Many Paths Lead to Discovery: Analogical Retrieval of Cancer Therapies

Trevor Cohen[1], Dominic Widdows[2], Lance De Vine[3], Roger Schvaneveldt[4], and
Thomas C. Rindflesch[5]

[1] University of Texas School of Biomedical Informatics at Houston
[2] Microsoft Bing
[3] Queensland University of Technology
[4] Arizona State University
[5] National Library of Medicine

Abstract. This paper addresses the issue of analogical inference, and its potential role as the mediator of new therapeutic discoveries, by using disjunction operators based on quantum connectives to combine many potential reasoning pathways into a single search expression. In it, we extend our previous work in which we developed an approach to analogical retrieval using the Predication-based Semantic Indexing (PSI) model, which encodes both concepts and the relationships between them in high-dimensional vector space. As in our previous work, we leverage the ability of PSI to infer predicate pathways connecting two example concepts, in this case comprising of known therapeutic relationships. For example, given that *drug x* TREATS *disease z*, we might infer the predicate pathway *drug x* INTERACTS_WITH *gene y* ASSOCIATED_WITH *disease z*, and use this pathway to search for drugs related to another disease in similar ways. As biological systems tend to be characterized by networks of relationships, we evaluate the ability of quantum-inspired operators to mediate inference and retrieval across multiple relations, by testing the ability of different approaches to recover known therapeutic relationships. In addition, we introduce a novel complex vector based implementation of PSI, based on Plate's Circular Holographic Reduced Representations, which we utilize for all experiments in addition to the binary vector based approach we have applied in our previous research.

Keywords: Distributional Semantics, Vector Symbolic Architectures, Literature-based Discovery, Abductive Reasoning

1 Introduction

The field of Literature-based Discovery (LBD) has been an important application area for quantum-inspired methodologies in recent years [1, 2]. In particular, the ability of quantum-inspired approaches to measure implicit relatedness between composite representations of concepts holistically offers advantages in scalability and efficiency over rule-based approaches that require the decomposition of conceptual representations into their atomic components. In previous work, we have shown that these holistic approaches can be used to facilitate analogical retrieval across a set of object-relation-object triplets, or predications extracted from the biomedical literature, to solve simple proportional analogy problems of the form "*A* is to *B* as *?* is to *C*" [2]. This mechanism provides the means to infer the predicate pathway connecting a disease to a drug that is known to treat it, and also to use the vector representation of this pathway to search for treatments connected to some other disease in the same way. However, the identification and re-use of individual pathways is of limited utility for the discovery of new therapies, as drugs tend to activate multiple pathways and targets simultaneously [3]. This suggests that modeling analogical retrieval across multiple pathways may facilitate the identification of novel therapeutic relationships. In this paper, we use quantum models of disjunction and superposition to achieve this end, allowing us to combine many compound stimuli to perform searches that would be brittle and computationally prohibitive using traditional symbolic methods. In doing so, we create a superposition of compound systems that has not been (and probably cannot be) represented as a product of two individual simple systems, a phenomenon known in the quantum literature as "entanglement". It is our hypothesis that modeling multiple pathways will improve the quality of analogical retrieval, and we evaluate this hypothesis by comparing the extent to which retrieval across individual and multiple pathways facilitates recovery of a held out set of cancer therapies. We evaluate these approaches in both binary and complex vector space, leveraging recent enhancements to the Semantic Vectors package [4].

