nature human behaviour

Article

https://doi.org/10.1038/s41562-023-01648-z

Accelerating science with human-aware artificial intelligence

Received: 17 August 2022

Accepted: 2 June 2023

Published online: 13 July 2023

Check for updates

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

Artificial intelligence (AI) models trained on published scientific findings have been used to invent valuable materials and targeted therapies, but they typically ignore the human scientists who continually alter the landscape of discovery. Here we show that incorporating the distribution of human expertise by training unsupervised models on simulated inferences that are cognitively accessible to experts dramatically improves (by up to 400%) AI prediction of future discoveries beyond models focused on research content alone, especially when relevant literature is sparse. These models succeed by predicting human predictions and the scientists who will make them. By tuning human-aware AI to avoid the crowd, we can generate scientifically promising 'alien' hypotheses unlikely to be imagined or pursued without intervention until the distant future, which hold promise to punctuate scientific advance beyond questions currently pursued. By accelerating human discovery or probing its blind spots, human-aware AI enables us to move towards and beyond the contemporary scientific frontier.

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 experimentally evaluate all candidate designs[1]. To face this challenge, researchers have initialized data-driven artificial intelligence (AI) models with published scientific results to create powerful prediction engines. These models have begun to assist human discovery by focusing scientific attention on the subset of discovery candidates most predicted to possess properties relevant to energy[2], human health[3] and other economic and societal values. In this way, AI intervenes in the discovery process by proposing efficient, model-based experiments that would require much longer for unassisted human scientists to identify. However, such efforts typically ignore the distribution of scientists and inventors[4]—the human prediction engines who continuously alter the landscape of discovery and invention. As we demonstrate below, incorporating knowledge of human researchers can dramatically improve predictions of future discoveries compared with AI methods that ignore them. Our work formalizes and demonstrates the critical importance of situated human expertise, communication and collaboration for unfolding scientific advance.

Previous studies have indicated that most new scientific discoveries emerge within neighbourhoods of prior findings[5,6]. Here we take a step further and demonstrate that the collective pattern of scientific attention is sufficient to boost the precision of future discovery forecasts. This generalizes the availability heuristic—the psychological tendency for individuals to evaluate event frequency on the basis of cognitive availability[7]. The availability heuristic is known to result in misjudgements and decision bias[8,9]. Here we consider how and when this aggregates in scenarios involving entire scientific communities[10]. The more scientists investigate a combination of topics, the more frequently other scientists from their community will observe it presented at conferences and read about it in literature. As that combination of ideas is spoken and written about, it becomes easy to imagine and consider by nearby scientists and so conditions future scientific consideration and investigation. Here we demonstrate that the distribution of scientists who author articles and their collaboration networks across topics and time is sufficient to foresee future discoveries and their discoverers with high precision, especially when research on the topic is sparse. This distribution, which can be recovered from publication metadata, represents a critical social fact that can stably improve our inferences about whether possible scientific relationships will soon be attempted. It can also inform our understanding of whether scientific possibilities will remain unimagined and unexplored until the more distant future[11].

[1]Department of Sociology, University of Chicago, Chicago, IL, USA. [2]Santa Fe Institute, Santa Fe, NM, USA. e-mail: jevans@uchicago.edu

We define scientific knowledge discovery as the first-time report of the relationship between an existing material and a well-defined property. An example of such pairwise relationships is 'vancomycin may be used to treat pneumonia', where vancomycin is the material and effective treatment of pneumonia is the property. Our approach draws on explicit measurement of the distribution of human scientists around each topic involved in candidate discoveries, using advances in unsupervised manifold learning[12–14] and drawing upon easily available publication meta-data. By programmatically incorporating information on the evolving distribution of human experts, our approach balances exploration and exploitation in experimental search that could be used to accelerate the realization of discoveries predicted to appear in future. We contrast our human-aware approach with precise replication of a recent, prominent content-only analysis[15] that trained a Word2Vec embedding model[12] over millions of abstracts from materials science publications. That study used the resulting word vectors to infer that materials closest to electrochemical properties in the embedding space will be discovered in the future to possess that property. Our models yield a ~100% increase in the precision of forecasts regarding future materials science discoveries. We extend this approach to identify a much broader matrix of materials and their functional properties[16], demonstrating comparable increases for predicting thousands of drugs to treat more than a hundred distinct human diseases, including vaccines and therapies for COVID-19.

Using human-aware AI, we can not only accelerate science by anticipating the human crowd; we can avoid that crowd to construct insights that punctuate human discovery with complementary hypotheses unlikely to be discovered by human scientists. If we model discovery as establishing novel links among otherwise disconnected concepts[11], it cannot occur until discoverers arise with viewpoints that bridge the fields required to imagine those conceptual connections (Fig. 1a). This diversity of scientific viewpoints was implicitly drawn upon by pioneering information scientist Swanson in his heuristic approach to knowledge generation. For example, he hypothesized that if Raynaud's disorder was linked to blood viscosity in one literature, and fish oil was known to decrease blood viscosity in another, then fish oil might lessen the symptoms of Raynaud's disorder but would probably not be arrived at in either field because no scientist was available to infer it[17–19]. This was one of several hypotheses later experimentally demonstrated[20–22]. Expansive opportunities for discovery persist as researchers crowd around past discoveries[6], neglecting to explore regions of knowledge cognitively distant from recent findings[23] (Extended Data Fig. 1). Our human-aware approach to complementary discovery scales and makes Swanson's heuristic continuous, identifying unstudied pairs of scientific entities likely to be scientifically and technologically relevant but unlikely to be imagined. This approach avoids scientific topics at the centre of collective attention and generates complementary hypotheses, which not only are unlikely to be considered by unassisted human experts but also outperform published discoveries. By staging intellectual arbitrage between isolated communities, our 'alien' predictions are unconstrained by the human incentive to flock together within fields. In this way, our human-aware framework provides opportunities for accelerating the normal pathway of human discovery by predicting human-accessible hypotheses and punctuating that path by predicting human-inaccessible hypotheses that complement it.

Incorporating human experts with hypergraph proximity

We model the distribution of inferences that are collectively and cognitively accessible to scientists by constructing a hypergraph over research publications. A hypergraph is a generalized graph where an edge connects a set of nodes rather than a node pair. Our research hypergraph is mixed, containing nodes corresponding not only to materials and properties mentioned in titles or abstracts but also to the researchers who investigate them (Fig. 1c, first step). Following the construction of this research hypergraph, we identify cognitively accessible inferences by generating random walk sequences over it. These walks suggest paths of inference available to active human scientists, which trace mixtures of diverse expertise sufficient for contemporary discoveries. If a valuable material property (for example, 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 later co-authors with another who has previously worked with sodium nitrite ($NaNO_2$, another ferroelectric material), that scientist is more likely to imagine whether sodium nitrite has the property through conversation than a scientist without the personal connection. In this way, the density of random walks over our research hypergraph is proportional to the density of cognitively plausible and conversationally attainable 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 only in one community with a material understood only in a disjoint one (Fig. 1a). We hypothesize that identifying topics with high human expert density around them provides us with an informative signal regarding near-future discoveries. These topics might be located far from one another in terms of the number of steps required to travel between them in the hypergraph, but a random walker—and the collective scientific mind—can easily travel between them if intermediate steps are socially dense, facilitating conversation and collaboration (Fig. 1a).

To generate each random walk sequence, our model (1) initiates the walk with a valued property (for example, ferroelectricity) as the first node in the sequence, (2) randomly selects an article (hyperedge) having mentioned that property, (3) randomly selects a material or author from that article as the next node (the end of the first step), and then starts the second step by randomly selecting another article with the newly selected material or author, and repeats this Markov process[5,14] a pre-specified number of times (see Fig. 1b for an example and the Supplementary Information for more details). Each random walk step can be viewed as a simulation of human actions: an author–author step mimics networking or conversation between two expert collaborators, an author–material or author–property step represents an author's deep familiarity with the selected material/property they have studied and published on, and a material/property–material/property step captures the potential for the transition to be realized by human scientists through reading a collection of scientific articles. Owing to the collaborative character of physical and biological science, author nodes in our hypergraph far outnumber materials. To compensate for this imbalance, we devised a non-uniform sampling distribution parameterized by α, which roughly determines the fraction of material-to-author nodes in the resulting sequences. Specifically, we define α when sampling a node from a paper (for example, in step 3 above) such that the probability of selecting a material is α times that of selecting an author (Supplementary Fig. 1). Larger values of α result in sampling materials/properties more frequently, suggesting that our simulated researcher will uncover new scientific possibilities predominantly through research and reading; smaller values result in higher frequencies of author selection, implying discovery through networking, conversation and collaboration with others in the field.

Random walks over the mixed hypergraph induce meaningful proximities between nodes. The proximity of two authors suggests they share similar research interests and experiences. The proximity of a material to a scientist assesses the likelihood that the scientist is or will become familiar with that material through research experience, related reading or social interaction. The proximity of materials to one another suggests that they may be substitutes or complements or share another more subtle relationship such as interaction or comparison. Finally, the proximity of a material to a property suggests the likelihood that the material may possess the property and that a scientist

https://doi.org/10.1038/s41562-023-01648-z

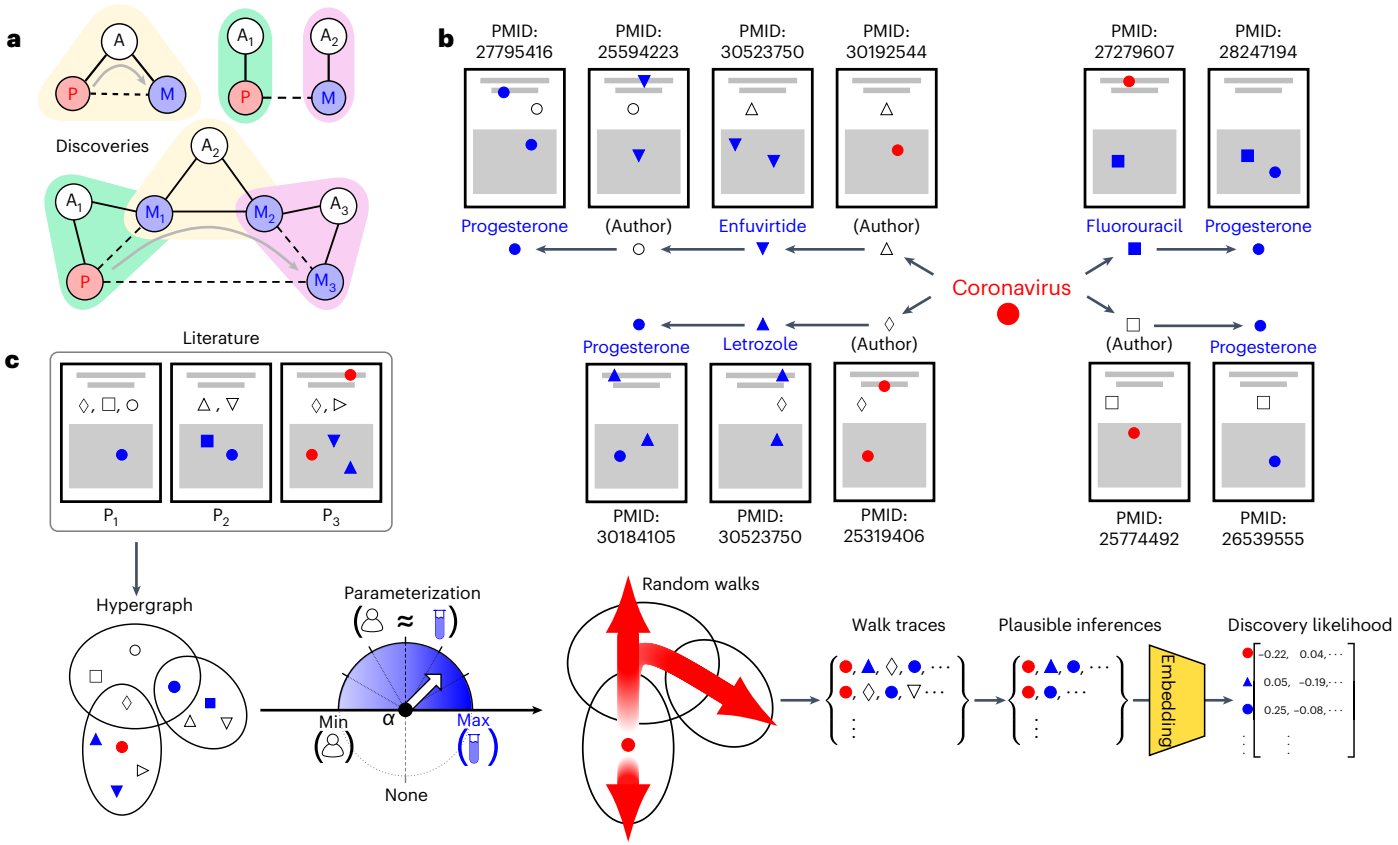


Fig. 1 | Motivation and design of our approach to simulate human-accessible scientific inferences. a, Three scenarios where a hidden underlying relationship between material M and property P waits to be discovered. The non-coloured circles represent non-overlapping populations of human experts, and the coloured nodes indicate materials (blue) or properties (red). The background colours represent overlapping disciplinary communities, within which scientists and topics are embedded. Solid lines between non-coloured and coloured nodes imply that experts studied or have experience with the material or property. Dashed lines represent property–material links that exist but have not yet been discovered by human scientists, and grey arrows represent new hypotheses proposed by our algorithm. The P–M relation in the upper left scenario is likely to be discovered and published in the near future and is proposed by our algorithm; the P–M relation in the upper right is likely to escape scientists' attention and also the notice of our algorithm, which simulates human-accessible hypotheses. Nevertheless, our algorithm also captures transitive inference as scientists do through research and conversation over time; let $P–M_1–M_2–M_3$ be a chain of materials connected to property P, and every consecutive pair $M_i–M_{i+1}$ are strongly connected either because they are already shown to be connected in published articles or because there is a group of researchers familiar with both, having studied both across their opus of research. Our algorithm walks over consecutive pairs and infers the existence of the $P–M_3$ relationship and its likelihood of discovery in future. **b**, Four examples of random walk paths starting from 'Coronavirus' (property) and ending at 'Progesterone' (a chemical under clinical trial investigation for COVID-19 therapeutic efficacy). Each arrow connecting two nodes indicates a sampling step, where the paper shown above the receiving node comprises the selected hyperedge for that step, which by construction contains both nodes sampled in the prior and current steps. **c**, Illustration of our hypergraph deepwalk algorithm. (1) We construct a hypothetical hypergraph based on literature represented by three papers. The non-coloured shapes represent authors, and the coloured shapes indicate properties (red) or materials (blue) mentioned in article titles or abstracts. (2) We perform classic or α-modified random walk sampling, which (3) results in a set of sequences consisting of authors, materials and the focal property. (4) We remove authors from the sequences, retaining only the materials on which discovery inference will be applied. (5) We train a word embedding model (for example, Word2Vec) on these sampled human-accessible sequences of material/property tokens, which results in (6) a vector representation of property and materials that we use to compute similarities for prediction.

will discover and publish it (Extended Data Fig. 1a,b). In this way, our hypergraph-induced proximities incorporate physical and material properties latent within literature, as well as the distribution of human scientists, which enables us to anticipate inferences by those scientists and predict upcoming discoveries. The distribution of human scientists is a factor available to and naturally 'read' by other competitive scientists when they attend conferences and survey their fields for promising new directions.

To foresee the potential discovery of materials with a valued property (for example, store energy, cure breast cancer or vaccinate against COVID-19), we utilize random-walk-induced node similarity metrics to capture the relevance between the targeted property and candidate materials. These metrics, evaluated between pairs of property/material nodes, reflect the human-inferable relatedness of corresponding nodes and are used to sort candidate materials and report those highest ranked as inferred to possess the property. A simple metric of this kind draws upon the local hypergraph structure to estimate the transition probability that a random walker travels from the property node to a material through intermediate author nodes within a fixed number of steps, denoted by s. We use Bayes' rules to calculate these probabilities without the need for actually running the random-walk sampler (Supplementary Fig. 2). Here we only consider two- and three-step transitions ($s = 2$ and $s = 3$). Our main choice of metric, however, is based on a popular, unsupervised neural-network-based embedding algorithm (deepwalk[13]), estimated over the random walks we generate. Like previous content-only methods[15], this method also entails the construction of a word embedding model[12]. Instead of abstract sentences as input, however, the embedding is constructed over our hypergraph, considering every random walk sequence a 'sentence' that links materials, experts and functional properties.

