Accelerating science with human versus alien artificial intelligences

Jamshid Sourati[a]
James Evans[a,b*]

[a]University of Chicago
1155 S. 60th Street
Chicago, IL 60637

[b]Santa Fe Institute
1399 Hyde Park Road
Santa Fe, NM 87501

Data-driven artificial intelligence models fed with published scientific findings have been used to create powerful prediction engines for scientific and technological advance, such as the discovery of novel materials with desired properties[1–3] and the targeted invention of new therapies and vaccines[4–6]. These AI approaches typically ignore the distribution of human prediction engines—scientists and inventors—who continuously alter the landscape of discovery and invention. As a result, AI hypotheses are designed to substitute for human experts[1], failing to complement them for punctuated collective advance. Here we show that incorporating the distribution of human expertise into self-supervised models by training on inferences cognitively available to experts dramatically improves AI prediction of future human discoveries and inventions. Including expert-awareness into models that propose (a) valuable energy-relevant materials increases the precision of materials predictions by ~100%, (b) repurposing thousands of drugs to treat new diseases increases precision by 43%, and (c) COVID-19 vaccine candidates examined in clinical trials by 260%. These models succeed by predicting human predictions and the scientists who will make them. By tuning AI to avoid the crowd, however, it generates scientifically promising "alien" hypotheses unlikely to be imagined or pursued without intervention, not only accelerating but punctuating scientific advance. By identifying and correcting for collective human bias, these models also suggest opportunities to improve human prediction by reformulating science education for discovery.

*Correspondence to jevans@uchicago.edu

Research across applied science and engineering, from materials discovery to drug and vaccine development, is hampered by enormous design spaces that overwhelm researchers' ability to evaluate the full range of potentially valuable candidate designs by simulation and experiment[7]. To face this challenge, researchers have initialized data-driven AI models with published scientific results to create powerful prediction engines. These models are being used to enable discovery of novel materials with desirable properties[2] and targeted construction of new therapies[4]. But such efforts typically ignore the distribution of scientists and inventors—human prediction engines—who continuously alter the landscape of discovery and invention. As a result, AI algorithms unwittingly compete with human experts, failing to complement them and augment collective advance. As we demonstrate below, incorporating knowledge of human experts and expertise can improve predictions of future discoveries by more than 100% above AI methods that ignore them. Nevertheless, with tens of millions of active scientists and engineers around the world, is the production of artificial intelligences that mimic human capacity our most strategic or ethical investment? By not mimicking, but rather avoiding human inferences we can design "alien" AIs that radically augment rather than replace human capacity. Identifying the bias of collective human discovery, we demonstrate how human-avoiding or alien algorithms broaden the scope of things discovered by identifying hypotheses unlikely for scientists and inventors to imagine or pursue with undiminished signs of scientific and technological promise.

Our analysis builds on insights underlying the wisdom of crowds[8], which hinges on the independence and diversity of crowd members' information[9] and approach[10]. In scientific crowds, findings established by more distinct methods and researchers are much more likely to replicate[11,12]. This diversity of scientific viewpoints was implicitly drawn upon by Donald Swanson in a heuristic approach to knowledge generation. He hypothesized that if Raynaud's disorder was linked to blood viscosity in one literature, and fish oil was known to decrease that viscosity in another, then fish oil might lessen the symptoms of Raynaud's disorder but would unlikely be arrived at by the sparse scientific community available to infer it[13–15], one of several hypotheses later experimentally demonstrated[16–18]. Our approach scales and makes this heuristic continuous, combining it with explicit measurement of the distribution of scientific expertise, and drawing upon advances in unsupervised manifold learning[19]. Recent efforts to generate scientific hypotheses rely heavily on scientific literature, but ignore equally available publication meta-data. By programmatically incorporating information on the evolving distribution of scientific expertise, our approach balances exploitation and exploration in experimental search that enables us to both (1) accelerate discoveries predicted to appear in the future and (2) punctuate advance by identifying promising experiments unlikely to be pursued without intervention.

Accounting for Human Experts in Machine Prediction

The distribution of research experts across topics and time represents a critical social fact that will stably improve our inference about whether surrounding facts have been tried and abandoned—and should be treated as negative knowledge—or remain available for profitable hypothesis generation[20]. First we do this alongside precise replication of a recent analysis in *Nature* that predicted materials having desirable electrochemical properties from prior literature encoded with unsupervised neural network methods[1], but ignorant of the distribution of human expertise. We show that by simply adding information about the location of scientists and their likely inferences, using a formally identical approach, we dramatically (~100%) improve predictions of future materials. Next, we extended this approach to identify a much broader matrix of materials and their functional properties[21], including drugs and vaccines. Finally, we use expert awareness to identify and validate the scientific and technological promise of research avenues unlikely to be explored by human experts unaided.

Specifically, we model the distribution of inferences cognitively available to scientists by constructing a hypergraph over research publications. A hypergraph is a generalized graph where an edge connects a set, rather than a pair, of nodes. Our research hypergraph is mixed, containing nodes corresponding not only to materials and

properties mentioned in title or abstract, but also the researchers who investigate them (Fig. 1b). Random walks over this hypergraph suggest paths of inference cognitively available to active scientists. If a valuable material property (e.g., ferroelectricity—reversible electric polarization useful in sensors) is investigated by a scientist who, in prior research, worked with lead titanate ($PbTiO_3$, a ferroelectric material), that scientist is more likely to consider whether lead titanate is ferroelectric than a scientist without the research experience. If that scientist coauthors with another who has previously worked with sodium nitrite ($NaNO_2$, also a ferroelectric material), that scientist is more likely to consider that sodium nitrite may have the property through conversation than a scientist without the personal connection. The density of the distribution of random walks over this research hypergraph will be proportional to the density of cognitively plausible inferences. If two literatures share no scientists, a random walk over our hypergraph will rarely bridge them, just as a scientist will rarely consider connecting a property valued in one with a material understood in another (Fig. 1a).

Our model (1) initiates a random walk over the research hypergraph with a valued property (e.g., ferroelectricity), then (2) randomly selects an article (hyperedge) with that property, then (3) randomly select a material or author from that article, then (4) randomly selects another article with that material or author, etc., following a Markov process[22,23]. Such a random walk induces similarity metrics that capture the relevance of nodes to one another. The first metric we use draws upon the local hypergraph structure to estimate the probability a random walker travels from one node to another in a fixed number of steps (see Supplementary Information). Our second metric is based on a popular, unsupervised neural network-based embedding algorithm (`deepwalk`[24]) over the generated random walks. This method is formally identical to the word embedding method used in replicated prior work that ignores the distribution of scientists[1], but which we apply to our hypergraph, considering every random walk sequence as a "sentence" linking materials, experts and functional properties (e.g., store energy; cure breast cancer, vaccinate against COVID-19). The resulting embedding maps every node to a numerical vector, with the dot-product between any pair reflecting the relatedness of corresponding nodes. We also created a comparable embedding space using deeper graph convolutional neural networks that did not change the pattern of results presented here (see Methods and Supplementary Information).

Accelerating science by predicting future discoveries

Pairwise relevances estimated across our mixed hypergraph reveal distinct phenomena. The relevance of a material to a scientist measures the likelihood that she is or will become familiar with that concept through research experience, related reading, or conversation. The co-relevance of materials suggests that they may be substitutes or complements within the same experiment. The relevance of a material to a property suggests both the likelihood that the material may possess the property, but also that a scientist will likely discover and publish it (Extended Data, Fig. 1a, 1b). In this way, our hypergraph-induced similarities incorporate physical and material properties latent within literature, but also the complementary distribution of scientists, enabling us to anticipate likely inferences and predict upcoming discoveries. We assessed the pool of materials available to scientists in the literature published prior to the prediction year, ranked materials in terms of their discovery likelihood based on transition probabilities and unsupervised embeddings, then compared those rankings with actual first-time published linkages between materials and properties in published research (see Methods for further details).

Energy-related Materials Prediction. To demonstrate the power of accounting for human experts, we considered the valuable electrochemical properties of thermoelectricity, ferroelectricity and photovoltaic capacity against a pool of 100K candidate compounds, contrasting our predictions with replicated prior work that did not account for human expertise[1]. We repeated identical analyses for 17 prediction periods, with prediction years ranging from 2001 to 2017, predicting future discoveries as a function of research publicly available to contemporary scientists. We computed annual precisions until the end of 2018, such that the longest precision array was nearly two decades (18 years, from 2001 to 2018) and the shortest was 2 (2017-2018, Extended Data,

Fig. 1c). Replicating the evaluation method of Tshitoyan et al. on the same dataset (1.5M articles about inorganic materials)[1], predictions that account for the distribution of materials scientists outperformed baselines for all properties and materials by an average of 100% (Fig. 2b-d).

Drug Repurposing. We used the same approach to explore the repurposing of ~4K existing FDA approved drugs to treat 100 critical human diseases. We used the MEDLINE database of biomedical research publications and set the prediction year to 2001. Ground-truth discoveries were based on drug-disease associations established by expert curators of the Comparative Toxicogenomics Database (CTD)[25], which chronicles the capacity of chemicals to influence human health. Figure 1a reports prediction precisions 19 years after the prediction year, revealing how accounting for the distribution of biomedical experts in our unsupervised hypergraph embedding yields predictions with 43% higher precision than identical models accounting for article content alone. Moreover, we found a strong correlation between prediction precision and drug occurrence frequency in literature ($r=0.74$, $p<0.001$), implying that our predictors work best for diseases with relevant drugs mentioned frequently in prior research.

COVID-19 Therapy and Vaccine Prediction. Finally, we considered therapies and vaccines to treat or prevent SARS-CoV-2 infection. Here prediction year was set to 2020, when the global search for relevant drugs and vaccines began in earnest. Following Gysi et al.[26], we considered a therapy relevant to COVID-19 if it amassed evidence to merit a COVID-related clinical trial, as reported by ClinicalTrials.gov. Results shown in Figure 1e indicate that 36% and 38% of the predictions made by deepwalk-based and transition probability metrics entered trials within 12 months of the date of prediction, respectively, 350 to 400% higher than the precision of discovery candidates generated by semantic content alone (10%). These precisions were even higher than a predictive model based on an ensemble of deep and shallow learning predictors trained on multiply measured protein interactions between COVID-19 and the pool of 3,948 relevant compounds from DrugBank[26], information to which our model was blind (see Extended Data, Fig. 2 for alternative measurement).

The success of these COVID-19 predictions suggests how fast-paced research on COVID therapies and vaccines increased the relevance of scientists' prior research experiences and relationships for the therapies and vaccines they would come to imagine, evaluate and champion in clinical trials. Consider the clinical trial of the female progesterone for treating COVID-19[27]. The trial was motivated by factors including the lower global death rate of women than men from COVID-19 and anti-inflammatory properties of progesterone that may moderate the immune system's overreaction to COVID-19 in men[28]. Random walks from our method frequently walked the path between "coronavirus" and "progesterone" literatures to predict clinical study of progesterone for coronavirus complications (Extended Data, Fig. 5). Our technique traced a pathway similar to the one articulated by researchers sponsoring the trial: 75% of trial-cited papers, published within the five-year period we considered in building our hypergraph (2015-2019), were identified and used by our prediction model, and 60% of scientists authoring those studies were sampled in our random walk sequences.

