Journal of Biomedical Informatics 90 (2019) 103094

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Rapamycin − mTOR + BRAF = ? Using relational similarity to find therapeutically relevant drug-gene relationships in unstructured text

Safa Fathiamini[a,*], Amber M. Johnson[b], Jia Zeng[b], Vijaykumar Holla[b], Nora S. Sanchez[b], Funda Meric-Bernstam[b,c,d], Elmer V. Bernstam[a,e,*], Trevor Cohen[f,*]

[a] School of Biomedical Informatics, The University of Texas Health Science Center at Houston, TX, United States
[b] Sheikh Khalifa Al Nahyan Ben Zayed Institute for Personalized Cancer Therapy, The University of Texas MD Anderson Cancer Center, Houston, TX, United States
[c] Department of Investigational Cancer Therapeutics, The University of Texas MD Anderson Cancer Center, Houston, TX, United States
[d] Department of Surgical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, United States
[e] Division of General Internal Medicine, Department of Internal Medicine, The University of Texas Health Science Center at Houston, TX, United States
[f] Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA, United States

ARTICLE INFO

Keywords:
Relational similarity
Attributional similarity
Neural word embeddings
Random Indexing
Precision oncology
Pharmacogenomics

1. Introduction

On account of the rapid growth in the biomedical literature, there is a pressing need for the development of new informatics technologies to identify clinically relevant assertions in this literature. This is particularly important in domains such as precision oncology, where new knowledge with immediate implications for patient care is emerging rapidly [1]. One approach to this problem involves applying Natural Language Processing (NLP) to find explicit assertions in the literature, such as “gene X is inhibited by drug Y”. As NLP is not perfect, it has been argued that methods that infer relationships between biomedical concepts from their distributional statistics across large text corpora present a robust and desirable alternative [2] because they draw conclusions from multiple observations of co-occurrence, rather than from an individual assertion.

Specifically, methods of distributional semantics derive similar representations for terms that occur in similar contexts in the literature [3]. Thus, two drugs with similar contextually-derived representations may be useful in the context of the same molecular aberration. Most research on distributional similarity has focused on the notion of *attributional similarity*, which estimates the similarity between entities (such as two drugs) based on the contexts in which they occur across a large corpus. In contrast, *relational similarity* between pairs of entities (such as two drug-gene pairs) is estimated from the contexts in which these entity pairs occur together. Relational similarity may help identify interesting relationships between biomedical concepts, but the relative utility of relational and attributional similarity has yet to be evaluated.

In this paper, we compare the performance of multiple relational and attributional similarity methods on the task of identifying drugs that may be therapeutically useful in the context of particular molecular aberrations. We compare automatically identified drugs to a gold standard created by a team of human experts. We hypothesize that relational similarity will be more effective than attributional similarity when applied to this task.

2. Background

Genomic profiling and drugs that target specific molecular aberrations found in patients’ tumors (i.e., targeted therapies) are increasingly available, enabling the practice of precision oncology. However, the knowledge required to practice precision oncology is evolving rapidly. Today’s “variant of unknown significance” may be tomorrow’s

[*] Corresponding authors at: School of Biomedical Informatics, The University of Texas Health Science Center at Houston, 7000 Fannin St., Suite 600, Houston, TX 77030, United States (S. Fathiamini, E.V. Bernstam). Department of Biomedical Informatics and Medical Education, University of Washington, 850 Republican Street, Seattle, WA 98109, United States (T. Cohen).

E-mail addresses: Safa.Fathiamini@uth.tmc.edu (S. Fathiamini), Elmer.V.Bernstam@uth.tmc.edu (E.V. Bernstam), cohenta@uw.edu (T. Cohen).

https://doi.org/10.1016/j.jbi.2019.103094
Received 4 September 2018; Received in revised form 30 November 2018; Accepted 27 December 2018
Available online 04 January 2019

preferred target. Further, many targeted therapies are investigational and are currently available primarily via clinical trials. Thus, there is an urgent need to develop informatics technologies to help curate clinically pertinent information in a timely fashion.

To address this problem, we previously introduced the Automated Identification of the Molecular Effects of Drugs (AIMED) system [4] that uses the knowledge-driven SemRep biomedical NLP system to extract clinically relevant pharmacogenomic relationships from the biomedical literature. While AIMED performed well with established drugs, performance with investigational agents was better overall when considering co-occurrence counts without the use of NLP (other than for concept extraction and normalization). However, while recall improved, precision decreased since extracted relationships were no longer constrained. These results revealed an underexplored area between the linguistic rules and semantic constraints that systems such as SemRep impose to identify specific relationships on the one hand (thus achieving higher precision), and the unconstrained associations defined by co-occurrence (evident by higher recall) on the other.

While methods of distributional semantics are commonly used to identify general associations between terms or concepts [3], these models can also capture information concerning the nature of the relationships between terms, either incidentally [5], or by design [6]. This raises the question as to whether such information might be used to identify drug/gene relationships to support the practice of precision oncology. More broadly, what sort of similarity is best-suited to the practical task of identifying relationships of interest? In this paper, we compare attributional similarity between entities (e.g., similarity to an effective drug) to relational similarity between entity pairs (e.g., similarity between drug:gene pairs).

2.1. Attributional similarity

Attributional similarity – similarity between objects or their properties [7] is relatively well-studied. Distributional methods estimate a quantitative measure of semantic relatedness between terms from the contexts in which they occur in across a large corpus of text. Geometric approaches to this problem involve the derivation of reduced-dimensional (i.e., with dimensionality less than the number of unique contexts, or context terms in a corpus) vector representations of terms from the contexts in which they occur [8], such that terms that occur in similar contexts will have similar vector representations. The distance between the resulting vectors provides a meaningful estimate of semantic similarity and relatedness. Such approaches include (but are not limited to) Latent Semantic Analysis (LSA) [9], and the Hyperspace Analogue to Language (HAL) [10], which have been used to find similarity between terms, as well as larger units of text, that correspond well to human judgment across multiple tasks [11,12]. Another method, Random Indexing [13] is more computationally efficient. It generates a reduced dimensional space and produces similar results to LSA in evaluations of term-term similarity such as synonym tests, and correspondence with free association norms [14,15]. More recently, neural word embeddings [5,16] have become a popular method of generating such reduced-dimensional representations, with improvements over prior distributional methods in some evaluations [17] (although some of these improvements have been shown to be contingent upon optimal configuration of model hyper-parameters in subsequent experiments [18]).

Reduced-dimensional vector representations of words and concepts (word/concept embeddings) have been applied to numerous biomedical informatics problems, including many applications that precede the current wave of popularity of word embeddings in biomedical NLP (for a review of applications prior to 2010, see [3]). The recognition that such representations provide a useful basis for machine learning models (neural network models in particular) of natural language (notably, [19]) led to their adoption as methods of biomedical NLP also, with a number of evaluations and applications in recent years [20–25]. For example, Nikfarjam et al. [26] applied embedding techniques to social media postings to model word similarity, and using a machine learning based approach, extracted adverse drug reaction mentions. Jiang et al. used biomedical entities, word stems, and syntactic chunks to create a biomedical embeddings model, which outperformed similar general domain models in such biomedical text mining tasks as drug-drug interaction extraction, and biomedical named entity recognition [27]. In a recent experiment, semantic data mapping and embedding techniques were combined to build a language model to analyze free-text clinical notes for estimating the short-term survival in metastatic cancer patients [28]. In most cases, the embedding techniques have been used to improve sentence level extraction. In general, this work has leveraged the attributional similarity provided by neural word embeddings – as words with similar meaning will have similar word embeddings, machine learning models can generalize from training to test examples even if they express the same idea in different terms. In contrast, our focus in this research, however, is on evaluating the ways in which analogical reasoning (in the form of relational similarity) can be leveraged for the purpose of finding useful relationships, and how such methods compare to their attributional counterparts.

2.2. Relational similarity

Relational similarity involves similarity between two pairs of concepts – if A's relationship to B is similar to C's relationship to D, then A::B is relationally similar to C::D. Relational similarity seems to be a fundamental component of analogical reasoning [7,29,30]. In seminal work, Turney and Littman developed a Vector Space Model (VSM) for calculating relational similarity [31]. Sixty-four "joining words" (such as "of", "to", etc.) were used to create patterns of both "A *join* B" and "B *join* A". Then, the relationship between two words (A and B) was characterized by counting the number of times they appeared together in those patterns across the corpus, resulting in a pair-by-pattern matrix. The relational similarity between any two given pairs of words was then represented by the cosine similarity between their corresponding vectors [31]. This work was then extended to develop Latent Relational Analysis (LRA) [32], a technique for measuring relational similarity that extends the VSM in three ways: (1) patterns are extracted from the corpus dynamically, (2) a thesaurus is used to extend the search space by including words that are synonymous with terms in the query pair, and (3) Singular Value Decomposition (SVD) is used to reduce the dimensionality of the pair-by-pattern matrix [6]. As such, LRA may be inconvenient to implement, particularly when pairs of interest change frequently and the corpus is large, and may scale poorly to large sets of concept pairs on account of the need for SVD.

