

ORIGINAL ARTICLE

Predicting High-Throughput Screening Results With Scalable Literature-Based Discovery Methods

T Cohen¹, D Widdows², C Stephan³, R Zinner⁴, J Kim⁵, T Rindflesch⁶ and P Davies³

The identification of new therapeutic uses for existing agents has been proposed as a means to mitigate the escalating cost of drug development. A common approach to such repurposing involves screening libraries of agents for activities against cell lines. *In silico* methods using knowledge from the biomedical literature have been proposed to constrain the costs of screening by identifying agents that are likely to be effective *a priori*. However, results obtained with these methods are seldom evaluated empirically. Conversely, screening experiments have been criticized for their inability to reveal the biological basis of their results. In this paper, we evaluate the ability of a scalable literature-based approach, discovery-by-analogy, to identify a small number of active agents within a large library screened for activity against prostate cancer cells. The methods used permit retrieval of the knowledge used to infer their predictions, providing a plausible biological basis for predicted activity.

CPT Pharmacometrics Syst. Pharmacol. (2014) 3, e140; doi:10.1038/psp.2014.37; published online 8 October 2014

The escalating cost of drug development and prolonged delay in bringing new drugs to clinical application limit the availability of new therapies for many devastating diseases. An efficient strategy for addressing this problem is the “repurposing” of existing drugs for novel therapeutic applications.¹ There are ~4,000 drugs approved for human use and 5,000 more investigational drugs registered for use, but not approved by regulatory agencies. These drugs represent a rich reservoir of potential therapeutics because much of the pharmacologic and toxicologic information necessary for their clinical use has already been acquired.

A common repurposing approach involves screening libraries of compounds for activity against a target of interest in high-throughput screening experiments.² To mitigate cost, a variety of “in-silico” approaches to identify repurposing opportunities have been developed.³ These include approaches based on the paradigm of literature-based discovery⁴ (LBD), developed by information scientist Don Swanson.⁵ The premise underlying LBD is that exploring the common concepts that a pair of biological entities relate to can identify meaningful implicit connections between these entities—including previously unrecognized therapeutic relationships. So, while the origin of LBD predated the popularization of the term “drug repurposing”, the discovery of new therapeutic applications for existing biological entities remains one of the primary goals of this field.

LBD is by now a mature discipline, with several review papers^{4,6,7} and a volume with contributions by prominent researchers already in print.⁸ A comprehensive review of LBD is beyond the scope of this paper, but important precedents for the current approach include using discrete concepts (such as those in the Unified Medical Language System (UMLS)),⁹ rather than terms as anchor points for discovery,^{10,11} and applying semantic relations extracted from the literature using

Natural Language Processing (NLP) to constrain the search for discoveries.^{12–14} Regardless of whether co-occurrence or some more restrictive relationship is considered, the large number of possible reasoning pathways between concepts presents a computational challenge. Consequently, LBD researchers have evaluated the ability of methods of distributional semantics¹⁵ to facilitate LBD.^{16,17} These methods derive vector representations of terms from the contexts in which they occur across large volumes of electronic text. Terms occurring in similar contexts will have similar representations. So, discovery candidates can be identified by comparing vector representations without exploring intermediate terms explicitly. An encouraging recent trend in LBD research has involved a move toward author-driven empirical validation of predictions made by LBD systems, including evaluating correlation with microarray data,¹⁸ and evaluation of selected predictions in cell-based¹⁹ and animal²⁰ models.

In this paper, we evaluate the ability of a scalable LBD methodology to identify agents active against prostate cancer cells in preclinical experiments. Our approach, *discovery-by-analogy*, is facilitated by a distributional method called Predication-based Semantic Indexing (PSI).^{21–25} For this work, PSI models were derived from SemMedDB,²⁶ a publicly available database of over sixty million concept-relationship-concept triples, or *semantic predications*, extracted from the biomedical literature by SemRep,²⁷ a biomedical NLP system. Additional details regarding PSI are provided in the **Supplementary Materials**.

For interpretation of the results, it is sufficient to know that PSI provides the means to encode the nature of the relationship between a pair of discrete concepts (e.g., *lovastatin* INTERACTS_WITH *nras_gene*) in a manner permitting efficient, albeit approximate inference. This is accomplished by representing concepts and relations in a high-dimensional

¹University of Texas School of Biomedical Informatics at Houston, Houston, Texas, USA; ²Microsoft Bing, Redmond, Washington, USA; ³Center for Translational Cancer Research, Texas A&M Health Sciences Center, Institute of Biosciences and Technology, Houston, Texas, USA; ⁴Department of Investigational Cancer Therapeutics, Division of Cancer Medicine, University of Texas MD Anderson Cancer Center, Houston, Texas, USA; ⁵Department of Genitourinary Medical Oncology, Division of Cancer Medicine, University of Texas MD Anderson Cancer Center, Houston, Texas, USA; ⁶National Library of Medicine, Bethesda, Maryland, USA. Correspondence: T Cohen (trevor.cohen@uth.tmc.edu)

Received 3 March 2014; accepted 20 July 2014; published online 8 October 2014. doi:10.1038/psp.2014.37

vector space, effectively converting the task of exploring multiple possible reasoning pathways into the task of measuring the similarity between vectors. Given an example pair consisting of a disease and an agent that is known to treat it, PSI can infer the reasoning pathways (such as *agent x INTERACTS_WITH y*; *y ASSOCIATED_WITH disease z*) that connect these concepts to one another. These analogical reasoning pathways can then be applied to another disease in order to search for possible treatments.

While our approach is generalizable, we focus our methods on the problem of identifying agents active against prostate cancer, a highly prevalent disease²⁸ with few effective therapies. Though androgen ablation proves effective briefly, relapse and castration-resistant disease inevitably result. Few therapeutic options exist at this stage, and newly approved agents add only a few months to the survival increase provided by long-time agent docetaxel, which remains standard first-line chemotherapy.²⁹ To evaluate our predictions empirically, we conducted a series of high-throughput screening experiments in which three libraries of pharmaceutical agents were evaluated for their activity against PC3 prostate cancer cells,³⁰ which are resistant to commonly used hormonal therapies. We then evaluated the extent to which the ranking of predicted therapeutic relationships between 1,398 agents in this set and the discrete concept “hormone-refractory prostate cancer” (HRPCA; the term HRPCA has been superseded by the term castration-resistant prostate cancer. As the controlled vocabulary used by SemRep contains an entry for HRPCA only, we used this concept as a starting point for the search for active agents) agreed with the average activity of these agents in our screening experiments.