2 Background

Distributional models of language, such as Latent Semantic Analysis (LSA) [5] derive human-like estimates of the semantic relatedness between terms from large volumes of unannotated natural language text. A desirable property of some distributional models is the ability to learn meaningful associations between terms that do not co-occur directly in the text concerned. This ability has been termed *indirect inference* and it has been argued that it is essential to LSA's human-like performance on a number of cognitive tasks [5]. Indirect inference is also a fundamental concern of the field of Literature-based Discovery (LBD), which aims to promote scientific discovery by identifying meaningful connections between terms, and concepts, in the scientific literature that have not yet occurred together in any published document [6], and several authors have explored the ability of distributional models to facilitate discoveries of this nature [7–9]. A limitation of the use of these models for LBD is that they capture general relatedness between terms or concepts only, without encoding the nature of the relationships concerned. As economic constraints limit the number of candidate therapies that can be advanced for further testing, there is a pressing need for the development of methods that selectively emphasize plausible therapeutic hypotheses. In recognition of the limitations of general relatedness, LBD researchers have recently begun exploring the notion of a *discovery pattern*, a sequence of relationship types that suggests a potential discovery [10]. For example, if a certain drug were known to inhibit a gene associated with a particular disease, it would follow that this drug may be a potential candidate therapy for this disease. These patterns have largely been pursued using rule-based approaches in which concepts, and the relationships between them, are represented as discrete entities each of which must be explored stepwise to find a pathway from treatment to disease (see for example [11]). However, given the rapid expansion of the number of logical connections between concepts in the biomedical literature [12, 13], the development of methods to directly identify meaningful connections across specific patterns of relationships is a desirable alternative. To this end, we have developed PSI [2, 14, 15], which encodes concepts and their relations as vectors in high-dimensional space, facilitating efficient search and indirect inference without the need to unpack and explore individual relationships. In previously published work, we have shown that PSI can be used to infer relationship paths (such as INHIBITS-ASSOCIATED WITH) from one concept to another, and that these inferred pathways can be used to direct search through predication space for concepts related to a third concept in the same way [2]. However, the identification and re-use of individual pathways is of limited utility for discovery of new therapies, as drugs tend to activate multiple pathways simultaneously [3]. In this paper, we evaluate the utility of the PSI model as a means to identify therapeutic relationships by accommodating drug-disease relationships that include multiple relationship paths. In some cases, the quantum disjunction operator [16] is applied to measure the relatedness between concepts that are connected across multiple relationship paths, and in others we use superposition of vectors to achieve this end. In the section that follows, we introduce the fundamental operations that mediate PSI, and the notation used to describe them. We then illustrate the way in which analogy occurs in PSI space, and proceed to describe the empirical component of this work, in which we use analogical relations drawn from one disease, or set of diseases, to seek treatments for another.

3 Mathematical Structure and Methods

The methods in this paper all use high-dimensional vectors to represent concepts. There are many ways of generating such representations. Ours is based upon the Random Indexing paradigm using terminology as described in [9] and developed in [2], in which *semantic vectors* are built as superpositions of randomly generated *elemental vectors*, during the process of training. Throughout this paper we will write $E(\mathrm{X})$ and $S(\mathrm{X})$ for the elemental and semantic vectors associated with the concept X. In addition to concept vectors, we include vectors for relations. For example, $E(\mathrm{R})$ would denote the elemental vector for the relation R. Many relationships are directional, and we will use $\mathrm{R}_{\mathrm{inv}}$ to denote the inverse of R, so that $\mathrm{A\ R\ B}$ and B $\mathrm{R}_{\mathrm{inv}}$ A carry the same external meaning (though they may in some cases be represented by different vectors). To encode typed relations into high-dimensional vector spaces, we utilize two members of a family of representational approaches collectively known as Vector Symbolic Architectures [17]. VSAs originated from Smolenksy's tensor-product based approach [18], but differ from it in that they depend on vector operations that produce products of the same dimensionality as the component vectors. The VSAs we will use in our experiments are Kanerva's Binary Spatter Code (BSP) [19], which uses high-dimensional binary vectors as a representational unit, and Plate's Circular Holographic Reduced Representation (CHRR) [20], which uses circular vectors, vectors in which each dimension represents an angle between $-\pi$ and π. CHRRs have recently been used to encode information related to word order in a distributional model [21]. Before we discuss further distinctions between these models, we will describe the fundamental operations of VSAs, which are common to both of them. The primary operations facilitated by VSAs are *binding* and *bundling*. Binding is a multiplication-like operator through which two vectors are combined to form a third vector C that is dissimilar from either of its component vectors A and B. We will use the symbol "$\otimes$" for binding, and the symbol "$\oslash$" for the inverse of binding throughout this paper. It is important that this operator be invertible: if C = A $\otimes$ B, then A $\oslash$ C = A $\oslash$ (A $\otimes$ B) = B. In some models, this recovery may be approximate, but the robust nature of the representation guarantees that A $\oslash$ C is similar enough to B that B can easily be recognized as the best candidate for A $\oslash$ C in the original set of concepts. Thus the invertible nature of the bind operator facilitates the retrieval of information encoded during the binding process. Bundling is an addition-like operator, through which superposition of vectors is achieved. For example, vector addition followed by normalization is a commonly employed bundling operator. Bundling results in a vector that is maximally similar to its component vectors. We will write the usual "+" for bundling, and the computer science "$+=$" for "bundle the left hand side with the right hand side and assign the outcome to the symbol on the left hand side." So for example, $S(\mathrm{A}) \mathrel{+}= E(\mathrm{B})$ could also be expressed as $S(\mathrm{A})$ = $S(\mathrm{A})$ + $E(\mathrm{B})$, and is a standard operation in training. Table 1 summarizes the differences between the binary (BSP) and complex (CHRR) vector implementations used in this work.