Recent work in the general domain has attempted to estimate relational similarity from term (rather than pair) vector representations directly, finding that word vectors derived from a relatively scalable neural network model can implicitly capture information of this sort. Specifically, Mikolov and his colleagues developed two neural network architectures, continuous bag of words (CBOW) (which learns to predict a word based on the words that surround it), and the continuous skip-gram model (which learns to predict context words based on an observed word). These "word embedding" architectures were used to train word representations from large corpora (billions of words). Surprisingly, the resulting word vectors were found to capture relationships between words, which could be recovered with simple geometric operations.

For example, using the resulting vector representations, $\overrightarrow{Paris} - \overrightarrow{France} + \overrightarrow{Germany} \cong \overrightarrow{Berlin}$ [5,16]. Training of these architectures occurs through a process of online learning [33], in which each training context – a "sliding window" of words surrounding each observed word – is considered independently, though global term frequency statistics inform subsampling strategies. This permits parallel implementation of the training process, enhancing scalability.

Alternatively, it is possible to capture such information without training a predictive model on an example-by-example basis. Pennington et al. introduced Global Vectors (GloVe), a model for unsupervised learning of word representations that utilizes global distributional statistics directly, while still capturing similar structural information to online neural-probabilistic methods. In some experiments, GloVe performed better than comparable neural network approaches in evaluations on pairwise analogies [34], however these advantages were not replicated in subsequent experiments in which hyperparameters and training corpora were consistent across models [18].

Some research on relational similarity exists in the biomedical domain. Predication-based Semantic Indexing (PSI) is a variant of Random Indexing that explicitly encodes relationships between concepts from a collection of semantic predications (such as those extracted by SemRep, for example *docetaxel* STIMULATES *akt*) into distributed vector representations of concepts [35–37]. Across several experiments [37–39], PSI inferred new relationships by using relational similarity [40]. Embedding of Semantic Predications (ESP) is a neural-probabilistic alternative to PSI that has shown advantages in predictive modeling experiments using estimates of relational similarity [41]. Both PSI and ESP use relations extracted by SemRep, and thus represent a different class of methods to those under consideration in the current work.

Percha and Altman developed a method that uses grammatical dependency paths in the sentences that contain a pair of concepts as contextual features [2]. An unsupervised clustering technique called Ensemble Biclustering for Classification (EBC) is then applied to the resulting pair-by-path matrix, such that drug-gene pairs are represented by their frequencies of co-clustering with every other pair across large numbers of stochastically-initialized clustering processes. As drug/gene pairs linked by similar dependency paths will cluster together, EBC leverages relational similarity drawn from distributional statistics. Because it does so in a largely unsupervised manner, EBC is not readily adaptable to the cue/response paradigm we will employ in the current evaluation, which is limited to methods that do not require parsing to reveal grammatical dependencies.

In this paper, we explore the application of relational and attributional similarity techniques in precision oncology, as an exemplar of a rapidly-changing biomedical domain, focusing specifically on drug-gene relationship extraction from Medline abstracts.

3. Materials and methods

3.1. Training corpus

We used Medline abstracts as the source of information for our models. Specifically, we used a locally-constructed (but publicly available) database, SemMedDB_UTH 2015 [4,42], which provides 144 M sentences from 23.4 M Medline abstracts dated up to Sep 2014, as well as a list of the Unified Medical Language System (UMLS) and EntrezGene concepts found in each sentence, their semantic types, and Concept Unique Identifiers (CUIs - SemRep relies upon MetaMap for concept extraction). We replaced the narrative descriptions of all concepts extracted by MetaMap from the abstracts with their CUIs, and removed stop words using the SMART stopword list [43]. For example, “Sialyl-Tn antigen expression was studied immunohistochemically in 211 primary advanced gastric carcinomas.” was transformed to “C0074480 C0185117 studied C1441616 211 C1335475”. We will refer to text so transformed as CUI-transplanted text. The result of this process was a set of 23,610,369 abstracts, with 4,288,491 unique terms, all saved in an Apache Lucene index [44] to facilitate search and retrieval. To extract explicit drug-gene pairs and their intervening terms, we further processed individual sentences from the CUI-transplanted abstracts, and whenever a drug co-occurred with a gene in a sentence, we extracted the words that lay between them. In this fashion, we identified 52,465,681 drug-gene pair co-occurrence events, and combining their intervening terms (including other CUIs and non-CUI terms) resulted in representations for 6,899,439 unique pairs, each with a “bag of words” (BOW) consisting of every term that occurred between their constituent CUIs in any sentence in the corpus.

3.1.1. Search space filters

Methods of distributional semantics produce continuous estimates of relatedness between entities, and as such, they are well suited toward rank-ordering potentially therapeutic agents. To construct a search space of potential therapies, we retained only concepts with UMLS semantic types *aapp (Amino Acid, Peptide, or Protein), antb (Antibiotic), clnd (Clinical Drug), horm (Hormone), imft (Immunologic Factor), nnon (Nucleic Acid, Nucleoside, or Nucleotide), opco (Organophosphorus Compound), orch (Organic Chemical), phsu (Pharmacologic Substance)* for drugs (*enzy (Enzyme), gngm (Gene or Genome),* and *aapp* were also retained to represent genes for the purpose of constructing cue vectors). These choices were informed by results produced by different configurations of AIMED [4]. Next, since the goal of the system was to find *clinically relevant* drugs, we used several filters, developed during the course of the AIMED project, to eliminate drugs that could not be given to patients in a clinical setting (i.e., drugs that were not either FDA approved, or available via a clinical trial). The *NCI drug filter* only includes drugs that are mentioned in the NCI terminology as a “Pharmacologic Substance”, the *CT filter* includes drugs mentioned in the clinicaltrials.gov database, and the *FDA filter* includes only FDA approved drugs. The retrieved entries had to exist in *either* the FDA *or* CT list, *and* the NCI filter to pass the drug filter. To ensure that the performance of pair and entity-based models was compared across the same search space, only drugs and genes that were represented in both the entity- *and* pair-based spaces were retained. To meet this last constraint, a drug would need to co-occur at least once with the gene in question. Fig. 1 shows a high-level data flow diagram providing an overview of the data sources and algorithms employed.

3.2. Evaluation reference set

As a reference set to test the system output and validate the results, we used the knowledge base provided and maintained by the Precision Oncology Decision Support (PODS) team at the MD Anderson Cancer Center, Sheikh Khalifa Bin Zayed Al Nahyan Institute for Personalized Cancer Therapy (IPCT), accessible with permission at http://personalizedcancertherapy.org. Each gene and its associated drugs (collectively known as a Gene Sheet – GS) included in this knowledgebase was deemed by the PODS team to have treatment implications for certain cancer types. This reference set was provided in 2014, and given the constantly evolving nature of the domain, it was important to use the corresponding version of SemMedDB_UTH (Sep 2014). With this version, the list included 17 genes (and some of their synonyms/CUI/Entrez_ID variations), and 430 drugs known to target them (and 1035 synonyms/CUI variations).

All the entries in this reference set were normalized to UMLS CUIs or EntrezGene IDs for genes, henceforth collectively referred to as CUIs in this manuscript for uniformity, using a slightly modified version of SemRep (SemRep_UTH [4]). Some of the drugs were excluded from evaluation, either because they were not identified as ‘drug’ by SemRep_UTH; or because they were not found in the drug filters (explained above). Also, following the practice explained in [24], if a drug had no representation in the search space, we disregarded it in the evaluation. This resulted in the GS for one gene (*KIT*) being removed from the reference set, as all its drugs were eliminated in the filtering process. Eventually, 16 genes and 163 drugs were included in the evaluation. Table 1 shows a list of the genes used for this purpose, and the number of therapeutically-relevant drugs associated with each gene. Thus, only drugs that met the co-occurrence constraint after filtering (bottom row of Table 1) were considered as positive examples. This co-occurrence constraint is a prerequisite to comparison between pair- and entity-