RESULTS

Activity of evaluated agents

Figure 1 shows the distribution of mean growth rates across at least three separate experiments (the number of experiments was larger for agents in multiple libraries) for each of the 1,398 agents that could be mapped to discrete concepts in SemMedDB. Most of the agents were inactive. Cells treated by a small number of agents, 68 in total, have a mean growth rate of less than or equal to 1.5 standard deviations below the average across all agents (growth rate ≤ 54.57). These agents were considered as active agents, and our models were evaluated for their ability to recover them from the list of 1,398 possibilities.

Discovered discovery patterns

PSI was used to infer reasoning pathways, or “discovery patterns”,¹² from therapeutic (TREATS) relationships in SemMedDB concerning cancers other than prostatic cancer. In one experiment, we derived a PSI space from all of the knowledge in SemMedDB subject to constraints described in the Methods section (all knowledge (AK)). In another, in order to simulate discovery of relationships that have not been identified previously, we withheld from the model knowledge of therapeutic (“TREATS”) relationships, and other direct relationships between a pharmaceutical agent and a type of cancer (This constraint was applied to all cancers, to ensure that the system did not infer reasoning pathways from other

cancers that would not apply to the restricted knowledge available for HRPCA.) (knowledge withheld (KW)). The five most frequently inferred two- and three-predicate reasoning pathways in each space were retained, and are shown in **Figure 2** with example bridging terms that populate instantiations of these pathways. These reasoning pathways illustrate the “thinking” of the system, and are often intuitively interpretable. For instance, path 2 for KW (**Figure 2**, right) suggests that those agents that INHIBIT a biological entity ASSOCIATED WITH a disease may treat it. For each PSI space, we evaluated performance with two-predicate pathways alone, and with two- and three-predicate pathways together.

Predicting activity against PC3 cells

Figure 3 shows the results of experiments in which these four configurations of discovery-by-analogy were used to rank the 1,398 candidate agents with respect to the strength of their therapeutic relationship to HRPCA. Both models effectively predict many of the active agents, with approximately a third of the active agents (sensitivity ~ 0.33) occurring in the top 10 percent of predictions (1-specificity ~ 0.1) when both two and three-predicate pathways are considered. Incorporating longer pathways improves performance, a finding consistent with simulated LBD experiments.²⁵ Differences in performance are summarized in **Table 1**. The best performing model returns around one in five active agents in the top 100 predictions, and discovers half of the

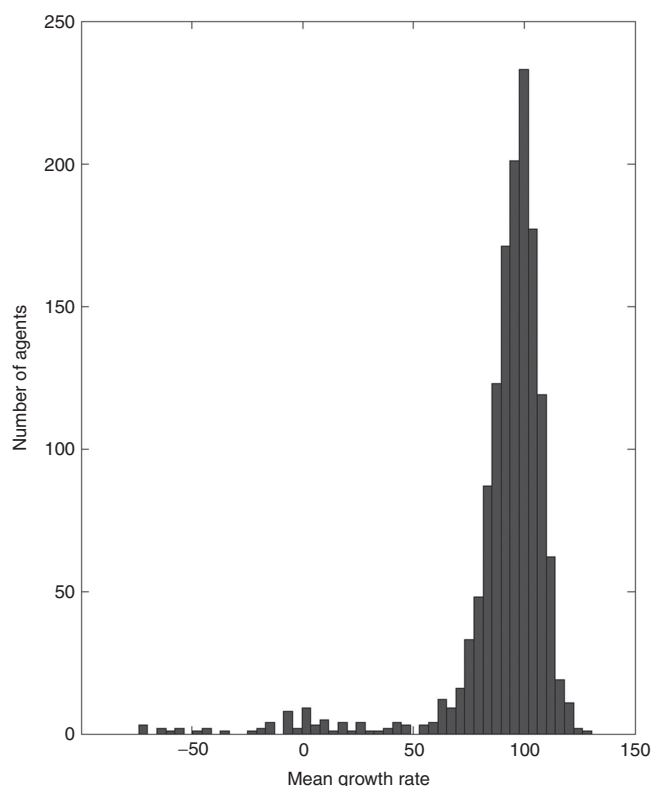


Figure 1 Histogram of growth rates of PC3 cells in response to 1,398 pharmaceutical agents. Positive values indicate the percentage of growth inhibition relative to negative controls. Negative values indicate the percentage of the originally seeded cells that remained after treatment.

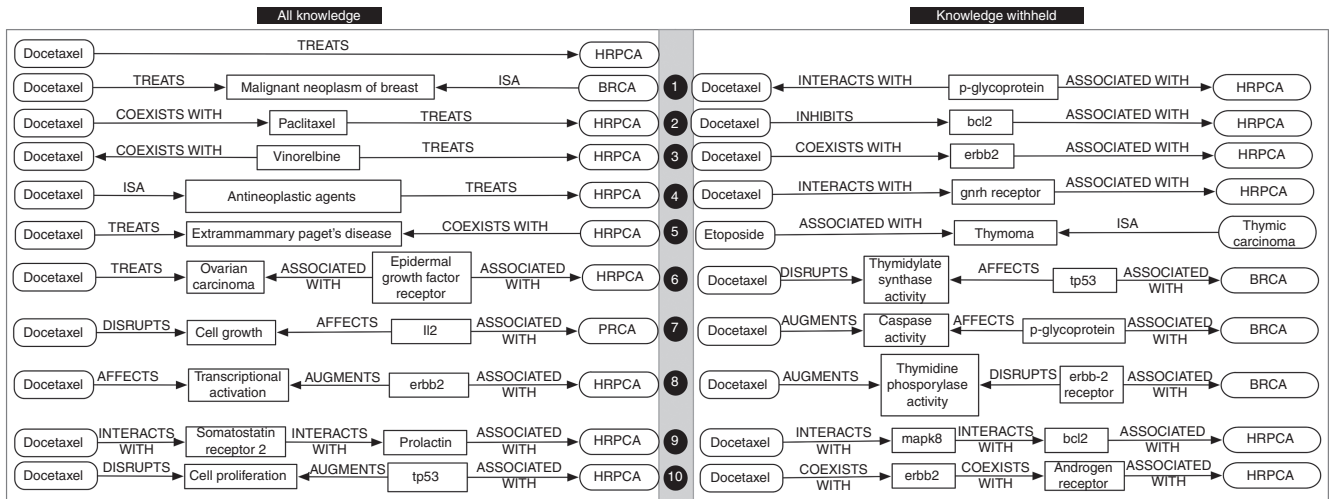


Figure 2 Inferred reasoning pathways with example bridging terms. BRCA, breast carcinoma; HRPCA, hormone-refractory prostate cancer; PRCA, prostate carcinoma.

