probability (SCP): the product of the conditional probabilities of one entity being mentioned in a sentence where another occurs.

- Normalized pointwise mutual information (NPMI): a measure of the independence of the mention occurrence distributions,

- Chi-squared (χ^2), Student's *t*-test (*t*-test) and log-likelihood ratio (LLR) are statistical tests measuring whether the mention distributions are independent of each other.

A number of alternatives for the scoring functions operating over the edge weights have also been implemented. For the aggregation function *f(g)*, the alternatives *min*, *avg*, and *max* are used. These functions assign the score for a path the minimum, mean, and maximum respectively of the edge weights on the path. For the accumulation function *f(c)*, the choices *sum* and *max* are supported. When multiple paths lead to the same node, the former sums the path score to obtain the node score while the latter simply uses the maximum score.

We focus on only the best performing methods for the mean and median metrics and report the relevant accumulator and aggregator functions in each experiment.

Neural approaches

Two neural link prediction methods are used for closed discovery and another two for open discovery. All approaches use node embeddings created with LINE with weighted edges, where weights are calculated using Jaccard Index. The embeddings were induced with the portion of the graph used for training, the pre-cutoff year period. The settings used are in Section 3 of the Supplementary Document (S1 File).

For each of the approaches described here, five node combination methods are used to determine how the nodes which constitute the link path are combined for input into the model, so models ending in '-A' refer to approaches which use Average to do this, '-C'—Concatenation, '-H'—Hadamard, '-W1'- Weighted-L1 and '-W2'- Weighted-L2.

Closed discovery neural model and approaches. In both of these approaches the model is a Multi-Layer Perceptron (MLP) which was effective in the similar task of neural link prediction on biomedical graphs [45]. The model contains a single hidden layer with ReLU [53] activation which led to a final layer with Softplus activation to allow for unrestricted positive scores. The model is trained as a classifier with the Cross Entropy loss.

As we use the model from [45], it is necessary to distinguish that work from this one. That paper presents a neural architecture for classifying whether a link exists between 2 nodes using their node embeddings; such an approach is not ABC LBD as is the focus of this paper. To perform ABC LBD the path(s) between A and C must be taken into account, which link prediction as proposed in [45] is unconcerned with. In this paper, the paths are taken into account in 2 different ways as reflected in the 2 approaches to Closed Discovery whose descriptions follow.

CD-1: The neural model is used to provide a score for each A-B or B-C link in the path. The scores are then used in aggregator functions as the baseline methods, so the difference here is that a neural network produces a score for the link instead of using one of the metric calculations described in 'Baselines'.

CD-2: In this approach, A-B-C embeddings are combined to create a single input to the model which then predicts a score for the entire A-B-C link. This negates the need for an aggregator function as in the baselines and CD-1 approach. This has the additional benefit of making it trivially easy to calculate a score regardless of the length of the path between A and C, by simply combining any additional node embeddings as the initial 3 and passing it to the neural network.

Open discovery neural models and approaches. The approaches used for open discovery are presented and explained here.

OD-1: The same model and a similar approach to CD-1 is used here: the neural model is used to provide a score for each A-B or B-C link in the path from A to each possible C. The scores are then used in aggregator functions. The difference in open discovery is that here the

scores are then also used in the accumulator functions to rank different paths which lead to the same C.

OD-2: A Convolutional Neural Network (CNN) model is used to implement an approach to open discovery which removes the need for aggregator and accumulator functions. As in CD-2, all the node embeddings of a path are combined into a single vector, however as this is open discovery, there will be many paths that lead to the same C. To obtain a score which uses information from all these paths, the combined vectors are stacked to create a window which we pass into a CNN which outputs a score indicative of the strength of the A-C links. This is analogous to passing an image to a CNN, but here the 'image' is produced by stacking vector representations of ABC links. The convolutional filter always slides down the stack of links, never across so that it always covers the entire link. The ABC links to be stacked are combined using the same 5 link combination methods as mentioned above.

The reader will perhaps note that the CNN expects a fixed size input and the amount of paths leading to a C will inevitably vary from case to case, creating varying input sizes which could exceed a fixed window size. To deal with this, we combine multiple windows into a single window using elementwise summation. As the total number of links will not always be a multiple of the window size, zero padding is used to fill any remaining gaps. For example: if a particular case has 175 paths and an input size of 50 is used, we will be able to sum 3 windows of 50 and as there will be only 25 paths in the final input 25 more paths will be zero padded to the input to make it 50.

