Enabling researchers to leverage systems to overcome the limits of human cognitive capacity.

BY TOM HOPE, DOUG DOWNEY, OREN ETZIONI, DANIEL S. WELD, AND ERIC HORVITZ

WE STAND AT the foot of a significant inflection in the trajectory of scientific discovery. As society continues its digital transformation, so does humankind’s collective scientific knowledge and discourse. The transition has led to the creation of a tremendous amount of information, opening exciting opportunities for computational systems that harness it. In parallel, we are witnessing remarkable advances in artificial intelligence, including large language models capable of learning powerful representations from unstructured text. The confluence of societal and computational trends suggests that computer science is poised to ignite a revolution in the scientific process itself.

At the heart of the scientific process, a basic behavior has remained unchanged for hundreds of years: We build on existing ideas to form new ideas. When faced with a new question or problem, we leverage knowledge from accumulated learnings and from external sources, and we perform synthesis and reasoning to generate insights, answers, and directions. But the last few decades have brought change. The explosion of digital information and steep acceleration in the production of scientific data, results, and publications—with more than one million papers added every year to the PubMed biomedical index alone—stand in stark contrast to the constancy of human cognitive capacity. While scientific knowledge, discourse, and the larger scientific ecosystem are expanding with rapidity, our human minds have remained static, with severe limitations in the capacity to find, assimilate, and manipulate information.

Herbert Simon’s reflection that “…a wealth of information creates a poverty of attention” aptly describes the limited attention of researchers in the modern scientific ecosystem. Even within narrow areas of interest, there is a vast space of potential directions to explore, while the keyhole of cognition admits only a tiny fraction of the broad landscape of information and

A Computational Inflection for Scientific Discovery

Recent advances in AI present great opportunities for augmenting human scientific reasoning. Future systems may work alongside humans throughout the scientific process: detecting and explaining relevant literature, generating hypotheses and directions, suggesting experiments and actions.

The scientific process may be decomposed into salient human tasks. Task-guided scientific knowledge retrieval systems retrieve and synthesize external knowledge in a manner that serves a task-guided utility of a researcher, while taking into consideration the researcher’s goals, state of inner knowledge, and preferences.

Progress has recently been made in building such systems yet fundamental challenges remain: in computational representation and synthesis of scientific knowledge, and in modeling the diversity of human tasks, contexts, and cognitive processes involved in consuming and producing scientific knowledge.

key insights
deliberates over small slices of possibility. The way we search through and reflect on information across the vast space—the areas we select to explore and how we explore them—is hindered by cognitive biases and lacks principled and scalable tools for guiding our attention. “Unknowns” are not just holes in science but important gaps in personal knowledge about the broader knowns across the sciences. We thus face an imbalance between the treasure trove of scholarly information and our limited ability to reach into it. Despite technological advances, we require new paradigms and capabilities to address this widening gap.

We see promise in developing new foundational capabilities that address the cognitive bottleneck, aimed at extending human performance on core tasks of research—for example, keeping abreast with developments, forming and prioritizing ideas, conducting experiments, and reading and understanding papers (see Table 1). We focus on a research agenda we call task-guided scientific knowledge retrieval, in which systems counter humans’ bounded capacity by ingesting corpora of scientific knowledge and retrieving inspirations, explanations, solutions, and evidence synthesized to directly serve task-specific utility. We present key concepts of task-guided scientific knowledge retrieval, including work on prototypes that highlight the promise of the direction and bring into focus concrete steps forward for novel representations, tools, and services. Then we review systems that help researchers discover novel perspectives and inspirations, help guide the attention of researchers toward opportunity areas rife with uncertainties and unknowns, and models that leverage retrieval and synthesis of scientific knowledge as part of machine learning and prediction. We conclude with a discussion of opportunities ahead with computational approaches that have the potential to revolutionize science. To set the stage, we begin by discussing some fundamental concepts and background for our research agenda.

Human-Centric Perspective

Extraordinary developments at the convergence of AI and scientific discovery have emerged in specific areas, including new kinds of analytical tools; the prominent example is AlphaFold, which harnesses deep neural models to dramatically improve the prediction of protein structure from amino-acid sequence information. Large language models (LLMs) have very recently made stellar progress in the ability to reason about complex tasks, including in the medical domain. The most advanced LLM at present—emerging before the ink has dried on this article—is GPT-4, which has exhibited jaw-dropping skill in handling clinical questions, mathematical problems, and computer coding tasks.

We view these developments as tremendous research opportunities for building computational approaches that accelerate scientific discovery. We take a human-centered, cognitive perspective: augmenting researchers by considering the diversity of tasks, contexts, and cognitive processes involved in consuming and producing scientific knowledge. Collectively, we refer to these as a researcher’s inner cognitive world (see Figure 1). The researcher interacts with the scientific ecosystem—literature, resources, discussions—to inform decisions and actions. Researchers have different uses for scholarly information, depending on the task at hand and the stage of exploration (see Table 1 and discussion in the section, “Task-Guided Retrieval”). We pursue a research agenda around assisting researchers in their tasks, guided by two main desiderata:

(1) Systems for augmenting human capabilities in the sciences need to enhance the effective flow of knowledge from the outer world of scientific information and discourse to the researcher’s inner cognitive world—countering humans’ bounded capacity by retrieving and synthesizing information targeted to enhance performance on tasks. Achieving this goal requires methods that build and leverage rich representations of scientific content and that can align computational representations with human representations, in the context of specific tasks and backgrounds of researchers.