Whereas a text-based embedding captures semantic relevance among words, our approach obtains word vectors while preserving hypergraph proximities among all nodes and therefore can be used to measure the human cognitive accessibility of each material with respect to a targeted property. Because inferred discoveries involve relevant materials, we train the deepwalk embedding model after excluding authors from our random walk sequences (Fig. 1c). Cosine similarity in the resulting embedding space can be used as a relevance metric. We use these two relevance metrics, transition probabilities and deepwalk similarities, as twin criteria for selecting materials most likely to emerge as the next discoveries. Additionally, we train deeper graph convolutional neural networks, which confirm the pattern of results obtained from deepwalk (Methods and Supplementary Information).

Note that our models do not use more data than traditional content-based methods but instead alter the type of data we feed them. Specifically, our approach extracts and adds authorship information but excludes the vast majority of textual content, excepting only material and property co-occurrences. In other words, our data are richer than traditional datasets in one dimension by adding human and social information, but less informative and dense in terms of content. Overall, our method possesses less data than the baselines against which we compare. In this way, our model's performance improvement, as shown below, reflects not more data but more informative data.

Results on anticipating human discoveries

To demonstrate the power of accounting for human experts, we use transition probability and deepwalk metrics to build two alternative discovery predictors. These algorithms assess the relevance of the focal property to each candidate material on the basis of literature published before a given prediction year (for example, 2001) by embedding the human-aware hypergraph. We contrast our predictions with a random baseline and predictions generated from precisely replicated prior work that used word embeddings based on the textual content of scientific literature without accounting for the distribution of human scientists[15]. This prior work measured property/material relevance with cosine similarity from a Word2Vec model[12] trained over the contents of scientific articles published prior to the prediction year. Our experiments and evaluation framework are identical to the settings of this study to facilitate precise replication. Each evaluated algorithm selects the 50 materials with the highest similarity to the focal property based on hypergraph or Word2Vec similarity metrics and reports them as discovery predictions. We evaluate prediction quality on the basis of their overlap with materials discovered and published after the prediction year (see the Methods for further details; for alternative evaluation metrics and prediction sizes, see Extended Data Fig. 2 and Supplementary Fig. 3).

Energy-related materials prediction

In our first set of experiments, we considered the valuable electrochemical properties of thermoelectricity, ferroelectricity and photovoltaic capacity against a pool of 100,000 candidate inorganic compounds. Following the evaluation regime of Tshitoyan et al. on the same dataset (1.5 million scientific articles about inorganic materials)[15], we ran prediction experiments with prediction year 2001 for all three properties, predicting future discoveries as a function of research publicly available to contemporary scientists. We computed annual precisions following the prediction year until the end of 2018 (Extended Data Fig. 1c) and visualized them in a cumulative manner (Fig. 2a–c). The results indicate that predictions accounting for the distribution of human scientists outperformed baselines for all properties and materials by an average of 100%.

Sensitivity analyses with α reveal that a deepwalk algorithm with $\alpha = 1$, which balances the likelihood of sampling materials and author nodes, had the highest and most consistent precision of prediction. Moreover, even for extremely large values of α (that is, $\alpha \to \infty$), where our random walk is ignorant of human experts, the deepwalk algorithm still substantially outperforms the Word2Vec model. We conjecture that this occurs because vague title and abstract words, irrelevant to future discoveries, add noise to the proximity of properties and materials. Our hypergraph method ignores these words, but they mislead Word2Vec, resulting in weaker predictions. This suggests a more specific conjecture. Material words alone are more relevant, specific and semantically local to other materials and properties mentioned within a paper. In this way, our hypergraph-based approach infers new discoveries in the vicinity of previous findings. Such a localized process aligns with how scientists make discoveries, leading to stronger predictions[5,6].

Drug repurposing prediction

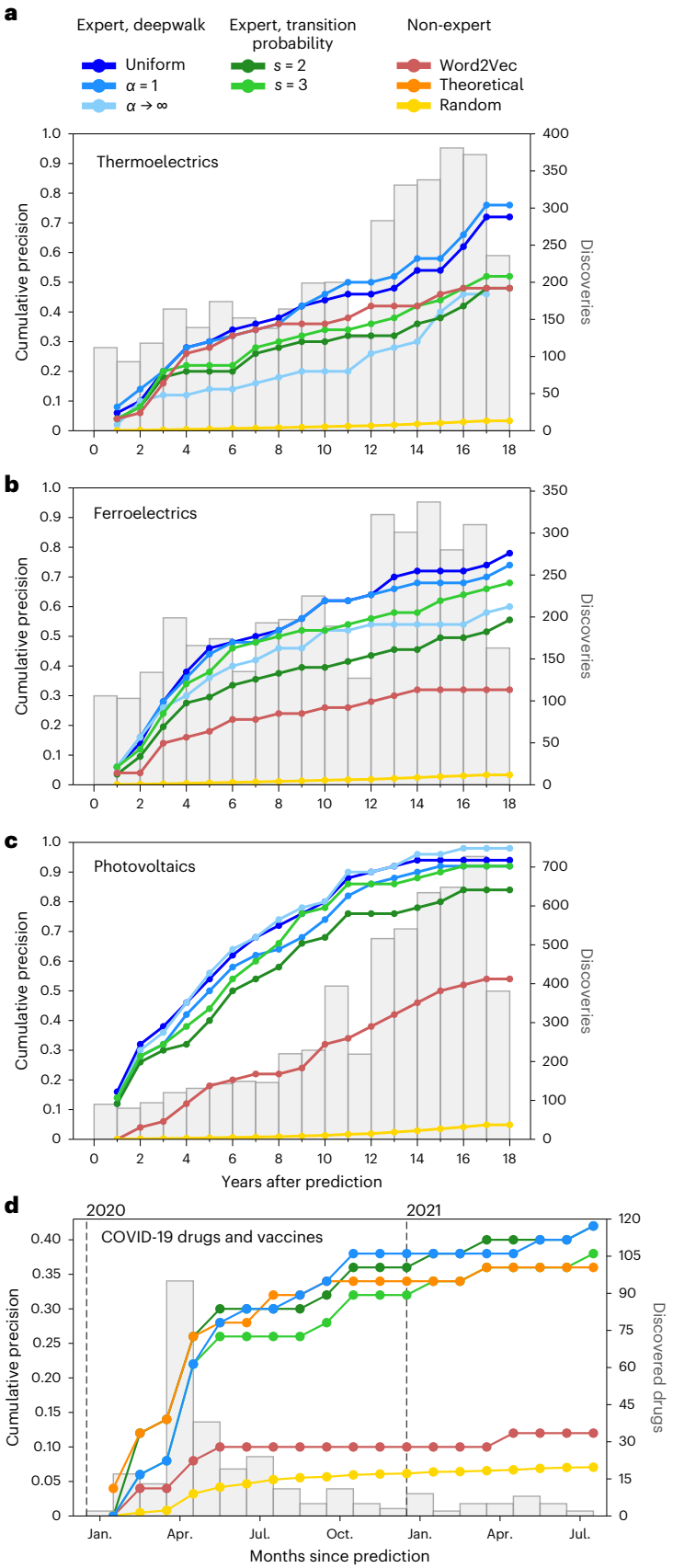
We used the same approach to explore the repurposing of ~4,000 existing Food and Drug Administration-approved drugs to treat 100 important human diseases. We used the MEDLINE database of biomedical research publications and set the prediction year to 2001 (Extended Data Fig. 1c). Ground-truth discoveries were based on drug–disease associations established by expert curators of the Comparative Toxicogenomics Database (CTD)[24], which chronicles the capacity of chemicals to influence human health. Figure 2e reports prediction precisions 18 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 research content alone. We found a strong correlation between our human-aware prediction precision and drug occurrence frequency in literature ($r = 0.74$, $P < 0.001$), implying that our approach works best for diseases whose relevant drugs are frequently mentioned in prior research.

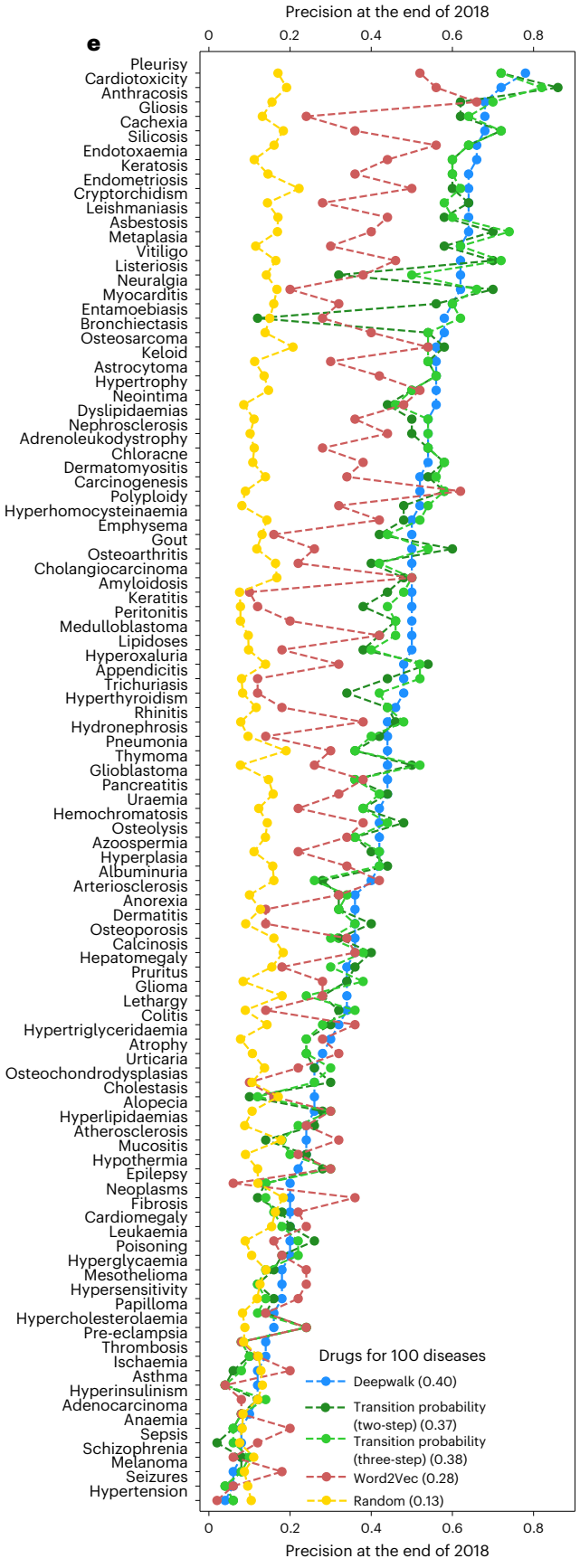
COVID-19 therapy and vaccine prediction

We also considered therapies and vaccines to treat or prevent SARS-CoV-2 infection. Here the prediction year was set to 2020 (Extended Data Fig. 1c), when the global search for relevant drugs and vaccines began in earnest. Following Morselli Gysi et al.[25], we considered a therapy relevant to COVID-19 if it amassed evidence to merit a COVID-19-related clinical trial, as reported by ClinicalTrials.gov. The results shown in Fig. 2d indicate that 36% and 38% of the predictions made by transition probability and deepwalk-based metrics, respectively, were selected by biomedical experts to evaluate using expensive clinical trials within 12 months of the prediction date (that is, the end of December 2020), which further increased to 42% by the end of July 2021. This is 350% to 400% higher than the precision of discovery candidates generated by scientific content alone (10% after the first 12 months and 12% in July 2021). These precisions were even higher than those of a recently proposed predictive model based on an ensemble of deep and shallow learning predictors trained on multiply measured

Fig. 2 | Evaluating human-accessible discovery predictions against various baselines. a–**e**, Precision rates for human-accessible discovery predictions regarding materials associated with different properties and prediction years: chemical compounds and electrochemical properties including thermoelectricity (**a**), ferroelectricity (**b**) and photovoltaic capacity (**c**), with prediction year 2001; therapeutics and vaccines for COVID-19 in prediction year 2020 (**d**); and general disease–drug associations for prediction year 2001 (**e**). Precisions reported for general disease–drug associations are individual rates computed 19 years after the prediction year, but precisions are computed annually for electrochemical properties and monthly for COVID-19 efficacy (Extended Data Fig. 1c). The grey bars in **a**–**d** indicate the number of actual new discoveries each month or year of the prediction period. The curve labelled 'theoretical' in the case of COVID-19 represents predictions generated on the basis of protein–protein interaction networks by Morselli Gysi et al.[25]. Predictions accounting for the distribution of human experts are far superior to those that ignore it.

https://doi.org/10.1038/s41562-023-01648-z





https://doi.org/10.1038/s41562-023-01648-z

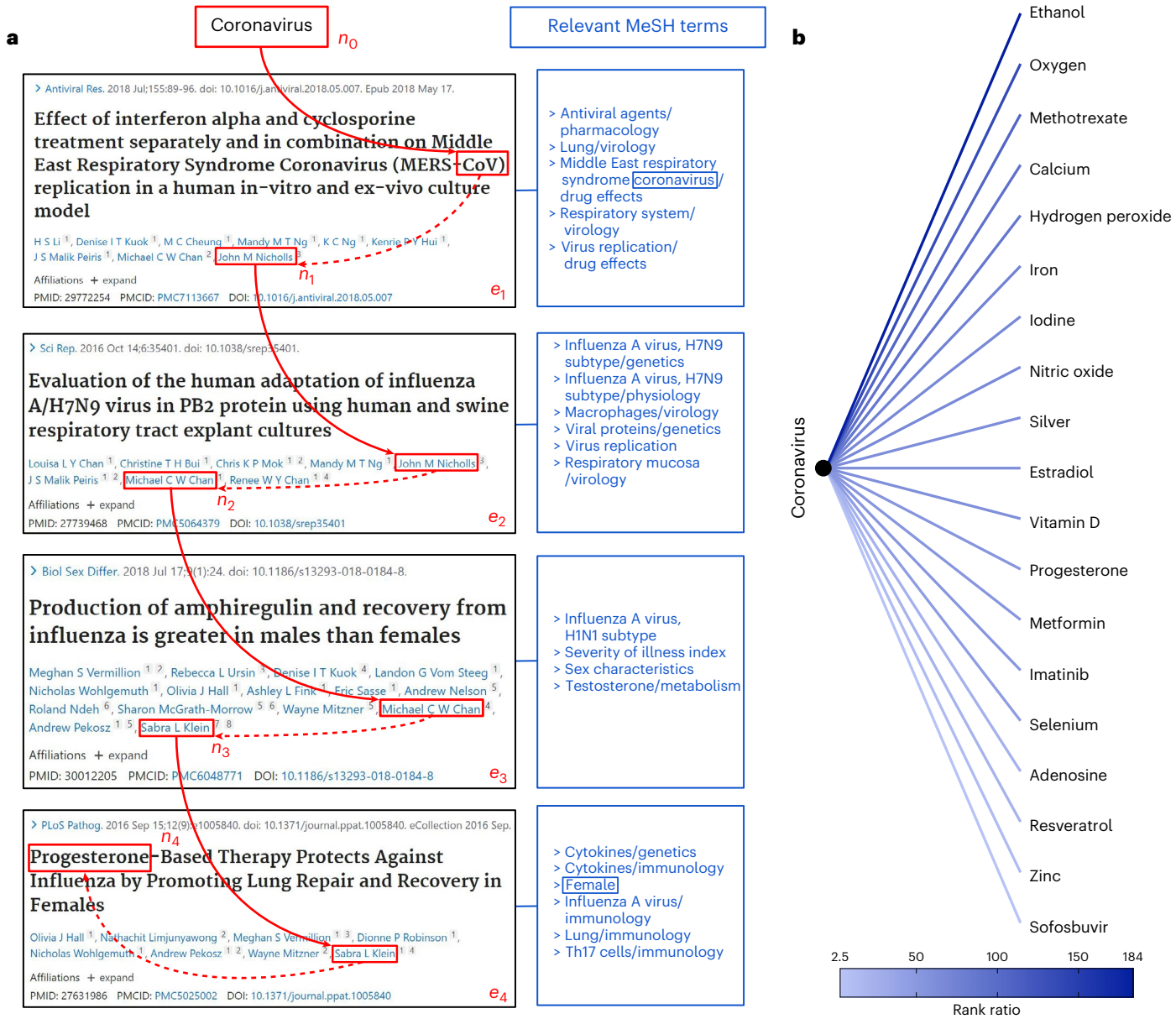


Fig. 3 | A prediction example of progesterone as a COVID-19 therapy. a, An example random walk from the property node 'Coronavirus' to the material node 'Progesterone', where selected hyperedges (papers) are shown in detail. Every article in this path is a hyperedge (denoted e_i in the ith step) connecting the prior to the subsequent node. The last article was cited by the University of Southern California clinical trial that investigated the effectiveness of progesterone for COVID-19 treatment. Relevant MeSH terms from the articles are shown to demonstrate their scope, indicating hints regarding the reasoning of human scientists championing the 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 in a published news interview[42]. **b**, True positive discovery predictions made by our human-accessible deepwalk algorithm, which were misclassified by the content-only predictor. Edge colours represent the ratio of $\text{rank}_{\text{Word2Vec}}$ to $\text{rank}_{\text{deepwalk}}$, where the numerator denotes the rank of the material in terms of our deepwalk scoring function that simulates the inferences made by human experts, and the denominator indicates the rank based on Word2Vec's scoring function that considers research content alone. Because we display only the true positives of the deepwalk algorithm, $\text{rank}_{\text{deepwalk}} \leq 50$ and $\text{rank}_{\text{Word2Vec}} > 50$ for all shown materials. A higher rank ratio reveals a larger disparity in the accuracy of algorithmic assessments. The largest ratio is associated with ethanol inhalation ($\text{rank}_{\text{deepwalk}} = 15$, $\text{rank}_{\text{Word2Vec}} = 2{,}762$), widely used in treating pulmonary oedema, and the smallest with sofosbuvir ($\text{rank}_{\text{deepwalk}} = 38$, $\text{rank}_{\text{Word2Vec}} = 102$), an antiviral used to treat hepatitis C.

protein interactions between COVID-19 and the pool of 3,948 relevant compounds from DrugBank[25], relevant information to which our model was blind.