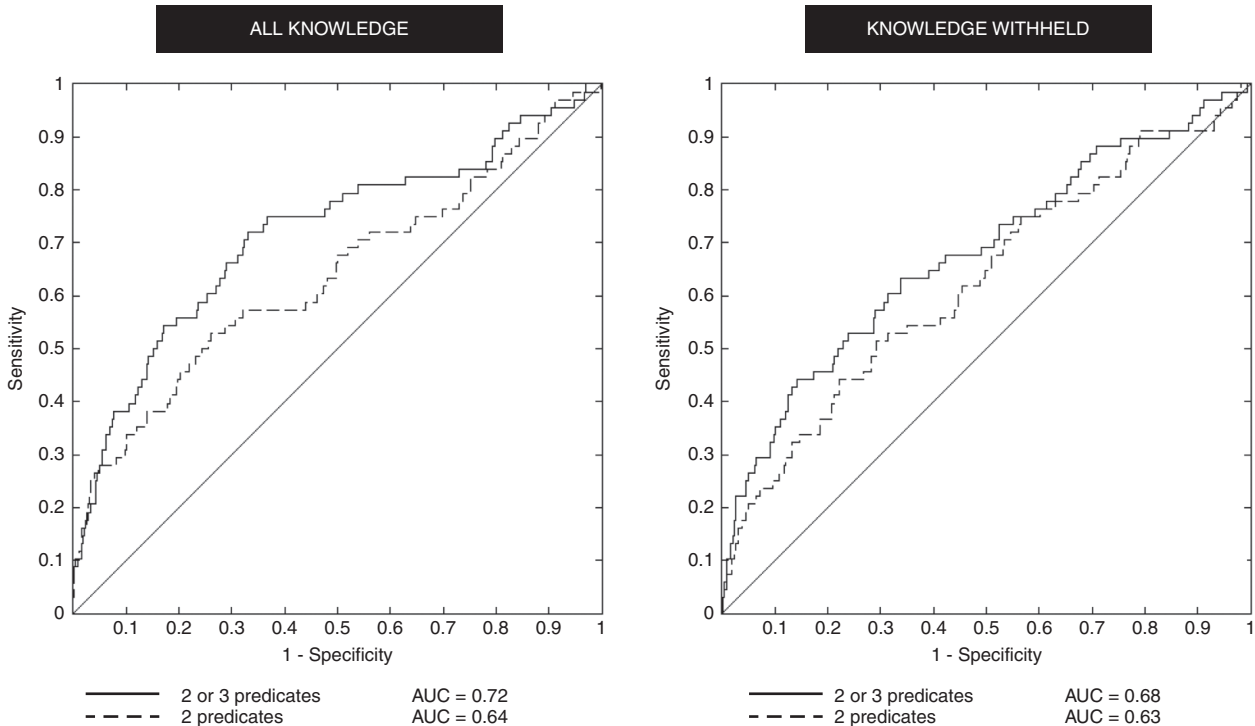


Figure 3 Results with all knowledge (left) and knowledge withheld (right). AUC, area under the curve; 2 predicates, dual-predicate pathways only; 2 or 3 predicates, dual-or triple-predicate pathways.

active agents within the first 240 predictions (around 17% of the agents evaluated).

Table 2 shows the ranks assigned to each agent by each model, and whether these agents occurred in “TREATS” relationships with HRPCA in SemMedDB. Identification of agents that did not occur in TREATS relationships, or when TREATS relationships were withheld from the model, suggests the possibility of inferring therapeutic approaches that have yet to be described in the literature. In certain cases withholding knowledge led to a higher ranking, suggesting the indirect

reasoning pathways these constraints emphasize may be advantageous at times.

Discovery-by-analogy appears better able to predict the activity of agents when more relevant knowledge is available to it. Figure 4 compares the rank of each active agent by the best-performing PSI model to the number of predications in which this agent appears as the subject. In the two quadrants on the left (less relevant knowledge), there are a number of agents in both the top (lower ranked) and bottom (higher ranked) quadrants. In the rightmost quadrants

Table 1 Performance metrics

	All knowledge (<i>n</i> = 1,398)		Knowledge withheld (<i>n</i> = 1,387 ^a)		Random (<i>n</i> = 1,398)
Predicates	2 or 3	2	2 or 3	2	$\bar{x}@n = 1,000$
AUC	0.7172	0.6380	0.6777	0.6260	0.4994
pAUC(0.1)	0.0260	0.0233	0.0244	0.0171	0.0050
pAUC(0.01)	9.1774e⁻⁰⁴	8.2928e ⁻⁰⁴	4.4597e ⁻⁰⁴	4.5712e ⁻⁰⁴	0.53527e ⁻⁰⁴
P@k=100	0.21	0.19	0.18	0.15	0.0480
Median rank T+	240	367.5	331	419.5	698.98

Best performance is shown in boldface.

AUC, area under curve; pAUC(*x*), partial AUC with FPR=*x*; P@k=*y*, precision at k = *y*; Median rank T+, median rank of the 68 active agents; Random, random permutation; $\bar{x}@n = 1,000$, mean across 1,000 iterations.

^aWithholding knowledge resulted in the elimination of 11 inactive agents for which no knowledge was available.

(more relevant knowledge), there is a preponderance of agents in the bottom right quadrant (higher ranked).

Retrieving the underlying assertions

Of the highly ranked active agents, the statins are of particular interest from a repurposing perspective, as it has recently been found that they reduce risk of cancer-related mortality,³¹ and a number of potential mechanisms related to their antiproliferative, anti-invasive, and proapoptotic effects have been proposed.³² As shown for lovastatin in **Table 3**, predications related to many of these mechanisms conform to the constraints of the KW discovery patterns, and would therefore have contributed toward the strength of this prediction. This table includes only a small number of the 7,883 unique reasoning pathways that conform to these constraints in this case. Assertions underlying these pathways derive from 164,649 sentences from the literature. So, while retrieving the middle terms that lie along these reasoning pathways is a simpler computational task than searching for an endpoint such as lovastatin, the presentation of such pathways and the excerpts that support them in a manner conducive to human review presents further challenges we are attempting to address in our ongoing work.