Expert-Sensitive Prediction

Our predictive models use the distribution of discovering experts to successfully improve discovery predictions. This is demonstrated by time to discovery, which is inversely proportional to the size of the expert population who studied both property and material in their research. If we define *expert density* between a property and material as the Jaccard index of experts who mentioned both in recent publications, higher densities suggest the two are cognitively available to more scientists, and that their underlying relationship (if any) is more likely to be investigated earlier. For materials, COVID-19 therapies and vaccines, and a majority of the 100 diseases we considered, correlations between discovery date and expert densities were negative, significant and substantial (Extended Data, Fig. 3) showing that materials considered by experts familiar with a property are discovered

sooner. Our predictive models efficiently incorporate these expert densities (Extended Data, Fig. 4). Similar results can be derived based on embedding proximities: Fig. 3 (top row) illustrates how our predictions cluster atop density peaks in a joint embedding space of experts and the materials they investigate. These expert-material proximities predict discoverers most likely to publish discoveries based on their unique research backgrounds and relationships. Computing the probability of transition from properties to experts through a single intermediate material across 17 prediction years (2001 to 2017), we found that 40% of the top 50 ranked potential discoverers became discoverers of thermoelectric and ferroelectric materials one year after prediction, and 20% of the top 50 discovered novel photovoltaics (Fig. 3, bottom; see also Extended Data, Fig. 6)

Punctuating science by predicting unlikely discoveries

As illustrated above, by identifying properties and materials cognitively available to human experts, we maximize the precision of predicting published material discoveries. Almost all published discoveries lie in close proximity to desired properties based on hypergraph induced from prior literature (Fig. 4a). By contrast, if we avoid the distribution of human experts, we can produce in-human, "alien" predictions designed to complement the scientific community. These predictions are cognitively unavailable to human experts based on the organization of scientific fields, prevailing scientific attention, and expert education, but nevertheless manifest strong mechanistic promise for possessing desired scientific properties (Fig. 4b). Here, we propose a generic framework for identifying disruptive discovery candidates expected to possess desired properties, but least likely to be studied by human scientists or discovered in the near future without machine recommendation (Fig. 1a, right).

Our framework combines two components: an alien component that measures the degree to which candidate materials are beyond the scope of human experts' research experiences and relationships, and a second that rules out those predicted scientifically irrelevant (Fig. 4c). Each component scores entities based on *human availability* and *scientific plausibility*. The two scores are then combined with a simple mixing coefficient β. Setting β=0 implies full emphasis on scientific plausibility, blind to the distribution of experts. Decreasing β imitates human experts and increasing β avoids them. At extremes, β=-1 and 1 yield algorithms that generate predictions very familiar or very strange to experts, regardless of their scientific merit. Non-zero positive βs balance exploitation of relevant materials with exploration of areas unlikely considered or examined by human experts. Materials are ranked by their final scores s with highest-ranked items reported candidates for disruptive discovery. Human availability is assessed with any graph distance metric varying with expert density (e.g., unsupervised neural embeddings, Markov transition probabilities, self-avoiding walks from Schramm-Loewner evolutions). Scientific merit is quantified through theory-driven simulation of material properties. For thermoelectricity, power factor (PF) represents an important component of the overall thermoelectric figure of merit, zT, calculated using density functional theory for candidate materials as a strong indication of thermoelectricity[29,30]. For COVID-19, proximity between SARS-CoV-2 and candidate compounds in protein-protein interaction networks suggests the likelihood a material will recognize and engage with the virus[26]. If theoretical predictions are unavailable, one may approximate scientific relevance with proximity in unsupervised literature embeddings[1].

Fig. 4d shows the results of running our hybrid model with different βs for thermoelectricity, and Extended Fig. 7 for COVID-19. In both, we normalize, rescale and linearly compose alienness derived from shortest-path distance, and scientific plausibility from word embedding proximity (see Methods). We reserve theory-driven indicators based on power factor and protein-protein network proximity to evaluate our predictions, rather than establish scientific plausibility as they would in a deployed system. Increasing β from zero to one, candidate materials were less likely to be conceived, discovered, and published, but PF and protein interaction likelihood remained strong for all but the most alien predictions. Intermediate βs resulted in a balanced trade-off with strong values of PF and protein-interaction even in distant and completely disconnected materials. This demonstrates the capability of our framework for punctuating scientific advance by proposing alien but scientifically promising candidate materials,

with only a naive combination and weighting system. More sophisticated metrics could be employed by incorporating all available prior scientific knowledge and learning combination metrics through self-supervised multi-headed graph convolutional neural networks.

Discussion

These models demonstrate the power of incorporating expert-awareness into artificial intelligence systems for accelerating and punctuating future discovery. Our models succeed by directly predicting human discoveries and the human experts who will make them, yielding an average of 100% improvement in prediction precision. By tuning these algorithms to avoid the crowd, however, they generate scientifically promising "alien" hypotheses unlikely to be imagined, pursued or published without machine recommendation. By identifying and correcting for collective patterns of human attention, formed by field boundaries and institutionalized education, these models complement the contemporary scientific community. A further class of alien predictions could be tuned to compensate not only for emergent bias, but universal cognitive constraints, such as limits on the human capacity to conceive or search through complex combinations (e.g., high-order drug cocktails[31]). Disorienting hypotheses from such a system will not be beautiful, but being inconceivable, they break unbroken ground and sidestep the path dependent "burden of knowledge" where scientific institutions require new advances built upon the old for ratification and support[32,33].

Our approach can also be used to identify individual and collective biases that limit productive exploration, and suggest opportunities to improve human prediction by reformulating science education for discovery. Insofar as research experiences and relationships condition the questions scientists investigate, education tuned to discovery would conceive of each student as a new experiment, recombining knowledge and opportunity in novel ways. Our investigation underscores the power of incorporating human and social factors to produce artificial intelligence that complements rather than substitutes for human expertise. By making AI hypothesis generation aware of human expertise, it can race with rather than against the scientific community to expand the scope of human imagination and discovery.

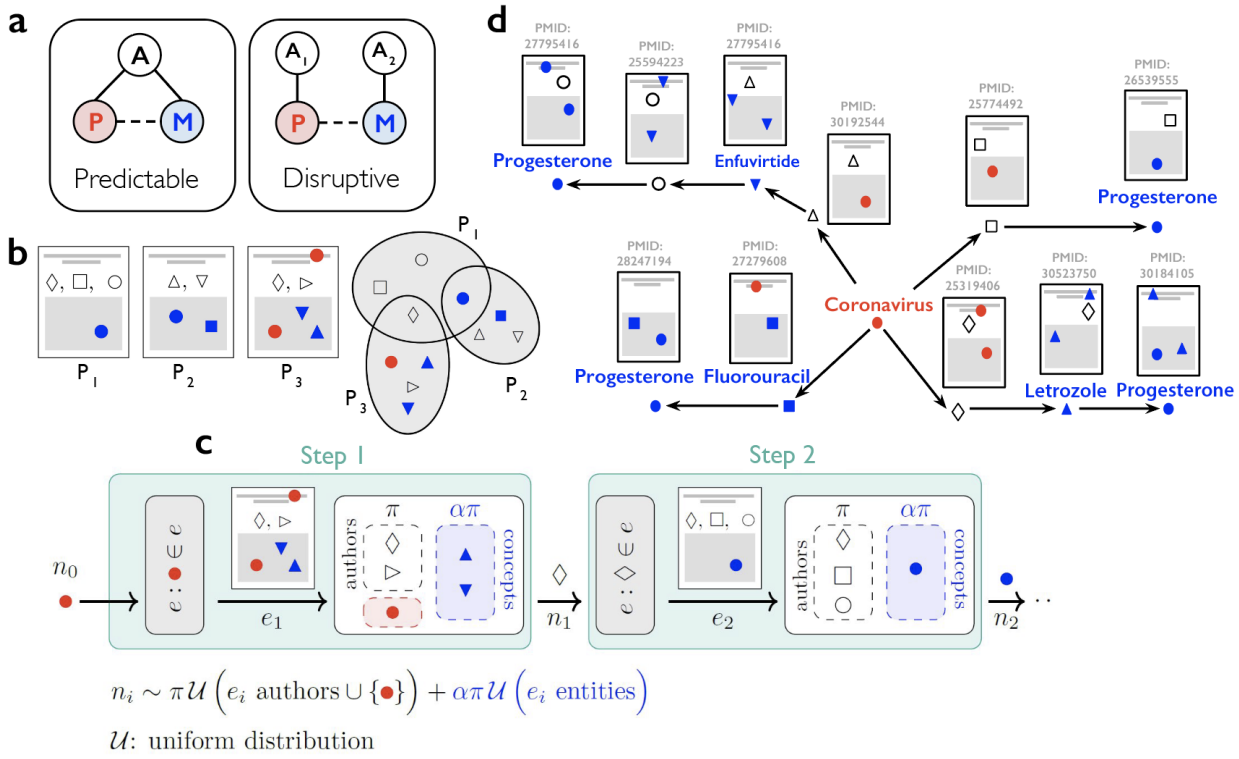


Fig. 1. (a) Two possible relations between experts, property and material nodes when there exists a hidden underlying relationship between the two (dashed line) to be discovered. Uncolored circles represent human experts and each colored node indicates a material (colored in blue denoted M) or a desirable property they possess (colored in red and denoted P). Solid lines show existing links between expert-material nodes and dashed lines represent existing property-material links that have not yet been discovered. The left case, where concepts P and M share a common collection of experts, is likely to be discovered and published in the near future—they are predictable by scientists, whereas the right case is likely to escape scientists' attention as there is no shared community of experts, and their pursuit would disrupt the current course of science. **(b)** Illustration of our mixed coauthorship hypergraph for three papers. Uncolored shapes represent authors and colored shapes represent properties (red) or materials (blue) mentioned in article titles and abstracts. These three papers constitute a hypergraph with three hyperedges traced by ellipses. **(c)** Two initial steps of a random walk process on the hypergraph shown in part (b). Blue and red shapes represent material and property keywords, respectively. Papers (hyperedges) are sampled uniformly whereas, if α is set, nodes are selected such that the probability of sampling an entity is α times the probability of sampling an author. Note that α is the only parameter of this non-uniform sampling (π can be uniquely determined from α). **(d)** Four example random walk paths starting from property "Coronavirus"-relevant and ending in Progesterone (a chemical under clinical trial investigation for therapeutic efficacy). Each arrow connecting two nodes indicates a sampling step, where the paper shown on top of the receiving node comprises a hyperedge containing that material and the property, author, or material from the prior step.