In this model, the input layer leads to a batchnormed convolutional layer with ReLU activation units, then a max pooling layer then a fully connected layer before the final layer with Softplus activation. Unlike the other models which are trained as classifiers, this model uses a pointwise approach, employing Mean Squared Error (MSE) loss, to learning the ranking function by using the Jaccard Index score of the AC link as the multi-level ratings (see [54] for a more detailed description of this). The model is depicted in Fig 1.

Datasets

The graphs we use were created from the following datasets. The graph details can be found in Table 1.

PubTator.   Biomedical entities recognised by PubTator mentioned in the titles and abstracts of PubMed publications from 1873 to 2017 were used to create this dataset. A link exists between two biomedical entities if they co-occur in a single sentence. The annotations were downloaded on June 20th, 2017. Instances of 'hallmarks of cancer' which identify fundamental cancer processes, identified in text using [44] were also added to the graph as entities.

Biological general repository for interaction datasets (BioGRID).   This is an open database created from manually curating experimentally-validated genetic and protein interactions that are reported in peer-reviewed publications [55]. The latest release [56] includes over 1 million Genetic and Protein interactions across all major organism species and humans. Links in this graph represent biomedical interactions from published, experimentally-validated genetic and protein interactions. We use version 3.4.167 of this dataset.

Experiments

As all approaches create ranked lists, the possibility of tied ranks exists. We use the median of the tied range to determine the rank of a gold item with ties, for example a gold ranked 10th with 10 ties is ranked the median of 10-20 range: 15th.
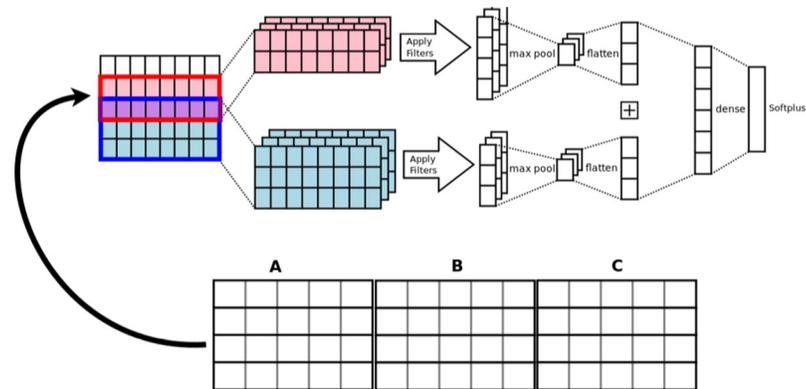
Fig 1. The open discovery 2 model.

Details of neural approaches

Unlike the baseline models, the neural approaches need negative examples for training. We create these by selecting either A-B or B-C links which did not form for a given A-C or A-C connections which do not exist for models which operate on the entire link path (i.e. CD-2 and OD-2).

All models are trained with batch size 100, training set size 200,000 for 150 epochs with the Adam optimiser [57], but the model is evaluated on the case after every 5 epochs and the best performance reported. For the BioGRID experiments, because evaluation is a lot more time-consuming, the models are evaluated every 25 epochs on the development set and the best performing model on MRR is evaluated on the held out test at the end. The CNN uses a learning rate of 10^{-5} while the MLPs use 10^{-4}. For CD-1, CD-2 and OD-1, there is a single hidden layer with 100 units. For OD-2, the input height is 50 and the width is the size of the combined vector dimensions. The convolution window height is 7 and the convolutional output size is 128.

Case discoveries

We use the data from [19] directly, so that our results will be directly comparable. The graphs are cut off at the relevant years before the publication date of the discovery.

Closed discovery on cancer discoveries. For CD-1, the model is given the A-B and B-C links and the scores it produces are used in the aggregator functions to rank the Bs. For CD-2 the model is fed all the A-B-C links for the given A and C in each triplet and the score it produces is used to rank the Bs.

Open discovery for cancer and Swanson discoveries. For OD-1, the model is given the A-B and B-C links and the scores it produces for each link were used in the aggregator functions to produce a score for each path. The different paths which lead to the same C are used in the accumulator functions to produce a score which is used to rank the Cs. For OD-2, the

Table 1. Graph details (undirected link count).

Dataset	Node Count	Link Count	Link Type
BioGRID	68,734	1,209,578	Published Interactions
PubTator	$\sim$194,691	$\sim$12,797,468	Literature Co-occurrences

model is given all the A-B-C links for the given A and C in each pair and the score it output is used to rank the Cs.

