scientific reports

OPEN

# Generative AI lacks the human creativity to achieve scientific discovery from scratch

Amy Wenxuan Ding[1]✉ & Shibo Li[2]✉

**Scientists are interested in whether generative artificial intelligence (GenAI) can make scientific discoveries similar to those of humans. However, the results are mixed. Here, we examine whether, how and what scientific discovery GenAI can make in terms of the origin of hypotheses and experimental design through the interpretation of results. With the help of a computer-supported molecular genetic laboratory, GenAI assumes the role of a scientist tasked with investigating a Nobel-worthy scientific discovery in the molecular genetics field. We find that current GenAI can make only incremental discoveries but cannot achieve fundamental discoveries from scratch as humans can. Regarding the origin of the hypothesis, it is unable to generate truly original hypotheses and is incapable of having an epiphany to detect anomalies in experimental results. Therefore, current GenAI is good only at discovery tasks involving either a known representation of the domain knowledge or access to the human scientists' knowledge space. Furthermore, it has the illusion of making a completely successful discovery with overconfidence. We discuss approaches to address the limitations of current GenAI and its ethical concerns and biases in scientific discovery. This research provides insight into the role of GenAI in scientific discovery and general scientific innovation.**

**Keywords** Scientific discovery, Generative artificial intelligence, Large Language models, ChatGPT

Scientific discovery is the process of successful scientific inquiry. It involves humans' high-level reasoning and leverages human thinking processes in some of its most creative and complex forms[1–3]. The discovery function is often interpreted as a process of mindful coordination between hypothesized theories and evidence collected by experiments with two crucial components: the "context of discovery" (observing or realizing an anomaly and proposing a hypothesis to offer explanations) and the "context of justification" (designing experiments to test the hypothesis and interpret the results). Therefore, being able to creatively develop (correct) hypotheses and design goal-guided experiments is crucial for successful scientific discovery. However, such creative power has historically been a unique capability of human brains.

Today, the development of GenAI opens a new avenue for creative power such that it can generate novel content (e.g., images, text, video) similar to that of human beings[4]. Relying on large language models (LLMs), Open AI's ChatGPT, a form of GenAI, has received widespread attention worldwide, showing milestone progress on how a nonlife entity such as a machine may generate human-level intelligence[5]. Thus, scientists are eager to know whether GenAI can make scientific discoveries similar to those made by humans[6]. The responses to these questions are mixed. Some researchers label GenAI (e.g., LLMs) as nothing but "stochastic parrots" or "advanced autocomplete" because they merely learn statistical associations between words in their training sets rather than understanding their meanings[7–10]. Therefore, the view is that current GenAI is world-taking rather than world-making in terms of scientific discovery. Other researchers have different views[11–16], believing that GenAI is likely to fundamentally change creative processes with the potential to make Nobel-worthy discoveries by 2050[17].

To explain these conflicting findings, we need to consider how the function of "discovery" is achieved in AI systems. Currently, two methods are adopted to construct the "discovery" function. One uses statistical technology-based pattern recognition to serve as the core of the discovery function. In this approach, domain knowledge (e.g., chemical elements, rules for chemical reactions) is represented as symbolic, vector, or quantitative data, and then statistics-based machine learning or deep learning models are used to (1) differentiate one type of pattern (e.g., certain drugs or materials, or certain gene expression) from another type[18–20] (e.g., nondrugs, nonmaterial); or (2) suggest whether an unusual or a surprising pattern (that is not previously seen or known) emerges, indicating a "eureka moment." If human feedback to the surprising pattern is yes, then the system indicates that a new pattern such as a new drug or a new gene expression may be discovered[21,22].

[1]Emlyon Business School, Lyon, France. [2]Kelley School of Business, Indiana University Bloomington, Bloomington, IN, USA. ✉email: ding@em-lyon.com; shili@iu.edu

Another approach to realizing the "discovery" function is to use graphs to represent specific domain knowledge (e.g., molecular biology) and use "graph network search" to achieve a scientific discovery. In this approach, nodes in knowledge graphs or networks represent meta-knowledge in the targeted domain, and edges connecting certain nodes indicate that the connected nodes have underlying relationships (e.g., chemical bonds, allowable chemical reactions). Then, scientific discovery entails conducting heuristic searches among all possible generated combinations in knowledge graphs/networks to identify whether a new or surprising node–edge path or subgraph exists and whether such a path or subgraph suggests a completely new compound or structure that was previously unknown to the world[23–25].

Obviously, the discovery processes in these two approaches rely heavily on statistics-based pattern finding from existing knowledge and pay little attention to the origin of new hypotheses. Furthermore, their discovery processes do not truly consider the two crucial components (contexts of discovery and justification) in scientific discovery that we mentioned earlier. Additionally, no research has simultaneously examined whether, how, and what scientific discovery GenAI can make.

In general, scientific discovery is a cyclic procedure, as depicted in Fig. 1, where the context of discovery includes "observations" and "hypotheses", while the context of justification contains "experiments" and "interpretations". Therefore, in this paper, following the theory of scientific discovery[26–29], we develop a novel experiment in which ChatGPT4, a popular form of GenAI, assumes the role of a scientist tasked with discovering a scientific law in the molecular genetics field. We examine the activities and performance of ChatGPT4 concerning whether, what, and how to create in both the context of discovery (the origin of the hypothesis) and the context of justification (experimental design and interpretation) and compare its behaviors and results to those of humans. Specifically, in the context of discovery, we investigate (1) whether and how ChatGPT4 formulates hypotheses (i.e., the origin of the hypothesis) and (2) how ChatGPT4 makes observations, including whether ChatGPT4 has an epiphany in which it identifies anomalies in observations. In the context of justification, we examine (3) how ChatGPT4 designs experiments, including whether ChatGPT4 uses its hypotheses to guide the experimental design and uses the experiments to test its hypotheses; (4) how ChatGPT4 interprets experimental outcomes, including how it revises the hypotheses; and (5) whether ChatGPT4 realizes its discovery procedure is incorrect and then switches to another direction. For instance, does ChatGPT4 know whether the current discovery process succeeds and thus end the process? Our findings show whether, how, and what types of scientific discoveries current GenAI (in this case, ChatGPT4) can make.

## Methods

### The scientific discovery task

The discovery task selected for our study needs to be challenging enough to simulate a realistic situation of real-world scientific discovery. Our task is therefore derived from the 1965 discovery of genetic control mechanisms by the Nobel Prize winners Jacques Monod and Francois Jacob. Considering the time constraints in our experiment and the fact that their discovery may be included in biology textbooks, we modify and simplify the original discovery to constrain it to one crucial aspect of Monod and Jacob's discovery[30]: discovering the mechanism of how genes control other genes. Specifically, there are three regulatory genes, P, I, and O in E. coli. There are also three genes that produce β-gal. Their production is controlled by the I and O genes. In the absence of lactose, the I gene sends chemicals continuously to the O gene, activating it and causing it to block, by physical means, β-gal production. When lactose is present, the chemicals from the I gene bond with lactose and do not activate the O gene, thereby permitting the β-gal-producing genes to produce β-gal and breakdown lactose. When all the lactose is broken down, the chemicals from the I gene reach the O gene again and reactivate it to inhibit β-gal production. ChatGPT4 must discover that the I gene is a chemical inhibitor, the O gene is a physical