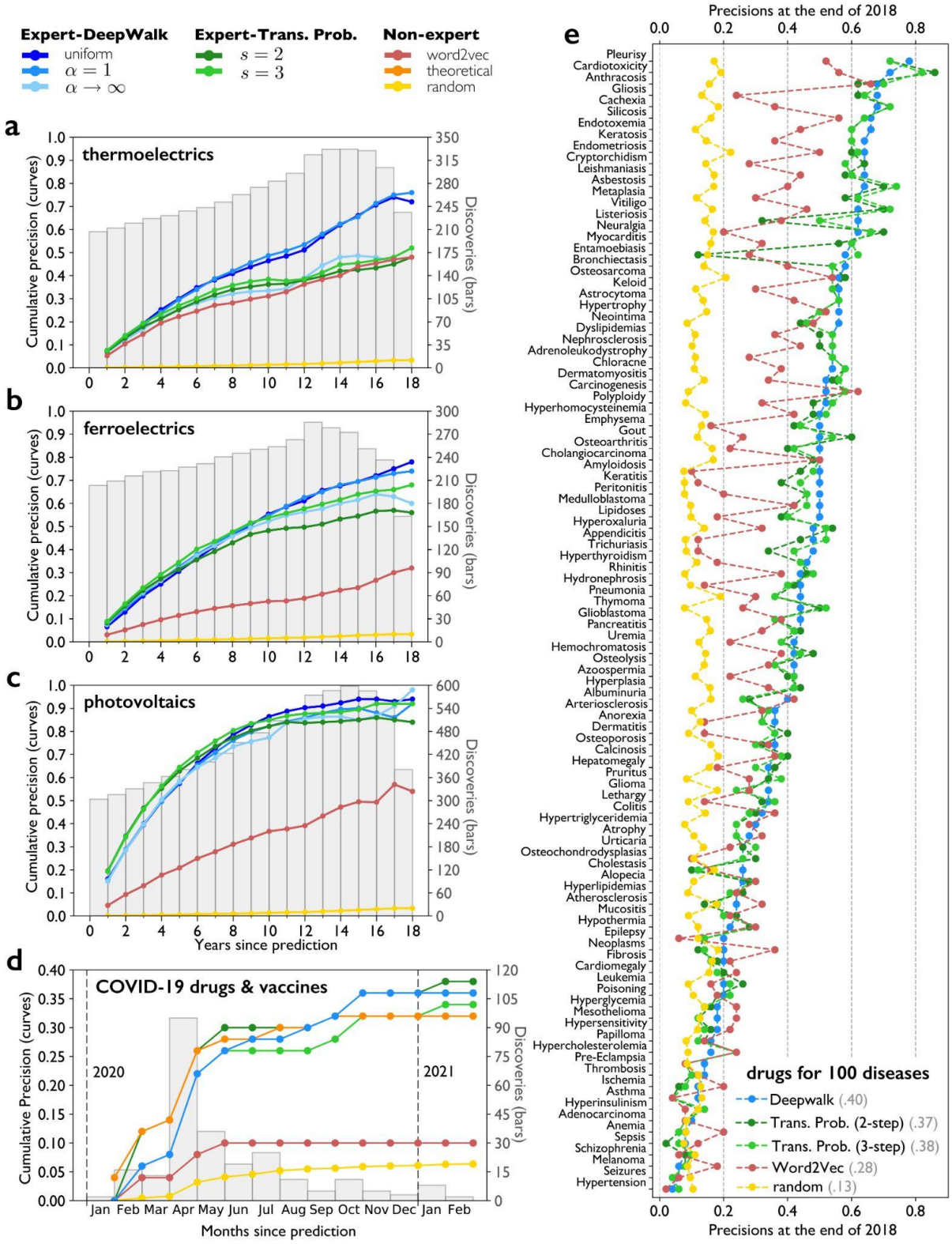

Expert-DeepWalk
uniform
$\alpha = 1$
$\alpha \to \infty$
Expert-Trans. Prob.
$s = 2$
$s = 3$
Non-expert
word2vec
theoretical
random
a
thermoelectrics
b
ferroelectrics
c
photovoltaics
Cumulative precision (curves)
Discoveries (bars)
Years since prediction
d
COVID-19 drugs & vaccines
2020
2021
Cumulative Precision (curves)
Discoveries (bars)
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec Jan Feb
Months since prediction
e
Precisions at the end of 2018
Pleurisy
Cardiotoxicity
Anthracosis
Gliosis
Cachexia
Silicosis
Endotoxemia
Keratosis
Endometriosis
Cryptorchidism
Leishmaniasis
Asbestosis
Metaplasia
Vitiligo
Listeriosis
Neuralgia
Myocarditis
Entamoebiasis
Bronchiectasis
Osteosarcoma
Keloid
Astrocytoma
Hypertrophy
Neointima
Dyslipidemias
Nephrosclerosis
Adrenoleukodystrophy
Chloracne
Dermatomyositis
Carcinogenesis
Polyploidy
Hyperhomocysteinemia
Emphysema
Gout
Osteoarthritis
Cholangiocarcinoma
Amyloidosis
Keratitis
Peritonitis
Medulloblastoma
Lipidoses
Hyperoxaluria
Appendicitis
Trichuriasis
Hyperthyroidism
Rhinitis
Hydronephrosis
Pneumonia
Thymoma
Glioblastoma
Pancreatitis
Uremia
Hemochromatosis
Osteolysis
Azoospermia
Hyperplasia
Albuminuria
Arteriosclerosis
Anorexia
Dermatitis
Osteoporosis
Calcinosis
Hepatomegaly
Pruritus
Glioma
Lethargy
Colitis
Hypertriglyceridemia
Atrophy
Urticaria
Osteochondrodysplasias
Cholestasis
Alopecia
Hyperlipidemias
Atherosclerosis
Mucositis
Hypothermia
Epilepsy
Neoplasms
Fibrosis
Cardiomegaly
Leukemia
Poisoning
Hyperglycemia
Mesothelioma
Hypersensitivity
Papilloma
Hypercholesterolemia
Pre-Eclampsia
Thrombosis
Ischemia
Asthma
Hyperinsulinism
Adenocarcinoma
Anemia
Sepsis
Schizophrenia
Melanoma
Seizures
Hypertension
drugs for 100 diseases
Deepwalk (.40)
Trans. Prob. (2-step) (.37)
Trans. Prob. (3-step) (.38)
Word2Vec (.28)
random (.13)
Precisions at the end of 2018

Fig. 2. Precision rates of discovery for materials associated with different properties and prediction years: **(a-c)** chemical compounds and electrochemical properties including thermoelectricity, ferroelectricity, and photovoltaic capacity, respectively, with prediction years varying from 2001 to 2017; **(d)** therapeutics and vaccines for COVID-19 for the prediction year 2020; **(e)** general disease-drug associations for prediction year 2001. Precisions reported for general disease-drug associations are individual rates computed 19 years after prediction year, but computed annually for electrochemical properties and monthly for COVID-19 efficacy. Gray bars in Figs. (a-d) indicate the number of new discoveries occurring in reality for each month or year of the prediction period.

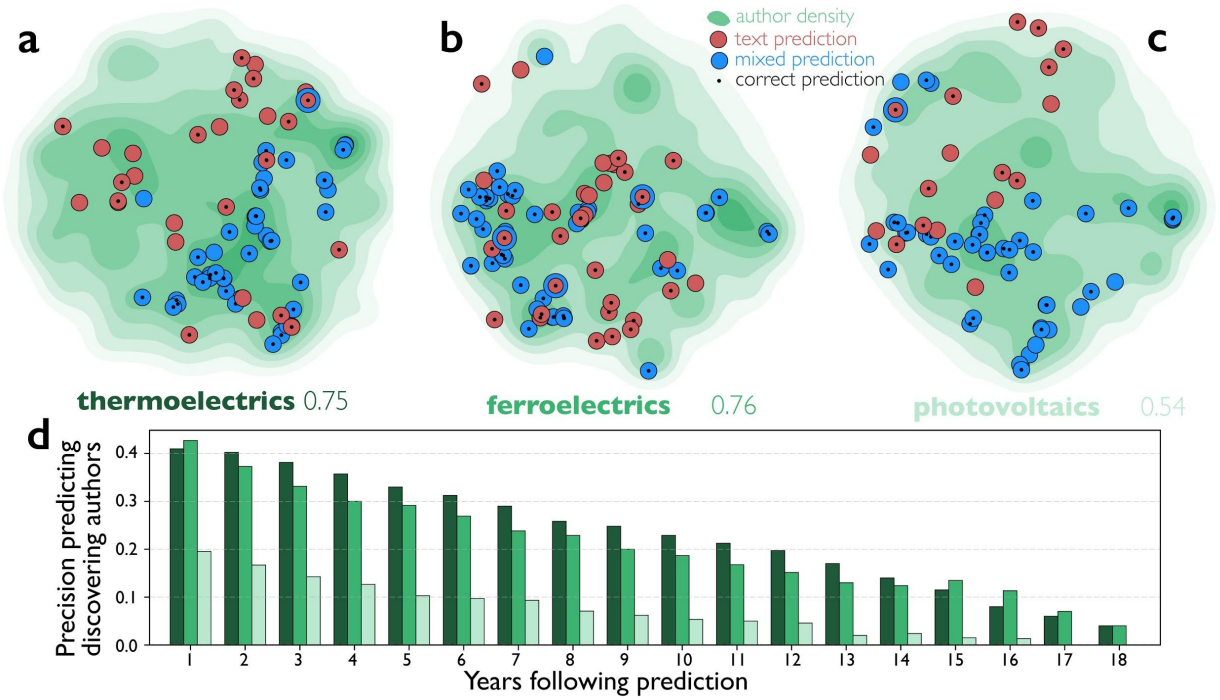


Fig. 3. (a) 2D projections of the expert-sensitive material predictions made by deepwalk (blue circles) and the content-exclusive word2vec model (red circles) for thermoelectricity (left), ferroelectricity (center) and photovoltaic capacity (right). Circles with center dots indicate true positive predictions discovered and published in subsequent years and empty circles are false positives, yet unpublished. Predictions are plotted atop the density of experts (topo map and contours estimated by Kernel Density Estimation) in a 2D tSNE-projected embedding space. Before applying tSNE dimensionality reduction, the original embedding was obtained by training a word2vec model over sampled random walks across the hypergraph of published science. Red circles are more uniformly distributed, but blue circles concentrate near peaks of expert density. **(b)** Precision rates for predicting discoverers of materials with electrochemical properties. Predictive models are built based on two-step transitions between property and expert nodes with an intermediate material in the transition path. Bars show average precision of expert predictions for individual years. An expert can publish a discovery in multiple years. Total precision rates are also shown near each property ignoring the repetition of discovering experts.