---

a We use the term researcher to include practitioners in science-driven areas, such as medical doctors and technological engineers, who require deep scientific knowledge.
(2) Research on such systems should be rooted in conceptual models of the inner cognitive world of a researcher. Shining a spotlight on this inner world brings numerous factors and questions to the fore. How do researchers form ideas? How do they decide which problems to look into? How do they find and assimilate new information in the process of making decisions? What cognitive representations and bottlenecks are involved? What computing services would best augment these processes?

Background and related themes. We leverage research in natural language processing (NLP), information retrieval, data mining, and human–computer interaction (HCI) and draw concepts from multiple disciplines. For example, efforts in metascience focus on sociological factors that influence the evolution of science, such as analyses of information silos that impede mutual understanding and interaction, of macro-scale ramifications of the rapid growth in scholarly publications, and of current metrics for measuring impact—work enabled by digitization of scholarly corpora. Metascience research makes important observations about human biases—work enabled by digitization of scholarly corpora. Metascience research makes important observations about human biases (desideratum 2) but generally does not engage in building computational interventions to augment researchers (desideratum 1). Conversely, work in literature-based discovery mines information from literature to generate new predictions (for example, functions of materials or drug targets) but is typically done in isolation from cognitive considerations; however, these techniques have great promise in being used as part of human-augmentation systems. Other work uses machines to automate aspects of science. Pioneering work from Herbert Simon and Pat Langley automated discovery of empirical laws from data, with models inspired by cognitive mechanisms of discovery. More recent work has focused on developing robot scientists that run certain experiments in biology or chemistry—not only formulating hypotheses but “closing the loop” through automated tests in a physical laboratory—where robots may use narrow, curated background knowledge (for example, of a specific gene regulatory network) and machine learning to guide new experiments. Related work explores automating scientific data analysis, which we discuss in the “Task-Guided Retrieval” section as a case of retrieval from scientific repositories to augment aspects of experiment and analysis (see Table 1).

We now turn to a discussion of central concepts: the ecosystem of science and the cognitive world. This presentation lays the foundations for our exposition of task-guided retrieval and research opportunities.

Outer world: Scientific ecosystem. We collectively name the scientific ecosystem and the digital representations of scientific knowledge as the outer world (see Figure 1). The outer world comprises scientific communities: a complex and shifting web of peers, concepts, methodologies, problems, and directions revolving around shared interests, understandings, and paradigms. This ecosystem generates digital information—digital “traces” of scientific thought and behavior—lying at the center of our attention as computer scientists interested in boosting human capacity to “reach into” the pool of scientific knowledge. This knowledge includes scholarly publications that appear in journals, conference proceedings, and online preprint repositories. Online publications are a main case of digital research artifacts; other examples of products of research include software, datasets, and knowledge bases. Research artifacts are also typically associated with signals of quality and interest, such as citations to a specific paper or downloads of a dataset. The specific context for why a paper or resource was cited or used is often reflected in natural language descriptions. Different types of signals include peer review prior to publication (mostly not shared publicly) and social media discussions, such as on Twitter, which has become a major virtual platform for academic dissemination and conversation. Along with the trend in society, private communication channels among researchers are

| Task/Activity | Description |
|---------------|-------------|
| Attention to areas of interest | A background process of keeping track of the latest developments in relevant scientific communities. Involves applying selective attention, perceiving relevance, and utility. |
| Problem identification and prioritization | Identifying new research questions and deciding which ones to work on. Involves factors such as subjective preferences and feasibility assessment. |
| Forming directions | Forming ideas to address a given a problem/question. Involves cognitive processes such as constructing mental models of a problem, problem reformulation, abstraction and decomposition, adaptation of relevant knowledge to new scenarios, and assessing the likelihood of success. |
| Literature search and review | Accessing and ingesting knowledge in the literature. Involves many processes including query formulation, skimming and assessing relevance, positioning ideas with comparisons and contrasts to existing work, and reading and summarization strategies. |
| Learning, understanding, sense-making | The cognitive processes and activities involved in assimilating new knowledge and concepts, and making sense of complex scientific information spaces. |
| Experimentation, analysis, action | A broad category referring to the many processes and activities involved in formulating and conducting experiments (for example, planning data collection and measurements), performing analyses (for example, understanding a set of data points, modeling and extrapolation, prediction, and evaluation), and producing artifacts, techniques, theories, decisions, policies, and actions. |
| Research communication | Writing research documents (papers, proposals, and analyses) and communicating with peers (feedback and review, collaboration, presentation). |
also digital—email, online calls, and messages. Similarly, note taking and writing—important activities across the scientific workflow—are done in digital form. This information is siloed in different platforms under privacy restrictions yet represents a treasure trove for tools for the augmentation of scientific reasoning and exploration.

**Inner world: Human cognition in science.** The way researchers decide to interact with information in the outer world and the way they process and use this information is governed by a complex array of cognitive processes, personal knowledge and preferences, biases, and limitations, which are only partially understood. We collectively name these the inner world, and briefly discuss several salient aspects.