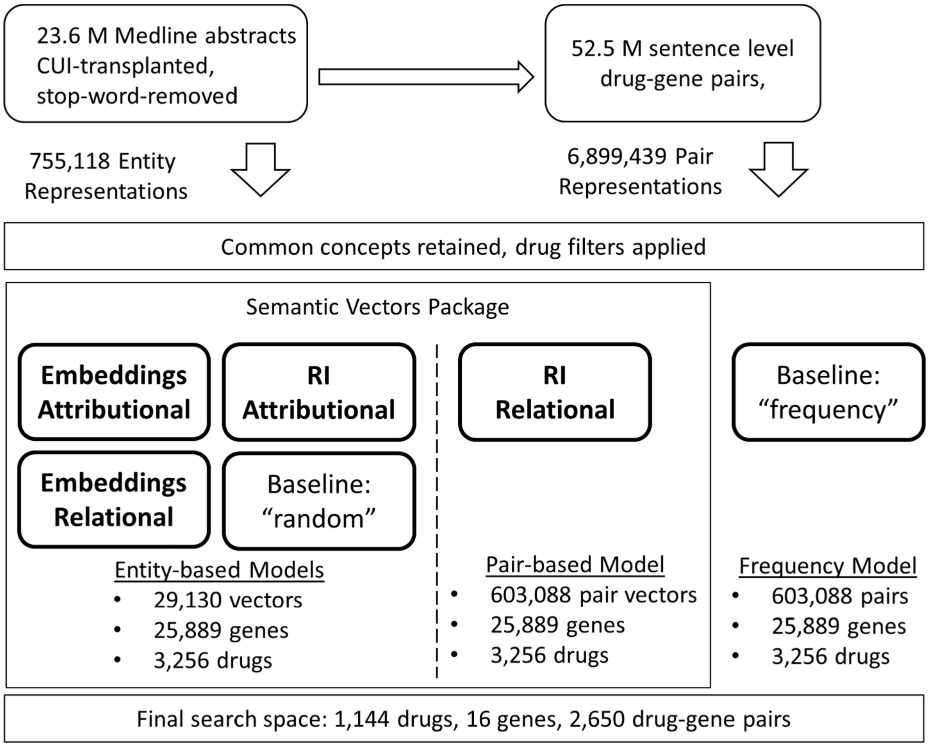


Fig. 1. High level data flow diagram from Medline abstracts to different models. RI = Random Indexing. CUI-transplanted Medline abstracts were used to create entity and pair representations. The drug filters were applied, and only concepts that had representatives in both spaces (common concepts) were retained. The open source Semantic Vectors package (Section 3.4 below) was used to create different vector models: RI Attributional (Sections 3.5.1, 3.5.2 below), RI Relational (Sections 3.5.3, 3.5.4 below), Embeddings Relational (Section 3.5.5 below), and Embeddings Attributional (Section 3.5.6 below). Two other models, "random", and "frequency" (Section 3.5.8 below) served as baselines for comparison.

based methods. However, it greatly constrains the number of drugs under consideration, a limitation we will subsequently discuss. This reduction in the number of therapeutically relevant drugs that could be considered for our experiments with the imposition of the co-occurrence constraint had a corresponding effect on the number of drugs remaining in the search space, reducing a total of 3,56 represented drugs (after filtering) to 1144. The proportion of drugs that were therapeutically relevant in at least one context was similar before (0.073) and after (0.087) this filtering. More details are presented in Table 3.

Many drugs that met the constraints for inclusion in the reference set were shared among two or more genes. That is to say, they were considered to be therapeutically active in the presence of an aberration to multiple genes. Out of the 16 genes in this set, five had all their drugs shared with other genes, and only one gene (SMO, targeted by only one drug) shared no drug with the others. Fig. 2 shows a summary of the drug overlap between any given gene and the rest of the genes. An important implication of this overlap is that sets of seed drugs (or seed drug-gene pairs) drawn from other gene sheets may, at times, include positive examples from the held-out gene sheet used at a particular point in the cross-validation procedure.

3.3. Search and evaluation process

We used known examples from the reference set as *seeds* and applied similarity measures to find *target* drugs in the search space, and the results were compared against the reference set. Based on the observation that in biomedicine there is often more than one correct answer to any given analogy question [45], and since distributional methods aim to prioritize results based on a continuous measure of similarity, we used standard ranked retrieval metrics to evaluate the results. The Average Precision (AP) is defined as:

$$AP = \frac{\sum_{k=1}^{n} P(k) \times IsRelevant(k)}{TR}$$

where

- n = number of results returned
- $IsRelevant$ = 1 for therapeutically-relevant drugs, otherwise 0
- TR = total number of relevant answers (whether they are returned or not)
- $P(k)$ = precision at the point at which the kth result was returned.

We also calculated Mean Average Precision (MAP) as the arithmetic mean of the AP values. The details and scope of the models involved in these evaluations are presented in the following sections.

3.4. Vector spaces and similarity models

We used variants of Random Indexing [13] and neural word embedding techniques [16,46] (as detailed in Section 3.5) to build our vector spaces. These operations were performed using the open source Semantic Vectors package[1] [47,48] which provides implementations of both of these approaches, eliminating the possibility of introducing bias on account of differences in pre-processing and tokenization of text.

3.4.1. Relational similarity

We used two approaches to model relational information. In the first, we *explicitly* identified drug-gene pairs, and created vector representations for them based on the terms that lie between them when they co-occur in our corpus. Relational similarity was estimated based on the cosine similarity between these *pair vectors.* A disadvantage of this approach is that all drug-gene pairs must be identified beforehand.

In contrast, in the second approach, we used the *implicit* relational information captured during the course of generating neural word embeddings, and performed geometric operations on the resulting *concept vectors* ($\overrightarrow{Drug}_{cue} - \overrightarrow{Gene}_{cue} + \overrightarrow{Gene}_{target} \cong ?$, as in the example $\overrightarrow{Paris} - \overrightarrow{France} + \overrightarrow{Italy} \cong \overrightarrow{Rome}$) [16]. Relational similarity was estimated as the cosine metric between the vector resulting from these arithmetic operations and the vector for each drug in the search space (as this will be high if $\overrightarrow{Drug}_{cue} - \overrightarrow{Gene}_{cue} \cong \overrightarrow{Drug}_{target} - \overrightarrow{Gene}_{target}$).

[1] https://github.com/semanticvectors/semanticvectors.

Table 1
List of genes and number of drugs used as the reference set for evaluation. Of note, there are fewer representations of drug-gene pairs than there are of therapeutically relevant drugs, as some therapeutically-relevant drugs did not co-occur with the gene in question, prohibiting the generation of a drug-gene pair representation. **Sum**: total number of therapeutically-relevant drug/gene pairs. **Total unique drugs**: total number of drugs that were considered therapeutically relevant in at least one context.

Gene	ABL1	AKT1	ALK	BRAF	CDK4	CDK6	EGFR	ERBB2	FGFR1	FGFR2	FLT3	KDR	PDGFRA	RET	ROS1	SMO	**Sum**	**Total unique drugs**
Number of therapeutically relevant drugs (TRD)	17	43	5	24	14	7	41	53	29	19	30	53	32	16	3	8	**394**	**237**
TRD found in entity-based spaces (ri_att, emb_att, emb_rel, rand-vec)[a]	11	23	3	10	4	3	22	26	15	10	19	33	24	11	2	1	**217**	**118**
TRD-gene pairs found in pair-based spaces (ri_rel, frequency)[a]	9	21	2	10	4	3	20	20	14	10	7	25	12	3	2	1	**163**	**99**

[a] A detailed description of the models is presented in the following sections.

3.4.2. Attributional similarity

To model attributional information, we used source abstracts as *documents* to build vector spaces, and measured the cosine similarity between *concepts*. Drugs known to be effective against particular genes were used as seeds to find other drugs by assessing their cosine similarity. In our first approach, we used Random Indexing to build the vector space, and in the second approach we used the same neural concept embeddings space from the relational similarity experiment, but instead of using relationships, individual drugs were used as seeds to find similar drugs.

3.5. Preliminary experiments and parameter selection

Preliminary experiments were performed to choose the optimal set of parameters for each model. All models used a minimum word frequency of 10. The vector dimensionality was 1000 for RI-based models (which tend to require relatively high dimensionality), and 500 for neural embedding models (which have been shown to perform well at relatively low dimensionalities).

3.5.1. Attributional similarity with Random Indexing (`ri_att-RI`)

In our first approach, we built a simple Random Indexing [13] space. A set of random vectors, one for each document in the corpus was generated by creating zero vectors of dimensionality 1000 and randomly assigning 10 of these values to either +1 or −1. The result is a set of document vectors with a high probability of being orthogonal, or close-to-orthogonal, on account of the statistical properties of high-dimensional space [13]. Term vectors were built by adding together the vectors for documents they occurred in.

3.5.2. Attributional similarity with Reflective Random Indexing (`ri_att-RRI`)

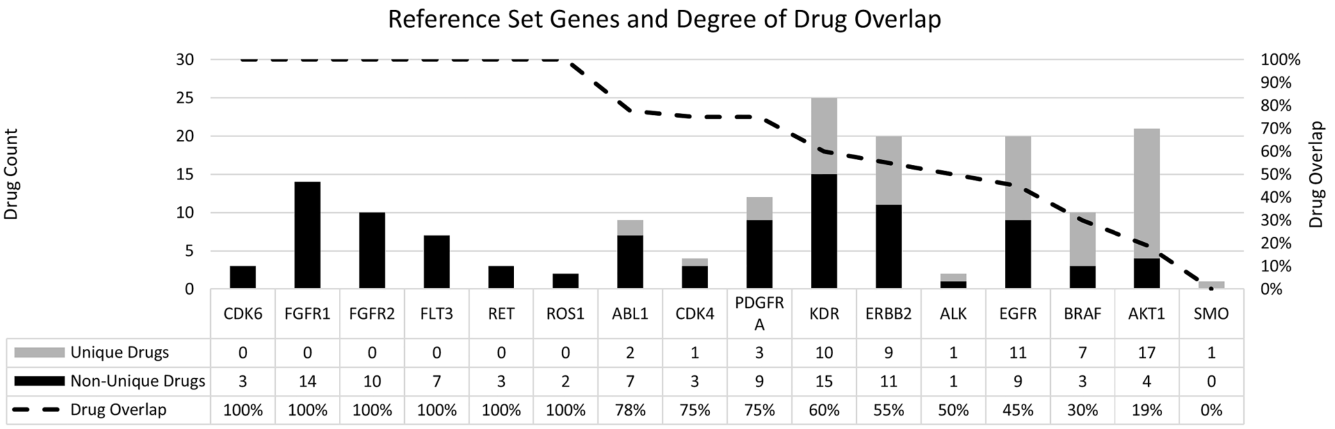
In this approach, a Term-based Reflective Random Indexing (TRRI) [49] space was built. In TRRI, random vectors are assigned to terms (a combination of terms and CUIs in our case), and added together to generate *document* vectors for documents containing those terms, which are subsequently normalized. Log entropy was used as the term-weighting scheme. This was the beginning of an iterative training procedure – new term vectors were generated by adding together the document vectors for documents in which they occurred, then the cycle was repeated. This provided a convenient way of estimating the relatedness between terms that do not co-occur in documents, as terms that co-occur with similar *other* terms will also have similar vectors. The `ri_att-RRI` was built with the same dimensionality and number of random values as the previously discussed `ri_att-RI` space, over a single iteration (random term vectors → document vectors → term vectors).