Table 1. Comparison between CHRR and BSP

Implementation	**Complex/Circular**	**Binary**
Semantic vectors $S(\mathrm{X})$	Complex (circular) vectors d $O(1000)$	Binary vectors d $O(10{,}000)$
Elemental vectors $E(\mathrm{X})$	Dense complex $[-\pi, \pi]$	Dense binary $\{0,1\}$
Bundling (Superposition)	Pairwise vector sum	Majority vote
Binding	Convolution (mod 2π addition of angles)	Pairwise XOR (mod 2 addition)
Release	Convolution with inverse	Pairwise XOR

In the case of the spatter code, pairwise exclusive or (XOR) is used as a binding operator: $X \otimes Y = X$ XOR Y. As it is its own inverse, the binding and decoding processes are identical ($\otimes$=$\oslash$). For bundling, the spatter code employs a majority vote: if the component vectors of the bundle have more ones than zeros in a dimension, this dimension will have a value of one, with ties broken at random (for example, bundling the vectors 011 and 010 may produce either 010 or 011 with equal probability). In the case of CHRR, binding is accomplished using circular convolution, accomplished by pairwise multiplication: $X \otimes Y = \{X_1Y_1, X_2Y_2, ...X_{n-1}Y_{n-1}, X_nY_n\}$, which is equivalent to addition of the phase angles of the circular vectors concerned, as they are of unit length. The inverse of binding is obtained by binding to the inverse of the vector concerned: $X \oslash Y = X \otimes Y^{-1}$, where the inverse of a vector Y, Y^{-1} is the vector with a phase angle that when added to that of Y produces a phase angle of 0. As each dimension in a circular vector can be represented as a vector on the unit circle, superposition is accomplished in a pairwise manner by adding the unit circle vectors in a given dimension, and normalizing the result for each circular component of the vector. In the implementation used in our experiments, normalization is delayed until after training has concluded, so that the sequence in which superposition occurs is not relevant. Once a vector representation for a concept has been built up by binding and/or bundling, it is possible to apply an operator that reverses the binding process to the vector as a whole, allowing us to direct search in PSI space without explicitly representing the individual relations of a concept. This property is appealing for the purpose of modeling analogy, as similarity is measured on the basis of a superposed product without the need to decompose it [20].

Predication-based Semantic Indexing: PSI takes as input sets of concept-relation-concept triplets, or predications. For these experiments, as well as those in our previous work, the PSI space is derived from a set of 22,669,964 predications extracted from citations added to MEDLINE over the past decade by the SemRep natural language processing system [22], which extracts predications from biomedical text using domain knowledge in the Unified Medical Language System [23]. For example, the predication "fluoxetine TREATS Major Depressive Disorder" (MDD) is extracted from "patients who have been successfully treated with fluoxetine for major depression." In a recent evaluation of SemRep, Kilicoglu et al. report .75 precision and .64 recall (.69 f-score) [24]. The first step in PSI is the generation of semantic and elemental vectors for each concept, $S(\mathrm{C})$ and $E(\mathrm{C})$. We also generate elemental vectors for each relation, or predicate $E(\mathrm{P})$. We then encode each predication in the set by binding $E(\mathrm{C}_1)$ to $E(\mathrm{P})$ and bundling this into $S(\mathrm{C}_2)$. The reverse of this process is also performed. In practice statistical weighting metrics are used to decrease the influence of frequently occurring concepts, and in some cases predicates. In the implementation we utilized for these experiments, we used inverse document frequency (*idf*) as a global weighting metric, and log(1+frequency of predication) as a local metric. For example, the predication "thalidomide INHIBITS cyclooxygenase 2" (cox2) would be encoded as follows:

$$S(\mathrm{thalidomide}) {+}{=} E(\mathrm{INHIBITS}) \otimes E(\mathrm{cox2}) \times \mathit{idf}(\mathrm{cox2}) \times gw$$

$$S(\mathrm{cox2}) {+}{=} E(\mathrm{INHIBITS_{inv}}) \otimes E(\mathrm{thalidomide}) \times \mathit{idf}(\mathrm{thalidomide}) \times gw$$

$$\mathit{idf}(\mathrm{C}) = \log \frac{\text{total predications}}{\text{predications containing C}}$$

$$\mathit{gw} = \log\left(1\ +\ \text{occurrences of thalidomide INHIBITS cox2}\right)$$

For the sake of brevity, we will describe future encoding operations without explicitly referring to *idf* or *gw*. This process is repeated across all of the predications in the database, to generate a set of trained semantic vectors for each concept.