The success of these COVID-19 predictions suggests how fast-paced research on COVID-19 therapies and vaccines increased the importance of scientists' prior research experiences and networks for the therapies and vaccines they would come to imagine, evaluate and champion in clinical trials. Consider the female hormone progesterone as a candidate material. Despite very few direct literature connections between 'Coronavirus' and 'Progesterone' before the rise of COVID-19, random walks from our method frequently walked the path between the two literatures through pre-COVID papers published in virology, immunology

https://doi.org/10.1038/s41562-023-01648-z

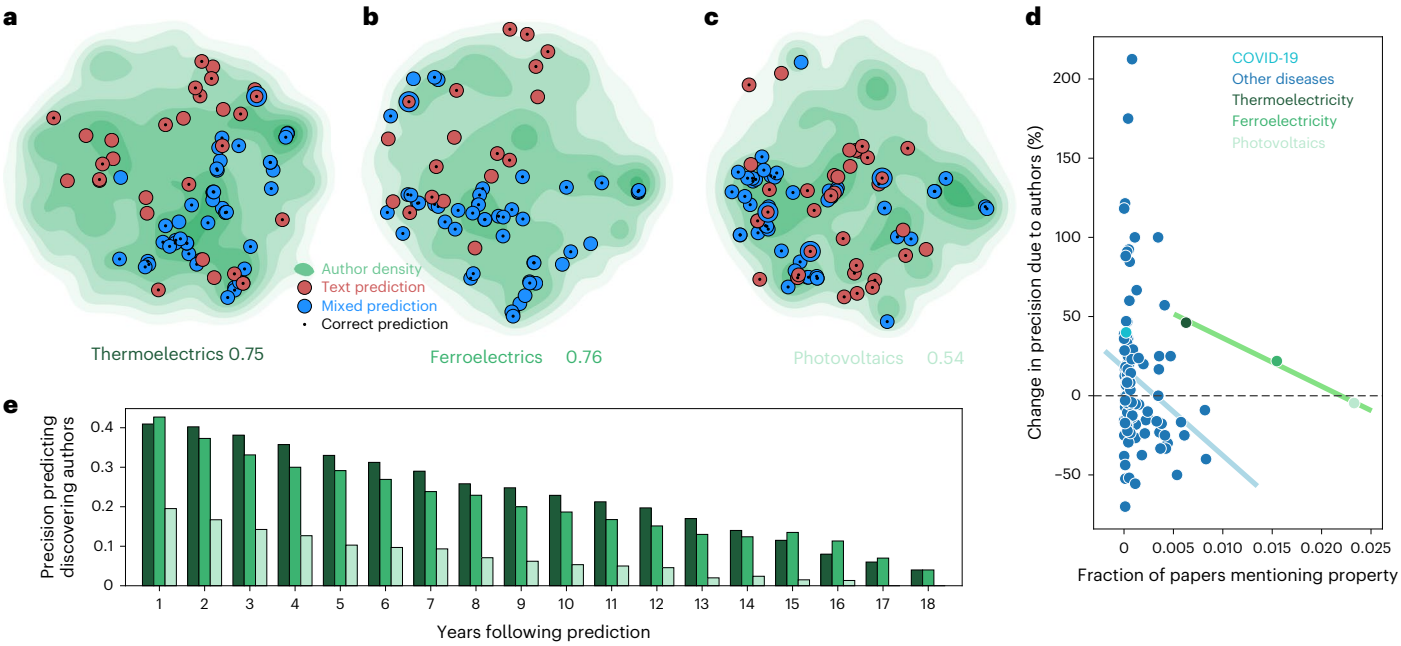


Fig. 4 | The contribution of human expert awareness for predicting discoveries and discoverers. a–**c**, 2D projections of human-accessible material predictions made by deepwalk (blue circles) and the content-exclusive Word2Vec model (red circles) for thermoelectricity (**a**), ferroelectricity (**b**) and photovoltaic capacity (**c**). Circles with centre dots indicate true positive predictions discovered and published in subsequent years, while empty circles represent false positives. The predictions are plotted atop the density of experts (topological 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 random walks generated across the hypergraph of published science (similar to our deepwalk procedure shown in Fig. 1 but without removing authors). The red circles are more uniformly distributed, but the blue circles concentrate near peaks of expert density. **d**, Precision shifts in predictions attributable to the inclusion of authors, defined as the percentage of precision change when switching from $\alpha \to \infty$ to $\alpha = 1$, plotted against the fraction of property-related papers within the literature. Higher density in the literature obviates the need for human author information. **e**, Precision rates for predicting discoverers of materials with electrochemical properties. The predictive models are built on the basis of two-step transitions between property and expert nodes with an intermediate material in the transition path. The bars show the average precision of human expert predictions for each year following prediction. Note that an expert can publish a discovery in multiple years. Total precision rates are also shown after each property, ignoring the repetition of discovering experts.

and studies regarding male/female characteristics of diseases and the female reproductive system (Fig. 3a and Extended Data Table 1). Shortly after the beginning of 2020 and in 2021, two clinical trials were initiated with similar motivations[26,27]: (1) the lower global death rate of women than that of men from COVID-19 and (2) the anti-inflammatory properties of progesterone that may moderate the immune system's overreaction to COVID-19 in men[26]. Our technique traced a pathway similar to the ones articulated explicitly by researchers sponsoring this trial: 75% of trial-cited papers, published within the five-year period preceding the prediction year we considered in building our hypergraph (2015–2019), were identified by our prediction model, and 60% of scientists authoring those studies were sampled in our random walk sequences. Progesterone and 18 other candidate materials were among the true positive predictions of our human-scientist-aware method that could not be captured by the content-only baseline (Fig. 3b). By contrast, only four true positives were exclusively made by content-only prediction (Extended Data Table 2), and these four materials had substantially fewer mentions than other predicted materials, confirming that human-aware prediction performs better when candidates are mentioned frequently in prior literature.

Human-sensitive prediction

Our predictive models use the distribution of discovering experts to successfully improve discovery prediction. To demonstrate this, we consider the time required by scientists to make a discovery starting from the prediction year. Materials cognitively close to the community of researchers who study a given property receive greater attention, and their relationships to that property are likely to be investigated, discovered and published earlier than those further from the community. In other words, the 'wait time' for discovery should be inversely proportional to the size of the expert population aware of both the property and the candidate material. We measure the size of this population by defining 'human expert density' between a property–material pair as the Jaccard index of two sets of human experts: those who mentioned the property and those who mentioned each candidate material in recent publications (Extended Data Fig. 3). This measures the overlap percentage between the property and material research communities. For all three electrochemical properties mentioned earlier, COVID-19 therapies and vaccines, and a majority of the 100 diseases we considered above, correlations between discovery date and expert density were negative, significant and substantial (Extended Data Fig. 4). This result confirms our hypothesis that materials receiving attention from a larger crowd of property experts are discovered sooner. Our predictive models efficiently leverage the hypergraph of past publications to incorporate these human expert densities (Extended Data Fig. 5). Similar results can be derived on the basis of embedding proximities: Fig. 4a–c illustrates how our predictions cluster atop density peaks in a joint embedding space of human experts and the materials they investigate. This further establishes that our human-aware approach is likely to select candidates more accessible to experts in the field.

We note that in some cases (for example, photovoltaics and silicosis), discovery prediction resulted in competitive performance when $\alpha \to \infty$, with the random walker ignoring authors and traversing only material nodes. Nevertheless, the human-ignorant algorithm performs well only when mentions of the targeted property are frequent in the literature (Fig. 4d). An abundance of property-related publications and their

https://doi.org/10.1038/s41562-023-01648-z

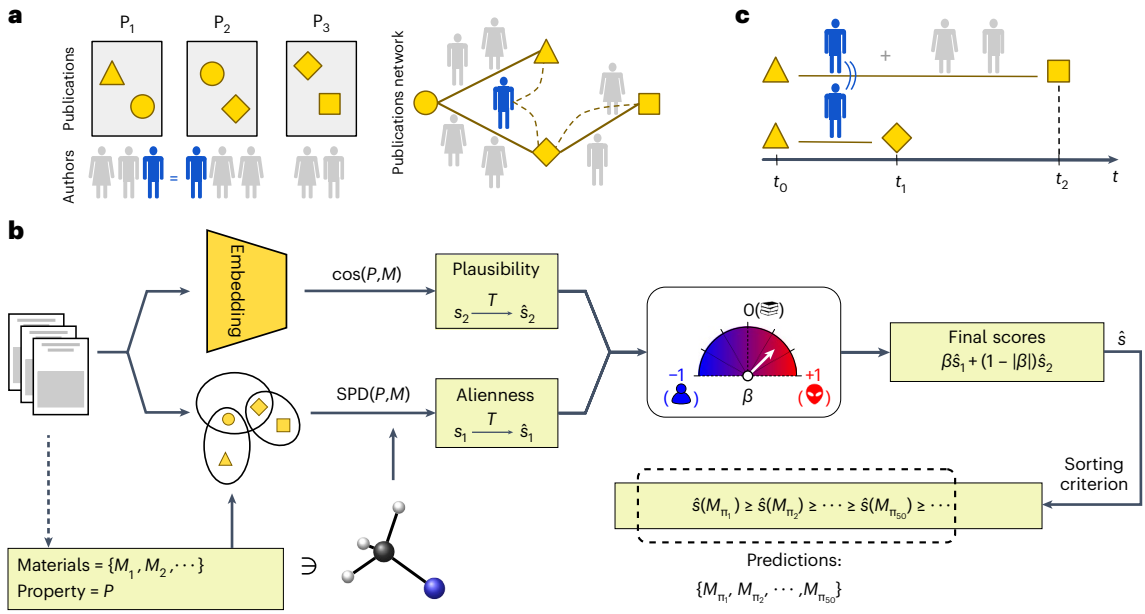


Fig. 5 | Motivation and design of our approach to generate complementary scientific predictions by avoiding human scientists. a, Distribution and overlap of experts investigating (and publishing on) topics represented by yellow geometric shapes. The dashed lines represent paths of more or less human cognitive availability between topics ('triangle', 'diamond' and 'square'). **b**, Overview of our complementary discovery prediction algorithm. Beginning with a scientific corpus and a targeted property, candidate materials are extracted from the corpus and used along with property mentions and authors to form the hypergraph. The algorithm follows two branches to compute plausibility from word embedding semantic similarities and human inaccessibility or 'alienness' from hypergraph SPDs. These two signals are combined after proper normalization and standardization through the mixing coefficient β to generate a prediction more or less complementary to the flow of human discovery (higher β values generate predictions that are more human inaccessible and so more complementary; lower β values generate ones that are more human accessible and so more in competition). Candidate materials are sorted on the basis of the resulting scores, and those with the highest rank are reported as proposed discoveries. **c**, Discovery wait times for relations between 'triangle'–'diamond' and 'triangle'–'square'. The time one needs to wait for a relationship to be discovered is proportional to the path length of human accessibility between the two relevant topics. The denser presence of human experts around the pair 'triangle'–'diamond' implies greater cognitive availability leading to earlier discovery and publication than that for 'triangle'–'square', where the connection requires a longer path.

availability to human scientists make the knowledge space more compact. This compactness enables scientists to infer future discoveries by simply taking in a redundant sample of papers, conference presentations or review articles without maintaining personal connections to relevant materials, properties or scientists. Expert awareness is critical for navigation when the knowledge is new or sparse. Even in these situations, however, the human-ignorant case $\alpha \to \infty$ performs much better in predicting discoveries than $\alpha = 0$ and other baselines, arguably because its inferences are local and in the vicinity of previous findings. This supports other evidence suggesting that scientists engage in localized search to make discoveries[6].

In addition to predicting discoveries, human-aware hypergraph proximities are able to predict discoverers most likely to publish discoveries on the basis of their unique configuration of research experiences and collaborations. Here discoverers are defined as all article authors associated with at least one discovery, disregarding author order. To identify potential discoverers of materials with a specific property, we compute the probability of random-walk transition from the targeted property to author nodes through a single intermediate material across our hypergraph (without rerunning the random-walk process). We then report potential discoverers to be those with the highest transition probabilities. Our calculations here are similar to the transition probabilities for discovery inferences described above, except that the destination nodes are authors and the intermediate nodes are materials (Supplementary Fig. 2). We evaluate these discoverer predictions against scientist authors who actually published discoveries following the prediction year. Calculating average precisions across 17 prediction years (2001 to 2017) for electrochemical properties, we found that 40% of the top 50 ranked potential authors became actual discoverers of thermoelectric and ferroelectric materials one year after prediction, and 20% of the top 50 predicted authors discovered novel photovoltaics (Fig. 4e). We also employ a method with slightly more subtlety to infer the identity of those predicted to discover a relationship between a targeted property and particular material (Extended Data Fig. 6).

Discoverer prediction serves as a validation of our main algorithm's operation—by implicitly identifying the people most likely to make the discovery. Strong precision values for both our discovery and discoverer predictions imply that discoveries are predominantly performed by individuals and teams familiar with and uniquely able to bridge otherwise disconnected topics and literatures. These results can also be viewed as an initial step towards predicting individuals and teams most likely qualified to achieve specific discoveries. They suggest the potential for a scientific service that recommends potential team members for recruitment on a targeted project.

