

Accelerating science with human-aware artificial intelligence

In the format provided by the
authors and unedited

Table of Contents

α -modified Random Walk	1
Multistep Transition Probabilities	1
Computational Specifications	2
Study of Hyperparameters	3
Experiments with Deep Neural Networks	4
Combining Scores for Complementary Predictions	5
Evaluating the Alien Prediction Algorithm	6
Theoretical Scoring of Candidate Hypotheses	8
References	10
Supplementary Figure 1	11
Supplementary Figure 2	12
Supplementary Figure 3	13
Supplementary Figure 4	14
Supplementary Figure 5	15
Supplementary Table 1	16

α -modified Random Walk

The number of authors and material nodes in our hypergraphs are not balanced in any of the data sets: (94% vs. 6%) in the energy-related materials data set and (>99.95% vs. <0.05%) in the drug repurposing data set. Hence, classic random walk with uniform node sampling in each step will result in sequences where author nodes severely outnumber materials. This especially can be seen in drug-disease cases. In order to mitigate this issue, we devised a non-uniform node sampling distribution that can be tuned through a positive parameter denoted by α . Depending on the value of α , the algorithm samples materials more or less frequently than the authors. This parameter is officially defined as the ratio of the probability of sampling a material (if any) in any given paper to the probability of sampling a non-material node (either author or property nodes). We implemented this algorithm by a mixture of set-wise uniform sampling distributions. Let us denote the set of all nodes existing in the paper (hyperedge) of the i -th random walk step as N_i . This set can be partitioned into material and non-material nodes denoted by M_i and A_i , respectively. While the standard random walk samples the next node n_{i+1} from a uniform distribution over the unpartitioned set of nodes, i.e., $n_{i+1} \sim U(N_i)$, the α -modified random walk selects the next node by sampling from the following distribution (assuming both sets M_i and A_i are non-empty):

$$n_{i+1} \sim \frac{1}{\alpha+1}U(A_i) + \frac{\alpha}{\alpha+1}U(M_i)$$

We illustrated the sampling procedure in Supplementary Figure 1. In the first step, one paper e_1 is uniformly selected from the set of publications containing the property keyword in their title or abstract. The set of nodes and its two partitions are shown as N_0 , M_0 and A_0 . With probability $\pi = \frac{1}{\alpha+1}$ the walker picks out the next sample n_1 from A_0 and with probability $\frac{\alpha}{\alpha+1}$ the next sample will be from M_0 . As α gets larger the probability of sampling materials, and therefore their frequency in the resulting random walk sequences, increases. In the limit as $\alpha \rightarrow \infty$, the walker only samples materials unless the sampled paper does not contain any material nodes, in which case the sampling process is terminated. On the other hand, setting $\alpha=0$ results in author-only sequences.

Multistep Transition Probabilities

A simple way of calculating node similarities in (hyper)graphs is based on transition probabilities. The transition may happen in one or multiple steps. One-step transitions would give non-zero probabilities only to pairs of nodes that have appeared in the same paper. In order to incorporate higher-order connectivities, we consider two- and three-step transitions (Supplementary Figure 2) when computing similarities between properties and other nodes. When predicting discoveries (i.e., materials possessing a targeted property), transitions are considered between the property node and the materials through author nodes (Supplementary Figure 2a-b) and their probabilities are computed as the property-material similarity scores. Moreover, when predicting discoverers (i.e., authors that make discoveries regarding the property), we compute the probability of transitions between the property and the candidate authors through a single material node (Supplementary Figure 2c). These probabilities can be easily formulated via matrix-vector multiplications, as we demonstrate below. However, note that

the computation costs can quickly explode and slow down the calculations for transitions with more than three steps even though our hypergraph's vertex matrix is highly sparse.

Formulation—Here, we briefly demonstrate the formulation for computing the probability of property-material transitions. Similar formula can be derived for property-author transitions. Let us denote the property node by P and the material node by M . We considered two- and three-step transitions with intermediate nodes conditioned to belong to the set of authoring experts (denoted by A). In each case, the starting node n_0 is set to the property node and we compute the probability that a random walker reaches M in two or three steps, i.e., $n_2 = M$ or $n_3 = M$, respectively. Therefore, the probability of a two-step transition through an intermediate author node is computed as:

$$\begin{aligned} Pr(n_2 = M, n_1 \in A \mid n_0 = P) &= \sum_{a \in A} Pr(n_2 = M, n_1 = a \mid n_0 = P) \\ &= \sum_{a \in A} Pr(n_1 = a \mid n_0 = P) \cdot Pr(n_2 = M \mid n_1 = a) \end{aligned}$$

where the second line draws on the independence assumptions implied by the Markovian process of random walks. Similar formulation could be derived for three-step transition. The individual transition probabilities in the second line are readily available based on our definition of a hypergraph random walk. For example, for a classic random walk with uniform sampling distribution, we get

$$Pr(n_1 = a \mid n_0 = P) = \frac{1}{d(P)} \sum_{e: \{P, a\} \in e} \frac{1}{d(e)}$$

where $d(P)$ is the degree of node P , i.e., the number of hyperedges (papers) it belongs to, and $d(e)$ is the size of hyperedge e , i.e., the number of distinct nodes inside it. The first multiplicand in the right-hand side of above equation accounts for selecting a hyperedge that includes P and the second computes the probability of selecting a from one of the common hyperedges (if any).

The above computations can be compactly represented and efficiently implemented through matrix multiplication. Let $\mathbf{P}$ represent the transition probability matrix over all nodes such that $\mathbf{P}_{ij} = Pr(n_t = j \mid n_0 = i)$. Then, two- and three-step transitions between nodes P and M could be computed via $\mathbf{P}(P, [A]) \cdot \mathbf{P}([A], M)$ and $\mathbf{P}(P, [A]) \cdot \mathbf{P}([A], [A]) \cdot \mathbf{P}([A], M)$, respectively, where $\mathbf{P}(P, [A])$ defines selection of the row corresponding to node P and columns corresponding to authors in set A .

Computational Specifications

The random walk sequences were generated in a high-performance cluster, where each of the compute nodes has a processor Intel E5-2680v4 2.4GHz. This task is highly parallelizable. For each experiment, we generated 250,000 sequences using 40 compute nodes. For electrochemical properties (e.g., thermoelectric), the running time for such a random walk is on the order of 2 hours. On the other hand, these experiments take longer for diseases due to the substantially

larger literature database than we used for energy-related properties. For example, generating 250,000 random walk sequences for COVID-19 (i.e., property="coronavirus") took approximately 16 hours on 40 nodes.

The word embedding models were trained on a local workstation with AMD Ryzen Threadripper 2970WX processor with 48 cores and 128GB of RAM. Training a word2vec model (through "gensim" package version 3.8.0 in Python) on 20 CPU cores took about 1 to 2 minutes for both electrochemical properties and diseases.

Study of Hyperparameters

Our predictive pipeline runs with a set of hyperparameters. Among them are the non-uniform sampling distribution parameter (denoted by α), the number of years prior to the prediction year considered when building our hypergraph (which we call "memory" here) and the number of high-ranked materials when reporting the predicted materials (denoted by k). Here, we study the sensitivity of our algorithm with respect to values of these parameters.