Early work in AI by Herbert Simon and Alan Newell and later efforts by Pat Langley and Paul Thagard focused on cognitive and computational aspects of problem solving, creativity, decision making, scientific reasoning, and discovery, seeking algorithmic representations to help understand and mimic human intelligence. Cognitive mechanisms that play important roles in scientific discovery include inductive and abductive reasoning, mental modeling of problems and situations, abstraction, decomposition, reformulation, analogical transfer, and recombination—for example, in analogical transfer, given a situation or problem being considered in our working memory, we retrieve prior analogous problems or situations from our long-term memory.

This cognitive machinery powers human ingenuity. However, the human mind also has severe limitations—bounded rationality in the words of Simon—that impede these powerful mechanisms. Our limitations and capabilities have been studied for more than 100 years with cognitive psychology. Our limitations manifest in bounded cognitive capacity and knowledge, as well as in the biases that govern our behaviors and preferences. These limitations are all tightly interrelated. The ability to generate ideas, for instance, directly relies on prior knowledge, but when a large volume of information from the outer world of science is met by insufficient cognitive capacity for processing and assimilating it, the result is information overload—a ubiquitous hindrance for researchers. Information overload in science strains the attentional resources of researchers, forcing them to allocate attention to increasingly narrow areas. This effect, in turn, amplifies a host of biases which researchers, just like all humans, suffer from. For example, scientists can be limited by confirmation bias, aversion to information from novel domains, homophily, and fixation on specific directions and perspectives without consideration of alternative views. More broadly, selection of directions and areas to work on is a case of decision making; as such, personal preference and subjective utility play fundamental roles. Our research decisions rely on subjective assessment of feasibility, long-term or short-term goals and interests, and even psychological factors—for example, tendencies for risk aversion. These factors are also impacted by biases. Clearly, the inner world of researchers is dauntingly complex. However, in the next section, we present encouraging results of applying computational methods to augment cognition in the sciences, helping to mitigate biases and limitations and enabling researchers to better use of their powerful creative mechanisms.

**Task-Guided Retrieval**

How might we widen and deepen the connection between the outer world of science and the limited cognitive worlds of researchers? We see a key bridge and research opportunity with developing tools for scientific task-guided knowledge retrieval. Drawing from discussions in literature on the process of scientific discovery, we enumerate in Table 1 salient scientific tasks and activities, such as problem identification, forming directions, learning, literature search and review, and experimentation. These tasks could benefit from augmentation of human capabilities but remain underexplored in computer science. Existing computational technologies to help humans discover scientific knowledge are under-invested in important aspects of the intricate cognitive processes and goal-oriented contexts of scholarly endeavors.

The dominant approach to information-retrieval research and systems can be summarized as “relevance first,” focusing on results that answer user queries as accurately as possible. Academic search engines assume users know what queries to explore and how to formulate them. For pinpointed literature search in familiar areas, this assumption may often suffice. But a broad array of other scholarly tasks, such as ideation or learning about a new topic, are very much underserved. At the same time, many voices in the information-retrieval community have discussed a different, broader view of utility-driven search situated in a wider context of information seeking by users with specific intents and tasks. Here, we adapt ideas and principles from this general paradigm.

We envision methods for task-guided scientific knowledge retrieval: systems that retrieve and synthesize outer knowledge in a manner that directly serves a task-guided utility of a researcher, while taking into consideration the researcher’s goals, state of inner knowledge, and preferences.

Consider the tasks in Table 1. For researchers engaged in experimentation or analysis, we envision systems that help users identify experiments and analyses in the literature to guide design choices and decisions. For researchers early stages of selecting problems to work on, we picture systems that support this decision with information from literature and online discussions, synthesized to obtain estimated impact and feasibility. As part of forming directions to address a problem, systems will help users find inspirations for solutions. Researchers who are learning about a new topic will be provided with retrieved texts and discussions that explain the topic in a manner tailored to personal knowledge. Importantly, task-guided knowledge retrieval follows the two considered previously introduced; namely, systems should enable users to find knowledge that directly assists them in core research tasks by augmenting their cognitive capacity and mitigating their biases, and computational representations and services should align with salient cognitive aspects of the inner world of researchers.
We focus on a research agenda we call task-guided scientific knowledge retrieval, in which systems counter humans’ bounded capacity.

Prototypes of task-guided retrieval. We present work on initial steps and prototypes, including representative work that we have done and the work of others, framed in alignment with task-guided knowledge retrieval and the tasks enumerated in Table 1. The main aim of this brief review is to stimulate discussion in the computer science community on tools for extending human capabilities in the sciences. Existing methods are far from able to realize our vision. For example, we see major challenges in representation and inferences about the inner world of knowledge and preferences and aligning these with representations and inferences drawn from the outer-world knowledge. Today’s prototypes are limited examples of our vision, using very rough proxies of inner knowledge and interest based on papers and documents written or read by the user, or in some cases only a set of keywords.

Forming directions. We have developed methods for helping researchers generate new directions. A fundamental pattern in the cognitive process of creativity involves detecting abstract connections across ideas and transferring ideas from one problem to another. Grounded in this cognitive understanding, we have pursued several approaches for stimulating creativity powered by retrieving outer knowledge. We developed and studied a system named Bridger, which connects researchers to peers who inspire novel directions for research. Bridger identifies matches among authors based on commonalities and contrasts, identifying peers who are both relevant and novel—working on similar problems but using very different methods to potentially inspire new solutions (see Figure 2). By doing so, Bridger helps mitigate the cognitive bias of fixation. In this setting, inner knowledge is represented as mentions of problems and methods extracted automatically from a researcher’s papers and weighted by term frequency. The outer knowledge being retrieved takes the form of other authors in computer science, following the same representation. For each retrieved author, the system displays salient problems, methods, and papers ranked by measures of relevance to the user. In studies with CS researchers, we found that Bridger dramatically boosted creative search and inspiration over state-of-the-art neural models employed by the Semantic Scholar search engine, surfacing useful connections across diverse areas. For example, one researcher drew novel connections between the mathematical area of graph theory and their own area of human-centered AI by exploring a recommended author who applies graph theory to decision making. The studies also surfaced important challenges, which we discuss later in this article.