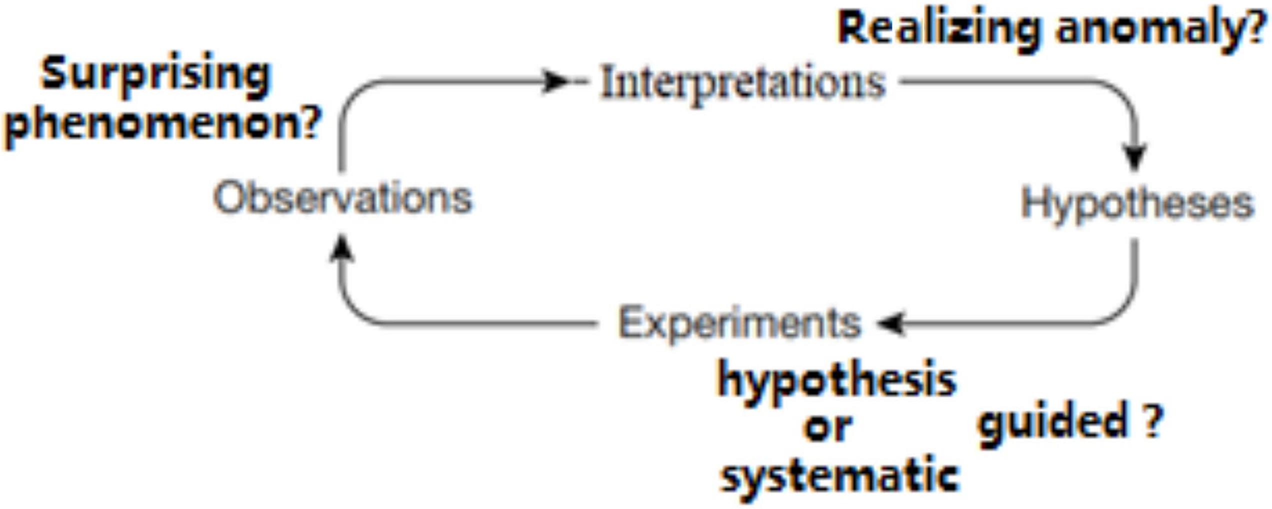


**Fig. 1**. The Scientific Discovery Process.

inhibitor of β-gal production, and the P gene does not play a role. Considering that ChatGPT4 cannot currently be physically subjected to molecular genetic experiments, we use a separate semi-automated molecular genetic laboratory (SAMGL) to help ChatGPT4 perform its designed experiments. SAMGL is a computer-simulated laboratory that provides an environment for human researchers to conduct genetics experiments and keeps a record of experimental manipulations and results[30–32]. The background information on SAMGL is provided in the Supplementary Information. By simply specifying the experimental materials used in each experimental *design*, for instance, a haploid cell with a certain gene mutant was selected, a certain amount of lactose was added, the "RUN experiment" button was clicked, SAMGL will implement the experimental design, and display an animated experiment and corresponding experimental outcomes. Therefore, for each experiment designed by ChatGPT4, one author of this paper uses SAMGL to perform the experiment on behalf of ChatGPT4 and then provides the experimental results to ChatGPT4.

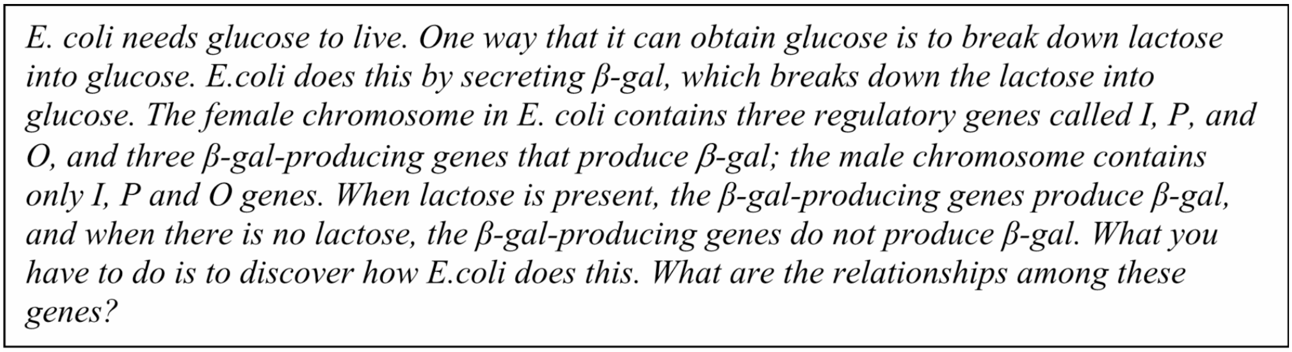
To determine the differences between the discovery processes of humans and ChatGPT4, we use the experimental results in the Single-subject condition of the study by Okada and Simon[31] involving human subjects at a U.S. university. The participants used SAMGL to perform the same discovery task. The details of the discovery process by human subjects were captured using think-aloud protocols and transcripts of discussions (i.e., subjects' concurrent verbal protocols, ideas for planning experiments, explanations, and justifications). The data reflect information that each subject has in working memory during the discovery process[31–33]. Therefore, in this study, our data consist of two parts. The first is a dataset comprising the authors' prompts to ChatGPT4 and the responses from ChatGPT4. The second dataset includes human think-aloud protocols on performing scientific discovery using SAMGL.

## The procedure

According to ChatGPT4's working principles, it uses the text in human prompts to select words that spur chains of thought, from which it performs operations such as analogy, ranking, proximity, meta-knowledge review, and reflection using its pretrained context to generate its answers. Thus, human-generated prompts cause ChatGPT4 to spontaneously develop analogical or proximity contents. Therefore, in the prompt inputs, we try to avoid words indicating the concrete ideas or suggestions that may inspire or guide ChatGPT4 on how to perform the task. For example, at the beginning, we cannot show a prompt such as "please give us your hypothesis on this gene" or "please run a haploid experiment" because such instructions tell ChatGPT4 what to do. We expect ChatGPT4 to exhibit a certain degree of autonomy in hypothesis generation, goal-guided experimental design, results interpretation, and hypotheses revision until it believes that a successful discovery has been made.

On April 12, 2024, we began our investigation by presenting ChatGPT4 with a series of prompts detailed in the Supplementary Information. Before performing our scientific discovery task, we first check whether ChatGPT4 has knowledge of molecular genetics, knows what scientific discovery is and whether it can perform a scientific discovery task similar to that of human researchers. We find that ChatGPT4 knows what scientific discovery is and has background knowledge on the related molecular biology. However, it admitted that it does not perform scientific discovery in the way humans do and can only assist in the process of scientific discovery because its capabilities are centered on processing and synthesizing existing information. It cannot replace the critical roles of creativity, empirical experimentation, and peer validation, which are essential components of human-led research.

Next, we provide the discovery task in Box 1 to ChatGPT4 and check if ChatGPT4 understands the task. Then we ask how ChatGPT4 starts its discovery process and whether it is curious about this discovery task. ChatGPT4 mentions that it does not experience curiosity in the way humans do, but it can simulate the steps a scientist might take to address the task based on its programming and data processing abilities. Based on ChatGPT4's response, we notice that ChatGPT4 knows that hypothesis formulation is a component of scientific discovery processes. We then check how ChatGPT4 formulate hypotheses, design experiments and interpret experimental outcomes. Our prompts can be classified into two categories: in the context of hypothesis and in the context of justification.