Non-uniform Sampling Distribution Parameter (α)—We selected two more intermediate values for the α parameter (i.e., 0.5 and 10) and re-ran all the deepwalk experiments with the new values. The resulting precision values in Supplementary Figure 3 show that increasing the value of α beyond the balanced case (i.e., $\alpha > 1$ —hence, sampling more materials than authors) had less detrimental effects on the inference's precision than making it less than one. In other words, sampling materials more frequently than authors would harm less than sampling authors with higher frequency. Therefore, if one is to break the balance between researching and networking (deviating from $\alpha = 1$), overemphasizing research harms less than overemphasizing networking in terms of discovering scientific knowledge. However, even with these perturbations, performance of deepwalk predictions are superior to the word2vec baseline in almost all cases (except for Photovoltaics, when $\alpha = 0.5$).

Number of Predictions (k)—In our main results, we chose $k = 50$ materials closest to the property node (in terms of one of the similarity metrics) as the algorithm's predictions. Here, we show the results when we report alternative number of materials (i.e., 25 and 100). Supplementary Figure 4 shows that using smaller number of predictions ($k = 25$) generally resulted in higher precision values. However, the order of algorithms in terms of the quality of predictions showed very minor changes. More importantly, the precision margin between our hypergraph-based methods (i.e., deepwalk and transition probability criteria) and the word2vec baselines showed negligible differences among these cases.

Memory—We have restricted our framework to the most recent literature prior to the prediction year. The question is how recent we want to stay. Here, we refer to the number of years that we consider before the prediction year as "memory". On one hand, decreasing the memory (hence, ignoring/forgetting all publications produced before that) might unnecessarily deprive us from precious information in the previous years. On the other hand, increasing the memory might bring in too many articles whose authors are not working in the area anymore, hence their experience might have already become irrelevant to the random walker. In order to probe the sensitivity of our algorithm to the memory hyperparameter, we re-ran our discovery prediction experiments on inorganic materials again with memories 2 and 10. Figure S5 shows the resulting

cumulative precision curves of these experiments in comparison with the value originally used in our main experiments (i.e., memory=5). In both scenarios, the prediction performance of our deepwalk algorithm slightly decayed from our original experiments, however, in all cases they still remained significantly superior to word2vec.

Experiments with Deep Neural Networks

We also evaluated the effect of incorporating the distribution of expertise in our predictive models after replacing our deepwalk method with graph convolutional neural networks.

Graph Neural Networks (GNNs) have become a popular tool for learning low-dimensional graph representations or solving high-level tasks such as classification of graph nodes ¹. They owe this popularity to their unique and efficient way of exploiting graph connectivities to propagate information between a central node and its neighborhood, their ability to incorporate feature vectors for nodes and/or edges, and their superior generalization to unseen (sub)graphs. Similar to deepwalk, these models build a low-dimensional embedding space where graph-based similarities are preserved. However, unlike deepwalk, they incorporate node feature vectors and directly utilize graph connectivities for message passing between nearby nodes when constructing the embedding space.

The embedding vector of a central node is constructed by sequentially processing messages passed from its local neighbors. There are numerous ways of aggregating the signals reaching out from neighbors. In our experiments, we use the Graph Sample and Aggregate (GraphSAGE) platform, which applies the aggregations function on a subset of neighbors to avoid computational overhead ². Let h_i^l denotes the message from the i -th node in the l -th step of this sequential procedure. Then, the representation of the i -th node at the next level will be computed as

$$h_i^{l+1} = \sigma(f_{AGG}(\{h_j^l\}_{j \in \tilde{N}(i)})W_l)$$

where f_{AGG} is an aggregation function (e.g., mean, pooling, etc) applied on the concatenation of the local neighborhood's messages $\{h_j^l\}_{j \in \tilde{N}(i)}$, and $\tilde{N}(i)$ is a subset of k_l uniformly sampled nodes from the immediate neighbors of the i -th node $N(i)$. The resulting aggregated messages will undergo a single-layer neural network parameterized by W_l (the bias term is ignored for simplicity) and the non-linear activation σ . The input messages in the first step, i.e., $h_i^0 \forall i$, are set to the provided node feature vectors. The final representation of the i -th node will be reached after L steps. We used the same set of hyperparameters as the original paper ²; we considered two steps ($L=2$) with samples sizes $k_1=25$ and $k_2=10$. We also used the mean aggregation function and applied the non-linearity through Rectified Linear Unit (ReLU) activation.

We approached the discovery prediction problem in an unsupervised manner through a graph autoencoder ³, where the encoder component was modeled using the GraphSAGE architecture and the decoder component consisted of a parameter-less inner-product of the encoder's output. This autoencoder was trained by minimizing a link-prediction loss function, which was

approximated with negative sampling. The approximate loss has two parts accounting for the similarity of positive samples (pairs of nearby nodes) and the dissimilarity of negative samples (pairs of unconnected nodes).

Our mechanism of sampling positive and negative pairs closely resembled that which was used in deepwalk: the former is formed by pairing central/contextual nodes within windows sliding over short random walks, and the latter by means of sampling from the unigram distribution raised to power $3/4$ over the full set of nodes ⁴. Once positive samples were drawn using a sliding window size of 8, we began minimizing the loss function in a mini-batch setting by iterating over pairs. We used a batch size of 1000, negative sampling size of 15 (per positive pair), learning rate of 5×10^{-6} and the Adam optimizer ⁵ with default parameters.

We trained our graph autoencoder in two different settings: (i) using our full hypergraph, and (ii) after dropping author nodes. In both settings, we only considered the material and property nodes. In the full setting, we took account of author nodes when drawing positive samples and computing the adjacency matrix. In this setting, the positive samples were drawn from deepwalk sequences associated with $\alpha = 1$, whereas the experiment without authors used sequences generated after excluding the author nodes ($\alpha \rightarrow \infty$). Moreover, connectivities between nodes were different for the two settings. In the author-less network, we connected two property or material nodes only if they appeared in the same paper. In the full setting, we kept these connections and added more edges between nodes with at least one common author neighbor even in the absence of papers in which they co-occur.

Combining Scores for Complementary Predictions

We combine two sources of information for measuring human availability and scientific plausibility, each of which separately scores the candidate materials. The two scores will then be combined through β to result in a single scalar score such that the magnitude of β determines the weight of whether human availability is encouraged (negative) or discouraged (positive) among predictions, as opposed to scientific plausibility (0). As β grows from one extreme (-1) to the other ($+1$) the algorithm lessens its objective of human accessibility, instead avoiding it after passing $\beta = 0$. Hence, at least three special operating points are distinguishable based on the value of β :

- ($\beta = -1$): scientific plausability is given zero attention and full weight lies on finding hypotheses near human reasoners, i.e., maximizing the likelihood that the algorithm's predictions will be discovered by human scientists in the near future.
- ($\beta = 0$): scientific plausibility is given full emphasis and human availability of predictions is ignored.
- ($\beta = +1$): scientific plausibility is ignored and the full weight rests on human avoidance, i.e., maximizing the likelihood that predictions will escape human scientists' collective attention.