We have also explored retrieving outer knowledge to enhance the human ability to find opportunities for analogical transfer. Extensive work in cognitive studies has highlighted the human knack for “analogical retrieval” as a central function in creativity—bringing together structurally related ideas and adapting them to a task at hand. We developed a search method that enables researchers to search through a database of technological inventions and find mechanisms that can be transferred from distant domains to solve a given problem. Given a textual description of an invention as input from the user, we retrieve ideas (inventions, papers) that have partial structural similarity to the input (for example, inventions with similar mechanisms) to facilitate discovery of analogical transfer opportunities. We found that the method could significantly boost measures of human creativity in ideation experiments, in which users were asked to formulate new ideas after viewing inspirations retrieved with our approach versus baseline information-retrieval methods. For example, a biomechanical engineering lab working on polymer stretching/folding for creating novel prototypes of task-guided retrieval.

We focus on a research agenda we call task-guided scientific knowledge retrieval, in which systems counter humans’ bounded capacity.

Figure 2. Matching researchers to authors with whom they are unfamiliar, to help generate directions. Author cards show key problems and methods extracted from their papers.
Information overload in science strains the attentional resources of researchers, forcing them to allocate attention to increasingly narrow areas.

Innovation may also involve traversing multiple levels of abstraction around a problem to “break out” of fixation on the details of a specific problem by exploring novel perspectives. Given as input a problem description written by the user (as a proxy summary of the user’s inner world of knowledge and purpose), we have pursued mechanisms that can retrieve diverse problem perspectives that are related to the focal problem, with the goal of inspiring new ideas for problem abstraction and reformulation\(^1\) (see Figure 3). Using NLP models to extract mentions of problems, we mine a corpus of technological-invention texts to discover problems that often appear together; we use this information to form a hierarchical problem graph that supports automatic traversal of neighboring problems around a focal problem, surfacing novel inspirations to users. In a study of the efficacy of the methods, more than 60% of “inspirations” retrieved this way were found to be useful and novel—a relative boost of 50%-60% over the best-performing baselines. For example, given an input problem of reminding patients to take medication, our system retrieves related problems, such as in-patient health tracking and alerting devices.

**Guiding attention and problem identification.** We see great opportunity in developing methods for guiding the attention of researchers to important areas in the space of ideas where there exists less knowledge or certainty (Figure 3). In one direction, we built a search engine that allows users to retrieve outer knowledge in the form of difficulties, uncertainties, and hypotheses in the literature. The key goals of this search mode are to bolster attention to rising and standing challenges of relevance to the user, to help overall with problem identification and selection. We performed experiments with participants from diverse research backgrounds, including medical doctors working in a large hospital. Using query topics as a proxy for the inner world of participants’ interests, we found the system could dramatically outperform PubMed search, the go-to biomedical search engine, at discovering important and interesting areas of challenges and directions. For example, while searching PubMed for the ACE2 receptor in the context of COVID-19 returns well-studied results, the prototype system by contrast focuses on finding statements of uncertainty, open questions, and initial hypotheses, such as a paper noting the possibility that ACE2 plays a role in liver damage in COVID-19 patients.

Another direction on biases and blind spots considers the long-term effort to identify protein-protein interactions (PPIs). A dataset of the growing graph of confirmed PPIs over decades was constructed and leveraged to identify patterns of scientific attention.\(^32\) A temporal analysis revealed a significant “bias of locality,” where explorations of PPIs are launched more frequently from those that were most recently studied, rather than following more general prioritization of exploration. While locality reflects an understandable focus on adjacent and connected problems in the biosciences, the pattern of attention leads to systematic blind spots in large, widely used PPI databases that are likely unappreciated—further exacerbating attentional biases. The study further demonstrated mechanisms for reprioritizing candidate PPIs based on properties of proteins and showed how earlier discoveries could be made using debiasing methods. The findings underscore the promise of tools that retrieve existing outer-world knowledge to guide attention to worthwhile directions. In this case, the outer-knowledge source is a...
PPI database, and a user-selected subgraph provides a proxy for inner-world knowledge and interests.

**Literature search and review.** A great body of work on literature search and review has deep relevance to task-guided retrieval in the sciences. In particular, we see great opportunity to build on recent advances in information retrieval to help biomedical researchers with domain-specific representations and to enhance scientific search by building new neural models. Specialized search systems have been developed for the biomedical domain, with the overall vision of harnessing natural language-understanding technologies to help researchers discover relevant evidence and expedit the costly process of systematic literature review. For example, Nye et al. built a search-and-synthesis system based on automated extraction of biomedical treatment-outcome relations from clinical trial reports. The system is found to assist in identifying drug-repurposing opportunities. As another recent example, the SPIKE system enables researchers to extract and retrieve facts from a corpus using an expressive query language with biomedical entity types and new term classes that the user can interactively define. Together, this work underscores the importance of extracting a semantically meaningful representation of out-world knowledge that aligns with core aspects of inner-world reasoning by researchers.