Results on complementing human discoveries

We can use our model of human cognitive availability to not only approach and mimic but also avoid and complement the distribution of human experts. Human concept linkages are guided by previous discoveries and their discoverers (Fig. 5a). To build human-aware AI that proposes concept linkages unlikely to be imagined by scientists, we invert a measurement of human cognitive accessibility using shortest-path distances (SPDs) between pairs of conceptual nodes interlinked by authors in our mixed hypergraph. To rule out candidate hypotheses that lack scientific promise, we couple cognitive unavailability with a signal of scientific plausibility. This signal could be provided by the content of the published research literature and quantified with unsupervised knowledge embedding models[12]. Alternatively, a signal of scientific

https://doi.org/10.1038/s41562-023-01648-z

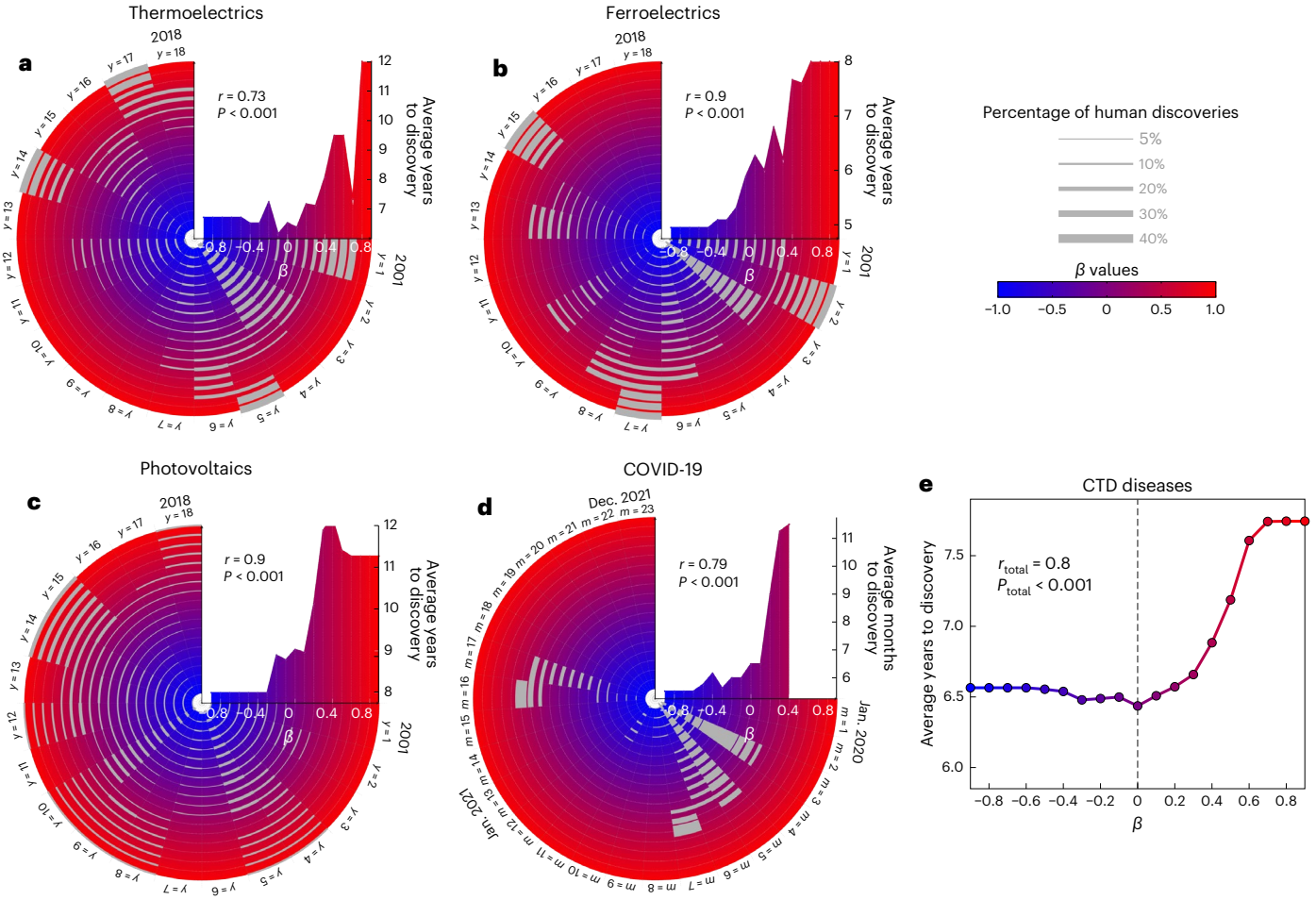


Fig. 6 | The wait time for published discoveries increases with human inaccessibility (higher β values). a–d, Average annual/monthly discovery wait times are shown as grey arcs, where thickness represents the percentage of materials discovered in the corresponding year/month. Each orbit is associated with a particular β value, with larger (redder) orbits representing larger β values and greater human inaccessibility as computed by our algorithm's human expert avoidance. The values we consider here vary between −0.8 (the smallest, bluest orbit) and 0.8 (the largest, reddest orbit). The plot in the upper right quarter of the orbits reveals the total average of discovery wait times including all years/months for each considered β value. **e**, Average wait times for discoveries across all human diseases (except COVID-19) from our experiments.

plausibility could be derived from theory-driven models of material properties. Here we use unsupervised knowledge embeddings for our algorithm, reserving theory-driven property simulations to evaluate the value and human complementarity of our predictions. Specifically, we forecast the scientific merit of any given hypothesis using the cosine similarity between embedding vectors of material and property nodes involved in that hypothesis.

Figure 5b provides a general overview of our algorithmic approach to identify discoveries that are both scientifically plausible and human inaccessible or complementary. After initializing with a pool of candidate materials extracted from literature, we compute human accessibility and scientific plausibility signals in an integrated fashion building on our prior analysis for generating human-like predictions. We use our unsupervised word embedding model over prior publications, measuring scientific relevance as cosine distance within the embedding. In parallel, we measure human accessibility by computing SPDs between the property and all materials across the hypergraph. We transform signals of plausibility and human accessibility into a unified scale and linearly combine them with a mixing coefficient β, which captures human complementarity (see the details in the Methods and Supplementary Information). With its expert awareness, we designed our algorithm to symmetrically generate either the most or least human-accessible hypotheses—those likely to compete with versus complement collective human capacity—depending on the sign of the mixing coefficient. Negative β values encourage high human accessibility, leading to predictions that mimic those of human experts in discovery. Positive values discourage human accessibility by producing hypotheses least similar to those human experts could plausibly infer, straddling socially disconnected but scientifically linked fields. At the extremes, $\beta = -1$ and 1 yield algorithms that generate predictions very familiar or very alien to human experts, regardless of scientific merit. Setting $\beta = 0$ (mid-range) implies exclusive emphasis on scientific plausibility, blind to the distribution of experts. This mode is equivalent to traditional discovery prediction methods exclusively based on previously published content. Intermediate positive β values balance the exploitation of relevant materials with the exploration of areas unlikely to be considered or connected by human experts. Each β value leads to a different model assigning a scalar score per material, which we use to sort candidate hypotheses. Materials with the highest resulting scores are reported as the algorithm's predictions corresponding to that specific β.

We evaluate our expert-avoiding algorithm with the same framework as before—that is, building our model using literature prior to a prediction year and evaluating inferred hypotheses on the basis of subsequent actual discoveries. In this section, we expand the drug

https://doi.org/10.1038/s41562-023-01648-z

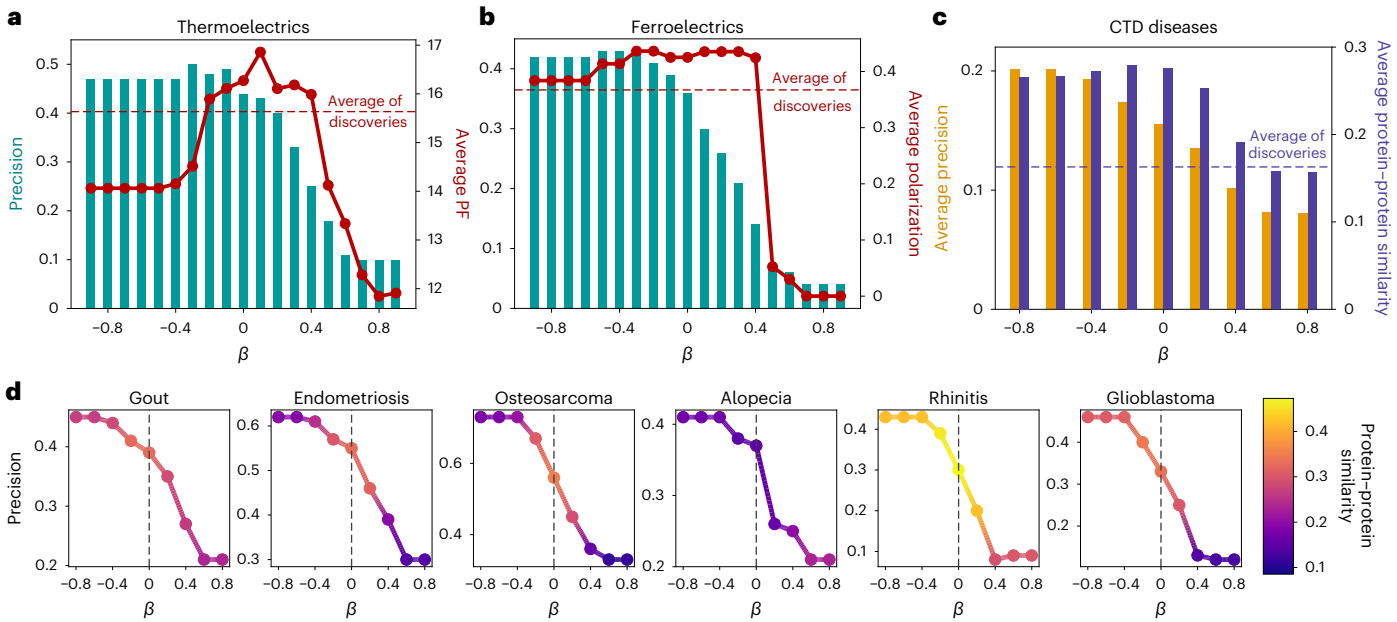


Fig. 7 | Precision in predicting human discovery falls before a comparable drop in theoretical expectations. a,**b**, The bars show the precision of complementary predictions with human-published discoveries, while the curves indicate theoretical expectations of first-principle simulations, which include average power factor (PF) for thermoelectricity (**a**) and spontaneous polarization for ferroelectricity (**b**). **c**, Precisions of complementary predictions for human discovery and average theoretical scores (that is, protein–protein interaction similarity scores) for therapeutic predictions. The horizontal dashed lines in all cases show the average theoretical scores computed for actual discoveries following the prediction year. **d**, Published discovery prediction versus average protein–protein similarity scores for six human diseases. The *y* axis indicates precision at predicting discovery, while the colour gradient represents the average theoretical scores for the predictions. In all cases, predictions of human discoveries fall much faster than theoretical expectations, which themselves are accessible to human experts and so represent a conservative estimate of scientific plausibility.

repurposing cases (properties) to include the treatment of 400 human diseases. We use the prediction year of 2001 for all properties except for COVID-19, for which we set the prediction year to 2020. The complementarity of these inferences is evaluated against human scientific knowledge by verifying (1) their distinctness from contemporary investigations and (2) their scientific promise. We anticipate that both features will simultaneously increase in ranges of β higher than those that characterize published science. Moreover, scientific merit will naturally reduce at the extremes of our interval [−1, 1], where the algorithm ignores the literature-based plausibility of candidate hypotheses. We expect to observe much higher plausibility in the intermediate ranges, which lead to strong complementarity for positive β values.

Evaluating discovered predictions

Our human-aware model is designed to allow us to dial up and down the degree to which predictions are similar to near-future human discoveries. As we increase β, the algorithm avoids human-accessible inferences that lie within regions of high expert density and focuses on candidate materials and properties that span disciplinary divides and evade human attention. As a result, we expect that generated hypotheses with large β values will (1) diverge from those pursued by the scientific community; (2) be less likely to become published; (3) if published, be discovered further into the future, after science has reorganized itself to consider them; and (4) manifest strong scientific performance as scientists conservatively crowd around areas of prior success. To verify these hypotheses, we first assess the discoverability of materials by computing the precision between our inferences and published discoveries. The results strongly confirm our expectation that materials inferred at higher β values are less discoverable by human scientists (Extended Data Fig. 7).

Moreover, materials distant from a given property in the hypergraph are expected to remain cognitively inaccessible to scientists in the property's proximity for longer (Fig. 5c). It takes more time for researchers in the field to broach knowledge gaps separating unfamiliar materials from valued properties. Among the inferences eventually discovered, we measure the discovery waiting time and expect to observe an increasing trend in wait times as we move from negative (human-competitive) to positive (human-complementary) β values in our predictions. Generating 50 hypotheses per β value and evaluating the resulting predictions indicates that for the majority of targeted properties, the average discovery wait times climb markedly when increasing β (Fig. 6) for energy-related chemical properties (Fig. 6a–c), COVID-19 prevention (Fig. 6d) and treatment for 70% of other human diseases (Fig. 6e). Averaging wait times across all human diseases manifests a clear increasing trend. For some cases, such as COVID-19 (Fig. 6d), none of the complementary predictions made with positive β values (larger than 0.4) come to be discovered by humans within the time frame we examine.

Evaluating undiscovered predictions

To evaluate the scientific merit of our algorithm's predictions, including those that remain undiscovered within the study period, we require data beyond the extant literature. Such hypotheses necessarily grow to comprise the vast majority of cases for large values of β. If science were an efficient market and experts optimally pursued scientific quality, then in human-avoiding high β hypotheses, we would observe a proportional decline in scientific promise and efficacy. In contrast, if scientists crowd together along the frontier of scientific possibility and their continued efforts yield diminishing marginal returns, we might observe an increase in promise as we move beyond them.

To evaluate the merit of undiscovered scientific inferences, we utilize first principles or data-driven models derived uniquely for each property on the basis of well-established theoretical principles within the field. Similar to our algorithms, such models also assign real-valued scores to candidate materials as a measure of their potential for possessing the targeted properties. These computations may be carried out without regard for whether materials have yet been discovered,