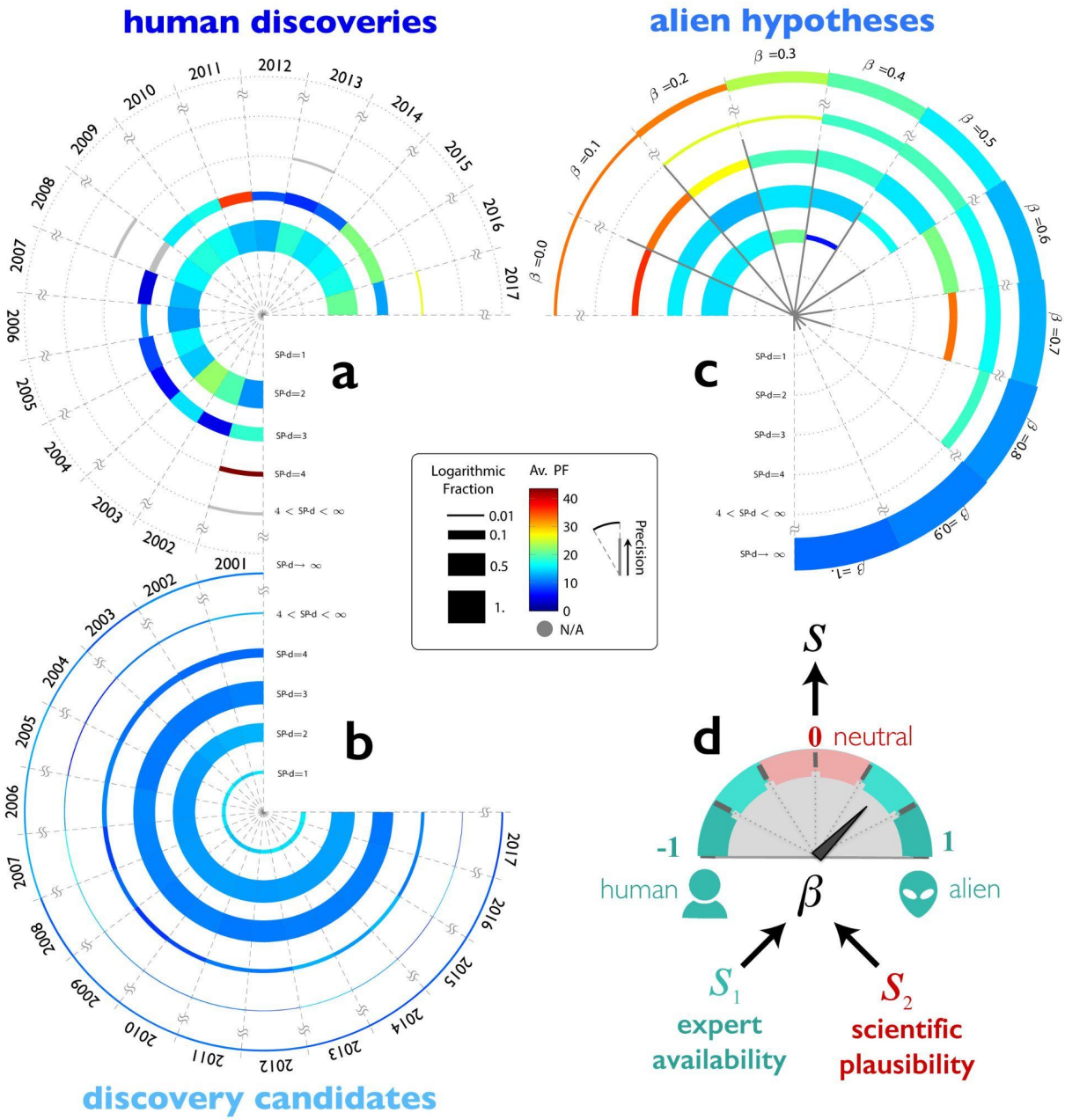


Fig. 4. Distribution of Power Factor (PF) as a simulated thermoelectricity score visualized on shortest-path distance (SP-*d*) levels from node thermoelectricity for **(a)** human discoveries, **(b)** all materials that have not been studied in context of thermoelectricity yet and are candidates of being discovered, and **(c)** hypotheses from our alien AI algorithm with parameter β varying from 0 to 1, where 50 hypotheses were generated with each value of β in the prediction year 2001 (evaluations performed after 18 years), as illustrated in **(d)**. The property node is located at the center; each concentric orbit represents a particular (range of) SP-*d*s, where the last orbit includes materials disconnected from the property (∞ SP-*d*). The size and color of each arc show the total number of compounds with corresponding SP-*d* from the property and their average PF scores, respectively. The further a compound is located from the property, the less cognitively available it is to scientists. **(a)** shows that human discoveries mainly lie in close proximity to the property node (highly cognitively available). Nevertheless, **(b)** shows that candidates of discoveries in each year are distributed more broadly across the network of scientists. Moreover, there exist materials with strong PF scores in distant orbits including the last one that is completely disconnected from the property. **(c)** indicates that our alien AI algorithm could capture cognitively unavailable hypotheses that are also evaluated to be scientifically plausible (strong PF values) except for very high β values. **(d)** Illustration of our general alien AI framework as a weighted combination of human (un)availability and scientific plausibility scores *s*. Combining these scores will result in a final ranking of materials

from which candidate hypotheses will be chosen. The mixing Coefficient β (varying in range [-1,1]) determines how much weight we give to the availability of hypotheses to human scientists. Setting β=0 implies full attention to scientific plausibility with no emphasis on human (un)availability. At the extremes, β=-1 and 1 sets the objective to be solely imitating and/or avoiding the expert distribution, respectively.

Methods

Experiments and Data Collection

Given a specific property and a pool of materials, each discovery prediction experiment consists of computing a set of scores based on the literature prior to the prediction year and selecting 50 materials with the highest scores. Precision of predictions could then be computed against ground-truth discoveries in subsequent months or years.

We collected several corpora of scientific articles and considered relationships between materials and various properties. Forming a mixed hypergraph requires a disambiguated set of authors for all scientific articles. Our testbed consisted of two datasets: a collection of ~1.5M articles published between 1937 and 2018 classified by Tshitoyan et. al (2019) relating to inorganic materials[1], and the MEDLINE database that includes more than 28M articles published in various biomedical fields over the span of more than two centuries. We downloaded the former using Scopus API provided by Elsevier (https://dev.elsevier.com/), which readily assigns unique codes to distinct authors. In order to author-disambiguate PubMed database, we used disambiguation results provided by PubMed Knowledge Graph (PKG)[34], which were obtained by combining information from the Author-ity disambiguation of PubMed[35] and the more recent semantic scholar database[36].

For energy-related materials science, we extracted the pool of materials from the collected 1.5M articles using Python Materials Genomics[37] and direct rule-based string processing. Material-property association was considered to be established if the material co-occurred with any of the property-related keywords. First-time co-occurrences were defined as ground-truth discoveries, following relevant prior work[1]. For the case of drug repurposing, we began with a pool of 7,800 approved candidate drugs downloaded from the DrugBank database. We then built our drug pool using approximately 4,000 drugs possessing simple names (e.g., by dropping complex names containing several numerical parts). We chose 100 diseases from the Comparative Toxicogenomics Database (CTD)[25] that had the largest number of relevant drugs from our drug pool. We searched for names of drugs and diseases in MEDLINE to detect their occurrence within papers to build a hypergraph. Ground-truth relevant drugs for the selected diseases were extracted from the associations curated by CTD. The discovery date for each of the disease-drug associations was set to the earliest publication reported by CTD for the curated or inferred relevance. We ran separate prediction experiments for each disease. The same pool of drugs and corpus of papers were used in case of COVID-19, where their relevance to COVID-19 were identified based on their involvement in COVID-related studies reported by ClinicalTrials.org in or after 2020, regardless of the studies' results. Date of discovery for each relevance was set to the date that the corresponding study was first posted, and if the drug was involved in multiple trials we considered the earliest date. There have been 4,899 trials posted as of March 3rd, 2021 (ignoring 32 trials dated before 2020), which included 251 drugs from our pool (~6%) included in their designs.

Random Walks and Relevance Metrics

In practice, coauthorships that occurred long before the time of prediction will neither be cognitively available nor perceived as continuingly relevant. Therefore, we restrict our prediction experiments to use literature produced in the 5 years prior to year of prediction. For each property, we took 250,000 non-lazy, truncated random walk sequences starting from the property node and terminating after 20 steps or after reaching a deadend node with no further connections. Without constraining the space, the majority of hypergraph nodes belong to experts—there are more authors on the average article than materials studied within it. We devised a biased random walk

algorithm to compensate for this imbalance, controlled by a parameter α, which defines the probability ratio of selecting conceptual (e.g., molecules or materials) to author nodes in any given paper. Larger α results in the higher frequency of selecting conceptual nodes and α=1 implies a balanced mixture of authors and entities (Fig. 1c, also see Methods). Note that deepwalk similarity is much more global than transition probability, provided the length of our walks (~20) are much longer than the transition steps considered (2-3), and it is more flexible as the walker's edge selection probability distribution can be easily modified to explore the network structure more deeply[38]. Note that authors heavily outnumbered materials in all our databases. To mitigate this imbalance, we introduced a non-uniform node sampling distribution parameterized by α, defined as the ratio of the probability of sampling a material or property to the probability of sampling an author in any given paper. A random walker with α=1 tends to select roughly equal number of authors and materials. In practice, we sampled from a mixture of two uniform distributions with weights $1/(1+\alpha)|A|$ and $\alpha/(1+\alpha)|M|$ assigned to authors in set A and materials/property in set M, respectively, where $|A|$ denotes the cardinality of set A.

Multistep transition probabilities are directly computed from transition matrices using Bayesian rules and Markovian assumptions (Supplementary Information). For deepwalk representation, we trained a skipgram Word2Vec model with embedding dimensionality of 200 over the truncated random walk sequences. In the task of discovery prediction, we discarded author nodes from the generated random walk sequences and training was performed over property/material tokens only. The training hyperparameters here were set equal to the ones used when training the Word2Vec baseline model, i.e., window size of 8, negative sampling size of 15 and learning rate of 0.01, which linearly decayed to 0.001 during iterations. The only exception is the number of epochs, which was 30 for baseline and 5 for the network representation. The size of the vocabularies produced in deepwalk sentences had much smaller tokens than the baselines, as a result they required less effort to capture inter-node relationships.

We also ran our prediction experiments after replacing deepwalk representation with a graph convolutional neural network. We used Graph Sample and Aggregate (GraphSAGE) model[39] with 400 and 200 as the dimensionality of hidden and output layers with Rectified Linear Units (ReLU) as the non-linear activation in the network. Convolutional models require feature vectors for all nodes but our hypergraph is inherently feature-less. Therefore, we utilized the word embeddings obtained by our Word2Vec baseline as feature vectors for materials and property nodes. A graph auto-encoder was then built using GraphSAGE architecture as the encoder and an inner-product decoder and its parameters were tuned by minimizing the unsupervised link-prediction loss function[40]. We took the output of the encoder as the embedded vectors and selected the top 50 discovery candidates by choosing entities with the highest cosine similarities to the property node. In order to evaluate the importance of the distribution of experts for our prediction power, we trained this model on our full hypergraph and also after withdrawing the author nodes (see Supplementary Information). Running the convolutional model on energy-related materials and properties yielded 62%, 58% and 74% precisions on the full graph, and 48%, 50% and 58% on the author-less graph for thermoelectricity, ferroelectricity and photovoltaics, respectively. These results show a similar pattern to those obtained from deepwalk although with a somewhat smaller margin, likely due to the use of Word2Vec-based feature vectors, which limit the domain of exploration by the new embedding model to within proximity of the baseline.

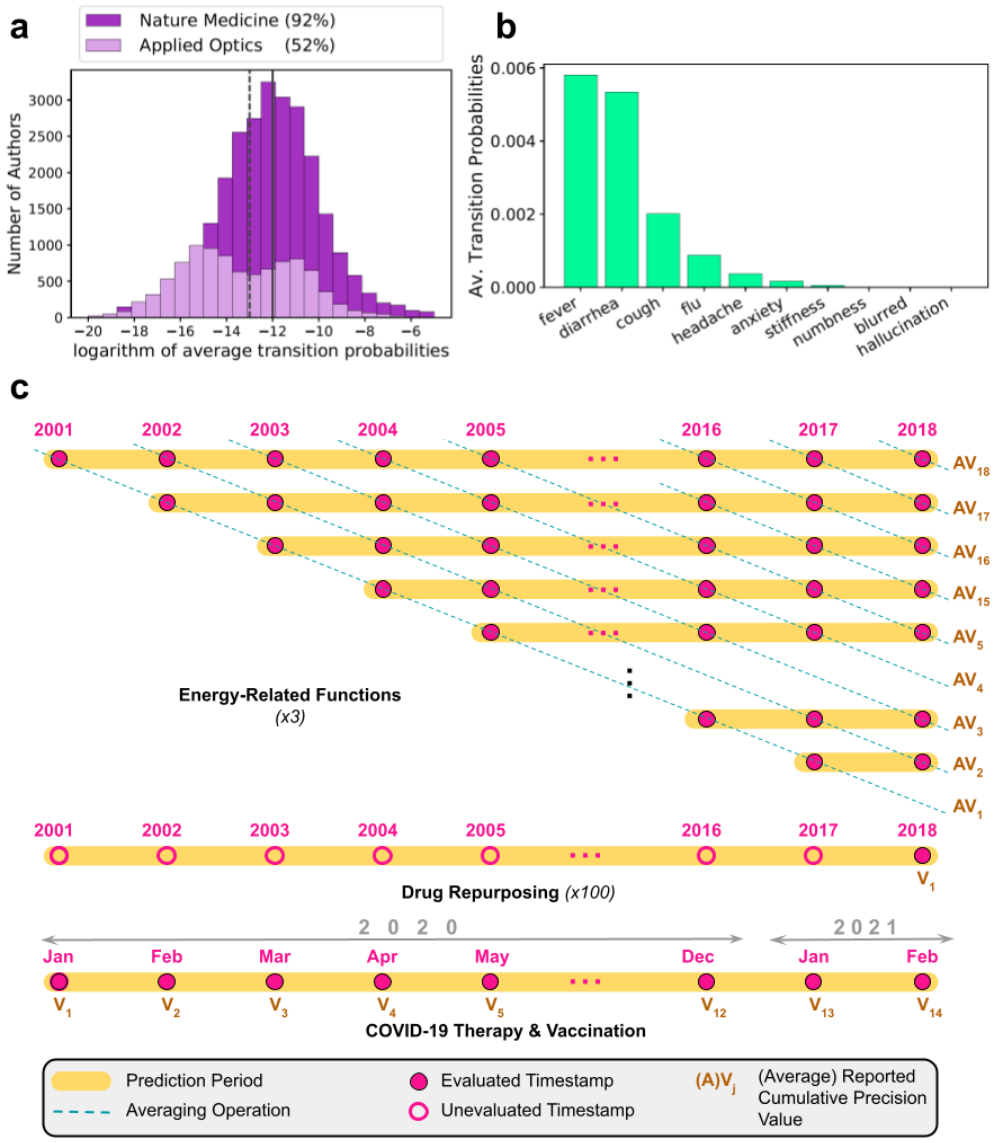
Alien Artificial Intelligence

Our alien knowledge discovery machine assigns human availability (or alienness) and scientific plausbility scores to each material with respect to a given property, which will be combined with a mixture weight β. In our AAI experiments, human unavailability was measured through shortest-path distance (SP-d) to the property node and scientific relevance was quantified by semantic similarities based on word embedding models (e.g., word2vec). The latter often yields continuous scores distributed similar to a Gaussian variable, but the former offers