Analogical Retrieval: As the binding process is invertible, it is possible to retrieve a dual-predicate path connecting two concepts:

Training:

$$S(\mathrm{multiple\_myeloma})(\mathrm{MM}) {+}{=} E(\mathrm{ASSOCIATED\_WITH}) \otimes E(\mathrm{cox2})$$

$$S(\mathrm{thalidomide}) {+}{=} E(\mathrm{INHIBITS}) \otimes E(\mathrm{cox2})$$

Inference:

$$\begin{aligned}S(\mathrm{MM}) \oslash S(\mathrm{thalidomide}) \approx & E(\text{ASSOCIATED\_WITH}) \otimes E(\mathrm{cox2}) \\ & \oslash (E(\text{INHIBITS}) \otimes E(\mathrm{cox2})) \\ \approx & E(\text{ASSOCIATED\_WITH}) \oslash E(\text{INHIBITS}) \\ & \otimes E(\mathrm{cox2}) \oslash E(\mathrm{cox2}) \\ \approx & E(\text{ASSOCIATED\_WITH}) \oslash E(\text{INHIBITS})\end{aligned}$$

These inferred relationships can then be used to find concepts relating to a third concept in the same way that these cue concepts relate to one another. The ability of VSAs to capture relational similarity has led to their utilization as a means to model aspects of analogical thought [25, 20, 26]. In previous work, we have shown that this facility of VSAs can be used to solve proportional analogy problems, by inferring predicate paths between cue concepts, and using the vector representations of these paths to direct search through predication space [2]. This is accomplished with either the retrieved path (e.g. $E(\text{ASSOCIATED\_WITH}) \oslash E(\text{INHIBITS})$) or the noisy approximation of it derived from the cue concept vectors (e.g. $S(\mathrm{MM}) \oslash S(\mathrm{thalidomide})$). The vector representations of these inferred paths can be applied to another concept to direct search through PSI space to facilitate analogical retrieval as follows:

Training:

$$\begin{aligned}S(\mathrm{fluoxetine}) {+}{=} & E(\text{INHIBITS}) \otimes E(\mathrm{serotonin}) \\ S(\mathrm{MDD}) {+}{=} & E(\text{ASSOCIATED\_WITH}) \otimes E(\mathrm{serotonin})\end{aligned}$$

Inference:

$$\begin{aligned}S(\mathrm{MDD}) & \oslash (E(\text{ASSOCIATED\_WITH}) \oslash E(\text{INHIBITS})) \\ \approx & E(\text{ASSOCIATED\_WITH}) \otimes E(\mathrm{serotonin}) \\ & \oslash (E(\text{ASSOCIATED\_WITH}) \oslash E(\text{INHIBITS})) \\ \approx & E(\text{ASSOCIATED\_WITH}) \oslash E(\text{ASSOCIATED\_WITH}) \\ & \otimes E(\text{INHIBITS}) \otimes E(\mathrm{serotonin}) \\ \approx & E(\text{INHIBITS}) \otimes E(\mathrm{serotonin}) \approx S(\mathrm{fluoxetine})\end{aligned}$$

4 Multiple Pathways and Quantum Disjunction

In previous work [2], we restricted our study of analogical retrieval to proportional analogies in which a single predicate path (consisting of one or two predicates) inferred from a cue pair (e.g. $S(\mathrm{MM}) \oslash S(\mathrm{thalidomide})$) is used to direct search toward concepts connected to a third target concept (e.g. $S(\mathrm{MDD})$ in the same way as the cue pair relate to one another (e.g. z INHIBITS y, y ASSOCATED_WITH x), thereby solving a proportional analogy problem of the form "what relates to MDD as thalidomide relates to MM". However, analogies used in science tend to have more complex structure than this [27], and drugs tend to be connected to the diseases they treat across networks involving multiple biological entities [3]. Consequently, in this paper we evaluate the ability of PSI to perform analogical inference and retrieval across multiple predicate paths. In order to do so, we require a way to measure the similarity between an individual vector, representing a potential treatment, and a set of vectors representing the permitted paths from the target disease to this vector. One approach we evaluate in this paper involves comparing candidate therapies to the superposition of a set of inferred predicate paths. However, as we would like to identify both treatments that are strongly connected across a single path (such as INHIBITS:ASSOCIATED_WITH) and treatments that are connected across multiple paths (such as INHIBITS:ASSOCIATED_WITH; INTERACTS_WITH:CAUSES), we also utilize for this purpose the span of vectors, described in logic as the quantum disjunction operator by Birkhoff and Von Neumann [28] and applied to information retrieval by Widdows and Peters [16]. This operator measures the proportion of a vector (in our case a treatment) that can be projected onto a subspace spanned by a set of component vectors (in our case the predicate paths of interest bound to the disease of interest). In addition, we introduce a binary vector approximation of this operator, compared with the continuous implementation in Table 2.