BioGRID

The graph is split at the year 2016. We randomly split the post-2016 links into development and test sections. The development set is used to determine which epoch has the best trained model for evaluation. Due to computational constraints, we have to restrict the amount of nodes we could evaluate on. We randomly select 1,000 entities from the test set to be A nodes and have the model score each node within two hops as the Cs. The scores are then used to rank the Cs. Like the Swanson cases, it is not possible to perform closed discovery on this dataset.

Results

The results of the neural approaches are the median ranks and mean ranks averaged over five runs. The standard deviations reported are of the mean ranks. The results of the baselines are simply the means of the method across all relevant cases (they were not run multiple times as the neural approaches were as they are not subject to per-run variances as the neural network methods are) and the standard deviations are over those ranks.

The best score for a metric is in **bold** and the best for an approach is <u>underlined</u>. We sought to determine what methods gave the lowest mean ranks and lowest variance, measured by standard deviation. 'Metric' refers to mean and median ranks. 'Approach' refers to the three approaches: Baselines (Jaccard, t-test etc.), neural discovery approach 1 (CD-1, OD-1) and neural discovery approach 2 (CD-2, OD-2). Thus each 'metric' column should have a bolded term and each approach category (delineated by horizontal lines in the tables) should have an underlined term.

To increase clarity in the tables, we selected only the best results for each approach to show here. Full experimental results can be found in Section 4 of the Supplementary Document (S1 File). Where applicable, the accumulator and aggregator functions (explained in the "Baselines" section) are listed in the "Details" column as 'Acc' and 'Agg' respectively.

Closed discovery on cancer discovery cases

The results for closed discovery performed on the five Cancer Discovery cases used to evaluate LION are in Table 2.

Open discovery on cancer discovery and Swanson cases

Open discovery on cancer discovery cases. The results for open discovery performed on the five Cancer Discovery cases used to evaluate LION are in Table 3.

Table 2. Closed discovery: Mean and median ranks on the cancer discovery cases.

Approach	Mean Rank	Std. Dev.	Median	Details
Jaccard	<u>214.8</u>	256.9	81.0	Agg: min
t -test	262.0	341.8	<u>56.0</u>	Agg: min
CD-1-A	<u>86.3</u>	52.0	93.8	Agg: min
CD-1-C	94.5	80.0	**<u>36.4</u>**	Agg: min
CD-2-C	**<u>48.7</u>**	19.5	<u>42.0</u>	-

Table 3. Open discovery: Mean and median ranks on the cancer discovery cases.

Approach	Mean Rank	Std. Dev.	Median	Details
NPMI	60.2	54.4	36.0	Acc: sum, Agg: max
Count	367.4	553.3	15.0	Acc: sum, Agg: min
OD-1-C	93.4	145.8	31.4	Acc: sum, Agg: min
OD-1-A	218.3	368.7	26.8	Acc: sum, Agg: min
OD-2-H	31.1	11.9	12.2	-

Table 4. Open discovery: Mean and median ranks on the Swanson cases.

Approach	Mean Rank	Std. Dev.	Median	Details
Doc-Count	2,199.8	4,216.7	31.0	Acc: max, Agg: avg
t-**test**	3,956.4	7,899.3	5.0	Acc: max, Agg: avg
OD-1-H	3,558.3	7,930.7	19.2	Acc: sum, Agg: min
OD-1-C	3,721.4	8,306.7	4.0	Acc: sum, Agg: min
OD-2-H	1,013.4	167.9	17.6	-

Open discovery on Swanson cases. The results for open discovery performed on the five Swanson cases used to evaluate LION are in Table 4.

Open discovery on cancer discovery and Swanson cases. The results for open discovery performed across the five Cancer Discoveries and five Swanson cases combined are in Table 5.

Open discovery on BioGRID published interactions

Results for open discovery performed on the BioGRID dataset. Performance across the 4 metrics explained in the "Metrics" Section are in Table 6.

Discussion

Closed discovery on cancer discovery cases

The results of this experiment can be found in Table 2. The neural approaches performed much better than the existing methods in these experiments. Rows 3 and 4 show that the performance doubled, by halving the mean ranks, simply by replacing the baseline scoring metrics with a small neural classifier to provide the scores instead. It almost doubled again by replacing the aggregation of individual path scores with combining the vectors of the nodes involved in the path (row 5). Performance on the median also increased though not as drastically.