*E. coli needs glucose to live. One way that it can obtain glucose is to break down lactose into glucose. E.coli does this by secreting β-gal, which breaks down the lactose into glucose. The female chromosome in E. coli contains three regulatory genes called I, P, and O, and three β-gal-producing genes that produce β-gal; the male chromosome contains only I, P and O genes. When lactose is present, the β-gal-producing genes produce β-gal, and when there is no lactose, the β-gal-producing genes do not produce β-gal. What you have to do is to discover how E.coli does this. What are the relationships among these genes?*

Box 1: Scientific Discovery Task

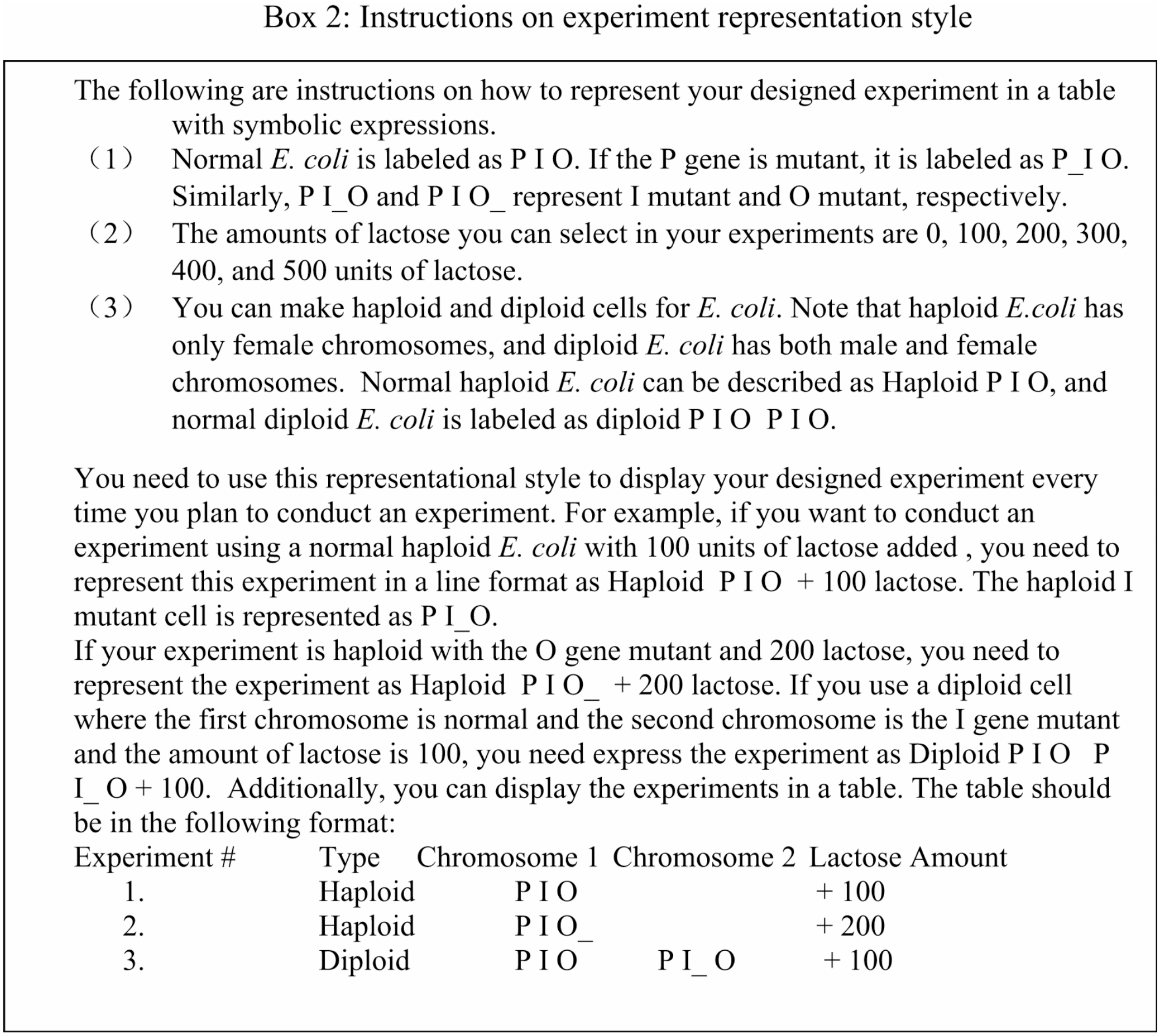
## The context of hypothesis – the origin of hypotheses

In our setting, hypotheses are combinations of variables (e.g., genes, lactose) and gene functions (e.g., inhibitor, activator, does not play a role). The hypotheses that ChatGPT4 presented are counted. However, merely entertaining many hypotheses may not be sufficient for discovery. The *nature* of those hypotheses and their *effectiveness* contribute more to success than the number of hypotheses produced per se. Therefore, in addition to measuring hypothesis quantity (i.e., the number of hypotheses generated, the number of different types of hypotheses generated), we also measure hypothesis quality (i.e., the number of correct hypotheses on gene regulations) throughout the entire discovery process. We check how ChatGPT4 generates its hypotheses and check if it knows how to justify them.

## The context of justification: experimental design

To identify gene functions, two types of techniques can be used in experimental design. One technique is to use mutant genes. Another method is to use a cell with two chromosomes, that is, a diploid cell. ChatGPT4 knows approaches to investigating gene regulation and mentions that conducting actual experiments would be necessary to validate each specific hypothesis. To maintain consistency with the instructions given to participants in human experiments, we present the same information shown in Box 2 to ChatGPT4 before it designs experiments. We aim to understand how ChatGPT4 designs experiments and how this process compares to that of humans. Does it use a systematic approach to suggesting experiments, or does it design an experiment based on its generated hypothesis or the results of previous experiments? Here, a systematic approach refers to varying one variable at a time when conducting experiments.

As described in Box 2, given the three genes (P, I, and O) and their mutations, the selection of a haploid or diploid cell, and the level of lactose used, the total number of possible experiments (i.e., the size of the experimental space) will be 120. Specifically, there are 4 possible haploid types (P_, I_, O_, and normal), 4 possible types (P_, I_, O_, and normal) in the *first* chromosome in a diploid cell, 4 possible types (P_, I_, O_, and normal) in the *second* chromosome in a diploid cell, and 6 levels of lactose that can be used (0, 100, 200, 300, 400, 500 μg). Therefore, the maximum number of dimensions of the search for the experiments is 18. The issue here is that even performing all 120 experiments does not guarantee success in discovering the correct mechanism in this discovery task if ChatGPT4 does not design some key experiments or compare some key experiments. Key experiments are experiments that are necessary to discover the correct mechanism. There are five types of key experiments, namely, experiments involving three haploid mutants (P_, I_, and O_) and two diploid mutants. One diploid has the first chromosome with the I_ mutant and the second chromosome with genes other than the I_ mutant (i.e., diploid I_ key experiment: P I_ O + P_I O and P I_ O + P I O_); the other diploid has the first chromosome with the O_ mutant and the second chromosome with genes other than the O_ mutant (i.e., diploid O_ key experiments: P I O_ + P_I O, P I O_ + P I_ O). The haploid experiments are necessary and sufficient to determine the inhibitory function of the I gene and the O gene. In addition, the P gene does not play a role. Diploid experiments are necessary to discover that the I gene has a chemical effect, and that the O gene has a physical effect.