In separate work, neural language models built via self-supervision on large corpora of biomedical publications have recently led to performance boosts and new features in literature search systems, such as support for natural language queries that provide users with a more natural way to formulate their informational goals. Neural models have also been trained to match abstract discourse aspects of pairs of papers (for example, sentences referring to methodologies) and automatically retrieve documents that are aspectually similar. By employing a representation that aligns with scientific reasoning across areas, this method achieves state-of-the-art results across biomedical and computer science literature.

**Experimentation, analysis, and action.** Beyond helping researchers via awareness and knowledge, we see great opportunities to use scientific corpora to construct task-centric inferential systems with automated models and tools for assisting with analysis, prediction, and decisions. We demonstrate these ideas by casting two different lines of work as cases of task-guided retrieval.

1. **Workflows** are multi-step computational pipelines used as part of scientific experimentation for data preparation, analysis, and simulation. Technically, this includes execution of code scripts, services and tools, querying databases, and submitting jobs to the cloud. In the life sciences, in areas such as genomics, there are specialized workflow-management systems to help researchers find and use workflows, enabled by a community that creates and publicly shares repositories of workflows with standardized interfaces, metadata, and functional annotations of tools and data. As discussed in Gil, machine-learning algorithms can potentially use these resources to automate workflow construction, learning to retrieve and

---

**Figure 4. Suggesting research opportunities for query concepts (for example, medical topics) by identifying blind spots, gaps in collective knowledge, and promising areas for exploration.**

- **Input:**
  - Items of interest (concepts, entities, topics...)

- **Retrieved Knowledge:**
  - Specific areas of
    - blindspots,
    - uncertainty,
    - difficulty, and
    - opportunity

- **Ingest scientific corpora, resources**

- **Automated mining, debiasing, inference**

**Figure 5. Leveraging medical corpora to enhance the precision of AI models for inference about patient outcomes.**

- **Admission Note**
  - 49-year-old male...
  - refractory hypoglycemia...
  - type 1 diabetic...

- **Literature Retrieval**
  - patient desc. + in-hospital mortality
  - Severe hypoglycemia...
  - not associated with the risk of mortality in adults with type 1 diabetes...

- **BEEP system integrates notes + literature**

- **More accurate prediction:** Patient survives!
synthesize data-analysis pipelines. In this setting, outer-world knowledge takes the form of workflow repositories, from which systems retrieve and synthesize modular building blocks; the user’s inner world is reflected via analysis objectives and constraints.

2. In our work on clinical predictions, the goal is to enhance prediction of medical outcomes of patients hospitalized in the intensive care unit (ICU), such as in-hospital mortality or prolonged length of stay. Our system, Biomedical Evidence Enhanced Prediction (BEEP), learns to make predictions by retrieving medical papers that are relevant to each specific ICU patient and to synthesize this outer knowledge in combination with internal EMR knowledge to form a final prediction. The primary envisaged user is a practice-oriented researcher—a medical doctor whose inner knowledge is given by a rough proxy in the form of internal clinical notes from which we extract “queries” issued over medical papers. We find BEEP to provide large improvements over state-of-the-art models that do not use retrieval.

Learning and understanding. We introduced a system for helping users learn about new concepts by showing definitions grounded in familiar concepts—for example, a new algorithm is explained as a variant of an algorithm familiar to the user. Cognitive studies have asserted that effective descriptions of a new concept ground it within the network of known concepts. Our system takes as input a list of source concepts reflecting the user’s inner knowledge as obtained from papers that they have written or read. When the user seeks a definition of a new target concept, we retrieve outer knowledge in the form of definitions appearing in scientific papers in which the target concept is explained in terms of the source concepts; a neural text-generation model then rewrites the text in a structured, templated form that relates the target to the source.

Opportunities Ahead
The challenges of task-guided retrieval in support of researchers frame a host of problems and opportunities. We focus on select challenges and directions (see also Table 2). We begin with an illustrative example, imagining a futuristic system to motivate the discussion.

Aspirations. We envision tools that flow outer-world knowledge to researchers based on inferences about their inner world—users’ knowledge, past and present goals and difficulties, and the tasks from Table 1 they are engaged in. The systems would use multiple signals for making inferences, including users’ papers, data, experiments, and communication channels; the systems also converse with the user to understand needs and suggest solutions, hypotheses, and experiments.

We foresee systems powered by rich representations of both inner and outer scientific knowledge. For a given concept, for example, a certain algorithm or organism, an aspirational system would ingest all papers on the subject to form a multi-faceted representation of concepts as objects with associated properties and functions. Using this representation, the system could assist in literature search and review, enabling expressive queries to outer-world information that target abstract aspects, such as functionalities, mechanisms, behaviors, and designs, in a manner that transcends field-specific jargon, abstracting away lexical differences that hindered historical search engines—for example, Google Scholar. To help users learn and understand new concepts they encounter, the system would explain them in relation to other concepts the user already knows. A future system might also help to automate experimentation, analysis, and action and to form directions, by forming compositions of concepts and predicting the resultant affordances—for example, matching a certain algorithm with a suitable problem based on the algorithm’s properties and the problem’s requirements, matching an organism with a specific method of measurement or modification, or recombining parts of two devices to form a new device. The system could help identify related problems in the literature, synthesizing from them useful suggestions for problem reformulations.