Comparison with a co-occurrence-based model

We also evaluated the performance of a co-occurrence-based distributional model, reflective random indexing³³ (RRI), on a subset of 1,295 agents represented by both PSI and RRI. On this subset, RRI achieved an AUC of 0.6868 and partial AUC (pAUC) with false-positive rate of 0.1 (pAUC(0.1)) and 0.01 (pAUC(0.01)) of 0.0255 and 3.4032e⁻⁰⁴, respectively. Discovery-by-analogy results were: AUC = 0.7017, pAUC(0.1) = 0.0255 and pAUC(0.01) = 8.5079e⁻⁰⁴ with AK, and AUC = 0.6880, pAUC(0.1) = 0.0219 and pAUC(0.01) = 5.2263e⁻⁰⁴ with KW (both with two- and three-predicate paths). So discovery-by-analogy generally outperformed RRI, even when knowledge was withheld from it. Restricting knowledge from RRI (see Methods) reduces its performance (AUC = 0.6693, pAUC(0.1) = 0.0249, pAUC(0.01) = 1.8839e⁻⁰⁴). It is still able to recover many active agents, supporting its utility as a means to derive therapeutically meaningful implicit associations between concepts that do not co-occur directly. However, RRI lacks the explanatory capability provided by the predications that mediate discovery-by-analogy.

Evaluation in simulation

As the KW space was constructed without encoding TREATS relationships, we evaluated the ability of discovery-by-analogy in this space to recover drugs in our evaluation set that occurred in TREATS relationships with 1,272 UMLS concepts of the semantic type “neop”. Across these 1,272 concepts, the median AUC was 0.8890, which supports the general applicability of our methods and suggests that recovering known relationships is an easier task. To evaluate the AK space, we used a reference standard constructed by Gottlieb *et al.*³⁴ We were able to identify 359 disease-related concepts and 554 drug-related concepts from this reference set within the AK space. Across these disease-related concepts, the median AUC was 0.8388, with better performance for concepts that are well represented in SemMedDB. As our methods necessitated using a subset of concepts and a different estimation procedure, our estimated global AUC of 0.7408 is not strictly comparable to Gottlieb *et al.*'s performance estimates. Nonetheless, it is lower than their reported AUC of ~0.9 across experiments. As their approach involves a classifier trained on features derived from an ensemble of drug–drug and disease–disease similarity metrics derived from multiple structured knowledge repositories and empirical expression data, it seems reasonable that it would perform well over a broader range of diseases than an approach based on a single similarity metric and knowledge source. As noted by the authors, their approach can be readily extended to incorporate additional similarity metrics, such as those described in this paper. Conversely, expansion-by-analogy could be configured to utilize additional structured knowledge presented in predicate form. For further details, see **Supplementary Material**.

Error analysis

To evaluate false-positive findings, we retrieved the top 25 predictions for the PSI models. These are provided in **Table S4** of the **Supplementary Material**, to complement this overview of the key findings. A number of the highly ranked predictions were true positives (i.e., active against PC3 cells—8/25 (AK) and 5/25 (KW)). In addition, many of the agents occurred in TREATS predications with hormone-refractory prostate cancer in SemMedDB (24/25 (AK) and 15/25 (KW)). This usually indicates that their activity against either this disease or a model of it has been asserted in the literature, though it may at times indicate a NLP error. A keyword search of clinicaltrials.gov for each drug name AND

Table 2 Ranking of active agents

AGENT	ALL23	ALL2	KW23	KW2
Docetaxel ^a	1	1	18	3
Vinorelbine ^a	2	2	28	38
Paclitaxel ^a	3	4	2	4
Gemcitabine ^a	4	5	3	8
Etoposide ^a	7	6	40	9
Topotecan hydrochloride	8	8	16	32
Bortezomib ^a	11	18	31	52
Dactinomycin	28	33	14	100
Camptothecin ^a	31	32	37	71
Parthenolide	35	155	183	1,076
Cytarabine	38	23	77	143
Triptolide	44	212	142	300
Colchicine ^a	48	83	74	174
Doxorubicin hydrochloride	59	51	416	790
Podophyllotoxin	70	61	216	307
Cyclosporine	72	157	9	15
Homoharringtonine	73	705	312	449
Cladribine	76	30	258	406
Lbh589	85	341	191	700
Simvastatin	91	765	42	161
Trichostatin_a	93	152	82	216
Atorvastatin	104	271	45	195
Lovastatin	105	381	50	194
Letrozole	116	445	338	752
Nocodazole	124	300	152	619
Vincristine_sulfate	127	47	169	179
Gestrinone	166	359	728	984
Thimerosal	184	129	414	301
Staurosporine	191	623	49	61
2-chloro-2-deoxyadenosine	203	264	102	110
Laq824	216	420	604	1,070
Disulfiram	217	54	422	405
Hexachlorophene	222	185	453	323
Artesunate	234	1,140	309	998
Epirubicin hydrochloride	246	70	1,269	1,309
Cycloheximide	260	1,164	17	49
Pyrimidinones	265	739	1,378	1,364
17-(dimethylaminoethylamino)-17-demethoxygeldanamycin	297	671	1,257	1,292
Protein c inhibitor	350	1,134	155	42
Mitoxantrone hydrochloride	354	49	777	1,106
Auranofin	378	1,251	193	500
Topotecan ^a	400	58	1,227	1,099
Piretanide	413	211	993	640
Irinotecan hydrochloride	424	324	926	1,051
Pxd101	429	292	560	321
Mebendazole	460	289	323	885
Parbendazole	474	376	695	630
Niclosamide	477	1,055	946	1,066
Lestaurtinib	487	981	864	1,351
Idarubicin hydrochloride	527	654	1,179	1,293
Digoxin	539	794	202	267
Monensin sodium	686	467	444	420
Naltrindole	698	912	489	269

Table 2 Continued

Albendazole	733	1,034	587	944
Haloprogin	771	683	917	847
Thiostrepton	893	1,233	143	79
Digitoxigenin	1,028	1,057	953	419
Benzethonium chloride	1,098	1,023	1,314	1,339
Chrysene	1,113	1,280	887	581
Fenbendazole	1,115	1,183	490	798
Isolanid	1,123	900	834	717
Daunorubicin hydrochloride	1,142	708	105	29
Vindesine sulfate	1,159	1,234	1,057	768
Compactin	1,191	1,325	352	383
Digoxigenin	1,268	1,397	739	719
Quinacrine dihydrochloride	1,326	711	743	631
Proscillaridin	1,352	1,099	1,238	747
Pyrvinium pamoate	1,358	1,269	976	688

Green/lighter shading, higher rank; red/darker shading, lower rank.
 ALL, all knowledge; KW, knowledge withheld; 2, 2 predicate paths only; 23, 2 and 3 predicate paths.
^aOccurrence in "TREATS" relationship with HRPDA in database.

prostate cancer was performed. Trials were found for all but two (AK) and four (KW) predictions, with a median of 10 (AK) and 9 (KW) trials per agent. So, other evidence suggesting the plausibility of inactive agents as treatments for prostate cancer (though not necessarily HRPDA) was usually found.