Of note here is that the neural approach which dispelled with the aggregator functions, instead opting to combine the inputs and obtaining a score for the entire path, was the best performer on mean ranks and the second best performer on median (row 5). This indicates

Table 5. Open discovery: Mean and median ranks on all open discovery Cases.

Approach	Mean Rank	Std. Dev.	Median	Details
Jaccard	1,634.4	4,733.9	21.0	Acc: sum, Agg: min
Count	1,925.8	5,171.3	11.5	Acc: sum, Agg: min
OD-1-C	1,907.4	5,859.4	18.2	Acc: sum, Agg: min
OD-2-H	522.2	89.9	14.9	-

Table 6. Open discovery on time-sliced BioGRID.

Approach	MR	MRR	R- Prec.	MAP	Details
Jaccard	1,197.3	<u>2.19</u>	<u>2.47</u>	<u>2.86</u>	Acc: sum, Agg: min
LLR	<u>1,132.9</u>	1.34	1.38	1.9	Acc: sum, Agg: max
OD-1-H	<u>1,907.5</u>	0.92	0.96	1.25	Acc: sum, Agg: max
OD-1-C	1,913.4	<u>0.94</u>	<u>1.01</u>	1.23	Acc: sum, Agg: max
OD-1-W2	1,908.3	0.92	0.98	<u>1.26</u>	Acc: sum, Agg: max
OD-2-C	**1,113.1**	**3.42**	**4.73**	**5.46**	-

that the information which the aggregator functions seek to provide to an approach is better provided by combining the vector representations of the nodes in the path.

Open discovery on cancer discovery and Swanson cases

Open discovery on cancer discovery cases. The results of this experiment can be found in Table 3. Despite the strong improvements seen in closed discovery by simply replacing the baseline scoring metrics with a neural classifier, that was not the case here for either mean or median rank (rows 3-4). However, the more complex CNN approach was able to produce results which approximately doubled performance on mean rank from a strong baseline (row 5). It also performed the best on median rank.

Analogous to the closed discovery experiments, the approach which dispelled with aggregators and accumulators outperformed on mean ranks (row 5). Additionally, it was the best median performer here, further validating it.

Open discovery on Swanson cases. The results of this experiment can be found in Table 4. A similar trend to the cancer cases was shown here: simply replacing the baseline scoring metrics with a neural classifier decreased performance on mean rank, although one such approach did perform the best on median rank (rows 3-4). The strong performance of the CNN continued as it again doubled performance on mean rank although it was only the third best on median rank (row 5). The trend of the approach which dispelled with aggregators and accumulators outperforming on mean ranks also continued.

Open discovery on both cancer discovery and Swanson cases. The results of this experiment can be found in Table 5. Given the results of the subset experiments, it is not surprising that the CNN was the best performer across all open discovery cases (row 4). Its performance on mean rank was approximately three times better than that of the best baseline and it was the second best on median, although the simple count baseline approach was the best.

General open discovery. In addition to its strong performance across the cases (Tables 3, 4 and 5), the OD-2-H approach is also the most stable as it showed the lowest variation in performance over multiple runs of the best performing methods as measured by the standard deviation shown in those tables.

A point in favour of the neural approaches over the baselines is their apparent consistency in performance over the subsets of the cancer and Swanson cases. The baseline methods which performed the best, shown in Tables 3, 4 and 5, varied while the best neural approaches recurred, demonstrating their invariability to the vagaries of the case subsets.

General case discoveries

Whether to use mean or median as average for these experiments is a valid question. [19] reported median and we do the same to allow for comparison, but also report the mean

because we believe that it is better suited to this situation. The median is robust to outliers and can give a more accurate picture of a system's performance when an outlier can radically affect the mean, as is the case with the Swanson cases used. However, the aim of this research is to find an approach which will aid researchers on totally novel data, so the worst-case performance of the system (even if it is rare) is of importance and the aim should be to use methods which give the best results across all cases. Thus, evaluating accurately should involve looking at performance in all available cases. Median ignores not only outliers, but effectively all performances beyond the median (approximately half the use cases). The argument can thus be made that the median does not give an accurate reflection of an approach's performance.

Taking mean as a preferable metric to median, the case of the neural methods is strengthened as they were the best performers across all the case experiments. Additionally, there was low variance among the best neural approaches. It was also pleasing to find that approaches which dispelled with the cumbersome aggregator and accumulator functions were the best. This indicates that when given the full path information, the neural models are able to discern how best to use it to improve performance.