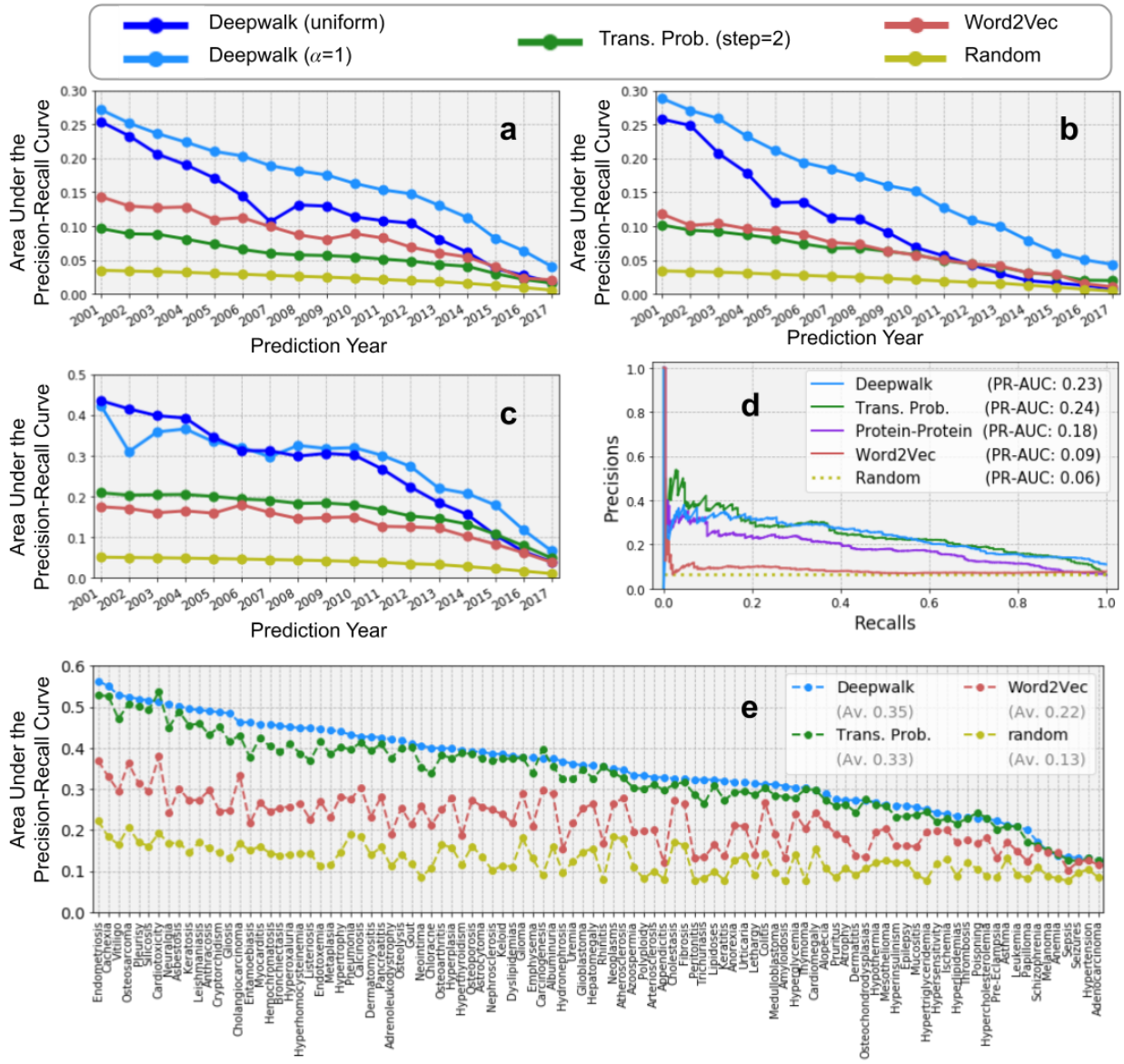
unbounded ordinal scores. This prevents us from directly combining them through Z-scores. To address this issue, we first transformed the two variables according to the Van der Waerden formulation[41] before taking the weighted average of their Z-scores (see Supplementary Materials for comparison to other combination methods). When evaluating AAI, we leveraged the property's theoretical scores obtained based on or prior knowledge from the relevant fields to assess scientific validation of candidates.

For thermoelectricity, we used Power Factor (PF) as a scalar score indicating how likely a material is thermoelectric based on theoretical simulations. PF is proportional to the electric conductivity and the absolute temperature and plays a key role in the more general metric of figure of merit (zT). Moreover, Tshitoyan et al. showed that materials that have been studied in conjunction with thermoelectricity in the literature tend to have higher PF scores. In our AAI experiments, we restricted the pool of entities to those for which there existed pre-calculated scores in the same database that Tshitoyan et al. had used[1,30], which formed 30% of the unstudied materials in the corresponding five-year period (1996 to 2000).

The theoretical scores that we used for evaluating our AAI method in case of COVID-19 were based on protein-protein interaction of the drugs with the SARS-CoV-2 viral target. Recently, Gysi et al. showed that existing drugs whose target proteins are within or in vicinity of the COVID-19 disease module are potentially strong candidates for repurposing[26]. They employed 12 network-based strategies individually and collectively to identify the most relevant candidate drugs, for which their rank-based combination yielded the best performance. We utilized the inverse of the aggregated ranks from their ensemble strategy as scores to theoretically measure material relevance to COVID-19. These scores were based on our prior knowledge of the target proteins associated with drugs and disease, to which our AAI method was blind.