In addition to these special conditions, we desire an algorithm where contribution from the two sources (plausibility and human accessibility) become equal when $|\beta| = 1/2$, and the output score varies continuously as β shifts.

For any given candidate material x , let us denote its shortest-path distance (SPD) and its semantic similarity with respect to the property node by $s_1(x)$ and $s_2(x)$, respectively. These scores have distinct units and vary in different scales, therefore a naive β -weighted averaging is inappropriate as it does not lead to equal contribution when $|\beta|=1/2$. Moreover, the SPD values are unbounded as they can become arbitrarily large for entities disconnected from the property node in our hypergraph. As a result, Z-scores could not be directly applied either. Instead, we applied a Van der Waerden transformation to first standardize the scores. Suppose S is a set of scores and $s(x) \in S$, then its Van der Waerden transformation of x , denoted by $\tilde{s}(x)$, is defined as

$$\tilde{s}(x) = \varphi\left(\frac{r(x)}{|S|+1}\right)$$

where φ is the quantile function of the normal distribution, $r(x)$ is the rank of $s(x)$ within the set S and $|S|$ denotes the cardinality of S . We then take the weighted average of Z-scores for the transformed signals $\tilde{s}_1(x)$ and $\tilde{s}_2(x)$ for each material x as the ultimate hybrid score to be used in our final ranking. We further normalize the resulting transformed scores by computing their Z-scores yielding $\hat{s}_1(x)$ and $\hat{s}_2(x)$ before combining them in the following weighted averaging:

$$s_{final}(x) = \beta \hat{s}_1(x) + (1 - |\beta|) \hat{s}_2(x)$$

Note that when $\beta < 0$, the algorithm tends to select materials with lower $\hat{s}_1(x)$, which in turn implies smaller SPD and materials with more contextual familiarity in terms of property. Alternatively, when $\beta > 0$, the algorithm scores higher those materials with a greater SPD and more unfamiliar or “alien” predictions will result.

Evaluating the Alien Prediction Algorithm

An ideal alien prediction algorithm generates discovery candidates that are complementary to those published by human scientists, i.e., unimaginable yet plausible candidates. Hence, running our alien AI at a certain prediction year y_{pred} , the majority of predictions are expected not to be discovered by human scientists in the following years. However, for the few cases that are actually discovered in reality, we show that the average year of their discoveries tends to grow, thus becoming less imaginable to those at the present moment as alienness (β) increases. In order to assess the full set of predictions that are dominated by undiscovered predictions for large β values, we evaluate their scientific plausibility as the likelihood that they are potential disruptive discoveries. For this purpose, we conservatively rely on the simulative and theoretical prior knowledge that exist in the field (see next section). Specifically, we show that while the percentage of predictions that become discovered after the prediction year sharply decays with β , the quality of the predictions remains high for longer. This observation, which we will measure by a criterion named the *expectations gap*, implies the existence of intervals of β where our alien AI approaches an ideal alien predictor by producing plausible undiscoverable predictions. Additionally, we explicitly estimated the joint probability of being plausible and undiscoverable at the same time to identify those intervals of β where the our alien algorithm approaches the ideal operation point.

Expectation Gap—Our goal is to define a single scalar score indicating whether our alien predictor is capable of generating unimaginable yet promising predictions. We define two distributions over β conditioned on (1) discoverability and (2) plausibility of predictions. The separation between these two distributions such that the latter is more concentrated on larger β values than the former will be our indicator, which will be quantified by the gap between their means (hence expectation gap).

For a fixed property and pool of candidate materials, let us denote the set of all materials that will be discovered after y_{pred} by D and the set of all plausible materials by P . Also, let H_β and h_β be the full set of predictions and a randomly selected prediction generated by our alien algorithm operating with β , respectively. Precision of the algorithm in terms of identifying near-future discoveries is defined as $Pr(h_\beta \in D | \beta)$, which can be simply computed by dividing the count of discoveries by the number of predictions: $|H_\beta \cap D| / |H_\beta|$. Now using a uniform prior distribution over β , i.e., $Pr(\beta) = \text{const.}$, and applying Bayes rule these precisions can be converted to $Pr(\beta | h_\beta \in D)$ by normalization across all β values such that they sum to unity.

Computing the second distribution is not as straightforward. The subtlety arises from the fact that we do not fully know P , instead we have one real-valued score per material characterizing the likelihood of its P -membership. These scores, denoted by τ , are obtained from field-related theoretical knowledge and first-principles laws (see next section). In the first step, we transform the theoretical scores to probabilities, such that for every material x , $\tau = \tau(x)$ goes to $Pr(x \in P)$. Let $\tau_{\min}$ and $\tau_{\max}$ be the global minimum and maximum of all the theoretical scores at hand. We engineered a monotonically increasing transformation T in the form of $\text{logit}[\tan(\pi(\hat{\tau} - \frac{1}{2}) + b)]$, where $\hat{\tau} = (\tau - \tau_{\min}) / (\tau_{\max} - \tau_{\min})$ such that

- $T(\tau_{\min}) = 0$
- $T(\tau_{\max}) = 1$
- $T(\tau_{\text{mid}}) = 1/2$, where $\tau_{\text{mid}} = \frac{1}{|D|} \sum_{x \in D} \tau(x)$, which is the average of theoretical scores attributed to discovered materials. This condition uniquely specifies parameter b .

The resulting probabilities will be thresholded by 1/2 to probabilistically indicate which materials belong to P . Setting the midpoint as above is a direct consequence of the assumption that the majority of the materials discovered by the scientists to have the targeted properties are plausible findings, hence we take their average theoretical scores as a baseline and every material with a higher τ will also be considered as plausible. Such probabilistic classification of a material x to P is done with a confidence level proportional to the distance between the probability $T(\tau(x))$ and the threshold 1/2. The confidence level of our decision regarding P -membership of sample x is

$$c(x) = \begin{cases} T(\tau(x)) & , \text{ if } T(\tau(x)) \geq 1/2 \\ 1 - T(\tau(x)) & , \text{ if } T(\tau(x)) < 1/2 \end{cases}$$

Now, for any prediction set H_β , we use weighted maximum likelihood estimation to compute the probability of being plausible given β :

$$Pr(h_\beta \in P|\beta) = \frac{\sum_{x \in H_\beta : T(\tau(x)) \geq 1/2} c(x)}{\sum_{x \in H_\beta} c(x)}$$

Finally, similar to the previous case, the likelihood of β given plausibility, $Pr(\beta|h_\beta \in P)$, can be obtained by simply normalizing these probabilities across all β values such that they sum to one.

The expectation gap is defined as the difference between the mean values of the two likelihoods described above:

$$\Delta E[\beta] := E[\beta|h_\beta \in P] - E[\beta|h_\beta \in D]$$

Having a positive gap suggests that theoretical plausibility is higher for more alien predictions than for those where the predictions are made and published. Therefore, because the discovery precision goes down with β , a positive expectation gap also means that there exists a non-empty gap when the alien algorithm approaches its ideal mission of human complementarity. Zero or negative gaps only occurred for a few human diseases.