It is also noteworthy that although methods which concatenated the node representations performed well, there were other approaches whose performance were comparable or better than it across these experiments. This is significant because unlike the concatenate combination method, which increases the input size linearly with the path length, the other node combination methods keep a fixed input size which makes them indifferent to the amount of hops between A and C. This feature makes them amenable to approaches to LBD beyond the simple two-hop ABC paradigm to the n-hop $AB_1 B_2 ... B_n C$ paradigm which it is generally agreed must be overcome for LBD to reach its true potential.

Time-sliced BioGRID

The reasons for undertaking these experiments were explained in the "Time Slicing" Section and the reasons for the multi-faceted evaluation in the "Metrics" Section. We will make use of and expand on these here.

The data used in this experiment represent experimentally-validated, human-curated interactions which were published in peer-reviewed publications. Thus the knowledge proposed by using it is of high quality. Additionally, the evaluation is time-sliced which is reflective of how knowledge discovery progresses in the real world and involves far more evaluation instances than a handful of cases, notwithstanding the very high quality of the cases.

LBD across a large amount of possible positives is a ranking problem because its proposals are usually costly to investigate. Thus priority should be given to approaches which can rank correct new associations at the very top of the list even if they rank more of them lower; the classic precision-recall trade-off. Performance too far down the list can effectively be ignored: when experimentally validating new knowledge proposals, whether it is ranked 200th or 900th is likely of little concern to a user; it is too far down the list.

Metrics like MAP, MRR and R-precision place value on higher ranked true positives but they do not do so equally. MAP and MRR are concerned with the entire list but MRR punishes lower-ranked correct items more when the retrieval space is large as it tends to be in LBD, especially open discovery. R-precision literally discards most of the returned results and reports results only on the best. Thus performance on metrics like R-precision and MRR give a better idea of the practical worth of an LBD system, especially on open discovery.

The results of this experiment can be found in Table 6. The OD-2-C method we introduce here performs approximately 1.5-1.9 times as good the baseline approaches on these metrics, in addition to strong performance on MAP and mean rank (row 6). It is a variant of the OD-

2-H method which showed vastly better performance on the cases experiments. The results here thus validates the OD-2 (CNN) approach to open discovery we presented in the Section "Open Discovery neural models and approaches".

The role of the node embeddings in the superior performance of the neural network methods may be difficult to isolate but we can surmise how they can contribute. The node embeddings utilise both first order and higher order proximities which incorporate information from a node's wider neighbourhood than the baseline scoring methods would. This additional information can aid in ranking a node and lead to improved performance.

While there is still lots of room for improvement, these results are dependable and demonstrate the potential for using neural networks to perform even traditional open and closed discovery within the ABC paradigm.

Conclusion

LBD aims to discover new knowledge automatically from large collections of literature. Scientific literature is growing exponentially, making it difficult for researchers to stay current in their discipline. LBD can solve this problem by facilitating hypothesis testing and generation to give tangible support to scientific research.

We proposed four neural network-based approaches to open and closed LBD. We compared our methods with those used by a state-of-the-art LBD system to replicate recently published findings in cancer biology and also applied them to a time-sliced dataset of human-curated, peer-reviewed biological interactions. In both cases, our methods showed a notable and significant improvement over the existing methods on metrics adapted to the situation.

Although there is scope for much improvement, these results strongly demonstrate the potential of using neural networks to perform open and closed LBD well within the ABC paradigm and in some cases using only sentence-level co-occurrence relationships. Combined with previous work on the viability of using neural link prediction for LBD, they indicate that neural networks can significantly improve performance on the increasingly important task of LBD. Immediate future work includes using the pairwise approach to learning the ranking function for the CNN approach, using more advanced graph embedding techniques to better capture the information present in graphs and applying attention to the neural approaches to determine which paths are contributing the most to its performance.

Supporting information

S1 File. Supplementary document. Contains additional results and formal definition of evaluation metrics which were left out of the paper in pursuit of conciseness.
(PDF)

Acknowledgments

Author Contributions

Conceptualization: Gamal Crichton, Simon Baker, Yufan Guo.

Data curation: Simon Baker.