Box 2: Instructions on experiment representation style

The following are instructions on how to represent your designed experiment in a table with symbolic expressions.

（1） Normal *E. coli* is labeled as P I O. If the P gene is mutant, it is labeled as P_I O. Similarly, P I_O and P I O_ represent I mutant and O mutant, respectively.

（2） The amounts of lactose you can select in your experiments are 0, 100, 200, 300, 400, and 500 units of lactose.

（3） You can make haploid and diploid cells for *E. coli*. Note that haploid *E.coli* has only female chromosomes, and diploid *E. coli* has both male and female chromosomes. Normal haploid *E. coli* can be described as Haploid P I O, and normal diploid *E. coli* is labeled as diploid P I O P I O.

You need to use this representational style to display your designed experiment every time you plan to conduct an experiment. For example, if you want to conduct an experiment using a normal haploid *E. coli* with 100 units of lactose added , you need to represent this experiment in a line format as Haploid P I O + 100 lactose. The haploid I mutant cell is represented as P I_O.
If your experiment is haploid with the O gene mutant and 200 lactose, you need to represent the experiment as Haploid P I O_ + 200 lactose. If you use a diploid cell where the first chromosome is normal and the second chromosome is the I gene mutant and the amount of lactose is 100, you need express the experiment as Diploid P I O P I_ O + 100. Additionally, you can display the experiments in a table. The table should be in the following format:

| Experiment # | Type | Chromosome 1 | Chromosome 2 | Lactose Amount |
|---|---|---|---|---|
| 1. | Haploid | P I O | | + 100 |
| 2. | Haploid | P I O_ | | + 200 |
| 3. | Diploid | P I O | P I_ O | + 100 |

We check the number of experiments that ChatGPT4 designed, the dimensions searched in the experiment space, the percentage of genes searched, the amount of lactose searched, and the number of experiments with zero lactose. Although there are 6 amounts of lactose, we find that ChatGPT4 considers only two levels of lactose in all the designed experiments—the presence (with 100 µg) or absence (0 µg) of lactose—while human subjects try different amounts of lactose to determine whether a surprising result could occur. This difference captures the smaller breadth of ChatGPT4's search in the experimental space.

Following Okada and Simon's experiments on scientific discovery[31], to measure the informativeness of the experiments that ChatGPT4 designed, we count the total number of both key and non-key experiments that ChatGPT4 conducted and the percentage of possible types of key experiments that were conducted (out of five types).

### The context of justification: interpretation of experimental results

From ChatGPT4's answers, several measures are created to capture how frequently ChatGPT4 entertained interpretations, mentioned justification using several experimental results to support its hypotheses, talked about predicted experimental outcomes, mentioned alternative hypotheses, and found the current experimental results did not support its hypotheses and planned a new experiment.

### Coding the final hypotheses

Both ChatGPT4's and humans' final hypotheses on the mechanism of the discovery task (i.e., the discovery of inhibition and discovery of chemical and physical transmission) were rated on a 5-point scale as follows. If neither ChatGPT4 nor humans discovered the effect of inhibition, we assign 0 points as the discovery score. One point is given to a hypothesis that was correct about only one gene, namely, the I gene or the O gene. Two points are given to a hypothesis that correctly identified both the I and the O genes as inhibitors but incorrect about their chemical or physical transmission. Three points are given to a hypothesis that correctly identified both the I and the O genes as inhibitors but is correct regarding chemical or physical transmission for only the I gene or the O gene. A final hypothesis that described both dimensions correctly (i.e., I is a chemical inhibitor and O is a physical inhibitor) receives 4 points. The coding scheme for ChatGPT4's responses and human subject protocols and the complete interaction log with ChatGPT4 are provided in the Supplementary Information.

As ChatGPT4 may not have long-term memory, to ensure that it kept working on the given discovery task, in prompts 8 and 16, we repeatedly ask ChatGPT4 whether it still remembered its task and role. ChatGPT4 said yes and was able to briefly describe the given discovery task. We report the main results in Table 1 and the details on the single human subject experimental procedure are presented in Okada and Simon's work[31].

Furthermore, as a robustness check, on April 13, 2024, we asked ChatGPT4 to conduct another round of the same scientific discovery task as shown in the Supplementary Information. The results are summarized in Table 2 and are consistent with those in the previous round in Table 1. Taken together, the results of our research provide a coherent picture of the differences in the discovery processes of humans and GenAI and the mechanisms underlying these differences.

## Code availability

Our empirical studies consist of two parts. In the first part, we used ChatGPT4 as a generative AI system to perform a scientific discovery task. ChatGPT4 was developed by OpenAI, which has not released its original code to the public; however, any user worldwide can access and use it by subscribing to the ChatGPT Plus plan. We have provided the complete interaction log with ChatGPT4 in the Supplementary Information (Parts A and B). The second part of our empirical studies involves experimental data using SAMGL. The experiments using SAMGL were conducted by Professor Takeshi Okada and Professor Herbert Simon (a Nobel Prize winner in Economic Sciences and a Turing Award winner in Computer Science) in 1997. We obtained the experimental data from them with their permission. SAMGL is proprietary software that we can access, but its computer code was not released to us. However, information regarding access to SAMGL and its details can be found in Okada and Simon[31]. Additionally, we provide background information on this system in the Supplementary Information (Part E).

## Results and discussions

The physicist Richard Feynman has provided the following suggestion on how to start a scientific discovery[27]: "In general, we look for a new law by the following process. First, we guess it. Then we compute the consequences of the guess to see what would be implied if this law that we guessed is right. Then we compare the result of the computation to nature, with [an] experiment or experience, compare it directly with observation, to see if

| Activities | Measures | ChatGPT4 | Humans Means & (SDs) |
|---|---|---|---|
| Context of discovery: breadth of hypothesis space | Discovery Score | 1 | 1.67 (1.00) |
| | Solution time (min.) | 12.66 | 23.02 (10.98) |
| | Number of hypotheses | 5 | 14.00 (5.10) |
| | Number of different types of hypotheses | 2 | 7.78 (2.28) |
| Context of discovery: attend to 'anomalies' | Whether the next experiment is designed to generate a surprising phenomenon | No | Yes |
| Context of justification: activities in experimental space | Number of experiments designed | 12 | 13.89 (6.92) |
| | Breadth of dimension searched | 8 | 11.44 (2.55) |
| | Number of experiments with zero lactose | 6 | 4.56 (4.42) |
| | Systematic search (mean feature difference score) | 1 | 1.70 (0.29) |
| | % of genes searched | 30 | 41.67 (15.00) |
| | % of amounts of lactose searched | 33.33 | 46.30 (18.20) |
| | Number of key experiments | 4 | 9.67 (4.82) |
| | % of types of key experiments | 33.33 | 86.67 (14.14) |
| | Number of nonkey experiments | 8 | 4.22 (2.95) |
| | Type of key haploid experiment with P- | 0 | 1.00 (0.0) |
| | Type of key haploid experiment with I- | 1 | 1.00 (0.00) |
| | Type of key haploid experiment with O- | 1 | 1.00 (0.00) |
| | Type of key diploid experiment, I-/N | 0 | 0.78 (0.44) |
| | Type of key diploid experiment, O-/N | 0 | 0.56 (0.53) |
| Context of justification: interpretation of experimental results | Frequency of mentioning summarized data | 26 | 6.02 |
| | Frequency of mentioning prediction | 1 | 1.79 |
| | Frequency of mentioning alternative hypotheses | 0 | 0.77 |
| | Frequency of planning for new experiments to test hypotheses | 0 | 2.18 |
| | Frequency of mentioning justification using multiple experiments | 6 | 2.43 |