Joint Probabilities—Our expectation gap provided us with a single evaluation score for the performance of the alien prediction. However, it does not say anything about the desired range of β where the algorithm operates closest to its mission to serve complementary, high-value predictions. Here, to provide a clearer overview of the performance of our alien prediction for different β values, we directly model and calculate the probability that it outputs unfamiliar yet scientifically promising (plausible) predictions. As is also described above, unfamiliarity of a random prediction h_β means its unimaginability in context of the considered property and therefore its undiscoverability (i.e., $h_\beta \notin D$). Thus, we calculate the probability that h_β is unfamiliar and plausible given a certain β by the joint distribution $Pr(h_\beta \notin D, h_\beta \in P|\beta)$. Applying Bayes rule, this joint probability decomposes into two simpler distributions:

$$Pr(h_\beta \notin D, h_\beta \in P|\beta) = Pr(h_\beta \notin D|\beta) Pr(h_\beta \in P|\beta, h_\beta \notin D)$$

where the first term in the right-hand-side is the complementary of discovery precision (i.e., $1 - Pr(h_\beta \in D|\beta)$) and the second term can be computed similarly to the probability of plausibility, $Pr(h_\beta \in P|\beta)$, described above, except that all computations are to be repeated on the predictions that are *not* discovered after the prediction year ($h_\beta \notin D$), hence replacing H_β with $H_\beta - D$ when doing the weighted maximum likelihood above.

Theoretical Scoring of Candidate Hypotheses

In order to assess the quality of predictions resulting from our algorithm, we used theoretically driven scores derived from first-principles equations or simulation models from the relevant disciplines. The underlying procedure for defining and curating such data naturally differs for distinct properties. In this section, we describe these theoretical scores for properties about which datasets could be found.

Thermoelectricity—Thermoelectricity is the property of producing electrical voltage when temperature varies on two sides of a material. An important measure of thermoelectricity that depends on the Seebeck coefficient and electrical conductivity is called Power Factor (PF). PF is also the major part of another common dimensionless criterion named its *figure of merit* (zT). First-principles methods based on Density Functional Theory (DFT) have been widely used to estimate energy-related chemical properties including PF⁶. Recently, DFT-based PF estimates have been used to evaluate content-based discovery predictions⁷, where it is shown that materials studied in connection with thermoelectricity have larger PF values on average than unstudied materials. Here, we use the same set of PF estimates, which have been prepared by taking the maximum of the average PFs computed for various temperatures, doping levels and semiconductor types. Among 12.6K materials that were mentioned at least once before the prediction year of 2001 (i.e., within the five-year period [1996,2000]), only about 3.5K materials were assigned a PF estimate.

Ferroelectricity—Ferroelectric materials are characterized by their spontaneous electric polarization, which is reversible in the presence of an external electric field. Smidt et al. developed an automated workflow that uses symmetry analysis and first-principles calculations to curate a list of ferroelectric materials⁸. Their final list included 255 candidates, where entries are rated based on the magnitude of their spontaneous polarization. There are 167 distinct chemical formulae in this list, and for some several ferroelectricity scores are reported corresponding to various crystal structures. For compounds with multiple values, we considered the maximum computed ferroelectricity over all available structures as the final score. Among the materials mentioned in the most recent five-year period preceding the prediction year of 2001, only 167 distinct compounds that have been reported as ferroelectric by Smidt et al. are assigned non-zero scores and the rest are considered to have zero ferroelectricity. This leads to a sparse scoring system but, as is shown in our results, remains nevertheless sufficient to demonstrate the performance of our method.

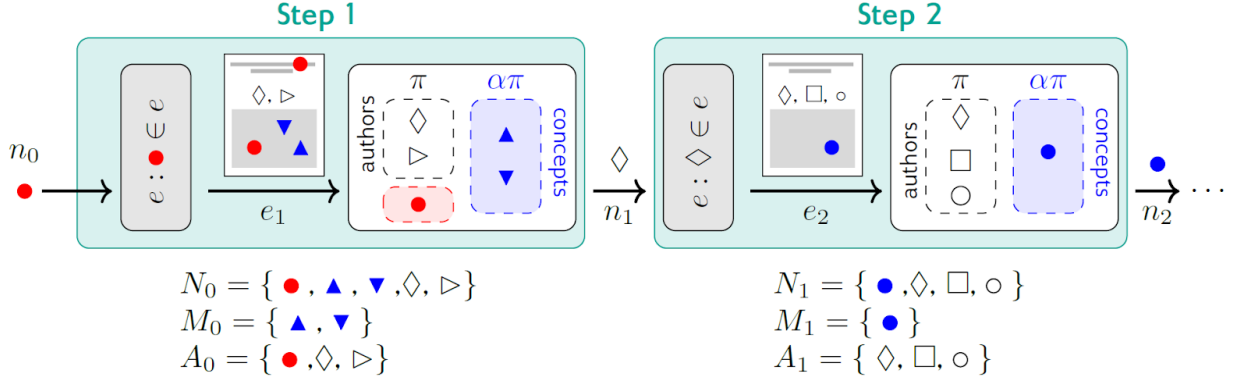
Human Diseases—The similarity between two sets of proteins targeted by a particular disease and a certain drug forms a basis for measuring their underlying association and therapeutic potential. Protein-protein interaction networks include protein, drug and disease nodes with pairwise interactions encoded within the edges. We utilize drug-disease proximity in such networks as the core of our theoretical scoring framework to assess drug candidates in terms of their relevance to the treatment of a disease. Recently, Gysi et al. showed that drugs whose target proteins are within or in vicinity of the COVID-19 disease module are potentially strong candidates for repurposing to treat or prevent the infection⁹. They used various proximity measures to compute network similarities and identify the most relevant drug candidates. Among these measures was cosine similarity of embedding vectors resulting from a pretrained Graph Neural Network (GNN) over the protein-protein network, which we used in our evaluations of drug-disease proposals.

The GNN-based embedding vectors of the interaction network produced and shared by Gysi et al.⁹ included 1.6K drug nodes and 2.5K disease nodes (including COVID-19), among which 1.5K drugs belonged to our pool of candidates and 176 diseases were common with the set of 400 human diseases we considered in our expert-avoiding experiments. We ran our complementary algorithm at prediction year 2001 for all diseases but COVID-19, where we restricted our pool of candidates to 1,179 drugs that existed in the protein-protein network and

also appeared in the literature between 1996 and 2000. For COVID-19, the prediction year is set to the beginning of 2020 with 1,436 candidate drugs that existed in both the interaction network and in the literature between 2015 and 2019.