The evaluation set includes three agents that are US Food and Drug Administration approved for treatment of prostate cancer. These are docetaxel, which was active against PC3 cells, and bicalutamide and prednisone, which were not. Both models recovered these agents, with ranks of 1,15,17 and 18,21,64 for AK and KW, respectively. Despite this Food and Drug Administration approval, the anti-androgen bicalutamide is not effective for HRPDA. As SemRep does not currently recognize mutant forms of this receptor that confer resistance to bicalutamide, eliminating this false positive would require increasing the granularity of concept extraction. We note also that the fifth KW reasoning pathway (Figure 2, right) was derived on account of failure to exclude UMLS concepts of type carbohydrate (carb) from the inference process. This pathway was of no value for prediction, as the KW space excluded direct relationships between pharmaceutical substances and HRPDA.

DISCUSSION

A substantial proportion of the active agents were highly ranked by our LBD models. For example, empirical screening of only the top 100 agents (around 7% of the set) suggested by the best-performing PSI model identified around 31% of active agents. This suggests that LBD methods provide a principled and cost-effective approach to selection of agents for screening. In addition, these methods can provide support for the biological plausibility of those agents that are active in screening experiments. Our methods can also be applied to find new therapeutic targets for an agent of interest, which has been identified as an important application for *in silico* methods.³⁵ Highly ranked predictions involve thousands of unique reasoning pathways drawn from a range of literature that is unlikely to fall within the

scope of reading of an individual researcher or research group. This provides access to knowledge that may otherwise be ignored, and broadens the scope of inquiry beyond the focus on individual mechanisms that has been identified as a limitation of contemporary drug discovery efforts.³⁶

Limitations

SemMedDB is not perfectly accurate. In a recent evaluation of SemRep, Kilicoglu *et al.*³⁷ report 0.75 precision and 0.64 recall (.69 *f*-score). However, SemMedDB contains a large number of predications extracted from MEDLINE with different frequencies. With PSI, the relative frequency of a predication affects its impact upon reasoning. So, perfectly accurate knowledge is not required. PSI is robust to infrequent errors, but vulnerable to systematic inaccuracies, though with concepts that occur infrequently rare errors may have greater consequences. As only the most popular inferred reasoning pathways across all other cancers were retained, reasoning pathways are unlikely to be affected by focal NLP errors. However, failure to constrain inference appropriately may lead to the generation of unproductive pathways. Regarding empirical evaluation, each agent was evaluated in at least three experiments, with a median standard deviation of the average growth rate of 10.76 across

these sets of three experiments. While this is unlikely to have affected the vast majority of inactive agents, further empirical evaluation of the small number of agents with growth rates in the vicinity of the threshold of 54.57 would increase confidence that they were correctly classified. Finally, it is likely that increasing the dimensionality of our PSI spaces would reduce information loss, improving results.

METHODS

PSI

Semantic vector representations in PSI are generated by superposing (adding) *elemental vector* representations of concepts, which are generated stochastically with a high probability of being dissimilar from one another. They serve as signatures for the concepts concerned, and remain distinguishable from one another despite distortion that occurs during training. Given the predication “esr1_protein ASSOCIATED WITH breast_carcinoma (BRCA)”, we wish to encode the elemental vector for “esr1_protein” ($E(\text{esr1_protein})$) into the semantic vector for BRCA ($S(\text{BRCA})$), and vice versa. We also wish to encode the nature of the relationships between these concepts. This is accomplished using a *binding operator* (\otimes),^{38,39} which combines vectors to generate a bound product that is dissimilar from its component vectors. If two vectors are bound, it is possible to retrieve one of these vectors by reversing binding (\oslash) using the other. For example, “esr1_protein ASSOCIATED WITH BRCA” is encoded into $S(\text{BRCA})$ as follows (-INV indicates directionality):

$$S(\text{BRCA}) += E(\text{ASSOCIATED_WITH-INV}) \otimes E(\text{esr1_protein}).$$

On account of the reversible nature of the binding operator, we would anticipate:

$$S(\text{BRCA}) \oslash E(\text{ASSOCIATED_WITH-INV}) \approx E(\text{esr1_protein}).$$

Indeed, in the KW space, $E(\text{esr1_protein})$ is the fourth nearest neighboring elemental vector to $S(\text{BRCA}) \oslash E(\text{ASSOCIATED_WITH-INV})$. This recovery will be approximate when the semantic vector concerned encodes other predications, but the dissimilarity between elemental vectors makes it highly unlikely that the resulting vector will be closer to a random vector than it is to $E(\text{esr1_protein})$. Thus, the binding operator facilitates querying a PSI space for concepts that relate to other concepts in particular ways. Queries can

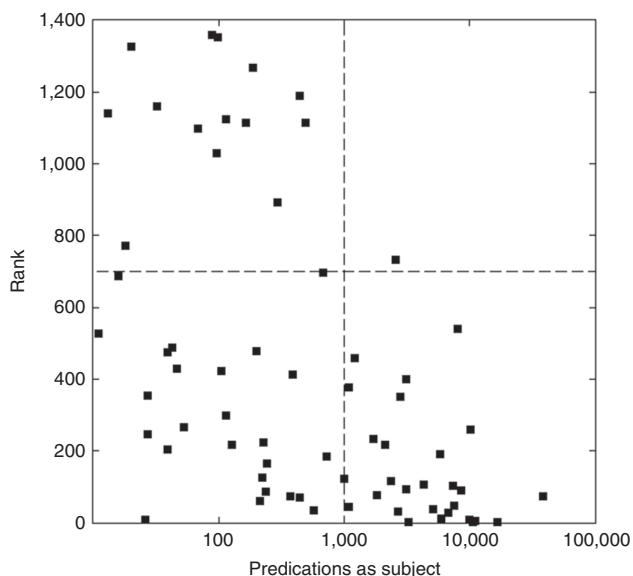


Figure 4 Rank vs. predications with agent as subject.