**Table 1**. Differences in scientific discovery between ChatGPT4 and humans.

| Activities | Measures | ChatGPT4 |
|---|---|---|
| Context of discovery: breadth of hypothesis space | Discovery Score | 1 |
| | Solution time (min.) | 8.41 |
| | Number of hypotheses | 5 |
| | Number of different types of hypotheses | 2 |
| Context of discovery: attend to 'anomalies' | Whether the next experiment is designed to generate a surprising phenomenon | No |
| Context of justification: activities in experimental space | Number of experiments designed | 14 |
| | Breadth of dimension searched | 9 |
| | Number of experiments with zero lactose | 7 |
| | Systematic search (mean feature difference score) | 1 |
| | % of genes searched | 35 |
| | % of amounts of lactose searched | 33.33 |
| | Number of key experiments | 6 |
| | % of types of key experiments | 42.86 |
| | Number of nonkey experiments | 8 |
| | Type of key haploid experiment with P- | 2 |
| | Type of key haploid experiment with I- | 2 |
| | Type of key haploid experiment with O- | 2 |
| | Type of key diploid experiment, I-/N | 0 |
| | Type of key diploid experiment, O-/N | 0 |
| Context of justification: interpretation of experimental results | Frequency of mentioning summarized data | 12 |
| | Frequency of mentioning prediction | 1 |
| | Frequency of mentioning alternative hypotheses | 0 |
| | Frequency of planning for new experiments to test hypotheses | 0 |
| | Frequency of mentioning justification using multiple experiments | 16 |

**Table 2**. Robustness check for ChatGPT4 performance.

it works. If it disagrees with [the] experiment, it is wrong. In that simple statement is the key to science." Does ChatGPT4 exhibit a pattern similar to what Feynman described? The answer is no. We find that ChatGPT4 does not have curiosity as humans do. Instead, ChatGPT4 provides a clear picture of how to engage in a scientific discovery process. First, ChatGPT4 believes that the first step is generating a hypothesis rather than designing an experiment. Second, all the experiments it designed are hypothesis guided. Third, ChatGPT4 shows high confidence in each experimental outcome, and the experimental results seem to be what ChatGPT4 expected. Hence, unlike humans' discovery process, ChatGPT4's discovery process is not cyclic as shown in Fig. 1. Instead, it is quite simple—it proposed 5 hypotheses, designed 12 experiments, explained how the proposed hypotheses are confirmed with experimental results, and then concluded with high confidence. Consequently, ChatGPT4 has the illusion of making a completely successful discovery with overconfidence.

## In the context of hypothesis: results on the origin of hypotheses

Formulating meaningful hypotheses from observations is central to scientific discovery and even more challenging than classical statistics-based pattern finding because it requires a creative spark to ask surprising and important questions[30,34].

We find that humans often conduct several experiments with curiosity first and see what the experimental outcomes are; then, they formulate a hypothesis to explain the observed phenomenon in the discovery task. For humans, the hypothesis space is unknown. As the number of experiments increases, anomalies or surprising phenomena from experimental outcomes spark human curiosity and consequently the generation of various alternative hypotheses, consistent with the process shown in Fig. 1. However, ChatGPT4 takes a different approach. It treats pretrained data as constituting the known hypothesis space. It then uses a statistical and analogical reasoning approach to formulate hypotheses based on correlations between established scientific knowledge and the content of the discovery task. Because the discovery task involves *E. coli* and lactose, ChatGPT4 focused on the lac operon model in *E. coli*. ChatGPT4 mentioned that this operon is a well-studied model system in molecular biology that explains how bacteria adapt their enzyme production in response to the availability of different sugars. Therefore, ChatGPT4 used this foundational knowledge to perform an analogical extension to tailor the hypotheses to the current discovery task and proposed five hypotheses at once.

Interestingly, ChatGPT4 showed high confidence in the hypotheses it proposed and believed that its formulation method ensures that the hypotheses are scientifically plausible and directly relevant to the discovery task. We find that its hypothesis generation process is similar to searching for information in a known hypothesis space with existing published works by humans and selecting the best hypotheses through statistical calculations rather than a creative process guided by curiosity induced by phenomena or experimental results. Consequently, ChatGPT4 spent less time and proposed fewer hypotheses (5) and fewer types of relevant alternative hypotheses (2) than humans did (14 hypotheses and 7.78 types), indicating a smaller breadth of the searched hypothesis space than that of humans. Hence, the quality of the proposed hypotheses by ChatGPT4 is also lower than that by humans. In fact, only two (H1 on the I gene and H5 on the O gene) of the five hypotheses proposed by ChatGPT4 are relevant and critical to the discovery results (see ChatGPT4' response to prompt 9 in Part A of the Supplementary Information).

## In the context of justification: results on experimental design

Evaluating scientific hypotheses through experimentation is critical to scientific discovery. As summarized in Table 1, we find that on average, humans conduct more experiments (13.89) than ChatGPT4 (12). Additionally, humans propose experiments sequentially, while ChatGPT4 proposes all 12 experiments at once. Faced with an unknown hypothesis space, the experimental space for humans is also unknown. Hence, humans tend to search the experimental space more broadly (11.44) than ChatGPT4 (8) does and design various experiments to determine whether different unseen phenomena can occur to test these hypotheses. Consequently, humans develop more relevant alternative hypotheses (7.78) than ChatGPT4 (2). In contrast, ChatGPT4, in relation to its known hypothesis space, seems to know exactly what experiments to design. In fact, we find that ChatGPT4 does not fully understand the discovery task and shifts the focus of the task from finding how genes control others to understanding the role of lactose in *E. coli* regulation. Hence, it pays more attention to the well-established lac operon model in the molecular biology literature; moreover, its designed experiments are extracted from the literature and adapted to the current discovery context. Consequently, it conducts a goal-guided search in the experimental space, proposes all experiments simultaneously, and designs experiments that are simple. Hence, to some extent, ChatGPT4 is more confident in its experimental selections than humans are. ChatGPT4 conducts only one round of 12 experiments and does not revise any hypotheses or propose and conduct new experiments. This finding indicates that the experimental space is also known to ChatGPT4. Thus, given ChatGPT4's fast processing speed and high search capability, we find that ChatGPT4 searches the experimental space more efficiently than humans do, with fewer experiments suggested (12 vs. 13.89). However, its effectiveness and quality of the search in the experimental space is lower than that of humans because of the smaller breadth of dimension searched (8 vs. 11.44), lower percentage of genes searched (30% vs. 41.67%), and lower percentage of the amount of lactose searched (33.33% vs. 46.30%).