3.5.3. Relational similarity with pair vectors and Random Indexing (`ri_rel-RI`)

As a relational counterpart to `ri_att-RI` above, we created vector representations of drug/gene pairs in accordance with the RI paradigm [14]. We treated each distinct BOW (see above) as a *pseudo-document*, generating pair vectors by adding together random vectors for the terms in each BOW, and normalizing the result. No term weighting scheme was used.

3.5.4. Relational similarity with pair vectors and Reflective Random Indexing (`ri_rel-RRI`)

This model was similar to `ri_rel-RI` in that we treated each distinct BOW as a *pseudo-document*, and created pair vectors by adding together vectors for terms in each BOW and normalizing the result. The difference, however, was that instead of using random vectors for terms, we used the term vectors trained in the process of TRRI for `ri_att-RRI` model explained above. We hypothesized that doing so would provide the means to assess the similarity between pair-based



	CDK6	FGFR1	FGFR2	FLT3	RET	ROS1	ABL1	CDK4	PDGFR A	KDR	ERBB2	ALK	EGFR	BRAF	AKT1	SMO
▨ Unique Drugs	0	0	0	0	0	0	2	1	3	10	9	1	11	7	17	1
■ Non-Unique Drugs	3	14	10	7	3	2	7	3	9	15	11	1	9	3	4	0
– – Drug Overlap	100%	100%	100%	100%	100%	100%	78%	75%	75%	60%	55%	50%	45%	30%	19%	0%

Fig. 2. Reference set genes and the percentage of therapeutically active drugs that they share with other genes.

pseudo-documents containing semantically related but non-identical terms.

3.5.5. *Relational similarity with concept embeddings (`emb_rel`)*

A second class of relational models were built using the Semantic Vectors implementation of the Skipgram-with-Negative-Sampling (SGNS) algorithm, following the descriptions provided in [16,46] for word embeddings, with the source abstracts (rather than sentences) as *documents*. With SGNS, a neural network is trained to predict the terms surrounding an observed term, within a sliding window (the "context") that is moved through the text. Consequently, each term has two vector representations, which correspond to the term-specific input weights (semantic vectors) and output weights (context vectors) of the neural network. These vectors are initialized stochastically, but become meaningfully similar to one another during the course of training (in contrast in RI only the semantic vectors change during training). The probability of a surrounding term given an observed term is estimated as the sigmoid function of the scalar product between the input weights of the observed term, and the output weights of the surrounding term.

3.5.6. *Attributional similarity with concept embeddings (`emb_att`)*

In this experiment, we used the same word embeddings space as the previous model to find drugs similar to known drugs from the reference set.

3.5.7. *Parameter variations with embeddings models*

Prior work has evaluated the effect of neural word embedding hyper-parameters on task performance in the biomedical domain [18,24]. We assessed two of those parameters: subsampling (`ss`: the process of ignoring instances of frequently occurring terms with some probability) at thresholds of 0.001 and 0.00001, and window size (`ws`: the number of words considered before and after the target word, in the context of a sliding window) at levels 5 and 8. Furthermore, based on the findings by Levy et al. [18] who showed that adding context vectors to word vectors (`w+c`) with SGNS could help improve performance, we tested models with and without context vectors. Overall, six versions of the embeddings search space were built using different combinations of these parameters, as summarized in Table 2.

3.5.8. *Baseline models*

To establish a baseline and to assess the effect of co-occurrence alone without any similarity measure, the original drug-gene pairs that were identified in the course of building the `ri_rel` models were sorted based on their frequency of co-occurrence across the entire search space. In this model (henceforth: "`frequency`" model), the more a drug co-occurred with a gene, the higher it ranked. For each gene of interest, the resulting ranked list of drugs was compared with the reference set for evaluation.

A second baseline model was built using a set of random vectors for individual concepts (henceforth: "`rand-vec`" model). In a manner similar to the attributional methods described above, drug vectors were used to find similar drugs, and the results were compared with the reference set. The intuition here was: since the vectors used in this model were randomly chosen, they have a high probability of being orthogonal or close-to-orthogonal to each other. Consequently, any performance observed must be attributable to random overlap between vectors (as they are not perfectly orthogonal), or because drugs overlap across reference sets (as discussed in Section 3.2). Thus, inclusion of the `rand-vec` model permits us to estimate the extent to which observed performance exceeds that produced by incidental overlap. Table 2 summarizes different models, and their variants, used for search.

3.6. *Cross-validation*

Both relational and attributional models require seed examples. Retrieval and ranking of target entries is based on similarity to these seeds. For attributional models the seed and target were drugs, and for relational models they were drug-gene pairs. With the `frequency` model the "seed" was just the gene in question, and we ranked the drugs that co-occurred with it based on frequency. To evaluate the pair-based models, rankings of retrieved pairs containing the reference set drugs were considered. For the sake of uniformity, we will refer to pair-based seeds and targets, simply as "drugs". We conducted our evaluation both at a single GS level (*InGene* – all cues and targets directly concerned the gene of interest), and across all the GSs (*ExGene* – the gene of interest served as the target, where all the other genes were used as seeds). Our hypothesis was that the *InGene* configuration would elicit the best performance from attributional models (as retrieved drugs would be similar to drugs that are known to be effective), while the *ExGene* configuration would elicit best performance from relational models (as the nature of the relationship between therapeutically relevant drugs and the genes they target may be consistent across genes).

3.6.1. *InGene models*

In the *InGene* model the scope of the cross validation was limited to one single GS at a time (given some drugs known to affect *this* gene, can we find others?) We used two cross validation strategies. With the first strategy, known as One-As-Seed ("*oas*"), we took one "target" drug at a time from the reference set and used all the other drugs *individually* as seeds to find it and calculate AP. Of note, since there was only one target drug to find, the AP was equivalent to reciprocal rank in this case. MAP for each gene was calculated by averaging the set of AP results (or rather, reciprocal ranks) obtained in this process. The second strategy, known as All-But-One ("*abo*"), involved using all the drugs (with vectors combined) to find a single held out drug. In this model the cue was the normalized superposition of the vector representations of

Table 2
Similarity models used for search.

	Attributional	Relational
Random Indexing	`ri_att`: Abstracts as *documents*, cosine similarity measured between *term vectors* – `ri_att-RI`: term vectors sum of random document vectors (RI) – `ri_att-RRI`: term vectors sum of document vectors trained on random term vectors (TRRI)	`ri_rel`: Drug-gene pairs-based BOW as *document*, cosine similarity measured between *pair (document) vectors* – `ri_rel-RI`: document vectors sum of random term vectors (RI) – `ri_rel-RRI`: document vectors sum of term vectors from `ri_att-RRI`
Word Embeddings	`emb_att`: Abstracts as *documents*, cosine similarity measured between *term vectors* – `emb_att-001_ws5`: ss = 10^{-3}, ws = 5 – `emb_att-001_ws8`: ss = 10^{-3}, ws = 8 – `emb_att-00001_ws8`: ss = 10^{-5}, ws = 8 All three variations above with `w+c` – `emb_att-001_ws5_w+c` – `emb_att-001_ws8_w+c` – `emb_att-00001_ws8_w+c`	`emb_rel`: Abstracts as *documents*, cosine similarity measured after *geometric operations on term vectors*: $\overrightarrow{CueDrug} - \overrightarrow{CueGene} + \overrightarrow{TargetGene} = ?$ – `emb_rel-001_ws5`: ss = 10^{-3}, ws = 5 – `emb_rel-001_ws8`: ss = 10^{-3}, ws = 8 – `emb_rel-00001_ws8`: ss = 10^{-5}, ws = 8 All three variations above with `w+c` – `emb_rel-001_ws5_w+c` – `emb_rel-001_ws8_w+c` – `emb_rel-00001_ws8_w+c`
Baseline	● `frequency`: drug-gene pairs sorted by the number of occurrences in the abstracts, search by gene returned drugs ● `rand-vec`: Abstracts as *documents*, cosine similarity measured between *random term vectors*	

all the cues concerned. For each gene, MAP was then calculated across this set of AP results (or more accurately, reciprocal ranks) (one for each held-out drug). As such, the main difference between "*oas*" and "*abo*" was that in the former, seed drugs were used *individually* as cues with the results averaged later, whereas in the latter, a *cumulative* seed vector was used as a cue. The motivation for this design was that in emerging domains, any single positive result would be useful toward finding other results and building the basis for further discoveries – hence the *oas* model. On the other hand, since in such domains the amount of available information is typically limited, one would try to maximize the robustness of the query vector by including in it as many existing positive answers as possible – hence the *abo* model. It has been shown that combining multiple examples as cues lead to better performance on analogical reasoning experiments [40,50]. As such, our hypothesis was that in any given class of experiments, the *abo* models would perform better than *oas*.