https://doi.org/10.1038/s41562-023-01648-z

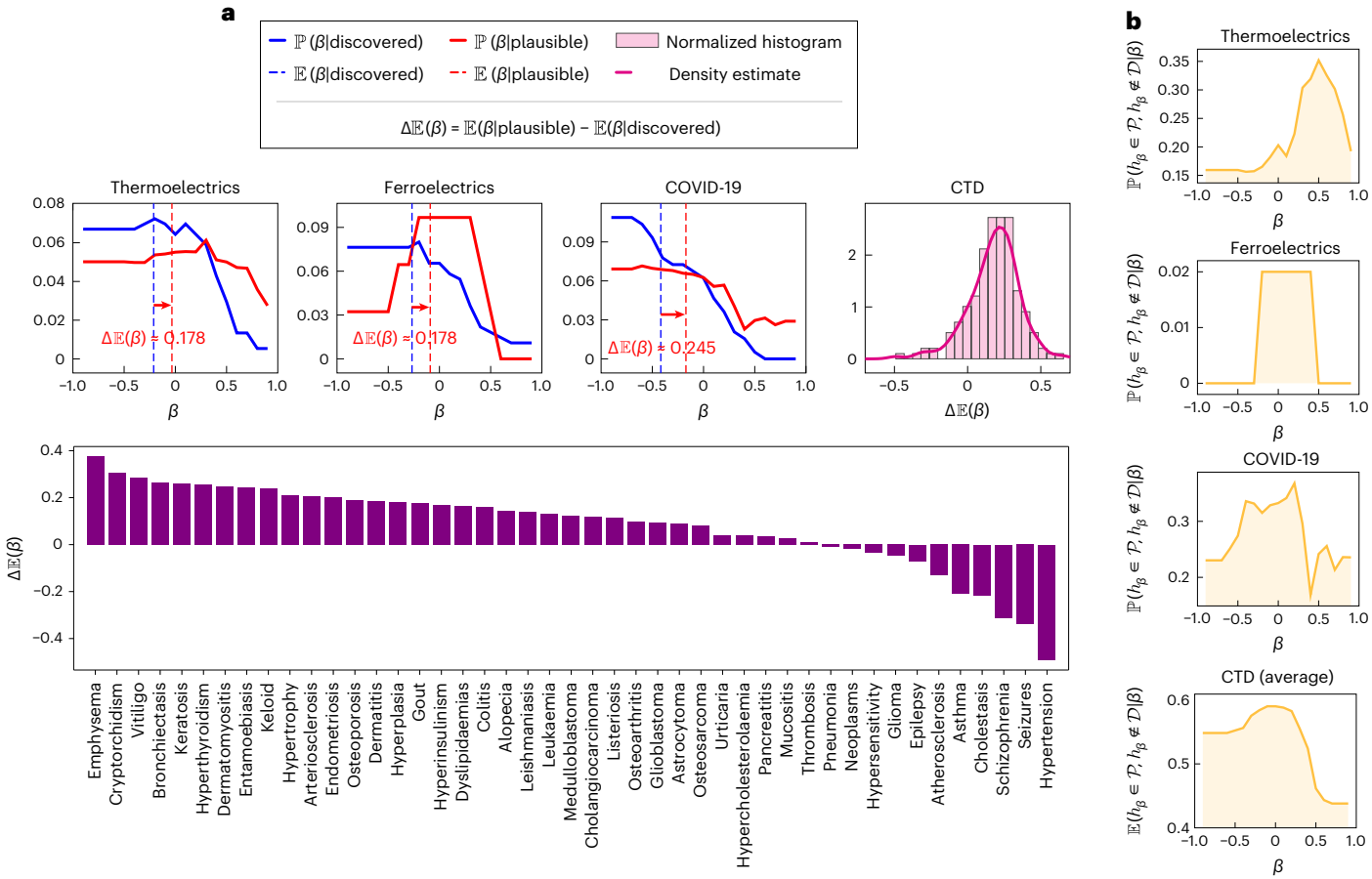


Fig. 8 | Complementary AI predictions outperform human discoveries. a, We formalize and estimate the expectation gap for properties with first-principle or data-driven theoretical plausibility scores. We plot the conditional distributions $\mathbb{P}(\beta|\text{plausible})$ and $\mathbb{P}(\beta|\text{discoverable})$ separately for materials with valuable energy-related properties of thermoelectricity and ferroelectricity, and for therapeutics fighting COVID-19 and hundreds of other human diseases (shown collectively in a normalized histogram). The first row demonstrates that the quality of our complementary hypotheses improves or maintains beyond those materials accessible to human scientists, as discovered and later published by them. The second row shows the individual gaps between discoverability by human scientists and plausible performance based on theoretical and data-driven models for a subset of human diseases. **b**, The joint probability of simultaneous undiscoverability and plausibility for different values of the alien parameter β, where low values ($\beta = -1$) indicate mimicry of human discovery; zero values ($\beta = 0$) indicate its ignorance of human discovery, attending only to the research literature; and positive values ($\beta = 1$) indicate its avoidance of human discoverability. Here, h_β denotes a randomly selected prediction generated with β, $\mathcal{D}$ represents the set of candidates discoverable by human scientists and $\mathcal{P}$ represents those scientifically plausible.

making them a suitable scoring function for evaluating undiscovered hypotheses. We produced such scores for approximately 45% of the properties we considered above using models based on first-principles understandings of the phenomenon or models based on databases curated with high-throughput protein screens. To evaluate thermoelectric promise, we used power factor as 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[28,29]. To evaluate ferroelectricity, estimates of spontaneous polarization obtained through symmetry analysis and relevant theoretical equations serve as a reliable metric[30]. For human diseases including COVID-19, proximity between disease agents (for example, SARS-CoV-2) and candidate compounds in protein–protein interaction networks suggests the likelihood that a material will recognize and engage with the disease agent[25] (for more details on how these theoretical scores are derived, see the Supplementary Information). We note that scores based on first-principles equations or simulations represent conservative estimates of scientific merit as they are based on widely accepted, scientist-crafted and theory-inspired models. Because these scores are potentially available to scientists in the area, they may be considered when guiding investigations such that experiments on these unevaluated hypotheses often lead to promising results. Nevertheless, in what follows we show that modestly positive β values manifest a marked improvement even on this conservative measure of quality.

We expect the average theoretical scores of hypotheses to decay sharply at the extremes of the β range [−1, 1], as at those points the algorithm ignores the merit signal, putting it at higher risk of generating scientifically irrelevant (or absurd) proposals. We expect, however, that this decay will occur more slowly than the decrease in hypothesis discovery and publication, which implies the existence of a β interval where proposals are not discoverable but highly promising—an ideal operating region for the generation of hypotheses that complement those from the human scientific crowd. To verify this, we contrasted changes in average theoretical scores with the discoverability of generated hypotheses for various β values. As illustrated in Fig. 7a–c, discoverability decreases near the transition of β from negative to positive values, but its decay is much sharper than average theoretical scores, which do not collapse until nearly $\beta = 0.4$. This holds for electrochemical properties and the majority of diseases. The results for certain individual diseases can be seen in Fig. 7d (for the full set of results, see Extended Data Fig. 8 and Supplementary Table 1). Moreover, note that for the cases investigated, the average theoretical

scores for inferred hypotheses grow higher than the average theoretical scores for actual, published discoveries (the dashed lines) before their eventual decay at high β values. For certain properties such as thermoelectricity or therapeutic efficacy against the disease alopecia, the theoretical merit of our inferences exhibits striking growth from negative (scientist-mimicking) to positive (scientist-avoiding) hypotheses, suggesting strong diminishing returns to following these scientific crowds, whose overharvested fields have become barren for new discovery.

To further compare the decay rate of discoverability and theoretical scores, we define and compute an expectation gap to measure the distance between expected values for two conditional distributions over β. These two conditionals are defined as two likelihoods over β given that a randomly selected prediction with that β value is (1) identified as promising on the basis of its corresponding first-principle score and (2) discoverable—that is, studied and published by a scientist following the prediction year (for details see the Methods and Supplementary Information). A positive expectation gap indicates that increasing β will preserve the quality of predictions while making them more complementary to human hypotheses. As shown in Fig. 8a, the vast majority of properties considered in this section yield substantially positive expectation gaps. Building on this, we use a probabilistic model to assess the complementarity of our algorithm's prediction with those of the scientific community for any value of β. This is done by explicitly computing the joint probability that a randomly selected prediction is plausible in terms of the desired property and beyond current scientists' scope of research (Supplementary Information). These probabilities specify the optimal β value to balance exploitation and exploration in augmenting collective human prediction. The results in Fig. 8b indicate that the optimal point varies for different properties, but one can distinguish the range 0.2–0.3 as the most consistently promising interval. In this interval, hypotheses are very unlikely to come from the scientific community but are very likely to yield successful scientific results.

Discussion

We demonstrate the power of incorporating human awareness into AI systems for accelerating future discovery. Our models succeed by directly predicting human discoveries and the human experts who will make them, yielding up to 400% improvement in prediction precision. These findings offer support for the influence of the human experience and social connection inscribed by our research hypergraph in driving scientific advance. This suggests that the search underlying materials and medical advance is dominated by local exploitation of the familiar over novel exploration of the unknown. Moreover, by tuning our algorithm to avoid the crowd, we generate promising hypotheses that are unlikely to be imagined, pursued or published without machine recommendation for years into the future. By identifying and correcting for collective patterns of human attention, formed by field boundaries and institutionalized education, these models complement the contemporary scientific community. This demonstrates that connectivities in our expert-aware hypergraph are useful not only for predicting and accelerating human discoveries in the near future but also for inferring disruptive discoveries that could be imagined by scientists only in the distant future.

Our analysis examined a limited space of scientific relationships—those involving a material possessing a valuable energy or therapeutic property. Many other scientifically meaningful relationships lie beyond this syntax, such as identity (that is, *a* is a *b*), composition (that is, *a* is a part of *b*), or any specific physical or logical relationship (for example, *a* chemically reacts with *b*; *a* genetically upregulates *b*). Using a hypergraph formalism, we could extend such relations beyond logical triples that connect a simple concept pair to larger sets of concepts connected by more complex relations. Another limitation involved our singular consideration of co-authorship as the relationship affecting the distribution of expertise. One could consider other relationships, such as scientist collocation within an institution, conference attendance or geographical proximity. Moreover, there are opportunities to technically improve our approach, such as combining content and human-aware information to amplify prediction accuracy, or inferring and exploiting the body of negative knowledge in science where researchers know that certain scientific claims are false[11,31].

Despite these limitations, our investigation underscores the power of incorporating human and social factors to produce AI that complements rather than substitutes for human expertise. Successful scientists competitively factor and follow the momentum of advances made by researchers around them in identifying the frontiers of science. When AI hypothesis generation is made aware of human expertise, it can accelerate discovery and liberate human scientists to steer science and technology in novel directions. Our system and its recommendations raise ethical concerns; they could be used as a 'scoop-machine' to leapfrog human scientists and seize on answers that they might otherwise ask and answer next. This would accelerate science but could augment some scientists' capacity at the expense of others. Such a concern would attenuate when scientific recommendation engines became ubiquitous, however, like recommendations for internet and social media searches. Moreover, we demonstrate how awareness of human scientific expertise could be used not only to mimic but to avoid it, generating insights that punctuate the current flow of discovery[32].

Our investigation also reveals the influence of human scientific institutions that crowd scientists along a shared frontier of likely discoveries. The success of our 'alien' or complementary hypotheses suggests that scientific departments and disciplines limit productive exploration and point to opportunities that could improve human prediction by reformulating science education for discovery. Insofar as research experiences and relationships condition the questions scientists investigate, education tuned to discovery might conceive of each student as a new experiment, recombining knowledge and opportunity in novel ways. However, we can build AI that reaches further. Our analysis demonstrates the benefit that comes from modelling human reasoning to explicitly complement it. In accounting for the complete distribution of human scientific experience and exposure, we can design AI systems that race with rather than against the scientific community to expand the scope of collective imagination and discovery.

Methods

Experiments and data collection

Each discovery prediction experiment consists of a target property and a pool of materials, where the materials are scored by a predictor and the 50 materials with the highest scores are selected as predictions. Each predictor scores an individual material through computing its similarity with the property. Similarity metrics for our hypergraph-based predictors are the transition probability between material and property nodes with one and two intermediate author nodes (hence two- and three-step transitions—that is, $s = 2$ and $s = 3$), and cosine similarity in the deepwalk embedding space. The former can be calculated through Bayes' rule without the need for generating random walks, but the latter requires an explicit set of random walk sequences over our hypergraph. The similarity metric from the replicated content-only baseline is the cosine similarity in the embedding space of a Word2Vec model trained on the corpus of publications produced before the prediction year. The corpus of publications and ground-truth discoveries are prepared differently for each set of properties and potential materials.

Our testbed consisted of two datasets: for the energy-related properties, we used a collection of ~1.5 million articles published between 1937 and 2018 classified by Tshitoyan et al. as related to inorganic materials[15], and for the diseases, we utilized the MEDLINE database, which includes more than 28 million articles published in various biomedical fields over the span of more than two centuries. Creating our hypergraph required us to have access to disambiguated authors

for all articles. We downloaded the database related to inorganic materials using Scopus API provided by Elsevier (https://dev.elsevier.com/), which readily assigns unique codes to distinct authors. To author-disambiguate the MEDLINE database, we used the disambiguation results provided by the PubMed Knowledge Graph[33], which were obtained by combining information from the Author-ity disambiguation of PubMed[34] and the more recent Semantic Scholar database[35]. This integrative method has a performance comparable to each of its individual components: 98.09% F_1-score, 98.62% precision and 97.56% recall. For this dataset, we restricted our experiments to 27.5 million papers with available abstracts, metadata (publication year) and disambiguated authors.

For energy-related properties, we extracted the pool of chemical compounds from the collected 1.5 million articles using Python Materials Genomics[36] and direct rule-based string processing. Material–property association was defined in terms of co-occurrence of materials with property-related keywords. First-time co-occurrences were considered ground-truth discoveries, following the replicated prior work[15]. 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 (that is, dropping complex names containing several numerical components). We chose 100 (or 400, when avoiding experts) diseases from the CTD[24] with the largest number of relevant drugs from our drug pool. To build our hypergraph, we searched for names of drugs and diseases in MEDLINE to detect their occurrence within papers. Ground-truth relevant drugs for the selected diseases (except COVID-19) were extracted from associations curated by CTD. The discovery date for each of the disease–drug associations was set to the earliest publication reported by CTD for curated relevance. We ran separate prediction experiments for each individual disease, where we define the property as drug efficacy in treating or preventing the selected disease. The same pool of drugs and corpus of articles were used for the case of COVID-19, where the ground-truth relevance of drugs to COVID-19 were identified on the basis of their involvement in COVID-19-related studies reported by ClinicalTrials.org in or after 2020, regardless of the studies' results, following the compared work by Morselli Gysi et al.[25]. The date of discovery for each relevance was set to the date the corresponding study was first posted, and if the drug was involved in multiple trials, we considered the earliest. There have been 6,280 trials posted as of 5 August 2021 (ignoring 37 trials dated before 2020), which included 279 drugs from our pool (~7%) included in their study designs.

Hypergraph random walks

In practice, research and co-authoring that occurred long before the time of prediction are unlikely to be cognitively available, socially accessible or perceived as being of continuing relevance. We therefore restrict our prediction experiments to use literature produced in the most recent five years prior to the year of prediction. For alternative time windows, the magnitude of precision curves slightly shifted, but their trend and ordering remained the same (Supplementary Fig. 4). For each property, we took 250,000 non-lazy, truncated random walks with and without α-modified sampling distribution sequences. All walks start from the property node and end either after 20 steps or after reaching a dead-end node with no further connections. The α-modified sampling algorithm is implemented as a mixture of two uniform distributions over authors and materials such that the mixing coefficient assigned to the latter is α times the coefficient of the former. Hence, α is the ratio of the probability of selecting a material to the probability of selecting an author node (see the Supplementary Information for more details). We tried three values for this parameter in our experiments: $\alpha = 1$, which implies an equal probability of sampling authors and materials; $\alpha \to \infty$, which samples only materials; and $\alpha = 0$, which samples only authors. The author-only mode yielded much weaker performance than the other two, and we do not include it in our results. This implies that mere networking with other human experts without reading and researching the literature does not typically lead to discoveries in practice. A further perturbation analysis of α showed that increasing it to values larger than 1 (for example, 10) is less harmful to precision levels than decreasing it below 1 (for example, 0.5). In other words, the balance point leads to the highest performance (that is, $\alpha = 1$), but if one breaks the balance between researching (for example, Googling and reading related research papers) and networking with nearby scientists, overemphasizing research exploration harms prediction less than overemphasizing social networking in predictions of knowledge discovery (see Supplementary Fig. 5 for a more thorough sensitivity analysis of our algorithm with regard to α).

Relevance metrics

Once the random walk sequences are drawn, we can compute our two hypergraph-induced similarities. Multi-step transition probabilities are directly computed from transition matrices using Bayes' rules and Markovian assumptions (Supplementary Information). Calculating probabilities for two- and three-step transitions from properties to materials requires us to sum the probability of all meta-paths with the form PAM and PAAM, where P, A and M stand for property, author and material nodes, respectively[37]. Alternatively, the meta-path that we considered for discoverer prediction was PMA. For our deepwalk representation, we trained a skipgram Word2Vec model with hyperparameter settings similar to the content-only prior work we replicated[15], including an embedding dimensionality set to 200. One exception is the number of epochs, which we reduced from 30 to 5. The size of vocabulary produced by deepwalk sampling is substantially smaller than the number of distinct words from literature. As a result, deepwalk training required less effort and lower iterations to capture the underlying internode relationships. Also note that deepwalk embedding similarity is more global than the transition probability metric, provided that the length of our walks (set to 20) is higher than the number of transition steps (2 or 3). Moreover, it is more flexible since the walker's edge selection probability distribution can be easily modified to explore the network structure more deeply[38]. Nevertheless, because the deepwalk Word2Vec is trained using a window of only length 8, only authors and materials that might find each other through conversation, seminars or conferences would be near one another in the resulting vector space.

We also ran our prediction experiments after replacing the deepwalk representation with a graph convolutional neural network. We used the Graph Sample and Aggregate (GraphSAGE) model[39] with 400 and 200 as the dimensionality of hidden and output layers, respectively, with rectified linear units as nonlinear activations in the network. Convolutional models require feature vectors for all nodes, but our hypergraph is inherently featureless. We therefore 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 the 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 the entities with the highest cosine similarities to the desired property. 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 author nodes (Supplementary Information). Running the convolutional model on energy-related materials and properties yielded precisions of 62%, 58% and 74% on the full graph and 48%, 50% and 58% on the authorless graph for thermoelectricity, ferroelectricity and photovoltaics, respectively. These results show a pattern similar to those obtained through the deepwalk model, although with a somewhat smaller margin due to the use of Word2Vec-based feature vectors, which limited the domain of exploration by the resulting embedding model to within proximity of the baseline.