Table 2. Continuous and Binary Implementations of Quantum Disjunction

Implementation Steps	**Continuous**	**Binary**
(1) Component vectors	Real/complex vectors $d\ O(1000)$	Binary vectors $d\ O(10{,}000)$
(2) Orthogonalize vectors	A - A's projection on B such that cos($\hat{A}$,B) = 0	Introduce/eliminate identical dimensions until HD($\hat{A}$,B) = $\frac{d}{2}$.
(3) Projection	Project into subspace	Compare with component vectors
(4) Comparison	Cosine of angle between projection and original vector	Count of overlap with orthogonalized component vectors

5 Evaluation

To evaluate PSI's ability to mediate analogical inference, we utilize the same set of 22,669,964 predications as in our previous work. From this, we extract predications involving predicates in the set $\{$AFFECTS; AUGMENTS; CAUSES; DISRUPTS; INHIBITS; PREDISPOSES; STIMULATES; ASSOCIATED_WITH; COEXISTS_WITH; INTERACTS_WITH$\}$, which were selected on the basis of their potential as justification for therapeutic hypotheses. Predications with the predicate TREATS, and any predications involving a direct relationship between a pharmaceutical substance (UMLS semantic type "**phsu**") and neoplastic process (UMLS semantic type "**neop**", which represents types of cancer), were excluded from training. In addition, predications involving a concept with a global frequency greater than or equal to 100,000 were excluded, as these concepts tend to be general in nature and relatively uninformative. From the remaining predications, we generated two PSI spaces, one of which utilized binary vectors with dimension 32,000, and one of which utilized complex vectors with dimension of 4,000. We will refer to these spaces as BSP and CHRR respectively, in accordance with the methodology used to generate them. As a test set, we extracted 1,158 types of cancer (or neoplastic processes: UMLS semantic type "**neop**") with the prerequisite that each extracted neoplastic process occur in a TREATS relationships with a pharmaceutical substance represented in our spaces. Inclusion in the set does not, however, guarantee that a dual-predicate pathway between the cancer concerned and this treatment exists. We use this set to evaluate analogical retrieval, with the following approaches.

Collective Cues: This is an approach we have pursued in our recent work [29], in which dual-predicate pathways are inferred from a set of 48,204 known TREATS relationships between diseases or syndromes (UMLS semantic type "**dsyn**") and pharmaceutical substances (UMLS semantic type "**phsu**"). For each pair, the dual-predicate path connecting the concepts concerned is inferred by generating the composite cue vector $S(\mathrm{dysn})$ $\oslash$ $S(\mathrm{phsu})$ and searching through the set of vectors generated by pairwise combination of the vectors representing individual predicate paths, $E(\textsc{pred}1) \oslash E(\textsc{pred}2)$. From the original set of seventeen predicate vectors (7 directional x 2 = 14 + 3 that commute = 17), a set of 136 binary ($\frac{17\times16}{2}$) and a set of 272 complex (17×16) dual-predicate path vectors were generated. With complex vectors, twice as many paths are generated, as unlike XOR, the convolution operator is not its own inverse - the order of application of operators is of importance. Paths connecting pharmaceutical substances and diseases or syndromes were inferred by retrieving dual-predicate path vectors with a similarity to the composite cue vector $S(\mathrm{dysn}) \oslash S(\mathrm{phsu})$ greater than 1 SD above the mean similarity between 1000 randomly generated vectors of the same vector type and dimensionality. The number of times each possible predicate path was retrieved with a similarity above this threshold to the cue vector was counted, and the five most popular paths for both binary and complex vector spaces were retained. These paths are illustrated in Table 3. Most paths are readily interpretable, as the ASSOCIATED_WITH predicate links diseases to related biological entities, and a drug that interacts with such entities may be a plausible therapy. Some pathways are more difficult to interpret, and we refer the interested reader to a related publication [29] concerned primarily with identification, interpretation and application of such pathways. Of interest for our present purposes, directionality of the predicate paths is encoded in the complex case only. So complex pathways are easier to interpret, and binary pathways are less constrained.

Individual Cues: Cues in this case consist of other neoplastic processes drawn from the set. For each neoplastic process, we draw at random another neoplastic process, *cue_neop*, and retrieve all of its TREATS relationships from the predication database. The dual predicate paths are compared to the subspace derived from this set of treatments using the quantum disjunction operator. The components of this subspace (prior to orthogonalization) consist of the set { $S(\mathrm{cue\_neop}) \oslash S(\mathrm{treatment_1})$ ...$S(\mathrm{cue\_neop})$ $\oslash$ $S(\mathrm{treatment_n})$ }. Only pathways with an association strength above empirically determined thresholds of 6SD (binary vectors) and 2.5SD (complex vectors) above the mean pairwise relatedness between 1000 randomly generated vectors of the same type and dimensionality are retained. Random cue selection is repeated until an example with more than one above-threshold predicate path is found.