3.6.2. ExGene models

In the case of the *ExGene* model (given drugs known to affect *other* genes, can we find those affecting this one?), the *oas* model was implemented by first adding (and normalizing) the vectors for individual drugs under each seed gene to form one *prototypical* drug vector for each GS (one *gene sheet* as seed), and then using that vector to find the drugs that target the gene. With *ExGene*, the *abo* model simply involved adding up the vectors for all the drugs under all the seed genes (and normalizing them afterwards) to use as the seed. Fig. 3 shows a diagram of the cross-validation configurations. We tested the models described in Table 2 with these cross-validation configurations, and report the median of MAP values for the genes in the reference set.

Also, as we explained previously, many genes in the reference set had drugs that were also mentioned in other Gene Sheets. We hypothesized that this drug overlap would affect the MAP results for ExGene models, since for those genes, seed and target sets have drugs in common. Positive correlation between model performance and the degree of drug overlap may explain the results. To this end, we ran a Spearman Rank Order test to evaluate the correlation between degree of drug overlap among genes in the reference set, and the MAP results for each gene/model combination.

3.6.3. Final filtering of result

In all of the evaluations explained above, a drug-gene co-occurrence filter was applied to each result set from entity-based models, before calculating the AP. For each such model, drugs that did not co-occur with the gene in question in at least one original source sentence were eliminated, so that entity and pair-based models could be compared against the same set of constraints. The final number of drugs, genes, and drug-gene pairs (where applicable) used post-filtering for all tests were 1144, 16, and 2650, respectively.

4. Results

The best performing models were `ri_att-RRI` for attributional RI, `emb_att-001_ws5_w+c` for attributional embeddings, `ri_rel-RI` for relational RI, and `emb_rel-00001_ws8_w+c` for relational embeddings. We henceforth report the results of these model configurations as the representatives for their respective categories (Table 3). A summary of the net effect of the hyper-parameters on model performance is presented in Table 4. The results of the correlation test that we performed to assess a potential link between some of the results, and the degree of drug overlap in the Gene Sheets are presented in Table 5.

As discussed previously, a substantial proportion of therapeutically relevant drugs were eliminated to facilitate comparison with pair-based models. To better estimate the practical utility of these approaches we tested the best performing models from the relational and attributional categories with the full set of available drugs in the reference set (394 therapeutic applications for drugs across 16 Gene Sheets, Table 1) with all the other constraints the same as the main experiment, and found the median MAP to drop by an average of 0.26 across those representative models (Table 6). In doing so, we are penalizing the models for not finding drugs that are not represented in the vector space. While these results do suggest considerable room for improvement it is notable that under these circumstances, the results are still consistent with our main hypothesis that relational models outperform the attributional models.

As distributional methods produce a continuous estimate of the association strength between pairs of concepts, metrics based on binary decisions (relevant vs. irrelevant) such as recall, precision and the F-measure cannot be applied without artificially imposing a threshold. To this end, we created a precision-recall curve for the best performing relational model, and its attributional counterpart (Fig. 4). To generate precision-recall curves, we first combined results from individual queries (16 result sets per model, one for each held-out gene), removed duplicate drug-gene pairs, sorted the results based on degree of similarity, and then calculated precision, recall, and F1 measure at the point at which each clinically relevant drug/gene pair was identified. Such aggregation of results across the genes was only possible with abo-ExGene configuration, as only one score per drug-gene pair was generated by this configuration.

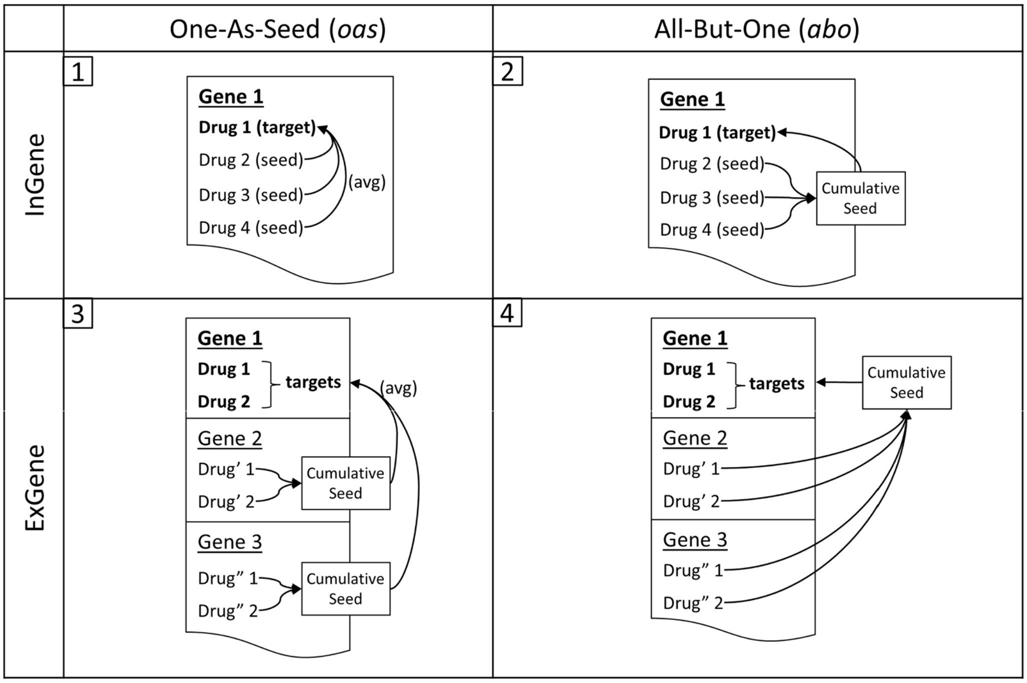


Fig. 3. Diagram of different cross validation models. (1) *oas-InGene*: Drugs in a Gene Sheet are used individually to find a target drug, (2) *abo-InGene*: Drugs in a Gene Sheet are combined (vectors superimposed, normalized), and used to find a target drug, (3) *oas-ExGene*: Gene Sheets are used individually (with drugs within each combined), to find drugs in a target Gene Sheet, (4) *abo-ExGene*: Gene Sheets are used in combination (all their drug vectors combined) to find drugs in a target Gene Sheet. In *oas* models, results from individual queries are averaged (shown as "(avg)" on the diagram) and reported as AP for the target drug(s).

5. Discussion

Our main hypothesis was that relational similarity would be more effective than attributional similarity in finding drugs that interact with the particular genes. To evaluate this hypothesis, for each category of relational similarity, we also developed an attributional counterpart. Our results indicate that models based on relational similarity generally outperform models based on attributional similarity on this task, providing strong support for the utility of analogical reasoning (exemplified by relational similarity) in the task of identifying clinically relevant relationships in natural language text.

A related hypothesis was that ExGene configurations would be advantageous for relational models, whereas attributional models may perform best with InGene. This hypothesis was supported in part by our results, as our Random Indexing based relational model exhibited its best performance in ExGene settings, leveraging relationships involving other genes (we did not compare relational embedding techniques for InGene configurations, as the `emb_rel` model is only defined for ExGene). However, we also anticipated that attributional models would perform worse in ExGene settings (where cue drugs interact with other genes than the target gene). This was exemplified by the `ri_att` model, with a performance drop from a MAP of 0.46 in abo-InGene to 0.14 in abo-ExGene. However, `emb_att` surprisingly displayed the opposite behavior, where its performance improved upon moving from InGene to ExGene (0.23–0.41). This paradoxical behavior may be due to the fact that in many cases the genes may be functionally related to one another, a hypothesis that is further supported by the drug overlap among Gene Sheets explained previously. Further investigation is needed to fully explain this phenomenon, as it is not clear why this would occur with one attributional model, but not the other.

A third hypothesis was that *abo* models would generally perform better than their *oas* counterparts. This hypothesis held true across the majority of the experiments (with one exception, `ri_att` oas-ExGene), suggesting that in emerging domains, where existing knowledge is limited, the best strategy for creating robust query vectors may be to use as many existing positive cues as possible. This finding is consistent with our previous work on analogical reasoning using distributed representations of semantic predications (concept-relation-concept triples) extracted from the biomedical literature using SemRep [40], as well as by subsequent work on analogical retrieval in the general domain [50]. As more positive examples are found, their addition to an existing query vector will progressively add to the robustness of the query.

Regarding the nature of the underlying representation, the `emb_rel` model consistently outperformed `ri_rel` both in *oas* and *abo* configurations. The `emb_att` model, however, is only marginally better than `ri_att` with oas-InGene, and in the case of abo-InGene, it falls short of this simpler model. This apparent disadvantage might be due to the context size for the two models. While the `ri_att` model uses the whole Medline abstract, `emb_att` only uses a small sliding window, which provides a limited scope, and may help explain the poor performance. Further research is needed to test this hypothesis, perhaps by providing a larger window for the neural embedding model, or adapting it to treat entire documents as contextual units.