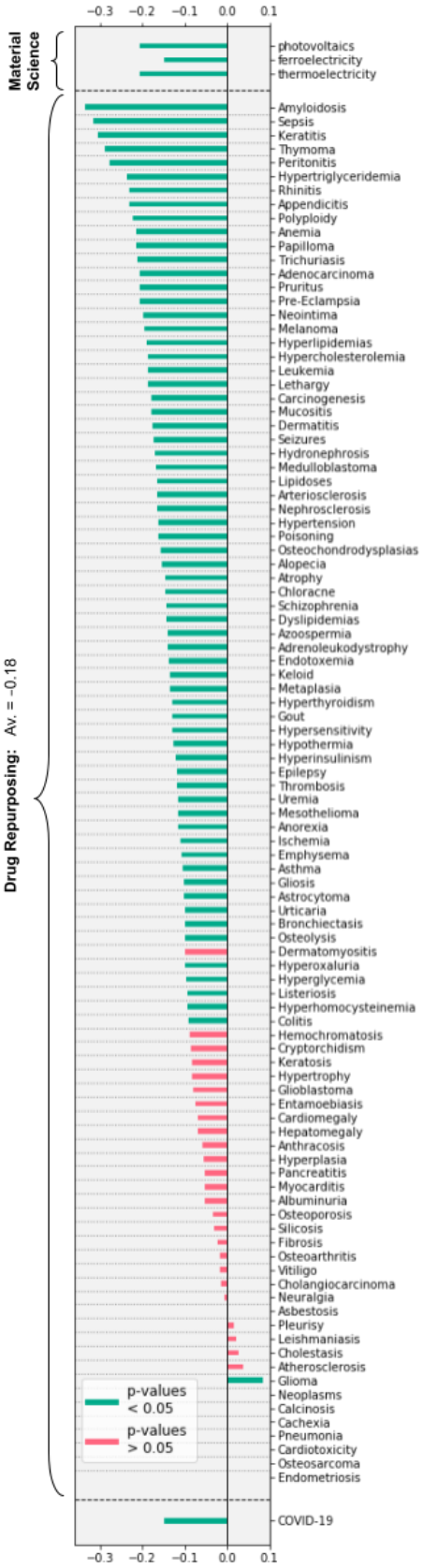


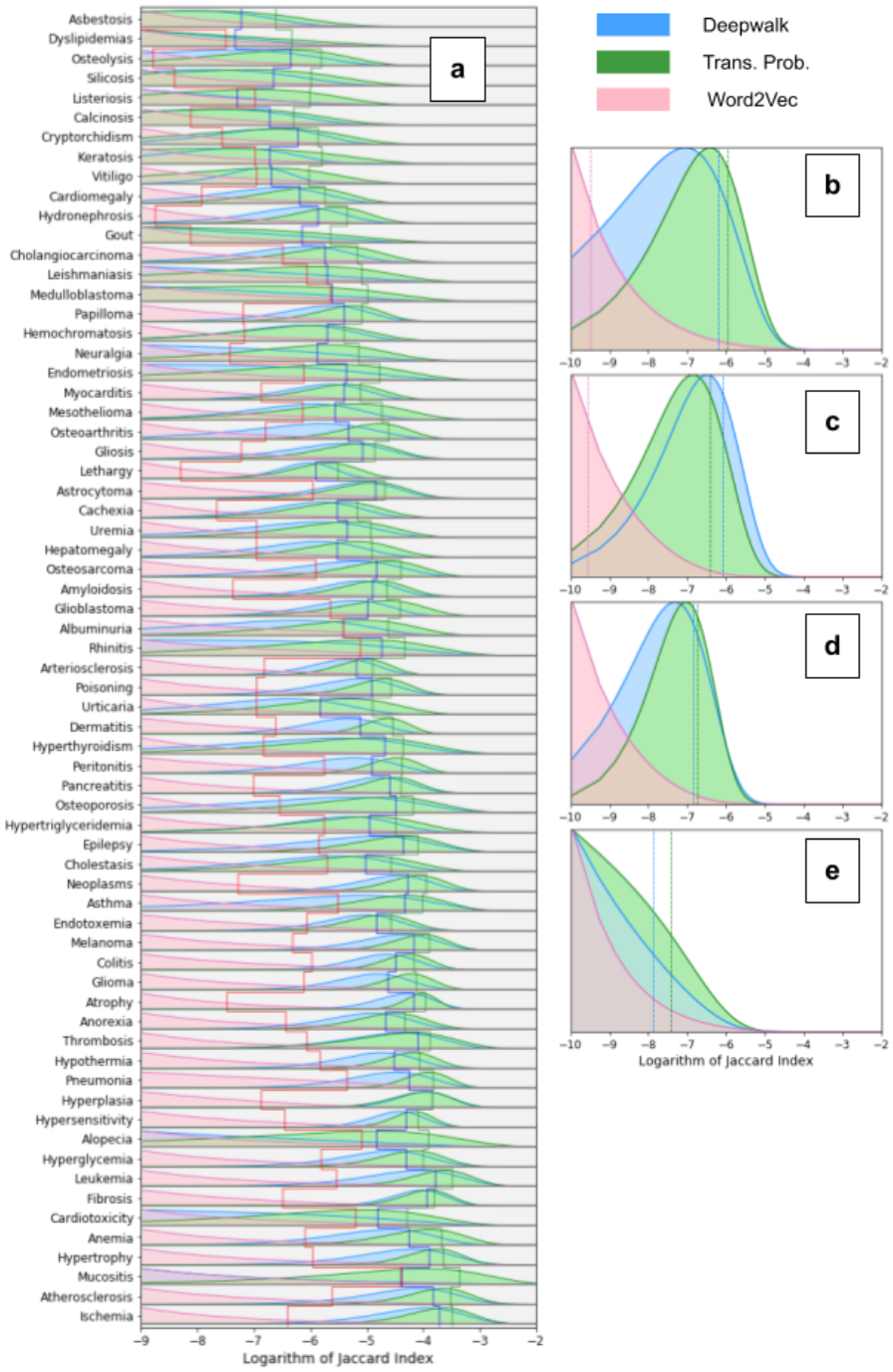
Extended Data Fig. 1. Sanity checks on our hypergraph-induced transition probability similarity metric. **(a) Between an author and a conceptual node:** Histogram of the similarities between nodes of two sets of authors and the node associated with the term "coronavirus". The two sets of authors are defined as authors of 5,000 randomly selected papers from journals Nature Medicine (red) and Applied Optics (turquoise) between 1990 and 2019. We computed similarities between the hypernodes as the logarithm of the average transition probabilities with one and two random walk steps. The histograms are plotted considering only non-zero transition probabilities: 92% of the authors of *Nature Medicine* (28,396 in total) and 51% of the selected *Applied Optics* authors (18,530 in total) had non-zero similarity values. Also, the average non-zero similarities associated with *Nature Medicine* authors (red dashed line) is almost 5 times larger than that of *Applied Optics* authors (blue dashed line), implying that based on the hypergraph-induce similarity metric the authors publishing in *Nature Medicine* write papers more relevant to coronavirus in comparison to those publishing in *Applied Optics*. **(b) Between two conceptual nodes:** similarities between several conceptual keywords shown on the x-axis and the node corresponding to "coronavirus". Similarities between the hypernodes are computed as the average transition probabilities with one and two intermediate nodes. The terms and symptoms known to be more relevant to coronavirus have larger average transition probabilities. **(c) Schematic of our experimental settings:** Starting and ending dates of the experiments are shown. As illustrated, we repeated the prediction experiments with 17 different starting dates for energy-related functions. Each predictor would be evaluated through (1) 18 average cumulative precision values for energy-related properties, (2) a single precision value for each disease in drug repurposing application, and (3) 14 cumulative precision values for COVID-19 therapy and vaccination.



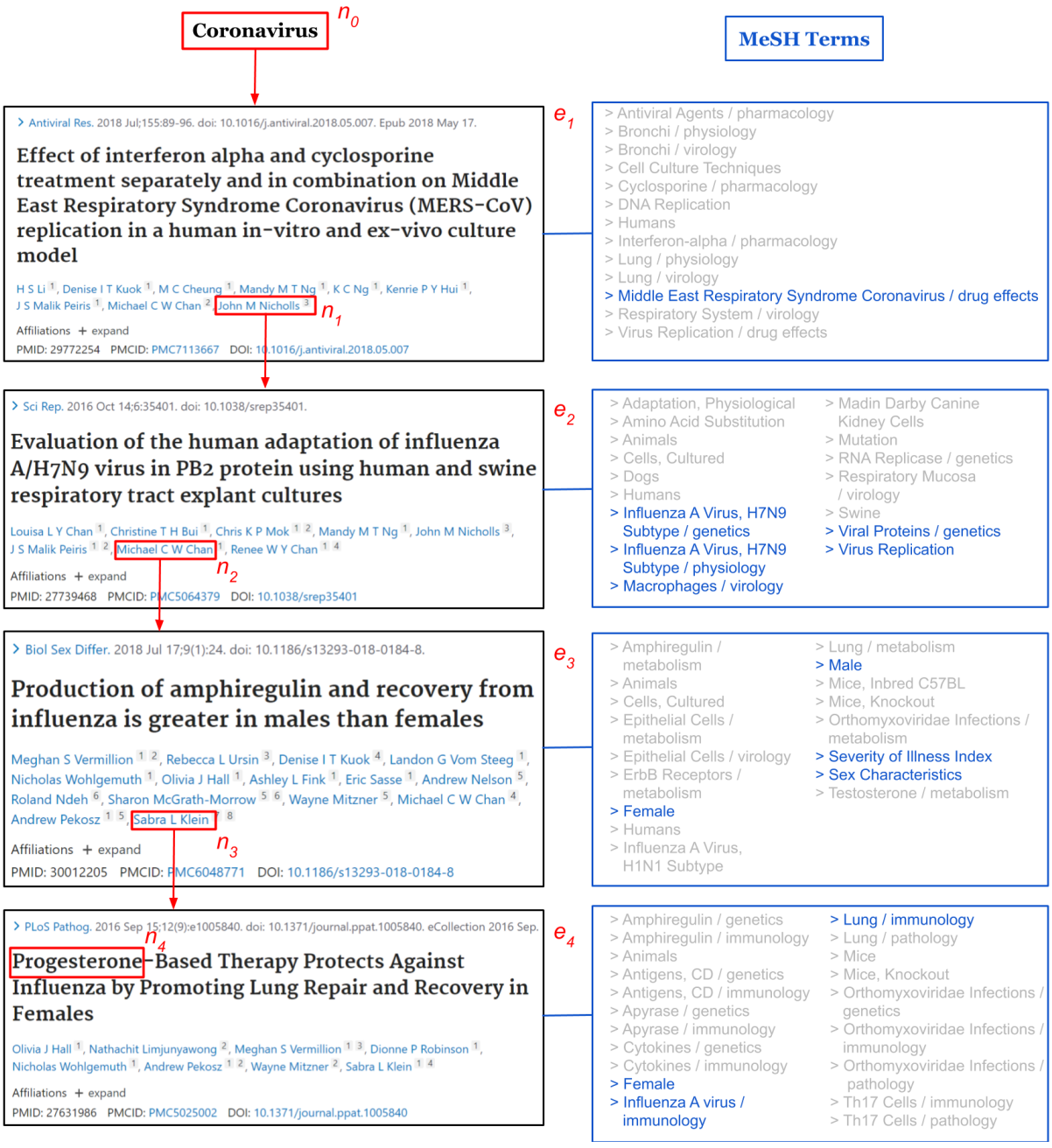
Extended Data Fig. 2. Precision-Recall (PR) curves and area under the curves for various predictors and databases: energy-related materials science properties, i.e., thermoelectrics (a), ferroelectrics (b) and photovoltaics (c), therapies and vaccines for COVID-19 (d), and generic drugs repurposing (e). Except for COVID-19, we only displayed the PR-AUC values for the selected prediction years skipping the PR curves themselves. Note that for Receiver Operating Curves (ROC) random predictors always result in AUC of 0.5, PR-AUC of the random baseline depends on the ratio of positive samples in the data set.

Extended Data Fig. 3. Spearman correlation coefficients of expert density (Jaccard index) between individual properties, the actual discoveries, and date of discovery. Negative correlations imply that entities with higher expert densities are likely to be discovered earlier than others. These results were obtained for discoveries after 2001 for energy-related properties and drugs repurposing applications, and after 2020 for COVID-19. The turquoise bars represent correlations with statistical significance (p-value<0.05) while the red bars had larger p-values indicating nonsignificant results. Moreover, for seven diseases in the CTD database all actual drugs repurposings, i.e., actual discoveries, occurred in a single year (we did not have reliable access to the month or day of discoveries from this database) and hence no correlation coefficients could be computed for them. The results indicate that the materials science properties and also COVID-19 showed strong negative correlations. In the case of CTD database, 67 out of 100 diseases (i.e., properties) showed statistically significant correlations, among which only one disease showed positive coefficient. The average of correlation coefficients across these 67 diseases was -0.18.

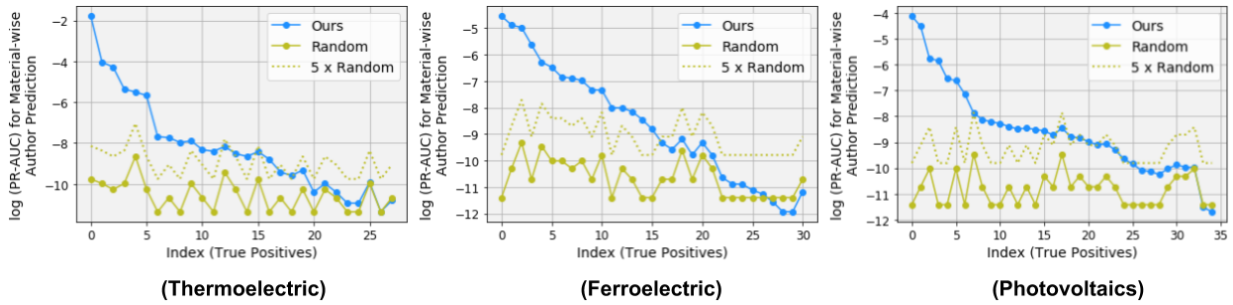




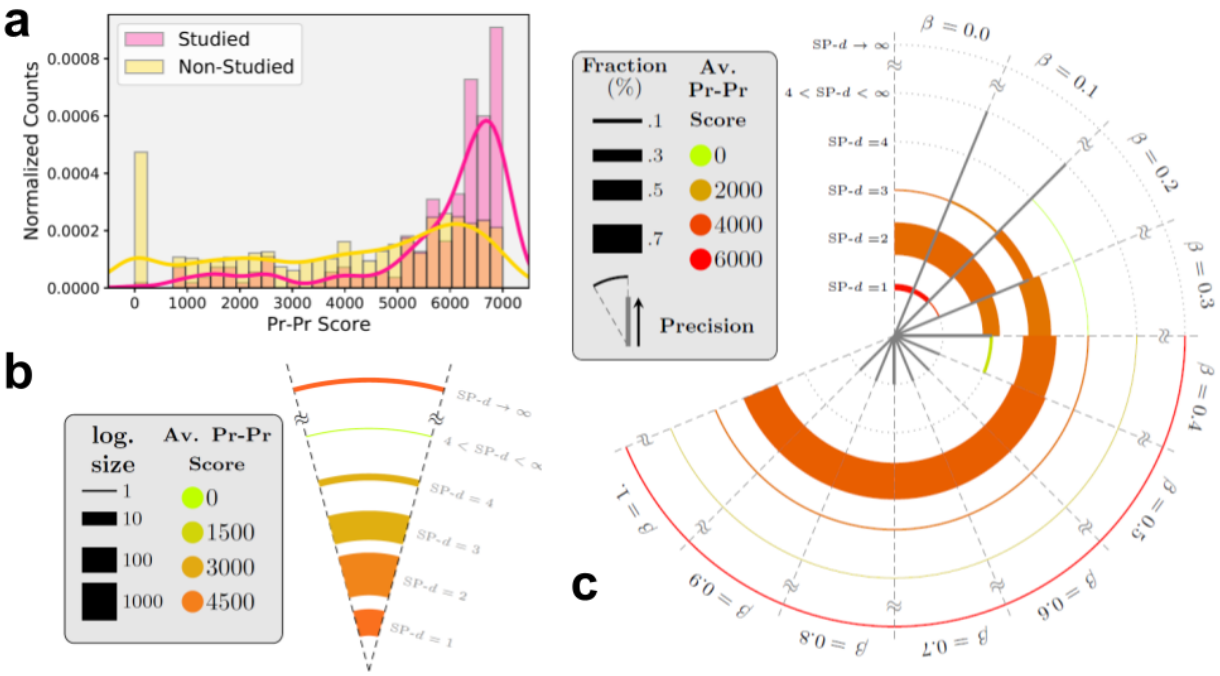
Extended Data Fig. 4. Distribution of expert densities between predicted discoveries and the corresponding properties: (a) drugs repurposing application (considering only the 67 diseases with statistically significant Spearman correlation coefficients, see Extended Data, Fig. 3); (b-d) energy-related materials science properties, i.e., thermoelectricity, ferroelectricity and photovoltaic capacity, respectively; and (e) therapies and vaccines for COVID-19. Curves measure normalized densities over the logarithm of Jaccard indices plotted by fitting a Beta distribution over expert densities for the 50 predictions. Solid and dashed lines represent mean values for the corresponding densities. It is clear that the distribution of expert densities for hypergarph-induced metrics (transition probability and deepwalk-based similarity) are concentrated around larger Jaccard index values than word embedding models tracing content alone. In content models, all estimated densities peak at zero ($0<a<1<b$, with a,b shape parameters of Beta distribution). CTD diseases are sorted by average expert similarity between them and the complete pool of drugs.