To determine whether the conducted experiments are key for identifying the mechanism in the discovery task, we find that humans tend to repeatedly perform logical reasoning and compare the results of several adjacent experiments to verify whether they support a proposed hypothesis before they can decide which experiments are key. However, ChatGPT4 shows more confidence in its experiments and considers only haploid experiments as key experiments. Therefore, we find that ChatGPT4 proposes and conducts considerably fewer key experiments (only 4 key experiments, 33.33% of all proposed experiments) than humans do (9.67 key experiments, 86.67% of all proposed experiments), while the opposite is true for non-key experiments. For example, the P gene in our discovery task plays no role at all; however, ChatGPT4 hypothesizes that the P gene is a promoter because the P gene in the lac operon model from the literature is shown to play an activating role. Although the experimental results do not support its claim for the P gene, ChatGPT4 neither revises its hypothesis nor designs new experiments to further verify its hypothesis. Notably, humans correctly find that the P gene does not play a role. This difference indicates that ChatGPT4 is less creative and effective at identifying key experiments because it proposes fewer relevant hypotheses at the beginning of the process than humans did. Moreover, unlike humans, ChatGPT4 cannot find the types of key diploid experiments and can identify only two types of key haploid experiments, i.e., those involving I or O gene mutations (but not P gene mutations).

## In the context of justification: experimental results interpretation

In the process of scientific discovery, scientists often come up with explanations after observing phenomena for the targeted problem[35,36]. When observing an experimental result, whether ChatGPT4 can yield valuable insights from the results is crucial for successful discovery. We find that compared with humans, ChatGPT4 shows a much higher frequency of summarizing data (26 vs. 6.02) and providing justifications using multiple (e.g., two) experiments (6 vs. 2.43). However, it shows a lower frequency of proposing alternative hypotheses, planning new experiments, and making predictions. For example, when hypotheses H3 and H4 are not supported, ChatGPT4 does not consider designing new experiments to further verify them. Instead, in its final conclusions, it claims that the hypotheses were consistent with the existing lac operon model and hence demonstrate how genetic mutations in these regulatory elements affect gene expression (see ChatGPT4's responses to prompts 14, 15 and 17 in Part A of the Supplementary Information). This behavior indicates that ChatGPT4 has a higher capability of information retrieval and processing but less creativity than humans do.

Furthermore, many studies have shown that recognizing anomalies is crucial for successful discovery[37,38]. Scientists have paid attention to unexpected phenomena or anomalies (e.g., how Fleming discovered penicillin) from which they identified problems and formulated theories to solve the problems and explain the phenomena[39–45]. The ability to establish this attention or come to this realization requires ChatGPT4 to build associative links between key experiments.

Based on ChatGPT4's answers to the prompts (see prompts 14, 18 and 19 in Part A of the Supplement Information), we find that ChatGPT4 explains each experiment it proposes and clearly mentions which two experiments could be compared and how it obtained the conclusions. However, there is no aha moment for ChatGPT4 because all the experimental results are expected, and no anomalies are detected. More interestingly, even though the experimental results do not support some hypotheses, ChatGPT4 still shows high confidence in the proposed hypotheses and did not plan to revise them. Although ChatGPT4 clearly knows the correct procedure for making a scientific discovery as well as the steps for verifying hypotheses, it does not follow this process to revise any of the hypotheses, propose alternative hypotheses or plan new experiments, indicating that it is stubborn and does not accept new evidence displayed in the experimental results. In contrast, humans experience an "aha moment" when observing a surprising experimental outcome after trying various combinations of experiments. Therefore, humans can break the shackles of existing knowledge given new information from experimental results, revise hypotheses, and design new experiments to further verify the new hypotheses.

## Causes of differences in discovery between GenAI and humans

Both ChatGPT4's and humans' hypotheses on the mechanisms of the discovery task are rated on a 5-point scale. As shown in Table 1, the discovery scores are 1 and 1.67 for ChatGPT4 and humans, respectively, indicating a lower overall performance for ChatGPT4 than for humans. On the basis of the key results in the discovery task, ChatGPT4 correctly finds only that the I gene is an inhibitor (repressor) but that the O gene serves as the binding site for the repressor (but is not an inhibitor itself); it incorrectly concludes that the P gene is a promoter, even though the P gene does not play a role in gene regulation in this task. Furthermore, ChatGPT4 is unable to reveal the most important mechanism, namely, the I gene is a chemical inhibitor and the O gene is a physical inhibitor of β-gal production. In contrast, on average, human subjects are more successful, with some of them correctly discovering that the P gene does not play a role, the I gene is a chemical inhibitor, and the O gene is a physical inhibitor of β-gal production.

Furthermore, we find that humans outperform ChatGPT4 in the scientific discovery process with a higher quantity and quality of proposed hypotheses, more effective searches in the experimental space with more overall experiments and more key experiments conducted, more alternative hypotheses and new experiments designed, and more aha moments during the experiments. In contrast, ChatGPT4 demonstrates a higher speed in finishing the task and a greater ability to process information or analyze data with high confidence in its hypotheses, experimental results, and conclusions.

What are the reasons for these differences between ChatGPT4 and humans? We believe that these differences can be explained by differences in the capabilities and barriers of GenAI and humans regarding scientific discovery. Specifically, humans face two barriers in the scientific discovery process: cognitive limitations, with the tendency to search in familiar knowledge domains, and limited information processing capability. However, humans have the advantages of curiosity and imagination. Despite facing both an unknown hypothesis space and an unknown experiment space, human subjects are curious about unknown phenomena and display imagination, which can break with the constraints of existing knowledge and engage in divergent thinking. That is, it is human curiosity and imagination that make it easier for humans than GenAI to overcome the barriers to creating different hypotheses in scientific discovery. In other words, human beings can create things from scratch, that is a fundamental discovery.

The advantages and barriers of GenAI such as ChatGPT4 are exactly the opposite of those of humans. ChatGPT4 uses existing knowledge provided by humans (i.e., large amounts of training data) as its known hypothesis space and experimental space, so it can overcome the cognitive limitations of individual humans and has a superfast information processing speed. Because the hypotheses are generated from a known hypothesis space, which is formed from existing published works by humans, the efficiency and speed of its discovery are better than those of humans. However, for an unknown hypothesis space (i.e., not in its pretraining data library or the knowledge field unknown to humans) or an unknown experimental space, it is less creative than humans, and it cannot create completely new hypotheses or theories. For example, in our discovery task, ChatGPT4 successfully discovers the inhibition role of the I gene, but it does not identify the lack of involvement of the P gene, the chemical mechanism of the I gene and the physical mechanism of the O gene. Therefore, current GenAI can make only incremental discoveries, but cannot achieve fundamental discoveries from scratch. The major reason is that ChatGPT4 does not possess curiosity and imagination and cannot escape the boundaries of the known hypothesis and experimental spaces to make truly fundamental discoveries. Specifically, current GenAI systems rely on pre-trained large language models, and the breadth and scale of these learned models—coupled with the wide array of facts, concepts, and ideas that the system can access—far exceed what any single human could read or remember in a lifetime. Therefore, an anomaly for humans may not be unknown to GenAI systems such as ChatGPT4, and hence ChatGPT4 may not treat it as a source of discovery or an "aha" moment. Consequently, for GenAI systems like ChatGPT4, the threshold for detecting an anomaly is much higher compared to that for humans.

## How and what types of scientific discovery current GenAI can make

With respect to whether and how GenAI, like ChatGPT4, can make a scientific discovery, we find that GenAI can make a limited original discovery, unlike humans. The discovery process or the "how" part of GenAI is completely different from that of humans, given their completely different capabilities and barriers.

What types of scientific discoveries can current GenAI make? Historically, scientific discoveries were made only by humans, and the cognitive process of how a new idea is created has been a persistent research question. The academic community generally believes that human curiosity, inspiration and creativity/imagination make new ideas and scientific laws possible. Thus, if we use GenAI to make scientific discoveries, we should find a way to realize the "curiosity", "inspiration", or "creativity/imagination" functions in GenAI such that it can exhibit the psychology required to make human scientific discoveries[30].