Table 3 Examples of evidence underlying lovastatin prediction

Mechanism	Relevant middle term (number of unique predications pathways)	Example predication pathway
Blocking G1-S transition in cell cycle	cell_cycle_proteins (24)	lovastatin INTERACTS_WITH cell_cycle_proteins INTERACTS_WITH cdc2_protein_kinase ASSOCIATED_WITH hormone-refractory_prostate_cancer
Ras inhibition	nras_gene (16); rho (24)	lovastatin INTERACTS_WITH rho INTERACTS_WITH nras_gene ASSOCIATED_WITH hormone-refractory_prostate_cancer
Rhoa inhibition	rhoa_gene (22); egf(106)	lovastatin INTERACTS_WITH rho INTERACTS_WITH egf_gene ASSOCIATED_WITH hormone-refractory_prostate_cancer
Hmg-coa reductase inhibition	3-hydroxy-3-methylglutaryl-coenzyme_a (2); hydroxymethylglutaryl-coa-reductase_inhibitors (44)	lovastatin INTERACTS_WITH 3-hydroxy-3-methylglutaryl-coenzyme_a INTERACTS_WITH p-glycoprotein ASSOCIATED_WITH hormone-refractory_prostate_cancer

extend across two predicates, as an extended path between semantic vectors can be specified as:

$$\begin{aligned} & (S(\text{BRCA}) \otimes E(\text{ASSOCIATED_WITH-INV})) \otimes \\ & E(\text{INTERACTS_WITH}) \\ & \approx E(\text{esr1_protein}) \otimes E(\text{INTERACTS_WITH}) \\ & \approx S(\text{tamoxifen}) \text{ (as "tamoxifen INTERACTS_WITH} \\ & \text{esr1_protein")} \end{aligned}$$

It is possible to infer dual predicate paths through which two concepts are connected, as follows:

$$\begin{aligned} & S(\text{tamoxifen}) \otimes S(\text{BRCA}) \\ & \approx (E(\text{esr1_protein}) \otimes E(\text{ASSOCIATED_WITH-INV})) \otimes (E \\ & (\text{INTERACTS_WITH}) \otimes E(\text{esr1_protein})) \\ & \approx E(\text{esr1_protein}) \otimes E(\text{esr1_protein}) \otimes \\ & E(\text{ASSOCIATED_WITH-INV}) \otimes E(\text{INTERACTS_WITH}) \\ & \approx E(\text{ASSOCIATED_WITH-INV}) \otimes E(\text{INTERACTS_WITH}) \end{aligned}$$

Thus, dual-predicate paths connecting drugs to diseases they treat can be inferred from their PSI vector representations. Once retrieved, these paths can then be used to evaluate potential therapies for a new disease by analogy. For example, the nearest semantic vector for an agent in our evaluation set to $S(\text{HRPCA}) \otimes E(\text{ASSOCIATED_WITH-INV}) \otimes E(\text{INTERACTS_WITH})$ in the KW space is $S(\text{bicalutamide})$. We extend this process in two ways that have improved results in prior simulations.^{25,40} To model three-predicate pathways, we generate *second order semantic vectors* for cancer types ($S_2(\text{cancer type})$) by superposing semantic vector representations of concepts they are ASSOCIATED_WITH, and use these vectors as a secondary starting point for a search and inference.²⁵ To search across multiple predicate pathways, we employ an adaptation of the quantum disjunction operator defined by Birkhoff and von Neumann⁴¹ and applied to information retrieval by Widdows and Peters.⁴² This operator serves as a vector space equivalent of the boolean "OR" operator, allowing us to combine multiple reasoning pathways into a single search expression. Consequently, pharmaceutical agents in the set can be ranked with respect to the strength of their relatedness across a set of dual- and/or triple-predicate reasoning pathways. See **Supplementary Material** for further details.

Discovery-by-analogy

For discovery-by-analogy, we utilized SemMedDB,²⁶ a publicly available repository of semantic predications extracted from the biomedical literature by the SemRep NLP system.²⁷ We used the June 2013 edition, containing 65,465,536 predications extracted from 13,537,476 MEDLINE citations. From this, we created a 32,000-dimensional binary-valued PSI space, using the open source Semantic Vectors package⁴³⁻⁴⁶ currently maintained and developed by authors DW and TC. Semantic vectors were generated for each concept occurring in 500,000 or fewer predications, and all predications involving these concepts and a set of

predicates of interest, {AFFECTS, ASSOCIATED_WITH, AUGMENTS, CAUSES, COEXISTS_WITH, DISRUPTS, INHIBITS, INTERACTS_WITH, ISA, PREDISPOSES, PREVENTS, SAME_AS, STIMULATES, TREATS}, were encoded during training.

In our experiments, we applied the inference process described previously to all TREATS relationships in SemMedDB involving represented neoplastic processes (UMLS semantic type *neop*) unrelated to prostate cancer, and retrieved the most strongly associated reasoning path in each case. Counting the number of times each dual-predicate path was retrieved, this way revealed the five most popular dual-predicate paths for each space (illustrated in **Figure 2**). To extend the range of search beyond two predicates, we substituted second-order semantic vectors for the original semantic vectors, and repeated the inference process to find the five most popular triple predicate paths ending with ASSOCIATED_WITH (also illustrated in **Figure 2**). To combine the dual- and triple-predicate pathways, we constructed a search subspace using the quantum disjunction operator, and measured the distance between this subspace and each of the agents to generate a ranked list.

Reflective random indexing (RRI)

Our RRI model was derived from the 2012 MetaMapped MEDLINE baseline (available at: <http://skr.nlm.nih.gov/resource/MetaMappedBaselineInfo.shtml>), which includes UMLS⁹ concepts extracted by the MetaMap package^{47,48} from 20,494,848 MEDLINE citations. This model was also constructed using the Semantic Vectors package.⁴³⁻⁴⁶ Document vectors were generated as weighted superpositions of 32,000-dimensional binary-valued elemental term vectors representing terms they contain. Terms with more than three non-alphabet characters, and terms occurring in more than 500,000 documents were excluded. The log-entropy weighting metric⁴⁹ was employed, to emphasize the effect of terms occurring focally in the corpus, and temper the effects of repeated mentions of a term within a document. Semantic vectors for all UMLS concepts occurring in 500,000 or fewer documents were generated by superposing the vectors representing documents they occur in. Concept vectors for pharmaceutical agents were compared to the concept vector for HRPCA to generate a ranked list of predictions (see **Supplementary Material**).

Evaluation set

The evaluation set was derived from results obtained by screening three libraries of agents (Custom Clinical (2010), Prestwick (2013) and NIH (2008)) for their activity against PC3 cells. Cells were seeded at ~250 cells, 50 μ l per well on day 0, and incubated for 24 h. Compounds were transferred at 50 nl vol from 1 mmol/l Stock using a Pin tool (V&P). Cells were then incubated for 72 h, and fixed with a DAPI stain, imaged at 4x on an INCELL6000 analyzer and segmented using GE Developer to obtain cell counts after 72 h of incubation time. The growth rates of cells were

compared to those of negative controls using the following formulas:

Formula A: For $T_i \geq T_z$

- $\{(T_i - T_z)/(C - T_z)\} \times 100 =$ % inhibition of growth compared to control growth

Formula B: For $T_i < T_z$

- $\{(T_i - T_z)/T_z\} \times 100 =$ -% of cells killed

where

- T_z = day 0 = cells seeded
- T_i = day growth = cells w/drugs-day of assay
- C = Control growth = cells with no drug-day of assay

To map the 1,622 unique evaluated agents to concepts represented by our models, we used local installations of the MetaMap^{47,48} and SemRep²⁷ systems to identify the preferred form of each agent used in the MetaMap baseline and SemMedDB respectively. These lists of concepts were then restricted to represented in the PSI space (without knowledge withheld), leaving 1,398 agents.