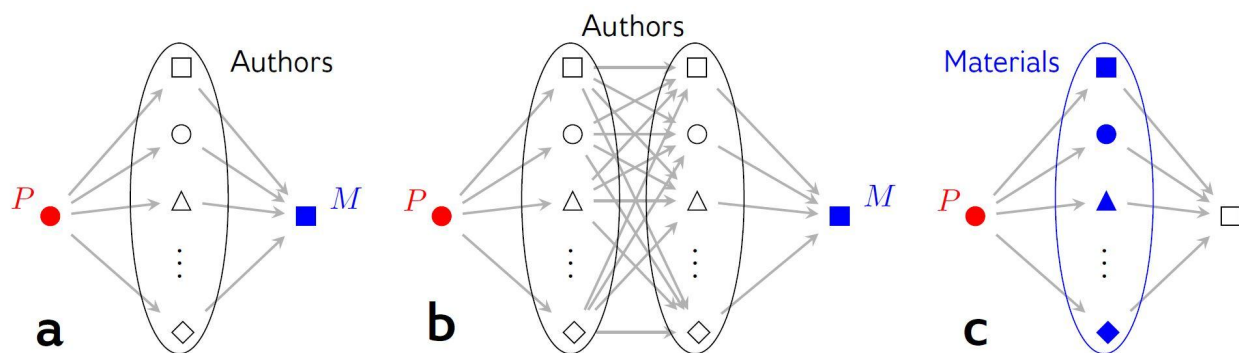
References

1. Chami, I., Abu-El-Haija, S., Perozzi, B., Ré, C. & Murphy, K. Machine Learning on Graphs: A Model and Comprehensive Taxonomy. *arXiv [cs.LG]* (2020).
2. Hamilton, W., Ying, Z. & Leskovec, J. Inductive Representation Learning on Large Graphs. in *Advances in Neural Information Processing Systems* (eds. Guyon, I. et al.) vol. 30 (Curran Associates, Inc., 2017).
3. Kipf, T. N. & Welling, M. Variational Graph Auto-Encoders. *Stat* **1050**, 21 (2016).
4. Mikolov, T., Yih, W.-T. & Zweig, G. Linguistic Regularities in Continuous Space Word Representations. in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 746–751 (Association for Computational Linguistics, 2013).
5. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *arXiv [cs.LG]* (2014).
6. Ricci, F. *et al.* An ab initio electronic transport database for inorganic materials. *Sci Data* **4**, 170085 (2017).
7. Tshitoyan, V. *et al.* Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571**, 95–98 (2019).
8. Smidt, T. E., Mack, S. A., Reyes-Lillo, S. E., Jain, A. & Neaton, J. B. An automatically curated first-principles database of ferroelectrics. *Sci Data* **7**, 72 (2020).
9. Morselli Gysi, D. *et al.* Network medicine framework for identifying drug-repurposing opportunities for COVID-19. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).



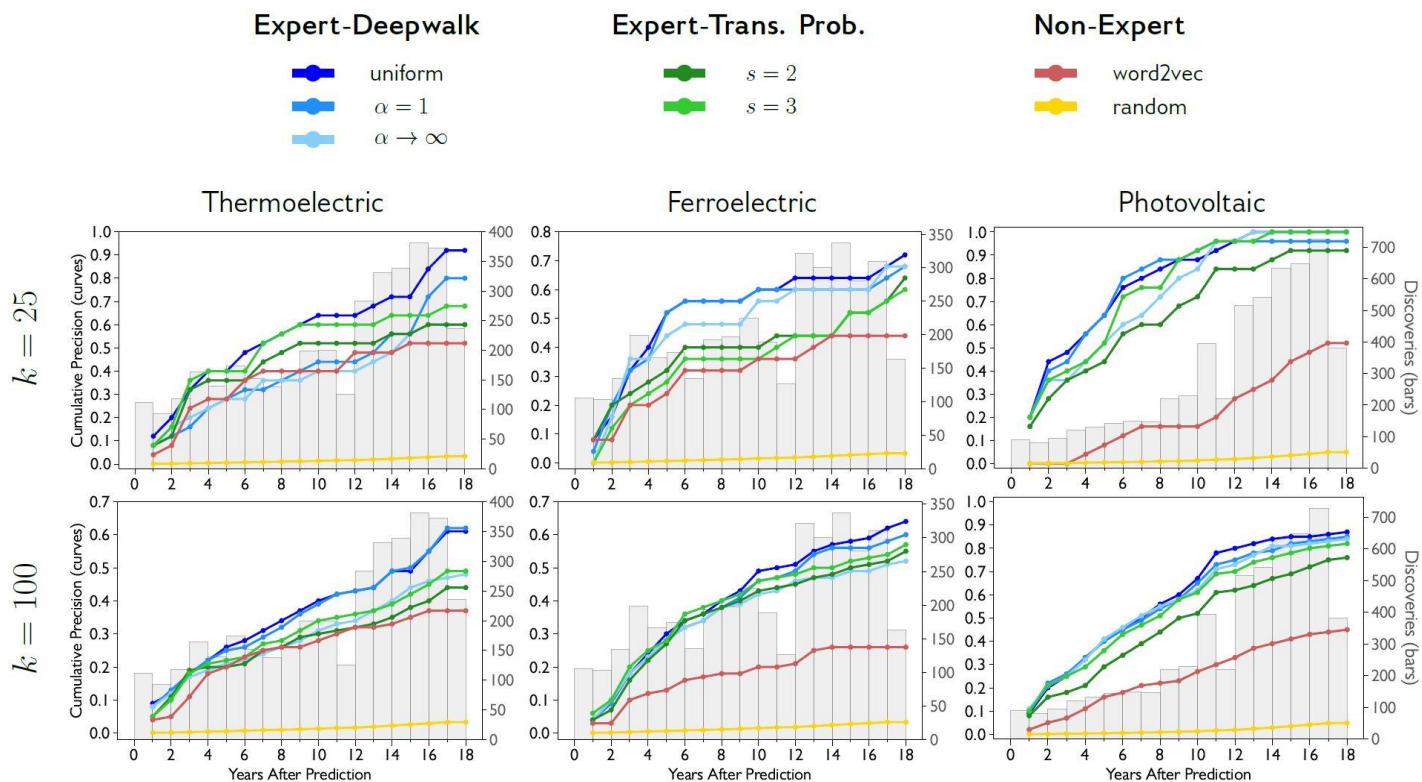
Supplementary Figure 1.

Two steps in an α -modified random walk. Blank shapes represent author nodes and colored shapes represent materials (blue) and the property (red). Papers (hyperedges) are sampled uniformly, whereas nodes are selected such that the probability of sampling a material node is α times the probability of sampling a non-material node. Here, π denotes the probability of sampling non-material nodes, which is uniquely determined by α itself. In each random walk step, the selected hyperedge is shown over the arrow (e_1 or e_2) and the hypernodes that it contains are listed below the figure (N_0 or N_1), which are in turn partitioned into material (M_0 or M_1) and non-material (A_0 or A_1) subsets. The output of the considered step will be a random draw from these hypernodes ($n_1 \in N_0$ or $n_2 \in N_1$).