Extended Data Fig. 5. An example random walk from the property node "Coronavirus" to the material node "Progesterone". Every article in this path is a hyperedge (denoted by e_i in the i-th step) connecting the prior to the next node. The last article was cited by the University of Southern California clinical trial that investigated the effectiveness of progesterone for COVID-19 treatment. MeSH Terms in the articles are shown to better demonstrate their scope, with colored terms indicating relevant hints to the reasoning of the human scientists behind the study of progesterone for COVID-19 treatment. The path indicates a clear transition from Coronavirus-related topics to male-female differences in pathological conditions and lastly to progesterone-based therapy. Similar bridges between topics were highlighted by the trial's investigator as the main motivation for her study.



Extended Data Fig. 6. Precision-Recall Area Under the Curve (PR-AUC) for predicting experts who will discover particular materials possessing specific properties. Materials were selected to be True Positives among the correct, top 50 predictions of our deepwalk-based predictor (α=1). The evaluation here compares scores assigned to candidate and actual discovering experts who ultimately discovered and published the property associated with True Positives. We developed a deepwalk-based scoring function for this purpose. Expert candidates are those sampled at least once in our deepwalk trajectories, produced over our five-year period hypergraph. For a fixed (discovered) material, scores were computed based on proximity of experts to both property and material. An expert is a good candidate discoverer if she is close (in cosine similarity) to both property and material nodes in the embedded space. Discovered associations whose discoverers were not present in sampled deepwalk trajectories were ignored. In order to summarize the two similarities and generate a single set of expert predictions, we ranked experts based on their proximity to the property (R_P) and the material (R_M) and combined the two rankings using average aggregation. This ranking was used as the final expert score in our PR-AUC computations. We compared the log-PR-AUC of this algorithm with random selection of experts and also with a curve simulating an imaginary method whose log-PR-AUC is five times higher than random baseline. Results reveal that predictions were significantly superior to random expert selection for all electrochemical properties.



Extended Data Fig. 7. Results of running our AAI for predicting therapies and vaccines for COVID-19. Scientific relevance of candidate drugs were measured based on protein-protein interaction networks, hence named Pr-Pr scores. The scores we use here comprise ordinal ascending ranks of drugs in terms of their similarity to COVID-19 in the underlying Pr-Pr network (hence greater scores imply higher relevance). This data was provided to us by Gysi et al, where they ranked drugs based on rank-aggregation for several similarity metrics including embedded similarity of graph neural networks[26]. **(a)** Comparison between two groups of drugs in terms of the distribution of their Pr-Pr scores: 251 drugs that have been involved in at least one clinical trial related to COVID-19 (pink), and 3,697 drugs that have not been studied in any such trial (blue). The former group clearly showed strong skewness towards large scores, whereas the latter indicated a more uniform distribution of Pr-Pr scores with a high spike at zero. **(b)** Distribution of the average Pr-Pr scores across various levels of shortest-path distance (SP-*d*) with respect to the property node. Each orbit includes a set of drugs of the corresponding SP-*d* value, with size proportional to width of the arcs. The color of each arc represents the average Pr-Pr scores of the drugs in the corresponding orbit. Our pool of materials consisted of 3,948 drugs and we used all PubMed publications between 2015 and 2019 (as the five-year interval before the base year of 2020) to build our hypergraph. This figure implies that there exist drugs with high relevance scores that are further than two SP-*d* from the property node. However, the majority of studied drugs which were involved in COVID-19 clinical trials in 2020 belonged to the first two SP-*d* orbits (specifically, 16% from the first orbit, 82.1% from the second and less than 2% from the third orbit). This motivated us to more extensively explore the space of drugs with our AAI algorithm in order to capture any potential candidate for disruptive discovery. **(c)** Distribution of protein-protein (Pr-Pr) theoretical scores and SP-*d* of 50 candidates generated by our AAI algorithm with different β values for prediction year 2020. SP-*d* and semantic similarities based on word2vec embedding model were used for measuring human unavailability and scientific relevance, respectively. Increasing β effectively activates the alien component of our algorithm and introduced candidates within further orbits. These predictions had promising Pr-Pr scores in average, but, as expected, they yielded lower precisions in terms of gaining attention and discovery by human scientists (i.e., championing COVID-19 clinical trials) as they were less cognitively available to them.

References

1. Tshitoyan, V. *et al.* Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571**, 95–98 (2019).
2. Sanchez-Lengeling, B. & Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **361**, 360–365 (2018).
3. Jose, R. & Ramakrishna, S. Materials 4.0: Materials big data enabled materials discovery. *Applied Materials Today* **10**, 127–132 (2018).
4. Smalley, E. AI-powered drug discovery captures pharma interest. *Nat. Biotechnol.* **35**, 604–605 (2017).
5. Lo, Y.-C., Rensi, S. E., Torng, W. & Altman, R. B. Machine learning in chemoinformatics and drug discovery. *Drug Discov. Today* **23**, 1538–1546 (2018).
6. Vamathevan, J. *et al.* Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* **18**, 463–477 (2019).
7. Khadherbhi, S. R. & Babu, K. S. Big Data Search Space Reduction Based On User Perspective Using Map Reduce. *International Journal of Advanced Technology and Innovative Research* **7**, 3642–3647 (2015).
8. Galton, F. Vox populi (the wisdom of crowds). *Nature* **75**, 450–451 (1907).
9. Surowiecki, J. The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business. *Economies, Societies and Nations* **296**, (2004).
10. Page, S. E. *The Diversity Bonus: How Great Teams Pay Off in the Knowledge Economy*. (Princeton University Press, 2019).
11. Danchev, V., Rzhetsky, A. & Evans, J. A. Centralized scientific communities are less likely to generate replicable results. *Elife* **8**, (2019).
12. Belikov, A. V., Rzhetsky, A. & Evans, J. Detecting signal from science:The structure of research communities and prior knowledge improves prediction of genetic regulatory experiments. *arXiv [cs.SI]* (2020).
13. Swanson, D. R. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspect. Biol. Med.* **30**, 7–18 (1986).
14. Swanson, D. R. Medical literature as a potential source of new knowledge. *Bull. Med. Libr. Assoc.* **78**, 29–37 (1990).
15. 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).
16. Evans, J. & Rzhetsky, A. Machine Science. *Science* **329**, 399–400 (2010).
17. Digiacomo, R. A., Kremer, J. M. & Shah, D. M. Fish-oil dietary supplementation in patients with Raynaud's phenomenon: A double-blind, controlled, prospective study. *The American Journal of Medicine* vol. 86 158–164 (1989).

18. Chiu, H.-Y., Yeh, T.-H., Huang, Y.-C. & Chen, P.-Y. Effects of Intravenous and Oral Magnesium on Reducing Migraine: A Meta-analysis of Randomized Controlled Trials. *Pain Physician* **19**, E97–112 (2016).

19. Mikolov, T., Yih, W. & Zweig, G. Linguistic regularities in continuous space word representations. *hlt-Naacl* (2013).

20. Rzhetsky, A., Foster, J. G., Foster, I. T. & Evans, J. A. Choosing experiments to accelerate collective discovery. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 14569–14574 (2015).

21. Burger, B. *et al.* A mobile robotic chemist. *Nature* **583**, 237–241 (2020).

22. Shi, F., Foster, J. G. & Evans, J. A. Weaving the fabric of science: Dynamic network models of science's unfolding structure. *Soc. Networks* **43**, 73–85 (2015).

23. Chitra, U. & Raphael, B. Random Walks on Hypergraphs with Edge-Dependent Vertex Weights. in *Proceedings of the 36th International Conference on Machine Learning* (eds. Chaudhuri, K. & Salakhutdinov, R.) vol. 97 1172–1181 (PMLR, 2019).

24. Perozzi, B., Al-Rfou, R. & Skiena, S. DeepWalk: online learning of social representations. in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* 701–710 (Association for Computing Machinery, 2014).

25. Davis, A. P. *et al.* The Comparative Toxicogenomics Database: update 2019. *Nucleic Acids Res.* **47**, D948–D954 (2019).

26. Gysi, D. M. *et al.* Network Medicine Framework for Identifying Drug Repurposing Opportunities for COVID-19. *ArXiv* (2020).

27. Ghandehari, S. Progesterone for the Treatment of COVID-19 in Hospitalized Men. *Identifier NCT04365127* https://clinicaltrials.gov/ct2/show/study/NCT04365127 (2020).

28. Ghandehari, S. *et al.* Progesterone in Addition to Standard of Care Versus Standard of Care Alone in the Treatment of Men Hospitalized with Moderate to Severe COVID-19: A Randomized, Controlled Pilot Trial. *Chest* (2021) doi:10.1016/j.chest.2021.02.024.

29. Mehdizadeh Dehkordi, A., Zebarjadi, M., He, J. & Tritt, T. M. Thermoelectric power factor: Enhancement mechanisms and strategies for higher performance thermoelectric materials. *Mater. Sci. Eng. R Rep.* **97**, 1–22 (2015).

30. Ricci, F. *et al.* An ab initio electronic transport database for inorganic materials. *Sci Data* **4**, 170085 (2017).

31. Gediya, L. K. & Njar, V. C. Promise and challenges in drug discovery and development of hybrid anticancer drugs. *Expert Opin. Drug Discov.* **4**, 1099–1111 (2009).

32. Jones, B. F. The Burden of Knowledge and the 'Death of the Renaissance Man': Is Innovation Getting Harder? *Rev. Econ. Stud.* **76**, 283–317 (2009).

33. Szell, M., Ma, Y. & Sinatra, R. A Nobel opportunity for interdisciplinarity. *Nat. Phys.* **14**, 1075–1078 (2018).

34. Xu, J. *et al.* Building a PubMed knowledge graph. *Sci Data* **7**, 205 (2020).

35. Torvik, V. I. & Smalheiser, N. R. Author Name Disambiguation in MEDLINE. *ACM Trans. Knowl. Discov. Data* **3**, (2009).