Complementary hypotheses generation

Our predictor consists of two scoring functions. The first measures the human inaccessibility (that is, alienness) of candidate materials via SPDs between the nodes corresponding to the targeted property and the candidates. The second measures scientific plausibility through the semantic cosine similarities of their corresponding keywords. For this purpose, we use our Word2Vec baseline pretrained over the literature (collected on inorganic materials for energy-related properties and MEDLINE for the diseases) produced prior to the prediction year. We combine the alienness and plausibility scores with a mixing coefficient, denoted by β, adjusting their contributions to obtain a final score for the candidate. The plausibility component yields continuous scores distributed close to Gaussian, whereas the alienness component offers unbounded ordinal SPD values. Simple normalization methods are insufficient to combine scores with such distinct characteristics. As a result, we first standardize the two scores to a unified scale by applying the van der Waerden transformation[41], followed by a Z-score normalization. The final step includes taking the weighted average of the resulting Z-scores with weights depending on β (see the Supplementary Information for more details).

We want our predictor to infer undiscoverable yet promising hypotheses. Setting β to a more positive value makes predictions less familiar and more alien—that is, less discoverable. Moreover, increasing β to the positive extreme (that is, +1) excludes scientific merit from the algorithm's objective in materials selection. Hence, increasing β causes both the discoverability and the plausibility of predictions to decay. What matters to us is that plausibility decreases more slowly than discoverability, suggesting that the predictor achieves a close-to-ideal state where predictions are simultaneously alien and promising. To verify this with a single number, we define the expectation gap criterion, computed as the difference between the expected values of the following two distributions over β: $\mathbb{P}(\beta|\text{plausible})$ and $\mathbb{P}(\beta|\text{discoverable})$. The terms 'plausible' and 'discoverable' on the conditional sides could be substituted by the precise statements 'a randomly selected inferred hypothesis is theoretically plausible' and 'a randomly selected inferred hypothesis is discoverable'—it will be published by scientists. While we know that both of these distributions reduce as β approaches +1, the expectation gap measures any positive shift in the mass of $\mathbb{P}(\beta|\text{plausible})$ against $\mathbb{P}(\beta|\text{discoverable})$. The likelihood of discovery, $\mathbb{P}(\beta|\text{discoverable})$, can be estimated through an empirical distribution of predictions discovered and published. Scientific plausibility can be estimated by leveraging properties' theoretical scores obtained from prior knowledge and first-principles equations and data from relevant fields. We estimate $\mathbb{P}(\beta = \beta_0|\text{plausible})$ in two steps: (1) converting theoretical scores to probabilities and (2) computing weighted maximum likelihood estimates of $\mathbb{P}(\beta = \beta_0|\text{plausible})$ given a set of predictions generated by our algorithm operated with β_0 (see the Supplementary Information for details). We restrict experiments in this section to only those properties for which we could obtain a reliable source of theoretical scores (see the Supplementary Information for details of the scores): thermoelectricity, ferroelectricity, COVID-19 and 175 other human diseases (178 of 404 total properties).

Finally, note that expectation gaps and average discovery dates (described above) say nothing about the β interval most likely to lead to better complementarity. We introduce an additional probabilistic criterion for this purpose, which explicitly and jointly models these two features and computes their likelihood for various β values, $\mathbb{P}(\text{undiscoverable, plausible} \mid \beta)$. One can use this distribution to screen the best operating point for complementary AI (Supplementary Information).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The DOIs of papers used for the electrochemical properties together with the PubMed identifiers of the MEDLINE entries used in our experiments can be found in our GitHub repository: https://github.com/jsourati/accelerate-discoveries. The abstracts of papers for electrochemical properties could not be shared due to copyright issues, but MEDLINE abstracts are accessible through their identifiers from the PubMed website. Source data are provided with this paper.

Code availability

All code for our algorithms can be found in the following GitHub repository: https://github.com/jsourati/accelerate-discoveries.

References

1. Khadherbhi, S. R. & Babu, K. S. Big data search space reduction based on user perspective using map reduce. *Int. J. Adv. Technol. Innov. Res.* **7**, 3642–3647 (2015).
2. Sanchez-Lengeling, B. & Aspuru-Guzik, A. Inverse molecular design using machine learning: generative models for matter engineering. *Science* **361**, 360–365 (2018).
3. Smalley, E. AI-powered drug discovery captures pharma interest. *Nat. Biotechnol.* **35**, 604–605 (2017).
4. Teruya, E., Takeuchi, T., Morita, H., Hayashi, T. & Ono, K. ARTS: autonomous research topic selection system using word embeddings and network analysis. *Mach. Learn. Sci. Technol.* **3**, 025005 (2022).
5. Shi, F., Foster, J. G. & Evans, J. A. Weaving the fabric of science: dynamic network models of science's unfolding structure. *Soc. Netw.* **43**, 73–85 (2015).
6. Singer, U., Radinsky, K. & Horvitz, E. On biases of attention in scientific discovery. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btaa1036 (2020).
7. Tversky, A. & Kahneman, D. Availability: a heuristic for judging frequency and probability. *Cogn. Psychol.* **5**, 207–232 (1973).
8. Evans, J. S. B. T. *Bias in Human Reasoning: Causes and Consequences* (Psychology Press, 1989).
9. Ehrlinger, J., Readinger, W. O. & Kim, B. in *Encyclopedia of Mental Health* 2nd edn (ed. Friedman, H. S.) 5–12 (Academic Press, 2016).
10. Chadwick, A. T. & Segall, M. D. Overcoming psychological barriers to good discovery decisions. *Drug Discov. Today* **15**, 561–569 (2010).
11. Rzhetsky, A., Foster, J. G., Foster, I. T. & Evans, J. A. Choosing experiments to accelerate collective discovery. *Proc. Natl Acad. Sci. USA* **112**, 14569–14574 (2015).
12. Mikolov, T., Yih, W.-T. & Zweig, G. Linguistic regularities in continuous space word representations. In *Proc. 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (eds Vanderwende, L. et al.) 746–751 (Association for Computational Linguistics, 2013).
13. Perozzi, B., Al-Rfou, R. & Skiena, S. DeepWalk: online learning of social representations. In *Proc. 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (eds Macskassy, S. et al.) 701–710 (Association for Computing Machinery, 2014).
14. Chitra, U. & Raphael, B. Random walks on hypergraphs with edge-dependent vertex weights. In *Proc. 36th International Conference on Machine Learning* (eds Chaudhuri, K. & Salakhutdinov, R.) 1172–1181 (PMLR, 2019).
15. Tshitoyan, V. et al. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature* **571**, 95–98 (2019).
16. Burger, B. et al. A mobile robotic chemist. *Nature* **583**, 237–241 (2020).
17. Swanson, D. R. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspect. Biol. Med.* **30**, 7–18 (1986).

18. Swanson, D. R. Medical literature as a potential source of new knowledge. *Bull. Med. Libr. Assoc.* **78**, 29–37 (1990).
19. 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).
20. Evans, J. & Rzhetsky, A. Machine science. *Science* **329**, 399–400 (2010).
21. 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. *Am. J. Med.* **86**, 158–164 (1989).
22. 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–E112 (2016).
23. Chu, J. S. G. & Evans, J. A. Slowed canonical progress in large fields of science. *Proc. Natl Acad. Sci. USA* **118**, e2021636118 (2021).
24. Davis, A. P. et al. The Comparative Toxicogenomics Database: update 2019. *Nucleic Acids Res.* **47**, D948–D954 (2019).
25. Morselli Gysi, D. et al. Network medicine framework for identifying drug-repurposing opportunities for COVID-19. *Proc. Natl Acad. Sci. USA* **118**, e2025581118 (2021).
26. 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* https://doi.org/10.1016/j.chest.2021.02.024 (2021).
27. Estradiol and progesterone in hospitalized COVID-19 patients https://clinicaltrials.gov/ct2/show/NCT04865029 (2022).
28. 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).
29. Ricci, F. et al. An ab initio electronic transport database for inorganic materials. *Sci. Data* **4**, 170085 (2017).
30. 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).
31. Belikov, A. V., Rzhetsky, A. & Evans, J. Prediction of robust scientific facts from literature. *Nat. Mach. Intell.* **4**, 445–454 (2022).
32. Sourati, J. & Evans, J. Complementary artificial intelligence designed to augment human discovery. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2207.00902 (2022).
33. Xu, J. et al. Building a PubMed knowledge graph. *Sci. Data* **7**, 205 (2020).
34. Torvik, V. I. & Smalheiser, N. R. Author name disambiguation in MEDLINE. *ACM Trans. Knowl. Discov. Data* **3**, 1–29 (2009).
35. Ammar, W. et al. Construction of the literature graph in Semantic Scholar. In *Proc. 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 3 (Industry Papers)* 84–91 (Association for Computational Linguistics, 2018).
36. 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).
37. Sun, Y., Han, J., Yan, X., Yu, P. S. & Wu, T. PathSim: meta path-based top-K similarity search in heterogeneous information networks. *Proc. VLDB Endow.* **4**, 992–1003 (2011).
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 *Proc. 31st International Conference on Neural Information Processing Systems* (eds Guyon, I. et al.) 1025–1035 (Curran Associates, 2017).
40. Kipf, T. N. & Welling, M. Variational graph auto-encoders. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1611.07308 (2016).
41. Coakley, C. W. Practical nonparametric statistics. *J. Am. Stat. Assoc.* **95**, 332–333 (2000).
42. Schaffer, R. Study examines progesterone to reduce inflammation in COVID-19. *Healio—EndocrineToday* https://www.healio.com/news/endocrinology/20200507/study-examines-progesterone-to-reduce-inflammation-in-covid19 (7 May 2020).

Acknowledgements

We thank our funders for their generous support: the National Science Foundation (grant no. 1829366), the Air Force Office of Scientific Research (grant nos. FA9550-19-1-0354 and FA9550-15-1-0162) and DARPA (grant no. HR00111820006). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. We thank L. Barabasi and D. Morselli Gysi for helpful data related to their network-based forecast of COVID-19 drugs and vaccines with protein–protein interactions[25], and A. Jain, V. Tshitoyan and A. Dunn for sharing data and code to help replicate their work on unsupervised word embeddings and latent knowledge about materials science[15]. We also thank the 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.

Author contributions

J.S.: conceptualization, methodology, software, validation, investigation, writing—original draft and visualization. J.A.E.: conceptualization, methodology, writing—original draft, visualization and funding acquisition.

Competing interests

The authors declare no competing interests.

Additional information

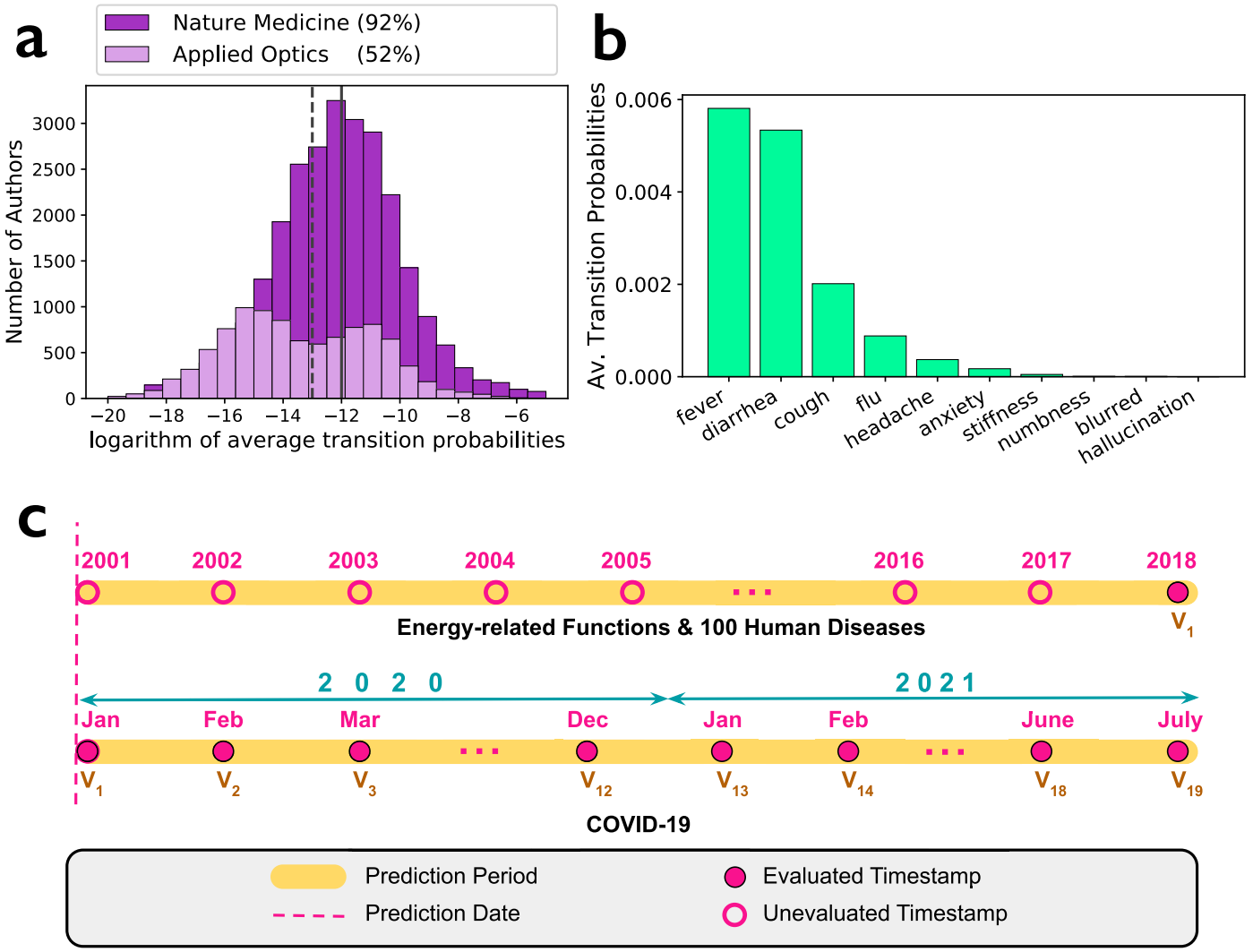
Extended data is available for this paper at https://doi.org/10.1038/s41562-023-01648-z.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-023-01648-z.

Correspondence and requests for materials should be addressed to James A. Evans.