Table 3. Most Popular Predicate Paths in Binary and Complex Space

Binary	**Count**	**Complex**	**Count**
ASSOCIATED_WITH COEXISTS_WITH	925	COEXISTS_WITH ASSOCIATED_WITH	900
ASSOCIATED_WITH INTERACTS_WITH	201	ASSOCIATED_WITH INTERACTS_WITH	827
ASSOCIATED_WITH INHIBITS	82	ASSOCIATED_WITH INHIBITS	284
COEXISTS_WITH CAUSES	71	ASSOCIATED_WITH COEXISTS_WITH	264
CAUSES-INV INTERACTS_WITH	69	COEXISTS_WITH AFFECTS	248

Application of Pathways: To evaluate the ability of our models to infer (i.e. rediscover) TREATS relationships pertinent to the types of cancer under evaluation, we generate a composite cue vector, or subspace, from the vector representing the target neoplastic process, $S(\mathrm{target\_neop})$, using three approaches. In the first of these, which we will term MAX, only the most strongly associated predicate path is utilized. The cue vector is constructed as $S(\mathrm{target\_neop}) \oslash E(\mathrm{predicate\ path_1})$. In the second, which we will term SUP, all of the relevant predicate paths (*n*=5 for composite cues, and $n >=$ 2 for individual cues) are superposed to generate a composite cue vectors constructed as $S(\mathrm{target\_neop}) \oslash E(\mathrm{predicate\ path_1})$ + $S(\mathrm{target\_neop}) \oslash E(\mathrm{predicate\ path_2})$ + ... + $S(\mathrm{target\_neop}) \oslash E(\mathrm{predicate\ path_n})$. In the third approach, which we will designate SUB, the same set of vectors used to generate SUP are combined, but rather than superposing these we generate a subspace from them using the quantum disjunction operator. For each of the 1,158 target neoplasms, the MAX, SUP and SUB cues are compared with the semantic vectors for all of the pharmaceutical substances in the PSI space (*n* = 16,337) . For each of the three cue types we retrieve all of the pharmaceutical substances with a similarity to the composite cues above a series of statistically determined thresholds of association for each of the 1,158 target neoplasms. This approach is used rather than a fixed number of nearest neighbors, as we anticipate that only a subset of target neoplasms will be connected in accordance with the dual predicate pathway cues. With a threshold, concepts connected in this way should be selectively retrieved.

6 Results and Discussion

Figures 1 and 2 show the results of our experiments in binary and complex space respectively. The *y* axis shows the total number of rediscovered therapeutic relationships at a given threshold for the set of 1,158 neoplastic processes. The *x* axis shows the mean number of candidate therapies retrieved at this threshold, so higher threshold values correspond to lower values on the *x* axis. Therefore, one interpretation of the results

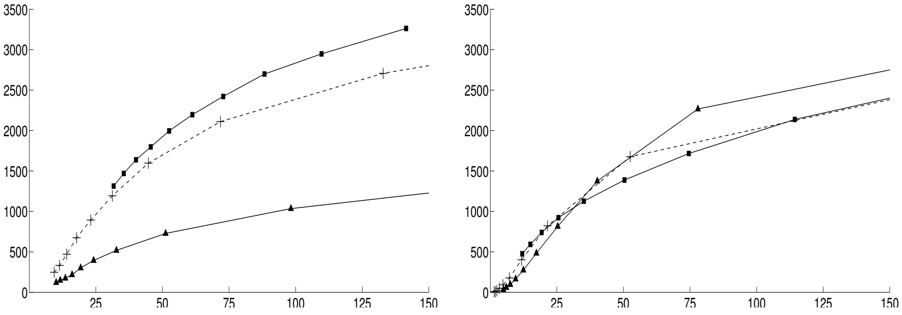

Fig. 1. Binary Vector Results. Left: Collective Cues. Right: Individual Cues. ■=SUB; +=SUP; ▲=MAX. Y axis = no. discoveries. X axis = mean no. retrieved.

in Figure 1 (left) is that the binary SUB model recovered approximately two treatments per disease in the test set while returning on average sixty results per search. However, this is not to say that treatments were found for every test case. The most productive models returned treatments in only around one third of the cases, even at the lowest thresholds tested. It may be the case that this approaches the proportion of this test set for which TREATS relationships corresponding to dual-predicate paths exist, and that models incorporating longer paths are required to recover the remaining treatments.