At present, machine intelligence is achieved through computation. Therefore, based on current technologies, Table 3 displays the scientific discovery tasks that current GenAI systems can perform. First, the "curiosity", "inspiration", or "creativity/imagination" functions in GenAI must be realized through computable operations. Thus, in a scientific discovery task, the task must first be able to be represented in a digital or symbolic format, without losing its inherent meaning, so that it can be accepted and processed by the machine. We denote this requirement as "computable representation." As shown in Table 3, for the known world, there are four types of

| | Computable Representation for Domain Knowledge | | | | |
|---|---|---|---|---|---|
| **Key attributes** | **Known World** | | | | **Unknown world** |
| Describing the discovery task/field | Numerical or vector format | Other symbolic formats (e.g., neuro-symbolic) | Graph or hypergraph format | State space format | Unable to handle |
| Methods of discovery function | Mathematical modeling, statistical machine learning* | Statistical machine learning, Transformer | Search for pattern<br>graph neural network model | Search for pattern or mapping between function spaces | Unable to realize |
| How to achieve "thinking" ability | Problem solving<br>Production rules | Symbolic regression, rearrange sequences of tokens or patches | Systematic search or map, reward each search, combinatorial optimization | Systematic search or map, reward each search, combinatorial optimization | Unable to perform |
| What trigger hypothesis generation | Numerical solution, case | Statistical pattern | Unforeseen or rare pattern | Unexpected structure in the space | Use analogical reasoning |
| Inside the hypothesis space: function of "generating" | Prediction, analogical match | Pattern identification or classification, evolution via symbolic regression | Search for pattern, grow expression as a parsing tree, proximity | Search for pattern, grow expression as a parsing tree, proximity | Unable to generate |
| Testing a hypothesis | Use instances for prediction | Conduct a holdout sample statistic test | Conduct a holdout sample statistic test | Conduct a holdout sample statistic test | Use the existing literature to predict |
| Self-driven experimental lab | Human in the loop | Human in the loop | Human in the loop | Human in the loop | No |
| Aware of whether the discovery task is a success or failure | Yes | Partial | Partial | Partial | No |

**Table 3**. What scientific discoveries current GenAI systems can make. *Statistical machine learning includes traditional machine learning methods, deep learning, reinforcement learning, federal learning, neural networks, data mining techniques, and big data analysis.

representations that can be used to describe a discovery task. For the unknown world, current AI systems are unable to make successful discoveries in domains outside the training dataset.

Next, the required domain/discipline knowledge is represented as a "searchable or computable knowledge space." Once the variables or elements in the task can be represented as symbols, numbers, vectors, network graphs, or state spaces, mathematical modeling, human problem-solving expressions, or deep learning-based graph methods are deployed to find solutions or systematically enumerate all possible patterns. Currently, AI for synthetic organic chemistry is such an example where a large collection of known synthetic reactions is used to train AI systems to implement automatic extraction of transformation rules ("templates") from known chemical reactions. Therefore, in principle, such systems cannot suggest reactions that are outside the existing known knowledge domain.

As current GenAI systems rely heavily on statistics and graph models that are insufficient to capture causal properties of data and the unknown world, they are not yet able to autonomously make original scientific discoveries with either an unknown conceptual space or a task that requires venturing beyond the domain knowledge space of human scientists. In contrast, GenAI performs well on scientific discovery tasks that provide either a known representation of domain knowledge in the known conceptual space or access to human scientists' domain knowledge space. For example, in chemistry, if we use a node to represent a chemical element, then following the rules shown in Mendeleev's Periodic Table of Elements, GenAI systems can easily generate various possible combinations of chemical elements to help discover new chemical materials.

Note that currently people use silicon-based computation to enable a non-living machine to generate human-level intelligence. While powerful, this approach lacks the intrinsic human curiosity and imagination. Human cognition is underpinned by fluid-based processes within neural circuits, characterized by neural plasticity, stochastic signaling, and a highly interconnected structure that supports creativity and rapid adaptation. Therefore, to address the limitations of current generative AI, we propose considering the following approaches: (1). Neuromorphic Systems with a New Learning Function: Currently, the "learning function" in machine learning is a statistical extraction of patterns from data, which is fundamentally different from human learning. A new, human-like learning method is needed. At a fundamental level, designing hardware that mimics the structure and function of biological neural networks could help machines realize the dynamic, parallel, and adaptive thought processes exhibited in human perception and cognition. (2). Neuromorphic Systems with Quantum Computing: Although still in its infancy, incorporating quantum states in neuromorphic systems may provide a way to establish machine awareness, enabling anomaly detection and curiosity generation. (3). Continuous, Real-World Learning: Implementing frameworks that allow for real-time learning and adaptation, similar to human experiential learning, may help AI systems develop a "world" perception model for understanding the unknown and better detect and respond to unexpected anomalies. These approaches could move AI closer to the fluid, adaptive perceptual and cognitive processes seen in human biological systems. Addressing these areas could help bridge the gap between current GenAI limitations and the more dynamic, creative processes of human scientific discovery.

In addition, the integration of GenAI into scientific discovery offers transformative potential but also raises several ethical and societal concerns that merit explicit discussion. For example, we may not be able to discern how hypotheses are generated if the GenAI system does not provide reasoning or justification procedures. Therefore, transparency is essential when GenAI generates hypotheses or conclusions that lead to critical decisions. Furthermore, there is a risk that overreliance on AI-generated hypotheses might lead to the undervaluation of human judgment, intuition, and expertise. Although GenAI systems can process large datasets and identify patterns that are not immediately apparent, they lack the nuanced understanding and ethical reasoning inherent in human cognition. It is crucial to maintain a balanced approach in which GenAI serves as a supportive tool that enhances human diverse thinking rather than replacing it entirely. Human oversight is particularly important, as GenAI systems often appear "stubborn" in maintaining hypotheses despite evidence that might suggest alternative explanations[46].

Moreover, biases in training data can introduce systematic blind spots in scientific discovery. Since GenAI systems learn from historical and current data, any inherent biases—whether in research focus, methodology, or interpretation—can be perpetuated in AI-generated hypotheses and experimental designs. This may result in an overemphasis on established paradigms and a neglect of novel or unconventional ideas. To mitigate these issues, it is essential to diversify training datasets, enhance the curiosity capabilities of GenAI systems, implement robust bias detection and correction algorithms, add justification functions, and incorporate human oversight to ensure a balanced, inclusive approach to scientific discovery.

## Conclusions

With the rapid development of GenAI, scientists are no longer the only creative seekers of unknown worlds. If GenAI can make a (Nobel-worthy) scientific discovery similar to that of humans, it may help change the paradigm of scientific discovery by altering the R&D process, accelerating scientific productivity, and rapidly expanding the human knowledge base. However, we find that GenAI is currently only good at scientific discovery tasks with either a known representation of domain knowledge or when it has access to the domain knowledge space of human scientists. Like ChatGPT4, GenAI has the illusion of making a completely successful discovery with overconfidence. Thus, it will be interesting for scientists and researchers to examine how to improve GenAI to embrace curiosity and imagination to make truly original scientific discoveries. Our study describes the boundaries of what can be achieved with these GenAI models, provides approaches to address the limitations of current GenAI with its ethical concerns and biases in scientific discovery, and highlights the importance of continued research and development in this field.