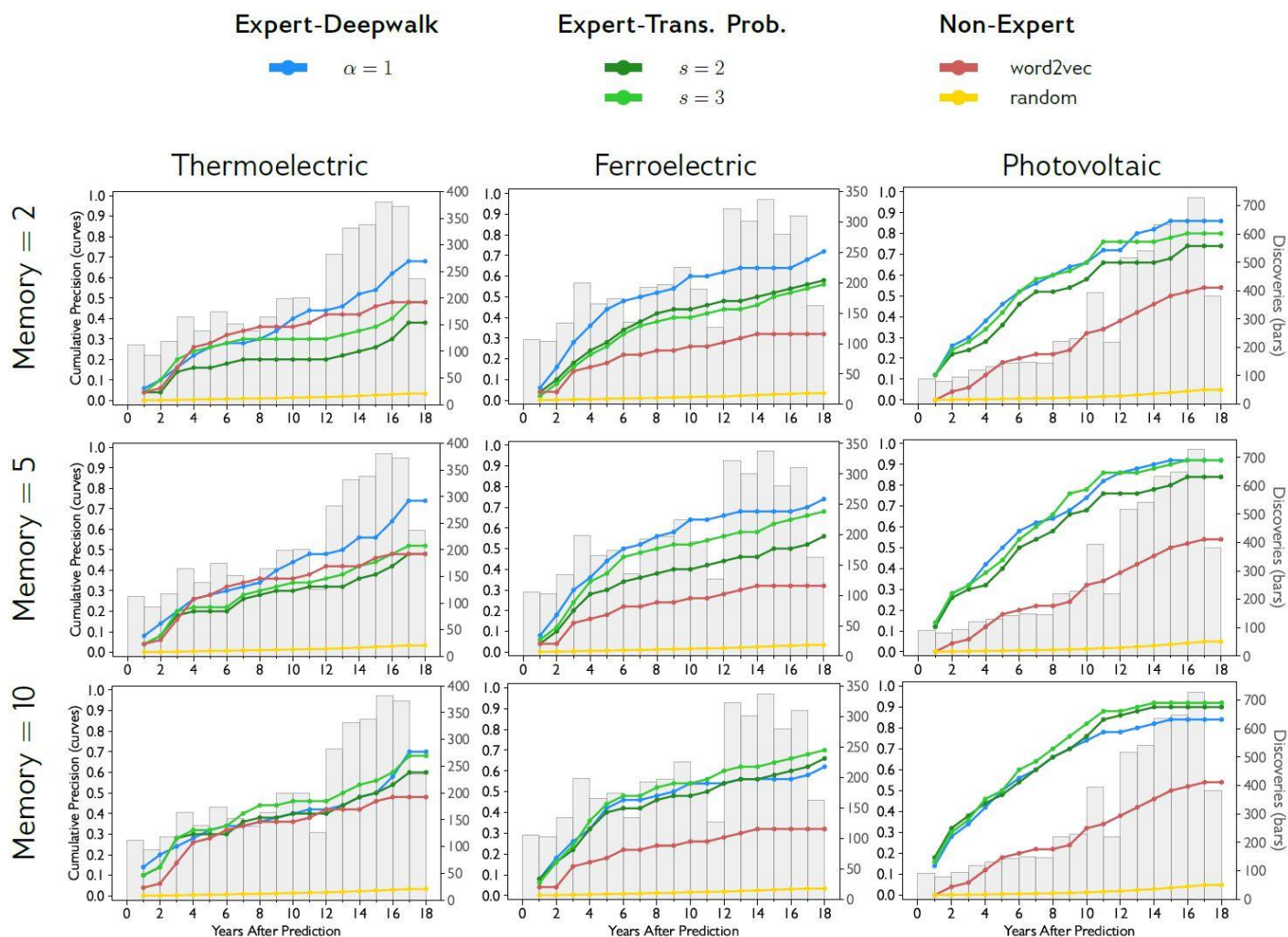
Supplementary Figure 2.

Using transition probabilities as the similarity criterion in our inferences. **(a)** Two-step transitions ($s=2$) between the property node P and a particular material M through a single intermediate author node. **(b)** Three-step transitions between ($s=3$) between the property node P and a particular material M through two intermediate author nodes. **(c)** Two-step transitions between property P and a particular author (uncolored square at the right) through a single material node. The probability of such property-author transitions is taken as the similarity between author and property nodes, which in turn, is used for selecting $k=50$ authors that are most similar to the property and reporting them as the potential discoverers of materials possessing property P .



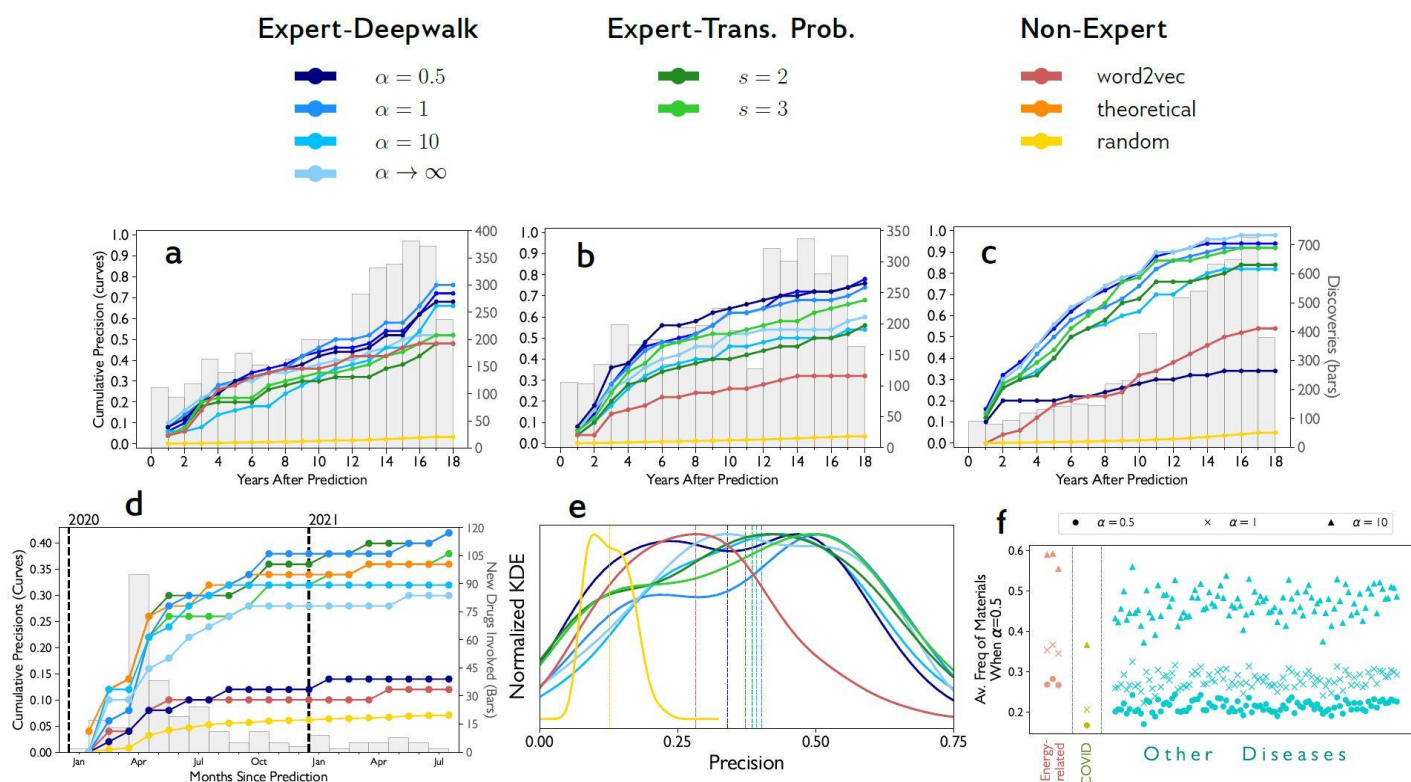
Supplementary Figure 3.