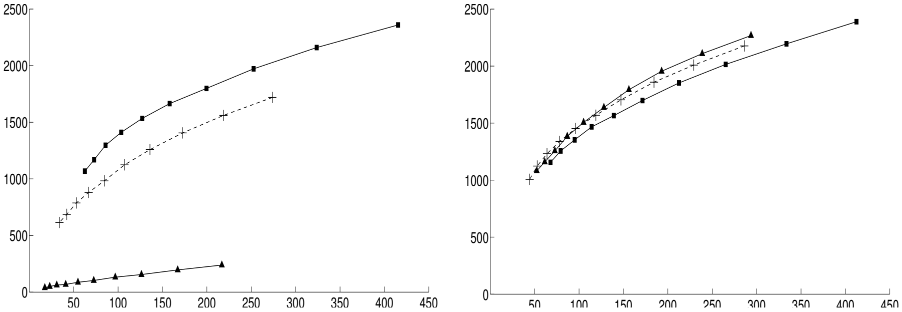

Fig. 2. Complex Vector Results. Left: Collective Cues. Right: Individual Cues. A. ■=SUB; +=SUP; ▲=MAX. Y axis = no. discoveries. X axis = mean no. retrieved.

With respect to the collective cues (left), there is a clear pattern of improved recovery for the models that capture connectedness across multiple pathways, with the quantum disjunction based SUB (■) model retrieving more treatments than the SUP (+) model, and both of these retrieving considerably more than MAX (▲). With individual cues (right) the distinction is less clear, with SUP and, in the binary case, SUB having a slight advantage over MAX at higher thresholds only, and MAX most productive at lower thresholds. This can be explained in part by the ASSOCIATED_WITH:INTERACTS_WITH pattern, which captures drug-gene-disease relationships. This was the second-ranked path for both collective cue sets, and consequently was not considered by MAX in these cases. However, this predicate path was usually the highest-ranked, and as such the predominant pathway used by MAX, with individual cues. One interpretation of this finding is that tight constraints on analogical retrieval are particularly hazardous when mapping from one domain (diseases other than cancer) to another. Overall, the quantum disjunction based SUB model with collective cues recovered the most treatments.

7 Conclusion

In this paper, we evaluate the ability of the PSI model to mediate retrieval across multiple relationships holistically and efficiently, without decomposing the representation of either the cue or the target. We find that models that facilitate retrieval across multiple predicate paths are better able to recover therapeutic relationships when the scope of these paths is relatively broad. The best performance was obtained with the quantum disjunction operator using collective cues derived from diseases other than cancer. As the predicate pathways concerned were not readily retrieved from individual cancer cues, the advantages of this model can be attributed to the application of relations derived from another domain, the hallmark of scientific analogy [27].

Acknowledgments: This research was supported by the US National Library of Medicine grant R21 LM010826, and the Intramural Research Program of the National Institute of Health, National Library of Medicine.

References

1. P. Bruza, “Semantic space: Bridging the divide between cognitive science, information processing technology and quantum mechanics.,” in *Proc Inform Symp on Inform Tech (ITsim ’08)*, pp. 1–9, 2008.
2. T. Cohen, D. Widdows, R. Schvaneveldt, and T. Rindflesch, “Finding schizophrenia’s prozac: Emergent relational similarity in predication space,” in *Proc 5th International Symposium on Quantum Interactions. Aberdeen, Scotland. Springer-Verlag Berlin, Heidelberg.*, 2011.
3. J. Dudley, E. Schadt, *et al.*, “Drug discovery in a multidimensional world: systems, patterns, and networks,” *J Cardiovasc Transl Res*, vol. 3, no. 5, pp. 438–47, 2010.
4. D. Widdows, T. Cohen, and L. De Vine, “Real, complex, and binary semantic vectors,” in *Proc Sixth Intl Symp on Quantum Interactions, Paris, France.*, 2012.
5. T. Landauer and S. Dumais, “A solution to plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge,” *Psych. Review*, vol. 104, pp. 211–240, 1997.
6. D. R. Swanson, “Fish oil, raynaud’s syndrome, and undiscovered public knowledge.,” *Perspect Biol Med*, vol. 30, no. 1, pp. 7–18, 1986.
7. M. D. Gordon and S. Dumais, “Using latent semantic indexing for literature based discovery,” *Journal of the American Society for Information Science*, vol. 49, pp. 674–685, 1998.
8. R. Cole and P. Bruza, “A bare bones approach to Literature-Based discovery: An analysis of the Raynaud’s/Fish-Oil and Migraine-Magnesium discoveries in semantic space,” *Hoffman, A. and Motoda, H. and Scheffer, T. (eds.) Discovery Science, 8th International Conference, DS 2005, Singapore, October 8-11, LNAI, Springer.*, vol. 3735, pp. 84–98, 2005.