References

1. Swanson DR. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. Perspectives in Biology and Medicine. 1986; 30(1):7–18. https://doi.org/10.1353/pbm.1986.0087 PMID: 3797213

2. DiGiacomo RA, Kremer JM, Shah DM. Fish-oil dietary supplementation in patients with Raynaud's phenomenon: a double-blind, controlled, prospective study. The American Journal of Medicine. 1989; 86 (2):158–164. https://doi.org/10.1016/0002-9343(89)90261-1 PMID: 2536517

3. Swanson DR. Migraine and Magnesium: eleven neglected connections. Perspectives in Biology and Medicine. 1988; 31(4):526–557. https://doi.org/10.1353/pbm.1988.0009 PMID: 3075738

4. Swanson DR. Somatomedin C and Arginine: implicit connections between mutually isolated literatures. Perspectives in Biology and Medicine. 1990; 33(2):157–186. https://doi.org/10.1353/pbm.1990.0031 PMID: 2406696

5. Smalheiser NR, Swanson DR. Linking estrogen to Alzheimer's disease an informatics approach. Neurology. 1996; 47(3):809–810. https://doi.org/10.1212/wnl.47.3.809 PMID: 8797484

6. Hristovski D, Rindflesch T, Peterlin B. Using literature-based discovery to identify novel therapeutic approaches. Cardiovascular & Hematological Agents in Medicinal Chemistry (Formerly Current Medicinal Chemistry-Cardiovascular & Hematological Agents). 2013; 11(1):14–24.

7. McDonald D, McNicoll I, Weir G, Reimer T, Redfearn J, Jacobs N, et al. The value and benefits of text mining. JISC Digital Infrastructure. 2012;.

8. Hunter L, Cohen KB. Biomedical Language Processing: What's Beyond PubMed? Molecular Cell. 2006; 21(5):589—594. https://doi.org/10.1016/j.molcel.2006.02.012 PMID: 16507357

9. Swanson DR. Medical literature as a potential source of new knowledge. Bulletin of the Medical Library Association. 1990; 78(1):29. PMID: 2403828

10. Swanson DR. Undiscovered public knowledge. The Library Quarterly. 1986; 56(2):103–118. https://doi.org/10.1086/601720

11. Kostoff RN, Briggs MB, Lyons TJ. Literature-related discovery (LRD): Potential treatments for multiple sclerosis. Technological Forecasting and Social Change. 2008; 75(2):239–255. https://doi.org/10.1016/j.techfore.2007.11.007

12. Kostoff RN, Briggs MB. Literature-Related Discovery (LRD): potential treatments for Parkinson's disease. Technological Forecasting and Social Change. 2008; 75(2):226–238. https://doi.org/10.1016/j.techfore.2007.11.006

13. Hristovski D, Kastrin A, Peterlin B, Rindflesch TC. Combining semantic relations and DNA microarray data for novel hypotheses generation. In: Linking literature, information, and knowledge for biology. Springer; 2010. p. 53–61.

14. Zhang R, Cairelli MJ, Fiszman M, Kilicoglu H, Rindflesch TC, Pakhomov SV, et al. Exploiting Literature-derived knowledge and semantics to identify potential prostate cancer drugs. Cancer Informatics. 2014; 13:CIN–S13889.

15. Banerjee R, Choi Y, Piyush G, Naik A, Ramakrishnan I. Automated suggestion of tests for identifying likelihood of adverse drug events. In: 2014 IEEE International Conference on Healthcare Informatics (ICHI). IEEE; 2014. p. 170–175.

16. Shang N, Xu H, Rindflesch TC, Cohen T. Identifying plausible adverse drug reactions using knowledge extracted from the literature. Journal of Biomedical Informatics. 2014; 52:293–310. https://doi.org/10.1016/j.jbi.2014.07.011 PMID: 25046831

17. Ahlers CB, Hristovski D, Kilicoglu H, Rindflesch TC. Using the literature-based discovery paradigm to investigate drug mechanisms. In: AMIA Annual Symposium Proceedings. vol. 2007. American Medical Informatics Association; 2007. p. 6.

18. Hristovski D, Kastrin A, Rindflesch TC. Semantics-based cross-domain collaboration recommendation in the life sciences: Preliminary results. In: Advances in Social Networks Analysis and Mining (ASO-NAM), 2015 IEEE/ACM International Conference on. IEEE; 2015. p. 805–806.

19. Pyysalo S, Baker S, Ali I, Haselwimmer S, Shah T, Young A, et al. LION LBD: a literature-based discovery system for cancer biology. Bioinformatics. 2018; p. bty845.