Simulating discovery

In some experiments, we withheld information to simulate a discovery scenario in which no known connection between the agents tested and HRPDA exists in the literature. For PSI (KW), we eliminated from training all TREATS relationships, and any predication containing a concept with UMLS semantic type neoplastic process (neop), and a concept with semantic type pharmaceutical substance (phsu), antibiotic (antb), or organic chemical (orch). This relatively severe constraint was imposed to ensure reasoning pathways inferred from other cancers would still apply to HRPDA. Consequently, reasoning related to all cancer-related concepts and therapies was handicapped. For RRI, we imposed a milder constraint, eliminating from training any document containing both HRPDA and a drug in the test set.

Acknowledgments. This research was supported in part by US National Library of Medicine grants (R21 LM010826) and (R01 LM011563); Cancer Prevention & Research Institute of Texas grants (R1307) and (RP110532-AC); and the Intramural Research Program of the US National Institutes of Health, National Library of Medicine.

Author Contributions. T.C., P.D., and J.K. wrote the manuscript. T.C., C.S., R.Z., and D.W. designed the research. T.C. and C.S. performed the research. T.R., D.W., and T.C. contributed new reagents/analytical tools. All authors read and approved the final manuscript.

Conflict of Interest. The authors declared no conflict of interest.

1. Ashburn, T.T. & Thor, K.B. Drug repositioning: identifying and developing new uses for existing drugs. *Nat. Rev. Drug Discov.* **3**, 673–683 (2004).
2. Persidis, A. High-throughput screening. *Nat. Biotechnol.* **16**, 488 (1998).
3. Ekins, S., Williams, A.J., Krasowski, M.D. & Freundlich, J.S. In silico repositioning of approved drugs for rare and neglected diseases. *Drug Discov. Today* **16**, 298–310 (2011).

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

- ✓ Methods of literature-based discovery (LBD) have been proposed as a means to facilitate drug repurposing, but evaluation of these methods has been limited to reproduction of historical discoveries, prediction of future co-occurrence from time delimited literature subsets, and qualitative evaluation of the novelty and plausibility of a small number of predictions.

WHAT QUESTION DID THIS STUDY ADDRESS?

- ✓ In this study we evaluate the extent to which scalable LBD methods can predict the activity of libraries of pharmaceutical agents against prostate cancer cell lines in *in vitro* experiments.

WHAT THIS STUDY ADDS TO OUR KNOWLEDGE

- ✓ We demonstrate that many of the small number of active agents in these experiments are highly ranked as potential therapies by LBD models. Models that encode the nature of the relationships between concepts appear better able to predict activity.

HOW THIS MIGHT CHANGE CLINICAL PHARMACOLOGY AND THERAPEUTICS

- ✓ LBD methods can be used to select agents of interest for empirical validation, to constrain development costs. The best performing of these methods can also provide plausible biological arguments to support the observed activity of agents in *in vitro* experiments.

4. Andronis, C., Sharma, A., Virvilis, V., Deftereos, S. & Persidis, A. Literature mining, ontologies and information visualization for drug repurposing. *Brief. Bioinform.* **12**, 357–368 (2011).
5. Swanson, D.R. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspect. Biol. Med.* **30**, 7–18 (1986).
6. Hristovski, D., Rindflesch, T. & Peterlin, B. Using literature-based discovery to identify novel therapeutic approaches. *Cardiovasc. Hematol. Agents Med. Chem.* **11**, 14–24 (2013).
7. Deftereos, S.N., Andronis, C., Friedla, E.J., Persidis, A. & Persidis, A. Drug repurposing and adverse event prediction using high-throughput literature analysis. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **3**, 323–334 (2011).
8. Bruza, P. & Weeber, M. Literature-based discovery. (Springer-Verlag New York Inc, New York, 2008).
9. Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* **32**, D267–D270 (2004).
10. Weeber, M., Klein, H., de Jong-van den Berg, L.T.W. & Vos, R. Using concepts in literature-based discovery: simulating Swanson's Raynaud-fish oil and migraine-magnesium discoveries. *J. Am. Soc. Inf. Sci. Technol.* **52**, 548–557 (2001).
11. Srinivasan, P. Text mining: generating hypotheses from MEDLINE. *J. Am. Soc. Inf. Sci. Technol.* **55**, 396–413 (2004).
12. Hristovski, D., Friedman, C., Rindflesch, T.C. & Peterlin, B. Exploiting semantic relations for literature-based discovery. *AMIA Annu. Symp. Proc.* 2006, 349–353 (2006).
13. Ahlers, C.B., Hristovski, D., Kilicoglu, H. & Rindflesch, T.C. Using the literature-based discovery paradigm to investigate drug mechanisms. *AMIA Annu. Symp. Proc.* 2007, 6–10 (2007).
14. Hristovski, D., Friedman, C., Rindflesch, T. & Peterlin, B. Literature-based knowledge discovery using natural language processing. In *Literature Based Discovery*. (eds. Bruza, P. and Weeber, M) 133–152 (Springer-Verlag, Berlin Heidelberg, 2008).