Considering the huge combinatorial space of potential suggestions, a system could assist in prioritization using estimated measures of interestingness, feasibility, and value by synthesizing historical and current signals in literature, online discussions, and knowledge bases.

Table 2. Directions with formulating and leveraging computational representations of scientific knowledge.

| Task/Activity                                      | Description                                                                                                                                 |
|---------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------|
| Task-aligned representations, scientific NLP      | How can we automatically and accurately extract conceptual representations of scientific knowledge that align with tasks that comprise the endeavor of science (Table 1)? How can we build NLP models that understand full scientific papers? |
| Computational algebra of ideas                    | Can we build representations of scientific knowledge that support composition of ideas—for example, inferring the result of recombining two concepts? |
| Identifying conceptual relationships across literature | How do we detect important relationships across scientific ideas and across related discussions in different communities and areas? How can we resolve challenges of diversity, ambiguity, and multiple levels of detail in scientific language? |
| Estimation of personal knowledge                  | How can we estimate a given researcher’s knowledge? What are useful, practical models of this knowledge? What concepts does a researcher know, which of their aspects, and to what technical extent? How do we account for latent knowledge? |
| Addressing knowledge gaps                         | Given an estimated model of a researcher’s knowledge, and given a specific task from Table 1, what new knowledge would be useful for the task at hand? |
| Estimation of preferences, goals, and interests    | How can we estimate the key latent preferences, interests, and subjective utilities of researchers? Using information in papers and discussions to infer factors behind researchers’ choices. |
| Prediction and prioritization                     | How might we identify promising sparse/unexplored areas in large “spaces of ideas” and prioritize directions that are novel, plausible, and valuable? |
| Developing new representations for learning and communicating | Might the way we read and write papers change to be more effective? Might we communicate with machine-actionable, interlinked representations of scholarly knowledge? Might we create personalized “living” documents that tailor their content to readers’ backgrounds? |
Envisioned systems would be designed as human-centric, focusing on the individual researcher. The systems would enable users to convey preferences, goals, and interests, and mediate the presentation of suggested directions and problem solutions based on personal prior knowledge—proposing concrete new directions grounded in representations that researchers can follow and assisting users in reading complex retrieved texts by editing their language to conform with concepts that users are familiar with.

**Research directions.** While we have witnessed remarkable strides in AI, the journey toward actualizing our vision requires further advancement. Envisioning such capabilities, however, can serve as a compass for directing research endeavors. An encouraging development can be seen in the recent progress with LLMs, which have demonstrated surprising capabilities with interpreting and generating complex texts and tackling technical tasks. The demonstrated proficiency of these models instills confidence that many of the possibilities we have discussed are attainable. We now elaborate on challenges and directions ahead, including limitations in representing scientific knowledge and making inferences about the inner worlds of researchers (see Table 2).

**Task-aligned representations and scientific NLP.** Paul Thagard writes: “Thinking can best be understood in terms of representational structures in the mind and computational procedures that operate on those structures.” We seek representations that can be aligned with human thinking—for insight-building, decision making, and communication. Can we go beyond textual representation toward representations that support such cognitive processes?

The quest for a universal schema representing scientific ideas goes back hundreds of years. Gottfried Leibniz and René Descartes were intrigued by the prospects of a universal codification of knowledge. Leibniz proposed the *characteristica universalis*, a hypothesized formal language of ideas enabling inferences with algebraic operators. While such a representation is not within reach, envisioning its existence—and how to roughly approximate it—points to important research directions. One exciting direction is obtaining representations that support a “computational algebra of ideas”—for example, modeling compositions of concepts and the affordances that would be formed as a result. Early work on learning vector representations of natural language concepts supported rudimentary forms of addition, subtraction, and analogy—for example, the Word2vec model.

Recently, LLMs have made striking progress in generating novel content and coherently combining concepts. Emerging evidence on GPT-4’s ability to reason not only in unstructured language but also with logical structures grounded in code suggests strong potential for generating novel ideas via compositionality and relational reasoning. Our early experiments with GPT-4 have revealed a constellation of promising abilities to assist with the scientific process, such as formulating hypotheses, recommending future research directions, and critiquing studies. Equipped with training and retrieval with access to millions of scientific papers, descendants of today’s models may be able to synthesize original scientific concepts with the in-depth technical detail at a level reported in high-quality scientific papers. We see great opportunity ahead to leverage LLMs to augment human scientific reasoning along the lines described in this paper.

One limitation with LLMs is that representations learned by these models are currently far from understood and lack “hooks” for control and interpretability, which are important in human-AI collaboration. In line with our focus on grounding representations of outer-world knowledge with inner-world cognitive aspects, we have pursued methods that “reverse-engineer” scientific papers to automatically extract, using NLP, structured representations that balance three desiderata:

1. Semantically meaningful representations, *aligned with a salient task* from the tasks in Table 1, grounded in cognitive research to guide us toward useful structures.

2. Representations with sufficient level of abstraction to generalize across areas and topics.
(3) Representations expressive enough for direct utility in helping researchers as measured in human studies.

For example, we have extracted representations of causal mechanisms and hierarchical graphs of functional relationships. This kind of decomposition of ideas has enabled us to perform basic analogical inference in the space of technological and scientific ideas, helping researchers discover inspirations. However, many richer structures should be explored—for example, of experimentation processes and methodologies to enable tasks in Table 1.