20. Wei CH, Kao HY, Lu Z. PubTator: a Web-based text mining tool for assisting Biocuration. Nucleic Acids Research. 2013; 41. https://doi.org/10.1093/nar/gkt441

21. Baker S, Ali I, Silins I, Pyysalo S, Guo Y, Högberg J, et al. Cancer Hallmarks Analytics Tool (CHAT): a text mining approach to organize and evaluate scientific literature on cancer. Bioinformatics. 2017; 33 (24):3973–3981. https://doi.org/10.1093/bioinformatics/btx454 PMID: 29036271

22. Sang S, Yang Z, Liu X, Wang L, Zhang Y, Lin H, et al. A Knowledge Graph based Bidirectional Recurrent Neural Network Method for Literature-based Discovery. In: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2018. p. 751–752.

23. Swanson D. Literature-based discovery? The very idea. In: Literature-based discovery. Springer; 2008. p. 3–11.

24. Henry S, McInnes BT. Literature based discovery: models, methods, and trends. Journal of Biomedical Informatics. 2017; 74:20–32. https://doi.org/10.1016/j.jbi.2017.08.011 PMID: 28838802

25. Kostoff RN, Solka JL, Rushenberg RL, Wyatt JA. Literature-related discovery (LRD): water purification. Technological Forecasting and Social Change. 2008; 75(2):256–275. https://doi.org/10.1016/j.techfore.2007.11.007

26. Gordon MD, Awad NF. The tip of the iceberg: the quest for innovation at the base of the pyramid. In: Literature-based Discovery. Springer; 2008. p. 23–37.

27. Smalheiser NR, Swanson DR. Using ARROWSMITH: a computer-assisted approach to formulating and assessing scientific hypotheses. Computer methods and programs in biomedicine. 1998; 57 (3):149–153. https://doi.org/10.1016/s0169-2607(98)00033-9 PMID: 9822851

28. Gordon MD, Lindsay RK. Toward discovery support systems: A replication, re-examination, and extension of Swanson's work on literature-based discovery of a connection between Raynaud's and fish oil. Journal of the American Society for Information Science. 1996; 47(2):116–128. https://doi.org/10.1002/(SICI)1097-4571(199602)47:2%3C116::AID-ASI3%3E3.0.CO;2-1

29. Cohen T, Schvaneveldt R, Widdows D. Reflective random indexing and indirect inference: A scalable method for discovery of implicit connections. Journal of Biomedical Informatics. 2010; 43(2):240–256. https://doi.org/10.1016/j.jbi.2009.09.003 PMID: 19761870

30. Sebastian Y, Siew EG, Orimaye SO. Emerging approaches in literature-based discovery: techniques and performance review. The Knowledge Engineering Review. 2017; 32. https://doi.org/10.1017/S0269888917000042

31. Eronen L, Hintsanen P, Toivonen H. Biomine: a network-structured resource of biological entities for link prediction. In: Bisociative Knowledge Discovery. Springer; 2012. p. 364–378.

32. Cameron D, Kavuluru R, Rindflesch TC, Sheth AP, Thirunarayan K, Bodenreider O. Context-driven automatic subgraph creation for literature-based discovery. Journal of Biomedical Informatics. 2015; 54:141–157. https://doi.org/10.1016/j.jbi.2015.01.014 PMID: 25661592

33. Swanson DR, Smalheiser NR. An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artificial Intelligence. 1997; 91(2):183–203. https://doi.org/10.1016/S0004-3702(97)00008-8

34. Weeber M, Klein H, Aronson AR, Mork JG, De Jong-van Den Berg L, Vos R. Text-based discovery in biomedicine: the architecture of the DAD-system. In: Proceedings of the AMIA Symposium. American Medical Informatics Association; 2000. p. 903.

35. Hristovski D, Džeroski S, Peterlin B, Rožić A. Supporting discovery in medicine by association rule mining of bibliographic databases. In: European Conference on Principles of Data Mining and Knowledge Discovery. Springer; 2000. p. 446–451.