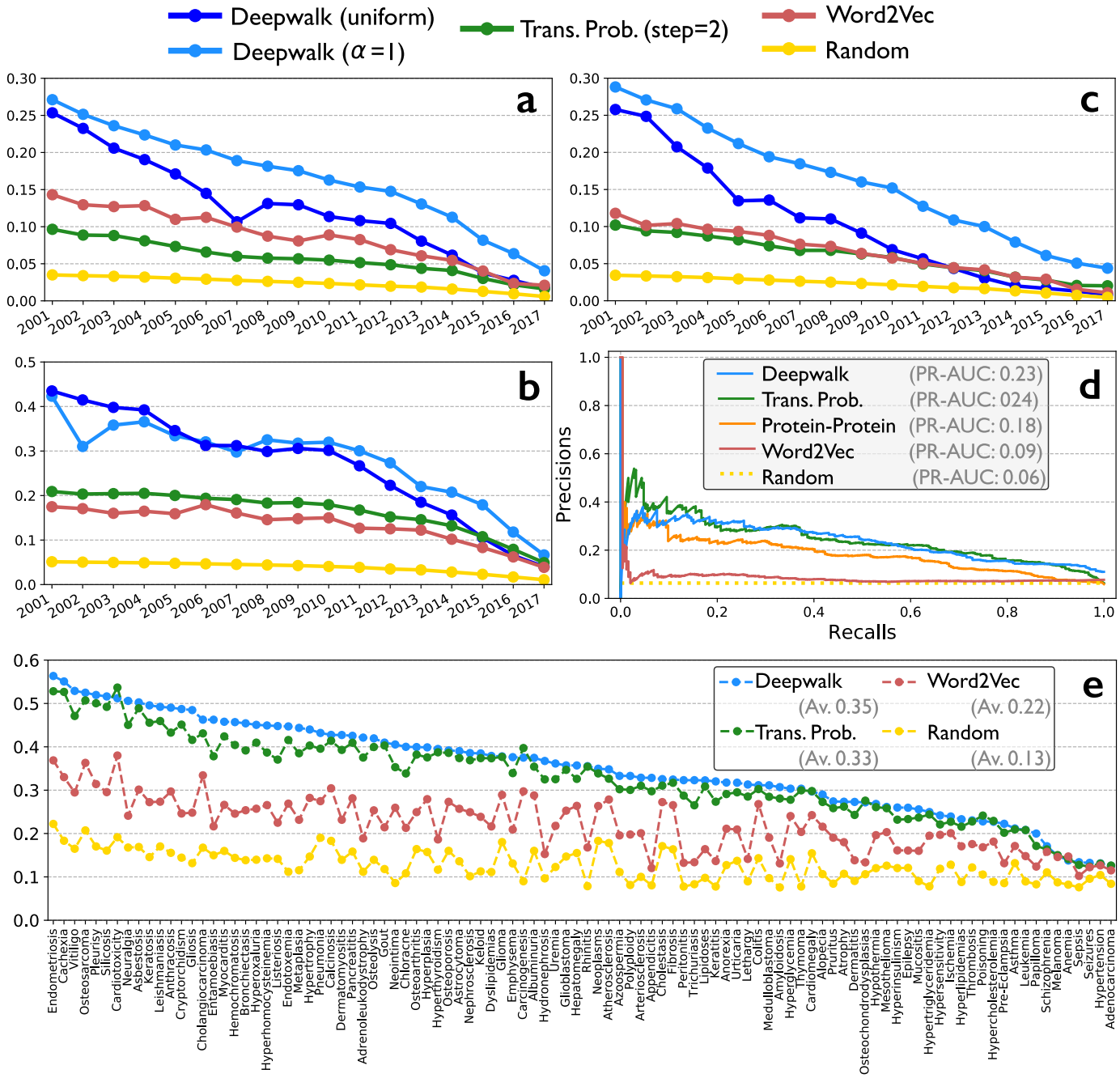
Peer review information *Nature Human Behaviour* thanks Chao Min, Roger Guimerà and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

https://doi.org/10.1038/s41562-023-01648-z



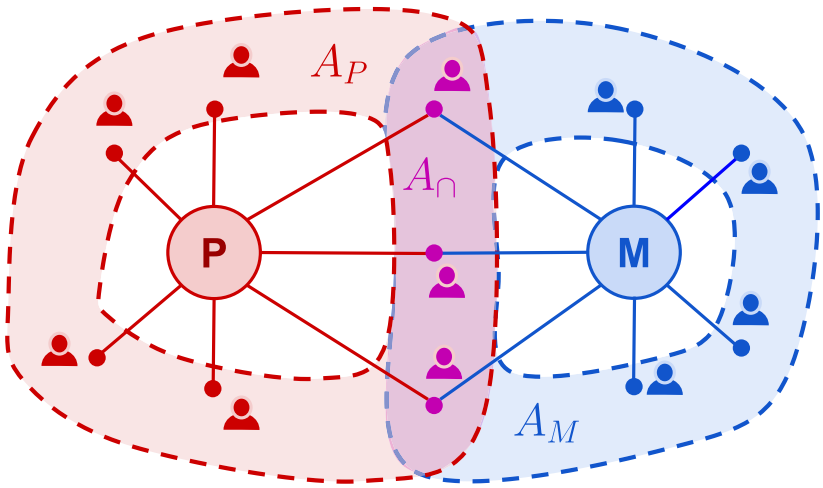
Extended Data Fig. 1 | Hypergraph-induced transition probabilities, and schematic of our experimental design. (a-b) Sanity checks for our hypergraph-induced transition probability similarity metric: **(a)** Between author and conceptual nodes: Histogram of similarities between nodes of two sets of authors and the conceptual node "coronavirus". The two sets of authors include the authors of 5,000 randomly selected papers from journals *Nature Medicine* (dark purple) and *Applied Optics* (light purple) between 1990 and 2019. Similarities between the hypernodes comprise the logarithm of the average transition probabilities with one and two random walk steps. Histograms are plotted considering only non-zero transition probabilities: 92% of the authors of *Nature Medicine* (28,396 in total) but only 51% of the selected *Applied Optics* authors (18,530 in total) had non-zero similarity values. Average non-zero similarities associated with *Nature Medicine* authors (red dashed line) is nearly 5 times larger than that of *Applied Optics* authors (blue dashed line), implying that based on our hypergraph-induce similarity metric, authors publishing in *Nature Medicine* write papers much more relevant to coronavirus in comparison with those publishing in *Applied Optics*. **(b)** Between two conceptual nodes: Similarities between conceptual keywords shown on the *x*-axis and "coronavirus". Similarities between the hypernodes are computed as the average transition probabilities with one and two intermediate nodes. Terms and symptoms known to be more relevant to coronavirus have larger average transition probabilities. **(c)** Schematic of our experimental design: Starting and ending dates of experiments are shown. For energy-related functions and 100 human diseases, we used the beginning of 2001 as prediction year and the end of 2018 as a single evaluation date (V1). For COVID-19, the prediction year is the beginning of 2020, and we cumulatively reported monthly precision values until July of 2021 (V1 to V19).

https://doi.org/10.1038/s41562-023-01648-z



Extended Data Fig. 2 | Precision-Recall (PR) curves for human-accessible predictions. Precision-Recall (PR) curves and area under the curves (AUCs) for various human-accessible predictions: energy-related material science properties, that is, thermoelectrics **(a)**, ferroelectrics **(b)** and photovoltaics **(c)**, therapies and vaccines for COVID-19 **(d)**, and generic drug 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 predictions always result in AUC of 0.5, but the PR-AUC of the random baseline depends on the ratio of positive samples in the data.

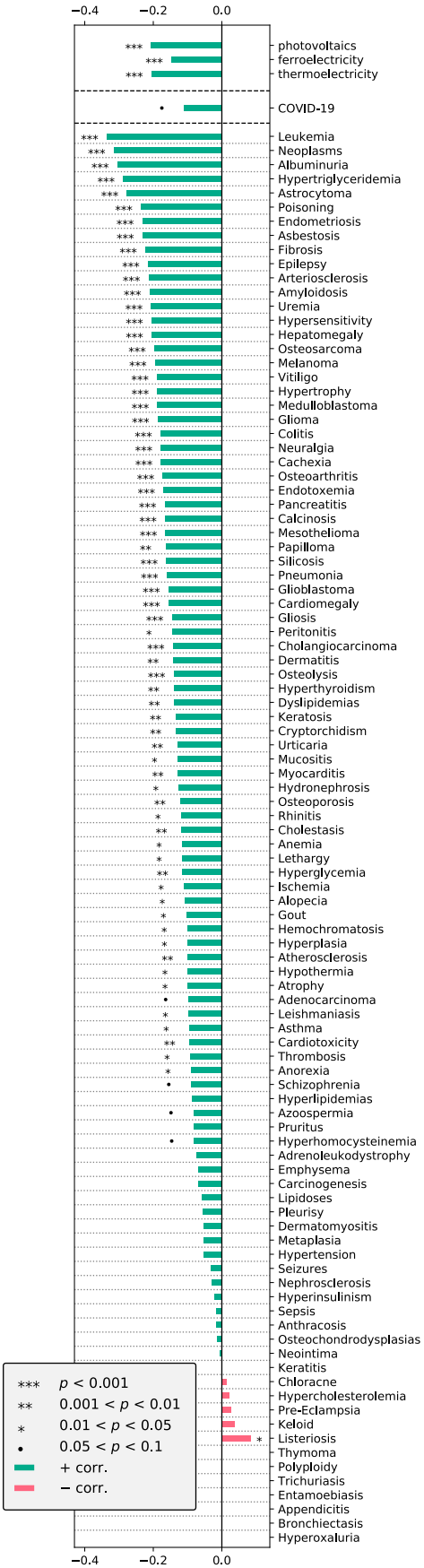
https://doi.org/10.1038/s41562-023-01648-z



$$\text{Expert Denisty}(P, M) = \frac{|A_P \cap A_M|}{|A_P \cup A_M|} = \frac{|A_\cap|}{|A_P \cup A_M|}$$

Extended Data Fig. 3 | Expert density calculataion. Calculation of expert density between property (node *P*) and each material (node *M*). Density is defined as the Jaccard index between the set of authors who have published on the property (denoted by A_P) and those who have mentioned the material in their publications (denoted by A_M). The Jaccard formulation involves taking the ratio of the size of the intersection (that is, the number of overlapping authors) denoted by $A_\cap$ to the size of the union of the two sets (that is, the total number of authors) denoted by $A_P \cup A_M$.

https://doi.org/10.1038/s41562-023-01648-z

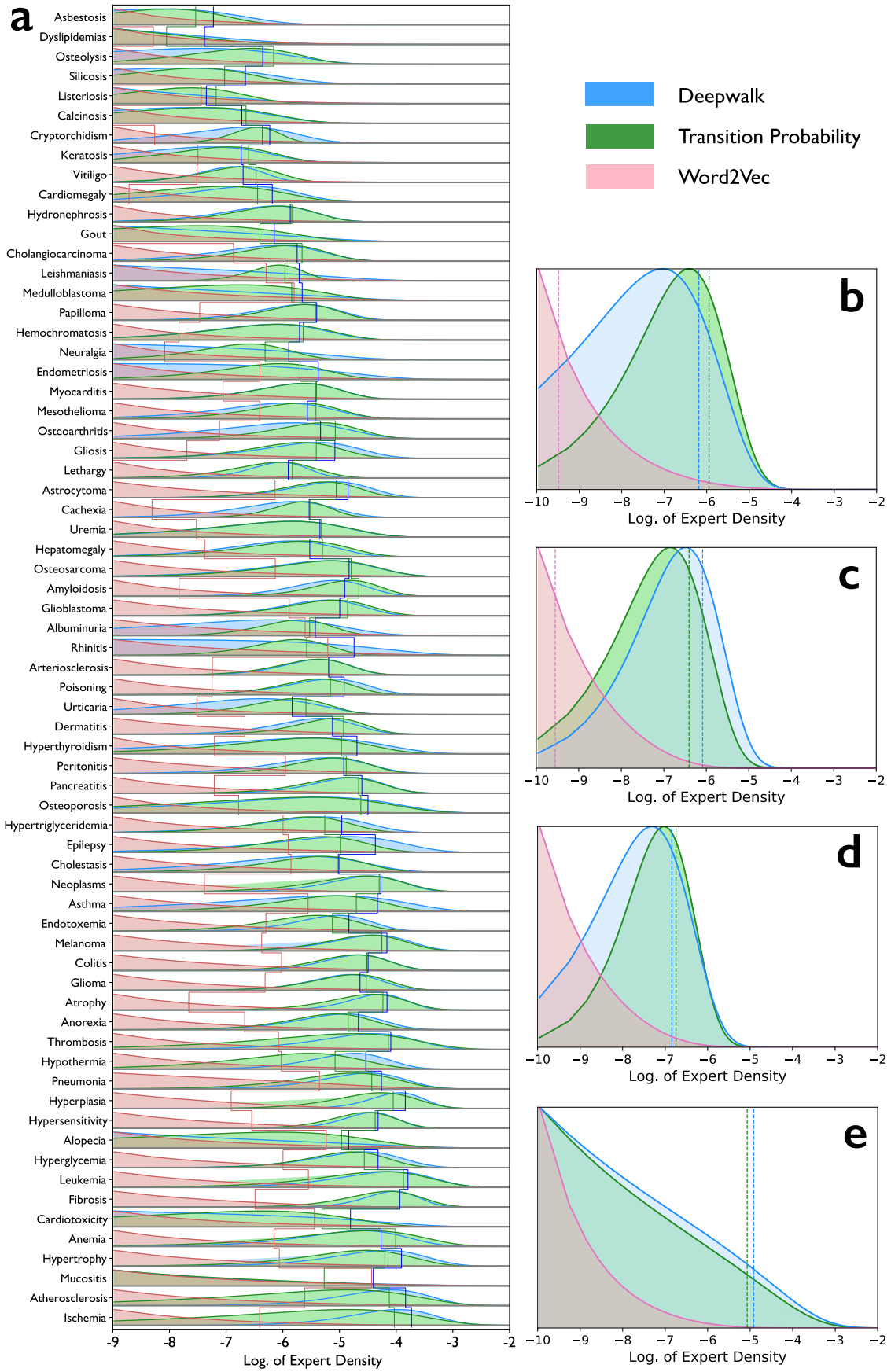


Extended Data Fig. 4 | See next page for caption.

https://doi.org/10.1038/s41562-023-01648-z

Extended Data Fig. 4 | Correlations between expert density and time to discovery. Spearman correlation coefficients between the human expert density (Jaccard index) linking properties with materials and their date of discovery if discovered. Negative correlations imply that materials with higher expert densities are likely to be discovered earlier than others. These results were obtained with the prediction year set to 2001 for energy-related properties and drug repurposing applications, and set to the beginning of 2020 for COVID-19. Turquoise and red bars represent negative and positive correlations, respectively. For seven diseases in the CTD (shown in the bottom of the figure), all discoveries were established in a single year and therefore no correlation coefficients could be obtained. This is because we did not have accurate access to the month or day of discoveries in our database. Results indicate that energy-related properties and COVID-19 all post strong negative correlations. In the case of CTD database, 67 out of 100 diseases (that is, properties) showed statistically significant correlations, among which only one disease had a positive coefficient. The mean correlation coefficients across these 67 diseases was −0.18.

https://doi.org/10.1038/s41562-023-01648-z

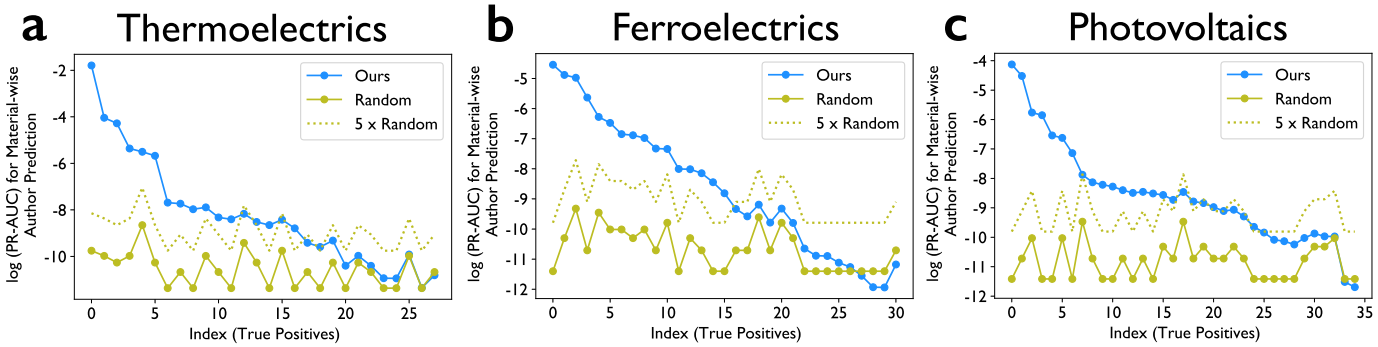


Extended Data Fig. 5 | See next page for caption.