A central challenge is that current models’ extraction accuracy is limited, and the diversity of scientific language leads to problems in generalization and normalization of terms and concepts. We have pursued construction of new datasets, models, and evaluations for identifying similarity between concepts and aspects across papers, with fundamental problems in resolving diversity, ambiguity, and hierarchy of language. As our results have highlighted, models tend to focus on surface-level lexical patterns rather than deeper semantic relationships. Generally, substantial advances are needed to handle challenges posed by scientific documents. We require NLP models with full document understanding, not only of text but of tables, equations, figures, and reference links. Open access corpora, such as S2ORC, provide a foundation to address this challenge.

New modes of writing and reading. Perhaps the way we write can be dramatically different, using machine-actionable representations? Beyond reporting and documentation, writing represents a channel between the inner and outer worlds, forcing us to communicate ideas in concrete language. This process often begets new questions and perspectives. Can systems accompany different phases of writing, suggesting new ideas? In parallel, there is the task of reading what others have written; a recent interactive PDF reader offers, for example, customized concept definitions. We imagine a future where every reader will see a different form of the same paper, rewritten to align with readers’ knowledge—for example, our person-alized concept definitions system will insert new wording and explanations grounded in readers’ knowledge.

Internal world of researchers. Grounding new concepts in readers’ knowledge suggests a wider and highly challenging problem. How can we enable researchers to specify their knowledge and preferences to direct systems to carry out tasks? Directly querying for these aspects burdens the researcher and may be prone to reporting biases. Digital traces present an opportunity for automatically estimating a researcher’s knowledge, objectives, needs, and interests—based on data. We are interested in using researchers’ papers to estimate what concepts users know and to what extent. We envision mixed-initiative interfaces in which approximations of the inner world are presented to researchers and refined in human-machine collaboration, to identify and fill personal gaps in knowledge for a specific task. Representations of interest and preference are central in Web commerce based on user-activity histories. We are encouraged by results highlighting the feasibility of rich user models—for example, in search personalization and dynamic inferences. Paul Samuelson wrote of “revealed preferences,” preferences revealed indirectly by the economic price people are willing to pay; while not equivalent, researchers’ digital traces may reveal preferences—for example, by working on one problem and not another.

Prediction and prioritization of directions. Whenever we decide to work on a research direction, we are implicitly making a prediction about an area in “idea space.” Can automated systems help make these predictions? This involves identifying promising areas and generating directions—hypotheses, ideas—in either natural or structured language, under constraints on users’ background knowledge. Directions should be ranked by estimated likelihood (feasibility, plausibility), utility, and novelty. Despite the great challenges involved, we are encouraged by advances in models trained for predicting specific targets—for example, protein structures; we see potential in building on these advances as part of our wider agenda, which considers the inner world of cognitive aspects.
and tasks and the outer world outside the context of a narrow dataset.

**Pursuing challenges of translation.** Finally, we note challenges for introducing new technologies into scientific workflows. In the context of systems for discovery, researchers interviewed in our studies reported being limited in time and resources, making them less likely to enter new areas and learn unfamiliar concepts, preventing them from discovering potentially promising ideas. More broadly, the sociotechnical environment in which AI models are deployed critically impacts their success. A pertinent example comes via reports on difficulties with translating IBM’s Watson Health systems into practice. The vision of the effort included systems providing insights about patients by mining research papers to suggest, for example, therapies or diagnostics. A prototype system faced difficulties ranging from data processing and availability problems to deeper perceived gaps between the system’s understanding of literature and that of physicians. Challenges such as these are fundamental to the fielding of new applications not only in healthcare but in any setting where humans are required to interact with AI systems. While issues such as data quality and privacy are orthogonal to our agenda, we see directions in modeling of human needs and limitations to inform the design of human-AI experiences within scientific workflows.

**Conclusion**

As the terrain of science widens at a fast pace, researchers are constrained by the limits of human cognition and lack principled methods to follow developments, guide attention, and formulate and prioritize directions. For the first time in history, essentially all scientific knowledge and discourse has moved into the digital space. At the time of this writing, dramatic advances in AI with LLMs are taking place at breathtaking speed. These shifts present tremendous opportunities for leveraging scientific corpora as databases from which solutions, insights, and inspirations can be gleaned. We see opportunity ahead for systems that can address the imbalance between the treasure trove of scholarly information and researchers’ limited ability to reach into it, harnessing humankind’s collective knowledge to revolutionize the scientific process. Numerous challenges stand in the way of the vision we have laid out. However, even small steps forward will unlock vast opportunities for making advances at the frontiers of science.

**Acknowledgments**

We thank the members of the Semantic Scholar team for stimulating discussions. Projects were supported by NSF Convergence Accelerator Grant 2132318, NSF RAPID grant 2040196, and ONR grant N00014-18-1-2193.