Another advantage of the `emb_rel` model over `ri_rel` was ease of generation, efficiency, and scalability. Embedding models represent individual concepts as vectors. To create our `ri_rel` search space, we had to first find and extract explicit drug-gene pairs from individual sentences, and then create bags-of-words from their intervening terms, a computationally demanding pre-processing step that took considerable effort to develop, and must be repeated whenever new information is added to the corpus. Furthermore, the resulting vector space is larger as each pair, rather than each entity, must be represented with a unique vector. Given both the level of development, execution effort, and overall performance, the concept-level `emb_rel` model offers clear advantages for relational retrieval.

Our results are not directly comparable to prior work in different domains. The literature is relatively sparse on the application of neural concept embeddings in precision oncology as compared with the general domain. In particular, we are aware of only one paper in the biomedical domain that concerns using neural word embeddings derived from unstructured text (as opposed to neural embeddings derived from semantic predications [41]) for analogical retrieval [45], and this work does not compare attributional and relational models. As mentioned previously, EBC provides an alternative method to `ri_rel` for estimating relational similarity, however it is not directly comparable to our work, since our corpus has not been parsed for grammatical dependencies. Future work, however, will include parsing the corpus to find dependency paths (or leveraging the set provided by the creators of

Table 3
MAP per gene-model combination, and the median MAP per model across all the genes are shown. Best results for each attributional or relational method are underlined, and best result for each gene sheet and overall are shown in **boldface**.

			Median MAP	ABL1	AKT1	ALK	BRAF	CDK4	CDK6	EGFR	ERBB2	FGFR1	FGFR2	FLT3	KDR	PDGFRA	RET	ROS1	SMO
Relational	`ri_rel`	oas-InGene	0.10	0.12	0.04	0.38	0.12	0.01	0.03	0.09	0.05	0.10	0.11	0.15	0.05	0.17	0.14	**1.00**	0.00
		abo-InGene	0.30	0.38	0.06	0.75	0.31	0.01	0.05	0.33	0.25	0.31	0.21	0.30	0.35	0.75	0.28	**1.00**	0.00
		oas-ExGene	0.34	0.37	0.12	0.63	0.40	0.18	0.29	0.34	0.33	0.36	0.28	0.32	0.29	0.38	0.33	0.66	0.74
		abo-ExGene	0.53	0.54	0.22	**1.00**	0.55	0.36	0.48	0.53	0.48	0.51	0.50	0.53	0.43	0.54	0.56	**1.00**	**1.00**
	`emb_rel`	oas-ExGene	0.72	0.71	0.38	**1.00**	0.83	**0.64**	0.74	0.65	**0.67**	0.42	0.53	0.34	0.93	0.92	0.73	0.93	**1.00**
		abo-ExGene	**0.75**	0.74	0.39	**1.00**	**0.85**	0.63	**0.76**	**0.75**	0.65	0.35	0.53	0.29	**0.95**	0.88	0.81	**1.00**	**1.00**
Attributional	`ri_att`	oas-InGene	0.16	0.20	0.09	0.53	0.23	0.27	0.38	0.11	0.09	0.09	0.14	0.14	0.08	0.20	0.52	0.17	0.00
		abo-InGene	0.46	0.69	0.31	0.75	0.68	0.47	0.55	0.47	0.32	0.17	0.27	0.45	0.25	0.58	0.78	0.17	0.00
		oas-ExGene	0.16	0.20	0.09	0.10	0.12	0.09	0.10	0.12	0.15	0.28	0.31	0.32	0.36	0.49	0.55	0.17	0.03
		abo-ExGene	0.14	0.20	0.09	0.05	0.14	0.06	0.05	0.11	0.13	0.40	0.45	0.57	0.47	0.66	0.79	0.15	0.03
	`emb_att`	oas-InGene	0.18	0.20	0.07	0.49	0.20	0.20	0.21	0.09	0.08	0.11	0.16	0.15	0.14	0.25	0.50	0.67	0.00
		abo-InGene	0.23	0.47	0.15	0.38	0.60	0.21	0.24	0.20	0.12	0.18	0.22	0.20	0.55	0.83	0.56	0.67	0.00
		oas-ExGene	0.40	0.32	0.12	0.28	0.51	0.40	0.32	0.24	0.28	0.39	0.42	0.36	0.87	0.94	0.58	0.75	0.58
		abo-ExGene	0.41	0.30	0.12	0.23	0.66	0.44	0.43	0.25	0.29	0.38	0.39	0.37	0.90	**0.96**	0.67	0.75	0.50
Baseline	frequency		0.35	0.46	0.11	0.70	0.36	0.30	0.52	0.34	0.31	0.15	0.17	0.09	0.32	0.48	0.53	0.83	**1.00**
	`rand-vec`	oas-InGene	0.02	0.02	0.01	0.03	0.04	0.01	0.00	0.01	0.02	0.02	0.04	0.03	0.03	0.12	0.14	0.00	0.00
		abo-InGene	0.02	0.02	0.01	0.02	0.02	0.01	0.01	0.02	0.01	0.02	0.04	0.01	0.07	0.23	0.28	0.00	0.00
		oas-ExGene	0.15	0.18	0.03	0.12	0.12	0.07	0.08	0.06	0.08	0.23	0.26	0.24	0.27	0.39	0.49	0.19	0.03
		abo-ExGene	0.27	0.45	0.03	0.21	0.23	0.18	0.14	0.06	0.09	**0.69**	**0.59**	**0.73**	0.60	0.73	**1.00**	0.31	0.03
[a]Included reference drugs				9	21	2	10	4	3	20	20	14	10	7	25	12	3	2	1
[b]Drugs with vector representations				3256															
Final search space unique drugs				1144															
[c]Drugs co-occurring with genes				213	797	50	231	160	183	501	302	295	141	240	177	54	24	33	76

[a] Number of drugs in the reference set copied from Table 1.

[†] Drugs in the vector space after applying filters explained in Section 3.1.1.

[c] Number of drugs available for search per gene concerned. The co-occurrence constraint explained in Section 3.6.3 effectively reduced the number of drugs available for search from 3256 to 1144 *unique* drugs, with an average of 217 available for consideration for each gene (searchable drugs are shared among the genes).

Table 4
Effect of different hyperparameters on model performance. Average increase/decrease is shown for each model across different configurations (abo/oas, InGene/ExGene). Adding context to word vectors consistently improved performance across embedding models, a finding shown in **boldface**. Some of the hyperparameters resulted in a decrease in performance, shown in *italics*.

Hyperparameter	`emb_rel`	`emb_att`	`ri_rel`	`ri_att`
Adding context to word vectors	**increase 40%**	**increase 25%**	n/a	n/a
Subsampling threshold from 0.001 to 0.00001	increase 21%	*decrease 3%*	n/a	n/a
Window size from 5 to 8	increase 17%	*decrease 2%*	n/a	n/a
Replacing RI with RRI	n/a	n/a	*decrease 23%*	increase 250%

EBC [51]) so that EBC can be used. As an attributional counterpart to EBC, Levy and Goldberg's dependency based embeddings [52] can be considered.

Another factor that complicates direct comparison with existing work involves exploration of the space of model hyperparameters, which often resulted in improved performance. Levy et al. provide an extensive description of the set of SGNS hyper-parameters that can be altered to improve the embedding results [18]. Among the many parameters they explain, we chose to examine three – window size, subsampling threshold, and adding context vectors to word vectors. In line with previous work [18], we found that adding context vectors to word vectors consistently improved word embedding results (across all the cross validation configurations). Future work will involve performing a more comprehensive experiment to determine the effect of these and other parameters.

A surprising finding amongst our results was the performance of our random vector based baseline model (`rand-vec`). We expected negligible performance, as random vectors are by design generated with a high probability of being mutually orthogonal or close-to-orthogonal, and as such are not meaningfully similar to one another. While we obtained the expected results with InGene models, those for ExGene were surprisingly productive, particularly the median MAP of 0.27 for abo-ExGene. We believe this phenomenon is explained by the overlap between drugs across gene sheets, providing the model with same vector both as a seed and as target. This theory is supported by the fact that using the `rand-vec` model, we obtained better results with genes that shared many drugs with other genes than those which did not (e.g., FGFR1, FGFR2, FLT3, KDR, PDGFRA, RET). As shown in Table 5, there is a high correlation between drug overlap and `rand-vec` results in the ExGene category, 0.6 and 0.63 for oas-ExGene and abo-ExGene, respectively. The other baseline model was `frequency`, which we compared to the relational models. While with a median MAP of 0.35, the `frequency` model seems relatively strong in terms of its ability to find gene-related drugs, it outperforms neither `ri_rel`, nor `emb_rel`, indicating that these models are more effective than a simple count of co-occurrence in finding the desired relationships.

Table 6
Effect of moving from using reference drugs that had representatives in the search spaces (Original) to the full reference set irrespective of whether the target drugs were represented in a space or not (Full Ref). Best results for attributional or relational categories are underlined, and best result overall is shown in **boldface**. On average the median MAP drops by 0.26. Only the results for best performing models in each category are shown.