Precision of predicting energy-related properties for inorganic materials when different numbers of materials are reported as the predictions ($k=25$ and 100).



Supplementary Figure 4.

Precision of predicting energy-related properties for inorganic materials for different memory values (i.e., the number of years considered prior to the prediction year when constructing our hypergraph).



Supplementary Figure 5.

Precision of discovery predicting with more intermediate values for α parameter. Subfigures (a-d) show cumulative precision curves for properties associated with being (a) thermoelectric, (b) ferroelectric, (c) photovoltaic and (d) effective in therapy of or vaccination against COVID-19. Subfigure (e) shows the KDE-estimated distribution of precision values for inferring materials that are useful for therapy of 100 diseases in CTD, done by different algorithms. Finally, subfigure (f) indicates average frequency of materials in random-walk sequences generated with three different intermediate values for α . In all cases, whereas performance of our deepwalk algorithm with $\alpha=10$ did not drastically change from $\alpha=1$, one can see a significant drop in precision when $\alpha=0.5$ is used.

Supplementary Table 1. Evaluating predicted discoveries by our complementary discovery prediction algorithms operated with β values 0, 0.2 and 0.4.

Disease	$\beta = 0$		$\beta = 0.2$		$\beta = 0.4$	
	Overlap (%)	Pr-Pr Sim.	Overlap (%)	Pr-Pr Sim.	Overlap (%)	Pr-Pr Sim.
Acromegaly	0	0.323	0	0.32	0	0.259
Adenocarcinoma	11	0.066	10	0.009	13	-0.133
Alcoholism	23	0.565	19	0.539	15	0.475
Alopecia	37	0.152	26	0.165	25	0.2
Amenorrhea	3	0.264	3	0.268	1	0.273
Amyloidosis	22	0.269	18	0.257	16	0.253
Anaphylaxis	4	0.367	3	0.369	1	0.335
Anemia	22	0.155	21	0.15	22	0.168
Angioedema	11	0.183	10	0.205	5	0.225
Appendicitis	19	0.281	13	0.222	8	0.145
Arteriosclerosis	41	0.441	35	0.423	28	0.406
Arthritis	13	0.309	9	0.276	6	0.129
Arthrogryposis	2	0.363	2	0.364	1	0.305
Ascites	6	0.142	5	0.176	1	0.245
Asthma	10	0.351	9	0.329	16	0.243
Astrocytoma	39	0.169	32	0.165	25	0.111
Ataxia	23	0.423	17	0.405	12	0.333
Atherosclerosis	33	0.387	35	0.4	32	0.281
Blepharospasm	2	0.305	1	0.259	1	0.14
Blindness	0	0.219	1	0.212	1	0.208
Brachydactyly	4	0.238	4	0.164	1	0.036
Bronchiectasis	36	0.408	34	0.387	20	0.341
Brucellosis	2	0.288	2	0.215	1	0.065
Candidiasis	6	0.225	5	0.155	0	0.069
Carcinoma	15	0.163	12	0.086	10	-0.176
Cardiomyopathies	15	0.393	15	0.379	11	0.383
Cataract	15	0.327	12	0.327	9	0.314
Cholangiocarcinoma	48	0.213	42	0.168	30	0.001
Cholangitis	10	0.217	8	0.164	6	0.097
Cholelithiasis	6	0.22	6	0.171	5	0.106

(To be continued)

Supplementary Table 1. (continued)

Disease	$\beta = 0$		$\beta = 0.2$		$\beta = 0.4$	
	Overlap (%)	Pr-Pr Sim.	Overlap (%)	Pr-Pr Sim.	Overlap (%)	Pr-Pr Sim.
Cholelithiasis	6	0.22	6	0.171	5	0.106
Cholestasis	23	0.255	26	0.241	32	0.239
Chorioamnionitis	17	0.274	13	0.236	11	0.194
Colitis	41	0.256	36	0.262	31	0.223
Constipation	19	0.334	16	0.329	10	0.304
Contracture	0	0.408	0	0.362	0	0.329
Cryptorchidism	43	0.35	35	0.337	17	0.26
Cystitis	11	0.361	12	0.362	9	0.283
Delirium	10	0.378	10	0.36	8	0.296
Dermatitis	17	0.311	13	0.252	8	0.206
Dermatomyositis	37	0.216	30	0.224	21	0.214
Diarrhea	8	0.267	10	0.283	8	0.251
Dizziness	6	0.551	6	0.535	3	0.433
Dysarthria	6	0.359	4	0.315	2	0.277
Dyskinesias	35	0.623	34	0.61	25	0.518
Dyslipidemias	35	0.377	28	0.356	14	0.26
Dyspnea	13	0.4	10	0.43	9	0.425
Eczema	7	0.225	6	0.215	6	0.126
Embolism	3	0.246	2	0.226	2	0.202
Emphysema	22	0.378	21	0.33	17	0.312
Endometriosis	55	0.327	46	0.314	39	0.195
Entamoebiasis	31	0.022	24	-0.05	12	-0.11
Enterocolitis	12	0.092	13	0.1	9	0.107
Eosinophilia	9	0.24	7	0.215	6	0.136
Epilepsy	17	0.457	14	0.436	15	0.331
Exanthema	11	0.108	10	0.092	9	0.083
Gallstones	4	0.216	2	0.181	1	0.101
Gastroenteritis	3	-0.081	2	-0.069	1	-0.006
Gastroparesis	15	0.18	12	0.18	9	0.186
Glaucoma	14	0.358	10	0.383	8	0.33

(To be continued)

Supplementary Table 1. (continued)

Disease	$\beta = 0$		$\beta = 0.2$		$\beta = 0.4$	
	Overlap (%)	Pr-Pr Sim.	Overlap (%)	Pr-Pr Sim.	Overlap (%)	Pr-Pr Sim.
Glioblastoma	33	0.336	25	0.301	13	0.143
Glioma	35	0.223	30	0.152	28	-0.002
Gliosarcoma	20	0.18	18	0.148	12	0.122
Glomerulonephritis	9	0.226	7	0.206	5	0.176
Goiter	3	0.009	2	-0.011	2	-0.002
Gout	39	0.327	35	0.286	27	0.264
Heartburn	3	0.159	3	0.147	2	0.142
Hemangioblastoma	17	-0.079	13	-0.1	5	-0.135
Hemangioma	5	0.172	5	0.174	3	0.131
Hemorrhage	4	0.263	4	0.263	3	0.233
Hemorrhoids	3	0.198	3	0.188	2	0.205
Hyperaldosteronism	6	0.429	4	0.418	2	0.398
Hyperalgesia	21	0.522	20	0.51	19	0.432
Hyperammonemia	4	0.206	4	0.21	3	0.205
Hypercholesterolemia	25	0.383	24	0.377	19	0.295
Hyperinsulinism	10	0.28	10	0.271	7	0.245
Hyperkalemia	8	0.182	5	0.171	5	0.183
Hyperlipidemias	32	0.423	27	0.397	13	0.277
Hyperparathyroidism	13	0.243	12	0.245	11	0.191
Hyperphosphatemia	7	0.413	6	0.408	4	0.376
Hyperpigmentation	11	0.234	10	0.224	7	0.198
Hyperplasia	34	0.165	35	0.172	31	0.198
Hyperprolactinemia	2	0.206	1	0.175	0	0.11
Hypersensitivity	19	0.23	20	0.22	24	0.189
Hypertension	3	0.553	5	0.49	9	0.361
Hyperthyroidism	18	0.04	16	0.039	16	0.027
Hypertrophy	41	0.394	38	0.394	25	0.32
Hyperuricemia	22	0.317	17	0.35	11	0.262
Hypoalbuminemia	19	0.185	18	0.175	14	0.124
Hypocalcemia	5	0.232	3	0.242	2	0.227