**References**

1. Bubeck, S. et al. Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv preprint arXiv:2303.12712 (2023).
2. Cattan, A. et al. Scico: Hierarchical cross-document co-reference for scientific concepts. Automated Knowledge Base Construction 2021.
3. Chan, J. et al. Solvent: A mixed-initiative system for finding analogies between research papers. In Proceedings of the ACM on Human-Computer Interaction 2 (2018), 1–21.
4. Chu, J.S.G. et al. Accelerating historical progress in large fields of science. In Proceedings of the National Academy of Sciences 118, 41 (2021).
5. Garcia-Villar, C. A critical review on atomics: Can we measure the social impact factor? Insights into Impacting 12, 1 (2021), 1–10.
6. Gil, Y. Will AI write scientific papers in the future? AI Magazine (2021).
7. Head, A. et al. Augmenting scientific papers with just-in-time, position-sensitive definitions of terms and symbols. In Proceedings of the 20th CHI Conf. on Human Factors in Computing Systems.
8. Hope, T., Chan, J., Kittur, A., and Shahaf, D. Accelerating innovation through analogy mining. In Proceedings of the 23rd ACM SIGKDD Conf. on Knowledge Discovery and Data Mining (2017).
9. Hope, T. et al. Scisight: Combining faceted navigation and research group detection for COVID-19 exploratory scientific search. In EMNLP (2020).
10. Hope, T. et al. Extracting a knowledge base of mechanisms from COVID-19 papers. In NAACL (2021).
11. Hope, T. et al. Scaling creative inspiration with fine-grained functional facets of product ideas. In CHI (2022).
12. Horvitz, E. Principles of mixed-initiative user interfaces. In Proceedings of the SIGCHI Conf. on Human Factors in Computing Sys. (1999), 159–166.
13. Horvitz, E. The future of biomedical informatics: Bottlenecks and opportunities. In Biomedical Information: Future Applications in Health Care and Biomedicine. E.H. Shortliffe, J.J. Cimino, et. Al. (Eds.), Springer (2021).
14. Horvitz, E. et al. The Lumiere project: Bayesian user modeling for inferring the goals and needs of software users. In Proceedings of the Conf. on Uncertainty in AI (1998), 256–265.
15. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. Nature 586, 7873 (2021), 583–589.
16. King, R.B. Functional genomic hypothesis generation and experimentation by a robot scientist. Nature 427, 6971 (2004), 247–252.
17. Kuhn, T.S. The Structure of Scientific Revolutions. Volume 111. University of Chicago Press (1970).
18. Lahav, D. A search engine for discovery of scientific challenges and directions. In AAAI (2022).
19. Langley, P. et al. Scientific Discovery: Computational Explorations of the Creative Processes. MIT Press (1987).
20. Lo, K. et al. S2ORDER: The Semantic Scholar open research corpus. In Proceedings of ACL (2020).
21. Lohr, S. What ever happened to IBM’s Watson. The New York Times 16, 7 (2021).
22. Murthy, S. Accord: A multi-document approach to generating diverse descriptions of scientific concepts. In EMNLP (2022).
23. Myure, S., Cohan, A., and Hope, T. Multi-vector models with textual guidance for fine-grained scientific document similarity. NAACL (2022).
24. Naik, A. Literature-assisted clinical outcome prediction. NAACL (2022).
25. Nori, H. et al. Capabilities of GPT-4 on medical challenge problems. arXiv preprint arXiv:2303.13375 (2023).
26. Nuzzo, R. et al. How scientists fool themselves—and how they can stop. Nature 526, 7572 (2015), 183–189.
27. Nye, B. et al. Understanding clinical trial reports: Extracting medical entities and their relations. In AMIA Annual Symposium Proceedings 2021, 485.
28. OpenAI. GPT-4 technical overview. 2022.
29. Portenoy, J. Bridger: Toward bursting scientific filter bubbles and boosting innovation via novel author discovery. CHI 2022.
30. Pyzer-Knapp, E.D. et al. Accelerating materials discovery using artificial intelligence, high performance computing and robotics. npj Computational Materials 2, 1–8 (2016).
31. Shah, C. and Bender, E.M. Situating search via automated analysis of interests and activities. In Proceedings of the 38th Annual Intern. ACM SIGIR Conf. on Research and Development in Information Retrieval (2005), 449–456.
32. Thagard, P. The Cognitive Science: Explanation, Discovery, and Conceptual Change. MIT Press (2012).
33. Thamba, An. and Guntherman, R.B. For Watson, solving cancer wasn’t so elementary: Prospects for artificial intelligence in radiology. Academic Radiology 29, 2 (2022), 312–314.
34. Valihora, D.A. Finding cultural holes: How structure and culture diverge in networks of scholarly communication. Sociological Science 1 (2014), 221.
35. Wang, Y. et al. Domain-specific pretraining for vertical search: Case study on biomedical literature. In Proceedings of the 27th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining (2021), 3247–3257.
36. Weid, D.S. and Bansal, G. The challenge of crafting intelligible intelligence. Communications of the ACM 62, 8 (2019), 70–79.

Tom Hope (tomh@alema.org) is an assistant professor at The Hebrew University of Jerusalem, Israel, and a Research Scientist at The Allen Institute for AI, Seattle, WA, USA.

Doug Downey is a director of Semantic Scholar Research at The Allen Institute for AI, Seattle, WA, USA, and professor at Northwestern University, Evanston, IL, USA.

Daniel S. Weid is chief scientist and general manager for Semantic Scholar at The Allen Institute for AI and Emeritus Professor at the University of Washington.

Oren Etzioni is founding CEO of the Allen Institute for Artificial Intelligence and Emeritus Professor at the University of Washington.

Eric Horvitz is Technical Fellow and chief scientific officer at Microsoft, Redmond, WA, USA.