Category	Model/configuration	Median MAP Original	Median MAP Full Ref	Drop
Relational	`ri_rel` abo-ExGene	0.53	0.25	0.28
Relational	`emb_rel` **abo-ExGene**	0.75	**0.34**	0.41
Attributional	`ri_att` abo-InGene	0.46	0.15	0.31
Attributional	`emb_att` abo-ExGene	0.41	0.16	0.25
Baseline	`frequency`	0.35	0.16	0.19
Baseline	`rand-vec` abo-ExGene	0.27	0.15	0.12

6. Limitations

We tested our assumptions and techniques using internal cross validation across a set constructed by a single team of PODS curators, and so the methods lack external validation, for example they have not been tested in other contexts or for similar tasks. However, the PODS curators constructed the reference standard independently of the computational work and the main goal of this research was to compare different similarity methods and paradigms.

We faced two problems when dealing with drug-gene relationships in precision oncology. The first problem concerned term to concept mapping (performed by MetaMap), and the other had to do with finding relationships of interest. In the current research, we specifically focused on the latter to fulfil the primary goal of this research – comparative evaluation of different similarity models. Some drugs (119 out of 237 or 50%, Table 1) in the reference set were excluded from evaluation, either because they had no representative in vector space (e.g. because they were not mapped to CUIs by MetaMap,), or because they did not pass the drug filters that we used (which were also based on CUIs). An additional 19 drugs were excluded because of the co-occurrence filter (Section 3.6.3). As such, some true positive results that would have been missed were excluded to allow a "fair" comparison of models.

However, to estimate practical utility, the full reference set should be used. As shown in Table 6, penalizing our models for missing drugs that they do not represent results in a substantial drop in performance. More work is needed to address the limited coverage of therapeutically relevant agents, an issue we hope to address by replacing the concept

Table 5
Spearman Rank-Order Correlation Coefficient values to show a possible link between genes with high drug overlap, and the MAP values for ExGene configurations. Some of the models show high correlation between their results and the degree of overlap (e.g., `rand-vec` abo-ExGene and `ri_att` oas-ExGene) which may help explain the surprisingly high MAP values for those models. MAP values are copied from Table 4 for comparison. High correlation values are shown in **boldface**.

	Relational				Attributional				Baseline		
Model	`emb_rel`		`ri_rel`		`emb_att`		`ri_att`			`rand-vec`	
Config	oas-ExGene	abo-ExGene	oas-ExGene	abo-ExGene	oas-ExGene	abo-ExGene	**oas-ExGene**	**abo-ExGene**	frequency	**oas-ExGene**	**abo-ExGene**
Overlap/MAP Correlation	−0.4	−0.39	−0.32	−0.25	−0.03	0	**0.55**	**0.51**	−0.32	**0.6**	**0.63**
MAP	0.72	0.75	0.34	0.53	0.40	0.41	0.16	0.14	0.35	0.15	0.27

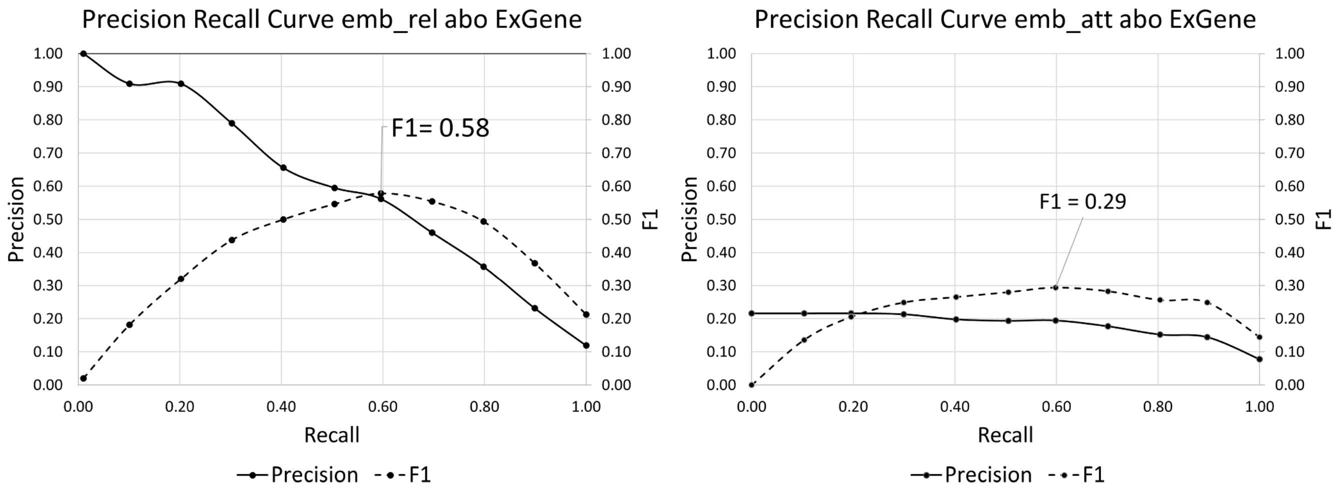


Fig. 4. Precision-recall curves for the best performing relational model, and its attributional counterpart. With abo-ExGene configuration, there is a single score for each potentially relevant drug/gene pair. The graphs also show changes of F1 measure at each interval, with the optimum values highlighted.

extraction component of the system in the future. This may involve further expansion of MetaMap vocabularies, or substitution of an alternative method for the recognition of drug and gene entities that is not dependent on curated knowledge resources, such as the machine learning-based NER used by the PubTator system [53]. This is likely to offer advantages in emerging domains such as precision oncology. While the dependence upon a vocabulary is a limitation that must be overcome to enhance the utility of these methods in such domains, we note that the methods are generally applicable and could be used to identify any sort of biomedical relationship in which the entities of interest can be accurately identified, and seed examples are available. Examples include identifying drug-drug and protein-protein interactions where known examples could be used to find new instances. Moreover, we considered drugs that target the gene in question directly *or* indirectly (interacting with a gene downstream of the main target). Restricting the evaluation to the simpler task of identifying direct drug-gene relationships only may lead to better overall performance, a direction we will explore in future work.

In addition, both the literature and the reference set used in this research were around 3–4 years old. Emerging domains by definition evolve at a rapid pace, and so should the search spaces and reference sets used in information retrieval research projects in these domains. The main goal of this project was to evaluate the utility of different methods in comparison to each other, and so this limitation does not adversely affect the main results of the research. This is especially true since the reference set used for evaluation was also obtained at the same timeframe as the source literature. Using Medline data from after the construction of the reference set could potentially underestimate performance, as newly discovered relationships would be judged as false positive. However, if the resulting models are to be practically useful, developing a system that utilizes the most recent versions of both source and reference material at any given time is an immediate priority.

Finally, while we tried to follow the current literature in selecting model hyper-parameters, the current work should not be considered an exhaustive test of these parameters. It is quite possible that other adjustments could further improve performance.

7. Conclusion

In this research, we compared relational to attributional measures of similarity across a range of representational approaches, for their ability to recover therapeutically important drug-gene relationships. Relational similarity performed better than attributional similarity for this task, demonstrating its utility as a means to identify clinically important biomedical relationships.

Acknowledgements

The work was supported in part by the U.S. National Library of Medicine Grant (R01LM011563), Using Biomedical Knowledge to Identify Plausible Signals for Pharmacovigilance, and by NIGMS grant R01 GM103859, NCI U01 CA180964, Sheikh Bin Zayed Al Nahyan Foundation, Cancer Prevention Research Institute of Texas (CPRIT) Precision Oncology Decision Support Core RP150535 and Data Science and Informatics Core for Cancer Research (RP170668), NCATS grant UL1 TR000371 (Center for Clinical and Translational Sciences), TL1TR001104 (UT Southwestern Center for Translational Medicine), the Bosarge Foundation, and the MD Anderson Cancer Center Support grant (NCI P30 CA016672).

References

[1] A. Johnson, J. Zeng, A.M. Bailey, et al., The right drugs at the right time for the right patient: the MD Anderson precision oncology decision support platform, Drug Discov. Today http://doi.org/10.1016/j.drudis.2015.05.013.

[2] B. Percha, R.B. Altman, Learning the structure of biomedical relationships from unstructured text, PLoS Comput. Biol. 11 (2015). http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4517797/ (accessed 28 Sep 2015).

[3] T. Cohen, D. Widdows, Empirical distributional semantics: methods and biomedical applications, J. Biomed. Inform. 42 (2009) 390–405, https://doi.org/10.1016/j.jbi.2009.02.002.

[4] S. Fathiamini, A.M. Johnson, J. Zeng, et al., Automated identification of molecular effects of drugs (AIMED), J. Am. Med. Inform. Assoc. 23 (2016) 758–765, https://doi.org/10.1093/jamia/ocw030.

[5] T. Mikolov, W. Yih, G. Zweig, Linguistic regularities in continuous space word representations, in: HLT-NAACL, 2013, pp. 746–751. http://www.aclweb.org/anthology/N13-1#page=784 (accessed 26 Oct 2015).

[6] P.D. Turney, Measuring semantic similarity by latent relational analysis, in: Proceedings of the 19th International Joint Conference on Artificial Intelligence, Morgan Kaufmann Publishers Inc, 2005, pp. 1136–1141. http://dl.acm.org/citation.cfm?id=1642475 (accessed 28 Sep 2015).