15. Cohen, T. & Widdows, D. Empirical distributional semantics: methods and biomedical applications. *J. Biomed. Inform.* **42**, 390–405 (2009).
16. Cole, R.J. & Bruza, P.D. A bare bones approach to literature-based discovery: an analysis of the Raynaud's/fish-oil and migraine-magnesium discoveries in Semantic space. In *Discovery Science, 8th International Conference, DS 2005, Singapore, October 8–11* (eds. Hoffman, A., Motoda, H and Scheffer, T) 84–98 (Springer-Verlag, Berlin Heidelberg, 2005).
17. Gordon, M.D. and Dumais, S. Using latent semantic indexing for literature based discovery. *J. Am. Soc. Inf. Sci.* **49**, 674–685 (1998).
18. Hu, Y. et al. Analysis of genomic and proteomic data using advanced literature mining. *J. Proteome Res.* **2**, 405–412 (2003).
19. Frijters, R., van Vugt, M., Smeets, R., van Schaik, R., de Vlieg, J. & Alkema, W. Literature mining for the discovery of hidden connections between drugs, genes and diseases. *PLoS Comput. Biol.* **6**, e1000943 (2010).
20. Lekka, E., Deftereos, S.N., Persidis, A., Persidis, A. & Andronis, C. Literature analysis for systematic drug repurposing: a case study from Biovista. *Drug Discov. Today Ther. Strat.* **8**, 103–108 (2011).
21. Cohen, T., Schvaneveldt, R.W. & Rindflesch, T.C. Predication-based semantic indexing: permutations as a means to encode predications in semantic space. *AMIA Annu. Symp. Proc.* **2009**, 114–118 (2009).
22. Cohen, T., Widdows, D., Schvaneveldt, R.W. & Rindflesch, T.C. Logical leaps and quantum connectives: forging paths through predication space. AAAI-Fall 2010 Symposium on Quantum Informatics for Cognitive, Social, and Semantic Processes, November 2010. pp 11–13.
23. Cohen, T., Widdows, D., Schvaneveldt, R. & Rindflesch, T. Finding Schizophrenia's Prozac: emergent relational similarity in predication space. In *QI'11 Proceedings of the 5th International Symposium on Quantum Interactions Aberdeen, Scotland*. (eds. Song, D. et al) (Springer-Verlag, Berlin Heidelberg, 2011).
24. Cohen, T., Widdows, D., Schvaneveldt, R., Davies, P. & Rindflesch, T. Discovering discovery patterns with predication-based Semantic indexing. *J. Biomed. Inf.* **45**, 1049–1065 (2012).
25. Cohen, T., Widdows, D., Schvaneveldt, R. & Rindflesch, T. Discovery at a distance: farther journey's in predication space. Proceedings of the First International Workshop on the Role of Semantic Web in Literature-Based Discovery (SWLBD2012), Philadelphia, PA, 2012.
26. Kilicoglu, H., Shin, D., Fiszman, M., Roseblat, G. & Rindflesch, T.C. SemMedDB: a PubMed-scale repository of biomedical semantic predications. *Bioinformatics* **28**, 3158–3160 (2012).
27. Rindflesch, T.C. & Fiszman, M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *J. Biomed. Inform.* **36**, 462–477 (2003).
28. Siegel, R., Naishadham, D. & Jemal, A. Cancer statistics, 2013. *CA. Cancer J. Clin.* **63**, 11–30 (2013).
29. Tannock, I.F. et al.; TAX 327 Investigators. Docetaxel plus prednisone or mitoxantrone plus prednisone for advanced prostate cancer. *N. Engl. J. Med.* **351**, 1502–1512 (2004).
30. Kaighn, M.E., Narayan, K.S., Ohnuki, Y., Lechner, J.F. & Jones, L.W. Establishment and characterization of a human prostatic carcinoma cell line (PC-3). *Invest. Urol.* **17**, 16–23 (1979).
31. Nielsen, S.F., Nordestgaard, B.G. & Bojesen, S.E. Statin use and reduced cancer-related mortality. *N. Engl. J. Med.* **367**, 1792–1802 (2012).
32. Chan, K.K., Oza, A.M. & Siu, L.L. The statins as anticancer agents. *Clin. Cancer Res.* **9**, 10–19 (2003).
33. Cohen, T., Schvaneveldt, R. & Widdows, D. Reflective Random Indexing and indirect inference: a scalable method for discovery of implicit connections. *J. Biomed. Inform.* **43**, 240–256 (2010).
34. Gottlieb, A., Stein, G.Y., Ruppig, E. & Sharan, R. PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Mol. Syst. Biol.* **7**, 496 (2011).
35. Chiang, A.P. & Butte, A.J. Systematic evaluation of drug-disease relationships to identify leads for novel drug uses. *Clin. Pharmacol. Ther.* **86**, 507–510 (2009).
36. Chen, B. & Butte, A.J. Network medicine in disease analysis and therapeutics. *Clin. Pharmacol. Ther.* **94**, 627–629 (2013).
37. Kilicoglu, H., Fiszman, M., Roseblat, G., Marimpietri, S. & Rindflesch, T.C. Arguments of nominals in semantic interpretation of biomedical text. Proceedings of the 2010 Workshop on Biomedical Natural Language Processing, 2010. pp 46–54.
38. Kanerva, P. Hyperdimensional computing: an introduction to computing in distributed representation with high-dimensional random vectors. *Cogn. Comput.* **1**, 139–159 (2009).
39. Gayler, R.W. Vector symbolic architectures answer Jackendoff's challenges for cognitive neuroscience. In *ICCS/ASCS International Conference on Cognitive Science* (ed. Slezak, P) 133–138 (University of New South Wales, Sydney, Australia, 2004).
40. Cohen, T., Widdows, D., Vine, L.D., Schvaneveldt, R. & Rindflesch, T.C. Many paths lead to discovery: analogical retrieval of cancer therapies. In *Quantum Interaction*. (eds. Busemeyer, JR et al) 90–101 (Springer, Berlin Heidelberg, 2012).
41. Birkhoff, G. & von Neumann, J. The logic of quantum mechanics'. *Ann. Math.* **37**, 823–843 (1936).
42. Widdows, D. & Peters, S. Word vectors and quantum logic experiments with negation and disjunction. Proceedings of Mathematics of Language 8, Bloomington, Indiana, 2003 (eds. Oehrlé, RT and Rogers, J).
43. Semanticvectors - Google Code [Internet]. <<http://code.google.com/p/semanticvectors/>>.
44. Widdows, D. & Ferraro, K. Semantic Vectors: a Scalable Open Source Package and Online Technology Management Application. Sixth International Conference on Language Resources and Evaluation, 2008.
45. Widdows, D. & Cohen, T. The Semantic Vectors Package: New Algorithms and Public Tools for Distributional Semantics. Semantic Computing (ICSC), 2010 IEEE Fourth International Conference on, Pittsburgh, PA, 22–24 September 2010. pp 9–15.
46. Widdows, D., Cohen, T. & DeVine, L. Real, Complex, and Binary Semantic Vectors. QI'12 Proceedings of the 6th International Symposium on Quantum Interactions, Paris, France, 2012.
47. Aronson, A.R. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc. AMIA Symp.* **2001**, 17–21 (2001).
48. Aronson, A.R. & Lang, F.M. An overview of MetaMap: historical perspective and recent advances. *J. Am. Med. Inform. Assoc.* **17**, 229–236 (2010).
49. Martin, D.I. & Berry, M.W. Mathematical foundations behind latent semantic analysis. In Landauer T, McNamara D, Dennis S, Kintsch W (Eds.), *Handbook of latent semantic analysis*, Lawrence Erlbaum Associates, Mahwah, NJ (2007).



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-nd/3.0/>

Supplementary information accompanies this paper on the *CPT: Pharmacometrics & Systems Pharmacology* website (<http://www.nature.com/psp>)