36. Preiss J, Stevenson M, Gaizauskas R. Exploring relation types for literature-based discovery. Journal of the American Medical Informatics Association. 2015; 22:987–992. https://doi.org/10.1093/jamia/ocv002 PMID: 25971437

37. Yetisgen-Yildiz M, Pratt W. A new evaluation methodology for literature-based discovery systems. Journal of Biomedical Informatics. 2009; 42(4):633–643. https://doi.org/10.1016/j.jbi.2008.12.001 PMID: 19124086

38. Yetisgen-Yildiz M, Pratt W. Evaluation of literature-based discovery systems. In: Literature-based discovery. Springer; 2008. p. 101–113. https://doi.org/10.1007/978-3-540-68690-3_7

39. Stegmann J, Grohmann G. Hypothesis generation guided by co-word clustering. Scientometrics. 2003; 56(1):111–135. https://doi.org/10.1023/A:1021954808804

40. Wren JD, Bekeredjian R, Stewart JA, Shohet RV, Garner HR. Knowledge discovery by automated identification and ranking of implicit relationships. Bioinformatics. 2004; 20(3):389–398. https://doi.org/10.1093/bioinformatics/btg421 PMID: 14960466

41. Benton A, Mitchell M, Hovy D. Multitask learning for mental health conditions with limited social media data. In: EACL 2017-15th Conference of the European Chapter of the Association for Computational Linguistics. EACL; 2017.

42. Crichton G, Pyysalo S, Chiu B, Korhonen A. A neural network multi-task learning approach to biomedical named entity recognition. BMC Bioinformatics. 2017; 18(1):368. https://doi.org/10.1186/s12859-017-1776-8 PMID: 28810903

43. Habibi M, Weber L, Neves M, Wiegandt DL, Leser U. Deep learning with word embeddings improves biomedical named entity recognition. Bioinformatics. 2017; 33(14):i37–i48. https://doi.org/10.1093/bioinformatics/btx228 PMID: 28881963

44. Baker S, Korhonen A, Pyysalo S. Cancer Hallmark Text Classification Using Convolutional Neural Networks. BioTxtM 2016. 2016; p. 1.

45. Crichton G, Guo Y, Pyysalo S, Korhonen A. Neural networks for link prediction in realistic biomedical graphs: a multi-dimensional evaluation of graph embedding-based approaches. BMC Bioinformatics. 2018; 19(1):176. https://doi.org/10.1186/s12859-018-2163-9 PMID: 29783926

46. Zeng T, Ji S. Deep Convolutional Neural Networks for Multi-instance Multi-task Learning. In: Proceedings of ICDM 2015; 2015. p. 579–588. Available from: http://dx.doi.org/10.1109/ICDM.2015.92.

47. Zhang W, Li R, Zeng T, Sun Q, Kumar S, Ye J, et al. Deep model based transfer and multi-task learning for biological image analysis. In: Proceedings of ACM SIGKDD 2015. ACM; 2015. p. 1475–1484.

48. Chen J, Yang L, Zhang Y, Alber M, Chen DZ. Combining fully convolutional and recurrent neural networks for 3d biomedical image segmentation. In: Advances in neural information processing systems; 2016. p. 3036–3044.

49. Oktay O, Ferrante E, Kamnitsas K, Heinrich M, Bai W, Caballero J, et al. Anatomically constrained neural networks (ACNNs): application to cardiac image enhancement and segmentation. IEEE transactions on medical imaging. 2017; 37(2):384–395. https://doi.org/10.1109/TMI.2017.2743464 PMID: 28961105

50. Shahid N, Rappon T, Berta W. Applications of artificial neural networks in health care organizational decision-making: A scoping review. PloS one. 2019; 14(2). https://doi.org/10.1371/journal.pone.0212356

51. Tang J, Qu M, Wang M, Zhang M, Yan J, Mei Q. LINE: Large-scale Information Network Embedding. In: Proceedings of WWW 2015. ACM; 2015.

52. Goyal P, Ferrara E. Graph embedding techniques, applications, and performance: A survey. Knowledge-Based Systems. 2018; 151:78–94. https://doi.org/10.1016/j.knosys.2018.03.022

53. Nair V, Hinton GE. Rectified linear units improve restricted boltzmann machines. In: Proceedings of ICML-10; 2010. p. 807–814.

54. Chen W, Liu TY, Lan Y, Ma ZM, Li H. Ranking measures and loss functions in learning to rank. In: Advances in Neural Information Processing Systems; 2009. p. 315–323.

55. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. Nucleic Acids Research. 2006; 34(suppl 1):D535–D539. https://doi.org/10.1093/nar/gkj109 PMID: 16381927

56. Chatr-aryamontri A, Oughtred R, Boucher L, Rust J, Chang C, Kolas NK, et al. The BioGRID interaction database: 2017 update. Nucleic Acids Research. 2017; 45(D1):D369–D379. https://doi.org/10.1093/nar/gkw1102 PMID: 27980099

57. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. In: Proceedings of 3rd International Conference on Learning Representations, ICLR, 2015; 2015.