[7] D.L. Medin, R.L. Goldstone, D. Gentner, Similarity involving attributes and relations: judgments of similarity and difference are not inverses, Psychol. Sci. 1 (1990) 64–69.

[8] P.D. Turney, P. Pantel, From frequency to meaning: vector space models of semantics, J. Artif. Intell. Res. 37 (2010) 141–188.

[9] T.K. Landauer, S.T. Dumais, A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge, Psychol. Rev. 104 (1997) 211.

[10] K. Lund, C. Burgess, Hyperspace analogue to language (HAL): a general model of semantic representation, in: Brain and Cognition, Academic Press Inc JNL-Comp Subscriptions 525 B ST, STE 1900, San Diego, CA 92101-4495, 1996, pp. 5–5.

[11] K. Lund, C. Burgess, Producing high-dimensional semantic spaces from lexical co-occurrence. http://csee.essex.ac.uk/staff/poesio/LAC/LAC03-04/lund_burgess_

96brmic.pdf (accessed 20 Oct 2015).

[12] T.K. Landauer, P.W. Foltz, D. Laham, An introduction to latent semantic analysis, Discourse Process 25 (1998) 259–284.

[13] M. Sahlgren, An Introduction to Random Indexing, in: Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE 2005, 2005. http://www.citeulike.org/group/3795/article/2227659 (accessed 17 Oct 2015).

[14] P. Kanerva, J. Kristoferson, A. Holst, Random indexing of text samples for latent semantic analysis, in: Proceedings of the 22nd Annual Conference of the Cognitive Science Society, 2000. http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.4.6523 (accessed 26 Oct 2015).

[15] M. Sahlgren, The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces, 2006.

[16] T. Mikolov, K. Chen, G. Corrado, et al., Efficient Estimation of Word Representations in Vector Space. ArXiv Prepr ArXiv13013781, Published Online First: 2013. http://seed.ucsd.edu/mediawiki/images/e/e3/Wordembeddings.pdf (accessed 14 Oct 2015).

[17] M. Baroni, G. Dinu, G. Kruszewski, Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2014, pp. 238–247.

[18] O. Levy, Y. Goldberg, I. Dagan, et al., Improving distributional similarity with lessons learned from word embeddings, Trans. Assoc. Comput. Linguist. 3 (2015) 211–225.

[19] R. Collobert, J. Weston, L. Bottou, et al., Natural language processing (almost) from scratch, J. Mach. Learn. Res. 12 (2011) 2493–2537.

[20] Y. Wu, J. Xu, M. Jiang, et al., A study of neural word embeddings for named entity recognition in clinical text, AMIA Annual Symposium Proceedings, American Medical Informatics Association, 2015, p. 1326.

[21] L. De Vine, G. Zuccon, B. Koopman, et al., Medical semantic similarity with a neural language model, Proceedings of the 23rd ACM International Conference on Information and Knowledge Management, ACM, 2014, pp. 1819–1822.

[22] Y. Choi, C.Y.-I. Chiu, D. Sontag, Learning low-dimensional representations of medical concepts, AMIA Summits Transl. Sci. Proc. 2016 (2016) 41.

[23] A.L. Beam, B. Kompa, I. Fried, et al., Clinical Concept Embeddings Learned from Massive Sources of Medical Data, ArXiv Prepr ArXiv180401486 2018.

[24] B. Chiu, G. Crichton, A. Korhonen, et al., How to train good word embeddings for biomedical NLP, in: Proceedings of the 15th Workshop on Biomedical Natural Language Processing, 2016, pp. 166–174.

[25] Y. Zhang, H.-J. Li, J. Wang, et al., Adapting word embeddings from multiple domains to symptom recognition from psychiatric notes, AMIA Summits Transl. Sci. Proc. 2017 (2018) 281–289.

[26] A. Nikfarjam, A. Sarker, K. O'connor, et al., Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features, J. Am. Med. Inform. Assoc. 22 (2015) 671–681.

[27] Z. Jiang, L. Li, D. Huang, et al., Training word embeddings for deep learning in biomedical text mining tasks, in: 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2015, pp. 625–628. http://doi.org/10.1109/BIBM.2015.7359756.

[28] I. Banerjee, M.F. Gensheimer, D.J. Wood, et al., Probabilistic prognostic estimates of survival in metastatic cancer patients (PPES-Met) utilizing free-text clinical narratives, Sci. Rep. 8 2018, https://doi.org/10.1038/s41598-018-27946-5.

[29] D.L. Medin, R.L. Goldstone, D. Gentner, Respects for Similarity, Psychol. Rev. 100 (1993) 254–278.

[30] K.J. Holyoak, P. Thagard, Analogical mapping by constraint satisfaction, Cogn. Sci. 13 (1989) 29–35.

[31] P.D. Turney, M.L. Littman, Corpus-based learning of analogies and semantic relations, Mach. Learn. 60 (2005) 251–278.

[32] P.D. Turney, Similarity of semantic relations, Comput. Linguist. 1 (1997).

[33] S. Shalev-Shwartz, Online learning and online convex optimization, Mach. Learn. 4 (2011) 107–194.

[34] R. JeffreyPennington, C. Manning, GloVe: Global Vectors for Word Representation, http://www-nlp.stanford.edu/projects/glove/glove.pdf (accessed 15 Oct 2015).

[35] D. Widdows, T. Cohen, Reasoning with vectors: a continuous model for fast robust inference, Log. J. IGPL 23 (2015). http://www.oxfordjournals.org/our_journals/igpl/content/current (accessed 14 Oct 2015).

[36] T. Cohen, R.W. Schvaneveldt, T.C. Rindflesch, Predication-based semantic indexing: permutations as a means to encode predications in semantic space, AMIA Annual Symposium Proceedings, American Medical Informatics Association, 2009, p. 114.

[37] T. Cohen, D. Widdows, R.W. Schvaneveldt, et al., Discovering discovery patterns with predication-based Semantic Indexing, J. Biomed. Inform. 45 (2012) 1049–1065.

[38] N. Shang, H. Xu, T.C. Rindflesch, et al., Identifying plausible adverse drug reactions using knowledge extracted from the literature, J. Biomed. Inform. 52 (2014) 293–310.

[39] T. Cohen, D. Widdows, C. Stephan, et al., Predicting high-throughput screening results with scalable literature-based discovery methods, CPT Pharmacomet. Syst. Pharmacol. 3 (2014) 1–9.

[40] T. Cohen, D. Widdows, R. Schvaneveldt, et al., Finding Schizophrenia's prozac emergent relational similarity in predication space, in: Quantum Interaction, Springer, 2011, pp. 48–59. http://link.springer.com/chapter/10.1007/978-3-642-24971-6_6 (accessed 22 Oct 2015).

[41] T. Cohen, D. Widdows, Embedding of semantic predications, J. Biomed. Inform. 68 (2017) 150–166.

[42] SemMedDB_UTH Database Outline. http://skr3.nlm.nih.gov/SemMedDB/index_uth.html (accessed 11 Aug 2015).

[43] G. Salton, The SMART Retrieval System—Experiments in Automatic Document Processing, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1971.

[44] Apache Lucene. http://lucene.apache.org/ (accessed 12 Apr 2018).

[45] D. Newman-Griffis, A.M. Lai, E. Fosler-Lussier, Insights into Analogy Completion from the Biomedical Domain, ArXiv170602241 Cs Published Online First: 7 June 2017. http://arxiv.org/abs/1706.02241 (accessed 7 Aug 2017).

[46] T. Mikolov, I. Sutskever, K. Chen, et al., Distributed representations of words and phrases and their compositionality, in: C.J.C. Burges, L. Bottou, M. Welling, et al. (Eds.). Advances in Neural Information Processing Systems 26. Curran Associates, Inc., 2013, pp. 3111–3119. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf (accessed 28 Sep 2017).

[47] D. Widdows, T. Cohen, The Semantic Vectors Package: New Algorithms and Public Tools for Distributional Semantics. language, 1, 43.

[48] D. Widdows, K. Ferraro, Semantic Vectors: a Scalable Open Source Package and Online Technology Management Application, in: LREC 2008, 2008.

[49] T. Cohen, R. Schvaneveldt, D. Widdows, Reflective Random Indexing and indirect inference: a scalable method for discovery of implicit connections, J. Biomed. Inform. 43 (2010) 240–256, https://doi.org/10.1016/j.jbi.2009.09.003.

[50] A. Drozd, A. Gladkova, S. Matsuoka, Word embeddings, analogies, and machine learning: beyond king-man + woman = queen, in: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, 2016, pp. 3519–3530.

[51] B. Percha, R.B. Altman, J. Wren, A global network of biomedical relationships derived from text, Bioinformatics (2018).

[52] O. Levy, Y. Goldberg, Dependency-Based Word Embeddings, in: ACL (2), 2014, pp. 302–308. http://www.aclweb.org/anthology/P/P14/P14-2050.pdf (accessed 28 Sep 2017).

[53] C.-H. Wei, H.-Y. Kao, Z. Lu, PubTator: a web-based text mining tool for assisting biocuration, Nucl. Acids Res. 41 (2013) W518–W522.