(To be continued)

Supplementary Table 1. (continued)

Disease	$\beta = 0$		$\beta = 0.2$		$\beta = 0.4$	
	Overlap (%)	Pr-Pr Sim.	Overlap (%)	Pr-Pr Sim.	Overlap (%)	Pr-Pr Sim.
Hypoglycemia	4	0.228	3	0.205	3	0.255
Hypogonadism	5	0.213	8	0.169	6	0.095
Hypokalemia	4	0.396	3	0.387	2	0.339
Hyponatremia	7	0.289	6	0.264	6	0.241
Hypoparathyroidism	2	0.217	2	0.243	1	0.177
Hypophosphatasia	16	-0.133	13	-0.124	7	-0.102
Hypophosphatemia	8	0.254	8	0.23	5	0.172
Hypopigmentation	3	0.091	3	0.103	2	0.175
Hypopituitarism	16	0.218	15	0.206	13	0.196
Hypotension	6	0.52	6	0.495	8	0.423
Hypothyroidism	7	0.235	3	0.208	2	0.188
Hypotrichosis	3	0.284	1	0.263	0	0.207
Hypoxia	12	0.361	10	0.373	8	0.332
Keloid	37	0.291	34	0.268	20	0.175
Keratitis	22	0.27	13	0.193	7	0.138
Keratosis	38	0.152	31	0.131	15	0.134
Leiomyoma	21	0.141	17	0.104	12	-0.002
Leiomyosarcoma	30	0.219	28	0.126	18	-0.008
Leishmaniasis	52	0.305	43	0.249	28	0.165
Leprosy	3	0.089	2	0.065	1	0.021
Leukemia	21	0.173	16	0.111	9	0.023
Liposarcoma	14	0.145	12	0.079	4	-0.119
Listeriosis	30	0.274	26	0.219	17	0.012
Lymphangioleiomyomatosis	10	0.028	8	-0.014	2	-0.144
Lymphoma	16	0.29	17	0.224	15	0.064
Malaria	7	0.241	4	0.22	0	0.129
Malnutrition	8	0.078	6	0.066	4	0.05
Medulloblastoma	35	0.291	28	0.224	20	0.093
Melanoma	24	0.257	19	0.232	15	0.145
Meningitis	3	0.169	3	0.105	1	0.042

(To be continued)

Supplementary Table 1. (continued)

Disease	$\beta = 0$		$\beta = 0.2$		$\beta = 0.4$	
	Overlap (%)	Pr-Pr Sim.	Overlap (%)	Pr-Pr Sim.	Overlap (%)	Pr-Pr Sim.
Methemoglobinemia	16	0.196	16	0.184	12	0.116
Mucositis	17	0.159	18	0.125	15	0.038
Myoclonus	15	0.647	12	0.627	10	0.515
Myositis	2	0.261	3	0.223	1	0.247
Narcolepsy	11	0.504	10	0.486	5	0.352
Neoplasms	23	0.2	26	0.09	24	0.006
Nephritis	9	0.219	11	0.226	4	0.178
Nephrolithiasis	4	0.415	6	0.4	2	0.332
Nephrosis	6	0.328	5	0.339	5	0.273
Neuroblastoma	13	0.253	11	0.267	10	0.096
Neurofibrosarcoma	3	0.151	2	0.11	2	-0.035
Neutropenia	11	0.228	9	0.171	8	0.089
Ophthalmoplegia	3	0.147	3	0.166	3	0.134
Osteoarthritis	30	0.342	27	0.311	27	0.152
Osteoporosis	35	0.398	33	0.372	26	0.291
Osteosarcoma	56	0.349	45	0.29	36	0.163
Pancreatitis	41	0.175	40	0.159	33	0.166
Paresis	1	0.339	0	0.305	0	0.273
Periodontitis	17	0.315	11	0.303	8	0.259
Peritonitis	16	0.346	15	0.324	12	0.2
Pheochromocytoma	14	0.278	13	0.275	9	0.26
Photophobia	4	0.244	2	0.225	2	0.204
Pneumonia	32	0.424	32	0.363	22	0.211
Polyneuropathies	10	0.36	8	0.346	3	0.303
Polyuria	5	0.396	3	0.387	2	0.351
Pre-Eclampsia	19	0.362	17	0.348	15	0.341
Prolactinoma	2	0.234	1	0.208	0	0.168
Prostatitis	10	0.477	9	0.421	4	0.285
Proteinuria	8	-0.022	10	0.016	10	0.088
Pruritus	30	0.294	31	0.29	22	0.238

(To be continued)

Supplementary Table 1. (continued)

Disease	$\beta = 0$		$\beta = 0.2$		$\beta = 0.4$	
	Overlap (%)	Pr-Pr Sim.	Overlap (%)	Pr-Pr Sim.	Overlap (%)	Pr-Pr Sim.
Psoriasis	5	0.324	7	0.287	10	0.165
Pyelonephritis	1	0.472	1	0.366	1	0.173
Rhabdomyosarcoma	18	0.292	14	0.218	9	0.066
Rhinitis	30	0.473	20	0.421	8	0.293
Sarcoidosis	10	0.2	8	0.194	7	0.12
Sarcoma	16	0.314	11	0.265	8	0.076
Schistosomiasis	6	0.153	6	0.147	5	0.105
Schizophrenia	18	0.659	18	0.573	23	0.463
Seizures	6	0.608	10	0.603	15	0.482
Seminoma	13	0.153	8	0.107	5	-0.101
Sinusitis	5	0.46	4	0.379	2	0.272
Stomatitis	9	0.064	7	0.095	2	0.079
Stroke	17	0.399	15	0.401	9	0.28
Synovitis	10	0.234	7	0.218	4	0.15
Tetany	3	0.162	3	0.165	2	0.126
Thrombocytopenia	14	0.252	11	0.256	5	0.234
Thromboembolism	8	0.215	8	0.183	5	0.141
Thrombosis	9	0.294	18	0.296	18	0.255
Toothache	2	0.17	2	0.155	0	0.046
Torticollis	4	0.462	3	0.409	1	0.263
Tremor	11	0.6	9	0.59	8	0.534
Trichuriasis	18	0.161	15	0.157	9	0.184
Tuberculosis	5	0.269	4	0.243	3	0.193
Urticaria	26	0.329	23	0.333	18	0.282
Uveitis	13	0.218	12	0.195	7	0.192
Vitiligo	46	0.167	39	0.15	25	0.095
Xerostomia	4	0.22	4	0.214	3	0.139