36. Ammar, W. *et al.* Construction of the Literature Graph in Semantic Scholar. *arXiv [cs.CL]* (2018).

37. Ong, S. P. *et al.* Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319 (2013).

38. Grover, A. & Leskovec, J. node2vec: Scalable Feature Learning for Networks. *KDD* **2016**, 855–864 (2016).

39. Hamilton, W. L., Ying, R. & Leskovec, J. Inductive representation learning on large graphs. in *Proceedings of the 31st International Conference on Neural Information Processing Systems* 1025–1035 (Curran Associates Inc., 2017).

40. Kipf, T. N. & Welling, M. Variational Graph Auto-Encoders. *Stat* **1050**, 21 (2016).

41. Coakley, C. W. Practical Nonparametric Statistics (3rd ed.). *J. Am. Stat. Assoc.* **95**, 332 (2000).

42. Brynjolfsson, E., Rock, D. & Syverson, C. *Artificial Intelligence and the Modern Productivity Paradox: A Clash of Expectations and Statistics*. https://www.nber.org/papers/w24001 (2017) doi:10.3386/w24001.

43. Freund, Y., Schapire, R. E. & Others. Experiments with a new boosting algorithm. in *icml* vol. 96 148–156 (Citeseer, 1996).

44. Fortunato, S. *et al.* Science of science. *Science* **359**, (2018).

45. Zhuang, Y.-T., Wu, F., Chen, C. & Pan, Y.-H. Challenges and opportunities: from big data to knowledge in AI 2.0. *Frontiers of Information Technology & Electronic Engineering* **18**, 3–14 (2017).

46. Zhou, Q. *et al.* Learning atoms for materials discovery. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E6411–E6417 (2018).

Acknowledgements

The authors wish to thank our funders for their generous support: National Science Foundation #1829366; Air Force Office of Scientific Research #FA9550-19-1-0354, #FA9550-15-1-0162; DARPA #HR00111820006. We also thank participants of the Santa Fe Institute workshop “Foundations of Intelligence in Natural and Artificial Systems”, the University of Wisconsin at Madison’s HAMLET workshop, and colleagues at the Knowledge Lab for helpful comments.

SUPPLEMENTARY INFORMATION: ACCELERATING SCIENCE WITH HUMAN VS. ALIEN ARTIFICIAL INTELLIGENCES

Jamshid Sourati[1], James A. Evans[1,2]

[1]Knowledge Lab, University of Chicago, Chicago, IL, USA

[2]Santa Fe Institute, Santa Fe, NM, USA

S1 Multistep Transition Probabilities

The first similarity metric we used based on our random walk settings was based on multistep transitions from the property node (denoted by P) to a target material (denoted by M). We considered two- and three-step transitions with intermediate nodes conditioned to belong to the set of authoring experts (denoted by $\mathcal{A}$). 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:

$$\begin{aligned}\mathbb{P}\big(n_2 = M, n_1 \in \mathcal{A} \,\big|n_0 = P\big) &= \sum_{A\in\mathcal{A}} \mathbb{P}\big(n_2 = M, n_1 = A\big|n_0 = P\big) \\ &= \sum_{A\in\mathcal{A}} \mathbb{P}\big(n_1 = A\big|n_0 = P\big)\cdot\mathbb{P}\big(n_2 = M\big|n_0 = A\big), \end{aligned} \tag{S1}$$

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

$$\mathbb{P}\big(n_1 = A\big|n_0 = P\big) = \frac{1}{d(P)}\sum_{e:\{P,A\}\in e}\frac{1}{d(e)}, \tag{S2}$$

where $d(P)$ is the degree of node P, i.e., the number of hyperedges 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 (S2) accounts for selecting a hyperedge that includes P and the second computes the probability of selecting A from one of the common hyperedeges (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 $\mathbf{P}_{ij} = \mathbb{P}\big(n_1 = j\big|n_0 = i\big)$. Then, two- and three-step transitions between nodes P and M could be computed via $\mathbf{P}\big(P,[\mathcal{A}]\big)\cdot\mathbf{P}\big([\mathcal{A}],M\big)$ and $\mathbf{P}\big(P,[\mathcal{A}]\big)\cdot\mathbf{P}\big([\mathcal{A}],[\mathcal{A}]\big)\cdot\mathbf{P}\big([\mathcal{A}],M\big)$, respectively, where $\mathbf{P}\big(P,[\mathcal{A}]\big)$ defines selection of the row corresponding to node P and columns corresponding to authors in set $\mathcal{A}$.

S2 Combining Scores for Human (Un)availability and Scientific Plausibility

Our Alternative or Alien Artificial Intelligence (AAI) algorithm combines two sources of information to generate candidates that are simultaneously alienated and scientifically plausible. The two signals we use to quantify these components in our experiments include the Shortest-Path distance (SP-d) of the materials to the property node (measuring human expert avoidance or cognitive unavailability) and their semantic similarities based on a word embedding model with regards to the property keyword (measuring relevance). Following[6], the word embedding model we used was the skipgram Word2Vec model trained over the literature in the five-year period preceding the prediction year. The mixing coefficient $\beta \in [0, 1]$ determines how much importance will be assigned to avoiding authors versus chasing current theoretical plausibility when combining scores. It is also desirable that the effect of the two sources become equal when $\beta = \frac{1}{2}$, and that the output score varies continuously as β changes. Let us denote human avoidance and scientific plausibility scores computed for an entity x by $s_1(x)$ and $s_2(x)$, respectively. In this section, we discuss several methods for combining these signals and compare them in the context of AAI's performance.

Simply weighted averaging of the scores through $\beta s_1(x) + (1 - \beta)s_2(x)$ is inappropriate for our experiment due to highly distinct scales of the human avoidance and scientific plausibility signals (preventing equal contribution when $\beta = 1/2$). Moreover, the SP-d 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. As a workaround, we applied Van der Waerden transformation over the scores. Suppose S is a set of scores and $s(x) \in S$, then its Van der Waerden transformation $\tilde{s}(x)$ is defined as

$$\tilde{s}(x) \;=\; \phi\left(\frac{r(x)}{|S|+1}\right), \tag{S3}$$

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 will 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.

Alternatively, geometric and harmonic means could be used for combining variables with unequal scales. In order to be able to use these measures, we replaced unbounded values of SP-d by an arbitrary finite value whose magnitude is larger than the other finite elements. We define the β-weighted versions of these means as below:

$$\text{geometric:} \quad s_{\text{GEO}}(x) = \left(s_1(x)^{\beta} \cdot s_2(x)^{1-\beta}\right)^{1/2} \tag{S4}$$

$$\text{harmonic:} \quad s_{\text{HRM}}(x) = 2 \Big/ \left(\frac{\beta}{s_1(x)} + \frac{1-\beta}{s_2(x)}\right) \tag{S5}$$

In order to make comparison between the above β-weighted combination techniques in the context of chemical compounds and the thermoelectric property, we set s_1 and s_2 to SP-d and Power Factor (PF) values, respectively. We ran AAI experiments varying β values between

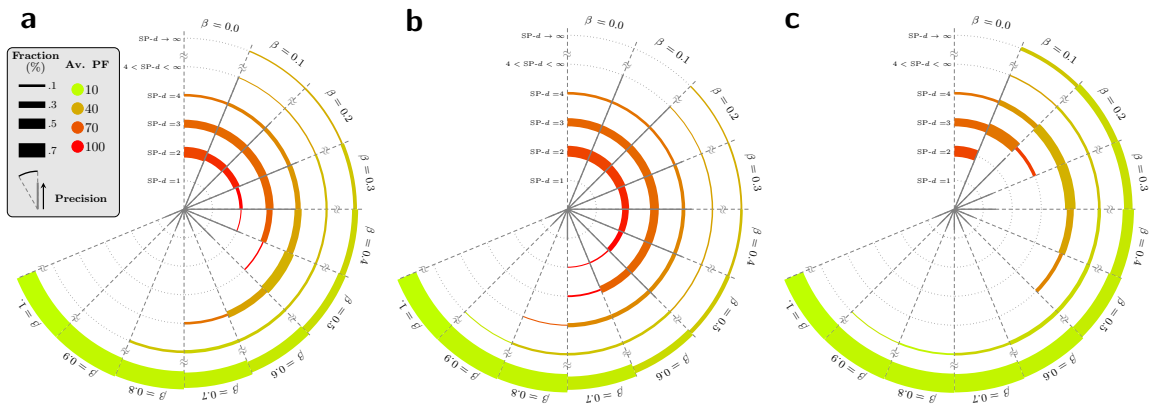


Figure S1: Distribution of SP-d values (alienness) and PF values (scientific relevance) of candidates generated by our AAI framework using various β-weighted score combinations: **(a)** Van der Waerden transformation followed by weighted averaging of the Z-scores, **(b)** weighted geometric mean, and **(c)** weighted harmonic mean.

0 and 1 with steps of 0.1 and used the same metrics to self-evaluate the human avoidance and scientific plausibility of candidates generated with each β value. Results indicate how rapidly changing the mixture weight shifts the attention from scientific relevance to human unavailability in various formulations. These results do not evaluate the performance of the AAI framework, but merely act as sanity checks for the appropriateness of score combination methods.

Results shown in fig. S1 demonstrate that using the harmonic mean leads to abrupt drops in SP-d values when β slowly departs from zero. By contrast, using the geometric mean causes oversmoothing and yields asymmetric behaviour around $\beta = 0.5$ overrating the scientific relevance. Applying Z-scores of the transformed scores resulted in a reasonably balanced and smooth β-weighted combination of scores.

S3 Experiments with Deep Neural Networks

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

S3.1 Graph 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 used the Graph Sample and Aggregate (GraphSAGE) platform, which applies the aggregations function on a subset of neighbors to avoid computational overhead[2]. Let $\mathbf{h}_i^\ell$ denote 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

$$\mathbf{h}_i^{\ell+1} \;=\; \sigma\left(\mathbf{f}_{\text{AGG}}\left(\{\mathbf{h}_j\}_{j\in\tilde{\mathcal{N}}(i)}\right)\mathbf{W}_\ell\right), \tag{S6}$$

where $\mathbf{f}_{\text{AGG}}$ is an aggregation function (e.g., mean, pooling, etc) applied on the concatenation of the local neighborhood's messages $\{\mathbf{h}_j\}_{j\in\tilde{\mathcal{N}}(i)}$, where $\tilde{\mathcal{N}}(i)$ is a subset of k_ℓ uniformly sampled nodes from the immediate neighbors of the i-th node $\mathcal{N}(i)$. The resulting aggregated messages will undergo a single-layer neural network parameterized by $\mathbf{W}_\ell$ (the bias term is ignored for simplicity) and the non-linear activation σ. The input messages in the first step, i.e., $\mathbf{h}_i^0\,\forall i$, 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[4], where the encoder component was modeled using the GraphSAGE architecture and the decoder component simply 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[5]. Once the positive samples were drawn using sliding window size of 8, we begin minimizing the loss function in a mini-batch setting by iterating over pairs. We used batch size of 1000, negative sampling size of 15 (per positive pair), learning rate of 5×10^{-6} and the Adam optimizer[3] with the default parameters.

S3.2 Experimental Settings

We trained our graph autoencoder in two different settings: (1) using our full hypergraph, and (2) 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 at the time we draw positive samples and compute 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 corresponding to $\alpha \to \infty$ so that no author nodes would be present in the random walk sequences. 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).

References

[1] I. Chami, S. Abu-El-Haija, B. Perozzi, C. Ré, and K. Murphy. Machine learning on graphs: A model and comprehensive taxonomy, 2021.

[2] W. L. Hamilton, R. Ying, and J. Leskovec. Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 1025–1035, 2017.

[3] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[4] T. N. Kipf and M. Welling. Variational graph auto-encoders. *arXiv preprint arXiv:1611.07308*, 2016.

[5] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013.

[6] V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K. A. Persson, G. Ceder, and A. Jain. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571(7763):95–98, 2019.