9. T. Cohen, R. Schvaneveldt, and D. Widdows, “Reflective random indexing and indirect inference: A scalable method for discovery of implicit connections,” *Journal of Biomedical Informatics*, vol. 43, pp. 240–256, Apr. 2010.
10. D. Hristovski, C. Friedman, T. Rindflesch, and B. Peterlin, “Literature-based knowledge discovery using natural language processing,” *in Bruza P, Weeber M (eds). Literature Based Discovery. Springer-Verlag Berlin Heidelberg.*, pp. 133–152, 2008.
11. C. Ahlers, D. Hristovski, H. Kilicoglu, and T. Rindflesch, “Using the Literature-Based discovery paradigm to investigate drug mechanisms,” *AMIA Annu Symp Proc.*, pp. 6–10, 2007.
12. J. Wren, “The ‘open discovery’challenge,” *Literature-based discovery*, pp. 39–55, 2008.
13. D. R. Swanson, “Medical literature as a potential source of new knowledge.,” *Bulletin of the Medical Library Association*, vol. 78, 1990.
14. T. Cohen, R. Schvaneveldt, and T. Rindflesch, “Predication-based semantic indexing: Permutations as a means to encode predications in semantic space,” *AMIA Annu Symp Proc.*, pp. 114–8, 2009.
15. T. Cohen, D. Widdows, R. W. Schvaneveldt, and T. C. Rindflesch, “Logical leaps and quantum connectives: Forging paths through predication space,” in *Proc AAAI Fall Symp on Quantum Informatics for Cognitive, Social, and Semantic Processes*, pp. 11–13, 2010.
16. D. Widdows and S. Peters, “Word vectors and quantum logic experiments with negation and disjunction,” in *Proc 8th Math. of Language Conference.*, (Bloomington, Indiana.), 2003.
17. R. W. Gayler, “Vector symbolic architectures answer jackendoff’s challenges for cognitive neuroscience,” in *In Peter Slezak (Ed.), ICCS/ASCS International Conference on Cognitive Science*, (Sydney, Australia. University of New South Wales.), pp. 133–138, 2004.
18. P. Smolensky, “Tensor product variable binding and the representation of symbolic structures in connectionist systems,” *Artificial intelligence*, vol. 46, no. 1-2, pp. 159–216, 1990.
19. P. Kanerva, “Binary spatter-coding of ordered k-tuples,” *Artificial Neural Networks—ICANN 96*, pp. 869–873, 1996.
20. T. A. Plate, *Holographic Reduced Representation: Distributed Representation for Cognitive Structures*. Stanfpord, CA.: CSLI Publications, 2003.
21. L. De Vine and P. Bruza, “Semantic oscillations: Encoding context and structure in complex valued holographic vectors,” *Proc AAAI Fall Symp on Quantum Informatics for Cognitive, Social, and Semantic Processes*, 2010.
22. T. C. Rindflesch and M. Fiszman, “The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text,” *Journal of Biomedical Informatics*, vol. 36, pp. 462–477, 2003.
23. O. Bodenreider, “The unified medical language system (UMLS): integrating biomedical terminology,” *Nucleic Acids Research*, vol. 32, no. Database Issue, p. D267, 2004.
24. H. Kilicoglu, M. Fiszman, G. Rosemblat, S. Marimpietri, and T. C. Rindflesch, “Arguments of nominals in semantic interpretation of biomedical text,” in *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, pp. 46–54, 2010.
25. C. Eliasmith and P. Thagard, “Integrating structure and meaning: A distributed model of analogical mapping,” *Cognitive Science*, vol. 25, no. 2, pp. 245–286, 2001.
26. P. Kanerva, “What we mean when we say “What’s the dollar of mexico?”: Prototypes and mapping in concept space,” in *Proc AAAI Fall Symp on Quantum Informatics for Cognitive, Social, and Semantic Processes*, 2010.
27. K. J. Holyoak and P. Thagard, *Mental Leaps: Analogy in Creative Thought*. MIT Press, 1995.
28. G. Birkhoff and J. V. Neumann, “The logic of quantum mechanics,” *The Annals of Mathematics*, vol. 37, no. 4, pp. 823–843, 1936.
29. T. Cohen, D. Widdows, R. Schvaneveldt, P. Davies, and T. Rindflesch, “Discovering discovery patterns with predication-based semantic indexing,” *Journal of Biomedical Informatics*, vol. [epub ahead of print], July 2012.