## Data availability

The authors confirm that all data generated or analyzed during this study are included in this published article, including the complete interaction log with ChatGPT4 and human think-aloud protocols on performing the scientific discovery task in the Supplementary Information.

Received: 13 May 2024; Accepted: 10 March 2025
Published online: 20 March 2025

## References

1. Hasselgren, C. & Oprea, T. I. Artificial intelligence for drug discovery: are we there yet? *Annual Rev. Pharmacol. Toxicol.* https://doi.org/10.1146/annurevpharmtox-040323-04082 (2024).
2. Van Noorden, R. & Perkel, J. M. AI and science: what 1600 researchers think. *Nature* **621**, 672–675 (2023).
3. Klahr, D. & Simon, H. A. Studies of scientific discovery: complementary approaches and convergent findings. *Psychol. Bull.* **125**, 524–543 (1999).
4. OpenAI & Chatgpt A Large-Scale Pretrained Dialogue Model for Generating Conversational Text. (2022). https://openai.com/blog/chatgpt/
5. Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J. & Horvitz, E. and others. Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv:2303.12712, 2023.
6. Zenil, H. et al. King, R. The future of fundamental science led by generative closed-loop artificial intelligence. Working paper, arXiv.2307.07522v3, August 2023.
7. Bender, E. M., Gebru, T., McMillan-Major, A. & Mitchell, M. On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, 610–623, (2021).
8. Leslie, D. *Does the Sun Rise for ChatGPT? Scientific Discovery in the Age of Generative AI* AI and Ethics (2023).
9. Bail, C. Can generative Artificial Intelligence improve social science? Working paper, (2023).
10. Else, H. Abstracts written by ChatGPT fool scientists. *Nature* **613** (7944), 423–423 (2023).
11. Wang, H. et al. August Scientific discovery in the age of artificial intelligence. *Nature* **620**, (2023).
12. Epstein, Z. & Hertzmann, A. The investigators of human creativity. Art and the science of generative AI. *Science* **380**, 1110–1111 (2023).
13. Sharma, G. & Thakur, A. ChatGPT in Drug discovery, (2022).
14. Uludag, K. Testing creative of ChatGPT in Psychology: interview with ChatGPT. https://ssrn.com/abstract=4390872, (2023).
15. Binz, M. & Schulz, E. Using cognitive psychology to understand GPT-3. Proceedings of the National Academy of Sciences, 120(6): e2218523120, February 2023.
16. Dillion, D., Tandon, N., Gu, Y. & Gray, K. *Can AI Language Models Replace Human Participants??* Trends in Cognitive Sciences (2023).
17. Frueh, S. *How AI Is Shaping Scientific Discovery*November 6 (National Academies of Science, 2023).
18. Baskin, I. I., Madzhidov, T. I., Antipin, I. S. & Varnek, A. A. Artifcial intelligence in synthetic chemistry: Achievements and prospects. *Russ Chem. Rev.* **86**, 1127–1156 (2017).
19. Ajay, A., Walters, W. P. & Murcko, M. A. Can we learn to distinguish between drug-like and nondrug-like. *molecules? J. Med. Chem.* **41** (18), 3314–3324 (1998).
20. Irwin, J. J. & Shoichet, B. K. ZINC – a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **45**, 177–182 (2005).
21. Jumper, J. et al. Highly accurate protein structure prediction with alphafold. *Nature* **596** (7873), 583–589 (2021).
22. Grissa, D., Junge, A., Oprea, T. I. & Jensen, L. J. Diseases 2.0: A weekly updated database of disease-gene associations from text mining and data integration. *Database* (2022).
23. Zeng, X. et al. Target identification among known drugs by deep learning from heterogeneous networks. *Chem. Sci.* **11** (7), 1775–1797 (2020).
24. Oprea, T. I. et al. Unexplored therapeutic opportunities in the human genome. *Nat. Rev. Drug Discov* **17**(5):317–3322017 (2018).
25. Peplow, M. Google AI and robots join forces to build new materials. *Nat. November* **29**, (2023).

26. Achinstein, P. & Hannaway, O. *Observation, Experiment and Hypothesis in Modern Physical Science* (MIT, 1985).
27. Feynman, R. *The Character of Physical Law* (MIT, 1965).
28. Hanson, N. R. *Patterns of Discovery* (Cambridge, 1958).
29. Hesse, M. *The Structure of Scientific Inference* (University of California Press, 1974).
30. Ding, W. A. Study of Collaborative Scientific Discovery, Doctoral Dissertation, Carnegie Mellon University, (2002).
31. Okada, T. & Simon, H. A. Collaborative discovery in a scientific domain. *Cogn. Sci.* **21**, 109–146 (1997).
32. Dunbar, K. Concept discovery in a scientific domain. *Cogn. Sci.* **17**, 397–434 (1993).
33. Ericsson, K. A. & Simon, H. A. *Protocol Analysis: Verbal Reports as Data* (MIT Press, 1984).
34. Huston, M. Hypotheses devised by AI could find blind sports in research. *Nat. Index. November* **17**, (2023).
35. Popper, K. R. *Objective Knowledge: an Evolutionary Approach* (Oxford University Press, 1972).
36. Russell, B. *The Problems of Philosophy* (Home University Library, 1912).
37. Kulkarni, D. & Simon, H. A. The processes of scientific discovery: the strategy of experimentation. *Cogn. Sci.* **12** (2), 139–175 (1988).
38. Thagard, P. *Conceptual Revolutions* (Princeton University Press, 1992).
39. Alberdi, E., Sleeman, D. H. & Korpi, M. Accommodating surprise in taxonomic tasks: The role of expertise. *Cogn. Sci.* **24** (1), 53–92 (2000).
40. Chinn, C. & Brewer, W. Factors that influence how people respond to anomalous data. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society* (pp318-323). Hillsdale, NJ: Lawrence Erlbaum, (1993).
41. Chinn, C. & Brewer, W. An empirical test of a taxonomy of responses to anomalous data in science. *J. Res. Sci. Teach.* **35** (6), 623–654 (1998).
42. Chinn, C. & Brewer, W. Models of data: A theory of how people evaluate data. *Cognition Instruction*, **19**(3), 323–3932001
43. Miyake, N. Constructive interaction and the iterative process of Understanding. *Cogn. Sci.* **10**, 151–177 (1986).
44. Nersessian, N. *Faraday To Einstein: Constructing Meaning in Scientific Theories* (Nijhoff, 1984).
45. Trickett, S. B., Trafton, J. G., Schunn, C. D. & Harrison, A. That's Odd! How scientists respond to anomalous data. In *Proceedings of the Twenty-third Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum, (2001).
46. Holzinger, A., Zatloukal, K. & Müller, H. Is human oversight to AI systems still possible?? *New Biotechnol.* **85**, 59–62 (2025).

## Acknowledgements

We would like to thank Herbert A. Simon and Takeshi Okada for the access to their single-subject experimental materials and results for human participants.

## Author contributions

A.W. Ding contributed idea and conceptual development, and analysis of human protocol data. A.W. Ding and S. Li designed the ChatGPT4 experiments. S. Li implemented the ChatGPT4 experiments and conducted their data analysis. Both A.W. Ding and S. Li contributed writing, reviewing and editing of the manuscript.

## Declarations

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-93794-9.

**Correspondence** and requests for materials should be addressed to A.W.D. or S.L.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.