https://doi.org/10.1038/s41562-023-01648-z

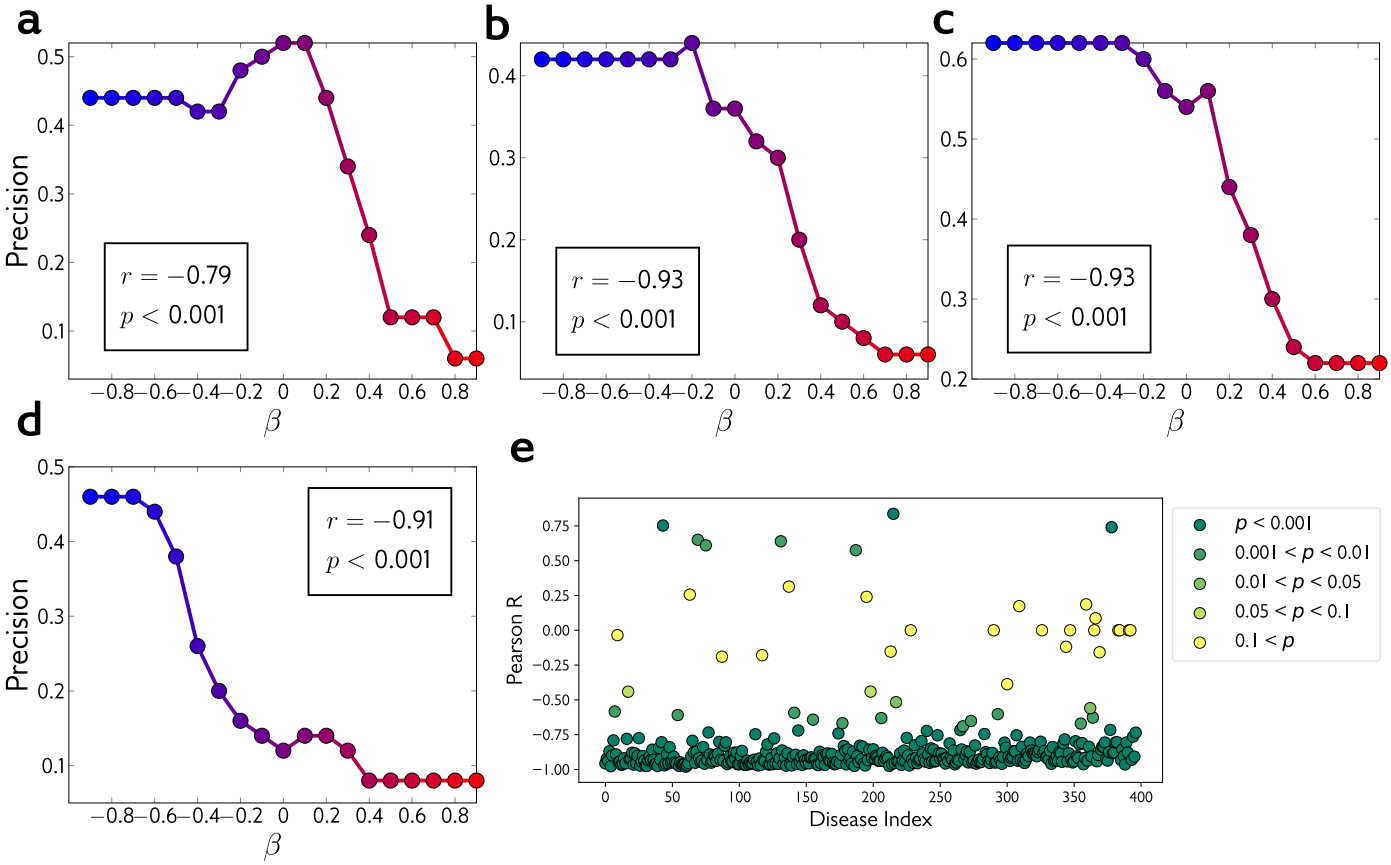
Extended Data Fig. 5 | Distribution of human expert densities. Distribution of human expert densities between discovery predictions and properties: **(a)** drug repurposing application (considering only the 67 diseases with statistically significant Spearman correlation coefficients, see Extended Data Fig. 3); **(b-d)** energy-related materials science properties, that is, thermoelectricity, ferroelectricity and photovoltaic capacity, respectively; and **(e)** therapies and vaccines for COVID-19. Curves measure normalized histograms over the logarithm of human expert densities plotted by fitting a Beta distribution over expert densities for predictions. Solid and dashed vertical lines represent mean values for corresponding densities. It is clear that the distribution of human 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 distributions). CTD diseases are sorted by average expert similarity between them and the complete pool of drugs.

https://doi.org/10.1038/s41562-023-01648-z



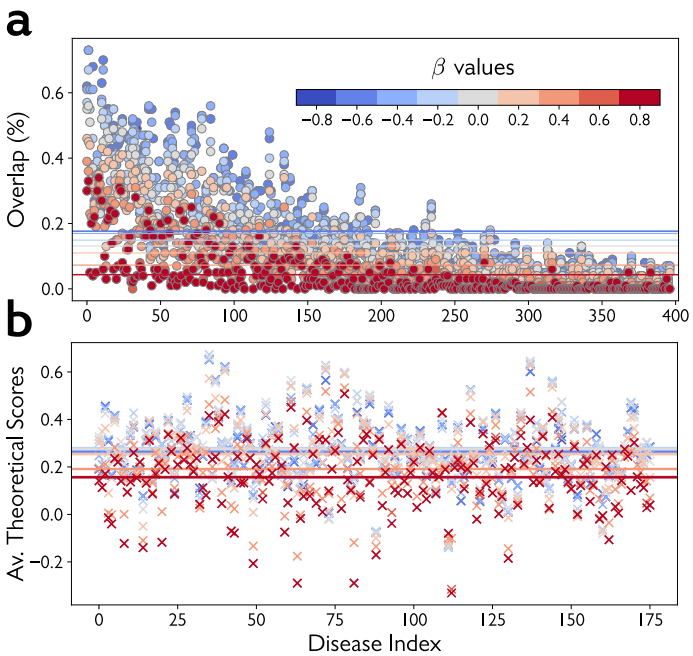
Extended Data Fig. 6 | Precision-Recall area under the curve for predicting human discoverers. Precision-Recall Area Under the Curve (PR-AUC) for predicting the human experts who will discover (discoverers of) materials possessing the following specific properties: **(a)** thermoelectrics, **(b)** ferroelectrics, and **(c)** photovoltaics. Materials selected were among True Positive discovery predictions of our deepwalk-based predictor (α=1). Our evaluation compares scores assigned to candidates 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 we considered here are those sampled at least once in deepwalk trajectories, produced over our five-year hypergraph. For a discovered material, scores were computed based on the 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 embedding 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 human expert predictions, we ranked experts based on their proximity to the property and the material 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 a random selection of experts and also with a curve simulating an imaginary method whose log-PR-AUC is five times higher than the random baseline. Results reveal that predictions were notably superior to random expert selection for all electrochemical properties.

https://doi.org/10.1038/s41562-023-01648-z



Extended Data Fig. 7 | Decaying discoverability in complementary predictions. Illustration of decaying discoverability for predictions as β, the parameter for human expert avoidance, increases. Discoverability of predictions is measured through computing the precision metric, that is, their overlapping percentage with respect to actual discoveries made after prediction year. Decreasing precision curves and their highly negative Pearson correlation coefficients are shown for **(a)** thermoelectricity, **(b)** ferroelectricity, **(c)** photovoltaics and **(d)** COVID-19. We also visualize these statistics for the remaining human diseases with a scatterplot of their Pearson correlation coefficients **(e)**.

https://doi.org/10.1038/s41562-023-01648-z



Extended Data Fig. 8 | Discoverability and scientific merit among drug repurposing predictions. Discoverability and scientific merit for predictions made with varying β values, our parameter for human expert avoidance, in research that repurposes drugs to treat human disease. **(a)** Precision values for predictions generated with eight levels of β and computed for all 400 human diseases we considered (except COVID-19). Diseases are sorted in terms of the number of relevant drugs. **(b)** Average theoretical scores measured through protein-protein similarity between diseases and candidate drugs for predictions generated with the same β values. We compute protein-based theoretical scores for 176 diseases out of 400 total cases (44%). In both subfigures, horizontal lines show average values across all diseases.

https://doi.org/10.1038/s41562-023-01648-z

Extended Data Table 1 | High-frequency MeSH terms appearing in COVID-19 random walks

No.	MeSH Descriptor	MeSH Qualifier	No.	MeSH Descriptor	MeSH Qualifier
1	Middle East Respiratory Syndrome Coronavirus	Immunology	21	Endometrium	
		Genetics	22	Gonadotropin-Releasing Hormone	
		Isolation & Purification	23	Vero Cells	
2	SARS Virus		24	Antibodies, Neutralizing	Immunology
3	Coronavirus	Diagnosis	25	Follicle Stimulating Hormone	
		Virology	26	Influenza A virus	
		Immunology	27	Viral Nonstructural Proteins	
		Epidemiology	28	Estrogen Receptor alpha	
		Prevention	29	Genome, Viral	
		Veterinary	30	Pregnancy Rate	
4	Progesterone	Administration & Dosage	31	Influenza, Human	Virology
		Blood	32	Receptor, ErbB-2	Metabolism
		Pharmacology	33	Zoonoses	
		Metabolism	34	RNA, Viral	Genetics
		Genetics	35	Triple Negative Breast Neoplasms	
5	Insemination, Artificial	Veterinary	36	Uterus	
6	Estrous Cycle		37	Carcinoma, Ductal, Breast	
7	Respiratory Tract Infections	Virology	38	Viral Proteins	Genetics
8	Swine Diseases	Virology	39	Testosterone	Blood
9	Progestins		40	Ovary	
10	Estradiol	Blood	41	Influenza, Human	Epidemiology
		Pharmacology	42	Virus Replication	
11	Ovulation		43	Antiviral Agents	Pharmacology
12	Feces	Virology	44	Seroepidemiologic Studies	
13	Ovarian Follicle		45	Breast Neoplasms	Mortality
14	Respiratory Tract Infections	Epidemiology			Metabolism
15	Viral Vaccines	Immunology			Genetics
16	Receptors, Estrogen	Metabolism			Pathology
17	Virus Internalization				Drug Therapy
18	Luteinizing Hormone				Diagnosis
19	Orthomyxoviridae Infections		46	Disease Outbreaks	
20	Antibodies, Viral	Immunology	47	Lactation	

List of 47 clinical MeSH terms that appeared with a frequency higher than random selection in the hyperedges of random walk sequences from property node “coronavirus” to candidate material node “progesterone”

https://doi.org/10.1038/s41562-023-01648-z

Extended Data Table 2 | True positive predictions for our expert-aware deepwalk algorithm and the word2vec baseline for COVID-19

No.	Candidate	$rank_{DW}$	$rank_{W2V}$	# Mentions	Clinical Trial Identifiers
1	Ethanol	15	2,762	18,155	NCT04554433, NCT04554433
2	Oxygen	13	1,763	90,304	NCT04842448, NCT04500626, NCT04398290 NCT04327505, NCT04425031, NCT04251871
3	Methotrexate	28	2,938	5,935	NCT04352465, NCT04610567
4	Calcium	12	1,114	45,221	NCT04379310, NCT04379310
5	Hydrogen Peroxide	17	1,476	9,252	NCT04603794, NCT04584684, NCT04723446 NCT04721457, NCT04659928
6	Iron	27	1,984	31,609	NCT04643691, NCT04424134, NCT04826822, NCT04345887
7	Iodine	47	3,136	7,173	NCT04603794, NCT04941131, NCT04410159, NCT04347954 NCT04344236, NCT04371965, NCT04449965, NCT04478019 NCT04364802, NCT04549376, NCT04446104, NCT04473261 NCT04473261, NCT04510402, NCT04721457, NCT04393792
8	Nitric Oxide	6	400	21,813	NCT04388683, NCT04383002, NCT04601077, NCT04460183 NCT04338828, NCT04398290, NCT04358588, NCT04476992 NCT04305457, NCT04312243, NCT04312243, NCT04312243 NCT04337918, NCT04842331, NCT04290858, NCT04306393
9	Silver	18	1,126	13,328	NCT04978025
10	Estradiol	48	2,932	8,455	NCT04853069, NCT04865029, NCT04359329
11	Vitamin D	21	1,267	15,063	NCT04664010, NCT04386850, NCT04709744, NCT04780061 NCT04482673, NCT04411446, NCT04525820 NCT04489628, NCT04621058
12	Progesterone	44	2,334	8,895	NCT04365127, NCT04865029
13	Metformin	46	2,374	6,374	NCT04604678, NCT04510194, NCT04625985, NCT04626089
14	Imatinib	42	2,167	2,736	NCT04394416, NCT04346147, NCT04422678 NCT04953052, NCT04794088
15	Selenium	49	2,179	4,491	NCT04869579
16	Adenosine	19	421	12,099	NCT04588441
17	Resveratrol	35	707	3,918	NCT04400890, NCT04799743
18	Zinc	25	289	19,993	NCT04370782, NCT04959786, NCT04446104, NCT04468139 NCT04377646, NCT04447534, NCT04472585
19	Sofosbuvir	38	102	1,585	NCT04532931, NCT04535869, NCT04443725, NCT04530422 NCT04561063, NCT04497649, NCT04498936 NCT04460443, NCT04773756
1	Amantadine	240	9	398	NCT04952519, NCT04854759, NCT04894617
2	Tenofovir Alafenamide	1,011	25	210	NCT04405271
3	Nitazoxanide	395	13	138	NCT04523090, NCT04532931, NCT04486313, NCT04348409 NCT04463264, NCT04959786, NCT04343248, NCT04359680 NCT04459286, NCT04746183, NCT04918927, NCT04435314 NCT04441398, NCT04423861, NCT04561063, NCT04552483 NCT04392427, NCT04498936, NCT04729491, NCT04561219 NCT04382846, NCT04788407, NCT04605588 NCT04920838, NCT04406246, NCT04341493
4	Danoprevir	1,540	40	20	NCT04345276

True positive candidate materials exclusively predicted by our human expert-aware deepwalk algorithm (unshaded rows) and content-only algorithm (shaded rows). $rank_{DW}$ and $rank_{W2V}$ denote the ranks assigned to materials by our deepwalk-based approach and the Word2Vec baseline, respectively. Scores generated by each prediction algorithm are defined as cosine similarity of the materials with respect to the targeted property ("coronavirus") based on their corresponding trained embedding, that is, word2vec models trained on random walk sequences (for $rank_{DW}$) and text from abstracts (for $rank_{W2V}$). A material will be reported as a discovery prediction if the algorithm ranks it higher than 50.

nature portfolio

Corresponding author(s): James Evans

Last updated by author(s): May 26, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Scopus API was used to retrieve titles, abstracts and meta data of papers related to electrochemical properties. For collecting MEDLINE (PubMed19) papers, no specific software was used.
Data analysis	In order to analyze the data, we mostly relied on our own codes in Python 3.6 and using its popular built-in packages (e.g., numpy 1.21.2, scipy 1.7.3, pandas 1.3.5, etc.). Other than that, we used MySQL 8.0.29 to organize the data, gensim 3.8 for training word2vec models, scikit-learn 1.0.2 for evaluating the results and torch-geometric 1.6.3 for building graph convolutional neural networks.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

DOIs of papers used for the electrochemical properties together with the PubMed identifiers of the MEDLINE entries used in our experiments can be found in our GitHub repository: https://github.com/jsourati/accelerate-discoveries.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	We did not have human research participants. No sex and gender information is collected.
Reporting on race, ethnicity, or other socially relevant groupings	We did not have human research participants. No socially relevant categorization variable is collected.
Population characteristics	We did not have human research participants. No population characteristic is reported
Recruitment	We did not have human research participants. No recruitment was done.
Ethics oversight	We did not have human research participants.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	This study quantitatively measures and incorporates expert density into a discovery prediction framework.
Research sample	The study used two datasets that include titles, abstracts and metadata (year of publication, authors, etc.) of scientific papers. The first dataset includes 1.5M papers in the field of materials science which are mostly related to inorganic materials (retrieved through Elsevier's Scopus API). The second dataset includes around 27.5M papers in biomedical and life sciences fields (downloaded from MEDLINE database).
Sampling strategy	We did not do any random sampling and used all papers for which we had sufficient materials (titles, abstracts, year of publication and authors) in the databases mentioned above.
Data collection	The data were collected digitally through Scopus API (for materials science papers) and the available MEDLINE database.
Timing	Such timing is not relevant to our type of data.
Data exclusions	We excluded papers that did not have one of the following information: title/abstract, year of publication, authors. This is because our algorithm could operate properly only if it has access to all these information. For example, about 0.5M papers were discarded from MEDLINE database.
Non-participation	Our research did not include any participants.
Randomization	Our research did not include any participants.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging